Statistics and Application 统计学与应用, 2025, 14(11), 299-311
Published Online November 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1411331

Hans汉斯

基于LSTM模型的空气质量指数预测

王嘉敏[1]，张亦舒[1]，苏旭颖[2]

[1]北方工业大学理学院，北京
[2]北方工业大学人工智能与计算机学院，北京

收稿日期：2025年10月22日；录用日期：2025年11月13日；发布日期：2025年11月24日

摘　要

随着经济的飞速发展和人类活动的日益频繁，环境问题逐渐成为全球关注的焦点，其中空气质量问题尤为突出。空气中污染物的种类和浓度不断增加，对人类健康和生态环境造成了严重威胁。本文基于LSTM模型来预测未来空气质量，使用处理后的数据对模型进行训练，确保模型的预测性能满足要求。准确计算和预测空气质量等级，能够及时向公众发布空气质量信息，引导公众采取有效的防护措施，减少污染物对健康的危害。此外，为环境管理部门制定科学合理的环境政策提供依据，有助于推动空气质量的改善。

关键词

长短期记忆模型，AQI，空气质量

Air Quality Index Prediction Based on LSTM Models

Jiamin Wang[1], Yishu Zhang[1], Xuying Su[2]

[1]School of Science, North China University of Technology, Beijing
[2]School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing

Received: October 22, 2025; accepted: November 13, 2025; published: November 24, 2025

Abstract

With rapid economic development and increasingly frequent human activities, environmental issues have gradually become a global focus, with air quality problems being particularly prominent. The types and concentrations of air pollutants continue to increase, posing a serious threat to human health and the ecological environment. This paper employs an LSTM model to forecast future air quality, training the model using processed data to ensure its predictive performance meets

文章引用: 王嘉敏, 张亦舒, 苏旭颖. 基于 LSTM 模型的空气质量指数预测[J]. 统计学与应用, 2025, 14(11): 299-311.
DOI: 10.12677/sa.2025.1411331

requirements. Accurate calculation and prediction of air quality levels enable timely dissemination of air quality information to the public, guiding individuals to take effective protective measures and reducing the health hazards posed by pollutants. Furthermore, it provides a basis for environmental management departments to formulate scientifically sound environmental policies, thereby contributing to the improvement of air quality.

Keywords

Long Short-Term Memory Model, AQI, Air Quality

1. 引言

随着全球工业化和城市化进程的迅猛推进，人类活动对大气环境产生了深远影响。工业活动释放大量污染物、汽车和飞机等交通工具的废气排放、家庭和商业活动中使用的煤炭和燃油释放的有害物质都是造成大气污染的主要来源。某些物质进入大气中，呈现出足够的浓度，达到足够的时间，并因此危害了人体的舒适、健康和福利或危害了生态环境。为此我国专门成立了城市空气质量检测部门，预测城市空气质量，保障人们安全的生活环境[1]。根据《环境空气质量指数(AQI)技术规定(试行)》空气质量指数(AQI)可用于判别空气质量等级，是反映大气环境质量水平的重要指标[2]。根据 AQI 值将空气质量划分为六个等级，不同空气质量等级对应着不同的健康影响和建议措施。此外，准确计算和预测空气质量等级，有助于公众及时了解空气质量状况，采取相应的防护措施，保护自身健康；同时也为政府环境管理部门制定科学合理的污染防控策略、优化产业布局、加强环境监管提供决策支持。

在空气质量预测研究领域，传统方法主要包括时间序列模型(如 ARIMA)和统计学习方法[3]。近年来，随着深度学习技术的发展，循环神经网络(RNN)及其变体长短期记忆网络(LSTM)在时间序列预测中展现出显著优势，能够有效捕捉数据中的长期依赖关系[4]。然而，标准 LSTM 模型在空气质量预测中的应用效果，特别是在融合季节性特征方面的系统优化，仍有待深入评估。本研究旨在系统评估 LSTM 模型在特定区域空气质量预测中的表现，重点探索季节性特征的引入对六种主要污染物预测精度的影响，以期为精准化空气质量预报提供一个优化的技术案例。

2. 模型假设

为了确保模型在预测任务中的有效性与适用性，我们做了以下假设。

六种污染物浓度序列具有时间依赖性，当前污染水平在一定程度上依赖于过去若干天的历史情况。

污染物的过去趋势可以用来预测未来趋势，未来污染变化可以由历史序列学习获得。

3. 模型的建立

3.1. 相关理论和技术

LATM 神经网络最早由 Hochreiter 和 Schmidhuber 提出[5]，它能够有效克服 RNN (recurrentneural network)中存在的梯度消失问题，使网络能够有效地处理长期时间序列数列[6]，并且在预测上所得误差明显低于其他方法。本文针对 6 种常规污染物的单日浓度值数据，提出了基于 LATM 魂环神经网络的预测

方法，包括输入层、隐藏层和输出层网络解构，并使用一种基于多层网格搜索的 LSTM 预测模型参数优选算法。

LSTM 是在 RNN 基础上进一步发展形成，包括四个核心部分：遗忘门、输入门、细胞状态更新和输出门(图 1)。其中遗忘门、输入门和输出门新增逻辑控制单元，通过这三个门控单元来控制信息的流动和更新，从而有效地捕捉和及异常序列中的关键信息。

