



Article

ConvFormer-KDE: A Long-Term Point–Interval Prediction Framework for PM_{2.5} Based on Multi-Source Spatial and Temporal Data

Shaofu Lin ^{1,†}, Yuying Zhang ^{1,†}, Xingjia Fei ¹, Xiliang Liu ^{1,*} and Qiang Mei ²

- ¹ Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; linshaofu@bjut.edu.cn (S.L.); zhangyuying@emails.bjut.edu.cn (Y.Z.); feixingjia@emails.bjut.edu.cn (X.F.)
- ² Navigation College, Jimei University, Xiamen 361021, China; meiqiang@jmu.edu.cn
- * Correspondence: liuxl@bjut.edu.cn
- † These authors contributed equally to this work.

Abstract: Accurate long-term PM_{2.5} prediction is crucial for environmental management and public health. However, previous studies have mainly focused on short-term air quality point predictions, neglecting the importance of accurately predicting the long-term trends of PM_{2.5} and studying the uncertainty of PM_{2.5} concentration changes. The traditional approaches have limitations in capturing nonlinear relationships and complex dynamic patterns in time series, and they often overlook the credibility of prediction results in practical applications. Therefore, there is still much room for improvement in long-term prediction of PM_{2.5}. This study proposes a novel long-term point and interval prediction framework for urban air quality based on multi-source spatial and temporal data, which further quantifies the uncertainty and volatility of the prediction based on the accurate PM_{2.5} point prediction. In this model, firstly, multi-source datasets from multiple monitoring stations are preprocessed. Subsequently, spatial clustering of stations based on POI data is performed to filter out strongly correlated stations, and feature selection is performed to eliminate redundant features. In this paper, the ConvFormer-KDE model is presented, whereby local patterns and short-term dependencies among multivariate variables are mined through a convolutional neural network (CNN), long-term dependencies among time-series data are extracted using the Transformer model, and a direct multi-output strategy is employed to realize the long-term point prediction of PM_{2.5} concentration. KDE is utilized to derive prediction intervals for PM_{2.5} concentration at confidence levels of 85%, 90%, and 95%, respectively, reflecting the uncertainty inherent in long-term trends of PM_{2.5}. The performance of ConvFormer-KDE was compared with a list of advanced models. Experimental results showed that ConvFormer-KDE outperformed baseline models in long-term point- and interval-prediction tasks for PM_{2.5}. The ConvFormer-KDE can provide a valuable early warning basis for future PM_{2.5} changes from the aspects of point and interval prediction.

Keywords: fine particulate matter; long-term point prediction; interval prediction; convolutional neural network; transformer; kernel density estimation



Citation: Lin, S.; Zhang, Y.; Fei, X.; Liu, X.; Mei, Q. ConvFormer-KDE: A Long-Term Point–Interval Prediction Framework for PM_{2.5} Based on Multi-Source Spatial and Temporal Data. *Toxics* **2024**, *12*, 554. <https://doi.org/10.3390/toxics12080554>

Academic Editor: Douglas Brugge

Received: 27 June 2024

Revised: 27 July 2024

Accepted: 29 July 2024

Published: 30 July 2024



Copyright: © 2024 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

With the rapid development of industrialization and urbanization, air quality issues have become the focus of social concern, especially in rapidly developing urban areas [1]. PM_{2.5}, a major factor affecting air quality, presents a serious threat to public health and environmental conservation [2]. Thus, accurate prediction of PM_{2.5} concentration is crucial for both government agencies and the public. The variations in PM_{2.5} concentrations are significantly influenced by socio-economic factors, human activities, and the spatial distribution of urban structures. Currently, most cities nationwide (such as Beijing, Guangzhou, and Haikou) have established air monitoring stations for monitoring hourly data on various air pollutants and meteorological factors [3]. However, although these monitoring stations

can provide real-time air pollution data, they cannot predict pollutant concentrations in advance. Consequently, accurate advance prediction of PM_{2.5} concentration has become essential for managing environmental health and preventing severe pollution events.

Previous studies on the prediction of PM_{2.5} concentration have mainly emphasized short-term point prediction. This approach focuses solely on specific momentary values of air pollutant concentration, overlooking the long-term trends and predictive uncertainty of PM_{2.5} concentration [4,5]. Such short-term point-prediction methods pose challenges in offering comprehensive information for decision making and constrain the deep comprehension of future air quality conditions [6]. Therefore, it is particularly important to develop an air quality prediction framework that can simultaneously consider long-term point and interval prediction. However, this is still a challenging topic, and its core issues can be summarized as follows:

- (1) How to fully exploit the interactions and impacts among air pollutants, meteorological factors, and spatial and temporal factors [7,8]. Meteorological factors have an important influence on the formation, transport, and dispersion of air pollutants. In addition, there is a degree of correlation between different monitoring stations. Therefore, it is crucial to fully consider the correlation between multiple monitoring stations and exploit the effects between multiple air pollutants and meteorological factors in the air quality prediction modelling process.
- (2) How to improve the accuracy and reliability of long-term predictions. Accurate long-term predictions can provide us with sufficient time to take measures against air pollution. However, there are complex nonlinear relationships among the factors affecting air pollutants, and the current prediction models applied to air pollution are mainly designed for short-term prediction tasks, which makes it challenging to capture the long-term dependences among air pollution time series effectively [9]. Therefore, fully exploiting the spatial and temporal effects between air pollutant concentration and meteorological factors is the key to achieving accurate PM_{2.5} prediction.
- (3) How to effectively use interval prediction to quantify uncertainty in PM_{2.5} concentration changes. Most previous studies on PM_{2.5} concentration prediction have focused on point prediction, but point prediction often has difficulty covering more fluctuating information (e.g., uncertainty, variability, and trends) [10]. The key to achieving interval prediction is modelling the point-prediction error distribution. Therefore, choosing an appropriate method to fit the point-prediction error distribution is the key to achieving interval prediction.

As we all know, the formation and variation of PM_{2.5} concentration are influenced by multiple factors, including meteorological conditions, environmental parameters, and human activities. For instance, specific meteorological conditions such as temperature and wind speed can significantly impact not only the transport and dispersion of pollutants but also determine the stability and reactivity of pollutants in the air [11]. Additionally, alterations in human activities and the distribution of points of interest (POIs) can have direct or indirect impacts on air quality [12]. Pollutant emissions from these activities can lead to correlated and synergistic PM_{2.5} concentration at different monitoring stations. However, many studies currently consider only the relationship between neighboring stations in the actual geographic area, ignoring geospatial similarity [13,14]. For example, two stations may be geographically distant, but they may exhibit similar patterns [15]. Therefore, it is imperative to fully consider the geographic similarity of all stations to enhance the accuracy of air quality prediction.

In recent years, machine learning and deep learning techniques have shown significant performance in short-term prediction of PM_{2.5} [8,16,17]. However, long-term prediction tasks present a higher challenge to existing models [18]. The core of long-term prediction modelling lies in choosing a multi-step prediction strategy [19]. The strategies commonly used in the current research can be categorized into recursive strategies [20] and direct multi-output strategies [21]. Recursive prediction strategies have the advantage of incorporating the extraction of the time dependence within the predicted sequence into the

modelling. However, introducing the predicted values leads to a severe error accumulation problem [22]. Conversely, the direct multi-output prediction strategy simultaneously generates predictions at multiple time points during the training process, effectively improving the prediction efficiency and mitigating the error accumulation problem [23]. However, this strategy typically relies on complex network architectures to capture long-term temporal dependencies between time series. Moreover, existing deep learning models for short-term prediction struggle to capture long-term dependencies in time series. Recently, the Transformer model has performed well in long-term time-series prediction owing to its advantages in capturing long-term dependencies, thus providing a new direction for long-term prediction of $PM_{2.5}$ [24]. It is worth noting that although the Transformer has significant advantages in establishing remote dependencies between data, it still has limitations in dealing with complex dependencies among multiple variables [25]. In addition, this study notes that CNNs have powerful grid-data processing capabilities to capture localized patterns and features in time series effectively. Therefore, combining a CNN with the Transformer model to construct a hybrid prediction framework that can effectively integrate multivariate and deeply mine long-term dependencies is crucial for improving the accuracy and reliability of $PM_{2.5}$ prediction.

Although point prediction of $PM_{2.5}$ concentration plays an important role in air pollution control, errors are inevitable in this type of prediction due to the volatility and non-stationarity of changes in $PM_{2.5}$ concentration [6]. In order to fully consider more uncertain information, interval prediction of $PM_{2.5}$ can effectively cover a range of $PM_{2.5}$ concentrations at different confidence levels, providing more practical information for decision makers. Currently, a commonly adopted strategy for interval prediction is to use deep learning models for point prediction and then model the distribution of prediction errors [26]. Error distribution analysis usually takes the form of a probability density function. Parametric [27] and nonparametric methods [10] are the two main techniques for extracting the probability density function of the error distribution. Parametric methods require specific presuppositions about the error distribution, such as normal distribution, exponential distribution, etc. However, in practice, the error distribution may be skewed, and these assumptions may lead to bias in estimating the error distribution. In contrast, nonparametric methods are more flexible and adaptable as they do not require specific assumptions about the error distribution but rather infer the shape of the error distribution directly from the data. Among them, kernel density estimation (KDE) is a commonly used nonparametric method, which has been widely used in wind power generation interval prediction [28], wave height interval prediction [29], and other fields.

According to the above analysis of the literature, a long-term point-and-interval-prediction framework for $PM_{2.5}$ concentration that integrates a convolutional neural network (CNN) and the Transformer model is introduced. The proposed approach comprehensively accounts for the interactions among air pollutants, meteorological factors, and $PM_{2.5}$ data from strongly correlated stations. The main contributions of this study are as follows:

- (1) In selecting influencing factors, this study considers both the interactions among various air pollutants and meteorological factors, as well as the correlations and synergies between monitoring stations across different geographic areas. The $PM_{2.5}$ concentrations at strongly correlated stations are used as one of the features to mine the potential relationship between them and the target station.
- (2) For long-term point prediction, this study notes the advantage of Transformer in mining the long-term dependence of time series. The overall structural design incorporates both CNN and Transformer models to effectively capture the long-term dependencies among multidimensional variables, thereby accomplishing stable and reliable $PM_{2.5}$ predictions.
- (3) In terms of long-term interval prediction, this study further utilizes KDE to obtain prediction intervals for $PM_{2.5}$ concentration at different confidence levels based on point-prediction results to provide more information about uncertainty levels.

The rest of this paper is organized as follows. Section 2 provides an overview of related work. Section 3 reviews the theoretical principles of these methods. Section 4 presents the dataset and experimental results. Section 5 summarizes the discussion. Finally, Section 6 gives conclusions and outlines future work.

