DOI:10.20040/j.cnki.1000-7709.2023.20222455

考虑协变量的随机森林降水融合算法在长江流域的应用

徐 志[1],牟亚莉[2],梁犁丽[1],王贺龙[3]

(1. 中国长江三峡集团有限公司科学技术研究院,北京 101199; 2. 中国水利水电科学研究院,北京 100038; 3. 浙江省水利河口研究院,浙江 杭州 310020)

摘要:高质量、高时空分辨率的降水数据对水文、气象等领域的研究具有重要意义。目前遥感、再分析降水数据应用普遍,但存在分辨率低、精度不确定性大等问题。对此,提出了考虑协变量的随机森林降水融合算法,对CMA、CN05、ERA5、GLDAS、TRMM、IMERG、PERSIANN七套降水产品进行融合,并选取长江流域的三个典型子流域(金沙江、三峡区间、鄱阳湖)进行随机森林融合数据(RFF)的效果检验。结果显示,对于降水产品准确性,随机森林融合数据精度相较于原始降水产品有所提升。对于不同降水事件的准确性评估,随着降雨强度增加,各降水产品风险评分(T_{TS})评分均呈减小趋势,RFF的T_{TS}评分优于原始降水产品。通过考虑协变量的随机森林模型对降水产品进行数据融合,可以提升降水数据的精度及不同降水事件发生的可靠性,为水文模拟等提供支持。

关键词:降水融合;随机森林;协变量;长江流域;精度评估

中图分类号:[TV11] **文献标志码**:A **文章编号**:1000-7709(2023)08-0001-04

1 概况

长江流域经济发达、人口众多,地形复杂,气候多变。流域年降水量时空分配不均,呈现出暴雨突发和洪水灾害频发的显著特点[1]。降水监测精度的提升对于流域水资源管理、水生态保护、防洪减灾等具有重要意义[2-4]。降水融合在长江流域的研究多基于统计关系,随着地学大数据发展,基于机器学习回归方法被应用于降水融合研究中[5,6],但对区域气象、地理因子的影响考虑不足。因此,本文以金沙江、三峡区间、鄱阳湖三个流域为典型流域,分析了流域降雨的主要影响因素,应用考虑协变量影响的随机森林模型,对CMA、CN05、ERA5、GLDAS、TRMM、IMERG、PERSIANN七套降水产品进行融合,并参照实测数据对融合结果进行验证分析,以期为长江流域高时空分辨率降雨数据的获取提供参考。

2 研究数据与方法

2.1 研究数据

为涵盖长江流域的基本特征,充分考虑长江流域地面气象站点分布及流域特点,分别选择长江流域上游、中游和下游的金沙江流域、三峡区间流域和鄱阳湖流域三个子流域作为研究区域,其位置和气象站点分布见图1。长江流域的气象站点分布不均,呈现明显的东部密集、西部稀疏的特点。

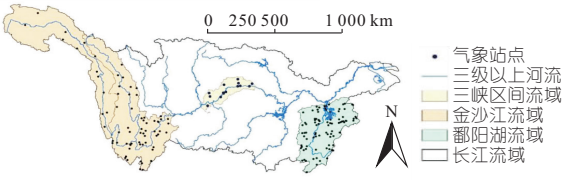


图1 研究区域及气象观测站点分布

Fig. 1 Distribution of study area and meteorological observation stations

采用的地面基准降水资料来自国家气象信息中心168个地面站点1984~2013年的实测日降水数据,金沙江、三峡区间、鄱阳湖三个流域分别有82、17、86个站点,站点位置分布见图1。采用的降水产品包括CMA、CN05、ERA5、GLDAS、TRMM、IMERG、PERSIANN七套数据集,具体信息见表1。

收稿日期:2022-11-21,**修回日期**:2023-02-28

基金项目:国家自然科学基金项目(U2040212);中国长江三峡集团有限公司科研项目(202103429,202203050)

作者简介:徐志(1990-),男,博士、工程师,研究方向为水文学及水资源,E-mail:xu_zhi1@ctg.com.cn

通讯作者:梁犁丽(1982-),女,博士、正高级工程师,研究方向为水文学及水资源,E-mail:liang_lilii@ctg.com.cn

表 1 降水数据时空分辨率、时间范围和来源

Tab. 1 Spatial and temporal resolution, time range and source of precipitation data

