基于集成学习的近实时FY-4A反演降水快速订正方法

吕毅[1,2]，雍斌[1,2]，沈哲辉[1]，李季[1,2]，梅俊[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

摘 要：卫星遥感反演是大范围快速获取近实时高分辨率降水信息的重要途径。2016年成功发射的风云四号卫星（FY-4A）是中国自主研发的新一代地球静止轨道定量遥感气象卫星，由FY-4A反演的中国区域近实时降水产品（FY-4A REGC）为天气监测、水文预报、气候分析等研究提供了高分辨率的近实时降水数据，但其精度与全球卫星降水观测计划的对标产品（IMERG-Early）仍有差距，降水反演的核心订正算法亟待改进与提高。本研究针对中国大陆区域，设定FY-4A REGC和IMERG-Early为模型输入和训练标定，采用最新的集成学习方法（LightGBM）动态构建了一种快速订正高分辨率FY-4A降水产品的新方法。以国家气象信息中心发布的气象地面自动站观测数据（CMPA）作为地面参考，将订正后的FY-4A降水产品（FY-4A Adj）与原始FY-4A REGC进行对比，结果表明新产品FY-4A Adj的相关系数、均方根系误差、相对误差等均有明显改善，而且订正算法有效地降低了原始FY-4A REGC数据对中国东南部区域降水的显著高估。综上，本文提出的基于集成学习的订正算法能够快速、有效地提高近实时风云降水数据FY-4A REGC的综合性能，为生产高精度高分辨率的国产卫星反演降水产品提供了新方法。

关键词：遥感，风云卫星，FY-4A，集成学习，近实时降水，降水订正

中图分类号：TP701/P2

引用格式：吕毅,雍斌,沈哲辉,李季,梅俊.2024.基于集成学习的近实时FY-4A反演降水快速订正方法.遥感学报,28(3):677-688

Lyu Y, Yong B, Shen Z H, Li J and Mei J. 2024. Rapid correction of near real-time FY-4A retrieval based on ensemble machine learning. National Remote Sensing Bulletin, 28(3):677-688[DOI:10.11834/jrs.20242566]

1 引 言

降水是全球水循环系统的重要组成部分，其时空分布的变化深刻影响着陆地水文变化过程（张建云，2010）。获取高时空分辨率降水信息，尤其是近实时降水数据，对径流预报、洪水预警、水库调度等与人民群众生命财产安全息息相关的重大科学问题起着关键作用（刘苏峡 等，2005），进一步深入影响着社会经济的稳定发展（刘志雨，2009）。

卫星反演具有不受下垫面限制、快速获取大范围降水信息、时空分辨率高的优点（刘元波 等，2011；唐国强 等，2015）。目前，新一代全球多卫星联合反演降水计划GPM（Global Precipitation Measurement）能提供整体质量好、时空分辨率高、应用途径广泛的卫星反演降水估计产品（陈晓宏 等，2017）。IMERG（Integrated Multi-satellitE Retrievals for GPM）作为GPM的核心产品之一，在中国大陆上已经多次被验证其具有较好的精度（任英杰 等，2019；张茹 等，2021）。”然而，由于中国未被列入GPM计划的核心研发成员国，国内科研人员难以获取GPM的底层观测信息和反演算法。

风云四号系列卫星是中国自主研发的新一代静止轨道运行的气象卫星（董瑶海，2016）。2016年，搭载着性能位于国际前列的静止轨道辐射成像仪

收稿日期：2022-11-15；**预印本：**2023-12-20

基金项目：国家重点研发计划(编号:2021YFB3900601);国家自然科学基金(编号:U2243229)

第一作者简介：吕毅,研究方向为水文遥感和机器学习。E-mail:lyuyi@hhu.edu.cn

通信作者简介：雍斌,研究方向为遥感水文学。E-mail:yongbin@hhu.edu.cn

AGRI（Advanced Geosynchronous Radiation Imager）的风云四号A星（FY-4A）成功发射并投入使用（唐世浩和毛凌野，2020）。AGRI可通过搭载的双扫描镜进行二维指向，首次实现了分钟级的区域快速扫描，可高频获取14个波段以上的地球云图（王淦泉，2004）。不再受限于单一可见光通道，FY-4A首次回传了更高质量的彩色卫星云图（陆风 等，2017）。

AGRI获取到的反照率、云顶亮温等数据，也为风云系列卫星降水反演提供了重要依据（钟宇璐，2021a）。从2018年3月起，国家气象中心（http：//www.nsmc.org.cn/［2022-11-15］）开始提供降水估计实时产品：FY-4A REGC（中国区域近实时降水估计产品）和FY-4A DISK（全圆盘近实时降水估计产品）。相较于FY-4A DISK覆盖整个亚洲地区，FY-4A REGC仅覆盖中国大陆区域，但有着更高的扫描频率（田昊，2021）。此外，FY-4A REGC没有融合雨量计信息，更能反映卫星反演降水的真实能力。GPM计划中，与FY-4A REGC对标的产品是IMERG的近实时版本IMERG-Early（钟宇璐，2021b）。迄今，部分研究已经评估了FY-4A REGC和IMERG-Early在中国区域近实时估计的精度，发现FY-4A REGC虽然较上代产品有了明显提升，但相较于IMERG-Early仍有差距。这主要因为IMERG-Early扫描时间更长、数据源更多、反演算法更成熟（高浩 等，2021）。与应用于气候研究的多源融合降水产品不同，近实时产品更多运用于水文预报、灾害预警等领域，对时效性要求高（龙柯吉 等，2020）。因此，如何快速订正FY-4A REGC，使其具有媲美IMERG-Early的精度，成为了亟待解决的问题。

目前，订正卫星反演降水产品的方法主要思路是建立历史卫星测雨估计与历史降水真值（一般是雨量计或雷达测量值）之间的线性先验关系模型。当获取到新的观测信息后，再利用上述关系反推订正后的降水（王超，2019）。然而，大量研究表明单纯的线性模型很难精准刻画卫星测雨与降水真值间的关系（魏义熊，2022；李昕潼 等，2023）。

集成学习是一种将几种元机器学习模型组合成一个模型的非线性算法（陈凯和朱钰，2007；何清 等，2014）。作为传统机器学习的凝练和提升，集成学习在偏差订正、方差减少、预测改进等领域取得了较大发展（徐继伟和杨云，2018）。其中，专注于偏差订正的Boosting算法或有潜力应用于卫星降水领域，这已经在宋蕾（2015）、陈浩等（2017）、王超（2019）、钟宇璐（2021a）的研究中有所体现。Boosting算法根据上一次训练得到的子模型结果，调整数据集样本分布，而后生成下一个子模型（于玲和吴铁军，2004）。每个子模型的重要度作为模型输出结果的权重，通过迭代的方式加权计算得出最终结果。根据模型结构设计的不同，产生了GBDT（Friedman，2001；Gradient Boosting Decision Tree，梯度提升决策树）、LightGBM（Ke等，2017；Light Gradient Boosting Machine，轻量级梯度提升机）、XGBoost（Chen和Guestrin，2016；eXtreme Gradient Boosting，极限梯度提升树）等重要分支算法，这些算法各有优势，在众多科学问题中都发挥了重要作用。

相较于深度学习，集成学习算法的模型训练速度更快、所需数据量更少、模型稳定性强（Shinde和Shah，2018；Chauhan和Singh，2018；Nguyen等，2019），更适合于近实时降水的研究。因此，本研究借助极具潜力的集成学习理论，选取并比较典型的集成学习模型LightGBM、XGBoost和Random Forest，从而快速高效地订正近实时FY-4A降水数据。

