

基于贝叶斯优化VMD-LSTM的成都市近地面臭氧浓度预测

孙玉鸿¹, 蒋露², 邱诚¹, 杨丽娇¹, 甘谨瑜¹

¹成都工业学院材料与环境工程学院, 四川 成都

²成都工业学院计算机工程学院, 四川 成都

收稿日期: 2026年8月11日; 录用日期: 2026年9月10日; 发布日期: 2026年9月17日

摘要

近地面臭氧已成为影响成都市夏季空气质量达标率和公众健康的首要污染物。臭氧的生成受气象条件、前体物排放及光化学反应等多重因素的非线性控制, 传统单一预测模型难以实现高精度预报。针对臭氧浓度的波动性与非线性特征, 本研究提出了一种基于贝叶斯优化(Bayesian Optimization, BO)与变分模态分解(Variational Mode Decomposition, VMD)相融合的长短期记忆网络(LSTM)臭氧预测模型(BO-VMD-LSTM)。利用成都市历史逐日空气质量与气象观测数据, 通过VMD技术将复杂的臭氧序列分解为多个相对平稳的子序列; 再利用LSTM网络捕捉时间的依赖性; 最后创新性地引入严格模式的贝叶斯优化, 对模型的四个参数(分解层数 K 、惩罚因子 α 、隐藏层节点数及学习率)进行全局联合寻优。实验结果表明, 在最优参数配置下, 模型在测试集上取得了优异的预测性能 ($R^2 = 0.9044 \pm 0.0021$, $RMSE = 10.0600 \pm 0.1096 \mu\text{g}/\text{m}^3$, $MAE = 7.7218 \pm 0.0925 \mu\text{g}/\text{m}^3$)。该模型为成都市重污染天气应急管控和公众健康防护提供了科学的高精度预警工具。

关键词

臭氧浓度预测, 变分模态分解, 长短期记忆网络, 贝叶斯优化, 成都市

Research on Prediction of Near-Surface Ozone Concentration in Chengdu Based on Bayesian Optimization VMD-LSTM

Yuhong Sun¹, Lu Jiang², Cheng Qiu¹, Lijiao Yang¹, Jinyu Gan¹

¹School of Materials and Environmental Engineering, Chengdu Technological University, Chengdu Sichuan

²School of Computer Engineering, Chengdu Technological University, Chengdu Sichuan

Received: August 11, 2026; accepted: September 10, 2026; published: September 17, 2026

文章引用: 孙玉鸿, 蒋露, 邱诚, 杨丽娇, 甘谨瑜. 基于贝叶斯优化 VMD-LSTM 的成都市近地面臭氧浓度预测[J]. 环境保护前沿, 2026, 16(9): 1466-1477. DOI: 10.12677/aep.2026.169148

Abstract

Near-surface ozone has become the primary pollutant affecting the summer air quality compliance rate and public health in Chengdu. The formation of ozone is non-linearly controlled by multiple factors such as meteorological conditions, precursor emissions, and photochemical reactions, making it difficult for traditional single-prediction models to achieve high-precision forecasting. To address the volatility and non-linear characteristics of ozone concentration, this study proposes an ozone prediction model (BO-VMD-LSTM) integrating Long Short-Term Memory (LSTM) network with Bayesian Optimization (BO) and Variational Mode Decomposition (VMD). Utilizing historical daily air quality and meteorological observation data in Chengdu, complex ozone sequences were decomposed into multiple relatively stable sub-sequences via VMD technology; the LSTM network was then employed to capture temporal dependencies; finally, a strict-mode Bayesian optimization was innovatively introduced to perform global joint optimization on four model parameters (decomposition level K , penalty factor α , number of hidden layer nodes, and learning rate). Experimental results show that under the optimal parameter configuration, the model achieves excellent prediction performance on the test set ($R^2 = 0.9044 \pm 0.0021$, $RMSE = 10.0600 \pm 0.1096 \mu\text{g}/\text{m}^3$, $MAE = 7.7218 \pm 0.0925 \mu\text{g}/\text{m}^3$). This model provides a scientific and high-precision early warning tool for emergency control of heavy air pollution and public health protection in Chengdu.

Keywords

Ozone Concentration Prediction, Variational Mode Decomposition (VMD), Long Short-Term Memory (LSTM) Network, Bayesian Optimization (BO), Chengdu

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近地面臭氧已成为成都市夏季的首要污染物, 对人体健康和生态系统均有显著危害[1]。臭氧属于二次光化学污染物, 其浓度并不由单一排放源直接决定, 而是前体物(如挥发性有机物和氮氧化物)在高温、强辐射等条件下发生光化学反应生成的[2] [3]。这种生成机制使臭氧序列天然带有强非线性和高波动性, 预报难度远大于一次污染物。

