姜晓飞,章丽娜,张昕,等.2024.青藏高原夏季FY-4A卫星对流初生产品的分类识别[J].暴雨灾害,43(2):214-223. JIANG Xiaofei, ZHANG Lina, ZHANG Xin, et al. 2024. Classification and identification of FY-4A convective initiation products in summer on the Qinghai-Tibet Plateau[J]. Torrential Rain and Disasters, 43(2):214-223 (in Chinese). doi:10.12406/byzh.2023-193

青藏高原夏季FY-4A卫星对流初生产品的分类识别

姜晓飞[1],章丽娜[1],张昕[1],姚爽[2]
(1. 中国气象局气象干部培训学院,北京 100081;2. 国家气象信息中心,北京 100081)

摘　要: 为了解并提升风云四号卫星A星(FY-4A)对流初生(Convective Initiation,CI)产品对青藏高原夏季降水的指示意义,基于FY-4A CI产品及全球降水测量计划(Global Precipitation Measurement,GPM)降水数据,根据青藏高原地区2020—2022年6—8月FY-4A CI产品识别出的CI样本与1 h后实际观测降水的对应关系,将CI样本划分为无降水CI、弱降水CI和强降水CI三类,并结合大气对流参数与地理位置等信息,利用决策树和随机森林两种机器学习算法建立CI类别识别模型并检验,结果表明:青藏高原地区对流初生后1 h内的降水情况存在明显区域差异,其西北部无降水比例高而东南部降水的比例高;利用抬升指数、云水总量、垂直风切变、中低层湿度、云底高度、零度层高度等大气对流参数信息,能较好区分青藏高原CI出现后是否有降水及降水的强弱;随机森林识别模型结果对于CI类别的识别效果优于决策树识别模型结果,利用随机森林识别模型可以更有效地对青藏高原夏季CI按照降水强度的分类进行识别。

关键词: FY-4A;对流初生;青藏高原;决策树;随机森林
中图法分类号: P409　　**文献标志码:** A　　**DOI:** 10.12406/byzh.2023-193

Classification and identification of FY-4A convective initiation products in summer on the Qinghai-Tibet Plateau

JIANG Xiaofei[1], ZHANG Lina[1], ZHANG Xin[1], YAO Shuang[2]
(1. *CMA Training Centre*, *Beijing* 100081; 2. *National Meteorological Information Centre*, *Beijing* 100081)

Abstract: This study aims to understand and enhance the indicative significance of convective initiation (CI) products from Fengyun-4A (FY-4A) satellite for summer precipitation over the Qinghai-Tibet Plateau. Using the convective initiation (CI) products of FY-4A and precipitation data from the Global Precipitation Measurement Program (GPM), and based on the correspondence between the CI samples identified by the FY-4A CI product and the actual observed precipitation one hour after identifying the CI in the Qinghai-Tibet Plateau region from June to August 2022, three categories of the CI samples, including no precipitation CI, weak precipitation CI, and strong precipitation CI, were divided. Then a CI class recognition model was established and the model performance testing was conducted by combining atmospheric convective parameters and geographic location information, two machine learning methods, decision tree and random forest. The results show that there are significant regional differences in the precipitation situation within one hour after the occurrence of CI in the Qinghai Tibet Plateau region, with a higher proportion of no precipitation in the northwest region and a higher proportion of precipitation in the southeast region. By utilizing atmospheric convective parameters such as lift index, total cloud water, wind shear, middle and low level humidity, cloud bottom height, zero degree layer height and so on, it is possible to better distinguish whether there is precipitation and the strength of precipitation after the appearance of CI in the Qinghai Tibet Plateau. The random forests identify model have better performance for CI classification than decision tree, and the use of random forests identify model can more effectively classify summer CI on the Qinghai Tibet Plateau according to precipitation intensity.

Key words: FY-4A; convective initiation; Tibet Plateau; decision tree; random forest

收稿日期: 2023-09-18;**定稿日期:** 2023-12-21
资助项目: 中国气象局风云卫星应用先行计划项目(FY-APP-20220102);国家自然科学基金重点项目(42030611);第二次青藏高原综合科学考察研究项目(2019QZKK0105);中国气象局干部培训学院重点项目(2023CMATCZDIAN06)
第一作者: 姜晓飞,主要从事人工智能在卫星资料中的应用研究。E-mail: 276708007@qq.com
通信作者: 章丽娜,主要从事强对流天气及相关方法研究。E-mail: zhangln@cma.gov.cn

引 言

对流初生(Convective Initiation，CI)是强对流天气开始活动的标志，CI的识别和追踪是提高局地突发强对流天气短时临近预报水平的关键，也是中尺度气象研究的重点和热点(崔新艳等，2021)。国内外相关研究中普遍采用的CI定义为多普勒天气雷达第一次检测到对流云的回波反射率因子≥35 dBz(寿绍文等，1993；Roberts and Rutledge，2003)，随着卫星遥感技术的不断发展，利用卫星对CI进行识别逐渐成为新的热点。多光谱地球静止气象卫星具有高时空分辨率和高光谱覆盖率的观测优势，可以比地面雷达更早地探测到积云和检测对流的发展，因此利用卫星识别和监测CI具有一定优势。

Mecikalski和Bedka (2006)、Mecikalski等(2010)最早提出了基于地球静止气象卫星资料的对流分析追踪方法(Satellite Convection Analysis and Tracking，SATCAST)来对CI进行识别，并在国内外CI的监测业务和研究中得到了广泛应用(Okabe et al.，2011；李五生等，2014)。国内基于风云四号卫星研发了快速对流监测(Rapid Developing Convection，RDC)产品用于识别监测CI以及强对流系统(Sun et al.，2019)，RDC识别CI的算法与SATCAST方法类似，已应用于气象卫星风云四号A星(FY-4A)产品系统中(Yang et al.，2017)。卫星识别CI的算法主要从卫星多通道的亮温数据中提取出包括云顶亮温、多通道亮温差、云顶亮温时间变化趋势等多个指标，通过阈值判断是否有CI出现。这些算法直接使用亮温数据进行CI识别，能充分发挥卫星高时间频次观测的优势，然而未考虑其他观测资料的结合问题(姚秀萍和曹晓敏，2023)。随着观测和预报数据的增加，利用机器学习算法融合多源数据，可以提高CI识别的时效性和精确性(崔林丽，2022)。Mecikalski等(2015)结合地球同步轨道环境卫星(Geostationary Operational Environmental Satellites，GOES)数据和基于数值天气预报资料计算的大气对流参数，使用逻辑回归和随机森林算法进行CI的识别，有效地识别出未产生降水的CI。Apke等(2015)分析了对流前环境变量对于CI的作用，表明利用对流有效位能和对流抑制能量对于识别出未产生降水的CI具有积极的作用。基于静止卫星的CI研究呈现出从仅使用卫星资料到使用多种观测资料，从关注CI过程本身到关注CI前环境条件和CI后续对流强度的发展趋势(黄亦鹏等，2019)。

