**Supporting Information**

# Incorporating Low-Cost Sensor Measurements into High-Resolution $PM_{2.5}$ Modeling at A Large Spatial Scale

Jianzhao Bi,[1] Avani Wildani,[2] Howard H. Chang,[3] Yang Liu[*,1]

[1]Department of Environmental Health, Emory University, Rollins School of Public Health, Atlanta, Georgia 30322, United States

[2]Department of Computer Science, Emory University, Atlanta, Georgia 30307, United States

[3]Department of Biostatistics and Bioinformatics, Emory University, Rollins School of Public Health, Atlanta, Georgia 30322, United States

Corresponding Author

*Mailing Address: Emory University, Rollins School of Public Health, 1518 Clifton Road NE, Atlanta, GA 30322, USA. E-mail: yang.liu@emory.edu.

1. Quality Control for PurpleAir $PM_{2.5}$ Measurements

Each PurpleAir sensor consists of two identical laser particle counters providing two sets of PM readings (Channel A & Channel B). PurpleAir $PM_{2.5}$ data were cleaned based on the dual-channel readings. We first discarded all hourly records with only one channel's reading since the outliers were hard to identify based on a single channel. Additionally, there were apparent outliers with $PM_{2.5}$ levels greater than 3,000 μg/m$^3$ in both channels which were also discarded. The discarded records accounted for ~3% of the total records. The remaining records (N = 5,658,772) still had a large dual-channel discrepancy with an $R^2$ of 0.32 and a slope of 0.60. We then used the absolute percentage bias (APB) computed from dual-channel readings (Eq. S1) to further filter out the outliers. In Eq. S1, $\mathrm{PM}_{2.5\mathrm{A}}$ denotes Channel A's reading and $\mathrm{PM}_{2.5\mathrm{B}}$ denotes Channel B's reading. A percentage threshold of APB was determined according to the improvement of overall dual-channel agreement. When setting the threshold to be 5%, *i.e.*, removing the records with top-5% largest APB values, all apparent outliers disappeared (Fig. S6). The remaining data (N = 5,375,833) had an excellent dual-channel agreement with an $R^2$ of 0.98 and a slope of 0.997. The final PurpleAir $PM_{2.5}$ measurements were the average of the dual-channel readings.

Eq. S1

$$\mathrm{APB} = \left|\frac{\mathrm{PM}_{2.5\mathrm{B}} - \mathrm{PM}_{2.5\mathrm{A}}}{\mathrm{PM}_{2.5\mathrm{A}}}\right| \times 100\%$$

2. Evaluation of PurpleAir $PM_{2.5}$ Measurements

The evaluation of PurpleAir was performed based on the paired hourly $PM_{2.5}$ measurements (N = 137,068). The hourly AQS measurements averaged 11.1 μg/m³ with an interquartile range (IQR) of 9.8 μg/m³ ($25^{th}$, $75^{th}$ percentiles: [5.0 μg/m³, 14.8 μg/m³], maximum: 369.0 μg/m³), while the corresponding PurpleAir measurements averaged 13.0 μg/m³ with an IQR of 14.8 μg/m³ ($25^{th}$, $75^{th}$ percentiles: [2.8 μg/m³, 17.6 μg/m³], maximum: 448.5 μg/m³). Compared to AQS, PurpleAir measured a higher overall $PM_{2.5}$ level by 1.9 μg/m³ and significantly higher peak values. Previous low-cost sensor evaluation studies based on the same sensor (Plantower PMS, Beijing Plantower Co., Ltd) also found that the sensor tended to overestimate $PM_{2.5}$ compared to reference-grade monitors [1, 2]. For example, Kelly, et al. [1] reported that PMS overestimated $PM_{2.5}$ concentrations when exceeding 10 μg/m³ during several cold-air pools (CAPs) in winter. Badura, et al. [2] reported that the raw outputs from PMS overestimated collocated tapered element oscillating microbalance (TEOM) data by a factor of 3.5 during a half-year field campaign.