在预测方法上, 早期的数值预报模型因计算成本高、排放清单更新不及时等问题, 难以满足日常业务需求[4]。随后, 研究者转向数据驱动的机器学习方法。许成丽等[1]用 SARIMA-BP 混合模型做城市臭氧预测, 验证了时序分解配合神经网络是可行的路子; 杨晓彤等[2]在四川盆地用随机森林建模, 发现集成学习的精度明显好于单一模型; 朱珈莹等[5]则用 LightGBM 在南京开展预测, 进一步证实了这个趋势。潘梦瑶等[6]把梯度提升和 SHAP 结合起来, 不光预测臭氧浓度, 还顺带量化了各个因子的贡献大小。路旭东等[3]也从可解释机器学习的角度, 分析了影响兰州臭氧的关键变量。

问题是, 这些方法说到底还是浅层模型。申子彬等[7]在宁波的工作也印证了, 单一统计模型碰到臭氧这种波动剧烈的序列, 峰值经常抓不住、整体精度也有限。即便是集成树模型, 对序列里那种跨越好几天的“长程依赖”也挖得不够深。所以, 把深度学习引进来, 再配合信号分解做预处理, 就成了一个值得尝试的方向。

顺着这个思路,本研究以成都市为对象,构建了一个 VMD、LSTM 和贝叶斯优化三者结合的预测模型。具体来说做了三件事:(1) 用 VMD 把原始臭氧序列拆成几个相对平稳的模态分量,从源头降低非平稳性;(2) 对每个分量分别训练 LSTM,靠门控机制去抓长程时序依赖;(3) 用贝叶斯优化把 VMD 的模态数 K 和惩罚因子 α 、LSTM 的隐藏单元数和学习率这四个参数捆在一起联合寻优,省去了人工调参的盲目性。我们希望这个模型能为成都市的臭氧重污染预警提供一个好用的工具,同时也给其他城市的同类工作提供一点参考。

2. 数据与方法

2.1. 研究区域与数据来源

本研究以成都市为研究区域,所采用的数据包括长序列逐日空气质量监测数据与同期气象观测资料。空气质量数据来源于国家空气质量实时监测平台,包含 O_3 、 $PM_{2.5}$ 、 PM_{10} 、 SO_2 、 NO_2 、 CO 等六项常规污染物浓度;气象数据来源于国家气象科学数据中心,包含气温、相对湿度、降水量、风速、日照时数、能见度、边界层高度及云量等要素。

原始数据存储在“空气数据 6 城市.xlsx”文件中,本研究选取工作表“成都”进行建模分析。数据读取范围为 F2:T3013,共包含 15 列特征变量,其中第 12 列为预测目标变量——臭氧 8 小时滑动平均值(O_{3_8h} ,单位: $\mu g/m^3$)。其余 14 列特征涵盖气象因子与污染物因子,共同构成模型输入矩阵。

2.2. 数据预处理与质量控制

原始监测数据在采集与传输过程中不可避免地存在少量缺失,为保障模型训练质量,本研究实施了以下预处理流程。

缺失值处理: 对每一列特征,检测并定位缺失值(NaN),采用该列非缺失值的均值进行填充。其数学表达为:

$$x_{i,j} = \frac{1}{N_{\text{valid}}} \sum_{k \in K_j} x_{k,j}, \text{ if } x_{i,j} \text{ is NaN}$$

其中, $x_{i,j}$: 数据矩阵中第 i 行、第 j 列的观测值。

NaN: Not a Number (非数字),即数据采集或传输过程中产生的无效空值。

K_j : 第 j 列所有有效观测值的索引集。

N_{valid} : 有效样本数。

公式释义: 若检测到某单元格数据为 NaN,则利用该特征列所有有效观测值的算术平均数进行填补,以确保时间序列的连续性。

归一化处理: 为消除不同特征间的差异对神经网络训练的不利影响,采用最大最小归一化(Min-Max Normalization)将各特征线性映射至[0, 1]区间:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

其中, $x_{\min}$ 与 $x_{\max}$ 分别为该特征在训练集中的最小值与最大值。归一化参数严格基于训练集计算,测试集仅应用相同的线性变换参数,以杜绝数据泄露。

时序数据划分: 按时间顺序将数据集划分为训练集(前 70%)与测试集(后 30%),测试集全程独立于模型训练与超参数寻优过程。采用滑动窗口技术来构建监督学习样本,设定时间步长 $T_w = 7$,即利用过去连续 7 天的多维特征预测第 8 天的臭氧浓度。经重构后,训练集可用于建模的有效序列样本数为 $n_{tr_seq} = n_{tr} - T_w$,测试集为 $n_{te_seq} = n_{te} - T_w$ 。

2.3. BO-VMD-LSTM 预测模型构建

针对臭氧时间序列强非线性、高非平稳性的特征,本研究构建了集成“信号分解-贝叶斯寻优-深度预测”于一体的 BO-VMD-LSTM 混合预测框架。该框架由三个核心模块构成:变分模态分解层、贝叶斯优化超参数寻优层以及长短期记忆网络预测层。

2.3.1. 变分模态分解(VMD)层

VMD 是一种完全非递归、自适应的信号分解方法,通过迭代搜寻约束变分模型的最优解,将复杂时间序列 $f(t)$ 自适应地解耦为 K 个具有有限带宽的本征模态函数(IMF) $u_k(t)$ [8]。其核心约束变分模型为:

$$\min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \text{ s.t. } \sum_{k=1}^K u_k = f(t)$$

