Science Advances
AAAS

Supplementary Materials for

Excess of COVID-19 cases and deaths due to fine particulate matter exposure during the 2020 wildfires in the United States

Xiaodan Zhou, Kevin Josey, Leila Kamareddine, Miah C. Caine, Tianjia Liu, Loretta J. Mickley, Matthew Cooper, Francesca Dominici*

*Corresponding author. Email: fdominic@hsph.harvard.edu

Published 13 August 2021, *Sci. Adv.* **7**, eabi8789 (2021)
DOI: 10.1126/sciadv.abi8789

This PDF file includes:

Sections S1 to S3
Figs. S1 to S14
Tables S1 to S5

S.1 Supplementary Text

S.1.1 Mathematical details of the statistical model

Let i denote the county and j denote the calendar day beginning on March 15, 2020. We have $i = 1, 2, \ldots, 92$, where the 92 counties are located California, Oregon, and Washington. For each county, we have daily time series of $\text{PM}_{2.5}$ levels, daily number of COVID-19 cases and deaths, and daily levels of weather variables. The index $j = 1, 2, \ldots, 277$ denotes the calendar days between March 15, 2020, and December 16, 2020.

Likelihood function accounting for overdispersion: Let Y_{ij} be the daily number of COVID-19 cases (or deaths) in county i on day j. We specify the following likelihood function $Y_{ij} \sim \mathcal{NB}(\pi_{ij}, \phi)$ where

$$\pi_{ij} = \frac{\phi}{\phi + (1 - A_{ij})\lambda_{ij}}. \tag{3}$$

The parameter λ_{ij} denotes the expected number of COVID-19 cases (or deaths) on a day j with a positive number of cases in county i. The parameter ϕ allows for overdispersion. In an analogous Poisson model without overdispersion, $\phi = 1$. We also included $A_{ij} \sim \text{Bern}(\psi_i)$ in equation (3) to account for county-specific zero inflation in the observed counts. ψ_i is the county-specific probability of observing a day with zero cases (or deaths).

County-specific regression model for time series data: We assume the following regression model for time series data:

$$\begin{aligned}\log(\lambda_{ij}) = \text{intercept}_i + \sum_{k=0}^{28} \theta_{ik}\text{PM}_{i(j-k)} + \text{day of week}_j + \text{ns}(j, df_1) \\ + \text{ns}(\text{temp}_{ij}, df_2) + \text{ns}(\text{humid}_{ij}, df_2) + \log(\text{pop}_i).\end{aligned} \tag{4}$$

The linear predictor in (4) includes a natural cubic spline basis of order df_1 for calendar days $j = 1, 2, \ldots, 277$. We also include variables that indicate the day of the week. After adjusting the model by the day of the week, the response measure should account for any variability attributable to weekly patterns associated with COVID-19 testing frequency and data reporting behavior. The spline bases $\text{ns}(\text{temp}_{ij}, df_2)$ and $\text{ns}(\text{humid}_{ij}, df_2)$ adjust for the daily temperature and relative humidity that are potentially associated with COVID-19 cases and deaths. The linear predictor in (4) also includes both a random intercept and a population offset $\log(\text{pop}_i)$. The values df_1 and df_2 are selected by fitting a series of models, similar to the one in (3) and (4), for a variety of combinations (df_1, df_2, df_2) and selecting the minimum pairing which stabilizes predictions about the cumulative PM$_{2.5}$ effects. Instead of using MCMC, however, this sequence of models fits with maximum likelihood estimation and assumes a common intercept and a common distributed lag function across counties. The results of this quick diagnostic analysis are featured in Figure S12.

County-specific distributed lag function: We specify a different vector of lag-specific regression coefficients θ_{ik} within each county. We also assume a constrained distributed lag model, where the parameters θ_{ik} are a smooth function of the lag k. More specifically, we assume $\theta_{ik} = \mathbf{U}_{k\cdot}\gamma_i$, where $\mathbf{U}_{k\cdot}$ is the kth row of the basis matrix $\mathbf{U}$ for a natural cubic spline of order four plus an intercept term. We assume $\gamma_{ih} \sim \mathcal{N}(\delta_h, \omega_h^2)$, for $h = 1, 2, 3, 4, 5.$. Note that $E_k(\theta_{ik}) = \eta_k = \mathbf{U}_{k\cdot}\delta$. In sensitivity analyses we also considered natural cubic splines of order 6 plus an intercept term see Figure 5 scenario D.

Prior Distributions: We assume non-informative priors on all the unknown model parameters. More specifically, we assume $\psi_i \sim \mathcal{B}(1, 1)$, $\phi \sim \mathcal{U}(0, 50)$; $\delta_h \sim \mathcal{N}(0, 10^{10})$;

$\omega_h \sim \mathcal{T}^+(2, 10^5)$. Note that $\mathcal{T}^+$ denotes the half-t distribution with two degrees of freedom, which is scaled by 10^5. The remaining parameters are fixed-effect coefficients for the the spline basis functions for calendar days, temperature, and humidity, as well as indicator variables for the day of the week. We assume a flat prior for all of these regression coefficients. Finally, for the county-specific intercept, we assume $\text{intercept}_i \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{N}(0, 10^{10})$ and $\sigma \sim \mathcal{T}^+(2, 10^5)$.

S.1.2 Simulation study

We conducted a simulation study to assess whether the statistical model recovers the values of the unknown parameters. We generated the data to mimic the real data situation. Let $i = 1, 2, \ldots, n$ index $n = 100$ counties and $j = 1, 2, \ldots, m$ index $m = 250$ time points (days). We generate PM$_{2.5}$ measurements PM$_{ij}$ using an autoregressive integrated moving average (ARIMA) mean centered around $8\mu g/m^3$ and variance of 1. The daily counts are generated from the model

$$\begin{aligned} Y_{ij} &\sim \mathcal{NB}\left(\frac{\phi}{\phi + (1 - A_{ij})\lambda_{ij}}, \phi\right) \\ \log(\lambda_{ij}) &= \text{intercept}_i + \sin\left[(j-1)\pi/100\right]/100 + \sum_{k=0}^{l} \theta_{ik} PM_{i(j-k)} + \log(\text{pop}_i). \end{aligned} \tag{5}$$

where $\text{pop}_i \sim \mathcal{U}(10000, 100000)$ and $\text{intercept}_i \sim \mathcal{N}(-10, 2)$ is the random intercept for county $i = 1, 2, \ldots, n$, respectively. We set $\phi = 1.5$. We assume that $A_{ij} \sim \text{Bern}(\psi_i)$ with $\text{logit}(\psi_i) = 10 - \log(\text{pop}_i)$.

The distributed lag coefficients in model (5) are generated assuming

$$\theta_{ik} \sim \mathcal{N}\left[\frac{\log(k+1)\sin(k\pi/4)}{10}, 0.01\right]$$

for $k = 0, 1, 2, \ldots, l$ where $l = 14$. The sinusoidal mean function of the θ_{ik} will generate values that appear to be noisy, when in fact the coefficients roughly follow a known smooth function.

We also specify a sinusoidal curve that relates calendar days to the outcome in an effort to simulate an overall non-linear trend for the outcome of interest concurrent with the lag PM$_{2.5}$ effects.

