基于随机森林和 MODIS 产品的逐日太阳辐射估算

温松楠,李 净 (西北师范大学地理与环境科学学院,甘肃兰州 730070)

摘要 由于基于站点数据的太阳辐射研究很难获得空间连续分布的日太阳辐射数据,因此引入 MODIS 遥感数据结合随机森林来获取空间连续分布的日太阳辐射数据。选取西北地区 2015 年 6 月 26 日、7 月 12 日、7 月 28 日、8 月 13 日、8 月 29 日、9 月 30 日共 6 d 的 MODIS 遥感数据和 25 个辐射站点的日太阳辐射数据,将 MODIS 的产品云量、地面反照率、云光学厚度、水汽、DEM 作为随机森林的输入参数,随机选取 20 个辐射站点的实测数据和对应输入参数的遥感数据作为随机森林的训练数据集,用其余 5 个辐射站点的日太阳辐射实测数据对随机森林模拟结果进行验证。验证结果表明 5 个站点的模拟效果都较好,说明遥感数据结合随机森林模型能够很好地获得空间连续分布的日太阳辐射数据。

关键词 日太阳辐射;MODIS 产品;随机森林;西北地区
中图分类号 P 422.1 **文献标识码** A
文章编号 0517-6611(2020)02-0006-04
doi:10.3969/j.issn.0517-6611.2020.02.002
开放科学(资源服务)标识码(OSID):

Simulation of Daily Solar Radiation Based on Random Forest and MODIS Products

WEN Song-nan, LI Jing (College of Geography and Environmental Science, Northwest Normal University, Lanzhou, Gansu 730070)

Abstract The research about solar radiation based on ground sites is difficult to obtain daily solar radiation data with continuous spatial distribution. Therefore, this paper introduced MODIS remote sensing data combined with random forest to obtain spatially distributed daily solar radiation data. This paper selected MODIS remote sensing data of 6 days on June 26, July 12, July 28, August 13, August 29, and September 30, 2015 and daily solar radiation data of 25 radiation sites in the northwest China. MODIS products including cloud fraction, surface albedo, cloud optical thickness, water vapor, DEM were used as input parameters of random forest, and the measured data of 20 radiation sites and corresponding remote sensing data input parameters were randomly selected. Remote sensing data was used as a training data of random forest, and the results of random forest simulations were verified using the measured data of solar radiation from the remaining five radiation sites. The verification results show that the simulation results of five stations are better, indicating that combining with remote sensing data and random forest model to simulate daily solar radiation data of spatially continuous distribution is a reasonable and effective way.

Key words Daily solar radiation; MODIS product; Random forest; Northwest China

太阳辐射是地气系统的能量来源,也是产生大气运动的主要动力[1-2],同时,太阳辐射数据是农作物模型、水文模型及气候变化模型等的重要参数[3-4]。虽然在局部区域可以通过辐射观测站准确地测量太阳辐射,但是由于太阳辐射站点分布稀疏,很难获得连续空间分布的太阳辐射[5],因此,模拟空间连续分布的太阳辐射对区域气候变化带来的影响具有重要的意义。

目前国内对于太阳辐射模拟的研究大多数是估算月或年的太阳辐射[6-8],日太阳辐射的模拟研究较少,国外已有少量模拟日太阳辐射的研究:Marzo 等[9]利用每日最低温度、最高温度和地外辐射,采用人工神经网络估算了全球 13 个沙漠地区的每日太阳辐射;Hassan 等[10]用 12 个不同的经验系数独立模型估算了每日太阳辐射;Yildir 等[11]用 ANN 模型和 Angström-Prescott 模型估算了土耳其东地中海地区每日太阳辐射;Jahani 等[12]用 4 种经验模型估算了伊朗的日太阳辐射;Kaba 等[13]选取地外辐射、日照时数、云量、最低温度和最高温度作为输入数据,利用深入学习方法估算了土耳其 30 个站点的日太阳辐射,深入学习模型在土耳其日太阳辐射模拟时取得了很好的效果,但这些研究都是基于站点数据估算日太阳辐射,很难获得空间连续分布的日太阳辐射。

