

DOI: 10.20040/j.cnki.1000-7709.2026.20250376

基于多源气象资料与机器学习的降水数据质量控制方法研究

李施颖^{1,2,3,4}, 黄晓龙^{1,2,3,4}, 罗岚心⁵, 吴 薇^{1,2,3}, 吕纯月^{1,2,3}

(1. 四川省气象探测数据中心, 四川 成都 610072; 2. 高原与盆地暴雨旱涝灾害四川省重点实验室, 四川 成都 610072; 3. 川西南(雅安)暴雨实验室, 四川 雅安 625099; 4. 温江国家气候观象台, 四川 成都 611130; 5. 国家气象信息中心, 北京 100081)

摘要: 为提高降水观测数据在公共气象服务中的应用价值, 基于多源气象资料及机器学习模型, 开展了降水数据的质量控制方法研究。研究利用气象站数据、多源融合实况分析产品、气象雷达卫星产品以及 XGBoost 模型, 并通过综合评分, 形成了一种多源协同的降水数据质量控制方法。采用人工植入误差法对该方法的准确性进行检验。结果表明, 随机植入降水异常值时, 错误数据的检出率为 0.85; 按降水量的不同等级植入误差时, 错误数据的检出率为 0.79。经协同质控后, 待检降水数据与邻近气象站的平均误差减小 15%, 相关系数提高至 0.7~1.0。因此, 该方法能有效筛选疑误降水数据, 为气象防灾减灾提供有力的数据支撑。

关键词: 降水; 雷达产品; 卫星产品; 机器学习; 质量控制

中图分类号: TV125; {P413.2⁺1}

文献标志码: A

文章编号: 1000-7709(2026)01-0015-05

Study on Precipitation Data Quality Control Method Based on Multi-source Meteorological Data and Machine Learning

LI Shi-ying^{1,2,3,4}, HUANG Xiao-long^{1,2,3,4}, LUO Lan-xin⁵, WU Wei^{1,2,3}, LV Chun-yue^{1,2,3}

(1. Sichuan Meteorological Observation and Data Center, Chengdu 610072, China; 2. Heavy Rain and Drought-Flood Disasters in Plateau and Basin Key Laboratory of Sichuan Province, Chengdu 610072, China; 3. Southwest Sichuan (Ya'an) Rainstorm Laboratory, Ya'an 625099, China; 4. Wenjiang National Climatology Observatory, Chengdu 611130, China; 5. National Meteorological Information Center, Beijing 100081, China)

Abstract: To enhance the application value of precipitation observation data in public meteorological services, this study investigates a quality control method for precipitation data based on multi-source meteorological data and machine learning models. Based on the meteorological station data, multi-source integrated analysis products, meteorological radar, and satellite products, as well as the XGBoost model, a multi-source collaborative quality control method for precipitation data was developed by comprehensive scoring. The accuracy of this method was evaluated using the artificial error implantation technique. The results indicate that when random precipitation anomalies are implanted, the detection rate of erroneous data is 0.85; When errors are implanted at different precipitation levels, the detection rate of erroneous data is 0.79. After collaborative quality control, the average error between the precipitation data to be tested and neighboring meteorological stations is reduced by 15%, and the correlation coefficient is increased to 0.7-1.0. Therefore, this method can effectively identify suspicious precipitation data, which can provide strong data support for meteorological disaster prevention and mitigation.

Key words: precipitation; radar products; satellite products; machine learning; quality control

收稿日期: 2025-03-03, **修回日期:** 2025-04-19, **接受日期:** 2025-04-26

基金项目: 国家气象信息中心 2023 年“揭榜挂帅”项目(13); 川西南(雅安)暴雨实验室科技发展基金项目(CXN-BYSYSWZD202406)

作者简介: 李施颖(1994-), 女, 工程师, 研究方向为气象资料质量控制及检验分析, E-mail: shiyingli0912@163.com

通讯作者: 黄晓龙(1981-), 男, 正高级工程师, 研究方向为气象资料质量控制、融合与产品评估, E-mail: xiaolong-huang116@163.com

