Understanding the source components captured by the Purple Air Network

Vijay Kumar,[†] S. Dinushani Senarathna,[†] Supraja Gurajala,[‡] William Olsen,[¶] Shantanu Sur,[§] Sumona Mondal,[†] and Suresh Dhaniyala[*,‖]

[†]*Department of Mathematics, Clarkson University*

[‡]*Department of Computer Science, State University of New York, Potsdam*

[¶]*Department of Civil and Environmental Engineering, Clarkson University*

[§]*Department of Biology, Clarkson University*

[‖]*Department of Mechanical and Aeronautical Engineering, Clarkson University*

E-mail: sdhaniya@clarkson.edu

Abstract

$PM_{2.5}$ has been linked to numerous pollution-mediated adverse health effects and their monitoring is key for taking preventative and mitigative measures. Accurate measurements of $PM_{2.5}$ concentrations are available at EPA sites, but such data lacks spatial resolution due to a limited number of monitoring locations. In recent years the deployment of low-cost sensor networks has opened up the possibility of acquiring air quality data at a high spatio-temporal resolution. However, the sensitivity, noise, and accuracy of data acquired by low-cost sensors remain a concern. Here, we studied $PM_{2.5}$ measurements made from EPA and Purple Air (PA) sensor networks in the Chicago area to understand the parameters influencing the performance characteristics of the low-cost sensor network. Using time series decomposition of $PM_{2.5}$ data into short-term and baseline components using Kolmogorov–Zurbenko (KZ) filter and analysis of

the extracted frequency signals, we determine that PA sensor data is more sensitive to meteorological conditions than anthropogenic activities in both short-term, and baseline components. We hypothesize that the low-cost sensor networks may have different sensitivity to aerosol from different sources and hence care must be taken in their calibrations and in their use for evaluating the impact of air quality mitigation policies.

Introduction

Air pollution is one of the world's leading risk factors for disease and premature death. An estimated 16% of total global deaths in 2015 can be attributed to diseases caused by air pollution.[1] Of particular concern is the mass concentration of Particulate Matter (PM) smaller than 2.5 μm, i.e. $PM_{2.5}$, or fine particles. Exposure to $PM_{2.5}$ has been directly correlated to diseases such as respiratory diseases and even mortality.[2–6] The high health impact of $PM_{2.5}$ is because of their ability to penetrate deep into the lungs and because their composition is often carcinogenic.[7] The European Study of Cohorts for Air Pollution Effects (ESCAPE) shows that exposure to high $PM_{2.5}$ concentrations are linked with a risk of developing lung cancer.[8] In addition to chronic diseases, exposure to $PM_{2.5}$ also impacts our response to acute diseases such as COVID-19.[9–12] Accurate knowledge of $PM_{2.5}$ exposure and efforts to mitigate it are critical to protecting public health.

In the United States, the Environmental Protection Agency (EPA) monitors air quality by measuring regulated or criteria pollutants including ambient $PM_{2.5}$ concentrations using Air Quality Monitoring Stations (AQMSs). The $PM_{2.5}$ measurements are made using a range of instruments classified as federal reference methods (FRMs) or federal equivalent methods (FEMs).[13] These methods ensure consistency and accuracy in measurements, but are expensive, and difficult to operate, requiring trained personnel and significant infrastructure. The strict maintenance and calibration routines followed in these stations ensure

high-quality data and comparability between different locations.[14] Even in the US, with over 5000 AQMSs, the geographic coverage of these monitoring sites is inadequate. The siting of AQMS is often biased towards populated areas, disadvantaging smaller cities and underdeveloped regions.[15] Even in populated areas, the limited number of sites do not capture the high spatial variation in $PM_{2.5}$ concentrations that are likely, resulting in an incorrect estimate of exposure and resultant health effects.[16]

For accurate exposure assessment, an air quality monitoring network providing measurements at high spatio-temporal resolution is required. To address this need, researchers, communities, organizations, and individuals have been deploying low-cost air quality sensors that provide air quality data at a granular level not possible with the EPA AQMSs.[17,18] One of these networks is composed of sensors from Purple Air (PA). The PA sensing platform incorporates a pair of Plantower PMS 5003 low-cost sensors, which use laser light scattering techniques to determine ambient aerosol concentrations. The PMS5003 reports a variety of particle concentration metrics including PM_1, $PM_{2.5}$, and PM_{10}.[19–21] Using two sensors for PM measurements allows for the robustness of data collection.[22] While the low-cost sensors have the advantage of deployment ease, their accuracy and precision are variable.[23]
The various PM sensors used in low-cost monitors are all subject to biases and calibration dependencies, with some factors accounted for with moderate success (e.g. meteorology, age of sensor) and others poorly (e.g. aerosol source, composition, refractive index).[24] The PA sensor measurements are often calibrated/corrected by co-location with a reference monitor at a regulatory site.[25–27] Additionally, researchers have developed correction models to account for the impact of environmental conditions on sensor performance.[15,28] The deployment of PA sensors has resulted in expanding the availability of $PM_{2.5}$ data and enabling a range of studies, including, validation of high resolution, large-scale regional modeling efforts[29] and understanding of the impact of wildfire smoke on local and regional air quality.[30]

Co-locating low-cost sensors with reference monitors provides a fast way for their calibration. Typically, this is done by co-locating the sensors for a period of time and then determining a scaling factor or equation based on a regression analysis. The time period for co-location is generally chosen to be around days to weeks and this allows for the calibration to be independent of data noise. The selection of the calibration time period can, however, bias the sensor data to be most sensitive to sources primarily responsible for pollutant concentration variability in that time period. Sources with shorter time periods, relative to the calibration period, are averaged out and inadequately accounted for in the calibration. Thus longer time scale events are completely lost in the calibration process.

Published studies on low-cost sensors have observed some of the above mentioned problems. The response characteristics of low-cost sensors are seen to be different from that advertised by their manufacturers, possibly because the aerosol size distributions and compositions differ with location.[31,32] As an example, low-cost sensor data are seen to be in better agreement with reference monitors at locations with low traffic than those at high-traffic locations.[14] To improve the quality of the reported data from low-cost sensor networks, we need to establish ideal field calibration principles for these units. For this, frequency based methods that have been previously used in air quality to find prominent temporal components can be used.[33–36] Time-series decomposition using low-pass filters can identify pollution sources that account for most of the measurement variation.[37,38] Here, using frequency-based analysis, the dependence of low-cost sensor $PM_{2.5}$ measurement accuracy on the calibration period will be established.

For this work, we chose our study area as Cook county, IL which includes the City of Chicago and a total population of nearly 10 million. Cook County is a major transportation hub lying at the crossroads of the country's rail, road, and air traffic, and an important industrial center, thus, there are a number of emission sources within the area. Despite a baseline long-term trend of improving air quality in Chicago, recent years show a worsening trend. $PM_{2.5}$ concentrations have nearly doubled since 2017, rising from 6.7 $\mu g/m^3$ in 2017

to 12.8 $\mu g/m^3$ in 2019, exceeding US EPA air quality standards (12 $\mu g/m^3$).[39] The changing air pollution levels have increased public interest in air quality monitoring, particularly using low-cost sensor networks. For the time period starting May 2018, the purple air network in Chicago and its surrounding neighborhoods have increased from a few sensors to more than 30 sensors now.

In this study, we used $PM_{2.5}$ data from EPA sites and PA sensors located in Cook County, IL to understand the differences in their data as a function of sensor location and time. Using spectral theory to extract the temporal signatures of the EPA and PA data, we analyze the short-term, and baseline components of air quality as measured by the two networks. The spectral analysis is used to understand the sources of biases in the PA network and provides guidance in improving the field calibration of these sensors.

Materials and Methods

Data Collection and Pre-processing

Monitoring Data

Cook County, IL, has 14 EPA air quality monitoring sites, providing data on criteria pollutants, including ambient $PM_{2.5}$ concentrations (`https://aqs.epa.gov/aqsweb/documents/data_api.html`). Hourly $PM_{2.5}$ measurements from EPA are available at 7 out of 14 monitoring sites in Cook County, IL. The PA network in Cook County consists of more than 30 PA low-cost sensors, that currently provide $PM_{2.5}$ data (`https://map.purpleair.com/1/mAQI/a10/p604800/cC0#11.44/41.8363/-87.6973`). Our analysis was conducted using data from a time period of October 2019 to September 2021. For this time period, hourly $PM_{2.5}$ data was only available at 10 out of 30 PA sensors. Further, after eliminating sites with more than 20% missing data, our analysis could only use data from 5 EPA sites and 9 PA sensors, as shown in (Figure 1).

It was observed that PA data included some outliers with very large $PM_{2.5}$ concentrations, which are likely erroneous data. To eliminate these outliers from our analysis, we chose a data range of [1,70] ug/m^3 as valid data.[15] In (Figure 1a) the sampling locations of EPA and PA are plotted on the map with the population density around the sampling locations in (Figure 1b). The population density was calculated in the census blocks, as defined by US Census Bureau,[40] using ArcgisPro 2.8. From a cursory analysis of the siting of sensors, it is clear that the PA sensors are located in urban areas where population densities on average are higher than what they are at the EPA sites except few sensors i.e P1, P5, and P8.

Meteorological Data

The impact of meteorological variables on both EPA and PA data is important to assess. In particular, it has been established that low-cost sensors are sensitive to meteorological parameters, especially relative humidity.[15,28] For this assessment, we collected meteorological data from 5 nearby stations of the National Oceanic and Atmospheric Administration (NOAA). The meteorological variables include temperature (T), relative humidity (RH), wind speed, and wind direction. The meteorological variables are used to identify and quantify the impact of weather on $PM_{2.5}$ concentrations coming from PA and EPA sensors.

