remote sensing

MDPI

Article

A Data-Fusion Approach to Assessing the Contribution of Wildland Fire Smoke to Fine Particulate Matter in California

Hongjian Yang [1,*], Sofia Ruiz-Suarez [2,3], Brian J. Reich [1], Yawen Guan [4] and Ana G. Rappold [5]

[1] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA; bjreich@ncsu.edu
[2] INIBIOMA-CONICET, National University of Comahue, Bariloche R8400, Rio Negro, Argentina; sofia.ruizsuarez@utoronto.ca
[3] Department of Statistical Science, University of Toronto, Toronto, ON M5R OA3, Canada
[4] Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA; yawen.guan@colostate.edu
[5] US Environmental Protection Agency, Durham, NC 27709, USA; rappold.ana@epa.gov
[*] Correspondence: hyang23@ncsu.edu

Abstract: The escalating frequency and severity of global wildfires necessitate an in-depth understanding and monitoring of wildfire smoke impacts, specifically its contribution to fine particulate matter ($PM_{2.5}$). We propose a data-fusion method to study wildfire contribution to $PM_{2.5}$ using satellite-derived smoke plume indicators and $PM_{2.5}$ monitoring data. Our study incorporates two types of monitoring data, the high-quality but sparse Air Quality System (AQS) stations and the abundant but less accurate PurpleAir (PA) sensors that are gaining popularity among citizen scientists. We propose a multi-resolution spatiotemporal model specified in the spectral domain to calibrate the PA sensors against accurate AQS measurements, and leverage the two networks to estimate wildfire contribution to $PM_{2.5}$ in California in 2020 and 2021. A Bayesian approach is taken to incorporate all uncertainties and our prior intuition that the dependence between networks, as well as the accuracy of PA network, vary by frequency. We find that 1% to 3% increase in $PM_{2.5}$ concentration due to wildfire smoke, and that leveraging PA sensors improves accuracy.

Keywords: Bayesian analysis; calibration; citizen science; spatiotemporal methods; spectral analysis

Citation: Yang, H.; Ruiz-Suarez, S.; Reich, B.J.; Guan, Y.; Rappold, A.G. A Data-Fusion Approach to Assessing the Contribution of Wildland Fire Smoke to Fine Particulate Matter in California. *Remote Sens.* **2023**, *15*, 4246. https://doi.org/10.3390/rs15174246

Academic Editors: Amy Braverman, Meredith Franklin and Emily Kang

Received: 30 June 2023
Revised: 14 August 2023
Accepted: 21 August 2023
Published: 29 August 2023

1. Introduction

Airborne particles are a serious environmental health risk globally, contributing in excess of 7 million premature deaths each year [1]. Fine particulate matter ($PM_{2.5}$, particles with a diameter of less than 2.5 micrometers) has been causally linked to cardiovascular morbidity and mortality [2] and are therefore regulated under the provisions of the Clean Air Act [3] to protect human health and wellbeing. As a result, the emissions of $PM_{2.5}$ from many antropogenic sources, such as transpiration and industry, have been on a steady decline [4] and wildfires have become the single largest source [5], potentially off setting reduction in emissions from other sources.

High concentrations of fine particles and gasses found in smoke have also produced alarming impacts on health [6,7]. During peak wildfire seasons, smoke exposure can exacerbate health problems, causing a spike in emergency department visits [8]. In an epidemiological study of health impacts by Thilakaratne et al., they estimated that 2.2% of annual respiratory health burden, or 92 ED visits per 100,000 people, is attributed to ambient particulate matter and that wildfire days account for over 15% of that burden [9]. However, providing a definite answer as to how much of particle pollution can be attributed to wildfires remains a challenging problem because instruments measure a total ambient concentration which is composed of natural, anthropogenic, and wildfire sources.

Previous research [10–12] has studied the contribution of wildfires on $PM_{2.5}$ concentrations by integrating remote sensing data on the location and extend of smoke plumes and $PM_{2.5}$ readings from Air Quality System (AQS) monitors deployed by the Environmental Protection Agency (EPA). These studies revealed that wildfires contribute to 40% of unhealthy days and substantially increase $PM_{2.5}$ concentrations [13,14]. Wildfire smoke impacts are dynamic and often affect areas without a monitoring station, as AQS monitors have limited spatial coverage due to the high cost and difficulty in installation. It is important to make air quality information available to the public quickly during wildfires, therefore AQS alone provides insufficient data source for monitoring wildfire emissions.

The increased incidence of days with poor air quality due to wildfires has created a demand and public interest for monitoring particulate pollution. Perhaps the most prevalent sensors are PurpleAir (PA), which are installed by members of the public, providing a real-time (every two minutes) monitoring of $PM_{2.5}$ with extensive spatial coverage [15]. However, it is known that PA sensors are less reliable compared to the AQS, and thus correction to the sensor readings is needed [16,17]. Barkjohn et al. developed a correction equation using meteorological conditions including relative humidity and temperature, as both measurements affect the accuracy of the instrument [15]; however, this calibration is developed for a US-wide correction and without smoke impacts. Another simple linear correction model under smoke impacted conditions was proposed by Holder et al. in [18]. As the sensor performance can be affected by geographic and environmental conditions, it is more reasonable to relax the assumption of a constant spatially varying bias, but rather capture the spatiotemporally varying bias.

Previous studies have either separated anthropogenic $PM_{2.5}$ from smoke emissions using chemical transport models or by subtracting out historically observed averages [19]. However, neither approach provides a definite answer as to how much of particle pollution can be attributed to wildfires. Data fusion is a widely used method that integrates information from different types of sensors to provide a robust and complete description of a process of interest [20,21]. It has been used extensively to estimate spatially and temporally resolved air quality surfaces. For example, Reich et al. [22], Warren et al. [23], and Friberg et al. [24,25] use data fusion method to study the complex relationship between monitoring data and outputs from Community Multi-Scale Air Quality (CMAQ), a deterministic chemical transport model. Nguyen et al. [26] combines observations from two noisy datasets to predict the true aerosol process. More recently, several researchers have exploited the usefulness of low-cost sensors such as Purple Air to map air quality and quantify the uncertainty of estimation [27–29]. Other spatiotemporal data fusion methods include Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) [30] and ST-Cokriging [31]. STARFM fuses spatial information from fine-resolution imagery and temporal information from coarse-resolution imagery. ST-Cokriging uses cross-variograms for prediction by assigning weights to observations from different sources. These methods cannot be directly applied to our analysis as both are more suitable for prediction than quantifying the contribution from wildland fires. ST-Cokriging uses numerical approach to solve for weight parameters, where it assigns weights to all nearby observations in a period. Similarly, STARFM employs a sliding window to assign weights to observations in a searching window, where the weights are determined by spectral, temporal, and location differences. In our case, it would be difficult to determine the spectral and location differences due to spatially misaligned AQS and PA readings. Most similar to our approach is Stein et al. [32], who also use a spectral transformation in time and spatial processes to capture dependence between stations for a single fixed monitoring network. We extend this approach to handle multiple data networks.

This study aims to provide an estimate of wildfire contribution on air quality in California by supplementing the remotely sensed smoke plume indicators with PA data. We propose a multi-resolution Bayesian approach fusing information from both AQS monitors and PA sensors to estimate the contribution to $PM_{2.5}$ caused by wildfires. We apply a Discrete Fourier Transform (DFT) to account for temporal correlation, transforming the data from the time domain to the frequency domain, and model the spatial correlation in the frequency domain. To quantify the relative increase in $PM_{2.5}$ concentrations due to wildfires, we propose regression and matching estimators, as discussed in Section 2.3. Our findings will not only enhance understanding of the relationship between wildfires and air pollution but also inform policy and decision-making related to wildfire management, public health, and climate change impacts.

2. Materials and Methods

2.1. Data Sources and Exploratory Analysis

Our analysis incorporates data from three distinct sources: satellite-derived smoke plume indicators obtained through the National Oceanic and Atmospheric Administration's Hazard Mapping System (HMS), $PM_{2.5}$ measurements from AQS monitoring stations, and $PM_{2.5}$ readings from PA monitoring stations. Figure 1 shows all three data sources for 20 September 2021 in California. We collect hourly data and average them to daily level from each source for 2020 and 2021 fire seasons, spanning 1 July to 31 October. We selected California because of its susceptibility to wildland fires, and 2020 and 2021 because these years have sufficient PA monitors. The original $PM_{2.5}$ readings from both AQS and PA stations are right-skewed and likely heteroskedastic so we apply the log transformation to all $PM_{2.5}$ readings.

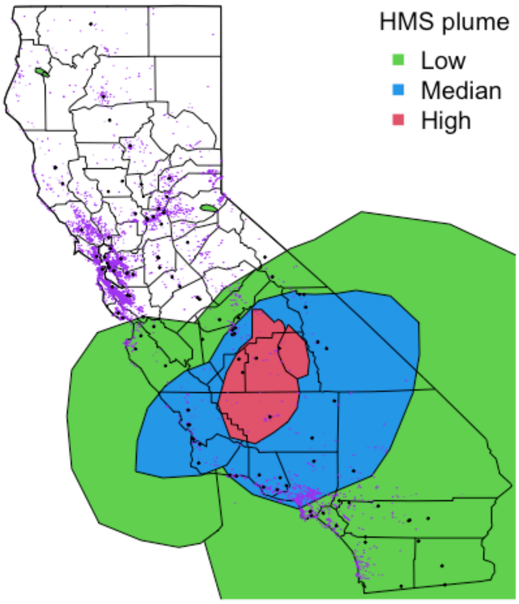


Figure 1. HMS smoke plume density (shaded regions) in California on 20 September 2021 and the locations of PA (purple dots) and AQS (black dots) monitoring stations.

