Machine Learning Models Accurately Model Ozone Exposure during Wildfire Events

Gregory L. Watson[a,*], Donatello Telesca[a], Colleen E. Reid[b], Gabriele G. Pfister[c], Michael Jerrett[d]

[a] *Department of Biostatistics, University of California, Los Angeles, California 90024, USA*
[b] *Department of Geography, University of Colorado Boulder, Boulder, Colorado 80309, USA*
[c] *Atmospheric Chemistry Observations and Modeling Laboratory, National Center for Atmospheric Research, Boulder, Colorado 80301, USA*
[d] *Department of Environmental Health Sciences, University of California, Los Angeles, California 90024, USA*

Abstract

Epidemiologists use prediction models to downscale (i.e., interpolate) air pollution exposure where monitoring data is insufficient. This study compares machine learning prediction models for ground-level ozone during wildfires, evaluating the predictive accuracy of ten algorithms on the daily 8-hour maximum average ozone during a 2008 wildfire event in northern California. Models were evaluated using a leave-one-location-out cross-validation (LOLO CV) procedure to account for the spatial and temporal dependence of the data and produce more realistic estimates of prediction error. LOLO CV avoids both the well-known overly optimistic bias of k-fold cross-validation on dependent data and the conservative bias of evaluating prediction error over a coarser spatial resolution via leave-k-locations-out CV. Gradient boosting was the most accurate of the ten machine learning algorithms with the lowest LOLO CV estimated root mean square error (0.228) and the highest LOLO CV $\hat{R}^2$ (0.677). Random forest was the second best performing algorithm with an LOLO CV $\hat{R}^2$ of 0.661. The LOLO CV estimates of predictive accuracy were less optimistic than 10-fold CV estimates for all ten models. The difference in estimated accuracy between the 10-fold CV and LOLO CV was greater for more flexible models like

*Corresponding author
Email address: gwatson@ucla.edu (Gregory L. Watson)

gradient boosting and random forest. The order of estimated model accuracy depended on the choice of evaluation metric, indicating that 10-fold CV and LOLO CV may select different models or sets of covariates as optimal, which calls into question the reliability of 10-fold CV for model (or variable) selection. These prediction models are designed for interpolating ozone exposure, and are not suited to inferring the effect of wildfires on ozone or extrapolating to predict ozone in other spatial or temporal domains. This is demonstrated by the inability of the best performing models to accurately predict ozone during 2007 southern California wildfires.

Capsule: Flexible machine learning methods model ozone well during a wildfire. LOLO CV more accurately estimates prediction error than 10-fold CV.

Keywords: Air Pollution, Exposure Model, Machine Learning, Ozone, Wildfire

1. Introduction

Ground-level ozone is toxic to humans, animals and plants and contributes significantly to climate change as the third most important greenhouse gas [1, 2, 3, 4, 5, 6, 7]. Short-term exposure is linked to increased mortality [3], decreased respiratory function, exacerbation of chronic obstructive pulmonary disease (COPD), bronchitis, emphysema and asthma [8, 9, 10, 11]. Long-term exposure has been linked with respiratory and cardiovascular mortality [12, 13], decreased lung function [14] and the progression of emphysema [15].

Wildfires contribute to the formation of ozone in the lower atmosphere (troposphere) by releasing volatile organic compounds (VOCs) and nitrogen oxides (NOx), which react in the presence of sunlight to form ozone [16, 17]. Fires upwind of large population centers can expose millions or even tens of millions of people to ozone and other pollutants [18]. Climate change is expected to intensify wildfires, which will likely increase the prevalence of wildfire-related ozone exposure.[17, 19]

The health effects of wildfire-induced ozone exposure are poorly understood.

A study of hospital admissions in Port, Portugal, in 2005 while wildfires were burning nearby, indicated ozone was significantly associated with cardiovascular disease admissions, but not with respiratory admissions [20]. This analysis, however, did not control for weather or land-use covariates. During a bushfire in southeastern Australia, respiratory emergency department visits were significantly associated with PM_{10} (particulate matter 10 μm or smaller in diameter), but not with ozone [21].

