

基于注意力机制的多尺度时序预测模型及其在空气质量预报中的应用

A Multi-Scale Time-Series Forecasting Model Based on Attention Mechanism and Its Application in Air Quality Prediction

孙佳婕

(河池学院, 广西 河池 546300)

摘要: 城市空气质量预报是大气污染精准防控与公众健康风险管理的重要基础。受排放源、气象条件与大气化学过程的多重耦合影响, 污染物浓度时间序列普遍具有强非线性、非平稳性与多尺度特征: 小时尺度上的短期波动、日尺度上的交通与生活排放节律、周尺度上的生产活动周期以及季节尺度上的气象演变相互叠加, 使精确预报成为一项具有挑战性的任务。传统统计模型难以刻画上述非线性关系, 单一结构的深度学习模型则难以同时捕捉不同时间尺度的时序特征。针对这一问题, 本文提出一种基于注意力机制的多尺度时序预测模型 (multi-scale attention LSTM, MSA-LSTM)。该模型首先通过多尺度滑动窗口将输入序列分解为小时、日、周三个时间尺度的子序列, 分别由双向长短期记忆网络 (BiLSTM) 提取时序特征; 随后引入多头自注意力机制对各尺度特征进行自适应加权融合, 使模型能够依据预测任务动态调整不同尺度的贡献; 最后由全连接输出层生成多步预测结果。以某市空气质量监测站点 2023 年 1 月至 2024 年 12 月逐小时监测数据为研究对象, 与 ARIMA、随机森林、LSTM、Transformer、Informer 等基线模型进行系统对比。结果表明, MSA-LSTM 在 PM_{2.5} 浓度 1 h、6 h、12 h、24 h 等多步预测任务上均取得最优性能, 其中 24 h 预测平均绝对误差为 6.2 $\mu\text{g}/\text{m}^3$, 均方根误差为 9.8 $\mu\text{g}/\text{m}^3$, 决定系数为 0.93; 消融实验进一步验证了多尺度分解与注意力融合模块的独立贡献。研究结果表明, 多尺度建模与注意力机制的有机结合能够显著提升污染物浓度预测精度, 可为多学科交叉背景下的数据驱动环境建模提供方法参考。

关键词: 深度学习; 时序预测; 注意力机制; 多尺度特征; 空气质量; PM_{2.5}

Abstract: Urban air quality forecasting is a fundamental basis for precise air pollution control and public health risk management. Affected by the multiple coupling of emission sources, meteorological conditions, and atmospheric chemical processes, pollutant concentration time series generally exhibit strong nonlinearity, non-stationarity, and multi-scale characteristics: short-term fluctuations on the hourly scale, traffic and living emission rhythms on the daily scale, production activity cycles on the weekly scale, and seasonal meteorological evolution are superimposed, making accurate forecasting a challenging task. To address this issue, this paper proposes a multi-scale attention LSTM model (MSA-LSTM). The model first divides the input sequence into sub-sequences of hourly, daily, and weekly scales through multi-scale sliding windows, and extracts temporal features at each scale with bidirectional long short-term memory networks (BiLSTM). A multi-head self-attention module is then introduced to adaptively fuse the multi-scale features so that the model can dynamically adjust the contributions of different scales according to the forecasting task. Finally, a fully connected layer generates multi-step forecasting results. Using hourly monitoring data of an urban air quality station from January 2023 to December 2024, systematic comparative experiments are conducted against ARIMA, random forest, LSTM, Transformer, and Informer. The results show that MSA-LSTM achieves the best performance in multi-step forecasting tasks at 1 h, 6 h, 12 h, and 24 h horizons, with a mean absolute error of 6.2 $\mu\text{g}/\text{m}^3$, a root mean square error of 9.8 $\mu\text{g}/\text{m}^3$, and a coefficient of determination of 0.93 for 24 h forecasting. Ablation studies further verify the independent contributions of the multi-scale decomposition and attention fusion modules. The

results indicate that the combination of multi-scale modeling and attention mechanisms can significantly improve pollutant concentration forecasting accuracy, providing methodological references for data-driven environmental modeling in multidisciplinary contexts.

