

基于集成学习方法的中国近地面臭氧浓度时空分布

宋世鹏^{1,2}, 范萌², 陶金花², 陈三明¹, 顾坚斌², 韩宗甫²,
梁晓霞⁴, 陆晓艳⁵, 王甜甜⁶, 张莹²

1. 桂林理工大学 地球科学学院, 桂林 541000;

2. 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;

3. 中国科学院大学, 北京 100049;

4. 首都师范大学 资源环境与旅游学院, 北京 100089;

5. 广西壮族自治区生态环境监测中心, 南宁 530028;

6. 江苏省环境监测中心, 南京 210019

摘要: 自2013年大气污染防治行动以来, $\text{PM}_{2.5}$ 、 PM_{10} 、 SO_2 、 NO_2 等空气污染物浓度都有不同程度下降, 但臭氧污染仍有上升趋势, 臭氧污染已成为制约中国空气质量持续改善的关键问题。地基点可以提供空间上特定点的臭氧浓度, 但无法获得近地面臭氧连续的空间分布。由于臭氧大量分布于平流层, 遥感卫星反演的臭氧柱浓度产品仅能反映整层臭氧柱浓度, 但整层臭氧柱浓度与近地面浓度无明显相关性, 因而无法体现近地面臭氧浓度。本文综合地基监测数据、再分析资料、卫星产品, 采用不同的模型方法, 得到近地面臭氧浓度的时空分布, 结果表明集成学习方法可以准确估算近地面臭氧在空间上的分布状况和在时间上的变化趋势。本文对比了梯度提升回归树 (GBRT)、极端随机树 (ERT)、极端梯度提升器 (XGBoost) 3种不同的集成学习方法在近地面臭氧污染估算的效果表现, 3种集成学习方法在2019年—2020两年的十折交叉验证 R^2 都在0.89以上, 极端梯度提升器 (XGBoost) 方法在RMSE、MAE指标上有最好的表现, 2019年—2020年两年的平均RMSE、MAE分别为 $15.77 \mu\text{g}/\text{m}^3$ 、 $10.53 \mu\text{g}/\text{m}^3$, 但基于极端随机树 (ERT) 方法获得的近地面臭氧空间分布更加连续自然。因此最终选择极端随机树 (ERT) 方法估算得到中国近地面臭氧浓度数据集, 并在此基础上进行时空分析。由于中国政府实施积极的减排措施的及疫情影响, 臭氧浓度多年来的上升趋势得到了逆转, 2020年臭氧年均值为 $107.41 \pm 18.6 \mu\text{g}/\text{m}^3$, 较去年平均值 $109.26 \pm 19.71 \mu\text{g}/\text{m}^3$, 有所减少。时间上, 每年的5—9月气温较高, 光化学反应剧烈, 因而臭氧高污染事件频繁发生。空间上, 京津冀地区、长三角地区、珠三角地区、成渝地区等城市群显著高于周围其他区域, 是臭氧污染防治的重点区域。

关键词: 遥感, 近地面臭氧, 集成学习, 极端随机树, 梯度回归提升树, 极端梯度提升器, TROPOMI, 新冠疫情
中图分类号: P2

引用格式: 宋世鹏, 范萌, 陶金花, 陈三明, 顾坚斌, 韩宗甫, 梁晓霞, 陆晓艳, 王甜甜, 张莹. 2023. 基于集成学习方法的中国近地面臭氧浓度时空分布. 遥感学报, 27(8): 1792–1806

Song S P, Fan M, Tao J H, Chen S M, Gu J B, Han Z F, Liang X X, Lu X Y, Wang T T and Zhang Y. 2023. Estimating ground-level ozone concentration in China using ensemble learning methods. National Remote Sensing Bulletin, 27(8): 1792–1806 [DOI: 10.11834/jrs.20231845]

1 引言

19世纪伦敦光化学烟雾事件的发生, 引起了人们对臭氧及其前体物研究的重视 (Haagen-Smit, 1952)。大气中90%的臭氧存在于平流层, 通过吸

收绝大部分太阳紫外辐射保护地球生物圈安全和调节气候变化。近地面臭氧除少量来源于平流层垂直输送外, 绝大部分由人为排放的氮氧化物 (NO_x) 和挥发性有机化合物 (VOCs) 在太阳辐射调节下经一系列光化学反应转化而来。长时间的

收稿日期: 2022-01-06; 预印本: 2023-04-19

基金项目: 国家重点研发计划 (编号: 2018YFC0214003); 国家自然科学基金 (编号: 41771391, 42171393, 41871254); 广西重点研发计划 (编号: AB20238015)

第一作者简介: 宋世鹏, 研究方向为大气环境遥感。E-mail: 18366840667@163.com

通信作者简介: 张莹, 研究方向为大气痕量气体定量遥感。E-mail: zhangying01@radi.ac.cn

高浓度臭氧暴露,可引起人类呼吸系统疾病、心血管疾病、神经系统疾病等,对公众健康造成重要影响(Bell, 2006)。2013年国务院颁布“大气污染防治行动计划”以来,中国大气污染防治工作取得了巨大成就,大部分地区年均细颗粒物浓度有大幅下降,然而臭氧污染却呈现快速上升和蔓延趋势。因此,开展大尺度区域近地面臭氧浓度监测和时空分布特征变化分析,将为实现臭氧与PM_{2.5}协同防治提供数据支持和决策支撑。

地基站点观测是常见的近地面臭氧浓度监测手段,Gardner和Dorling(2000)最早通过线性回归、回归树、多层感知机等统计学方法建立气象要素及前体物与臭氧之间的关系。陈仁杰等(2010)利用上海市环境保护部门的近地面臭氧监测数据,与居民健康相结合,研究了上海市近地面臭氧对健康的影响。王闯等(2015)对沈阳市臭氧浓度变化特征进行了分析,并将其与气象资料结合用以判断气象因子与臭氧浓度的关系。

相比地面监测来说,卫星监测臭氧柱浓度不仅可以不受地面监测站点的选址的限制,而且可以全天时、全天候获取信息,具有地面监测站点无可比拟的时间和空间覆盖度(陈良富等, 2019)。1970年,美国研制的最早一代臭氧垂直分布探测仪BUV(Solar Backscatter Ultraviolet Instruments)搭载在雨云4(Nimbus-4)卫星上。1995年欧洲航天局研制了第一台高光谱臭氧观测仪—全球臭氧监测试验装置GOME(Global Ozone Monitoring Experiment)。美国于2002年和2004年相继又发射了Aqua和Aura卫星,其分别搭载了大气红外探测仪AIRS(Atmospheric Infrared Sounder)和臭氧监测仪OMI(Ozone Monitoring Instrument)。中国在2008年发射了搭载有紫外臭氧垂直探测仪SBUS(Solar Backscatter Ultraviolet Sounder)和臭氧总量探测仪TOU(Total Ozone Unit)的新一代极地轨道气象卫星风云三号(FY-3)。Duncan等(2010)利用OMI数据,划分了美国不同城市为VOCs或NO_x控制区,为臭氧控制策略提供了依据。张艳等(2013)利用风云三号卫星搭载的紫外臭氧总量探测仪(TOU)对1979年—2011年北半球春季臭氧特征进行了分析,发现北极臭氧年际变化显著,与平流层温度变化一致。

目前存在的估算近地面臭氧浓度的估算方法有空间内插法、化学传输模型和基于统计学模型

的方法。空间插值方法是通过已有站点的污染物浓度分布直接对站点外进行估算,优点是计算成本低,但通常会产生阶梯状的分布,不能充分地表征局部可变性。化学传输模型可以很好的考虑臭氧的产生和消耗过程,但需要大量的特征输入和复杂的运算过程,并且由于排放源清单数据的滞后性往往会导致化学传输模型精度欠佳。统计学模型不直接针对污染物扩散途中复杂的物理和化学过程建模,而是凭借足量的历史监测数据,选取特征如空间位置、时间、气象、人类活动、污染物前体物等对近地面污染物浓度进行建模估算。传统的近地面臭氧统计模型包括线性回归模型、广义可加模型、支持向量机方法等。近年来,一些新兴机器学习方法和深度学习方法被广泛用于近地面臭氧估算研究中,Zhan等(2018)利用随机森林算法对中国近地面臭氧浓度做了估算,验证精度 $R^2=0.69$,RMSE=26 $\mu\text{g}/\text{m}^3$,并分析了影响臭氧浓度的气象因素。方韬(2020)利用卷积神经网络对中国近地面臭氧浓度进行了重建,验证精度 R^2 为0.65。

由于平流层臭氧浓度在整层臭氧浓度中占据主导地位,整层臭氧柱浓度产品在估算近地面臭氧浓度中只能提供极少量信息,这使得通过卫星柱浓度产品获得近地面臭氧浓度无法实现。本文将与臭氧化学过程相关的NO₂、HCHO、CO等卫星对流层柱浓度产品引入并辅助详细的气象参数和地面信息,进行近地面臭氧反演工作。利用集成学习方法开展中国地区2019年和2020年近地面臭氧浓度估算,并分析讨论了中国重点城市群地区以及新冠肺炎疫情管控前中后近地面臭氧浓度的时空变化特征。

2 数据与预处理

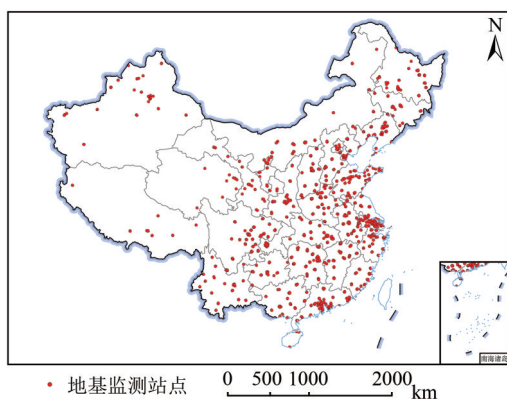
2.1 地面观测数据

近地面臭氧浓度地面观测数据来自于中国环境监测总站CNEMC(China National Environmental Monitoring Center)1473个地基站点(图1)的臭氧浓度实测值,时间分辨率为一小时,站点覆盖全国(台湾省资料暂缺)、自治区、直辖市。

