**Reviewer 2:**

**Comments:**

We would like to thank the reviewer for their comments.

Overall, this paper is well-written and was completed in a systematic manner. The main drawback is it fails to identify and highlight innovation in the work. The stated aim of the paper was to “identifying robust calibration protocols”, which feels like a step in completing the QAQC process. As a whole, the Plantower/purple-air pm sensor's accuracy and precision are well-studied. I recommend that the authors revisit the abstract and introduction in order to highlight the contribution to the scientific body of knowledge that this work provides.

We thank the reviewer for the opportunity to clarify our objectives and novel contributions. We have reworked the manuscript abstract accordingly.

> Lower-cost air pollution sensors can fill critical air quality data gaps in India, which experiences very high fine particulate matter ($PM_{2.5}$) air pollution but has sparse regulatory air monitoring. Challenges for low-cost $PM_{2.5}$ sensors in India include high aerosol mass concentrations and pronounced regional and seasonal gradients in aerosol composition. Here, we report on a detailed long-time performance evaluation of a popular sensor, the Purple Air PA-II, at multiple sites in India. We established 3 distinct sites in India across land-use categories and population density extremes (North India: Delhi [urban], Hamirpur [rural]; South India: Bangalore [urban]), where we collocated the PA-II with reference beta-attenuation monitors. We evaluated the performance of uncalibrated sensor data, and then developed, optimized, and evaluated calibration models using a comprehensive feature selection process with a view to reproducibility in the Indian context. We assessed the seasonal and spatial transferability of sensor calibration schemes, which is especially important in India because of the paucity of reference instrumentation. Without calibration, the PA-II was moderately correlated with the reference signal ($R^2$: 0.55 - 0.74) but was inaccurate (NRMSE $\geq$ 40%). Relative to uncalibrated data, parsimonious annual calibration models improved PA performance at all sites (cross-validated NRMSE 20-30%, $R^2$: 0.82-0.95), and greatly reduced seasonal and diurnal biases. Because aerosol properties and meteorology vary regionally, the form of these long-term models differed among our sites, suggesting that local calibrations are desirable when possible. Using a moving-window calibration, we found that using seasonally-specific information improves performance relative to a static annual calibration model, while a short-term calibration model generally does not transfer reliably to other seasons. Overall, we find that the PA-II can provide reliable $PM_{2.5}$ data with better than ± 25% precision and accuracy when paired with a rigorous calibration scheme that accounts for seasonality and local aerosol composition.

Abstract:

I suggest strengthening the aims in your opening line since there are now a lot of calibration papers for the Plantower/purple air. I.e., what are you adding to this body of literature?

We agree and have strengthened the title as well as the abstract.

> Seasonally Optimized Calibrations Improve Low-cost Sensor Performance: Long-term Field Evaluation of PurpleAir Sensors in Urban and Rural India

Clarify why these three are distinct and what they add to the study (e.g., urban, suburban, background, forested, etc.)

We agree these distinctions are important and have included the terms “urban” and “rural” in the abstract. We further describe each site in detail in the methods section and SI.

Can you clarify what a “major season” is?

We agree this term may be confusing and removed the line as it is extraneous.

It would be useful to the reader if you briefly state how the aerosol and meteorology vary by the site since I assume they capture unique environments.

We agree and have described them in the methods section.

> The National Capital Region along with the rest of North India experiences dynamic meteorology with cold wet winters, warm drier post-monsoons and pre-monsoons, and hot wet monsoons (SI Fig. S4)… During the course of our campaign, Delhi experienced extreme $PM_{2.5}$ concentrations during the post-monsoon agricultural burning seasons and characteristic winter inversion layers, with a relatively low-pollution monsoon season consistent with expected seasonal trends…
>
> …Although campaign median $PM_{2.5}$ concentrations at the site (Table 1) are high in the global context, this site’s remote location outside of both cities and villages means that concentrations do not reach the same peaks as in Delhi. However, there are still many local sources of aerosol air pollution in rural North India such as biomass burning for cooking and heating…
>
> In Bangalore, emissions are dominated by traffic and dust resuspension… Compared to Delhi and Hamirpur, winters are milder, and the climate is more consistent year-round in Bangalore (SI Fig. S6).