其中, ω_k 为各模态的中心频率, α 为二次惩罚因子,控制模态带宽。

VMD 的分解效果高度依赖于两个关键参数——模态数 K 与惩罚因子 α 。 K 过小将导致模态混叠,过大则产生虚假冗余分量; α 直接影响各模态的带宽与抗噪性。传统方法多依赖人工经验设定固定参数,本研究将 K 与 α 释放为待优化变量,交由贝叶斯优化模块自动搜索最优组合。

2.3.2. 贝叶斯优化(BO)超参数寻优层

在 VMD-LSTM 混合建模中,涉及多个对预测性能影响显著的核心超参数:VMD 的模态数 K 与惩罚因子 α ,以及 LSTM 的隐藏层单元数与初始学习率。传统的人工调参或网格搜索方法存在效率低、易陷入局部最优等局限。本研究引入贝叶斯优化算法进行全局联合寻优[9]。

贝叶斯优化通过高斯过程(Gaussian Process, GP)对目标函数 $f(\theta)$ 进行概率代理建模,并利用采集函数(Acquisition Function)在探索与利用之间实现平衡,从而以较少的评估次数逼近全局最优解。

本研究四参数联合搜索空间设定如下:

- 模态数 K : 整数型, 搜索范围[8, 12]。
- 惩罚因子 α : 整数型, 搜索范围[2000, 4000]。
- LSTM 隐藏单元数: 整数型, 搜索范围[16, 128]。
- 初始学习率: 对数型, 搜索范围[10^{-4} , 10^{-2}]。

在本研究中,超参数搜索空间的设定并非凭空假设,而是基于特征解耦效能、模型容量与收敛稳定性进行的科学界定:

(1) **模态数 K (搜索范围[8, 12]):** 若 K 取值过小,则难以将臭氧序列错综复杂的非线性波动充分剥离,导致模态混叠;若 K 过大,则会产生高频冗余分量并引入虚假噪声。设定该区间旨在兼顾信号分解的充分性与计算效率。

(2) **惩罚因子 α (搜索范围[2000, 4000]):** α 决定了本征模态函数(IMF)的带宽。传统经验常固定为较小值,但在面对大气环境高噪声数据时极易导致噪声滤除不彻底。较大的搜索区间赋予了算法更强的带宽收窄能力,从而在频域实现更纯净的正交解耦。

(3) **LSTM 隐藏单元数 H (搜索范围[16, 128]):** 该参数决定了网络的非线性表达能力。单元数过少会导致网络容量不足以拟合复杂映射,过多则在逐日样本下极易引发过拟合。设定此范围平衡了模型的泛化能力与拟合精度。

(4) **初始学习率 η (搜索范围[10^{-4} , 10^{-2}]):** 学习率直接关系到 Adam 优化器的收敛速度。采用对数均匀分布的搜索范围,可全面覆盖从细粒度微调到快速收敛的有效区间,确保梯度平稳下降,避免训练陷入震荡或停滞。

严格模式下的目标函数设计: 为防止超参数筛选过程中的数据泄露, 本研究设计了一种严格限定于训练集内部的验证策略。对于贝叶斯优化器生成的每一组候选超参数 $\theta = (K, \alpha, H, \eta)$, 目标函数的计算流程如下:

(1) **训练集内部 VMD 分解:** 仅使用训练集臭氧序列进行 VMD 分解, 得到 K 个模态分量, 测试集信息绝对不可见。

(2) **内部二次划分:** 将训练集按时间顺序重新划分为 80% 子训练集与 20% 内部验证集。

(3) **模态独立建模与验证:** 对每个模态分量, 利用子训练集构建 LSTM 网络, 在内部验证集上进行预测并反归一化, 将所有模态预测值求和得到最终预测。

(4) **损失计算:** 以内部验证集的均方根误差(RMSE)作为目标函数值, 返回给贝叶斯优化器:

$$f_{obj}(\theta) = \sqrt{\frac{1}{n_{val}} \sum_{i=1}^{n_{val}} (y_i - \hat{y}_i)^2}$$

贝叶斯优化共运行 15 次目标函数评估, 最终筛选出使内部验证 RMSE 最小的最优超参数组合 $(\hat{K}, \hat{\alpha}, \hat{H}, \hat{\eta})$ 。

2.3.3. 长短期记忆网络(LSTM)预测层

LSTM 是一种专为序列建模设计的循环神经网络变体, 通过引入遗忘门、输入门和输出门三种门控机制, 有效解决了传统 RNN 在处理长序列时面临的梯度消失与梯度爆炸问题[10][11]。其核心计算可形式化表示为:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned}$$

其中, f_t 、 i_t 、 o_t 分别为遗忘门、输入门、输出门的输出, C_t 为细胞状态(长期记忆), h_t 为隐藏状态(短期记忆)。

在本框架中, 经 VMD 分解得到的最优模态数 $\hat{K}$ 个分量, 逐一构建独立的 LSTM 预测子模型。每个子模型的网络结构为: 序列输入层 → LSTM 层(隐藏单元数 $\hat{H}$, 输出模式 last, 即仅取序列末尾的输出) → 全连接层(1 个神经元) → 回归层。优化器选用 Adam, 最大训练轮次为 20, 批量大小为 128, 初始学习率为 $\hat{\eta}$ 。各模态预测值经反归一化后进行线性求和, 得到最终的臭氧浓度预测值。

