引 用 格 式:Gao Shuai,Hou Xuehui,Wang Yun,*et al*.Remote Sensing Monitoring of Terrestrial Ecosystem Carbon Budget based on Machine Learning and Big Data Platform[J].Remote Sensing Technology and Application,2022,37(5):1190-1197.[高帅,侯学会,汪云,等.基于机器学习和大数据平台的陆地生态系统碳收支遥感监测[J].遥感技术与应用,2022,37(5):1190-1197.]
DOI:10.11873/j.issn.1004-0323.2022.5.1190

基于机器学习和大数据平台的陆地生态系统碳收支遥感监测

高 帅[1],侯学会[2],汪 云[3],王 倩[4],陈 悦[1],邢 瑞[1],王 晶[1,5]
(1.中国科学院 空天信息创新研究院 遥感科学国家重点实验室,北京 100101;
2.山东省农业科学院 农业信息与经济研究所,山东 济南 250100;
3.北京林业大学 园林学院,北京 100083; 4.天津师范大学 地理与环境科学学院,天津 300387;
5.中国地质大学地球科学与资源学院,北京 100083)

摘要:陆地生态系统碳收支是全球碳循环研究的重要指标,也是气候变化的重要参数。针对该指标估测的不确定性,基于陆地生态系统通量观测研究网络的实测碳通量数据及遥感卫星观测数据产品,利用机器学习方法进行建模研究。研究选用随机森林算法自动从高质量的星—地训练数据集中学习特征、挖掘数据中的隐含信息以及时序间依赖关系的差异,建立了基于随机森林算法的碳收支参数GPP(Gross Primary Production)、NEP(Net Ecosystem Production)估算模型,并选择标准指标利用验证数据集对模型进行了客观评价。结果分析表明:与MODIS GPP产品相比,该方法在估算精度上有了提高,其中落叶阔叶林预测结果最优,决策系数为R^2为0.82,均方根误差为1.93 $gCm^{-2}d^{-1}$,在其他植被类型上也明显优于传统光能利用率模型产品,更接近于地面通量观测数据。基于相同方法建立的NEP模型也得到了较好的估测结果,落叶阔叶林预测模型的输出结果与通量塔获得的NEP相关关系R^2为0.70,RMSE=1.75 g C $m^{-2}d^{-1}$。GPP和NEP模型精度差异也表明,在进行机器学习建模时,训练数据集自变量的选择仍然需要机理模型支持。为进行陆地生态系统碳收支大范围快速估算,本研究进行了陆地生态系统碳收支遥感监测平台的搭建,该平台以GEE(Google Earth Engine)大数据平台作为数据存储与计算后端,Django和Nginx作为Web服务框架,OpenLayers和jQuery作为前端框架,从而实现了碳收支参数长时间序列大范围的快速计算、结果实时显示等功能。基于该平台和模型获取的2002—2016年全球(60°N—60°S)逐年GPP结果表明,全球平均GPP存在明显的空间差异,显著增加的区域主要集中在亚洲东部地区及北美洲森林地区等。研究表明,基于机器学习和大数据平台进行碳收支参数遥感监测,能够快速提供与地面真实观测较为一致的陆地生态系统区域和全球尺度碳收支遥感监测结果,该流程在一定程度避免了生理过程模型复杂的参数设置,减少了区域和全球大尺度碳收支监测的不确定性。

关 键 词:机器学习;大数据平台;碳收支;随机森林;时空扩展
中图分类号:X16;X87 **文献标志码**:A **文章编号**:1004-0323(2022)05-1190-08

1 前 言

近年来随着工业革命的发展,各种化石燃料的燃烧释放了大量二氧化碳,引起了温室效应等一系列全球气候问题[1],面对这一问题,各国政府经过艰难协商在2016年签订了《巴黎协定》(The Paris

收稿日期:2021-08-17;**修订日期**:2022-07-07
基金项目:国家重点研发计划项目(2017YFA0603004),国家自然科学基金项目(42171377),高分项目(30-Y20A15-9003-17/18),天津市高等学校科技发展计划项目(2018KJ154)。
作者简介:高帅(1983—),男,山东高密人,副研究员,主要从事数据挖掘和激光雷达研究。E-mail:gaoshuai@aircas.ac.cn

Agreement),为了履行协定条款,需要对全球生态系统碳循环进行准确快速的监测。全球生态系统碳循环包括陆地、海洋、大气3个部分,陆地生态系统碳储量大约为大气碳储量的3倍[2],同时也是碳循环不确定性的主要来源[3],其中植被总初级生产力GPP(Gross Primary Production)是陆地生态系统最主要的固碳参量,而其他指标,例如生态系统呼吸ER(Ecosystem Respiration)是最重要的碳损耗量,净生态系统生产力NEP(Net Ecosystem Production)能够量化碳汇的大小[4-7],因此,要厘清陆地生态系统碳循环的总体状况,必须对上述指标进行准确监测。