遗忘门：

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

输入门：

$$i_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

单元：

$$\tilde{c}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \tag{4}$$

输出门：

$$o_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

最终输出：

$$h_t = o_t \tanh\left(c_t\right) \tag{6}$$

分别使用 i、f 和 o 来表示输入、遗忘和输出门，W 和 b 表示网络的权重矩阵和偏置向量。遗忘门决定前一时刻的细胞状态有多少信息需要以往；输入门决定当前输入信息有多少需要更新到细胞状态中；输出门决定当前时刻的细胞装填有多少信息需要输出到隐藏状态中。

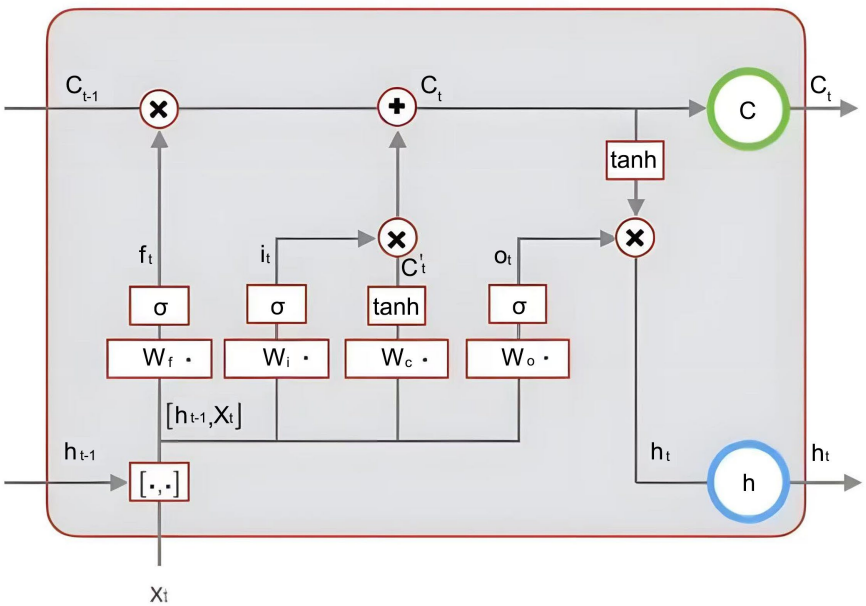

Figure 1. LSTM unit structure
图 1. LSTM 单元结构

3.2. 数据预处理

数据的可靠性分析，其可靠性分析包括检验数据的完整性和准确性。

数据的完整性指数据中是否存在断电和缺口，并尝试从其他监测站点、历史数据或相关研究中补充数据。对于附件并不存在不连续的时间，故附件数据具有完整性。

数据的准确性是指数据中是否曾在缺失值和异常数据。对于缺失值，我们在对附件遍历时并未发现缺失值；对于异常值，我们采用 3σ 原则，通过计算均值和标准差确定正常范围的上限和下限(图 2)。遍历数据集，识别任何落在正常之外数据点，起被视为异常值。之后我们采用上下数据的均值代替异常值，我们发现附件中的数据存在异常值。因此，附件数据不具有准确性，需要对数据进行清醒，将异常值替换后可用于后续探究分析。

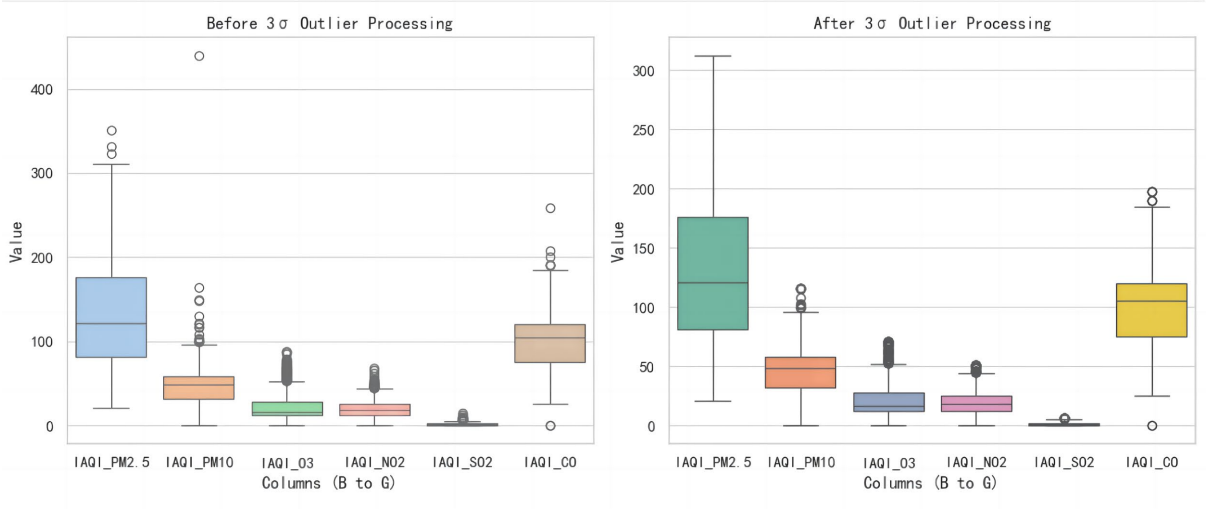

Figure 2. Data processing comparison chart
图 2. 数据预处理对比图

3.3. LSTM 模型的构建

选取 6 种污染物单日浓度值按照时间顺序由远及近才形成一个时间序列，将搜集到数据中 2019 年 4 月 16 日到 2022 年月 5 日 10 期间的各污染物浓度数据依次排列。以上数据导入 MATLAB 后使用 figure 函数对该时间序列构图，形成一个时间相关的图形。输出各污染物浓度值时间序列图。

