**Methods**

\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 \*\*\*, 2018. We predicted daily estimates of PM\textsubscript{2.5} at the ZIP code and county 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 (\*\*put in all variables here) %Ellen.

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 XX daily PM\textsubscript{2.5} observations, which represent XX locations. %Ellen

\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.

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.

%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 American Community Survey at the XXX (spatial resolution) for each year or five year averages?

%Ellen, could you add to this about population density - I don't think that it is in our documentation

To account for seasonality in PM\textsubscript{2.5} data, we created the following predictor variables: cosine of day-of-year and cosine of month. %Ellen, can you explain why you chose cosine for these?

We also created dummy variables for each state and month 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.

%Let's move this comment below to the machine learning methods section - done

%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 ZIP code and county for each day in the study domain.

\subsection\*{Machine learning modeling and mapping}

Within the training dataset, we first created 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 on new data. The reason for using a completely held-out testing set is that this data was not used to inform the choice of the model, and therefore represents new data (in our prediction set) better than data that was used to develop the model.

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]. Randomly selecting data points for each of the cross-validation folds, however, violates the assumption of independence between folds because of repeated observations (on different days) from the same locations (PM2.5 sensor locations). Thus, spatial block cross-validation is a more appropriate tool for evaluating the accuracy of our model when predicting PM2.5 at new locations [2]. Instead of randomly selecting data points for our testing and training sets, we randomly selected 10% of the monitoring locations for the held-out test set monitoring and used the remaining monitoring locations in the testing set, also ensuring that all observations from each monitoring location were in the same fold. We refer to the model and results generated using this procedure as the spatial-folds model and results. For comparison to previous studies that used random observations in their folds rather than clustering spatially, we ran a sensitivity analysis using random (non-spatial) folds, referring to these as the random-folds model and results.

, for both the 10-fold cross-validation and for the left-out testing data setAlso for comparison to other studies, we calculated the “spatial R2” and “temporal R2” metrics used by Di et al. (2019). 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.

To model PM2.5 exposures across the western US, we employed ensemble machine learning. Specifically, we used a generalized linear model (GLM) to combine (“ensemble”) the results from a random forest model and a gradient boosting model. All analyses were run using R version 3.5.2, and all machine learning models utilized the R packages caret [3] and caret ensemble [4].

Preliminary analysis on random subsets of the data (exploring many different types of algorithms compatible with the caret package) suggested that a random forest implementation called *ranger* and a gradient boosting implementation called *xgbTree* (extreme gradient boosting tree), both available within the caret package, performed the best. This aligns with the findings of Xu et al. (2018), who found that tree-based models (using random forest, gradient boosting, and cubist algorithms) performed the best in this kind of land-use regression [5]. Then, we used the same random subsets of the data to tune hyperparameters for each algorithm via a grid-search (see code in Supplementary Materials).

One of the biggest challenges in this study was developing a model that worked well spatially and temporally across the study domain. To help capture nonlinear spatiotemporal effects, we included nested levels of spatiotemporal variables. Temporally, we included 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. Spatially, we included variables for region (within the 11 western states: northwest, southwest, four corners, and northern mountain states); 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” [6]. We also investigated the use of spatio-temporal kriging of the model residuals for 2009-2010, but ultimately found that it was not useful for these years and abandoned the approach.

We did a sensitivity analysis in which we examined whether a different model would perform better for low values than high values

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. In addition to the fact that we had a lot more data at low values than at 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 developed a preliminary classification model to split the data into “high” versus “low” values, with 15µg/m3 being the most plausible and accuratecut-point. This did not prove to significantly improve predictive performance. The results of the split analysis (“high” versus “low”) are in the Supplementary Materials.

We ran all of the models on the 2008-2018 data without CMAQ as well as on the 2008-2016 data with CMAQ.

**Results**

What tables and figures will we want to include in the body of the paper? FVO? Map(s)?

We note that because our models tend to dramatically underpredict the highest observations (above 300 µg/m3), which tend to be associated with wildfires, RMSE is a more illustrative metric than R2.

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

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

[5] 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.

[6] 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.