基于机器学习的河南省短波辐射数值预报订正方法研究[1]

程凯琪[1,2]　魏璐[1,2]　李伊吟[1,2]　孙睿藻[1,2]　张凡[1,3]

1.中国气象局·河南省农业气象保障与应用技术重点实验室，郑州 450003

2.河南省气象服务中心，郑州 450003

3.河南省人工影响天气中心，郑州 450003

提要：使用 2022 年河南省 23 个辐射观测站总辐照度数据和 CMA-WSP2.0 模式产品，通过 LASSO 回归选取特征变量，建立训练数据集和测试数据集，利用训练数据集采用机器学习方法（随机森林、XGBoost、LightGBM）进行模型训练，订正河南省 CMA-WSP2.0 模式预报总辐照度，并对订正结果分站点和区域、分季节、总辐照度分级检验，结论如下：随机森林、XGBoost、LightGBM 三种机器学习方法订正效果良好，相较于 CMA-WSP2.0 模式预报结果，平均绝对误差和均方根误差显著降低，24 h 的准确率和合格率显著提升。其中 LightGBM 订正效果最优，平均绝对误差相较于 CMA-WSP2.0 模式预报减小了 18.32~32.91 W·m^{-2}，平均绝对误差减小比例在 38%~56%，均方根误差减小比例在 36%~52%；24 h 的平均准确率和平均合格率较 CMA-WSP2.0 模式预报分别提升了 7.3%、5.7%。区域统计与站点统计结果较为一致，对于 5 个区域而言，豫西区域订正效果最好。三种机器学习方法订正后的偏差范围相比 CMA-WSP2.0 模式预报集中范围更窄，偏差分布在零值附近的概率更大。在各季节检验结果中三种方法对于冬季订正效果更为显著。对于不同的总辐照度等级，三种机器学习方法均有效改善了 CMA-WSP2.0 模式预报，随着总辐照度等级的增加，订正效果总体呈逐渐减弱的趋势。研究结果可为提高河南省总辐照度预报能力提供有益参考。

关键词:CMA-WSP2.0模式，机器学习，总辐照度，分区建模

中图分类号：P456　　文献标志码:A

Short Wave Radiation Forecast Correction Based on Machine Learning in Henan Province

CHENG Kaiqi[1,2]　WEI Lu[1,2]　LI Yiyin[1,2]　SUN Ruizao[1,2]　ZHANG Fan[1,3]

1 CMA Henan Key Laboratory of Agrometeorological Support and Applied Technique, Zhengzhou 450003

2 Henan Meteorological Service Center, Zhengzhou 450003,

3 Weather Modification Center of Henan Province, Zhengzhou 450003

Abstract: Using the total irradiance data from 23 radiation observation stations in Henan Province in 2022 and CMA-WSP2.0 model products,,characteristic variables were selected by Lasso regression, training data sets and

[1] 中国气象局 河南省农业气象保障与应用技术重点实验室应用技术研究基金（KQ202320）、河南省科技研发计划联合基金(222103810093、232103810092)共同资助

第一作者：程凯琪，主要从事数值预报、新能源发电技术研究.E-mail:qchengkai@163.com

通讯作者：魏璐，主要从事电网防灾减灾，新能源发电技术研究.E-mail:2422591103@qq.com

test data sets were established, and machine learning methods (Random forest, XGBoost, LightGBM) were used to train the model using the training data set, and the total irradiance forecast by CMA-WSP2.0 model in Henan Province was revised. The revised results were tested by site and region, season and total irradiance classification, and the following conclusions were obtained: The three machine learning methods of random forest, XGBoost, and LightGBM have good correction effects. Compared with the CMA-WSP2.0 model prediction results, the average absolute error and root mean square error are significantly reduced, and the 24-hour accuracy and 24-hour qualification rate are significantly improved. The average absolute error decreases by 18.32~32.91 $W \cdot m^{-2}$, the average error decreases by 38~56%, and the root mean square error decreases by 36~52%. The 24-hour average accuracy and 24-hour average qualification rate increased by 7.3% and 5.7%. The results of regional statistics are consistent with those of the stations. For the five regions, the correction effect of western Henan is the best. The corrected deviation range of the three machine learning methods is narrower than that of the CMA-WSP2.0 simulation set, and the probability of the deviation distribution near 0 is greater. Among the seasonal test results, the three methods have more significant correction effect in winter. For different total irradiance levels, the three machine learning methods can effectively improve the CMA-WSP2.0 model prediction, and the correction effect tends to gradually weaken with the increase of total irradiance levels. The results can provide useful reference for improving the ability of total irradiance forecast in Henan Province.
Key words: CMA-WSP2.0 model, machine learning, total irradiance, partition modeling

引言

太阳能因其具有清洁、无污染、分布广泛等优势，在“碳达峰、碳中和”的目标驱动下，得到了更为广泛的关注和应用。在太阳能的众多利用方式中，光伏发电因其转换效率高、使用期长，装机容量迅速增长，截至 2023 年 11 月底，河南省新能源装机容量达 6021 万 kW，首次突破 6000 万 kW，新能源装机占全省电源总装机的 44%，稳居河南省第二大电源，其中分布式光伏装机容量增速为全国第一。随着光伏并网的快速增长，光伏功率预报已成为制约光伏发电并网消纳的重要瓶颈，大规模的光伏并网会对电网的稳定性造成冲击，增加电网计划和调度的难度，影响电网稳定运行（Delannoy et al, 2021）。

地表太阳总辐射作为影响光伏功率预报的关键因子，其在时间变化上具有不连续、不确定性，如何提高太阳短波辐射精细化预报的准确率是亟待解决的问题（Haupt and Kosovic, 2017；李遥等，2020）。许多学者利用 MOS 订正方法对模式预报总辐照度进行订正（白永清等，2013；顾婷婷等，2022；孙朋杰等，2015），研究表明该方法对太阳辐射的预报精度有一定程度的改善。随着计算机性能的提高以及海量数据的收集，越来越多的研究者开始使用机器学习方法对预报模拟偏差进行订正，研究表明机器学习方法较传统统计学方法订正效果更为明显（普智勇等，2023；Belmahdi and el Bouardi, 2024）。王雪洁等（2022）基于随机森林算法对 ERA5 总辐射产品进行了订正，结果表明经过随机森林订正后精度有明显的提高，随机森林模拟精度高，有较高的稳定性。陈昱文等（2020）利用 4 个气象站点数据，挖掘观测数据的时序特征并结合气温预报训练机器学习模型，发现集成学习算法在数值模式预报结果订正中具有较大的应用潜力。机器学习在气温数值预报（陈有龙等，2020；李韬等，2022；方鸿斌等，2024；智协飞等，2020）、空气质量（芦华等，2020）以及海洋环境预报（许立兵等，2020）、风速（孙全德等，2019；徐景峰等，2023）、

强对流监测预报（周康辉等，2021；李文娟等，2024）、灾害等级评估（刘淑贤等，2024；Zhang et al，2019）等方面都有应用且效果较好。

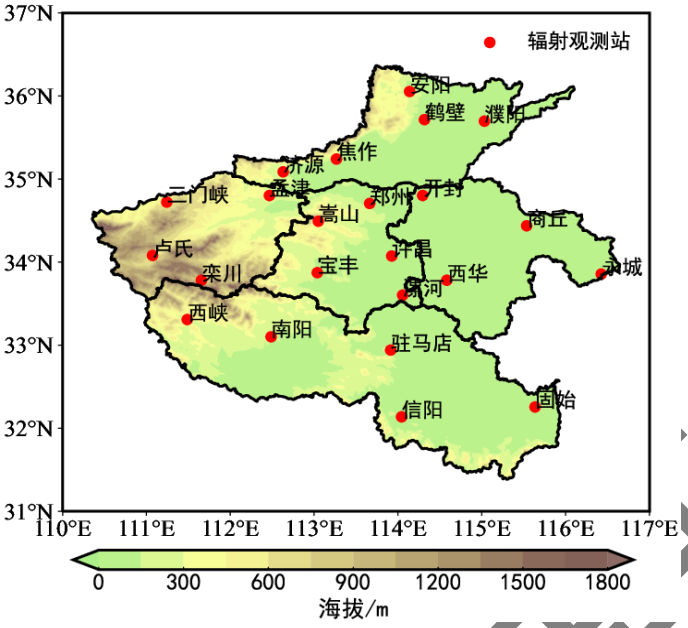
中国气象局风能太阳能预报系统（CMA-WSP2.0）预报产品，其时间分辨率为 15 min，空间分辨率为 9 km，包含总辐照度、温度、湿度、气压等与光伏功率预报相关的要素（万超等，2023）。但由于模式初始场、动力和微物理参数化方案的不完善等多种原因，导致模式预报结果偏差较大，目前该预报产品在河南省的预报准确性和适用性有待检验和提高。因此为提高 CMA-WSP2.0 预报总辐照度产品质量，提升该产品在河南省适用性，本研究采用多种机器学习算法（随机森林、XGBoost、LightGBM），结合地理特征、日变化特征、模式偏差及融合多种影响总辐照度的物理因子，对 CMA-WSP2.0 预报总辐照度进行订正。将河南省划分为豫东、豫西、豫南、豫北、豫中 5 个区域（张凡和程凯琪，2024），基于 LASSO 回归进行相关特征要素的选择，进行机器学习（随机森林、XGBoost、LightGBM）的训练建模，对 CMA-WSP2.0 预报总辐照度进行订正，并通过准确率、合格率、均方根误差、平均绝对误差等相关统计量比较不同订正方法的修正效果。

