Comments on: amt-2022-140 **"**Low-Cost Air Quality Sensor Evaluation and Calibration in Contrasting Aerosol Environments**"**

Summary

The objective of this work is to test and evaluate their existing MOMA calibration methodology for the $PM_{2.5}$ and PM_{10} sensors in the AQY sensor package using two approaches in three regions in southern California. The authors find that the best method to select a proxy reference site for building the calibration model(s) is via shortest distance. Both the monthly and drift detection framework calibration approaches appear to improve the quality of the uncalibrated PM sensor data, with the drift detection approach being slightly superior. They find this is likely due to temporal changes in meteorological and aerosol conditions at their sites over the study period, which the drift detection framework can more readily detect and adjust to.

Strengths

- There is a clear and straightforward objective and the methodology to address it is sound. The results and figures support their conclusions, and they are generally presented in a way that the reader can identify the main takeaway. The authors also provide more in-depth examples to illustrate a few of their more detailed findings, which are helpful in interpreting the results of the MOMA evaluation.
- The analysis clearly suggests that calibrating at a monthly frequency would not be sufficient under these conditions, and that a drift detection framework improves the PM sensor data quality by adjusting the calibration as dominant aerosol conditions change within the region.

Weaknesses

- The authors fail to discuss their results or conclusions in the context of their application (which is presumably to operate and maintain a network of AQY sensor packages in southern California to supplement the existing regulatory network), which would help the reader understand if the MOMA calibration methodology is feasible and widely applicable to others with a similar goal. They also do not really address the challenges or shortcomings of their approach, compare/contrast their method to other calibration methodologies, or indicate future research directions.
- The details of their methodology are not readily available (i.e., number and location of collocation sites and AQY sensors, duration of the study/data completeness, etc.) and the work of linking the AMS abbreviations/codes to each of the three study regions is often left up to the reader. There are missed opportunities to use color and other visual aids to help make it obvious which reference sites and AQY sensor IDs are in which region. Further, the text could focus more on framing their general findings in terms of the three regions (LA, IE, SC desert) by comparing/contrasting the MOMA performance in each of these settings/seasons. Presently, it is difficult for a reader unfamiliar with the region, the typical conditions and sources, and AMS abbreviations to follow some of their statements.

General Technical Comments

- The number and location of AQY-FRM/FEM collocated pairs and proxies should be more clearly indicated. For example, in Figure 1, which sites indicate collocation sites and which sites represent proxy sites? Further, a very clear description for "collocation site", "calibration site", "proxy site", needs to be provided and the language should remain consistent throughout (e.g., Figure 4 mentions "the site of interest" - this should be replaced with one of the previously defined terms). From my

understanding, the proxy site is used to develop the calibration, and the collocation site is used to validate or evaluate the performance of that calibration. This info is currently summarized in Section 2.5, but it should be moved into Section 2.3 and made more explicit (i.e., "We use MOMA at the proxy sites to develop an AQY calibration using two different approaches. Then, to evaluate the performance of the MOMA calibrations, we compare the calibrated AQY sensor data (N=6) to reference instrument data at the two collocation sites in each of the three regions/networks.").

- The authors designate three study regions (LA, IE, and RC desert) but do not indicate which of the FEM/FRM sites (e.g., CELA, CMPT, INDIO, etc.) are in which region until Section 3.1. Ideally, this should happen in the Methods. It would also help if some visual aids (like in Figure 3) were consistently used to distinguish sites within the three regions (i.e., different color shading, text labels on the figures, or at least grouping the sites within the same region/network adjacent to one another on Fig. 5-8). Also, in Figure 3 these three regions are referred to as "Networks". The terminology should be defined early and stay consistent throughout.
- Different study durations are given in Sections 2.5 and 2.3 – it is not clear which parameters were collected when and which subsets were used to assess the different calibration approaches. This should be spelled out. For example, it seems that the AQY have been deployed since April 2020 to the present day and PM_{10} sensors were added sometime in 2021 (this should be indicated in Section 2.2), but the period used for the analysis was only from Aug 2020 to either January 2021 or February 2022 (both timelines are given in Section 2.3 – unclear which is correct). The authors should provide an explanation for why different subsets of the data were used for different purposes (or if there were gaps or if certain time periods – like the fog episodes - were deliberately excluded from the analysis).
- There are various performance metrics introduced throughout, sometimes for the first time in the Results section, sometimes redundantly, and sometimes inconsistently (e.g., "K-S test" and "KS test"). Consider adding a new section to the Methods called "Assessment metrics", to define and describe just once which metrics are used in the analysis (MAE, K-S test, R^2, etc.) and what they are used for.

