Comment on AMT-2022-140
Anonymous Referee #1


General Comments

This manuscript describes the development and performance of a machine learning algorithm that used PM$_{2.5}$ concentrations reported by PurpleAir sensors, temperatures reported by PurpleAir sensors, and relative humidity values reported by PurpleAir sensors to predict the PM$_{2.5}$ concentration that would be reported by a PM$_{2.5}$ monitor with the U.S. Environmental Protection Agency Federal Equivalent Method (FEM) designation. Two different machine learning algorithms were trained and testing using data from Raleigh, NC, USA and Delhi, India, respectively. It's nice that the authors' examined the performance of the PurpleAir sensors in two locations with very different levels of PM$_{2.5}$ pollution. I'm most concerned that a) the authors might not have used appropriate techniques to split their data into training and validation sets and b) model performance metrics seem to be presented for predictions calculated using the same data that were used to train the model. I'm also concerned that most of the PurpleAir sensors tested in Raleigh were not collocated with the FEM and it's not clear what was done with the data from the non-collocated sensors.

Specific Comments

1. Page 7, Lines 15-19: “In order to develop the final sensor calibration algorithm, we used the following steps: 1) quality control the PA data; 2) randomly divide the data into training (75%) and validation (25%) datasets; 3) train the algorithm using the training dataset and validate using the validation dataset; and 4) repeat steps 2 and 3 ten times (i.e., 10-fold cross validation) using random data selection.”

My first concern is that steps 3 and 4 seem inconsistent with each other and with the authors’ presentation of their results. 10-fold cross validation typically implies that: i) the data were split into ten folds, ii) nine folds (90% of the data) were used to train the model
while the remaining fold (10% of the data) was used to test the model, and then iii) step ii was repeated for all ten folds. I don’t see how a 75% training/25% validation split corresponds to 10-fold cross validation. Why are results presented for both the training and validation data in Table 1 and Figure 5? Performance metrics should not be shown for predictions generated using the same data that were used to train the model. How come training/validation data are shown separately in Figures 4 and 5, but together in Figures 6, 7, 8, and 9? Were the results shown in Figures 6–9 generated using 10-fold cross validation? Or were 75% of the data points shown in Figures 6–9 generated from the same data used to train the models?

My second concern is that the authors should not use random data selection to split the data into 75% training/25% validation or to split the data into folds for cross-validation. The assumption that observations are independent and identically distributed does not apply to grouped data or to time series data. The authors’ data are both grouped and time-series. The data are grouped by PurpleAir sensor: the authors have multiple PurpleAir sensors and multiple observations from each of those sensors. Data from each individual PurpleAir sensor might be dependent upon properties of that sensor. Time-series data are autocorrelated: each observation is related to some number of observations recorded immediately before and immediately after that observation. One approach for time-series data is to split up the data such that the training and validation sets do not overlap in time. If the training and validation data sets overlap in time, then the metrics reported for the validation set represent an overly-optimistic picture of the model’s performance.

References:

- https://scikit-learn.org/stable/modules/cross_validation.html#group-cv

2. Section 3.2: Only 5 of the PurpleAir sensors were collocated with the FEM in Research Triangle Park for two weeks. The other 84 PurpleAir sensors collected data at a residential location in Apex, NC. Were data from the 84 PurpleAir sensors that were never collocated with the FEM also used to train and validate the model? Or did the authors only use data from the 5 sensors that were collocated with the FEM to train and validate the model? Are data from the 84 sensors that were not collocated with the FEM included in any of the results presented in the manuscript (outside of Figure S4)? If data from those 84 sensors were used to train and validate the model, how were those data treated? Were those 84 sensors assumed to measure the same pollution as the FEM? How far from the FEM was the location in Apex? Please clarify the methods in the manuscript.

4. Page 3, Line 42 through Page 4, Line 1: "The primary sensor-reported data include PM$_1$, PM$_{2.5}$, and PM$_{10}$ concentrations with a factory-specified correction factor for ambient measurements (CF=ATM), concentrations with CF=1 factor recommended by the manufacturer for use in indoor environments..." I don't think it's sufficient or appropriate to simply state the manufacturer-provided specifications here given that: a) there doesn't seem to be any scientifically-valid reason to use the CF=ATM stream instead of the CF=1 stream for ambient monitoring and b) more detailed information on the PMS 5003 outputs is available from other sources. For sensor-reported concentrations below 30 μg m$^{-3}$, the PM$_{2.5}$ CF=ATM stream is equal to the PM$_{2.5}$ CF=1 stream. For sensor-reported concentrations above 30 μg m$^{-3}$, the PM$_{2.5}$ CF=ATM stream has a nonlinear response and reports lower concentrations than the PM$_{2.5}$ CF=1 stream.

See the information provided under the heading “Comparison between Std. Particle and Std. Atmosphere” on the aqicn.org page for the PMS 5003 (https://aqicn.org/sensor/pms5003-7003/).

