北京大学学报(自然科学版) 第 57 卷 第 5 期 2021 年 9 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 57, No. 5 (Sept. 2021)
doi: 10.13209/j.0479-8023.2021.070

# 基于机器学习方法的臭氧和$PM_{2.5}$污染潜势预报模型——以成都市为例

王馨陆[1] 黄冉[1,†] 张雯娴[1] 吕宝磊[2] 杜云松[3] 张巍[3] 李波兰[3] 胡泳涛[4]

1. 杭州矮马科技有限公司, 杭州 311121; 2. 华云升达(北京)气象科技有限责任公司, 北京 102299; 3. 四川省生态环境监测总站, 成都 610091; 4. School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332; † 通信作者, E-mail: ranhuang2019@163.com

**摘要** 以成都市为例, 以多项可能影响污染物时空分布的变量为潜在预报因子, 筛选关键入模因子, 利用2016—2018年数据为训练集, 采用多元线性回归、BP神经网络和随机森林算法, 建立成都市夏季(4—8月)臭氧及冬季(11—2月)$PM_{2.5}$污染潜势模型, 并利用2019年数据对模型的中长期污染潜势浓度的预报性能进行评估。结果表明, 建立的多元线性回归、BP神经网络和随机森林模型对成都市臭氧及$PM_{2.5}$的短期(1~3天)污染潜势都具有良好的预报效果, 对7~15天的中长期潜势预报表现稳定。其中, 多元线性回归模型和随机森林模型分别对臭氧和$PM_{2.5}$表现出相对最佳的预报性能。

**关键词** 多元线性回归; BP神经网络; 随机森林; 中长期潜势预报

# Forecasting Ozone and $PM_{2.5}$ Pollution Potentials Using Machine Learning Algorithms: A Case Study in Chengdu

WANG Xinlu[1], HUANG Ran[1,†], ZHANG Wenxian[1], LÜ Baolei[2], DU Yunsong[3], ZHANG Wei[3], LI Bolan[3], HU Yongtao[4]

1. Hangzhou AiMa Technologies, Hangzhou 311121; 2. Huayun Sounding Meteorological Technology Company, Ltd., Beijing 102299; 3. Sichuan Bio-Environmental Monitoring Center, Chengdu 610091; 4. School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332; † Corresponding author, E-mail: ranhuang2019@163.com

**Abstract** Potential forecast models have been developed for air pollution of summertime (Apr.–Aug.) ozone and wintertime (Nov.–Feb.) $PM_{2.5}$ in Chengdu using the multiple linear regression (MLR), back-propagation (BP) neural network (NN) and random forest (RF) algorithms. The key predicting factors for each of the models are selected from various potential factors that may impact the spatiotemporal distribution of pollutions. The models are trained and established with 2016–2018 datasets and evaluated with a data-withheld method and further with independent 2019 dataset. The results show that the MLR, NN and RF models are all capable to accurately predict $O_3$ and $PM_{2.5}$ pollution potentials in short lead-time (1–3 days) in Chengdu. The models are also found having quite stable performances in medium- and long-term (7–15 days lead time) forecasts. Among the three models, the MLR model performs the best in prediction of $O_3$, while RF model performs the best for $PM_{2.5}$.

**Key words** multiple linear regression; BP neural network; random forest; medium- and long-term air pollution potential forecast

环境空气质量的好坏对公众健康有着显著影响, 不论是极端重污染事件还是长期暴露于低浓度空气污染环境中, 均会直接增加人体心血管和呼吸系统等多种疾病的发病率[1–2]。近年来, 我国大多数城市的空气质量持续改善, 尤其是秋冬季细颗粒物($PM_{2.5}$, 空气动力学直径小于或等于2.5 μm的气

国家重点研发计划(2018YFC0214004)、四川省环境保护科技计划(2019HB03)和四川省重大科技专项(2018SZDZX0023)资助
收稿日期: 2020–07–31; 修回日期: 2020–09–15

溶胶粒子)污染程度下降明显[3−4]，但春夏季臭氧污染呈现上升趋势[5−6]。空气污染物源排放是影响空气质量的决定性因素，天气形势及气象条件亦为关键因素。气象条件的变化直接或间接地影响大气中污染物的化学反应、传输、扩散稀释和沉降等过程[7−11]，对空气质量的影响呈现多时空尺度、影响大及变化快的特点[12]。对一定的区域而言，如果短期内污染源排放相对稳定，其空气质量则主要取决于气象条件[13−14]。当出现静稳天气等不利气象条件时，污染物浓度容易在短时间内出现大幅增长，造成严重的空气污染事件[15−18]。因此，研究天气形势及气象条件对污染物在大气中传输和转化的影响，开展空气污染潜势预报预警，对评估气象条件对空气污染的贡献以及辅助大气环境精细化管理和科学决策具有重要意义。

污染潜势预报是在假定污染源排放不变的情况下，以可能影响污染物时空分布的天气形势及气象条件为主要依据，对未来气象条件下的空气污染状况进行预测[19−21]。其特点在于忽略不确定的污染源排放速率的变化，重点关注有利或不利于污染物扩散稀释等过程的气象因素[22]，将气象因素对空气质量的影响分离出来[23]，是评估气象条件对污染物浓度影响及贡献的重要方法之一。众多研究采用逐步多元线性回归的方法建立气象因子(如风速、相对湿度等)与污染物浓度(如 $PM_{2.5}$ 和臭氧)之间的污染潜势模型[24−28]，量化气象条件变化对污染物浓度变化的贡献。Zhai 等[27]以中国地面气象观测日值数据及 MERRA2 再分析数据中的风速、降水、相对湿度、气温和 850 hPa 经向风等作为潜在预报变量，采用逐步多元线性回归法建立 2013—2018 年中国主要地区的 $PM_{2.5}$ 污染潜势预报模型，定量分析气象条件对 $PM_{2.5}$ 污染变化的贡献，结果表明在中国 $PM_{2.5}$ 浓度下降的趋势中，气象贡献占 12%。张小曳等[29]利用国家自动气象站逐小时地面气象观测数据及欧洲中期天气预报中心的再分析数据，对与气溶胶浓度密切相关的气象要素(如风速、风向和大气稳定度等)进行诊断和参数化分析，得到可定量反映停滞–静稳型天气程度的“污染–气象条件”指数(PLAM 指数)，建立气溶胶浓度与气象要素之间的量化关系，并分析评估了 2013 年《大气污染防治行动计划》实施以来气象条件变化对 $PM_{2.5}$ 污染变化的影响。

数值预报计算量大，计算成本高，依赖于大量输入数据(如源排放清单和气象场)的驱动，与之相比，基于各种机器学习算法的空气污染潜势预报较为简单易行，且无需源排放清单，已广泛应用于各项研究中[30−31]，具有较好的预报效果。不同于数值预报模式中以大气污染物转化扩散的化学和物理机制为基础[32]，基于统计方法的污染潜势预报主要利用大量污染监测历史数据及同期气象观测资料，分析污染物浓度与相关辅助因子之间的统计关系，建立从简单相关到复杂多参数的模型，从而进行未来空气质量的预测[19,22−23]。常见的潜势预报方法包括多元线性回归[33−35]、支持向量机[36−37]、决策树[30,38]、随机森林[39−40]和人工神经网络[41−43]等。Lightstone 等[44]利用 2016 年 NCEP/NARR 再分析资料及 NYSDEC 地面监测网的 $PM_{2.5}$ 数据，建立纽约市 $PM_{2.5}$ 神经网络预报模型，并与 CMAQ 12 km 网格数值模式模拟结果进行对比，结果表明神经网络模型准确性更好，尤其是对传输引起的污染浓度快速变化时段的模拟。

本研究利用成都市 2016—2019 年 WRF 模式回溯模拟气象场及同期空气质量观测数据，以影响污染物转化、扩散和传输的主要气象因子及相关辅助因子为潜在预报因子，通过筛选关键入模变量，利用多元线性回归、随机森林及 BP (back-propagation)神经网络等机器学习算法，建立成都市夏季(4—8 月) $O_3$ 及冬季(11 月—来年 2 月) $PM_{2.5}$ 浓度污染潜势预报模型，对比分析各模型对成都市 $O_3$ 及 $PM_{2.5}$ 污染的预测效果，并检验建立的污染潜势模型的中长期预报能力。

# 1 研究数据与方法

## 1.1 研究数据

### 1.1.1 空气质量数据

本研究使用的 2016—2019 年成都市逐日臭氧及 $PM_{2.5}$ 环境浓度观测数据来自四川省空气质量监测网络管理平台(http://www.scnewair.cn:3389)。成都市 2016—2019 年 $O_3$ 日最大 8 小时浓度在每年的 4—8 月达到污染高峰期，$O_3$ 超标事件($O_3$≥160 μg/m$^3$)频发(图 1)，4—8 月的多年累月平均浓度分别为 114.5，128.2，126.2，131.2 和 143.7 μg/m$^3$。$PM_{2.5}$ 日均浓度的污染高峰期主要发生在每年的 11 月至来年 2 月(图 1)，11—2 月的多年累月平均浓度分别为 65.2, 89.9, 93.5 和 69.7 μg/m$^3$。

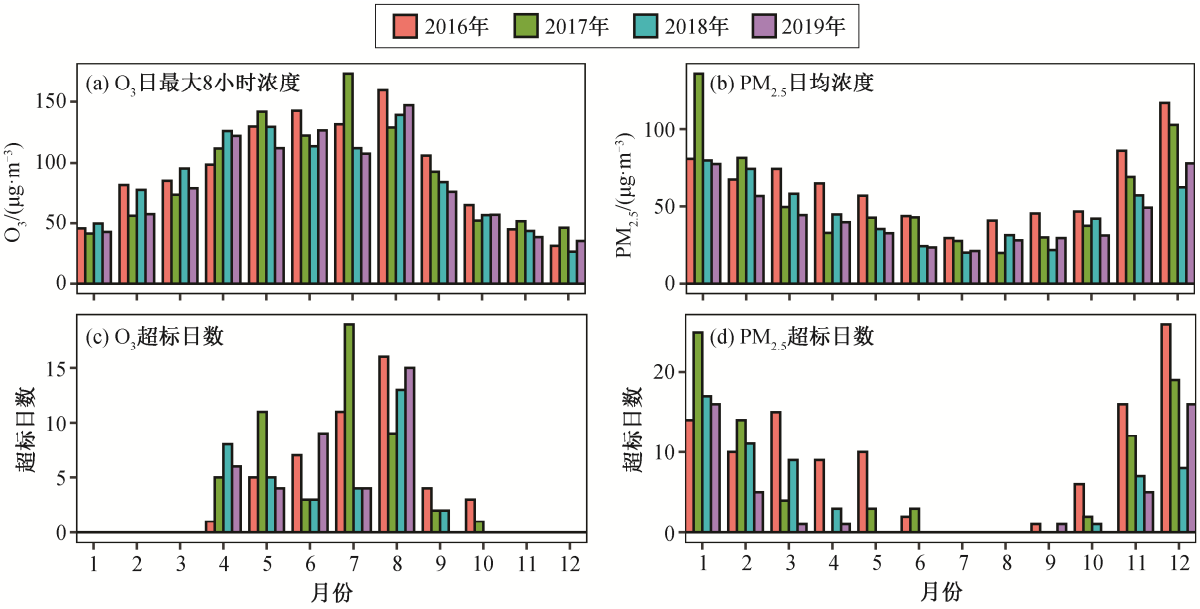



**图 1 成都市 2016—2019 年 $O_3$ 日最大 8 小时浓度及 $PM_{2.5}$ 日均浓度的月平均及每月超标日数 ($O_3 \geq 160$ μg/m$^3$, $PM_{2.5} \geq 75$ μg/m$^3$)统计**

Fig. 1 Monthly mean of daily maximum 8-hr average $O_3$ and daily average $PM_{2.5}$ concentrations, and monthly number of exceedance days ($O_3 \geq 160$ μg/m$^3$, $PM_{2.5} \geq 75$ μg/m$^3$)

考虑到成都市 $O_3$ 和 $PM_{2.5}$ 污染以及各气象因子的显著季节波动, 为提高所建模型的可靠性、准确性和实用性, 本文针对成都市每年 4—8 月和 11—2 月分别建立 $O_3$ 和 $PM_{2.5}$ 的污染潜势预报模型。

### 1.1.2 气象数据

本研究使用的成都市 2016—2019 年气象数据来自中尺度预报模式 WRF (Weather Research and Forecast Model, 版本 3.6)[45]的气象回溯模拟结果。该回溯模拟采用基于 Lambert 投影坐标的 36 km、12 km 和 4 km 水平分辨率的 3 重嵌套网格(范围见附录 1, 分别为紫色、蓝色和红色矩形区域), 最外层网格覆盖包括青藏高原在内的所有中国地区和东亚以及部分东南亚和印度次大陆, 次内层网格包括四川省全省及西南地区各省(市、区)的大部分区域, 最内层网格覆盖四川盆地的主要城市, 垂直方向采用从地面到 50 hPa 共 35 个 $\sigma$ 层。模拟中以 NCEP GDAS/FNL 0.25°×0.25° 全球再分析资料作为初始条件和边界条件, 主要物理过程采用 Lin 微物理参数化方案[46]、Kain-Fritsch 积云方案[47]、YSU 边界层参数化方案[48]以及 NOAH+MOSAIC 陆面模式[49]。此外, 在模拟过程中启用 Grid Nudging 同化技术[50−51], 利用 NCEP ADP 全球地面及探空气象观测数据, 对逐 6 小时猜测场进行“校正”, 并在 WRF 计算过程中通过同化技术优化模拟结果。利用中国地面气象观测站逐小时数据, 对 2016—2019 年 WRF 回溯模拟结果进行评估(详见附录 2), 各评估统计指标都处于合理的可接受范围[52], 表明气象回溯模拟数据可进一步用于成都市污染潜势预报模型的建立及后续的预报能力评估。