1 引言

随着我国气象事业的发展,气象与水利部门积极推进降水数据共享,以实现跨行业应用。气象站点多分布于城镇,服务天气预报、农业气象等业务^[1];水利站点集中于水库、流域等,用于收集水体气象资料^[2-3],部门间数据应用场景互补。目前,四川省气象部门获取的水文站降水数据为原始观测值,未经质量控制,存在误差。为提高数据可靠性,需进行质量控制。气象部门对地面气象站数据建立了严格的质量控制规范,并通过实时质控确保准确性^[4]。尽管雷达和卫星数据存在一定的不确定性,但其在空间覆盖和实时监测方面的优势使其成为降水监测的重要数据源^[5-6]。此外,多源融合实况降水产品(CMPAS_RT)在降水事件中精度较高^[7],机器学习在降水数据质量提升中也得到广泛应用^[8]。传统降水质控算法依赖单一质控源,在山区孤站降水中,漏检和错检现象频发。为此,多源降水资料协同质控成为解决这一问题的有效途径。本文以地形复杂的四川省为研究区域,结合邻近站点质控方法,引入雷达、卫星及实况降水产品,基于统计与机器学习方法,提出多源协同质量控制技术,通过多源数据之间的相互验证,有效降低单一数据源的不确定性,提升质控的准确性和可靠性。该方法解决了传统质控算法的局限性,推动了跨部门降水数据的深度融合,可为公共气象服务提供高质量的数据支撑。

2 数据与方法

2.1 数据来源

以四川省气象部门获取的 2023 年 1~12 月逐小时原始水文站降水观测资料为质控对象,共 4 377 个水文站。所有参与质控的数据(表 1)均来源于四川省气象探测数据中心,水文站与气象站的空间分布相似,集中于四川盆地,川西高原站

表 1 研究所需数据详细信息

Tab. 1 Detailed information on the data used in this study

数据资料	数据来源	时间 分辨率	空间 分辨率	单位
水利部站点数据	站点	逐小时	站点	mm
气象台站数据	站点	逐小时	站点	mm
CMPAS_RT	地面—雷 达—卫星	逐小时	0.01°	mm
雷达组合反射率	雷达	6 min	0.05°	dBZ
葵花 8 号卫星红外亮温(H8 TBB) 卫星		20 min	0.05°	K
风云 4A 卫星红外亮温(FY4A TBB) 卫星		4~30 min	0.05°	K

点相对稀疏,99%的水文站 10 km 范围内均有气象站点(图 1)。

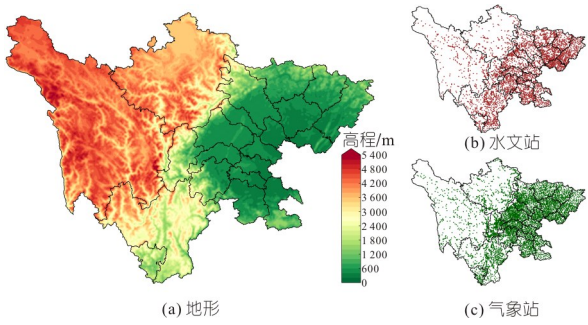


图 1 四川省地形、水文站与气象站空间分布
Fig. 1 The spatial distribution map of topography, hydrological station and meteorological station in Sichuan Province

2.2 多源协同质量控制方法

2.2.1 利用气象站数据质控

采用气象站降水数据对待检水文站降水数据进行质控,步骤如下。

步骤 1 气候学界限值检查就是将降水值超出 0~240 mm/h 的数据判为可疑。

步骤 2 空间一致性检查就是将待检站邻近气象站待检小时及前 3 个小时的降水值组成时间序列,计算 25%、50%、75%分位的值,并计算质控系数 f :

$$f = (P_{QC_A} - P_{QC_{p50}}) / (P_{QC_{p75}} - P_{QC_{p25}}) \quad (1)$$

式中, P_{QC_A} 为待检站待检小时降水值; $P_{QC_{p25}}$ 、 $P_{QC_{p50}}$ 、 $P_{QC_{p75}}$ 分别为邻近站时间序列 25%、50%、75%分位的降水值。

若 $|f|$ 大于判断阈值,则判定为疑误值。

步骤 3 时间一致性检查就是选取待检小时及前 6 个小时的数据,计算待检水文站与最邻近的气象站的协方差 Cov :

$$Cov = \frac{1}{N} \sum (G_i - \bar{G})(R_i - \bar{R}) \quad (2)$$

式中, G_i 、 R_i 分别为水文站和气象站的小时观测值; $\bar{G}$ 、 $\bar{R}$ 均为对应样本均值; N 为样本个数。

若 $Cov < 0$,表示水文站和气象站点降水量变化趋势不一致,则判定为可疑数据。此外,若水文站连续 6 h 降水量无变化且大于 0 mm/h,需结合其他检查结果进一步判定。