2.2 卫星观测资料

考虑到不同下垫面类型对臭氧生成、扩散过

程差异性较大, 选择MODIS归一化植被指数产品MYD13A3和地表反射率产品MYD09A1作为地表特征参量来表征下垫面的差异性。其中, MYD13A3数据为1 km空间分辨率的归一化植被指数月均值产品。MYD09A1数据为500 m空间分辨率的8 d地表反射率产品, 该产品选择8 d内高观测覆盖率、低视角、没有云或云阴影以及气溶胶负载的像元作为最佳观测。



审图号:GS京(2023)0780号

图1 中国臭氧监测站点空间分布图

Fig. 1 Spatial distribution of ozone monitoring stations in China

由于卫星臭氧柱浓度监测结果中能够反映的近地面信息非常有限, 本文在选择特征参量时舍弃了臭氧柱浓度, 而是选择了与臭氧生成密切相关的前体物 NO_2 和HCHO, 以及具有一定催化作用的CO (Chin等, 1994)。 NO_2 、HCHO和CO训练数据分别来自于Sentinel-5P/TROPOMI的对流层 NO_2 和HCHO柱浓度产品、整层CO柱浓度产品。其中, NO_2 和HCHO数据空间分辨率为 $5.5 \text{ km} \times 3.5 \text{ km}$, CO数据空间分辨率为 $7 \text{ km} \times 5.5 \text{ km}$ 。

2.3 模式再分析资料

为表征辐射和云对臭氧光化学反应的影响, 以及复杂气象条件对臭氧扩散传输的影响, 本研究获取了欧洲中期天气预报中心发布的第五代全球气候再分析数据集ERA5 ([https://cds.climate.copernicus.eu/\[2022-01-06\]](https://cds.climate.copernicus.eu/[2022-01-06])), 并提取其中的辐射、气象、云和气体参数: 地表太阳辐射下方向(ssrd)、地表热辐射下方向(strd)、地表紫外辐射(uvb)、相对湿度(rh)、地表两米温度(t2 m)、地表压强(sp)、10 m风U分量(u10)、10 m风V分量(v10)、蒸发(e)、总降雨(tp)、云底高(uvb)、云层液态水柱浓度(tclw)、云覆盖总量(tcc)、臭氧整层柱浓

度(tco3)、臭氧东向通量垂直积分(vieo)、臭氧北向通量垂直积分(vino)。ERA5再分析资料时间分辨率为1 h, 空间分辨率为 0.25° 。

2.4 数据预处理

构建特征样本集的时间范围为2019年—2020年, 所有用于模型训练的备选特征参量详细信息如表1所示。地基监测数据根据环境空气质量评价技术规范(HJ663-2013), 本文将每天的8 h滑动 O_3 平均值的最大值作为集成学习模型的标签值。TROPOMI数据筛选 NO_2 、HCHO和CO柱浓度产品数据质量字段大于50的有效像元。ERA5数据选择与卫星过境时间邻近的北京时间13、14和15等3个时刻的数据, 并做空间均值处理。TROPOMI、MODIS、ERA5数据均重采样至 0.1° , 并统一至WGS-84地理坐标系。

3 研究方法

3.1 传统方法

本文采用的传统统计方法包括线性回归方法(Linear Regression), K近邻方法(k Nearest Neighbor)和弹性网络方法(Elastic Net)。其中线性回归方法是最常用最简单的回归方法, 通过不同的权重值建立多个特征变量与目标之间的线性方程, 线性回归模型有充分的可解释性。K近邻算法的思想核心是寻找与预测样本最为接近的 k 个已知样本, 通过这 k 个已知样本去实现对预测样本的回归问题。K近邻算法核心是 k 值的选择, 过小的 k 值会导致过拟合问题, k 值过大时与输入实例较远的(不相似)训练实例也会对预测起作用, 降低了预测的准确性。弹性网络是一种使用L1、L2范数作为先验正则项训练的线性回归模型, 并且集合了岭回归和Lasso回归两种方法的, 并且弹性网络还有解恒有效性和收敛耗时少的优势。

3.2 集成学习方法

集成学习方法是把多个使用给定学习算法构建的基估计器的预测结果结合起来, 从而获得比单个估计器预测更好的泛化能力和鲁棒性。集成学习的常用的方法有Bagging方法和Boosting方法。Bagging方法对所有样本采取有放回性的抽取方法, 得到 k 个训练集。对这 k 个训练集, 独立训练 k 个弱学习器, 最后对 k 个模型的均值作为最后的结

果。Boosting 算法是用初始权重训练初始弱学习器，对此分类器进行 n 次迭代，每次迭代会减小在上一轮正确的样本的权重，增大错误样本的权重，最终得到一个强分类器。

表1 特征信息表
Table 1 Feature information table

数据类别	简称	变量名称	英文名称	单位	时间分辨率	空间分辨率	数据源
地面监测	O ₃	臭氧近地面浓度	Ozone	μg/m ³	1 h	—	CNEMC
地表特征	NDVI	归一化植被指数	Normalized difference vegetation index	—	1 月	500 m	MODIS
	sr	地表反射率	Surface reflectance	—	8 d	500 m	MODIS
云特征	cbh	云底高	Cloud base height	m	1 h	0.25°×0.25°	ERA5
	tc1w	云层液态水柱浓度	Total column cloud Liquid water	kg/m ⁻²	1 h	0.25°×0.25°	ERA5
	tcc	云覆盖总量	Total cloud cover	—	1 h	0.25°×0.25°	ERA5
气象特征	rh	相对湿度	Relative humidity	—	1 h	0.25°×0.25°	ERA5
	t2m	地表两米温度	2 m temperature	K	1 h	0.25°×0.25°	ERA5
	sp	地表压强	Surface pressure	Pa	1 h	0.25°×0.25°	ERA5
	u10	10 米风 U 分量	10 m U-component of wind	m/s ⁻¹	1 h	0.25°×0.25°	ERA5
	v10	10 米风 V 分量	10 m V-component of wind	m/s ⁻¹	1 h	0.25°×0.25°	ERA5
	blh	边界层高度	Boundary layer height	m	1 h	0.25°×0.25°	ERA5
	e	蒸发	Evaporation	mm	1 h	0.25°×0.25°	ERA5
	tp	总降雨	Total precipitation	m	1 h	0.25°×0.25°	ERA5
	ssrd	地表太阳辐射下方向	Surface solar radiation downwards	J/m ⁻²	1 h	0.25°×0.25°	ERA5
辐射特征	strd	地表热辐射下方向	Surface thermal radiation downwards	J/m ⁻²	1 h	0.25°×0.25°	ERA5
	uvb	地表紫外辐射下方向	Downward UV radiation at the surface	J/m ⁻²	1 h	0.25°×0.25°	ERA5
气体特征	NO ₂	对流层 NO ₂ 柱浓度	NO ₂ tropospheric column concentration	mol/cm ²	1 d	5.5 km×3.5 km	TROPOMI
	HCHO	对流层 HCHO 柱浓度	HCHO tropospheric column concentration	mol/cm ²	1 d	5.5 km×3.5 km	TROPOMI
	CO	整层 CO 柱浓度	CO total column concentration	mol/cm ²	1 d	7 km×5.5 km	TROPOMI
	vieo	臭氧东向通量的垂直积分	Vertical integral of eastward ozone flux	kg/m ⁻¹ s ⁻¹	1 h	0.25°×0.25°	ERA5
	vino	臭氧北向通量的垂直积分	Vertical integral of northward ozone flux	kg/m ⁻¹ s ⁻¹	1 h	0.25°×0.25°	ERA5
	tco3	臭氧整层柱浓度	Total column ozone	kg/m ⁻²	1 h	0.25°×0.25°	ERA5

极端随机树（ERT）是一种重要的基于 Bagging 的集成学习方法（Geurts 等，2006），可以用来处理分类、回归等问题。随机森林从原始训练样本集 N 中有放回地重复随机抽取 k 个样本生成新的训练样本集合，然后根据自助样本集生成 k 个分类树组成随机森林，新数据的分类结果按分类树投票结果决定（Breiman，2001）。极端随机树同样是一种多棵决策树集成的分类器，与随机森林分类器比较，主要有两点不同：第一极端随机树一般不采用随机采样，即每个决策树采用原始训练集。第二是端随机树特征是随机选取的，因为分裂是随机的，所以在某种程度上比随机森林得到的结果更加好。梯度提升回归树（GBRT）属于 Boosting 方法，使用了前向分布算法，但是弱学习器限定了只能使用 CART 回归树模型（Friedman 等，2000）。前向分布算法的思想是基于当前模型和拟合函数选择合适的决策树函数，从而最小

化损失函数。极端梯度提升器（XGBoost）模型 Boosting 算法中的一种（Chen 和 Guestrin，2016）。对于一组数据，使用过于复杂的模型去拟合，往往会发生过拟合，因而可以用 XGBoost 引入正则化项来限制模型复杂度。XGBoost 中，限制了叶子节点的增长，使得每棵树都是“弱”的，同时还引入了学习速率，进一步降低了每棵树的影响。代价是树的总数目会增加，但是如此可以取得更好的预测效果。

3.3 技术流程图

基于集成机器学习模型估算近地面臭氧浓度的方法流程如图 2 所示。通过将所有备选特征参量数据集进行时空匹配，构建样本集；基于特征参量重要性分析，筛选可参与建模的特征参量；利用筛选后特征参量，进行不同机器学习模型训练，优化调整模型参数，估算中国地区近地面臭氧浓度。