1 数据和方法

1.1 数据

本文使用的实测总辐照度数据为河南省 23 个辐射观测站（图 1）逐时观测总辐照度，数据来源于河南省 “天擎”平台，其中郑州、固始、南阳提供总辐照度观测较早，分别于 1957 年、1960 年、1990 年开始进行总辐照度观测，其他辐射观测站于 2018 年提供总辐照度观测，各气象站的观测运维及数据质量控制由各地市级气象局和河南省气象探测数据中心负责。观测设备使用 TBQ-2-B 型号总辐射表，该设备能够捕捉到波长介于 0.3～3 μm 的太阳总辐射。对于地面总辐照度的观测，一般通过每分钟至少进行 6 次的总辐照度数据采集，并取这 1 分钟内的数据平均值，以此作为该分钟的总辐照度值。

河南省地形复杂，地势西高东低，北、西、南三面被太行山、伏牛山、桐柏山、大别山环绕，中东部为平原，为更好探索不同订正方法效果，将河南省划分为豫东、豫西、豫南、豫北、豫中 5 个区域，23 个辐射观测站也进行了划分（图 1，表 1）。



注：填色为地形高度。

图 1 河南省辐射观测站分布及区域划分

Fig.1 Distribution and regional division of radiation observation stations in Henan Province

表 1 河南省辐射观测站区域划分

Table 1 Regional division of radiation observation stations in Henan Province

区域	辐射观测站
豫东	开封、商丘、永城、西华
豫西	三门峡、孟津、卢氏、栾川
豫南	西峡、南阳、驻马店、信阳、固始
豫北	安阳、鹤壁、濮阳、焦作、济源
豫中	嵩山、郑州、宝丰、许昌、漯河

模式数据为 CMA-WSP2.0 预报产品，模式每日 20 时起报，时间分辨率为 15 min，空间分辨率为 9 km，模式数据来源于中国气象局业务下发，时间范围为 2022 年 1 月 1 日至 12 月 31 日。

提取 2022 年 1 月 1 日至 12 月 31 日 CMA-WSP2.0 模式每日 20 时起报的 0~84 h 的逐时预报要素场，使用临近插值法把模式预报场插值到 23 个辐射观测站位置，构建总辐照度订正模型。初步选取要素包括：时间、总辐照度、法向短波辐射、地表向下直接辐射、晴空地表向下直接辐射、地表散射辐射、地表向下长波辐射、10 m 纬向风、10 m 经向风、70 m 纬向风、70 m 经向风、80 m 纬向风、80 m 经向风、100 m 纬向风、100 m 经向风、120 m 纬向风、120 m 经向风、2 m 温度、2 m 比湿、地面温度、地面气压、雪水当量、积云对流累积降水、网格尺度累积降水、地表反照率、边界层高度，共 26 个预报要素场，使用 LASSO 回归对 26 个预报要素场做特征变量筛选，并进行机器学习算法模型的训练，订正模式预报总辐照度。

1.2 方法

1.2.1 LASSO 回归

Tibshiran（2011）提出的 LASSO 回归，是一种用于线性回归的正则化方法，通过在损失函数中添加 L1 正则项来促使模型参数稀疏化，得到一个更为精炼简单的模型，将一部分系数进行压缩，一些系数设定为 0，因此保留子集收缩的优点。

在本文总辐照度订正中，给定有 m 个自变量的特征向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$，其中$x_m$为$x$在第$i$个特征上的取值，通过 m 个特征量的线性组合来预测总辐照度，公式如下：

$$f(x)=w_1x_1+w_2x_2+\cdots+w_mx_m \tag{1}$$

式中$\boldsymbol{w}=(w_1,w_2,\cdots,w_m)$是 m 个特征向量的权重。

损失函数的定义为

$$loss(\boldsymbol{w})=\|f(\boldsymbol{x})-y\|^2+\alpha\|\boldsymbol{w}\| \tag{2}$$

式中：y表示总辐照度实测值，$\alpha\|\boldsymbol{w}\|$是正则化项，不仅有助于降低过拟合风险，还具有特征选择的作用。通过对$loss(\boldsymbol{w})$求最小值，求得$\boldsymbol{w}$，从而 LASSO 回归模型得以确定。

1.2.2 随机森林

随机森林是一种经典的机器学习算法，最早由 Breiman（2001）提出。它是基于集成学习的一种方法，通过组合多个决策树进行预测，在回归问题中取其平均值。随机森林算法在数据挖掘各个领域具有较广的应用性，对于异常值和噪声具有较好的容忍度，不容易出现过拟合。

随机森林一般采用以下步骤进行训练：第一步，数据准备，对原始数据集进行构建；第二步，随机抽样，构建多个决策树进行预测，每个决策树的训练样本通过随机抽样得到，进行重复训练后得到多个决策树，组成随机森林；第三步，随机森林的预测，采用平均的方式，最终结果为每个决策树的预测值取平均。

1.2.3 XGBoost

XGBoost 是一种基于梯度提升决策树的机器学习算法，它具有快速训练时间、自动处理数据不平衡、自动选择最佳特征等优势。XGBoost 算法通过不同的目标函数、正则化以及损失函数来训练模型，它以树模型为基础，可以使模型自动学习特征的权重，并且具有较高的准确率。

1.2.4 LightGBM

LightGBM 是微软于 2016 年开源的一种将决策树作为机器学习的梯度提升机器学习的框架，是对梯度提升算法的高效实现，原理上和 XGBoost 类似，都采用损失函数的负梯度作为当前决策树的残差近似值，去拟合新的决策树。

1.2.5 检验方法

检验指标主要包括平均绝对误差（MAE）、均方根误差(RMSE)、相关系数（R）、准确率(A)、合格

率(*Q*)，MAE 和 RMSE 主要反映真实值与预测值之间的偏离情况，反映模式的预测能力和准确性。*A* 则是反映一段时间内连续的太阳辐射预报值与实测值之间接近程度的指标，*Q* 是反映一段时间内到达基本评判要求的太阳辐射预报占比的指标（张敏等，2024）。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|I_f^i-I_o^i\right| \tag{3}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_f^i-I_o^i)^2} \tag{4}$$

$$R=\frac{\sum_{i=1}^{n}(I_o^i-\overline{I_o})(I_f^i-\overline{I_f})}{\sqrt{\sum_{i=1}^{n}(I_o^i-\overline{I_o})^2}\sqrt{\sum_{i=1}^{n}(I_f^i-\overline{I_f})^2}} \tag{5}$$

$$A=\left(1-\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{I_f^i-I_o^i}{I})^2}\right)\times 100\% \tag{6}$$

$$Q=\frac{1}{n}\sum_{i=1}^{n}Q_i\times 100\% \tag{7}$$

$$Q_i=\begin{cases}1 & \frac{|I_f^i-I_o^i|}{I}<0.3\\ 0 & \frac{|I_f^i-I_o^i|}{I}\geq 0.3\end{cases} \tag{8}$$

式中：*n* 为样本总量，I_f^i为当日第 i 时刻的预报总辐照度，I_o^i为当日第 i 时刻的实测总辐照度，$\overline{I_f}$为当日预测平均总辐照度，Q_i为 i 时刻的预测合格率判定结果。计算 *A* 和 *Q* 时，针对不同情况对 *I* 取不同值，当日实测平均总辐照度（$\overline{I_o}$）＞250 W·m^{-2}时，*I* 取值为当日实测总辐照度的最大值；当日$\overline{I_o}$≤250 W·m^{-2}时，*I* 取值为 600。

2 订正模型构建

在订正模型构建之前，首先对数据进行预处理，剔除实测总辐照度及模式预报要素中的错误值及缺测值，将实测总辐照度与模式预报结果进行匹配，同时为了避免训练时各要素值数值小而贡献小的问题，对各要素进行标准化处理。将处理后的各个站点数据分成 2 个部分（原始训练集和原始测试集），考虑模型训练更适用于样本量较大的数据集，同时避免因客观条件对训练与测试产生影响，将数据集打乱后随机选取 80%的数据作为原始训练集，其中豫东、豫西、豫南、豫北、豫中样本量分别为 123862、154939、123959、154976、154652 条，剩下 20%数据作为原始测试集，豫东、豫西、豫南、豫北、豫中样本量分别为 24280、24280、31025、31025、31025 条。