## 3. Nonlinearity of PurpleAir Systematic Bias

The nonlinearity of PurpleAir systematic bias was examined by locally weighted scatterplot smoothing (LOWESS). LOWESS fits a low-degree polynomial at each point of the data set where the data near the point are given higher weights [3]. LOWESS is a non-parametric strategy for finding a curve of best fit without assuming the distribution of data. An important hyperparameter of LOWESS is the smoothing span controlling the degree of smoothing. This hyperparameter was tuned with 10-fold cross-validation (CV). A smoothing span of 10% was the

optimal value in this analysis. The LOWESS showed an almost linear curve coinciding with the curve of linear regression (Fig. S7), indicating that linear calibration was satisfactory for PurpleAir data.

4. Validation of Scale Factor ρ

The scale factor ρ was mainly used as a proxy of implicit factors which may impact the prediction quality and further reduce the relative importance of PurpleAir in the model. This factor was set to be a multiplicative term within a range (0, 1). Intuitively, for a set of perfect measurements, *i.e.*, the data quality is identical to reference-grade data, this data-driven scale factor should be close to 1. In order to validate this assumption, we pretended the PurpleAir measurements had a perfect quality and used them as ground truth with AQS measurements in the prediction model. The trend of CV RMSPE with different ρ values is shown in Fig. S8(a). We can see that CV RMSPE reaches its minimum when ρ is closer to 1. This result indicates the reasonability of our assumption, *i.e.*, ρ is a physically meaningful parameter with the value closer to 1 for a perfect data set such as reference-grade data and closer to 0 for a data set with large uncertainty such as low-cost sensor measurements. The optimal ρ value of the weighted prediction model was also tuned based on the 10-fold CV (Fig. S8(b)). The CV RMSPE shows a U-curve with a minimum at a ρ value of ~0.23. The range of CV RMSPE is as large as 0.2 μg/m³, indicating the large influence of this scale factor on the model performance.

**References**

1. Kelly, K. E.; Whitaker, J.; Petty, A.; Widmer, C.; Dybwad, A.; Sleeth, D.; Martin, R.; Butterfield, A., Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environ Pollut* **2017,** *221*, 491-500.
2. Badura, M.; Batog, P.; Drzeniecka-Osiadacz, A.; Modzel, P., Evaluation of Low-Cost Sensors for Ambient PM2. 5 Monitoring. *Journal of Sensors* **2018,** *2018*.
3. Cleveland, W. S.; Devlin, S. J., Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association* **1988,** *83*, (403), 596-610.

Table S1 Summary statistics of the absolute differences between fully calibrated PurpleAir data and the calibrated data based on subsets of collocated AQS/PurpleAir sites. The total number of collocated AQS/PurpleAir sites is 26 and the subsets are randomly selected from these 26 sites with different proportions from 90% to 10%. This analysis is based on a subset of 10,000 randomly selected PurpleAir measurements.

| **Proportion*** | **Mean (μg/m³)** | **Q1**** **(μg/m³)** | **Median (μg/m³)** | **Q3**** **(μg/m³)** | **Max (μg/m³)** |
|---|---|---|---|---|---|
| 90% | 0.02 | 0.00 | 0.01 | 0.02 | 0.77 |
| 80% | 0.35 | 0.02 | 0.08 | 0.44 | 11.51 |
| 70% | 0.47 | 0.07 | 0.20 | 0.52 | 12.58 |
| 60% | 0.84 | 0.32 | 0.65 | 1.10 | 13.39 |
| 50% | 1.06 | 0.37 | 0.76 | 1.34 | 20.26 |
| 40% | 1.75 | 0.72 | 1.43 | 2.29 | 20.70 |
| 30% | 1.87 | 0.83 | 1.62 | 2.50 | 20.40 |
| 20% | 2.26 | 1.04 | 2.03 | 3.11 | 19.59 |
| 10% | 2.35 | 1.01 | 2.02 | 3.22 | 23.49 |
| Raw*** | 4.59 | 1.13 | 2.51 | 4.89 | 158.62 |

* The (gross) proportion of collocated AQS sites being kept.

