We thank the reviewers for their time reviewing our manuscript and their feedback. We have taken the time to make a number of significant improvements to the manuscript based on the reviewers' comments. We have included responses in blue below where the line numbers refer to the manuscript with track changes enabled.

Referee #1

The paper Development and Application of a United States wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor by Barkjohn et al. is an evaluation of the PA units using reference sensors for $PM_{2.5}$ values. Overall, the paper is well written and clear, I have some concerns and comments for things that hopefully you will be able to address and clarify

General comments

1. My biggest concern is about the fact you used the T and RH from the PA itself to perform your analysis. Even the company itself does not recommend using these values for ambient conditions for the same reason you mention in your paper. Also, you cite Holder et al. 2020 but even they state that "the PA temperature and RH measurement are interpreted as an internal rather than ambient measurement". You also enhance this comment in line 375.

We acknowledge that the temperature and RH sensors are not ambient but are internal measurements that run typically hotter and dryer than ambient measurements. Past work has shown that these measurements are strongly correlated with reference measurements making them good candidates for inclusion in the correction. Many PurpleAir sensors are located far from meteorological sites, so if we used reference meteorological data this would make our correction difficult to apply at rural sites where sensors have the most potential to increase spatial resolution on the map. In addition, these measurements may actually be more relevant in understanding hygroscopic growth since particles may be warmed and dried inside the sensor before being measured. Our work is showing that we don't need an ambient measurement of RH to improve the PurpleAir measurements but that the RH reported by the sensor is able to improve the measurements. This question was already addressed in the text:

Lines 392-403: It is important to note that the meteorological sensor in the PurpleAir sensor is positioned above the particle sensors nestled under the PVC cap resulting in temperatures that are higher (2.7 to 5.3°C) and RH that is drier (9.7% to 24.3%) than ambient conditions (Holder et al., 2020;Malings et al., 2020). In addition, these internal measurements have been shown to be strongly correlated with reference temperature and RH measurements with high precision (Holder et al., 2020;Tryner et al., 2020a;Magi et al., 2019). The well characterized biases and strong correlations between PurpleAir and ambient meteorological parameters mean that the coefficients using these terms in a correction equation account for the differences between the ambient and PurpleAir measured meteorology. Although not as accurate as the reference measurements, the PurpleAir temperature and RH measurements are good candidates for inclusion in a linear model because they are well correlated with reference measurements and may more closely represent the particle drying that occurs inside the sensor. In addition, using onboard measurements and information that would be available for all PurpleAir sensors, allows us to gather corrected air quality data from all PurpleAirs, even those in remote areas far from other air monitoring or meteorological sites.

2. Information on the location of co-located units (lat/long) will help the community to understand which purple air sensor you used (pubic), also if you could provide their name that will be important, at least for the public one.

Latitude and Longitude information is available from AQS associated with the AQS ID. We have included them in the SI as well for convenience as requested by the reviewer. The public IDs have also been included for the PurpleAirs (See SI Tables S1, S2)

3. Why 50 meters for distance for collocated units, 1 km is not good for that, other works used a distance larger than 50 m for collocated units. I am wondering how many more collocated units you could have gotten if you had a larger distance between units.

Our goal in this work was to leverage the partnerships air monitoring agencies have with the EPA in order to identify true collocations where a PurpleAir was actually running at the air monitoring site as opposed to close by. As stated in the text, 50m was selected to account for the typical footprint of an air monitoring station. This helps us to constrain that most of the error is due to inaccuracies in the sensor as opposed to localized sources or inaccurate location information.

This was addressed in the text on **lines 95-96:** The 50-meter distance was selected because it is large enough to cover the footprint of most AQS sites and small enough to exclude most PurpleAir sensors in close proximity, but not collocated with, an AQS site.

And we have added additional text **lines 103-110:** Much past work using public data from PurpleAir has used public sensors that appear close to a regulatory station on the map (Ardon-Dryer et al., 2020;Bi et al., 2020). However, there is uncertainty in the reported location of PurpleAir sensors as this is specified by the sensor owner. In some cases, sensors may have the wrong location. Known examples include owners who forgot to update the location when they moved, take the sensor inside for periods to check their indoor air quality, or specifically choose an incorrect location to protect their privacy. In addition, without information on local sources of $PM_{2.5}$, it can be unclear how far away is acceptable for a "collocation" since areas with more localized sources will need to be closer to the reference monitor to experience similar $PM_{2.5}$ conditions. By limiting this work to true collocations operated by air monitoring agencies, we eliminate one source of uncertainty. We can conclude that the PurpleAir errors measured in this work are not due to poor siting or localized sources and can focus on other variables that influence error (e.g. RH).