General Editorial Comments

- The first few sentences of the Introduction are a bit disjointed and lacking in references. For example, the mention of negative health effects and different sources can be more explicitly linked to the sizes of PM that are federally regulated/focused on in this study (i.e., small particles are largely to blame for health effects; Different sources lead to different sizes of particles: combustion sources → small particles, volcanoes/sea salt/mining → larger particles, etc.). In the second paragraph, more attention could be given to recent research that has shown how well/poorly these types of optical PM sensors perform at the different size ranges focused on in this study (see work by Ouimette et al 2022: https://doi.org/10.5194/amt-15-655-2022)
- The Methods section contains too much background information that would be better in the Introduction and lacks specific details that the reader needs to understand what the authors did. It could also be reorganized. The authors should first introduce what datasets they used and for what purpose (Sections 2.2 and 2.6). The mention of PM speciation data does not occur until Section 2.6 – the authors should indicate earlier on why these data were collected and what they were used for. Then follow with a description of the study area: Section 2.1 could include more information on the site-specific characteristics of the three regions: why were these selected and how are they different/similar? This could also be a good place to link the AMS site codes to the three regions (some of this detail can be moved from Section 2.4). Finally, end with the calibration approaches

and proxy selection criteria (Section 2.3 and 2.4). Section 2.5 could be eliminated, and the information therein could be incorporated into Section 2.3 to improve clarity.

- There are a few typographical and grammatical errors throughout that need to be corrected (see specific comments for a few examples). Further, the authors frequently use vague language instead of quantitative statements (i.e., "considerable discrepancies", "better improvements", "positive impact", etc.)
- There are a lot of figures. The authors may consider combining some. For example, Fig. 5 and Fig. 7 could be combined if $PM_{2.5}$ and PM_{10} were plotted in the same panel and point color/opacity was used to distinguish them (instead of coloring the data by count, which doesn't seem relevant to any of the discussed results). Fig. 6 and 8 could also potentially be combined by placing the $PM_{2.5}$ and PM_{10} results side-by-side. There is also an inconsistent use of color between figures which makes it more difficult to link sites/regions from one figure to another.

Specific Comments

Abstract

Best practice is to define "MOMA" acronym at first use.

Introduction

Line 35: "*…opportunities to measure PM with much denser networks and making them popular choices for citizen projects and…"* typographical error

Line 38-39: "*The relationship between scattered light, particle size and number, and the PM mass is dependent on the properties of the particles, which include size, shape, refractive index, and composition*." The relationship between scattered light and particle size and number depends on particle size? This sentence should be rephrased for clarity – the authors know what they are talking about, but the details need to be explained more carefully/explicitly.

Line 41: "*…change with particle type or properties changes over time*." Typo

The study duration should be mentioned somewhere in the last paragraph.

2.1 Study area

Can Table S1 indicate which AMS are in which region (LA, IE, or RC desert)?

Lines 74-79: This information may fit better in the Introduction.

2.2 Air Quality Sensors

How many sensors are in this network? And how many are in each of the three study regions?

Can you provide a reference or description for the humidity correction algorithm?

Lines 92-93: "*The AQY PM measurements were evaluated by South Coast AQMD's Air Quality Sensor Performance Evaluation Centre*." What do the authors want the reader to take away from this statement? At least a summary of the performance evaluation(s) (and under what conditions the evaluation was performed) should be provided here.