数据类型	数据名称	空间分辨率/(°)	时间分辨率	时间范围	数据来源
再分析数据	CMA	0.25	逐时	1984～2020 年	国家气象信息中心(NMC)
	CN05	0.25	逐时	1961～2020 年	http://data.cma.cn/search/uSearch.html? keywords=CRA
	ERA5	0.25	逐日	1979～至今	欧洲中期天气预报中心(ECMWF)https://ccrc.iap.ac.cn/search? title=CN05.1
	GLDAS	0.25	3 h	2000～至今	美国国家航空航天局(NASA)https://ldas.gsfc.nasa.gov/gldas
遥感数据	TRMM	0.25	逐日	1998～2019	美国国家航空航天局(NASA)https://disc.gsfc.nasa.gov/datasets? keywords=TRMM&-page=1
	IMERG	0.10	逐日	2000～至今	美国国家航空航天局(NASA)https://disc.gsfc.nasa.gov/datasets? keywords=TRMM&-page=1
	PERSIANN	0.25	逐日	1983～至今	美国国家海洋和大气管理局(NOAA) https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00854/html

2.2 研究方法

2.2.1 协变量获取方法

在降水产品融合的过程中,为提高模型模拟精度,考虑将影响降水的气象、地理要素及包含地理和季节性信息的经纬度和时间作为模型协变量参与融合模型构建。本文先采用双线性插值法将各套降水产品空间分辨率统一插值到 0.1°,然后确定区域内实测站点经纬度所对应的中心格点位置(图 2),提取每个格点处的气象、地理、时间要素作为协变量。选取的气象要素包括降水、气温、风速、比湿及向下长短波辐射。为保持降水结构的大尺度空间依赖性,除考虑中心格点处的降水外,还将中心格点四周(上下左右四个方向)的降水同时作为协变量;地理要素包括高程、坡向。协变量类型见表 2。

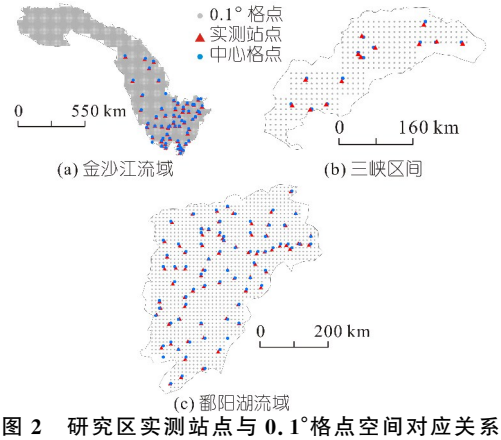


图 2 研究区实测站点与 0.1°格点空间对应关系

Fig. 2 The spatial correspondence between the observation station and the 0.1° grid point in the study area

2.2.2 考虑协变量的随机森林模型

随机森林(RF)是一种机器学习算法,具有优秀的处理变量间复杂非线性关系的能力,同时能最大程度降低过拟合问题,操作简单,功能强大[2]。以提取的协变量作为模型输入值,中心格点降水及周边格点降水提取自七套降水产品(共 35 个),其他协变量共 11 个(表 2),共计 46 个协变量作为随机森林模型的输入,以对应的气象站点的实测降水作为模型输出值,构建随机森林融合模型。

表 2 协变量类型表

Tab. 2 Table of covariable types

变量类型	变量名称	变量简写	说明
气象要素	中心格点降水	pre	实测站点对应的各套降水产品的中心格点降水数据及上、下、左、右四个方向的格点降水数据
	上方格点降水	upre	
	下方格点降水	dpre	
	左方格点降水	lpre	
	右方格点降水	rpre	
	气温	temp	CMFD 再分析数据集,空间分辨率 0.1°
	风速	wind	
	比湿	shum	
	降水速率	prec	
	向下短波辐射	srad	
	向下长波辐射	lard	
地理要素	高程	dem	资源环境科学与数据中心全国 DEM 1 km 数据,经双线性插值法插值至 0.1°;坡向通过 ArcGIS 计算得到
	坡向	aspect	
辅助要素	时间	year、month、day	
	经度	lon	
	纬度	lat	

