atmosphere

MDPI

Article

Enhanced Sequence-to-Sequence Attention-Based $PM_{2.5}$ Concentration Forecasting Using Spatiotemporal Data

Baekcheon Kim, Eunkyeong Kim, Seunghwan Jung, Minseok Kim, Jinyong Kim and Sungshin Kim *

Department of Electrical and Electronics Engineering, Pusan National University, Busan 46241, Republic of Korea; 1001000@pusan.ac.kr (B.K.); kimeunkyeong@pusan.ac.kr (E.K.); shjung0091@pusan.ac.kr (S.J.); minskey@pusan.ac.kr (M.K.); skes1234@pusan.ac.kr (J.K.)
* Correspondence: sskim@pusan.ac.kr

Abstract: Severe air pollution problems continue to increase because of accelerated industrialization and urbanization. Specifically, fine particulate matter ($PM_{2.5}$) causes respiratory and cardiovascular diseases, and according to the World Health Organization (WHO), millions of premature deaths and significant health burdens annually. Therefore, $PM_{2.5}$ concentration forecasting is essential. This study proposed a method to forecast $PM_{2.5}$ concentrations one hour after using Sequence-to-Sequence Attention (Seq2Seq-attention). The proposed method selects neighboring stations using minimum redundancy maximum relevance (mRMR) and integrates their data using a convolutional neural network (CNN). The proposed attention score and Seq2Seq are used on the integrated data to forecast $PM_{2.5}$ concentration after one hour. The performance of the proposed method is validated through two case studies. The first comparison evaluated the performance of the conventional attention score against the proposed attention scores. The second comparison evaluated the forecasting results with and without considering neighboring stations. The first study showed that the proposed attention score improved the performance index (Root Mean Square Error (RMSE): 3.48%p, Mean Absolute Error (MAE): 8.60%p, R^2: 0.49%p, relative Root Mean Square Error (rRMSE): 3.64%p, Percent Bias (PBIAS): 59.29%p). The second case study showed that considering neighboring stations' data can be more effective in forecasting than considering that of a standalone station (RMSE: 5.49%p, MAE: 0.51%p, R^2: 0.67%p, rRMSE: 5.44%p, PBIAS: 46.56%p). This confirmed that the proposed method can effectively forecast the $PM_{2.5}$ concentration after one hour.

Keywords: $PM_{2.5}$ concentration forecasting; minimum redundancy maximum relevance; Sequence-to-Sequence; attention method

Citation: Kim, B.; Kim, E.; Jung, S.; Kim, M.; Kim, J.; Kim, S. Enhanced Sequence-to-Sequence Attention-Based $PM_{2.5}$ Concentration Forecasting Using Spatiotemporal Data. *Atmosphere* **2024**, *15*, 1469. https://doi.org/10.3390/atmos15121469

Academic Editors: Rana Muhammad Adnan, Ozgur Kisi and Mo Wang

Received: 15 November 2024
Revised: 5 December 2024
Accepted: 5 December 2024
Published: 9 December 2024

1. Introduction

With accelerating industrialization and urbanization, air pollution continues to be an increasing environmental problem worldwide [1–3]. Air pollution directly and indirectly affects health, the ecosystem, and climate change. In particular, fine particulate matter ($PM_{2.5}$) is considered a serious threat. $PM_{2.5}$ are fine particles with a diameter of 2.5 µm or less, which can long remain airborne and be easily inhaled. Consequently, $PM_{2.5}$ can penetrate deep into the lungs, causing respiratory and cardiovascular diseases, and even premature death [4–7]. According to the World Health Organization (WHO), the disease burden caused by $PM_{2.5}$ causes millions of deaths annually, mainly affecting weak groups such as the elderly, children, and those with chronic diseases [8]. In addition to health problems, $PM_{2.5}$ causes economic losses such as deterioration of urban appearance, damage to buildings, and decreased agricultural productivity. In particular, high $PM_{2.5}$ concentration problems are recognized as a severe social problem in major cities worldwide, in addition to rapidly industrializing countries, e.g., China and India [9,10]. Accordingly, many countries are making policy efforts to manage air quality and reduce particulate matter. For this, the accurate forecasting of $PM_{2.5}$ concentrations and the development of real-time monitoring technology are emerging as important research topics. This study proposes a forecasting model that utilizes spatiotemporal data. Its goal is to improve forecasting accuracy and support effective air quality management.

The changing concentrations of $PM_{2.5}$ are determined by various atmospheric and other influences. A combination of complex factors such as meteorological conditions, geographical characteristics, traffic volume, and industrial activities hamper accurate forecasting [11,12]. In particular, $PM_{2.5}$ concentrations are temporally and spatially extremely variable, and the concentration appears differently depending on diverse regional characteristics. Therefore, a forecasting method that considers both temporal and spatial characteristics is needed [13].

Traditional forecasting methods that consider temporal characteristics include physics-based and statistical models [14]. Physics-based models mathematically model atmospheric dynamics and chemical reactions to forecast the transport and diffusion of $PM_{2.5}$ concentrations [15–17]. Statistical models analyze patterns based on past data to forecast $PM_{2.5}$ concentrations [18–21]. However, these traditional models cannot accurately reflect the interactions of various complex variables, such as meteorological conditions, anthropogenic factors, and geographical factors. In addition, their performance may deteriorate when data are insufficient or when dealing with complex nonlinear relationships [13,14].

More recently, machine- and deep learning-based models have been introduced to solve the problems in forecasting $PM_{2.5}$ concentrations [14]. Machine learning models produce forecasts by learning from data and have strengths in learning nonlinear and complex patterns. Furthermore, deep learning has attracted particular attention in the field of air quality forecasting due to its ability to process and learn from large amounts of data [22–30]. Compared with statistical models, deep learning models can extract useful features from multidimensional data and learn nonlinear relationships more accurately, often resulting in significantly improved performance [14].

Among deep learning models, the Recurrent Neural Network (RNN) structure that processes time-series data is suitable for sequential data processing and is widely used in $PM_{2.5}$ concentration forecasting [31–33]. RNN is advantageous for time-series data because it can reflect past information in current forecasts but has limitations in processing long-term dependencies. Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) were developed to overcome these shortcomings. LSTM processes long-term dependencies more effectively through cell states and gate structures, and GRUs have more compact structures while providing similar functions [34,35]. These models enable forecasting that considers pattern changes over time, improving $PM_{2.5}$ concentration forecasting accuracy. The Sequence-to-Sequence (Seq2Seq) structure is being used for sophisticated time-series forecasting. The Seq2Seq model was initially developed for tasks like machine translation and speech recognition. Recently, it has also attracted interest in time-series forecasting for its ability to transform input sequences into output sequences [36]. Forecasting models utilizing the Seq2Seq structure can learn complex data patterns with a more flexible structure than simple time-series models.

In addition, attention mechanisms have become widely used in time-series data forecasting. An attention mechanism is a method in which the model gives more weight to important information under the premise that not all input data have the same importance [37]. Since data at a specific time point can significantly impact the overall forecasting, many forecasting studies have used the attention mechanism [38,39].

The method that considers spatial characteristics is relatively more complex than the method that considers temporal characteristics [11]. In prior studies, forecasting methods mainly focused on neighboring stations when considering spatial characteristics [40–45]. These methods are challenging to implement because they involve selecting neighboring stations for forecasting. In particular, excluding the target station and selecting too many stations increases the complexity of the model, which may actually deteriorate its performance. Therefore, it is crucial to select appropriate stations from many stations. Methods that consider neighboring stations include distance-based methods, statistical methods, and data-based methods. Distance-based methods consider the distance between the target and neighboring stations. This intuitive and simple method is limited in reflecting the characteristics of the data because it only considers the distance between stations. Statistical methods better represent data characteristics than distance-based methods by considering spatial correlations between neighboring stations. However, they face challenges in addressing nonlinear relationships between the target and neighboring stations. Data-driven methods can model nonlinear relationships by learning the connections between the target and neighboring stations. However, their black-box characteristics make them difficult to interpret. Significant advancements have been made in employing hybrid methods to forecast $PM_{2.5}$ concentrations by simultaneously accounting for temporal and spatial information. These methods combine two or more models to capitalize on their strengths, improving forecasting accuracy. For example, hybrid approaches often integrate deep learning models with statistical techniques or utilize convolutional neural networks (CNNs) for spatial pattern recognition alongside RNNs or LSTM networks to address temporal dependencies. Such approaches are particularly effective in capturing the intricate dynamics of air quality variations and managing the heterogeneous nature of environmental data.

The challenges in effectively incorporating spatial characteristics into forecasting models highlight the importance of selecting appropriate neighboring stations and accurately modeling their relationships with the target station. While distance-based, statistical, and data-driven methods have strengths, their limitations in capturing complex, nonlinear relationships underscore the need for advanced approaches that balance interpretability and performance.

To address these challenges, this study employs a combination of minimum redundancy maximum relevance (mRMR) and a CNN. By selecting relevant neighboring stations based on correlations between the target station and its neighbors and among the selected stations, mRMR minimizes redundancy in the input data, ensuring that only the most informative stations are considered. Additionally, CNNs effectively capture the nonlinear relationships between the target station and its neighbors, providing a robust framework for integrating spatial information into forecasting models. This combined approach enhances forecasting accuracy and mitigates the shortcomings of previous spatial modeling methods.

This study proposed a method to forecast $PM_{2.5}$ concentration after one hour by considering spatiotemporal aspects using mRMR and a CNN and using Seq2Seq combined with improved attention.The contributions of this study can be outlined as follows:

1. In conventional attention mechanisms, weights are calculated using the dot product, which assigns high weights to inputs with high similarity. This approach affects vectors with high similarity but not those with low similarity. The proposed attention score calculates similarity using both the dot product and the inverse of Euclidean distance, allowing the model to consider both similarity and dissimilarity. Therefore, the proposed attention score enables consideration of both similarity and dissimilarity. This enhancement directly addresses the shortcomings of previous attention mechanisms, enabling better performance in scenarios where dissimilar inputs significantly impact the forecast. By reducing redundancy and accurately modeling nonlinear relationships, the proposed method is expected to outperform traditional and hybrid approaches, as evidenced in the experimental results.
2. When considering neighboring stations, prior studies have usually selected the distance between the target and neighboring stations to consider spatial characteristics or the correlation between stations. However, when using mRMR and a CNN, one can reduce redundancy in the influence of neighboring stations and the nonlinear relationship between the target and neighboring stations can be considered.

The remainder of this paper is organized as follows. Section 2 introduces the related work forecasting on $PM_{2.5}$ concentrations. Section 3 provides a detailed description of the study site, the data used in this study, and the proposed $PM_{2.5}$ concentration forecasting method. Section 4 presents the verification experiments along with the experimental results and performance indices. Finally, Section 5 discusses the experimental results and the conclusions drawn from the study.

2. Related Work

Many studies have been conducted on forecasting $PM_{2.5}$ concentrations to protect human health. The traditional physical method [15–17] is a $PM_{2.5}$ concentration dispersion model based on atmospheric chemistry and aerodynamic theory. This method is heavily dependent on theoretical assumptions, restricting its applicability to specific regions. Furthermore, its substantial computational requirements limit its feasibility for real-time forecasting. In contrast, statistical and data-driven methods are more suitable for real-time applications.

Statistical methods define the relationship between pollutant concentrations, meteorological variables, and air pollution data. Several studies have applied statistical methods to forecast $PM_{2.5}$ concentrations. Zhang et al. [19], Badicu et al. [18], and Wang et al. [21] utilized the Auto Regression Int1egrated Moving Average (ARIMA) model to forecast $PM_{2.5}$ concentrations in Fuzhou, China, Bucharest, Romania, and California, USA, respectively. Similarly, Amnuaylojaroen [20] employed a Multivariate Linear Regression (MLR) model to forecast $PM_{2.5}$ concentrations in Chiang Mai, Lampang, and Nan, Thailand. Although these statistical methods are effective in modeling linear relationships, they are limited in their ability to adequately represent the nonlinear interactions between meteorological conditions and air pollution.

To overcome the limitations of statistical methods, data-driven models have emerged as a promising approach for $PM_{2.5}$ forecasting. These models are particularly effective in capturing the nonlinear relationships between inputs and outputs. Among them, Artificial Neural Networks (ANNs) have been widely utilized to improve forecasting accuracy. Chen et al. [46] applied an ANN to analyze meteorological factors such as temperature, humidity, and wind speed, successfully forecasting $PM_{2.5}$ concentrations in Fuling, Chongqing, China. Similarly, Lightstone et al. [47] employed an ANN to foreacst $PM_{2.5}$ concentrations in New York, USA, demonstrating its superiority over the Community Multiscale Air Quality (CMAQ) model. Additionally, Bera et al. [48] demonstrated that the ANN outperformed the MLR statistical model in forecasting $PM_{2.5}$ concentrations. Collectively, these studies underscore the potential of ANNs to model the nonlinear interactions between meteorological conditions and air pollution, effectively addressing the shortcomings of traditional statistical approaches.

Due to the rapid improvement in computing performance, deep learning models capable of processing large amounts of data simultaneously have become available, with many studies conducted on forecasting $PM_{2.5}$ concentrations using deep learning models. In particular, the RNN, a deep learning model, is widely used for forecasting $PM_{2.5}$ concentrations because it considers historical information. M. Oprea et al. [31] performed a $PM_{2.5}$ concentration forecast around Munich, Germany to compare the performance of an ANN and an RNN. They confirmed that the RNN, which considers the past, produced superior performance. However, despite its strengths, the RNN faces challenges such as long-term dependency issues and gradient vanishing problems [34]. To solve this problem, a study was recently conducted on forecasting using LSTM, a model that modifies RNN cells. Y.T. Tsai et al. [49] used LSTM to forecast the $PM_{2.5}$ concentration one hour in advance after measuring meteorological and air pollution data at 20 monitoring stations in Taiwan for the preceding 72 h. Zhao et al. [50] used an ANN to define the relationship with neighboring stations and forecasted $PM_{2.5}$ concentration in Beijing, China, using LSTM. Ayturan Y.A. et al. [51] used GRU, a modified LSTM model, to forecast $PM_{2.5}$ concentration in Keçiören district, Ankara, Turkey, 1∼3 h in advance.

In addition to these RNN-based models, studies have also been conducted that utilize Seq2Seq models while maintaining the cell structure of the RNN series. Seq2Seq efficiently compresses information in long sequences. It is effective for time-series forecasts as it generates necessary sequences from the compressed data. Wang et al. [52] applied an LSTM-based Seq2Seq model to forecast $PM_{2.5}$ and roadside CO concentrations in Shanghai, China. The model used $PM_{2.5}$ and CO concentration data from the past seven days, previous-day meteorological and air pollution data, and same-day PM_2.5 and CO concentrations. Yan et al. [53] forecasted $PM_{2.5}$ concentrations in Tianjin, China, using a GRU-based Seq2Seq model. The model utilized 72 h of previous air pollution and meteorological data as inputs. It showed superior performance compared to ANN and GRU models.

