Reviewer 1:

Comments:

We would like to thank Dr. Subramanian for his comments and transparency.

Disclosure: I will be joining CSTEP in July, where two coauthors were working during this study (and one continues to be part of CSTEP and will be reporting to me). CSTEP is also one of the three sites in this study.

Overall, this is a thorough study on LCS performance and correction at three different sites across India, nicely presented. My comments/questions are mostly minor. However, I am not sure whether the title is entirely correct - monthly corrections aren't a great improvement and don't even always carry over to other months of the same season. The final recommendation is to collocate at least one sensor with a reference for the entire study duration. As for site-specific, the Hamirpur and Delhi corrections appear reasonably interchangeable (see comment 20).

With feedback from all reviewers, we have decided to remove the phrase "site-specific" from the title, in line with the more recent direction in literature focusing on calibration time-periods (Levy Zamora et al 2023)

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

Comments:

1. Abstract uses both Pearson r and R^2; please use one or the other for consistency. The R^2 values for "raw" (I call it uncorrected) data (0.55-0.74, Fig S16 - which show the final corrections in much better light!) show sensor performance that is not as impressive as r >= 0.9. Incidentally, the Pearson r result only appears in the abstract and not in the main text.

For the purposes of consistency, we have amended the text to uniformly use the Coefficient of Determination (R^2) rather than Pearson's r.

> Without calibration, the PA-IIs were moderately well correlated with the reference signal (R2: 0.55 - 0.74)…

2. Showing a table of fit statistics for the uncorrected and final corrected data (and maybe the spatial transferability results) in the main text would improve clarity. Currently, these key results are discussed in the text but only presented graphically in Figure S16.

Thanks for this suggestion. We have improved clarity by promoting Figure S16 to the main text as the new Fig. 3. Upon transferring the figure, we identified a transcription error in the NRMSE matrix resulting in the incorrect numbers displayed in the Bangalore panel as well as minor rounding issues in the other panels. This correction has no impact on interpretation.

3. Line 43: The Plantower sensors do sense particles above 0.8 micron, even up to 2 micron - just not very efficiently. Kuula et al. say 0.8 micron but based on "Valid detection ranges...defined as the upper half of the detection efficiency curve" - which seems different from "failing to characterize". Maybe something like "do not adequately characterize".

We agree and have rephrased.

> "…do not adequately characterize fine particles above 0.8 microns…"

4. From Wallace et al. (2021): "The ALT method is based on the number of particles per deciliter reported by the PMS 5003 sensors in the PurpleAir instrument for the three size categories less than 2.5 μm in diameter." It is unclear that these are independent measures; Kuula, He/Dhaniyala, Ouimette, Andy May, and others have shown the size distribution is not real. Since the ALT method isn't finally used, maybe move these results to SI and improve clarity by focusing on the two metrics (CF1 and ATM) that most people use anyway.

We agree that the citations in the reviewer's comment are strong evidence against treating the size resolved data from Plantower sensors as true size distribution measurements. Nonetheless, since publications with the ALT method have grown in popularity and is featured on the PurpleAir map, we sought to highlight negative findings as evidence against its application. Therefore, we believe it would be best to keep these findings in the text as a point of contrast to the CF1 and ATM data. We have increased the level of detail in which we describe the differences between CF1, ATM, and ALT as well as provided a brief justification for reporting calibration results for each data channel.

> 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. Although the CF1 and ATM data have dominated most calibration efforts (Malyan et al., 2023; Puttaswamy et al., 2022; Barkjohn et al., 2021; McFarlane et al., 2021; Magi et al., 2020; Malings et al., 2019), the usage of ALT data continues to propagate in peer-reviewed literature (Wallace and Zhao, 2023; Wallace and Ott, 2023). Therefore we use CF1, ATM, and ALT in our study to work towards harmonizing a calibration approach for PA-II in India.

5. Line 83: "while can the BAMs provide" should be "while the BAMs can provide"

We have fixed this grammatical error.

6. Line 150: Instead of "block averaged", recommend using "hourly averages of" - because I think that's what is being done. "Block averaging" is not otherwise clarified in the manuscript and "hourly averaging" is easily understood.

We have reviewed the entirety of the text and replace “block average” with “hourly average.”

7. Line 160: "the quotient of the mean and standard deviation" seems the inverse of the CV - might relative standard deviation be easier to understand?

Thanks for catching this error in our text. We reviewed our code pipeline and ensured that we employed the Coefficient of Variation as it is conventionally defined: $CV = \frac{\sigma}{\mu}$, therefore this is a typo in the text. We have rephrased.

> “… the quotient of standard deviation and mean…”

8. Eq. 1 is an unusual formulation, so perhaps the original study that used this formulation (as far back as I can track it!) should be cited?
    1. Zhang et al. (1994) https://doi.org/10.1080/1073161X.1994.10467244 (Their Figure 4 was used in a workshop report Laulainen et al. 1993 that was then cited by Chakrabarti et al. 2004.)