本研究以可能影响 $O_3$ 及 $PM_{2.5}$ 污染的气象及相关辅助因子为潜在预报变量, 建立污染潜势模型, 重点在于识别影响空气质量的关键预报因子。瞬时多变的天气形势及气象条件对空气质量的影响极为复杂, 不同气象条件和相关辅助因子对不同污染物的作用各不相同, 又相互影响。为了尽可能准确地识别影响 $O_3$ 和及 $PM_{2.5}$ 污染的关键预报因子, 本研究拟定 39 个潜在的预报因子(详见附录 3), 主要包含污染持续性因子(如前一日的污染物浓度)、节假日和工作日信息[53−55]以及相关气象条件因子(如风速、气温、湿度和云量等)[12−13,22,24,56], 并利用 WRF 回溯模拟结果建立潜在预报因子数据集, 以便后续关键预报因子的筛选。

## 1.2 研究方法

以成都市 2016—2019 年 $O_3$ 及 $PM_{2.5}$ 的日值观测数据和 1.1.2 节建立的包含 39 个潜在预报因子的数据集为基础, 筛选关键预报因子, 并分别建立训练、测试和评估数据集。采用多元线性回归(Multiple Linear Regression, MLR)[57]、随机森林(Random

Forest, RF)[58]以及 BP 神经网络(Back-Propagation Neural Network, NN)[43]3 种机器学习算法, 建立成都市夏季 $O_3$ 及冬季 $PM_{2.5}$ 污染潜势预报模型, 并进行验证和评估。图 2 为建立污染潜势预报模型的技术路线。

### 1.2.1 关键入模变量的筛选

首先进行预报因子的筛选, 确认影响成都市夏季 $O_3$ 及冬季 $PM_{2.5}$ 浓度的关键入模变量。采用基于随机森林算法的变量重要性分析工具进行潜在变量的初步筛选, 然后根据相关性及不同组间的差异性分析, 最终选定入模变量。

1) 以潜在预报因子数据集中的 39 个变量为自变量, 分别以成都市 2016—2019 年的 $O_3$ 及 $PM_{2.5}$ 浓度为因变量, 利用随机森林算法进行潜在预报因子的重要性分析, 降序排列选择其中前 25 个变量为初步选定的潜在入模因子。分别计算上述步骤初步选定的 $O_3$ 及 $PM_{2.5}$ 的 25 个入模变量间的相关系数矩阵(详见附录 4 和 5), 可见其中存在大量高度相关的相似变量, 进一步剔除相关系数高于 0.7 的相对不重要变量, 达到删除多余相似变量的目的, 避免高度相关变量进入模型中可能导致的严重的多重共线性问题[59–60]并减少模型训练过程中的计算量。

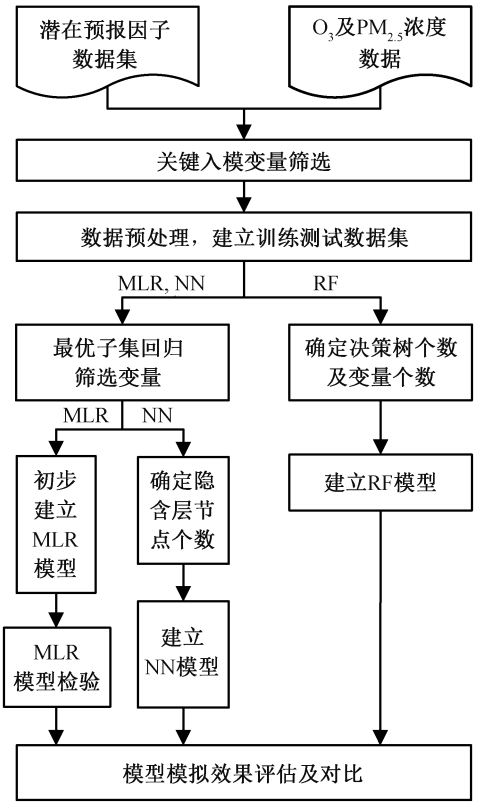



**图 2　建立污染潜势预报模型的技术路线**

Fig. 2　Flow chart of building the air pollution potential forecasting models

2) 分别分析 $O_3$ 及 $PM_{2.5}$ 浓度与上一步筛选出的对应潜在入模因子的相关性, 并根据国家一级及二级标准(GB/T 3095—2012 环境空气质量标准), 分别划分 $O_3$ 和 $PM_{2.5}$ 污染的清洁日($O_3$<100 μg/m$^3$, $PM_{2.5}$<35 μg/m$^3$)和污染日($O_3$>160 μg/m$^3$, $PM_{2.5}$>75 μg/m$^3$), 利用 t 检验对在清洁日与污染日潜在入模因子的差异性进行分析, 选择具有显著相关性及显著差异的因子分别作为 $O_3$ 和 $PM_{2.5}$ 潜势预报模型的最终关键入模因子。

通过上述步骤, 最终选定成都市臭氧污染的关键入模变量为 T_MAX(地面每日最高气温)、PBL_MAX(每日边界层高度最大值)、O3_YEST(前一日臭氧平均浓度)、HCC(每日平均高云量)、MCC(每日平均中云量)、WS850(850 hPa 每日平均风速)、WS_AFTE(地面下午时段平均风速)、PR(每日降水总量)、PS_DELTA_YEST(前一日 24 小时变压)、WD(地面每日最多风向)及 WD700(700 hPa 每日最多风向)。$PM_{2.5}$ 的关键入模变量为 PM2.5_YEST(前一日 $PM_{2.5}$ 平均浓度)、PBL(每日平均边界层高度)、WS(地面每日平均风速)、T700_MAX(700 hPa 每日最高气温)、PS_DELTA(当日 24 小时变压)、WD_CHANGE(风向日变化因子)、PS_DELTA_YEST(前一日 24 小时变压)、PR(每日降水总量)、WS500(500 hPa 每日平均风速)、GHT500 (500 hPa 每日平均位势高度)及 WD(地面每日最多风向)。

### 1.2.2 数据预处理

在正式建立预报模型之前, 需要对数据进行预处理, 包括归一化处理、污染物浓度对数化处理及风向相关变量特殊处理等。

1) $O_3$ 及 $PM_{2.5}$ 浓度数据为对数正态分布, 对相关变量($PM_{2.5}$, $O_3$, $PM_{2.5}$_YEST 和 $O_3$_YEST)进行自然对数化处理, 处理完成后的数据主要用于 MLR 及 NN 模型的建立。

2) 为消除量纲的影响, 对各变量数据做归一化处理, 处理完成后的数据用于 MLR, NN 及 RF 模型的建立。

3) 针对类别型变量(WD, WD700, IF_HOLIDAY 和 IF_WEEK)进行特殊处理。在 RF 模型的建立中, 对上述 4 个变量进行因子化处理; 在 MLR 及 NN 模型的建立中, 则分别构建新的虚拟变量, 如 WD 变量共包含 17 个因子水平(N, NNE, NE, ENE, E, ESE,

SE, SSE, S, SSW, SW, WSW, W, WNW, NW, NNW 和 C), 因此新建 16 个虚拟变量(WD.N, WD.NNE, WD.NE, WD.ENE, WD.E, WD.ESE, WD.SE, WD.SSE, WD.S, WD.SSW, WD.SW, WD.WSW, WD.W, WD.WNW, WD.NW 和 WD.NNW)。若所有虚拟变量为 0, 则代表 WD 为 C; 若 WD.N 为 1 且其他虚拟变量为 0, 则代表 WD 为 N; 依此类推。

此外, 随机抽取 2016—2018 年 75%的数据作为模型训练数据集, 剩余 25%的数据为测试数据集, 保留 2019 年数据为回顾预报数据集, 用于模型建立完成后对预报效果进行独立评估。

#### 1.2.3 模型训练及优化

本研究利用建立的训练数据集, 分别采用多元线性回归(MLR)、BP 神经网络(NN)和随机森林(RF)3 种方法训练, 建立成都市夏季臭氧及冬季 $PM_{2.5}$ 污染潜势预报模型。

1) MLR 模型: 在数据预处理过程中, 针对类别型变量新建了大量的虚拟变量, 首先利用最优子集回归法进一步筛选变量, 基于马洛斯 Cp 准则、贝叶斯信息量准则和修正 $R^2$ 选择最佳的变量组合, 建立初步的 MLR 模型, MLR 模型建立完成后, 进行模型的诊断及显著性检验, 并利用方差膨胀因子进行共线性分析和模型优化, 确定相对最优的 MLR 模型。

2) NN 模型: 采用最优子集回归法确定最佳变量组合, 建立 NN 模型。设置隐含层层数为 1, 采用十折交叉检验确定隐含层神经元个数, 建立相对最优的 NN 模型。

3) RF 模型: 采用筛选的关键入模变量建立 RF 模型, 通过诊断测试抽样的特征个数和森林决策树的个数等参数对 RF 模型的影响, 确定最优的参数组合, 建立相对最优的 RF 模型。

#### 1.2.4 模拟和预报效果评估

对建立的“最优”MLR, NN 和 RF 模型在训练集和测试集中的表现进行评估, 并分析模型的泛化能力; 利用建立的模型对 2019 年的 $O_3$ 及 $PM_{2.5}$ 浓度进行回顾预报, 进一步验证 3 种模型的预报模拟能力。用于评估模拟效果的统计量包括相关系数($R$)、平均偏差(Bias)、平均绝对误差(GE)、均方根误差(RMSE)以及分类误判率。

## 2 结果与讨论

### 2.1 成都市臭氧污染潜势预报模型

在成都市 2016—2018 年数据中随机选取 75%作为训练数据集, 剩余 25%的数据作为测试集, 利用多元线性回归、BP 神经网络及随机森林算法进行模型训练, 分别建立成都市臭氧污染潜势预报 MLR, NN 及 RF 模型(各模型的参数设置列于附录 6), 并评估各模型在训练数据集和测试数据集中的模拟表现(表 1 和附录 7)。MLR 和 NN 两个模型在训练集和测试集中的表现相对稳定, 性能接近。与训练集相比, 两个模型在测试集中的相关性有所降低, 误差值略有增大, 但仍处于合理的可接受范围内。RF 模型在训练集中的综合表现最优, 其相关系数高达 0.98, BIAS, GE, RMSE 和分类误判率分别为−0.22, 9.09, 11.98 和 8.93, 均明显优于 MLR 及 NN 模型在训练集中的模拟表现。在测试集中, RF 模型的相关系数显著降低 22.4%, GE, RMSE 和分类误判率等误差指标分别增加 148%, 150%和 300%, 模拟能力显著降低, 但仍与 MLR 及 NN 模型在测试集中的评估结果接近。可见, 尽管 RF 模型存在明显的过拟合问题, 但依旧保持较好的模拟能力。综上所述, 利用多元线性回归、BP 神经网络、随机森林算法训练建立的 MLR, NN 以及 RF 模型的模拟表现较为接近, 都能够对成都市夏季臭氧污染进行良好的预测。

利用上述建立的 MLR, NN 及 RF 模型, 对成都市 2019 年 4—8 月的臭氧污染进行回顾预报模拟, 对模型的独立预报能力进行评估(表 1 和图 3)。该回顾预报可理解为提前一天(1-day lead)的污染潜势预报。MLR, NN 及 RF 模型在回顾预报集中的模拟值与观测值的相关系数位于 0.75~0.77 之间, 除 BIAS 指标外, GE, RMSE 及分类误判率等误差结果

**表 1 成都市臭氧污染潜势模型模拟效果评估**

Table 1 Evaluation of the ozone pollution potential forecast models in Chengdu

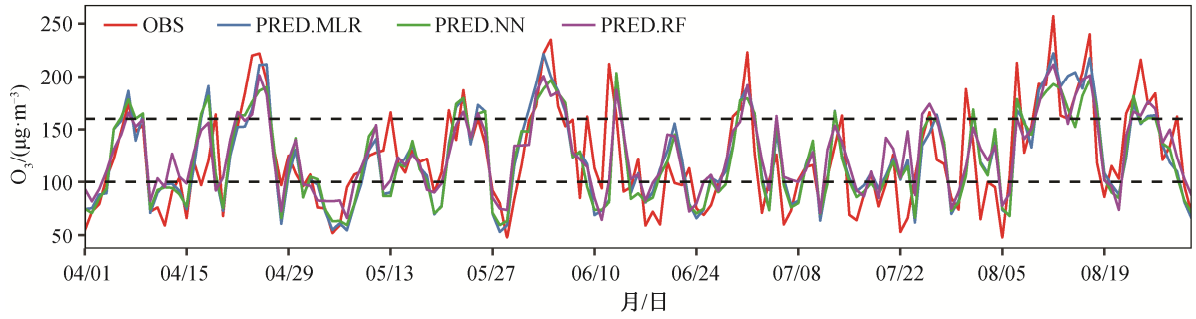
| 方法 | 数据集 | $R$ | BIAS | GE | RMSE | 分类误判率/% |
|---|---|---|---|---|---|---|
| MLR | 训练集 | 0.83 | −2.63 | 20.29 | 25.77 | 26.80 |
| | 测试集 | 0.73 | −1.39 | 23.79 | 31.48 | 32.14 |
| | 回顾预报集 | 0.77 | 0.16 | 23.58 | 30.45 | 37.91 |
| NN | 训练集 | 0.82 | −2.77 | 20.58 | 26.46 | 27.38 |
| | 测试集 | 0.73 | −1.36 | 23.05 | 31.17 | 32.14 |
| | 回顾预报集 | 0.75 | −1.29 | 23.96 | 31.08 | 39.22 |
| RF | 训练集 | 0.98 | −0.22 | 9.09 | 11.98 | 8.93 |
| | 测试集 | 0.76 | 3.88 | 22.62 | 29.96 | 35.71 |
| | 回顾预报集 | 0.77 | 3.53 | 23.05 | 29.68 | 35.95 |