多种子集成策略: 为降低神经网络训练过程中随机权重初始化的影响并量化预测不确定性, 本研究采用多种子集成策略。固定最优超参数后, 使用 5 组不同的随机种子(42, 123, 270, 424, 888)独立初始化 LSTM 网络并分别训练, 得到 5 组测试集预测结果。取 5 次预测的均值作为最终预测输出, 标准差作为预测不确定性的度量:

$$\hat{y}_{\text{final}} = \frac{1}{5} \sum_{s=1}^5 \hat{y}^{(s)}, \sigma = \sqrt{\frac{1}{5} \sum_{s=1}^5 (\hat{y}^{(s)} - \hat{y}_{\text{final}})^2}$$

2.4. 模型评价指标

为定量评估模型预测性能, 本研究选用以下三个统计学指标:

决定系数(R^2):

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

R^2 越接近 1, 表明模型对数据变异的解释能力越强。

均方根误差(RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

RMSE 对较大误差敏感, 数值越小代表整体预测精度越高。

平均绝对误差(MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

MAE 直观反映预测误差的平均幅度, 对异常值具有较好的鲁棒性。

其中, y_i 为第 i 个样本的实测值, $\hat{y}_i$ 为模型预测值, $\bar{y}$ 为实测值均值, n 为测试集样本总数。最终报告各指标在 5 组随机种子实验下的均值与标准差, 以全面反映模型的预测精度与稳定性。

3. 结果与分析

为全面检验 BO-VMD-LSTM 模型的实际预测性能, 本章从超参数寻优、VMD 分解形态、时序预测拟合度、散点相关性以及横向对比分析五个维度进行系统评估。所有实验结果均基于独立测试集获得, 测试集数据全程未参与模型训练与超参数筛选。

3.1. 贝叶斯超参数寻优结果

模型性能的优劣高度依赖于超参数的合理配置。本研究采用贝叶斯优化算法对 VMD 模态数 K 、惩罚因子 α 、LSTM 隐藏层单元数 H 及初始学习率 η 进行四参数联合寻优。优化过程共进行 15 次目标函数评估, 每次评估均在训练集内部执行严格的 80%/20% 划分验证, 以内部验证集的均方根误差(RMSE)作为目标函数值。

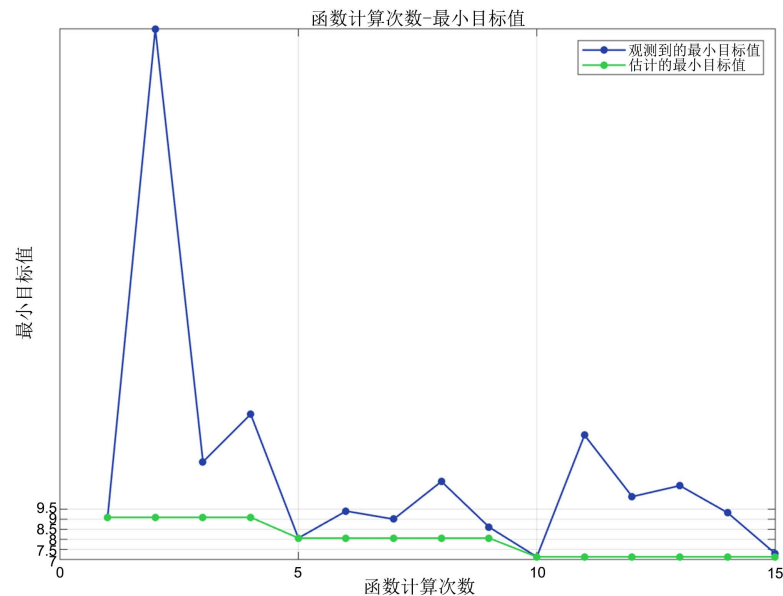


Figure 1. Bayesian optimization iteration curve
图 1. 贝叶斯优化迭代曲线

图 1 展示了目标函数值随迭代步数的变化趋势。随着迭代推进，贝叶斯优化器的高斯过程代理模型逐步逼近真实的目标函数空间，验证 RMSE 呈稳定下降态势，表明优化过程收敛良好。

最终确定的最优超参数组合如表 1 所示。

Table 1. Bayesian optimization for optimal hyperparameter configuration
表 1. 贝叶斯优化最优超参数配置

超参数	含义	最优值	搜索范围
K	VMD 模态数	8	[8, 12]
α	VMD 惩罚因子	3157	[2000, 4000]
H	LSTM 隐藏单元数	42	[16, 128]
η	初始学习率	0.0099737	$[10^{-4}, 10^{-2}]$