目前,通过涡度相关技术在地面可以获得测量准确、时间连续的GPP、ER、NEP等参数,但是这类地面通量站点全球分布有限,难以实现区域或者全球大尺度空间范围的观测[8-11]。目前空间大尺度估算这些参量主要有两种方法:生理过程模型和卫星遥感模型。基于生理过程的模型可以动态模拟植被生理过程[12],以GPP为例,包括BIOME-BGC模型[13](Biome-BioGeochemical Cycles)、InTEC模型[14](Integrated Terrestrial Ecosystem C-budget)和LPJ-DGVM(Lund-Potsdam-Jena Dynamic Global Vegetation)模型[15]等,这类过程模型具有较强机理性,但过程复杂,关键参数依赖经验设置。基于卫星遥感的模型通常具有较高的空间分辨率,包括TG(Temperature and Greenness)模型[16]、GR(Greenness and Radiation)模型[17]和VPM(Vegetation Photosynthesis)模型[18]等,然而这些模型在全球大范围估算时往往具有较大的不确定性[19]。

因此,如何将地面站点准确的通量观测与大范围的遥感数据空间覆盖相结合,实现大数据驱动的碳收支参数大范围准确估算,具有极大的现实需求。近年来,大数据技术不断发展,机器学习模型,例如支持向量机[20](SVM)、模型树集合[21](MTE)和随机森林回归[22](RFR)模型等,可以从观测样本出发寻找规律,建立模型进行预测,已经被广泛应用在生态学研究中。另一方面,当前对地观测卫星遥感数据呈现海量爆发式增长,原有的单机或服务器批处理的数据处理方式已经不能满足需要。遥感图像具有非结构化,数据量大的特点,适合利用分布式平台进行存储和计算,因此对云计算(cloud computing)的需求应运而生。云计算是目前广泛采用的一种分布式计算方式,指的是通过网络将巨大的数据计算处理程序分解成无数个小程序,通过多部服务器组成的系统对数据进行动态易扩展处理,最后将这些小程序得到结果合并返回给用户[23]。目前,各个国家和私营大公司,也都开发了各种基于云计算的平台,例如美国谷歌公司(Google Inc.)就推出了GEE(Google Earth Engine)云计算平台,该平台存储了Landsat,MODIS,Sentinel-1和Sentinel-2等海量卫星数据集及其他数据,同时平台内置了多种大数据处理算法,集成了数据处理接口API(Application Programming Interface),可以快速、批量的处理海量的数据。用户通过GEE设计各种应用算法,可以预测作物相关产量,监测旱情长势变化,监测全球森林变化等[24-26]。

因此,研究将探索利用海量的遥感数据产品和地面真实观测数据,通过机器学习算法进行碳收支相关参数估算模型研究,同时结合大数据平台,实现碳收支大范围、长时间的快速计算,探索避免模式和经验模型的不确定性、实现区域和全球碳收支的遥感直接估算的策略,从而服务于碳循环时空变化与气候变化响应的科学认知。

2　方法与数据

研究基于卫星遥感观测数据、气象数据和地面站点通量观测数据,利用随机森林算法挖掘数据中的隐含特征以及时序间依赖关系,从而自动从高质量的训练数据中学习特征,建立基于数据驱动的陆地生态系统碳收支遥感监测站点模型,并选择指标对模型进行客观评价。基于模型,以大数据存储和计算平台为支撑进行陆地生态系统碳收支区域和全球尺度时空扩展,通过建立交互式网络接口,实现可定制、高时效的产品生产和共享,具体技术路线如图1所示。

2.1　数据

2.2.1　通量数据

研究所用的通量数据为覆盖全球范围的212个通量塔站点的FLUXNET2015数据集(http://fluxnet.fluxdata.org),包括实测碳通量数据和相关通量的衍生数据产品。这些站点覆盖的范围较广,涵盖较长的时间跨度,具有很强的利用价值和代表性。植被覆盖类型广泛,包括落叶阔叶林(Deciduous Broadleaf Forest,DBF)、常绿阔叶林(Evergreen Broadleaf Forest,EBF)、常绿针叶林(Evergreen Needleleaf Forest,ENF)、混合林(Mixed Forest,MF)、稀树

图 1 技术路线

Fig.1 Workflow of the research

灌丛(Open Shrubland,OSH)、草原(Grassland,GRA)、稀树草原(Savannah,SAV)、多树草原(Woody Savannas,WSA)、耕地(Cropland,CRO)、湿地(Wetland,WET)等13种类型。在研究中将获得的每半小时NEE数据和GPP、ER、NEP等数据产品,按照8 d的尺度进行积分累加,从而获得8 d的时间分辨率的值,以对应遥感数据的时间间隔[27-28]。

