顾桃峰,岳海燕,熊梓立,等.基于 AI 的大气二氧化碳数据填补与重构算法研究 [J]. 中国环境科学, 2026,46(3):1288-1297.
Gu T F, Yue H Y, Xiong Z L, et al. Research on AI-based algorithms for atmospheric carbon dioxide data imputation and reconstruction [J]. China Environmental Science, 2026,46(3):1288-1297.

基于 AI 的大气二氧化碳数据填补与重构算法研究

顾桃峰[1],岳海燕[2],熊梓立[1],梁建平[3],吴艳玲[4],刘立新[5*] (1.广州市气象综合保障中心,广东 广州 511430；2.广州市气象台,广东 广州 511430；3.桂林市气象局,广西 桂林 541000；4.龙凤山区域大气本底站,黑龙江 五常 150001；5.中国气象局气象探测中心,北京 100081)

摘要：针对大气 CO_2 浓度部分观测数据缺失的挑战,利用龙凤山区域大气本底站(简称龙凤山站)2009 年 1 月～2022 年 6 月 30m(L1 层)与 80m(L2 层)高度层的 CO_2 浓度观测数据,评估 12 种主流机器学习算法在数据填补中的性能,提出以 K 近邻-距离加权(KNeighborsDist)算法为核心框架的优化组合机器学习方法.该方法利用同址不同高度站点均有数据记录重建了缺失数据.经对重建数据集数据测试分析验证,数据集重建效果优异(R^2≥0.98),充分表明使用测试数据集重建的温室气体数据与实际测量数据之间具有良好的相关性,验证了所构建模型在温室气体数据重建任务上的有效性和准确性.利用该方法成功填补了龙凤山站原始数据缺失部分,构建了逐 5min 连续的 CO_2 数据集,填补后,龙凤山站数据完整率达到 100%,填补方法在保留 CO_2 长期趋势/季节/月/日变化规律上表现良好.本研究提出的 AI 数据重构算法可在其他本底站应用推广,并可为区域气候变化特征分析、碳排放核算评估及反演研究提供高质量数据支持.

关键词：CO_2；机器学习算法；数据填补；数据集；龙凤山
中图分类号：X511 **文献标识码**：A **文章编号**：1000-6923(2026)03-1288-10

Research on AI-based algorithms for atmospheric carbon dioxide data imputation and reconstruction. GU Tao-feng[1], YUE Hai-yan[2], XIONG Zi-li[1], LIANG Jian-ping[3], WU Yan-ling[4], LIU Li-xin[5*] (1.Guangzhou Meteorological Comprehensive Guarantee Center, Guangzhou 511430, China；2.Guangzhou Meteorological Observatory, Guangzhou 511430, China；3.Guilin Meteorological Observatory, Guilin 541000, China；4.Longfengshan Regional Atmospheric Backbone Station, Wuchang 150001, China；5.China Meteorological Bureau Meteorological Detection Center, Beijing 100081, China). *China Environmental Science*, 2026,46(3)：1288~1297

Abstract：To address the challenge of partial missing data in atmospheric CO_2 concentration observations, this study systematically evaluated the performance of 12 mainstream machine learning algorithms for data imputation, utilizing CO_2 measurement data collected at the 30m (L1 layer) and 80m (L2 layer) heights of the Longfengshan Regional Atmospheric Background Station over the period from January 2009 to June 2022. An optimized ensemble machine learning method was proposed, with the distance-weighted K-Nearest Neighbors (KNeighborsDist) algorithm as its core framework. This method reconstructs missing data by leveraging synchronous observational records from the two co-located height layers. Validation analysis of the reconstructed dataset yielded excellent performance (R^2≥0.98), which confirms a strong correlation between the reconstructed greenhouse gas data and in-situ measurements, as well as the effectiveness and accuracy of the proposed model in greenhouse gas data reconstruction tasks. The method successfully filled the data gaps in the original dataset of Longfengshan Station, generating a continuous 5minute interval CO_2 dataset with a completeness rate of 100%. Additionally, it well preserved the long-term trends and seasonal, monthly, and diurnal variation characteristics of CO_2. The proposed AI-based data reconstruction algorithm is applicable to other atmospheric background stations and can provide high-quality data support for regional climate change analysis, carbon emission accounting, and inversion research.

Key words：carbon dioxide；machine learning algorithm；data filling；data set；Longfengshan

自 20 世纪 90 年代起,中国气象局陆续在北京上甸子、黑龙江龙凤山和青海瓦里关等地区持续开展大气温室气体浓度本底观测,在长期观测过程中,由于设备故障、观测员误操作、天气影响、周边环境影响等因素,导致观测数据的不连续问题时有发生,一旦数据中断时间过长,会严重干扰温室气体浓度长期变化趋势的分析,进而阻碍对温室气体管理措

收稿日期：2025-08-09
基金项目：科技部重点研发计划项目(2022YFF0606400);中国气象局创新发展专项(CXFZ2023J066);广州市气象局科研项目(M202405);中国气象局“温室气体观测关键技术研发及应用”青年创新团队(CMA2023QN13)
* 责任作者, 正研级工程师, liulx@cma.gov.cn

施的有效评估[1].

近年来,各国学者从不同维度运用传统统计模型开展缺失数据填补与评估工作.MASARIE 等[2]在相关研究中聚焦 CO_2 浓度数据,考虑到采样点稀疏且缺乏连续性,提出将有限观测数据通过插值和外推方法拓展为完整全球数据集,以此摆脱对大气模型的依赖.该方法助力研究人员克服数据不完整难题,构建出全球一致的 CO_2 数据记录,为全球 CO_2 浓度研究提供了基础.世界气象组织温室气体全球数据分析方法技术报告[3]中介绍了温室气体的长期趋势和季节性变化的分析方法,通过去除短期局部效应,运用数字滤波等手段,先以线性趋势拟合去除季节性影响,再滤除高频信号,最终得到平滑的长期趋势线.此方法在计算和预测全球、半球及分区平均浓度方面应用广泛,但存在明显局限,仅能填补小于 2 个月的缺失数据,且数据填补精度仅至逐小时,难以满足更长时间跨度和更高精度数据填补需求.闫丽莉等[4]借助改进的线性回归模型,对缺测和错误的气温整点值数据进行填补.此方法有效缓解了长期连续观测数据缺测的问题,为气温数据的完整性和连续性提供了保障,在气温数据研究与分析中发挥了积极作用.司鹏等[5]针对河北保定气象站 1913~2014 年月平均气温、最高气温和最低气温资料,采用标准化序列法和多元线性回归法进行填补.通过交叉检验法分析,标准化序列法填补得到的气温序列效果更佳,且其气候统计特征与同区域周边站的研究结果一致性更高,这表明标准化序列法在该气象站气温数据填补中具有较好的适用性和准确性.线性插值作为一种常见的数据估算方法,其原理是在两个已知数据点之间估算新数据点,并假设两点间变化为线性.然而,该方法在处理大气 CO_2 数据时存在明显不足.大气 CO_2 数据具有显著的日变化和季节变化趋势,同时多年长期变化趋势近乎线性增长,线性插值法简单假设的线性变化无法精准反映这种复杂且多变的实际变化趋势,难以满足对大气 CO_2 数据准确填补和分析的需求.

