Reviewer 3:

Comments:

We would like to thank the reviewer for their comments.

In this manuscript, the authors analyze standard BAM and Purple Air data from three sites in India to identify the optimum calibration procedures. The analysis may include some important and useful results on use of low cost sensor data, but at present, I had a hard time following the authors procedures and conclusions. There is some extraneous results in the manuscript that is distracting (like the theory driven calibrations that are not used) and too little clarity on what the authors did and how their calibrations performed compared to some of the standard published calibrations equations.

We thank the reviewer for asking us to streamline the main research questions we are trying to answer. We substantially edited the manuscript to address the general suggestions to refine our discussion, clarify our methods, and highlight our key contributions.

We agree that theory-driven results may seem distracting since they are negative findings, however, they remain an area of active investigation in the community, so we intend to keep them in the paper. Nonetheless we take this review's constructive feedback and valuable perspective seriously. We have added a section to the SI describing the derivation of the theory-driven equations and have added a new figure to the main text to contextualize the theory-driven results.

It is certainly reasonable to expect that site and season specific calibrations will do better than generic calibration equations. This is stated as a conclusion in the abstract, but its not clearly shown in the manuscript (except in Figure 5). The authors need to provide a table showing R2, NRMSE and bias for the SFS generated values along with those from a more generic calibration, such as that used by the US EPA, either Barkjohn 2021 or the updated EPA equation given here: https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=353088&Lab=CEMM

This would be key to showing that site specific calibrations actually matter. The magnitude matters here. One could argue that improvements of a percent or so in NRMSE or 0.01 in the R2 are rather trivial.

We thank the reviewer for pointing us to add context to the significance of our results. We have added to the SI (Tables S4-S5) demonstrating the relevance of our model forms compared to applying the EPA correction from the peer-reviewed Barkjohn et al 2021. We have added the following details to the Results section.

> In Barkjohn et al 2021a, the large sample size of PA-II across the continental United States was used to derive a similar calibration regression. In SI Tables S4-S5 we compare the NRMSE and MBE for our best CF1 model forms from the SFS procedure (up to 3 parameters), theory-driven CF1 model, and Barkjohn et al 2021 model output. We have found from our seasonally-balanced test dataset that our models perform moderately

> better (ΔNRMSE of about 5% across sites) than the EPA model, which is perhaps intuitive given the differences in PM composition and concentrations in India relative to the US. Furthermore, our site-specific models' MBEs are close to 0 μg/m3, while the Barkjohn et al 2021 model systemically underestimates mass concentrations, with an MBE as high as 22 μg/m3 in Delhi, compared to an MBE of -0.7 μg/m3 using the Delhi site-specific model or 3.25 μg/m3 using the Hamirpur model on the Delhi test dataset. Overall, while the site-specific models we develop here clearly outperform the model of Barkjohn et al (2021a) for these three Indian sites, it is nonetheless striking that this US-developed calibration still performs quite well at these three Indian sites.

In addition, I don't understand how the authors went from the more complex calibration relationships shown in Table 2, to the much simpler relationships shown by equations 2, 3 and 4.

We agree that more details should be shared to demonstrate our model selection methods using more transparently the SFS procedure. We have added text and citations more clearly describing the procedure. We have also added a figure to the SI (the new SI Fig. S16), showing the impact of adding model parameters on adjusted R^2 – which we have already shown to be marginal after adding a second parameter.

> To iterate across all possible arrangements of predictors - including additive terms, interaction terms, as well as polynomial terms up to order 3 – we implemented Sequential Feature Selection (SFS) using the Python package scikit-learn 0.24.2. SFS uses a "greedy" approach to converge on the best-performing model for a user-defined number of parameter (Raschka and Mirjalili, 2019; James et al., 2013; Ferri et al., 1994). For example, if a user wanted a 2-parameter model from a set of 10 features, SFS would iteratively compare 90 models, the set of all possible 2-parameter feature permutations, using a robust regression metric (such as adjusted R^2 or Bayesian Information Criterion [BIC]). In our approach, we first use SFS to define the best-performing n-parameter model starting with all possible parameters (n=34). We then compare adjusted R^2 across best-performing n-parameter models to measure the impact of model complexity. If increasing parameters results in only marginal improvements ($\Delta R^2 \approx 0.01$), then it is unnecessary to use those additional features. The overall most robust model, therefore, reflects both the best possible selection of features as well as feature parsimony.

> Temperature and dewpoint terms only imparted marginal improvements to calibration models ($\Delta R^2 \approx 0.01$, see SI Fig. S16), and it is not determinable if the models are deriving a spurious correlation or detecting underlying aerosol or instrument properties.

Other comments:

Line 76: need to provide equations for "Alt" corrections in SI.

The ALT data is now common in literature as well as directly available from PurpleAir, therefore we direct readers to Wallace et al 2021 and Wallace et al 2022. SI Fig S1, panel b illustrates the relationship between the ALT values and CF1 data. We have added detail to the methods section as well.

> Briefly, the ALT method adds all the particle counts from bins less than 2.5 µm, and calculates the particle volume concentration assuming spherical particles. Particle volume concentration is then multiplied by unit density (1 g/cm^3) to estimate PM$_{2.5}$ mass concentration.

