Atmospheric Environment 242 (2020) 117761

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: http://www.elsevier.com/locate/atmosenv

Statistical field calibration of a low-cost $PM_{2.5}$ monitoring network in Baltimore

Abhirup Datta [a,*], Arkajyoti Saha [a], Misti Levy Zamora [b,c], Colby Buehler [c,d], Lei Hao [b], Fulizi Xiong [c,d], Drew R. Gentner [c,d], Kirsten Koehler [b,c]

[a] Department of Biostatistics, Johns Hopkins University, Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD, 21205, USA
[b] Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD, 21205, USA
[c] SEARCH (Solutions for Energy, Air, Climate and Health) Center, Yale University, New Haven, CT, USA
[d] Department of Chemical & Environmental Engineering, Yale University, School of Engineering and Applied Science, New Haven, CT, 06511, USA

HIGHLIGHTS

- 32 low-cost monitors were operating in Baltimore which has 1 regulatory PM2.5 site.
- Colocation at the regulatory site showed that raw low-cost data overestimated PM2.5
- Multiple linear regression was used for field-calibration using co-located data.
- Field calibration (24-hr avg. RMSE of 2 μg/m^3) outperformed laboratory correction.
- Calibrated data revealed spatiotemporal differences in PM2.5 across Baltimore.

ARTICLE INFO

Keywords:
Baltimore
Field colocation
Gain-offset model
Linear regression
Low-cost monitors
$PM_{2.5}$

ABSTRACT

Low-cost air pollution monitors are increasingly being deployed to enrich knowledge about ambient air-pollution at high spatial and temporal resolutions. However, unlike regulatory-grade (FEM or FRM) instruments, universal quality standards for low-cost sensors are yet to be established and their data quality varies widely. This mandates thorough evaluation and calibration before any responsible use of such data. This study presents evaluation and field-calibration of the $PM_{2.5}$ data from a network of low-cost monitors currently operating in Baltimore, MD, which has only one regulatory $PM_{2.5}$ monitoring site within city limits. Co-location analysis at this regulatory site in Oldtown, Baltimore revealed high variability and significant overestimation of $PM_{2.5}$ levels by the raw data from these monitors. Universal laboratory corrections reduced the bias in the data, but only partially mitigated the high variability. Eight months of field co-location data at Oldtown were used to develop a gain-offset calibration model, recast as a multiple linear regression. The statistical model offered substantial improvement in prediction quality over the raw or lab-corrected data. The results were robust to the choice of the low-cost monitor used for field-calibration, as well as to different seasonal choices of training period. The raw, lab-corrected and statistically-calibrated data were evaluated for a period of two months following the training period. The statistical model had the highest agreement with the reference data, producing a 24-h average root-mean-square-error (RMSE) of around 2 $\mu g\, m^{-3}$. To assess transferability of the calibration equations to other monitors in the network, a cross-site evaluation was conducted at a second co-location site in suburban Essex, MD. The statistically calibrated data once again produced the lowest RMSE. The calibrated $PM_{2.5}$ readings from the monitors in the low-cost network provided insights into the intra-urban spatiotemporal variations of $PM_{2.5}$ in Baltimore.

* Corresponding author.
E-mail address: abhidatta@jhu.edu (A. Datta).

https://doi.org/10.1016/j.atmosenv.2020.117761
Received 2 March 2020; Received in revised form 12 June 2020; Accepted 6 July 2020
Available online 22 July 2020

1. Introduction

Evidence of the harmful effects of air pollution on human health and morbidity is now overwhelming. Recent studies have attributed around 7–9 million annual global deaths to ambient air pollution (Lelieveld et al., 2019; World Health Organization, 2016), and approximately 88,000 annual deaths in the United States (Cohen et al., 2017). Fine particulate matter ($PM_{2.5}$) is now an established carcinogen and is linked to increased morbidity and mortality (Hoek et al., 2013; Loomis et al., 2013). Studies focused on the United States also consistently associate $PM_{2.5}$ with increased incidence of cardiovascular diseases (Dominici et al., 2006; Powell et al., 2015). Attributable burdens of $PM_{2.5}$ on preterm births (Trasande et al., 2016) and mortality have been quantified (Dominici et al., 2007; Fann et al., 2012). Even exposure to air pollutants at concentrations below the current air quality standards in the United States has been shown to elevate mortality rates (Di et al., 2017).

In the United States, $PM_{2.5}$ levels are measured at designated sites with high precision monitors. These monitors adhere to the Environmental Protection Agency (EPA)'s Federal Reference Method (FRM) or Federal Equivalent Method (FEM) standards and their data are considered as gold standard. The data from this network of regulatory monitors are used to assess compliance with National Ambient Air Quality Standards (NAAQS) and in studies assessing impact of air pollution on health (Samet et al., 2000). The scope of spatial studies using this network of high-quality monitoring sites is limited owing to their sparse geographical coverage. Most urban centers in the United States have less than 5 regulatory monitors (Apte et al., 2017) for any particular pollutant. The spatial resolution offered by gold standard data fails to discriminate between exposure levels across communities proximal to the same monitor. This results in exposure misclassification. Chemistry transport models like CMAQ (Byun and Schere, 2006) offer improved spatial resolution. These projections often have crude temporal resolution prohibiting insight into the short-term fluctuations of pollutant levels.

Recent years have witnessed a surge in usage of diverse alternative technologies to measure air pollution at spatial- and temporal-resolutions higher than what is offered by the regulatory monitors. Low-cost monitors have been deployed in spatially-dense stationary networks (Gao et al., 2015), car-mounted mobile networks (Apte et al., 2017; Hasenfratz et al., 2015; Lim et al., 2019), and as wearable devices for personal monitoring (Cai et al., 2014). Other technologies like silicone wristbands (O'Connell et al., 2014), etc., are also being explored. Data from novel low-cost devices enable block-level and high-frequency source apportionment studies (Shah et al., 2018), neighborhood-level association analysis of exposure and health (Hajat et al., 2013), and can directly measure activity- or source-specific personal exposures (Dons et al., 2017). Data from a spatially dense network of low-cost monitors may provide a better surrogate for individual-level ambient exposure in cohort studies (Szpiro et al., 2010). This an important application of low-cost monitors as regional monitors often do not reflect personal exposure levels (Levy Zamora et al., 2018). The hyper-local, high-frequency, and individual-level characterizations of exposure from low-cost monitoring networks cannot be achieved by the sparse network of regulatory monitors.

Data quality of low-cost sensors vary widely as they are prone to measurement errors and missing data. The out-of-sample R^2 of low-cost monitoring data (against reference measurements) can range from less than 1% to over 75% depending on the type of monitor (United States Environmental Protection Agency n.d.). The review article (Morawska et al., 2018) contains assessment of 17 low-cost monitoring studies initiated between 2010 and 2017. It highlights lack of consistent information about accuracy and precision of these monitors as well lack of consensus about required performance standards. This is in contrast to FRM or FEM instruments which have well established accuracy and precision. Until such universal standards are established for low-cost monitors, thorough validation and calibration exercises are mandated for any study using low-cost monitors.

The Solutions to Energy, Air, Climate, and Health (SEARCH) Center, an EPA-funded "Air in a Changing Environment" Center, is currently conducting a study using up to 45 multi-pollutant low-cost stationary monitors in Baltimore city. There is only one location within the city limits with a FEM $PM_{2.5}$ monitor. Hence information on intra-urban variations of $PM_{2.5}$ in Baltimore is lacking. Contingent upon successful validation and calibration of its data, the SEARCH low-cost network, promises insight into the $PM_{2.5}$ variations in the city at an unprecedented spatial and temporal resolution.

Most low-cost sensors for $PM_{2.5}$, including the sensors used in the SEARCH project, are optical sensors. They estimate concentrations using light-scattering principles and their performance is known to be biased by meteorological variables like relative humidity (RH). Optical sensors overestimate hygroscopic particles, and consequently overestimate the pollutant concentration. To a lesser extent, temperature and pressure also affect these sensors (Feenstra et al., 2019). Such biases can be studied in lab-settings, and correction equations are often determined based on the experimental data from this preliminary lab-testing phase (Levy Zamora et al., 2019). Many low-cost units contain in-built RH and temperature sensors. So, these corrections can be subsequently applied to calibrate the raw data from each monitor in the field. However, the meteorological conditions used in lab-experiments may not cover the wide range of outdoor conditions that is encountered when monitors are eventually deployed (Castell et al., 2017; Piedrahita et al., 2014). Lab-experiments also struggle to create test-scenarios like long periods of low ambient concentrations, as is often the case in outdoor deployment (Morawska et al., 2018). Long-term temporal drifts are often observed in low-cost sensors (Kelleher et al., 2018). Short- or medium-term lab-evaluations cannot inform about such drifts.

Field calibration of low-cost monitors using co-located or proximally-located high-precision regulatory monitors is commonly used to supplement or replace lab-correction (Carvlin et al., 2017; Topalović et al., 2019; Zimmerman et al., 2018). One method is to use simple correction factors to reduce bias in low-cost monitor data (Apte et al., 2017; Van den Bossche et al., 2015; Clougherty et al., 2013; Dons et al., 2012; Larson et al., 2009; Lim et al., 2019). These factors are derived based on the correlation between the low-cost monitoring data and one or many nearby regulatory monitors. This approach is pragmatic for mobile monitoring networks where exact co-location for a prolonged period is not possible. However, correlation between data from the regulatory monitor and a low-cost monitor is expected to depend on the location of the latter, as PM2.5 sources and compositions vary with location. Use of a single correction factor for all low-cost monitors in the network disregards this variation which can introduce bias in the calibration. Another popular approach is the gain-offset model (Balzano and Nowak, 2007) which quantifies the deviation between the true pollutant level and the measured level by the low-cost monitors in terms of an additive bias and a multiplicative bias. More generally, gain-offset models are a subclass of linear regression models that are popularly used to calibrate low-cost monitoring data using co-located data (Deffner et al., 2016). As the linearity assumption is often not adequate, non-linear approaches like random forests (Zimmerman et al., 2018), neural networks (Topalović et al., 2019), stacking (Lim et al., 2019), gradient boosting regression trees (Johnson et al., 2018a, 2018b), support vector regression (De Vito et al., 2018), etc. Are increasingly being used.