We fit to the data the BH-ZINB-DL model described above. In the first stage, we assume

$$\log(\lambda_{ij}) = \text{intercept}_i + \text{ns}(j, 4) + \sum_{k=0}^{l} \theta_{ik} \text{PM}_{i(j-k)} + \log(\text{pop}_i). \tag{6}$$

where $\text{ns}(j, 4)$ represents a natural cubic spline function with four degrees of freedom constructed from the vector of sequential calendar days. We fit the data by assuming that: 1) the θ_{ik} are unconstrained; and 2) the θ_{ik} are a smooth function of the lags.

Figure S13 shows the posterior distributions of the lag-specific coefficients (θ_{ik}) from counties $i = 1, 25, 50, 75, 100$. The black dots are the real values, the blue dots are estimates under an unconstrained distributed lag, and the red dots are the estimates under constrained distributed lags. The panel at the bottom right shows the same estimates but pooled across all counties.

S.1.3 Hazard Mapping System smoke product: description and validation

The smoke product generated by NOAA's Hazard Mapping System (HMS) has been used since 2005 as a proxy for ground-level smoke extent and density across North America. Air quality forecasts rely on the HMS smoke product and, more recently, scientific studies have utilized HMS to distinguish smoke from other types of particulate matter (*2,20,31,32*). The system takes advantage of observations from the NOAA Geostationary Operational Environmental Satellites (GOES) (*2, 31, 32*). Using GOES imagery as a visual guide, HMS analysts use GIS software to manually draw the maximum extent of smoke plumes and qualitatively categorize each plume's density, or apparent opacity as light/thin (0-10 $\mu g/m^3$), medium (10-21 $\mu g/m^3$), or heavy/thick (21-32 $\mu g/m^3$).

First, we validate the HMS smoke product by comparing HMS smoke plumes against the NOAA Integrated Surface Data (ISD; https://www.ncdc.noaa.gov/isd/data-access) collected at four airports in California, Oregon, and Washington. More specifically, we calculate the accuracy of the HMS smoke product when used as a binary proxy for the presence of ground-level smoke on a daily basis during the fire season (here defined as July-November) over 2007-2019 and also for the year 2020. Through our validation efforts, we increase confidence in the utility of HMS to represent ground-level smoke.

Smoke observations are retrieved for the ISD by automated sensors at each airport. We first aggregate the hourly airport data to daily scale by assigning each day with a binary label: smoke or no smoke. We define a smoke day as a day when more than 5% of the hourly observations at a given airport include smoke. This criterion limits the impact of days with spurious smoke observations that may be confused with haze or dust. For HMS data, we define smoke days using the HMS densities as criteria. We test three methods with HMS smoke days defined as days with at least one observation of (1) heavy smoke, (2) either medium or heavy smoke, or

(3) any density of smoke, including light smoke. We then compare how HMS smoke days align with airport smoke days for each of the three methods differentiating smoke from non-smoke days with HMS data.

Table S4 shows a comparison of average HMS smoke days compared to airport smoke days from 2007-2019, using these different methods of differentiating smoke days from non-smoke days. We find the best accuracy between the two datasets when HMS smoke days are defined by heavy density observations, though this definition may result in a conservative estimate of smoke days. When averaged across the four airports, the accuracy with this method is 95%, where accuracy is defined as the number of days categorized as smoke days or non-smoke days by both HMS and the airport. The high accuracy is misleading, however, as its calculation is skewed toward true negatives (non-smoke days) rather than true positives (smoke days). Thus, we also calculated two other metrics, Cohen's kappa and Matthew's correlation coefficient, which account for agreement beyond chance. When averaged across airports, Cohen's kappa is 0.2 and Matthew's correlation coefficient is 0.23, which suggests a rather weak agreement between airport data and HMS smoke days. This highlights the uncertainty in categorizing days as smoke or non-smoke based on the satellite-derived HMS product, but can also indicate errors or biases in the airport smoke data.

Table S5 shows the cross-tabulation of smoke and non-smoke days defined using airport versus HMS data, averaged across the four airports. Spatially, we also find greater agreement between HMS smoke plumes and ground-level measurements in Portland, OR, and Seattle, WA, than in the two California sites, Redding and Los Angeles. In summary, our long-term validation using the four airports shows that using all categories of HMS smoke to define smoke days leads to the highest true positive rate but introduces more false positives, or days categorized as non-smoke by airport data but smoke by HMS (Table S5). In contrast, using heavy HMS smoke to define smoke days maximizes overall accuracy and minimizes false-positive rates.

Second, for the year 2020 and when we use only heavy HMS smoke, we found that difference in magnitude of airport- and HMS-derived smoke days is minimized across all airports in the western U.S. (CA, OR, WA). The median of the absolute difference in airport- and HMS-derived smoke days is as follows for each HMS smoke definition: heavy-only (6 days), medium and heavy (22 days), and all densities (46 days) (Figure S14). In particular, incorporating light and medium HMS smoke often results in severe overestimation of smoke days. The correlation between airport data and HMS smoke days is similar for the three definitions of HMS smoke days (r = 0.61 to 0.63).

Both HMS and airport data are subject to errors and limitations. For example, HMS smoke polygons are drawn by NOAA analysts based on daytime GOES satellite images, which are spatially coarse (0.5 - 1 km). Errors arise at smoke polygon edges and from heterogeneity within polygons. Moreover, smoke seen by satellites in certain cases may be aloft and therefore would not impact surface air quality. On the other hand, airport sensors may confuse smoke with other sources of pollution, such as dust or haze.

S.2 Figs. S1 to S14

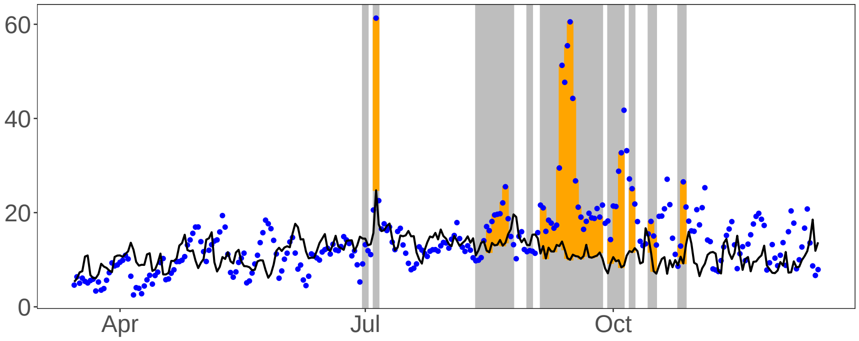

Figure S1: Daily increase in PM$_{2.5}$ ($\mu g/m^3$) attributable to a wildfire day for Los Angeles County, CA. The blue dots are daily PM$_{2.5}$ levels for 2020 and the black lines are median of daily PM$_{2.5}$ levels for the same days and in the same county but for the years 2016-2019. The vertical grey bars identify the wildfire days. When the 2020 values are higher than historical median values, the gap is colored in orange.