2.1.1. Satellite-Derived Smoke Plume Indicators

Exposure to wildland fire smoke is assessed using smoke plume indicators supplied by the HMS [33]. This automated data product integrates observations from multiple polar and geostationary satellites to generate polygons representing smoke plume extents on a daily basis. Distinct polygons are provided for low-, medium-, and high-density plumes. These smoke plume indicators tend to underestimate the actual intensity of smoke, as they primarily rely on satellite imagery with an approximate spatial resolution spanning several miles [34,35]. Additionally, smoke visibility is limited to daytime hours, resulting in a significant underestimation of smoke levels during the night. While the HMS data are among the most reliable widely-available measures of plume extent, Ref. [35] shows that it may underestimate wildland fire contribution to $PM_{2.5}$.

2.1.2. AQS Monitoring Stations

The AQS monitoring stations, deployed by the US Environmental Protection Agency (EPA) and state, local, and tribal air pollution control agencies, provide precise $PM_{2.5}$ measurements. However, their distribution is spatially sparse due to the high cost and complexity associated with their installation and maintenance.

2.1.3. PA Sensors

PA sensors are low-cost monitoring devices deployed by individuals and organizations for continuous ambient air pollutant tracking. Even during wildland fire events, PA sensors have been show to strongly correlate with gold standard measurements [36]. Despite their affordability and ease of installation, PA sensors offer less accurate $PM_{2.5}$ readings and are significantly influenced by environmental factors, such as temperature and humidity [15]. We use bias corrected data for all analyses. However, this initial bias correction based on Barkjohn et al. in [15] may be insufficient because it only depends on a linear trend in temperature and humidity and is constant across space and time. Therefore, our Bayesian data fusion model adds a more flexible spatiotemporal bias correction term.

Before fitting the statistical model, we implemented several pre-processing steps on the PA data and standardized temperature and humidity. PA stations feature two independent channels, Channel A and Channel B, both of which measure ambient $PM_{2.5}$ independently. To achieve a more accurate estimation of the actual ambient $PM_{2.5}$ concentration, we discarded readings where the measurement difference exceeded 200 $\mu g/m^3$. We discarded readings where the daily readings have constant high $PM_{2.5}$ readings over 2000 $\mu g/m^3$. We choose the threshold of value of 2000 $\mu g/m^3$ because some PA stations have a constant reading around 2000 $\mu g/m^3$, and all other stations have values at most at 800 $\mu g/m^3$, which suggests a data collection error. Subsequently, the mean reading from Channel A and Channel B was considered as the PA measurement.

A majority of PA stations measure temperature (in Fahrenheit) and relative humidity. Because the temperature and humidity are spatially smooth, we employ a 10-nearest-neighbor approach to impute stations with missing temperature and humidity values and unobserved sites in California.

In 2021, more than 7800 outdoor PA sensors were operational in California. We only use outdoor sensor for comparison with AQS stations. We have included only those PA stations that reported fewer than 18 missing days during the fire seasons, resulting in a total of 1080 for 2020 and 712 PA stations for 2021. Figure 2 displays the distribution of $PM_{2.5}$, aggregated across stations for 2021, by smoke plume intensity. A similar pattern is observed in both PA stations and AQS stations where $PM_{2.5}$ measurements escalate in the presence of a smoke plume.

Figure 3 shows one AQS and a nearby PA monitor daily readings over the fire season in 2021. For this stations, the two types of monitors have a high degree of correlation, and both monitors' readings are elevated when under the high smoke plume. Figure 4 investigates the relationship between AQS stations and their corresponding nearby PA sites across California. For each AQS site, we compute its correlation with the closest PA site. Figure 4 plots these correlations, binned by the distance between the AQS and PA sites. The correlation is high when the stations are close and decreases with distance, suggesting that PA data will be a useful supplement to the spatial model.

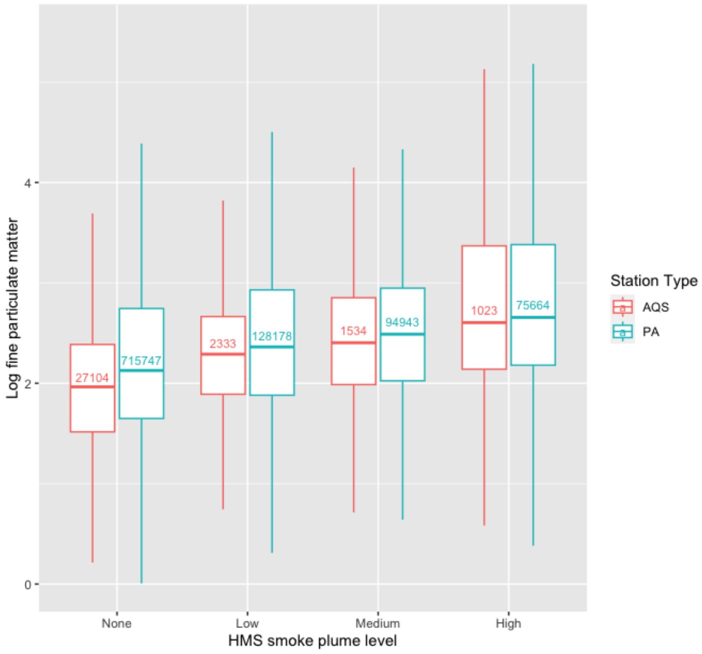


Figure 2. Distribution of log $PM_{2.5}$ ($\mu g/m^3$) by smoke plume level for PA and AQS stations. Four smoke plume levels from left to right are: no smoke, low, medium, and high plume density. The number of observations for each smoke plume level and each sensor type is displayed in the box.

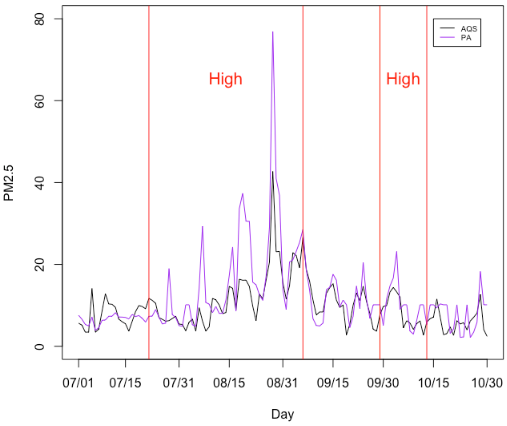


Figure 3. One AQS monitor at 37°20′N, 121°53′W (downtown San Jose) and a nearby PA monitor $PM_{2.5}$ readings over the fire season in 2021. Dates from 07/18 to 09/03 and 09/29 to 10/13 are covered in high smoke plume region and are indicated by "High" above.

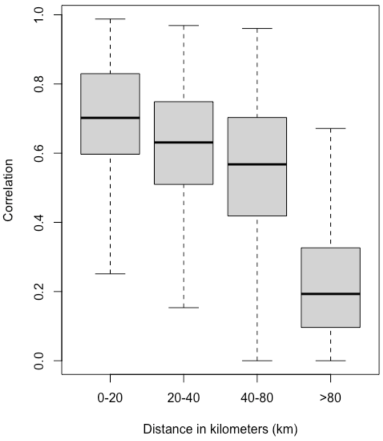


Figure 4. Sample correlation between AQS and nearby PA stations versus the distance (km).

2.2. Statistical Model

We propose a multi-resolution Bayesian model for modeling AQS and PA measurements jointly in the spectral domain. Let $Y_{1t}(\mathbf{s})$ and $Y_{2t}(\mathbf{s})$ be AQS and PA measurements, respectively, for spatial location $\mathbf{s}$ at time (day) $t \in \{1, \ldots, n_t\}$, and $\mathbf{X}_t(\mathbf{s}) = \{X_{0t}(\mathbf{s}), \ldots, X_{pt}(\mathbf{s})\}$ be a corresponding vector of covariates with $X_{0t}(\mathbf{s}) = 1$ for the intercept. The $p = 5$ covariates are temperature, relative humidity and indicators of low, medium and high density smoke plumes at site $\mathbf{s}$ and day t. We note that temperature and relative humidity are standardized to have mean zero and variance one and that the AQS and PA measurements are not taken at the same spatial locations.