Field calibration efforts, whenever possible, use multiple reference sensors located at diverse ambient conditions for colocation (Steinle et al., 2015; Wang et al., 2019). This strategy cannot be adopted in Baltimore as there is only one location for reference $PM_{2.5}$ data within the city limits. Co-locating all units periodically with this reference monitor to calibrate for monitor-specific bias is also not pragmatic. This would have involved extensive manual efforts and substantially delayed deployment of the units at their designated locations. Instead, two of the

SEARCH low-cost monitors are co-located with the FEM $PM_{2.5}$ monitor at Oldtown, Baltimore continually throughout the duration of the project. A parsimonious multiple linear regression-based gain-offset calibration model was estimated using this co-located data. In a gain-offset model the parameters are sometimes estimated by assuming some similarity in the moments (mean and variance) of the reference data and the calibrated data (Miskell et al., 2018). This strategy only works when the gain and offset are assumed to be constants. A more general approach was adopted here that allows gains and offsets to vary with other covariates (meteorological variables, seasonality, etc.). The gain-offset was framed as a multiple linear regression problem ensuring that all the parameters can be estimated using standard least squares approach. Recasting gain-offset model as linear regression allowed seamless comparison of different models for the gain and the offset using standard comparison metrics like AIC and BIC. This helped to select explanatory variables for the calibration. An extensive set of evaluation studies were conducted. The statistical calibration considerably outperformed both the raw data and the lab-corrected data from the monitors. The model calibrated data from the low-cost network offered intriguing insights into the spatiotemporal variations of $PM_{2.5}$ in Baltimore city.

2. Data and methods

2.1. Regulatory $PM_{2.5}$ data in Baltimore

The Maryland Department of the Environment (MDE) measures $PM_{2.5}$ levels at 12 locations in Maryland as part of the EPA's State or Local Air Monitoring Stations (SLAMS) Network. Among these 12 sites, only the location at Oldtown lies within the Baltimore city limits. The Oldtown site is in an urban setting, located in the city center adjacent to a major traffic intersection with high traffic volumes. An FEM Beta Attenuation Monitor (BAM) hosted at the site measures hourly $PM_{2.5}$. Besides Oldtown, the suburban Essex site is the nearest source of regulatory $PM_{2.5}$ data, measuring daily average $PM_{2.5}$ concentration every 6 days using a manual gravimetric FRM. The Essex site is located around 8 miles away from the Oldtown site. General information about the two sites is provided in Table S1 of the Supplement. All this information is publicly available ("Maryland Department of the Environment" 2020) along with the corresponding $PM_{2.5}$ data.

2.2. SEARCH low-cost monitoring network

Each multi-pollutant monitor in SEARCH includes sensors for $PM_{2.5}$, PM_{10} (coarse particulate matter), PM_1 (particles with size less than 1 μm), ozone, nitrous oxide, nitrogen dioxide, carbon monoxide, carbon dioxide, methane, temperature (T), and relative humidity (RH). The $PM_{2.5}$ sensor is a Plantower A003. When deployed outdoors, the monitors are encased inside a protective shell to guard against weather hazards. The monitors are equipped with a SIM card and antenna enabling wireless transmission of data to a remote server. 10 s averages of the recorded data are transmitted to the wireless server every 10 s. Raw data are also stored locally several times a second in an SD card inside each monitor. The SD card data are periodically collected and transferred to a data repository to supplement the server data. During these periodic visits, faulty individual sensors identified by frequent manual monitoring of the online data are also removed and replaced. The multipollutant monitors have been deployed in batches since December 2018.32 monitors were active during the study period and the network is expected to expand to 45 monitors. The locations were determined by weighted random sampling. The weights were based on distance from major roadways, population density, presence of industry and other point sources including airports and power plants, etc. Final locations were determined by the availability of an individual or an entity consenting to hosting a low-cost monitor. Due to the wide variety of properties hosting the monitors and the constraint of requiring a power source, the monitors are installed at different heights, but most are within 3 m of the ground. Two of these monitors, henceforth referred to as Ot1 and Ot2, have been co-located with the reference monitor at Oldtown since December 20, 2018. Another two monitors (Ex1 and Ex2) have been co-located with the reference monitor at Essex since January 31, 2019.

The raw data from SEARCH were obtained both continually from the remote server as well as periodically from the SD cards inside each monitor. The SD card data were used to recover any missing data due to cellular network or server issues. Subsequently, the two sources of data were integrated and aggregated to produce hourly averages. There was a drift diagnosed in the time-stamps associated with the SD card readings of about 40 min per day if cellular connection was not available. These drifts were corrected by matching the diurnal maxima and minima of the T and RH time-series from the monitors with the analogous data available from the weather station from the nearest MDE site.

2.3. Lab-based RH/T correction factors

Prior to deployment of the first monitors, an extensive set of lab-experiments were conducted to assess the quality of the raw $PM_{2.5}$ data from the low-cost monitors. The experiment settings were similar to those outlined in Levy Zamora et al. (2019). 8 monitors were used to assess how the bias in the raw data from these monitors varied in different RH and T conditions. The sensors were placed inside a custom-built steel chamber (0.71 m × 1.35 m x 0.89 m), equipped with a filtered air inlet, vacuum exhaust, two internal fans, and three sampling ports. A personal DataRAM™ pDR-1200 (Thermo Scientific Corp., Waltham, Mass.) with a single-stage $PM_{2.5}$ impactor along with an external pump (BGI 400, Mesa Labs, Inc.) was also set up in the chamber. A 37-mm Teflon filter was used to collect all particles sampled by the pDR for subsequent analysis and gravimetric correction (Pall Corporation, Ann Arbor, MI). The time-resolved pDR $PM_{2.5}$ mass concentrations were gravimetrically corrected for known RH and T biases. The corrected values were then compared to the raw $PM_{2.5}$ sensor data. The monitors were exposed to RH ranging between about 5 and 85%, temperatures ranging between about 25 and 40 °C, and $PM_{2.5}$ mass concentrations between 0 and 1000 $\mu g/m^3$. All of the variables were changed independently. The experiments yielded the following correction equations for monitor u:

$$CF_u = \frac{PM}{0.00025599(RH_u)^2 - 0.002648(RH_u) + 0.88732}$$

$$\mathrm{PM} = \frac{CF_u}{0.00020883 \cdot T_u^2 - 0.012708 \cdot T_u + 1.1235}$$

where temperature (T) and relative humidity (RH) is measured inside the monitor (not ambient).

2.4. Regression models

A generic gain-offset model (Balzano and Nowak, 2007) to calibrate the low-cost sensors is given by:

$$PM_{2.5\ ref,u}(t) = a_0 + b_0 PM_{2.5\ u}(t) + e(t), \quad (1)$$

where u denotes the low-cost monitor ID, $PM_{2.5\ u}(t) = PM_{2.5\ RH/T\ corrected\ u}(t)$ is the lab-corrected $PM_{2.5}$ recorded by monitor u at time t, $PM_{2.5\ ref,u}(t) = PM_{2.5\ MDE\ u}(t)$ is the ambient $PM_{2.5}$ concentration at the location of monitor u measured by a reference instrument, and $e(t)$ is the random error. The parameters a_0 and b_0 respectively denote the gain (additive bias) and offset (multiplicative bias). Model (1) assumes both the gain and offset are constants, i.e., homogeneity of both biases across time which might be inappropriate if there are temporal trends or seasonality in the bias. Additionally, for this study, reference data were available at only one location in the city ruling out estimating monitor-

specific gain and offset parameters. This would have required periodic co-location of every device in the network with the reference monitor. Hence, it was essential to consider monitor-and-time-specific covariates including meteorological variables recorded by the monitors, time of the day, seasonality, etc.

For this analysis, the gain-offset model was generalized to model the gains and offsets as linear functions of the set of covariates $x_u(t)$, i.e.,

$$PM_{2.5\ ref,u}(t) = a^T x_u(t) + b^T x_u(t) PM_{2.5\ u}(t) + e(t) \quad (2)$$

For model (1), the constant gain and offset parameters can be estimated by matching moments of the two-sets of co-located time-series $PM_{2.5\ ref,u}(t)$ and $PM_{2.5\ u}(t)$ (Miskell et al., 2018). This estimation strategy does not work for model (2). With covariates, matching of higher moments will be required whose closed form expressions are difficult to derive. A much simpler solution for parameter estimation is to recast (2) as a multiple linear regression model

$$PM_{2.5\ ref,u}(t) = z_u(t)^T \beta + e(t)$$

where

$$z_u(t) = \left(x_u(t)^T, x_u(t)^T PM_{2.5\ u}(t)\right)^T$$

and

$$\beta = \left(a^T, b^T\right)^T.$$

The model can thus be fitted using simple least squares. This also makes it feasible to use standard model comparison measures like AIC or BIC for variable selection.

Four models of increasing complexity, i.e., increasing number of covariates for the gains and offsets were considered. The models are presented in Table 1. In model 3, *daytime* is a binary variable which is 1 between 5AM and 7PM (approximating the daytime hours) and *weekend* is a binary variable which is 1 on Saturdays and Sundays. In model 4, *hod* is a 23-level categorical variable corresponding to hour of the day from 1 to 23 (hour 0 is the baseline), and *dow* is a 6-level categorical for each day of the week from Tuesday to Sunday (Monday being the baseline). The models were considered based on the exploratory analysis outlined in Section 3.3.

The final model was chosen based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). For out-of-sample evaluation, 10% of the data from the training period (Dec 20, 2018 to July 31, 2019) were randomly chosen and held-out during estimation. The held-out data were used to compare the performance of the statistical model with the raw and lab-corrected readings. For the comparison, root-mean-square-error (RMSE) and mean-absolute-error (MAE) were used for evaluating point predictions.

In addition to the point predictions, reporting error bounds or confidence intervals around the predicted value is essential to quantify the uncertainty of these predictions. Little attention has been paid in the literature on prediction inference (reporting prediction error bounds and evaluating them) of low-cost monitors. To do this, mean Coverage Probability (CP) and mean Confidence Interval Width (CIW) were used for evaluating the 95% prediction intervals. More details about these metrics and other aspects of the methodology are presented in Section S1 of the Supplement.

Table 1
Four different gain-offset models considered for statistical calibration of the low-cost $PM_{2.5}$ data (*Model 3 was the selected model based on AIC and BIC and was used in the remainder of the study).

