doi: 10.6046/zrzyyg.2021151
引用格式: 王雪洁,施国萍,周子钦,等. 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,2022,34(2):105-111. (Wang X J, Shi G P, Zhou Z Q, et al. Revision of solar radiation product ERA5 based on random forest algorithm[J]. Remote Sensing for Natural Resources, 2022, 34(2): 105-111.)

基于随机森林算法对ERA5太阳辐射产品的订正

王雪洁[1], 施国萍[2], 周子钦[1], 甄 洋[1]
(1. 南京信息工程大学长望学院,南京 210044; 2. 南京信息工程大学地理科学学院,南京 210044)

摘要: 为了进一步提高太阳辐射量空间分布资料的精度,利用2013年93个中国太阳辐射逐时资料,对欧洲中期天气预报中心(ECMWF)ERA5平均地表太阳下行短波辐射产品(0.25°×0.25°)进行多尺度的误差分析,并利用多种相关的气象、地理等要素训练随机森林模型,对ERA5总辐射产品进行订正与分析,最后利用该模型得到订正后的逐时辐射量空间分布,使得再分析资料更好地应用于农业、电力和城市建设等行业。研究结果表明: ①2013年ERA5太阳辐射量与站点观测量的*MAE*,*RMSE*和*R*分别为27.60 W/m^2,29.87 W/m^{2}和0.97,且ERA5值比站点实测值偏高; ②利用随机森林订正后精度得到提高,校正后ERA5太阳辐射量与站点实测值的*MAE*,*RMSE*,*R*分别为3.34 W/m^2,3.85 W/m^2,1.00,相关性明显提高; ③订正前后的辐射量的空间宏观分布规律一致,但是ERA5太阳辐射量在局部地区有明显的下降。

关键词: ERA5产品; 太阳辐射量; 随机森林订正
中图法分类号: P 422.1 **文献标志码**: A **文章编号**: 2097-034X(2022)02-0105-07

0 引言

太阳能作为地球最主要的能量来源和基本动力,推动了地表的几乎全部自然地理过程,使地理环境得以形成和有序发展[1]。也是气候形成和演变过程中重要的外参数[2],是陆面过程的主要驱动因子。太阳辐射影响大气圈、水圈、陆地圈层中的物质与能量交换,对太阳辐射研究可以促进对碳循环、水循环等的研究,对全球气候变化也有重要意义。

随着科技的进步,农业、电力和城市建设等行业对太阳辐射的研究提出了新的要求[3-5]。我国太阳辐射观测台站较少,且测站空间分布不均匀。而再分析数据,即经过对太阳辐射观测资料(包括地面观测、卫星,还有雷达、探空等)的质量控制,再同化入全球模式后得到的数据,有着时间序列长、空间分布广的特点,可以极大弥补地面观测数据的不足。但是,再分析数据受资料源、模式等影响,无法完全达到真实模拟大气的程度,具有一定程度的偏差,建立订正模型对于使用再分析数据来说显得极为重要。目前已有相关研究利用地面站点地表辐射数据对再分析辐射数据进行了多尺度的验证。研究表明太阳辐射与云量、气溶胶、水汽等有关[6-7]。再分析资料与我国太阳辐射站点观测资料相比,绝大部分高于台站数据[8-10]。6种再分析地表辐射产品(NCEP-NCAR reanalysis, NCEP-DOE reanalysis, Climate Foreast System Reanalysis (CFSR), ECMWF Interim Reanalysis (ERA Interim), Modem-Era Retrospective Analysis for Research and Applications (MERRA) reanalysis, The Japanese 55-year reanalysis),全球月均偏差为11.25~49.80 W/m^2,并且发现在中国范围上,云量和气溶胶的低估都可能导致再分析地表辐射的高估,夏秋季节明显好于春冬季节[8-9]。

随着机器学习的发展,越来越多的研究者开始使用机器学习进行不同地区的太阳辐射的预报偏差的订正。大量的研究表明,机器学习模型效果较理论参数模型、经验模型更加准确。陈昱文等[11]利用气象站点的4个观测要素,挖掘观测数据的时序特征并结合气温预报结果训练机器学习模型,对结果进行偏差订正,发现集成学习方法在数值模式预报结果订正中具有较大的应用潜力; 李净等[12]利用

收稿日期: 2021-05-18; **修订日期**: 2021-08-12
基金项目: 国家自然科学基金青年基金项目"基于SUNFLUX辐射参数化计算方案的起伏地形云天实际地表太阳辐射分布式模拟研究及其在陆面过程中的应用"(编号: 41805083)资助。
第一作者: 王雪洁(1999-),女,本科,主要从事3S集成与气象应用研究。Email: 201883330052@nuist.edu.cn。
通信作者: 施国萍(1984-),女,博士,副教授,主要从事3S集成与气象应用研究。Email: shiguopingnj@163.com。

