Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.230033

Estimation of PM_{10} and $PM_{2.5}$ Using Backscatter Coefficient of Ceilometer and Machine Learning

Bu-Yo Kim[*], Joo Wan Cha, Yong Hee Lee

Research Applications Department, National Institute of Meteorological Sciences, Seogwipo, Jeju 63568, Korea

ABSTRACT

Air quality issues, including health and environmental challenges, have recently become more relevant in urban areas with large populations and active industries. Therefore, particulate matter (PM) estimation with high accuracy using various methods is required. In this study, PM_{10} and $PM_{2.5}$ in Cheongju city, South Korea, were estimated using the attenuated backscatter coefficient of the ceilometer and meteorological observation data from an automatic weather station with supervised machine learning (ML). The backscatter coefficient data were obtained from the vertical layer with the highest correlation with PM_{10} and $PM_{2.5}$. The estimation methods utilized were tree-, vector-, neural-, and regularization-based supervised ML. The extreme gradient boosting method yielded the highest PM estimation accuracy. The estimation of PM_{10} and $PM_{2.5}$ for the test data set was more accurate than that in previous studies that used satellite and ground-based meteorological data (bias = 0.10 $\mu g\ m^{-3}$, root mean square error (RMSE) = 14.44 $\mu g\ m^{-3}$, and $R = 0.92$ for PM_{10}; and bias = 0.12 $\mu g\ m^{-3}$, RMSE = 7.16 $\mu g\ m^{-3}$, and $R = 0.91$ for $PM_{2.5}$). Particularly, the correlation coefficient was the highest for the estimation results for strong haze cases (1 km < visibility ≤ 5 km) ($R = 0.95$ for PM_{10}; $R = 0.89$ for $PM_{2.5}$). Therefore, PM estimation using meteorological observation data can help obtain meteorological and PM information simultaneously, making it useful for air quality monitoring.

Keywords: PM_{10}, $PM_{2.5}$, Ceilometer, Backscatter coefficient, Machine learning, Extreme gradient boosting

OPEN ACCESS

Received: February 12, 2023
Revised: September 1, 2023
Accepted: October 20, 2023

[*] Corresponding Author:
kimbuyo@korea.kr

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Aerosols in the atmosphere consist of small solid particles or liquid droplets that are either naturally occurring or artificially produced. These particulate matter (PM) are defined based on size into $PM_{2.5}$ (particles < 2.5 µm) and PM_{10} (particles < 10 µm). Increase in PM concentrations can lead to respiratory and cardiovascular diseases, cancer, and even death in severe cases (Czernecki *et al.*, 2021; Minh *et al.*, 2021). Smaller PM particles are more harmful to human health because they can penetrate deep into the respiratory system (Pappa and Kioutsioukis, 2021). In recent years, human activity such as population growth and increased energy consumption and industrial activities has led to continuous generation of high levels of air pollutants (Oh *et al.*, 2015). China consumes approximately 40% and 70% of global coal and energy, respectively, and emits a large amount of pollutants, deteriorating the air quality of neighboring countries and cities according to atmospheric pressure patterns (Peterson *et al.*, 2019; Oh *et al.*, 2020). Moreover, these PMs contain large amounts of harmful organic and inorganic chemicals, such as sulfur dioxide (SO_2) and lead (Pb) (Kim and Lee, 2018). In addition, PM in the air—especially very small PM (< 2.5 µm)—frequently causes haze by scattering visible light and causes human and material damage via traffic accidents (Biswas *et al.*, 2020; Ma *et al.*, 2020). Haze is frequently generated under calm atmospheric conditions (such as low wind speed, stable steady state, and low mixing layer height) and floats in the atmosphere, further polluting the atmosphere by producing secondary aerosols (Gao *et al.*, 2015). Therefore, many countries and cities are making efforts to reduce air pollutant levels to preserve the environment and improve the quality of human life (Lim *et al.*, 2022). Various studies are being conducted to improve PM monitoring, estimation, and prediction accuracy (Yang *et al.*, 2020; Czernecki *et al.*, 2021; Kim *et al.*, 2022b).

$PM_{2.5}$ and PM_{10} are generally observed in real-time using the beta-ray method with equipment such as FH62C14 (Thermo Fisher Scientific Inc., USA) and BAM 120 (Met One Instrument Inc., USA). However, since PM variations are largely spatial and temporal, they must be accompanied by dense observations locally. However, observations are very restricted by the limitations of establishing the observatory (Ma *et al.*, 2021; Kim *et al.*, 2022a). Therefore, PM is estimated using data from satellite and ground-based observation instruments (visibility sensors and ceilometers). However, although satellite data provide wider coverage, the observation period is not continuous depending on the characteristics of the satellite platform (polar or geostationary orbit), the spatial resolution is large, and PM can only be estimated in cloud-free regions (Huang *et al.*, 2018; Kim *et al.*, 2018; Shin *et al.*, 2020b). Therefore, satellite data are mainly used for daily mean PM estimations (Chen *et al.*, 2018; Danesh Yazdi *et al.*, 2020). In visibility sensors, PM can be estimated using the exponential relationship between the visibility or extinction coefficient and PM for each relative humidity range (Ji *et al.*, 2020; Sun *et al.*, 2020). However, because an exponential relationship is used under limited meteorological conditions, PM estimation accuracy according to various meteorological conditions is not high. In addition, because the instrument error of the visibility sensor is ± 10–20%, depending on the visibility range, the uncertainty of PM estimation can be high (Du *et al.*, 2013; Zhang *et al.*, 2017).

Ceilometers are well-known instruments for observing cloud base height, cover, and structures (multi-layer clouds) (Kotthaus *et al.*, 2016). However, ceilometers record backscattered laser pulses in the vertical atmosphere from atmospheric clouds, droplets, and aerosols, as raw data. Therefore, aerosols can be detected and PM can be estimated using the vertical attenuated backscatter coefficient of the ceilometer. In other words, the attenuated backscatter coefficient of the upper atmosphere is used to estimate variables, such as mixing layer height, volcanic ash, and aerosol transport, and the attenuated backscatter coefficient of the lower atmosphere is used to estimate precipitation, fog, and drizzle characteristics (Kotthaus *et al.*, 2016; de Arruda Moreira *et al.*, 2020). Previous studies (Münkel *et al.*, 2007; Li *et al.*, 2017; Parde *et al.*, 2020; Jung and Um, 2022) have estimated PM considering ceilometer data through linear and exponential relationships between the vertical attenuated backscatter coefficient and PM. However, these empirical methods are limited to estimating PM variations that occur under various meteorological conditions (Kim *et al.*, 2022b). Therefore, in this study, we present a novel method for estimating PM_{10} and $PM_{2.5}$ using the vertical attenuated backscatter coefficient of the ceilometer and machine learning (ML). ML effectively describes the nonlinear relationships among meteorological factors, leading to higher accuracy and utilization than those of previous empirical methods (Shin *et al.*, 2020a; Kim *et al.*, 2021a). The results of PM estimation in this study suggest that air quality monitoring network can be expanded and improved by incorporating additional PM data into the PM observation network.

2 METHODS

2.1 Research Data

In this study, we estimated PM_{10} and $PM_{2.5}$ using raw data (vertical profile) from the ceilometer and meteorological data from the automatic weather station (AWS) in Cheongju (Station No. 131, 36.639°N, 127.441°E, altitude: 58 m) of the Korea Meteorological Administration (KMA). This study used data observed at 1 h intervals between 1 January 2020 and 31 December 2021. The ceilometer used in this study is Vaisala CL31, which observes the backscatter coefficient at 5 m intervals up to 7500 m vertically and records data at 15 s intervals. The AWS data used were temperature (T, °C), dew point depression (T_{dep}, K), pressure (P, hPa), wind direction (WD, °), and wind speed (WS, m s^{-1}). PM data from the air-quality measurement station (AMS) in Cheongju (Station No. 533112, 36.645°N, 127.437°E, altitude: 57 m) of the Ministry of Environment (MOE) were used as label data for ML and to verify the estimation results. This AMS was closest to the AWS with an aerial distance of approximately 0.70 km, and the observational altitude was similar.

PM largely varies spatially and temporally; therefore, data from the nearest region can reduce estimation errors (Kim *et al.*, 2020). In this study, we attempted to clarify the relationship between the AWS and AMS data. Therefore, data from the closest AWS and AMS in South Korea were used in this study. PM_{10} and $PM_{2.5}$ of the AMS were measured using Met One BAM-1020, quality controlled according to the MOE guidelines (MOE, 2021), and the data were provided to AirKorea (www.airkorea.or.kr). Missing cases among the collected PM data were not used in this study, and PM was not observed between 27 May 2020 and 30 June 2020, owing to the replacement of the PM observation instrument. The raw ceilometer and meteorological data were randomly sampled without replacement at 5:3:2 to construct datasets for ML training, validation, and testing. The ML methods used in this study optimized each hyperparameter using training and validation sets (Kim *et al.*, 2021b). The estimated PM_{10} and $PM_{2.5}$ were compared with the AMS data, and the ML method with the highest estimation accuracy was used to evaluate the results of the test set. The comparison of the estimated and observed PM_{10} and $PM_{2.5}$ (PM_{est} and PM_{obs}, respectively) is presented in Eqs. (1)–(3) and evaluated using the bias, root mean square error (RMSE), and correlation coefficient (*R*).

$$\text{bias} = \sum_{i=1}^{N} \frac{\left(\text{PM}_{est_i} - \text{PM}_{obs_i}\right)}{N} \tag{1}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{\left(\text{PM}_{est_i} - \text{PM}_{obs_i}\right)^2}{N}} \tag{2}$$

$$\text{R} = \frac{\left(\text{PM}_{est_i} - \overline{\text{PM}_{est}}\right)\left(\text{PM}_{obs_i} - \overline{\text{PM}_{obs}}\right)}{\sqrt{\sum_{i=1}^{N} \frac{\left(\text{PM}_{est_i} - \overline{\text{PM}_{est}}\right)^2}{N} \sum_{i=1}^{N} \frac{\left(\text{PM}_{obs_i} - \overline{\text{PM}_{obs}}\right)^2}{N}}} \tag{3}$$

Here, N means the number of data.