相较于传统统计模型,机器学习算法优势显著.Grange 等[6]指出,机器学习算法在捕捉大数据中的非线性特征方面表现更优,且具备更强的鲁棒性与准确性[7-8].在环境数据重建领域,多种机器学习算法得到广泛应用.在重建表面 $PM_{2.5}$ 或 O_3 浓度方面,随机森林(RF)被其他研究者等采用[9-11];极端梯度增强(XGBoost)由 Liu 等[12]用于相关研究;分类增强(CatBoost)在 Ding 等[13]的研究中发挥作用;光梯度增强机(LightGBM)被应用于数据重建[14-16];神经网络也被 Di 等[17]用于此目的.除了历史空气质量数据重建,机器学习在空间数据空白填充方面也成果颇丰.有研究者利用机器学习进行空间数据空白填充[18-24],Chen 等[25]使用两阶段机器学习模型在中国构建网格化 $PM_{2.5}$ 数据集,还利用随机森林和深度神经网络成功实现 AOD 的空间数据缺口填充.然而,在温室气体数据重建领域,利用机器学习算法开展的研究却鲜有文献报道.

针对大气 CO_2 浓度观测数据的部分缺失的,本研究利用龙凤山区域大气本底站 2009 年 1 月~2022 年 6 月 30m(L1 层)与 80m(L2 层)高度层的 CO_2 浓度观测数据,通过评估 12 种主流机器学习算法在数据填补中的性能,提出以 K 近邻−距离加权(Kneighbors DistK)算法为核心框架的优化组合机器学习方法模型,探索并构建一套适用于大气 CO_2 浓度完整序列历史数据集方法,以期提升气候变化监测与评估工作的准确性和连续性.

1　数据与方法

1.1　观测站点

龙凤山站(44°44′N,127°36′E,330.5m)位于黑龙江省五常市境内的龙凤山,距离五常市 40km,距离省会哈尔滨市直线距离 185km.该站周边环境优良,站址方圆 40km 无大城市和工业区,四周森林覆盖良好,属典型的温带季风气候.该站是世界气象组织全球大气观测网(WMO/GAW)的区域大气本底站之一[26],国家大气成分本底野外观测研究站,其观测结果能够代表东北区域大气温室气体浓度水平.

1.2　观测方法

龙凤山站建有 83m 温室气体梯度观测塔,气体进样口位于观测塔 30m(L1)和 80m(L2)两个高度,进气口处安装有⩽7μm 的颗粒物过滤器,L1/L2 层自然空气通过无油惰性隔膜泵和多级冷凝除水,再通过多口样品选择阀及自动流量和压力控制系统后,交替压入 PicarroG2401 型 CO_2 分析主机.由程序控制的多口选择阀处同时连接高浓度工作标气(高于环境空气中 CO_2 浓度)、低浓度工作标气

(低于环境空气中 CO_2 浓度)和目标气(与环境空气中 CO_2 浓度接近)进入主机,通过标气外标法,利用标气的标称值和仪器测量的响应平均值进行最小二乘法线性拟合,得到线性回归方程,在方程中带入实测和目标气的响应平均值,计算得到空气和目标气的校正值,目标气实测值与标称值的差值应小于 0.1×10^{-6}.

1.3 数据处理方法

借助开源的自动化机器学习库 AutoGluon 中的 12 个模型[27-30]进行模型训练、验证、测试以及指标评估,这 12 种算法分别为轻量级梯度提升机(LightGBM)、类别特征梯度提升(CatBoost)、轻量级梯度提升机-极端随机树混合(LightGBMXT)、轻量级梯度提升机-大规模数据版(LightGBMLarge)、XGBoost、神经网络-PyTorch 实现(NeuralNetTorch)、神经网络-FastAI 实现(NeuralNetFastAI)、加权集成模型-二级(WeightedEnsemble_L2)、随机森林-均方误差准则(RandomForestMSE)、极端随机树-均方误差准则(ExtraTreeMSE),KNeighborsDist,K 近邻-均匀加权(KNeighborsUnif).这 12 种机器学习算法分为 4 类,即梯度提升类算法、神经网络类算法、集成学习类算法和最近邻类算法.梯度提升算法是一种广泛应用于分类和回归问题的监督式机器学习算法.其核心原理在于借助决策树来探寻因变量与自变量之间的内在关系.它本质上属于集成学习技术范畴,通过将多个性能较弱的分类器或回归器进行有机结合,显著提升模型的整体性能.在本研究中,本文选用了 LightGBM、CatBoost、XGBoost、LightGBMXT 和 LightGBMLarge 这 5 种基于梯度增强原理的算法.神经网络类算法是一种极具影响力的算法,它通过模拟人工神经元构建起一个并行计算系统,专门用于数据处理与预测任务.本研究中采用的 NeuralNetTorch 和 NeuralNetFastAI 便是基于神经网络原理构建的算法.集成学习类算法的工作机制是在原始训练数据集的随机子集上构建多棵决策树,随后将这些决策树进行整合,形成一个具备高度鲁棒性和泛化能力的模型.对 WeightedEnsemble_L2、RandomForestMSE 和 ExtraTreeMSE 这 3 种基于集成学习的算法进行了测试.K-最近邻(KNN)算法是一种常见的监督学习算法,它依据待预测样本与已知样本之间的距离来进行预测.本文还评估了 KNeighborsDist 和 KNeighborsUnif 这 2 种基于 KNN 原理的算法.12 种机器学习算法的基本描述见表 1.