2.2.2 *遥感数据*

研究中使用的遥感数据包括增强型植被指数(EVI)(MCD43A4, Version 6)、陆地表面温度(LST, ℃)(Version 6,MOD11A2)、(Global Land Data Assimilation System, GLDAS)Noah 2.7.1模型中短波辐射(SWR,W·m^{-2})产品[29]、NOAA/PERSIANN-CDR降水数据集[30](PREC,mm)等。遥感数据都取通量点以及周围的1 km×1 km区域,并且在相应区域内的平均值作为该站点的值,并将以上各种数据重采样成与EVI时间分辨率一致的8 d合成的分辨率。对于遥感观测数据,如果缺少全年的长序列数据,将这些数据标记为缺失,如果缺少短时间数据,将用无缺失日的平均值来代替。同时,在研究中使用MODIS GPP数据产品MOD17A2H与本方法监测结果进行对比分析。

2.2 随机森林模型

在综合比较各种模型的基础上,研究采用随机森林模型进行建模[30]。随机森林是一种基于无参数回归算法的集成学习方法,其主要思想为对原始的训练数据集T(公式(1)),从中随机抽样获得n个子样本集(公式(2)),再对每个子样本集分别建立决策树回归(CART,Classification and Regression Tree)模型$h_i(x)$,所有的决策树互不相关,经过训练后,得到一个回归树模型序列$\{h_1(x),h_2(x),\cdots,h_n(x)\}$,对任意给定的新样本,它的预测结果是对$n$个结果的平均汇总(公式(3))。当需要根据属性对一个对象进行分类时,随机森林中的每棵决策树都对其进行判断,最终随机森林将输出权重次数最多的分类项。

$$T=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\} \tag{1}$$

其中:(x_i,y_i)是一个训练样本;x_i为自变量;y_i为因变量。

$$y=\{c_1,c_2,...,c_n\} \tag{2}$$

c_i是第i个样本子集合。

$$f_r(x)=\frac{1}{n}\sum_{i=1}^{n}h_i(x) \tag{3}$$

其中:$f_r(x)$表示随机森林回归模型的结果;$h_i(x)$是单个回归树模型的结果。

在随机森林算法中,最初必须设定两个参数,分别是决策树中树的数量N和每个分类节点上进行分裂时要考虑使用的特征变量的个数m。N数值越大越好,但计算时间也会变长,通常N最佳参数值应始终在进行交叉验证时产生。在回归问题中,m默认值是变量总数的平方根[32]。由于本研究中特征量不多,且测试表明随着N和m增大,精度提升,因此本模型不限制树的数量和最大特征数。

在对每种植被类型进行随机森林模型建模时,首先将训练数据集和测试数据集按照约2∶1比例进行分割,数据集中每条数据中增强植被指数数据产品,陆地表面温度数据产品,短波辐射数据产品,降水数据产品为自变量x_i,相对应地面通量站点的GPP/NEP数据产品为因变量y_i。在训练中,使用十折交叉验证方法调整模型参数,当迭代产生的残差最小时的模型确定为预测模型。以图2所所示,当迭代次数为190次时,阔叶林GPP预测模型残差值达到最小值5.92。

