

Article

Not peer-reviewed version

---

# A Study on LSTM-Based PM2.5 Forecasting with Increased Training Data Volume in Seoul, Korea

---

[Seoyeon Kim](#) , [Seung-Hee Eun](#) <sup>\*</sup> , [Ki-Hong Shin](#) , [Sung-Chul Hong](#) , [Jae-Bum Lee](#)

Posted Date: 26 June 2025

doi: 10.20944/preprints202506.2098.v1

Keywords: artificial intelligence (AI); long short-term memory (LSTM); PM2.5; air quality forecast; big data



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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

*Article*

# A Study on LSTM-Based PM<sub>2.5</sub> Forecasting with Increased Training Data Volume in Seoul, Korea

Seoyeon Kim <sup>1</sup>, Seung-Hee Eun <sup>1,\*</sup>, Ki-Hong Shin <sup>1</sup>, Sung-Chul Hong <sup>1,2</sup> and Jae-Bum Lee <sup>1</sup>

<sup>1</sup> Air Quality Forecasting Center, National Institute of Environmental Research (NIER), Incheon, 22689, Republic of Korea

<sup>2</sup> Department of Environmental Engineering, Inha University, Incheon, 22212, Republic of Korea

\* Correspondence: eunshee@korea.kr

## Abstract

PM<sub>2.5</sub> air pollution has become a critical environmental concern in Korea, requiring accurate forecasting systems for effective air quality management and public health protection. Numerical models face limitations including high computational costs, complex processes, and parameterization uncertainties. This study developed an LSTM-based PM<sub>2.5</sub> forecasting model for the Seoul area using integrating observational data and numerical model outputs to overcome the limitations. To address input data gaps in future time steps where observational data are unavailable, WRF-CMAQ model outputs are incorporated as supplementary inputs. Three LSTM models with different training periods are developed: T3V19(3-year), T5V21(5-year), and T6V22(6-year training). Performance evaluation during January-March 2023 demonstrated significant improvements over the CMAQ model. The T6V22 model achieves a 96% improvement in NMB (1.3 vs. 32% for CMAQ), meeting “Goal” benchmark criteria. The correlation coefficient increased from 0.79 to 0.85, while NME decreases from 43.3% to 22.6%. LSTM models consistently outperformed conventional numerical models across all forecast lead times (D+0, D+1, D+2). The results suggest that as input data volume increases, model performance becomes more superior and enables more stable air quality predictions, providing a promising framework for operational forecasting systems.

**Keywords:** artificial intelligence (AI); long short-term memory (LSTM); PM<sub>2.5</sub>; air quality forecast; big data

## 1. Introduction

Air pollution has now become one of the most significant environmental concerns due to its detrimental effects on human health such as disease and mortality and the environment in the world, especially in Korea (WHO, 2016). Therefore, real-time PM<sub>2.5</sub> concentration predictions in advance have significant practical and social values, which play a significant role in making atmospheric management decisions to control air pollution and are important information needed to facilitate sustainable development.

PM<sub>2.5</sub> forecasting methods can be classified into numerical models and statistical models. Numerical models for forecasting PM<sub>2.5</sub> are mainly used the Community Multiscale Air Quality (CMAQ) model (Zhang et al., 2022; Chang et al., 2021; TT de Almeida et al., 2018; Tao et al., 2020), CMAS (Flemming et al., 2017), WRF-Chem models (Reyes-Villegas et al., 2023; Li et al., 2023; Ma et al., 2018; Wang et al., 2020) and GEOS-Chem models (Yang and Zhao, 2023; Kumar et al., 2021; Travis et al., 2022). However, due to the complex pollutant diffusion mechanism, some problems produced by these models result in usage limitations, such as colossal calculation workload, the complexity of the process, uncertainty of parameters (Gao and Zhou, 2024; Huang et al., 2021; Ferreira et al., 2020; Sokhi et al., 2021; Baklanov et al., 2020).

Statistical models for forecasting PM<sub>2.5</sub> are based on statistics and use historical time series data to predict future PM<sub>2.5</sub> concentration. As historical observational data of air quality continues to

accumulate and the demand for more efficient and accurate forecasting grows, machine learning methods—recognized for their strong predictive capabilities—have emerged in forecasting PM<sub>2.5</sub> in recent years research (Do et al., 2023; Karimian et al., 2019; Vignesh et al., 2023). Compared with numerical models, the statistical prediction models not only avoid the complicated mechanism so that the model forecasts PM<sub>2.5</sub> concentration faster, and the cost of each prediction can be reduced but also can achieve an almost improved level of PM<sub>2.5</sub> concentration prediction accuracy of the physical prediction models.

Recently, long short-term memory neural network (LSTM; Hochreiter and Schmidhuber, 1997) has been used extensively for forecasting time-series data as a type of machine learning method due to its capability of simulating long and short-term tendencies simultaneously. Qadeer et al. (2020) shows that LSTM demonstrated the outstanding accuracy compared to gradient tree boosting models, RNN, and CNN in forecasting PM<sub>2.5</sub> concentrations. Li et al. (2017) constructed a method of using the historical concentrations of all sites as the inputs of the LSTM and integrating auxiliary variables by a fully connected layer. As prediction of PM<sub>2.5</sub> concentration increasingly adopts deep learning (DL) approaches, it is becoming evident that the performance and applicability of DL-based models depend more critically on quality and consistency of datasets than on the choice of algorithms.

In this study, we develop a model employing LSTM algorithm to overcome the limitations of numerical models and improve the accuracy of PM<sub>2.5</sub> forecasts. We evaluated the model's performance with respect to datasets enhancement for PM<sub>2.5</sub> forecasting.

## 2. Input Data Collection and Preprocessing

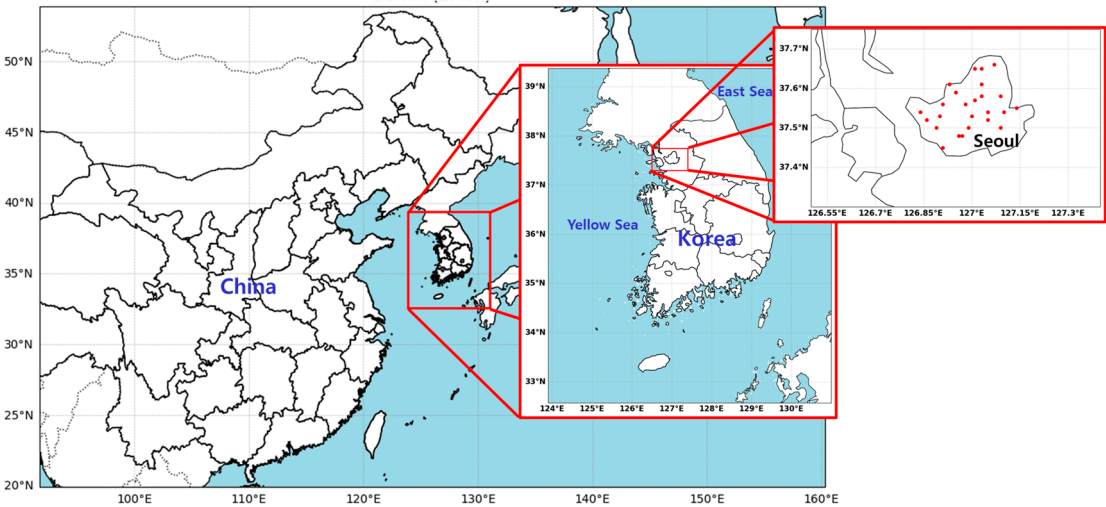
### 2.1. Observational Meteorological and Air Quality Data

