PNAS
www.pnas.org

Supplementary Information for
Wildfire smoke impacts on indoor air quality assessed using crowdsourced data in California

Yutong Liang[a]*, Deep Sengupta[a], Mark J. Campmier[b], David M. Lunderberg[c], Joshua S. Apte[b,d], , Allen H. Goldstein[a,b]*

[a]Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA 94720, USA.
[b]Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, CA 94720, USA.
[c]Department of Chemistry, University of California at Berkeley, Berkeley, CA 94720, USA.
[d]School of Public Health, University of California at Berkeley, Berkeley, CA 94720, USA.

*Yutong Liang or Allen H. Goldstein

Email: yutong.liang@berkeley.edu or ahg@berkeley.edu.

This PDF file includes:

Extended Materials and Methods

Data Sources and Study Regions. The PurpleAir sensors report the mass of size-resolved particulate matter, as well as environmental parameters such as temperature and relative humidity (RH). Data from many of these sensors are voluntarily shared online by the owners (including the citizens, and government agencies like the California Air Resources Board, Bay Area Air Quality Management District and Southern California Air Resource Board). For this study we downloaded the 10-min average $PM_{2.5}$ concentration data from the PurpleAir website (https://www2.purpleair.com/). For the NC 2020 case, we used data from the areas boxed by latitudes [38.77º N, 38.04º N] and longitudes [123.19º W, 121.15º W]; [38.04º N, 37.98º N] and [123.19º W, 121.60º W]; [37.98º N, 37.67º N] and [122.69º W, 121.90º W]; [37.67º N, 37.21º N] and [122.47º W, 121.36ºW] for August and September 2020 (Fig. S2). These boxes cover most of the San Francisco Bay Area and part of the Sacramento County. In this period, residents in this area experienced smoky days caused by the LNU Lightning Complex Fire, the August Complex Fire, the SCU Lightning Complex Fires, the CZU Lightning Complex Fires, and at the end of September the Glass Fire, as well as the massive fires in Oregon (https://www.fire.ca.gov/incidents/2020/). The same study area was used in the NC 2018 case, although fewer sensors were operating at that time. The study area for the SC 2020 case is boxed by [33.47º N, 34.50º N] and [116.85º W, 119.40º W], as shown in Fig. S3.

Selection of Sensor Correction Models. Plantower sensors (Plantower Technology) used by PurpleAir measure the mass of particulate matter by measuring light scattering at 680±10 nm (1). The manufacturer has a proprietary algorithm to convert the light scattering signal to the mass concentration of particulate matter. Each sensor is also embedded with a BME 280 sensor (Bosch Sensortec) to measure the temperature, pressure, and relative humidity in real time. The performance of low-cost $PM_{2.5}$ sensors is dependent on humidity, temperature and level of particulate matters (2–7). Many corrections have been proposed to convert the raw $PM_{2.5}$ data ($PM_{2.5}$ CF=1) measured by Plantower sensors to values consistent with research grade instruments. In our analysis, hourly average primary $PM_{2.5}$ data measured by 16 EPA Air Quality Measurement Stations (AQMSs) in August and September 2020 in the study area was downloaded from the EPA AirNow's API website (https://docs.airnowapi.org/). According to the California Air Resource Board (https://ww2.arb.ca.gov/our-work/programs/ambient-air-monitoring-regulatory/annual-

monitoring-network-report), the primary $PM_{2.5}$ monitors in these sites are MetOne BAM (beta-ray attenuation) continuous monitors. For each EPA measurement site, we compared the data measured by outdoor PurpleAir sensors within 5 km (using at most 50 sensors near each EPA site to avoid data being skewed towards a small number of sites). We excluded outdoor sensors that (a) reporting less than 4 weeks of data. (b) had weak correlation with the EPA station's measurement ($r < 0.8$) because it might be affected by other local pollution sources or it could be listed as an outdoor sensor by mistake, (c) reported $PM_{2.5}$ larger than 800 μg m^{-3} and sensors always reporting data lower than 10 μg m^{-3} as they were either malfunctioning or were operating outside of the recommended limits of detection. In total, data from 446 outdoor sensors surrounding the 16 EPA sites were included in the correction factor evaluation.

To get correction factors for converting PurpleAir sensor measurements to federal reference/equivalent method measurements, some studies performed a linear regression of $PM_{2.5}$ measured by the PurpleAir sensors with data from nearby EPA regulatory instruments (2), while others also considered the effect of temperature and relative humidity on the sensor's performance (5, 8, 9). There are two main types of PurpleAir sensors available for purchase on the PurpleAir website (https://www2.purpleair.com/collections/air-quality-sensors). The PA-I sensors only have one channel (Plantower PMS 1003) for PM measurement. Each PA-II PurpleAir sensor has two Plantower PMS 5003 sensors inside (Channel A and Channel B). Ideally, it is good to average the values reported by the two sensors and to remove some abnormal data because of sensor failures that can be captured by the difference of PM reported for the two channels. However, many sensors did not report $PM_{2.5}$ data from Channel B, presumably because they were the indoor PA-I sensor model. To incorporate as many sensors (buildings) as possible in the analysis, we only used Channel A data if data from both channels are available. According to the evaluation by Barkjohn et al. (8), the $PM_{2.5}$ concentrations reported by Channels A and B agree well. In line with this prior result, we compared 42 sensors with fully available Channel A and B data and found excellent agreement [slopes of linear fit between two channels' $PM_{2.5}$ data have IQR of (0.97, 1.06) with median at 1.01; R^2 of fit between two channels' $PM_{2.5}$ are all above 0.95]. More broadly, we believe that many instances of abnormal data are reliably excluded by our other QA/QC procedures (described in "Other QA and QC" section below). The sensors report both $PM_{2.5}$ CF = 1 data and $PM_{2.5}$ CF = ATM (atmospheric) data. It is not known how the CF = 1 data are converted to CF = ATM data in the

proprietary algorithm from the manufacturer. However, it is known that the ATM data can result in a nonlinearity for concentrations below and above around 20-40 μg m^{-3} (10, 11), while CF = 1 data do not have this problem. The $PM_{2.5}$ CF = 1 data have also been shown to correlate better with the EPA federal reference methods or federal equivalent methods. The $PM_{2.5}$ CF = 1 data were therefore chosen as the raw input data in our calibration.

Seven correction methods were compared in our analysis, with their performance summarized in Table S1. Method 1 and 2 are based on linear regressions (ordinary least square method) of the EPA $PM_{2.5}$ data with the PurpleAir $PM_{2.5}$ CF = 1 data (of individual sensors, not the average of all sensors within 5 km of each AQMS). Method 3 uses an orthogonal distance regression (ODR) with zero intercept. The Barkjohn *et al.*(9) US fire correction was based on comparison of PurpleAir measurement data with collocated federal equivalent methods in 7 sites across the United States affected by prescribed fires, ambient aged fires, woodstove fires and wildfires. It considers the effect of relative humidity on the measurement. A similar field comparison was performed by Holder *et al.* (3). The correction factors from these two studies were also evaluated here for our dataset. We also constructed a "New fit incorporating RH" correction by a multivariate regression of EPA $PM_{2.5}$ against $PM_{2.5}$ and RH measured by nearby outdoor PurpleAir sensors using data in August and September 2020 from the San Francisco Bay Area. In some studies, a nonlinear RH term $RH^2/(1\text{-}RH)$ was used (5, 8). However, recently it has been demonstrated that a linear term of RH can perform even better than the non-linear term (6). Therefore, the linear RH function is used in our "New fit incorporating RH" correction. Finally, using EPA $PM_{2.5}$ as the response, and $PM_{2.5}$ and RH reported by the nearby PurpleAir sensors as input, we trained a binary decision tree for regression model using the Statistics and Machine Learning Toolbox in MATLAB. The temperature term was not included in our correction models because it has been shown that including the temperature term can only negligibly improve the performance of such correction models (5, 6). The commonly used Lane Regional Air Protection Agency (LRAPA) correction, which uses the CF = ATM data in the correction equation (6), was not compared here.

Adding a non-zero intercept to the model did not substantially improve the R^2 or reduce the root mean square error (RMSE). A major disadvantage of adding such an intercept is it can lead to an overestimation when the $PM_{2.5}$ concentration is very low. We also evaluated whether the linear regression

of the EPA $PM_{2.5}$ data with the PurpleAir $PM_{2.5}$ (CF = 1) data are sensitive to the distance threshold. Table S2 shows that the regression coefficients are not very sensitive to the distance threshold from 2 km to 20 km.

In ordinary least square regression, it is assumed that the independent variable is free from errors (13). However, this assumption may not be true for PurpleAir sensor measurements. We therefore also calculated the slope using orthogonal distance regression (ODR). The ODR minimizes the sum of orthogonal distances of the data points from the regression line (13). Using the ODR method changed the correction factor by only 0.01 and increased the RMSE (Table S1). Adding RH in the linear regression also only made an almost negligible improvement. We therefore chose the linear regression without intercept correction. In this case, the fitted correction factor is 0.53. Fig. S6 displays the hourly concentration time profiles of $PM_{2.5}$ measured by each EPA monitor in the San Francisco Bay Area, and the average concentrations of PurpleAir sensors (after correction with CF = 0.53) within 5 km in August and September 2020. They agree reasonably well with each other.

It is important to note that the correction equations evaluated here are only applicable for this analysis, and they should not be generalized to other places and/or at other times. As shown in Holder *et al*. (3), even the correction factors for wildfire smoke from different fires in the US can differ by a factor of more than 2. It is also worth noting that our analysis is not heavily dependent on the exact correction factor because the concentration ratios are the targets. The correction factors only affected which peaks were defined as indoor source peaks. When wildfire smoke affected a region, the composition of indoor and outdoor $PM_{2.5}$ were expected to be similar because wildfire particles dominated even in the indoor environments (Table 1). Therefore, it is reasonable to use the same correction for both indoor and outdoor $PM_{2.5}$, especially we focus on the indoor/outdoor ratios, as suggested by Bi et al. (14). .

As shown in Table S1 and Fig. S15, the binary decision tree method can improve the correlation of PurpleAir data with EPA measurements. Results from the same analysis with this correction are shown in Fig. S16. The trend of the result is the same as the no-correction case, but the difference between the fire days and non-fire days are larger, which is probably due to a non-zero intercept in the correction.

We also performed regression for the correction of Greater Los Angeles Area sensors (SC 2020 case). Based on linear regression of EPA monitor data with the nearby PurpleAir sensor using the same

approach as in the NC 2020 case, a correction factor of β_1 = 0.58 was adopted (NRMSE = 0.42, see Figures S8-S9). Similar analysis has been performed by Delp and Singer (2) for San Francisco Bay Area sensors in November 2018. A correction factor of β_1 = 0.48 was adopted in our analysis.

Other QA and QC. We selected indoor sensors that had measurement value for at least 1/6 of the time (~10 days) in the two-month period considered in our study. We found 1459 indoor monitors in this region meeting this criterion. For each indoor sensor selected, we used its longitude and latitude to locate the nearest outdoor sensor. More than 2000 outdoor sensors in this region reported at least 10 days data during the period considered, compared with only 16 EPA Air Quality Monitoring Stations (AQMS) in this region. The geometric mean (GM) distance from an indoor sensor to the nearest AQMS is 6.7 km, but it is only 0.21 km to the nearest outdoor sensor (Fig. S17). The substantially reduced distance allows much more accurate evaluation of indoor/outdoor concentration relationships. To prevent the possibility that the nearest outdoor sensor was located near major pollution sources, when the nearest outdoor sensor is more than 500 m away from the indoor sensor, we required the 50th percentile concentration at this outdoor node when it was not affected by wildfires to be below 25 μg m^{-3}, according to the levels and spatial decay rate of $PM_{2.5}$ measured near roads (15–18). We further required the outdoor sensor to cover at least 85% of the time when the indoor sensor reported data. If the $PM_{2.5}$ concentration measured by an "indoor" sensor is correlating too well with a nearby outdoor sensor (r^2 > 0.8), it is likely that this sensor was placed outdoors. This mislabeled or dislocated sensor is therefore not used in the analysis. Fig. S18 shows an example of an indoor node discarded for this reason. We removed 165 "indoor" sensors from the analysis because of this problem. Another 20 indoor sensors were not considered because we could not find a nearby outdoor sensor that reported data for more than 85% of the time when the indoor sensor reported data. With all these criteria in place, data from N = 1274 indoor sensors in this region could be used. The same procedure was applied to data in the NC 2018 and SC 2020 cases. Negative values of $PM_{2.5}$ concentration were also discarded.