2 研究区、研究数据和评估方法

2.1 研究区

研究区域为中国（香港、澳门、台湾数据缺失）。研究区地处亚欧大陆东部，太平洋西岸，南北跨度近50°，地势西高东低且地形复杂。研究区降水的空间分布不均匀，年平均降水量呈现由东南沿海向西北内陆递减的趋势（左洪超 等，2004）。由于对季风活动响应较强，中国的降水季节性变化显著，呈现出冬季降水少，夏季降水多的特性（翟盘茂和潘晓华，2003），其中夏季降水是造成中国洪涝灾害的主要原因。

2.2 研究数据

2.2.1 FY-4A REGC

风云四号A星是风云二号C星（中国第一代静止气象卫星第一颗业务卫星）的迭代产品。除了具有通过静止轨道观测云、水汽、植被、地表的

基础功能，FY-4A还具备了捕捉气溶胶、雪的能力，并且能清晰区分云的不同相态和中、高层水汽（范存群 等，2018）。FY-4A的AGRI每小时完成一次全圆盘观测，每15 min在观测空隙进行定位定标观测，覆盖范围为亚太地区；当无全圆盘观测时每5 min进行一次中国区域观测，覆盖范围为3°N—55°N，60°E—137°E（张环宇和唐伯惠，2021）。

国家气象中心于2018年3月12日发布降水反演产品FY-4A REGC和FY-4A DISK。本研究使用FY-4A REGC作为模型输入。FY-4A REGC的原始时空分辨率为5 min（不连续）/4 km。在本研究中，将FY-4A REGC的时空分辨率重采样至1 h/0.1°以匹配地面观测分辨率。数据的时间范围为2018年6月1日至2019年9月30日，覆盖两年的夏季（6、7、8月）。

2.2.2 IMERG-Early

IMERG是全球卫星降水观测计划GPM的代表性卫星反演降水产品之一，其核心卫星上搭载的微波成像仪（GMI）和支持Ku波段（13.6 GHz）和Ka波段（35.5 GHz）的双频降雨雷达（DPR）提供了时空采样更精密的信息源，再通过其反演算法得到满足不同时效和质量需求的全球降水数据集（Smith等，2007）。作为GPM时代的重要成果，IMERG使用的算法由TRMM（Tropical Rainfall Measuring Mission）时代3套主流的降水反演算法（TMPA、GSMaP和PERSIANN）融合发展而产生，它同时引进了卡尔曼滤波和云移动矢量传播两种算法（Draper等，2015）。IMERG系统在近实时阶段运行两次，先后得到IMERG-Early和IMERG-Late（Skofronick-Jackson等，2017）。其中IMERG-Early仅使用了云移动矢量传播算法中的前向传播算法以相对快速地提供结果。IMERG-Early的原始时空分辨率为30 min/0.1°。为与地面参考、FY-4A降水数据匹配，将IMERG-Early的时间分辨率重采样到1 h。

2.2.3 CMPA

CMPA（中国自动站与CMORPH融合的逐时降雨量0.1°网格数据集）使用地面和卫星两个来源的降雨数据：地面观测降雨资料来自全国3万多个自动观测站（包括国家级自动站和区域自动站）逐时降雨量，卫星反演降雨产品选用由美国环境预测中心开发的实时卫星反演降雨产品，应用了概率密度匹配和最优插值算法分两步融合数据（张强 等，2007）。在本研究中，仅使用地面自动站观测数据，将其作为卫星降水数据质量检验的真值。以上3套数据的信息已在表1中给出。

表1 研究使用数据

Table 1 Data used in this research

数据名称	空间分辨率	时间分辨率	延迟时间
FY-4A REGC	4 km	5 min	1 h
IMERG-Early	0.1°	30 min	约4 h
CMPA	0.1°	1 h	实测

2.3 评估方法

本研究为定量评估订正结果的表现采用了3种常用的精度指标（廖荣伟 等，2015；曾岁康和雍斌，2019），其中包括：（1）相关系数CC（Correlation Coefficient）用于量化降水数据与实测数据之间的线性相关程度，最优值为1；（2）均方根误差RMSE（Root Mean Square Error）用于量化降水数据与实测数据之间的离散程度，最优值为0；（3）相对偏差Bias（relative Bias）用于反映卫星降水数据与实测数据之间的偏差程度，最优值为0。各指标计算表达式和最优值见表2。

表2 统计评估参数

Table 2 Statistical evaluation parameters

参数名称	表达式	最优值
相关系数(CC)	$CC=\frac{\sum_{i=1}^{n}(G_i-\overline{G})(S_i-\overline{S})}{\sqrt{\sum_{i=1}^{n}(G_i-\overline{G})^2}\times\sqrt{\sum_{i=1}^{n}(S_i-\overline{S})^2}}$	1
均方根误差（RMSE）	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i-G_i)^2}$	0
偏差(Bias)	$Bias=\frac{\sum_{i=1}^{n}(S_i-G_i)}{\sum_{i=1}^{n}G_i}\times 100\%$	0

注：n代表样本总量；G与$\overline{G}$代表地面降水观测值与其均值；S与$\overline{S}$代表卫星反演降水产品与其均值。

3 基于集成学习的快速订正算法

3.1 LightGBM

LightGBM是集成学习中经典Boosting方法GBDT的改进。LightGBM在传统的梯度提升树的基础上引入直方图决策算法、单边梯度采样和互斥特征捆绑算法（Lundberg等，2019）。在样本数据

量和特征量增长的情况下，LightGBM的精度却不受影响，并且能够有效提升模型训练速度。

直方图决策算法通过构建直方图得到分集。将连续的输入值离散化成k个整数并构造一个宽度为k的直方图，遍历直方图的值以找最优分割点，有效减少了候选分裂点数量。由于目标函数增益主要来自于梯度绝对值较大的样本，因此单边梯度采样只考虑梯度绝对值小于一定阈值的样本，保留绝对值较大的样本。互斥特征捆绑算法则可以通过对某些特征的取值重新编码，将多个互斥的特征绑定为一个新特征，以降低计算复杂度（Ke等，2017）。这使得该算法在保证训练精度的同时，极大提升了算法的运行速度。

3.2 XGBoost

XGBoost是经典Boosting方法GBDT的另一种改进。相较于GBDT，XGBoost基于二阶泰勒公式并引入了正则化方法。对于一般模型，目标函数可以表示为

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{3.1}$$

式中，$L(\theta)$是训练损失函数，$\Omega(\theta)$是正则化项。$L(\theta)$的常见选择是均方根误差，它由下式给出：

$$L(\theta) = \sum_{i=1}(y_i - \overline{y_i})^2 \tag{3.2}$$

式中，y_i是样本，$\overline{y_i}$是样本均值。

正则化方法定义了模型复杂度：

$$\Omega(\theta) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega^2 \tag{3.3}$$

式中，T是决策树的叶子数，γ是折算系数，是$\frac{1}{2}\lambda\sum_{j=1}^{T}\omega^2$叶子结点对应的值向量的L2范数。

XGBoost运行的一般步骤是：首先，从深度为0的树开始，对每个叶子节点枚举所有可用特征。其次，针对每个特征，把属于该节点的训练样本根据该特征值升序排列，通过线性扫描的方式来决定该特征的最佳分裂点，并记录该特征的最大收益。然后，选择收益最大的特征作为分裂特征，用该特征的最佳分裂点作为分裂位置，并为每个新节点关联对应的样本集。最后，反复递归执行到满足特定条件为止（Chen和Guestrin，2016）。

3.3 Random Forest

Random Forest是集成学习中Bagging方法的代表模型之一。其一般步骤是：从训练集中有放回地抽样，取样多次形成一个新训练用子集$\boldsymbol{D}$，随机选择m个特征。使用新的训练集$\boldsymbol{D}$和m个特征，学习出一个完整的决策树，反复进行多次，最后得到随机森林（Breiman，2001）。