然后基于 LASSO 回归算法进行特征量的选择，每个区域输入的自变量特征共 26 个，利用原始训练集数据，通过 LASSO 回归训练，得到豫东、豫西、豫南、豫北、豫中 26 个特征要素与总辐照度的权重，根据权重的绝对值从高到低进行排列，对 5 个区域进行训练，得到河南省 5 个区域原始测试集中 RMSE 随特

征量维度变化的曲线（图 2）。可以看出，当特征量维度到达一定数量时，RMSE 不再显著降低，达到一个稳定的水平。通过对河南省 5 个区域特征量进行分析，最后选取时间、总辐照度、法向短波辐射、地表向下直接辐射、晴空地表向下直接辐射、地表向下长波辐射、10 m 纬向风、10 m 经向风、2 m 温度、2 m 比湿、地面气压、网格尺度累积降水、地表反照率、边界层高度共 14 个特征量进行随机森林、XGBoost、LightGBM 的训练，通过特征量的选择可以降低计算和所用的存储，减少模型训练时间。

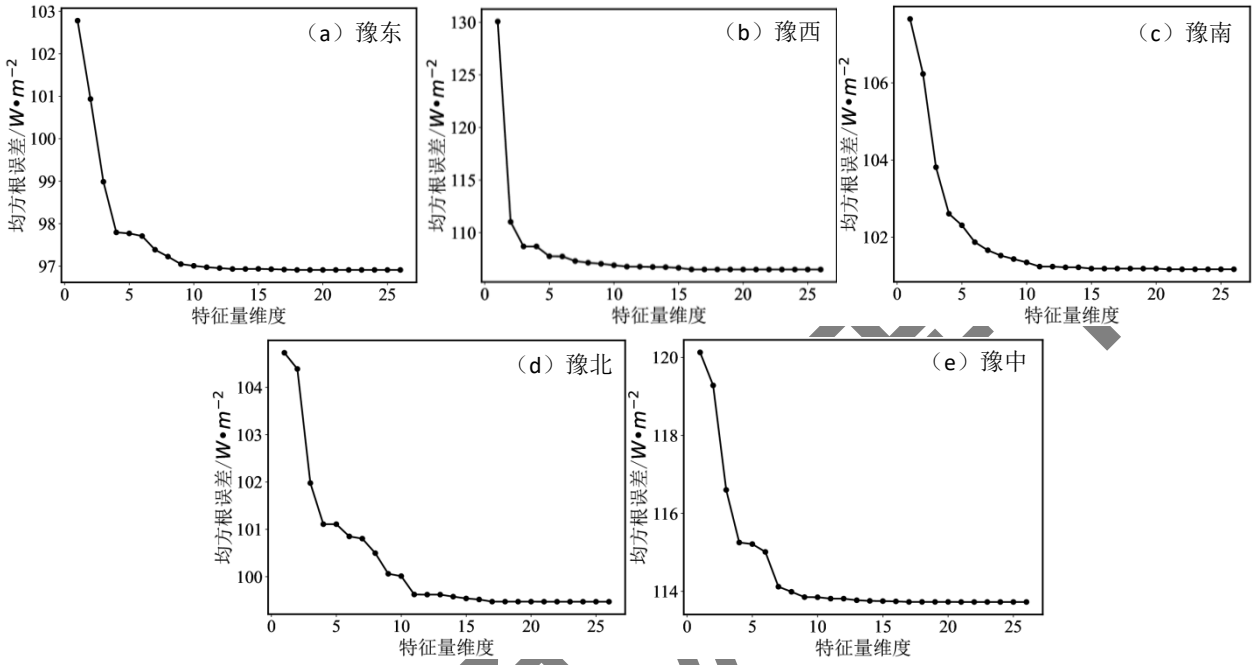


图 2 2022 年 1 月 1 日至 12 月 31 日原始测试集中河南省 5 个分区基于 LASSO 回归预测总辐照度 RMSE 随特征量维度的变化

Fig.2 RMSE of total irradiance prediction with feature dimension based on Lasso regression in five regions of Henan Province in the original test set from January to December 2022

根据 LASSO 回归筛选出的 14 个特征量，分别从原始训练集和原始测试集中提取对应的特征列，重新构建训练集和测试集。将训练集进行随机森林、XGBoost 和 LightGBM 三种机器学习的模型训练，并在硬件环境和时间允许的范围内对三种机器学习算法中关键参数通过网格搜索寻找最优参数，得到每种机器学习方法的最优参数模型。使用最优参数模型对测试集进行订正，并对订正结果进行检验分析。

3 订正结果检验

3.1 分站点和区域评估

统计三种机器学习模型订正后及 CMA-WSP2.0 模式预报的河南省 23 个辐射观测站总辐照度与观测总辐照度的 MAE、*R* 及 RMSE，以及三种订正方法较 CMA-WSP2.0 模式预报总辐照度的改善效果（图 3，图 4）。由图 3a 可见，CMA-WSP2.0 模式预报 MAE 在 45.88~61.79 $W \cdot m^{-2}$，嵩山、宝丰、西峡、栾川 MAE 较大，在 60 $W \cdot m^{-2}$以上，河南东部地区的 MAE 较小，最小值为永城站 45.88 $W \cdot m^{-2}$，这可能与河南西高东低的地形分布有关，在地形复杂区域，模式偏差较大。经随机森林、XGBoost、LightGBM 订

正后的 MAE 范围分别为 26.36~39.07 W·m^{-2}、25.05~37.17 W·m^{-2}、23.25~34.02 W·m^{-2}。三种机器学习算法订正后的 MAE 明显降低，其中 LightGBM 订正方法订正效果最好，MAE 相较于 CMA-WSP2.0 模式预报减小 18.32~32.91 W·m^{-2}，MAE 减小比例在 38%~56%，XGBoost 订正效果次之，MAE 减小比例在 34%~53%，相较于 XGBoost 订正和 LightGBM 订正，随机森林订正效果最弱，但 MAE 减小比例也在 32%~52%，三种机器学习订正算法均对 CMA-WSP2.0 模式预报总辐照度进行了改善。

从 RMSE 统计结果分析，经过三种机器学习方法订正后的 RMSE 显著降低，CMA-WSP2.0 模式预报总辐照度与观测值的 RMSE 为 106.10~133.04 W·m^{-2}，其中随机森林、XGBoost、LightGBM 订正后的 RMSE 分别为 62.52~92.65、58.25~86.44、53.88~81.25 W·m^{-2}。LightGBM 和 XGBoost 订正后的 RMSE 减小值和减小比例大于随机森林订正（图 4），LightGBM 的 RMSE 减小比例最大，范围在 36%~52%。23 个辐射观测站 CMA-WSP2.0 模式预报与实测值的相关系数为 0.85~0.95，经过订正之后，三种方法相关系数大多都在 0.95 以上。