The meteorological data used in the model is based on real-time ASOS (Automated Surface Observing System) data obtained via API system of KMA (Korea Meteorological Administration). Although ASOS instruments collect a wide range of meteorological data, this study selected six key variables that are most relevant to the formation and dispersion of PM<sub>2.5</sub>. The first variable considered is air temperature, which exhibits a distinct diurnal cycle and can influence the increase or decrease of PM<sub>2.5</sub> (Li et al., 2022). Increasing temperatures contribute to the enhanced chemical formation and accumulation of PM<sub>2.5</sub>. Dew temperatures and relative humidity, which serve as indicators of atmospheric moisture content, have been suggested to positively influence PM<sub>2.5</sub> concentrations, as increased atmospheric humidity provides favorable conditions for aerosol formation and hygroscopic growth (Cheng et al., 2015). Wind speed, one of the key meteorological variables influencing the formation and dispersal of PM<sub>2.5</sub>, is negatively correlated with PM<sub>2.5</sub> concentrations (Yu et al., 2018, Eun et al., 2020). Fujino and Miyamoto (2022) suggest that the scavenging efficiency was 2.03% for light precipitation, 28.15% for moderate precipitation, and 26.75% for heavy precipitation, suggesting that moderate or stronger rainfall is more effective in removing PM<sub>2.5</sub> concentration. This implies that wet deposition processes such as rainout and washout play a significant role in reducing PM<sub>2.5</sub> concentration during precipitation events. Precipitation duration, along with intensity, exerts a direct influence on PM<sub>2.5</sub> concentration and should therefore be regarded as a key meteorological variable in PM<sub>2.5</sub>. The synoptic system which influences PM<sub>2.5</sub> can be represented by sea level pressure (SLP). Chang et al. (2020) identified high-pressure systems as dominant synoptic patterns during high PM<sub>2.5</sub> episodes in Seoul accounting for 83% of the cases. These conditions were associated with atmospheric stagnation and enhanced secondary aerosol formation, contributing to elevated PM<sub>2.5</sub> concentration. In this study, the 6 variables are utilized as input for machine learning model (Table 1).

**Table 1.** Variables of training and validation data for the LSTM model.

Data Type		Variable
Observation	Meteorological Variables (surface)	Temperature (K), Dew point (K), Wind component (U, V; m/s), Sea level pressure (hPa), Precipitation (mm), Solar radiation (W/m <sup>2</sup> )
	Air Quality variables	PM10 (µg/m <sup>3</sup> ), PM2.5 (µg/m <sup>3</sup> ), O <sub>3</sub> (ppm), NO <sub>2</sub> (ppm), CO (ppm), SO <sub>2</sub> (ppm)
Numerical data	Meteorological Variables (surface)	Temperature (K), Dew point (K), Wind component (U, V; m/s), Sea level pressure (hPa), Precipitation (mm), Solar radiation (W/m <sup>2</sup> )
	Meteorological Variables (925 hPa, 850 hPa, 700 hPa, 500 hPa)	Temperature (K), , Relative humidity (%), Wind component (U, V; m/s), Geopotential height (m)
	Air Quality variables	PM10 (µg/m <sup>3</sup> ), PM2.5 (µg/m <sup>3</sup> ), O <sub>3</sub> (ppm), NO <sub>2</sub> (ppm), CO (ppm), SO <sub>2</sub> (ppm)
	Back trajectory pattern	Local wind group, Northwestern wind group, Western wind group, Northern wind group, Short flow
	• Cosine Similarity (1000 hPa, 925 hPa, 850 hPa, 700 hPa, 500 hPa, 300 hPa)	Temperature (K), Relative humidity (%), Wind components (U, V; m/s), Geopotential height (m)

Observational air quality data is obtained from both urban ambient monitoring stations and background sites (Figure 1). SO2 and NO2 serve as precursors to PM2.5 and directly affect PM2.5 concentration. Ozone is a secondary pollutant that is primarily formed through photochemical reactions. It is generated when nitrogen oxides (NOx) react with VOCs under the influence of sunlight. NOx and SO2 can undergo chemical reactions in the presence of O3. These reactions result in the formation of nitrate and sulfate. Nitrate and sulfate are recognized as major components of PM2.5. In addition, VOCs can react with O3. This process contributes to the production of secondary organic aerosols. CO can react with hydroxyl radicals Known as OH. Through this reaction, O3 is produced by oxidation. As a result, CO indirectly contributes to the formation of secondary aerosols and leads to an increase in PM2.5 concentration.



**Figure 1.** The study area with WRF-CMAQ domain for 27 km horizontal grid. The red dots donate observed sites (25) of air quality in Seoul.



## 2.2. Numerical Models

WRF (Weather Research and Forecasting) model is a mesoscale numerical model designed to simulate complex atmospheric phenomena. It incorporates a variety of physical and dynamical parameterizations and utilizes high resolution of land use data. Through these features, the model can capture regional characteristics and provide detailed atmospheric prediction (Skamarock et al., 2019). The WRF model in this study is used for simulating meteorological prediction data including air temperature, relative humidity, precipitation, geopotential height, pressure, and wind speed (U and V components) with altitude. Cosine similarity between meteorological variables is utilized to investigate the vertical structure of the atmosphere and to analyze characteristic patterns in meteorological behavior (Grell et al., 2005). Forecasting data generated by WRF model is widely recognized as important input for training AI-based prediction models, as they contribute significantly to improving predictive accuracy (Baklanov et al., 2014). The predicted input data for air quality variables are obtained using the Community Multiscale Air Quality (CMAQ) model. CMAQ incorporates a series of preprocessing steps-including ICON, BCON, MCIP, and JPROC-to prepare initial and boundary conditions. The CMAQ includes key atmospheric processes such as advection and diffusion in both horizontal and vertical directions, gas-phase and aqueous-phase chemistry, cloud mixing, aerosol dynamics and size distribution, velocity of deposition, etc. This model simulates pollutant distribution in a multivariate framework that accounts for interactions between emission sources (Skamarock and Klemp, 2008). Forecasting data from numerical models is used to prevent input gaps from the future time steps where observational data are no longer available in future time steps. DL model using both observed data and predicted data from numerical model significantly improved forecast accuracy at time steps without observed data, maintaining reduced errors (Chen et al., 2024). Input data utilizes PM<sub>10</sub>, PM<sub>2.5</sub>, O<sub>3</sub>, NO<sub>2</sub>, CO, and SO<sub>2</sub> from CMAQ outputs such as Table 1.

