

Article

Outdoor Particulate Matter Correlation Analysis and Prediction Based Deep Learning in the Korea

Minsu Chae ¹, Sangwook Han ¹ and HwaMin Lee ^{2,*} 

¹ Department of Computer Science, Soonchunhyang University, Asan 31538, Korea; cms@sch.ac.kr (M.C.); sanguk@sch.ac.kr (S.H.)

² Department of Computer Software Engineering, Soonchunhyang University, Asan 31538, Korea

* Correspondence: leehm@sch.ac.kr; Tel.: +82-41-530-4761

Received: 26 May 2020; Accepted: 30 June 2020; Published: 15 July 2020



Abstract: Particulate matter (PM) has become a problem worldwide, with many deleterious health effects such as worsened asthma, affected lungs, and various toxin-induced cancers. The International Agency for Research on Cancer (IARC) under the World Health Organization (WHO) has designated PM as a group 1 carcinogen. Although Korea Environment Corporation forecasts the status of outdoor PM four times a day, whichever is higher among PM₁₀ and PM_{2.5}. Korea Environment Corporation forecasts for the stages of PM. It remains difficult to predict the value of PM when going out. We correlate air quality and solar terms, address format, and weather data, and PM in the Korea. We analyzed the correlation between address format, air quality data, and weather data, and PM. We evaluated performance according to the sequence length and batch size and found the best outcome with a sequence length of 7 days, and a batch size of 96. We performed PM prediction using the Long Short-Term Recurrent Unit (LSTM), the Convolutional Neural Network (CNN), and the Gated Recurrent Unit (GRU) models. The CNN model suffered the limitation of only predicting from the training data, not from the test data. The LSTM and GRU models generated similar prediction results. We confirmed that the LSTM model has higher accuracy than the other two models.

Keywords: particulate matter; prediction; air quality; analysis; deep learning

1. Introduction

Recently, PM has become a problem worldwide. The causes of PM are industries, etc [1]. PM₁₀ and PM_{2.5} mean that the length of the diameter is 10 μm or less and 2.5 μm or less, respectively. Air pollution causes 7 million deaths annually worldwide, including children [2]. Air pollution affects the respiratory symptoms, causes breathing problems, and decreases respiratory function [3–12]. In particular, PM worsens asthma and affect lung [13–16]. In 2013, the IARC under the WHO designated PM as a group 1 carcinogen. In 2018, to prevent the deterioration of health caused by PM, the Korea strengthened the definition for the level of PM to that of advanced countries (Table 1). PM is generated during the process of power plants and vehicles [17]. PM₁₀ reduced through the filter of the power plant, but PM_{2.5} is not reduced through the filter of power plant [17]. The causes of PM in the Korea are domestic, and international from neighboring countries such as China [17–19]. The Government of South Korea estimates that it is 30 to 50% of ultrafine dust from China as a cause of air pollution [20]. Deputy Mayor for Climate & Environment in the Korea estimates that the ratio of ultrafine dust generated in Korea is 50–70% [21]. According to a report by the Korea Environmental Industry & Technology Institute (KEITI) [22], it is estimated that 51% of ultra-fine dust in Seoul is generated. A high concentration of PM occurs every day in the Korea [21].

Table 1. Air Quality Standard.

Countries	1hour Average on PM ₁₀	24hour Average on PM ₁₀	An annual Average on PM ₁₀	24hour Average on PM _{2.5}	An Annual Average on PM _{2.5}
The Korea	-	100 µg/m ³	50 µg/m ³	35 µg/m ³	15 µg/m ³
USA	-	150 µg/m ³	-	35 µg/m ³	15 µg/m ³
Japan	200 µg/m ³	100 µg/m ³	-	35 µg/m ³	15 µg/m ³
China	-	150 µg/m ³	70 µg/m ³	75 µg/m ³	35 µg/m ³
WHO	-	50 µg/m ³	20 µg/m ³	25 µg/m ³	10 µg/m ³
EU	-	50 µg/m ³	40 µg/m ³	-	25 µg/m ³

In addition, Korea Environment Corporation forecasts the status of PM, whichever is higher among PM₁₀ and PM_{2.5}, for times a day (5 AM, 11 AM, 5 PM, and 11 PM) according to four stages: Good, Normal, Bad, and Very bad (Table 2). Korea Environment Corporation forecasts for the stages of PM. Therefore, it is difficult to predict the value of PM when going out. Recently, the prediction of PM in the Korea has been studied [23–25].

Table 2. Stages of PM in the Korea.

Status	PM ₁₀	PM _{2.5}
Good	0~30 µg/m ³	0~15 µg/m ³
Normal	31~80 µg/m ³	16~35 µg/m ³
Bad	81~150 µg/m ³	36~75 µg/m ³
Very bad	150 µg/m ³ over	76 µg/m ³ over

Yi et al. [26] used air quality information and PM, and included air quality data, PM data, and location information in the image, and used location information using a grid pattern through the ConvLSTM model. However, it was applied to only one city, and only air quality data and PM data were used. The PM was correlated with weather data, but Yi et al. did not consider the weather data. Their study also suffered the limitation of needing to perform pre-processing to create a grid image based on time series data. Vong et al. [27] used SVM and MLM models to predict PM in Macau, China. Vong et al. used air quality data and weather data. When using 11 variables, the accuracy of the ELM model is approximately 81%, and the accuracy of the SVM model is approximately 79%. When using 52 variables, the accuracy of the ELM model is approximately 75%, and the accuracy of the SVM model is approximately 78%. PM_{2.5} had more adverse effects on health, but Vong predicted only PM₁₀. Xayasouk et al. [25] studied PM with Deep AutoEncoder (DAE) and LSTM model in Seoul, the Korea. The LSTM model was more accurate than the DAE model. Accordingly, we consider location information and air quality data and predict PM_{2.5} as well as PM₁₀. We consider the Recurrent Neural Network (RNN) model such as the LSTM model. We analyze the correlation between address format, air quality data, and weather data, and PM and develop the PM prediction model.

2. Materials

2.1. Air Pollution Stations

In the Korea, air pollution stations are installed in eight cities and nine provinces (Table 3). The Korea has seven metropolitan cities, one self-governing city, eight provinces, and one self-governing province. We use air quality data from 2015 to 2019. Air quality data were collected by AirKorea [28]. The collected air quality data include PM₁₀, PM_{2.5}, O₃, CO, SO₂, and NO₂ and were measured every hour; however, some data were missing due to communication problems in the air pollution stations.

Table 3. The number of air pollution stations in each city and province.

City Name	The Number of Stations	Province Name	The Number of Stations
Seoul	40	Gyeonggi-do	108
Incheon	25	Gangwon-do	25
Daejeon	12	Chungcheongsam-do	34
Daegu	15	Chungcheongbuk-do	29
Ulsan	17	Gyeongsangnuk-do	42
Busan	29	Gyeongsangnam-do	38
Gwangju	11	Jeollabuk-do	32
Sejong	4	Jeollanam-do	38
		Jeju Island	88

2.2. Weather Stations

In the Korea, weather stations are installed in eight cities and nine provinces (Table 4). We used only the weather data measured at stations with the Automated Surface Observing System (ASOS) installed. The Korea has seven metropolitan cities, one self-governing city, eight provinces, and one self-governing province. We use weather data from 2015 to 2019. Weather data were collected by the Korea Meteorological Administration [29]. The collected weather data were temperature, humidity, wind speed, wind direction, snowfall, precipitation, and atmospheric pressure, and were measured every hour; however, some data were missing due to communication problems in the air pollution stations.

Table 4. The number of weather stations in each city and province.

City name	The Number of Stations	Province Name	The Number of Stations
Seoul	2	Gyeonggi-do	5
Incheon	3	Gangwon-do	15
Daejeon	1	Chungcheongsam-do	6
Daegu	2	Chungcheongbuk-do	5
Ulsan	1	Gyeongsangnuk-do	14
Busan	1	Gyeongsangnam-do	10
Gwangju	1	Jeollabuk-do	16
Sejong	1	Jeollanam-do	14
		Jeju Island	4

2.3. Data Description

Air quality data and weather data are important for predicting PM. We used Korea's air quality data and weather data from 1 January 2015 to 30 November 2019, because air quality data provided by Air Korea did not have final measurement data as of December 2019. We added weather data for each region to air quality data. In some locations, air quality data were available, but weather data were not, so these locations were deleted. In addition, we deleted data in areas where air quality data were measured, but weather data were not. We combined weather data and air quality data based on address and time.

2.3.1. Weather Data

Weather data is meaningful data in predicting PM. Temperature affects air quality data [30]. Low humidity and low temperature are related to the PM_{2.5} concentration [31]. Also, the wind carries PM and affects the PM concentration [30,31]. In addition, rain removes air pollution [32,33].