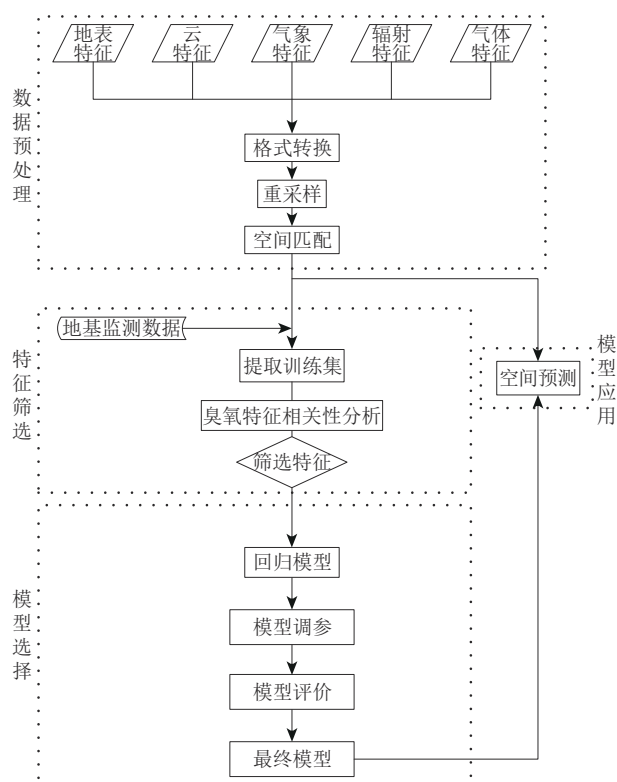


图2 技术流程图

Fig. 2 Technology flow chart

4 特征与模型选择

4.1 臭氧特征相关性分析

通过计算臭氧与所有特征参量间的皮尔森相关系数（图3）、显著性 P 值（图4）和方差膨胀因子（图5）。针对 O_3 前体物，HCHO可以表征VOCs参与光化学反应的程度（Sillman, 1995）， O_3 与HCHO呈正相关关系， $(r=0.62, p<0.01)$ 。 NO_2 柱浓度可以表征一个区域的污染企业生产密集程度，因此 NO_2 与臭氧存在正相关 $(r=0.56, p<0.01)$ 。在所有气象特征中， O_3 与t2 m相关性最强为0.34 $(p<0.01)$ ，因为温度影响臭氧的光化学生成速率，温度越高越有利于光化学反应生成（Dueñas等, 2002）。 O_3 与ssrd的相关系数分别为0.27 $(p<0.01)$ ，ssrd为 O_3 的光化学反应提供能量，加速光化学反应的发生（Feister和Balzer, 1991）。当两个特征参量间相关性过高会造成特征冗余，通过计算方差膨胀因子，ssrd和uvb的方差膨胀因子都超过20，舍弃与 O_3 相关性较低的uvb。此外，由于NDVI、blh、cbh和tco3与 O_3 无明显相关性，也被从样本集中剔除。

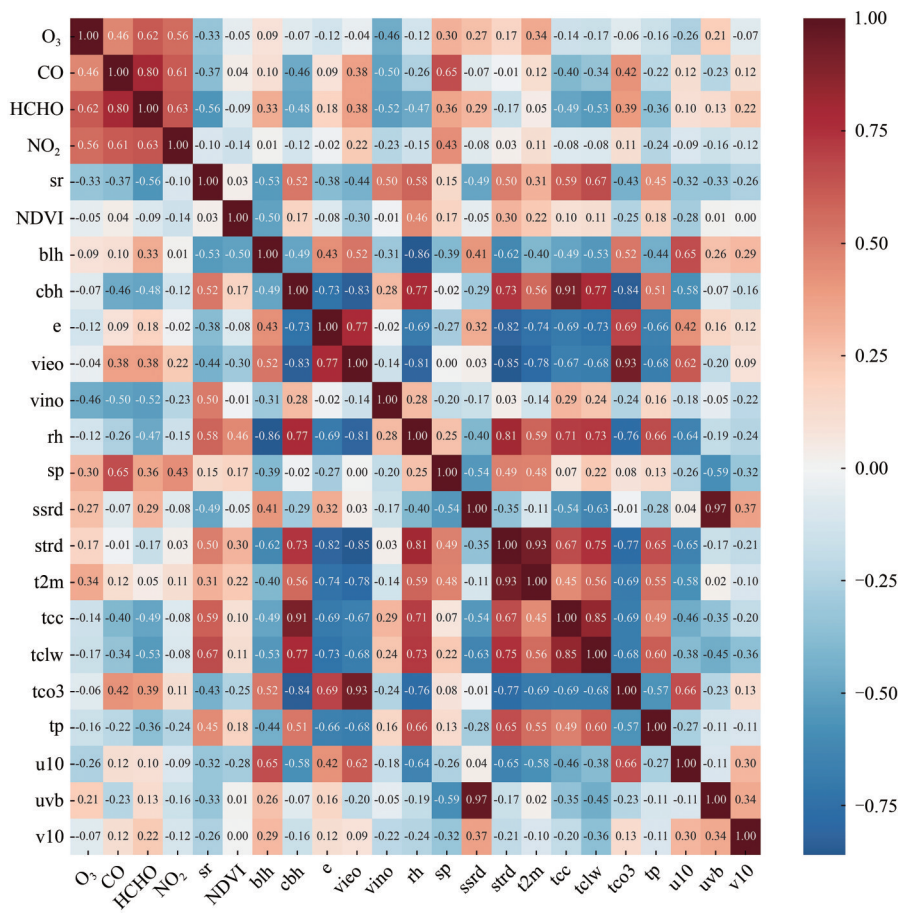


图3 臭氧与特征相关性矩阵

Fig. 3 Ozone and feature correlation matrix

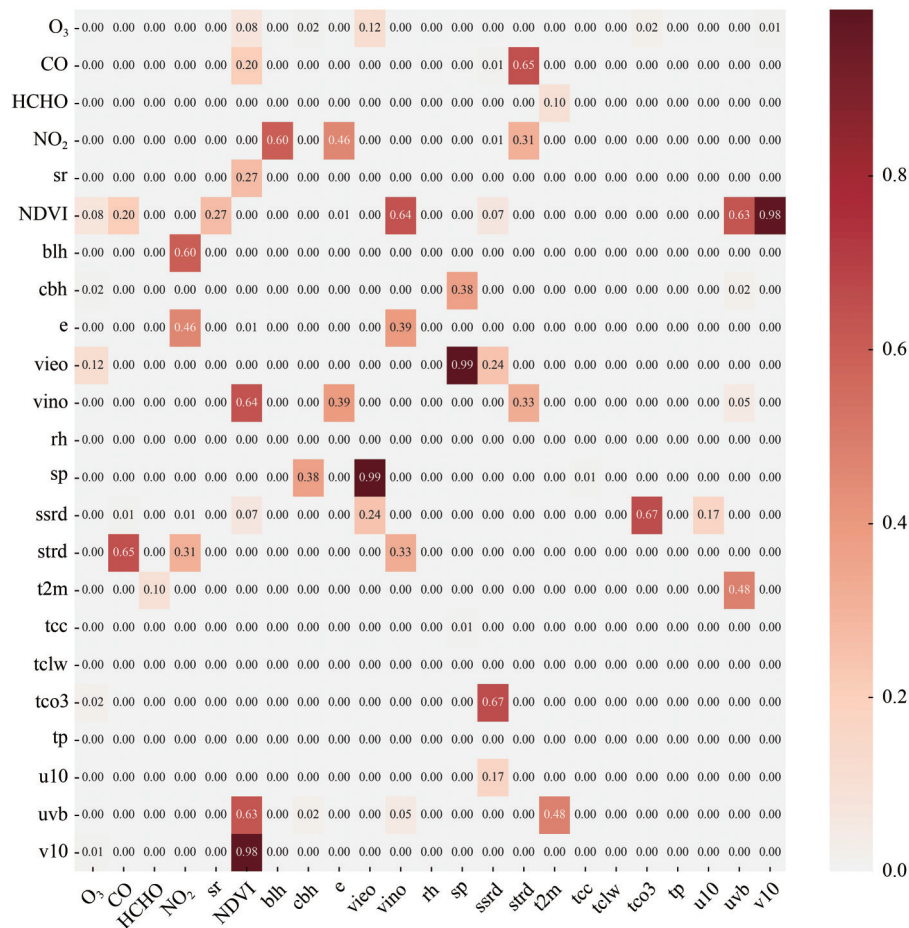


图4 臭氧与特征显著性矩阵

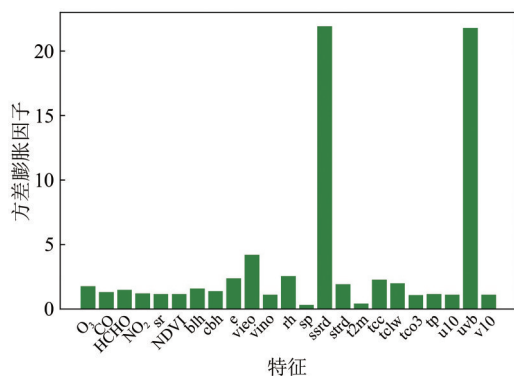
Fig. 4 Ozone and feature p value matrix

图5 臭氧与特征方差膨胀因子

Fig. 5 Variance inflation factor of ozone and feature

4.2 模型对比分析

基于上述特征筛选, 本研究最终选择了CO、HCHO、NO₂、sr、e、v10、vino、rh、sp、ssrd、t2m、tcc、tclw、tp、u10和v10共16个特征参量, 并增加month、day、lon和lat共4个表征时空信息的特征参量参与模型训练。为评价不同的算法模型在反演近地面O₃的精度, 将数据分为训练集和测试

集, 并采用十折交叉验证方法进行精度验证。在模型初步筛选过程中, 发现集成学习方法整体效果显著高于传统方法, 因此在此基础上对集成学习模型的精细调参, 最终得到3个模型的主要参数的最优选择, ERT (n_estimators=200, max_depth=25, min_samples_split=4, min_samples_leaf=1)、GBRT (n_estimators=450, max_depth=14, min_samples_split=8, min_samples_leaf=5)、XGBoost (n_estimators=450, max_depth=15, min_child_weight=8, gamma=0.2)。如图6所示, ERT模型、GBRT模型、XGBoost模型在2019年—2020两年的十折交叉验证 R^2 分别为0.8968、0.904、0.9086。基于XGBoost模型估算的2019年—2020年中国地区近地面臭氧浓度的RMSE、MAE指标综合最优, 2019年—2020年平均RMSE和MAE分别为15.77 $\mu\text{g}/\text{m}^3$ 和10.53 $\mu\text{g}/\text{m}^3$ 。GBRT模型有最大的斜率(0.90)和最小的截距指标(11.34), 这表明GBRT模型在整体拟合效果上优于其他模型。如表2所示, 集成学习方法在模型