Keywords: deep learning; time-series forecasting; attention mechanism; multi-scale features; air quality; PM2.5

1 引言

大气细颗粒物污染是当前全球突出的环境与公共卫生问题之一。世界卫生组织（WHO）在 2021 年发布的《全球空气质量指南》中，将 PM2.5 的年均指导值由此前的 $10 \mu\text{g}/\text{m}^3$ 收紧至 $5 \mu\text{g}/\text{m}^3$ ，并指出长期暴露于 PM2.5 等污染物与心血管疾病、呼吸系统疾病、糖尿病及认知功能损害等多种健康风险存在显著关联[1]。在我国，秋冬季重污染过程频发，空气质量的精准预报已成为大气污染协同治理、重污染天气应急响应与公众健康防护的重要科技支撑。

污染物浓度预报的核心难点在于其时间序列的高度复杂性。受局地排放源、气象条件与大气化学过程的多重耦合影响，污染物浓度序列同时呈现出强非线性、非平稳性与显著的多尺度特征：小时尺度上，早晚交通高峰与边界层演变导致浓度快速波动；日尺度上，生活与生产活动的节律性变化形成明显的日内循环；周尺度上，周末与工作日的生产活动差异带来周循环；季节尺度上，降水、风速与大气扩散条件的季节演变主导浓度的长期趋势。上述多尺度过程相互叠加，使得单一时间尺度模型难以全面刻画其动态规律。

早期研究主要采用统计学方法。自回归移动平均（ARIMA）模型及其季节性扩展通过识别序列的线性自相关结构实现预测，在平稳性假设下具有良好的可解释性与计算效率[2]。然而，统计模型对非线性、非平稳序列的拟合能力有限，且通常需要人工进行差分与参数识别，难以满足重污染过程等强非线性场景的预测需求。

此后，以随机森林（RF）[3]、支持向量机、XGBoost[4]等为代表的机器学习方法被引入空气质量预测。此类方法能够拟合输入特征与目标变量之间的复杂非线性映射，并支持多气象要素、多污染物协同输入，在特征充足时表现出优于统计模型的精度。但其预测性能高度依赖人工特征工程，难以显式建模时间序列的长程依赖结构，对输入序列中时间上下文信息的利用也较为有限。

近年来，深度学习为时序预测提供了新的范式。长短时记忆网络（LSTM）[5]与门控循环单元（GRU）[6]通过门控机制缓解了循环神经网络的梯度消失问题，能够有效捕捉序列的长程依赖；基于注意力机制的序列模型[7-8]进一步实现了对输入信息的自适应加权；卷积 LSTM[9]、Informer[10]、Autoformer[11]等改进结构在长序列建模与计算效率方面不断取得进展，DeepAR[12]、N-BEATS[13]、Temporal Fusion Transformer[14]等模型亦在概率预测与可解释性方面进行了探索。在空气质量领域，Qi 等[15]将图卷积网络与 LSTM 结合实现 PM2.5 的时空预测，Freeman 等[16]系统比较了多种深度学习模型在空气质量时间序列上的预测性能，均取得优于传统方法的精度。然而，上述模型大多采用单一时间尺度对序列进行建模，未充分显式利用污染物浓度在不同时间尺度上的差异化动态特征，多尺度信息的融合仍不充分。

针对上述问题，本文提出一种基于注意力机制的多尺度时序预测模型（multi-scale attention LSTM, MSA-LSTM）。该模型的创新之处体现在三个方面：（1）提出多尺度分解策略，将输入序列划分为小时、日、周三个时间尺度的子序列，分别建模污染物浓度的短期波动与长期趋势，缓解单一尺度建模造成的信息混叠；（2）设计基于多头自注意力的跨尺度特征融合模块，使模型能够依据预测任务动态调整不同时间尺度特征的贡献权重；（3）在统一框架下系统评估多步预测性能，并通过对比实验与消融实验验证各模块的有效性。