77: what does three refer to?

We have clarified be replacing "three" with "CF1, ATM, and ALT."

83, 90: grammar issues.

We have fixed these grammatical errors.

150: What are "unreasonably large"? What is "block average"? Same as hourly average?

"Unreasonably large" is defined in Line 153 as the operational range (5- 500 µg/m^3)of the instrument as described in the cited literature.

We have reviewed the entirety of the text and replaced "block average" with "hourly average."

155: Not sure what you mean by statistically paired.

Statistically paired refers to independent datasets for each observation that can be considered "coupled" or matched because they are from the same location and are therefore expected to draw from the same underlying distribution. We have rephrased for clarity.

> Analyses of PurpleAir data typically report the percent error between channels A and B for a given unit to remove imprecise points, treating them as joint measurements and all other nodes as independent…

164: I don't think BME 280 is defined anywhere.

BME280 is not an initialism, it simply refers to the model's name. We have ensured that is properly contextualized here.

> The Adafruit model BME280…

177: Not just RH, many other factors.

We have added another sentence explaining the roles of other key factors such as particle size distribution and index of refraction – although RH strongly moderates these factors, especially within the context of LCS observations.

> "…calibration procedures attempt to account for bias due to RH, index of refraction, and mischaracterizing the particle size distribution…"

211: Grammar.

We have fixed this grammatical error.

227: Suggest a citation for defs of these std relationships. (MBE, NMBE, etc)

We have used the citation Simon et al. 2012, and James et al. 2013 to clarify these points.

235: This is not a bias.

We understand this term may be confusing and instead have referred to it as mean difference.

> "… a mean difference from the regulatory network average of…"

284: Suggest this ref for quantitative analysis of dust with Pas

https://doi.org/10.5194/amt-16-1311-2023

Thank you for sharing, we have added it to our discussion on the impact of dust in the results and conclusion sections.

292: Table 2 does not summarize the procedure, but rather results.

We have clarified this in caption by replacing "procedure" with "results."

294:   This does not seem to be true for Bangalore, PM x temp is the most relevant, right?

We have clarified that all sites selected a term which included a PM parameter.

> " The form of the most robust Bangalore model is different from the North India sites with an interaction term between temperature and ALT…"

295: I assume this refers to RH SQUARED, right?

Thanks, we have updated our formatting to ensure that the superscripts are properly formatted.

323-325: Not sure what this refers to.

We have added detail to the methods section to document the different data channels more precisely and clearly.

PurpleAir reports mass concentrations from PA-IIs in three forms, referred to as CF1, ATM, and ALT. CF1 ("Correction Factor 1") is the "uncorrected" data from the Plantower. The CF1 data has been demonstrated to strongly correlate with collocated integrating nephelometer data (Ouimette et al., 2021). ATM or "Atmospheric Corrected" data uses a piece-wise function to attempt to account for overestimation. SI Figure S1 illustrates this function across the full dynamic range for data collected in Delhi. Between 0 - 25 µg/m3, the CF1 and ATM data are 1:1, between 25 - 40 µg/m3 the ATM to CF1 ratio transitions from 1:1 to approximately 0.7:1, and at greater than 40 µg/m3 the ATM to CF1 ratio is stable at 0.7:1. Although it is reasonable to hypothesize the ATM data may better represent exposure ambient PM2.5 than the CF1 data, there is no transparent reasoning in the user manual for this design choice (Wallace et al., 2021; Zhou and Zheng, 2016). Finally, the ALT data represents a reconstruction of the PM2.5 data from the particle number data reported by the Plantower….Wallace et al. (2021) and Wallace et al. (2020) used this data to develop calibration relationships, reporting the ALT data as more transparent than using the CF1 or ATM data. However, the particle number data is known not to reflect the actual ambient size distribution since the Plantower PMS5003 is not a particle sizing instrument, but rather reflects a modeled size distribution using assumptions for relationships between size bins that is not always accurate for atmospheric conditions (Ouimette et al., 2021; Hagan and Kroll, 2020; He et al., 2020; Kuula et al., 2020). SI Figure S1 shows the ALT to CF1 ratio is approximately 0.15:1.

350 and equations 2,3 and 4:

After describing the details of a multi-linear SFS model, and showing in Table 2 how various permutations of the parameters give the best fits, I don’t understand how you arrive at these much simpler relationships.   These look like very standard PA calibration equations that have been developed by others.

As described in detail in our response to major comments, the SFS approach is “greedy” – it considers all possible permutations for a given number of features. Perhaps unsurprisingly given the underlying instrument and aerosol properties, when we limited models to only 2 parameters, we arrived at the same model formulation in Delhi and Hamirpur as other studies. Using more than 2 parameters via SFS offered only marginal improvements (see the new SI Figure S16). Therefore, although we started with many features and possible model forms, we converged at a form common to the literature. The broad continuity of the calibration form across geographies is valuable information for the community in India, since $PM_{2.5}$ concentrations and speciation are much different than in North America, Europe, and Australia.