Decomposition of Indoor $PM_{2.5}$ We separated the indoor $PM_{2.5}$ from indoor and outdoor origins by removing short-term indoor $PM_{2.5}$ peaks that were unlikely due to penetration. A very similar approach has

been demonstrated in previous studies by Allen *et al.* (19, 20). According to high time-resolution measurements of particulate matter in previous indoor studies, the major indoor emission processes (mainly cooking and cleaning) typically last for half an hour to an hour, and after that a longer period is needed for the $PM_{2.5}$ perturbation to decay to less than half of its peak value (21). When these processes happen, the indoor level of $PM_{2.5}$ was at least 30 μg m^{-3}. We therefore selected all the peaks with half-prominence width (*w*) between 1 hour and 4 hour and prominence level above 30 μg m^{-3} as indoor-source peaks. It is possible that in some buildings the windows were opened for around an hour during the fires and created peaks that meet this criterion. Out of the 1274 buildings considered, we identified these large indoor source peaks in 834 buildings. Buildings without such peaks might be commercial buildings without large indoor $PM_{2.5}$ sources, or the sensor in that building was placed in a location free from large indoor emissions. We assumed the indoor $PM_{2.5}$ other than that caused by these large peaks to be infiltrated $PM_{2.5}$. We reconstructed the infiltrated $PM_{2.5}$ by linearly interpolating indoor $PM_{2.5}$ concentration 3*w* before and after these large peaks with respect to time. The long 3*w* window was chosen to ensure that the indoor source peaks can be more thoroughly removed. For data outside of this window, the indoor concentration was assumed to be equal to the infiltrated $PM_{2.5}$. As a QA/QC step, if the calculated non-cooking indoor concentration was higher than outdoor concentration, that data point was removed from the analysis.

Mass Balance Model and Total Indoor Particle Loss Rate Constant Calculation. The indoor concentration of $PM_{2.5}$ depends on infiltration, indoor emission, and loss. We explored the dynamics of indoor $PM_{2.5}$ with a box model. If we assume the $PM_{2.5}$ is well-mixed indoors, the mass balance of $PM_{2.5}$ in a building can be written as:

$$V\frac{dC_{in}}{dt} = aPVC_{out} - aVC_{in} - k_{loss}VC_{in} + S \quad \text{(S1)}$$

where *V* is the volume of the room, *a* is the air exchange rate, *P* is the penetration factor of particles, k_{loss} is the loss rate constant including deposition and indoor filtration, and *S* is the indoor emission rate. C_{in} and C_{out} are the indoor and outdoor concentrations, respectively (22, 23). Dividing by *V* on both sides, we can simplify the equation to:

$$\frac{dC_{in}}{dt} = aPC_{out} - (a + k_{loss})C_{in} + \frac{S}{V} \quad \text{(S2)}$$

When the indoor and outdoor particles are in steady state, and S is small, we have:

$$\frac{dC_{in}}{dt} = 0 = aPC_{out} - (a + k_{loss})C_{in} \Rightarrow F_{in} = \frac{C_{in}}{C_{out}} = \frac{aP}{a + k_{loss}} \quad \text{(S3)}$$

where F_{in} is the infiltration factor. For particulate matter, F_{in} can be obtained from the ratio of indoor/outdoor concentration when there are no outdoor sources (23, 24). Another way to estimate F_{in} is to regress the indoor $PM_{2.5}$ on outdoor values (25). However, this method has been shown to underestimate the infiltration factor while overestimating the indoor background (23), or produce infiltration factors outside [0,1] (14). Therefore, the ratio method was used for our analysis.

During the peak of cooking-like indoor particle release events, the indoor $PM_{2.5}$ resulting from cooking is much larger than the infiltrated smoke. When the indoor emission event is over, we assume the indoor source term becomes 0, and we have:

$$\frac{dC_{in}}{dt} = -(a + k_{loss})C_{in} \Rightarrow C_{in}(t - t_{peak}) = C_{in,peak}e^{-(a+k_{loss})(t-t_{peak})} \quad \text{(S4)}$$

Therefore, (a+k_{loss}) can be estimated by fitting the curve of $C_{in}(t)$ (26). We define (a+k_{loss}) as the total indoor particle loss rate constant (λ_t). A peak prominence of 30 μg m^{-3} (20 μg m^{-3} in the SC 2020 case to incorporate more peaks) was used as the threshold to find large indoor peaks that were subsequently used in the particle loss rate constant calculation. If windows were opened and then closed, the decay of resulted indoor $PM_{2.5}$ can also be described by Equation S4. Those peaks were also included because the decrease of indoor $PM_{2.5}$ under that circumstance can also be described by the exponential decay. The decay rate constant is also not substantially affected by the correction factor used because the correction factor affects C_{in} and $C_{in,\ peak}$ in the same way. To get total particle loss rate $\lambda_t = a + k_{loss}$, Equation S4 can be rewritten as:

$$-\ln\frac{C_{in}(t)}{C_{in}(t_{peak})} = \lambda_t(t - t_{peak}) \quad \text{(S5)}$$

We then linearly fitted this equation by least square method to get slope λ_t for the decay of each peak of indoor $PM_{2.5}$. In this part, we no longer require the width of the peak to be above 1 hour. In this way, indoor $PM_{2.5}$ peaks resulting from short-time window opening were also used to get λ_t. The 95% confidence interval of λ_t was also calculated. To ensure the exponential decay model is applicable, if the lower bound of the confidence interval of λ_t for a peak was below zero, this peak was not used as data for Fig. 5.

The decrease of indoor $PM_{2.5}$ concentration can also be caused by the decrease of outdoor $PM_{2.5}$ concentration. In such cases, the assumption that incoming outdoor $PM_{2.5}$ source is stable no longer holds. Therefore, if the indoor $PM_{2.5}$ was decaying together with the outdoor $PM_{2.5}$ measured by the nearest sensor (r^2 > 0.8), this peak was excluded from the analysis. For the 1274 buildings considered in the NC 2020 case, we observed such decay peaks in 1000 buildings. On average, 4.7 decay peaks were captured in each building in the two-month period.

Uncertainty of the infiltration ratios and the decay rate constants

We roughly estimated the uncertainty of the infiltration ratios of individual sensor pairs, based on the idea that disagreement among any two paired sensors would lead to an uncertain estimate of the ratio of concentrations between those sensors. Thus, we gain a magnitude estimate of the uncertainty of the indoor/outdoor concentration ratio by examining the disagreement among a large number of paired nearby *outdoor* sensors across the PurpleAir dataset in our domain. We consider two timescales: (1) the uncertainty of the indoor/outdoor ratios of the 10-min data, reflecting the transient noise at short time scales, and (2) the uncertainty of the infiltration ratio for a building over the two-month period in the analysis, reflecting the possible range of persistent-sensor-to-sensor bias. To do so, we first found the outdoor sensors that were used to calculate indoor/outdoor ratios. Since it is possible that the nearest outdoor sensor of multiple indoor sensors is the same sensor, for the 1274 pairs of sensors, there are only 784 outdoor sensors used. For each sensor, we tried to find the nearest outdoor sensor within 1 km, which was successful for 775 sensors. For each pair of sensors i at time j, we calculate the ratios of $x_{i,i}/y_{i,j}$, where $x_{i,I}$ is the concentration of the i^{th} of the 775 used outdoor sensors at time t, and $y_{i,j}$ is the concentration of its nearest outdoor sensor. We make $\boldsymbol{R_i}$ as:

$$\boldsymbol{R_i} = \left[\frac{x_{i,1}}{y_{i,1}}, \frac{x_{i,2}}{y_{i,2}}, \ldots \frac{x_{i,n}}{y_{i,n}} \right]^T \quad \text{(S6)}$$

Then we concatenate the $\boldsymbol{R_i}$ array into $\boldsymbol{R_{all}}$ array by:

$$\boldsymbol{R_{all}} = \left[\boldsymbol{R}_1 \ \boldsymbol{R}_2 \ \ldots \boldsymbol{R}_n \right]^T \quad \text{(S7)}$$

The uncertainty of the indoor/outdoor ratio of 10-minute data is reflected by the variation of $\boldsymbol{R_{all}}$, which yields 0.886, 1.005, and 1.138 as 25^{th}, 50^{th}, and 75^{th} percentiles values, respectively.

The uncertainty of the infiltration ratio for a building over the two-month period can be roughly estimated by the statistics of $median(\boldsymbol{R}_i)$, which has 0.955, 1.035, and 1.131 as 25th, 50th, and 75th percentiles values, respectively. Therefore, we can conclude that the uncertainty of the infiltration ratio for a building over the two-month period is less than ±10%.

The decay rate calculation should have a very small uncertainty due to any bias in PA sensors because it uses measurements only from a single indoor sensor. We were fitting the decay curves of individual sensors by:

$$-\ln\frac{C_{in}(t)}{C_{in}(t_{peak})}=\lambda_t(t-t_{peak}) \quad \text{(S5)}$$

in which $C_{in}(t)$ ratios by the same sensor (especially in the same peak) should have very small uncertainty.

Given the reasons stated above, we expect the exposure reduction calculations have even lower uncertainties because we are averaging the exposure reduction of the 1274 buildings. Assume uncertainty of the infiltration ratio for a building over the two-month period is 10%, the average of infiltration ratios of all the buildings will have an uncertainty of $10\%/\sqrt{1274}=0.28\%$ following central limit theorem. More conservatively, the median uncertainty of the infiltration ratio over two months, as reported for 774 sensor pairs above, was 1.035, or 3.5%. In either case, this uncertainty is quite small. We expect that the average exposure reduction would have a quantified uncertainty of similar or better magnitude to the I/O ratio, in other words, well less than 5%. Other unquantifiable uncertainties – e.g., differential or non-linear response of the PurpleAir to time-varying aerosol properties – add additional uncertainties that are more difficult to directly estimate, but we believe that these uncertainties do not fundamentally undermine the validity of our qualitative results.

Building information. Property data for PurpleAir Indoor-Outdoor comparison analysis were obtained by matching coordinates to addresses, verifying the addresses, looking up the addresses on publicly available property listing services, and finally quality control of the resulting data. The latitude-longitude coordinates were obtained from the publicly available PurpleAir database formally from a PurpleAir JSON file (purpleair.com/json – defunct as of December 2020), now available through the official PurpleAir API (api.purpleair.com). The coordinates contain 6 decimal places of precision and thus are accurate to under

10 meters, however, the placement is based on the available WiFi signal and can be edited by the sensor owner to be located anywhere on the map. As such, there is some uncertainty introduced into the reverse geocoding process, but since citizen scientists are interested in air quality within their own homes and research groups require spatial fidelity it can be assumed these coordinates are approximately correct.