我国新一代静止气象卫星FY-4A搭载了多项世界先进探测仪器，具有高时空分辨率和较高光谱分辨率，能提供包括CI在内的多种对流天气监测产品。FY-4A CI产品提供了初生的对流识别结果，可以用于临近预报，然而该结果只表明是否识别到CI，无法提供CI出现后是否会产生降水以及降水强弱等信息，这也影响了该产品的应用效果。对临近预报而言，识别和监测哪些对流能发展起来并最终会形成降水系统尤为重要(覃丹宇和方宗义，2014)，在识别出CI的基础上，结合与对流相关的大气对流参数，了解识别对流初生未来可能对应的降水情况，在预报业务中具有实际意义。青藏高原地形复杂，天气、气候和环流独特(李国平和张万诚，2019；赵思雄和孙建华，2019)，夏季对流发生频率高、预报难度大，而在高原地区缺乏雷达资料和其他观测资料，卫星是实时监测CI的最佳选择，因此有必要对卫星识别出的高原CI与CI出现后降水的关系进行研究，以提高卫星CI产品在高原地区的适用性，为高原降水以及强降水的预报预警提供参考。

本文首先利用全球降水测量计划(Global Precipitation Measurement，GPM)降水数据，统计分析了青藏高原地区2020—2022年6—8月FY-4A CI产品识别出的CI样本与识别出CI后1 h实际观测降水的对应关系，根据CI出现后1 h的降水量大小将CI样本分为CI发生后无降水、CI发生后有弱降水和CI发生后有强降水三类(以下简称无降水CI、弱降水CI和强降水CI)，并利用决策树和随机森林两种常用的机器学习算法，基于与对流发生发展关系密切的大气对流参数等信息，对CI样本的分类结果进行学习，建立基于大气对流参数等信息的CI类别识别模型并对识别结果进行检验，以期能在FY-4A识别出青藏高原夏季CI后进一步结合大气对流参数等信息得到CI出现后可能的降水情况，为高原局地对流降水短临预报预警提供参考。

1 资料与方法

1.1 资料说明

本文的研究区域为75—105°E，25—40°N范围内海拔3 000 m以上的青藏高原地区，资料时段为2020—2022年夏季(6—8月)共9个月。

本文使用的CI产品是FY-4A先进的静止轨道辐射成像仪(AGRI) 2级产品中的对流初生实时产品，产品内的数据主要包括CI判识的二维矩阵、对流云顶降温率强度等级以及相关的地理信息数据，其中CI判识的二维矩阵给出了每个扫描点的CI识别结果，当判识值为-1时代表卫星在对应扫描点监测到CI发生。该产品包含全圆盘、区域两种空间范围，时间分辨率最高可达到5 min左右，星下点分辨率4 km，产品从2019

年8月开始提供。本文共收集了2020—2022年夏季(6—8月)的FY-4A CI产品中的CI判识矩阵，由于CI判识矩阵对应的是各扫描点数据，为了方便与再分析数据进行匹配，采用最邻近插值法(龙四春等，2015)，将CI判识矩阵处理成时间分辨率为1 h、空间分辨率为0.25°的等经纬度投影格点数据，当某个格点CI判识值为-1时代表识别出该格点有CI发生，即为一个CI样本。

由于青藏高原地面观测站稀疏，因此使用GPM降水数据作为实际观测降水结果，GPM降水数据是利用多传感器、多卫星、多算法，结合卫星网络和雨量计反演得到的更高精度降水数据(曲学斌等，2020)，其提供1级、2级和3级产品数据。其中，3级产品数据是在2级产品基础上对固定时间和空间尺度进行插值，数据完整性和一致性较高(Smith，2007)。Ma 等(2016)评估了GPM降水资料在青藏高原上的表现，认为该降水产品整体表现较好，因此本文以GPM 3级降水产品作为观测降水量，当某个时次、某格点的CI判识值为-1时，根据该格点周围75 km半径范围内未来1 h的降水量大小按照阈值对该CI样本进行分类，将CI样本分为无降水CI、弱降水CI和强降水CI三类。GPM 3级降水产品的时间分辨率为0.5 h，空间分辨率为0.1°，空间上将其插值为0.25°，时间分辨率的处理及降水量划分阈值详见1.4节。

大气中常用物理量来自欧洲中期天气预报中心第五代再分析资料，其时间分辨率为1 h，空间分辨率为0.25°。此外，基于该资料进行了大气对流参数的计算。

1.2 机器学习算法

CI出现后的降水情况涉及对流的产生与触发问题，本文利用影响对流的热力、动力、水汽等大气对流参数与地理位置信息，基于机器学习算法对无降水CI、弱降水CI和强降水CI三种分类情况进行学习，以得到CI出现后对降水影响重要的大气对流参数，并利用机器学习算法建立基于大气对流参数等信息的CI类别识别模型，本文使用决策树和随机森林两种算法分别对高原的东南部和西北部建立CI类别识别模型，并利用网格搜索和多轮调参进行超参数优化，得到模型的最优参数，利用最优参数模型对CI的类别进行学习和效果检验。

1.2.1 决策树算法

决策树是一种典型的分类算法，本文利用决策树在树形结构的基础上直接模仿现实生活中人类做决策的过程特点，尝试得到对CI分类影响较大的对流参数及阈值。首先提取所有CI样本对应的类别以及大气对流参数等信息，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。决策树是一个树结构，其每个非叶节点表示一个特征属性(对流参数等信息)上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。本文使用C4.5决策树算法(Quinlan，1993)进行CI类别的学习，C4.5决策树算法会遍历每个特征和阈值，计算划分数据集后的信息增益率并选择信息增益率最大的特征对数据集进行划分，之后递归地处理被划分后的所有子数据集，最终可以得到每个节点的特征和阈值。

1.2.2 随机森林算法

随机森林本质上属于机器学习中的集成学习，是将许多棵决策树整合成森林从而实现一个预测效果更好的集成分类器(刘杰等，2024)。随机森林采用Bagging的思想(Breiman，2001)，即每次有放回地从训练集中取出n个训练样本，组成新的训练集。利用新的训练集，训练得到M个子模型，对于分类问题，采用投票的方法，得票最多子模型的分类类别为最终的类别。随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元共线性不敏感，对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用，被誉为当前最好的算法之一(Iverson et al.，2008)。

1.3 检验方法

本文建立了基于大气对流参数等信息的CI三分类识别机器学习模型，将所有样本的80%作为训练集进行学习，剩余20%的作为测试集进行效果检验。将模型在测试集上基于大气对流参数等信息判断出的CI类别与通过降水观测得到的CI类别合成混淆矩阵，对模型的判断效果进行检验。表1给出了二分类情况下的混淆矩阵，混淆矩阵的每一列代表了模型的判断类别及判断为该类别的数据数目，每一行代表了数据的真实归属类别及该类别的数据实例数目。