A key challenge facing epidemiological analyses of air pollution exposure is quantifying pollution concentrations where people live, which may be distant from regulatory monitoring sites. This is particularly difficult for wildfire-related pollution, because wildfires often ignite far from urban regulatory monitoring sites, and satellite evidence indicates that traditional monitoring networks are too sparse to capture smoke plume variation and dynamics [22]. Epidemiologists attempt to overcome this difficulty by constructing exposure models to predict pollution concentration at unmonitored locations and times. The prediction of a quantity across a domain, such as a spatial region, based on observations of that quantity at discrete locations within that domain is referred to as downscaling or interpolation, and is an example of infill prediction.

The simplest downscaling exposure models rely upon the tendency of nearby observations to be more similar than those farther apart to interpolate between pollution monitor observations without the use of additional information (i.e., without covariates). This tendency is an example of spatial (or space-time) dependence. Kriging is a very commonly used method for interpolating observations, modeling air pollution concentrations as the best linear unbiased prediction (BLUP) given the data and mean and covariance functions selected by the researcher and estimated from the data [23]. Kriging tends to perform relatively well when monitoring data is dense, but model accuracy degrades at locations or times distant from monitor observations.

To improve accuracy, especially when monitoring density is sparse, researchers have employed regression models that incorporate ancillary information as covariates. These models are referred to as land use regression in the literature,

but the covariates need not pertain to land use, and increasingly include satellite retrievals, meteorological data, and less frequently the output of atmospheric chemistry numerical simulation models. Land use regression models have been used for modeling air pollution exposure at least since the Small Area Variations in Air quality and Health (SAVIAH) study in 1997 [24] with numerous examples appearing subsequently. While these regression models include covariate information, they have often assumed linear, additive covariate effects. This assumption makes the effects easy to interpret but is too stringent to predict air pollution concentrations accurately. These models cannot accommodate nonlinear effects or interactions between covariates unless they are specified by the analyst a priori. They also lack a mechanism for variable selection, requiring the analyst to manually select covariates or employ a separate variable selection procedure.

These limitations have prompted the development of more flexible models that allow for nonlinear effects including spatially or spatiotemporally varying-coefficient models [25]. These models generally fit covariate effects with smooth functionals that need not be linear and may be indexed by space or space-time, which allows the covariate effects to differ across space and time. These models are an improvement over the very stringent restrictions set on covariate effects in linear, additive models, but they often rely on research code, making them less accessible to other researchers. In our experience we have also found that these more sophisticated models do not scale well to realistic space-time data settings.

These challenges have motivated researchers to turn to more flexible models that do not require such stringent assumptions, including a variety of machine learning algorithms, which have been shown to be very useful for prediction [26], especially random forest [27], gradient boosting [28] and neural networks [29]. They may lack the straightforward interpretability of linear regression, but this is of secondary concern when prediction is the primary objective, and variable importance scores have been developed for many such methods to quantify the contribution of each covariate.

Generalized additive models [30], support vector machines [31, 32], gradient boosting [33] and deletion/substitution/addition [34] have been used to model particulate matter exposure. Neural networks [35, 36] and random forest [37, 38, 39] have been used to model both particulate matter and ozone concentrations. A comparison of 11 machine learning models indicated that random forest, gradient boosting and bagged trees predict $PM_{2.5}$ (particulate matter smaller than 2.5 μm in diameter) concentrations well during a wildfire event [22]. Random forest, boosting and Cubist performed well in a comparison of 8 machine learning tools predicting $PM_{2.5}$ in British Columbia [40]. Here we conduct a similar analysis using ten machine learning algorithms to model ozone exposure during a wildfire air pollution event for the first time, evaluating their predictive accuracy for use as land use regression models.