The observations are decomposed as $Y_{jt}(\mathbf{s}) = Z_{jt}(\mathbf{s}) + \varepsilon_{jt}(\mathbf{s})$ for $j \in \{1, 2\}$, where $j = 1$ and $j = 2$ indicate AQS and PA monitors, respectively, $Z_{1t}(\mathbf{s})$ and $Z_{2t}(\mathbf{s})$ are spatiotemporal processes, and $\varepsilon_{jt}(\mathbf{s}) \overset{indep}{\sim} \text{Normal}(0, \tau_j^2)$ is error. The time span of our data is relatively short, therefore, its reasonable to assume the spatiotemporal processes are stationary within the modeling period. We will apply Fourier transformation to the spatiotemporal processes $Z_{jt}(\mathbf{s})$ with respect to time to remove the temporal dependence. The resulting spectral processes $Z^*_{jl}(\mathbf{s})$ capture periodicity, are independent over frequency $\{\omega_l, l = 1, \ldots, n_t\}$ and spatially correlated. For time series observed at equal time intervals, we can apply the DFT. The spectral processes at frequency ω_l is

$$Z^*_{jl}(\mathbf{s}) = \sum_{t=1}^{n_t} \exp(-it\omega_l) Z_{jt}(\mathbf{s}) \tag{1}$$

and measures the variation in $Z_{jt}(\mathbf{s})$ at frequency ω_l. Terms with small ω_l (low frequency) represent long-term trends such as month-to-month averages and terms with large ω_l (high frequency) represent short-term trends such as day-to-day variation. Let $\{Z^*_{j1}(\mathbf{s}), \ldots, Z^*_{jn_t}(\mathbf{s})\}$ be the unique real components of the DFT of $\{Z_{j1}(\mathbf{s}), \ldots, Z_{jn_t}(\mathbf{s})\}$ at frequencies $\{\omega_1, \ldots, \omega_{n_t}\}$ with $\omega_1 \leq \cdots \leq \omega_{n_t}$.

The spectral processes $Z^*_{jl}(\mathbf{s})$ are dependent across $j = 1, 2$, as they represent the two networks measuring the same underlying PM$_{2.5}$ process. They are also spatially dependent processes as locations nearby may exhibit similar periodicity. We model the cross network dependence and spatial dependence for each ω_l as

$$Z^*_{1l}(\mathbf{s}) = U_l(\mathbf{s}) \text{ and } Z^*_{2l}(\mathbf{s}) = A_l U_l(\mathbf{s}) + V_l(\mathbf{s}), \tag{2}$$

where spatial process $U_l(\mathbf{s})$ is the true PM$_{2.5}$ concentration for frequency l. The PA stations are assumed to be measuring a biased and noisy version of the true PM$_{2.5}$ with discrepancy $V_l(\mathbf{s})$. The coefficient A_l controls the dependence across networks. Both the bias $V_l(\mathbf{s})$ and cross-dependence A_l vary by ω_l to allow for a multi-resolution calibration of the two networks. We model A_l linearly as $A_l = \beta_{A0} + \beta_{A1} \cdot \omega_l$, where β_{A0} and β_{A1} are unknown coefficients. This allows the correlation between the processes to vary stochastically with frequency. For example, if PA is more reliable for long-term trends than day-to-day variation, then we expect larger (smaller) correlation between networks for small (large) ω_l.

The true process $U_l(\mathbf{s})$ and discrepancy term $V_l(\mathbf{s})$ are both regressed onto the covariates. Since we are developing a model in the spectral domain, we will also apply DFT to each covariate in $\mathbf{X}_t(\mathbf{s})$ with respect to time and denote this as $\mathbf{X}^*_j(\mathbf{s}) = \{X^*_{0l}(\mathbf{s}), \ldots, X^*_{pl}(\mathbf{s})\}$ Define the covariates for the true process U_l as $\mathbf{X}^*_{ul}(\mathbf{s}) = \mathbf{X}^*_j(\mathbf{s})$, containing all five covariates, and define $\mathbf{X}_{vl}(\mathbf{s}) = \{X^*_{0l}(\mathbf{s}), X^*_{1l}(\mathbf{s}), X^*_{2l}(\mathbf{s})\}$ to include only temperature and relative humidity for bias correction [15]. We model $U_l(\mathbf{s})$ and $V_l(\mathbf{s})$ as independent (with each other and over l) Gaussian processes with means $\text{E}\{U_l(\mathbf{s})\} = \mathbf{X}^*_{ul}(\mathbf{s})\boldsymbol{\beta}_u$ and $\text{E}\{V_l(\mathbf{s})\} = \mathbf{X}^*_{vl}(\mathbf{s})\boldsymbol{\beta}_v$, variances $\text{Var}\{U_l(\mathbf{s})\} = \sigma^2_{ul}$ and $\text{Var}\{V_l(\mathbf{s})\} = \sigma^2_{vl}$, and spatial correlations $\text{Cor}\{U_l(\mathbf{s}), U_l(\mathbf{s}')\} = \exp(-||\mathbf{s} - \mathbf{s}'||/\rho_u)$ and $\text{Cor}\{V_l(\mathbf{s}), V_l(\mathbf{s}')\} = \exp(-||\mathbf{s} - \mathbf{s}'||/\rho_v)$.

The regression coefficients $\boldsymbol{\beta}_u = (\beta_{u0}, \ldots, \beta_{up})^T$ control the effects of the covariates on the true PM$_{2.5}$ process U. Although we specify the model in the spectral domain, the

DFT is a linear operator and thus the covariates can be interpreted as usual in the spatial domain since the mean AQS response is

$$\mathrm{E}\{Y_{1t}(\mathbf{s})\} = \mathbf{X}_t(\mathbf{s})\boldsymbol{\beta}_u \tag{3}$$

Therefore, $\boldsymbol{\beta}_u$ is of primary interest. In particular, the components of $\boldsymbol{\beta}_u$ that correspond to the smoke plume indicators are used to summarize the wildland fire contribution to $\mathrm{PM}_{2.5}$.

The regression coefficients $\boldsymbol{\beta}_v = (\beta_{v0}, \beta_{v1}, \beta_{v2})^T$ control the effect of the covariates on the discrepancy term V, and thus the contribution of the covariates to the PA bias. By allowing the covariance parameters σ^2_{ul} and σ^2_{vl} to vary by frequency (l), we allow for a different degree of dependence between the networks at different temporal scales, with

$$\mathrm{Cor}\{Z^*_{1l}(\mathbf{s}), Z^*_{2l}(\mathbf{s})\} = \frac{A_l}{\sqrt{A_l^2 + \sigma^2_{vl}/\sigma^2_{ul}}}. \tag{4}$$

The prior for the variance components is

$$\sigma^2_{ul} \sim \mathrm{InvGamma}(a_{ul}, b_{ul}) \quad \text{and} \quad \sigma^2_{vl} \sim \mathrm{InvGamma}(a_{vl}, b_{vl}) \tag{5}$$

where the hyperparameters are modelled as log-linear in frequency, e.g., $\log(a_{ul}) = \gamma_{au1} + \gamma_{au2} \cdot \omega_l$ the prior captures the intuition that the variance is higher in month-to-month variation than day-to-day variation, and the correlation between two sources vary over frequencies.

2.3. Quantifying the Wildland Fire Contribution

To estimate the $\mathrm{PM}_{2.5}$ contribution from wildfire, given the estimated parameters above, we consider two metrics based on either regression or matching. For the regression metric, let $\mathbf{X}^0_t(\mathbf{s})$ be the covariate vector with three plume indicators fixed at zero. For the matching estimator, define $\mathcal{P}(\mathbf{s})$ as the set of days for which site $\mathbf{s}$ is in a smoke plume (any density) and $\bar{\mathcal{P}}(\mathbf{s})$ as the set of non-plume days. We match each plume day with a non-plume day with similar meteorology and time period. Let $\mathcal{A}_t(\mathbf{s}) = \bar{\mathcal{P}}(\mathbf{s}) \cap \{t - 30, \ldots, t + 30\}$ be the set of non-plume days within 30 days of plume day t. For each plume day, we selected the matching day $m_t(\mathbf{s})$ as

$$m_t(\mathbf{s}) = \arg\min_{d \in \mathcal{A}_t(\mathbf{s})} \; |\mathrm{temp}_t(\mathbf{s}) - \mathrm{temp}_d(\mathbf{s})| + \phi|\mathrm{humidity}_t(\mathbf{s}) - \mathrm{humidity}_d(\mathbf{s})| \tag{6}$$

where ϕ above is a scaling factor adjusting the magnitude of humidity and temperature, we set $\phi = 1$ so that the best matching station has equal weights on temperature and humidity. Then at site $\mathbf{s}$ the estimated contribution from wildland fires per day are

1. Regression estimator: $\delta_1(\mathbf{s}) = \frac{1}{n_t}\sum_{t=1}^{n_t}\{\mathbf{X}_t(\mathbf{s}) - \mathbf{X}^0_t(\mathbf{s})\}\boldsymbol{\beta}_u$
2. Matching estimator: $\delta_2(\mathbf{s}) = \frac{1}{n_t}\sum_{t\in\mathcal{P}(\mathbf{s})}\{Z_{1t}(\mathbf{s}) - Z_{1t'}(\mathbf{s})\}$ for $t' = m_t(\mathbf{s})$.

In the matching estimator, $Z_{1t}(\mathbf{s})$ is the true $\mathrm{PM}_{2.5}$, the transformed pairs of $Z^*_{1l}(\mathbf{s})$ in (2) obtained by inverse DFT, and thus this estimator accounts for spatiotemporal bias and correlation. Since the analysis is on the log-scale, we plot $\exp\{\delta_1(\mathbf{s})\}$ and $\exp\{\delta_2(\mathbf{s})\}$ which estimate the multiplicative effect, i.e., $\exp\{\delta_1(\mathbf{s})\} = 1.05$ corresponds to a 5% increase in $\mathrm{PM}_{2.5}$ in the presence of a smoke plume.