2. Related Works

2.1. Point Prediction

2.1.1. Traditional Models

Traditional prediction models for $PM_{2.5}$ can be roughly categorized into three types: deterministic methods [30], statistical methods [31], and machine learning methods [32]. Deterministic methods are a type of modeling based on the transport, dispersion, and chemical transformation processes of pollutants in the atmosphere. The most commonly used deterministic methods include the Community Multiscale Air Quality (CMAQ) model [30], the Weather Research and Forecasting (WRF) model [33], and others. However, these methods usually have limitations such as high computational complexity, parameter uncertainty, and high data requirements when confronted with complex atmospheric environments and $PM_{2.5}$ concentration variations [18]. To address these challenges, statistical models have been proposed, such as the autoregressive integrated moving average (ARIMA) model [31], geographically weighted regression (GWR) [34], and the ridge regression (RR) model [35]. Unlike deterministic methods, statistical methods do not rely on complex theoretical understanding and provide a faster, simpler, and less costly implementation [36]. However, the prediction performance of statistical methods can be limited because they may struggle to effectively capture the complex nonlinear relationships between the data [9]. Therefore, for $PM_{2.5}$ concentration prediction, some studies have chosen machine learning methods that can better handle the complex correlations between data. Common methods include the K-nearest neighbor algorithm (KNN) [37], random forest (RF) [18] and linear regression models [16]. These machine learning-based models capture the nonlinear relationships in the data more accurately. Nonetheless, they may struggle to capture long-term dependencies, leading to a rapid decrease in prediction accuracy as the time step of prediction increases.

2.1.2. Deep Learning Models

In recent years, with the rapid development of artificial intelligence and big data technologies, deep learning has emerged as a pivotal area of research for air quality prediction [38–40]. Among them, the structure of the recurrent neural network (RNN) is well adapted to the highly nonlinear nature of air pollution data and is widely used in air quality prediction [41]. However, these RNN-based models face limitations, including gradient vanishing and time-consuming iteration propagation problems [42]. To overcome this limitation, LSTM, a variant of RNN, is gradually being applied to air pollution prediction. For instance, Zhang proposed a multi-scale $PM_{2.5}$ prediction method based on bidirectional LSTM, yielding promising results in hourly $PM_{2.5}$ prediction tasks in Beijing [43]. Gao et al. constructed a water quality parameter prediction model based on the results of driver analysis of an interpretable LSTM model [38]. The experimental results of these studies show that LSTM can alleviate problems such as RNN gradient drop. However, there are still limitations in capturing the long-term dependence of time series [44]. The Transformer model has recently been proposed to provide new ideas for time-series prediction [24]. For example, Li et al. applied the Transformer to time-series prediction and proposed an improvement scheme to solve the localization and memory bottleneck problems of the Transformer in time-series prediction applications, which laid a foundation for the subsequent advancements of the Transformer in this field [45]. Zuo et al. devised a Transformer-based THP model, leveraging self-attention mechanisms to capture long-term dependencies in event sequence data, thereby enhancing time-series prediction accuracy and computational efficiency [46]. Unlike traditional RNNs and LSTMs, the Transformer employs a self-attention mechanism that is independent of positional information, thereby enhancing its ability to capture information within lengthy sequences [45]. Currently, the

Transformer and its variants remain among the advanced models for time-series prediction, especially in long sequence prediction applications. However, the above models, including Transformer, overlook the relationship between multiple variables. For instance, Zhang et al. have stated that the Transformer model makes it difficult to capture the correlations between different variable sequences in a multivariate time series [25]. However, time-series relationships between multivariate variables are crucial for time-series prediction. Many multivariate time-series prediction models have been developed to tackle this issue, with a majority adopting a hybrid modeling approach that combines two distinct modeling paradigms [47]. Hybrid approaches combine multiple prediction models, leveraging the strengths of each model structure, resulting in improved accuracy and stability. For example, Rick et al. constructed a deep learning architecture model combining LSTM and a CNN. LSTM was used to process time-series dependencies, and the CNN was used to capture spatial features in time-series data [28]. Experimental results demonstrated superior performance to traditional temporal convolutional network (TCN) methods in energy prediction tasks. Similarly, Kumar et al. introduced a multi-view CNN-BiLSTM model architecture for predicting time-series data of multiple pollutants in a highly polluted city. They demonstrated that it significantly outperformed traditional deep learning models in terms of performance [48]. CNN models can accurately extract local features [49] and can be used to extract relationships between long multivariable time series through sliding windows and convolution operations. Motivated by these insights, this study integrates the CNN and Transformer models for long-term prediction of PM_{2.5} concentration.

2.2. Interval Prediction

Compared with point prediction, interval prediction can provide a more accurate measure of the underlying uncertainty in the prediction [26]. Interval prediction results include upper and lower bounds that can be shown to be within a certain confidence level. Compared with point prediction, interval prediction provides more reliable and comprehensive information [10]. Currently, interval-prediction methods can be divided into two categories: directly predicting the upper and lower bounds of the intervals and estimating the probability density based on the point-prediction results [26]. Direct prediction methods often require the specification of a fixed interval width, making it challenging to calculate the uncertainty of prediction results [50]. Therefore, probabilistic prediction methods based on point-prediction results are more widely used for interval prediction [51]. These methods can be further categorized into non-parametric and parametric methods [27]. In practice, the difficulty in predetermining the accurate potential distribution of prediction errors makes the implementation of parametric methods challenging [29]. Non-parametric methods do not rely on specific error distribution assumptions and can accurately quantify the range of fluctuations [52], as in quantile regression (QR) [4] and KDE [28] methods. Xu et al. implemented forecasting of renewable energy generation and buildings' electricity loads using quantile regression methods [53]. Li et al. proposed an hourly PM_{2.5} prediction system and utilized the KDE method to quantify the uncertainty of the prediction results [6]. Since the KDE method can directly provide the probability density function, it has become a more widely used probabilistic prediction method. For example, Niu et al. employed kernel density estimation with the Gaussian kernel function to obtain wind-power prediction intervals with different confidence levels and validated the method's practicality and reliability through several experiments [28]. Among the various kernel functions of KDE, the Gaussian kernel function is mostly used for time-series prediction. The Gaussian kernel function exhibits a faster decaying tail, which aids in reducing the variance of the estimate and enhancing the accuracy of the density estimate. Therefore, the current study employs a Gaussian kernel function fitted to the KDE for long-term interval prediction of PM_{2.5} concentration.

3. Methodology

3.1. The Overall Framework

This study combines air pollutant and meteorological data from target stations and strongly correlated stations to exploit intricate spatial and temporal relationships for long-term point and interval prediction of PM_{2.5}. The overall framework is depicted in Figure 1. First, multi-source data are collected and preprocessed. POI data in the study area are used to perform spatial clustering analysis of all monitoring stations to screen for strongly correlated stations. The Pearson correlation coefficient is used to analyze the correlation between all features to determine the feature variables for final input into the model. For model training and testing, the dataset is separated into three sets: training, validation, and test, in a 7:1:2 ratio. Second, a hybrid deep learning model based on a convolutional neural network and the Transformer is applied to achieve accurate long-term point predictions of PM_{2.5}. Finally, KDE-based interval prediction is performed based on point-prediction error estimation to obtain the prediction intervals of PM_{2.5} at different confidence levels.

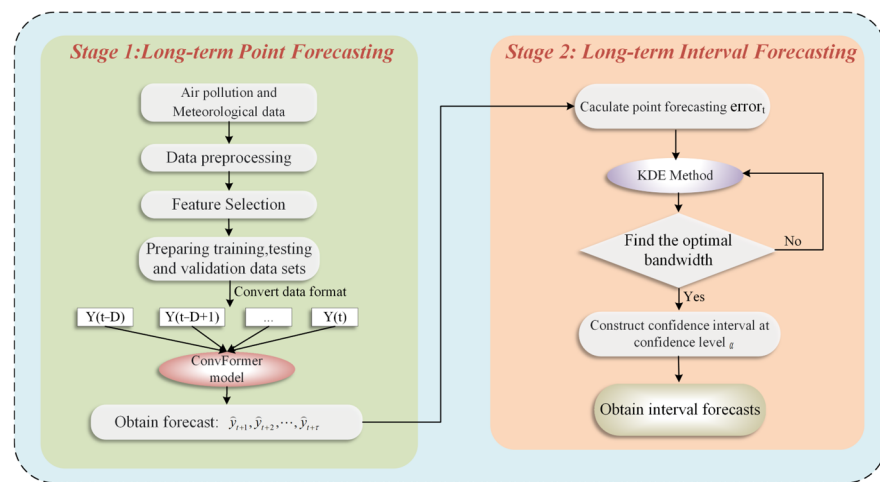


Figure 1. The overall framework of the proposed approach.

3.2. Preliminaries

Assume that there are N stations in the study area, denoted by the set $S = \{S_1, S_2, \dots, S_N\}$. Each station contains three attributes, including station id, longitude, and latitude. The count of different types of POIs around each station is denoted by $S^* \in R^{N \times K}$, where N denotes the number of stations and K denotes the total number of categories of POIs. Let $X_i \in R^{T \times M}$ represent all the features of station S_i at historical time T , encompassing air pollution data (PM_{2.5}, PM₁₀, CO, etc.), meteorological data (wind speed, wind direction, temperature, etc.), and PM_{2.5} concentration at strongly correlated stations. PM_{2.5} is the target pollutant in this study. For the target station S_i , the historical observation data $X' \in R^{D \times M}$ are used to predict the point and interval concentration of PM_{2.5} for the future time interval from t to $t + \tau$, where D denotes the historical time step $D \in \{t - D, t - D + 1, t\}$. The point prediction of PM_{2.5} concentration from t to $t + \tau$ is denoted by $\hat{Y}_{t+\tau}^{point} = \{\hat{Y}_{t+1}^{point}, \hat{Y}_{t+2}^{point}, \dots, \hat{Y}_{t+\tau}^{point}\}$, $\hat{Y}_{t+\tau}^{point} \in R^{\tau \times 1}$. For a given confidence interval α , the interval prediction is denoted by $\hat{Y}_{\alpha, t+\tau}^{interval} = [L_{\alpha, t+\tau}, U_{\alpha, t+\tau}]$, $\hat{Y}_{\alpha, t+\tau}^{interval} \in R^{\tau \times 2}$.

