

Article

An Approach to Spatiotemporal Air Quality Prediction Integrating SwinLSTM and Kriging Methods

Jiangquan Xie ¹, Fan Liu ², Shuai Liu ¹ and Xiangtao Jiang ^{1,*} 

¹ College of Computer Science and Mathematics, Central South University of Forestry and Technology, Changsha 410004, China; 20221200532@csuft.edu.cn (J.X.); 20231100431@csuft.edu.cn (S.L.)

² Gansu Electric Power Changle Power Generation Co., Ltd., Lanzhou 730000, China; liufan@csuft.edu.cn

* Correspondence: xtjiang@csuft.edu.cn

Abstract: Air pollution has become a major environmental issue, posing severe threats to human health and ecosystems. Accurately predicting future regional air quality is crucial for effective air pollution control and management strategies. This study proposes a novel deep learning-based approach. First, Kriging interpolation was applied to meteorological indicators such as temperature, humidity, and wind speed, as well as climate-altering gas indicators like CO₂, SO₂, and NO₂ recorded at monitoring stations to obtain their spatial distributions over the entire region. Subsequently, a long short-term memory neural network (SwinLSTM) incorporating Swin Transformer feature extraction was employed to learn the correlations from regional meteorological data and historical air quality records. This model overcomes the limitation of traditional CNNs by capturing long-range spatial dependencies when processing two-dimensional meteorological data through its sliding window attention mechanism. Ultimately, it outputs air quality predictions in both spatial and temporal dimensions. This study collected data from 29 stations across four cities surrounding China's Dongting Lake for experimentation. Predictions for PM_{2.5} and PM₁₀ levels over the entire lake area were made for 1, 6, and 24 h. The results demonstrate that the proposed SwinLSTM architecture significantly outperforms the current mainstream ConvLSTM architecture, with an average R-squared improvement of 5%, establishing a new state-of-the-art model for spatiotemporal air quality prediction.

Keywords: air quality prediction; deep learning; swin transformer; LSTM; Kriging interpolation



Academic Editor: Giovanni Gualtieri

Received: 18 February 2025

Revised: 14 March 2025

Accepted: 17 March 2025

Published: 25 March 2025

Citation: Xie, J.; Liu, F.; Liu, S.; Jiang, X. An Approach to Spatiotemporal Air Quality Prediction Integrating SwinLSTM and Kriging Methods. *Sustainability* **2025**, *17*, 2918. <https://doi.org/10.3390/su17072918>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The spatiotemporal prediction of air pollution has long been a major topic in environmental science and atmospheric chemistry research [1]. With rapid urbanization and industrial development, air pollution issues have become increasingly severe, posing serious threats to human health and ecological environments [2]. Therefore, accurately forecasting future air quality levels in specific regions through indicators such as particulate matter (PM_{2.5}), inhalable particles (PM₁₀), and the comprehensive Air Quality Index (AQI), which reflects the pollution status of an area, holds significant value for government authorities to implement targeted measures and raise public environmental awareness [3].

Traditional air quality prediction methods primarily include numerical and statistical approaches. Numerical methods are based on physical descriptions of atmospheric chemical processes and meteorological circulations, employing numerical simulations and differential equation solvers to predict future pollutant concentrations. For instance, Ghude et al. [4] proposed a coupled regional meteorology–atmospheric chemistry online model

(WRF-Chem) that describes the feedback between atmospheric aerosols and meteorology through mathematical formulations, enabling the prediction of future PM_{2.5} levels using various ecosystem emission data. Statistical methods, on the other hand, establish models that characterize the correlations between historical data and air pollutant concentrations for regression-based predictions. Ni et al. [5] used multi-source meteorological data for correlation analysis to determine the relationships among different indicators and applied an Autoregressive Integrated Moving Average (ARIMA) time series model for PM_{2.5} time series forecasting. Wang et al. [6] utilized two classic machine learning techniques, Random Forest (RF) and Support Vector Machines (SVMs), to achieve hybrid predictions of daily PM₁₀ and SO₂ concentrations. Although these traditional methods have practical applications, they generally assume known emission sources and atmospheric circulation conditions and struggle to handle complex non-linear problems, limiting their prediction accuracy and spatiotemporal resolution [7].

In recent years, deep learning techniques have gained widespread application in spatiotemporal data analysis, demonstrating powerful modeling and prediction capabilities. An increasing number of studies have attempted to apply deep learning techniques to spatiotemporal air pollution prediction problems. For example, Chen et al. [8] proposed an air quality prediction model integrating dual long short-term memory (LSTM) networks with a sequence-to-sequence approach, outperforming various classic machine learning models, including Support Vector Regression (SVR), Ridge Regression, and XGBoost. Gurumoorthy et al. [9] introduced an air quality prediction method using reinforced swarm optimization (RSO) and Bidirectional Gated Recurrent Units (Bi-GRU). Although these methods achieved high prediction accuracy, they were limited to predicting one-dimensional time series data from single monitoring stations. To address this limitation, some studies have attempted to fuse Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to construct ConvLSTM hybrid models for the simultaneous extraction of spatial and temporal features [10–12]. Building upon this model, Liu et al. [13] incorporated an attention mechanism into ConvLSTM, enabling global PM_{2.5} prediction using monitoring station data. Additionally, other studies have explored incorporating auxiliary information, such as remote sensing imagery [14,15] and environmental images [16,17], into deep learning models to further improve prediction accuracy.

Despite the progress made by deep learning models in regional air quality prediction, challenges remain regarding the effective capture of complex spatiotemporal correlations between meteorological conditions and air pollutant concentrations. A critical gap in the existing research is the difficulty in simultaneously addressing spatial heterogeneity and temporal continuity in air quality data. Most current models either focus predominantly on temporal patterns using recurrent architectures or prioritize spatial relationships using convolutional approaches but rarely integrate both dimensions optimally. Additionally, the majority of existing models fail to fully leverage the spatial relationships between monitoring stations, particularly in regions with sparse monitoring networks. These limitations result in prediction inaccuracies, especially for medium-to-long-term forecasting and in areas with complex terrain or meteorological conditions.

The spatiotemporal air quality prediction task is akin to video prediction, requiring the extraction of temporal and spatial dynamic patterns from continuous time series data. In air quality prediction, the focus is on the diffusion and evolution trends of pollutant concentrations across spatial and temporal dimensions. In the field of computer vision (CV), the Swin Transformer has recently demonstrated outstanding performance in various vision tasks by effectively integrating local spatial information and global contexts through its innovative shifted window attention mechanism, which computes self-attention within local windows [18,19]. Inspired by this, this study proposes a novel spatiotemporal

air quality prediction method that integrates Kriging interpolation with a SwinLSTM model. First, Kriging interpolates the collected one-dimensional station data onto a two-dimensional spatial plane. Then, a prediction network integrating Swin Transformer and LSTM modules is constructed to capture more comprehensive spatiotemporal dependencies in the air quality prediction task.

The main contributions of this paper are summarized as follows:

(1) The Kriging interpolation is introduced to the deep learning-based air quality prediction task, interpolating sparse monitoring station data over the entire spatial region, providing spatial distribution input for the deep learning model, and compensating for the shortcomings of single-model spatiotemporal modeling.