4. In table 1 you used 50 PA units but in Fig 1 you show more than 50, it is confusing. Maybe have in table 1 the full number of units

We have clarified this in the in the caption of Table 1 and have removed the first reference to table 1.

Table 1. Summary of the dataset used to generate the U.S-wide PurpleAir correction equation after 3 sensors with large A, B channel discrepancies were removed. $PM_{2.5}$ concentrations from both the FEM or FRM and the PurpleAir (PA), temperature (T) and relative humidity (RH) are summarized as median (min, max).

Line 116-117: In total, 53 PurpleAir sensors at 39 unique sites across 16 states were ideal candidates and were initially included in this analysis

5. Why did you remove data from a station like Iowa, why not using as much data as you have as you could have a full range of T ad RH conditions?

As shown in Figure S1 after subsetting the Iowa data, the full range of temperature and relative humidity conditions are still represented. Using the entire Iowa dataset could lead to a correction that is weighted too heavily towards the aerosol and meteorological conditions experienced in Iowa. This is addressed in the text:

Lines 122-125: Initially, there were 10,907 pairs of 24-hr averaged collocated data from Iowa which was 55% of the entire collocated dataset. In order to better balance the dataset among the states, and to avoid building a correction model that is weighted too heavily towards the aerosol and meteorological conditions experienced in Iowa, the number of days from Iowa was reduced to equal the size of the California dataset, the state with the next largest amount of data (29% of the entire collocated dataset).

6. PA company record on using the cf_atm so why also testing the other type?

Although the cf_atm may more closely reflect $PM_{2.5}$ concentrations without correction, it was important for this work to explore which correction was more strongly correlated and more linear with reference $PM_{2.5}$ concentrations. It is not clear that PurpleAir comprehensively evaluated which correction would be best suited for the large range of ambient conditions that exists across the United States. Since PurpleAir provides both it seems like there is no real reason to exclude one from consideration. This is addressed in the text:

Lines 368-370: For PM, we considered $PM_{2.5}$ concentrations from both the [cf_1] and [cf_atm] data columns as model terms. Previous work has found different columns to be more strongly correlated under different conditions (Barkjohn et al., 2020a;Tryner et al., 2020a).

7. In Section 3.5 you mention that the use of T for correction should be used on a local basis, but don't you think this is an important factor. In your work, you covered locations with wide T conditions from cold Alaska to warm Florida is it possible that only changes in RH can represent the way to correct the PA data across the US?

We did not write that a local T correction "should" be used just that it has been used in past work and that it may address local factors. It is well known that when relative humidity increases, particles undergo hygroscopic growth (dependent on composition and other particle properties), causing them to scatter more light. Since FEM and FRM methods based on a consistent, relatively low RH range (30-40%), this leads to errors in estimating "dry" $PM_{2.5}$. With temperature there is no similar mechanism that is widely understood to impact the performance of optical measurements. It is unclear what is being accounted for when a temperature term is included in a correction for a single location. It could be accounting for different sources during different parts of the year or something else. Although we have a large dataset, we do not have a large enough dataset to build corrections for each region of the U.S. This is addressed in the text:

Lines 678-690: This work indicates that only an RH correction is needed to reduce the error and bias in the nationwide dataset. Some previous single site studies found temperature to significantly improve their $PM_{2.5}$ prediction as well (Magi et al., 2019;Si et al., 2020). Humidity has known impacts on the light scattering of particles; no similar principle exists for explaining the influence of temperature on particle light scattering. Instead, the temperature factor may help account for some local seasonal or diurnal patterns in aerosol properties within smaller geographical areas. These more local variations may be why temperature does not substantially reduce error and bias in the nationwide dataset. More work should be done to better understand this influence. These previous models also did not include a term accounting for the collinearity between temperature and relative humidity that may have been present. **Error! Reference source not found.** shows the residual error in each 24-hour corrected PurpleAir $PM_{2.5}$ measurement compared with the temperature, RH, and FRM or FEM $PM_{2.5}$ concentrations. Error has been reduced compared to the raw dataset (Figures S8, and S9) and is unrelated to temperature, RH, and $PM_{2.5}$ variables. Some bias at very low temperature < -12°C and potentially high concentration (> 60 µg m^{-3}) may remain, but more data are needed to further understand this relationship.