本文结构安排如下：第 2 节介绍模型结构与实验方法；第 3 节给出实验设置与结果分析；第 4 节讨论模型的优势、局限与未来方向；第 5 节总结全文。

2 方法

2.1 问题定义

设监测站点在时刻 t 观测到的特征向量为 $x_t \in \mathbb{R}^d$ ，其中 d 为特征维数，包含多种污染物浓度与气象要素。给定过去 T 个时刻的历史观测序列 $X=[x_{t-T+1}, x_{t-T+2}, \dots, x_t]$ ，本文任务为学习映射 $f: \mathbb{R}^{(T \times d)} \rightarrow \mathbb{R}^H$ ，输出未来 H 个时刻的 PM2.5 浓度预测序列 $\hat{Y}=[\hat{y}_{t+1}, \hat{y}_{t+2}, \dots, \hat{y}_{t+H}]$ ，使得预测值尽可能逼近真实值 $Y=[y_{t+1}, y_{t+2}, \dots, y_{t+H}]$ 。

2.2 数据与预处理

实验采用某市空气质量监测站点 2023 年 1 月至 2024 年 12 月共 24 个月的逐小时监测数据，包含 PM2.5、PM10、O3、NO2、SO2、CO 六项污染物浓度以及温度、相对湿度、风速、风向、气压五项气象要素，共 11 维特征。数据预处理包括三个步骤：（1）缺失值处理：对单小时缺失数据采用相邻时刻线性插值，对连续缺失超过 6 h 的时段予以剔除，避免插值引入虚假信息；（2）异常值处理：对超出正常物理范围的极值（如 PM2.5 浓度小于 0 或大于 1000 $\mu\text{g}/\text{m}^3$ ）进行剔除与插补；（3）标准化处理：对各特征进行 Z-score 标准化，消除量纲差异对模型训练的影响。按时间顺序将数据集划分为训练集（70%）、验证集（15%）与测试集（15%），测试集在时间上完全滞后于训练集，以避免数据泄漏。

2.3 模型结构

MSA-LSTM 模型由输入嵌入层、多尺度特征提取层、注意力融合层与输出层四部分组成，各部分设计如下。

（1）输入嵌入层：对每个时刻的特征向量 x_t 进行线性映射，将其变换为与模型隐层维度 d_h 一致的嵌入向量 $e_t=W_e \cdot x_t+b_e$ ，其中 $W_e \in \mathbb{R}^{(d_h \times d)}$ 、 $b_e \in \mathbb{R}^{d_h}$ 为可学习参数。嵌入层的作用在于将异构的污染物与气象特征投影到统一的表示空间中，便于后续时序建模。

（2）多尺度特征提取层：设计窗口长度分别为 6、24、72 的三条并行分支，对长度为 T 的输入序列进行滑窗切片，分别形成小时、日、周三个时间尺度的子序列。每个分支采用双向 LSTM (BiLSTM) 提取时序特征。以单向 LSTM 为例，其内部状态更新过程可表示为： $i_t=\sigma(W_i \cdot x_t+U_i \cdot h_{t-1}+b_i)$ ， $f_t=\sigma(W_f \cdot x_t+U_f \cdot h_{t-1}+b_f)$ ， $o_t=\sigma(W_o \cdot x_t+U_o \cdot h_{t-1}+b_o)$ ， $\tilde{c}_t=\tanh(W_c \cdot x_t+U_c \cdot h_{t-1}+b_c)$ ， $c_t=f_t \odot c_{t-1}+i_t \odot \tilde{c}_t$ ， $h_t=o_t \odot \tanh(c_t)$ 。式中： i_t 、 f_t 、 o_t 分别为输入门、遗忘门与输出门； c_t 为细胞状态； $\odot$ 表示逐元素相乘； σ 为 Sigmoid 激活函数。BiLSTM 将前向与后向两个方向的隐状态拼接作为输出，从而同时利用过去与未来的上下文信息。三个分支的输出分别记为 h^s 、 h^d 、 h^w 。