Monitoring Data Summary

This study uses 2 years of $PM_{2.5}$ data from 5 EPA sites and 9 PA sensors. Sample time series trend in $PM_{2.5}$ from a set of EPA and PA sites (EPA site E2 and PA sensor P6) that are in close vicinity (within 2 km) to each other are shown in (Figure 2a). High temporal variations are seen in the data from both networks' sensors, along with some seasonal trends over longer timescales. Comparing such time series data from the EPA and PA sites in close vicinity to each other (Figure 1) shows that the PA data overestimates $PM_{2.5}$ relative to EPA measurements at all locations (Figure 2b). The gap in the total time

series of $PM_{2.5}$ data around April of 2020 in E2 and September-October, 2021 in P6 is due to missing observations in the time series in (Figure 2a). The major causes for missing air pollutant data in reference monitor includes monitor malfunctions and errors, power outages, computer system crashes, pollutant levels lower than detection limits, and filter changes.[41,42] Whereas in low-cost sensors approximately 40 % of the data generated is missing, most likely because of extreme weather events, battery failure, and disruption in internet accessibility at sensors location.[43,44] Looking at the graph of the original time series of both networks in (Figure 2a), the peaks in July of both years show the effect of independence day fireworks, the $PM_{2.5}$ concentrations on July 4 and 5 are greater than on the two days in July before and after, which are considered control days. On the national average, the increases are largest (21 $\mu g/m^3$) at 9–10 pm on July 4.[45] The $PM_{2.5}$ concentrations are lower in July 2020, as compared to those of July 2021 in low-cost sensors, as well as reference monitors, which maybe due to COVID-19 restrictions. In fact, overall air quality slightly improved in 2020 as compared to 2021 maybe due to COVID-19 lockdowns. The $PM_{2.5}$ concentrations are high in winter in (Figure 2a) due to a combination of continuous temperature inversions and a thin mixed boundary layer throughout the winter months made it difficult for pollutants to scatter into the atmosphere. This led to an increase in anthropogenic emissions related to the demand for heating[46]

The distribution of the $PM_{2.5}$ data at each of the EPA and PA sites over the entire time period of analysis is shown in (Figure 3). The median values of $PM_{2.5}$ from the PA sites are always higher than that from nearby EPA sites, again suggesting that the data from the sensor network is an overestimation of the local $PM_{2.5}$ values. Additionally, the overall distribution of the $PM_{2.5}$ data, irrespective of the sampling location, is narrower for the EPA sites than for the PA sites.

There can be many reasons for the overestimation of $PM_{2.5}$ by the PA sensors. The PA $PM_{2.5}$ values are obtained from the light scattering signal based on calibration using Beijing air. As the size distribution of particles in Chicago is likely different from that in Beijing,

the calibration may not be entirely valid. Also, temperature and relative humidity influence particle size and optical properties, and, thus, PA measurements do not affect EPA measurements due to thermal control in their measurements.[27,47,48] Lastly, PA sensor measurements are sensitive to particle composition and, thus, when there are a large number of particle sources, and hence particle compositions, such as in an urban area like Chicago, then the mass distributions reported by an optical sensor can be broader than that obtained gravimetrically. It must be noted that the differences in the measurement techniques affect not just the reported magnitudes of $PM_{2.5}$ concentrations, but also the overall average value. The annual average of $PM_{2.5}$ values from EPA show a reduction from 12.8 $\mu g/m^3$ in 2019, to 8.8$\mu g/m^3$ in 2020, while the average PA $PM_{2.5}$ values from PA is 12.97 $\mu g/m^3$, more than the NAAQS standard.

Recently a US-wide correction model for PA sensors that takes into account the contribution of ambient conditions on sensor performance was introduced.[28] The model was built using data from 53 PA sensors, with data spanning the time period of September 2017 to January 2020, at 39 distinct sites spread throughout 16 states. From an evaluation of several models using temperature and relative humidity, they suggested a final model only considering the effect of relative humidity (RH) on PA sensor data. Using this US-wide correction model and data from EPA sites in the vicinity of the sensors, we corrected the PA data of our study location of Cook County, IL.

In (Figure 4), we show the weekly averages of corrected PA data from sensors P6 and P1 in comparison with nearby EPA data from sites E2 and E1, respectively. The data is split into weekdays and weekends for comparison. The time series trends seem to suggest that the correction results in the EPA and PA data largely overlap with each other over the period of study. However, a two sample t-test between the EPA and corrected PA data shows that the difference between the two data sets is statistically significant (p-value < 0.05) on weekdays but not on weekends. The better match of the two data sets on weekends could

be because of the lower contribution of traffic to local air quality on weekends compared to weekdays. Previous studies[14] have shown that low-cost sensor measurements more closely match reference monitors at locations with low traffic than at high traffic locations.[49]

To understand the underlying differences between two data sets and identify any drivers of PA data inaccuracy, a frequency-based analysis proves to be helpful. As aerosol sources have distinct time-period signatures, frequency analysis can help determine the relative response of PA sensors to different emission sources.

Spectral Analysis: Methods

In meteorology and air quality studies, spectrum-based analysis has often been used to extract and examine different temporal components in the obtained data.[33–36] Here, using spectral analysis of the PA and EPA data, we identify similarities and differences among the sources contributing to these data sets.

In general, a time series X_t of length N can be represented as a linear combination of harmonic functions with frequencies f_j and amplitudes A_j and B_j:

$$X_t = \mu + \sum_{j=1}^{[N/2]} \Big[A_j \cos(2\pi f_j t) + B_j \sin(2\pi f_j t) \Big], t = 1, 2, ...N \tag{1}$$

where μ is a constant, $[N/2]$ is the greatest integer less than or equal to $N/2$, and the frequencies f_j are related to the sample size N by

$$f_j = j/N, 1 \leq j \leq [N/2] \tag{2}$$

Thus, for a measurement resolution of 1 hour, a wave with a period of 2 hours or more is required (Nyquist theorem).

The discrete Fourier transform, $X(k)$, of hourly time series X_t, can be calculated using

the Fast Fourier transform (FFT) algorithm. The spectral density for a finite time series can then be calculated as the squared magnitude of $X(k)$:

$$\Phi(v_k) = |X(k)|^2 = \left| \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} X_t e^{-2\pi i v_k t} \right|^2 \tag{3}$$

where $k = 0, 1, ..., (N-1)$. N is the number of observations and $v_k = \frac{k}{N}$.

FFT, which performs spectral analysis, needs successive, equal length sequences in order to ensure that no data points are missed.[50] Here we replace the missing data in all EPA sites, and PA sensors using the ARIMA model with Kalman filter.[51–54] The power spectral density of each EPA and PA hourly time series of $PM_{2.5}$ data was then calculated using the stats package in R.

Spectral Analysis: Results and Discussions

The power spectral density (PSD) of $PM_{2.5}$ data from each EPA site and PA sensor were calculated and averaged in the same type of monitor in order to understand and quantify the differences between the EPA and PA network measurements in various temporal periods, as illustrated in (Figure 5). The PSD shows the distinct frequency peaks corresponding to components that provide higher contributions to the total variance in $PM_{2.5}$ data. These peaks correlate to higher frequencies with corresponding time periods of 24 hours, to 4 hours, as well as low frequencies corresponding to 6-month, monthly, and weekly time periods for all monitoring stations of EPA as well as low-cost sensors. This pattern is related to the frequency of emission sources, short-term weather patterns, and long-term seasonal changes.[34,36] It has been observed that the proportion of low frequency variance in relation to the total variance is similar in low-cost sensor data compared to reference monitors. (Figure 5a). Conversely, the total variance in higher frequencies is greater in reference monitors than in low-cost sensors. The difference in the proportion of variance in frequencies relative

to the total variance between low-cost sensors and reference monitors may be due to low-cost sensors overestimating observations at low frequencies. However, at higher frequencies < 24 hours, low-cost sensors' observations and peaks are smaller than those of reference monitors, which may be due to low-cost sensors' lower sensitivity to local short-term changes such as traffic. In general, if we compare the overall contribution of variance in specific frequencies to the total variance of the two networks, we can hypothesize that either the PA sensors are less responsive to short-term changes caused by anthropogenic activity or more sensitive to weather variations.

To analyze the pattern of PSD in both networks, the average PSD ratio of the PA sensors with the average PSD of EPA sites was calculated and plotted in (Figure 5b). For comparison of PSD of data from the same network, We displayed the ratios of four PA sensors with four other PA sensors and two EPA sites with two other EPA sites in (Figure A1). The PSD ratio of PA sensors tends to be similar to that of other PA sensors, and the PSD ratio of EPA sites tends to be similar to that of other EPA sites, with both hovering around 1. This indicates that measurements taken from the same type of network tend to have relatively similar PSD values. Whereas, for different networks in (Figure 5b), The average PSD ratio decreases below 1 at frequencies below 24 hours, indicating that PA data and EPA data have similar contributions to the baseline of $PM_{2.5}$ concentrations at low frequencies but PSD of PA is lower at high frequency such as below 24 hours. If the responses in both networks were similar over the entire time series, the ratio would be closer to 1 at all frequencies due to corrections. However, the pattern of the ratio declining at frequencies below 24 hours suggests that PA sensors are less able to capture short-term changes compared to EPA instruments, but are more effective at capturing baseline changes due to weather. The low-cost sensors in the PA network are highly correlated with each other, indicating that the measurements and algorithms used by each device are consistent across all sensors.[55] In contrast, reference monitors in the EPA network are not highly correlated with each other. (Table A4.