3.3. Spatial Clustering Based on POIs

In order to examine the similarity and geographical association patterns across monitoring stations, this method obtains POI data for the study area using Baidu’s open API. The POI data include a range of geographical entities, such as business areas, cultural facilities, transportation hubs, and more. Afterwards, hierarchical clustering is utilized to spatially group all monitoring stations. Hierarchical clustering algorithms create a dendrogram by grouping into clusters stations that are both spatially close and similar in nature. This approach eliminates the requirement to pre-determine the number of clusters, making it

easier to explore potential geographical patterns within the research region without prior information. Examining the clustering outcomes enhances our overall comprehension of the geographical connections between monitoring stations, uncovering groups of stations that display close spatial correlations with mutually beneficial changes. The spatial clustering module is represented by pseudo-code in Algorithm 1, and the formulas used in the algorithm are provided in Equations (1)–(3).

Algorithm 1 Proposed spatial clustering approach

Input: $S = \{S_1, S_2, \dots, S_n\}$ ($n \in 1, \dots, N$); $P = \{P_1, P_2, \dots, P_m\}$ ($m \in 1, \dots, M$). // S_n, P_m represent station location information and POI information, respectively;

Output: C ;

- 1: $S^* = \{S_1^*, S_2^*, \dots, S_n^*\}$ // initialize S^* to a matrix of $n \times k$ dimensions, S_n^* to a matrix of $1 \times k$ dimensions;
 - 2: **for** S_i in $\{S_1, S_2, \dots, S_n\}$ **do**
 - 3: **for** P_j in $\{P_1, P_2, \dots, P_m\}$ **do**
 - 4: compute $d(S_i, P_j)$. according to Equation (1);
 - 5: **if** $d(S_i, P_j) < 1$ km **do**
 - 6: update S_i^* ;
 - 7: $C = \{C_1, \dots, C_n\}$ // Each S_n^* is regarded as a separate cluster;
 - 8: **while** $n > 1$ **do**
 - 9: **for** C_i in $\{C_1, \dots, C_n\}$ **do**
 - 10: **for** C_j in $\{C_1, \dots, C_n\}$ **do**
 - 11: $M(i, j) = D(C_i, C_j)$ according to Equations (2) and (3);
 - 12: $M(j, i) = M(i, j)$;
 - 13: find the most similar clusters: C_{i^*} and C_{j^*} ;
 - 14: merge C_{i^*} and C_{j^*} : $C_{i^*} = C_{i^*} \cup C_{j^*}$;
 - 15: **for** $k = j^* + 1, j^* + 2, \dots, n$ **do**
 - 16: $C_k = C_{k+1}$;
 - 17: $n = n - 1$;
 - 18: **return** C ;
-

$$d(S_i, P_j) = 2\arcsin \sqrt{\sin^2 \frac{(plat - slat)}{2} + \cos(plat) \times \cos(slat) \times \sin^2 \frac{(p \ln g - s \ln g)}{2}} \times 6378.137 \tag{1}$$

$$dist(p_a, q_b) = \sqrt{\sum_{k=1}^n \left(\frac{x_k^{(p_a)} - x_k^{(q_b)}}{s_k} \right)^2} \tag{2}$$

$$D(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{p_a \in C_i, q_b \in C_j} dist(p_a, q_b) \tag{3}$$

where $p \ln g$, $plat$ denote the latitude and longitude of the POI and $s \ln g$, $slat$ denote the latitude and longitude of the monitoring stations. The value 6378.137 is the radius of the Earth’s equator in kilometers. $dist(p_a, q_b)$ represents the normalized Euclidean distance between data points p_a and q_b , n represents the number of dimensions of the data point, $x_k^{(p_a)}$ and $x_k^{(q_b)}$, respectively, represent the values of data points p_a and q_b in the k dimension, s_k is the standard deviation on the k dimension, $D(C_i, C_j)$ represents the similarity between clusters C_i and C_j , and $|C_i|$ and $|C_j|$, respectively, represent the number of samples in each cluster.

3.4. ConvFormer Network

The structure of the proposed ConvFormer network is shown in Figure 2. The Transformer has a significant advantage in capturing long-term dependencies between time series. However, it is difficult for it to capture relationships between multivariate variables. Therefore, this study combines CNN to mine local patterns and short-term dependencies

among multivariate variables and the Transformer to obtain long-term dependencies among time series. Additionally, this study adopts a direct multi-output strategy for long-term point prediction.

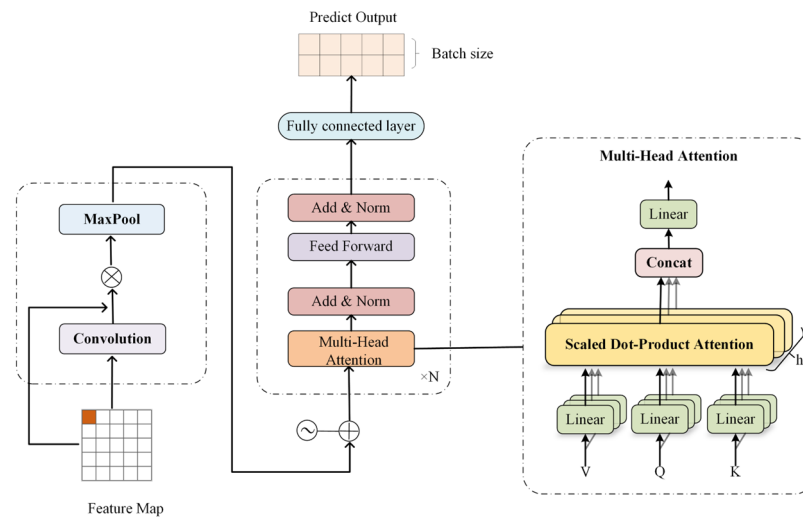


Figure 2. ConvFormer network architecture.

CNNs, representing a robust deep learning model, have demonstrated successful applications in image analysis, natural language processing, and various other fields. In the field of multivariate time-series prediction, a CNN automatically learns complex patterns and regularities in time-series data through its structure of convolutional and pooling layers, which can effectively deal with the interactions and temporal relationships among multiple variables. Therefore, this proposed method utilizes a CNN to process historical observation data. The input multivariate time-series data are converted into two-dimensional feature variables $X' \in R^{D \times M}$, and the convolution operation is performed to obtain the feature map $X'' \in R^{D \times M}$ where D denotes the sliding window step size and M denotes the dimension of the input features. The computation process of each element in the feature map is shown in Equation (4). Then, the maximum pooling operation is utilized to retain the most significant features in the multivariate data and ignore the less important information. The final matrix $X''' \in R^{D \times 1}$ is obtained as an input to the Transformer.

$$X''_{i,j} = f_{conv}(\sum_{m=0}^p \sum_{n=0}^q w_{m,n} x'_{i+m,j+n} + b) \tag{4}$$

where $X''_{i,j}$ denotes the feature output value of row i and column j of the feature map, $x'_{i+m,j+n}$ denotes the value in row $i + m$ and column $j + n$ of the input feature matrix, $f_{conv}(\cdot)$ denotes the chosen activation function, $w_{m,n}$ denotes the weight value of the row m and column n of the convolution kernel, b denotes the deviation of the convolution kernel.

The Transformer model is a feed-forward neural network architecture. Its core is its self-attention mechanism, which can be utilized to effectively capture the relationship between any two points in a time series. In particular, the self-attention mechanism computes correlation weights between each position in the input sequence and other locations, and then applies these weights to generate a representation of each position. The self-attention mechanism is defined by the following formula:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

where Q, K, V denote the query vector, key vector, and value vector, respectively, d_k denotes the dimensionality of the key, and $Softmax$ is the activation function that transforms the

input to the interval $[0, 1]$. The self-attention mechanism derives the attention weights by evaluating the similarity between the query and the keys, and it produces the final representation through a weighted sum.

In contrast to the original Transformer architecture, this model eliminates the need for final probability calculations using *Softmax*. Instead, the final predicted value of the target pollutant concentration at the station at time t is derived by mapping the generated feature maps to the output values.

3.5. Interval-Prediction Method: Non-Parametric Kernel Density Estimation

Interval prediction of $PM_{2.5}$ relies on point prediction, followed by the delineation of upper and lower bounds to define the prediction intervals. This approach quantifies the uncertainty in $PM_{2.5}$ concentration changes, offering comprehensive early warning information for future $PM_{2.5}$ variations. Non-parametric KDE is widely applied in interval prediction due to its independence from specific probability distribution assumptions. KDE, as a non-parametric estimation method, is not constrained by the specific form of probability distribution, which enables it to fit sample data accurately and reliably. Therefore, the proposed method uses KDE to quantitatively analyze and estimate point-prediction results for $PM_{2.5}$. First, the error sequence $error = [error_1, error_2, \dots, error_n]$ is initially derived based on the difference between predicted and actual values within the training set. The optimal bandwidth h_{opt} of the KDE is then determined based on a grid search and a five-fold cross-validation approach. Based on the obtained optimal bandwidth h_{opt} , the KDE model is fitted on the error sequence $error$, where the estimation function of KDE is described as follows:

$$\hat{f}(error) = \frac{1}{Nh_{opt}} \sum_{i=1}^N K\left(\frac{error - error_i}{h_{opt}}\right) \quad (6)$$

where N denotes the number of samples and $K(\cdot)$ denotes the kernel function. Commonly used kernel functions include the Gaussian kernel function, Epanechnikov kernel function, rectangular kernel function, etc. Compared with other kernel functions, the Gaussian kernel function can generate a smoother density estimation curve, which is conducive to capturing the overall characteristics of the data distribution. Therefore, the Gaussian kernel function is used here, and its expression is as follows:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \quad (7)$$

Based on the fitted KDE model, the probability density function (PDF) and cumulative distribution function (CDF) of the error are calculated. For a given confidence level α , the lower and upper bounds of the confidence interval are $l_{\alpha,t+\tau}$ and $u_{\alpha,t+\tau}$. Finally, the interval-prediction result $\hat{Y}_{\alpha,t+\tau}^{interval}$ for the test set is obtained through Equation (8).

$$\hat{Y}_{\alpha,t+\tau}^{interval} = \left[\hat{Y}_{t+\tau}^{point} + l_{\alpha,t+\tau}, \hat{Y}_{t+\tau}^{point} + u_{\alpha,t+\tau} \right] \quad (8)$$

4. Experiment

4.1. Dataset Description and Preprocessing

4.1.1. Description

The study area selected was Haikou City, located in Hainan Province. Initially, air pollution concentration data ($PM_{2.5}$, PM_{10} , O_3 , CO , etc.) and meteorological data (wind speed, temperature, pressure, etc.) for the same period were collected from the monitoring stations in Haikou. The dataset on air quality included hourly data spanning from 30 October 2020 to 26 December 2023. The distribution of monitoring stations is marked in blue in Figure 3. The target station for this study was S9, which is indicated by the red marking in Figure 3. S9, being in the heart of the city and surrounded by a multitude of

stores, provides a better response to the influence of spatial–temporal correlation on the model’s predictions and is relatively representative. Additionally, the POI data obtained through Baidu’s open API included 14 different categories, as shown in Table 1. Each POI point contained six attributes, from which first-level classification, longitude, and latitude were selected for the application, resulting in a total of 92,108 POIs.

