# Data Descriptor Template

### Title

*Daily PM2.5 concentration estimates by County, ZIP code, and census tract in 11 western states 2008-2018*

### Authors

Colleen E. Reid1,2,3, Ellen M. Considine2, Melissa M. Maestas2, Gina Li1,2

**Affiliations**

1. Geography Department, University of Colorado Boulder

2. Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder

3. Institute of Behavioral Sciences, University of Colorado Boulder

Corresponding author(s): Colleen E. Reid (Colleen.Reid@colorado.edu)

### Abstract

*170 words maximum*

The Abstract should succinctly describe the study, the assay(s) performed, the resulting data, and their reuse potential, but should not make any claims regarding new scientific findings.

This dataset provides daily PM2.5 concentration estimates at the centroids of each county, ZIP code, and census tract across the western US, from 2008-2018. We used ensemble machine learning models trained on 24-hour PM2.5 measurements from a wide variety of monitoring station data from across 11 states in the western US with predictor variables that included satellite, land cover, chemical transport model, and meteorological data. We evaluated the models using spatial cross-validation techniques to estimate accuracy at locations where there are no PM2.5 monitors. We present data from a model for 2008-2016 that includes output from the Community Multiscale Air Quality (CMAQ) chemical transport model as a predictor variables and data for 2008-2018 on a model without CMAQ output. The 2008-2016 model achieved a 10-fold cross-validated (CV) R2 of 0.659 on the training set and an RMSE of 5.420 µg/m3 on the completely held-out training set. The 2008-2018 model achieved a 10-fold CV RMSE of 6.576 µg/m3 and R2 of 0.598 and a testing set RMSE of 6.599 and R2 of 0.593. These data can be used for understanding spatiotemporal patterns in PM2.5 and its associated health impacts in the western US where PM2.5 levels have been heavily impacted by wildfire smoke over this time period.

### Background & Summary

*700 words maximum*

The Background & Summary should provide an overview of the study design, the assay(s) performed, and the data generated, including any background information needed to put this study in the context of previous work and the literature, and should reference literature as needed. The section should also briefly outline the broader goals that motivated collection of the data, as well as their potential reuse value. **We also encourage authors to include a figure that provides a schematic overview of the study and assay(s) design**.

Fine particulate matter PM2.5 air pollution is increasingly associated with numerous adverse health outcomes including, but not limited to, mortality,1 respiratory and cardiovascular morbidity2,3, negative birth outcomes4, and lung cancer5. Although PM2.5 concentrations have been declining in many parts of the United States due to policies to limit emissions of air pollutants6, PM2.5 levels have been increasing in parts of the western US7. This increase has been shown to be associated with wildfire smoke7,8, which can cause PM2.5 concentrations that are several times higher than the Environmental Protection Agency's (EPA’s) daily PM2.5 National Ambient Air Quality Standard (NAAQS) in areas downwind of the wildfires for several days at a time9.

Estimates of PM2.5 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 most US counties do not contain a regulatory air pollution monitor.10 There is therefore not enough temporal and spatial coverage from FRM monitors to obtain a good estimate of the air pollution exposures where every person lives. 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 null11.

To improve population exposure assessment of PM2.5, researchers have increasingly been using methods to estimate PM2.5 exposures in the temporal and spatial gaps between regulatory monitoring data using data from satellites (such as aerosol optical depth (AOD) or polygons of smoke plumes or air pollution models10,12) 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 PM2.5 in space and time. Various methods of blending have been used including spatiotemporal regression kriging13, geographically-weighted regression14, and machine learning methods15–18.

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 PM2.5 where people breathe. These models have been implemented in various locations around the world at city, regional, and national scales19. 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. (2016; 2019)16,17 and Hu et al. (2017)18 have separately used machine learning algorithms to create fine-resolution daily PM2.5 estimates for the continental US. These models, however, have performed poorly in the western US16,18 and particularly the mountain west17 compared to the rest of the country. Given the increasing trends in PM2.5 concentrations in parts of the western US and the importance of wildfires as a source of PM2.5 there, it is important to have a model that is tailored to this region to capture the variability in PM2.5 concentrations in space and time in this region.

The dataset we describe here improves upon previous daily estimates of PM2.5 concentrations from machine learning models in the following ways: (1) use of a more extensive monitoring station network that captures more spatial locations and also proximity to wildfires, a key driver of PM2.5 in the western US, (2) use of an ensemble of machine learning algorithms which have been shown to improve model performance over just one machine learning algorithm, (3) retaining high monitoring data to train on thus allowing our models to predict the high PM2.5 values that occur during wildfire episodes, and (4) making the data available in a public repository which the above cited papers have not done. The data is available as daily PM2.5 concentration estimates at census tract, ZIP-code, and county scales with the aim that they be used by researchers to understand the societal impacts of air pollution exposure in the western US, where wildfires are a significant contributor to PM2.5 concentrations.