This sounds like a more innovative part of the work, expand?
We used a comprehensive feature selection process to create optimized site-specific calibrations.

We agree and have added relevant details in the methods section.

> 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.

Since the form varies by site, do you make a recommendation to other users on what to include in their model?

We agree this is valuable information for the community. From our analysis, we show that the calibration equation is very similar for our two sites in North India region, so in our results and conclusions we endorse using this form across distinct settings in this region. Furthermore, we demonstrate that the general form for North India only marginally degrades the best fitting parameters in Bangalore, so it is acceptable for usage based on our work. We discuss this point at length in the results and conclusions.

Can you clarify how it's “successful” if it does not work overall?

By “successful” we mean robust performance metrics within the training season. We have clarified the language to be clear we are referring to within-training season performance in this sentence.

> In contrast, we demonstrate that a short-term calibration exercise for one season with robust metrics within the season may not transfer reliably to other seasons.

Introduction:

Ln 39: Is “mischaracterize” the correct word here? I am confused by the goal of the statement.

We have edited to use more precise language.

> Optical sensors inaccurately estimate mass from aerosol scattering properties, since $PM_{2.5}$ is a mixture of particle sizes and chemical compositions thus resulting in spatial-temporal variability in optical properties…

Line 41-45 – This fails to cite enough work to support this statement. I suggest including more work from the past 2-3 years. This will also help you identify the innovative part of this work. Several LCS PM networks have thorough publications on calibration methods based on long-term field data.

We appreciate the rapid growth of research efforts in developing low-cost sensors networks, and have added more recent citations. It is worth noting the references in the first edition of the manuscript were strategically selected to emphasize efforts within India, given the difference in both aerosol regimes and infrastructure between the US (where a plurality LCS literature is based) and India.

The end of the introduction feels more like concluding remarks.

We recognize that this may be a matter of differing styles or tastes, but we believe that a brief paragraph summarizing of the approach and results in the introduction section is reasonable and provides a structure that improves the accessibility and readability of the rest of the paper. We'd welcome input from the AMT editors if they disagree with this stylistic choice.

Methods:

150: Please define "block-averaged" in the text.

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

Ln 151: How do you determine what constitutes "imprecise points"?

> CV using the 6 nodes?

We calculated the Coefficient of Variation on all available Plantower signals, so for 3 collocated PA-II that would be the CV of 6 data points.

> … if we had three PA-IIs at a site, we averaged the six values together – two from each unit – to estimate a single data point. We established 80% completeness criteria (or 24 2-minute data points) for each hourly average, and at least 2 valid Plantower hourly averages for the resulting site PA datapoint. Imprecise site points were removed using the coefficient of variation (CV), the quotient of the standard deviation, and the mean of the collocated Plantower sensors for a given 2-min raw sample. CV values greater than 0.2 were removed, broadly consistent with approaches used by other studies…

Ln 160: "the quotient of the mean and standard deviation of the sensors" – How are these values were used?

We agree this point is unclear. There is a typo in the text, we have rephrased.

> We established 80% completeness criteria (or 24 2-minute data points) for each hourly average, and at least 2 valid Plantower hourly averages for the resulting site PA data point. Imprecise site points were removed using the coefficient of variation (CV), the quotient of the standard deviation, and the mean of the collocated Plantower sensors for a given 2-min raw sample. CV values greater than 0.2 were removed, broadly consistent with approaches used by other studies.

Ln 185: I think this is partially correct. Testing has shown that it exponentially overestimates at high RH, but these conditions are less likely to be sustained in real world environments.

We have adjusted our language accordingly.