ERA5(ECMWF Reanalysis 5)等产品,将人工神经网络、支持向量机和随机森林 3 种机器学习模拟黄土高原地区的太阳辐射并对 3 种方法进行比较; Benali 等[13]将智能持久性、人工神经网络和随机森林这 3 种方法进行比较,预测了法国的太阳辐射,二者结果都表明随机森林的模拟精度最高; Yu 等[14]利用 4 种机器学习方法,包括梯度提升回归树、随机森林、多元自适应回归样条和人工神经网络对地面 96 个站点日、月尺度的太阳辐射进行模拟评估,验证了训练数据集基于随机森林方法的太阳辐射估计值与地面测量值的相关性最好。随机森林的订正方法在海洋环境预报[15]、气温数值预报[16]、空气质量[17]等方面都有应用且精度很高。Babar 等[18]利用 ERA5 和云、反照率、辐射数据集(CLARA - A2)建立随机森林回归模型对挪威地区日平均全球水平辐照度(GHI)进行估计,发现随机森林模型的估计值较原来的预报值更精确,能够更好地估计。但是目前很少有研究对中国范围内的 ERA5 的高精度太阳辐射数据进行空间订正,没有连续的空间分布订正资料。

本文利用 2013 年全国 93 个辐射站的总辐射逐时观测资料对 ERA5 同期再分析辐射资料进行了评估,并选择相关气象要素及地理要素作为随机森林学习的输入量,对全国 93 个站点上 ERA5 辐射量的值进行了订正,进而对 ERA5 辐射产品进行空间分布上的订正,得到订正后的逐时辐射空间分布图。利用 ERA5 辐射产品作为输入变量进行随机森林回归,对中国范围内的高空间分辨率格网数据进行逐时数值订正,解决了太阳辐射站点不均的问题,为高精度太阳辐射量空间分布资料的获取提供一种方法。

1 数据源与研究方法

1.1 数据源

①使用欧洲中心天气预报中心(European Centre for Medium - Range Weather Forecasts, ECMWF)发布的第五代再分析数据集 ERA5(ECMWF Reanalysis 5)中相关数据进行分析与订正,主要包括太阳下行短波辐射(mean surface downward shot - wave radiation flux, MSDWSWRF)、地表反照率、水汽、总云量、臭氧、高云、低云、中云、冰云、水云产品,时间为 2013 年,时间分辨率为 1 h,空间分辨率为 0.25° × 0.25°; ②中国 93 个辐射站点信息(经度、纬度和海拔),以及 2013 年逐时的太阳总辐射量。

1.2 研究方法

1.2.1 数据评估指标

采用绝对误差(absolute error, AE)、平均绝对误差(mean absolute error, MAE)、均方根误差(root mean squared error, RMSE)和相关系数(R)来分别描述偏离程度、不确定性、准确性和相关性。表达式分别为:

$$AE = | x - y | \quad , \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | x_i - y_i | \quad , \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \quad , \tag{3}$$

$$R = \frac{Cov(x,y)}{\sqrt{D(x)}\ \sqrt{D(y)}} \quad , \tag{4}$$

式中: x 为估计值; y 为观测值; n 为样本数; i 为样本序号, $i = 1, 2, \cdots, n$; $Cov(\cdot)$ 为协方差函数; $D(\cdot)$ 为方差函数。

本文主要运用 MAE, $RMSE$ 和 R 进行逐时资料的再分析资料与实测值、逐时订正值与实测值之间的误差对比分析。AE 用于比较 ERA5 再分析资料的模拟值和随机森林回归值与中国地面站点辐射量的月均实测值之间的差异。

1.2.2 随机森林回归

1)随机森林算法。该算法是基于多棵决策树的一种集成学习算法,且森林中的每一棵决策树之间没有关联,模型的最终输出由森林中的每一棵决策树共同决定。随机森林用于分类时,采用 N 个决策树分类,将分类结果采用简单投票法得到最终分类,提高分类准确率。选取与太阳辐射有关的因子,即时间、经纬度、地表反照率、海拔、天顶角余弦、总云量、臭氧、水汽、低云、中云、高云、冰云、水云作为输入数据,输出量为每小时辐射值。算法步骤为: ①用有抽样放回的方法(bootstrap)从样本集中选取 n 个样本作为一个训练集; ②用抽样得到的样本集生成一棵决策树,在每一个决策树节点不重复地选择 m 个特征,使用基尼指数找到最佳的划分特征; ③重复第二步 N 次之后生成 N 棵决策树; ④用训练得到的随机森林对测试样本进行预测,并采用票选法决定预测的结果。