(2) Swin Transformer is integrated into the LSTM module, proposing a novel SwinLSTM structure that addresses the limited capability of convolutional structures in extracting spatial distribution features of air quality-related indicators, significantly improving prediction accuracy.

(3) A novel end-to-end spatiotemporal air quality prediction architecture is designed that divides the interpolated meteorological spatial dimension indicators into patches for encoding. The SwinLSTM Cell inherits features from the previous moment and propagates the current moment's features to the next moment, finally decoding the corresponding air quality spatial distribution prediction results and providing an end-to-end solution for this task.

The remainder of this paper is organized as follows: In Section 2, the Kriging interpolation algorithm, LSTM structure, and Swin Transformer structure relevant to the current research are presented. In Section 3, the proposed SwinLSTM and its application to the air quality prediction task are described in detail. In Section 4, prediction experiments with multiple indicators and time horizons are conducted using three years of meteorological data from China's Dongting Lake region to demonstrate the effectiveness of the model. Section 5 presents the paper's conclusion and future work.

2. Materials and Methods

2.1. Kriging Interpolation

Due to the often spatially unbalanced distribution of monitoring stations, it was necessary to employ reasonable spatial interpolation techniques to estimate the spatial distribution across the entire study area based on data from known monitoring sites [20]. Among these methods, Kriging interpolation has been widely applied in this field owing to its rigorous mathematical foundation and its property of providing the best linear unbiased estimate [21,22]. Unlike simpler techniques such as Inverse Distance Weighting (IDW) that assume values are solely determined by distance, or spline interpolation that focuses on smoothness, Kriging was selected for this research due to several key advantages. First, Kriging accounts for both the distance and direction of measured points, making it particularly suitable for handling the spatial heterogeneity common in air pollutant distribution. Second, it provides not only predicted values but also error estimates (prediction variance), offering valuable uncertainty quantification not available with deterministic methods. Third, Kriging demonstrates superior robustness against outliers compared to polynomial-based methods, which is an important consideration given the occasional anomalous readings in air quality monitoring data. Finally, Kriging's ability to incorporate spatial structure through variogram analysis allows it to adapt to different spatial correlation patterns across various pollutants and geographical regions.

The core idea of Kriging is to perform the weighted averaging of known samples to estimate the value at an unknown location, ensuring that the estimated value has the same mathematical expectation as the true value and minimizes the variance. Based on

the variogram theory and structural analysis, this method achieves the optimal unbiased spatial interpolation of regionalized variables; hence, it is also known as the spatial local interpolation method.

Let the observed values at a series of observation points $t_i (i = 1, 2, \dots, n)$ in the region be $Z(t_i)$; then, the value $Z^*(t_0)$ at the to-be-estimated point t_0 can be expressed as a weighted sum of the surrounding known values with the following computation formula:

$$Z^*(t_0) = \sum_{i=1}^n \lambda_i Z(t_i) \quad (1)$$

where N is the total number of known points, and λ_i represents the weight coefficients that need to be solved. In the Kriging method, the calculation of λ_i must satisfy two conditions: one is that $Z^*(t_0)$ must be an unbiased estimate, and the other is that the total variance must be minimized.

To satisfy the unbiased estimation, the mathematical expectation of the bias should be 0, i.e.,

$$\mathbb{E} \left[\sum_{i=1}^n \lambda_i Z(t_i) - Z(t_0) \right] = 0 \quad (2)$$

When satisfying the optimality condition, the minimum variance of the residual mean squared error is expressed using the covariance, which can be written as follows:

$$\sigma_E^2 = c(t_0, t_0) + \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j c(t_i, t_j) - 2 \sum_{i=1}^n \lambda_i c(t_0, t_i) \quad (3)$$

where $c(t_i, t_j)$ is the covariance between points t_i and t_j , $c(t_0, t_0)$ represent the variance at the to-be-estimated point t_0 , and $c(t_0, t_i)$ is the covariance between the known and unknown points. Using the Lagrange multiplier method to find the extremum and taking the derivative, the following can be obtained:

$$\begin{cases} \sum_{j=1}^n \lambda_j c(t_i, t_j) + \mu = c(t_0, t_i) \\ \sum_{i=1}^n \lambda_i = 1 \end{cases} \quad (4)$$

where μ is the Lagrange multiplier.

Since the Kriging variance σ_E^2 is a function of the number of known points n and the spatial correlation between points, the spatial correlation needs to be modeled during the computation process. This is typically performed using the semivariogram function to describe the spatial correlation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(t_i) - Z(t_i + h)]^2 \quad (5)$$

where h is the spatial lag distance, and $N(h)$ is the number of data pairs separated by the distance h . The semivariogram function can describe the spatial variation pattern of the regionalized variable, typically increasing with distance until reaching a stable value called the sill.

In practical applications, the semivariogram function usually needs to be fitted with a theoretical model. Commonly used models include the spherical model, exponential model, and Gaussian model, among others [23]. The fitted theoretical semivariogram model can then be used to calculate the covariance between any two points, further solving the Kriging equation system to obtain the final interpolation weights λ_i .

2.2. Long Short-Term Memory

Recurrent Neural Networks (RNNs) [24] are network structures specialized for processing sequence data. Building upon feedforward neural networks, RNNs introduce recurrent connections between hidden layer nodes, allowing information to propagate across time steps and better capture dynamic dependencies in sequential data. However, during backpropagation, the gradients of long sequences need to be computed through the multiplication of derivatives of the activation function and weight matrices. This cumulative multiplication effect can lead to exponential amplification or the decay of gradient values. When sequences are long, gradients can easily become saturated or approach zero, impacting the model's ability to capture long-range dependencies.

To address this issue, Hochreiter et al. [25] proposed the long short-term memory (LSTM) model in the 1990s, as illustrated in Figure 1. The core of this model lies in the introduction of gating mechanisms and state-carrying memory, aiming to endow RNNs with stronger long-term memory and selective forgetting capabilities. Through the gating mechanism, LSTMs can flexibly control the flow of information, retaining important information and forgetting irrelevant information, allowing key information to be efficiently propagated across time steps, thereby mitigating the vanishing gradient problem and improving the model's ability to capture long-range dependencies.

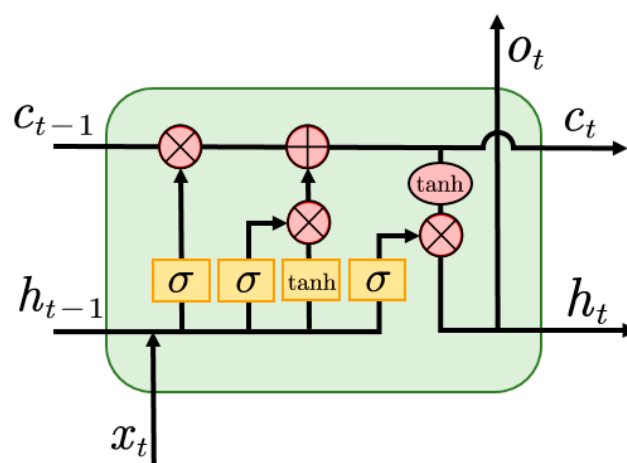


Figure 1. Structure of long short-term memory neural network module.