However, the Seq2Seq model faces limitations in equally processing the significance of all time points in the information compression process. Recent studies have successfully applied the attention mechanism to improve this by assigning weights according to the importance of specific time points in time-series data. This is gaining recognition as an especially effective method for forecasting $PM_{2.5}$ concentrations. Liu et al. [38] forecasted the $PM_{2.5}$ concentration 24 h in advance at the Olympic Center and Dongsi Observatory in Beijing, China, using a Seq2Seq model applying an attention mechanism. Bai et al. [39] forecasted the $PM_{2.5}$ concentration in Beijing by combining the attention mechanism with a GRU-based Seq2Seq model.

In addition to methods considering temporal characteristics, many forecast methods considering spatial characteristics have also been developed. Vargas-Campos et al. [54] used Inverse Distance Weight (IDW) to forecast $PM_{2.5}$ concentration considering spatial characteristics in Beijing, China, and Yang et al. [55] used the k-Nearest Neighbor (kNN) algorithm. Yeo et al. [56] forecasted $PM_{2.5}$ concentration in Seoul considering correlation with neighboring stations. By considering neighboring stations, using a convolutional neural network (CNN), and correlation, Zhang et al. [57] forecasted the $PM_{2.5}$ concentrations in densely populated urban areas in the Yangtze River Delta, China.

Related studies have extensively studied temporal and spatial factors for forecasting $PM_{2.5}$ concentrations. However, effectively integrating temporal and spatial dimensions to forecast remains a challenge. Temporal forecasting methods, particularly those using Seq2Seq models with attention mechanisms, have significantly improved by capturing the relative importance of specific time points [38,39]. Similarly, spatial methods utilizing correlations with neighboring stations and advanced deep learning techniques, such as CNNs, have enhanced the ability to model spatial relationships [56,57]. Despite these advancements, many studies treat temporal and spatial characteristics independently, which limits their ability to fully capture the complex dynamics of air pollution.

This paper overcomes these limitations by introducing an integrated approach that considers temporal and spatial dimensions simultaneously. Unlike related studies, this method uses mRMR to select neighboring stations. It reduces redundant data and includes only the most relevant spatial information. CNN also integrates data from neighboring stations and models nonlinear relationships between neighboring stations. Temporal information is enhanced through a Seq2Seq model combined with an improved attention mechanism. By integrating the dot product and inverse Euclidean distance, the proposed attention mechanism accounts for both similarity and dissimilarity, addressing a critical gap in the previous attention-based models.

3. Materials and Method

Figure 1 presents a flowchart of the proposed method for forecasting $PM_{2.5}$ concentration. The process is divided into two main steps: the preprocessing step, denoted by red dotted lines, and the forecasting step, denoted by blue dotted lines.

The preprocessing steps include outlier detection and elimination, data interpolation, data normalization, data separation between the main station and neighbor stations, and the selection of neighbor stations. Missing values are handled through linear interpolation. Min-max normalization is applied to the data, scaling the values to a range between -1 and 1. Neighbor stations are selected based on their relevance using the mRMR method.The forecast step uses the processed data to forecast the $PM_{2.5}$ concentration one hour ahead. The data from neighbor stations are first integrated through the CNN. Subsequently, the Seq2Seq-attention mechanism is applied to the integrated data, along with data from the main station, to forecast the $PM_{2.5}$ concentration one hour ahead.

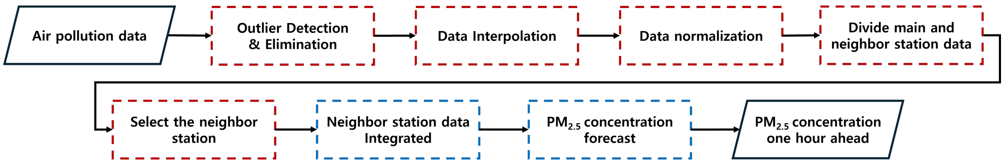

Figure 1. Flowcharts for $PM_{2.5}$ concentration forecasting models.

3.1. Dataset

Due to rapid industrialization and urbanization, the $PM_{2.5}$ concentration in Beijing has increased [58,59]. As a result, numerous studies have focused on forecasting $PM_{2.5}$ concentrations in this region to mitigate the adverse health and environmental impacts [38,39,50,54]. The severity of air pollution in Beijing, combined with the growing body of research, makes it a crucial area for studying and developing more accurate forecasting models. For these reasons, this study has chosen Beijing as the target region to explore advanced forecasting techniques for $PM_{2.5}$ concentrations. In this study, we used data provided by the Microsoft Research Urban Computing Team to forecast $PM_{2.5}$ concentrations in the Beijing area. The data were provided the Microsoft Research Urban Computing Team [60]. The data were recorded every hour from 1 May 2014, to 30 April 2015. These data comprise six air pollution categories($PM_{2.5}$, PM_{10}, sulfur dioxide (SO_2), ozone (O_3), nitrogen dioxide (NO_2), and carbon monoxide (CO)) along with six categories of meteorological data (temperature, pressure, humidity, wind speed, wind direction, and weather). Meteorological data are provided at the district level, whereas air pollution data are available as monitoring data from individual observation stations. Due to the inherent redundancy associated with district-level meteorological data, this study excludes their use to mitigate potential redundancy issues. Instead, the forecasting used only air pollution data collected from monitoring stations. This approach ensures finer spatial resolution and enhances the accuracy and relevance of the forecasting model. Table 1 provides the units and formats of the data utilized for forecasting.

Figure 2 shows the locations of the stations on the map, while Table 2 lists the coordinates for each station. Figure 3 displays the missing ratio of air pollution data for each station, where the x-axis represents the station ID and the y-axis denotes the percentage of missing data. Blue, orange, green, red, purple, and brown correspond to $PM_{2.5}$, PM_{10}, NO_2, CO, O_3, and SO_2, respectively. The bars represent the missing ratio for each pollutant, while the line indicates the mean missing ratio across all pollutants. As shown in Figure 3, $PM_{2.5}$ exhibits the lowest missing rate, whereas PM_{10} shows the highest. Furthermore, Station 1022 has the highest missing rate across all pollutants except for $PM_{2.5}$. As a result, this study excludes Station 1022 and utilizes data from the remaining stations. For these stations, missing data were interpolated and used in the forecast.

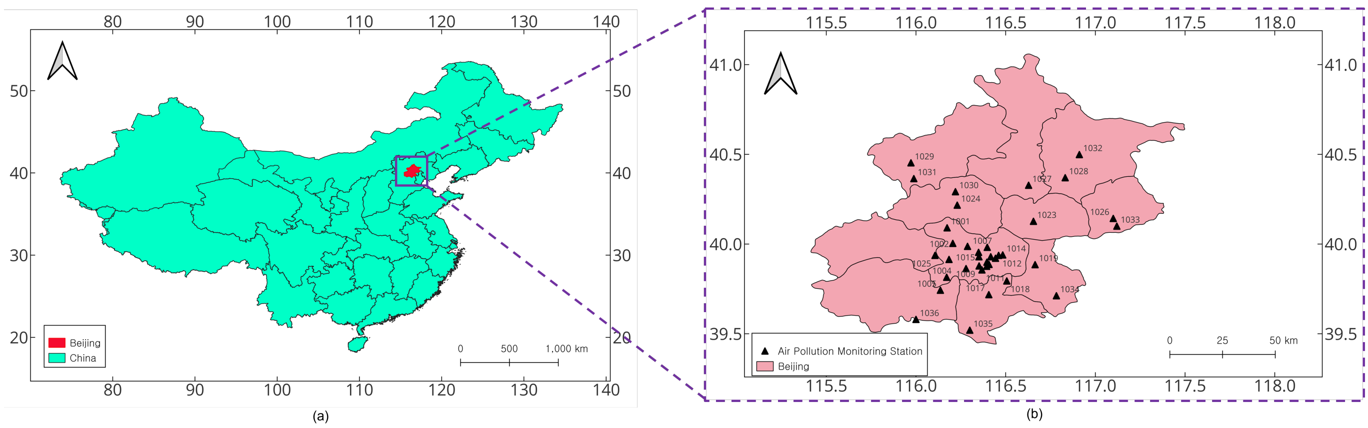

Figure 2. Locations of air pollution monitoring stations: (**a**) Beijing's location within China. (**b**) Air pollution monitoring station locations in Beijing.

Table 1. Sampling time and units for data variables.

Variable Type	Variable Name	Data Type	Unit	Sampling Time
Air pollution	PM_{10}	Numerical	$\mu g/m^3$	1 h
	$PM_{2.5}$	Numerical	$\mu g/m^3$	1 h
	O_3	Numerical	$\mu g/m^3$	1 h
	SO_2	Numerical	$\mu g/m^3$	1 h
	NO_2	Numerical	$\mu g/m^3$	1 h
	CO	Numerical	$\mu g/m^3$	1 h

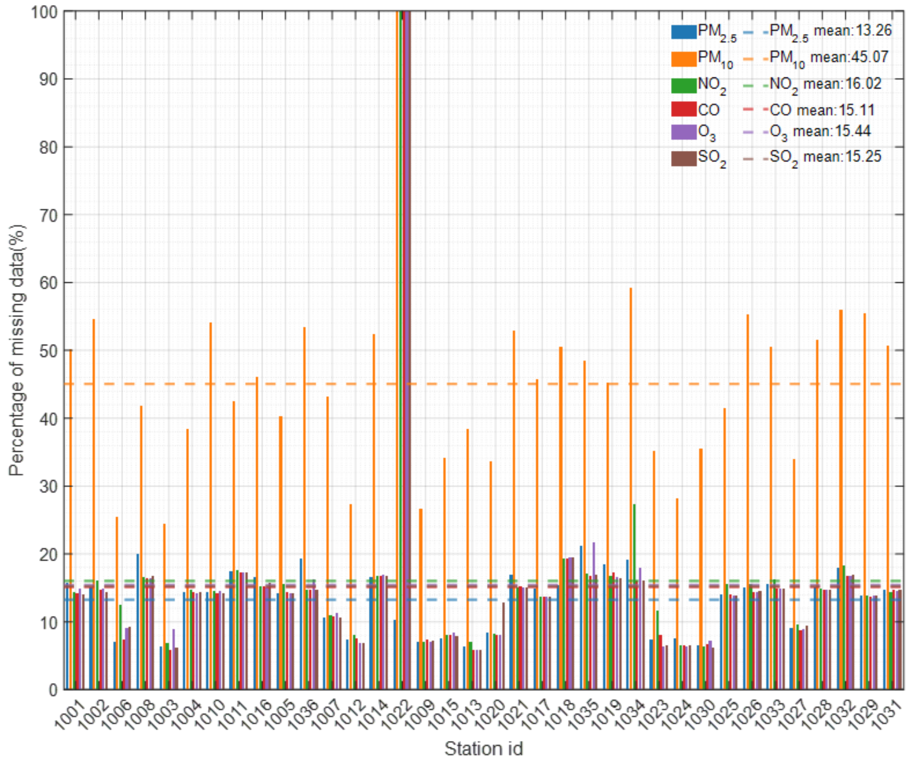

Figure 3. Missing data rate by monitoring stations.

Table 2. Information of air pollution monitoring stations.

Station ID	Latitude	Longitude	Station ID	Latitude	Longitude
1001	40.090679	116.173553	1019	39.885241	116.664162
1002	40.00395	116.20531	1020	39.886491	116.407355
1003	39.914409	116.184239	1021	39.899135	116.395383
1004	39.815128	116.17115	1022	39.920993	116.443448
1005	39.742767	116.136045	1023	40.127	116.655
1006	39.987313	116.287451	1024	40.216999	116.23
1007	39.982053	116.3974	1025	39.936999	116.105999
1008	39.954047	116.348991	1026	40.143	117.099999
1009	39.878193	116.351974	1027	40.328	116.628
1010	39.876184	116.394009	1028	40.369999	116.831999
1011	39.855958	116.36781	1029	40.453	115.971999
1012	39.937119	116.460742	1030	40.292	116.22
1013	39.929287	116.416883	1031	40.365	115.988
1014	39.939554	116.483746	1032	40.499	116.911
1015	39.929302	116.351029	1033	40.1	117.12
1016	39.86347	116.279082	1034	39.712	116.783
1017	39.718147	116.406155	1035	39.52	116.3
1018	39.794491	116.506319	1036	39.579999	116

3.2. Minimum Redundancy Maximum Relevance

$PM_{2.5}$ concentration exhibits a strong spatial correlation with the pollution levels of neighboring stations, as well as temporal variation [61,62]. Forecasting $PM_{2.5}$ concentration using data from a main station alone results in lower performance. In contrast, incorporating data from neighboring stations improves forecasting accuracy [63]. Therefore, it is essential to consider the influence of neighboring stations when forecasting $PM_{2.5}$ concentration. In this study, the mRMR method was employed to select relevant neighboring stations. The mRMR algorithm evaluates the correlations between the standalone station and neighboring stations, effectively reducing redundancy and simplifying the model [64]. Algorithm 1 outlines the process for calculating mRMR. In this algorithm, $I(f_k|Y_*)$ represents the correlation coefficient between the k-th neighboring station and the main station. Equation (1) defines the correlation coefficient between two variables, where COV(f_k, Y_*) is the covariance between the $PM_{2.5}$ concentrations of the k-th neighboring station and the main station, and σ_{f_k} and σ_{Y_*} represent the variances of the $PM_{2.5}$ concentrations at the k-th station and the main station, respectively. In Algorithm 1, the number of subsets of neighboring stations is defined by the user. In this paper, mRMR scores were calculated to two decimal places. After selecting the stations according to the scores, we performed a forecasting experiment to select the optimal number of stations.

Algorithm 1: minimum Redundancy Maximum Relevance (mRMR)

Input: The training dataset D with the neighbor monitoring station's $PM_{2.5}$ concentration set $F = \{f_1, f_2, \ldots, f_n\}$, main monitoring station $PM_{2.5}$ concentration Y_*, and the required number of neighboring stations T.
Output: The selected neighbor monitoring station subset S_T and mRMR score set ρ_T.

```
S_T ← ∅;
ρ_T ← ∅;
for f_i in F do
    MI_i = I(f_i; Y_*);
end
f ← arg max(MI);
S_T ← S_T ∪ {f};
ρ_T ← ρ_T ∪ max(MI);
F ← F − {f};
for i ← 2 to T do
    for each f_i in F do
        J(f_i) = I(f_i|Y_*) − (1/|S_T|) Σ_{f_j ∈ S_T} I(f_j; f_i);
    end
    Select f_s from J(f_i) with the largest value;
    S_T ← S_T ∪ {f_s};
    ρ_T ← ρ_T ∪ J(f_i);
    F ← F − {f_s};
end
return S_T, ρ_T
```

$$I(f_k|Y_*) = \frac{\mathrm{COV}(f_k, Y_*)}{\sigma_{f_k}\sigma_{Y_*}}. \quad (1)$$

3.3. Proposed Network Architecture

This study used the CNN-Seq2Seq-attention model to forecast the main station considering the selected neighboring stations. Figure 4 shows the structure of the proposed network. The proposed network consists of CNN and Seq2Seq-attention. The CNN integrates the selected neighbor station data using mRMR. The Seq2Seq-attention model forecasts the $PM_{2.5}$ concentration one hour ahead using the data of the main station and the neighbor station data integrated through the CNN.

