Post-print of Research on PM2.5 Concentration Prediction and Application in Arid Cities Based on Improved Deep Learning Models

Authors: Shengjie Wang, Zhang Qinghong, Mingjian Sang

Date: 2026-02-27T22:08:33+00:00

Abstract

As global urbanization accelerates, PM2.5 pollution in arid cities exhibits strong non-stationarity and complex spatiotemporal characteristics due to unique geographical and climatic conditions, making it difficult for traditional prediction models to effectively capture its dynamic patterns. To address this challenge, a hybrid prediction framework— "Complete Ensemble Empirical Mode Decomposition with Adaptive Noise-Ivy Algorithm-Kolmogorov-Arnold Network-Bidirectional Long Short-Term Memory" (CEEMDAN-IVY-KAN-BiLSTM)—was constructed to enhance the prediction accuracy of PM2.5 concentrations. This framework extracts multi-scale features through a combination of noise-reduction decomposition and parameter optimization, integrating the strong nonlinear fitting and bidirectional temporal modeling capabilities of the KAN-BiLSTM model to effectively improve prediction performance.

The results indicate that: (1) From 2021 to 2024, PM2.5 concentrations in Urumqi exhibited significant seasonal fluctuations, with the mean reaching 41.97 $\mu g \cdot m^{-3}$ in winter due to coal-fired heating and temperature inversion layers, while concentrations dropped to near annual lows in summer due to enhanced atmospheric convection, showing an overall downward trend year by year. (2) Importance ranking of the data reveals that PM2.5 is strongly positively correlated with the Air Quality Index, PM10, CO, and NO2, and negatively correlated with temperature and dew point temperature, indicating that coal-fired emissions, motor vehicle exhaust, and meteorological dispersion conditions are the primary influencing factors; furthermore, the model effectively separated high-frequency fluctuations (such as dust events) from low-frequency trends (seasonal changes) in the PM2.5 sequence, reducing the impact of data non-stationarity. (3) Experiments conducted based on daily air quality data from Urumqi between 2021 and 2024 show that this framework achieved a coefficient

of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) of 0.991, 1.391, and 1.881, respectively, all of which are significantly superior to traditional machine learning and common deep learning models. This validates the applicability of the "decomposition-optimization-integration" deep learning framework for PM2.5 prediction in arid cities.

Full Text

Preamble

ARID LAND GEOGRAPHY Vol. 49 No. 2 Feb. 2026

Prediction and Application of PM2.5 Concentration in Arid Zone Cities Based on an Improved Deep Learning Model

WANG Shengjie, ZHANG Qinghong, SANG Mingjian *(School of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi, Xinjiang, China)*

Abstract: With the acceleration of global urbanization, air pollution in arid zone cities presents strong non-stationarity and complex spatio-temporal characteristics due to unique geographical and climatic conditions, making it difficult for traditional prediction models to effectively capture dynamic patterns. To address this challenge, this study constructs a hybrid prediction framework: "Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) -Ivy Algorithm (IVY) -Kolmogorov-Arnold Network (KAN) -Bidirectional Long Short-Term Memory (BiLSTM)." This framework enhances the prediction accuracy of PM2.5 concentrations by combining multi-scale feature extraction through denoising decomposition and parameter optimization with the strong nonlinear fitting and bidirectional temporal modeling capabilities of the KAN-BiLSTM model. The results indicate that: (1) PM2.5 concentrations in Urumqi showed significant seasonal fluctuations from 2015 to 2024, with an overall downward trend. In winter, the mean concentration reached 41.97 $\mu g \cdot m^{-3}$ due to coal-fired heating and temperature inversions, while in summer, concentrations dropped to annual lows as atmospheric convection intensified. (2) Importance ranking of the data reveals that PM2.5 is highly correlated with the Air Quality Index (AQI) and PM10, and negatively correlated with air temperature and dew point temperature. This suggests that coal emissions, vehicle exhaust, and meteorological dispersion conditions are the primary influencing factors. (3) The CEEMDAN model effectively separates high-frequency fluctuations (such as dust events) from low-frequency trends (seasonal changes) in the PM2.5 sequence, reducing the impact of data non-stationarity. (4) Experiments conducted on daily air quality data from Urumqi demonstrate that this framework achieves R^2, MAE, and RMSE values that significantly outperform traditional machine learning and common deep learning models. This validates the applicability of the "decomposition-ensemble" deep learning framework for PM2.5 prediction in arid zone cities.

Keywords: PM2.5 concentration prediction; CEEMDAN decomposition; IVY optimization algorithm; KAN-BiLSTM model; Deep learning; Arid zone cities

1 Introduction

In recent years, with the acceleration of global urbanization, air quality issues have become increasingly prominent. As a primary pollutant, PM2.5 poses a serious threat to human health [**?**]. Consequently, monitoring and predicting PM2.5 concentrations has become a critical task for air pollution control. Urumqi faces severe air quality challenges due to the combined influence of its natural geographical environment, industrial development, and energy structure [**?**]. To address these challenges, the Urumqi municipal government has issued policies such as the "Urumqi Atmospheric Environment Remediation Three-Year Action Plan," continuously promoting the optimization of industrial, energy, and transportation structures to reduce PM2.5 concentrations and improve overall air quality, thereby supporting high-quality urban development [**?**]. In this context, accurate prediction of PM2.5 concentrations holds significant practical importance: scientific forecasting not only allows for the advance assessment of air pollution levels, providing a quantitative basis for government departments to formulate precise governance measures, but also helps the public take protective measures through early warning information, thereby reducing health risks. Therefore, PM2.5 concentration prediction serves as both scientific support for policy-making and a key factor in safeguarding resident health and improving quality of life [**?**].

