Leveraging Machine Learning Algorithms to Advance Low-Cost Air Sensor Calibration in Stationary and Mobile Settings

An Wang[1], Yuki Machida[1], Priyanka deSouza[2], Simone Mora[1,3]*, Tiffany Duhl[4], Neelakshi Hudda[4], John L. Durant[4], Fabio Durate[1], and Carlo Ratti[1]

1. Senseable City Lab, Department of Urban Studies and Planning, Massachusetts Institute of Technology

2. Department of Urban and Regional Planning, University of Colorado Denver

3. Department of Computer Science, Norwegian University of Science and Technology

4. Department of Civil and Environmental Engineering, Tufts University

* corresponding author

ABSTRACT

Low-cost air sensing is changing the paradigm of ambient air quality management research and practices. However, consensus on a structured low-cost sensor calibration and performance evaluation framework is lacking. Our study aims to devise a standardized low-cost sensor calibration protocol and evaluate the performance of various calibration algorithms. Extensive collocation data were collected in stationary and mobile settings in two American cities, New York and Boston. We trained the calibration models using stationary data aggregated at various intervals to examine the performance of several commonly used calibration algorithms described in the literature. Linear models provide consistently satisfactory calibration results, indicating linear responses from the low-cost sensors in our stationary test environment. Its simplicity is recommended for citizen science and education usages. Models that can account for non-linear relationships, especially random forest, perform well and transfer between sensors better than generalized linear regression models for $PM_{2.5}$ calibration, which should be adopted for regulatory and scientific purposes. Data collected in a mobile validation campaign in Boston were passed through the best-performing calibration models to assess their transferability. The results indicate that models trained with data from a different urban environment and season in the stationary setting did not transfer well to a mobile setting. It is recommended that low-cost sensors should be calibrated more often than suggested in Environmental Protection Agency's air sensor performance evaluation guidelines and used in an environment that is as similar as possible to the calibration environment.

Keywords: low-cost sensor calibration, $PM_{2.5}$, NO_2, machine learning, mobile monitoring, environmental justice

1 INTRODUCTION

Lacking the resources to deploy high-quality monitors, low- and middle-income countries are held back from making effective measurements to inform air pollution management. The rapid development of low-cost sensors (< $2500(US EPA, 2014)) in recent years provides a unique opportunity to shift the current air quality monitoring paradigm. While low-cost sensors have been widely adopted in high-income regions to supplement traditional air quality monitoring and mapping of local air quality (Castell et al., 2017; Crawford et al., 2021; Gressent et al., 2020; Miskell et al., 2018), their applications in low-income countries are limited (Brauer et al., 2019; deSouza et al., 2020; SM et al., 2019). The benefits of incorporating low-cost sensors in the existing monitoring network are multifold, including improving public awareness of air pollution, pushing for environmentally just decision-making and reducing *data colonialism* in air quality monitoring, where government agencies and high-income countries claim ownership of collected air quality data (Duarte & deSouza, 2020).

Despite the advantages in cost efficiency, flexibility, and ease of use, low-cost sensors suffer from constant data quality and stability issues; therefore, sensor collocation and calibration are of utmost importance prior to field deployment. Collocation is the process of deploying sensors side-by-side with reference monitors, and calibration involves adjusting raw sensor readings using collocation data and mathematical methods. We summarized seven issues that affect low-cost sensor performance from low-cost sensor calibration literature within the last five years: inter-sensor variability, intra-sensor variability, drift, aging, response time, cross-sensitivity, and sensitivity to environmental factors. Table S1 presents detailed findings from each study.

Inter-sensor variability refers to the variability in measurements using multiple identical sensors under the same testing environment. The calibration for inter-sensor variability is crucial for low-cost sensors as it is the foundation of large-scale sensor deployment and data transferability. Intra-sensor variability describes the variability in consecutive measurements made by a given sensor under a similar testing environment. Drift is the gradual change in sensor response over time. Aging refers to the continuous deterioration of sensor performance over time. Unlike reference sensors, low-cost sensors are prone to drifting and have a much shorter lifetime. Thus, they require routine sensor calibration and replacement. Response time reflects the lag before sensors reach stable readings in the test environment, parity between response time and temporal scale of change in signal is critical for time-resolved sampling. Cross-sensitivity, an issue exclusive to gas sensors, denotes a sensor's false response to gases other than the target gas. Finally, sensitivity to environmental factors, including temperature, humidity, wind, barometric pressure, and particle composition, is ubiquitous in both low-cost and reference sensors. These environmental factors are commonly identified as the main explanatory features in low-cost sensor calibration models. Acknowledging that there might be no single solution to all seven issues in given applications, it is essential to characterize the performance of various calibration algorithms and identify the proper context for their usage.

In light of low-cost sensor performance issues and the surging interest in using sensors for air quality monitoring in areas where regulatory monitors are absent, the US Environmental Protection Agency (EPA) has been engaging local communities, subject matter experts, and air sensor manufacturers to develop guidelines for low-cost sensor usage. EPA published two performance target reports for O_3 and particulate matter (PM) sensors to support low-cost sensor non-regulatory supplemental and informational monitoring (NSIM) applications (US EPA, 2021a, 2021b). These reports provide low-cost sensor collocation, calibration, performance evaluation, and deployment protocols. The reports suggest using linear regression to calibrate low-cost sensor data against the reference station for ease of use; however, linear calibration may not meet performance targets in many cases (Kelly et al., 2017; Malings et al., 2019).

The three main categories of low-cost sensor calibration algorithms currently in use include physical mechanism-based models, parametric models such as linear and non-linear regression models, and machine learning models (Liang, 2021). To date, physical mechanism-based models have targeted the problem that relative humidity can change PM hygroscopic size and refractive index, influencing conversion from PM number count to mass concentrations. Crilley et al. (2018) and Zheng et al. (2018) applied κ-Köhler theory and scattering efficiency, respectively, to adjust relative humidity correction factors for better low-cost PM sensor calibration. Parametric models, including linear and non-linear regression models, are more widely used in the literature, as shown in Table S1. They can consider more features affecting low-cost sensor readings, while physical mechanism-based methods mainly calibrate for relative humidity. As state-of-the-art calibration algorithms, machine learning approaches can provide higher accuracy in data-rich environments

(Liang, 2021; Malings et al., 2019; Wang et al., 2021). Nonetheless, some machine learning algorithms lack interpretability and transferability, which hinders their broader implementation.

Given that there is no all-in-one solution to low-cost sensor calibration, it is necessary to explore various calibration approaches in different applications and compare their performance. To this end, we tested multiple commonly used low-cost sensor calibration algorithms in stationary and mobile settings. The calibration models trained with stationary collocation data were further examined with mobile collected data for transferability tests. We were motivated to simulate a common scenario deploying sensors calibrated in one location in a different location that does not have high-quality air instruments to enable local colocation and calibration. This is especially useful for sensor applications in low-income regions and countries to address the existing air quality data gap and related environmental justice issues. This study is among the first to incorporate the standardized EPA air sensor evaluation frameworks and the effects of data aggregation in stationary and mobile settings. Beyond simply comparing model performance by investigating algorithm behaviors and the impact factors of sensor performance, this study explores the appropriate application scenarios for multiple calibration algorithms to account for non-linearity in low-cost sensor performance. Finally, our study provides valuable information on algorithm transferability by exploiting two large air monitoring datasets collected in different locations under stationary and mobile conditions.