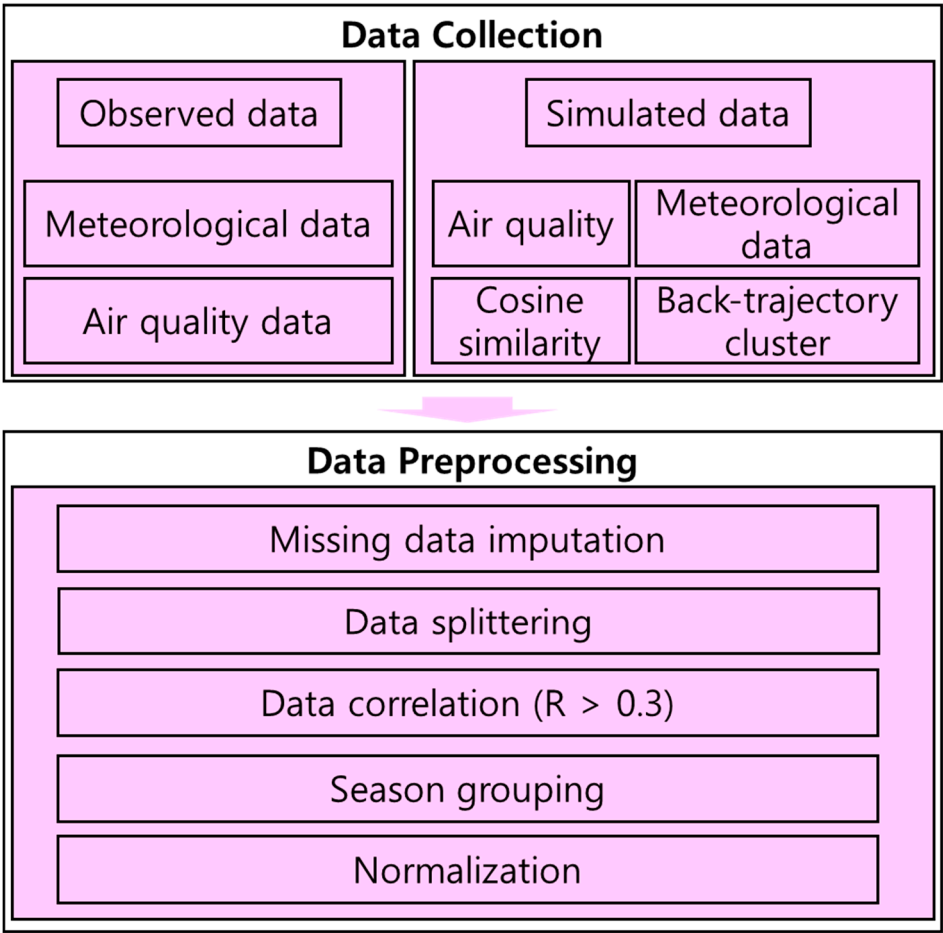
FLEXPART (FLEXible PARTicle dispersion model) is a sophisticated lagrangian transport and dispersion model suitable for the simulation of a large range of atmospheric transport processes. It supports both forward and backward simulations: forward simulations are used to calculate pollutant transport and dispersion trajectories, while backward simulations are employed to identify the source and receptor locations of the potential emission (Bakels et al., 2024). Furthermore, the resulting back-trajectories are grouped using a clustering method based on the Euclidean distances between neighboring clusters (Lee et al., 2011). The FLEXPART model enables backward trajectory analysis that more accurately captures particle dispersion and turbulent mixing processes than conventional single-line trajectory approaches.

In this study, input data for forecasting PM<sub>2.5</sub> concentration are combined with cluster analysis based on back-trajectory data from the FLEXPART model. Cluster analysis is performed on the backward trajectories generated by the FLEXPART model and identifies 5 representative transport patterns. By incorporating characteristic transport pathways and velocities for each cluster, the analysis provides valuable information for use as input data in DL-based PM<sub>2.5</sub> prediction models.

## 2.3. Data Preprocessing and Data Set

Data preprocessing consisted of data missing imputation, data splitting, data correlation, seasonal grouping and normalization to enable the machine learning models to learn (Figure 2). Observed data of meteorological and air quality often contain missing value due to equipment malfunctions, extreme weather conditions, network outages and errors in the data collection process, resulting in datasets that can be missing periods. To solve this problem, missing values of observed data are imputed by values nearby. The data splitting procedure of input data at 6 hourly interval is to divide the dataset of training and validation sets (Figure 2). PM<sub>2.5</sub> concentration has seasonal characteristics (Park et al., 2018; Jeong et al., 2022). Accordingly, the dataset is partitioned into 12 independent groups, each spanning three consecutive months. These groups are independently used for model training and evaluation. The Machine learning models work better when the range of numerical input variables is regulated to a standard value. But each meteorological and air quality

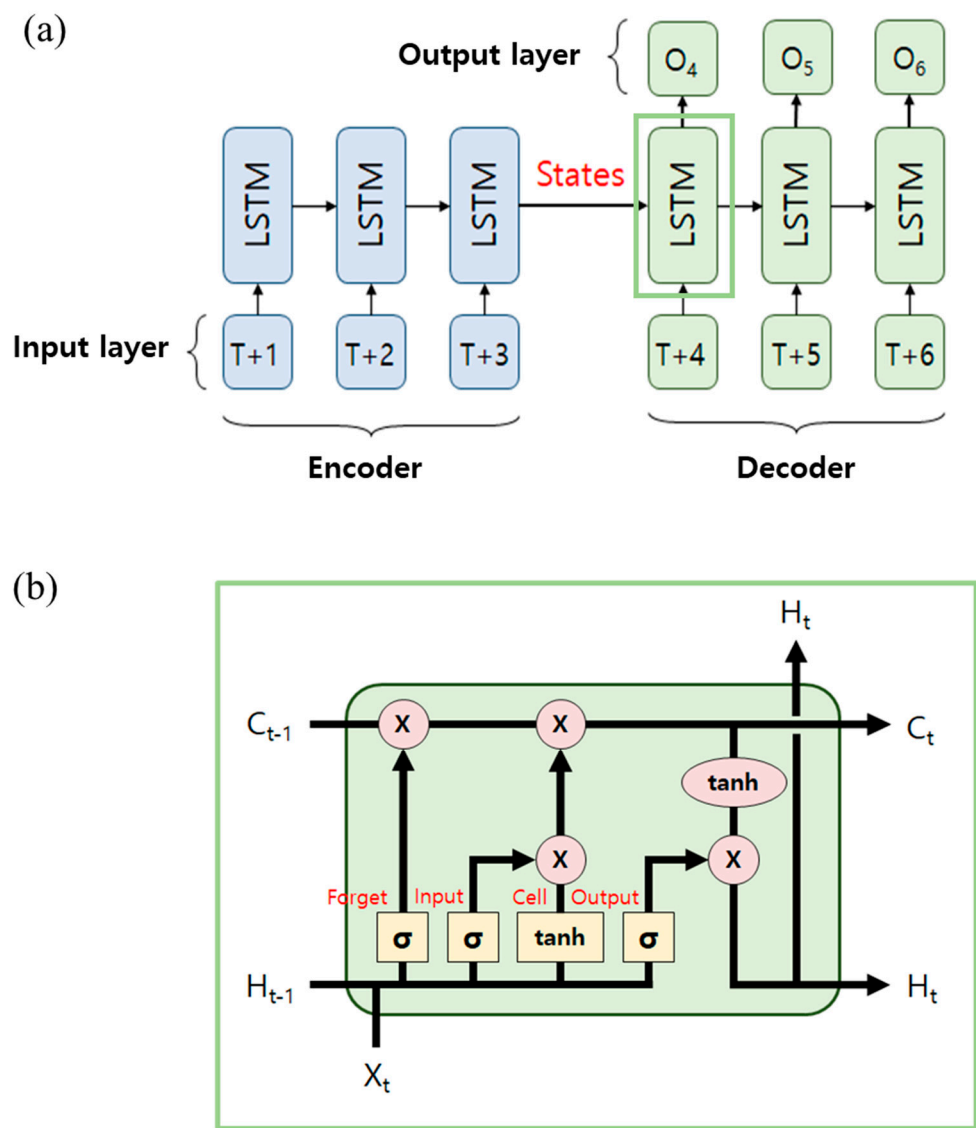
data has different dimensions and magnitudes. To reduce the error and speed up the model training, all datasets were normalized in the range of 0-1 using the minimum and maximum values of the training set. The Machine learning models work better when the range of numerical input variables is regulated to a standard value. But each meteorological and air quality data has different dimensions and magnitudes. To reduce the error and speed up the model training, all datasets were normalized in the range of 0-1 using the minimum and maximum values of the training set.