PM2.5 concentration prediction is essentially a time-series problem aimed at establishing mathematical models by mining historical temporal information. With the rapid development of artificial intelligence, machine learning models have demonstrated high flexibility and fault tolerance in PM2.5 concentration prediction [**?**]. For example, Kang Junfeng et al. [**?**] used various machine learning models to predict hourly PM2.5 in Ganzhou, Jiangxi Province, showing that these models can effectively predict concentration changes within unit time. Another study [**?**] utilized a Random Forest model combined with training data provided by chemical transport models, though the prediction capability of this framework remained limited with relatively low precision. Similarly, Jiang Yuyan et al. [**?**] used an improved Grey Wolf Optimizer to tune hyperparameters for a Gradient Boosting Decision Tree (GBDT) model to predict PM2.5 in Beijing, further enhancing the accuracy of machine learning models. However, PM2.5 time-series data contain dynamic and nonlinear relationships, and traditional machine learning models have limited capacity to capture long-term temporal dependencies.

Deep learning, with its powerful nonlinear modeling capabilities, can effectively capture complex dynamic relationships in time-series data and has shown excellent performance across various forecasting tasks, making it the preferred method for addressing PM2.5 concentration prediction challenges [**?**]. For instance, Long Short-Term Memory (LSTM) networks and Convolutional Neural

Networks (CNN) can capture temporal and spatial features in data, performing particularly well for long-duration series. Liu En et al. [**?**] proposed a Transformer prediction model integrated with an attention mechanism; however, they did not include meteorological factors such as pressure, temperature, and wind speed as auxiliary variables, nor did they clarify the model' s adaptability to extreme pollution scenarios or broader regions. Gu Kuo et al. [**?**] proposed a composite model to analyze monitoring data from air quality stations in Zibo City, which integrated Grey Relational Analysis prior to model prediction.

Funding Projects: National Natural Science Foundation of China (No. [NUMBER]); National Social Science Fund of China (No. [NUMBER]); Graduate Research and Innovation Project of Xinjiang University of Finance and Economics (24XTJ003, XJUFE2025B011).

About the Authors: WANG Shengjie (1998-), male, PhD candidate, primarily engaged in environmental statistics research in arid zones. **Corresponding Author:** ZHANG Qinghong (1974-), female, PhD, Professor, primarily engaged in resources and environmental statistics research. E-mail: wsj12252025@163.com; zhqh@xjufe.edu.cn.

分析和改进的自适应噪声完备集合经验模态分解

Data preprocessing was performed; however, the model only considered a single direction of the time series, which may lead to the omission of critical information and a subsequent reduction in model accuracy. To address this, an attention mechanism for feature selection was integrated into the architecture to capture dynamic relationships within the data. Furthermore, \cite{REFERENCE_{ID}} proposed a Long Short-Term Memory (LSTM) network based on application strategies, which achieved favorable results; nevertheless, the complexity of the model remains a concern.

1.1 研究区概况

Urumqi is located in northwestern China, situated in the heart of the Eurasian continent. As the capital of the Xinjiang Uyghur Autonomous Region, it lies between 86°37′33″-88°58′24″ E and 42°45′32″-44°08′00″ N, covering a total area of $13,787.9\ \mathrm{km}^2$. The urban area is positioned at the northern foot of the Tianshan Mountains and experiences a mid-temperate continental arid climate characterized by cold, dry winters and hot summers. The topography is primarily basin-like, surrounded by the Tianshan mountain range.

With the advancement of urbanization and industrialization, the issue of $PM_{2.5}$ pollution in Urumqi has become particularly prominent. This is especially evident during the winter heating season, when $PM_{2.5}$ concentrations rise sharply due to coal-fired heating and industrial emissions. Air pollution not only disrupts the daily lives of residents but also poses significant health threats, particularly to children, the elderly, and individuals with respiratory

diseases. To improve air quality, Urumqi has actively implemented comprehensive measures, including reducing coal consumption, optimizing the energy structure, and strengthening pollution control efforts. Although these governance initiatives have made some progress, $PM_{2.5}$ pollution prevention and control still face numerous challenges. Consequently, constructing an efficient and accurate prediction and early-warning system is of great significance for further addressing air pollution, enhancing the effectiveness of governance, and safeguarding public health.

1.2 数据来源

The data used in this study consist of daily observations from Urumqi. Air quality data, specifically $PM_{2.5}$ concentrations and the Air Quality Index (AQI), were obtained from the National Urban Air Quality Real-time Publishing Platform of the China Environmental Monitoring Center. these data provide a comprehensive reflection of pollutant concentrations and air quality conditions within the research area. Meteorological data include key variables such as temperature, air pressure, precipitation, dew point temperature, and wind speed, sourced from the National Climatic Data Center (NCDC).

Urumqi, a typical city in an arid region, is equipped with several national air quality automatic monitoring stations [Figure 1: see original paper]. To address the non-linearity and dynamic complexity inherent in $PM_{2.5}$ concentration prediction, which often leads to significant increases in computational time and resource consumption, this study proposes a hybrid model: "Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) -Ivy Check Optimization Algorithm (ICOA) -Kolmogorov-Arnold Network (KAN) -Bidirectional Long Short-Term Memory (BiLSTM)." Experimental results demonstrate that this model effectively improves the prediction accuracy of $PM_{2.5}$ concentrations. It handles non-linear relationships and spatio-temporal dependencies more efficiently than traditional methods, showing superior performance particularly in complex environments. This provides more precise predictive support for air quality management and policy formulation.

The monitoring stations—including the Midong District Environmental Protection Bureau, No. 31 Middle School, Railway Bureau, Toll Station, Xinjiang Academy of Agricultural Sciences Farm, Hongguangshan Area, Monitoring Station, Xinjiang Normal University Wenquan Campus, Dalu Valley, Training Base, and Dabancheng District Environmental Protection Bureau—are primarily distributed across the urban and suburban areas. These locations cover transportation hubs, residential areas, and local industrial zones, effectively capturing the variation characteristics of air quality in built-up areas. According to pollution source apportionment data, the areas surrounding these stations are dominated by motor vehicle exhaust and winter coal-fired heating emissions, characterizing them as primary source-dominated environments. During winter, temperature inversion phenomena also promote the formation of secondary aerosols. Overall, $PM_{2.5}$ pollution levels are primarily driven by primary emissions.