表1 二分类混淆矩阵

Table1 Binary classification confusion matrix

观测	判断	
	负例	正例
负例	真负例T_N (True Negative)	假正例F_P (False Positive)
正例	假负例F_N (False Negative)	真正例T_P (True Positive)

根据混淆矩阵的真负例、假负例、真正例、假正例，可以计算准确率A_C (Accuracy)，命中率P_{OD} (Probability of Detection)，精确率P_R (Precision)用于评价模型的分类的性能，计算公式如下

$$A_C = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{1}$$

$$P_{OD}=\frac{T_P}{T_P+F_N} \tag{2}$$

$$P_R=\frac{T_P}{T_P+F_P} \tag{3}$$

其中A_C代表模型判断正确的样本占总样本的百分比，反映模型对正负样本整体判断的能力；P_{OD}为判断正确的正样本占总的正样本的比例，反映模型正确判断正样本全度的能力；P_R代表判断的正样本中有多少比例是真实的正样本，反映模型正确判断正样本精度的能力。对于本文处理的三分类问题，将关注的类别作为一类，其他所有类作为另一类，即转变成二分类问题。 例如在关注第一类别(无降水CI)识别效果时，将第一类别作为正例，第二和第三类别均作为负例，即可通过上述准确率、命中率、精确率的计算对第一类别的识别情况进行检验，即检验机器学习模型利用大气对流参数等信息对无降水CI类别的识别效果。

1.4 数据预处理

1.4.1 CI产品处理

参考Walker等(2012)提出的CI模糊检验方法对FY-4A CI产品判识出的CI样本进行分类。具体做法为：首先将识别出CI的时刻记为t_0，在以CI识别点为圆心、75 km为半径范围内的GPM降水总和(t_0+0.5 h和t_0+1 h两个时刻)作为与该CI匹配的降水量，根据降水量是否超过0.1 mm先将CI样本分为有降水和无降水两类；然后在有降水的样本中，以75百分位数作为阈值，将有降水的样本分为弱降水和强降水(弱降水和强降水阈值分别为小时降水量0.1 mm和12.3 mm)；最终将识别出的CI样本根据CI出现后1 h降水量将CI分为无降水CI，弱降水CI和强降水CI三类，对这三类CI样本分别用‘0，1，2’标签表示。

利用收集整理的FY-4A CI判识数据，当一个格点CI判识值为-1时记为一条CI样本，共得到39 840个CI总样本。基于上述阈值分类方法，无降水CI样本13 648个占比为34.3%，弱降水CI样本和强降水CI样本的占比分别为49.3%和16.4%。

1.4.2 预报因子选取

根据CI发生的动力、热力学过程，选取了37个与CI关系较为密切的大气对流参数和位置参数作为预报因子。其中，水汽类参数有8个，包括整层大气可降水量，整层水汽散度，300、400、500 hPa水汽散度，300、400、500 hPa温度露点差。不稳定度类参数有9个，包括沙氏指数(SI)、K指数、抬升指数(LI)、最大抬升指数(BI)、全总指数(TT)、500与250 hPa假相当位温差、400与250 hPa假相当位温差、对流有效位能、对流抑制。动力类参数有13个，包括300、400、500 hPa垂直速度，300、400、500 hPa涡度，300、400、500 hPa位势涡度，300、400、500 hPa与地面垂直风切变，250与450 hPa垂直风切变。特殊高度厚度类的参数有4个，包括云底高度、零度层高度、自由对流高度、自由对流高度温度。此外，还选取了3个与地理位置相关的参数，分别是经度、纬度、地形高度。

由于青藏高原海拔较高，有些对流参数需要重新定义和计算才能使用。本文将SI指数的定义修改为气块从500 hPa开始沿干绝热过程上升至抬升凝结高度，再从抬升凝结高度沿湿绝热过程抬升至250 hPa环境温度与气块的温度之差(王凌云和李国平，2017)。K、LI、BI、TT等指数参考了尤伟等(2012)和赵定池等(2017)的方法进行重新计算。其他热动力参数则直接来源于欧洲中期天气预报中心第五代再分析数据或者按照通常的计算方法。

2 青藏高原FY-4A CI空间分布及分区

图1为2020—2022年6—8月青藏高原海拔3 000 m以上地区CI样本个数的空间分布和各类样本占比的空间分布，整个青藏高原海拔3 000 m以上地区CI的空间分布呈现出中部偏多、东部和西部偏少的特征(图1a)。无降水CI和有降水CI的空间分布特征各不相同。青藏高原西北部的无降水CI样本较多(图1b)，以蓝线为分界线，将青藏高原海拔3 000 m以上地区分为西北部和东南部两个区域，由表2统计情况可知，西北部无降水CI的平均占比为50.0%，而东南部无降水CI的平均占比为27.7%。与无降水CI的空间分布特征相反，弱降水CI和强降水CI的占比分布则是东南部明显高于西北部(图1c、d)，这也与高原夏季降水量的空间分布一致。

整体而言，FY-4A在青藏高原夏季识别出的CI样本中，有一定比例的无降水CI，因此在利用FY-4A CI产品进行降水分析和预报时，有必要结合其他资料有效区分出CI后的降水情况，为预报员提供参考。考虑到青藏高原范围较大，不同地区有不同的气候和降水特征，并且三类CI样本占比也具有明显的地区差异，因此将青藏高原按照图1中蓝线分为西北部和东南部两个区域，进行后续的建模分析。

3 结果与分析

基于CI分类结果与提取的大气对流参数等信息，首先利用决策树算法建立CI类别识别模型并得到决策树的分类规则，然后利用随机森林提升模型的效果，并分别对决策树和随机森林两种机器学习算法建立的CI类别识别模型进行效果检验。

