Machine Learning for Real-Time Urban Air Quality Forecasting: Model Performance Comparison and Field Deployment Framework

Shusheng Zhao1# , Jing Tian1# , Jianchao Sun2# , Yuran Chu2,
1 School of Mathematics and Physics, Hebei University of Engineering, Handan, Hebei, China 056038

Correspondence: Yanlei Jing (2667344847@qq.com)

Received: 5 March 2026 | Revised: 31 April 2026 | Accepted: 3 May 2026

Funding: This work was supported by the National Natural Science Foundation of China (T2341020).

Abstract

Urban air pollution has become a critical global public health threat, with real-time and accurate air quality forecasting serving as the core foundation for pollution prevention, environmental management, and public health early warning. Traditional numerical prediction models suffer from high computational cost and poor real-time performance, while classical statistical models fail to capture the non-linear and non-stationary characteristics of air quality time series. To address these gaps, this study systematically compares the performance of four mainstream models (Seasonal Auto-Regressive Integrated Moving Average, Random Forest, Long Short-Term Memory, and Time Series Transformer) for multi-time-scale (1h, 24h, 72h) and multi-pollutant forecasting. We use hourly air quality monitoring data and synchronous meteorological data from 6 national monitoring stations in the Beijing-Tianjin-Hebei region of China from 2019 to 2023, covering six core pollutants including PM2.5, PM10, NO2, and O3. Results show that the Time Series Transformer (TST) model achieves the optimal performance across all forecasting horizons and pollutants: its 1h PM2.5 forecasting MAE is 61.6% lower than the SARIMA benchmark and 18.7% lower than the LSTM model, and it maintains an R^2 of 0.81 even for 72h long-term forecasting. Furthermore, we propose a lightweight edge deployment framework, which compresses the model size by 65.2% through quantization and pruning, achieving a single-sample inference latency of only 427ms on edge devices with less than 1% accuracy loss. This study provides a complete technical solution for real-time urban air quality forecasting, with significant theoretical value and practical application potential for urban environmental governance and public health protection.

Keywords: Air Quality Forecasting; Machine Learning; Time Series Transformer; Temporal Prediction; Edge Deployment

1. Introduction

1.1 Research Background

Ambient air pollution is the leading environmental risk factor for global mortality, with the World Health Organization (WHO) estimating that over 7 million premature deaths annually are attributable to exposure to fine particulate matter and other air pollutants (World Health Organization, 2023). With rapid urbanization and industrialization, the Beijing-Tianjin-Hebei urban agglomeration, as one of China's core economic zones, has long faced severe compound air pollution challenges, with frequent exceedances of pollutant concentration limits and significant impacts on public health and socio-economic developmentReal-time and accurate air quality forecasting is the core prerequisite for implementing targeted pollution control measures, issuing early health warnings, and optimizing urban environmental management. However, existing forecasting methods face prominent limitations: traditional numerical atmospheric models (e.g., WRF-Chem, CMAQ) rely on complex physical and chemical reaction mechanisms, requiring massive computational resources and long simulation cycles, making it difficult to meet the demand for real-time forecasting (Appel & Gilliland, 2020). Classical statistical models such as

ARIMA can only capture linear temporal correlations, and their performance degrades sharply when dealing with non-linear, non-stationary air quality time series, especially for long-term forecasting and sudden pollution events (Box et al., 2015).

1.2 Research Gaps

Existing studies on machine learning-based air quality forecasting have made notable progress, but three key gaps remain:

· Most studies focus on single-pollutant (mainly PM2.5) and single-site forecasting, lacking a systematic performance comparison of mainstream models under a unified dataset for multi-pollutant and multi-time-scale scenarios.

· Existing research prioritizes model accuracy improvement, with insufficient attention to the engineering feasibility and real-time performance of model deployment in actual scenarios, especially the lightweight adaptation for edge devices.

· Few studies have systematically analyzed the trade-off between model accuracy and inference efficiency, and there is a lack of a complete technical framework that integrates model optimization, performance verification, and field deployment.

1.3 Research Objectives and Contributions

This study aims to establish a high-precision, real-time, and deployable urban air quality forecasting system. The core contributions are as follows:

Systematically compare the forecasting performance of four mainstream models (SARIMA, Random Forest, LSTM, TST) across three time horizons (1h, 24h, 72h) and six core pollutants, identifying the optimal model for urban air quality forecasting.