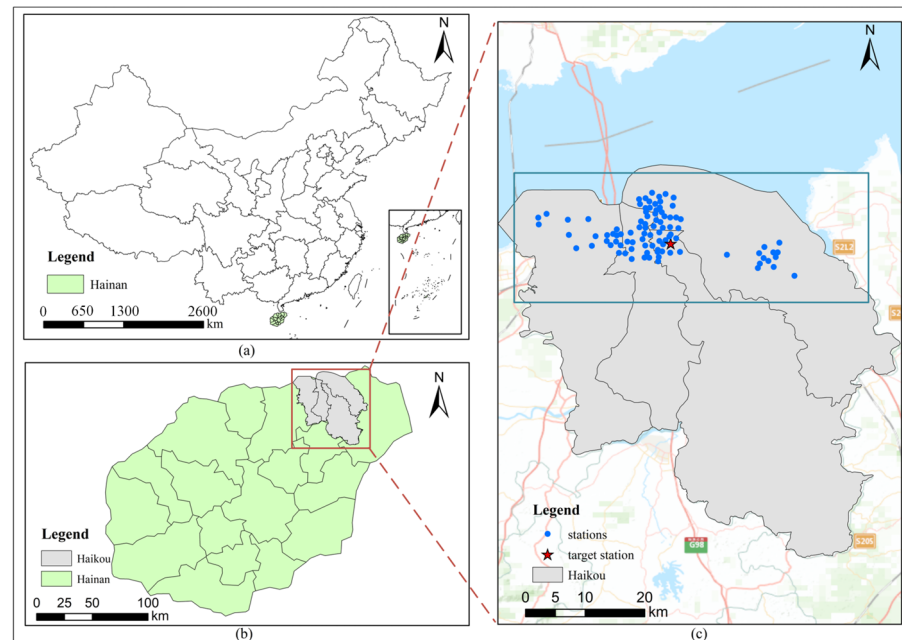


Figure 3. Study area and spatial distribution of air monitoring stations: (a) The green part is the boundary map of Hainan province, (b) The gray part is the boundary map of Haikou city, (c) Distribution of air monitoring stations in Haikou (zoom figure).

Table 1. POI categories.

| First Level Classification | Second Level Classification |
|----------------------------|--|
| Education and Training | Higher education institutions, secondary schools, elementary schools, kindergartens, adult education, parent–child education, special education schools, study abroad agencies, research institutions, training institutions, libraries, science and technology centers, others. |
| Medical | General hospitals, specialist hospitals, clinics, pharmacies, medical centers, sanatoriums, emergency centers, disease control centers, others. |
| Transportation Facilities | Parking lots, service areas, bus stations, wharves, train stations, ferries, toll stations, airports, coach stations, others. |
| Sports and Fitness | Stadiums, extreme sports venues, fitness centers, others. |
| Tourist Attractions | Town squares, zoos, botanical gardens, amusement parks, museums, aquariums, heritage sites, churches, scenic spots, others. |
| Finance | Banks, ATMs, credit unions, investment banking, pawnshops, others. |
| Automobile Services | Automobile sales, automobile repair, automobile cosmetics, automobile parts, automobile rental, automobile inspection yard, others. |
| Life Services | Logistics, public toilet, post office, salons, hairdressers, bath and massage, laundry, public utilities, others. |
| Food | Chinese restaurants, foreign restaurants, snack and fast-food restaurants, cake and dessert stores, cafes, cafeterias, bars, and others. |
| Hotels | Star hotels, fast hotels, apartment hotels, others. |
| Shopping | Shopping centers, department stores, supermarkets, convenience stores, home building materials, home appliances and digital stores, bazaars, duty-free stores, others. |

Table 1. Cont.

| First Level Classification | Second Level Classification |
|----------------------------|---|
| Leisure and Entertainment | Resorts, open farms, cinemas, karaoke halls, theaters, dance halls, Internet cafes, gaming arcades, bath and massage, leisure plazas, others. |
| Company Enterprise | Companies, factories, others. |
| Real Estate | Office buildings, residential areas, dormitories, neighborhoods, villages, community centers, others. |

4.1.2. Dataset Preprocessing

Figure 4 depicts the hourly $PM_{2.5}$ concentration sequence from the target station in 2022. The figure shows that the $PM_{2.5}$ concentration exhibited significant volatility and instability. Notably, a continuous missing value was evident, represented by the red mark labeled A in the figure. Data gathering can be hindered by problems including equipment breakage and transmission faults, which can result in outliers and missing results. These missing values can negatively affect the prediction of $PM_{2.5}$ concentration. Thus, preprocessing was necessary before constructing the prediction model. The forward filling method was specifically employed to address missing values within short time periods (e.g., up to 4 h) in the original data. For medium and long time periods (e.g., more than 4 h but less than 72 h), the multiple interpolation method was utilized to fill in the missing values. However, missing values in long time periods (e.g., exceeding 72 h) were simply deleted and handled accordingly. In order to eliminate the effect of magnitude between different features, a normalization method was finally used to scale the data within the values (0, 1).

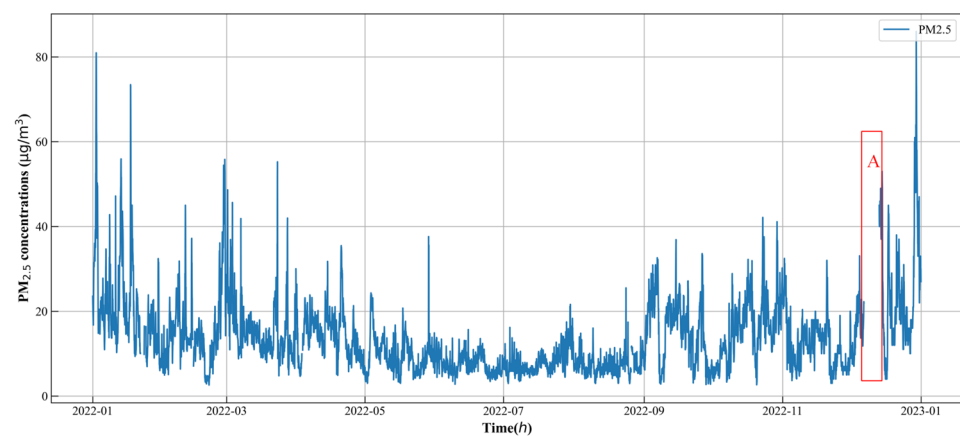


Figure 4. Time series of hourly $PM_{2.5}$ at the S9 station and a large number of missing values in time period A.

After data preprocessing, a total of 26,451 data samples were obtained. The dataset was separated into the training set, validation set, and test set according to the ratio 7:1:2. The training set contained 18,526 sample points for model training and parameter optimization. The validation set consisted of 2635 sample points to assess the training and generalization performance of the model. The test set included 5290 sample points to validate and evaluate the long-term prediction performance of the model. The statistical descriptions of the training, validation, and test sets are shown in Table 2.

Table 2. Statistical descriptions of different data sets.

| Data Set | Numbers | Maximum | Minimum | Mean | Standard Deviation |
|----------------|---------|---------|---------|-------|--------------------|
| Train set | 18,526 | 103.48 | 2.53 | 15.84 | 11.52 |
| Validation set | 2635 | 78.20 | 4.0 | 19.45 | 9.66 |
| Test set | 5290 | 97.0 | 2.62 | 16.83 | 12.78 |

4.2. Evaluation Metrics

4.2.1. Evaluation Metrics of Point Prediction

To comprehensively evaluate the performance of the proposed approach in this study, three evaluation metrics including root mean square error (RMSE), mean absolute error (MAE), and R-square (R^2) were used. These metrics were calculated as shown in Equations (9)–(11). RMSE denotes the sample standard deviation of the difference between predicted and observed values. MAE denotes the mean of absolute errors between predicted and observed values. The MAE is a linear score in which all individual differences are equally weighted on the mean. RMSE penalizes high variance more compared with MAE. R^2 was used to assess the fitting ability of the model, with values closer to 1 indicating a better fit of the predicted learning results.

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2} \tag{9}$$

$$MAE = \frac{1}{N} \sum_{t=1}^N |y_t - \hat{y}_t| \tag{10}$$

$$R^2 = 1 - \frac{\sum_{t=0}^N (y_t - \hat{y}_t)^2}{\sum_{t=0}^N (y_t - \bar{y})^2} \tag{11}$$

where N denotes the total number of samples, y_t denotes the true value, \hat{y}_t denotes the predicted value, and \bar{y} denotes the mean value.

4.2.2. Evaluation Metrics of Interval Prediction

A good interval-prediction model should ensure that observations align with the prediction interval to the closest degree possible. At the same time, these prediction intervals should be as narrow as possible to improve the accuracy of the prediction. Therefore, in order to comprehensively assess the performance of the interval-prediction model proposed in this study, prediction interval coverage probability (PICP) and prediction interval normalized averaged width (PINAW) metrics were used. These metrics were calculated as shown in Equations (12)–(14). PICP is used to reflect the probability that an actual observation falls within the prediction interval at a given confidence level. It serves as a key metric for assessing the efficacy of an interval-prediction model. PINAW is employed to reflect the normalized average width of all prediction intervals. Typically, smaller PINAW values indicate narrower prediction intervals, i.e., higher accuracy.

$$PICP = \frac{1}{N} \sum_{i=1}^N C_i^\alpha \tag{12}$$

$$C_i^\alpha = \begin{cases} 0, & Y_i^{point} \notin [L_{\alpha,i}, U_{\alpha,i}] \\ 1, & Y_i^{point} \in [L_{\alpha,i}, U_{\alpha,i}] \end{cases} \tag{13}$$

$$PINAW = \frac{1}{N(\max(Y^{point}) - \min(Y^{point}))} \sum_{i=1}^N (U_{\alpha,i} - L_{\alpha,i}) \tag{14}$$

where N represents the number of samples, α represents the confidence level, Y^{point} and \widehat{Y}_i^{point} represent the actual observed and predicted values of the point prediction, respectively, C_i^α represents a Boolean value, where 1 indicates that the observation falls within the prediction interval, and 0 otherwise, $U_{\alpha,i}$ and $L_{\alpha,i}$ represent the upper and lower bounds of the prediction interval at the confidence level α , respectively.

