# Machine learning-derived daily wildfire and non-wildfire $PM_{2.5}$ concentration estimates over the western US, 2008-2018

Colleen Reid[2*], Ellen Considine[1], Melissa M Maestas[1], Gina Li[1]

September 30, 2019

1. Cooperative Institute for Research in Environmental Sciences, Earth Lab and 2. Department of Geography, University of Colorado Boulder, Boulder, Colorado, USA *corresponding author(s): Colleen Reid (Colleen.Reid@Colorado.edu)

**Abstract**

Fine particulate matter ($PM_{2.5}$) levels are declining in many areas of the US due to policies and enforcement of the Clean Air Act. However, in much of the western US, $PM_{2.5}$ concentrations have been increasing, likely due to the increased presence of wildfires in this region. There is growing evidence of various health impacts of $PM_{2.5}$ exposures, even at levels below the federal standard. Health studies of $PM_{2.5}$ in the western US are limited by spatial sparseness of monitoring data. To improve population exposure assessment of $PM_{2.5}$, researchers are increasingly using statistical methods to "blend" information from multiple data sources to better estimate $PM_{2.5}$ in space and time. Some studies have created daily fine-resolution estimates of $PM_{2.5}$ for the whole US, but they perform poorly in the western US. We have tailored a machine learning model to the western US, combining satellite, meteorological, monitoring, land use and other spatiotemporal data to estimate daily $PM_{2.5}$ estimates at the census tract, ZIP code, and county levels during 2008-2018. Our methods improve upon previous models by: use of a more extensive monitoring station network, which captures more spatial locations and proximity to wildfires; use of ensembles of machine learning algorithms, which have been shown to improve model performance; and coverage of a longer period of time. We are making our data publicly available for use in future studies of the health impacts of fine particulate air pollution in the western US.

## Background & Summary

Fine particulate matter ($PM_{2.5}$) air pollution is increasingly associated with numerous adverse health outcomes including, but not limited to, mortality [1], respiratory and cardiovascular morbidity [20, 16], negative birth outcomes [10], and lung cancer [7]. Although $PM_{2.5}$ concentrations have been declining in many parts of the United States due to policies to limit emissions of air pollutants [6], $PM_{2.5}$ levels have been increasing in parts of the northwestern US [14]. This increase has been shown to be associated with wildfire smoke [14, 15], which can cause $PM_{2.5}$ concentrations that are several times higher than the Environmental Protection Agency's (EPA's) daily $PM_{2.5}$ National Ambient Air Quality Standard (NAAQS) in areas downwind of the wildfires for several days at a time [18].

Estimates of $PM_{2.5}$ concentrations for health studies have traditionally been derived from data from stationary air quality monitors placed in and around populated areas for regulatory purposes. In the US, the EPA's Federal Reference Method (FRM) monitors often only measure every third or sixth day and do not provide enough spatial coverage to obtain a good estimate of the air pollution exposures where every person lives. In fact, most US counties do not contain a regulatory air pollution monitor [3]. Using solely monitoring data in health studies leads to exposure misclassification, which often, but not always, drives effect estimates of the association between air pollution and health towards the null [21].

To improve population exposure assessment of $PM_{2.5}$, epidemiological researchers have increasingly been using methods to estimate $PM_{2.5}$ exposures in the temporal and spatial gaps between regulatory monitors using a data from satellites (such as aerosol optical depth (AOD) or polygons of smoke plumes) or air pollution models [3, 12]) over the past two decades. Each of these data sources has its own benefits and limitations, and researchers are increasingly statistically "blending" information from a combination of data sources to better estimate $PM_{2.5}$ in space and time. Various methods of blending have been used including spatiotemporal regression kriging (e.g., [8]), geographically-weighted regression (e.g., [11]), and machine learning methods (e.g., [17, 9, 4]).

Machine learning methods train large auxiliary datasets, often including satellite AOD, meteorological data, chemical transport model output, and land cover and land use data to provide optimal estimates of $PM_{2.5}$ where people breathe. These models have been implemented in various locations around the world at city, regional, and national scales [2]. Some epidemiological questions can only be addressed in longitudinal studies with large sample sizes. Exposure models with large spatial and temporal domains will help enable such studies. Within the US, Di et al. [4, 5] and Hu et al. [9] have separately used machine learning algorithms to create fine-resolution daily $PM_{2.5}$ estimates for the continental US. These models, however, have performed poorly in the western US [4, 9] and particularly the mountain west [5] compared to the rest of the country. Given the increasing trends in $PM_{2.5}$ concentrations in parts of the western US and the importance of wildfires as a source of $PM_{2.5}$ there, it is important to have a model that is tailored to this region to capture the variability in space and time in this region.