2.3.2. Air Quality Data

NO₂ and CO are related to automobile emissions [33], and SO₂ and PM are related to factory emissions. O₂ is produced when there are NO₂ and NO in the atmosphere [34]. SO₂, NO₂, and O₃ have a high correlation with PM [34], and a low correlation between CO and PM [35].

2.3.3. Address Format

PM is affected by the surrounding cities because it moves in the wind [34,36]. The address format used a total of three methods. The first method was divided into (1) city name, (2) district name or county name or ward name, and (3) dong name or eup name or myeon name. The second method was divided into (1) province name, (2) city name or county name or and (3) dong name or eup name or, and myeon name. The third method is the address of the station. We removed the building name from the address of the air quality station. Accordingly, we used the address format using the first method or the second method and the third method. We converted the address format using the one-hot encoding library.

2.3.4. Twenty-Four Solar Terms

The 24 solar terms are divided into seasons. Twenty-four solar terms are significant in time series data affected by weather [37]. We divided the 24 solar terms by dates provided by the Korea Meteorological Administration [38]. We converted the 24 solar terms using the one-hot encoding library.

2.3.5. Wind_x, Wind_y

The wind is expressed in terms of wind speed and wind direction. The wind direction represents the direction from 0 to 360 degrees, and the wind speed represents the distance per second of the wind. The accuracy is higher when calculating wind_x and wind_y than when using wind speed and wind direction [39].

$$\text{wind_x} = \text{wind}_{\text{direction}} \times \cos(\text{wind}_{\text{speed}}) \quad (1)$$

$$\text{wind_y} = -\text{wind}_{\text{direction}} \times \sin(\text{wind}_{\text{speed}}) \quad (2)$$

2.3.6. Data Parameters

We predict PM₁₀ and PM_{2.5} using air quality data and weather data. Air quality data include SO₂, O₃, CO, and NO₂. Weather data include temperature, atmospheric pressure, humidity, precipitation, snowfall, precipitation, and snowfall. We performed deep learning using four air quality data, seven weather data, wind_x, wind_y, and 24 solar terms, and address format. The following shows the input data, measurement unit, and range (Table 5). We used data from 2015 to 2019. In the dataset, 70% were used as training data, 24% were used as verification data, and 6% were used as test data. We corrected the missing values for precipitation and snowfall to zero. CO and NO₂ were corrected using the average. Resampling was performed for SO₂ and O₃, temperature, air pressure, wind speed, wind direction, and humidity. SO₂ has 631319 missing values, CO has 636586, O₃ has 474455, and NO₂ has 492315. Temperatures have 470 missing values, wind speeds have 5242, wind directions have 19926, humidities have 6052, and atmospheric pressures have 1554. If the value of IsRain is 0, it means that it does not rain, and if it is 1, it means rain. If the value of IsSnow is 0, it means no snow, and 1 means snow.

Table 5. Input data.

Variable	Unit	Range	Variable	Unit	Range
Address1	Integer	≥ 0	PM ₁₀	$\mu\text{g}/\text{m}^3$	0–400
Address2	Integer	≥ 0	temperature	$^{\circ}\text{C}$	–25–45
Address3	Integer	≥ 0	precipitation	mm	≥ 0
Full Address	Integer	≥ 0	wind_x	Float	–12–12
Solar Terms	Integer	0–23	wind_y	Float	–12–12
SO ₂	ppm	≥ 0	humidity	%	0–100
CO	ppm	≥ 0	atmospheric pressure	hPa	≥ 0
O ₃	ppm	≥ 0	snowfall	cm	≥ 0
NO ₂	ppm	≥ 0	IsRain	Integer	0 or 1
PM _{2.5}	$\mu\text{g}/\text{m}^3$	0–180	IsSnow	Integer	0 or 1

We correlated the input data (Figure 1). Correlation analysis excludes Address1, Address2, Address3, Full Address, IsRain, and IsSnow, which are values expressed as integers, because only continuous time-series data has meaning.

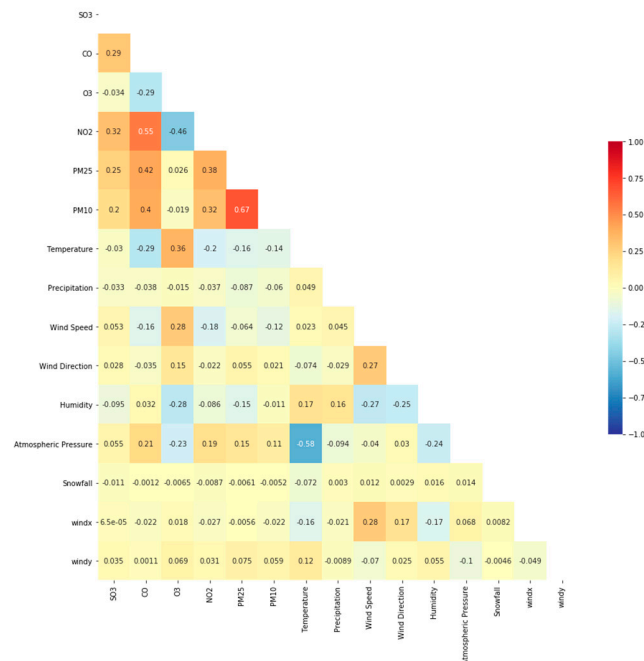


Figure 1. The correlation analysis result.

3. Methods

In previous studies [40], the CNN-LSTM hybrid model showed better performance than the CNN and LSTM models. However, we excluded the hybrid models due to its low accuracy. We used the GRU model that is similar to the LSTM model, as well as the CNN model and LSTM model. We applied the CNN-LSTM and CNN-GRU hybrid models.

3.1. Long Short-Term Recurrent Unit

The LSTM model is one of the RNN models. The RNN model takes the previous results as input, causing long-term dependency problems (Figure 2).

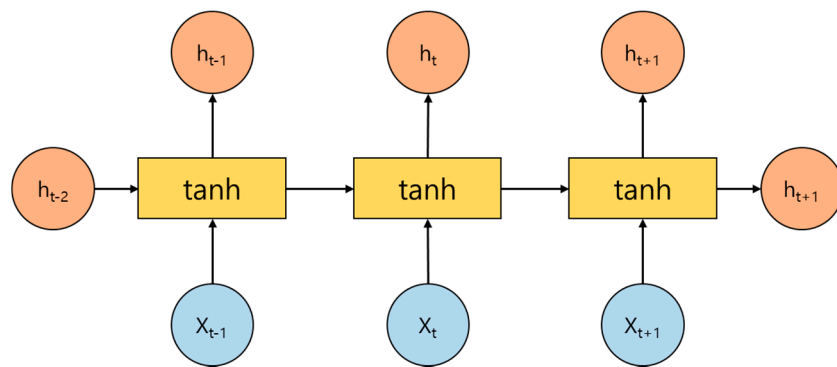


Figure 2. Structure of the RNN model.

The LSTM model is one of the models that solves the long-term dependency problem in the RNN model. Below is the structure of the LSTM model (Figure 3). The LSTM model has high accuracy for the prediction of time-series data.

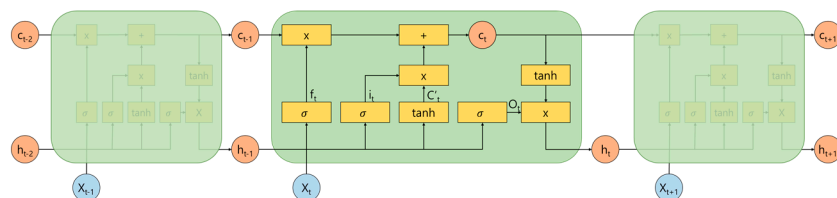


Figure 3. Structure of the LSTM model.

Table 6 show the variables used in LSTM model. The LSTM model implements the logistic sigmoid function through Equation (3).

$$\sigma = \frac{1}{1 + e^{-1}} \quad (3)$$

Table 6. Variables of the LSTM model.

Variable	Description
σ	Logistic sigmoid function
t	current state. $t - 1$ means the previous state.
i	unit of input gate. i_t means a unit of input gate of the current state.
f	unit of forget gate. f_t means a unit of forgetting gate of the current state.
o	unit of output gate. o_t means a unit of output gate of the current state.
x	Input value. x_t means the input value of the current state.
b	Bias value. b_f means bias of forget gate.
W	Weight matrix. W_f means weight matrix of forget gate.
c	Cell for long term memory.
	c_t means cell for long term memory of the current state.
h	hidden state for short term memory
	h_t means hidden state for short term memory of the current state.

The LSTM model uses Equation (4) to determine which data to keep and which to delete.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (4)$$

The LSTM model determines the data to be updated with long-term memory through Equations (5) and (6). The LSTM model updates long-term memory through Equation (7).