较为一致(表 1)。对比在测试集中的表现，3 个模型在回顾预报集中的评估指标结果并无明显差异，可见 MLR，NN 及 RF 模型的表现均较为稳定。此外，虽然 MLR，NN 及 RF 模型的模拟结果存在一定的定量方面问题(图 3)，表现在对臭氧高峰值存在一定的低估(如 8 月 5—19 日期间的 3 个高峰值)或漏报(如 5 月 13 日)或 1~2 天的迟滞(如 6 月 12 日)，对低谷时段则存在一定的高估(如 6 月 17—19 日)，但模拟值与观测值之间的时间变化趋势保持良好的一致性，可见 3 个模型都能对成都市 2019 年夏季臭氧进行较好的模拟。模型之间相较而言，MLR 及 RF 模型在定量方面能够更好地再现臭氧高污染时段，更接近污染高峰观测值，其中 RF 模型虽具有更小的 GE，RMSE 及分类误判率，但在整体上存在一定的高估(其在测试集及回顾预报集中的 BIAS 分别为 3.88 和 3.53)，在某些时段的变化趋势识别上不如 MLR 模型精准。整体而言，在 3 个模型中，MLR 模型具有最好的预报能力。

## 2.2 成都市 $PM_{2.5}$ 污染潜势预报模型

同样针对成都市冬季(11—2 月) $PM_{2.5}$ 污染建立 MLR，NN 及 RF 潜势预报模型，模型在训练集和测试集中的结果详见 附录 8 以及表 2。在训练集和测试集中，MLR 及 NN 模型的各项评估结果较为接近，且 MLR 和 NN 模型在测试集中的模拟能力反映在相关系数上与训练集无明显差别，GE，RMSE 和分类误判率则略有降低。RF 模型在训练集中的相关系数最大，GE，RMSE 及分类误判率等各项误差最小。RF 模型在测试集中的表现整体上与 MLR 和 NN 模型相似，但对比其在训练集中的表现，相关性明显降低，各项误差(GE，RMSE 和分类误判率)显著增大，可见 RF 模型依旧存在一定程度的过拟合问题。MLR，NN 和 RF 模型对 $PM_{2.5}$ 污染潜势的模拟能力较为相似，表现稳定，能够对成都市冬季 $PM_{2.5}$ 污染进行较好的模拟，且模拟效果(表 2)优于其在臭氧污染潜势模拟中的表现(表 1)。

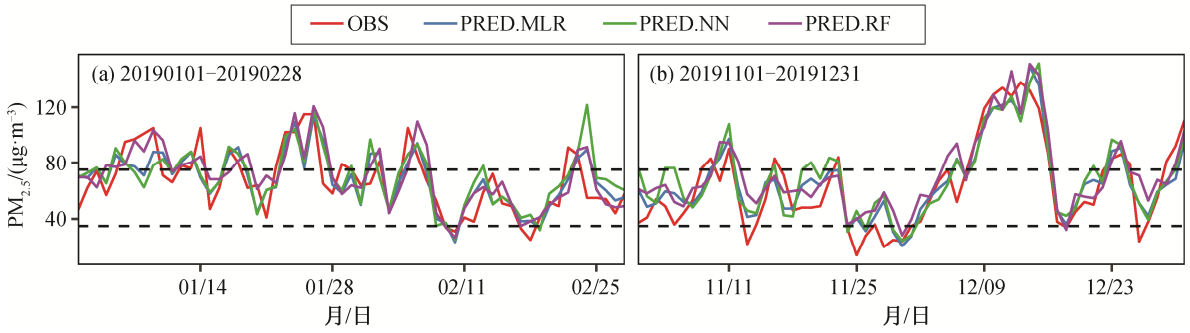
对成都市 2019 年 1—2 月和 11—12 月的 $PM_{2.5}$ 浓度进行回顾模拟，评估建立的污染潜势预报模型的预报能力(表 2 和图 4)。MLR 及 RF 模型的预报性能整体上较为稳定，与测试集中的评估结果接近。这两个模型预测值与观测值的相关系数分别为 0.83




黑色虚线分别代表臭氧浓度为 100 μg/m$^3$ 及 160 μg/m$^3$

**图 3　成都市 2019 年夏季 $O_3$ 浓度观测值及 MLR, NN 和 RF 模型模拟值时间序列**

Fig. 3　Timeseries of $O_3$ concentrations: observed versus simulated by MLR, NN and RF models for Chengdu in summer 2019




黑色虚线分别代表 $PM_{2.5}$ 浓度为 35 μg/m$^3$ 及 75 μg/m$^3$

**图 4　成都市 2019 年冬季 $PM_{2.5}$ 污染观测值及 MLR, NN 和 RF 模型模拟值时间序列**

Fig. 4　Timeseries of $PM_{2.5}$ concentrations: observed versus simulated by MLR, NN and RF models for Chengdu in winter 2019

**表 2 成都市 $PM_{2.5}$ 污染潜势预报模型模拟效果评估**

Table 2 Evaluation of the $PM_{2.5}$ pollution potential forecast models in Chengdu

| 方法 | 数据集 | *R* | BIAS | GE | RMSE | 分类误判率/% |
|---|---|---|---|---|---|---|
| MLR | 训练集 | 0.88 | −3.45 | 17.47 | 22.94 | 24.51 |
| | 测试集 | 0.88 | −4.05 | 16.18 | 22.44 | 21.88 |
| | 回顾预报集 | 0.85 | 1.75 | 11.83 | 14.65 | 23.33 |
| NN | 训练集 | 0.86 | −3.66 | 18.71 | 24.68 | 27.27 |
| | 测试集 | 0.86 | −3.19 | 17.04 | 23.66 | 22.92 |
| | 回顾预报集 | 0.78 | 4.01 | 14.09 | 17.96 | 31.67 |
| RF | 训练集 | 0.98 | −0.04 | 7.89 | 10.52 | 13.44 |
| | 测试集 | 0.83 | −1.82 | 16.04 | 25.45 | 22.92 |
| | 回顾预报集 | 0.83 | 5.25 | 12.96 | 16.44 | 26.67 |

和 0.85, GE, RMSE 及分类误判率等误差值也都保持在同一水平, 但 RF 模型的 BIAS 高于 MLR 模型, 说明 RF 模型的高估程度更大。NN 模型的预报能力相较于测试集显著降低, 其预测值与观测值的相关系数降至 0.78, 虽然其 GE 和 RMSE 值与 MLR 和 RF 模型较为接近, 但 BIAS 为 4.01, 说明 NN 模型在回顾预报集中亦存在一定程度的高估, 且分类误判率比测试集中的 22.92 增加 38.2%。MLR, NN 及 RF 模型的模拟结果与观测时间序列皆较为吻合(图 4), 对 $PM_{2.5}$ 的变化趋势都能够进行较好的模拟, 且都能够识别主要的高浓度时段(如 12 月 8—15 日的连续重污染时段)。对比 MLR, NN 及 RF 模型的预报性能, NN 模型的相关系数相对较低, 分类误判率误差较高, 在时间序列中也存在更多的不一致; MLR 及 RF 模型具有更好的模拟能力。虽然 MLR 模型预测结果与观测值的相关性最强, 各项误差皆较低, 但在各项评估指标与 MLR 模型相近的情况下, RF 模型对 $PM_{2.5}$ 的重污染时段具有更好的识别能力(如 1 月 6—9 日和 2 月 5 日)。从整体上看, RF 模型对成都市冬季 $PM_{2.5}$ 污染的预报性能最佳。

## 2.3 中长期潜势预报

### 2.3.1 臭氧中长期潜势预报

本研究选定的成都市臭氧及 $PM_{2.5}$ 污染潜势预报模型的关键入模变量主要为相关气象因子(基于 WRF 回溯模拟结果)及前一日污染浓度变量(基于观测数据)。在 2.1 及 2.2 节的提前一天(1-day lead)污染潜势预报中, 我们利用 WRF 当日气象回溯模拟结果及前一日污染浓度观测结果对当日臭氧和 $PM_{2.5}$ 污染潜势进行预测, 而通过迭代预报结果生成前一日污染物浓度变量(即利用当天的浓度预报值作为下一天预报中的前一日污染物浓度值), 则可对未来 2~15 天(2–15-day lead)的污染潜势进行提前更长时间的预报(详见附录 9)。利用建立的 MLR, NN 及 RF 模型, 对成都市 2019 年夏季(4—8 月)臭氧及冬季(1—2 月及 11—12 月) $PM_{2.5}$ 的污染潜势进行提前 1~15 天的预报, 评估 MLR, NN 及 RF 模型对中长期污染潜势预报的性能。

在 MLR 模型的中长期臭氧潜势预报结果(图 5 和 6)中, 不同提前天数的预报浓度数值非常接近, 除提前 1~3 天(1–3-day lead)的预报结果外, 其余提前各天(4–15-day lead)的预报浓度时间序列几乎完全重叠, 且都能与实测浓度数据的变化趋势较好地吻合(图 5)。当从提前 1 天增加至提前 3 天预报时, MLR 模型预报结果与实测值的相关性有所下降(由 0.77 降至 0.73), 各项误差指标有所增加, 但不显著(GE, RMSE 和分类污染率分别增加 5.6%, 6.0% 和 6.7%); 当提前预报时间延长至 7~15 天(7–15-day lead)时, 各项误差指标保持稳定, 不再发生明显的变化, 始终保持较高的预报性能(图 6)。在 NN 及 RF 模型中也观察到短期预报(提前 1~3 天)误差微弱增加、中长期(提前 7~15 天)预报趋于稳定的特征(图 6), 可见 3 个模型在中长期臭氧潜势预报中都有较好的预报性能, 其中 MLR 模型能够更准确地识别臭氧重污染时段(图 5), 在定量上与观测结果更接近, 中长期污染潜势预报性能最佳。

考虑到在提前 1~15 天的臭氧污染潜势预报测试中, 各模型关键预报因子中的相关气象因子均无变化(基于当日 WRF 回溯模拟气象场), 仅前一日臭氧浓度预报因子(O3_YEST)由预报模拟值迭代重新生成, 在不考虑 WRF 模拟气象场的不确定条件下, 臭氧中长期污染潜势预报的准确性差异主要受 O3_YEST 变量的影响。由前面的分析可知, O3_YEST 变量对中长期污染潜势预报模拟的影响极为有限, 表现在提前 1~3 天的预报中 O3_YEST 的改变对预报性能影响较小, 而当预报时间超过 3 天时, O3_YEST 变量的影响几乎消失。为进一步验证 O3_YEST 变量对成都市夏季臭氧污染潜势预报性能的影响, 去除 O3_YEST 变量后重新构建 MLR, NN 及 RF 潜势预报模型(预报评估结果见 附录 10 和 11)。对比包含 O3_YEST 变量的模型预报效果(2.1 节), 不包含 O3_YEST 变量的 MLR, NN 及 RF 模型在回顾预报集中的相关性分别略为下降至 0.72, 0.71 和

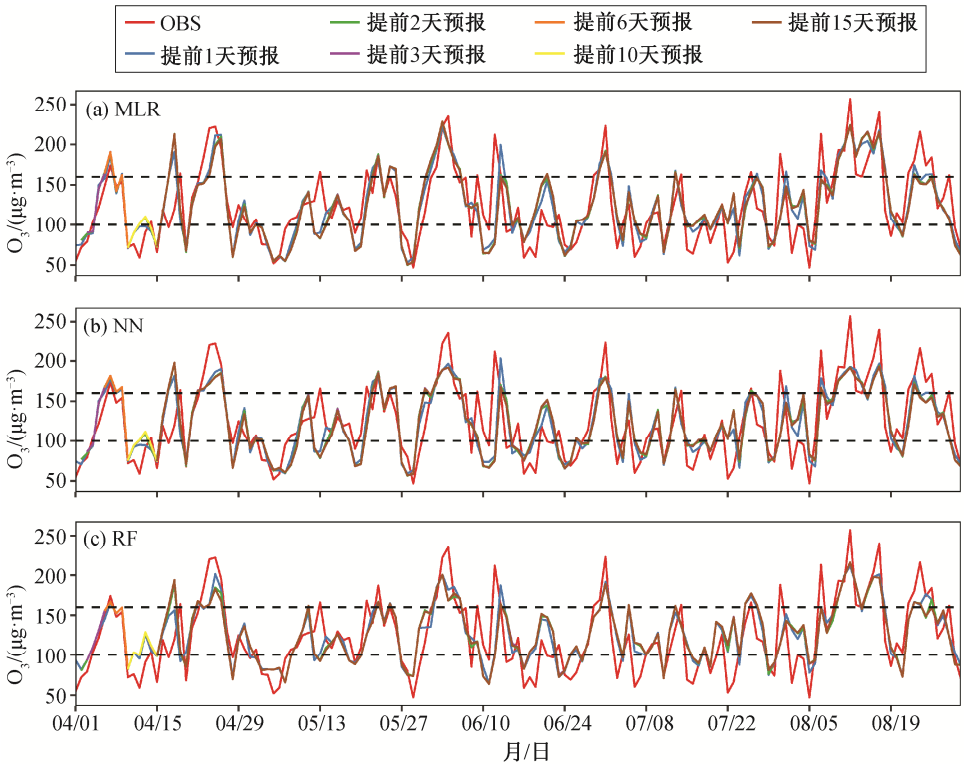



**图 5　成都市 2019 年夏季臭氧污染潜势提前 1~15 天预报值和观测值时间序列**

Fig. 5　Timeseries of 1–15-day lead $O_3$ pollution potential forecasts versus observations for Chengdu in summer 2019

