Multiple PM Low-Cost Sensors, Multiple Seasons’ Data, and Multiple Calibration Models

Srishti S1, Pratyush Agrawal1, Padmavati Kulkarni2, Hrishikesh Chandra Gautam1, Meenakshi Kushwaha2, V. Sreekanth1

1 Center for Study of Science, Technology & Policy, Bengaluru 560094, India
2 ILK Labs, Bengaluru 560046, India

ABSTRACT

In this study, we combined state-of-the-art data modelling techniques (machine learning [ML] methods) and data from state-of-the-art low-cost particulate matter (PM) sensors (LCSs) to improve the accuracy of LCS-measured PM2.5 (PM with aerodynamic diameter less than 2.5 microns) mass concentrations. We collocated nine LCSs and a reference PM2.5 instrument for 9 months, covering all local seasons, in Bengaluru, India. Using the collocation data, we evaluated the performance of the LCSs and trained around 170 ML models to reduce the observed bias in the LCS-measured PM2.5. The ML models included (i) Decision Tree, (ii) Random Forest (RF), (iii) eXtreme Gradient Boosting, and (iv) Support Vector Regression (SVR). A hold-out validation was performed to assess the model performance. Model performance metrics included (i) coefficient of determination ($R^2$), (ii) root mean square error (RMSE), (iii) normalised RMSE, and (iv) mean absolute error. We found that the bias in the LCS PM2.5 measurements varied across different LCS types (RMSE = 8–29 $\mu g m^{-3}$) and that SVR models performed best in correcting the LCS PM 2.5 measurements. Hyperparameter tuning improved the performance of the ML models (except for RF). The performance of ML models trained with significant predictors (fewer in number than the number of all predictors, chosen based on recursive feature elimination algorithm) was comparable to that of the ‘all predictors’ trained models (except for RF). The performance of most ML models was better than that of the linear models. Finally, as a research objective, we introduced the collocated black carbon mass concentration measurements into the ML models but found no significant improvement in the model performance.

Keywords: Plantower, Beta attenuation monitor, Support vector regression

1 INTRODUCTION

Over the last decade, air pollution low-cost sensors (LCSs) have become popular and are complementing the existing air pollution monitoring capacity around the world (Kumar et al., 2015; Rai et al., 2017; Gupta et al., 2018; Morawska et al., 2018). The number of air quality studies using LCSs has tremendously increased in recent years. LCSs are easier to handle, install, and maintain than reference-grade monitors. Further, they are capable of measuring air pollutants at a high temporal resolution and can improve the granularity of monitoring. Strategically placed LCSs can provide detailed information on air quality and its variability within a region.

The affordability and simplicity of LCSs, however, come with the trade-off of accuracy. In general, LCS measurements of air pollutants are often less accurate than reference-grade measurements (Clements et al., 2017). In case of particulate matter (PM) LCSs, most LCSs quantify PM mass concentrations using the light scattering (nephelometric) principle. This technique is sensitive to aerosol microphysical properties and environmental factors (e.g., aerosol size distribution, aerosol refractive index, and humidity) in addition to the particle mass concentration. Moreover, LCSs suffer from declining sensing accuracy with age. These aspects can introduce bias in PM measurements,
thereby requiring evaluation and correction to ensure accuracy. A common practice for evaluating the performance of LCSs and deriving correction factors for LCS-measured pollutant concentrations is to collocate the LCS and reference-grade instrument and analyse the collocation data. Several studies have applied a range of training-based models (from simple linear regression to machine learning [ML] algorithms) to the collocation data to derive regression coefficients/functions that have been used to correct the LCS-measured PM$_{2.5}$ (e.g., Barkjohn et al., 2020; deSouza et al., 2022). Studies have shown that data correction increases the accuracy of LCS PM$_{2.5}$ measurements and the corrected values are comparable to reference-grade PM$_{2.5}$ measurements (Tryner et al., 2020; McFarlane et al., 2021; deSouza et al., 2022; Sreekanth et al., 2022).