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (5)$$

$$c'_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (6)$$

$$c_t = f_t \times c_{t-1} + i_t \times c'_t \quad (7)$$

The LSTM model calculates the results through Equations (8) and (9).

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t \times \tanh(o_t) \quad (9)$$

3.2. Gated Recurrent Unit

The GRU model is an improved model of complex equations in the LSTM model. The following is the structure of the GRU model (Figure 4). The GRU model consists of the reset gate and the update gate. The GRU model is a model that improves the computation speed of the LSTM model.

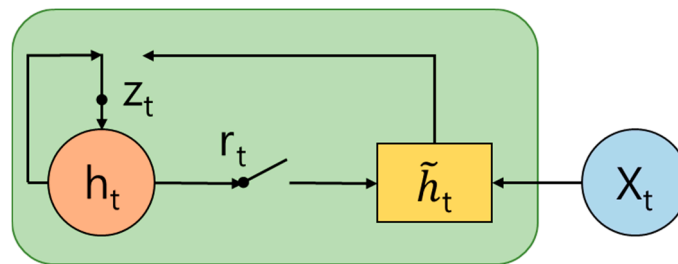


Figure 4. Structure of the GRU model.

Table 7 show the variables used in GRU model. In the GRU model, the logistic sigmoid function was implemented through Equation (3). The GRU model does not update or update h depending on the r_t the value obtained through Equation (10). Through this, the long-term memory modification in the LSTM was improved.

$$r_j^t = \sigma([W_r x]_j + [U_r h_{t-1}]_j) \quad (10)$$

Table 7. Variables of the GRU model.

Variable	Description
σ	Logistic sigmoid function
t	current state. $t - 1$ means the previous state.
j	Index of hidden unit
r	Unit of reset gate. r_j^t means unit of reset gate of current state when calculating the j -th hidden unit. It decides whether the previous hidden state is ignored.
z	Unit of update gate. z_j^t means a unit of update gate of current state when calculating the j -th hidden unit. It selects whether the hidden state is to be updated with a new hidden state.
U	Weight matrix. U_r means weight matrix of reset gate.
W	Weight matrix. W_r means weight matrix of the reset gate.
$\tilde{h}$	A new hidden state. $\tilde{h}_j^t$ means a new hidden state of current state when calculating the j -th hidden unit.
h	The hidden state. h_j^t means the hidden state of the current state when calculating the j -th hidden unit. The output is the hidden state.

The GRU model uses h_{t-1} or $\tilde{h}$ according to the z_t value through Equation (11). It updates $\tilde{h}$ through Equation (12). The predicted value is output through Equation (13).

$$z_j^t = \sigma([W_z x]_j + [U_z h_{t-1}]_j) \quad (11)$$

$$\tilde{h}_j^t = \tanh([Wx]_j + [U(r_t \odot h_{t-1})]_j) \quad (12)$$

$$h_j^t = z_j^t h_j^{t-1} + (1 - z_j^t) \tilde{h}_j^t \quad (13)$$

3.3. Convolutional Neural Network

The CNN model is used to extract hidden patterns through the convolution and pooling stages.

4. Experiments and Result

4.1. Evaluation Methods

We evaluated the performance using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). In deep learning, regression is verified via errors. MAE and RMSE are expressions of errors, and the lower the value, the better the performance. MAE and RMSE are calculated using Equations (14) and (15), respectively. MAE and RMSE use the difference between actual and predicted values to determine how many errors have occurred. The lower the values of MAE and RMSE, the higher the accuracy. Table 8 are variables used in MAE and RMSE.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i| \quad (14)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2} \quad (15)$$

Table 8. Variables.

Variable	Description
n	The number of data
i	The index of elements
y_i	Real value
y'_i	Predicted value

4.2. Data Prepare

We corrected outliers and missing values in the collected data. When the ranges of PM_{2.5} and PM₁₀, humidity, temperature, precipitation, and snowfall were outside the range of Table 5, the average was corrected. The missing value of precipitation and snowfall was corrected to zero because it means that there was no rain or snow at that time. We resampled the missing values using the LSTM model to process the missing values of air quality data and weather data. When resampling air pollution information, we corrected the missing value using the average value (Table 9). The small values of the units of air quality data value made it difficult to understand whether it is appropriately predicted only by MAE and RMSE.

Table 9. Results of MAE and RMSE of air quality data.

Evaluation Method	SO ₂	CO	NO ₂	O ₃
MAE	0.0019767759774657213	0.13297266375870512	0.004597223937342291	0.00520195667252871
RMSE	0.002783199108899503	0.19131779740725147	0.006546173996365833	0.007070109851593688

Accordingly, we evaluated the air pollution information by referring to a comparison graph between the actual (blue line) and predicted values (orange line; Figure 5). As a result of the comparison,

NO_2 and O_3 are predicted similarly, but SO_2 and CO are not. We resampled the missing values of NO_2 and O_3 using the predicted values. Because SO_2 and CO are not predicted accurately, the missing values were corrected using the average.

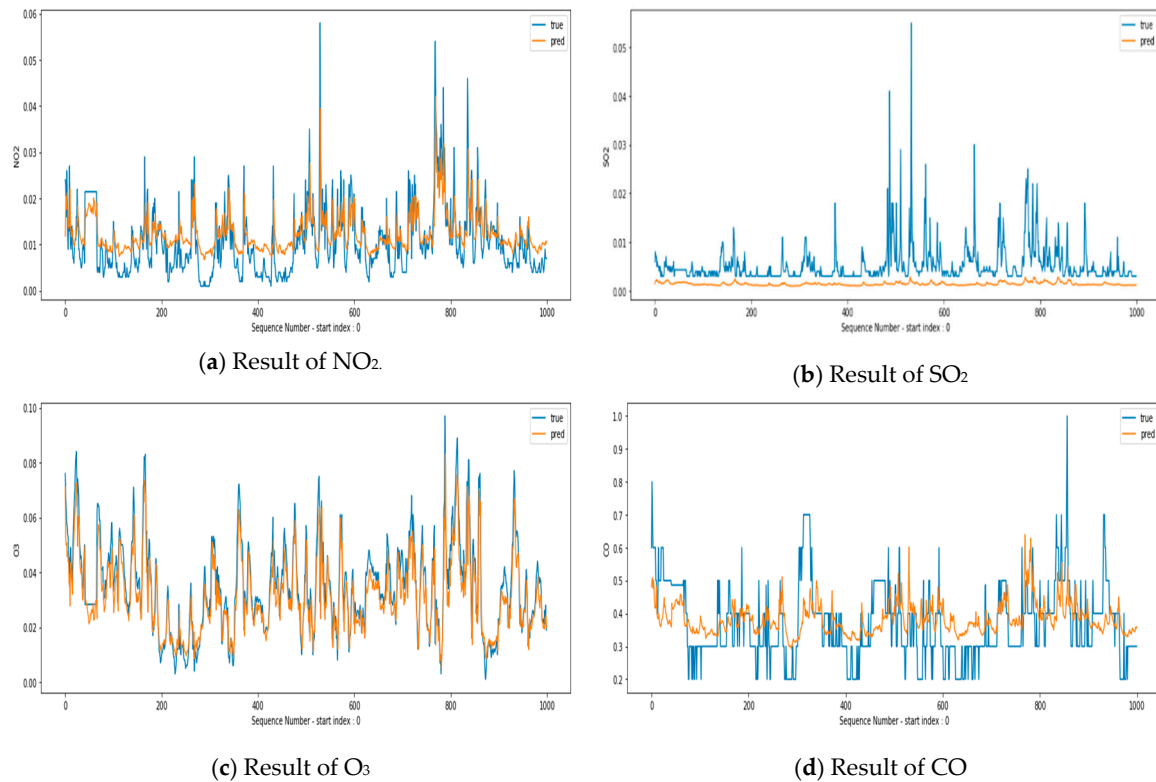


Figure 5. A comparison graph of actual and predicted values for NO_2 , SO_2 , O_3 , and CO. (a) Predicted and actual value of NO_2 ; (b) predicted and actual values of SO_2 ; (c) predicted and actual values of O_3 ; (d) predicted and actual values of CO.

When resampling the weather data, missing values were corrected using the average value. We predicted weather data for evaluation (Table 10) and compared the graph between the predicted (orange line) value and the actual (blue line) value (Figure 6). As a result of the comparison, atmospheric pressure, temperature, wind speed, wind direction, and humidity were predicted similarly. We have resampled the missing values of weather data using predicted values.

Table 10. Results of MAE and RMSE of air quality data.

