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

Bu-Yo Kim$^*$, Joo Wan Cha$^1$, Yong Hee Lee$^1$

$^1$Research Applications Department, National Institute of Meteorological Sciences, Seogwipo, Jeju 63568, South Korea

* Corresponding author. Tel: +82-64-780-6695
E-mail address: kimbuyo@korea.kr

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 µg m$^{-3}$, root mean square error (RMSE) = 14.44 µg m$^{-3}$, and R = 0.92 for PM$_{10}$; and bias = 0.12 µg m$^{-3}$, RMSE = 7.16 µg 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
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 \( \text{PM}_{2.5} \) (particles < 2.5 µm) and \( \text{PM}_{10} \) (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, Ninh 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).

\( \text{PM}_{2.5} \) and \( \text{PM}_{10} \) 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., 2020). 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., 2020; 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\textsubscript{10} and PM\textsubscript{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\textsubscript{10} and PM\textsubscript{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\textsubscript{10} and PM\textsubscript{2.5} (PM\textsubscript{est} and PM\textsubscript{obs}, respectively) is presented in Equations (1)–(3) and evaluated using the bias, root mean square error (RMSE), and correlation coefficient (R).

\[
\text{bias} = \frac{\sum_{i=1}^{N} (PM_{\text{est}i} - PM_{\text{obs}i})}{N}
\]  
\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (PM_{\text{est}i} - PM_{\text{obs}i})^2}{N}}
\]  
\[
R = \frac{\sum_{i=1}^{N} (PM_{\text{est}i} - PM_{\text{est}})(PM_{\text{obs}i} - PM_{\text{obs}})}{\sqrt{\sum_{i=1}^{N} (PM_{\text{est}i} - PM_{\text{est}})^2 \sum_{i=1}^{N} (PM_{\text{obs}i} - PM_{\text{obs}})^2}}
\]

Here, N means the number of data.

2.2 ML Methods

In this study, PM\textsubscript{10} and PM\textsubscript{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; Kim et al., 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., 2022). 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. 1a), as illustrated in Fig. 1b, LGB has relatively fast learning and prediction speed by growing the tree leaf-wise (Al Banna et al., 2020; Lu and Ma, 2020).

### 2.2.2 Vector-based methods

SVR regresses data based on the $\varepsilon$-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. 1c (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. 1d, 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. 1e and 1f (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 Equations (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 \((N \times k)\) extracted from the component set, \( D \) \((k \times n)\) and \( Q \) \((k \times n)\).
1) are the loading matrix and vector for $X$ and $Y$, and $E$ ($N \times n$) and $F$ ($N \times 1$) are the matrix and vector of residuals, respectively.

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. 2a. 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 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 (Fig. 3a and 3c) and PM and between the mean time and layer-mean backscatter coefficient for each layer (Fig. 3b and 3d) (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; Kim et al., 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, $T_{\text{dep}}$, 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 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 $T_{\text{dep}}$ (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).

### 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. 6a and 6b, 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 $\geq 200$ 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; Kim et al., 2022a; Kim et al., 2022b; Kim et al., 2022c).

The results of the daily mean time series (Fig. 6c) also showed estimation similar to that of the observed PM, with the estimation accuracy for PM$_{10}$ having a bias of $-0.27 \mu g \ m^{-3}$, an RMSE of $5.98 \mu g \ m^{-3}$, and an $R$ of 0.97, and that for PM$_{2.5}$ having a bias of $-0.18 \mu g \ m^{-3}$, an RMSE of 3.18 $\mu g \ m^{-3}$, 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. (2022) ($R = 0.91$ for PM$_{10}$), who estimated the mean daily PM using satellite data.

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; Gue 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.
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 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


in short-term forecasting of PM10 and PM2.5 concentrations in selected Polish agglomerations. Aerosol Air Qual. Res. 21, 200586–200586. https://doi.org/10.4209/aaqr.200586


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

<table>
<thead>
<tr>
<th>Algorithm (R library)</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RF</strong> (ranger)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>num.trees = 940, mtry = 4, min.node.size = 4</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>num.trees = 890, mtry = 3, min.node.size = 3</td>
</tr>
<tr>
<td><strong>XGB</strong> (xgboost)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>nrounds = 1010, max_depth = 5, eta = 0.10</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>nrounds = 920, max_depth = 9, eta = 0.10</td>
</tr>
<tr>
<td><strong>LGB</strong> (lightgbm)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>nrounds = 620, max_depth = 7, learning_rate = 0.19</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>nrounds = 490, max_depth = 9, learning_rate = 0.19</td>
</tr>
<tr>
<td><strong>SVR</strong> (e1071)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>cost = 5, gamma = 0.11, epsilon = 0.09</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>cost = 15, gamma = 0.11, epsilon = 0.10</td>
</tr>
<tr>
<td><strong>kNN</strong> (kknn)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>k = 8</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>k = 22</td>
</tr>
<tr>
<td><strong>ANN</strong> (nnet)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>maxit = 980, size = 7, decay = 0.5</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>maxit = 950, size = 9, decay = 0.5</td>
</tr>
<tr>
<td><strong>ELM</strong> (Kim et al., 2022a)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>nhidden = 1360</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>nhidden = 650</td>
</tr>
</tbody>
</table>
Table 2. Estimated accuracy of monthly PM$_{10}$ and PM$_{2.5}$ and monthly mean meteorological variables for the test set. The unit of $\beta$ (backscatter coefficient) is $10^{-5}$ sr$^{-1}$ km$^{-1}$, and the unit of mean, bias, and RMSE is $\mu$g m$^{-3}$.

