王馨仪, 吴楚仪, 吴森森, 等. 高分辨率海水表层二氧化碳分压重构——以大西洋为例[J]. 海洋学报, 2023, 45(3): 147–158, doi:10.12284/hyxb2023048

Wang Xinyi, Wu Chuyi, Wu Sensen, et al. Reconstruction of sea surface pCO_2 with high resolution: A case study of the Atlantic Ocean[J]. Haiyang Xuebao, 2023, 45(3): 147–158, doi:10.12284/hyxb2023048

高分辨率海水表层二氧化碳分压重构
——以大西洋为例

王馨仪[1, 2, 3]，吴楚仪[1, 2, 3]，吴森森[1, 2, 3]，陈奕君[1, 2, 3]，杜震洪[1, 2, 3*]

(1. 浙江大学 地球科学学院, 浙江 杭州 310027; 2. 浙江省资源与环境信息系统重点实验室, 浙江 杭州 310030; 3. 浙江大学 地理与空间信息研究所, 浙江 杭州 310027)

摘要: 海洋是自然界中重要的碳汇, 海–气二氧化碳通量通常利用大气和海水表层的二氧化碳分压（pCO_2）差进行估算。受制于时空分布不均匀的观测样本和预测数据, 目前已有海水表层二氧化碳分压的重构结果在空间分辨率上仍有较大可提升空间。为在高空间分辨率下更好地拟合时空变化, 基于表层大洋二氧化碳地图（SOCAT）的海水表层二氧化碳逸度（fCO_2）数据集和遥感卫星等多源数据, 利用 XGBoost 模型建立了海水表层二氧化碳分压值与海洋物理、生物、光学等要素的非线性关系, 并根据样本时空频率构建权重模型, 最终重构了 2000–2018 年大西洋 0.041 7°×0.041 7°下月度海水表层二氧化碳分压分布。预测结果的相关系数为 0.966, 均方根误差为 8.087 μatm, 平均偏差为 4.012 μatm, 与同类重构结果相比, 海水表层二氧化碳分压的时空变化趋势一致性强, 且在空间分辨率上具有优势。

关键词: 海水表层二氧化碳分压; 遥感卫星; 高空间分辨率

中图分类号: P732.6　　**文献标志码:** A　　**文章编号:** 0253–4193(2023)03–0147–12

1　引言

近代以来, 化石燃料的燃烧和树木砍伐等人类活动使全球碳排放量迅速增加, 大气中二氧化碳（CO_2）浓度大幅上升[1]。浓度不断升高的大气 CO_2 所带来的温室效应引起了全球气候变化, 对人类社会经济以及地球未来的生存环境产生了巨大威胁。碳循环中, 海–气界面的二氧化碳交换贡献了最大的自然碳通量, 近 10 年来海水每年约吸收碳 2.78 Gt, 约占人为碳排放总量的 26%[2]。因此, 海洋是自然界中重要的碳汇, 在调控地球生态系统及气候变化中起到了关键作用, 对于海–气二氧化碳通量的监测和量化能够帮助人类了解全球碳的流动和变化趋势, 为实现碳中和、碳达峰的目标提供参考。

海–气界面的二氧化碳通量通常利用大气和海水表层的二氧化碳分压（pCO_2）差进行估算。其中, 由于海水表层二氧化碳分压实测数据大多依赖于船舶测量, 受制于航线、航次的影响, 时空分辨率较低且分布不均匀, 使得海洋碳汇估算结果仍存在很大的不确定性和挑战。

海水表层二氧化碳分压与海水表面、内部变化和大气交换密切相关, 主要控制因素包括与温度相关的热力学效应、生物化学效应、水团之间的混合以及海气交换等[3]。目前已有丰富的再分析及模式数据提供

收稿日期: 2022–04–22; **修订日期:** 2022–10–12。

基金项目: 国家自然科学基金（201300001）。

作者简介: 王馨仪（1997—）, 女, 重庆市人, 研究方向为遥感与地理信息系统、遥感反演。E-mail: 21938031@zju.edu.cn

*** 通信作者:** 杜震洪（1981—）, 男, 教授, 博士生导师, 研究方向为遥感与地理信息系统、时空大数据与人工智能、大数据与地球–海洋系统。E-mail: duzhenhong@zju.edu.cn

了相关的海洋环境信息，但由于其空间分辨率较低，无法用于细尺度的海水表层二氧化碳分压重构。与之相比，卫星遥感数据具有长时序稳定、空间分辨率高的特点，能够提供海洋表面的物理、生物和光学特性信息，在海洋碳通量监测评估及海洋碳循环研究中具有极大的优势。

目前已有大量研究利用海洋卫星数据重构区域性海水表层二氧化碳分压分布，包括建立多元线性回归（Multivariate Linear Regression, MLR）[4]、多元非线性回归（Multi-variate Nonlinear Regression, MNR）[5]、多元多项式回归（Multi-variate Polynomial Regression, MPR）[6] 等回归方程，基于机理和组分的半分析方法[3] 以及随机森林回归模型（Random Forest Regression, RFR）[7]、支持向量机（Support Vector Machines, SVM）算法[8]、神经网络[9] 等机器学习方法。基于卫星遥感、原位或航线测量的数据，大部分聚焦于区域尺度的重构，具有较高的准确度和空间分辨率。但在大空间尺度下，当前研究主要以再分析与模式数据为主导，对高空间分辨率遥感卫星数据的利用有限。Takahashi 等[10] 基于表层海水的扩散和平流输送公式在空间和时间上进行插值，建立了 4°×5°海水表层二氧化碳分压月平均全球分布数据。Landschützer 等[11] 建立了 SOM-FFNN 两步神经网络，结合已有低分辨率分压数据集和叶绿素 *a*（chlorophyl *a*, Chl *a*）浓度、海表面温度（Sea Surface Temperature, SST）、海表面盐度（Sea Surface Salinity, SSS）、海洋混合层深度（Mixed Layer Depth, MLD）等再分析与模式数据集，第一步使用自组织映射神经网络（Self-Organizing Feature Map, SOM）方法将全球海洋分区，再分别针对每个区建立环境驱动因素 SST、SSS、干燥大气 CO_2 摩尔分数（xCO_2）、MLD 与观测值的前馈神经网络（Feed Forward Network, FFN），构建了 1°×1°的全球月平均数据集，均方根误差约为 20 μatm（1 μatm=0.101 325 Pa）。Krishna 等[12] 分析并证实了 Chl *a* 浓度、SST、SSS 与海水表层二氧化碳分压的密切联系，并分别利用遥感卫星和原位测量数据，建立了测量值与各因子的分段多元非线性回归关系，重构了海水表层二氧化碳分压的全球尺度分布，但该重构结果受制于 SSS 数据集的影响，最终重构空间分辨率为 1°×1°。

当前研究在重构海水表层二氧化碳分压的连续分布方面已取得了很大进展，但在空间分辨率上仍有不足。空间分辨率的提高能够充分利用已有观测数据并减少观测样本时空匹配到格网中的误差，且能更好地表现海水表层二氧化碳分压的细节变化。本研究以遥感卫星数据为主导，将遥感波段及衍生产品的原始反射率、海表光学特性、SST 和 Chl *a* 浓度作为表征海水表面物理、生物及其他过程的要素，以大西洋为研究区域重构高空间分辨率下的海水表层二氧化碳分压。但由于遥感卫星产品难以直接反映海水内部变化及与大气的交互过程，且受到天气、硬件波动的影响，存在数据缺失，因此将上采样后的 SSS、MLD、xCO_2 以及海表 10 m 风速（u_{10}）再分析及模式数据作为辅助要素，分别表征海洋环境中的化学、物理要素及与大气界面的交换过程。在此基础上，结合时间和空间特征以提高对海水表层二氧化碳分压时空变异性的拟合能力。