训练时间消耗上整体高于传统方法，其中GBRT模型训练时间最长，约为ERT和XGBoost的8—9倍，ERT模型在单位时间内的内存消耗量最大。如图7所示，基于3种模型得出2019年臭氧浓度均值空间分布，在估算近地面臭氧时，基于Boosting思想的XGBoost模型及GBRT模型，空间分布出现明显的经向和纬向的条纹，而基于Bagging思想的ERT模型基本不存在条纹现象，臭氧浓度空间分布更连续自然。此现象出现的原因可能是：（1）GBRT模

型与XGBoost模型都是将弱分类器提升为一个强分类器的模型，最后只有一个分类器参与决策，而ERT模型是若干子模型加权决策最后得到臭氧空间估算值。（2）随机森林特征是随机选取的，因为分裂是随机的所以减小了模型对于经度纬度两个空间特征对近地面臭氧估算的依赖程度。3个集成学习模型在指标差距相对较小，但ERT估算得到的近地面臭氧空间分布更加连续自然，因此，最后选择ERT模型作为最后的预测模型。

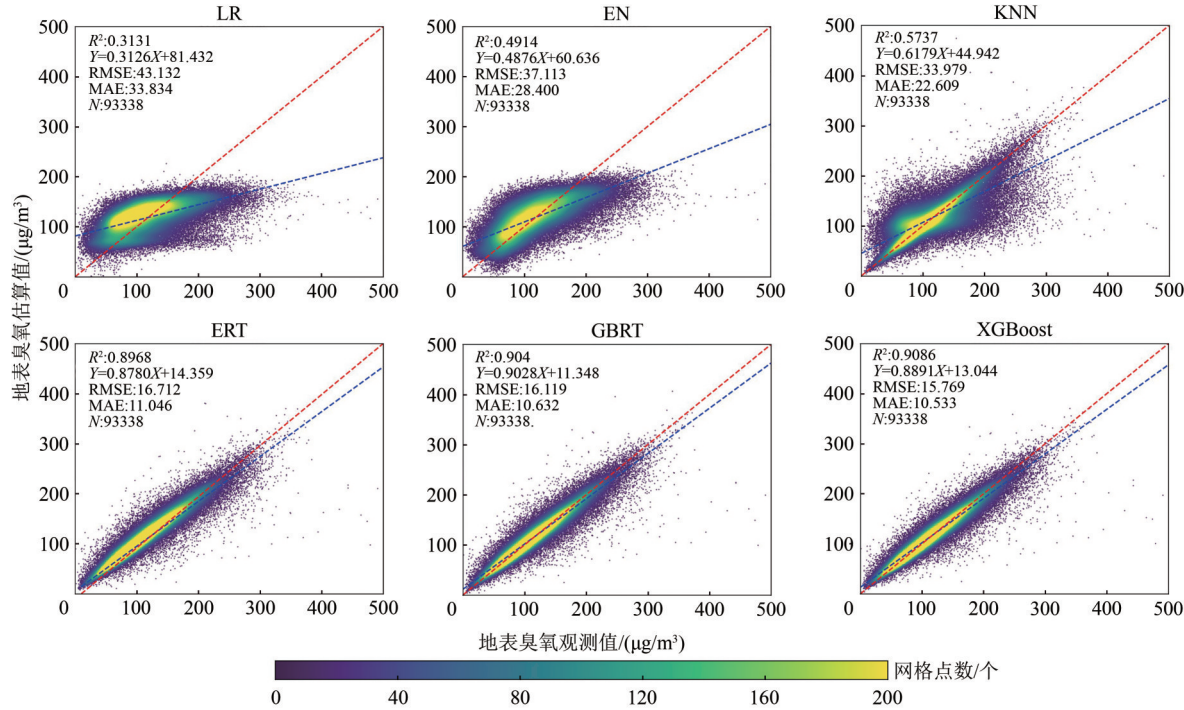


图6 模型估算近地面臭氧浓度散点图
Fig. 6 Scatter plot of near-ground ozone concentration estimated by different models

表2 模型训练时间与内存消耗

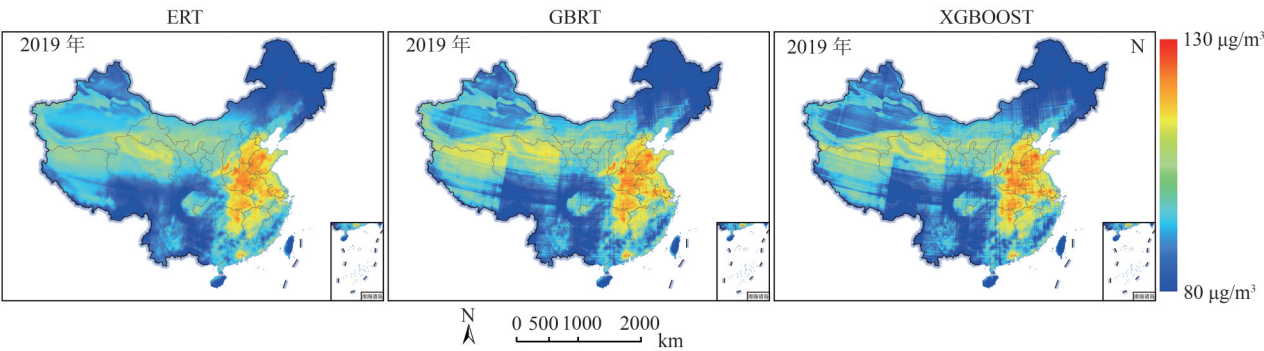
参数	模型					
	LR	EN	KNN	ERT	GBRT	XGBoost
训练用时/s	159	154	143	420	3470	361
平均内存消耗/(mb/s)	401	403	496	2569	489	509

5 结果与讨论

5.1 时空精度验证

为细化分析模型估算近地面臭氧的精度，从时间和空间上对模型的估算能力进行有效的验证。为评价模型在区域尺度上的估算效果，提取全国四大典型区域的训练集，分别进行精度验证，结

果如图8所示，京津冀地区模型估算近地面臭氧的十折交叉验证 R^2 指标最佳为0.94，成渝地区的RMSE和MAE指标最佳，分别为14.81 $\mu\text{g}/\text{m}^3$ 和9.99 $\mu\text{g}/\text{m}^3$ 。为评价模型在时间尺度上的估算效果，分别提取4个季节训练集，进行精度验证，结果如图9所示，臭氧污染高发季节的夏秋季模型 R^2 高于冬春季。



审图号:GS京(2023)0780号

图7 基于ERT、GBRT和XGBoost模型的近地面臭氧浓度空间分布图(2019年均值)

Fig. 7 Spatial distribution of ground-level ozone concentration based on ERT (left), GBRT (middle) and XGBoost (right) models (average value in 2019)

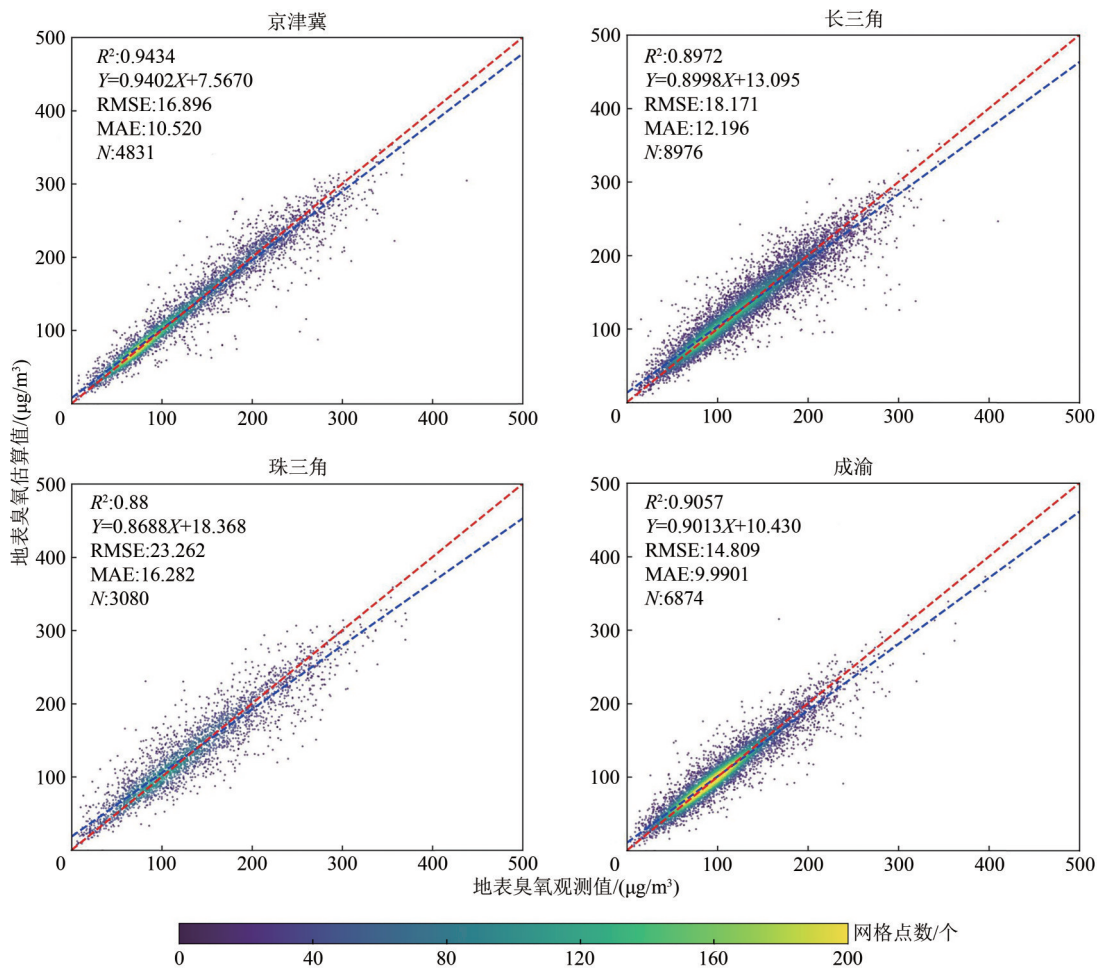


图8 4大城市群近地面臭氧估算散点图

Fig. 8 Scatter plot of estimated near-ground ozone concentration in four major city agglomerations