The dataset we describe here improves upon previous daily estimates of $PM_{2.5}$ concentrations from machine learning models in the following ways: (1)

use of a more extensive monitoring station network than used in previous models that captures more spatial locations and also proximity to wildfires, a key driver of $PM_{2.5}$ in the western US, (2) use of an ensemble of machine learning algorithms which have been shown to improve model performance [5], (3) better temporal prediction through the use of a nonlinear function (cosine) on day of year, (4) allowance for different prediction models for fire-affected and non-fire affected days to better capture and predict high $PM_{2.5}$ levels during wildfires, and (5) incorporation of errors in prediction back into daily estimates through spatial interpolation. We are making these data available as daily estimates of $PM_{2.5}$ exposures at census tract, ZIP-code, county scales in a public repository, which the above cited papers have not done, to be used in future studies of the societal impacts of air pollution exposure in the western US, where wildfires are a significant contributor to $PM_{2.5}$ concentrations.

[insert Figure 1: monitor locations (points) and state boundaries]

[insert Table 1: list variables]

# Methods

## Study Area

Our study area includes 11 western US states: Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming.[What other descriptions should we put? - square kilometers? climate zones? topography? other?]

Example of citation: [13]

## $PM_{2.5}$ Measurements

[Write short description of each $PM_{2.5}$ data source.]

We downloaded the 2008-2018 pre-generated daily summary files for $PM_{2.5}$ (88101 and 88502 parameter codes) (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily) as well as the spreadsheet listing all AQS monitors with datums (https://aqs.epa.gov/aqsweb/airdata/aqs_monitors.zip) from the United States Environmental Protection Agency (US EPA).

All available $PM_{2.5}$ data in the Fire Cache Smoke Monitor Archive (https://wrcc.dri.edu/cgi-bin/smoke.pl) was downloaded for the years 2008-2018.

$PM_{2.5}$ data from the Uintah Basin, Utah were provided by Seth Lyman at Utah State University (personal communication).

$PM_{2.5}$ data from the Persistent Cold Air Pool Study (PCAPS) [19] conducted in the Salt Lake Valley, Utah in January–February, 2011 were provided by Dr. Geoff Silcox in Chemical Engineering at the University of Utah

## Predictors

[Write short description of each predictor data set and refer to Table 1]

### Machine learning modelling and mapping

[Write description of ML modelling approach]

### Code availability

[Insert brief description of how to access code on GitHub.] The code was written and annotated in R [version number] and Python [version number] and is available from GitHub [doi citation link]. The key package for implementing the ML model was [caretEnsemble?].

## Data Records

All data are freely available from [repository name, data doi citation]. We provide ... [reference Figure 2]

[insert Figure 2: choropleths at zip code level - 4-panel: a) highest year $PM_{2.5}$, Aug or Sept, b) highest year $PM_{2.5}$, Jan/Feb, c) lowest year $PM_{2.5}$, Aug or Sept, d) lowest year $PM_{2.5}$, Jan/Feb.]

[Insert Table 3: list of files]

## Technical Validation

[Write description of goodness of fit methods/metrics - out-of-bag data, RMSE, R2, models run on subsets of data, etc.]

[Insert Figure 4: a) out-of bag observed $PM_{2.5}$ vs predicted, b) full model observed $PM_{2.5}$ vs predicted, c-j) various subsets of data - oob and full model plots (see figure 5 of example paper)]

[Write discussion about variable importance, possibly referring to the suggested figure of variable importance panel figure. Could make an observation or two about the complexity of the variables, e.g., $PM_{2.5}$ can be highest at highest and lowest temperatures (summer fire season and winter inversions), etc.]

[Thoughts - insert figure of predicted $PM_{2.5}$ vs predictor variable for the 8 (or so) most important variables (panel figure)]

Thoughts: compare to $PM_{2.5}$. Concerned comparing to HMS will take too long?

## Usage Notes

[Write brief description of things the provided code can be adapted to do, such as making plots of specific years, use in health/pollution studies.]

## Acknowledgements

[Write acknowledgements text here.]

## Author contributions

[Write brief description of contribution from each author.]

## Competing interests

The authors declare not competing interests.

## Figures and figures legends

[All figures go here and are referred to in the text]

## Tables

[All tables go here and are referred to in the text - read template text for tables]

## References


[1] Souzana Achilleos, Marianthi-Anna Kioumourtzoglou, Chih-Da Wu, Joel D. Schwartz, Petros Koutrakis, and Stefania I. Papatheodorou. Acute effects of fine particulate matter constituents on mortality: A systematic review and meta-regression analysis. *Environment International*, 109:89–100, 2017.