2.2.2 利用 1 km 多源融合实况分析产品质控

通过邻近插值将 CMPAS_RT 产品与待检水文站降水数据匹配。将小时降水量分为 6 个等级,依次为 0.1~1.9、2~4.9、5~9.9、10~19.9、20~49.9、 >50 mm/h,分级计算两者之间的差值 d 。计算差值序列平均值 $\bar{d}$ 和标准差 s ,采用 3 倍标准差法判断异常数据:

$$d = G'_i - O_i \quad (3)$$

$$\bar{d} = \frac{1}{N} \sum_{i=1}^N |(G'_i - O_i)| \quad (4)$$

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N [(G'_i - O_i) - \bar{d}]^2} \quad (5)$$

式中, G'_i 为实况网格产品插值到站点的数据; O_i 为水文站点数据; N 为样本总数。

若 d 落在 $[\bar{d} - 3s, \bar{d} + 3s]$ 区间内为正常数据, 否则为可疑数据。

2.2.3 利用雷达和卫星产品质控

选取雷达组合反射率产品, 以小时内的最大值代表该小时的反射率。根据雷达组合反射率与降水的关系^[5], 判定以下情况为可疑数据: ①降水量为 0, 组合反射率大于 25 dBZ。②降水量大于 0, 组合反射率小于 10 dBZ。③降水量小于 1 mm/h, 组合反射率大于 35 dBZ。④降水量大于 1 mm/h, 雷达组合反射率为 10~15 dBZ。

利用 FY4A TBB 和 H8 TBB 产品, 亮温值在 214~265 K^[9] 时判定为有降水发生, 以下情况为可疑数据: ①降水量为 0, 亮温显示有降水。②降水量大于 0, 亮温显示无降水。

2.2.4 利用机器学习方法质控

采用 XGBoost 算法, 选取距离待检站点最近的 8 个气象站作为邻近站, 以邻近站的逐小时降水量和距离为特征值, 预测待检站点降水事件是否发生。模型在测试数据集上预测准确率达到 0.96, 表明构建的模型具有较高的准确性。因此, 对以下情况判定为可疑数据: ①降水量为 0, 模型预测有降水。②降水量大于 0, 模型预测无降水。

2.2.5 综合评分法

对每种质控结果赋予不同的权重分值, 将不同气象资料对待检站质控的结果进行融合, 得出最终的质控结果。综合评分 (S_{Score}) 满分为 100 分, 计算公式为:

$$S_{\text{Score}} = S_1 R_1 + S_2 R_2 + S_3 R_3 + S_4 R_4 + S_5 R_5 + S_6 R_6 \quad (6)$$

式中, S_1 、 S_2 、 S_3 、 S_4 、 S_5 、 S_6 分别为气象站、CMPAS_RT、雷达组合反射率、FY4A TBB、H8 TBB 和机器学习方法的权重分值; R_1 、 R_2 、 R_3 、 R_4 、 R_5 、 R_6 分别为每种质控方法的结果。

判别正确为 2, 可疑为 1, 错误为 0。

各质控方法的权重分值通过 AHP 层次分析法确定, 具体步骤如下。

步骤 1 构建层次结构模型(图 2), 包括目标层、准则层和方案层。

步骤 2 通过两两对比确定各质控方法的权

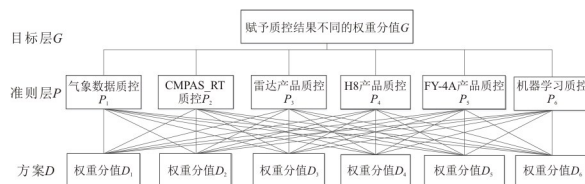


图 2 权重评价层次结构图

Fig. 2 Weight evaluation hierarchy structure

重, 并构建 G-P 判断矩阵(表 2)。

表 2 G-P 判断矩阵

Tab. 2 G-P judgment matrix

G	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	权重
P ₁	1	4	5	5	5	2	0.40
P ₂	1/4	1	2	2	2	1/3	0.12
P ₃	1/5	1/2	1	1	1	1/4	0.08
P ₄	1/5	1/2	1	1	1	1/4	0.08
P ₅	1/5	1/2	1	1	1	1/4	0.08
P ₆	1/2	3	4	4	4	1	0.24

