Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220076

What Influences Low-cost Sensor Data Calibration? - A Systematic Assessment of Algorithms, Duration, and Predictor Selection

Lu Liang[1*], Jacob Daniels[2]

[1] Department of Geography and the Environment, University of North Texas, Denton, TX 76203, USA
[2] Department of Electrical Engineering, University of North Texas, Denton, TX 76203, USA

ABSTRACT

The low-cost sensor has changed the air quality monitoring paradigm with the capacity for efficient network expansion and community engagement. The surge in its use has sparked new research interests in understanding its data quality. Many studies have employed field calibration to improve sensor agreement with co-located reference monitors. Yet, studies that systematically examine the performance of different calibration techniques are limited in scope and depth. This study comprehensively assessed ten widely used data techniques, namely AdaBoost, Bayesian ridge, gradient tree boosting, K-nearest neighbors, Lasso, multivariable linear regression, neural network, random forest, ridge regression, and support vector machine. We compared their performance using a standardized baseline dataset and their responses to various parameter combinations. We further assessed the training sample size effect to understand the optimal duration of field calibration for achieving good accuracy. Finally, we tested different predictor combinations to address whether the inclusion of more predictors will lead to better performance. Using baseline data, the neural network achieved the best performance, followed by the four regression-based methods, showing very consistent and stable performance. While confirming that the latest research tendency is deep learning, regression is still a viable option for studies with limited effort in parameter tuning and method selection, especially considering its computational efficiency and simplicity. The sample size effect is most evident when the sample size drops below 30%, which is equivalent to six weeks of continuously collected hourly data. Although algorithms react differently to the number of predictors, their performance was typically boosted by adding more predictors, especially the particle count and humidity. Our study not only describes an approach of sophisticated data-driven calibration for practical applications, but also provides insights into the compounding impacts of parameters, samples, and predictors in algorithm performance.

Keywords: PurpleAir, Machine learning, Particulate matter, $PM_{2.5}$, Air quality

OPEN ACCESS

Received: February 14, 2022
Revised: June 6, 2022
Accepted: June 23, 2022

* **Corresponding Author:**
lu.liang@unt.edu

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

1 INTRODUCTION

Air pollution is one of the global leading mortality risk factors (Apte *et al.*, 2017; Liang and Gong, 2020). Even at low concentrations, fine particulate matter with aerodynamic diameters smaller than 2.5 μm ($PM_{2.5}$) is significantly associated with an increased health hazard (Bell *et al.*, 2011) and adverse social-environmental effects (Sager, 2019). Increasing evidence proves that socio-economically disadvantaged communities suffer more from higher levels of air pollution (Colmer *et al.*, 2020; Gray *et al.*, 2013; Peled, 2011). There is a critical need to characterize the spatial-temporal patterns of $PM_{2.5}$ at the granular level to better estimate and mitigate those risks at the individual or community level.

A paradigm shift in granular-level air monitoring is the growing usage of low-cost sensors (LCSs) to supplement conventional sparsely located regulatory stations (Mao *et al.*, 2019; Snyder *et al.*, 2013).

Because of their affordability (Jiao *et al.*, 2016), LCSs can detect fine-scale spatial-temporal $PM_{2.5}$ variability (Giordano *et al.*, 2021; Hart *et al.*, 2020). Some widely adopted sensors include the Plantower PMS series (module used in a variety of integrated sensor systems, such as PurpleAir), Alphasense OPC-N2, Panasonic $PM_{2.5}$ sensors, NovaFitness SDS, Shinyei PPD series, Samyoung DSM series, and Sharp GP and DP series. Those sensors have been used under various scenarios, such as environmental regulation (Bi *et al.*, 2020), hotspot detection (Mousavi *et al.*, 2021), traffic-related studies (Amegah *et al.*, 2022; McFarlane *et al.*, 2021), and health assessment (Liang *et al.*, 2019; Tsou *et al.*, 2021). This change is significant in transforming environmental governance, especially in historically underserved countries or regions, where access to the air quality data is insufficient to act on pollution trends (Pope *et al.*, 2018). Additionally, their user-friendly operation and crowd-sourcing capacity can well support network expansion and community engagement (Morawska *et al.*, 2018). The easy integration of LCS data with multi-source information (such as remotely sensed data) further extends its utilization in multi-disciplinary studies at various scales (e.g., Gupta *et al.*, 2021).

Despite the increasing popularity of LCSs in the scientific, industrial, and civilian domains (Liang, 2021), a rising concern is that their out-of-the-box data quality is generally lower than in the laboratory. Since most low-cost $PM_{2.5}$ sensors are light scattering based (Morawska *et al.*, 2018), they show larger uncertainty than reference instruments with degrading performance (Masson *et al.*, 2015). Under natural conditions, LCSs show non-linear responsiveness to their interfering environments, such as meteorology (Feinberg *et al.*, 2019) and background target and non-target gas interference (Castell *et al.*, 2017). Thus, direct usage of LCS data without proper calibration can lead to undesirable outcomes (Rai *et al.*, 2017).

LCSs typically require field calibration before wide-scale deployment (Austin *et al.*, 2015). A common approach is to collocate the sensors with a regulatory instrument in places where they will be deployed (Giordano *et al.*, 2021) and use data-driven or empirical methods to adjust the drift of LCS data to reference data (Liang, 2021). Physical mechanism-based models, such as κ-Köhler theory or scattering efficiency derived relative humidity correction factor, correct the biased conversion from light-scattering to particle mass concentration due to humidity (Crilley *et al.*, 2020; Zheng *et al.*, 2018). Regression is one of the earliest data-driven methods used because of its simplicity and a high degree of method scalability (Liang, 2021). Recently, the advanced machine learning (ML) methods, such as neural networks, are leading the trend because of their problem-specific and robust performance (Giordano *et al.*, 2021; Johnson *et al.*, 2018; Mahajan and Kumar, 2020; Morawska *et al.*, 2018; Zimmerman *et al.*, 2018). However, little work has been done to understand essential questions during the procedures, such as how broadly applicable those methods are, how long the sensors should be collocated to provide enough calibration datasets, and what variables should be accounted for to achieve sufficient accuracy.

Given the implications of LCSs in environmental monitoring and the importance of LCS calibration in their applications, this study aims to systematically compare different data-driven methods by quantitatively analyzing the coupling effects of algorithms, sample sizes, and explanatory variables on the calibration performance. In particular, we are driven by three research objectives: 1) synthesize and compare the various mainstream data-driven algorithms in LCS field calibration. 2) examine the influencing variables during calibration. 3) explore the sample size effects on the calibration model performance.

2 METHODS

2.1 Low-cost Sensors and Calibration Setup

2.1.1 PurpleAir sensors

We chose the PurpleAir (PA) sensors given their good performance and wide deployment among a broad spectrum of groups (Barkjohn *et al.*, 2021). At a size of 22,530 registered sensors (as of April 22, 2022), this massive global network requires immediate attention to the data quality. Many field calibrations of PA sensors have been conducted on many continents, including Africa (McFarlane *et al.*, 2021), Asia (Kim *et al.*, 2019), Australia (Robinson, 2020), Europe (Stavroulas *et al.*, 2020), and North America where most deployments were carried out (Ardon-Dryer *et al.*, 2020; Bi *et al.*, 2020; Feenstra *et al.*, 2019; Magi *et al.*, 2020; Malings *et al.*, 2019; Ouimette *et al.*, 2022; Tryner *et al.*, 2020).