基金项目 国家自然科学基金项目(41261016,41761083,41561016);西北师范大学青年教师科研能力提升计划项目(NWNU-LKQN-14-4)。
作者简介 温松楠(1995—),女,甘肃天水人,硕士研究生,研究方向:定量遥感。
收稿日期 2019-07-25;**修回日期** 2019-08-12

由于遥感数据可以很好地用于模拟空间连续分布的太阳辐射[14-16],再加上空间连续分布的日太阳辐射是水文模型、农作物模型及气候变化模型等的重要参数,因此该研究在前人日太阳辐射模拟的基础上,引入遥感手段来模拟空间连续分布的日太阳辐射。由于随机森林模型对于太阳辐射有很好的模拟效果,所以笔者以西北地区为研究区,利用 MODIS 遥感数据和随机森林来模拟获得空间连续分布的日太阳辐射。

1 数据来源与研究方法

1.1 数据来源 选取西北地区 25 个辐射站点(图 1)的太阳日辐射数据,数据来源于中国气象数据网(http://data.cma.cn/),主要用于随机森林模拟太阳辐射时的训练数据和太阳辐射模拟结果的验证。选取的遥感数据 MODIS 产品来源于 NASA 官网,所用数据如表 1 所示。

1.2 研究方法

1.2.1 随机森林。随机森林是 2001 年由 Leo Breiman 和 Cutler Adele 开发的一种数据挖掘方法,是一种现代分类与回归的机器学习技术,同时也是一种组合式的机器学习技术[17]。与人工神经网络、支持向量机等机器学习方法相比,随机森林算法具有运算量小、容纳样本数量大等优点。随机森林的基本组成单元是决策树,其优越性体现在同等运算率下的高预测精度,对非线性数据有更好的拟合效果,减少了均方根误差、提高了模型的预估精度[18]。

通过 Python 中的 Pandas 准备数据框数据,导入 Sklearn 工具包,在 Sklearn 模块库中,与随机森林算法相关的函数都

位于集成算法模块 Ensemble 中,利用一系列运算代码实现随机森林模型预测太阳辐射。选取与太阳辐射有关的因子云量、地面反照率、云光学厚度、水汽、DEM 为自变量,因变量为日太阳辐射。随机森林具体算法步骤如下[19-20]:①用 Bootstrap 法在 N 个总样本中有放回地随机抽取 n 次,得到 n 个自助样本集作为训练集,未抽取的部分组成袋外数据。Bagging 是早期组合树方法之一,又称自助聚集(Bootstrap Aggregating),是一种从训练集中随机抽取部分样本(不一定有放回抽样)来生成决策树的方法。②将每个训练集都单独作为一棵决策树,决策树节点从自变量中选择 M 个(M 小于自变量个数),并按照节点不纯洁度最小原则进行分支生长。③重复步骤②n 次,得到 n 棵决策树组成随机森林。对于每一棵决策树,都可以得到一个 OOB 误差估计,将森林中所有决策树的 OOB 误差估计取平均,可得到随机森林的泛化误差估计。

