A model for rapid wildfire smoke exposure estimates using routinely-available data

Sean Raffuse[1], Susan O'Neill[2], and Rebecca Schmidt[3]

[1]Air Quality Research Center, University of California Davis, Davis, CA, United States
[2]Pacific Northwest Research Station, USDA Forest Service, Seattle, WA, United States
[3]Department of Public Health Sciences, MIND Institute, University of California Davis School of Medicine, Davis, CA, United States

Correspondence: Sean Raffuse (sraffuse@ucdavis.edu)

Abstract. Urban smoke exposure events from large wildfires have become increasingly common in California and through the western United States. The ability to study the impacts of high smoke aerosol exposures from these events on the public is limited by the availability of high-quailty, spatially-resolved estimates of aerosol concentrations. Methods for the assigning aerosol exposure often employ multiple data sets that are time consuming and expensive to create and difficult to reproduce. As these events have gone from occasional to nearly annual in frequency, the need for rapid smoke exposure assessments has increased. The rapidfire R package provides a suite a tools for developing exposure assigments using data sets that are routinely generated and publicly available within a month of the event. Specifically, rapidfire harvests official air quality monitoring, satellite observations, meteorological modeling, operational predictive smoke modeling, and low-cost sensor networks. A machine learning approach (random forests regression) is used to fuse the different data sets. Using rapidfire, we produced estimates of ground-level 24-hour average $PM_{2.5}$ over for several large wildfire smoke events in California from 2017-2021. These estimates show excellent agreement with independant measures of $PM_{2.5}$ from filter-based networks.

1 Introduction

Changes in climate in the western United States, and elsewhere, are driving larger, more intense fires with greater smoke impacts on larger populations (Burke et al., 2021), and these trends are projected to continue (Hurteau et al., 2014). The wildfire seasons of 2020 and 2021 produced some of the highest concentrations of $PM_{2.5}$ ever observed in monitoring stations around California, some for several days or weeks. Despite reductions in ambient $PM_{2.5}$ driven by air pollution regulations, areas of the western United States are seeing increasing concentrations (McClure and Jaffe, 2018).

The health impacts of exposure to wildfire smoke are wide-ranging.

{Rebecca's paragraphs here}

Despite these findings, many gaps remain in our understanding of the linkages between wildfire smoke and human health (Black et al., 2017). A critical challenge is in characterizing personal or population exposures during high-intensity events. There are many methods for estimating exposure to ambient pollution, including spatial interpolation of measured values, chemical transport modeling, remote sensing, land-use regression modeling, data fusion and machine learning, and combinations of all of these approaches (e.g., Reid et al. (2015), Zhang et al. (2020), Al-Hamdan et al. (2014), Cleland et al. (2020), Hoek et al. (2008)). The rapidly changing conditions during wildfire smoke events can confound otherwise high-performing approaches (O'Neill et al., 2021). There are several barriers to the adoption of existing methods for exposure assignment. These can include data availability for the study location, data latency, and high-performance computing requirements. The combination of increasing frequency of smoke events and the proliferation of smoke exposure human health studies drives a need for exposure modeling that is rapid and inexpensive.

There has been a rapid proliferation of low-cost sensors for air quality within the past decade. While these sensors do not measure $PM_{2.5}$ with the same fidelity as the regulatory monitoring conducted by federal and local air quality agencies, they represent a new resource for $PM_{2.5}$ assessment with relatively dense spatial coverage. Many low-cost PM sensors operate with similar principles, using a laser to count particles that scatter light in the optical range, with a bias toward accumulation mode particles (PM with aerodynamic diameter < 0.3, Ouimette et al. (2022)). Recent studies have shown the value of incorporating low-cost sensor networks into $PM_{2.5}$ exposure modeling (Bi et al., 2020).

Past work has shown that a data fusion approach that combines ground-based air quality monitors, transport modeling that incorporates wildfire emissions, satellite observations, and meteorological variables can be effective in predicting $PM_{2.5}$ exposure during large wildfire events (Zou et al., 2019, and O'Neill et al. (2021)).