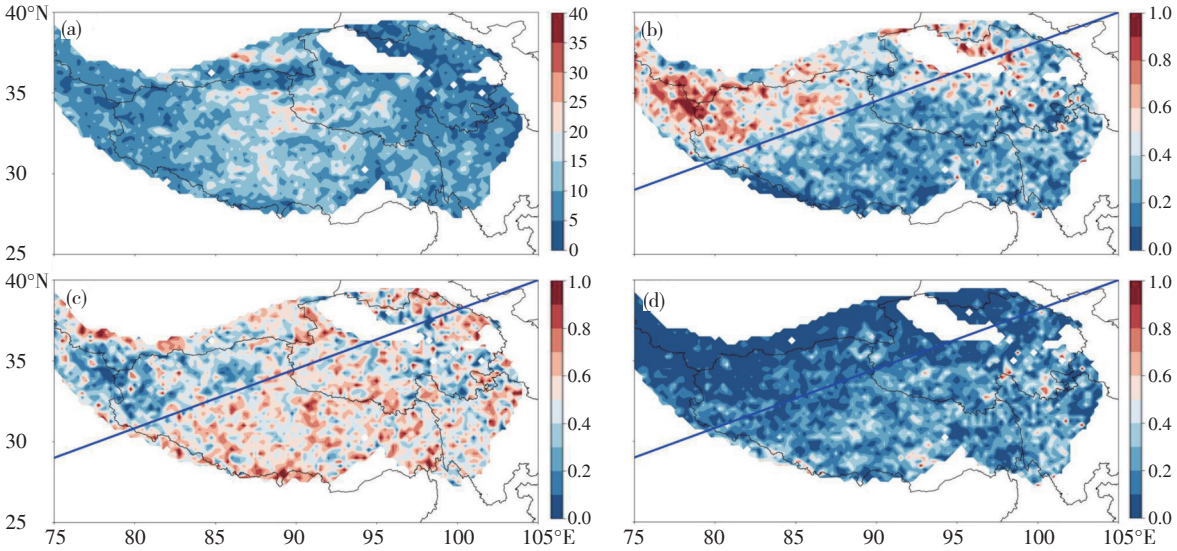

图1 2020—2022年6—8月青藏高原地区海拔3 000 m以上区域总CI样本数(a,单位:个数),无降水CI占比(b,单位:%),弱降水CI占比(c,单位:%),强降水CI占比(d,单位:%)的空间分布
(图中蓝线表示按照CI分布特征分为东南部和西北部分界线)

Fig.1 The spatial distribution of (a) total CI (unit: number), (b) proportion of no precipitation CI (unit: %), (c) proportion of weak precipitation CI(unit: %), (d) proportion of heavy precipitation CI (unit: %) in the Qinghai–Tibet Plateau over 3 000 m from June to August 2020–2022 (the blue line represents the boundary between the southeast and northwest parts according to the distribution characteristics of CI)

表2 青藏高原不同区域三类CI样本统计情况

Table2 Statistics of three types CI samples in different regions of Qinghai–Tibet Plateau

区域	样本数				
	总样本数	无降水CI	弱降水CI	强降水CI	空报率
西北	11 745	5 868	5 120	757	50.0%
东南	28 095	7 780	14 504	5 811	27.7%
总计	39 840	13 648	19 624	6 568	34.3%

3.1 决策树识别模型结果分析

利用决策树识别模型,基于测试集数据对高原西北部和东南部两个地区的CI类别进行识别,得到的混淆矩阵如图2所示。图2中从左上到右下对角线上的样本数越高,代表分类效果越好。总体而言,决策树识别模型对无降水CI(标签0)和弱降水CI(标签1)的识别能力较好,而对于强降水CI(标签2)的识别能力较差,这主要由于样本数量不平衡导致(强降水CI样本少)。对于西北部地区(图2a),模型测试样本准确率为61%,其中无降水CI和弱降水CI这两个类别识别效果

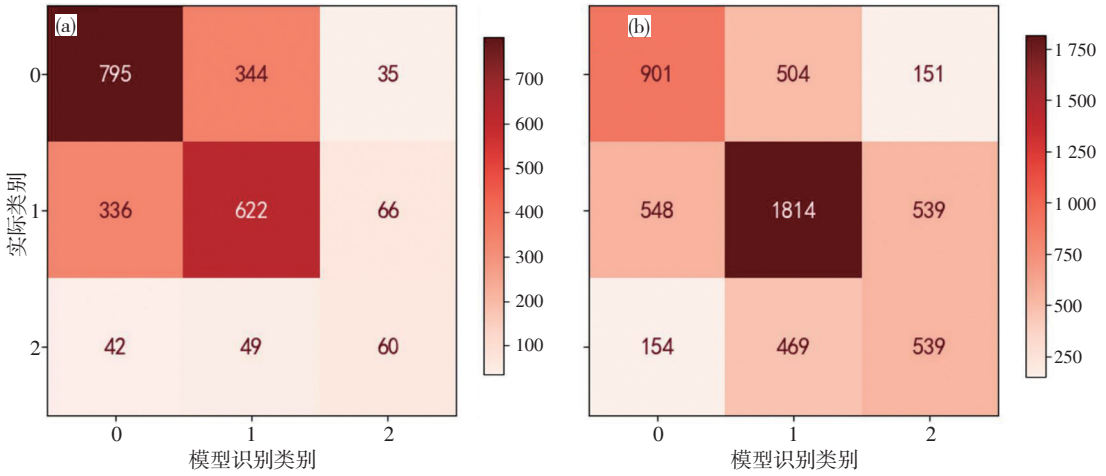

图2 青藏高原西北部(a)和东南部(b)决策树识别模型对测试集识别得到的混淆矩阵
(横纵坐标中,0代表无降水CI,1代表弱降水CI,2代表强降水CI)

Fig.2 Confusion matrix obtained by decision tree identify model for test samples in (a) northwest, (b) southeast of the Qinghai–Tibet Plateau(In the horizontal and vertical coordinates, 0 represents no precipitation CI, 1 represents weak precipitation CI, and 2 represents strong precipitation CI)

略好，命中率分别为68%和61%，而强降水CI命中率较低(40%)，三个类别的识别精确率分别为68%、61%、37%。对于高原东南部地区(图2b)，决策树识别模型的准确率为58%，无降水CI、弱降水CI和强降水CI的识别命中率分别为58%、63%、46%，精确率分别为56%、65%、44%。

综上，决策树识别模型对西北部地区无降水CI的样本识别最好，这表明结合FY-4A CI产品以及大气对流参数及阈值可以有效识别有CI出现后无降水的情况。而决策树识别模型对东南部地区弱降水CI的区分效果最好，命中率和精确率均高于其他类别，表明结合FY-4A CI产品以及大气对流参数及阈值可对CI出现后产生弱降水的情况进行有效识别。

3.2 决策树识别模型的决策结果

决策树识别模型对CI出现后是否会产生降水有一定的识别能力，因此可以进一步利用决策树识别模型学习到的决策结果得到有参考价值的决策信息，在实际工作中为是否产生降水提供依据。在决策树识别模型中，从根节点到每一个叶子节点均可以通过一系列条件判断得到，并且可以给出相应的判断阈值。训练好的高原西北部和高原东南部决策树识别模型示意图分别如图3和图4所示，图中叶子节点用方形表示，方形中的数字代表落在该叶子结点中的样本数。在决策树节点中的参数都是对分类影响重要的因子，越接近根节点参数的重要性越高。为了展示方便，图3和图4只给出了前3层决策树的结果。

对于高原西北部地区(图3)，最重要的参数是云水总量。当云水总量小于等于8.42 kg·m^{-2}时，绝大多数为无降水样本。当云水总量大于8.42 kg·m^{-2}并且抬升指数小于1.15 K时均为有降水(包括强和弱)样本，