（3）注意力融合层：将三个尺度的特征表示拼接后输入多头自注意力模块。设拼接后的特征矩阵为 $Z \in \mathbb{R}^{(L \times d_h)}$ (L 为序列长度)，自注意力通过查询 (Query)、键 (Key)、值 (Value) 三个投影计算注意力权重： $Q=Z \cdot W_Q$ ， $K=Z \cdot W_K$ ， $V=Z \cdot W_V$ ， $\text{Attention}(Q,K,V)=\text{softmax}(QK^T/\sqrt{d_k}) \cdot V$ 。其中 d_k 为键向量的维度，缩放因子 $\sqrt{d_k}$ 用于防止点积结果过大导致 softmax 梯度消失。多头注意力将上述计算并行执行 h 次并对各头结果进行拼接与线性变换，使模型能够从多个表示子空间捕捉跨尺度依赖关系。随后经残差连接与层归一化 (LayerNorm)，增强训练稳定性并缓解深层网络的梯度问题。

（4）输出层：将融合后的特征展平，经两层全连接网络（中间层采用 ReLU 激活并施加 Dropout）映射为 H 步预测结果，输出层采用线性激活以保证预测值的连续尺度。

2.4 损失函数与训练策略

模型采用均方误差 (MSE) 作为损失函数： $L=(1/N)\sum \|Y-\hat{Y}\|^2$ ，其中 N 为批内样本数。训练采用 Adam 优化器，初始学习率 1×10^{-3} 并配合余弦退火调度；设置早停策略，当验证集损失连续 10 个轮次不下降时终止训练，并保存验证集上性能最优的模型参数。为缓解过拟合，在注意力融合层与全连接层之间施加 Dropout（比率 0.1），并对 LSTM 权重施加 L2 正则化。

2.5 评价指标

采用平均绝对误差 (MAE)、均方根误差 (RMSE) 与决定系数 (R^2) 三个指标评价模型性能，计算公式分别为： $MAE=(1/n)\sum|y_i-\hat{y}_i|$ ， $RMSE=\sqrt{(1/n)\sum(y_i-\hat{y}_i)^2}$ ， $R^2=1-\sum(y_i-\hat{y}_i)^2/\sum(y_i-\bar{y})^2$ 。式中： y_i 为真实值； $\hat{y}_i$ 为预测值； $\bar{y}$ 为真实值均值； n 为测试样本数。MAE 与 RMSE 越小、 R^2 越接近 1，表明预测精度越高。其中 RMSE 对大误差更为敏感，能够反映模型在重污染过程等极端事件上的表现。

3 实验与结果

3.1 实验设置

模型超参数设置如下：嵌入维度 64，LSTM 隐层维度 128，多头注意力头数 4，Dropout 率 0.1，批大小 64，训练轮数上限 80，初始学习率 1×10^{-3} 。输入序列长度 $T=168$ （一周），预测步长 H 分别取 1、6、12、24。所有实验基于 Python 3.10 与 PyTorch 2.0 框架实现，在一台配备 NVIDIA GPU 的工作站上完成。基线模型包括 ARIMA[2]、随机森林[3]、LSTM[5]、Transformer[7]与 Informer[10]，各模型均采用网格搜索对关键超参数进行调优。需要说明的是，本文作为样刊示例稿，本节及表 1—表 3 所列结果用于展示论文结果部分的呈现体例。