Evaluation Method	Atmospheric Pressure	Temperature	Wind Speed	Wind Direction	Humidity
MAE	9.709005228526323	2.452505929222188	0.9497748647253894	9.97904782589953	4.971006381407262
RMSE	12.06379300207477	3.124994324296823	1.1238794798143990	4.41206318590864	6.291656963869701

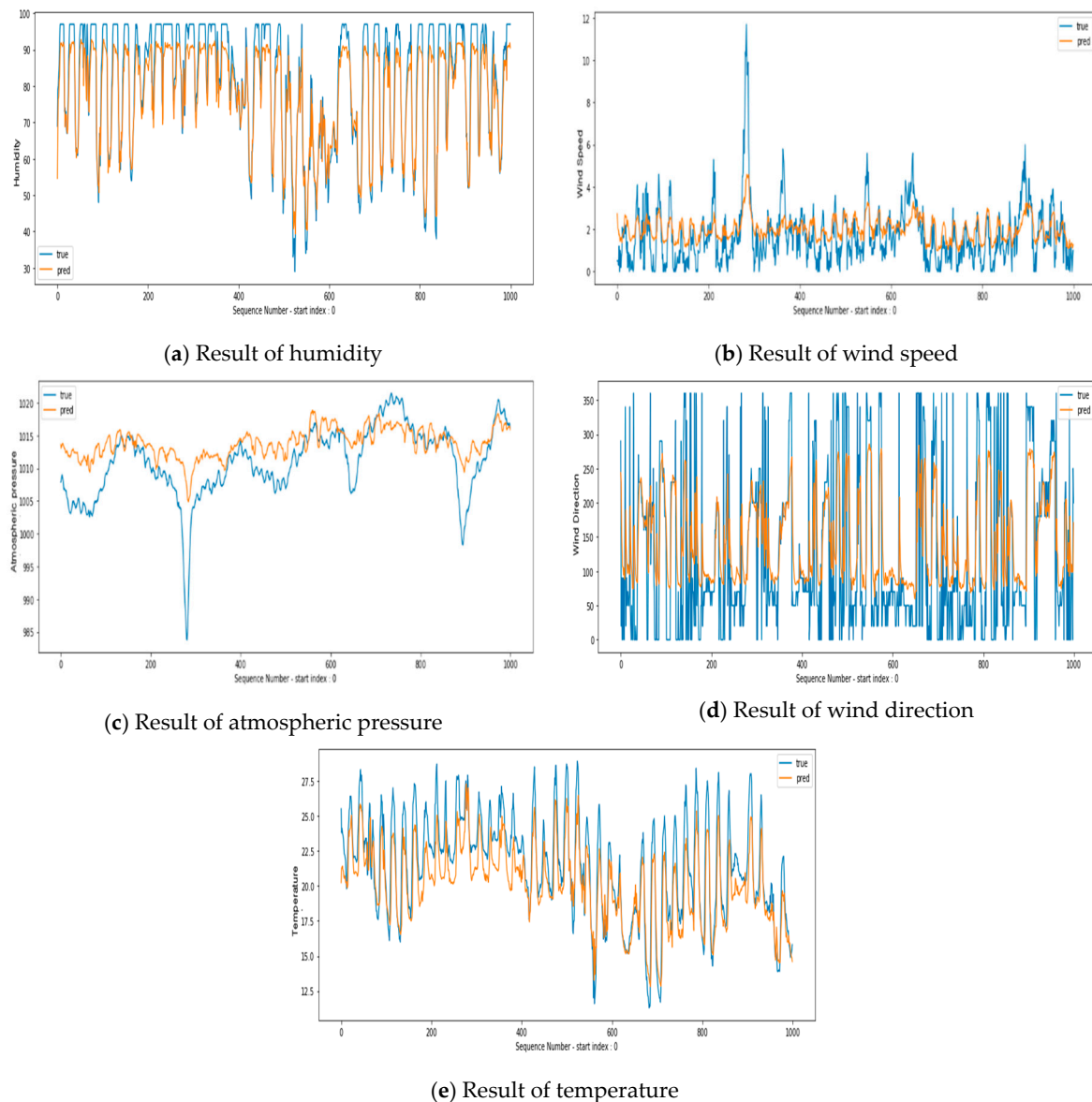


Figure 6. A comparison graph of actual and predicted values for atmospheric pressure, temperature, wind speed, wind direction, and humidity. (a) Predicted and actual values of humidity; (b) predicted and actual values of wind speed, (c) predicted and actual values of atmospheric pressure; (d) predicted and actual values of wind direction; (e) predicted and actual values of temperature.

4.3. Comparison of PM Prediction Accuracy According to SO_2

We performed PM prediction using only weather data to compare by air quality data (Table 11). Based on the results in Table 11, we analyzed the correlation by air quality data. Forecasts were made for PM_{10} and $PM_{2.5}$ using only weather data (Figure 7). The orange line is the predicted value, and the blue line is the actual value.

Table 11. Evaluation of the predicted PM using weather data.

$PM_{2.5}$ RMSE	$PM_{2.5}$ MAE	PM_{10} RMSE	PM_{10} MAE
11.424509950830299	7.386990585039382	7.810868049063216	5.192483483811323

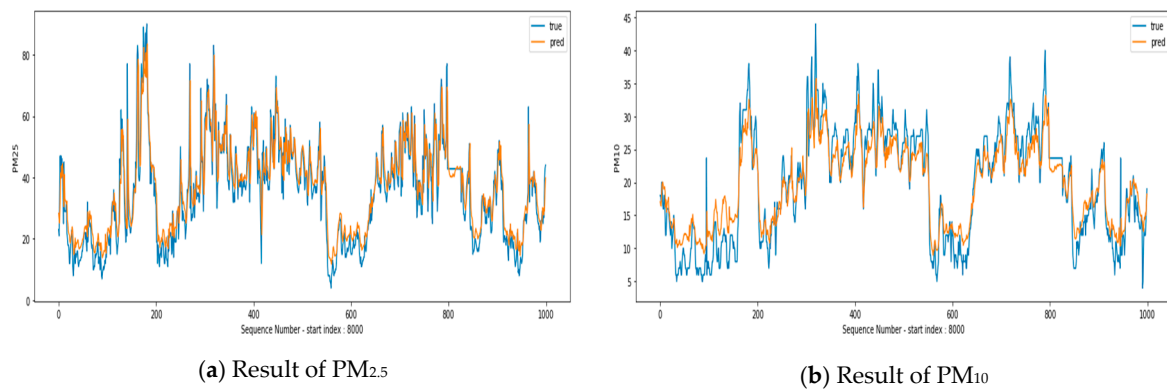


Figure 7. A graph of the actual and predicted values of PM₁₀ and PM_{2.5}. (a) Predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

We analyzed the correlation between SO₂ and PM by comparing weather data and weather data with SO₂. As a result, it was confirmed that RMSE and MAE of PM_{2.5} and PM₁₀ were lowered (Table 12), and hence SO₂ are related to PM_{2.5} and PM₁₀. The actual and predicted values of PM₁₀ and PM_{2.5} were compared (Figure 8). The orange line is the predicted value, and the blue line is the actual value. We confirmed that the accuracy of PM prediction is improved via SO₂.

Table 12. Evaluation of the predicted PM using weather data with SO₂.

PM _{2.5} RMSE.	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
11.150591658178262	7.212869946039611	7.266045101108083	5.150813651956774

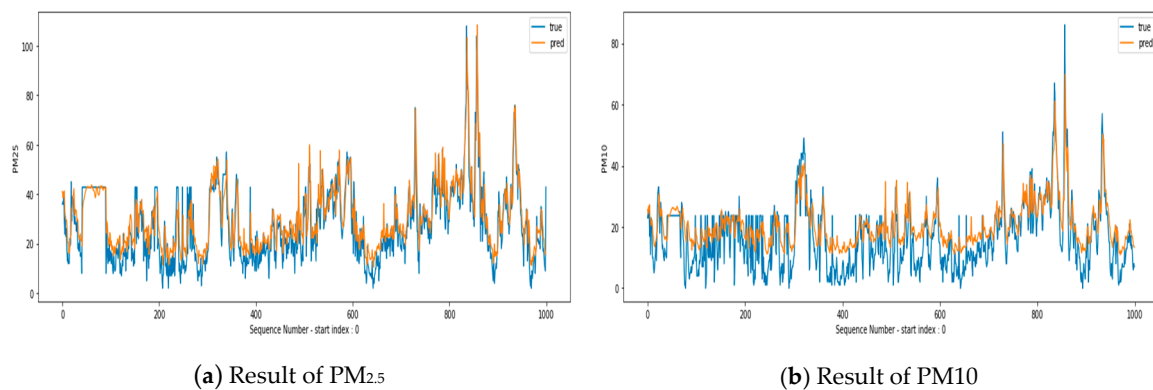


Figure 8. A graph of the actual and predicted values of PM₁₀ and PM_{2.5} using weather data with SO₂. (a) The predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

4.4. Comparison of PM Prediction Accuracy According to CO

We analyzed the correlation between CO and PM by comparing weather data and weather data with CO. As a result, it was confirmed that RMSE and MAE of PM_{2.5} and PM₁₀ were lowered (Table 13), and hence CO is related to PM_{2.5} and PM₁₀. The actual and predicted values of PM₁₀ and PM_{2.5} were compared (Figure 9). The orange line is the predicted value, and the blue line is the actual value. We confirmed that the accuracy of PM prediction is improved via CO.