** The 25th and 75th percentiles

*** Uncalibrated PurpleAir data

Table S2 The ten-most important variables of the AQS-based model based on which the HAC was performed.

| | **$PM_{2.5}$-Related Variables Used in HAC** |
|---|---|
| 1 | $PM_{2.5}/PM_{10}$ ratio |
| 2 | Elevation |
| 3 | Visibility |
| 4 | Gap-filled Aqua AOD |
| 5 | 10-meter meridional wind speed |
| 6 | Gap-filled Terra AOD |
| 7 | Percentage of shrublands |
| 8 | 2-meter specific humidity |
| 9 | Population |
| 10 | Nearest distance to roads |

Table S3 Numbers and densities of continuous AQS stations (capable of providing hourly $PM_{2.5}$ measurements) in 47 states of the Contiguous United States (CONUS) (without California). The rows in green are the states with densities of continuous AQS stations greater than 5 per 100,000 $km^2$.

| Rank | CONUS State | N of AQS | State Area ($km^2$) | N of AQS per 100,000 $km^2$ |
|---|---|---|---|---|
| **1** | Rhode Island | 5 | 4,001 | 124.97 |
| **2** | Delaware | 5 | 6,446 | 77.57 |
| **3** | Massachusetts | 16 | 27,336 | 58.53 |
| **4** | Connecticut | 8 | 14,357 | 55.72 |
| **5** | New Jersey | 12 | 22,591 | 53.12 |
| **6** | Maryland | 11 | 32,131 | 34.24 |
| **7** | Washington | 61 | 184,661 | 33.03 |
| **8** | Pennsylvania | 39 | 119,280 | 32.70 |
| **9** | New Hampshire | 6 | 24,214 | 24.78 |
| **10** | Florida | 39 | 170,312 | 22.90 |
| **11** | Ohio | 26 | 116,098 | 22.40 |
| **12** | New York | 30 | 141,297 | 21.23 |
| **13** | Indiana | 19 | 94,326 | 20.14 |
| **14** | Vermont | 4 | 24,906 | 16.06 |
| **15** | Tennessee | 16 | 109,153 | 14.66 |
| **16** | North Carolina | 20 | 139,391 | 14.35 |

| | | | | |
|---|---|---|---|---|
| **17** | Illinois | 21 | 149,995 | 14.00 |
| **18** | Kentucky | 14 | 104,656 | 13.38 |
| **19** | South Carolina | 11 | 82,933 | 13.26 |
| **20** | Wisconsin | 18 | 169,635 | 10.61 |
| **21** | Alabama | 14 | 135,767 | 10.31 |
| **22** | Idaho | 22 | 216,443 | 10.16 |
| **23** | Oklahoma | 18 | 181,037 | 9.94 |
| **24** | Minnesota | 22 | 225,163 | 9.77 |
| **25** | Georgia | 15 | 153,910 | 9.75 |
| **26** | Maine | 8 | 91,633 | 8.73 |
| **27** | Virginia | 9 | 110,787 | 8.12 |
| **28** | Louisiana | 10 | 135,659 | 7.37 |
| **29** | Missouri | 13 | 180,540 | 7.20 |
| **30** | Arizona | 21 | 295,234 | 7.11 |
| **31** | Iowa | 10 | 145,746 | 6.86 |
| **32** | Utah | 15 | 219,882 | 6.82 |
| **33** | Michigan | 17 | 250,487 | 6.79 |
| **34** | Mississippi | 8 | 125,438 | 6.38 |
| **35** | Texas | 44 | 695,662 | 6.33 |
| **36** | Colorado | 17 | 269,601 | 6.31 |
| **37** | North Dakota | 11 | 183,108 | 6.01 |

| | | | | |
|---|---|---|---|---|
| **38** | Montana | 19 | 380,831 | 4.99 |
| **39** | Wyoming | 12 | 253,335 | 4.74 |
| **40** | Nevada | 13 | 286,380 | 4.54 |
| **41** | South Dakota | 8 | 199,729 | 4.01 |
| **42** | Arkansas | 5 | 137,732 | 3.63 |
| **43** | West Virginia | 2 | 62,756 | 3.19 |
| **44** | New Mexico | 9 | 314,917 | 2.86 |
| **45** | Oregon | 6 | 254,799 | 2.35 |
| **46** | Kansas | 5 | 213,100 | 2.35 |
| **47** | Nebraska | 2 | 200,330 | 1.00 |

