Evaluation methods for low-cost particulate matter sensors

Jeffrey K. Bean

Phillips 66, Bartlesville, OK 74003, United States

Correspondence to: jeff.bean@p66.com

Abstract. Understanding and improving the quality of data generated from low-cost sensors is a crucial step in using these sensors to fill gaps in air quality measurement and understanding. This paper shows results from a 10-month long campaign that included side-by-side measurements and comparison between EPA-approved reference instruments and low-cost particulate matter sensors in Bartlesville, Oklahoma. At this rural site in the Midwestern United States the instruments typically encountered only low (under 20 $\mu g/m^3$) concentrations of particulate matter, however higher concentrations (50-400 $\mu g/m^3$) were observed on three different days during what were likely agricultural burning events. This study focused on methods for understanding and improving data quality for low-cost particulate matter sensors. The data offered insights on how averaging time, choice of reference instrument, and the observation of higher pollutant concentrations can all impact performance indicators (R^2 and root mean square error) for an evaluation. The influence of these factors should be considered when comparing one sensor to another or when determining whether a sensor can produce data that fits a specific need. Though R^2 and root mean square error remain the dominant metrics in sensor evaluations, an alternative approach using a prediction interval may offer more consistency between evaluations and a more direct interpretation of sensor data following an evaluation. Ongoing quality assurance for sensor data is needed to ensure data continues to meet expectations. Observations of trends in linear regression parameters and sensor bias were used to analyze calibration and other quality assurance techniques.

1 Introduction

Traditional particulate matter measurements are taken using stationary instruments that cost tens, if not hundreds of thousands of dollars. The high cost limits data collection to certain entities such as government agencies and research institutions that take measurements through field campaigns and through networks of stationary sensors. However, research has shown that these traditional measurements do not capture the spatial variations in particulate matter (Apte et al., 2017; Mazaheri et al., 2018). Low-cost sensors are increasingly being used in attempts to better map the spatial and temporal variations in particulate matter (Ahangar et al., 2019; Bi et al., 2020; Gao et al., 2015; Li et al., 2020; Zikova et al., 2017). Governments, citizen scientists, and device manufacturers are connecting these low-cost devices to build large air quality measurement networks. Understanding and improving the quality of this type of data is crucial in determining its appropriate applications. Though there has been a significant amount of research in recent years on the topic (Feenstra et al., 2019; Holstius

30 et al., 2014; Jiao et al., 2016; Malings et al., 2020; Papapostolou et al., 2017; Williams et al., 2019; Williams et al., 2018), there is an ongoing effort to understand 1) how to concisely describe the performance of a low-cost sensor, and 2) what best practices can maximize data quality while keeping costs down. Rather than presenting evaluation results for specific low-cost sensors, this study focuses on evaluation methods that can improve the use of all low-cost sensors.

Much of the performance characterization has focused on correlation (R^2) and root mean square error (RMSE)
35 (Karagulian et al., 2019; Williams et al., 2019; Williams et al., 2018). However, these performance metrics can be influenced by the conditions during a sensor evaluation. Higher concentration episodes during an evaluation can impact R^2 and RMSE (Zusman et al., 2020). The choice of instrument for comparison can also be a factor (Giordano et al., 2021; Mukherjee et al., 2017; Stavroulas et al., 2020; Zusman et al., 2020) as some reference instruments are more inherently similar to the low-cost sensors and will likely show better comparisons. Finally, the averaging time can be a significant factor in performance metrics
40 (Giordano et al., 2021). Some of these evaluation inconsistencies (instrument comparison choice, averaging time) can be mitigated by implementing standard evaluation protocols. Other inconsistencies, such as influence of observed concentration range, may be better managed by shifting away from R^2 and RMSE. While these metrics can be useful in comparing one sensor to another, they are not as useful in interpreting future sensor measurements. An alternative evaluation method using prediction interval is outlined in Sect. 3.2.

45 A past evaluation of a sensor is a useful predictor of future data quality, but quality assurance techniques are needed to ensure data quality continues to meet expectations. Calibrations are an important component of quality assurance. ~~The root of a calibration for low cost particulate matter sensors is simple:~~During a calibration low-cost~~–~~ sensors and reference instruments measure the same mass of air for a period and then adjustments are made to better align sensor measurements. Though laboratory comparisons would be more consistent, only location-specific field comparisons are able to capture the full
50 variety of particle sizes and compositions that a sensor will encounter once deployed (Datta et al., 2020; Jayaratne et al., 2020). However, there are different calibration techniques with varying cost (Hasenfratz et al., 2015; Holstius et al., 2014; Malings et al., 2020; Stanton et al., 2018; Williams et al., 2019) and the needed requirements are not always clear for a successful field calibration. This technical gap is explored in this study by evaluating changes in linear regression parameters over time and their dependence on the amount of data that is included. A recent publication from the United States Environmental Protection
55 Agency (Duvall et al., 2021) begins to address these issues and standardize evaluation practices, though they acknowledge that this is an evolving topic.

This study of low-cost particulate matter sensors was conducted in a rural area of the Midwestern United States (Bartlesville, Oklahoma). This area is interesting for evaluation as it typically sees lower concentrations of $PM_{2.5}$ but occasionally encounters much higher concentrations, such as during agricultural burning events. Data was collected for a total
60 of 10 months in 2018 and 2019. This large, mixed dataset allowed exploration of both evaluation and quality assurance techniques. These techniques are crucial in finding ways to fill existing knowledge gaps in spatial and temporal air quality variation using data from low-cost sensors.

2 Experimental methods and materials

2.1 Site description

65   Data was collected at the Phillips 66 Research Center in Bartlesville, Oklahoma. Bartlesville is approximately 47 miles north of Tulsa and has a population of approximately 36,000. Measurements were collected over 9 months from May 2018 to January 2019, and for one additional month in April 2019. Particulate matter concentrations were typically low (under 20 µg/m^3 for 1-hour averaged data), which is characteristic of many rural areas. The exception is during times when agricultural burning emissions are observed, in which case concentrations of PM$_{2.5}$ observed were as high as 400 µg/m^3 for 1-hour averaged
70 data.

2.2 Instrumentation

        Reference measurements were collected using a Met One Beta Attenuation Monitor 1020 (BAM) and a Teledyne T640 (T640). Though both instruments are considered Federal Equivalent Methods (FEM) by the United States Environmental Protection Agency (EPA), the BAM uses beta ray attenuation to measure the mass of PM$_{2.5}$ collected on filter tape, while the
75 T640 uses an optical counting method that is more similar to the method used by the low-cost particulate matter sensors. The BAM was used throughout the entire period of evaluation but often struggled to maintain sample relative humidity below 35%, which is a FEM requirement. Any data from the BAM which did not meet this relative humidity criterion was removed prior to analysis. The T640 was only available for approximately one month of comparison in April 2019, but still provided a useful dataset as it employs a different sampling technique and samples at a higher frequency (one sample per minute).
80   Low-cost particulate matter sensors were evaluated by comparing samples taken within 12 feet of the reference instruments. Four brands of low-cost (less than $300/sensor) nephelometric-type particulate matter sensors were initially evaluated through comparison with reference measurements for one month in May 2018. The specific brands of sensors are not identified here, as the primary goal of this work is exploration of different methods of evaluation. Each sensor provided measurements in near real-time, but was averaged to 1-hour intervals for comparison with the BAM. The correlation (R^2)
85 between sensors and the BAM during May 2018 testing was 0.67, 0.53, 0.24, and 0.12 for the four different sensors. Though this one-month test may not have definitively identified the best of the four sensors, it allowed selection of a useful sensor for additional testing and method exploration, which is the primary purpose of this study. After this initial testing, long-term testing continued only for the best performing sensor (R^2 = 0.67), which was evaluated over a total of 10 months. The remainder of this study focuses on the 10 months of data from the best performing sensor.