Time Series Decomposition: Methods

Recognizing that the correction models do not uniformly account for the contribution of all sources to the PA data, particularly at higher frequencies. Therefore, we plan to investigate the source components impacting both lower and higher frequencies. To further examine this, we have separated the data into short-term and long-term/baseline components. The short-term component includes high-frequency data that is influenced by local anthropogenic sources such as traffic and short-term weather events. The baseline component, on the other hand, includes low frequency data that are related to seasonal changes in weather and meteorology, and changes in emission rates over time.[56–58]

The time series data is separated into short-term and baseline components using the Kolmogorov–Zurbenko (KZ) filter technique.[56] Recent studies have applied the KZ filter to determine the source component in short-term and baseline components of $PM_{2.5}$.[37,38,59,60] The KZ filter is a low-pass filter produced through repeated iterations of a moving average with parameters moving window (m), and iterations (p) also known as $KZ_{m,p}$:

$$Y_t = \frac{1}{m} \sum_{j=-k}^{k} X_{t+j} \tag{4}$$

where Y_t is a filtered time sequence; X_t is the input time series; k is the number of values included on each side of the targeted value, $m = 2k+1$ is window length; t is the time index, and j is the time point of sliding. The output of the first pass then becomes the input for the next pass. Adjusting the window length and the number of iterations makes it possible to control the filtering of different scales of motion.[61,62] To filter a period of fewer than N days, the following criterion is applied to determine the filter's effective width:[58]

$$m \cdot^{1/2} \leq \mathrm{N} \tag{5}$$

Also, the filter can be used to remove frequencies below a desired cutoff frequency w_0:[57]

$$w_0 \approx \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2p}}{m^2-(1/2)^{1/2p}}} \tag{6}$$

The cutoff period can be obtained by $\frac{1}{w_0}$. For our study, we have used $KZ_{15,3}$, as it is designed to remove short-term cycles of noisy data.[38]

The baseline $PM_{2.5}$ ($PM_{2.5,B}$) and meteorological (M_B) components are obtained as:

$$PM_{2.5,B}(t) = KZ_{15,3}PM_{2.5}(t) \tag{7}$$

$$M_B(t) = KZ_{15,3}M(t) \tag{8}$$

The short term $PM_{2.5}$ ($PM_{2.5,S}$) and meteorological (M_S) components are defined as:

$$PM_{2.5,S}(t) = PM_{2.5}(t) - KZ_{15,3}PM_{2.5}(t) \tag{9}$$

$$M_S(t) = M(t) - KZ_{15,3}M(t) \tag{10}$$

Relative Contributions of Temporal Components

By separating the data into short-term and baseline components, we can analyze and examine how each component contributes to the overall variance of the time series data for both EPA and PA..[63] The relative contributions of temporal components are obtained as:

$$Relative\ \ contribution\ (\%) = \frac{Var(i(t))}{Var(X(t))} \cdot 100 \tag{11}$$

where $Var(i(t))$ is variance of short-term, or baseline component, and $Var(X(t))$ is variance of total $PM_{2.5}$ time series.

$PM_{2.5}$ Contributions from Meteorology and Anthropogenic Emissions

The short-term and baseline components of $PM_{2.5}$ can be combined with short-term and baseline components of meteorology to quantify the effect of meteorology and relatively estimate the effect of anthropogenic activities on $PM_{2.5}$ data.[38,64,65] The $PM_{2.5}$ data can be approximated as short-term and baseline $PM_{2.5}$ measurements as:

$$PM_{2.5} = PM_{2.5,B}(t) + PM_{2.5,S}(t), \tag{12}$$

The MLR models for short-term and baseline components of $PM_{2.5}$ in relation to short-term and baseline of meteorology and anthropogenic activities can be written as:

$$PM_{2.5} = PM_{2.5,B}(t)+PM_{2.5,S}(t) = \Big[a_0+\sum_{i=1}^{6} a_i M_{Si}(t)\Big]+\Big[b_0+\sum_{i=1}^{6} b_i M_{Bi}(t)\Big]+(\epsilon_B+\epsilon_S), \tag{13}$$

where,

$$PM_{2.5,S}(t) = \Big[a_0 + \sum_{i=1}^{6} a_i M_{Si}(t)\Big] + \epsilon_S, \text{ and } PM_{2.5,B}(t) = \Big[b_0 + \sum_{i=1}^{6} b_i M_{Bi}(t)\Big] + \epsilon_B. \tag{14}$$

$M_{2.5,Si}(t)$, and $M_{2.5,Bi}(t)$ are time series of the i^{th} meteorological variable for short-term and base-line components, respectively, and a_0, b_0, a_i, and b_i are regression model parameters to be estimated using a stepwise algorithm in MLR model. The residuals ϵ_S, ϵ_B represent changes in $PM_{2.5}$ concentrations that cannot be attributed to meteorological variables present in the model and are mainly due to anthropological activities in the short-term and baseline components, respectively.[38,64,66] To estimate the impact of metrology and anthropogenic impact on both short-term and baseline $PM_{2.5}$, we built models considering $PM_{2.5}$ data

as the response variable and meteorological data from nearby NOAA site as the predictor variable for each EPA site and PA sensor.

Relative Importance of Predictors

As a result of MLR models, we could quantify the overall impact of meteorology on $PM_{2.5}$ measurements of both EPA and PA networks in short-term and baseline components. However, the question of which predictor is most influencing the data of both networks has no trivial answer due to the presence of many predictors. Correlation analysis is often used to examine the relationship between two variables. However, when there are many predictors, correlation analysis is not the best method to use." Here, we use the LMG measure proposed by Lindeman, Merenda and Gold,[67] and popularized by[68] to determine the relative importance of predictors.

The LMG measure uses sequential R^2, but it accounts for the dependence on orderings by averaging over all possible orderings. According to[69,70] the variance decomposition for a linear model with k predictors can be defined as:

$$E(Y|X) = X\beta, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{15}$$

and

$$V(Y) = \sum_{j=1}^{p} \beta_j^2 + 2\sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_j \beta_k \sqrt{\nu_j \nu_k} \rho_{jk} + \sigma^2, \tag{16}$$

where ν_j, and ν_k are variance of each predictor, $\rho_j k$ =covariance of predictor , $j = 1, 2, 3, ..., p$, $k = j + 1, ..., p$.

The R^2 for a model with predictors in set S is given as,

$$R^2(S) = \frac{\text{Model sum of square}}{\text{Total sum of square}} \tag{17}$$

The additional R^2 adding set M to a model with the predictors in set S,

$$seqR^2(M|S) = R^2(M \cup S) - R^2(S), \tag{18}$$

where S and M are disjoint sets of predictors.

$$seqR^2(x_k|S_k(r)) = R^2(x_k \cup S_k(r)) - R^2(S_k(r)), \tag{19}$$

where, r denotes permutations, $r = 1, 2, ..., p!$; $seqR^2(x_k|r)$ denotes the sequential sum of squares for the predictors x_k in the ordering of the predictors in the r-th permutation.
The LMG measure for k-th predictor x_k based on sequential sums of squares from all possible (p!) orderings for p predictors;

$$LMG(x_k) = \frac{1}{p!} \sum_{rpermutations} seqR^2(x_k|r) \tag{20}$$

For example, for three explanatory variables (p=3), there are six different orderings (3!) and six different estimations (sequential sum of squares) for each explanatory variable. The relative importance of each explanatory variable is the mean of the six estimations. The R package "relaimpo" developed by[71] can be used to calculate the relative importance of predictor variables in multiple regression using the LMG measure and bootstrap confidence intervals.

Time Series Decomposition: Results and Discussions

To investigate the source components influencing the lower and higher frequencies of low-cost sensor $PM_{2.5}$ data and to compare the outcomes with $PM_{2.5}$ data from reference monitors, we used the KZ filtering approach to separate the short-term and baseline components of the $PM_{2.5}$ time series at each selected EPA site and PA sensor, as well as for meteorological

variables from a nearby NOAA station, in accordance with (Equations (9) to (8)). For one combination of EPA and PA data sets (EPA site E2 and nearby PA sensor P6), the short-term and baseline components are shown in (Figure 6. The total $PM_{2.5}$ time series in (Figure 6a) for EPA site E2 has a range, from $0-30ug/m^3$, but raw data from PA sensor P6 has an almost double the range from $0-60ug/m^3$. Those high observations can be scaled down after correcting PA data as shown in the P6 corrected graph in (Figure 6). The short-term variations in $PM_{2.5}$ concentrations (Figure 6b) are due to local temporal sources such as traffic, and short-term weather variations and baseline component in (Figure 6c) shows the effect of baseline emission and climate changes as described by.[56–58]

Relative Contributions of Temporal Components

Our analysis of the PSD revealed that the differences in the short-term (high frequency) and baseline (low frequency) components in both networks, especially those below 24 hours, are distinct even after corrections. By measuring the variation in each temporal component, we can quantify the proportion of variation in each component to the total variation of the EPA and PA $PM_{2.5}$ time series data and compare the two networks to determine if the correction method was only effective for baseline data.

The results of the relative contribution of short and baseline components to total data are presented in (Figure 7). The short-term component is influenced by short-term emissions due to local sources such as traffic, or other anthropogenic activities, and intraday weather variations, the more fluctuations in the data due to such sources create more variability in the data. The relative contribution in the short-term component to total variance is greater in original PA data (Figure 7a), which are scaled down to EPA data in (Figure 7b). After correcting the data, the short-term component of the PA sensor shows a greater number of outliers and a narrow distribution, indicating that the short-term variance is largely uniform in this sensor and may be due to the capture of a local source that is independent of the sensor's location as it was observed in (Table A4). On the other hand, the EPA sensor

exhibits a broader variation in its short-term component, indicating that it captures local sources based on its location. In the original data, the EPA sensor has a greater relative contribution to the baseline compared to the PA sensor, but after correcting both distributions, the gap between them narrows. This may be due to the influence of location on the EPA sensor's broader variations. Looking more closely into numbers in (Table A5), in the short-term component, PA sensors which are located in urban areas near the lake i.e PA sensors P2, P3, P7, P8, and P9 have more relative contribution than other PA sensors. These higher values in the relative contribution of low-cost sensors to total variance at most of the locations, specifically at highly populated areas near the lake can be due to the effect of weather, as this pattern was not seen in the PSD of high frequency in (Figure 5), which are mainly signals due to anthropogenic activities. Weather can influence low-cost sensors' performance at these locations in capturing particle size and optical properties, which might cause higher variations.[27,47,48] In fig (Figure 7b), and (Table A6), significant weather dependent corrections were observed in sensors located in areas with high population density, near the lake in both the short-term and baseline components. Therefore analyzing the impact of meteorological conditions on low-cost sensor performance in both short-term and baseline components can better explain why low-cost sensors are missing high frequency signals, and identify the meteorological parameters that are most influential in affecting sensor performance in both high and low frequencies.