表 1 几种代表性机器学习算法的基本描述

Table 1 Brief descriptions of typical machine learning algorithms

算法类型	描述	作者,时间	本研究中使用的具体算法
Gradient Boosting	一种监督学习算法,利用决策树建立因变量与自变量之间的关系,通过集成多个弱分类器或回归器来提升模型性能	Friedman,2001	LightGBM, CatBoost, XGBoost, LightGBMXT、LightGBMLarge,
Neural Networ	一种流行算法,使用人工神经元构建并行计算系统,用于数据处理和预测	Bishop,1994	NeuralNetTorch,NeuralNetFastAI
Random Forest	通过在原始训练数据的随机子集上构建多个决策树并将其组合,以形成一个强大且泛化的模型	Breiman,2001	RandomForestMSE, ExtraTreeMSE, WeightedEnsemble_L2
K-Nearest Neighbor (KNN)	一种常用的监督学习算法,根据最近邻居的距离进行预测	Cover,1967	KNeighborsDist, KNeighborsUnif

1.4 试验方法

1.4.1 模型训练与优化 为确保观测数据质量,在将其应用于机器学习之前,需结合值班记录剔除由于系统故障及其它因素导致的明显不合理数据(例如采样点周围的人为活动、系统维护和测试等),并对测量数据进行标准化处理.将 CO_2_L1/CO_2_L2 层同一时刻都有观测要素的部分按照 8:2 比例分成 2 部分(80%划分为训练集/验证集,20%作为测试集)[31],训练集用于构建训练模型;验证集在训练过程中用于评估模型表现并进行调整(如超参数调优);测试集则用于最终独立地评估训练好的模型的泛化能力和可靠性.在每一个部分分别训练不同的模型,并在各自的测试集中进行模型测试.通过充分的迭代,使用 R^2 作为验证度量获得优化后的模型参数.重建 CO_2 浓度数据的整个过程如图 1 所示,使用训练/验证数据来训练机器学习模型,训练过程完成后,使

用测试数据(重叠数据的20%)在所有12种机器学习算法中评估训练后的模型的性能,得出最佳模型.对于 L1 层的缺失数据,利用以 L1 数据为目标变量、L2数据为特征变量的模型,输入对应时间段的L2层数据,从而预测得到 L1 完整数据;对于 L2 层的缺失数据,利用以 L2 数据为目标变量、L1 数据为特征变量的模型,输入对应时间段的L1层数据,从而预测得到 L2 完整数据.

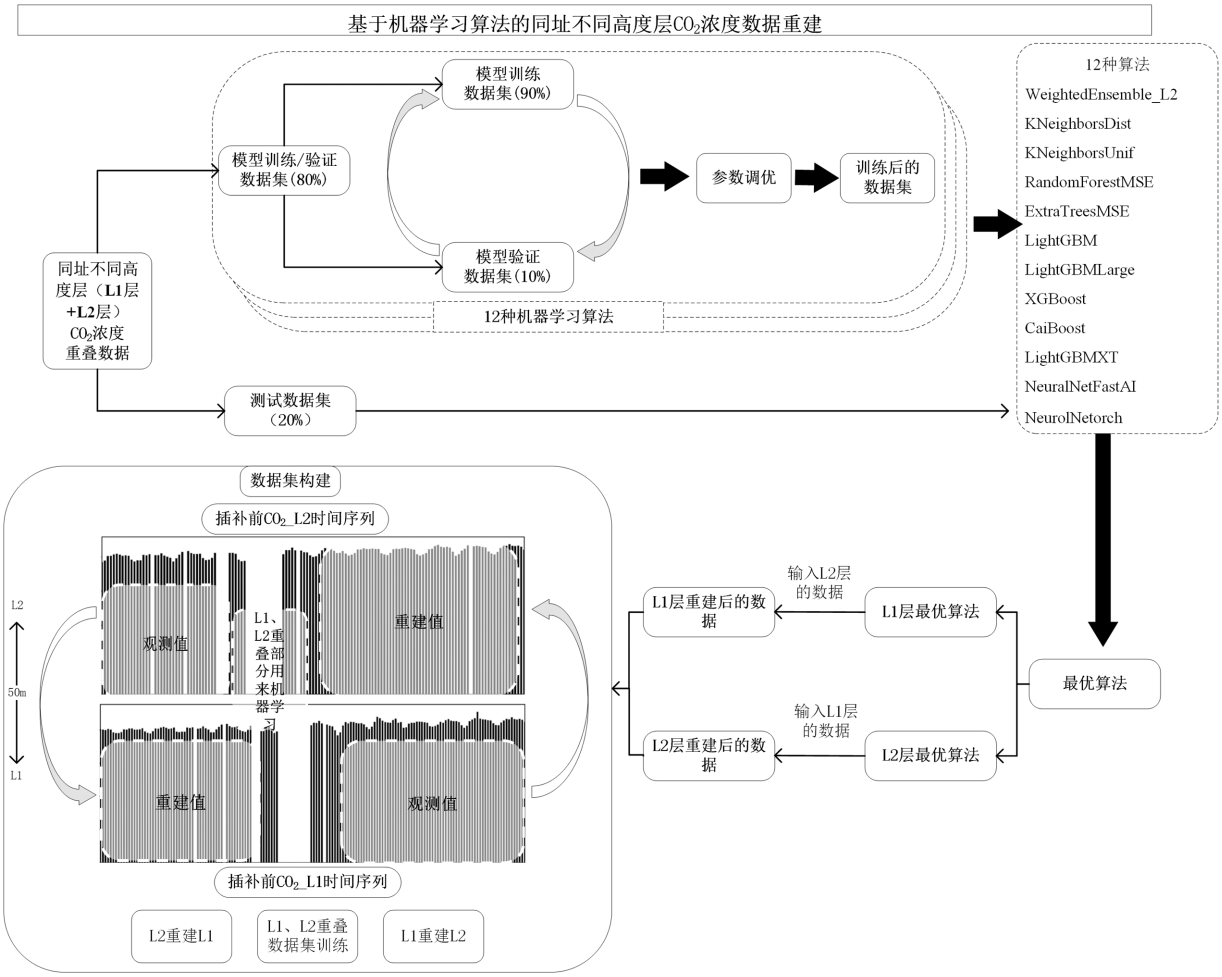


图 1　CO_2 浓度数据重建及数据集构建流程

Fig.1　Flowchart of CO_2 concentration data reconstruction and dataset establishment

1.4.2　模型评估和基准测试　采用 10 个评价指标,包括决定系数(R^2)、均方根误差(RMSE)、平均绝对误差(MAE1)、中位数绝对误差(MAE2)、平均偏差(MD)、精度(Precision)、效率(Efficiency)、平均绝对误差变异系数(CvMAE)、戴明回归斜率(DRS)和戴明回归截距(DRI).

