ACS ES&T Air
pubs.acs.org/estair
Article
Open Access
This article is licensed under CC-BY-NC-ND 4.0

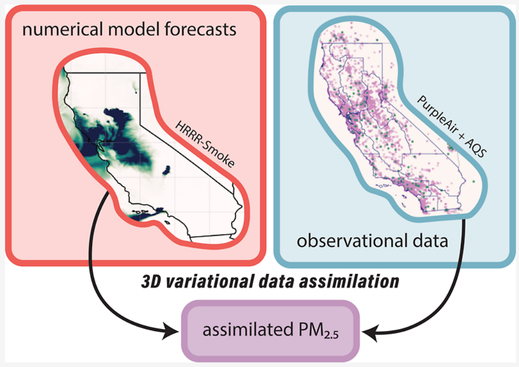
Hourly $PM_{2.5}$ Estimates across California from 2018 to 2023

Sienna R. White,* Rebecca A. Sugrue, Léo Guillotte, Eric James, Hongli Wang, Ravan Ahmadov, Neeta Thakur, and Fotini K. Chow

Cite This: ACS EST Air 2026, 3, 968–980
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This study presents a new data set of hourly $PM_{2.5}$ concentrations across California from 2018 to 2023 at a three-kilometer resolution. This data set was developed by assimilating observations from PurpleAir and the U.S. EPA Air Quality System monitors into wildfire smoke forecasts from the High-Resolution Rapid Refresh Smoke (HRRR-Smoke) model using the Gridpoint Statistical Interpolation (GSI) three-dimensional variational data assimilation framework. Archived forecasts of modeled wildfire smoke $PM_{2.5}$ from HRRR-Smoke create the background field for assimilation, which is then corrected using surface observations of total $PM_{2.5}$. The resulting reanalysis from GSI provides an estimate of total $PM_{2.5}$ that minimizes error from both the observational and the model data. Validation results indicate strong performance, with monthly R^2 values ranging from 0.73 to 0.91 across the six-year data set, comparable to other $PM_{2.5}$ data sets. Case studies are presented for three major fire events, the 2018 Camp Fire, 2019 Kincade Fire, and 2020 Lightning Complex Fires to demonstrate the data set's fidelity in resolving plume dynamics and local exposure patterns. Root-mean-squared error averaged over each month scales with average $PM_{2.5}$ concentrations, resulting in a low error under typical conditions but higher absolute errors during extreme smoke events. This is the first long-term, hourly $PM_{2.5}$ data set of its kind for California and enables the generation of subdaily exposure metrics, such as peak hourly concentrations, exceedance durations, and time-of-day exposure peaks. The novelty and strong validation of this data set make it a compelling resource for future studies on the impact and significance of subdaily $PM_{2.5}$ exposure.



KEYWORDS: *air pollution, $P_{M2.5}$, wildfire smoke, data assimilation, machine learning*

■ INTRODUCTION

Fine particulate matter ($PM_{2.5}$) poses a significant health risk to all populations. $PM_{2.5}$ can penetrate deep into the lungs, leading to local injury as well as entering the bloodstream, triggering inflammation, exacerbating chronic diseases, and contributing to premature mortality.[1,2] These impacts are especially severe for children, older adults, and individuals living in communities disproportionately burdened by air pollution.[3−7]

While $PM_{2.5}$ has historically been associated with combustion sources such as vehicles and industrial emissions, wildfires have become an increasingly dominant $PM_{2.5}$ contributor in the western United States, particularly in California.[8,9] Climate change has further intensified fire seasons,[10,11] leading to more frequent and severe smoke episodes, often affecting communities far from the fire source. Emerging evidence now suggests that exposure to wildfire smoke $PM_{2.5}$ may pose even greater health impacts than other $PM_{2.5}$ sources, especially in communities already facing high rates of respiratory disease and hospitalizations.[12,13] This paper seeks to expand the data available for health studies by creating an hourly $PM_{2.5}$ data set at a 3 km resolution across California from 2018 to 2023.

There have been significant efforts to estimate historical daily $PM_{2.5}$ levels in California for health exposure studies. Generalized additive models using observational and land-use data were successfully deployed by Li et al.,[14] Gao et al.,[15] and Di et al.[16] Other approaches include mixed-effect models,[17] Bayesian maximum entropy methods combining modeled and observational data,[18] and machine learning methods such as random forest,[19−22] gradient boosted trees,[23] neural networks,[21] and combined ensemble models.[24] The spatial scale of these data sets varies from the one-kilometer grid of Lee,[17] Zhang et al.,[22] and Cleland et al.[18] to the zip-code level of Aguilera et al.[24] While some studies, such as Park et al.,[21] have developed hourly $PM_{2.5}$ data over shorter time windows, all of the multiyear studies have focused on predicted daily averaged $PM_{2.5}$.

There is growing interest in using subdaily $PM_{2.5}$ data to better understand how acute exposure events affect health outcomes. Hourly data enable additional exposure assessments,

Received: September 22, 2025
Revised: February 13, 2026
Accepted: February 13, 2026
Published: February 21, 2026

Downloaded via ACS Test Institution on April 20, 2026 at 18:49:28 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications
© 2026 The Authors. Published by American Chemical Society

including time-of-exposure estimates and alternative metrics, such as maximum concentrations over rolling windows. These subdaily metrics can be combined with health data (typically available on a daily scale) to determine the impact of acute exposures (e.g., maximum daily concentration) on health outcomes (e.g., daily emergency department visits). Hourly data also allow investigations into the impact and efficacy of public health awareness and response. Consider, for example, analysis of acute health events or public response to changing $PM_{2.5}$ (e.g., event cancellation or keeping schoolchildren indoors), which all occur on a subdaily scale.

The primary $PM_{2.5}$ data available on subdaily time scales are ground-level observational, numerical model, or satellite (aerosol optical depth) data, all of which carry unique strengths and weaknesses. Satellite data provide column-integrated measurements and are limited in time or space resolution. We therefore center our discussion on numerical models and observational monitor data, which provide direct estimates of ground-level $PM_{2.5}$ concentrations.

The most accurate available observational network is the Federal Reference Method and Federal Equivalent Method Air Quality System, herein termed AQS, which is maintained by the Environmental Protection Agency (EPA). While the AQS network records hourly $PM_{2.5}$ concentrations, the spatial coverage of these monitoring sites is limited, making it difficult to estimate $PM_{2.5}$ concentrations in areas far from individual monitoring sites. There is also a denser $PM_{2.5}$ data set available with the rise of the PurpleAir network in the past decade.[25−28] PurpleAir monitors provide crowd-sourced, real-time air quality monitoring at a two min resolution through a network of low-cost devices that use laser particle counters to estimate $PM_{2.5}$ concentrations. These monitors provide $PM_{2.5}$ data in areas not covered by traditional monitoring networks such as AQS. After data correction,[29] PurpleAir data are quite accurate; however, the locations of PurpleAir monitors in California tend to be heavily linked to population density and relative financial wealth, limiting data availability in rural and less wealthy regions.[30,31]

Numerical weather prediction models can also provide hourly estimates of $PM_{2.5}$; however, there are limited archives of modeled $PM_{2.5}$ available at both high spatial and temporal resolutions. We focus our discussion on the National Oceanic and Atmospheric Administration's High-Resolution Rapid Refresh Smoke model (herein, HRRR-Smoke), which is the highest-resolution operational wildfire smoke model in the United States.[32−34] HRRR-Smoke predicts wildfire smoke $PM_{2.5}$ transport over the U.S. at a 3 km resolution and hourly time scale. NOAA has maintained an archive of modeled wildfire smoke $PM_{2.5}$ since 2018. To the authors' knowledge, there are no other archives that predict total or wildfire smoke $PM_{2.5}$ at a similar spatial and temporal resolution across our domain of study (specifically California, from 2018 to 2023).

High-fidelity estimates of $PM_{2.5}$ exposure require the benefits of both methods: the accuracy of observational data, in addition to the spatial resolution and coverage of numerical model results. Machine learning presents one methodology for leveraging both observational and model data to predict $PM_{2.5}$; however, many machine learning methods lack the ability to explicitly consider both measurement and model error covariance in their prediction. By contrast, data assimilation methods have been used successfully for decades in weather prediction to merge observational and model data with explicit consideration of both measurement error and systematic model bias.

Thus, to produce high-resolution, subdaily estimates of total historical $PM_{2.5}$, we apply three-dimensional variational data assimilation to merge numerical model estimates of wildfire smoke $PM_{2.5}$ from HRRR-Smoke with ground-level observations of total $PM_{2.5}$ from PurpleAir and AQS monitors. Though HRRR-Smoke predicts only wildfire smoke $PM_{2.5}$, it provides a first-order prediction of $PM_{2.5}$ during wildfire events, especially in areas not covered by surface monitors. This estimate is critical given wildfire smoke's large contribution to total $PM_{2.5}$.[22] When there are no active wildfires, the HRRR-Smoke model fields predict zero wildfire smoke $PM_{2.5}$; however, the assimilation of observed data into the numerical model field produces a nonzero estimate of total $PM_{2.5}$. In this sense, the assimilated total $PM_{2.5}$ during periods with no wildfire smoke represents a weighted interpolation of observational data across the state, while the assimilated total $PM_{2.5}$ during periods with wildfire smoke incorporates HRRR-Smoke predictions. The resulting data set presents a best-case, or lowest error, fusion of both observational and model data.

This is, to the authors' knowledge, the first multiyear estimate of total $PM_{2.5}$ across California that spans both a high spatial (three km) and temporal (hourly) resolution for long-term exposure estimates. Nevada and parts of other neighboring states are included in the data set, but because of sparse observational data in those areas, we focus on California. The data set covers a six-year period (from 2018 to 2023) and will enable ongoing health studies to investigate the impact of $PM_{2.5}$ exposure in California at a variety of temporal and spatial scales.

■ METHODS

Variational data assimilation integrates the observed data into numerical model predictions to create an improved estimate. To accommodate data availability for both PurpleAir and archived HRRR-Smoke data, our data set spans from January 1, 2018 through December 31, 2023 for a domain including all of California and Nevada. This six-year data set encompasses several major wildfire events in California, such as the Camp Fire in November 2018, the Kincade Fire in October 2019, and the severe fire seasons of 2020 and 2021.