2)特征重要性分析。基尼指数用作对随机森林训练样本特征进行重要性分析,比较每个特征在随机森林中的每棵树所做的贡献。基尼指数越小表示集合中被选中的样本被分错的概率越小,也就是说集合的纯度越高,反之,集合越不纯。设 VIM 为变量重要性评分, GI 为基尼指数,假设有 m 个特征 $X_1, X_2, \cdots, X_m$,则 GI 的计算公式为:

$$GI_m = \sum_{k=1}^{|K|}\sum_{k'\neq k} p_{mk}p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \quad , \quad (5)$$

式中：K 为类别数；p_{mk} 为节点 m 中类别 k 所占的比例。在节点 m 上，特征 X_i 在节点 m 分支前后的基尼指数变化为：

$$VIM_{im} = GI_m - GI_g - GI_h \quad , \quad (6)$$

式中 GI_g 和 GI_h 为分支后新的基尼指数。特征 X_i 在第 j 棵决策树的重要性为特征 X_i 在决策树 j 中出现的节点 m 的基尼指数变化量的和，记为 VIM_{ij}。若随机森林模型中有 N 棵树，特征 X_i 的变量重要性评分 VIM_i 为出现的所有决策树 j 中的 VIM_{ij} 的总和。最后，把所有求得的重要性评分做一个归一化处理即可。

3）K 折交叉验证。该方法用于模型调优，可以减少过拟合的问题。为了进一步检验模型泛化能力，基于独立样本数据，因训练集较大选择 5 折交叉验证以降低训练成本。步骤为：将数据分为 5 组，每次从训练集中，抽取出 5 份中的一份数据作为验证集，剩余 4 组作为测试集，重复 5 次。测试结果采用 5 组数据的测试误差的平均值作为最后精度评价。原始数据集划分成训练集和测试集以后，其中测试集除了用作调整参数，也用来测量模型的好坏。K 折交叉验证对网格搜索（GridSearchCV）是很重要的，用来选择模型的最优参数，本文将全部数据集按照 7∶3 划分为训练集和测试集进行 5 折交叉验证。

1.3　随机森林模型

1.3.1　模型参数取值

使用随机森林模型进行学习时，参数对模型准确度意义重大。网格搜索算法（GridSearchCV）是一种通过遍历给定的参数组合来优化模型表现的方法，再利用 K 折交叉验证，得到最优模型。随机森林算法参数众多，最终优化模型参数取值如表 1 所示。

表 1　模型参数取值

Tab. 1　Value of model parameters

模型参数	参数取值
子树数量	56
衡量分裂质量的性能	GI
最佳分裂点时考虑的属性数目	auto
树的最大深度	30
叶节点最小样本数	5
分割内部节点的最小样本数	8
Bootstrap 抽样	TRUE
是否使用袋外样本	FALSE
随机数生成器使用的种子	None

随机森林模型参数优化的一般步骤是：先保持其他参数为默认值，对待定参数设置范围，然后不断缩小范围，最终确定参数值。子树数量对模型的准确性影响最大，设置过低会导致模型不准确，设置过高会增加模型复杂度，所以首先确定子树数量。设置范围从[50,70]缩小为[50,57]，最终确定子树数量为 56。其他参数仍然利用网格搜索方法，得到最终模型。

1.3.2　5 折交叉验证模型的精度

经过 5 折交叉验证模型，结果如表 2 所示，决定系数平均值为 0.855，说明建立的随机森林模型的精度较高，模型模拟较优，稳定性良好。

表 2　5 折交叉验证模型结果

Tab. 2　Result of model 5 – fold cross – validation

次数	决定系数 R^2
1	0.865
2	0.851
3	0.854
4	0.868
5	0.837
平均值	0.855

1.3.3　特征重要性分析

特征重要性可以看出每个输入量对模型预报所做的贡献，将时间、经纬度、地表反照率、海拔、天顶角、水汽、总云量、臭氧、高云、低云、中云、冰云、水云作为最终输入量，输出量为每小时辐射量。利用基尼指数作为评价指标来衡量特征重要性。由图 1 可以看出，天顶角数据重要性最大，为 0.325，高云重要性最小，为 0.004。说明天顶角对地表太阳辐射量影响较大，高云对地表太阳辐射量影响很小。