[2] Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osmar Zaiane, and Alvaro Osornio-Vargas. A systematic review of data mining and machine learning for air pollution epidemiology. *Bmc Public Health*, 17:907, 2017. WOS:000416433100002.

[3] Cole Brokamp, Eric B. Brandt, and Patrick H. Ryan. Assessing Exposure to Outdoor Air Pollution for Epidemiological Studies: Model-based and Personal Sampling Strategies. *The Journal of Allergy and Clinical Immunology*, May 2019.

[4] Q. Di, I. Kloog, P. Koutrakis, A. Lyapustin, Y. Wang, and J. Schwartz. Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ Sci Technol*, 50(9):4712–21, May 2016.

[5] Qian Di, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M. Benjamin Sabath, Christine Choirat, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, Loretta J. Mickley, and Joel Schwartz. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 130:104909, July 2019.

[6] Neal Fann, Sun-Young Kim, Casey Olives, and Lianne Sheppard. Estimated Changes in Life Expectancy and Adult Mortality Resulting from Declining PM2.5 Exposures in the Contiguous United States: 1980-2010. *Environmental Health Perspectives*, 125(9):097003, 2017.

[7] Ghassan B. Hamra, Neela Guha, Aaron Cohen, Francine Laden, Ole Raaschou-Nielsen, Jonathan M. Samet, Paolo Vineis, Francesco Forastiere, Paulo Saldiva, Takashi Yorifuji, and Dana Loomis. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environmental Health Perspectives*, 122(9):906–911, September 2014.

[8] Hongda Hu, Zhiyong Hu, Kaiwen Zhong, Jianhui Xu, Feifei Zhang, Yi Zhao, and Pinghao Wu. Satellite-based high-resolution mapping of ground-level PM2.5 concentrations over East China using a spatiotemporal regression kriging model. *The Science of the Total Environment*, 672:479–490, April 2019.

[9] Xuefei Hu, Jessica H. Belle, Xia Meng, Avani Wildani, Lance A. Waller, Matthew J. Strickland, and Yang Liu. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology*, 51(12):6936–6944, June 2017.

[10] Petra Klepac, Igor Locatelli, Sara Korošec, Nino Künzli, and Andreja Kukec. Ambient air pollution and pregnancy outcomes: A comprehensive review and identification of environmental public health challenges. *Environmental Research*, 167:144–159, 2018.

[11] William Lassman, Bonne Ford, Ryan W. Gan, Gabriele Pfister, Sheryl Magzamen, Emily V. Fischer, and Jeffrey R. Pierce. Spatial and temporal estimates of population exposure to wildfire smoke during the washington state 2012 wildfire season using blended model, satellite, and in situ data. *GeoHealth*, 1(3):106–121, 2017.

[12] Jia C. Liu, Gavin Pereira, Sarah A. Uhl, Mercedes A. Bravo, and Michelle L. Bell. A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. *Environ Res*, 136:120–132, 2015.

[13] Y. Liu, J. A. Sarnat, V. Kilaru, D. J. Jacob, and P. Koutrakis. Estimating ground-level PM2.5 in the eastern United States using satellite remote sensing. *Environ Sci Technol*, 39(9):3269–78, May 2005.

[14] Crystal D. McClure and Daniel A. Jaffe. Us particulate matter air quality improves except in wildfire-prone areas. *Proc Natl Acad Sci U S A*, pages 1–6, 2018.

[15] Katelyn O'Dell, Bonne Ford, Emily V. Fischer, and Jeffrey R. Pierce. The contribution of wildland-fire smoke to US PM2.5 and its influence on recent trends. *Environmental Science & Technology*, January 2019.

[16] Sanjay Rajagopalan, Sadeer G. Al-Kindi, and Robert D. Brook. Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*, 72(17):2054–2070, October 2018.

[17] C. E. Reid, M. Jerrett, M. L. Petersen, G. G. Pfister, P. E. Morefield, I. B. Tager, S. M. Raffuse, and J. R. Balmes. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environ Sci Technol*, 49(6):3887–96, March 2015.

[18] Colleen E. Reid, Ellen M. Considine, Gregory L. Watson, Donatello Telesca, Gabriele G. Pfister, and Michael Jerrett. Associations between respiratory health and ozone and fine particulate matter during a wildfire event. *Environment International*, 129:291–298, August 2019.

[19] Geoffrey D. Silcox, Kerry E. Kelly, Erik T. Crosman, C. David Whiteman, and Bruce L. Allen. Wintertime pm2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley. *Atmospheric Environment*, 46:17 – 24, 2012.

[20] Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian. The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease*, 8(1):E69–74, January 2016.

[21] S.L. Zeger, D. Thomas, F. Dominici, J.M. Samet, J. Schwartz, D. Dockery, and A. Cohen. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect*, 108(5):419–426, 2000.