Modeled and Observational Data

HRRR-Smoke. HRRR-Smoke is a three-dimensional weather model that provides hourly forecasts of wildfire smoke concentrations at a three-kilometer spatial resolution.[32] The model was developed by NOAA's Global Systems Laboratory in collaboration with the Cooperative Institute for Research in Environmental Sciences (CIRES) and Cooperative Institute for Research in the Atmosphere and is an extension of the HRRR weather forecast system.[33]

A smoke tracer was added to the HRRR framework in 2016. Data from instruments on polar-orbiting satellites (VIIRS and MODIS) are used to detect fires and estimate the fire radiative power.[35] Estimates of smoke emissions are then derived from fire radiative power, and a plume-rise algorithm predicts the vertical penetration of the estimated smoke plume.[36,37] Once injected, smoke acts primarily as a passive tracer in the model, with the exception of feedback with the model's atmospheric radiation scheme and wet and dry deposition. It is important to note that HRRR-Smoke does not assimilate any air quality observations, although there is advanced assimilation of meteorological observations during each hourly refresh. While this maintains accurate meteorology, the lack of operational air quality assimilation allows errors in the smoke field to persist during wildfire events.

The HRRR-Smoke model has been operational since December 2020, and forecast products are accessible through NOAA, as well as several sources outlined on the HRRR webpage. Like many numerical weather prediction models, errors in the surface $PM_{2.5}$ estimates are introduced by errors in observational forcing data such as satellite-predicted fire radiative power. Additional errors arise from model parametrizations, such as for plume rise.[38,39]

This work uses six years of hourly HRRR-Smoke forecasts (2018–2023) from the NOAA archives. The lowest-available forecast hour was used to represent each time, meaning that the zero forecast hour was used whenever available, followed by the 1 h forecast, or 2 h forecast, etc., up to 24 h. Higher forecast hours are associated with higher error, including potential errors from latency due to infrequent fire detection, which at some latitudes may occur only twice per day. For example, a 12 h forecast that starts at midnight (00:00) may not account for a fire that starts at 8AM (08:00). Meteorological fields are also generally more accurate for shorter forecast lengths.

The availability of the archived HRRR-Smoke data is shown in Figure 1. Continuous historical $PM_{2.5}$ data sets—especially at high temporal resolution—are rare, whether observed or modeled. HRRR-Smoke represents a unique archive of historical $PM_{2.5}$ estimates, with comparatively low data gaps compared to other data sets (such as remotely sensed $PM_{2.5}$ concentrations) and 24 h coverage. On average, 2019, 2020, 2022, and 2023 had the best data availability, with a low average forecast hour. There was relatively good coverage from 2019 onward, with over 96% of hours per year represented from 2019 to 2023. HRRR-Smoke was still under development until 2020. 2018 has both the lowest coverage (89% of hours in a year represented) and highest average forecast hour (six), indicating potentially higher error in the HRRR-Smoke estimates for that time period due to outages in the experimental forecasting system prior to operational implementation. In contrast, 2022 and 2023 have the highest data coverage, at 99%, with an average forecast hour close to zero.

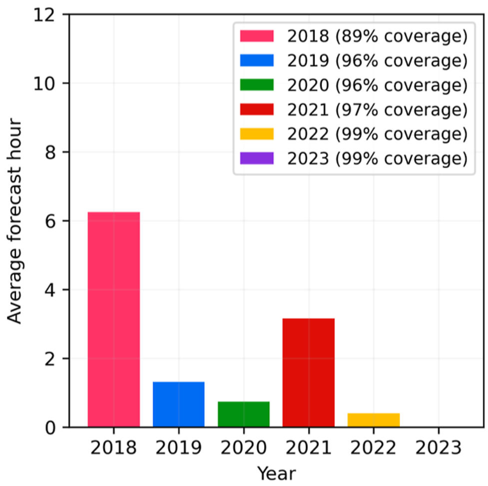


Figure 1. Average forecast hour of the HRRR-Smoke data used in the assimilation process (bar graph). Lower forecast hours are preferred. The percent of overall data availability is listed in the legend as a fraction of hours with data over the total available hours in a given year.

To prepare HRRR-Smoke data for assimilation, the archived GRIB files were converted to netcdf using the WRF Preprocessing System (consisting of ungrib, geogrid, and metgrid). The archived data were cropped to the domain shown in Figure 2, which centers on California.

Air Quality System (AQS) Data. The AQS from the EPA is a comprehensive repository of ambient air pollution data collected from hundreds of monitoring stations across the United States. These stations are managed by the EPA in combination with state, local, and tribal air pollution control agencies and provide data on pollutant concentrations and meteorological conditions. The data are subject to rigorous quality assurance and control measures to ensure high reliability.

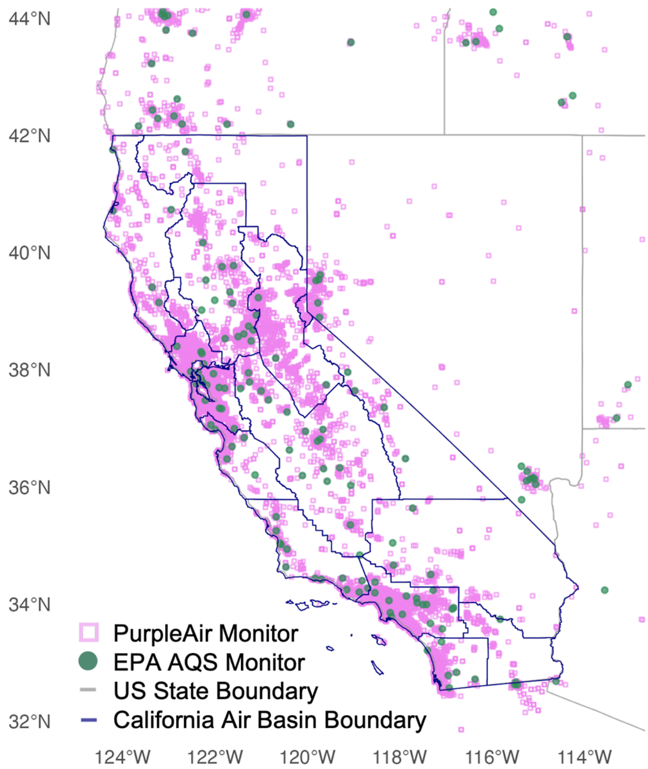


Figure 2. Map of monitoring sites from AQS (green circles) and PurpleAir (pink squares) that fall within the spatial domain of this study. State (gray) and California air basin boundaries (blue) are shown.

AQS data were obtained from the Air Quality System Data Mart,[40] which contains yearly files of hourly $PM_{2.5}$ data from all official monitoring sites across the US. Because a data quality procedure is in place for all AQS monitoring sites, no additional quality assurance steps are taken here. The data were limited to the spatial domain of interest before being used in data assimilation. AQS data covered January 1, 2018 through December 31, 2023.

PurpleAir Data. Hourly data were obtained from outdoor PurpleAir monitors from January 1, 2018 through December 31, 2023. These data were used for both data assimilation and data validation. The $PM_{2.5}$ CF1 and humidity data outputs from channels A and B were used in this study.

The sensors housed within a PurpleAir monitor are sensitive to fluctuations in ambient $PM_{2.5}$ pollution and humidity; therefore, quality assurance measures and correction factors need to be applied.[41−44] Following methods from Barkjohn et al., we implemented a quality assurance protocol to evaluate monitor consistency and ensure accurate data performance by the PurpleAir monitors. Note that our data are at an hourly time scale, whereas Barkjohn et al. used daily averaged data. The quality assurance protocol starts with evaluating the absolute and percent differences between dual-channel $PM_{2.5}$ CF1 measurements (channels A and B) for each monitor and hourly time point. We removed data if both channels were missing CF1 data or if the absolute difference between the channels exceeded 5 $\mu g/m^3$ and the percent difference exceeded 70% (slightly higher than Barkjohn's 61% threshold). Our approach also included a monthly check, in which we averaged the hourly channel A and B CF1 values over a month and excluded any data with an average difference greater than 25%. For the correction, all PurpleAir data were adjusted using the U.S.-wide and wildfire smoke calibration methods from Barkjohn et al.,[29,42] which adjusts the average CF1 PM2.5 value with the relative humidity levels. Lastly, corrected PurpleAir data were removed if they lacked at least six consecutive hours of data in a given month.

To benchmark PurpleAir monitor performance, for each month, every monitor was matched with its nearest AQS $PM_{2.5}$ monitor based on location. We used correlation coefficients (R^2) between the matched PurpleAir and AQS monitors to flag monitors if their mean monthly values substantially deviated (more than 50%) from the nearest AQS and the mean of all PurpleAir monitors matched to the same AQS monitor. Data were excluded if a flagged monitor also had an $R^2 \leq 0.2$ when compared to the nearest AQS monitor. Because some PurpleAir monitors may be a substantial distance from the nearest AQS monitor, data were excluded only if the values both deviated from the nearby PurpleAir monitors and if there was a low R^2 ($\leq$0.2). After following this quality assurance and data correction procedure, the hourly PurpleAir and AQS data have a high R^2 (0.81) with a slope of 1.02 for matched monitoring sites within 1 km of each other.

Validation Data and Nomenclature. To clearly differentiate data sets in the following discussion, we present the nomenclature for our validation in Table 1.

Table 1. Nomenclature Used in the Validation Process

observed $PM_{2.5}$ (PurpleAir)	observational corrected PurpleAir data. 95% of available PurpleAir data were included in the assimilation process
Observed $PM_{2.5}$ (AQS)	Observational EPA AQS data. All available data were included in the assimilation process
HRRR-Smoke	Modeled wildfire smoke $PM_{2.5}$ concentrations from the HRRR-Smoke model. All available data were included in the assimilation process
Assimilated $PM_{2.5}$	Estimated total $PM_{2.5}$. Produced by assimilating observed $PM_{2.5}$ from PurpleAir and AQS into HRRR-Smoke data
Validation data	The random 5% of corrected PurpleAir data withheld from the assimilation process. Each month has a different 5% withheld

Corrected PurpleAir observations are used for validation to ensure high spatial coverage in both the assimilation and validation data sets. Using AQS monitors for validation would limit both the number of validation sites (<30) and the spatial coverage of the data available for assimilation in an already sparse AQS network. Therefore, corrected PurpleAir sites are used. Each month, a random 5% of PurpleAir data was withheld from the assimilation process. On average, this ranged from 60 to 300 monitors depending on the time frame. Because the number of installed PurpleAir monitors has increased over time, later years in our study have more validation sites per month than in earlier years.