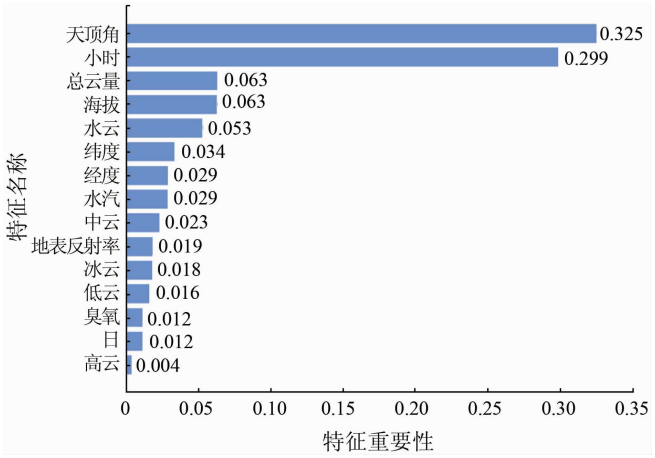


图 1　特征重要性比较

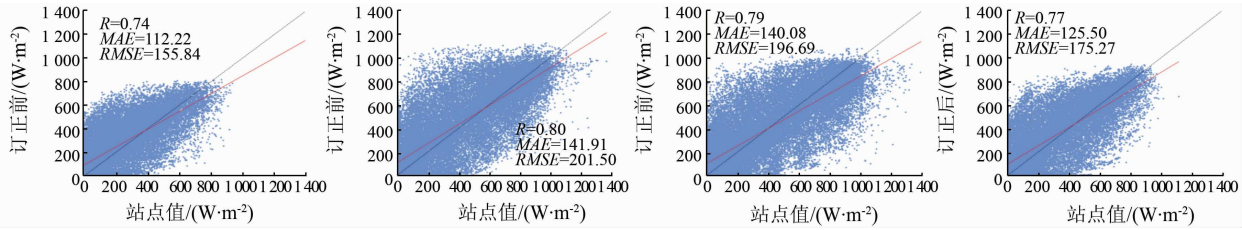
Fig. 1　Comparison of feature importance

2　结果与分析

2.1　订正前后的精度分析

图 2 为 1 月、4 月、7 月、10 月 ECMWF 再分析数据与地面站观测数据小时地表辐射的订正前（（a）—（d））后（（e）—（h））逐时辐射量散点

分布对比图。订正前，*MAE* 分别为 112.22 W/m^2，141.91 W/m^2，140.08 W/m^2 和 125.50 W/m^2，*RMSE* 分别为 155.84 W/m^2，201.50 W/m^2，196.69 W/m^2 和 175.27 W/m^2，*R* 分别为 0.74，0.80，0.79，0.77。结果表明不同月份的小时辐射数据误差不同且较大，1 月和 10 月的离散程度小，相关程度也较小，4 月和 7 月的离散程度大，相关程度也较大。订正后，各月份的订正值与站点值的离散程度减小，相关性明显提高，*MAE* 分别为 47.99 W/m^2，78.77 W/m^2，96.44 W/m^2 和 58.38 W/m^2，*RMSE* 分别为 87.90 W/m^2，133.53 W/m^2，160.59 W/m^2 和 102.29 W/m^2，*R* 分别为 0.91，0.91，0.88 和 0.92；1 月、4 月、7 月、10 月的各误差指标变化幅度不同，*MAE* 分别降低了 57.24%，44.49%，31.15% 和 53.48%，*RMSE* 分别降低了 43.60%，33.73%，18.35% 和 41.64%，*R* 分别提高了 0.17，0.11，0.09 和 0.15，可见 4 个月中 1 月的 ERA5 地表太阳辐射值订正效果最好。



（a）1 月订正前 （b）4 月订正前 （c）7 月订正前 （d）10 月订正前

（e）1 月订正后 （f）4 月订正后 （g）7 月订正后 （h）10 月订正后

图 2 订正前后 4 个月逐时辐射量散点分布

Fig. 2 Scatter diagram of hourly radiation before and after revision in four months

2.2 月均辐射值变化

2013 年 ERA5 再分析资料与中国气象站点资料订正前后的月均辐射变化如图 3 所示。可以看出，订正前 ERA5 太阳辐射量的值较站点值偏高，总体规律都是夏秋辐射量高、春冬辐射量低。订正后的值接近站点实测值，误差较小。

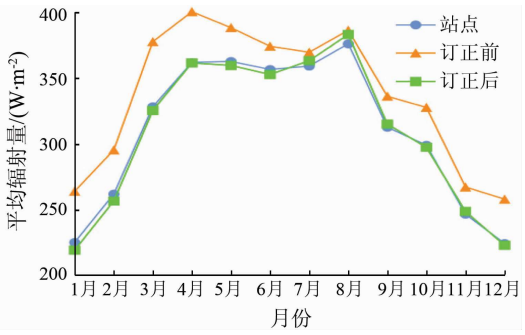


图 3 订正前后月辐射均值

Fig. 3 Monthly radiation mean value before and after revision

误差指标比较如表 3 所示，订正前的 *MAE*，*RMSE* 和 *R* 分别是 27.60 W/m^2，29.87 W/m^2 和 0.97，订正后三者值分别为 3.34 W/m^2，3.85 W/m^2 和 1.00。*MAE* 下降了 87.90%，*RMSE* 下降了 87.11%，*R* 提高了 0.03。