We recognize by skipping the exact derivation of our form, we may have omitted some key details on how Eq. 1 is related to the form cited in Chakrabarti et al 2004. We have added this section to the SI:

> The correction equations used in Laulainen et al. 1993 and Chakrabarti et al. 2004 take the form:
>
> (1) $CF = 1 + 0.25\frac{RH^2}{1-RH}$
>
> (2) $C_{corrected} = \frac{C_{raw}}{CF}$
>
> Where RH represents the fractional RH, expressed on scale of 0 – 1, CF represents the hygroscopic correction factor, C_{raw} represents the light scattering instrument PM mass concentration, and $C_{corrected}$ represents the PM mass concentration corrected for hygroscopic growth. Combining the two equations yields:
>
> (3) $C_{corrected} = \frac{C_{raw}}{1+0.25\frac{RH^2}{1-RH}}$
>
> From Laulainen et al. 1993, the selection of 0.25 in the denominator of (3) was found to vary with chemical composition, and in fact suggests a value of 0.328 at a site dominated by ammonium sulfate. As speciation across India is known to be strongly variable dependent based on seasonal and diurnal factors, we instead allowed for a best-fit approach described in the text to select the best fitting factor yielding the following:

(4) $C_{corrected} = \frac{C_{raw}}{1+\beta\frac{RH^2}{1-RH}}$

Furthermore, given that (1) was derived for use with integrating nephelometers, we chose to include an additional term (α) in the numerator to account for the differences in instrumentation of the truncated nephelometer. Therefore, we arrived at:

(5) $C_{corrected} = \frac{\alpha \times C_{raw}}{1+\beta\frac{RH^2}{1-RH}}$

Finally, we replaced C_{raw} with P to represent any $PM_{2.5}$ mass concentration signal (CF1, ATM or ALT) from the PurpleAir, and simplified $C_{corrected}$ to C.

(6) $C = \frac{\alpha \times P}{1+\beta\frac{RH^2}{1-RH}}$

9. Line 216 says the rolling OLS performance was compared against other two-week periods, but the results suggest monthly evaluations. Please clarify.

This is a typo from an earlier iteration of the analysis and has been corrected from "2-week periods" to "4-week periods."

10. Lines 263-264: 15% and 14% seem not that different to warrant an explanation.

We agree, this sentence has been abridged for conciseness.

The CV test removed about 15% from each site.

11. Lines 267-268: Why is Delhi so unusually lossy? Both IGP sites are significantly lossier (data recovery <40%) than the CSTEP site (75%), which is surprising and not well explained given e.g. my comment #11 that even a 1% difference in the results was explained even if not seemingly necessary.

We agree it is necessary to offer summary comment on the dataset losses. First, the BAM in Bangalore failed less within the collocation periods than the BAMs in the IGP (Delhi and Hamirpur. Also, the CSTEP site in Bangalore was much better staffed allowing for fewer gaps due to power or Wi-Fi fluctuations, and maintenance such as BAM tape replacement. We have added these details to the end of the paragraph.

The smaller number of data points available for the Delhi and Hamirpur sites principally arose because of relatively more downtime of the BAM instruments at these two locations.

12. Lines 280-282: This is unclear from the figure, which I interpreted as "the PA line is mostly above or close to the BAM line for the pre-monsoon period at Hamirpur; it underestimates about half the time for Delhi, maybe."

We agree this phrasing requires some streamlining and have accepted similar language to the reviewer's suggested wording.

13. Line 282: coarse aerosol are particles larger than 2.5 micron. $PM_{2.5}$ is called fine PM, e.g. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

We agree with the spirit of this comment and have rephrased the sentence for clarity, as presented below.

Here, we intended to refer to the fact that the coarse-*mode* of a particle size distribution typically is not monodisperse, but rather includes a lower tail of ~ 1-2.5 µm sized particles that contributes to $PM_{2.5}$ mass. While this is not generally a major contributor to $PM_{2.5}$ loadings under ordinary conditions, dust storms or other events that produce predominantly coarse aerosol can lead to elevated $PM_{2.5}$. See, for example, the archetypal size distributions of Seinfeld and Pandis 2016 (3rd Edition of Atmospheric Chemistry and Physics) Figure 8.10, which illustrate that while the median of the coarse mode is >2.5 microns, the tail of the coarse mode extends well below 2.5 microns.

> Although both the Delhi and Hamirpur sites feature relatively low bias in the pre-monsoon period, they underestimate mass concentrations in this season, perhaps due to the influence of wind-blown mineral dust, as observed elsewhere in field and lab evaluations…
>
> While crustal material does not generally dominate $PM_{2.5}$ mass, during dust storms the lower tail of the coarse mode aerosol can lead to substantially elevated $PM_{2.5}$ concentrations in India.