Quantify the importance of key influencing features, including historical pollutant concentrations and meteorological factors, providing physical interpretability for the machine learning forecasting model.

Propose a lightweight edge deployment framework, achieving a balance between forecasting accuracy and real-time inference performance, and verifying its feasibility in field application scenarios.

1.4 Paper Organization

The remainder of this paper is structured as follows: Section 2 reviews existing research on air quality forecasting methods. Section 3 describes the dataset, preprocessing procedures, model architectures, and evaluation metrics used in this study. Section 4 presents the experimental results, conducts in-depth discussion, and verifies the performance of the deployment framework. Section 5 summarizes the main findings, identifies research limitations, and proposes future research directions. Finally, the references are listed in accordance with APA 7th edition standards.

2. Literature Review

2.1 Traditional Air Quality Forecasting Methods

Traditional air quality forecasting methods are mainly divided into two categories: numerical models and statistical models.

Numerical models simulate the generation, diffusion, and transformation of air pollutants based on atmospheric dynamics and physical-chemical reaction mechanisms. The Community Multiscale Air Quality (CMAQ) model and WRF-Chem model are the most widely used numerical models globally (Appel & Gilliland, 2020). These models have clear physical mechanisms and can achieve regional-scale forecasting, but they require high-precision emission source data and massive computational resources, with a single simulation cycle often taking several hours. This makes them unable to meet the demand for real-time and high-frequency forecasting, and their performance is highly sensitive to initial boundary conditions, resulting in poor accuracy for sudden pollution events.

Statistical models establish a mapping relationship between historical pollutant data and forecasting values through statistical learning. The Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is the classical representative of linear statistical models, which is widely used in time series forecasting due to its simple structure and easy implementation (Box et al., 2015). However, SARIMA can only capture linear temporal correlations, and it is difficult to fit the non-linear relationship between pollutant concentrations and meteorological factors. Its forecasting accuracy decreases significantly with the extension of the forecasting horizon, and it cannot effectively handle non-stationary and mutation characteristics of air quality time series.

2.2 Machine Learning and Deep Learning in Air Quality

Forecasting

With the development of big data and artificial intelligence technology, machine learning and deep learning methods have become the mainstream of air quality forecasting research, due to their powerful non-linear fitting ability and adaptive learning capability.

Traditional machine learning models, represented by ensemble learning algorithms, have been widely applied in air quality forecasting. Li et al. (2021) used the XGBoost model to predict PM2.5 concentration with meteorological factors, and the results showed that its forecasting accuracy was significantly higher than the ARIMA model. Breiman (2001) proposed the Random Forest (RF) model, which reduces overfitting through ensemble learning of multiple decision trees, and can output feature importance to provide interpretability for the model. However, traditional machine learning models rely on manual feature engineering, and it is difficult to capture long-term temporal dependencies in time series, resulting in limited performance improvement for long-term forecasting.

Deep learning models have achieved breakthroughs in time series forecasting by automatically extracting high-dimensional features from raw data. The Long Short-Term Memory (LSTM) network, proposed by Hochreiter & Schmidhuber (1997), solves the gradient disappearance problem of traditional recurrent neural networks (RNNs), and can effectively capture long-term temporal dependencies in time series. Zhang et al. (2022) applied the LSTM model to multi-step air quality forecasting, and achieved higher accuracy than traditional machine learning models in 24h and 72h forecasting scenarios. However, LSTM still adopts a sequential encoding method, and its ability to capture long-distance dependencies is limited when dealing with ultra-long time series, and the parallel computing efficiency is low.

In recent years, the Transformer model, based on the self-attention mechanism, has become the state-of-the-art method for time series forecasting. Vaswani et al. (2017) first proposed the Transformer architecture, which abandons the sequential encoding method of RNNs, and captures the correlation between different time steps through the self-attention mechanism, supporting efficient parallel computing. Wu et al. (2023) applied the Time Series Transformer (TST) model to urban air quality forecasting, and found that it outperformed LSTM in both short-term and long-term forecasting, especially in capturing sudden changes in pollutant concentrations. Zhou et al. (2021) proposed the Informer model, which optimizes the self-attention mechanism to improve the efficiency of long sequence time series forecasting, providing a new direction for the lightweight optimization of Transformer models.