3 Results and discussion

3.1 Data quality

Eight replicas of the best performing brand of low-cost particulate matter sensor were placed with the BAM from May 2018 to January 2019. The overall correlation (R^2) between these sensors and the BAM during this time was as high as 0.55 for sensors that performed well, but for a few sensors the correlation was 0.15 or lower. Upon inspection it was found that poor correlation often resulted from just a small handful of odd measurements. For example, one sensor logged 10 measurements of 500-2000 $\mu g/m^3$ during a time when the BAM reported below 10 $\mu g/m^3$. A relatively inexpensive way to identify these erroneous data points is by collocating two sensors in any deployment. Figure 1 shows impact to data quality when pairing sensors together with increasingly stringent requirements for data agreement. Each point in the figure is the correlation between reference data for the entire time period and the average of any two data points that meet the allowable different requirement shown on the x-axis. For perspective, 95% of sensor measurements were below 30 $\mu g/m^3$. No data is removed for points on the left side of the graph, but the average of two sensors is often much less correlated with reference measurements. An allowed difference between sensor measurements of 50 $\mu g/m^3$ results in only a small portion of the data being removed (bottom of Fig. 1) but has a significant positive impact on RMSE and correlation with the reference. As more stringent data agreement requirements are put in place (moving to the right in Fig. 1) there are not significant improvements in correlation. Thus, a pair of sensors and a loose requirement for data agreement may serve as a quality assurance check to greatly improve data quality and spot erroneous measurements. This can also help identify defective sensors that need to be replaced (Bauerová et al., 2020).

Though only an absolute ($\mu g/m^3$) data agreement is considered here, a requirement for percentage agreement between sensors could also be considered (Barkjohn et al., 2021; Tryner et al., 2020). R^2 begins to decrease with stricter data agreement requirements in Fig. 1 (5, 2, and 1 $\mu g/m^3$), which is the result of higher concentration measurements being unnecessarily removed. A combination of a percentage and absolute data agreement would prevent this data from being unnecessarily removed. However, in this case just a generous absolute data agreement requirement (50 $\mu g/m^3$) works well since this method easily catches the most egregious measurements without deleting data unnecessarily. Agreement between sensor measurements will likely depend on both the sensor and the type of measurement, but this type of analysis can be performed inexpensively at any location to determine what type of data agreement is necessary to filter out any odd measurements.

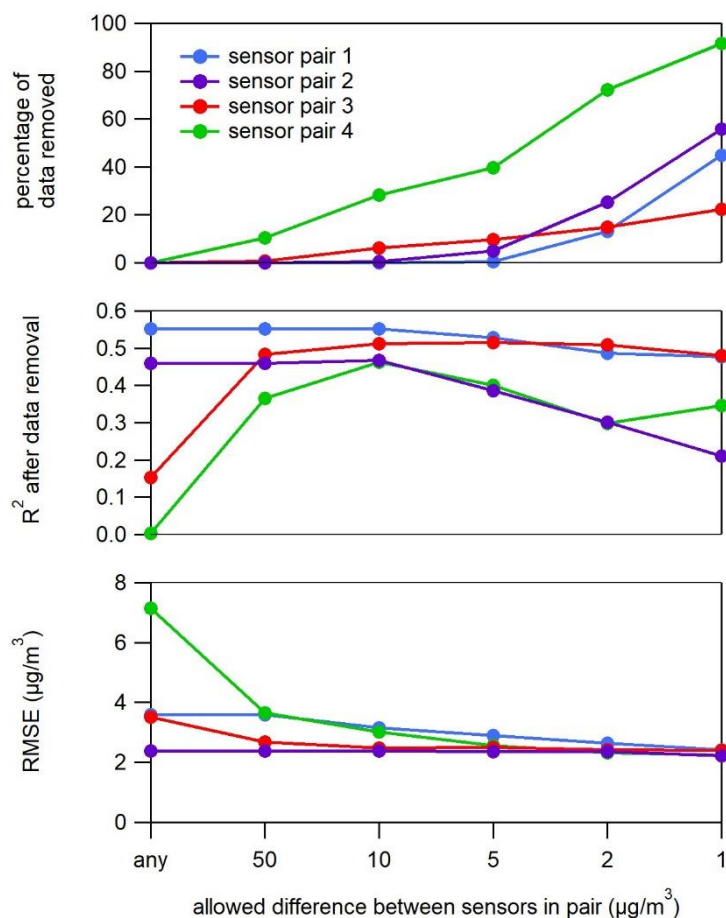



Figure 1. The eight sensor replicas were divided into four pairs with different measurement agreement requirements for the data from the sensors in each pair (first averaged to 1-hour intervals). The x-axis shows the allowed difference between paired measurements. The top figure shows the percentage of data that is removed from the sensor pair for different data agreement

120 **requirements. The ~~bottom~~ middle figure shows how R^2 between the sensor pair (average of the pair) and reference measurements changes ~~as~~ when these disagreeing data points are removed. The bottom figure shows how RMSE is impacted as disagreeing data points are removed.**

Figure 1 shows that good agreement was observed between measurements from duplicate sensors. Figure 2A supports

125 this observation, showing close correlation between bias, defined as $C_{sensor} - C_{reference}$, between duplicates of a sensor. Figure 2

(y axes) shows that sensor bias typically ranges from 10 µg/m³ below to 40 µg/m³ above the reference measurement. It is

noteworthy that sensor measurements correlated so closely from one sensor to another (Figure 2A) despite such a large range

of variation from reference measurements. Comparison between only two sensors is shown in Fig. 2A for simplicity, but

similarly strong correlation was observed for other pairs of sensors. Others have observed similar correlation between