The core design of LSTMs involves three types of gating mechanisms: the forget gate, input gate, and output gate, which finely control the flow of information into and out of the cell state and hidden state. First, the forget gate determines which information from the previous cell state should be forgotten based on the previous hidden state and current input with the following computation formula:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (6)$$

where σ is the sigmoid activation function, which limits each scalar value between 0 and 1, with 0 representing complete forgetting and 1 representing complete retention. W_f and b_f are the learnable weight and bias parameters.

Next, LSTM needs to decide which information to extract from the current input and previous hidden state and which information to update into the new cell state. This process is controlled by the input gate using the following formulas:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (9)$$

First, the input gate i_t determines how much new information is added. $\tilde{C}_t$ is a vector containing candidate values that can be added to the new cell state. The final new cell state C_t is obtained by adding the retained information from the old state C_{t-1} and the new state $\tilde{C}_t$. With the new cell state C_t , LSTM also needs to control which information from it should be output as the hidden state h_t for downstream tasks. This is determined by the output gate o_t , with the following formulas:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t * \tanh(C_t) \quad (11)$$

where o_t controls how much information is obtained from the cell state C_t is contained in the hidden state, and the $\tanh(\cdot)$ function limits the values of C_t within the range $[-1, 1]$.

Through this design, LSTMs can flexibly control the memory and forgetting of long-term states, overcoming the weakness of traditional RNNs in effectively capturing long-range dependencies. The forgetting gate decides which information to discard, the input gate decides which new information to add, and the output gate decides which states are to be output as the hidden state. The coordinated operation of these three “gates” enables LSTMs to efficiently model long sequences. Furthermore, by introducing additive state updates and multiplicative gating mechanisms, LSTMs also avoid the vanishing and exploding of gradient problems, greatly enhancing model performance. This innovative design has established LSTMs as an important model in the field of sequence modeling.

2.3. Swin Transformer

The Swin Transformer [26], introduced by Microsoft in 2021, is an innovative general-purpose vision model. While retaining the powerful feature representation capability of the Transformer, it incorporates a series of key improvements and optimizations tailored for vision tasks, enabling it to surpass the ViT model [27] in feature extraction capability.

As shown in Figure 2, the input vector within the Swin Transformer module needs to pass through both the window-based multi-head self-attention (W-MSA) and the shifted-window-based multi-head self-attention (SW-MSA), a mechanism proven to effectively enhance feature extraction ability. Based on this window-partitioning mechanism, the process of computing feature maps in consecutive Swin Transformer blocks is as follows:

$$\hat{z}^l = W - \text{MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1} \quad (12)$$

$$z^l = \text{MLP}\left(\text{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l \quad (13)$$

$$\hat{z}^l = W - \text{MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1} \quad (14)$$

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \quad (15)$$

where $\hat{z}^l$ represents the output of W-MSA, $\hat{z}^{l+1}$ represents the output of SW-MSA, z^l represents the output of the first MLP module, and z^{l+1} represents the output of the second MLP module.

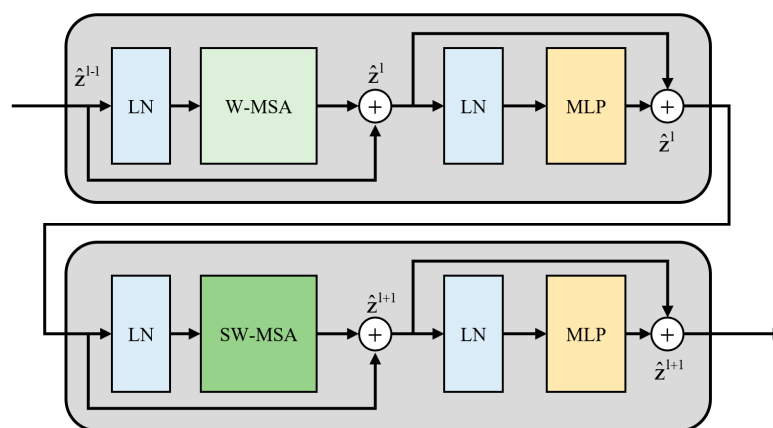


Figure 2. Internal structure of Swin Transformer module.

Although this architecture has been proven effective in enhancing the model's image feature extraction capability, the following significant issue arises: SW-MSA leads to windows of uneven lengths. To ensure computational efficiency, cyclic shifting is employed here, maintaining the same number of batched windows as regular window partitioning. This allows the use of a masking method to add relative position biases $B \in \mathbb{R}^{\hat{M} \times \hat{M}}$, thereby restricting the self-attention computation within each sub-window. Consequently, the improved attention computation formula is as follows:

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(z^{l+1}\right)\right) + z^{l+1} \quad (16)$$

where $Q, K, V \in \mathbb{R}^{M \times d}$ are the query vector, key vector, and value vector, respectively, and d is the self-attention feature dimension. In the relative position bias matrix B , for any element $b_{i,j}$, if i and j belong to the same sub-window, then $b_{i,j} = 0$; if i and j belong to different sub-windows, then $b_{i,j} \rightarrow -\infty$. After passing through the SoftMax function, the attention weights between different sub-windows approach 0, while the attention weights within the same sub-window are retained with a larger proportion.

3. Main Method

The overall architecture of the proposed air quality spatiotemporal prediction algorithm is illustrated in Figure 3. Initially, based on the location information of monitoring stations, a two-dimensional meteorological data distribution image X_t for each time step t was obtained through Kriging interpolation. For the variogram modeling in this study, a spherical model was selected after comparing the fitness with exponential and Gaussian alternatives, as it best captured the spatial correlation structure of the air quality data in our study region while maintaining computational efficiency. Subsequently, the image was segmented into non-overlapping patches. After applying cosine positional encoding [28] to each patch, they were input into the SwinLSTM Block for spatial feature extraction. Combining the hidden state H_{t-1} and cell state C_{t-1} extracted from the previous time step, the block generated the current hidden state H_t and cell state C_t . H_t was duplicated, with one copy used as a feature for predicting the meteorological data distribution $\hat{X}_{t+1}$ at the next time step, while the other was transmitted to the SwinLSTM Block of the subsequent time step along with C_t to maintain and update temporal information. The specific details of the model will be elaborated in the following sections.