One of the most extensive and systematic evaluation of LCS measurements has been performed by the Air Quality Sensor Performance Evaluation Center (AQ-SPEC; www.aqmd.gov/aqspect) of the South Coast Air Quality Management District, United States. The program evaluated 39 PM LCSs based on chamber experiments and field collocations and found that the performance of the LCSs considerably varied among manufacturers and models (AQ-SPEC 2019). However, similar institutional-level LCS evaluation facilities are lacking in other countries, especially in developing countries where regulatory PM$_{2.5}$ monitoring devices are scarce and sparsely located (Brauer et al., 2019). Given the extreme pollutant concentrations and heterogeneous sources in LMICs that differ from those in high-income countries with a capacity of extensive LCS testing, it is critical that LCSs are rigorously evaluated in different LMIC settings and the measurements are corrected accordingly for accuracy.

In India, studies have applied statistical models to correct LCS measurements of PM$_{2.5}$ (Puttaswamy et al., 2022; Sreekanth et al., 2022), but a limited number of studies have used ML (Kumar and Sahu, 2021) methods. In this study, we investigated the performance of multiple PM$_{2.5}$ LCSs and trained several ML models using collocation data to correct the hourly mean LCS PM$_{2.5}$. The collocation experiment was conducted in Bengaluru city (south India), and the study period covered all major seasons (December 2021–August 2022). To our knowledge, this is one of the first studies from LMICs to evaluate multiple PM$_{2.5}$ LCSs at one geographical location. As a case study, we also introduced collocated black carbon (BC) mass concentration measurements as an additional predictor in the ML models to investigate any possible improvement in model performance.

2 MATERIALS AND METHODS

In total, nine (Fig. S1) PM$_{2.5}$ LCSs were collocated with a beta attenuation monitor (BAM, a reference instrument for measuring PM$_{2.5}$) on the roof terrace of the Center for Study of Science, Technology, and Policy (CSTEP) building. CSTEP (13.04°N, 77.57°E) is located in the northern part of Bengaluru city. Bengaluru is the administrative capital of Karnataka and is located at an elevation of 900 m above mean sea level. The city experiences a tropical savanna climate around the year, with an annual rainfall of ~960 mm (June, July, August, and September are the monsoon months). The annual PM$_{2.5}$ concentrations are about ~27 µg m$^{-3}$, with higher values during winter (~35 µg m$^{-3}$), followed by pre-monsoon, post-monsoon, and monsoon seasons (Prabhu et al., 2022). In this study, collocated PM$_{2.5}$ measurements from BAM and the LCSs were obtained for 9 months from December 2021 to August 2022.

2.1 BAM

In the current study, we used a BAM (BAM1022; Met One Instruments, Inc., Grants Pass, USA) to measure hourly mean ambient PM$_{2.5}$ levels. BAM1022 is a United States Environmental Protection Agency-certified Federal Equivalent Method class instrument for measuring PM$_{2.5}$ levels. BAM1022 uses C$^{14}$ as a beta particle source and operates at a nominal flow rate of 16.67 litres per minute. Based on the difference in the attenuation of the glass fibre filter tape before and after PM$_{2.5}$ loading and the flow rate, BAM estimates the PM$_{2.5}$ mass concentration. The PM$_{2.5}$ measurement range of the BAM is between ~15 µg m$^{-3}$ and 10,000 µg m$^{-3}$. The BAM comprises a heater that removes moisture from the sampled ambient airflow. It is equipped with a meteorological sensor that is capable of measuring ambient temperature, relative humidity (RH), and pressure. More details on the BAM and the precision of its PM$_{2.5}$ measurements have
been reported previously (Kushwaha et al., 2022). PM$_{2.5}$ data from the hourly channel of the BAM was used for our analyses and model training.