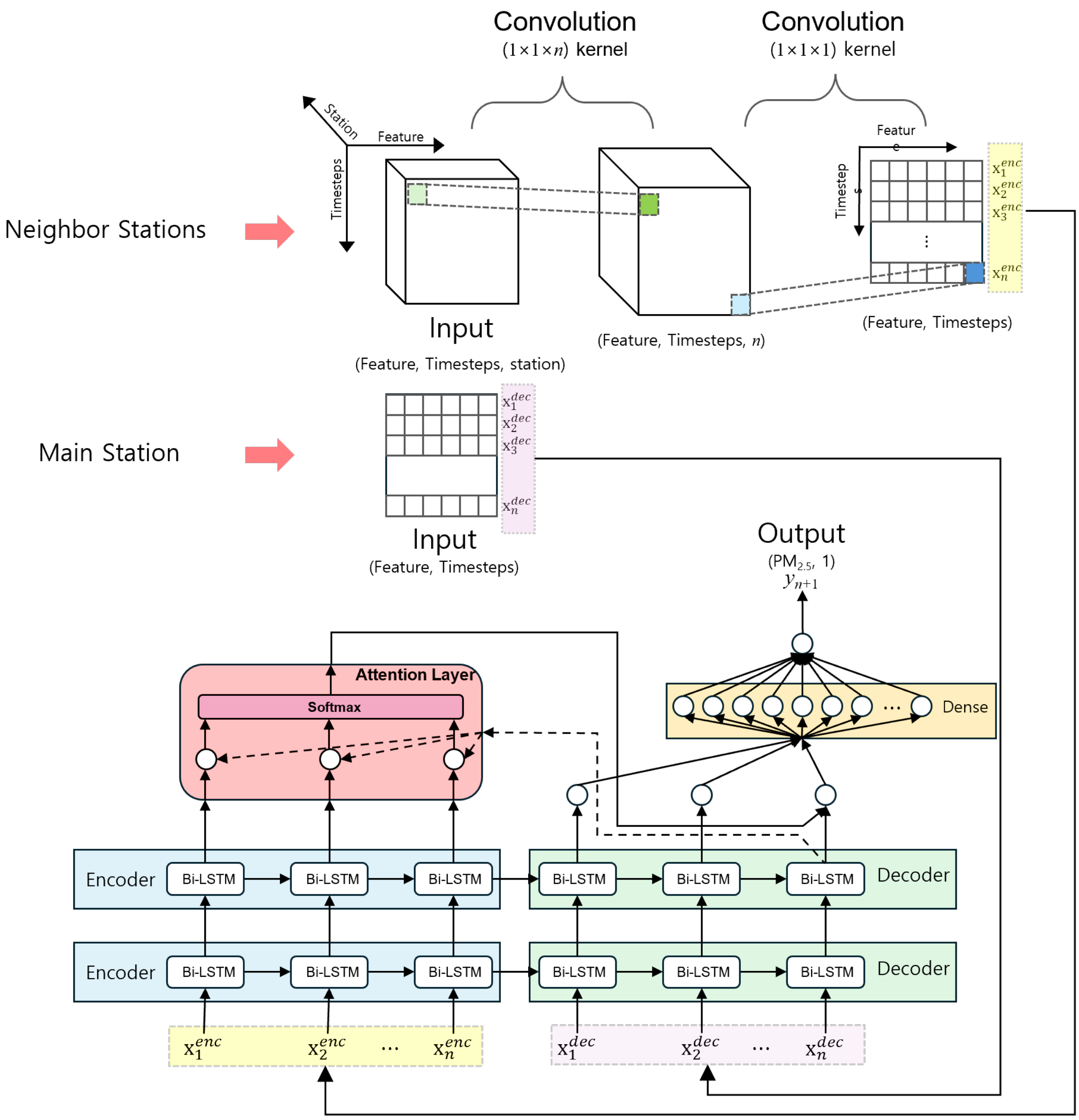

Figure 4. The proposed network architecture.

A CNN was used to integrate the data from neighboring stations selected through the mRMR method. A 1 × 1 filter was utilized in the CNN to process air pollution data commonly collected at stations. The use of 1 × 1 filters allows for the flexible adjustment of the number of data channels, which can be either increased or reduced based on the number of filters applied. This approach enables the model to be effectively deepened through the addition of layers [65–67]. In this study, two convolutional layers were used for data integration. In the first convolutional layer, the number of filters exceeded the number of selected neighboring stations, allowing for an expansion of the station data. In the second convolutional layer, the number of filters was reduced to one. This compression integrated the expanded information from the previous layer and combined the data from neighboring stations. Algorithm 2 presents the pseudocode for integrating data from neighboring stations using a CNN.

Algorithm 2: Data Integration with CNN for Selected Neighboring Stations

Input: Air pollution data from selected neighboring stations $D_{in} \in \mathbb{R}^{T \times n}$, where T is the time dimension and n is the number of selected stations.
Output: Integrated neighboring station data $D_{out} \in \mathbb{R}^{T \times 1}$.
Step 1: Apply the first 1×1 convolution layer
Set the number of filters $f_1 > n$ to expand the feature dimension.
$D_{conv1} \leftarrow \text{Conv1D}(D_{in}, \text{kernel size} = 1, \text{filters} = f_1)$;
$D_{conv1} \leftarrow \text{ReLU}(D_{conv1})$;
Step 2: Apply the second 1×1 convolution layer
Set the number of filters $f_2 = 1$ to compress and integrate the features.
$D_{conv2} \leftarrow \text{Conv1D}(D_{conv1}, \text{kernel size} = 1, \text{filters} = f_2)$;
$D_{conv2} \leftarrow \text{ReLU}(D_{conv2})$;
Step 3: Output the integrated data
$D_{out} \leftarrow D_{conv2}$;
return D_{out}

To forecast the $PM_{2.5}$ concentration at the main station using the integrated data from neighboring stations and the data from the main station, the Seq2Seq-attention model was applied. In the Seq2Seq-attention model, the proposed attention score was applied. In the conventional attention score, the dot product is used to compare the similarity between the encoder output and the decoder output. The decoder output is weighted according to the compared similarity to influence the decoder output. However, this method can emphasize similarity but fails to penalize dissimilar outputs. To overcome this limitation, this paper proposes a method containing the inverse Euclidean distance to simultaneously consider both similarity and dissimilarity. This approach emphasizes both aspects, enabling the model to maintain accurate forecasting performance while also accounting for dissimilar outputs. The pseudocode for the Seq2Seq-attention model proposed in this study is presented in Algorithm 3. In Algorithm 3, the Seq2Seq component of the Seq2Seq-attention model was implemented using Bidirectional-LSTM (Bi-LSTM).

Algorithm 3: Seq2Seq-attention Mechanism

Input: Input sequence $\mathbf{X_{enc}} = [x_1, x_2, \ldots, x_T]$, decoder input sequence $\mathbf{X^{dec}} = [x_1^{dec}, x_2^{dec}, \ldots, x_T^{dec}]$.

Output: Forecasted $PM_{2.5}$ concentration sequence $\hat{\mathbf{Y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T]$.

Step 1: Encoder Bi-LSTM

Initialize encoder hidden states $\mathbf{h^{enc}}$ and cell states $\mathbf{c^{enc}}$ for forward and backward LSTM.

for $t \leftarrow 1$ ***to*** T **do**

$\overrightarrow{\mathbf{h}_t^{enc}}, \overrightarrow{\mathbf{c}_t^{enc}} \leftarrow \overrightarrow{\mathrm{LSTM}}(x_t, \overrightarrow{\mathbf{h}_{t-1}^{enc}}, \overrightarrow{\mathbf{c}_{t-1}^{enc}});$

$\overleftarrow{\mathbf{h}_t^{enc}}, \overleftarrow{\mathbf{c}_t^{enc}} \leftarrow \overleftarrow{\mathrm{LSTM}}(x_t, \overleftarrow{\mathbf{h}_{t+1}^{enc}}, \overleftarrow{\mathbf{c}_{t+1}^{enc}});$

end

Concatenate the final hidden and cell states:

$\mathbf{z} = [\overrightarrow{\mathbf{h}_T^{enc}}, \overrightarrow{\mathbf{c}_T^{enc}}, \overleftarrow{\mathbf{h}_1^{enc}}, \overleftarrow{\mathbf{c}_1^{enc}}];$

Step 2: Decoder with Bi-LSTM and Attention

Initialize decoder hidden states $\mathbf{h^{dec}}$ and cell states $\mathbf{c^{dec}}$ using $\mathbf{z}$.

for $t \leftarrow 1$ ***to*** T **do**

Step 2.1: Attention Mechanism

Compute similarity scores for each encoder hidden state:

$e_n^t \leftarrow (\mathbf{h}_n^{enc} \cdot \mathbf{h}_t^{dec}), \quad \forall n \in [1, T];$

Compute dissimilarity scores (inverse Euclidean distance):

$d_n^t \leftarrow \frac{1}{\sqrt{\sum_{i=1}^{l} (\mathbf{h}_{n,i}^{enc} - \mathbf{h}_{t,i}^{dec})^2}}, \quad \forall n \in [1, T];$

Combine similarity and dissimilarity:

$b_n^t \leftarrow e_n^t \odot d_n^t, \quad \forall n \in [1, T];$

Normalize to compute attention weights:

$\alpha_n^t \leftarrow \frac{b_n^t}{\sum_{i=1}^{T} b_i^t}, \quad \forall n \in [1, T];$

Compute the attention vector:

$\boldsymbol{\alpha}^t \leftarrow \sum_{n=1}^{T} \alpha_n^t \cdot \mathbf{h}_n^{enc};$

Step 2.2: Decoder LSTM

Update decoder hidden state using Bi-LSTM:

$\overrightarrow{\mathbf{h}_t^{dec}}, \overrightarrow{\mathbf{c}_t^{dec}} \leftarrow \overrightarrow{\mathrm{LSTM}}(x_t^{dec}, \overrightarrow{\mathbf{h}_{t-1}^{dec}}, \overrightarrow{\mathbf{c}_{t-1}^{dec}});$

$\overleftarrow{\mathbf{h}_t^{dec}}, \overleftarrow{\mathbf{c}_t^{dec}} \leftarrow \overleftarrow{\mathrm{LSTM}}(x_t^{dec}, \overleftarrow{\mathbf{h}_{t+1}^{dec}}, \overleftarrow{\mathbf{c}_{t+1}^{dec}});$

Combine forward and backward decoder states:

$\mathbf{h}_t^{dec} \leftarrow [\overrightarrow{\mathbf{h}_t^{dec}}; \overleftarrow{\mathbf{h}_t^{dec}}];$

Step 2.3: Generate Output

Concatenate attention vector and decoder hidden state:

$\mathbf{v}_t \leftarrow [\boldsymbol{\alpha}^t; \mathbf{h}_t^{dec}];$

Integrate attention distribution and decoder hidden state through weights:

$\tilde{\mathbf{s}}_t \leftarrow tanh(\mathbf{w}_c \cdot \mathbf{v}_t);$

Compute weighted output by applying a fully connected layer:

$\hat{y}_t \leftarrow \mathbf{w}_o \cdot \tilde{\mathbf{s}}_t;$

Append output $\hat{y}_t$ to the result sequence $\hat{\mathbf{Y}}$.

end

return $\hat{\mathbf{Y}}$

---

4. Experiments and Results

In this section, we perform two case studies to validate the performance of our proposed model. The first case involved comparing the implemented attention score, which integrates both similarity and dissimilarity, with the conventional attention mechanism introduced by Bahdanau et al. [37]. The second case evaluated the forecasting performance by contrasting a standalone model that excludes neighboring station data with a model that incorporates neighboring stations selected through mRMR.

4.1. Experimental Settings

In this paper, we selected four stations (1013, 1018, 1005, 1023) to verify the proposed method. The data were divided into training and testing periods at an 8:2 ratio by month to ensure accurate model evaluation. This strategy can prevent the loss of seasonal characteristics.

Table 3 presents the statistical parameters for data of each station. Despite similar overall data values, each station exhibits unique characteristics. Station 1013 shows high variability in PM_{10} and $PM_{2.5}$ levels. Station 1018 displays distinct patterns in O_3 and SO_2 concentrations. Station 1005, with elevated PM_{10} and $PM_{2.5}$ values, represents areas with higher pollution. Lastly, Station 1023, with lower pollutant levels, reflects regions with cleaner air. These varying features make the selected stations ideal for comprehensive testing of the model.

Table 3. Statistical parameters of applied data for each station.

Station	Variable	min (Train/Test)	max (Train/Test)	mean (Train/Test)	skew (Train/Test)	kurt (Train/Test)	std (Train/Test)
1013 Train: 7032 Test: 1728	PM_{10}	5/5	1000/1000	123.69/152.33	1.78/2.23	8.75/11.48	103.68/126.47
	$PM_{2.5}$	3/3	461/446	81.24/89.15	1.52/0.94	5.27/3.91	80.04/70.88
	O_3	2/2	262/302	55.81/61.84	1.28/1.15	4.42/3.77	51.41/58.49
	SO_2	2/2	253/149	17.81/18.88	2.17/2.34	9.99/10.69	19.92/19.04
	NO_2	2/6	237/163	54.45/55.08	0.91/0.65	4.01/2.59	33.3/31.86
	CO	0.1/0.2	7.5/5.8	1.3/1.39	2/1.53	8.06/5.13	1.08/1.03
1018 Train: 7032 Test: 1728	PM_{10}	5/6.7	1000/1000	131.16/161.48	1.92/1.88	9.99/9.9	105.01/123.86
	$PM_{2.5}$	3/3	732/545	95.04/109.52	1.79/1.62	7.2/6.61	89.61/90.36
	O_3	2/2	340.5/271.4	60.8/65.98	1.22/1.01	4.34/3.26	56.4/60.52
	SO_2	2/2	298.8/154	19.48/19.9	2.45/2.6	11.46/10.28	26.82/25.03
	NO_2	3/5.8	244.5/211.2	53.42/55.34	1.03/1.14	4.47/4.01	34.22/37.95
	CO	0.1/0.2	8.5/9	1.33/1.48	2.19/2.31	8.67/9.38	1.22/1.29
1005 Train: 7032 Test: 1728	PM_{10}	5/5	1000/1000	140.57/175.36	1.9/1.82	9.78/8.52	107.56/130.83
	$PM_{2.5}$	3/3	730/582	90.78/107.99	1.76/1.8	6.95/7.32	86.23/96.13
	O_3	2/2	500/348.7	47.17/53	1.42/1.55	5.79/5.68	46.12/55.84
	SO_2	2/2	295.4/146.9	16.75/14.96	4.82/2.53	42.71/11.48	26.12/17.91
	NO_2	2/2	219.4/210.8	56.59/60.2	0.87/1.02	3.81/4.13	33.89/35.58
	CO	0.1/0.1	10.5/12.5	1.42/1.59	2.22/2.54	9.03/10.99	1.33/1.55
1023 Train: 7032 Test: 1728	PM_{10}	5/5	1000/1000	109.74/138.65	2.8/2.49	19.82/16.76	94.62/105.17
	$PM_{2.5}$	3/3	641/331	77.08/83.64	1.75/0.82	7.54/3.4	74.47/64.39
	O_3	2/2	330/328.1	58.92/64.56	1.48/1.44	5.22/4.8	57.13/64.81
	SO_2	2/2	161/102	12.33/13.21	2.73/1.98	12.4/6.98	17.3/15.6
	NO_2	2/3.1	193.9/204	43.67/44.34	0.96/1.01	3.7/4.36	29.87/29.5
	CO	0.1/0.1	8.9/8	1.09/1.19	2.2/1.94	10/9.14	0.98/0.97