We developed a method a suite of tools for rapidly predicting $PM_{2.5}$ exposure, particularly during wildfire smoke events, using readily available data with low latency (less than one month). The tools are contained within a package written in the R programming language called rapidfire (relatively accurate particulate information derived from information retrieved easily). rapidfire adapts and builds upon the methods of Zou et al. (2019) and O'Neill et al. (2021), replacing retrospective chemical transport modeling and other data sets developed for research with smoke forecast modeling and "off-the-shelf" data sets that are routinely available and easily acquired. A major addition is the incorporation of low-cost sensor data. This paper describes the data sets and algorithms used in the rapidfire package and presents an example case study during four recent extreme wildfire seasons in California.

2 Methods

2.1 Input Data Sets

2.1.1 Official and Temporary Monitoring

Hourly $PM_{2.5}$ observations are available from monitoring stations across the United States via the AirNow network [ref]. Within California, about [number] of monitors were operating during the study period. These permanent monitors are a mixture of federal reference method or federal equivalent method instruments, meaning that they are approved by the US EPA to calculate and report air quality to the public.

During wildfires, temporary monitors are also deployed by several government agencies, such as the California Air Resources Board (CARB), and the USDA Forest Service (USFS). These are mostly [what]. Though they are not as accurate as the AirNow monitors [ref], they are deployed in regions where smoke impacts are significant and permanent monitoring is sparse or absent. The locations of permanent and temporary monitors as of September 1, 2021 is shown in Figure 1. The permanent monitors are concentrated in the coastal and valley regions, while temporary monitors are focused in areas of complex terrain where most wildfires are located.

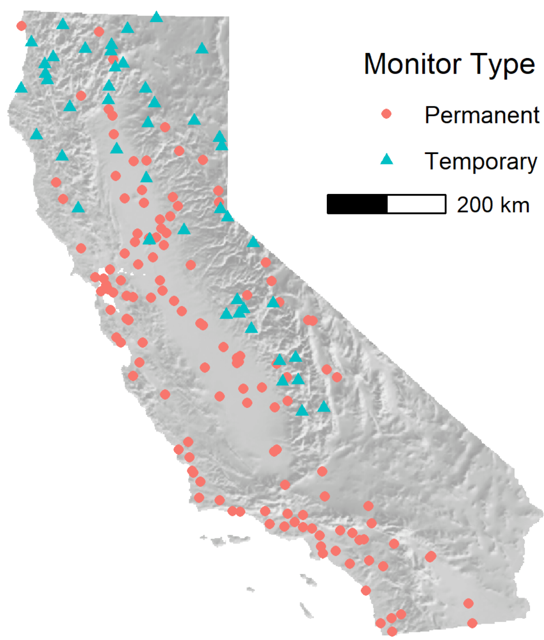

Figure 1. Map of permanent and temporary California monitor locations as of September 1, 2021.

Hourly $PM_{2.5}$ concentrations from both the permanent and temporary monitors were acquired using the `rapidfire::get_airnow_` and `rapidfire::get_airsis_daterange` functions. These wrap the `monitor_subset` function from the `PWFSLSmoke` R package [Mazama Science]. `rapidfire::recast_monitors` was then used to calculate daily 24-hr averages from the hourly data. At least 16 hours are required to produce an average.

The daily average data from both the permanent and temporary monitors were combined into a single data set. The spatial extent of the monitors used in this analysis are shown in Figure [xx]. Portions of this monitor data set were withheld for development and validation of the model. $PM_{2.5}$ observations were log-transformed and interpolated to estimate concentrations at locations away from the monitors using ordinary kriging. 30% of the monitoring data were withheld as test data to develop model variograms using `rapidfire::create_airnow_variograms`.

2.1.2 Smoke Modeling

Air quality models provide ground-level estimates of $PM_{2.5}$ on an output grid. We processed daily average values acquired from the BlueSky Daily Run Viewer (Websky), developed by the USFS AirFire Team. Depending on the event year, different model runs were available. Modeling from Websky was chosen because it is available operationally, is high spatial resolution, and is focused specifically on modeling smoke aerosols from wildland fires; however, other air quality modeling could be substituted. [Susan, can you help me here to describe which models were used and their references?] On some days, the model did not run successfully. For those days, data were backfilled by using the second our third day of a previous day's 72-hr model run.