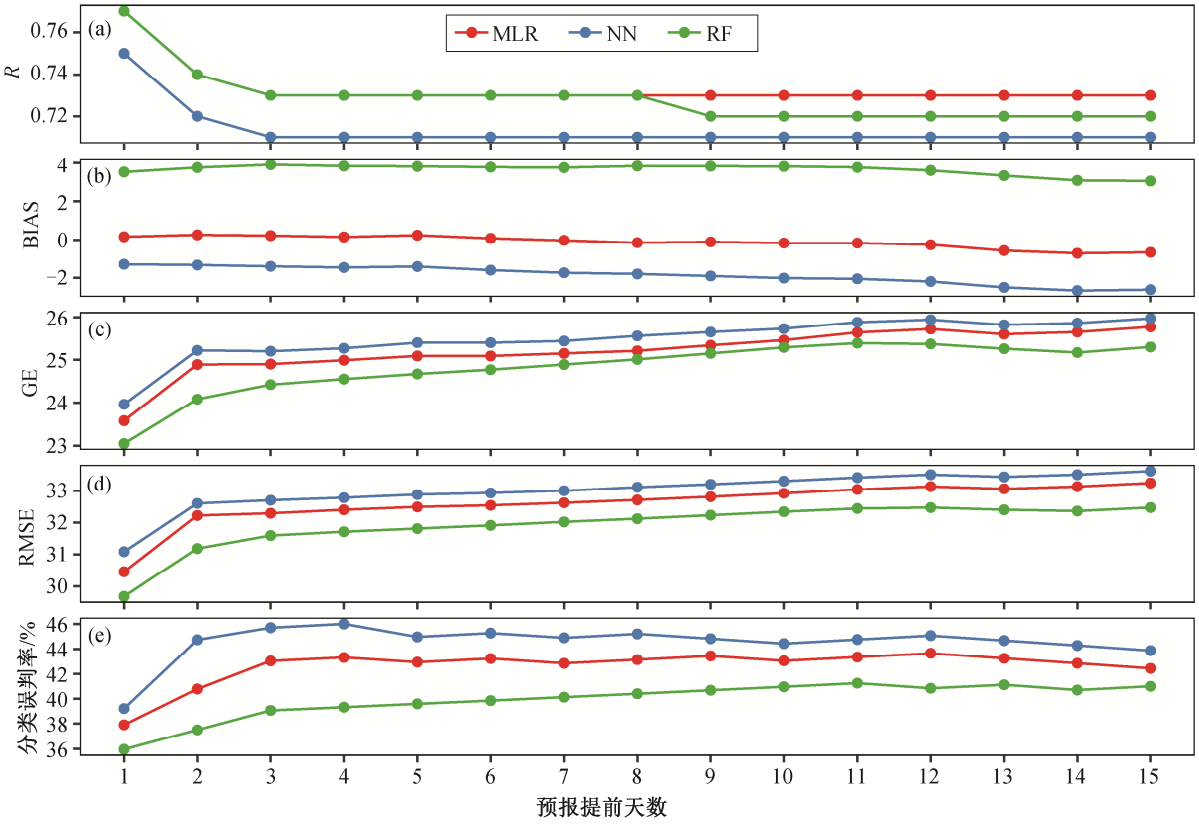



**图 6　成都市 2019 年 $O_3$ 污染潜势提前 1~15 天预报性能评估结果**

Fig. 6　Evaluation of 1–15-day lead forecasts of $O_3$ pollution potential for Chengdu in summer 2019

0.74(见附录 11, 与表 1 对比), GE, RMSE 及分类误判率都小幅度增加(GE, RMSE 及分类误判率增幅分别为 7.9%~10.6%, 6.4%~8.9%和 1.7%~10.3%)。3 个模型预报结果的时间序列变化趋势也都依旧保持与观测值良好的一致性(图见附录 10)。可见在模型建立的过程中, 虽然 O3_YEST 变量对臭氧潜势模型预报性能的提升起到一定的作用, 但效果有限, 成都市 $O_3$ 污染潜势预报模型的预报效果主要受各相关气象因子的影响。

**2.3.2 $PM_{2.5}$ 中长期潜势预报**

同样地, 利用建立的 MLR, NN 及 RF 模型对成都市 2019 年 1—2 月和 11—12 月的 $PM_{2.5}$ 污染进行提前 1~15 天的预报模拟, 结果如图 7 和 8 所示。MLR 模型在提前 1~3 天(1–3-day lead)的预报测试中相关性降低 17.6%, BIAS 由 1.75 增至 5.2, GE, RMSE 及分类误判率分别增加 41.9%, 41.3%和 58.9%(图 8), 模型误差显著增加, MLR 模型的预报效果明显下降。当延长至提前 7~15 天(7–15-day lead)的预报时, 各误差指标(GE, RMSE 及分类误判率)依旧存在一定程度的增长趋势, 且 BIAS 持续增加说明高估问题更加显著, 但各误差指标仍处于可接受范围内。从图 7 预报值的时间序列中亦可见, 1~3 天预报结果之间的差异较为显著, 随着预报提前时间的延长, 对 $PM_{2.5}$ 的高估愈加明显(如 2019 年 2 月 15 及 11 月 25 日前后)。当延长至提前 7~15 天时, 预报值时间序列出现很大程度的重叠, 但依旧与观测值的时间变化趋势大体上保持一致。同样的结果在 NN 及 RF 模型的中长期潜势预报测试中亦可见, 尤其是 NN 模型, 其性能变差更为显著。对比 3 个模型对提前 1~15 天预报的性能评估结果, 可见 RF 模型的预报效果更为稳定, 与观测结果的时间序列保持更好的一致性, 具有最好的预报性能。针对成都市 $PM_{2.5}$ 污染建立的 MLR, NN 及 RF 模型对中长期 $PM_{2.5}$ 污染潜势预报的性能均随提前预报时长的增加而明显地下降, 其中 NN 模型的预报性能下降最严重, MLR 和 RF 模型预报性能的下降幅度较小。综合来看, 3 个模型的预报性能都仍处于可接受的范围[61]。

在 $PM_{2.5}$ 模型中, 相关气象预报因子数据不变的情况下, 成都市 $PM_{2.5}$ 中长期污染潜势预报模拟效果的显著降低主要受前一日 $PM_{2.5}$ 浓度(PM2.5_YEST)变量的影响。尤其在提前 1~3 天的短期预报中, PM2.5_YEST 的影响极为显著, 而当延长至提前 7~15 天时, 其预报性能趋于相对稳定, PM2.5_YEST 的影响显著变小。去除 PM2.5_YEST 后重新建立成都市冬季 $PM_{2.5}$ 污染潜势 MLR, NN 和 RF 模型, 进行预报效果测试(评估结果见附录 11 和附录 12)。与包含 PM2.5_YEST 变量的模型预报效果(表 2)相比, 去除 PM2.5_YEST 变量后, 新建的 $PM_{2.5}$ 模型预报性能显著地下降(附录 11), 各模型的预报值在回顾预报集中与观测值的相关系数由原来的 0.78~0.85 下降至 0.38~0.47, 且各项误差指标(GE, RMSE 及分类误判率等)的增幅都达到 90%~130%, 尤其是分类误判率皆达到 50%以上, 不论是在定性还是在定量方面, 各模型模拟值的时间序列(附录 12)都与观测值存在很大的差异。PM2.5_YEST 对 $PM_{2.5}$ 污染潜势预报模型的建立具有显著影响, 该变量能够明显地提升模型的预报性能, 可见 $PM_{2.5}$ 潜势模型的预报性能随提前预报时长的增加而显著降低主要是对 PM2.5_YEST 这一变量的依赖所致。

## 3 结论

本文以成都市为例, 利用 2016—2019 年 WRF 模式回溯模拟气象场及同期 $O_3$ 及 $PM_{2.5}$ 日值观测数据, 利用影响污染物转化、扩散和传输的主要气象条件及相关因子建立潜在预报因子数据集。通过筛选影响成都市夏季(4—8 月) $O_3$ 及冬季(11 月—来年 2 月) $PM_{2.5}$ 污染的关键预报因子, 利用多元线性回归、随机森林以及 BP 神经网络等机器学习算法, 分别建立夏季 $O_3$ 及冬季 $PM_{2.5}$ 污染潜势预报模型。对比分析各模型对成都市 $O_3$ 及 $PM_{2.5}$ 浓度的预报效果, 讨论基于机器学习方法建立的污染潜势预报模型的中长期预报能力。

基于多元线性回归、BP 神经网络、随机森林等算法建立的 MLR, NN 及 RF 模型对成都市夏季臭氧浓度均具有良好的预报性能, 模型泛化能力较好, 能够准确地识别成都市典型的臭氧高污染时段。在不考虑气象模拟准确性的情况下, 建立的潜势模型亦能够较好地应用于成都市夏季臭氧中长期(提前 7~15 天)污染潜势预报。随着预报提前时间延长, 模型预报性能并未显著降低, 表现稳定, 主要原因是各模型都对前一日臭氧浓度变量的依赖性较小。其中, MLR 模型对成都市臭氧浓度具有相对最佳的预报性能, 臭氧高值更接近观测结果, 且与观测结果的时间变化趋势更加吻合。

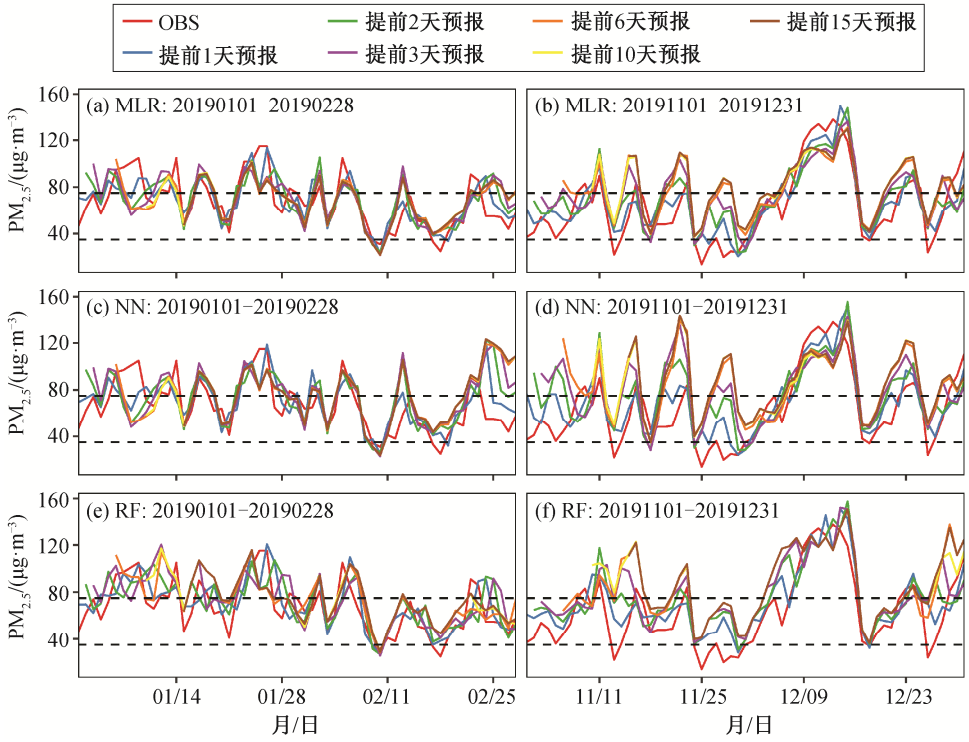



**图 7　成都市 2019 年冬季 $PM_{2.5}$ 污染潜势提前 1~15 天预报值和观测值时间序列**

Fig. 7　Timeseries of 1–15-day lead forecasts of $PM_{2.5}$ pollution potential versus observations for Chengdu in winter 2019

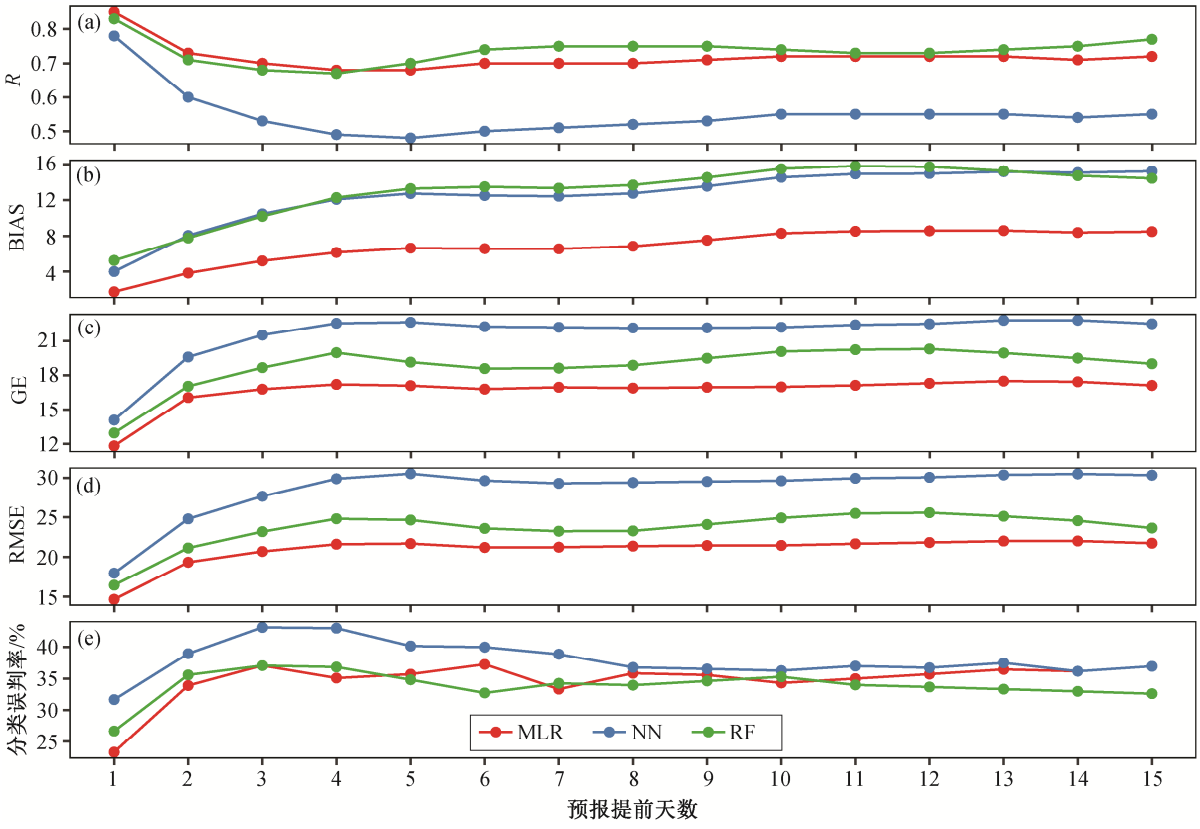



**图 8　成都市 2019 年冬季 $PM_{2.5}$ 提前 1~15 天污染潜势预报效果评估**

Fig. 8　Evaluation of 1–15-day lead forecasts of $PM_{2.5}$ pollution potential for Chengdu in winter 2019