> Although the theory-driven model should produce the most transferable models since theory should apply in all environments, the underlying data processing of the Plantower - a truncated nephelometer … may result in a bias structure better explained by a linear RH correction than an non-linear correction for the dynamic range of RH under real-world conditions.

Can you also provide local regulatory values in addition to WHO?

We agree this will add valuable context.

> While the annual average is low in comparison to cities in North India as well as the Indian National Ambient Air Quality Standard of 40 μg/m$^3$ it exceeds the WHO annual guideline value of 5 μg/m$^3$ and hourly winter concentrations often exceed 50 μg/m$^3$. Consequently, Bangalore has been designated for air quality improvement under the Indian National Clean Air Programme…

Please provide a reason for “We removed all raw PM2.5 data points outside of the range 5 – 500 µg/m3”

Strongly contradicts “with peak daily (hourly) in excess of 500 µg/m3” line 103

From the specification sheet for the instrument as well as previous performance characterizations cited in text, we identified 5- 500 μg/m$^3$ as the operational range of the instrument on an hourly basis. We acknowledge our assertion that the daily maximums in exceedance of 500 μg/m$^3$ was used as motivational text to emphasize the urgency of PM$_{2.5}$ air pollution in India could cause confusion with our later implementation of this quality assurance procedure. However, from our collocation study, applying this filter only removed about 1% of hourly data points from the raw dataset. Therefore, we have clarified the text by removing line 103.

How many minutes were required for the data to be averaged up to 1 hr? Is that the 80%?

Yes, the 80% refers to the number of 2-min PA-II data points required for a valid hourly average. We have clarified the text this corresponds to 24 datapoints/hour.

> We established 80% completeness criteria (i.e., 24 2-minute data points) for each hourly average…

You should cite the sources that influenced you to choose these covariates (e.g., ln 176, 188, etc.)

We have ensured the relevant sources are properly cited throughout the text.

Ln 215: A similar method was employed in "Identifying optimal co-location calibration periods for low-cost sensors." Compare results?

We agree there are some similarities in our approaches, we will ensure this paper is referenced, and used to contextualize our method. We agree with the findings of Levy Zamora et al 2023, that there are optimal collocation periods – in our case the Post-Monsoon – which are more transferrable than other periods. We have added these details to our results.

> Previously Levy Zamora et al. (2023) identified diminishing returns in improvements to calibration regressions after about 4 weeks of collocation in Baltimore, USA, if that period encapsulated a representative range of $PM_{2.5}$ and RH conditions. Here we build on this work by seeking to identify which 4-week period is ideal at our sites in India since annual median $PM_{2.5}$ concentrations at Delhi and Hamirpur sites are about 10× higher than Baltimore and reflect a different mixture of chemical composition and aerosol properties.

Ln 230. Please clarify "US EPA's data reduction process"

We have added a citation to the EPA SOP for maintaining and managing the BAM-1020 at embassy and consulate sites, which describes the process used by the State Department AirNow network in compliance with EPA protocols.

Ln 249: missing word?

We removed it to fix this grammatical error.

A general comment on the results: it is very acronym heavy, and I think it sometimes takes longer to mentally decrypt than it would be if it was just written out.

We have reviewed the text and clarified as necessary. For example, instead of referring to both the site and the location (ie, IGP-CARE vs Hamirpur) we have simply referred to the location.

Ln 363 – extra comma

We have fixed this grammatical error.

Ln 369 – Can you add the reasons for this trend here?

How do you think the notable differences in data likely influenced some of the transferability?

We add details to better describe this trend.

> Monsoon meteorological conditions contrast with other seasons – it is humid, windy, cloudy, hot, and frequently rains (SI Figs. S4-S6). These conditions result in lower emissions (i.e., less biomass burning for heating relative to winter), as well as act to suppress emissions (i.e., wet deposition) resulting in lower average seasonal mass concentrations in the Monsoon (SI Figs. S3 and S7). Consequently, models trained in the monsoon poorly translate to other seasons.