Gridpoint Statistical Interpolation (GSI)

We applied three-dimensional variational data assimilation (3D-Var) using Gridpoint Statistical Interpolation (GSI) program. GSI was deployed by the NOAA in 2006 as an operational tool. Originally designed to assimilate weather variables such as wind speed, humidity, and surface temperature, GSI has since been expanded to assimilate a host of other variables and observation data types, including surface measurements of $PM_{2.5}$ and PM_{10} for WRF-Chem.[45] In this work, GSI was used to assimilate surface $PM_{2.5}$ observations from AQS and PurpleAir into smoke fields simulated by HRRR-Smoke, following the approach set forth by Pagowski et al.[45]

3D-Var consists of solving for the "analysis field" (herein referred to as x) that minimizes a cost function ($J(x)$) by comparing a model result (often referred to as the "background field") with observational data, which include both PurpleAir and AQS data. The relevant terms involved in this process are summarized in Table 2. The cost function, $J(x)$, is given by

$$J(x) = \frac{1}{2}((x - x_b)^T B^{-1}(x - x_b) + (H(x) - y_0)^T R^{-1}(H(x) - y_0)) \quad (1)$$

Table 2. Definition of Terms Used in the 3D-Var Cost Function

term	definition
x	Analysis field (the output of the assimilation)
x_b	Background field (HRRR-Smoke model field)
$H(x)$	Analysis field projected into observation space
y_0	Observational data (PurpleAir and AQS)
B	Background error covariance matrix (symmetric)
R	Observation error covariance matrix (symmetric)

More details on the cost function, including alternative forms, can be found in Wu et al.[46] and Huang et al.[47] eq 1 can be interpreted as a weighted average between the difference of the analysis field to model estimate and the difference of the analysis field to observations. The first term in eq 1 is the weighted background error, which is the difference between the analysis field x and the background state x_b (HRRR-Smoke model field). The second term is the weighted observation error, which contains the difference between the set of observations y_0 and the converted analysis field $H(x)$. H represents the operator that projects the analysis field x into the observation space, which describes the associated units, location, and time of an observation. For single-point measurements, this involves interpolating model results in time and space to the location of an instantaneous monitor measurements. These analysis-to-background and analysis-to-observations costs are then weighted by the relative uncertainty in the observations and background field. GSI minimizes the cost function using a preconditioned conjugate gradient algorithm, which iterates until a convergence threshold or the maximum number of iterations is reached.[48]

The observational error covariance matrix R is a diagonal matrix as observation errors are assumed to be uncorrelated. The background error covariance matrix B, meanwhile, carries more complexity. This matrix represents the degree to which errors in the model are correlated. One key simplification made in this study is the use of a static, constant background error, following work by Wang et al.;[49] reasons for this are discussed below.

Case Studies for Calibration and Validation

To calibrate several key parameters in the assimilation process as well as characterize behavior in the resulting assimilated $PM_{2.5}$ field, we selected three distinct months with active wildfire events that occurred during the study period: November 2018 (the Camp Fire), October 2019 (the Kincade Fire), and September 2020 (the Creek, Bobcat, Slater, and Devil Fires). These fire months contained significant smoke plumes that affected communities across California. The fire perimeters,[50] PurpleAir monthly validation monitors, and NOAA's Hazard Mapping System (HMS) (herein referred to as HMS[51]) smoke plumes for each case study month are shown in Figure 3.

November 2018. The Camp Fire started on November 8, 2018 in Butte County and lasted 17 days. It was the most destructive fire in California history, destroying over 20,000 homes and structures[52] and resulting in 95 fatalities, with the vast majority of damage affecting the community of Paradise, CA. The smoke plume from the Camp Fire traveled across northern California, causing the San Francisco Bay Area and Central Valley to experience nearly 2 weeks of high $PM_{2.5}$ concentrations (averaging over 50 μg/m^3, with highs over 200 μg/m^3). During a similar time frame (November 8−16, 2018), the Woolsey Fire in Los Angeles and Ventura[53] produced smoke that affected nearby communities in Southern California.

October 2019. October 2019 was marked by the Kincade Fire, which was the largest fire of the 2019 California wildfire season. The Kincade Fire burned approximately 78,000 acres in Sonoma County in Northern California, including 374 structures, from October 23 to November 6, 2019. Smoke from the Kincade fire spurred a number of air quality warnings throughout the San Francisco Bay Area, from Sonoma to Santa Clara county.[54,55]

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968−980

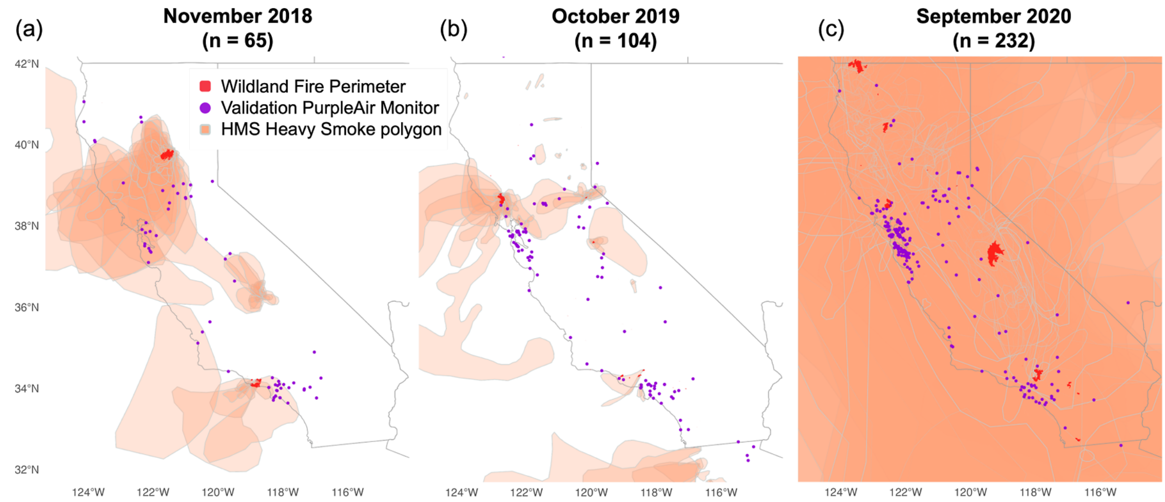


Figure 3. Map of validation PurpleAir monitors, wildfire fire perimeters, and heavy smoke polygons from NOAA's HMS for each of the three case studies months. The number of validation PurpleAir monitors is listed, which represents the 5% of available PurpleAir monitors that were withheld from the data assimilation that given month. The HMS smoke polygons represent "heavy" smoke events as determined by NOAA. The smoke polygons are shown for each day of each case study month.

September 2020. September 2020 was the largest wildfire year in California's recorded history and consisted of several major wildfire events that occurred simultaneously across the state. Among these were the Creek Fire (September 4 to December 24 in Fresno and Madera Counties in Central California), the Bobcat Fire (September 6 to November 27 in Los Angeles County), and the Slater and Devil Fires (September 7 to November 16 in Siskiyou and Del Norte Counties in Northern California). Collectively, these events burned approximately 661,000 acres across the state and caused over $653 million in damage.[56,57] The lingering air quality concerns about the health impact of the prolonged smoke exposure, which for some communities lasted nearly two months, prompted air quality management districts throughout the state to revisit air quality education and outreach.[58]

Parameter Selection for GSI

Four parameters were adjusted in GSI while the assimilation process was developed: the gross error, the maximum error (ermax), the minimum error (ermin), and the horizontal length scale (L_c), described below. The gross, maximum, and minimum errors are all used to evaluate whether or not an observation is included in the assimilation or withheld. GSI evaluates the ratio of the difference between the observed and background $PM_{2.5}$ values to the standard deviation of observation error as

$$R = \frac{(\mathrm{PM}_{\mathrm{obs}} - \mathrm{PM}_{\mathrm{bkg}})}{\max(\mathrm{ermin},\ \min(\mathrm{ermax},\ \mathrm{observed\ error}))} \quad (2)$$

If this ratio R is greater than the gross error, then the observation is withheld from the data assimilation process. Higher gross errors and higher maximum errors allow GSI to assimilate observational data with large differences from the background field, which may be critical to accurately adjust $PM_{2.5}$ predictions during wildfire smoke events. Note that the gross error for assimilating $PM_{2.5}$ is several orders of magnitude higher than the gross error for standard meteorological variables (such as wind and temperature). This is because errors in chemical transport models are not Gaussian and instead may be on the same order of magnitude as the predicted value itself.

GSI uses three recursive filters to model the horizontal scale of the background error covariance. These horizontal correlation length scales (L_c) are weighted to allow users to mimic non-Gaussian length scale behavior while retaining computational efficiency.[59] The resulting structure is incorporated into the background model error covariance matrix B (shown in eq 1). Larger values of L_c result in a larger radius for increments or adjustments to the background field. We selected values of L_c that provided a horizontal correlation of about 30 km, consistent with work from Wang et al.[49]

The first round of calibration focused on selecting the gross error, ermax, and ermin parameters. After testing a range of parameters, those that minimized RMSE for the November 2018, October 2019, and September 2020 study periods were selected. To distinguish performance between the two highest performing parameter sets, a 10-fold cross-validation was conducted across these three case study periods. The full PurpleAir observational data set was randomly split into 10 different segments, and GSI was run 10 times, withholding one data segment at a time for model validation. This process ensured that the chosen parameter set was not biased toward a specific set of calibration data. The assimilated $PM_{2.5}$ field was compared to validation data by bilinearly interpolating the assimilated $PM_{2.5}$ field to the coordinates of each PurpleAir validation site. Both the RMSE and correlation coefficients (R^2) were used to assess the assimilation performance. The resulting best-case parameter set (lowest RMSE and highest R^2) is shown in Table 3.

Table 3. Final Parameter Set Used in GSI

gross error	ermin	ermax	L_c
900	1.5	0.75	(0.05, 0.2, 0.2)

Sensitivity to Observational Data Coverage. Using PurpleAir data in the assimilation process in addition to AQS data increases the spatial coverage of observational data in California significantly. Only 2% of the geographical area of California is within 5 km of an AQS monitor, but over 27% of California is within 5 km of a PurpleAir monitor. To further explore the additional information content provided by the PurpleAir network, we compared assimilation results using just AQS data with results that also included PurpleAir for the November 2018 case study. This exercise demonstrates that the PurpleAir data (in addition to AQS data) are critical to accurately estimating $PM_{2.5}$ in California (shown in the Supporting Information, Figure S1). All other analyses in this study are for results using both PurpleAir and AQS data.