针对高空间分辨率下的不平衡大数据集，本研究使用 XGBoost（eXtreme Gradient Boosting）模型[13] 建立特征要素与海水表层二氧化碳分压的非线性关系实现重构。XGBoost 模型是以梯度提升树（Gradient Boosting Tree, GBT）为基础的优化树模型，具有精度高、抗噪声强、可并行处理等优点。XGBoost 模型对稀疏特征值的高效学习策略能够应对遥感数据的缺失问题，基于树模型的可解释性进行特征筛选能够提高训练效率。此外，研究中针对海水表层二氧化碳分压观测样本时空分布不均匀的问题建立了样本时空权重模型，以提高模型学习的平衡性。

2 数据来源

本研究将表层大洋二氧化碳地图（The Surface Ocean CO_2 Atlas, SOCAT）[14] 实测数据集作为海水表层二氧化碳分压观测样本数据来源。该数据集由超过 100 个国际海洋碳研究团体进行质量控制，至今已发布从 1957 年至 2020 年对全球海洋和沿海海域的 3 060 万次观测数据。由于 SOCAT 数据集仅提供海水表层二氧化碳逸度（fCO_2）数据，需修正至二氧化碳分压值[15]。

$$pCO_2 = fCO_2 \times \exp\left[-\frac{P_{atm}(B+2\delta)}{RT}\right], \tag{1}$$

式中，pCO_2 表示海水表层二氧化碳分压；fCO_2 表示海水表层二氧化碳逸度；P_{atm} 表示大气压，单位为 Pa；R 为理想气体常数，值为 8.314 J/(mol·K)；B 与 δ 表示与温度 T（K）相关的修正系数，单位为 m^3/mol，计算公式为

$$B = \left(-1\,636.75 + 12.040\,8T - 3.279\,57\times10^{-2}T^2 + 3.165\,28\times10^{-5}T^3\right)\times10^{-6}, \tag{2}$$

$$\delta = (57.7 - 0.118T)\times10^{-6}. \tag{3}$$

本研究使用 SOCAT version 2020 数据集，观测数

据在大西洋 1°×1°网格上的空间分布与数量的时间分布如图 1 和图 2 所示。由图可知, 大西洋观测点的时空分布极不均匀。从空间上看, 观测点数量北半球多于南半球, 沿海多于大洋中心。其中北大西洋沿岸分布最为密集, 南大西洋的中心海域缺失较为严重。观测点的数量在时间序列中呈不断上升的趋势, 分布的较不均匀。

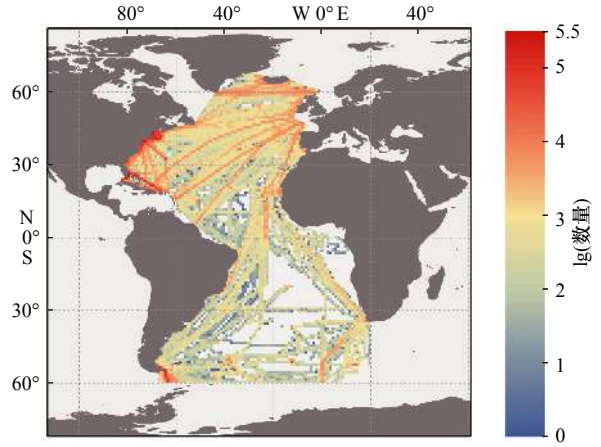


图 1　大西洋观测点空间分布

Fig. 1　Space contribution of observations in Atlantic

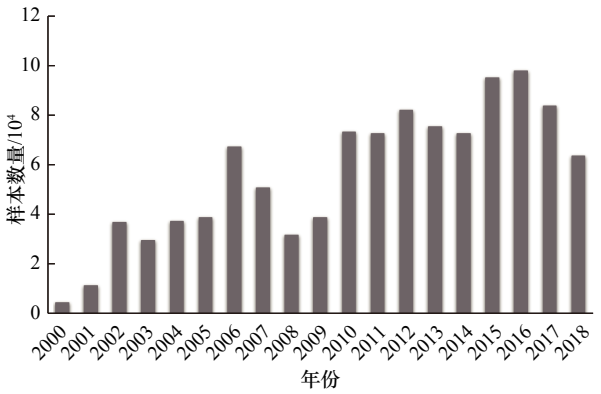


图 2　大西洋观测点数量时间分布

Fig. 2　Numbers of observations in Atlantic in time serial

研究使用 ESA(The European Space Agency) OC-CCI (Ocean-Colour Climate Change Initiative)融合产品[16] 作为高空间分辨率卫星遥感数据。该产品融合了 MERIS(MEdium Spectral Resolution Imaging Spectrometer)、MODIS-Aqua(Moderate-resolution Imaging Spectroradiometer-Aqua)、VIIRS(Visible and Infrared Imaging Radiometer Suite)、SeaWiFS(Sea-viewing Wide-Field-of-view Sensor)4 种遥感数据, 并优化现有算法提供了一套全球尺度的高空间分辨率(0.041 7°)水色遥感反射率及其衍生产品数据集, 在空间和时间序列上实现了数据补全, 为研究长时序海洋生物、化学、气象的季节性和年际尺度变化与相互作用提供支持。表 1 列举了本研究中所用数据波段及说明。

表 1　ESA OC-CCI 使用波段说明

Table 1　Introduction of bands used in ESA OC-CCI

遥感产品	波长/nm	遥感产品	波长/nm	遥感产品	波长/nm
黄色物质和碎屑吸收系数(a_{dg})	412	总吸收系数(a_{tot})	412	遥感反射率(R_{rs})	412
	443		443		443
	490		490		490
	510		510		510
	560		560		560
	665		665		665
浮游植物吸收系数(a_{ph})	412	粒子后向散射系数(b_{bp})	412	向下漫射衰减系数(K_d)	490
	443		443		
	490		490	叶绿素a(Chl a)浓度	–
	510		510		
	560		560		
	665		665		

注:“–”代表空值。

其他数据(SST、SSS、MLD、u_{10}、xCO_2)来源及说明如表 2 所示。(1)SST 使用 MODIS 卫星的月平均 L3 产品, 由于 Aqua 传感器对应 SST 数据集时序相对较短(始于 2002 年), 本研究使用 Terra 传感器的 0.041 7° SST 数据(https://oceandata.sci.gsfc.nasa.gov/directaccess/MODIS-Terra/Mapped/Monthly/4km/sst4);(2)SSS 和 MLD 使用 ECCO2(Estimating the Circulation and Climate of the Ocean, Phase II) Cube92[17] 每日数据(https://apdrc.soest.hawaii.edu/erddap/griddap);(3)海面 10 m 风速和用于补全的 SST 数据使用 ERA5 (5th generation ECMWF reanalysis)[18] 单层月平均数据集(https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means);(4)xCO_2 数据集使用 GML(Global Monitoring Laboratory) CarbonTracker CT2019B[19] 每日数据(https://gml.noaa.gov/aftp/

表 2　辅助数据来源

Table 2　Source of ancillary data

数据类型	数据来源	数据集	空间分辨率
遥感数据	MODIS Terra传感器	SST	0.041 7°×0.041 7°
模式数据	ECCO2 Cube92	SSS	0.25°×0.25°
		MLD	
	GML CarbonTracker CT2019B	xCO_2	3°×2°
再分析数据	ERA5 单层月均数据集	SST	0.25°×0.25°
		u_{10}	

products/carbontracker/co2/CT2019B/molefractions/xCO2_1330LST)；以上数据除 MODIS SST 数据外均利用最邻近线性插值上采样至 0.041 7°×0.041 7°分辨率，SSS、MLD、xCO_2 月度数据为日数据的月平均值。由于 MODIS 数据起始于 2000 年，CarbonTracker CT2019B 数据仅更新至 2019 年，因此本研究的时间范围为 2000−2018 年。