Also see Section 2.2.1 on page 4619 of Barkjohn et al. (https://doi.org/10.5194/amt-14-4617-2021): “The two data columns have a [cf_atm] / [cf_1] = 1 relationship below roughly 25 μg m$^{-3}$ (as reported by the sensor) and then transition to a two-thirds ratio at higher concentration ([cf_1] concentrations are higher).” Additionally, Barkjohn et al. found that the PM$_{2.5}$ CF=1 stream was more strongly correlated with 24-hour average ambient PM$_{2.5}$ concentrations measured using FRM and FEM instruments than the PM$_{2.5}$ CF=ATM stream.

Did the authors evaluate whether their model performed any better or worse if trained with the PM$_{2.5}$ CF=1 stream instead of the PM$_{2.5}$ CF=ATM stream?

5. Page 10, Line 10: Is there a US EPA AQS site identification number for this location?

6. Page 10, Line 25: "...suggesting an overall underestimation by PA in clean conditions (PM$_{2.5}$ < 10 μg m$^{-3}$)." I don't think this generalization is appropriate. Other U.S.-based studies (Barkjohn et al.: https://doi.org/10.5194/amt-14-4617-2021; Tryner et al.: https://doi.org/10.1016/j.atmosenv.2019.117067) found that PurpleAir sensors overestimated ambient PM$_{2.5}$ concentrations, so while it is true that the mean bias was < 0 for the authors' dataset, similar results are not always observed at low concentrations.

7. Page 10, Lines 31-32: "It is also important to note that the chemical composition of particles in Delhi and Raleigh is expected to be different." This statement is true: There are likely to be differences in the sources of particles in Delhi and Raleigh and, as a result, the chemical composition of the particulate matter is likely to vary between these two locations, but I think differences in particle size distribution are likely to affect PurpleAir accuracy more than differences in optical properties and particle density. See Hagan and Kroll (https://doi.org/10.5194/amt-13-6343-2020) and Ouimette et al. (https://doi.org/10.5194/amt-15-655-2022).
8. Page 10, Lines 32–33: Hagan et al. (https://doi.org/10.1021/acs.estlett.9b00393) also report data on the composition of ambient PM in Delhi in the winter.

9. Page 13, Line 8: “We also evaluated the importance of each input parameter...” Please include the methods used for this analysis in the manuscript.

10. Page 21, Lines 6-7: “The performance of the model for other seasons (and thus for other weather conditions) will need to be evaluated as part of future work.” Differences in temperature and humidity are not the only reason why the model might perform differently in other seasons. I suggest the revising this text to reflect the fact that differences in model performance across seasons is likely to also be influenced by differences in pollution sources. Several prior publications have described how pollution sources and PurpleAir performance vary across seasons in locations around the world:

- Sayahi et al., 2019, Salt Lake City, Utah, USA: https://doi.org/10.1016/j.envpol.2018.11.065
- McFarlane et al., 2021, Accra, Ghana: https://doi.org/10.1021/acsearthspacechem.1c00217
- Raheja et al., 2022, Lomé, Togo: https://doi.org/10.1021/acsearthspacechem.1c00391

Technical Corrections

11. Page 2, Lines 22-23: Barkjohn et al., 2021 (https://doi.org/10.5194/amt-14-4617-2021) seems to be missing from this list of citations.

12. Figures 2, 3, 4, 5, 6, 7, 9, 10, and 11: The two sites are referred to as “Delhi” and “Raleigh” throughout the text, but as “India” and “NC” in the figures. Please update the plot labels to also refer to the two sites as Delhi and Raleigh.

13. Figures 2, 3, 4, 5, 6, 8, 9, 10, and 11: All x- and y-axes should have descriptive labels with units (not just variable names from the authors’ code).

- I assume the units are μg m⁻³ on all x- and y-axes in Figures 2, 3, 4, 5, and 6?
- The “PA (ML CALIBRATED – Daily)” labels on the y axes in Figure 6 make it look like
one value has been subtracted from another, but I don’t think that’s what the authors are showing in those plots. How about “24-h average ML-calibrated PA PM$_{2.5}$ ($\mu$g m$^{-3}$)”?
- I assume the units are $\mu$g m$^{-3}$ on the y-axes in Figures 8 and 9?
- I assume that all PM$_{2.5}$ concentrations shown on in Figures 10 and 11 are in $\mu$g m$^{-3}$?

14. Figures 4, 5, and 6: “The color scale represents the density of data points.” I do not understand what this means.

15. Figure S1: What do the red lines represent? +/- 1 standard deviation? Please explain in the figure caption.

16. Figure S3.1: This image is not legible. Please provide a higher-resolution image.

17. Figure S4: Please use more informative axis labels. Why does the caption say “ADD THIS FIGURE”? 