\subsection\*{Study Area}

Our study area includes 11 western US states: Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming (Figure \ref{fig:MonitorLocations}). Our temporal domain were all days between January 1, 2008 and December 31, 2018. We predicted daily estimates of PM\textsubscript{2.5} at the county, ZIP code, and census tract levels from machine learning ensembles trained on observed daily PM\textsubscript{2.5} values from monitoring stations from a variety of sources (\*\*put in all PM\textsubscript{2.5} data sources). The predictor variables for the machine learning ensemble included PM\textsubscript{2.5} 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 at 2 meters, dew point temperature at 2 meters, U- and V- components of wind speed at 10 meters, surface pressure, pressure reduced to mean sea level, and "DZDT\_850\_mb" and "DZDT\_700\_mb" %Melissa, what were these??

More information on the sources of these data can be found in Table 1 \ref{tab:Table1}.

%Example of citation: \cite{liu\_estimating\_2005}

\subsection\*{PM\textsubscript{2.5} Measurements}

To get a more comprehensive set of locations and time points of PM\textsubscript{2.5} measurement throughout the western US, we did an extensive search for as many PM\textsubscript{2.5} monitoring data within our spatial and temporal study area as we could find. We downloaded PM\textsubscript{2.5} data from the US EPA AQS Air Data Query Tool \cite{EPAAirData2017} for the 11-state region (Figure \ref{fig:MonitorLocations}) including any of the following parameter codes: 88101, 88500, 88502, 81104 \cite{EPANPM25Memo2017,EPANPM25Parameters2017,EPANAllParameters2017}. These data include the IMPROVE monitors that capture air quality information in more rural areas \cite{EPANPM25IMPROVE2017}. We also retrieved all available PM\textsubscript{2.5} data in the Fire Cache Smoke Monitor Archive (\url{https://wrcc.dri.edu/cgi-bin/smoke.pl}), which includes U.S. Forest Service monitors that were deployed to capture air quality impacts during wildfire events.

Some states have additional PM\textsubscript{2.5} 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 PM\textsubscript{2.5} data from California Air Resources Board and the Utah Department of Environmental Quality. We only included data that was in addition to the monitors in those states that was part of the U.S. EPA's AQS and IMPROVE data.

We also reached out to researchers who may have had their own monitoring networks of PM\textsubscript{2.5} throughout the region. We were able to obtain data from the Uintah Basin, Utah from Seth Lyman at Utah State University, and PM\textsubscript{2.5} measurements from the Persistent Cold Air Pool Study (PCAPS) \cite{Silcox\_wintertime\_2012} conducted in the Salt Lake Valley, Utah in January--February, 2011 from Dr. Geoff Silcox at the University of Utah.

All of this yielded a total of 1,591,533 daily PM\textsubscript{2.5} observations, which represent 7,754 locations and 4,006 days.

\subsection\*{Predictor Variables}

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

Satellite Aerosol Optical Depth (AOD) is a measure of particle loading in the atmosphere from the ground to the satellite. We obtained daily estimates of %Aerosol Optical Depth (AOD)from

AOD from the MODIS Terra and Aqua combined Multi-angle Implementation of Atmospheric Correction (MAIAC) dataset \url{https://ladsweb.modaps.eosdis.nasa.gov/archive/allData/6/MCD19A2/}. 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 nearest neighbor value at each PM\textsubscript{2.5} monitoring location. MAIAC AOD has been shown to better predict PM\textsubscript{2.5} than coarser resolution AOD \cite{chudnovsky\_spatial\_2012} and has been used in many studies in various geographic regions in blended models to predict daily PM\textsubscript{2.5} \cite{lee\_benefits\_2019, geng\_satellite-based\_2018-1, li\_using\_2018}.