2 METHODOLOGY

Our study included two main components: developing calibration models and evaluating the transferability of the models. As shown in Figure 1, reference station data, raw low-cost sensing data, and meteorological data during the collocation were first aggregated at various temporal aggregations to test the effects of data aggregation on the algorithms. We employed a k-fold time series cross-validation to avoid overfitting and performance overestimation, where past observations were used to predict future observations but not the other way around. Two major categories of calibration algorithms were tested: the generalized linear regressions (yellow) and the machine learning algorithms (red). The best-performing model of each calibration algorithm was passed to the transferability test using mobile data collected with research-grade mobile sensors.

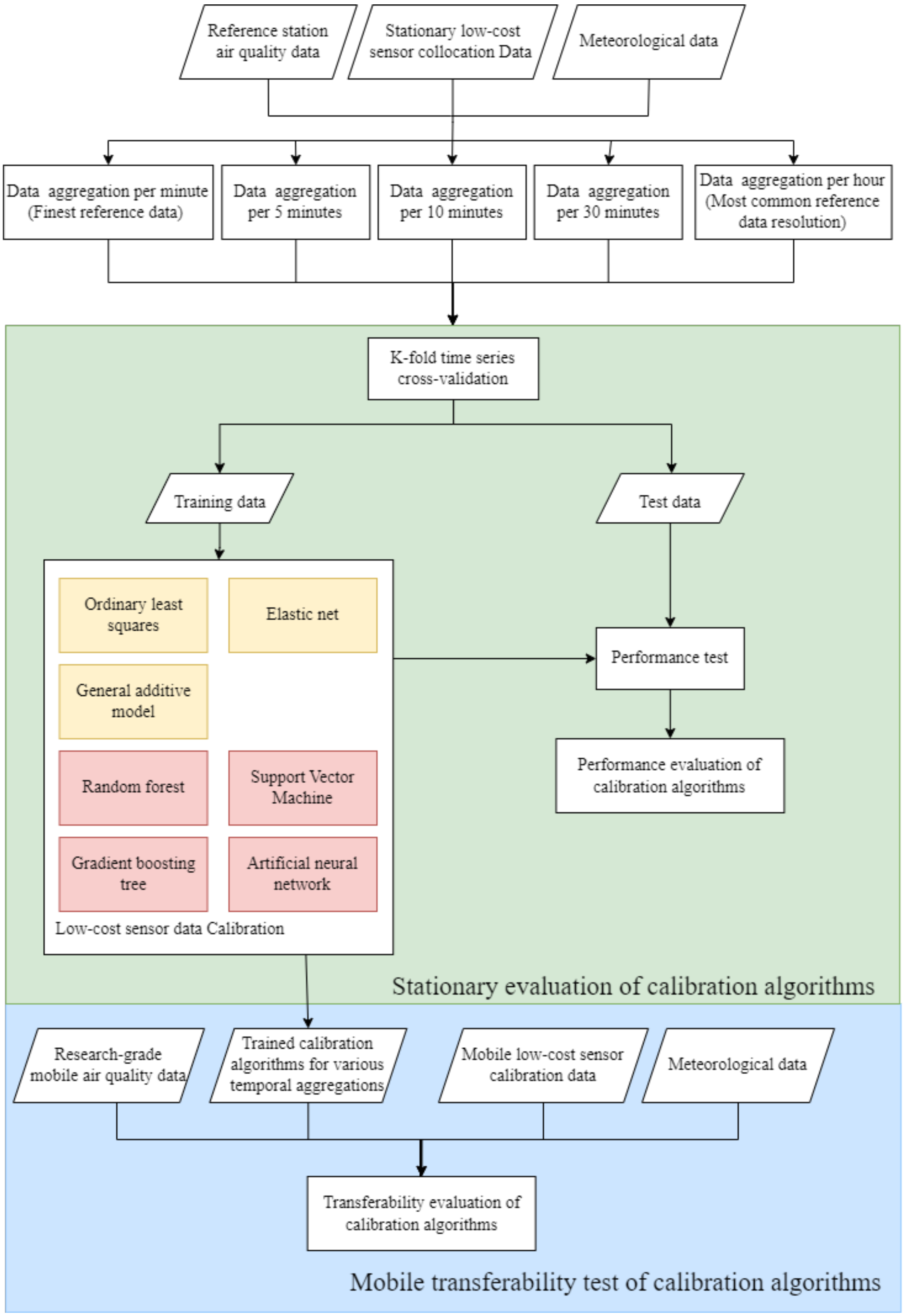


Figure 1 Methodology flow chart

2.1 Low-Cost Sensors

The mobile low-cost sensing platform we evaluate is developed as part of the City Scanner initiative at the Senseable City Lab, Massachusetts Institute of Technology. It is designed to address big data needs for urban development and planning while helping advance environmental justice in global cities. For simplicity, the sensing platform is referred to as the 'City Scanner' or 'CS' hereafter. Each CS unit is equipped with an Alphasense OPC-N3 and an Alphasense NO_2-A43F that measure particle counts and nitrogen dioxide, respectively. The Alphasense OPC-N3 is an optical particle counter with a nominal particle monitoring size range from 0.35 to 40 micrometers, which has been widely adopted in multiple previous studies (Bezantakos et al., 2018; Crilley et al., 2018, 2020; Sousan et al., 2016). It provides particle counts in 24 size bins and estimates PM_1, $PM_{2.5}$, and PM_{10} mass concentrations based on particle shape and density presumptions. Alphasense NO2-A43F is an electro-chemical gas sensor that can operate in a variety of ambient environments but is known for its cross-sensitivity issue with NO and CO_2, while its cross-sensitivity with ozone is minimized with an ozone scrubber. The NO_2 readings are given as electric signals (millivoltages) and need to be converted to parts per billion by volume. The low-cost sensors are integrated into the CS, which is a standalone sensing platform with data storage and remote monitoring capability. The platform is powered by a battery that can be charged through a power supply or a solar panel. The data from the sensors are stored on an onboard SD card, which can be accessed remotely through an LTE connection for routine checks on the data from sensors and the device's status.

2.2 Stationary and Mobile Data Collection

As per EPA's air sensor calibration guidelines (US EPA, 2021a, 2021b), the stationary collocation lasted four weeks, from August 13 to September 10, 2021. With permission from the New York State Department of Environmental Conservation, five CS units were placed next to a reference station collecting PM and NO_2 data every five seconds. The reference station (Site ID: IS 52/MS 302) is located on the roof of New York City Department of Education Public School 52/Middle School 302 in South Bronx. The station measures criteria pollutants, particulate matter speciation, hydrocarbons, and volatile organic chemicals continuously or intermittently. We obtained continuous minute-by-minute measurements of $PM_{2.5}$ and NO_2 at the station from

NYSDEC, measured by Teledyne T640 at 5.0 liter per minute (Federal Equivalent Method 236) and Thermo Environmental Instruments 42C (Federal Reference Method 074), respectively.

In addition to the stationary collocation, a mobile validation comparison campaign was carried out using the Tufts Air Pollution Laboratory (TAPL), which measured real-time particulate matter and NO_x concentrations (Hudda et al., 2020; Padró-Martínez et al., 2012). TAPL is housed in an electric vehicle to avoid self-pollution and equipped with research-grade instruments, including a DustTrak™ DRX Aerosol Monitor (Model 8533, TSI, Shoreview, MN) for size-resolved mass fraction concentrations and a NO-NO_2-NO_x Analyzer (Model 42i, Thermo Fisher, Waltham, MA) for NO/NO_2/NO_x concentrations. TAPL instruments were synchronized with CS units using GPS time, reporting $PM_{2.5}$ and NO_2 readings every second and every ten seconds, respectively. TAPL Dusttrak was calibrated against a regulatory site in Boston managed by the Massachusetts Department of Environmental Protection (Nubian Square, site ID 250250042). Detailed information on this calibration is documented in the SI. Two CS units were mounted on the TAPL for mobile validation in Boston. The mobile validation campaign occurred in neighborhoods north of Boston Logan International Airport and lasted from February 11 to April 8, 2022, with data collected on 16 different days. It is worth noticing that between the New York collocation and Boston mobile validation with TAPL, from September 13 to December 16, we continuously used these CS units for opportunistic mobile air quality for four months (unattended in an open environment) on municipal service vehicles. This study does not report these data as they are irrelevant to sensor calibration.

