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

- California smoke impact on the rise (and throughout the west)
- Expected to continue
- Health impact of wildfire smoke and need for research
- Need for rapid, inexpensive exposure modeling (increasing frequency and intensity of events)
- pedigree (NASA HAQAST and Yufei's work)
- growth of low-cost sensor networks and their fidelity for fires
- Compared with other methods (land use regression), this may be more suitable for fires because of the regional nature of the source

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. Hourly $PM_{2.5}$ concentrations from both the permanent and temporary monitors were acquired using the `rapidfire::get_airnow_daterange` 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`.

– figure of Permanent and Temporary monitors.

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. [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 [ref]. We used AOD from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) project [ref]. 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 [Lyapustin et al, 2011a,b; 2021; 2018]. Past work has shown that thick smoke is often mistaken for clouds in the standard MODIS algorithms [ref], 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 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`. [more to describe the figure]

– figure illustrating the gap filling approach

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. While PurpleAir estimates of $PM_{2.5}$ concentration have been shown to be biased, and are dependent on humidity and aerosol type [ref], they still strongly 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 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 [ref].

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 $2\mathrm{m}$ above ground level from NARR
uwnd.10m	Daily average u component of wind at $10\mathrm{m}$ above ground level from NARR
vwnd.10m	Daily average v component of wind at $10\mathrm{m}$ above ground level from NARR
rhum.2m	Daily average relative humidity at $2\mathrm{m}$ above ground level from NARR
apcp	Daily total precipitation amount from NARR
hpbl	Daily average height of the planetary boundary layer from NARR

We developed models for five large wildfire smoke events from 2017-2021 in Northern California (Table 2). The data

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 Model Training and Validation

2.3.1 IMPROVE, CSN, CA FRM

3 Code?

4 Results

4.1 Model development plots? (Test vs. Training)

4.2 Comparison to filter measurements

4.2.1 crossplots

4.2.2 time series?

4.3 Maps

4.4 Compare to other method (interpolation alone)

5 Discussion

- Importance plots
- application for health studies
- advantages over existing methods
- limitations

6 Content section with citations

See the R Markdown docs for bibliographies and citations.

7 Content section with R code chunks

You should always use `echo = FALSE` on R Markdown code blocks as they add formatting and styling not desired by Copernicus. The hidden workflow results in 42.

You can add verbatim code snippets without extra styles by using ```` ``` ```` without additional instructions.

```
sum <- 1 + 41
```

8 Content section with list

If you want to insert a list, you must

- leave

- empty lines

- between each list item

because the `\tightlist` format used by R Markdown is not supported in the Copernicus template. Example:

```
- leave


- empty lines


- between each list item
```

9 Examples from the official template

9.1 FIGURES

When figures and tables are placed at the end of the MS (article in one-column style), please add

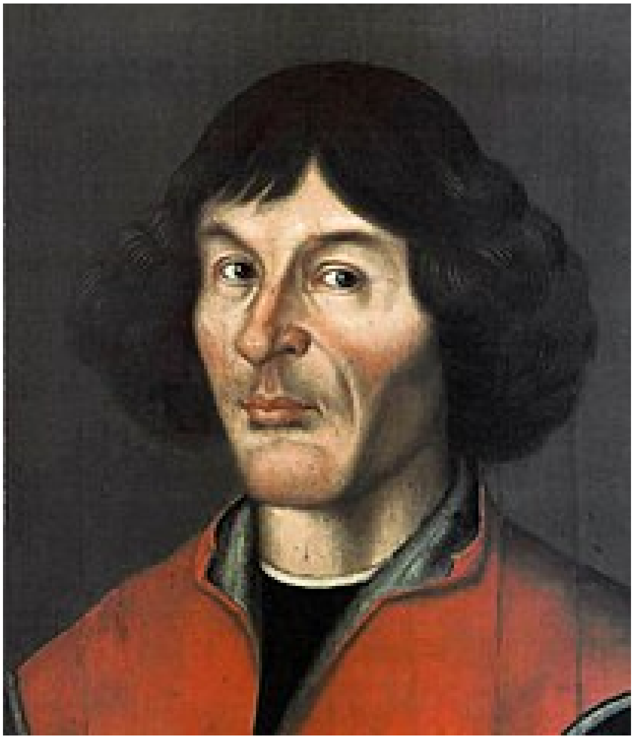

Figure 1. one column figure

between bibliography and first table and/or figure as well as between each table and/or figure.