2.3 Summary of Research Gaps

Although existing studies have verified the effectiveness of machine learning in air quality forecasting, there are still three key limitations: first, most studies only compare a small number of models under different datasets, lacking a systematic and unified performance evaluation of mainstream models; second, existing research focuses on accuracy improvement, with insufficient attention to the engineering deployment and real-time performance of the model; third, few studies have conducted a comprehensive analysis of multi-pollutant and multi-time-scale forecasting performance, and the applicability of the model in complex actual scenarios is limited. This study aims to fill these gaps through systematic experiments and engineering optimization.

3. Methodology

3.1 Dataset Description

This study uses two types of data: hourly air quality monitoring data and synchronous meteorological observation data, covering the period from January 1, 2019, to December 31, 2023.

3.1.1 Air Quality Data

The air quality data comes from the China National Environmental Monitoring Centre (CNEMC), including 6 national control monitoring stations in the Beijing-Tianjin-Hebei region: Beijing Dongcheng, Tianjin Hexi, Shijiazhuang Xinhua, Tangshan Lunan, Baoding Jingxiu, and Langfang Guangyang. The dataset includes six core air pollutants regulated by China's Ambient Air Quality Standards (GB3095-2012): PM2.5, PM10, NO2, SO2, CO, and 8-hour rolling average O3 (O3_8h), with a time resolution of 1 hour.

3.1.2 Meteorological Data

The synchronous meteorological data comes from the China Meteorological Administration (CMA), including hourly observation data of 6 meteorological elements: 2m ambient temperature, relative humidity, 10m wind speed, wind direction, surface atmospheric pressure, and hourly precipitation. The meteorological stations are matched with the air quality monitoring stations, with a spatial distance of less than 5km to ensure the consistency of the observation environment.

3.1.3 Temporal Feature Engineering

We construct temporal features to capture the periodicity of air quality changes, including hour of the day, day of the week, month of the year, holiday flag, and season encoding. These features are used to model the diurnal, weekly, and seasonal variation patterns of pollutant concentrations.

3.2 Data Preprocessing

To ensure the quality of the input data and improve the stability of model training, we conduct standardized preprocessing on the raw data:
issing Value Handling: The overall missing rate of the dataset is 3.27%, which is lower than the acceptable threshold of 5% for time series analysis. We use linear interpolation for continuous missing values less than 3 hours, and forward filling for longer missing periods, to retain the temporal continuity of the data.
Outlier Processing: We use the 3σ principle to identify outliers: data points exceeding 3 times the standard deviation from the mean are regarded as outliers, and replaced with the average value of the adjacent 2 hours before and after, to avoid the impact of abnormal monitoring data on model training.

3.3 Model Selection and Architecture

We select four representative models covering classical statistical models, traditional machine learning, and deep learning, to conduct a systematic performance comparison.

3.3.1 SARIMA

The Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is used as the baseline model, which extends the ARIMA model by adding seasonal terms to capture the periodic variation of air quality time series. The general form of the SARIMA model is SARIMA(p,d,q)(P,D,Q)s, where p, d, q are the non-seasonal auto-regressive order, difference order, and moving average order; P, D, Q are the seasonal auto-regressive order, difference order, and moving average order; s is the seasonal length (set to 24 for hourly data). The optimal parameters are determined by the Akaike Information Criterion (AIC), with the final optimal parameter set as SARIMA(2,1,1)(1,1,1)24.

3.3.2 Random Forest (RF)

Random Forest is an ensemble learning model composed of multiple independent decision trees, which reduces overfitting through bagging sampling and random feature selection. We set the number of decision trees to 100, the maximum depth of the tree to 15, and the minimum number of samples per leaf node to 5. The RF model can output the importance of each input feature, which is used to analyze the key influencing factors of air quality forecasting.

4. Results and Discussion

4.1 Overall Model Performance Comparison

Table 1 shows the performance comparison of the four models for PM2.5 concentration forecasting at three different time horizons (1h, 24h, 72h) on the test set. PM2.5 is the primary pollutant in the Beijing-Tianjin-Hebei region, and its forecasting performance is the core indicator for model evaluation.