In the study titled "Research and Application of $PM_{2.5}$ Concentration Prediction in Arid Cities Based on Improved Deep Learning Models" by Wang Shengjie et al., more than ten items were analyzed to achieve a fine-grained resolution of multi-temporal scale features. Given that $PM_{2.5}$ concentration data exhibit significant seasonality and complex non-linear characteristics, traditional analytical methods struggle to accurately capture their patterns of change. However, the CEEMDAN method can effectively separate seasonal trends from short-term fluctuations, clearly extracting variation trends across different time scales.

1.3.2 常春藤优化算法常春藤优化算法（

The Ivygrowth optimization algorithm analyzes the ecological adaptive growth mechanisms of ivy plants through mathematical modeling. By transforming these biological behaviors into efficient search strategies, it provides an innovative roadmap for solving complex optimization problems.

[Figure 1: see original paper] Fig. 1 Distribution of air quality monitoring stations in Urumqi City

As shown in Figure 1, the distribution of air quality monitoring stations in Urumqi ensures that the collected data maintains strong urban representativeness.

1.3.1 自适应噪声完全集合经验模态分解自适应

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) serves as a cutting-edge method for processing nonlinear and non-stationary signals. As an advanced iteration of Empirical Mode Decomposition (EMD), CEEMDAN effectively addresses the mode mixing problem inherent in traditional EMD while significantly reducing the reconstruction error and computational cost associated with standard Ensemble Empirical Mode Decomposition (EEMD).

The core innovation of CEEMDAN lies in the strategic addition of adaptive white noise at each stage of the decomposition process. By calculating a unique residue for each mode, the algorithm ensures that the intrinsic mode functions (IMFs) are more physically meaningful and numerically stable. This approach allows for a more precise extraction of signal components across different frequency scales, making it particularly suitable for complex engineering applications where signal integrity is paramount.

[Figure 1: see original paper]

In the CEEMDAN framework, the decomposition process is defined by the following mathematical formulation. Let $x(t)$ be the original signal and $E_k(\cdot)$ be

the operator that produces the k-th mode obtained by EMD. The first IMF, IMF_1, is calculated as the average of the first EMD modes of the signal plus white noise:

$$\mathrm{IMF}_1 = \frac{1}{N}\sum_{i=1}^{N} E_1(x(t) + \epsilon_0 w^i(t))$$

where $w^i(t)$ represents the i-th realization of white noise and ϵ_0 is the noise amplitude. The corresponding first residue is then defined as $r_1(t) = x(t) - \mathrm{IMF}_1$. For subsequent stages $k = 2, 3, \ldots, K$, the k-th IMF is extracted as:

$$\mathrm{IMF}_k = \frac{1}{N}\sum_{i=1}^{N} E_1(r_{k-1}(t) + \epsilon_{k-1} E_{k-1}(w^i(t)))$$

This iterative process continues until the residue $r_k(t)$ no longer contains enough extrema to be further decomposed. Compared to its predecessors, CEEMDAN provides a more complete decomposition, ensuring that the sum of the IMFs and the final residue perfectly reconstructs the original signal, as

方法，通过引入自适应噪声迭代分解，有效解决了

Traditional Empirical Mode Decomposition (EMD) is often hindered by the issues of mode mixing and boundary effects. This method decomposes the original signal into a set of Intrinsic Mode Functions (IMFs) and a residual component. The operational mechanism of the algorithm is based on the following core modules: First, a population of ivy individuals is initialized using a randomization strategy, where the coordinates of each individual in the solution space are mapped to candidate solutions of the optimization problem, thereby constructing an initial search sample set. Second, a dynamic growth model is constructed using differential equations to accurately characterize the adaptive growth dynamics of ivy individuals based on environmental feedback. Third, an information interaction mechanism among individuals is established based on a topological neighborhood structure. This simulates the phototropism and climbing growth characteristics of ivy, achieving directional search and spatial diffusion of the population through local optimal guidance. Fourth, the algorithm iteratively executes growth simulations and position updates until a preset convergence criterion is met (such as reaching the maximum number of iterations or the rate of change in the fitness function falling below a specific threshold).

Fifth, the fitness of individuals within the population is evaluated using an objective function. Through an elite retention strategy, the principle of survival of the fittest is implemented, driving the population to progressively converge toward the global optimal solution during the evolutionary process.

1.3.3 Kolmogorov-Arnold 网络

is a novel neural network architecture based on the Kolmogorov-Arnold theorem (Figure [Figure 2: see original paper]). This model replaces the fixed activation functions found in traditional Multi-Layer Perceptron (MLP) frameworks with learnable spline functions. By representing the network as a linear combination of curves, the spline function formula is defined as follows:

The multilayer perceptron (MLP) architecture has been reconstructed by replacing fixed weights with optimizable univariate spline functions.

The core of this approach lies in decomposing multidimensional functions into a combination of univariate functions, thereby balancing high expressive power with excellent interpretability. Furthermore, the network allows activation functions to be dynamically adjusted during the training process to flexibly capture complex nonlinear relationships. By leveraging this characteristic, the model can efficiently fit complex structures with fewer parameters. It outperforms traditional architectures in both nonlinear representation and computational efficiency, fully demonstrating the theoretical and practical value of the Kolmogorov-Arnold representation theorem. The relevant formulas for the model are as follows:

Kolmogorov-Arnold Networks and Residual Activation Strategies

The Kolmogorov-Arnold representation theorem states that a multivariate continuous function can be represented as a finite composition of continuous functions of a single variable and the addition operation. Specifically, the general form is expressed as:

$$f(x) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$$

In this expression, x_n represents an n-dimensional input vector, Φ_q denotes the outer nonlinear activation functions, and $\phi_{q,p}$ are the univariate functions that map the input variables.