与GBDT相比：Random Forest是并行生成的，而GBDT是串行生成的；Random Forest的结果是多数表决形成的，而GBDT的结果则是多棵树累加所得。本研究中，主要使用Random Forest与两种Boosting方法模型LightGBM和XGBoost对比。

3.4 算法流程

本研究提出的基于集成机器学习的快速订正算法如图1（a）所示。具体可分为4个步骤。

步骤一，数据处理。本步骤首先将FY-4A REGC和IMERG-Early的时空分辨率重采样至0.1°/1 h，以匹配CMPA的自动站观测数据。为确保在没有其他气象（降水）数据输入的情况下仍能完成订正任务，本研究仅针对FY-4A REGC估计有雨时的数据。与此同时，CMPA则仅使用有自动观测站点的格点。在本研究中，我们选取了80%的FY-4A REGC作为训练集的输入，并将IMERG-Early作为训练数据集标定。然后，训练集将按不同数量级进一步分割，具体分割方式如图1（b）所示。不同的分割方法将产生不同的模型参数和运行时间。此处，2^{20}个样本数量级约包含3 h的数据特征，而2^{25}个样本数量级约包含4 d的数据特征。分割后，剩余的20%的FY-4A REGC将用作验证集，输入训练完成的模型以获得订正结果。此外，研究未打乱输入数据的时间顺序。因此训练集大约对应2018年6月1日至2019年6月30日，而验证集大约对应2018年7月1日至2019年9月30日。最后将CMPA中的自动站观测数据用作验证真值，以评估订正效果。评估结果的时间范围与验证集相同。

步骤二，模型比较。本研究选取了两种Boosting方法模型LightGBM和XGBoost以及一种Bagging方法模型Random Forest。本研究通过综合评估回归准确率、时间复杂度与输入数据量的关系，获取在默认参数设置下，最适合当前任务的集成学习模型。一旦确定被选模型，我们将使用网格搜索方法对其超参数进行进一步优化。

图2使用泰勒图比较了FY-4A REGC和IMERG-Early在2018年夏季和2019年夏季的表现。

在泰勒图中，估计点距离“观测值”越近，说明数据集越接近观测值。结果显示，IMERG-Early在2018年夏季和2019年夏季的表现几乎相同，而FY-4A REGC则有明显不同：2019年夏季点与观测点的距离较2018年夏季更小。这表明FY-4A REGC的数据质量随着时间的推移有明显的提升。这主要是由于官方对反演算法和定标结果进行了调整。为了使模型能够清晰反映FY-4A REGC和IMERG-Early之间的隐含关系，我们提出了一种滚动输入最新数据并不断更新模型参数的运行方法。

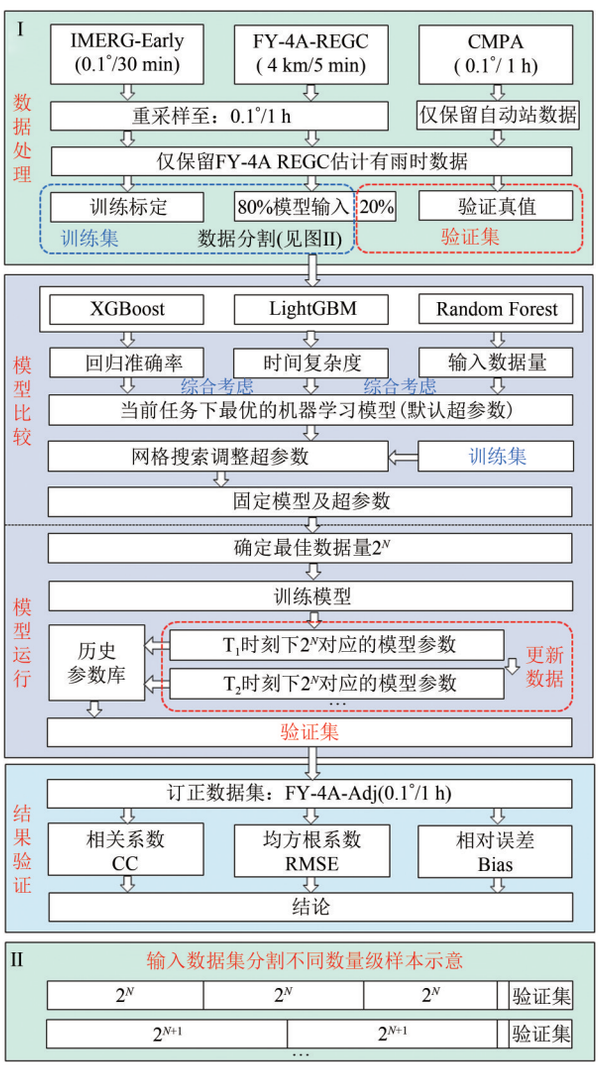


图1 基于集成学习的快速订正算法

Fig. 1 Fast correction algorithm based on ensemble machine learning

步骤三，模型运行。首先，固定模型及其超参数，确定模型训练合适的输入数据量2^N（N为待确定值）。其次，训练模型并获得T_1时刻下2^N对应的模型参数并记录到历史参数库。然后，当获取到新数据时，记为T_2时刻。此时删除最旧的数据并加入新数据，始终保持数据总量为2^N。重复训练模型的过程。最后，获得模型参数库，加载最接近参数库所载时间T_i（i=1，2，3，…）的模型参数以运行模型。

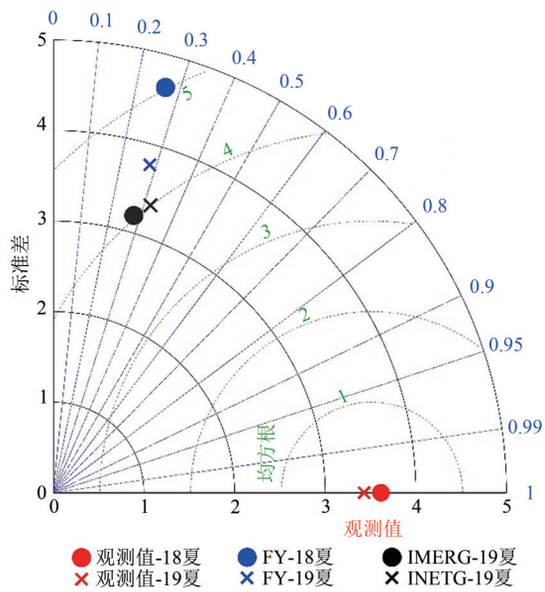


图2 2018年、2019年夏季3套产品统计性能泰勒图

Fig. 2 Taylor diagram of statistical performance for three datasets of products in the summers of 2018 and 2019

图3展示了随着时间推移，在训练集上不同模型参数更新后的输出结果评估对比。结果显示，随着时段数的增加，输出结果的评估效果明显改善。

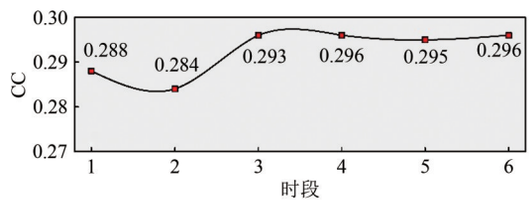
步骤四，结果验证。我们固定模型并输入验证集，将模型输出的数据作为输出订正数据集FY-4A Adj（时空分辨率为1 h/0.1°）。最后，通过计算CC、RMSE、Bias等指标以评估模型效果。

