李树文,赵桂香,王一颉,等.基于机器学习的多源实况分析产品和观测数据融合应用试验[J].海洋气象学报,2023,44(1):1-10.
LIShuwen, ZHAO Guixiang, WANG Yijie, et al. Fusion and application experiment of machine learning based multi-source real-time analysis products and observation data[J]. Journal of Marine Meteorology,2023,44(1):1-10. DOI:10.19513/j.cnki.issn2096-3599.2024.01.002.(in Chinese)

基于机器学习的多源实况分析产品和观测数据融合应用试验

李树文[1],赵桂香[2],王一颉[2],陈霄健[3],闫慧[2]
(1. 太原市气象局,山西 太原 030002;2. 山西省气象台,山西 太原 030006;3.山西省气象信息中心,山西 太原 030006)

摘要:利用中国气象局公共气象服务中心地面实况专业服务产品(CARAS_SUR1 km,简称CAR)、国家气象信息中心多源融合实况分析数据(ART_1 km,简称ART)、全国雷达反演降水产品(简记为“RAD”)、风云四号卫星反演降水产品(简记为“SAT”)以及全国气象观测站逐小时资料,应用机器学习方法建立了基于选定位置气温、降水、风向、风速要素的实况融合应用模型(简记为“GBDT模型”)。15 d逐时GBDT融合产品的全国分区域检验结果表明:GBDT气温融合产品在东北、华北、西北、华中、新疆、西藏6个区域较CAR产品和ART产品均有改进,在西藏的改进最明显,在华东和西南GBDT融合产品优于ART产品而逊于CAR产品,在华南和内蒙古GBDT融合产品误差较ART产品、CAR产品略有增加;GBDT降水融合产品在样本偏少的内蒙古较ART产品、CAR产品误差略有增加,其他区域有改进或基本相当;GBDT风速、风向融合产品较ART产品、CAR产品均有较大改进。试验结果表明机器学习方法可应用于融合多源实况分析产品和观测数据开展选定位置气温、降水、风向、风速要素的实况气象信息服务。

关键词:机器学习;多源数据;动态模型;误差分析
中图分类号:P457 **文献标志码**:A **文章编号**:2096-3599(2024)01-0000-00
DOI:10.19513/j.cnki.issn2096-3599.2024.01.000

Fusion and application experiment of machine learning based multi-source real-time analysis products and observation data

LI Shuwen[1], ZHAO Guixiang[2], WANG Yijie[2], CHEN Xiaojian[3], YAN Hui[2]
(1. *Taiyuan Meteorological Bureau*, *Taiyuan* 030002, *China*; 2. *Shanxi Meteorological Observatory*, *Taiyuan* 030006, *China*; 3. *Shanxi Meteorological Information Center*, *Taiyuan* 030006, *China*)

Abstract: Based on the machine learning, an application model (GBDT model) of real-time fusion on temperature, precipitation, wind direction, and wind speed at selected locations is developed by using the professional service product (CAR) of Public Meteorological Service Centre of China Meteorological Administration, the multi-source fusion observation analysis data (ART) of National Meteorological Information Centre, the nationwide radar precipitation retrieval product (RAD), the Fengyun-4 satellite precipitation retrieval product (SAT), and the hourly data of nationwide meteorological observation stations. The regional inspection results of the 15-d hourly GBDT fusion product throughout the country

收稿日期:2023-03-30;**修回日期**:2023-08-02
基金项目:山西省基础研究计划自然科学研究面上项目(202203021211081),山西省气象局面上项目(SXKMSTQ20226305)
第一作者简介:李树文,硕士,高级工程师,主要从事天气预报技术和机器学习方法在气象的应用研究,lsw1989@163.com。
通信作者简介:赵桂香,硕士,正高级工程师,主要从事中尺度数值诊断和灾害天气预报技术研究,liyun0123@163.com。

are as follows. The GBDT temperature fusion product improves compared to CAR and ART in 6 regions: Northeast China, North China, Northwest China, Central China, Xinjiang, and Tibet, with the most significant improvement in Tibet. In East China and Southwest China, GBDT fusion product is superior to ART, but inferior to CAR, and its error slightly increases compared to ART and CAR in South China and Inner Mongolia. The error of GBDT precipitation fusion product has a slight increase compared to ART and CAR in Inner Mongolia, where there are fewer samples, while in other areas, there are improvements or the two are basically equivalent. The GBDT wind speed and direction fusion products have significant improvements compared to ART and CAR. The experiment results indicate that the machine learning method can be applied to fuse multi-source real-time analysis products and observation data, providing real-time meteorological information service of temperature, precipitation, wind direction, and wind speed at selected locations.

