**Frontiers in Computing and Intelligent Systems**
ISSN: 2832-6024 | Vol. 15, No. 3, 2026

# A Hybrid GCN–PCA–LSTM Framework for Accurate Spatiotemporal Prediction of PM2.5 Concentrations

**Qian Yu, Chunli Kan, Zhiyuan Hou, Man Yang, Ni Zheng, Hongwu Yuan ***

School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui, 230088, China
*** Corresponding author:** Hongwu Yuan

**Abstract:** PM2.5 pollution has become a critical environmental issue affecting air quality and public health. Accurate concentration prediction is of great significance for pollution early warning and control. Considering that PM2.5 concentration variations exhibit both temporal dependence and spatial correlation, along with the presence of redundant features in multi-source data, a spatiotemporal prediction model integrating a Graph Convolutional Network (GCN), Principal Component Analysis (PCA), and a Long Short-Term Memory network (LSTM) is proposed. Specifically, the GCN is first employed to characterize the spatial dependencies among air quality monitoring stations and to extract spatial features. Subsequently, PCA is utilized to reduce the dimensionality of high-dimensional features, thereby eliminating redundant information and improving computational efficiency. Finally, the LSTM is adopted to model temporal sequence features, enabling dynamic prediction of PM2.5 concentrations. Based on air pollutant and meteorological data collected from ten monitoring stations in Hefei from 2018 to 2019, sliding time window samples are constructed to predict PM2.5 concentrations for the next hour. LSTM, GCN, and GCN-LSTM models are selected as baseline methods for comparison. Experimental results demonstrate that the proposed GCN-PCA-LSTM model outperforms the comparative models in terms of RMSE, MAE, and $R^2$ metrics, achieving an RMSE of 7.94, an MAE of 5.79, and an R2 of 89.36%. The model is capable of more accurately capturing the variation trends of PM2.5 concentrations. Moreover, it maintains strong fitting performance during periods of high pollution and rapid fluctuations, indicating robust spatiotemporal modeling capability and stability. In summary, the integration of spatial feature extraction, feature dimensionality reduction, and temporal sequence modeling effectively enhances PM2.5 prediction performance, providing a feasible approach for urban air quality forecasting and refined environmental management.

**Keywords:** PM2.5 Prediction; Spatiotemporal Modeling; GCN; LSTM; PCA.

## 1. Introduction

With the continuous advancement of urbanization and industrialization, air pollution has become increasingly severe. Among various pollutants, fine particulate matter (PM2.5), characterized by its small particle size and long atmospheric residence time, can be readily inhaled into the human body, posing significant risks to the respiratory and cardiovascular systems. It has therefore emerged as a critical factor affecting public health. Previous studies[1] have demonstrated that short-term exposure to PM2.5 is significantly associated with increased incidence of respiratory diseases and mortality. Consequently, accurate prediction of PM2.5 concentrations is of great importance for air pollution early warning and environmental management.

PM2.5 concentration dynamics are influenced by multiple factors, including emission sources, meteorological conditions, and geographical environments, exhibiting pronounced nonlinearity and complexity. Existing prediction approaches can be broadly categorized into numerical simulation methods and data-driven methods. Numerical simulation models, such as CMAQ and WRF-Chem[2], are grounded in atmospheric physical and chemical mechanisms and offer strong interpretability; however, they are computationally intensive and structurally complex. In contrast, data-driven methods rely on historical data and provide advantages in modeling flexibility and computational efficiency, making them widely adopted in practice.

Within data-driven approaches, traditional statistical models, such as ARIMA[3] and wavelet decomposition[4], exhibit limitations in handling nonlinear relationships. Machine learning methods, including Random Forest[5] and Extreme Learning Machine[6], can improve prediction performance to some extent, yet they often fail to adequately capture temporal dependencies. In recent years, deep learning techniques have demonstrated superior performance in time series forecasting. Among them, the Long Short-Term Memory (LSTM) network, with its gated architecture, is capable of effectively capturing long-term dependencies and has become one of the mainstream approaches for PM2.5 prediction.