4 结果与讨论

4.1 模型的选取

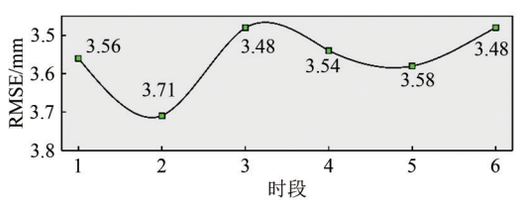
鉴于FY-4A REGC的数据质量随着时间推移有明显提升的特性，滚动输入新观测信息以更新模型的最优参数有其必要性。此外，FY-4A REGC是近实时降水反演产品，更新订正模型参数的过程必须考虑时效性。因此，挑选一种在输入数据量级逐步提升条件下，仍能兼顾运行时间和订正精度的模型成了本研究的首要问题。图4通过热力图的形式，给出了每种集成学习模型运行各数据量级的输入数据后，训练模型的回归精度和所需

时间（运行平台如表3所示）。需要指出的是，模型使用默认超参数且运行同一模型时仅改变输入数据的量级。



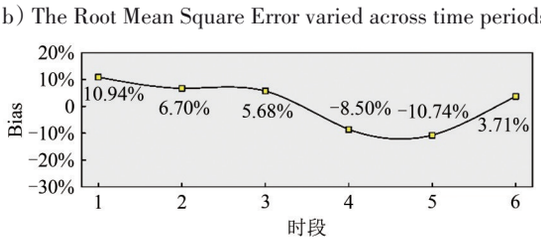
（a）相关系数随时间变化

（a）The Correlation Coefficient varied across time periods



（b）均方根误差随时间变化

（b）The Root Mean Square Error varied across time periods



（c）相对误差随时间变化

（c）The Bias varied across time periods

图3 模型内不同参数随时段变化输出结果的评估对比（基于训练集和CMPA）

Fig. 3 Evaluation comparison of model output with different parameters varied across time periods (based on training datasets and CMPA)

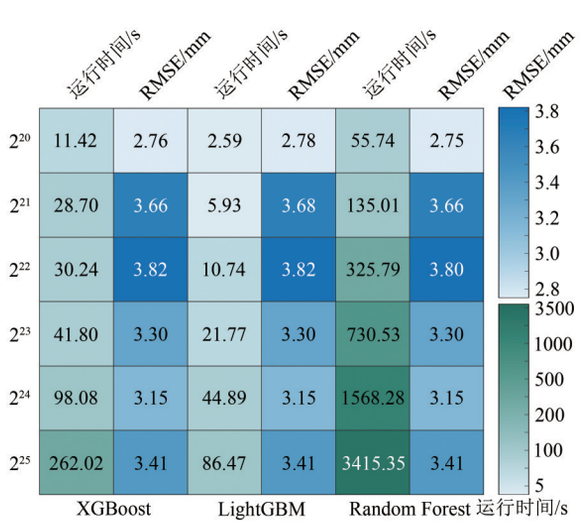


图4 3种模型不同数量级数据输入下的回归精度和运行时间

Fig. 4 Regression accuracy and execution time of three models with varying magnitudes of data input

表3 实验平台

Table 3 Experiment platform

参量	指标
中央处理器	英特尔酷睿 i5-9500 处理器
内存	三星 8 G DDR4 2666 MHZ *2
硬盘	西部数据 BLACK SN750 SSD
操作系统	Windows 10 企业版
运行平台	PyCharm Community Edition
PYTHON 版本	Python 3.9
SKLEARN 版本	1.1.2

根据训练精度表现可知，在数据量提升至2^{23}之前，代表Bagging算法的Random Forest模型要优于代表Boosting算法的XGBoost和LightGBM模型，但当XGBoost和LightGBM在2^{23}数据输入时，它们的训练效果与Random Forest持平。从训练时间方面来看，尽管3种模型获得了类似的训练效果，但XGBoost在2^{20}的训练时间为LightGBM的4.4倍，而Random Forest所需的训练时间更是为LightGBM的21.5倍。在数据量进一步增加至2^{25}后，所需时间更是增长到了LightGBM的39.5倍。

经过上述分析，可以得出以下结论：随着训练数据量的增加，所有集成学习模型的训练时间都呈线性增长趋势。训练精度相对稳定，受训练数据量级的影响不大。Bagging算法在数据量较少时略好于Boosting算法，但随着数据量的增加，Bagging算法的运行复杂度显著增加，而Boosting算法则只需要延拓部分误差传播模型即可。在Boosting算法中，LightGBM的直方图决策算法、单边梯度采样和互斥特征捆绑算法对维持训练精度和提升训练速度起到了显著作用。在样本数据量和特征量增长的情况下，LightGBM不但能保持训练精度，而且模型训练速度明显更快。因此，当数据量较少时，更推荐使用包括Random Forest模型在内的Bagging算法，而当数据量较大时，更推荐使用Boosting算法，尤其是LightGBM模型，因为它可以兼顾训练精度和训练时间。在本研究中，我们将选取LightGBM作为快速订正近实时降水反演产品FY-4A REGC的主要方法。

4.2 分割输入量级的选取

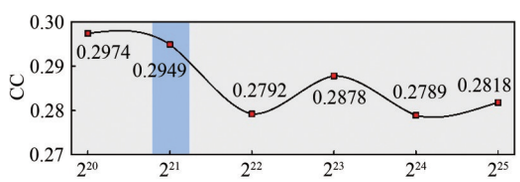
此外，对新生成的产品FY-4A Adj进行了分析，该产品是未参与训练的验证集数据，输入经

过网格搜索法调参后的LightGBM模型产生的输出结果。我们将FY-4A Adj与CMPA观测资料进行比较，以确定最适合的输入量级。表4列出了网格搜索法调整的超参数。图5则展示了训练结果与IMERG-Early计算所得的均方根误差值RMSE。

表4 网格搜索法调整超参数

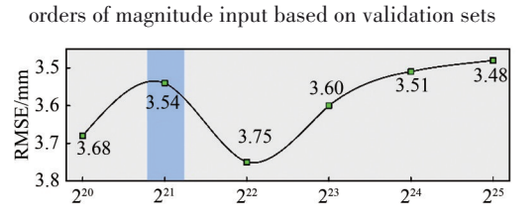
Table 4 Grid search for hyper-parameters tuning

超参数名称	搜索步长	搜索范围	搜索结果
n_estimators	50	[200,500]	450
min_child_weight	1	[1,10]	5
num_leaves	1	[1,10]	8
min_data_in_leaf	1	[1,30]	16
max_bin in	10	[1,101]	21



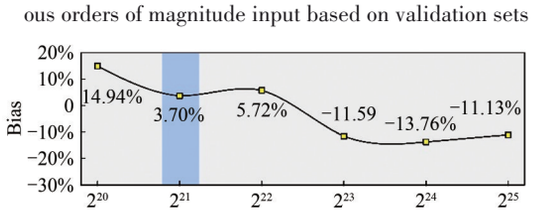
（a）相关系数的变化趋势基于验证集

（a）The variation of Correlation Coefficient leaded by various orders of magnitude input based on validation sets



（b）均方根误差的变化趋势基于验证集

（b）The variation of Root Mean Square Error leaded by various orders of magnitude input based on validation sets



（c）Bias的变化趋势基于验证集

（c）The variation of Bias leaded by various orders of magnitude input based on validation sets

图5 各数量级输入导致评估结果（蓝色代表优选数据量）

Fig. 5 Performance changing trends (The blue areas represent the best)

如图5所示，我们展示了不同输入量级下FY-4A Adj对比验证真值的评估结果。总体而言，随着输入量级的增加，相关系数CC下降，均方根误差RMSE波动上升，偏差Bias轻微下降。这表明，在相似的模型结构和参数设置下，当数据输入量过多时，模型的泛化能力可能会降低。这主要是由于Boosting算法框架下，模型误差是通过生成和累积决策树来实现的。输入量级的增加可能会导致决策树结构更加复杂，从而产生更多的不确定性。因此，集成模型需要考虑数据输入量以获得更好的效果。这与深度学习要求更多的数据输入相反（Bottou和Bousquet，2007）。