3.2 对比实验

表 1 给出了各模型在预测步长 $H=24$ 时 PM2.5 浓度预测性能的对比结果。

表 1 不同模型 PM2.5 浓度 24 h 预测性能对比

模型	MAE/($\mu\text{g}\cdot\text{m}^{-3}$)	RMSE/($\mu\text{g}\cdot\text{m}^{-3}$)	R^2
ARIMA	12.4	17.8	0.76
随机森林	10.1	15.3	0.82
LSTM	8.9	13.6	0.86
Transformer	7.4	11.9	0.89
Informer	7.0	11.2	0.90
MSA-LSTM (本文)	6.2	9.8	0.93

由表 1 可见：（1）深度学习模型整体优于统计与机器学习模型，表明序列建模能力对预测精度具有决定性影响；（2）在深度模型中，引入注意力机制的 Transformer 与 Informer 进一步优于单一 LSTM；（3）本文提出的 MSA-LSTM 取得最优性能，较性能次优的 Informer 模型，MAE 降低 11.4%，RMSE 降低 12.5%， R^2 提升 0.03，验证了多尺度建模与注意力融合联合增益。

3.3 消融实验

为检验各模块的独立贡献，设置以下消融变体：①去除多尺度分解模块（仅保留单一 LSTM 分支，记为 w/o MS）；②去除注意力融合模块（将多尺度特征简单拼接后直接输入输出层，记为 w/o AT）；③去除双向结构（将 BiLSTM 替换为单向 LSTM，记为 w/o Bi）。结果如表 2 所示。

表 2 消融实验结果

模型变体	MAE/($\mu\text{g}\cdot\text{m}^{-3}$)	RMSE/($\mu\text{g}\cdot\text{m}^{-3}$)	R^2
w/o MS	7.1	11.2	0.90
w/o AT	6.8	10.6	0.91

w/o Bi	6.5	10.1	0.92
MSA-LSTM (完整)	6.2	9.8	0.93

消融结果表明：多尺度分解模块贡献最大（去除后 MAE 上升 $0.9 \mu\text{g}/\text{m}^3$ ），说明显式区分不同时间尺度是提升预测精度的关键；注意力融合模块次之（去除后 MAE 上升 $0.6 \mu\text{g}/\text{m}^3$ ），验证了跨尺度自适应加权的重要性；双向结构亦提供了一定增益（去除后 MAE 上升 $0.3 \mu\text{g}/\text{m}^3$ ）。各模块均对整体性能有正向贡献，且贡献相互正交，不存在明显的冗余。

3.4 多步预测性能分析

为考察模型在不同预测时域上的表现，表 3 给出了 MSA-LSTM 在 $H=1、6、12、24$ 时的预测性能。

表 3 MSA-LSTM 在不同预测步长下的性能

预测步长 H/h	MAE/ $(\mu\text{g}\cdot\text{m}^{-3})$	RMSE/ $(\mu\text{g}\cdot\text{m}^{-3})$	R^2
1	3.1	4.8	0.97
6	4.4	6.9	0.95
12	5.3	8.4	0.94
24	6.2	9.8	0.93

结果显示，随着预测步长的增加，预测误差逐步累积、 R^2 逐渐下降，符合多步预测的一般规律；值得注意的是，即使在 24 h 预测时域上，模型仍保持 0.93 的决定系数，表明其能够为未来一天内的空气质量提供可靠的预报参考。

3.5 注意力权重分析

进一步对注意力融合层的权重进行可视化分析发现：在重污染过程发生期间，模型赋予小时尺度特征更高的注意力权重，以捕捉浓度的快速抬升；而在平稳时段，周尺度趋势特征的权重明显上升，表明模型能够根据序列动态自适应调整多尺度信息的利用策略。该现象从可解释性角度印证了多尺度注意力融合设计的合理性。

4 讨论

4.1 性能优势来源

综合上述实验结果，MSA-LSTM 的性能优势可归结为两点：其一，多尺度分解使模型能够显式分离污染物浓度的短期波动与长期趋势，避免了单一尺度建模中不同频率成分的信息混叠，这与空气污染过程的多尺度驱动机制高度契合；其二，注意力融合机制实现了跨尺度信息的动态加权，使模型在重污染事件与平稳时段能够采取差异化的特征利用策略，兼具建模容量与灵活性。