**Figure 2.** Flowchart of the training process of the PM2.5 The network structure diagram of LSTM.

As discussed in Section 2.1, PM2.5 concentrations are influenced by a variety of variables, including meteorological condition and air quality. To identify the most significant variables, we conduct a set of correlation analyses between the input variables and PM2.5 concentrations. The correlation is calculated using input variables averaged over 6-hour intervals, and a predefined threshold ( $R > 0.3$ ) is applied. Variables with correlation values exceeding the threshold are retained as input features.

Table 2 summarizes the dataset utilized in this study. An LSTM-based model is employed to predict PM2.5 concentrations using three distinct datasets. Three LSTM-based models are developed using different training and validation periods. The first model, referred to as T3V19, is trained on data from 2016 to 2018 and validated using 2019 data. The second model, denoted as T5V21, is trained in five years of data (2016-2020) and validated on 2021 data. The third model (T6V22) extended the T5V21 dataset by including data from 2021 for training and used 2022 data for validation. All models use the same set of input variables; only the temporal coverage of the datasets differed.



**Figure 3.** The basic structure of the LSTM. (a) LSTM. (b) Schematic of LSTM.

**Table 2.** The configuration for CMAQ model in this study.

CMAQ v5.0.2 Module	Option
Horizontal advection (ModHadv)	Hyamo
Vertical advection (ModVadv)	YAMO
Horizontal diffusion (ModHdiff)	Multiscale
Vertical diffusion (ModVdiff)	Eddy
Aerosol module (ModAero)	AERO5
Gas-phase chemistry solver (ModChem)	EBI
Deposition velocity calculation (ModDepv)	M3dry
Cloud module (ModCloud)	RADM
Gas-phase chemistry mechanism (Mechanism)	SAPRC99

**Table 3.** Volume of data for the LSTM model.

Integrated Model Name	T3V19	T5V21	T6V22
Target Region		Seoul	

Training Period	2016 ~ 2018 (3 years)	2016 ~ 2020 (5 years)	2016 ~ 2021 (6 years)
Evaluation Period	2019 (1 year)	2021 (1 year)	2022 (1 year)
Training Data number	450 million	760 million	910 million
Forecasting period	January 2023 ~ March 2023 (3 months)		

3. Description of LSTM Model

3.1. LSTM Model for Forecasting PM2.5 Concentration

Long Short-Term Memory (LSTM) is a sophisticated neural network introduced by Hochreiter and Schmidhuber (1997). An LSTM structure consists of input, output, and forget gates, each of which performs an individual function. The forget gate selectively removes unnecessary information from the previous cell state:

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

The sigmoid function outputs values between 0 and 1 to determine information retention, where the 128-dimensional hidden state  $h_{t-1}$  and input  $x_t$  are concatenated as input. The input gate selects information to update while generating new candidate values:

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

$$\widetilde{g}_t = \tanh(W_g[h_{t-1}, x_t + b_g])$$

This study employs the sine function instead of tanh, leveraging its periodic characteristics to better model cyclical patterns in time-series data. The cell state combines selective forgetting and input:

$$c_t = f_t \times c_{t-1} + i_t \times \widetilde{g}_t$$

Each element of the 128-dimensional cell state vector is independently updated, integrating long-term and short-term information. The Output gate regulates the final output:

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

$$h_t = o_t \times \tanh(c_t)$$

The output gate determines which portions of the cell state are output, generating the final hidden state through sine modulation. The model training employs the Backpropagation Through Time (BPTT) algorithm. To prevent overfitting, an early stopping technique is implemented to automatically terminate training when validation loss ceases to improve.

3.2. Model Evaluation

The performance statistics proposed by Emery et al. (2017) were used as the evaluation criteria. Emery et al. (2017) conducted a review of 38 peer-reviewed air quality modeling studies published in SCI-indexed journals and proposed criteria for evaluation of models using normalized mean bias (NMB), normalized mean error (NME), and the correlation coefficient (R). NMB indicates the degree to which model predictions systematically overestimate or underestimate, while a negative value indicates underestimation. NME measures the magnitude of the error by taking the absolute differences between predicted and observed values, with lower values indicating better model



performance. R represents the strength and direction of the linear relationship between the predicted and observed values, ranging from -1 to 1.

According to Emery et al. (2017), the 33rd and 67th percentiles divide the overall distribution into three performance categories: studies within the 33rd quantile successfully meet the “goal” benchmarks that the best performing models are expected to achieve; studies between the 33rd and 67th quantiles successfully meet the “criteria” that the majority of modeling studies achieve; and studies that fall outside the 67th quantile indicate relatively poor performance for that specific metric. Emery et al., (2017) recommended evaluating model performance using normalized mean bias (NMB), normalized mean error (NME), and correlation coefficient (R). According to the “Goal” benchmarks, acceptable performance is defined as  $NMB \leq 10\%$ ,  $NME \leq 35\%$ , and  $R \geq 0.7$ . The “criteria” are defined as  $NMB \leq 30\%$ ,  $NME \leq 50\%$ , and  $R \geq 0.4$ . Root Mean Square Error (RMSE) is used to evaluate the overall error of model by comparing its forecasting data to observed data, which is used to measure the standard deviation of prediction error. RMSE is more sensitive to outliers and a lower RMSE implies the improvement of the model performance. IOA is a correlation-based measure of the differences between simulated and observed data in terms of the mean and variance. IOA varies between 0 and 1 with a value of 1 indicating a perfect match, and 0 no agreement at all.