However, PM2.5 concentration variations exhibit not only temporal dependence but also significant spatial correlation. Due to geographical proximity, pollutant concentrations at different monitoring stations are often interrelated. Consequently, increasing attention has been devoted to spatiotemporal joint modeling in PM2.5 prediction. Some studies employ Convolutional Neural Networks (CNNs)[7] or attention mechanisms[8] to extract spatial features; however, these approaches typically require complex data structures or incur relatively high computational costs. As a deep learning model capable of handling non-Euclidean data, the Graph Convolutional Network (GCN)[9] can effectively model spatial relationships among monitoring stations via graph structures, demonstrating strong potential in air quality prediction tasks.

Despite these advances, several limitations remain. On the one hand, spatial correlations have not been fully exploited; on the other hand, multi-source data are often high-dimensional, leading to feature redundancy that may degrade model performance. Therefore, how to effectively capture spatial dependencies while reducing feature redundancy

remains a critical issue in PM2.5 prediction research.

To address these challenges, this study proposes a PM2.5 concentration prediction model that integrates GCN, Principal Component Analysis (PCA), and Long Short-Term Memory (LSTM). Specifically, the model first employs GCN to extract spatial features, then applies PCA for dimensionality reduction to mitigate feature redundancy, and finally utilizes LSTM to model temporal sequences and achieve PM2.5 concentration prediction.

The remainder of this paper is organized as follows. Section 2 introduces the model architecture and methodological principles. Section 3 describes the study area and data processing methods. Section 4 presents the experimental analysis. Finally, Section 5 concludes the paper and outlines future research directions.

# 2. Method

To simultaneously capture the spatial and temporal characteristics of PM2.5 concentration variations, a spatiotemporal prediction model integrating GCN, PCA, and LSTM is developed in this study. The model first employs GCN to extract spatial dependencies among monitoring stations, then applies PCA for feature dimensionality reduction, and finally utilizes LSTM to model temporal dynamics, thereby achieving PM2.5 concentration prediction. The overall architecture of the proposed model is illustrated in Fig. 1.

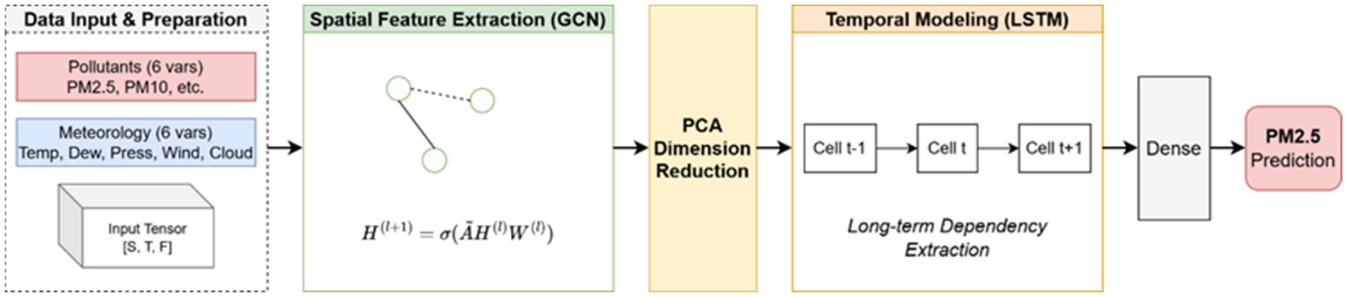


**Fig 1.** Structure of the GCN-PCA-LSTM prediction model

## 2.1. Spatial Feature Extraction Using GCN

PM2.5 pollution exhibits significant spatial correlation. To characterize the spatial relationships among monitoring stations, each station is represented as a node in a graph structure, and the graph is defined as $G = (V, E)$.

An adjacency matrix $A$ is constructed based on the geographical distances between stations, with weights defined as:

$$A_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right)$$

where $d_{ij}$ denotes the distance between station $i$ and station $j$, and $\sigma$ is the distance decay parameter. A threshold can be introduced to attenuate the influence of long-distance nodes.

The GCN aggregates node features through the adjacency matrix, and its propagation process is expressed as:

$$H^{(l+1)} = \sigma\left(AH^{(l)}W^{(l)}\right)$$

where $H^{(l)}$ represents the feature matrix at the l-th layer and $W^{(l)}$ denotes the trainable weight matrix. Through multiple graph convolution layers, the model effectively integrates information from neighboring monitoring stations, thereby extracting spatial features of PM2.5 concentration variations.

