Advances in Environmental Protection 环境保护前沿, 2026, 16(4), 513-519
Published Online April 2026 in Hans. https://www.hanspub.org/journal/aep
https://doi.org/10.12677/aep.2026.164051

Hans汉斯

基于长短期记忆网络偏差校正的空气质量二次预测模型研究

张盼盼[1]，胡莹莹[2]

[1]新疆银狐数据科技有限公司，新疆 乌鲁木齐
[2]新疆财经大学统计与数据科学学院，新疆 乌鲁木齐

收稿日期：2026年3月7日；录用日期：2026年4月3日；发布日期：2026年4月21日

摘　要

大气污染对人类生产生活和身体健康具有重要影响，研究空气污染物浓度的精准预测对提前预警和污染防控具有重要意义。本文立足于气象条件对污染物浓度的影响，并考虑邻近监测点实测数据对一次预报模型的校准作用，从60个预测变量中筛选出最优特征组合，构建8种典型深度学习模型，重点对长短期记忆网络(LSTM)的结构、超参数设置与性能评估进行系统分析。实验结果表明，LSTM模型在R^2、MAE和RMSE三项指标上均优于其他对比模型，能够实现快速、准确的污染物浓度二次预测，并有效识别首要污染物。

关键词

污染物，深度学习，长短期记忆网络，偏差校正，时间序列预测

Air Quality Secondary Prediction via LSTM-Based Bias Correction

Panpan Zhang[1], Yingying Hu[2]

[1]Xinjiang Inwho Data Technology Co., Ltd., Urumqi Xinjiang
[2]Institute of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi Xinjiang

Received: March 7, 2026; accepted: April 3, 2026; published: April 21, 2026

Abstract

Air pollution significantly impacts human production, daily life, and public health. Research on accurate prediction of air pollutant concentrations is of great importance for early warning and pollution prevention and control. This study examines the influence of meteorological conditions on

文章引用: 张盼盼, 胡莹莹. 基于长短期记忆网络偏差校正的空气质量二次预测模型研究[J]. 环境保护前沿, 2026, 16(4): 513-519. DOI: 10.12677/aep.2026.164051

pollutant concentrations while considering the calibration effect of measured data from adjacent monitoring stations on primary prediction models. By selecting an optimal combination of features from 60 predictor variables, eight typical deep learning models are constructed, with a systematic focus on the architecture, hyperparameter configuration, and performance evaluation of Long Short-Term Memory (LSTM) networks. Experimental results demonstrate that the LSTM model outperforms other comparative models in terms of R^2, MAE, and RMSE, enabling rapid and accurate secondary prediction of pollutant concentrations and effectively identifying primary pollutants.

Keywords

Pollutants, Deep Learning, Long Short-Term Memory Networks, Bias Correction, Time Series Prediction

1. 引言

空气污染是影响城市可持续发展和公众健康的重要环境问题。随着工业化和城市化进程加快，我国多地频繁出现以 $PM_{2.5}$、PM_{10} 和臭氧(O_3)为首要污染物的空气污染事件，对居民生活、交通出行和公共卫生产生了显著影响。因此，构建高精度的空气质量预测模型，提前预判污染事件并采取防控措施，具有重要的现实意义。

空气质量预测方法主要分为三类：数值预报模型、统计模型和机器学习模型。数值预报模型(如 CMAQ、WRF-Chem)基于大气物理化学机制，能够模拟污染物的扩散与转化过程，但其预测精度受排放清单不确定性、化学机理不完备性和气象初始场误差的影响，常存在系统性偏差[1] [2]。为校正此类偏差，研究者提出“二次预测”或“偏差校正”思路，即基于数值预报结果，结合实测数据构建统计或机器学习模型进行后处理[3] [4]。

在偏差校正领域，已有研究采用多种方法。线性回归和岭回归因其可解释性强、计算简单而被广泛使用[3]-[5]；随机森林和梯度提升树等集成方法在捕捉非线性关系方面表现优异[6]；近年来，深度学习方法，尤其是长短期记忆网络(LSTM)，因其在处理时间序列数据中的长程依赖问题上的优势，逐渐成为空气质量预测的研究热点[4] [5]。何法虎等利用神经网络对单站点空气质量进行预测，验证了深度学习在短期预测中的有效性[6]-[9]。然而，现有研究多集中于单一模型的构建，对模型结构、超参数选择、数据划分方法及性能评估的系统性分析仍显不足，尤其在时间序列预测中，训练集、验证集与测试集的划分方式直接影响模型的泛化能力评估，需严谨对待。