Also, from Table S2, it seems most of your RH were < 100, only 3 PA units reach RH of 100 is it possible the lack of high RH harm your analysis.

Lines 341-343: There was limited data above 80% RH as measured by the PurpleAir RH sensor likely due to the warmer and dryer conditions inside the PurpleAir as mentioned previously.

8. Would be nice to see the distribution of RH and T from all dataset (not just for Iowa)

Thank you for this suggestion. We have added it to the SI (Figure S5).

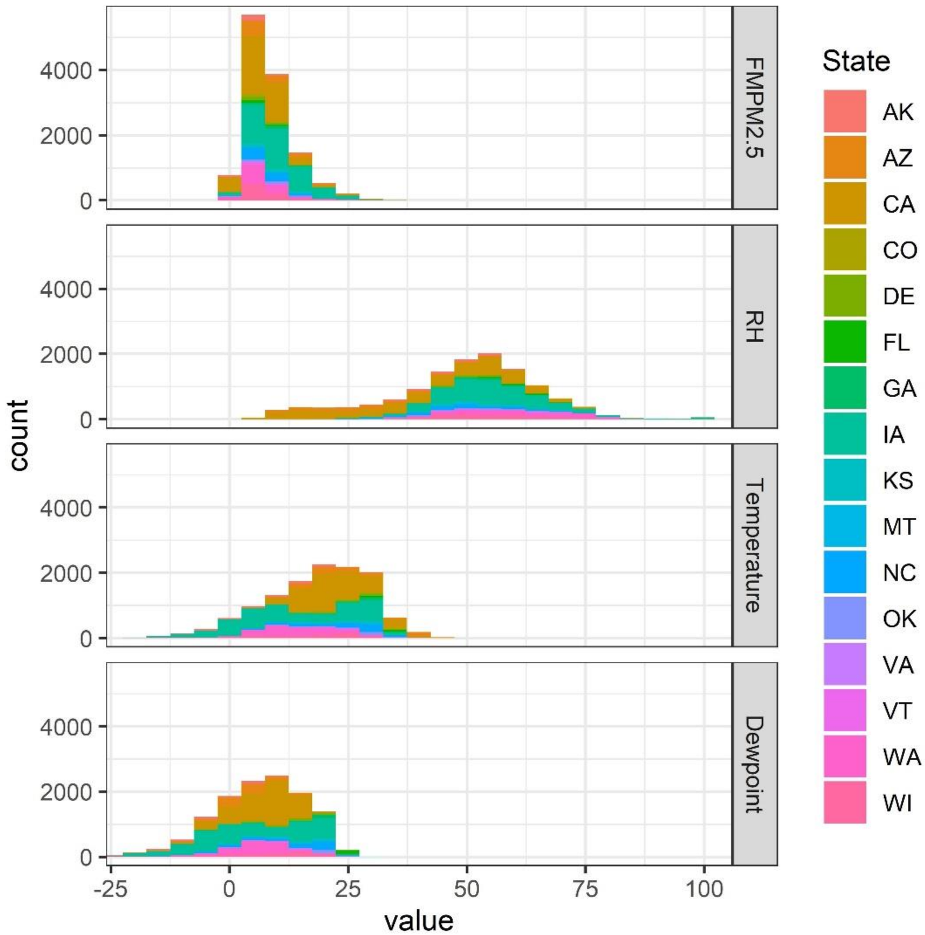

Figure S5. Distribution of full dataset after subsetting Iowa

Comments of figures and tables

9. Many of your figures are pixelated

My apologies. I believe this is from the way I export the pdf. I will make sure to export at higher quality for the next draft.