Keywords: machine learning; multi-source data; dynamic model; error analysis

引言

目前,气象监测系统不断完备、监测数据日趋精密,形成了以地面观测、大气探空、天气雷达、气象卫星为主的多位一体探测布局,极大地提升了气象服务与保障能力。在重大社会活动、气象灾害应急、个性化商业等服务与保障中,往往需要某一确定经纬度的实况数据。而依赖于地面观测站点布网的传统资料离散化程度较高,难以满足任意位置实况数据的气象服务需求。

为发展无缝隙、全覆盖的高分辨率实况产品,科研人员做出很多努力。早期的研究多以站点观测数据为主,运用数学插值方法形成格点化产品,这些产品在站点密集区效果较好,但在地形复杂、站点稀少的区域并不理想。20世纪90年代随着卫星技术的发展,有学者使用地面降水实况对多卫星集成降水产品进行订正,研发了早期卫星融合降水产品[1-2]。21世纪以来,随着概率密度函数(probability density function,PDF)匹配、最优插值(optimal interpolation,OI)、卡尔曼滤波(Kalman filter,KF)等方法在卫星资料校正中的成熟应用,卫星融合降水产品得到显著改善[3-5]。随着气象雷达的广泛应用,将雷达定量降水评估(quantitative precipitation estimation,QPE)产品[6]与站点降水实况相结合,发展了基于卡尔曼滤波、最优插值、距离反比加权(inverse distance weighted,IDW)等方法的系统误差订正和局部偏差订正技术[7],逐步形成雷达降水融合产品。2014年中国气象局气象探测中心将"概率密度函数+贝叶斯模型平均(Bayes model averaging,BMA)+最优插值"方法引入雷达定量估测产品[8],研制了地面、卫星、雷达三源融合降水产品。与此同时,随着计算机技术的发展,国内外借助数值模式,将站点、雷达和卫星等观测数据进行融合,取得了很多成果[9-13]。

目前逐小时1 km × 1 km高分辨率的格点化实况产品已有了多种选择。2020年7月国家气象信息中心研发的中国区域1 km × 1 km多源融合实况分析产品(简记为"ART_1 km")[14-15]业务试运行,2021年7月根据应用评估成果[16-18]完成产品质量和时效优化。2021年1月中国气象局公共气象服务中心研发的逐小时滚动生成的全国1 km × 1 km地面实况专业服务产品(简记为"CARAS-SUR1 km")业务运行。但在日益精细化的气象业务与服务中,还缺少综合应用这两种分析产品制作任意位置的气象要素实况客观工具方法。如果进一步提高现有格点产品的分辨率,带来的计算量将呈指数级增长。那么,这些产品的日常应用效果如何[19-20]?能否将这些产品融合使用或者在此基础上进一步优化?这方面的研究目前还较少[21-22]。本研究旨在充分利用各类已有的实况分析数据,运用机器学习方法[23-26],研究多源资料实况融合算法,建立基于任意位置的逐时实况分析(气温、降水、风速、风向)模型,并进行对比检验,为实况分析服务提供基础支撑。

1 资料与方法

1.1 资料

文章所用资料为2020年8月1—15日由国家气象信息中心提供的5类全国范围逐小时数据:国家气象信息中心多源融合实况分析产品(ART_1 km,简称ART)、中国气象局公共气象服务中心地面

实况专业服务产品(CARAS_SUR1 km,简称 CAR)、全国雷达反演降水产品(简记为“RAD”)、风云四号卫星反演降水产品(简记为“SAT”)以及实况观测数据。其中:ART 产品包括气温、降水、风速、U 分量和 V 分量,水平分辨率为 0.01°×0.01°,单要素单文件存储;CAR 产品包括气温、降水、露点温度、相对湿度、平均风速、平均风向、平均风 U 分量、平均风 V 分量、极大风 U 分量、极大风 V 分量和地表气压,分辨率为 0.01°×0.01°,单文件多要素存储;RAD 即全国天气雷达定量估测降水,分辨率为 0.01°×0.01°;SAT 即风云四号卫星降水估计实时产品,原始数据平均分辨率为 4 km,按照卫星行列号存储,换算为经纬度后,在中国区域其分辨率约为 0.01°×0.01°。将 ART、CAR、RAD、SAT 等 4 类格点产品作为自变量,实况观测数据作为因变量来构建模型,并将模型输出产品与实况观测数据对比分析检验。

需要特别说明的是,降水是离散数据,模型构建时样本内可能不存在降水。因此,在降水样本选取时,先用观测数据对研究区域内降水要素做筛选得到降水时段,确保取样时段内该区域存在降水。

1.2　方法

1.2.1　梯度提升决策树算法

梯度提升决策树(gradient boosting decision tree,GBDT)是机器学习中一种基于决策树的集成算法(简称“GBDT 算法”),其主要思想是利用弱分类器(决策树)迭代训练以得到最优算法,该算法具有训练效果好、不易过拟合等优点。核心是将预测样本逐次输入到 k 个回归决策树的基分类算法,每次迭代过程用梯度下降减小损失,再由基分类算法的分类条件得到叶子结点值,乘以权重,最后累加得出结果。表达式为:

$$F(x,P_{\text{arm}})=\sum_{k=0}^{K}\alpha_k T_k(x,P_{\text{arm}})=\sum_{k=0}^{K}f_k(x,P_{\text{arm}}),\quad(1)$$

式中:x 为训练样本点,P_{arm} 为 GBDT 算法参数;T_k 为回归决策树;α_k 为每棵决策树的权重系数。k 为第 k 棵子回归决策树($k=0,1,\cdots K$)。

1.2.2　模型参数优化

基于 GBDT 算法的模型(简称“GBDT 模型”)建立后,分别对损失函数、权重缩减系数、最大迭代次数、子采样比例、树节点最大深度等参数调优,以更好地拟合训练数据集,提高模型拟合精度。