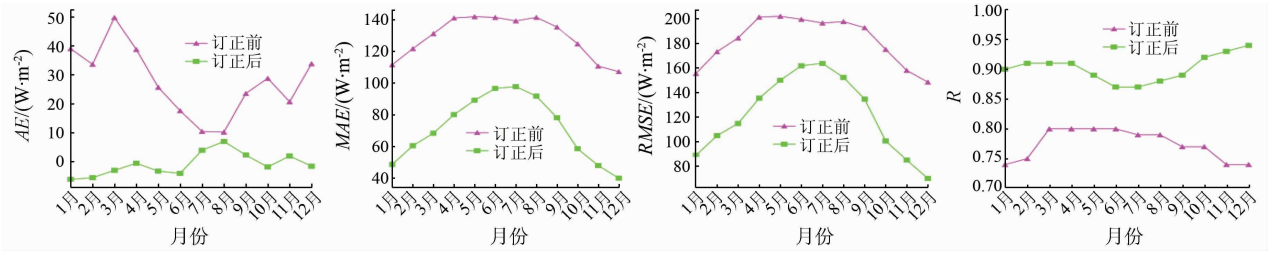
表 3 3 种误差指标比较

Tab. 3 Comparison of three error indexes

误差指标	订正前	订正后
MAE/(W · m^{-2})	27.60	3.34
RMSE/(W · m^{-2})	29.87	3.85
R	0.97	1.00

2.3 订正前后分月的误差分布规律

图 4(a) 为 ERA5 再分析资料与中国地面站点辐射量的月均值绝对误差比较，图 4(b)—(d) 为对每个月的小时数据求 *MAE*，*RMSE* 和 *R* 的分月误差比较。得出的结果是，在订正前，ERA5 数据与地面站点的 *AE* 在 10.28 ~ 49.53 W/m^2 之间，且夏秋季绝对误差小，春冬季绝对误差大；*MAE* 在 107.80 ~ 142.75 W/m^2 之间，*RMSE* 在 148.85 ~ 202.15 W/m^2 之间，*R* 在 0.74 ~ 0.80 之间；订正后，*AE* 在 −5.91 ~ 7.08 W/m^2 之间，*MAE* 在 40.00 ~ 98.31 W/m^2 之间，*RMSE* 在 70.98 ~ 164.07 W/m^2 之间，*R* 在 0.87 ~ 0.94 之间。此结果说明：随机森林模型对 ERA5 地表太阳辐射量的订正效果较好；*MAE* 和 *RMSE* 随着时间的变化也有所规律，明显看出夏秋季 2 种误差指标较大，春冬季较小。



（a）*AE*　（b）*MAE*　（c）*RMSE*　（d）*R*

图 4　订正前后的 *AE*，*MAE*，*RMSE* 和 *R* 比较

Fig. 4　Comparison of *AE*，*MAE*，*RMSE* and *R* before and after revision

2.4　简单交叉验证

为了进一步验证随机森林模型的稳定性，在中国范围内采用均匀分布的方法选取 6 个站点（表 4），分别为阿克苏、刚察、长春、绵阳、建瓯和北海，将其 2013 年所有样本数据作为模型的验证数据集，不参与训练，其余站点的样本数据作为训练数据集，进行训练，订正结果的比较如表 5 所示。从表 5 中看出，对于本身离散程度大、相关性弱的站点数据，经过随机森林的订正后精度有明显的提高，而本身离散程度小、相关性较高的站点数据，经过随机森林订正后能够保持精度或者有小幅度的提高。说明随机森林的模拟精度高，有较好的稳定性。

表 4　6 个站点的信息

Tab. 4　Information of six stations

站号	站点	纬度/(°)	经度/(°)
51628	阿克苏	N41.17	E80.23
52754	刚察	N37.33	E100.13
54161	长春	N43.90	E125.22
56196	绵阳	N31.45	E104.75
58737	建瓯	N27.05	E118.32
59644	北海	N21.45	E109.13