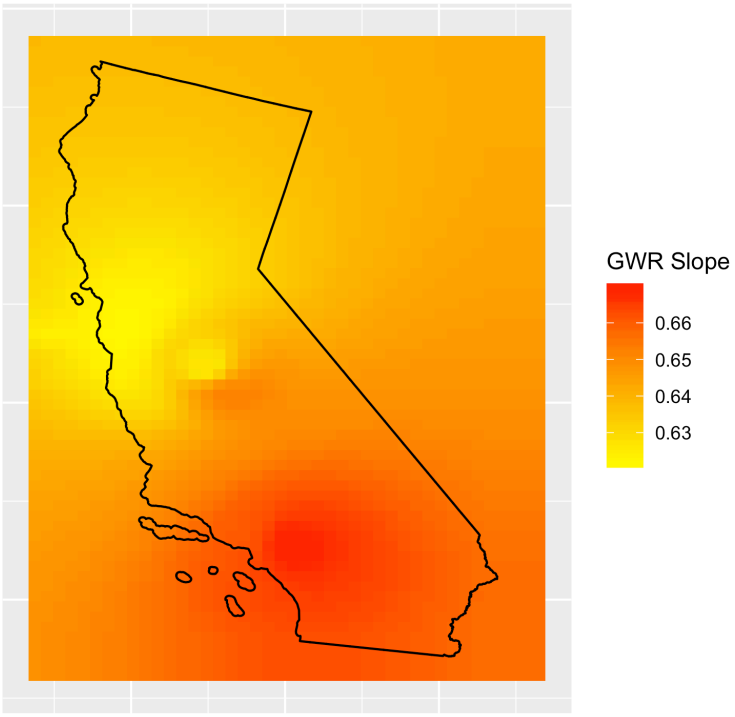



Fig. S1 The spatial distribution of GWR slopes.