2.2 LCSs

All LCSs used in the study were compact, Internet of Things (IoT)-based devices comprising a laser PM and meteorological sensor. The following LCSs were used: Aerogram (https://aerogram.in/), Airveda (https://www.airveda.com/), Atmos I and Atmos II (http://urbansciences.in/), BlueSky (https://tsi.com/), PAQS (https://paqs.biz/), Prana Air (https://www.pranaair.com/), Prkuti (https://www.prkruti.com/), and PurpleAir (https://www2.purpleair.com). The internal laser PM sensor consists of a micro fan to draw the ambient air inside the optical chamber, where particles are detected using the light scattering technique. Data logging and averaging intervals of these LCSs varied between 30 sec and 30 min. All LCSs were equipped with a meteorological sensor capable of measuring the temperature and RH. The PM$_{2.5}$ measurement range of the LCSs was between 0 and 1,000 µg m$^{-3}$. Most LCSs are equipped with a microSD card, which stores data locally in addition to cloud storage. Two versions of Atmos were used in this study, with a different basic laser PM sensor in each. Plantower-based Atmos was named Atmos I, whereas Sensirion-based Atmos was named Atmos II. Most LCSs investigated in this study had Plantower (PMS5003/PMS7003) as the internal laser PM sensor, whereas other LCSs were equipped with Sensirion, Nova, Honeywell, PAS-OUT-01, and Winsen laser sensors. PurpleAir was equipped with dual Plantower sensors and output PM$_{2.5}$ data in two channels labelled CF_1 and CF_ATM (Barkjohn et al., 2020). We trained individual ML models for the PM$_{2.5}$ values from PurpleAir for both CF_1 and CF_ATM channels. Technical and operational details of the LCSs are listed in Table S1.

2.3 Aethalometer (AE33)

We used a rack-mount Aethalometer (AE33, Aerosol Co. Ljubljana, SI) to measure BC mass concentrations. AE33 measures filter attenuation (before and after aerosol loading) at seven wavelengths and is capable of providing high temporal-resolution PM absorption mass concentration data. Values measured at 880-nm wavelength were considered BC mass concentrations. AE33 uses DualSpot™ technology that compensates for loading errors, which are commonly observed in most filter-based optical analysers. The instrument was configured to operate at a flow rate of 2 litres per minute and log 1-min average concentrations. A 2.5-micron cut cyclone was installed in the inlet of AE33 to allow particles smaller than 2.5 µm into the detection chamber.

2.4 ML Models

We explored four different ML models: Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR). All ML models (except XGBoost) were trained using the scikit-learn package in Python programming language. We used the xgboost package for training the XGBoost models.

2.4.1 DT

DT is a rule-based model that divides the entire dataset into homogeneous groups. Briefly, it iteratively splits the dataset into regions based on predictors, which result in a maximum reduction of the error. A new sample is estimated as the average value of the training set true value at the terminal node, assigned based on the rule defined by the tree on predictors. The DT regression models have few limitations: (i) high variance, (ii) less predictive performance if the relation between predictors and the response variable is not defined accurately, and (iii) limited predicted values (based on the number of terminal nodes).

2.4.2 RF

RF is an ensemble method where parallel DTs are grown on bootstrapped random samples, i.e., subset samples drawn with replacement from the dataset (Breiman, 2001). Each DT is built on a subset of predictors, which introduces a reduction of correlation between trees. For a new sample, the prediction from each tree is averaged. One of the drawbacks of RF regression is that the predictions are always within the range of the training samples' true values.
2.4.3 XGBoost

The XGBoost is an optimised distributed gradient boosting library which implements machine learning algorithms under gradient boosting framework also known as Gradient boosting machine (GBM). It incorporates ensemble model which grow in a stage wise manner where each weak learner (regression tree) is fitted on the residual from the previous learner (Chen and Guestrin, 2016). Unlike DTs, the leaves of each of the regression trees are assigned a score. Based on the rule of the DT, each new observation is classified to the leaves of the regression trees at each iteration, and the final prediction is the sum of the corresponding scores at the leaf of each tree. To prevent overfitting, an additional regularisation parameter is added to the objective function. The newly added leaf scores are shrunk by a factor eta, and a column subsample is used to find the best feature to split.

2.4.4 SVR

Linear SVR estimates the relation between input and output variables by finding a linear function such that the deviation of the training points from the true value is less than or equal to a specified margin called the maximum error epsilon (\(\varepsilon\)) and the function be as flat as possible (Smola and Schölkopf, 2004). The deviation \(\varepsilon\) from the true value is described as a tube around the function. The deviations outside the tube are accounted for errors. In case of non-linear SVR, the input space is mapped to a new feature space and linear SVR is applied to that space. The kernel functions of the support vector machine represent the dot product in the new feature space. The drawback of a support vector machine is that it is scale invariant and is more suited for smaller datasets as the computation cost or time increases with an increase in training data points. In our study, z-transformed datasets were used for training the non-linear SVR models.