1.3　资料分析与处理

1.3.1　数据读取与插值

对数据统一用 Python 处理,逐小时实况观测数据为文本文件,用 pandas 库直接读入气温、降水、风速、风向 4 类要素;ART 和 CAR 是 grib2 格式,使用 xarray、cfgrib、pygrib 库读入,其中 ART 读入 5 类要素,CAR 读入 11 类要素;RAD 为二进制(bin)格式,按照编码方式对 bin 进行解析读入;SAT 为 nc 格式,用 xarray 读入。

用 Python 处理时,为尽可能降低插值方法带来的误差,对 ART、CAR、RAD 和 SAT 等 4 类格点数据统一使用 xarray 库内置插值方法,其中 ART、CAR 直接使用 xarray,RAD 解析后转为 xarray 类型;SAT 在行列号的等距网格中完成插值与提取。

1.3.2　耗时与并行策略设计

由表 1 可见,单时次 5 类数据的读取耗时主要集中在两类 grib2 格式的数据,其中 ART 是单文件单要素,采用并行可提升效率;CAR 是单文件多要素,不宜采用并行。考虑到建模数据需要多个时次,将并行策略用于多时次上,采用多 python 终端带参数执行,而单时次只用线性处理。

表 1　资料读取耗时情况

Table1　Data reading time

数据类型	格式	单时次文件数量	读取要素	读取耗时(1 000 目标为例)
实况观测	csv	1	气温、降水、风速、风向	10 s
ART	grib2	5	气温、降水、风速、风 U 和 V 分量	50 min
CAR	grib2	1	气温、降水、露点、相对湿度、风速、风向、风 U 和 V 分量、极大风 U 和 V 分量、地表气压	40 min
RAD	bin	1	降水	4 min
SAT	nc	1	降水	10 s

需要说明的是,从机器学习 GBDT 算法的理论来讲,适度增加特征数量有利于 GBDT 算法发挥其决策树的优势。但当特征维度超过一定界限后,性能会随特征维度增加而下降,此时需要去除冗余和无关特征。在本文中,特征数量远远小于维度界限,增加特征有利于得到更好的结果。因此,表 1 中提取了露点、相对湿度和地表气压 3 个与气温、降水、风相关但无直接联系的要素作为增加特征量。

2 模型建立

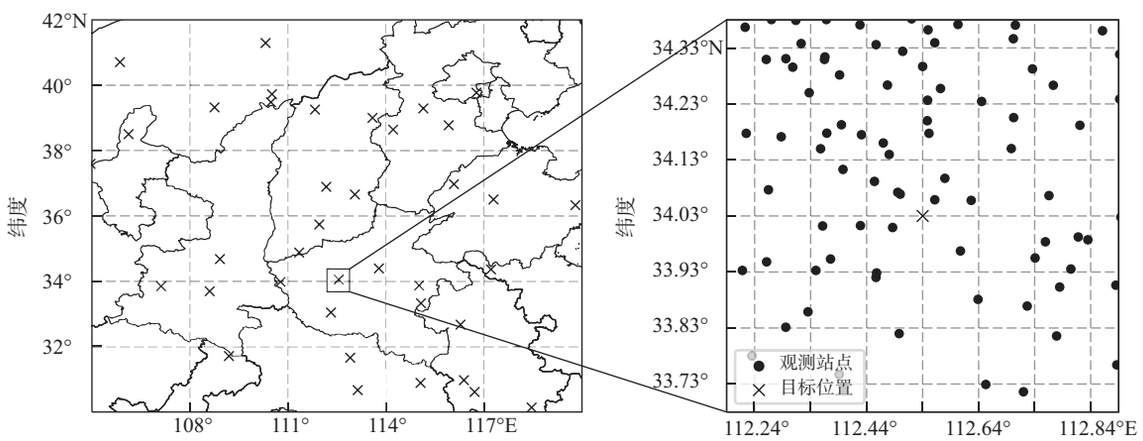
2.1 算法设计策略

2.1.1 研究区域确定

建立基于任意位置的实况分析模型,首先要确定研究区域。经过大量试验发现,以任意位置的目标点为中心,向四周外扩 0.35°形成 0.7°×0.7°的区域作为目标位置的研究区域效果最好,对降水的分析尤为明显。这主要因为:一是可以获取到区域内相关度高的站点,二是可以降低因范围太大产生的噪音影响。

2.1.2 算法设计思路

图 1a 为目标位置分布,图 1b 为以目标位置为中心的研究区域,圆点为研究区域中的站点。确定了任意目标位置所在的区域(图 1b)后,分别提取区域内站点实况数据和站点所在位置对应的 ART、CAR、RAD 和 SAT 等 4 类融合数据。站点实况数据为因变量,其余 4 类融合数据为自变量,构建机器学习模型,最后将目标点的 4 类融合数据带入到建立的模型中得到目标点的最终实况融合数据。



(a) 目标位置分布 (b) 以目标位置为中心的研究区域

图 1 研究区域的确定

Fig.1 Research region determination

在使用 GBDT 算法构建模型时,对于气温、降水、风速采用直接建模方法;而对于风向,构建了 U 和 V 分量两个模型,再将 U 和 V 分量合成,得出结果。

2.1.3 建模数据时次选取

基于不同时次数据为自变量构建的融合模型,经 10 次随机试验,计算其平均绝对误差(表 2)。由表可见,气温在 1~3 h 模型中的融合效果明显优于 6 h 以上;随着时间的延长,降水误差增大明显,2 h 模型融合效果较好;风速、风向也是 2 h 模型融合效果较好。因此,对于任意位置的实况分析模型,选取 2 h 数据建模,即当前时次和上一时次。