2.1.3 Satellite Aerosol Optical Depth

Satellite aerosol optical depth (AOD) is a measure of the total columnar aerosol light extinction from the satellite sensor to the ground. AOD is indirectly related to $PM_{2.5}$, with the relationship depending on aerosol type, humidity, and aerosol vertical profile (Li et al., 2015). We used AOD from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) project (**?**). MAIAC is an algorithm that uses time series analysis and additional processing to improve aerosol retrievals, atmospheric correction, and, importantly, cloud detection from the MODerate-resolution Imaging Spectroradiometers (MODIS) onboard NASA's Terra and Aqua satellites. Past work has shown that thick smoke is often mistaken for clouds in the standard MODIS algorithms (van Donkelaar et al., 2011), which hampers their use in wildfire conditions.

The `rapidfire::maiac_download` function can be used to acquire the 1-km daily atmosphere product (MCD19A2) which contains AOD. Clouds prevent the retrieval of AOD, and there are sometimes clouds present even in the hot, dry conditions during California wildfires. The data fusion algorithm requires a complete data set, so a placeholder value must be used to gap-fill in locations under clouds. Previous work has used model-simuluated AOD, along with meteorological variables in a data fusion approach to gap-fill satellite-observed AOD (Zou et al., 2019). For this work, where clouds cover less of the domain, we took a simpler approach. Missing AOD values were filled using a three-stage focal average, available in `rapidfire::maiac_fill_gaps_complete`, and illustrated in Figure 2. In the first stage, a focal mean of a 5-by-5 pixel square (5 km) is used. In the second state, the window is increased to 9-by-9 and to 25-by-25 in the final stage. Any values that are still missing after the final stage are filled with the median value for the entire scene.

2.1.4 Low-cost Sensors

There has been a proliferation of low-cost sensors that estimate $PM_{2.5}$ deployed by the public across the world in the last decade. We used data from the PurpleAir network, which has grown to over 6500 outdoor sensors in California as of 2021. Figure 3 shows the locations of PurpleAir sensors reporting data on September 1, 2021. Coverage in populated areas is exten-

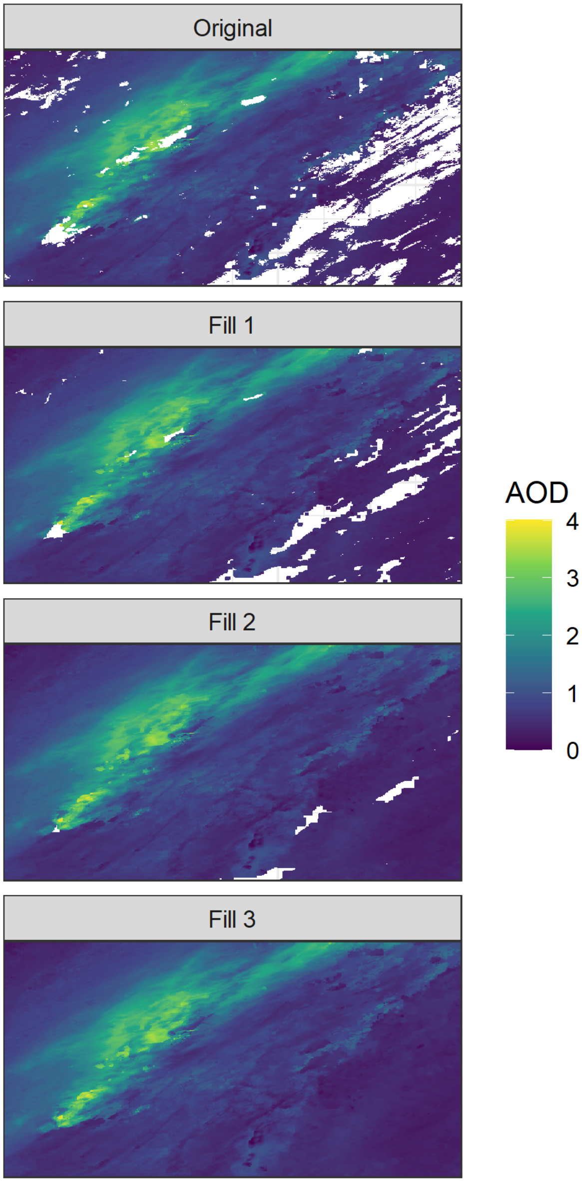

Figure 2. Illustration of MAIAC AOD gap filling

sive.