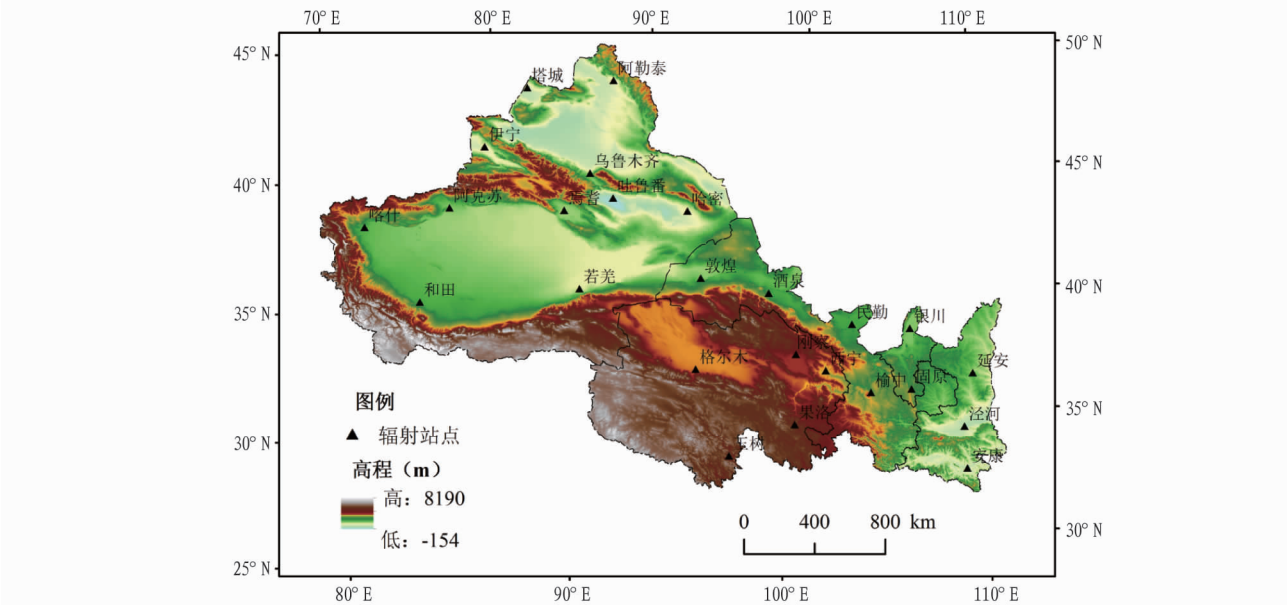

图 1　西北地区辐射站点

Fig.1　Radiation sites in northwest China

表 1　所用 MODIS 数据

Table 1　MODIS data used in this paper

MODIS 产品 MODIS products	数据 Data	时间分辨率 Time resolution	空间分辨率 Spatial resolution
MOD08-D3	云量、云光学厚度	每天	1°
MOD09GA	地面反照率	每天	1 km
MOD09CMA	水汽	每天	0.05°

1.2.2　精度评价指标。采用相关系数(R)、平均偏差(MBE)、平均绝对偏差(MABE)、均方根误差(RMSE)这 4 种精度评价指标对模型结果进行验证[21]。

$$R=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

$$\mathrm{MBE}=\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i) \tag{2}$$

$$\mathrm{MABE}=\frac{1}{n}\sum_{i=1}^{n}|x_i-y_i| \tag{3}$$

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

式(1)、(2)、(3)、(4)中,X_i 代表第 i 个模拟值;y_i 代表第 i 个实测值;$\bar{x}$ 和 $\bar{y}$ 分别表示模拟值和实测值的平均值;n 为样本个数。R 为 x_i 和 y_i 的相关系数,R 越大,模拟值与实测值的相关程度越高;MBE、MABE 和 RMSE 越小,表示模拟值越接近观测值。

1.2.3　太阳辐射模拟参数的选取。MODIS 能够提供的大气和陆地日产品有气温、云量、水汽、气溶胶、地表温度、地面反照率、植被指数,该研究需要获得空间连续的日太阳辐射分布,而 MODIS 提供的日气溶胶空间连续性很差,日气温与日太阳辐射相关性较低,最终选取与太阳辐射相关性较高且空间连续性好的云量、地面反照率、云光学厚度、水汽以及 DEM 共 5 种参数来模拟逐日太阳辐射。

2　日太阳辐射模拟

2.1　日太阳辐射模拟结果验证　考虑到遥感数据的完整性和质量好坏,选择 2015 年 6 月 26 日、7 月 12 日、7 月 28 日、8 月 13 日、8 月 29 日、9 月 30 日共 6 d 的数据,随机选取西北地区 20 个辐射站点的实测数据和对应输入参数的遥感数据作为随机森林训练的数据集,将太阳辐射日辐射作为随机森林的输出,从而模拟得到每日的太阳辐射,最后用其余 5 个辐射站点的日太阳辐射实测数据对随机森林模拟结果进行验证,结果如表 2 所示。

从站点验证结果来看,5 个验证站点的相关系数都大于 0.89,哈密站点和泾河站点实测值与模拟值的相关系数最大,达 0.98,吐鲁番站点实测值与模拟值的相关系数最小,为0.89;5 个验证站点的平均偏差(MBE)控制在-2.5~2.0MJ/(m^2 · d)波