3 研究方法

3.1 XGBoost 模型

本研究利用 XGBoost 模型构建海水表层二氧化碳分压与多源特征数据之间的非线性关系实现预测重构。XGBoost 模型采用了机器学习中集成学习和梯度下降的思想，通过多个特征生成多个树基模型，将所有预测结果相加以提升决策效果。模型中每一个基模型拟合的是上一个基模型的残差，即预测值和真实值的损失函数。同时，计算上一棵树损失函数的负梯度作为新树生成的依据，以快速降低系统误差的大小。其次，XGBoost 模型中集成了样本权重、稀疏特征学习和防过拟合策略，并通过并行优化、缓冲处理和核外计算大大提高了模型的训练和预测效率，适用于高空间分辨率下海水表层二氧化碳分压的大型非平衡数据集，并能够有效应对遥感数据部分缺失的情况。对于具有 n 个样本、m 个特征的数据集 $D=\{(x_i,y_i)\}$ $(i=1,2,\cdots,n)$，其中 x_i $(x_i\in\mathbb{R}^m)$ 表示样本的特征集合，y_i $(y_i\in\mathbb{R})$ 表示样本标签值。XGBoost 模型由 K 个基模型的加法模型实现预测：

$$\hat{y}_i=\varphi(x_i)=\sum_{k=1}^{K}f_k(x_i),\quad f_k\in F, \tag{4}$$

式中，$\hat{y}_i$ 表示模型的预测值；K 表示树模型的数量；f_k 表示第 k 个树模型；F 表示基模型的假设空间。为使得模型的预测值准确率提高并具有尽量大泛化能力，定义目标函数 $L(\varphi)$ 为

$$L(\varphi)=\sum_{i=1}^{n}l\left(\hat{y}_i,y_i\right)+\sum_{k=1}^{K}\Omega(f_k), \tag{5}$$

式中，$l\left(\hat{y}_i,y_i\right)$ 表示第 i 个样本的预测误差；$\Omega(f_k)$ 表示第 k 个树模型复杂度，定义为

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2, \tag{6}$$

式中，T 表示叶节点的个数；w_j 表示叶子的权重分数；γ 和 λ 表示惩罚项权重系数。复杂度惩罚项的设置能够防止模型过拟合，且在函数结构上进行了简化而更易于并行。

此外，考虑到现实数据集中存在的样本特征稀疏性问题，XGBoost 模型采用了一致性的处理策略，即在每个树节点设置稀疏值的默认方向，并与未丢失特征一起参与枚举并从数据中学习最优解，提高了处理效率。

3.2 基于 XGBoost 模型的高空间分辨率海水表层二氧化碳分压回归反演

本研究融合遥感、模式和再分析多源数据，挖掘海水表层二氧化碳分压与表征海洋表面、内部变化及大气交换等过程要素的时空关联，重构高空间分辨率下大西洋月度海水表层二氧化碳分压分布。研究步骤如下：

(1)构建样本数据集。修正 SOCAT 二氧化碳逸度数据集得到海水表层二氧化碳分压观测样本，采用 SOCAT 在 1°×1°月平均格网数据集中按航线进行加权平均的方法进行网格化得到 0.041 7°×0.041 7°海水表层二氧化碳月度格网数据集[20]。由于从不同航次收集的测量数据时间分辨率可能存在较大差距，网格单元内的测量平均值将偏向于更高时间分辨率航次的测量结果，造成一定的偏差。基于航次的加权平均值首先对网格单元内给定航次的所有测量数据进行平均，再对所有航次的平均值进行第二次平均。因此，尽管初始时间分辨率存在较大差距，加权平均依然能够尽可能使不同航次的观测数据拥有相同的权重。将观测数据集按年、月、0.041 7°×0.041 7°格网划分子数据集 $Y^{a,mon,r,c}$（r 为格网行号，c 为格网列号），再按照数据集内样本采集的 C 条航线划分为 $\{Y_j^{a,mon,r,c}\mid j=(1,2,\cdots,C)\}$，则月平均值 $\bar{y}^{a,mon,r,c}$ 可由式(7)、式(8)计算：

$$\bar{y}^{a,mon,r,c}=\frac{\sum_{j=1}^{C}\bar{Y}_j^{a,mon,r,c}}{C}, \tag{7}$$

$$\bar{Y}_j=\frac{\sum_{i=1}^{a}y_i}{a},\quad y_i\in Y_j. \tag{8}$$

样本数据时空匹配到月度 0.041 7°×0.041 7°格网的标准差为 1.78 μatm，与 SOCAT 1°×1°网格产品的标准差 4.9 μatm[20] 相比有所下降，说明在高空间分辨率下观测数据网格化所引起的误差有所下降。月度格网观测集再与上采样至 0.041 7°×0.041 7°后的模式、再分析数据以及遥感数据进行空间匹配，加入经纬度作为空间特征，月份作为时间特征，得到样本特征集。最终，样本特征集含有 106 万条数据，以 8∶2 的比例将数据随机划分为训练集和测试集。

(2)构建多源特征数据集。由于遥感、再分析与模式数据时空分辨率存在差异，首先通过上采样、时

序平均将所有数据统一至月尺度, 0.041 7°×0.041 7°分辨率。其次, 针对遥感卫星受传感器内部硬件故障或恶劣大气条件影响引起的数据缺失问题, 尽管 XGBoost 模型集成了应对稀疏特征值的策略, 但考虑到海表面温度对于海水表层二氧化碳分压具有较直接和显著的影响, 利用 ERA5 的全覆盖再分析数据对 MODIS 的海表面温度数据进行简单补全, 以提高重构结果的覆盖度和精确度。

(3)构建训练样本权重模型。由于样本数据在时空上分布不均匀, 在 1°尺度下建立时空权重调整模型训练的平衡性。将月度观测数据集按年、月、1°×1°格网划分为子数据集$Y'^{a,mon,r,c}$, 按月、1°×1°格网划分子数据集$Y'^{mon,r,c}$, 则$Y'^{a,mon,r,c}$内样本的空间权重由数据集内样本个数决定; 考虑到海水表层二氧化碳的周期性、季节性变化较强, 时间权重由该像元在对应月份的子数据集个数决定, 即:

$$w^{a,mon,r,c}=\varepsilon\times w_{\text{space}}^{a,mon,r,c}\times w_{\text{time}}^{a,mon,r,c}, \tag{9}$$

$$w_{\text{space}}^{a,mon,r,c}=\frac{1}{count(y_i)},\quad y_i\in Y'^{a,mon,r,c}, \tag{10}$$

$$w_{\text{time}}^{a,mon,r,c}=\frac{1}{count(Y'^{mon,r,c})}, \tag{11}$$

式中, $w^{a,mon,r,c}$为某时刻某单元内样本的权重; $w_{\text{space}}^{a,mon,r,c}$为某时刻某单元内样本的空间权重; $w_{\text{time}}^{a,mon,r,c}$为某时刻某单元内样本的时间权重; ε为权重缩放系数。

(4)构建并优化模型。为提高模型训练效率和精度, 通过预训练模型的特征重要性对特征要素进行筛选。使用十折交叉验证法将训练数据集平均划分为 10 份, 每次将 1 份作为验证集、9 份作为训练集训练 XGBoost 模型, 并使用网格搜索对模型参数进行优化, 过程中以表征模型拟合度的相关系数 R^2 为评价模型的标准。对于真实值$\{y_1,y_2,\cdots,y_n\}$及预测值$\{p_1,p_2,\cdots,p_n\}$, 相关系数 R^2 由回归平方和与总体总离差的比值决定, 即:

$$R^2=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2}=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2}, \tag{12}$$