5.2 臭氧时空分析

基于2019年和2020年ERT训练模型，分别开展2019年和2020年中国地区单日近地面臭氧浓度估算。在单日近地面估算中，由于云污染现象的存在，在反演时首先需要判断像元是否为有云像

元，若该像元为有云像元，则放弃反演，提高反演效率并且也避免了有云状态下的错误反演，如图10所示，云覆盖区域存在臭氧反演空缺值，其余区域估算近地面O₃与地基点臭氧分布空间变化趋势一致。对单日反演结果进行均值计算得到

月份和季节平均臭氧浓度数据集, 单日有云像元不参与运算, 最终结果如图 11 和图 12 所示。中国近地面臭氧浓度存在显著的季节性差异, 冬季 (1—2 月和 12 月) 基本维持在较低水平, 两年中月均最低值分别为 2019 年 12 月 $81.81 \pm 3.29 \mu\text{g}/\text{m}^3$ 和 2020 年 12 月 $76.33 \pm 2.96 \mu\text{g}/\text{m}^3$ 。春季 (3—4 月) 随着温度回升近地面臭氧浓度在全国范围内开始逐渐升高。臭氧高值通常出现在 5—9 月, 其中, 2019 年、2020 年近地面臭氧月均浓度最高值分别出现在 6 月 ($131.16 \pm 3.93 \mu\text{g}/\text{m}^3$) 和 5 月 ($127.49 \pm$

$5.23 \mu\text{g}/\text{m}^3$)。10 月近地面臭氧浓度高值区由西北向东南开始消退, 11 月回退到中国东南部地区。空间上, 京津冀地区、长三角地区、珠三角地区、成渝地区等城市群显著高于周围其他区域, 是臭氧污染防治的重点区域。除四大城市群外, 中国北方的山东省西北部、河南省西北部、陕西的关中平原区域、山西的汾河谷地; 中国南方的武汉市及周围、长沙市及周围、合肥市及周围也存在较高程度的臭氧污染。

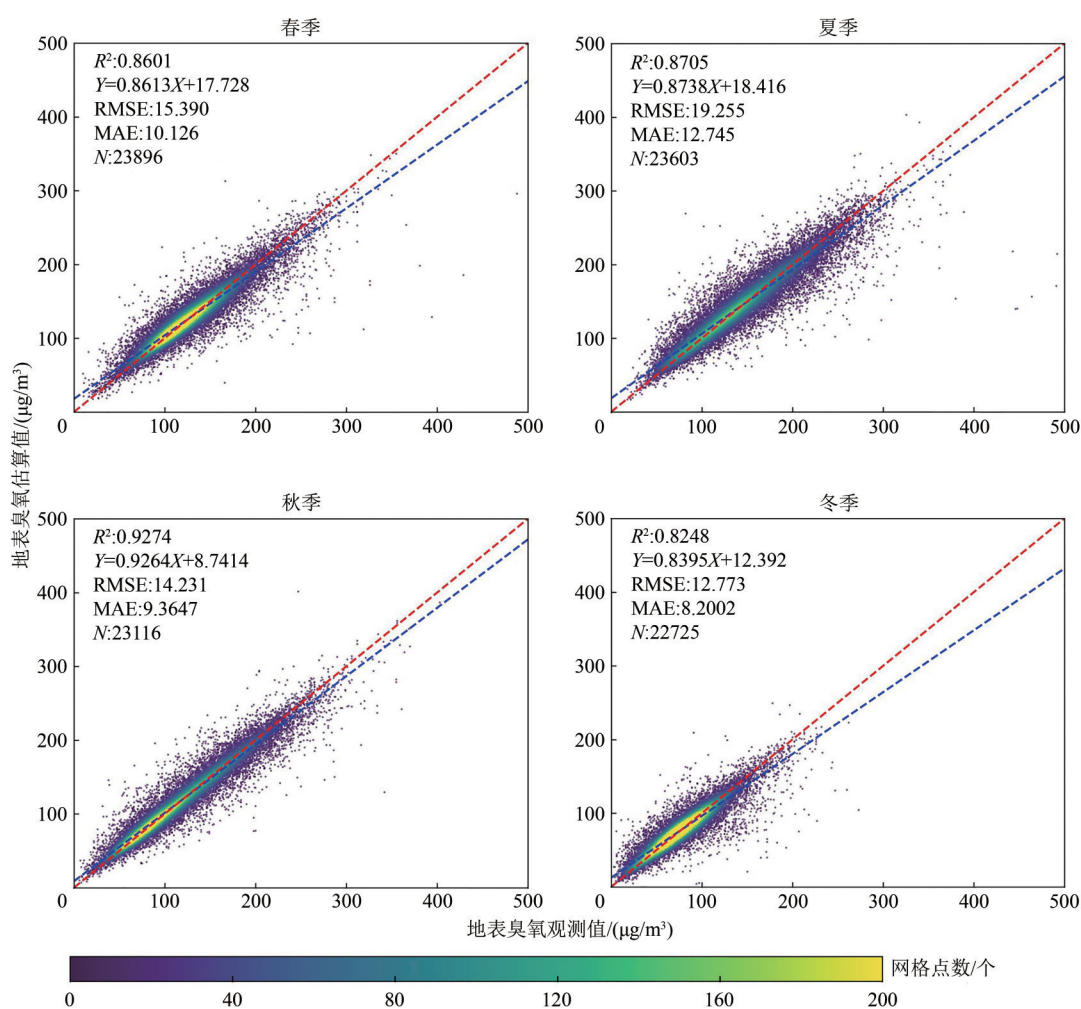


图9 全国不同季度近地面臭氧估算散点图

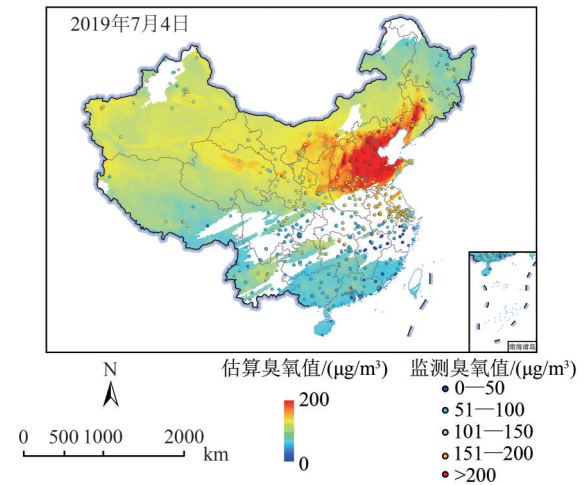
Fig. 9 Scatter plot of estimated near-surface ozone in different seasons across the country

提取全国和 4 个城市群臭氧浓度空间月均值, 并制作月均变化折线图如图 13 所示。臭氧浓度全国平均值变化平缓且有明显的周期性变化, 臭氧浓度多年来的上升趋势得到了逆转, 2020 年臭氧年均值为 $107.41 \pm 18.6 \mu\text{g}/\text{m}^3$, 较去年平均值 $109.26 \pm 19.71 \mu\text{g}/\text{m}^3$, 有所减少。京津冀地区臭氧污染折线图呈现单峰型, 峰值出现在每年的 6 月份。2019 年

6 月份京津冀地区臭氧月均浓度在 4 个城市群最高, 最高值 $190.58 \pm 37.69 \mu\text{g}/\text{m}^3$, 远超全国 6 月份均值 $131.16 \pm 3.93 \mu\text{g}/\text{m}^3$ 。2020 年 6 月京津冀地区臭氧月均浓度值为 $189.16 \pm 37.58 \mu\text{g}/\text{m}^3$, 较 2019 年同期下降 0.7%。长三角地区两年内臭氧污染折线图呈现双峰型, 峰值出现在每年的 5 月和 9 月, 2019 年 5 月和 9 月臭氧浓度分别为 $160.19 \pm 30.69 \mu\text{g}/\text{m}^3$ 、 $155.31 \pm$

31.67 $\mu\text{g}/\text{m}^3$, 2020 年分别为 162.16 $\pm$ 23.26 $\mu\text{g}/\text{m}^3$ 、156.07 $\pm$ 29.48 $\mu\text{g}/\text{m}^3$ 。珠三角地区近地面臭氧浓度月变化在 2019 年呈现“单峰”型, 2020 年呈现“双峰”型。长三角地区臭氧则在这两年均呈现显著的“双峰”特征, 原因是长三角地区在 7 月份属于多雨季节降水较多, 云层遮盖太阳辐射较少, 导致在 7 月份相对于 6 月份和 8 月份有臭氧低值(程麟钧, 2018)。2019 年, 珠三角地区在 10 月份出现臭氧最高值 165.22 $\pm$ 36.04 $\mu\text{g}/\text{m}^3$ 。2020 年, 珠三角地区的第一个臭氧峰值(4 月, 136.63 $\pm$ 28.23 $\mu\text{g}/\text{m}^3$)早于长三角地区(5 月, 162.16 $\pm$ 23.26 $\mu\text{g}/\text{m}^3$), 但其第二个峰值(11 月, 138.44 $\pm$ 32.97 $\mu\text{g}/\text{m}^3$)则晚于长三角地区(9 月, 156.07 $\pm$ 29.48 $\mu\text{g}/\text{m}^3$)。成渝地区近地面臭氧浓度仅在 2019 年 8 月、2020 年 5 月显著高于全国平均值, 分别达到 146.23 $\pm$ 23.39 $\mu\text{g}/\text{m}^3$ 和 147.90 $\pm$ 24.19 $\mu\text{g}/\text{m}^3$, 其他月份则均与全国平均