Figure 1: PM2.5 Monitoring Locations by Source of Monitoring Data

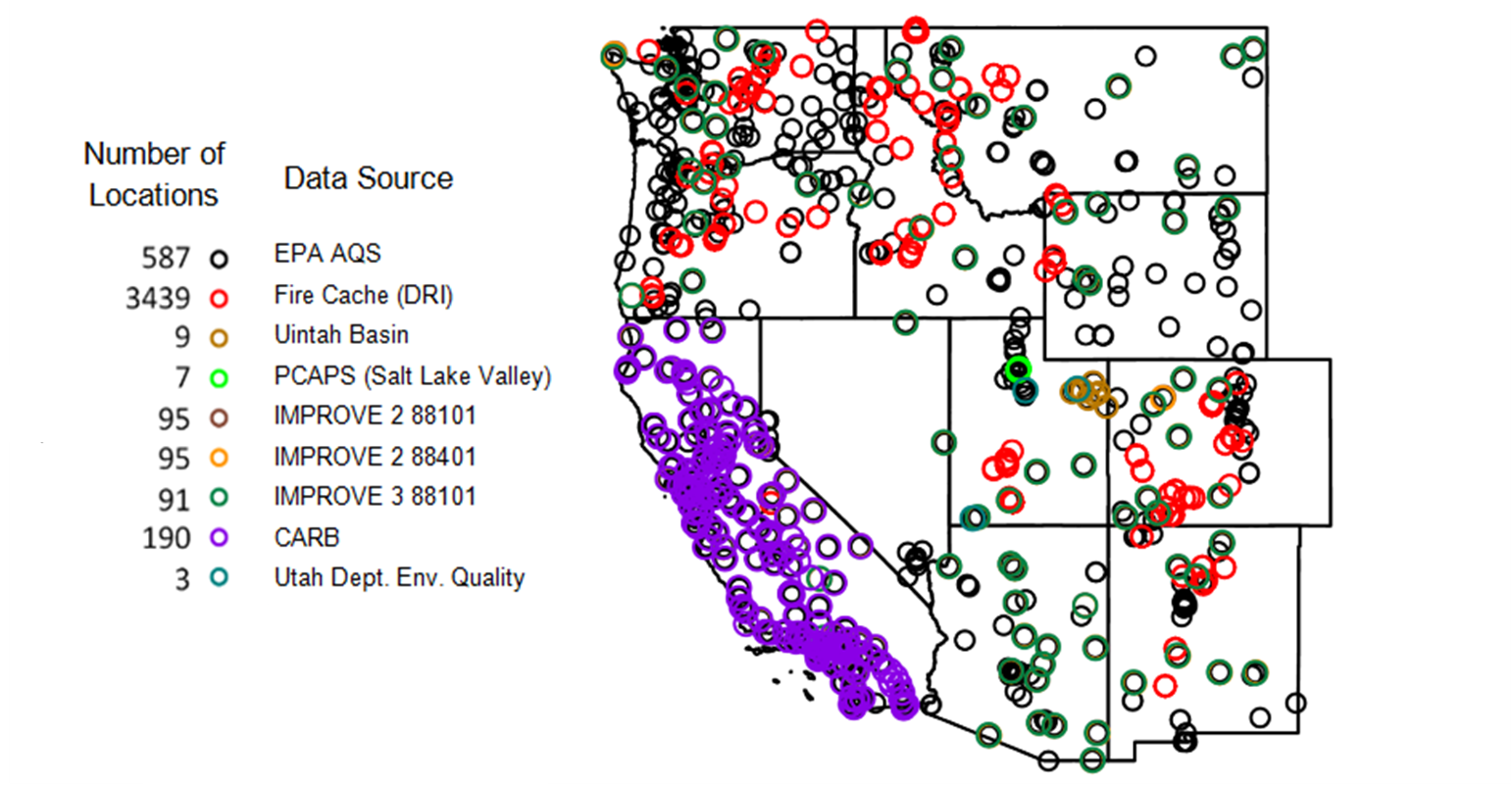


Table 1: Input Variables and their descriptions/sources and names in the Testing and Training Datasets

|  |  |  |
| --- | --- | --- |
| **Variable** | **Source / Description** | **Name in data file** |
| **Spatial variables** | | |
| Coordinates in degrees (Longitude, Latitude) |  | Lon, Lat |
| State |  | State |
| Region within the western US | Northwest = WA, OR; Southwest = CA, NV; Four corners = AZ, CO, NM, UT; and Northern mountain states = WY, MT, ID | Region |
| Population Density | US Census 2010; spatial resolution = census tract level | Pop\_density |
| Percent urban landcover in circular buffers of radius 1km, 5km, 10km (%) | Derived from the USGS National Map Program NLCD, derived from 2011 Landsat imagery | NLCD\_1km, NLCD\_5km, NLCD\_10km |
| Lengths of arterial and collector roads in circular buffers of radius 100m, 250m, 500m, 1000m (mi) | Derived from the US Federal Highway Administration National Highway Planning Network | Both\_100, Both\_250, Both\_500, Both\_1000 |
| Elevation (m) | USGS 3D Elevation Program, National Elevation Database (NED) | elevation |
| **Temporal variables** | | |
| Date |  | Date |
| Season | Summer = June-August, Fall = September-November, Winter = December-February, Spring = March-May | Season |
| Cosine of month | Cosine (2\*pi\*month/12) | CosMonth |
| Cosine of day of year | Cosine (2\*pi\*day of year/365) | CosDOY |
| Cosine of day of week | Cosine (2\*pi\*day of week/7) | CosDOW |
| Day of week |  | DayOfWeek |
| Year |  | Year |
| Middle of the study period | Years 2013-2016 | Mid\_Study |
| Late in the study period | Years 2017-2018 | Late\_Study |
| **Spatio-temporal variables** | | |
| PM2.5 observations (µg/m3) | Federal and state air quality databases | PM2.5\_Obs |
| Aerosol optical depth | NASA MAIAC; spatial resolution = 1km, temporal resolution = generally two observations per day are found but additional ones are possible near the edges of the scan | MAIAC\_AOD |
| Chemical transport model simulation | US EPA CMAQ; ; spatial resolution = 4km, temporal resolution = daily | all\_CMAQ |
| Vegetation index | NASA MODIS NDVI; spatial resolution = 1km, temporal resolution = monthly | ndvi |
| Active fire points within the last week, weighted inversely by distance (up to 500km) | Derived from NASA MODIS thermal anomalies; spatial resolution = 1km, temporal resolution = daily | Lag0, Lag1, Lag2, Lag3, Lag4, Lag5, Lag6, Lag7 |
| Indicator for whether there was at least one fire in the last week within 500km | Derived from NASA MODIS thermal anomalies; spatial resolution = 1km, temporal resolution = daily | Binary\_fire |
| Temperature 2m above the ground (K) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | TMP\_2m |
| Relative humidity 2m above the ground (%) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | RH\_2m |
| Dew point temperature 2m above the ground (K) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | DPT\_2m |
| Height of the planetary boundary layer (m) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | HPBL\_surface |
| Surface pressure (Pa) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | PRES\_surface |
| Pressure reduced to mean sea level (Pa) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | PRMSL\_mean\_sea\_level |
| East-West and North-South components of windspeed 10m above the ground (m/s) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | Ugrd\_10m, Vgrd\_10m |
| Vertical component of windspeed, 700 and 850 mb (m/s) | North American Mesoscale Analysis (NAM); spatial resolution = 12km, temporal resolution = 6 hours | DZDT\_700\_mb, DZDT\_850\_mb |
| Interaction indicators between region and time span | 4 regions, 3 timespans | Region\_Mid\_Study, Region\_Late\_Study |

### Methods

The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). See the [detailed section in our submission guidelines](https://www.nature.com/sdata/publish/submission-guidelines#sec-5) for advice on writing a transparent and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below).

Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section.

Study Area

Our study area includes 11 western US states: Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming (Fig. 1). Our temporal domain were all days between January 1, 2008 and December 31, 2018. We predicted daily estimates of PM2.5 at the county, ZIP code, and census tract levels from machine learning ensembles trained on observed daily PM2.5 values from monitoring stations from a variety of sources (EPA Air Now data20, EPA IMPROVE Network21, California Air Resources Board (stationary and mobile monitoring network)22, Federal Land Manager Environmental Database23, Fire Cache Smoke Monitor Archive24, Utah State University25, Utah Department of Environmental Quality26, and the University of Utah27). The predictor variables for the machine learning ensemble included PM2.5 observations, latitude, longitude, date, year, cosine of month, cosine of day of year, cosine of day of week, day of week, season, state, region, mid-study (2013-2016), late-study (2017-2018), region-mid-study (interaction), region-late-study (interaction), active fire lags 0 through 7, binary fire variable, elevation, sum of arterial and collector roads within circles of radius 100, 250, 500, and 1000 meters, percent of urban land cover within circles of radius 1, 5, and 10 kilometers, population density, CMAQ output, MAIAC AOD, NDVI, planetary boundary layer height, temperature at 2 meters, relative humidity at 2 meters, dew point temperature at 2 meters, U- (east-west) and V- (north-south) components of wind speed at 10 meters, surface pressure, pressure reduced to mean sea level, and vertical wind velocity at an altitude of 850 mb and vertical wind velocity at an altitude of 700 mb.

More information on the sources of these data can be found in Table 1.

PM2.5 Measurements

To get a more comprehensive set of locations and time points of PM2.5 measurement throughout the western US, we did an extensive search for as many PM2.5 monitoring data within our spatial and temporal study area as we could find, stated above. We downloaded PM2.5 data from each of these sources for the 11-state region (Figure 1) including any of the following parameter codes: 88101, 88500, 88502, 81104. These data include the IMPROVE monitors that capture air quality information in more rural areas21 PM2.5 data in the Fire Cache Smoke Monitor Archive24, which includes U.S. Forest Service monitors that were deployed to capture air quality impacts during wildfire events. Our models, therefore, captured a more comprehensive monitoring network than has been used in other previous models of daily PM2.5.