Model 1	$a = a_0,\ b = b_0$
Model 2	$a = a_0 + a_{RH}{*}RH_u + a_T{*}T_u,\quad b = b_0 + b_{RH}{*}RH_u + b_T{*}T_u$
Model 3*	$a = a_0 + a_{RH}{*}RH_u + a_T{*}T_u + a_{daytime}{*}daytime + a_{weekend}{*}weekend$ $b = b_0 + b_{RH}{*}RH_u + b_T{*}T_u + b_{daytime}{*}daytime + b_{weekend}{*}weekend$
Model 4	$a = a_0 + a_{RH}{*}RH_u + a_{Temp}{*}T_u + a^T_{hod}hod + a^T_{dow}dow$ $b = b_0 + b_{RH}{*}RH_u + a_{Temp}{*}T_u + b^T_{hod}hod + b^T_{dow}dow$

3. Results

3.1. Summary statistics of raw and lab-corrected data

In Table 2, the summary statistics for the different variables are presented for the training period of December 20, 2018 to July 31, 2019. The variables are the reference $PM_{2.5}$ data at Oldtown ($PM_{2.5\ MDE}$), the raw $PM_{2.5}$ ($PM_{2.5\ raw}$), relative humidity (RH) and temperature (T) data from the co-located monitors Ot1 and Ot2, and the lab-corrected data ($PM_{2.5\ RH/T\ corrected}$). Monitor Ot1 did not produce accurate RH readings during part of the measurement period and those were replaced by contemporaneous RH readings from Monitor Ot2.

The mean $PM_{2.5}$ level recorded by the reference monitor was 8 $\mu g\, m^{-3}$, whereas the mean of the raw $PM_{2.5}$ data from monitors Ot1 and Ot2 were respectively 12.3$\mu g\, m^{-3}$ and 13.6 $\mu g\, m^{-3}$. This substantial overestimation by the raw data is a known issue reported in Levy Zamora et al. (2019). After the lab-based RH/T correction, the mean $PM_{2.5}$ level for these two monitors are 8.8 $\mu g\, m^{-3}$ and 9.7 $\mu g\, m^{-3}$, respectively, indicating that the correction considerably diminishes the bias. The standard deviations and the inter-quartile ranges indicate the large variability in the readings of the low-cost sensors. Compared to the standard deviation of 6 $\mu g\, m^{-3}$ for the reference data, the standard deviations for the RH/T corrected $PM_{2.5}$ readings from these two monitors were 8.6 $\mu g\, m^{-3}$ and 8.9 $\mu g\, m^{-3}$ respectively (the raw $PM_{2.5}$ standard deviations were 13.1 $\mu g\, m^{-3}$ and 13.5 $\mu g\, m^{-3}$ respectively). The summary statistics for all the variables were similar for the two monitors.

3.2. Exploratory analysis

The time series of hourly raw and lab-corrected $PM_{2.5}$ data from monitors Ot1 and Ot2 are plotted in Fig. 1 along with the reference $PM_{2.5}$ data. The plot is for the first month of data from Dec 20, 2018 to Jan 20, 2019. The raw readings consistently overestimated the true $PM_{2.5}$ levels. The RH/T correction reduced the magnitude of overestimation. The full time series of daily (24-hr average concentrations) for the Dec 2018 to July 2019 window is provided in Fig. S1 of the Supplementary materials. It shows that this overestimation is prevalent throughout the period of the study.

The biases of the corrected readings ($PM_{2.5\ RH/T\ corrected} - PM_{2.5\ MDE}$) were analyzed for periodicity. Mean biases for each hour of the day and for each day of the week are presented in Fig. 2. Biases were generally higher during the daytime than at night. The hours from 8pm to 4am generally had lower biases than the hours from 5am to 7pm. The

Table 2
Mean, standard deviation and inter-quartile range for $PM_{2.5}$ recorded by the reference monitor at Oldtown and for $PM_{2.5}$, RH and T data recorded by the two monitors Ot1 and Ot2 co-located at Oldtown for the period of Dec 20, 2018 to July 31, 2019.

Variable	Mean (Standard Deviation)		Inter-quartile range	
	Monitor Ot1	Monitor Ot2	Monitor Ot1	Monitor Ot2
T (degree Celsius)	20.6 (11.1)	20.1 (11.6)	(11.7–29.3)	(10.4–29.3)
RH (%)	–	46.2 (15.2)	–	(33.7–58.5)
$PM_{2.5\ raw}$ (μgm^{-3})	12.3 (13.1)	13.6 (13.5)	(3.0–17.6)	(3.8–19.7)
$PM_{2.5\ RH/T\ corrected}$ (μgm^{-3})	8.8 (8.6)	9.7 (8.9)	(2.3–12.9)	(3.0–14.4)
$PM_{2.5\ MDE}$ (μgm^{-3}) (reference)	8.0 (6.0)		(4.0–11.0)	

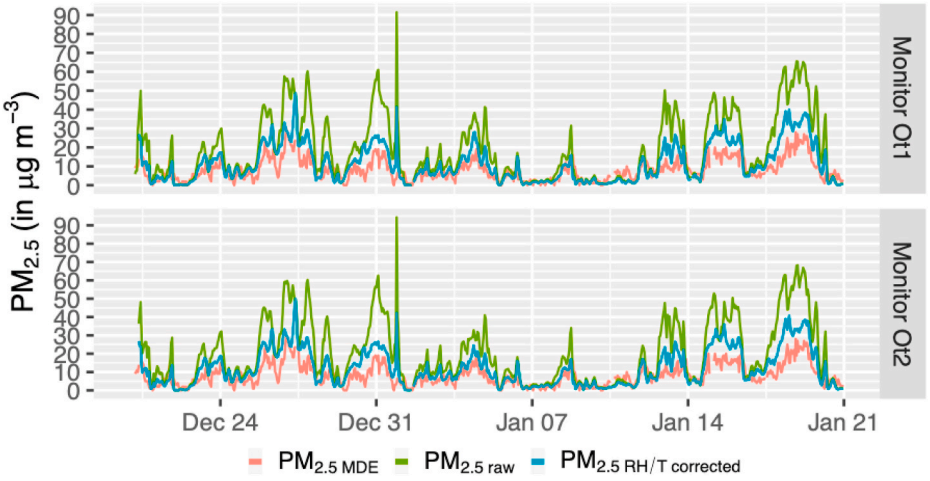

Fig. 1. Hourly time series of reference $PM_{2.5}$ (in $\mu g m^{-3}$) readings ($PM_{2.5\ MDE}$) at Oldtown along with the raw ($PM_{2.5\ raw}$) and lab-corrected ($PM_{2.5\ RH/T\ corrected}$) readings from the two co-located monitors (Ot1 and Ot2) from Dec 20, 2018 to Jan 20, 2019.

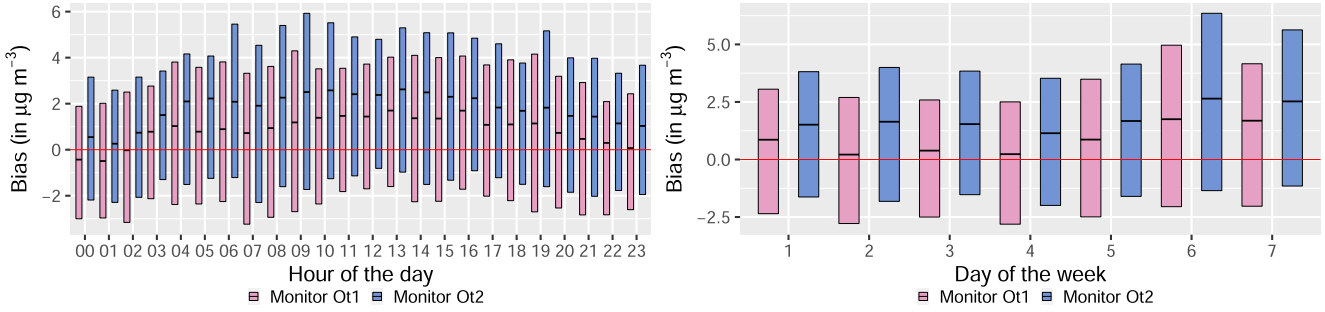

Fig. 2. Intra-quartile ranges and means of the biases ($PM_{2.5\ RH/T\ corrected}$ - $PM_{2.5\ MDE}$) in $\mu g/m^{-3}$ for each hour of the day (left) and each day of the week (right).

differences in mean biases between the two periods was approximately 1.5 $\mu g\, m^{-3}$. Higher mean biases were observed on weekends. The difference between the mean bias on Saturdays and on Thursdays is approximately 1.5 $\mu g\, m^{-3}$. These findings were consistent among the two monitors although the magnitudes of the biases were different. The diurnal and weekday-weekend differences in bias might have been caused by temporal variation of the $PM_{2.5}$ composition (Levy Zamora et al., 2019). This is due to the by variation in motor vehicle fleet composition (i.e. gasoline vs. diesel-powered vehicles) (Gentner et al., 2012), atmospheric chemistry conditions (Marr and Harley, 2002), and/or operating hours of local industry (Orozco et al., 2015).

Low-cost sensor data often drift over time leading to progressively worse biases (Miskell et al., 2018). The drift can be caused by aging components, intensity of LED dying out, sensors becoming less sensitive due to accumulation of dust, etc. The 8-month window for this study was well within the manufacturer estimated lifetime of 3 years for these Plantower sensors. No drift was detected during the study period.

3.3. *Regression results*

Four gain-offset regression models (presented in Table 1) were considered for the statistical calibration. Model 1 was the simplest with a constant gain and offset. Model 2 added the meteorological variables RH and T. Models 3 and 4 tried to fit the daily and weekly temporal patterns in the bias documented in Fig. 2. Hourly data from Dec 20, 2018 to July 31, 2019 were used to train each of these models and compared them using AIC and BIC. The model comparison metrics are presented in Table 3.