式中, $\bar{y}_i$表示真实值平均值; $\hat{y}_i$表示p_i和y_i的回归方程, 即$\hat{y}_i=\hat{a}+\hat{b}p_i$, 由最小二乘法可解得:

$$\hat{b}=\frac{\sum_{i=1}^{n}p_iy_i-n\bar{p}_i\bar{y}_i}{\sum_{i=1}^{n}p_i^2-n\bar{p}_i}, \tag{13}$$

$$\hat{a}=\bar{y}_i-\hat{b}\bar{p}_i. \tag{14}$$

模型参数确定后, 利用测试集验证模型的有效性。除 R^2 外, 均方根误差(Root Mean Squard Error, RMSE)、平均偏差(Average Deviation, AD)也将作为验证模型精确度的指标。

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-p_i)^2}{n}}, \tag{15}$$

$$\text{AD}=\frac{\sum_{i=1}^{n}|y_i-p_i|}{n}. \tag{16}$$

(5)利用构建模型反演 2000 年至 2018 年海水表层二氧化碳分压数据集。为进行进一步比较, 对 $x\text{CO}_2$ 进行水汽校正[15] 得到大气二氧化碳分压$p\text{CO}_2^{\text{atm}}$。

$$p\text{CO}_2^{\text{atm}}=x\text{CO}_2\times(1-p\text{H}_2\text{O}^{\text{sea}}), \tag{17}$$

$$p\text{H}_2\text{O}^{\text{sea}}=p\text{H}_2\text{O}^{\text{pure}}\times\exp(-0.018\varphi m_B), \tag{18}$$

式中, $p\text{H}_2\text{O}^{\text{sea}}$表示海水蒸汽压; $p\text{H}_2\text{O}^{\text{pure}}$表示纯水蒸汽压; m_B表示溶解物质的总浓度; φ表示海水渗透系数。在温度为 T(K)的纯水中:

$$p\text{H}_2\text{O}_T^{\text{pure}}=P_c\times\exp\left[\frac{T_c}{T}\left(a_1\vartheta+a_2\vartheta^{1.5}+a_3\vartheta^3+a_4\vartheta^{3.5}+a_5\vartheta^4+a_6\vartheta^{7.5}\right)\right], \tag{19}$$

式中, $T_c=647.096$ K; $\vartheta=(1-T/T_c)$; $P_c=22.064$ MPa; $a_1=-7.859\,517\,83$; $a_2=1.844\,082\,59$; $a_3=-11.786\,649\,7$; $a_4=22.680\,741\,1$; $a_5=-15.961\,871\,9$; $a_6=1.801\,225\,02$。

在 25℃ 海水中:

$$m_B=\frac{31.998\text{SSS}}{1\,000-1.005\text{SSS}}, \tag{20}$$

$$\varphi=0.907\,99-0.089\,92\left(\frac{m_B}{2}\right)+0.184\,58\left(\frac{m_B}{2}\right)^2-0.073\,95\left(\frac{m_B}{2}\right)^3-0.002\,21\left(\frac{m_B}{2}\right)^4. \tag{21}$$

尽管渗透系数φ理论上与温度相关, 但当 0℃≤T≤40℃ 时其数值变化小于 1%, 因此在研究中忽略此部分变化。

4　实验结果

4.1　XGBoost 模型构建

模型训练的硬件环境为 GeForce RTX 2080Ti, 操作系统为 Ubuntu, 语言环境为 Python 3.9。

预训练中, 将除时间和空间外所有特征在模型中被用作分割样本的特征次数作为评判标准筛选重要性较低的特征要素, 为避免训练的随机性, 取多次预训练的平均结果。由于现有研究普遍认为叶绿素 a 浓度为影响海水表层二氧化碳分压的重要因素, 最终以叶绿素 a 浓度的特征重要性为阈值保留 24 个特征。图 3 中深色标识的特征为正式训练所用特征集。

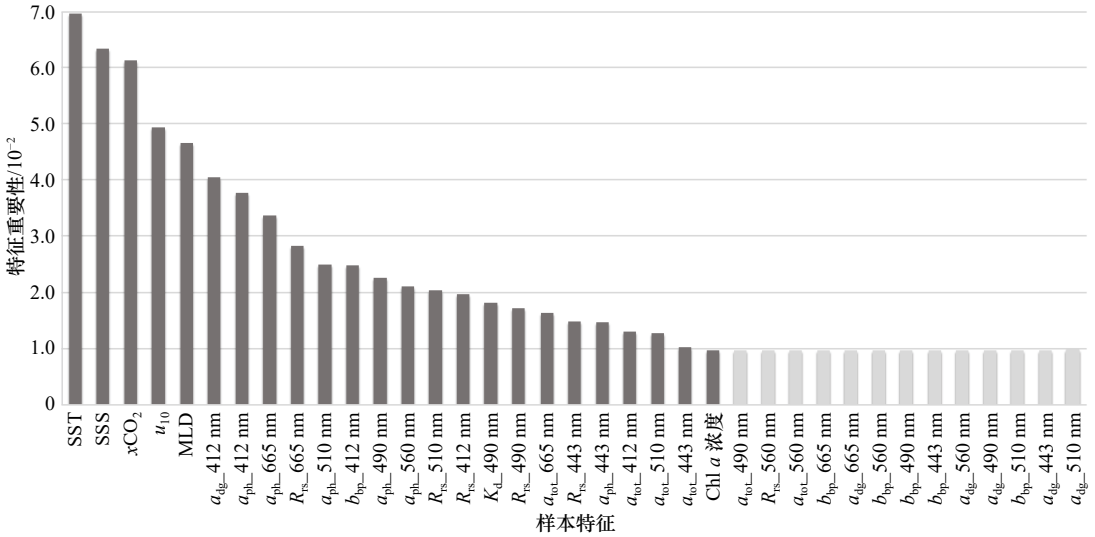


图 3 预训练模型样本特征重要性

Fig. 3 Feature importance of pre-training model

模型训练过程中设置学习率为 0.01, 权重缩放系数ε设置为 1 000, 使用网格搜索选择最佳迭代次数(K)、最大树深度(max_depth)。由于数据量较大, 两者数值的增加均会使模型拟合度提高, 但达到一定数值后提升效果有限。考虑到模型训练和预测的效率以及防止过拟合, 最终设置迭代次数为 10 000, 最大树深度为 15。训练模型在样本测试集上的 R^2=0.966, RMSE=8.078 μatm, AD=4.012 μatm, 拟合程度及精度均较高。图 4 展示了正式训练中所有特征的重要性排序。总体上看, 海表面温度和叶绿素 a 浓度的重要性高于其他特征, 说明热力学和生物效应是影响全球海水表层二氧化碳分压的最主要因素。同时, 除了与海洋生物、物理、化学过程直接相关的特征, 412 nm 粒子后向散射系数、510 nm 总吸收系数和 510 nm 浮游植物吸收系数等海面光学特性也对海水表层二氧化碳分压的预测具有较为重要的贡献。

