경제학석사 학위논문

Applying PCA to Deep Learning Forecasting Models for Predicting Concentration of Fine Particulate Matter ($PM_{2.5}$)

딥러닝 시계열 알고리즘 기반 초미세먼지 예측 모델의 주성분분석 적용

2021년 8월

서울대학교 대학원

농경제사회학부 지역정보학전공

최상원

Applying PCA to Deep Learning Forecasting Models for Predicting Concentration of Fine Particulate Matter ($PM_{2.5}$)

딥러닝 시계열 알고리즘 기반 초미세먼지 예측 모델의 주성분분석 적용

지도 교수 Hong Sok(Brian) Kim
이 논문을 경제학석사 학위논문으로 제출함
2021년 6월

서울대학교 대학원
농경제사회학부 지역정보학전공
최 상 원

최상원의 경제학석사 학위논문을 인준함
2021년 6월

위 원 장 김 의 준

부위원장 HONG SOK(BRIAN) KIM

위 원 유 도 일

Abstract

Applying PCA to Deep Learning Forecasting Models for Predicting Concentration of Fine Particulate Matter($PM_{2.5}$)

SangWon Choi
Program in Regional Information Studies
Dept. of Agricultural Economics and
Rural Development
The Graduate School
Seoul National University

Fine Particulate Matter($PM_{2.5}$) is a global air pollution problem that many metropolitan cities are experiencing. A $PM_{2.5}$ concentration of one country is influenced by not only internal but also external factors such as air quality of adjacent countries. Hence, data of both a country of interest and its surrounding countries are needed in order to estimate measures needed to design mitigation strategies and policies. However, there is a chance for 'curse of dimensionality' which occurs when there are more variables than observations in a data set; thereby, reducing the predictive power.

This study aims to estimate the daily $PM_{2.5}$ concentration in eight major cities in South Korea using deep learning time-series models. To do so, it uses each city's air quality, meteorological factors, and $PM_{2.5}$ concentration along with adjacent Chinese cities' $PM_{2.5}$ concentration in the course of five years. Here, PCA was applied in order to prevent 'curse of dimensionality', as mentioned earlier. In estimating $PM_{2.5}$ concentration, time series models such as RNN, LSTM, BiLSTM were used. By dividing the model' s function into ones with PCA and ones without PCA, RMSE and MAE were reflected for a better comparison.

As a result, the overall performance of both LSTM and BiLSTM was better after the application of PCA. The performance of LSTM with PCA was higher than that without PCA by up to 16.6% and 33.3% in terms of RMSE and MAE, respectively. Similarly, BiLSTM with PCA outperformed that without PCA by up to 16.7% and 31.6% in terms of RMSE and MAE, respectively. Hence, it can be inferred that the application of PCA enhances the function of deep learning time series models and provides a more accurate estimation for designing a better mitigation policy.

Keywords : Principal Components Analysis(PCA), Fine Particulate Matter($PM_{2.5}$), Recurrent Neural Network(RNN), Long Short-Term Memory(LSTM), Bidirectional

LSTM(BiLSTM)
Student Number: 2019-26828

CONTENTS

List of Tables

List of Figures

Chapter 1. Introduction

Fine particulate matter($PM_{2.5}$) indicates particles with an aerodynamic diameter of 2.5 μm or less. It is not a single chemical compound, such as sulfur oxide(SO_x) or nitrogen oxide (NO_x), but a mixture of particles of varying sizes, components, and shapes. For example, typical substances that form $PM_{2.5}$ include elemental carbon (EC), organic carbon (OC), NO_x, volatile organic compounds (VOC), ozone (O_3), ammonia (NH_3), SO_x, condensate particles, metal particles, mineral particles, etc. Because of its small size, it can easily penetrate into the human body through the respiratory tract, causing inflammation or damaging organs (Gong, 2012). Their potential health effects can be detrimental; thus, the WHO considers $PM_{2.5}$ as a major environmental risk factor that causes cardiovascular and respiratory diseases, and various types of cancer. Figure 1 shows the potential effects of $PM_{2.5}$ on human body.

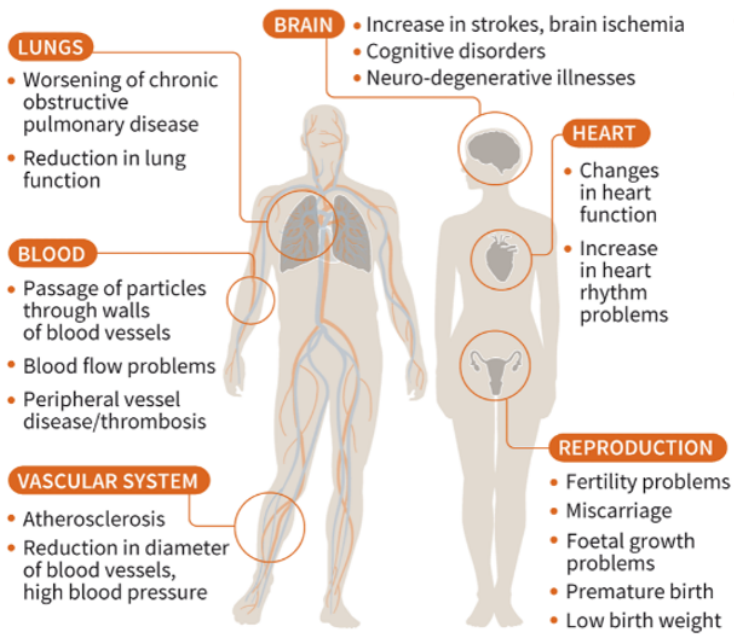


Figure 1 Effects of fine particulate matter on human body
Source: French National Health Agency,
InVS(Institut de veille sanitaire),

European Environment Agency and AFP

South Korea's $PM_{2.5}$ concentration was the highest among the 37 OECD(Organization for Economic Cooperation and Development) countries in 2019. On a related note, some studies have shown that it has a negative effect on human health. Han et al. (2018) stated that 1763 early deaths cases in Seoul in 2015 were closely related to $PM_{2.5}$. Hwang & Son (2018) found that, when the average annual concentration of $PM_{2.5}$ in Seoul increases by $10\ \mu g/m^3$, the risk of death of the elderly population(over 65 years) from cardiovascular and respiratory diseases (ischemic heart disease, chronic obstructive lung disease, lung cancer, and cerebrovascular diseases) increases by 13.9%. This aligns with the major causes of death for South Koreans in 2019. In addition, Statistics Korea(2020) showed that cancer (158.2 deaths per 100,000 people), cardiovascular diseases (60.4 deaths per 100,000 people), and pneumonia (45.1 deaths per 100,000 people) are the three major causes of death in 2019. This suggests that $PM_{2.5}$ concentration is highly correlated to the main cause of death for Koreans.

The South Korean government is making great efforts to reduce $PM_{2.5}$ concentration to protect people's health. The government has divided the crisis into three stages based on the current concentration level and the predicted level of $PM_{2.5}$ concentration and has devised a strategic manual for local governments for each stage. The government also aims to reduce the annual average concentration of $PM_{2.5}$ by 35% compared to 2016 by establishing a 'five-year plan' for $PM_{2.5}$ concentration reduction. To achieve this goal, the government has selected 15 major tasks by evaluating its potential reduction amount, cost effectiveness, linkage with other policies, and social impact and these tasks are implemented by each local government (Joint association of related Korean ministries

of Korea, 2019).

Table 1 shows Korea's crisis stage standard for $PM_{2.5}$ concentration, which reflects the concentration of $PM_{2.5}$ of the current period and future forecast values. It suggests that the accurate prediction of $PM_{2.5}$ concentration is needed in both short and long terms. In this regard, several studies have conducted an air quality prediction using deep learning methods with Korea's domestic data (wind speed, NO_2, SO_2, temperature, etc.), and new deep learning models have been developed to show higher performance in air quality prediction(Xayasouk, T. et al., 2020, Mengara, A.M. et al., 2020). However, foreign factors should also be considered in predicting $PM_{2.5}$ concentration in South Korea, as the concentration of $PM_{2.5}$ in the Shandong region of China was also found to affect South Korea's $PM_{2.5}$ concentration (Park & Shin, 2017). However, China's past $PM_{2.5}$ concentration was not recorded on an hourly basis. Therefore, Korea's daily data were used instead of hourly data for deep learning $PM_{2.5}$ prediction in order to match units of observations. Here, it should be noted that this data composition can cause a "curse of dimensionality" due to the small number of observations compared to variables, which can reduce the performance of the model.

Table 1 Crisis stage standard.

Crisis Stages	Criteria	Main Contents
Stage 1	150 μg/m³ for 2 h or longer +75 μg/m³ for the following day	Strengthening the current system
Stage 2	200 μg/m³ for 2 h or longer +150 μg/m³ for the following day	Strengthening public sector measures
Stage 3	400 μg/m³ for 2 h or longer +200 μg/m³ for the following day	Strengthening private sector measures /disaster response

Thus, this study aims to show that applying principal

component analysis to the deep learning time series models (RNN, LSTM, BiLSTM) for predicting $PM_{2.5}$ concentration can produce better performance by comparing the root mean square error (RMSE) and mean absolute error (MAE) with the same models with and without PCA application.;

Chapter 2. Literature Review

In this chapter, a series of literature review on health effects of $PM_{2.5}$ and prediction methodologies of $PM_{2.5}$ concentration is discussed to assist further understanding in this study. After careful examination of prior researches, this study chose deep learning models for predicting $PM_{2.5}$ concentration, which will also be discussed in this chapter.

Section 1. The associated diseases with $PM_{2.5}$

Several studies have shown the association between $PM_{2.5}$ and lung and cardiovascular disease (CVD). Wang et al. (2016) reported that CVD is the one of the main mortality factors of elderly population. It was found that the ambient $PM_{2.5}$ concentration is related to several CVDs by linking $PM_{2.5}$ exposure and CVD based on multiple pathophysiological mechanisms. César et al. (2016) showed that the exposure to $PM_{2.5}$ can cause hospitalizations for pneumonia and asthma in children younger than 10 years old through an ecological study of time series and a generalized additive model of Poisson regression. Kim et al. (2020) reported associations of short-term $PM_{2.5}$ exposure with acute upper respiratory infection and bronchitis among children of 0-4 years old through a difference-in-differences approach generalized to multiple spatial units (regions) and time periods (day) with distributed lag non-linear models. Vinikoor-Imler et

al. (2011) studied the relationship between $PM_{2.5}$concentration, lung cancer incidence, and mortality by linear regression and concluded that there is a possibility of an association between them. Choe et al. (2015) reported that the effect of changes in $PM_{2.5}$ emissions on changes in internal visits and hospitalization probabilities due to respiratory diseases was estimated through Probit and Tobit models. If $PM_{2.5}$ emissions change by 1%, the probability of visitation due to respiratory diseases increases from 0.755% to 1.216%, and the probability of hospitalization increases from 0.150% to 0.197%.

Section 2. Prediction of $\mathbf{PM_x}$ concentration

The need for prediction of PM_x concentration research is emerging, and various studies are underway on prediction of PM_x concentration. Ross et al. (2007) developed the land use regression model to predict $PM_{2.5}$ concentration in New York City and showed that urbanization factors such as traffic volume and population density have a high explanatory power in predicting $PM_{2.5}$ concentration. Rob Beelen et al. (2009) compared the performance of ordinary kriging, universal kriging, and regression mapping in developing EU-wide maps of air pollution and showed that universal kriging performs better in mapping NO_2, PM_{10}, and O_3. In addition, Vikas Singh et al. (2011) suggested a cokriging based approach and interpolated PM_{10} concentration in areas not observed in the network in PM_{10} concentration monitoring based on the suggested method with secondary variable from the results of a deterministic chemical transport model (CTM) simulation. Eventually, the results showed that the proposed method provides flexibility in collecting $PM_{2.5}$ data. Furthermore, several academic

discussions have taken place in time series prediction. For instance, Zhang et al (2018) showed the seasonality and prediction range of $PM_{2.5}$ concentration in Fuzhou, China through the Autoregressive Integrated Moving Average (ARIMA). On the other hand, Pozza et al (2010) used Seasonal Autoregressive Integrated Moving Average (SARIMA) and Winter models to predict $PM_{2.5}$ concentration in Sao Carlos, Brazil. Another example is Thaweephol & Wiwatwattana's prediction of $PM_{2.5}$ concentration in Bangkok(2019). This study compared the performance of Autoregressive Integrated Moving Average with Exogenous Regressor (SARIMAX) and LSTM, a deep learning time series model. Here, it was proven that higher prediction was performed by LSTM than that was done by SARIMAX. Overall, it is suggested that deep learning time series models can perform better than conventional time series analysis models can. In addition, it can be inferred that utilization of deep learning models can be further enhanced in future time series predictions, in which high accuracy is important.

Section 3. Deep learning & machine learning prediction of $\mathbf{PM_{2.5}}$ concentration

Other studies have shown examples of predicting $PM_{2.5}$ concentration through machine learning and deep learning. Zhao et al. (2019) predicted the $PM_{2.5}$ contamination of stations in Beijing using long short-term memory-fully connected (LSTM-FC), LSTM, and an artificial neural network (ANN) with historical air quality data, meteorological data, weather forecast data, and the day of the week data. They showed that the LSTM-FC model outperforms LSTM and the ANN, with MAE = 23.97-50.13 and RMSE = 35.82-69.84 over 48 h.

Karimian et al. (2019) also predicted Tehran's $PM_{2.5}$ concentration by implementing multiple additive regression trees (MARTs), a deep feedforward neural network (DFNN), and a new hybrid model DFNN-LSTM with meteorological data (temperature, surface-level pressure, relative humidity, etc.). The best model in this research was DFNN-LSTM in 12, 24, and 48 h prediction, with RMSE = 7.03-11.73 $\mu g/m^3$, and MAE = 5.59-8.41$\mu g/m^3$. Moreover, Qadeer et al. (2020) used XGBoost (XGB), the light gradient boosting machine (LGBM), the gated recurrent unit (GRU), convolutional neural network-LSTM (CNNLSTM), BiLSTM, and LSTM to predict $PM_{2.5}$ concentration of eight sites in Seoul and Gwangju with community multiscale air quality (CMAQ) data. The result showed that LSTM performs the best, with MAE = 3.5847 $\mu g/m^3$, RMSE = 4.8292 $\mu g/m^3$, R = 0.8989, and IA = 0.9368 of the mean in all sites.

After thorough scrutinization of previous studies, RNN, LSTM, and BiLSTM models were chosen for this particular study, because previous studies have shown that the deep learning sequence model performs better in prediction. Thus, the preceding daily local meteorological data, air quality data and $PM_{2.5}$ concentration data were used as input variables to predict $PM_{2.5}$ concentration of succeeding day, as shown in previous studies, and used as predictive input variable. The regional $PM_{2.5}$ concentration of China was also used as predictive input variable, which was found to affect $PM_{2.5}$ concentration in South Korea.

Chapter 3. Research Data

Section 1. Spatial area

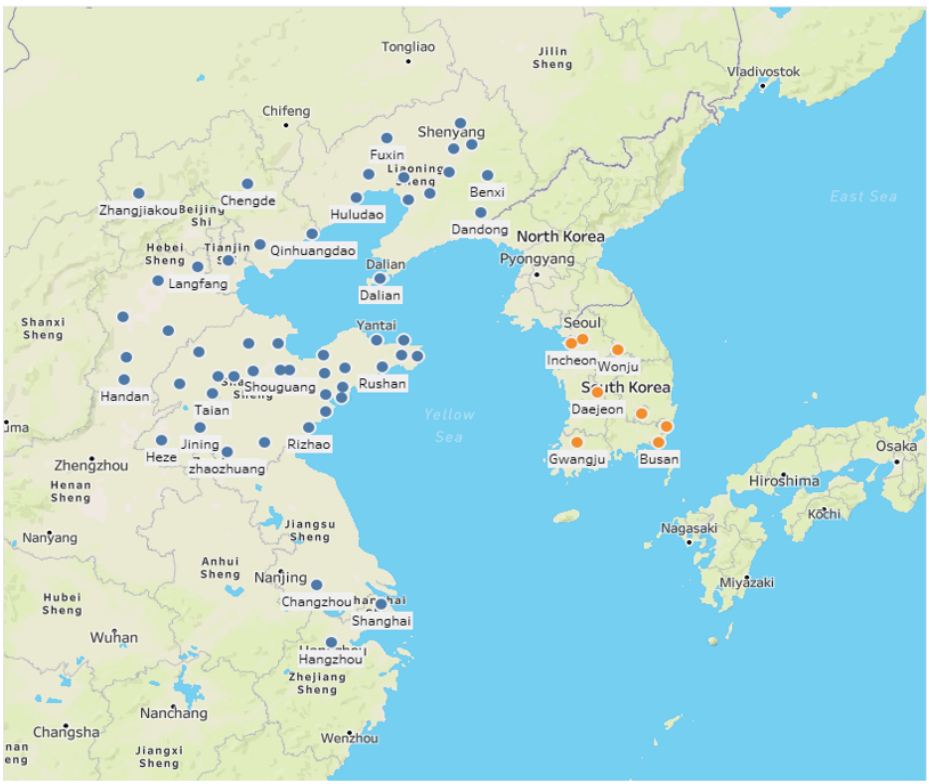


Figure 2 Spatial range of the research

Figure 2 shows the spatial range of this research. A total of eight cities in South Korea was selected for analysis. Eight selected cities are composed of six metropolitan cities (Busan, Daejeon, Daegu, Gwangju, Incheon, and Ulsan) representing each province, one capital city (Seoul), and one populous city in Kangwon province (Wonju) since Kangwon province does not have a metropolitan city. For each city, daily air quality data ($PM_{2.5}$, SO_2 , O_3 , NO_2 , and CO and meteorological data (temperature, wind speed, wind direction, humidity, precipitation, etc.) were collected from Air Korea and Korea Meteorological Agency in consideration of the internal factors of $PM_{2.5}$ formation. Air quality data were collected within 5 km radius of each city's meteorological data observatory.