2.2 ML Methods

In this study, PM_{10} and $PM_{2.5}$ were estimated using four types of supervised ML methods: tree-based (RF: random forest, XGB: extreme gradient boosting, LGB: light gradient boosting), vector-based (SVR: support vector regression, kNN: k-nearest neighbor), neural-based (ANN: artificial neural network, ELM: extreme learning machine), and regularization-based (PLSR: partial least squares regression). These methods have been widely used in aerosol research. However, PM estimation accuracy may vary depending on the characteristics of the input data, such as the type, amount, frequency, and local meteorological conditions (Kim *et al.*, 2022b, 2022c). Thus, the most appropriate ML method was selected by comparing the estimated with the observed PM_{10} and $PM_{2.5}$ using various methods. The details of each ML method used in the study are described in Sections 2.2.1–2.2.4. The hyperparameter settings are presented in Table 1.

2.2.1 Tree-based methods

The RF method constructs N decision trees and ensembles the predicted results from each tree to obtain the final prediction results (Wright and Ziegler, 2017; Wright *et al.*, 2020). In this process, each tree grows leaves to the maximum depth and randomly samples input variables from each leaf to create symmetrical leaves. The LGB and XGB methods are similar to the RF method. However, they use boosting instead of bagging to resample and ensemble (Chen *et al.*, 2022b). In other words, these methods improve the predictive power by performing sequential reinforcement learning on trees with weak predictive power. The predicted estimates (f_1(x)-f_N(x)) from each tree were weighted means to obtain the final result (f(x)). Unlike the XGB method (level-wise, Fig. 1(a)), as illustrated in Fig. 1(b), LGB has relatively fast learning and prediction speed by growing the tree leaf-wise (Al Banna *et al.*, 2020; Lu and Ma, 2020).

Table 1. Hyper-parameter settings of machine learning methods for estimating PM_{10} and $PM_{2.5}$.

Algorithm (R library)	Setting	
RF (ranger)	PM_{10}	num.trees = 940, mtry = 4, min.node.size = 4
	$PM_{2.5}$	num.trees = 890, mtry = 3, min.node.size = 3
XGB (xgboost)	PM_{10}	nrounds = 1010, max_depth = 5, eta = 0.10
	$PM_{2.5}$	nrounds = 920, max_depth = 9, eta = 0.10
LGB (lightgbm)	PM_{10}	nrounds = 620, max_depth = 7, learning_rate = 0.19
	$PM_{2.5}$	nrounds = 490, max_depth = 9, learning_rate = 0.19
SVR (e1071)	PM_{10}	cost = 5, gamma = 0.11, epsilon = 0.09
	$PM_{2.5}$	cost = 15, gamma = 0.11, epsilon = 0.10
kNN (kknn)	PM_{10}	k = 8
	$PM_{2.5}$	k = 22
ANN (nnet)	PM_{10}	maxit = 980, size = 7, decay = 0.5
	$PM_{2.5}$	maxit = 950, size = 9, decay = 0.5
ELM (Kim *et al.*, 2022a)	PM_{10}	nhidden = 1360
	$PM_{2.5}$	nhidden = 650

2.2.2 Vector-based methods

SVR regresses data based on the ε-insensitive loss function by finding a hyperplane composed of support vectors that can classify the margin of the distance between vectors, as illustrated in Fig. 1(c) (Taghizadeh-Mehrjardi *et al.*, 2017). The optimal hyperplane has weights and biases that minimize the mapping function (Meyer *et al.*, 2021). This study used the radial basis function kernel for high-dimensional mapping. The SVR method was optimized by determining the constraint's margin threshold and violation level. As presented in Fig. 1(d), the kNN method finds the k-nearest neighbors to a query in the high-dimensional data feature space and predicts the query based on the weighting of these neighbors according to the Euclidean distance (Zhang *et al.*, 2018; Martínez *et al.*, 2019).

2.2.3 Neural-based methods

ANN and ELM consist of an input layer that accepts the input data, an output layer that outputs the prediction results, and a hidden layer between both measurements, as presented in Figs. 1(e) and 1(f) (Rosa *et al.*, 2020). The input data determines the weight and bias of the logistic regression through the activation function in the hidden layer. In the ANN method, the output layer provides prediction results using the weights and biases determined in the hidden layer through the sigmoid function as an activation function (Rosa *et al.*, 2020). In ELM, a rectified linear function was the activation function. Unlike ANNs, ELM generates weights and biases between the input and hidden layers and between the hidden and output layers with multiple perceptrons to provide prediction results at the output layer (Huang *et al.*, 2006).

2.2.4 Regularization-based method

PLSR models the relationships between data and correlated predictors. The observed value (Y) was estimated by extracting k latent vectors with high covariance with the predictor (X) from the component set and finding the most suitable regression function (Zhou *et al.*, 2019). This process involves generalizing the principal component analysis (PCA) to decompose X and find the component that best explains Y. The PM was estimated based on the relationship between Eqs. (4) and (5). In this study, the' pls' package in R was used.

$$X = TD + E \tag{4}$$

$$Y = UQ + F \tag{5}$$

Here, T and U are k score matrices (N × k) extracted from the component set, D (k × n) and Q (k × 1) are the loading matrix and vector for X and Y, and E (N × n) and F (N × 1) are the matrix and vector of residuals, respectively.

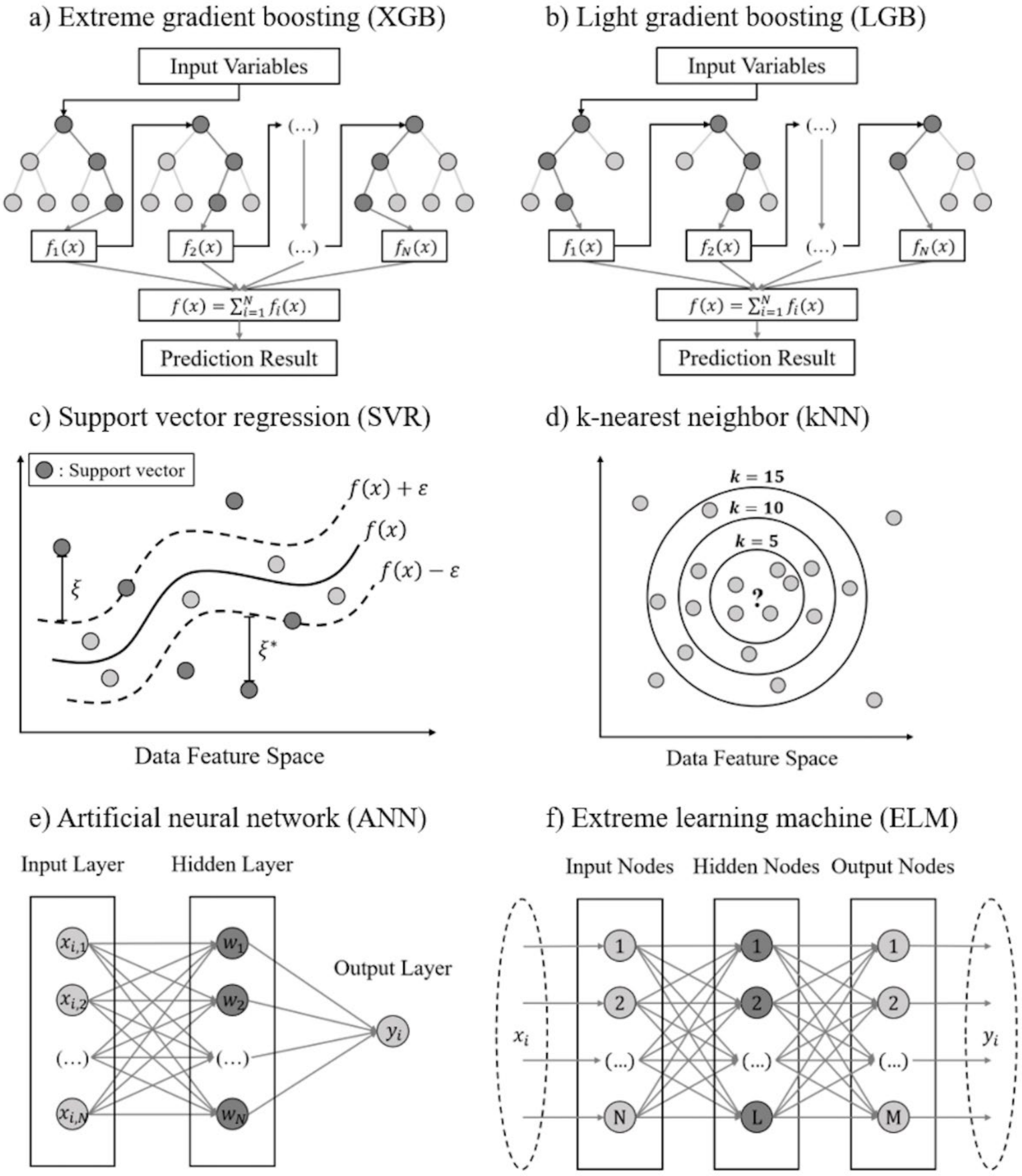

Fig. 1. Schematic diagram of each machine learning method: a) XGB, b) LGB, c) SVR, d) kNN, e) ANN, and f) ELM (Kim *et al.*, 2021b, 2022a, 2022b).