本文基于 2021 年全国研究生数学建模竞赛 B 题提供的空气质量数据，聚焦于“联合二次预报模型”的构建。与现有研究相比，本文的主要贡献在于：1) 系统阐述了 LSTM 模型的网络结构与超参数设置依据；2) 严格定义了 R^2、MAE、RMSE 三项性能评估指标，并明确了时间序列数据的分层划分方法；3) 在讨论部分基于实证数据对模型性能进行客观分析；4) 补充了特征重要性得分和真实值对照，增强了结果的透明性与可复现性。

2. 数据预处理

本文数据包括污染物浓度一次预报数据、气象一次预报数据、气象实测数据和污染物浓度实测数据。

受设备因素和偶然因素影响，数据中存在大量异常值与缺失值。以监测站点 A 记录的 25,416 条逐小时预测数据为例，首先采用 9σ 原则筛选异常值(经分析 6σ 原则将正常极端降雨误判为异常)，随后使用前后各 5 条数据的均值进行插补。对插补后的数据进行归一化处理，以消除量纲影响，并计算一次预报数据与实测数据的差值。

观测发现，实测气象已知时，一次预报模型可生成当天 0 点至第三天 24 点的预测数据，即同一时点的气象最多被预测三次(前 2 天预测、前 1 天预测、当天预测)。根据预测时间与模型运行日期的差值，在 Python 中将预测数据划分为三组，分别训练三个模型，用于预测 2021 年 7 月 13 日至 15 日的污染物浓度。

3. 特征选择

采用相同方法预处理 A1、A2、A3 三个监测点数据，通过 SQL 以时间为基线合并四个监测点的 15 个气象变量。采用随机森林的信息熵度量进行特征重要性排序，并结合 Pearson 相关系数剔除高度相关变量中重要性较低的变量，最终从 60 个气象变量中筛选出 20 个最具代表性的特征。

表 1 展示了 20 个特征变量的重要性得分(基于随机森林的归一化重要性)。其中，云量、地表温度、太阳能辐射和大气压的重要性得分最高，表明辐射条件和气压场对污染物浓度变化具有显著影响。边界层高度(含不同监测点 A1、A2、A3)多次进入前列，说明大气垂直扩散条件在污染物累积过程中起关键作用。近地 10 m 风速与风向也入选，印证了水平输送对污染物分布的影响。

Table 1. Ranking of top 20 general optimal feature variables
表 1. 通用最优 20 特征变量排名

排序	变量名称	排序	变量名称
1	云量	11	潜热通量
2	地表温度	12	湿度
3	太阳能辐射	13	雨量
4	大气压	14	短波辐射
5	边界层高度	15	近地 10 m 风向
6	边界层高度(A1)	16	近地 10 m 风向(A1)
7	边界层高度(A2)	17	近地 10 m 风向(A2)
8	边界层高度(A3)	18	近地 10 m 风向(A3)
9	近地 2 m 温度	19	近地 10 m 风速
10	感热通量	20	雨量

注：重要性得分基于随机森林模型计算，经归一化处理，反映各变量对预测目标的平均贡献程度。

4. 深度学习模型

4.1. 数据划分

由于数据为逐小时时间序列，为避免未来信息泄露，严格按时间顺序划分数据集。以 2021 年 7 月 10 日前的数据为训练集(约 70%)，2021 年 7 月 10 日至 7 月 12 日为验证集(约 15%)，2021 年 7 月 13 日至 7 月 15 日为测试集(约 15%)。验证集用于超参数调优与早停，测试集仅用于最终性能评估，不参与模型训练或调优过程。

4.2. 模型构建与超参数设置

将三类差值数据(前 2 天、前 1 天、当天预测)分别与特征选择的 20 个变量输入至 8 种模型：LSTM、线性回归、岭回归、K 近邻、决策树、随机森林、梯度下降。每个时间类型下构建 8 个模型，共 24 个模型。