PA sensors are equipped with two laser scattering particle counters (Plantower PMS5003) that report independently at approximately a 120 s interval. The Plantower sensors use a fan to draw air through an inlet past the laser, producing a scattering effect that is detected by the photodiode. A proprietary algorithm developed by Plantower was applied to convert the amount of light scatter detected into particle sizes, and then from particle count ($\mu m\ dl^{-1}$) into mass concentration ($\mu g\ m^{-3}$). Because the indoor and outdoor conversion options are different, only data calculated using the outdoor conversion method was used. The mass concentration for PM_1, $PM_{2.5}$, and PM_{10} are reported, all of which are average for the two channels. If the outdoor particle values reported for the two channels drift apart, the PurpleAir system will downgrade one of the channels and exclude the channel from the data average. Raw particle count is also reported in six size bins ranging from 300 nm to 10 μm, separately particle sizes greater than 0.3 μm diameter, 0.5 μm, 1.0 μm, 2.5 μm, 5.0 μm, and 10 μm. PA sensors also use a Bosch BME280 sensor to estimate relative humidity (RH), temperature, dew point, and pressure. The data transmission and storage are enabled by its Wi-Fi module for real-time data transmission and a built-in SD card as a backup solution to internet disconnection.

2.1.2 Reference instrument and calibration system

Reference instruments typically refer to federal reference methods (FRMs) and federal equivalent methods (FEMs) that provide National Ambient Air Quality Standards (NAAQS) in the U.S. (U.S. EPA, 2011), or similar sampling technologies in other countries (Cao *et al.*, 2013). FRMs and FEMs commonly use more sophisticated and regularly maintained technologies for particle mass measurement such as direct gravimetric methods, beta attenuation, and oscillating microbalance methods (Schmidt-Ott and Ristovski, 2003). Despite their gold standard role in air quality monitoring, the implementation and operational costs are high. For instance, it costs approximately $50 million to maintain U.S. national ambient air quality monitoring system per year (U.S. GAO, 2020). Besides, the site selection is primarily based on population density, with less consideration of other factors such as social inequality (Watson *et al.*, 1997).

2.2 Data Collection and Cleaning

Here, we employed a US-wide PurpleAir correction dataset from a previous EPA work to make the results generic enough to avoid any location-specific biases (Barkjohn *et al.*, 2021). Part of the collocation data was obtained from sensor calibration experiments that were operated by air monitoring agencies. Another portion of the data came from privately owned sensors that are within 30 m of an active EPA Air Quality System site reporting $PM_{2.5}$ and have been confirmed by a local air monitoring agency for their identities. A thorough data cleaning was performed to ensure data quality following these steps (Fig. 1): 1) One Iowa dataset that constituted 55% of the entire collocated dataset was thinned from 10,907 to 3,762 data points to better balance the datasets among the states and to avoid building a final model that is Iowa dependent. All high-concentration data ($\geq$ 25 $\mu g\ m^{-3}$) were retained and low concentration data were randomly drawn; 2) A 90% completion threshold was applied to data to enable a true representation of daily averages; 3) Extremely high and low values in $PM_{2.5}$, temperature ($>$ 540°C), and RH ($>$ 100%) collected by PA were removed; and 4) Each PA units has two identical Plantower sensors (refer to as channels hereafter), and the agreement between the data collected from both channels can indicate potential data outliers. We first calculated the absolute and percentage differences between two PA channels using their 24-hour average. Percentage is the absolute difference divided by the average of the two channel readings. The percentage difference was used to deal with channel disagreement under a high concentration scenario that can not be captured by absolute difference. Records with an absolute difference of 5 $\mu g\ m^{-3}$ or fall outside of two standard deviations of the entire percent difference dataset were removed.

Because no Texas site was included in the national dataset, we supplemented it with the field calibration data that we collected at the Texas Commission on Environmental Quality (TCEQ) Denton Airport South station (EPA site number: 481210034, Lat: 33.2190759, Long: –91.19962841). From April 12, 2020 to September 17, 2020, four PA sensors were placed at a close distance ($<$ 5 m) to a FEM regulatory instrument (BAM). To reduce data redundancy, we picked only one sensor with $R^2 > 0.9$ between the two channels and with the highest agreement with other units during

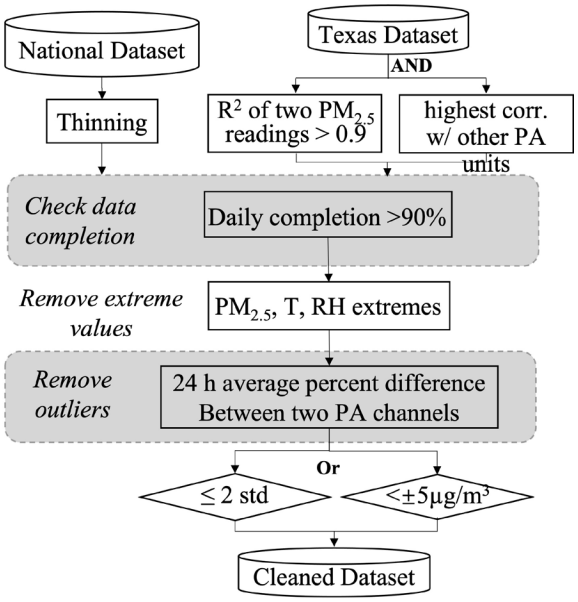

Fig. 1. Flow of data cleaning.

the same deployment period. After that, the same data download and cleaning procedure were applied.

The final dataset contains 50 PA sensors that were located in 16 states across 39 sites (Fig. 2), with a total of 12,705 records. California and Iowa have 19 sensors and account for almost 60% of the total number of data records. The longest data collection period was 833 days and the shortest one was only two days. Thirty-eight sites contribute over 100 records, which is equivalent to approximately three-month period of data collection (Fig. S1). Overall, the PA sensors are in good agreement with the reference data, with the mean R^2 as 0.88 using linear regression, but tend to overestimate the ambient $PM_{2.5}$ level (Fig. 2). The mean R^2 between the PA and reference data for all sites is 0.88, with the highest agreement as 0.996 and the lowest as 0.468. Detailed site information and data summary can be found in the supplementary file. One example of the time-series comparison between the PA and reference data collected in the Texas site is displayed in Fig. S2.

2.3 Data Experiments

2.3.1 Testing the effects of different algorithms and parameter combinations

We tested ten widely applied and openly accessible machine learning algorithms that can be roughly divided into four groups: regression-based, distance-based, network-based, and ensemble (Table 1).

Regression-based algorithms. As one of the earliest methods being tested, multivariate linear regression (MLR) takes the linear form of one response variable and a set of explanatory variables. In LCS calibration studies, the readouts of the reference instrument are the response variable and the LCS data is the main explanatory variable. Other influencing factors, including environmental or mechanical ones (e.g., temperature, *RH*, sensor age), have also been widely used under the assumption that all factors respond linearly to the reference data. Ordinary least squares is often used by default in MLR to estimate the coefficients by minimizing the sum of the squared residuals. The final selected US-wide correction model for PA sensor adopted the MLR form Barkjohn *et al.* (2021):

$$PM_{2.5} = 0.524 \times PA_{cf} - 0.0862 \times RH + 5.75 \quad (1)$$

Ridge, Bayesian ridge, and Lasso are all extensions of MLR, with additional regularization parameter that aims to minimize complexity. Ridge regression uses a tunable additive L2 norm penalty term—the sum of squares of coefficients—in the optimization. Alpha is the parameter that