$$NMB = \frac{\sum_{i=1}^N (M_i - O_i)}{\sum_{i=1}^N O_i}$$

$$NME = \frac{1}{n} \sum_{i=1}^N \frac{|M_i - O_i|}{O_i} \times 100$$

$$R = \frac{\sum_{i=1}^N (M_i - \bar{M}) \times (O_i - \bar{O})}{\sqrt{\sum_{i=1}^N (M_i - \bar{M})^2 \times \sum_{i=1}^N (O_i - \bar{O})^2}}$$

$$RMSE = \sqrt{\frac{1}{N} \times \sum_{i=1}^N (M_i - O_i)^2}$$

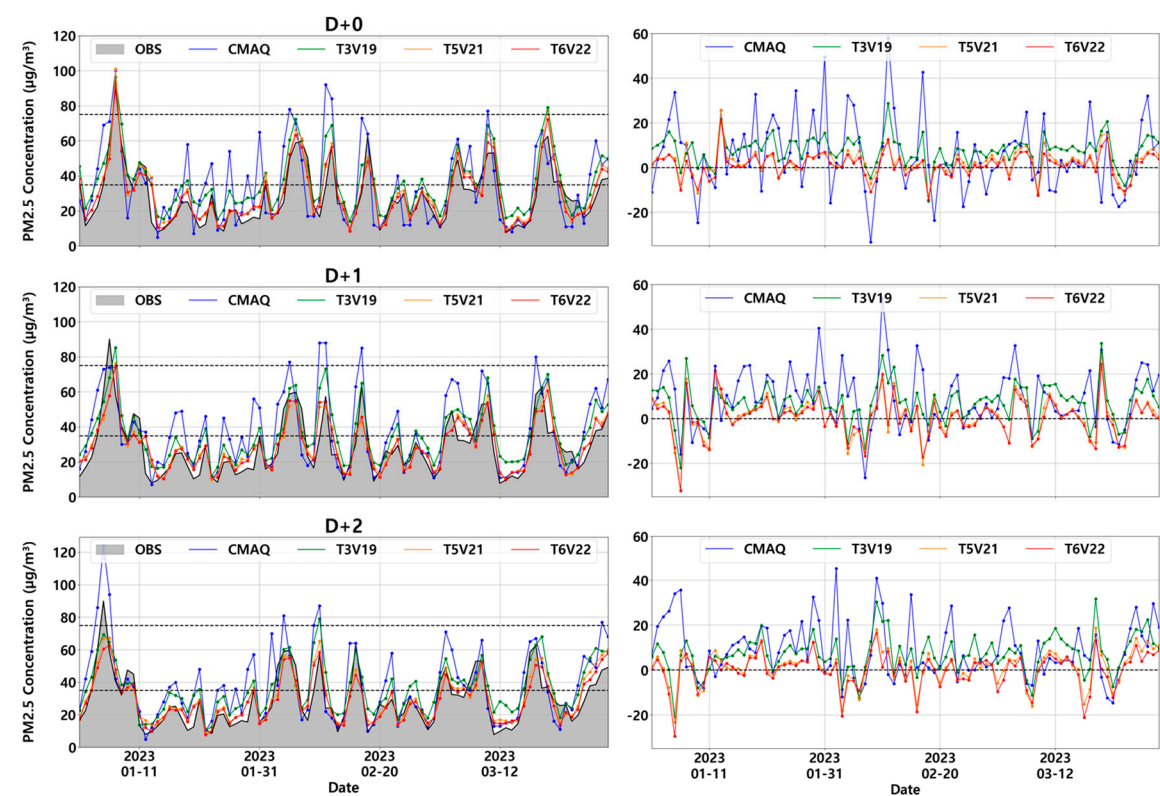
$$IOA = 1 - \frac{\sum_{i=1}^N (M_i - O_i)^2}{\sum_{i=1}^N (|M_i - \bar{O}| + |O_i - \bar{O}|)^2}$$

Here, M and O represent the modeled and observed PM<sub>2.5</sub> concentrations, respectively, and N denotes the total number of data points. This study evaluated the forecasted PM<sub>2.5</sub> concentrations for D+0, D+1, and D+2.

#### 4. Results

Evaluation of the overall performance of the LSTM models is conducted by assessing their D+0, D+1, and D+2 forecasting accuracy and comparing it with that of CMAQ over Seoul during the period from January 1 to March 30, 2023. The time series of daily forecasting and observed PM<sub>2.5</sub> concentrations for the D+0, D+1, and D+2 are shown in Figure 4. All models are capable of accurately simulating the overall observed trend, but those trained on larger datasets tend to approximate the observed values more closely. The CMAQ results noticeably overestimate relative to the DL-model outputs across all lead times. The right panel of Figure 4 shows the differences between CMAQ, T3V19, T5V21, and T6V22-simulated and observed values. Like the PM<sub>2.5</sub> concentration time series, the numerical model exhibits the largest difference, whereas the DL-based models show reduced differences as the data volume increases. In terms of predictive performance against CMAQ data, the three DL-based models (T3V19, T5V21, and T6V22) demonstrate more accurate forecasting performance over the entire forecasting period than CMAQ-based PM<sub>2.5</sub> concentrations. This

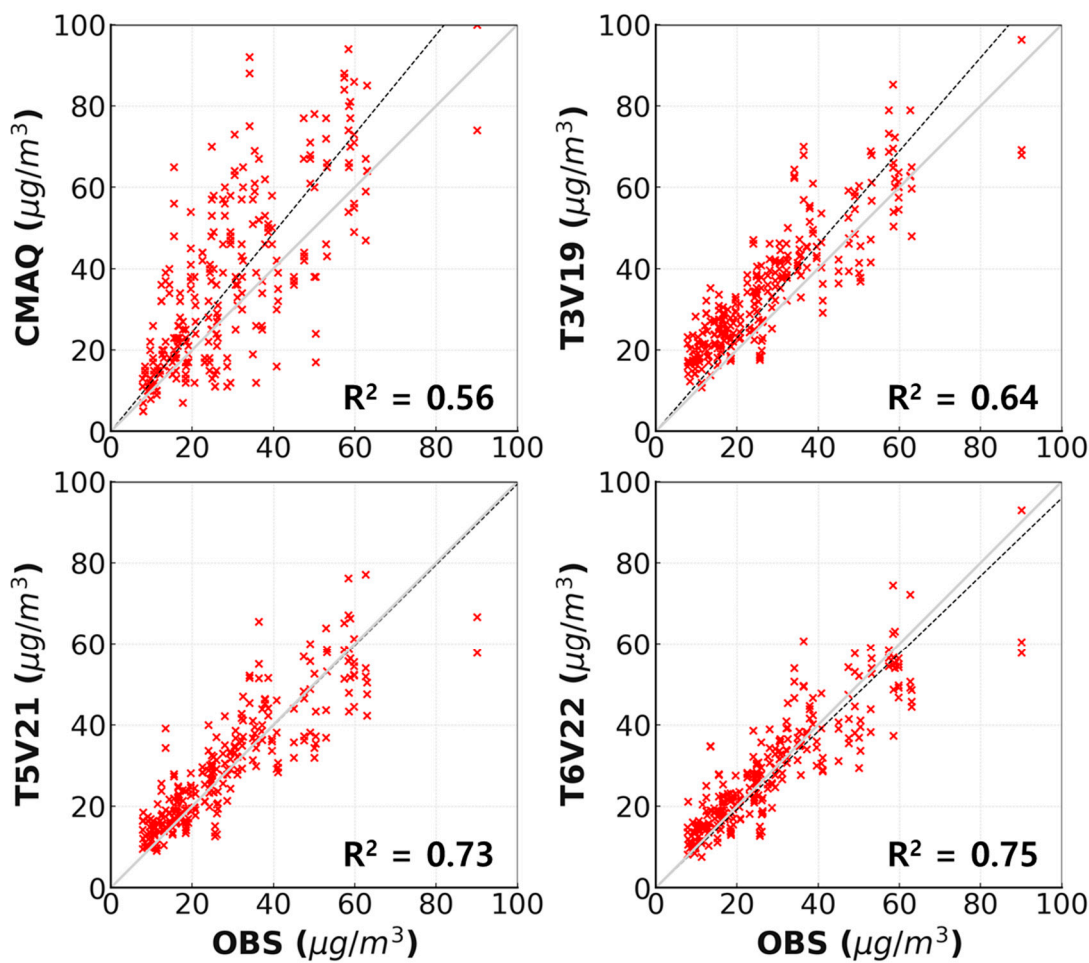
indicates that the DL-based models are capable of accurately exhibiting the characteristic rapid trends-both increases and decreases-in PM2.5 concentrations.