100 While PurpleAir estimates of $PM_{2.5}$ concentration have been shown to be biased, and are dependent on humidity and aerosol type (Barkjohn et al., 2021), they still correlate with $PM_{2.5}$ observed at FEM monitors and provide invaluable spatial and temporal information that is not available with the relatively sparse network of monitors. Because these sensors are not quality

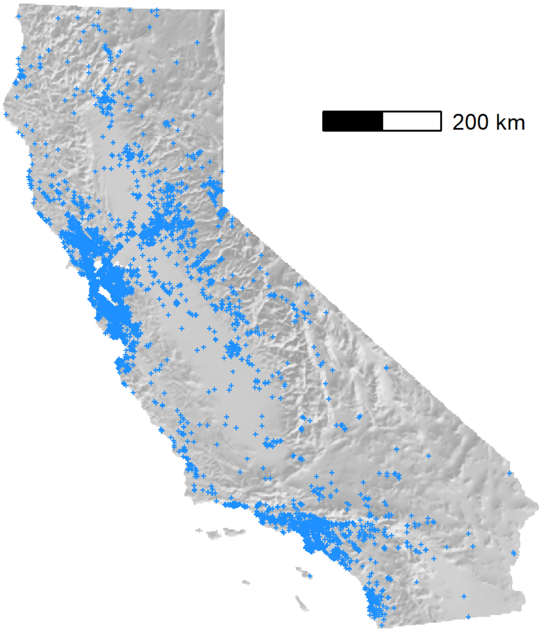

Figure 3. Map of California PurpleAir outdoor sensor locations as of September 1, 2021.

controlled or validated, and their sighting may be suspect, care must be taken when using them in modeling.

rapidfire takes advantage of the AirSensor R package [Mazama Science] for discovering and acquiring PurpleAir sensor data from sensors designated as “outdoor.” `rapidfire::create_purpleair_archive` was used to download and preprocess PurpleAir data from two-channel, 1-minute estimates to single 24-hr average values. The two channels were compared and data were only kept if both values were low ($< 2\,\mu\mathrm{g\,m^{-3}}$) or were within a scaled relative difference (SRD) between channels A and B of 0.5. A daily mean was calculated for both channels and those were then averaged to produce a final daily estimate for the sensor.

$$SRD = \frac{A-B}{\sqrt{2}} / \frac{A+B}{2}. \tag{1}$$

In addition to the channel comparison, we also employed a spatial test to remove sensors that were significantly different from their neighbors. rapidfire::purpleair_clean_spatial_outliers removes any sensors that are more that two standard deviations away from the median of all sites within $10\mathrm{km}$. PurpleAir estimates used in data fusion were log-transformed and then interpolated using ordinary kriging.

2.1.5 Meteorology

Meteorological conditions can help explain the relationships between our inputs and observed $PM_{2.5}$. For example, the PurpleAir sensor is sensitive to relative humidity. AOD is sensitive to humidty and planetary boundary layer height. Following Zou et al. (2019), we included several meteorological variables in our model, including temperature, winds, humidity, boundary layer height, and rainfall. These variables were acquired from the North American Regional Reanalysis (NARR) data set (Mesinger et al., 2006).

2.2 Data Fusion

We developed event specific models using random forests regression (RF). RF is a technique that uses a large number of randomly generated regression trees (Breiman, 2001). Each tree is constructed using a random subset of the training data and each node uses a random subset of the potential predictive variables. New values are estimated as the mean prediction of the individual trees. For each RF run, 500 trees were grown. A single tuning parameter, the number of variables selected at each node, was varied between 2 and 5. The model was trained using 10-fold cross-validation. Internally, `rapidfire::develop_model` uses the randomForest R package.

For the final model, 10 predictor variables were used (Table 1). $PM_{2.5}$ from the monitors was used as both a predictor and a target variable. A random subset of 30% of the monitoring data was withheld for model validation.

Table 1. Predictor variables used in the rapidfire RF model.