Comparing and evaluating prediction models for dependent data is challenging. Cross-validation (CV) and the bootstrap are commonly used model evaluation procedures that repeatedly fit a model to a training subset of the data and evaluate the accuracy of its predictions on a different, test subset, combining the performance across multiple test subsets into a nonparametric estimate of prediction error. For data that are spatially and temporally dependent (i.e., autocorrelated), however, these procedures can be overly optimistic, because of the dependence between training and test subsets [41].

In the case of daily air pollution monitoring observations, we wish to estimate the average error made by a downscaling model when predicting at a new location within the spatial domain of the data. Including observations in the training data that were taken at the same monitors as the test set observations provides an unrealistic amount of information on the test data, because of the strong correlation between observations taken at the same location. This produces estimates of prediction error that are biased downward, especially for flexible models which tend to overfit to a greater degree than less flexible models when trained on dependent data. The true prediction error associated with predicting at a new location would be greater than these estimates, because the model could not have been trained on any observations recorded at a new

location.

When dependence is restricted to observations within the same group or cluster, consistent (i.e., asymptotically unbiased) estimates of prediction error can be recovered by resampling groups rather than individual observations. It is unrealistic, however, to assume that spatially dependent data are nested within independent groups of observations. Modified cross-validation schemes have been used on pollution exposure data that partition the data into spatial grid cells [33] or monitor locations [42, 43]. Such approaches attempt to reduce the dependence between training and test data sets by placing all observations at a particular location or within a particular region into the same cross-validation fold. In this vein, we use leave-one-location-out (LOLO) cross-validation, which defines each CV fold as the observations recorded at a single monitor location [44]. This does not partition the data into independent groups, but estimates the error associated with predicting the time series of ozone observations at a new location, conditioning upon the observed monitor data. By using all the observations at a single location as the test set, LOLO CV ensures that no observations from this location appear in the training set, which would result in unrealistically low estimates of prediction error. It also avoids the overly conservative bias of leave-k-locations-out CV, which tends to overestimate prediction error because it uses substantially fewer observations for model training.

2. Materials and Methods

2.1. Data

One hundred ground-based ozone monitors administered by the United States Environmental Protection Agency (EPA) made hourly observations from which the daily maximum 8-hour averages were computed across northern California between May 6, 2008 and September 26, 2008 for a total of 13,487 observations. We selected this time period with the goal of estimating ozone exposures before, during, and after a spate of wildfires that afflicted northern California in late June and July of 2008. The mean maximum 8-hour average concen-

Table 1: Covariates used to predict ozone.

Covariate	Data Source
Monitor Latitude	U. S. Environmental Protection Agency
Monitor Longitude	U. S. Environmental Protection Agency
Elevation (m)	National Digital Elevation Model
Date	U. S. Environmental Protection Agency
Dew Point (°K)	Rapid Update Cycle
Boundary Layer Height (m)	Rapid Update Cycle
Surface Pressure (Pa)	Rapid Update Cycle
Relative Humidity (%)	Rapid Update Cycle
Temperature at 2 m (°K)	Rapid Update Cycle
U-Component of Wind Speed (m/s)	Rapid Update Cycle
V-Component of Wind Speed (m/s)	Rapid Update Cycle
Inverse Distance to Nearest Fire (m^{-1})	Fire Inventory from NCAR v1.5
Annual Average Traffic within 1 km	Dynamap 2000, TeleAtlas
Agricultural Land Use within 1 km (%)	2006 National Land Cover Database
Urban Land Use within 1 km (%)	2006 National Land Cover Database
Vegetation Land Use within 1 km (%)	2006 National Land Cover Database
Normalized Difference Vegetation Index	Landsat Data
Nitrogen Dioxide (log molecules/cm^2)	Ozone Monitoring Instrument Satellite
WRF-Chem Carbon Monoxide (log moles/day)	WRF-Chem
WRF-Chem $PM_{2.5}$ (log kg/day)	WRF-Chem
WRF-Chem Ozone (log 8 Hour Maximum)	WRF-Chem