Table 13. Evaluation of the predicted PM using weather data with CO.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
10.876065670272803	6.952868615163375	6.646710365161875	4.482592977233096

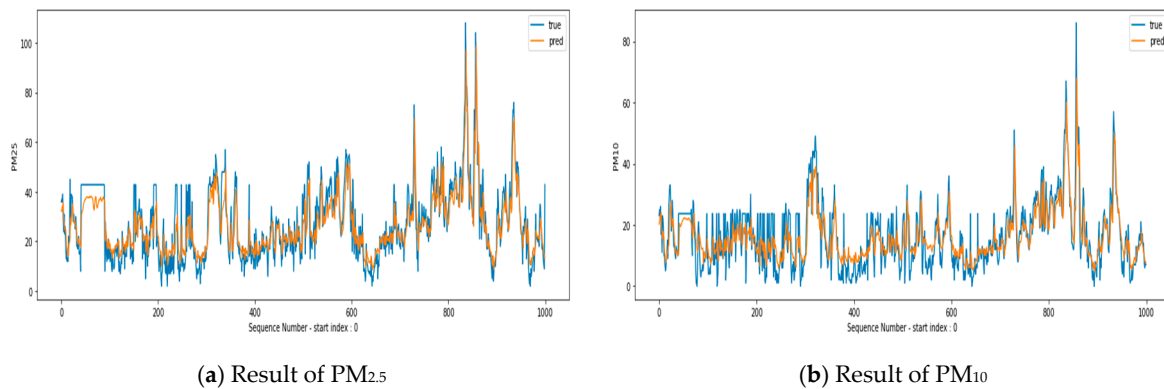


Figure 9. A graph of the actual and predicted values of PM₁₀ and PM_{2.5} using weather data with CO.

(a) Predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

4.5. Comparison of PM Prediction Accuracy According to NO₂

We analyzed the correlation between NO₂ and PM by comparing weather data and weather data with NO₂. As a result, it was confirmed that RMSE and MAE of PM_{2.5} and PM₁₀ were lowered (Table 14), and hence NO₂ is related to PM_{2.5} and PM₁₀. The actual and predicted values of PM₁₀ and PM_{2.5} were compared (Figure 10). The orange line is the predicted value, and the blue line is the actual value. We confirmed that the accuracy of PM prediction is improved via NO₂.

Table 14. Evaluation of the predicted PM using weather data with NO₂.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
10.997660334582932	7.11836606573322	6.981789199218698	4.811013752937683

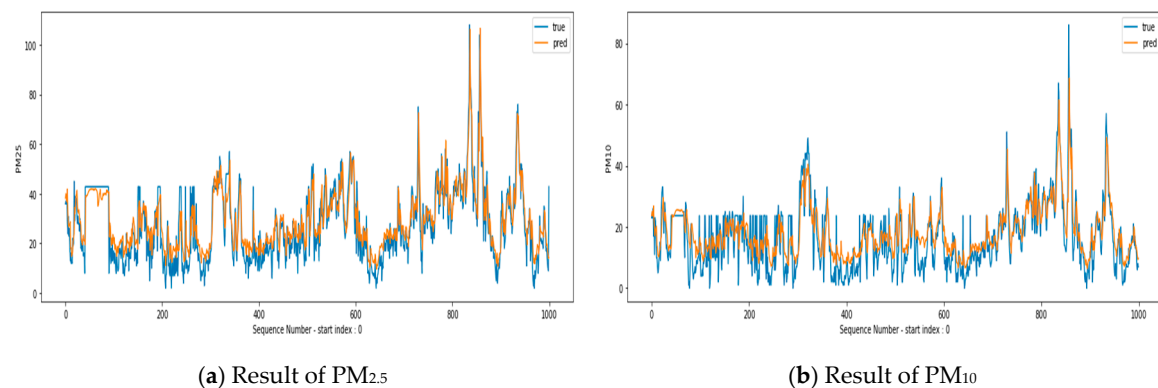


Figure 10. A graph of the actual and predicted values of PM₁₀ and PM_{2.5} using weather data with NO₂:

(a) predicted and actual values of PM_{2.5}, and (b) predicted and actual values of PM₁₀.

4.6. Comparison of PM Prediction Accuracy According to O₃

We analyzed the correlation between O₃ and PM by comparing weather data and weather data with O₃, and found that RMSE and MAE of PM₁₀ were lowered (Table 15), MAE of PM_{2.5} was higher (Table 15), and hence O₃ is related to PM₁₀. The actual and predicted values of PM₁₀ and PM_{2.5} were compared (Figure 11). The orange line is the predicted value, and the blue line is the actual value. We confirmed that the accuracy of PM₁₀ prediction is improved via O₃.

Table 15. Evaluation of the predicted PM using weather data with O₃.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
11.609210552459992	7.574959231025852	6.802652715247799	4.745732709395488

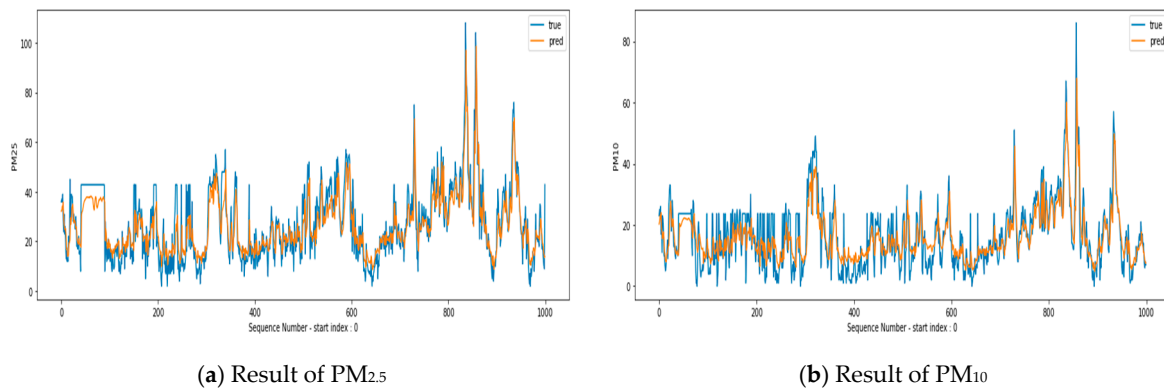


Figure 11. A graph of the actual and predicted values of PM₁₀ and PM_{2.5} using weather data with O₃. (a) predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

4.7. Comparison of PM Prediction Accuracy According to Atmospheric Pressure

We analyzed the correlation between atmospheric pressure and PM by comparing weather data and weather data with atmospheric pressure. As a result, it was confirmed that RMSE and MAE of PM₁₀ were lowered (Table 16), RMSE and MAE of PM_{2.5} were higher (Table 16), and hence atmospheric pressure is related to PM₁₀. The actual and predicted values of PM₁₀ and PM_{2.5} were compared (Figure 12). The orange line is the predicted value, and the blue line is the actual value. We confirmed that the accuracy of PM₁₀ prediction is improved via atmospheric pressure.

Table 16. Evaluation of the predicted PM using weather data with atmospheric pressure.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
11.418551994654374	7.6519155502608305	7.099408392719633	5.154189156530305

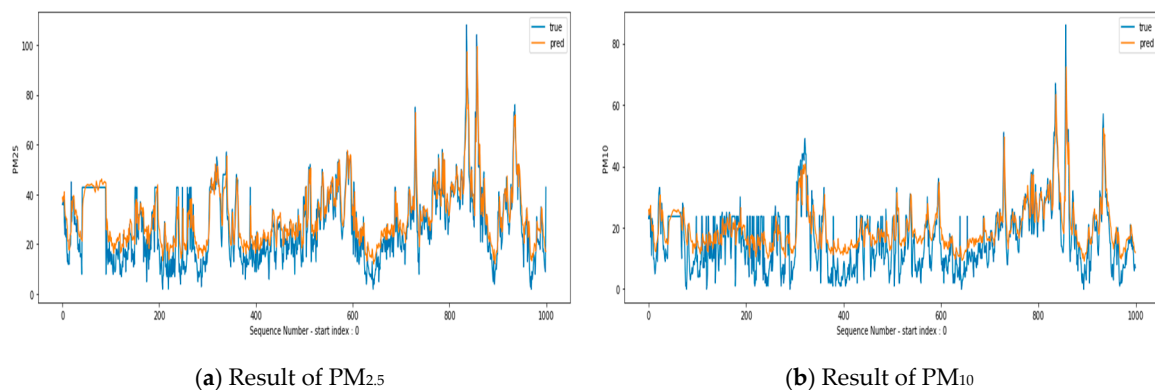


Figure 12. A graph of the actual and predicted values of PM₁₀ and PM_{2.5} using weather data with atmospheric pressure. (a) Predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

4.8. Comparison of PM Prediction Accuracy According to Wind_x, Wind_y

We confirmed that the accuracy was improved through Equations (1) and (2) discussed in Section 2.3.5. After removing wind speed and wind direction, we evaluated the performance by adding wind_x and wind_y (Table 17). We confirmed that the accuracy of PM prediction is improved via wind_x and wind_y (Figure 13).