After obtaining the list of coordinates, the Google and ArcGIS geocoding engines performed reverse geocoding scripted using the Python library OSMnx 1.0.1 (osmnx.readthedocs.io/en/stable/). About 38.5% of addresses in the SF Bay region disagreed between Google and ArcGIS lookups. The reasons for the disagreements are due to placement in homes leading to low confidence assigning addresses to lots such as on a street corner. The sensor labels and manual searches on Google Maps were used to confirm the address. If the sensor label contains the address or a partial address or is obvious from the Google Map manual search, the confidence to the matched address is high. If the reverse geocoded searches match, then the confidence is medium, otherwise, it is assigned low confidence. From this analysis of valid address (n=1274), 13% were assigned high confidence, 73% were assigned medium confidence, and 14% were assigned low confidence. For low confidence addresses, the ArcGIS address was used.

The list of addresses was then manually inputted to Zillow, a publicly accessible website which offers data on homes and apartments using multiple listing services and county databases including building age, HVAC information, and livable area. Zillow furthermore uses existing publicly available information as well as a proprietary algorithm to derive an estimate of the current (as of December 2020) evaluation of the home or apartment (rent if a rental unit) termed a "Zestimate®." If the address matches an apartment complex, the first listed unit was then used to find the year of construction, HVAC information, and a bell-weather of the typical price and area of apartments since these can vary within complexes. From the 1274 address, 79.5% returned the year of construction, 83.6% returned HVAC data, 76.7% returned a price estimate, and 72.2% returned the area. Out of the 1274 buildings analyzed, 1112 (87%) buildings were found to be residential. Among these residential buildings, 80%, 13%, and 4% were matched to single-family houses, condominiums or multi-family buildings, and apartments, respectively.

As an additional sensitivity analysis, we restricted our dataset to the 87% of buildings that could unambiguously by ascertained to be residential. For this restricted dataset, the mean infiltration ratios on both fire days and non-fire days changed by less than 0.01.

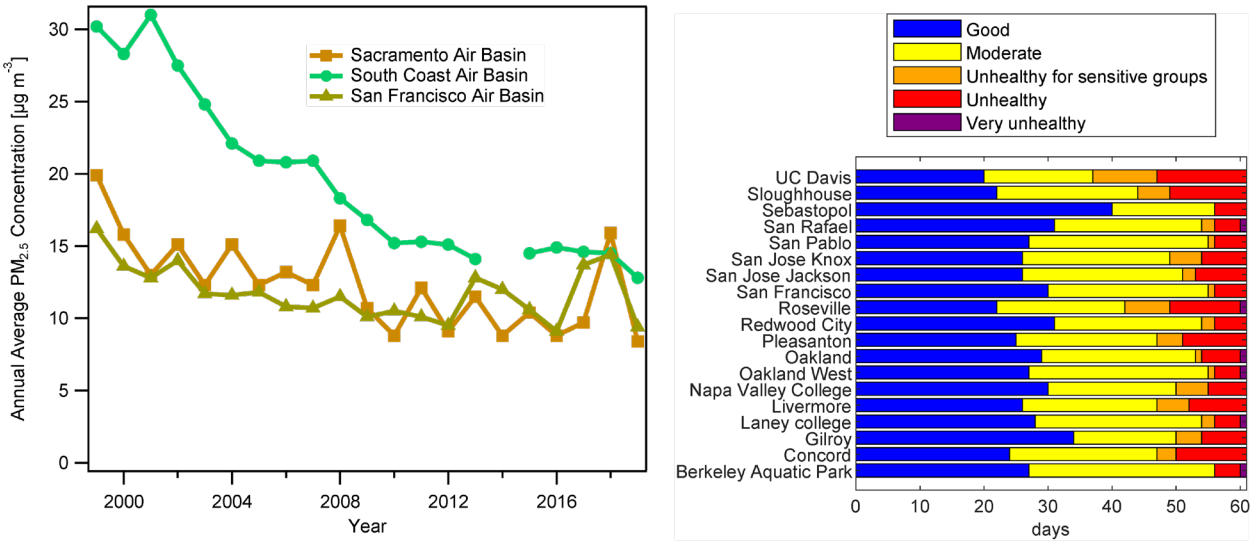

Fig. S1. A. Annual average $PM_{2.5}$ concentrations in 3 Air Basins in California from 1999 to 2019 (Data retrieved from California Air Resource Board website https://www.arb.ca.gov/adam). The missing point is because of insufficient data available to determine the value. **B.** San Francisco Bay Area Air Quality Index (AQI) category in August and September 2020 based on 24-hour average level of $PM_{2.5}$ at each EPA Air Quality Measurement Station. 0 - 15.4 μg/m^3: Good; 15.5 - 35.4 μg/m^3: Moderate; 35.5 - 55.4 μg/m^3: Unhealthy for sensitive groups; 55.5 - 150.4 μg/m^3: Unhealthy; 150.5 - 250.4 μg/m^3: Very unhealthy.

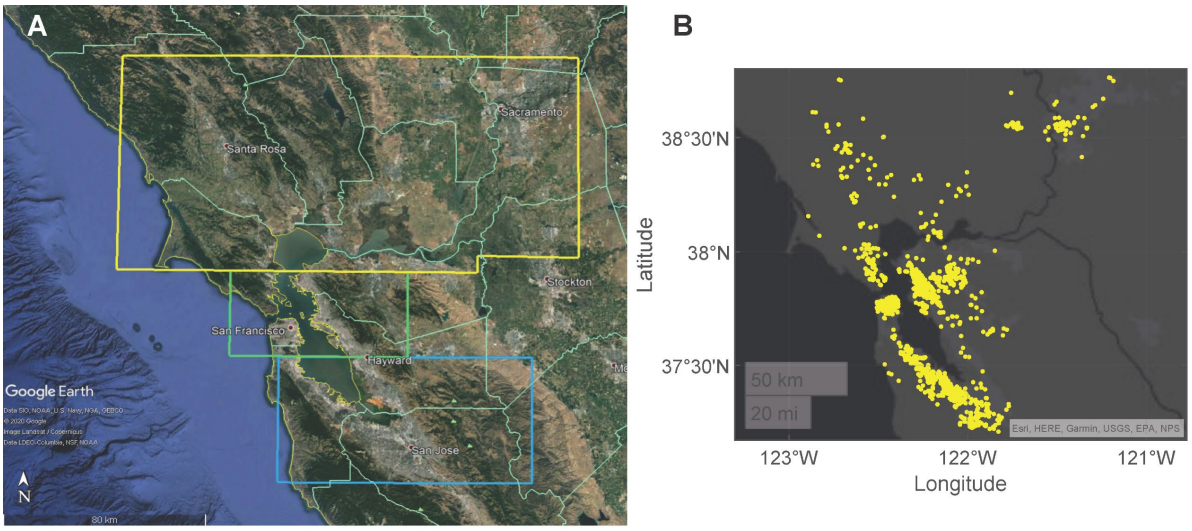

Fig. S2. A. Study regions in the San Francisco Bay Area. Google Earth imagery © 2020 Google. PurpleAir sensors in the three boxes were analyzed together. **B.** Locations of all the indoor PurpleAir sensors included in the NC 2020 case.

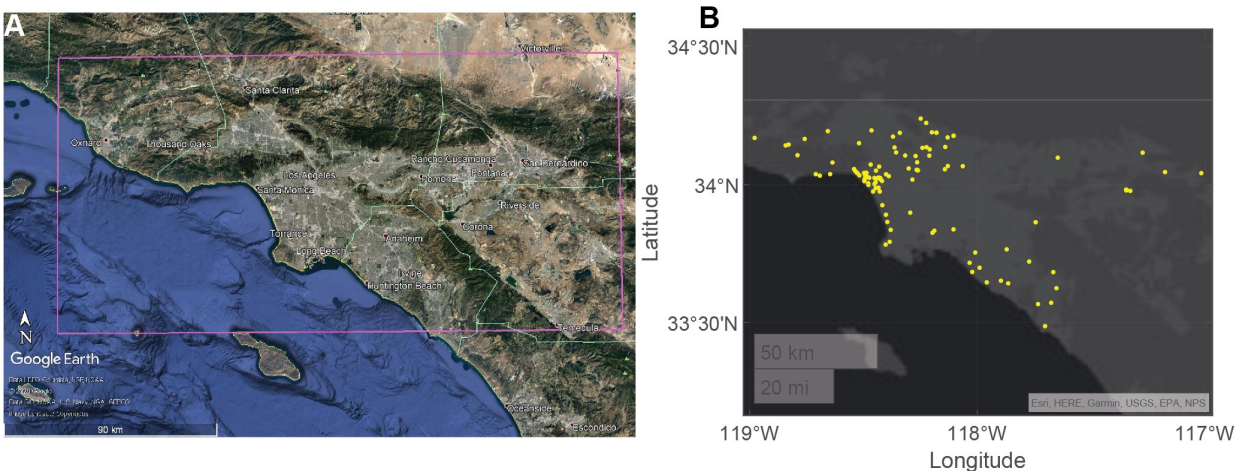

Fig. S3. A. Study region in the Greater Los Angeles Area. Google Earth imagery © 2020 Google. **B.** Locations of all the indoor PurpleAir sensors included in the SC 2020 case.

335

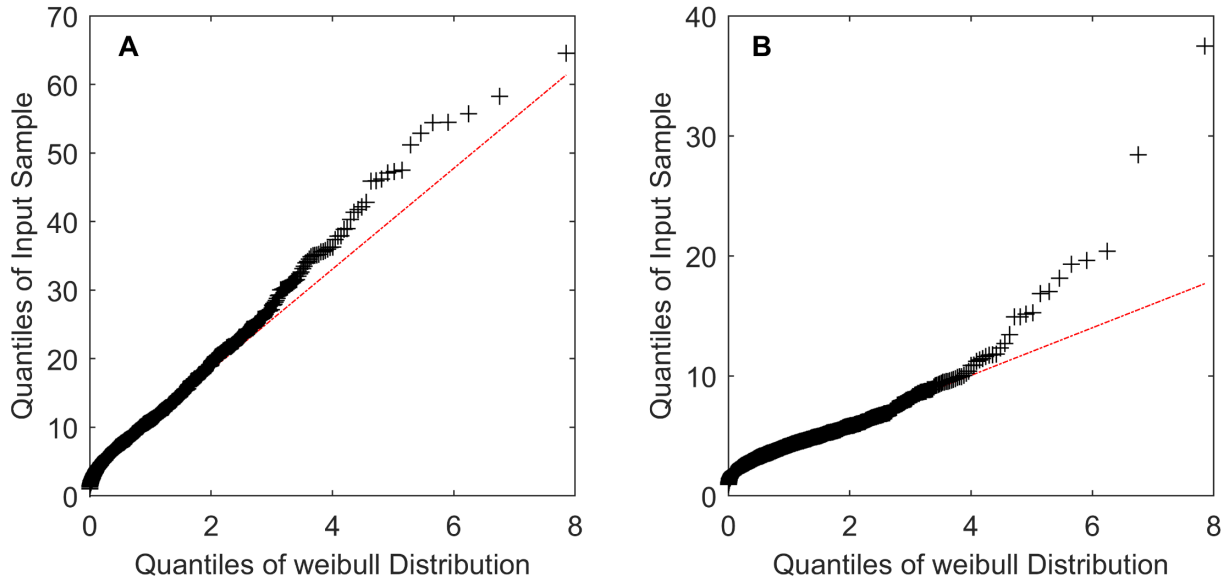

Fig. S4. Quantile-quantile plots of mean indoor $PM_{2.5}$, on the fire days **(A)** and non-fire days **(B)** against Weibull distribution. The reference line represents the theoretical Weibull distribution.