$PM_{2.5}$ Contributions from Meteorology and Anthropogenic Emissions

To understand and quantify the effect of meteorological conditions on the original data from PA, and to compare with the data from EPA, we used multiple linear regression models with the stepwise forward selection algorithm. We included temperature (T), relative humidity (RH), wind speed (WS), and wind direction (WD) as meteorological predictors, and short-term and baseline $PM_{2.5}$ data as response variables in our analysis. Short-term and baseline

$PM_{2.5}$ data were used as response variables, and meteorological parameters were used as the predictors in the analysis. In the final model, only variables that were significant according to the stepwise forward selection algorithm were included. This technique involves adding variables one at a time based on their p-value and determines the optimal set of parameters for the model. The model performance was compared using the R^2 values. We noted from the previous section of the relative contributions of short-term were higher in PA sensors in (Figure 7) but looking at MLR models, the meteorology has a greater impact on PA sensors' short-term variations in (Figure 8), which implies that higher variations were infact due to weather in the short-term component of low-cost sensors. The short-term component of the PA sensors in (Figure 8) have the highest R^2 around 0.33 to 0.42, which is on average 11 % more R^2 than EPA sites. Similarly, the baseline component of $PM_{2.5}$ data of PA sensors have a similar higher association with the meteorology in (Figure 8), despite its lesser contributions in the baseline component. The baseline component of PA sensors in (Figure 8) has higher R^2 around 0.23 to 0.67 which is on average 18 % more R^2 than EPA sites. This shows that weather is a higher contributor to variation of $PM_{2.5}$ in both short and baseline components in low-cost sensors as compared to reference monitors.

The anthropogenic activities can be measured through residual term of (eq. (11)), where the R^2 not explained by meteorological conditions is mainly due to anthropogenic impact on the variability of $PM_{2.5}$.[38,64] The short-term, and baseline component of all of PA sensor's $PM_{2.5}$ data has higher R^2 with meteorological parameters, which implies the PA sensors are more influenced by the weather but less responsive to anthropogenic emissions from traffic, and other sources relative to EPA instruments. In general, the low-cost PA sensor was revealed to be less responsive to local sources of pollution such as traffic, and more sensitive to regional changes such as weather conditions from both PSD and KZ filtering analysis.

Relative Importance of Predictors

We used MLR models to discover that low-cost sensors are more sensitive to weather parameters compared to reference monitors. However, we don't have a way to quantify the specific impact of each meteorological predictor using MLR because the meteorological predictors are correlated with each other. Thus we used the LMG measure to find the relative importance of each predictor, the output of $LMG(x_k)$ is partial R^2 for the variable that adds up to 1 for all predictors x_k, for $k = 1, 2, ..., n$. The $LMG(x_k)$ measure was calculated using 20, and results of $LMG(x_k)$ measure for total time series, short-term, and baseline component is presented in (Table A9), (Figure 9a), (Table A10), and (Figure 9b), (Table A11) respectively. The results of the LMG measure suggest that wind speed is the most influential factor for the time series of both PA and EPA before decomposing but if we remove wind speed, the relative humidity is the most influential factor as discussed by.[28] However, in the short-term component of almost half of the PA sensors namely (P1, P4, P5, and P6), the wind speed is the most influential factor. In the rest of the PA sensor's short-term components, the wind speed is the second most influential factor where the temperature is the first. On the other hand, the temperature is the most important factor in the short-term component of all EPA sites except E1. In baseline, the relative humidity is a consistently important factor of PA sensors except for P6, and P8 but it is the second most important factor, Whereas, in more than half of EPA sites, the WS is the most important factor. The impact of meteorological factors varies in both networks. It is worth noting that the RH is the only useful factor for baseline (low-frequency) components but not for high frequency components but corrections have been done using RH in both frequency components. On the other hand, the weather has been reported to have an impact on low-cost sensor concentrations by,[15,28,72] but only relative humidity and temperature were taken into account for corrections. Additionally, the wind speed was not taken into account in previous studies, despite the fact that sensor performance depends on the location it is deployed. The wind speed impact can be due to the PA sensor's inlet orientation 90° to the wind, upward flow, and the low inlet velocity

through the sampling holes that can result in significant aspiration losses of larger particles as reported by.[20,73] Aspiration losses are greater at higher wind speeds because it is more difficult for the larger particles to follow the streamlines into the low-velocity PA sensor's inlet. This can result in a lower concentration of larger particles entering the PA sensor than are in the ambient air. In summary, it is likely that the laser in the PA sensor is sampling a lower concentration of particles $\geq 2\ m$ diameter than in the ambient air. Based on the literature and calculations, the dominant coarse aerosol loss mechanism maybe aspiration, not internal losses.[20]

Conclusions

It is evident from a lot of studies that PA low-cost sensor data are comparatively less accurate than gold standard EPA data, but these differences have not been categorized yet. However, it is reported that the correction models of PA data have been constructed to rectify and correct PA data using EPA data and environmental conditions.[15,28,72] The $PM_{2.5}$ time series of both EPA and PA air quality networks has been examined by applying two different approaches i.e spectral theory, and time series decomposition using KZ filtering. This study suggests that analysis of $PM_{2.5}$ time series can be significant if time series are decomposed and disintegrated into low and high-frequency components. This analysis has also helped to determine that as compared to EPA, the relative contribution out of total variance of short-term components of PA sensors is higher suggesting PA sensor is sensitive to source components. However, if we look into the source components of these contributions to these variations, the PA is more sensitive to meteorological conditions. Similarly, the relative contribution out of total variance in baseline components of PA sensors is overall less than EPA, but if we look into source components of these contributions to variations, the PA sensors are again more sensitive to meteorological conditions in baseline component.

Considering these differences it can also be assumed that both networks can possibly have a difference in sensitivity to aerosol from various sources specifically PA is more sensitive to weather as compared to EPA but less efficient to capture anthropogenic emission. Our analysis used the LMG measure to find out important weather parameter influencing low-cost sensor data, we found that more than half of the PA sensors data is influenced by wind speed which is an additional finding to previous studies where relative humidity and temperature was considered as influential factor. The influence of wind speed on PA sensors was recently discussed by[20] but no one has used frequency analysis to analyze the wind speed in temporal components or suggested the wind speed as an important correction factor. Therefore, any modeling and calibration should be incorporated based on local conditions in the surrounding after decomposing the time series into different frequency components. Wind speed must be tested and included in the correction model for robust correction of low-cost sensor data. Future studies will focus on building correction models in short-term and baseline components using wind speed, temperature, and relative humidity, and traffic information.

Coding Language and Libraries

For the entire workflow (reading and organizing data, descriptive analysis, and data analyses) we used the R software (R: A Language and Environment for Statistical Computing) (version 4.2.0), along with the following libraries in our coding: readxl, dplyr, tidyr, ggplot2, car, qqplotr, kza, stats, relaimpo.

Data Availability Statement

The datasets used for this study are available at and can be accessed through the following github repository.

`https://github.com/IVijaykumar/Airquality-Spectral-Analysis.`

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

VK: Writing-original draft, conceptualization, methodology, editing, investigation, analysis.

DS: Data curation, visualization.

SG: Conceptualization, validation, editing.

WO: GIS, maps.

SS: Supervision, conceptualization, methodology, validation, editing.

SM: Supervision, conceptualization, methodology, validation, editing.

SD: Writing–review draft, conceptualization, methodology, formal analysis, project administration.

All authors contributed to the article and approved the submitted version.

Supporting Information Available

References

(1) Landrigan, P. J.; Fuller, R.; Acosta, N. J.; Adeyi, O.; Arnold, R.; Baldé, A. B.; Bertollini, R.; Bose-O'Reilly, S.; Boufford, J. I.; Breysse, P. N., et al. The Lancet Commission on pollution and health. *The lancet* **2018**, *391*, 462–512.

(2) Li, T.; Hu, R.; Chen, Z.; Li, Q.; Huang, S.; Zhu, Z.; Zhou, L.-F. Fine particulate matter (PM2. 5): The culprit for chronic lung diseases in China. *Chronic diseases and translational medicine* **2018**, *4*, 176–186.

(3) Xing, Y.-F.; Xu, Y.-H.; Shi, M.-H.; Lian, Y.-X. The impact of PM2. 5 on the human respiratory system. *Journal of thoracic disease* **2016**, *8*, E69.

(4) Samoli, E.; Analitis, A.; Touloumi, G.; Schwartz, J.; Anderson, H. R.; Sunyer, J.; Bisanti, L.; Zmirou, D.; Vonk, J. M.; Pekkanen, J., et al. Estimating the exposure–response relationships between particulate matter and mortality within the APHEA multicity project. *Environmental health perspectives* **2005**, *113*, 88–95.

(5) Ostro, B.; Broadwin, R.; Green, S.; Feng, W.-Y.; Lipsett, M. Fine particulate air pollution and mortality in nine California counties: results from CALFINE. *Environmental health perspectives* **2006**, *114*, 29–33.

(6) Lewis, T. C.; Robins, T. G.; Dvonch, J. T.; Keeler, G. J.; Yip, F. Y.; Mentz, G. B.; Lin, X.; Parker, E. A.; Israel, B. A.; Gonzalez, L., et al. Air pollution–associated changes in lung function among asthmatic children in Detroit. *Environmental Health Perspectives* **2005**, *113*, 1068–1075.

(7) Li, L.; Losser, T.; Yorke, C.; Piltner, R. Fast inverse distance weighting-based spatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter PM2. 5 in the contiguous US using parallel programming and kd tree. *International journal of environmental research and public health* **2014**, *11*, 9101–9141.

(8) Raaschou-Nielsen, O.; Andersen, Z. J.; Beelen, R.; Samoli, E.; Stafoggia, M.; Weinmayr, G.; Hoffmann, B.; Fischer, P.; Nieuwenhuijsen, M. J.; Brunekreef, B., et al. Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *The lancet oncology* **2013**, *14*, 813–822.