We obtained meteorological data from the North American Mesoscale (NAM) Analysis meteorological model \url{https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-mesoscale-forecast-system-nam} because it includes all of the standard meteorological variables, including planetary boundary layer height, which play a role in PM\textsubscript{2.5} levels and can be important to help scale AOD values to ground-level estimates of PM\textsubscript{2.5} \cite{liu\_estimating\_2005}. 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.

machine learning models for PM2.5 \cite{di\_\*\*\*, reid\_2015\*\*\*\*} We were able to obtain daily estimates of PM\textsubscript{2.5} at XXXX spatial resolution from runs of the CMAQ (Community Multi-scale Air Quality) model from the U.S. EPA for the years 2008-2016 \cite{\*\*\*}.

Because one of the reasons that PM\textsubscript{2.5} concentrations have been increasing in the western US is the increasing number and magnitude of wildfires, we wanted to have 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) \cite{Giglio2006,Hawbaker2017}.

%We collected daily data about fire detection locations, size, and fire radiative power from the MODIS Thermal Anomalies/Fire Daily L3 Global 1km product (MOD14 and MYD14) \cite{Giglio2006,Hawbaker2017}.

As fires in closer proximity are likely to influence PM\textsubscript{2.5} 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.

%since we did not use FRP, should we remove this? Is FRP a variable in our model?

Elevation can influence PM\textsubscript{2.5} concentrations. For example, PM\textsubscript{2.5} can accumulate in mountain valleys during persistent cold air pools (commonly referred to as inversions)

during winter \cite{Whiteman2014}. 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 \cite{USGSElevation2017}.

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) \cite{Homer2017} 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 MOD13A3 \url{https://lpdaac.usgs.gov/products/mod13a3v006/} at 1 km resolution by month as another measure of vegetation that was not just a measure of agricultural vegetation but all vegetation.

%To estimate emissions from vehicles,

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 Network \url{https://www.fhwa.dot.gov/planning/processes/tools/nhpn/index.cfm} which contains spatial information on over 450,000 miles of highways in the United States.

We included population density as an additional proxy for emisions 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 PM\textsubscript{2.5} 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. One of the biggest challenges in this study was developing a model that worked well spatially and temporally across the study domain. 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)) and state; and 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” [8].

%Finally, we created an indicator variable for whether there were one or more fires within 500 km of a monitor in the last week.

\subsection\*{Data merging}

We created three datasets: one dataset to train the model and two prediction datasets. The training dataset merged all predictor variables to each 24-hour average PM\textsubscript{2.5} 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.

\subsection\*{Machine learning modeling }

For the machine learning modeling, we took the full training data set and separated it into separate training and testing data sets. With this setup, the results of 10-fold cross-validation (with no resampling) from training are used as validation metrics, while the results of each model applied to the testing set are a measure of how well the model will perform when predicting at new locations and that were not part of the training data, as will be the case with our prediction data set.

Using 10-fold cross-validation (with no resampling) for this kind of land-use regression is standard practice, as shown in Di et al. (2019) [1]. This type of cross validation selects data points randomly from the training set for each of the cross-validation folds. This method, however, violates the assumption of independence between folds because of repeated observations (on different days) from the same locations (PM2.5 sensor locations). Spatial block cross-validation, whereby all observations from a given monitoring site are within the same fold, is a more appropriate tool for evaluating the accuracy of a model when predicting PM2.5 at new locations [2]. We also randomly selected 10% of the monitoring locations for the held-out test set monitoring. We refer to the model and results generated using this procedure as the spatial-folds model and results. We also ran our models using random observations for the test-set and within the 10 folds solely for comparison of results to previous studies that only reported results from random folds. We refer to these as the random-folds model and results.

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 block folds and random folds. 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). 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* [3].

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 (ranger)and a gradient boosting model (xgbTree). These models performed best on preliminary analyses of random subsets of our dataset. All analyses were run using R [4], and all machine learning models utilized the R packages caret [5] and caret ensemble [6]. Variable importance was calculated using the “permutation” importance algorithm in the caret package.

The CMAQ dataset was only available from 2008-2016. Because of our interest in the years 2017 and 2018, when there were many large wildfires in the western US, we ran all the models in this analysis on the 2008-2018 data without CMAQ as well as on the 2008-2016 data with CMAQ.