2　结果与分析

2.1　低、高层数据互补解析

通过龙凤山站的温室气体梯度观测塔,可以获取不同空间尺度的观测数据信息.近地面观测主要反映近地面气团的影响,受局地地表特征(如植被、建筑)影响显著,覆盖范围较小.随着高度增加,可代表更厚的混合层气团[32-33],尤其在白天对流较强时垂直范围更大.在地形相对平坦均匀的情况下,观测点水平代表性通常为塔高的 10~100 倍[34-35].实际水平和垂直代表性范围受风速、湍流混合程度影响.夜间,地面因长波辐射冷却后,热通量是向下的,空气处于稳定层结状态.白天,由于地表接收太阳辐射后被加热,边界层内的湍流运动使得这些热量向上传递;在白天尤其午后时段,对流比较旺盛,垂直方向上湍流较强,气团混合更为均匀.因此,如果高层或低层设备出现故障而导致数据缺测,部分时段观测数据可以互为补充.

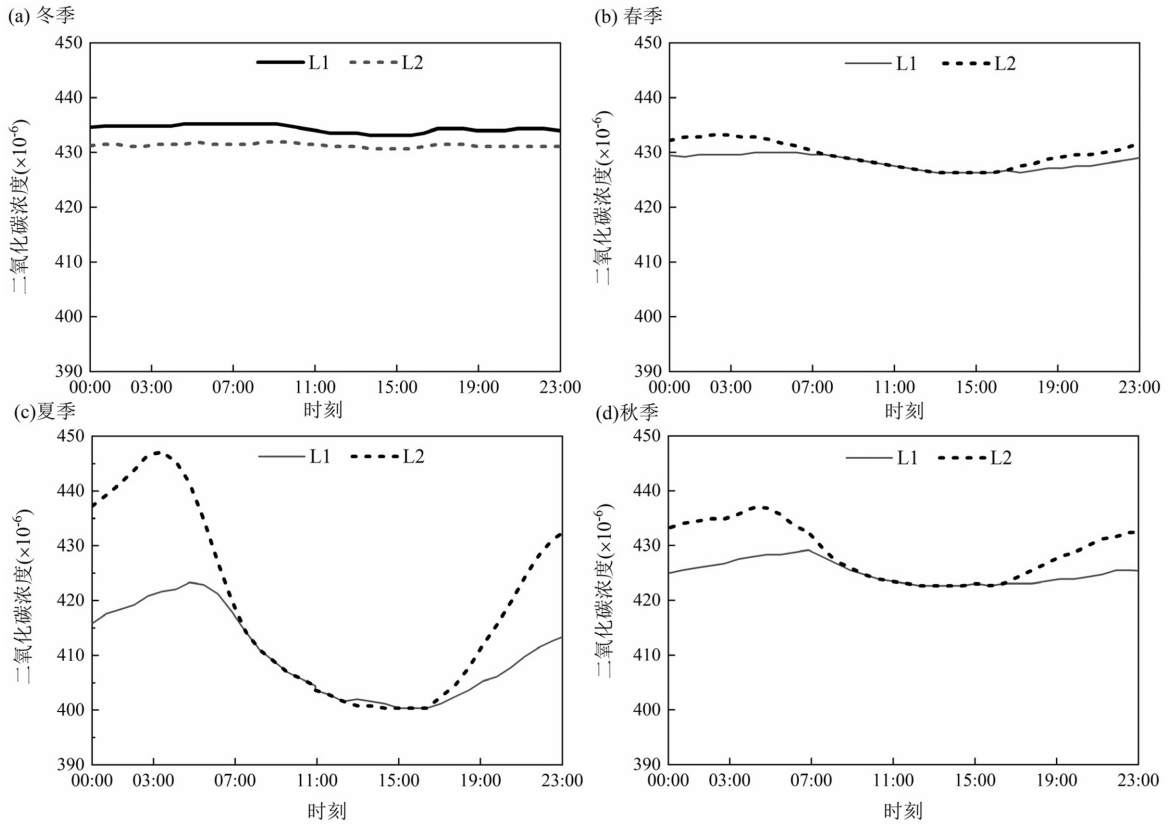


图 2 2021 年龙凤山站不同季节高低层大气 CO_2 浓度昼夜变化

Fig.2 Day−night change chart of CO_2 concentration at high and low levels in Longfengshan Station in different seasons in 2021

由图 2 可见,除冬季由于局地和区域人为源排放影响较大,导致高低层昼夜分层明显外,春夏秋 3 个季节,傍晚至第二凌晨时段,受边界层逆温等气象因素影响,高低层分层较为明显,而在日出至日落时段,大气边界层垂直混合较为均匀,高低层 CO_2 浓度值较为接近,可以互为补充.

2.2 最优模型评估结果

根据 Liang 方法,采用 10 个评估分数的等权来综合评估模型的性能.将 10 个评估分数相加,得到每个模型的总分.利用该评分系统,可以定量地确定填补 CO_2 的最佳模型(各模型综合评分表见表 2).

表 2 各模型综合评分

Table 2 Comprehensive evaluationscores of each model

要素	评估模型	R^2	RMSE	MAE1	MAE2	MD	Precision	CvMAE	DRS	DRI	Efficiency	综合评分
CO_2_L1	KNeighborsDist	100	100	100	100	97.87	100	100	100	100	100	97.87
	KNeighborsUnif	97.73	95.11	95.63	96.6	97.94	95.06	95.82	98.76	98.76	97.35	96.86
	WeightedEnsemble_L2	99.96	99.87	99.75	99.69	97.83	99.88	100	99.46	99.44	50.75	94.67
	RandomForestMSE	90.39	82.79	81.96	81.52	95.59	82.58	82.15	92.02	92.02	1.79	78.72
	ExtraTreesMSE	85.92	76.92	72.66	70.46	99.48	76.7	73.04	83.34	83.38	5.01	72.37
	LightGBMLarge	53.91	44.75	48.06	61.24	97.88	44.13	48.73	78.49	78.47	0.71	55.63
	XGBoost	52.46	43.62	45.44	57.75	100	42.98	46.46	77.59	77.52	1.23	54.08
	LightGBM	49.31	41.21	40.41	51.07	97.89	40.65	41.52	70.62	70.61	1.67	50.02
	CatBoost	42.34	36.12	31.11	37.79	96.47	35.72	32.03	60.32	60.29	0.13	42.76
	LightGBMXT	34.74	30.68	24.35	31.55	94.88	30.39	25.57	55.14	55.11	1.03	37.88
	NeuralNetFastAI	4.04	11.24	0.43	0	98.91	11.93	2.41	20.08	19.77	0.24	16.7
	NeuralNetTorch	0	0	0	8.7	0	0	0	0	0	0	0.87