图5中用蓝色标识的部分，是总体表现最好的模型。因此，本研究将使用由2^{21}作为分割输入数量级以训练生成的模型。

4.3 订正产品FY-4A Adj的空间分布

图6展示了FY-4A Adj、FY-4A REGC、IMERG-Early等3种降水估计产品在中国大陆各区域空间分布表现。可以发现，这3种产品的降水分布趋势大致相似，均能反映出雨季的降水地域性分异特征。其中，IMERG-Early的表现整体更加精细，而FY-4A系列产品则表现出明显的插值特征。在中国东南部地区，FY-4A REGC和IMERG-Early的降水估计存在明显的差异，FY-4A REGC相比IMERG-Early有显著的高估现象。在中国西北部地区，两者的表现差异不大。

由于算法只考虑FY-4A REGC估计有降水的区域，而降水事件数量远少于非降水事件，因此在平均小时降水尺度上很难反映两套产品的差距。图6中所示的FY-4A Adj和FY-4A REGC降水空间分布较为相似。因此，本研究还提供了FY-4A Adj减去FY-4A REGC的结果，如图6所示。可以看出，在与IMERG-Early的估计有较大分歧的地区，FY-4A Adj几乎都进行了降水量上的调整，使其更接近IMERG-Early。整体而言，FY-4A Adj进行了许多正向的调整：在中国中部和北部进行了一些上调；而在中国的西部，进行了轻微的下调。在中国的东南部，FY-4A Adj进行了较大程度的下调，使其更加接近IMERG-Early。

4.4 订正产品FY-4A Adj的地面验证

图7展示了FY-4A Adj、FY-4A REGC、IMERG-Early和CMPA自动站数据之间的散点关系图。从图7（b）和图7（c）可以看出：在研究时间段内，IMERG-Early和FY-4A REGC估计的小时降水主要分布在0—5 mm，且分布相对均匀，接近45°线。

然而，IMERG-Early的表现明显优于FY-4A REGC，因为FY-4A REGC有更多的数据点分布在接近坐标轴的区域，这意味着CMPA观测到的降水量较FY-4A REGC估计的降水量偏移较多。此外，在45°线上，IMERG-Early有更多的数据点，呈现出“凸型”，而FY-4A REGC则较为分散，呈现出“凹型”。从精度指标来看，IMERG-Early的CC和RMSE都明显优于FY-4A REGC，这与散点图的结果一致。

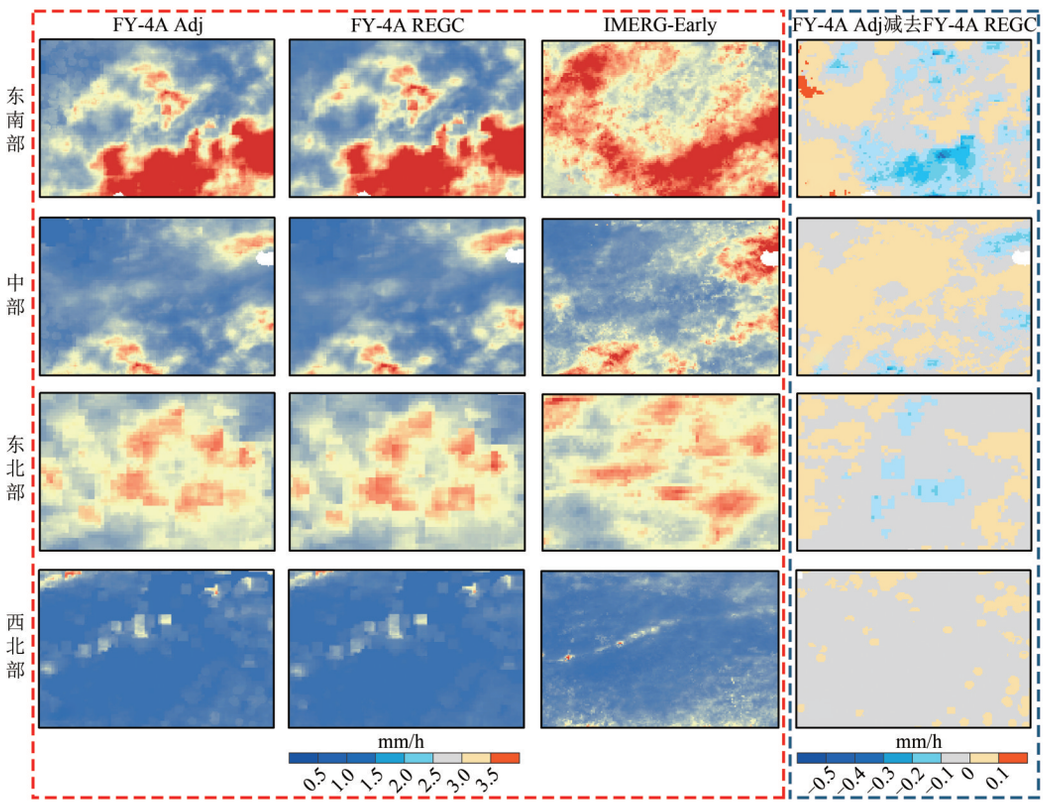
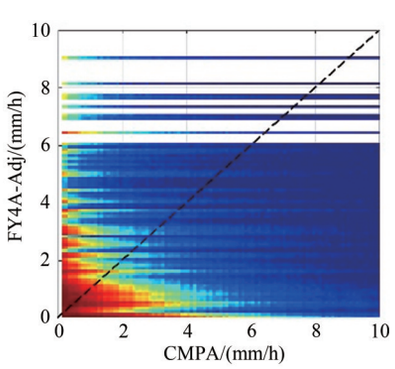


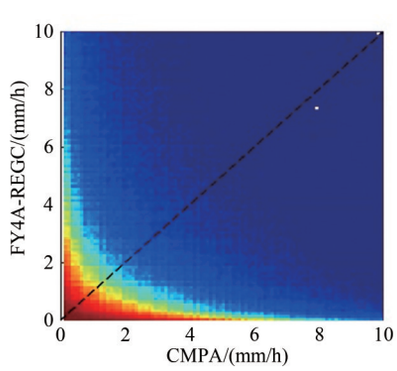
图6 中国东南部、中国中部、中国东北部、中国西北部各区域上FY-4A Adj，FY-4A REGC，IMERG-Early，FY-4A Adj减去FY-4A REGC的小时平均降水量(时间范围:2019年7月1日至2019年9月30日,即验证集)

Fig. 6 Southeast China, Central China, Northwest China, Northwest China: Average hourly precipitation of FY-4A Adj, FY-4A REGC, IMERG-Early, FY-4A Adj minus FY-4A REGC in other regions of China (Time range: July 1, 2019 to September 30, 2019, i.e. based on validation datasets)

图7（a）展示了FY-4A Adj和CMPA自动站数据之间的散点关系图。值得注意的是，经过订正后，更多的数据点集中在45°线上（尤其是在0—2 mm降水区间范围内），这表明本方法对FY-4A REGC的订正在雨强较小时效果显著。



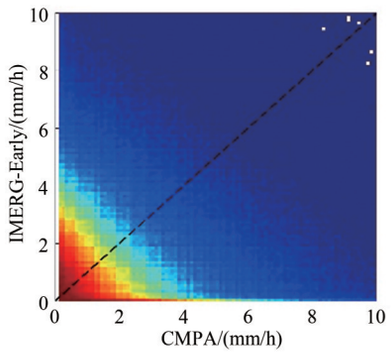
（a）FY-4A Adj



（b）IMERG-Early

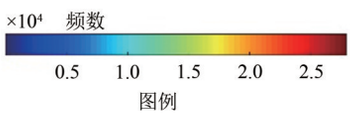
（a）FY-4A Adj

（b）IMERG-Early



表：主要数据对比评估

数据	基准	CC	RMSE/mm	Bias/%
FY-4A REGC	CMPA	0.27	5.03	14.15
IMERG-Early	CMPA	0.31	3.98	17.15
FY-4A Adj	IMERG-Early	0.82	2.95	1.01
FY-4A Adj	CMPA	0.29	3.54	3.70