动,平均偏差为负值表示随机森林模型的低估,正值表示随机森林模型的高估,固原站点对太阳辐射有轻幅度低估,其余 4 个站点的模拟值稍有高估;5 个站点的平均绝对偏差(MABE)都控制在 3.5 MJ/(m^2 · d)以内,哈密站点的平均绝对偏差最小,吐鲁番站点的平均绝对偏差最大;5 个站点的均方根误差(RMSE)都控制在 4.5 MJ/(m^2 · d)以内,哈密站点的均方根误差最小,吐鲁番站点的均方根误差最大。总体上,5 个站点的模拟效果都较好,其中哈密站点的模拟效果最好,吐鲁番站点模拟效果一般,模拟效果一般的原因是 6 月 26 日输入遥感参数中云量偏高导致太阳辐射的低估,9 月 30 日输入遥感参数中云量偏低造成太阳辐射的高估,总体上,吐鲁番站点模拟的太阳辐射有 1.25 MJ/(m^2 · d)的轻幅高估。

表 2 站点结果验证

Table 2 Result verification of site

辐射站点 Radiation site	R	MBE MJ/(m^2 · d)	MABE MJ/(m^2 · d)	RMSE MJ/(m^2 · d)
伊宁 Yining	0.89	0.94	2.97	3.54
吐鲁番 Turpan	0.90	1.25	3.38	4.15
哈密 Hami	0.98	0.75	1.52	2.10
固原 Guyuan	0.96	−2.11	3.01	3.57
泾河 Jinghe	0.98	1.63	2.08	2.67