The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R^2, relative Root Mean Squared Error (rRMSE), and Percent Bias (PBIAS) were used to evaluate the performance of the proposed method in forecasting [68–72]. These performance indices evaluate the accuracy and reliability of the model forecast. RMSE means the average magnitude of the error between the forecasted and actual values. A lower RMSE indicates better model performance. MAE calculates the average of the absolute errors, providing an understanding of the overall magnitude of the forecast error. R^2 represents the proportion of the variance in the actual values that the model successfully captures in its forecast. A higher R^2 value suggests a better fit of the model. rRMSE normalizes the RMSE by the range or mean of the observed data, making it interpretable across different datasets. PBIAS calculates the overall bias in the forecasts. A negative PBIAS indicates under-forecasting, while a positive value suggests over-forecasting. In Equations (2)–(6), y_t and $\hat{y}_t$ represent the actual value and the forecasted value at time t. $\bar{y}$ represents the average of the actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}. \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|y_t - \hat{y}_t|). \tag{3}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}. \tag{4}$$

$$rRMSE = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\frac{1}{n}\sum_{t=1}^{n}y_t}. \tag{5}$$

$$PBIAS = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)}{\sum_{t=1}^{n}y_t}. \tag{6}$$

4.2. Case Study 1: Compare Forecasting Result Based on Conventional and Proposed Attention Scores

In Case Study 1, we compared the results of forecasting $PM_{2.5}$ concentrations according to attention scores. Figures 5 and 6 show the $PM_{2.5}$ concentration forecasting results for stations 1013, 1018, 1005, and 1023. The x-axis represents time, and the y-axis represents $PM_{2.5}$ concentration. The black line shows the actual observed values. The blue dotted line represents the conventional attention score forecast. The red dashed line represents the proposed attention score forecast. Figures 5 and 6 (a,d) present the forecasting results for the test data period. Subfigures (b,e) show enlarged views of periods with average $PM_{2.5}$ concentrations below 100, where frequent changes occur. Similarly, (c,f) focus on periods with $PM_{2.5}$ concentrations above 100, which also exhibit frequent changes.

Both methods produced forecast errors at Station 1013. In Figure 5b, the proposed method shows lower error than the conventional method due to over-forecasting. Specifically, the proposed method exhibits lower error during the increasing periods (40–68). In Figure 5c, both methods show errors caused by over-forecasting.

At Station 1018, the proposed method mainly under-forecasts when the $PM_{2.5}$ concentration is less than 100. This leads to many errors. Conversely, the conventional method over-forecasts and causes errors, as shown in Figure 5. In the 1140–1170 periods, the error from the proposed method is lower than the error from the conventional method. In Figure 5, where the $PM_{2.5}$ concentration exceeds 100, both methods cause over-forecasting errors. However, the proposed method results in lower errors.

In the case of $PM_{2.5}$ at Station 1005, Figure 6a shows that both the proposed method and the conventional method produced errors mainly due to over-forecasting. Figure 6b confirms that the conventional method produced a larger error due to over-forecasting compared to the proposed method. During the period 48–68, where the $PM_{2.5}$ concentration gradually increases, the conventional method exhibited a more significant over-forecasting error than the proposed method. In Figure 6c, both methods caused errors due to under-forecasting. However, in the 1050–1060 period, where the $PM_{2.5}$ concentration exceeded 400, the proposed method produced a smaller under-forecasting error than the conventional method.

In the case of Station 1023, the error mainly occurs due to under-forecasting, unlike the other stations. In the periods 230–246 of Figure 6e, the proposed method shows a lower error due to under-forecasting compared to the conventional method. In Figure 6f, the error caused by the proposed method's under-forecasting is also lower than that of the conventional method.

Table 4 shows the forecasting results for Case Study 1 as performance indices. Bold in the table indicates better performance. The PBIAS does not indicate the magnitude based on its sign; therefore, the average of PBIAS is calculated by converting each value to its absolute value and then computing the mean of those absolute values. The RMSE decreased by about 2.01% from 17.2571 to 16.9103 at Station 1013 and approximately 4.25% from 22.4380 to 21.4842 at Station 1018. It decreased by 3.22% from 22.8176 to 22.0825 at Station 1005 and 4.30% from 17.0031 to 16.2714 at Station 1023. The forecast error decreased as the average of all stations decreased by about 3.48% from 19.8789 to 19.1871 based on RMSE. The improved performance can also be seen in the MAE. It decreased by approximately 3.09% from 10.5938 to 10.2671 at Station 1013 and improved by 12.51% from 14.0076 to 12.2573 at Station 1018. It decreased by 9.16% from 14.2673 to 12.9601 at Station 1005 and 8.25% from 10.9377 to 10.0363 at Station 1023. On average, it decreased by 8.60% from 12.4516 to 11.3802, resulting in a lower overall forecast error. The R^2 shows an improvement in explanatory power with a slight increase in value at each station. It increased by 0.27% from 0.9376 to 0.9401 at Station 1013 and by 0.60% from 0.9324 to 0.9380 at Station 1018. In addition, it increased by 0.44% from 0.9350 to 0.9391 at Station 1005 and by 0.66% from 0.9275 to 0.9336 at Station 1023, confirming that the model better explains variations in the data. The R^2 indicator increased by an average of 0.49% from 0.9331 to 0.9377, confirming that the model's explanatory power for explaining the data variances has improved. The rRMSE decreased at some stations, indicating improved performance. At Station 1013, it decreased by approximately 2.04% from 0.2059 to 0.2017; at Station 1023, it decreased by approximately 4.26% from 0.2251 to 0.2155. In addition, at Stations 1018 and 1005, rRMSE decreased by 0.0075 and 0.0092, respectively, showing about 3.24% and 4.30% decreases. On average, it decreased by approximately 3.64% from 0.2192 of the conventional model to 0.2115 of the proposed model, confirming that the forecast performance improved overall. The PBIAS index showed varying results across stations, indicating differences in forecasting bias. At Station 1013, the PBIAS improved from −6.0629 to 1.3573, reflecting a shift from over-forecasting to under-forecasting. At Station 1018, PBIAS improved from −4.0316 to 2.6637, showing a similar change. At Station 1005, the conventional method had a PBIAS of −0.7858, which improved to −0.0378 with the proposed method, indicating a reduction in bias. At Station 1023, the conventional method's PBIAS of 4.2174 decreased to 2.1168 with the proposed method, showing a reduction in under-forecasting. On average, PBIAS decreased from 3.7744 for the conventional method to 1.5364 for the proposed method, indicating an overall improvement in forecast accuracy and a reduction in forecast bias.

Table 4. Performance Indices of $PM_{2.5}$ Concentration Forecasting for Case Study 1.

Station ID	RMSE		MAE		R^2		rRMSE		PBIAS	
	Conv.	Prop.	Conv.	Prop.	Conv.	Prop.	Conv.	Prop.	Conv.	Prop.
1013	17.2571	**16.9103**	10.5938	**10.2671**	0.9376	**0.9401**	0.2059	**0.2017**	−6.0629	**1.3573**
1018	22.4380	**21.4842**	14.0076	**12.2573**	0.9324	**0.9380**	0.2251	**0.2155**	−4.0316	**2.6637**
1005	22.8176	**22.0825**	14.2673	**12.9601**	0.9350	**0.9391**	0.2318	**0.2243**	−0.7858	**−0.0378**
1023	17.0031	**16.2714**	10.9377	**10.0363**	0.9275	**0.9336**	0.2138	**0.2046**	4.2174	**2.1168**
Average	19.8789	**19.1871**	12.4516	**11.3802**	0.9331	**0.9377**	0.2192	**0.2115**	3.7744	**1.5364**

Conv.: conventional method, Prop.: proposed method, Bold values indicate better performance.

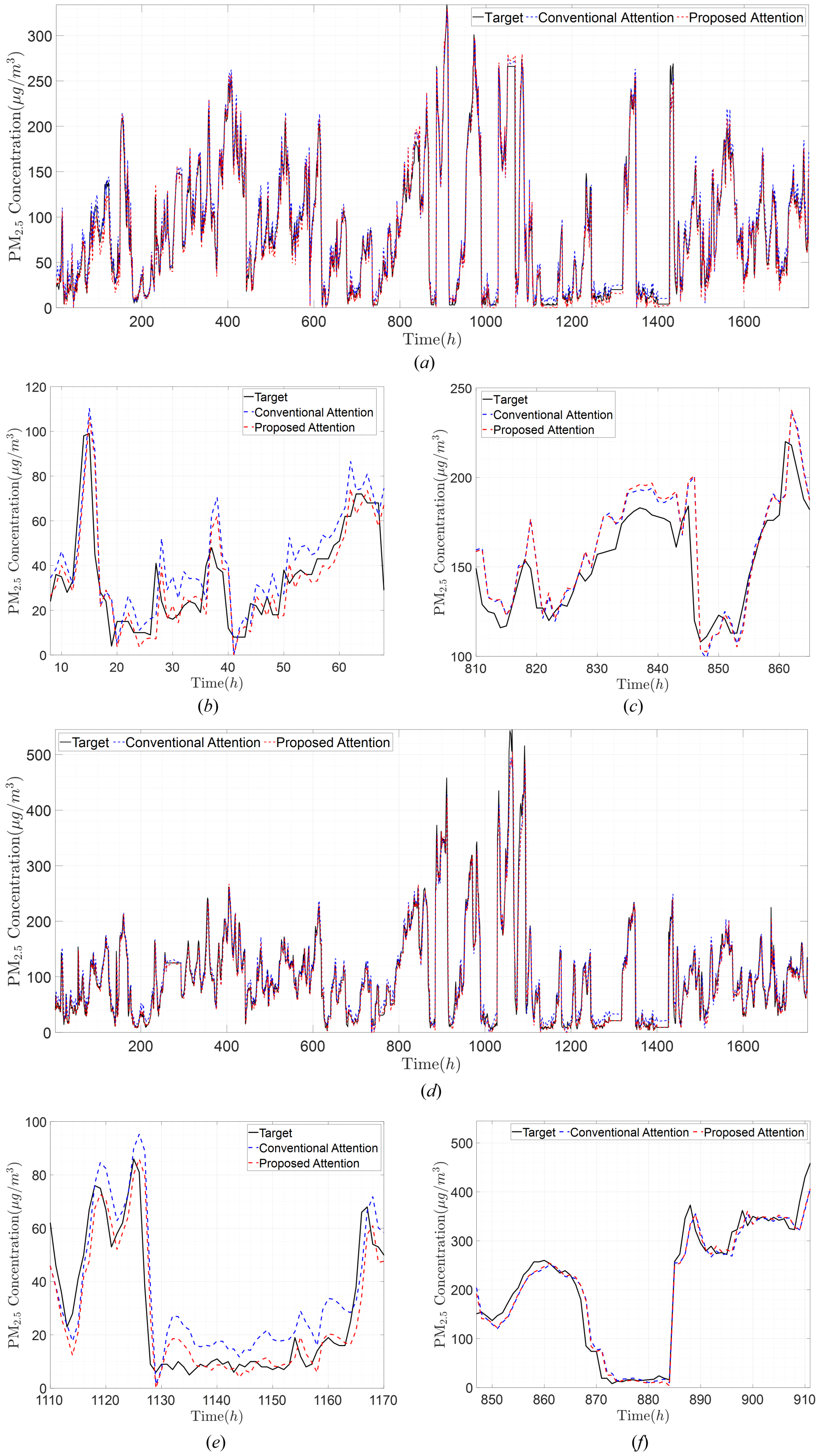

Figure 5. Results of $PM_{2.5}$ concentration forecast for Station 1013: (**a**) Forecasting results for the test data period. (**b**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**c**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100. Result of $PM_{2.5}$ concentration forecast for Station 1018: (**d**) Forecasting results for the test data period. (**e**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**f**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100.

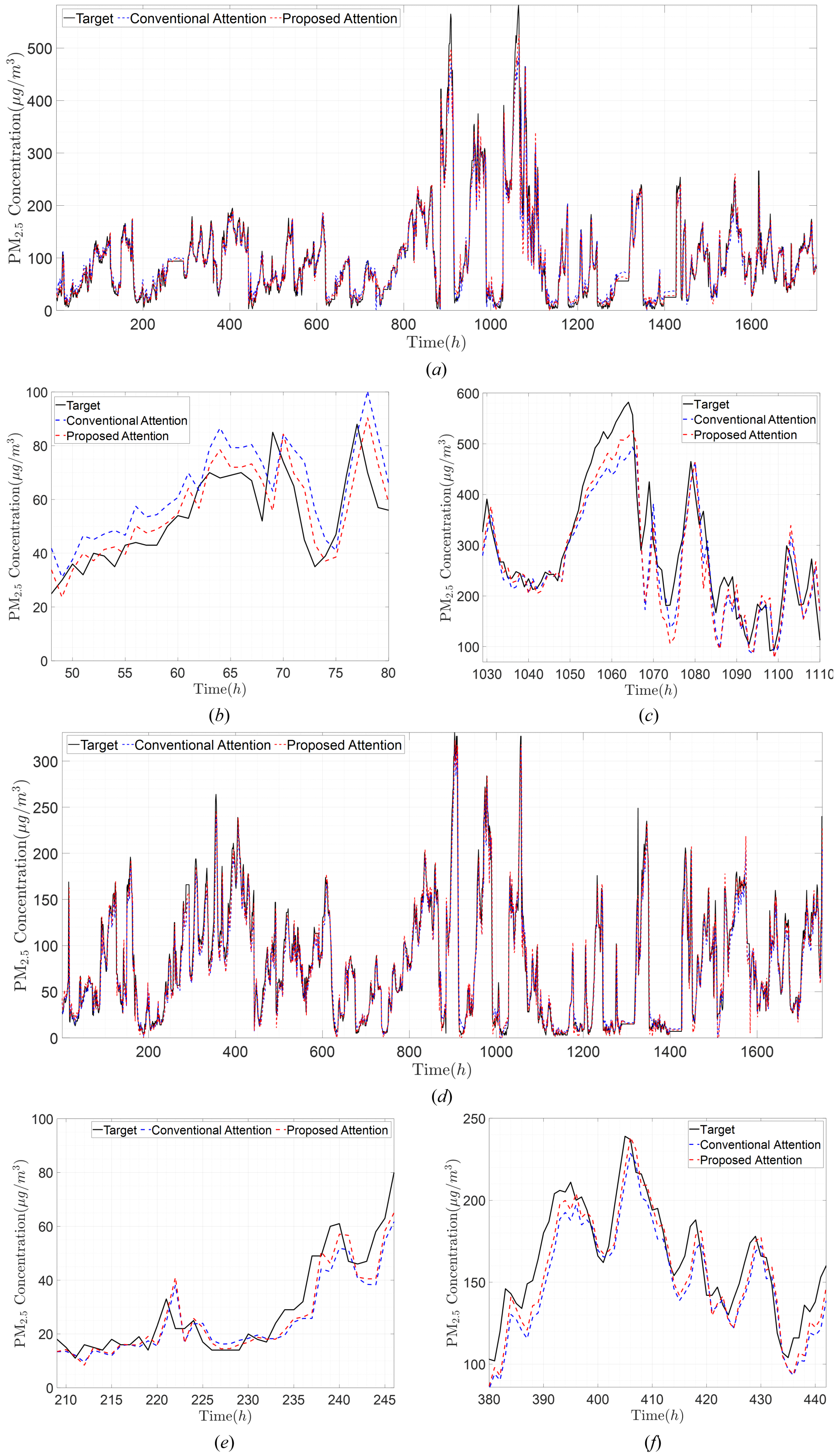

Figure 6. Result of $PM_{2.5}$ concentration forecast for Station 1005: (**a**) Forecasting results for the test data period. (**b**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**c**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100. Result of $PM_{2.5}$ concentration forecast for Station 1023: (**d**) Forecasting results for the test data period. (**e**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**f**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100.