基于关键气象因子和前一日 $PM_{2.5}$ 浓度变量建立的 MLR, NN 及 RF 模型能够较好的预测成都市冬季 $PM_{2.5}$ 浓度的变化趋势, 与观测时间序列保持较好的一致性, 各项误差指标较低, 3 个模型均具有较优的预报性能。通过迭代生成前一日 $PM_{2.5}$ 浓度变量, 可利用建立的 MLR, NN 及 RF 模型, 对 $PM_{2.5}$ 污染的中长期潜势进行预报。受前一日 $PM_{2.5}$ 浓度变量的影响, 随着提前时长的增加, 各模型的预报性能均有所降低, 但仍处于可接受范围。其中, RF 模型在保持良好误差指标的同时, 在定量上对 $PM_{2.5}$ 的高浓度数值有更好的表现, 具有相对最优的预报能力。

## 参考文献


[1] 白志鹏, 蔡斌彬, 董海燕, 等. 灰霾的健康效应. 环境污染与防治, 2006, 28(3): 198–201

[2] 廖志恒, 范绍佳. 2006—2012 年珠江三角洲地区 $O_3$ 污染对人群健康的影响. 中国环境科学, 2015, 35(3): 897–905

[3] 姜磊, 周海峰, 赖志柱, 等. 中国城市 $PM_{2.5}$ 时空动态变化特征分析: 2015—2017 年. 环境科学学报, 2018, 38(10): 3816–3825

[4] 杨兴川, 赵文吉, 熊秋林, 等. 2016 年京津冀地区 $PM_{2.5}$ 时空分布特征及其与气象因素的关系.生态环境学报, 2017, 26(10): 1747–1754

[5] 孟晓艳, 宫正宇, 张霞, 等. 全国及重点区域臭氧污染现状. 中国环境监测, 2017, 44(4): 17–25

[6] 张倩倩, 张兴赢. 基于卫星和地面观测的 2013 年以来我国臭氧时空分布及变化特征. 环境科学, 2019, 40(3): 1132–1142

[7] Wang Yuesi, Yao Li, Wang Lili, et al. Mechanism for the formation of the January 2013 heavy haze pollution episode over central and eastern China. Science China Earth Sciences, 2014, 57(1): 14–25

[8] Elminir H K. Dependence of urban air pollutants on meteorology. Science of the Total Environment, 2005, 350(1/2/3): 225–237

[9] Liao Hong, Chang Wenyuan, Yang Yang. Climatic effects of air pollutants over China: a review. Advances in Atmospheric Sciences, 2015, 32(1): 115–139

[10] 王冠岚, 薛建军, 张建忠. 2014 年京津冀空气污染时空分布特征及主要成因分析. 气象与环境科学, 2016, 39(1): 34–32

[11] Leung D M, Tai A P K, Mickley L J, et al. Synoptic meteorological modes of variability for fine particulate matter ($PM_{2.5}$) air quality in major metropolitan regions of China. Atmospheric Chemistry and Physics, 2018, 18(9): 6733–6748

[12] 张小玲, 熊亚军, 徐敬, 等. 气象条件对京津冀区域细粒子浓度增加与改善的影响分析//中国颗粒学会气溶胶专业委员会. 第八届全国大气细及超细粒子技术研讨会暨 $PM_{2.5}$ 源谱交流会论文集. 黄石, 2015: 5

[13] Mao Lu, Liu Run, Liao Wenhui, et al. An observation-based perspective of winter haze days in four major polluted regions of China. National Science Review, 2019, 6(3): 515–523

[14] Zheng Guangjie, Duan Fengkui, Su Hang, et al. Exploring the severe winter haze in Beijing: the impact of synoptic weather, regional transport and heterogeneous reactions. Atmospheric Chemistry and Physics, 2015, 15(6): 2969–2983

[15] 徐晓峰, 李青春, 张小玲. 北京一次局地重污染过程气象条件分析. 气象科技, 2005, 33(6): 543–547

[16] 梅鹏蔚. 稳定气象条件对天津市环境空气质量的影响. 城市环境与城市生态, 2006, 19(4): 37–39

[17] 李展, 陈建文, 杜云松, 等. 成都及周边城市一次区域性空气污染过程特征分析. 环境科学与技术, 2015, 38(3): 125–130

[18] 瞿华, 朱彬, 赵雪婷, 等. 长江三角洲初冬一次重污染天气成因分析. 中国环境科学, 2018, 38(11): 4001–4009

[19] 张美根, 韩志伟, 雷孝恩. 城市空气污染预报方法简述. 气候与环境研究, 2001, 6(1): 113–118

[20] 王迎春, 孟燕军, 赵习方. 北京市空气污染业务预报方法. 气象科技, 2001, 4(11): 42–46

[21] Niemeyer L E. Forecasting air pollution potential. Monthly Weather Review, 1960, 88(3): 88–96

[22] 黄晓娴, 王体健, 江飞. 空气污染潜势–统计结合预报模型的建立及应用. 中国环境科学, 2012, 32(8): 1400–1408

[23] Yu Mingyuan, Cai Xuhui, Xu Chunmeng, et al. A climatological study of air pollution potential in China. Theoretical and Applied Climatology, 2019, 136: 627–638

[24] Tai A P K, Mickley L J, Jacob D J. Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: implications for the sensitivity of $PM_{2.5}$ to climate change. Atmospheric Environment, 2010, 44(32): 3976–3984

[25] Otero N, Sillmann J, Mar K A, et al. A multi-model

comparison of meteorological drivers of surface ozone over Europe. Atmospheric Chemistry and Physics, 2018, 18(16): 12269–12288

[26] Li Ke, Jacob D J, Liao Hong, et al. Anthropogenic drivers of 2013–2017 trends in summer surface ozone in China. Proceedings of the National Academy of Sciences, 2019, 116(2): 422–427

[27] Zhai Shixian, Jacob D J, Wang Xuan, et al. Fine particulate matter ($PM_{2.5}$) trends in China, 2013–2018: separating contributions from anthropogenic emissions and meteorology. Atmospheric Chemistry and Physics, 2019, 19(16): 11031–11041

[28] Chen Xi, Zhong Buqing, Huang Fuxiang, et al. The role of natural factors in constraining long-term tropospheric ozone trends over Southern China. Atmospheric Environment, 2020, 220: 117060

[29] 张小曳, 徐祥德, 丁一汇, 等. 2013—2017年气象条件变化对中国重点地区 $PM_{2.5}$ 质量浓度下降的影响. 中国科学: 地球科学, 2020, 50(4): 483–500

[30] 丁愫, 陈报章, 王瑾, 等. 基于决策树的统计预报模型在臭氧浓度时空分布预测中的应用研究. 环境科学学报, 2018, 38(8): 3229–3242

[31] Cabaneros S M S, Calautit J K, Hughes B R. A review of artificial neural network models for ambient air pollution prediction. Environmental Modelling and Software, 2019, 119: 285–304

[32] Byun D, Schere K L. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale AIR Quality (CMAQ) modeling system. Applied Mechanics Reviews, 2006, 59(2): 51–77

[33] 赵惠芳, 陈雅莲, 唐会荣, 等. 晋江城市空气质量污染潜势统计预报方法初探. 气象与环境学报, 2009, 25(5): 27–30

[34] Barrero M A, Grimalt J O, Canton L. Prediction of daily ozone concentration maxima in the urban atmosphere. Chemometrics and Intelligent Laboratory Systems, 2006, 80(1): 67–76

[35] Stadlober E, Hormann S, Pfeiler B. Quality and performance of a $PM_{10}$ daily forecasting model. Atmospheric Environment, 2008, 42(6): 1098–1109

[36] 苏筱倩, 安俊琳, 张玉欣, 等. 支持向量机回归在臭氧预报中的应用. 环境科学, 2019, 40(4): 1697–1704

[37] Lu Wei-Zhen, Wang Wen-Jian. Potential assessment of the “support vector machine” method in forecasting ambient air pollutant trends. Chemosphere, 2005, 59(5): 693–701

[38] Deters J K, Zalakeviciute R, Gonzalez M, et al. Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering, 2017, 2017: 5106045

[39] 常恬君, 过仲阳, 徐丽丽. 基于 Prophet–随机森林优化模型的空气质量指数规模预测. 环境污染与防治, 2019, 41(7): 758–766

[40] Singh K P, Gupta S, Rai P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment, 2013, 80: 426–437

[41] Feng Xiao, Li Qi, Zhu Yajie, et al. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment, 2015, 107: 118–128

[42] Li Xiang, Peng Ling, Yao Xiaojing, et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution, 2017, 231: 997–1004

[43] 刘银超. 基于BP神经网络和随机森林的空气污染物浓度预测研究[D]. 秦皇岛: 燕山大学, 2017

[44] Lightstone S D, Moshary F, Gross B. Comparing CMAQ forecasts with a neural network forecast model for $PM_{2.5}$ in New York. Atmosphere, 2017, 8(9): 161

[45] Skamarock W C, Klemp J B, Dudhia J, et al. A description of the advanced research WRF Version 3. NCAR Technology Note: NCAR/TN-475+STR, 2008

[46] Chen Shu-Hua, Sun Wen-Yih. A one-dimensional time dependent cloud model. Journal of the Meteorological Society of Japan, 2002, 80(1): 99–118

[47] Kain J S. The Kain-Fritsch convective parameterization: an update. Journal of Applied Meteorology, 2004, 43(1): 170–181

[48] Hong Songyou, Noh Y, Dudhia J. A new vertical diffusion package with an explicit treatment of entrainment processes. Monthly Weather Review, 2006, 134(9): 2318–2341

[49] Tewari M, Chen F, Wang W, et al. Implementation and verification of the unified NOAH land surface model in the WRF model // 20th conference on weather analysis and forecasting/16th conference on numerical

weather prediction. Seattle, 2004: 11–15

[50] Stauffer D R, Seaman N L. Multiscale four-dimensional data assimilation. Journal of Applied Meteorology, 1994, 33(3): 416–434

[51] Liu Y, Warner T T, Bowers J F, et al. The operational mesogamma-scale analysis and forecaste system of the US army test and evaluation command. Part 1: overview of the modeling system, the forecast products. Journal of Applies Meteorology, 2008, 47(4): 1077–1092

[52] Emery C A, Tai E, et al. Enhanced meteorological modeling and performance evaluation for Two Texas Ozone Episodes, project report prepared for the Texas Natural Resource conservation commission [R]. Texas: Texas Natural Resource Conservation Commission, 2011

[53] 雷瑜, 张小玲, 唐宜西, 等. 北京城区 $PM_{2.5}$ 及主要污染气体“周末效应”和“假日效应”研究. 环境科学学报, 2015, 35(5): 1520–1528

[54] 石玉珍, 徐永福, 王庚辰, 等. 北京市夏季 $O_3$、$NO_x$ 等污染物“周末效应”研究. 环境科学, 2009, 30(10): 2832–2838

[55] 罗进奇, 黄小娟, 张军科, 等. 成都市成都大气污染特征及其假日效应研究. 四川环境, 2018, 37(6): 21–37

[56] US Environmental Protection Agency. Guideline for developing an air quality (ozone and $PM_{2.5}$) forecasting program. Research Triangle Park: EPA-456/R-03-002, 2003

[57] Miller A J. Subset selection in regression. 2nd ed. London: Chapman and Hall, 2002

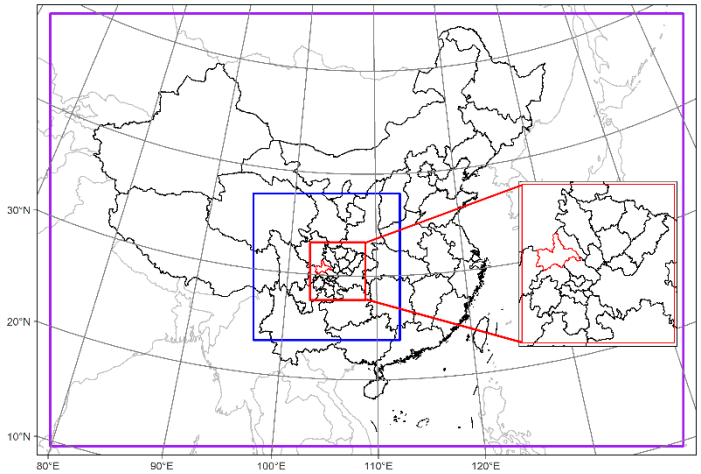
[58] Breiman L. Random Forests. Machine Learning, 2001, 45(1): 5–32