2.3 Remote Network Calibration

Line 98 and Line 112: Was the study period for the monthly calibration approach 13 months longer than the drift detection approach? Or is this a typo? The study duration is unclear.

For the monthly calibration approach, were the 7-days in the calibration window required to be continuous? Or were 7 non-continuous days selected from the most recent two-week window? Is that calibration then applied for the next 30 days or from the first to last calendar day of the subsequent month?

A definition of "gain" and "offset" should be provided to the reader.

Lines 109-112: For the drift detection approach, why are data from a 3-day period compared to predetermined thresholds from a 5-day period? This should be rephrased to improve clarity. The authors then provide the thresholds that they used to determine sensor drift and say: "*These thresholds can be adjusted to explore test sensitivity to drift detection*." Were they adjusted in this study or are they just saying that it could be done? Did they only use the thresholds they just gave? Again, the methodology should be described more clearly.

2.4 Proxy site selection

Lines 120-128 belong in the Introduction, integrated with lines 52-57. This should be used to motivate and frame the objective of this paper: to identify what proxy site selection and calibration approach(es) worked best for the PM sensors, in comparison to what the authors have already learned about their O_3 and NO_2 sensors.

Lines 132-135: Which of these sites are in which of the three previously described regions (LA, IE, and RC desert)?

2.5 Evaluating the performance of MOMA

Table 1 is located too far away to be first mentioned in Line 148 – Just reference Figure 1.

2.6 Speciation data

This is the first mention that speciation data was collected. For what purpose was this data used? It is not entirely clear which data were obtained using RAQSAPI and which data the authors collected themselves using integrated filter trains and/or the MetOne SASS. It should be spelled out exactly what parameter was measured by which instrument or where it was obtained. This section is also lacking detail - how was the total amount of OC and EC determined? Over what period was this speciation data collected? For which sites was this data collected?

3.1 General characteristics of the data

Figure 2 could be an SI figure.

Figure 3 – can you indicate here which sites are the collocation sites in each region and which are proxy sites? Maybe using line boldness or line style. Also, can this figure be larger? The $PM_{2.5}$ and PM_{10} panels might look better side-by-side rather than stacked.

3.2 Proxy selection criteria

Criteria should generally be covered in the methods, not the results. Lines 188-190 should go in Section 2.4 or a new "Assessment metrics" section.

Line 189-190: "*By using data from the reference network any uncertainties related to sensor performance are eliminated*." Can you be more explicit?

Line 191: "…*proxy site rather than the site with the most similar land use is the most suitable proxy resulting in the lowest and highest R2*…" lowest MAE? Missing word?

Figure 4: What is meant by "*the site of interest*"? It is a bit unclear what the authors are trying to find out via this figure – which potential proxy site is most like their collocation site? The caption should also provide directions for how to interpret the figure (i.e., "data points in the upper left-hand corner indicate best performance based on these metrics".) The caption also says, "*the site with the most similar land use is labelled with a "*".*" How was this determined? Again, this should be covered in the methods. Lastly, what is a "*facet*"? Suggest using the word "panel" instead.

Line 207-208: "*Overall, the nearest proxy generally resulted in the most similar distribution with the smallest MAE and largest R2*". A few lines above, the authors say "*lowest MAE and highest R2*" (This is the better way to phrase this – 'small' and 'large' are not the best adjectives for these metrics). Either way, try to stay consistent throughout.

Figure 1 makes it seem as if the proxy sites were selected as part of the methods, but it appears that the best proxy site was determined as part of the results (i.e., Section 3.2 and Figure 4). Perhaps the proxy sites should not yet be indicated in the Figure 1 map – maybe the collocation sites can be highlighted there instead?

3.3.1 PM2.5

Line 225: "*Also, the PM sensor does not exhibit significant instrumental drift over the 12-month period*." Where is this shown or how was this determined?