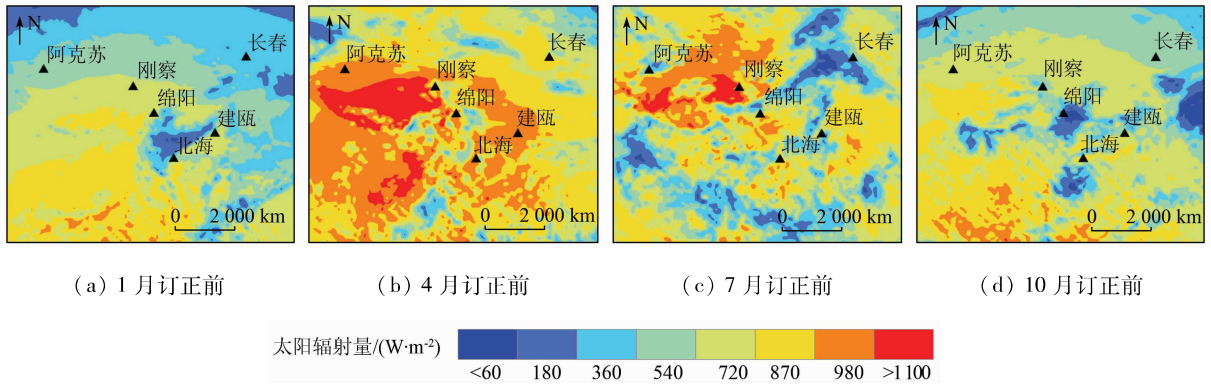
表 5　6 个站点前后订正的误差指标分析

Tab. 5　Analysis of error indexes in six stations before and after revision

站点	订正前			订正后			*MAE* 改善率/%	*RMSE* 改善率/%	*R* 提高量
	MAE	*RMSE*	*R*	*MAE*	*RMSE*	*R*			
阿克苏	237.83	288.28	0.50	62.49	96.82	0.94	73.72	66.42	0.44
刚察	170.13	208.48	0.75	74.80	117.00	0.92	56.03	43.88	0.17
长春	65.40	97.32	0.93	59.50	96.12	0.93	9.01	1.23	0
绵阳	109.09	148.87	0.83	74.47	113.13	0.89	31.74	24.01	0.06
建瓯	74.66	120.87	0.90	72.25	112.71	0.92	3.23	6.75	0.02
北海	121.11	165.75	0.84	85.21	129.37	0.89	29.64	21.95	0.05

2.5　订正前后的空间分布变化

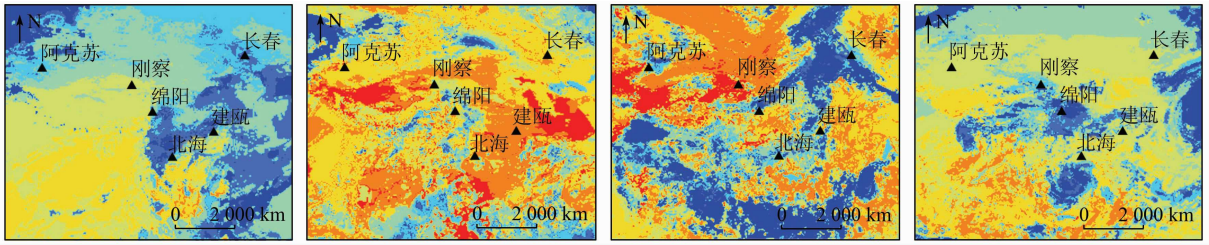
图 5 为利用上述建立的随机森林模型对 1 月、4 月、7 月、10 月的北京时间 15 日 13 时 ERA5 总辐射进行订正前后的空间分布图。图 5（a）—（d）和（e）—（h）分别为订正前后的空间分布结果。由图 5 可见，太阳辐射量订正前后的宏观分布规律一致，订正后 ERA5 太阳辐射量在局部地区有明显的下降，对 ERA5 太阳辐射量偏高的情况有所改进，通过随机森林订正后的分布图局部特征更加明显，精度得到提高。



（a）1 月订正前　（b）4 月订正前　（c）7 月订正前　（d）10 月订正前

图 5－1　订正前后太阳辐射的空间分布

Fig. 5－1　Spatial distribution of solar radiation before and after revision



(e) 1 月订正后　(f) 4 月订正后　(g) 7 月订正后　(h) 10 月订正后

图 5 – 2　订正前后太阳辐射的空间分布

Fig. 5 – 2　Spatial distribution of solar radiation before and after revision

3　结论与讨论

1)本文首先对 2013 年的再分析资料 ERA5 和地面观测的太阳总辐射数据进行了对比。从总体上看,两者有较大的差异,ERA5 的地表太阳辐射量要高于地面观测数据,这与前人研究一致; ERA5 辐射量与站点值的 *AE* 夏秋季小,春冬季大; 1 月和 10 月的离散程度小,相关程度也较小,4 月和 7 月的离散程度大,相关程度也较大。订正前,2013 年 *MAE*, *RMSE* 和 *R* 的值分别是 27.60 W/m^2,29.87 W/m^{2}和 0.97,对小时数据分月比较,ERA5 数据与地面站点的 *AE* 在 10.28 ~ 49.53 W/m^2 之间,且夏秋季 *AE* 小,春冬季 *AE* 大; *MAE* 在 107.80 ~ 142.75 W/m^2 之间,*RMSE* 在 148.85 ~ 202.15 W/m^2 之间,*R* 在 0.74 ~ 0.80 之间。