Both AIC and BIC penalize model complexity. Use of BIC generally tends to prevent overfitting as the penalty is stronger. Model 4 produced the best (lowest) AIC but it also produced the second highest BIC as it had 64 parameters. Model 3 only had 10 parameters and produced the lowest BIC and the second lowest AIC. To avoid potential overfitting by Model 4, Model 3 was selected as the final model. The in-sample R^2 for Model 3 is 65%. The coefficient estimates from Model 3 are plotted in Fig. S2. Three sets of estimates are presented corresponding to three choices of training data (i.e., training using data from only monitor Ot1, only monitor Ot2, or both monitors). Most of the coefficient estimates were consistent across the three choices.

Table 3
Comparison of the four gain-offset models of Table 1 using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Models with lower AIC or BIC is preferable.

Model	Number of Parameters	AIC	BIC
1	2	26,886	26,906
2	6	26,697	26,742
3	10	26,570	26,642
4	64	26,434	26,857

Based on the randomly selected 10% hold-out data, the hourly RMSE for the statistical model (Model 3) was 3.7 $\mu g\, m^{-3}$. For the same set of hold-out time points, the hourly RMSE for the raw data was 11.9 $\mu g\, m^{-3}$ and the RH/T corrected data was 5.8 $\mu g\, m^{-3}$. The mean Coverage Probability (CP) for the statistical model (Model 3) was 94.7% (very close to the nominal level of 95%). The CP for the raw data (89.2%) and the RH/T corrected data (91.7%) indicated slight under-coverage. The mean Confidence Interval Width (CIW) for the statistical model

predictions (Model 3) were much tighter (13.9 $\mu g\, m^{-3}$) compared to CIWs of 36.5 $\mu g\, m^{-3}$ and 20.9 $\mu g\, m^{-3}$ respectively for the raw and lab-corrected readings.

3.4. Seasonal analysis

In addition to the full training window of December 2018 to July 2019, the analysis was repeated over smaller time windows to see if the out-of-sample performance trends are consistent over time. The windows were December 2018 to February 2019, March 2019 to May 2019, and June 2019 to July 2019. These roughly corresponded to the three seasons (winter, spring and summer respectively).

The out-of-sample performance metrics for all four windows are provided in Fig. 3. The results for the three shorter time windows were consistent with that from the full analysis. The hourly RMSE for statistical model (Model 3) was generally higher (4.1 $\mu g\, m^{-3}$) in the winter months of December to February) than in Spring (2.9 $\mu g\, m^{-3}$) or summer (2.7 $\mu g\, m^{-3}$). The same trend was observed for the lab-corrected readings. For each of the three windows, the RMSE for the statistical model ware substantially lower than those from the lab-corrected readings (7.1 $\mu g\, m^{-3}$ in winter, 4.4 $\mu g\, m^{-3}$ in spring, and 4.4 $\mu g\, m^{-3}$ *in*summer) and the raw data (16.1 $\mu g\, m^{-3}$ in winter, 8.1 $\mu g\, m^{-3}$ in spring, and 6.1 $\mu g\, m^{-3}$ in summer).The mean Coverage Probabilities (CP) for the statistical model were generally close to the nominal level of 95% for all three windows. For the lab-corrected readings, they were close to 95% for spring and summer but had slight undercoverage (88.6%) during winter. The CP for the raw data were generally lower (around 90%) except for summer.The mean Confidence Interval Widths (CIW) were once again the narrowest for the statistical model (Model 3). The lab-corrected and raw data respectively produced 50% and 150% wider intervals.

The performance of the statistical model using hold-out-data was robust to the choice of training data being monitor-specific or pooled across monitors Ot1 and Ot2 (Fig. S3). Across all the time windows and choices of training data, the statistical model produced uniformly lowest RMSE suggesting that it has the highest predictive accuracy. It also produced tightest and well-calibrated prediction intervals.

3.5. Performance beyond co-location period

An out-of-sample performance evaluation of the regression model was conducted for the period of August to September 2019, i.e., for almost two months subsequent to the training period. The purpose of this analysis was to understand the accuracy of the statistical model beyond the period of co-location. The daily time series of the reference $PM_{2.5}$ data, the raw ($PM_{2.5\ raw}$) and lab-corrected readings ($PM_{2.5\ RH/T\ corrected}$) from the monitors Ot1 and Ot2, and the predictions from the statistical model ($PM_{2.5\ stat\ model}$) are presented in Fig. 4. The raw and the lab-corrected readings overestimated the $PM_{2.5}$ levels, as observed earlier in the year. The overestimation is more acute on days of elevated $PM_{2.5}$. One example is the data just before August 15, where the raw and the lab-corrected readings overestimated the actual peak respectively by around 20 $\mu g\, m^{-3}$ and 7 $\mu g\, m^{-3}$. The time-series of predictions from the regression model did not suffer from such overestimation. The predictions aligned much more closely to the true ambient $PM_{2.5}$ levels, accurately identifying both the times and the magnitudes of the peaks.

For this period, the hourly RMSE for the raw readings, the lab-corrected readings and the statistical model predictions for monitor Ot1 were 7.2 μgm^{-3}, 4.8 μgm^{-3} and 3.4 μgm^{-3} respectively The corresponding RMSEs for monitor Ot2 were 8.0 μgm^{-3}, 4.7 μgm^{-3} and 3.1 μgm^{-3} respectively.

3.6. Accuracy at different time-scales

The hourly predictions from the model were aggregated into daily and weekly predictions and RMSEs were calculated for these aggregation time-scales. In Fig. 5, the RMSEs for the three aggregation time-scales and the two choices of hold-out-data are provided.

For the statistical model, the hourly, 24-hr and weekly average RMSEs for the hold-out data within the training period were respectively 3.6 $\mu g\, m^{-3}$, 1.9 $\mu g\, m^{-3}$, and 1.5 $\mu g\, m^{-3}$ for monitor Ot1 . The corresponding numbers for the lab-corrected readings were 5.5 $\mu g\, m^{-3}$, 3.7 $\mu g\, m^{-3}$ and 2.5 $\mu g\, m^{-3}$. For the raw data, they were 10.6 $\mu g\, m^{-3}$, 8.7 $\mu g\, m^{-3}$and 6.3 $\mu g\, m^{-3}$. For all three series, RMSEs decreased as the length of the averaging time-window increased. This trend was consistent for both sets of hold-out data and for both monitors Ot1 and Ot2. It suggests that longer-term average levels of $PM_{2.5}$ were more accurately estimated by Plantower sensors. For all three averaging time scales and two hold-out periods, once again the predictions from the statistical model uniformly outperformed the raw or lab-corrected readings.

3.7. Transferability of calibration to a different MDE site

Before applying the estimated calibration equations from the statistical model to data from other monitors in the network, the cross-site and cross-monitor transferability of these equations were assessed. The calibration equation from Model 3 with parameters estimated from co-located data at Oldtown were used to calibrate the hourly readings from Monitor Ex1 at the suburban Essex site co-located with reference instrument. The 24hr-average RMSE for the period of February 2019 to August 2019 for the raw, lab-corrected, and statistically calibrated data were respectively 8.9 $\mu g\, m^{-3}$, 3.4 $\mu g\, m^{-3}$and 2.1 $\mu g\, m^{-3}$. Once again, the

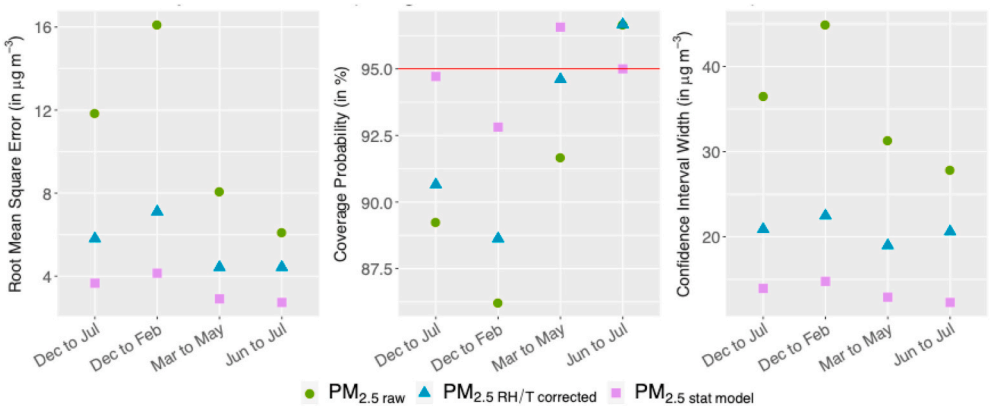

Fig. 3. Model comparison metrics using combined data from Monitors Ot1 and Ot2 with 10% hold-out data from Dec 2018 to July 2019 at Oldtown for different choices of training season. The left, middle and right panel respectively plots the RMSE, CP and CIW).Within each panel, the x-axis indicates the months of data used as training window for the model (e.g. Dec to Jul for the full training window of Dec 2018 to July 2019).

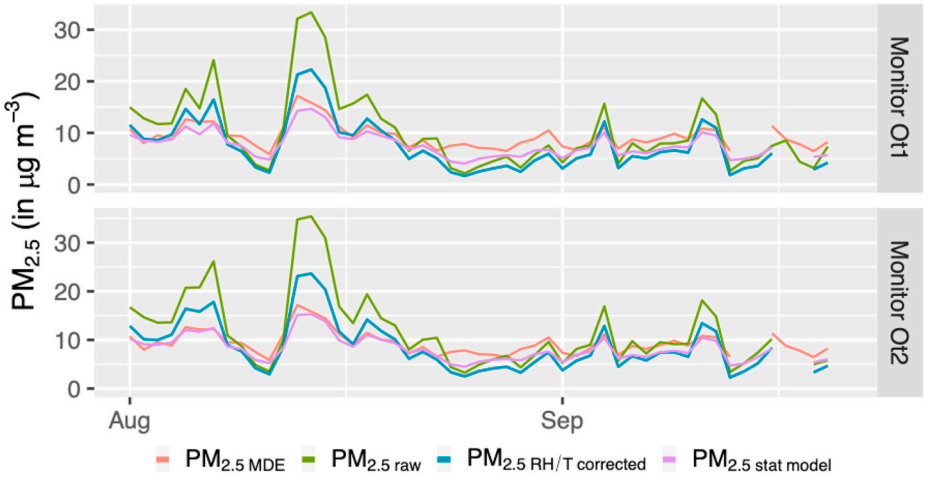

Fig. 4. 24-hr average $PM_{2.5}$ predictions ($PM_{2.5\ stat\ model}$) in $\mu g\ m^{-3}$ from Monitors Ot1 and Ot2 for the hold-out period of August and September 2019 from the statistical model (Model 3) trained with data from Dec 2018 to July 2019. The raw ($PM_{2.5\ raw}$) and lab-corrected data ($PM_{2.5\ RH/T\ corrected}$), and reference readings from the co-located MDE monitor at Oldtown ($PM_{2.5\ MDE}$) are also plotted for the same period.