（c）IMERG-Early和CMPA小时降水散点对比及评估结果

（c）IMERG-Early and CMPA about hourly precipitation

图7 FY-4A Adj、FY-4A REGC、IMERG-Early和CMPA小时降水散点对比及评估结果（2019年7月1日至2019年9月30日，即基于验证集）

Fig. 7 Scatter comparison and evaluation results of FY-4A Adj、FY-4A REGC、IMERG-Early and CMPA about hourly precipitation (From July 1, 2019 to July 30, 2019, i.e. based on validation dataset)

然而，当降水强度超过6 mm时，散点图中出现了部分“断层”。这主要是因为在中、高雨强下，输入的训练样本过少，导致模型会笼统地把一定范围内的输入都映射到同一个标定值附近。因此，需要对中、高雨强的样本进行强化训练。但由于中、高雨强仍然占少数，因此上述因素对订正结果的质量影响有限。从另一方面来看，尽管对中、高雨强的订正仍有明显缺陷，但经本方法订正后的FY-4A降水数据更接近IMERG-Early的质量，证明了本方法的潜力。此外，整个降水分布出现了一定的“倾斜”现象，说明本算法对整体降水估计进行了调整。这样的调整有利于订正结果，使得FY-4A Adj对于降水总量的估计更加准确（Bias由14.15%降至3.70%，超过IMERG-Early）。

5 结 论

本研究提出了一种基于集成学习的快速订正算法，实现了基于地面站点观测的近实时FY-4A卫星反演降水数据的快速校正。经评估分析表明，该方法能够有效且快速地提升FY-4A REGC的精度，使其达到了全球降水观测计划近实时产品IMERG-Early的数据质量。具体结论如下：

（1）FY-4A Adj相较于FY-4A REGC，评估指标CC、RMSE和Bias值有明显提升，有效降低了FY-4A REGC在中国南部的显著高估，改善了风云卫星反演降水估计的准确性。

（2）集成学习算法的选取会受到输入数据量级的影响。对于数据量较少的情况，建议使用Bagging算法，如Random Forest；而对于数据量较大的情况，建议使用Boosting算法，如LightGBM模型，以兼顾精度和运行时间。

（3）输入训练集数据数量的增加并不一定能够提高集成学习模型的精度。在本研究中，2^{21}个样本量是训练模型参数的最佳数量级。

（4）由于缺乏中高雨强的样本数据，本算法对于此类情况的预测存在偏向同一值的问题。在获取更多样本数据后，可以考虑使用强化学习算法对中、高雨强的情况进行训练，以提高模型的准确性。

参考文献（References）

Bottou L and Bousquet O. 2007. The tradeoffs of large scale learning// Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc.: 161-168

Breiman L. 2001. Random forest. Machine Learning, 45(1): 5-32 [DOI: 10.1023/A:1010933404324]

Chauhan N K and Singh K. 2018. A review on conventional machine learning vs deep learning//2018 International Conference on Computing, Power and Communication Technologies (GUCON). Greater Noida: IEEE: 347-352 [DOI: 10.1109/GUCON.2018.8675097]

Chen H, Ning C, Nan Z T, Wang Y D, Wu X B and Zhao L. 2017. Correction of the daily precipitation data over the Tibetan Plateau with machine learning models. Journal of Glaciology and Geocryology, 39(3): 583-592 (陈浩, 宁忱, 南卓铜, 王玉丹, 吴小波, 赵林. 2017. 基于机器学习模型的青藏高原日降水数据的订正研究. 冰川冻土, 39(3): 583-592) [DOI: 10.7522/j.issn.1000-0240.2017.0065]

Chen K and Zhu Y. 2007. A summary of machine learning and related algorithms. Statistics and Information Forum, 22(5): 105-112 (陈凯, 朱钰. 2007. 机器学习及其相关算法综述. 统计与信息论坛, 22(5): 105-112) [DOI: 10.3969/j.issn.1007-3116.2007.05.021]

Chen T Q and Guestrin C. 2016. XGBoost: a scalable tree boosting system//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM: 785-794 [DOI: 10.1145/2939672.2939785]

Chen X H, Zhong R D, Wang Z L, Lai C G and Chen J C. 2017. Evaluation on the accuracy and hydrological performance of the latest-generation GPM IMERG product over South China. Journal of Hydraulic Engineering, 48(10): 1147-1156 (陈晓宏, 钟睿达, 王兆礼, 赖成光, 陈家超. 2017. 新一代GPM IMERG卫星遥感降水数据在中国南方地区的精度及水文效用评估. 水利学报, 48(10): 1147-1156) [DOI: 10.13243/j.cnki.slxb.20170202]

Dong Y H. 2016. FY-4 meteorological satellite and its application prospect. Aerospace Shanghai, 33(2): 1-8 (董瑶海. 2016. 风云四号气象卫星及其应用展望. 上海航天, 33(2): 1-8) [DOI: 10.19328/j.cnki.1006-1630.2016.02.001]

Draper D W, Newell D A, Wentz F J, Krimchansky S and Skofronick-Jackson G M. 2015. The Global Precipitation Measurement (GPM) Microwave Imager (GMI): instrument overview and early on-orbit performance. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(7): 3452-3462 [DOI: 10.1109/JSTARS.2015.2403303]

Fan C Q, Lin M Y, Zhao X G, Xie L Z, Wei L and Guo P. 2018. Research on parallelization of Fengyun satellite precipitation estimation daily end product algorithm//The 35th Annual Conference of the Chinese Meteorological Society S21 Satellite Meteorology and Ecological Remote Sensing. Hefei: [s.n.]: 33-38 (范存群, 林曼筠, 赵现纲, 谢利子, 卫兰, 国鹏. 2018. 风云卫星降水估计日收工产品算法并行化研究//第35届中国气象学会年会 S21 卫星气象与生态遥感. 合肥: [s.n.]: 33-38)

Friedman J H. 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics, 29(5): 1189-1232 [DOI: 10.1214/AOS/1013203451]

Gao H, Tang S H and Han X Z. 2021. China's Fengyun (FY) meteorological satellites, development and applications. Science and Technology Review, 39(15): 9-22 (高浩, 唐世浩, 韩秀珍. 2021. 风云气象卫星发展及其应用. 科技导报, 39(15): 9-22) [DOI: 10.3981/j.issn.1000-7857.2021.15.001]

He Q, Li N, Luo W J and Shi Z Z. 2014. A survey of machine learning algorithms for big data. Pattern Recognition and Artificial Intelligence, 27(4): 327-336 (何清, 李宁, 罗文娟, 史忠植. 2014. 大数据下的机器学习算法综述. 模式识别与人工智能, 27(4): 327-336) [DOI: 10.3969/j.issn.1003-6059.2014.04.007]

Ke G L, Meng Q, Finley T, Wang T F, Chen W, Ma W D, Ye Q W and Liu T Y. 2017. LightGBM: a highly efficient gradient boosting decision tree//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 2149-3157

Li X T, Li Z L and Han R C. 2023. Evaluations of different bias correction methods on the GCM precipitation data. Journal of China Hydrology, 43(3): 93-100, 117 (李昕潼, 李占玲, 韩孺村. 2023. 不同偏差校正法对GCM降水数据的应用效果分析. 水文, 43(3): 93-100, 117) [DOI: 10.19797/j.cnki.1000-0852.20220060]