2)利用 5 折交叉验证和网格搜索选择模型参数,得到模型最优参数和交叉验证的模型得分并评价模型的稳定性,从得分可以验证模型的模拟较优,稳定性较好。将时间、经纬度、地表反照率、海拔、天顶角、水汽、云量、高云等作为输入参数进行随机森林训练。从 2013 年总体上看,*MAE* 下降了 24.26 W/m^2,*RMSE* 下降了 26.02 W/m^2,*R* 提高了0.03,说明随机森林回归模型取得了相对有效的订正结果。对小时数据处理并分月比较,订正后的 *AE* 在 – 5.91 ~ 7.08 W/m^2 之间,*MAE* 在 40.00 ~ 98.31 W/m^2 之间,*RMSE* 在 70.98 ~ 164.07 W/m^{2}之间,*R* 在 0.87 ~ 0.94 之间,订正后的离散程度减小,相关性明显提高。*MAE* 和 *RMSE* 随着时间的变化有所规律,明显看出夏秋季 2 种误差指标较大,春冬季较小。1 月、4 月、7 月、10 月的各误差指标增长幅度不同,4 个月中 1 月的 ERA5 地表太阳辐射量订正效果最好。

3)利用简单交叉验证,进一步验证随机森林模型的稳定性。阿克苏、刚察、长春、绵阳、建瓯、北海站点的太阳辐射量订正后的 *MAE*, *RMSE* 和 *R* 均有提高。结果表明对于本身离散程度大、相关性弱的站点数据,经过随机森林的订正后精度有明显的提高,而本身离散程度小、相关性较高的站点数据,经过随机森林订正后能够保持精度或者有小幅度的提高。说明随机森林的模拟精度高,有较好的稳定性。

4)通过随机森林的回归,对 ERA5 进行了空间分布上的订正,订正前后的宏观规律一致,订正后 ERA5 太阳辐射量在局部地区有明显的下降,精度得到提高。随机森林模型对 ERA5 地表太阳辐射量能够进行有效地订正,在实现大样本数据训练时能够保证速度,训练的模型精度较高,实现较为方便快捷,能够更好地进行太阳辐射产品的数据融合,得到的全国范围的连续的太阳辐射数据能够为农业、电力和城市建设等行业研究提供基础数据。

参考文献(References):

[1] 刘 钊,于学峰. 太阳驱动地球环境变化研究进展[J]. 自然杂志,2012,34(3):154 – 156,160.
Liu Z, Yu X F. The progress in study of the effects of the solar variation on the earth environment[J]. Chinese Journal of Nature, 2012,34(3):154 – 156,160.

[2] Roberto R, Renzo R. Distributed estimation of incoming direct solar radiation over a drainage basin[J]. Journal of Hydrology, 1995, 166(3):461 – 478.

[3] 赵 康,桂雪晨,葛 坚. 高大空间中太阳辐射对热舒适的影响及室内参数设计[J]. 太阳能学报,2019,40(9):2655 – 2662.
Zhao K, Gui X C, Ge J. Influence of solar radiation on thermal comfort in large spaces and corresponding design of indoor parameters[J]. Acta Energiae Solaris Sinica, 2019, 40(9):2655 – 2662.

[4] 张欣欣,景 丽. 太阳能热水系统在建筑给排水设计中的应用[J]. 智能建筑与智慧城市,2021(2):88 – 89.
Zhang X X, Jing L. Application of solar water heating system in the design of building water supply and drainage[J]. Intelligent Building and City Information, 2021(2):88 – 89.

[5] 张佳飞. 多时间尺度太阳辐射估算模型[D]. 重庆:西南大学, 2013.
Zhang J F. Estimation models of solar radiation at different time scales[D]. Chongqing: Xi'nan University, 2013.

[6] 吕 宁. 近年来中国地表太阳辐射时空变化及影响因素分析[D]. 北京:中国科学院地理科学与资源研究所,2009.
Lyu N. Analysis of spatio – temporal variation of surface downward

shortwave radiation and associated effecting factors over China in recent years[D]. Beijing: Institute of Geographic Sciences and Natural Resources Research, CAS, 2009.