Assimilation Validation

We evaluated the performance of our assimilated $PM_{2.5}$ field using PurpleAir monitor validation data (which were excluded from the assimilation process). $PM_{2.5}$ data with values of over 500 $\mu g/m^3$ were withheld from error calculations to exclude outlying events. Events with $PM_{2.5}$ > 500 $\mu g/m^3$ represent less than 0.03% of the data, but concentrations during high $PM_{2.5}$ events (usually wildfire smoke

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968−980

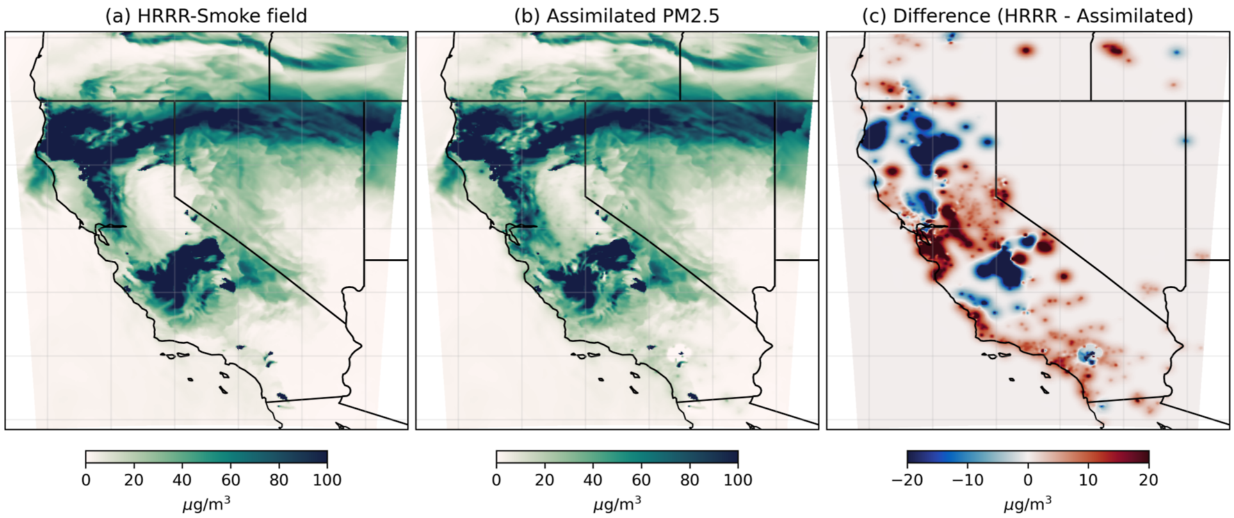


Figure 4. Spatial concentrations of (a) HRRR-Smoke $PM_{2.5}$, (b) assimilated $PM_{2.5}$, and (c) the respective difference on September 7, 2020 at midnight (00:00 UTC).

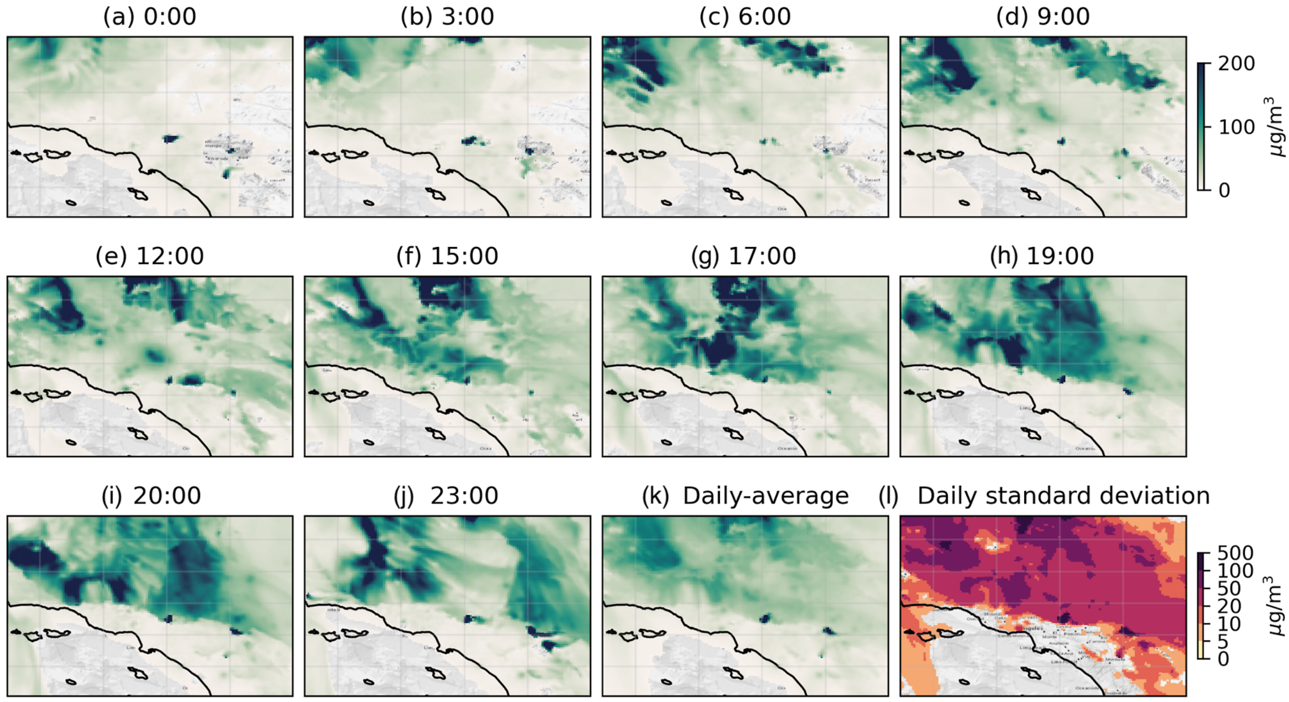


Figure 5. Smoke concentrations at 00:00, 3:00, 6:00, 9:00, 12:00, 15:00, 17:00, 19:00, 20:00, and 23:00 UTC on September 7, 2020 in panels (a–j), respectively. The daily average is shown in panel (k), and the standard deviation for that day is shown in (l).

events) can vary by orders of magnitude over narrow stretches of both time and space, biasing statistics. RMSE and R^2 values were calculated for all remaining validation PurpleAir $PM_{2.5}$ data, and model performance was determined over the entire spatial domain per month and per year. In addition, we assessed the performance of the assimilated $PM_{2.5}$ data during and outside of wildfire smoke events to see if our estimates of total $PM_{2.5}$ are consistently accurate. Smoke plumes from NOAA's HMS were used to determine if validation data fell within a smoke event in space and time. While HMS data are a commonly used proxy for determining the presence of smoke,[24,60] it can carry errors when categorizing an hourly data set, both from its coarser temporal frequency (daily) and the possibility that observed smoke is in fact aloft, and not at ground level.[61] Once validation data were delineated as "inside" or "outside" a smoke event period, performance statistics were calculated across the entire spatial domain.

RESULTS AND DISCUSSION

We used GSI to assimilate six years (2018–2023) of HRRR-Smoke, AQS, and corrected PurpleAir $PM_{2.5}$ data. GSI was adjusted using the finalized parameter set shown in Table 3. This effort used 32 CPUs on UC Berkeley's Savio high-performance computing cluster, with two CPUs assigned to each MPI task. The assimilation process for each time step required about 50 s of wall-clock time for a total computational cost of 19,456 CPU hours.

The validation of the resulting assimilated data set is discussed below. Given the size and scale of this hourly data set, we focus on performance during the three case study months, with performance statistics discussed broadly across all six years.

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968–980

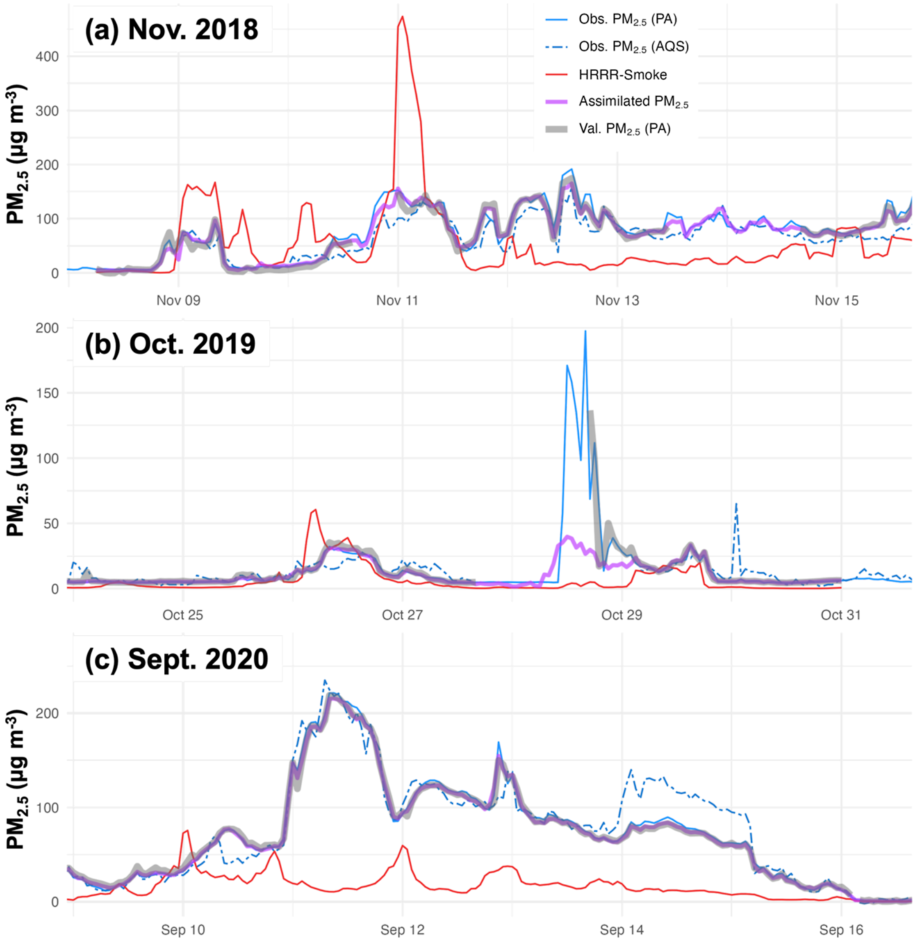


Figure 6. Time series of observational (both PA and AQS), HRRR-Smoke, assimilated, and validation $PM_{2.5}$ data at a location in Richmond, CA, for (a) November 2018, (b) October 2019, and (c) September 2020. The observational AQS and validation PA monitors are within 5 km of the observational PA monitor. The observational data were used by GSI to produce the assimilated $PM_{2.5}$ field. Note that there are some periods with missing data.