2.5 Predictors

We used three types of predictor variables to train the ML models: (i) continuous variables, (ii) categorical variables, and (iii) cyclic variables. The continuous variables included mass concentrations of all PM size fractions obtained from the LCSs, meteorological parameters (temperature and RH) from the LCSs, and BAM PM2.5 measurements. As a case study, we used collocated BC data as a predictor, which is a continuous variable. Using the timestamp of the data, we created new categorical variables related to (i) hour of the day, (ii) month of the year, (iii) season, and (iv) weekend/weekday. These variables were transformed before using them in the modelling exercise. We converted weekend/weekday and season-related variables to dummy variables; ‘hour of the day’ and ‘month of the year’ were converted to cyclic values using both sine and cosine transformations. Reference PM2.5 (BAM PM2.5) was used as the response variable.

2.6 Hyperparameter Tuning

All four types of ML models were investigated for their performance in correcting the LCS PM2.5 measurements with and without hyperparameter tuning. ML models with default hyperparameters in the scikit-learn/xgboost packages were termed as untuned models, whereas ML models with hyperparameter tuning performed using the Grid Search algorithm were termed as tuned models. To evaluate the effect of tuning, all ML models using all predictors were trained with and without tuning. The hyperparameter space for the different LCS and ML models is presented in Table S2.

2.7 Recursive Feature Elimination (RFE)

Simple ML models with significant and uncorrelated features can run more efficiently by occupying less computational space and having less execution time. Therefore, we defined three sets of predictors for each of the LCSs for the ML model training: (i) all predictors, (ii) uncorrelated predictors, and (iii) significant predictors. To arrive at the uncorrelated predictor sets for each LCS, we calculated the pairwise Karl Pearson correlation coefficient for all predictors from the complete list and dropped one predictor if the coefficient was greater than the defined threshold of 0.9. In the next step, we used the RFE feature of the scikit-learn library to further narrow down the predictor list for each of the LCSs. We selected the optimal number of predictors (significant predictors) by comparing the R^2 value of linear regression models trained on different numbers
of predictors selected using RFE. The significant predictors for each LCS were finalised when insignificant improvement in $R^2$ was observed even after adding further predictors to the linear regression. The list of predictors for all three types of predictor sets for each of the LCSs is provided in Table S3.

### 2.8 Cross-validation and Performance Metrics

To understand a model’s performance on unseen data, we performed a hold-out validation exercise for all models trained, wherein 75% of the data were used for model training and remaining 25% were used for testing. In addition, the test data were unseen by the model hyperparameter tuning. The accuracy of model-corrected PM$_{2.5}$ was quantified based on the (i) coefficient of determination ($R^2$), (ii) root mean square error (RMSE), (iii) normalised root mean square error (NRMSE), and (iv) mean absolute error (MAE).

\[
R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\]

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}}
\]

\[
NRMSE = \frac{RMSE}{\bar{Y}}
\]

\[
MAE = \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{n}
\]

where $Y_i$, $\bar{Y}$, and $\hat{Y}_i$ represent true value, mean of the true value, and estimated value, respectively, and $n$ is the number of paired data points. An increase in $R^2$ indicates improvement in the performance, whereas a decrease in RMSE, NRMSE, and MAE indicates improvement in the performance.

### 3 RESULTS

The PM and meteorological data from the LCSs and BC from AE33 were averaged to 1-h intervals to match the temporal resolution of hourly BAM datasets. PurpleAir PM$_{2.5}$ data were quality checked based on the difference between their values from the dual Plantower sensors, following Barkjohn et al. (2020). Fill values from all devices were removed from the analysis. The hourly PM$_{2.5}$ data availability chart is shown in Fig. S2. Prkruti PM$_{2.5}$ had the highest amount of data unavailability (due to instrument malfunction and IoT issues), and Atmos II was installed in February 2022. Records having data for all predictor variables and response variable were only considered for the model training.