2.4. Computational Algorithm

To complete the Bayesian model, we specify uninformative prior distributions for the model parameters. The regression coefficients have Gaussian priors $\boldsymbol{\beta}_u, \boldsymbol{\beta}_v \sim$ Normal $(\mathbf{0}, c^2\mathbf{I}_{p+1})$. The variance parameters have conjugate priors $\tau^2_j \sim \mathrm{InvGamma}(a, b)$. The hyperpameters have Gaussian priors $\gamma_{au1}, \gamma_{au2}, \gamma_{av1}, \gamma_{av2} \sim \mathrm{Normal}(0, c^2)$. To give uninformative priors we set $a = b = 0.01$ and $c = 10$. Due to poor convergence, the dependence parameters β_{A0} and β_{A1} were fixed based on cross-validation to minimize mean squared prediction error for AQS stations.

The main computational bottleneck of spatial modeling is manipulating spatial covariance matrices to estimate the range parameters ρ_u and ρ_v. Given the large size of the air pollution dataset, a reasonable simplification is to estimate the range parameters using variogram and then assume they are fixed for the purpose of fitting the final model. The estimated spatial range from variograms are $\rho_u = 177$ and $\rho_v = 111$ kilometers.

Given the range parameters are fixed, the remaining parameters are estimated using Markov Chain Monte Carlo (MCMC) methods. In particular, we perform Gibbs sampling steps for most parameters and Metropolis sampling for some hyperparameters. We generate 8000 posterior samples and discard the first 5000 as burn-in. The MCMC details are relegated to the Appendices A–C. Appendix A gives the details of each MCMC step. A simulation study is included in the Appendix B to verify the algorithm produces reliable parameter estimates. Convergence is monitored using trace plots for several representative parameters shown in Appendix C.

3. Results

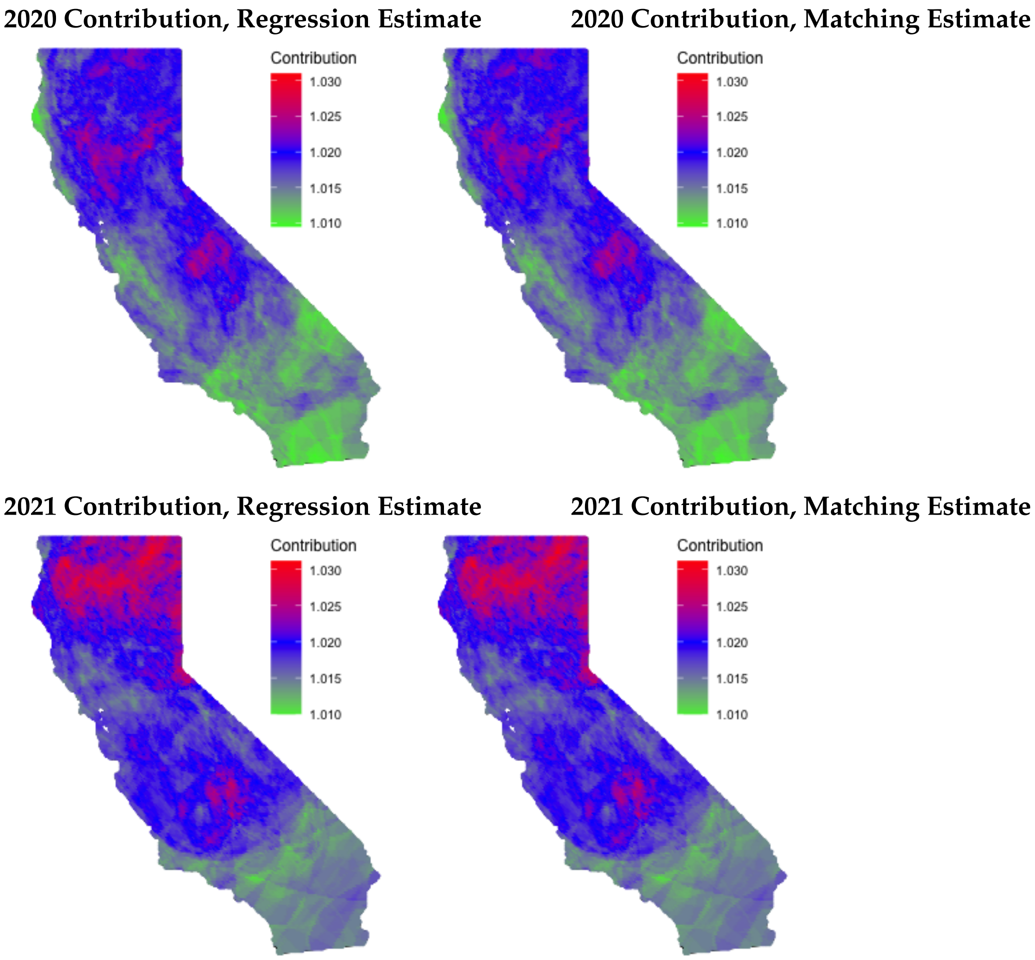
3.1. Summary of the Fitted Model

Table 1 gives the estimates of the regression coefficients for both the true process β_u and bias correction term β_v. All three smoke plume levels positively affect $PM_{2.5}$ concentrations, with high smoke plumes having the greatest impact, followed by medium and low smoke plumes. These results are consistent between 2020 and 2021. The bias correction terms, however, are not significant. Given that PA readings have already been corrected as per [15] using temperature and relative humidity, it is reasonable that these variable do not explain trends in bias. We note that our model does include more general spatiotemporal bias correction in $V_l(\mathbf{s})$ and including this bias term leads to improved results, as discussed below.

Table 1. Posterior mean (95% interval) for the model parameters. The regression coefficients are given separately for the true $PM_{2.5}$ process ($\boldsymbol{\beta}_u$) and bias correction ($\boldsymbol{\beta}_v$). A "***" indicates that the 95% interval excludes zero.

2020 Fire Season		
Parameter	**True $PM_{2.5}$**	**Bias Correction**
Temperature	0.115 (0.106,0.125) ***	−0.002 (−0.009,0.005)
Humidity	0.064 (0.048,0.080) ***	0.012 (−0.002,0.035)
Plume—Low	0.007 (0.003,0.011) ***	/
Plume—Medium	0.022 (0.012,0.032) ***	/
Plume—High	0.049 (0.033,0.065) ***	/
2021 Fire Season		
Parameter	**True $PM_{2.5}$**	**Bias Correction**
Temperature	0.006 (0.004,0.008) ***	0.006 (−0.003,0.015)
Humidity	0.000 (−0.001,0.001)	−0.011 (−0.026,0.003)
Plume—Low	0.011 (0.001,0.021) ***	/
Plume—Medium	0.018 (0.007,0.029) ***	/
Plume—High	0.041 (0.031,0.051) ***	/

Figure 5 plots the estimated wildland fire contribution both years and both metrics. The estimated wildland fire contribution ranges from a 1–3% increase in $PM_{2.5}$, depending on the location. Both metrics yield similar estimates of contribution and spatial patterns. The impact of wildfires varies across the state and years. In 2020, both Northern and Central California experienced significant wildfire impacts, while only Northern California faced major effects in 2021. This is in line with the fact that 2020 had the highest frequency of wildfires across all states, whereas 2021 witnessed a single, massive wildfire in Northern California [37]. Figure 6 shows the posterior standard deviation of the contribution. The uncertainty of estimation in 2020 is generally smaller than 2021. Moreover, both estimators give roughly the same undertainty estimation, with matching estimator only slightly more stable than regression estimate.


2020 Contribution, Regression Estimate

2020 Contribution, Matching Estimate

2021 Contribution, Regression Estimate

2021 Contribution, Matching Estimate

Figure 5. Smoke contribution to $PM_{2.5}$. Contributions are exponentiated to reflect actual percentage contribution. For example, 1.01 and 1.03 mean wildfire contributes to roughly a 1% to 3% increase in $PM_{2.5}$.

In addition to covariate effects, the data-fusion model provides an evaluation of the concordance between AQS and PA stations. Equation (4) defines the correlation between the two networks as a function of the spectral frequency, ω_l. Figure 7 plots the correlation between AQS and PA by period, i.e., $1/\omega_l$. For example, period 7 (30) corresponds to variation that occurs on a weekly (monthly) scale. Figure 7 shows that the correlation between AQS and PA stations increases from short-term, such as day-to-day variation, to long-term, such as month-to-month variation. In the short-term, the correlation is lower since the readings are taken at different spatial locations and are subject to small scale variability. Over the long run, the correlation is higher as both sources estimate ambient unbiased $PM_{2.5}$ readings.