从图中可以清晰度的观察到，污染物在整个时间段内呈现出明显的波动特征。浓度值并非保持稳定，而是在不同的时间点有较大的起伏，说明空气质量在不同时间存在明显差异，具有较大的不稳定性。因此，为实现对污染物浓度的有效预测，我们引入 LSTM 网络模型，构建 LSTM 污染浓度预测模型。LSTM 是一种特殊的循环神经网络(RNN)，它通过精心设计的门控机制，能够有效地捕捉时间序列数据中的长期依赖关系。构建的 LSTM 网络结构图如图 3 所示。

由于 LSTM 模型对输入数据的尺度比较敏感，需要对单日浓度数据进行归一化处理，将数据线性映射至 0~1 范围内。隐藏层采用 LSTM 细胞和 Dropout 搭建双层循环神经网络(图 4)。由于 LSTM 神经网络模块的层数越多，其学习能力越强，但是层数过多又会造成网络训练难以收敛，因此训练过程中网络的层数一般不超过 3 [7]，本文采用两层。隐藏层采用多步预测法，输出层使用了全连接层对结果进行降低维度，并将得到预测数据后进行了反归一化，最终得到预测结果。

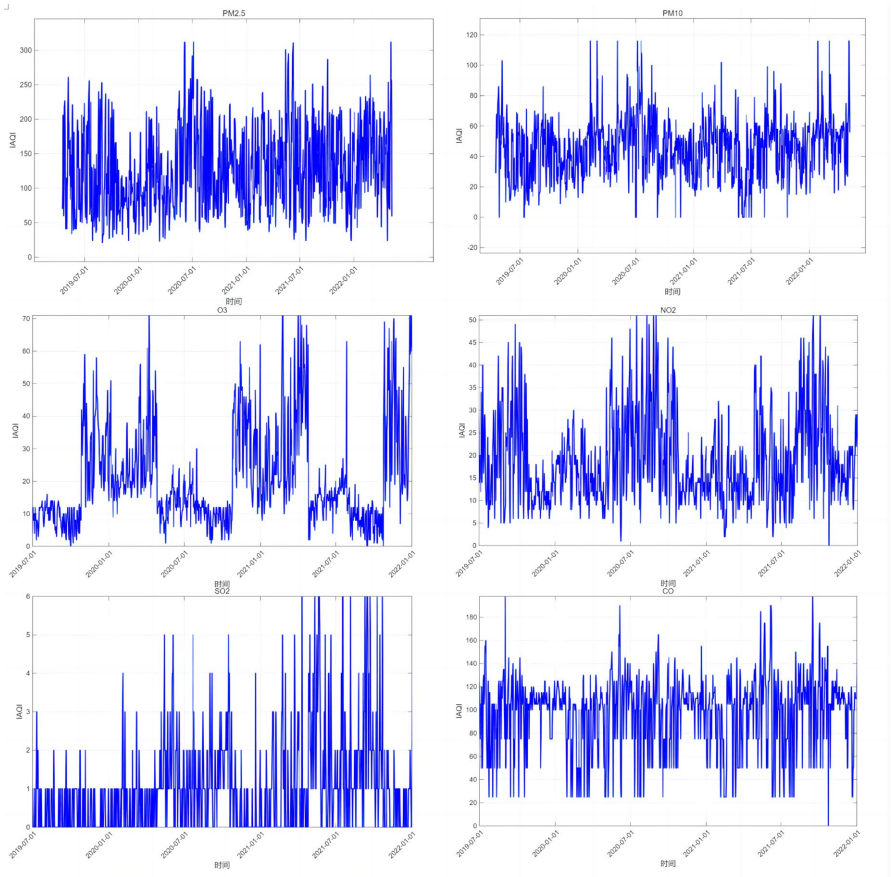

Figure 3. Time series charts of concentrations of six pollutants
图 3. 6 种污染物浓度时间序列图

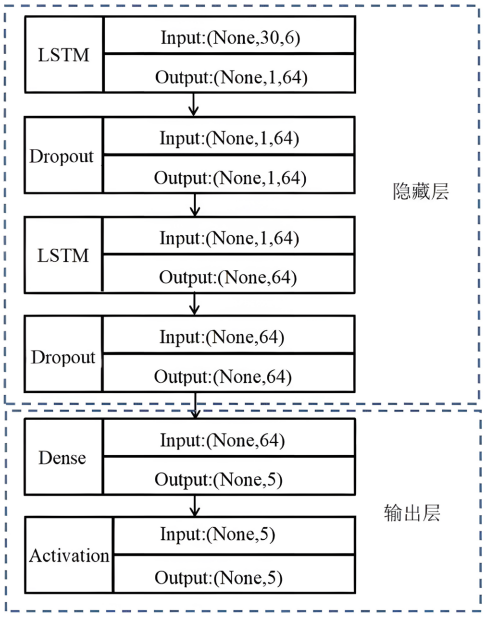

Figure 4. LSTM network structure for pollutant concentration
图 4. 污染物浓度的 LSTM 网络结构