Variable	Description
PM25_log_ANK	Log-transformed, interpolated $PM_{2.5}$ from permanent and temporary monitors
PM25_log_PAK	Log-transformed, interpolated $PM_{2.5}$ estimates from PurpleAir sensors
PM25_bluesky	Daily average ground-level $PM_{2.5}$ predictions from BlueSky smoke model
MAIAC_AOD	Gap-filled daily AOD from MAIAC
air.2m	Daily average ambient temperature at 2m above ground level from NARR
uwnd.10m	Daily average u component of wind at 10m above ground level from NARR
vwnd.10m	Daily average v component of wind at 10m above ground level from NARR
rhum.2m	Daily average relative humidity at 2m above ground level from NARR
apcp	Daily total precipitation amount from NARR
hpbl	Daily average height of the planetary boundary layer from NARR

2.3 Model Validation

We developed models for five large wildfire smoke events from 2017-2021 in Northern California (Table 2) and validated the modeling against two data sets of $PM_{2.5}$ observations, 1) the permanent and temporary hourly monitors described above, and 2) 24-hr filter-based measurements from the IMPROVE and CSN networks.

Table 2. Modeled time periods and major Northern California wildfires

Year	Time Period	Major Fires
2017	October	Atlas, Nuns, Pocket, Redwood Valley, Tubbs
2018	July 15 - September 15; November	Carr, Camp
2019	October 15 - November 15	Thomas
2020	August - October	August, Creek, LNU Lightning, North, SCU Lightning
2021	August - October	Antelope, Caldor, Dixie, Monument, River

2.3.1 Hourly Monitors

Model predicted $PM_{2.5}$ values were compared against withheld measurements from the permanent and temporary monitoring networks using rapidfire and three other modeling techniques: 1) ordinary kriging (OK) interpolation of AirNow monitors, 2) OK interpolation of PurpleAir sensors, and 3) multiple linear regression (MLR) using the same inputs as those used for the rapidfire modeling. Comparative model performance metrics are presented in Table 3. For these wildfire events, rapidfire provides good correlation with low error and bias, offering advantages over classical MLR or interpolation of the ground monitors alone. Graphical results of the validation (Figure 4) show the tighter distribution around the 1:1 line for the rapidfire modeling, especially for higher concentrations.

Table 3. Performance metrics for four modeling methods

Model	R^2	RMSE	Median Bias	Normalized Bias	Median Error	Normalized Error
rapidfire	0.74	21.5	0.083	0.76	2.13	18.6
MLR	0.68	23.8	0.056	0.49	2.59	22.6
AirNow OK	0.63	25.7	0.133	1.22	2.63	23.0
PurpleAir OK	0.38	33.3	-0.095	-1.04	3.75	32.8

2.3.2 IMPROVE, CSN, CA FRM

rapidfire results were also compared with available 24-hr integrated filter-based observations from the IMPROVE and CSN networks. They represent a challenging test of the method as they are 100% independent of the model inputs, accurate estimates of $PM_{2.5}$ concentration, and, for IMPROVE especially, located far from other monitors in remote locations with complex terrain. Downsides of using these data are that the networks are more sparse and the days with the highest concentrations are often not available as the IMPROVE sampler can clog in very heavy smoke situations. Paired results are plotted in Figure 5, showing good agreement across the concentration range with a few outliers at IMPROVE sites. Model performance metrics for the filter-based comparison are shown in Table 4. As expected, the CSN sites are better predicted, as they are typically located in urban areas with nearby AirNow monitors and PurpleAir sensors.

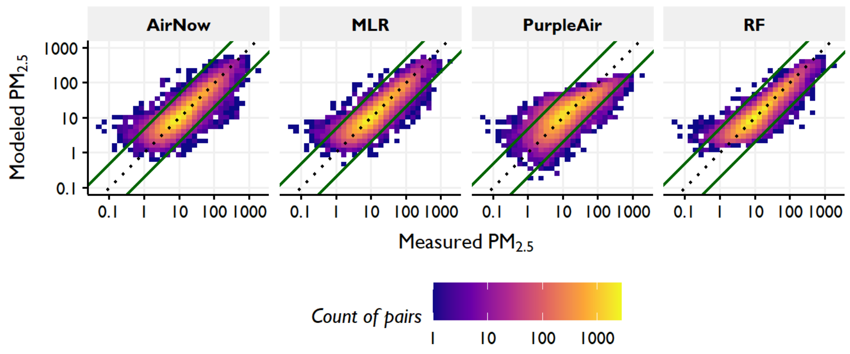

Figure 4. Model comparison against measured PM 2.5 from the IMPROVE and CSN filter networks.