3 碳收支参数遥感监测平台

本研究的碳收支参数遥感监测平台以GEE为核心数据存储和计算平台,基本架构如图3所示。平台后端使用Python调用GEE提供的API进行影

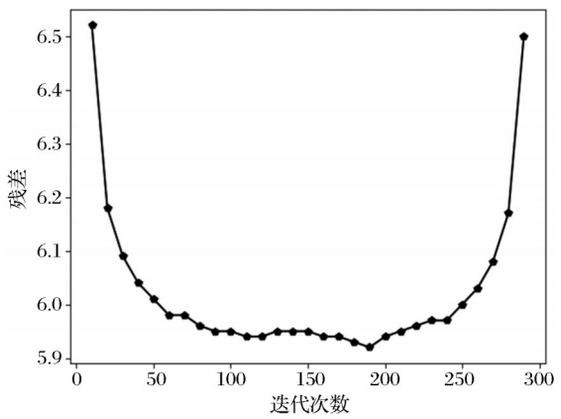

图 2　模型迭代次数与残差的关系

Fig.2　The relationship between iterations and residuals of the model

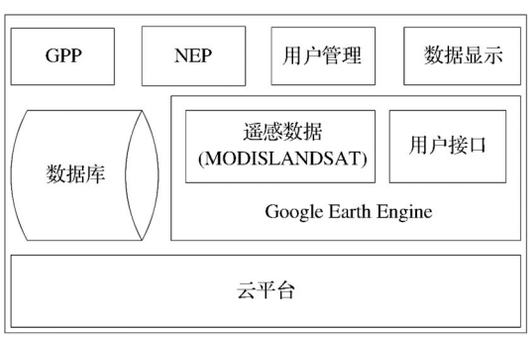

图 3　碳收支监测平台基本架构

Fig.3　Basic architecture of carbon budget monitoring platform

像处理,实现了复杂的叠加分析和空间分析及碳收支算法等功能。其中,服务端使用Django(https://www.djangoproject.com/)框架,以动态地响应前端请求,同时使用Nginx作为网页资源、Shapefile、GeoJSON等静态资源文件的响应服务器。数据库则选用了PostgreSQL(https://www.postgresql.org/)。平台前端使用OpenLayers作为地图加载、渲染、图形绘制框架,使用jQuery作为请求发起、页面元素获取、样式设置框架。基于以上技术,平台实现了碳收支参数数据的一体化组织管理、快速查询、计算和展示。

如图4所示,碳收支监测平台网页前端包括一个标题栏、一个状态栏、一组地图操作控件(放大与缩小)和一个控制面板。所有的数据查询参数设置均在控制面板上完成,数据的展示则在地图上叠加显示。核心部件“控制面板”,包括“环境参数”、“查询区域”、“查询日期”“图层管理”以及“查询”等部分。在“碳循环量详细参数设置窗口”中,可以对模型参数进行设置,既可以使用默认的参数,也可

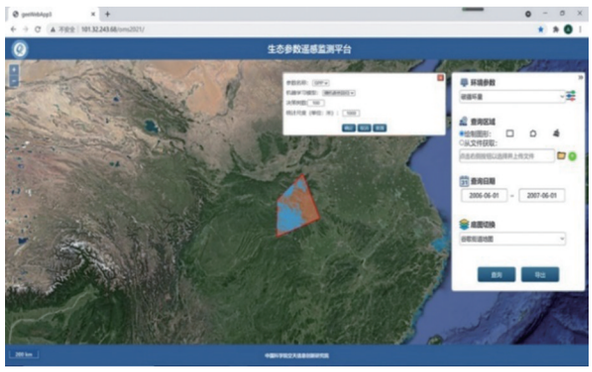

图 4　云平台 Web 应用程序整体界面

Fig.4　Cloud platform Web application overall interface

以打开详细参数设置窗口进行参数设置;可以在地图上通过绘制“矩形”或“多边形”的方式创建查询区域或者上传矢量文件到服务器,系统将根据区域图形进行计算。

4　结果与讨论

4.1　站点模型

研究基于随机森林模型对每种植被类型进行了GPP,NEP等碳收支参数的建模。在获得模型预测结果后,以R^2、RMSE为评价指标分别分析估测GPP/NEP的数值和未参与训练通量塔站点GPP/NEP数据的拟合关系,并与遥感数据产品开展比较。

以GPP为例,全球范围内土地覆盖类型是落叶阔叶林站点的个数为24个,从中随机选择了16个站点构建训练样本进行模型迭代,通过前述的随机森林算法确定了阔叶林GPP模型。模型预测所得GPP(GPP_RFR)与未参与训练的8个站点的GPP数值(GPP_EC)比较表明,两者总体相关性R^2为0.81,RMSE为2.02 gC m^{-2} d^{-1}。将8个站点的GPP_RFR值与MODIS GPP产品(GPP_MODIS)比较表明,在所有站点GPP_RFR的表现均优于GPP_MODIS。除了落叶阔叶林,其他植被类型的建模也表明GPP_RFR的预测结果较GPP_MODIS结果更接近于地面观测数据(表1)。对于稀树草原,尽管GPP_RFR与GPP_EC的相关关系R^2=0.43,但相对于GPP_MODIS数据产品(R^2=0.19)也有较大的改进。

类似地,基于随机森林模型对每种植被类型开展了NEP建模,结果表明,落叶阔叶林模型预测模型的输出结果更加靠近通量塔实测NEP数据产品,相关关系R^2为0.70,RMSE=1.75 g C m^{-2} d^{-1},常绿

表 1 GPP模型、MODIS GPP产品、NEP模型分别与通量塔站点对比

Table 1 The GPP model, MODIS GPP product and NEP model compared with flux tower sites

IGBP	GPP_RFR		GPP_MODIS		NEP_RFR	
	R^2	RMSE ($g\ C\ m^{-2}\ d^{-1}$)	R^2	RMSE ($g\ C\ m^{-2}\ d^{-1}$)	R^2	RMSE ($g\ C\ m^{-2}\ d^{-1}$)
DBF	0.81	2.02	0.69	2.69	0.70	1.75
GRA	0.78	1.77	0.6	2.51	0.37	1.44
WSA	0.78	1.12	0.48	1.72	0.41	1.03
OSH	0.71	0.64	0.53	1.17	0.34	0.76
CRO	0.69	3.01	0.41	4.47	0.55	2.37
ENF	0.68	1.92	0.61	2.29	0.35	1.67
MF	0.68	2.03	0.61	2.31	0.43	1.64
WET	0.61	2.27	0.48	2.69	0.43	1.54
EBF	0.59	2.05	0.44	2.57	0.18	1.90
SAV	0.43	1.87	0.19	2.42	0.24	1.51