[7] 蔡子颖. 我国华东、中南地区地面太阳辐射变化规律及其原因分析[D]. 南京:南京信息工程大学,2009.
Cai Z Y. Analysis on the variation of surface solar radiation in East China, Central South China and its causes[D]. Nanjing: Nanjing University of Information Science and Technology, 2009.

[8] 张星星,吕 宁,姚 凌,等. ECMWF 地表太阳辐射数据在我国的误差及成因分析[J]. 地球信息科学学报,2018,20(2):254 - 267.
Zhang X X, Lyu N, Yao L, et al. Error analysis of ECMWF surface solar radiation data in China[J]. Journal of Geo - Information Science, 2018, 20(2): 254 - 267.

[9] Zhang X T, Liang S L, Wang G X, et al. Evaluation of the reanalysis surface incident shortwave radiation products from NCEP, ECMWF, GSFC, and JMA using satellite and surface observations[J]. Remote Sensing, 2016, 8(3): 225 - 237.

[10] Wang A H, Zeng X B. Evaluation of multireanalysis products with in situ observations over the Tibetan Plateau[J]. Journal of Geophysical Research: Atmospheres, 2012, 117: D05102.

[11] 陈昱文,黄小猛,李 熠,等. 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,2020,31(4):494 - 503.
Chen Y W, Huang X M, Li Y, et al. Ensemble learning for bias correction of station temperature forecast based on ECMWF products[J]. Journal of Applied Meteorological Science, 2020, 31(4): 494 - 503.

[12] 李 净,温松楠. 基于 3 种机器学习法的太阳辐射模拟研究[J]. 遥感技术与应用,2020,35(3):615 - 622.
Li J, Wen S N. Simulation of solar radiation based on three machine learning methods[J]. Remote Sensing Technology and Application, 2020, 35(3): 615 - 622.

[13] Benali L, Notton G, Fouilloy A, et al. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components[J]. Renewable Energy, 2019, 132: 871 - 884.

[14] Yu W, Zhang X T, Hou N, et al. Estimation of surface downward shortwave radiation over China from AVHRR data based on four machine learning methods[J]. Solar Energy, 2019, 177: 32 - 46.

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

[16] 陈有龙,宁雨珂,唐荣年,等. 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),2020,38(4):356 - 364.
Chen Y L, Ning Y K, Tang R N, et al. Tropical temperature correction for numerical forecast in Hainan based on spatiotemporal independence random forest model[J]. Natural Science Journal of Hainan University, 2020, 38(4): 356 - 364.

[17] 芦 华,谢 旻,吴 钲,等. 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,2020,40(12):4419 - 4431.
Lu H, Xie M, Wu Z, et al. Adjusting $PM_{2.5}$ prediction of the numerical air quality forecast model based on machine learning methods in Chengyu region[J]. Acta Scientiae Circumstantiae, 2020, 40(12): 4419 - 4431.

[18] Babar B, Luppino L T, BostrÖm T, et al. Random forest regression for improved mapping of solar irradiance at high latitudes[J]. Solar Energy, 2019, 198: 81 - 92.

Revision of solar radiation product ERA5 based on random forest algorithm

WANG Xuejie[1], SHI Guoping[2], ZHOU Ziqin[1], ZHEN Yang[1]

(1. *Changwang School of Honors, Nanjing University of Information Science and Technology, Nanjing* 210044, *China*; 2. *School of Geographical Sciences, Nanjing University of Information Science and Technology, Nanjing* 210044, *China*)

Abstract: This study performed a multi - scale error analysis of the mean surface downward shortwave radiation flux product ERA5 (0.25° × 0.25°) of the European Centre for Medium - Range Weather Forecasts (ECMWF) using 93 pieces of solar radiation hourly data in 2013 of China. Subsequently, this study revised and analyzed the total radiation product ERA5 by training the random forest model using various relevant elements such as meteorological and geographic ones. Finally, the model was used to obtain the map of revised hourly radiation spatial distribution. As a result, the reanalyzed data can be better applied in industries such as agriculture, electric power, and urban construction. The results are as follows. ① The *MAE*, *RMSE*, and *R* values between the ERA5 solar radiation and the measured values of stations in 2013 were 27.60 W/m^2, 29.87 W/m^2, and 0.97 respectively. Moreover, the ERA5 values were higher than the measured values of stations. ② The accuracy was improved after the revision using the random forest model. After revision, the *MAE*, *RMSE*, and *R* values between the ERA5 solar radiation and the measured values of stations were 3.34 W/m^2, 3.85 W/m^2, and 1.00, respectively, indicating that correlation was significantly improved. ③ The spatial macroscopic distribution patterns of radiation before and after revision were consistent, but the ERA5 radiation value significantly decreased in local areas.

Keywords: ERA5 products; solar radiation; random forest revision

(责任编辑: 陈 理)