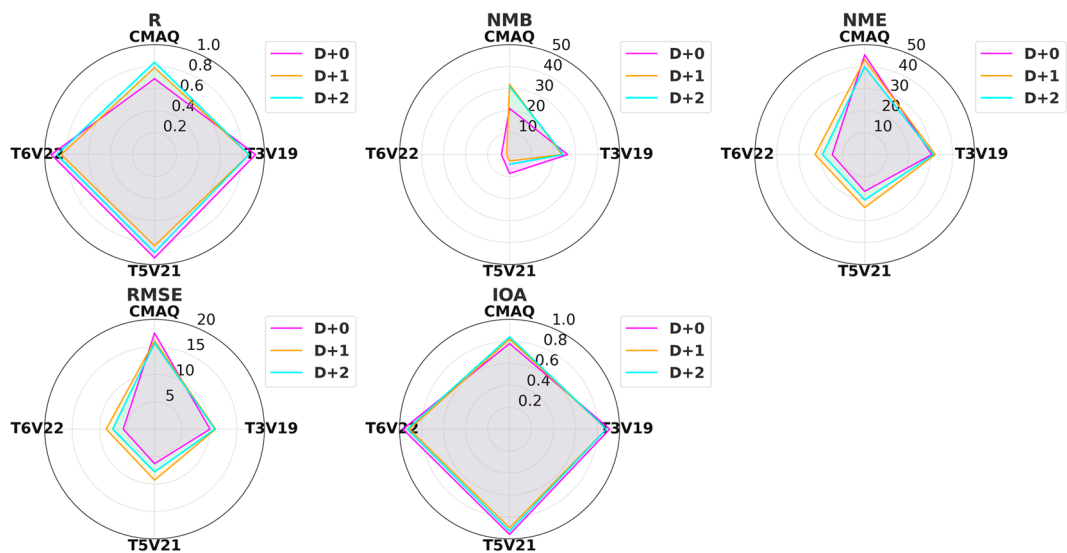
**Figure 4.** Time series of daily PM2.5 concentration over 25 sites in Seoul from January and March 2023. The shaded gray (left) represents observational PM2.5 concentration while the blue, green, orange, and red represent PM2.5 concentrations of CMAQ, T3V19, T5V21, and T6V22. The dashed lines denote ‘bad’ and ‘very bad’ grades.

Figure 5 shows scatter plots of daily averaged PM2.5 concentration between observation and models across all lead times at Seoul. The DL-models achieved the best performance, particularly T6V22, with an  $R^2$  value of 0.75. The numerical model consistently overestimates PM2.5 concentration and shows limited ability to capture temporal patterns, which is indicated by its large regression slope. In contrast, the DL-based models demonstrate progressively improved pattern reproduction as the volume of the training dataset increases.

To evaluate the accuracy of CMAQ, T3V19, T5V21, and T6V22, Table 4 and Figure 6 present evaluation metrics for PM2.5 forecasting performance over a three-day (D+0, D+1, and D+2) period using five statistical metrics (R, IOA, NMB, NME, RMSE) in Seoul. The correlation coefficient evaluates how well the numerical and LSTM models represent the temporal trend of the observed values. The correlation coefficient (R) increases from 0.79 in the CMAQ model to 0.85 in the T6V22\_6HR model.



**Figure 5.** The scatter plots between CMAQ, T3V19, T5V21, and T6V22 and observed data for PM<sub>2.5</sub> concentrations in Seoul during all lead time (D+0, D+1, and D+2). The black dash lines indicate slope.



**Figure 6.** Radar chart of D+0, D+1, and D+2 for R, NMB, NME, RMS, and IOA in Seoul.

**Table 4.** Performance comparison of the 5 different evaluation measures between T3V19 T5V21, and T6V22-based PM2.5 prediction models for daily PM2.5 predictions.

Forecast lead time	Models	R	NMB (%)	NME (%)	RMSE	IOA
D+0	CMAQ	0.69	21.0	45.3	17.5	0.78
	T3V19	0.92	26.4	30.5	10.1	0.91
	T5V21	0.94	8.6	16.7	6.3	0.96
	T6V22	0.94	3.7	14.8	5.7	0.97
D+1	CMAQ	0.79	31.9	43.2	16.1	0.82
	T3V19	0.86	23.8	32.1	11.1	0.88
	T5V21	0.83	2.9	24.1	9.3	0.90
	T6V22	0.85	1.3	22.6	8.8	0.91
D+2	CMAQ	0.84	31.0	39.9	15.7	0.84
	T3V19	0.86	24.4	31.3	11.1	0.88
	T5V21	0.89	4.4	20.6	7.8	0.93
	T6V22	0.89	-0.2	19.2	7.6	0.93

The numerical model (CMAQ) records a Normalized Mean Bias (NMB) of 32.0%. This value exceeds the “Criteria” threshold of  $\pm 30\%$  suggested by Emery et al. (2017) recommendations, which indicates low forecasting accuracy. When the training data covers three years, the NMB decreases to 23.8%. This value meets the Criteria threshold. When the data period extends to five years, the NMB drops sharply to 2.9%. This result is close to the “Goal” benchmark of  $\pm 10\%$ . With six years of data, the NMB further decreases to 1.3%. This change results in approximately 96% improvement compared to the CMAQ model. The Normalized Mean Error (NME) decreases from 43.3% in the CMAQ model to 22.6% in the T6V22\_6HR model. This reduction means that the prediction error becomes nearly half of its original value. RMSE, which penalizes large errors more heavily, shows the outstanding performance of the DL-models across all lead times compared to the numerical model. The RMSE values of the DL models range from  $5.7 \mu\text{g}/\text{m}^3$  to  $11.1 \mu\text{g}/\text{m}^3$ , indicating up to 67.4% improvement over the numerical model. This suggests that the DL-models result in fewer large deviations from observed values than the numerical model. The Index of Agreement (IOA) also increases from 0.82 to 0.91. This change shows that the agreement between the predicted and observed values improves. These results confirm that the developed DL-based model shows better performance than the CMAQ model in Seoul in terms of overall accuracy. The expansion of the dataset improves all performance indicators. The improvements in NMB and NME are particularly significant.

5. Concluding Remarks

This study develops an LSTM-based PM2.5 prediction model for the Seoul area by integrating diverse observational data and numerical model data to overcome the limitations of CMAQ numerical models in PM2.5 forecasting, including high computational costs, emission inventory limitations, uncertainties in meteorological and chemistry assumptions, and enormous time requirements for numerical model improvements. The PM2.5 prediction model for the Seoul region is developed utilizing various meteorological and air quality observed data, numerical model outputs, and secondary variables.