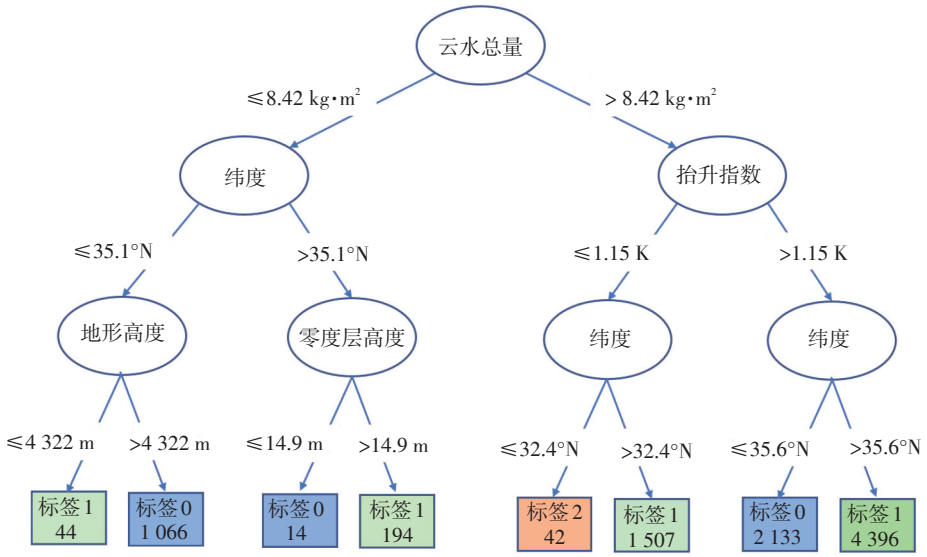

图3 青藏高原西北部CI分类决策树识别示意图

Fig.3 Decision tree identify diagram of CI classification in the northwest of the Qinghai-Tibet plateau

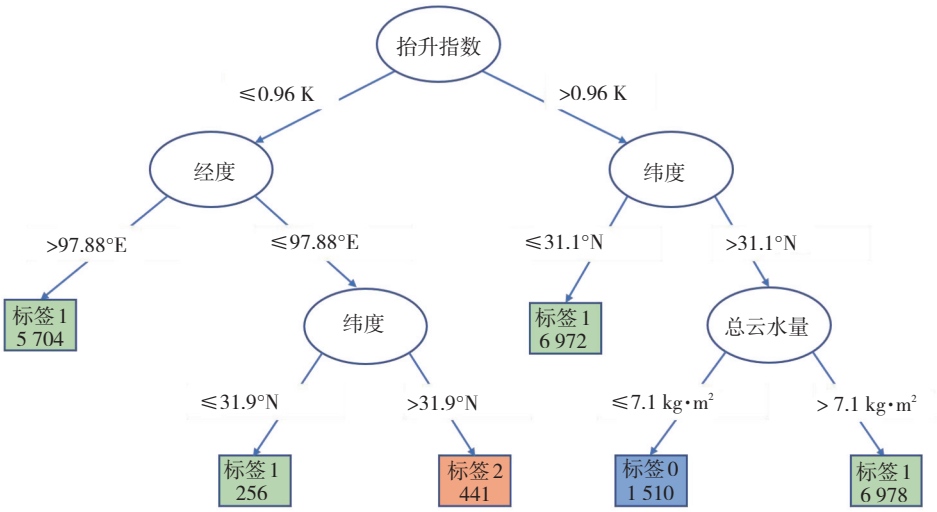

图4 青藏高原东南部CI分类决策树识别示意图

Fig.4 Decision tree identify diagram of whether CI classification in the southeast of the Qinghai-Tibet plateau

表明水汽条件和不稳定条件较好时更容易产生降水。当云水总量大于8.42 kg·m^{-2}并且抬升指数大于1.15 K,即水汽条件较好而不稳定条件一般时,需进一步结合纬度等信息对是否有降水进行区分。对于高原东南部地区(图4),最重要的参数是抬升指数。当抬升指数小于等于0.96 K,会产生弱降水或强降水。当抬升指数大于0.96 K、纬度大于31.1°N、总云水量小于等于7.1 kg·m^{-2}时,为无降水样本。表明当不稳定程度较低、水汽含量较少时,即使卫星识别了CI,东南部地区大概率也不会产生降水。

总体而言,对于整个高原,代表不稳定程度的抬升指数,代表水汽含量的总云水量以及代表当地气候特点的位置信息是CI分类的重要依据。

3.3 随机森林识别模型结果分析

为了进一步提升分类效果和模型的泛化能力,使用随机森林识别模型采用有放回抽样构建500棵决策树,利用这500棵树的投票结果对青藏高原CI进行分类识别。随机森林识别模型对测试集识别得到的混淆矩阵如图5所示。对于高原西北部地区(图5a),随机森林识别模型准确率为75%,无降水CI、弱降水CI和强降水CI的命中率分别为77%、78%、34%,精确率分别为79%、71%、65%。这表明模型对无降水CI和弱降水CI有较好的识别能力。强降水CI的命中率较差而精确率较好,表明虽然有些强降水CI无法识别,但对于识别出的强降水CI准确率较高。对于东南部地区(图5b),随机森林识别模型准确率为75%,其中无降水CI、弱降水CI和强降水CI的命中率分别为66%、88%、51%,精确率分别为83%、72%、77%。这表明可以识别出大多数弱降水CI的样本,而无降水CI和强降水CI样本识别的精确率较好。

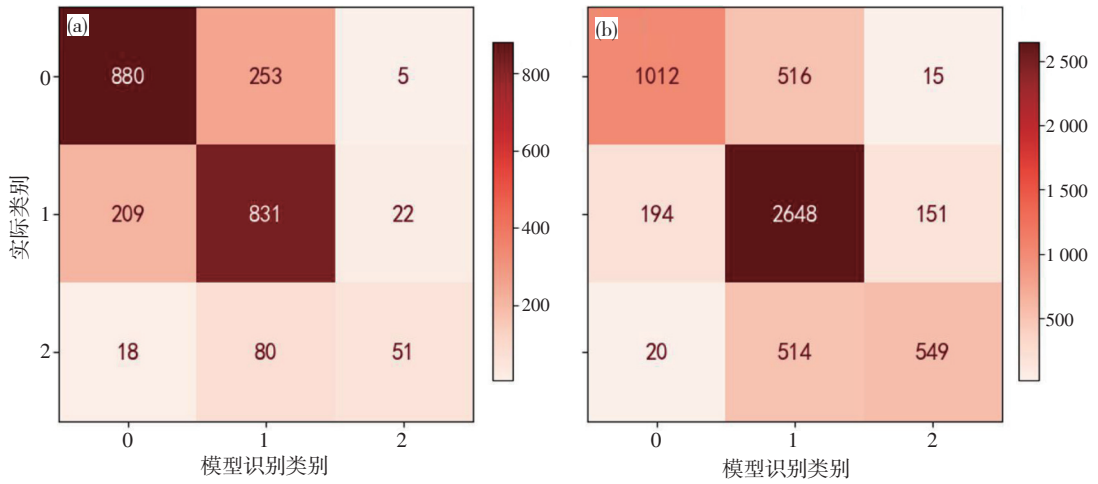

图5 青藏高原西北部(a)和东南部(b)随机森林识别模型对测试集识别得到的混淆矩阵

Fig.5 Confusion matrix obtained by random forest identify model for test samples in (a) northwest and (b) northwest of the Qinghai-Tibet plateau