2.3 Data Processing and Calibration

2.3.1 Data quality assurance and aggregation

In total, during the stationary collocation, 1.6 million unique 5-second resolution data points were collected by the five CS units in New York. During the mobile validation, 130,000 unique 5-second-interval data points were recorded by two CS units in Boston. The CS data were then filtered by two criteria before they were aggregated. Fifteen percent of raw CS data from stationary collocation were excluded in the process, where 12% percent were removed due to high relative humidity (> 90% or raining), while the other 3% due to extreme readings (< 1 μg/m^3 or >1000 μg/m^3 for $PM_{2.5}$, < 200 mv or > 900 mv for NO_2 electro-signal).

The filtered raw CS data were aggregated into five temporal resolutions, including minute-level, 5-minute-level, 10-minute-level, 30-minute-level, and hour-level. Minute-by-minute reference air quality data were aggregated in the same intervals. Hourly meteorological data were gathered from the nearest LaGuardia Airport, about 4 kilometers away. Raw CS data, reference data, and meteorology were synchronized and matched, yielding five datasets of various temporal aggregations to explore data aggregation's effect on sensor calibration. Similarly, mobile validation CS data were cleaned using the same criteria as in the stationary collocation. The filtered CS and TAPL data were matched and aggregated to the five temporal resolutions. Hourly meteorological data were obtained from the Boston Logan International Airport (within 5 km of the driving area) and matched with air quality data.

2.3.2 Development of calibration models

Our study evaluated seven distinct algorithms in two categories, the generalized linear regressions and machine learning algorithms, which frequently appear in the existing literature. The algorithms were selected in light of a study that assessed the air quality prediction performance by Kerckhoffs et al. (2019). Ordinary least squares (OLS) is the most widely adopted and straightforward algorithm in sensor calibration, which is also recommended by the EPA air sensor performance evaluation guidelines. It is an effective tool for sensor calibration providing intuitive results and straightforward interpretation. Elastic net is essentially an extension of linear regression with penalty terms (Lasso and/or Ridge regularization) for correlated explanatory features. While we acknowledge that there is great variability and capability in linear regression algorithms, we only adopted the simplest form of these algorithms as recommended by the EPA guidelines for citizen science and educational purposes without considering feature transformations. In addition, feature transformation and smoothing are taken into account in the generalized additive model (GAM), which estimates the target feature with linear combinations of smooth functions of the features. GAM models in this study did not consider interactions between the features for better interpretability.

Regarding machine learning algorithms, supporter vector regression (SVR) is a commonly used regression method that uses a kernel function to transform the data into a higher dimension and finds the hyperplane that fits the data with a desirable error margin. Random forest and Light Gradient Boosting Machine (LightGBM) belong to the bigger ensemble modeling family but employ bagging and boosting techniques, respectively. Specifically, LightGBM is considered the state-of-the-art gradient boosting framework originally developed by Microsoft, which has faster training speed, higher efficiency, and lower memory usage while yielding better accuracy (Ke et al., 2017). Artificial neural network (ANN) is popular in regression for prediction purposes, yet its black box nature hinders its wider adoption in explanatory tasks.

For each temporally aggregated dataset, we adopted a k-fold time series cross-validation approach to evaluate calibration models' performance (Pedregosa et al., 2011). The approach is designed to avoid predicting past observations from future ones, limiting data leakage and performance overestimation. Each dataset is first ranked temporally and divided into five sequential sections with the same number of observations. In this case, the first k folds are used to train the models, and the k+1 fold split is used to validate and test the model. The train-test split is illustrated in Figure 2. In the training and test sets, the CS readings for $PM_{2.5}$ and NO_2, and meteorological factors, including temperature in Celsius, relative humidity in percentage, dew point in Celsius, and air pressure in kilopascal, were treated as explanatory features, and the reference pollutant readings were the target feature. Other meteorological factors, such as wind speed, wind direction, and feel temperature, were also tested in the models but excluded due to statistical insignificance. Detailed feature descriptions and model-tuning information are documented in the SI. All features were tested for their normality and transformed if not following a normal distribution. It is worth noting that unlike the PM sensor, which provides mass concentration, the NO_2 sensor yields electric signals instead of parts per billion by volume as the raw readings. Conventionally, one converts electro-signals to gas concentrations using sensitivity factors provided by the manufacturer from factory calibration, which is conducted for every batch of sensors. Then, the converted gas concentrations are calibrated with reference concentrations collected in the field. Our approach directly mapped the electric signals to reference concentrations rather than using factory-calibrated sensitivity factors, which can be different between sensor batches and taking average within each batch.

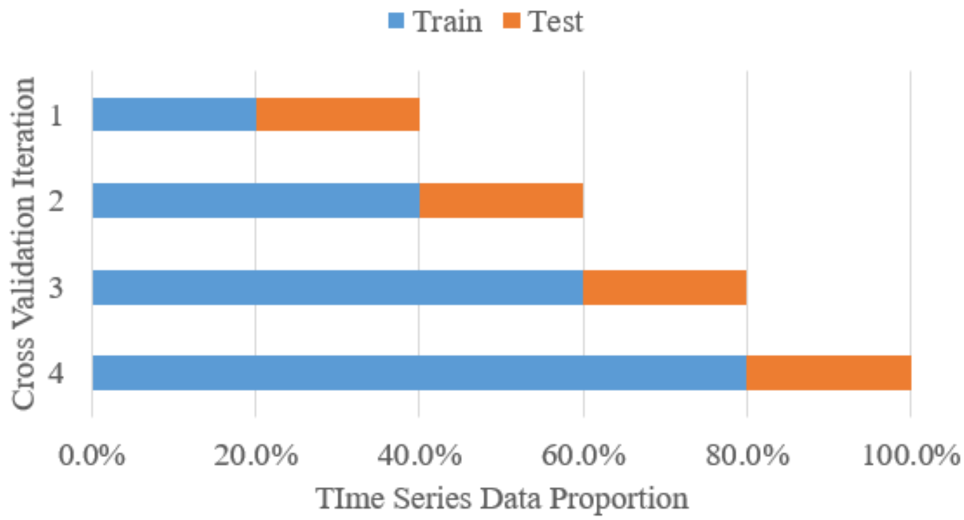


Figure 2. Illustration of k-fold time series cross-validation splitCalibration Algorithm Performance Evaluation and Transferability tests

2.3.3 Performance measures

In each fold of the time series cross-validation, the trained model was assessed against the test dataset using two performance measures, r^2 and the root mean squared error (RMSE) as shown in Equations 1 and 2. r^2 simply represents the square of the Pearson correlation coefficient between predictions and observations. It measures the correlation between calibrated CS readings and reference readings, while RMSE quantifies the absolute difference between them. It is worth noting that even though we train models with log-transformed CS and reference air pollutant readings, both performance measures are calculated using back-transformed predictions and observations.

$$r^2 = \left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}\right)^2 \qquad \text{Eq.1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \qquad \text{Eq.2}$$

Where:

y_i and $\hat{y}_i$ are observed and predicted target features in the test dataset;

$\bar{y}$ and $\bar{\hat{y}}$ are mean values of observed and predicted target features in the test dataset;

n is the size of the test dataset.