续表 2

要素	评估模型	R^2	RMSE	MAE1	MAE2	MD	Precision	CvMAE	DRS	DRI	Efficiency	综合评分
CO_2_L2	KNeighborsDist	100	100	100	100	99.26	100	100	100	100	100	90.71
	WeightedEnsemble_L2	100	100	100	100	99.26	100	100	100	100	94.35	90.14
	KNeighborsUnif	97.3	94.67	95.44	96.58	99.25	94.69	95.66	99.11	99.09	97.46	87.71
	RandomForestMSE	90.62	83.91	82.84	81.83	100	83.78	82.15	93.6	93.67	1.85	70.2
	ExtraTreesMSE	86.49	78.54	75.31	73.06	99.22	78.39	74.92	87.65	87.84	5.48	65.47
	XGBoost	55.36	45.92	51.26	62.16	94.65	45.94	50.32	85.02	85.25	0.86	48.5
	LightGBMLarge	54.73	45.45	52.07	63.6	96.21	45.45	50.8	83.73	83.97	0.66	48.48
	LightGBM	50.44	42.14	45.16	55.52	96.65	42.09	44.53	78.04	78.35	1.64	44.26
	CatBoost	43.78	37.09	36.4	43.73	96.64	37.05	35.37	71.26	71.72	0.11	38.12
	LightGBMXT	36.24	31.57	30.74	39.55	98.58	31.53	29.1	66.98	67.46	1	34.06
	NeuralNetFastAI	6.48	11.92	7.94	14.91	91.49	12.12	5.95	35.24	36.08	0.24	13.09
	NeuralNetTorch	0	0	0	0	0	0	0	0	0	0	10

从表 1 的性能评估结果可见,在众多参与评估的算法中,KNeighborsDist 算法的综合得分最高.进一步结合图 3,KNeighborsDist 算法的表现远超其他算法,这充分表明该算法在处理当前温室气体数据相关任务时具备显著优势.基于上述性能评估结果,本文将 KNeighborsDist 算法选定为核心算法,并围绕其开展后续的模型优化训练工作,以期充分发挥该算法的优异性能,提升模型对温室气体数据的处理精度和预测能力.与此同时,为深入探究 CO_2_L1 与 CO_2_L2 同时刻均有数据之间的关系,借助 sklearn 工具包对这两组数据进行了相关性检查.检查结果显示,二者之间的 R^2 值为 0.65.该值在一定程度上表明 CO_2_L1 与 CO_2_L2 同时刻均有数据之间存在一定的相关性,这一相关性分析结果为后续模型训练和数据处理提供了重要参考,有助于更全面地理解数据特征,优化模型参数,提高模型对温室气体数据的处理效果.

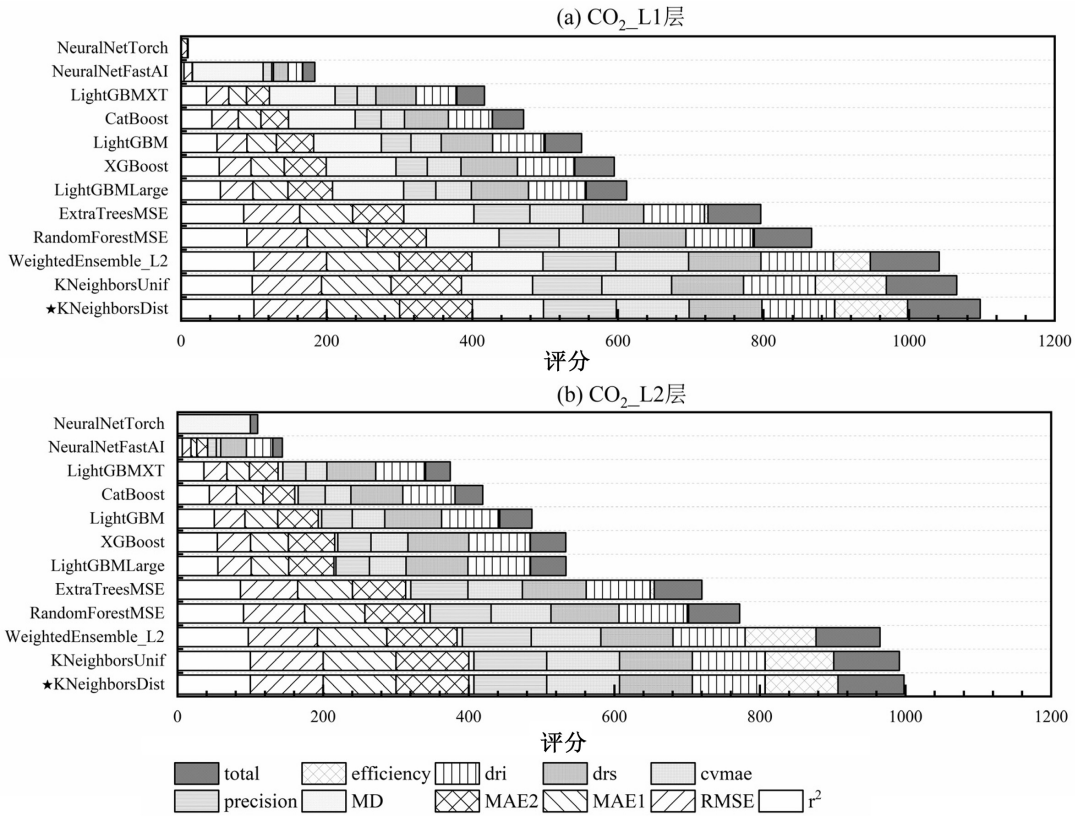


图 3　利用 12 种机器学习算法对 CO_2_L1 和 CO_2_L2 层数据的性能评估

Fig.3　Comprehensive performance evaluation scores of CO_2_L1 and CO_2_L2 layer data using 12 machine learning algorithms

为评估基于选定核心算法KNeighborsDist构建的模型对温室气体数据的重建效果,本研究利用测试数据集开展了针对性分析.该测试数据集选取自2014年5月12日~2015年12月31日的重叠期间,通过随机抽样选取了其中20%的数据点.通过绘制重建数据与测量数据的散点图(图4),可以直观地观察二者之间的分布关系.数据呈现出高度的线性相关趋势,且 R^2 均大于等于 0.98,充分表明使用测试数据集重建的温室气体数据与实际测量数据之间具有极好的相关性,也验证了所构建模型在温室气体数据重建任务上的有效性和准确性.