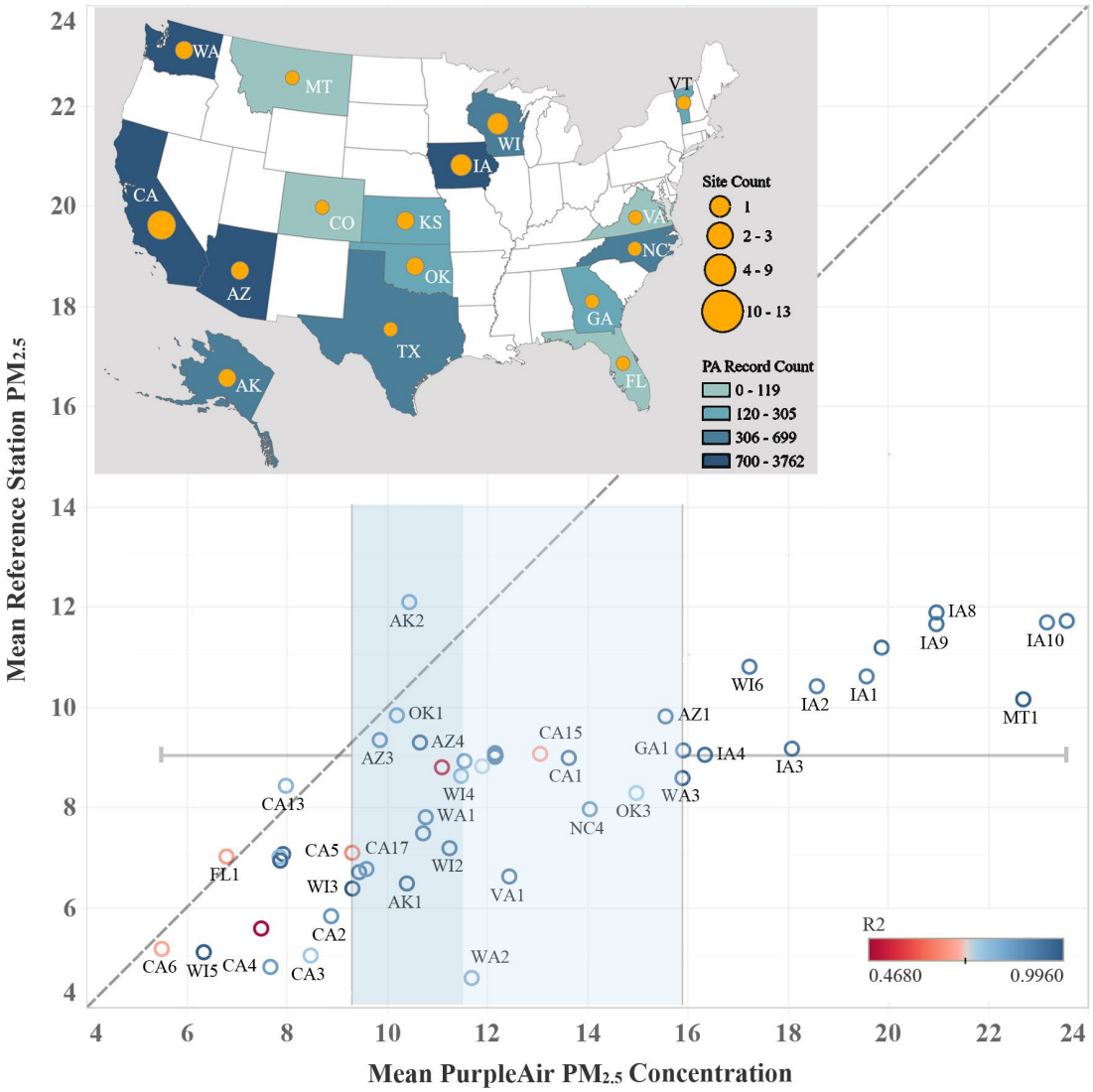

Fig. 2. The map shows the collocation sites. The color symbol indicates the total number of PA data records and the size symbol indicates the total number of collocation sites for each state. The scatter plot shows the relationship between the mean $PM_{2.5}$ concentration reported by the PA sensors and their corresponding reference station during the calibration period. The marks are labeled by site. Color represents the R^2 between those two $PM_{2.5}$ values. The box plot displays the five-number summary of the mean PA data.

balances the minimization of the residual sum of squares and the magnitude of coefficients. The model complexity tends to reduce as the alpha value increases. An optimal alpha provides a trade-off between significant overfitting at low alpha values and underfitting at high alpha values. Bayesian ridge regression uses regularization in probabilistic terms. The model estimation is conducted by iteratively maximizing the marginal log-likelihood of the observations (Pedregosa *et al.*, 2011). Lasso performs L1 regularization by adding a factor of the sum of absolute value of coefficients in the optimization process. The alpha works similar to that of ridge regression.

Support vector machine (SVM) regression finds the best fit line as the hyperplane that has a maximum number of points. SVM uses kernel functions, including linear, polynomial, and gaussian radial basis kernel function, to convert low dimensional data space into a better dimensional space, so data points can be better separated.

Distance-based algorithm. K-nearest neighbors (KNN) is a distance-based method that uses the mean of all the nearest neighbors' values to predict the value of new data. K indicates the count of the nearest neighbors. The weights of neighbors could be assigned in two ways: uniform treats all neighbors equally, whereas distance-based weighting assigns higher weights to the closer neighbors.

Table 1. Algorithms, parameter settings, and paper citation. The default values for each parameter used in their corresponding package were underlined.

Algorithms	Abbrev.	Parameters	Parameter settings	Citation
AdaBoost	AB	Number of Estimators	25, 50, 75, 100, 150, 200, 300	Freund and Schapire, 1997
Bayesian Ridge	BR	-	-	MacKay, 1992
Gradient Tree Boosting	GTB	Number of Boosting Stages	25, 50, 75, 100, 150, 200, 300	Friedman, 2001
		Fraction of Samples	0.7, 0.9, 1.0	
K-Nearest Neighbors	KNN	Number of Neighbors	1, 2, 3, 4, 5, 7, 10	Fix and Hodges, 1989
		Weight Function	Uniform, Distance-based	
Least absolute shrinkage and selection operator	Lasso	Alpha	0.5, 0.75, 1.0, 1.5, 2.0	Tibshirani, 1996
Multivariable Linear Regression	MLR	-	-	Mardia *et al.*, 1979
Neural Network	NN	Number of Layers	1, 2, 3	Hopfield, 1982
		Neurons per Layer	50, 100, 256, 512	
Random Forest	RF	Number of Trees	20, 40, 60, 80, 100, 120, 140, 160, 180, 200	Breiman, 2001
Ridge Regression	RR	Alpha	0.5, 0.75, 1.0, 1.5, 2.0	Hilt and Seegrist, 1977
Support Vector Machine	SVM	Kernel Type	Linear, Polynomial, Radial Basis Function	Cortes and Vapnik, 1995

Note: algorithms are listed in alphabetical order. Abbrev.: abbreviation.

Network-based algorithms. Neural network (NN) is relatively new but attractive to users because of its superior performance (Okafor *et al.*, 2020; Yamamoto *et al.*, 2017). One previous study has reported a 10% increase in R^2 from MLR to NN, with the improvement attributable to its ability in capturing the data variation (Mahajan and Kumar, 2020). A NN is an architectural structure consisting of highly interconnected processing units (neurons) that are organized in layers. The weight of neurons is tuned and optimized through the supervised learning process.

Ensemble methods. Ensemble techniques are typically built upon many weaker classifiers to create a strong classifier. AdaBoost is the first generation of boosting algorithms and another successful example is the random forests that build decision trees independently and combine results at the end. Both methods have a main parameter—the number of estimators or trees—controlling the structure. Generally, a larger quantity of estimators can lead to better performance but longer training time. Additionally, the accuracy will plateau after a certain number of estimators. Gradient boosting differs by building one tree at a time and combining results along the way in a forward stage-wise fashion. A larger number of boosting stages usually results in better performance. The fraction of samples is fitting the individual base learners. A fraction less than one may lead to a reduction of variance and an increase in bias.

Since there is no golden standard for choosing the optimal parameter, we tested a range of parameters that are recommended by the algorithm documentation or close to the default values picked by the sourcing code Scikit Learn (Pedregosa *et al.*, 2011). Python 3.9.7 was used to implement those algorithms. All codes are available at: https://github.com/unt-geo/Calibration

2.3.2 Sample size effect

A critical question in data-driven techniques is to determine how much training data is needed to achieve a specific performance goal. In the context of the LCS field calibration, we aim to answer two questions: 1) As training data grows, will performance continue to improve? 2) Does the sample size effect vary by algorithms?