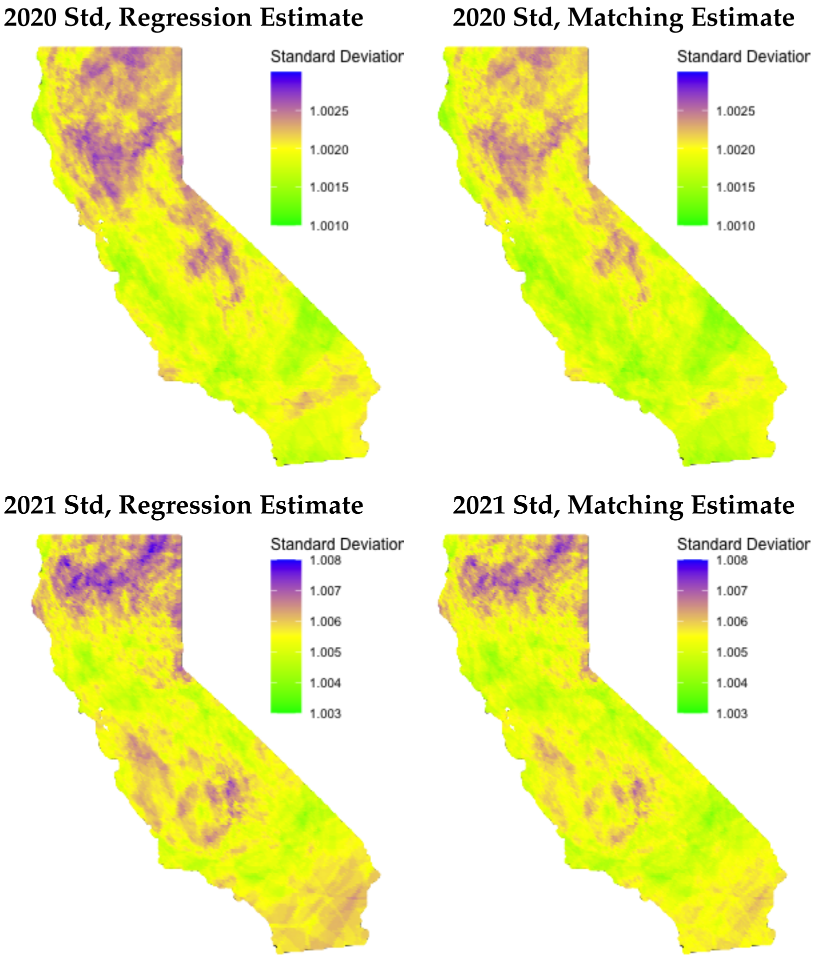


Figure 6. Posterior standard deviation of smoke contribution to $PM_{2.5}$. The posterior standard deviations are not exponentiated, and they show uncertainty estimation on the original scale.

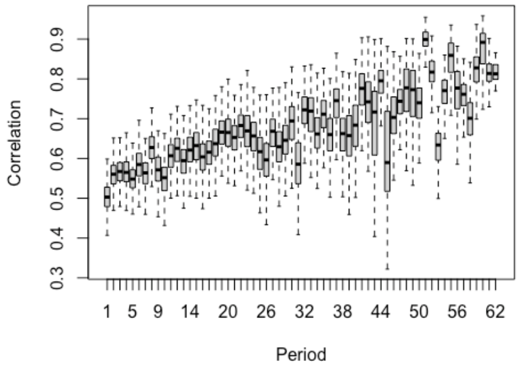


Figure 7. Posterior distribution of the correlation between AQS and PA by period. Small periods capture short-term variation, such as day-to-day variation, while large periods capture long-term variation, such as monthly trends.

3.2. Model Comparisons

To assess the effectiveness of integrating additional PA readings, we compared the proposed data-fusion model (“Data fusion”) with two simpler alternatives. The first uses only AQS data (“AQS only”) and discards the PA data (i.e., sets $A_l = 0$ for all l). The second naively (“Naive”) combines AQS and PA data and treats them as a single source without spatiotemporal bias adjustment (i.e., sets $A_l = 1$ and $V_l(\mathbf{s}) = 0$ for all $\mathbf{s}$, and includes an indicator variable in the regression term, $\boldsymbol{\beta}_u$, to distinguish two types of data).

The estimated parameters for each model, along with the corresponding posterior standard deviations, are presented in Table 2. Clearly, incorporating PA monitors significantly reduces the posterior standard deviation. For many of the parameters the reduction in uncertainty is striking, with the standard deviation being 2–4 times smaller for the data-fusion model. Also, with the AQS-only model, only high smoke plumes exhibit a significant contribution due to a higher standard deviation. In contrast, when merging AQS and PA data, both medium and high smoke plume levels show significant contributions.

Table 2. Posterior mean (standard deviation) for the model parameters $\boldsymbol{\beta}_u$ for the CA data using the proposed data-fusion model, the model that uses only AQS data, and the naive data-fusion model that ignores bias in the PA data. A "***" indicates that the 95% interval excludes zero.

2020 Fire Season			
Parameter	**Data Fusion**	**AQS Only**	**Naive**
Temperature	0.115 (0.005) ***	0.105 (0.024) ***	−0.418 (0.066) ***
Humidity	0.064 (0.008) ***	0.086 (0.022) ***	−1.125 (0.052) ***
Plume—Low	0.007 (0.002) ***	0.005 (0.012)	0.107 (0.078)
Plume—Medium	0.022 (0.005) ***	0.020 (0.014)	0.271 (0.052) ***
Plume—High	0.049 (0.008) ***	0.042 (0.016) ***	0.637 (0.079) ***
2021 Fire Season			
Parameter	**Data Fusion**	**AQS Only**	**Naive**
Temperature	0.006 (0.001) ***	0.015 (0.003) ***	−0.014 (0.006) ***
Humidity	0.000 (0.000)	0.008 (0.002) ***	−0.039 (0.003) ***
Plume—Low	0.011 (0.004) ***	−0.001 (0.014)	−0.330 (0.032) ***
Plume—Medium	0.018 (0.004) ***	0.023 (0.016)	0.230 (0.074) ***
Plume—High	0.041 (0.005) ***	0.054 (0.017) ***	0.980 (0.071) ***

Furthermore, to verify that our proposed methodologies not only improve parameter estimation but also lead to accurate $PM_{2.5}$ predictions, we performed a 5-fold cross-validation for the three models using data from 2021. We randomly split the AQS stations into five folds. For each fold, we build predictive models based on the other AQS stations and all PA stations and make predictions at the test sites. Performance was compared based on three key metrics: Root mean squared error, 95% prediction coverage, and prediction variance. For all models, we fix the spatial range parameters (ρ_u and ρ_v) based on the variogram analysis of the full dataset. The cross-dependence parameter A_l is fixed at 0.2.

The results in Table 3 show that the performance of the AQS-only analysis is fairly similar to the proposed data-fusion approach, with slightly smaller prediction mean squared error and larger average prediction variance. Therefore, carefully including the additional PA data mainly reduces the prediction variance. However, naively including the PA data gives much higher prediction errors and low coverage.

Table 3. Root mean squared error ("RMSE"), coverage of 95% prediction intervals ("Coverage") and average prediction variance ("Ave Var") for the cross-validation study comparing the proposed data fusion model to models that ignore PA data ("AQS only") and includes PA data without bias correction ("Naive").

Model	**RMSE**	**Coverage**	**Ave Var**
Data Fusion	0.42	0.89	0.13
AQS only	0.40	0.91	0.16
Naive	0.66	0.73	0.18

In summary, the AQS-only and data fusion model produce fairly similar out-of-sample prediction accuracy, therefore the main benefit of including the PA data is reducing uncertainty in parameter estimates. Also, the Naive model gives a 50% larger RMSE and low coverage, emphasizing the need for a careful data fusion approach.

4. Discussion

In this study, we examine the impact of wildland fires on $PM_{2.5}$ concentrations in California during the fire seasons of 2020 and 2021. As we can see from Figure 5, $PM_{2.5}$ contributes to about a 3% increase in parts of California that are heavily affected by wildland fires in both 2020 and 2021; in most other areas the increase ranges from 1.0% to 2.3%. To obtain precise estimates, we combine remotely-sensed smoke-plume indicators with AQS and PA measurement networks. To model the spatiotemporal correlation of $PM_{2.5}$ concentration and relationship between AQS and PA monitors, we first transform the data from spatial domain to frequency domain, and then use a data-fusion approach to model spatial correlations while accounting for biases in the PA data. Furthermore, we use a Bayesian approach to compute posterior distributions of the quantities of interest to fully characterize uncertainty.

As shown in Table 2, we find that including PA monitors significantly increases the precision of the estimated contribution of wildland fire smoke to total $PM_{2.5}$. Using only AQS data we find that medium and high smoke plume levels significantly contribute to $PM_{2.5}$ concentration with standard deviations as large as 0.017, and the data fusion approach that supplements AQS with PA data gives similar parameter estimation, with standard deviation as small as 0.004. Moreover, the data fusion model also estimates a significant low smoke plume level contribution. However, as we can see from Table 3, since $PM_{2.5}$ concentration is relatively smooth across space and AQS stations are evenly distributed across the state, incorporating PA readings does not improve prediction performance even for the data-fusion approach. Comparing prediction performance does reveal that simple data fusion model such as the model that ignores bias in the PA data gives inferior prediction results. Based on Table 1, with our model, all three smoke plume levels demonstrate a significant contribution to $PM_{2.5}$ concentration, and the impact varies across different regions depending on the year. This study highlights the value of utilizing both AQS and PA data in understanding the impact of wildfires on air quality and informs future monitoring and management efforts.