Assimilated $PM_{2.5}$ Output

A representative illustration of the spatial differences between the HRRR-Smoke model and the assimilated $PM_{2.5}$ field is shown in Figure 4. The map depicts September 7, 2020 at 00:00 UTC when several wildfires were burning across the state. An additional map of relative (or percentage) difference for this date is shown in Figure S4, and similar maps for November 15, 2018 are in Figures S2 and S3. Recall that while HRRR-Smoke predicts wildfire smoke $PM_{2.5}$, the assimilation process corrects this using observation data to produce an estimate of total $PM_{2.5}$. In Figure 4, the difference between panels (a) and (b) is shown in panel (c), highlighting the spatially heterogeneous changes to the background $PM_{2.5}$ field. Many of the positive increments are because HRRR-Smoke only predicts wildfire smoke concentrations, while our observational data captures total $PM_{2.5}$, including nonwildfire anthropogenic sources and secondary aerosol production. Thus, in areas with ambient nonwildfire $PM_{2.5}$ levels, the assimilation process results in a net increase of $PM_{2.5}$ compared to the HRRR-Smoke result, as the assimilated $PM_{2.5}$ represents total $PM_{2.5}$. For example, this occurs with industrial- or transportation-related air pollution in Southern California. On September 7, 2020, 00:00 (Figure 4), the assimilated $PM_{2.5}$ has corrected for HRRR-Smoke's overestimation of wildfire smoke concentrations in Northern California and for underestimation of $PM_{2.5}$ concentrations in the Bay Area and Los Angeles area, with most increments on the order of 0−20 $\mu g/m^3$. Meanwhile, the shift between HRRR-Smoke and the assimilated $PM_{2.5}$ on November 15, 2018 (Figure S2) includes large absolute (∼20 $\mu g/m^3$) increments within the smoke plume in Northern California.

We show the evolution of total $PM_{2.5}$ across the Los Angeles basin during a representative high pollution day on September 7, 2020. The greater basin includes both major urban areas as well as wildland−urban interface communities. There is clear time variability in the $PM_{2.5}$ concentration field across the entire basin, and the hourly data retain sharper spatial gradients and structure than the daily average. At 00:00 UTC, there are low $PM_{2.5}$ concentrations across the region. However, by 9:00, a heavy smoke plume moved in from the west. Pollution levels peak around 17:00−19:00. By 23:00, the smoke has partially cleared, though some regions still experience high $PM_{2.5}$ levels. The daily standard deviation is high throughout the Los Angeles basin (>50 $\mu g/m^3$) with the exception of several coastal communities (such as Long Beach and Anaheim), and the daily average looks significantly different from the hourly $PM_{2.5}$ levels. Figure 5 demonstrates that a daily average often cannot reflect the hourly scale variations in $PM_{2.5}$ concentrations, which depending on the time of day can be much higher or much lower than a daily

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968−980

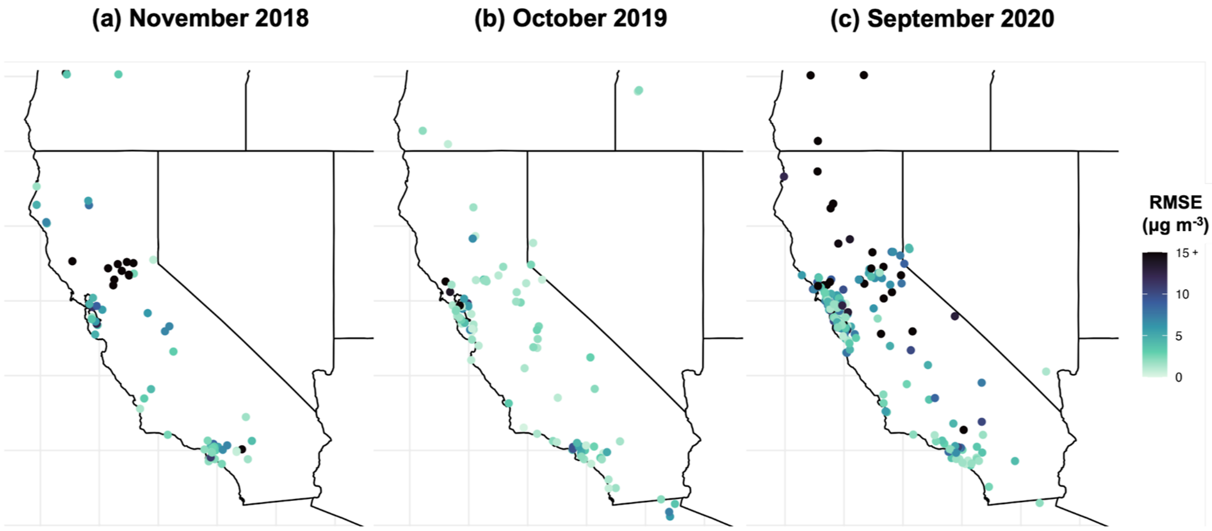


Figure 7. Root-mean-squared error (RMSE) for all validation sites during the three case studies. RMSE was calculated using the hourly data and then averaged over each month. RMSE is in units of $\mu g/m^3$.

Table 4. Root-Mean-Squared Error (RMSE) and R^2 Values for HRRR-Smoke and Assimilated $PM_{2.5}$ for Each Case Study (November 2018, October 2019, September 2020)[a]

Case Study	RMSE (HRRR-Smoke)	RMSE (Assimilated $PM_{2.5}$)	R^2 (HRRR-Smoke)	R^2 (Assimilated $PM_{2.5}$)
Nov-2018	35.7	10.7	0.17	0.90
Oct-2019	11.8	6.24	0.10	0.55
Sep-2020	37.5	9.22	0.25	0.94

[a]RMSE is in units of $\mu g/m^3$.

average. This high variability is common during wildfire smoke events and across wildland–urban interfaces, since smoke plumes can shift rapidly as a function of meteorology, swinging air quality from relatively good to highly polluted conditions over the course of a few hours. Figure 5 also demonstrates the extent to which local topography and meteorology can affect $PM_{2.5}$ levels over short spatial and temporal scales. Additional plots are shown in the Supporting Information, showing rapidly changing $PM_{2.5}$ levels in the greater Bay Area during representative days during the case study months (Figures S5 and S6).

Validation

Across the Case Studies. Assimilated $PM_{2.5}$ concentrations are first validated during the three case study periods. We present a time series of HRRR-Smoke, observational (PurpleAir and AQS), assimilated, and validation (PurpleAir) $PM_{2.5}$ data in Figure 6. These data come from a site in Richmond, CA, with a PurpleAir monitor that was active across all three case study time periods and used in the assimilation. This location was within 5 km of both an AQS monitor (also used in the assimilation) and a validation PurpleAir monitor (not used in the assimilation). Note that while the specific validation PurpleAir monitor varied month to month (due to the random selection of validation sites), there was consistently a validation site available within 5 km (see Figure S8).

As seen in Figure 6a, during November 2018, the assimilated $PM_{2.5}$ output diverges from the original HRRR-Smoke signal and has largely collapsed onto the observed data, capturing both the magnitude and temporal pattern of the observed and the validation $PM_{2.5}$ signal. The assimilated signal (purple line) tracks with the validation data (gray line), indicating excellent GSI model performance. October 2019, shown in Figure 6b, shows the assimilated $PM_{2.5}$ largely tracking both the observed and validation $PM_{2.5}$ signals, with a notable divergence during a high smoke event on October 28. The third case study, Figure 6c, September 2020, behaves similarly to that of November 2018, panel (a). Here, there is a strong validation performance, with the assimilated $PM_{2.5}$ closely tracking the validation signal. From September 14 to 15, the AQS monitor logs an increase in $PM_{2.5,}$ while the other PurpleAir monitors (observational and validation) stay relatively constant. The assimilated product, which has been interpolated to the site of the PurpleAir validation monitor, does not follow this increase shown in AQS data and instead closely aligns with the observational PurpleAir data from that time period, indicating that local variations in $PM_{2.5}$ are important even over shorter (5 km) scales and that these variations are captured by the data assimilation process. In all three case studies, the assimilated $PM_{2.5}$ field has high agreement with validation data, especially compared to the original HRRR-Smoke fields, which can differ from observed $PM_{2.5}$ by hundreds of $\mu g/m^3$ values during heavy smoke events.

We show a spatial plot of the RMSE at the validation sites for each case study in Figure 7. The areas with high RMSE differ across the three case studies, demonstrating that the specific characteristics of each smoke event (such as plume location and intensity) influence error behavior more than distance from assimilated observational data. Instead of seeing higher RMSE in areas far from population centers, where observational data are limited, error is instead highest closer to wildfire boundaries, where the smoke is most intense. For example, during November 2018 (Figure 7a), the highest RMSE is close to the Camp Fire itself (recall Figure 3, which depicts the fire perimeter). Similarly, in October 2019 (Figure

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968−980

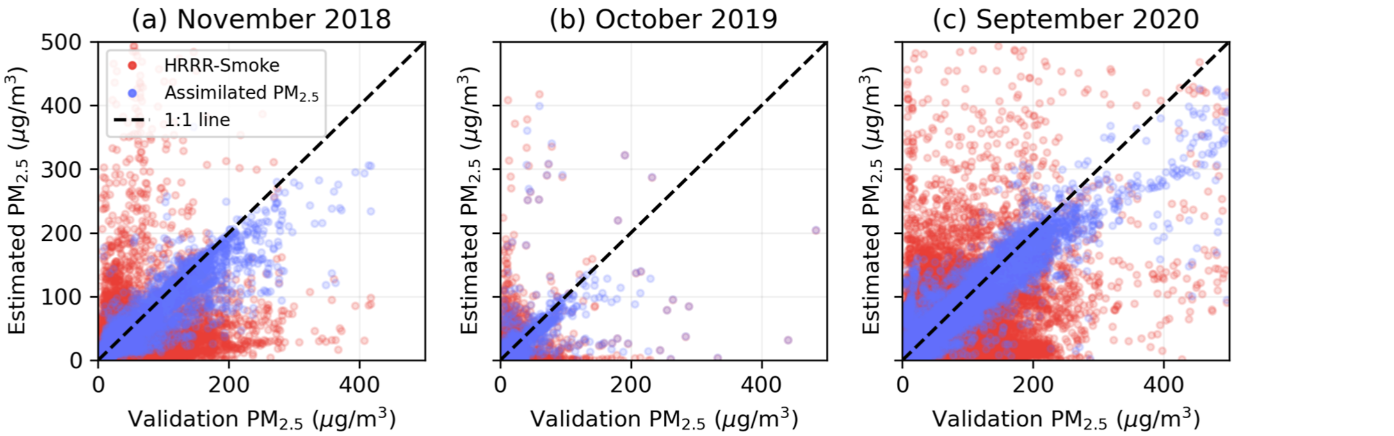


Figure 8. Scatter plots of HRRR-Smoke (red) and assimilated (blue) $PM_{2.5}$ estimates vs validation $PM_{2.5}$ from PurpleAir monitors for the three case studies: (a) November 2018, (b) October 2019, and (c) September 2020. The black dashed line represents the 1:1.