阔叶林、稀树草原等的预测结果较差(表 1)。与GPP模型对比表明,之前4种自变量能够较好地预测GPP,但是对于NEP效果相对较差,究其原因,在于GPP是固碳参量,而其净生态系统生产力则不仅受到固碳量参数影响,也受到损耗量ER等影响,体现了碳汇的大小,因此其影响因素更多,这表明在进行机器学习建模时,训练数据集自变量的选择仍然需要机理模型的支持。

4.2 时空扩展

为了将GPP和NEP两种碳收支遥感监测变量扩展到全球,本研究将机器学习模型嵌入碳收支遥感监测云平台。在平台中利用该算法和平台提供的API,调用“增强植被指数”、“地表温度”、“降雨”、“短波辐射”4个变量对应的遥感数据,实现2002—2016年全球GPP和NEP的计算,影像的相关预处理与模型训练数据集预处理保持一致性。基于该平台,利用机器学习模型计算获取了2002—2016年全球逐年GPP数据,空间范围为60° N—60° S,下图为全球平均GPP空间分布(图5)。

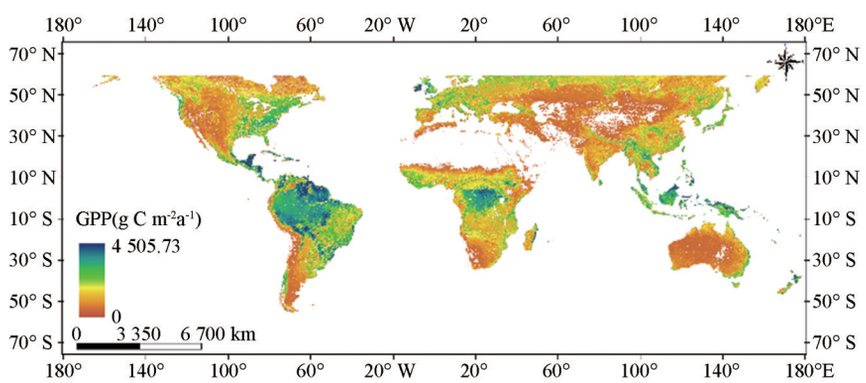

图 5 基于机器学习模型的2002—2016年全球平均GPP空间分布

Fig.5 Spatial distribution of global average GPP from 2002—2016 based on machine learning model

从图(5)可以看出,2002—2016年全球平均GPP存在明显的空间差异,热带雨林区年均GPP总体上呈现较高值,例如南美洲亚马逊热带雨林、非洲中部刚果盆地和东南亚热带森林地区等,而亚欧大陆北部、北美洲高山及高寒地区及大洋洲大部分区域年均GPP值相对较低。

为研究全球GPP时间分布特征,本研究计算了2002—2016年全球GPP逐年变化趋势,并统计了变化显著性水平(图6)。研究发现,2002—2016年间,全球55.59%的区域GPP呈现下降趋势,44.41%的区域GPP呈现增加趋势,但变化幅度较小,年变化量介于±15g $Cm^{-2}a^{-1}$的区域占研究区域的86.90%,年变化量介于±5g $Cm^{-2}a^{-1}$的区域占研究区域的59.19%,说明2002—2016年间全球大部分地区GPP的变化趋势并不明显,发生显著变化区域仅占研究区域的11.2%($p < 0.01$ 和 $p<0.05$ 的区域所占比例分别为:4.89%、6.31%),主要集中在亚洲东部地区及北美洲森林地区等,与亚洲变绿的观点一致[31]。

5 结 论

本研究基于遥感观测数据和地面通量观测数据相结合进行陆地生态系统碳收支的监测,建立了数据驱动的基于机器学习模型和大数据系统的陆地生态系统碳收支参数监测的方法,并基于云计算平台建立碳收支监测系统,可以实现碳收支监测的时空扩展。