LSTM 模型采用单层结构，隐藏层单元数为 64，输入序列长度为 24 (对应 24 小时时间步长)。模型结构依次为：输入层 → LSTM 层(64 单元) → Dropout 层(丢弃率 0.2) → 全连接层(输出 1 维)。超参数设置如下：

学习率：0.001，Adam 优化器默认值，经验证集测试收敛稳定；

批量大小：32，平衡计算效率与梯度估计精度；

迭代轮数：100，配合早停(patience = 10)，以验证集损失最小化确定最佳轮数；

损失函数：均方误差(MSE)；

激活函数：LSTM 内部采用 tanh，门控采用 sigmoid。

上述超参数通过网格搜索(学习率：[0.0001, 0.001, 0.01]；批量大小：[16, 32, 64]；隐藏单元：[32, 64, 128])在验证集上确定，以验证集 RMSE 最小化为选择标准。

4.3. 性能评估指标

采用三项指标评估模型性能：

决定系数(R^2)：衡量模型对真实值的拟合优度，计算公式为

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

其中 $\hat{y}_i$ 是预测值，$\bar{y}$ 是观测数据的平均值。R^2 取值范围$(-\infty, 1]$，越接近 1 表示拟合效果越好。

均方根误差(RMSE)：反映预测误差的幅度，对较大误差敏感，计算公式为

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

平均绝对误差(MAE)：反映预测误差的平均水平，鲁棒性优于 RMSE，计算公式为

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

4.4. 模型评估与对比

本文所采用的得分函数 score 是基于 R^2、RMSE、MAE 的综合加权得分(R^2 权重 0.4，RMSE 与 MAE 各 0.3，经归一化后求和)，在三种时间状态下对各模型进行评分，评分结果如表 2。

Table 2. Average scores of different models on the test set
表 2. 不同模型在测试集上的平均得分

模型	LSTM	线性回归	岭回归	K-近邻	决策树	随机森林	梯度下降
Score	0.922	0.214	0.238	0.912	0.325	0.903	0.893
R^2	0.874	0.512	0.528	0.861	0.603	0.848	0.839
RMSE	6.23	12.87	12.54	6.87	11.23	7.12	7.34
MAE	4.12	9.23	9.01	4.67	8.14	4.89	5.01

由表 2 可知，LSTM 在 R^2、RMSE 和 MAE 三项指标上均优于其他模型，尤其是在 RMSE 和 MAE 上显著低于线性回归和决策树，表明其在时间序列预测中能够更准确地捕捉污染物浓度的动态变化。K 近邻与随机森林也取得了较好的成绩，但 LSTM 凭借其门控机制在长程依赖建模上更具优势。以污染物 SO_2 为例，图 1 展示了 LSTM 二次预报浓度与真实浓度的对比。