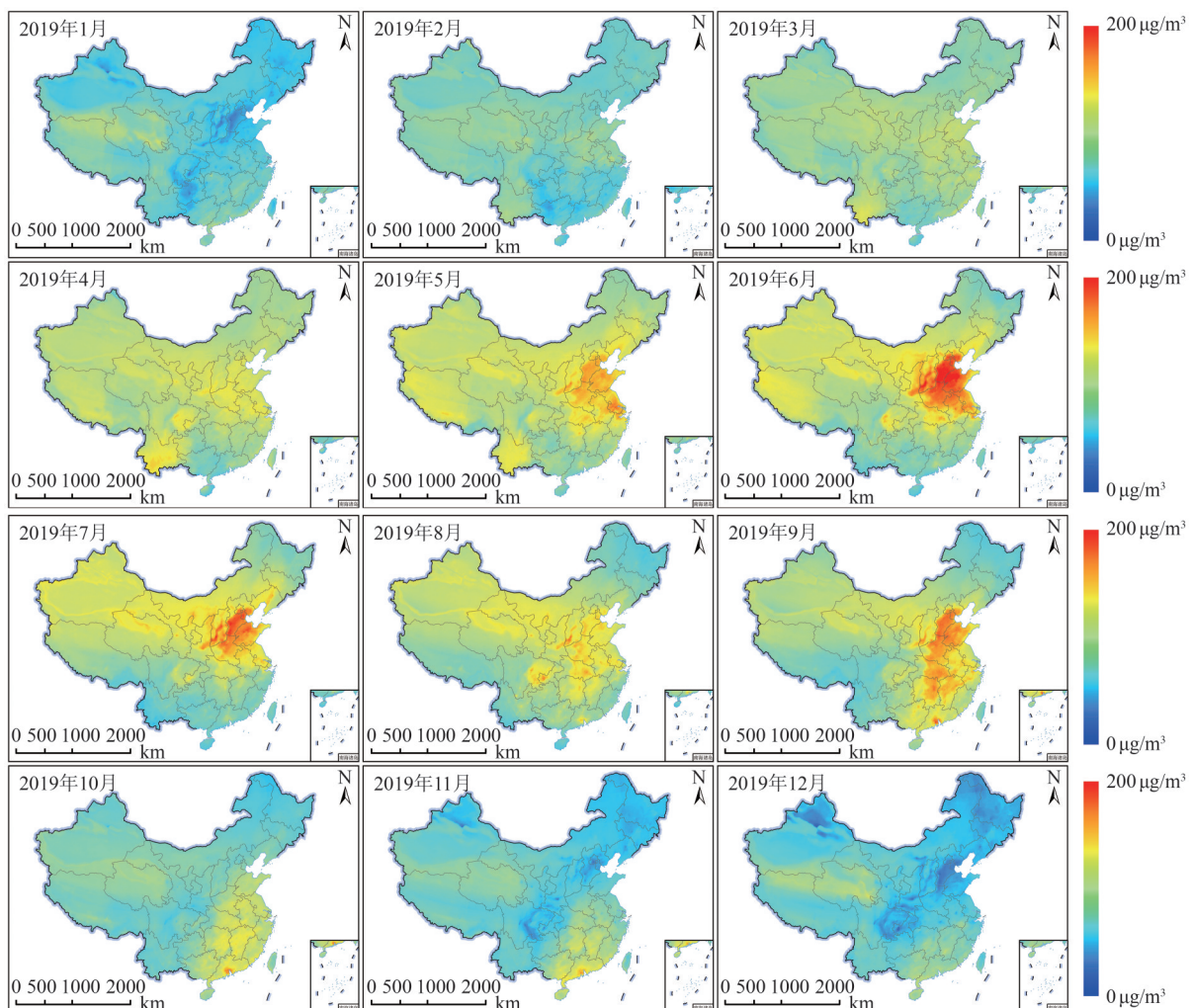
水平基本持平。



审图号:GS京(2023)0780号

图 10 单日近地面臭氧空间分布图(2019-07-04)

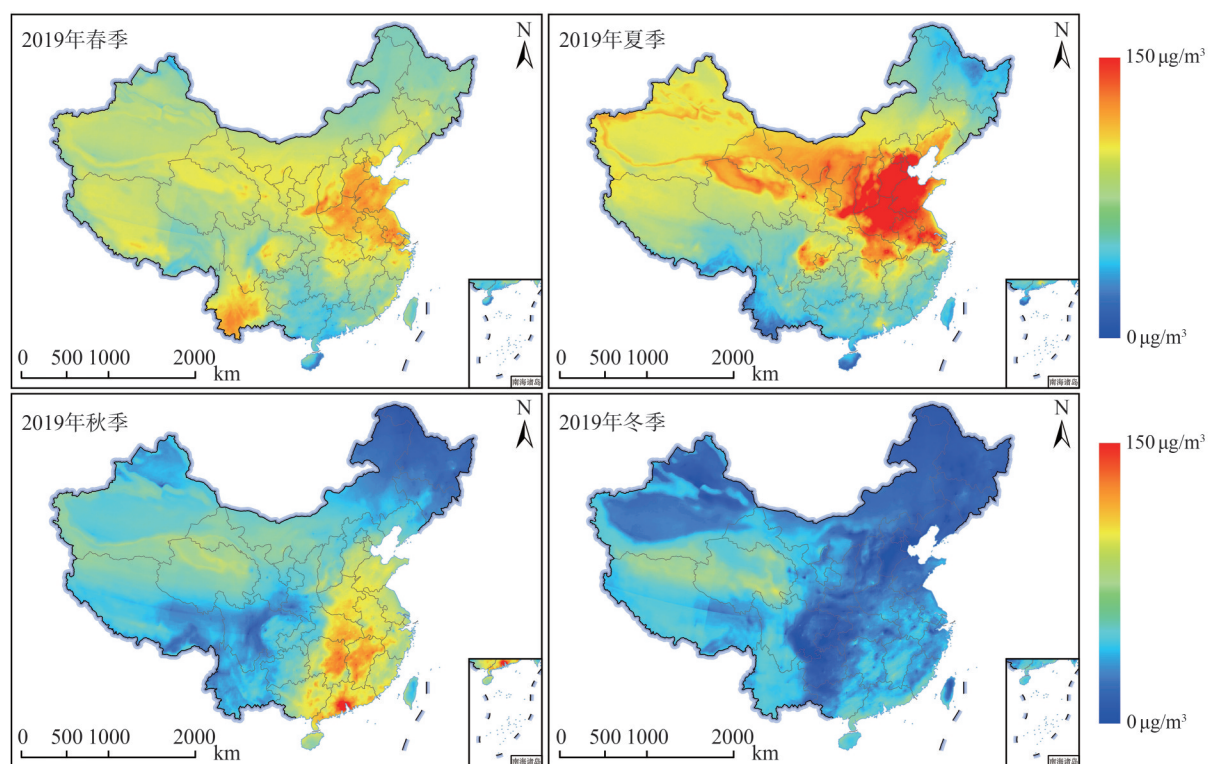
Fig. 10 Spatial distribution map of near-ground ozone in a single day (2019-07-04)



审图号:GS京(2023)0780号

图 11 2019 年—2020 年臭氧月均值空间分布图

Fig. 11 Spatial distribution of monthly average near-ground ozone during 2019—2020



审图号:GS 京(2023)0780 号

图 12 2019 年—2020 年臭氧季均值空间分布图

Fig. 12 Spatial distribution of seasonal average near-ground ozone during 2019—2020