步骤 3 计算判断矩阵各行元素乘积的 n 次方根 $\bar{W}_i$, 并进行归一化处理, 得到权重值。

$$\bar{W}_i = \sqrt[n]{\prod_{j=1}^n a_{ij}} \quad (7)$$

$$W_i = \bar{W}_i / \sum_{i=1}^n \bar{W}_i \quad (8)$$

式中, $\bar{W}_i$ 为判断矩阵第 i 行元素的乘积的 n 次方根; n 为判断矩阵阶数; a_{ij} 为判断矩阵中的元素; W_i 为相应指标权重, 见表 2。

步骤 4 计算最大特征根 $\lambda_{\max}$ 和一致性指标 C_1 , 检验其一致性比例 C_R 。

$$\lambda_{\max} = \sum_{i=1}^n \frac{B_i W}{n W_i} \quad (9)$$

$$C_1 = (\lambda_{\max} - n) / (n - 1) \quad (10)$$

$$C_R = C_1 / R_1 \quad (11)$$

式中, B_i 为判断矩阵的第 i 个行向量。

当 $C_R < 0.1$ 时, 权重结果有效。经计算, $C_R = 0.008 < 0.1$, 满足一致性要求。

最终权重分值为气象站 20 分, CMPAS_RT 6 分, 雷达组合反射率 4 分, FY4A TBB 4 分, H8 TBB 4 分, 机器学习方法 12 分。根据总分判断数据的准确性: >64 分为正确数据, $24 \sim 64$ 分为可疑数据, <24 分为错误数据。总体质控流程见图 3。

2.3 质控结果检验指标

通过人工植入误差检验多源协同质控方法的效果。误差分两种情况植入: ①当站点无降水时, 随机植入 0.5~1.0 mm 的异常值; ②当站点有降水时, 按原降水量的 100%~300% 植入误差。采用错误数据检出率 R 定量评估质控效果:

$$R = m / M \quad (12)$$

式中, m 为正确捕捉到错误或可疑数据的数量;

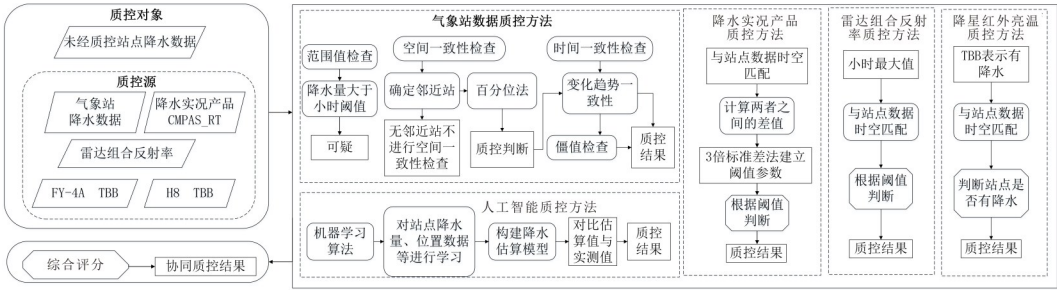


图 3 总体质控流程

Fig. 3 Flowchart of the collaborative quality control method

M 为错误数据总数。

若该质控方法将植入误差后的数据判定为可疑或错误,则认为质控达成。

同时,利用平均误差(M_{ME})、均方根误差(R_{RMSE})、相关系数(C_{COR})来检验质控前后水文站与邻近气象站小时降水量的一致性。

3 结果与分析

3.1 多源协同质量控制方法检验

为检验多源协同质控方法的准确性,对四川省 156 个国家气象站逐小时降水数据人工植入误差,错误数据检出率见表 3。当原站点无降水时,随机植入 0.5~1.0 mm 降水量,错误数据检出率达 85%;当原站点有降水时,随着植入误差的增加,检出率逐渐提高,当误差为原始降水量的 3 倍时,检出率达 84%,平均检出率约为 79%。无降水时(图 4(a)),各站点检出率为 0.77~0.99,表明该方法对虚假降水处理效果较好。有降水时(图 4(b)~(f)),误差增至 300%,所有站点的检出率均高于 0.7,表明该方法对显著误差的识别能力较强。

表 3 多源协同质控方法错误数据检出率

Tab. 3 Error data detection rate of multi-source collaborative quality control method

植入误差情况	错误数据检出率	植入误差情况	错误数据检出率
无降水随机植入(A)	0.85	按降水量的 200%植入(D)	0.80
按降水量的 100%植入(B)	0.73	按降水量的 250%植入(E)	0.82
按降水量的 150%植入(C)	0.77	按降水量的 300%植入(F)	0.84