Figure 3 shows that Korea is mainly influenced by north and west winds. As a result, the air quality of Korea can be directly and indirectly affected by the air quality of China, a country located in the northwestern part of Korea. For example, Figure 4, A1-A4(Nullschool) show the concentration of $PM_{2.5}$ in Korea and China before and after the outbreak of COVID-19. According to Bao et al. (2020), it can be interpreted that the lockdown of Chinese factories after the COVID-19 outbreak actually improved China's air quality. To sum up, we can see that the air quality of Korea is highly affected by the air quality of China. In order to reflect this kind of relationship between South Korea's $PM_{2.5}$ concentration and that of China, daily $PM_{2.5}$ concentration in 55 areas in China close to Korea were selected as input variables in this study. This includes the $PM_{2.5}$ concentration in Shandong province, which was found to increase $PM_{2.5}$ concentration in Korea.

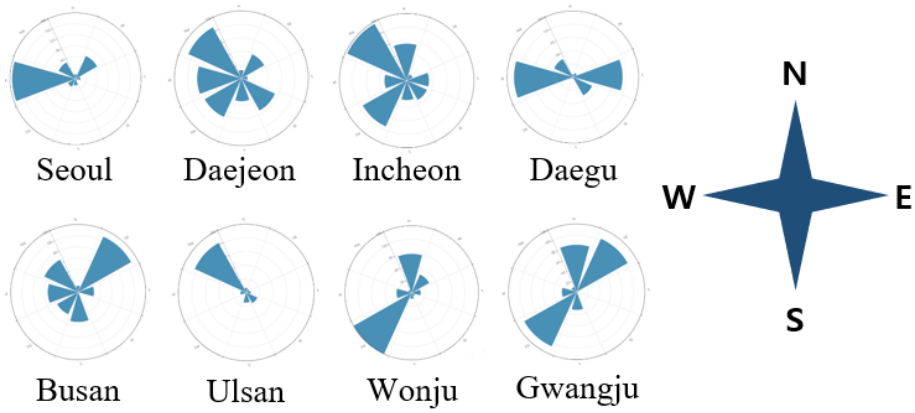
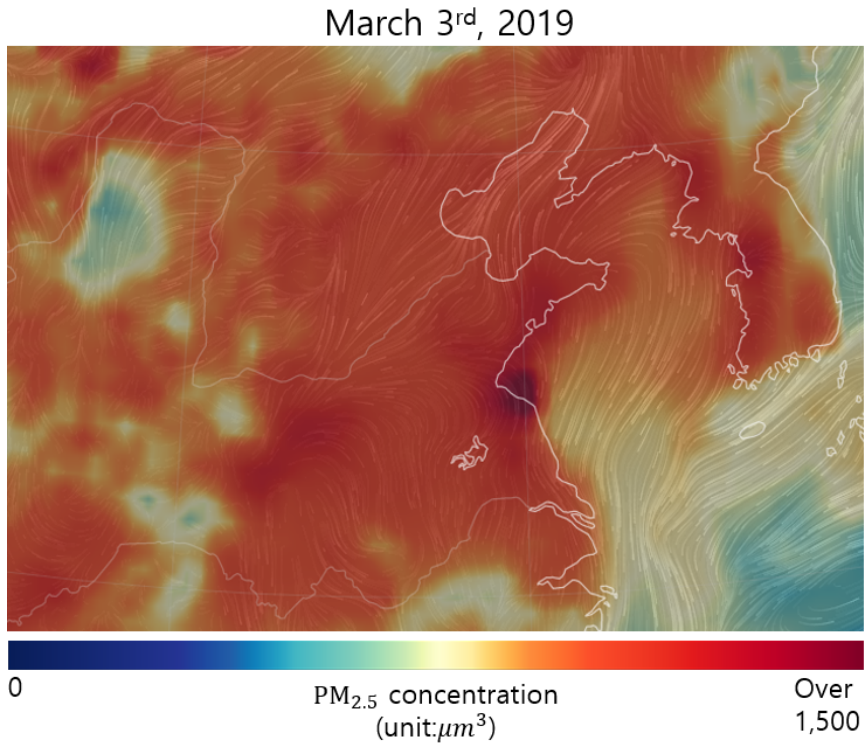


Figure 3 The wind direction frequency of Korea's selected cities in 2019



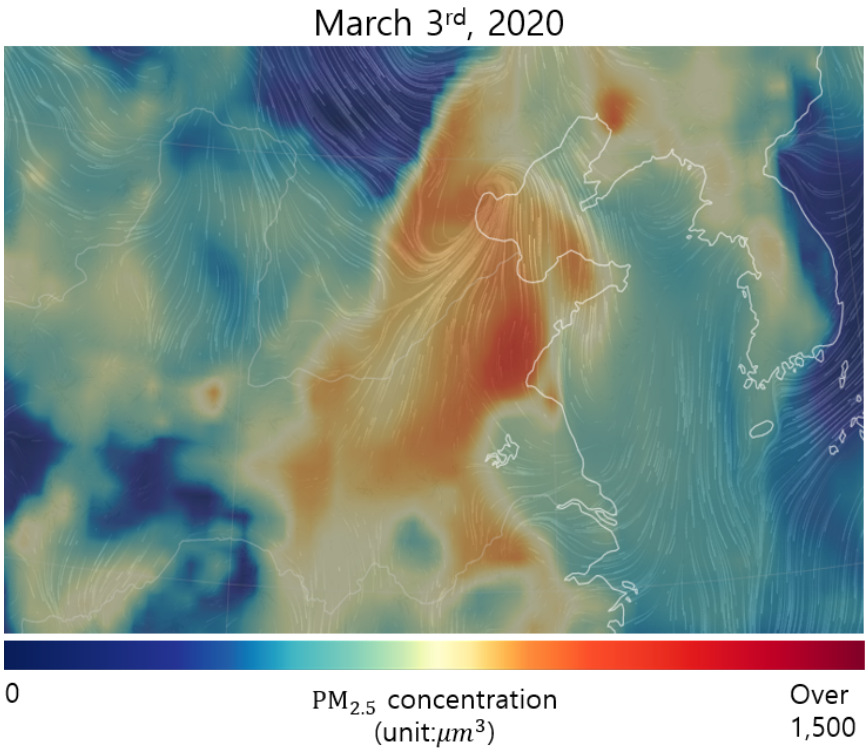


Figure 4 **$PM_{2.5}$** concentration distribution maps
(before and after the COVID-19 outbreak**:** 2019 **&** 2020**)**

Section 2. Data preprocessing

All variables had a time range from January 1, 2015 to December 31, 2019 and were collected as daily data. There were missing values in some variables, and these missing values were processed by the exponentially weighted moving average (EWMA) using the imputeTS package of the R software (Moritz, S. et al., 2017). The EWMA gives higher weights to the latest data, reducing the weight of older values. The formula for EWMA imputation suggested by Hunter, J. S. (1986) is as follows:

$$\hat{S}_t = \hat{S}_{t-1} + \alpha e_{t-1} \quad (1)$$

$$= \hat{S}_{t-1} + \alpha(S_{t-1} - \hat{S}_{t-1}) \quad (2)$$

$$= \alpha S_{t-1} + (1 - \alpha)\hat{S}_{t-1} \tag{3}$$

$$\vdots$$

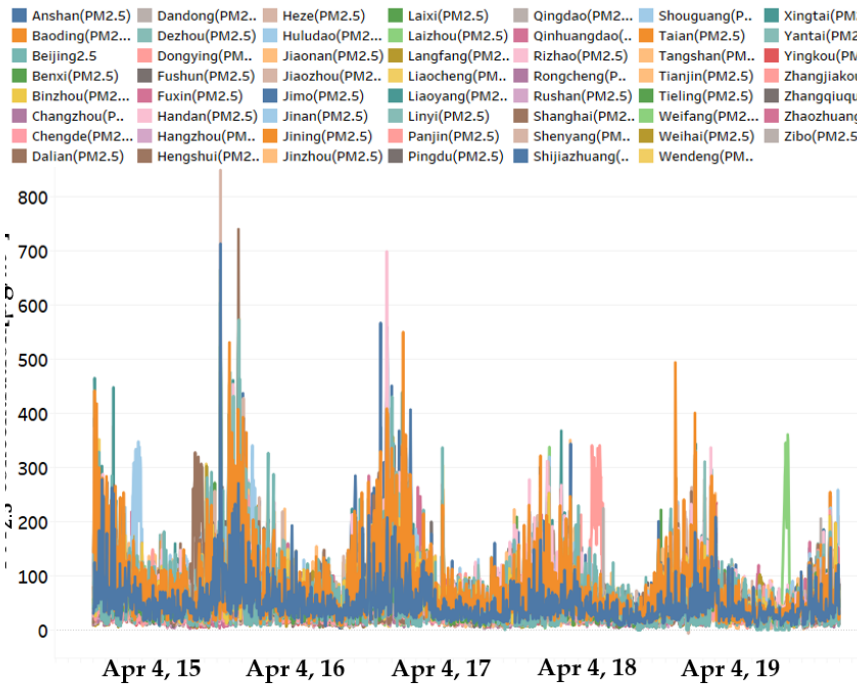
$$= \alpha \sum_{k=1}^{t-2} (1-\alpha)^{k-1} S_{t-k} + (1-\alpha)^{t-2} S_2 \tag{4}$$

$$* \ \alpha = \frac{2}{n+1}, \ n = \text{Moving Average Period}, k \in \{1,2,\ldots\}, t \geq 2 \tag{5}$$

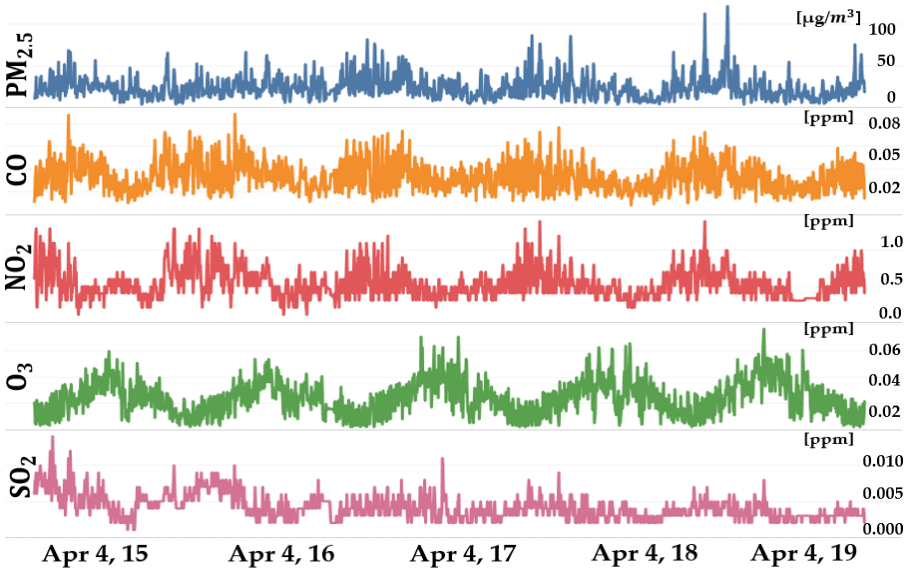
$\hat{S}_t$ is the predicted value at time t, S_t is the observed value at time t, e_t is the observed error at time t, and α is a constant value which refers to a weight from zero to one. Therefore, the higher the α value is, the less it reflects past data.

Figure 5 shows the concentration of $PM_{2.5}$ in China (Figure 5a) and that in Seoul (Figure 5b) with its other air quality factors, and Figure A5-A7 show the meteorological data of Seoul. Each variable shows the values in a different range due to the differences in units of measurement and the characteristics within the region. In the case of Chinese data, the concentration of $PM_{2.5}$ in each city over time seems to be constant, but some cities have outliers.

If one variable has a relatively greater value, or a wider range of values than the others in the composition of the data, it can exert a significant impact on the predicted value, regardless of the predictive importance of the variable.



(a) The concentration of $\mathbf{PM_{2.5}}$ in China cities



(b) $\mathbf{PM_{2.5}}$ concentration and air quality data in Seoul

Figure 5 Visualization of China's and Seoul's air quality data set

To solve these problems, the scope of the variables should be adjusted through normalization. In this study, maximum-minimum normalization was carried out to every data of each city as shown in the following equation:

$$\text{Normalized Variable's value} = \frac{\text{Variable's Original value} - \text{Variable's Minimum value}}{\text{Variable's Maximum value} - \text{Variable's Minimum value}} \quad (6)$$

Because the wind direction data were collected as 16 cardinal points, these are labels encoded to transform direction data into numerical data.

Section 3. Variable correlation analysis

As mentioned above, the prediction target of this study is the level of $PM_{2.5}$ concentration. The efficiency of the forecast results in deep learning, and machine learning depends on the correlation between the dependent and the independent variables. It is important to add variables that have strong negative or positive correlation with the dependent variable and the independent variable. In addition, the results of correlation are necessary for data analysis because they provide a basis for determining the influence of each independent variable on a dependent variable. In this study, the Pearson correlation coefficient was calculated, which is expressed as the covariance and standard deviation of the variables, as shown in the following equations in the case of observation vector $X = (X_1, X_2, X_3 \ldots, X_n)$:

Correlation Matrix

$$\begin{bmatrix} \frac{Cov(X_1,X_1)}{\sqrt{Var(X_1)}\sqrt{Var(X_1)}} & \frac{Cov(X_1,X_2)}{\sqrt{Var(X_1)}\sqrt{Var(X_2)}} & \cdots & \frac{Cov(X_1,X_n)}{\sqrt{Var(X_1)}\sqrt{Var(X_n)}} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{Cov(X_n,X_1)}{\sqrt{Var(X_n)}\sqrt{Var(X_1)}} & \frac{Cov(X_n,X_2)}{\sqrt{Var(X_n)}\sqrt{Var(X_2)}} & & \frac{Cov(X_n,X_n)}{\sqrt{Var(X_n)}\sqrt{Var(X_n)}} \end{bmatrix} \quad (7)$$

Each element in the correlation matrix has a value between −1 and 1, showing that a value greater than 0 is a positive correlation and a value less than 0 is a negative correlation. The correlation matrix is symmetric, and all of the diagonal elements of the matrix have a value of 1 based on the condition, $Cov(X_i, X_i) = Var(X_i)$, $i \in \{1,2,\ldots n\}$.

Figure 6 is a visualization of the correlation between eight factors with the highest values in Seoul, Korea. Appendix A Table A2-A9 show the correlation between $PM_{2.5}$ concentration

and the meteorological air quality factors of each city in Korea. Overall, the factors that have a strong positive correlation with $PM_{2.5}$ concentration are all air quality factors except, O_3.

$PM_{2.5}$ concentration also appears to have a positive correlation with local air pressure (LAP), sea-level pressure (SP), wind direction, and relative humidity. Conversely, temperature, wind speed, O_3, wind flow sum (wind flow sum refers to the distance that the air flows, and the Korea Meteorological Administration produces a day-to-day wind flow sum (24 h wind flow sum).), whereas daily precipitation is found to have a negative correlation with $PM_{2.5}$ concentration. However, the variables that have a relatively weak correlation with $PM_{2.5}$ concentration change the sign of the correlation depending on the region.

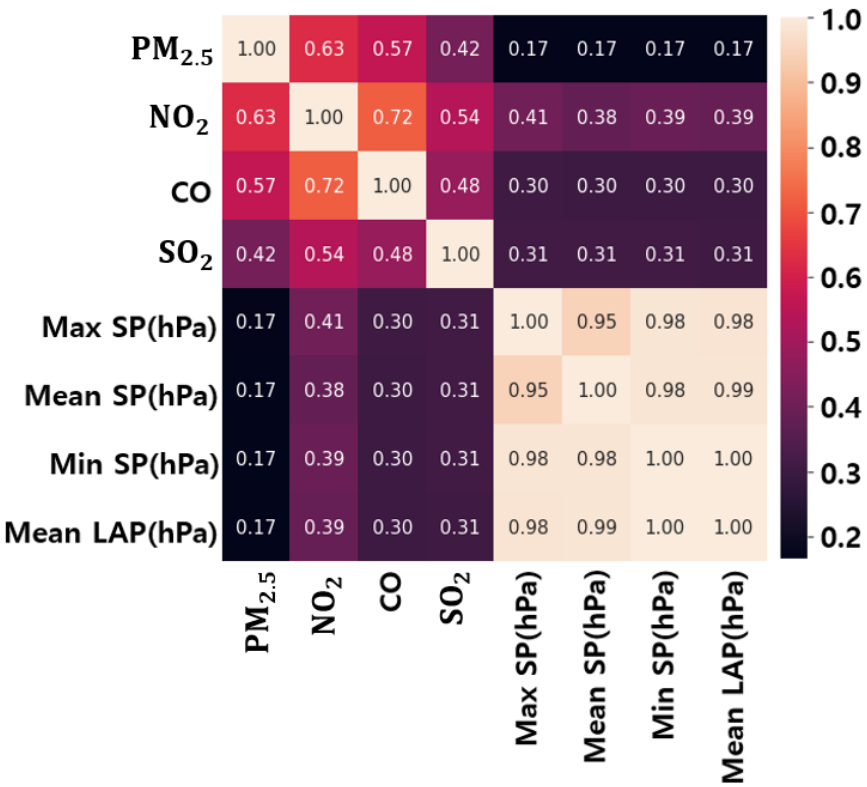


Figure 6 Correlation between the highest eight factors in Seoul, Korea

Figure 7 shows an origin-destination map of $PM_{2.5}$

correlations between Chinese (China National Environmental Monitoring Centre) and Korean cities. The correlations between $PM_{2.5}$ concentration in each Chinese city and $PM_{2.5}$ concentration in each Korean city vary. As shown in Table 2, the correlation range is from 0.13 to 0.55, which is along the same line with Park & Shin (2017). Comparing this with the factors inside the Korean cities, it can be interpreted that the $PM_{2.5}$ concentration of each city in China is related with the $PM_{2.5}$ concentration in Korea as much as it is related with the air quality data inside the city. Therefore, this implies that China's $PM_{2.5}$ concentration could serve as an important independent variable in predicting $PM_{2.5}$ concentration in Korea.