## 2.2. Feature Dimensionality Reduction Using PCA

Air quality data are typically characterized by high dimensionality and strong inter-feature correlations, which may lead to feature redundancy when directly used for modeling. To improve computational efficiency, Principal Component Analysis (PCA) is employed for dimensionality reduction.

Let the data matrix be denoted as $X$, and its covariance matrix is defined as:

$$C = \frac{1}{n}X^T X$$

Eigenvalue decomposition is then performed on the covariance matrix to obtain eigenvalues $\lambda_i$ and their corresponding eigenvectors. The cumulative variance contribution rate is calculated as:

$$CR_k = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i}$$

Principal components are selected based on the cumulative contribution rate to achieve feature compression. In this study, PCA is applied to the spatial features extracted by the GCN, aiming to reduce redundant information and enhance training efficiency.

## 2.3. Temporal Feature Learning Using LSTM

PM2.5 concentrations exhibit strong temporal dependencies; therefore, LSTM is employed to model the time series. LSTM utilizes a gating mechanism to regulate information flow, and its core computational processes are described as follows:

Forget gate:

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

Input gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

Cell state update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

Hidden state output:

$$h_t = o_t \cdot \tanh(C_t)$$

where $x_t$ denotes the input, $h_t$ represents the hidden state, and $C_t$ is the cell state. Through this architecture, the model can effectively capture long-term dependencies in time series data, thereby improving the prediction accuracy of PM2.5 concentrations.

# 3. Data Sources and Preprocessing

## 3.1. Data Sources

In this study, Hefei City, Anhui Province, China, is selected as the study area. Located in the middle and lower reaches of

the Yangtze River, Hefei experiences a subtropical humid monsoon climate. During winter, stagnant atmospheric conditions and temperature inversions frequently occur, which are unfavorable for pollutant dispersion. As a result, PM2.5 pollution in this region is relatively representative, making it a suitable case for analysis.

Air pollutant data are obtained from the air quality monitoring platform released by the China National Environmental Monitoring Center. Hourly observations from ten monitoring stations in Hefei are selected, including six pollutants: PM2.5, PM10, SO2, NO2, O3, and CO. The spatial distribution of the monitoring stations is shown in Fig. 2, and their geographic coordinates are listed in Table 1.

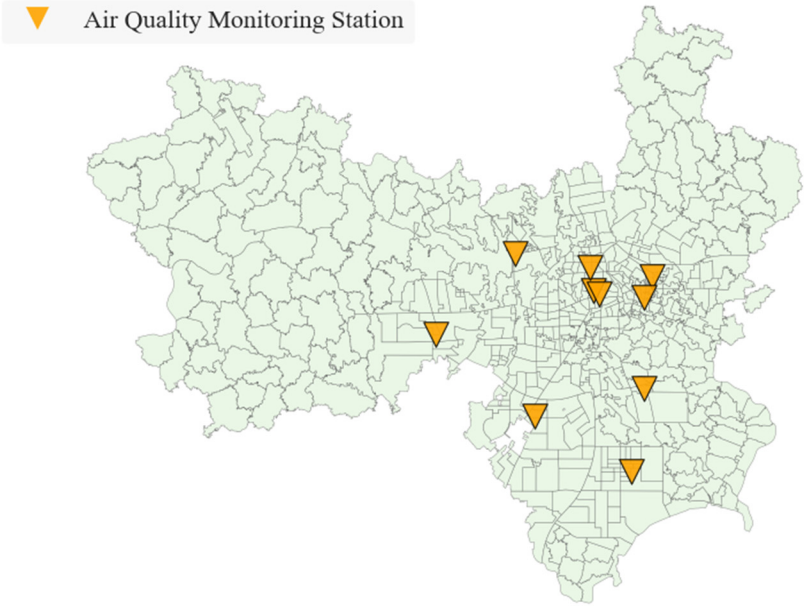


**Fig 2.** Spatial distribution of air quality monitoring stations in Hefei

**Table 1.** Geographic coordinates of air quality monitoring stations in Hefei

| Station Code | Station Name | Longitude | Latitude |
|---|---|---|---|
| 1270A | Mingzhu Square | 117.2252 | 31.7799 |
| 1271A | Sanlijie | 117.3229 | 31.8787 |
| 1272A | Hupo Villa | 117.2744 | 31.8689 |
| 1273A | Dongpu Reservoir | 117.2082 | 31.8952 |
| 1274A | Changjiang Middle Rd. | 117.2790 | 31.8665 |
| 1275A | Luyang District | 117.2714 | 31.8848 |
| 1276A | Yaohai District | 117.3160 | 31.8637 |
| 1277A | Baohe District | 117.3164 | 31.7992 |
| 1278A | Binhu New District | 117.3059 | 31.7404 |
| 1279A | High-tech Zone | 117.1422 | 31.8375 |