3.2 质控前后站点数据可用性分析

采用多源协同质控方法对原始水文站降水数据进行质控,将数据分为正确、可疑和错误三类,并分析逐月正确率、可疑率和错误率(图 5)。全年各月正确率均高于 70%,可疑率均低于 30%,错误率不超过 0.2%。无降水时,逐月正确率均高于 99%,表明水文站降水漏报极少。有降水时,正确率存在波动,6 月最高(75%),12 月最低

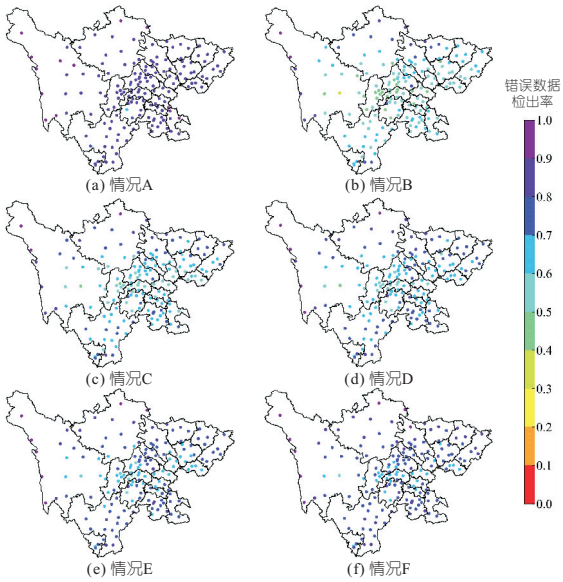


图 4 情况 A~F 的错误数据检出率分布图

Fig. 4 The error data detection rate distribution maps of condition A-F

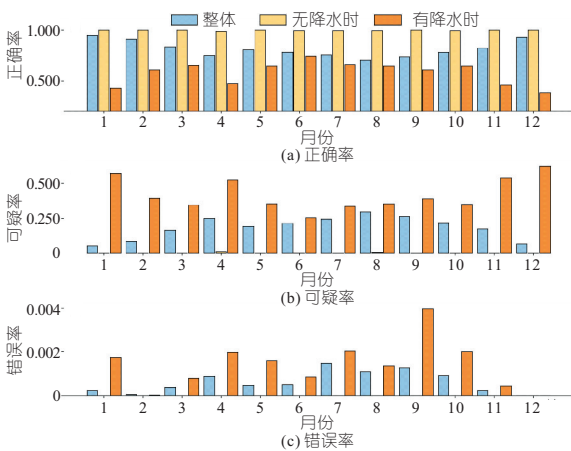


图 5 待检水文站经协同质控后正确率、可疑率、错误率情况

Fig. 5 Correct rate, suspicious rate, error rate after quality control at hydrological stations

(39%)。冬季可疑率较高,主要因为水文站记录少量降水(0.5 mm/h),而气象数据未显示降水。有降水时错误率始终低于 0.4%,表明水文站与气象站数据之间存在显著差异的情况极少。

为检验质控前后水文站与邻近气象站小时降水量的一致性,选取站点距离在 1 km 以内的

1 637 对水文与气象站点进行对比(图 6)。质控前,水文站与邻近气象站点的平均误差为 $-0.7 \sim 0.8$ mm/h,均方根误差为 $0 \sim 4$ mm/h,相关系数为 $0.5 \sim 1.0$ 。质控后,平均误差降至 $-0.5 \sim 0.6$ mm/h(减少 15%),均方根误差降至 $0 \sim 3.4$ mm/h(减少 13%),相关系数提升至 $0.7 \sim 1.0$ 。结果表明,质控后的待检站数据剔除了异常值,与邻近气象站的一致性显著提升,更能准确反映降水情况。

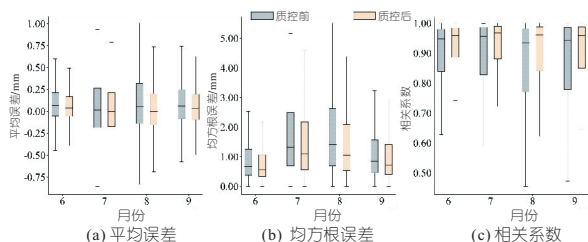


图 6 待检水文站与 1 km 内邻近气象站质控前后平均误差、均方根误差、相关系数对比分析

Fig. 6 Comparison of M_{ME} , R_{RMSE} , and C_{COR} between hydrological stations and neighboring meteorological stations within 1 km before and after quality control