340

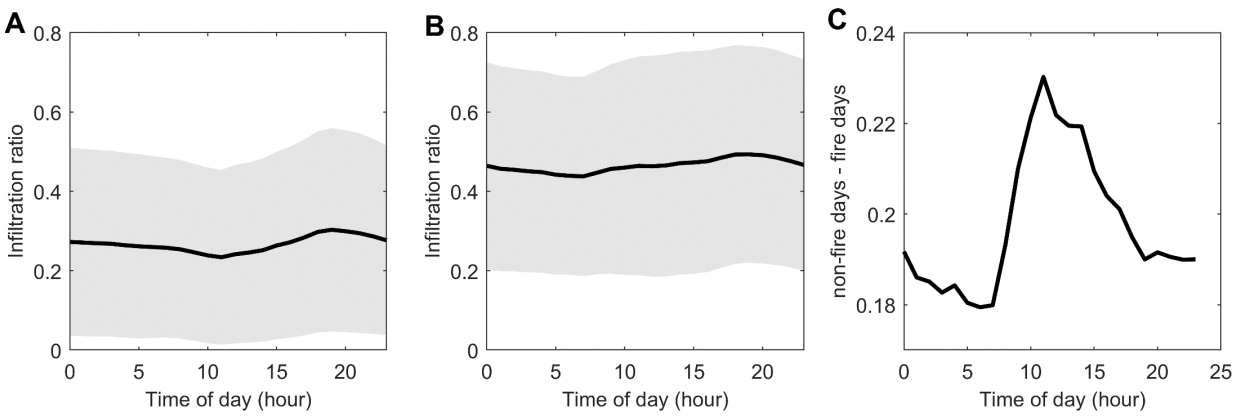

Fig. S5. Diel plots (local time) **A.** Infiltrated $PM_{2.5}$ /outdoor $PM_{2.5}$ on fire days and **B.** Infiltrated $PM_{2.5}$ / outdoor $PM_{2.5}$ on non-fire days **C.** Diel plot of the difference in infiltrated $PM_{2.5}$ / outdoor $PM_{2.5}$ (non-fire days – fire days). Gray shading in A & B shows the standard deviation. Data are average of all the PurpleAir sensors in the NC 2020 case. The difference in mean infiltration ratio between fire days and non-fire days are most apparent in the daytime, consistent with more ventilation typically occurring during daytime.

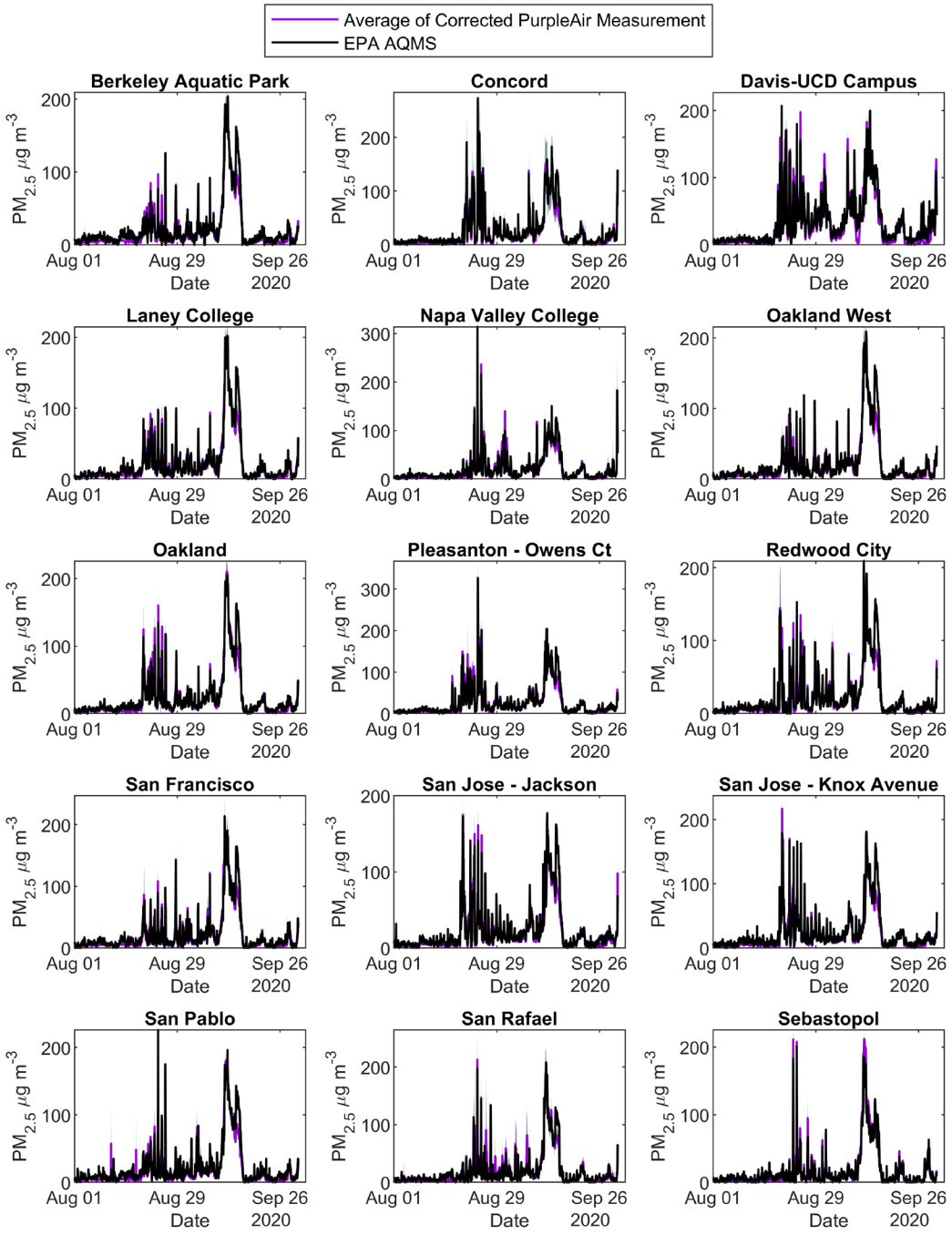

Fig. S6. Hourly time profile of $PM_{2.5}$ concentration of the EPA monitors (black) and mean (purple) ± standard deviation (gray) of $PM_{2.5}$ (corrected) measured by nearby PurpleAir sensors in the San Francisco Bay Area in August and September 2020. The plots only include EPA monitoring stations having at least three outdoor PurpleAir sensors within 5 km of them. The EPA measurement and nearby PurpleAir sensors measurement agree reasonably well with each other.

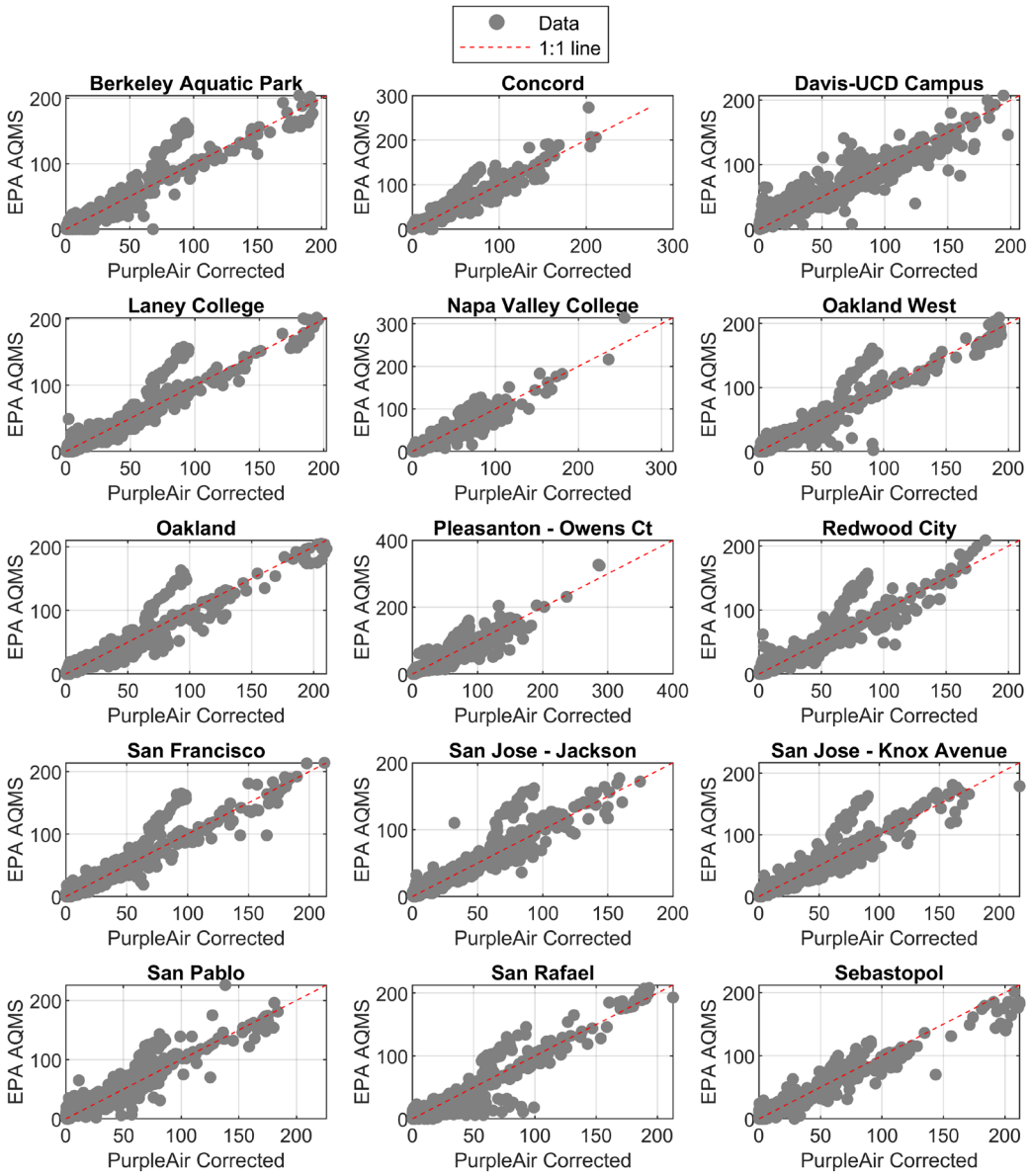

Fig. S7. Scatter plot of $PM_{2.5}$ (μg m^{-3}) of the EPA monitors and mean $PM_{2.5}$ (corrected) measured by nearby PurpleAir sensors in the San Francisco Bay Area in August and September 2020.