(9) Wu, X.; Nethery, R. C.; Sabath, M. B.; Braun, D.; Dominici, F. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study. **2020**,

(10) Zhou, X.; Josey, K.; Kamareddine, L.; Caine, M. C.; Liu, T.; Mickley, L. J.; Cooper, M.; Dominici, F. Excess of COVID-19 cases and deaths due to fine particulate matter exposure during the 2020 wildfires in the United States. *Science Advances* **2021**, *7*, eabi8789.

(11) Mondal, S.; Chaipitakporn, C.; Kumar, V.; Wangler, B.; Gurajala, S.; Dhaniyala, S.; Sur, S. COVID-19 in New York state: Effects of demographics and air quality on infection and fatality. *Science of The Total Environment* **2022**, *807*, 150536.

(12) Chaipitakporn, C.; Athavale, P.; Kumar, V.; Sathiyakumar, T.; Budisic, M.; Sur, S.; Mondal, S. COVID-19 in the United States during pre-vaccination period: Shifting impact of sociodemographic factors and air pollution. *Frontiers in Epidemiology* **2022**, *2*, 48.

(13) Noble, C. A.; Vanderpool, R. W.; Peters, T. M.; McElroy, F. F.; Gemmill, D. B.; Wiener, R. W. Federal reference and equivalent methods for measuring fine particulate matter. *Aerosol science & technology* **2001**, *34*, 457–464.

(14) Castell, N.; Dauge, F. R.; Schneider, P.; Vogt, M.; Lerner, U.; Fishbain, B.; Broday, D.; Bartonova, A. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment international* **2017**, *99*, 293–302.

(15) Ardon-Dryer, K.; Dryer, Y.; Williams, J. N.; Moghimi, N. Measurements of PM 2.5 with PurpleAir under atmospheric conditions. *Atmospheric Measurement Techniques* **2020**, *13*, 5441–5458.

(16) Wang, Y.; Li, J.; Jing, H.; Zhang, Q.; Jiang, J.; Biswas, P. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. *Aerosol Science and Technology* **2015**, *49*, 1063–1077.

(17) Commodore, A.; Wilson, S.; Muhammad, O.; Svendsen, E.; Pearce, J. Community-

based participatory research for the study of air pollution: A review of motivations, approaches, and outcomes. *Environmental monitoring and assessment* **2017**, *189*, 1–30.

(18) Woodall, G. M.; Hoover, M. D.; Williams, R.; Benedict, K.; Harper, M.; Soo, J.-C.; Jarabek, A. M.; Stewart, M. J.; Brown, J. S.; Hulla, J. E., et al. Interpreting mobile and handheld air sensor readings in relation to air quality standards and health effect reference values: Tackling the challenges. *Atmosphere* **2017**, *8*, 182.

(19) Sayahi, T.; Kaufman, D.; Becnel, T.; Kaur, K.; Butterfield, A.; Collingwood, S.; Zhang, Y.; Gaillardon, P.-E.; Kelly, K. Development of a calibration chamber to evaluate the performance of low-cost particulate matter sensors. *Environmental Pollution* **2019**, *255*, 113131.

(20) Ouimette, J. R.; Malm, W. C.; Schichtel, B. A.; Sheridan, P. J.; Andrews, E.; Ogren, J. A.; Arnott, W. P. Evaluating the PurpleAir monitor as an aerosol light scattering instrument. *Atmospheric Measurement Techniques* **2022**, *15*, 655–676.

(21) He, M.; Kuerbanjiang, N.; Dhaniyala, S. Performance characteristics of the low-cost Plantower PMS optical sensor. *Aerosol Science and Technology* **2020**, *54*, 232–241.

(22) PurpleAir, PurpleAir.: PublicLab. `https://publiclab.org/wiki/purpleair`, 2020.

(23) Kuula, J.; Mäkelä, T.; Hillamo, R.; Timonen, H. Response characterization of an inexpensive aerosol sensor. *Sensors* **2017**, *17*, 2915.

(24) Giordano, M. R.; Malings, C.; Pandis, S. N.; Presto, A. A.; McNeill, V.; Westervelt, D. M.; Beekmann, M.; Subramanian, R. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science* **2021**, *158*, 105833.

(25) Wallace, L.; Bi, J.; Ott, W. R.; Sarnat, J.; Liu, Y. Calibration of low-cost PurpleAir

outdoor monitors using an improved method of calculating PM2. 5. *Atmospheric Environment* **2021**, *256*, 118432.

(26) Stavroulas, I.; Grivas, G.; Michalopoulos, P.; Liakakou, E.; Bougiatioti, A.; Kalkavouras, P.; Fameli, K. M.; Hatzianastassiou, N.; Mihalopoulos, N.; Gerasopoulos, E. Field Evaluation of Low-Cost PM Sensors (Purple Air PA-II) Under Variable Urban Air Quality Conditions, in Greece. *Atmosphere* **2020**, *11*, 926.

(27) Kelly, K.; Whitaker, J.; Petty, A.; Widmer, C.; Dybwad, A.; Sleeth, D.; Martin, R.; Butterfield, A. Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environmental pollution* **2017**, *221*, 491–500.

(28) Barkjohn, K. K.; Gantt, B.; Clements, A. L. Development and application of a United States-wide correction for PM 2.5 data collected with the PurpleAir sensor. *Atmospheric Measurement Techniques* **2021**, *14*, 4617–4637.

(29) Bi, J.; Wildani, A.; Chang, H. H.; Liu, Y. Incorporating low-cost sensor measurements into high-resolution PM2. 5 modeling at a large spatial scale. *Environmental Science & Technology* **2020**, *54*, 2152–2162.

(30) Gupta, P.; Doraiswamy, P.; Levy, R.; Pikelnaya, O.; Maibach, J.; Feenstra, B.; Polidori, A.; Kiros, F.; Mills, K. Impact of California fires on local and regional air quality: the role of a low-cost sensor network and satellite observations. *GeoHealth* **2018**, *2*, 172–181.

(31) Kuula, J.; Mäkelä, T.; Aurela, M.; Teinilä, K.; Varjonen, S.; González, Ó.; Timonen, H. Laboratory evaluation of particle-size selectivity of optical low-cost particulate matter sensors. *Atmospheric Measurement Techniques* **2020**, *13*, 2413–2423.

(32) Tryner, J.; Mehaffy, J.; Miller-Lionberg, D.; Volckens, J. Effects of aerosol type and simulated aging on performance of low-cost PM sensors. *Journal of Aerosol Science* **2020**, *150*, 105654.

(33) Hies, T.; Treffeisen, R.; Sebald, L.; Reimer, E. Spectral analysis of air pollutants. Part 1: elemental carbon time series. *Atmospheric Environment* **2000**, *34*, 3495–3502.

(34) Marr, L. C.; Harley, R. A. Spectral analysis of weekday–weekend differences in ambient ozone, nitrogen oxide, and non-methane hydrocarbon time series in California. *Atmospheric Environment* **2002**, *36*, 2327–2335.

(35) Choi, Y.-S.; Ho, C.-H.; Chen, D.; Noh, Y.-H.; Song, C.-K. Spectral analysis of weekly variation in PM10 mass concentration and meteorological conditions over China. *Atmospheric Environment* **2008**, *42*, 655–666.

(36) Tchepel, O.; Borrego, C. Frequency analysis of air quality time series for traffic related pollutants. *Journal of Environmental Monitoring* **2010**, *12*, 544–550.

(37) Zhang, Z.; Kim, S.-J.; Ma, Z. Significant decrease of PM2. 5 in Beijing based on long-term records and Kolmogorov-Zurbenko filter approach. **2018**,

(38) Bai, H.; Gao, W.; Zhang, Y.; Wang, L. Assessment of health benefit of PM2. 5 reduction during COVID-19 lockdown in China and separating contributions from anthropogenic emissions and meteorology. *Journal of Environmental Sciences* **2022**, *115*, 422–431.

(39) IQAIR, Air quality in Chicago.: Public Database. `https://www.iqair.com/us/usa/illinois/chicago`, 2020.

(40) Bureau, U. C. US Census Bureau: Public Database. `https://www.census.gov/geo/maps-data/data/tallies/tractblock.html`, 2021.

(41) Imtiaz, S. A.; Shah, S. L. Treatment of missing values in process data analysis. *The Canadian Journal of Chemical Engineering* **2008**, *86*, 838–858.

(42) Hirabayashi, S.; Kroll, C. N. Single imputation method of missing air quality data for i-tree eco analyses in the conterminous united states. *Retrieved January* **2017**, *1*, 2021.

(43) Kim, T.; Kim, J.; Yang, W.; Lee, H.; Choo, J. Missing Value Imputation of Time-Series Air-Quality Data via Deep Neural Networks. *International Journal of Environmental Research and Public Health* **2021**, *18*, 12213.

(44) Rivera-Muñoz, L. M.; Gallego-Villada, J. D.; Giraldo-Forero, A. F.; Martinez-Vargas, J. D. Missing data estimation in a low-cost sensor network for measuring air quality: A case study in Aburrá Valley. *Water, Air, & Soil Pollution* **2021**, *232*, 1–15.

(45) Seidel, D. J.; Birnbaum, A. N. Effects of Independence Day fireworks on atmospheric concentrations of fine particulate matter in the United States. *Atmospheric Environment* **2015**, *115*, 192–198.

(46) Sun, X.; Luo, X.-S.; Xu, J.; Zhao, Z.; Chen, Y.; Wu, L.; Chen, Q.; Zhang, D. Spatio-temporal variations and factors of a provincial PM2. 5 pollution in eastern China during 2013–2017 by geostatistics. *Scientific reports* **2019**, *9*, 1–10.

(47) Zheng, T.; Bergin, M. H.; Johnson, K. K.; Tripathi, S. N.; Shirodkar, S.; Landis, M. S.; Sutaria, R.; Carlson, D. E. Field evaluation of low-cost particulate matter sensors in high-and low-concentration environments. *Atmospheric Measurement Techniques* **2018**, *11*, 4823–4846.

(48) Magi, B. I.; Cupini, C.; Francis, J.; Green, M.; Hauser, C. 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 Science and Technology* **2020**, *54*, 147–159.