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

<table>
<thead>
<tr>
<th>Month</th>
<th>PM$_{10}$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Bias</td>
<td>RMSE</td>
<td>R</td>
<td>Mean</td>
<td>Bias</td>
<td>RMSE</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>46.00</td>
<td>0.91</td>
<td>10.55</td>
<td>0.90</td>
<td>29.04</td>
<td>0.36</td>
<td>6.87</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>49.89</td>
<td>0.03</td>
<td>12.56</td>
<td>0.88</td>
<td>31.89</td>
<td>-0.29</td>
<td>7.08</td>
<td>0.93</td>
</tr>
<tr>
<td>3</td>
<td>58.10</td>
<td>1.84</td>
<td>25.16</td>
<td>0.84</td>
<td>30.37</td>
<td>0.50</td>
<td>9.13</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>45.35</td>
<td>-0.35</td>
<td>16.36</td>
<td>0.75</td>
<td>18.73</td>
<td>0.58</td>
<td>6.77</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>62.20</td>
<td>-1.04</td>
<td>25.51</td>
<td>0.96</td>
<td>23.85</td>
<td>-0.15</td>
<td>8.45</td>
<td>0.92</td>
</tr>
<tr>
<td>6</td>
<td>42.11</td>
<td>-0.57</td>
<td>9.80</td>
<td>0.89</td>
<td>24.62</td>
<td>0.06</td>
<td>6.67</td>
<td>0.85</td>
</tr>
<tr>
<td>7</td>
<td>25.91</td>
<td>1.11</td>
<td>7.53</td>
<td>0.86</td>
<td>15.41</td>
<td>0.83</td>
<td>5.79</td>
<td>0.85</td>
</tr>
<tr>
<td>8</td>
<td>31.19</td>
<td>-0.33</td>
<td>7.71</td>
<td>0.84</td>
<td>15.58</td>
<td>-0.17</td>
<td>5.80</td>
<td>0.81</td>
</tr>
<tr>
<td>9</td>
<td>31.38</td>
<td>-0.15</td>
<td>8.76</td>
<td>0.71</td>
<td>14.28</td>
<td>-0.09</td>
<td>6.28</td>
<td>0.57</td>
</tr>
<tr>
<td>10</td>
<td>47.89</td>
<td>1.58</td>
<td>9.72</td>
<td>0.90</td>
<td>27.00</td>
<td>0.46</td>
<td>6.66</td>
<td>0.89</td>
</tr>
<tr>
<td>11</td>
<td>53.88</td>
<td>-0.86</td>
<td>13.16</td>
<td>0.90</td>
<td>30.74</td>
<td>0.14</td>
<td>7.93</td>
<td>0.92</td>
</tr>
<tr>
<td>12</td>
<td>48.51</td>
<td>-1.56</td>
<td>10.60</td>
<td>0.91</td>
<td>29.82</td>
<td>-0.75</td>
<td>7.21</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Table 3. Estimation accuracy of PM$_{10}$ and PM$_{2.5}$ according to the visibility range for the test data set.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Visibility range (km)</th>
<th>N</th>
<th>Mean ($\mu g m^{-3}$)</th>
<th>Bias ($\mu g m^{-3}$)</th>
<th>RMSE ($\mu g m^{-3}$)</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{10}$</td>
<td>10-20</td>
<td>1908</td>
<td>36.67</td>
<td>0.51</td>
<td>11.81</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>1-10</td>
<td>716</td>
<td>70.20</td>
<td>-1.01</td>
<td>19.82</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>1-5</td>
<td>226</td>
<td>88.70</td>
<td>-0.32</td>
<td>25.65</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>10-20</td>
<td>1908</td>
<td>18.37</td>
<td>0.43</td>
<td>6.35</td>
<td>0.78</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>1-10</td>
<td>716</td>
<td>41.80</td>
<td>-0.73</td>
<td>8.95</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>1-5</td>
<td>226</td>
<td>53.57</td>
<td>-0.69</td>
<td>9.72</td>
<td>0.89</td>
</tr>
</tbody>
</table>
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; Kim et al., 2022a; Kim et al., 2022b).
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 ($\beta$, mean of 2nd–20th layers), and precipitation, b) temperature (T), dewpoint depression ($T_{\text{dep}}$), and relative humidity (RH), and c) pressure, wind speed (WS), and wind direction (WD).
Fig. 3. Correlation coefficients between the backscatter coefficient and PM$_{10}$ and PM$_{2.5}$ by each mean time (a, c) and between the backscatter coefficient and PM$_{10}$ and PM$_{2.5}$ by each mean time and mean vertical layers (b, d). β means the backscatter coefficient.
Fig. 4. Taylor plot of PM$_{10}$ (a, b) and PM$_{2.5}$ (c, d) estimation results for training and validation.
Fig. 5. Relative importance of each input variable in PM$_{10}$ (a) and PM$_{2.5}$ (b) estimation based on the XGB method.
Fig. 6. Hourly scatter plot (log scale) (a, b) and daily mean time series (c) 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.
Fig. 7. PM$_{10}$ (a) and PM$_{2.5}$ (b) 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.