In my experience, LCS struggle more at high concentrations. Can you discuss concentration ranges more?

We explore this trend in SI Figure S17 panel a, which is referenced in the main text. While we do see performance degradation as mass concentrations increase, our calibration residuals are not sensitive to mass loading. We have added an additional sentence in Ln 347 around the discussion of SI Fig S17 and the impact of mass loading.

> The calibrated residuals distributions demonstrate marked improvements across the full range of mass concentrations, unlike the raw residuals which show increasing uncertainty at high mass concentrations.

How do you think the difference in complexities affected the transferability? Since they are quite different, they may be overfitting and that reduced the transferability too. It would be interesting to see something like Figure 5 all with the same model structures.

We agree model complexity likely impacts transferability. Figure 5 (Figure 6 in the new revised manuscript) already uses the same model structure for each model design according to equations 2-4. We have clarified this point in the figure caption.

> Performance evaluation metrics of Eq. (2-4) with the training site on the x-axis and the test site on the y-axis…

Figure 1: To clarify, is 779 the number of data points total (could be 10 points for 1 am and 100 points for 9 am) or the number for each hour? If it's the first, did you check that there are a comparable number of points per hour?

Yes, it is the total number of hourly averages. We appreciate the concern that some hours may not be represented properly, however, we found close to uniform coverage, with hourly totals contributing no more than ~ 7% to the total number of data points on a given plot, compared to the ideal ~ 4% (1/24). Therefore, we do not expect a systemic bias based on data availability of one hour in a season versus another. We have amended the figure caption accordingly.

> No single hour of day represents more than about 7% of the total dataset shown in the bottom left corner of each plot.

Figure 1 would also be nice with the range highlights like Figure 2.

Thanks for this nice idea. We agree that including the range helps to better illustrate the statistical properties of the distributions, however in Figure 1, there is overlap in the diurnal profiles at some key periods (such as in Delhi, panel b). We experimented with adding this shading, but ultimately concluded that adding the shading for the distribution ranges added so much visual clutter that it would have made the figure less interpretable.

Figure 3 – Check this figure for visual accessibility.

Thanks for the suggestion. We have updated all our figures to ensure visual accessibility using the colormaps recommended by Crameri et al 2018.

Figure 5 – Is there a training/testing split in time for each site? I am confused as to why the site applied to itself changes.

The test-train split here is the same split used to derive and evaluate Eq. 2-4, such that the error metrics reflect evaluating each model from the y-axis exclusively on the test data described in the methods section, rather than on the training data used to generate each model. We have amended the figure caption accordingly (now called Figure 6 in main text).

> At each site, we compute performance metrics by comparing the calibration model output to an independent test set that was held out from model training.

Conclusions:

Ln 416 - extra comma

We have fixed this grammatical error.

Ln 423 – Dust storms were not discussed with these numbers in the text. Please add a discussion so it is appropriate as a conclusion.

We agree the language suggests stronger than presented findings on the influence of dust storms. We therefore have rephrased.

> We identified periods of low-cost sensor signal underestimation by a factor of 2 – 6× in the Pre-Monsoon period at the Delhi and Hamirpur sites, when supra-micron wind-blown dust particles are relatively abundant.

Ln 430 – Some of the thoughts on seasonal and location transferability have been described elsewhere, but the discussion on the difference in the PM composition of the sites is interesting. Do you have thoughts on how the community can account for these differences if co-location is not feasible?

We agree that PM composition plays a key role in spatial transferability. We have added these relevant details to the text.

> Based on our analysis, we hypothesize that it is better to use a model developed at a background site such as Hamirpur to correct data from an urban environment such as Delhi, since the composition of PM in Hamirpur represents a good subset of the variability in Delhi. On the other hand, since there are PM species only found in some urban environments in India using models from these industrial microenvironments are less likely to produce reliable results outside of the training location.