Some states have additional PM2.5 monitors beyond those required by the U.S. EPA. We reached out to the department charged with air quality in every state within our study domain and obtained additional PM2.5 data from California Air Resources Board and the Utah Department of Environmental Quality. Any data that was repeated from multiple sources was de-duplicated.

We also reached out to researchers who may have had their own monitoring networks of PM2.5 throughout the region. We were able to obtain data from the Uintah Basin, Utah from Seth Lyman at Utah State University, and PM2.5 measurements from the Persistent Cold Air Pool Study (PCAPS)27 conducted in the Salt Lake Valley, Utah in January--February, 2011 from Dr. Geoff Silcox at the University of Utah.

This yielded a total of 1,591,533 daily PM2.5 observations, which represent 7,754 locations and 4,006 days.

Predictor Variables

Satellite AOD is a measure of particle loading in the atmosphere from the ground to the satellite. We obtained daily estimates of AOD from the MODIS Terra and Aqua combined Multi-angle Implementation of Atmospheric Correction (MAIAC) dataset28. This is the finest resolution (1 km) AOD dataset currently available and was available for our whole time period and spatial domain. After downloading each Hierarchical Data Format (HDF) file from the online repository, we calculated the average daily AOD values at each location, and took the value from the nearest neighbor at each PM2.5 monitoring location. MAIAC AOD has been shown to better predict PM2.5 than coarser resolution AOD29 and has been used in many studies in various geographic regions in blended models to predict daily PM2.5.30–32

We obtained meteorological data from the North American Mesoscale (NAM) Analysis meteorological model33 because it includes all of the standard meteorological variables, including planetary boundary layer height, which play a role in PM2.5 levels and can be important to help scale AOD values to ground-level estimates of PM2.534. We calculated 24-hour averages from 6-hourly data for temperature, relative humidity, sea level pressure, surface pressure, planetary boundary layer height, dew point temperature, precipitation, snow coverage, and the U and V components of wind speed. NAM has 12 km resolution.

Output from chemical transport models has been shown to be an important input to machine learning models for PM2.5.15,17 We obtained daily estimates of PM2.5 at spatial resolution from runs of the CMAQ (Community Multi-scale Air Quality) model from the U.S. EPA for the years 2008-2016.35

Because one of the reasons that PM2.5 concentrations have been increasing in the western US is the increasing number and magnitude of wildfires, we included variables about the proximity of a location to an active fire. We collected daily data about fire detection locations and size from the MODIS Thermal Anomalies/Fire Daily L3 Global 1km product (MOD14 and MYD14).36,37 As fires in closer proximity are likely to influence PM2.5 more than fires further away, we calculated the number of active fires in radial buffers of 25, 50, 100, and 500 km radii around each monitoring location, on the current day as well as each of the past seven days. Then we calculated an inverse-distance-weighted average for each lag. Finally, we created an indicator variable for whether there were one or more fires within 500 km of a monitor in the last week.

Elevation can influence PM2.5 concentrations. For example, PM2.5 can accumulate in mountain valleys during persistent cold air pools (commonly referred to as inversions) during winter.38 We obtained elevation data from the 3D Elevation Program, which has a resolution of 1 arc-second, which is approximately 30 m north/south and varies east/west with latitude.39

Surrounding land cover can be a proxy for air pollution emissions not from wildfires. We used the land cover class information from the Landsat-derived National Land Cover Dataset (NLCD)40 to calculate the percentage of urban development (codes 22, 23, and 24), agriculture (codes 81 and 82), and vegetated area other than agricultural land (codes 21, 41, 42, 43, 52, and 71) within buffer radii of 1 km, 5 km, and 10 km around each monitor. NLCD 2011 has a spatial resolution of 30 m and uses circa 2011 Landsat satellite data. We obtained the Normalized Difference Vegetation Index (NDVI) from the MODIS satellite product MOD13A341 at 1 km resolution by month as another measure of vegetation that was not just a measure of agricultural vegetation but all vegetation.

As a proxy indicator of emissions from vehicles, we calculated the sum of all road lengths of type "Arterial" and "Collector" within 100, 250, 500, 1000 m buffers of each monitoring location. Arterial roads are high-capacity urban roads. Collector roads are low-to-moderate capacity roads. The road data came from the National Highways Planning Network42 which contains spatial information on over 450,000 miles of highways in the United States.

We included population density as an additional proxy for emissions as areas with higher population have more sources of air pollution emissions. Population density was obtained from the 2010 U.S. Census.

To account for seasonality in PM2.5 data, we created the following predictor variables: cosine of day-of-week, cosine of day-of-year and cosine of month. This ensures that day/month values at the end/beginning of the week and year align.

We also created dummy variables for each state, region, and time period (2008-2012, 2013-2016, 2017-2018) in our study domain to allow for spatial and temporal variation in the data that could not be explained by any of the other spatial, temporal, or spatiotemporal variables. Using nested levels of spatiotemporal variables helped capture nonlinear spatiotemporal effects. Temporal variable nesting consisted of variables to indicate the periods 2008-2012, 2013-2016, and 2017-2018 (the periods when CMAQ simulation availability changed); year; season; cosine of month; and cosine of day of year. Spatial variable nesting consisted of dummy variables for region (within the 11 western states: northwest (i.e., WA, OR), southwest (i.e., CA, NV), four corners (i.e., AZ, CO, NM, UT), and northern mountain states (i.e., WY, MT, ID)), state; latitude; and longitude. We also included interaction terms for time period (grouping of years) and region. This type of nesting has been referred to as a “multiresolution basis”.43

Data merging

We created three datasets: one dataset to train the model and three prediction datasets (county, ZIP code, and census tract). The training dataset merged all predictor variables to each 24-hour average PM2.5 monitoring observation by linking the data temporally (using date) and spatially (by selecting the nearest observation for each predictor variable). Similarly, the prediction datasets were created by spatially and temporally linking all predictor variables to the population-weighted centroid of each count, ZIP code, and census tract for each day in the study domain.

Machine learning modelling

For the machine learning modelling, we performed 10-fold spatial cross-validation whereby all observations from a given monitoring site are within the same fold. This , is a more appropriate tool for evaluating the accuracy of a model when predicting PM2.5 at new locations.44 Before the 10-fold spatial cross-validation, we also removed a randomly selected 10% of the monitoring locations for a testing data set. With this setup, the results of 10-fold cross-validation (CV)(with no resampling) from training are used as validation metrics, while the results of each model applied separately to the testing set are a measure of how well the model will perform when predicting at new locations that were not part of the training data, as will be the case with our prediction data set. Most previous studies to create daily PM2.5 estimates using machine learning16–18 present results from non-spatial 10-folds, which violates the assumption of independence between folds because of repeated observations (on different days) from the same locations (PM2.5 sensor locations). Additionally, they do not give performance metrics for predictions at a left-out test set which provides overly optimistic performance metrics given that they are predicting at locations that were not in the training data set.