表 2 不同时次的多次随机试验平均绝对误差对比

Table 2 Comparison of mean absolute errors of multiple random experiments at different times

时次	平均绝对误差			
	气温 /℃	降水 /mm	风速 /(m·s^{-1})	风向 /(°)
1 h	0.201	0.682	1.251	71.734
2 h	0.197	0.528	1.150	49.988
3 h	0.191	0.859	1.793	51.343
6 h	0.257	1.336	1.564	77.758
24 h	0.234	1.730	1.645	80.561

2.1.4 模型选择

为探讨模型的区域适用性,设计了静态单一模型、静态分区多模型、动态单目标模型、动态全目标模型共 4 类进行试验。其中静态模型是固定模型,

为提前建模,后期将目标位置相关数据输入模型即可;而动态模型需每次重新建模。两类静态模型使用 8 月 1—12 日的全部时次数据建模,而两类动态模型仅使用当前时次和上一时次 2 h 数据建模。区域划分采用全国气象区域。分量级是在降水和风速建模时,采用自然断点法,按照目标位置的 CAR 产品要素划分级别后建模。表 3 为 4 类模型构建思路对比。

表 3　4 类模型对比

Table 3　Comparison of 4 types of models

模型	模型输出	分区	分量级	建模时次	核心思路
静态单一模型	是	是	是	全部时次	网格到点,用观测站点实况做目标,用其所在位置的格点数据做训练
静态分区多模型	是	是	是	全部时次	在静态单一模型基础上,增加分区域
动态单目标模型	否	否	否	2 h	周边点求中心点,以目标点为中心,确定研究区域,区域内站点数据做目标集,其对应位置的格点数据做训练集
动态全目标模型	否	否	否	2 h	在动态单目标模型的基础上,取消单目标单区域的建模,即使用所有研究区的数据做训练集

表 4　4 类模型平均绝对误差对比

Table 4　Comparison of mean absolute errors of 4 types of models

模型	平均绝对误差			
	气温/℃	降水/mm	风速/(m·s^{-1})	风向/(°)
静态单一模型	0.53	0.52	3.83	80.56
静态分区多模型	0.46	0.33	3.25	77.78
动态单目标模型	0.20	0.41	2.63	49.98
动态全目标模型	0.19	0.27	1.48	27.30

表 4 为 4 类模型试验的平均绝对误差结果,由表可见,4 类模型的融合效果从高到低依次是:动态全目标模型、动态单目标模型、静态分区多模型、静态单一模型,选择动态全目标模型作为实况融合算法(简记为“GBDT 模型”)。

2.2　算法技术路线

图 2 所示,建模到运行共 4 步:第一步,数据提取目标确定;第二步,提取数据;第三步,方法选取并构建模型;第四步,结果输出。

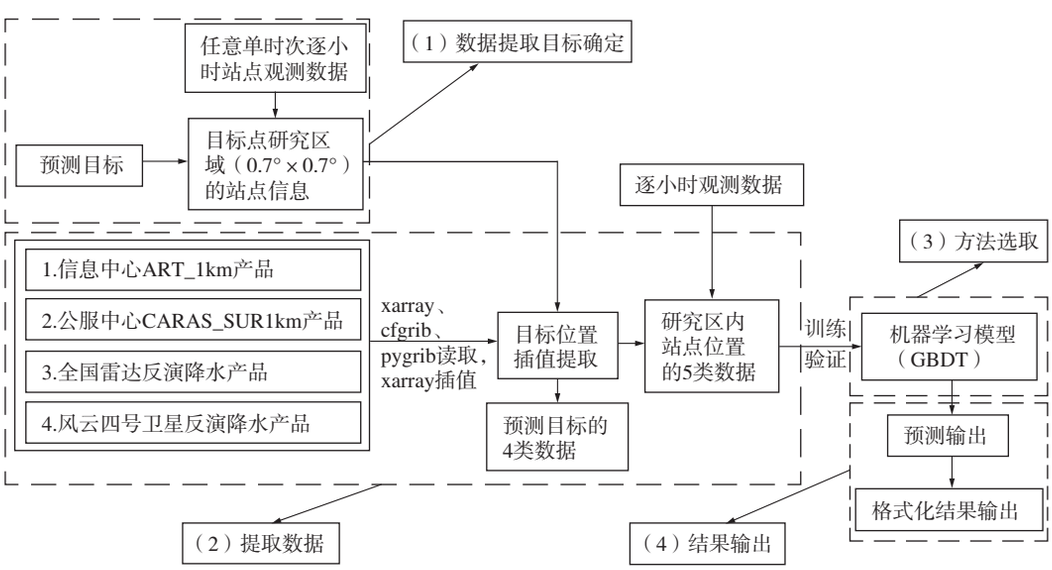


图 2　算法技术流程图

Fig.2　Technical flow chart of algorithm

3　结果分析

利用 ART、CAR、RAD、SAT 以及实况观测资料,构建 GBDT 模型后,将预测目标的 ART、CAR、RAD、SAT 等 4 类数据作为自变量代入模型,输出最终融合产品。为验证 GBDT 模型的融合效果以及在不同

区域的适用性，将 GBDT 模型输出的融合产品和 ART、CAR 分别与逐小时实况数据做全国分区域的误差分析。误差分析方法采用计算平均绝对误差、最大绝对误差和均方根误差。