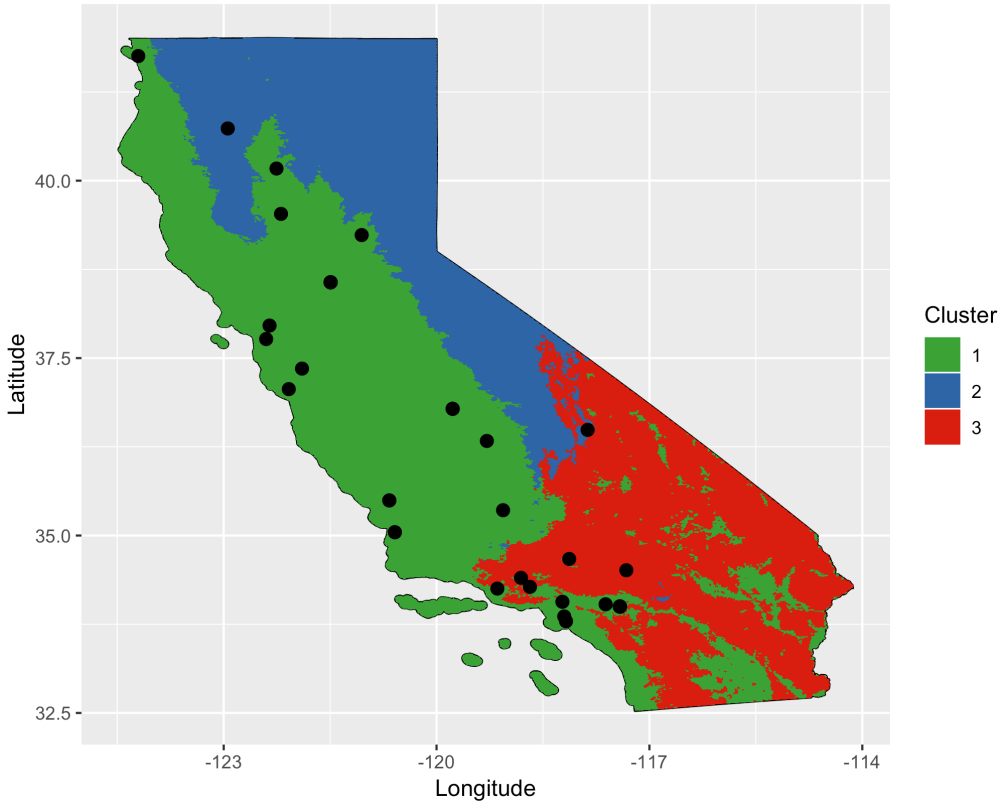
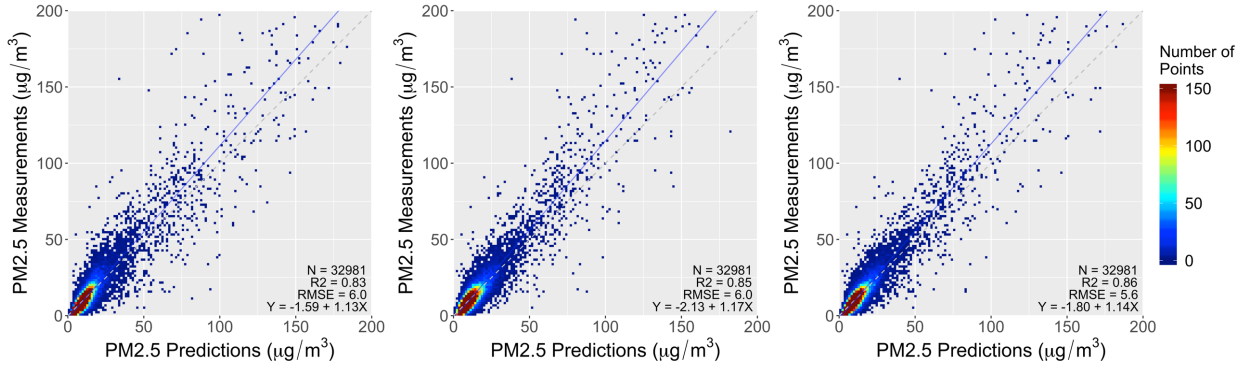