Table 4. Performance metrics for rapidfire at IMPROVE sites, CSN sites, and IMPROVE and CSN sites combined

Network	R^2	RMSE	Median Bias	Normalized Bias	Median Error	Normalized Error
CSN	0.82	5.18	0.42	3.93	1.96	15.3
IMPROVE	0.76	8.47	2.48	46.5	3.19	49.6
Both	0.77	7.52	1.84	22.4	2.70	29.9

3 Results

The results are plotted across California for two wildfire seasons: August - October, 2020 (6) and August - October, 2021 (7). In each case, daily average $PM_{2.5}$ reaches values greater than $200\,\mu g\,m^{-3}$, with very strong spatial and temporal variability. The 2020 case shows three widespread peaks, in August, September, and October. In the 2021 case, concentrations were highest in northern locations in August, while values were higher further south in September and early October. These two cases highlight the complexity of these smoke events, which are controlled by multiple wildfires burning in and around the state.

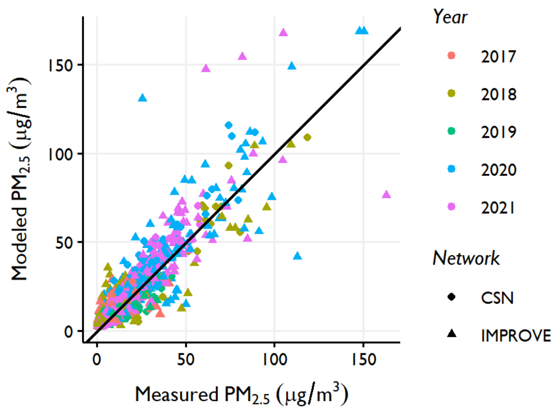

Figure 5. Model comparison against measured PM 2.5 at IMPROVE and CSN monitors

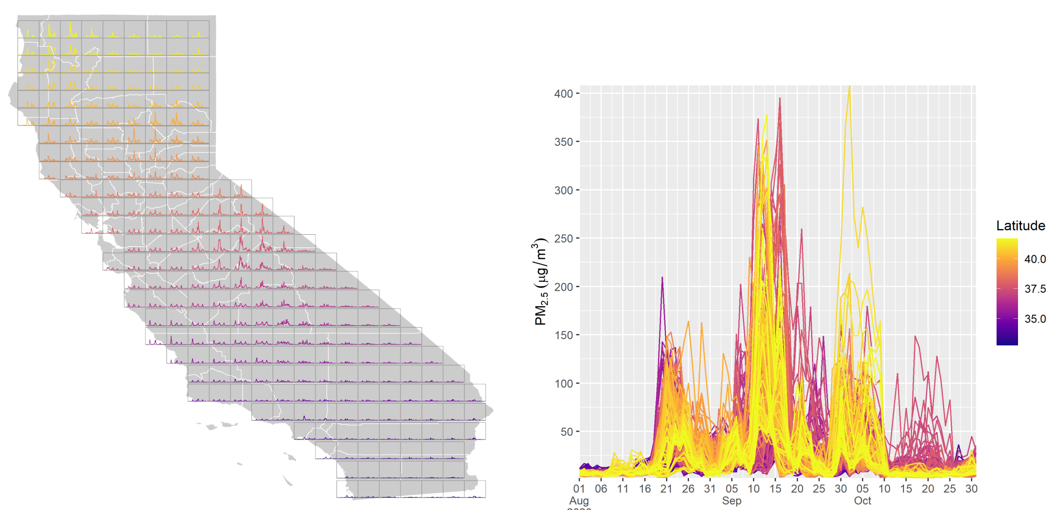

Figure 6. 2020

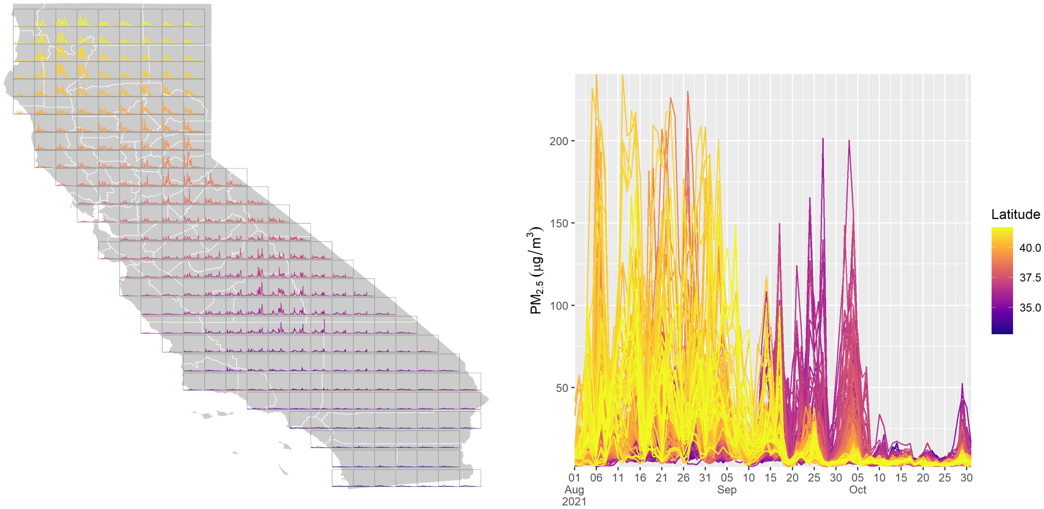

Figure 7. 2021

4 Code?

5 Discussion

5.1 Model input importance

Although the random forest model uses all of the provided predictor variables, the most explanatory variables are selected more often at each node. The relative importance of each variable can calculated by adding noise to each predictor variable in turn and examining the impact on model prediction errors (Breiman, 2001). A plot of relative importance of each variable is shown in . . . .

5.2 Application for health studies

The rapidfire modeling has been applied, and is being applied, in several epidemiological studies, including B-SAFE, WHAT-NOW, others? [need help from Rebecca here. What can we put that has already been published?] These have found interesting things, such as A, B, and C.

5.3 Advantages over existing methods

The primary advantages of the rapidfire methods. - open code - use of production data sets - rapid development - adaptability to other regions, time periods, and input data sets - focus on inputs important for wildfire smoke

5.4 Limitations

- Need for training data
- black box nature of random forests
- current only includes data sets in the US

6 Conclusions