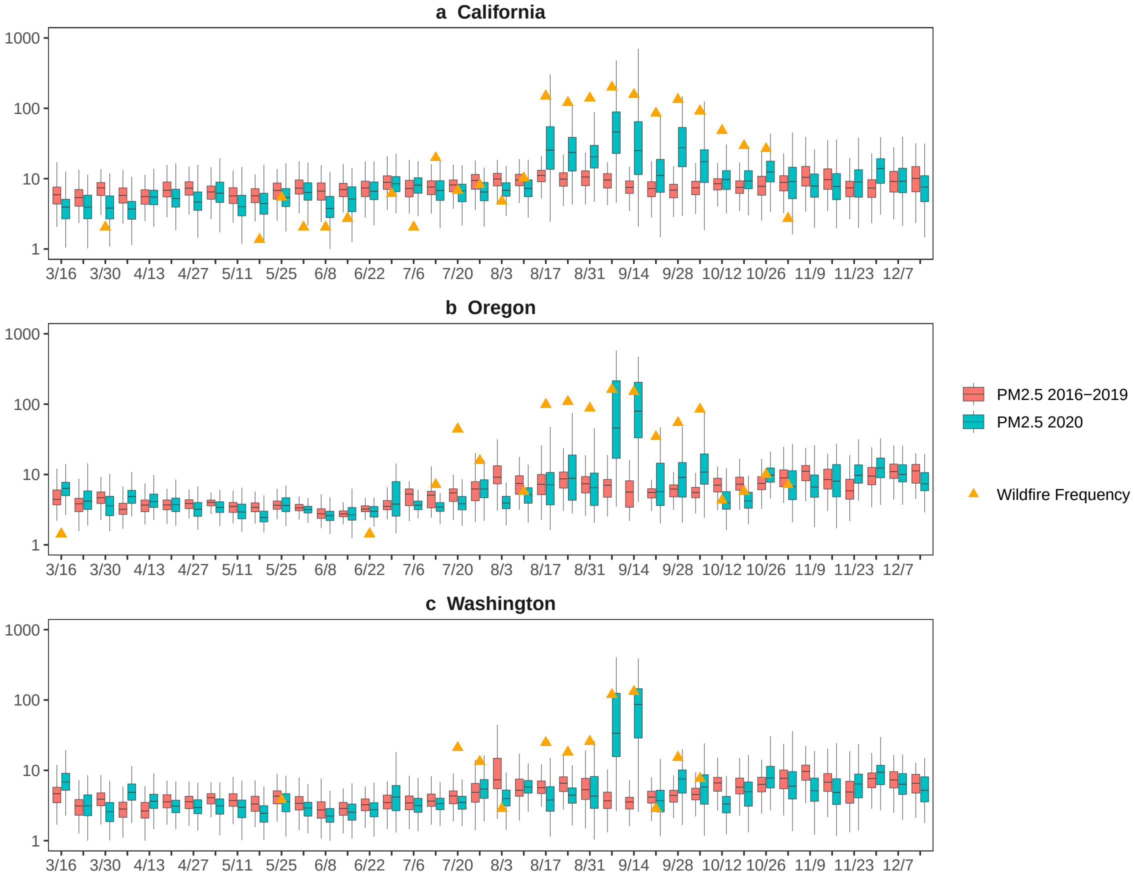

Figure S2: Boxplots of the distribution of the weekly levels of $\text{PM}_{2.5}$ ($\mu g/m^3$) for all counties combined and separately for each state (**a California**, **b Oregon**, **c Washington**). Green boxplots are the weekly values for the period from March 15, 2020, to December 16, 2020, and the red boxplots are the counterpart for 2016-2019. The height of the orange triangles denotes the percentage of counties-days that has wildfires (value zero not shown).

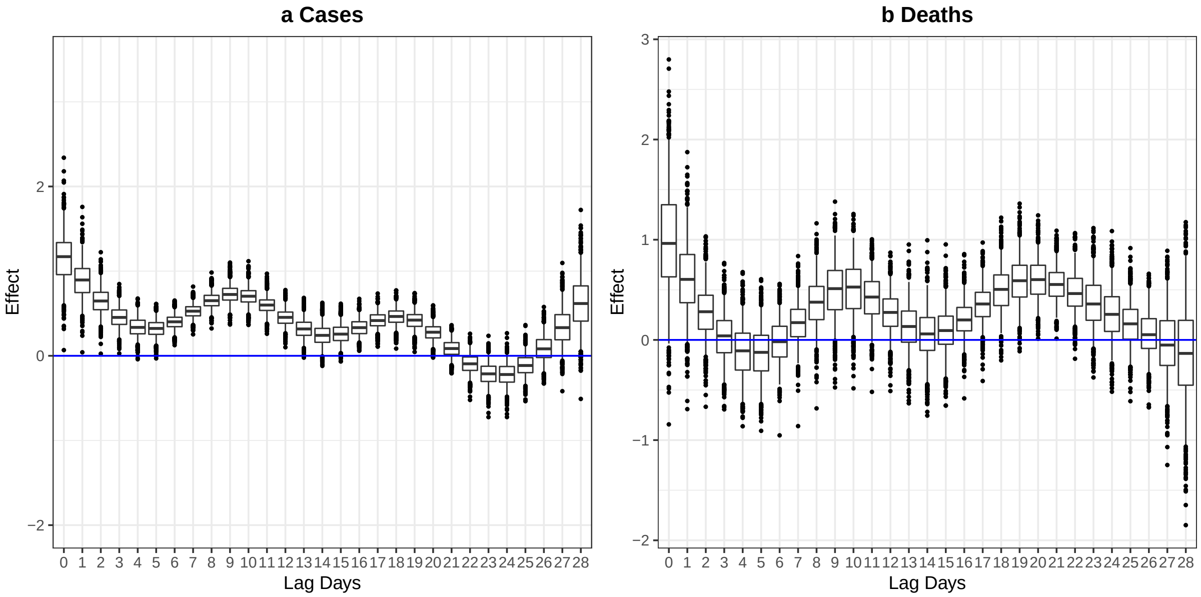

Figure S3: Posterior distributions of the lag-specific effects pooled across all counties. These are posterior distribution of percentage increase in COVID-19 cases (**a Cases**) and COVID-19 deaths (**b Deaths**) associated with a daily increase of 10 $\mu g/m^3$ in PM$_{2.5}$ separately for each lag, up to lag 28.

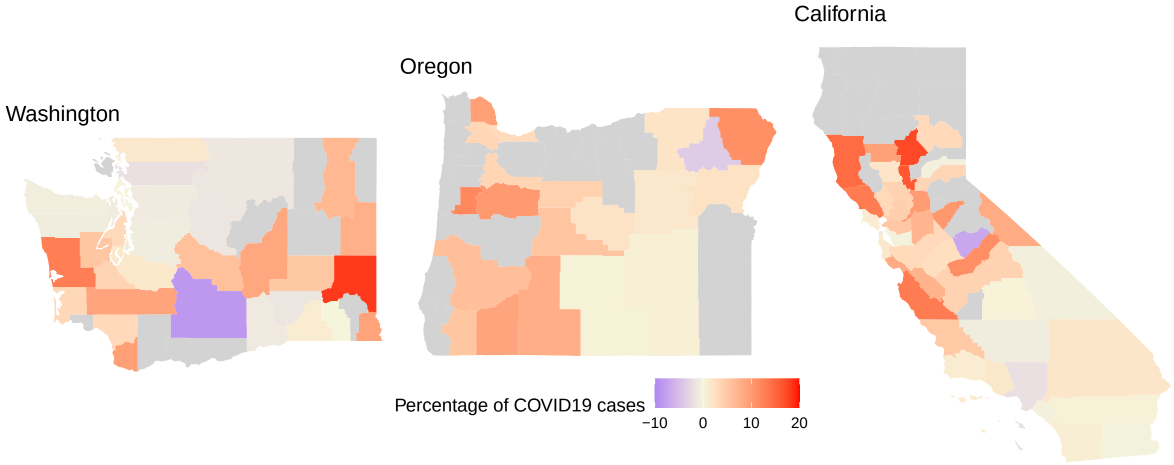

Figure S4: Percentage of total COVID-19 cases attributable to observed high levels of $PM_{2.5}$ on wildfire days.