值得关注的是，优化算法自主选择了搜索空间内较低的模态数($K = 8$)与较高的惩罚因子($\alpha = 3157$)。相比于传统研究中凭经验固定 $\alpha = 1500$ 的做法，更大的惩罚因子倾向于产生带宽更窄的本征模态函数(IMF)，有助于在高噪声大气环境数据中分离出更纯净的振荡成分，从信号处理层面为后续 LSTM 建模奠定更优基础。这一结果表明，人工经验设定的参数往往并非理论最优，采用自动化联合寻优策略具有显著必要性。

3.2. VMD 分解形态分析

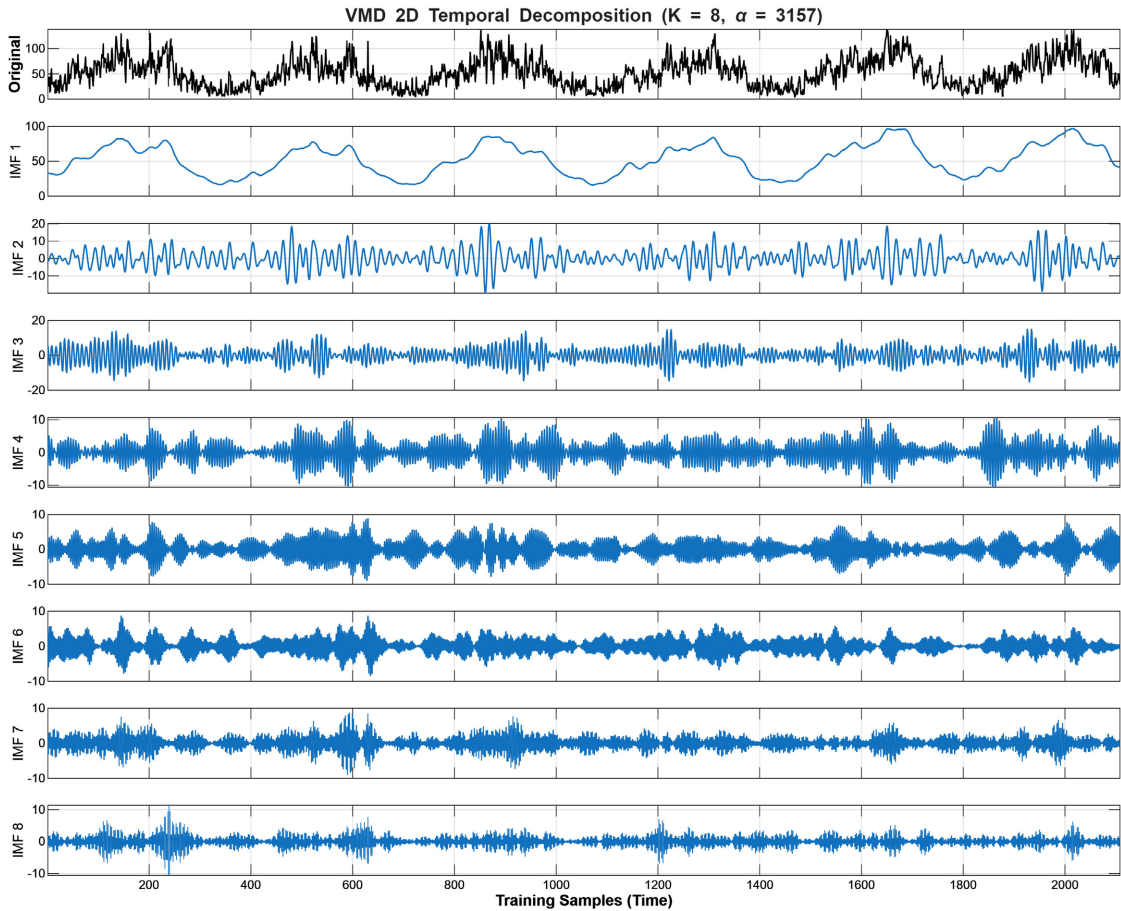


Figure 2. VMD 2D time series decomposition diagram ($K = 8, \alpha = 3157$)
图 2. VMD 2D 时序分解图($K = 8, \alpha = 3157$)

为了直观展示变分模态分解(VMD)在处理成都市臭氧非平稳时间序列时的实际效能, 图 2 给出了基于贝叶斯优化锁定最优参数($K=8, \alpha=3157$)得到的各本征模态函数(IMF)分量的时序分布图。

从图 2 的 2D 时序分解结果可以看出, 原本高度复杂的原始臭氧序列(Original)经过 VMD 自适应分解后, 被解耦为 8 个具有不同中心频率与有限带宽的平稳分量。具体而言:

(1) **低频趋势项**: 模态 1 (IMF 1)的波形最为平缓, 波动幅度大且无高频毛刺, 它完美捕获了成都市臭氧浓度随季节交替呈现的宏观演变趋势与长期周期性规律。

(2) **中频周期项**: 模态 2 至模态 4 的频率适中, 展现了臭氧在特定月份或季节转换过程中的准周期性波动特征, 体现了气象条件阶段性变化对浓度序列的调制作用。

(3) **高频噪声与突变项**: 模态 5 至模态 8 的频率较高、幅值较小且多呈零均值的随机振荡状态, 主要表征了大气环境中的随机噪声、观测误差以及局地突发性污染事件引起的短期剧烈扰动。