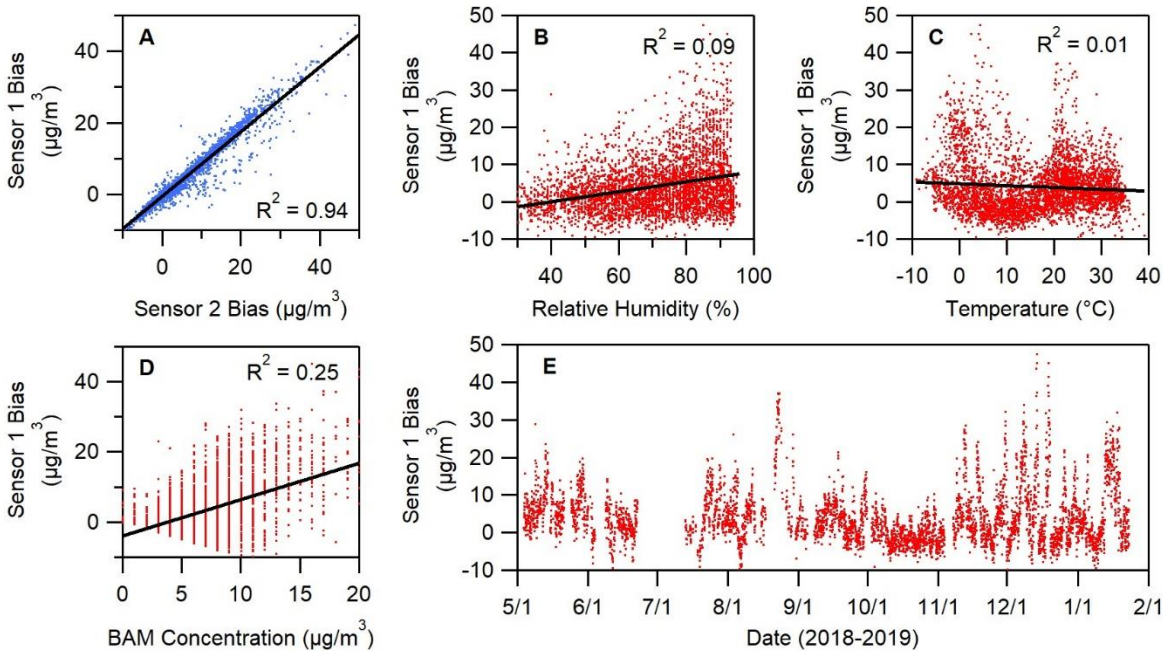
130 measurements from duplicates of low-cost sensors (Feenstra et al., 2019; Zamora et al., 2020). Because these measurements

came from separate sensors (same brand and model) in separate housing, the large, yet correlated bias suggests that an external,

correctable factor influences how accurately sensor measurements correlate to the reference measurement. However, Fig. 2B-E shows that only slight correlation was seen between this bias and easily observable external factors like humidity, temperature, particulate matter concentration, or time. Solar radiation, wind direction, wind speed, and rain were also

135 measured, and ~~no correlation~~similarly little correlation was observed between bias and these factors. Other research has observed improved PM$_{2.5}$ predictions when parameters such as temperature and relative humidity are included in analysis (Datta et al., 2020; Di Antonio et al., 2018; Gao et al., 2015; Kumar and Sahu, 2021; Levy Zamora et al., 2019; Zou et al., 2021b). Though some improvement in PM$_{2.5}$ predictions is still possible using the same approach here, Figure 2 suggests that these meteorological parameters are not the primary cause of the similar bias that is observed from one sensor to another.

140   The lack of correlation suggests a different external factor, such as particle properties, may influence sensor measurements. Previous research has observed the impact of particle composition on the accuracy of low-cost sensors (Giordano et al., 2021; Kuula et al., 2020; Levy Zamora et al., 2019). Particle size has also been observed to influence measurements (Stavroulas et al., 2020; Zou et al., 2021a). Very small particles go undetected and other particles can be incorrectly sized by the optical detectors used in low-cost particulate matter sensors. Regardless of the cause in varying, yet

145 correlated sensor response, data here suggests that low-cost measurements of meteorology will not be sufficient to improve low-cost sensor data. It may be possible to improve sensor data through measurements of particle properties, but the high cost of these measurements would undo the benefit of the low sensor price.

150 **Figure 2. (A) shows the correlation between the bias ($C_{sensor} - C_{reference}$) of two sensor replicas. (B)-(D) show correlation between sensor bias and meteorological factors. (E) shows that bias varies over time but not in a consistent pattern. Measurement gaps in (E) are the result of the BAM not meeting its internal relative humidity specification (35%).**

3.2 Performance evaluation

The T640 was available for comparisons for approximately one month in April 2019. In contrast to the BAM, which reports data only in 1-hour intervals, the T640 was programmed to report a measurement every minute, allowing comparisons to the high time-resolution data offered by sensors. In addition, the month of April provided useful comparisons as elevated concentrations of particulate matter were observed on three different days, likely due to nearby agricultural burning. Under different evaluation conditions, the R^2 and RMSE of a linear regression were calculated. Impacts to R^2 and RMSE from different averaging time, reference instrument, and higher concentrations are shown in Fig. 3.

One-hour averaged data comparisons between a sensor and the T640 are used as a baseline for comparison (Fig. 3B). To highlight differences, the three days that included high particulate matter concentrations were not included, except when indicated (Fig. 3E). Figure 3A-C shows that less averaging (higher time-resolution data) negatively impacted R^2 and RMSE while more averaging improved both these metrics. Though higher time-resolution is often considered an advantage of low-cost sensors compared to some reference measurements, this data shows that time resolution comes at a cost in the ability of sensors to predict concentration. If a specific averaging period becomes the standard for evaluations, such as 1-hour averaging (Duvall et al., 2021), then sensor users will need to carefully assess how the quality of data changes if it is necessary to average to a different interval, such as 5 or 1-minute intervals.

Figure 3D shows the difference in comparison due to reference instrument. The EPA considers both the BAM and the T640 as FEM instruments, but both the R^2 and the RMSE are negatively impacted when evaluating with the BAM instead of the T640. The T640 is an optical particle counter, which dries and then counts individual particles. It differs from the evaluated low-cost sensors, which take a nephelometric measurement un-dried, bulk particle concentrations. However, the T640 is still an optical measurement that and is more similar to the method used by low-cost particulate matter sensors. This comparison shows that the choice of a reference instrument in evaluation of sensors can impact results.

Figure 3E is the only chart on Fig. 3 that includes data from the three days in April 2019 in which higher concentrations of particulate matter were observed. Most 1-hour measurements were below 50 µg/m^3 during this month, but within these three days, there were 24 observations of 1-hour particulate matter concentrations between 50 and 400 µg/m^3. During this month, about 565 1-hour data points were captured, but Fig. 3E shows the influence of just a few measurements at higher concentrations. R^2 in the baseline chart is 0.80, but with the addition of these points it increases all the way to 0.97. In contrast, RMSE increases from 3.2 to 5.9 µg/m^3 with the addition of this higher concentration data, suggesting decreased sensor performance. At high concentrations, small percent differences between sensor and reference measurements translate into larger errors when expressed in µg/m^3.

Figure 3 shows that the circumstances surrounding an evaluation such as averaging time, reference instrument, and the presence of high particulate matter concentrations can be very influential on the performance results for a sensor, even with other factors being held equal. The averaging time and the choice of reference instrument could become smaller issues as standard evaluation procedures are developed, such as those recently proposed by the United States Environmental Protection

Agency (Duvall et al., 2021). However, the influence of concentration range on R^2 and RMSE is a challenge in evaluating sensors, as it suggests that evaluation location and random circumstances such as high concentration events are influential on evaluation results. In addition, R^2 and RMSE are not particularly suited to interpreting a new measurement from a sensor once an evaluation has been completed. As an alternative to R^2 and RMSE, a prediction interval can be considered as an evaluation tool for low-cost sensors.