4.3. Feature Selection

The hierarchical clustering method and POI data were employed to cluster all 95 monitoring stations, resulting in four distinct clusters, as illustrated in Figure 5. The stations in Cluster 1 and Cluster 3 were located in the city center, characterized by dense architectural facilities, such as prominent urban structures such as commercial buildings, cultural institutions, and retail centers. The stations in Cluster 2 were located in the outside region, predominantly flanked by educational institutions and recreational areas. The stations in Cluster 3 were located in central locations, with a convergence of various activity facilities nearby. According to the clustering results, it was observed that the target monitoring station (S9) was part of Cluster 1. The stations in Cluster 1 underwent additional refinement to identify those that correlated with the target station, in conjunction with spatial variability analysis. Figure 5c illustrates the selection of stations that were highly correlated, such as S2, S3, S8, S39, and S45. Ultimately, the model's prediction was informed by the PM_{2.5} concentrations from these stations that were highly correlated.

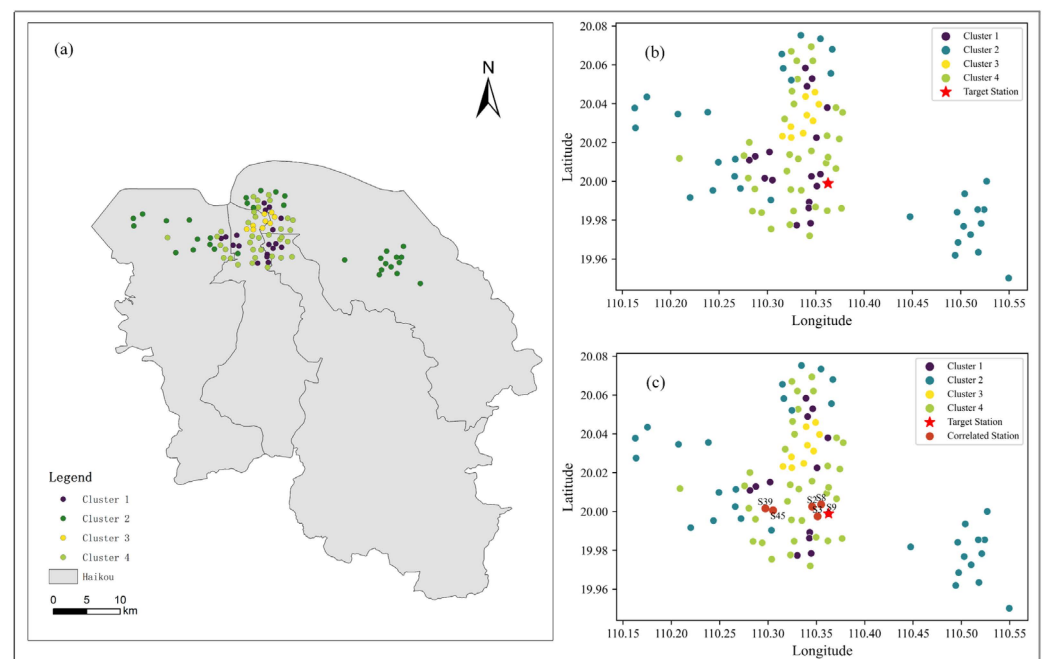


Figure 5. The selection results of the correlated stations: (a,b) Spatial clustering results based on POIs, (c) Results of spatial anisotropy analysis: stations S2, S3, S8, S39, and S45 are strongly correlated with the target station.

Considering that redundant features negatively affect the model performance, Pearson's correlation coefficient was further used in this study to analyze the influences of air pollutant concentration and meteorological factors on PM_{2.5} concentration at strongly correlated stations. The results are presented in Figure 6. Pearson's correlation coefficient ranges from -1 to 1 , where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear correlation between the variables. The final results of the calculations are displayed in Figure 1, revealing the strongest correlation between PM_{2.5} and PM₁₀ at the target station, followed by the selected strongly correlated stations. Notably, NO₂, SO₂, and humidity exhibited weak correlations with PM_{2.5}, with

correlation coefficients $|r| < 0.1$. Therefore, this study excluded NO_2 , SO_2 , and humidity from the prediction modelling process.

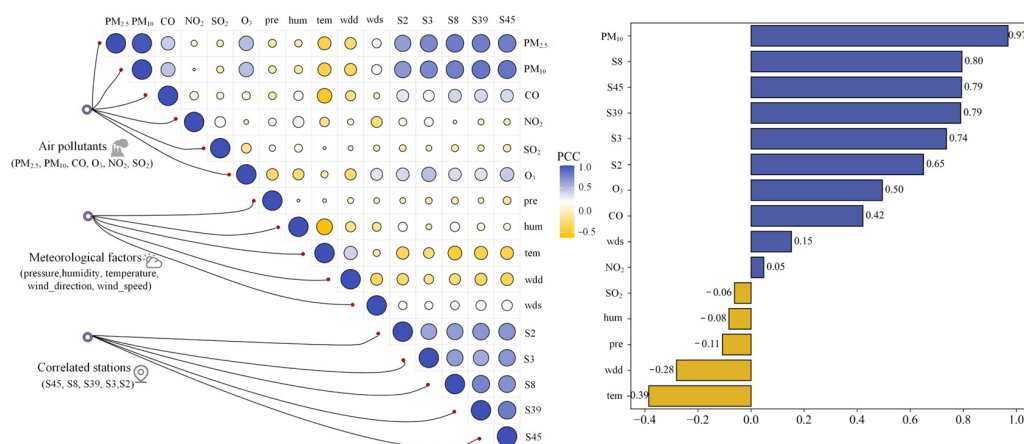


Figure 6. Correlations between the influencing factors.

4.4. Point-Prediction Performance Analysis

To verify the effectiveness of ConvFormer for long-term prediction of $\text{PM}_{2.5}$, this study analyzed the predictive performance of six baseline models (ns_Transformer [54], Informer [55], Autoformer [56], Reformer [57], Pyraformer [58], and LightTS [59]) in comparison with the proposed approach, validating its effectiveness. These baseline models currently achieve good results in long-term series prediction tasks. Comparison with these models enabled a better assessment of the performance advantages of the ConvFormer model in the task of long-term prediction of $\text{PM}_{2.5}$.

The chosen station for the experiment was S9. Experiments were performed to predict the $\text{PM}_{2.5}$ concentration at the station for time intervals of 24, 48, and 96 h. The accuracy of the prediction models was assessed by employing metrics such as MAE, RMSE, and R^2 . These metrics measured the disparity between the observed and predicted values. The experimental results are summarized in Table 3. In summary, the ConvFormer model had superior performance in predicting $\text{PM}_{2.5}$ levels, with the lowest errors in terms of RMSE and MAE, and the highest R^2 score. Specifically, compared with the baseline models, in the prediction tasks of $t + 24$, $t + 48$, and $t + 96$, the MAE of ConvFormer decreased by 8.44%, 9.31%, and 8.13% on average, the RMSE was reduced by 7.94%, 10.89%, and 9.12% on average, and the R^2 increased by 10.29%, 26.2%, and 27.47% on average. Among the baseline models, the ns_Transformer, Informer, Autoformer, and Reformer models are modifications based on the Transformer, and perform well in long-term prediction tasks [24]. However, in our air pollution prediction experiments, the Transformer model significantly outperformed these models.

Table 3. Comparison of the performance of different point-prediction models.

| Model | $t+24$ | | | $t+48$ | | | $t+96$ | | |
|----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | MAE | RMSE | R^2 | MAE | RMSE | R^2 | MAE | RMSE | R^2 |
| ns_Transformer | 4.956 | 8.075 | 0.603 | 6.432 | 10.648 | 0.352 | 7.237 | 11.179 | 0.232 |
| Informer | 5.282 | 8.45 | 0.566 | 7.436 | 10.778 | 0.336 | 6.749 | 10.443 | 0.331 |
| Autoformer | 5.206 | 8.664 | 0.544 | 6.333 | 10.001 | 0.428 | 7.359 | 11.417 | 0.198 |
| Reformer | 4.960 | 7.817 | 0.628 | 6.179 | 9.549 | 0.479 | 6.390 | 10.489 | 0.324 |
| Pyraformer | 5.092 | 8.120 | 0.599 | 5.700 | 8.879 | 0.549 | 6.146 | 9.458 | 0.442 |
| LightTS | 4.616 | 7.516 | 0.656 | 5.856 | 9.134 | 0.523 | 6.170 | 9.845 | 0.405 |
| ConvFormer | 4.595 | 7.463 | 0.661 | 5.799 | 8.760 | 0.561 | 6.132 | 9.516 | 0.444 |

Figures 7–9 visually illustrate the changes in MAE, RMSE, and R^2 values for each model in prediction tasks over different time periods. It is evident from these figures that the prediction accuracies of all models decreased as the prediction time increased. One possible explanation for this phenomenon is that as the prediction time lengthened, it became increasingly challenging for the models to capture the long-term dependencies between historical time-series information, thus affecting prediction accuracy.

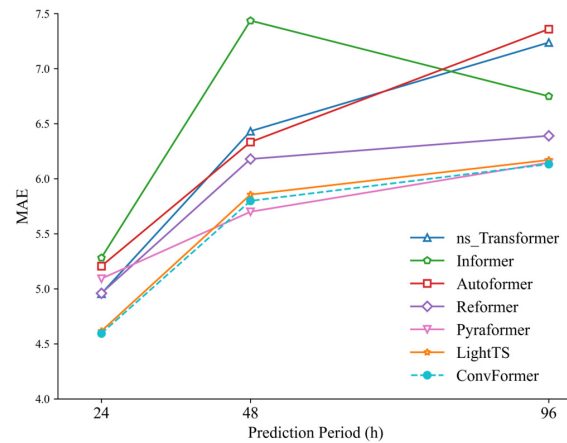


Figure 7. MAE of ConvFormer and the baseline models.

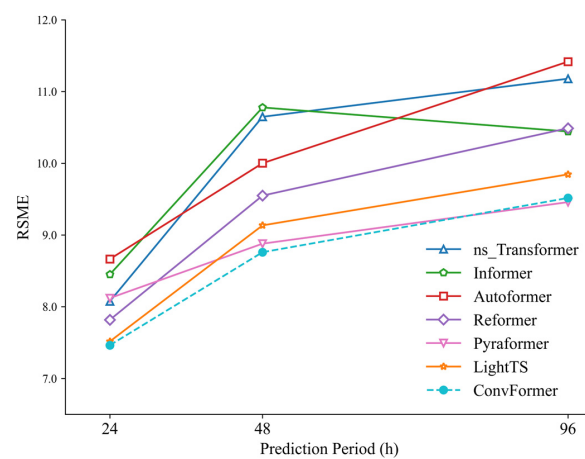


Figure 8. RMSE of ConvFormer and the baseline models.