The best-performing models and temporally aggregated dataset were selected for each algorithm (7 algorithms × 5 temporal resolutions = 35 models). Mobile CS data collected by TAPL were passed to the models, whose transferability was evaluated using the same performance measures.

2.3.4 Model interpretation

In addition to performance evaluation, we explored the impact of meteorological factors on sensor and model performance. We focus the model interpretation on the ensemble algorithms, as they are widely adopted in the existing calibration literature to account for non-linear relationships with an established global interpretation method. We employed the Shapley Additive exPlanations (SHAP) method (Lundberg et al., 2018; Lundberg & Lee, 2017) to interpret the trained random forest and LightGBM model. The method is based on Shapley values, a commonly used metric to measure contributions from players in cooperative game theory. It treats each explanatory feature as a player and calculates its importance by comparing the model performance with and without the feature. Another useful visualization, the SHAP summary plot, sorts the features by their global impact and calculates each feature's impact on the model output of each test sample, illustrating complex associations between the target and explanatory features.

3 RESULTS

3.1 Descriptive Analysis of Stationary Collocation and Mobile Validation

We first examined the raw data collected by CS units compared to the reference. As shown in Table 1, the average $PM_{2.5}$ concentrations during the sampling periods were 7.9 and 4.7 μg/m^3, respectively, in New York City and Boston. All five CS units underestimated $PM_{2.5}$ concentrations in stationary collocation by 62.1% on average in New York City. During the mobile validation in Boston, CS Unit 3 overestimated $PM_{2.5}$ concentrations, while CS Unit 5 still underestimated $PM_{2.5}$. The discrepancy is mainly caused by the different particle characteristics in New York and Boston and inter-sensor sensing variability for these particles. The inter-sensor variability of CS units is exhibited, which indicates the importance of developing unit-specific calibration models. The specifically high inter-sensor variability in the Boston campaign is mainly attributable to the continuous deployment in an unattended and open environment in NYC. The average reference NO_2 concentrations were 10.2 and 14.3 μg/m^3, respectively, in New York City and Boston, which is not included in Table 1 as CS provides only raw millivoltage readings.

Table 1. Raw $PM_{2.5}$ levels in μg/m^3 collected by City Scanner and reference monitors during stationary collocation in New York City and mobile comparison in Boston

	New York City (μg/m^3)			Boston (μg/m^3)		
	25th percentile	50th percentile	75th percentile	25th percentile	50th percentile	75th percentile
CS Unit 1	1.6	2.6	4.1	-	-	-
CS Unit 2	2.8	4.1	6.2	-	-	-
CS Unit 3	1.7	2.8	4.4	3.9	5.4	8.1
CS Unit 4	2.3	3.6	5.6	-	-	-
CS Unit 5	2.2	3.6	5.5	0.2	0.5	1.1
Reference	5.6	7.3	10.3	1.2	2.2	3.0

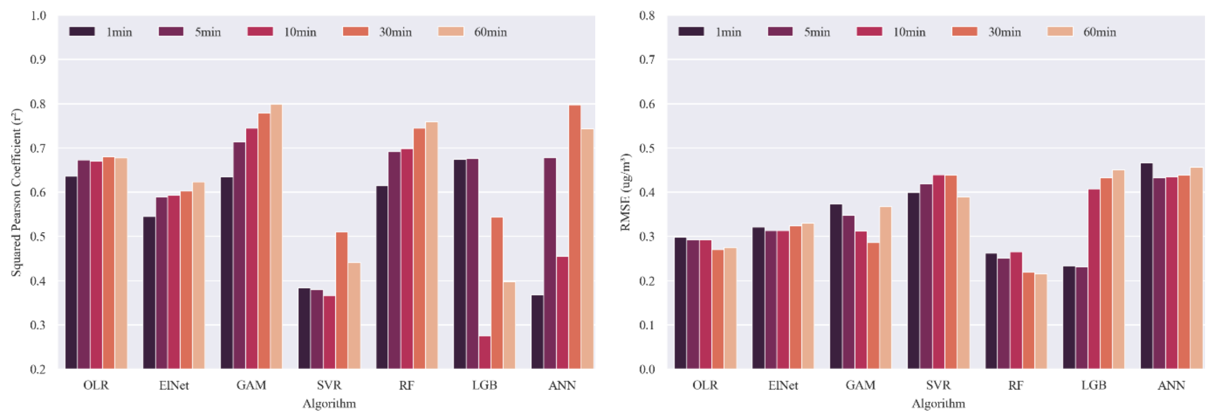
The weather conditions were different during the NYC stationary collocation and Boston mobile validation. Table 2 demonstrates the meteorological factors that were employed in the calibration algorithms. While relative humidity and air pressure were relatively consistent regionally across the two monitoring campaigns, air temperature and dew point were about 20 °C lower during the Boston mobile campaign than NYC.

Table 2. Meteorological conditions during the stationary collocation and mobile validation

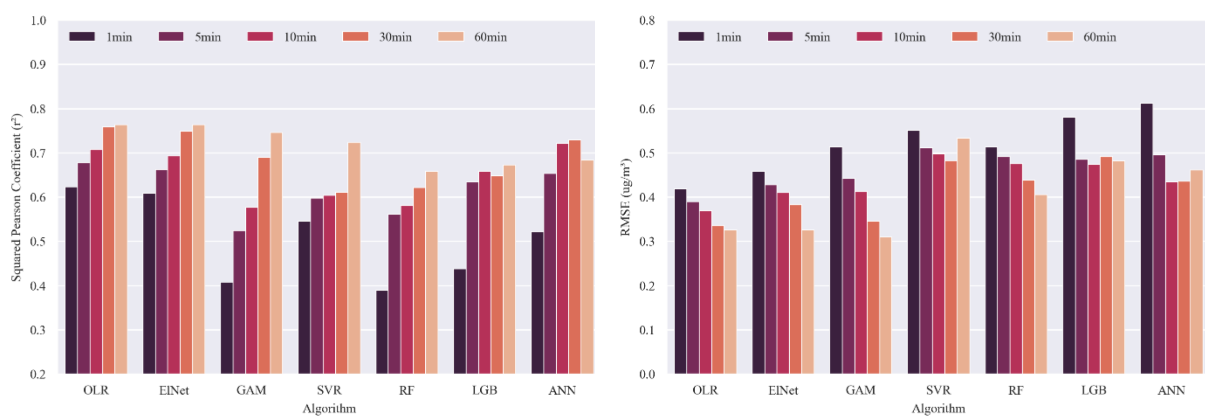
Duration and meteorology	**New York average (Standard deviation)**	**Boston average (Standard deviation)**
Duration	August 13 - September 10, 2021	February 11 - April 8, 2021
Temperature (Celsius)	25.6 (±3.5)	4.1 (±6.4)
Relative humidity (%)	64.5 (±14.6)	58.7 (±21.8)
Dew point (Celsius)	17.9 (±3.5)	-4.1 (±8.6)
Air pressure (kPa)	101.5 (±0.5)	101.6 (±1.0)