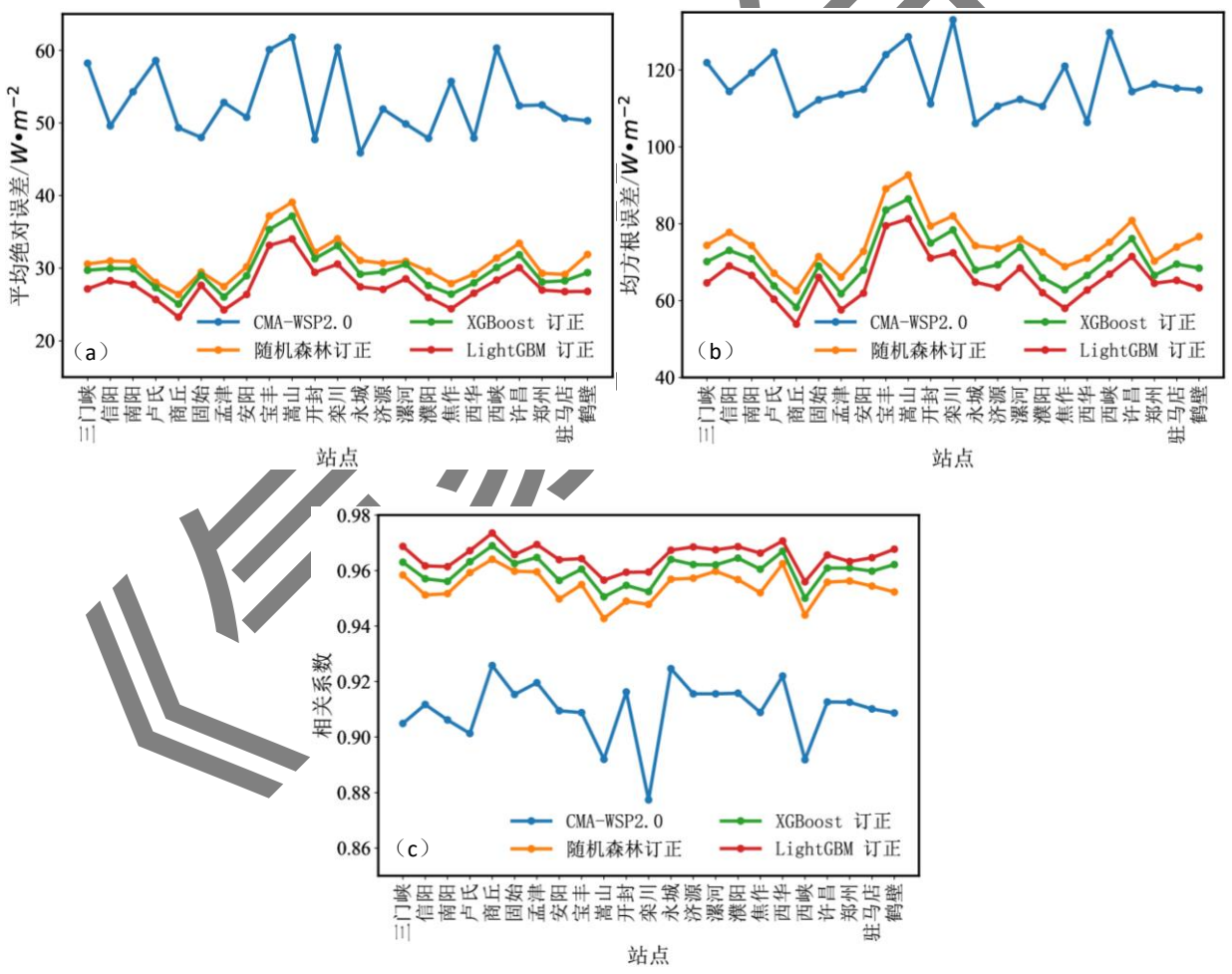


图 3 2022 年 1—12 月河南省各站点测试集中 CMA-WSP2.0、随机森林、XGBoost、LightGBM 预报的总辐照度（a）MAE、（b）RMSE、（c）*R*

Fig.3 The (a) MAE, (b) RMSE and (c)*R* of total irradiance predicted by CMA-WSP2.0, Random Forest, XGBoost and LightGBM at various stations in Henan Province in the test set from January to December 2022

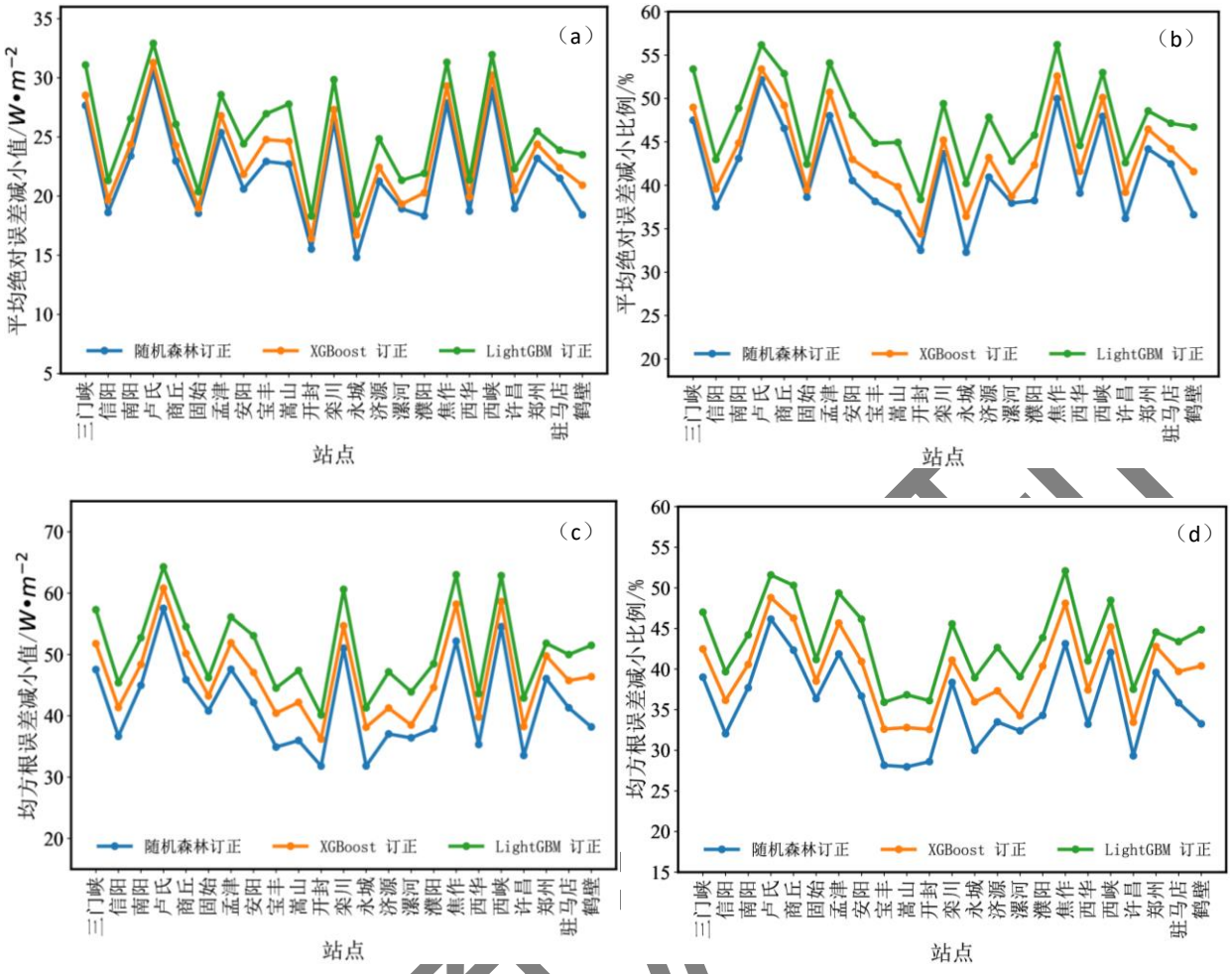


图 4 2022 年 1—12 月河南省各站点测试集中随机森林、XGBoost、LightGBM 预报的总辐照度（a）MAE 减小值、（b）MAE 减小比例、（c）RMSE 减小值、（d）RMSE 减小比例

Fig.4 The (a) reduction values and (b) reduction proportion of MAE, and (c) reduction values and (d) reduction proportion of RMSE of total irradiance predicted by Random Forest, XGBoost and LightGBM at various stations in Henan Province in the test set from January to December 2022

A 和 *Q* 是电网调度考核的两个重要指标，GB/T 40607—2021(国家市场监督管理总局和国家标准化管理委员会，2021)中对光伏预测性能指标要求光伏短期（24 h）功率预测 *A* 高于 85%以上。图 5 给出了河南省各站点总辐照度模式预报与订正后的 24 h 的 *A*、*Q* 检验结果，CMA-WSP2.0 模式预报总辐照度 24 h 的 *A* 在 81.2%~87.1%，经过随机森林订正后 *A* 在 88.3%~91.5%，XGBoost 订正后 *A* 在 89.1%~92.0%，LightGBM 订正后 *A* 在 89.8%~92.5%，相较于 CMA-WSP2.0 模式预报，三种机器学习订正方法均显著提升了 24 h 的 *A*，其中 LightGBM 订正 *A* 最高，24 h 的平均 *A* 较 CMA-WSP2.0 模式预报提升了 7.3%。三种机器学习订正方法对 24 h 的 *Q* 也较 CMA-WSP2.0 模式预报有了明显的提升，LightGBM 提升效果最好，24 h 的平均 *Q* 较 CMA-WSP2.0 模式预报提升了 5.7%，XGBoost 次之，提升了 5.2%。从 MAE、RMSE、24 h 的 *A* 和 *Q* 及 *R* 等统计量分析结果而言，三种机器学习方法订正效果良好，相较于 CMA-WSP2.0 模式预报结果，MAE 和 RMSE 显著降低，24 h 的 *A*、*Q* 显著提升，其中 LightGBM 订正效果最好，XGBoost 订正次之，随机森