To improve the stability and representational power of these networks, researchers have proposed a residual activation strategy. The formulation for this approach is as follows:

$$\text{act}(x) = w \cdot b(x) + \phi(x)$$

In this context, $\phi(x)$ represents a learnable univariate activation function, while $b(x)$ serves as a basis function for a residual connection. The term w denotes the weight assigned to this basis function.

Furthermore, the spline component of the activation function, often denoted as spline(x), is typically constructed using a linear combination of B-splines:

$$\text{spline}(x) = \sum_i c_i B_i(x)$$

Where c_i represents the i-th weight coefficient and $B_i(x)$ denotes the i-th B-spline basis function, with $i = 1, 2, \ldots, k$. This structure allows the model to learn complex, smooth univariate mappings by adjusting the coefficients of the spline curves.

1.3.4 双向长短期记忆网络双向长短期记忆网络

Bidirectional Long Short-Term Memory (BiLSTM) is an extension of the traditional Long Short-Term Memory (LSTM) network, designed to simultaneously capture both forward and backward temporal dependencies within sequential data.

By introducing the forgetting gate, input gate, and output gate, LSTM effectively addresses the gradient vanishing problem encountered by traditional Recurrent Neural Networks (RNNs) during long-sequence learning. However, a unidirectional LSTM layer can only utilize forward information, failing to fully capture the complete context. BiLSTM builds upon this by adding a backward layer, allowing the input sequence to be processed in both chronological and reverse chronological order. The outputs from both directions are then concatenated to obtain more comprehensive contextual information. Compared to a standard LSTM, BiLSTM possesses stronger feature representation and contextual information capture capabilities, making it particularly suitable for processing complex temporal dependencies and non-stationary sequence data.

The model formulas are summarized as follows: $h_t = [W_h h_t + \ldots]$. In the above formulas, h_t represents the hidden state of the bidirectional LSTM at time step t; h'_t is the hidden state at time step t in the backward pass; $b(x)$ and $spline(x)$ are spline functions; and w represents the weights of the spline function.

The term is redundant because it can be absorbed into $b(x)$ and $spline(x)$ by the BiLSTM. However, the formula still includes these factors to provide better control over the overall magnitude of the activation function.

The basis function formula in the above equation is as follows: $b(x) = \text{silu}(x) = \frac{x}{1+e^{-x}}$. Here, h_t is the backward state; h'_t is the state in the backward pass; and x is the input variable.

In most cases, σ represents the Sigmoid activation function; e is the natural constant; $spline(x)$ is parameterized accordingly; y_t is the final output of the output layer at different time steps t; W_h is the weight matrix of the output layer; b_h is the bias vector of the output layer; and σ is the activation function.

Note: X represents the elements in the input sequence; y represents the output elements; h_t is the hidden state of the bidirectional LSTM at time step t; T is the final time step. The specific meanings of the letters are the same as those in the equations.

[Figure 3: see original paper] BiLSTM model structure

Wang Shengjie et al.: Research on PM2.5 Concentration Prediction and Application in Arid City Based on Improved Deep Learning Model

1.4.1 CEEMDAN-IVY-KAN-BiLSTM 预测模型的

Construction of the Prediction Model

To further improve the accuracy of concentration prediction and reduce the impact of time-series fluctuations, this study proposes a robust modeling framework. By leveraging advanced machine learning techniques, we aim to capture the complex non-linear relationships inherent in the data. The integration of deep learning architectures allows for the extraction of high-level features from raw input sequences, which is essential for handling the stochastic nature of concentration levels over time.

Optimization of Time-Series Forecasting

The primary objective is to minimize the error metrics associated with long-term dependencies in the temporal data. Traditional methods often struggle with the vanishing gradient problem when processing extended sequences; therefore, we implement specialized recurrent structures or attention mechanisms to maintain information flow. By refining the loss functions and incorporating regularization techniques, the model can achieve superior generalization performance, ensuring that the predicted concentrations remain stable even under varying environmental conditions.

2 结果与分析

To address the non-stationarity and complexity of $PM_{2.5}$ concentration data, this paper constructs an ensemble deep learning prediction model. The methodology begins with comprehensive data preprocessing, including data cleaning, missing value imputation, and standardization of both concentration and meteorological data. Subsequently, the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is employed to decompose the $PM_{2.5}$ sequence into several Intrinsic Mode Functions (IMFs) and a residual component, effectively separating high-frequency and low-frequency features.

To optimize the model performance, hyperparameters are tuned for each decomposed component. These components are then fed into the prediction architecture, which leverages Bi-directional Long Short-Term Memory (BiLSTM) networks. This approach utilizes adaptive univariate spline functions to capture

complex non-linear features, while the bidirectional structure of the BiLSTM extracts long-term temporal dependencies. Finally, the individual predictions from each component are weighted and reconstructed to obtain the overall predicted value for $PM_{2.5}$ concentration.

1.4.2 模型评价指标本文选择均方根误差（

To evaluate the performance of the models, we utilize Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) to measure the discrepancy between the predicted values and the original data. Additionally, the Coefficient of Determination (R^2) is selected to evaluate the fitting effect of the model on the target variable. Generally, smaller values for RMSE and MAE indicate superior predictive performance, while an R^2 value closer to 1 signifies a better model fit. The calculation formulas are as follows:

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

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

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

In these equations, y_i represents the true value of the i-th $PM_{2.5}$ sample; $\hat{y}_i$ represents the predicted value of the i-th $PM_{2.5}$ sample; N is the total sample size; and $\bar{y}$ is the mean of the true $PM_{2.5}$ values.