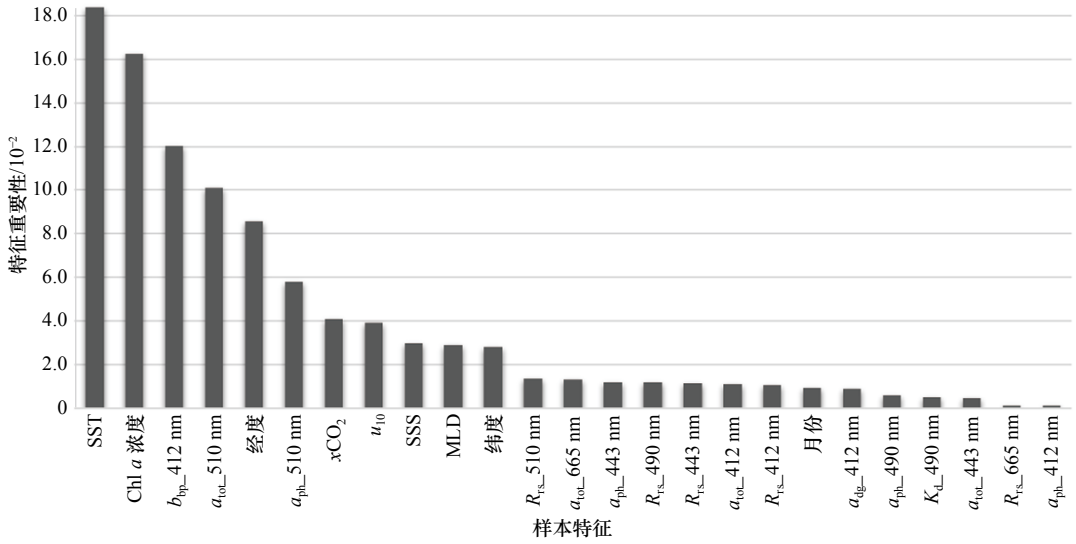


图 4 训练模型特征重要性

Fig. 4 Feature importance of model

由于训练样本数量在南北半球存在较大差异, 如图 5 将大西洋沿赤道划分为南北子区域分别进行误差统计, 结果如表 3、图 6 所示。结果显示, 模型在空间分布、时间序列上的精度和拟合度均较高, 未出现部分拟合现象, 说明模型对于海水表层二氧化碳在高空间分辨率及大洋尺度下的时空变异性具有较好的拟合能力。此外, 利用 SOCAT 数据集中未覆盖的 BATS(Bermuda Atlantic Time-series Study)[21] 及 ESTOC(The European Station for Time Series in the Ocean Canary Islands)[22] 站点在研究时间内的观测数据与重

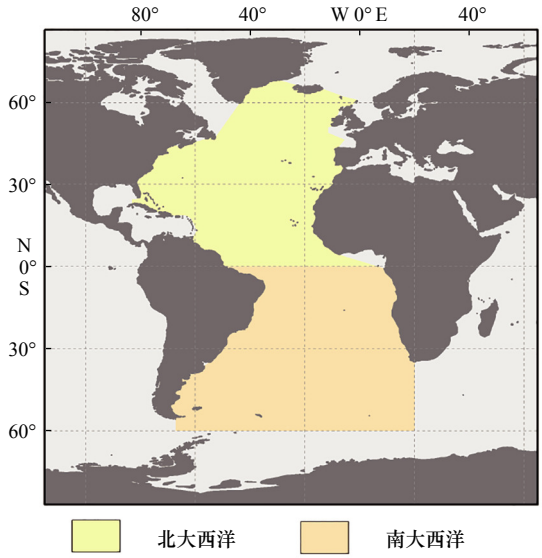


图 5　洋区划分

Fig. 5　Partition of the Ocean

表 3　洋区模型验证

Table 3　Verification of model for ocean area

洋区	RMSE/μatm	AD/μatm	R^2
北大西洋	6.001	2.143	0.984
南大西洋	5.295	1.987	0.991

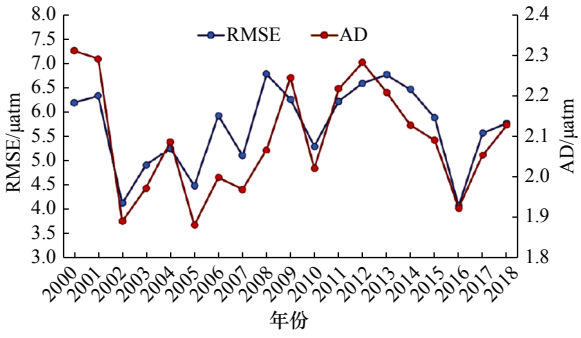


图 6　模型验证

Fig. 6　Verification of model

构结果对比以进一步检验重构结果的准确性, 在时序上的变化趋势分别如图 7、图 8 所示, 精度统计结果如表 4 所示。由表可知, 重建结果与两站点观测样本的各项误差均高于 SOCAT 预测集的误差, 其中与 ESTOC 站点数据的误差更小。从时间序列上看, 重构结果与 ESTOC 站点测量值的变化趋势一致性强, 与 BATS 站点测量值的整体变化趋势相似, 但在拟合部分冬季低异常值和夏季高异常值时存在一定偏差。

4.2　2000−2018 年大西洋海水二氧化碳分压反演结果

由图 9 可知, 总体上看, 大西洋大气及海水表层二氧化碳分压均呈上升趋势, 且在时序上的变化具有一定的相关性。使用趋势线拟合可知, 大气二氧化碳

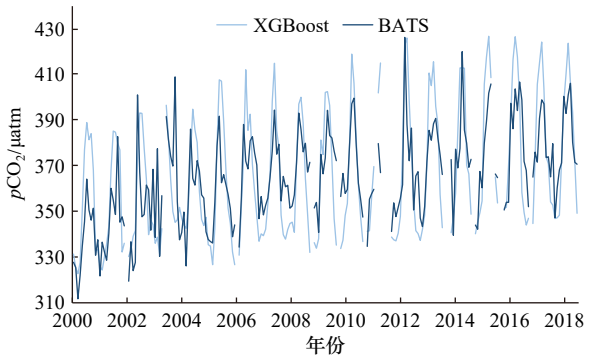


图 7　观测站点对比

Fig. 7　Comparison with observation stations

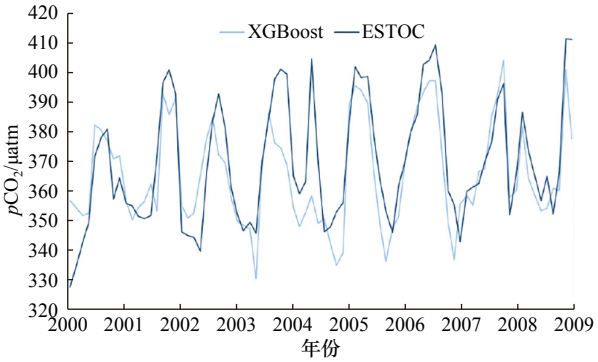


图 8　观测站点对比

Fig. 8　Comparison with observation stations

表 4　观测站点误差

Table 4　Error with observation stations

站点	位置	RMSE/μatm	AD/μatm
BATS	31.66°N, 64.16°W	18.90	14.37
ESTOC	29.04°N, 15.50°W	11.95	4.04

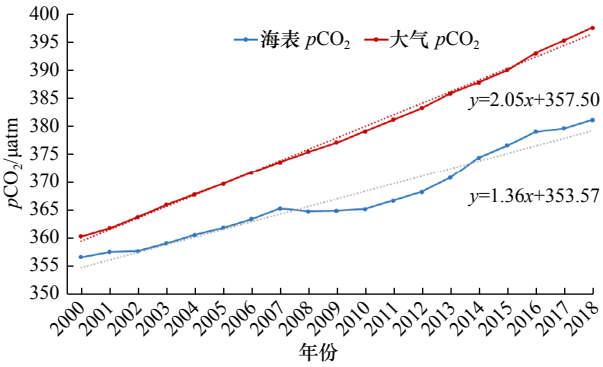


图 9　大西洋大气与海洋 pCO_2

Fig. 9　pCO_2 of air and sea in Atlantic

分压的增长速率高于海洋(约为海洋的 1.51 倍), 这表明尽管大西洋对大气中二氧化碳的吸收与大气二氧化碳分压同时升高, 但随着时间的迁移, 大气二氧化碳浓度可能仍将持续提升, 造成更严重的温室效应。

为了进一步分析海水表层二氧化碳分压的年际

变化，分别对南北海域计算年平均值，结果如图 10 所示。两个半区海水表层二氧化碳分压年间变化偶有起伏，但整体呈现上升趋势。从局部看，南大西洋分压平均值高于北大西洋，但北大西洋的分压值增长速度更快，在数值上逐渐接近南大西洋。