4 结论

a. 通过邻近气象站一致性检验、多源融合产品降水量级划分、雷达和卫星产品降水判断,以及机器模型,构建了针对原始地面观测降水数据的多源协同质控方法。该方法对虚假降水和异常大降水识别效果显著,错误数据检出率最高达 85%。

b. 将多源协同质控方法应用于未经质控的水文站降水数据后,与邻近已质控站点的对比结果显示,平均误差和均方根误差分别减少 15%、13%,相关系数达 $0.7 \sim 1.0$ 。这表明异常值被有效剔除,数据一致性显著提高,能够更真实地反映实际降水情况。

c. 提出的多源协同质控方法有效弥补了单一质控源的不足,显著提升了降水数据质量,促进了跨部门降水数据的联合应用。但由于部门间观测设备、数据采样方式和精度等差异,未来需进一步探索数据校正方法,以提高数据一致性。

参考文献:

[1] 王建凯,陈汝龙,侯威,等. 中国地面气象观测业务的发展历程与展望[J]. 气象科技进展, 2022, 12(5): 10-18. (WANG J K, CHEN R L, HOU W, et al. Development and prospect of surface meteorological observations in China[J]. Advances in meteorological science and technology, 2022, 12(5): 10-

18. (in Chinese))

[2] 黄少平,赖亮,刘海燕,等. 基于格点场的江西省气象和水文降水资料对比分析[J]. 气象与减灾研究, 2020, 43(2): 130-136. (HUANG S P, LAI L, LIU H Y, et al. Comparative analysis of meteorological and hydrological precipitation data based on spatial interpolation in Jiangxi Province[J]. Meteorology and disaster reduction research, 2020, 43(2): 130-136. (in Chinese))

[3] 李显凤,张玮,黄少平,等. 江西省气象-水文-雷达小时雨量信息融合试验结果分析[J]. 暴雨灾害, 2020, 39(3): 276-284. (LI X F, ZHANG W, HUANG S P, et al. Analysis of fusion test results on hourly precipitation from meteorological and hydrological stations and radar[J]. Torrential rain and disasters, 2020, 39(3): 276-284. (in Chinese))

[4] 任芝花,张志富,孙超,等. 全国自动气象站实时观测资料三级质量控制系统研制[J]. 气象, 2015, 41(10): 1268-1277. (REN Z H, ZHANG Z F, SUN C, et al. Development of three-step quality control system of real-time observation data from AWS in China[J]. Meteorological monthly, 2015, 41(10): 1268-1277. (in Chinese))

[5] 姜明,崔明,党岳,等. 雷达组合反射率在降水数据质量控制中的应用[C]//第 32 届中国气象学会年会论文集. 天津, 2015: 353-362. (JIANG M, CUI M, DANG Y, et al. Application of radar composite reflectivity in precipitation data quality control[C]//Proceedings of the 32nd Annual Meeting of the Chinese Meteorological Society. Tianjin, 2015:353-362. (in Chinese))

[6] 王根,陆雅君,王悦,等. KNN 不同距离度量对 FY-4A/AGRI 红外亮温反演降水的影响研究[J]. 红外, 2020, 41(4): 41-48. (WANG G, LU Y J, WANG Y, et al. Research on influence of different distance measurements of KNN on precipitation retrieval by AGRI infrared bright temperatures of FY-4A[J]. Infrared, 2020, 41(4): 41-48. (in Chinese))

[7] LI S Y, HUANG X L, WU W, et al. Evaluation of CMPAS precipitation products over Sichuan, China [J]. Atmospheric and oceanic science letters, 2022, 15(2): 100129.

[8] 谢祥洲,刘军龙,霍斐斐,等. 基于机器学习模型的关中地区 GPM IMERG 降水数据订正方法[J]. 水电能源科学, 2022, 40(2): 6-9. (XIE X Z, LIU J L, HUO F F, et al. Correction method of GPM IMERG precipitation data in Guanzhong Region using machine learning models[J]. Water resources and power, 2022, 40(2): 6-9. (in Chinese))

[9] 张婷. 云顶亮温与义乌市降水关系研究[D]. 金华: 浙江师范大学, 2009. (ZHANG T. Research on Yiwu precipitation background based on TBB[D]. Jin-hua: Zhejiang Normal University, 2009. (in Chinese))