10. Table 1 - I assume PA data on table 1 is uncorrected. What do you mean by the wide PurpleAir correction equation?

We have hyphenated U.S.-wide consistently throughout the text to clarify this.

11. Fig 1 - The quality of the fig itself is not great it is very pixelated, also R^2 values on each plot would have helped to see the comparison between A and B. you should more h=than 50

I have added pearson correlation (since correlation is discussed in the text instead of R^2) to each plot and exported at higher resolution as requested. See the updated figure below:

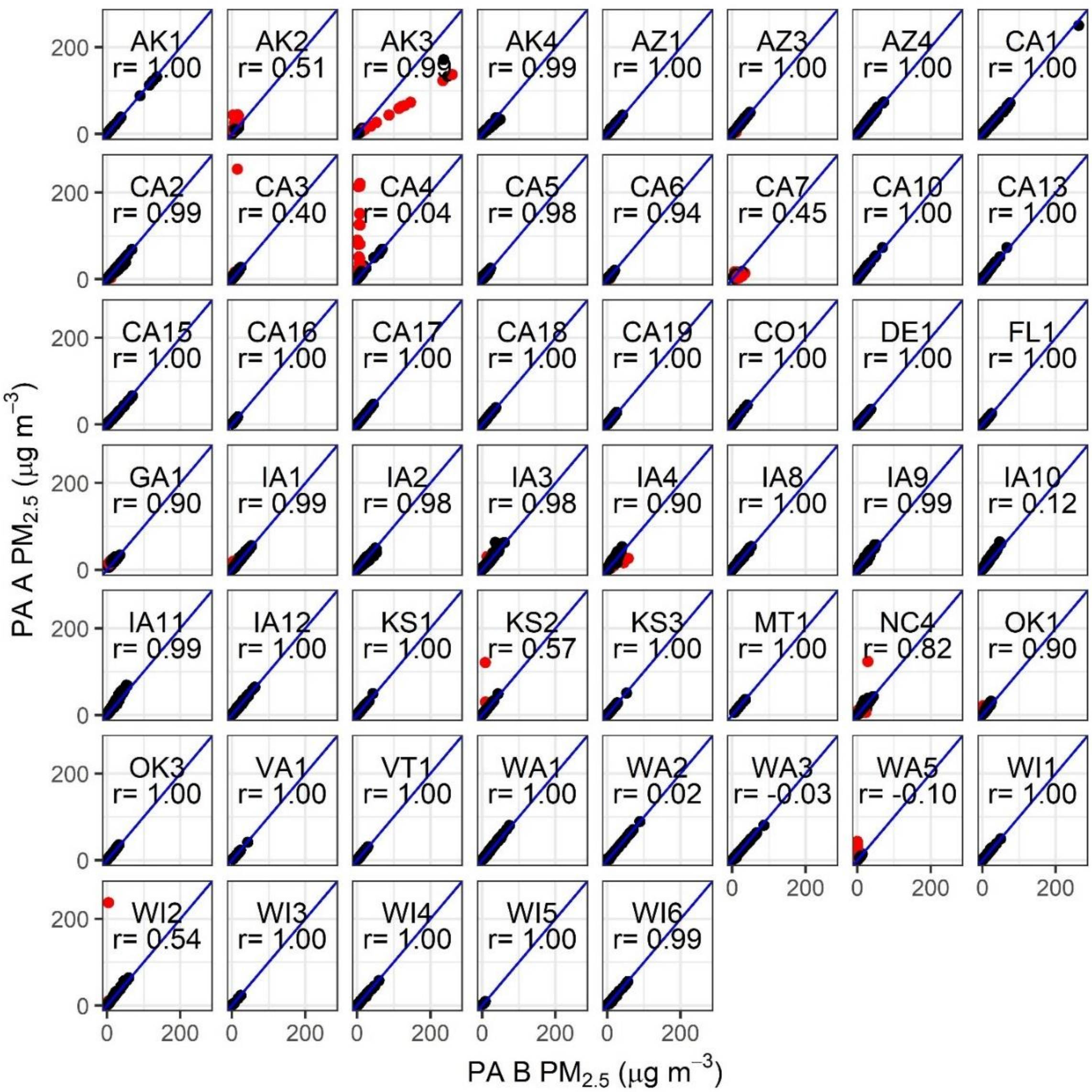

Figure 1. Comparison of 24-hour averaged $PM_{2.5}$ data from the PurpleAir A and B channels. Excluded data (2.1%) are shown in red and represent data points where channels differed by more than 5 μg m^{-3} and 61%. AK3, CA7, WA5 were excluded from further analysis. Pearson correlation (r) is shown on each plot.