5 个验证站点 6 d 的实测值与模拟值的散点图如图 2 所示,R 为 0.92,说明 5 个辐射站点的模拟值和实测值非常接近,模拟效果较好。

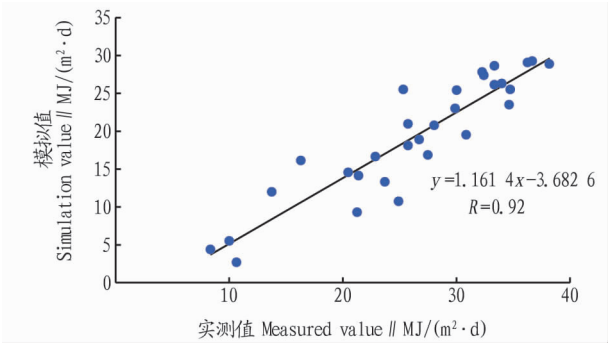

图 2 验证站点的散点图

Fig.2 Scatter plot of verification sites

2.2 日太阳辐射模拟 采用随机森林算法模拟的日太阳辐射结果如图3所示,2015年6月26日新疆与西安太阳辐射

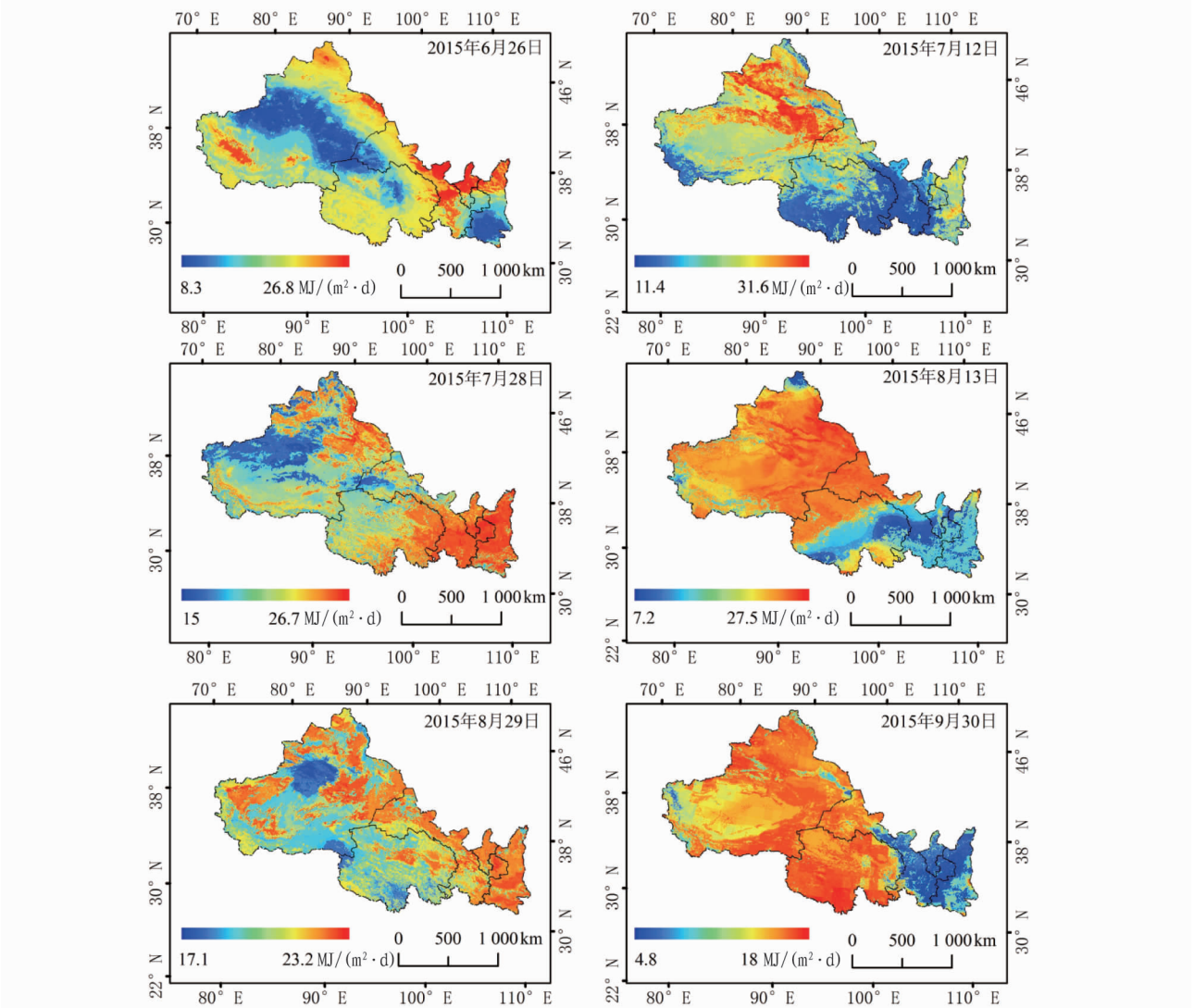

图 3 日太阳辐射模拟结果

Fig.3 Daily solar radiation simulation results

偏低,在甘肃、宁夏太阳辐射较高,2015 年 7 月 12 日、8 月 13 日和 9 月 30 日太阳辐射的空间分布很相似,在西北部较高,东南部较低,2015 年 7 月 28 日和 8 月 29 日太阳辐射的空间分布很接近,在新疆局部地区较低,其余地区较高,日太阳辐射的增加或降低受天气状况影响较大,主要是由云量多少以及水汽含量的多少共同导致的。

3　结论

该研究基于 MODIS 遥感数据和随机森林模型估算了西北地区 2015 年 6 月 26 日、7 月 12 日、7 月 28 日、8 月 13 日、8 月 29 日、9 月 30 日共 6 d 的日太阳辐射,选取 MODIS 的云量、地面反照率、云光学厚度、水汽以及 DEM 作为随机森林的输入参数,选取西北地区 20 个辐射站点的实测数据和对应输入参数的遥感数据作为随机森林模型训练的数据集,用其余 5 个辐射站点的日太阳辐射实测数据对随机森林模拟结果进行验证,最后得到空间连续分布的日太阳辐射数据,研究得出以下主要结论。

(1)站点验证结果表明,5 个验证站点的相关系数都大于 0.89,说明模拟值和实测值相关程度较高;5 个验证站点的平均偏差(MBE)都控制在-2.5~2.0 MJ/(m^2 · d);平均绝对偏差(MABE)都控制在 3.5 MJ/(m^2 · d)以内;均方根误差(RMSE)都控制在 4.5 MJ/(m^2 · d),总体上伊宁、吐鲁番、哈密、固原、泾河 5 个验证站点的模拟效果都较好。