为尽可能让试验具有可对比性，在分区（图 3）试验取样中（表 5），对于时次尽可能选取该区域内出现降水且量级变化大、空间分布不均的时段。对于目标数，东北、华北和西北 3 个区域内各省随机目标取 10 个，突出纬向分布检验；而华中和华南区域内各省目标取 5 个，突出经向分布检验；华东和西南各省目标取 3 个，突出东部与西部差异的检验；新疆、西藏和内蒙古都是单省分区，新疆和西藏取 40 个、内蒙古取 20 个目标，相互可形成对比检验。

图 3 全国气象分区图

Fig.3 Nationwide meteorological zoning map

表 5 分区试验取样说明

Table 5 Sampling instructions for zone test

区域	时刻	单省目标数	目标总数	样本总数
东北	01T09	10	30	980
华北	12T16	10	40	2 400
西北	11T16	10	40	1 930
华中	08T14	5	15	700
华南	12T16	5	15	739
华东	10T10	3	18	1 610
西南	07T19	3	12	998
内蒙古	12T14	20	20	337
新疆	10T14	40	40	636
西藏	12T16	40	40	268

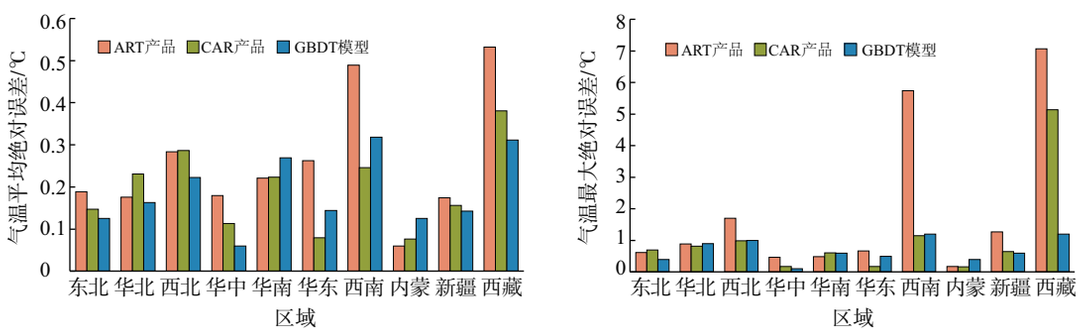
注：“时刻”指日及该日的时刻，如“01T09”指 1 日 09 时。

3.1 不同区域气温的检验

由 GBDT 融合产品与 ART 产品、CAR 产品的气温平均绝对误差对比图可见（图 4a），东北、华北、西北、华中、新疆、西藏 6 个区域 GBDT 融合产品误差为0.06～0.31 ℃，ART 产品、CAR 产品误差范围分别是0.17～0.53 ℃、0.11～0.38 ℃，GBDT 融合产品效果优于 ART 产品与 CAR 产品。华南、华东、西南、内蒙古 4 个区域 GBDT 融合产品误差为0.12～0.31 ℃，ART 产品、CAR 产品误差范围分别为0.06～0.49 ℃、0.07～0.24 ℃，GBDT 融合产品表现略逊，其中华东和西南 GBDT 融合产品优于 ART 产品而逊于 CAR

产品,华南和内蒙古 GBDT 融合产品误差增加,但幅度小于0.06 ℃。由最大绝对误差来看(图 4b),GBDT 融合产品在西藏改善幅度最大,较 ART 产品、CAR 产品误差降幅分别达5.87 ℃和3.94 ℃;东北、华中、新疆 GBDT 融合产品误差为0.10~0.60 ℃,ART 产品、CAR 产品误差分别为0.47~1.26 ℃、0.17~0.69 ℃,误差均有小幅减小,表现略优;华北、华南三者误差差别极小;西北 GBDT 融合产品与 CAR 产品相近,优于 ART 产品;华东、西南 GBDT 融合产品优于 ART 产品而逊于 CAR 产品;内蒙古 GBDT 融合产品误差约增加0.14 ℃。由气温均方根误差来看(图 5),GBDT 融合产品在西藏提升幅度最大,均方根误差小于0.20 ℃,表明误差分布较为集中;东北、华北、西北、华中、新疆误差为0.01~0.09 ℃,ART 产品、CAR 产品误差分别为0.02~0.20 ℃、0.01~0.15 ℃,GBDT 融合产品优于二者;华东、西南 GBDT 融合产品优于 ART 产品而逊于 CAR 产品;华南、内蒙古 GBDT 融合产品误差较二者略有增大,这与平均绝对误差在各区域的表现是一致的。

总体上,气温检验中 GBDT 融合产品在西藏效果最好,尽管空间分布上来看误差仍为最大,但从最大绝对误差与均方根误差来看,该区域误差整体减小幅度明显,且误差相对集中,较 ART 产品、CAR 产品均有较大改进。这与该地区站点布网偏少有关,ART 产品、CAR 产品在当地质量不高,使 GBDT 模型优势得以凸显。在其他区域,最大绝对误差接近或明显小于1.00 ℃,平均绝对误差较 ART 产品、CAR 产品在 60%的区域均有减小。



(a) 平均绝对误差　　(b) 最大绝对误差

图 4　不同区域气温平均绝对误差和最大绝对误差分布

Fig.4　Mean absolute error and maximum absolute error of air temperature in different zones