2.1 数据统计特征分析

The selected data dimensions are $(1277, 13)$, indicating that the dataset consists of 1,277 rows and 13 columns (including the time column). In Urumqi, the $PM_{2.5}$ mass concentration ranges from 4.9 to $316.68\mu g \cdot m^{-3}$, with a mean value of $41.97\mu g \cdot m^{-3}$. The daily variation trend of $PM_{2.5}$ concentration in Urumqi exhibits distinct seasonal fluctuations and a long-term decreasing characteristic [FIGURE:N]. Every winter (December to February) serves as a period of concentrated high values, with typical peaks occurring in January, where concentrations have exceeded $250\mu g \cdot m^{-3}$. This reflects the dual influence of coal-fired emissions and temperature inversion phenomena during the heating season, which cause particulate matter to accumulate easily near the ground and form heavy pollution events. In contrast, concentrations during the summer (June to August) remain at the annual low, with fluctuations mostly ranging between 15 and $30\mu g \cdot m^{-3}$. This is attributed to enhanced atmospheric convection under high-temperature conditions and relatively weakened pollution emissions, which significantly improve dispersion conditions. Additionally, the

figure shows a small number of short-term abnormal peaks during periods such as April and May, which may be related to unfavorable seasonal meteorology or natural disturbance events. However, as the current dataset does not include records for dust storms or specific meteorological events, these factors cannot yet be quantitatively incorporated into the analysis.

From the perspective of the overall trend, $PM_{2.5}$ concentrations have shown a year-on-year decline since 2019. The concentration levels in 2022 decreased significantly compared to the same period in 2019, indicating that Urumqi has achieved phased governance results in recent years regarding energy structure adjustment, industrial emission control, and the implementation of environmental policies.

2.2.1 数据预处理为提升模型性能与泛化能力，

The raw data must undergo preprocessing to ensure quality and consistency. First, for time-series data exhibiting seasonality and strong temporal correlation, forward and backward filling methods are employed to handle missing values, thereby ensuring the continuity of the sequence. Second, an autoencoder-based neural network model is constructed to detect outliers by calculating reconstruction errors, using a threshold defined as the mean plus two standard deviations. These outliers are then replaced with smoothed values calculated via a sliding window moving average. Finally, the data features are mapped to a unified range (such as $[0, 1]$ or standardization) through normalization to eliminate dimensional effects, enhancing feature comparability and the convergence stability of the model.

The normalization formula is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x represents a specific data sample in the original dataset, $\min(x)$ denotes the minimum value of the entire original dataset, and $\max(x)$ denotes the maximum value of the entire original dataset.

[Figure 4: see original paper]

Figure 4 illustrates the changes in $PM_{2.5}$ concentration. The analysis suggests that under unfavorable meteorological conditions—such as temperature inversions and cold, low-humidity environments—pollutants are more likely to accumulate. Temperature inversions inhibit vertical atmospheric dispersion, while low temperatures are typically associated with increased heating activities; these dual mechanisms jointly drive the rise in pollution levels. Some variables (such as $PM_{2.5}$ and certain meteorological factors) exhibited low correlation coefficients and failed to pass significance tests. Consequently, the final selection of input features for model construction was based on two criteria: correlation strength and statistical significance.

2.2.2 数据相关性分析

$PM_{2.5}$ concentrations are susceptible to fluctuations in the concentrations of other pollutants and exhibit significant correlations with meteorological factors. In this study, Python software was employed to generate heatmaps to analyze the correlations between $PM_{2.5}$ levels and various influencing factors [Figure 1: see original paper], supplemented by statistical tests on the dataset . The heatmaps utilize color intensity to represent the complex interactive relationships between $PM_{2.5}$ and various meteorological factors.

First, there is a significant correlation with $PM_{2.5}$ concentrations; second, there are clear environmental mechanistic explanations, such as source emission characteristics or meteorological dispersion conditions. This strategy ensures the sufficiency of the model' s input information while effectively controlling for dimensional redundancy and the risks of multicollinearity. Consequently, it establishes a foundation for constructing a predictive model that possesses both physical rationality and robust generalization capabilities.

2.2.3 CEEMDAN 分解为降低

To enhance prediction performance by addressing the volatility and complexity of the PM2.5 concentration time series, a decomposition-based processing approach is employed. Given that the original sequence is non-stationary and exhibits significant oscillation amplitudes, it is necessary to decompose the data into more manageable components.

方法将其分解为多个

Complexity decreases as the high-frequency, medium-frequency, and low-frequency components, along with the trend component derived from CEEMDAN, intuitively present the strength of correlations. These components respectively capture short-term fluctuations, medium-term cycles, and long-term trends. This decomposition clearly analyzes the multi-time scale characteristics of the original sequence, providing a rich set of features for subsequent modeling, which helps improve the precision and robustness of the predictive model.

According to the statistical test results, there is a significant linear correlation between $PM_{2.5}$ and certain pollutants as well as meteorological factors. Specifically, $PM_{2.5}$ exhibits a high degree of positive correlation, indicating that these variables maintain consistency in their numerical trends.

2.3 模型结果与分析

To verify the superior predictive performance of the CEEMDAN-BiLSTM model, this study utilizes various machine learning models, such as Random Forest, Gradient Boosting, AdaBoost, XGBoost, and CatBoost. These models

are used to analyze the significant factors influencing particulate matter concentration changes in Urumqi. Among these, indicators of incomplete combustion products and transportation emissions—specifically gases associated with coal-fired emissions—align with the urban characteristics of Urumqi, where winter heating relies heavily on coal and natural gas, and motor vehicle exhaust emissions are frequent.

$PM_{2.5}$ exhibits a significant negative correlation with linear regression, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Trees, Random Forest, AdaBoost, XGBoost, and CatBoost. In the study titled "Research on $PM_{2.5}$ Concentration Prediction and Application in Arid Cities Based on an Improved Deep Learning Model" by Wang Shengjie et al., the variables analyzed include Air Quality Index (AQI), temperature, atmospheric pressure, wind speed, precipitation, and dew point temperature.

[Figure 5: see original paper] Heat map analysis Statistical test Note: * and ** represent significance at the 5% and 1% levels, respectively.