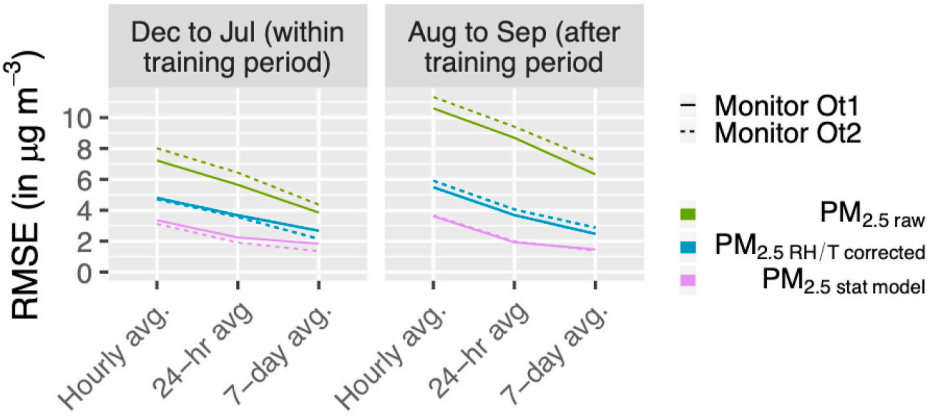

Fig. 5. Hourly, daily (24-hr), and weekly (7-day) averaged RMSEs in $\mu g\ m^{-3}$for the raw ($PM_{2.5\ raw}$), RH/T corrected ($PM_{2.5\ RH/T\ corrected}$) and model-predicted ($PM_{2.5\ stat\ model}$) $PM_{2.5}$ data. The RMSEs are based on two sets of hold-out data: within the training period of Dec 2018 to July 2019 (left), and Aug to Sep 2019, i.e., after the training period (right).

RMSE for the statistical model was substantially lower than those for the raw and the lab-corrected readings. Monitor Ex2, also co-located at Essex, yielded data almost identical to monitor Ex1 as shown in Fig. S4 and the analysis was not repeated for Monitor Ex2.

The time-series of daily reference and predicted $PM_{2.5}$ levels at Essex are presented in Fig. 6. The predictions from the statistical model quite accurately matched the ambient $PM_{2.5}$ levels for the entire period. The predictions both identified the peaks as well as correctly estimate the magnitude of the peaks. Both the raw and the lab-corrected readings overestimated the magnitude of the peaks.

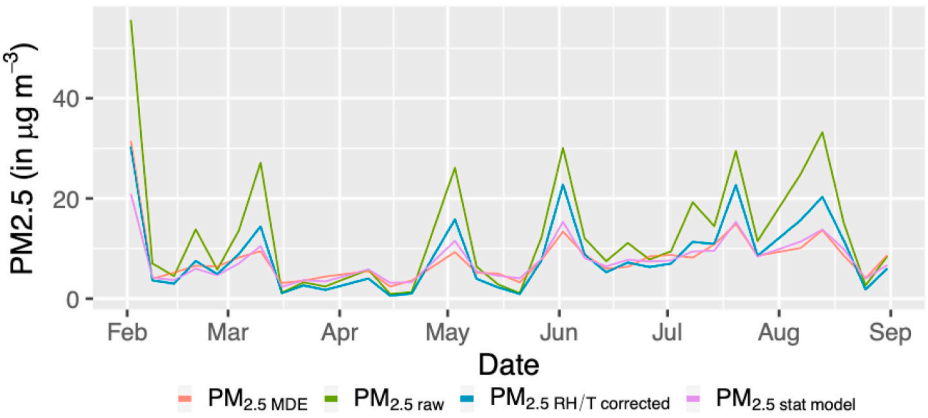

Fig. 6. Time-series of 24 h average reference $PM_{2.5}$ readings at Essex ($PM_{2.5\ MDE}$) available on every 6th day, along with corresponding raw ($PM_{2.5}$ raw), lab-corrected ($PM_{2.5\ RH/T\ corrected}$) and statistically calibrated ($PM_{2.5\ stat\ model}$) data from co-located low-cost sensor (monitor Ex1).

This cross-site and cross-monitor evaluation did not involve any retraining of the calibration model based on the Essex data. The model trained using Oldtown data was simply applied to calibrate readings from monitor Ex1 at Essex. The Essex site is located in a suburban setting outside of Baltimore city contrasting it with the urban Oldtown site in the city center. This exercise thus tested the transferability of the calibration equation to a different location and site type in the area. All the other monitors in the SEARCH network are in the same city and had similar $PM_{2.5}$, RH and T trends as the Monitors in Oldtown and Essex (Fig. S6). The performance of the statistical calibration at Essex (24 hr-average RMSE of around 2 $\mu g\, m^{-3}$ for a window of around 6–7 months) offered confidence about its applicability to the other monitors.

3.8. $PM_{2.5}$ levels in Baltimore on July 4, 2019

The statistically calibrated $PM_{2.5}$ readings from all the monitors in the SEARCH network provided insights into the spatial variations of $PM_{2.5}$ in Baltimore city. A case-study on the levels of $PM_{2.5}$ in the city around July 4, 2019 is presented here. The day was chosen as there is evidence suggesting Independence Day fireworks are associated with elevated levels of $PM_{2.5}$ in the US (Seidel and Birnbaum, 2015). The maps of calibrated $PM_{2.5}$ data for July 3, 2019 at 8pm and July 4, 2019 at 8pm are presented in Fig. 7. The two maps show differences in the levels of $PM_{2.5}$. On July 3, 25 low-cost units were operational. The calibrated $PM_{2.5}$ levels ranged from 8.1 $\mu g\, m^{-3}$ to 17.1 $\mu g\, m^{-3}$ with an average of 10.7 $\mu g\, m^{-3}$. For the same 25 units, 24 h later, $PM_{2.5}$ values ranged from 4.5 $\mu g\, m^{-3}$to 39.3 $\mu g\, m^{-3}$ with an average of 18.1 $\mu g\, m^{-3}$. A video showing changes in the $PM_{2.5}$ levels for each hour during this period is provided as a.gif file in the online Supplementary materials.

The video reveals that the elevated $PM_{2.5}$ levels primarily occurred during the hours of 7pm to midnight.

The increase in $PM_{2.5}$ was not uniform across the city. The monitor in Southwest Baltimore (Monitor 41) showed an increase from 9.6 $\mu g\, m^{-3}$ on July 3, 8pm to 39.3 $\mu g\, m^{-3}$ 24 h later. The spike recorded at the Maryland Science Center (Monitor 57) located on the Inner Harbor in Baltimore where a major fireworks display occurs was relatively moderate (9.6 μgm^{-3} on July 3 to 16.4 μgm^{-3} on July 4). The difference at Oldtown based on data from Monitor Ot2 was even more modest (8.1 μgm^{-3} on July 3 to 9.2 μgm^{-3} on July 4). The time-series of raw, lab-corrected and statistically calibrated readings for these three units during this period is presented in Fig. 7 (right panel). All three timeseries display a spike on July 4th. However, the magnitude and timing of the spikes differ considerably. Monitors 41 and 57 spiked roughly around the evening and night (7pm-midnight) on July 4th. The spike at the Maryland Science Center only went only up to 20 $\mu g\, m^{-3}$ the spike at the Southwest Baltimore site went up to almost 50 $\mu g\, m^{-3}$. The Oldtown site did not show substantial spike during this period but spiked later up to 15–20 $\mu g\, m^{-3}$ in the early hours of July 5. This case study nicely highlights the spatial and temporal variation of $PM_{2.5}$ within Baltimore which is impossible to capture with a single regulatory monitor.

4. Discussion

4.1. Comparison with performances of other on-field calibrated low-cost $PM_{2.5}$ networks

A comparison of the performance of the statistically calibrated SEARCH $PM_{2.5}$ monitors with other low-cost monitors is presented in Table 4. The comparison focused on four low-cost PM2.5 studies conducted in the United States that also reported hourly RMSE or Mean Average Error (MAE). US studies were used to reduce potential interferences due to factors that may vary across countries. $PM_{2.5}$ compositions can differ across countries due to differences in distributions of sources or fuel types (e.g. gasoline vs diesel) and in air quality standards. Such differences in $PM_{2.5}$ compositions could affect Plantower response factors (Levy Zamora et al., 2019). Typical ambient concentrations can also vary widely across countries and can impact sensor performance. For example, the performance of Plantower sensors have been reported to degrade at extremely high concentrations common in some countries (Sahu et al., 2020)).