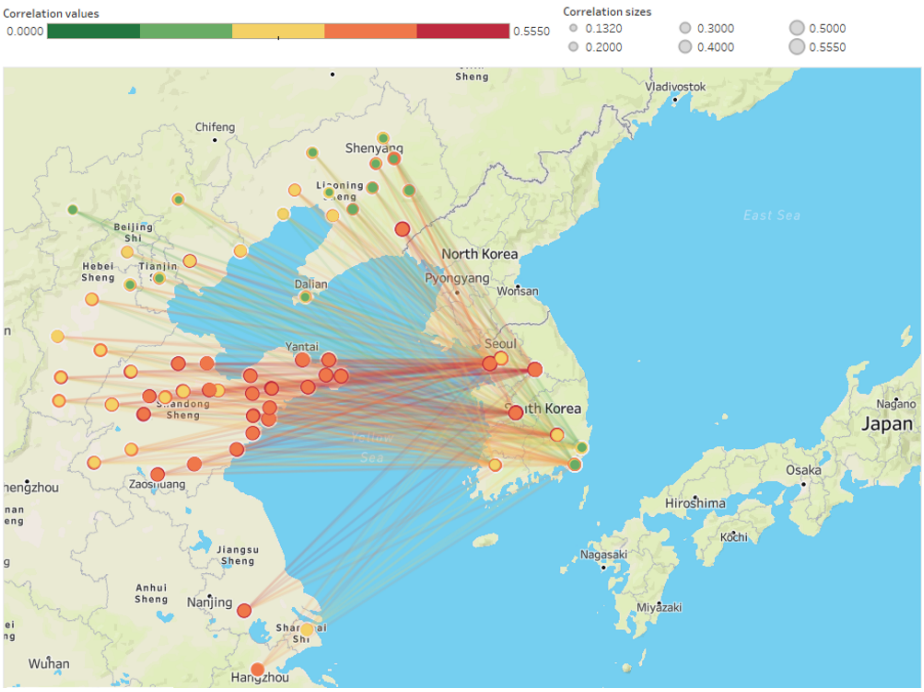


Figure 7 Origin–destination map of $PM_{2.5}$ concentration correlation between Chinese cities and Korean cities

Table 2 Correlation range from Chinese cities to Korean cities.

Cities	Minimum	Maximum

Seoul	0.1993	0.4994
Gwangju	0.1446	0.4035
Daegu	0.1813	0.5087
Daejeon	0.1839	0.5087
Busan	0.1320	0.5084
Ulsan	0.1419	0.5394
Wonju	0.1824	0.5550
Incheon	0.2556	0.5415

Chapter 4. Analysis Methods

Section 1. Principal components analysis(PCA)

PCA reduces dimensions by linear combinations of variables with high explanatory power of the overall data variability, explaining variation in high-dimension data in low dimensions. Vectors with p variables can have total p principal components, and the principal components of vector $\mathbf{x}(1 \times p)$, which has a covariance matrix of $\Sigma(p \times p)$, can be generated as follows:

$$PC = a^T x = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p \tag{8}$$

$$Var(a^T x) = a^T Var(x) a = a^T \Sigma a \tag{9}$$

$$\mathcal{L} = a^T \Sigma a - \lambda(a^T a - 1) \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial a} = 2\Sigma a - 2\lambda a = 0 \tag{11}$$

$$\Sigma a = \lambda a \tag{12}$$

$$\mathrm{Var}(PC) = a^T \Sigma a = a^T(\lambda a) = \lambda \tag{13}$$

$$PC_i = a_i^T x = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p \tag{14}$$

$$\mathrm{Var}(PC_i) = \lambda_i \tag{15}$$

Because the principal component is a linear combination of X, it can be expressed as Equation (8), and the variance of this linear combination can be expressed as Equation (9). The PCA has to preserve the variance of the original data as much as possible, so Equation (11) should also be maximized. Therefore, the method of generating principal components can be transformed into the problem of obtaining $a(p \times 1)$, which maximizes $a^T \Sigma a$ under the condition $a^T a = 1$. Equation (12) was derived by applying Lagrange's multiplier method to Equation (9). Equation (12) was made by Equation (11), which partially differentiates Equation (10) by a. Equation (12) shows that λ is the eigenvalue of Σ, and a is the eigenvector of Σ. As a result, a linear combination that maximizes Equation (9), i.e., the principal component, can be expressed as Equation (8). In addition, Equation (9), which is the variance of the principal component, can be expressed as λ under the condition $a^T a = 1$. Therefore, in vectors with p variables, the i-th principal component is Equation (14), and the variance is Equation (15).

Subsequently, the number of principal components is selected by the principal components where the sum of the principal components is more than 80% to 90% of the total variance. For example, the i number of principal components has to be selected out of principal components p. After application of Equation (16), the yielded results have to be more than 80% to 90%:

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_i}{\lambda_1+\lambda_2+\lambda_3+\lambda_4+\cdots+\lambda_p} \quad (16)$$

Section 2. Recurrent neural network

The RNN is a deep learning model for processing sequence data, such as stock charts (Hsieh, T.-J. et al., 2011), music (Franklin, J.A., 2006) and natural language processes (Goldberg, Y., 2017). It remembers the state entered from the previous time point (t - 1) through the hidden layer and passes the hidden layer state at that specific time point (t) to the next time point (t + 1). In other words, the status at the previous time point affects the state at the present time point, and the state at the present time point affects the status at the next time point. This procedure is repeated until result values are optimized, which refers to "Recurrent Neural Network."

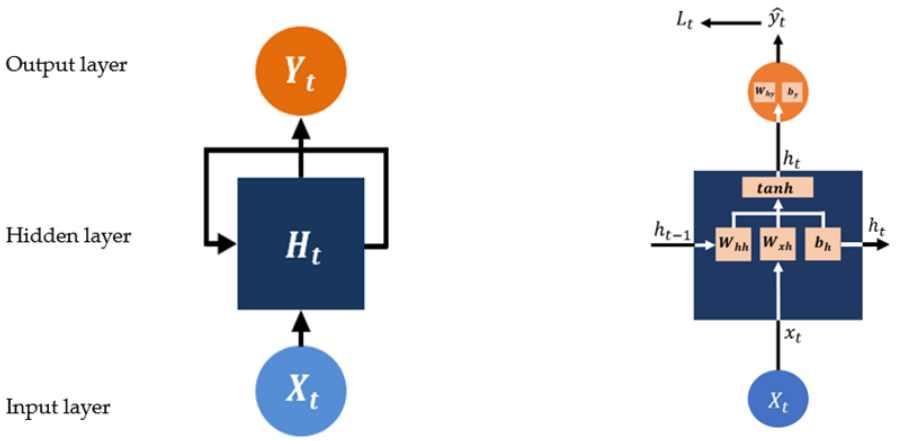
$$h_{t-1} = \tanh(W_{hh}h_{t-2} + W_{xh}x_{t-1} + b_h) \quad (17)$$

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \quad (18)$$

$$\hat{y_t} = W_{hy}h_t + b_y \quad (19)$$

$$L_t = \mathrm{MSE} = \frac{\sum (y_t - \hat{y}_t)^2}{n} \quad (20)$$

Figure 8b is the unrolled and inner structure of Figure 8a. In Equations (17)-(19), x_t is an input, and h_t is a hidden state at time t. W_{ij} is the weight from layer i to layer j, and b_i is the bias in each layer. In Equation (20), L_t is the loss at time t, and y_t and $\hat{y_t}$ are the actual and predicted values at time point t, respectively.



(a) Basic structure of RNN (b) Internal structure of RNN

Figure 8 Internal structure of recurrent neural network (RNN)

The RNN model shares the weights and biases at all time points and circulates the input data to output the results. Model training is repeated until the loss value is minimized by gradient descending in the loss function, with information of specific previous time steps. At the same time, the weights are updated to find optimum values. This is called backpropagation through time (BPTT) and in RNN, it can be expressed as follows (Chen, G. A., 2016).:

$$\text{Updated } W_{xh} = \text{Existing } W_{xh} - \eta \sum_{t=1}^{n} \sum_{k=0}^{n} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W_{xh}} \quad (21)$$

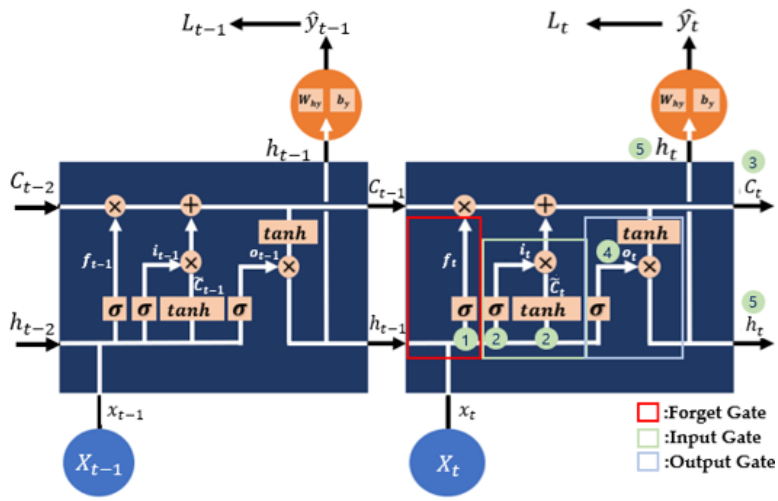
$$\text{Updated } W_{hh} = \text{Existing } W_{hh} - \eta \sum_{t=1}^{n} \sum_{k=0}^{n} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W_{hh}} \quad (22)$$

$$\text{Updated } W_{hy} = \text{Existing } W_{hy} - \eta \sum_{t=1}^{n} \frac{\partial L_t}{\partial W_{hy}} \quad (23)$$

$$* \eta = \text{learning rate } [0,1]$$

Section 3. Long short-term memory and Bidirectional LSTM

In an RNN, tanh is used as an activation function to train the model in a non-linear way. However, there is a long-term dependency problem caused by a "vanishing gradient" problem in the RNN's BPTT, in which the gradient (weights update rate) disappears as the value (derivative value of the tanh function with respect to h_t) less than 1 continues to multiply. Thus, the state of a relatively distant past time point has almost no effect on an output of the present time point. As a result, the model relies only on short-term data and has a limit in achieving the best performance. In order to solve this limitation, Schmidhuber & Hochreiter(1997) suggested the LSTM model.



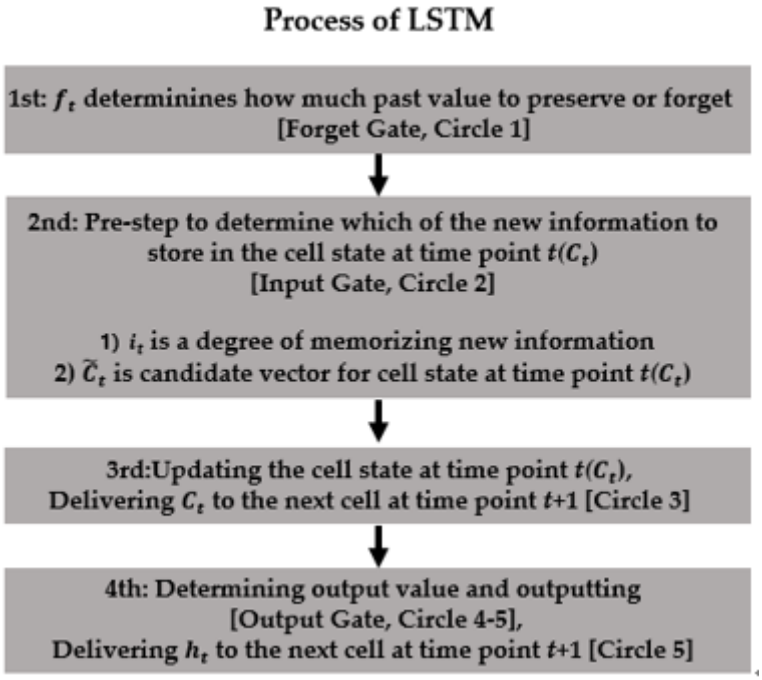


Figure 9 Internal structure of long short-term memory (LSTM)

Figure 9 shows the internal structure of LSTM and its process. LSTM is the model that incorporates forgetting and memory (f_t), the input (i_t), the inner cell state candidate ($\tilde{C}_t$), the conveying and inner cell state at time point t (C_t), and the output (o_t) into the RNN model. Especially, C_t, which pass-through all-time points, greatly contributes to solving the long-term dependency problem. The order of each part and the internal algorithm can be explained by the following process:

$$f_t = \sigma(W_{xh(f)} x_t + W_{hh(f)} h_{t-1} + b_{h(f)}) \quad (24)$$

$$i_t = \sigma(W_{xh(i)} x_t + W_{hh(i)} h_{t-1} + b_{h(i)}) \quad (25)$$

$$\tilde{C}_t = \tanh(W_{xh(\tilde{C}_t)} x_t + W_{hh(\tilde{C}_t)} h_{t-1} + b_{h(\tilde{C}_t)}) \quad (26)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (27)$$

$$o_t = \sigma(W_{xh(o)} x_t + W_{hh(o)} h_{t-1} + b_{h(o)}) \quad (28)$$

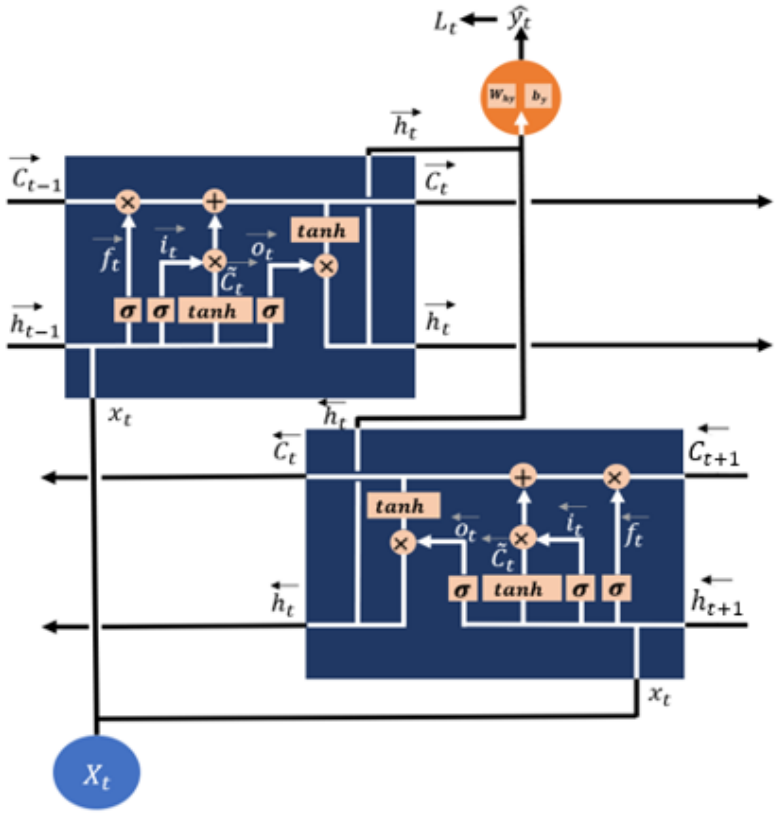
$$h_t = o_t \odot \tanh(C_t) \quad (29)$$

$$* \odot = \text{Hadamard product}, \sigma = \text{sigmoid function} = \frac{1}{1 + e^{-x}}$$

Equation (24), output of the forget gate, determines whether the historical state is forgotten by the combination of x_t and h_{t-1}. The output value of this step is converted to a number between 0 and 1 by the sigmoid function and multiplied by C_{t-1} (memory of past data, i.e., historical state) to determine how much past data to preserve or forget. A value of 0 indicates forgetfulness, and 1 indicates memorization of past data. Equations (25) and (26) are involved in the storage of the inner cell state of time point t. Equation (25), output of the input gate, determines how much data of time point t are memorized. In other words, it has a value between 0 and 1, indicating the degree of memorization of the new information. Simultaneously, Equation (26) generates the inner cell state candidate of time point t. Then, Equation (27) generates the new cell state at time point t and passes it on to the LSTM cell at the next time point (t + 1). In other words, LSTM solves the RNN's long-term dependency problem by adjusting the memorization and forgetfulness of the past, and presents the state through Equations (24)- (27). Ultimately, the output is decided by Equations (28) and (29). More precisely, Equation (28), output of the output gate, decides which part of the new cell state will be the output. A value of the new cell status is converted through the tangent function and calculated with the result value from Equation (28) to produce the final output of time point t, as shown in Equation (29).