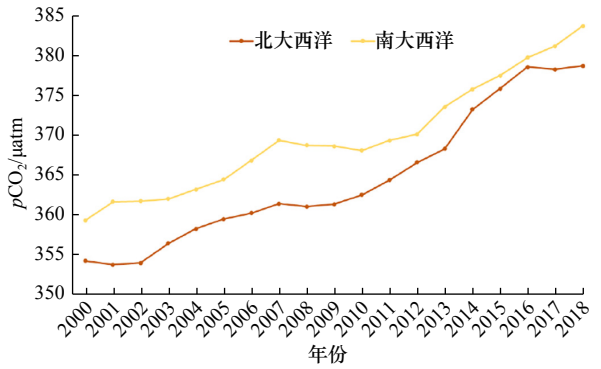
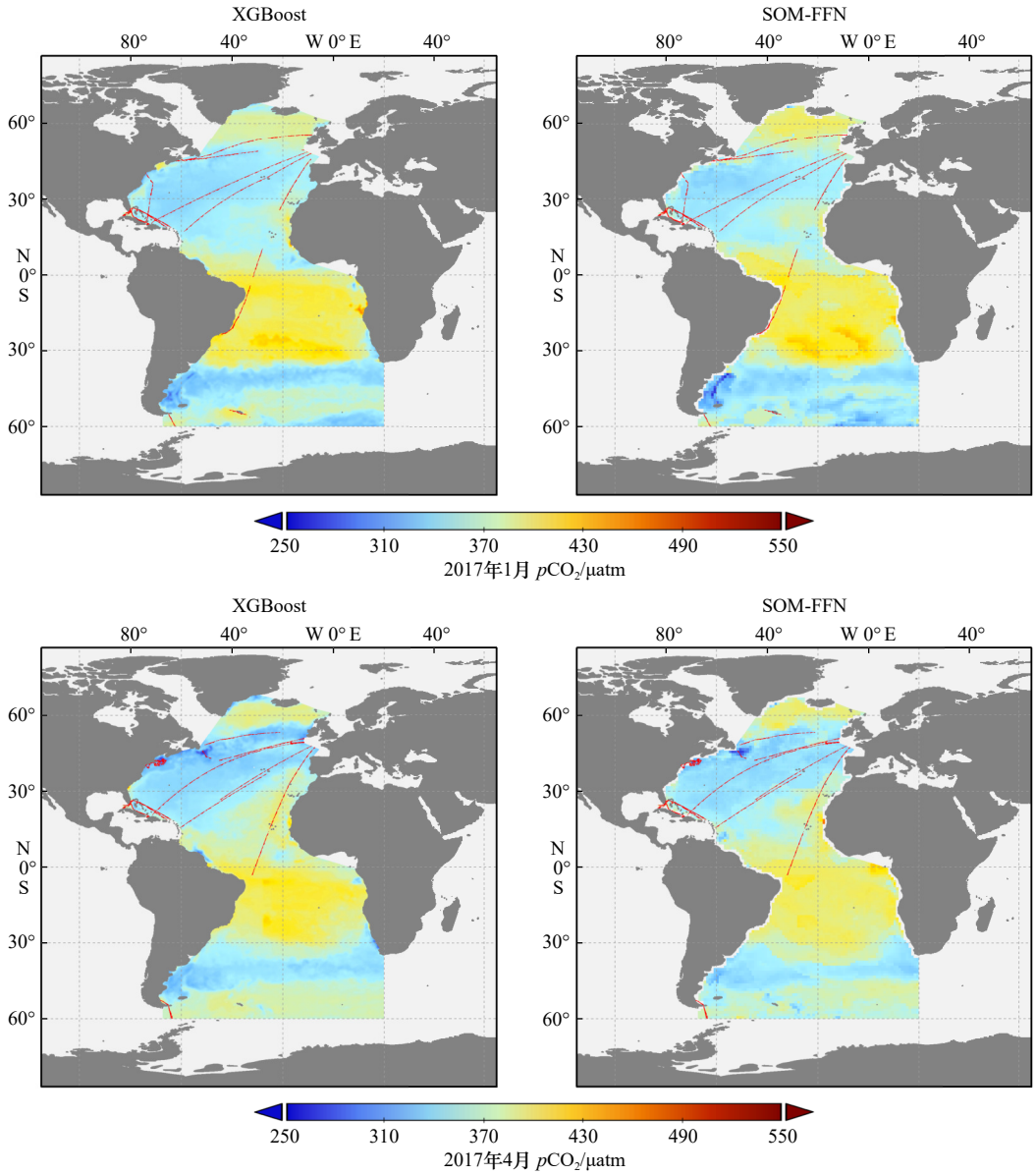


图 10 2000−2018 年南北海域海水二氧化碳分压

Fig. 10 pCO_2 of sea water for south and north sea area

图 11 展示了 2017 年大西洋海水表层二氧化碳分压观测样本与重构结果在 4 个季节（春：4 月、夏：7 月、秋：10 月、冬：1 月）的时空变化以及与 SOM-FFN[11] 方法重构结果的对比。从空间分布上看，分压值在赤道低纬地区较高，在南半球中纬地区较低；高压值也普遍分布在靠近大陆的沿岸区域。在时序变化方面，大西洋内部存在较大的季节性差异。大西洋赤道沿线分压值全年保持在较高水平，数值和范围与太阳直射点的变化相关；高压值也稳定分布在北美大陆西南岸和非洲大陆西岸的近岸海域。南大西洋中纬度海洋中心海域则稳定保持较低水平，一定程度上归因于较少的观测点以及远离陆地带来的影响。其他区域则存在明显的季节性变化，表现为中低纬地区



