Evaluation methods for low-cost particulate matter sensors (amt-2021-154)

Jeff Bean

The author thanks the reviewers for the comments. The manuscript has been updated and more information on the updates is provided below. Author responses are shown in italics and a tracked-changes version of the manuscript shows changes in red.

RC1:

This paper presents and discusses an improved air sensor evaluation method the prediction interval. Overall, this appears to be a valuable addition to statistics more commonly used to evaluate air sensor performance. However, the methods are lacking to understand how to implement this method in future work. Some of the discussion and conclusions are also lacking.

Major:

It seems like this paper would be more helpful to the air sensor community if the sensor type was provided.

Response: The sensor models are withheld in part to allow focus on evaluation methods, rather than specific sensors, but models are also withheld to avoid giving the impression that the author endorses or disparages one sensor company over another. Additional details on sensor measurements have been added to the manuscript, as requested, but the specific makes and models are still withheld for the reasons stated.

Line 98: I think this method is super helpful to think about determining how to pick a threshold to remove outliers based on precision moving forward! Did you consider whether percentage and absolute concentration criteria would be more useful than just 50 ug/m3 by itself (as used in https://doi.org/10.5194/amt-14-4617-2021)? What was the range of concentrations experienced by the sensors during this period (as reported by the sensors since that is the criteria you are using to exclude)?

Response: That is a good suggestion about using a percentage data agreement criteria. This was considered, but in the end we chose to only use an absolute criteria for simplicity and because it worked well. Some discussion on this has been added to the manuscript (starting line 107).

Figure 1: How much does removing disagreeing datapoints clean up the bias? Or does it only clean up the R2 as shown in this figure? It would be helpful to include a 3rd panel on Fig 1 that would be average concentration reported by the sensor versus the allowed difference and also the average of the FEM as well.

Response. Good question. Improvements in RMSE are also observed as disagreeing datapoints are removed. These results have been added to Figure 1.

Line 120: How was no correlation defined? How did you decide which parameters to include in the figures and which not to? What are the correlations of the parameters in the figures? It would be helpful to also show the correlation between PM10 or PM10-2.5 and the error since as you say in the paper particle size can influence the accuracy.

Response: "no correlation" was a strong choice of words and that has been changed (line 134). The key message was that extremely close correlation was observed between sensors, even at times when bias was very high, but despite this there was no indication that this was caused by factors such as humidity, temperature or any other measured parameter. Unfortunately, PM_{10} concentrations were not available during this measurement campaign.

Figure 2 seems overly complicated and challenging to interpret. It seems like 2A would be easier to interpret if it just showed sensor 1 versus sensor 2 (not the bias) and then if figure 2d showed sensor 1 versus the BAM concentration. I think this would be valuable even if you wanted to add a figure showing the basic plots and then a second figure showing all the bias plots if you feel you need both.

Response: The reason we focus on bias in the panels of the figure is to draw attention to the fact that these sensors often read well above or below the actual concentration, but in a very consistent way from one sensor to another. We believe that 2A and 2E draw attention to the fact that sensors will read higher or lower than they should in a way that suggests that it should be correlated with an external factor, though that factor was not observed in the measurements available during this campaign.

Prediction Interval: I don't understand how to calculate this based on the information you have included in the text. Please cite additional references and include the equations in the SI if needed. Please include a citation for the R package if one was used. What are the 3 diagonal lines shown on Figure 4? Can you label them?

Response: The three diagonal lines are the best fit and upper/lower prediction intervals. A prediction interval can be provided by the standard linear model (lm) package in R. For example:

```
sensor_model <- lm(data, formula = reference ~ sensor)

prediction <- predict(sensor_model, data, interval="predict")
```

Line 186: "Through examination it was found that residual trends were best eliminated by raising both the sensor and reference data to the 0.4 power." More scientific explanation is needed. How would we repeat this method in the future would others always just use 0.4 since that is what you "examined" and found or would they need to examine their datasets and come to different conclusions? Your figure doesn't show transformed data so it is unclear how you would determine this. Please provide additional explanation.

Response: Some additional clarification has been added to the manuscript (starting line 207). There are many transformation methods that could be applied but the key criteria before accepting a linear regression is that the residuals should be evenly spaced across the sampling domain. In our case raising both reference and sensor data to the 0.4 power eliminated any trends in residuals. Raising data to a power is just one option for any data that has uneven residuals. Future applications of this type of method might find other ways to ensure consistent residuals.