对比随机森林识别模型和决策树识别模型,高原西北部地区在无降水CI和弱降水CI的分类命中率和精确率均提升10%以上,其中对弱降水CI命中率的提升较大(17%),而强降水CI的命中率有所下降,但精确率提升28%。这表明随机森林识别模型能识别出更多的无降水和弱降水样本,同时能减少各类别的空报情况,尤其是大幅降低了强降水的空报。对高原东南部地区,随机森林识别模型在各类别CI的命中率和精确率均有所提升,其中弱降水CI的命中率提升明显(提升22%),即能识别出更多的弱降水。无降水CI和强降水CI的精确率提升较大,分别为27%、33%,大幅减少了决策树识别模型无降水CI和强降水CI的空报。对比表明,随机森林识别模型比决策树识别模型有更好的分类识别效果,可以提升FY-4A CI产品在青藏高原地区的实用性并且对FY-4A CI产品的细化提供参考。

3.4 随机森林模型特征重要性分析

为了评估各个参数在随机森林识别模型分类中的重要性,利用训练好的模型,将某一个参数的数据进行随机打乱后再重新进行预测,模型性能的衰减量代表该参数的重要程度。使用该方法对每一个参数进行重要性的计算,并将每个参数的重要性除以所有参数重要性的总和,得到的前十个重要参数百分比,如图6所示。在37个参数中,对高原西北部CI分类判断最重要的参数依次为抬升指数、纬度、总云水量、250与450 hPa风切变、沙氏指数、300 hPa与地面风切变、经度、300 hPa温度露点差、500与250 hPa相当位温差、全总指数。可以看到纬度和经度也对高原西北部CI分类有重要的影响。对高原东南部CI分类判断最重要的参数依次为纬度、总云水量、云底高度、抬升指数、300 hPa温度露点差、250与450 hPa风切变、500 hPa与地面风切变、500 hPa温度露点差、经度、零度层高度。

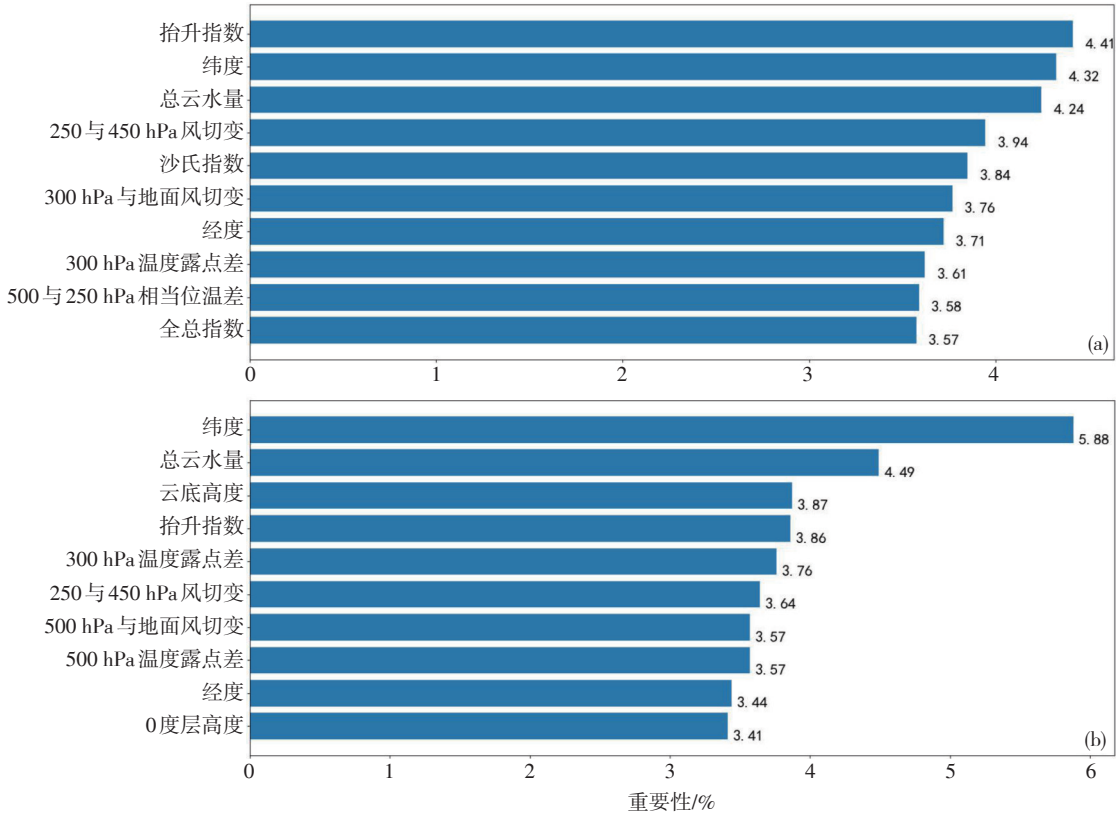

图6 青藏高原西北部(a)和东南部(b)随机森林模型得到的参数重要性示意图

Fig.6 Feature importance from random forest in (a) northwest and (b) southeast of the Qinghai-Tibet plateau

随机森林识别模型得到的参数重要性顺序与决策树(图略)略有区别，但整体看来是一致的，主要区别在与参数的排序。由于高原西北部地区出现无降水的概率大，对该地区结合参数对CI分类时，更多的是在判断有无降水，因此参数重要性排序为：不稳定参数、水汽含量和垂直风切变。其中，强降水类对应最强的是不稳定程度、总云水含量和垂直风切变。而高原东南地区出现降水的概率更高，因此在结合参数对CI进行分类时，不稳定度仍是非常重要的参数，但表示水汽条件的参数更为重要，除了总云水量，还需要考虑大气的饱和程度等。

4 结论和讨论

本文基于2020—2022年6—8月GPM降水数据和FY-4A CI产品，统计分析了青藏高原地区夏季FY-4A CI产品识别出的CI样本与1 h后实际观测降水的对应关系，根据CI出现后1 h的降水量大小将CI样本分为无降水CI、弱降水CI和强降水CI三类，并结合与对流相关的大气对流参数和地理信息参数，采用决策树和随机森林两种常用的机器学习分类算法建立CI的三分类识别模型并进行效果检验，并寻找对青藏高原CI出现对降水或强降水影响重要的参数。由于青藏高原夏季无降水CI样本数呈现西北高于东南的分布特征，为尽可能消除地理位置的影响，分别对高原西北部和东南部建立CI识别模型并对比识别后的结果，得到主要结论如下：