There are some limitations of our current work. First, as mentioned above, the satellite-derived smoke plume levels might underestimate the actual smoke level, which may lead to underestimation of wildfires' contribution to $PM_{2.5}$ [34]. Second, due to computational limitations and poor MCMC convergence, we fixed the spatial correlation range parameters for both AQS and PA monitors and parameters that control the relationships between AQS and PA data. The analysis would more fully quantify uncertainty if we are able to implement a fully Bayesian analysis. Our analysis of the smoke contribution is also limited because we only consider temperature and relative humidity and no other meteorological variables or anthropogenic sources. Another limitation is that we use only HMS smoke indicators to denote fire smoke, which has known limitations [35]. Although we estimate the relationship between HMS and PM2.5 concentration using the data, HMS may fail to capture the smoke contribution from some fires.

We have taken a purely statistical approach to estimating the contribution of wildland fires on ambient air pollution. An area of future work is to incorporate numerical models to simulate the process. Dispersion models, e.g., HySPLIT [38], combine the location and size of fires and meteorological conditions in a mathematical model to track particulate matter emanating from a fire. Of course, numerical models also have bias and other limitations [39], but combining their output within our statistical framework would likely further refine our estimates. Further, instead of using one range parameter for all frequencies, it is possible to get variogram estimates of ranges over frequencies. Similarly, instead of assuming the same β_u and β_v for all locations, it may be better to estimate spatially-varying β_u and β_v, although this would be computationally intensive. To extend the current work, we can estimate the contribution over the entire U.S., although more efficient computational methods would be required for this analysis.

Author Contributions: Conceptualization, A.G.R. and B.J.R.; methodology, H.Y., Y.G., B.J.R.; validation, H.Y., S.R.-S.; formal analysis, H.Y.; data curation, H.Y., S.R.S.; writing-original draft preparation, H.Y.; writing—review and editing, S.R.-S., B.J.R., Y.G., A.G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health (R01ES031651-01) and the National Science Foundation (DMS2152887).

Data Availability Statement: AQS data is a publicly available dataset, which is part of this study. This data can be found on EPA website https://aqs.epa.gov/aqsweb/airdata/download_files.html (accessed on 1 April 2023). PA data is a 3rd party data and restrictions apply to the availability of these data. Data was obtained from Purple Air and are available from PurpleAir API https://community.purpleair.com/t/making-api-calls-with-the-purpleair-api/180 (accessed on 1 April 2023) with the permission of Purple Air. HMS smoke plume data is publicly available and can be downloaded at Office of Satellite and Product Operations website https://www.ospo.noaa.gov (accessed on 1 April 2023). The codes to download and analyze data in this paper is available at this GitHub repo https://github.com/hyang199723/PAFusion (uploaded on 30 June 2023).

Acknowledgments: This work does not represent EPA views.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. MCMC Algorithm

Assume the n_1 AQS monitors are at spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_{n_1}$ and the n_2 PA monitors are located at $\mathbf{s}_{n_1+1}, \ldots, \mathbf{s}_{n_s}$ for $n_s = n_1 + n_2$. The observations can be written as the vectors $\mathbf{Y}_{1t} = [Y_{1t}(\mathbf{s}_1), \ldots, Y_{1t}(\mathbf{s}_{n_1})]^T$, $\mathbf{Y}_{2t} = [Y_{2t}(\mathbf{s}_{n_1+1}), \ldots, Y_{1t}(\mathbf{s}_{n_s})]^T$ and $\mathbf{Y}_t = (\mathbf{Y}_{1t}^T, \mathbf{Y}_{2t}^T)^T$. Similarly, for frequency l let $\mathbf{Y}_{jl}^*$, $\mathbf{U}_{jl}$ and $\mathbf{V}_{jl}$ be vectors of length n_j and $\mathbf{Y}_l^*$, $\mathbf{U}_l$ and $\mathbf{V}_l$ be vectors of length n_s, analogous to $\mathbf{Y}_t$. The covariate matrices of size $n_j \times p$ are denoted $\mathbf{X}_{jl}^*$ and $\mathbf{X}_l^*$ is the $n_s \times p$ matrix that stacks $\mathbf{X}_{1l}^*$ and $\mathbf{X}_{2l}^*$. Then the model in the spectral domain is

$$\mathbf{Y}_{1l}^* = \mathbf{U}_l + \mathbf{E}_{1l} \quad \text{and} \quad \mathbf{Y}_{2l}^* = A_l \mathbf{U}_l + \mathbf{V}_{2l} + \mathbf{E}_{2l} \tag{A1}$$

where $\mathbf{E}_{jl} \overset{indep}{\sim} \text{Normal}(\mathbf{0}, \tau_j^2 \mathbf{I}_{n_j})$. Using this notation, the spatial models are defined by $\text{E}(\mathbf{U}_{jl}) = \mathbf{X}_{jl}^* \boldsymbol{\beta}_u$, $\text{E}(\mathbf{V}_{jl}) = \mathbf{X}_{jl}^* \boldsymbol{\beta}_v$, $\text{Cov}(\mathbf{U}_{jl}, \mathbf{U}_{kl}) = \sigma_{ul}^2 \boldsymbol{\Sigma}_{ujk}$ and $\text{Cov}(\mathbf{V}_{jl}, \mathbf{V}_{kl}) = \sigma_{vl}^2 \boldsymbol{\Sigma}_{vjk}$. The full $n_s \times n_s$ spatial correlation matrices are denoted $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_v$.

Each MCMC iteration we impute missing data and update the error variance parameters in the spatial domain, and then update all remaining parameters in the spectral domain. The missing values are simply drawn from the univariate normal distribution

$$Y_{jt}|\text{rest} \sim \text{Normal}(Z_{jt}(\mathbf{s}), \tau_j^2) \tag{A2}$$

independently over j and t. The error variances are drawn from full conditional distribution $\tau_1^2|\text{rest} \sim \text{InvGamma}[n_1 n_t/2 + a, \sum_{i=1}^{n_1} \sum_{t=1}^{n_t} \{(Y_{1t}(\mathbf{s}_i) - Z_{2t}(\mathbf{s}_i)\}^2/2 + b]$ and $\tau_2^2|\text{rest} \sim \text{InvGamma}[n_2 n_t/2 + a, \sum_{i=n_1+1}^{n_s} \sum_{t=1}^{n_t} \{(Y_{2t}(\mathbf{s}_i) - Z_{2t}(\mathbf{s}_i)\}^2/2 + b]$.

After imputation in the spatial domain, the data are complete and can be projected into the spectral domain where they are independent over time. The spatial processes are updated as

$$\begin{aligned} \mathbf{U}_l|\text{rest} &\sim \text{Normal}\left\{\boldsymbol{\Omega}_{ul}\left(\mathbf{T}\mathbf{A}_l^1(\mathbf{Y}_l^* - \mathbf{V}_l) + \frac{1}{\sigma_{ul}^2}\boldsymbol{\Sigma}_u^{-1}\mathbf{X}_l^*\boldsymbol{\beta}_u\right), \boldsymbol{\Omega}_{ul}\right\} \\ \mathbf{V}_{2l}|\text{rest} &\sim \text{Normal}\left\{\boldsymbol{\Omega}_{vl}\left(\frac{1}{\tau_2^2}(\mathbf{Y}_{2l}^* - A_l\mathbf{U}_{2l}) + \frac{1}{\sigma_{vl}^2}\boldsymbol{\Sigma}_{v22}^{-1}\mathbf{X}_{2l}^*\boldsymbol{\beta}_v\right), \boldsymbol{\Omega}_{vl}\right\} \end{aligned} \tag{A3}$$

where $\mathbf{A}_l^k$ is diagonal with first n_1 elements equal one and the remaining n_2 elements equal $\mathbf{A}_l^k$, $\mathbf{T}$ is diagonal with first n_1 elements equal τ_1^{-2} and the remaining n_2 elements equal τ_2^{-2}, $\mathbf{V}_l$ is the vector with n_1 zeros followed by $\mathbf{V}_{2l}$, $\mathbf{\Omega}_{ul}^{-1} = \mathbf{T}\mathbf{A}_l^2 + \frac{1}{\sigma_{ul}^2}\mathbf{\Sigma}_u^{-1}$ and $\mathbf{\Omega}_{vl}^{-1} = \frac{1}{\tau_2^2}\mathbf{I}_{n_2} + \frac{1}{\sigma_{vl}^2}\mathbf{\Sigma}_{v22}^{-1}$.