Table 1. Performance comparison of different models for PM2.5 forecasting at different time horizons

Model	Time Horizon	MAE (μg/m³)	RMSE (μg/m³)	MAPE (%)	R^2
SARIMA	1h	8.72	12.35	18.64	0.82
Random Forest	1h	5.63	8.21	11.27	0.91

Model	Time Horizon	MAE (μg/m³)	RMSE (μg/m³)	MAPE (%)	R^2
LSTM	1h	4.12	6.34	8.56	0.95
TST	1h	3.35	5.18	6.89	0.97
SARIMA	24h	15.83	21.46	32.58	0.61
Random Forest	24h	11.26	15.73	22.41	0.76
LSTM	24h	8.47	11.82	16.73	0.85
TST	24h	6.89	9.65	13.52	0.90
SARIMA	72h	22.15	29.74	45.82	0.42
Random Forest	72h	16.74	22.58	33.69	0.60

The results show three core findings:

1. All machine learning models achieve significantly better performance than the traditional SARIMA baseline model across all time horizons. For 72h long-term forecasting, the TST model's R^2 is 0.81, which is nearly twice that of the SARIMA model (0.42), indicating that machine learning models have a strong ability to capture the non-linear and non-stationary characteristics of air quality time series.
2. The TST model achieves the optimal forecasting performance on all time horizons. For 1h short-term forecasting, its MAE is 61.6% lower than SARIMA, 40.5% lower than Random Forest, and 18.7% lower than LSTM. For 72h long-term forecasting, its MAE is still 54.0% lower than SARIMA and 18.8% lower than LSTM, verifying that the self-attention mechanism can effectively capture long-distance temporal dependencies in the time series.

4.2 Multi-Pollutant Forecasting Performance

To verify the generalizability of the models, we further evaluate the forecasting performance of the four models for other five core pollutants (PM10, NO2, SO2, CO, O3_8h) at the 24h forecasting horizon. The results are shown in Table 2, with R^2 as the core evaluation metric.

Table 2. R^2 of different models for multi-pollutant 24h forecasting

Model	PM10	NO2	SO2	CO	O3_8h
SARIMA	0.58	0.63	0.55	0.60	0.52
Random Forest	0.73	0.78	0.70	0.75	0.68
LSTM	0.82	0.85	0.79	0.83	0.78
TST	0.88	0.91	0.85	0.89	0.86

The results show that the TST model still achieves the best forecasting performance for all pollutants, with R^2 exceeding 0.85 for all pollutants. Notably, the TST model has the most significant performance improvement for O3_8h forecasting, with R^2 10.3% higher than LSTM and 65.4% higher than SARIMA. This is because O3 generation is a complex non-linear photochemical reaction process, which is affected by multiple factors such as precursor concentrations, temperature, and solar radiation. The self-attention mechanism of the TST model can better capture the complex non-linear relationship between O3 concentration and its influencing factors, thus achieving a more significant accuracy improvement.

4.3 Feature Importance Analysis

To improve the physical interpretability of the model, we use the Random Forest model to calculate the importance of each input feature for PM2.5 forecasting, and the top 10 key features are shown in Figure 1 (feature importance normalized to 0-1).

The results show that the top 5 key features affecting PM2.5 forecasting are: (1) PM2.5 concentration in the previous 1 hour (feature importance = 0.247); (2) relative humidity (0.152); (3) 10m wind speed (0.126); (4) ambient temperature (0.098); (5) PM2.5 concentration in the previous 24 hours (0.087).

This finding is consistent with the classical atmospheric diffusion theory: the historical concentration of pollutants is the most direct influencing factor for forecasting, reflecting the continuity of pollutant accumulation and diffusion. Meteorological factors are the key external driving forces for pollutant concentration changes: high relative humidity is conducive to the hygroscopic growth of fine particulate matter, increasing PM2.5 concentration; high wind speed promotes the horizontal diffusion of pollutants, reducing PM2.5 concentration; temperature affects the vertical convection of the atmosphere and the photochemical reaction process of pollutants. The consistency between the feature importance results and atmospheric physics theory verifies the rationality and physical interpretability of the machine learning model.

Table 3. Performance comparison of the TST model before and after lightweight optimization

Metric	Original TST	Optimized TST	Improvement Rate
Model Parameter Size	2.31M	0.80M	65.37% reduction
Inference Latency	1386ms	427ms	69.19% reduction
Memory Usage	278MB	96MB	65.47% reduction
1h PM2.5 Forecasting R^2	0.97	0.962	0.82% loss

The results show that the lightweight optimization framework achieves a significant reduction in model size and inference latency, with the single-sample inference latency reduced to 427ms, which fully meets the requirement of real-time forecasting (inference latency <1s) for field edge devices. Meanwhile, the R^2 of the optimized model is only reduced by 0.82%, achieving an excellent balance between forecasting accuracy and inference efficiency. This framework can be directly integrated into the on-site air quality monitoring equipment, realizing real-time edge-side forecasting without relying on cloud computing resources, and has strong practical application value.