结合多源数据和随机森林算法,提出估算全球站点尺度GPP的数据驱动方法,能够得到不同植被类型的GPP预测结果,模型预测结果较MODIS

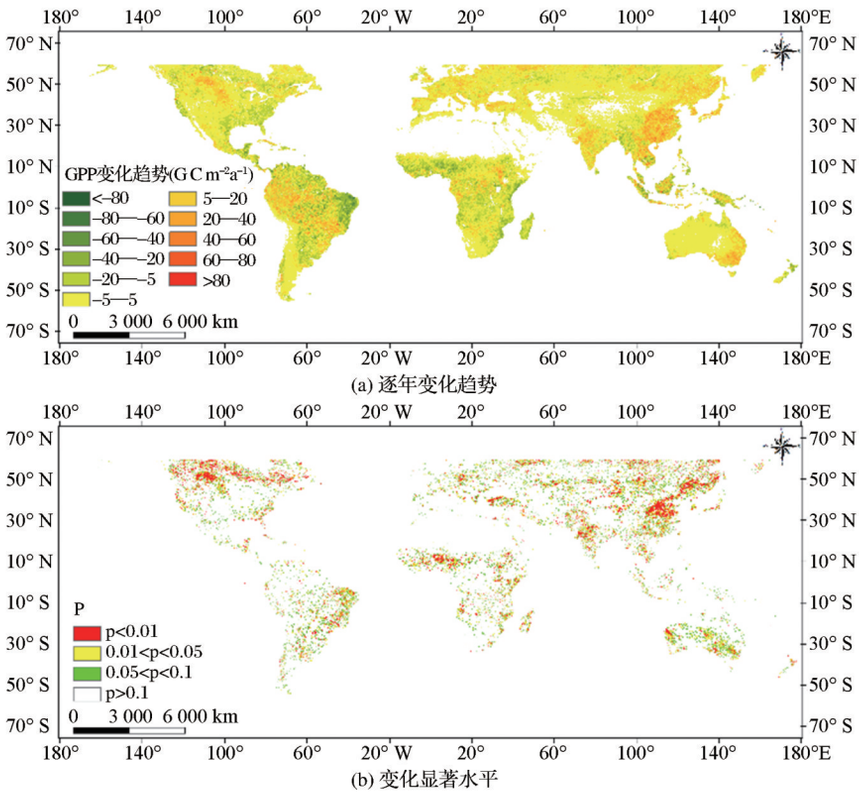

图 6　2002—2016 年全球 GPP

Fig.6　Global GPP from 2002—2016

GPP 产品具有更高的精度,也更接近于地面观测数据。将方法运用在 NEP 的估测研究,也得出较好的估测结果。对比表明,增强型植被指数、陆地表面温度、短波辐射、降水数据等 4 种自变量能够较好的预测 GPP,但是对于 NEP 效果相对较差,究其原因,主要在于 GPP 是固碳参量,NEP 是固碳量与损耗量的差值,体现了碳汇的大小,因此其影响因素更多。

为进行站点模型的时空扩展,本研究基于 GEE 云计算平台,使用 Python 调用 GEE 提供的 API,在该平台实现了核心算法,同时以此为计算后端,基于 Django 和 Nginx 搭建了 Web 服务平台,并基于 jQuery、OpenLayers 等网页技术进行平台前端的编写,从而实现数据一体化组织管理、碳收支结果快速计算和显示等。基于该平台和算法,本研究开展了 GPP 和 NEP 两种碳收支全球遥感监测,计算获取了 2002—2016 年全球(60° N—60° S)逐年 GPP 数据,研究表明全球平均 GPP 存在明显的空间差异,热带雨林区年均 GPP 总体上呈现较高值,而亚欧大陆北部、北美洲高山及高寒地区及大洋洲大部分区域年均 GPP 值相对较低。时间趋势研究表明,全球大部分地区 GPP 的变化趋势并不明显,发生显著变化区域仅占研究区域的 11.2%,主要集中在亚洲东部地区及北美洲森林地区等。

本研究通过机器学习和大数据平台等工具,进行了碳收支遥感监测,对比表明机器学习模型能够明显提高传统卫星遥感模型的精度,同时避免生理过程模型复杂的参数设置,减少区域和全球大尺度碳收支监测的不确定性。研究表明,项目通过数据挖掘、机器学习等方法,基于高质量卫星及地面观测数据产品,可以建立数据驱动的碳收支遥感监测方法,一定程度上避免碳循环模式模拟的不确定性,快速提供全球和区域碳收支的直接估算结果。

参考文献(References):

[1] Solomon S, Qin D, Manning M, *et al*. Climate Change 2007: The physical science basis. contribution of working group I to the fourth 16 assessment report of the intergovernmental panel on climate change[R]. Summary for Policy Makers, Geneva, SwitzerLand: Intergovernmental Panel on Climate (IPCC), 2007:18.

[2] Falkowski P, Scholes R, Boyle E, *et al*. The global carbon cycle: A test of our knowledge of earth as a system[J]. Science, 2000, 290(5490): 291-296.

[3] Xia J, Chen J, Piao S, *et al*. Terrestrial carbon cycle affected by non-uniform climate warming [J]. Nature Geoscience, 2014, 7(3): 173-180.

[4] Pourtaghi Z S, Pourghasemi H R, Rossi M. Forest fire sus-

ceptibility mapping in the minudasht forests, Golestan Province, Iran[J]. Environmental Earth Sciences, 2015, 73(4): 1515-1533.