When examining the predicted PM2.5 values compared with the observed PM2.5 values, we noted that the models were performing much worse on high values than on low 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 did a sensitivity analysis in which we examined whether a different model would perform better for low values than high values. This did not prove to significantly improve predictive performance. A more detailed description along with the results of the split analysis (“high” versus “low”) are in the Supplementary Material.

**Results**

Table XX shows the mean and quantiles of PM2.5 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Mean** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
| 2008 | 8.998 | 0 | 3.8 | 6.6 | 11.4 | 200.2 |
| 2009 | 8.329 | 0 | 3.7 | 6.3 | 10.6 | 195.583 |
| 2010 | 7.441 | 0 | 3.3 | 5.7 | 9.6 | 114 |
| 2011 | 8.316 | 0 | 3.5 | 6.2 | 10.625 | 208.025 |
| 2012 | 8.156 | 0 | 3.6 | 6.2 | 10.2 | 705.458 |
| 2013 | 8.559 | 0 | 3.7 | 6.5 | 10.635 | 452.792 |
| 2014 | 7.85 | 0 | 3.458 | 6 | 9.813 | 504.542 |
| 2015 | 7.648 | 0 | 3.2 | 5.5 | 9.2 | 830.792 |
| 2016 | 6.929 | 0 | 3 | 5.2 | 8.7 | 804.5 |
| 2017 | 8.938 | 0 | 3.2 | 5.7 | 10 | 811.792 |
| 2018 | 9.193 | 0 | 3.6 | 6 | 10 | 826.292 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **State** | **Mean** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
| Arizona | 6.365 | 0 | 3.3 | 5.3 | 7.9 | 199.3 |
| California | 10.269 | 0 | 4.95 | 8.2 | 12.625 | 791.625 |
| Colorado | 5.66 | 0 | 2.3 | 4.5 | 7.3 | 781.455 |
| Idaho | 7.949 | 0 | 3 | 5.3 | 8.9 | 519.391 |
| Montana | 7.431 | 0 | 2.5 | 4.8 | 8.3 | 641.9 |
| Nevada | 6.829 | 0 | 3.5 | 5.7 | 8.4 | 230 |
| New Mexico | 5.231 | 0 | 2.5 | 4.15 | 6.6 | 263 |
| Oregon | 7.5 | 0 | 3.1 | 4.8 | 8.15 | 811.792 |
| Utah | 7.202 | 0 | 3.3 | 5.2 | 8.2 | 225.25 |
| Washington | 6.546 | 0 | 2.95 | 4.7 | 7.6 | 830.792 |
| Wyoming | 4.311 | 0 | 1.8 | 3.2 | 5.2 | 513.417 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **State** | **Mean** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
| Fall | 8.643 | 0 | 3.7 | 6.375 | 10.5 | 830.792 |
| Spring | 6.144 | 0 | 3.1 | 5 | 7.9 | 189.9 |
| Summer | 8.76 | 0 | 4 | 6.4 | 10.25 | 712.917 |
| Winter | 9.299 | 0 | 3.043 | 6.5 | 12.2 | 557 |

Table XX shows the performance metrics (RMSE and R2) of our ensemble machine learning models with spatial folds and random folds. The models with the 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Training (10-fold CV) RMSE (µg/m3)** | **Training (10-fold CV) R2** | **Testing RMSE (µg/m3)** | **Testing R2** |
| CMAQ, spatial folds | 5.061 | 0.659 | 5.420 | 0.589 |
| Non-CMAQ, spatial folds | 6.576 | 0.598 | 6.599 | 0.593 |
| CMAQ, random folds | 4.482 | 0.732 | 4.642 | 0.715 |
| Non-CMAQ, random folds | 5.482 | 0.719 | 5.954 | 0.680 |