Past work has shown that sensors may respond nonlinearly at high concentrations (**https://doi.org/10.1111/ina.12621**, https://doi.org/10.1016/j.envpol.2018.11.065). Would a prediction interval still be appropriate in a case like this? Would you be able to remove the residual trends in a dataset like this?

Response: A prediction interval is especially relevant if there is nonlinearity under different concentration regimes. The method in the manuscript would probably work, but if it didn't then an equally thorough alternative would be to split the data into different ranges and fit separate prediction intervals for each.

Line 189: "This did not change the outcome significantly". Please define what you mean by this. How did you decide to use 70 ug/m3 as your split to have equal number of data points above and below? Does Figure 4 show the full dataset or the subset? How important is having a well-balanced dataset to getting accurate results?

Response: 70 µg/m3 was chosen because it approximately split the 5% of data during high concentration events from the rest of the data. The random sampling below this line meant that the model and prediction intervals were a result of 50% of data below 70 µg /m3 and 50% of data above that point. If the model instead used 95% of data below 70 µg /m3 and only 5% above then the model is fit in a way that is more weighted towards lower concentration. If the transformation is done well and the residuals are the same across the domain then the weight of lower/higher concentrations does not matter. However since residual distribution is not perfect it is best to fit the model using an equal amount of data across the range of observed concentrations. Some clarification has been added to the manuscript at line 213.

The need to balance a dataset seems like a limitation of this method as compared to R2 or RMSE that has not been discussed.

Response: Balancing a dataset can improve the robustness of any model, as it ensures the model is built equally using the entire domain over which it will predict. This can improve R2 and RMSE in a standard linear model as well. However as discussed here, the result of doing this is small if residuals are correctly accounted for, so doing that should be the priority in building any model that predicts PM predictions from sensors.

Figure 4: I would recommend coloring the arrows uniquely and including them next to the text in the bottom right corner as a legend so the figure can be more quickly interpreted.

Response: Great suggestion. Additional labels have been added to make this quicker to interpret.

Figure 5: Is this only for T640 data? If so, why not also show BAM data as shown in the other figure?

Response: Figure 5 was built only for T640 data as this allowed 5-minute resolution in comparison with 1-hour or 24-hour. The BAM is limited to 1-hour resolution.

Figure 5: Did you consider whether uncertainty as a % would be more stable?

Response: Uncertainty as a percentage is an interesting idea. When applied to 5-minute data in Figure 5 it results in something like an exponential decay with uncertainty as high as 400% for low concentrations (1 µg/m3) that decays towards ~45% uncertainty at 100 µg/m3. We believe that absolute concentrations are a little easier/quicker to interpret, but percent uncertainty is an idea that could be interesting to explore in the future as well.

Line 219: “It allows for better comparison between sensors, as the evaluation results are not biased by the range of concentrations observed during evaluation.” I think more explanation is needed here. It doesn’t seem to me this is one of the findings of your analysis since you only showed the results from one dataset covering the full range. It might be interesting to show another data subset with a different concentration range to understand how the range doesn’t impact the results but maybe there is another way to explain since as I said above I don’t really understand how you are calculating the prediction interval.

Response: More explanation has been added to that section to clarify: “Uncertainty at any given concentration can be compared from one brand of sensor to another and is not impacted by the range of concentrations observed, in contrast to RMSE or R2. In other words, the uncertainty of a sensor at 35 µg/m3 does not change depending on whether concentrations of 100 µg/m3 were also measured during evaluation, though the overall R2 or RMSE of that evaluation can be influenced by the 100 µg/m3 measurements, as shown in Fig. 3.”

Lines 306-308: “Two of the most popular evaluation metrics, R2 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.” Figure 5 shows that prediction interval is also influenced by averaging time. Are you saying it isn’t influenced by reference instrument? If so, I think you need more results to show that. Overall, this statement seems misleading.

Response: The reviewer is correct that the statement here was misleading. A few additional sentences have been added to clarify that the choice of reference instrument still needs to be standardized and that evaluations would be more simple if averaging time was also standardized (starting line 354).

Figure 6: With much of the data below 5 ug/m3 did you consider how LOD of the sensor and reference influence your results?

Response: A LOD was not provided for the sensors being evaluated and was not explored in this work. Future work could consider the connection between uncertainty, as measured here, and LOD, which is a similar concept.