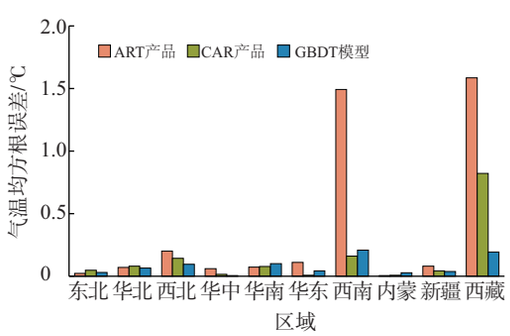


图 5　气温均方根误差

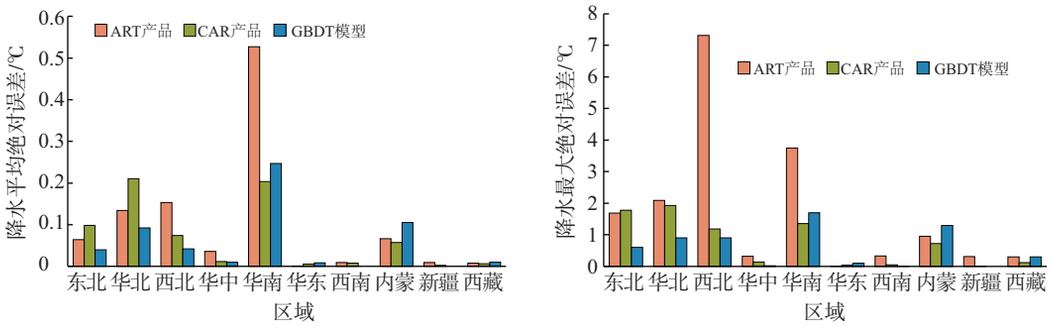
Fig.5　Root mean square error of air temperature

3.2　不同区域降水的检验

2020年 8 月 1—4 日、5—7 日、8—10 日、11—15 日先后受东北冷涡、副热带高压、台风、西南涡、蒙古气旋、热带季风槽等系统影响,全国各地均有降水,除新疆、西藏以及西北北部降水较少外,其他各地均有短时强降水出现,其中西北东部、华北、东北频次最多。因此,利用 8 月 1—15 日的多源资料进行降水融合试验是有意义的。

由 GBDT 融合产品与 ART 产品、CAR 产品的降水平均绝对误差(图 6a)、最大绝对误差(图 6b)和均方根误差(图 7)对比来看,三者差异在各区域表现一致。东北、华北和西北 GBDT 融合产品改进较明显,其平均绝对误差为0.04~0.09 mm,最大绝对误差为0.60~0.90 mm,均方根误差为0.02~0.04 mm,而 ART 产品、CAR 产品的平均绝对误差分别为0.06~0.15 mm、0.07~0.21 mm,最大绝对误差分别为1.68~7.31 mm、1.19~1.92 mm,均方根误差分别为0.09~0.77 mm、0.05~0.19 mm。华中、华东、西南、新疆、西藏三者差异极小,GBDT 融合产品与 CAR 产品表现相当;华南 GBDT 融合产品优于 ART 产品而逊于 CAR 产品;在内蒙古 GBDT 融合产品平均绝对误差为0.11 mm、最大绝对误差为1.30 mm、均

方根误差为0.09 mm,3 类误差较 ART 产品、CAR 产品均略有增大。



(a) 平均绝对误差 (b) 最大绝对误差

图 6 不同区域降水平均绝对误差和最大绝对误差分布

Fig.6 Mean absolute error and maximum absolute error of precipitation in different zones

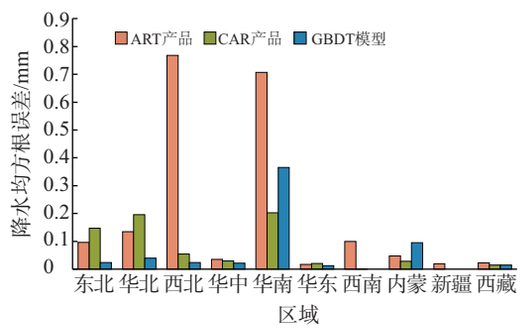


图 7 降水均方根误差

Fig.7 Root mean square error of precipitation

可见,在降水检验中,除华南和内蒙古外,GBDT 融合产品都取得了较好效果。由平均绝对误差来看,GBDT 融合产品较 ART 产品、CAR 产品在东北、华北和西北误差减小明显,在华中、华东、西南、新疆、西藏 5 个区域三者基本相当,在华南和内蒙古效果略差。对内蒙古而言,可能与其东西跨度太大,试验中有限数量的取样,得到的样本特征不一致,导致建模效果不理想。

3.3 不同区域风速和风向的检验

由 GBDT 融合产品与 ART 产品、CAR 产品的风速平均绝对误差和最大绝对误差对比(图 8)来看,GBDT 融合产品在各分区融合效果都很好,较 ART 产品、CAR 产品均有明显改进。GBDT 融合产品与真值相比,平均绝对误差小于或接近 1.0 $m\cdot s^{-1}$,最大绝对误差介于 1.5~4.5 $m\cdot s^{-1}$,均方根误差(图略)小于 1.5 $m\cdot s^{-1}$。