The conclusion goes here. You can modify the section name with `\conclusions[modified heading if necessary]`.

Code and data availability. use this to add a statement when having data sets and software code available

Sample availability. use this section when having geoscientific samples available

Video supplement. use this section when having video supplements available

Appendix A: Figures and tables in appendices

Regarding figures and tables in appendices, the following two options are possible depending on your general handling of figures and tables in the manuscript environment:

A1 Option 1

If you sorted all figures and tables into the sections of the text, please also sort the appendix figures and appendix tables into the respective appendix sections. They will be correctly named automatically.

A2 Option 2

If you put all figures after the reference list, please insert appendix tables and figures after the normal tables and figures.

To rename them correctly to A1, A2, etc., please add the following commands in front of them: `\appendixfigures` needs to be added in front of appendix figures `\appendixtables` needs to be added in front of appendix tables

Please add `\clearpage` between each table and/or figure. Further guidelines on figures and tables can be found below.

Author contributions. Daniel wrote the package. Josiah thought about poterry. Markus filled in for a second author.

Competing interests. The authors declare no competing interests.

Disclaimer. We like Copernicus.

Acknowledgements. Thanks to the rticles contributors!

References

Al-Hamdan, M. Z., Crosson, W. L., Economou, S. A., Jr, M. G. E., Estes, S. M., Hemmings, S. N., Kent, S. T., Puckett, M., Quattrochi, D. A., Rickman, D. L., Wade, G. M., and McClure, L. A.: Environmental public health applications using remotely sensed data, Geocarto International, 29, 85–98, https://doi.org/10.1080/10106049.2012.715209, 2014.

Barkjohn, K. K., Gantt, B., and Clements, A. L.: Development and application of a United States-wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor, Atmospheric Measurement Techniques, 14, 4617–4637, https://doi.org/10.5194/amt-14-4617-2021, 2021.

Bi, J., Wildani, A., Chang, H. H., and Liu, Y.: Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale, Environmental Science & Technology, 54, 2152–2162, https://doi.org/10.1021/acs.est.9b06046, pMID: 31927908, 2020.