14. Lines 297-301: I appreciate this decision. Good call.

Thanks, glad you agree!

15. Line 306: I don't think T & dewpoint coefficients were reported anywhere. Are these statistically different from zero?

While the coefficients are not statistically different from zero, adding terms with temperature and dew point only resulted in marginal improvements ($\Delta R^2 \approx 0.01$). Therefore, we have clarified this comment by changing the phrasing.

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

16. Lines 326-327: Text is unclear ("CFs"??) and not presenting these key results in the main text (ideally as the table requested earlier) is not helping.

We agree the text requires clarification. In addition to adding SI Fig S16 to the main text we have rephrased the sentence.

> The theory-driven hygroscopic growth correction consistently improved performance from the uncalibrated baseline data across sites by 12% for ATM and 60% for CF1, on average (Fig. 3).

17. Fig S16 uses "RHC" to indicate the theory-driven fRH-esque approach, which is odd. What does RHC stand for? Maybe just say "theory".

We have reviewed the entirety of the text and replaced "RHC" with "theory-driven."

18. The discussion of Fig S16 doesn't align with the results shown. The Bengaluru theory-driven performance has the same R^2 as the 1-parameter model, lower than the 2-parameter model. NRMSE is lower for theory model than any of the empirical models. (One might quibble about differences of 0.02 and argue that is "comparable", but the surrounding text touts differences of 2-3% so...)

We agree this text should be more consistent. In addition to promoting SI Fig S16 to the main text (as the new Fig. 3), we have harmonized the analysis of the results it illustrates to clarify we identified differences of more than 3% NRMSE or R^2 as clearly more robust. Therefore, we have rewritten this section to state the theory-driven models are more less robust than the data-driven in North India (Delhi, Hamirpur), and offer only marginal benefits in Bangalore. Therefore, we choose to uniformly apply a data-driven model.

> "…the 2 parameter CF1 models in Delhi and Hamirpur, with their additive RH terms, outperformed theory-driven by at least 3%. In Bangalore, the theory-driven model performance was comparable to the data-driven models (about 1% NRMSE, see Fig. 3). This contrast in performance between the two methods in Delhi and Hamirpur is likely a result of the less seasonally variable meteorology and source mixtures in Bangalore, leading to less dynamic aerosol hygroscopicity."

19. Eq 2-4 - is RH used as a fraction in these equations? Please specify. RH is usually reported as % and can be used directly in such equations, so maybe use that convention instead.

To maintain consistency across data-driven and theory-driven approaches we choose to use the fractional form. For consistency with other models from literature, we have re-written these calibration relationships in terms of %.

20. Line 408: Unclear if this parenthetical is really the case. Applying Hamirpur correction to Delhi or vice-versa produces relatively similar R^2 and NRMSE 0.82/39% or 0.78/35% with no clear winner.

We agree the is not strong enough to fully support the assertion and have removed the phrase.

> Clearly, the differences in composition of the Delhi and Bangalore aerosols prevents exchange between models at these two sites, but with enough preserved from the regional contribution to support some support from the Hamirpur model to the Delhi data.

21. Lines 422-423: Dust storms were not identified nor discussed elsewhere in the text. Dust is only hypothesized as a potential explanation for a result (lines 280-284).

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

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

22. Lines 436-437 - check the sentence.

We will fix the grammar.

23. Data availability: insert data repository link.

We have setup a repository on Dryad and will add the link before final publication.

24. Vos et al. is almost three pages of authors for one reference that doesn't actually contribute much to this specific manuscript. Can you just say "GBD 2019 Diseases and Injuries Collaborators" as the group is known on the paper?

We agree the Vos et al citation full form is unnecessarily large and have replaced it as recommended as well as Pandey et al.

25. Fig 4 caption is really long, but has a simpler explanation of the results than what is in the text...

We have refined the Fig. 5 (figure formerly referred to as Fig. 4) caption to describe the take-home message of the figure more concisely.

> Assessment of inter-seasonal transferability of seasonal models. Panel (a) depicts box plots of the distribution of Normalized Mean Bias Error (NMBE) for a given model starting month of a 4-week rolling ordinary least-squares (ROLS) model on all other windows. The bottom, solid line, and top of the boxes represent the 25th, 50th, and 75th

percentiles respectively. Panel (b) presents the median NMBE of a 4-week ROLS model trained to start in the month (colored by season) on the x-axis and evaluated on all other windows as binned by starting month on the y-axis. Gray boxes represent months without sufficient data. Models trained in the Pre-Monsoon underpredicted in other seasons, contrary to the typical pattern of overprediction – this pattern is consistent at Delhi and Hamirpur. As a point of comparison, we present the performance of our long-term calibration in individual months at each site in column (b) titled "All." Consistent with our observation that 4-week models trained in a single month generally do not perform as well in other months.