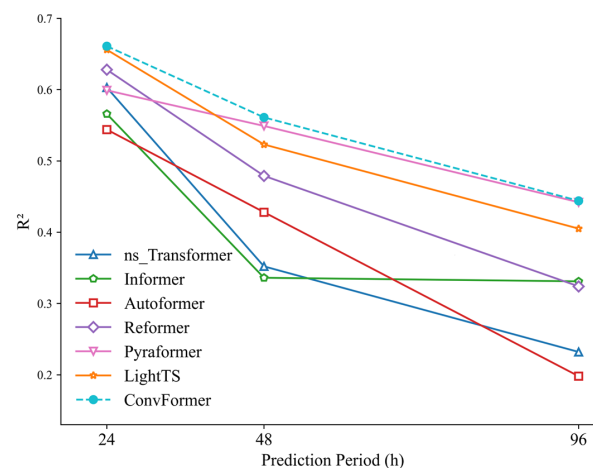


Figure 9. R^2 of ConvFormer and the baseline models.

4.5. Interval-Prediction Performance Analysis

In order to compare the interval-prediction effectiveness of different models based on the KDE method, this study compared the interval-prediction results of different models using the same training set and test set. The experimental results are summarized in Table 4. For the various point-prediction models, the PICP of all models exceeded the preset confidence level at different confidence levels and prediction times. Figures 10–12 depict the PICP and PINAW values for each model under different prediction t tasks. Specifically, the ConvFormer-KDE model achieved the highest PICP and the lowest PINAW in the $t + 24$ prediction task at the same confidence level. PICP measures the proportion of actual observations in the prediction interval, typically ranging from 0 to 1, with values closer to 1 indicating more accurate predictions that cover more actual observations. PINAW measures the width of the prediction intervals. Combining PICP and PINAW can assess the overall performance of interval prediction. As a result, in the $t + 24$ prediction task, the ConvFormer-KDE demonstrated the best performance, making it able to provide more accurate PM_{2.5} information for the public and air pollution prevention workers. In the $t + 48$ and $t + 96$ prediction tasks, the ConvFormer-KDE also achieved relatively high PICP and PINAW. It is worth noting that while the Reformer ranked in the middle of all models in terms of predictive performance in the $t + 96$ point-prediction task, it achieved the best PICP values at all confidence levels in the $t + 96$ interval-prediction task. However, it did not have the lowest PINAW. This suggests that the Reformer model was overly conservative in the $t + 96$ prediction task, resulting in extensive prediction intervals. Although this ensured a high PICP, the prediction uncertainty remained high. Therefore, overall, the ConvFormer-KDE demonstrated the best interval-prediction performance.

Table 4. Comparison of the performance of different interval-prediction models under different confidence levels.

| Confidence Levels | Model | $t+24$ | | $t+48$ | | $t+96$ | |
|-------------------|----------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | | PICP | PINAW | PICP | PINAW | PICP | PINAW |
| $\alpha = 95\%$ | ns_Transformer | 0.9517 | 0.3661 | 0.9510 | 0.4633 | 0.9512 | 0.5328 |
| | Informer | 0.9518 | 0.3874 | 0.9513 | 0.4740 | 0.9512 | 0.4835 |
| | Autoformer | 0.9522 | 0.3798 | 0.9523 | 0.4515 | 0.9507 | 0.5535 |
| | Reformer | 0.9524 | 0.3544 | 0.9517 | 0.4097 | 0.9559 | 0.4368 |
| | Pyraformer | 0.9516 | 0.3713 | 0.9520 | 0.3823 | 0.9526 | 0.4251 |
| | LightTS | 0.9539 | 0.3352 | 0.9520 | 0.3974 | 0.9522 | 0.4535 |
| | ConvFormer-KDE | 0.9542 | 0.3351 | 0.9531 | 0.3946 | 0.9527 | 0.4340 |
| $\alpha = 90\%$ | ns_Transformer | 0.905 | 0.2589 | 0.9035 | 0.3234 | 0.9025 | 0.3827 |
| | Informer | 0.9051 | 0.2718 | 0.9039 | 0.3435 | 0.9041 | 0.3400 |
| | Autoformer | 0.9072 | 0.2634 | 0.9054 | 0.3230 | 0.9033 | 0.3846 |
| | Reformer | 0.9072 | 0.2558 | 0.9056 | 0.2958 | 0.9119 | 0.3159 |
| | Pyraformer | 0.9038 | 0.2572 | 0.9069 | 0.2730 | 0.9058 | 0.3067 |
| | LightTS | 0.9106 | 0.2414 | 0.9053 | 0.2876 | 0.9053 | 0.3248 |
| | ConvFormer-KDE | 0.9125 | 0.2361 | 0.9079 | 0.2871 | 0.9072 | 0.3057 |
| $\alpha = 85\%$ | ns_Transformer | 0.8603 | 0.2039 | 0.8568 | 0.2532 | 0.8544 | 0.3011 |
| | Informer | 0.8589 | 0.2146 | 0.8565 | 0.2745 | 0.8594 | 0.2685 |
| | Autoformer | 0.8637 | 0.2080 | 0.8603 | 0.2519 | 0.8569 | 0.3000 |
| | Reformer | 0.8636 | 0.2051 | 0.8616 | 0.2371 | 0.8701 | 0.2552 |
| | Pyraformer | 0.8617 | 0.1980 | 0.8627 | 0.2192 | 0.8589 | 0.2437 |
| | LightTS | 0.8684 | 0.1933 | 0.8609 | 0.2340 | 0.8606 | 0.2558 |
| | ConvFormer-KDE | 0.8711 | 0.1900 | 0.8638 | 0.2300 | 0.8663 | 0.2434 |

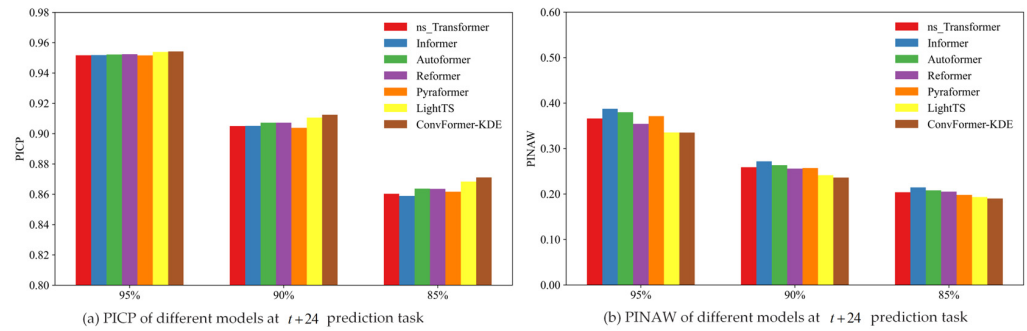


Figure 10. The interval-prediction performance of different models at $t + 24$ prediction task.

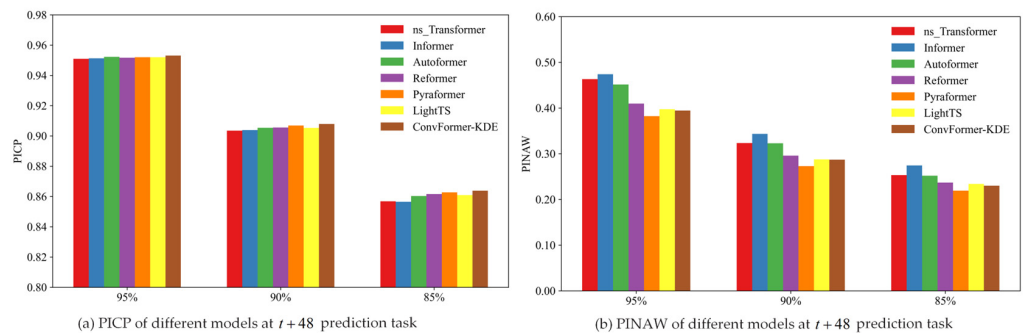


Figure 11. The interval-prediction performance of different models at $t + 48$ prediction task.

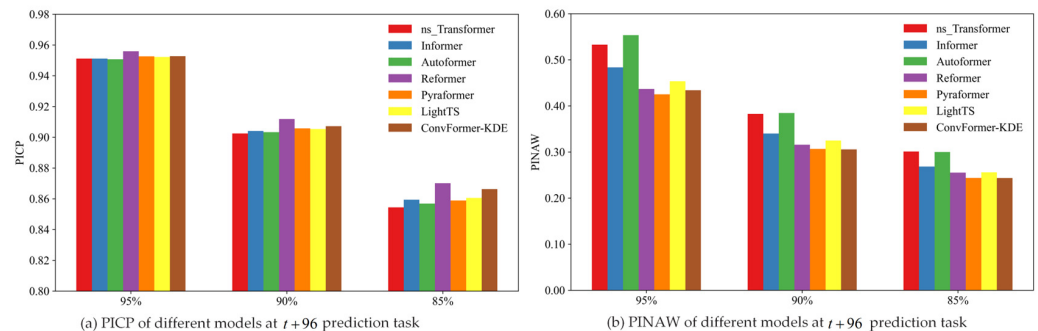


Figure 12. The interval-prediction performance of different models at $t + 96$ prediction task.

4.6. The Result of Ablation Experiment

To further evaluate the effectiveness of ConvFormer-KDE in this study, the employed network modules (e.g., convolutional neural networks and transformers) were thoroughly tested to assess their effectiveness in extracting input features. In the experiments, the proposed method was compared with a single CNN and Transformer model, respectively. During the experiment, the input variables of the three models remained consistent. The experimental results are shown in Tables 5 and 6. The results showed that the combination of CNN and Transformer improved the prediction learning performance compared with relying only on the CNN or Transformer for prediction learning. Specifically, in terms of point prediction, the ConvFormer-KDE improved the R^2 by 2.16%, 4.66%, and 8.78%, respectively, compared with CNN in the prediction tasks of $t + 24$, $t + 48$, and $t + 96$. Compared to Transformer, the ConvFormer-KDE improved the R^2 by 4.09%, 15.67%, and 26.85% for the prediction tasks of $t + 24$, $t + 48$, and $t + 96$. In terms of interval prediction, the performance of ConvFormer-KDE was generally superior that of the CNN and Transformer models at different confidence levels. Therefore, the experimental results demonstrate the effectiveness of the ConvFormer-KDE model combining a CNN with the Transformer, indicating that it is highly capable of predicting the long-term $PM_{2.5}$ concentration with significant advantageous performance.