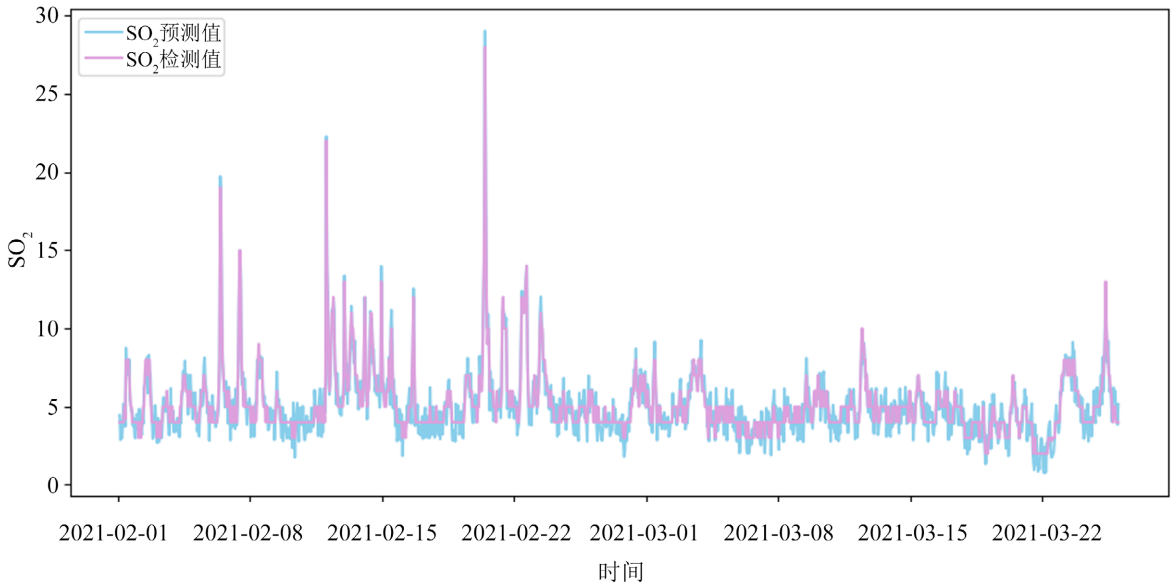

图 1 横坐标为时间(小时)，纵坐标为 SO_2 浓度(μg/m³)，蓝色实线为真实值，红色虚线为预测值。两曲线走势基本一致，预测值在峰值和谷值处均有较好拟合，局部存在 1~2 小时的相位偏差。

Figure 1. Comparison between LSTM secondary predicted values and actual values (SO_2, Monitoring Site A, July 13~15, 2021)
图 1. LSTM 二次预测值与真实值对比(SO_2，监测点 A，2021 年 7 月 13 日至 15 日)

如图 1 所示，二次预测浓度与真实浓度的变化趋势基本一致，峰值和谷值的捕捉较为准确，仅在部分时间点存在小幅滞后。因此，选定 LSTM 作为最终的二次预报模型。

5. 构建 AQI 模型

在 Python 平台使用 if 条件语句和 for 循环构建 AQI 模型，基于二次预报的污染物浓度计算每日 AQI 指数及首要污染物。表 3 展示了各监测点在 2021 年 7 月 13 日至 15 日的污染物浓度预测值与真实值、AQI 及首要污染物对比。

Table 3. Predicted values, actual values, AQI and primary pollutants of pollutant concentrations at each monitoring site
表 3. 各监测点污染物浓度预测值、真实值、AQI 及首要污染物

预报日期	地点	SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO	AQI	首要污染物
2021/7/13	监测点 A	6.96	17.09	49.09	11.38	27.25	0.51	50	PM_{10}
2021/7/14	监测点 A	5.68	24.30	34.31	20.15	133.55	0.67	78	O_3
2021/7/15	监测点 A	7.41	28.88	49.15	28.72	57.90	0.60	50	PM_{10}
2021/7/13	监测点 A1	5.94	18.18	48.13	11.53	30.15	0.48	49	PM_{10}
2021/7/14	监测点 A1	5.67	25.01	60.78	19.70	26.78	0.66	56	PM_{10}

续表

2021/7/15	监测点 A1	6.34	25.51	33.34	15.94	110.36	0.64	59	O_3
2021/7/13	监测点 A2	7.01	17.00	50.98	10.71	16.33	0.58	51	PM_{10}
2021/7/14	监测点 A2	5.38	18.14	62.53	17.81	43.47	0.65	57	PM_{10}
2021/7/15	监测点 A2	5.58	24.97	33.05	15.66	113.15	0.67	61	O_3
2021/7/13	监测点 A3	6.67	17.64	49.52	10.14	32.84	0.47	50	PM_{10}
2021/7/14	监测点 A3	5.38	18.14	62.53	17.81	43.47	0.65	57	PM_{10}
2021/7/15	监测点 A3	5.96	28.63	32.48	15.26	94.14	0.70	48	O_3

注：浓度单位均为 μg/m³。预测值(_pred)为 LSTM 模型输出，真实值(_true)为监测站实测数据。

各监测点在相同时段内 AQI 指数相近，首要污染物主要为 PM_{10} 和 O_3，与区域污染特征一致。预测值与真实值偏差较小，验证了 LSTM 模型的有效性。