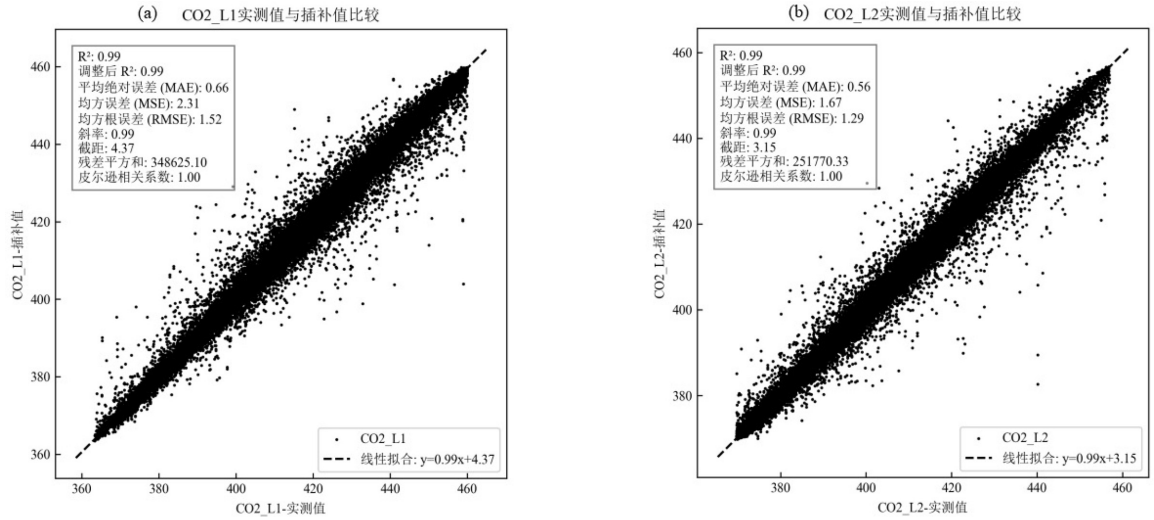


图4 CO_2_L1及CO_2_L2的实测值与填补值的相关性

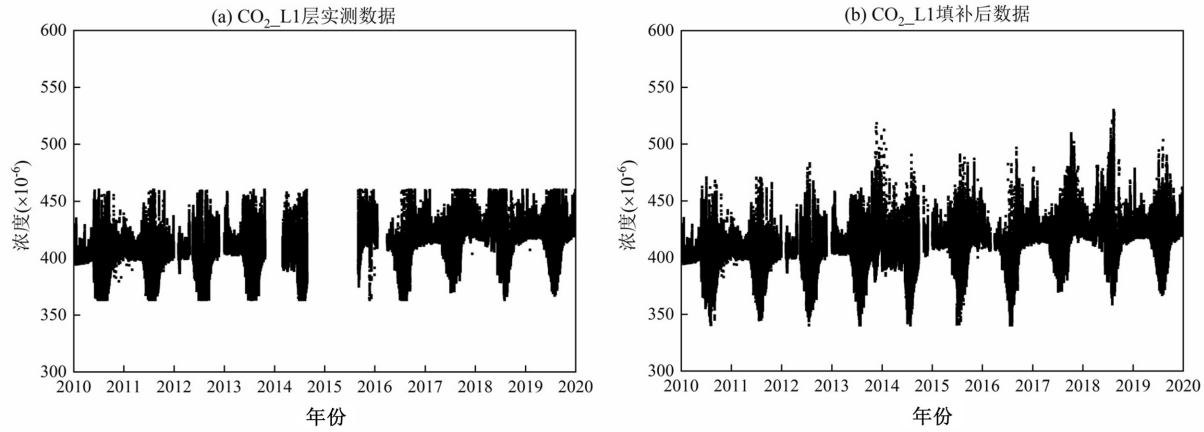
Fig.4 Correlation between measured and imputed CO_2 concentrations at L1 and L2 layers

2.3 龙凤山站本底大气 CO_2 浓度数据集重构

2.3.1 实测数据对比 据统计(图5),CO_2_L1层测量数据有995071条,填补后数据达到1417443条,填补比例为29.8%;CO_2_L2层测量数据有846280条,填补后数据达到1417443条,填补比例为40.3%;L1/L2层同一时刻都有数据的有432704条,机器学习模型补充比例为35.05%.

2.3.2 填补前后大气 CO_2 浓度年、月变化对比 从图6可以看出,原始实测值与填补后大气CO_2浓度的年均值均呈现显著的上升趋势,表明龙凤山站大气CO_2浓度长期持续增长.

从图7可以看出,各月份浓度变化趋势在填补后保持高度一致,未出现明显偏差.填补方法对月尺度波动的还原效果较好,说明数据填补过程考虑了月际差异.它精准地捕捉到了不同月份之间大气 CO_2 浓度的细微差异与变化规律,说明在数据填补的整个过程中,充分且细致地考虑到了月际差异这一关键因素,从而确保了填补后数据能够真实、准确地反映各月份大气 CO_2 浓度的实际变化情况.



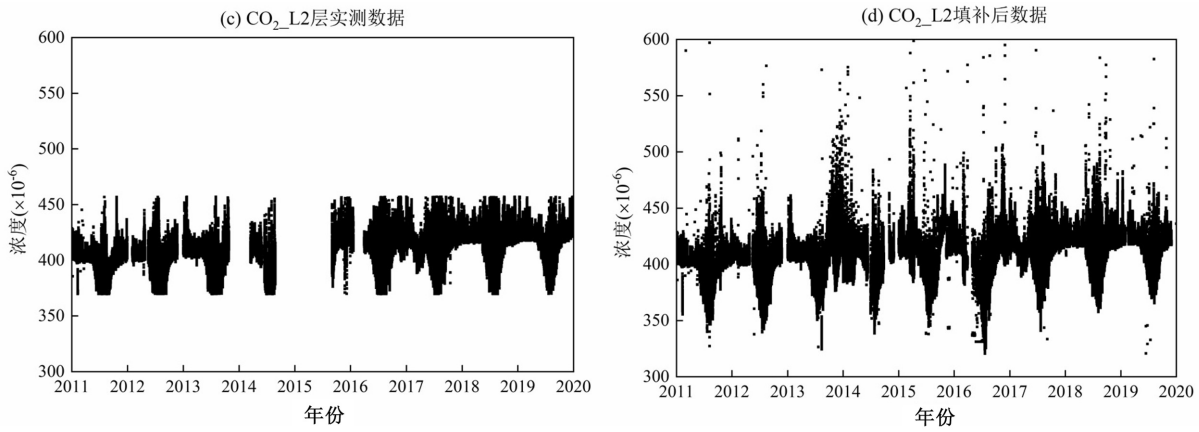


图 5　CO_2_L1 及 CO_2_L2 层实测数据及填补后数据时序图

Fig.5　Time sequence diagram of the actual measured data and the data after the reconstruction of the L1 and L2 level of Longfengshan Station