[59] 鲁茂, 贺昌政. 对多重共线性问题的探讨. 统计与决策, 2007(8): 6–9

[60] 张凤莲. 多元线性回归中多重共线性问题的解决办法探讨[D]. 华南理工大学, 2010

[61] Emery C, Liu Zhen, Russell A G, et al. Recommendations on statistics and benchmarks to assess photochemical model performance. Journal of the Air & Waste Management Association, 2017, 67(5): 582–598

## 附录




**附录 1　WRF 模拟 36-公里(紫色矩形框区域)、12-公里(蓝色矩形框区域)和 4-公里(红色矩形框区域)的 3 重嵌套网格范围**

Appendix 1　Three nested domains of the WRF simulation. The purple-, blue- and red-line framed regions indicate the 36-, 12- and 4-km grids, respectively

**附录 2　2016—2019 年 WRF 回溯模拟效果评估**

Appendix 2　Performance Evaluation Statistics of the WRF Simulated Meteorological Fields Against Hourly Observations from Surface Meteorological Station in China during 2016–2019

| 年份 | Hmd Bias | Hmd GE | TMP Bias | TMP RMSE | WD Bias | WD GE | WS Bias | WS RMSE |
|---|---|---|---|---|---|---|---|---|
| 2016 | −0.07 | 1.14 | −1.48 | 3.29 | 4.64 | 63.63 | 0.87 | 1.73 |
| 2017 | −0.1 | 1.14 | −1.32 | 3.17 | 4.44 | 63.39 | 0.85 | 1.69 |
| 2018 | −0.04 | 1.14 | −1.18 | 3.07 | 4.52 | 61.89 | 0.88 | 1.72 |
| 2019 | −0.04 | 1.15 | −1.18 | 3.14 | 4.69 | 62.54 | 0.86 | 1.72 |
| 平均 | −0.06 | 1.14 | −1.29 | 3.17 | 4.57 | 62.86 | 0.87 | 1.72 |

说明: 表中平均偏差(Bias)、平均绝对误差(GE)、均方根误差(RMSE)等统计指标计算公式为: $\mathrm{Bias}=\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)$, $\mathrm{GE}=\frac{1}{N}\sum_{i=1}^{N}|P_i-O_i|$, $\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)^2}$, 其中: N—样本数; P—模拟值; O—观测值。

**附录 3**

本研究共拟定 3 大类共 39 个潜在的预报因子, 详细介绍如下:

1) 污染持续性因子: 考虑污染物滞留累积等过程的持续影响, 将前一日 $O_3$ 浓度(O3_YEST)及前一日 $PM_{2.5}$ 浓度(PM2.5_YEST)列为潜在预报因子。尤其是在 $PM_{2.5}$ 的短期预报中, 前一日污染物浓度能够显著提高 $PM_{2.5}$ 浓度预测的准确性。

2) 节假日、工作日信息因子: 受人类活动变化规律所影响, 污染源排放存在显著的时序变化, 如受“周末效应”、“假日效应”等影响的机动车排放以及工业源排放。受排放时间变化的影响, 城市污染浓度变化亦存在明显的“周末效应”及“假日效应”[53–55]。因此, 将节假日信息(IF_HOLIDAY)及工作日信息(IF_WEEK)列为潜在预报因子。

3) 气象条件相关因子: 气象条件相关因子主要考虑影响污染物转化、稀释、扩散和沉降

等过程的气象因子, 以及部分气象变量重新组合后的二次因子。引入的主要潜在气象预报因子如下:

① 风场: 风场主要影响着污染物的水平扩散、稀释及输送过程。地面风速越大, 一方面影响着污染前体物, 使得其无法充分混合, 另一方面影响着 $O_3$ 及 $PM_{2.5}$ 污染物的疏散过程; 此外, 考虑 $O_3$ 及 $PM_{2.5}$ 污染及其前体物的日变化特征, 不同时段的风速对空气质量的影响亦各不相同, 如上午时段的风速影响 $O_3$ 前体物 NOx 和 VOC 的混合, 而下午风速大小则影响着 $O_3$ 的扩散稀释。首先引入每日地面平均风速(WS)、上午时段地面平均风速(WS_MORN)、午后时段地面平均风速(WS_AFTE)等风场相关因子。

风向则主要影响着污染物水平迁移、扩散传输的方向, 下风向区域的空气质量受上风向污染物传输的影响较为显著, 且高空风场影响着污染物的区域传输。故引入地面每日最多风向(WD)以及不同高度层(850hPa、700hPa、500hPa)的风向、风速相关变量(WS850、WS700、WS500 和 WD700)。

此外, 引入风向日变化因子[16]用以表征风向的集中程度, 计算公式为

$$\bar{v} = \frac{1}{24}\sum_{i=1}^{24} v_i \tag{1}$$

$$\bar{u} = \frac{1}{24}\sum_{i=1}^{24} u_i \tag{2}$$

$$\mathrm{WD\_}CHANGE = \sqrt{\frac{1}{24}\sum_{i=1}^{24}(u_i-\bar{u})^2 + \frac{1}{24}\sum_{i=1}^{24}(v_i-\bar{v})^2} \tag{3}$$

其中: $v_i$, $u_i$ 分别为每日第 $i$ 时刻的纬向风、经向风。

② 气温: 气温的变化直接或间接的影响着污染物的排放和化学转化反应过程。如高温常会加快地表蒸发、植物蒸腾作用等, 低温则间接的影响着污染的排放(如北方冬季采暖等), 且气温的上升会使得臭氧反应速率加快, 每日最高气温与臭氧浓度和臭氧生成高度相关。此外, 不同高度层的气温变化影响着污染物的垂直混合等过程。基于以上考虑, 引入每日地面最高气温(T_MAX)、每日地面最低气温(T_MIN)、每日地面平均气温(T)及不同时段和不同高度层的气温预报因子(T_MORN、T850、T850_MAX、T700、T700_MAX、T500、T500_MAX)。

③ 云量: 云量的变化对辐射有着重要影响。如低云的增加将会显著减少短波辐射, 而高云则会影响长波辐射, 从而影响 O3 及 PM2.5 生成过程中的光化学反应。故此针对云量这一变量引入高云量(HCC)、中云量(MCC)及低云量(LCC)作为潜在预报因子。

④ 降水: 降水对大气中的污染物起到重要的湿清除作用, 故引入每日降水总量(PR)。

⑤ 相对湿度: 高湿条件下, 气溶胶颗粒易吸湿增长, 对硫酸盐及硝酸盐的非均相反应过程有着重要影响, 因此将每日平均相对湿度(RH)以及上午时段相对湿度(RH_MORN)纳入潜在预报因子。

⑥ 短波辐射: 短波辐射是光化学反应的关键因素, 对 $PM_{2.5}$ 及 $O_3$ 的生成都有重要影响, 因此引入日短波辐射总量(SR)及最大短波辐射量(SR_MAX)。

⑦ 500hPa 位势高度: 高空大尺度环流系统的移动影响着各种气象条件的变化, 采用 500hPa 位势高度(GHT500)作为天气形势的表征变量。

⑧ 边界层高度: 作为影响污染物扩散稀释的重要因素之一, 引入每日平均边界层高度(PBL)及每日最大边界层高度(PBL_MAX)作为潜在预报变量。

⑨ 气压: 引入地面平均气压(PS), 同时将当日 24 小时变压(PS_DELTA)及前一日 24 小时变压(PS_DELTA_YEST)列为潜在变量, 用以表征天气形势的发展变化情况等。

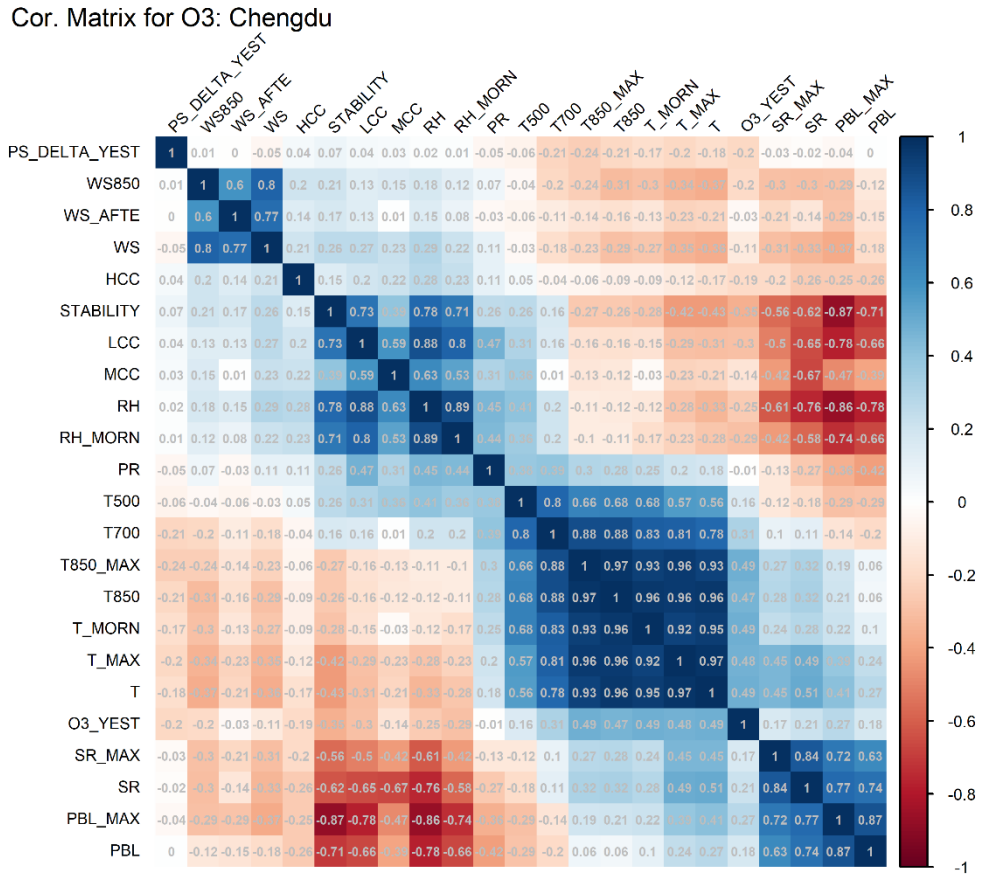
⑩ 稳定度: 大气层结稳定度对污染物的扩散能力和扩散方式起到重要作用。大气层结稳定时, 大气湍流受到抑制, 不利于污染物扩散; 当层结不稳定时, 大气湍流相对活跃, 污染物的扩散也相对增强。利用 700hPa 高度 17: 00 的气温与地面日最高气温的差值来表征大气层结的稳定度(STABILITY)。

**$O_3$ 和 $PM_{2.5}$ 污染潜势预报的潜在预报因子**

Potential Prediction Factors for Forecasting the $O_3$ and $PM_{2.5}$ Pollution Potentials