林订正效果最弱。

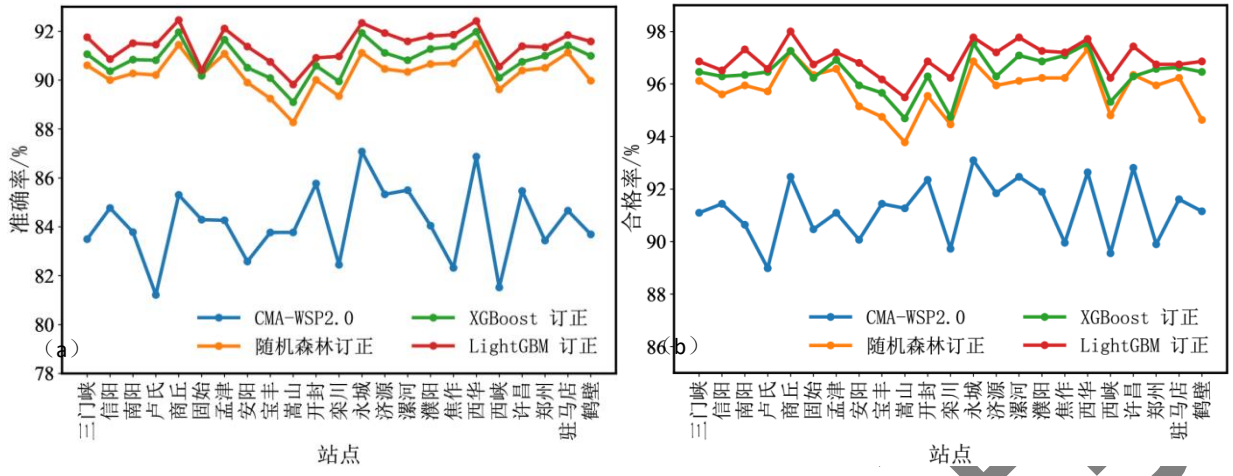


图 5 2022 年 1—12 月河南省各站点测试集中 CMA-WSP2.0、随机森林、XGBoost、LightGBM 预报的总辐照度 24 h（a）*A*、（b）*Q*

Fig.5 The (a)24-hour *A* and (b) 24-hour *Q* of total irradiance predicted by CMA-WSP2.0, Random Forest, XGBoost and LightGBM at various stations in Henan Province in the test set from January to December 2022

为进一步探索不同机器学习算法订正效果，对河南省 5 个区域订正效果进行检验评估。表 2 给出了河南省 5 个区域总辐照度 CMA-WSP2.0 模式预报及三种机器学习方法订正后的 MAE 和 RMSE，CMA-WSP2.0 模式预报的 5 个区域中豫西区域 MAE 和 RMSE 最大，分别为 57.51 W·m^{-2}、123.50 W·m^{-2}；豫东区域最小，分别为 47.71 W·m^{-2}、108.40 W·m^{-2}。豫西区域 4 个辐射观测站平均海拔高度为 576.4 m，豫东区域 4 个辐射观测站平均海拔高度较低（52.3 m），CMA-WSP2.0 模式对于地形复杂、海拔高度较高区域模拟结果差于地势平坦区域。三种订正方法均有效降低了 MAE 和 RMSE，其中，随机森林订正后 MAE 和 RMSE 减小比例在 38%~48%、31%~41%，LightGBM 和 XGBoost 订正优于随机森林订正，LightGBM 又优于 XGBoost 订正，LightGBM 订正后 MAE 和 RMSE 减小比例分别在 44%~53%、39%~48%。对于 5 个区域而言，豫西区域的订正效果最好（表 3）。

表 2 2022 年 1—12 月河南省 5 个分区测试集中 CMA-WSP2.0、随机森林、XGBoost、LightGBM 预报总辐照度的 MAE 和 RMSE

Table 2 The MAE and RMSE of total irradiance predicted by CMA-WSP2.0, Random Forest, XGBoost and LightGBM in five regions of Henan Province in the test set from January to December 2022

区域	**MAE**/ W·m^{-2}				**RMSE**/ W·m^{-2}			
	CMA-WSP2.0	随机森林	XGBoost	LightGBM	CMA-WSP2.0	随机森林	XGBoost	LightGBM
豫东	47.71	29.70	28.38	26.66	108.04	72.06	67.19	63.41
豫西	57.51	30.02	29.03	26.90	123.50	72.68	68.82	63.98
豫南	52.57	30.37	29.45	27.76	118.34	74.55	70.72	66.75
豫北	51.32	30.03	28.37	26.12	114.44	72.92	66.91	61.77
豫中	55.32	33.98	32.59	30.54	119.30	82.17	77.63	73.31

表 3 2022 年 1—12 月河南省 5 个分区测试集中随机森林、XGBoost、LightGBM 预报总辐照度 MAE 和 RMSE 的减小比例

Table 3 The reduction values and reduction proportion of the MAE and RMSE of total irradiance predicted by Random Forest, XGBoost and LightGBM in five regions of Henan Province in the test set from January to December 2022

区域	MAE 减小比例/%			RMSE 减小比例/%		
	随机森林	XGBoost	LightGBM	随机森林	XGBoost	LightGBM
豫东	38	41	44	33	38	41
豫西	48	50	53	41	44	48
豫南	42	44	47	37	40	44
豫北	41	45	49	36	42	46
豫中	39	41	45	31	35	39

进一步分析预报值与实际观测值之间的偏差在整个数据集上的分布情况。由于夜晚总辐照度观测为 0 $W \cdot m^{-2}$，在数据中占比较大，为了更好地分析订正效果，去除了模拟和观测同时为零值的数据。图 6 给出了河南省 5 个区域模式预报及三种机器学习订正后的总辐照度与观测总辐照度的偏差概率分布特征，结果显示，三种机器学习订正后的偏差概率分布更表现出正态分布的特征，而 CMA-WSP2.0 模式预报的偏差概率更多分布在正值，模式预报较观测数值偏大。经三种机器学习算法订正后的偏差出现在零值附近的概率更高，且偏差分布集中范围更窄，其中 LightGBM 订正效果更好。

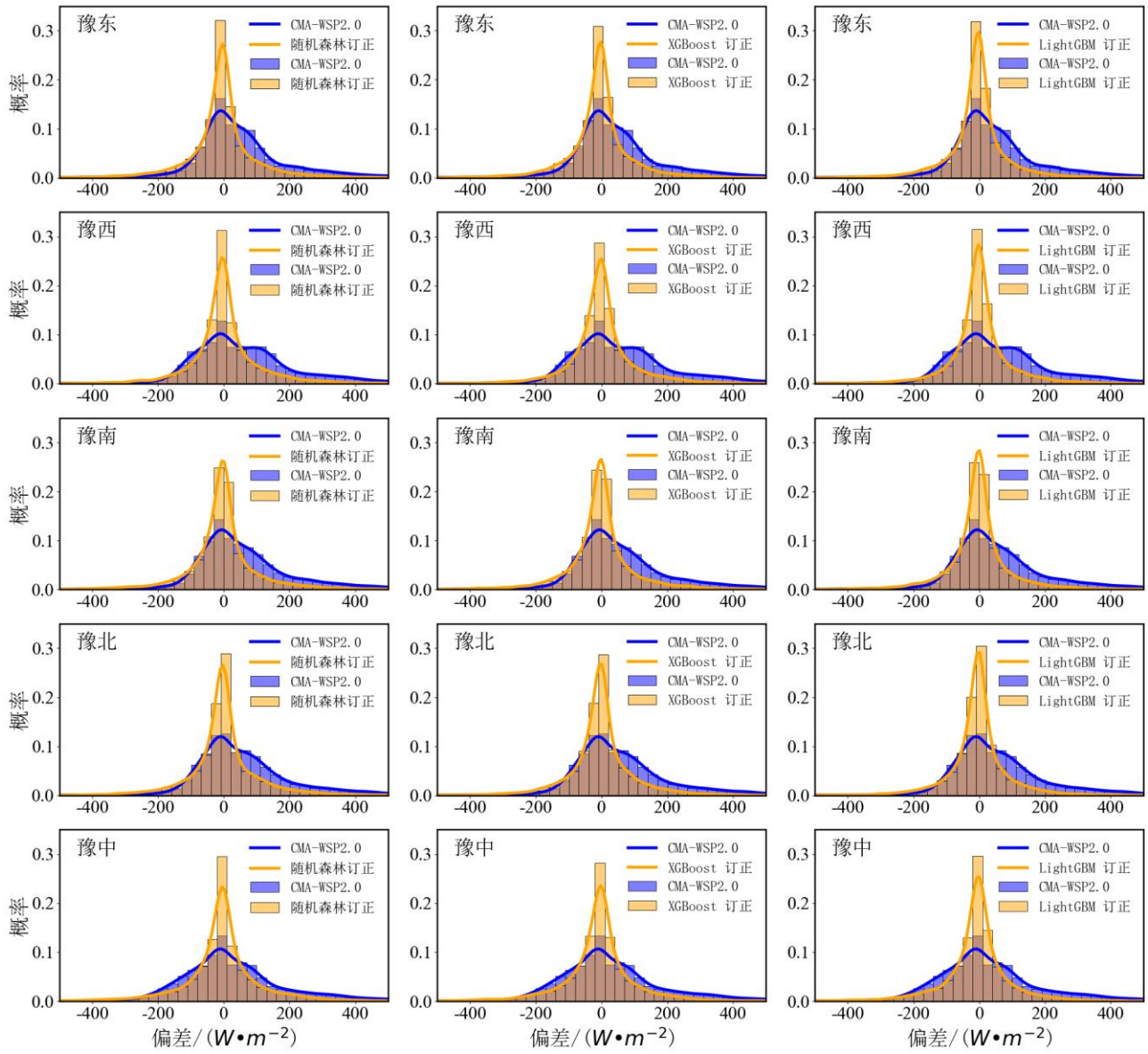


图 6 2022 年 1—12 月河南省 5 个分区测试集中模式预报及三种机器学习订正后的总辐照度与观测的偏差概率分布

Fig.6 The probability distribution of deviation of total irradiance predicted by CMA-WSP2.0 and three types of machine learning in five regions of Henan Province in the test set from January to December 2022