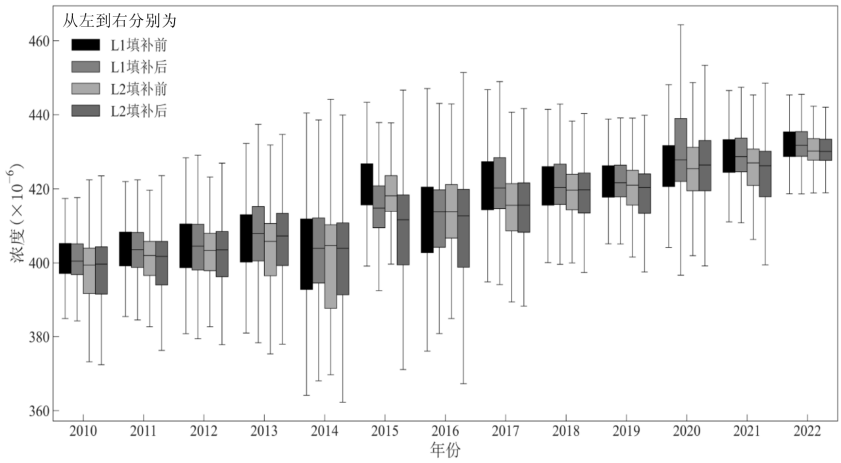


图 6　填补前后 CO_2年均值长期趋势时间序列

Fig.6　Long−Term trend of annual average CO_2 concentrations before and after imputation

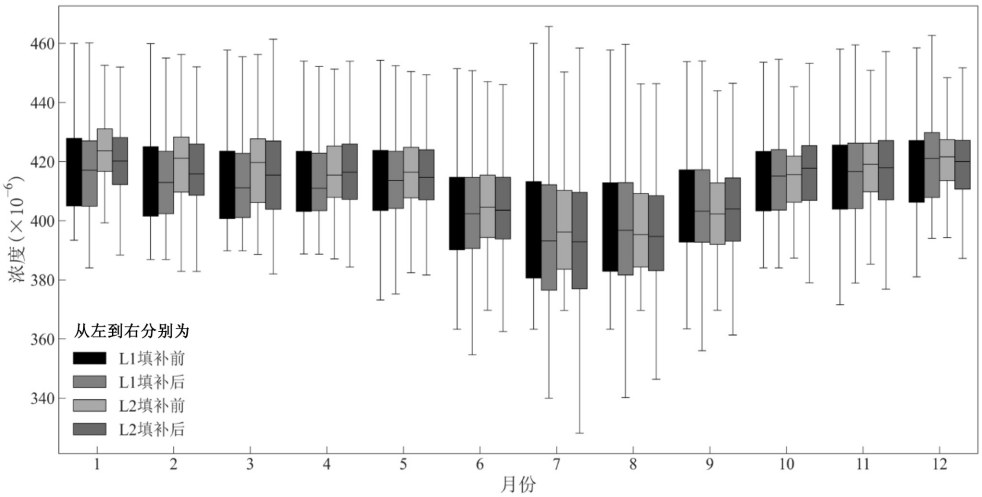


图 7　填补前后 CO_2多年均值月变化

Fig.7　Average monthly changes in CO_2 before and after filling

2.3.3　填补前后大气 CO_2浓度日变化对比　从图 8　　可见,填补处理后的数据所呈现的日变化规律(例如

早晚时段出现的浓度高峰)与原始数据高度吻合.这一结果清晰地表明,此次数据填补工作并未对日内周期性特征造成破坏,很好地保留了大气 CO_2 浓度在一天之内随时间变化的固有规律.特别值得一提的是,对于部分原本存在数据缺失的时段(比如夜间),通过所采用的模型进行填补后,得到了合理且符合实际情况的估计值.这也验证了模型在处理缺失数据方面的有效性.

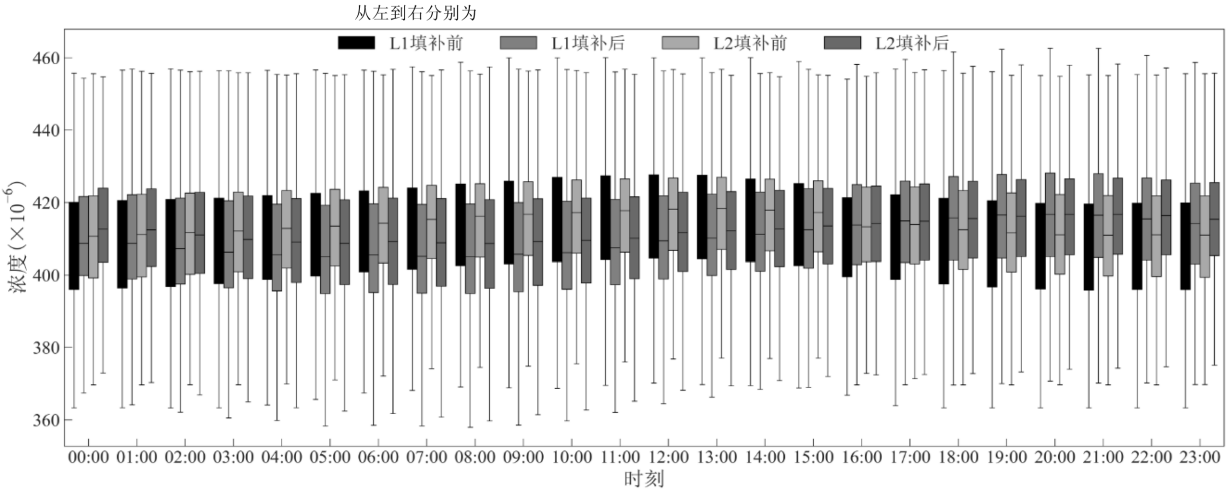


图 8 填补前后 CO_2多年均值日变化

Fig.8 Average daily changes in CO_2 before and after filling

3 结论

3.1 本研究通过系统评估多种机器学习算法,提出了基于 KNeighborsDist 算法的优化组合机器学习方法.

3.2 评估基于 KNeighborsDist 算法构建的模型对温室气体数据的重建效果,表明模型重建效果优异(R^2≥0.98),充分说明使用测试数据集重建的温室气体数据与实际测量数据之间具有良好的相关性,也验证了所构建模型在温室气体数据重建任务上的有效性和准确性.

3.3 填补后,龙凤山站数据完整率达到 100%,填补方法在保留 CO_2 长期趋势/季节/月/日变化规律上表现良好.

参考文献:

[1] World Meteorological Organization (WMO). The state of greenhouse gases in the atmosphere based on global observations through 2023 [R]. Geneva: World Data Centre for Greenhouse Gases, 2024.

[2] Masarie K A, Tans P P. Extension and integration of atmospheric carbon dioxide data into a globally consistent measurement record [J]. J. Geophys. Res., 1995,100(D6):11593-11610.