Black, C., Tesfaigzi, Y., Bassein, J. A., and Miller, L. A.: Wildfire smoke exposure and human health: Significant gaps in research for a growing public health issue, Environmental Toxicology and Pharmacology, 55, 186–195, https://doi.org/10.1016/j.etap.2017.08.022, 2017.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Burke, M., Driscoll, A., Heft-Neal, S., and Wara, M.: The Changing Risk and Burden of Wildfire in the United States, PNAS, 118, 1–6, https://doi.org/10.1073/pnas.2011048118, 2021.

Cleland, S. E., West, J. J., Jia, Y., Reid, S., Raffuse, S., O'Neill, S., and Serre, M. L.: Estimating Wildfire Smoke Concentrations during the October 2017 California Fires through BME Space/Time Data Fusion of Observed, Modeled, and Satellite-Derived PM2.5, Environmental Science & Technology, 54, 13 439–13 447, https://doi.org/10.1021/acs.est.0c03761, 2020.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D.: A review of land-use regression models to assess spatial variation of outdoor air pollution, Atmospheric environment, 42, 7561–7578, https://doi.org/10.1016/j.atmosenv.2008.05.057, 2008.

Hurteau, M., Westerling, A., Wiedinmyer, C., and Bryant, B.: Projected Effects of Climate and Development on California Wildfire Emissions through 2100, Environmental Science and Technology, 48, 2298–2304, https://doi.org/10.1021/es4050133, 2014.

Li, J., Carlson, B. E., and Lacis, A. A.: How well do satellite AOD observations represent the spatial and temporal variability of PM2.5 concentration for the United States?, Atmospheric Environment, 102, 260–273, https://doi.org/10.1016/j.atmosenv.2014.12.010, 2015.

McClure, C. and Jaffe, D.: US Particulate Matter Air Quality Improves Except in Wildfire-prone Areas, PNAS, 115, 7901–7906, https://doi.org/10.1073/pnas.1804353115, 2018.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W.: North American Regional Reanalysis, Bulletin of the American Meteorological Society, 87, 343 – 360, https://doi.org/10.1175/BAMS-87-3-343, 2006.

O'Neill, S. M., Diao, M., Raffuse, S., Al-Hamdan, M., Barik, M., Jia, Y., Reid, S., Zou, Y., Tong, D., West, J. J., Wilkins, J., Marsha, A., Freedman, F., Vargo, J., Larkin, N. K., Alvarado, E., and Loesche, P.: A multi-analysis approach for estimating regional health impacts from the 2017 Northern California wildfires, Journal of the Air & Waste Management Association, 71, 791–814, https://doi.org/10.1080/10962247.2021.1891994, 2021.

Ouimette, J. R., Malm, W. C., Schichtel, B. A., Sheridan, P. J., Andrews, E., Ogren, J. A., and Arnott, W. P.: Evaluating the PurpleAir monitor as an aerosol light scattering instrument, Atmospheric Measurement Techniques, 15, 655–676, https://doi.org/10.5194/amt-15-655-2022, 2022.

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M., and Balmes, J. R.: Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning, Environmental science & technology, 49, 3887–3896, https://doi.org/10.1021/es505846r, 2015.

van Donkelaar, A., Martin, R. V., Levy, R. C., da Silva, A. M., Krzyzanowski, M., Chubarova, N. E., Semutnikova, E., and Cohen, A. J.: Satellite-based estimates of ground-level fine particulate matter during extreme events: A case study of the Moscow fires in 2010, Atmospheric Environment, 45, 6225–6232, https://doi.org/10.1016/j.atmosenv.2011.07.068, 2011.

Zhang, H., Wang, J., García, L. C., Ge, C., Plessel, T., Szykman, J., Murphy, B., and Spero, T. L.: Improving Surface PM2.5 Forecasts in the United States Using an Ensemble of Chemical Transport Model Outputs: 1. Bias Correction With Surface Observations in Nonrural Areas, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 293, https://doi.org/10.1029/2019JD032293, 2020.

Zou, Y., O'Neill, S. M., Larkin, N. K., Alvarado, E. C., Solomon, R., Mass, C., Liu, Y., Odman, M. T., and Shen, H.: Machine Learning-Based Integration of High-Resolution Wildfire Smoke Simulations and Observations for Regional Health Impact Assessment, International Journal of Environmental Research and Public Health, 16, https://doi.org/10.3390/ijerph16122137, 2019.