Figure 2 seems to show that the bias is much more variable at high RH. How can you take that into account using prediction interval?

Response: Yes, a prediction interval can also be found for a multiple linear regression. The approach to do so in R is similar as with a single linear regression, as shown below. Exploring a prediction interval based on many predicting variables would be interesting and valuable, but would greatly complicate the figures in this work. To keep the concept simple we focus just on one prediction variable.

```
sensor_model <- lm(data, formula = reference ~ sensor + RH)

prediction <- predict(sensor_model, data, interval="predict")
```

Did you consider how the precision of the sensor influences prediction interval? I'm assuming that Figure 4 is for all the sensors but if it is for a group of sensors that would be helpful to clarify in the caption/text.

Response: Figure 4 is from a single sensor. The Figure 4 analysis could either be repeated for replicas of a sensor or data from multiple sensors could be combined to create an analysis such as Figure 4. Additional numbers are now included for a repeat of this analysis of a second sensor and more discussion has been added about how one might approach this for multiple sensors and how sensor precision will impact this (starting line 250).

It would be helpful to add the prediction interval for all of the sensors you tested not just the best sensor so that readers could compare the R2/RMSE/PI more closely across devices and understand how they could use this in the future.

Response: Good suggestion. The analysis has been added for a 2nd sensor to show how they compare and more discussion has been added about how one might approach doing this analysis for multiple sensors. Comparisons between sensor uncertainty for different sensors have been added as well (starting line 259).

Have you thought about how you could report this PI as something more easily to compare across papers than a plot (which may have different axis labels etc.)? For example, fitting a function or reporting the 95% uncertainty at various AQI breakpoints, etc?

Response: This is a nice suggestion and the following text has been added to the manuscript: "Picking a single comparison point allows users to quickly compare measurement uncertainty between different sensor types, as they might currently using R2 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/m3."

This work is missing relevant citations. Examples: Giordano 2021 calibration review paper https://doi.org/10.1016/j.jaerosci.2021.105833, Zheng 2018 similar discussion of averaging interval and the precision of the reference https://doi.org/10.5194/amt-11-4823-2018, some others included in my other responses.

Response: These works have now been appropriately cited and the author appreciates the reviewer pointing them out.

Minor:

Figure 1: Could you include the averaging interval you are using for exclusion in the figure caption?

Response: They were averaged to 1-hour intervals for this figure and this has now been included in the caption.

Line 45: "The root of a calibration for low-cost particulate matter sensors is simple: sensors and reference instruments measure the same mass of air for a period and then adjustments are made to better align sensor measurements.". I'm not sure "root" is the clearest way to express this.

Response: This sentence has been adjusted to read: "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."

Line 255: US EPA recommends at least 30 days for their PM2.5 sensor evaluations. https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=542106&Lab=CEMM

Response: This has been added to the manuscript (line 296).

RC2:

General Comments:

In this work, the author examines the impact of controllable factors, such as averaging time and type of reference instrument, on performance metrics used to evaluate low-cost sensors. In addition, the author demonstrates that pairing two sensors together with a data agreement requirement can act as an easy quality assurance check. Also, the author proposes and tests using a prediction interval as a method to evaluate low-cost sensor performance. Both of which would enhance our current methods of evaluating and comparing low-cost particulate matter sensors. However, the methods used in this paper are not well described and further work could be done to strengthen some of the conclusions.

Major:

Line 31: The EPA recently published a report on performance testing protocols, metrics, and target values for PM2.5 low-cost air sensors. In this report they recommend using various indicators to evaluate sensor performance and offer performance target values for those indicators. This report may better address the concerns of using R2 and RMSE to evaluate sensor performance in addition to other performance metrics.

Duvall, R., A. Clements, G. Hagler, A. Kamal, Vasu Kilaru, L. Goodman, S. Frederick, K. Johnson Barkjohn, I. VonWald, D. Greene, AND T. Dye. 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. U.S. EPA Office of Research and Development, Washington, DC, EPA/600/R-20/280, 2021.

Response: This newer work from the EPA has now been included in the discussion (see lines 54, 184). This newer work helps to address some of the issues with standardization of evaluations, though there is a continued need to discuss whether R^2 and RMSE are the best metrics for evaluations.

Line 66: Multiple times in this paper agricultural burning emissions are listed as the cause of the high PM events, what other evidence can be provided to confirm these burning events? Do particles generated by these events have different optical properties than those present in ambient air?