A comparative analysis of the R^2 values and performance of the proposed model on the dataset was conducted. To prevent model overfitting caused by multivariate inputs, the following strategies were implemented during the training process: regularization terms (weight penalties) were introduced into the deep learning models; a Dropout mechanism was employed to randomly mask a portion of neurons in each training round; and an early stopping mechanism was used to prevent performance degradation on the validation set. All input features were standardized to ensure data consistency during training. To ensure a fair comparison, all machine learning models underwent the same data preprocessing workflow prior to training. Note: IMF refers to Intrinsic Mode Functions.

[Figure 6: see original paper] Results of CEEMDAN decompositions The study employed a training and testing set strategy based on chronological order, combined with 5-fold time-series cross-validation to evaluate model stability. This approach ensures the scientific rigor of the training process and the generalization capability of the predictive performance (Table). Box plots of R^2 (Figure [Figure 7: see original paper]) and Taylor diagrams of the performance evaluation indicators (Figure [Figure 8: see original paper]) demonstrate that the proposed model significantly outperforms traditional machine learning and deep learning models in terms of predictive accuracy and fitting ability. Compared to traditional models like CatBoost, XGBoost (61.18%), and Random Forest (54.93%), the CEEMDAN-BiLSTM model shows prominent advantages.

Note: CEEMDAN-IVY-KAN-BiLSTM refers to the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise combined with the Ivy-optimized Kolmogorov-Arnold Network and Bidirectional Long Short-Term Memory network; CEEMDAN-IVY-KAN-LSTM refers to the same architecture using a standard LSTM; IVY-KAN-BiLSTM refers to the Ivy-optimized Kolmogorov-Arnold Bidirectional Long Short-Term Memory network; CNN-LSTM refers to the Convolutional Neural Network-Long Short-Term Memory network; Grid-

KAN-BiLSTM refers to the grid-search optimized Kolmogorov-Arnold Bidirectional Long Short-Term Memory network; Linear Regression, KNN, SVR, Decision Tree, and Gradient Boosting Model are standard baseline models.

[Figure 7: see original paper] Five-fold cross-validation chart [Figure 8: see original paper] Taylor diagram of performance evaluation indicators Indicators for evaluating the performance of different models Deep learning models performed better across all indicators, suggesting they can effectively capture complex relationships within the data, providing higher predictive accuracy and stability. According to the Taylor diagram analysis in Figure [Figure 8: see original paper], where the horizontal and vertical axes represent the standard deviation (indicating the dispersion of predicted values), a smaller standard deviation signifies more stable results. The arcs represent the correlation coefficient; the closer the arc is to the upper right corner, the stronger the correlation between predicted and actual values. The CEEMDAN-BiLSTM model proposed in this study demonstrates a significant advantage in capturing time-series features for $PM_{2.5}$ prediction. By introducing the Ivy optimization algorithm and constructing a hybrid deep learning framework, this study validates the applicability of the CEEMDAN-BiLSTM model for $PM_{2.5}$ forecasting in arid urban regions.

The CEEMDAN decomposition strategy effectively addresses the strong non-stationarity and multi-scale fluctuation issues of $PM_{2.5}$ concentration sequences in Urumqi (such as sudden concentration changes caused by winter cooling and frequent dust storms). Separating high- and low-frequency components significantly enhances the model' s adaptability to complex meteorological conditions. The hybrid architecture, which integrates BiLSTM and KAN, possesses both the ability to capture long-term and short-term dependencies and the flexibility of learnable activation functions for non-linear expression. This is particularly suitable for modeling the coupling relationship between $PM_{2.5}$ and meteorological factors such as local circulation and inversion layers in arid regions. Furthermore, the global hyperparameter optimization tailored to the characteristics of arid region data (e.g., seasonal dust inputs and stable winter weather) provides greater accuracy than grid search, avoiding the local optima traps of traditional parameter tuning methods in complex meteorological scenarios.

Note: * represents the ablation models of the proposed model; R^2 is the coefficient of determination; MAE is the Mean Absolute Error; RMSE is the Root Mean Square Error. Model names are consistent with Figure [Figure 7: see original paper].

The CEEMDAN-BiLSTM model proposed in this paper exhibits excellent predictive stability. Its standard deviation is near zero, significantly outperforming other models such as CatBoost, Gradient Boosting, Random Forest, XGBoost, and standard LSTM/BiLSTM variants. This indicates that the dispersion of the model' s prediction results is extremely low, demonstrating high reliability and a strong correlation between predicted and actual values.

3 讨论

Urumqi, the capital of the Xinjiang Uyghur Autonomous Region, is geographically situated in an arid region significantly influenced by its topography (such as the Tianshan Mountains) and climate (including harsh winters and extreme weather conditions). These factors lead to substantial fluctuations in $PM_{2.5}$ concentrations. During the winter months, Urumqi relies heavily on coal-based heating; the resulting concentration of energy consumption and pollutant emissions during the heating season causes $PM_{2.5}$ levels to rise sharply. Conversely, the summer season is typically characterized by strong convective weather, which effectively disperses ground-level pollutants and reduces $PM_{2.5}$ concentrations. Predicting $PM_{2.5}$ levels is a complex time-series problem involving non-linear, non-stationary, and multi-scale dynamic characteristics. Traditional models, such as multiple regression [**?**], show limited performance when facing complex time-series issues. Furthermore, standard machine learning models like XGBoost [**?**] and certain hybrid machine learning models are often unable to adequately handle sequential modeling, high-frequency signal capture, and spatial correlations.

However, deep learning models [**?**] have demonstrated significant advantages due to their powerful non-linear modeling capabilities. As the economic center of the Xinjiang Uyghur Autonomous Region and a "bridgehead" for the "Belt and Road" initiative, Urumqi has experienced rapid economic growth that has accelerated regional industrialization and urbanization. With this rapid development, industrial emissions—particularly from the energy, metallurgical, and chemical sectors—along with transportation emissions, have become primary sources of $PM_{2.5}$. Nevertheless, analysis indicates that $PM_{2.5}$ concentrations in the Urumqi area are currently showing a downward trend. This suggests that government policies aimed at improving public health, enhancing the living environment, and promoting sustainable economic development have been effective. These efforts have successfully improved the overall well-being of society and enhanced the regional environment.