#### 3.1 Bias in LCS PM$_{2.5}$

The average value of the BAM PM$_{2.5}$ for the study period was $\sim$32 µg m$^{-3}$. Detailed statistics on the hourly concentrations of PM$_{2.5}$ from BAM and LCSs are given in Table S4. Scatter plots between hourly uncorrected LCS and BAM PM$_{2.5}$ revealed that the bias of the LCS PM$_{2.5}$ was different across various sensors (Fig. 1). All LCS PM$_{2.5}$ values maintained a linear relationship with BAM PM$_{2.5}$. Across the LCSs, the $R^2$ values of the linear fit varied between 0.63 and 0.89. The bias of the LCS PM$_{2.5}$ (in terms of RMSE) varied between 8 µg m$^{-3}$ and 29 µg m$^{-3}$. The NRMSE of the
LCS PM$_{2.5}$ ranged between 0.26 and 0.89. The observed bias could be because of the differences in the geometry of the optics chamber and the wavelengths used in the laser PM sensors (e.g., Hapidin et al., 2019). The highest bias was observed in PAQS PM$_{2.5}$, while the lowest was in Atmos I PM$_{2.5}$. PAQS highly underestimated the reference PM$_{2.5}$. Comparably, Plantower-based LCSs (Aerogram, Atmos I, and PurpleAir) performed better. Atmos II and BlueSky consist of the same laser PM sensor (Sensirion), and their RMSE values ranged between 11 µg m$^{-3}$ and 15 µg m$^{-3}$. The performance metrics of the uncorrected LCS PM$_{2.5}$ are listed in Table 1. Further, the performance of all LCSs was evaluated on a seasonal scale. The calendar year was divided into four seasons, namely, winter (JF), pre-monsoon (MAM), monsoon (JJAS), and post-monsoon (OND). In the current study, OND comprised only of the December 2021 data. The PM$_{2.5}$ values were lower during the monsoon season, followed by pre-monsoon, winter, and post-monsoon. Scatter plots between uncorrected LCS PM$_{2.5}$ and BAM PM$_{2.5}$ (data obtained during different seasons are shown in different colours) are shown in Fig. S3, and the corresponding performance metrics are given in Tables S5 and S6. Box plots of season-wise performance metrics of uncorrected LCS PM$_{2.5}$ are given in Fig. 2. Relative to other seasons, the performance (in terms of NRMSE) of the LCSs during monsoon season was poor. No seasonality in the performance was observed during other seasons. The observed bias in the LCS PM$_{2.5}$ was consistent with that in previous laboratory and field evaluations (Badura et al., 2018; Feenstra et al., 2019; Kim et al., 2019; Levy Zamora et al., 2019).

Based on a multi-season field evaluation of PurpleAir sensors, Magi et al. (2020) reported an RMSE (MAE) of ~7.5 µg m$^{-3}$ (5.8 µg m$^{-3}$) for the uncorrected PM$_{2.5}$. Feenstra et al. (2019) presented the field evaluations of 12 commercially available LCSs under ambient conditions as a part of the AQ-SPEC sensor evaluation program spanning over a 3-year period. Their performance evaluation revealed that 6 of 12 sensors performed with an average $R^2 > 0.70$ and MAE ranging between 4.4 µg m$^{-3}$ and 7.0 µg m$^{-3}$ (for PM$_{2.5}$ concentration range < 50 µg m$^{-3}$).

We also compared the temperature and RH measurements by the LCSs with the BAM ambient meteorological measurements. Compared with the BAM measurements, most of the LCSs overestimated the temperature and underestimated the RH. Temperature (RH) measurements by PAQS (Prana Air) were more accurate than those by other LCSs. The RMSE values of LCS temperature measurements ranged between 3°C and 7°C, whereas those of LCS RH measurements ranged between 7% and 30% (see Figs. S4 and S5). This could be due to the placement of the meteorological sensor with respect to the LCS electronics. In most of the LCSs, the electronics were compactly packed, and the heat emitted by these electronics could impact the temperature and RH measurements.
Table 1. Performance metrics of uncorrected hourly LCS PM2.5.