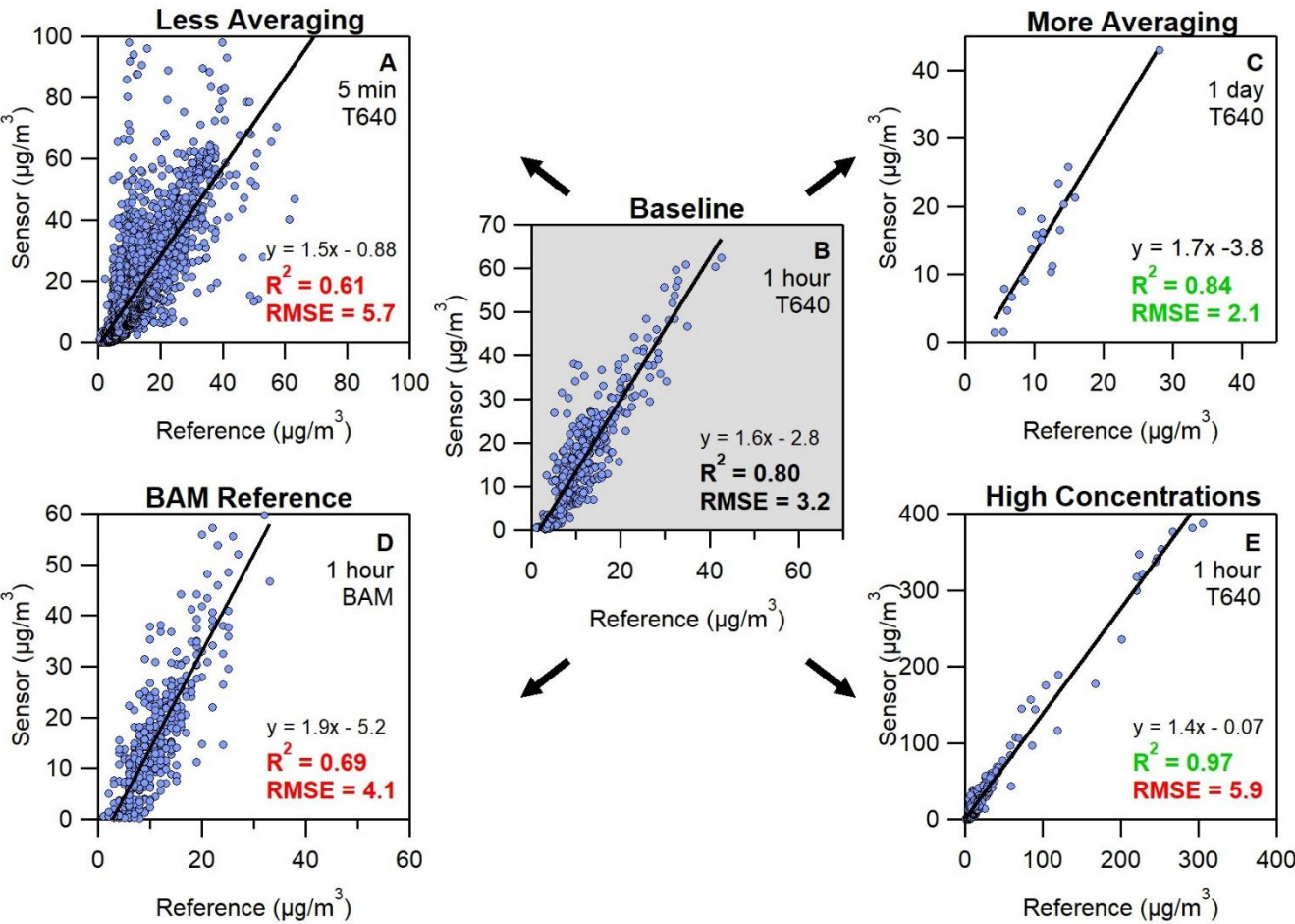



Figure 3. The changes to R^2 and RMSE from a baseline condition depending on evaluation conditions. (B) shows the baseline of 1-hour averaged data, T640 as a reference instrument, and without three days of high concentrations. (A) and (C) show the impact of less and more data averaging, respectively. (D) shows the impact of switching to the BAM instead of the T640. (E) shows the effect of high concentration by including three additional days during which higher concentrations were observed.

A prediction interval (PI) between sensor and reference data offers a robust, yet straightforward interpretation of sensor measurements. A 95% PI suggests that one can be 95% confident that any new measurement will be within its bounds, thus a new sensor reading can be converted to a range of estimates with statistical confidence. The width of this PI is a useful way to show the performance of the sensor. Though a PI is calculated from a linear regression just like R^2 and RMSE, it requires a few extra details. The most important of these details is that the residuals of the linear fit need to be even across the

range of observed values. In the data described here, and likely for other low-cost sensor data, this will require a transformation to the data.

Figure 4 shows the PI for data that was collected at 5-minute intervals in April 2019. After fitting a linear regression to this data, it was found that residuals generally increased with increasing concentration, suggesting that bias can be higher as concentrations increase. This is also seen in the Fig. 2D comparison between low-cost sensor bias and BAM measurements. In order to ensure a correct linear fit and PI, these trends in residuals were eliminated by transforming both the sensor and reference data prior to the fit. Through examination it was found that residual trends were best eliminated by raising both the sensor and reference data to the 0.4 power. Future applications of this method to various sensors may find different powers or transformation methods are needed to eliminate trends that are observed in residuals. Even duplicates of the same sensor may require different transformations if taking measurements in different locations. A detailed analysis of residuals is an important step in all model development. In addition to the transformation, measurements less than 70 $\mu g/m^3$ were randomly sampled to capture an equal number of data points below and above 70 $\mu g/m^3$. This sampling did not change the outcome significantly, but helped ensure that the linear regression and PI were equally weighted to the entire range of observed measurements. Before this sampling only 5% of the data was above 70 $\mu g/m^3$. The R software suite was used to calculate the linear regression and PI for the transformed data and these curves were then reverse-transformed (raised to the 0.4^{-1} power) to create the graph shown in Fig. 4.