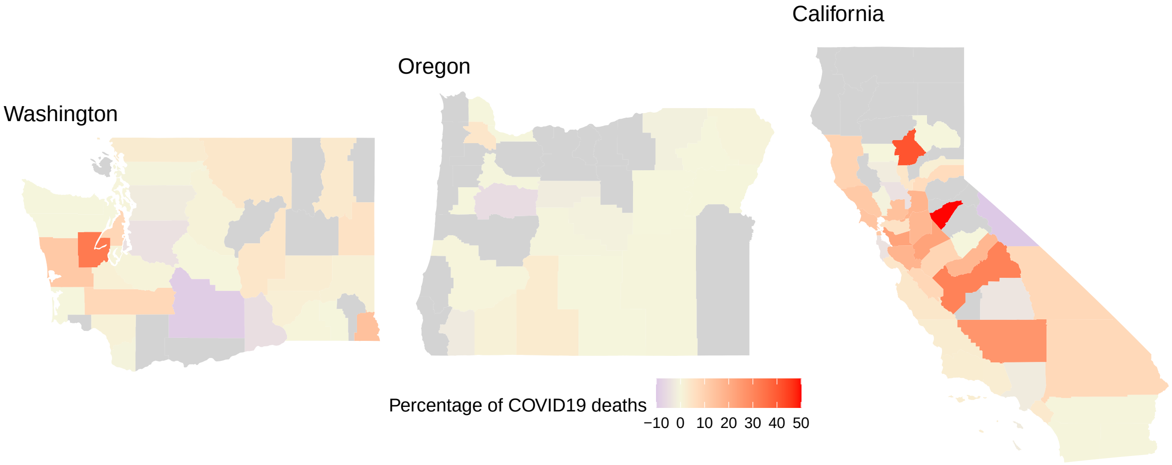

Figure S5: Percentage of total COVID-19 deaths attributable to observed high levels of $PM_{2.5}$ on wildfire days.

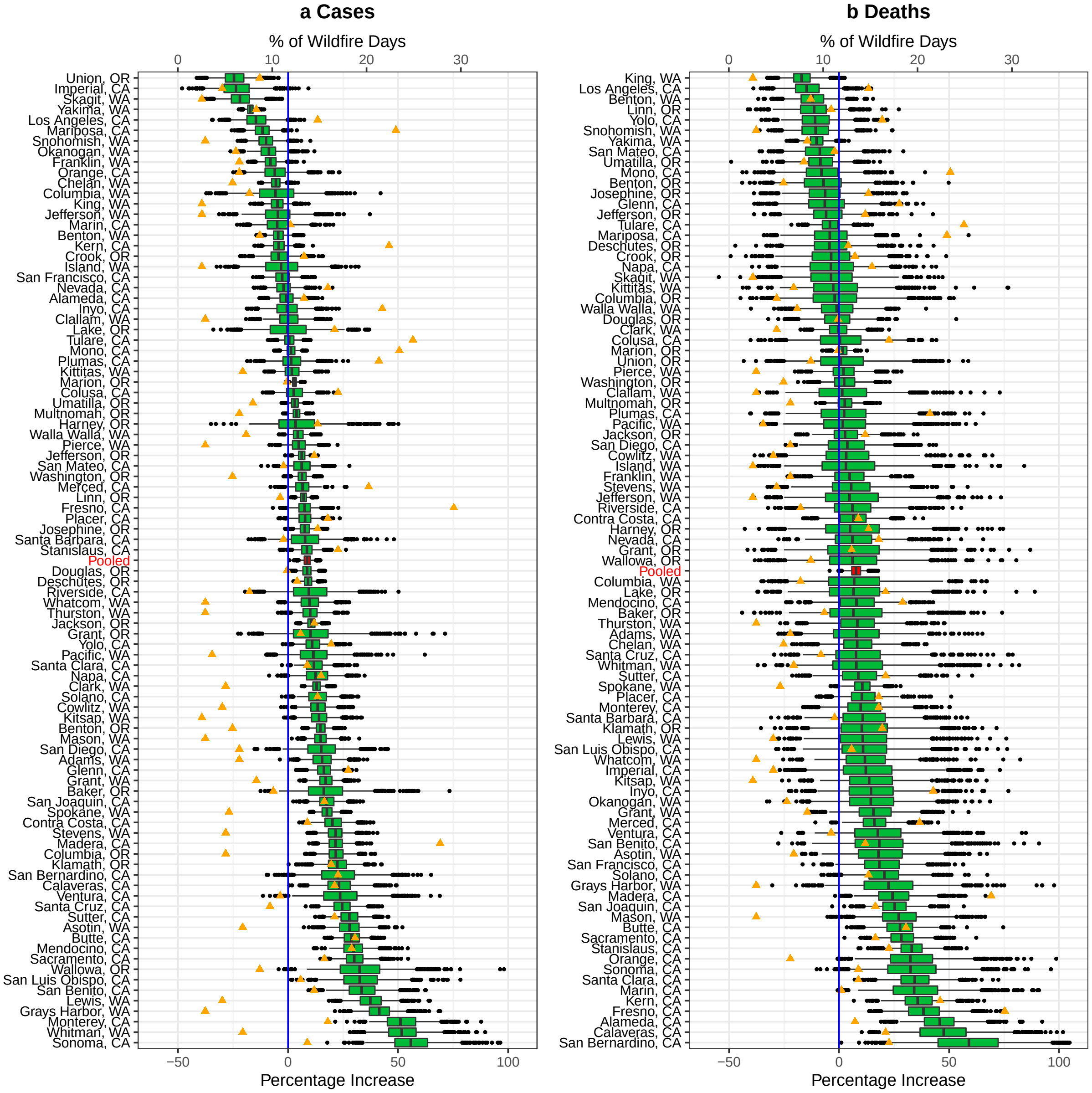

Figure S6: Posterior distributions of the cumulative effects at three weeks: the percentage increase in COVID-19 cases (**a Cases**) and COVID-19 deaths (**b Deaths**) associated with a daily increase of 10 $\mu g/m^3$ in PM$_{2.5}$ for 21 subsequent days, cumulatively up to three weeks. The orange dots represent the percentage of wildfire days (out of 277 days in the study period) for each county.