Table 5. The results of ablation experiments for point prediction.

| Model | <i>t</i> +24 | | | <i>t</i> +48 | | | <i>t</i> +96 | | |
|----------------|--------------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
| | MAE | RMSE | R ² | MAE | RMSE | R ² | MAE | RMSE | R ² |
| CNN | 4.671 | 7.623 | 0.647 | 5.706 | 9.014 | 0.536 | 6.247 | 9.851 | 0.405 |
| Transformer | 4.79 | 7.745 | 0.635 | 5.871 | 9.492 | 0.485 | 6.603 | 10.282 | 0.350 |
| ConvFormer-KDE | 4.595 | 7.463 | 0.661 | 5.799 | 8.760 | 0.561 | 6.132 | 9.516 | 0.444 |

Table 6. The results of ablation experiments for interval prediction.

| Confidence Levels | Model | <i>t</i> +24 | | <i>t</i> +48 | | <i>t</i> +96 | |
|-------------------|----------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | | PICP | PINAW | PICP | PINAW | PICP | PINAW |
| $\alpha = 95\%$ | CNN | 0.9532 | 0.3336 | 0.9521 | 0.3922 | 0.9518 | 0.4561 |
| | Transformer | 0.9525 | 0.3335 | 0.9531 | 0.4027 | 0.9514 | 0.4641 |
| | ConvFormer-KDE | 0.9542 | 0.3351 | 0.9531 | 0.3946 | 0.9527 | 0.4340 |
| $\alpha = 90\%$ | CNN | 0.9088 | 0.2399 | 0.9048 | 0.2848 | 0.9042 | 0.3292 |
| | Transformer | 0.9077 | 0.2388 | 0.9080 | 0.2813 | 0.9037 | 0.3256 |
| | ConvFormer-KDE | 0.9125 | 0.2361 | 0.9079 | 0.2871 | 0.9072 | 0.3057 |
| $\alpha = 85\%$ | CNN | 0.8650 | 0.1907 | 0.8582 | 0.2355 | 0.8591 | 0.2577 |
| | Transformer | 0.8636 | 0.1918 | 0.8643 | 0.2252 | 0.8570 | 0.2576 |
| | ConvFormer-KDE | 0.8711 | 0.1900 | 0.8638 | 0.2300 | 0.8663 | 0.2434 |

5. Discussion

This study proposes a prediction framework based on the ConvFormer-KDE model, which combines CNN, Transformer, and KDE techniques to obtain long-term point and interval prediction for PM_{2.5} concentration. In selecting influencing factors, some meteorological factors and other pollution factors cannot be ignored. Therefore, PM₁₀, CO, O₃, wind speed, temperature, pressure, and wind direction, which are highly correlated to PM_{2.5}, were included in the modelling of this study. In addition, PM_{2.5} values from stations that were strongly correlated with the target station were used as input to the model. In the modelling process, unlike previous studies where the final prediction results are obtained after integrating the prediction results of two separate models, this study converts the CNN-extracted features into the input dimensions required by the Transformer model. Subsequently, the Transformer is utilized to mine the long-term dependencies in the time series to obtain the prediction results. The ConvFormer-KDE takes full advantage of different deep-learning modules. CNN is able to learn temporal relationships and interactions between multivariate variables, and the multi-attention mechanism of the Transformer enables the model to track each data point in relation to another specific data point, allowing it to capture long-term dependencies between temporal sequences. In terms of output strategy selection, this method employs direct output of multiple predicted duration values simultaneously rather than recursively training multiple models. The key to recursive multistep prediction methods lies in continuously updating the dataset and utilizing the updated dataset to make predictions. These methods have the problem of error accumulation becoming worse as the prediction time increases, since each prediction builds on the previous one. The direct multi-output approach chosen in this study can alleviate this problem, and the model structure is simpler and more efficient in terms of computational efficiency.

In terms of interval prediction, directly predicting the upper and lower bounds of intervals often necessitates specifying a fixed interval width, making it challenging to calculate the uncertainty of prediction results. Therefore, implementing interval prediction based on point-prediction results is more widely used. The point-prediction model used in this study plays an important role in interval prediction. A preliminary analysis of the point-prediction error is performed and a probability density function (PDF) of the point-prediction error is constructed using the KDE method. Subsequently, the cumulative

distribution function (CDF) is employed to depict the distribution of the error at a specific confidence level, and the upper and lower bounds of the interval prediction are ultimately derived at the designated confidence level. In KDE, the selection of the kernel function holds significant importance as it directly impacts the level of smoothing and the bias–variance trade-off of the estimation. In this study, the Gaussian function was selected as the kernel function for KDE. Typically, the Gaussian function offers smoother characteristics compared with other kernel functions, resulting in a more continuous and smoother distribution of weights within the observations and yielding a smaller bias.

6. Conclusions and Future Directions

6.1. Summary of Experimental Results

This study proposes a method that combines a CNN, the Transformer model, and kernel density estimation to achieve long-term point and interval prediction of PM_{2.5} concentration. The effectiveness and stability of this model were verified with data from Haikou. Compared with a range of baseline models, including ns_Transformer, Informer, Autoformer, Reformer, Pyraformer, and LightTS, the experimental results showed that the ConvFormer-KDE provided results that were closer to the actual values and the prediction performance was better than the baseline models. The experimental results and analysis demonstrated that the ConvFormer-KDE performed better on the task of long-term point and interval prediction of PM_{2.5}, with good prediction generalization ability and robustness, providing a new direction for PM_{2.5} prediction.

6.2. Caveats and Future Directions

There are still some limitations in this study. The ConvFormer-KDE in this study requires a large amount of data for training to adequately capture long-term dependencies in sequences. Therefore, the model's prediction performance with small quantities of sample data is poor. Secondly, this study applied the proposed approach only to PM_{2.5} concentration prediction, and the ability to make accurate long-term predictions of multiple air pollutants simultaneously is an important target for future research. Future research will focus on the above research directions, and the model proposed in this study will play an important role in these follow-up works.

Author Contributions: Conceptualization, X.L.; methodology, X.L. and Y.Z.; software, Y.Z. and X.F.; formal analysis, S.L. and X.L.; investigation, Y.Z. and X.F.; resources, S.L.; data curation, Y.Z.; writing—original draft preparation, Y.Z. and X.F.; writing—review and editing, S.L., X.L. and Q.M.; visualization, Y.Z. and X.F.; supervision, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program, grant number 2020YFB2104400.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Acknowledgments: The authors would like to express sincere gratitude to the anonymous reviewers for their valuable feedback and constructive comments on the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Peden, D.B. Respiratory Health Effects of Air Pollutants. *Immunol. Allergy Clin. N. Am.* **2024**, *44*, 15–33. [[CrossRef](#)] [[PubMed](#)]
2. Petrou, I.; Psistaki, K.; Kassomenos, P.A.; Dokas, I.M.; Paschalidou, A.K. Studying the Economic Burden of Premature Mortality Related to PM_{2.5} and O₃ Exposure in Greece between 2004 and 2019. *Atmos. Pollut. Res.* **2024**, *15*, 101978. [[CrossRef](#)]
3. Ding, J.; Ren, C.; Wang, J.; Feng, Z.; Cao, S.-J. Spatial and Temporal Urban Air Pollution Patterns Based on Limited Data of Monitoring Stations. *J. Clean. Prod.* **2024**, *434*, 140359. [[CrossRef](#)]

4. Ding, Z.; Chen, H.; Zhou, L.; Wang, Z. A Forecasting System for Deterministic and Uncertain Prediction of Air Pollution Data. *Expert Syst. Appl.* **2022**, *208*, 118123. [[CrossRef](#)]
5. Fang, W.; Zhu, R.; Lin, J.C.-W. An Air Quality Prediction Model Based on Improved Vanilla LSTM with Multichannel Input and Multiroute Output. *Expert Syst. Appl.* **2023**, *211*, 118422. [[CrossRef](#)]
6. Li, H.; Yu, Y.; Huang, Z.; Sun, S.; Jia, X. A Multi-Step ahead Point-Interval Forecasting System for Hourly PM_{2.5} Concentrations Based on Multivariate Decomposition and Kernel Density Estimation. *Expert Syst. Appl.* **2023**, *226*, 120140. [[CrossRef](#)]
7. Nguyen, G.T.H.; La, L.T.; Hoang-Cong, H.; Le, A.H. An Exploration of Meteorological Effects on PM_{2.5} Air Quality in Several Provinces and Cities in Vietnam. *J. Environ. Sci.* **2024**, *145*, 139–151. [[CrossRef](#)] [[PubMed](#)]
8. Yang, H.; Liu, Z.; Li, G. A New Hybrid Optimization Prediction Model for PM_{2.5} Concentration Considering Other Air Pollutants and Meteorological Conditions. *Chemosphere* **2022**, *307*, 135798. [[CrossRef](#)] [[PubMed](#)]
9. Maltare, N.N.; Vahora, S. Air Quality Index Prediction Using Machine Learning for Ahmedabad City. *Digit. Chem. Eng.* **2023**, *7*, 100093. [[CrossRef](#)]
10. Wang, W.; Wang, B.; Chau, K.; Xu, D. Monthly Runoff Time Series Interval Prediction Based on WOA-VMD-LSTM Using Non-Parametric Kernel Density Estimation. *Earth Sci. Inf.* **2023**, *16*, 2373–2389. [[CrossRef](#)]
11. Handhayani, T. An Integrated Analysis of Air Pollution and Meteorological Conditions in Jakarta. *Sci. Rep.* **2023**, *13*, 5798. [[CrossRef](#)] [[PubMed](#)]
12. Li, X.; Li, S.; Tian, S.; Guan, Y.; Liu, H. Air Quality and the Spatial-Temporal Differentiation of Mechanisms Underlying Chinese Urban Human Settlements. *Land* **2021**, *10*, 1207. [[CrossRef](#)]
13. Wang, C.; Zhu, Y.; Zang, T.; Liu, H.; Yu, J. Modeling Inter-Station Relationships with Attentive Temporal Graph Convolutional Network for Air Quality Prediction. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Virtual, 8 March 2021; Association for Computing Machinery: New York, NY, USA; pp. 616–634.
14. Seng, D.; Zhang, Q.; Zhang, X.; Chen, G.; Chen, X. Spatiotemporal Prediction of Air Quality Based on LSTM Neural Network. *Alex. Eng. J.* **2021**, *60*, 2021–2032. [[CrossRef](#)]
15. Sui, S.; Han, Q. Multi-View Multi-Task Spatiotemporal Graph Convolutional Network for Air Quality Prediction. *Sci. Total Environ.* **2023**, *893*, 164699. [[CrossRef](#)] [[PubMed](#)]
16. Park, S.-Y.; Yoon, D.-K.; Park, S.-H.; Jeon, J.-I.; Lee, J.-M.; Yang, W.-H.; Cho, Y.-S.; Kwon, J.; Lee, C.-M. Proposal of a Methodology for Prediction of Indoor PM_{2.5} Concentration Using Sensor-Based Residential Environments Monitoring Data and Time-Divided Multiple Linear Regression Model. *Toxics* **2023**, *11*, 526. [[CrossRef](#)] [[PubMed](#)]
17. Li, H.; Yang, T.; Du, Y.; Tan, Y.; Wang, Z. Interpreting Hourly Mass Concentrations of PM_{2.5} Chemical Components with an Optimal Deep-Learning Model. *J. Environ. Sci.* **2025**, *151*, 125–139. [[CrossRef](#)]
18. Jin, X.-B.; Yang, N.-X.; Wang, X.-Y.; Bai, Y.-T.; Su, T.-L.; Kong, J.-L. Deep Hybrid Model Based on EMD with Classification by Frequency Characteristics for Long-Term Air Quality Prediction. *Mathematics* **2020**, *8*, 214. [[CrossRef](#)]
19. Nguyen, N.P.; Duong, T.A.; Jan, P. Strategies of Multi-Step-Ahead Forecasting for Chaotic Time Series Using Autoencoder and LSTM Neural Networks: A Comparative Study. In Proceedings of the 5th International Conference on Image Processing and Machine Vision, Macau, China, 13–15 January 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 55–61.
20. Sundararajan, A.; Olama, M.; Ferrari, M.; Ollis, B.; Chen, Y.; Liu, G. Recursive Blind Forecasting of Photovoltaic Generation and Consumer Load for Microgrids. In Proceedings of the 2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 16–19 January 2023; pp. 1–5.
21. Dolgintseva, E.; Wu, H.; Petrosian, O.; Zhadan, A.; Allakhverdyan, A.; Martemyanov, A. Comparison of Multi-Step Forecasting Methods for Renewable Energy. *Energy Syst.* **2024**. [[CrossRef](#)]
22. Harrykissoon, K.; Hosein, P. Recursive vs. Direct Forecasting of Crop Prices. In Proceedings of the 2023 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), Sharjah, United Arab Emirates, 4–6 November 2023; pp. 1–6.
23. Aslam, M.; Kim, J.-S.; Jung, J. Multi-Step Ahead Wind Power Forecasting Based on Dual-Attention Mechanism. *Energy Rep.* **2023**, *9*, 239–251. [[CrossRef](#)]
24. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. *arXiv* **2022**, arXiv:2211.14730. [[CrossRef](#)]
25. Zhang, Y.; Yan, J. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
26. Yang, X.; Ma, X.; Kang, N.; Maihemuti, M. Probability Interval Prediction of Wind Power Based on KDE Method with Rough Sets and Weighted Markov Chain. *IEEE Access* **2018**, *6*, 51556–51565. [[CrossRef](#)]
27. Sun, M.; Feng, C.; Chartan, E.K.; Hodge, B.-M.; Zhang, J. A Two-Step Short-Term Probabilistic Wind Forecasting Methodology Based on Predictive Distribution Optimization. *Appl. Energy* **2019**, *238*, 1497–1505. [[CrossRef](#)]
28. Niu, D.; Sun, L.; Yu, M.; Wang, K. Point and Interval Forecasting of Ultra-Short-Term Wind Power Based on a Data-Driven Method and Hybrid Deep Learning Model. *Energy* **2022**, *254*, 124384. [[CrossRef](#)]
29. Wang, M.; Ying, F. Point and Interval Prediction for Significant Wave Height Based on LSTM-GRU and KDE. *Ocean Eng.* **2023**, *289*, 116247. [[CrossRef](#)]