3.3. 臭氧浓度时序预测结果

采用表 1 所列最优超参数, 在独立测试集上对成都市臭氧 8 小时滑动平均浓度(O_{3_8h})进行回溯预测。为降低神经网络随机初始化引入的不确定性, 固定最优参数后使用 5 组随机种子(42, 123, 270, 424, 888)分别独立训练并预测, 取均值作为最终预测结果, 标准差作为预测不确定性度量。

图 3 展示了测试集上的预测值与实测值时序对比。

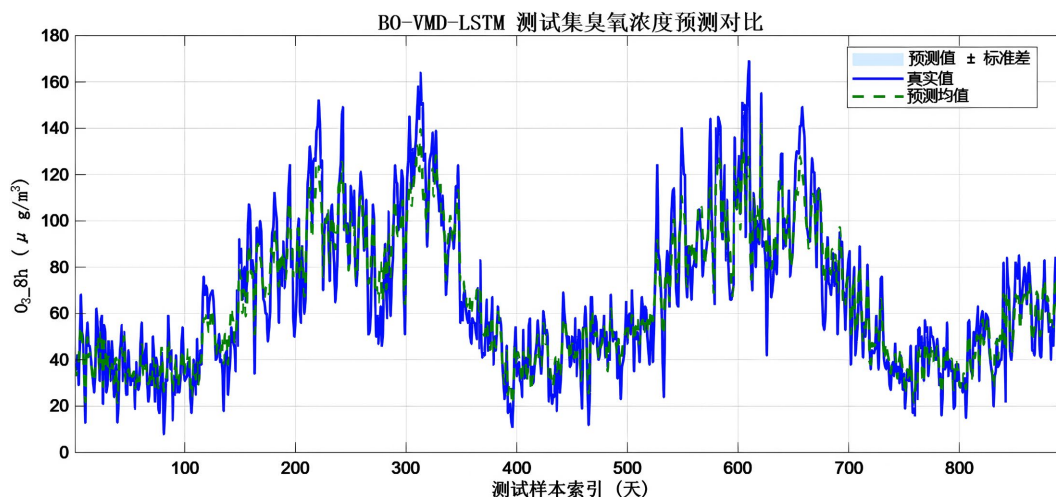


Figure 3. Ozone concentration time series prediction comparison diagram

图 3. 臭氧浓度时序预测对比图

从图 3 中可以直观观察到以下特征:

(1) **整体趋势追踪**: 绿色虚线的预测均值曲线与蓝色实线的实测值曲线在全时段内保持了高度一致, 未出现明显的相位漂移或系统性高估/低估, 说明模型有效学习到了成都市臭氧浓度的季节性演变和逐日波动规律。

(2) **峰值捕捉能力**: 在臭氧浓度急剧攀升的高污染时段(浓度突破 $140 \mu\text{g}/\text{m}^3$), 模型预测曲线依然紧密贴合实测曲线, 未表现出传统模型中常见的峰值滞后或过度平滑现象。这一优势得益于 VMD 将强非平稳序列预先解构为相对平稳的模态分量, 使得 LSTM 能够针对不同频率分量进行精细化建模。

(3) **不确定性量化**: 图中浅蓝色阴影区域表示 5 次集成预测的 ± 1 标准差带宽。该带宽在全测试时段内保持极窄, 表明不同随机初始化条件下模型的预测输出高度一致, 模型架构具有优良的稳健性。

量化评估方面:

表 2 汇总了 5 组随机种子实验的统计指标。本模型在测试集上取得了优异表现：决定系数 $R^2 = 0.9044 \pm 0.0021$ ，均方根误差 $RMSE = 10.0600 \pm 0.1096 \mu\text{g}/\text{m}^3$ ，平均绝对误差 $MAE = 7.7218 \pm 0.0925 \mu\text{g}/\text{m}^3$ 。

Table 2. Test set prediction performance indicators (mean $\pm$ standard deviation)
表 2. 测试集预测性能指标(均值 $\pm$ 标准差)

评价指标	均值	标准差
R^2	0.9044	0.0021
RMSE ($\mu\text{g}/\text{m}^3$)	10.0600	0.1096
MAE ($\mu\text{g}/\text{m}^3$)	7.7218	0.0925

R^2 高达 0.9044 意味着模型能够解释成都市臭氧浓度实测方差超过 90%，而极小的标准差(R^2 标准差仅为 0.0021)有力证明了在最优超参数锁定后，该预测框架具有极强的可复现性和稳定性，为业务化预报提供了可靠保障。