4.3. Case Study 2: Comparison of Forecasting Using Neighboring Stations vs. Standalone Station

In Case Study 2, we experimented with comparing the performance of a model using a standalone station with a model that considers neighboring stations. The model used in Case Study 2 has the same structure as the model implemented in Case Study 1, utilizing the proposed attention score. Table 5 presents the RMSE values based on the mRMR score for considering neighboring stations. The first row of the table lists the station ID, with bold text highlighting the selected mRMR score, the number of neighboring stations, and the RMSE for each station. Except for Station 1023, fewer than 10 neighboring stations were selected.

Table 5. RMSE by mRMR score by station.

1013			1018			1005			1023		
mRMR Score	**Number of Station**	**RMSE**	**mRMR Score**	**Number of Station**	**RMSE**	**mRMR Score**	**Number of Station**	**RMSE**	**mRMR Score**	**Number of Station**	**RMSE**
0.96	1	16.5196	0.94	1	20.8046	0.93	1	22.9925	0.89	1	17.2837
0.13	2	16.3046	0.07	2	20.0653	**0.07**	**2**	**21.7751**	0.08	2	15.7653
0.12	3	16.5344	**0.06**	**3**	**19.8194**	0.06	3	24.0116	0.07	3	17.1782
0.09	4	16.3244	0.05	6	20.9996	0.05	4	24.1725	0.06	4	16.9708
0.07	5	16.3560	0.04	8	20.1663	0.04	7	24.2891	0.04	7	15.6276
0.06	**6**	**16.1756**	0.03	12	20.5897	0.03	15	25.2465	0.03	9	16.7848
0.05	7	16.5681	0.02	17	20.8992	0.02	19	25.2250	0.02	10	17.0142
0.04	10	17.2918	0.01	19	22.1383	0.01	23	25.2691	0.01	11	16.4986
0.03	13	17.2950	0	23	23.9987	0	25	25.4812	**0**	**15**	**15.3535**
0.02	18	16.3739	−0.01	25	24.0091	−0.01	29	27.3180	−0.01	17	16.3688
0.01	27	17.7325	−0.02	26	25.0417	−0.02	30	27.4363	−0.02	22	17.2486
0	31	18.5435	−0.03	29	25.0799	−0.03	31	27.2852	−0.03	28	19.7211
−0.01	34	17.4558	−0.04	30	24.5889	−0.04	33	27.6138	−0.04	33	18.3389
			−0.05	33	25.3252	−0.06	34	24.8485	−0.06	34	18.3402
			−0.06	34	27.0870						

Bold values indicate better performance.

Figures 7 and 8 show the results of forecasting $PM_{2.5}$ concentration by station. In Figures 7 and 8 (a,d), the results of forecasting $PM_{2.5}$ concentration for the test data period are shown. In (b,e), the focus is on the periods where the average $PM_{2.5}$ concentration is below 100 and changes frequently, while in (c,f), the focus is on the periods where the $PM_{2.5}$ concentration is above 100 and changes frequently.

Figure 7a–c shows the forecasting results for the $PM_{2.5}$ concentration at Station 1013. In Figure 7b, it can be seen that both the standalone method and the method considering neighboring stations have errors due to over-forecasting. In periods 36–40, both methods are over-forecasting, but the error is smaller when neighboring stations are considered. When the $PM_{2.5}$ concentration decreases to approximately 40, the forecast is more accurate when neighboring stations are considered. As shown in Figure 7c, the over-forecasting error is smaller when neighboring stations are considered, particularly during the periods 830–848, where the error is reduced.

The forecast results of $PM_{2.5}$ concentration at Station 1018 are shown in Figure 7d–f. In Figure 7e, during the periods from 1116 to 1130, the forecast method using neighboring stations has a lower error than the method using a standalone station. In Figure 7f, the method using a standalone station mainly under-forecasts, while the method using neighboring stations tends to over-forecast. Specifically, for the periods between 847 and 870, the forecast considering neighboring stations has a lower error. During the periods from 885 to 911, both methods generally under-forecast, but the method considering neighboring stations produces a smaller error.

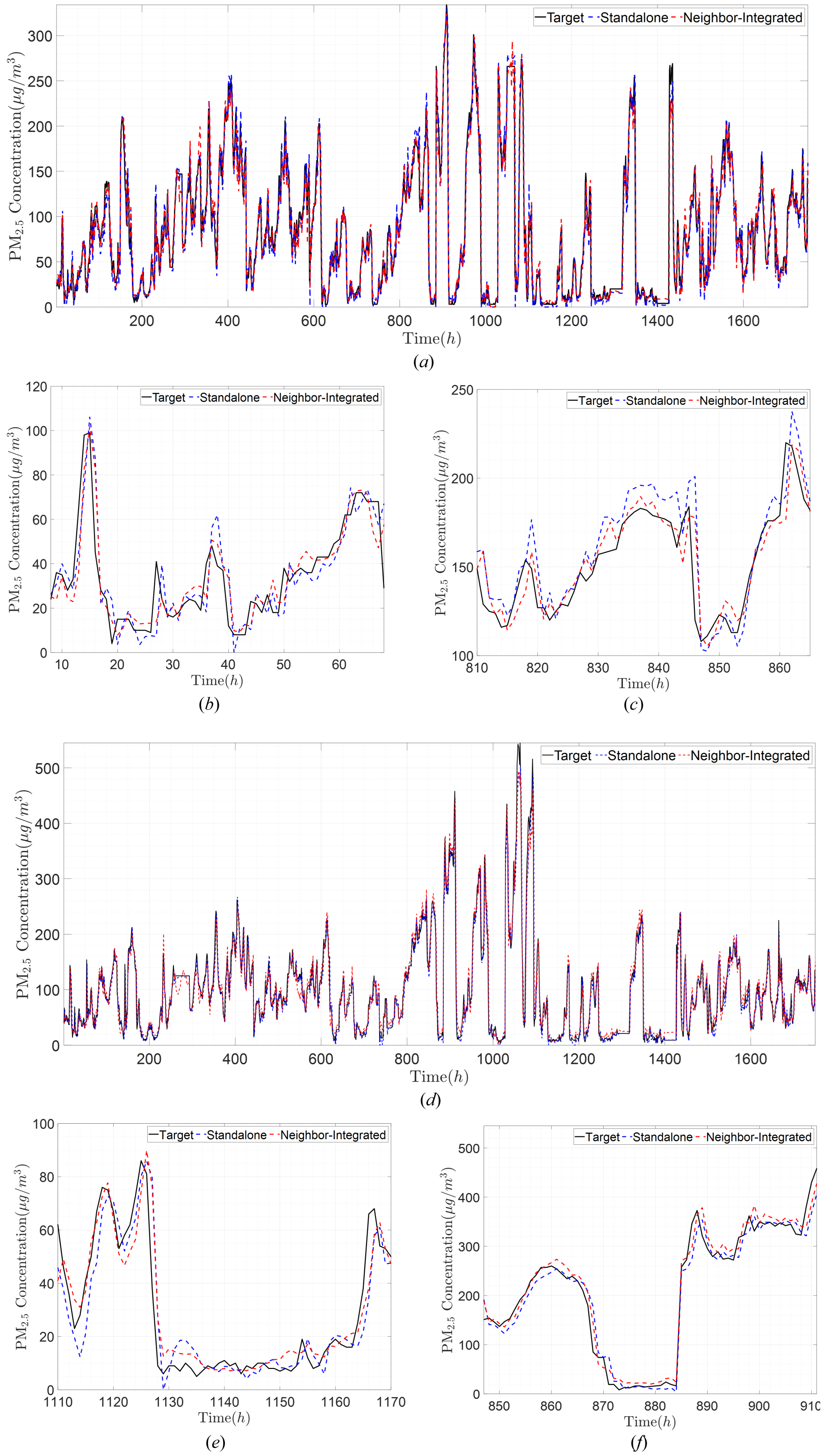

Figure 7. Result of $PM_{2.5}$ concentration forecasting for Station 1013: (**a**) Forecasting results for the test data period. (**b**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**c**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100. Result of $PM_{2.5}$ concentration forecasting for Station 1018: (**d**) Forecasting results for the test data period. (**e**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**f**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100.

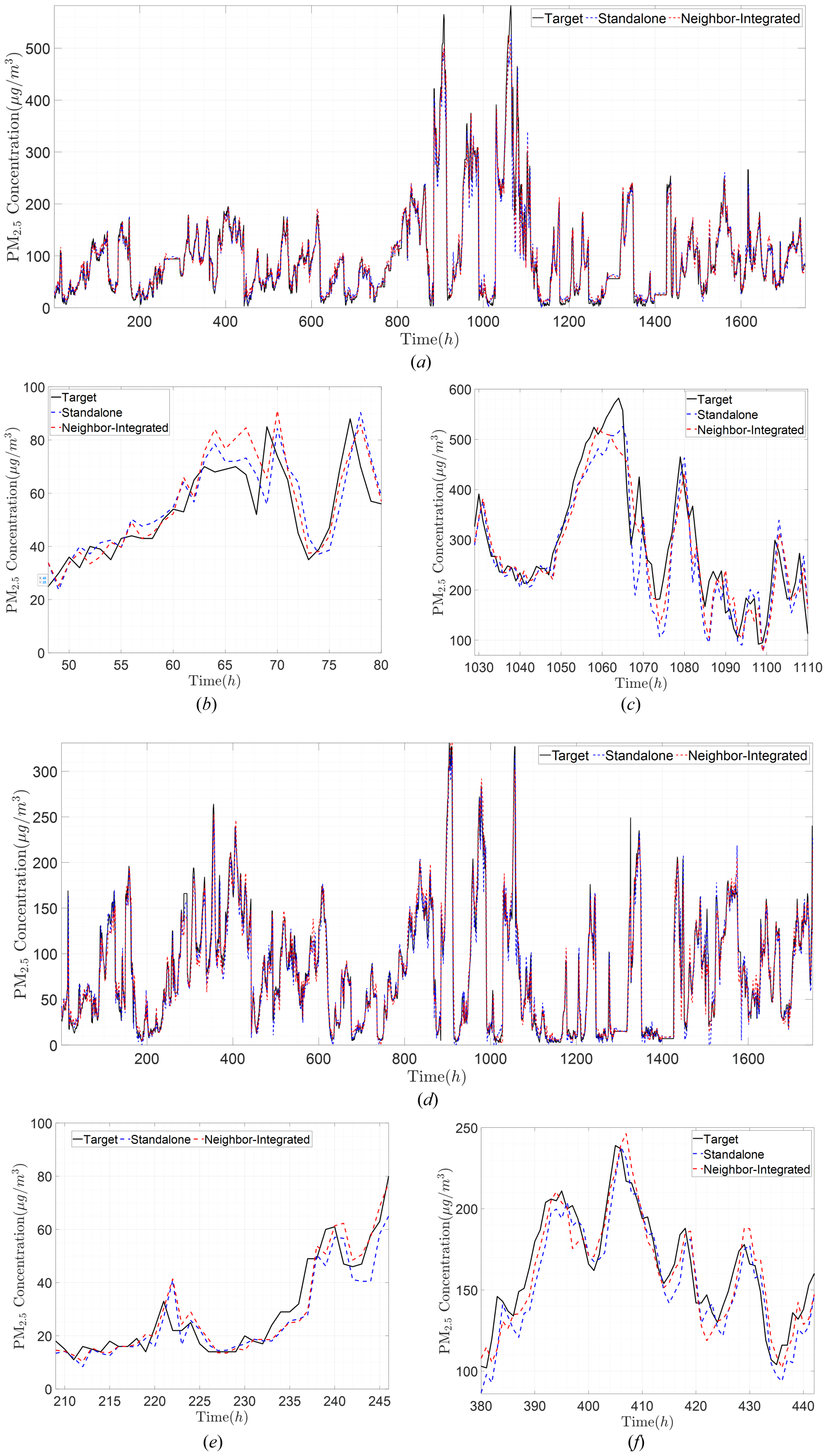

Figure 8. Result of $PM_{2.5}$ concentration forecasting for Station 1005: (**a**) Forecasting results for the test data period. (**b**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**c**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100. Result of $PM_{2.5}$ concentration forecasting for Station 1023: (**d**) Forecasting results for the test data period. (**e**) Forecasting results for periods with $PM_{2.5}$ concentrations of 100 or less. (**f**) Forecasting results for periods with $PM_{2.5}$ concentrations exceeding 100.

In Figure 8, when the $PM_{2.5}$ concentration was forecasted by considering the neighbors of Station 1005, the error due to over-forecasting was lower than the result forecasted by considering a standalone station. In particular, in Figure 8b, in the 48–68 periods where the $PM_{2.5}$ concentration gradually increases from 20 to 75, the forecasting performance improved when the neighboring stations were considered. In Figure 8c, both methods incur errors due to under-forecasting. When considering the neighboring stations in the periods 1050–1065, the $PM_{2.5}$ concentration forecasting performance is superior.

As Figure 8d shows, the method considering neighboring stations at Station 1023 and the method considering a standalone station have errors due to under-forecasting. In Figure 8e, both methods are under-forecasting in all periods except for the 220–225 periods. The method considering neighboring stations forecasts higher $PM_{2.5}$ concentrations than a standalone station. Therefore, the under-forecasting error is lower than the forecast considering a standalone station. In Figure 8f, both methods under-forecast and the forecasting error considering neighboring stations is lower than that of the forecast considering a standalone station.