Feenstra et al. (2019) conducted a comprehensive performance evaluation of twelve low-cost $PM_{2.5}$ sensors with respect to an FEM BAM instrument in Riverside, CA. The hourly RMSEs of the monitors ranged from 5.4 $\mu g\ m^{-3}$ to 18.0 $\mu g\ m^{-3}$. The Kaiterra LaserEgg (RMSE: 5.4 μg

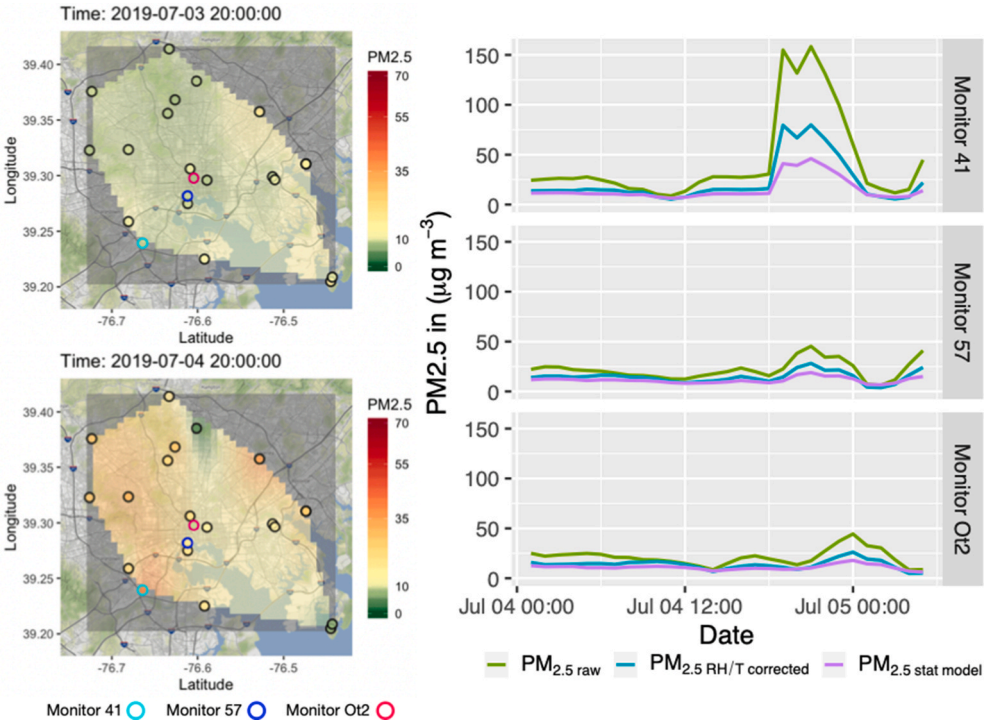

Fig. 7. $PM_{2.5}$ levels in $\mu g\, m^{-3}$ in Baltimore during July 4, 2019. Statistically calibrated $PM_{2.5}$ measurements from the SEARCH network on 8pm July 3, 2019 (top-left) and on 8pm July 4, 2019 (bottom-left). Time-series of the raw, lab-corrected and statistically calibrated data at Maryland Science Center (Monitor 57), southwest Baltimore (Monitor 41) and Oldtown (Monitor Ot2) from 12 a.m. July 4, 2019 to early morning of July 5, 2019. These three monitors are also highlighted in the maps on the left.

Table 4
Performance comparison of the statistically calibrated sensors in this study with those reported in recent studies of $PM_{2.5}$ low-cost sensors in the United States (all comparisons are based on hourly averages).

Study	Location	Reference Instrument	Sensor	Hourly RMSE[a] (or MAE[a]) in $\mu g\ m^{-3}$
SEARCH, Baltimore	Baltimore, MD	FEM BAM	Plantower PMS A003	3.6 (2.6)
Feenstra et al. (2019)	Riverside, CA	FEM BAM	Shinyei PM Evaluation Kit	6.7
			Alphasense OPC-N2	7.2
			TSI AirAssure	7.6
			Hanvon N1	21.9
			Airboxlab Foobot	8.6
			Kaiterra LaserEgg	5.4
			PurpleAir PA-II	10.0
			HabitatMap Air Beam 1	10.6
			SainSmart Pure Morning P3	7.8
			IQAir AirVisual Pro	5.8
			Uhoo	18.0
			Aeroqual AQY	6.1
Johnson et al. (2018a)	Atlanta, GA roadside	FEM TEOM	Shinyei PPD20V	6.2[b]
	Atlanta, GA rooftop		Shinyei PPD60PV	5.3[b]
Malings et al. (2019)	Pittsburgh, PA	FEM BAM	Met-One Monitor NPM	(3–4)
			PurpleAir PA-II	
Magi et al. (2019)	Charlotte, NC	FEM BAM	PurpleAir PA-II	4.1 (3.2)

[a] Mean or median RMSE and MAE when multiple units were studied.
[b] RMSE was calculated based on the bias and standard error reported.

m^{-3}) and IQAir AirVisual Pro (RMSE: 5.8 $\mu g\ m^{-3}$) sensors had the lowest RMSE. Johnson et al. (2018a) compared two Shinyei sensors against FEM TEOM reference data at two locations in Atlanta and reported hourly RMSE of 5 $\mu g\ m^{-3}$ to 6 $\mu g\ m^{-3}$. The hourly RMSE of around 3.4 $\mu g\ m^{-3}$ for Monitors Ot1 and Ot2 for almost two-months of hold-out testing period (August–September 2019) was around 2 $\mu g\ m^{-3}$ lower than the best performing sensors in these two studies. Malings et al. (2019) tested Met-One Monitor Neighborhood Particle Monitors (NPM) and a Purple Air $PM_{2.5}$ low-cost monitors against a FEM BAM in four sites around Pittsburg, PA. They reported an hourly MAE of 3–4 $\mu g\ m^{-3}$ which was comparable to the MAE for the SEARCH study (2.6 $\mu g\ m^{-3}$ for Monitor Ot1 and 2.4 $\mu g\ m^{-3}$ for Monitor Ot2). Similar performance of Purple Air $PM_{2.5}$ low-cost monitors tested with an FEM BAM was also reported in Magi et al. (2019) (hourly RMSE of 4.2 $\mu g\ m^{-3}$ and MAE of 3.4 $\mu g\ m^{-3}$). The similarity in performance between this study and the studies using Purple Air monitors can be partially attributed to the use of similar Plantower sensors in each design (SEARCH monitor: Plantower PMSA003; Purple Air monitor: Plantower PMS5003/PMS1003).

Co-location design: Oldtown was the only location providing reference $PM_{2.5}$ values in the city limits of Baltimore. Monitors Ot1 and Ot2 are permanently co-located there for the entire duration of the project. This helps to continuously monitor data-quality instead of having to find an optimal co-location-window. It also will allow updating of the regression estimates periodically in the future as more data are collected at Oldtown.

Training a statistical model based on co-location at a single site and using only two low-cost units has limitations. The co-located data can only identify biases that change with time. This limits the choice of covariates for the model to only time-varying variables T and RH, day of the week, hour of the day, etc. Biases in the Plantower sensors have been shown to vary with the PM composition (Levy Zamora et al., 2019) and this composition varies in space depending on proximity to different sources. Owing to the lack of multiple locations with reference $PM_{2.5}$ data, such spatially varying biases cannot be identified and corrected for. Success of the cross-site evaluation at Essex suggests that such biases may not be substantial within Baltimore.

The alternative to co-locating two monitors throughout the period of the project would have been individually co-locating each monitor used in the network at Oldtown for shorter periods. Monitor-specific biases that are constant over time, arising due to factors like hardware issues, could have been accounted for by this approach. However, for such a strategy, finding the optimal co-location window is challenging. Also, this would have delayed the deployment of the network by several weeks and require substantial manual efforts. Such a design would have considerably mitigated the utility of a low-cost network in producing instantaneous data. There was considerable evidence that at least for the length of this study (first 10 months of the network), such monitor specific-biases were less of a concern. First, the lab-corrections for RH and T was done using different groups of monitors and the calibration equations did not show much variability. Thus, a universal RH/T correction was used for all monitors. Second, the pairs of monitors at a site generally displayed extremely high correlation with each other (both for Oldtown and Essex). Third, the cross-site evaluation of calibrating readings from Monitor Ex1 at Essex using the model trained at Oldtown was successful. These provided evidence that the pragmatic strategy of only permanently co-locating two monitors at a reference site yielded an accurate field calibration model that could be applied to other monitors in the same geographical region.

Statistical Modeling: The selection of covariates in the statistical model was based on a hybrid approach.

Inclusion of covariates like T and RH despite using the lab-corrected readings is based on prior evidence that the range of ambient conditions for in-field co-location is wider than what is simulated in a laboratory. Hence, corrections only based on lab-experiments can be insufficient (Castell et al., 2017). Table 3 clearly revealed the utility of adding these meteorological variables as covariates. Model 2 which only includes RH and T had improved AIC and BIC over Model 1 which does not include them. Including covariates like day of the week, and hour of the day were predicated on the exploratory analysis (Fig. 2) and resulted in improved model performance. These variables partially reveal the daily and weekly periodicity in $PM_{2.5}$ levels which is possibly linked to the differential $PM_{2.5}$ compositions both within and between days. Future versions of the calibration model, using multi-year data, can use months or seasons as covariates to estimate annual seasonality. Additional temporal terms may also be needed to accommodate for potential drifts as the network starts to age.

A parsimonious gain-offset model, recast as a multiple linear regression, was used in this study. This allowed easy parameter estimation and model comparison. In the future, non-linear machine-learning approaches can be considered directly using the raw data as the dependent variable. Whether such methods can eliminate the need for the lab-correction step needs to be assessed. More discussion on this is presented in Section S1 of the Supplement.

Some in-field calibration methods adopt a multi-pollutant approach, calibrating a monitor's target pollutant using concentrations of other pollutants measured by the same monitor, along with other covariates (Topalović et al., 2019; Zimmerman et al., 2018). As described in Section 2.1, the SEARCH units simultaneously measure many pollutants and gases. A multipollutant approach was not used in this study. They can induce inaccurate statistical correlation between pollutants which can potentially interfere with the results of multipollutant studies using this data. To avoid skewing spatiotemporal patterns, studies should only consider applying such techniques when essential and with sufficient laboratory evidence to isolate and confirm the cross interferences. Also, reliance of the calibration on data on other pollutants measured by the monitors implies that malfunctioning of any of these sensors would

result in lost calibration data.

The statistical analysis presented in this study was primarily restricted to the exact locations where the low-cost units were hosted. During the early part of the study period the network had less than 10 low-cost monitors. Even in the later months (June or July 2019) around 20–30 monitors were active. Exploratory analysis (not included) suggested that due to the small number of data locations, sophisticated methods like kriging (Datta et al., 2016; Saha and Datta, 2018) did not work well for interpolation of the calibrated $PM_{2.5}$ data at an arbitrary location. The maps in Fig. 7 used *bsplines* for the spatial interpolation and is for visualization only. More discussion about spatial modeling and inference is provided in Section S1 of the Supplement.

Field testing routine: The set of evaluation procedures considered in this study can be used as a template for evaluating other stationary low-cost monitoring networks using only one or few co-located reference monitors. The evaluations consisted of out-of-sample validation using hold-out-data both within and outside of the training window. In addition to evaluating point predictions using RMSE or MAE, the prediction intervals were also evaluated using Coverage Probability and Confidence Interval Width. The robustness of the results to choices of co-location monitor and season of co-location were assessed. Evaluation of the data aggregated at different time scales (hourly, daily, and weekly) was conducted. This helped to assess the utility of the predicted $PM_{2.5}$ concentrations for future studies with data collected at those timescales. Finally, the transferability of the calibration equations to other units and locations was evaluated by co-location analysis at a second site with reference data.