3.2 Algorithm Performances in a Stationary Setting

We summarized the calibration algorithm performance (r^2 and RMSE) of CS Unit 1 for $PM_{2.5}$ and NO_2 in Figure 3. Performance summary plots for other CS units can be found in the Supporting Information. After calibration, all five CS units present good agreement with the reference data. In general, the algorithms provide good calibration results for both $PM_{2.5}$ and NO_2, while PM calibration models have slightly better performance than NO_2 ones across all CS units. Best PM models' r^2 values reach over 0.8, compared to those of NO_2 models around 0.7. Still, it indicates the feasibility of calibrating CS NO_2 readings directly from millivoltage signals to reference concentrations. Among all algorithms, ordinary least squares and elastic net have demonstrated consistently satisfactory performance in $PM_{2.5}$ calibration, r^2 ranging from 0.6 to 0.8. This phenomenon is also observed in NO_2 calibration, indicating good linear responses between the low-cost sensors and reference sensors. RMSE values are very low after calibration, mostly below 0.5 μg/m^3 and 0.8 ppb for $PM_{2.5}$ and NO_2, respectively, considering that the median concentrations for $PM_{2.5}$ and NO_2 are 7.3 μg/m^3 and 9.0 ppb in the stationary collocation. Machine learning models can also deliver moderate calibration results but sometimes fail due to overfitting the training data. Across all units and algorithms, we do not observe an obvious effect of temporal aggregation interval on model performance.



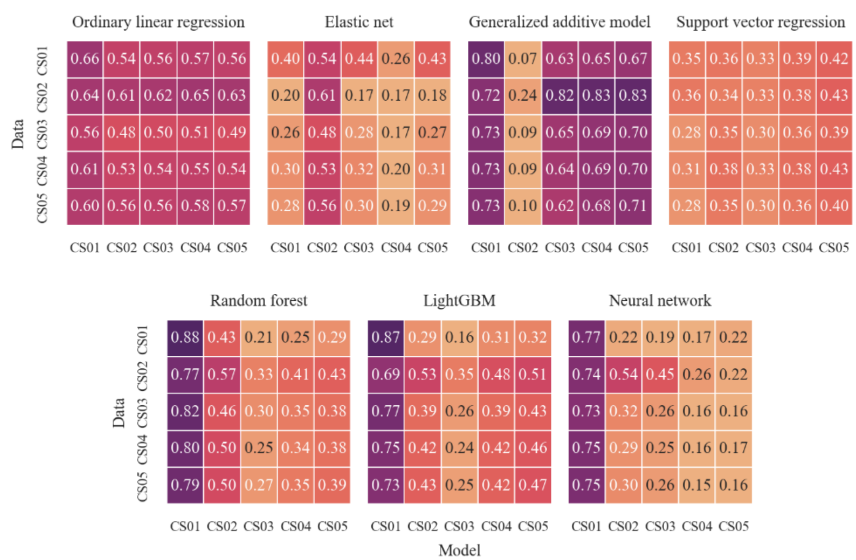
(a) City Scanner Unit 1 $PM_{2.5}$ performance



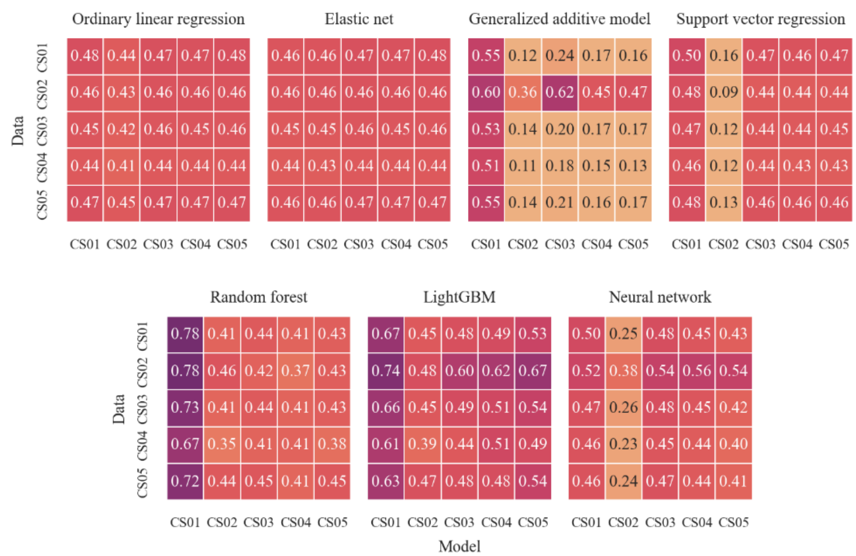
(b) City Scanner Unit 1 NO_2 performance

Figure 3. K-fold cross validation model performances (a) $PM_{2.5}$ and (b) NO_2 in a stationary setting for City Scanner Unit 1.

Figure 4 tests the inter-sensor algorithm transferability for the best-performing $PM_{2.5}$ and NO_2 models. Test data from a CS unit were passed through models developed using data from another CS unit, which process was repeated for all five CS units and all calibration algorithms. 1-min aggregated data and models were used as they provide the worst-case scenario in transferability and are most prone to overfitting data from a specific sensor. It is hypothesized that linear regressions should have the best transferability across sensors as they are the least likely to overfit. Nevertheless, for $PM_{2.5}$ models, the generalized additive model and random forest are observed to have the best inter-sensor transferability, while the transferability of linear regressions is moderate. For NO_2, ordinary least squares and elastic net models perform, on average, as well as random forest and LightGBM models. But the latter two's transferability still outperforms generalized linear models developed for CS Unit 1. It is worth noting that models developed for CS Unit 2 have the worst transferability, and those developed for CS Unit 1 have the best transferability for both pollutants and all other algorithms.