Statistical evaluation shows that all LSTM models (T3V19, T5V21, T6V22) perform better than the conventional CMAQ model in terms of NMB, NME, and R across all lead times (D+0 to D+2). The CMAQ model shows relatively poor performance, with NMB ranging from 21.0% to 31.9%, NME from 39.9% to 45.3%, and R from 0.69 to 0.84 exceeding the “Criteria” thresholds for NMB ( $\pm 30\%$ ) at D+1 and D+2. In contrast, the T3V19 meets the “Criteria” level, with NMB ranging from 23.8% to 26.4%, NME from 30.5% to 32.5%, and R from 0.86 to 0.92. The T5V21 and T6V22 satisfy “Goal”

benchmarks ( $\text{NMB} \leq \pm 10\%$ ,  $\text{NME} \leq \pm 35\%$ ,  $R \geq 0.7$ ) across all lead times. Specifically, T5V21 shows NMB values between 2.9% and 8.6%, NME between 16.7% and 24.1%, and R between 0.83 and 0.94. T6V22 demonstrated the best performance, with NMB values between -0.2% and 3.7%, NME from 14.8% to 22.6%, and R values between 0.85 and 0.94. Compared to CMAQ, T6V22 reduced NMB by up to 96% and NME by up to 50%, demonstrating substantial improvement in forecast accuracy. These consistent results are maintained when lead time increases.

The LSTM model developed in this study can provide fast and accurate air quality forecast information to the public while compensating for the shortcomings of numerical models. Furthermore, given high-quality observational data and numerical model data continue to accumulate in the future, the performance of deep learning-based PM<sub>2.5</sub> forecasting models is expected to be further enhanced. Future studies should focus on developing and evaluating models for various regions because each region has different characteristics of PM<sub>2.5</sub>. They should also assess the forecasting performance of the models over longer periods of ensure stability and accuracy.

**Acknowledgments:** This work was supported by a grant from the National Institute of Environmental Research (NIER), funded by the Ministry of Environment (MOE) of the Republic of Korea (NIER-2024-01-01-011).

**Conflicts of Interest:** The authors declare that they have no conflict of interest. (*required*).

## References

1. Bakels, L., Tatsii, D., Tipka, A., Thompson, R., Dütsch, M., Blaschek, M., Seibert, P., Baier, K., Bucci, S., Cassiani, M. and Eckhardt, S., 2024. FLEXPART version 11: Improved accuracy, efficiency, and flexibility. *Geoscientific Model Development*, 17(21), pp.7595-7627.
2. Baklanov, A. and Zhang, Y., 2020. Advances in air quality modeling and forecasting. *Global Transitions*, 2, pp.261-270.
3. Baklanov, A., Schlünzen, K., Suppan, P., Baldasano, J., Brunner, D., Aksoyoglu, S., Carmichael, G., Douros, J., Flemming, J., Forkel, R. and Galmarini, S., 2014. Online coupled regional meteorology chemistry models in Europe: current status and prospects. *Atmospheric Chemistry and Physics*, 14(1), pp.317-398.
4. Chang, A., 2020. Cleaning and disinfectant chemical exposures and temporal associations with COVID-19—National poison data system, United States, January 1, 2020–March 31, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69.
5. Chen, G., Chen, S., Li, D. and Chen, C., 2025. A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention. *Scientific Reports*, 15(1), p.3685.
6. Chen, Y., Huang, L., Xie, X., Liu, Z. and Hu, J., 2024. Improved prediction of hourly PM<sub>2.5</sub> concentrations with a long short-term memory and spatio-temporal causal convolutional network deep learning model. *Science of The Total Environment*, 912, p.168672.
7. Cheng, F.Y., Feng, C.Y., Yang, Z.M., Hsu, C.H., Chan, K.W., Lee, C.Y. and Chang, S.C., 2021. Evaluation of real-time PM<sub>2.5</sub> forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted PM<sub>2.5</sub> forecasts in Taiwan. *Atmospheric Environment*, 244, p.117909.
8. de Almeida Albuquerque, T.T., de Fátima Andrade, M., Ynoue, R.Y., Moreira, D.M., Andreão, W.L., Dos Santos, F.S. and Nascimento, E.G.S., 2018. WRF-SMOKE-CMAQ modeling system for air quality evaluation in São Paulo megacity with a 2008 experimental campaign data. *Environmental Science and Pollution Research*, 25, pp.36555-36569.
9. Do, K., Mahish, M., Yeganeh, A.K., Gao, Z., Blanchard, C.L. and Ivey, C.E., 2023. Emerging investigator series: a machine learning approach to quantify the impact of meteorology on tropospheric ozone in the inland southern California. *Environmental Science: Atmospheres*, 3(8), pp.1159-1173.
10. Emery, C., Liu, Z., Russell, A.G., Odman, M.T., Yarwood, G. and Kumar, N., 2017. Recommendations on statistics and benchmarks to assess photochemical model performance. *Journal of the Air & Waste Management Association*, 67(5), pp.582-598.