4 结论

In 2022, $\mathrm{PM}_{2.5}$ concentrations in Urumqi exhibited significant seasonal fluctuations. During the winter, influenced by coal-fired heating and temperature inversion layers, the mean $\mathrm{PM}_{2.5}$ concentration reached 41.97 $\mu\mathrm{g}\cdot\mathrm{m}^{-3}$. Conversely, in the summer, enhanced convection improved dispersion conditions, causing concentrations to drop to their annual minimum. Overall, $\mathrm{PM}_{2.5}$ levels have shown a downward trend year by year, with a decreasing frequency of extreme high-value events (> 250 $\mu\mathrm{g}\cdot\mathrm{m}^{-3}$), indicating that government mitigation measures—such as the optimization of the energy structure—have begun to yield results. Through heatmap analysis and model feature importance ranking, $\mathrm{PM}_{2.5}$ was found to be strongly positively correlated with dew point temperature and negatively correlated with air temperature, suggesting that coal emissions, vehicle exhaust, and meteorological dispersion conditions are the primary

influencing factors.

The CEEMDAN decomposition effectively separated the high-frequency fluctuations of the $PM_{2.5}$ series (associated with extreme weather such as dust storms) from the low-frequency trends (seasonal variations). As discussed in Wang Shengjie et al.: *Research on Prediction and Application of $PM_{2.5}$ Concentration in Arid Cities Based on Improved Deep Learning Models*, global hyperparameter tuning (e.g., number of units, learning rate) enhanced the model' s adaptability to the complex meteorological conditions characteristic of arid regions.

In the $PM_{2.5}$ prediction for Urumqi, the CEEMDAN-based model proposed in this study achieved the best performance, with an R^2 of 0.787, an $RMSE$ of 19.159, and an MAE of 11.051. This represents a significant improvement over traditional machine learning models such as XGBoost ($R^2 = 0.706$, $RMSE = 21.711$) and standard deep learning models like BiLSTM.

Ablation experiments demonstrated that the bidirectional temporal modeling capability of the BiLSTM network and the adaptive activation functions improved the R^2 by 54.93% and 40.36%, respectively, compared to traditional architectures. This underscores the advantages of the hybrid architecture in modeling complex non-linear relationships. Finally, the results validate that the "decomposition-integration" deep learning framework proposed in this study is highly effective for $PM_{2.5}$ prediction in arid urban environments.

参考文献（References）

Shi Yu, Wu Di, Yilihamu Yilipa, et al. Construction of an air quality health index for respiratory diseases in Urumqi [J]. Journal of Environmental and Occupational Medicine, 2024, 41(3): 276-281. Yu Zhixiang, Yu Xiaojing, Li Xia, et al.

Spatiotemporal variations of $PM_{2.5}$ in Xinjiang during 2000–2022 revealed by the China High Air Pollutants (CHAP) reanalysis dataset [J]. Journal of Arid Land Resources and Environment, 2025, 39(4): 132-144. Yan Jinye, Ma Zhengquan, Sun Xuanxuan, et al. Spatiotemporal variations and potential source analysis of $PM_{2.5}$ in the "Urumqi-Changji-Shihezi" urban agglomeration [J]. Arid Land Geography, 2025, 48(3): 405-420.

Spatiotemporal variations and potential sources of $PM_{2.5}$ and PM_{10} in the Urumqi-Changji-Shihezi urban agglomeration from 2015 to 2023 [J]. Arid Land Geography, 2025, 48(3): 405-420. Palida Yahefu. Research on the variation characteristics and potential sources of near-surface $PM_{2.5}$ in Urumqi and Kashgar [J]. Arid Zone Research / Acta Scientiae Circumstantiae, 2025, 45(1): 388-399. Wang Jiamin, Yang Wenzhu, Jiao Yan, et al. Regional transport and source apportionment of $PM_{2.5}$ in urban agglomerations along the Yellow River Basin in Inner Mongolia [J]. China Environmental Science, 2025, 45(10): 5338-5356.

Wang Jiamin, Yang Wenzhu, Jiao Yan, et al. Regional transport and source

analysis of $PM_{2.5}$ in urban agglomerations along the Yellow River Basin in Inner Mongolia [J]. China Environmental Science, 2025, 45(10): 5338-5356. Pan Mengyao, Ren Ying, Wang Siyuan, et al. Prediction of $PM_{2.5}$ and ozone concentration in Shijiazhuang and analysis of influencing factors based on gradient boosting algorithm and SHAP [J]. Acta Scientiae Circumstantiae, 2024, 44(7): 402-409. Yang Guoliang, Yu Huasheng, Huang Cong. $PM_{2.5}$ concentration prediction based on an optimization combination model [J]. Computer Engineering and Design, 2023, 44(10): 3132-3137. Kang Junfeng, Huang Liexing, Zhang Chunyan, et al. Hourly $PM_{2.5}$ prediction and comparative analysis under multiple machine learning models [J]. China Environmental Science, 2020, 40(5): 1895-1905. [10] Bi J Z, Knowland K E, Keller C A, et al. Combining machine learning and numerical simulation for high-resolution $PM_{2.5}$ concentration forecast [J]. Environmental Science & Technology, 2022, 56(3): 1544-1556.

Application of an improved Grey Wolf Optimizer to optimize GBDT in $PM_{2.5}$ concentration prediction [J]. Journal of Safety and Environment, 2024, 24(4): 1569-1580. Jiang Yuyan, Fu Jie, Gan Rumeijiang, et al. Application of improved Grey Wolf Algorithm to optimize GBDT in $PM_{2.5}$ prediction [J].

Journal of Safety and Environment, 2024, 24(4): 1569-1580.] [12] Zhou S, Wang W, Zhu L, et al. Deep learning architecture for PM2.5 concentration prediction: A review[J]. Environmental Science and Ecotechnology, 2024: 100400, doi: 10.1016/j.ese.2024.100400.