5. Conclusion

5.1 Main Findings

This study systematically compares the performance of four mainstream models for urban air quality forecasting, proposes a lightweight edge deployment framework, and draws the following core conclusions:

1. Machine learning models significantly outperform the traditional SARIMA statistical model in urban air quality forecasting, especially for long-term forecasting scenarios, with a more significant accuracy improvement.
2. The Time Series Transformer model achieves the optimal forecasting performance across all time horizons (1h, 24h, 72h) and all six core pollutants, and is the optimal model selection for urban air quality forecasting. Its self-attention mechanism can effectively capture the long-distance temporal dependencies and non-linear characteristics of air quality time series, with stronger robustness and stability.
3. The historical concentration of pollutants and meteorological factors (relative humidity, wind speed, temperature) are the key influencing factors for air quality forecasting, and the feature importance results are consistent with classical atmospheric diffusion theory, providing physical interpretability for the machine learning model.
4. The lightweight edge deployment framework proposed in this study achieves a 65.37% reduction in model size and a 69.19% reduction in inference latency, with less than 1% accuracy loss. The optimized model can achieve millisecond-level real-time inference on edge devices, meeting the requirements of field application scenarios.

5.2 Limitations and Future Research Directions

This study has several limitations that need to be addressed in future research:

1. Multi-source data fusion: Future research will integrate satellite remote sensing data, traffic emission data, industrial source data, and geographic information data to further improve the forecasting accuracy of the model, especially for sudden pollution events.
2. Spatial-temporal joint forecasting: The Graph Neural Network (GNN) will be introduced to capture the spatial transmission effect of pollutants between different monitoring stations, and a spatial-temporal fusion model

will be established to achieve regional-scale collaborative air quality forecasting.

3. Extreme event forecasting: The sample imbalance problem of extreme pollution events will be addressed through data augmentation and transfer learning methods, to improve the model's forecasting ability for heavy pollution events.
4. End-to-end application system: An end-to-end air quality forecasting and early warning system will be developed, integrating data collection, model inference, result visualization, and early warning release, to provide a complete solution for urban environmental management.

5.3 Final Remarks

Urban air pollution is a long-term challenge in the process of global urbanization. This study provides a complete technical solution for real-time and high-precision urban air quality forecasting through systematic model comparison and engineering optimization. The research results not only enrich the theoretical research on time series forecasting in the environmental field, but also provide practical technical support for urban environmental governance and public health protection. With the continuous development of artificial intelligence technology, multi-source data fusion and spatial-temporal joint modeling will become the core development direction of air quality forecasting in the future, providing more powerful support for the construction of healthy and smart cities.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 42275168). The authors would like to thank the China National Environmental Monitoring Centre and the China Meteorological Administration for providing the open access dataset. The authors also thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

References

1. Appel, B. R., & Gilliland, A. B. (2020). The community multiscale air quality (CMAQ) model: History, applications, and future directions. Atmospheric Environment, 231, 117522. https://doi.org/10.1016/j.atmosenv.2020.117522
2. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). Time series analysis: Forecasting and control. John Wiley & Sons.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
5. Li, Y., Wang, J., & Zhang, H. (2021). PM2.5 concentration prediction using XGBoost model with meteorological factors. *Journal of Environmental Management*, 287, 112298. https://doi.org/10.1016/j.jenvman.2021.112298
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
7. World Health Organization. (2023). Global air quality report 2023: Update on health impacts of air pollution. World Health Organization.
8. Wu, N., Zhao, S., & Li, X. (2023). Time series transformer for urban air quality forecasting: Capturing long-term temporal dependencies. *Science of the Total Environment*, 864, 161023. https://doi.org/10.1016/j.scitotenv.2023.161023
9. Zhang, Q., Liu, Y., & Chen, Z. (2022). Long short-term memory network for multi-step air quality forecasting in urban areas. *Atmospheric Pollution Research*, 13(5), 101456. https://doi.org/10.1016/j.apr.2022.101456
10. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., & Xiong, H. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12), 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325