<table>
<thead>
<tr>
<th>LCS</th>
<th>$R^2$</th>
<th>NRMSE</th>
<th>RMSE ($\mu g m^{-3}$)</th>
<th>MAE ($\mu g m^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerogram</td>
<td>0.86</td>
<td>0.30</td>
<td>9.7</td>
<td>6.9</td>
</tr>
<tr>
<td>Airveda</td>
<td>0.78</td>
<td>0.39</td>
<td>13.5</td>
<td>9.8</td>
</tr>
<tr>
<td>Atmos I</td>
<td>0.86</td>
<td>0.26</td>
<td>8.3</td>
<td>5.8</td>
</tr>
<tr>
<td>Atmos II</td>
<td>0.87</td>
<td>0.39</td>
<td>11.3</td>
<td>9.4</td>
</tr>
<tr>
<td>BlueSky</td>
<td>0.89</td>
<td>0.45</td>
<td>14.6</td>
<td>12.0</td>
</tr>
<tr>
<td>PAQS</td>
<td>0.66</td>
<td>0.89</td>
<td>29.0</td>
<td>24.3</td>
</tr>
<tr>
<td>Prana Air</td>
<td>0.66</td>
<td>0.43</td>
<td>14.5</td>
<td>10.6</td>
</tr>
<tr>
<td>Prkruti</td>
<td>0.63</td>
<td>0.65</td>
<td>22.4</td>
<td>17.4</td>
</tr>
<tr>
<td>PurpleAir_ATM</td>
<td>0.87</td>
<td>0.29</td>
<td>9.4</td>
<td>6.9</td>
</tr>
<tr>
<td>PurpleAir_CF1</td>
<td>0.87</td>
<td>0.75</td>
<td>24.1</td>
<td>17.5</td>
</tr>
</tbody>
</table>

Fig. 2. Season-wise box plots of the $R^2$, NRMSE, RMSE, and MAE of the various LCS PM2.5. The central line of the box represents the median, the box represents the inter-quartile range (IQR), and the lower and upper whisker caps represent the Q1-1.5*IQR and Q3+1.5*IQR, respectively.
3.2 Performance of the Untuned versus Tuned Models

A comparison of the performance metrics of the untuned and tuned ML models in correcting the LCS PM2.5 measurements is depicted in Fig. 3. The metrics are provided in Tables S7 and S8. The untuned and tuned ML models were trained using the ‘all predictors’ training dataset (75% of the total data length). All performance metrics of the corrected PM2.5 were derived based on the testing data (remaining 25% of the total data length). Except for RF, all tuned models performed better in correcting the LCS PM2.5 than the untuned models. For RF, a marginal decrease was observed in the performance metrics, indicating that the scikit-learn default hyperparameters were more efficient than the Grid Search-based hyperparameters in controlling the learning. This could be due to the difference in the range of the hyperparameter space that we used and the scikit-learn default hyperparameters. The highest improvement in model performance due to the hyperparameter tuning was observed in SVR models, followed by DT models and XGBoost models. The $R^2$ of the untuned SVR models ranged between 0.77 and 0.91, whereas that of the tuned models ranged between 0.84 and 0.95. The NRMSE of the corrected PM2.5 using untuned SVR models ranged between 0.19 and 0.30, whereas that using the tuned SVR models was between 0.14 and 0.24. The highest reduction in NRMSE (~40%) of the model-corrected PM2.5 (testing

Fig. 3. Comparison of $R^2$, NRMSE, RMSE, and MAE of the corrected PM2.5 across the untuned and tuned ML models.
data) due to tuning was observed for Prana Air PM$_{2.5}$. The RMSE (MAE) of the corrected PM$_{2.5}$ using tuned SVR models ranged between 4.3 $\mu$g m$^{-3}$ (2.9 $\mu$g m$^{-3}$) and 8.0 $\mu$g m$^{-3}$ (5.0 $\mu$g m$^{-3}$).