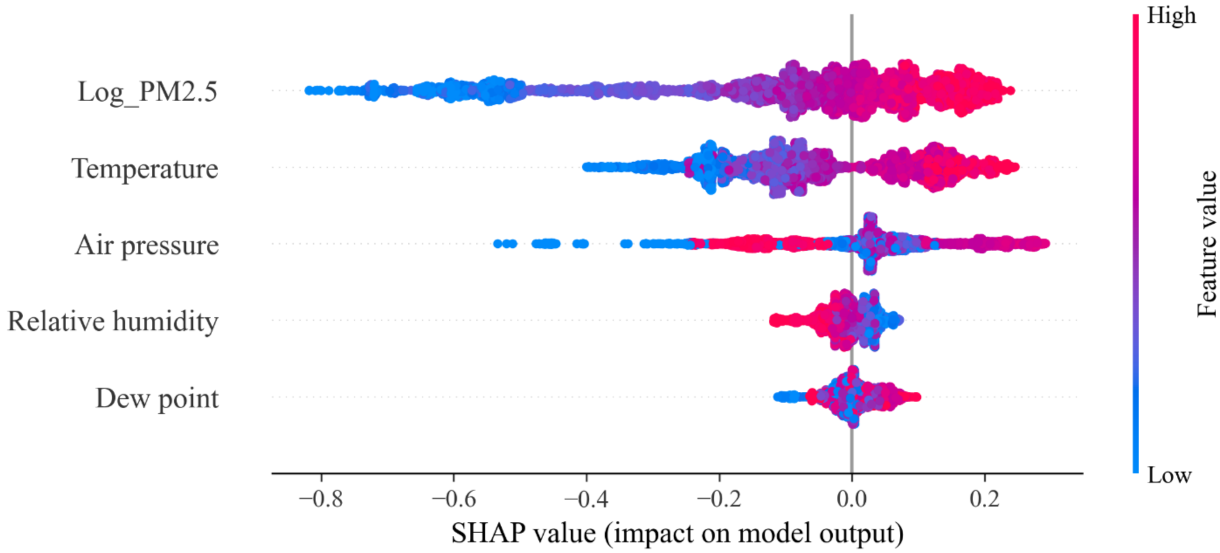
(a) $PM_{2.5}$ inter-sensor test



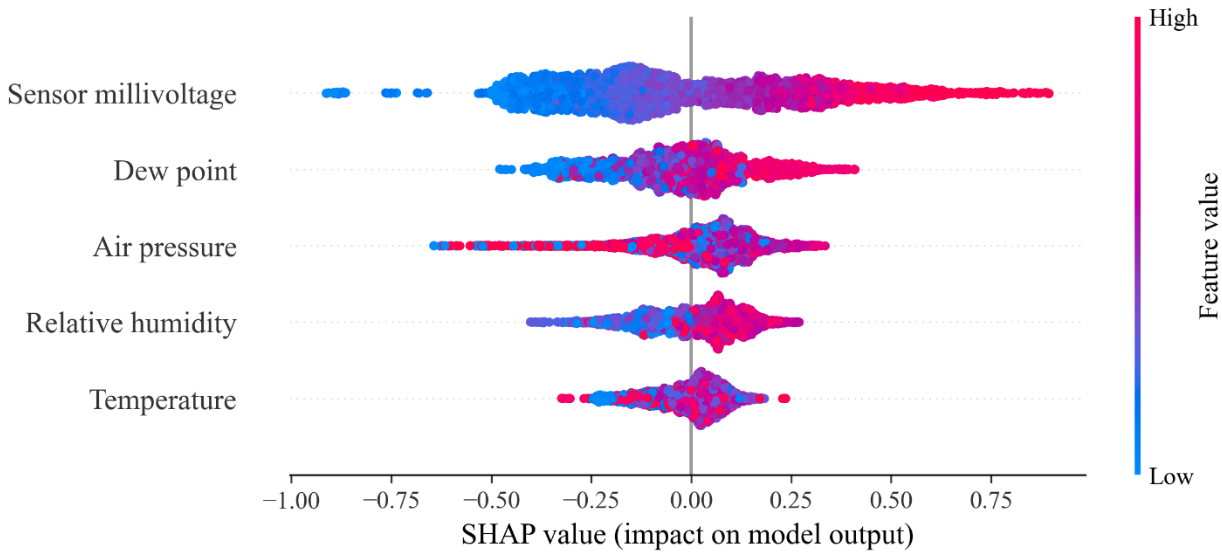
(b) NO_2 inter-sensor test

Figure 4. Inter-sensor variability tests of best-performing 1-min (a) $PM_{2.5}$ and (b) NO_2 models. Values in the tables are r^2 values calculated from passing sensor-specific data (y-axis) to sensor-specific models (x-axis); darker color represents higher values.

Finally, we examine the model behaviors in the stationary setting from the best-performing random forest models for $PM_{2.5}$ and NO_2. The results are visualized using SHAP summary plots, which consist of marginal contributions of all data points of each explanatory feature, represented by their SHAP values. Positive values denote positive effects, and higher absolute values denote larger effects on the output and vice versa. Data points are color-coded from blue to red, indicating low to high values. Data points of each feature are stacked vertically if different values have the same impact as in a violin plot. Features are ranked by the absolute average SHAP values of all observations from high to low, indicating high to low feature importance. In Figure 5, log-transformed $PM_{2.5}$ readings from CS units have the highest feature importance. Moreover, its impact on the predictions is monotonic and positive as the SHAP value (x-axis) increases as log $PM_{2.5}$ values increase with a smooth color change from blue to red (low to high feature values). A similar phenomenon is found in the NO_2 calibration model, even though we mapped the electro-signal directly to reference readings. It is desirable that low-cost sensor readings can reflect air quality changes. Other than CS unit readings, the temperature is the second most important that positively correlates with reference $PM_{2.5}$ readings, while in the NO_2 model, the dew point feature plays the same role. The effect of relative humidity is not significant in the $PM_{2.5}$ and NO_2 models, where it is positively correlated with NO_2 readings but negatively correlated with $PM_{2.5}$. This observation is opposite to previous literature (Crilley et al., 2018; di Antonio et al., 2018), which can be attributed to a difference in particle hygroscopicity from city to city, especially in the collocation region. We observe that the relationship between air pressure and model output is non-linear. It is worth noting that air pressure can be merely a good predictor rather than a good explainer for sensor readings. It is in line with our main purpose to develop high-accuracy calibration models. Dew point in the $PM_{2.5}$ model and temperature in the NO_2 model demonstrate similar behaviors, proving the necessity of using non-linear models.