2.3 Variation Trend of Meteorological and PM Variables in Cheongju

The trends of the daily mean time series variation of meteorological and PM data and backscatter coefficient data collected between 2020 and 2021 are presented in Fig. 2. For the backscatter coefficient, the mean results of the 2nd–20th vertical layers (< 100 m) are presented in Fig. 2(a). In the backscatter coefficient, the lower layer below 100 m height has less sunlight and electronic noise than the higher layer; thus, the accuracy of the backscatter coefficient is high, and no separate data quality control is required. Notably, the first layer is usually not used due to electronic noise (Kotthaus *et al.*, 2016; Parde *et al.*, 2020). PM_{10} and $PM_{2.5}$ concentrations in Cheongju tended to increase in spring and winter. During this period, PM concentration increases as Asian dust is generated in urban areas around China and Mongolia due to westerly winds and air pollutants generated from fossil fuels and industrial activities (Peterson *et al.*, 2019; Oh *et al.*, 2020; Hur *et al.*, 2021; Filonchyk, 2022; Filonchyk and Peterson, 2022). In addition, a high PM concentration is maintained for an extended period according to atmospheric pressure patterns, and the PM concentration increases. In particular, in the dry season (December–May), high PM concentrations

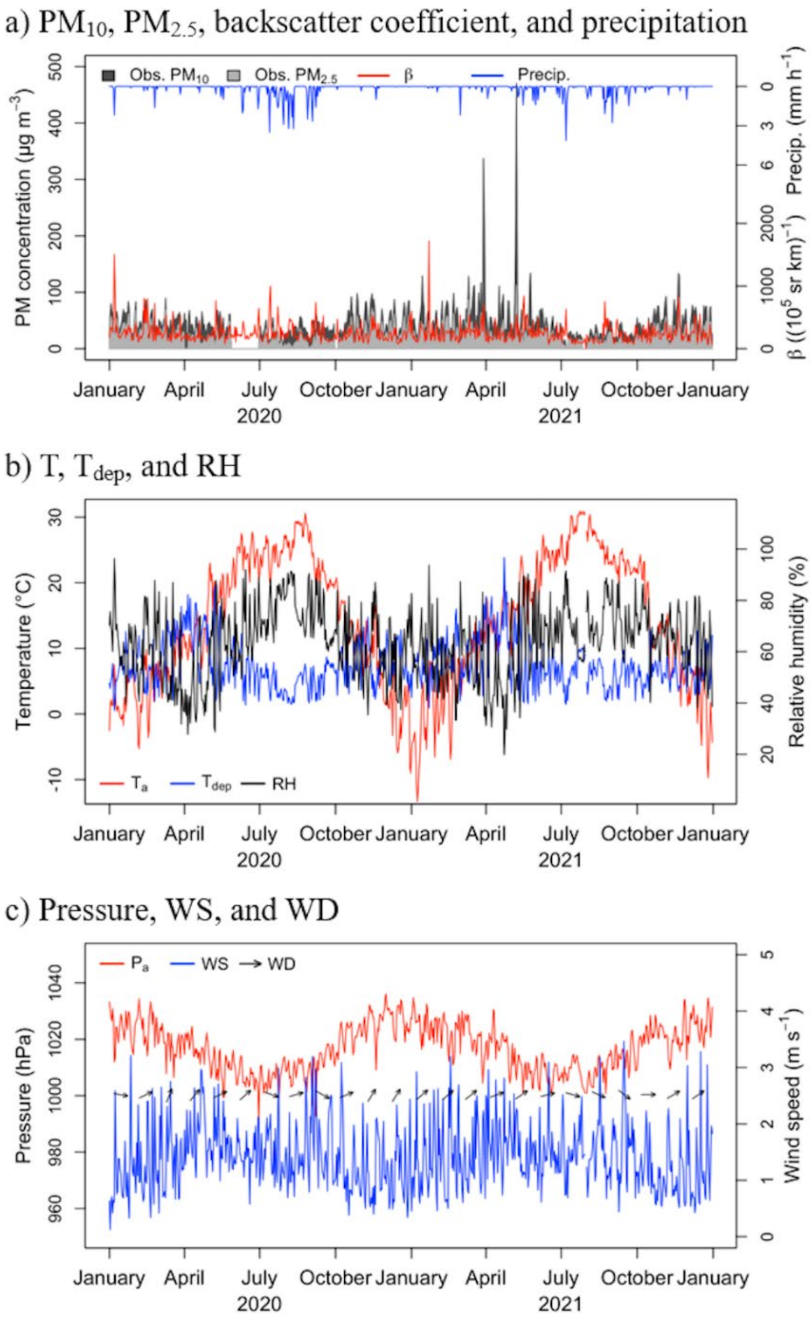

Fig. 2. Daily mean meteorological and particulate matter variables and backscatter coefficient time-series in 2020–2021 (a) PM_{10}, $PM_{2.5}$, backscatter coefficient (β, mean of 2nd–20th layers), and precipitation, b) temperature (T), dewpoint depression (T_{dep}), and relative humidity (RH), and c) pressure, wind speed (WS), and wind direction (WD).

are presented in the Korean Peninsula due to the western high and eastern low pressure arrangement and the Siberian high pressure and Aleutian low pressure (Kim *et al.*, 2022b). The foreign country's PM contribution to the Korean Peninsula is 40%–50% (Oh *et al.*, 2020). High concentrations of PM_{10} were introduced on 29–30 March and 7–8 May 2021. Moreover, Cheongju City is topographically a basin form surrounded by mountains. When the atmosphere is stable or WS weakens, air pollutants do not diffuse well, resulting in high PM concentrations (Kim and Moon, 2020). In summer and autumn, PM concentration tends to decrease due to the wet deposition of PM due to precipitation (Lee *et al.*, 2013; Kim and Kim, 2020). In addition, when the atmosphere is too dry or too humid, the PM concentration tends to decrease due to dry deposition (Lou *et al.*, 2017).

A ceilometer detects backscattered light reflected by fog, clouds, precipitation, and aerosols by emitting laser pulses in the atmosphere (Münkel *et al.*, 2007). In particular, backscattering by water droplets is strong; therefore, it produces a large vertical backscatter coefficient for precipitation, clouds, and fog (Wiegner and Gasteiger, 2015; Parde *et al.*, 2020). Therefore, estimating the PM_{10} and $PM_{2.5}$ in these meteorological phenomena is challenging. In this study, to estimate PM using the vertical backscattering coefficient data of the ceilometer, precipitation, lower layer clouds with a cloud base height of < 200 m, and fog cases were excluded (Münkel *et al.*, 2007; Li *et al.*, 2017; Parde *et al.*, 2020; Jung and Um, 2022). The backscatter coefficient detected aerosol characteristics using only data of < 2000 10^{-5} sr^{-1} km^{-1} (Chan *et al.*, 2018). In this study, the backscatter coefficient data of the layer with the highest correlation coefficient between the vertically observed backscatter coefficient and PM and with each mean time (1 min, 10 min, 30 min, and 60 min) were used to estimate PM. Therefore, correlation coefficients were estimated between the mean time backscatter coefficient for each layer (Figs. 3(a) and 3(c)) and PM and between the mean time and layer-mean backscatter coefficient for each layer (Figs. 3(b) and 3(d)) (the correlation coefficient of the 10th layer means the mean backscatter coefficient of 2nd–10th layers) and PM. In this study, the correlation coefficients of the backscatter coefficient and PM_{10} and $PM_{2.5}$ were the highest at the 9th layer (45 m) at 0.60 and 0.72, respectively, among the 1 min mean results. This result revealed the highest correlation coefficient at a 45 m height, similar to a study using the Lufft CHM 15k ceilometer (Li *et al.*, 2017). Therefore, in this study, PM_{10} and $PM_{2.5}$ were estimated using backscatter coefficient data of the 9th layer-mean over 1 min, meteorological data of AWS, and ML methods.

3 RESULTS AND DISCUSSION

3.1 Training and Validation Results for Each ML Method

Fig. 4 shows a Taylor plot of PM_{10} and $PM_{2.5}$ estimations for each ML method with hyper-parameters optimized using training and validation sets. In this figure, the blue dotted line on the x- and y-axes represents the standard deviation (SD) of the observed and estimated PM, respectively. The open black circle and solid line indicate the SD of the observed value. In addition, the dotted black line is the correlation coefficient, and the solid green line is the centered RMSE. The closed circles refer to the results of the ML methods used in this study. Therefore, proximity between the closed and open black circles correlated with a smaller difference between the centered RMSE and the SD of each estimation result and a higher correlation coefficient (Taylor, 2001). In the Taylor plot results, PM_{10} and $PM_{2.5}$ estimation in each training and validation set revealed that the PM estimation performance of the tree-based method (especially XGB and LGB) was high. However, in PLSR and vector-based methods, PM estimation performance was relatively low. This means that all the ML methods used allowed optimal PM estimation through hyper-parameter optimization. However, the bagging and boosting methods of tree-based MLs are suitable for PM estimation (Kim *et al.*, 2022a, 2022b). Among the tree-based methods, the XGB method had the best performance, with a bias of –0.17 µg m^{-3}, RMSE of 14.80 µg m^{-3}, and R of 0.89 for PM_{10} and a bias of 0.20 µg m^{-3}, RMSE of 7.15 µg m^{-3}, and R of 0.90 for $PM_{2.5}$. The RMSE was the smallest, and the R was the highest. Therefore, in this study, the estimation accuracy of PM_{10} and $PM_{2.5}$ was evaluated using the XGB method, which was the most accurate.