We used the metrics root-mean-squared error (RMSE) and R2 to report accuracy, for both the 10-fold cross-validation and for the left-out testing data set, for spatial folds and random folds (to provide better comparison to results from previous studies). Also, for comparison of our random-folds models to those in other studies, we calculated the “spatial R2” and “temporal R2” metrics used by Di et al. (2019).17 In that study, spatial R2 is calculated by regressing the annual mean PM2.5 at location *i* against the annual mean predicted PM2.5 at location *i*. Temporal R2 is calculated by regressing the difference between the actual PM2.5 at location *i* and the annual mean PM2.5 at location *i* against the difference between the predicted PM2.5 at location *i* and the annual mean predicted PM2.5 at location *i*.45

We employed ensemble machine learning to model PM2.5 exposures across the western US. Specifically, we used a generalized linear model (GLM) to combine the results from two machine learning algorithms: a random forest model and a gradient boosting model. These models performed best on preliminary analyses of random subsets of our dataset, which aligns with a previous study that found that tree-based models (using random forest, gradient boosting, and cubist algorithms) performed best in air pollution modelling.46 Then, we used the same random subsets of the data to tune hyperparameters for each algorithm via a grid-search (see code and final parameters in the Supplementary Material).

All analyses were run using R,47 and all machine learning models utilized the R packages caret48 and caret ensemble.49 Variable importance was calculated using the “permutation” importance algorithm in the caret package.

Our main model and prediction data set for 2008-2016. Because of our interest in the years 2017 and 2018, when there were many large wildfires in the western US, we created another model for 2008-2018 that does not contain CMAQ as a predictor variable.

When observed that our models, like most models, underpredict at very high values. We hypothesized that some of the higher values were being generated by a fundamentally different process than the lower values, most likely wildfires. We therefore did a sensitivity analysis in which we examined whether we would observe better performance by having different models for low values that are not likely influenced by wildfires and another for high values which are likely influenced by wildfires. This did not yield better performance but a more detailed description along with the results of the split analysis (“high” versus “low”) can be found in the Supplementary Material.

Daily PM2.5 Prediction Creation

Similar to the training dataset, the prediction input datasets were created by spatially and temporally linking all predictor variables to the centroid of each county, ZIP code, and census tract for each day in the study domain. After the merge, we observed some missingness in the predictors that required imputation to create daily PM2.5 predictions, however, missingness was observed for fewer than 1% of the location-days within each state, except for the meteorological variables within Wyoming, for which nearly 10% of the location-days were missing. Our approach to dealing with missing data was to use the missRanger50 package in R to impute the missing data for each state, based on all the available data for that state for all years in the given model (2008-2016 for the model including CMAQ and 2008-2018 otherwise).

Post-imputation, the full models (trained with the entire training dataset including the left-out 10% training data) were applied to the prediction input datasets to make the final PM2.5 predictions.

### Data Records

The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record.

Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each. They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

Table 2 lists the names and descriptions of the datafiles that are available on XXX. We provide our input data for the machine learning models (ML\_inputs.RData) which was aggregated at the monitoring location, the input data for the prediction data sets by state (Prediction\_inputs\_[state]\_2.RData), and files by state for the predictions using the model with CMAQ (Ensemble\_preds\_with\_CMAQ\_[state]\_2.RData) and files by state for the predictions using the model without CMAQ (Ensemble\_preds\_no\_CMAQ\_[state]\_2.RData). All prediction data sets have predictions at three spatial resolutions: county, ZIP code, and census tract. In the prediction data sets, ranger\_preds refers to the predictions made by the Ranger algorithm, xgbt\_preds to the predictions made by the extreme gradient boosting tree (XGBT) algorithm, and Ens\_preds to the predictions made by the GLM ensemble of Ranger and XGBT. Missing\_vars indicates if there was at least one missing input variable (and subsequent imputation). In the prediction inputs files, there are individual columns to show which of these variables (which could have been for NDVI, NAM, or MAIAC) was missing for each observation. Missing\_CMAQ indicates if CMAQ was missing; all of these were for the years that CMAQ was not available 2017 and 2018.

Table 2: Prediction Data Set Information

|  |  |  |  |
| --- | --- | --- | --- |
| **File Name(s)** | **Data Frame(s)** | **Years of Daily PM2.5 Estimates** | **Variables** |
| Ensemble\_preds\_no\_CMAQ\_[state]\_2.RData | DF | 2008-2016 | County\_FIPS, Tract\_code, ZCTA5\_code, Lon, Lat, Date, ranger\_preds, xgbt\_preds, Missing\_vars, Ens\_pred |
| Ensemble\_preds\_with\_CMAQ\_[state]\_2.RData | DF | 2008-2018 | County\_FIPS, Tract\_code, ZCTA5\_code, Lon, Lat, Date, CMAQ\_ranger\_preds, CMAQ\_xgbt\_preds, Missing\_vars, Missing\_CMAQ, Ens\_pred |
| Prediction\_inputs\_[state]\_2.RData | DATA | 2008-2018 | Variables from Table 1; Missing\_NDVI, Missing\_NAM, Missing\_MAIAC, Missing\_CMAQ, Missing\_vars |
| ML\_inputs.RData | DF\_with\_CMAQ, DF\_no\_CMAQ | 2008-2018 | Variables from Table 1 |

### Technical Validation

The Technical Validation section should present any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. *This is a required section*; authors must provide information to justify the reliability of their data.

Possible content **may include:**

* experiments that support or validate the data-collection procedure(s) (e.g. negative controls, or an analysis of standards to confirm measurement linearity)
* statistical analyses of experimental error and variation
* general discussions of any procedures used to ensure reliable and unbiased data production, such as blinding and randomization, sample tracking systems, etc.
* any other information needed for assessment of technical rigour by the referees

Generally, this **should not include:**

* follow-up experiments aimed at testing or supporting an interpretation of the data
* statistical hypothesis testing (e.g. tests of statistical significance, identifying differentially expressed genes, trend analysis, etc.)
* exploratory computational analyses like clustering and annotation enrichment (e.g. GO analysis).