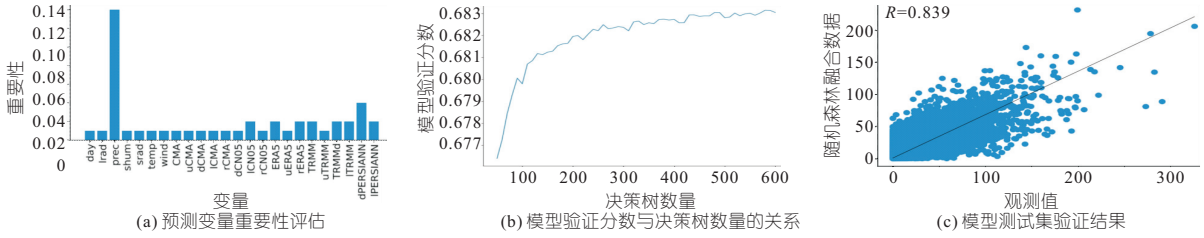
2.2.3 评估指标

根据各气象站点的经纬度确定对应的数据格点,提取对应格点的日降水量,计算日累计降水量,进而利用 6 个指标对降水产品进行评估,分析降水产品与观测降水数据之间的相关性和差异。选取的评估指标主要包括 Pearson 相关系数(r)、均方根误差(R_{RMSE})。其中,r 用来反映站点实测值与降水产品的一致性程度;R_{RMSE} 用来反映降水产品的降水序列整体误差水平和波动情况。用来评价降水产品对降水的捕捉能力的指标包括风险评分(T_{TS})、误报率(F_{FAR})、命中率(P_{POD})及偏差评分(B_{BIAS})。评估指标相关描述见文献[7]。

3 结果分析

3.1 协变量敏感性

图 3(a)为各套降水产品的随机森林融合模型构建中各变量的重要性评估,该重要性评估由 Random Forest 模块的 Feature_importances 函数获取。重要性指标值越大,该变量在预测中越重要。由于协变量较多,仅展示了重要性大于 0 的协变量。由图 3(a)可知,协变量中的降水速率(prec)重要性最高,其余协变量重要性都较低,其中 CN05、ERA5、TRMM、PERSIANN 重要性稍高,其余产品则评分很低。



(a) 预测变量重要性评估　(b) 模型验证分数与决策树数量的关系　(c) 模型测试集验证结果

图 3　模型验证

Fig. 3　Model validation

3.2　模型率定验证

将各套数据集共有的 2000～2018 年的日数据均用于模型,并根据站点数量比例 8∶2 随机划分训练集和测试集,训练集和测试集均有整个时间序列的数据,采取 Bootstrap 方法对训练数据集进行训练。通过 3 折交叉验证的方法确定模型决策树的数量,见图 3(b)。模型决策树到 400 棵后模型验证分数无明显提升,因此选取 400 为模型最终决策树数量(模型参数 n_estimators=400)。