Fig. 5 presents the relative importance of each input variable of the XGB method used to estimate PM_{10} and $PM_{2.5}$. In this study, the backscatter coefficient variable from the ceilometer had the highest relative importance, at 54.67% for PM_{10} and 62.84% for $PM_{2.5}$. This is the directly observed backscatter intensity for suspended matter in the atmosphere; therefore, it is of the highest importance in PM estimation (Münkel *et al.*, 2007). T and T_{dep} can increase local PM concentrations under meteorological conditions with weak WS (Whalley and Zandi, 2016). Additionally, low temperatures are associated with emissions of local air pollutants, such as heating. Moreover, T_{dep}, similar to the RH parameter, is related to the wet and dry PM deposition, according to atmospheric condensation and humidity (Gao *et al.*, 2019). High pressure generates a downdraft and reduces the PM vertical mixing. PM accumulates near the ground, and the PM concentration tends to

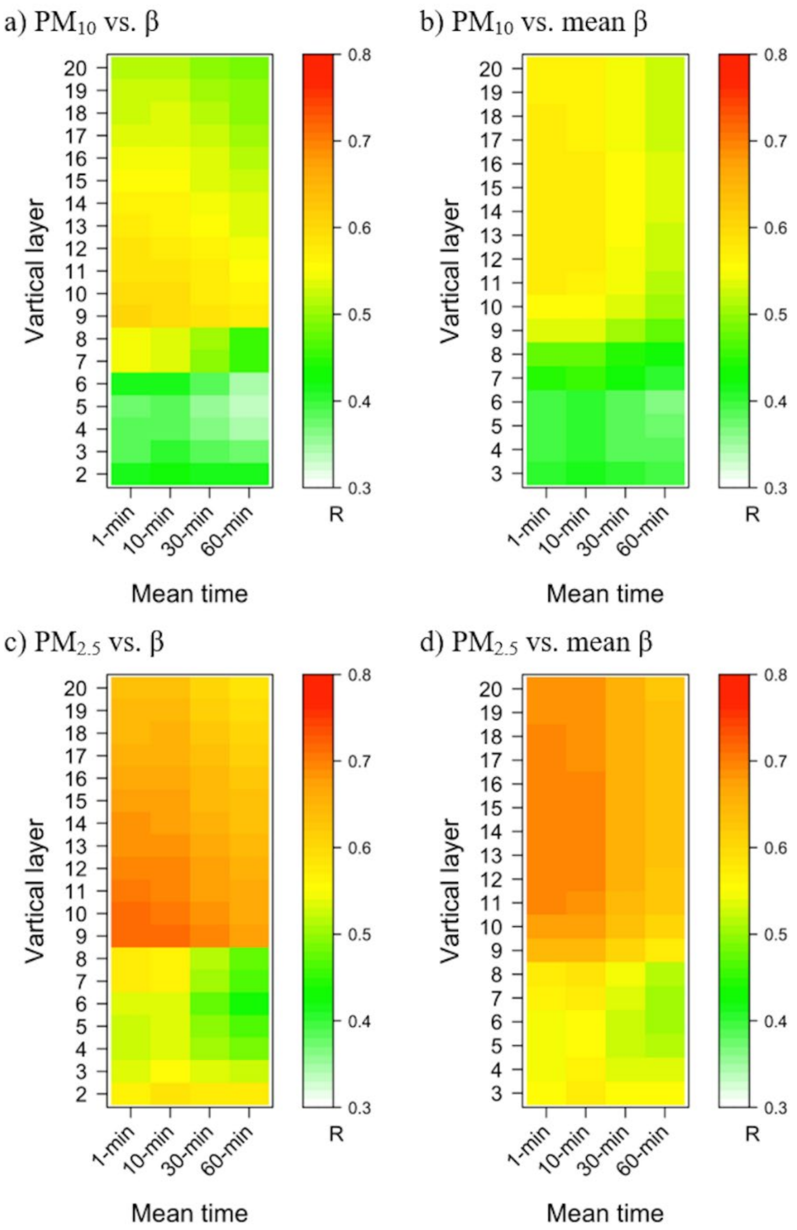

Fig. 3. (a, c) Correlation coefficients between the backscatter coefficient and PM_{10} and $PM_{2.5}$ by each mean time and (b, d) between the backscatter coefficient and PM_{10} and $PM_{2.5}$ by each mean time and mean vertical layers. β means the backscatter coefficient.

increase (Wang *et al.*, 2010; Wen *et al.*, 2018). In this study, the relative importance of P for $PM_{2.5}$ was 8.18%—the second highest relative importance. However, RH (22.49%) and T_{dep} (7.48%) were relatively more important for PM_{10} than P in estimating PM. This is because large particles, such as PM_{10}, are relatively more affected by deposition (Lou *et al.*, 2017). The WD was highly related to concentrations of PM transported from other regions. Nonetheless, its relative importance was low. Julian days, weeks, and hours, which are periodic variables, can reflect the periodic variation in PM and the effect of periodic meteorological variation patterns not considered in this study. However, in this study, the relative importance of the backscatter coefficient variable and other meteorological variables was relatively high; therefore, the relative importance of these periodic variables was relatively low. The relative importance demonstrated in this study may produce different results depending on the ML method and the data characteristics (type, amount, frequency, and local meteorological conditions) (Kim *et al.*, 2022b; Kim *et al.*, 2022c).

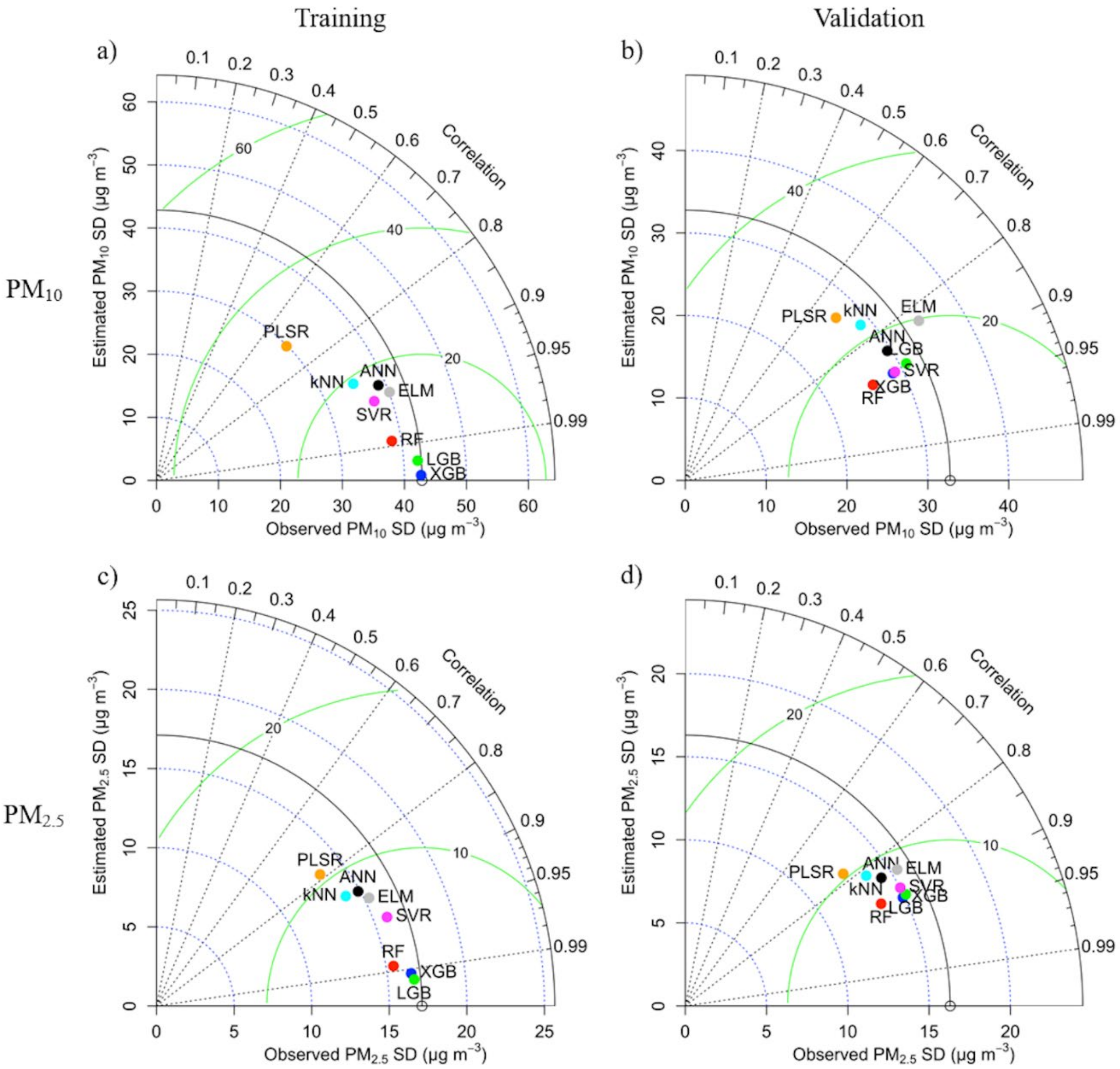

Fig. 4. Taylor plot of (a, b) PM$_{10}$ and (c, d) PM$_{2.5}$ estimation results for training and validation.