Table 17. Evaluation of the predicted PM using weather data with wind_x and wind_y.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
11.2317932705464	7.316346260707396	6.877921416196969	4.751750217462248

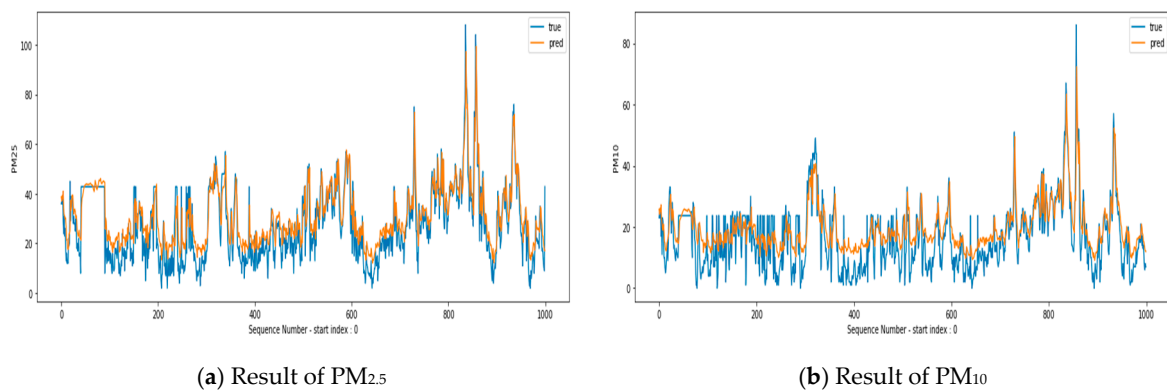


Figure 13. This figure is a graph of the actual and predicted values of PM₁₀ and PM_{2.5} using weather data with wind_x and wind_y: (a) predicted and actual values of PM_{2.5}, and (b) predicted and actual values of PM₁₀.

4.9. Comparison of PM Prediction Accuracy According to Address Format

We used the address format as discussed in the Section 2.3.3. The results were compared with Table 9 and confirmed to be improved (Table 18), and hence, the city and road are separated and affected by adjacent cities and provinces. These are the actual and predicted values of PM₁₀ and PM_{2.5} (Figure 14). The orange line is the predicted value, and the blue line is the actual value. We confirmed that the accuracy of PM prediction is improved via address format.

Table 18. Evaluation of the predicted PM using weather data with the address format.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
11.188162098223694	7.312010079856926	7.047761546009768	4.76084091141095

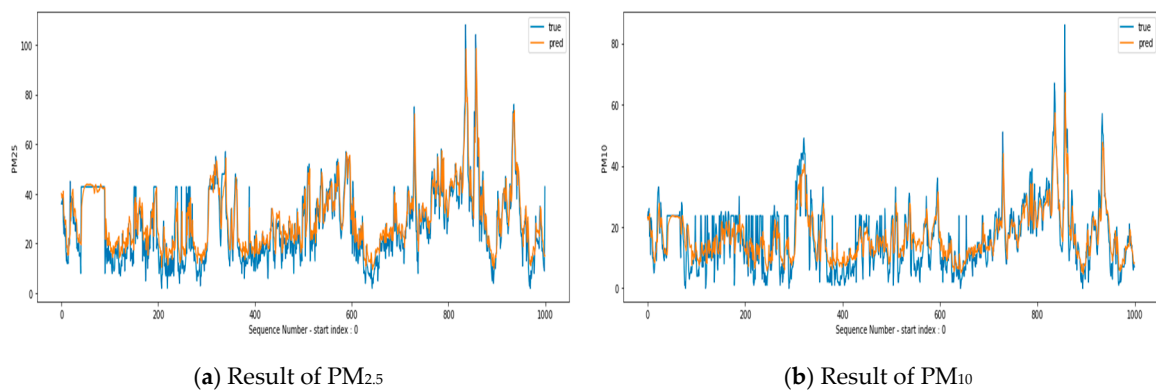


Figure 14. A graph of the actual and predicted values of PM₁₀ and PM_{2.5}. (a) Predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

4.10. Comparison of PM Prediction Accuracy According to 24 Solar Terms

We used the 24 solar terms as discussed in Section 2.3.4. The results were compared with Table 7 and confirmed to be improved (Table 19). We confirmed that the accuracy of PM prediction is improved via 24 solar terms (Figure 15).

Table 19. Evaluation of the predicted PM using weather data with 24 solar terms.

PM _{2.5} RMSE	PM _{2.5} MAE	PM ₁₀ RMSE	PM ₁₀ MAE
11.378345949349278	7.2755643802599455	6.75767487545403	4.663318307737441

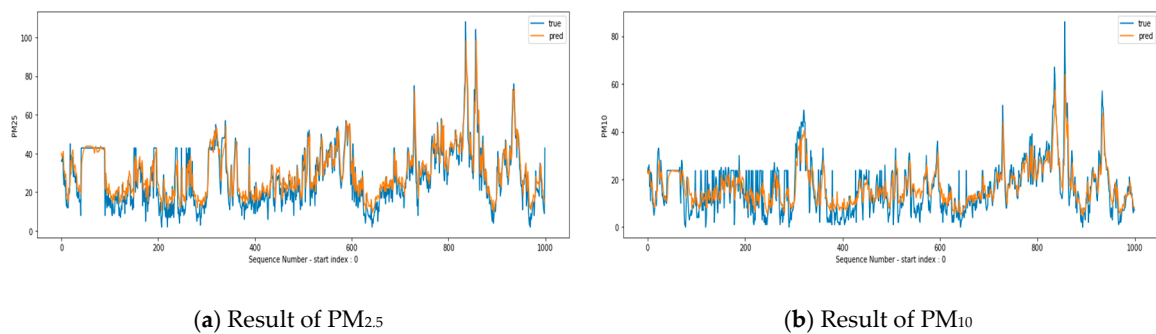


Figure 15. A graph of the actual and predicted values of PM₁₀ and PM_{2.5}. (a) Predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

4.11. Comparison of PM Prediction Accuracy According to Sequence Length

We evaluated the performance according to the sequence length with the data parameters in Table 5. The sequence length is 24 per day. The comparison confirmed that the best result was a sequence length of 7 days (Table 20), which generated the lowest values for PM_{2.5} MAE, PM_{2.5} RMSE, PM₁₀ MAE, and PM₁₀ RMSE. We highlighted that MAE and RMSE were reduced.

Table 20. Performance evaluation results according to the sequence length.

Sequence Length	PM _{2.5} MAE	PM _{2.5} RMSE	PM ₁₀ MAE	PM ₁₀ RMSE
1 Day	6.8723638736498796	10.872907681846236	4.889641018450285	7.023668509637945
2 Days	7.07283570669722	10.922789343087404	4.752256456662703	6.864557932586567
3 Days	8.010269844202515	11.565469091024756	5.419169106578719	7.81227016304491
4 Days	7.193962617691575	11.139297312830582	5.014263292572295	7.125953867868762
5 Days	7.0712063786053045	10.885098257926721	4.683081878761333	6.8245960495513325
6 Days	6.950957594294552	10.877398546601832	4.8897148765323655	7.075202116985077
7 Days	6.748574488524663	10.666811784105274	4.576222601673498	6.706854128280908
8 Days	8.311487394638375	12.524398614693366	6.037762553211165	9.065500584144592
9 Days	7.45985392084471	11.087943511761466	4.869146536925853	6.957451893754408
10 Days	7.068727429495773	10.9847680613304	5.32557239845155	8.05212499374417
11 Days	7.486351519908187	11.437778947196147	4.839375773430798	7.135375508314753
12 Days	7.060299407943606	11.125659651922488	6.116040476671633	9.085695877353128
13 Days	7.35881634959586	11.374645196090356	6.7099764973762905	10.786191153184818
14 Days	7.057825475352884	10.854893642056501	4.967073756780628	6.993131797131667

4.12. Comparison of PM Prediction Accuracy According to Batch Size

We evaluated the performance according to the batch size and found the best result with a batch size of 96 (Table 21). We highlighted that MAE and RMSE were reduced.