Liao R W, Zhang D B and Shen Y. 2015. Validation of six satellite-derived rainfall estimates over China. Meteorological Monthly, 41(8): 970-979 (廖荣伟, 张冬斌, 沈艳. 2015. 6种卫星降水产品在中国区域的精度特征评估. 气象, 41(8): 970-979) [DOI: 10.7519/j.issn.1000-0526.2015.08.006]

Liu S X, Xia J and Wo X G. 2005. Advances in predictions in ungauged basins. Water Resources and Hydropower Engineering, 36(2): 9-12 (刘苏峡, 夏军, 莫兴国. 2005. 无资料流域水文预报(PUB计划)研究进展. 水利水电技术, 36(2): 9-12) [DOI: 10.3969/j.issn.1000-0860.2005.02.003]

Liu Y B, Fu Q N, Song P, Zhao X S and Dou C C. 2011. Satellite retrieval of precipitation: an overview. Advances in Earth Science, 26(11): 1162-1172 (刘元波, 傅巧妮, 宋平, 赵晓松, 豆翠翠. 2011. 卫星遥感反演降水研究综述. 地球科学进展, 26(11): 1162-1172) [DOI: 10.11867/j.issn.1001-8166.2011.11.1162]

Liu Z Y. 2009. Research progress and prospect of flood forecasting technology in China. China Flood and Drought Management, 19(5): 13-16 (刘志雨. 2009. 我国洪水预报技术研究进展与展望. 中国防汛抗旱, 19(5): 13-16) [DOI: 10.16867/j.cnki.cfdm.2009.05.005]

Long K J, Gu J X, Shi C X, Pan Y and Huang X L. 2020. Quality Assessment of several merged precipitation products in a heavy rainfall process in Sichuan. Plateau and Mountain Meteorology Research, 40(2): 31-37 (龙柯吉, 谷军霞, 师春香, 潘旸, 黄晓龙. 2020. 多种降水实况融合产品在四川一次强降水过程中的评估. 高原山地气象研究, 40(2): 31-37) [DOI: 10.3969/j.issn.1674-2184]

Lu F, Zhang X H, Chen B Y, Liu H, Wu R H, Han Q, Feng X H, Li Y and Zhang Z Q. 2017. FY-4 geostationary meteorological satellite imaging characteristics and its application prospects. Journal of Marine Meteorology, 37(2): 1-12 (陆风, 张晓虎, 陈博洋, 刘辉, 吴荣华, 韩琦, 冯小虎, 李云, 张志清. 2017. 风云四号气象卫星成像特性及其应用前景. 海洋气象学报, 37(2): 1-12) [DOI: 10.19513/j.cnki.issn2096-3599.2017.02.001]

Lundberg S M, Erion G G and Lee S I. 2019. Consistent individualized feature attribution for tree ensembles. arXiv: 1802.03888 [DOI: 10.48550/arXiv.1802.03888]

Nguyen G, Dlugolinsky S, Bobák M, Tran V, García Á L, Heredia I, Malík P and Hluchý L. 2019. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 52(1): 77-124 [DOI: 10.1007/s10462-018-09679-z]

Ren Y J, Yong B, Lu D K and Chen H Q. 2019. Evaluation of the Integrated Multi-satellitE Retrievals (IMERG) for Global Precipitation Measurement (GPM) mission over the Mainland China at multiple scales. Journal of Lake Sciences, 31(2): 560-572 (任英杰, 雍斌, 鹿德凯, 陈汉清. 2019. 全球降水计划多卫星降水联合反演IMERG卫星降水产品在中国大陆地区的多尺度精度评估. 湖泊科学, 31(2): 560-572) [DOI: 10.18307/2019.0224]

Shinde P P and Shah S. 2018. A review of machine learning and deep

learning applications//2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). Pune: IEEE: 1-6 [DOI: 10.1109/ICCUBEA.2018.8697857]

Skofronick-Jackson G, Petersen W A, Berg W, Kidd C, Stocker E F, Kirschbaum D B, Kakar R, Braun S A, Huffman G J, Iguchi T, Kirstetter P E, Kummerow C, Meneghini R, Oki R, Olson W S, Takayabu Y N, Furukawa K and Wilheit T. 2017. The Global Precipitation Measurement (GPM) mission for science and society. Bulletin of the American Meteorological Society, 98(8): 1679-1695 [DOI: 10.1175/BAMS-D-15-00306.1]

Smith E A, Asrar G, Furuhama Y, Ginati A, Mugnai A, Nakamura K, Adler R F, Chou M D, Desbois M, Durning J F, Entin J K, Einaudi F, Ferraro R R, Guzzi R, Houser P R, Hwang P H, Iguchi T, Joe P, Kakar R, Kaye J A, Kojima M, Kummerow C, Kuo K S, Lettenmaier D P, Levizzani V, Lu N M, Mehta A V, Morales C, Morel P, Nakazawa T, Neeck S P, Okamoto K, Oki R, Raju G, Shepherd J M, Simpson J, Sohn B J, Stocker E F, Tao W K, Testud J, Tripoli G J, Wood E F, Yang S and Zhang W J. 2007. International Global Precipitation Measurement (GPM) program and mission: an overview//Levizzani V, Bauer P and Turk F J, eds. Measuring Precipitation from Space: Eurainsat and the Future. Dordrecht: Springer: 611-653 [DOI: 10.1007/978-1-4020-5835-6_48]

Song L. 2015. Study of Regional Precipitation Product with High Spatial-Temporal Resolution over the Tibetan Plateau based on TRMM 3B43. Nanjing: Nanjing University of Information Science and Technology (宋蕾. 2015. 基于TRMM 3B43青藏高原区域性高时空分辨率降水探究. 南京: 南京信息工程大学)

Tang G Q, Wan W, Zeng Z Y, Guo X L, Li N, Di L and Hong Y. 2015. An overview of the Global Precipitation Measurement (GPM) mission and it's latest development. Remote Sensing Technology and Application, 30(4): 607-615 (唐国强, 万玮, 曾子悦, 郭晓林, 李娜, 龙笛, 洪阳. 2015. 全球降水测量(GPM)计划及其最新进展综述. 遥感技术与应用, 30(4): 607-615) [DOI: 10.11873/j.issn.1004-0323.2015.4.0607]

Tang S H and Mao L Y. 2020. "Fengyun" application for 50 years: leapfrog development, serve the world. Satellite Application, (11): 8-13 (唐世浩, 毛凌野. 2020. "风云"应用50年: 跨越发展, 服务全球. 卫星应用, (11): 8-13) [DOI: 10.3969/j.issn.1674-9030.2020.11.004]

Tian H. 2021. Typhoon Rainfall Inversion Technology and Error Analysis based on Fengyun-4 Sensor. Nanjing: Nanjing University of Information Science and Technology (田昊. 2021. 基于风云四号传感器的台风降雨反演技术及误差分析. 南京: 南京信息工程大学) [DOI: 10.27248/d.cnki.gnjqc.2021.000586]

Wang C. 2019. Research on Quality Evaluation and Correction Method of Satellite Telemetry Precipitation Data. Chongqing: Chongqing Jiaotong University (王超. 2019. 卫星遥测降水数据的质量评估与校正的方法研究. 重庆: 重庆交通大学) [DOI: 10.27671/d.cnki.gcjtc.2019.000157]

Wang G Q. 2004. Fengyun-4 meteorological satellite multi-channel scanning imaging radiometer//Mr. Daheng's 90th Birthday Collection and Proceedings of the 2004 Academic Conference of the Chinese Optical Society. Hangzhou: Zhejiang University Press: 1436-1439 (王淦泉. 2004. 风云四号气象卫星多通道扫描成像辐射计//大珩先生九十华诞文集暨中国光学学会2004年学术大会论文集. 杭州: 浙江大学出版社: 1436-1439)