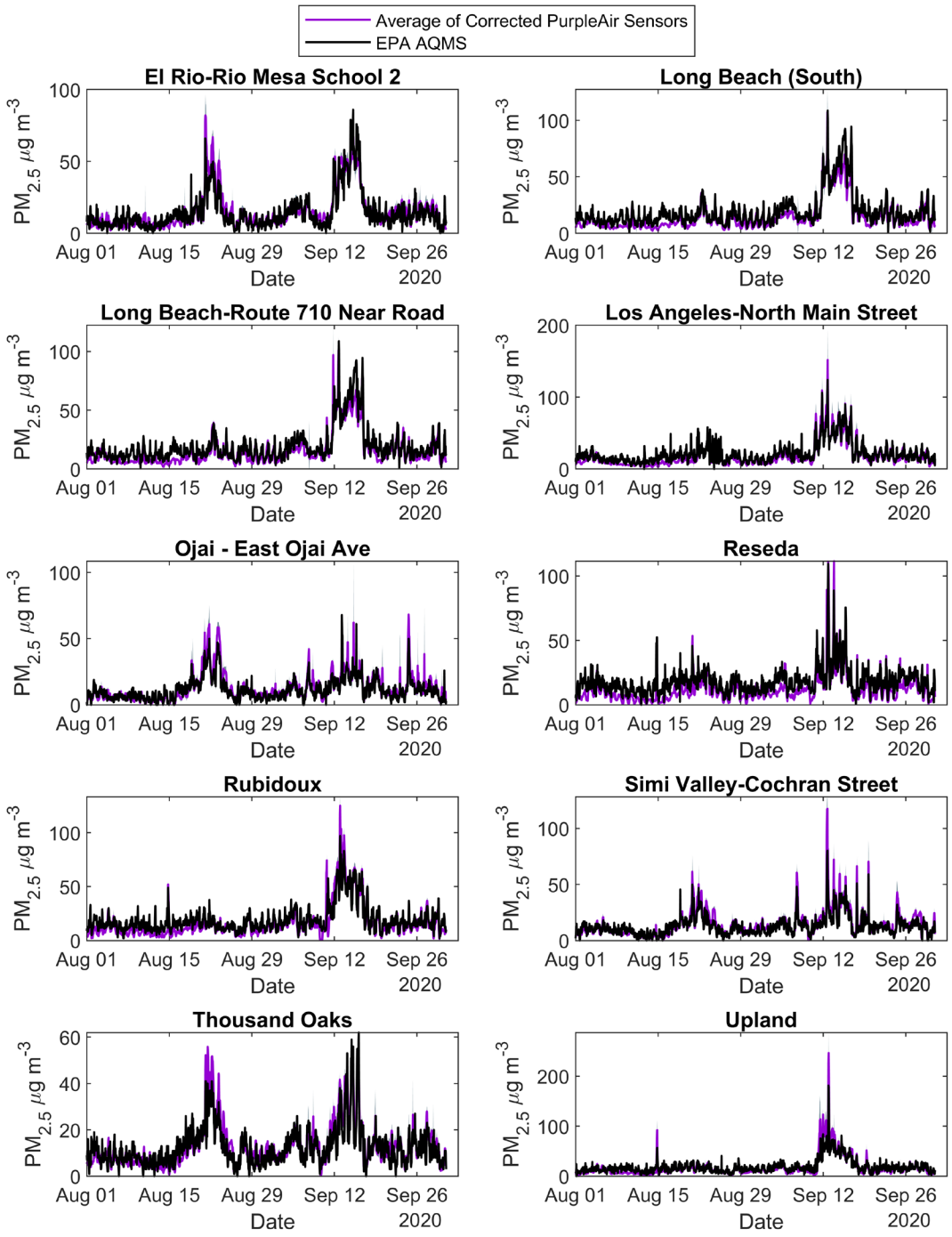

Fig. S8. Hourly time profile of $PM_{2.5}$ concentration of the EPA monitors and mean (purple) ± standard deviation (gray) of $PM_{2.5}$ (corrected) measured by nearby PurpleAir sensors in the Greater Los Angeles Area in August and September 2020. The plots only include EPA monitoring stations having at least three outdoor PurpleAir sensors within 5 km of them.

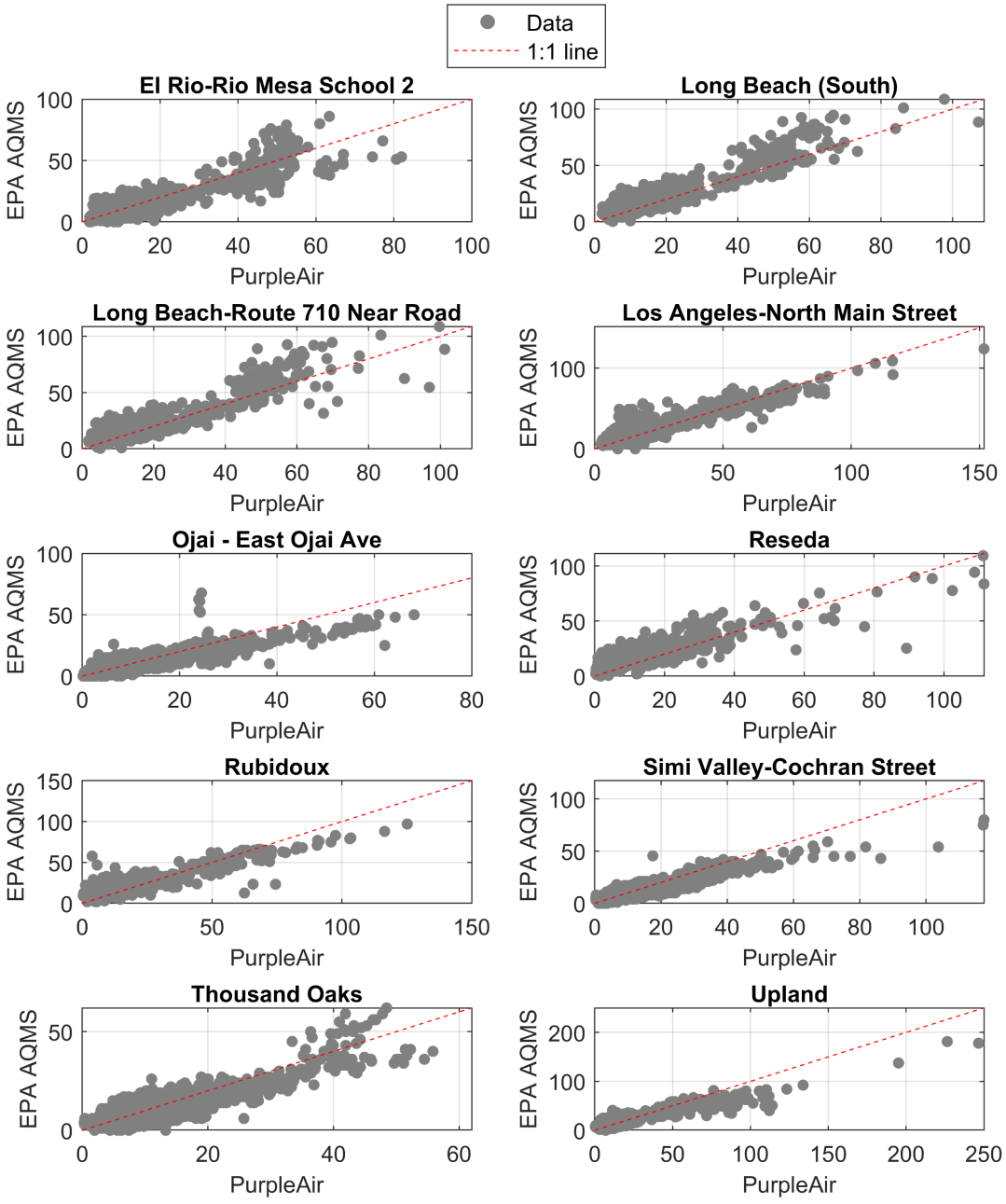

Fig. S9. Scatter plot of $PM_{2.5}$ (μg m^{-3}) of the EPA monitors and $PM_{2.5}$ (corrected) measured by nearby PurpleAir sensors in the Greater Los Angeles Area in August and September 2020

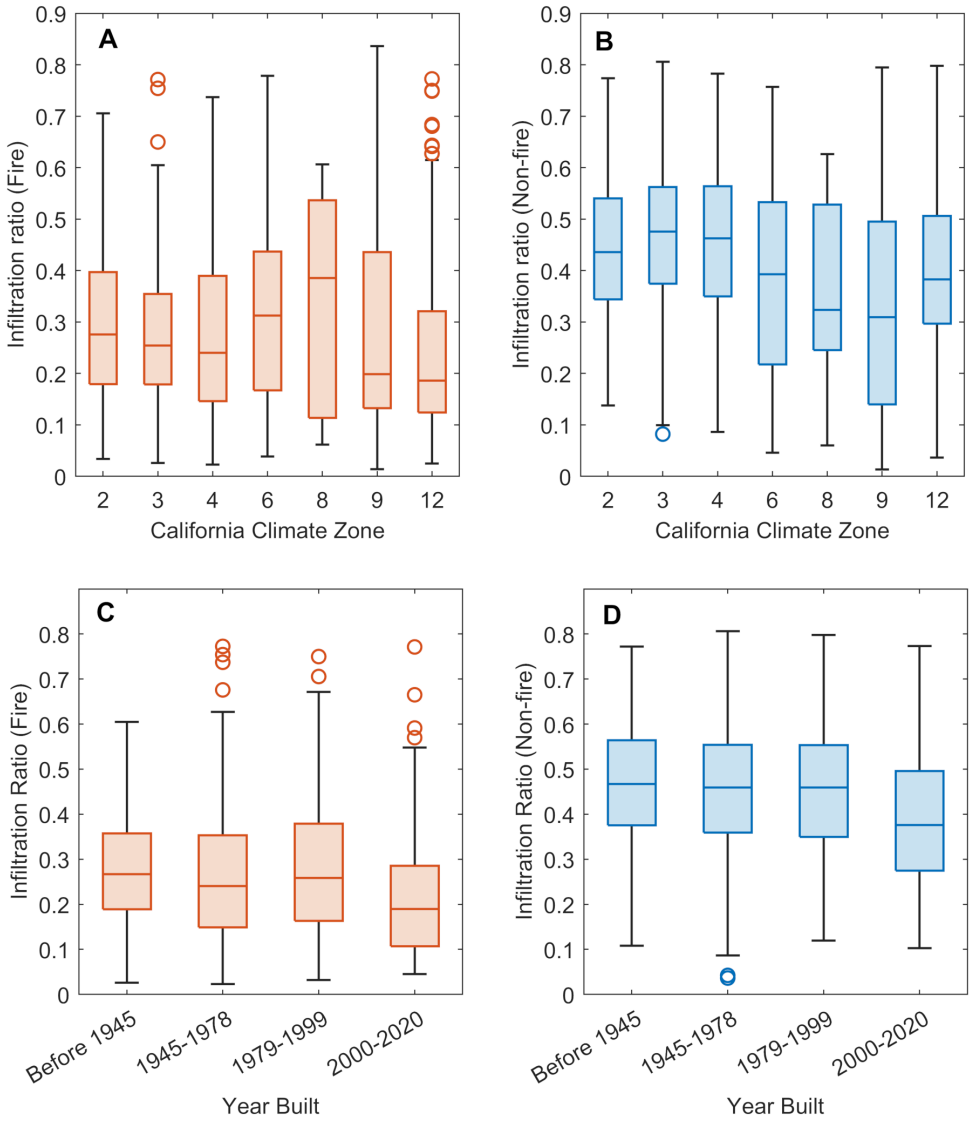

Fig. S10. Infiltration ratio of buildings in different climate zones **A.** on fire days (ANOVA $p = 0.004$); **B.** non-fire days (ANOVA $p < 10^{-3}$) in August and September 2020. Only climate zones with at least 10 indoor sensors being analyzed are included in this figure. Reference cities for different climate zones (which were included in our study) are: Zone 2-Napa, Zone 3-San Francisco & Oakland, Zone 4-San Jose, Zone 6-Los Angeles (LAX), Zone 8-Long Beach, Zone 9-Los Angeles (Civic Center), Zone 12-Sacramento (27). Infiltration ratio of residential buildings (NC 2020 case) built in different periods **C**. on fire days (ANOVA $p = 0.004$); and **D.** on non-fire days (ANOVA $p < 10^{-3}$). Only residential buildings are considered in C and D. Buildings in Zone 12 had lower infiltration ratios than other Northern California climate zones considered.