(49) Romano, S.; Pichierri, S.; Fragola, M.; Buccolieri, A.; Quarta, G.; Calcagnile, L. Characterization of the PM2. 5 aerosol fraction monitored at a suburban site in south-eastern Italy by integrating isotopic techniques and ion beam analysis. *Frontiers in Environmental Science* **2022**, 1300.

(50) Dilmaghani, S. *Spectral analysis of air quality data*; University of Southern California, 2007.

(51) Hadeed, S. J.; O'Rourke, M. K.; Burgess, J. L.; Harris, R. B.; Canales, R. A. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment* **2020**, *730*, 139140.

(52) Afrifa-Yamoah, E.; Mueller, U. A.; Taylor, S.; Fisher, A. Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications* **2020**, *27*, e1873.

(53) Wijesekara, W.; Liyanage, L. Comparison of imputation methods for missing values in air pollution data: Case study on sydney air quality index. Future of Information and Communication Conference. 2020; pp 257–269.

(54) Saputra, M.; Hadi, A.; Riski, A.; Anggraeni, D. Handling Missing Values and Unusual Observations in Statistical Downscaling Using Kalman Filter. Journal of Physics: Conference Series. 2021; p 012035.

(55) Potyrailo, R. A.; Go, S.; Sexton, D.; Li, X.; Alkadi, N.; Kolmakov, A.; Amm, B.; St-Pierre, R.; Scherer, B.; Nayeri, M., et al. Extraordinary performance of semiconducting metal oxide gas sensors using dielectric excitation. *Nature Electronics* **2020**, *3*, 280–289.

(56) Rao, S. T.; Zurbenko, I. G. Detecting and tracking changes in ozone air quality. *Air & waste* **1994**, *44*, 1089–1092.

(57) Rao, S.; Zurbenko, I.; Neagu, R.; Porter, P.; Ku, J.; Henry, R. Space and time scales in ambient ozone data. *Bulletin of the American Meteorological Society* **1997**, *78*, 2153–2166.

(58) Wise, E. K.; Comrie, A. C. Meteorologically adjusted urban air quality trends in the Southwestern United States. *Atmospheric Environment* **2005**, *39*, 2969–2980.

(59) Fang, C.; Qiu, J.; Li, J.; Wang, J. Analysis of the meteorological impact on PM2.5 pollution in Changchun based on KZ filter and WRF-CMAQ. *Atmospheric Environment* **2022**, *271*, 118924.

(60) Sá, E.; Tchepel, O.; Carvalho, A.; Borrego, C. Meteorological driven changes on air quality over Portugal: a KZ filter application. *Atmospheric Pollution Research* **2015**, *6*, 979–989.

(61) Eskridge, R. E.; Ku, J. Y.; Rao, S. T.; Porter, P. S.; Zurbenko, I. G. Separating different scales of motion in time series of meteorological variables. *Bulletin of the American Meteorological Society* **1997**, *78*, 1473–1484.

(62) Milanchus, M. L.; Rao, S. T.; Zurbenko, I. G. Evaluating the effectiveness of ozone management efforts in the presence of meteorological variability. *Journal of the Air & Waste Management Association* **1998**, *48*, 201–215.

(63) Botlaguduru, V. S.; Kommalapati, R. R.; Huque, Z. Long-term meteorologically independent trend analysis of ozone air quality at an urban site in the greater Houston area. *Journal of the Air & Waste Management Association* **2018**, *68*, 1051–1064.

(64) Li, K.; Jacob, D. J.; Liao, H.; Shen, L.; Zhang, Q.; Bates, K. H. Anthropogenic drivers of 2013–2017 trends in summer surface ozone in China. *Proceedings of the National Academy of Sciences* **2019**, *116*, 422–427.

(65) Zhai, S.; Jacob, D. J.; Wang, X.; Shen, L.; Li, K.; Zhang, Y.; Gui, K.; Zhao, T.; Liao, H. Fine particulate matter (PM 2.5) trends in China, 2013–2018: Separating contributions from anthropogenic emissions and meteorology. *Atmospheric Chemistry and Physics* **2019**, *19*, 11031–11041.

(66) Ibarra-Berastegi, G.; Madariaga, I.; Elıas, A.; Agirre, E.; Uria, J. Long-term changes of ozone and traffic in Bilbao. *Atmospheric Environment* **2001**, *35*, 5581–5592.

(67) Lindeman, R. H. *Introduction to bivariate and multivariate analysis*; 1980.

(68) Kruskal, W. Relative importance by averaging over orderings. *The American Statistician* **1987**, *41*, 6–10.

(69) Bi, J. A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies* **2012**, *27*, 87–101.

(70) Grömping, U. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics* **2015**, *7*, 137–152.

(71) Grömping, U. Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software* **2007**, *17*, 1–27.

(72) Mei, H.; Han, P.; Wang, Y.; Zeng, N.; Liu, D.; Cai, Q.; Deng, Z.; Wang, Y.; Pan, Y.; Tang, X. Field evaluation of low-cost particulate matter sensors in Beijing. *Sensors* **2020**, *20*, 4381.

(73) Hangal, S.; Willeke, K. Overall efficiency of tubular inlets sampling at 0–90 degrees from horizontal aerosol flows. *Atmospheric Environment. Part A. General Topics* **1990**, *24*, 2379–2386.

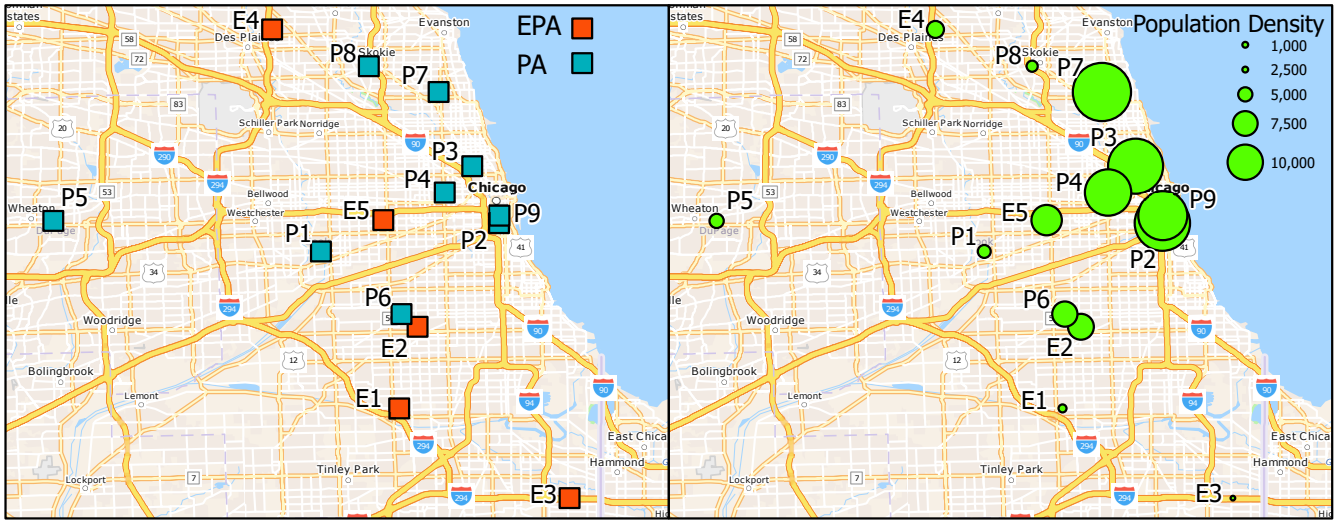

Figure 1: (a) EPA and PA sampling locations with (b) population density in the block defined by US census bureau in Cook County IL

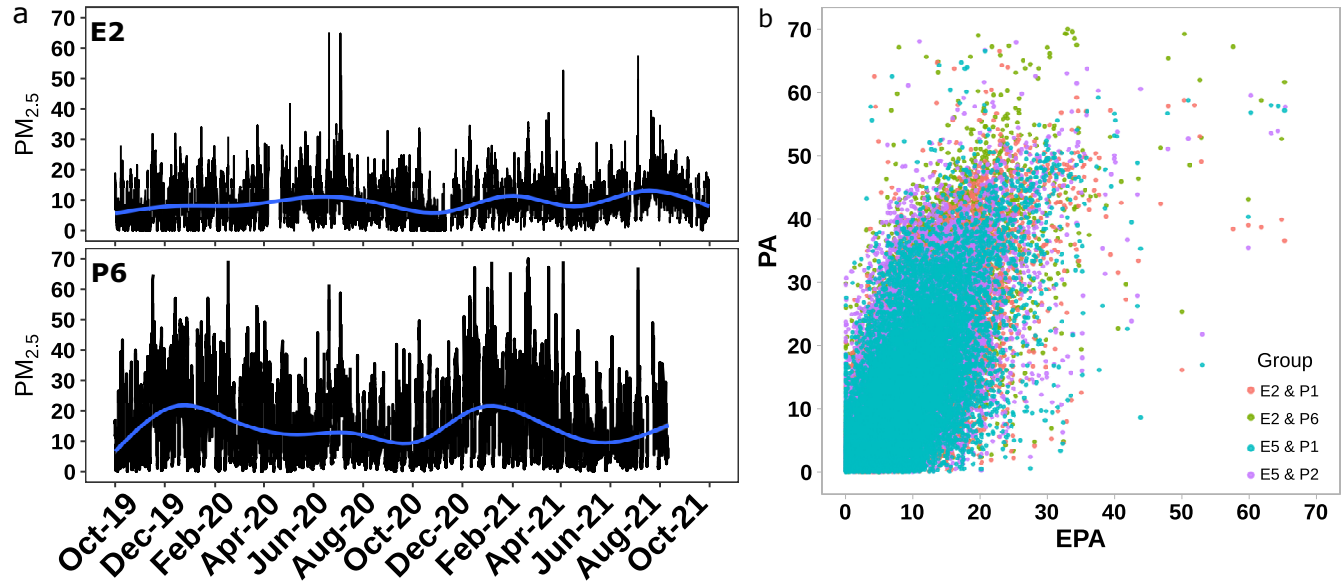

Figure 2: (a) Time series plot of hourly $PM_{2.5}$ measurements from EPA site E2 and PA sensor P6, (b) the scatter plot of hourly $PM_{2.5}$ measurements from several neighboring pairs of EPA and PA sites