3.3.1 PM10

Line 290-291: *"…interestingly the GRIMM and T640 PM10 MAE is the highest - the opposite of the PM2.5 result."* Can the authors suggest a few potential explanations?

Fig 6. And 8 captions need some polishing: e.g., "*a) monthly calibrated and drift calibrated PM2.5 data as well as for the collocated reference data versus the proxy reference*." This is clunky phrasing.

3.4 Drift detection triggers

Line 301: "*MV-intercept test and MV-slope test"* This is the first appearance of these specific acronym/terms. Again, there should be an "Assessment metrics" section in the Methods to define all of these in advance of the results.

Line 305-307: "*In the IE (RIVR: AQY BD-1146) alarms were related to changes in the MV-slope and clustered around February, May, and September/October indicating more frequent changes in environmental conditions (e.g., RH) or particle composition and size during these months (discussed in sect. 3.5).*" It also appears that there were multiple alarms due to K-S test exceedances, with a much less obvious temporal pattern – why don't the authors mention/discuss these alarms? Further, how can the authors tell that the alarms indicate changes to environmental conditions or particle composition? Maybe they should use the word "suggest" here, instead of "indicate", until clear evidence (sect 3.5) is presented to confirm their suspicions.

Figure 9 – the pattern of the alarms (black dots) is not consistent with the blue line straying outside the threshold (dotted grey line). For example, there are many more instances of the blue line being above 0.05 on the K-S test panel than there are black dots. The authors should discuss the reason for this. The caption should also explain what the dotted grey line indicates.

Figure 10 – The caption should indicate what the smooth lines and the block/cityscape lines indicate.

3.5 Particle composition variability

Figure 11 – Data sets included in your analysis should typically be introduced in the methods, rather than in the caption of a figure. Was this data collected at the RIVR site or just somewhere nearby?

Figure 12 – why is there no data for SE? Indicate what the box and whiskers represent in the caption (mean, median, percentile, min, max, etc.). Months are shown as numbers here, but as their abbreviation in Figure 11. It would be better to use the same style throughout all the figures.

Lines 344-345: "*This is in agreement with observations from Fig. 12 which shows higher concentrations of Crustal Material and Elemental Carbon during N/NE and NW, reaching a maximum in November*." Organic Carbon also appears to follow the same trend for WD as Elemental Carbon, and a somewhat similar trend for the monthly concentration (also peaks in Nov.) Why isn't this discussed?

Figure 13b – this panel should at least be discussed in the text for 1-2 sentences.

Conclusions

It would be helpful for the authors to comment on whether they think either/both MOMA calibrations increase the data quality enough such that the calibrated AQY PM sensor would be acceptable/suitable/useful for supplemental monitoring in this region. For example, under what conditions should a user be cautious when interpreting data from this sensor? And when can a user be confident in the data quality from this sensor? Are there are general trends/red flags the authors noticed?

Did the authors consider using different proxy sites for different times of year? For example, for a collocation site at MLVB, the wind direction data in Fig. 11 suggests that RIVR (located to E/NE) would be a better proxy site for Oct-Mar (SAW), but ANA/CMPT/CELA (any of the sites to the W) might be better proxy sites for Apr-Sep. It seems that allowing the proxy site to vary (with temporally predictable changes in meteorological trends or dominant source emissions) would be a more immediately practical way to allow for a more flexible choice of proxies (versus the authors' suggestion that reference/regulatory monitoring stations be re-sited or that additional ones should be sited in diverse locations to allow for more flexibility in choice of proxies).

This would also be a good place for the authors to compare/contrast with their previous work – how did the best approach for $PM_{2.5}$ and PM_{10} compare to the best approach for O_3 and NO_2? Was MOMA more or less effective at improving the PM sensor data quality compared to the gas sensor data? It would also be helpful for the authors to comment on the ease of use/feasibility for calibrating each of the different PM/gas sensors within the AQY sensor package using different approaches - is this approach manageable for sensor network managers in terms of complexity, computational and human resources, and time? What are the barriers/limitations and benefits of this approach?