(a) $PM_{2.5}$ best performing 1-min model

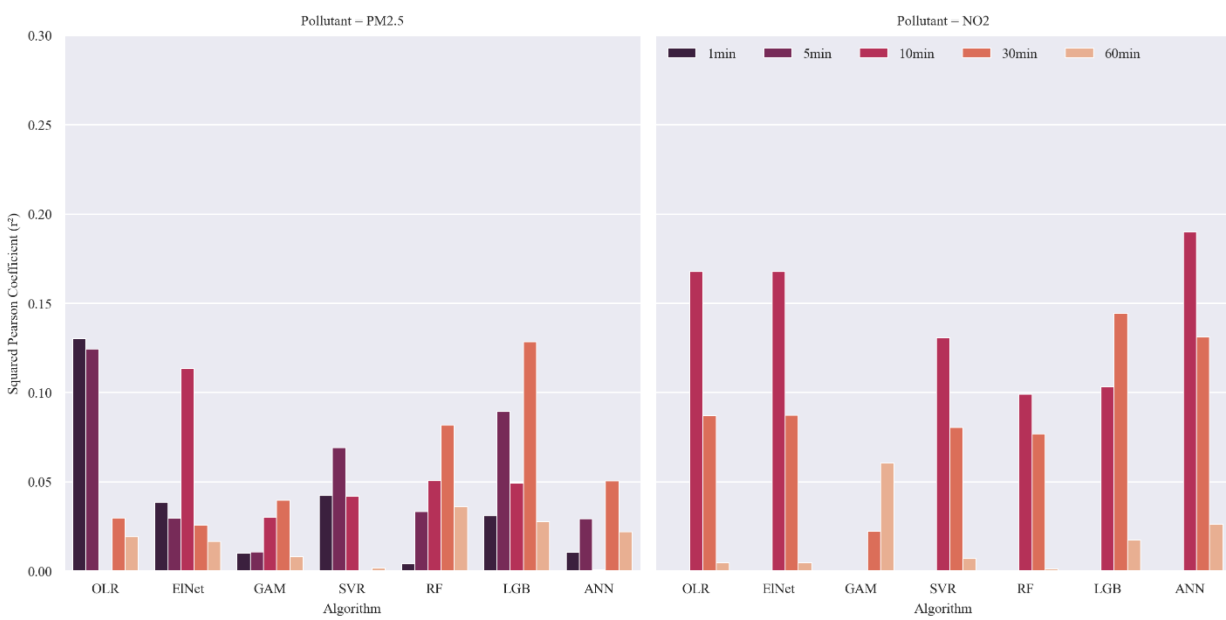


(b) NO_2 best performing 1-min model

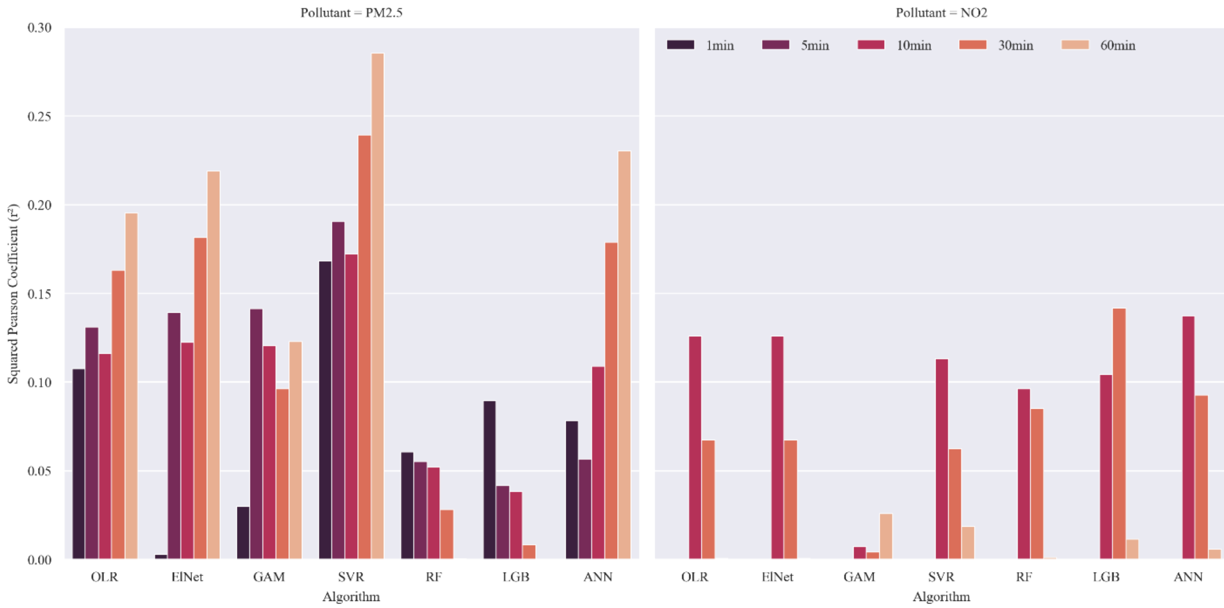
Figure 5. SHAP summary plots of best performing 1-min random forest models on a City Scanner unit for (a) $PM_{2.5}$ and (b) NO_2. Dots represent explanatory feature values and color-coded from low to high values in blue to red; data points with the same SHAP value are stacked vertically; all features are ranked from high to low feature importance on the y-axis.

3.3 Algorithm Transferability in a Mobile Setting

We further transfer the best-performing models trained with stationary collocation data in NYC to predict mobile validation observations in Boston. We observe in Figure 6 significantly deteriorated performance across all algorithms for both CS units used in Boston (Units 3 and 5). The r^2 values are low in the range of 0.1 to 0.3. Multiple possible reasons lead to unsatisfactory model transferability. On the one hand, mobile validation has brought extra challenges to sensor performance. In mobile validation, the sensors can be exposed to a variety of emission sources at different distances in a short amount of time. Such drastic changes in particle properties and pollutant concentrations are difficult to account for in stationary calibration. Also, the sensors being mobile validated were used for another mobile measurement campaign, which has led to accelerated drifting and aging due to vibration than promised in the manual. On the other hand, some problems exist ubiquitously for low-cost sensor calibration models in both stationary and mobile settings. First, the weather in NYC during stationary collocation and in Boston during the mobile validation are drastically different, especially the temperature and dew points, which play important roles in the $PM_{2.5}$ and NO_2 models, respectively. Some weather feature values are unseen in the training data, thus, hard to extrapolate when transferred to mobile data prediction. Second, while we use reference stationary data to train our model, the target data to predict during the Boston mobile validation are from research-grade instruments rather than reference-grade ones. Third, it is common that the performance of low-cost sensors degrades, and their behaviors change after long-time outdoor usage, given that the stationary collocation and mobile validation were conducted more than six months apart.



(a) City Scanner Unit 3



(b) City Scanner Unit 5

Figure 6. Model performances in a mobile setting for two City Scanner units and five data aggregations. Box colors illustrate 1 min to 60 min temporal aggregations from dark to light.

4 DISCUSSION

This study systematically evaluated the effects of temporal aggregation, calibration algorithms, and meteorological conditions on low-cost sensor calibration in both stationary and mobile settings.

We observed that linear regression models, including ordinary least squares and elastic net, perform consistently well in both $PM_{2.5}$ and NO_2 calibrations. Meanwhile, algorithms that consider non-linear relationships, including random forest and GAM, provide good prediction accuracy and better transferability between CS units than those that do not when calibrating $PM_{2.5}$. This phenomenon is not found in NO_2 calibration models, where linear regression transfers as well as machine learning models. It indicates that Alphasense NO2-A43F sensor provides consistent linear responses in most circumstances with little inter-sensor variability. It is recommended to employ algorithms that can account for non-linear interactions between meteorology and low-cost sensor behaviors, especially for PM calibration. For most educational and citizen science projects, generalized linear regressions are sufficient for low-cost sensor calibration as they are easy to apply and interpret. Robust results can be reached if we properly follow the procedures recommended in the EPA guidelines for field tests. We also observed that random forest and GAM models are highly transferable between $PM_{2.5}$ sensors, second by linear regression models. It supports adopting low-cost air sensors at a large scale to obtain big air quality data in urban areas.

By interpreting the best $PM_{2.5}$ and NO_2 model behavior using SHAP plots, we confirmed the non-linear relationships between meteorological features and model outputs, especially for PM calibration. Moreover, we developed all models using only the CS readings and four meteorological features without further feature engineering. It indicates that capturing non-linearity and feature interactions might be more effective in improving calibration models than incorporating and compositing more features.

We observed that calibration model performance deteriorates when transferred from stationary to mobile use. This is mainly caused by several reasons, some of which are ubiquitous to all low-cost sensors, while others are unique to mobile measurement. The ubiquitous ones include weather differences between stationary collocation and mobile validation periods, using research-grade instead of reference-grade monitors in the mobile validation, and performance degradation of the low-cost sensors, including drifting and aging. Drifting and aging are common among all types of low-cost sensors (Kim et al., 2018; Malings et al., 2019) and can happen after 2 to 6 months of usage (Van Zoest et al., 2019). Moreover, mobile deployment between the stationary collocation and mobile validation reported in this study has led to accelerated aging and drifting. It is suggested to calibrate low-cost sensors before and validate after each deployment in an environment as similar to the actual monitoring as possible if it lasts more than six months and is unsupervised, as in our case. Regular collocations and calibrations are recommended for low-cost sensors performed no more than six months apart. Our work also demonstrates the risk of deploying low-cost sensors in an environment different from the calibration environment. The lack of local reference instruments to perform calibration can constrain low-cost sensor applications in low-income regions and countries; alternatively, the viability of calibration (spatial and temporal differences in calibration and deployment locations) should be acknowledged in data analysis.

We acknowledge several limitations in our study. First, the $PM_{2.5}$ and NO_2 levels were low during the stationary collocation and mobile validation periods (< 20 µg/m^3 and < 50 ppb, respectively). Ideally, an as wide as possible air quality range should be included in the calibration process to account for the variability of air pollution in mobile on-road deployments. Kelly et al.

(2017) reported that some low-cost sensors begin to exhibit a non-linear response only when in high PM concentrations (>40 μg /m^3). Second, not all CS units are tested in the mobile comparison due to limited space and resources. Lastly, while it is more rigorous to calibrate low-cost sensors against Federal Reference Methods instruments, our study used data from Federal Equivalent Methods instruments as they are available minute-by-minute, which can better serve the purpose of our study.