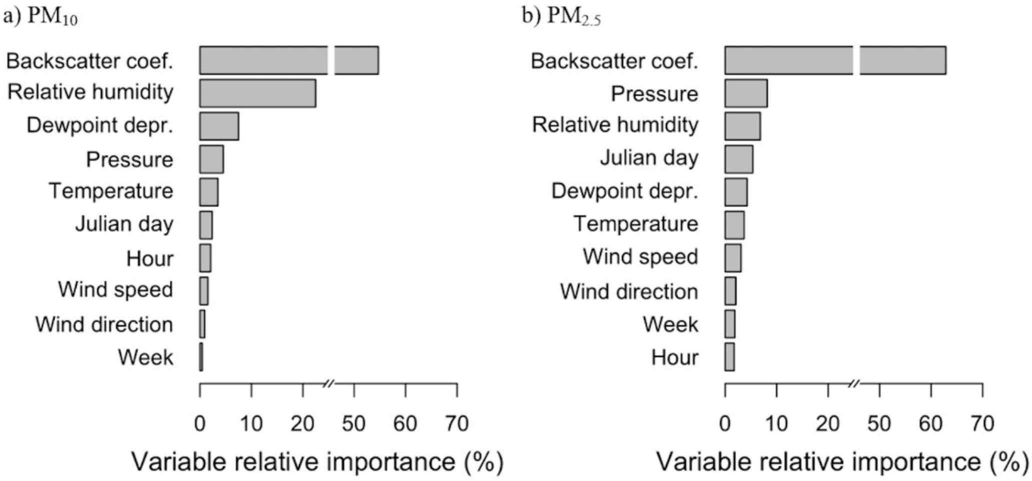

Fig. 5. Relative importance of each input variable in (a) PM$_{10}$ and (b) PM$_{2.5}$ estimation based on the XGB method.

3.2 Test Data Set Estimation

The hourly scatter plot (log scale) and daily mean time series of PM estimated using the XGB method and the observed PM for the test dataset are illustrated in Fig. 6. In Figs. 6(a) and 6(b), the observed and estimated PM on the 1:1 line tended to correlate. Regarding PM_{10}, the estimation error was minimal even for high concentrations of ≥ 200 µg m^{-3}. The estimation results for PM_{10} revealed bias = 0.10 µg m^{-3}, RMSE = 14.44 µg m^{-3}, and R = 0.92 and those for $PM_{2.5}$ revealed bias = 0.12 µg m^{-3}, RMSE = 7.16 µg m^{-3}, and R = 0.91. These results had a higher correlation coefficient than the those reported by Münkel *et al.* (2007) (R = 0.83 for PM_{10}), Li *et al.* (2017) (R = 0.75 for $PM_{2.5}$), Parde *et al.* (2020) (R = 0.82 for PM_{10} and R = 0.84 for $PM_{2.5}$), and Jung and Um (2022) (R = 0.79 for $PM_{2.5}$), who estimated PM empirically using a raw data of ceilometer. In other words, using the ML methods used in this study, such as the XGB method, we can accurately estimate PM through the nonlinear relationship between the observed PM and input variables under various atmospheric conditions (Kim *et al.*, 2021a, 2022a, 2022b, 2022c). The results of the daily mean time series (Fig. 6(c)) also showed estimation similar to that of the observed PM, with the estimation accuracy for PM_{10} having a bias of –0.27 µg m^{-3}, an RMSE of 5.98 µg m^{-3}, and an R of 0.97, and that for $PM_{2.5}$ having a bias of –0.18 µg m^{-3}, an RMSE of 3.18 µg m^{-3}, and an R of 0.97. These results had a higher correlation coefficient than those reported by Ferrero *et al.* (2019) (R = 0.91 for PM_{10} and $PM_{2.5}$), Chen *et al.* (2021) (R = 0.91 for $PM_{2.5}$), and Chen *et al.* (2022a) (R = 0.91 for PM_{10}), who estimated the mean daily PM using satellite data.

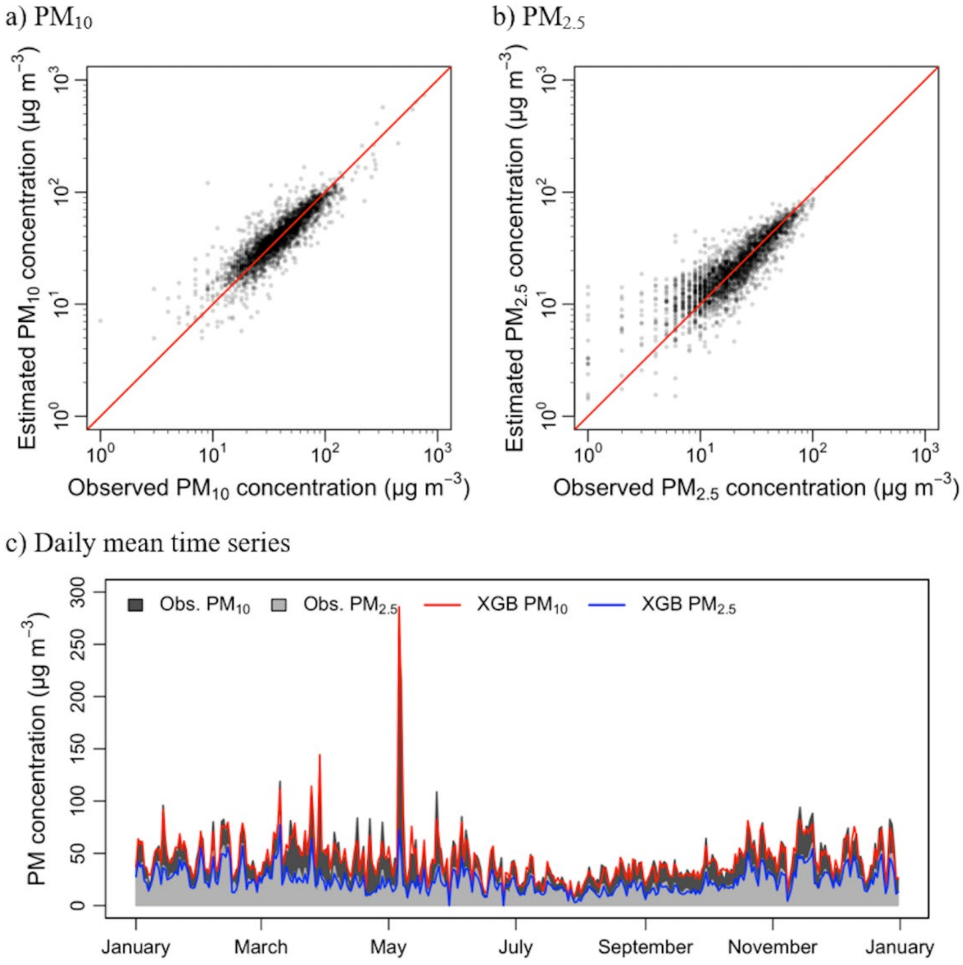

Fig. 6. (a, b) Hourly scatter plot (log scale) and (c) daily mean time series of PM estimated using the XGB method and observed PM for the test data set. The red line on the scatter plot is a 1:1 line. Dark gray and gray indicate observed PM_{10} and $PM_{2.5}$ in the time series, respectively, and the red and blue lines indicate estimated PM_{10} and $PM_{2.5}$, respectively.

Table 2 presents the estimated accuracy of monthly PM_{10} and $PM_{2.5}$ and monthly mean meteorological variables for the test data set. In this table, PM estimation accuracy was relatively low monthly in the dry season (April) with low RH or wet season (July–September) with high RH. This result is because the estimation accuracy of PM differed for each RH interval, as presented in Fig. 7. For PM_{10} and $PM_{2.5}$ estimation, RMSE was relatively large. R was low in dry atmospheric conditions with RH $<$ 40% or wet atmospheric conditions with RH $>$ 90%. Under these RH conditions, the estimation accuracy of PM_{10} was lower than that of $PM_{2.5}$ because more deposition and a relatively large change in optical properties appear in matter containing larger particles (Lou *et al.*, 2017; Guo *et al.*, 2020). Meanwhile, in September, PM_{10} and $PM_{2.5}$ had the lowest estimation accuracy. In September, the correlation coefficient between the backscatter coefficient and PM was 0.29 and 0.30, respectively, for PM_{10} and $PM_{2.5}$, revealing low correlation coefficients. In contrast, in May and February, where the estimation accuracy of PM_{10} and $PM_{2.5}$ was the highest, the correlation coefficient of the backscatter coefficient and PM_{10} and $PM_{2.5}$ were 0.67 and 0.81, respectively. The lower the correlation between the backscatter coefficient and the PM, the lower the estimation accuracy. Furthermore, September is the Jangma (June–October) period in South Korea, a season in which the wet deposition of PM occurs due to frequent and high-intensity rainfall (Founda *et al.*, 2016; Kim *et al.*, 2021a). This study excluded precipitation and fog cases from the collected data. However, since these phenomena can affect the backscattering coefficient before and after the observation, they may contain relatively more estimation errors than the other months (Wiegner and Gasteiger, 2015). Nevertheless, based on the visibility of the test set, the estimated accuracy of PM_{10} and $PM_{2.5}$ for each clear (10 km $<$ visibility $\leq$ 20 km), haze (1 km $<$ visibility $\leq$ 10 km), and strong haze (1 km $<$ visibility $\leq$ 5 km) case is presented in Table 3. These results revealed that the correlation coefficient between the observed PM and the

Table 2. Estimated accuracy of monthly PM_{10} and $PM_{2.5}$ and monthly mean meteorological variables for the test set. The unit of β (backscatter coefficient) is 10^{-5} sr^{-1} km^{-1}, and the unit of mean, bias, and RMSE is μg m^{-3}.