3.2 分季节评估

选取 1 月、4 月、7 月、10 月分别代表冬、春、夏、秋，对河南省 5 个区域不同季节总辐照度的订正效果进行分析检验，从 MAE 和 RMSE 减小比例可知（图 7,，图 8），三种机器学习订正方法对河南省不同区域各季节的总辐照度模拟结果均有一定的提升效果。四季总辐照度的订正效果整体表现为冬季订正效果最好，秋季次之，对于夏季的订正效果最弱。在三种机器学习订正方法中，LightGBM 订正和 XGBoost 订正优于随机森林订正，LightGBM 订正又优于 XGBoost 订正。随机森林 MAE 和 RMSE 减小比例范围分别为 25%~69%、26%~64%，而 LightGBM 则分别为 32%~72%、33%~70%。

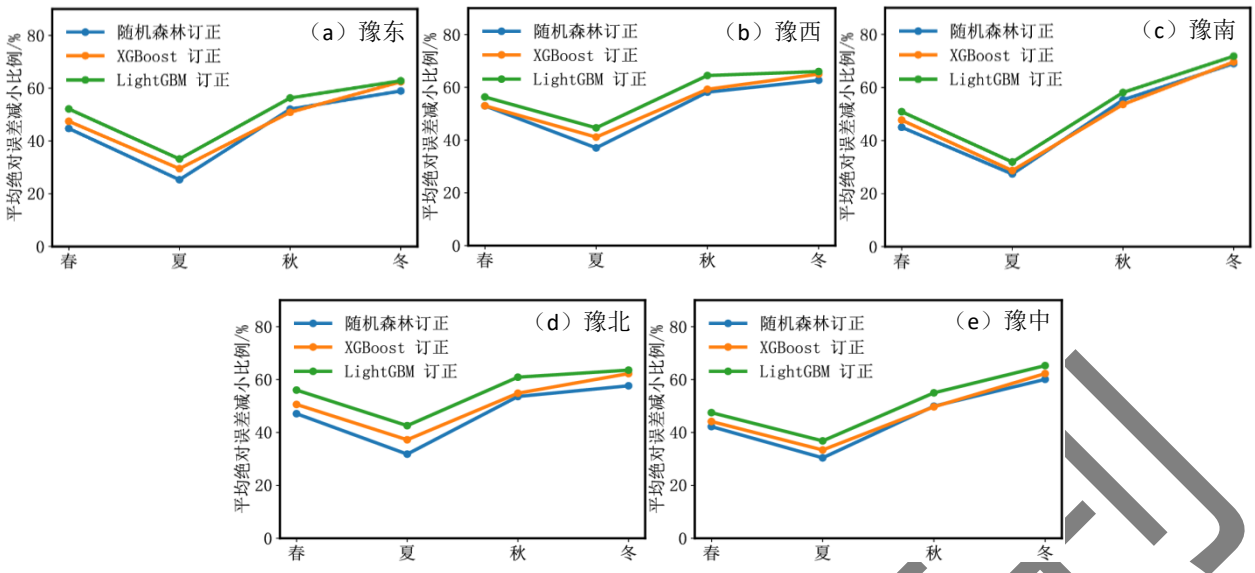


图 7 2022 年 1—12 月河南省 5 个分区测试集中随机森林、XGBoost、LightGBM 预报总辐照度在河南省 5 个分区不同季节 MAE 减小比例

Fig.7 The reduction proportion of the MAE of total irradiance predicted by Random Forest, XGBoost and LightGBM in different seasons in five regions of Henan Province in the test set from January to December 2022

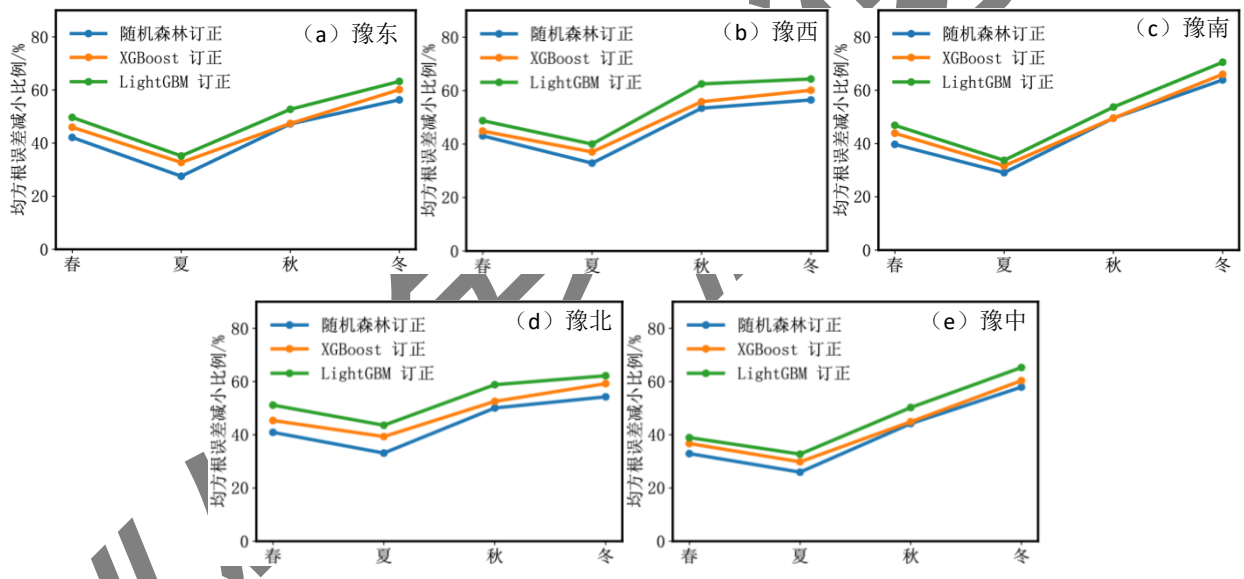


图 8 2022 年 1-12 月河南省 5 个分区测试集中随机森林、XGBoost、LightGBM 预报总辐照度不同季节 RMSE 减小比例

Fig.8 The reduction proportion of the RMSE of total irradiance predicted by Random Forest, XGBoost and LightGBM in different seasons in five regions of Henan Province in the test set from January to December 2022

3.3 总辐照度分级评估

将地面观测总辐照度划分为以下等级（0，100）、［100，200）、［200，300）、［300，400）、［400，500）、［500，600）、［600，700）、［700，800）、［800，+∞）$W \cdot m^{-2}$（徐丽娜等，2021；杨宣等，2024），分析各区域 CMA-WSP2.0 模式预报及不同订正方法订正效果随总辐照度不同等级的分布特征。从 MAE 分布可见（图 9），CMA-WSP2.0 模式预报 MAE 在[58.2，191.0] $W \cdot m^{-2}$区间内，在不同的总辐

照度等级，MAE 明显不同，当总辐照度<300 W·m^{-2}时，MAE 随总辐照度的增大而增大，之后 MAE 逐渐减小，在总辐照度 700 W·m^{-2}左右 MAE 又开始增大，各个区域在总辐照度［200，300）等级的 MAE 最大。经过三种机器学习方法订正后 MAE 明显减小，MAE 减小比例随着总辐照度的增大总体呈减小趋势，LightGBM 订正效果最好，MAE 减小比例在 12%~65%，对于总辐照度<500 W·m^{-2}的等级，MAE 减小比例均在 40%以上，而对于总辐照度在[800,+∞)等级则在 10%左右。RMSE（图 10）显示出与 MAE 相似的规律，三种机器学习方法均有效改善了 CMA-WSP2.0 模式预报，对于不同的总辐照度等级，改善效果存在有差别，RMSE 减小比例随着总辐照度的增大总体呈减小趋势。