将训练好的随机森林融合模型经测试集进行模型验证,验证结果见图 3(c),其相关系数 R 为 0.839,模型精度较好。

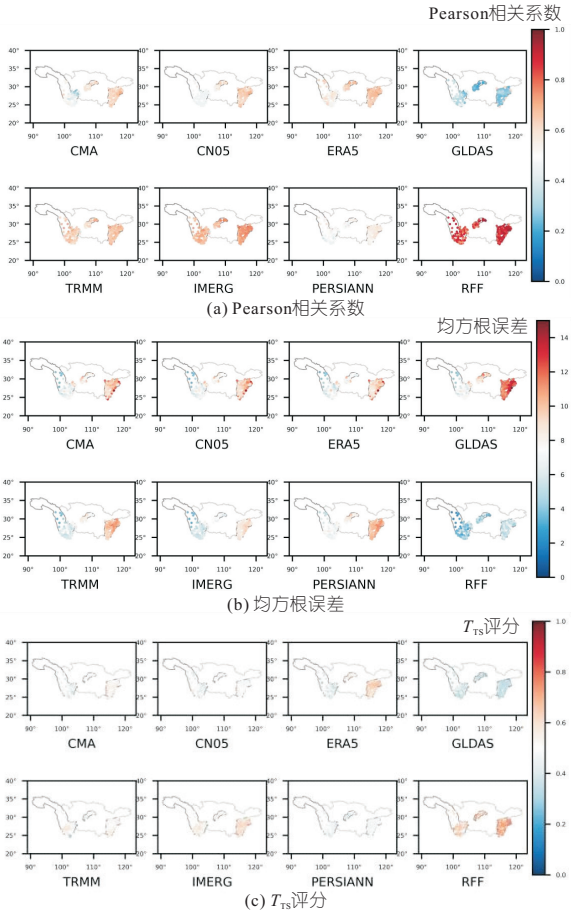
3.3　融合效果评估

3.3.1　降水数据的准确性

根据预测的格点数据提取实测站点对应的降水,并与观测数据进行比较。对比分析各降水数据的精度及一致性特点,由七套降水产品及随机森林融合数据(RFF)的 6 项精度评估指标可知,RFF 的 6 个指标相对于其他降水数据均有所提升。RFF 的 Pearson 值介于 0.75～1,明显高于其他降水产品的评分;RFF 的 R_{RMSE} 值介于 1.2～7.3,三个流域自西向东增加,与我国降水“东多西少”的空间分布一致,且明显小于其他降水数据;RFF 的 T_{TS}、F_{FAR}、B_{BIAS}、P_{POD} 评分效果优于其他降水产品,其中 P_{POD} 值为 1,说明融合数据对于降水事件的命中能力有所提升,能够命中所有降水事件,实测降水事件均能被融合数据捕捉到。

3.3.2　降水事件空间的准确性分析

为了更直观地对比七套降水数据和随机森林融合数据的精度及一致性的空间分布特点,图 4 展示了 Pearson、R_{RMSE}、T_{TS} 三个指标在长江三个子流域的站点分布。由图 4 可知,八套数据的 Pearson 分布特征具有一致性,金沙江流域的下游和三峡区间上游区域的 Pearson 相对较低,鄱阳湖流域的 Pearson 整体较高。在 R_{RMSE} 对比中,八套数据同样表现出相似的空间分布特点,即自西向东逐渐递增,这与长江流域上下游实际降水强度和频次的差异有关。同时,八套数据的 T_{TS}



(a) Pearson相关系数　(b) 均方根误差　(c) T_{TS}评分

图 4　降水产品的 Pearson 相关系数、均方根误差、T_{TS} 评分在金沙江流域、三峡区间、鄱阳湖流域的站点分布

Fig. 4　Distribution of pearson correlation coefficient R_{RMSE}, T_{TS} scores of precipitation products in Jinsha River basin, Three Gorges area and Poyang Lake basin

空间分布具有一致性,与 R_{RMSE} 空间表现一致,金沙江流域评分较低(0.2～0.4),三峡和鄱阳湖流域较高(0.4～0.6)。总体看,随机森林融合数据精度相较于原始的七套降水产品的精度均有所提升。

3.3.3　不同降水事件的发生频率及精准性评估

以 0.1、10、25、50 mm/d 作为阈值将不同降水事件定义为小雨(0.1～10 mm/d)、中雨(10～25 mm/d)、大雨(25～50 mm/d)、暴雨(50～100 mm/d),分别评估八套降水数据对应不同等级降水事件的精准性。

图5为不同的降水事件在长江子流域的发生

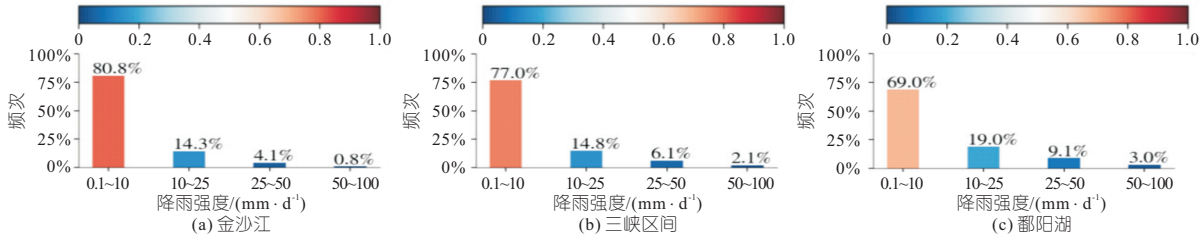


图 5 不同降水事件在长江子流域的发生频次

Fig. 5 Frequency of different precipitation events in each basin

频次。各个子流域降水事件均以小雨为主，小雨在三个流域占比分别为 80.8%、77%、69%，准确地评估小雨事件是衡量降水产品精度的关键。长江流域从上游到下游，小雨事件发生的占比逐渐减小，中到大雨的发生频次逐渐增加。金沙江流域和三峡区间的中雨占比约为 14%，大到暴雨的占比不超过 9%，鄱阳湖的中雨占比接近 20%，大到暴雨的占比达到 12%。由此可见，大雨事件不可忽略，尤其在中、大雨发生频次较高的鄱阳湖流域，降水产品对于该类事件的准确预估同样重要。