Wei Y X. 2022. Downscaling Correction of Multi-Source Satellite Remote Sensing Precipitation Products in Pingtang Basin and its Applicability in Runoff Simulation. Nanning: Guangxi University (魏义熊. 2022. 多源卫星遥感降水产品在平塘流域的降尺度校正及径流模拟应用研究. 南宁: 广西大学) [DOI: 10.27034/d.cnki.ggxiu.2022.000853]

Xu J Y and Yang Y. 2018. A survey of ensemble learning approaches. Journal of Yunnan University (Natural Sciences Edition), 40(6): 1082-1092 (徐继伟, 杨云. 2018. 集成学习方法: 研究综述. 云南大学学报(自然科学版), 40(6): 1082-1092) [DOI: 10.7540/j.ynu.20180455]

Yu L and Wu T J. 2004. Assemble learning: a survey of Boosting algorithms. Pattern Recognition and Artificial Intelligence, 17(1): 52-59 (于玲, 吴铁军. 2004. 集成学习: Boosting算法综述. 模式识别与人工智能, 17(1): 52-59) [DOI: 10.3969/j.issn.1003-6059.2004.01.010]

Zeng S K and Yong B. 2019. Evaluation of the GPM-based IMERG and GSMaP precipitation estimates over the Sichuan region. Acta Geographica Sinica, 74(7): 1305-1318 (曾岁康, 雍斌. 2019. 全球降水计划IMERG和GSMaP反演降水在四川地区的精度评估. 地理学报, 74(7): 1305-1318) [DOI: 10.11821/dlxb201907003]

Zhai P M and Pan X H. 2003. Change in extreme temperature and precipitation over Northern China during the second half of the 20th century. Acta Geographica Sinica, 58(S1): 1-10 (翟盘茂, 潘晓华. 2003. 中国北方近50年温度和降水极端事件变化. 地理学报, 58(S1): 1-10) [DOI: 10.3321/j.issn:0375-5444.2003.z1.001]

Zhang H Y and Tang B H. 2021. Remote sensing retrieval of total precipitable water under clear-sky atmosphere from FY-4A AGRI data by combining physical mechanism and random forest algorithm. National Remote Sensing Bulletin, 25(8): 1836-1847 (张环宇, 唐伯惠. 2021. 融合物理机理与随机森林算法的FY-4A AGRI数据晴空大气可降水量遥感反演. 遥感学报, 25(8): 1836-1847) [DOI: 10.11834/jrs.20211217]

Zhang J Y. 2010. Review and reflection on China's hydrological forecasting techniques. Advances in Water Science, 21(4): 435-443 (张建云. 2010. 中国水文预报技术发展的回顾与思考. 水科学进展, 21(4): 435-443) [DOI: 10.14042/j. cnki. 32.1309.2010.04.019]

Zhang Q, Tu M H, Ma S Q, Yang Z B and Luo Y C. 2007. Quality assessment of the observational data of automatic precipitation stations in China. Journal of Applied Meteorological Science, 18(3): 365-372 (张强, 涂满红, 马舒庆, 杨志彪, 罗永春. 2007. 自动雨量站降雨资料质量评估方法研究. 应用气象学报, 18(3): 365-372) [DOI: 1038781/j.issn.1006-9895.1403.13295]

Zhang R, Yong B and Zeng S K. 2021. Evaluation of GPM satellite precipitation products over Mainland China. Yangtze River, 52(5): 50-59 (张茹, 雍斌, 曾岁康. 2021. GPM卫星降水产品在中国大陆的精度评估. 人民长江, 52(5): 50-59) [DOI: 10.16232/j.cnki.1001-4179.2021.05.009]

Zhong Y L. 2021a. Based on the Observation of the Fengyun-4 Satellite AGRI, a Random Forest Algorithm is Used to Invert Ground Precipitation. Nanjing: Nanjing University of Information Science and Technology (钟宇璐. 2021a. 基于风云四号卫星AGRI观测用随机森林算法反演地面降水. 南京: 南京信息工程大学) [DOI: 10.27248/d.cnki.gnjqc.2021.000618]

Zhong Y L. 2021b. Evaluation and verification of FY-4A satellite quantitative precipitation estimation product. Journal of Agricultural Catastropholgy, 11(3): 96-98 (钟宇璐. 2021b. 风云四号卫星定量降水估计产品的检验评估. 农业灾害研究, 11(3): 96-98) [DOI: 10.3969/j.issn.2095-3305.2021.03.041]

Zuo H C, Lü S H and Hu Y Q. 2004. Variations trend of yearly mean air temperature and precipitation in China in the last 50 years. Plateau Meteorology, 23(2): 238-244 (左洪超, 吕世华, 胡隐樵. 2004. 中国近50年气温及降水量的变化趋势分析. 高原气象, 23(2): 238-244) [DOI: 10.3321/j.issn:1000-0534.2004.02.017]

Rapid correction of near real-time FY-4A retrieval based on ensemble machine learning

LYU Yi[1,2], YONG Bin[1,2], SHEN Zhehui[1], LI Ji[1,2], MEI Jun[1,2]

1.National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing 210098, China;
2.College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

Abstract: Satellite remote sensing retrieval is an important way to solve the problem of obtaining near-real-time high-resolution precipitation information. Fengyun-4A (FY-4A) is outfitted with the Advanced Geosynchronous Radiation Imager (AGRI), which boasts world-leading performance. The dual scanning mirrors of AGRI enable precise 2-D pointing, allowing for minute-level regional scans—a groundbreaking achievement. This advanced instrument can capture high-frequency images of the Earth's cloud cover in more than 14 spectral bands. It can generate the official FY-4A REGC (Regional Precipitation Estimation Near-real-time Product for China), which is one of the precipitation estimates information that China can independently obtain from satellite remote retrieval. However, the accuracy of FY-4A REGC still lags behind that of IMERG-Early, the counterpart product of the Global Precipitation Measurement (GPM). Currently, the prevailing approach for correcting satellite-derived precipitation products involves constructing linear prior relationship models between historical satellite rainfall estimates and corresponding ground truth measurements, typically obtained from rain gauges or radar systems. When new observational data become available, this relationship is utilized to derive corrected precipitation values. However, linear models struggle to precisely capture the intricate relationship between satellite rainfall estimates and ground truth measurements. We have observed that ensemble learning methods offer nonlinear models that exhibit advantages such as faster training, reduced data requirements, and robust model stability. In this study, a correction method for official FY-4A precipitation estimates is dynamically constructed using an ensemble machine learning method (LightGBM) with FY-4A REGC as the model input and IMERG-Early as the training calibration for the mainland China region. The revised FY-4A precipitation product (FY-4A Adj) was compared with the original FY-4A REGC using the CMPA automatic gauge observations as the ground reference. The Correlation Coefficient (CC), Root Mean Square Error (RMSE), and relative bias (Bias) of FY-4A Adj were found to be improved significantly compared with those of FY-4A REGC. The revised algorithm effectively reduced the significant overestimation of the original FY-4A REGC in southern China. Our investigation revealed that choosing the correct order for training information significantly enhances model accuracy, with this study opting for training order 2^{21}. In practical applications, the ensemlde learning model can continually optimize its model parameters and performance by dynamically adjusting to the latest training data in real time. We also conducted a comparative analysis of two classes of methods employing ensemble learning, namely, bagging and boosting. Our findings indicate that the Random Forest method performs better when working with limited data volumes, while LightGBM is the recommended choice for large datasets. In conclusion, the correction method based on ensemble machine learning proposed in this paper can quickly and effectively improve the near-real-time Precipitation estimates of FY-4A REGC. This method provides guidance for producing high-quality satellite precipitation products based on FY-4A.

Key words: remote sensing, Fengyun satellite, FY-4A, ensemble learning, near real-time precipitation estimates, precipitation estimates correction

Supported by National Key Technologies Research and Development Program of China (No.2021YFB3900601); National Natural Science Foundation of China (No.U2243229)