The proper test of the sample size effect requires a geographically and size balanced dataset. Otherwise, the assessment may be misleading. To reduce the bias, we first adjusted the whole daily average dataset by selecting all 38 sites with more than 100 days of data, and further randomly selecting 100 data points from each site. The final dataset with 3,800 records was used to conduct the sample size experiment, which was randomly split into 90% for the training set

and 10% for the test set. The training data was used to fit the model and the test data was to provide an unbiased evaluation of the model fit on the training dataset. We further prepared various training datasets at different sample sizes. Specifically, we constructed 10 sets of training samples with 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of the entire dataset in the order of data collection time.

2.3.3 Predictor selection

Many previous works are focused on using a single variable—PA $PM_{2.5}$ concentration—to run the calibration model. However, because the responsive rate of LCSs is controlled by a range of internal and ambient environmental factors, it can be beneficial to include additional influencing variables in the modeling process as evidenced in the previous literature (Gao *et al.*, 2015). Similar to the sample size effect, the key questions lie in whether more predictors will lead to better performance, and whether a plateau effect exists in the variable selection for certain algorithms.

In this experiment, we picked seven variables that are commonly used in calibration studies, separately $PM_{2.5}$ concentration ($PM_{2.5}$ conc), $PM_{2.5}$ count ($C_{2.5}$), PM_1 count (C_1), $PM_{5.0}$ count (C_5), PM_{10} count (C_{10}), humidity (RH), and temperature (T). $PM_{2.5}$conc is the mass concentration generated by the proprietary algorithm developed by the laser counter manufacturer Plantower, which incorporates assumptions about potentially varying density and shape of the particles. However, because the information on the assumptions is unrevealed, it is unlikely that the assumed particle properties would be similar to those observed in the fields. With this consideration, we included the other type of PA output values—the particle counts in different sizes, which are the raw reporting of airborne particle numbers. Some studies have found that particle counts explain well in the calibration model (Zusman *et al.*, 2020).

We first tested the effects of each single predictor on explaining the variance of reference data using univariate linear regression. We then tested the combined effects of multiple predictors. Datasets 2–5 incorporated the RH and T to account for the known sensitivity of sensors to fluctuations in meteorological conditions (Castell *et al.*, 2017). RH influences the LCS readings by changing the particle size and the refractive index when water condenses onto particles (Di Antonio *et al.*, 2018; Molnár *et al.*, 2020). The water moistening effect also partially explains the typical overestimation of LCSs, which is especially evident when RH exceeds 75%. Temperature interferes with the nature of the aerosol samples and impacts the sensor performance, especially in the ambient environment (Olivares and Edwards, 2015). However, how the sensors respond to the temperature is less studied and still unexplained.

The seven variables were combined into five datasets (Table 2). For example, Dataset 3 included three variables while Dataset 4 used seven variables. Dataset 0 that uses $PM_{2.5}$ concentration as the single explanatory variable was used as the baseline for comparison. Other variables were gradually included according to their importance values obtained from the single variable test (Tables S3 and S4).

2.3.4 Accuracy metrics

We used the coefficient of determination (R^2) for quantifying the portion of the variation in the dependent variable that can be predicted from the model and the independent variables. Root mean squared error (RMSE) was used as indices of the respective average absolute error. In this paper, we reported how those algorithms respond to adjustments in training data and parameters. Accuracy values were used as an indicator for the degree of response. However, we

Table 2. Variables as predictors and the dataset constitution.

Dataset	Variables Used	Acronomy
0	$PM_{2.5}$ concentration	$PM_{2.5}$
1	$PM_{2.5}$ concentration, $PM_{2.5}$ count	$PM_{2.5} + C_{2.5}$
2	$PM_{2.5}$ concentration, $PM_{2.5}$ count, humidity	$PM_{2.5} + C_{2.5} + RH$
3	$PM_{2.5}$ concentration, $PM_{2.5}$ count, humidity, temperature	$PM_{2.5} + C_{2.5} + RH + T$
4	$PM_{2.5}$ concentration, $PM_{2.5}$ count, humidity, temperature, PM_1 count, PM_5 count, PM_{10} count	$PM_{2.5} + C_{2.5} + RH + T + C_1 + C_5 + C_{10}$

intend to avoid listing those specific accuracy values, as different studies and datasets may reach varying results. Interested readers can find those values in figures, tables, and supplementary files.

3 RESULTS AND DISCUSSION

3.1 Effects of Algorithm and Parameter Settings

We compared the effects of different algorithms on sensor data calibration by using the baseline Dataset 0 and tested with the default and most optimal parameter setting (Fig. 3, Table 3). Between the two major categories, the regression-based methods achieve overall high accuracies, except for Lasso. The ensemble methods show the largest discrepancies in their performance, with GTB proving to be the best and AB the worst. NN slightly outperforms some models, although at the higher computational cost.

On average, the KNN models tend to perform best using a uniform weight function with a lower number of input features and a distance-based weight function with a higher number of variables (Fig. S3). For NN, neuron count and layer count seem to have similar levels of impact on the performance; both increase the model's ability to create a representation of the input, but more neurons increase the amount of information gained while the number of layers increases attention to increasingly fine details (Fig. 4).

In the ensemble methods, AdaBoost performs best on average when using a smaller number of classifiers (Fig. S4). The RF is very sensitive to the number of trees when only a few trees (seven) are used. The performance largely stagnates with an increasing number of trees (Fig. S5). This is because the trends in the data can be largely accounted for using only 7 or more trees. Most outliers are eliminated and overfitting to a particular input is diminished, so increasing the number of

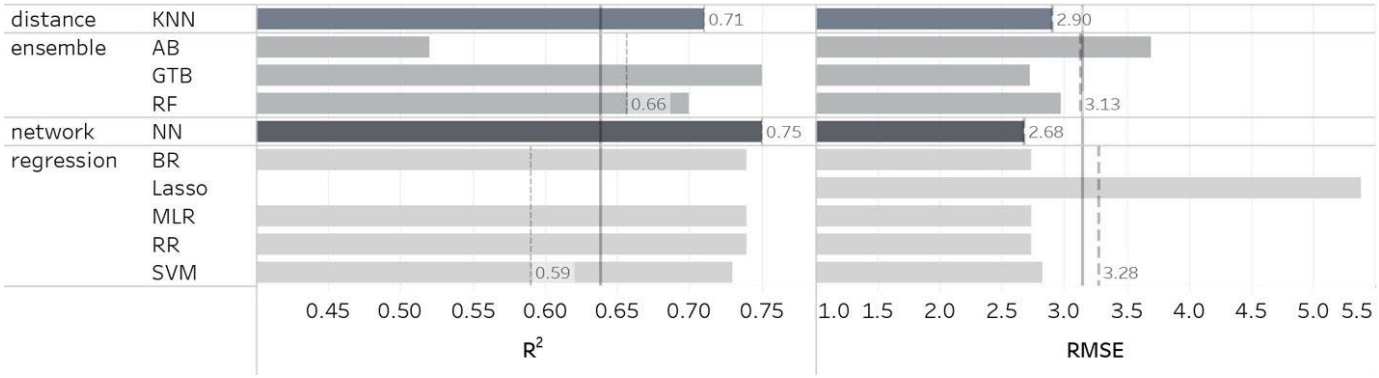

Fig. 3. The accuracy metrics of algorithms when tested on baseline Dataset 0 and using the default parameter setting.

Table 3. Optimal parameter values and performance by algorithms.