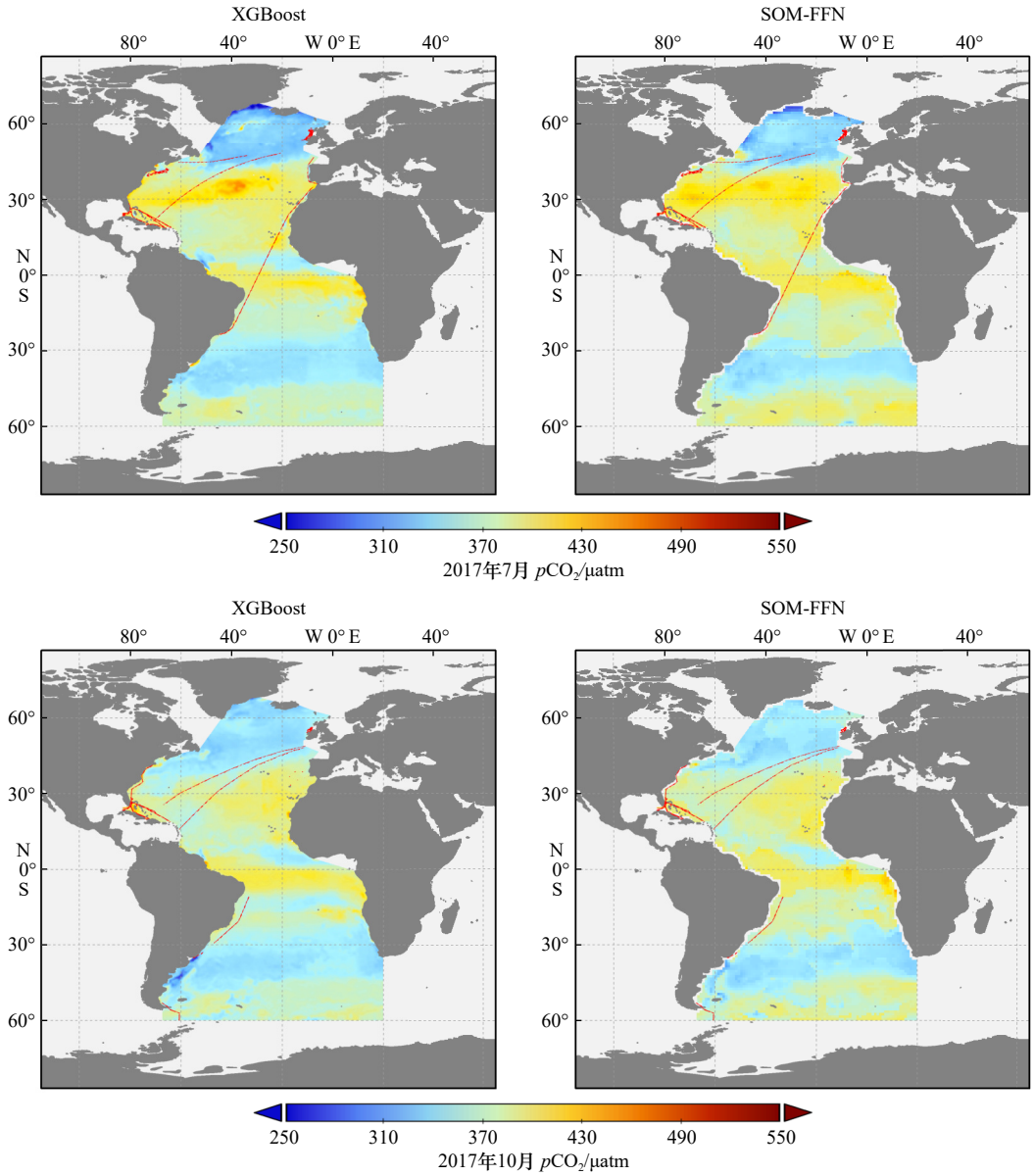


图 11　大西洋海水表层二氧化碳分压观测数据及重构结果(红色点表示观测样本)

Fig. 11　Obeservations and reconstruction result of sea surface pCO_2 in Atlantic (red points present for obeservation samples)

分压值夏高冬低, 高纬度地区则趋势相反。其中, 北大西洋高纬区域的季节性变化最为显著。与 SOM-FFN 产品对比, 可以发现两者在空间分布和变化趋势上十分一致, 佐证了重构结果数值分布的合理性。

表 5 列举了 XGBoost 与 SOM-FFN 重构结果与观测数据在 2017 年 4 个时次的均方根误差, 结果显示 XGBoost 在 4 个季节的误差值均较低。其次, 在预测效果上, 高空间分辨率重构结果能够更清晰、更全面地拟合二氧化碳分压的时空变异性。

图 12 展示了 29°N 以南、60°W 以西观测样本数量较多的巴哈马及北美大陆西部近岸海域的重构结果及与区域内预测样本的绝对偏差(Bias)。整体上看, XGBoost 与 SOM-FFN 在此局部区域的重构结果同样具有较高的一致性, 但 XGBoost 重构结果的绝对偏差小于 SOM-FFN 重构结果的绝对偏差, 且能够更细致地表现海水表层二氧化碳分压在时间和空间上

表 5　重构结果均方根误差

Table 5　RMSE of reconstruction result

时次	RMSE/μatm	
	XGBoost	SOM-FFN
2017年1月	4.48	14.97
2017年4月	4.99	16.21
2017年7月	5.44	16.29
2017年10月	3.31	11.76

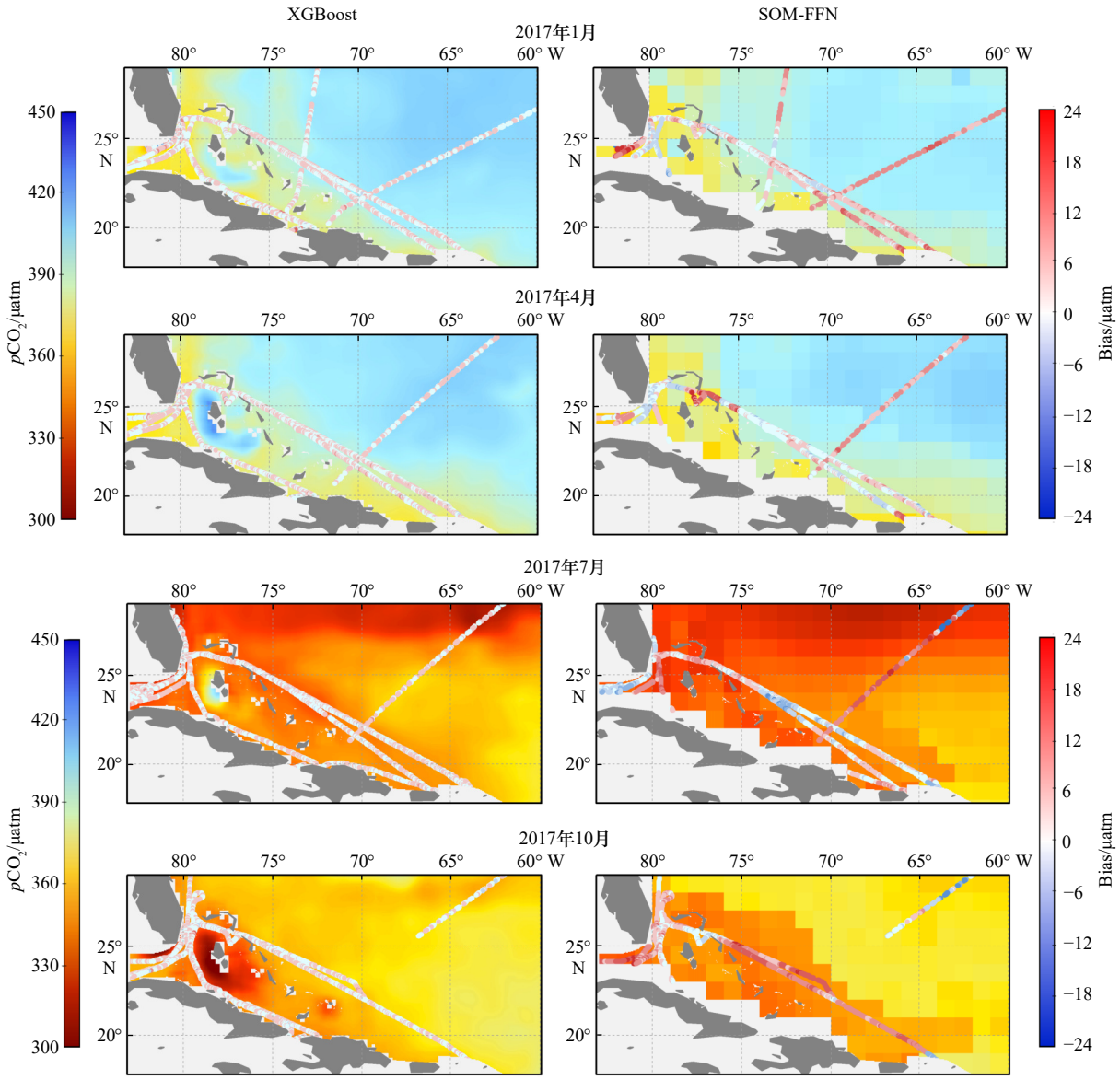


图 12 巴哈马海域海水表层二氧化碳分压重构结果(点数据表示绝对偏差)

Fig. 12 Reconstruction result of sea surface pCO_2 in Bahamas sea area (points present the Bias)

的变化。其次,高空间分辨率下重构结果在近岸海域的覆盖度有所提高,能够更加充分利用近岸的观测样本。同时,由于 SOM-FFN 是基于子区域划分的预测,重构结果中存在一定的突变边界和图斑化现象。

5 结论

本文以高空间分辨率遥感卫星数据为核心,结合再分析、模式等多源数据,基于 XGBoost 模型重构了 2000−2018 年大西洋 0.041 7°×0.041 7°海水表层二氧化碳分压分布。基于 XGBoost 精度高、可解释性强、可处理特征稀疏值等特点,建立了海表面温盐及部分光学特性、混合层深度、叶绿素 *a* 浓度、干燥大气二氧化碳摩尔分数、海平面 10 m 风速以及时空要素与海水二氧化碳分压值的非线性关系重构。其次,针对海水表层二氧化碳观测样本分布不均匀的问题建立了样本的时空权重模型以提高模型学习的平衡性。最终,模型预测的 R^2=0.966, RMSE=8.078 μatm, AD=4.012 μatm,且在不同区域和时次上均取得了较高的精度。在高空间分辨率下,观测数据格网化造成的误差有所下降,重建结果与同类低空间分辨率产品的对比证明了结果的可靠性以及在表现海水表层二氧化碳分压时空变化方面的优势。