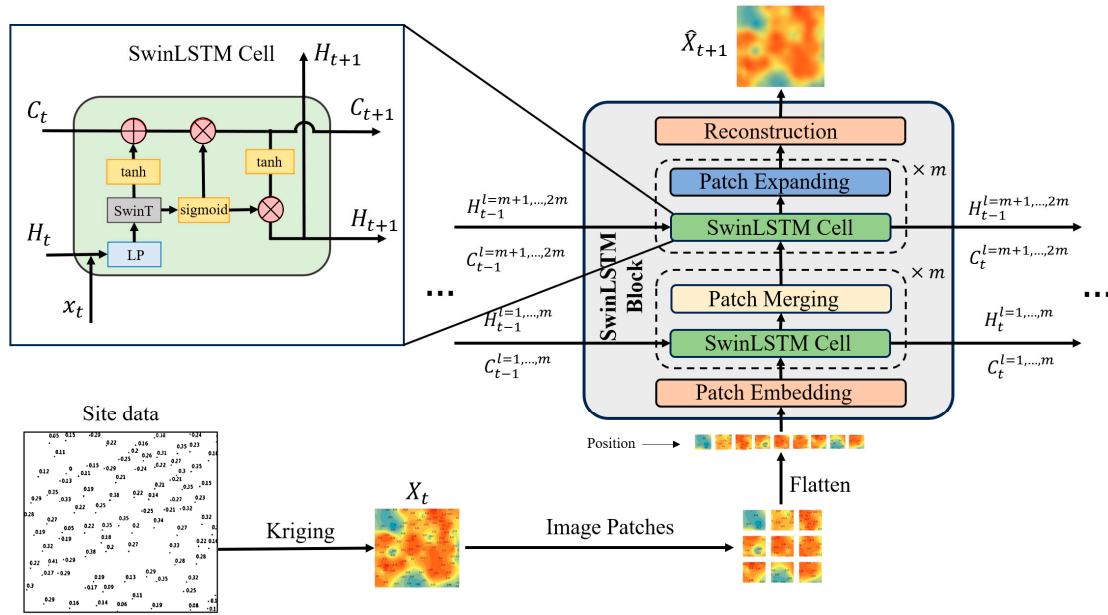


Figure 3. The overall architecture of air quality spatiotemporal prediction combining Kriging interpolation and SwinLSTM.

3.1. SwinLSTM Cell

The SwinLSTM Cell, serving as the core spatiotemporal feature extraction module in this model, inherits its concept from the LSTM structure introduced in Section 2.2 and is inspired by ConvLSTM [28].

In SwinLSTM, the hidden state H_{t-1} inherited from the previous time step was first concatenated with the current two-dimensional meteorological data. This combined feature was then input into the Swin Transformer module, as shown in Figure 2, to extract spatial features F_t , expressed as follows:

$$F_t = \sigma(\text{SwinT}(\text{LP}(x_t; H_t))) \quad (17)$$

where $\text{LP}(\cdot)$ denotes the linear projection layer and $\text{SwinT}(\cdot)$ represents the Swin Transformer feature extraction module.

Subsequently, F_t was mapped to the cell state C_{t+1} and the updated hidden state H_{t+1} of the next time step through the forget gate and output gate structures, respectively:

$$C_{t+1} = F_t \circ (\tanh(\text{SwinT}(\text{LP}(x_t; H_t)))) + C_t \quad (18)$$

$$H_{t+1} = F_t \circ \tanh(C_{t+1}) \quad (19)$$

where $\circ$ represents the Hadamard product.

3.2. SwinLSTM Block

SwinLSTM Cell is merely a basic structural unit capable of temporal feature transmission. To achieve dynamic spatiotemporal modeling and the prediction of two-dimensional meteorological data distributions, this study designed a SwinLSTM Block comprising multiple SwinLSTM Cells. In addition to the fundamental SwinLSTM Cell, the use of Block incorporates several key modules for processing the data and features input.

Specifically, after the input of two-dimensional meteorological data, X_t was segmented into multiple non-overlapping patches, and the iRPE [29] method was first employed to perform positional encoding on these patches. This introduced relative positional and contextual information to each patch, ensuring that their relative relationships and

overall structure were not lost during the partitioning process. Following iRPE encoding, the patches were mapped to a high-dimensional feature space by a Patch Embedding layer with a 1×1 convolution kernel. Given that the input and output formats of this meteorological spatiotemporal prediction task remained consistent, this study adopted the encoder–decoder paradigm inspired by the classic U-Net [30] architecture. The features first underwent m downsampling operations, with each downsampling including a SwinLSTM Cell for spatiotemporal feature extraction and a Patch Merging module that combined features of adjacent patches to achieve spatial downsampling. This path corresponded to the hierarchical extraction of spatiotemporal features, capturing representations from fine to coarse granularity. After obtaining the final bottom-level feature representation through downsampling, m upsampling operations followed. Each upsampling operation consisted of a SwinLSTM Cell and a Patch Expansion module. The SwinLSTM Cell is responsible for generating new spatiotemporal feature representations, while the Patch Expansion module expands features in the spatial dimension. This process restored feature resolution gradually, transitioning from abstract semantic features to concrete pixel-level predictions. Finally, after m upsampling operations, the features were mapped back to two-dimensional space through a 1×1 convolution layer, outputting the predicted meteorological distribution map $\hat{X}_{t+1}$.

3.3. Loss Function

To simultaneously capture the disparities between predicted and actual meteorological spatiotemporal results, this study proposes a comprehensive loss function to guide model training. This loss function comprises two components: pixel-level regression loss and spatiotemporal gradient loss.

The pixel-level regression loss aims to minimize pixel value differences between predicted and actual values, utilizing a smooth L1 loss function, which is calculated as follows:

$$\mathcal{L}_{\text{pixel}} = \frac{1}{MNT} \sum_{t=1}^T \sum_{i=1}^M \sum_{j=1}^N \text{SmoothL1}(\hat{X}_{t,i,j} - X_{t,i,j}) \quad (20)$$

where T denotes the total prediction time length, M and N represent the horizontal and vertical coordinate ranges of the Kriging-interpolated two-dimensional meteorological distribution, respectively. $\hat{X}_{t,i,j}$ and $X_{t,i,j}$ indicate the predicted and actual values at time t , row i , and column j . $\text{SmoothL1}(\cdot)$ is the smooth L1 loss function, which, compared to traditional norm losses, facilitates stable gradient updates. It is specifically calculated as follows:

$$\text{SmoothL1}(x) = \begin{cases} 0.5 \times x^2, & |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \quad (21)$$

Furthermore, this study introduces spatiotemporal gradient loss, defined as follows:

$$\mathcal{L}_{\text{gradient}} = \|\nabla_t \hat{X}_{t,i,j} - \nabla_t X_{t,i,j}\|_2^2 + \|\nabla_i \hat{X}_{t,i,j} - \nabla_i X_{t,i,j}\|_2^2 + \|\nabla_j \hat{X}_{t,i,j} - \nabla_j X_{t,i,j}\|_2^2 \quad (22)$$

where ∇_t , ∇_i , and ∇_j represent gradient calculations in temporal, horizontal, and vertical spatial dimensions, respectively. This loss term penalizes the gradient differences between predicted and actual values in both time and space, facilitating the model's learning of spatiotemporal evolution patterns.

By linearly combining the aforementioned loss terms, the comprehensive loss function is derived as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{pixel}} + \lambda_2 \mathcal{L}_{\text{gradient}} \quad (23)$$

where λ_1 and λ_2 are hyperparameters balancing the two loss terms.

During model training, minimizing this comprehensive loss function yields SwinLSTM with robust spatiotemporal modeling capabilities. The resulting predictions not only closely approximate actual values at the pixel level but also align well with real data in terms of the spatiotemporal evolution structure.

4. Experiments

The model code was developed in a PC environment using Python 3.9 and PyTorch 2.0.1+cu118 deep learning framework, compiled on the Jupyter interactive platform. The hardware configuration included a 13th Gen Intel(R) Core(TM) i5-13600K CPU and an NVIDIA GeForce RTX 3060 GPU.