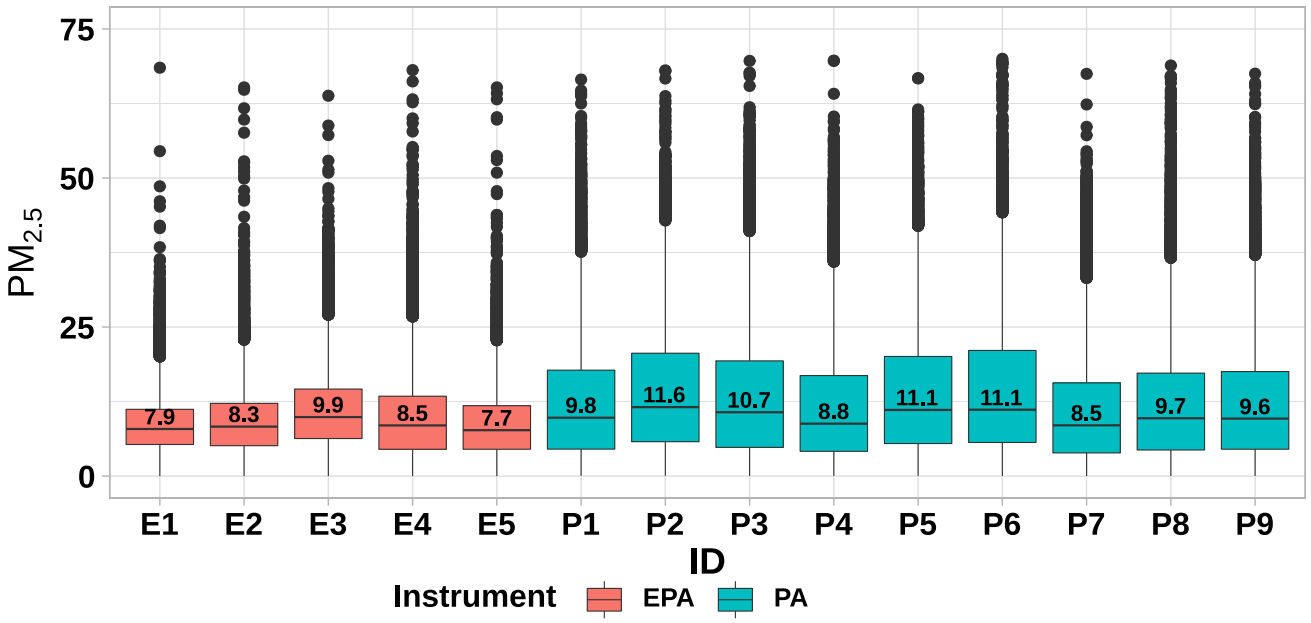

Figure 3: Box plots representing hourly $PM_{2.5}$ measurements from EPA sites and PA sensors located in Cook County IL

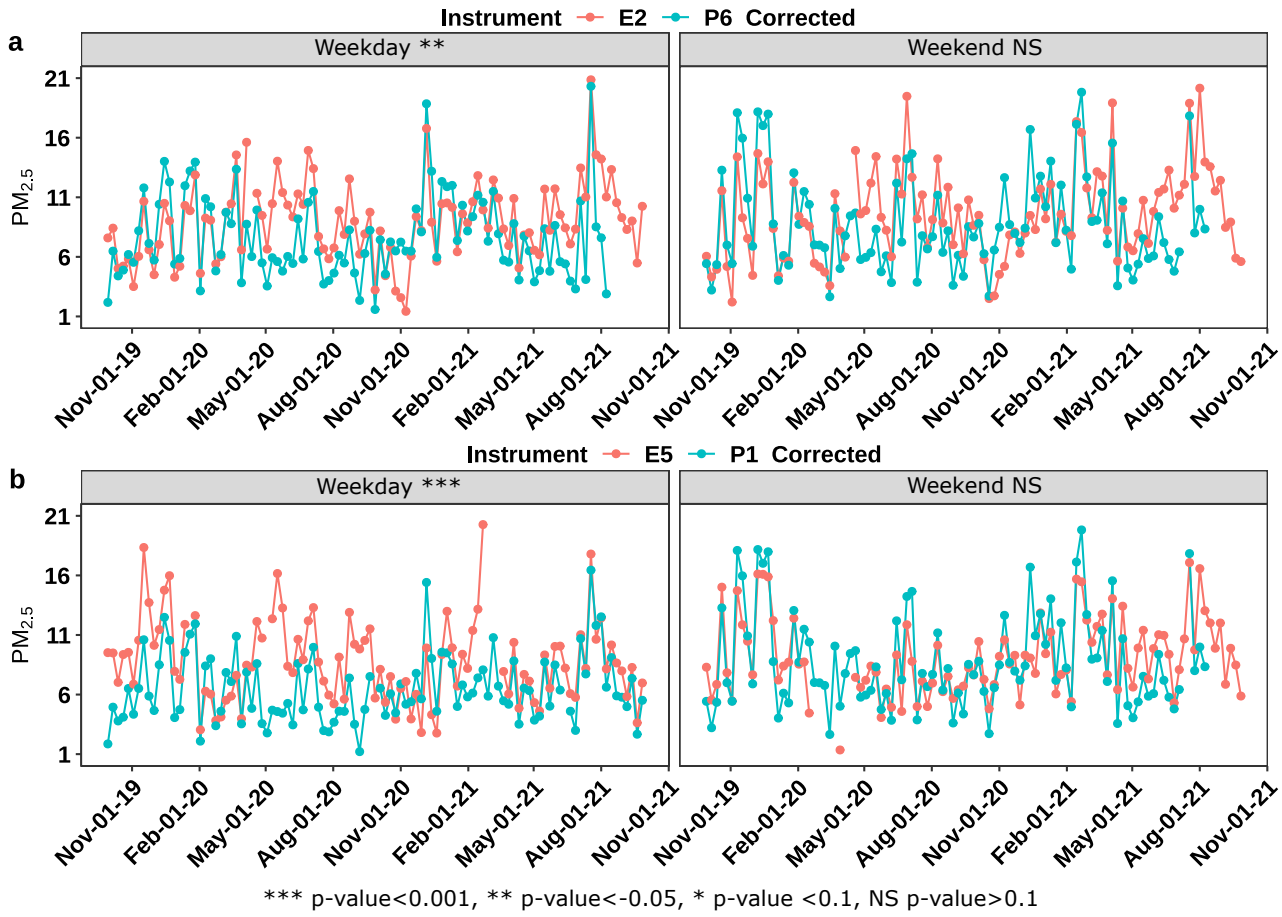

Figure 4: Corrected PA sensor $PM_{2.5}$ measurements in weekdays and weekends in with nearby located EPA sites (a) E2, P6, and (b) E5, P1 with t-test statistic

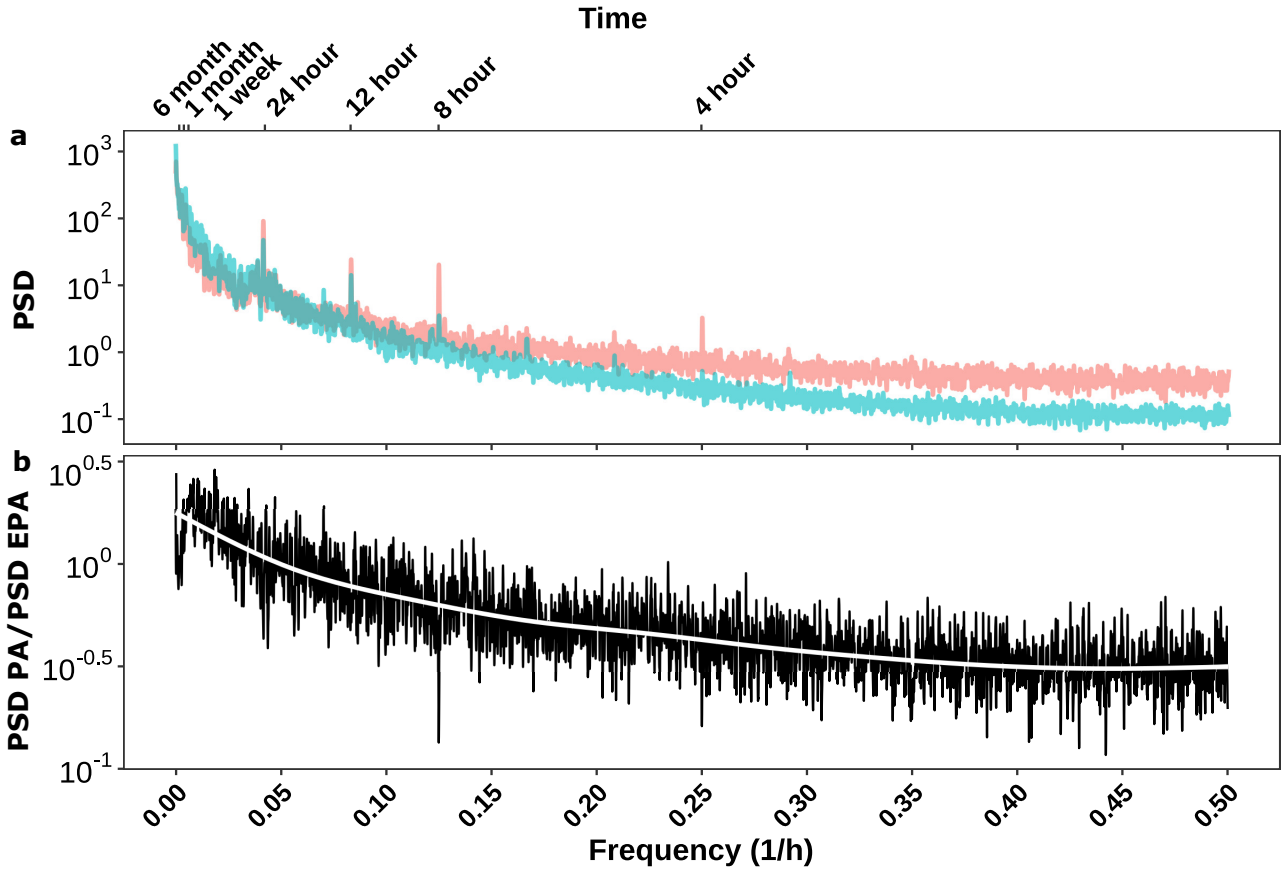

Figure 5: (a) Average PSD of $PM_{2.5}$ data from all EPA sites, and all PA sensors with baseline peaks of one week to 6 months, and short-term peaks of 24 hours to 4 hours, (b) average ratio of PA PSD to EPA PSD