Meteorological data are collected from the China Meteorological Data Service Center. Hourly observations from the Hefei meteorological station (ID: 58321) are used, including variables such as air temperature, dew point temperature, atmospheric pressure, wind direction, wind speed, and cloud cover.

The study period spans from January 1, 2018, to December 31, 2019, with hourly resolution. The variables and their corresponding units are summarized in Table 2.

## 3.2. Data Preprocessing

To improve model performance, the raw data are preprocessed, including missing value imputation, outlier handling, and data normalization.

(1) Missing value imputation

Based on the missing patterns, the data are categorized into short-term missing values and consecutive missing intervals. For isolated missing values, the neighboring mean method is applied:

$$X_t = \frac{X_{t-1} + X_{t+1}}{2}$$

For consecutive missing segments, linear interpolation is adopted:

$$X_t = X_{t_0} + \frac{X_{t_1} - X_{t_0}}{t_1 - t_0}(t - t_0)$$

(2) Outlier handling

Outliers may arise from measurement errors or sudden pollution events. Considering that extreme PM2.5 values may carry meaningful information, this study retains outliers to preserve data authenticity.

**Table 2.** Description of the dataset

| Data Type | Variable | Unit |
|---|---|---|
| Air pollutants | $PM_{2.5}$ | $\mu g/m^3$ |
| | $PM_{10}$ | $\mu g/m^3$ |
| | $SO_2$ | $\mu g/m^3$ |
| | $NO_2$ | $\mu g/m^3$ |
| | $O_3$ | $\mu g/m^3$ |
| | CO | $mg/m^3$ |
| Meteorological data | Temperature | ℃ |
| | Dew point | ℃ |
| | Pressure | $hPa$ |
| | Wind direction | - |
| | Wind speed | $m/s$ |
| | Cloud cover | - |

(3) Data normalization

To eliminate scale differences among variables, the Z-score normalization method is applied:

$$X' = \frac{X - \mu}{\sigma}$$

After normalization, each variable has a mean of 0 and a standard deviation of 1, which helps improve model stability and prediction accuracy.

# 4. Experimental Setup and Results Analysis

## 4.1. Input Structure and Evaluation Metrics

(1) Model Input Structure

PM2.5 concentration exhibits strong temporal dependence; therefore, a sliding window approach is employed to construct input samples during model training. Specifically, 24 consecutive hours of historical data are used as input to predict the PM2.5 concentration in the next hour.

Let the time series be defined as:

$$X = \{x_1, x_2, x_3, \dots, x_T\}$$

where $x_t$ denotes the observation at time t. Using the sliding window method, the training samples can be constructed as follows:

$$X_i = (x_i, x_{i+1}, \dots, x_{i+23}), Y_i = x_{i+24}$$

where $X_i$ represents the input sequence and $Y_i$ denotes the prediction target. As the window continuously moves forward, a large number of training samples can be generated, thereby improving model performance.

In this study, the High-tech Zone air quality monitoring station is selected as the target site for PM2.5 prediction. Data from other monitoring stations are incorporated to construct spatial correlation features, enhancing the model's ability to capture air pollution diffusion patterns.

(2) Evaluation Metrics

To comprehensively evaluate the predictive performance of the model, three metrics are employed: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$). Their formulations are given as follows:

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

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

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

where $y_i$ denotes the observed value, $\hat{y}_i$ represents the predicted value, and $\bar{y}$ is the mean of the observed values. Smaller RMSE and MAE values indicate lower prediction errors, while an $R^2$ value closer to 1 implies better model fitting performance.