5 CONCLUDING REMARKS

Our work presents two major takeaways regarding effective and robust low-cost calibration with a special focus on mobile air quality monitoring. We first demonstrated that PM calibration should consider the complex relationship between sensor responses, pollutant concentrations, and meteorological factors. Therefore, algorithms that consider non-linear relationships should be adopted in this case.

Models trained in the stationary setting can hardly be transferred to the mobile setting in a different urban environment and climate; additionally, sensors degrade after long-term outdoor mobile deployment. It is necessary to calibrate low-cost sensors in an environment that is similar to real-world deployment. It is recommended to calibrate or assess comparability both before and after mobile deployments of durations exceeding six months. Our recommendation exceeds existing EPA guidelines that are primarily focused on stationary monitoring.

The findings in our study are important to citizen scientists, air quality researchers, and practitioners interested in advancing low-cost sensor applications, such as raising awareness and community engagement in scientific analysis. Further research is needed on other commonly used low-cost particulate matter and gas sensors and low-cost sensor performance in temperature, humidity, and concentration range not included in this study. Finally, there is a need to assess low-cost sensor performance in mobile settings given the rapidly increasing use of and demands for mobile environmental sensing.

REFERENCE

Bezantakos, S., Schmidt-Ott, F., & Biskos, G. (2018). Performance evaluation of the cost-effective and lightweight Alphasense optical particle counter for use onboard unmanned aerial vehicles. *Aerosol Science and Technology*, *52*(4), 385–392. https://doi.org/10.1080/02786826.2017.1412394/SUPPL_FILE/UAST_A_1412394_SM0236.ZIP

Brauer, M., Guttikunda, S. K., K A, N., Dey, S., Tripathi, S. N., Weagle, C., & Martin, R. v. (2019). Examination of monitoring approaches for ambient air pollution: A case study for India. *Atmospheric Environment*, *216*, 116940. https://doi.org/10.1016/J.ATMOSENV.2019.116940

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? *Environment International*, *99*, 293–302.

Crawford, B., Hagan, D. H., Grossman, I., Cole, E., Holland, L., Heald, C. L., & Kroll, J. H. (2021). Mapping pollution exposure and chemistry during an extreme air quality event (the 2018 Kīlauea eruption) using a low-cost sensor network. *Proceedings of the National Academy of Sciences*, *118*(27). https://doi.org/10.1073/PNAS.2025540118

Crilley, L. R., Shaw, M., Pound, R., Kramer, L. J., Price, R., Young, S., Lewis, A. C., & Pope, F. D. (2018). Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring. *Atmospheric Measurement Techniques*, *11*(2), 709–720. https://doi.org/10.5194/AMT-11-709-2018

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. *Atmospheric Measurement Techniques*, *13*(3), 1181–1193. https://doi.org/10.5194/AMT-13-1181-2020

deSouza, P., Kahn, R. A., Limbacher, J. A., Marais, E. A., Duarte, F., & Ratti, C. (2020). Combining low-cost, surface-based aerosol monitors with size-resolved satellite data for air quality applications. *Atmospheric Measurement Techniques*, *13*(10), 5319–5334. https://doi.org/10.5194/AMT-13-5319-2020

Duarte, F., & deSouza, P. (2020). Data Science and Cities: A Critical Approach. *Harvard Data Science Review*, *2*(3). https://doi.org/10.1162/99608F92.B3FC5CC8

Gressent, A., Malherbe, L., Colette, A., Rollin, H., & Scimia, R. (2020). Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. *Environment International*, *143*, 105965. https://doi.org/10.1016/J.ENVINT.2020.105965

Hua, J., Zhang, Y., Foy, B. de, Mei, X., Shang, J., Zhang, Y., Sulaymon, I. D., & Zhou, D. (2021). Improved PM2.5 concentration estimates from low-cost sensors using calibration models categorized by relative humidity. *Https://Doi.Org/10.1080/02786826.2021.1873911*, *55*(5), 600–613. https://doi.org/10.1080/02786826.2021.1873911

Hudda, N., Simon, M. C., Patton, A. P., & Durant, J. L. (2020). Reductions in traffic-related black carbon and ultrafine particle number concentrations in an urban neighborhood during the COVID-19 pandemic. *Science of The Total Environment*, *742*, 140931. https://doi.org/10.1016/J.SCITOTENV.2020.140931

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kelly, K. E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R., & Butterfield, A. (2017). Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environmental Pollution*, *221*, 491–500. https://doi.org/10.1016/J.ENVPOL.2016.12.039

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., & Vermeulen, R. C. H. (2019). Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. *Environmental Science & Technology*, *53*(3), 1413–1421. https://doi.org/10.1021/acs.est.8b06038

Liang, L. (2021). Calibrating low-cost sensors for ambient air monitoring: Techniques, trends, and challenges. *Environmental Research*, *197*, 111163. https://doi.org/10.1016/J.ENVRES.2021.111163

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent Individualized Feature Attribution for Tree Ensembles*.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. *Section 2*, 1–10.

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. *Atmospheric Measurement Techniques*, *12*(2), 903–920. https://doi.org/10.5194/AMT-12-903-2019

Miskell, G., Salmond, J. A., & Williams, D. E. (2018). Use of a handheld low-cost sensor to explore the effect of urban design features on local-scale spatial and temporal air quality variability. *Science of The Total Environment*, *619–620*, 480–490. https://doi.org/10.1016/J.SCITOTENV.2017.11.024

Padró-Martínez, L. T., Patton, A. P., Trull, J. B., Zamore, W., Brugge, D., & Durant, J. L. (2012). Mobile monitoring of particle number concentration and other traffic-related air pollutants in a near-highway neighborhood over the course of a year. *Atmospheric Environment*, *61*, 253–264. https://doi.org/10.1016/J.ATMOSENV.2012.06.088

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

SM, S. N., Reddy Yasa, P., MV, N., Khadirnaikar, S., & Pooja Rani. (2019). Mobile monitoring of air pollution using low cost sensors to visualize spatio-temporal variation of pollutants at urban hotspots. *Sustainable Cities and Society*, *44*, 520–535.

Sousan, S., Koehler, K., Hallett, L., & Peters, T. M. (2016). Evaluation of the Alphasense optical particle counter (OPC-N2) and the Grimm portable aerosol spectrometer (PAS-1.108). *Aerosol Science and Technology*, *50*(12), 1352–1365. https://doi.org/10.1080/02786826.2016.1232859/SUPPL_FILE/UAST_A_1232859_SM0343.ZIP

US EPA. (2014). *Air Sensor Guidebook.* https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NERL&dirEntryId=277996

US EPA. (2021a). *Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors: Use in Ambient, Outdoor, Fixed Sites, Non-Regulatory Supplemental and Informational Monitoring Applications*. https://doi.org/EPA/600/R-20/280

US EPA. (2021b). *Performance Testing Protocols, Metrics, and Target Values for Ozone Air Sensors: Use in Ambient, Outdoor, Fixed Site, Non-Regulatory and Informational Monitoring Applications*. https://doi.org/EPA/600/R-20/279

Wang, S., Ma, Y., Wang, Z., Wang, L., Chi, X., Ding, A., Yao, M., Li, Y., Li, Q., Wu, M., Zhang, L., Xiao, Y., & Zhang, Y. (2021). Mobile monitoring of urban air quality at high spatial resolution by low-cost sensors: Impacts of COVID-19 pandemic lockdown. *Atmospheric Chemistry and Physics*, *21*(9), 7199–7215. https://doi.org/10.5194/acp-2020-1169

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. *Atmospheric Measurement Techniques*, *11*(8), 4823–4846. https://doi.org/10.5194/AMT-11-4823-2018