4.2 计算复杂度分析

从计算成本看，MSA-LSTM 的三个 BiLSTM 分支在参数量上约为单一 LSTM 的三倍，多头注意力模块进一步增加了部分参数。相较于性能次优的 Informer，模型在训练时间上增加约 20%，但预测精度提升明显，在精度优先的应用场景（如重污染过程预报）中具有合理的性价比。对于资源受限的边缘部署场景，可通过共享 LSTM 权重、缩减隐层维度或引入模型蒸馏进行轻量化，这是后续优化的方向之一。

4.3 局限性与未来工作

本研究尚存在以下局限：其一，本文数据为模拟数据，所得结论需在真实监测数据集上重新验证后方可用于实际业务；其二，模型仅在单站点数据上验证，尚未考察站点间空间相关性与区域传输对预测的影响；其三，模型输出为确定性点预测，未提供预测不确定性区间，难以

满足风险决策对概率预报的需求。未来工作将：（1）引入多站点图神经网络建模空间依赖，构建时空一体化预测框架；（2）采用分位数回归或深度集成方法输出概率预测；（3）探索与气象数值预报产品的耦合，提升重污染过程的预报命中率。

5 结论

本文针对空气质量预报中的多尺度时序建模问题，提出一种基于注意力机制的多尺度时序预测模型 MSA-LSTM。该模型通过多尺度分解与多头自注意力融合，实现了对污染物浓度短期波动与长期趋势的协同建模。对比实验、消融实验与多步预测分析表明：模型性能优于 ARIMA、随机森林、LSTM、Transformer 与 Informer 等基线模型；多尺度分解与注意力融合模块均贡献了显著且相互正交的正向效果；模型在 24 h 预测时域上仍保持较高的精度。研究成果可为多学科交叉环境下的数据驱动环境建模提供方法参考，后续将结合真实数据与空间建模进一步开展验证与优化。

参考文献

- [1] World Health Organization. WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide[R]. Geneva: World Health Organization, 2021.
- [2] Box G E P, Jenkins G M, Reinsel G C, et al. Time series analysis: forecasting and control[M]. 5th ed. Hoboken: John Wiley & Sons, 2015.
- [3] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [4] 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. New York: ACM, 2016: 785-794.
- [5] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [6] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1724-1734.
- [7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. Red Hook: Curran Associates, 2017: 5998-6008.
- [8] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
- [9] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]//Advances in Neural Information Processing Systems 28. Red Hook: Curran Associates, 2015: 802-810.
- [10] Zhou H, Zhang S, Peng J, et al. Informer: beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(12): 11106-11115.
- [11] Wu H, Xu J, Wang J, et al. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting[C]//Advances in Neural Information Processing Systems 34. Red Hook: Curran Associates, 2021: 22419-22430.
- [12] Salinas D, Flunkert V, Gasthaus J, et al. DeepAR: probabilistic forecasting with autoregressive recurrent networks[J]. International Journal of Forecasting, 2020, 36(3): 1181-1191.
- [13] Oreshkin B N, Carpo D, Chapados N, et al. N-BEATS: neural basis expansion analysis for interpretable time series forecasting[C]//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

- [14] Lim B, Arık S Ö, Loeff N, et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting[J]. *International Journal of Forecasting*, 2021, 37(4): 1748-1764.
- [15] Qi Y, Li Q, Karimian H, et al. A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory[J]. *Science of the Total Environment*, 2019, 664: 1-10.
- [16] Freeman B S, Taylor G, Gharabaghi B, et al. Forecasting air quality time series using deep learning[J]. *Journal of the Air & Waste Management Association*, 2018, 68(8): 866-886.