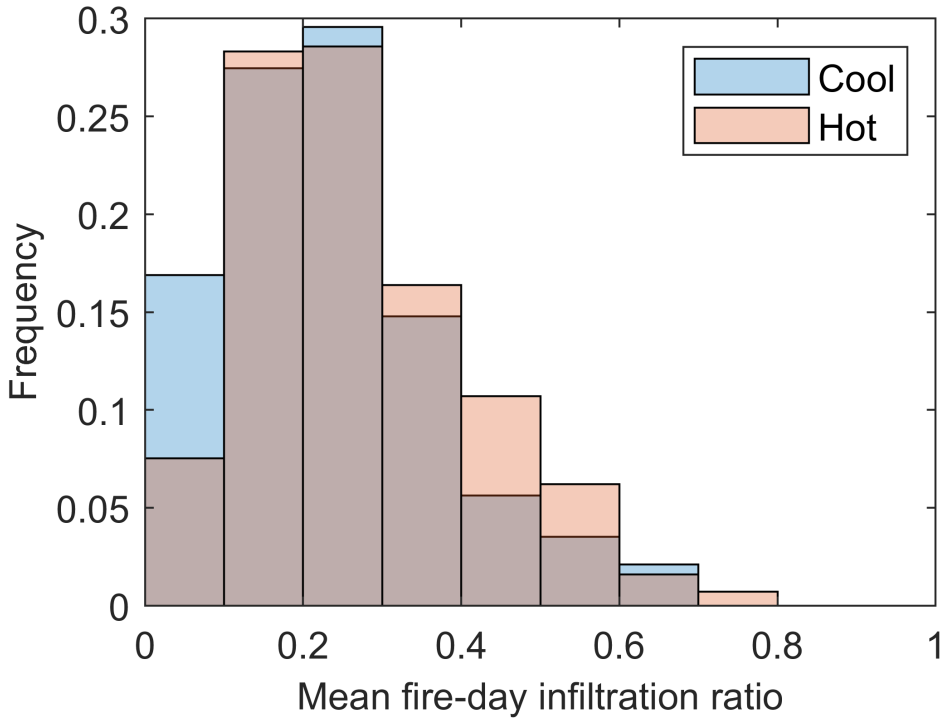

Fig. S11. Infiltration ratio on fire days for cool buildings (95th percentile indoor temperature < 30ºC, *N* = 142) and hot buildings (95th percentile indoor temperature ≥ 30ºC, *N* = 1132) in the San Francisco Bay Area in August and September 2020. The cool buildings have significantly lower fire-day infiltration ratios than the hot ones ($p < 0.01$), and 17% of cool buildings had extremely low infiltration ratios (< 0.1).

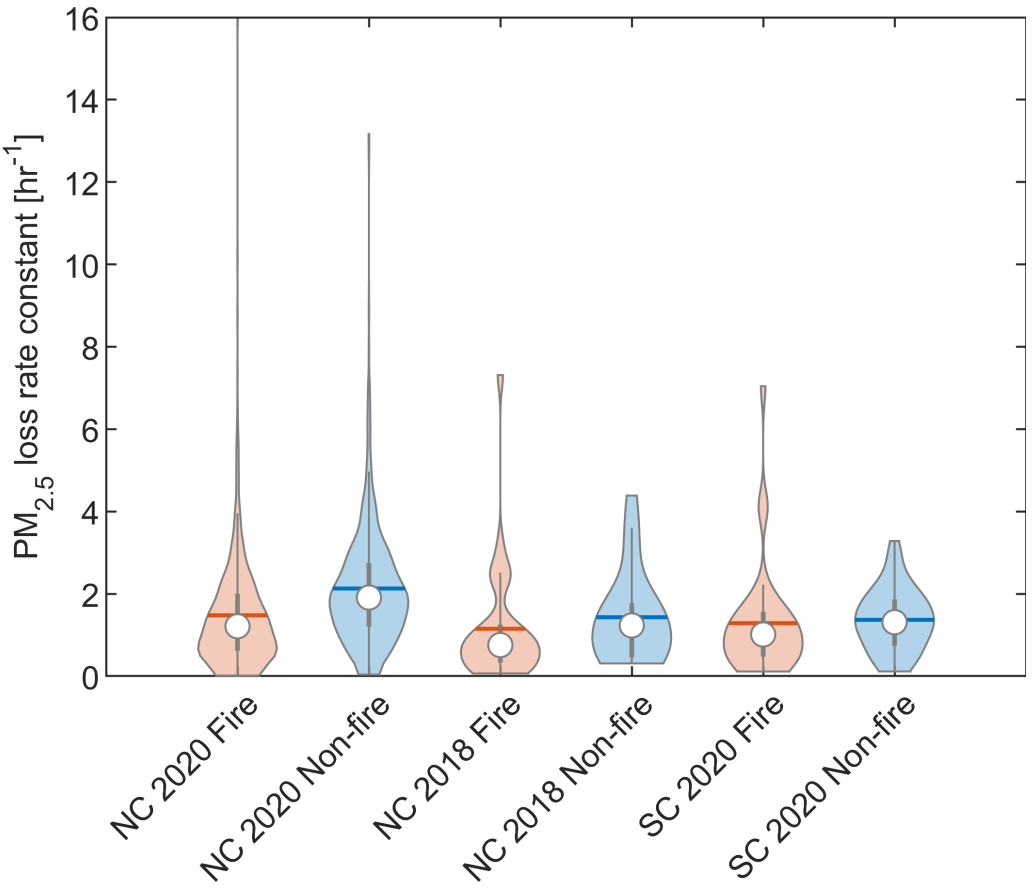

Fig. S12. Violin plot of total particle loss rate constant in buildings in on the fire days and non-fire days. NC = San Francisco Bay Area, SC = Los Angeles Area. Each violin plot shows the probability density of the total $PM_{2.5}$ decay rate and a boxplot of interquartile range with whiskers extended to 1.5 times the interquartile range. Circles indicate the median, and horizontal lines indicate the mean.

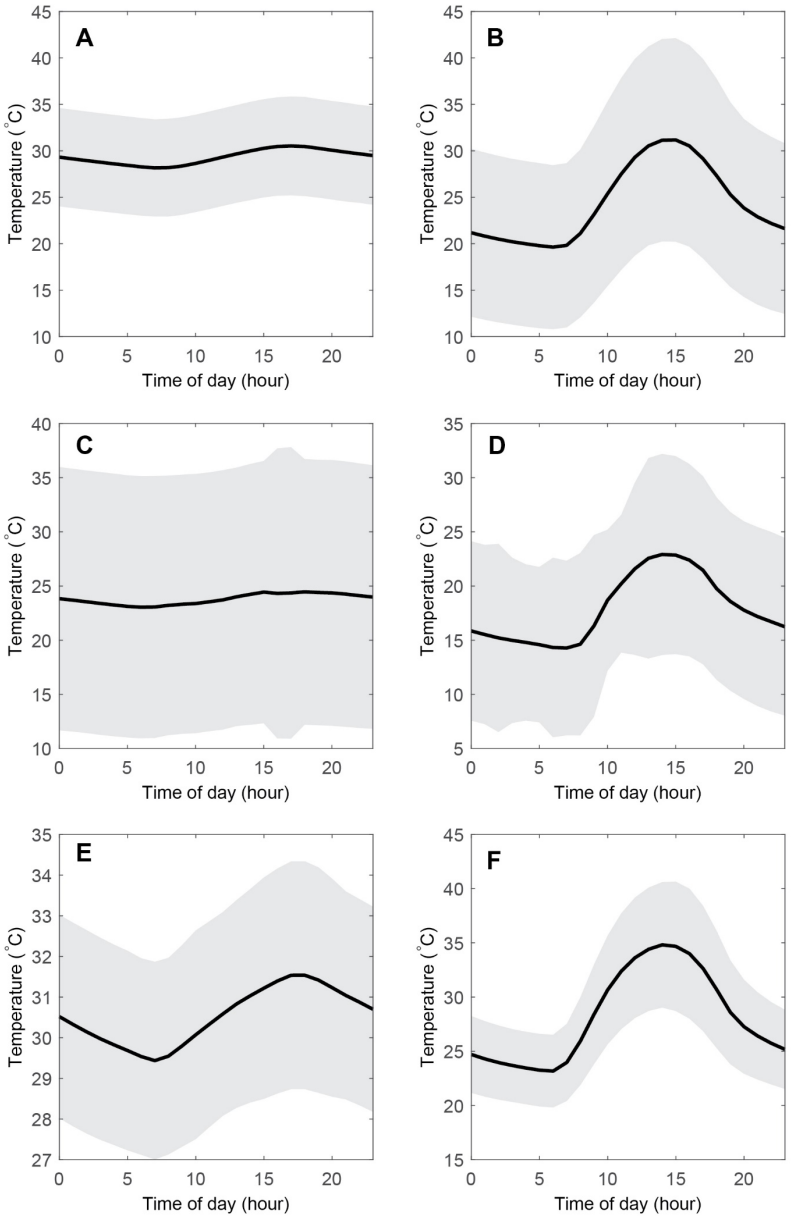

Fig. S13. Diel plots (local time) of average temperature measured by PurpleAir sensors in the San Francisco Bay Area in August-September 2020 (**A**. Indoor **B**. Outdoor) and November 2018 (**C**. Indoor **D**. Outdoor); and in August-September 2020 in Greater Los Angeles Area (**E**. Indoor **F**. Outdoor). Gray shading shows the standard deviation. In the Summer 2020 cases, the difference in daytime indoor/outdoor temperature alternated between positive and negative values. In the NC November 2018 case, the indoor temperature was almost always higher than the outdoor temperature.

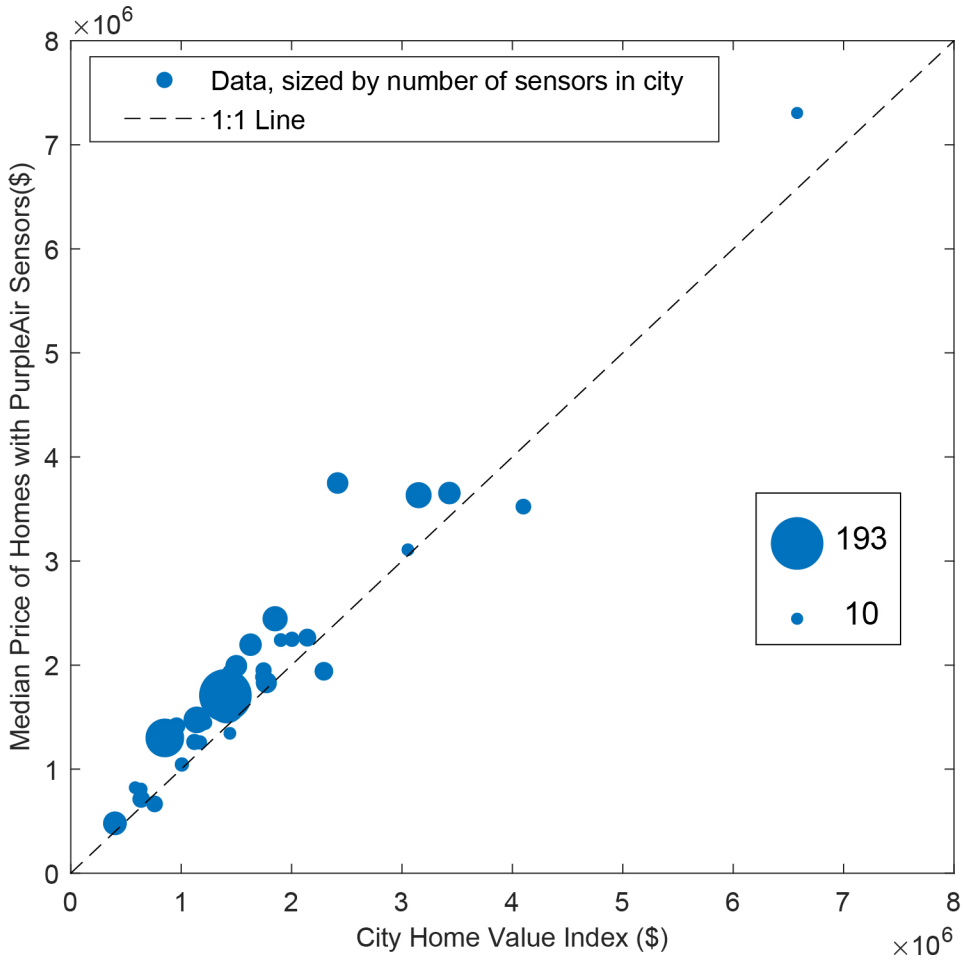

Fig. S14. Median price of homes with indoor PurpleAir sensors vs. Median Housing Price in that city, sized by the number of indoor sensors in that city (only showing data from cities with at least 10 buildings with valid indoor sensors in the NC 2020 case). PurpleAir owners live in homes with estimated average property values 21% greater than the median property value for their cities.