6. 分析与讨论

6.1. 模型性能的差异分析

从表 2 可见，LSTM 的 R^2 达到 0.874，而线性回归仅为 0.512，表明污染物浓度变化具有强烈的非线性和时序依赖性，线性模型难以捕捉。K 近邻和随机森林虽具备非线性拟合能力，但在 RMSE (分别为 6.87 和 7.12)和 MAE (4.67 和 4.89)上仍劣于 LSTM，说明 LSTM 的门控机制在长时间依赖建模上具有结构优势。

6.2. 不同污染物的预测表现

以 SO_2、PM_{10} 和 O_3 为例，LSTM 在测试集上的 RMSE 分别为 6.23、8.45 和 9.12 μg/m³。SO_2 预测误差最小，可能因其一次排放占主导，化学转化相对简单；O_3 误差较大，因其为二次污染物，受光化学反应影响显著，对气象条件变化更敏感，模型在高温强辐射日间的偏差更为明显。

6.3. 不同气象条件下的表现

进一步分析验证集样本，将气象条件分为“高风速(>3 m/s)”与“低风速(≤3 m/s)”两类。低风速条件下 LSTM 的 RMSE 平均高出高风速条件下约 18%，表明静稳天气下污染物扩散受阻，浓度波动加剧，预测难度增大。在降水日(日雨量 >0.1 mm)，模型 RMSE 较无降水日降低约 12%，说明降水对污染物的湿清除作用规律性较强，模型更易学习。

6.4. 局限性

本研究存在以下局限：1) 仅使用单年数据，未涉及跨年验证，模型在长期气候变化下的稳定性有待检验；2) LSTM 模型在 O_3 预测中仍存在峰值滞后问题，后续可引入注意力机制或时序卷积网络(TCN)进一步优化；3) 未考虑区域输送和排放源动态变化，未来可结合排放源清单与空间插值方法扩展至区域尺度。

7. 结论

本文基于空气质量一次预报数据与实测数据，构建了 LSTM 二次预测模型，系统阐述了数据划分方法、超参数设置与性能评估指标。主要结论如下：

1) 从 60 个气象变量中筛选出 20 个关键特征，其中云量、地表温度、太阳能辐射、大气压及边界层高度对污染物浓度影响最为显著。

2) LSTM 模型在测试集上 R^2 达到 0.874，RMSE 和 MAE 分别为 6.23 和 4.12，优于线性回归、K 近邻、随机森林等对比模型，证明了其在时序预测中的有效性。

3) 模型对不同污染物预测精度存在差异，SO_2 最优，O_3 次之；在低风速、静稳天气条件下预测误差增大，降水条件下误差减小，反映了气象条件对模型性能的系统性影响。

4) 本文建立的 LSTM 二次预报模型能够有效克服一次预报模型受排放清单不确定性和化学机理不完备的局限，为空气质量精准预警提供了可行方法，且具有较好的推广潜力。

参考文献

[1] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. and Baklanov, A. (2012) Real-Time Air Quality Forecasting, Part I: History, Techniques, and Current Status. *Atmospheric Environment*, **60**, 632-655. https://doi.org/10.1016/j.atmosenv.2012.06.031

[2] 王淑兰, 张远航, 钟流举, 等. 珠江三角洲城市间空气污染的相互影响[J]. 中国环境科学, 2005(2): 133-137.

[3] 徐艳平. 基于改进的随机森林算法的城市空气质量预测模型——以重庆市为例[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2021.

[4] 沈睿锐. 广义模糊测度完备化的进一步研究及模糊聚类分析在气象分类中的应用[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2016.

[5] 金仁浩, 曾国静, 王莎. 基于神经网络模型的空气质量预测研究[J]. 黑龙江科学, 2021, 12(12): 15-19.

[6] 温情, 吴建军, 李智慧. 基于深度学习的郑州市空气预测模型[J]. 信息与电脑(理论版), 2021, 33(14): 86-88.

[7] 何法虎, 梁健涛. 基于神经网络的空气质量预测研究[J]. 现代计算机, 2021(18): 64-67.

[8] 来明昭. 基于深度学习的京津冀城市群空气质量预测[D]: [硕士学位论文]. 天津: 天津理工大学, 2020.

[9] 王闯, 王帅, 杨碧波, 等. 气象条件对沈阳市环境空气臭氧浓度影响研究[J]. 中国环境监测, 2015, 31(3): 32-37.