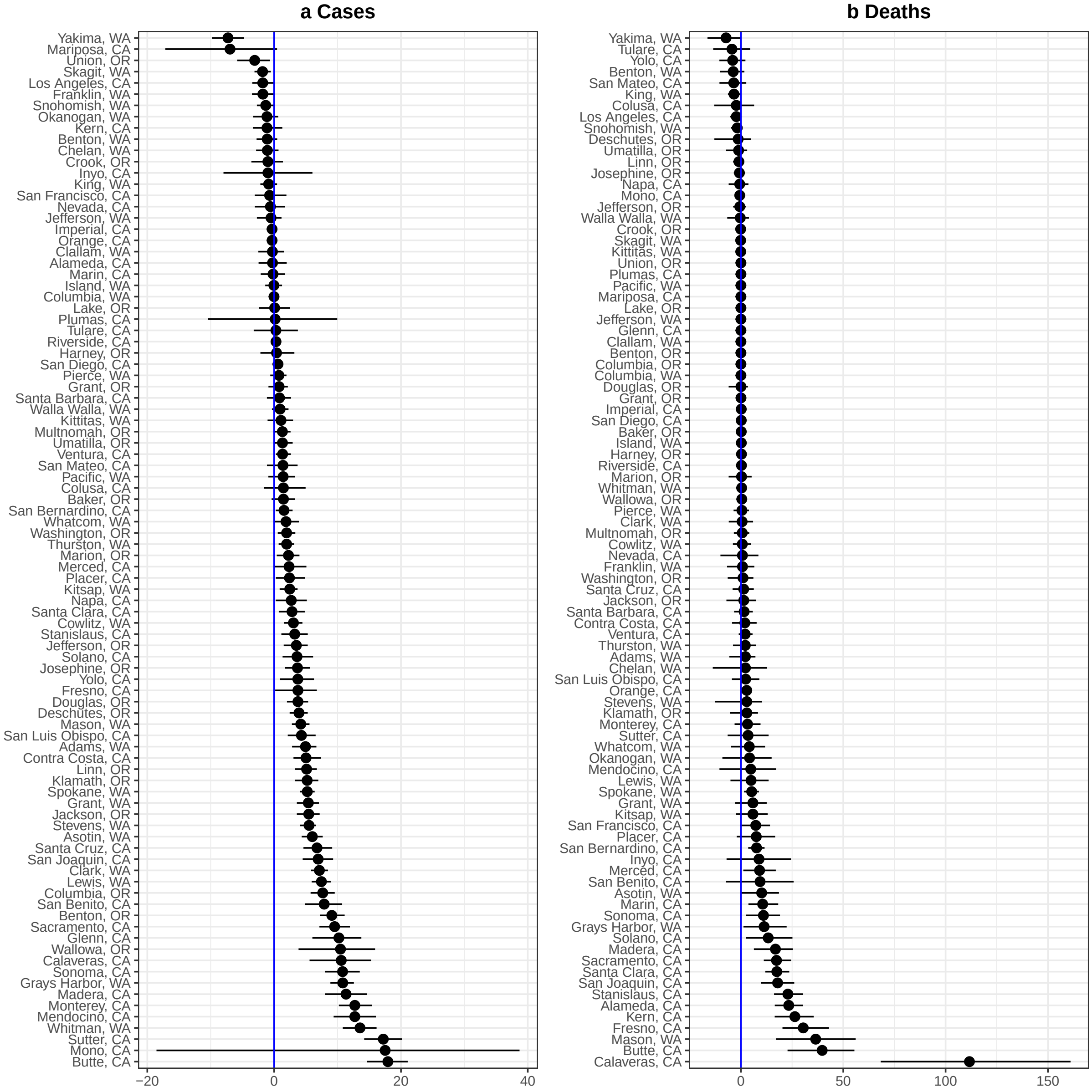

Figure S7: Percentage of total COVID-19 cases (**a Cases**) and deaths (**b Deaths**) attributable to the high levels of $PM_{2.5}$ on wildfire days using 21 lag days during the period from March 15 to November 26, 2020. The error bars represent the 95% CI.

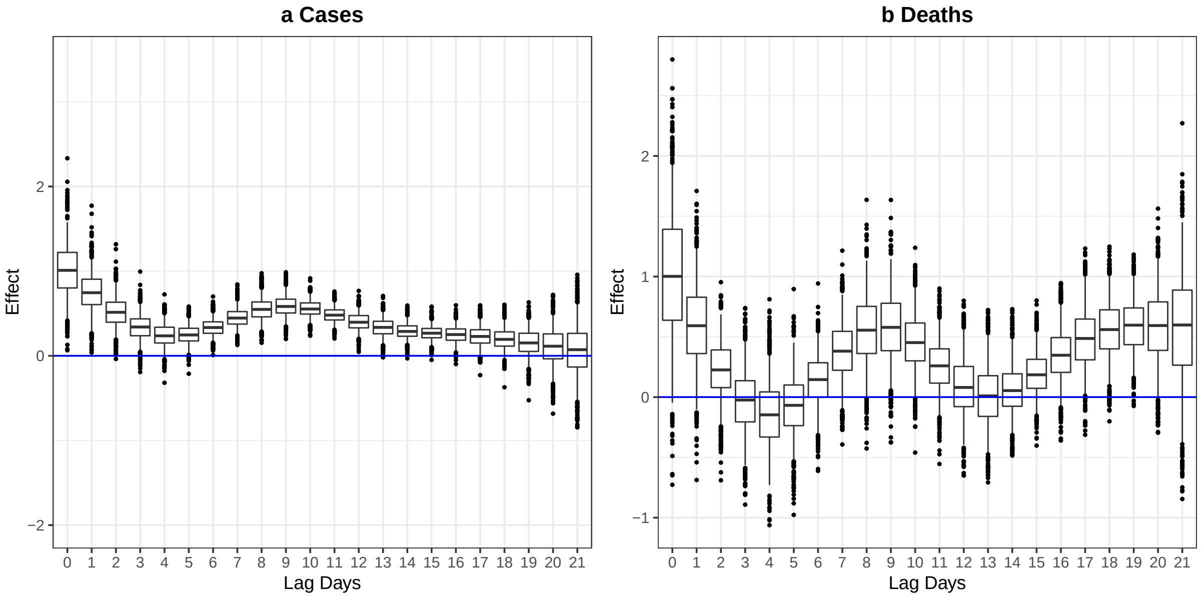

Figure S8: Posterior distributions of the lag-specific effects pooled across all counties. These are posterior distribution of percentage increase in COVID-19 cases (**a Cases**) and COVID-19 deaths (**b Deaths**) associated with a daily increase of 10 $\mu g/m^3$ in PM$_{2.5}$ separately for each lag up to lag 21.