The regression coefficients and bias parameters are updated as

$$\begin{aligned} \boldsymbol{\beta}_u|\text{rest} &\sim \text{Normal}\left\{\mathbf{P}_u\left(\sum_{l=1}^{n_t}\frac{1}{\sigma_{ul}^2}\mathbf{X}_l^{*T}\mathbf{\Sigma}_u^{-1}\mathbf{U}_l\right), \mathbf{P}_u\right\} \\ \boldsymbol{\beta}_v|\text{rest} &\sim \text{Normal}\left\{\mathbf{P}_v\left(\sum_{l=1}^{n_t}\frac{1}{\sigma_{vl}^2}\mathbf{X}_{2l}^{*\,T}\mathbf{\Sigma}_{v22}^{-1}\mathbf{V}_{2l}\right), \mathbf{P}_v\right\} \end{aligned} \tag{A4}$$

where $\mathbf{P}_u^{-1} = \sum_{l=1}^{n_t}\frac{1}{\sigma_{ul}^2}\mathbf{X}_l^{*T}\mathbf{\Sigma}_u^{-1}\mathbf{X}_l^* + \frac{1}{c^2}\mathbf{I}_p$ and $\mathbf{P}_v^{-1} = \sum_{l=1}^{n_t}\frac{1}{\sigma_{vl}^2}\mathbf{X}_{2l}^{*\,T}\mathbf{\Sigma}_{v22}^{-1}\mathbf{X}_{2l}^* + \frac{1}{c^2}\mathbf{I}_p$. The remaining hyperparameters are updated as

$$\begin{aligned} \sigma_{ul}^2|\text{rest} &\sim \text{InvGamma}\left(\frac{n_s}{2} + a_{ul}, \frac{(\mathbf{U}_l - \mathbf{X}_l^*\boldsymbol{\beta}_u)^T\mathbf{\Sigma}_u^{-1}(\mathbf{U}_l - \mathbf{X}_l^*\boldsymbol{\beta}_u)}{2} + b_{ul}\right) \\ \sigma_{vl}^2|\text{rest} &\sim \text{InvGamma}\left(\frac{n_2}{2} + a_{vl}, \frac{(\mathbf{V}_{2l} - \mathbf{X}_{2l}^*\boldsymbol{\beta}_v)^T\mathbf{\Sigma}_{v22}^{-1}(\mathbf{V}_{2l} - \mathbf{X}_{2l}^*\boldsymbol{\beta}_v)}{2} + b_{vl}\right). \end{aligned} \tag{A5}$$

Finally, γ_{au1}, γ_{au2}, γ_{av1} and γ_{av2} are updated using a Metropolis step with Gaussian candidate distribution tuned to give acceptance rate around 0.4.

Appendix B. Simulation Results

We conduct a simulation study to demonstrate the reliability of the MCMC algorithm. The regression parameters, $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_v$, are fixed at the mean of the 2021 model output in Table 1. We generate a total number of 80 AQS stations and 500 PA stations with 60 time steps. The spatial locations are randomly sampled from the region $(0, 15)^2$. The data was generated in the frequency domain using the following equations:

$$Y_{1l}(\mathbf{s}) = U_l(\mathbf{s}) + \epsilon_1(\mathbf{s}) \quad \text{and} \quad Y_{2l}(\mathbf{s}) = A_l U_l(\mathbf{s}) + V_{2l}(\mathbf{s}) + \epsilon_2(\mathbf{s}). \tag{A6}$$

The variables U_l and V_l are drawn from Gaussian processes as described in (2). The range parameters are set to $\rho_u = 2$ and $\rho_v = 4$. The error variances of $\epsilon_1(\mathbf{s})$ and $\epsilon_2(\mathbf{s})$ are set to 1.6 and 3.6, respectively. The values of A_l are fixed at the best A_l selected from the real data which is $A_l = 0.2$.

To simulate realistic smoke plume frequencies, we assigned percentages to represent the occurrence of low, medium, and high smoke plume levels. Specifically, 20% of the days corresponded to low smoke plume levels, 15% to medium levels, and 10% to high levels. Temperature and humidity values were randomly generated from standard normal distributions.

The covariates were initially generated in the time domain and then transformed to the frequency domain. The values of σ_{ul} form a decreasing sequence ranging from 50 to 10, with larger values assigned to lower frequencies. Similarly, σ_{vl} follows a decreasing sequence from 40 to 10. Finally, the values of β_u and β_v are the mean values from Table 1.

We generate 50 datasets from this model. For each simulated dataset, we fit the model with ρ_u, ρ_v and A_l fixed at the true values and generate 8000 MCMC iterations and discard the first 5000 as burn-in. Since our main interest is in the covariate effects, for each dataset we record the effective sample size of the MCMC algorithm [40] and the posterior mean estimator and 95% posterior interval.

For each dataset and each parameter, we compute the posterior mean, standard deviation and 95% interval and measure MCMC convergence using the effective sample size. The average of the posterior means, standard deviations and effective samples sizes, and the empirical coverage of 95% intervals are shown in Table A1. The posterior means show small bias, the coverage is near the nominal level and the effective sample size coefficients indicate reasonable convergence.

Table A1. True value used for the fixed effects for the true $PM_{2.5}$ (β_u) and bias (β_v) to simulate data and the average (SD) over the 50 datasets of the posterior mean estimators ("Ave post mean"), coverage of 95% posterior intervals and average (SD) effective sample size based on 3000 MCMC iterations.

Type	Covariate	True Value	Average Post Mean	Coverage	ESS
$PM_{2.5}$	Temperature	0.118	0.117 (0.013)	100%	420.23 (0.14)
	Humidity	0.064	0.069 (0.022)	96%	307.27 (0.10)
	Plume-Low	0.007	0.006 (0.132)	100%	875.99 (0.29)
	Plume-Medium	0.022	0.020 (0.037)	98%	376.91 (0.13)
	Plume-High	0.049	0.050 (0.176)	100%	480.22 (0.16)
Bias	Temperature	−0.002	0.003 (0.019)	92%	168.75 (0.06)
	Humidity	0.012	0.009 (0.041)	96%	176.97 (0.06)

Appendix C. MCMC Convergence

We display several representative trace plots of the data fusion model to verify the convergence of our MCMC algorithm for the 2021 CA analysis. After burn-in, the MCMC chains appear to have converged.

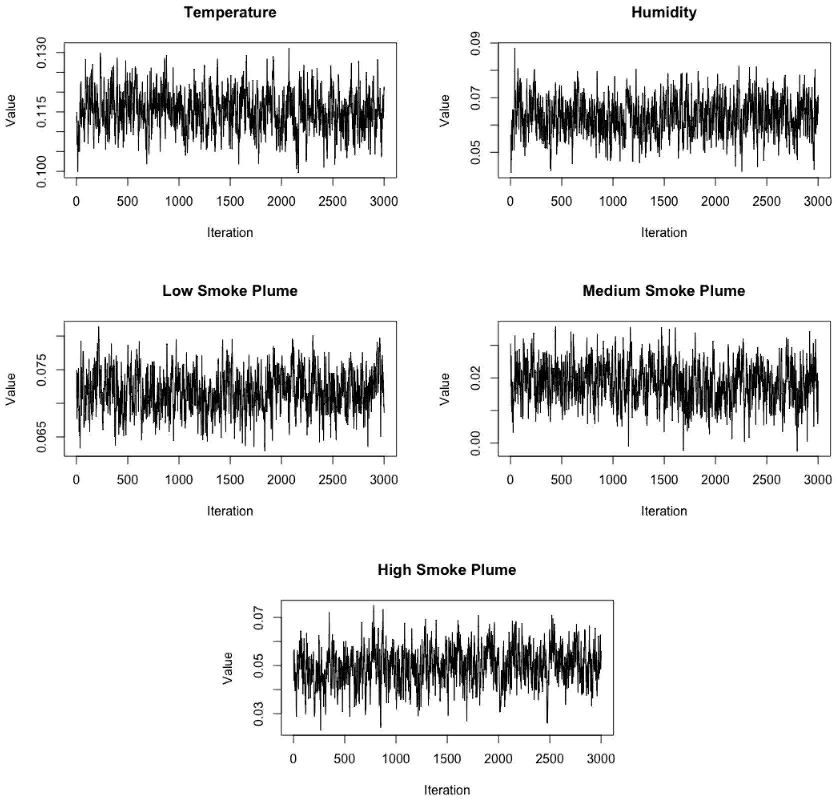


Figure A1. Trace plots of parameters of interest (β_u) for the 2021 California data analysis.