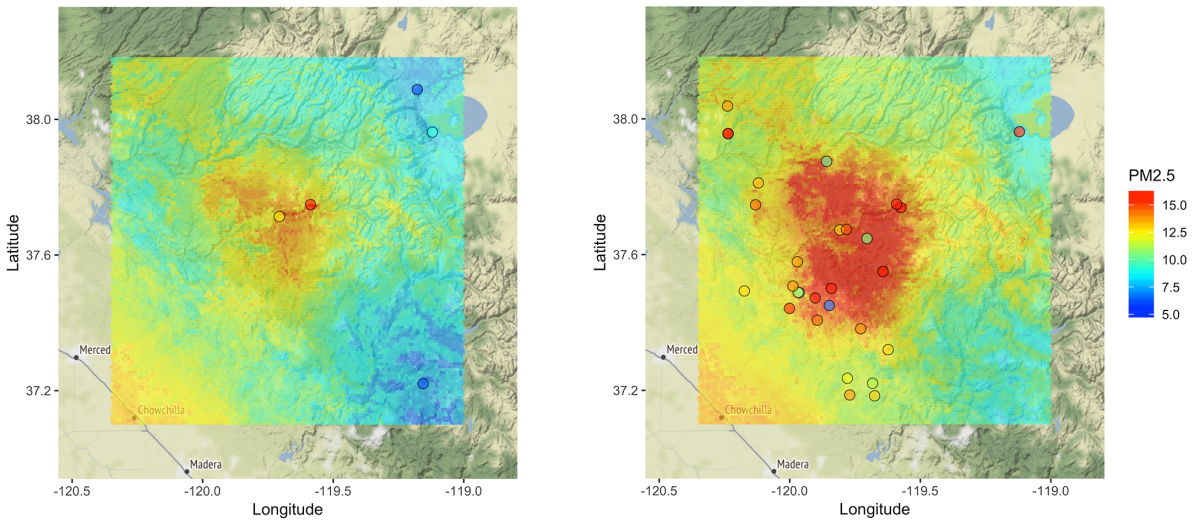