Zhu et al. [**?**] proposed a prediction model for $PM_{2.5}$ in subway stations based on an attention mechanism and CNN-ILSTM. Ma et al. [**?**] analyzed the characteristics and potential sources of surface O_3 variation in Urumqi and Kashgar. Furthermore, Mei et al. [**?**] conducted an analysis of the variation trends and influencing factors of $PM_{2.5}$ and O_3 (8-hour) on the northern slope of the Tianshan Mountains from 2015 to 2023. Peng et al. [**?**] developed a $PM_{2.5}$ concentration prediction model based on deep learning and random forests. Additionally, Fan and Xuan [**?**] established a prediction model for $PM_{2.5}$ mass concentration using neural networks.

Research has also explored $PM_{2.5}$ recurrent prediction networks combined with Transformer attention mechanisms [**?**]. Ghasemi et al. [**?**] introduced the Ivy algorithm, an optimization method based on the smart behavior of plants, for engineering applications.

Gu et al. [**?**] investigated $PM_{2.5}$ concentration prediction using a composite LSTM model. Liu et al. [**?**] recently introduced Kolmogorov-Arnold Networks (KAN) as a novel architecture for neural computing.

Xie et al. [**?**] focused on the prediction of $PM_{2.5}$ concentration in Xi'an based on a CEEMDAN-BiLSTM model. Shi et al. [**?**] developed an improved attention-based integrated deep neural network for $PM_{2.5}$ concentration prediction, while Lin et al. [**?**] applied LSTM-based strategies for short-term $PM_{2.5}$ prediction in urban environments. Finally, Zhang et al. [**?**] conducted research on the characteristics and prediction models of $PM_{2.5}$ in Urumqi.

Temporal and spatial distribution characteristicsand prediction scitotenv.2023.167892.

Yang Changchun, Nie Qianqian. Fusion model of time series decomposition and machine learning for PM2.5 forecasting [J]. Journal of Safety and Environment, 2023, 23(12): 4600-4608. Yang Changchun, Nie Qianqian. Fusion model of PM2.5 in Urumqi [J]. Journal of Shihezi University (Natural Science), 2020, 38(5): 648-654. Zhang Shijie, Dou Yan. Application of feature analysis based on XGBoost and SHAP models in PM2.5 concentration prediction [J]. Journal of Chifeng University (Natural Science Edition), 2022, 38(12): 10-17. Wei Jiang, Zhao Caixin, Wang Guohua, et al. Characteristics and sources of water-soluble ion components in PM2.5 in the urban area of Urumqi City [J]. Arid Land Geography, 2025, 48(4): 623-631.

Wei Jiang, Song Dandan, Zhao Lili, et al. Analysis of carbon component pollution characteristics and sources of PM2.5 in Urumqi City [J]. Arid Zone Research, 2024, 41(8): 1323-1330. Ding Chengliang, Zheng Hongbo. Study of PM2.5 concentration prediction model based on improved machine learning [J]. Journal of Dalian University of Technology, 2024, 64(4): 353-360. Li Mingming, Wang Xiaolan, Yue Jiang, et al. PM2.5 prediction based on EOF decomposition and CNN-LSTM neural network [J]. Environmental Science, 2025, 46(2): 715-726. [23] Torres M E, Colominas M A, Schlotthauer G, et al. A complete ensemble empirical mode decomposition with adaptive noise [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Prediction and Application of Urban PM2.5 Concentration in Arid Zones Based on Improved Deep Learning Models

WANG Shengjie, ZHANG Qinghong, SANG Mingjian *(School of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumqi 830012, Xinjiang, China)*

Abstract: With the acceleration of global urbanization, severe PM2.5 pollution in arid zone cities—driven by unique geographical and climatic conditions—exhibits strong non-stationarity and complex spatiotemporal characteristics. These factors make it difficult for traditional prediction models to effectively capture dynamic patterns. To address this challenge, this study proposes a hybrid prediction framework: "Complete Ensemble Empirical Mode Decomposition with Adaptive Noise - Ivy Optimization Algorithm - Kolmogorov-Arnold Network - Bidirectional Long Short-Term Memory" (CEEMDAN-IVY-KAN-BiLSTM). This framework aims to enhance the prediction accuracy of PM2.5 concentrations by jointly extracting multi-scale features through noise reduction decomposition and parameter optimization. It integrates the robust nonlinear fitting capabilities of the KAN with the bidirectional time-series modeling of the BiLSTM, effectively improving overall prediction performance.

The results reveal that PM2.5 concentrations in Urumqi from 2021 to 2024 show significant seasonal fluctuations. Concentrations average $41.97 \mu g/m^3$ in winter due to coal heating and the influence of temperature inversion layers,

dropping to 14.04$\mu g/m^3$ in summer due to enhanced atmospheric convection. Furthermore, an overall annual decreasing trend was observed. Feature importance analysis indicates that PM2.5 is significantly positively correlated with the Air Quality Index (AQI), PM10, CO, and NO2, and negatively correlated with temperature and dew point temperature. This suggests that coal emissions, vehicle exhaust, and meteorological diffusion conditions are the primary influencing factors. The model effectively separates high-frequency fluctuations (such as dust storm events) from low-frequency trends (seasonal changes) in the PM2.5 sequence, reducing the impact of data non-stationarity. Finally, experiments based on daily air quality data in Urumqi from 2021 to 2024 demonstrate that the proposed model achieves a coefficient of determination (R^2) of 0.991, a Mean Absolute Error (MAE) of 1.391, and a Root Mean Square Error (RMSE) of 1.881. These results significantly outperform conventional machine learning and standard deep learning models, verifying the applicability of the "decomposition-optimization-integration"deep learning framework for air quality prediction in arid zone cities.

Keywords: PM2.5 concentration prediction; CEEMDAN decomposition; IVY optimisation algorithm; KAN- BiLSTM model; deep learning; urban PM2.5 concentration in arid zones

Note: Figure translations are in progress. See original paper for figures.