9.1.1 ONE-COLUMN FIGURES

Include a 12cm width figure of Nikolaus Copernicus from Wikipedia with caption using R Markdown.

9.1.2 TWO-COLUMN FIGURES

You can also include a larger figure.

9.2 TABLES

You can ad LaTeXtable in an R Markdown document to meet the template requirements.

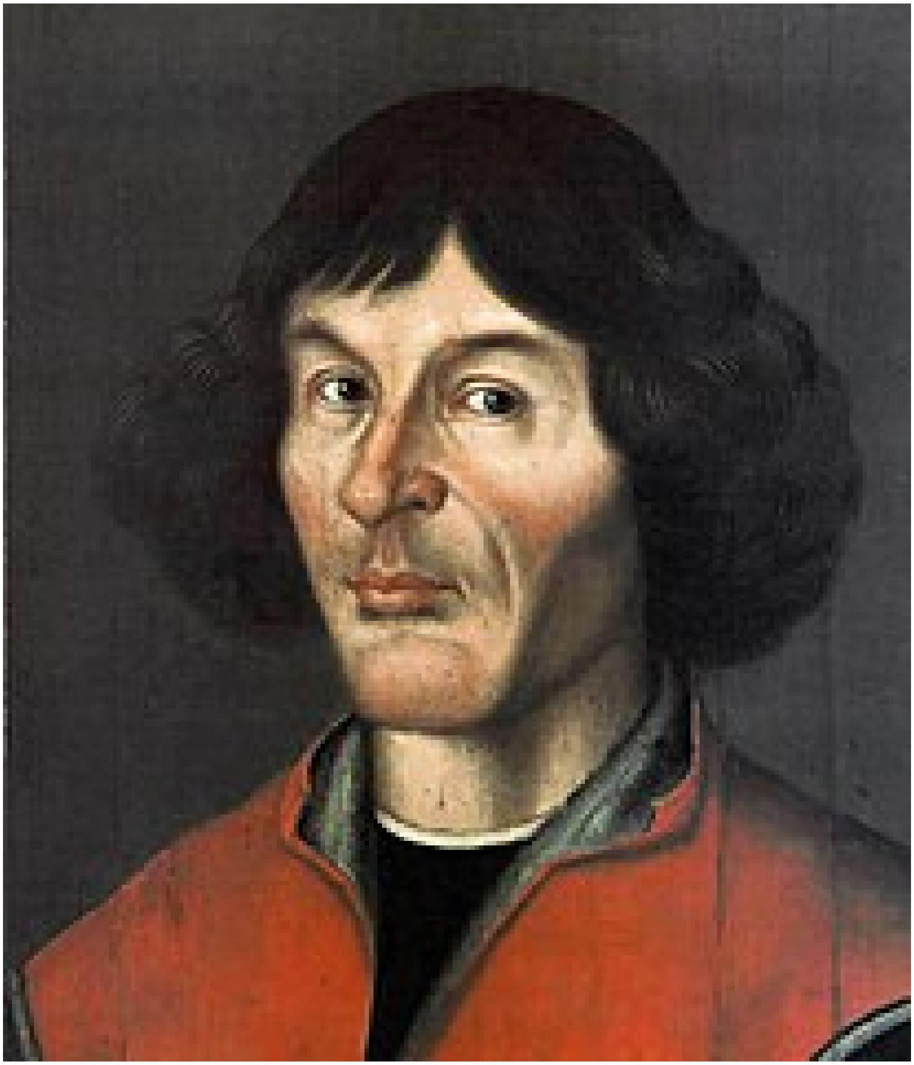

Figure 2. two column figure

Table 3. TEXT

a	b	c
1	2	3

Table Footnotes

Table 4. TEXT

a	b	c
1	2	3

Table footnotes

9.2.1 ONE-COLUMN TABLE

9.2.2 TWO-COLUMN TABLE

9.3 MATHEMATICAL EXPRESSIONS

All papers typeset by Copernicus Publications follow the math typesetting regulations given by the IUPAC Green Book (IUPAC: Quantities, Units and Symbols in Physical Chemistry, 2nd Edn., Blackwell Science, available at: http://old.iupac.org/publications/book 1993).

Physical quantities/variables are typeset in italic font (t for time, T for Temperature)

Indices which are not defined are typeset in italic font (x, y, z, a, b, c)

Items/objects which are defined are typeset in roman font (Car A, Car B)

Descriptions/specifications which are defined by itself are typeset in roman font (abs, rel, ref, tot, net, ice)

Abbreviations from 2 letters are typeset in roman font (RH, LAI)

Vectors are identified in bold italic font using $\boldsymbol{x}$

Matrices are identified in bold roman font

Multiplication signs are typeset using the LaTeX commands `\times` (for vector products, grids, and exponential notations) or `\cdot`

The character * should not be applied as mutliplication sign

9.4 EQUATIONS

9.4.1 Single-row equation

Unnumbered equations (i.e. using `$$` and getting inline preview in RStudio) are not supported by Copernicus.

$$1 \times 1 \cdot 1 = 42 \tag{2}$$

$$A = \pi r^2 \tag{3}$$

$$x = \frac{2b \pm \sqrt{b^2 - 4ac}}{2c}. \tag{4}$$

9.4.2 Multiline equation

$$3 + 5 = 8 \tag{5}$$