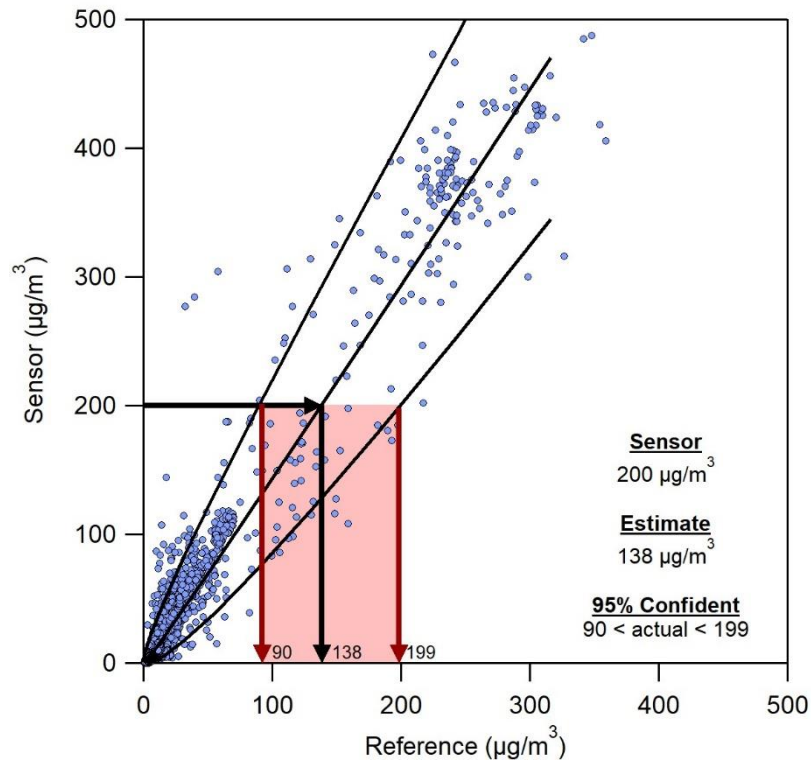



Figure 4. An example of a prediction interval evaluation for 5-minute data from a single sensor in April 2019 that includes periods of high concentration. The curved lines are the upper and lower limits of the 95% prediction interval. A visual interpretation of a new sensor measurement of 200 μg/m³ is also shown.

Once data has been analyzed using the method shown in Fig. 4, the interpretation of new sensor data becomes easy. As shown in Fig. 4, a new sensor measurement at 200 μg/m³ suggests the most likely actual concentration is 138 μg/m³. However, more importantly, it can be said with 95% confidence that the true concentration is between 90 and 199 μg/m³. Replicating this analysis for similar data from another sensor suggests that for a sensor reading of 200 μg/m³ the most likely concentration is 148 μg/m³, with a 95% prediction interval between 97 and 213 μg/m³. These estimates are close to those from the first sensor, showing again that precision is good for these sensors. Data from multiple replicas of a sensor could also be combined to provide a more general prediction interval that applies broadly to sensors of that type. If sensors are not very precise then the combined data will have higher variance, which will lead to an appropriately broader prediction interval and higher uncertainty in estimates for future sensor measurements. Any non-linearity in sensor response (Zheng et al., 2018) will also result in a broader prediction interval if it has not been accounted for in the calibration model.

A range can be provided for any new sensor measurement following this method, with 95% confidence. In some cases that uncertainty range may limit the ability to distinguish one concentration from another. For example, in Fig. 4 the estimated ranges for 5-minute averaged sensor measurements of 150 μg/m³ and 200 μg/m³ will overlap significantly, so it would not be clear whether those measurements capture different concentrations or multiple measurements at the same concentration.

235 However, 5-minute averaged sensor measurements of 30 µg/m³ and 200 µg/m³ will clearly show a difference between measurements, even with the uncertainty in those measurements. It is notable that the relationship between uncertainty and sensor concentration is non-linear. This non-linearity is a detail that would not be captured using RMSE or normalized RMSE to describe uncertainty in measurements.

Figure 5 shows the dependence of average uncertainty on the concentration (sensor estimate) and on the averaging

240 time. For the sake of simplicity Fig. 5 only shows the average difference between the sensor estimate and both the upper and lower PI. In the example in Fig. 4 uncertainty would be calculated as (199 – 138 µg/m³)/2 + (138 – 90 µg/m³)/2. Generally, sensor uncertainty increases with concentration, though it does so non-linearly. For 5-minute averaged data the uncertainty is ±22 µg/m³ when actual concentration is 30 µg/m³, but this rises to ±88 µg/m³ for measurements of 300 µg/m³. When averaging times are lengthened and more data is included in each measurement the uncertainty ranges can change significantly. For

245 example, a measurement of 20 µg/m³ has uncertainty of ±17 µg/m³, ±12 µg/m³, and ±4.4 µg/m³ for averaging times of 5-minutes, 1-hour, and 1-day.

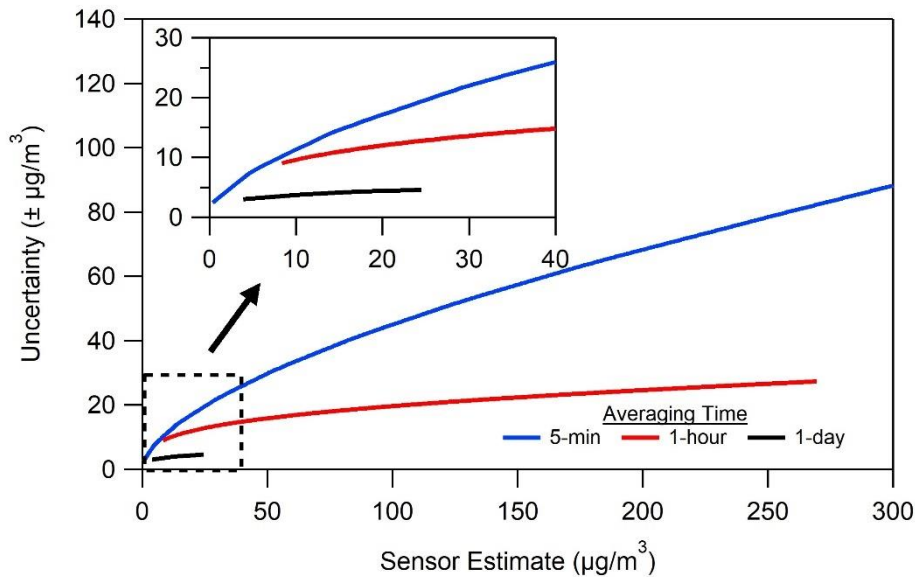



Figure 5. An alternative view of the prediction interval which shows how this interval varies with concentration and with averaging time.

250 An PI evaluation such as that shown in Fig. 4 and Fig. 5 offers more information about what to expect from future sensor measurements versus R^2 or RMSE. A future 5-minute averaged sensor measurement of 200 µg/m³ is more meaningful if it can be quantified with the statement that there is a 95% probability that the actual concentration is between 90 and 199 µg/m³. It An evaluation with a PI also allows for better comparison between sensors, as the evaluation results are not biased influenced by the range of concentrations observed during evaluation. Uncertainty at a specific concentration can be compared

255 from one brand of sensor to another and is not impacted by the range of concentrations observed, in contrast to RMSE or R^2.

In other words, the uncertainty of a sensor at 35 µg/m^3 does not change depending on whether concentrations of 100 µg/m^3 were also measured during evaluation, though the overall R^2 or RMSE of that evaluation can be influenced by the 100 µg/m^3 measurements, as shown in Fig. 3. ~~If presented with this analysis for different sensors, a user can quickly compare the expected uncertainty that comes with each sensor. The~~ An analysis such as that shown in Fig. 5 also allows a user ~~can also~~to see how averaging time and concentration changes the meaning of a sensor measurement.