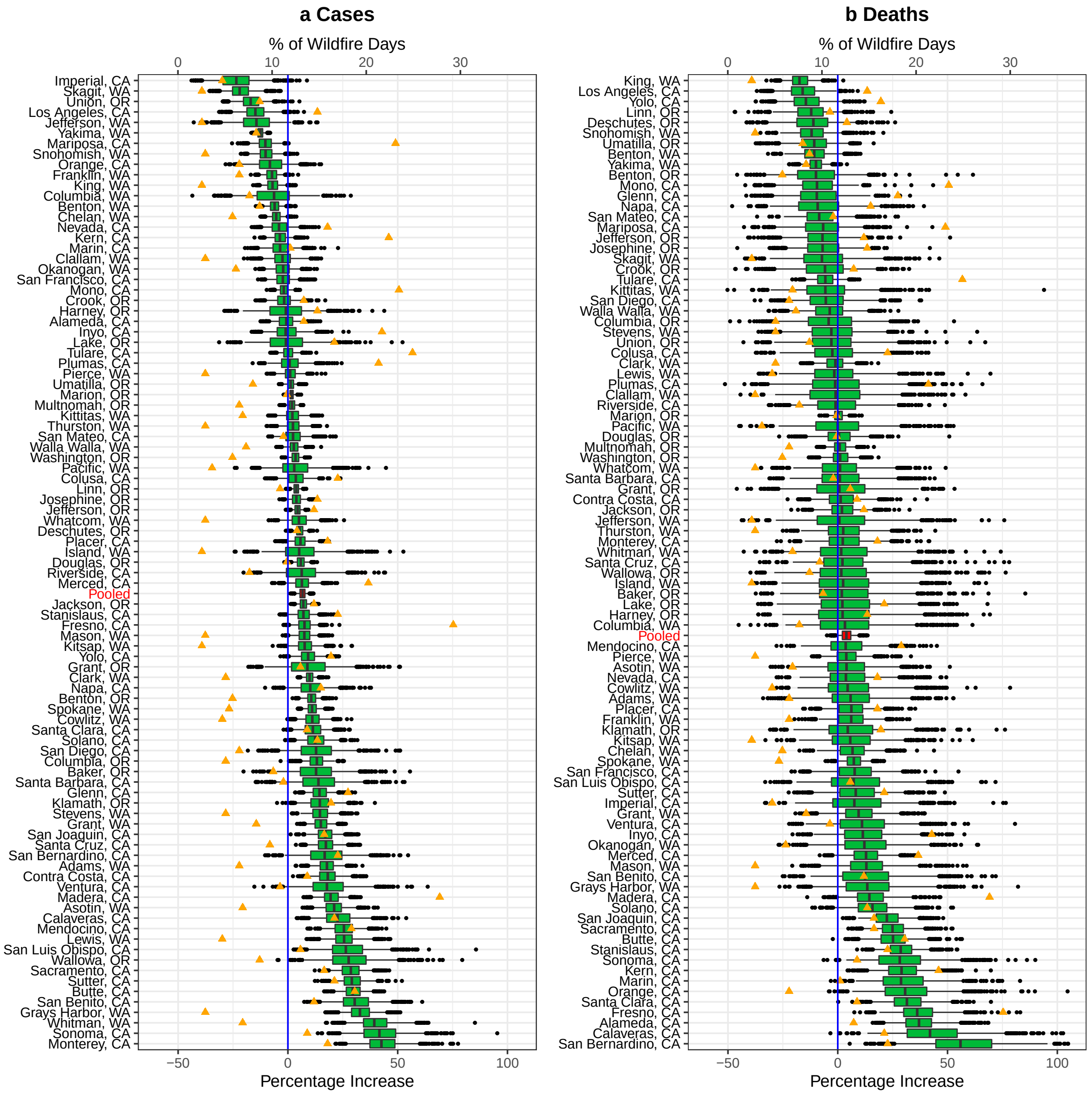

Figure S9: Posterior distributions of the cumulative effects at two weeks: the percentage increase in COVID-19 cases (**a Cases**) and COVID-19 deaths (**b Deaths**) associated with a daily increase of 10 $\mu g/m^3$ in PM$_{2.5}$ for 14 subsequent days, cumulatively up to two weeks. The orange dots represent the percentage of wildfire days (out of 277 days in the study period) for each county.

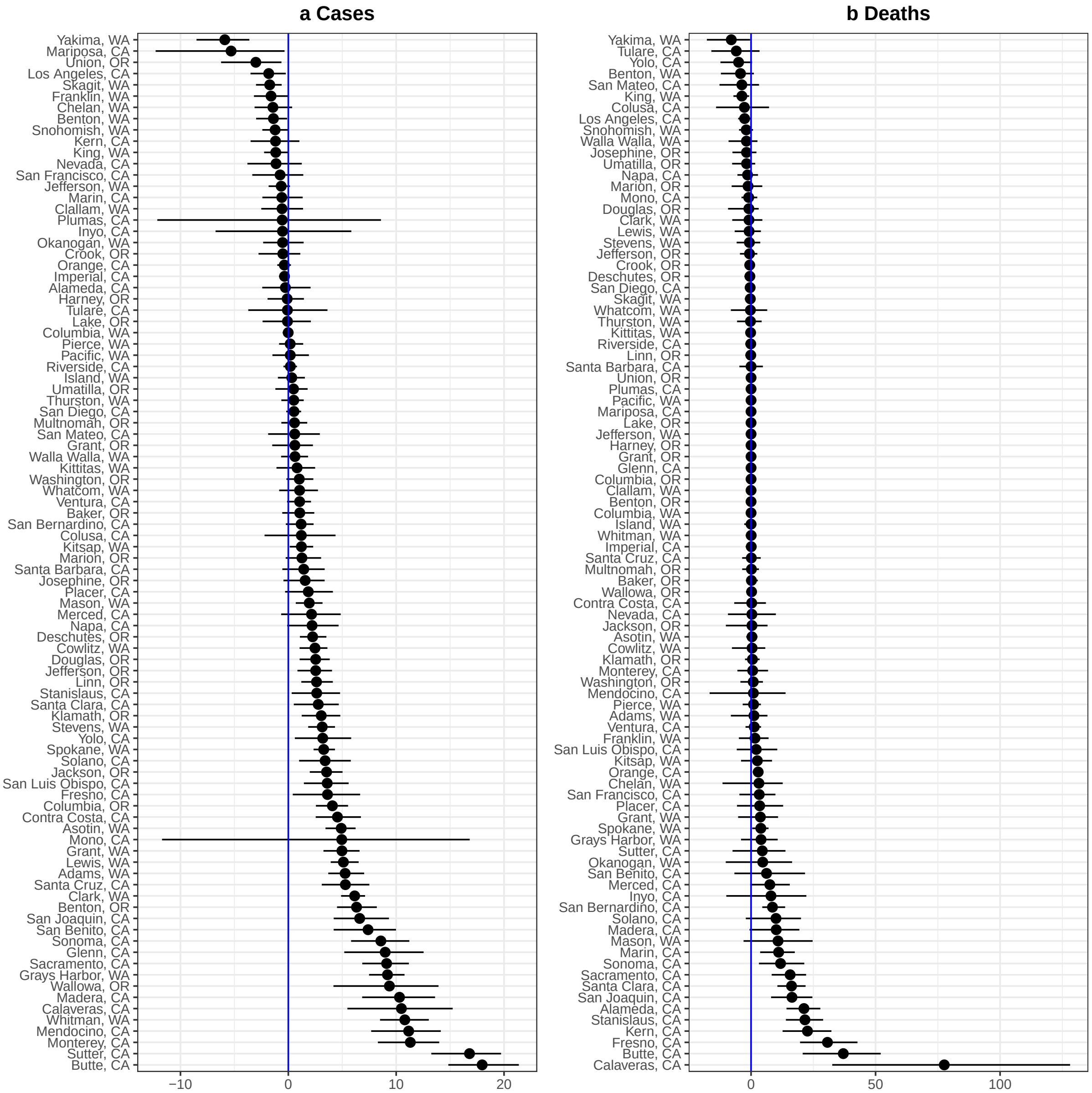

Figure S10: Percentage of total COVID-19 cases (**a Cases**) and deaths (**b Deaths**) attributable to the high levels of $PM_{2.5}$ on wildfire days using 14 lag days during the period from March 15 to November 26, 2020. The error bars represent the 95% CI.