**\*\*time series, observed-expected plots, tables with R2 and RMSE, descriptive data on prediction data**

Table 2 shows the mean and quantiles of PM2.5 monitoring observations across our study domain. We observe that the maximum values of PM2.5 are much higher in 2012 and 2015-2018 than in the other years, that California has much higher values of PM2.5 than the other states (although all of the states have right-skewed distributions), and that spring has much lower values of PM2.5 than the other seasons.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Subset** | **N** | **Mean (µg/m3)** | **Min (µg/m3)** | **Q1 (µg/m3)** | **Median (µg/m3)** | **Q3 (µg/m3)** | **Max (µg/m3)** |
| **Year** |  |  |  |  |  |  |  |
| 2008 | 121,396 | 8.998 | 0 | 3.8 | 6.6 | 11.4 | 200.2 |
| 2009 | 131,397 | 8.329 | 0 | 3.7 | 6.3 | 10.6 | 195.583 |
| 2010 | 143,040 | 7.441 | 0 | 3.3 | 5.7 | 9.6 | 114 |
| 2011 | 147,828 | 8.316 | 0 | 3.5 | 6.2 | 10.625 | 208.025 |
| 2012 | 156,946 | 8.156 | 0 | 3.6 | 6.2 | 10.2 | 705.458 |
| 2013 | 163,096 | 8.559 | 0 | 3.7 | 6.5 | 10.635 | 452.792 |
| 2014 | 162,365 | 7.85 | 0 | 3.458 | 6 | 9.813 | 504.542 |
| 2015 | 137,573 | 7.648 | 0 | 3.2 | 5.5 | 9.2 | 830.792 |
| 2016 | 140,581 | 6.929 | 0 | 3 | 5.2 | 8.7 | 804.5 |
| 2017 | 150,213 | 8.938 | 0 | 3.2 | 5.7 | 10 | 811.792 |
| 2018 | 137,098 | 9.193 | 0 | 3.6 | 6 | 10 | 826.292 |
| **State** |  |  |  |  |  |  |  |
| Arizona | 79,964 | 6.365 | 0 | 3.3 | 5.3 | 7.9 | 199.3 |
| California | 680,549 | 10.269 | 0 | 4.95 | 8.2 | 12.625 | 791.625 |
| Colorado | 61,453 | 5.66 | 0 | 2.3 | 4.5 | 7.3 | 781.455 |
| Idaho | 59,456 | 7.949 | 0 | 3 | 5.3 | 8.9 | 519.391 |
| Montana | 90,837 | 7.431 | 0 | 2.5 | 4.8 | 8.3 | 641.9 |
| Nevada | 39,311 | 6.829 | 0 | 3.5 | 5.7 | 8.4 | 230 |
| New Mexico | 63,717 | 5.231 | 0 | 2.5 | 4.15 | 6.6 | 263 |
| Oregon | 150,682 | 7.5 | 0 | 3.1 | 4.8 | 8.15 | 811.792 |
| Utah | 83,290 | 7.202 | 0 | 3.3 | 5.2 | 8.2 | 225.25 |
| Washington | 230,120 | 6.546 | 0 | 2.95 | 4.7 | 7.6 | 830.792 |
| Wyoming | 51,128 | 4.311 | 0 | 1.8 | 3.2 | 5.2 | 513.417 |
| **Season** |  |  |  |  |  |  |  |
| Fall | 399,121 | 8.643 | 0 | 3.7 | 6.375 | 10.5 | 830.792 |
| Spring | 399,129 | 6.144 | 0 | 3.1 | 5 | 7.9 | 189.9 |
| Summer | 405,596 | 8.76 | 0 | 4 | 6.4 | 10.25 | 712.917 |
| Winter | 387,687 | 9.299 | 0 | 3.043 | 6.5 | 12.2 | 557 |

Table 3 shows the performance metrics (RMSE and R2) of our ensemble machine learning models with spatial folds and random folds for the cross validation (CV). The models that included CMAQ output included as a predictor variable were run for the years 2008-2016 (the years for which we have CMAQ data), and the models without CMAQ were run for 2008-2018 to include 2017 and 2018 because there were so many large wildfires in the western US during these years. The results for the training data are based on predictions for those observations when they are in the left-out fold within the 10-fold cross validation. The results for the training sets are metrics based on predictions at locations that were XXXX, whereas the testing RMSE and R2 are for the completely left-out 10% of the data, a good test of how our model will do predicting at locations on which it did not train.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Training (10-fold CV) RMSE (µg/m3)** | **Training (10-fold CV) R2** | **Testing RMSE (µg/m3)** | **Testing R2** |
| 2008-2016 model with CMAQ, spatial folds | 5.061 | 0.659 | 5.420 | 0.589 |
| 2008-2018 model without CMAQ, spatial folds | 6.576 | 0.598 | 6.599 | 0.593 |
| 2008-2016 model with CMAQ, random folds | 4.482 | 0.732 | 4.642 | 0.715 |
| 2008-2018 model without CMAQ, random folds | 5.482 | 0.719 | 5.954 | 0.680 |

Overall, models including CMAQ perform better (have lower RMSE and higher R2 values) than models without CMAQ. This may be due to the additional information provided by the CMAQ output or could be because the models without CMAQ include two additional years of data that more days with high PM2.5, which are much harder to predict accurately than lower values.For comparison, results for a model without CMAQ on the years 2008-2016 only gave test set RMSE = 4.747 µg/m3 and R2 = 0.702, slightly better than the model without CMAQ that included 2017 and 2018, but worse than the model with CMAQ for years 2008-2016.

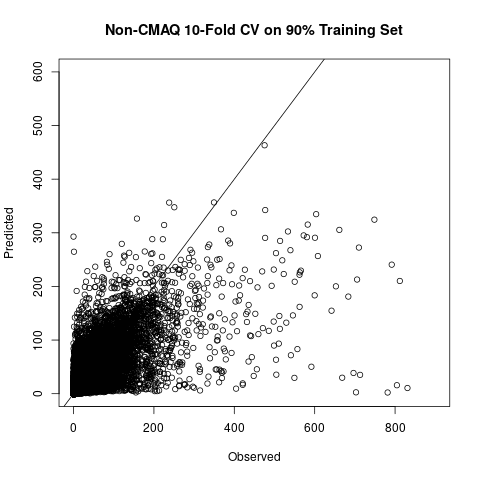
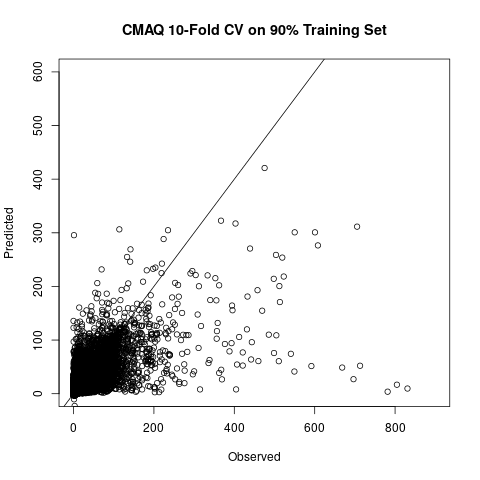
The spatial-folds models perform worse than the random-folds models. This is not surprising because the spatial folds do not allow for observations from the same location to be in multiple folds, therefore the models are predicting to locations that they did not train on, whereas random folds have likely trained on observations at all locations, thus are more likely to predict values better for those locations. Using solely random folds can therefore be misleading as to the performance of the models when they are predicting at locations without monitoring data. Thus, we posit that most of the models presented in the literature previously that use random folds are reporting R2 values that are likely higher than their predictive performance at non-sampled locations.

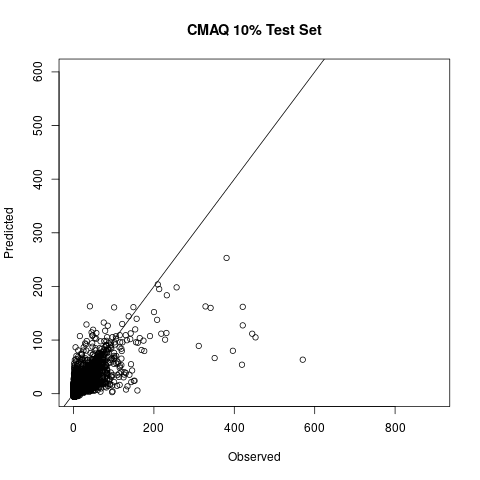
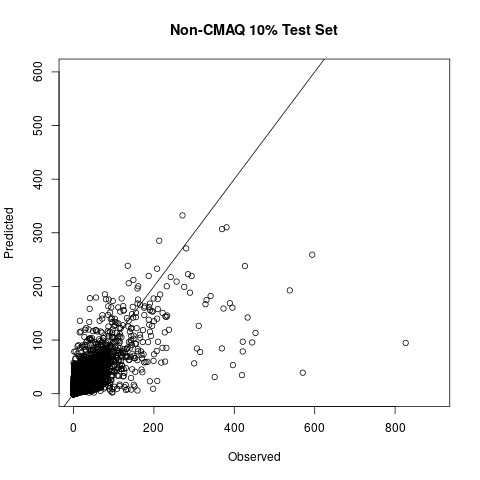
Performance of our models on our completely left-out testing data set provide worse metrics than their training (10-fold CV) counterparts. Some of the discrepancy between training and testing set results is because the testing data set was not used to inform the development of the model; some of the discrepancy is because of random chance of a given monitoring site being in the testing data set.