Fig. S2 Three clustered sub-domains with the locations of AQS/PurpleAir pairs (black points): 1 - agricultural/developed areas; 2 - mountainous areas; 3 - arid areas.




(a) CV scatters of the AQS-based model (b) CV scatters of the non-weighted model (c) CV scatters of the weighted model

Fig. S3 10-fold CV scatter plots of (a) the AQS-based model, (b) the non-weighted model, and (c) the weighted model.




(a) AQS-based predictions (b) Weighted predictions

Fig. S4 Locations (with annual mean $PM_{2.5}$ levels) of (a) AQS and (b) PurpleAir sites and the annual mean $PM_{2.5}$ distributions derived by the (a) AQS-based and (b) weighted models in the region of Ferguson Fire in 2018.

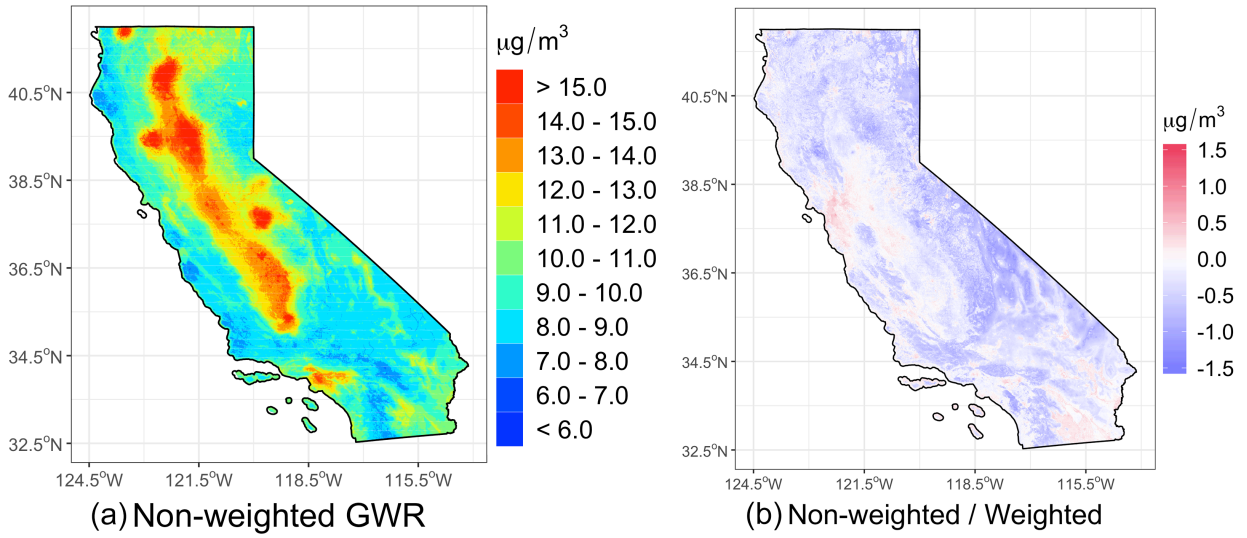
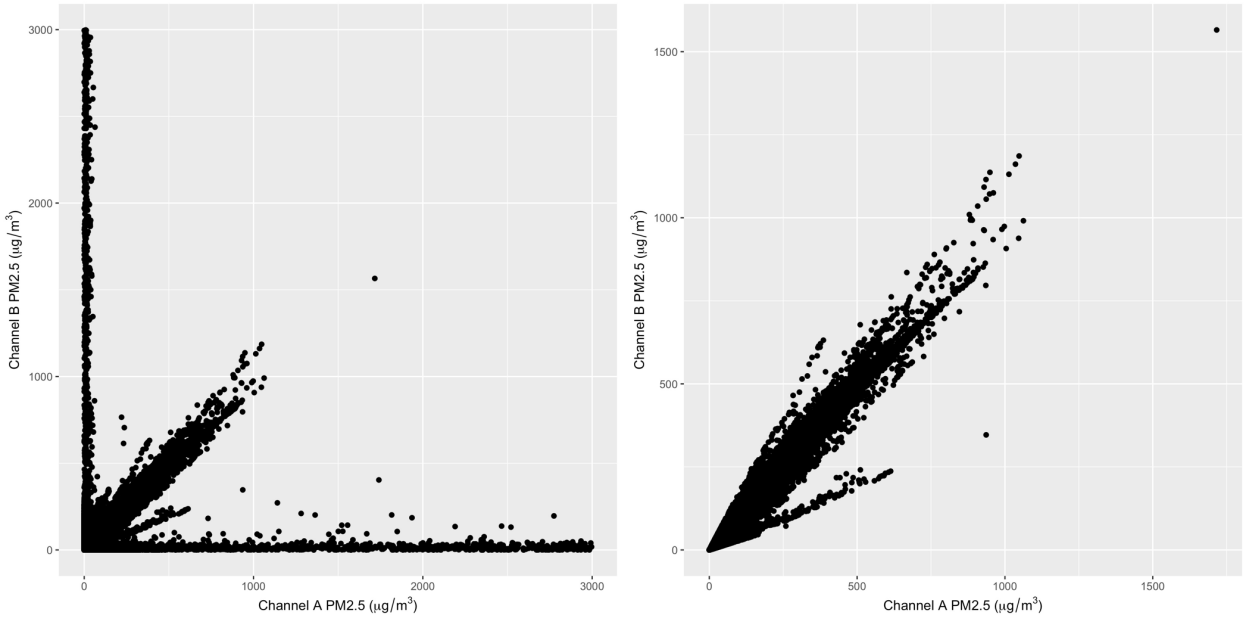