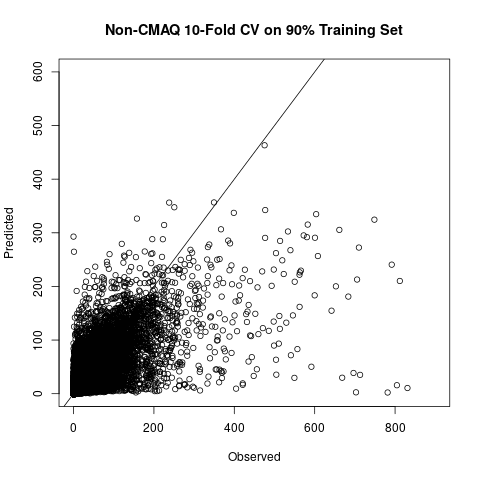
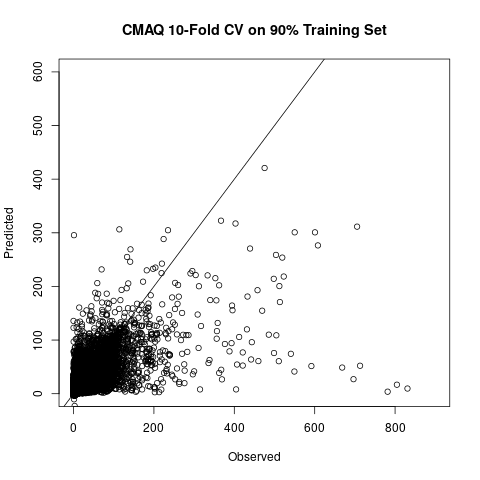
Overall, the models including CMAQ perform better (have lower RMSE and higher R2 values) than the models not including CMAQ. This may be due to the additional information provided by the CMAQ output or could be because the models without CMAQ include two additional years of data that had more high PM2.5 days, which are much harder to predict accurately than lower values. When we ran a model without CMAQ on the years 2008-2016 only, the results were slightly better than those from a model without CMAQ that included 2017 and 2018, but still worse than the model with CMAQ.

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, however, can 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 of the testing data set not being used to inform the model; some of the discrepancy is because of random chance of a given monitoring site being in the testing data set.

The performance metrics for our full models (without any cross-validation folds and that we use for prediction) on the CMAQ and non-CMAQ datasets are, respectively, RMSE = 1.726 and R2 = 0.960; RMSE = 2.027 and R2 = 0.961. These are much better performance metrics than any of those in the tables below because the full models train on the whole data set, and thus are not realistic representations of how accurately our model will be able to predict at locations outside of our training set, which is best represented by the spatial folds models. 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 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.



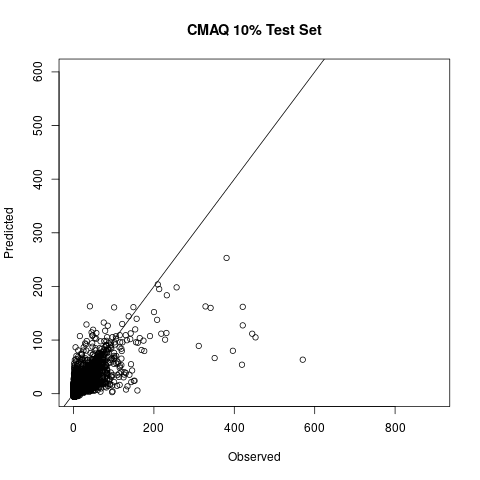
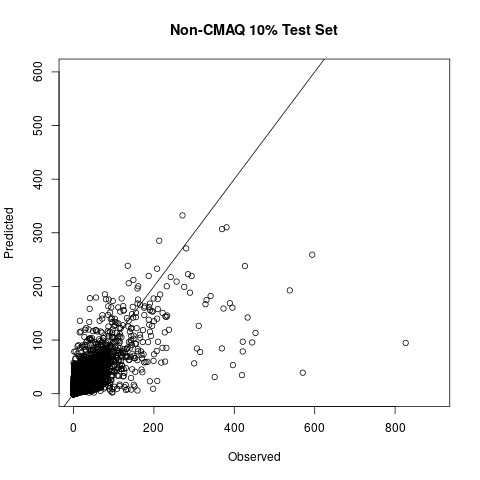
 

Table XX shows the RMSE values of our models on different levels of PM2.5, years, states, and seasons. Similar tables with R2 values for these data can be found in the Supplementary Material. We observe that the models with CMAQ always perform better than the models without CMAQ. There is not such a clear pattern when comparing the results on the training (10-fold CV) and testing (completely held-out) data sets. This likely represents the reality that our models will perform better in some places and worse in other places in our prediction set.