Table 6 shows the $PM_{2.5}$ concentration forecast results of Case Study 2 as a performance indicator. The values in bold indicate cases where the performance exceeds that of the other models.

Table 6. Performance Indices of $PM_{2.5}$ Concentration Forecasting for Case Study 2.

Station ID	RMSE		MAE		R^2		rRMSE		PBIAS	
	Standalone	Neighbor	Standalone	Neighbor	Standalone	Neighbor	Standalone	Neighbor	Standalone	Neighbor
1013	16.9103	**16.1756**	10.2671	**9.8101**	0.9401	**0.9452**	0.2017	**0.1930**	1.3573	**0.3105**
1018	21.4842	**19.6820**	**12.2573**	12.7890	0.9380	**0.9480**	0.2155	**0.1974**	2.6337	**−0.5673**
1005	22.0825	**21.3221**	12.9601	**12.8822**	0.9391	**0.9433**	0.2243	**0.2166**	**−0.0378**	−1.9433
1023	16.2714	**15.3535**	10.0363	**9.8057**	0.9336	**0.9409**	0.2046	**0.1931**	2.1168	**0.4267**
Average	19.1871	**18.1333**	11.3802	**11.3217**	0.9377	**0.9444**	0.2115	**0.2000**	1.5364	**−0.8120**

Bold values indicate better performance.

Regarding the RMSE, the forecast error decreased at all stations. The RMSE of Station 1013 decreased by approximately 4.34% from 16.9103 to 16.1756 when forecasting considered the neighboring stations, compared to when forecasting considered a standalone station, and that of Station 1018 decreased by approximately 8.39% from 21.4842 to 19.6820. In addition, that of Station 1005 decreased by approximately 3.44% from 22.0825 to 21.3221, and that of Station 1023 decreased by approximately 5.64% from 16.2714 to 15.3535. The average RMSE of all stations decreased by approximately 5.49% from 19.1871 to 18.1333 when considering the neighboring stations, confirming that the performance improved overall. There are also some improvements in the MAE. The MAE of Station 1013 decreased by approximately 4.45% from 10.2671 to 9.8101, and there were also slight decreases in MAE at Stations 1005 and 1023. However, the MAE at Station 1018 increased slightly from 12.2573 to 12.7890; overall, the average MAE decreased slightly from 11.3802 to 11.3217, showing a decrease in forecasting error. The R^2 shows that the overall explanatory power of the model is improved by considering neighboring stations. Station 1018 showed the most considerable improvement, increasing from 0.9380 to 0.9480, and Station 1023 also increased from 0.9336 to 0.9409. The average R^2 of all stations increased by approximately 0.67% from 0.9377 to 0.9444, indicating that the model better explains data variation. The rRMSE also decreased at most stations, indicating an improved forecasting performance. It decreased from 0.2017 to 0.1930 at Station 1013, from 0.2155 to 0.1974 at Station 1018, and from 0.2046 to 0.1931 at Station 1023. The overall average decreased by approximately 5.44% from 0.2115 for the conventional model to 0.2000 for the proposed model, which can be evaluated as an improvement in model performance. Finally, The PBIAS exhibited variations across stations, reflecting differences in forecasting bias. At Station 1013, the PBIAS improved significantly from 1.3573 to 0.3105, indicating a notable reduction in over-forecasting bias. At Station 1018, the PBIAS shifted from 2.6337 to −0.5673, demonstrating a transition from over-forecasting to under-forecasting with reduced bias. At Station 1005, the standalone station had a PBIAS of −1.0378, which changed to −1.9433 with the neighbor stations, suggesting an increase in under-forecasting bias. Similarly, at Station 1023, the PBIAS decreased from 2.1168 to 0.4267, highlighting a reduction in forecasting bias. On average, the PBIAS improved from 1.5364 with the conventional method to −1.8120 with the proposed method, confirming a reduction in overall forecast bias and improved accuracy.

Multiple visualization techniques were employed to comprehensively analyze the experimental results of Case Studies 1–2. Figure 9 shows the results for Cases Studies 1–2 as boxplots. Tables 7 and 8 present the boxplot parameters for $PM_{2.5}$ concentration forecasting results from each station. In Figure 9 and Tables 7 and 8, "Conv." refers to the conventional attention method model used in Case Study 1. "Standalone" represents the standalone stations model employing the proposed attention method in Case Study 2. "Neighbor" denotes the neighbor-integrated stations model utilizing the proposed attention method in Case Study 2. The tables include the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values, comparing the target values with the forecasting results made by the models (Conv., Standalone, and Neighbor). For Station 1013, the Neighbor model provided the closest forecasting results to the target values, showing similar results in Q1 and Q2. The Conv. and Standalone models tended to over-forecast, with the minimum value for the Neighbor model being the only one below the target, indicating some under-forecasting. At Station 1018, the Neighbor model again yielded the closest forecasting results, with the Conv. model showing slightly higher values than the target. In Q1 and Q2, the Neighbor model's forecasting results were closer to the target, while the other models showed higher values. For Station 1005, the Neighbor model also performed better, with forecasting results close to the target in Q1 and Q2. Both Conv. and Standalone models had a wider range of forecasting results, with the maximum values showing significant discrepancies from the target. At Station 1023, the range of forecasting results was narrower compared to the other stations, and the Neighbor model once again provided forecasting results closest to the target. In Q1, Q2, and Q3, all models showed similar results to the target, but the Conv. model tended to slightly underestimate the values. In conclusion, across all stations, the Neighbor model generally provided forecasting results closer to the target values compared to the Conv. and Standalone models. The Neighbor model showed more consistent and accurate results in Q1, Q2, and Q3, indicating that incorporating neighboring station data improves forecasting accuracy. Meanwhile, the Conv. and Standalone models exhibited larger variability in their forecasting results, often over-forecasting or under-forecasting the target values.

Table 7. Boxplot Parameters of $PM_{2.5}$ Concentration Forecasting for Stations 1013 and 1018.

Station	**1013**				**1018**			
Statistics	**Target**	**Conv.**	**Standalone**	**Neighbor**	**Target**	**Conv.**	**Standalone**	**Neighbor**
Minimum	3	0	0	−0.35	3	0	0	3.55
Q1	125	131.86	123.6	126.18	132	133.09	127.51	132.47
Q2 (Median)	72	76.19	67.09	70.21	84	87.44	79.92	82.47
Q3	22	26.26	19.85	22.27	33	40.82	32.97	35.89
Maximum	334	326.96	329.77	324.21	545	497.86	505.58	494.02

Table 8. Boxplot Parameters of $PM_{2.5}$ Concentration Forecasting for Stations 1005 and 1023.

Station		**1005**				**1023**		
Statistics	**Target**	**Conv.**	**Standalone**	**Neighbor**	**Target**	**Conv.**	**Standalone**	**Neighbor**
Minimum	3	0	0	9.62	3	0	0	0
Q1	129	126.64	126.02	130.6	123	115.18	120.18	120.67
Q2 (Median)	78	81.66	78.93	79.77	66	64.25	65.94	67.29
Q3	35	41.92	39.87	38.88	23	25.22	23.65	23.72
Maximum	582	493.93	525.78	524.64	331	315.85	322.48	354.19

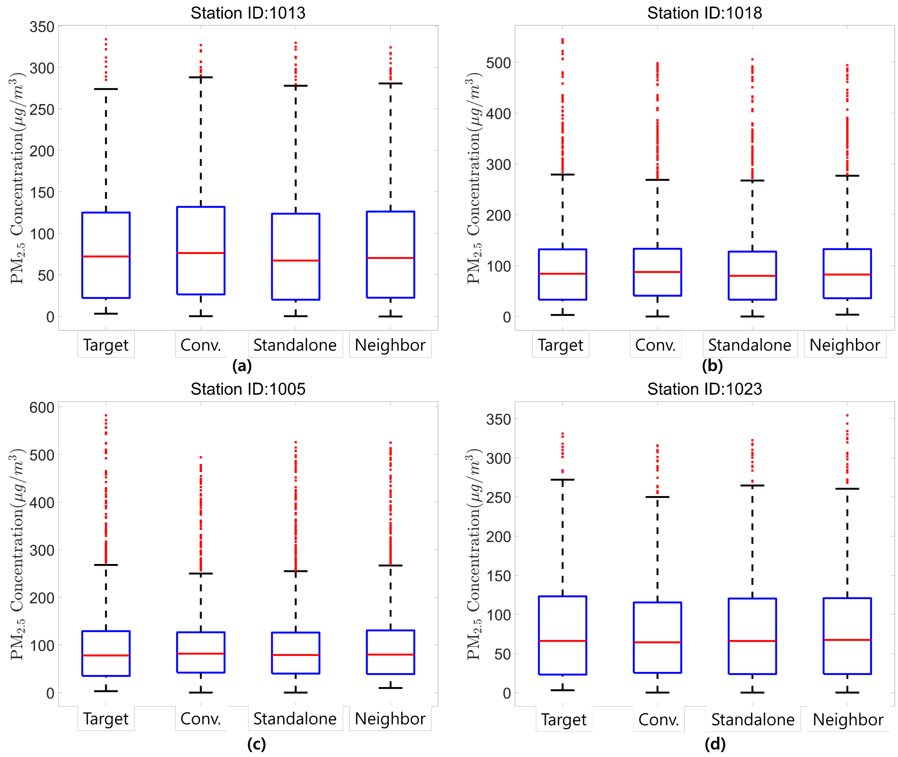

Figure 9. Boxplot of $PM_{2.5}$ concentration forecasting result for Station: (**a**) 1013. (**b**) 1018. (**c**) 1005. (**d**) 1023.

Figure 10 shows the results for Case Studies 1–2 as Taylor diagrams. The Tables 9 and 10 present the Taylor diagram parameters for $PM_{2.5}$ concentration forecasting results from each station. Figure 10 and Tables 9 and 10 present that the Neighbor model provides the most accurate $PM_{2.5}$ concentration forecasts. Its standard deviation (STD) and root mean square difference (RMSD) are closest to the target values across all stations. All models achieved high correlation coefficients of 0.97, showing a strong relationship with the target data. However, the Neighbor model consistently minimized forecast errors more effectively. At Station 1013, the Neighbor model's STD was 68.02, closely matching the target of 69.07. Its RMSD was the lowest at 16.17. In comparison, the Conv. and Standalone models had STDs of 69.65 and 71.11 and RMSDs of 16.45 and 16.82. At Station 1018, the Neighbor model showed an STD of 83.67 and an RMSD of 19.67, outperforming the Conv. model (STD: 82, RMSD: 22.07) and the Standalone model (STD: 83.97, RMSD: 21.31). For Station 1005, the Neighbor model showed the closest STD to the target, with a value of 85.27 compared to the target of 89.51. It also had the lowest RMSD of 21.24. The Conv. and Standalone models showed larger deviations, with STDs of 79.92 and 83.14 and RMSDs of 22.80 and 22.08. At Station 1023, the Neighbor model recorded an STD of 62.75 and an RMSD of 15.35, which are better than the Conv. model (STD: 58.29, RMSD: 16.66) and the Standalone model (STD: 61.17, RMSD: 16.12). The Neighbor model is close to the target values, reducing forecast errors and variability.

It outperforms the Conv. and Standalone models. This demonstrates that incorporating neighboring station data improves forecasting accuracy for $PM_{2.5}$ concentrations.

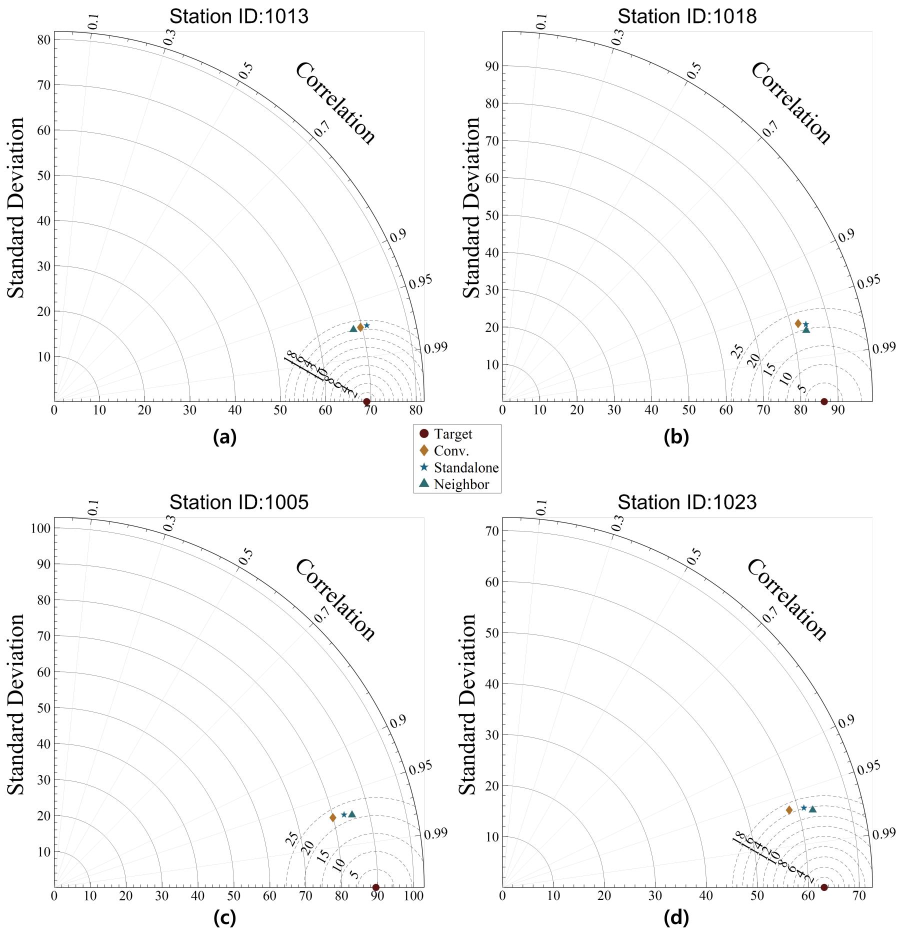

Figure 10. Taylor Diagram of $PM_{2.5}$ Concentration Forecasting Result for Station: (**a**) 1013. (**b**) 1018. (**c**) 1005. (**d**) 1023.

Table 9. Taylor Diagram Parameters of $PM_{2.5}$ Concentration Forecasting for Stations 1013 and 1018.

Station		**1013**				**1018**		
Statistics	**Target**	**Conv.**	**Standalone**	**Neighbor**	**Target**	**Conv.**	**Standalone**	**Neighbor**
STD	69.07	69.65	71.11	68.02	86.29	82	83.97	83.67
RMSD	0	16.45	16.82	16.17	0	22.07	21.31	19.67
Correlation	1	0.97	0.97	0.97	1	0.97	0.97	0.97

Table 10. Taylor Diagram Parameters of $PM_{2.5}$ Concentration Forecasting for Stations 1005 and 1023.