(2)基于随机森林算法,选取 MODIS 遥感数据云量、地面反照率、云光学厚度、水汽以及 DEM 作为太阳辐射的影响因子,可以用于模拟日太阳辐射且模拟效果较好。

(3)利用遥感数据结合随机森林模型可以很好地模拟日太阳辐射,能够得到空间连续的、高分辨率的逐日太阳辐射数据。

参考文献

[1] SUN H W,ZHAO N,ZENG X F,et al.Study of solar radiation prediction and modeling of relationships between solar radiation and meteorological variables[J].Energy conversion and management,2015,105:880-890.

[2] CHEN J L,LI G S,WU S J.Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration[J].Energy convers manage,2013,75:311-318.

[3] 施国萍,邱新法,曾燕.中国三种太阳辐射起始数据分布式模拟[J].地理科学,2013,33(4):385-392.

[4] 黎微微,胡斯勒图,陈洪滨,等.利用 MODIS 资料计算不同云天条件下的地表太阳辐射[J].遥感技术与应用,2017,32(4):643-650.

[5] JOURNÉE M,BERTRAND C.Improving the spatio-temporal distribution of surface solar radiation data by merging ground and satellite measurements [J].Remote sensing of environment,2010,114(11):2692-2704.

[6] 刘剑,曹美燕,高治军,等.一种基于随机森林的太阳能辐射预测模型[J].控制工程,2017,24(12):2472-2477.

[7] 张春桂,文明章.利用卫星资料估算福建晴空太阳辐射[J].自然资源学报,2014,29(9):1496-1507.

[8] 罗悦,俞文政,袁真艳.淮北平原太阳总辐射的估算及时空特征分析[J].长江流域资源与环境,2018,27(5):1031-1042.

[9] MARZO A,TRIGO-GONZALEZ M,ALONSO-MONTESINOS J,et al.Daily global solar radiation estimation in desert areas using daily extreme temperatures and extraterrestrial radiation[J].Renewable energy,2017,113:303-311.

[10] HASSAN M A,KHALIL A,KASEB S,et al.Independent models for estimation of daily global solar radiation:A review and a case study[J].Renewable and sustainable energy reviews,2018,82:1565-1575.

[11] YILDIR M H B,ÇELIK Ö,TEKE A,et al.Estimating daily Global solar radiation with graphical user interface in Eastern Mediterranean region of Turkey[J].Renewable and sustainable energy reviews,2018,82:1528-1537.

[12] JAHANI B,DINPASHOH Y,NAFCHI A R.Evaluation and development of empirical models for estimating daily solar radiation[J].Renewable and sustainable energy reviews,2017,73:878-891.

[13] KABA K,SARIGÜL M,AVCI M,et al.Estimation of daily global solar radiation using deep learning model[J].Energy,2018,162:126-135.

[14] EROL A Ö ,FILIK Ü B.Estimation methods of global solar radiation,cell temperature and solar power forecasting:A review and case study in Eski Ṣ ehir[J].Renewable and sustainable energy reviews,2018,91:639-653.

[15] FALLAHI S,AMANOLLAHI J,TZANIS C G,et al.Estimating solar radiation using NOAA/AVHRR and ground measurement data[J].Atmospheric research,2018,199:93-102.

[16] YAO W X,ZHANG C X,HAO H D,et al.A support vector machine approach to estimate global solar radiation with the influence of fog and haze [J].Renewable energy,2018,128:155-162.

[17] BREIMAN L.Random forests[J].Machine learning,2001,45(1):5-32.

[18] 华俊玮,祝善友,张桂欣.基于随机森林算法的地表温度降尺度研究[J].国土资源遥感,2018,30(1):78-86.

[19] BREIMAN L.Bagging predictors[J].Machine learning,1996,24(2):123-140.

[20] 方匡南,吴见彬,朱建平,等.随机森林方法研究综述[J].统计与信息论坛,2011,26(3):32-38.

[21] QUEJ V H,ALMOROX J,IBRAKHIMOV M,et al.Empirical models for estimating daily global solar radiation in Yucatán Peninsula,Mexico[J].Energy conversion and management,2016,110:448-456.