Algorithms	Parameters	Parameter Values	R^2	RMSE
AB	Number of Estimators	25	0.64	3.22
BR	-	-	0.81	2.33
GTB	Number of Boosting Stages	300	0.85	2.05
	Fraction of Samples	0.7		
KNN	Number of Neighbors	5	0.86	2.02
	Weight Function	Distance		
Lasso	Alpha	0.5	0.50	3.83
MLR	-	-	0.81	2.33
NN	Number of Layers	2	**0.89**	**1.79**
	Neurons per Layer	512		
RF	Number of Trees	200	0.86	1.99
RR	Alpha	Tie	0.81	2.33
SVM	Kernel Type	RBF	0.82	2.26

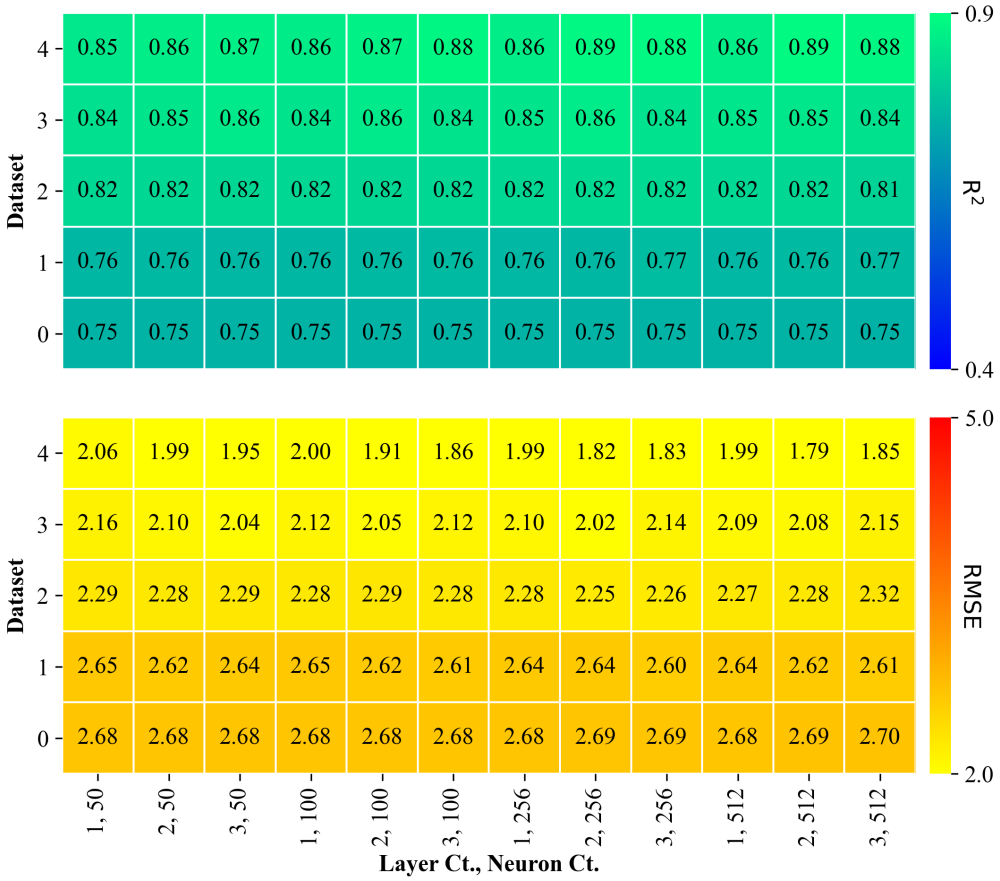

Fig. 4. The performance of Neural Network with varying datasets and model parameters.

trees has little effect. Gradient Tree Boosting shows improving performance with the number of trees increasing until past 50. However, the sample fraction shows mild performance changes (Fig. S6).

Ridge regression, a modification of the linear regression model, performs nearly identically to MLR (Figs. S7 and S8). It appears that introducing a small amount of bias to the linear regression model does not significantly change the performance. Lasso regression performs the worst among all models. As the value of alpha increases, the model performs even worse. As the alpha value gets lower, meaning the lasso regression is approaching regular linear regression, the model better fits to the data (Fig. S9). Bayesian ridge regression performs similarly to regular ridge regression, with a very slight increase in R^2 for the final dataset (Fig. S10). For SVM, the kernel plays a big role in determining the model performance. Default RBF outperforms the linear and polynomial kernels. The linear kernel increases in performance relatively slower compared to the other two kernels. RBF shows good performance whereas the polynomial kernel performs poorly, indicating that it is not a good fit for this dataset (Fig. S11). The underperformance of linear kernel is likely because the dataset is not linearly separable due to the nature of $PM_{2.5}$. Similarly, the relatively simple 3rd-degree polynomial kernel used in this study does not fit well, especially to the datasets with fewer variables as these are likely more linearly separable, as shown by the similar performance of RBF and linear kernels with the less-variable datasets.

3.2 Sample Size Effects

Most algorithms show positive responses to increased training sample sizes, except for Lasso (Fig. 5). The algorithm most affected by the training sample size is AB, of which the R^2 raised to 100% from using one-tenth to 80% of the whole data. SVM, Lasso, and NN are the least affected. With a very small dataset (i.e., two weeks of hourly data, about 340 data points in this study), SVM, NN, and Lasso can produce relatively good results. When the dataset is rich (i.e., half a

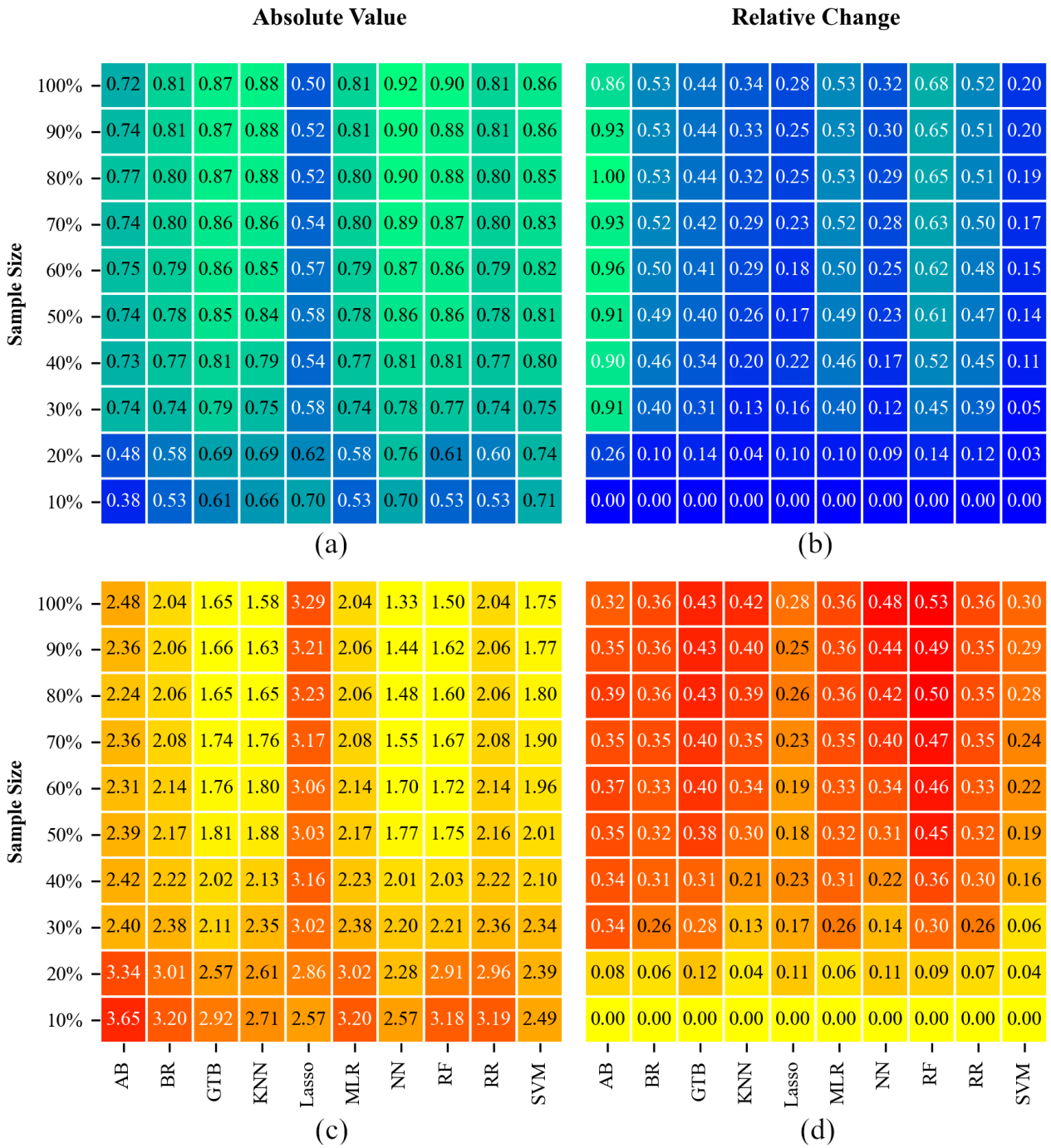

Fig. 5. R^2 (top) and RMSE (bottom) with relative change compared to 10% of data (right). (a) is R^2, (b) is the relative change in R^2, (c) is RMSE, and (d) is the relative change in RMSE.