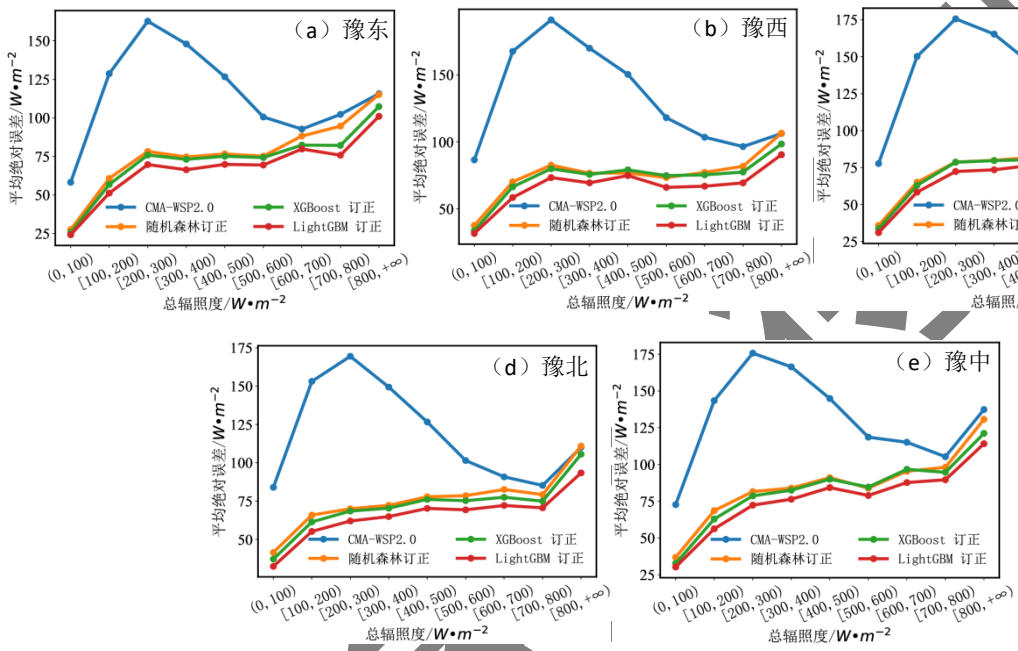
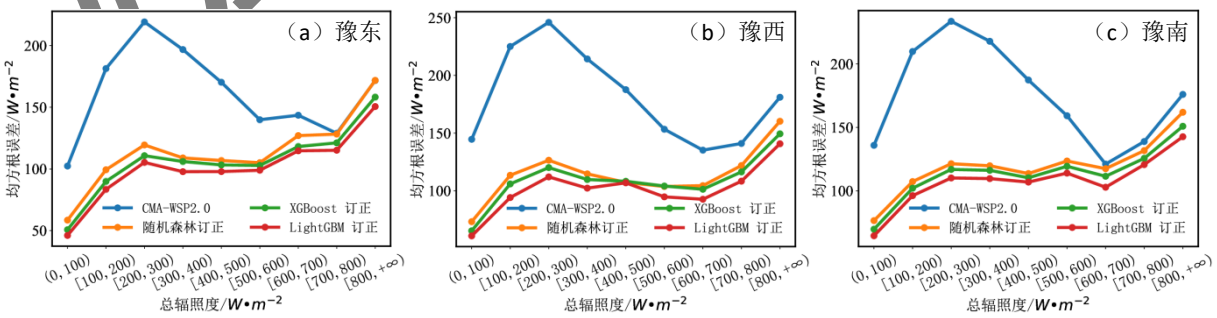


图 9 2022 年 1—12 月河南省 5 个分区测试集中 CMA-WSP2.0、随机森林、XGBoost、LightGBM 预报总辐照度在不同总辐照度等级下的 MAE

Fig.9 The MAE in different total irradiance levels of total irradiance predicted by CMA-WSP2.0, Random Forest, XGBoost and LightGBM in five regions of Henan Province in the test set from January to December 2022



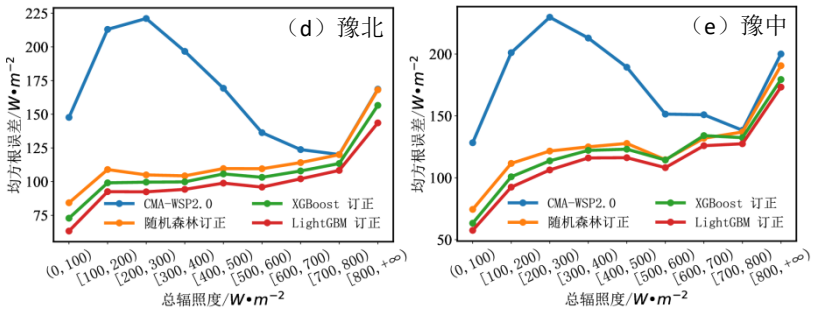


图 10 2022 年 1—12 月河南省 5 个分区测试集中 CMA-WSP2.0、随机森林、XGBoost、LightGBM 预报总辐照度在不同总辐照度等级下的 RMSE

Fig.10 The RMSE in different total irradiance levels of total irradiance predicted by CMA-WSP2.0, Random Forest, XGBoost and LightGBM in five regions of Henan Province in the test set from January to December 2022

4 结论

使用 2022 年河南省 23 个辐射观测站总辐照度数据和 CMA-WSP2.0 模式产品，通过 LASSO 回归选取特征变量，建立训练数据集和测试数据集，并利用训练数据集采用机器学习方法（随机森林、XGBoost、LightGBM）进行模型训练，订正河南省 CMA-WSP2.0 模式预报总辐照度，并对订正结果分站点、分区域、分季节、总辐照度分级进行检验，主要得到以下结论：

（1）河南省 23 个辐射观测站统计检验表明，随机森林、XGBoost、LightGBM 订正后的总辐照度相比 CMA-WSP2.0 模式预报结果，与观测总辐照度的 MAE 和 RMSE 显著降低，24 h 的 A 和 Q 显著提升。其中 LightGBM 订正效果最优，MAE 相较于 CMA-WSP2.0 模式预报减小 18.32~32.91 $W \cdot m^{-2}$，MAE 减小比例在 38%~56%，RMSE 减小比例在 36%~52%；24 h 的平均 A、Q 较 CMA-WSP2.0 模式预报提升了 7.3%、5.7%。

（2）区域统计与站点统计结果较为一致，三种订正方法均有效降低了 MAE 和 RMSE，其中，LightGBM 订正后 MAE 和 RMSE 减小比例在 44%~53%、39%~48%，LightGBM 和 XGBoost 订正优于随机森林订正，LightGBM 又优于 XGBoost 订正。对于 5 个区域而言，对于豫西区域订正效果最好。偏差的概率分布表明，三种机器学习方法订正后的偏差出现在零值附近的概率更高，且偏差分布集中范围更窄，其中 LightGBM 订正效果更为显著。

（3）四个季节检验评估结果表明，冬季订正效果最好，秋季次之，对于夏季的订正效果最弱。在三种机器学习订正方法中，LightGBM 订正效果最好，MAE 和 RMSE 减小比例在 32%~72%、33%~70%。

（4）对于不同的总辐照度等级，三种机器学习方法均有效改善了 CMA-WSP2.0 模式预报，随着总辐照度等级的增加，订正效果总体呈逐渐减弱的趋势。

由上述分析可知，本文运用的三种机器学习方法对 CMA-WSP2.0 模式总辐照度预报进行了较好的改善，未来将进一步研究生成总辐照度格点预报订正产品，从而达到更好的预报效果。

《气象》待刊

参考文献

白永清, 陈正洪, 王明欢, 等, 2013. 关于 WRF 模式模拟到达地表短波辐射的统计订正[J]. 华中师范大学学报(自然科学版), 47(2): 292-296. Bai Y Q, Chen Z H, Wang M H, et al, 2013. Statistical correction on the surface shortwave radiation forecasted by WRF model[J]. J Huazhong Norm Univ (Nat Sci), 47(2): 292-296 (in Chinese).

陈有龙, 宁雨珂, 唐荣年, 等, 2020. 基于时空独立的随机森林模型对海南热带气温数值预报的订正[J]. 海南大学学报(自然科学版), 38(4): 356-364. Chen Y L, Ning Y K, Tang R N, et al, 2020. Tropical temperature correction for numerical forecast in Hainan based on spatiotemporal independence random forest model[J]. Nat Sci J Hainan Univ, 38(4): 356-364 (in Chinese).

陈昱文, 黄小猛, 李熠, 等, 2020. 基于 ECMWF 产品的站点气温预报集成学习误差订正[J]. 应用气象学报, 31(4): 494-503. Chen Y W, Huang X M, Li Y, et al, 2020. Ensemble learning for bias correction of station temperature forecast based on ECMWF products[J]. J Appl Meteor Sci, 31(4): 494-503 (in Chinese).

方鸿斌, 王珊珊, 王晓玲, 等, 2024. 基于机器学习的格点气温预报订正方法[J]. 气象, 50(1): 103-114. Fang H B, Wang S S, Wang X L, et al, 2024. Gridded temperature forecast correction method based on machine learning[J]. Meteor Mon, 50(1): 103-114 (in Chinese).

顾婷婷, 潘娅英, 张加易, 2022. 浙江省中尺度数值预报系统的地表太阳辐射预报订正方法[J]. 干旱气象, 40(2): 327-332. Gu T T, Pan Y Y, Zhang J Y, 2022. Correction method of surface solar radiation forecast based on ZJWARMS[J]. J Arid Meteor, 40(2): 327-332 (in Chinese).