5. Conclusions

The raw $PM_{2.5}$ data from the co-located low-cost monitors showed large biases (up to 25 $\mu g\, m^{-3}$). The raw data is particularly inaccurate when the true $PM_{2.5}$ levels are elevated. Generally, the RH was also elevated for these periods (Fig. S5), so the overestimation may be partially attributed to the environmental conditions (hygroscopy). The RH/T corrections based on lab-experiments prior to co-location improved both the accuracy and the precision of the readings. However, even after lab-correction, tendency of overestimating peaks prevailed, albeit to a lesser extent. Variability of the lab-corrected data was also higher than the reference data. Exploratory analysis confirmed that the biases were exacerbated during daytime hours and on weekends. The statistical model (hourly RMSE 3.7 $\mu g\, m^{-3}$) using these periodic variables along with RH and T out-performs the lab-corrected data (hourly RMSE 5.8 $\mu g\, m^{-3}$) and the raw data (hourly RMSE 11.9 $\mu g\, m^{-3}$). The statistical model also offered well-calibrated (Coverage Probability close to 95%) and tighter confidence intervals (width of 13.9 $\mu g\, m^{-3}$ compared to 20.9 $\mu g\, m^{-3}$ when using only lab-correction). The improvement due to the statistical calibration was consistent across out-of-sample validation exercises within and after the co-location period. The results were robust to different choices of the co-location monitor, and the co-location season. The low-cost data aligned better with the reference data when averaged over longer time periods (i.e., hourly RMSE > 24hr-average RMSE > weekly RMSE). The 24-hr average RMSE for the statistical model was 2 $\mu g\, m^{-3}$. External validation exercise at the co-location site at suburban Essex, MD confirmed that the model trained at Oldtown can be successfully used to calibrate other low-cost units in the network. A case-study was presented on the calibrated $PM_{2.5}$ data from the network on July 4th, 2019. The analysis revealed that $PM_{2.5}$ levels were higher that evening than the previous one, and that the magnitude and timing of the concentration spikes varied across the city. The case-study demonstrated the utility of the SEARCH low-cost network in understanding the spatio-temporal variations of $PM_{2.5}$ in Baltimore city.

CRediT authorship contribution statement

Abhirup Datta: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Arkajyoti Saha:** Methodology, Validation, Formal analysis, Visualization. **Misti Levy Zamora:** Investigation, Data curation, Writing - original draft, Writing - review & editing. **Colby Buehler:** Data curation, Writing - original draft. **Lei Hao:** Data curation. **Fulizi Xiong:** Data curation. **Drew R. Gentner:** Supervision, Resources, Writing - original draft, Writing - review & editing. **Kirsten Koehler:** Supervision, Resources, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: D. R.G. and F.X. have externally funded research projects on low-cost air quality monitoring networks, and Yale University has licensed out their technology.

Acknowledgements

This publication was developed under Assistance Agreement no. RD835871 awarded by the U.S. Environmental Protection Agency to Yale University. It has not been formally reviewed by the Environmental Protection Agency (EPA). The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. The EPA does not endorse any products or commercial services mentioned in this publication. A.D. and A.S. was supported by the Johns Hopkins Bloomberg American Health Initiative Spark Award. A.D. is supported by National Science Foundation DMS-1915803. C.B. is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1752134. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. M.L.Z. was also supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under awards number 1K23ES029985-01 and K99ES029116. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank the Maryland Department of the Environment Air and Radiation Management Administration for allowing collocation of the sensors with their instruments at the downtown Baltimore site. D. R.G., and F.X. would like to thank HKF Technology for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2020.117761.

References

Apte, Joshua S., Messier, Kyle P., Gani, Shahzad, Brauer, Michael, Kirchstetter, Thomas W., Lunden, Melissa M., Marshall, Julian D., Portier, Christopher J., Vermeulen, Roel C.H., Hamburg, Steven P., 2017. High-resolution air pollution mapping with google street view cars: exploiting big data. Environ. Sci. Technol. 51 (12), 6999–7008.

Balzano, Laura, Nowak, Robert, 2007. Blind calibration of sensor networks. In: Proceedings of the 6th International Conference on Information Processing in Sensor Networks. IPSN, pp. 79–88.

Byun, Daewon, Schere, Kenneth L., 2006. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale Air quality (CMAQ) modeling system. Appl. Mech. Rev. 59 (2), 51–77.

Cai, Jing, Yan, Beizhan, Ross, James, Zhang, Danian, Kinney, Patrick L., Perzanowski, Matthew S., Hwa Jung, Kyung, Miller, Rachel, Chillrud, Steven N., 2014. Validation of MicroAeth® as a black carbon monitor for fixed-site measurement and optimization for personal exposure characterization. Aerosol Air Qual. Res. 14 (1), 1–9.

Carvlin, Graeme N., Humberto, Lugo, Luis, Olmedo, Ester, Bejarano, Alexa, Wilkie, Dan, Meltzer, Michelle, Wong, Galatea, King, Amanda, Northcross, Michael, Jerrett,