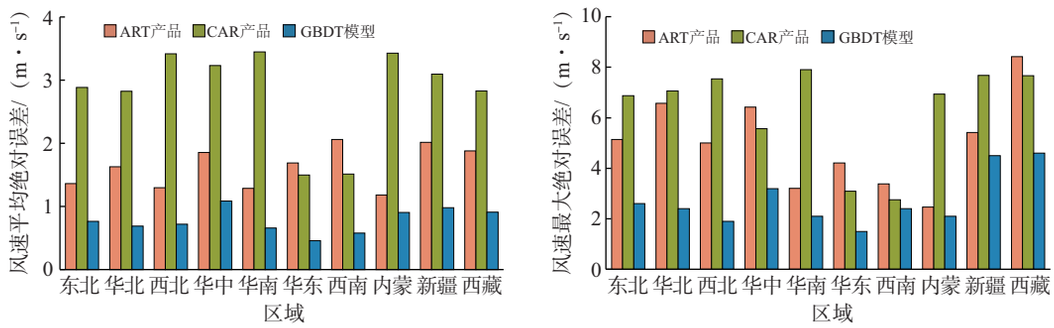
由 GBDT 融合产品与 ART 产品、CAR 产品的风向对比(图 9)来看,平均绝对误差在各区域都有减小,表明风向整体较 ART 产品、CAR 产品均有改进。在最大绝对误差中,GBDT 融合产品在东北、西北、华中、华南、华东、内蒙古、新疆、西藏 8 个区域较 ART 产品、CAR 产品有改进,而在华北与西南两个区域没有明显改善,其中华北区域较 ART 产品、CAR 产品误差均略有增大,西南区域优于 ART 产品而逊于 CAR 产品。深入分析最大绝对误差在西南区域中明显偏离 CAR 产品的样本结果,发现这些样本风速分布在 2.3~2.5 $m\cdot s^{-1}$,而样本总体的风速分布在 0~15.0 $m\cdot s^{-1}$,负订正样本占比为 0.32%,并且风向出现的负订正并非由强风或静风引起。同时也发现,GBDT 融合产品风向负订正超 CAR 产品 50°以上的样本数有且仅有 1 站次,而绝对误差中次大值为 30°,与 CAR 产品相近。

综上,GBDT 模型在风速和风向的分析中总体较 ART 产品、CAR 产品有较大的改进,尤其是在对于风速的分析中平均绝对误差较 ART 产品、CAR 产品分别减小 23%~73%、61%~80%,对于风向的分析中平均绝对误差分别减小 5%~37%、28%~63%。

4 结论

本文从气象业务综合应用多源实况分析产品制作任意位置实况数据的需求出发,应用逐时 ART、CAR、RAD、SAT 格点实况分析产品和观测数据,基于 GBDT 机器学习方法构建了动态融合应用模型(GBDT 模型),对 15 d 的 GBDT 气温、降水、风向、风速融合产品进行了全国分区检验,得到如下结论:

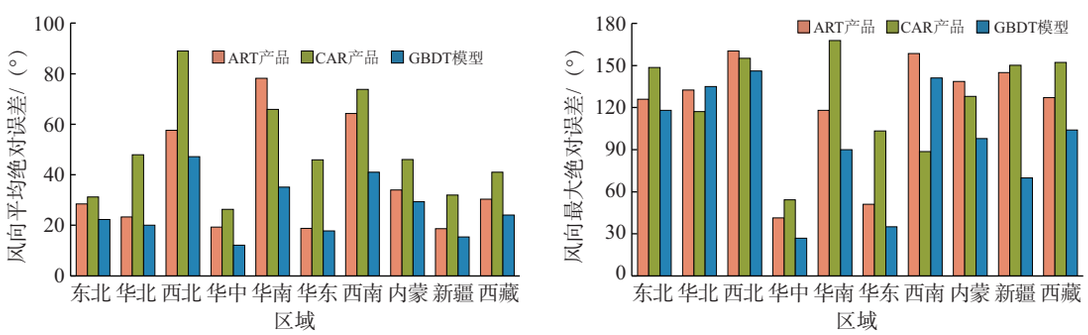
(1)GBDT 气温融合产品在东北、华北、西北、华中、新疆、西藏 6 个区域较 ART 产品、CAR 产品均有改进,在华东和西南 GBDT 融合产品优于 ART 产品



（a）平均绝对误差　　（b）最大绝对误差

图 8　风速平均绝对误差和最大绝对误差分布

Fig.8　Mean absolute error and maximum absolute error of wind speed



（a）平均绝对误差　　（b）最大绝对误差

图 9　不同区域风向平均绝对误差和最大绝对误差分布

Fig.9　Mean absolute error and maximum absolute error of wind direction in different zones

而逊于 CAR 产品,在华南和内蒙古 GBDT 融合产品误差较 ART 产品、CAR 产品略有增大,幅度小于 0.06℃。考虑到气温在实际应用中,只保留一位小数,其融合意义仅在西藏区域较大。

(2)GBDT 降水融合产品在东北、华北和西北 3 个区域平均绝对误差较 ART 产品、CAR 产品改进明显,在华中、华东、西南、新疆、西藏 5 个区域三者基本相当,在华南 GBDT 融合产品优于 ART 产品逊于 CAR 产品;在样本偏少的内蒙古较 ART 产品、CAR 产品误差略有增大。

(3)GBDT 风速、风向融合产品较 ART 产品、CAR 产品均有较大改进。风速融合产品平均绝对误差较 ART 产品、CAR 产品分别减小 23%~73%、61%~80%,风向融合产品平均绝对误差分别减小 5%~37%、28%~63%。

初步试验结果表明基于机器学习方法的动态全目标模型(GBDT 模型)可应用于融合多源实况分析产品和观测数据开展选定位置气温、降水、风向、风速要素的实况气象信息服务,但有待利用更长时间序列资料进行检验并不断完善模型。

参考文献:

[1] HUFFMAN G J, ADLER R F, ARKIN P, et al. TheGlobal Precipitation Climatology Project (GPCP) combined precipitation dataset[J]. Bull Amer Meteor Soc,1997,78(1):5-20.