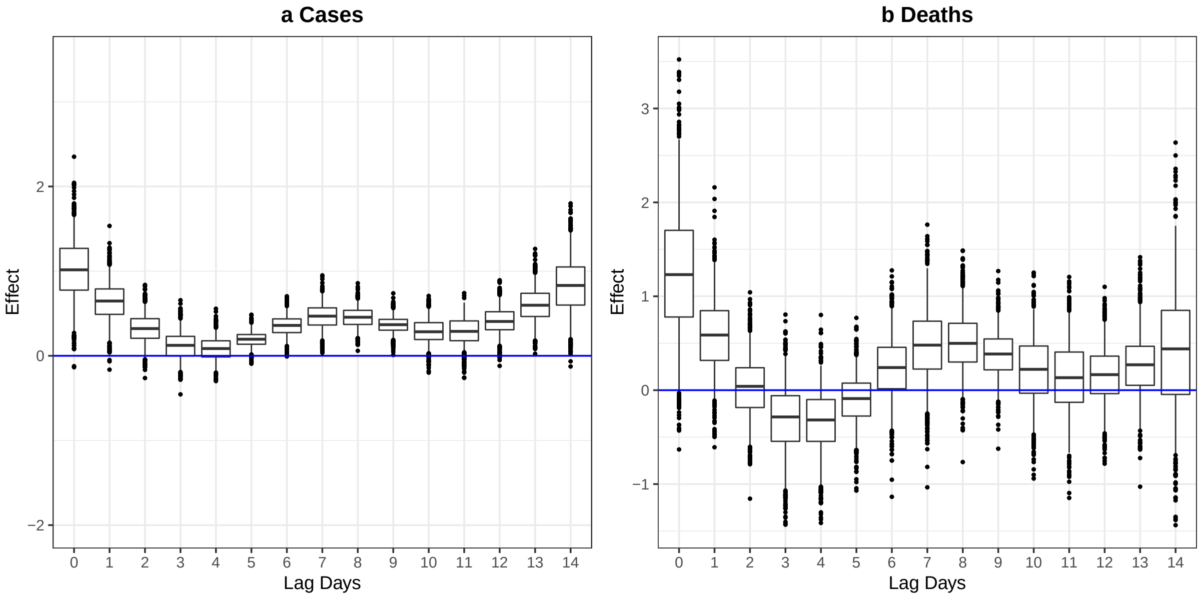

Figure S11: Posterior distributions of the lag-specific effects pooled across all counties. These are posterior distribution of percentage increase in COVID-19 cases (**a Cases**) and COVID-19 deaths (**b Deaths**) associated with a daily increase of 10 $\mu g/m^3$ in PM$_{2.5}$ separately for each lag up to lag 14.

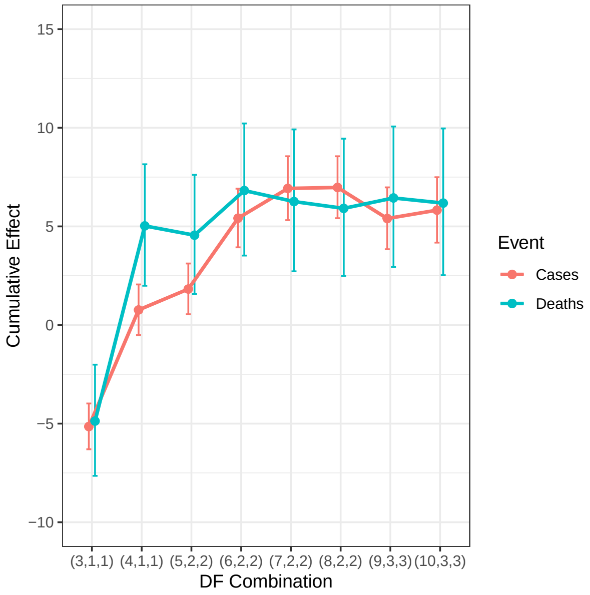

Figure S12: Point estimates and 95% CI of the percentage increase in COVID-19 cases (deaths) associated with a 10 $\mu g/m^3$ increase in $\mathrm{PM}_{2.5}$ for 28 subsequent days, cumulatively across lag days and pooled across counties. We are plotting the posterior means and 95% CI obtained using different approaches for adjusting for seasonality and long-term trends. More specifically, we varied the degrees of freedom for calendar time, temperature and humidity from (3, 1, 1) to (10, 3, 3) respectively. In our main analysis, we chose (6, 2, 2).

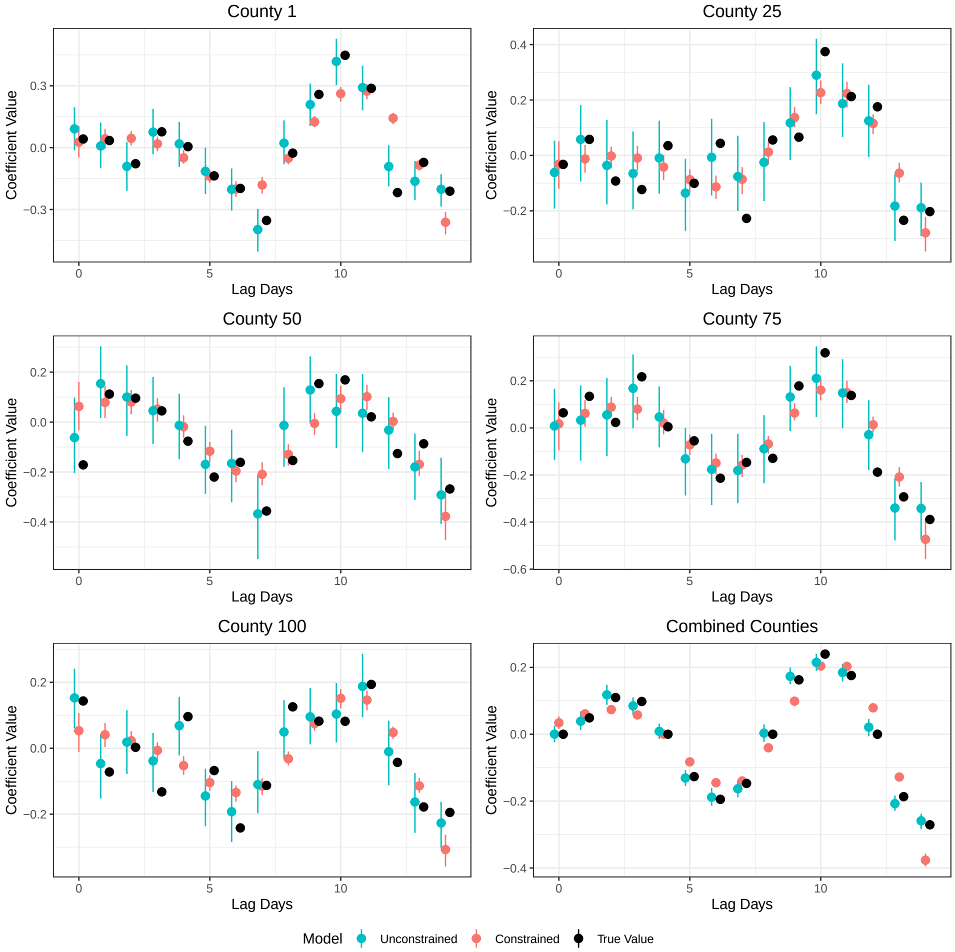

Figure S13: Estimates of θ_{ik} from counties $i = 1, 25, 50, 75, 100$. The error bars represent the 95% CI.