Station		**1013**				**1018**		
Statistics	**Target**	**Conv.**	**Standalone**	**Neighbor**	**Target**	**Conv.**	**Standalone**	**Neighbor**
STD	89.51	79.92	83.14	85.27	63.15	58.29	61.17	62.75
RMSD	0	22.80	22.08	21.24	0	16.66	16.12	15.35
Correlation	1	0.97	0.97	0.97	1	0.97	0.97	0.97

Figure 11 shows the results for Case Studies 1–2 as radar charts. RMSE, MAE, and rRMSE are performance indices representing the magnitude of forecasting errors, with smaller values indicating better model performance. On the other hand, the better the performance of the model matches the actual data, the higher R^2 increases. Additionally, PBIAS values closer to zero indicate smaller discrepancies between the forecasted and actual values. Because of these differences, using all these performance indices simultaneously in a radar chart could lead to interpretational confusion. Therefore, only RMSE, MAE, and rRMSE are used in the radar chart to provide an intuitive and consistent performance evaluation. The results from the models applied to each station, as reflected in the radar chart, demonstrate that the Neighbor model consistently outperforms the Conv. and Standalone models in terms of RMSE and rRMSE across all stations. At Station 1013, the Neighbor model showed the lowest RMSE (16.1756), MAE (9.8101), and rRMSE (0.1930), indicating that it provides the most accurate forecasting with the smallest errors. A similar trend was observed at Station 1018, where the Neighbor model outperformed the Conv. model in RMSE (19.682) and rRMSE (0.1974), but the Standalone model showed the best MAE (12.2573). At Station 1005, the Neighbor model had the lowest RMSE (21.3221) and rRMSE (0.2166), while the Standalone model showed the best MAE (12.9601), demonstrating the variability of the models' performance across different metrics. Finally, at Station 1023, the Neighbor model exhibited the best performance with the lowest RMSE (15.3535), MAE (9.8057), and rRMSE (0.1931). These results consistently show that incorporating neighboring station data leads to more accurate and reliable $PM_{2.5}$ concentration forecasting across different stations, especially in terms of RMSE and rRMSE.

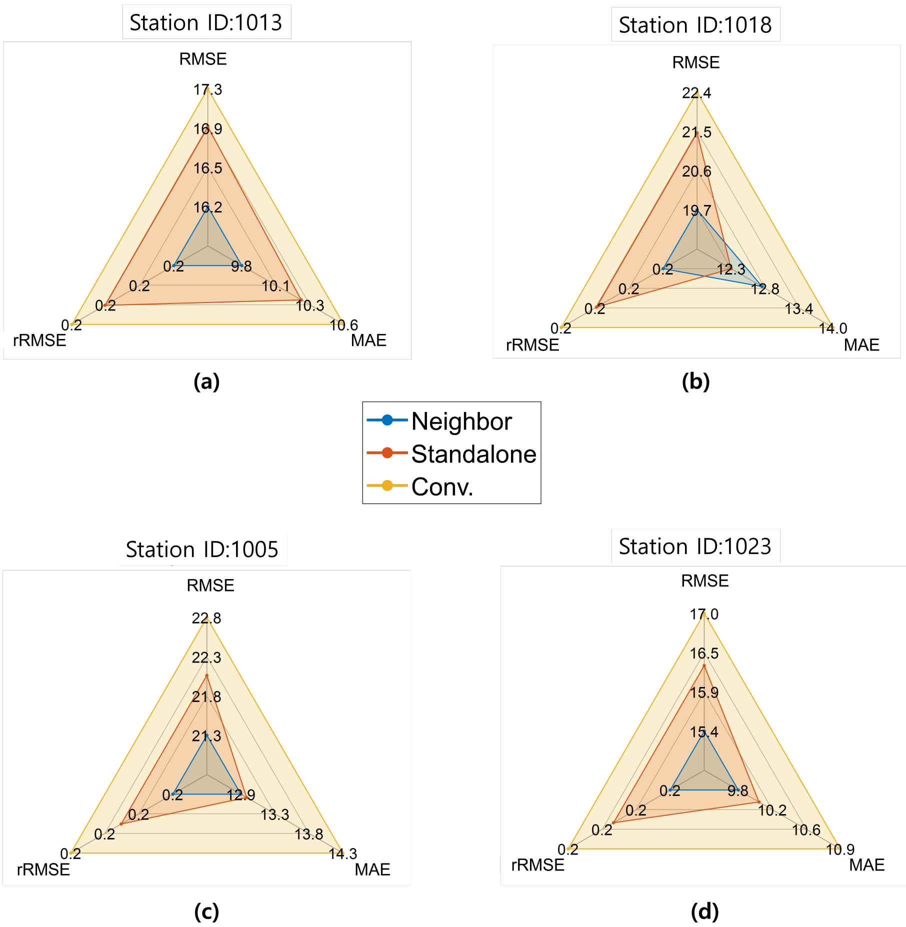

Figure 11. Radar chart of $PM_{2.5}$ concentration forecasting result for Station: (**a**) 1013. (**b**) 1018. (**c**) 1005. (**d**) 1023.

Figure 12 shows the results for Case Studies 1–2 as scatter plots. In Figure 12, the red line represents the R^2 line. R^2 represents the linear regression fit, showing the relationship between the forecast and actual values. The Neighbor model achieved the highest R^2 value of 0.9723 at Station 1013, slightly surpassing the Conv. model (0.9719) and the Standalone model (0.9716). All three models demonstrated a strong linear correlation with the actual values, indicating reliable forecast accuracy. However, the Neighbor model showed the closest alignment with the observed values. For Station 1018, the Neighbor model exhibited superior performance, achieving an R^2 value of 0.9737. This result outperformed both the Conv. model (0.9669) and the Standalone model (0.969). These findings highlight that integrating data from neighboring stations improves forecasting accuracy and better matches actual data. At Station 1005, the Neighbor model recorded an R^2 value of 0.9716, narrowly exceeding the Conv. model (0.9701) and the Standalone model (0.97). Although the performance differences among the three models were small, the Neighbor model demonstrated a distinct advantage in capturing the underlying patterns in the data. Station 1023 also saw the Neighbor model achieving the best R^2 value of 0.9703, outperforming the Standalone (0.9669) and Conv. (0.9655) models. This consistent performance across stations underscores the effectiveness of incorporating spatial data for more accurate forecasting.

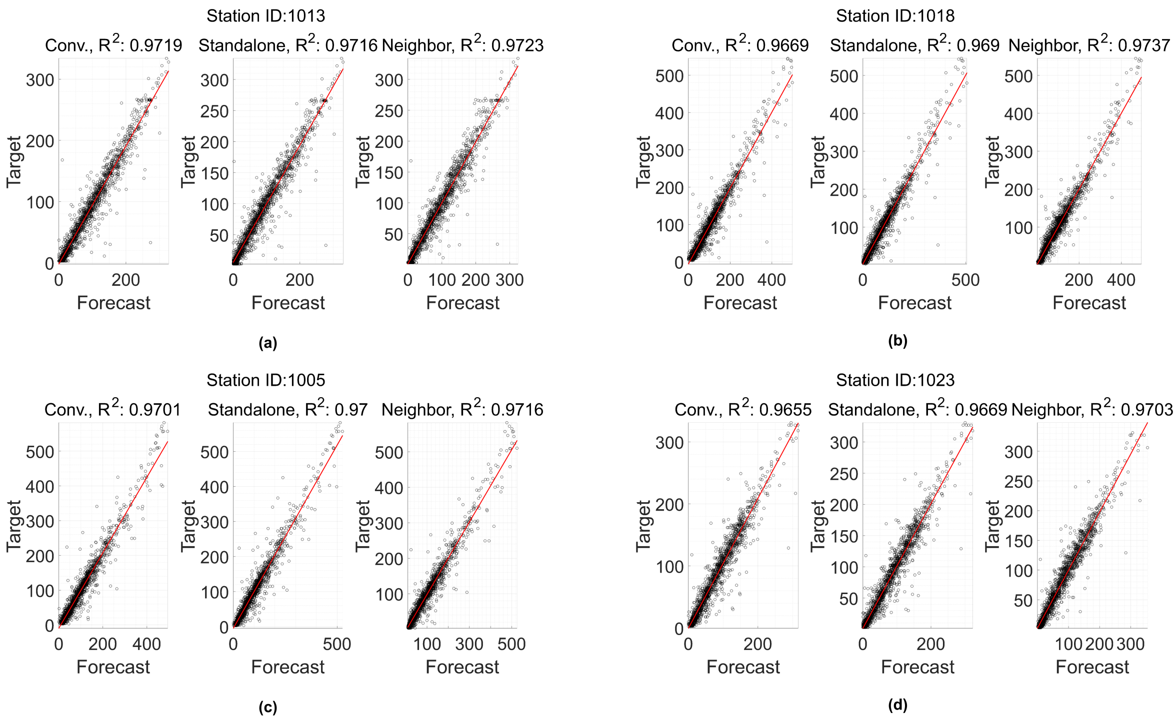

Figure 12. Scatter Plot of $PM_{2.5}$ Concentration Forecasting Result for Station: (**a**) 1013. (**b**) 1018. (**c**) 1005. (**d**) 1023.

Zhang et al.'s [63] study addressed a similar challenge in forecasting $PM_{2.5}$ concentrations. Table 11 presents the forecasting results for each monitoring station reported in [63]. The proposed model shows superior forecasting performance compared to the previous study across stations. At Station 1023, the proposed model reduced the RMSE from 23.741 to 15.3535. Similarly, the MAE decreased from 14.037 to 9.8057. The R^2 value also improved from 0.931 to 0.9409. These results indicate a better alignment between the forecasted and actual values. For Station 1005, the proposed model exhibited slightly better performance. It achieved an RMSE of 21.3221 compared to 18.055 in the previous study. The MAE was marginally lower, with values of 12.8822 versus 10.977. Additionally, the R^2 value increased from 0.938 to 0.9433, demonstrating improved correlation with the actual values. The proposed model demonstrated notable improvements over the previous study at Station 1018. It achieved an RMSE of 19.682, which is markedly lower than the 22.742 reported previously. The MAE also showed a notable reduction, decreasing from 13.979 to 12.789. Additionally, the R^2 value increased from 0.939 to 0.9480, indicating enhanced precision in forecasting $PM_{2.5}$ concentrations. This comparison shows that the proposed method achieves superior accuracy. It eliminates redundancy and effectively uses data from neighboring stations. Consequently, the enhanced performance of the proposed method validates its advantage over previous studies.

Table 11. Forecast Results from Related Studies

Station ID	RMSE	MAE	R^2
1023	23.741	14.037	0.931
1005	18.055	10.977	0.938
1018	22.742	13.979	0.939

4.4. Discussion

Accurately forecasting $PM_{2.5}$ concentrations is critical to mitigating the severe impacts of air pollution on human health and the environment. Providing reliable forecasting supports effective air quality management and informed policy decisions. This study introduces a novel approach to improving $PM_{2.5}$ forecasting accuracy by leveraging spatiotemporal data from neighboring monitoring stations. The proposed method integrates multiple advancements: selecting relevant neighboring stations using the mRMR algorithm, modeling nonlinear interactions with a CNN, and enhancing attention mechanisms with a refined scoring strategy.

Two case studies conducted at four monitoring stations in Beijing, China, demonstrated the effectiveness of the proposed method. When compared to conventional attention mechanisms, the enhanced attention score improved RMSE by 3.48%p, MAE by 8.60%p, R^2 by 0.49%p, rRMSE by 3.64%p, and PBIAS by 59.29%p. Furthermore, when data from neighboring stations are considered, we see even greater improvements, with RMSE decreasing by 5.49%p, MAE increasing by 0.51%p, R^2 increasing by 0.67%p, rRMSE improving by 5.44%p, and PBIAS improving by 46.56%p.

5. Conclusions

To mitigate the long-term adverse health effects of $PM_{2.5}$, this study forecasts $PM_{2.5}$ concentrations one hour ahead by incorporating both temporal and spatial information. The forecasting process consists of two steps. In the first step, mRMR is used to select neighboring stations, which incorporates spatial information, and a CNN integrates the data from these stations. In the second step, Seq2Seq-attention is used to forecast $PM_{2.5}$ concentrations based on the integrated data. Previous studies have used correlation coefficients to account for neighboring stations or have forecasted by considering all neighboring stations. These methods increase model complexity, potentially degrading forecast performance. In contrast, mRMR was chosen in this study because it enhances the correlation between the main and neighboring stations while reducing redundancy. This reduces model complexity and ensures that only relevant station data are used. Data integration is performed with a CNN to account for the nonlinearity in the selected data. Finally, Seq2Seq-attention is used for forecasting $PM_{2.5}$ concentrations, with the attention score emphasizing both similarity and dissimilarity. Two case studies were conducted to evaluate the performance of the proposed method. Case Study 1 (Section 4.2) compared forecasting performance with attention score improvement, while Case Study 2 (Section 4.3) evaluated forecasting performance based on whether or not data from neighboring stations were considered. The experimental results showed that the improved attention score led to better forecast performance by accounting for dissimilarity. Furthermore, incorporating both spatial and temporal information from neighboring stations significantly improved performance compared to the standalone station model. This study adopts a comprehensive approach to air quality forecasting. It combines advanced feature selection, nonlinear data integration, and refined attention mechanisms to address spatial and temporal complexities effectively. The results suggest that considering spatiotemporal information from neighboring stations can improve forecasting performance compared to conventional methods. However, the model has certain limitations. It forecasts $PM_{2.5}$ concentrations only one hour ahead. Additionally, it considers data from neighboring stations to be static, not accounting for potential dynamic changes in station relationships over time. This static approach may not fully capture the evolving nature of air quality patterns. Future work will improve the accuracy and applicability of the model in real-time air quality management by considering dynamic relationships between stations and extending the forecast range.