[3] Tsutsumi Y, Mori K, Ikebe T, et al. Technical report of global analysis method for major greenhouse gases by the world data center for greenhouse gases [R]. Tokyo: Global Atmosphere Watch Programme (GAW), 2009.

[4] 闫丽莉,温少妍,高文晶,等.整点气温缺测的插补方法研究及其初步应用 [J]. 震灾防御技术, 2019,14(2):446-455.
Yan L L, Wen S Y, Gao W J, et al. Interpolating method for missing data of integral point temperature and its preliminary application [J]. Technology for Earthquake Disaster Prevention, 2019,14(2):446-455.

[5] 司 鹏,郝立生,罗传军,等.河北保定气象站长序列气温资料缺测记录插补和非均一性订正 [J]. 气候变化研究进展, 2017,13(1):41-51.
Si P, Hao L S, Luo C J, et al. The interpolation and homogenization of Long-Term temperature time series at Baoding observation station in Hebei province [J]. Climate Change Research, 2017,13(1):41-51.

[6] Grange S K, Carslaw D C, Lewis A C, et al. Random forest meteorological normalisation models for swiss pm10 trend analysis [J]. Atmospheric Chemistry and Physics, 2018,18(9):6223-6239.

[7] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction [M]. New York: Springer, 2009.

[8] James G, Witten D, Hastie T, et al. An introduction to statistical learning with applications in R [M]. New York: Springer, 2013.

[9] Li X, Peng L, Yao X, et al. Long-term reconstruction of pm2.5 concentrations in China using a random forest model [J]. Environmental Pollution, 2020,263:114422.

[10] Wei J, Li Z, Lyapustin A, et al. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China [J]. Atmospheric Chemistry and Physics, 2019,19(13):8839-8855.

[11] 刘 冰,时凯歌,陈燕敏,等.基于可解释机器学习的地表水 DOM 光化学性质预测 [J]. 中国环境科学, 2025,45(7):3925-3934.
Liu B, Shi K G, Chen Y M, et al. Prediction of photochemical properties of surface water DOM based on interpretable machine learning [J]. Chinese Environmental Science, 2025,45(7):3925-3934.

[12] Liu Y, Cao G, Zhao N, et al. Improve ground-level $PM_{2.5}$ concentration

mapping using a random forest-based geostatistical approach [J]. Environmental Pollution, 2018,235:272−282.

[13] Ding N, Wang H, Xu T, et al. Estimating $PM_{2.5}$ concentrations using CatBoost and satellite remote sensing data [J]. Science of The Total Environment, 2021,750:141234.

[14] Wei J, Huang W, Li Z, et al. Estimating 1-km daily $PM_{2.5}$ concentrations across China using the lightGBM model [J]. Environmental Science & Technology, 2020,54(11):6796−6805.

[15] Dorogush A V, Ershov V, Gulin A. Cat Boost: gradient boosting with categorical features support [J]. arXiv.org, 2018,29:1189−1232.

[16] Wu C, Lin H, Xiao Q, et al. High-resolution mapping of $PM_{2.5}$ using lightGBM and satellite data [J]. Remote Sensing of Environment, 2021,254:112256.

[17] Di Q, Rowland S, Koutrakis P, et al. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States [J]. Journal of the Air & Waste Management Association, 2016,67(1): 39−52.

[18] Brokamp C, Jandarov R, Rao M B, et al. Exposure assessment models for particulate matter using machine learning [J]. Environmental Health Perspectives, 2017,125(6):067010.

[19] Di Q, Amini H, Shi L, et al. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution [J]. Environment International, 2019,130: 104909.

[20] Wu C, Yu J Z. Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting [J]. Atmospheric Measurement Technigues, 2018,11(2):1233−1250.

[21] Just A C, Arfer K B, Schwartz J, et al. A new method for estimating missing air pollution data using machine learning [J]. Atmospheric Environment, 2020,224:117323.

[22] Reid C E, Jerrett M, Petersen M L, et al. Spatiotemporal prediction of fine particulate matter using machine learning [J]. Environmental Science & Technology, 2015,49(6):3870−3879.

[23] van Donkelaar A, Martin R V, Brauer M, et al. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors [J]. Environmental Science & Technology, 2016,50(7):3762−3772.

[24] Kloog I, Ridgway B, Koutrakis P, et al. Long- and short-term exposure predictions using satellite-based $PM_{2.5}$ estimates and machine learning [J]. Environmental Health Perspectives, 2014,122(10):1069−1075.

[25] Chen G, Li S, Knibbs L D, et al. A two-stage machine learning model for estimating daily $PM_{2.5}$ concentrations in China [J]. Environment International, 2018,114:129−137.

[26] 栾　天,周凌晞,方双喜,等.龙凤山本底站大气 CO_2 数据筛分及浓度特征研究 [J]. 环境科学, 2014,35(8):2864−2870.
Luan T, Zhou L X, Fang T X, et al. Atmospheric CO_2 data filtering nethod and characteristics of the molar fractions at the longfengshan wmo/gaw regional station in China [J]. Environmental Sciences, 2014, 35(8):2864−2870.

[27] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient bosting decision tree [J]. Advances in Neural Information Processing Systems, 2017,30:3146−3154.

[28] Dorogush A V, Ershov V, Gulin A. CatBoost: unbiased boosting with categorical features [J]. Advances in Neural Information Processing Systems, 2018,31:6638−6648.

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

[30] Erickson N, Mueller J, Shirkov A, et al. AutoGluon-Tabular: Robust and accurate autoML for structured data [J]. arXiv preprint arXiv: 2003.06505, 2020.

[31] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning [M]. Springer, 2009.

[32] Finn D, Clawson K L, Carter R G, et al. Tracer studies to characterize the effects of complex terrain on dispersion [J]. Journal of Applied Meteorology, 2010,49(7):1420−1436.

[33] Stull R B. An introduction to boundary layer meteorology [M]. Springer Science & Business Media, 1988.

[34] Horst T W, Weil J C. How far is far enough: The fetch requirements for micrometeorological measurement of surface fluxes [J]. Journal of Atmospheric and Oceanic Technology, 1994,11(4):1018−1025.

[35] Leclerc M Y, Thurtell G W, Kidd G E, et al. Measurements of the footprint of scalar sources using a laser anemometer [J]. Boundary−Layer Meteorology, 1990,52(3):247−258.

作者简介：顾桃峰(1982−),江苏苏州人,副研级工程师,硕士,主要从事气象探测技术、气象信息技术的研究.发表论文 35 篇.gutaofeng@gz.gov.cn.