Picking a single comparison point allows users to quickly compare measurement uncertainty between different sensor types, as they might currently using R^2 or RMSE. The breakpoints in the United States Air Quality Index (AQI) could be considered as standard comparison points. For example, the United States AQI transitions to "Unhealthy for Sensitive Groups" at 35 µg/m^3. At this level three sensor replicas showed uncertainty of 14, 16, and 20 µg/m^3.

3.3 Trends over time

The analysis shown in Fig. 4 and Fig. 5 relies on a user collecting enough data to predict the PI bounds that capture 95% of future data points. Without sufficient data the PI will be incorrect or will only cover measurements over a limited range. This is illustrated in the 1-day averaged uncertainty in Fig. 5, where uncertainty is only calculated for concentrations ranging from approximately 5 to 25 µg/m^3 due to limited data. The amount of time it will take to collect enough data to build a PI analysis for sensor data will vary from one location to another and is explored for this location in Fig. 6. This figure outlines an approach to determine the amount of time needed to calibrate a sensor using a PI as described in Sect. 3.2, or any other method.

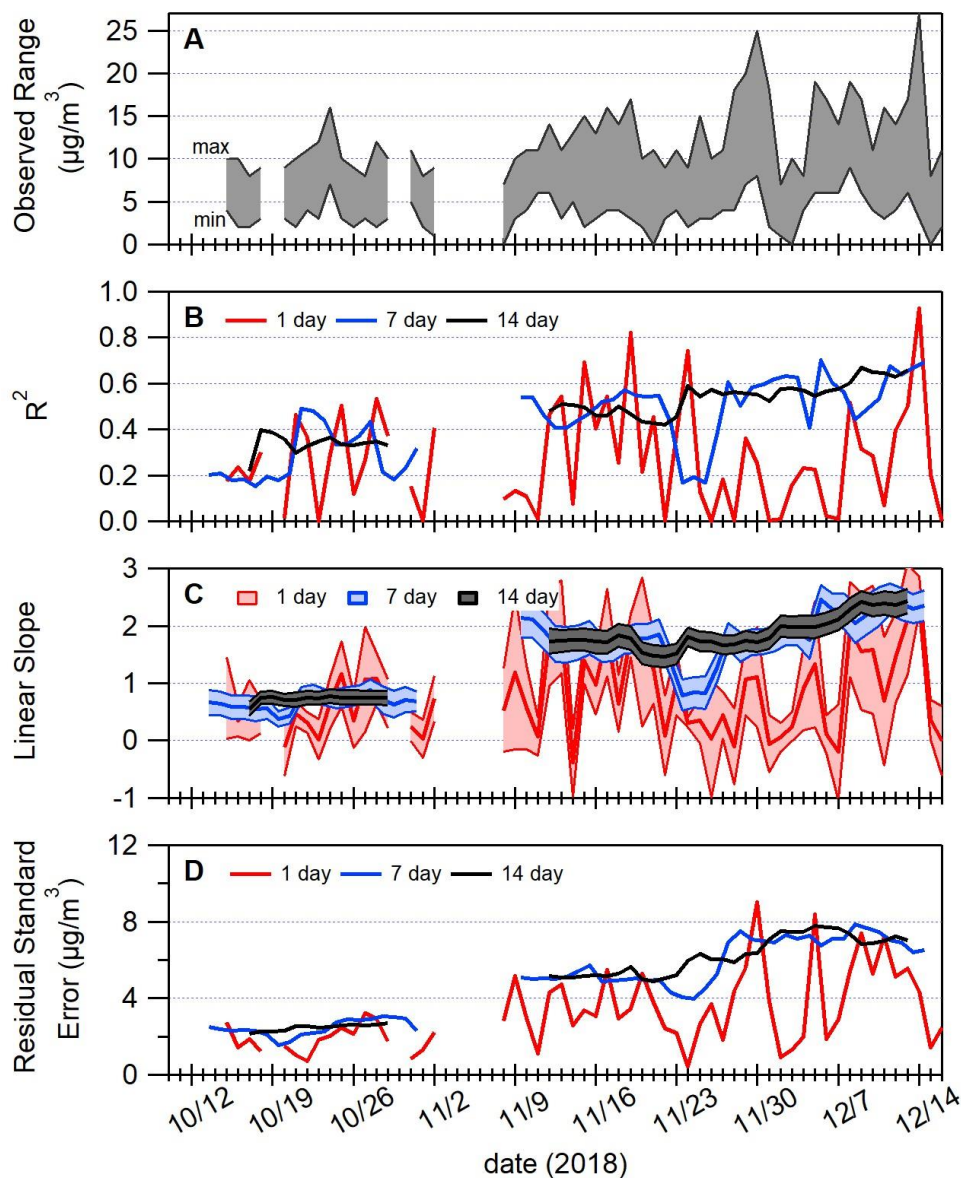



Figure 6. Linear regression parameters from 1-hour measurements over a two-month period in 2018. (A) shows the range of measurements observed on each day. (B)-(D) show the R^2, slope, and residual standard error, respectively. In (B)-(D) a rolling collection of 1, 7, or 14 days of data was used to find linear regression parameters.

Figure 6 shows variations to linear fit parameters between 1-hour averaged sensor and reference measurements collected during a two-month period. Figure 6A shows the range of values observed each day during this period. In Fig. 6B-D a linear regression is calculated for 1-hour measurements in a rolling timeframe of 1, 7, or 14 days. Figure 6B shows that with just one day of hour-averaged data, R^2 is sporadic from one day to the next. It is much more consistent over time when 7 days of data are used, but even then there are periods that are inconsistent with the rest of the data. For example, the R^2 for a 7 day

comparison between sensor and reference data centered around November 24-26th which is lower than at other times. This inconsistency is eliminated when using 14 days of data.

Just as for R^2, Fig. 6C shows a sporadic trend is also observed for slope when fitting a linear regression on just one day of data at a time. Notably, the confidence interval on the slope (shown as shaded area) often dips below 0, suggesting a lack in statistical confidence that a trend has been observed. Slope is much more consistent using 7 days of data, though as with R^2 there is a 7 day period in November (centered around the 24th-25th) in which slope is different. Using 14 days of data eliminates the abnormal period in November, but a difference is still observed between the slope in October and November. This suggests that 14 days is an improvement compared to 7 days, but even 14 days may not be enough collocation time for a thorough calibration.

Figure 6D shows residual standard error over time, which can be used as a surrogate for the width of a PI. Calibrations using 1 day of data often have low error, as they include only a narrow range of concentrations. The residual standard error changes over time, even using 7 and 14 day intervals. All variations in residual standard error should be captured for a thorough calibration. It is especially important to capture the maximum in residual standard error to ensure that the prediction interval captures the full range of uncertainty in future measurements.