### 3.3 Performance of the Models Trained Using ‘All Predictors’ versus ‘Uncorrelated Predictors’ versus ‘Significant Predictors’

Fig. 4 depicts the performance of the tuned ML models trained using ‘all predictors’, ‘uncorrelated predictors’, and ‘significant predictors’. The metrics are given in Tables S9 and S10. The performance metrics provided in the tables were derived based on the testing dataset. The performance of the models trained using ‘all predictors’, ‘uncorrelated predictors’, and ‘significant predictors’ in correcting the LCS PM$_{2.5}$ was comparable. In terms of $R^2$, the degradation in the models’ performance due to its training by ‘significant predictors’ was around 10%, compared with the performance of models trained using ‘all predictors’. In terms of NRMSE, the degradation in the performance of the ML models trained using ‘significant predictors’ was < 25% compared with that of the models trained using ‘all predictors’. The highest increase (from ‘all predictors’-trained models) in the NRMSE values for ‘significant predictors’-trained models was observed for RF-corrected Aerogram PM$_{2.5}$. The performance of ML models trained using ‘uncorrelated predictors’ was intermediate between that of ‘all predictors’- and ‘significant predictors’-trained ML models.

![Fig. 4. Comparison of $R^2$, NRMSE, RMSE, and MAE of the corrected PM$_{2.5}$ across the tuned ‘all predictors’ models, tuned ‘uncorrelated predictors’ models, and tuned ‘significant predictors’ models.](image-url)
Table 2. LCS-wise best performing ML model.

<table>
<thead>
<tr>
<th>LCS</th>
<th>Best performing model</th>
<th>Percentage reduction in NRMSE of corrected PM$_{2.5}$ versus uncorrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerogram</td>
<td>SVR trained with ‘uncorrelated predictors’</td>
<td>37</td>
</tr>
<tr>
<td>Airveda</td>
<td>SVR trained with ‘uncorrelated predictors’</td>
<td>62</td>
</tr>
<tr>
<td>Atmos I</td>
<td>XGBoost trained with ‘all predictors’</td>
<td>46</td>
</tr>
<tr>
<td>Atmos II</td>
<td>SVR trained with ‘all predictors’</td>
<td>62</td>
</tr>
<tr>
<td>BlueSky</td>
<td>SVR trained with ‘uncorrelated predictors’</td>
<td>71</td>
</tr>
<tr>
<td>PAQS</td>
<td>SVR trained with ‘uncorrelated predictors’</td>
<td>74</td>
</tr>
<tr>
<td>Prana Air</td>
<td>SVR trained with ‘all predictors’</td>
<td>67</td>
</tr>
<tr>
<td>Prkruti</td>
<td>SVR trained with ‘all predictors’</td>
<td>74</td>
</tr>
<tr>
<td>PurpleAir_ATM</td>
<td>SVR trained with ‘significant predictors’</td>
<td>48</td>
</tr>
<tr>
<td>PurpleAir_CF1</td>
<td>SVR trained with ‘significant predictors’</td>
<td>81</td>
</tr>
</tbody>
</table>

For each of the LCSs, the best performing models are listed in Table 2. The best performing model was chosen based on NRMSE values of the corrected LCS PM$_{2.5}$. If two models were characterised with the same NRMSE values, $R^2$ was chosen as the criteria. Out of ten LCSs, SVR emerged as the best performing model for nine and XGBoost as the best performing for one. Of note, the differences in the performances of the XGBoost and SVR models were marginal. The best performing models were ‘all predictors’-trained models for four LCSs, ‘uncorrelated predictors’-trained models for three LCSs, and ‘significant predictors’-trained models for the other three LCSs. The scatter plots between the predicted hourly PM$_{2.5}$ by the LCS-wise best performing ML models and the corresponding hourly BAM PM$_{2.5}$ are shown in Fig. 5. As shown in Fig. S6, the quantile-quantile plots revealed that the residuals were normally distributed. The NRMSE of the best-performing models-corrected LCS PM$_{2.5}$ was improved by 37%–81% compared with that of the uncorrected PM$_{2.5}$. The highest improvement was observed for PurpleAir_CF1 and the least for Aerogram. PurpleAir_CF1 PM$_{2.5}$ also showed the highest improvement in terms of RMSE (~77%, ~19 µg m$^{-3}$ reduction in RMSE).