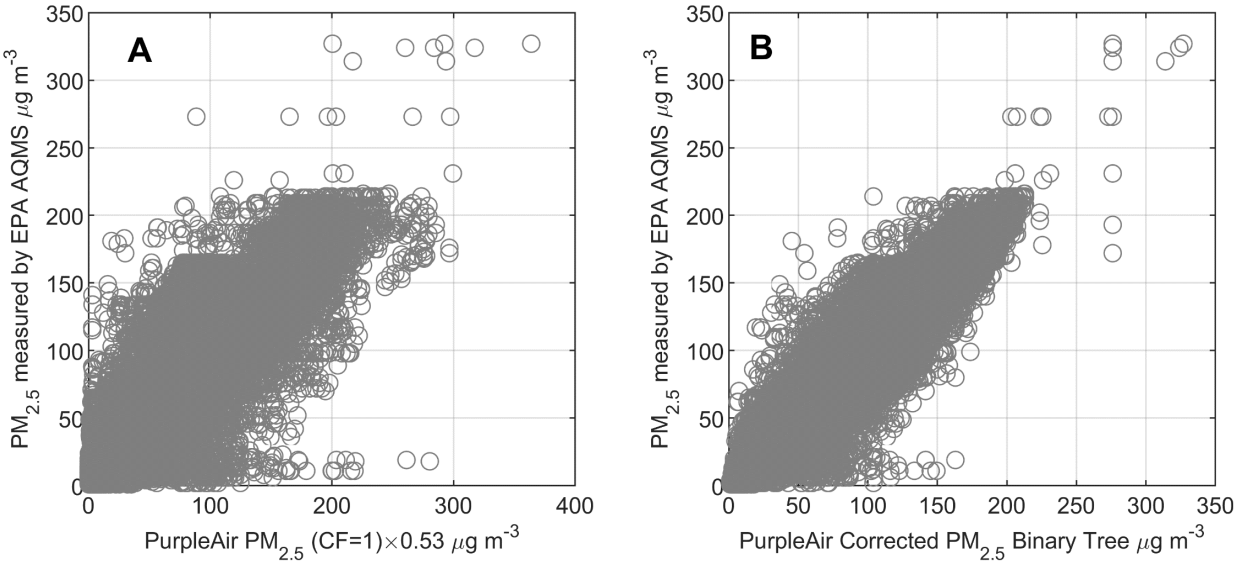

Fig. S15. A. Hourly $PM_{2.5}$ measured by EPA AQMS against the linearly corrected (correction factor = 0.53) $PM_{2.5}$ data measured by nearby PurpleAir sensors; **B.** Hourly $PM_{2.5}$ measured by EPA AQMS against $PM_{2.5}$ measured by the PurpleAir sensors after the binary tree correction, both for data in San Francisco Bay Area in August and September 2020. This figure demonstrates the binary tree model can improve the precision and accuracy of the sensors.

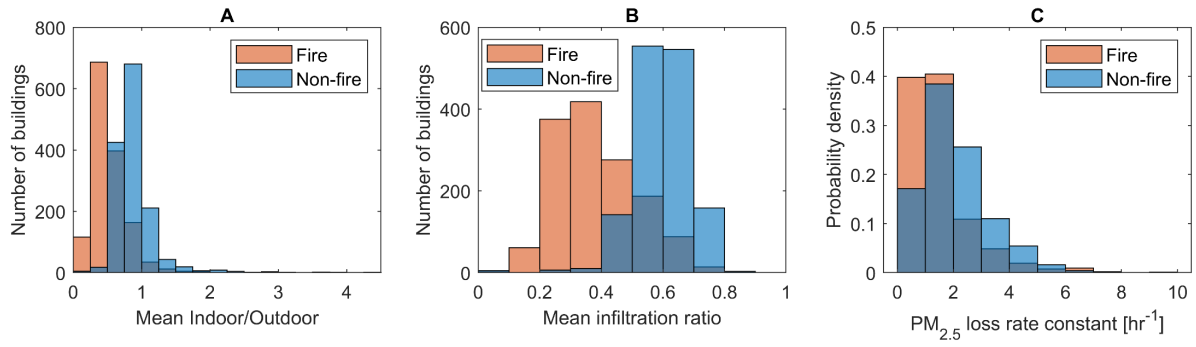

Fig. S16. Binary tree $PM_{2.5}$ correction case. **A.** Distribution of mean Indoor/Outdoor $PM_{2.5}$ ratio during fire days and non-fire days for the buildings; **B.** Distribution of Infiltrated/Outdoor $PM_{2.5}$ ratio during fire days and non-fire days for the buildings. **C.** Probability density distribution of total indoor particle loss rate constants of $PM_{2.5}$ for the NC 2020 case. This figure demonstrates the binary tree correction does not meaningfully affect the fire day/non-fire day comparison.

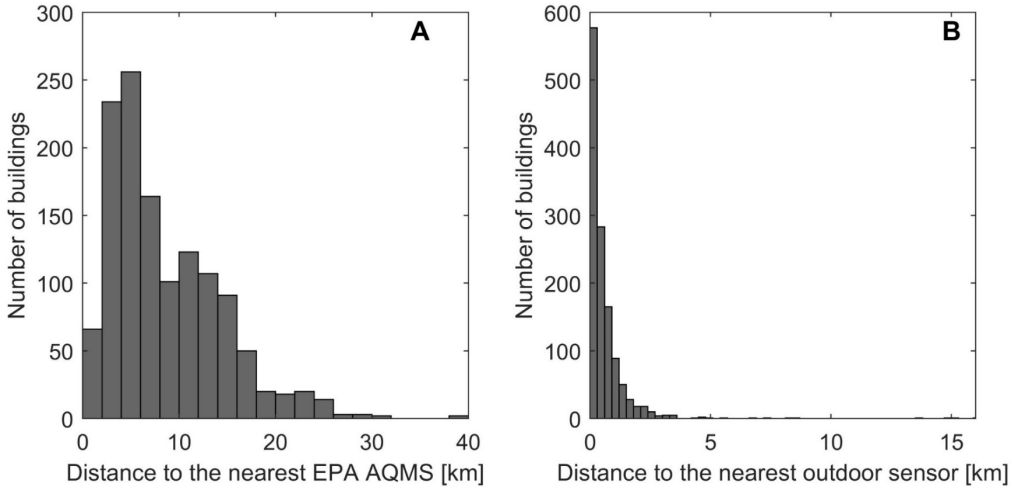

Fig. S17. Distribution of distance from the indoor sensor to **A.** the nearest EPA air quality measurement station and **B.** the nearest outdoor PurpleAir sensor in the NC 2020 case. The geometric mean (GM) distance from an indoor sensor to the nearest AQMS is 6.7 km, but it is only 0.21 km to the nearest outdoor PurpleAir sensor.

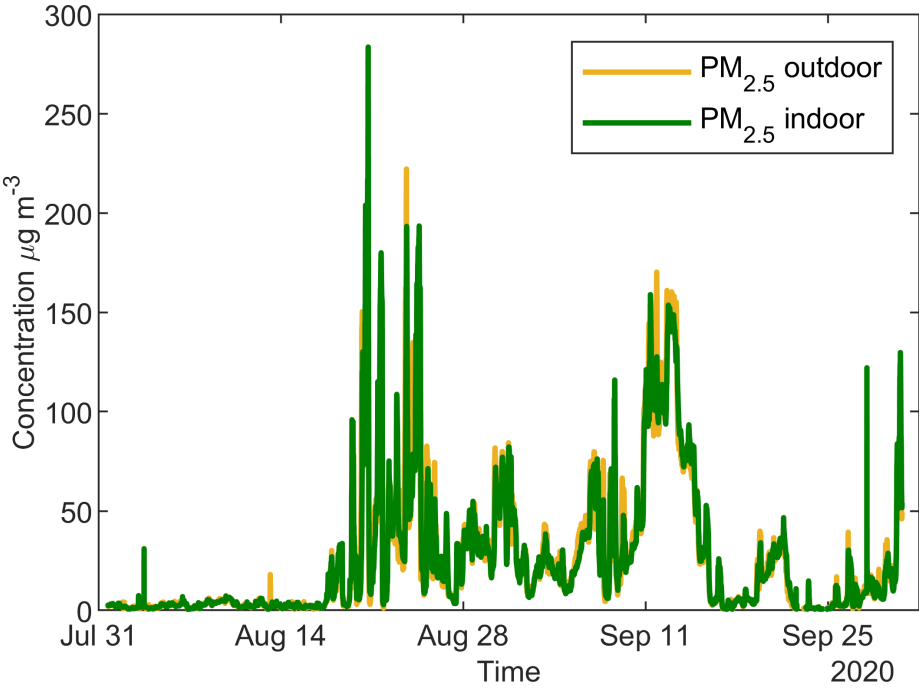

Fig. S18. Concentration timelines of $PM_{2.5}$ reported by an "indoor" sensor and the nearest outdoor sensor. Because the indoor concentration measured is too close to and too well correlated with the outdoor concentration, this sensor might be placed outdoors. This node was therefore not used in this analysis.

Table S1. Parameters and performance of 7 correction methods for the outdoor sensors in the San Francisco Bay Area in August and September 2020 (NC 2020 case). Parameters are for the correction equation $PM_{2.5,corrected} = \beta_0 + \beta_1 PM_{CF1} + \beta_2 RH$. RH is between 0 and 1.

	Linear regression with intercept	Linear regression no intercept	Linear regression no intercept (ODR)	Barkjohn *et al.*(9) US fire correction	Holder *et al.* (3) wildfire correction	New fit incorporating RH	Binary decision tree with RH
β_0 [μg m^{-3}]	3.52	n/a	n/a	5.60	-3.21	3.92	n/a
β_1	0.50	0.53	0.54	0.53	0.51	0.50	n/a
β_2	n/a	n/a	n/a	-0.084	n/a	-0.80	n/a
RMSE[a] [μg m^{-3}]	12.2	12.6	12.6	12.8	13.9	12.2	7.7
NRMSE[b]	0.48	0.50	0.50	0.51	0.55	0.48	0.39
Regression R^2	0.88	n/a	n/a	n/a	n/a	n/a	n/a
R^2 of calibrated data against EPA reference measurements	0.88	0.87	0.87	0.88	0.88	0.88	0.95

a. The root mean square error (RMSE, in [μg m^{-3}]) is calculated by

$$RMSE = \sqrt{\frac{1}{N}\sum_{h=1}^{N}(x_h - R_h)^2},$$

where N is the number of 1-hour $PM_{2.5}$ [μg m^{-3}] data points. x_h is hourly averaged sensor $PM_{2.5}$ concentration [μg m^{-3}] for hour h after correction. R_h is the hourly concentration of $PM_{2.5}$ [μg m^{-3}] measured by the EPA AQMS.

b. The root mean squared error normalized to the observed mean (NRMSE) is calculated by:

$$NRMSE = \frac{RMSE}{\overline{R_h}},$$

where R_h is the mean $PM_{2.5}$ [μg m^{-3}] observed by reference EPA AQMS.

Table S2. Corrections based on linear regression of EPA monitor $PM_{2.5}$ measurements with PurpleAir Sensors within certain distance in the San Francisco Bay Area in August and September 2020 (NC 2020 case). Number of sensors refer to the total number of sensors near EPA monitoring sites that meets the requirements described in the "Selection of correction method" section in the Supplement. At most 50 sensors near each EPA site were included. Parameters are for the correction equation.