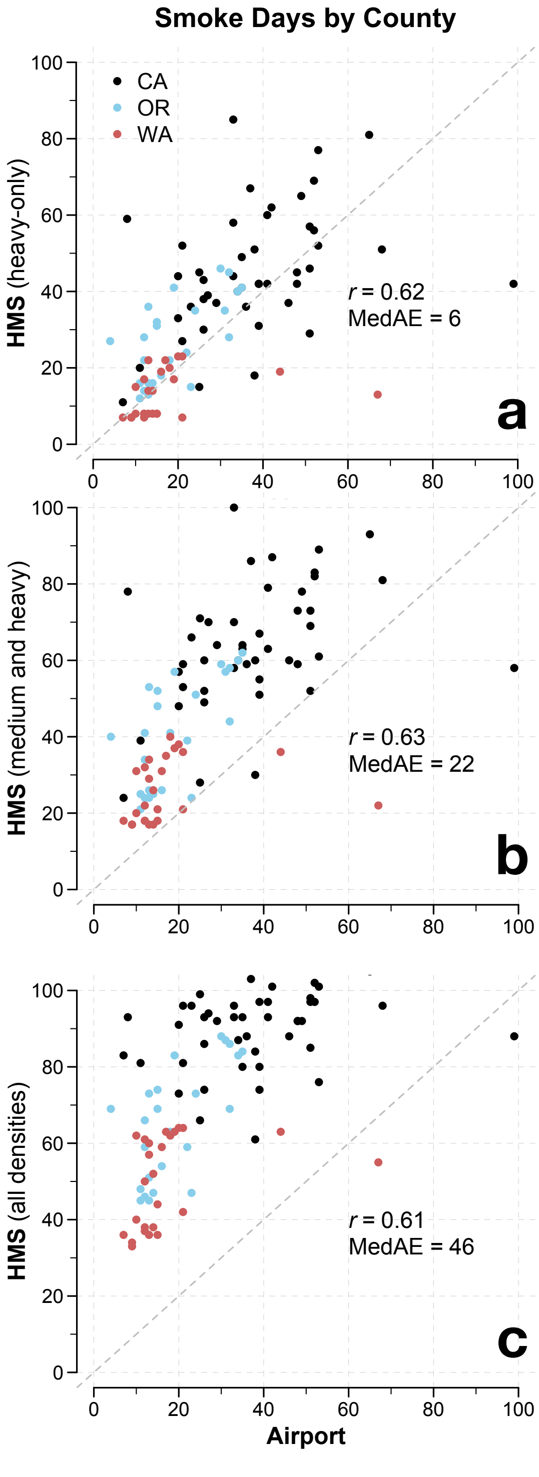

Figure S14: HMS and airport-derived smoke days by county in the western U.S. from July-November, 2020. Three definitions of HMS smoke days are used to estimate smoke days: (**a**) only heavy smoke, (**b**) medium and heavy smoke, and (**c**) all densities (light, medium, and heavy smoke). Inset are the correlation coefficients (r) and the median absolute error (MedAE) between HMS and airport-derived smoke days.

S.3 Tables S1 to S5

Description	**Formulation**
County- and lag-specific effect of PM$_{2.5}$ on the log-linear rate of daily COVID-19 cases (or deaths).	$\theta_{ik} = \mathbf{U}_{k\cdot}\gamma_i$
County-specific percentage increase in COVID-19 cases (or deaths) associated with a 10 $\mu g/m^3$ increase in PM$_{2.5}$ for each lag day.	$100\% \times [\exp(10 \times \theta_{ik}) - 1]$
County-specific percentage increase in COVID-19 cases (or deaths) associated with a 10 $\mu g/m^3$ increase in PM$_{2.5}$ for 28 subsequent days, cumulatively after four weeks.	$100\% \times \left[\exp\left(10\sum_{k=0}^{28}\theta_{ik}\right) - 1\right]$
Lag-specific effect of PM$_{2.5}$ on the log-linear rate of daily COVID-19 cases (or deaths) pooled across counties.	$\eta_k = \mathbf{U}_{k\cdot}\delta$
Percentage increase in COVID-19 cases (or deaths) associated with a 10 $\mu g/m^3$ increase in PM$_{2.5}$ for each lag day, pooled across counties.	$100\% \times [\exp(10 \times \eta_k) - 1]$
Percentage increase in COVID-19 cases (or deaths) associated with a 10 $\mu g/m^3$ increase in PM$_{2.5}$ for 28 subsequent days, cumulatively at four weeks, and pooled across counties.	$100\% \times \left[\exp\left(10\sum_{k=0}^{28}\eta_k\right) - 1\right]$

Table S1: Description and mathematical formulations of the key parameters of interest.

	Sources	Data
Exposure	AirNow.gov (API) NOAA Hazard Mapping System	Daily observed $PM_{2.5}$ Daily smoke density
Health	USAFacts.org	Daily COVID-19 cases and deaths
Meteorological	GRIDMET	Daily maximum temperature Daily maximum relative humidity
Confounders	USAFacts.org Facebook	County population Daily traffic volume as a percent of Feb 2020 (mobility)

Table S2: Publicly available data sources

	Cumulative Pooled Effect for Cases (95% CI)	$\Pr > 0$	Cumulative Pooled Effect for Deaths (95% CI)	$\Pr > 0$
A	11.7 (8.2, 16.0)	1.000	8.4 (2.1, 15.4)	0.996
B	8.7 (4.9, 12.6)	1.000	7.7 (1.3, 13.1)	0.997
C	6.6 (3.2, 9.8)	1.000	4.0 (-0.8, 9.7)	0.932
D	9.9 (6.6, 14.0)	1.000	6.2 (-0.2, 12.9)	0.967
E	11.6 (8.0, 15.4)	1.000	7.9 (1.6, 14.6)	0.993
F	13.3 (9.2, 17.3)	1.000	3.4 (-3.2, 10.0)	0.839
G	11.6 (7.6, 15.7)	1.000	8.2 (1.4, 14.7)	0.995
H	11.7 (7.7, 15.8)	1.000	8.1 (1.4, 14.4)	0.993

Table S3: Cumulative effects across all lags and pooled across counties. Posterior means and 95% CI of the percentage increase in COVID-19 cases (or deaths) associated with a $10\mu g/m^3$ increase in PM$_{2.5}$ for 28 subsequent days, cumulatively at four weeks and pooled across counties. Results are shown under different scenarios of sensitivity analysis A - Original analysis; B - Decreasing the number of lag days under consideration from four to three weeks along with a decrease in number of spline bases for approximating the distributed lag function from six to five; C – Decreasing the number of lag days in consideration to two weeks along with a decrease in the number of natural spline basis functions to four; D - Adjusting for mobility data, which consequently omits six counties; E - Increasing the number of spline basis functions for the distributed lag function from six to eight; F - More aggressively adjusting for temperature, humidity, and seasonality trends; G - Omitting Calaveras, CA, and Mono, CA, counties; H - Dropping adjustments for the day of week.

City	Airport	HMS*		
		Heavy	Medium/Heavy	Light/Medium/Heavy
Los Angeles, CA	13	1	2	15
Seattle-Tacoma, WA	6	3	9	28
Portland, OR	3	2	7	30
Redding, CA	13	5	17	44

Table S4: Average smoke days during the fire season (July-November) from 2007-2019 inferred from airport and HMS data at four cities. *For HMS, the subcategories refer to the three definitions for smoke days based on HMS densities: only heavy smoke, only medium and heavy smoke, and all densities (light, medium, heavy)

	Heavy smoke only		
	HMS-Smoke	HMS- No Smoke	
Airport- Smoke	20	84	19.2%
Airport- No Smoke	17	1869	99.1%
	54.1%	95.7%	94.9%
	Heavy and medium smoke only		
	HMS- Smoke	HMS- No smoke	
Airport- Smoke	36	68	34.6%
Airport- No Smoke	68	1817	96.4%
	34.6%	96.4%	93.2%
	All smoke categories		
	HMS- Smoke	HMS- No Smoke	
Airport- Smoke	57	47	54.8%
Airport- No Smoke	293	1593	84.5%
	16.3%	97.1%	82.9%

Table S5: Confusion matrices of smoke days from 2007-2019, defined using HMS versus airport data, averaged across four airports: Portland, OR, Seattle, WA, Los Angeles and Redding, CA. For each airport condition (smoke or no smoke), the tables report the total number of smoke or non-smoke days that different categories of the HMS product agree or disagree with that condition. These categories are “heavy”, “medium”, and “light”, based on plume opacity. “All” refers to HMS smoke in all three categories.