Author Contributions: Conceptualization, B.K., J.K. and E.K.; methodology, B.K.; software, B.K., S.J. and M.K.; validation, B.K., E.K., S.J. and M.K.; formal analysis, J.K.; investigation, J.K.; resources, J.K.; data curation, J.K.; writing—original draft preparation, B.K.; writing—review and editing, B.K., E.K., S.J. and M.K.; visualization, B.K. and J.K.; supervision, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by BK21FOUR, Creative Human Resource Education and Research Programs for ICT Convergence in the 4th Industrial Revolution.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Kim, H.C.; Kim, S.; Kim, B.U.; Jin, C.S.; Hong, S.; Park, R.; Son, S.W.; Bae, C.; Bae, M.; Song, C.K.; et al. Recent increase of surface particulate matter concentrations in the Seoul Metropolitan Area, Korea. *Sci. Rep.* **2017**, *7*, 4710. [CrossRef] [PubMed]
2. Han, L.; Zhou, W.; Li, W. Growing urbanization and the impact on fine particulate matter (PM2.5) dynamics. *Sustainability* **2018**, *10*, 1696. [CrossRef]
3. Zhang, Z.; Wang, S.; Yu, W.; Wang, P.; Zhang, H. Health Impacts of Fine Particulate Matter Shift Due to Urbanization in China. *Environ. Sci. Technol.* **2024**, *58*, 15732–15740. [CrossRef] [PubMed]
4. Pope Iii, C.A.; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA* **2002**, *287*, 1132–1141. [CrossRef]
5. Henning, R.J. Particulate matter air pollution is a significant risk factor for cardiovascular disease. *Curr. Probl. Cardiol.* **2024**, *49*, 102094. [CrossRef]
6. Lin, S.; Xue, Y.; Thandra, S.; Qi, Q.; Hopke, P.K.; Thurston, S.W.; Croft, D.P.; Utell, M.J.; Rich, D.Q. PM2.5 and its components and respiratory disease healthcare encounters–unanticipated increased exposure-response relationships in recent years after environmental policies. *Environ. Pollut.* **2024**, *360*, 124585. [CrossRef]
7. Sharma, R.; Kurmi, O.; Hariprasad, P.; Tyagi, S. Health implications due to exposure to fine and ultra-fine particulate matters: A short review. *Int. J. Ambient. Energy* **2024**, *45*, 2314256. [CrossRef]
8. World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; World Health Organization: Geneva, Switzerland, 2021.
9. Maji, K.J.; Ye, W.F.; Arora, M.; Nagendra, S.S. PM2.5-related health and economic loss assessment for 338 Chinese cities. *Environ. Int.* **2018**, *121*, 392–403. [CrossRef]
10. Wang, J.; Zhang, L.; Niu, X.; Liu, Z. Effects of PM2.5 on health and economic loss: Evidence from Beijing-Tianjin-Hebei region of China. *J. Clean. Prod.* **2020**, *257*, 120605. [CrossRef]
11. Liu, B.; Yan, S.; Li, J.; Li, Y.; Lang, J.; Qu, G. A spatiotemporal recurrent neural network for prediction of atmospheric PM2.5: A case study of Beijing. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 578–588. [CrossRef]
12. Vardoulakis, S.; Fisher, B.E.; Pericleous, K.; Gonzalez-Flesca, N. Modelling air quality in street canyons: A review. *Atmos. Environ.* **2003**, *37*, 155–182. [CrossRef]
13. Xiao, F.; Yang, M.; Fan, H.; Fan, G.; Al-Qaness, M.A. An improved deep learning model for predicting daily PM2.5 concentration. *Sci. Rep.* **2020**, *10*, 20988. [CrossRef]
14. Liao, Q.; Zhu, M.; Wu, L.; Pan, X.; Tang, X.; Wang, Z. Deep learning for air quality forecasts: A review. *Curr. Pollut. Rep.* **2020**, *6*, 399–409. [CrossRef]
15. Wang, L.; Wei, Z.; Wei, W.; Fu, J.S.; Meng, C.; Ma, S. Source apportionment of PM2.5 in top polluted cities in Hebei, China using the CMAQ model. *Atmos. Environ.* **2015**, *122*, 723–736. [CrossRef]
16. Geng, G.; Zhang, Q.; Martin, R.V.; van Donkelaar, A.; Huo, H.; Che, H.; Lin, J.; He, K. Estimating long-term PM2.5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model. *Remote. Sens. Environ.* **2015**, *166*, 262–270. [CrossRef]
17. Minh, V.T.T.; Tin, T.T.; Hien, T.T. PM2.5 forecast system by using machine learning and WRF model, a case study: Ho Chi Minh City, Vietnam. *Aerosol Air Qual. Res.* **2021**, *21*, 210108. [CrossRef]

18. Badicu, A.; Suciu, G.; Balanescu, M.; Dobrea, M.; Birdici, A.; Orza, O.; Pasat, A. PMs concentration forecasting using ARIMA algorithm. In Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 25–28 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
19. Zhang, L.; Lin, J.; Qiu, R.; Hu, X.; Zhang, H.; Chen, Q.; Tan, H.; Lin, D.; Wang, J. Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. *Ecol. Indic.* **2018**, *95*, 702–710. [CrossRef]
20. Amnuaylojaroen, T. Prediction of PM2.5 in an urban area of northern Thailand using multivariate linear regression model. *Adv. Meteorol.* **2022**, *2022*, 3190484. [CrossRef]
21. Wang, W.; Guo, Y. Air pollution PM2.5 data analysis in Los Angeles long beach with seasonal ARIMA model. In Proceedings of the 2009 International Conference on Energy and Environment Technology, Guilin, China, 16–18 October 2009; IEEE: Piscataway, NJ, USA, 2009; Volume 3, pp. 7–10.
22. Xing, H.; Wang, G.; Liu, C.; Suo, M. PM2.5 concentration modeling and prediction by using temperature-based deep belief network. *Neural Netw.* **2021**, *133*, 157–165. [CrossRef]
23. Xing, Y.; Yue, J.; Chen, C.; Xiang, Y.; Chen, Y.; Shi, M. A deep belief network combined with modified grey wolf optimization algorithm for PM2.5 concentration prediction. *Appl. Sci.* **2019**, *9*, 3765. [CrossRef]
24. Samal, K.K.R.; Babu, K.S.; Das, S.K. Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: A deep learning approach. *Urban Clim.* **2021**, *36*, 100800. [CrossRef]
25. Li, L.; Girguis, M.; Lurmann, F.; Pavlovic, N.; McClure, C.; Franklin, M.; Wu, J.; Oman, L.D.; Breton, C.; Gilliland, F.; et al. Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. *Environ. Int.* **2020**, *145*, 106143. [CrossRef] [PubMed]
26. Luo, Z.; Huang, F.; Liu, H. PM2.5 concentration estimation using convolutional neural network and gradient boosting machine. *J. Environ. Sci.* **2020**, *98*, 85–93. [CrossRef]
27. Chae, S.; Shin, J.; Kwon, S.; Lee, S.; Kang, S.; Lee, D. PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network. *Sci. Rep.* **2021**, *11*, 11952. [CrossRef]
28. Chen, Y.; Hu, C. Hourly PM2.5 Concentration Prediction Based on Empirical Mode Decomposition and Geographically Weighted Neural Network. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 79. [CrossRef]
29. Yang, H.; Wang, W.; Li, G. Multi-factor PM2. 5 concentration optimization prediction model based on decomposition and integration. *Urban Clim.* **2024**, *55*, 101916. [CrossRef]
30. Liu, Z.; Ji, D.; Wang, L. PM2. 5 concentration prediction based on EEMD-ALSTM. *Sci. Rep.* **2024**, *14*, 12636.
31. Oprea, M.; Popescu, M.; Mihalache, S.F. A neural network based model for PM 2.5 air pollutant forecasting. In Proceedings of the 2016 20th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, Romania, 13–15 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 776–781.
32. Chen, Y.C.; Lei, T.C.; Yao, S.; Wang, H.P. PM2.5 prediction model based on combinational hammerstein recurrent neural networks. *Mathematics* **2020**, *8*, 2178. [CrossRef]
33. Sepúlveda, T.; Nicolis, O.; Peralta, B. Predictions of PM2.5 concentrations and critical events in Santiago, Chile using Recurrent Neural Networks. In Proceedings of the 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Valparaiso, Chile, 13–27 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–7.
34. Hochreiter, S. Long Short-term Memory. In *Neural Computation*; MIT-Press: Cambridge, MA, USA, 1997.
35. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Gated feedback recurrent neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, Lille, France, 7–9 July 2015; pp. 2067–2075.
36. Sutskever, I. Sequence to Sequence Learning with Neural Networks. *arXiv* **2014**, arXiv:1409.3215.
37. Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
38. Liu, B.; Yan, S.; Li, J.; Qu, G.; Li, Y.; Lang, J.; Gu, R. An attention-based air quality forecasting method. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando FL, USA, 17–20 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 728–733.
39. Bai, W.; Li, F. PM2.5 concentration prediction using deep learning in internet of things air monitoring system. *Environ. Eng. Res.* **2023**, *28*, 210456. [CrossRef]
40. Pak, U.; Ma, J.; Ryu, U.; Ryom, K.; Juhyok, U.; Pak, K.; Pak, C. Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Sci. Total. Environ.* **2020**, *699*, 133561. [CrossRef]
41. Wang, W.; Zhao, S.; Jiao, L.; Taylor, M.; Zhang, B.; Xu, G.; Hou, H. Estimation of PM2.5 concentrations in China using a spatial back propagation neural network. *Sci. Rep.* **2019**, *9*, 13788. [CrossRef]
42. Liu, B.; Yan, S.; Li, J.; Li, Y. Forecasting PM2.5 concentration using spatio-temporal extreme learning machine. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 950–953.
43. 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), Kottayam, India, 13–14 February 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
44. Zhan, Y.; Luo, Y.; Deng, X.; Chen, H.; Grieneisen, M.L.; Shen, X.; Zhu, L.; Zhang, M. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.* **2017**, *155*, 129–139. [CrossRef]

45. Ma, J.; Ding, Y.; Gan, V.J.; Lin, C.; Wan, Z. Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM. *IEEE Access* **2019**, *7*, 107897–107907. [CrossRef]
46. Chen, Y. Prediction algorithm of PM2.5 mass concentration based on adaptive BP neural network. *Computing* **2018**, *100*, 825–838. [CrossRef]
47. Lightstone, S.D.; Moshary, F.; Gross, B. Comparing CMAQ forecasts with a neural network forecast model for PM2.5 in New York. *Atmosphere* **2017**, *8*, 161. [CrossRef]
48. Bera, B.; Bhattacharjee, S.; Sengupta, N.; Saha, S. PM2.5 concentration prediction during COVID-19 lockdown over Kolkata metropolitan city, India using MLR and ANN models. *Environ. Chall.* **2021**, *4*, 100155. [CrossRef]
49. Tsai, Y.T.; Zeng, Y.R.; Chang, Y.S. Air pollution forecasting using RNN with LSTM. In Proceedings of the 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), Athens, Greece, 12–15 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1074–1079.
50. Zhao, J.; Deng, F.; Cai, Y.; Chen, J. Long short-term memory-Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. *Chemosphere* **2019**, *220*, 486–492. [CrossRef]
51. Ayturan, Y.A.; Ayturan, Z.C.; Altun, H.O.; Kongoli, C.; Tuncez, F.D.; Dursun, Ş.; Ali, Ö. Short-term prediction of PM2.5 pollution with deep learning methods. *Glob. Nest J.* **2020**, *22*, 126–131.
52. Wang, D.; Wang, H.W.; Li, C.; Lu, K.F.; Peng, Z.R.; Zhao, J.; Fu, Q.; Pan, J. Roadside air quality forecasting in shanghai with a novel sequence-to-sequence model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 9471. [CrossRef] [PubMed]
53. Yan, L.; Wu, Y.; Yan, L.; Zhou, M. Encoder-decoder model for forecast of PM2.5 concentration per hour. In Proceedings of the 2018 1st International Cognitive Cities Conference (IC3), Okinawa, Japan, 7–9 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 45–50.
54. Vargas-Campos, I.R.; Villanueva, E. Comparative Study of Spatial Prediction Models for Estimating PM 2.5 Concentration Level in Urban Areas. In Proceedings of the Annual International Conference on Information Management and Big Data, Lima, Peru, 1–3 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 169–180.
55. Yang, M.; Fan, H.; Zhao, K. PM2.5 prediction with a novel multi-step-ahead forecasting model based on dynamic wind field distance. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4482. [CrossRef]
56. Yeo, I.; Choi, Y.; Lops, Y.; Sayeed, A. Efficient PM2.5 forecasting using geographical correlation based on integrated deep learning algorithms. *Neural Comput. Appl.* **2021**, *33*, 15073–15089. [CrossRef]
57. Zhang, K.; Yang, X.; Cao, H.; Thé, J.; Tan, Z.; Yu, H. Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning. *Environ. Int.* **2023**, *171*, 107691. [CrossRef]
58. Zíková, N.; Wang, Y.; Yang, F.; Li, X.; Tian, M.; Hopke, P.K. On the source contribution to Beijing PM2.5 concentrations. *Atmos. Environ.* **2016**, *134*, 84–95. [CrossRef]
59. Yan, D.; Lei, Y.; Shi, Y.; Zhu, Q.; Li, L.; Zhang, Z. Evolution of the spatiotemporal pattern of PM2.5 concentrations in China–A case study from the Beijing-Tianjin-Hebei region. *Atmos. Environ.* **2018**, *183*, 225–233. [CrossRef]
60. Zheng, Y.; Yi, X.; Li, M.; Li, R.; Shan, Z.; Chang, E.; Li, T. Forecasting Fine-Grained Air Quality Based on Big Data. In Proceedings of the 21th SIGKDD conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; Available online: https://www.microsoft.com/en-us/research/publication/forecasting-fine-grained-air-quality-based-on-big-data/?from=https://research.microsoft.com/apps/pubs/?id=246398&type=exact (accessed on 27 November 2024)
61. Cheng, Z.; Li, L.; Liu, J. Identifying the spatial effects and driving factors of urban PM2.5 pollution in China. *Ecol. Indic.* **2017**, *82*, 61–75. [CrossRef]
62. Zeng, M.; Du, J.; Zhang, W. Spatial-temporal effects of PM2.5 on health burden: Evidence from China. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4695. [CrossRef]
63. Zhang, L.; Na, J.; Zhu, J.; Shi, Z.; Zou, C.; Yang, L. Spatiotemporal causal convolutional network for forecasting hourly PM2.5 concentrations in Beijing, China. *Comput. Geosci.* **2021**, *155*, 104869. [CrossRef]
64. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
65. Lin, M.; Chen, Q.; Yan, S. Network In Network. *arXiv* **2013**, arXiv:1312.4400.
66. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
67. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
68. Salehie, O.; Jamal, M.H.B.; Shahid, S. Characterization and prediction of PM2.5 levels in Afghanistan using machine learning techniques. *Theor. Appl. Climatol.* **2024**, *155*, 9081–9097. [CrossRef]
69. Adnan, R.M.; Keshtegar, B.; Abusurrah, M.; Kisi, O.; Alkabaa, A.S. Enhancing solar radiation prediction accuracy: A hybrid machine learning approach integrating response surface method and support vector regression. *Ain Shams Eng. J.* **2024**, *15*, 103034. [CrossRef]
70. Ghimire, S.; AL-Musaylh, M.S.; Nguyen-Huy, T.; Deo, R.C.; Acharya, R.; Casillas-Pérez, D.; Yaseen, Z.M.; Salcedo-Sanz, S. Explainable deeply-fused nets electricity demand prediction model: Factoring climate predictors for accuracy and deeper insights with probabilistic confidence interval and point-based forecasts. *Appl. Energy* **2025**, *378*, 124763. [CrossRef]

71. Tao, H.; Abba, S.I.; Al-Areeq, A.M.; Tangang, F.; Samantaray, S.; Sahoo, A.; Siqueira, H.V.; Maroufpoor, S.; Demir, V.; Bokde, N.D.; et al. Hybridized artificial intelligence models with nature-inspired algorithms for river flow modeling: A comprehensive review, assessment, and possible future research directions. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107559. [CrossRef]
72. Samantaray, S.; Sahoo, A.; Paul, S.; Ghose, D.K. Prediction of bed-load sediment using newly developed support-vector machine techniques. *J. Irrig. Drain. Eng.* **2022**, *148*, 04022034. [CrossRef]