Table 5. Root-Mean-Squared Error (RMSE) and R^2 for HRRR-Smoke and Assimilated $PM_{2.5}$ During Fire and Nonfire Periods[a]

Event	RMSE (HRRR-Smoke)	RMSE (Assimilated $PM_{2.5}$)	R^2 (HRRR-Smoke)	R^2 (Assimilated $PM_{2.5}$)
Fire	12.2	3.6	0.21	0.76
Nonfire	8.1	2.5	0.10	0.75

[a]RMSE is in units of $\mu g/m^3$.

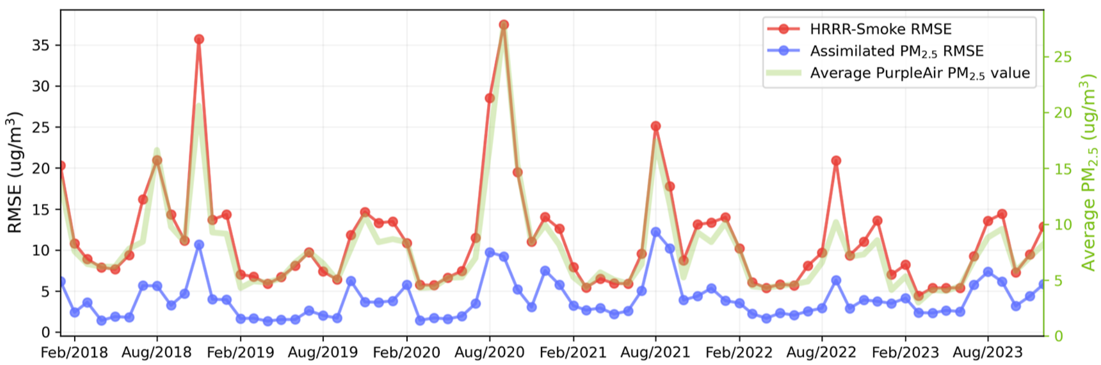


Figure 9. Time series of monthly RMSE (left-hand axis) compared to average $PM_{2.5}$ from the validation data set (right-hand axis).

7b), error is highest in Sonoma County and generally low (<5 $\mu g/m^3$) throughout the rest of the state. These patterns highlight the challenge of estimating $PM_{2.5}$ in areas experiencing sharp, rapidly evolving gradients and large $PM_{2.5}$ values associated with wildfire smoke.

We present error statistics in Table 4 and a scatter plot of model performance for both HRRR-Smoke and assimilated $PM_{2.5}$ in Figure 8. The assimilated output yields an overall improvement in RMSE and coefficient of determination R^2 when compared to the original $PM_{2.5}$ concentrations estimated from HRRR-Smoke. As seen in Figure 8, the correlation between validation $PM_{2.5}$ and the assimilated estimate is much stronger at lower $PM_{2.5}$ concentrations than at higher values. RMSE in November 2018 and September 2020 was reduced approximately 70−75%, while the RMSE reduction is lower for October 2019 at about 50%. The R^2 also increases significantly for all three case studies, with November 2018 and September 2020 having $R^2 \geq 0.90$, indicating excellent agreement. PurpleAir data are critical to this validation performance due to the high spatial coverage across California (as seen in Figure S1). The lower values for October 2019 are likely because wildfire smoke $PM_{2.5}$ contributed much less than ambient $PM_{2.5}$ to values seen across California, producing a weaker $PM_{2.5}$ signal (Figure 3).

Fire vs Nonfire. We evaluate fire and nonfire seasons separately over the six year analysis period. In the absence of wildfire smoke, ambient $PM_{2.5}$ values are typically far lower (e.g., 0−15 $\mu g/m^3$) than when wildfire smoke plumes are present, which can produce $PM_{2.5}$ values ranging from 50 to over 500 $\mu g/m^3$. Validation statistics for the assimilated $PM_{2.5}$ during both fire and nonfire periods are presented in Table 5.

The improvement in the assimilated $PM_{2.5}$ field compared to that in HRRR-Smoke during both fire and nonfire events is very similar, with about a 70% reduction in RMSE and a resulting correlation R^2 of approximately 0.75 for both periods. The RMSE is higher during the fire season than during the nonfire season (3.6 $\mu g/m^3$ compared to 2.5 $\mu g/m^3$, respectively).

Error in the assimilated $PM_{2.5}$ is highest during wildfire events. As seen in Table 5 and Figure 9, RMSE is smaller during nonfire time periods compared to time periods dominated by wildfire smoke. This difference can be attributed to several factors. The first is the simplest: RMSE tends to be smaller when $PM_{2.5}$ concentrations are lower, because the absolute differences between observed and modeled values are

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968−980

typically smaller. Wildfire smoke events are harder to model, and model errors can persist during the assimilation process. During wildfire smoke events, errors in HRRR-Smoke may result in plumes of high $PM_{2.5}$ that advect slightly off in time or space compared to true events,[39] resulting in observational monitors capturing high $PM_{2.5}$ in areas where HRRR-Smoke concentrations may be low or even zero. While the assimilation process effectively corrects for the phasing of these smoke events, it tends to underestimate the magnitude. The use of a static background error covariance field compounds these model errors. Future studies focusing on just wildfire events could reduce the error by generating case-dependent background error covariances. When considering events like the historic fire season of 2020, a case-dependent field could account for the high variance in $PM_{2.5}$ concentration across time and space. However, generating case-specific error covariances is impractical for a six year period with hundreds of different wildfire events, as it would require rerunning the HRRR smoke model[62] in order to customize the error covariance field to a daily resolution. This would significantly increase the computational cost of our approach.

Overall, high-$PM_{2.5}$ events tend to carry a higher RMSE in the assimilated $PM_{2.5}$, even if other performance metrics are strong. For example, note the relatively high RMSE in the assimilated $PM_{2.5}$ during November 2018 (10.7 $\mu g/m^3$) despite the excellent R^2 value (0.90). Lower performance at high $PM_{2.5}$ concentrations was also observed in the data set developed by Aguilera et al.[24] and demonstrates the difficulty in capturing both low-level and extreme $PM_{2.5}$ concentrations with the same method.

Summary across All Years. Finally, we review validation performance across the full six year data set spanning 2018–2023. Yearly performance statistics are provided in Table 6.

Table 6. Model Performance across 2018–2023[a]

Year	RMSE	R^2
2018	5.41	0.89
2019	3.08	0.83
2020	5.65	0.91
2021	6.07	0.79
2022	3.36	0.89
2023	4.52	0.73

[a]RMSE is in units of $\mu g/m^3$.

Note that additional tables with detailed yearly statistics, as well as scatter plots comparing the HRRR-Smoke and assimilated $PM_{2.5}$ estimates to validation data, are provided as Supporting Information.

A monthly time series of the RMSE is shown in Figure 9. The RMSE of the assimilated $PM_{2.5}$ data set is nearly entirely in phase with the average monthly $PM_{2.5}$ signal from the observed PurpleAir data across California. RMSE is lowest when $PM_{2.5}$ concentrations are low and rises during fire seasons when there are high $PM_{2.5}$ concentrations with strong spatial and temporal gradients. For 94% of the six year period, the RMSE remains below 10 $\mu g/m^3$. The highest RMSE occurs during the fire seasons in 2018, 2020, 2021, and 2022. The assimilation process substantially improves performance compared to HRRR-Smoke across all three case studies and the six year period. Compared to the original HRRR-Smoke model, which frequently misses the location or timing of high $PM_{2.5}$ during heavy smoke events, the assimilated product tends to capture both the magnitude and the phasing of observed pollution peaks. This improvement is quantified by a 61% reduction of average RMSE.

With hourly statistics, it is generally harder to get better agreement because of the need to capture temporal variability in the $PM_{2.5}$ signal. Nonetheless, across the six year data set, the average hourly RMSE is 4.13 $\mu g/m^3$ and the R^2 is 0.82 (see Table 6 for a yearly breakdown). This performance compares well to other $PM_{2.5}$ data sets from the literature, shown in Table 7. Some years within the assimilated $PM_{2.5}$ data set, such as 2020, exceed the R^2 values shown in Table 7. This is especially noteworthy given that our statistics are at hourly time resolution compared to daily data used by most other studies. Statistics from the hourly data set of Park et al.[21] also compare well.

Table 7. Model Statistics for Datasets of $PM_{2.5}$ in California[a]

Data set	Frequency	Time span	RMSE	R^2
Assimilated $PM_{2.5}$	**Hourly**	**2018–2023**	**4.13**	**0.82**
Di et al.[16]	Daily	2000–2015	4.04	0.80
Aguilera et al.[24]	Daily	2006–2020	4.37	0.83
Cleland et al.[18]	Daily	October 2017	n/a	0.71
Li et al.[14]	Daily	1990–2013	n/a	0.79
Reid et al.[63]	Daily	2008–2018	4.03	0.76
Park et al.[21]	Hourly	2023	6.14	0.80

[a]Values were calculated by the respective authors and span different time periods. RMSE and R^2 values are included when available for the portions of datasets spanning California. Statistics for our study's dataset are shown in bold.

Our approach has several limitations. First, due to the low density of regulatory monitors, we relied on low-cost PurpleAir data for validation. As a result, our validation is biased toward urban, higher-income areas, where PurpleAir monitors tend to be concentrated. Furthermore, the quality control and correction algorithms used here constrain, but do not fully eliminate, observational errors. There are also clear opportunities to improve the data assimilation process. In particular, fine-scale adjustments to the estimated background error covariance would strengthen the variational data assimilation performance during wildfire events. There is also promise in experimenting with appropriate observational error covariances for both AQS and PurpleAir monitors.