本文目标是预测 6 种污染物单日浓度值故选取均方误差(Mean Square Error, MSE)作为损失函数

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_1\right)^2 \tag{7}$$

通过计算预测值与真实值之间的差值的平方的平均值，能够直观的量化模型预测的准确程度。由于污染物浓度在不同时间存在明显波动，MSE 对这种波动具有较高的敏感性。较大的预测误差会导致 MSE 值显著增加，这能够及时反映出模型在处理污染浓度预测的性能，进而针对性的改进模型。下面 6 种污染物损失函数曲线图(图 5~10)。

其次，选取 tanh 函数作为激活函数

$$\tanh\left(x\right)=\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}} \tag{8}$$

优化器采用 Adam 优化器(Adaptive Moment Estimation，适应性矩估计)进行优化训练。Adam 优化器由 Kingma 和 Ba (2015)提出[8]，它结合了 AdaGrad 和 RMSProp 优化器的优点，能够高效地处理大规模数据和复杂模型，并在各种深度学习任务中取得良好的效果。

Figure 5. O_3 loss function curve
图 5. O_3 损失函数曲线图

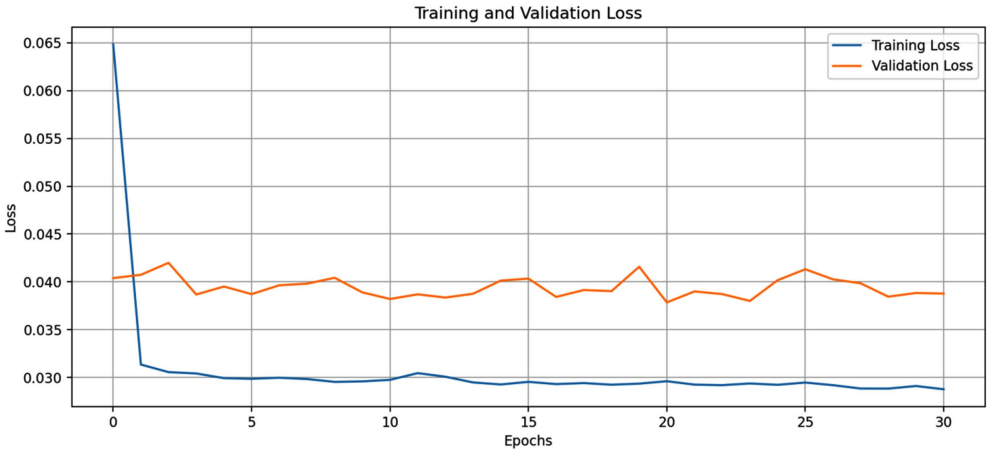

Figure 6. $PM_{2.5}$ loss function curve
图 6. $PM_{2.5}$ 损失函数曲线图

Figure 7. NO_2 loss function curve
图 7. NO_2 损失函数曲线图

Figure 8. SO_2 loss function curve
图 8. SO_2 损失函数曲线图

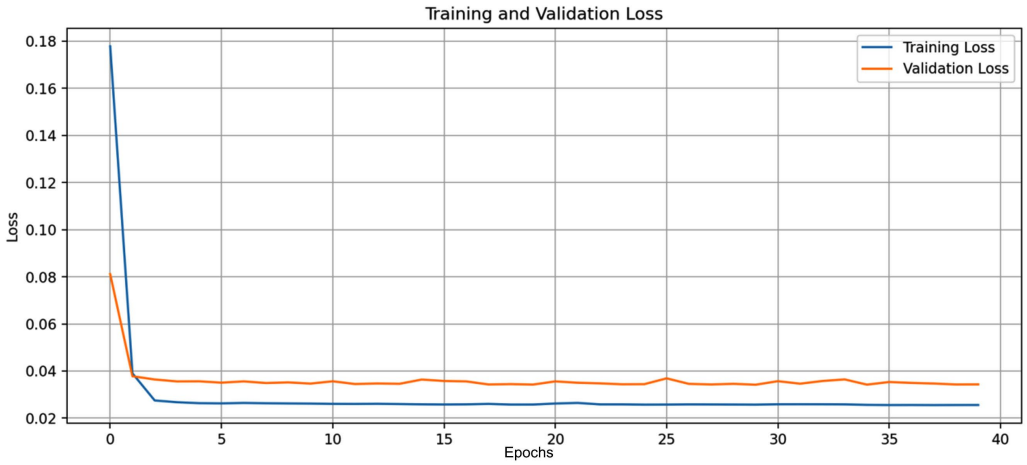

Figure 9. CO loss function curve
图 9. CO 损失函数曲线图

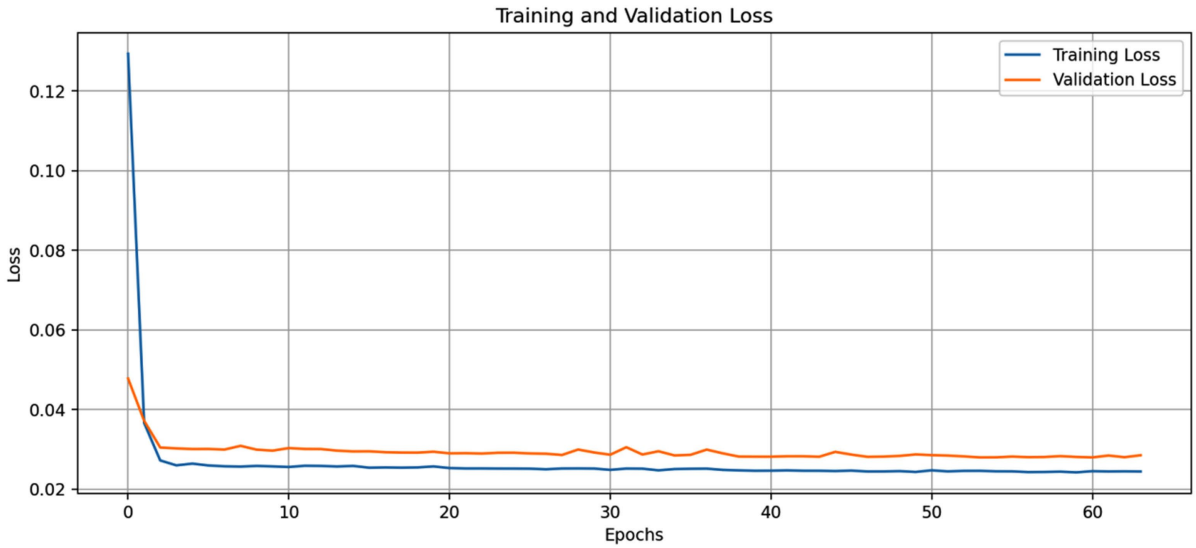

Figure 10. PM_{10} loss function curve
图 10. PM_{10} 损失函数曲线图