[5] Running S W, Nemani R R, Heinsch F A, *et al*. A continuous satellite-derived measure of global terrestrial primary production[J]. Bioscience, 2004, 54(6): 547-560.

[6] Raich J W, Schlesinger W H. The Global Carbon-Dioxide Flux in soil respiration and its relationship to vegetation and climate[J].Tellus Series B-Chemical and Physical Meteorology, 1992,44(2):81-99.

[7] Li Kerang,Huang Mei,Tao Bo.Process-based modeling on the response and adaptation of chinese terrestrial ecosystems to global change[M]. Beijing: China Meteorological Press,2009.[李克让, 黄玫, 陶波.中国陆地生态系统过程及对全球变化响应与适应的模拟研究[M].北京:气象出版社,2009.]

[8] Baldocchi D, Falge E, Gu L H, *et al*. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities [J]. Bulletin of the American Meteorological Society, 2001, 82(11): 2415-2434.

[9] Baldocchi D. Measuring fluxes of trace gases and energy between ecosystems and the atmosphere-the state and future of the eddy covariance method[J].Global Change Biology,2014, 20(12): 3600-360.

[10] Geng Shaobo,Lu Shaowei,Rao Liangyi, *et al*. Research progress of measurement of land surface carbon budget based on eddy covariance technology[J]. World Forestry Research,2010, 23(3): 24-28.[耿绍波, 鲁绍伟, 饶良懿,等.基于涡度相关技术测算地表碳通量研究进展.世界林业研究, 2010,23(3): 24-28.]

[11] Jensen R, Herbst M ,Friborg T. Direct and indirect controls of the interannual variability in atmospheric CO_2 exchange of three contrasting ecosystems in denmark[J]. Agricultural and Forest Meteorology,2017,233:12-31.DOI:10.1016/j.agrformet2016.10.023.

[12] Zhu A X, Mackay D S. Effects of spatial detail of soil information on watershed modeling[J]. Journal of Hydrology, 2001, 248(1): 54-77.

[13] White M A, Thornton P E, Running S W *et al*. Parameterization and sensitivity analysis of the BIOME-BGC terrestrial ecosystem model: Net primary production controls[J]. Earth Interactions, 2000,4(3):1-84.

[14] Chen W, Chen J, Liu J,*et al*. Approaches for reducing uncertainties in regional forest carbon balance[J]. Global Biogeochemical Cycles, 2000, 14(3): 827-838.

[15] Sitch S, Smith B, Prentice I C, *et al*. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model[J].Global Change Biology, 2003, 9(2): 161-185.

[16] Sims D A, Rahman A F, Cordova V D, *et al*. A new model of gross primary productivity for North American ecosystems based solely on the Enhanced Vegetation Index and Land Surface Temperature from MODIS[J]. Remote Sensing of Environment, 2008, 112(4): 1633-1646.

[17] Gitelson A A, Vina A, Verma S B, *et al*. Relationship between gross primary production and chlorophyll content in crops: Implications for the synoptic monitoring of vegetation productivity[J]. Journal of Geophysical Research-Atmospheres,2006,111:D08S11. DOI:10.1029/2005JD006017.

[18] Xiao X, Hollinger D, Aber J, *et al*. Satellite-based modeling of gross primary production in an evergreen needleleaf forest [J]. Remote Sensing of Environment,2004,89(4):519-534.

[19] Heinsch F A, Zhao M, Running S W, *et al*.Evaluation of remote sensing based terrestrial productivity from MODIS using regional tower eddy flux network observations[J].IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(7): 1908-1925.

[20] Yang F, Ichii K, White M A, *et al*. Developing a continental-scale measure of gross primary production by combining MODIS and ameriflux data through support vector machine approach[J]. Remote Sensing of Environment, 2007, 110(1): 109-122.

[21] Jung M, Reichstein M, Bondeau A. Towards global empirical upscaling of fluxnet eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model [J]. Biogeosciences,2009,6(60):2001-2013.

[22] Liu X, Guanter L, Liu L, *et al*. Downscaling of solar-induced chlorophyll fluorescence from canopy level to photosystem level using a random forest model[J]. Remote Sensing of Environment,2019,231:110772.DOI:10.1016/j.rse.2018. 05.035.

[23] Xu Ziming , Tian Yangfeng. The development history and application of cloud computing[J]. Information Recording Materials, 2018,19(8):66-67.[许子明,田杨锋.云计算的发展历史及其应用[J].信息记录材料,2018,19(8):66-67.]

[24] Chen B, Xiao X, Li X, *et al*. A mangrove forest map of China in 2015: Analysis of time series Landsat 7/8 and Sentinel-1A imagery in Google Earth Engine cloud computing platform [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2017, 131: 104-120. DOI:10.1016/j.isprsjprs.2017.07.011.