国家市场监督管理总局, 国家标准化管理委员会, 2021. GB/T 40607-2021 调度侧风电或光伏功率预测系统技术要求[S]. 北京: 中国标准出版社. State Administration for Market Regulation, National Standardization Administration, 2021. GB/T 40607-2021 Technical requirements for dispatching side forecasting system of wind or photovoltaic power[S]. Beijing: Standards Press of China (in Chinese).

李韬, 王磊, 李月英, 等, 2022. 基于随机森林的 EC 气温预报订正研究[J]. 农业灾害研究, 12(6): 95-97. Li T, Wang L, Li Y Y, et al, 2022. Study on EC temperature forecast revision based on random forest[J]. Journal of agricultural catastrophology, 12(6): 95-97 (in Chinese). **(查阅网上资料，未找到本条文献刊名缩写，请确认)**

李文娟, 郦敏杰, 马昊, 等, 2024. 基于 XGBoost 分类和数值模式“配料”的浙江强对流预报方法[J]. 气象, 50(11): 1343-1358. Li W J, Li M J, Ma H, et al, 2024. Severe convection prediction method based on XGBoost classified algorithm and numerical model ingredients[J]. Meteor Mon, 50(11): 1343-1358 (in Chinese).

李遥, 李照荣, 王小勇, 等, 2020. 基于斜面辐射算法的短期光伏功率预测方法研究[J]. 干旱气象, 38(5): 869-877. Li Y, Li Z R, Wang X Y, et al, 2020. Prediction methods of short-term photovoltaic power based on inclined plane solar radiation algorithm[J]. J Arid Meteor, 38(5): 869-877 (in Chinese).

刘淑贤, 张立生, 刘扬, 等, 2024. 基于机器学习的热带气旋灾害等级评估模型构建及其活动特征分析[J]. 气象, 50(3): 331-343. Liu S X, Zhang L S, Liu Y, et al, 2024. Construction of tropical cyclone disaster grade assessment model based on machine learning and analysis of its activity characteristics[J]. Meteor Mon, 50(3): 331-343 (in Chinese).

芦华, 谢旻, 吴铖, 等, 2020. 基于机器学习的成渝地区空气质量数值预报 $PM_{2.5}$ 订正方法研究[J]. 环境科学学报, 40(12): 4419-4431. Lu H, Xie M, Wu Z, et al, 2020. Adjusting $PM_{2.5}$ prediction of the numerical air quality forecast model based on machine learning methods in Chengyu region[J]. Acta Sci Circumst, 40(12): 4419-4431 (in Chinese).

普智勇, 夏攀, 张璐, 等, 2023. 机器学习与统计方法在太阳能预报中的比较性分析[J]. 太阳能学报, 44(7): 162-167. Pu Z Y, Xia P, Zhang L, et al, 2023. Comparative analysis of machine learning and statistical methods in solar energy prediction[J]. Acta Energ Solaris Sin, 44(7): 162-167 (in Chinese).

孙朋杰, 陈正洪, 成驰, 等, 2015. 一种改进的太阳辐射 MOS 预报模型研究[J]. 太阳能学报, 36(12): 3048-3053. Sun P J, Chen Z H, Cheng C, et al, 2015. Improved MOS model for solar radiation forecasting[J]. Acta Energiae Solaris Sin, 36(12): 3048-3053 (in Chinese).

孙全德, 焦瑞莉, 夏江江, 等, 2019. 基于机器学习的数值天气预报风速订正研究[J]. 气象, 45(3): 426-436. Sun Q D, Jiao R L, Xia J J, et al, 2019. Adjusting wind speed prediction of numerical weather forecast model based on machine learning methods[J]. Meteor Mon, 45(3): 426-436 (in Chinese).

万超, 刘涛, 曾莉萍, 等, 2023. 基于 CMA-WSP1.0 的贵州短波辐射检验分析[J]. 科技创新与应用, 13(30): 104-107. Wan C, Liu T, Zeng L P, et al, 2023.

Analysis of Guizhou shortwave radiation testing based on CMA-WSP1.0[J]. Technol Innov Appl, 13(30): 104-107 (in Chinese).

王雪洁, 施国萍, 周子钦, 等, 2022. 基于随机森林算法对 ERA5 太阳辐射产品的订正[J]. 自然资源遥感, 34(2): 105-111. Wang X J, Shi G P, Zhou Z Q, et al, 2022. Revision of solar radiation product ERA5 based on random forest algorithm[J]. Remote Sens Nat Resour, 34(2): 105-111 (in Chinese).

徐景峰, 宋林烨, 陈明轩, 等, 2023. 冬奥会复杂山地百米尺度 10m 风速预报的机器学习订正对比试验[J]. 大气科学, 47(3): 805-824. Xu J F, Song L Y, Chen M X, et al, 2023. Comparative machine learning-based correction experiment for a 10 m wind speed forecast at a 100 m resolution in complex mountainous areas of the Winter Olympic Games[J]. Chin J Atmos Sci, 47(3): 805-824 (in Chinese).

许立兵, 王安喜, 汪纯阳, 等, 2020. 基于机器学习的海洋环境预报订正方法研究[J]. 海洋通报, 39(6): 695-704. Xu L B, Wang A X, Wang C Y, et al, 2020. Research on correction method of marine environment prediction based on machine learning[J]. Mar Sci Bull, 39(6): 695-704 (in Chinese).

徐丽娜, 申彦波, 李忠, 等, 2021. 基于概率密度匹配方法的 FY-4A 地表入射太阳辐射订正[J]. 高原气象, 40(4): 932-942. Xu L N, Shen Y B, Li Z, et al, 2021. Correction of FY-4A surface solar irradiance based on probability density function matching method[J]. Plateau Meteor, 40(4): 932-942 (in Chinese).

杨宣, 魏璐, 马百胜, 等, 2024. FY-4A 卫星地表太阳辐射产品在河南省适用性评估与订正方法研究[J]. 高原气象, 43(6): 1600-1613. Yang X, Wei L, Ma B S, et al, 2024. Assessment of the applicability and calibration methods of FY-4A satellite surface solar radiation products in Henan Province[J]. Plateau Meteor, 43(6): 1600-1613 (in Chinese).

智协飞, 王田, 季焱, 2020. 基于深度学习的中国地面气温的多模式集成预报研究[J]. 大气科学学报, 43(3): 435-446. Zhi X F, Wang T, Ji Y, 2020. Multimodel ensemble forecasts of surface air temperature over China based on deep learning approach[J]. Trans Atmos Sci, 43(3): 435-446 (in Chinese).

周康辉, 郑永光, 韩雷, 等, 2021. 机器学习在强对流监测预报中的应用进展[J]. 气象, 47(3): 274-289. Zhou K H, Zheng Y G, Han L, et al, 2021. Advances in application of machine learning to severe convective weather monitoring and forecasting[J]. Meteor Mon, 47(3): 274-289 (in Chinese).

张敏, 袁心仪, 张顾, 等, 2024. CMA-WSP2.0 在江苏地表太阳辐射预报中的检验评估[J]. 气象研究与应用, 45(1): 17-22. Zhang M, Yuan X Y, Zhang G, et al, 2024. Validation and evaluation of CMA-WSP2.0 in surface solar radiation forecasting in Jiangsu[J]. Joural of Meteorological Research and Application , 45(1): 17-22 (in Chinese).

张凡, 程凯琪, 2024. 基于降雪识别的河南省降雪及降雪比率时空变化特征[J]. 气象与环境科学, 47(4): 13-22. Zhang F, Cheng K Q, 2024. Spatio-temporal variation characteristics of snowfall and snowfall ratio in Henan Province based on snowfall Identification[J]. Meteorological and Environmental Sciences, 47(4): 13-22 (in Chinese).

Belmahdi B, el Bouardi A, 2024. Short-term solar radiation forecasting using machine learning models under different sky conditions: evaluations and comparisons[J]. Environ Sci Pollut Res, 31(1): 966-981.

Breiman L, 2001. Random forests[J]. Mach Learn, 45(1): 5-32.

Delannoy L, Longaretti P Y, Murphy D J, et al, 2021. Peak oil and the low-carbon energy transition: a net-energy perspective[J]. Appl Energ, 304: 117843.

Haupt S E, Kosovic B, 2017. Variable generation power forecasting as a big data problem[J]. IEEE Trans Sustain Energ, 8(2): 725-732.

Tibshirani R, 2011. Regression shrinkage and selection via the lasso: a retrospective[J]. J Roy Stat Soc Ser B: Stat Methodol, 73(3): 273-282.

Zhang Y H, Ge T T, Tian W, et al, 2019. Debris flow susceptibility mapping using machine-learning techniques in Shigatse Area, China[J]. Remote Sens, 11(23): 2801.