应用 LSTM 神经网络对每个污染物预测思路为：用过去 30 天的浓度值数据信息来对未来 5 天各污染物浓度值进行预测。因为大气污染物具有时间序列特性，通常对于未来 5 天浓度值，使用前天 30 的浓度数据进行预测且 30 天的数据包含足够信息。所有污染物的 LSTM 神经网络输出均是未来 5 天的各污染物的预测值。在参数选取上指定 LSTM 神经网络单元有 64 个隐藏单元，并进行 100 论的训练，用于捕捉输入序列中的长期依赖关系；设初始学习率为 0.001，Dropout 率为 0.2 用于防止过拟合，再经过 100 论训练。设置以上参数进行性调试，以确保模型能够准确的捕捉污染物浓度变化规律，提高预测的准确性。

3.4. 实验评估

本文针对 6 种污染物浓度值进行预测，因此构建相对误差指标进行测试集预测效果评估。预测精度评估采用平均绝对误差(Mean Absolute Error, MAE)，MAE 值越低表示模型的而预测值更接近实际值，模型的预测性能更好；均方根误差(Root Mean Square Error, RMSE)，RMSE 能衡量预测值与真实值之间的平均误差幅度，反应不同污染物对空气质量的误差度，其值越小表示模型的预测精度越高。MAE 提供了预测误差的平均水平，而 RMSE 则强调了大误差的影响。通过二者的指标能更的判断模型的拟合效果。MAE、RMSE 的及计算分别如下所示：

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{10}$$

根据实验，对各个污染物进行了测试数据集与模型预测数据对比(如图 11)。可以观察到，部分区域拟合效果较好，担任有部分预测值与实际数据相差较大。根据 6 种污染物 RMSE、MAE 值(表 1)，可以看出 NO_2、SO_2 的 RMSE 和 MAE 值均较低，与真实值相比误差较小，LSTM 模型对两者预测精度看较高；CO、PM_{10}、O_3 的 RMSE 和 MAE 的值相对较大；$PM_{2.5}$ 的实验值相比较下误差较大，预测精度较差。由此可见，该模型的预测结果较为良好。

Figure 11. Fitting chart of 6 pollutants
图 11. 6 种污染物拟合图

Table 1. RMSE and MAE values of 6 pollutants
表 1. 6 种污染物 RMSE、MAE 值

	$PM_{2.5}$	NO_2	SO_2	CO	PM_{10}	O_3
RMSE	33.58	5.31	1.07	16.73	12.02	13.26
MAE	26.49	3.87	0.81	13.51	9.06	9.9

3.5. 模型优化

从上述实验评估分析中，可以看到模型的拟合程度还有待提高，预测精度略有偏差。跟据时间序列图发现，污染物浓度存在季节性变化。如臭氧的时间序列图(图 12)中，可以看出在臭氧在冬季节浓度值整张幅度变大，在春末到初秋季节浓度相对较低。由此得出，污染物浓度呈现周期性季节性变化。因此，考虑将季节特征列入对 LSTM 模型的影响因素中。