BiLSTM is a variant of the bidirectional RNN proposed by Schuster & Paliwal(1997). Figure 10 shows an example of applying a bidirectional way to sentence learning. If (A) is taught in the model and "went" is set as the target, (B)

predicts in a forward way and (C) predicts in both a forward and a backward way. If LSTM uses a historical state to predict the value of time point t, bidirectional LSTM predicts the value of time point t by adding an LSTM layer that reads data from a future state. The computations within the model are the same as those of LSTM, and LSTM and BiLSTM update their weights in the training model as RNN(Gonzalez & Yu, 2018).



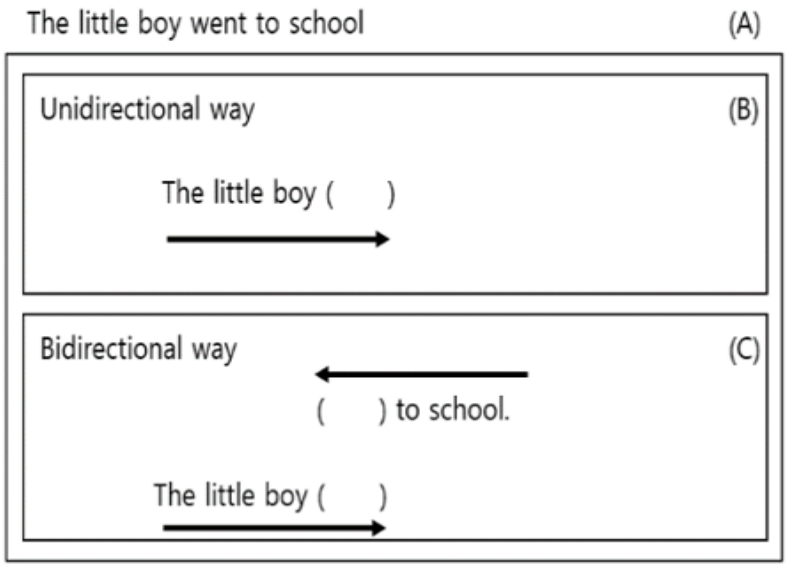


Figure 10 Internal structure of bidirectional long short-term memory **(**BiLSTM**)** and an example

Section 4. Model training process

In this study, the learning applied to the models is a supervised learning that provides input values and actual output values in training models. In supervised learning, the whole data is divided into two parts: training set and test set. Here, the models learn the training set which includes input values and actual output values. Part of the training set is allocated to the validation set, which is used to increase the efficiency of models' learning. If the validation set does not exist, the overfitting problem can occur and the performance can drop significantly. The assigned validation set allows the models to conduct another test while the models learn the training set. This contributes to supervision on degree of models' learning, and selection of the optimal parameters within the models. Overall, when the models receive the input values and export the output values through own operation, they proceed learning

by comparing the output values to the given actual values at every point in time to reduce the loss values during the training process. After training, the models generate output values in the test set that is solely composed of input values. Then, the models compare generated output values to the actual output values which is not part of the test set, to measure the models' performance.

Figure 11 Data set division

All models train to receive a series of historical values from a particular point in time and then output the values at the next point in time. Based on the time point t, the output value of the time t+1 is calculated by entering the past time steps. Hence, the seven-time steps data are set as input units in this particular study. And each time step contains all relevant data of that time step. For example, when predicting the concentration of $PM_{2.5}$ on day 8 in a selected region of South Korea, $PM_{2.5}$ concentration from day 1 to day 7 of that region is used as input variables. Besides, $PM_{2.5}$ concentration, air quality, weather factors of adjacent Chinese cities are also used as input variables. This means that the structure of one unit of input values is equivalent to "the number of variables × the number of time steps".

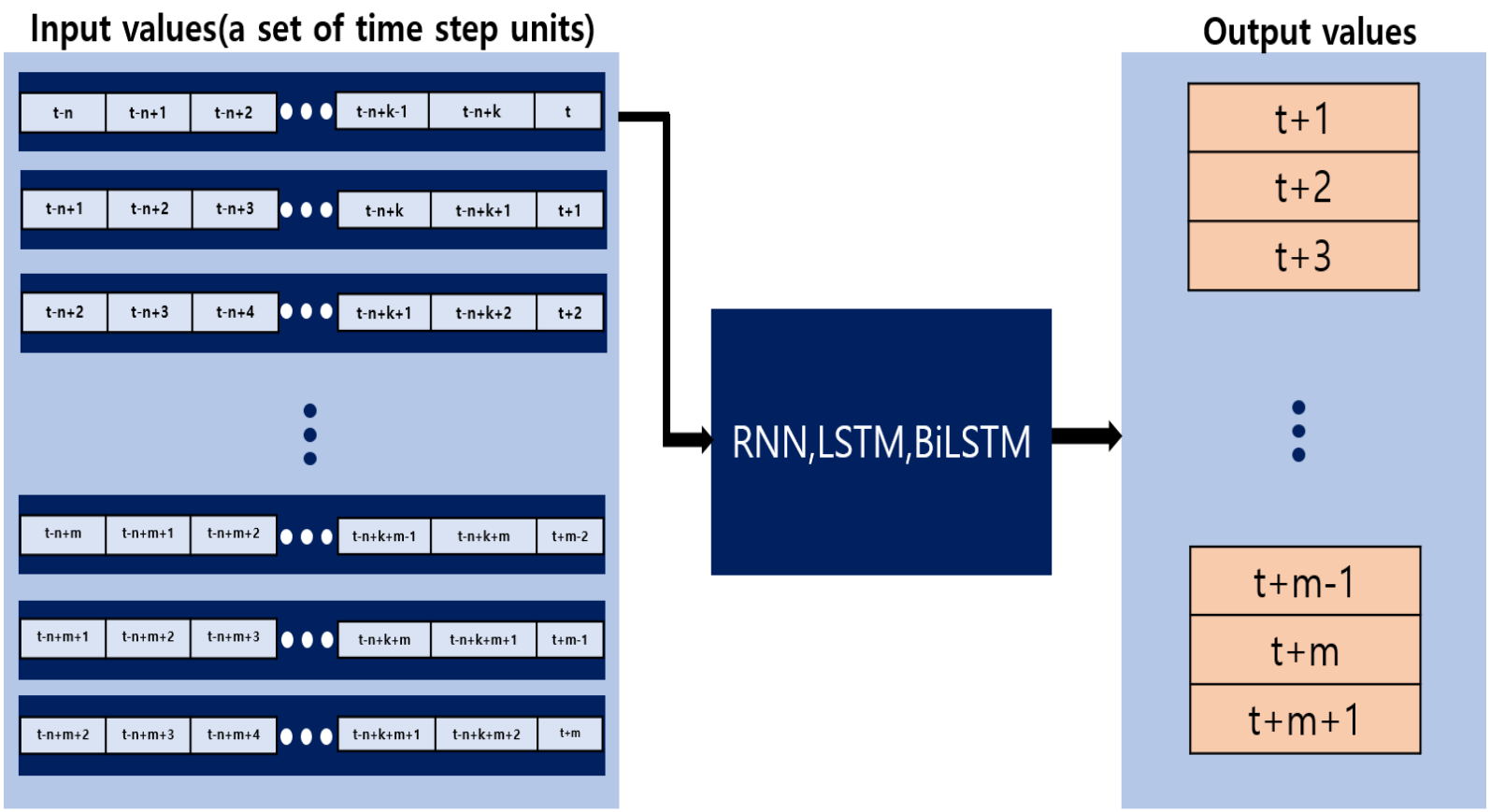


Figure 12 Data input output structure in models

Section 5. Evaluation model performance

In this study, MAE(Mean Absolute Error) and RMSE(Root Mean Square Error) were used as evaluation indicators to compare the performance of each model with and without PCA application. The calculations of each indicator are expressed as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y_t}| \tag{30}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y_t})^2} \tag{31}$$

Section 6. Research procedure

This study was conducted in four stages: data collection, data preprocessing, prediction, and evaluation (Figure 12). The application of PCA was used in the data preprocessing stage, aiming to reduce the number of variables and increase the performance of model predictions. Thus, the data preprocessing stage was divided into two cases. Case 1 was set as a prediction without a PCA application, whereas Case 2 was set as a prediction with the PCA application. Afterwards, a comparison analysis of two cases was carried out using evaluation indicators (MAE and RMSE).

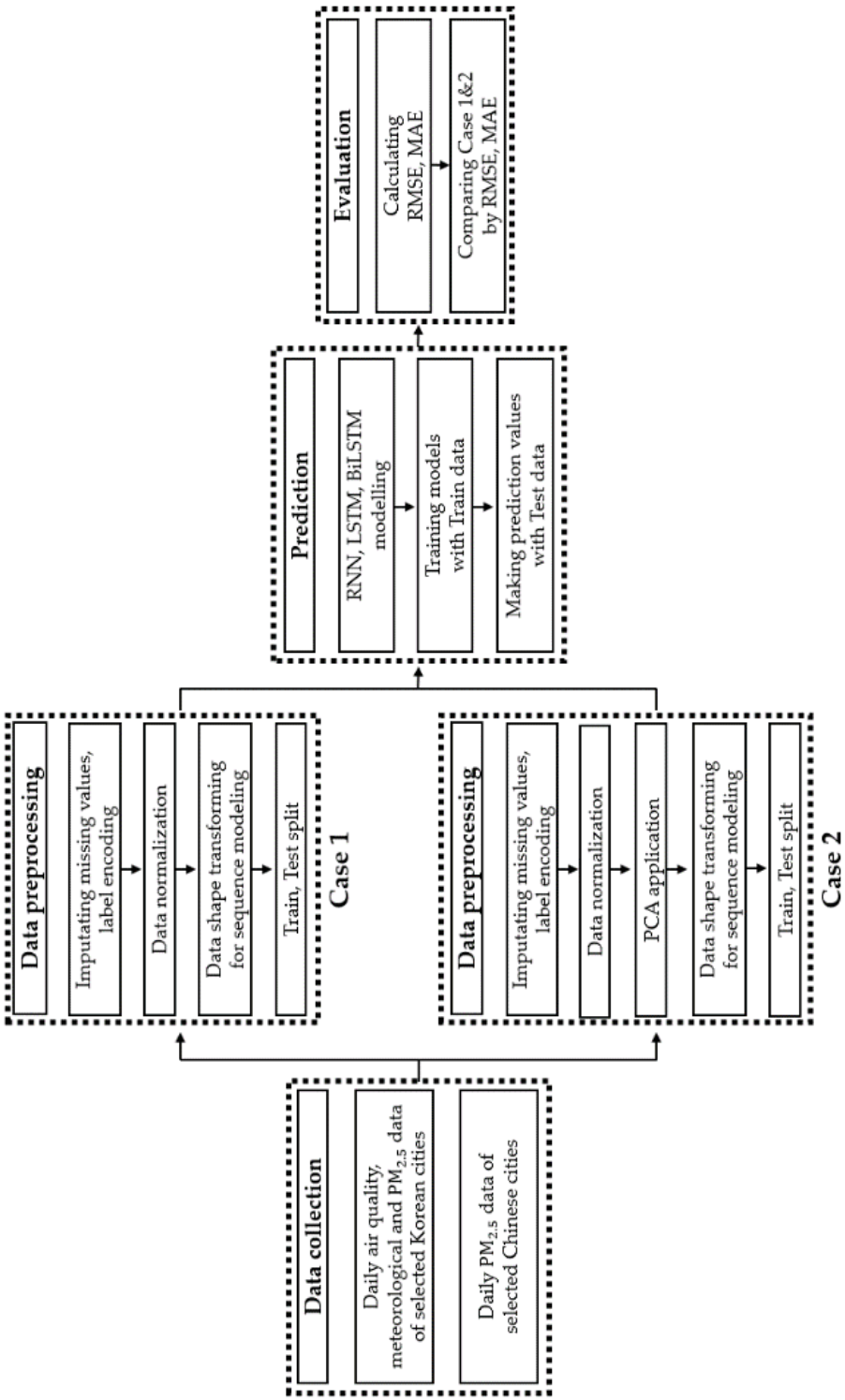


Figure 13 Research procedure of the PCA application deep learning model for predicting $\mathbf{PM_{2.5}}$ concentration

Chapter 5. Result

Section 1. Principal components selection

PCA was performed on the input variables of each city, excluding $PM_{2.5}$ concentration. The variance of each city's input variable data was explained by a relatively small number of principal components. Ultimately, a different set of five principal components was drawn for each city. This reduced the number of input variables to about 1/16. To elaborate, Table 3 shows the degree of how much overall variation of each city can be explained by selected five principal components.

Table 3 The ratio of variance explained
by five principal components in each city

Cities	Cumulative Variance
Seoul	0.9631(=96.31%)
Gwangju	0.9553(=95.53%)
Daegu	0.9770(=97.70%)
Daejeon	0.9539(=95.39%)
Busan	0.98102(=98.102%)
Ulsan	0.9655(=96.55%)
Wonju	0.9366(=93.66%)
Incheon	0.98123(=98.123%)

Section 2. Setup and case comparison

China's daily $PM_{2.5}$ concentration and Korea's air quality and meteorological data were collected from January 1 2015 to December 31 2019 to predict the $PM_{2.5}$ concentration in eight Korean cities. In total, 85% of the collected data were allocated to the training set and 15% to the test set. To specify, the three models have 256 units in the layer, a tanh activation function, a batch size of 64, 200 epochs, and an adaptive moment estimation (ADAM) optimizer (Kingma & Ba, 2014)

To avoid overfitting, 30% of the training set was designated as a validation set, and 30% dropout regulation was used between the input layer and the output layer. In terms of model learning, earlystopping, one of the callback functions of Keras, was applied. This was used to stop learning if optimal learning is achieved before 200 epochs.

Figure 14 shows the predicted and actual values of $PM_{2.5}$ concentration for each case and model in Seoul. Figure A8-A14 show the prediction of $PM_{2.5}$ concentration of remaining cities. On one hand, RNN shows relatively low predictive power in both Case 1 and Case 2. The RNN without PCA seems to be corresponding more closely to the trend and show relatively higher performance than the RNN with PCA does.

On the other hand, LSTM and BiLSTM's performance outweighed that of RNN in terms of its compliance to actual $PM_{2.5}$ concentration trend as shown in Figure 14. Minor differences in predictive power may exist between cities; however, LSTM and BiLSTM show better overall performance regardless of application of PCA. Moreover, it can be seen that PCA application in all cities corrects the difference between the predicted values and actual values that may have existed if PCA was not applied. Furthermore, it also appears to have produced more accurate results in terms of predicting peak values.

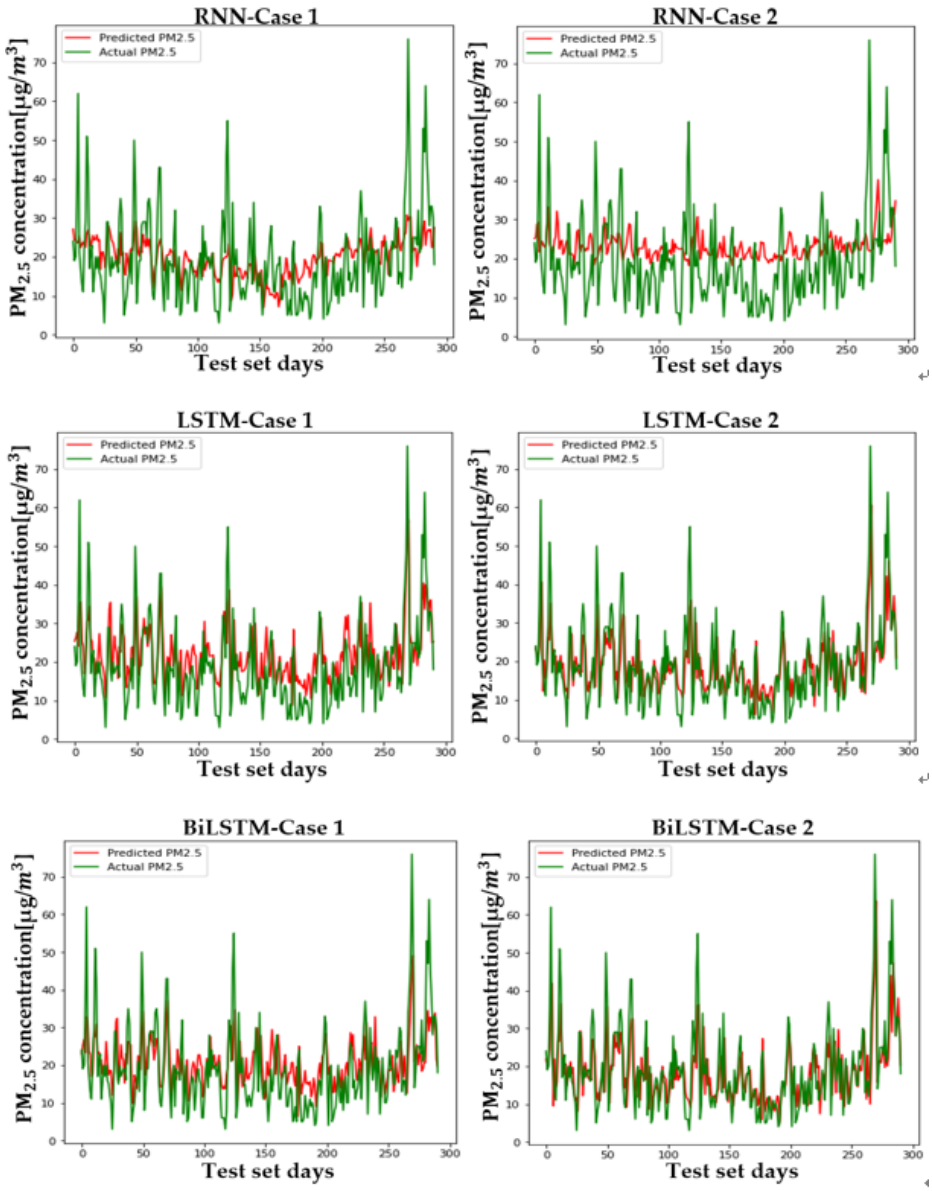


Figure 14 The $\mathbf{PM_{2.5}}$ concentration prediction in Seoul by two cases

Table 4 and 5 are numerical representations of Figure 14. As noted above, it is understood that the reduction in dimension in all cities leads to a relatively low performance in the RNN, except for Daegu in terms of the MAE. In other

words, reducing variables does not help model learning for RNN. Instead, providing a large amount of information in a short period of time can lead to a better performance, depending on the feature of the model that relies on short term information. Since RNN has low overall accuracy, results that should be considered are those from LSTM and BiLSTM rather than those from RNN. Unlike the RNN, PCA application to LSTM and BiLSTM show better results in RMSE and MAE evaluation, which aligns with visual results from Figure 14. The level of performance is ranked from the highest to the lowest in the following order: Busan > Daejeon > Gwangju > Daegu > Seoul > Ulsan > Wonju > Incheon. Additionally, the degree of improvement in MAE and RMSE is ranked from the highest to the lowest in the following order: Busan > Incheon > Gwangju > Seoul > Ulsan > Daegu > Daejeon > Wonju. Overall, LSTM shows high performance in Daejeon, Daegu, and Busan, while BiLSTM showed higher performance in the cities excluding Daejeon, Daegu and Busan.