Distance (km)	**2**	**5**	**10**	**20**
Number of sensors	104	442	624	750
		Intercept $\beta_0 \neq 0$		
β_0 [μg m^{-3}]	3.26	3.52	3.77	4.05
β_1	0.50	0.50	0.50	0.49
R^2	0.89	0.88	0.87	0.85
RMSE [μg m^{-3}]	11.5	12.2	12.7	13.5
NRMSE	0.46	0.49	0.51	0.53
		No intercept ($\beta_0 = 0$)		
β_1	0.53	0.53	0.51	0.52
R^2	n/a	n/a	n/a	n/a
RMSE [μg m^{-3}]	11.8	12.6	13.1	13.9
NRMSE	0.47	0.50	0.52	0.55

Table S3. Median prices of homes with PurpleAir sensors compared to Home Value Index in cities with at least 10 buildings with valid indoor sensors in the NC 2020 case, as of December 2020. Prices were rounded to nearest thousand.

	Median price of homes with PurpleAir sensors	Number of buildings with PurpleAir sensors	Zillow Home Value Index of that city	Price Difference [a]
Alameda	$1,143,000	18	$1,119,000	2%
Albany	$1,257,000	14	$1,170,000	7%
Atherton	$7,306,000	10	$6,579,000	11%
Belmont	$2,240,000	13	$1,902,000	18%
Berkeley	$1,616,000	93	$1,411,000	14%
Campbell	$1,344,000	11	$1,441,000	-7%
Davis	$666,000	19	$759,000	-12%
El Cerrito	$1,045,000	14	$ 1,006,000	4%
Emeryville	$822,000	11	$ 583,000	41%
Lafayette	$1,992,000	33	$ 1,499,000	33%
Los Altos	$3,653,000	35	$ 3,429,000	7%
Los Gatos	$2,264,000	22	$ 2,142,000	6%
Menlo Park	$3,645,000	32	$2,417,000	51%
Mill Valley	$1,911,000	18	$1,746,000	9%
Moraga	$1,886,000	11	$1,726,000	9%
Mountain View	$2,317,000	44	$1,851,000	25%
Oakland	$1,300,000	104	$851,000	53%
Orinda	$1,941,000	24	$2,292,000	-15%
Palo Alto	$3,594,000	47	$3,151,000	14%
Portola Valley	$3,523,000	17	$4,099,000	-14%
Redwood City	$2,196,000	35	$1,628,000	35%
Richmond	$807,000	12	$635,000	27%
Sacramento	$469,000	39	$400,000	17%
San Carlos	$2,248,000	16	$2,003,000	12%
San Francisco	$1,696,000	193	$1,400,000	21%
San Jose	$1,460,000	49	$1,141,000	28%
San Mateo	$1,919,000	28	$1,461,000	31%
San Rafael	$1,447,000	15	$1,214,000	19%
Santa Rosa	$713,000	21	$637,000	12%
Saratoga	$3,109,000	11	$3,053,000	2%
Sunnyvale	$1,654,000	31	$1,771,000	-7%
Walnut Creek	$1,414,000	21	$958,000	48%

[a]Price difference = (Median price of homes with PurpleAir sensors - Median City Home Value)/ Median City Home Value

Table S4. Mean ± standard deviation of fire-day infiltration ratios and the number of buildings with fire-day infiltration ratios below 0.14 or above 0.40 in cities with at least 10 buildings with valid indoor sensors in the NC 2020 case.

City	Number of buildings with PurpleAir sensors	Mean ± SD of Fire-day infiltration ratio	No. of Buildings with Fire-day infiltration ratio < 0.14	No. of Buildings with Fire-day infiltration ratio > 0.40
Alameda	18	0.19±0.09	5	1
Albany	14	0.31±0.11	0	3
Atherton	10	0.31±0.12	1	3
Belmont	13	0.27±0.08	1	1
Berkeley	93	0.27±0.10	10	13
Campbell	11	0.24±0.13	0	2
Davis	19	0.17±0.16	11	2
El Cerrito	14	0.27±0.10	2	1
Emeryville	11	0.26±0.16	3	1
Lafayette	33	0.23±0.11	5	4
Los Altos	35	0.33±0.19	7	12
Los Gatos	22	0.34±0.15	2	7
Menlo Park	32	0.27±0.12	3	5
Mill Valley	18	0.36±0.18	1	6
Moraga	11	0.16±0.10	4	0
Mountain View	44	0.25±0.14	12	5
Oakland	104	0.25±0.12	14	11
Orinda	24	0.24±0.19	11	5
Palo Alto	47	0.28±0.17	12	15
Portola Valley	17	0.27±0.12	2	3
Redwood City	35	0.30±0.15	8	11
Richmond	12	0.25±0.16	4	2
Sacramento	39	0.29±0.19	10	10
San Carlos	16	0.20±0.10	5	1
San Francisco	193	0.28±0.12	24	35
San Jose	49	0.26±0.14	14	10
San Mateo	28	0.26±0.11	3	2
San Rafael	15	0.31±0.16	3	5
Santa Rosa	21	0.31±0.15	4	6
Saratoga	11	0.30±0.18	0	2
Sunnyvale	31	0.22±0.15	8	7
Walnut Creek	21	0.21±0.14	9	1

Table S5. Weibull parameters of the concentration indoor/outdoor ratios for buildings with PurpleAir sensors in August-September 2020 in the San Francisco Bay Area (35 μg m^{-3} daily average $PM_{2.5}$ concentration measured at the nearest EPA measurement site was used as the threshold for fire days and non-fire days). *N* = 1274. Unhealthy days are defined as days with daily average EPA $PM_{2.5}$ concentration above 55.4 μg/m^3.

	Mean indoor conc μg m^{-3}		**Indoor/outdoor ratio**		**Infiltration ratio**	
	γ	*β*	γ	*β*	γ	*β*
Non-fire days	4.65	1.82	1.00	1.35	1.00	1.35
Fire days	12.4	1.50	0.45	1.26	0.30	2.00
Unhealthy days	14.9	1.40	0.34	1.19	0.26	1.74

Quantile-quantile plots (*SI Appendix,* Fig. S4) show the mean concentration of indoor $PM_{2.5}$ in all the buildings can be satisfactorily described by the Weibull distribution. The scale parameter and shape parameter of the Weibull fit are γ and *β,* respectively. The probability distribution function for *x* is $f(x) = \frac{\beta}{\gamma}(\frac{x}{\gamma})^{\beta-1}e^{-(x/\gamma)^{\beta}}$, where *x* > 0. Parameters of the SC 2020 and NC 2018 cases are not shown here due to the small sample sizes, which are less representative of all the buildings in these areas at that time.

SI References

1. T. Sayahi, A. Butterfield, K. E. Kelly, Long-term field evaluation of the Plantower PMS low-cost particulate matter sensors. *Environ. Pollut.* **245**, 932–940 (2019).
2. W. W. Delp, B. C. Singer, Wildfire smoke adjustment factors for low-cost and professional PM2.5 monitors with optical sensors. *Sensors* **20**, 1–21 (2020).
3. A. L. Holder, *et al.*, Field evaluation of low-cost particulate matter sensors for measuring wildfire smoke. *Sensors* **20**, 1–17 (2020).
4. K. Ardon-Dryer, Y. Dryer, J. N. Williams, N. Moghimi, Measurements of PM2.5 with PurpleAir under atmospheric conditions. *Atmos. Meas. Tech.* **13**, 5441–5458 (2020).
5. T. Zheng, *et al.*, Field evaluation of low-cost particulate matter sensors in high-and low-concentration environments. *Atmos. Meas. Tech.* **11**, 4823–4846 (2018).
6. K. K. Barkjohn, B. Gantt, A. L. Clements, Development and Application of a United States wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmos. Meas. Tech. Discuss.* (2020) https:/doi.org/10.5194/amt-2020-413 (January 5, 2021).
7. J. Bi, A. Wildani, H. H. Chang, Y. Liu, Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale. *Environ. Sci. Technol.* **54**, 2152–2162 (2020).
8. K. K. Barkjohn, *et al.*, Real-time measurements of PM2.5 and ozone to assess the effectiveness of residential indoor air filtration in Shanghai homes. *Indoor Air* **31**, 74–87 (2021).
9. K. K. Barkjohn, A. Holder, S. Frederick, G. Hagler, A. Clements, PurpleAir PM2.5 U.S. Correction and Performance During Smoke Events in *International Smoke Symposium*, (2020).
10. K. E. Kelly, *et al.*, Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environ. Pollut.* **221**, 491–500 (2017).
11. C. Malings, *et al.*, Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Sci. Technol.* **54**, 160–174 (2020).
12. M. Levy Zamora, *et al.*, Field and Laboratory Evaluations of the Low-Cost Plantower

Particulate Matter Sensor. *Environ. Sci. Technol.* **53**, 838–849 (2019).

13. C. Wu, J. Zhen Yu, Evaluation of linear regression techniques for atmospheric applications: The importance of appropriate weighting. *Atmos. Meas. Tech.* **11**, 1233–1250 (2018).
14. J. Bi, L. A. Wallace, J. A. Sarnat, Y. Liu, Characterizing outdoor infiltration and indoor contribution of PM2.5 with citizen-based low-cost monitoring data. *Environ. Pollut.* **276** (2021).
15. P. Tiitta, *et al.*, Measurements and modelling of PM2.5 concentrations near a major road in Kuopio, Finland. *Atmos. Environ.* **36**, 4057–4068 (2002).
16. T. Suvendrini Lena, V. Ochieng, M. Carter, J. Holguín-Veras, P. L. Kinney, Elemental carbon and PM2.5 levels in an urban community heavily impacted by truck traffic. *Environ. Health Perspect.* **110**, 1009–1015 (2002).
17. A. A. Karner, D. S. Eisinger, D. A. Niemeier, Near-roadway air quality: Synthesizing the findings from real-world data. *Environ. Sci. Technol.* **44**, 5334–5344 (2010).
18. J. S. Apte, *et al.*, High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environ. Sci. Technol.* **51**, 6999–7008 (2017).
19. R. Allen, T. Larson, L. Sheppard, L. Wallace, L. J. S. Liu, Use of real-time light scattering data to estimate the contribution of infiltrated and indoor-generated particles to indoor air. *Environ. Sci. Technol.* **37**, 3484–3492 (2003).
20. R. Allen, L. Wallace, T. Larson, L. Sheppard, L. J. S. Liu, Estimated hourly personal exposures to ambient and nonambient particulate matter among sensitive populations in seattle, washington. *J. Air Waste Manag. Assoc.* **54**, 1197–1211 (2004).
21. S. Patel, *et al.*, Indoor Particulate Matter during HOMEChem: Concentrations, Size Distributions, and Exposures. *Environ. Sci. Technol.* **54**, 7107–7116 (2020).
22. C. Chen, B. Zhao, Review of relationship between indoor and outdoor particles: I/O ratio, infiltration factor and penetration factor. *Atmos. Environ.* **45**, 275–288 (2011).
23. L. Wallace, R. Williams, Use of personal-indoor-outdoor sulfur concentrations to estimate the infiltration factor and outdoor exposure factor for individual homes and persons.

Environ. Sci. Technol. **39**, 1707–1714 (2005).

24. S. Bhangar, N. A. Mullen, S. V Hering, N. M. Kreisberg, W. W. Nazaroff, Ultrafine particle concentrations and exposures in seven residences in northern California. *Indoor Air* **21**, 132–144 (2011).

25. W. Ott, L. Wallace, D. Mage, Predicting particulate (PM10) personal exposure distributions using a random component superposition statistical model. *J. Air Waste Manag. Assoc.* **50**, 1390–1406 (2000).

26. B. Stephens, J. A. Siegel, Penetration of ambient submicron particles into single-family residences and associations with building characteristics. *Indoor Air* **22**, 501–513 (2012).

27. California Energy Commission, *California Building Climate Zone Areas* (2018) (January 26, 2021).