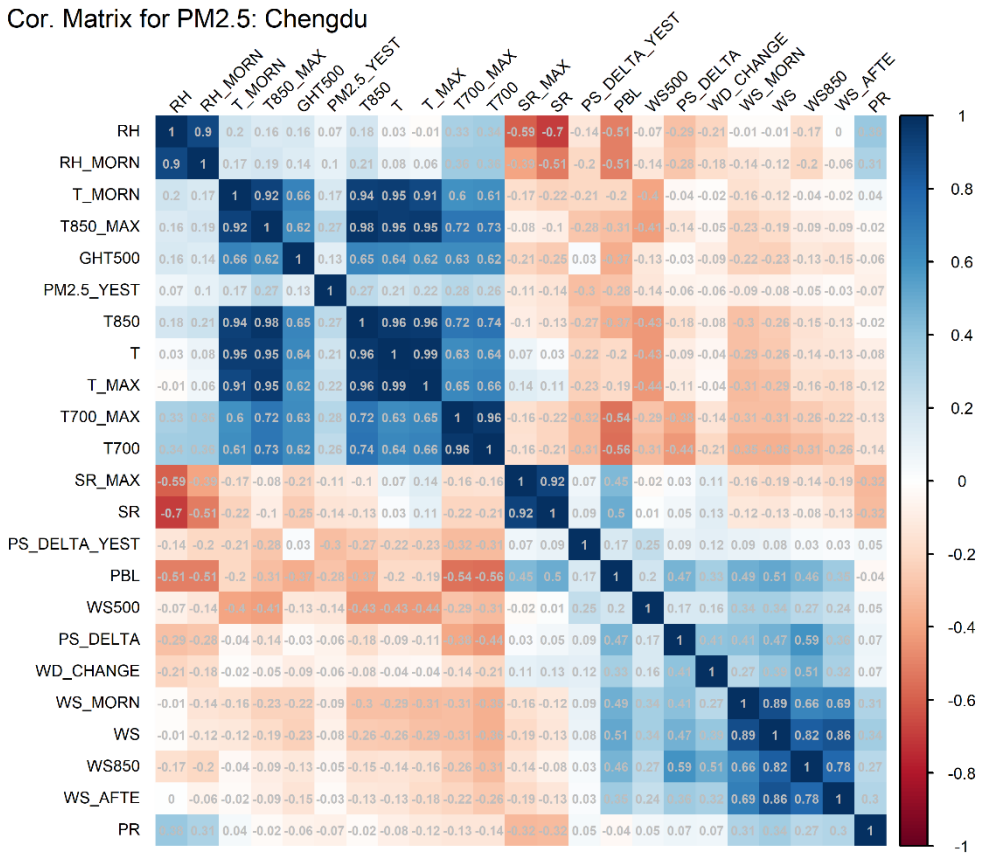
| 代码 | 变量名 | 单位 | 计算标准 |
|---|---|---|---|
| T_MAX | 地面每日最高气温 | K | 每日 00:00-23:00 2m 气温小时值-最大值 |
| T_MIN | 地面每日最低气温 | K | 每日 00:00-23:00 2m 气温小时值-最小值 |
| T_MORN | 地面上午时段平均气温 | K | 每日 08:00-11:00 2m 气温小时值-平均值 |
| T | 地面每日平均气温 | K | 每日 00:00-23:00 2m 气温小时值-平均值 |
| T850 | 850hPa 每日平均气温 | K | 每日 11:00-15:00 850hPa 气温小时值-平均值 |
| T850_MAX | 850hPa 每日最高气温 | K | 每日 00:00-23:00 850hPa 气温小时值-最大值 |
| T700 | 700hPa 每日平均气温 | K | 每日 11:00-15:00 700hPa 气温小时值-平均值 |
| T700_MAX | 700hPa 每日最高气温 | K | 每日 00:00-23:00 700hPa 气温小时值-最大值 |
| T500 | 500hPa 每日平均气温 | K | 每日 11:00-15:00 500hPa 气温小时值-平均值 |
| T500_MAX | 500hPa 每日最高气温 | K | 每日 00:00-23:00 500hPa 气温小时值-最大值 |
| WS | 地面每日平均风速 | m/s | 每日 00:00-23:00 10m 风速小时值-平均值 |
| WS_MORN | 地面上午时段平均风速 | m/s | 每日 07:00-10:00 10m 风速小时值-平均值 |
| WS_AFTE | 地面下午时段平均风速 | m/s | 每日 12:00-18:00 10m 风速小时值-平均值 |
| WS850 | 850hPa 每日平均风速 | m/s | 每日 00:00-23:00 850hPa 风速小时值-平均值 |
| WS700 | 700hPa 每日平均风速 | m/s | 每日 00:00-23:00 700hPa 风速小时值-平均值 |
| WS500 | 500hPa 每日平均风速 | m/s | 每日 00:00-23:00 500hPa 风速小时值-平均值 |
| WD | 地面每日最多风向 | | 每日 00:00-23:00 10m 风向方位小时值-众数 |
| WD700 | 700hPa 每日最多风向 | | 每日 00:00-23:00 700hPa 风向方位小时值-众数 |
| LCC | 每日平均低云量 | % | 每日 11:00-15:00 低云量小时值-平均值 |
| MCC | 每日平均中云量 | % | 每日 11:00-15:00 中云量小时值-平均值 |
| HCC | 每日平均高云量 | % | 每日 11:00-15:00 高云量小时值-平均值 |
| RH | 地面每日平均相对湿度 | % | 每日 11:00-15:00 2m 相对湿度小时值-平均值 |
| RH_MORN | 地面上午时段平均相对湿度 | % | 每日 08:00-11:00 2m 相对湿度小时值-平均值 |
| SR | 地面每日向下短波辐射总量 | MJ $m^{-2}$ | 每日 00:00-23:00 向下短波辐射小时值-总和 |
| SR_MAX | 地面每日向下短波辐射最大值 | MJ $m^{-2}$ | 每日 00:00-23:00 向下短波辐射小时值-最大值 |
| PS | 地面每日平均气压 | Pa | 每日 00:00-23:00 地面气压小时值-平均值 |
| PS_DELTA | 地面每日平均气压差值 | Pa | 当日地面平均气压与前一日地面气压差值 |
| STABILITY | 稳定度 | K | 当日 17:00 700hPa 每日平均气温与地面当日最高气温差值 |
| PBL | 每日平均边界层高度 | M | 每日 00:00-23:00 边界层高度小时值-平均值 |

| | | | |
|---|---|---|---|
| PBL_MAX | 每日边界层高度最大值 | M | 每日 00:00-23:00 边界层高度小时值-最大值 |
| PR | 每日降水总量 | Mm | 每日 00:00-23:00 降水量小时值-总和 |
| GHT500 | 500hPa 每日平均位势高度 | Gpm | 每日 00:00-23:00 500hP 位势高度小时值-平均值 |
| IF_WEEK | 是否为工作日 | | 1-工作日; 0-非工作日。 |
| IF_HOLIDAY | 是否为节日 | | 1-节日; 0-非节日。 |
| SLP | 每日平均海平面气压 | Pa | 每日 00:00-23:00 海平面气压小时值-平均值 |
| WD_CHANGE | 风向日变化因子 | | 每日 00:00-23:00 风向变化 |
| PS_DELTA_YEST | 前一日 24 小时变压 | Pa | 前一日地面平均气压与前二日地面气压差值 |
| PM2.5_YEST | 前一日 $PM_{2.5}$ 平均浓度 | $\mu g/m^3$ | - |
| O3_YEST | 前一日 $O_3$ 平均浓度 | $\mu g/m^3$ | - |




**附录 4　基于随机森林重要性分析初步选定的臭氧入模变量的相关系数矩阵**

Appendix 4　Correlation matrix of the preliminary variables selected for Ozone model based on importance analysis of Random Forest



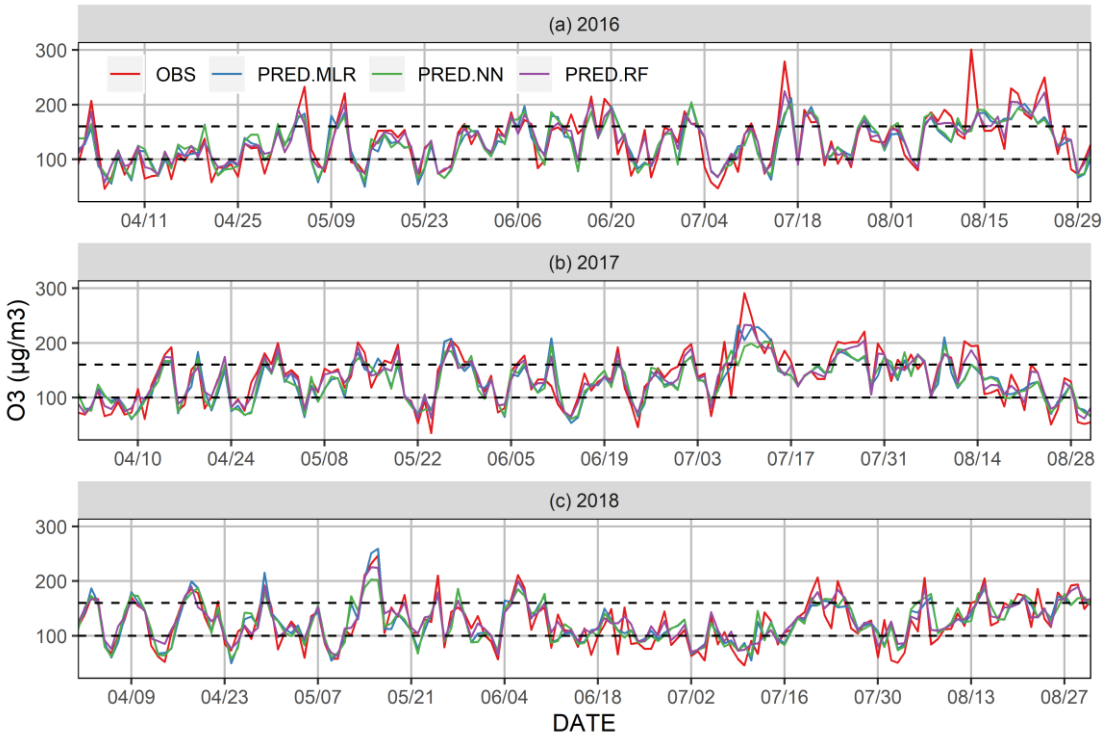

**附录 5　基于随机森林重要性分析初步选定的 $PM_{2.5}$ 入模变量的相关系数矩阵**

Appendix 5　Correlation matrix of the preliminary variables selected for $PM_{2.5}$ model based on importance analysis of Random Forest

**附录 6　成都市臭氧及 $PM_{2.5}$ 污染潜势预报模型参数设置**

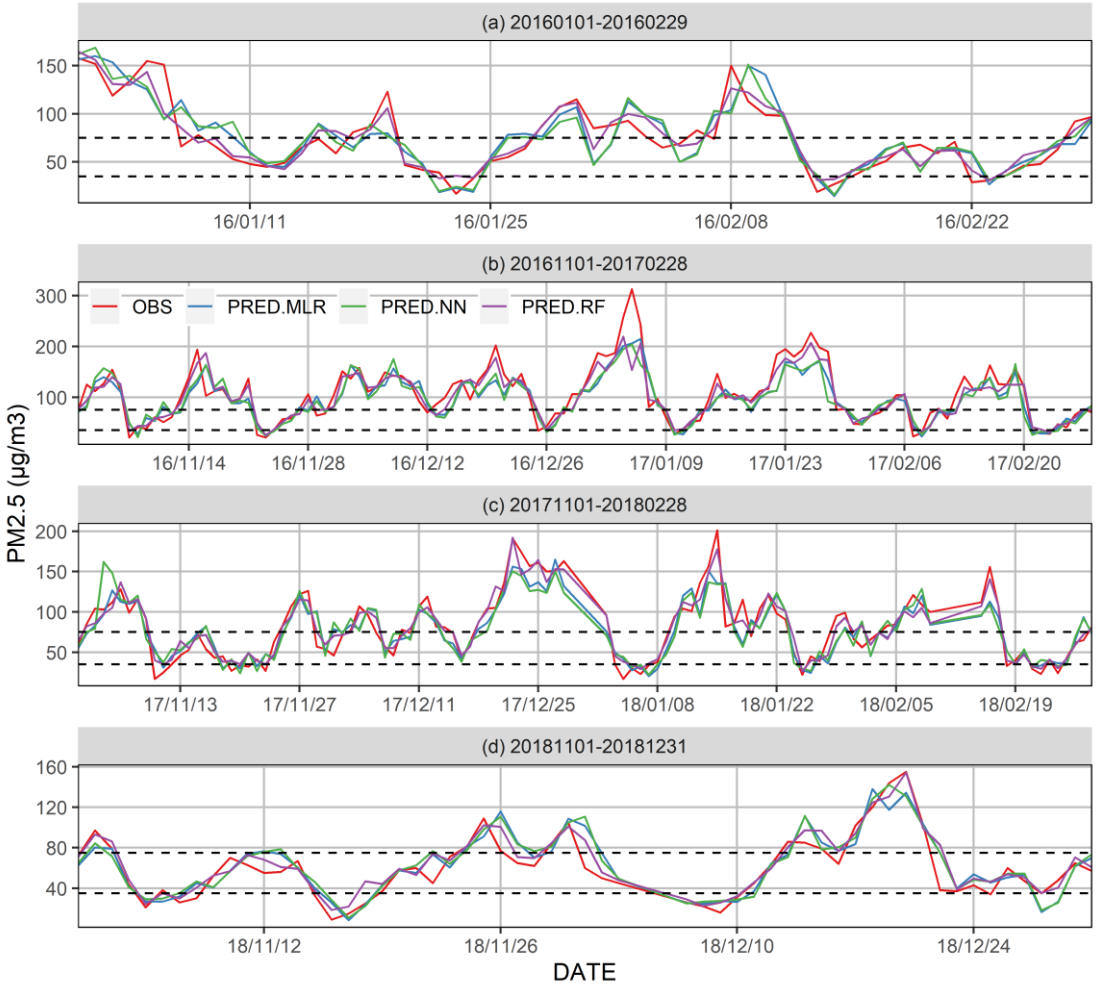
Appendix 6　The pollution potential forecast models for Ozone and $PM_{2.5}$ in Chengdu with list of selected predictor variables and setting of parameterization

| 污染物 | 模型 | 因变量 | 自变量 | 参数设置 |
|---|---|---|---|---|
| 臭氧 | MLR | $O_3$ 浓度(自然对数值) | T_MAX, PBL_MAX, HCC, MCC, WS850, PS_DELTA_YEST, WD.SSW, WD.SW, WD700.NNE, O3_YEST(自然对数值) | - |
| | NN | $O_3$ 浓度(自然对数值) | T_MAX, PBL_MAX, HCC, MCC, WS850, PS_DELTA_YEST, WD.SSW, WD.SW, WD700.NNE, O3_YEST(自然对数值) | 隐含层节点个数: 5 |
| | RF | $O_3$ 浓度 | T_MAX, PBL_MAX, O3_YEST, HCC, MCC, WS850, WS_AFTE, PR, PS_DELTA_YEST, WD, WD700 | 决策树个数: 40<br>抽取变量个数: 6 |
| $PM_{2.5}$ | MLR | $PM_{2.5}$ 浓度(自然对数值) | WS, PS_DELTA, WD.NNE, WD.NE, PM2.5_YEST(自然对数值) | - |
| | NN | $PM_{2.5}$ 浓度(自然对数值) | PBL, WS, T700_MAX, PS_DELTA, PR, GHT500, WD.NNE, WD.NE, PM2.5_YEST(自然对数值) | 隐含层节点个数: 6 |
| | RF | $PM_{2.5}$ 浓度 | PM2.5_YEST, PBL, WS, T700_MAX, PS_DELTA, WD_CHANGE, PS_DELTA_YEST, PR, WS500, GHT500, WD | 决策树个数: 70<br>抽取变量个数: 6 |