year's hourly data), nine out of eleven algorithms reach the R^2 higher than 0.8, with NN and RF especially high (over 0.9). Generally speaking, the sample size effect is most evident when the sample size drops below 30%.

Calibration duration has been recognized as a non-neglectable factor in calibrating LCSs. The sample size effect can also provide insights into the optimal time length to co-locate LCS sensors with a reference instrument. Using our compiled national dataset, there is a consensus among various algorithms that the accuracy improves the most when the sample size increases to approximately 1000 records, which is equivalent to six weeks of continuously collected hourly data. Passing this threshold, the accuracy improves more slowly or remains stable.

3.3 Effects of Predictors

Figs. 4 and S3–S11 displayed the results of the predictor selection. As a comparison, we applied the US-wide correction model (Formula (1)) to our dataset, which obtained R^2 as 0.76 and RMSE as 2.63. For KNN, the input variables play a more significant role in the performance than the model parameters as would be expected, with sharp performance increases between the first, second, and third datasets and a moderate performance boost between the third and final datasets (Fig. S2). As a non-parametric method, more variables create a higher dimensional space for the distance calculation, which typically leads to more refined predictions. As the dimension gets higher, the advantages of multivariate distance calculation become weaker. In NN, the effect of the number of layers and neurons depends on the sample size. Higher numbers of both layers and neurons improve the model's performance with higher values beginning to stagnate in a performance increase. This contrasts with the results at low levels of data, where the lower values perform better. This is likely because, at lower sample sizes, the larger neural networks are more likely to overfit the small amount of training data since there is not enough data to get a good generalization with that number of details.

AdaBoost performs best on average when using a smaller number of estimators. This performance trend is especially apparent for Datasets 3 and 4 where performance drastically decreases as the number of estimators increases (Fig. S3). Gradient Tree Boosting shows significant increases in performance with the larger datasets than with the smaller ones.

The MLR model performance increases slowly after the second dataset is introduced. For both Ridge and Bayesian ridge regression, the dataset used does not significantly increase performance, except for going from the first dataset to the second. For Lasso, the dataset used makes almost no difference in the poor performance. For SVM, the inclusion of more predictors can lead to about a 10% increase in R^2 from Dataset 0 to Dataset 4, regardless of the kernel type used.

In general, the inclusion of humidity and $PM_{2.5}$ count can improve model performance, as these two factors demonstrated the heaviest weights of coefficients in the four regression models (Table S3). $PM_{2.5}$ also obtained the highest importance score in the three tree-based models (Table S4). Temperature and particle count at other sizes only slightly influence the outcome. Although when the single variable was evaluated against the reference data, temperature shows a slightly better correlation than RH (Table S2). This can be attributed to two reasons. First, ambient temperature has not been proved to significantly influence the physicochemical property of PM particles. Second, all particle counts are strongly correlated. Including highly correlated variables can introduce multicollinearity and data redundancy issues to the model. Particle counts at the different sizes all show high correlation (Table S1) and they can be good proxies for PA concentration data when the count to concentration conversion formula is not publicly available.

We need to note that some predictors that may be important are not included in the analysis due to data limitation, such as sensor age. Dust sensors lose sensitivity and the accuracy drifts over time (De Vito *et al.*, 2020; Jiao *et al.*, 2016), which becomes another potential source of measurement artifact (Hasenfratz *et al.*, 2012). PurpleAir sensor has a shorter shelf life than high-end reference instruments and the accuracy is found to degrade after 1 to 1.5 years after deployment (informal communication through PurpleAir User Group). Other meteorological factors influencing LCS performance include wind speed, sensor temperature, and sensor type (Liang, 2021).

4 CONCLUSIONS

Failure to invest in calibration may leave large uncertainties in retrieving reliable LCS data that further hinders its broader applications. As a result, field calibration of LCS has been recognized by a larger user group as a critical and necessary step before the LCS deployment for evaluating their reliability and improving the accuracy. Despite the increasing interest, there is an evident knowledge gap on how data-driven algorithms affect calibration performance. This paper aims to provide a first-hand report on the performances of each algorithm, and the impacts of sample size and predictor selection. The key findings are summarized below.

Algorithms respond differently to the baseline dataset and there exists a large variation. While this study implies that NN and GTB slightly outperform the other methods, the users should test

the algorithms on their own as the datasets behave differently. Regression-based methods show the most consistent high accuracy, and we thus recommend it as a viable option for studies with limited effort in parameter tuning and method selection.

The sample size effect is evident in our experiment, especially when the sample size is small. Regardless of the algorithm type, the accuracy drops significantly when the calibration model was trained using less than 1000 records, which is equivalent to six weeks of continuously collected hourly data. However, more training data doesn't always lead to higher accuracy. The accuracy plateaued when the training sample reaches a certain level, which varies slightly among algorithms.

More predictors lead to better accuracies, but the boosting is most evident when $PM_{2.5}$ particle count and humidity were added to the data models. Temperature and particle counts at other sizes play a minor role. Considering the tradeoffs between computational efficiency and more predictors, we suggest the inclusion of $PM_{2.5}$ concentration, particle count, and humidity in the model establishment.

ACKNOWLEDGMENTS

Lu Liang and Jacob Daniels are supported by the National Science Foundation (BCS-2117433). The authors wish to thank Sean Hickey and Rooney Phillips for setting up the calibration system at the Texas site. We thank Dr. Karoline Barkjohn from the US Environmental Protection Agency for sharing with the authors the national field calibration dataset.

DISCLAIMER

Reference to any companies or specific commercial products does not constitute or imply its endorsement by the authors.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.220076

REFERENCES

Amegah, A.K., Dakuu, G., Mudu, P., Jaakkola, J.J.K. (2022). Particulate matter pollution at traffic hotspots of Accra, Ghana: levels, exposure experiences of street traders, and associated respiratory and cardiovascular symptoms. J. Exposure Sci. Environ. Epidemiol. 32, 333–342. https://doi.org/10.1038/s41370-021-00357-x

Apte, J.S., Messier, K.P., Gani, S., Brauer, M., Kirchstetter, T.W., Lunden, M.M., Marshall, J.D., Portier, C.J., Vermeulen, R.C.H., Hamburg, S.P. (2017). High-resolution air pollution mapping with google street view cars: Exploiting big data. Environ. Sci. Technol. 51, 6999–7008. https://doi.org/10.1021/acs.est.7b00891

Ardon-Dryer, K., Dryer, Y., Williams, J.N., Moghimi, N. (2020). Measurements of $PM_{2.5}$ with PurpleAir under atmospheric conditions. Atmos. Meas. Tech. 13, 5441–5458. https://doi.org/10.5194/amt-13-5441-2020