Nonetheless, the strong performance of the assimilated data set, with an average R^2 value of 0.82 across the six year period and an average RMSE of 4.13 $\mu g/m^3$, demonstrates that 3D-Var can produce analysis fields with reasonable accuracy across not only wildfire smoke events but also nonfire-related pollution periods. In combination with its coverage of six years of hourly $PM_{2.5}$ levels across California, this data set provides a valuable opportunity for health studies to examine the effects of subdaily $PM_{2.5}$ exposure on a range of outcomes. One area to explore further would be quantifying how much PurpleAir monitor data are required to adequately improve $PM_{2.5}$ estimates. Specifically, it would be useful to identify a subset of the monitoring network that maximizes the information value for GSI while minimizing the cost required to purchase PurpleAir data. Lastly, using this data set, future studies could investigate the efficacy of subdaily metrics for exposure estimates, as well as explore how local meteorology

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968–980

and climate variability affect the range of $PM_{2.5}$ that individuals may be exposed to on a subdaily scale.

ASSOCIATED CONTENT

Data Availability Statement

This study utilized the Gridpoint Statistical Interpolation (GSI) data assimilation system developed by NOAA, which is publicly available at: https://dtcenter.org/community-code/gridpoint-statistical-interpolation-gsi. The HRRR-Smoke model outputs used in this study are publicly available from the National Oceanic and Atmospheric Administration (NOAA) and can be accessed at https://rapidrefresh.noaa.gov/hrrr/. EPA Air Quality System (AQS) monitoring data are publicly available from the U.S. Environmental Protection Agency at: https://www.epa.gov/aqs/. PurpleAir data were obtained through a paid data access agreement and are subject to licensing restrictions; therefore, they are not publicly available. Researchers interested in accessing these data may contact PurpleAir directly. Data sets and analysis files generated during the current study are available upon request from the corresponding author (siennaw@berkeley.edu). Additional information on this research initiative can be found at https://clear.ucsf.edu/protect.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestair.5c00372.

Additional figures and discussion reviewing the before-and-after impacts of data assimilation; validating assimilated $PM_{2.5}$ with and without assimilated PurpleAir data; figures depicting time-varying $PM_{2.5}$ (including daily standard deviation) during October 26, 2019 and November 10, 2018; comparison of the assimilated $PM_{2.5}$ to validation data; and tables of monthly performance metrics from 2018 to 2023 (PDF)

AUTHOR INFORMATION

Corresponding Author

Sienna R. White – *Civil and Environmental Engineering, University of California, Berkeley, Berkeley, California 94720, United States;* orcid.org/0000-0002-9587-8662; Email: siennaw@berkeley.edu

Authors

Rebecca A. Sugrue – *Department of Medicine, University of California, San Francisco, San Francisco, California 94110, United States;* orcid.org/0000-0001-7886-7681

Léo Guillotte – *Civil and Environmental Engineering, University of California, Berkeley, Berkeley, California 94720, United States*

Eric James – *NOAA Global Systems Laboratory, Boulder, Colorado 80305, United States;* orcid.org/0000-0002-6507-4997

Hongli Wang – *Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309, United States; NOAA Global Systems Laboratory, Boulder, Colorado 80305, United States*

Ravan Ahmadov – *NOAA Global Systems Laboratory, Boulder, Colorado 80305, United States;* orcid.org/0000-0002-6996-7071

Neeta Thakur – *Department of Medicine, University of California, San Francisco, San Francisco, California 94110, United States*

Fotini K. Chow – *Civil and Environmental Engineering, University of California, Berkeley, Berkeley, California 94720, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestair.5c00372

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This project was supported by the Environmental Protection Agency (EPA) assistance agreement R8404810, NOAA cooperative agreement NA22OAR4320151, and the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0024386. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect those of the EPA, NOAA, or the Department of Commerce. Thank you to PurpleAir for their support and sponsorship of this work. Lastly, the authors thank Brian Blaylock for his invaluable work maintaining the Herbie HRRR-Smoke archives.

REFERENCES

(1) Xing, Y.-F.; Xu, Y.-H.; Shi, M.-H.; Lian, Y.-X. The impact of PM2.5 on the human respiratory system. *J. Thorac. Dis.* **2016**, *8*, No. E69.

(2) Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and health impacts of air pollution: a review. *Front. Public Health* **2020**, *8*, 14.

(3) Tessum, C. W.; Paolella, D. A.; Chambliss, S. E.; Apte, J. S.; Hill, J. D.; Marshall, J. D. PM2.5 polluters disproportionately and systemically affect people of color in the United States. *Sci. Adv.* **2021**, *7*, No. eabf4491.

(4) Clark, L. P.; Harris, M. H.; Apte, J. S.; Marshall, J. D. National and intraurban air pollution exposure disparity estimates in the United States: impact of data-aggregation spatial scale. *Environ. Sci. Technol. Lett.* **2022**, *9*, 786−791.

(5) Lane, H. M.; Morello-Frosch, R.; Marshall, J. D.; Apte, J. S. Historical redlining is associated with present-day air pollution disparities in US cities. *Environ. Sci. Technol. Lett.* **2022**, *9*, 345−350.

(6) Colmer, J.; Hardman, I.; Shimshack, J.; Voorheis, J. Disparities in PM2.5 air pollution in the United States. *Science* **2020**, *369*, 575−578.

(7) Jbaily, A.; Zhou, X.; Liu, J.; Lee, T.-H.; Kamareddine, L.; Verguet, S.; Dominici, F. Air pollution exposure disparities across US population and income groups. *Nature* **2022**, *601*, 228−233.

(8) Burke, M.; Childs, M. L.; de la Cuesta, B.; Qiu, M.; Li, J.; Gould, C. F.; Heft-Neal, S.; Wara, M. The contribution of wildfire to PM2.5 trends in the USA. *Nature* **2023**, *622*, 761−766.

(9) Ford, B.; Val Martin, M.; Zelasky, S.; Fischer, E.; Anenberg, S.; Heald, C. L.; Pierce, J. Future fire impacts on smoke concentrations, visibility, and health in the contiguous United States. *GeoHealth* **2018**, *2*, 229−247.

(10) Abatzoglou, J. T.; Williams, A. P. Impact of anthropogenic climate change on wildfire across western US forests. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 11770−11775.

(11) Burke, M.; Driscoll, A.; Heft-Neal, S.; Xue, J.; Burney, J.; Wara, M. The changing risk and burden of wildfire in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2011048118.

(12) Aguilera, R.; Corringham, T.; Gershunov, A.; Benmarhnia, T. Wildfire smoke impacts respiratory health more than fine particles

from other sources: observational evidence from Southern California. *Nat. Commun.* **2021**, *12*, 1493.

(13) Heaney, A.; Stowell, J. D.; Liu, J. C.; Basu, R.; Marlier, M.; Kinney, P. Impacts of fine particulate matter from wildfire smoke on respiratory and cardiovascular health in California. *GeoHealth* **2022**, *6*, No. e2021GH000578.

(14) Li, L.; Wu, A. H.; Cheng, I.; Chen, J.-C.; Wu, J. Spatiotemporal estimation of historical PM2.5 concentrations using PM10, meteorological variables, and spatial effect. *Atmos. Environ.* **2017**, *166*, 182−191.

(15) Gao, Z.; Ivey, C. E.; Blanchard, C. L.; Do, K.; Lee, S.-M.; Russell, A. G. Emissions, meteorological and climate impacts on PM2.5 levels in Southern California using a generalized additive model: Historic trends and future estimates. *Chemosphere* **2023**, *325*, 138385.

(16) Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M. B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. others An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environ. Int.* **2019**, *130*, 104909.

(17) Lee, H. J. Benefits of high resolution PM2.5 prediction using satellite MAIAC AOD and land use regression for exposure assessment: California examples. *Environ. Sci. Technol.* **2019**, *53*, 12774−12783.

(18) Cleland, S. E.; West, J. J.; Jia, Y.; Reid, S.; Raffuse, S.; O'Neill, S.; Serre, M. L. Estimating wildfire smoke concentrations during the October 2017 California fires through BME space/time data fusion of observed, modeled, and satellite-derived PM2.5. *Environ. Sci. Technol.* **2020**, *54*, 13439−13447.

(19) Vu, B. N.; Bi, J.; Wang, W.; Huff, A.; Kondragunta, S.; Liu, Y. Application of geostationary satellite and high-resolution meteorology data in estimating hourly PM2.5 levels during the Camp Fire episode in California. *Rem. Sens. Environ.* **2022**, *271*, 112890.

(20) Lu, Y.; Giuliano, G.; Habre, R. Estimating hourly PM2.5 concentrations at the neighborhood scale using a low-cost air sensor network: A Los Angeles case study. *Environ. Res.* **2021**, *195*, 110653.

(21) Park, S.; Sayeed, A.; Seo, J.; Henderson, B. H.; Naeger, A. R.; Gupta, P. Hour by Hour $PM_{2.5}$ Mapping Using Geostationary Satellites. *ACS ES&T Air* **2025**, *2*, 1816.

(22) Zhang, D.; Wang, W.; Xi, Y.; Bi, J.; Hang, Y.; Zhu, Q.; Pu, Q.; Chang, H.; Liu, Y. Wildland fires worsened population exposure to PM2. 5 pollution in the contiguous united states. *Environ. Sci. Technol.* **2023**, *57*, 19990−19998.

(23) Childs, M. L.; Li, J.; Wen, J.; Heft-Neal, S.; Driscoll, A.; Wang, S.; Gould, C. F.; Qiu, M.; Burney, J.; Burke, M. Daily local-level estimates of ambient wildfire smoke PM2. 5 for the contiguous US. *Environ. Sci. Technol.* **2022**, *56*, 13607−13621.

(24) Aguilera, R.; Luo, N.; Basu, R.; Wu, J.; Clemesha, R.; Gershunov, A.; Benmarhnia, T. A novel ensemble-based statistical approach to estimate daily wildfire-specific PM2. 5 in California (2006−2020). *Environ. Int.* **2023**, *171*, 107719.

(25) Raheja, G.; Harper, L.; Hoffman, A.; Gorby, Y.; Freese, L.; O'Leary, B.; Deron, N.; Smith, S.; Auch, T.; Goodwin, M.; et al. Community-based participatory research for low-cost air pollution monitoring in the wake of unconventional oil and gas development in the Ohio River Valley: Empowering impacted residents through community science. *Environ. Res. Lett.* **2022**, *17*, 065006.