(1) 季节对污染物的影响

季节因素包括当季温度、湿度、风速和风向等对污染物的影响。对于温度，夏季温度高，大气对流活动相对旺盛，有利于污染物的扩散和稀释，对空气质量有哦一定的改善作用；冬季温度低，大气层相对稳定，会抑制污染物的垂直扩散，导致污染物浓度升高。对于湿度，夏季湿度较大有利于气态污染物(如：二氧化硫、氮氧化物等)的湿沉降；冬季湿度较低，不利于污染物的湿沉降，浓度容易积累升高。由

此可见，冬季容易是污染物积累，会导致浓度升高；夏季有利于污染物的扩散和稀释。

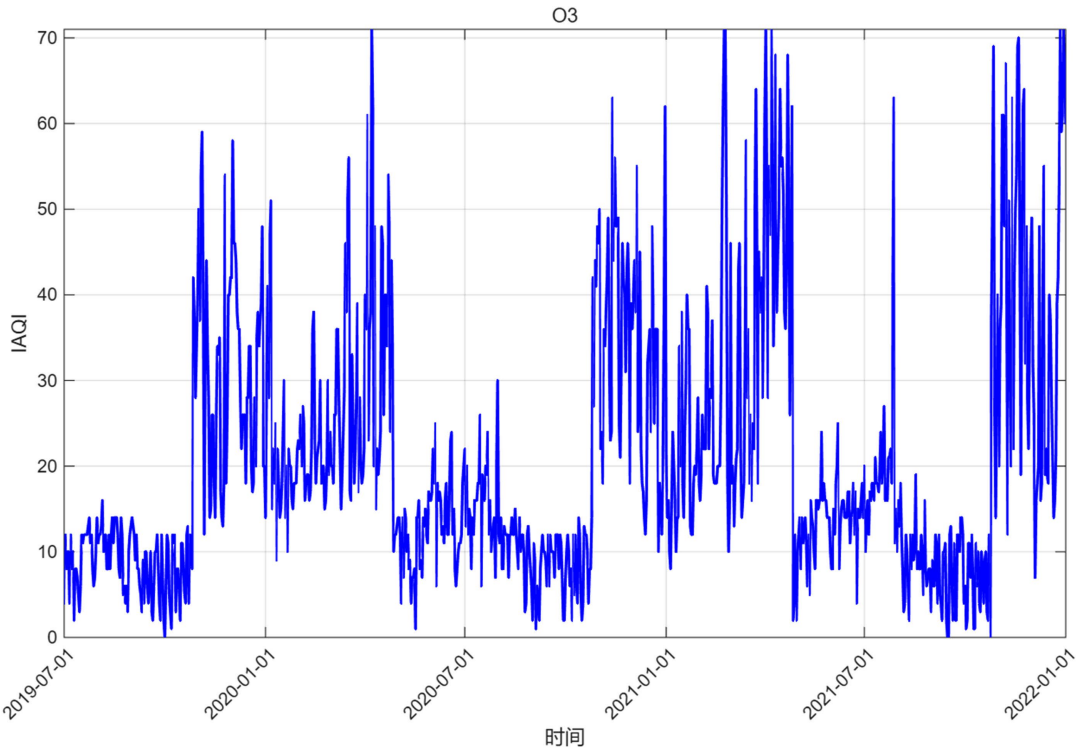

Figure 12. Ozone time series chart
图 12. 臭氧时间序列图

(2) 数据处理

考虑季节对模型的影响，首先需要对污染物浓度数据的时间进行的提取。空气污染具有明显的周期性，因此从数据的时间字段中提取月和星期，并进一步通过正余弦变换引入周期编码，以增强模型对循环变化的识别。

其次，由问题一计算得到各个污染物的 IAQI。空气污染指数通常存在较大的波动，模型易受极端值干扰。为加快网络收敛速度并防止梯度爆炸问题，采用 Min-Max 归一化方法污染物主变量进行缩放落于区间[0, 1]内。

(3) 参数优化

本文采用超参数优化方法，使用 Keras Tuner 工具进行超参数自动搜索，结合随机搜索算法(Random Search)实现优化。Keras Tuner 是一个用于深度学习模型超参数搜索的高效工具。与传统的网格搜索(Grid Search)相比，随机搜索可以在相同计算预算下更可能找到更优的超参数组合。

Keras Tuner 在每一次试验中从超参数搜索空间中采样一组超参数组合，构建模型并训练若干轮，通过验证集误差评估该组合效果，最终记录最优结果。最终，调参过程返回使验证误差最小的超参数组合。该组合用于构建最终预测模型，并在测试集上进行误差评估

(4) LSTM 模型构建

LSTM 神经网络模块的层数为 2 层，采用滑动窗口方法构造输入输出模型。LSTM 层的输入为过去 30 天的特征序列，输出捕捉时间依赖特征；全连接输出层为 LSTM 输出映射到未来 5 天的预测值。并选取均方误差(Mean Square Error, MSE)作为损失函数。构建的 LSTM 网络结构图如图 13 所示。