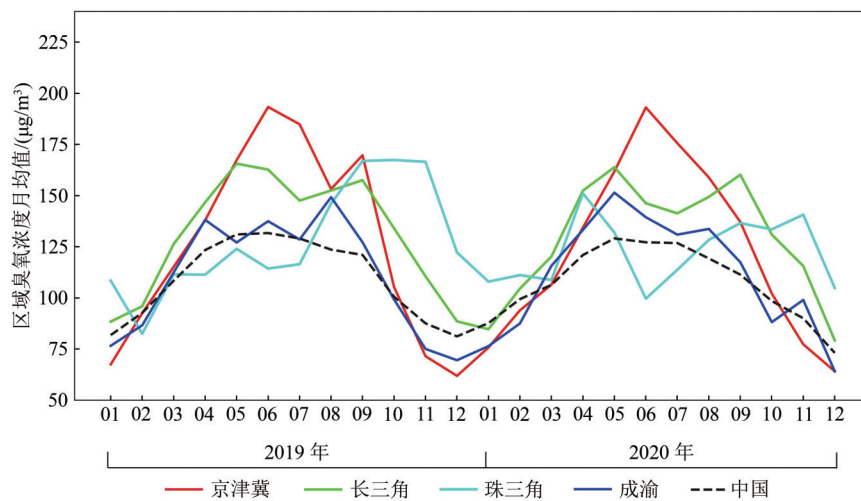


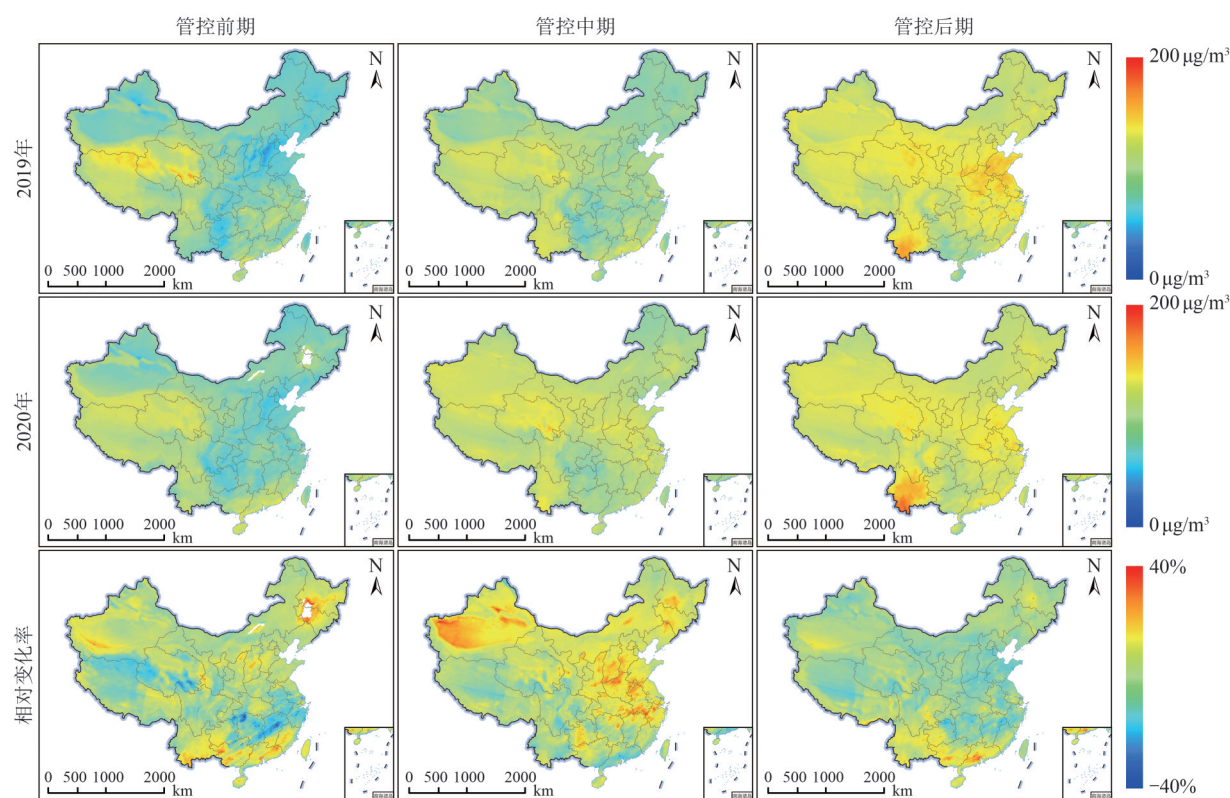
图 13 全国及 4 大城市群月平均臭氧浓度

Fig. 13 Monthly average ozone concentration of the whole country and the four major city agglomerations

5.3 疫情期间臭氧分析

2020 年初新冠肺炎疫情开始在全球范围内蔓延, 为保障人民生命健康, 中国政府实施了严格的管控政策, 期间大量企业停工停产, 公共交通出行也大幅减少(陶金花等, 2020)。为评估疫情管控前后不同阶段近地面臭氧浓度变化, 本文分管控前期(2020 年 1 月 1 日至 25 日)、管控中期

(2020 年 1 月 26 日至 2 月 17 日) 和管控后期(2020 年 2 月 18 日至 3 月 31 日) 3 个阶段开展分析。绘制 2020 年管控前中后期和 2019 年同期的臭氧空间分布图及同一时期的 2020 年较 2019 年臭氧浓度相对变化率图, 成图如图 14 所示。并计算全国、4 大城市群和武汉 3 个时期的相对变化率, 成图如图 15 所示。



审图号:GS 京(2023)0780 号

图 14 疫情管控前中后与去年同期臭氧浓度空间分布图

Fig. 14 Spatial distribution of ozone concentration before, during and after the epidemic control and the same period last year

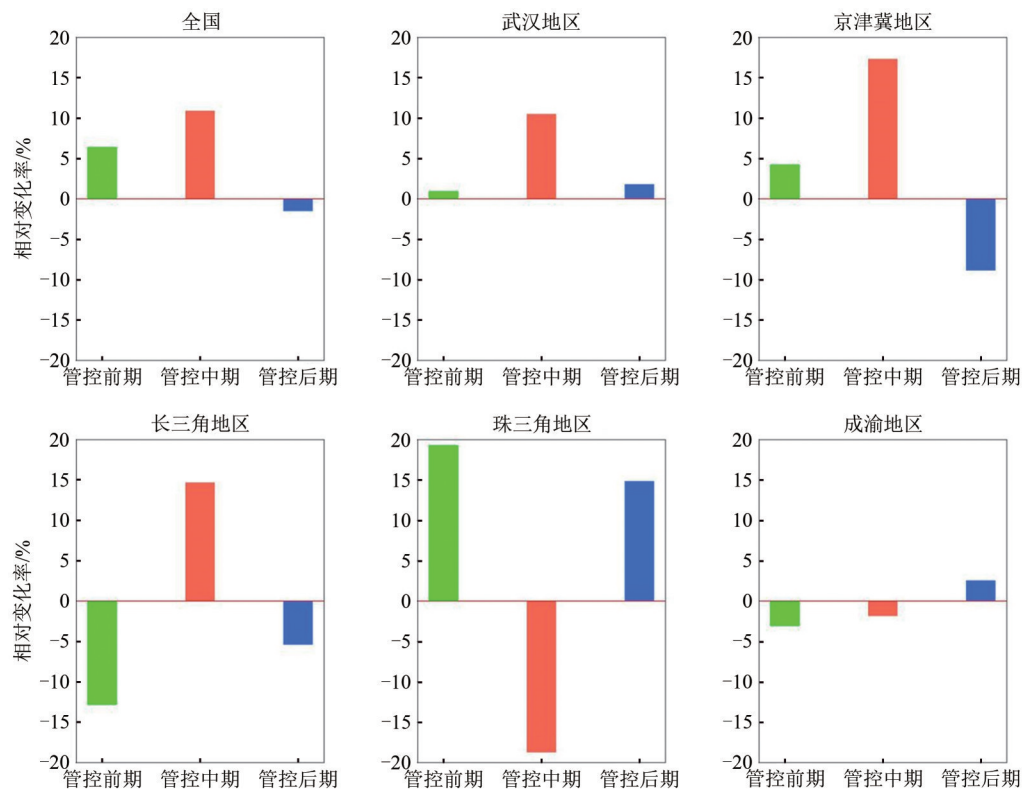


图 15 不同区域近地面臭氧浓度管控前中后期较去年同期的相对变化率

Fig. 15 Relative change rates of near-ground ozone concentration in different regions before, during and after the control period compared with the same period of last year

如图 14 所示, 全国范围内近地面臭氧浓度在 2020 年管控前、中、后 3 个时期总体呈逐渐上升趋势, 但上升程度低于 2019 年同期, 中国东部地区的臭氧浓度在管控后期显著低于 2019 年同期。如图 15 所示, 管控前期, 京津冀地区、成渝地区和武汉地区的近地面臭氧浓度与 2019 年同期相当, 相对变化率分别为 4.23%、-3.1%、0.95%, 长三角地区近地面臭氧浓度较 2019 年同期降低 12.86%, 珠三角地区近地面臭氧浓度较 2019 年同期升高 19.28%。管控中期, 全国、京津冀地区、长三角地区和武汉地区近地面臭氧浓度较 2019 年有较大幅度升高, 相对变化率分别为 10.9%、17.34%、14.69%、10.51%, 珠三角地区近地面臭氧浓度较 2019 年同期降低 18.74%。在管控后期, 京津冀地区和长三角地区的近地面臭氧浓度较 2019 年同期有较大幅度降低, 分别降低 8.9%、5.4%。

在严格的管控措施下, 中国很多地区近地面臭氧浓度并未像 $\text{PM}_{2.5}$ 和 NO_2 一样大幅减低, 而是呈现出上升趋势, 这与 Wei 等 (2022) 和 Song 等 (2021) 研究结果基本一致。其原因主要在于 1—4 月温度回升和太阳辐射增强必然会导致臭氧浓度的上升。此外, 管控期间大气中颗粒物的减少导致太阳辐射在经过大气时损耗较少, 会进一步加强了臭氧光化学反应, 以及 NO_x 的大量减少会导致 NO 对臭氧的滴定效应减少, 也都是导致近地面臭氧浓度上升的重要原因。

6 结 论

本文基于 2019 年—2020 年地面、卫星和模式再分析资料, 利用 6 种不同的模型方法开展了近地面臭氧浓度估算精度评估, 选取了最优模型, 建立了 2019 年—2020 年中国地区近地面臭氧浓度数据集, 重点分析了其时空变化特征, 以及新冠肺炎疫情管控前后对臭氧浓度变化的影响。研究结果表明:

(1) 前体物因素是影响近地面臭氧浓度的主导因素, NO_2 、HCHO 和 CO 与近地面臭氧相关性分别为 0.56、0.62 和 0.46; 气象因素是影响近地面臭氧浓度的重要因素, 其中地表两米温度和地表太阳辐射与近地面臭氧相关性分别为 0.34、0.27。

(2) 集成学习模型整体精度优于传统模型, 基于 ERT 模型、XGBoost 模型、GBRT 模型估算的近

地面臭氧浓度 R^2 指标差距, 分别为 0.8968、0.904、0.9086。但 ERT 模型估算得到的近地面臭氧浓度空间分布更加连续。

(3) 中国臭氧浓度多年来的上升趋势得到了逆转, 2020 年臭氧年均值为 $107.41 \pm 18.6 \mu\text{g}/\text{m}^3$, 较去年平均值 $109.26 \pm 19.71 \mu\text{g}/\text{m}^3$, 有所减少。京津冀地区臭氧高值主要出现在 6 月 ($190.58 \pm 37.69 \mu\text{g}/\text{m}^3$), 明显高于全国均值和其他区域均值。

(4) 新冠肺炎疫情管控期间, 全国、京津冀地区、长三角地区和武汉地区近地面臭氧浓度较 2019 年有较大幅度升高, 相对变化率分别为 10.9%、17.34%、14.69%、10.51%, 珠三角地区近地面臭氧浓度较 2019 年同期降低 18.74%。