[2] XIE PP, ARKIN P A. Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs[J]. Bull Amer Meteor Soc, 1997,78(11):2539-2558.

[3] HUFFMAN G J,ADLER R F, BOLVIN D T, et al. The TRMM Multisatellite Precipitation Analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales[J]. J Hydrometeorol,2007,8(1):38-55.

[4] JOYCE R J, JANOWIAK J E, ARKIN P A, et al. CMORPH: a method that produces global precipitation estimates from passive microwave and infrared data at

high spatial and temporalresolution [J]. J Hydrometeorol, 2004,5(3):487-503.

[5] USHIO T, SASASHIGE K, KUBOTA T, et al. AKalman filter approach to the global satellite mapping of precipitation (GSMaP) from combined passive microwave and infrared radiometric data[J]. J Meteor Soc Japan,2009,87A: 137-151.

[6] 毕宝贵,代刊,王毅,等.定量降水预报技术进展[J].应用气象学报,2016,27(5):534-549.

[7] SEO D J, BREIDENBACH J P. Real-time correction of spatiallynonuniform bias in radar rainfall data using rain gauge measurements[J]. J Hydrometeorol,2002,3(2): 93-111.

[8] 潘旸,沈艳,宇婧婧,等.基于最优插值方法分析的中国区域地面观测与卫星反演逐时降水融合试验[J].气象学报,2012,70(6):1381-1389.

[9] RASMY M, KOIKE T, BOUSSETTA S, et al. Development of a satellite land data assimilation system coupled with a mesoscale model in the Tibetan Plateau [J]. IEEE TransGeosci Remote Sens, 2011, 49(8): 2847-2862.

[10] XIA Y L, MITCHELL K, EK M, et al. Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System Project phase 2 (NLDAS-2): 2. Validation of model-simulatedstreamflow [J]. J Geophys Res, 2012, 117 (D3):D03109.

[11] ALBERGEL C, DORIGO W, BALSAMO G, et al. Monitoring multi-decadal satellite earth observation of soil moisture products through land surfacereanalyses [J]. Remote Sens Environ,2013,138:77-89.

[12] 陈冠宇,艾未华,程玉鑫,等.基于星载SAR数据和模式资料的海面风场变分融合方法研究[J].海洋气象学报,2017,37(4):65-74.

[13] 周强,陈洁,李玉华.基于Himawari-8卫星的自适应阈值火点判识算法适用性分析[J].海洋气象学报,2020,40(1):127-133.

[14] 师春香,潘旸,谷军霞,等.多源气象数据融合格点实况产品研制进展[J].气象学报,2019,77(4):774-783.

[15] 张璐,潘旸,谷军霞,等.国际主流多源融合降水实况产品的研究进展与展望[J].气象科技进展,2022,12(6):16-27.

[16] 崔园园,张强,李威,等.CLDAS融合土壤相对湿度产品适用性评估及在气象干旱监测中的应用[J].海洋气象学报,2020,40(4):105-113.

[17] 刘维成,徐丽丽,朱姜韬,等.再分析资料和陆面数据同化资料土壤湿度产品在中国北方地区的适用性评估[J].大气科学学报,2022,45(4):616-629.

[18] 刘莹,师春香,王海军,等.CLDAS气温数据在中国区域的适用性评估[J].大气科学学报,2021,44(4):540-548.

[19] 殷笑茹,焦圣明,喜度,等.基于多源数据的地面降水质量控制方法研究[J].气象科学,2022,42(4):539-548.

[20] 李奇临,旷兰,魏麟骁,等.不同分辨率的气温格点实况分析产品在重庆的对比检验[J].气象科技进展,2022,12(6):91-96.

[21] 潘旸,谷军霞,师春香,等.中国北方冬季降水的多源资料产品评估和融合优化[J].气象学报,2022,80(6):953-966.

[22] 董春卿,郭媛媛,张磊,等.基于CLDAS的格点温度预报偏差订正方法[J].干旱气象,2021,39(5):847-856.

[23] 孙全德,焦瑞莉,夏江江,等.基于机器学习的数值天气预报风速订正研究[J].气象,2019,45(3):426-436.

[24] 李昕蓓,张苏平,衣立,等.基于循环神经网络的单站能见度短临预报试验[J].海洋气象学报,2019,39(2):76-83.

[25] 王萌,刘合香,卢耀健,等.基于模糊时间序列的华南台风登陆时最大风速极值预测模型[J].海洋气象学报,2019,39(4):68-74.

[26] 任萍,陈明轩,曹伟华,等.基于机器学习的复杂地形下短期数值天气预报误差分析与订正[J].气象学报,2020,78(6):1002-1020.