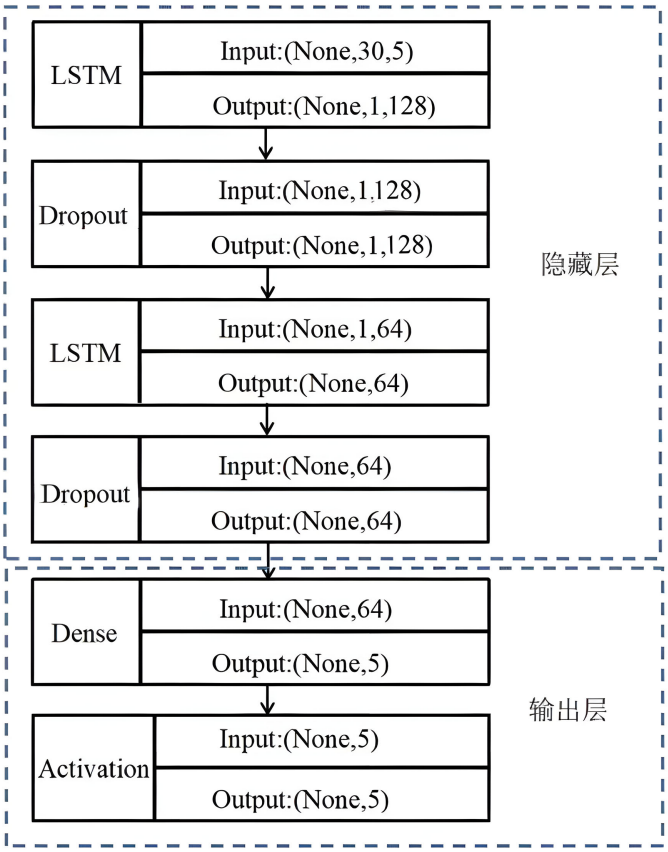

Figure 13. LSTM network structure diagram
图 13. LSTM 网络结构图

(5) 模型评估

对于改进后的模型，仍采用相对误差指标进行测试集预测效果评估。预测精度评估采用平均绝对误差(Mean Absolute Error, MAE)和均方根误差(Root Mean Square Error, RMSE)对数据进行精度预测。MAE 与 RMSE 值越低表示模型的而预测值更接近实际值，模型的预测性能更好。

通过对 6 种污染物展开全面测试，并将测试数据集与模型预测数据进行对比，从图 14 中可以直观且清晰地观察到，预测数据集曲线几乎与实际测试数据集曲线达到完全重合程度。这一对比表明当前模型在污染物浓度预测方面随具备的卓越性。加入季节特征因素的分析后，与优化前模型对比(图 11)，模型能够准确地捕捉到污染物浓度随随时间的变化趋势，无论是指数的峰值、谷值还是波动情况，6 种污染物的预测精度都大大提高。此外，根据 MAE 和 RMSE 值与优化前模型数据对比大部分污染物的值有所降低(如：$PM_{2.5}$、O_3、SO_2)。尽管有下数据 RMSE 和 MAE 值变化不大甚至有所提高(表 2)，如：CO、NO_2，表明该污染物受季节影响变化不大。