30. Chen, J.; Lu, J.; Avise, J.C.; DaMassa, J.A.; Kleeman, M.J.; Kaduwela, A.P. Seasonal Modeling of PM_{2.5} in California's San Joaquin Valley. *Atmos. Environ.* **2014**, *92*, 182–190. [[CrossRef](#)]
31. Zhao, L.; Li, Z.; Qu, L. Forecasting of Beijing PM_{2.5} with a Hybrid ARIMA Model Based on Integrated AIC and Improved GS Fixed-Order Methods and Seasonal Decomposition. *Heliyon* **2022**, *8*, e12239. [[CrossRef](#)]
32. Guo, Q.; He, Z.; Wang, Z. Predicting of Daily PM_{2.5} Concentration Employing Wavelet Artificial Neural Networks Based on Meteorological Elements in Shanghai, China. *Toxics* **2023**, *11*, 51. [[CrossRef](#)] [[PubMed](#)]
33. Cao, Q.; Shen, L.; Chen, S.-C.; Pui, D.Y.H. WRF Modeling of PM_{2.5} Remediation by SALSCS and Its Clean Air Flow over Beijing Terrain. *Sci. Total Environ.* **2018**, *626*, 134–146. [[CrossRef](#)] [[PubMed](#)]
34. Peng, B.; Xie, B.; Wang, W.; Wu, L. Enhancing Seasonal PM_{2.5} Estimations in China through Terrain–Wind–Rained Index (TWRI): A Geographically Weighted Regression Approach. *Remote Sens.* **2024**, *16*, 2145. [[CrossRef](#)]
35. Tao, H.; Ahmadianfar, I.; Goliatt, L.; Ul Hassan Kazmi, S.S.; Yassin, M.A.; Oudah, A.Y.; Homod, R.Z.; Togun, H.; Yaseen, Z.M. PM_{2.5} Concentration Forecasting: Development of Integrated Multivariate Variational Mode Decomposition with Kernel Ridge Regression and Weighted Mean of Vectors Optimization. *Atmos. Pollut. Res.* **2024**, *15*, 102125. [[CrossRef](#)]
36. Gokul, P.R.; Mathew, A.; Bhosale, A.; Nair, A.T. Spatio-Temporal Air Quality Analysis and PM_{2.5} Prediction over Hyderabad City, India Using Artificial Intelligence Techniques. *Ecol. Inform.* **2023**, *76*, 102067. [[CrossRef](#)]
37. Kumar, V.; Sahu, M. Evaluation of Nine Machine Learning Regression Algorithms for Calibration of Low-Cost PM_{2.5} Sensor. *J. Aerosol Sci.* **2021**, *157*, 105809. [[CrossRef](#)]
38. Gao, Z.; Chen, J.; Wang, G.; Ren, S.; Fang, L.; Yinglan, A.; Wang, Q. A Novel Multivariate Time Series Prediction of Crucial Water Quality Parameters with Long Short-Term Memory (LSTM) Networks. *J. Contam. Hydrol.* **2023**, *259*, 104262. [[CrossRef](#)] [[PubMed](#)]
39. Luo, Z.; Huang, F.; Liu, H. PM_{2.5} Concentration Estimation Using Convolutional Neural Network and Gradient Boosting Machine. *J. Environ. Sci.* **2020**, *98*, 85–93. [[CrossRef](#)] [[PubMed](#)]
40. Bakht, A.; Sharma, S.; Park, D.; Lee, H. Deep Learning-Based Indoor Air Quality Forecasting Framework for Indoor Subway Station Platforms. *Toxics* **2022**, *10*, 557. [[CrossRef](#)] [[PubMed](#)]
41. Wang, B.; Kong, W.; Zhao, P. An Air Quality Forecasting Model Based on Improved Convnet and RNN. *Soft. Comput.* **2021**, *25*, 9209–9218. [[CrossRef](#)]
42. Lipton, Z. A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv* **2015**, arXiv:1506.00019.
43. Zhang, L.; Liu, P.; Zhao, L.; Wang, G.; Zhang, W.; Liu, J. Air Quality Predictions with a Semi-Supervised Bidirectional LSTM Neural Network. *Atmos. Pollut. Res.* **2021**, *12*, 328–339. [[CrossRef](#)]
44. Massaoudi, M.; Chihi, I.; Sidhom, L.; Trabelsi, M.; Refaat, S.S.; Abu-Rub, H.; Oueslati, F.S. An Effective Hybrid NARX-LSTM Model for Point and Interval PV Power Forecasting. *IEEE Access* **2021**, *9*, 36571–36588. [[CrossRef](#)]
45. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; Yan, X. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Proc. Adv. Neural Inf. Process. Syst.* **2019**, *32*. [[CrossRef](#)]
46. Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; Zha, H. Transformer Hawkes Process. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 21 November 2020; pp. 11692–11702.
47. Wu, C.; He, H.; Song, R.; Zhu, X.; Peng, Z.; Fu, Q.; Pan, J. A Hybrid Deep Learning Model for Regional O₃ and NO₂ Concentrations Prediction Based on Spatiotemporal Dependencies in Air Quality Monitoring Network. *Environ. Pollut.* **2023**, *320*, 121075. [[CrossRef](#)] [[PubMed](#)]
48. Kumar, S.; Kumar, V. Multi-View Stacked CNN-BiLSTM (MvS CNN-BiLSTM) for Urban PM_{2.5} Concentration Prediction of India's Polluted Cities. *J. Clean. Prod.* **2024**, *444*, 141259. [[CrossRef](#)]
49. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
50. Ghobadi, F.; Yaseen, Z.M.; Kang, D. Long-Term Streamflow Forecasting in Data-Scarce Regions: Insightful Investigation for Leveraging Satellite-Derived Data, Informer Architecture, and Concurrent Fine-Tuning Transfer Learning. *J. Hydrol.* **2024**, *631*, 130772. [[CrossRef](#)]
51. Du, B.; Huang, S.; Guo, J.; Tang, H.; Wang, L.; Zhou, S. Interval Forecasting for Urban Water Demand Using PSO Optimized KDE Distribution and LSTM Neural Networks. *Appl. Soft Comput.* **2022**, *122*, 108875. [[CrossRef](#)]
52. Sun, M.; Feng, C.; Zhang, J. Conditional Aggregated Probabilistic Wind Power Forecasting Based on Spatio-Temporal Correlation. *Appl. Energy* **2019**, *256*, 113842. [[CrossRef](#)]
53. Xu, C.; Sun, Y.; Du, A.; Gao, D. Quantile Regression Based Probabilistic Forecasting of Renewable Energy Generation and Building Electrical Load: A State of the Art Review. *J. Build. Eng.* **2023**, *79*, 107772. [[CrossRef](#)]
54. Liu, Y.; Wu, H.; Wang, J.; Long, M. Non-Stationary Transformers: Exploring the Stationarity in Time Series Forecasting. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9881–9893.
55. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [[CrossRef](#)]
56. Chen, M.; Peng, H.; Fu, J.; Ling, H. AutoFormer: Searching Transformers for Visual Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12270–12280.
57. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The Efficient Transformer. *arXiv* **2020**, arXiv:2001.04451.

-
58. Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A.X.; Dustdar, S. Pyraformer: Low-Complexity Pyramidal Attention for Long-Range Time Series Modeling and Forecasting. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
 59. Campos, D.; Zhang, M.; Yang, B.; Kieu, T.; Guo, C.; Jensen, C.S. LightTS: Lightweight Time Series Classification with Adaptive Ensemble Distillation. *Proc. ACM Manag. Data* **2023**, *1*, 171. [[CrossRef](#)]

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.