Response: These agricultural burning events are very common in this area and plumes of smoke can often be seen rising from fields while driving around in the area. Further downwind, the plumes become less distinct and become a broad haze. The particles likely have unique properties, but identification properties were not measured in this work. While it would be nice to have a way to express more certainty about the origins of these high concentration events, the origin of these events is not crucial to the analysis in this manuscript.

Line 72: "…while the T640 uses an optical counting method that is more similar to the method used by the low-cost particulate matter sensors." Please provide a little more detail on the similarities and differences between the T640 and the low-cost PM sensors used in this study.

Response: More details have been added to clarify the key similarities and differences: "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."

Line 73: "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 which did not meet this criterion was removed prior to analysis." Please clarify which data was removed, is this the cause of the gaps in Figure 2E?

Response: Any data for which the internal RH measurement of the BAM exceeded 35% was removed. This is the cause of all gaps in Figure 2E and a note explaining this has been added to the caption for that figure.

Line 78: More information should be provided on the initial testing of the 4 low-cost PM sensors of different brands. Why is this one-month test indicative of how the sensors will perform over the ten-month test? Why was only R2 used to determine the best-performing brand?

Response: It has now been clarified in the manuscript that a definitive identification of the "best" sensor from this one-month trial was not crucial to the study. The purpose of the one-month trial was to identify a useful sensor that could be explored with additional analysis, as described in the rest of the manuscript (starting line 84).

Line 79: It would be helpful to state the brands of the low-cost PM sensors in addition to the OEM of the optical sensors inside the devices. Additionally, more information such as sampling/averaging time of these sensors should be provided.

Response: The sensor models are withheld in part to allow focus on evaluation methods, rather than specific sensors, but models are also withheld to avoid giving the impression that the author endorses or disparages one sensor company over another. Additional details on sensor measurements have been added to the manuscript, as requested, but the specific makes and models are still withheld for the reasons stated.

Line 113: "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." Figure 2A only shows the correlation for 2 sensors, is this same trend seen when comparing all 8 sensors? Did all sensors have the same response to the environmental conditions?

Response: For simplicity only the comparison between 2 sensors is shown but the result was similar for other combinations of sensors. This has been clarified in the manuscript: "Comparison between only two sensors is shown in Fig. 2A for simplicity, but similarly strong correlation was observed for other pairs of sensors."

Line 218: The paper states that the prediction interval can be used to evaluate/compare multiple sensors, however data is only provided for 1 sensor when 8 were tested. It would be interesting to examine the PI between all 8 of the same brand and even compare the prediction intervals of the 4 brands initially tested. Including the results of these tests would strengthen the argument to include prediction interval as a performance evaluation metric for low-cost PM sensors.

Response: This is a good suggestion. Unfortunately, there was an insufficient range of concentrations during the 1 month of initial testing to produce a meaningful prediction interval. However a comparison of uncertainty between other sensor replicas at 35 µg/m3 has now been included (line 262).

Line 280: "The question remains on how much distance can be allowed between the reference sensor and the network sensors before this method fails." The EPA recommends mounting sensors within 20m horizontal from the FRM/FEM monitor. See above citation which also includes recommendations on setting up low-cost sensors at test sites.

Response: This statement was referring to sensors that are deployed in the field (away from a reference) but still calibrated using a sensor that is collocated with the reference. It has been clarified in the manuscript to indicate this (line 323).

Minor:

Line 98: "As more stringent data agreement requirements are put in place (moving to the right in Fig. 1) there are not significant improvements in correlation." Change not to no.

Line 107: "The bottom figure shows how R2 between the sensor pair (average of the pair) and reference measurement changes as when these disagreeing data points are removed." Reword this sentence in the Figure 1 caption. Suggestion: "The bottom figure shows how the R2 between the sensor pair (average of the pair) and reference measurements changes when disagreeing data points are removed."

Line 273: "Figure 6 suggests that this method will have mixed results if calibrating over short time periods but is reliable if enough time is allowed to capture all variations in slope, concentration, and residual standard error." Reword second part of this sentence. Suggestion: "…but can be reliable given enough time to capture all variations in slope, concentration, and residual standard error."

Response: These minor suggestions have all been implemented and the author thanks the reviewer for pointed them out.

RC3:

This manuscript describes a co-location calibration of low-cost optical PM2.5 sensors. Overall the manuscript is topically relevant and well-written.