4.1. Dataset Collection and Preprocessing

To validate the effectiveness of the proposed research framework, this study collected air quality data from 29 monitoring stations across four cities surrounding China's Dongting Lake, including Changsha, Yueyang, Changde, and Yiyang, over a three-year period from 1 January 2020 to 31 December 2022. Each monitoring station recorded local air quality indicators at hourly intervals, including the PM_{2.5}, PM₁₀, SO₂ concentration, NO₂ concentration, O₃ concentration, CO concentration, and AQI index. The distribution of the 29 monitoring stations is illustrated in Figure 4, where brown dots represent individual monitoring stations, and blue areas indicate water systems or reservoirs. The meteorological research and prediction area covers the region from 111.60° E to 113.90° E longitude and from 28.00° N to 30.00° N latitude, as indicated by the red area in the figure.

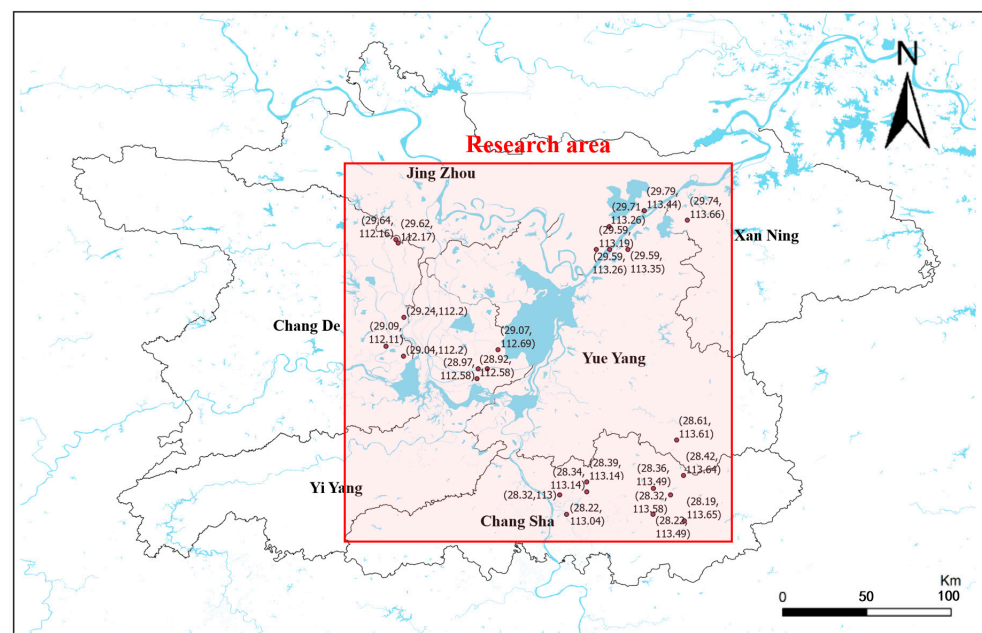


Figure 4. Distribution of monitoring stations around Dongting Lake in China.

Following data collection, preprocessing was conducted, including the imputation of missing values caused by station malfunctions. The resulting PM_{2.5} value variation curves collected from multiple stations over the three-year period are shown in Figure 5. Notably, from a long-term macroscopic perspective, PM_{2.5} variations exhibit an annual cyclical pattern, with higher levels in winter and lower levels in summer. Microscopically, PM_{2.5} values also demonstrate a daily pattern, with higher levels during the day and lower levels at night. Additionally, it was observed that PM_{2.5} values from some stations were significantly higher than the group average during certain periods. To mitigate

the influence of outliers in the dataset, this experiment employed robust standardization processing. Data scaling based on quartiles was applied to stabilize the data distribution, facilitating the unification of data across different scale ranges while avoiding the impact of extreme values on model training, i.e.:

$$x_i^* = \frac{(x_i - \mu)}{Q_3(x) - Q_1(x)} \quad (24)$$

where x_i represents the collected raw value, x_i^* denotes the standardized value, μ represents the mean of the specific air quality indicator, $Q_1(x)$ represents the first quartile, and $Q_3(x)$ represents the third quartile.

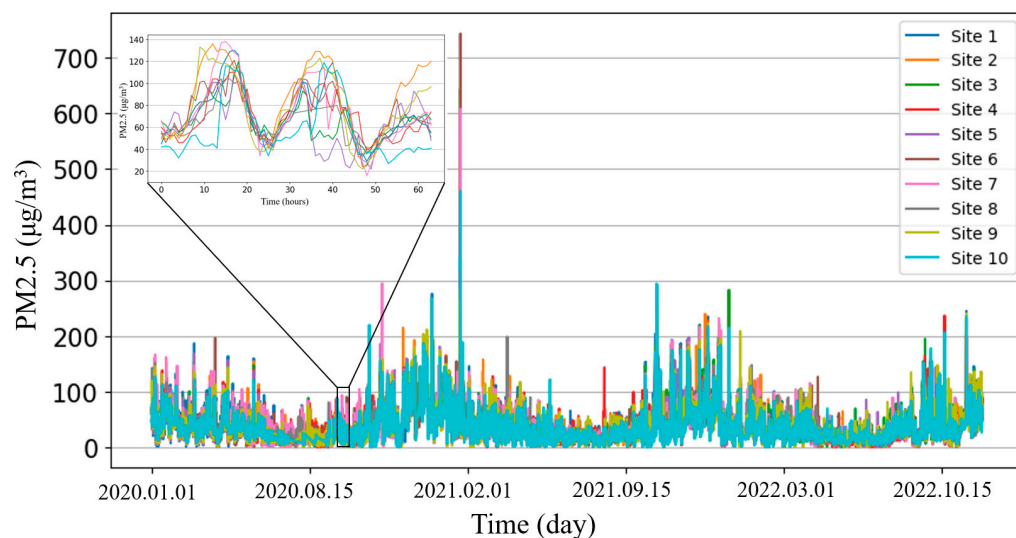


Figure 5. PM2.5 concentration change curve at 10 stations from 1 January 2020 to 31 December 2022.

4.2. Data Partitioning and Parameter Settings

For the PM2.5 spatiotemporal prediction task, a rolling window modeling approach was adopted. Specifically, for the temporal dimension T , the time series was divided into overlapping sample windows, each containing Δt time steps. The first time steps (a) served as the input feature set $Y = \{X_{\Delta t+1}, \dots, X_{\Delta t+b}\}$, while the subsequent time steps (b) were used as the prediction target $Y = \{X_{\Delta t+1}, \dots, X_{\Delta t+b}\}$. It should be noted that each time step comprised multi-dimensional input features, including not only the air quality indicators to be predicted but also meteorological conditions (temperature, humidity, and wind speed) as auxiliary features, which were aligned with the X features before being input into the model. The dataset was then split into training and testing sets. To satisfy the principle of lagged prediction in the time series models, data from 1 January 2020 to 31 December 2021 (two years) were selected as the training set, while data from 1 January 2022 to 31 December 2022 (one year) were used as the testing set.