Acknowledging differences in performance and performance improvements of each city is meaningful in that characteristics of each city can cause regional differences, which can lead to different performances from the same model. To clarify, this means that a certain model would perform better depending on the city' s regional characteristics. In order to overcome this potential issue, multidisciplinary considerations are required for further studies.

Table 4 Evaluation results from $\mathbf{PM_{2.5}}$ concentration prediction in each Korean city **(**Case 1**)**

City	Model	RMSE	MAE
Seoul	RNN	9.730	7.328
	LSTM	8.020	6.374
	BiLSTM	8.101	6.168
Daegu	RNN	10.171	8.110
	LSTM	7.654	6.223
	BiLSTM	7.707	6.193
Daejeon	RNN	9.361	7.497
	LSTM	7.042	5.753
	BiLSTM	7.231	5.927
Busan	RNN	11.603	9.208
	LSTM	8.718	6.520
	BiLSTM	8.459	6.251
Gwangju	RNN	9.002	7.472
	LSTM	7.7415	5.797
	BiLSTM	8.300	6.590
Busan	RNN	8.410	7.224
	LSTM	7.770	6.504
	BiLSTM	7.897	6.578
Ulsan	RNN	10.558	8.988
	LSTM	8.660	6.959
	BiLSTM	8.383	6.772
Incheon	RNN	13.686	11.408
	LSTM	11.900	9.828
	BiLSTM	10.393	8.285

Table 5 Evaluation results from $PM_{2.5}$ concentration prediction in each Korean city (Case 2)

City	Model	RMSE	MAE
Seoul	RNN	11.680(20%↑)	9.310(27%↑)
	LSTM	7.667(4.6%↓)	5.455(16.8%↓)
	BiLSTM	7.567(7.1%↓)	5.368(14.9%↓)
Daegu	RNN	10.208(0.4%↑)	7.824(3.5%↓)
	LSTM	7.491(2.2%↓)	5.664(9.9%↓)
	BiLSTM	7.552(2.1%↓)	5.703(8.6%↓)
Daejeon	RNN	9.602(2.6%↑)	7.824(4.4%↑)
	LSTM	6.967(1.1%↓)	5.374(7.1%↓)
	BiLSTM	7.098(1.9%↓)	5.537(7%↓)
Busan	RNN	12.132(4.6%↑)	9.758(6%↑)
	LSTM	8.424(3.5%↓)	6.251(4.3%↓)
	BiLSTM	8.345(1.4%↓)	6.137(1.9%↓)
Gwangju	RNN	9.492(5.4%↑)	7.746(3.7%↑)
	LSTM	7.148(8.3%↓)	5.541(4.6%↓)
	BiLSTM	7.110(16.7%↓)	5.455(20.8%↓)
Busan	RNN	9.924(18%↑)	8.316(15.1%↑)
	LSTM	6.668(16.5%↓)	4.881(33.3%↓)
	BiLSTM	6.779(16.5%↓)	4.999(31.6%↓)
Ulsan	RNN	11.160(5.7%↑)	9.389(4.5%↑)
	LSTM	8.021(8%↓)	6.251(11.3%↓)
	BiLSTM	7.871(6.5%↓)	5.993(13%↓)
Incheon	RNN	14.744(7.7%↑)	12.427(8.9%↑)
	LSTM	10.205(16.6%↓)	8.000(22.9%↓)
	BiLSTM	9.709(7%↓)	7.354(12.7%↓)

Chapter 6. Conclusion

Performance degradation can occur in deep learning and machine learning due to the curse of dimensionality. To prevent such outcome, this study proposes a PCA-applied model. The performance comparison with a non-PCA model demonstrates that PCA applications produce better results in deep learning time series prediction. This finding suggests that such a performance improvement technique can be used to increase the efficiency of the government system by providing better forecasts. Thereby, this will provide a foundation and rationale for issuing crisis alerts and designing air pollution reduction policies in the future.

As the correlation analysis shows, the concentration of $PM_{2.5}$ in China appears to have positive correlations with the concentration of $PM_{2.5}$ in Korea. This indicates that we have to consider China's air pollution factors when predicting the concentration of $PM_{2.5}$ in Korea. Furthermore, it justifies the current setup of real-time air pollution databases between the two countries from the ongoing joint research between Korea and China (Ministry of Environment, 2020).

Despite PCA application's ability to improve model performance, the results show relatively weak predictions on predicting the minimum and maximum $PM_{2.5}$ concentration for each city. Such outcome can be attributed to small number of observations since daily observations replaced hourly observations. In other words, future joint cross-border research performance can be easily enhanced with collection of more observations. Moreover, some meteorological data in each Korean city show a relatively weak correlation with concentration, so it seems necessary to find variables that have causality or strong correlation within academic areas other than deep learning. For example, if spatial factors (spatial

homogeneity, autocorrelation, etc.) in Chinese cities and Korean cities are added to the model as input variables, it is expected that the model will produce better performance in terms of learning time and spatial features of data.

In addition, there were limitations in data collection over time units of data. Along with weather factors, factors related to the occurrence of $PM_{2.5}$ in Korea include number of thermoelectric power plants and power generation, secondary generators of $PM_{2.5}$, number of vehicles, and old diesel cars. But this research could not include all factors above due to differences in data recording units (days, months, years). In general, $PM_{2.5}$ concentration between Korea and adjacent Chinese cities showed higher correlation. However, low correlation appeared in some northern regions despite their close proximity to Korea. This suggests that Chinese regional factors, such as the number of chemical and thermoelectric power plants should also be included in the data set when predicting $PM_{2.5}$ concentration in Korea. Therefore, further study with higher performance can take place if data such as daily power generation and traffic volume, daily public transportation usage, and daily generators of $PM_{2.5}$ by combining chemicals in each region can be acquired and reflected in models.

This research will proceed to improve the prediction performance of deep learning models by increasing number of observations and optimizing models. At the same time, it will apply new algorithms and find additional variables that have causality with $PM_{2.5}$ concentration in the field of econometrics and spatial econometrics.

References

Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. Science of the Total Environment, 731, 139052.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., & Briggs, D. J. (2009). Mapping of background air pollution at a fine spatial scale across the European Union. Science of the Total Environment, 407(6), 1852-1867.

China National Environmental Monitoring Centre [Website]. (2020, 03. 01). Retrieved from http://www.cnemc.cn/sssj/

Chen, G. (2016). A gentle tutorial of recurrent neural network with error backpropagation. arXiv preprint arXiv:1610.02583.

César, A. C. G., Nascimento, L. F. C., Mantovani, K. C. C., & Vieira, L. C. P. (2016). Fine particulate matter estimated by mathematical model and hospitalizations for pneumonia and asthma in children. Revista Paulista de Pediatria (English Edition), 34(1), 18-23.

Franklin, J. A. (2006). Recurrent neural networks for music computation. INFORMS Journal on Computing, 18(3), 321-338.

French National Health Agency, InVS, European Environment Agency [Website]. (2020. 03. 03). Retrieved from https://news.yahoo.com/micro-pollution-ravaging-china-south-asia-study-031634307.html

Gonzalez, J., & Yu, W. (2018). Non-linear system modeling using LSTM neural networks. IFAC-PapersOnLine, 51(13), 485-489.

Goldberg, Y. (2017). Neural network methods for natural language processing. Synthesis lectures on human language technologies, 10(1), 1-309.

Hunter, J. S. (1986). The exponentially weighted moving average. Journal of quality technology, 18(4), 203-210.

Han, C., Kim, S., Lim, Y. H., Bae, H. J., & Hong, Y. C. (2018). Spatial and temporal trends of number of deaths attributable to ambient PM2. 5 in the Korea. Journal of Korean medical science, 33(30).

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. Neural Comput, 9(8), 1735-1780.

Thaweephol, K., & Wiwatwattana, N. (2019, November). Long Short-Term Memory Deep Neural Network Model for PM2. 5 Forecasting in the Bangkok Urban Area. In 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE) (pp. 1-6). IEEE.

Hsieh, T. J., Hsiao, H. F., & Yeh, W. C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. Applied soft computing, 11(2), 2510-2525.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., ... & Sachdeva, S. (2019). Evaluation of different machine

learning approaches to forecasting PM2. 5 mass concentrations. Aerosol and Air Quality Research, 19(6), 1400-1410.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kim, K. N., Kim, S., Lim, Y. H., Song, I. G., & Hong, Y. C. (2020). Effects of short-term fine particulate matter exposure on acute respiratory infection in children. International journal of hygiene and environmental health, 229, 113571.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. R J., 9(1), 207.

Mengara Mengara, A. G., Kim, Y., Yoo, Y., & Ahn, J. (2020). Distributed Deep Features Extraction Model for Air Quality Forecasting. Sustainability, 12(19), 8014.

Nullschool [Website]. (2020, 01. 30). Retrieved from https://earth.nullschool.net/ko/

OECD Air pollution exposure [Website]. (2019. 11. 11). Retrieved from https://data.oecd.org/air/air-pollution-exposure.htm

Pozza, S. A., Lima, E. P., Comin, T. T., Gimenes, M. L., & Coury, J. R. (2010). Time series analysis of PM2. 5 and PM10- 2.5 mass concentration in the city of Sao Carlos, Brazil. International Journal of Environment and Pollution, 41(1-2), 90-108.

Qadeer, K., Rehman, W. U., Sheri, A. M., Park, I., Kim, H. K., & Jeon, M. (2020). A Long Short-Term Memory (LSTM) Network for Hourly Estimation of PM2. 5

Concentration in Two Cities of South Korea. Applied Sciences, 10(11), 3984.

Ross, Z., Jerrett, M., Ito, K., Tempalski, B., & Thurston, G. D. (2007). A land use regression for predicting fine particulate matter concentrations in the New York City region. Atmospheric Environment, 41(11), 2255-2269.

Singh, V., Carnevale, C., Finzi, G., Pisoni, E., & Volta, M. (2011). A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations. Environmental Modelling & Software, 26(6), 778-786.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.

Vinikoor-Imler, L. C., Davis, J. A., & Luben, T. J. (2011). An ecologic analysis of county-level PM2. 5 concentrations and lung cancer incidence and mortality. International journal of environmental research and public health, 8(6), 1865-1871.

Wang, C., Tu, Y., Yu, Z., & Lu, R. (2015). PM2. 5 and cardiovascular diseases in the elderly: an overview. International journal of environmental research and public health, 12(7), 8187-8197.

WHO Health Organization. Ambient (Outdoor) Air Pollution [Website]. (2019. 12. 08). Retrieved from https://www.who.int/
news-room/fact-sheets/detail/ambient-(outdoor)-airquality-

and-health

Xayasouk, T., Lee, H., & Lee, G. (2020). Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. Sustainability, 12(6), 2570.

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., ... & Wang, J. (2018). Trend analysis and forecast of PM2. 5 in Fuzhou, China using the ARIMA model. Ecological indicators, 95, 702-710.

Zhao, J., Deng, F., Cai, Y., & Chen, J. (2019). Long short-term memory-Fully connected (LSTM-FC) neural network for PM2. 5 concentration prediction. Chemosphere, 220, 486-492.

공성용. (2012). 초미세먼지(PM2.5)의 건강영향 평가 및 관리정책 연구 1. 기본연구보고서, 2012, 1-209.

기상청 기상자료개발포털 [웹사이트]. (2019. 02. 15). Retrieved from https://data.kma.go.kr/cmmn/main.do

관계부처 합동. (2019), 미세먼지 관리 종합계획(2020~2024)

박순애, & 신현재. (2017). 한국의 초미세먼지 (PM 2.5)의 영향요인 분석: 풍향을 고려한 계절성 원인을 중심으로. 환경정책, 25(1), 227-248.

에어코리아 [웹사이트]. (2020. 01. 30). Retrieved from www.airkorea.or.kr

최종일, & 이영수. (2015). 초미세먼지 (PM 2.5) 배출량이 호흡기계 질환에 미치는 영향 연구. 환경정책, 23(4), 155-172.

통계청 (2020, 09. 21). 2019년 사망원인통계 결과.

보도자료.

환경부 (2020. 01. 22). 한중 공동연구단, 미세먼지 저감 위한 마중물 수행. 보도자료.

황인창, & 손원익. (2019). 서울시 미세먼지 관리 정책의 사회경제적 편익: 지불용의액 중심으로. 한국정책학회 춘계학술발표논문집, 2019, 9-26.

Appendix

Appendix 1. The correlation coefficient of the mete −orological and air quality factors between $PM_{2.5}$ concentration in each city

Table A1 shows the acronym list of Table A2-A9 and Figure A5-A7.

Table A1 Acronym list

Acronym	Meaning
Min Temp	Minimum temperature (° C)
Max Temp	Maximum temperature (° C)
Mean Temp	Mean temperature (° C)
Daily prep	Daily precipitation (mm)
Max inst WS	Maximum instantaneous wind speed (m/s)
Max inst WSD	Maximum instantaneous wind speed directions (16 cardinal points)
Max WS	Maximum wind speed (m/s)
Max WSD	Maximum wind speed directions (16 cardinal points)
Mean WS	Mean wind speed (m/s)
WFS	Wind flow sum (100 m)
Max freq WD	Maximum frequent wind directions (16 cardinal points)
Mean DP	Mean dew point (° C)
Mean RH	Mean relative humidity (%)

Mean LAP	Mean local atmospheric pressure (hPa)
Max SP	Maximum sea-level pressure (hPa)
Min SP	Minimum sea-level pressure (hPa)
Mean SP	Mean sea-level pressure (hPa)
Min RH	Minimum relative humidity (%)

Table A2-A9 show the correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in each city.