3.4. 预测值与实测值散点相关性分析

图 4 以散点图形式展示了测试集所有样本的预测值与实测值对应关系。横坐标为实测臭氧浓度，纵坐标为模型预测浓度，红色对角线代表理想预测线($y = x$)。

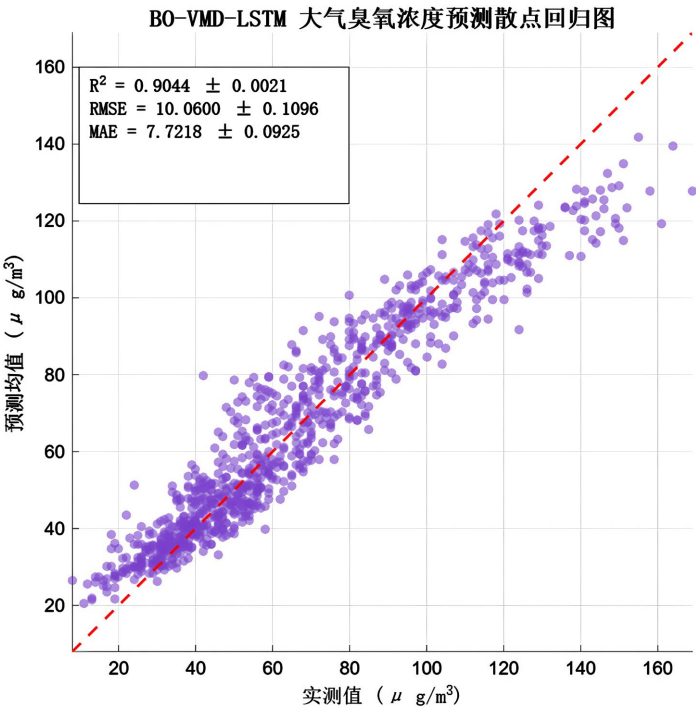


Figure 4. Scatter regression plot of predicted values vs measured values
图 4. 预测值与实测值散点回归图

从散点整体分布来看，绝大部分样本点紧密聚集于对角线两侧，无显著系统性偏倚。在中等浓度区间(50~150 $\mu\text{g}/\text{m}^3$)，散点的集聚度最高，表明模型对成都市常规气象-化学条件下臭氧演变规律的学习已趋于充分。在低浓度区间(<50 $\mu\text{g}/\text{m}^3$)，散点同样维持了较窄的分布带，未出现明显的高估倾向。在高浓度

区间(>150 $\mu\text{g}/\text{m}^3$), 散点离散度相对有所增加, 这主要归因于极端重污染事件在全年数据集中占比较小, 属于深度学习模型在样本分布不均衡条件下的共性特征。尽管如此, 高值区散点仍基本维持在对角线附近, 未出现偏离对角线的大幅度预测崩溃, 充分证实 VMD 分解有效缓解了非平稳性导致的极端值预测难题。

上述散点图拟合特征与表 2 中的指标互为印证: R^2 接近 1 反映散点沿对角线集中, RMSE 和 MAE 的较小数值则量化为预测误差的绝对幅度。

3.5. 讨论与对比分析

为客观评估本模型在同类研究中的性能定位, 表 3 将本研究结果与近年文献中报道的相关工作进行了横向对比。

Table 3. Comparison of ozone prediction performance of different models

表 3. 不同模型臭氧预测性能对比

模型	研究区域	R^2	RMSE	数据来源
随机森林	四川盆地	0.87	16.70	杨晓彤等[2]
决策树统计预报	徐州	0.42	27.42	丁慷等[4]
渐进梯度回归树	宁波	0.64	19.24	申子彬等[7]
BO-VMD-LSTM (本研究)	成都	0.9044	10.06	本文

综合分析, 本模型性能的提升主要源于以下三个层面的协同作用:

(1) **VMD 的频率解耦效应**[8]: 臭氧浓度序列受局地光化学生成、区域输送及气象扩散条件等多尺度因素耦合控制, 直接对原始非平稳序列建模难度极大。VMD 通过变分约束将高度复杂的序列自适应地拆分为有限个相对平稳的 IMF 分量, 每个分量对应特定的频率尺度。LSTM 针对不同频率分量分别学习其演变规律, 有效避免了噪声与趋势相互干扰导致的过拟合或欠拟合, 从信号处理层面降低了建模复杂度。

(2) **贝叶斯优化的全局寻优能力**[9]: 传统研究中 VMD 的 K 和 α 通常依赖经验设定, LSTM 的隐藏单元数与学习率也多靠试凑。本研究将上述四个核心参数统一纳入 BO 的四维搜索空间, 在训练集内部执行严格交叉验证, 自动挖掘出人工难以企及的最优参数组合。特别是优化器自主选择了较高的 $\alpha = 3157$, 其窄带分离效应对含噪大气数据的适应性优于传统默认值, 这一发现对同类研究具有参考价值。

(3) **严格防泄漏验证与集成预测策略**[10] [11]: BO 寻优过程中的 VMD 分解和验证严格限定在训练集内部, 杜绝了测试集信息泄露导致的评价虚高; 多种子集成策略在保证精度的同时, 将预测标准差控制在极小范围内, 增强了业务化应用的可靠性。LSTM 门控机制对时间序列长程依赖关系的有效捕获, 则为各模态分量的精准建模提供了算法基础[10] [11]。

综上, BO-VMD-LSTM 在预测精度、峰值捕捉和稳健性三方面均展现出明显优势, 为成都市臭氧污染预警提供了可靠的技术方案。

4. 结论与展望

4.1. 主要结论

本研究以成都市的逐日空气质量和气象数据为基础, 针对臭氧序列强非线性、高波动的问题, 构建并验证了融合 VMD、贝叶斯优化和 LSTM 的 BO-VMD-LSTM 混合预测模型。主要结论如下:

(1) VMD 分解把原始臭氧序列拆成了几个相对平稳的模态分量, 相当于从源头把非平稳性降了下来。

这样一来,后面的 LSTM 网络不需要同时处理趋势、周期和噪声,可以分别对不同频率的分量去学习,整体建模难度明显降低。

(2) 贝叶斯优化帮我们绕开了人工一个个试参数的麻烦。它把模态数 K 、惩罚因子 α 、隐藏单元数和学习率这四个参数放在一起联合搜索,最终锁定了 $K=8$ 、 $\alpha=3157$ 、 $H=42$ 、 $\eta=0.0099737$ 这一组较优的配置。而且整个寻优过程都关在训练集内部做验证,没有让测试集信息“偷跑”进去,结果更经得起推敲。

(3) 在独立测试集上,模型 R^2 超过了 0.90, RMSE 和 MAE 都压得比较低,预测的标准差带也很窄。从时序图上看,预测曲线和实测值跟得很紧,重污染天气的峰值基本都抓住了,没有出现传统方法那种明显滞后或者偏低的情况。散点图也进一步印证了这一点,绝大部分点都老老实实趴在对角线附近。

(4) 总的来看,“信号分解-智能寻优-深度预测”这套技术路线在成都的臭氧预测上是跑通了,精度明显优于文献里一些传统的单一模型或集成树模型。这套框架是模块化搭的,以后换到别的城市、别的污染物,思路可以直接搬过去用。

4.2. 不足与展望

本研究虽然取得了一些结果,但回过头看,还有几个方面值得继续往下做:

(1) **数据时间分辨率有待提升**。目前用的是逐日数据,看宏观趋势还行,但臭氧的小时级峰值突变就抓不到了。后面有条件的话,可以把小时级监测数据也融进来,试试多时间尺度嵌套建模,看看能不能把日内的快速波动也捕捉到。

(2) **空间泛化能力尚需检验**。模型现在只在成都跑过,换到别的城市效果怎么样其实不清楚。下一步可以往成渝经济圈的其他城市推一推,或者挑几个排放结构和气候条件差别比较大的地方做测试,看看这套框架的适应性到底如何。

(3) **模型结构仍有优化空间**。现在用的是单层单向 LSTM,结构比较基础。后面可以试试双向 LSTM、加注意力机制,甚至换 Transformer,看看能不能再往上提一点精度。另外,VMD 分解在序列两端会有边界效应,可能影响边界样本的预测稳定性,后面可以考虑用镜像延拓这类方法缓解一下。

(4) **模型可解释性有待加强**。这篇论文主要精力都放在提高精度上了,但模型到底是怎么作出判断的、哪个因子贡献大、哪个基本没用,其实没有深究。后面可以引入 SHAP 这类工具[6],把各气象和污染因子的贡献度算出来,既能让模型更透明,也能给臭氧成因分析提供一点线索。

基金项目

2026 年国家级大学生创新训练计划项目(202611116008)。

参考文献

- [1] 许成丽,郑朝阳. 基于 SARIMA-BP 神经网络的城市近地面臭氧预测[J]. 环境科学, 2026, 47(3): 1389-1399.
- [2] 杨晓彤,康平,王安怡,等. 基于随机森林模型的四川盆地臭氧污染预测[J]. 环境科学, 2024, 45(5): 2507-2515.
- [3] 路旭东,张定海,黄天豪,等. 基于可解释机器学习的兰州市近地面臭氧浓度影响因素分析[J/OL]. 环境科学: 1-17. <https://doi.org/10.13227/j.hjlx.202509063>, 2026-01-06.
- [4] 丁慷,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[J]. 环境科学学报, 2018, 38(8): 3229-3242.
- [5] 朱珈莹,安俊琳,冯悦政,等. 基于轻量级梯度提升机的南京大气臭氧浓度预测[J]. 环境科学, 2023, 44(7): 3685-3694.
- [6] 潘梦瑶,任瑛,王思源,等. 基于梯度提升算法和 SHAP 的石家庄 PM2.5 和臭氧浓度预测及影响因素分析[J]. 环境科学学报, 2024, 44(7): 402-409.

-
- [7] 申子彬, 庞乐, 王文典, 等. 基于机器学习的臭氧污染预报研究——以宁波市镇海区为例[J]. 四川环境, 2025, 44(3): 28-33.
 - [8] Dragomiretskiy, K. and Zosso, D. (2014) Variational Mode Decomposition. *IEEE Transactions on Signal Processing*, **62**, 531-544. <https://doi.org/10.1109/tsp.2013.2288675>
 - [9] Snoek, J., Larochelle, H. and Adams, R.P. (2012) Practical Bayesian Optimization of Machine Learning Algorithms. 2012 *Advances in Neural Information Processing Systems*, Nevada, 3-8 December 2012, 2951-2959.
 - [10] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
 - [11] Gers, F.A., Schmidhuber, J. and Cummins, F. (2000) Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, **12**, 2451-2471. <https://doi.org/10.1162/089976600300015015>