此外,本研究仍有许多值得展望之处。首先,尽管遥感卫星数据在海洋环境相关研究中具有较大优势,但云覆盖、算法失效的影响会造成一定的数据缺失。其次,高空间分辨率遥感数据大多诞生于 21 世

纪后, 可获得数据的时间序列较短, 因此无法在高空间分辨率下重构更早期的海水表层二氧化碳分压分布。而在观测数据集方面, 尽管 SOCAT 已收集整理了大量海水表层二氧化碳观测样本, 但在时间和空间上的分布不均匀为重构长时序的海水表层二氧化碳分压分布带来了困难, 如何更好地提取不平衡数据集中的信息以更好拟合高空间分辨率下海水表层二氧化碳分压的时空变化是重构工作的重点与难题。

参考文献:

[1] Hannah L. Chapter 2—The Climate System and Climate Change[M]. London: Academic Press, 2011: 13−52.

[2] Friedlingstein P, Jones M W, O'Sullivan M, et al. Global carbon budget 2021[J]. Earth System Science Data, 2022, 14(4): 1917−2005.

[3] Bai Yan, Cai Weijun, He Xianqiang, et al. A mechanistic semi-analytical method for remotely sensing sea surface pCO_2 in river-dominated coastal oceans: a case study from the East China Sea[J]. Journal of Geophysical Research: Oceans, 2015, 120(3): 2331−2349.

[4] 邱爽, 叶海军, 张玉红, 等. 基于航次观测和再分析资料的南海海表二氧化碳分压反演及变化机制分析[J]. 热带海洋学报, 2022, 41(1): 106−116.
Qiu Shuang, Ye Haijun, Zhang Yuhong, et al. Multi-linear regression of partial pressure of sea-surface carbon dioxide in the South China Sea and its mechanism[J]. Journal of Tropical Oceanography, 2022, 41(1): 106−116.

[5] Chen Shuangling, Hu Chuanmin, Byrne R H, et al. Remote estimation of surface pCO_2 on the West Florida Shelf[J]. Continental Shelf Research, 2016, 128: 10−25.

[6] Nurdjaman S. Estimation of partial pressure of CO_2 (pCO_2) around mount Krakatau waters, Sunda Straits, Indonesia[J]. Borneo Journal of Marine Science and Aquaculture (BJoMSA), 2021, 5(1): 25−31.

[7] Chen Shuangling, Hu Chuanmin, Barnes B B, et al. A machine learning approach to estimate surface ocean pCO_2 from satellite measurements[J]. Remote Sensing of Environment, 2019, 228: 203−226.

[8] Dixit A, Lekshmi K, Bharti R, et al. Net sea-air CO_2 fluxes and modeled partial pressure of CO_2 in open ocean of bay of Bengal[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(7): 2462−2469.

[9] Wang Yanjun, Li Xiaofeng, Song Jinming, et al. Carbon sinks and variations of pCO_2 in the Southern Ocean from 1998 to 2018 based on a deep learning approach[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 3495−3503.

[10] Takahashi T, Sutherland S C, Wanninkhof R, et al. Climatological mean and decadal change in surface ocean pCO_2, and net sea-air CO_2 flux over the global oceans[J]. Deep-Sea Research Part II: Topical Studies in Oceanography, 2009, 56(8/10): 554−577.

[11] Landschützer P, Gruber N, Bakker D C E. Decadal variations and trends of the global ocean carbon sink[J]. Global Biogeochemical Cycles, 2016, 30(10): 1396−1417.

[12] Krishna K V, Shanmugam P, Nagamani P V. A multiparametric nonlinear regression approach for the estimation of global surface ocean pCO_2 using satellite oceanographic data[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 6220−6235.

[13] Chen Tianqi, Guestrin C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 785−794.

[14] Bakker D C E, Pfeil B, Landa C S, et al. A multi-decade record of high-quality fCO_2 data in version 3 of the Surface Ocean CO_2 Atlas (SOCAT)[J]. Earth System Science Data, 2016, 8(2): 383−413.

[15] Dickson A, Sabine C L, Christian J R. Guide to best practices for ocean CO_2 measurements[R]. Sidney: North Pacific Marine Science Organization, 2007.

[16] Sathyendranath S, Brewin R J W, Brockmann C, et al. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (OC-CCI)[J]. Sensors, 2019, 19(19): 4285.

[17] Menemenlis D, Campin J M, Heimbach P, et al. ECCO2: High resolution global ocean and sea ice data synthesis[J]. Mercator Ocean Quarterly Newsletter, 2008, 31: 13−21.

[18] Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis[J]. Quarterly Journal of the Royal Meteorological Society, 2020, 146(730): 1999−2049.

[19] Peters W, Jacobson A R, Sweeney C, et al. An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker[J]. Proceedings of the National Academy of Sciences, 2007, 104(48): 18925−18930.

[20] Sabine C L, Hankin S, Koyuk H, et al. Surface Ocean CO_2 Atlas (SOCAT) gridded data products[J]. Earth System Science Data, 2013, 5(1): 145−153.

[21] Bates N R. Interannual variability of the oceanic CO_2 sink in the subtropical gyre of the North Atlantic Ocean over the last 2 decades[J]. Journal of Geophysical Research: Oceans, 2007, 112(C9): C09013.

[22] González-Dávila M, Santana-Casiano J M. Carbon dioxide, temperature, salinity and other variables collected via time series monitoring from METEOR, POSEIDON and others in the North Atlantic Ocean from 1995−10−02 to 2009−11−25 (NCEI Accession 0100064)[Z]. Dataset: NOAA National Centers for Environmental Information, 2012.

Reconstruction of sea surface pCO_2 with high resolution: A case study of the Atlantic Ocean

Wang Xinyi [1, 2, 3], Wu Chuyi [1, 2, 3], Wu Sensen [1, 2, 3], Chen Yijun [1, 2, 3], Du Zhenhong [1, 2, 3]

(1. *School of Earth Sciences*, *Zhejiang University*, *Hangzhou* 310027, *China*; 2. *Zhejiang Provincial Key Laboratory of Geographic Information System*, *Hangzhou* 310030, *China*; 3. *Department of Geographic and Spatial Information Science*, *Zhejiang University*, *Hangzhou* 310027, *China*)

Abstract: Ocean is an important carbon sink in nature. The sea-air carbon dioxide flux is usually estimated by the difference of partial pressure of carbon dioxide (pCO_2) between the atmosphere and the sea surface. Due to the imbalance of observation data on temporal and spatial distribution and datasets used for prediction, there is still large room for improvement in spatial resolution for present reconstruction of pCO_2 on sea surface. In order to fit the temporal and spatial variability under high spatial resolution better, based on the sea surface fugacity of carbon dioxide (fCO_2) observations of the Surface Ocean CO_2 Atlas (SOCAT) and other multi-source data including remote sensing data, the nonlinear relationship between sea surface pCO_2 and physical, biological, optical factors was established by a XGBoost model and a weight model was built based on spatiotemporal frequency of samples. A 0.041 7°×0.041 7° monthly sea surface pCO_2 dataset in Atlantic from 2000 to 2018 was finally constructed with correlation coefficient of 0.966, mean squared error of 8.087 μatm and mean error of 4.012 μatm on prediction dataset. The reconstruction is highly consistent to other similar reconstruction results on temporal and spatial trend and also gains advantage in spatial resolution.

Key words: sea surface carbon dioxide partial pressure; remote sensing; high spatial resolution