Paul, B English, Donald, Hammond, Edmund, Seto, 2017. Development and field validation of a community-engaged particulate matter air quality monitoring network in Imperial, California, USA. J. Air Waste Manag. Assoc. 67 (12), 1342–1352.
Castell, Nuria, Franck, R., Schneider, Dauge Philipp, Vogt, Matthias, Lerner, Uri, Fishbain, Barak, Broday, David, Bartonova, Alena, 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302.
Clougherty, Jane E., Kheirbek, Iyad, Eisl, Holger M., Ross, Zev, Grant, Pezeshki, Gorczynski, John E., Johnson, Sarah, Markowitz, Steven, Kass, Daniel, Matte, Thomas, 2013. Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the New York city community air survey (NYCCAS). J. Expo. Sci. Environ. Epidemiol. 23 (3), 232–240.
Cohen, Aaron J., Michael Brauer, Richard Burnett, Anderson, H. Ross, Joseph, Frostad, Estep, Kara, Balakrishnan, Kalpana, Brunekreef, Bert, Dandona, Lalit, Dandona, Rakhi, Feigin, Valery, Freedman, Greg, Hubbell, Bryan, Jobling, Amelia, Kan, Haidong, Knibbs, Luke, Liu, Yang, Martin, Randall, Morawska, Lidia, Arden Pope, C., Shin, Hwashin, Kurt, Straif, Gavin, Shaddick, Thomas, Matthew, van Dingenen, Rita, van Donkelaar, Aaron, Vos, Theo, Christopher, J., Murray, L., Forouzanfar, Mohammad H., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. Lancet 389 (10082), 1907–1918.
Datta, Abhirup, Banerjee, Sudipto, Finley, Andrew O., Gelfand, Alan E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Am. Stat. Assoc. 111 (514), 800–812.
De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., Di Francia, G., 2018. Calibrating chemical multisensory devices for real World applications: an in-depth comparison of quantitative machine learning approaches. Sensor. Actuator. B Chem. 255, 1191–1210.
Deffner, Veronika, Küchenhoff, Helmut, Maier, Verena, Pitz, Mike, Cyrys, Josef, Breitner, Susanne, Schneider, Alexandra, Gu, Jianwei, Geruschkat, Uta, Peters, Annette, 2016. Personal exposure to ultrafine particles: two-level statistical modeling of background exposure and time-activity patterns during three seasons. J. Expo. Sci. Environ. Epidemiol. 26 (1), 17–25.
Di, Qian, Dai, Lingzhen, Wang, Yun, Zanobetti, Antonella, Choirat, Christine, Schwartz, Joel D., Dominici, Francesca, 2017. Association of short-term exposure to air pollution with mortality in older adults. JAMA - J. Am. Med. Assoc. 318 (24), 2446–2456.
Dominici, Francesca, Peng, Roger D., Bell, Michelle L., Pham, Luu, McDermott, Aidan, Zeger, Scott L., Samet, Jonathan M., 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. J. Am. Med. Assoc. 295 (10), 1127–1134.
Dominici, Francesca, Peng, Roger D., Zeger, Scott L., White, Ronald H., Samet, Jonathan M., 2007. Particulate air pollution and mortality in the United States: did the risks change from 1987 to 2000? Am. J. Epidemiol. 166 (8), 880–888.
Dons, Evi, Luc Int, Panis, Martine Van, Poppel, Jan, Theunis, Geert, Wets, 2012. Personal exposure to black carbon in transport microenvironments. Atmos. Environ. 55, 392–398.
Dons, Evi, Laeremans, Michelle, Pablo Orjuela, Juan, Avila-Palencia, Ione, Carrasco-Turigas, Glòria, Cole-Hunter, Tom, Anaya-Boig, Esther, Standaert, Arnout, De Boever, Patrick, Tim Nawrot, Götschi, Thomas, De Nazelle, Audrey, Nieuwenhuijsen, Mark, Int Panis, Luc, 2017. Wearable sensors for personal monitoring and estimation of inhaled traffic-related air pollution: evaluation of methods. Environ. Sci. Technol. 51 (3), 1859–1867.
Fann, Neal, Lamson, Amy D., Anenberg, Susan C., Wesson, Karen, Risley, David, Hubbell, Bryan J., 2012. Estimating the national public health burden associated with exposure to ambient PM2.5and ozone. Risk Anal. 32 (1), 81–95.
Feenstra, Brandon, Papapostolou, Vasileios, Hasheminassab, Sina, Zhang, Hang, Berj Der Boghossian, Cocker, David, Polidori, Andrea, 2019. Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. Atmos. Environ. 216.
Gao, Meiling, Cao, Junji, Seto, Edmund, 2015. "A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. Environ. Pollut. 199, 56–65.
Gentner, Drew R., Gabriel, Isaacman, Worton, David R., Chan, Arthur W.H., Dallmann, Timothy R., Davis, Laura, Liu, Shang, Day, Douglas A., Russell, Lynn M., Wilson, Kevin R., Weber, Robin, Guha, Abhinav, Harley, Robert A., Goldstein, Allen H., 2012. Elucidating secondary organic aerosol from diesel and gasoline vehicles through detailed characterization of organic carbon emissions, 45th ed. 109 Proc. Natl. Acad. Sci. U. S. A. 18318–18323.
Hajat, Anjum, Diez-Roux, Ana V., Adar, Sara D., Auchincloss, Amy H., Lovasi, Gina S., O'Neill, Marie S., Sheppard, Lianne, Kaufman, Joel D., 2013. Air pollution and individual and neighborhood socioeconomic status: evidence from the multi-ethnic study of atherosclerosis (MESA). Environ. Health Perspect. 121 (11–12), 1325–1333.
Hasenfratz, David, Saukh, Olga, Walser, Christoph, Hueglin, Christoph, Martin, Fierz, Arn, Tabita, Beutel, Jan, Thiele, Lothar, 2015. Deriving high-resolution urban air pollution maps using mobile sensor nodes. B 16. Pervasive Mob. Comput. 268–285.
Hoek, Gerard, Krishnan, Ranjini M., Beelen, Rob, Peters, Annette, Ostro, Bart, Brunekreef, Bert, Joel, D., Kaufman, 2013. Long-term air pollution exposure and cardio-respiratory mortality: a review. Environ. Health: A Global Access Science Source 12.
Johnson, Karoline K., Bergin, Michael H., Russell, Armistead G., Hagler, Gayle S.W., 2018a. Field test of several low-cost particulate matter sensors in high and low concentration urban environments. Aerosol Air Qual. Res. 18 (3), 565–578.
Johnson, Nicholas E., Bonczak, Bartosz, Kontokosta, Constantine E., 2018b. 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.
Kelleher, Scott, Quinn, Casey, Miller-Lionberg, Daniel, Volckens, John, 2018. A low-cost particulate matter (PM2.5) monitor for wildland fire smoke. Atmos. Meas. Tech. 11 (2).
Larson, Timothy, Henderson, Sarah B., Brauer, Michael, 2009. Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression. Environ. Sci. Technol. 43 (13), 4672–4678.
Lelieveld, Jos, Klingmüller, Klaus, Pozzer, Andrea, Ulrich, Pöschl, Fnais, Mohammed, Daiber, Andreas, Münzel, Thomas, 2019. Cardiovascular disease burden from ambient air pollution in europe reassessed using novel hazard ratio functions. Eur. Heart J. 20, 1590–1596.
Levy Zamora, Misti, Misti, Fulizi Xiong, Drew, Gentner, Kerkez, Branko, Kohrman-Glaser, Joseph, Koehler, Kirsten, 2019. Field and laboratory evaluations of the low-cost plantower particulate matter sensor. Environ. Sci. Technol. 53 (2), 838–849..
Levy Zamora, Misti, Pulczinski, Jairus C., Johnson, Natalie, Garcia-Hernandez, Rosa, Rule, Ana, Carrillo, Genny, Zietsman, Josias, Sandragorsian, Brenda, Vallamsundar, Suriya, Askariyeh, Mohammad H., Koehler, Kirsten, 2018. Maternal exposure to PM2.5 in South Texas, a pilot study. Sci. Total Environ. 628, 1497–1507.
Lim, Chris C., Ho Kim, M.J., Vilcassim, Ruzmyn, Thurston, George D., Gordon, Terry, Chen, Lung-Chi, Lee, Kiyoung, Heimbinder, Michael, Kim, Sun-Young, 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ. Int. 131.
Loomis, Dana, Grosse, Yann, Lauby-Secretan, Béatrice, El Ghissassi, Fatiha, Bouvard, Véronique, Benbrahim-Tallaa, Lamia, Guha, Neela, Baan, Robert, Mattock, Heidi, Kurt, Straif, 2013. The carcinogenicity of outdoor air pollution. Lancet Oncol. 14 (13), 1262–1263.
Magi, Brian I., Cupini, Calvin, Francis, Jeff, Green, Megan, Hauser, Cindy, 2019. Evaluation of PM2.5 measured in an urban setting using a low-cost optical particle counter and a federal equivalent method Beta attenuation monitor. Aerosol. Sci. Technol. 54 (2), 147–159.
Malings, Carl, Tanzer, Rebecca, Hauryliuk, Aliaksei, Saha, Provat K., Robinson, Allen L., Presto, Albert A., Subramanian, R., 2019. Fine particle mass monitoring with low-cost sensors: corrections and long-term performance evaluation. Aerosol. Sci. Technol. 54 (2), 160–174.
Marr, Linsey C., Harley, Robert A., 2002. Modeling the effect of weekday - weekend differences in motor vehicle emissions on photochemical air pollution in central California. Environ. Sci. Technol. 36 (19), 4099–4106.
Maryland Department of the Environment, 2020. AMBIENT AIR MONITORING NETWORK PLAN for CALENDAR YEAR 2020. https://Mde.Maryland.Gov/Programs/Air/AirQualityMonitoring/Documents/MDNetworkPlanCY2020.Pdf.
Miskell, Georgia, Salmond, Jennifer A., Williams, David E., 2018. Solution to the problem of calibration of low-cost air quality measurement sensors in networks. ACS Sens. 3 (4), 832–843.
Morawska, Lidia, Thai, Phong K., Liu, Xiaoting, Asumadu-Sakyi, Akwasi, Godwin, Ayoko, Bartonova, Alena, Bedini, Andrea, Chai, Fahe, Christensen, Bryce, Dunbabin, Matthew, Gao, Jian, Hagler, Gayle S.W., Jayaratne, Rohan, Kumar, Prashant, Alexis, K., Lau, H., Louie, Peter K.K., Mazaheri, Mandana, Ning, Zhi, Motta, Nunzio, Mullins, Ben, Rahman, Md Mahmudur, Ristovski, Zoran, Shafiei, Mahnaz, Tjondronegoro, Dian, Westerdahl, Dane, Williams, Ron, 2018. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? Environ. Int. 116, 286–299.
O'Connell, Steven G., Kincl, Laurel D., Anderson, Kim A., 2014. Silicone wristbands as personal passive samplers. Environ. Sci. Technol. 48 (6), 3327–3335.
Orozco, Daniel, Delgado, Ruben, Wesloh, Daniel, Powers, Richard J., Hoff, Raymond, 2015. Aerosol particulate matter in the Baltimore metropolitan area: temporal variation over a six-year period. J. Air Waste Manag. Assoc. 65 (9), 1050–1061.
Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., Li, K., Dick, R.P., Lv, Q., Hannigan, M., Shang, L., 2014. The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. Atmos. Meas. Tech. 7 (10), 3325–3336.
Powell, Helen, Krall, Jenna R., Wang, Yun, Bell, Michelle L., Peng, Roger D., 2015. Ambient coarse particulate matter and hospital admissions in the medicare cohort air pollution study, 1999–2010. Environ. Health Perspect. 123 (11), 1152–1158.
Saha, Arkajyoti, Datta, Abhirup, 2018. BRISC: bootstrap for rapid inference on spatial covariances. Stat 7 (1), e184.
Sahu, Ravi, Kumar Dixit, Kuldeep, Mishra, Suneeti, Kumar, Purushottam, Shukla, Ashutosh Kumar, Sutaria, Ronak, Tiwari, Shashi, Tripathi, Sachchida Nand, 2020. Validation of low-cost sensors in measuring real-time PM10 concentrations at two sites in Delhi national capital region. Sensors 20 (5), 1347 (Switzerland).
Samet, J.M., Zeger, S.L., Dominici, F., Curriero, F., Coursac, I., Dockery, D.W., Schwartz, J., Zanobetti, A., 2000. The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and Mortality from Air Pollution in the United States, , second ed.vol. 94. Health Effects Institute, pp. 5–79. Research Report.
Seidel, Dian J., Birnbaum, Abigail N., 2015. Effects of independence day fireworks on atmospheric concentrations offine particulate matter in the United States. Atmos. Environ. 115, 192–198.
Shah, Rishabh U., Robinson, Ellis S., Gu, Peishi, Robinson, Allen L., Apte, Joshua S., Presto, Albert A., 2018. High-spatial-resolution mapping and source apportionment of aerosol composition in Oakland, California, using mobile aerosol mass spectrometry. Atmos. Chem. Phys. 18 (22).
Steinle, Susanne, Reis, Stefan, Sabel, Clive E., Semple, Sean, Twigg, Marsailidh M., Braban, Christine F., Leeson, Sarah R., Heal, Mathew R., Harrison, David, Lin, Chun, Wu, Hao, 2015. Personal exposure monitoring of PM2.5 in indoor and outdoor microenvironments. Sci. Total Environ. 508, 383–394.

Szpiro, Adam A., Sampson, Paul D., Sheppard, Lianne, Lumley, Thomas, Adar, Sara D., Kaufman, Joel D., 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21 (6), 606–631.

Topalović, Dušan B., Davidović, Miloš D., Jovanović, Maja, Bartonova, A., Ristovski, Z., Jovašević-Stojanović, Milena, 2019. In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: comparison of linear, multilinear and artificial neural network approaches. Atmos. Environ. 213, 640–658.

Trasande, Leonardo, Malecha, Patrick, Teresa, M Attina, 2016. Particulate matter exposure and preterm birth: estimates of U.S. Attributable burden and economic costs. Environ. Health Perspect. 124 (12), 1913–1918.

United States Environmental Protection Agency. Evaluation of emerging air pollution sensor performance. https://www.epa.gov/air-sensor-toolbox/evaluation-emerging-air-pollution-sensor-performance n.d.

Van den Bossche, Joris, Peters, Jan, Jan, Verwaeren, Botteldooren, Dick, Jan, Theunis, De Baets, Bernard, 2015. Mobile monitoring for mapping spatial variation in urban air quality: development and validation of a methodology based on an extensive dataset. Atmos. Environ. 105, 148–161.

Wang, Yanwen, Du, Yanjun, Wang, Jiaonan, Li, Tiantian, 2019. Calibration of a low-cost PM2. 5 monitor using a random forest model. Environ. Int. 133, 105161.

World Health Organization, 2016. Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease. *World Health Organization*.

Zimmerman, Naomi, Presto, Albert A., Kumar, Sriniwasa P.N., Gu, Jason, Hauryliuk, Aliaksei, Robinson, Ellis S., Robinson, Allen 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 (1).