(1) 决策树识别模型对青藏高原夏季CI出现后是否产生降水有一定的识别能力，利用决策树每个分支的条件判断，可以总结出易于理解、符合对流天气发生原理的、有参考价值的CI分类判断阈值及规则。

(2) 随机森林识别模型对青藏高原夏季CI分类效果相比决策树识别模型有进一步提升，这表明结合FY-4A CI产品和其他大气对流参数等信息，可以较好地判断有无降水及降水强度等情况，对于FY-4A CI产品在青藏高原地区的应用及CI产品的分类有一定参考意义。

(3) 综合决策树和随机森林识别模型结果，对青藏高原地区夏季基于降水特征的CI分类影响重要的特征主要包括：位置、抬升指数、云水总量、各层风切变、低层温度露点差、0 ℃层高度等，这些结果也可以对青藏高原初生对流的发生机理和局地降水预报预警提供一定参考。

本文对青藏高原地区夏季基于降水特征的CI分类影响重要的特征研究表明，位置信息对CI分类影响较大，未来可以进一步细化研究范围，对更小的区域或者单站建立机器学习模型，不考虑经纬度等位置的影响，可能会得到更好的分类效果以及对当地分类影响较大的因子。

参考文献(References):

崔林丽.2022.新型静止气象卫星对流初生识别展望[J].气象科技进展,12(5):80-84. Cui L L.2022. Prospect of convective initiation identification based on new generation geostationary meteorological satellite [J]. Advances in Meteorological Science and Technology,12(5):80-84. doi:10.3969/j.issn.2095-1973.2022.05.011

崔新艳,陈明轩,秦睿,等.2021.对流初生机理的研究进展[J].气象,47(11):1297-1318. Cui X Y, Chen M X, Qin R, et al.2021. Research advances in the convective initiation mechanisms [J]. Meteorological Monthly, 47(11):1297-1318. doi: 10.7519/j.issn.1000-0526.2021.11.001

黄亦鹏, 李万彪, 赵玉春, 等.2019.基于雷达与卫星的对流触发观测研究和临近预报技术进展[J].地球科学进展,34(12):1273-1287. Huang Y P, Li W B, Zhao Y C, et al.2019. A review of radar- and satellite-based observational studies and nowcasting techniques on convection initiation [J]. Advances in Earth Science,34(12):1273-1287. doi:10.11867/j.issn.1001-8166.2019.12.1273

刘杰,刘高平,安晶晶,等.2024.基于机器学习的模式温度预报订正方法研究[J].沙漠与绿洲气象,18(3):1-11. Liu J, Liu G P, An J J, et al. 2024. Research on correction method of model temperature forecast based on machine learning [J]. Desert and Oasis Meteorology,18(3) : 1-11. doi:10.12057/j.issn.1002-0799.2024.03.001

龙四春,周威,文佳胜,等.2015.雷达地形测绘DEM空洞插补方法研究[J].遥感信息,30(4):20-24. Long S C, Zhou W, Wen J, et al. 2015. SRTM DEM voids interpolation method based on MATLAB [J]. Remote Sensing Information,30(4):20-24. doi:10.3969/j.issn.1000-3177.2015.04.004

李国平,张万诚.2019.高原低涡、切变线暴雨研究新进展[J]. 暴雨灾害,38(5):464-471. Li G P, Zhang W C. 2019. Recent advances in the research of heavy rain associated with vortices and shear lines come from the Tibetan Plateau [J]. Torrential Rain and Disasters,38(5):464-471. doi:10.3969/j.issn.1004-9045.2019.05.008

李五生,王洪庆,王玉,等.2014.基于卫星资料的对流初生预报及效果评估[J]. 北京大学学报(自然科学版),50(5):819-824. Li W S, Wang H Q, Wang Y, et al. 2014. Convective initiation forecasting and statistical evaluation based on satellite data [J]. Acta Scientiarum Naturalium Universitatis Pekinensis,50(5):819-824. doi:10.13209/j.0479-8023.2014.084

覃丹宇,方宗义.2014.利用静止气象卫星监测初生对流的研究进展[J].气象,40(1):7-17. Qin D Y, Fang Z Y. 2014. Research progress of geostationary Satellite-Based convective initiation [J]. Meteorological Monthly,40(1):7-17. doi:10.7519/j.issn.1000-0526.2014.01.002

曲学斌,付亚男,袁秀芝,等.2020. GPM-IMERG日降水数据在内蒙古地区的适用性分析[J].暴雨灾害, 39(3):293-299. Qu X B, Fu Y N, Yuan X Z, et al. 2020. Applicability analysis of GPM-IMERG daily precipitation data in Inner Mongolia [J]. Torrential Rain and Disasters,39(3):293-299. doi:10.3969/j.issn.1004-9045.2020.03.010

寿绍文,杜秉玉,肖稳安,等. 1993.中尺度对流系统及其预报[M].北京:气象出版社:182. Shou S W, Du B Y, Xiao W A, et al. 1993. Mesoscale convective systems and their forecasting [M]. Beijing: Meteorological Press:182 (in Chinese)

王凌云,李国平. 2017.应用AIRS卫星资料对一次青藏高原东南部MCSs的对流指数进行分析[J].云南大学学报(自然科学版),39(1):88-97. Wang L Y, Li G P. 2017. An analyze convective index of MCSs in southeastern Tibetan Plateau with AIRS data [J].Journal of Yunnan University (Natural Sciences Edition),39(1):88-97. doi:10.7540/j.ynu.20160348

姚秀萍,曹晓敏.2023.大气对流初生的研究进展与展望[J].大气科学学报,46(6):940-949. Yao X P, Cao X M. 2023. Research progress and prospect of atmospheric convection initiation [J]. Transactions of Atmospheric Sciences,46(6):940-949. doi:10.13878/j. cnki. dqkxxb.20230224001

尤伟,臧增亮,潘晓滨,等.2012.夏季青藏高原雷暴天气及其天气学特征的统计分析[J].高原气象,1(6): 1523-1529. You W, Zang Z L, Pan X B, et al. 2012. Statistical analyses on characteristic and environmental aspect of summer thunderstorm over the Tibetan Plateau [J]. Plateau Meteorology,31(6):1523-1529. doi:CNKI:SUN:GYQX.0.2012-06-007.