[25] Alonso A, Munoz-Carpena R, Kennedy R E, *et al*. Wetland lndscape spatio-temporal degradation dynamics using the new Google Earth Engine cloud-based platform: Opportunities for non-specialists in remote sensing[J]. Transactions of the Asabe, 2016, 59(5): 1331-1342.

[26] Chen Y, Shen W, Gao S, *et al*. Estimating deciduous broadleaf forest gross primary productivity by remote sensing data using a random forest regression model[J]. Journal of Applied Remote Sensing, 2019, 13(3):038502. DOI: 10.1117/1.JRS.13.038502.

[27] Baldocchi D D.Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems:Past,present and future[J].Global Change Biology,2003,9(4):479-492.

[28] Papale D, Reichstein M, Aubinet M, *et al*. Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: Algorithms and uncertainty estimation[J]. Biogeosciences, 2006, 3(15): 571-583.

[29] Hamill T M, Dentremont R P, Bunting J T. A description of the air force real-time nephanalysis model[J]. Weather and Forecasting, 1992, 7(2): 288-306.

[30] Hsu K L, Gupta H V, Gao X G ,*et al*.Estimation of physical variables from multichannel remotely sensed imagery using a neural network: Application to rainfall estimation[J]. Water

Resources Research, 1999, 35(5): 1605-1618.

[31] Zhang K, Liu N, Chen Y, *et al*. Comparison of different machine learning method for GPP estimation using remote sensing data[J] IOP Conference Series: Materials Science and Engineering, 2019, 490(6): 062010. DOI: 10.1088/1757-899X/490/6/062010.

[32] Fabian Gieseke, Christian Igel. Training big random forests with little resources [C] //2018: 1445-54. DOI: 10.1145/3219819.3220124.

[33] Chi C, Park T, Wang X, *et al*. China and India lead in greening of the world through land-use management[J]. Nature sustainability, 2019, 2(2): 122-129. DOI:10.1038/s41893-019-0220-7.

Remote Sensing Monitoring of Terrestrial Ecosystem Carbon Budget based on Machine Learning and Big Data Platform

Gao Shuai[1], Hou Xuehui[2], Wang Yun[3], Wang Qian[4], Chen Yue[1], Xing Rui[1], Wang Jing[1,5]

(1.*State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing* 100101, *China*;
2.*Institute of Agricultural Information and Economy, Shandong Academy of Agricultural Sciences, Jinan* 250100, *China*;
3.*The College of Forestry of Beijing Forestry University, Beijing* 100083, *China*;
4.*School of Geographic and Environmental Sciences, Tianjin Normal University, Tianjin* 300387, *China*;
5.*School of Earth Sciences and Resources, China University of Geosciences, Beijing* 100083, *China*)

Abstract: The carbon budget of terrestrial ecosystems is an important indicator of global carbon cycle research and an important parameter of climate change. Based on the terrestrial ecosystem flux observation and remote sensing satellite observation data, machine learning methods are applied for carbon budget estimation. In this study, random forest algorithm is established to automatically learn features from training data and differences in time series dependencies, and carbon related parameters (Gross Primary Production, GPP; Net Ecosystem Production, NEP) could be estimated. Finally, standard indicators are selected to objectively evaluate the model using the validation data set. The result analysis shows that compared with MODIS GPP products, this method has greatly improved the estimation accuracy. Among them, the prediction result of deciduous broad-leaved forest is the best, the decision coefficient R^2 is 0.82, and the root mean square error is 1.93 $gCm^{-2}\ d^{-1}$.It is also significantly better than traditional light energy utilization model products in other vegetation types. The NEP machine learning model established based on the same method has also obtained good estimation results. The correlation between the output results of the deciduous broad-leaved forest model prediction model and the NEP obtained by the flux tower is 0.70 and RMSE=1.75 g C $m^{-2}\ d^{-1}$. The difference in accuracy between GPP and NEP models indicates that when machine learning modeling is performed, the selection of independent variables in the training data set still needs to consider theoretical model. In order to quickly estimate the carbon budget of the terrestrial ecosystem, a remote sensing monitoring platform is established. The platform uses the GEE (Google Earth Engine) big data platform as the data storage and computing backend, and Django, HTML, CSS, JavaScript, etc. as the front-end, in order to quick calculation, real-time visualization and other functions. Based on the platform and algorithm, the global (60° N—60° S) GPP results obtained from 2002 to 2016 show that there are obvious spatial differences in the global average GPP, and the significant increase is mainly concentrated in eastern Asia and forested areas in North America. Research shows that remote sensing monitoring of carbon budget parameters based on machine learning and big data platforms can quickly provide regional and global-scale carbon storage and the results are consistent with true ground observations. The obtained estimation results avoid the complicated parameter setting of the physiological process model, and reduce the uncertainty of regional and global large-scale carbon budget monitor.

Key words: Machine learning; Big data platform; Carbon budget; Random forest; Spatio-temporal expansion