For comparison: the performance metrics for full models (without any cross-validation folds) on the 2008-2016 CMAQ and the 2008-2018 non-CMAQ datasets are, respectively, RMSE = 1.726 µg/m3 and R2 = 0.960; RMSE = 2.027 µg/m3 and R2 = 0.961. These are much better performance metrics than any of those in the tables below because the full models are overfitting the training data, and thus are not realistic representations of how accurately our model will be able to predict at locations outside of our training set.

Henceforth all results in this section refer to the spatial-folds analysis. Results from the random-folds analysis are in the Supplementary Material.

The predicted-versus-observed plots in Figure XX illustrate the variation in both predictions from our models and observations of PM2.5. It is clear from these plots that there were many more high values in the years 2017 and 2018 (on the non-CMAQ model plots). Also, all models tend to dramatically underpredict values of PM2.5 higher than 200 µg/m3, which is likely because there are fewer high values than low values in the training set.



Tables XX shows the RMSE values of our models on different levels of PM2.5, years, states, and seasons. Similar tables with R2 values for these data can be found in the Supplementary Material.

The models with CMAQ always perform better than the models without CMAQ. When investigating variable importance in each model, MAIAC AOD rises in variable importance when CMAQ output is not in the model (Supplemental Material). (Note that while collinearity between variables does not matter for prediction with random forest, it most likely reduces the variable importance calculations via permutation.51)

In the spatiotemporal subsets, our data show that we have better predictive performance at lower levels of PM2.5. This is likely because a much higher number of observations at lower values allowed the model to be better trained at those values. We also observed higher RMSE for the years 2012 and 2015-2018, which have some of the highest PM2.5 values. The patterning of results by state is less clear, although it is notable that the RMSE values for California are lower than might be expected given the state’s higher-than average PM2.5 levels. This is likely because there are so many monitoring locations and thus observations from California. Finally, the RMSE values for Spring are lower than those from the other seasons, which is likely due to the predominance of lower PM2.5 values in the spring.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PM2.5 (µg/m3)** | **CMAQ Ensemble Training RMSE (µg/m3)** | **CMAQ Ensemble Testing RMSE (µg/m3)** | **Non-CMAQ Ensemble Training RMSE (µg/m3)** | **Non-CMAQ Ensemble Testing RMSE (µg/m3)** |
| Below 35 | 3.346 | 3.905 | 4.201 | 4.151 |
| Below 60 | 3.664 | 4.179 | 4.651 | 4.499 |
| Below 150 | 4.010 | 4.526 | 5.154 | 5.009 |
| Below 300 | 4.319 | 4.646 | 5.635 | 5.376 |
| Below 500 | 4.618 | 5.247 | 6.044 | 6.118 |
| Below 1000 | 5.061 | 5.420 | 6.576 | 6.599 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **CMAQ Ensemble Training RMSE (µg/m3)** | **CMAQ Ensemble Testing RMSE (µg/m3)** | **Non-CMAQ Ensemble Training RMSE (µg/m3)** | **Non-CMAQ Ensemble Testing RMSE (µg/m3)** |
| 2008 | 4.037 | 4.818 | 4.920 | 4.903 |
| 2009 | 3.755 | 4.416 | 4.603 | 4.611 |
| 2010 | 3.537 | 3.834 | 4.372 | 4.013 |
| 2011 | 4.016 | 4.264 | 4.840 | 4.396 |
| 2012 | 5.459 | 8.006 | 6.180 | 8.456 |
| 2013 | 4.990 | 5.927 | 5.836 | 6.022 |
| 2014 | 4.816 | 5.175 | 5.728 | 5.359 |
| 2015 | 5.881 | 5.107 | 6.723 | 5.423 |
| 2016 | 7.239 | 5.296 | 7.794 | 5.627 |
| 2017 | N/A | N/A | 9.348 | 8.868 |
| 2018 | N/A | N/A | 8.553 | 10.435 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **State** | **CMAQ Ensemble Training RMSE (µg/m3)** | **CMAQ Ensemble Testing RMSE (µg/m3)** | **Non-CMAQ Ensemble Training RMSE (µg/m3)** | **Non-CMAQ Ensemble Testing RMSE (µg/m3)** |
| Arizona | 3.139 | 4.094 | 3.795 | 3.894 |
| California | 4.753 | 4.026 | 6.577 | 5.197 |
| Colorado | 5.940 | 3.539 | 9.143 | 3.758 |
| Idaho | 7.016 | 11.539 | 7.789 | 10.034 |
| Montana | 5.428 | 5.192 | 7.642 | 7.183 |
| Nevada | 3.639 | 3.630 | 4.273 | 3.809 |
| New Mexico | 3.628 | 9.732 | 3.732 | 10.823 |
| Oregon | 5.239 | 10.063 | 8.441 | 12.081 |
| Utah | 4.667 | 4.979 | 5.223 | 4.584 |
| Washington | 5.642 | 4.409 | 6.814 | 8.412 |
| Wyoming | 6.983 | 3.859 | 6.405 | 3.839 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Season** | **CMAQ Ensemble Training RMSE (µg/m3)** | **CMAQ Ensemble Testing RMSE (µg/m3)** | **Non-CMAQ Ensemble Training RMSE (µg/m3)** | **Non-CMAQ Ensemble Testing RMSE (µg/m3)** |
| Fall | 5.960 | 6.465 | 7.896 | 8.505 |
| Spring | 3.168 | 3.227 | 3.871 | 3.293 |
| Summer | 5.745 | 5.760 | 7.771 | 7.629 |
| Winter | 4.682 | 5.632 | 5.787 | 5.620 |

### Usage Notes

*This section is optional*

The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our [code availability policy](http://www.nature.com/sdata/policies/editorial-and-publishing-policies#code-avail) for advice on supplying custom code alongside Data Descriptor manuscripts.

For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

All of our data files are available in the RData format. An RData file can be opened in R using the “load” command.

Note that the state data frames range in size from about 2 million rows to about 34 million rows. Especially when reading in the prediction input file for California, the size of this data frame is likely to crash R on a regular laptop. A computer with lots of memory and/or cloud computing may be necessary.

### Code Availability

For all studies using custom code in the generation or processing of datasets, a statement must be included under the subheading "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset.

### Acknowledgements

This work was supported by Earth Lab through the University of Colorado Boulder’s Grand Challenge Initiative. We also thank the following individuals and groups who helped us obtain additional monitoring data: Denise Odenwalder, Charles Pearson, and Joseph McCormack at the California Air Resources Board; Seth Lyman at Utah State University for the Uintah Basin PM2.5 data; Geoff Silcox at the University of Utah for the PM2.5 data from the Persistent Cold Air Pool Study (PCAPS).