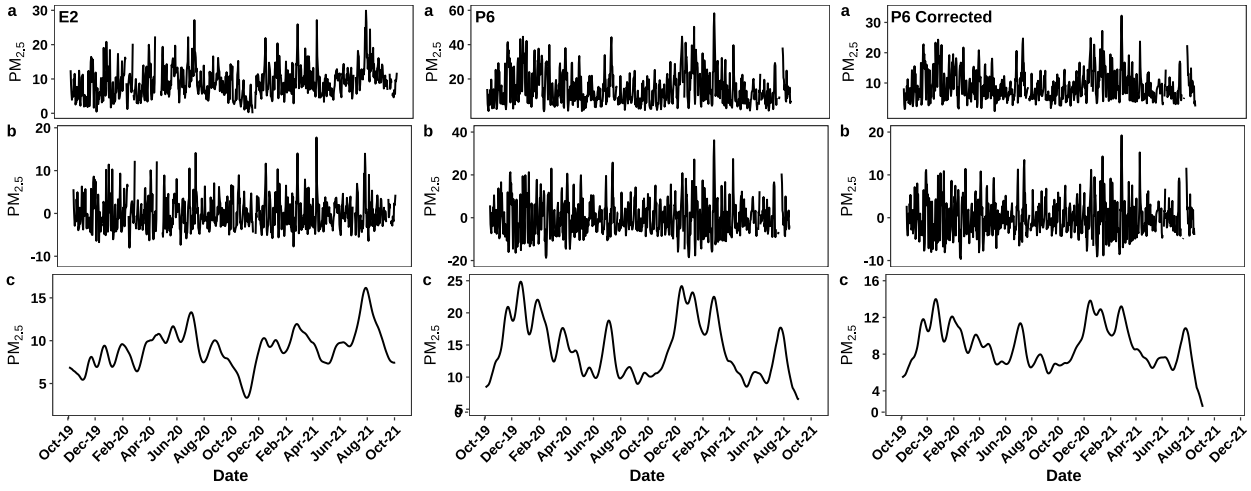

Figure 6: (a) Measured $PM_{2.5}$ time series data from EPA site E2, and corrected $PM_{2.5}$ time series data from PA sensor P6 (b) Extracted short-term component for the two data sets (c) Extracted baseline component for the two data sets

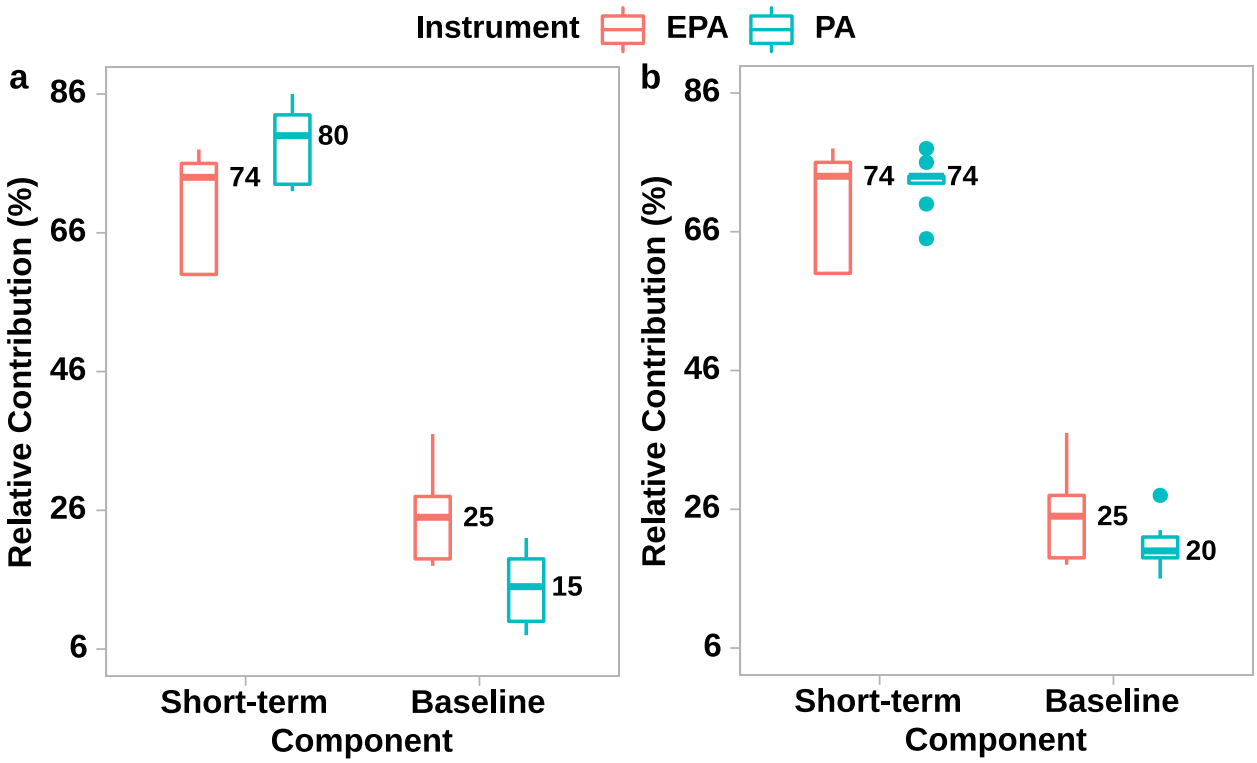

Figure 7: Relative contribution of temporal components to total variations of $PM_{2.5}$; (a) short-term, and baseline components of original PA data with EPA data (b) short-term, and baseline components of corrected PA data with EPA data

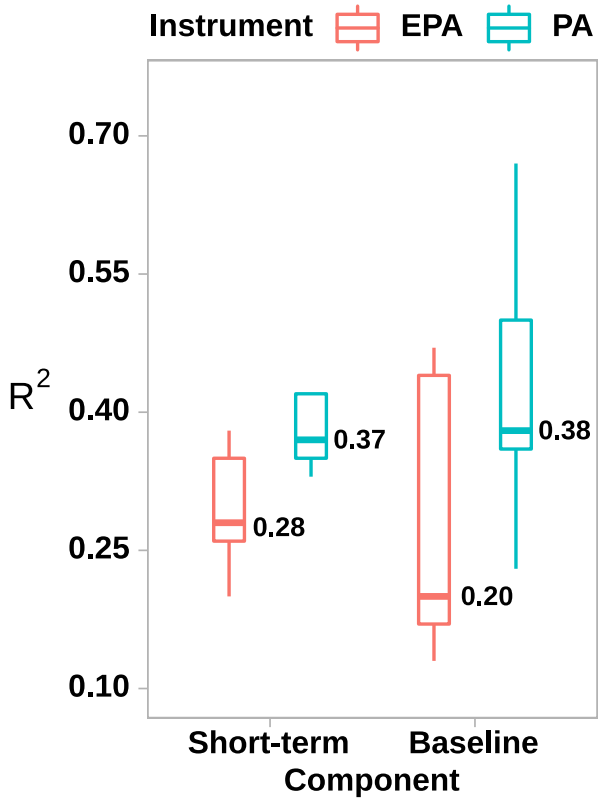

Figure 8: Impact of meteorological conditions in short-term and baseline component on both EPA and PA $PM_{2.5}$ data

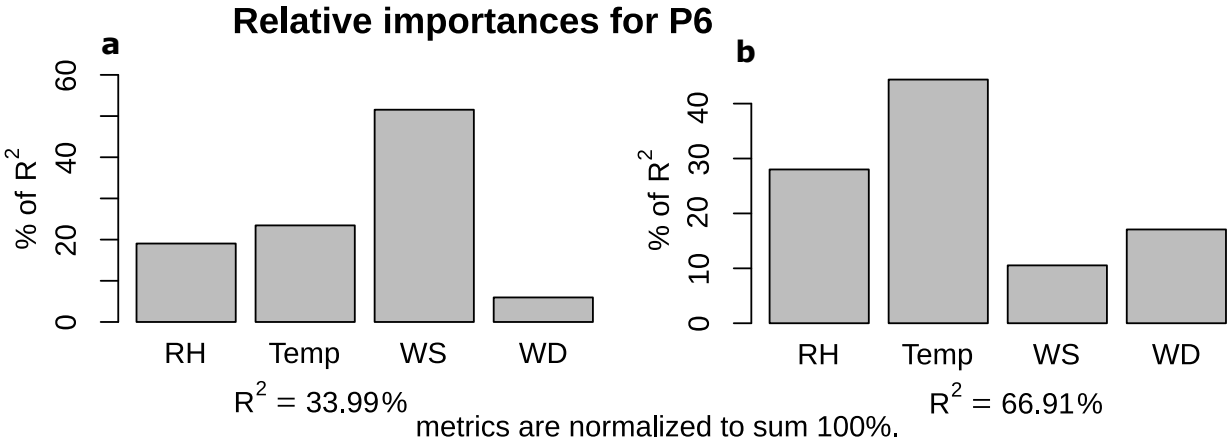

Figure 9: Relative Importance of meteorological variables for PA sensor P6 in (b) short-term and (b) baseline component of $PM_{2.5}$ data