(26) Durkin, A.; Gonzalez, R.; Isaksen, T. B.; Walker, E.; Errett, N. A. Establishing a community air monitoring network in a wildfire smoke-prone rural community: the motivations, experiences, challenges, and ideas of clean air methow's clean air ambassadors. *Int. J. Environ. Res. Public Health* **2020**, *17*, 8393.

(27) Lu, T.; Liu, Y.; Garcia, A.; Wang, M.; Li, Y.; Bravo-Villasenor, G.; Campos, K.; Xu, J.; Han, B. Leveraging citizen science and low-cost sensors to characterize air pollution exposure of disadvantaged communities in Southern California. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8777.

(28) Robinson, D. L.; Goodman, N.; Vardoulakis, S. Five years of accurate PM2. 5 measurements demonstrate the value of low-cost PurpleAir monitors in areas affected by woodsmoke. *Int. J. Environ. Res. Public Health* **2023**, *20*, 7127.

(29) Barkjohn, K. K.; Holder, A. L.; Frederick, S. G.; Clements, A. L. Correction and accuracy of PurpleAir PM2.5 measurements for extreme wildfire smoke. *Sensors* **2022**, *22*, 9669.

(30) Desouza, P.; Kinney, P. L. On the distribution of low-cost PM2.5 sensors in the US: demographic and air quality associations. *J. Exposure Sci. Environ. Epidemiol.* **2021**, *31*, 514−524.

(31) Sun, Y.; Mousavi, A.; Masri, S.; Wu, J. Socioeconomic disparities of low-cost air quality sensors in California, 2017−2020. *Am. J. Public Health* **2022**, *112*, 434−442.

(32) National Oceanic and Atmospheric Administration, *The High-Resolution Rapid Refresh Model*. https://rapidrefresh.noaa.gov/hrr/.

(33) Dowell, D. C.; Alexander, C. R.; James, E. P.; Weygandt, S. S.; Benjamin, S. G.; Manikin, G. S.; Blake, B. T.; Brown, J. M.; Olson, J. B.; Hu, M.; et al. The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather Forecast.* **2022**, *37*, 1371−1395.

(34) James, E. P.; Alexander, C. R.; Dowell, D. C.; Weygandt, S. S.; Benjamin, S. G.; Manikin, G. S.; Brown, J. M.; Olson, J. B.; Hu, M.; Smirnova, T. G.; et al. The High-Resolution Rapid Refresh (HRRR): an hourly updating convection-allowing forecast model. Part II: Forecast performance. *Weather Forecast.* **2022**, *37*, 1397−1417.

(35) Ahmadov, R.; Grell, G.; James, E. Using VIIRS fire radiative power data to simulate biomass burning emissions, plume rise, and smoke transport in a real-time air quality modeling system. In *Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp 2806−2808.

(36) Freitas, S. R.; Longo, K. M.; Chatfield, R.; Latham, D.; Silva Dias, M. A. F. d.; Andreae, M.; Prins, E.; Santos, J.; Gielow, R.; Carvalho, J., Jr Including the sub-grid scale plume rise of vegetation fires in low resolution atmospheric transport models. *Atmos. Chem. Phys.* **2007**, *7*, 3385−3398.

(37) Freitas, S.; Longo, K.; Trentmann, J.; Latham, D. Sensitivity of 1-D smoke plume rise models to the inclusion of environmental wind drag. *Atmos. Chem. Phys.* **2010**, *10*, 585−594.

(38) Benjamin, S.; Weygandt, S.; Brown, J.; Hu, M.; Alexander, C. R.; Smirnova, T. G.; Olson, J. B.; James, E. P.; Dowell, D. C.; Grell, G. A.; et al. A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Mon. Weather Rev.* **2016**, *144*, 1669−1694.

(39) Chow, F. K.; Yu, K. A.; Young, A.; James, E.; Grell, G. A.; Csiszar, I.; Tsidulko, M.; Freitas, S.; Pereira, G.; Giglio, L.; et al. High-resolution smoke forecasting for the 2018 Camp Fire in California. *Bull. Am. Meteorol. Soc.* **2022**, *103*, E1531−E1552.

(40) *US Environmental Protection Agency*. accessed October 2022.

(41) Tryner, J.; L'Orange, C.; Mehaffy, J.; Miller-Lionberg, D.; Hofstetter, J. C.; Wilson, A.; Volckens, J. Laboratory evaluation of low-cost PurpleAir PM monitors and in-field correction using co-located portable filter samplers. *Atmos. Environ.* **2020**, *220*, 117067.

(42) Barkjohn, K. K.; Gantt, B.; Clements, A. L. Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmos. Meas. Tech.* **2021**, *14*, 4617−4637.

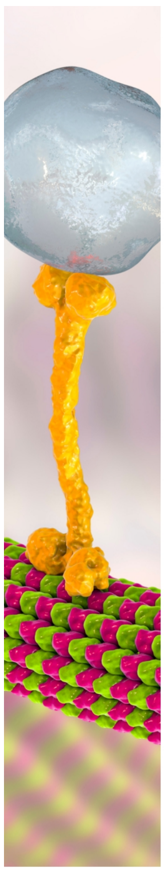
(43) Jaffe, D.; Miller, C.; Thompson, K.; Nelson, M.; Finley, B.; Ouimette, J.; Andrews, E. An evaluation of the U.S. EPA's correction equation for Purple Air Sensor data in smoke, dust and wintertime urban pollution events. *Atmos. Meas. Tech. Discuss.* **2023**, *16*, 1311.

(44) Wallace, L.; Bi, J.; Ott, W. R.; Sarnat, J.; Liu, Y. Calibration of low-cost PurpleAir outdoor monitors using an improved method of calculating PM2. 5. *Atmos. Environ.* **2021**, *256*, 118432.

(45) Pagowski, M.; Liu, Z.; Grell, G.; Hu, M.; Lin, H.-C.; Schwartz, C. Implementation of aerosol assimilation in Gridpoint Statistical Interpolation (v. 3.2) and WRF-Chem (v. 3.4. 1). *Geosci. Model Dev.* **2014**, *7*, 1621−1627.

(46) Wu, W.-S.; Purser, R. J.; Parrish, D. F. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Weather Rev.* **2002**, *130*, 2905−2916.

(47) Huang, X.-Y.; Xiao, Q.; Barker, D. M.; Zhang, X.; Michalakes, J.; Huang, W.; Henderson, T.; Bray, J.; Chen, Y.; Ma, Z.; et al. others Four-dimensional variational data assimilation for WRF: Formulation and preliminary results. *Mon. Weather Rev.* **2009**, *137*, 299−314.
(48) Derber, J.; Rosati, A. A global oceanic data assimilation system. *J. Phys. Oceanogr.* **1989**, *19*, 1333−1347.
(49) Wang, H.; Weygandt, S.; Ahmadov, R.; Li, R.; Romero-Alvarez, J.; Li, H.; Grell, G.; Pagowski, M.; Tang, Y.; Martin, C.; McQueen, J.; Farooqui, M. *Assimilation of Surface Particulate Matter Observations in the Experimental Rapid Refresh Forecast System Coupled with Smoke and Dust Model*; CU/CIRES Rendezvous, 2023. https://insidecires.colorado.edu/rendezvous/uploads/Rendezvous_2023_7732_1683821981.pdf.
(50) Fire; Program, R. A. *Historical Fire Perimeters*. https://www.fire.ca.gov/what-we-do/fire-resource-assessment-program/fire-perimeters, accessed: 2025−07.
(51) Oceanic, N.; Administration, A. *Hazard Mapping System Fire and Smoke Product*. https://www.ospo.noaa.gov/products/land/hms.html#data, accessed: 2025−07.
(52) Wallingford, N. *Camp Incident Damage Inspection Report: CABTU 016737*, 2018.
(53) Smith, F. P. C.; Leidos, I. *Evaluation of the Effects of the Woolsey Fire on the Analysis in the Final Environmental Impact: Statement for Remediation of Area IV and the Northern Buffer Zone of the Santa Susana Field Laboratory*, 2019.
(54) Reyes-Velarde, A. *Smoke from the Kincade Fire is Spreading Across the Bay Area, Bringing Unhealthful Air Quality*; Los Angeles Times, 2019.
(55) Galbraith, K.; Hernandez, L. *Bay Area Air Quality: Kincade Fire Smoke Could be Blown from Region*; San Francisco Chronicle, 2019.
(56) Cart, J. *California's 2020 Fire Siege: Wildfires by the Numbers*; Cal Matters, 2021.
(57) Krishnakumar, P.; Kannan, S. *The Worst Fire Season Ever. Again*; Los Angeles Times, 2020.
(58) Sacks, B. *California's 2020 Smoke Storm was Horrific. what Did the State Learn?*; Washington Post, 2023.
(59) Purser, R. J.; Wu, W.-S.; Parrish, D. F.; Roberts, N. M. Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. *Mon. Weather Rev.* **2003**, *131*, 1524−1535.
(60) Brey, S. J.; Ruminski, M.; Atwood, S. A.; Fischer, E. V. Connecting smoke plumes to sources using Hazard Mapping System (HMS) smoke and fire location data over North America. *Atmos. Chem. Phys.* **2018**, *18*, 1745−1761.
(61) Liu, T.; Panday, F. M.; Caine, M. C.; Kelp, M.; Pendergrass, D. C.; Mickley, L. J.; Ellicott, E. A.; Marlier, M. E.; Ahmadov, R.; James, E. P. Is the smoke aloft? Caveats regarding the use of the Hazard Mapping System (HMS) smoke product as a proxy for surface smoke presence across the United States. *Int. J. Wildland Fire* **2024**, *33*..
(62) Wang, H.; Huang, X.-Y.; Sun, J.; Xu, D.; Zhang, M.; Fan, S.; Zhong, J. Inhomogeneous background error modeling for WRF-Var using the NMC method. *J. Appl. Meteorol. Climatol.* **2014**, *53*, 2287−2309.
(63) Reid, C. E.; Considine, E. M.; Maestas, M. M.; Li, G. Daily PM2. 5 concentration estimates by county, ZIP code, and census tract in 11 western states 2008−2018. *Sci. Data* **2021**, *8*, 112.



CAS BIOFINDER DISCOVERY PLATFORM™
BRIDGE BIOLOGY AND CHEMISTRY FOR FASTER ANSWERS
Analyze target relationships, compound effects, and disease pathways
Explore the platform
CAS
A division of the American Chemical Society

https://doi.org/10.1021/acsestair.5c00372
ACS EST Air 2026, 3, 968−980