Table A2 The correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in Seoul

Air quality factors	O_3 (ppm)	−0.021	**Meteorological factors**	Min Temp	−0.175	Max inst WS	−0.196	Mean WS	−0.174	Mean RH	0.013	Mean SP	0.168
	CO (ppm)	0.565		Max Temp	−0.185	Max inst WSD	0.09	WFS	−0.175	Mean LAP	0.166	Min RH	−0.047
	NO_2 (ppm)	0.627		Mean Temp	−0.156	Max WS	−0.098	Max freqWD	0.041	Max SP	0.17		
	SO_2 (ppm)	0.417		Daily prep	−0.143	Max WSD	0.118	Mean DP	−0.141	Min SP	0.169		

Table A3 The correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in Gwangju

Air quality factors	O_3 (ppm)	0.108	**Meteorological factors**	Min Temp	−0.223	Max inst WS	−0.226	Mean WS	−0.28	Mean RH	−0.164	Mean SP	0.192
	CO (ppm)	0.532		Max Temp	−0.122	Max inst WSD	0.108	WFS	−0.281	Mean LAP	0.192	Min RH	−0.235
	NO_2 (ppm)	0.562		Mean Temp	−0.179	Max WS	−0.214	Max freq WD	0.11	Max SP	0.186		
	SO_2 (ppm)	0.276		Daily prep	−0.212	Max WSD	0.102	Mean DP	−0.2	Min SP	0.196		

Table A4 The correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in Daegu

Air quality factors	O_3 (ppm)	−0.113	**Meteorological factors**	Min Temp	−0.291	Max inst WS	−0.305	Mean WS	−0.373	Mean RH	−0.056	Mean SP	0.256
	CO (ppm)	0.665		Max Temp	−0.193	Max inst WSD	0.156	WFS	−0.374	Mean LAP	0.252	Min RH	−0.128
	NO_2 (ppm)	0.702		Mean Temp	−0.244	Max WS	−0.317	Max freq WD	0.053	Max SP	0.26		
	SO_2 (ppm)	0.437		Daily prep	−0.157	Max WSD	0.14	Mean DP	−0.214	Min SP	0.253		

Table A5 The correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in Daejeon

Air quality factors	O_3 (ppm)	−0.086	**Meteorological factors**	Min Temp	−0.299	Max ins tWS	−0.241	Mean WS	−0.265	Mean RH	−0.101	Mean SP	0.272
	CO (ppm)	0.535		Max Temp	−0.236	Max inst WSD	0.121	WFS	−0.265	Mean LAP	0.271	Min RH	−0.173
	NO_2 (ppm)	0.483		Mean Temp	−0.272	Max WS	−0.239	Max freq WD	0.211	Max SP	0.27		
	SO_2 (ppm)	0.41		Daily prep	−0.18	Max WSD	0.1	Mean DP	−0.271	Min SP	0.272		

Table A6 The correlation coefficient of the meteorological and air quality factors between $\mathbf{PM_{2.5}}$ concentration in Busan

Air quality factors	O_3 (ppm)	0.029	**Meteorological factors**	Min Temp	−0.139	Max inst WS	−0.231	Mean WS	−0.162	Mean RH	−0.126	Mean SP	0.104
	CO (ppm)	0.32		Max Temp	−0.095	Max inst WSD	0.196	WFS	−0.162	Mean LAP	0.102	Min RH	−0.187
	NO_2 (ppm)	0.554		Mean Temp	−0.119	Max WS	−0.07	Max freqWD	0.178	Max SP	0.086		
	SO_2 (ppm)	0.366		Daily prep	−0.17	Max WSD	0.249	Mean DP	−0.125	Min SP	0.124		

Table A7 The correlation coefficient of the meteorological and air quality factors between $\mathbf{PM_{2.5}}$ concentration in Ulsan

Air quality factors			Meteorological factors										
	O_3 (ppm)	0.095		Min Temp	−0.084	Max inst WS	−0.198	Mean WS	−0.318	Mean RH	−0.125	Mean SP	0.032
	CO (ppm)	0.665		Max Temp	0.053	Max inst WSD	0.023	WFS	−0.319	Mean LAP	0.064	Min RH	−0.233
	NO_2 (ppm)	0.667		Mean Temp	−0.015	Max WS	−0.166	Max freq WD	−0.055	Max SP	0.016		
	SO_2 (ppm)	0.525		Daily prep	−0.167	Max WSD	0.014	Mean DP	−0.055	Min SP	0.051		

Table A8 The correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in Wonju

Air quality factors			Meteorological factors										
	O_3 (ppm)	−0.129		Min Temp	−0.384	Max inst WS	−0.187	Mean WS	−0.257	Mean RH	−0.018	Mean SP	0.309
	CO (ppm)	0.686		Max Temp	−0.339	Max inst WSD	0.171	WFS	−0.259	Mean LAP	0.299	Min RH	−0.077
	NO_2 (ppm)	0.675		Mean Temp	−0.366	Max WS	−0.187	Max freq WD	0.077	Max SP	0.318		
	SO_2 (ppm)	0.575		Daily prep	−0.184	Max WSD	0.14	Mean DP	−0.326	Min SP	0.302		

Table A9 The correlation coefficient of the meteorological and air quality factors between $PM_{2.5}$ concentration in Incheon

Air Quality Factors	O_3 (ppm)	−0.102	**Meteorological Factors**	Min Temp	−0.15	Max inst WS	−0.288	Mean WS	−0.308	Mean RH	0.214	Mean SP	0.149
	CO (ppm)	0.621		Max Temp	−0.122	Max inst WSD	0.045	WFS	−0.309	Mean LAP	0.143	Min RH	0.091
	NO_2 (ppm)	0.667		Mean Temp	−0.142	Max WS	−0.254	Max freq WD	0.07	Max SP	0.155		
	SO_2 (ppm)	0.559		Daily prep	−0.144	Max WSD	0.049	Mean DP	−0.054	Min SP	0.151		

Appendix 2. $\mathbf{PM_{2.5}}$ concentration distribution maps (Before and after the COVID-19 outbreak: 2019 & 2020)

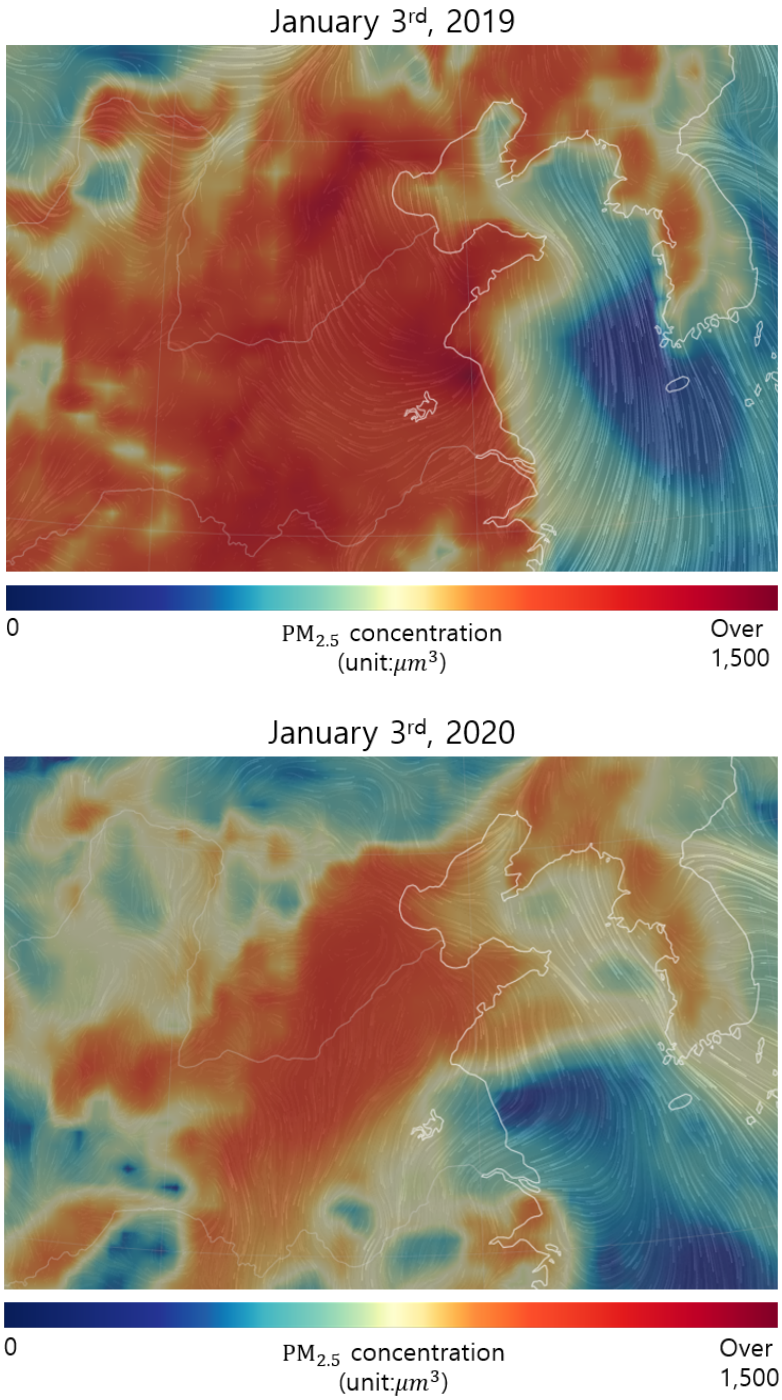


Figure A1 $\mathbf{PM_{2.5}}$ concentration distribution maps

(January **3rd,** 2019 **&** January **3rd,** 2020**)**

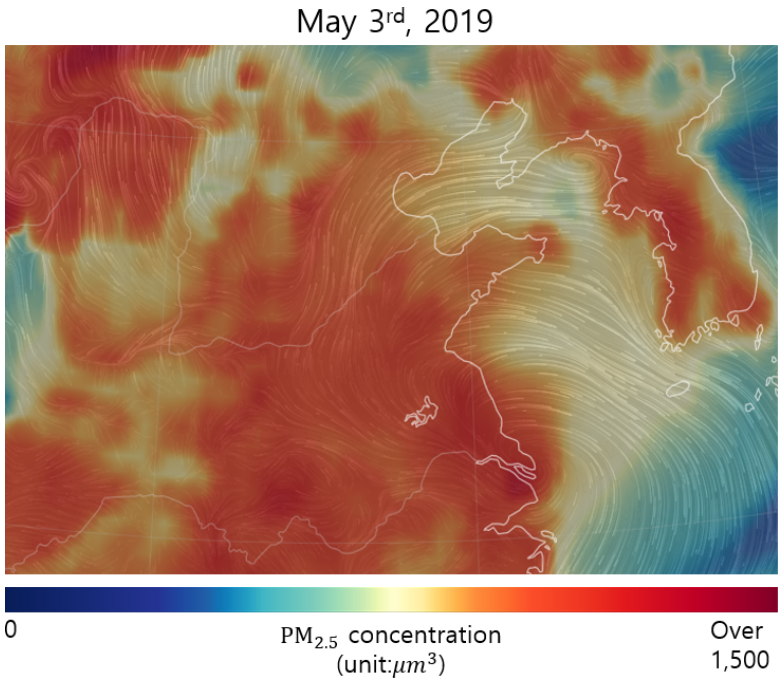


Figure A2 **$PM_{2.5}$** concentration distribution maps
(May **3rd,** 2019 **&** May **3rd,** 2020**)**

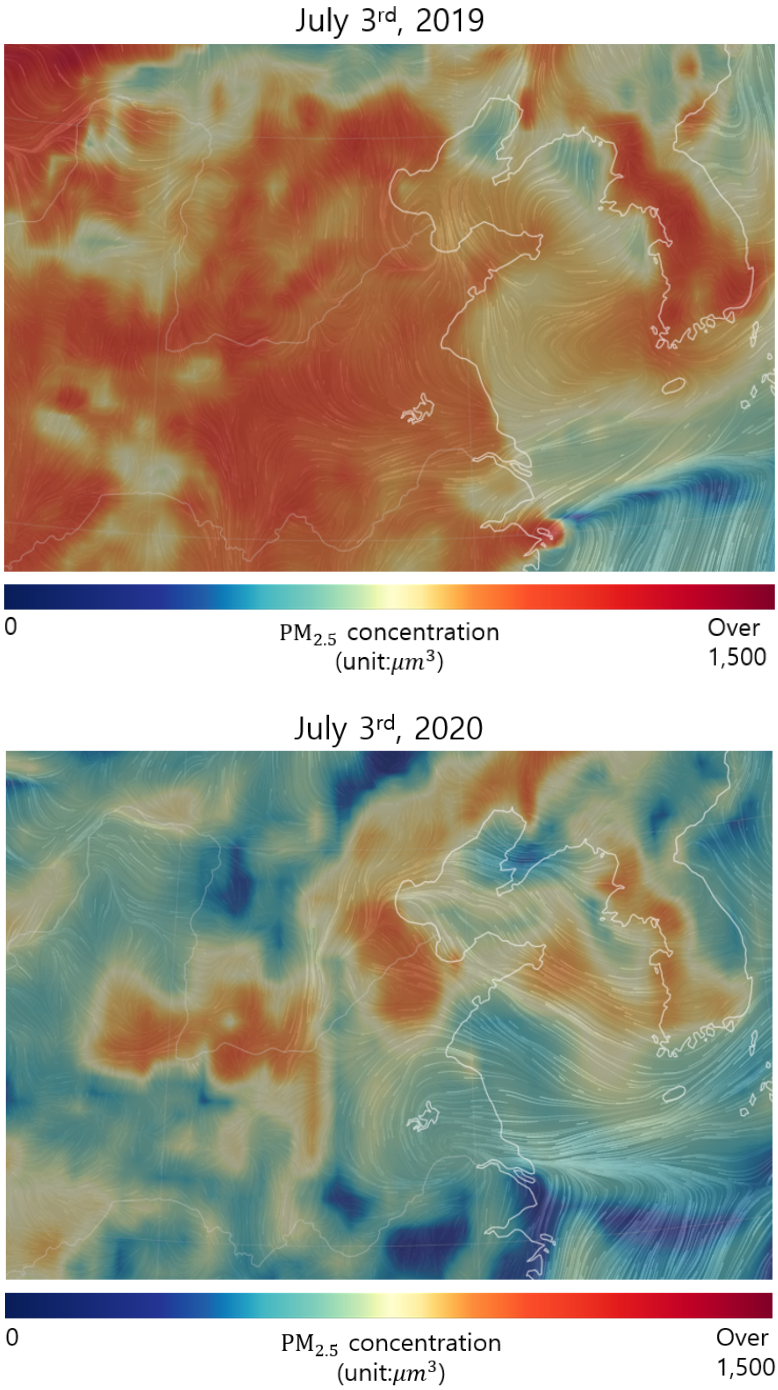


Figure A3 **$PM_{2.5}$** concentration distribution maps
(July **3rd,** 2019 **&** July **3rd,** 2020**)**

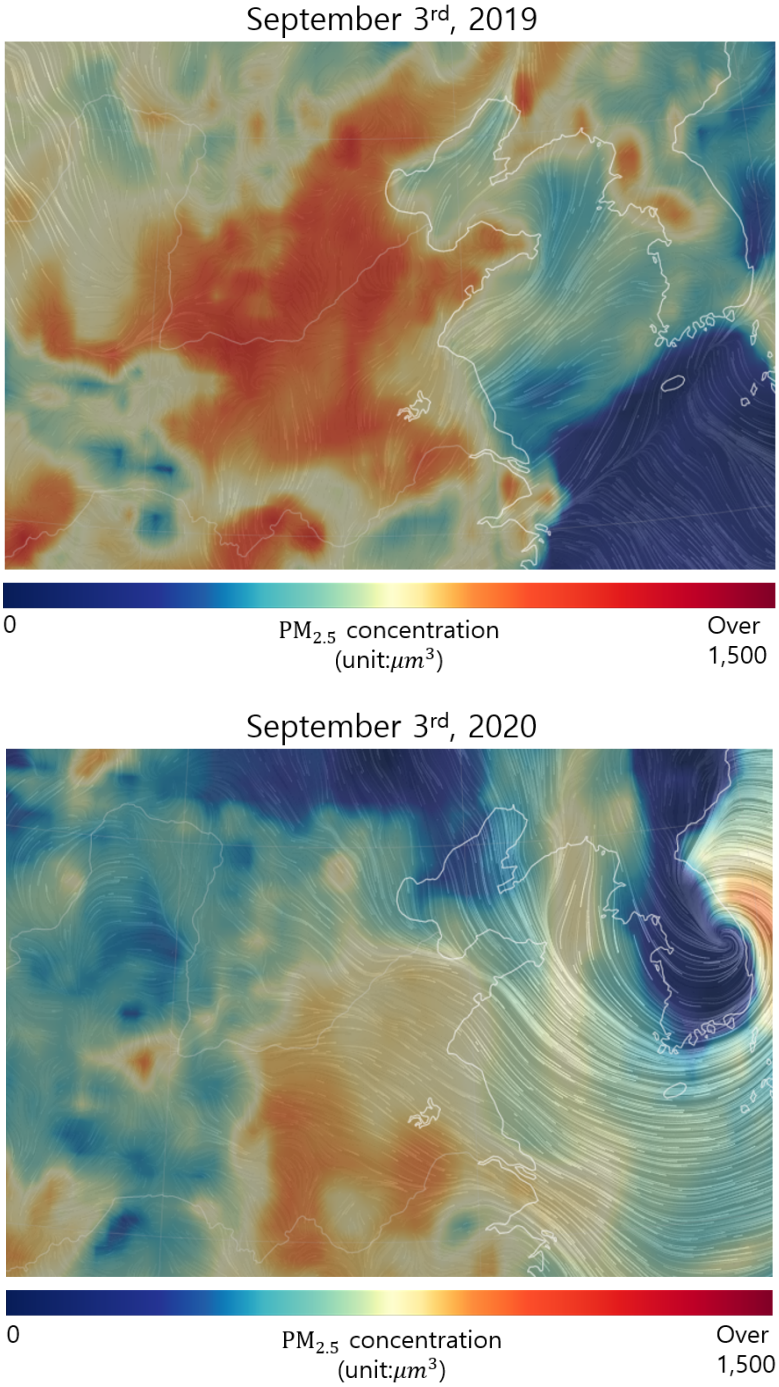


Figure A4 **$PM_{2.5}$** concentration distribution maps
(September **3rd,** 2019 **&** September **3rd,** 2020**)**