My major concern is that the Methods need more detail. The author does not reveal the brand of sensor used in this study, so a lot of the methods seem like a black box. For example, there needs to be some detail on the conversion of signal to PM concentration. Does the author simply rely on the factory calibration? Is there any compensation for humidity or other meteorological parameters?

Response: The sensor models are withheld in part to allow focus on evaluation methods, rather than specific sensors, but models are also withheld to avoid giving the impression that the author endorses or

disparages one sensor company over another. Additional details on sensor measurements have been added to the manuscript, as requested, but the specific makes and models are still withheld for the reasons stated.

A second important consideration is the range of PM concentrations shown in many of the figures. There seems to be a lot of emphasis on high concentrations (e.g., Fig 4), but I think it's more important to understand the performance of these sensors at typical ambient PM concentrations. Several examples are given in the specific comments below.

Response: More references to performance at ambient concentrations have now been added as discussed in the responses below and in responses to other reviewers.

Comments:

Line 90 discusses the value of co-locating two sensors to catch instances of erroneous measurements from a single sensor. Some commercially available sensor packages, like the Purple Air, already do this.

Response: It is true that Purple Air already includes two sensors in each package. The discussion in this manuscript could easily be applied to a setup such as this to eliminate data that fails a quality assurance step.

What is the time resolution of Figure 1? Also, it seems strange that the R^2 generally decreases as the allowed difference shrinks. Is this perhaps because a small allowable absolute difference ends up scrubbing data from higher concentration events?

Response: Data was averaged to 1-hour intervals in this analysis for comparison to the BAM. The reviewer is correct that higher concentration data tends to get removed with stricter data agreement requirements. This is an important point and an argument for using a combination of absolute and percent data agreement requirements. This has now been noted in the manuscript (starting line 107).

Line 187 - why was 70 ug/m^3 selected as the cutoff point?

Response: 70 µg/m3 was chosen because it approximately split the 5% of data during high concentration events from the rest of the data. The random sampling below this line meant that the model and prediction intervals were a result of 50% of data below 70 µg /m3 and 50% of data above that point – a relatively even distribution of data. If the model instead used 95% of data below 70 µg /m3 and only 5% above then the model is fit in a way that is more weighted towards lower concentration. If the transformation is done well and the residuals are the same across the domain then the weight of lower/higher concentrations does not significantly impact the resulting model. However, since residual distribution is not perfect it is ideal to fit the model using an equal amount of data across the range of observed concentrations.

The example in Figure 4 focuses on a very high concentration. Perhaps it would be more useful for readers to show this example for a more typical PM2.5 concentration.

Response: The high concentrations in Figure 4 are useful as it provides a clear picture of how uncertainty changes as concentration does. The reviewer is correct that more discussion of lower concentrations can add relevance to the discussion. Examples have been added of how this analysis applies to uncertainty at relevant concentrations (35 µg/m3) and how that varied between different sensors (starting line 259).

The analysis in Figure 4 relies on a transform of the data to the 0.4 power. I assume that this is sensor-specific, and if someone wants to repeat this analysis they will need to find a transform that works for their sensor. Revealing the specific sensor used in this work would help others try to repeat the same transform.

Response: It is true that this is sensor specific and could even be location specific since different particles may result in different sensor responses. This has been clarified in the manuscript (starting line 207).

I like Figure 5, but I wish I could see the typical ambient range (let's say up to 40 ug/m^3) better. I think that understanding the uncertainty for sensor measurements at typical ambient conditions is important, because in many cases these sensors will be deployed to examine neighborhood-level variations in PM concentration. Those variations can be small.

Line 225-226 note that in Fig 5, the daily average PM2.5 PI's can only be calculated for concentrations between 5 and 25 ug/m^3. This echoes my comments above - the performance of the sensors at low concentration is very important.

Response: This is a good recommendation and a zoomed-in window has been added to Figure 5 to better show how uncertainty varies at lower concentrations.

Figure 6 - I'm a little confused by exactly what is plotted in this figure. However it seems to me that the results are showing noise reduction via signal averaging. It's not clear to me why the R^2 is so low for the 1-day case (though errors are similar to, or slightly lower than, the 7- and 14-day cases).

Response: 1-day errors can be very low if the day contained only a narrow range of concentrations, especially if all concentrations were very low. In these cases there may be little correlation with the reference but errors are small.