## 4.2. Experimental Results Analysis

**Table 3.** Performance comparison of different models for PM2.5 prediction

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| LSTM | 11.23 | 8.38 | 78.7 |
| GCN | 11.13 | 8.87 | 79.08 |
| GCN-LSTM | 9.06 | 6.79 | 86.12 |
| **GCN-PCA-LSTM** | **7.94** | **5.79** | **89.36** |

To verify the effectiveness of the proposed model, several representative models are selected for comparative experiments, including the LSTM model, GCN model, and GCN-LSTM model, alongside the proposed GCN-PCA-LSTM model. The experimental results are presented in Table 3.

As shown in Table 3, there are significant differences in the performance of the models for the PM2.5 prediction task. First, although the standalone LSTM model can capture temporal sequence features, its prediction accuracy is relatively limited due to the lack of spatial information modeling. Second, the GCN model primarily focuses on spatial feature extraction but does not adequately account for temporal dynamics, resulting in only marginal performance improvement.

When GCN is combined with LSTM, the model is able to jointly learn spatial and temporal features, leading to a notable enhancement in prediction performance. For instance, the RMSE of the GCN-LSTM model decreases from 11.23 to 9.06, while the R2 increases to 86.12%. Furthermore, by incorporating PCA-based dimensionality reduction into the GCN-LSTM framework, the prediction performance is further improved. The proposed GCN-PCA-LSTM model achieves an RMSE of 7.94, an MAE of 5.79, and an R2 of 89.36%, outperforming all baseline models in overall prediction accuracy. These results demonstrate that the integration of spatial feature extraction, feature dimensionality reduction, and temporal sequence modeling can effectively enhance PM2.5 prediction performance.

To further evaluate the predictive capability of the model, a comparative analysis between the predicted values and the observed PM2.5 concentrations is conducted. Fig. 3 illustrates the temporal variation trends of both the predicted and actual values.