由八套数据在不同降水事件下的 T_{TS} 评分情况可知，整体上各降水产品对降水的捕捉能力与降水强度呈反比，不同降水事件的 T_{TS} 评分从高到低依次为小雨、中雨、大雨、暴雨。小雨情况下 T_{TS} 评分均为 1，中雨情况下 T_{TS} 总体评分在 0.6 以上，RFF 优于其他七套降水产品。CMA、CN05、ERA5 的 T_{TS} 评分优于 GLDAS、TRMM、IMERG、PERSIANN，各产品在三个子流域的评分相似；由于大雨、暴雨事件频次较少，导致七套产品的 T_{TS} 评分较低（大部分站点评分集中在 0～0.2，少部分接近 0.4），RFF 的评分介于 0.4～0.6，优于其他产品。金沙江流域大量站点未发生或极少发生暴雨事件，T_{TS} 评分可认定为 0；其他两个子流域，各降水产品的 T_{TS} 有一定差异，与大雨评分表现一致，CMA、CN05、ERA5 的总体评分约为 0.4，高于 GLDAS、TRMM、IMERG、PERSIANN 的评分（均在 0.2 以下）。

4 结论

提出的考虑协变量的随机森林降水融合算法，考虑了区域气象、地理因子的影响，形成降水融合产品，提高了降水产品精度，可为水文模拟提供支持。本文主要进行融合精度指标分析，后续将进一步研究融合算法的误差及融合精度随地理、气候特征的变化规律。

参考文献：

[1] 许冠宇，李琳琳，田刚，等. 国家级降水融合产品在长江流域的适用性评估[J]. 暴雨灾害，2020，39(4)：400-408.

[2] 徐彬仁，魏瑗瑗. 基于随机森林算法对青藏高原TRMM降水数据进行空间统计降尺度研究[J]. 国土资源遥感，2018，30(3)：181-188.

[3] 王文，汪小菊，王鹏. GLDAS月降水数据在中国区的适用性评估[J]. 水科学进展，2014，25(6)：769-778.

[4] 肖楠，叶磊，吴剑，等. 降雨对山丘区小流域洪峰模拟不确定性的影响[J]. 中国农村水利水电，2018，429(7)：35-38.

[5] 尹家波，郭生练，王俊，等. 基于贝叶斯模式平均方法融合多源数据的水文模拟研究[J]. 水利学报，2020，51(11)：1335-1346.

[6] 李运龙，熊立华，闫磊. 基于地理加权回归克里金的降水数据融合及其在水文预报中的应用[J]. 长江流域资源与环境，2017，26(9)：1359-1368.

[7] 夏昕然，田烨，谭伟丽，等. 多种卫星降水产品在中国的精度评估[J]. 水利水电技术(中英文)，2022，53(8)：29-40.

Application of Stochastic Forest Precipitation Fusion Algorithm With Covariates in Yangtze River Basin

XU Zhi[1], MOU Ya-li[2], LIANG Li-li[1], WANG He-long[3]

(1. Science and Technology Research Institute, China Three Gorges Corporation, Beijing 101199, China;
2. China Institute of Water Resources and Hydropower Research, Beijing 100038, China;
3. Zhejiang Institute of Hydraulics & Estuary, Hangzhou 310020, China)

Abstract: The precipitation data with high quality and high spatial and temporal resolution is of great significance to the research of hydrology, meteorology and other fields. At present, remote sensing and reanalysis of precipitation data are widely used, but there are problems such as low resolution, high uncertainty of accuracy, etc. In this paper, a random forest precipitation fusion algorithm considering covariates is proposed to fuse seven sets of precipitation products, namely CMA, CN05, ERA5, GLDAS, TRMM, IMERG and PERSIANN. Three typical sub-basins of the Yangtze River basin (Jinsha River, Sanxiaqujian and Poyang Lake) are selected to test the effect of random forest fusion data (RFF). The results show that for the accuracy of precipitation products, the accuracy of random forest fusion data is improved compared with the original precipitation products. For the accuracy assessment of different precipitation events, with the increase of rainfall intensity, the T_{TS} score of each precipitation product shows a decreasing trend, and the T_{TS} score of RFF is better than the original precipitation product. Data fusion of precipitation products through random forest model considering covariates can improve the accuracy of precipitation data and the reliability of different precipitation events, which provides support for hydrological simulation.

Key words: precipitation fusion; random forest; covariant; Yangtze River Basin; accuracy evaluation