Month	N	T (°C)	T_{dep} (K)	RH (%)	P (hPa)	WD (°)	WS (m s^{-1})	β
1	266	0.07	7.55	59.08	1025.32	197.88	1.20	252.81
2	227	3.49	8.32	56.75	1024.18	215.49	1.29	260.12
3	235	8.23	9.89	53.44	1020.56	228.84	1.37	272.39
4	218	13.34	13.17	44.02	1017.88	224.43	1.89	184.57
5	200	18.53	9.06	59.22	1010.21	214.74	1.67	263.17
6	120	23.72	7.23	65.81	1008.39	188.85	1.44	229.26
7	176	26.27	6.15	70.21	1008.10	167.77	1.57	162.87
8	212	26.77	5.29	73.98	1009.48	174.54	1.44	187.51
9	215	22.02	6.30	69.14	1013.84	145.41	1.55	175.84
10	263	14.73	7.34	63.85	1022.13	187.29	1.18	231.77
11	233	8.34	7.44	61.70	1023.89	231.96	1.15	259.77
12	259	0.90	8.27	56.17	1026.73	227.44	1.29	219.32

Month	PM_{10}				$PM_{2.5}$			
	Mean	Bias	RMSE	R	Mean	Bias	RMSE	R
1	46.00	0.91	10.55	0.90	29.04	0.36	6.87	0.90
2	49.89	0.03	12.56	0.88	31.89	–0.29	7.08	0.93
3	58.10	1.84	25.16	0.84	30.37	0.50	9.13	0.86
4	45.35	–0.35	16.36	0.75	18.73	0.58	6.77	0.75
5	62.20	–0.94	25.51	0.96	23.85	–0.15	8.45	0.92
6	42.11	–0.57	9.80	0.89	24.62	0.06	6.67	0.85
7	25.91	1.11	7.53	0.86	15.41	0.83	5.79	0.85
8	31.19	–0.33	7.71	0.84	15.58	–0.17	5.80	0.81
9	31.38	–0.15	8.76	0.71	14.28	–0.09	6.28	0.57
10	47.89	1.58	9.72	0.90	27.00	0.46	6.66	0.89
11	53.88	–0.86	13.16	0.90	30.74	0.14	7.93	0.92
12	48.51	–1.56	10.60	0.91	29.82	–0.75	7.21	0.91

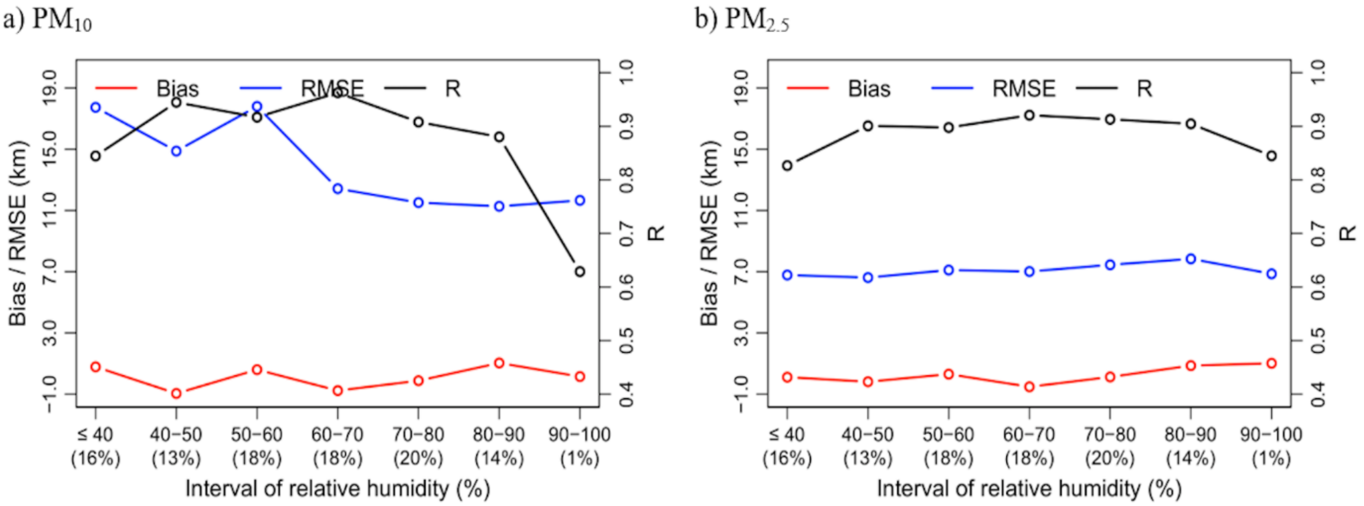

Fig. 7. (a) PM_{10} and (b) $PM_{2.5}$ estimation accuracy for each RH interval (bias: red line, RMSE: blue line, and R: black line). The number in parentheses below the interval represents the data ratio (%) to the test data set.

Table 3. Estimation accuracy of PM_{10} and $PM_{2.5}$ according to the visibility range for the test data set.

Variable	Visibility range (km)	N	Mean (μg m^{-3})	Bias (μg m^{-3})	RMSE (μg m^{-3})	*R*
PM_{10}	10–20	1908	36.67	0.51	11.81	0.80
	1–10	716	70.20	−1.01	19.82	0.93
	1–5	226	88.70	−0.32	25.65	0.95
$PM_{2.5}$	10–20	1908	18.37	0.43	6.35	0.78
	1–10	716	41.80	−0.73	8.95	0.89
	1–5	226	53.57	−0.69	9.72	0.89

estimated PM increased as the air quality worsened from clear to strongly haze. The related RMSE for PM_{10} and $PM_{2.5}$ in the haze was approximately 29% (*R* = 0.94) and 20% (*R* = 0.89), respectively.

4 CONCLUSIONS

In this study, a method for estimating PM_{10} and $PM_{2.5}$, which are the raw data of the ceilometer, was presented using the vertical attenuated backscatter coefficient data. We estimated PM using the backscatter coefficient of the layer with the highest correlation coefficient between the ceilometer and the PM, the meteorological data of the AWS, and the ML method. The meteorological data of the AWS were the input data for ML to estimate PM variation that can occur under various meteorological conditions. The estimation accuracy of ML methods can vary depending on the characteristics of the input data. Therefore, a method suitable for PM estimation was determined by comparing the PM estimation accuracy through hyperparameter optimization for each ML method. Among the tree-based (RF, XGB, and LGB), vector-based (SVR and kNN), neural-based (ANN and ELM), and regularization-based (PLSR) ML methods used in this study, the XGB method was the most accurate for PM estimation.

The PM estimation results for the test set revealed low accuracy in meteorological conditions where the correlation between PM and the backscattering coefficient was low. In particular, the estimation accuracy was relatively low in cases of low or high relative humidity during the Jangma season. Nevertheless, the estimation accuracy for PM_{10} was bias = 0.10 μg m^{-3}, RMSE = 14.44 μg m^{-3}, and R = 0.92, and for $PM_{2.5}$, the estimation accuracy of bias = 0.12 μg m^{-3}, RMSE = 7.16 μg m^{-3}, and R = 0.91. In particular, the estimation accuracy of haze cases with visibility of < 10 km was high. In the daily mean results, PM_{10} and $PM_{2.5}$ had high correlation coefficients of 0.97 or higher. These results revealed a higher correlation than the PM estimation results using a linear or exponential relationship between PM and the backscatter coefficient (or meteorological data) of the ceilometer. In other words, estimating PM using ML has described the nonlinear relationship between PM and various meteorological conditions. Therefore, the ML-based estimation results of PM_{10} and $PM_{2.5}$ using the backscatter coefficient data of the ceilometer can expand and improve the PM observation network.

ACKNOWLEDGMENTS

This work was funded by the Korea Meteorological Administration Research and Development Program “Research for Weather Modification and Cloud Physics” under Grant (KMA2018-00224).