In the spatial dimension, Kriging interpolation was applied to the original observational data using monitoring station location information, resulting in 64×64 two-dimensional air quality concentration distribution images. Through the aforementioned dataset construction method, the input was defined as a combination of meteorological features and air quality indicators with a feature set step length (a) of 48 h (past two days). The output labels were set as the prediction results for PM2.5 and PM10 indicators at 1, 6, and 24 h ahead to validate the model's performance across different prediction time spans.

Regarding the model parameters, the number of backbone feature upsampling and downsampling layers m was set to two. The Swin Transformer's patch size was set to six pixels, with a window size of four patches, an embedding dimension of 64, and four

attention heads. For the training configuration, the AdamW optimizer, most compatible with transformer architectures, was employed with a base learning rate of 1×10^{-4} . An exponential decay learning rate strategy based on training steps was adopted, with a decay rate of 0.999 and a step size of one. The total number of training iterations was set to 1000.

4.3. Evaluating Indicator

As air quality spatiotemporal prediction is a regression task, this study implemented three different evaluation metrics: mean absolute error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). The calculation formulas for these metrics are shown in Equations (25), (26), and (27), respectively.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (25)$$

$$MSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (26)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (27)$$

where n represents the sample size, y_i is the true value of the i -th sample, $\hat{y}_i$ is the corresponding predicted value, and $\bar{y}$ is the mean of all true values. The R^2 value ranges from 0 to 1, with values closer to 1 indicating better model predictions and values closer to 0 suggesting poorer predictive performance.

4.4. Result Analysis

Figure 6 presents the experimental results of the proposed SwinLSTM model for the 24 h prediction task compared with the currently mainstream ConvLSTM algorithm.

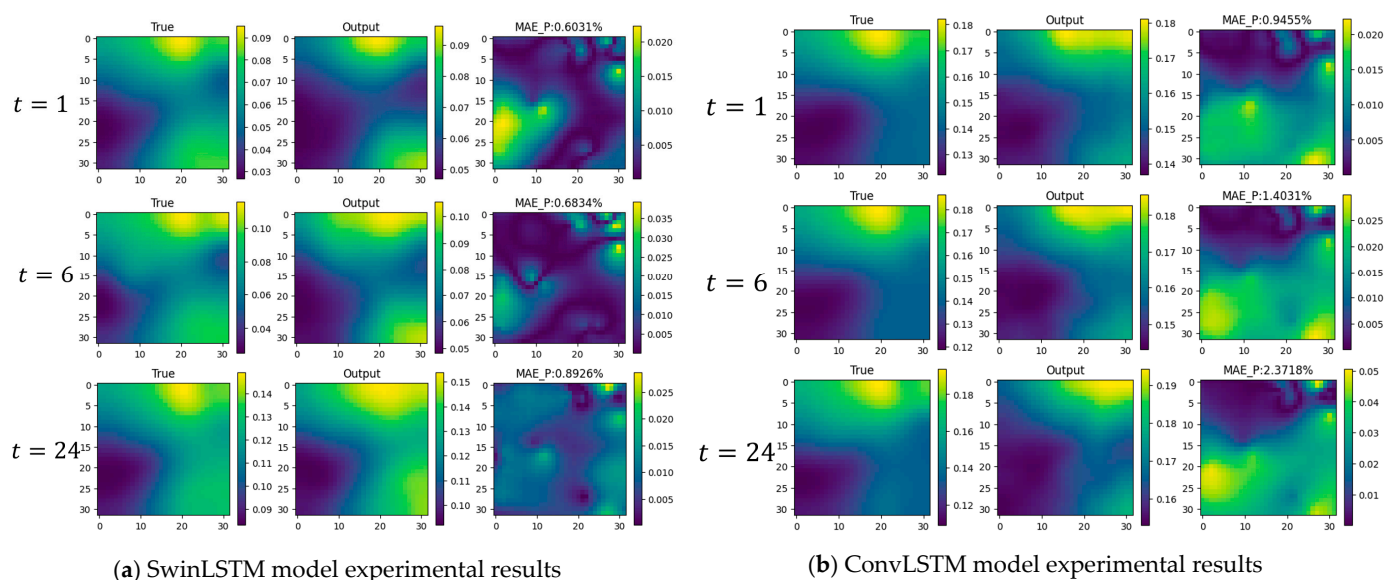


Figure 6. Comparison of the effects of SwinLSTM and ConvLSTM in the 24 h PM_{2.5} concentration prediction task (concentration unit: $\mu\text{g}/\text{m}^3$). The left column (“True”) shows the reference 2D PM_{2.5} for the spatial distribution obtained through the Kriging interpolation of actual monitoring station measurements, representing the ground truth. The middle column (“Output”) displays the model-predicted PM_{2.5} spatial distribution based on historical data. The right column (“MAE”) visualizes the pixel-wise mean absolute error between predicted and true values, with MAE_P indicating the percentage of mean absolute error relative to the true concentration range.

The results show that the PM2.5 indicator of distribution predicted by the SwinLSTM model more closely aligns with the actual values. This is evident from the mean absolute error percentage (MAEP), where for short-term prediction ($t = 1$ h), SwinLSTM and ConvLSTM yield errors of 0.60% and 0.95%, respectively. While SwinLSTM leads in terms of accuracy, the difference is not substantial. However, when extending the prediction time step to 24 h, SwinLSTM maintains a high prediction accuracy with an error percentage below 0.9%, while ConvLSTM's error percentage increases to 2.37%. This demonstrates that the proposed model can better capture spatiotemporal dependencies, thereby improving prediction accuracy.

To evaluate the model's air quality prediction accuracy at fixed locations, samples were taken from the final output PM2.5 indicator spatial distribution map at different times for a single station location. A comparison between the proposed SwinLSTM model and the ConvLSTM model is shown in Figure 7, using Station 1 as a benchmark, with the red box highlighting typical comparison areas.

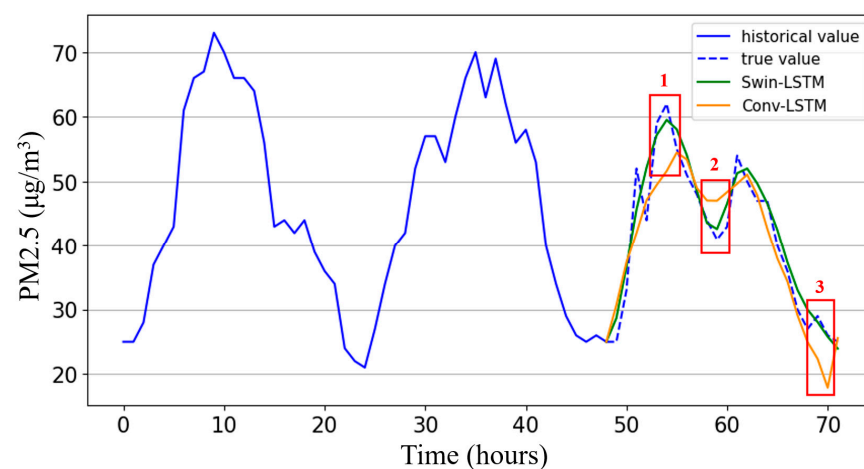


Figure 7. Comparison of the performance of SwinLSTM and ConvLSTM models in the 24 h PM2.5 concentration prediction task at site No.1.