Table 21. Performance evaluation results according to the batch size.

Batch Size	PM _{2.5} MAE	PM _{2.5} RMSE	PM ₁₀ MAE	PM ₁₀ RMSE
64	6.8723638736498796	10.872907681846236	4.889641018450285	7.023668509637945
96	6.915248555917008	10.872228585931706	4.5699693535437085	6.6472971974889905
128	7.290778784389358	11.358100746070686	5.090586659374237	7.652284409114551
160	7.195339101694315	10.9563082562599	4.820049060024904	6.8492290512882485
192	7.0080317971032	10.925634792582278	5.250878940846936	7.90720787936262
256	7.627937784358025	11.255332451776745	4.7917052769598865	6.9397277896506875
512	6.833282805161927	10.740158388237752	4.691628873916771	6.85637822970579

4.13. Compare of LSTM Model, CNN Model, and GRU Model

Based on the experimental results, we evaluated the performance with the LSTM and CNN models, the GRU model with a sequence length of 7 days, and a batch size of 96 (Table 22). As a result, the LSTM model appeared to be highly accurate in the training data (Figure 16), but the test data confirmed that the CNN model did not show similarity (Figure 17). The gray line is the actual value and the red line is the prediction value of the GRU model. The green line is the CNN prediction value, and the blue line is the LSTM model prediction value. We highlighted that lowest MAE and RMSE.

Table 22. Performance evaluation result according to each model.

Batch Size	PM _{2.5} MAE	PM _{2.5} RMSE	PM ₁₀ MAE	PM ₁₀ RMSE
CNN	10.21723294956235	13.501415151238904	9.280525187180793	13.293511747301356
LSTM	7.201965056645242	10.964201214885806	4.672849256285137	6.92387169617792
GRU	7.796999000699376	11.986331910014759	6.552875746664709	9.338689775858853

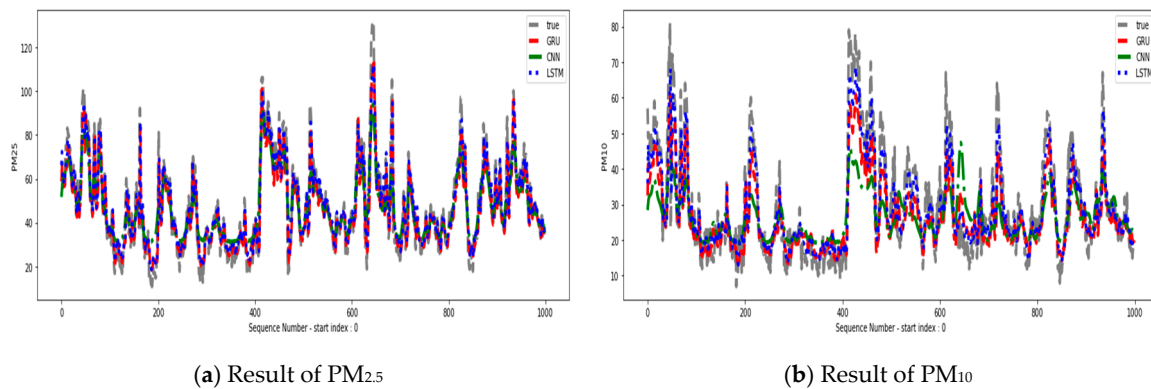


Figure 16. A graph of actual and predicted values of PM₁₀ and PM_{2.5} using training data. (a) Predicted and actual values of PM_{2.5}; (b) predicted and actual values of PM₁₀.

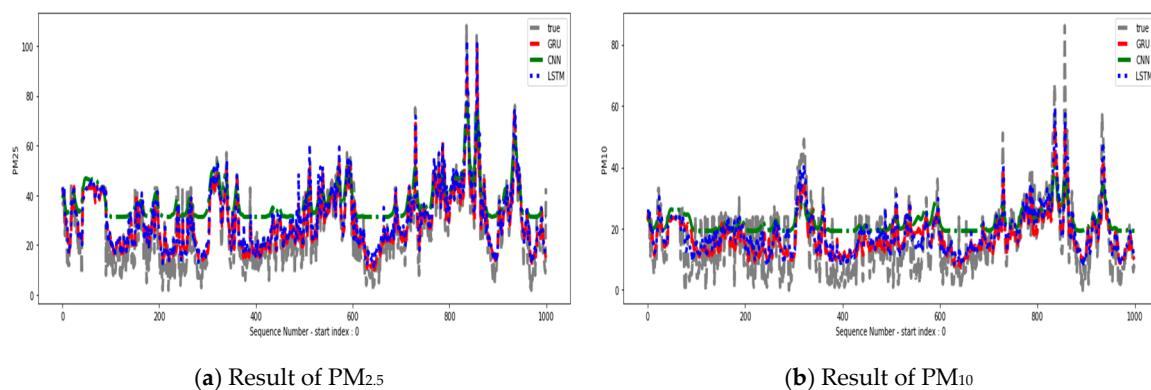


Figure 17. A graph of actual and predicted values of PM₁₀ and PM_{2.5} using test data: (a) predicted and actual values of PM_{2.5}, (b) predicted and actual values of PM₁₀.

The difference between the actual and predicted values in each model was confirmed to differ from the training data and the test data. In particular, the training data of the CNN model had some similarities, but the test data could not be predicted. We confirmed that the highest accuracy was obtained using the LSTM model (Table 22).

5. Conclusions

We analyzed other PM prediction models. We analyzed the correlation between address format, air quality data, and weather data, and PM. We developed a PM prediction model through the 24

solar terms provided by the Korea Meteorological Administration, weather data, and air quality data provided by AirKorea. Our paper makes the following key contributions to the literature. (1) It was confirmed that the address format improves the accuracy when developing prediction models that are affected by regions such as PM and weather. (2) When developing a forecasting model that is affected by seasons such as PM and weather, it was confirmed that accuracy is improved when 24 solar terms are used. (3) It was confirmed that the air quality data improves the accuracy when developing the PM prediction model. (4) It was confirmed that the atmospheric pressure improves the accuracy when developing the PM prediction model. For future research, the authors will apply hybrid models, such as convolutional neural networks and recurrent neural networks. For future research, the authors will apply visualization such as a map. For future research, the authors will apply hybrid models of the LSTM model and other deep learning models (e.g., RNN, CNN, GRU, DAE, Q-Networks) to improve the accuracy of PM prediction.

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

Funding: This research supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2015-0-00403) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and by Soonchunhyang Research Fund.

Acknowledgments: We appreciate the air quality indices data provided by the Korea Environment Corporation (<http://www.airkorea.or.kr/>) and the weather data provided by the Korea Meteorological Administration (<http://www.kma.go.kr/>). This manuscript was proofread and edited by the professional English editors at HARRISCO (<http://en.harrisco.net/>).

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

References

1. Park, D.U.; Kwon, C.H. Characteristics of PM₁₀, PM_{2.5}, CO₂ and CO monitored in interiors and platforms of subway train in Seoul, Korea. *Environ. Int.* **2008**, *34*, 629–634. [[CrossRef](#)] [[PubMed](#)]
2. Ten Threats to Global Health in 2019. Available online: <https://www.who.int/news-room/feature-stories/ten-threats-to-global-health-in-2019> (accessed on 3 February 2020).
3. Ebi, K.L.; McGregor, G. Climate change, tropospheric ozone and particulate matter, and health impacts. *Environ. Health Perspect.* **2008**, *116*, 1449–1455. [[CrossRef](#)] [[PubMed](#)]
4. Samet, J.; Buist, S.; Bascom, R.; Garcia, J.; Lipsett, M.; Mauderly, J.; Mannino, D.; Rand, C.; Romieu, I.; Utell, M.; et al. What constitutes an adverse health effect of air pollution. *Am. J. Respir. Crit. Care Med.* **2000**, *161*, 665–673. [[CrossRef](#)]
5. Mannucci, P.M.; Harari, S.; Martinelli, I.; Franchini, M. Effects on health of air pollution: A narrative review. *Intern. Emerg. Med.* **2015**, *10*, 657–662. [[CrossRef](#)]
6. Janssen, N.A.H.; Fischer, P.; Marra, M.; Ameling, C.; Cassee, F.R. Short-term effects of PM_{2.5}, PM₁₀ and PM_{2.5–10} on daily mortality in the Netherlands. *Sci. Total Environ.* **2013**, *463*, 20–26. [[CrossRef](#)] [[PubMed](#)]
7. Gauderman, W.J.; Avol, E.; Gilliland, F.; Vora, H.; Thomas, D.; Berhane, K.; McConnell, R.; Kuenzli, N.; Lurmann, F.; Rappaport, E.; et al. The effect of air pollution on lung development from 10 to 18 years of age. *N. Engl. J. Med.* **2004**, *351*, 1057–1067. [[CrossRef](#)] [[PubMed](#)]
8. Terada, N.; Maesako, K.I.; Hiruma, K.; Hamano, N.; Houki, G.; Konno, A.; Ikeda, T.; Sai, M. Diesel exhaust particulates enhance eosinophil adhesion to nasal epithelial cells and cause degranulation. *Int. Arch. Allergy Immunol.* **1997**, *114*, 167–174. [[CrossRef](#)]
9. Gent, J.F.; Triche, E.W.; Holford, T.R.; Belanger, K.; Bracken, M.B.; Beckett, W.S.; Leaderer, B.P. Association of low-level ozone and fine particles with respiratory symptoms in children with asthma. *JAMA* **2003**, *290*, 1859–1867. [[CrossRef](#)]
10. Holz, O.; Mücke, M.; Paasch, K.; Böhme, S.; Timm, P.; Richter, K.; Magnussen, H.; Jörres, R.A. Repeated ozone exposures enhance bronchial allergen responses in subjects with rhinitis or asthma. *Clin. Exp. Allergy* **2002**, *32*, 681–689. [[CrossRef](#)]