$$3 + 5 = 8 \tag{6}$$

$$3 + 5 = 8 \tag{7}$$

9.5 MATRICES

$$\begin{matrix} x & y & z \\ x & y & z \\ x & y & z \end{matrix}$$

9.6 ALGORITHM/PROGRAMMING CODE

If you want to use algorithms, you need to make sure yourself that the LaTeXpackages `algorithms` and `algorithmicx` are installed so that `algorithm.sty` respectively `algorithmic.sty` can be loaded by the Copernicus template. Both need to be available through your preferred LaTeX distribution. With TinyTeX (or TeX Live), you can do so by running `tinytex::tlmgr_install(c("algorithms", "algorithmicx"))`

Copernicus staff will no accept any additional packages from your LaTeX source code, so please stick to these two acceptable packages. They are needed to use the example below

9.7 CHEMICAL FORMULAS AND REACTIONS

For formulas embedded in the text, please use `\chem{}`, e.g. $\mathrm{A} \rightarrow \mathrm{B}$.

The reaction environment creates labels including the letter R, i.e. (R1), (R2), etc.

Algorithm 1 Algorithm Caption

```
$i \leftarrow 10$
if $i \geq 5$ then
    $i \leftarrow i - 1$
else
    if $i \leq 3$ then
        $i \leftarrow i + 2$
    end if
end if
```

  - `\rightarrow` should be used for normal (one-way) chemical reactions
  - `\rightleftharpoons` should be used for equilibria
  - `\leftrightarrow` should be used for resonance structures

$$A \rightarrow B \tag{R1}$$

$$Coper \rightleftharpoons nicus \tag{R2}$$

$$Publi \leftrightarrow cations \tag{R3}$$

9.8 PHYSICAL UNITS

Please use `\unit{}` (allows to save the math/`$` environment) and apply the exponential notation, for example $3.14\,\mathrm{km\,h^{-1}}$ (using LaTeX mode: `\( 3.14\,\unit{...} \)`) or $0.872\,\mathrm{m\,s^{-1}}$ (using only `\unit{0.872\,m\,s^{-1}}`).

10 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

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