Austin, E., Novosselov, I., Seto, E., Yost, M.G. (2015). Laboratory evaluation of the Shinyei PPD42NS low-cost particulate matter sensor. PLoS One 10, e0137789. https://doi.org/10.1371/journal.pone.0137789

Barkjohn, K.K., Gantt, B., Clements, A.L. (2021). Development and application of a United States-wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor. Atmos. Meas. Tech. 14, 4617–4637. https://doi.org/10.5194/amt-14-4617-2021

Bell, J.N.B., Power, S.A., Jarraud, N., Agrawal, M., Davies, C. (2011). The effects of air pollution on urban ecosystems and agriculture. Int. J. Sustainable Dev. World Ecol. 18, 226–235. https://doi.org/10.1080/13504509.2011.570803

Bi, J., Wildani, A., Chang, H.H., Liu, Y. (2020). Incorporating low-cost sensor measurements into

high-resolution $PM_{2.5}$ modeling at a large spatial scale. Environ. Sci. Technol. 54, 2152–2162. https://doi.org/10.1021/acs.est.9b06046

Breiman, L. (2001). Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cao, J., Chow, J.C., Lee, F.S.C., Watson, J.G. (2013). Evolution of $PM_{2.5}$ measurements and standards in the U.S. and future perspectives for China. Aerosol Air Qual. Res. 13, 1197–1211. https://doi.org/10.4209/aaqr.2012.11.0302

Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302. https://doi.org/10.1016/j.envint.2016.12.007

Colmer, J., Hardman, I., Shimshack, J., Voorheis, J. (2020). Disparities in $PM_{2.5}$ air pollution in the United States. Science 369, 575–578. https://doi.org/10.1126/science.aaz9353

Cortes, C., Vapnik, V. (1995). Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018

Crilley, L.R., Singh, A., Kramer, L.J., Shaw, M.D., Alam, M.S., Apte, J.S., Bloss, W.J., Hildebrandt Ruiz, L., Fu, P., Fu, W., Gani, S., Gatari, M., Ilyinskaya, E., Lewis, A.C., Ng'ang'a, D., Sun, Y., Whitty, R.C.W., Yue, S., Young, S., Pope, F.D. (2020). Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. Atmos. Meas. Tech. 13, 1181–1193. https://doi.org/10.5194/amt-13-1181-2020

De Vito, S., Esposito, E., Castell, N., Schneider, P., Bartonova, A. (2020). On the robustness of field calibration for smart air quality monitors. Sens. Actuators, B 310, 127869. https://doi.org/10.1016/j.snb.2020.127869

Di Antonio, A., Popoola, O.A.M., Ouyang, B., Saffell, J., Jones, R.L. (2018). Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter. Sensors 18, 2790. https://doi.org/10.3390/s18092790

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Boghossian, B.D., Cocker, D., Polidori, A. (2019). Performance evaluation of twelve low-cost $PM_{2.5}$ sensors at an ambient air monitoring site. Atmos. Environ. 216, 116946. https://doi.org/10.1016/j.atmosenv.2019.116946

Feinberg, S.N., Williams, R., Hagler, G., Low, J., Smith, L., Brown, R., Garver, D., Davis, M., Morton, M., Schaefer, J., Campbell, J. (2019). Examining spatiotemporal variability of urban particulate matter and application of high-time resolution data from a network of low-cost air pollution sensors. Atmos. Environ. 213, 579–584. https://doi.org/10.1016/j.atmosenv.2019.06.026

Fix, E., Hodges, J.L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. Int. Stat. Rev. 57, 238. https://doi.org/10.2307/1403797

Freund, Y., Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451

Gao, M., Cao, J., Seto, E. (2015). A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of $PM_{2.5}$ in Xi'an, China. Environ. Pollut. 199, 56–65. https://doi.org/10.1016/j.envpol.2015.01.013

Giordano, M.R., Malings, C., Pandis, S.N., Presto, A.A., McNeill, V.F., Westervelt, D.M., Beekmann, M., Subramanian, R. (2021). From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. J. Aerosol Sci. 158, 105833. https://doi.org/10.1016/j.jaerosci.2021.105833

Gray, S.C., Edwards, S.E., Miranda, M.L. (2013). Race, socioeconomic status, and air pollution exposure in North Carolina. Environ. Res. 126, 152–158. https://doi.org/10.1016/j.envres.2013.06.005

Gupta, A., Bherwani, H., Gautam, S., Anjum, S., Musugu, K., Kumar, N., Anshul, A., Kumar, R. (2021). Air pollution aggravating COVID-19 lethality? Exploration in Asian cities using statistical models. Environ. Dev. Sustain. 23, 6408–6417. https://doi.org/10.1007/s10668-020-00878-9

Hart, R., Liang, L., Dong, P. (2020). Monitoring, mapping, and modeling spatial-temporal patterns of $PM_{2.5}$ for improved understanding of air pollution dynamics using portable sensing technologies. Int. J. Environ. Res. Public Health 17, 4914. https://doi.org/10.3390/ijerph17144914

Hasenfratz, D., Saukh, O., Thiele, L. (2012). On-the-Fly Calibration of Low-Cost Gas Sensors, in:

Picco, G.P., Heinzelman, W. (Eds.), Wireless Sensor Networks, Springer, Berlin, Heidelberg, pp. 228–244. https://doi.org/10.1007/978-3-642-28169-3_15

Hilt, D.E., Seegrist, D.W. (1977). Ridge: A computer program for calculating ridge regression estimates. Research Note NE-236. Upper Darby, U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station, USA. https://www.nrs.fs.fed.us/pubs/9260

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. PNAS 79, 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S., Buckley, K. (2016). Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. Atmos. Meas. Tech. 9, 5281–5292. https://doi.org/10.5194/amt-9-5281-2016

Johnson, N.E., Bonczak, B., Kontokosta, C.E. (2018). Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. Atmos. Environ. 184, 9–16. https://doi.org/10.1016/j.atmosenv.2018.04.019

Kim, S., Park, S., Lee, J. (2019). Evaluation of performance of inexpensive laser based $PM_{2.5}$ sensor monitors for typical indoor and outdoor hotspots of South Korea. Appl. Sci. 9, 1947. https://doi.org/10.3390/app9091947

Liang, L., Gong, P., Cong, N., Li, Z., Zhao, Y., Chen, Y. (2019). Assessment of personal exposure to particulate air pollution: the first result of City Health Outlook (CHO) project. BMC Public Health 19, 711. https://doi.org/10.1186/s12889-019-7022-8

Liang, L., Gong, P. (2020). Urban and air pollution: A multi-city study of long-term effects of urban landscape patterns on air quality trends. Sci. Rep. 10, 18618. https://doi.org/10.1038/s41598-020-74524-9

Liang, L. (2021). Calibrating low-cost sensors for ambient air monitoring: Techniques, trends, and challenges. Environ. Res. 197, 111163. https://doi.org/10.1016/j.envres.2021.111163

MacKay, D.J.C. (1992). Bayesian interpolation. Neural Comput. 4, 415–447. https://doi.org/10.1162/neco.1992.4.3.415

Magi, B.I., Cupini, C., Francis, J., Green, M., Hauser, C. (2020). Evaluation of $PM_{2.5}$ measured in an urban setting using a low-cost optical particle counter and a Federal Equivalent Method Beta Attenuation Monitor. Aerosol Sci. Technol. 54, 147–159. https://doi.org/10.1080/02786826.2019.1619915

Mahajan, S., Kumar, P. (2020). Evaluation of low-cost sensors for quantitative personal exposure monitoring. Sustain. Cities Soc. 57, 102076. https://doi.org/10.1016/j.scs.2020.102076

Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S.P.N., Zimmerman, N., Kara, L.B., Presto, A.A., Subramanian, R. (2019). Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. Atmos. Meas. Tech. 12, 903–920. https://doi.org/10.5194/amt-12-903-2019

Mao, F., Khamis, K., Krause, S., Clark, J., Hannah, D.M. (2019). Low-cost environmental sensor networks: Recent advances and future directions. Front. Earth Sci. 7, 221. https://doi.org/10.3389/feart.2019.00221

Mardia, K., Kent, J.T., Bibby, J. (1979). Multivariate analysis. Academic Press, London.