In short, Fig. 6 shows that even with up to 7 or 14 days of data, variations in linear regression parameters are still observed. Considine et al. (2021) observed that approximately 3 weeks were needed for a simple, linear sensor calibration, and 7-8 weeks were required if using a more complicated correction such as machine learning. Duvall et al. (2021) recommended 30 days to develop a calibration. The goal of any calibration for low-cost particulate matter sensors should be to observe all variations in particle size, composition, and concentration. Thorough calibrations may require lengthy periods of collocation to capture all variations in these parameters and in the range of concentrations observed. An analysis of linear fit parameters over time such as Fig. 6 can be helpful in determining if the full range of situations has been observed and captured by the resulting calibration model.

3.4 Calibration implications

Understanding the behavior of low-cost particulate matter sensors over time is important in planning for the amount of data required to calibrate the sensor. The results presented in this study are useful in analyzing the strengths and weaknesses of different calibration methods. Stanton et al. outlined four methods that could be used to capture sensor and reference measurements for calibration (Stanton et al., 2018):

- Routine Collocation: sensors are placed near the reference instrument for a period before deployment. They may be brought back periodically for re-calibration.
- Permanent Collocation: one sensor is placed next to a reference instrument with the assumption that any correction needed for that sensor applies to the others in the network.
- Mobile Collocation: a reference instrument is place on a vehicle, and all sensors receive a single point calibration update when the reference comes within proximity.

315   •   "Golden Sensor": one sensor is calibrated via collocation with a reference instrument and then slowly moved throughout the network to calibrate the other sensors.

The Routine Collocation method is a useful starting point in any sensor network. A period of side-by-side sensor and reference data captures the slope, intercept, and typical error that is associated with sensor measurements. Figure 6 suggests that this method will have mixed results if calibrating over short time periods but can be reliable given enough time to capture

320 all variations in slope, concentration, and residual standard error~~is reliable if enough time is allowed to capture all variations in slope, concentration, and residual standard error~~. The method does not capture the location bias that may be observed if sensors respond differently to the particles at the sensor location compared to those at the calibration location. Despite these limitations, Fig. 6 suggests that this method can likely improve sensor data if utilized appropriately.

The precision between sensors (see Fig. 2) suggests there is potential for a Routine Collocation method to greatly

325 improve data, though there may be difficulties in accounting for location bias. The question remains on how much distance can be allowed between the reference sensor and the field-deployed network sensors before this method fails. This allowable distance will depend on how quickly particle properties change between the reference instrument and the network of sensors.

A mobile approach to calibration is attractive in that it can be used to capture variations to sensor responsiveness at different locations with different types of particles compared to those at a fixed reference location. The challenge with this

330 approach will be capturing enough data to thoroughly understand the slope and prediction interval of sensor measurements. Figure 6 shows that even using 24 1-hour data points can lead to unusual estimates for the slope between sensor and reference measurements, and a single-measurement spot check between a mobile reference and a sensor would likely be even more sporadic. It is possible that these single-measurement spot checks could slowly improve an existing calibration over time, but whether that improvement happens within a reasonable time frame is something that would require more exploration.

335 The "Golden Sensor" method relies on a calibrated sensor to calibrate other sensors in the network. The challenge with this approach is that actual concentration is not known once the sensor has left the reference instrument. At this point measurements are just an estimate with a 95% confidence interval. The level of uncertainty in measurements (see Fig. 4) would make it very challenging to pass a calibration from one sensor to another without greatly increasing uncertainty.

Regardless of the choice of calibration method, it is important to consider variations in sensor data when conducting

340 a calibration and when interpreting future sensor results. As Fig. 6 shows, calibrations may take weeks or longer in order to capture all variations in the external factors that influence sensor response. If these variations are not captured correctly, then the resulting calibration may miss important changes to sensor response that occur due to changing environmental variables. Reliance on a calibration that does not account for all variance in measurements may make sensor data less reliable. A calibration that also provides a PI ensures that future results can be interpreted with statistical rigor.

4 Conclusions

Low-cost sensors have potential to provide a better understanding of temporal and spatial trends of pollutants like particulate matter. Evaluations of low-cost particulate matter sensors alongside reference instruments in Bartlesville, Oklahoma have been used to identify methods that ensure more consistent evaluation and interpretation of sensor data.

Bias in sensor measurements varied over time but was very closely correlated from one sensor to the next (see Fig. 2). Because bias is so closely correlated, sensors can be deployed in pairs as a simple way to identify erroneous measurements (see Fig. 1). Finding ways to efficiently and effectively determine sensor performance is critical as sensors become more widely adopted. Two of the most popular evaluation metrics, R^2 and RMSE, can be influenced by averaging time, choice of reference instrument, and the range of concentrations observed (see Fig. 3). This study shows how a prediction interval can be used as a more statistically thorough evaluation tool. A PI offers a more robust method of sensor evaluation and a statistical confidence interval for interpretation of future sensor measurements (see Fig. 4). When properly applied, this method can show how uncertainty in sensor measurements varies as a non-linear function of observed concentrations and also varies with the averaging time for measurements (see Fig. 5). A standard ambient $PM_{2.5}$ concentration could be chosen for simple comparisons of uncertainty between sensors. For example, uncertainty measurements from multiple sensors at 35 $\mu g/m^3$ would allow convenient comparisons. Standardization of comparison concentration, reference instruments, and averaging times would allow more thorough decisions about which sensor is best suited to a proposed task. It is important to consider these impacts when interpreting sensor measurements. Building an effective prediction interval, linear regression, or any other calibration model depends on capturing the necessary data (see Fig. 6), and careful thought is required in planning the method and length of time for a calibration. The work presented here shows how adjustments to low-cost particulate matter sensor evaluations can greatly improve the interpretation of future data.

5 Competing Interests

The author declares that he has no conflict of interest.

6 Acknowledgements

The author thanks John Gingerich and Irby Bailey for their assistance in carrying out the experiments described in this study.

7 References

Ahangar, F. E., Freedman, F. R., and Venkatram, A.: Using Low-Cost Air Quality Sensor Networks to Improve the Spatial and Temporal Resolution of Concentration Maps, International Journal of Environmental Research and Public Health, 16, 1252, 2019.

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. H., and Hamburg, S. P.: High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data,
375 Environmental Science & Technology, doi: 10.1021/acs.est.7b00891, 2017. 2017.

Barkjohn, K. K., Gantt, B., and Clements, A. L.: Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor, Atmos. Meas. Tech., 14, 4617-4637, 2021.

Bauerová, P., Šindelářová, A., Rychlík, Š., Novák, Z., and Keder, J.: Low-Cost Air Quality Sensors: One-Year Field Comparative Measurement of Different Gas Sensors and Particle Counters with Reference Monitors at Tušimice Observatory,
380 Atmosphere, 11, 492, 2020.

Bi, J., Stowell, J., Seto, E. Y. W., English, P. B., Al-Hamdan, M. Z., Kinney, P. L., Freedman, F. R., and Liu, Y.: Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA, Environmental Research, 180, 108810, 2020.

Considine, E. M., Reid, C. E., Ogletree, M. R., and Dye, T.: Improving accuracy of air pollution exposure measurements:
385 Statistical correction of a municipal low-cost airborne particulate matter sensor network, Environmental Pollution, 268, 115833, 2021.