Fig. S5 (a): Annual mean $PM_{2.5}$ distribution for 2018 from the non-weighted model. (b): Annual mean $PM_{2.5}$ differences between the non-weighted and weighted predictions (non-weighted minus weighted).




(a) Before quality control (b) After quality control

Fig. S6 Scatter plots of PurpleAir dual-channel hourly measurements (a) before and (b) after removing the 5% largest absolute percentage biases (APBs).

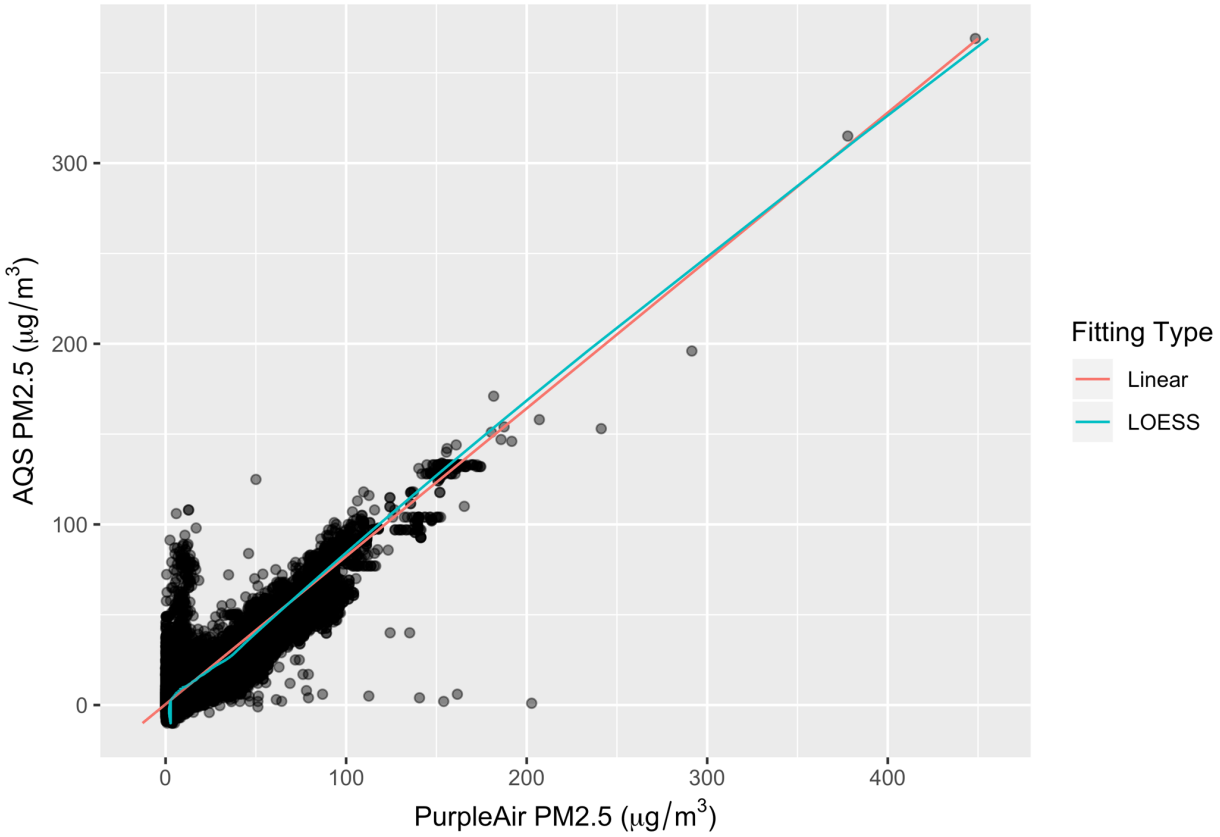



Fig. S7 LOWESS (green) and linear (red) fitting curves of the paired AQS/PurpleAir hourly measurements (black scatters).

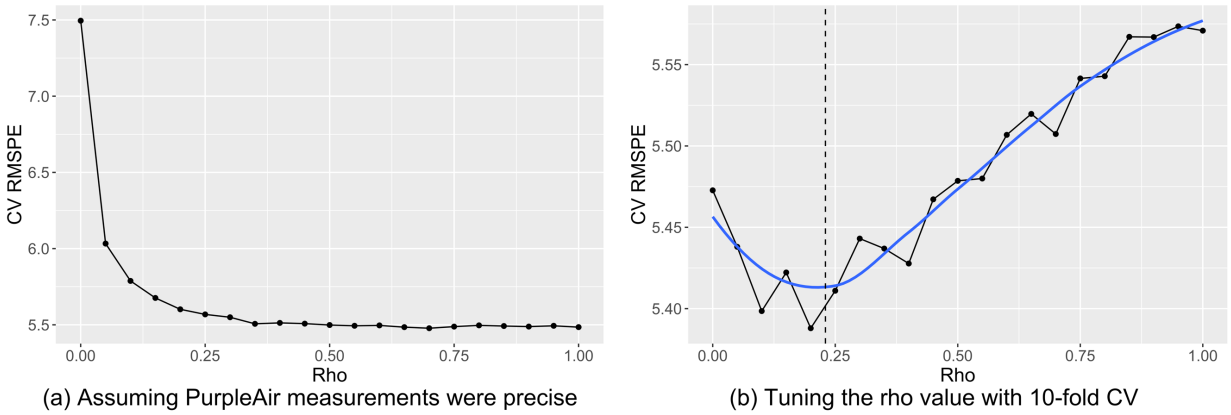



Fig. S8 (a): The trend of the 10-fold CV RMSPE with ρ within a range [0, 1] when assuming PurpleAir measurements were precise. (b): The trend of the 10-fold CV RMSPE with ρ within the range [0, 1] in the real case. The blue curve is the smoothed fitting curve, showing the minimum of CV RMSPE at a ρ value of ~0.23.