Masson, N., Piedrahita, R., Hannigan, M. (2015). Quantification method for electrolytic sensors in long-term monitoring of ambient air quality. Sensors 15, 27283–27302. https://doi.org/10.3390/s151027283

McFarlane, C., Isevulambire, P.K., Lumbuenamo, R.S., Ndinga, A.M.E., Dhammapala, R., Jin, X., McNeill, V.F., Malings, C., Subramanian, R., Westervelt, D.M. (2021). First measurements of ambient $PM_{2.5}$ in Kinshasa, Democratic Republic of Congo and Brazzaville, Republic of Congo using field-calibrated low-cost sensors. Aerosol Air Qual. Res. 21, 200619. https://doi.org/10.4209/aaqr.200619

Molnár, A., Imre, K., Ferenczi, Z., Kiss, G., Gelencsér, A. (2020). Aerosol hygroscopicity: Hygroscopic growth proxy based on visibility for low-cost PM monitoring. Atmos. Res. 236, 104815. https://doi.org/10.1016/j.atmosres.2019.104815

Morawska, L., Thai, P.K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G.S.W., Jayaratne, R., Kumar, P., Lau, A.K.H., Louie, P.K.K., Mazaheri, M., Ning, Z., Motta, N., Mullins, B., *et al.* (2018). Applications of low-

cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? Environ. Int. 116, 286–299. https://doi.org/10.1016/j.envint.2018.04.018

Mousavi, A., Yuan, Y., Masri, S., Barta, G., Wu, J. (2021). Impact of 4th of July fireworks on spatiotemporal $PM_{2.5}$ concentrations in California based on the PurpleAir sensor network: Implications for policy and environmental justice. Int. J. Environ. Res. Public. Health 18, 5735. https://doi.org/10.3390/ijerph18115735

Okafor, N.U., Alghorani, Y., Delaney, D.T. (2020). Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach. ICT Express 6, 220–228. https://doi.org/10.1016/j.icte.2020.06.004

Olivares, G., Edwards, S. (2015). The Outdoor Dust Information Node (ODIN) – development and performance assessment of a low cost ambient dust sensor. Atmos. Meas. Tech. Discuss. 8, 7511–7533. https://doi.org/10.5194/amtd-8-7511-2015

Ouimette, J.R., Malm, W.C., Schichtel, B.A., Sheridan, P.J., Andrews, E., Ogren, J.A., Arnott, W.P. (2022). Evaluating the PurpleAir monitor as an aerosol light scattering instrument. Atmos. Meas. Tech. 15, 655–676. https://doi.org/10.5194/amt-15-655-2022

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. (2011). Scikit-learn: Machine learning in Python. JMLR 12, 2825–2830.

Peled, R. (2011). Air pollution exposure: Who is at high risk? Atmos. Environ. 45, 1781–1785. https://doi.org/10.1016/j.atmosenv.2011.01.001

Pope, F.D., Gatari, M., Ng'ang'a, D., Poynter, A., Blake, R. (2018). Airborne particulate matter monitoring in Kenya using calibrated low-cost sensors. Atmos. Chem. Phys. 18, 15403–15418. https://doi.org/10.5194/acp-18-15403-2018

Rai, A.C., Kumar, P., Pilla, F., Skouloudis, A.N., Di Sabatino, S., Ratti, C., Yasar, A., Rickerby, D. (2017). End-user perspective of low-cost sensors for outdoor air pollution monitoring. Sci. Total Environ. 607–608, 691–705. https://doi.org/10.1016/j.scitotenv.2017.06.266

Robinson, D.L. (2020). Accurate, low cost $PM_{2.5}$ measurements demonstrate the large spatial variation in wood smoke pollution in regional Australia and improve modeling and estimates of health costs. Atmosphere 11, 856. https://doi.org/10.3390/atmos11080856

Sager, L. (2019). Estimating the effect of air pollution on road safety using atmospheric temperature inversions. J. Environ. Econ. Manage. 98, 102250. https://doi.org/10.1016/j.jeem.2019.102250

Schmidt-Ott, A., Ristovski, Z.D. (2003). Measurement of Airborne Particles, in: Morawska, L., Salthammer, T. (Eds.), Indoor Environment, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 56–81. https://doi.org/10.1002/9783527610013.ch2b

Snyder, E.G., Watkins, T.H., Solomon, P.A., Thoma, E.D., Williams, R.W., Hagler, G.S.W., Shelow, D., Hindin, D.A., Kilaru, V.J., Preuss, P.W. (2013). The changing paradigm of air pollution monitoring. Environ. Sci. Technol. 47, 11369–11377. https://doi.org/10.1021/es4022602

Stavroulas, I., Grivas, G., Michalopoulos, P., Liakakou, E., Bougiatioti, A., Kalkavouras, P., Fameli, K.M., Hatzianastassiou, N., Mihalopoulos, N., Gerasopoulos, E. (2020). Field evaluation of low-cost PM sensors (Purple Air PA-II) under variable urban air quality conditions, in Greece. Atmosphere 11, 926. https://doi.org/10.3390/atmos11090926

Tsou, M.C.M., Lung, S.C.C., Shen, Y.S., Liu, C.H., Hsieh, Y.H., Chen, N., Hwang, J.S. (2021). A community-based study on associations between $PM_{2.5}$ and PM_1 exposure and heart rate variability using wearable low-cost sensing devices. Environ. Pollut. 277, 116761. https://doi.org/10.1016/j.envpol.2021.116761

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tryner, J., L'Orange, C., Mehaffy, J., Miller-Lionberg, D., Hofstetter, J.C., Wilson, A., Volckens, J. (2020). Laboratory evaluation of low-cost PurpleAir PM monitors and in-field correction using co-located portable filter samplers. Atmos. Environ. 220, 117067. https://doi.org/10.1016/j.atmosenv.2019.117067

U.S. Environmental Protection Agency (U.S. EPA) (2011). Reference and Equivalent Method Applications: Guidelines for Applicants. National Exposure Research Laboratory, U.S. Environmental Protection Agency, USA. https://www.epa.gov/sites/production/files/2017-02/documents/frmfemguidelines.pdf

US Government Accountability Office (U.S. GAO) (2020). Opportunities to better sustain and modernize the national air quality monitoring system. https://www.gao.gov/assets/720/711027.pdf

Watson, J.G., Chow, J.C., DuBois, D., Green, M., Frank, N. (1997). Guidance for the network design and optimum site exposure for $PM_{2.5}$ and PM_{10} (Technical Report No. PB-99-157513/XAB; EPA-454/R-99/022 TRN: 92291323). Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, USA. https://www3.epa.gov/ttnamti1/files/ambient/pm25/network/r-99-022.pdf

Yamamoto, K., Togami, T., Yamaguchi, N., Ninomiya, S. (2017). Machine learning-based calibration of low-cost air temperature sensors using environmental data. Sensors 17, 1290. https://doi.org/10.3390/s17061290

Zheng, T., Bergin, M.H., Johnson, K.K., Tripathi, S.N., Shirodkar, S., Landis, M.S., Sutaria, R., Carlson, D.E. (2018). Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments. Atmos. Meas. Tech. 11, 4823–4846. https://doi.org/10.5194/amt-11-4823-2018

Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, E.S., Robinson, A.L., Subramanian, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos. Meas. Tech. 11, 291–313. https://doi.org/10.5194/amt-11-291-2018

Zusman, M., Schumacher, C.S., Gassett, A.J., Spalt, E.W., Austin, E., Larson, T.V., Carvlin, G., Seto, E., Kaufman, J.D., Sheppard, L. (2020). Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. Environ. Int. 134, 105329. https://doi.org/10.1016/j.envint.2019.105329