The SwinLSTM model's predictions more closely follow the trend of actual observed values, especially when judging daily PM2.5 concentration peaks and troughs. ConvLSTM typically exhibits larger errors at these extreme values, while SwinLSTM maintains higher accuracy. This comprehensively demonstrates that the proposed model, by combining multi-head self-attention mechanisms with LSTM, achieves efficient feature fusion, resulting in more powerful spatial modeling capabilities and the better capture of dynamic patterns in the temporal dimension.

Table 1 presents comparative results of the SwinLSTM model for PM10 indicators across three different prediction time steps, alongside quantitative comparisons with previously proposed models.

It is evident that the method of first using RNN time series models for single-dimension predictions at each monitoring station, followed by Kriging interpolation to extend predictions to the entire Dongting Lake area, is notably inferior to the approach used in this study. The latter uses Kriging interpolation first, followed by full-area prediction based on two-dimensional meteorological distribution maps. Furthermore, the proposed SwinLSTM exhibited the best performance across all prediction time steps. Compared to ConvLSTM, it showed a 2.75% improvement in the R^2 value for 3 h predictions and a 6.47% improvement for 24 h predictions.

Table 1. PM10 index prediction comparison table.

Method	Model	1 h Forecast			6 h Forecast			24 h Forecast		
		MAE	MSE	R ²	MAE	MSE	R ²	MAE	MSE	R ²
Time Series + Kriging	LSTM [31]	0.073	0.295	0.877	0.164	0.390	0.806	0.274	0.436	0.673
	GRU [32]	0.068	0.248	0.898	0.156	0.317	0.821	0.287	0.449	0.668
Kriging + Spatiotemporal	PredRNN [33]	0.026	0.104	0.913	0.085	0.145	0.884	0.149	0.197	0.805
	ConvLSTM [34]	0.010	0.095	0.944	0.037	0.114	0.915	0.089	0.140	0.834
	SwinLSTM	0.009	0.074	0.970	0.015	0.092	0.953	0.028	0.117	0.898

4.5. Discussion

The superior performance of the SwinLSTM model can be attributed to several key innovations that address limitations in existing methodologies. Unlike traditional LSTM or GRU models that only capture temporal patterns at individual stations, this study's approach integrates spatial interpolation with advanced spatiotemporal modeling. The performance gap is particularly evident in medium- to long-term forecasting, where the model maintains high accuracy while pure time series approaches deteriorate rapidly. This improvement stems from SwinLSTM's ability to model the complex diffusion patterns of pollutants across space. Second, compared to CNN-based architecture like ConvLSTM, the SwinLSTM model demonstrates superior feature extraction capabilities. The shifted window attention mechanism enables both local detail preservation and global context integration, overcoming the fixed receptive field constraints of conventional convolutions. This is particularly important for capturing distant correlations in air quality patterns that may be influenced by regional meteorological phenomena. Third, the hierarchical feature representation in Swin Transformer components allows the model to detect multiscale spatiotemporal dependencies that single-scale models often miss. This is reflected in the 6.47% improvement in R² value for 24 h predictions compared to ConvLSTM, highlighting SwinLSTM's enhanced capability to maintain prediction accuracy over longer horizons.

The superior long-term prediction capability of SwinLSTM suggests that attention mechanisms more effectively capture persistent spatiotemporal dependencies compared to convolutional approaches. This has significant implications for air quality management, as more accurate 24 h forecasts provide authorities with a crucial lead time for implementing mitigation measures. The model's ability to capture peak pollution events more accurately than comparison models indicates that the hierarchical representation learning in Swin Transformer components successfully models the complex, non-linear relationships that drive extreme pollution episodes. This capability is particularly valuable for public health applications, as high-concentration episodes pose the greatest health risks.

The performance variations observed also provide guidance for model deployment strategies. For instance, the model could be dynamically configured to emphasize different aspects of its architecture based on the prediction horizon—potentially emphasizing the LSTM components for short-term predictions and the Swin Transformer components for longer-term forecasts. These insights not only validate the theoretical advantages of the proposed SwinLSTM architecture but also offer practical guidance for both model refinement and operational implementation in air quality management systems.

5. Conclusions

This research addresses the crucial environmental monitoring issue of regional air quality spatiotemporal prediction by proposing an innovative deep learning-based method. This method first uses Kriging interpolation to extend multi-dimensional meteorological and pollutant indicators recorded at monitoring stations to two-dimensional distribution

images of the entire region. It then employs the proposed SwinLSTM model to predict future regional air quality in both spatial and temporal dimensions.

The experimental results from the Dongting Lake area in China demonstrate that the proposed SwinLSTM model not only achieves excellent results in short-term PM_{2.5} and PM₁₀ prediction tasks, significantly outperforming the current mainstream ConvLSTM model but also shows more pronounced advantages in medium- to long-term prediction tasks. This proves that the proposed model can better capture the spatiotemporal correlations and evolution patterns of regional air quality.

Despite the promising results, several limitations of this study should be acknowledged. Data availability constraints restricted the analysis to a three-year period and specific pollutants; a longer time series and additional pollutants could enhance the model's robustness. The model's performance is dependent on the specific geographical and meteorological conditions of the Dongting Lake region, potentially limiting direct transferability to regions with substantially different characteristics without retraining or adaptation. The computational resources required for the SwinLSTM architecture may present implementation challenges for real-time prediction systems with limited processing capabilities.

Future work will explore more comprehensive and efficient spatiotemporal feature fusion methods, such as considering geographical factors like terrain, lakes, and altitude, to enable the model to better capture geographical conditions affecting air quality dispersion. Cross-attention mechanisms for multimodal fusion will be used to model spatiotemporal features. Additionally, the incorporation of forecast data from numerical weather models into input features will be attempted to further improve prediction fidelity. Furthermore, empirical studies on meteorological predictions will be conducted over larger regional scales, and predictions for more meteorological indicators will be made. This will provide technical support for formulating more comprehensive environmental control policies, demonstrating important theoretical value and practical significance across a wider range of fields.

Author Contributions: Data curation, X.J.; Formal analysis, X.J.; Funding acquisition, X.J.; Investigation, X.J., J.X., F.L. and S.L.; Methodology, X.J.; Project administration, J.X.; Resources, F.L.; Software, S.L.; Supervision, X.J., J.X. and F.L.; Validation, X.J.; Visualization, S.L.; Writing—original draft, J.X.; Writing—review and editing, X.J., F.L. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the science and technology innovation program of Hunan Province (2023JJ50058).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: All participants in this study gave their informed consent for participation.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: Author Fan Liu was employed by the Gansu Electric Power Changle Power Generation Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Maleki, H.; Sorooshian, A.; Goudarzi, G.; Baboli, Z.; Tahmasebi Birgani, Y.; Rahmati, M. Air pollution prediction by using an artificial neural network model. *Clean Technol. Environ. Policy* **2019**, *21*, 1341–1352. [[CrossRef](#)] [[PubMed](#)]
2. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Model. Softw.* **2019**, *119*, 285–304.
3. Mohammadi, F.; Teiri, H.; Hajizadeh, Y.; Abdollahnejad, A.; Ebrahimi, A. Prediction of atmospheric PM_{2.5} level by machine learning techniques in Isfahan, Iran. *Sci. Rep.* **2024**, *14*, 2109.