赵定池,李毅,尤伟,等.2017.拉萨地区夏季夜间雷暴的物理量指数分析[J].气象与环境科学,40(1):114-119. Zhao D C, Li Y, You W, et al.2017. Physical index analysis of summer night thunderstorms in Lhasa area [J]. Meteorological and Environmental Sciences,40(1):114-119. doi:10.16765/j.cnki.1673-7148.2017.01.016

赵思雄,孙建华. 2019.我国暴雨机理与预报研究进展及其相关问题思考[J].暴雨灾害,38(5):422-430. Zhao S X, Sun J H.2019. Progress of mechanism and forecast for heavy rain in China in recent 70 years [J]. Torrential Rain and Disasters,38(5):422-430. doi:10.3969/j.issn.1004-9045.2019.05.004

Apke J M, Nietfeld D, Anderson M R. 2015.Environmental analysis of GOES-R proving ground convection-initiation forecasting algorithms [J]. Journal of Applied Meteorology and Climatology,54(7):1637-1662. doi:10.1175/JAMC-D-14-0190.1

Breiman L. 2001. Random forests [J]. Machine Learn,45(1):5-32. doi:10.1023/A:1010933404324

Iverson L R, Prasad A M, Matthews S N, et al. 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios [J]. Forest Ecology and Management,254(3):390-406. doi:10.1016/j.foreco.2007.07.023

Ma Y Z, Tang G Q, Long D, et al. 2016. Similarity and error intercomparison of the GPM and its predecessor-TRMM multisatellite precipitation analysis using the best available hourly gauge network over the Tibetan Plateau [J].Remote Sensing,8(7):1-17. doi:10.3390/rs8070569

Mecikalski J R, Bedka K M. 2006. Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery [J]. Monthly Weather Review,134: 49-78. doi:10.1175/MWR3062.1

Mecikalski J R, Bedka K M, Paech S J, et al. 2010. A statistical evaluation of GOES cloud-top properties for nowcasting convective initiation [J]. Monthly Weather Review,136:4899-4914. doi:10.1175/2008MWR2352.1

Mecikalski J R, Williams J, Jewett C, et al. 2015. Probabilistic 0-1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data [J]. Journal of Applied Meteorology and Climatology,54:1039-1059. doi:10.1175/JAMC-D-14-0129.1

Okabe I, Imai T, Izumikawa Y. 2011. Detection of rapidly developing cumulus areas through MTSAT rapid scan operation observations [J]. Meteorological Satellite Centre Technical Note, 55: 69-91

Quinlan J R. 1993.C4.5: Promgrams for Machine Learning [M]. San Mateo: Morgan Kaufmann Publishers,235

Roberts R D, Rutledge S. 2003. Nowcasting storm initiation and growth using GOES-8 and WSR-88D Data[J]. Weather and Forecasting, 18(4): 562-584. doi: 10.1175/1520-0434(2003)018<0562:nsiagu>2.0.co;2

Smith E A. 2007. International Global Precipitation Measurement (GPM) program and mission: an overview [J]. Measuring Precipitation from Space Eurainsat & the Future,18(3):10. doi:10.1007/978-1-4020-5835-6_48

Sun F L, Qin D Y, M M, et al. 2019. Convective initiation nowcasting over China from Fengyun-4A measurements based on TV-L1 Optical flow and BP_Adaboost neural network algorithms [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 12 (11):4284-4296. doi:10.1109/JSTARS.2019.2952976

Walker J R, Mackenzie W M J, Mecikalski J R, et al. 2012. An enhanced geostationary satellite-based convective initiation algorithm for 0-2-h nowcasting with object tracking[J].Journal of Applied Meteorology and Climatology, 51(11):1931-1949. doi:10.1175/JAMC-D-11-0246.1

Yang J, Zhang Z Q, Wei C Y, et al. 2017. Introducing the new generation of Chinese geostationary weather satellites, Fengyun-4 [J]. Bulletin of the American Meteorological Society.98(8):1637-1658. doi:10.1175/BAMS-D-16-0065.1

（责任编辑　张　文）

《暴雨灾害》征稿简则

《暴雨灾害》是中国气象局武汉暴雨研究所主办的中文学术期刊，主要刊载与暴雨、强对流天气及其次生灾害相关的最新研究成果和综合评述。本刊为中国科技核心期刊(中国科技论文统计源期刊)双月刊，欢迎国内外从事大气科学研究和气象业务、服务的科技工作者投稿。

1 来稿要求

1) 论点明确、文字精练、数据可靠。其书写顺序为：中文题名(不超过20个汉字)、作者姓名、单位、摘要(400字左右，一般应包含研究目的、资料、方法、结果和结论)、关键词(4～8个为宜)；英文题名、作者姓名、单位、摘要、关键词；正文；结论与讨论；参考文献。

2) 论文具体格式请参考本刊网站下载中心的《暴雨灾害》论文投稿模板，或参考本刊网站首页最新发表的论文。投稿时，只需上传论文word文档，“版权转让协议及伦理声明”待论文录用后另行提交。

3) 论文首页脚页处附各类资助项目名称(编号)和第一作者姓名、主要研究领域、E-mail。第一作者下方可列一位通信作者姓名、主要研究领域、E-mail，通信作者对该文负有学术责任。

4) 来稿中计量单位一律采用中华人民共和国法定计量单位。

5) 所附插图、表格宜少而精，图表应随文给出，并提供英文图(表)题。一篇论文插图不宜超过10幅，要求图像清晰；附表一般应制成三线表(表中数据宜采用word文档录入)。图(表)中的量和单位用“/”隔开，物理量用斜体，并注明图(表)号、图(表)题、图(表)注以及符号文种、大小写、正斜体、黑白体、上下脚码等。

6) 本刊参考文献采用著者-出版年制，具体参见本刊网站下载中心的“《暴雨灾害》参考文献引用和著录格式”。

2 投稿须知

1) 请勿一稿多投。稿件初审时限为40 d。修改稿务必在45 d内修回(另有约定除外)；否则，视作自动撤稿。若6个月内未见刊用通知，稿件可自行处理。

2) 请作者通过本刊网站(http://www.byzh.org.cn)提交稿件并查询稿件处理进展。

3) 本刊编辑可能会对录用稿件作适当的文字性和技术性删改处理，不同意者，请书面说明。

4) 本刊已入编中国学术期刊(光盘版)、中国期刊网、数字化期刊群万方数据库、中文科技期刊数据库、博看网及中国核心期刊(遴选)数据库。来稿一经录用，将同时被光盘版和数据库收录。若作者不同意收录，请来稿时说明。

5) 本刊对录用稿件收取论文版面费。印刷出版后一次性支付稿酬(包含纸质版、数字版稿酬以及网络版著作权使用费)，并赠送当期《暴雨灾害》2册。

3 通联方式

来信(函)请寄至“430205湖北省武汉市东湖高新技术开发区金融港二路6号《暴雨灾害》编辑部”，也可通过邮箱(byzh7939@163.com)或电话(027-81804908，81804915，81804935)联系。

微信扫一扫
关注《暴雨灾害》