Appendix 3. The meteorological data distribution

of Seoul

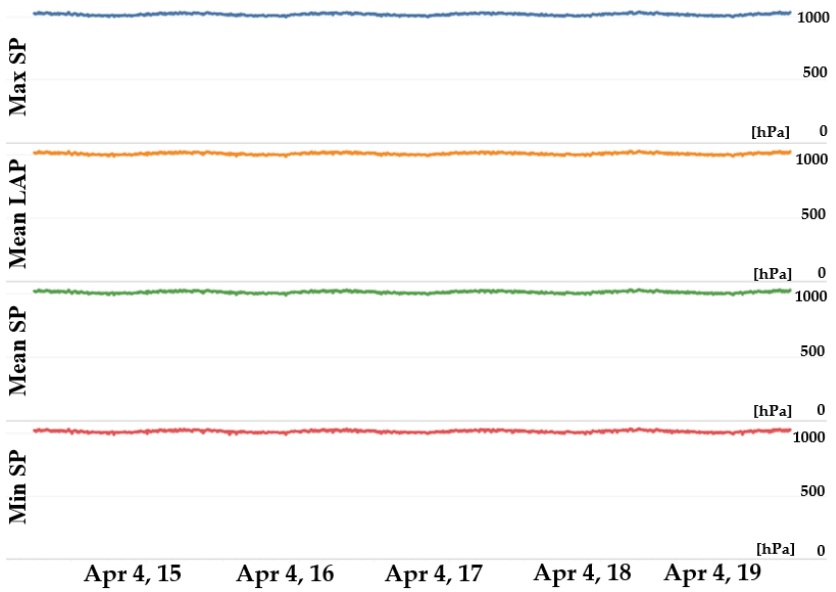


Figure A5 The meteorological data of Seoul
(atmospheric data, sea-level pressure data)

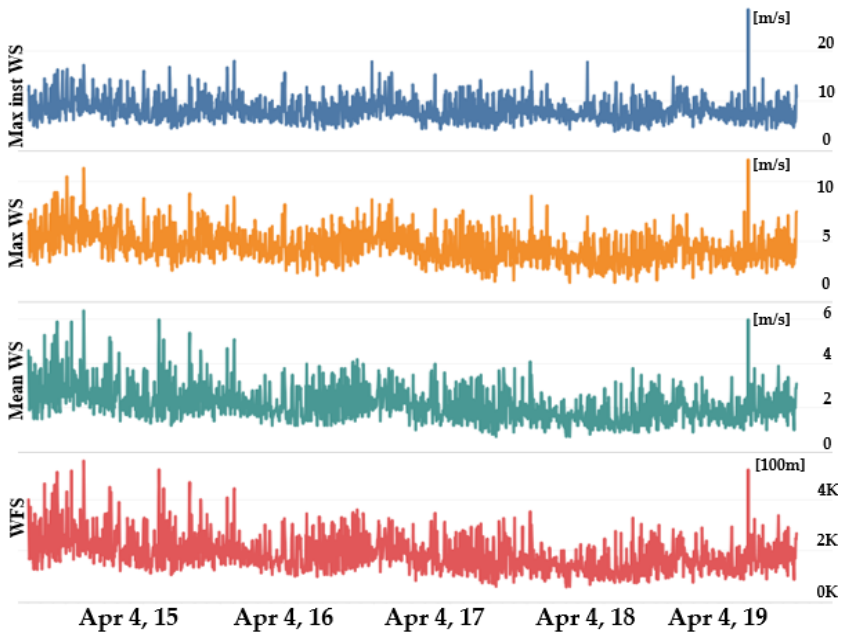


Figure A6 The meteorological data of Seoul (wind data)

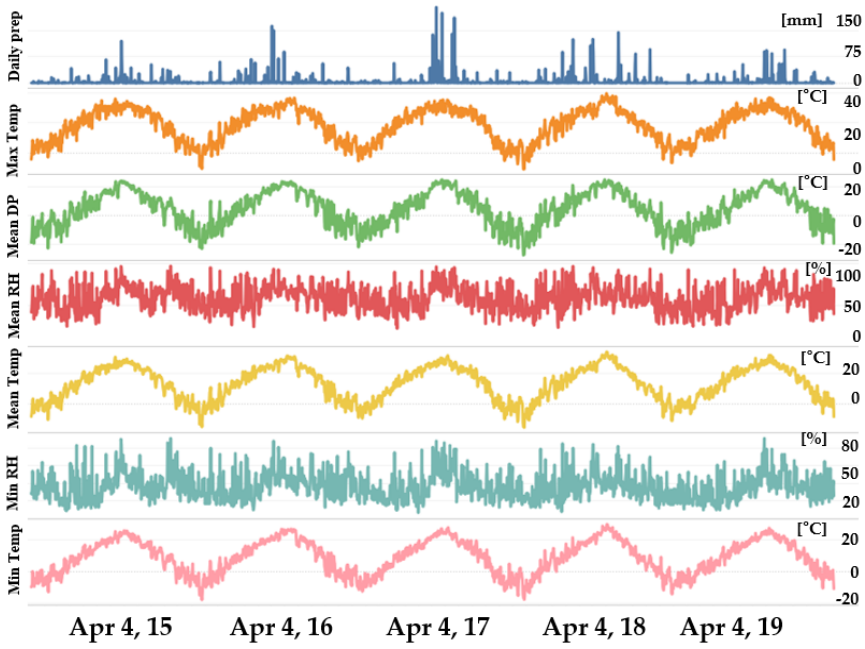


Figure A7 The meteorological data of Seoul
(temperature data, relative humidity data, precipitation data)

Appendix 4. The $\mathbf{PM_{2.5}}$ concentration prediction in each city by two cases

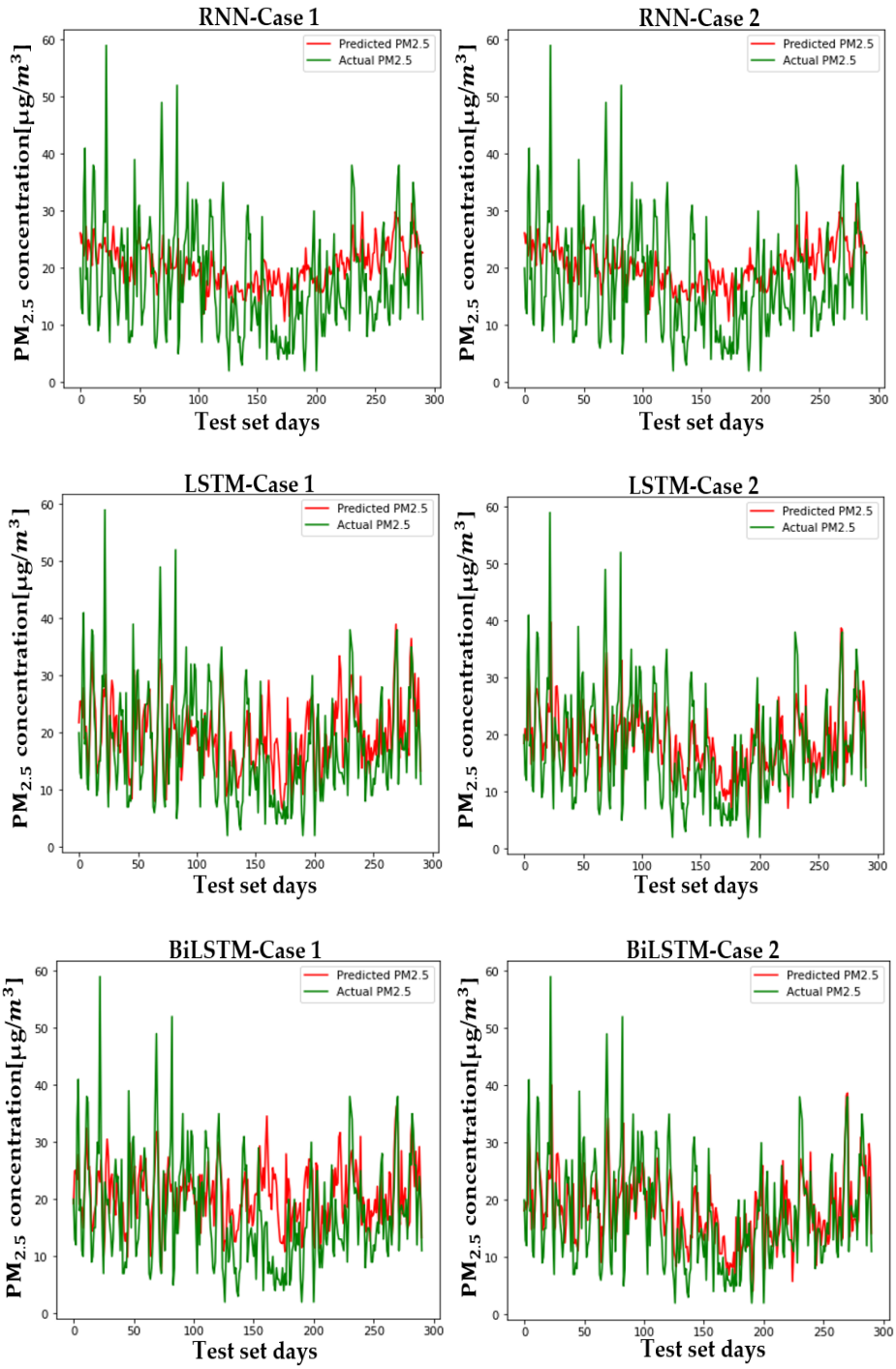


Figure A8 The $\mathbf{PM_{2.5}}$ concentration prediction in Gwangju by two cases

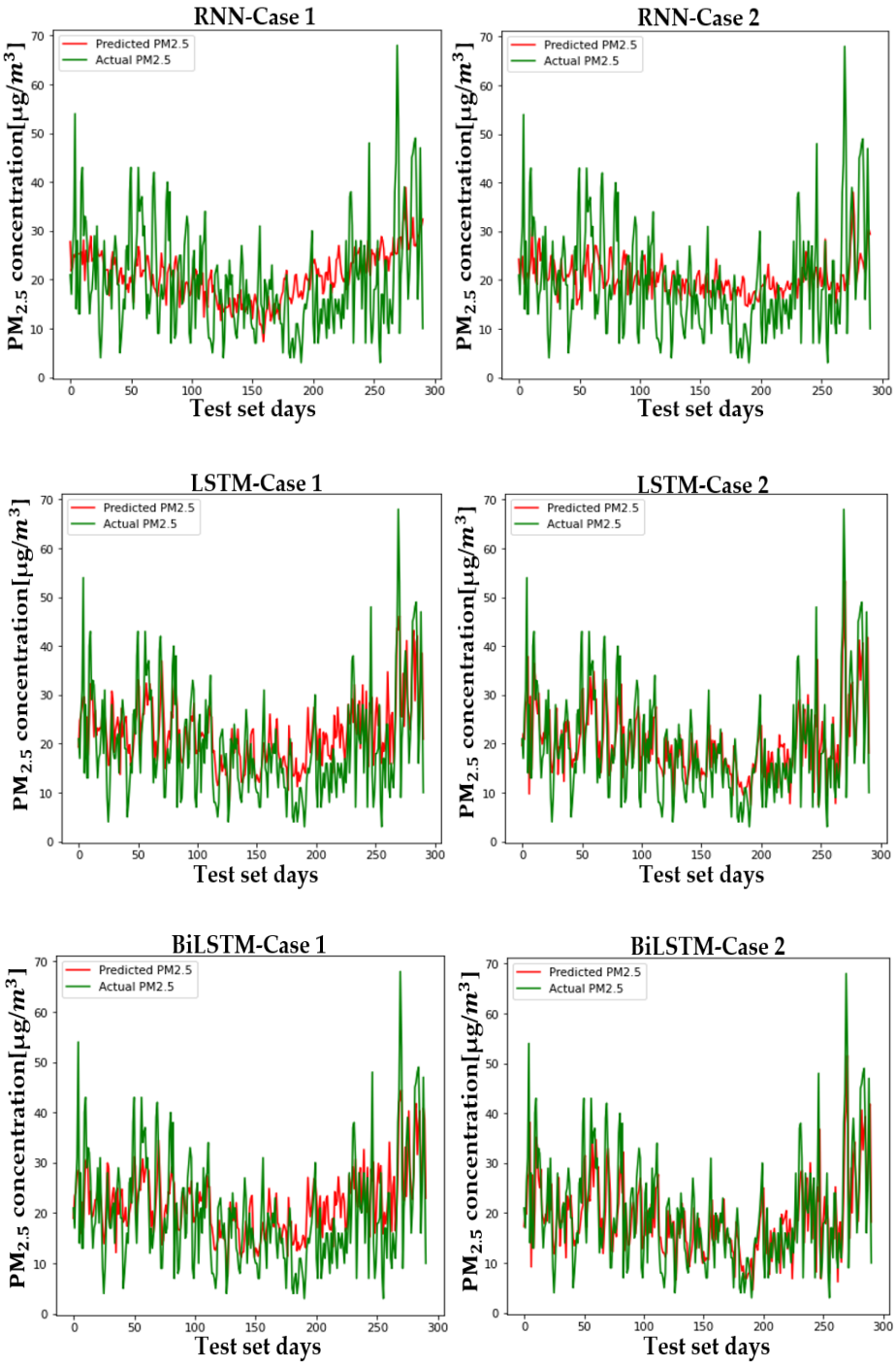


Figure A9 The $\mathbf{PM_{2.5}}$ concentration prediction in Daegu by two cases

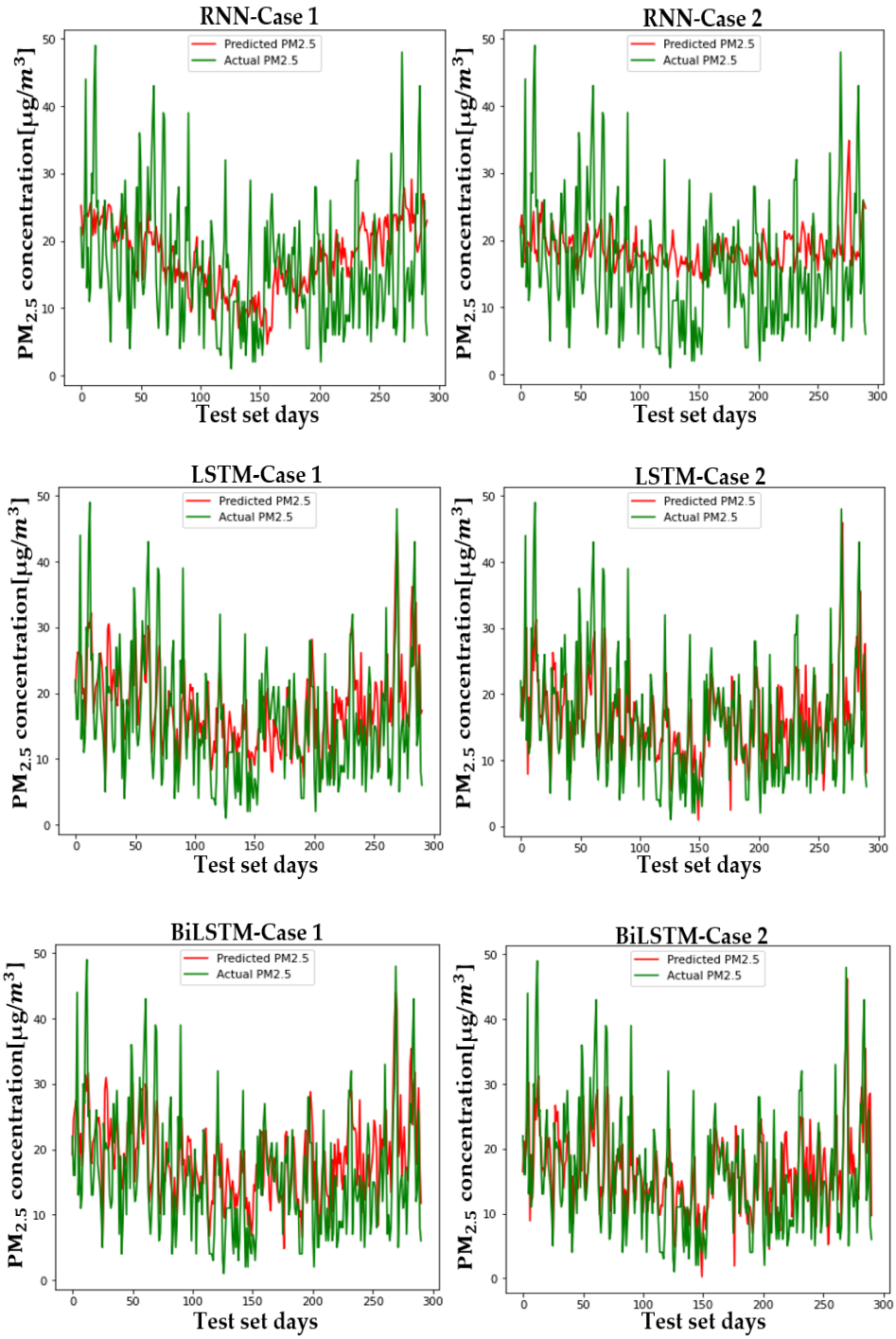


Figure A10 The $\mathbf{PM_{2.5}}$ concentration prediction in Daejeon by two cases