DISCLAIMER

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Al Banna, M.H., Taher, K.A., Kaiser, M.S., Mahmud, M., Rahman, M.S., Hosen, A.S., Cho, G.H. (2020). Application of artificial intel-ligence in predicting earthquakes: State-of-the-art and future challenges. IEEE Access 8, 192880–192923. https://doi.org/10.1109/ACCESS.2020.3029859

Biswas, K., Chatterjee, A., Chakraborty, J. (2020). Comparison of air pollutants between Kolkata and Siliguri, India, and its relationship to temperature change. J. Geovis. Spat. Anal. 4, 25. https://doi.org/10.1007/s41651-020-00065-4

Chan, K.L., Wiegner, M., Flentje, H., Mattis, I., Wagner, F., Gasteiger, J., Geiß, A. (2018). Evaluation of ECMWF-IFS (version 41R1) operational model forecasts of aerosol transport by using ceilometer network measurements. Geosci. Model Dev. 11, 3807–3831. https://doi.org/10.5194/gmd-11-3807-2018

Chen, B., Song, Z., Huang, J., Zhang, P., Hu, X., Zhang, X., Guan, X., Ge, J., Zhou, X. (2022a). Estimation of atmospheric PM_{10} concentration in China using an interpretable deep learning model and top-of-the-atmosphere reflectance data from China’s new generation geostationary meteorological satellite, FY-4A. J. Geophys. Res. 127, e2021JD036393. https://doi.org/10.1029/2021JD036393

Chen, G., Li, S., Knibbs, L.D., Hamm, N.A., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y. (2018). A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251

Chen, G., Li, Y., Zhou, Y., Shi, C., Guo, Y., Liu, Y. (2021). The comparison of AOD-based and non-AOD prediction models for daily $PM_{2.5}$ estimation in Guangdong province, China with poor AOD coverage. Environ. Res. 195, 110735. https://doi.org/10.1016/j.envres.2021.110735

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., XGBoost contributors (2022b). Package ‘xgboost’. R Reference Document, pp. 1–66. https://cran.r-project.org/web/packages/xgboost/xgboost.pdf (accessed 1 February 2023).

Czernecki, B., Marosz, M., Jędruszkiewicz, J. (2021). Assessment of machine learning algorithms in short-term forecasting of PM_{10} and $PM_{2.5}$ concentrations in selected Polish agglomerations. Aerosol Air Qual. Res. 21, 200586–200586. https://doi.org/10.4209/aaqr.200586

Danesh Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., Lyapustin, A., Katsouyanni, K., Schwartz, J. (2020). Predicting fine particulate matter ($PM_{2.5}$) in the greater london area: An ensemble approach using machine learning methods. Remote Sens. 12, 914. https://doi.org/10.3390/rs12060914

de Arruda Moreira, G., Guerrero-Rascado, J.L., Bravo-Aranda, J.A., Foyo-Moreno, I., Cazorla, A., Alados, I., Lyamani, H., Landulfo, E., Alados-Arboledas, L. (2020). Study of the planetary boundary layer height in an urban environment using a combination of microwave radiometer and ceilometer. Atmos. Res. 240, 104932. https://doi.org/10.1016/j.atmosres.2020.104932

Du, K., Wang, K., Shi, P., Wang, Y. (2013). Quantification of atmospheric visibility with dual digital cameras during daytime and nighttime. Atmos. Meas. Tech. 6, 2121–2130. https://doi.org/10.5194/amt-6-2121-2013

Ferrero, L., Riccio, A., Ferrini, B.S., D'Angelo, L., Rovelli, G., Casati, M., Angelini, F., Barnaba, F., Gobbi, G.P., Cataldi, M., Bolzacchini, E. (2019). Satellite AOD conversion into ground PM_{10}, $PM_{2.5}$ and PM_1 over the Po valley (Milan, Italy) exploiting information on aerosol vertical profiles, chemistry, hygroscopicity and meteorology. Atmos. Pollut. Res. 10, 1895–1912. https://doi.org/10.1016/j.apr.2019.08.003

Filonchyk, M. (2022). Characteristics of the severe March 2021 Gobi Desert dust storm and its impact on air pollution in China. Chemosphere 287, 132219. https://doi.org/10.1016/j.chemosphere.2021.132219

Filonchyk, M., Peterson, M. (2022). Development, progression, and impact on urban air quality of the dust storm in Asia in March 15–18, 2021. Urban Clim. 41, 101080. https://doi.org/10.1016/j.uclim.2021.101080

Founda, D., Kazadzis, S., Mihalopoulos, N., Gerasopoulos, E., Lianou, M., Raptis, P.I. (2016). Long-term visibility variation in Athens (1931–2013): a proxy for local and regional atmospheric aerosol loads. Atmos. Chem. Phys. 16, 11219–11236. https://doi.org/10.5194/acp-16-11219-2016

Gao, B., Ouyang, W., Cheng, H., Xu, Y., Lin, C., Chen, J. (2019). Interactions between rainfall and fine particulate matter investigated by simultaneous chemical composition measurements in downtown Beijing. Atmos. Environ. 218, 117000. https://doi.org/10.1016/j.atmosenv.2019.117000

Gao, J., Tian, H., Cheng, K., Lu, L., Zheng, M., Wang, S., Hao, J., Wang, K., Hua, S., Zhu, C., Wang, Y. (2015). The variation of chemical characteristics of $PM_{2.5}$ and PM_{10} and formation causes during two haze pollution events in urban Beijing, China. Atmos. Environ. 107, 1–8. https://doi.org/10.1016/j.atmosenv.2015.02.022

Guo, B., Wang, Y., Zhang, X., Che, H., Zhong, J., Chu, Y., Cheng, L. (2020). Temporal and spatial variations of haze and fog and the characteristics of $PM_{2.5}$ during heavy pollution episodes in China from 2013 to 2018. Atmos. Pollut. Res. 11, 1847–1856. https://doi.org/10.1016/j.apr.2020.07.019

Huang, G.B., Zhu, Q.Y., Siew, C.K. (2006). Extreme learning machine: theory and applications. Neurocomputing 70, 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Huang, Y., Yan, Q., Zhang, C. (2018). Spatial–temporal distribution characteristics of $PM_{2.5}$ in China in 2016. J. Geovis. Spat. Anal. 2, 1–18. https://doi.org/10.1007/s41651-018-0019-5

Hur, S.K., Ho, C.H., Kim, J., Oh, H.R., Koo, Y.S. (2021). Systematic bias of WRF-CMAQ PM_{10} simulations for Seoul, Korea. Atmos. Environ. 244, 117904. https://doi.org/10.1016/j.atmosenv.2020.117904

Ji, D., Deng, Z., Sun, X., Ran, L., Xia, X., Fu, D., Song, Z., Wang, P., Wu, Y., Tian, P., Huang, M. (2020). Estimation of $PM_{2.5}$ mass concentration from visibility. Adv. Atmos. Sci. 37, 671–678. https://doi.org/10.1007/s00376-020-0009-7

Jung, H., Um, J. (2022). Calculations of surface $PM_{2.5}$ concentrations using data from ceilometer backscatters and meteorological variables. J. Environ. Sci. Int. 31, 61–76. https://doi.org/10.5322/JESI.2022.31.1.61

Kim, B.Y., Lee, K.T., Jee, J.B., Zo, I.S. (2018). Retrieval of outgoing longwave radiation at top-of-atmosphere using Himawari-8 AHI data. Remote Sens. Environ. 204, 498–508. https://doi.org/10.1016/j.rse.2017.10.006

Kim, B.Y., Cha, J.W., Chang, K.H., Lee, C. (2021a). Visibility prediction over South Korea based on random forest. Atmosphere 12, 552. https://doi.org/10.3390/atmos12050552

Kim, B.Y., Cha, J.W., Chang, K.H. (2021b). Twenty-four-hour cloud cover calculation using a ground-based imager with machine learning. Atmos. Meas. Tech. 14, 6695–6710. https://doi.org/10.5194/amt-14-6695-2021

Kim, B.Y., Cha, J.W., Chang, K.H., Lee, C. (2022a). Estimation of the visibility in Seoul, South Korea, based on particulate matter and weather data, using machine-learning algorithm. Aerosol Air Qual. Res. 22, 220125. https://doi.org/10.4209/aaqr.220125

Kim, B.Y., Lim, Y.K., Cha, J.W., (2022b). Short-term prediction of particulate matter (PM_{10} and

$PM_{2.5}$) in Seoul, South Korea using tree-based machine learning algorithms. Atmos. Pollut. Res. 13, 101547. https://doi.org/10.1016/j.apr.2022.101547

Kim, B.Y., Belorid, M., Cha, J.W. (2022c). Short-term visibility prediction using tree-based machine learning algorithms and numerical weather prediction data. Weather Forecasting 37, 2263–2274. https://doi.org/10.1175/WAF-D-22-0053.1

Kim, D.B., Moon, Y.S. (2020). Causes of high $PM_{2.5}$ concentrations in Cheongju owing to non-Asian dust events. J. Korean Earth Sci. Soc. 41, 557–574. https://doi.org/10.5467/JKESS.2020.41.6.557

Kim, M., Lee, K., Lee, Y.H. (2020). Visibility data assimilation and prediction using an observation network in South Korea. Pure Appl. Geophys. 177, 1125–1141. https://doi.org/10.1007/s00024-019-02288-z

Kim, S.U., Kim, K.Y. (2020). Physical and chemical mechanisms of the daily-to-seasonal variation of PM_{10} in Korea. Sci. Total Environ. 712, 136429. https://doi.org/10.1016/j.scitotenv.2019.136429

Kim, Y.P., Lee, G. (2018). Trend of air quality in Seoul: Policy and science. Aerosol Air Qual. Res. 18, 2141–2156. https://doi.org/10.4209/aaqr.2018.03.0081

Kotthaus, S., O'Connor, E., Münkel, C., Charlton-Perez, C., Haeffelin, M., Gabey, A.M., Grimmond, C.S.B. (2016). Recommendations for processing atmospheric attenuated backscatter profiles from Vaisala CL31 ceilometers. Atmos. Meas. Tech. 9, 3769–3791. https://doi.org/10.5194/amt-9-3769-2016