11. Ferreira, J., Lopes, D., Rafael, S., Relvas, H., Almeida, S.M. and Miranda, A.I., 2020. Modelling air quality levels of regulated metals: Limitations and challenges. *Environmental Science and Pollution Research*, 27, pp.33916-33928.
12. Flemming, J., Benedetti, A., Inness, A., Engelen, R.J., Jones, L., Huijnen, V., Remy, S., Parrington, M., Suttie, M., Bozzo, A. and Peuch, V.H., 2017. The CAMS interim reanalysis of carbon monoxide, ozone and aerosol for 2003–2015. *Atmospheric Chemistry and Physics*, 17(3), pp.1945-1983.
13. FLEXPART version 11 : improved accuracy, efficiency, and flexibility (2024).
14. Fujino, R. and Miyamoto, Y., 2022. PM<sub>2.5</sub> decrease with precipitation as revealed by single-point ground-based observation. *Atmospheric Science Letters*, 23(7), p.e1088.
15. Gao, X. and Li, W., 2021. A graph-based LSTM model for PM<sub>2.5</sub> forecasting. *Atmospheric Pollution Research*, 12(9), p.101150. Yang, J. and Zhao, Y., 2023. Performance and application of air quality models on ozone simulation in China—A review. *Atmospheric Environment*, 293, p.119446.
16. Gao, Z. and Zhou, X., 2024. A review of the CAMx, CMAQ, WRF-Chem and NAQPMS models: Application, evaluation and uncertainty factors. *Environmental Pollution*, 343, p.123183.
17. Grell, G.A., Peckham, S.E., Schmitz, R., McKeen, S.A., Frost, G., Skamarock, W.C. and Eder, B., 2005. Fully coupled “online” chemistry within the WRF model. *Atmospheric environment*, 39(37), pp.6957-6975.
18. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* 9(8), 1735–1780 (1997).
19. Huang, L., Zhu, Y., Zhai, H., Xue, S., Zhu, T., Shao, Y., Liu, Z., Emery, C., Yarwood, G., Wang, Y. and Fu, J., 2021. Recommendations on benchmarks for numerical air quality model applications in China—Part 1: PM<sub>2.5</sub> and chemical species. *Atmospheric Chemistry and Physics*, 21(4), pp.2725-2743.
20. Jeong, D., Yoo, C., Yeh, S.W., Yoon, J.H., Lee, D., Lee, J.B. and Choi, J.Y., 2022. Statistical seasonal forecasting of winter and spring PM<sub>2.5</sub> concentrations over the Korean Peninsula. *Asia-Pacific Journal of Atmospheric Sciences*, 58(4), pp.549-561.
21. Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X. and Sachdeva, S., 2019. Evaluation of different machine learning approaches to forecasting PM<sub>2.5</sub> mass concentrations. *Aerosol and Air Quality Research*, 19(6), pp.1400-1410.
22. Kumar, N., Park, R.J., Jeong, J.I., Woo, J.H., Kim, Y., Johnson, J., Yarwood, G., Kang, S., Chun, S. and Knipping, E., 2021. Contributions of international sources to PM<sub>2.5</sub> in South Korea. *Atmospheric Environment*, 261, p.118542.
23. Li, N., Zhang, H., Zhu, S., Liao, H., Hu, J., Tang, K., Feng, W., Zhang, R., Shi, C., Xu, H. and Chen, L., 2023. Secondary PM<sub>2.5</sub> dominates aerosol pollution in the Yangtze River Delta region: Environmental and health effects of the Clean air Plan. *Environment International*, 171, p.107725.
24. Li, X. et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environ. Pollut.* 231(Pt 1), 997–1004 (2017).
25. Li, Z., Zhang, X., Liu, X. and Yu, B., 2022. PM<sub>2.5</sub> pollution in six major Chinese urban agglomerations: spatiotemporal variations, health impacts, and the relationships with meteorological conditions. *Atmosphere*, 13(10), p.1696.
26. Li, Z., Zhang, X., Liu, X. and Yu, B., 2022. PM<sub>2.5</sub> pollution in six major Chinese urban agglomerations: spatiotemporal variations, health impacts, and the relationships with meteorological conditions. *Atmosphere*, 13(10), p.1696.
27. Ma, X., Sha, T., Wang, J., Jia, H. and Tian, R., 2018. Investigating impact of emission inventories on PM<sub>2.5</sub> simulations over North China Plain by WRF-Chem. *Atmospheric Environment*, 195, pp.125-140.
28. Park et al., Variation of PM<sub>2.5</sub> Chemical Compositions and their Contributions to light extinction in Seoul, *Aerosol and Air Quality Research*, 18, 2220-2229, 2018
29. Park, E.H., Heo, J., Hirakura, S., Hashizume, M., Deng, F., Kim, H. and Yi, S.M., 2018. Characteristics of PM<sub>2.5</sub> and its chemical constituents in Beijing, Seoul, and Nagasaki. *Air Quality, Atmosphere & Health*, 11, pp.1167-1178.
30. Qadeer, K., Rehman, W.U., Sheri, A.M., Park, I., Kim, H.K. and Jeon, M., 2020. A long short-term memory (LSTM) network for hourly estimation of PM<sub>2.5</sub> concentration in two cities of South Korea. *Applied Sciences*, 10(11), p.3984.

31. Retrieving ground-level PM<sub>2.5</sub> concentrations in China (2013–2021) with a numerical-model-informed testbed to mitigate sample-imbalance-induced biases.
32. Reyes-Villegas, E., Lowe, D., Johnson, J.S., Carslaw, K.S., Darbyshire, E., Flynn, M., Allan, J.D., Coe, H., Chen, Y., Wild, O. and Archer-Nicholls, S., 2023. Simulating organic aerosol in Delhi with WRF-Chem using the volatility-basis-set approach: exploring model uncertainty with a Gaussian process emulator. *Atmospheric Chemistry and Physics*, 23(10), pp.5763–5782.
33. Rybarczyk and Zalkeviciute, Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review (2018).
34. Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M.G., Huang, X.Y., Wang, W. and Powers, J.G., 2008. A description of the advanced research WRF version 3. NCAR technical note, 475(125), pp.10–5065.
35. Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Liu, Z., Berner, J., Wang, W., Powers, J.G., Duda, M.G., Barker, D.M. and Huang, X.Y., 2019. A description of the advanced research WRF model version 4. National Center for Atmospheric Research: Boulder, CO, USA, 145(145), p.550.
36. Sokhi, R.S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., Halenka, T. and Ketzel, M., 2021. Advances in air quality research—current and emerging challenges. *Atmospheric Chemistry and Physics Discussions*, 2021, pp.1–133.
37. Statistical Seasonal Forecasting of Winter and Spring PM<sub>2.5</sub> Concentrations over the Korean Peninsula, 2022, 58.
38. Tao, H., Xing, J., Zhou, H., Pleim, J., Ran, L., Chang, X., Wang, S., Chen, F., Zheng, H. and Li, J., 2020. Impacts of improved modeling resolution on the simulation of meteorology, air quality, and human exposure to PM<sub>2.5</sub>, O<sub>3</sub> in Beijing, China. *Journal of Cleaner Production*, 243, p.118574.
39. Travis, K.R., Crawford, J.H., Chen, G., Jordan, C.E., Nault, B.A., Kim, H., Jimenez, J.L., Campuzano-Jost, P., Dibb, J.E., Woo, J.H. and Kim, Y., 2022. Limitations in representation of physical processes prevent successful simulation of PM<sub>2.5</sub> during KORUS-AQ. *Atmospheric Chemistry and Physics*, 22(12), pp.7933–7958.
40. Vignesh, P.P., Jiang, J.H. and Kishore, P., 2023. Predicting PM<sub>2.5</sub> concentrations across USA using machine learning. *Earth and Space Science*, 10(10), p.e2023EA002911.
41. Wang, P., Qiao, X. and Zhang, H., 2020. Modeling PM<sub>2.5</sub> and O<sub>3</sub> with aerosol feedbacks using WRF/Chem over the Sichuan Basin, southwestern China. *Chemosphere*, 254, p.126735.
42. WHO Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease, 9789241511353 (2016) <https://www.who.int/publications/i/item/9789241511353>.
43. Y. S. Kim, G. H. Kwak, K. D. Lee, S. I. Na, C. W. Park, & N. W. Park: Performance Evaluation of Machine Learning and Deep Learning Algorithms in Crop Classification: Impact of Hyper-parameters and Training Sample Size. *Korean Journal of Remote Sensing*, 2018, 34(5), 811–827pp .
44. Yu, M., Cai, X., Song, Y. and Wang, X., 2019. A fast forecasting method for PM<sub>2.5</sub> concentrations based on footprint modeling and emission optimization. *Atmospheric Environment*, 219, p.117013.
45. Zhang, H., Wang, J., García, L.C., Zhou, M., Ge, C., Plessel, T., Szykman, J., Levy, R.C., Murphy, B. and Spero, T.L., 2022. Improving surface PM<sub>2.5</sub> forecasts in the United States using an ensemble of chemical transport model outputs: 2. Bias correction with satellite data for rural areas. *Journal of Geophysical Research: Atmospheres*, 127(1), p.e2021JD035563..
46. Zhou, L., Sun, L., Luo, Y., Xia, X., Huang, L., Liao, Z. and Yan, X., 2023. Air pollutant concentration trends in China: correlations between solar radiation, PM<sub>2.5</sub>, and O<sub>3</sub>. *Air Quality, Atmosphere & Health*, 16(8), pp.1721–1735.

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