In the spatiotemporal subsets, our data show that we have better predictive performance at lower levels of PM2.5. This is likely because of a much higher number of observations at lower values that 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 |

**Discussion**

This dataset provides daily PM2.5 estimates at the centroids of each county, ZIP code, and census tract across the western US, from 2008-2018. We used completely open-source code and publicly available data sets to make these estimates. Our ensemble machine learning models were evaluated using spatial cross-validation techniques to estimate accuracy at locations where there are no PM2.5 monitors. For the 2008-2016 model including CMAQ, we achieved a 10-fold CV (training set) RMSE of 5.061 and R2 of 0.659 and a testing set RMSE of 5.420 and R2 of 0.589. For the 2008-2018 model without CMAQ, we achieved a 10-fold CV RMSE of 6.576 and R2 of 0.598 and a testing set RMSE of 6.599 and R2 of 0.593.

All models including CMAQ performed better than their counterpart models without CMAQ, even after removing the years 2017 and 2018 from the non-CMAQ model. Thus, we conclude that CMAQ is an important variable in this land-use regression. When CMAQ is not available, MAIAC AOD is the closest proxy. This intuition is supported by the fact that the variable importance of MAIAC AOD drops much more substantially than any of the other variables after adding CMAQ into the model. (Note that while collinearity between variables does not matter for prediction with random forest, it most likely influences the variable importance calculations via permutation (Gregorutti et al., 2017).)

One of the strengths of this study is that we focused on estimating PM2.5 values associated with wildfires. However, our models still dramatically underpredict high observations of PM2.5 (above 200 µg/m3), likely because we have so much less training data at these high values. To address this, we did a sensitivity analysis in which we examined whether a different model would perform better for high values than for low values. Although we were not able to see significant improvements, this may still be a promising direction for future research.

Another method that would likely improve our results would be to incorporate a two-step model such as that used by Di et al. [1], where spatially- and temporally-lagged estimates (from machine learning regression) of PM2.5 were fed into a second machine learning regression model. We did investigate the use of spatio-temporal kriging of the model residuals for years 2009-2010, but ultimately found that it was not useful for these years and abandoned the approach.

**Conclusion**

[Summarize and conclude that there is still need for improved modelling of high PM2.5 levels, but this dataset should help to inform research about air pollution in the western US over a decade with increasing wildfires and smoke.]

**References**

[1] Q. Di, “An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution,” *Environ. Int.*, p. 13, 2019.

[2] D. R. Roberts *et al.*, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, vol. 40, no. 8, pp. 913–929, Aug. 2017, doi: 10.1111/ecog.02881.

[3] I. Kloog, P. Koutrakis, B. A. Coull, H. J. Lee, and J. Schwartz, “Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements,” *Atmos. Environ.*, vol. 45, no. 35, pp. 6267–6275, Nov. 2011, doi: 10.1016/j.atmosenv.2011.08.066.

[4] R Core Team, *R software*. .

[5] M. Kuhn, *caret package (R)*. Journal of Statistical Software, 28(5), 2008.

[6] Z. A. Deane-Mayer and J. E. Knowles, *caret ensemble package (R)*. 2019.

[7] Y. Xu *et al.*, “Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5,” *Environ. Pollut.*, vol. 242, pp. 1417–1426, Nov. 2018, doi: 10.1016/j.envpol.2018.08.029.

[8] M. Katzfuss, “A Multi-Resolution Approximation for Massive Spatial Datasets,” *J. Am. Stat. Assoc.*, vol. 112, no. 517, pp. 201–214, Jan. 2017, doi: 10.1080/01621459.2015.1123632.

*Fix these references above, later:*

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

M. Kuhn. Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan and T. Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

Z.A. Deane-Mayer and J.E. Knowles (2016). caretEnsemble: Ensembles of Caret Models. R package version 2.0.0. https://CRAN.R-project.org/package=caretEnsemble