Lee, S., Ho, C.H., Lee, Y.G., Choi, H.J., Song, C.K. (2013). Influence of transboundary air pollutants from China on the high-PM_{10} episode in Seoul, Korea for the period October 16–20, 2008. Atmos. Environ. 77, 430–439. https://doi.org/10.1016/j.atmosenv.2013.05.006

Li, S., Joseph, E., Min, Q., Yin, B., Sakai, R., Payne, M.K. (2017). Remote sensing of $PM_{2.5}$ during cloudy and nighttime periods using ceilometer backscatter. Atmos. Meas. Tech. 10, 2093–2104. https://doi.org/10.5194/amt-10-2093-2017

Lim, Y.K., Kim, B.Y., Chang, K.H., Cha, J.W., Lee, Y.H. (2022). Analysis of PM_{10} reduction effects with artificial rain enhancement using numerical models. Atmosphere 32, 341–351, https://doi.org/10.14191/Atmos.2022.32.4.341

Lou, C., Liu, H., Li, Y., Peng, Y., Wang, J., Dai, L. (2017). Relationships of relative humidity with $PM_{2.5}$ and PM_{10} in the Yangtze River Delta, China. Environ. Monit. Assess. 189, 1–16. https://doi.org/10.1007/s10661-017-6281-z

Lu, H., Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169

Ma, J., Yu, Z., Qu, Y., Xu, J., Cao, Y. (2020). Application of the XGBoost machine learning method in $PM_{2.5}$ prediction: A case study of Shanghai. Aerosol Air Qual. Res. 20, 128–138. https://doi.org/10.4209/aaqr.2019.08.0408

Ma, Y., Zhu, Y., Liu, B., Li, H., Jin, S., Zhang, Y., Fan, R., Gong, W. (2021). Estimation of the vertical distribution of particle matter ($PM_{2.5}$) concentration and its transport flux from lidar measurements based on machine learning algorithms. Atmos. Chem. Phys. 21, 17003–17016. https://doi.org/10.5194/acp-21-17003-2021

Martínez, F., Frías, M.P., Charte, F., Rivera, A.J. (2019). Time series forecasting with KNN in R: the tsfknn Package. R J. 11, 229–242. https://doi.org/10.32614/RJ-2019-004

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C. (2021). Package ‘e1071’. R Reference Document, pp. 1–66. https://cran.r-project.org/web/packages/e1071/e1071.pdf (accessed 1 February 2023).

Minh, V.T.T., Tin, T.T., Hien, T.T. (2021). $PM_{2.5}$ forecast system by using machine learning and WRF model, a case study: Ho Chi Minh City, Vietnam. Aerosol Air Qual. Res. 21, 210108. https://doi.org/10.4209/aaqr.210108

Ministry of Environment (MOE) (2021). Air pollution monitoring network installation and operation manual. Ministry of Environment, Korea. pp. 1–676. https://www.me.go.kr/home/file/readDownloadFile.do?fileId=222422&fileSeq=1 (accessed 1 February 2023).

Münkel, C., Eresmaa, N., Räsänen, J., Karppinen, A. (2007). Retrieval of mixing height and dust concentration with lidar ceilometer. Boundary Layer Meteorol. 124, 117–128. https://doi.org/10.1007/s10546-006-9103-3

Oh, H.R., Ho, C.H., Kim, J., Chen, D., Lee, S., Choi, Y.S., Chang, L.S., Song, C.K. (2015). Long-range transport of air pollutants originating in China: A possible major cause of multi-day high-PM_{10} episodes during cold season in Seoul, Korea. Atmos. Environ. 109, 23–30. https://doi.org/10.1016/j.atmosenv.2015.03.005

Oh, H.R., Ho, C.H., Koo, Y.S., Baek, K.G., Yun, H.Y., Hur, S.K., Choi, D.R., Jhun, J.G., Shim, J.S. (2020). Impact of Chinese air pollutants on a record-breaking PMs episode in the Republic of Korea for 11–15 January 2019. Atmos. Environ. 223, 117262. https://doi.org/10.1016/j.atmosenv.2020.117262

Pappa, A., Kioutsioukis, I. (2021). Forecasting particulate pollution in an urban area: From Copernicus to Sub-Km scale. Atmosphere 12, 881. https://doi.org/10.3390/atmos12070881

Parde, A.N., Ghude, S.D., Pithani, P., Dhangar, N.G., Nivdange, S., Krishna, G., Lal, D., Jenamani, R., Singh, P., Jena, C., Karumuri, R., Safai, P., Chate, D. (2020). Estimation of surface particulate matter ($PM_{2.5}$ and PM_{10}) mass concentrations from ceilometer backscattered profiles. Aerosol Air Qual. Res. 20, 1640–1650. https://doi.org/10.4209/aaqr.2019.08.0371

Peterson, D.A., Hyer, E.J., Han, S.O., Crawford, J.H., Park, R.J., Holz, R., Kuehn, R.E., Eloranta, E., Knote, C., Jordan, C.E., Lefer, B.L. (2019). Meteorology influencing springtime air quality, pollution transport, and visibility in Korea. Elem. Sci. Anth. 7, 57. https://doi.org/10.1525/elementa.395

Rosa, J.P.S., Guerra, D.J.D., Horta, N.C.G., Martins, R.M.F., Lourenço, N.C.C. (2020). Overview of Artificial Neural Networks, in: Rosa, J.P.S., Guerra, D.J.D., Horta, N.C.G., Martins, R.M.F., Lourenço, N.C.C. (Eds.), Using Artificial Neural Networks for Analog Integrated Circuit Design Automation, Springer International Publishing, Cham, pp. 21–44. https://doi.org/10.1007/978-3-030-35743-6_3

Shin, J.Y., Kim, B.Y., Park, J., Kim, K.R., Cha, J.W. (2020a). Prediction of leaf wetness duration using geostationary satellite observations and machine learning algorithms. Remote Sens. 12, 3076. https://doi.org/10.3390/rs12183076

Shin, M., Kang, Y., Park, S., Im, J., Yoo, C., Quackenbush, L.J. (2020b). Estimating ground-level particulate matter concentrations using satellite-based data: A review. GIScience. Remote Sens. 57, 174–189. https://doi.org/10.1080/15481603.2019.1703288

Sun, X., Zhao, T., Liu, D., Gong, S., Xu, J., Ma, X. (2020). Quantifying the influences of $PM_{2.5}$ and relative humidity on change of atmospheric visibility over recent winters in an urban area of East China. Atmosphere 11, 461. https://doi.org/10.3390/atmos11050461

Taghizadeh-Mehrjardi, R., Neupane, R., Sood, K., Kumar, S. (2017). Artificial bee colony feature selection algorithm combined with machine learning algorithms to predict vertical and lateral distribution of soil organic matter in South Dakota, USA. Carbon Manage. 8, 277–291. https://doi.org/10.1080/17583004.2017.1330593

Taylor, K.E. (2001). Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106. 7183–7192. https://doi.org/10.1029/2000JD900719

Wang, F., Chen, D.S., Cheng, S.Y., Li, J.B., Li, M.J., Ren, Z.H. (2010). Identification of regional atmospheric PM_{10} transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis. Environ. Modell. Software 25, 927–934. https://doi.org/10.1016/j.envsoft.2010.02.004

Wen, X., Zhang, P., Liu, D. (2018). Spatiotemporal variations and influencing factors analysis of $PM_{2.5}$ concentrations in Jilin Province, Northeast China. Chin. Geogr. Sci. 28, 810–822. https://doi.org/10.1007/s11769-018-0992-0

Whalley, J., Zandi, S. (2016). Particulate Matter Sampling Techniques and Data Modelling Methods, in: Sallis, P. (Ed.), Air Quality, IntechOpen, Rijeka, p. Ch. 2. https://doi.org/10.5772/65054

Wiegner, M., Gasteiger, J. (2015). Correction of water vapor absorption for aerosol remote sensing with ceilometers. Atmos. Meas. Tech. 8, 3971–3984. https://doi.org/10.5194/amt-8-3971-2015

Wright, M.N., Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01

Wright, M.N., Wager, S., Probst, P. (2020). Package ‘ranger’. R Reference Document, pp. 1–25. https://cran.r-project.org/web/packages/ranger/ranger.pdf (accessed 1 February 2023).

Yang, G., Lee, H., Lee, G. (2020). A hybrid deep learning model to forecast particulate matter concentration levels in Seoul, South Korea. Atmosphere 11, 348. https://doi.org/10.3390/atmos11040348

Zhang, S., Cheng, D., Deng, Z., Zong, M., Deng, X. (2018). A novel kNN algorithm with data-drive k parameter computation. Pattern Recognit. Lett. 109, 44–54. https://doi.org/10.1016/j.patrec.2017.09.036

Zhang, Z., Wu, W., Wei, J., Song, Y., Yan, X., Zhu, L., Wang, Q. (2017). Aerosol optical depth retrieval from visibility in China during 1973–2014. Atmos. Environ. 171, 38–48. https://doi.org/10.1016/j.atmosenv.2017.09.004

Zhou, Y., Lu, Z., Cheng, K. (2019). Sparse polynomial chaos expansions for global sensitivity analysis with partial least squares and distance correlation. Struct. Multidiscip. Optim. 59, 229–247. https://doi.org/10.1007/s00158-018-2062-8