4. Ghude, S.D.; Kumar, R.; Jena, C.; Debnath, S.; Kulkarni, R.G.; Alessandrini, S.; Biswas, M.; Kulkrani, S.; Kelkar, S.; Sajjan, V.; et al. Evaluation of PM_{2.5} forecast using chemical data assimilation in the WRF-Chem model. *Curr. Sci.* **2020**, *118*, 1803–1815.
5. Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of PM_{2.5} concentrations in Beijing based on multi-source data. *Atmos. Environ.* **2017**, *150*, 146–161.
6. Wang, P.; Liu, Y.; Qin, Z.; Zhang, G. A novel hybrid forecasting model for PM₁₀ and SO₂ daily concentrations. *Sci. Total Environ.* **2015**, *505*, 1202–1212. [[PubMed](#)]
7. Bose, A.; Chowdhury, I.R. Towards cleaner air in Siliguri: A comprehensive study of PM_{2.5} and PM₁₀ through advance computational forecasting models for effective environmental interventions. *Atmos. Pollut. Res.* **2024**, *15*, 101976.
8. Chen, H.; Guan, M.; Li, H. Air quality prediction based on integrated dual LSTM model. *IEEE Access* **2021**, *9*, 93285–93297.
9. Gurumoorthy, S.; Kokku, A.K.; Falkowski-Gilski, P.; Divakarachari, P.B. Effective air quality prediction using reinforced swarm optimization and bi-directional gated recurrent unit. *Sustainability* **2023**, *15*, 11454. [[CrossRef](#)]
10. Lee, S.; Shin, J. Hybrid model of convolutional LSTM and CNN to predict particulate matter. *Int. J. Inf. Electron. Eng.* **2019**, *9*, 34–38. [[CrossRef](#)]
11. Abirami, S.; Chitra, P.; Madhumitha, R.; Kesavan, S.R. Hybrid spatio-temporal deep learning framework for particulate matter PM_{2.5} concentration forecasting. In Proceedings of the 2020 International Conference on Innovative Trends in Information Technology (ICITIIT) IEEE, Kottayam, India, 13–14 February 2020; pp. 1–6.
12. Zhang, G.; Lu, H.; Dong, J.; Poslad, S.; Li, R.; Zhang, X.; Rui, X. A framework to predict high-resolution spatiotemporal PM_{2.5} distributions using a deep-learning model: A case study of Shijiazhuang, China. *Remote Sens.* **2020**, *12*, 2825. [[CrossRef](#)]
13. Liu, P.; Yao, E.; Liu, T.; Kong, L.; Tang, X.; Tan, G. Improvement of AI forecast of gridded PM_{2.5} forecast in China through ConvLSTM and Attention. *CCF Trans. High Perform. Comput.* **2022**, *4*, 104–119. [[CrossRef](#)]
14. Xue, T.; Zheng, Y.; Geng, G.; Zheng, B.; Jiang, X.; Zhang, Q.; He, K. Fusing observational, satellite remote sensing and air quality model simulated data to estimate spatiotemporal variations of PM_{2.5} exposure in China. *Remote Sens.* **2017**, *9*, 221. [[CrossRef](#)]
15. Tariq, S.; Mariam, A.; Mehmood, U. Spatial and temporal variations in PM_{2.5} and associated health risk assessment in Saudi Arabia using remote sensing. *Chemosphere* **2022**, *308*, 136296. [[PubMed](#)]
16. Kar, A.; Ahmed, M.; May, A.A.; Le, H.T. High spatio-temporal resolution predictions of PM_{2.5} using low-cost sensor data. *Atmos. Environ.* **2024**, *326*, 120486.
17. Fan, Z.; Zhao, Y.; Hu, B.; Wang, L.; Guo, Y.; Tang, Z.; Tang, J.; Ma, J.; Gao, H.; Huang, T.; et al. Enhancing urban real-time PM_{2.5} monitoring in street canyons by machine learning and computer vision technology. *Sustain. Cities Soc.* **2024**, *100*, 105009.
18. Li, B.; Zhang, Y.; Xu, H.; Yin, B. CCST: Crowd counting with swin transformer. *Vis. Comput.* **2023**, *39*, 2671–2682. [[CrossRef](#)]
19. Chen, T.; Mo, L. Swin-fusion: Swin-transformer with feature fusion for human action recognition. *Neural Process. Lett.* **2023**, *55*, 11109–11130.
20. Wong, P.Y.; Su, H.J.; Lung, S.C.C.; Wu, C.D. An ensemble mixed spatial model in estimating long-term and diurnal variations of PM_{2.5} in Taiwan. *Sci. Total Environ.* **2023**, *866*, 161336.
21. Lee, Y.M.; Lin, G.Y.; Le, T.C.; Hong, G.H.; Aggarwal, S.G.; Yu, J.Y.; Tsai, C.J. Characterization of spatial-temporal distribution and microenvironment source contribution of PM_{2.5} concentrations using a low-cost sensor network with artificial neural network/kriging techniques. *Environ. Res.* **2024**, *244*, 117906.
22. Zhang, H.; Zhan, Y.; Li, J.; Chao, C.Y.; Liu, Q.; Wang, C.; Jia, S.; Ma, L.; Biswas, P. Using Kriging incorporated with wind direction to investigate ground-level PM_{2.5} concentration. *Sci. Total Environ.* **2021**, *751*, 141813.
23. Liu, A.; Li, Z.; Wang, N.; Zhang, Y.; Krankowski, A.; Yuan, H. SHAKING: Adjusted spherical harmonics adding KrigING method for near real-time ionospheric modeling with multi-GNSS observations. *Adv. Space Res.* **2023**, *71*, 67–79.
24. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10012–10022.
27. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [[PubMed](#)]
28. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
29. Wu, K.; Peng, H.; Chen, M.; Fu, J.; Chao, H. Rethinking and improving relative position encoding for vision transforme. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10033–10041.
30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

31. Kim, S.K.; Oh, T.I. Real-time PM₁₀ concentration prediction LSTM model based on IoT streaming sensor data. *J. Korea Acad. Ind. Coop. Soc.* **2018**, *19*, 310–318.
32. Becerra-Rico, J.; Aceves-Fernández, M.A.; Esquivel-Escalante, K.; Pedraza-Ortega, J.C. Airborne particle pollution predictive model using Gated Recurrent Unit (GRU) deep neural networks. *Earth Sci. Inform.* **2020**, *13*, 821–834.
33. Zhao, W.; Zhou, Y.; Tang, W. Novel Convolution and LSTM Model for Forecasting PM2.5 Concentration. *Int. J. Perform. Eng.* **2019**, *15*, 1528.
34. Guo, R.; Zhang, Q.; Yu, X.; Qi, Y.; Zhao, B. A deep spatio-temporal learning network for continuous citywide air quality forecast based on dense monitoring data. *J. Clean. Prod.* **2023**, *414*, 137568.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.