### Author contributions

C.E.R. conceptual design, project supervision, writing and editing. E.M.C. data retrieval, data analysis, data and code curation, writing and editing. M.M.M. data retrieval, data analysis, writing and editing. G.L. data retrieval.

### Competing interests

The authors claim no conflict of interest.

### Figures

Figure images should be provided as separate files and should be referred to using a consistent numbering scheme through the entire Data Descriptor. In most cases, a Data Descriptor should not contain more than three figures, but more may be allowed when needed. We discourage the inclusion of figures in the Supplementary Information – all key figures should be included here in the main Figure section.

For initial submissions, authors may choose to supply a single PDF with embedded figures.

Authors are encouraged to consider creating a figure that outlines the experimental workflow(s) used to generate and analyse the data output(s).

### Figure Legends

### Figure legends begin with a brief title sentence summarizing the purpose of the figure as a whole, and continue with a short description of what is shown in each panel and an explanation of any symbols used. Legends must total no more than 350 words, and may contain literature references. The first sentence of the legend will be used as the title for the figure. It should contain no references of any kind, including to specific figure panels, bibliographic citations or references to other figures or panels.

### Tables

Authors are encouraged to provide one or more tables that provide basic information on the main ‘inputs’ to the study (e.g. samples, participants, or information sources) and the main data outputs of the study; also see the additional information on providing metadata on page 6. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Authors may provide tables within the Word document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included in the Word document. Generally, a Data Descriptor should have fewer than ten tables, but more may be allowed when needed. Tables may be of any size, but only tables that fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as ‘Online-only’ tables and numbered separately from the main table list e.g. ‘Table 1, Table 2, Online-only Table 1’ etc.

### References

1. Achilleos, S. *et al.* Acute effects of fine particulate matter constituents on mortality: A systematic review and meta-regression analysis. *Environ. Int.* **109**, 89–100 (2017).

2. Xing, Y.-F., Xu, Y.-H., Shi, M.-H. & Lian, Y.-X. The impact of PM2.5 on the human respiratory system. *J. Thorac. Dis.* **8**, E69-74 (2016).

3. Rajagopalan, S., Al-Kindi, S. G. & Brook, R. D. Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review. *J. Am. Coll. Cardiol.* **72**, 2054–2070 (2018).

4. Klepac, P., Locatelli, I., Korošec, S., Künzli, N. & Kukec, A. Ambient air pollution and pregnancy outcomes: A comprehensive review and identification of environmental public health challenges. *Environ. Res.* **167**, 144–159 (2018).

5. Hamra, G. B. *et al.* Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environ. Health Perspect.* **122**, 906–911 (2014).

6. Fann, N., Kim, S.-Y., Olives, C. & Sheppard, L. Estimated Changes in Life Expectancy and Adult Mortality Resulting from Declining PM2.5 Exposures in the Contiguous United States: 1980-2010. *Environ. Health Perspect.* **125**, 097003 (2017).

7. McClure, C. D. & Jaffe, D. A. US particulate matter air quality improves except in wildfire-prone areas. *Proc. Natl. Acad. Sci.* 201804353 (2018) doi:10.1073/pnas.1804353115.

8. O’Dell, K., Ford, B., Fischer, E. V. & Pierce, J. R. The contribution of wildland-fire smoke to US PM2.5 and its influence on recent trends. *Environ. Sci. Technol.* (2019) doi:10.1021/acs.est.8b05430.

9. Reid, C. E. *et al.* Associations between respiratory health and ozone and fine particulate matter during a wildfire event. *Environ. Int.* **129**, 291–298 (2019).

10. Brokamp, C., Brandt, E. B. & Ryan, P. H. Assessing Exposure to Outdoor Air Pollution for Epidemiological Studies: Model-based and Personal Sampling Strategies. *J. Allergy Clin. Immunol.* (2019) doi:10.1016/j.jaci.2019.04.019.

11. Zeger, S. L. *et al.* Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ. Health Perspect.* **108**, 419–426 (2000).

12. Liu, J. C., Pereira, G., Uhl, S. A., Bravo, M. A. & Bell, M. L. A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. *Environ. Res.* **136**, 120–132 (2015).

13. Hu, H. *et al.* Satellite-based high-resolution mapping of ground-level PM2.5 concentrations over East China using a spatiotemporal regression kriging model. *Sci. Total Environ.* **672**, 479–490 (2019).

14. Lassman, W. *et al.* 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**, 106–121 (2017).

15. Reid, C. E. *et al.* Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Env. Sci Technol* **49**, 3887–96 (2015).

16. Di, Q. *et al.* Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Env. Sci Technol* **50**, 4712–21 (2016).

17. Di, Q. *et al.* An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environ. Int.* **130**, 104909 (2019).

18. Hu, X. *et al.* Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ. Sci. Technol.* **51**, 6936–6944 (2017).

19. Bellinger, C., Jabbar, M. S. M., Zaiane, O. & Osornio-Vargas, A. A systematic review of data mining and machine learning for air pollution epidemiology. *Bmc Public Health* **17**, 907 (2017).

20. EPA, U.S. AirData website File Download page. https://aqs.epa.gov/aqsweb/airdata/download\_files.html#Daily.

21. US EPA, O. PM 2.5 Visibility (IMPROVE) | Ambient Monitoring Technology Information Center | US EPA. https://www3.epa.gov/ttnamti1/visdata.html.

22. California Air Resources Board. AQMIS 2 - Air Quality and Meteorological Information System. https://www.arb.ca.gov/aqmis2/aqmis2.php.

23. University, Colorado State. Federal Land Manager Environmental Database. http://views.cira.colostate.edu/fed/DataWizard/Default.aspx.

24. Fire Cache Smoke Monitor Archive. https://wrcc.dri.edu/cgi-bin/smoke.pl.

25. Lyman, S. UBAQR\_2018\_AnnualReport.pdf. https://usu.app.box.com/s/rigadr7yt7ipir4gzj75vfaazoe8u8mt.

26. Utah DEQ: AMP. http://www.airmonitoring.utah.gov/dataarchive/archpm25.htm.

27. Silcox, G. D., Kelly, K. E., Crosman, E. T., Whiteman, C. D. & Allen, B. L. Wintertime PM2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley. *Atmos. Environ.* **46**, 17–24 (2012).

28. LAADS DAAC. https://ladsweb.modaps.eosdis.nasa.gov/archive/allData/6/MCD19A2/.

29. Chudnovsky, A. A., Kostinski, A., Lyapustin, A. & Koutrakis, P. Spatial scales of pollution from variable resolution satellite imaging. *Env. Pollut* **172C**, 131–138 (2012).

30. Lee, H. J. Benefits of High Resolution PM2.5 Prediction using Satellite MAIAC AOD and Land Use Regression for Exposure Assessment: California Examples. *Environ. Sci. Technol.* (2019) doi:10.1021/acs.est.9b03799.

31. Geng, G. *et al.* Satellite-Based Daily PM2.5 Estimates During Fire Seasons in Colorado. *J. Geophys. Res.-Atmospheres* **123**, 8159–8171 (2018).

32. Li, R., Ma, T., Xu, Q. & Song, X. Using MAIAC AOD to verify the PM2.5 spatial patterns of a land use regression model. *Environ. Pollut. Barking Essex 1987* **243**, 501–509 (2018).