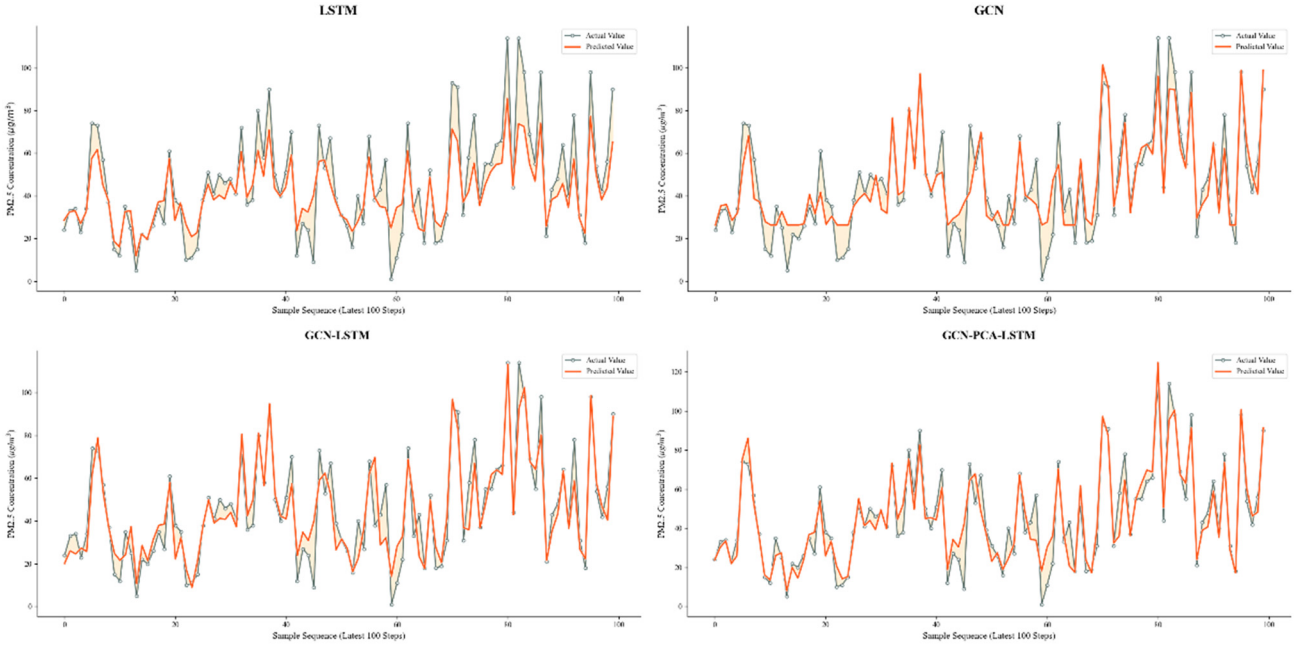


**Fig 3.** Comparison between predicted and observed PM2.5 concentrations

It can be observed from the figure that the GCN-PCA-LSTM model is able to closely track the temporal variation of PM2.5 concentrations. Even during periods of high pollution levels or rapid fluctuations, the predicted results maintain a good fit with the observed values, indicating strong dynamic prediction capability.

## 5. Conclusion

This study addresses the characteristics of PM2.5 concentration variations in air quality prediction, which exhibit both significant temporal dependence and spatial correlation, and proposes a hybrid prediction model integrating a GCN, PCA, and a LSTM. The proposed model first employs GCN to extract spatial correlation features among monitoring stations, followed by PCA to reduce the dimensionality of high-dimensional features, thereby mitigating feature redundancy and improving training efficiency. Finally, LSTM is utilized to model temporal sequence features, enabling effective prediction of PM2.5 concentration trends.

In the experimental analysis, the High-tech Zone air quality monitoring station is selected as the target prediction site. A sliding time window approach is adopted to construct input samples, and the proposed model is compared with several baseline models, including LSTM, GCN, and GCN-LSTM. The results demonstrate that the GCN-PCA-LSTM model achieves superior performance in PM2.5 prediction tasks. Specifically, it outperforms the comparison models in terms of RMSE, MAE, and R2, and is capable of more accurately capturing the dynamic variation trends of PM2.5 concentrations.

The findings indicate that the integration of spatial feature extraction, feature dimensionality reduction, and temporal sequence modeling can effectively enhance the performance of air quality prediction models, providing a feasible technical approach for urban air pollution forecasting and environmental management.

Future research can further incorporate additional meteorological factors and regional pollution transport information, and explore advanced spatiotemporal prediction frameworks based on deep learning models such as Transformers, with the aim of further improving prediction accuracy.

# Acknowledgments

This study was supported by the Youth Research Program of Anhui Xinhua University [Grant Number 2025zrqyb03], the General Program of Natural Sciences of Anhui Xinhua University [Grant Number 2024zr017], the Key Program on Spatial Correlation and Urban Collaborative Governance of Air Pollution in Anhui Province based on Complex Networks [Grant Number 2025rwzda04], the Outstanding Young Teachers Training Key Project [Grant Number YQZD2025092], and the Research on Ozone Disinfection and Real-Time Monitoring Technology for Outdoor Camping Equipment [Grant Number 2025AHGXZK31078].


# References


[1] Orellano, Pablo, et al. "Short-term exposure to particulate matter (PM10 and PM2. 5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis." Environment international 142 (2020): 105876.

[2] Gao, Zhaoqi, and Xuehua Zhou. "A review of the CAMx, CMAQ, WRF-Chem and NAQPMS models: Application, evaluation and uncertainty factors." Environmental Pollution 343 (2024): 123183.

[3] Zhao, Lingxiao, Zhiyang Li, and Leilei Qu. "Forecasting of Beijing PM2. 5 with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition." Heliyon 8.12 (2022).

[4] Zeng, Yongkang, et al. "Air quality forecasting with hybrid LSTM and extended stationary wavelet transform." Building and Environment 213 (2022): 108822.

[5] Ma, Xin, et al. "Time series-based PM2. 5 concentration prediction in Jing-Jin-Ji area using machine learning algorithm models." Heliyon 8.9 (2022).

[6] Bernacki, Jaroslaw, and Rafał Scherer. "A Comprehensive Review of Data-Driven Techniques for Air Pollution Concentration Forecasting." Sensors 25.19 (2025): 6044.

[7] Duan, Jiahui, et al. "Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer." Scientific Reports 13.1 (2023): 12127.

[8] Wang, Jingyang, et al. "An air quality prediction model based on CNN-BiNLSTM-attention." Environment, Development and Sustainability 27.10 (2025): 24705-24720.

[9] Wang, Shuo, et al. "Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting." Proceedings of the 28th international conference on advances in geographic information systems. 2020.