Datta, A., Saha, A., Zamora, M. L., Buehler, C., Hao, L., Xiong, F., Gentner, D. R., and Koehler, K.: Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore, Atmospheric Environment, 242, 117761, 2020.

Di Antonio, A., Popoola, O. A. M., Ouyang, B., Saffell, J., and Jones, R. L.: Developing a Relative Humidity Correction for
390 Low-Cost Sensors Measuring Ambient Particulate Matter, Sensors (Basel), 18, 2790, 2018.

Duvall, R., Clements, A., Hagler, G., Kamal, A., Kilaru, V., Goodman, L., Frederick, S., Barkjohn, K. J., VonWald, I., Greene, D., and Dye, T.: Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors: Use in Ambient, Outdoor, Fixed Site, Non-Regulatory Supplemental and Informational Monitoring Applications, United States Environmental Protection AgencyEPA/600/R-20/280, 2021.

395 Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Boghossian, B. D., Cocker, D., and Polidori, A.: Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site, Atmospheric Environment, 216, 116946, 2019.

Gao, M., Cao, J., and Seto, E.: A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China, Environmental Pollution, 199, 56-65, 2015.

Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., Beekmann, M., and Subramanian,
400 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, 158, 105833, 2021.

Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., Arn, T., Beutel, J., and Thiele, L.: Deriving high-resolution urban air pollution maps using mobile sensor nodes, Pervasive and Mobile Computing, 16, 268-285, 2015.

Holstius, D. M., Pillarisetti, A., Smith, K. R., and Seto, E.: Field calibrations of a low-cost aerosol sensor at a regulatory
405 monitoring site in California, Atmos. Meas. Tech., 7, 1121-1131, 2014.

Jayaratne, R., Liu, X., Ahn, K.-H., Asumadu-Sakyi, A., Fisher, G., Gao, J., Mabon, A., Mazaheri, M., Mullins, B., Nyaku, M., Ristovski, Z., Scorgie, Y., Thai, P., Dunbabin, M., and Morawska, L.: Low-cost PM2.5 Sensors: An Assessment of Their Suitability for Various Applications, Aerosol and Air Quality Research, doi: 10.4209/aaqr.2018.10.0390, 2020. 2020.

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

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., and Borowiak, A.: Review of the Performance of Low-Cost Sensors for Air Quality Monitoring, Atmosphere, 10, 506, 2019.

Kumar, V. and Sahu, M.: Evaluation of nine machine learning regression algorithms for calibration of low-cost PM2.5 sensor,
415 Journal of Aerosol Science, 157, 105809, 2021.

Kuula, J., Friman, M., Helin, A., Niemi, J. V., Aurela, M., Timonen, H., and Saarikoski, S.: Utilization of scattering and absorption-based particulate matter sensors in the environment impacted by residential wood combustion, Journal of Aerosol Science, 150, 105671, 2020.

Levy Zamora, M., Xiong, F., Gentner, D., Kerkez, B., Kohrman-Glaser, J., and Koehler, K.: Field and Laboratory Evaluations
420 of the Low-Cost Plantower Particulate Matter Sensor, Environmental Science & Technology, 53, 838-849, 2019.

Li, J., Zhang, H., Chao, C.-Y., Chien, C.-H., Wu, C.-Y., Luo, C. H., Chen, L.-J., and Biswas, P.: Integrating low-cost air quality sensor networks with fixed and satellite monitoring systems to study ground-level PM2.5, Atmospheric Environment, 223, 117293, 2020.

425   Malings, C., Tanzer, R., Hauryliuk, A., Saha, P. K., Robinson, A. L., Presto, A. A., and Subramanian, R.: Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation, Aerosol Science and Technology, 54, 160-174, 2020.

Mazaheri, M., Clifford, S., Yeganeh, B., Viana, M., Rizza, V., Flament, R., Buonanno, G., and Morawska, L.: Investigations into factors affecting personal exposure to particles in urban microenvironments using low-cost sensors, Environment International, 120, 496-504, 2018.

430   Mukherjee, A., Stanton, L. G., Graham, A. R., and Roberts, P. T.: Assessing the Utility of Low-Cost Particulate Matter Sensors over a 12-Week Period in the Cuyama Valley of California, Sensors, 17, 1805, 2017.

Papapostolou, V., Zhang, H., Feenstra, B. J., and Polidori, A.: Development of an environmental chamber for evaluating the performance of low-cost air quality sensors under controlled conditions, Atmospheric Environment, doi: https://doi.org/10.1016/j.atmosenv.2017.10.003, 2017. 2017.

435   Stanton, L. G., Pavlovic, N. R., DeWinter, J. L., and Hafner, H.: Approaches to Air Sensor Calibration, Oakland, CA2018.

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

Tryner, J., L'Orange, C., Mehaffy, J., Miller-Lionberg, D., Hofstetter, J. C., Wilson, A., and Volckens, J.: Laboratory evaluation
440   of low-cost PurpleAir PM monitors and in-field correction using co-located portable filter samplers, Atmospheric Environment, 220, 117067, 2020.

Williams, R., Duvall, R., Kilaru, V., Hagler, G., Hassinger, L., Benedict, K., Rice, J., Kaufman, A., Judge, R., Pierce, G., Allen, G., Bergin, M., Cohen, R. C., Fransioli, P., Gerboles, M., Habre, R., Hannigan, M., Jack, D., Louie, P., Martin, N. A., Penza, M., Polidori, A., Subramanian, R., Ray, K., Schauer, J., Seto, E., Thurston, G., Turner, J., Wexler, A. S., and Ning, Z.:
445   Deliberating performance targets workshop: Potential paths for emerging PM2.5 and O3 air sensor progress, Atmospheric Environment: X, 2, 100031, 2019.

Williams, R., Nash, D., Hagler, G., Benedict, K., MacGregor, I., Seay, B., Lawrence, M., and Dye, T.: Peer Review and Supporting Literature Review of Air Sensor Technology Performance Targets, EPA 600/R-18/324, 2018.

Zamora, M. L., Rice, J., and Koehler, K.: One year evaluation of three low-cost PM2.5 monitors, Atmospheric Environment,
450   235, 117615, 2020.

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

Zikova, N., Masiol, M., Chalupa, D., Rich, D., Ferro, A., and Hopke, P.: Estimating Hourly Concentrations of PM2.5 across a
455   Metropolitan Area Using Low-Cost Particle Monitors, Sensors, 17, 1922, 2017.

Zou, Y., Clark, J. D., and May, A. A.: Laboratory evaluation of the effects of particle size and composition on the performance of integrated devices containing Plantower particle sensors, Aerosol Science and Technology, 55, 848-858, 2021a.

Zou, Y., Clark, J. D., and May, A. A.: A systematic investigation on the effects of temperature and relative humidity on the performance of eight low-cost particle sensors and devices, Journal of Aerosol Science, 152, 105715, 2021b.

460   Zusman, M., Schumacher, C. S., Gassett, A. J., Spalt, E. W., Austin, E., Larson, T. V., Carvlin, G., Seto, E., Kaufman, J. D., and Sheppard, L.: Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study, Environment International, 134, 105329, 2020.