12. Fig 7 - Does the data represent the entire data set or one location?

This is the full dataset. The caption has been updated as follows:

Figure 7. 24-hr AQI categories as measured by the corrected PurpleAir and the FEM or FRM for the full dataset generated with the models built using LOSO withholding.

Specific comments

13. Line 56-60: it seems you have too many references can you reduce the amount or at least separate them into multiple sentences

I sorted them into a few sentences for clarity.

Lines 56-67: Previous work has explored the performance and accuracy of PurpleAir sensors under outdoor ambient conditions in a variety of locations across the United States including in Colorado (Ardon-Dryer et al., 2020;Tryner et al., 2020a), Utah (Ardon-Dryer et al., 2020;Kelly et al., 2017;Sayahi et al., 2019), Pennsylvania (Malings et al., 2020), North Carolina (Magi et al., 2019), and in California where the most work has occurred to date (Ardon-Dryer et al., 2020;Bi et al., 2020;Feenstra et al., 2019;Mehadi et al., 2020;Schulte et al., 2020;Lu et al., 2021). Their performance has been explored in a number of other parts of the world as well including in Korea (Kim et al., 2019), Greece (Stavroulas et al., 2020), and Australia (Robinson, 2020). Additional work has been done to evaluate their performance under wildland fire smoke impacted conditions (Bi et al., 2020;Delp and Singer, 2020;Holder et al., 2020), indoors (Wang et al., 2020b), and during laboratory evaluations (Kelly et al., 2017;Kim et al., 2019;Li et al., 2020;Mehadi et al., 2020;Tryner et al., 2020a;Zou et al., 2020a;Zou et al., 2020b). The performance of their dual Plantower PMS5003 laser scattering particle sensors has also been explored in a variety of other commercial and custom built sensor packages (He et al., 2020;Tryner et al., 2019;Kuula et al., 2019;Ford et al., 2019;Si et al., 2020;Zou et al., 2020b;Tryner et al., 2020b).

Line 85: you state that the data was until 2018 but some of your data was for 2019

Since this analysis occurred in 2018 we identified sensors that were online in 2017 and 2018. This section has been rewritten to improve clarity

Lines 85-95: Data for this project came from 3 sources: 1) PurpleAir sensors sent out by EPA for collocation to capture a wide range of regions and meteorological conditions, 2) privately operated sensor data volunteered by state, local and tribal (SLT) air monitoring agencies independently operating collocated PurpleAir sensors, and 3) publicly available sensors located near monitoring stations and confirmed as true collocation by air monitoring agency staff. In order to identify publicly available collocated sensors, in August of 2018, a survey of sites with potentially collocated PurpleAir sensors and regulatory $PM_{2.5}$ monitors was performed by identifying publicly available PurpleAir sensor locations within 50 meters of an active EPA Air Quality System (AQS) site reporting $PM_{2.5}$ data in 2017 or 2018.

14. Lines 110-114: the entire section is unclear, I do not understand what you mean, can you rewrite this part to make it clearer

We have clarified this section.

Lines 260-265: The 2-minute (or 80-second) $PM_{2.5}$ data were averaged to 24-hours (representing midnight to midnight local standard time). A 90% data completeness threshold was used based on channel A, since both channels were almost always available together (i.e. 80-second averages required at least 0.9*1080 points before 5/30/2019 or 2-minute averages required at least 0.9*720 points after 5/30/2019). This methodology ensured that the averages used were truly representative of daily averages reported by regulatory monitors. A higher threshold of completeness was used for the PurpleAir data as it likely has more error than FEM or FRM measurements.

15. Line 235: what do you mean - one ran multiple sensors in series

We have clarified the meaning of series in this context:

Lines 333-335: Some sites had several PurpleAir sensors running simultaneously (N=9) and one ran multiple sensors in series (i.e. one sensor failed, was removed, and another sensor was put up in its place).