References

1. Dennekamp, M.; Abramson, M.J. The effects of bushfire smoke on respiratory health. *Respirology* **2011**, *16*, 198–209. [CrossRef]
2. Dennekamp, M.; Straney, L.D.; Erbas, B.; Abramson, M.J.; Keywood, M.; Smith, K.; Sim, M.R.; Glass, D.C.; Del Monaco, A.; Haikerwal, A.; et al. Forest fire smoke exposures and out-of-hospital cardiac arrests in Melbourne, Australia: A case-crossover study. *Environ. Health Perspect.* **2015**, *123*, 959–964. [CrossRef] [PubMed]

3. Melnick, R.S. *Regulation and the Courts: The Case of the Clean Air Act*; Brookings Institution Press: Washington, DC, USA, 2010.
4. Sager, L.; Singer, G. Clean Identification? The Effects of the Clean Air Act on Air Pollution, Exposure Disparities and House Prices. 2022. Available online: https://www.lse.ac.uk/granthaminstitute/wp-content/uploads/2022/05/working-paper-376-Sager-Singer_May-2023.pdf (accessed on 1 May 2023).
5. McClure, C.D.; Jaffe, D.A. US particulate matter air quality improves except in wildfire-prone areas. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 7901–7906. [CrossRef] [PubMed]
6. Johnston, F.H.; Henderson, S.B.; Chen, Y.; Randerson, J.T.; Marlier, M.; DeFries, R.S.; Kinney, P.; Bowman, D.M.; Brauer, M. Estimated global mortality attributable to smoke from landscape fires. *Environ. Health Perspect.* **2012**, *120*, 695–701. [CrossRef] [PubMed]
7. Rappold, A.G.; Stone, S.L.; Cascio, W.E.; Neas, L.M.; Kilaru, V.J.; Carraway, M.S.; Szykman, J.J.; Ising, A.; Cleve, W.E.; Meredith, J.T.; et al. Peat bog wildfire smoke exposure in rural North Carolina is associated with cardiopulmonary emergency department visits assessed through syndromic surveillance. *Environ. Health Perspect.* **2011**, *119*, 1415–1420. [CrossRef] [PubMed]
8. Haikerwal, A.; Akram, M.; Sim, M.R.; Meyer, M.; Abramson, M.J.; Dennekamp, M. Fine particulate matter ($PM_{2.5}$) exposure during a prolonged wildfire period and emergency department visits for asthma. *Respirology* **2016**, *21*, 88–94. [CrossRef]
9. Thilakaratne, R.; Hoshiko, S.; Rosenberg, A.; Hayashi, T.; Buckman, J.R.; Rappold, A.G. Wildfires and the changing landscape of air pollution–related gealth burden in California. *Am. J. Respir. Crit. Care Med.* **2023**, *207*, 887–898. [CrossRef]
10. Li, L.; Girguis M.; Lurmann, F.; Pavlovic, N.; McClure, C.; Franklin, M.; Wu, J.; Oman, L.; Breton, C.; Gilliland, F. Ensemble-based deep learning for estimating $PM_{2.5}$ over California with multisource big data including wildfire smoke. *Environ. Int.* **2020**, *145*, 106143. [CrossRef]
11. Romanov, A.A.; Tamarovskaya A.N.; Gusev B.A.; Leonenko, E.V.; Vasiliev, A.S.; Krikunov, E.E. Catastrophic $PM_{2.5}$ emissions from Siberian forest fires: Impacting factors analysis *Environ. Pollut.* **2022**, *306*, 119324. [CrossRef]
12. Ikeda, K.; Tanimoto, H. Exceedances of air quality standard level of $PM_{2.5}$ in Japan caused by Siberian wildfires *Environ. Res. Lett.* **2015**, *10*, 105001.
13. Larsen, A.E.; Reich, B.J.; Ruminski, M.; Rappold, A.G. Impacts of fire smoke plumes on regional air quality, 2006–2013. *J. Expo. Sci. Environ. Epidemiol.* **2018**, *28*, 319–327. [CrossRef] [PubMed]
14. Matz, C.J.; Egyed, M.; Xi, G.; Racine, J.; Pavlovic, R.; Rittmaster, R.; Henderson, S.B.; Stieb, D.M. Health impact analysis of $PM_{2.5}$ from wildfire smoke in Canada (2013–2015. 2017–2018). *Sci. Total Environ.* **2020**, *725*, 138506. [PubMed]
15. Barkjohn, K.; Gantt, B.; Clements, A. Development and Application of a United States wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor. *Atmos. Meas. Tech. Discuss.* **2020**, *2020*, 7304881. [CrossRef]
16. 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. [CrossRef]
17. Wallace, L.; Bi, J.; Ott, W.R.; Sarnat, J.; Liu, Y. Calibration of low-cost PurpleAir outdoor monitors using an improved method of calculating $PM_{2.5}$. *Atmos. Environ.* **2021**, *256*, 118432. [CrossRef]
18. Holder, A.L.; Mebust, A.K.; Maghran, L.A.; McGown, M.R.; Stewart, K.E.; Vallano, D.M.; Elleman, R.A.; Baker, K.R. Field evaluation of low-cost particulate matter sensors for measuring wildfire smoke. *Sensors* **2020**, *20*, 4796. [CrossRef]
19. Kosmopoulos, G.; Salamalikis, V.; Pandis, S.; Yannopoulos, P.; Bloutsos, A.; Kazantzidis, A. Low-cost sensors for measuring airborne particulate matter: Field evaluation and calibration at a South-Eastern European site. *Sci. Total Environ.* **2020**, *748*, 141396. [CrossRef]
20. Durrant-Whyte, H.; Henderson, T.C. Multisensor data fusion. In *Springer Handbook of Robotics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 867–896.
21. Luo, R.C.; Kay, M.G. A tutorial on multisensor integration and fusion. In Proceedings of the IECON'90: 16th Annual Conference of IEEE Industrial Electronics Society, Pacific Grove, CA, USA, 27–30 November 1990, pp. 707–722.
22. Reich, B.J.; Chang, H.H.; Foley, K.M. A spectral method for spatial downscaling. *Biometrics* **2014**, *70*, 932–942. [CrossRef]
23. Warren, J.L.; Miranda, M.L.; Tootoo, J.L.; Osgood, C.E.; Bell, M.L. Spatial distributed lag data fusion for estimating ambient air pollution. *Ann. Appl. Stat.* **2021**, *15*, 323. [CrossRef]
24. Friberg, M.D.; Zhai, X.; Holmes, H.A.; Chang, H.H.; Strickland, M.J.; Sarnat, S.E.; Tolbert, P.E.; Russell, A.G.; Mulholland, J.A. Method for fusing observational data and chemical transport model simulations to estimate spatiotemporally resolved ambient air pollution. *Environ. Sci. Technol.* **2016**, *50*, 3695–3705. [CrossRef]
25. Friberg, M.D.; Kahn, R.A.; Holmes, H.A.; Chang, H.H.; Sarnat, S.E.; Tolbert, P.E.; Russell, A.G.; Mulholland, J.A. Daily ambient air pollution metrics for five cities: Evaluation of data-fusion-based estimates and uncertainties. *Atmos. Environ.* **2017**, *158*, 36–50. [CrossRef]
26. Nguyen, H.; Cressie, N.; Braverman, A. Spatial statistical data fusion for remote sensing applications. *J. Am. Stat. Assoc.* **2012**, *107*, 1004–1018. [CrossRef]
27. Gressent, A.; Malherbe, L.; Colette, A.; Rollin, H.; Scimia, R. Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. *Environ. Int.* **2020**, *143*, 105965. [CrossRef] [PubMed]
28. Datta, A.; Saha, A.; Zamora, M.L.; Buehler, C.; Hao, L.; Xiong, F.; Gentner, D.R.; Koehler, K. Statistical field calibration of a low-cost $PM_{2.5}$ monitoring network in Baltimore. *Atmos. Environ.* **2020**, *242*, 117761. [CrossRef] [PubMed]

29. Lin, Y.C.; Chi, W.J.; Lin, Y.Q. The improvement of spatial-temporal resolution of PM2. 5 estimation based on micro-air quality sensors by using data fusion technique. *Environ. Int.* **2020**, *134*, 105305. [CrossRef]
30. Gao, F.; Masek, J.; Schwaller, M.; Hall, F. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Trans. Geosci. Remote. Sens.* **2006**, *44*, 2207–2218.
31. Hu, D.G.; Shu, H. Spatiotemporal interpolation of precipitation across Xinjiang, China using space-time CoKriging. *J. Cent. South Univ.* **2019**, *26*, 684–694. [CrossRef]
32. Stein, M.L. Statistical methods for regular monitoring data. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2005**, *67*, 667–687. [CrossRef]
33. National Oceanic and Atmospheric Administration. Hazard Mapping System Fire and Smoke Product. Available online: https://www.ospo.noaa.gov/Products/land/hms.html (accessed on 15 October 2022).
34. O'Dell, K.; Ford, B.; Fischer, E.V.; Pierce, J.R. Contribution of wildland-fire smoke to US $PM_{2.5}$ and its influence on recent trends. *Environ. Sci. Technol.* **2019**, *53*, 1797–1804. [CrossRef]
35. Buysse, C.E.; Kaulfus, A.; Nair, U.; Jaffe, N.A. Relationships between particulate matter, ozone, and nitrogen oxides during urban smoke events in the western US. *Environ. Sci. Technol.* **2019**, *53*, 12519–12528. [CrossRef]
36. Barkjohn, K.K.; Holder, A.L.; Frederick, S.G.; Clements, A.L. Relationships between particulate matter, ozone, and nitrogen oxides during urban smoke events in the western US. *Sensors* **2022**, *22*, 9669. [PubMed]
37. California Department of Forestry and Fire Protection. Top 20 Largest California Wildfires. Available online: https://www.fire.ca.gov/our-impact/statistics (accessed on 1 February 2023).
38. Draxler, R.; Rolph, G. *HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) Model Access via NOAA ARL READY*; NOAA Air Resources Laboratory: Silver Spring, MD, USA, 2010; Volume 25. Available online: https://www.ready.noaa.gov/HYSPLIT.php (accessed on 1 May 2023).
39. Su, L.; Yuan, Z.; Fung, J.C.; Lau, A.K. A comparison of HYSPLIT backward trajectories generated from two GDAS datasets. *Sci. Total Environ.* **2015**, *506*, 527–537. [CrossRef] [PubMed]
40. Geyer, C.J. Introduction to Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2011; Volume 20116022, p. 45.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.