11. McDonnell, W.F.; Abbey, D.E.; Nishino, N.; Lebowitz, M.D. Long-term ambient ozone concentration and the incidence of asthma in nonsmoking adults: The AHSMOG Study. *Environ. Res.* **1999**, *80*, 110–121. [\[CrossRef\]](#)
12. D'amato, G.; Liccardi, G.; D'Amato, M. Environmental risk factors (outdoor air pollution and climatic changes) and increased trend of respiratory allergy. *J. Investig. Allergol. Clin. Immunol.* **2000**, *10*, 123–128. [\[PubMed\]](#)
13. Salvador, P.; Artíñano, B.; Querol, X.; Alastuey, A. A combined analysis of backward trajectories and aerosol chemistry to characterise long-range transport episodes of particulate matter: The Madrid air basin, a case study. *Sci. Total Environ.* **2008**, *390*, 495–506. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Delfino, R.J.; Quintana, P.J.E.; Floro, J.; Gastañaga, V.M.; Samimi, B.S.; Kleinman, M.T.; Liu, L.J.S.; Bufalino, C.; Wu, C.F.; McLaren, C.E. Association of FEV1 in asthmatic children with personal and microenvironmental exposure to airborne particulate matter. *Environ. Health Perspect.* **2004**, *112*, 932–941. [\[CrossRef\]](#)
15. Riedl, M.; Sanchez, D.D. Biology of diesel exhaust effects on respiratory function. *J. Allergy Clin. Immunol.* **2005**, *115*, 221–228. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Behndig, A.F.; Mudway, I.S.; Brown, J.L.; Stenfors, N.; Helleday, R.; Duggan, S.T.; Wilson, S.J.; Boman, C.; Cassee, F.R.; Frew, A.J.; et al. Airway antioxidant and inflammatory responses to diesel exhaust exposure in healthy humans. *Eur. Respir. J.* **2006**, *27*, 359–365. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Song, G.J.; Moon, Y.H.; Joo, J.H.; Lee, A.Y.; Lee, J.B. Distribution of Hazardous Heavy Metal in TSP, PM10 and PM2.5 Emitted from Coal-fired Power Plants. *J. Korean Soc. Environ. Anal.* **2018**, *21*, 172–180.
18. Kim, Y.P. Air pollution in Seoul caused by aerosols. *J. Korean Soc. Atmos. Environ.* **2006**, *22*, 535–553.
19. Kim, Y.W.; Lee, H.S.; Jang, Y.J.; Lee, H.J. How does media construct particulate matter risks?: A news frame and source analysis on particulate matter risks. *Korean J. Commun. Stud.* **2015**, *59*, 121–154.
20. Oh, Y.M. Countermeasures for Fine Dust Management. The Government of the Korea: Seoul, Korea, 2013.
21. Lee, J.S. *Climate and Environment Headquarters Plan*; Deputy Mayor for Climate & Environment in the Korea: Seoul, Korea, 2015.
22. Park, S.B. *Cause of fine Dust and Response Policy Issues*; Korea Environmental Industry & Technology Institute: Seoul, Korea, 2018.
23. Lee, D.K.; Lee, S.W. Hourly Prediction of Particulate Matter (PM2.5) Concentration Using Time Series Data and Random Forest. *Kips Trans. Softw. Data Eng.* **2020**, *9*, 129–136. [\[CrossRef\]](#)
24. Yang, G.; Lee, H.M.; Lee, G.Y. A Hybrid Deep Learning Model to Forecast Particulate Matter Concentration Levels in Seoul, South Korea. *Atmosphere* **2020**, *11*, 348. [\[CrossRef\]](#)
25. Xayasouk, T.; Lee, H.M.; Lee, G.Y. Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. *Sustainability* **2020**, *12*, 2570. [\[CrossRef\]](#)
26. Yi, H.; Bui, K.H.N.; Seon, C.N. A Deep Learning LSTM Framework for Urban Traffic Flow and Fine Dust Prediction. *Korean Inst. Inf. Sci. Eng.* **2020**, *47*, 292–297.
27. Vong, C.M.; Ip, W.F.; Wong, P.K.; Chiu, C.C. Predicting minority class for suspended particulate matters level by xstream learning machine. *Neurocomputing* **2014**, *128*, 136–144. [\[CrossRef\]](#)
28. Airkorea. Available online: <http://www.airkorea.or.kr> (accessed on 3 February 2020).
29. Home-Korea Meteorological Administration. Available online: <http://www.kma.go.kr/> (accessed on 3 February 2020).
30. Papanastasiou, D.K.; Melas, D.; Kioutsioukis, I. Development and assessment of neural network and multiple regression models in order to predict PM10 levels in a medium-sized Mediterranean city. *Water Air Soil Pollut.* **2007**, *182*, 325–334. [\[CrossRef\]](#)
31. Tai, A.P.K.; Mickley, L.J.; Jacob, D.J. Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *Atmos. Environ.* **2010**, *44*, 3976–3984. [\[CrossRef\]](#)
32. Yoo, Y.M.; Lee, Y.R.; Kim, D.; Jeong, M.J.; Stockwell, W.R.; Kundu, P.K.; Oh, S.M.; Shin, D.B.; Lee, S.J. New indices for wet scavenging of air pollutants (O₃, CO, NO₂, SO₂, and PM10) by summertime rain. *Atmos. Environ.* **2014**, *82*, 226–237. [\[CrossRef\]](#)
33. Yuxuan, W.; Brendon, M.M.; William, M.J.; Jiming, H.; Hong, M.; Chris, N.; Yaosheng, C. Variations of O₃ and CO in Summertime at a Rural Site Near Beijing. *Atmos. Chem. Phys.* **2008**, *8*, 6355–6363. [\[CrossRef\]](#)
34. Grivas, G.; Chaloulakou, A. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmos. Environ.* **2006**, *40*, 1216–1229. [\[CrossRef\]](#)

35. Xie, Y.; Zhao, B.; Zhang, L.; Luo, R. Spatiotemporal variations of PM_{2.5} and PM₁₀ concentrations between 31 Chinese cities and their relationships with SO₂, NO₂, CO and O₃. *Particuology* **2015**, *20*, 141–149. [[CrossRef](#)]
36. Sayegh, A.S.; Munir, S.; Habeebullah, T.M. Comparing the performance of statistical models for predicting PM₁₀ concentrations. *Aerosol Air Qual. Res.* **2014**, *14*, 653–665. [[CrossRef](#)]
37. Xie, J.; Hong, T. Load forecasting using 24 solar terms. *J. Mod. Power Syst. Clean Energy* **2018**, *6*, 208–214. [[CrossRef](#)]
38. Weather Data Open Portal [Climate Statistics Analysis: Meteorological Events: 24 Solar Terms]. Available online: <https://data.kma.go.kr/climate/solarTerms/solarTerms.do> (accessed on 24 February 2020).
39. Feng, X.; Li, Q.; Zhu, Y.; Hou, J.; Jin, L.; Wang, J. Artificial neural networks forecasting of PM_{2.5} pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos. Environ.* **2015**, *107*, 118–128. [[CrossRef](#)]
40. Chae, M.S.; Han, S.W.; Lee, H.W. CNN and LSTM models for predicting particulate matter in the Korea. In Proceedings of the 8th International Conference on Information, System and Convergence Applications, Ho Chi Minh, Vietnam, 12–15 February 2020.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).