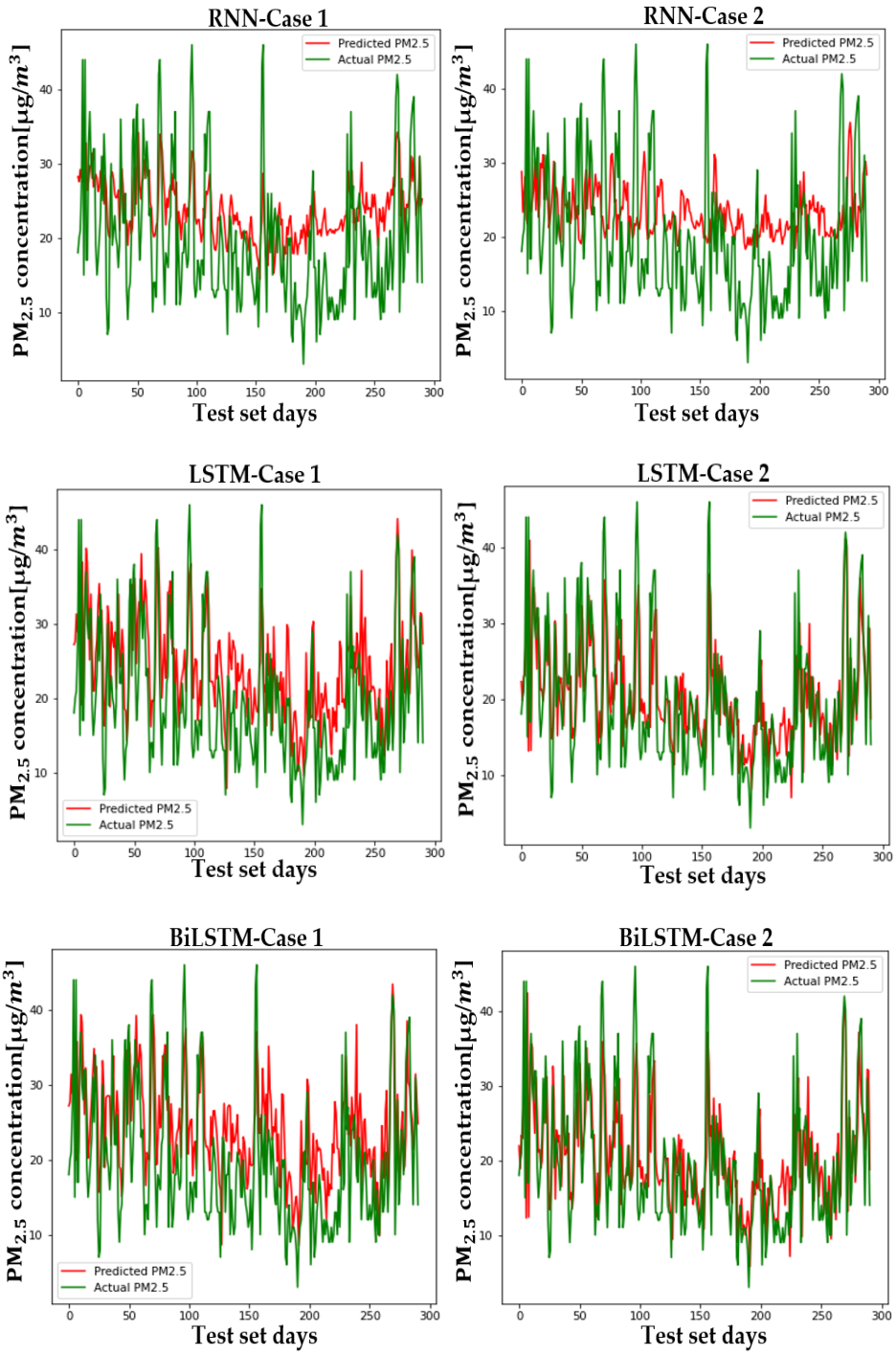


Figure A11 The $\mathbf{PM_{2.5}}$ concentration prediction in Busan by two cases

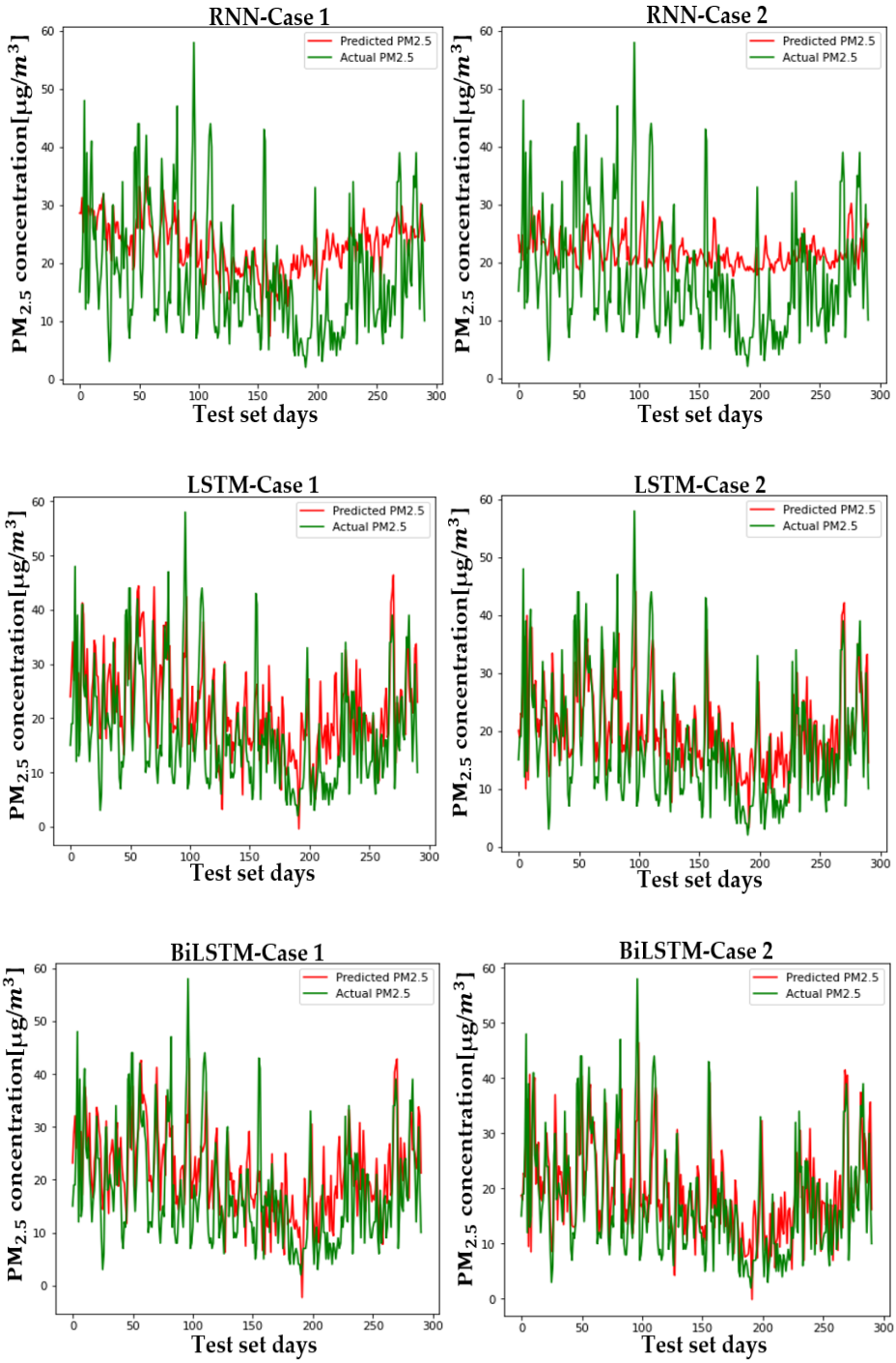


Figure A12 The $\mathbf{PM_{2.5}}$ concentration prediction in Ulsan by two cases

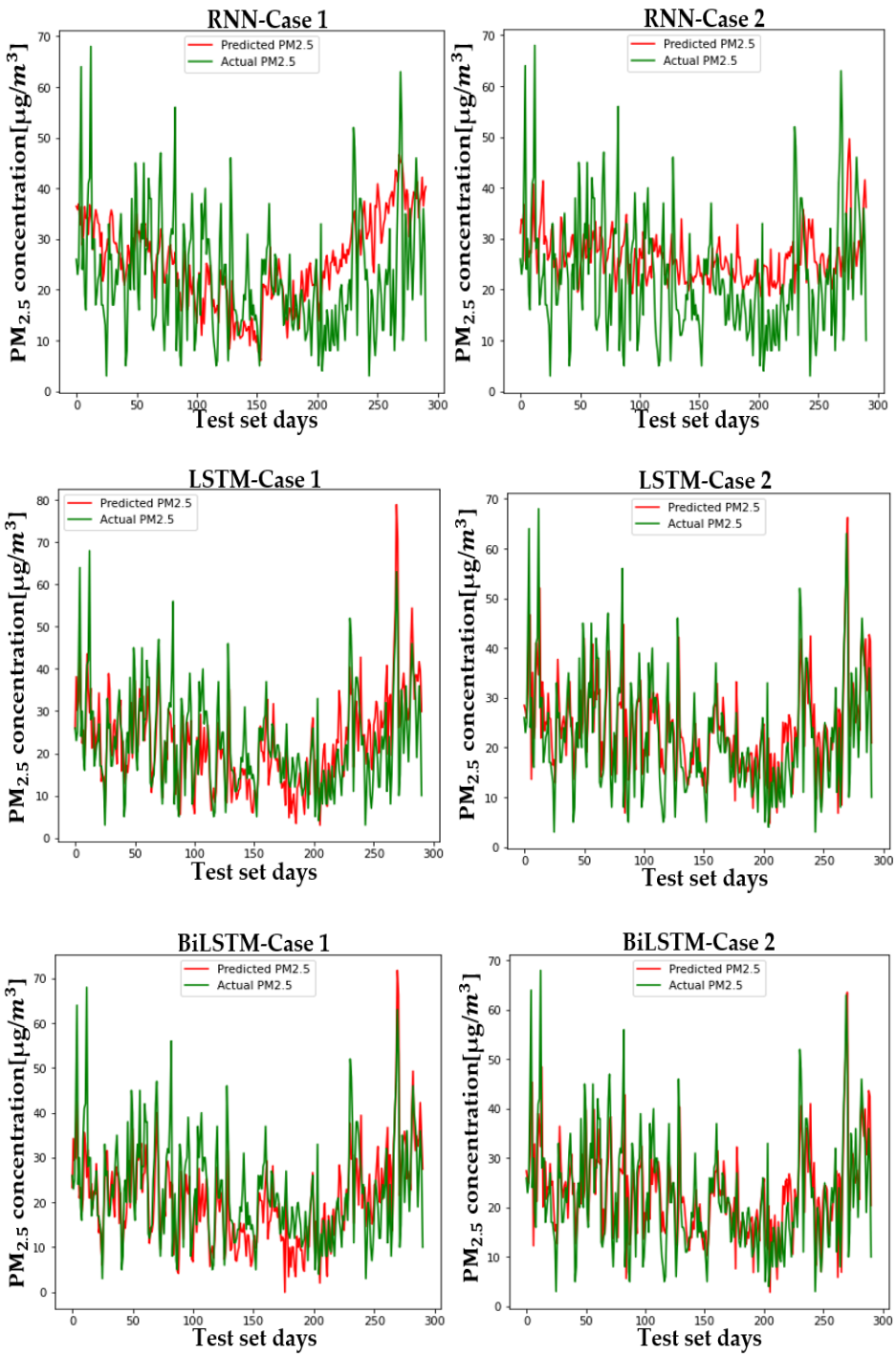


Figure A13 The $\mathbf{PM_{2.5}}$ concentration prediction in Wonju by two cases

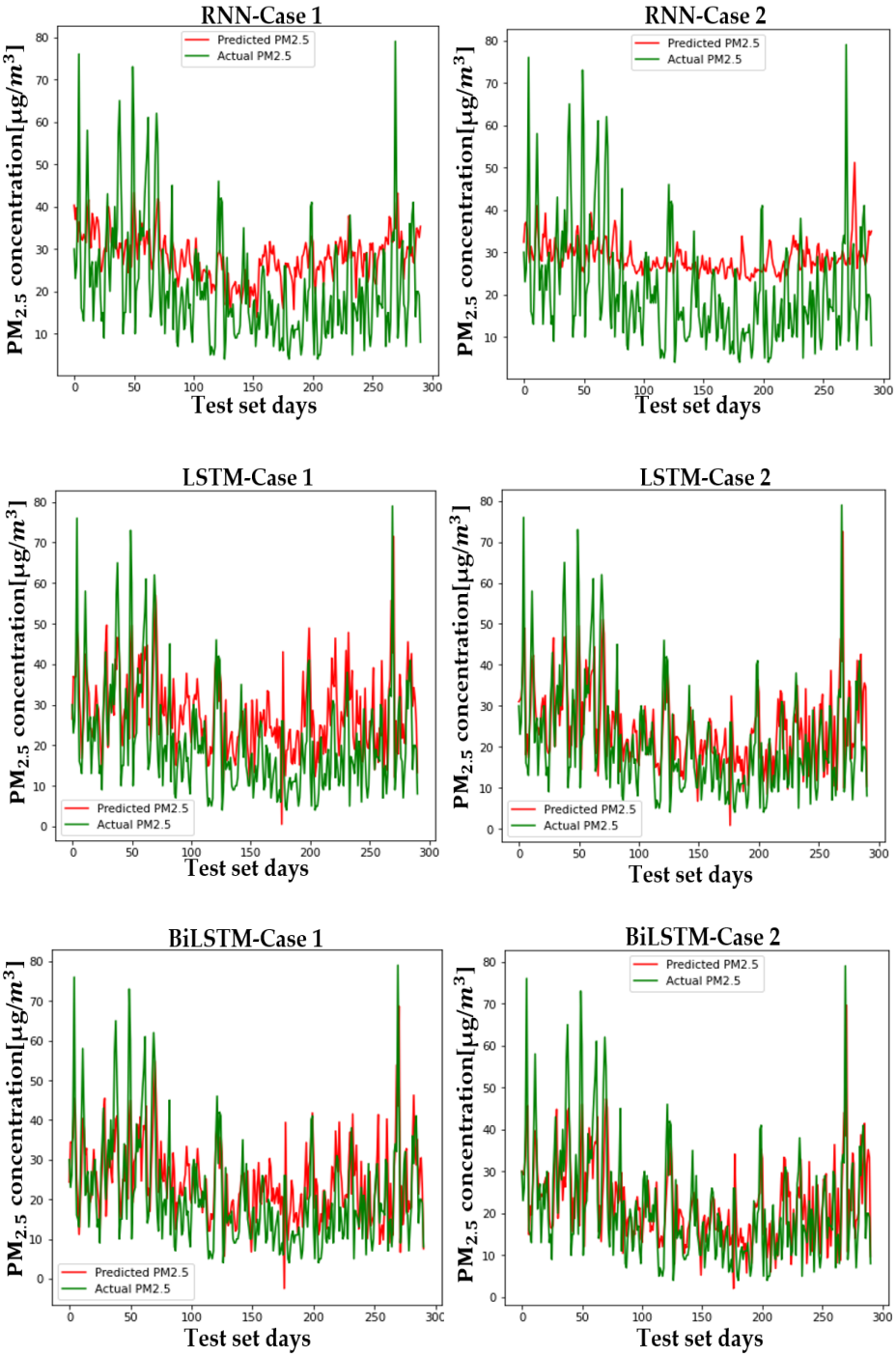


Figure A14 The $\mathbf{PM_{2.5}}$ concentration prediction in Incheon by two cases

국문초록

딥러닝 시계열 알고리즘 기반 초미세먼지($PM_{2.5}$) 예측 모델의 주성분분석 적용

최 상 원
농경제사회학부 지역정보학전공
서울대학교 대학원

초미세먼지는 세계 주요도시에서 발생하고 있는 대기오염 문제이다. 한 국가의 초미세먼지는 국가 내부 요인에만 영향을 받는 것이 아닌 인접국가의 대기질에도 영향을 받는다. 그러므로 초미세먼지 저감정책 및 계획 수립을 위한 예측치 산출에는 국가 내/외부 자료를 활용할 필요가 있다. 그러나 관측치에 비해 비교적 많은 변수 데이터셋은 차원의 저주를 유발할 수 있으며, 이는 예측력 저하의 주된 요인이 될 수 있다.

본 연구는 한국 주요 8개 도시의 일별 초미세먼지 농도를 딥러닝 시계열 모델로 예측함에 있어 해당 도시의 대기질 및 기상, 초미세먼지 농도와 한국과 근접한 중국 도시들의 초미세먼지 농도를 각각 과거 5년치 데이터를 사용하였다. 이때 발생할 수 있는 차원의 저주로 인한 예측력 하락문제 해결을 위해 데이터 셋에 주성분분석을 실시하여 고차원 데이터를 저차원 데이터로 변환하였다. 초미세먼지 예측에 있어 순환신경망, 장단기 기억모델, 양방향 장단기 기억 모델과 같은 딥러닝 시계열 모델을 사용하였으며, 각 모델의 성

능을 주성분분석을 적용한 경우와 그렇지 않은 경우로 나누어 평균제곱근오차, 평균절대오차를 활용하여 비교를 진행하였다.

그 결과 주성분분석을 적용한 장단기 기억모델의 성능은 그렇지 않은 경우보다 RMSE, MAE에서 각각 최대 16.6%, 33.3% 더 나은 성능을 보였음을 알 수 있었다. 또한 양방향 장단기 기억모델은 RMSE, MAE에서 각각 최대 16.7%, 31.6% 더 나은 성능을 보였음을 알 수 있었다. 이를 통해 주성분분석의 적용은 딥러닝 시계열 모델 성능 향상을 도출할 수 있음과 동시에, 향후 초미세먼지 저감정책 수립에 있어 보다 정확한 예측치를 제공할 수 있음을 알 수 있었다.

주요어: 주성분분석, 초미세먼지, 순환신경망, 장단기기억모델, 양방향장단기기억모델
학번 : 2019-26828