33. North American Mesoscale Forecast System (NAM) | National Centers for Environmental Information (NCEI) formerly known as National Climatic Data Center (NCDC). https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-mesoscale-forecast-system-nam.

34. Liu, Y., Sarnat, J. A., Kilaru, V., Jacob, D. J. & Koutrakis, P. Estimating ground-level PM2.5 in the eastern United States using satellite remote sensing. *Env. Sci Technol* **39**, 3269–78 (2005).

35. Rappold, A. G., Reyes, J., Pouliot, G., Cascio, W. E. & Diaz-Sanchez, D. Community Vulnerability to Health Impacts of Wildland Fire Smoke Exposure. *Environ. Sci. Technol.* **51**, 6674–6682 (2017).

36. Giglio, L., Csiszar, I. & Justice, C. O. Global distribution and seasonality of active fires as observed with the Terra and Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) sensors. *J. Geophys. Res. Biogeosciences* **111**, (2006).

37. Hawbaker, T. J. *et al.* Mapping burned areas using dense time-series of Landsat data. *Remote Sens. Environ.* **198**, 504–522 (2017).

38. Whiteman, C. D., Hoch, S. W., Horel, J. D. & Charland, A. Relationship between particulate air pollution and meteorological variables in Utah’s Salt Lake Valley. *Atmos. Environ.* **94**, 742–753 (2014).

39. About 3DEP Products & Services. https://www.usgs.gov/core-science-systems/ngp/3dep/about-3dep-products-services.

40. Homer, C. *et al.* Completion of the 2011 National Land Cover Database for the Conterminous United States—Representing a Decade of Land Cover Change Information. *Photogramm. Eng. Remote Sens.* (2017).

41. Didan, K. MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. (2015).

42. National Highway Planning Network - Tools - Processes - Planning - FHWA. https://www.fhwa.dot.gov/planning/processes/tools/nhpn/index.cfm.

43. Katzfuss, M. A multi-resolution approximation for massive spatial datasets. *J. Am. Stat. Assoc.* **112**, 201–214 (2017).

44. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

45. Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J. & Schwartz, J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* **45**, 6267–6275 (2011).

46. Xu, Y. *et al.* Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environ. Pollut.* **242**, 1417–1426 (2018).

47. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2018).

48. Kuhn, M. & Contributions from Jed Wing, S. W., Andre Williams, Chris Keefer and Allan Engelhardt. *caret: Classification and Regression Training.* (2012).

49. Dean-Mayer, Z. A. & Knowles, J. E. *caretEnsemble: Ensembles of Caret Models*. (2019).

50. Mayer, M. *missRanger: Fast Imputation of Missing Values.* (2019).

51. Gregorutti, B., Michel, B. & Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **27**, 659–678 (2017).

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference. See ‘Citing Data’ below for further details.

### Additional Formatting Information

**Referencing Figures, Tables, and other content**

**The Word document may reference Figures (e.g. Fig. 1), Tables (e.g. Table 1), online-only tables (e.g. Online-only Table 1) and Supplementary Information (e.g. Supplementary Table 1, or Supplementary File 2, etc.). When information from metadata documents must be referred to, it should also be included in the main manuscript as Tables, and formatted in a way that suits human readability. To refer to the ISA-Tab metadata records within the manuscript, use the phrase “see associated Metadata Record”.**

**Citation format**

All references should be numbered sequentially, first throughout the text, then in tables, followed by figures and, finally, boxes; that is, references that only appear in tables, figures or boxes should be last in the reference list. Only one publication is given for each number. Only papers that have been published or accepted by a named publication or recognized preprint server should be in the numbered list; preprints of accepted papers in the reference list should be submitted with the manuscript. Published conference abstracts, numbered patents, and archived code with an assigned DOI may be included in the reference list. Grant details and acknowledgments are not permitted as numbered references. Footnotes are not used.

Scientific Data uses standard Nature referencing style. All authors should be included in reference lists unless there are six or more, in which case only the first author should be given, followed by ‘et al.’. Authors should be listed last name first, followed by a comma and initials (followed by full stops, '.') of given names. Article titles should be in Roman text; only the first word of the title should have an initial capital and the title should be written exactly as it appears in the work cited, ending with a full stop. Book titles should be given in italics and all words in the title should have initial capitals. Journal names are italicized and abbreviated (with full stops) according to common usage. Volume numbers and the subsequent comma appear in bold. The full page range should be given where appropriate. See the examples below:

**Journal Article**:

1. Schott, D. H., Collins, R. N. & Bretscher, A. Secretory vesicle transport velocity in living cells depends on the myosin V lever arm length. *J. Cell Biol*. **156**, 35‐39 (2002).

**Book** ‐ Book titles should be given in italics and all words in the title should have initial capitals:

1. Hogan, B. *Manipulating The Mouse Embryo: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, 1994)

**Publicly available preprint:**

1. Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of nonlocal single photon. Preprint at http://arXiv.org/quant-ph/0208066 (2002).

**Code:**

1. Gallotti, R. & Barthélemy, M. Source code for: The multilayer temporal network of public transport in Great Britain. *Figshare* https://dx.doi.org/10.6084/m9.figshare.1249862.v1 (2014).

**Online material** ‐ Stable documents hosted on the web may be cited in the main reference list, using the format below. Websites or dynamic web resources should be cited by embedding the URL in the main article text:

1. Manaster, J. Sloth squeak. *Scientific American Blog Network* http://blogs.scientificamerican.com/psi-vid/2014/04/09/sloth-squeak (2014).

**Technical or government report:**

1. Akutsu, T. *Total Heart Replacement Device.* Report No. NIH-NHLI-69 2185-4 (National Institutes of Health, 1974).

## Citing Data

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference.

An author list (formatted as above) and title for the dataset should be included in the data citation, and should reflect the author(s) and dataset title recorded at the repository. If author or title is not recorded by the repository, these should not be included in the data citation. The name of the data-hosting repository, URL to the dataset and year the data were made available are required for all data citations. For DOI-based (e.g. figshare or Dryad) repositories the DOI URL should be used. For repositories using accessions (e.g. SRA or GEO) an [identifiers.org](https://identifiers.org/) URL should be used where available. For first submissions, authors may choose to include just the accession number. Scientific Data staff will provide further guidance after peer-review. Please refer to the following examples of data citation for guidance:

1. Zhang, Q-L., Chen, J-Y., Lin, L-B., Wang, F., Guo, J., Deng, X-Y. Characterization of ladybird Henosepilachna vigintioctopunctata transcriptomes across various life stages. figshare <https://doi.org/10.6084/m9.figshare.c.4064768.v3> (2018).
2. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRP121625> (2017).
3. Barbosa, P., Usie, A. and Ramos, A. M. Quercus suber isolate HL8, whole genome shotgun sequencing project. GenBank<http://identifiers.org/ncbi/insdc:PKMF00000000> (2018).
4. DNA Data Bank of Japan <http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA004814> (2016).

**Depositing your data to an appropriate repository**

Your *Scientific Data* manuscript will not be sent to review unless the dataset(s) described therein have been deposited in an appropriate public repository ([please see our list of recommended repositories](http://www.nature.com/sdata/policies/repositories)). Should a specific repository not be available for your field or data-type, or should the repository of your choice not permit confidential peer-review, you may upload your data to one of our [recommended generalist repositories](https://www.nature.com/sdata/policies/repositories#general). Integrated submission systems are available for both figshare and Dryad.