参考文献 (References)

- Bell M L. 2006. The use of ambient air quality modeling to estimate individual and population exposure for human health research: a case study of ozone in the Northern Georgia Region of the United States. *Environment International*, 32(5): 586-593 [DOI: 10.1016/j.envint.2006.01.005]
- Breiman L. 2001. Random forests. *Machine Learning*, 45(1): 5-32 [DOI: 10.1023/A:1010933404324]
- Chen L F, Wang Y P, Zhang X X, Wang Z F, Tao J H, Wang L L and Zhang Y. 2019. Satellite remote sensing monitoring of ozone and its precursors for regional secondary pollution risk control. *Environmental Monitoring and Forewarning*, 11(5): 13-21 (陈良富, 王雅鹏, 张欣欣, 王子峰, 陶金花, 王莉莉, 张莹. 2019. 面向区域二次污染风险控制的臭氧及其前体物卫星遥感监测. *环境监控与预警*, 11(5): 13-21) [DOI: 10.3969/j.issn.16746732.2019.05.003]
- Chen R J, Chen B H and Kan H D. 2010. Health impact assessment of surface ozone pollution in Shanghai. *China Environmental Science*, 30(5): 603-608 (陈仁杰, 陈秉衡, 阚海东. 2010. 上海市近地面臭氧污染的健康影响评价. *中国环境科学*, 30(5): 603-608)
- Chen T Q and Guestrin C. 2016. XGBoost: a scalable tree boosting system//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: ACM: 785-794 [DOI: 10.1145/2939672.2939785]
- Cheng L J. 2018. A Study on the Evolution of Ozone Pollution in China and Regional Management Methods. Beijing: China University of Geosciences (程麟钧. 2018. 中国臭氧污染特征及分区管理方法研究. 北京: 中国地质大学)
- Chin M, Jacob D J, Munger J W, Parrish D D and Doddridge B G. 1994. Relationship of ozone and carbon monoxide over North America. *Journal of Geophysical Research: Atmospheres*, 99(D7): 14565-14573 [DOI: doi.org/10.1029/94JD00907]
- Dueñas C, Fernández M C, Cañete S, Carretero J and Liger E. 2002. Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. *Science of the Total Environment*, 299(1/3): 97-113 [DOI: 10.1016/S0048-9697(02)00251-6]
- Duncan B N, Yoshida Y, Olson J R, Sillman S, Martin R V, Lamsal L,

- Hu Y T, Pickering K E, Retscher C, Allen D J and Crawford J H. 2010. Application of OMI observations to a space-based indicator of NO_x and VOC controls on surface ozone formation. *Atmospheric Environment*, 44(18): 2213-2223 [DOI: 10.1016/j.atmosenv.2010.03.010]
- Fang T. 2020. Study on Estimation and Prediction of Near-Surface Ozone based on Neural Network. Shanghai: Shanghai Normal University (方韬. 2020. 基于神经网络的近地面臭氧估算和预测研究. 上海: 上海师范大学) [DOI: 10.27312/d.cnki.gshsu.2020.002120]
- Feister U and Balzer K. 1991. Surface ozone and meteorological predictors on a subregional scale. *Atmospheric Environment. Part A. General Topics*, 25(9): 1781-1790 [DOI: 10.1016/0960-1686(91)90262-6]
- Friedman J, Hastie T and Tibshirani R. 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2): 337-407 [DOI: 10.1214/aos/1016218223]
- Gardner M W and Dorling S R. 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, 34(1): 21-34 [DOI: 10.1016/S1352-2310(99)00359-3]
- Geurts P, Ernst D and Wehenkel L. 2006. Extremely randomized trees. *Machine Learning*, 63(1): 3-42 [DOI: 10.1007/s10994-006-6226-1]
- Haagen-Smit A J. 1952. Chemistry and physiology of Los Angeles smog. *Industrial and Engineering Chemistry*, 44(6): 1342-1346 [DOI: 10.1021/ie50510a045]
- Sillman S. 1995. The use of NO_x , H_2O_2 , and HNO_3 as indicators for ozone- NO_x -hydrocarbon sensitivity in urban locations. *Journal of Geophysical Research*, 100: (D7): 14175-14188 [DOI: 10.1029/94JD02953]
- Song Z G, Bai Y, Wang D F, Li T and He X Q. 2021. Satellite retrieval of air pollution changes in central and Eastern China during COVID-19 lockdown based on a machine learning model. *Remote Sensing*, 13(13): 2525 [DOI: 10.3390/rs13132525]
- Tao J H, Fan M, Gu J B and Chen L F. 2020. Satellite observations of the return-to-work over China during the period of COVID-19. *Journal of Remote Sensing (Chinese)*, 24(7): 824-836 (陶金花, 范萌, 顾坚斌, 陈良富. 2020. 新冠病毒疫情期间复工复产卫星遥感监测. *遥感学报*, 24(7): 824-836) [DOI: 10.11834/jrs.20200098]
- Wang C, Wang S, Yang B B, Zhang Z H, Wang L and Liu M. 2015. Study of the effect of meteorological conditions on the ambient air ozone concentrations in Shenyang. *Environmental Monitoring in China*, 31(3): 32-37 (王闯, 王帅, 杨碧波, 张丽辉, 王磊, 刘闯. 2015. 气象条件对沈阳市环境空气臭氧浓度影响研究. *中国环境监测*, 31(3): 32-37) [DOI: 10.19316/j.issn.1002-6002.2015.03.007]
- Wei J, Li Z Q, Li K, Dickerson R R, Pinker R T, Wang J, Liu X, Sun L, Xue W H and Cribb M. 2022. Full-coverage mapping and spatiotemporal variations of ground-level ozone (O_3) pollution from 2013 to 2020 across China. *Remote Sensing of Environment*, 270: 112775 [DOI: 10.1016/j.rse.2021.112775]
- Zhan Y, Luo Y Z, Deng X F, Grieneisen M L, Zhang M H and Di B F. 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environmental Pollution*, 233: 464-473 [DOI: 10.1016/j.envpol.2017.10.029]
- Zhang Y, Wang W H, Zhang X Y and Lawrence F. 2013. Interannual variations of Arctic ozone and their relationship to the polar vortex. *Journal of Remote Sensing*, 17(3): 527-540 (张艳, 王维和, 张兴赢, Lawrence F. 2013. 北极臭氧年际变化特征及其与极涡的关系. *遥感学报*, 17(3): 527-540) [DOI: 10.11834/jrs.20132037]

Estimating ground-level ozone concentration in China using ensemble learning methods

SONG Shipeng^{1,2}, FAN Meng², TAO Jinhua², CHEN Sanming¹, GU Jianbin², HAN Zongfu²,
LIANG Xiaoxia⁴, LU Xiaoyan⁵, WANG Tiantian⁶, ZHANG Ying²

1. College of Earth Sciences, Guilin University of Technology, Guilin 541000, China;

2. State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China;

3. University of the Chinese Academy of Sciences, Beijing 100049, China;

4. College of Resource Environment and Tourism, Capital Normal University, Beijing 100089, China;

5. Guangxi Eco-Environmental Monitoring Center, Nanning 530028, China;

6. Jiangsu Provincial Environmental Monitoring Center, Nanjing 210019, China

Abstract: Following the successful implementation of the Air Pollution Prevention and Control Action Plan (2013—2017) and the Three-Year Action Plan to Win the Blue Sky Defense War (2018—2020), the concentrations of five major pollutants (i.e., $\text{PM}_{2.5}$, PM_{10} , SO_2 , NO_2 , and CO), except for ozone, significantly dropped for most cities in China. The increasing ground-level ozone concentrations have been a key factor restricting the improvement in ambient air quality, especially during summer.

Compared with the measurements from ground-based monitoring sites, satellite remote sensing technology can obtain spatially continuous total column ozone. However, given that ozone is abundantly distributed in the stratosphere, ground-level ozone has a very low

contribution to the total column ozone observed from space. Therefore, the satellite total column ozone product provides limited information for estimating ground-level ozone concentrations. In this study, by combining TROPOMI ozone precursor (NO_2 and HCHO) products, ERA5 meteorological parameters, and ground-based monitoring data, a machine learning model was developed to estimate the daily maximum 8-hour average ground-level ozone concentration over China from years 2019 to 2020.

By comparing the performance of three ensemble learning methods, namely, extreme gradient boosters (XGBoost), extreme random trees (ERT), and gradient boost regression tree (GBRT), the averaged overall 10-fold cross-validation R^2 of 2019 and 2020 are all larger than 0.89. Although the results estimated by XGBoost showed the best agreement between the model predictions and observations with an average RMSE and MAE of $15.77 \mu\text{g}/\text{m}^3$ and $10.53 \mu\text{g}/\text{m}^3$, respectively, the ERT method was eventually selected to model the daily maximum 8-hour average ground-level ozone concentration by considering the rationalization of spatial distribution.

Due to the proactive emission reduction measures implemented by the Chinese government and the impact of the COVID-19 pandemic, the rising trend of ozone concentration over the years has been reversed. The annual average value of ground-level ozone concentration in 2020 reached $107.41 \pm 18.6 \mu\text{g}/\text{m}^3$ over China, which is $1.85 \mu\text{g}/\text{m}^3$ less than that recorded in 2019 ($109.26 \pm 19.71 \mu\text{g}/\text{m}^3$). Severe surface ozone pollution events frequently occur from May to September of every year because the high temperatures during these months can promote photochemical reactions. The estimated ground-level ozone concentrations in the Beijing-Tianjin-Hebei, Yangtze River Delta, Pearl River Delta, and Chengdu-Chongqing regions are significantly higher than those in their surrounding areas, making these regions the key areas for ozone pollution prevention and control.

Key words: remote sensing, ground-level ozone, Ensemble Learning, ERT, GBRT, XGBoost, TROPOMI, COVID-19

Supported by National Key Research and Development Program of China (No. 2018YFC0214003); National Natural Science Foundation of China (No. 41771391, 42171393, 41871254); Key Research and Development Program of Guangxi (No. AB20238015)