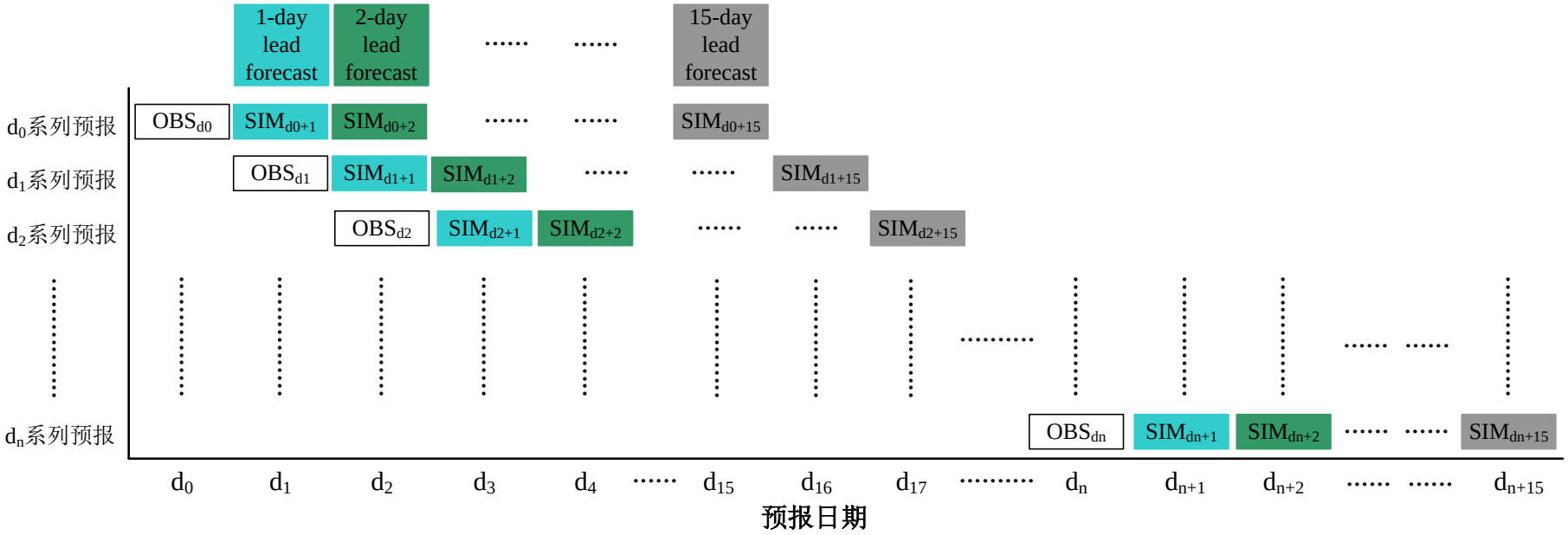

**附录 7　成都市 2016—2018 年夏季 $O_3$ 浓度观测值及模拟值时间序列**

Appendix 7　Timeseries of $O_3$ observed and simulated concentrations in Chengdu in summer during 2016–2018




**附录 8　成都市 2016—2018 年冬季月份 $PM_{2.5}$ 浓度观测值及模拟值时间序列**

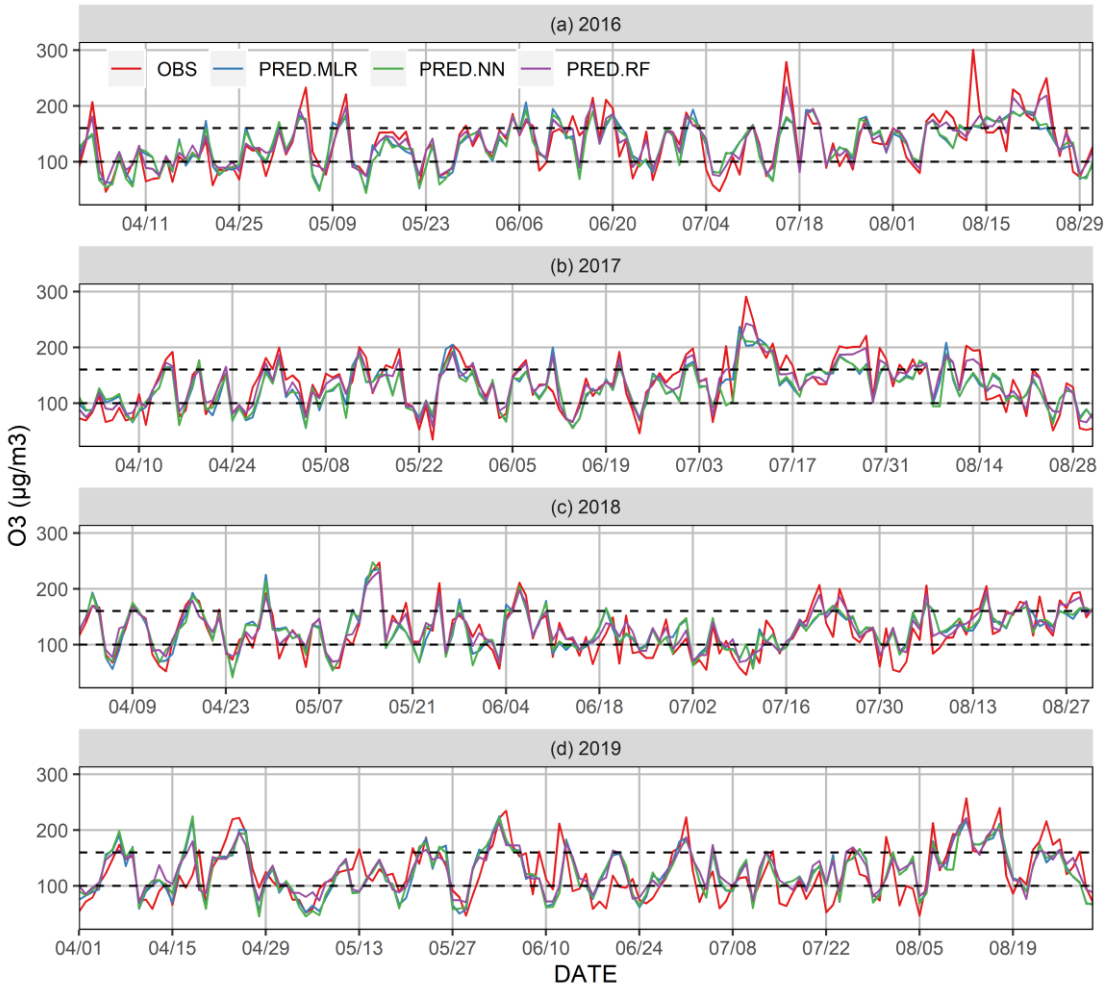
Appendix 8　Timeseries of $PM_{2.5}$ observed and simulated concentrations in Chengdu in winter months during 2016–2018




$d_k(k = 0\sim n)$系列提前 1~15 天预报：以 OBS$d_k$ 为起始的预报系列，分别生成预报浓度 SIM$_{dk+1}$, SIM$_{dk+2}$, ⋯, SIM$_{dk+15}$；其中除 SIM$_{dk+1}$ 预报值以 OBS$_{dk}$ 为 PM25_YEST(O3_YEST)获得外，其余各天的提前预报均以前一日预报值为 PM25_YEST(O3_YEST)获得

**附录 9　臭氧和 $PM_{2.5}$ 提前 1~15 天污染潜势预报示意图**

Appendix 9　Diagram of producing 1–15 -day lead forecasts for $O_3$ and $PM_{2.5}$



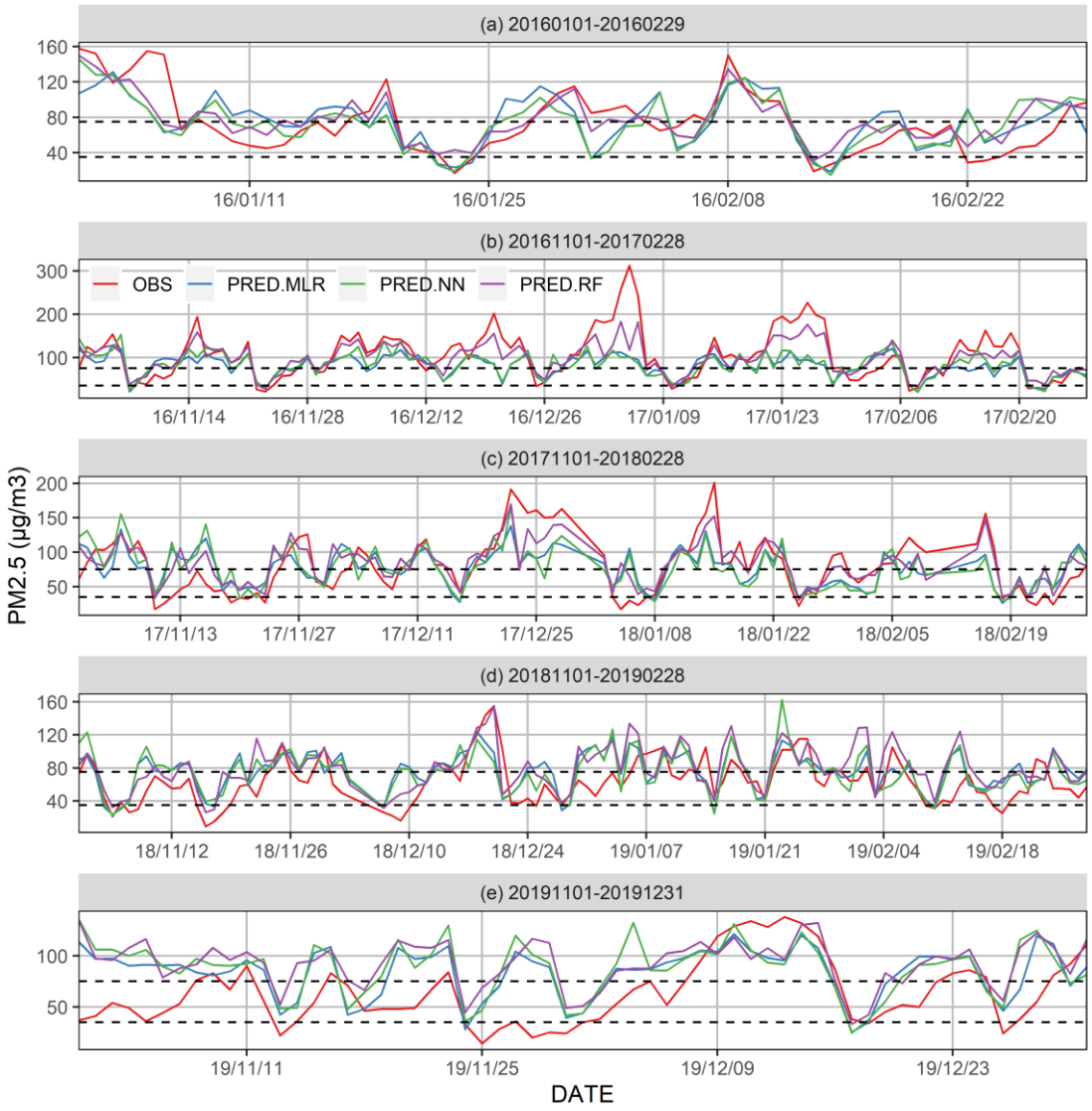

**附录 10　成都市夏季臭氧污染潜势预报值及观测值时间序列(去除预报变量 O3_YEST)**

Appendix 10 Timeseries of ozone pollution potential forecasts (with variable O3_YEST removed from models) versus observations for Chengdu

**附录 11　成都市臭氧及 $PM_{2.5}$ 污染潜势预报模型预报效果评估(去除前一日浓度变量)**

Appendix 11　Evaluation of the ozone and PM2.5 pollution potential forecast models in Chengdu (with O3_YEST/PM2.5_YEST removed)

| 污染物 | 方法 | 数据集 | R | BIAS | GE | RMSE | 分类误判率 |
|---|---|---|---|---|---|---|---|
| $O_3$ | MLR | 训练集 | 0.81 | -2.95 | 21.51 | 27.06 | 30.26 |
| | | 测试集 | 0.74 | -1.55 | 23.61 | 31.1 | 33.93 |
| | | 回顾预报集 | 0.72 | 1.7 | 25.81 | 33.09 | 41.83 |
| | NN | 训练集 | 0.82 | -3.02 | 20.9 | 26.2 | 27.95 |
| | | 测试集 | 0.72 | -2.74 | 23.64 | 32.12 | 34.82 |
| | | 回顾预报集 | 0.71 | 1.98 | 26.49 | 33.85 | 39.87 |
| | RF | 训练集 | 0.98 | 0.22 | 9.33 | 11.94 | 9.51 |
| | | 测试集 | 0.77 | 4.11 | 22.45 | 29.58 | 34.82 |
| | | 回顾预报集 | 0.74 | 5.1 | 24.87 | 31.58 | 37.25 |
| $PM_{2.5}$ | MLR | 训练集 | 0.56 | -8.16 | 29.54 | 39.86 | 36.36 |
| | | 测试集 | 0.53 | -4.22 | 26.86 | 38.98 | 37.5 |
| | | 回顾预报集 | 0.47 | 15.27 | 24.93 | 30.34 | 53.33 |
| | NN | 训练集 | 0.57 | -7.55 | 28.74 | 39.54 | 36.36 |
| | | 测试集 | 0.5 | -3.55 | 27.86 | 39.91 | 41.67 |
| | | 回顾预报集 | 0.38 | 16.17 | 27.09 | 34.18 | 53.33 |
| | RF | 训练集 | 0.96 | 0.49 | 13.33 | 17.66 | 21.74 |
| | | 测试集 | 0.5 | 4.3 | 29.56 | 39.77 | 65.62 |
| | | 回顾预报集 | 0.43 | 23.9 | 30.03 | 36.45 | 58.33 |




**附录 12　成都市冬季月份 $PM_{2.5}$ 污染潜势预报值和观测值时间序列(去除预报变量 PM2.5_YEST)**

Appendix 12　Timeseries of $PM_{2.5}$ pollution potential forecasts (with variable PM2.5_YEST removed from models) versus observations for Chengdu