Table 2. RMSE and MAE values of 6 pollutants
表 2. 6 种污染物 RMSE、MAE 值

	$PM_{2.5}$	NO_2	SO_2	CO	PM_{10}	O_3
RMSE	28.65	5.68	0.65	23.67	12.57	6.5
MAE	22.82	4.3	0.46	18.08	9.25	4.56

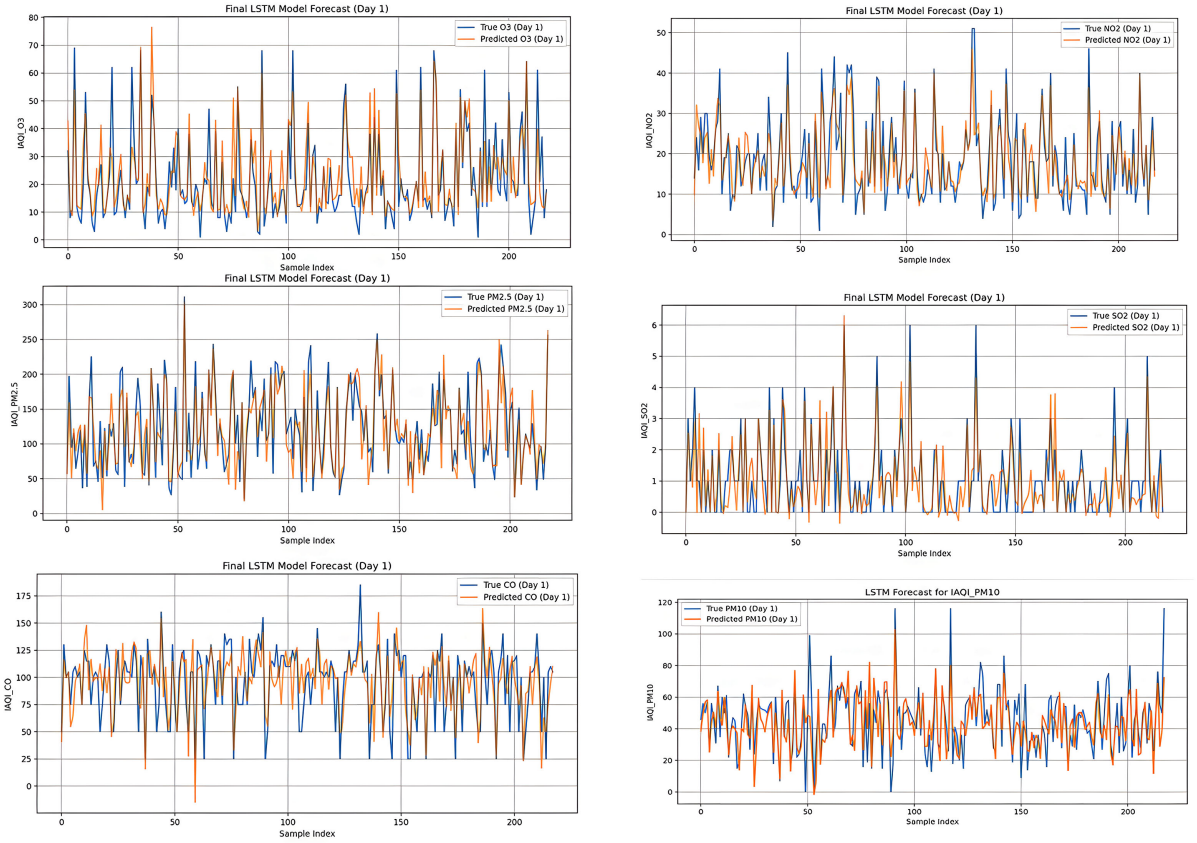

Figure 14. Fitting chart of 6 pollutants
图 14. 6 种污染物拟合图

3.6. 实验结果

采用 LSTM 神经网络模型对 6 种污染物浓度和空气质量指数进行预测。通过时间序列图发现污染物浓度的变化与季节性变化有着密切的关系。考虑季节因素的影响，我们对模型进行了进一步的优化，通过绘制原始数据与预测数据集拟合图发现数据重合成都极高，表明模型的预测值较为精确具有现实意义，并具备对未来一段时期内空气质量指数预报能力，能够为提前发现并有效防止空气污染的治理措施提供精准、可靠的数据支撑。

4. 结论与展望

4.1. 研究结论

本研究系统评估了 LSTM 模型在六种主要污染物浓度预测中的应用效果，并通过引入季节性特征对模型进行了优化，得出以下结论。

首先，对于模型适用性 LSTM 模型在空气质量预测中整体表现良好，相比其他模型，在大多数污染物上展现出更高的预测精度，证明了其在处理复杂时间序列数据方面的优势。其次，季节性影响差异其季节性特征对不同污染物的预测效果影响显著：对 O_3、$PM_{2.5}$ 和 SO_2 预测改善明显，优化后 MAE 分别下降 53.9%、13.8%和 43.2%，对 NO_2 和 CO 预测改善有限甚至下降，表明这些污染物受非季节性因素影响更大。此外，优化后的 LSTM 模型能够有效捕捉污染物浓度的长期趋势和季节性波动，在峰值和谷值预测方面表现稳定，为空气质量预警提供了可靠技术支撑。但该模型仍有不足之处，模型对受突发排放事

件和局部人为活动影响较大的污染物(如 NO_2、CO)预测精度仍有提升空间；同时存在轻微过拟合现象，需进一步优化正则化策略。

4.2. 未来展望

针对本文的研究成果与局限性，为进一步提升模型的预测性能、实用性和可持续性，可从以下几个方面进行优化。

首先，可收集多源数据并将其融合，整合实时气象数据(如风速、风向、温度、湿度、气压、降水量)、地理信息数据(如海拔、土地利用类型)以及人类活动数据等。构建一个多变量的 LSTM 或 Transformer 模型，以更全面地捕捉影响污染物扩散与生成的外部驱动因素。

其次，对于不确定性量化可采用分位数回归、蒙特卡洛 Dropout 或贝叶斯神经网络等方法，为预测结果提供置信区间，而不仅仅是点估计。还可利用 SHAP (SHapley Additive exPlanations)、LIME (Local Interpretable Model-agnostic Explanations)等工具分析模型决策的原因，识别影响污染物预测的关键历史时间点和关键特征，增强模型的透明度和可信度，为环境治理提供更深入的洞见。

最后，将优化后的模型部署为可扩展的云端服务，构建一个集数据采集、模型推理、结果可视化于一体的实时空气质量预测与预警系统。通过 Web 端或移动端应用向公众发布未来数天的空气质量预报和健康建议，实现研究成果的社会价值转化。

参考文献

[1] 李明. 科技发展引领未来[N]. 光明日报, 2024-10-01(1).

[2] 夏起铁. 基于机器学习技术的城市空气质量预测研究[J]. 信息记录材料, 2020, 21(12): 89-90.

[3] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., *et al*. (2015) Time Series Analysis: Forecasting and Control. John Wiley & Sons.

[4] Sutskever, I., Vinyals, O. and Le, Q.V. (2014) Sequence to Sequence Learning with Neural Networks. *NIPS*'14: *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Vol. 2, 3104-3112.

[5] 李小飞, 张明军, 王圣杰, 等. 中国空气污染指数变化特征及影响因素分析[J]. 环境科学, 2012, 33(6): 1936-1943.

[6] 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

[7] 王鑫, 吴际, 刘超, 等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报, 2018, 44(4): 772-784.

[8] 陈振宇, 刘金波, 李晨, 等. 基于 LSTM 与 XGBoost 组合模型的超短期电力负荷预测[J]. 电网技术, 2020, 44(2): 614-620.