3.4 Performance of Linear Models versus ML Models

To investigate the improvement in the performance of the ML models over statistical models, we trained multi-linear regression (MLR) models using ‘all predictors’ for all LCSs and compared their performances against the best-performing ML models (see Fig. 6 and Table S11). The MLR...
models also improved the performance of LCS PM$_{2.5}$. However, except for Aerogram, we observed that ML models performed better in correcting the LCS PM$_{2.5}$ than the statistical models. For example, when PurpleAir_CF1 PM$_{2.5}$ was corrected using the MLR model, its RMSE improved by 74%, whereas it improved by around 77% when corrected using the tuned SVR model trained using ‘significant predictors’. For Aerogram, the performance metrics of the MLR model and SVR model were almost similar. Compared with MLR, the highest improvement in the ML model performance was observed for Prana Air (NRMSE improved by ~42%), followed by Airveda and Prkruti. ML models can capture more complex nonlinear effects that simple statistical models cannot. Earlier studies (Liu et al., 2019; Considine et al., 2021; Liang, 2021; Gupta et al., 2022; deSouza et al., 2022) have also demonstrated that ML models could perform better than statistical models in correcting the LCS PM$_{2.5}$.

### 3.5 Case Study

As a case study, we included the collocated BC data as an additional predictor to the best performing ML models and investigated if there was any improvement in the model performance. As all LCSs quantify PM based on light scattering, the inclusion of information on the light absorbing PM (BC) can impact the model performance. A marginal improvement was observed in BC-added
4 CONCLUSIONS

We collocated nine different PM LCSs with a BAM in Bengaluru and observed a range of bias in the LCS-measured PM$_{2.5}$. ML models performed considerably better in improving the LCS PM$_{2.5}$ accuracies than statistical models. We also observed that the performance of ML models (in correcting the LCS PM$_{2.5}$) trained using RFE-shortlisted predictors was comparable to that of the ‘all predictors’ models. In case of RF models, the scikit-learn default hyperparameters performed better than the tuned hyperparameters. It is recommended to explore the default hyperparameters and conduct a thorough exploratory data analysis to eliminate insignificant predictors from the list. In this study, the effects of hyperparameter tuning and the choice of predictors on different ML models (Fig. 7). For example, the RMSE of corrected LCS PM$_{2.5}$ using best-forming ML models ranged between 4.2 $\mu$g m$^{-3}$ and 7.7 $\mu$g m$^{-3}$, while it was between 3.5 $\mu$g m$^{-3}$ and 7.5 $\mu$g m$^{-3}$ for the models in which BC was also included as an additional predictor (see Table S12 for more details). With the addition of BC, the highest improvement in the model performance was observed for PAQS (22% in NRMSE), followed by BlueSky, and Prana Air. No improvement was observed for Aerogram and PurpleAir.

Fig. 7. Comparison of $R^2$, NRMSE, RMSE, and MAE of the corrected PM$_{2.5}$ across the LCS-wise best performing models and BC-included best performing ML models.
LCS PM$_{2.5}$ were different. Further, the model performance improved when variables related to the periodicities in the continuous variables were included. We created new variables related to the hour of the day, weekday/weekend, month, and season. The inclusion of collocated optical absorption-based BC mass concentration measurements in the ML models did not significantly improve the ML model’s performance in correcting the LCS PM$_{2.5}$.

The study has few limitations. Our study is limited to one geography; the bias in the uncorrected LCS PM$_{2.5}$ and the performance of calibration models might vary for other geographies. The amount of data available from each of the sensors varied due to the intermittent malfunctioning/IoT issues of a few LCSs. It should be noted that the version of PAQS LCS used in the study was intended for indoor air pollution measurements. For the post-monsoon season, only December data were available. As the LCS technology is continuously evolving, a few of the LCSs used in the study were upgraded to new versions in terms of the internal laser PM sensors.

**ACKNOWLEDGEMENTS**

All authors are thankful to the MacArthur Foundation and Google for providing funding support to conduct air pollution studies at the Center for Study of Science, Technology and Policy (CSTEP).

**SUPPLEMENTARY MATERIAL**

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.220428

**REFERENCES**