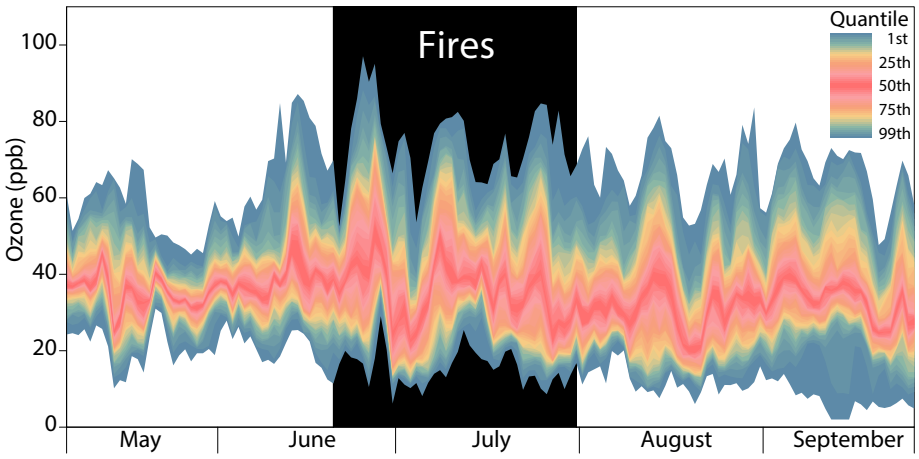


Figure 1: The daily empirical distribution of maximum daily 8-hour average ozone between May 6 and September 26, 2008 at 100 northern California monitoring sites.

tration was 36.2 ppb, and the standard deviation was 13.6 ppb. During the study, the maximum 8-hour average exceeded 70 ppb 236 times and exceeded 75 ppb 107 times. Most exceedances occurred while the fires were burning (153 and 75 respectively), although this time period is one of high solar intensity when high concentrations of ozone are expected. It is not our objective, however, to quantify the contribution of wildfires to ozone formation, but simply to downscale ozone concentrations during a wildfire event for subsequent epidemiological analysis. Figure 1 depicts the temporal evolution of monitor observations throughout this time period.

Twenty-one covariates were also collected for the monitor locations, including location, elevation, date, atmospheric weather data (dew point, boundary layer height, surface pressure, relative humidity, temperature, and wind speed), inverse distance to the nearest fire, traffic, land use information (agricultural, urban, and vegetation), tropospheric nitrogen dioxide (NO_2) vertical column density and predictions of daily total carbon monoxide concentration (CO), particulate matter ($PM_2.5$) and daily maximum 8-hour average ozone. Table 1 lists the covariates and their sources.

Monitor elevation was determined from the 2010 National Elevation Dataset

for California. The date of each observation was encoded as the continuous covariate, Julian date. The U.S. National Centers for Environmental Prediction's Rapid Update Cycle atmospheric prediction model provided hourly predictions of dew point, planetary boundary layer height, surface pressure, relative humidity, temperature, and the U and V components of wind speed, which were averaged into daily values [45].

Inverse distance to the nearest fire was included as a covariate. The Fire Inventory from NCAR (FINN) v1.5 provided estimates of fire point locations in California during the study period [46]. Fire points occurring within 5 km of each other were clustered and circumscribed by a polygon using the ArcGIS Aggregate Points tool, and the distance between each monitoring site and the closest point on the nearest fire cluster polygon was determined on each day using the ArcGIS Near tool. On days with no fire in California, distance to the nearest fire was undefined. Conceptualizing this undefined distance as equivalent to the nearest fire cluster being infinitely far away, inverse distance to fire was defined as 0 for observations taken on days with no fires in California and as the inverse of the distance to the nearest fire cluster otherwise.

Dynamap 2000, a TeleAtlas product, was used to compute the annual average of roadway traffic within 1 km of each monitor [22]. The National Land Cover Database for 2006 [47] was used to calculate the percentage of urban development (codes 22, 23, and 24), agriculture (codes 81 and 82) and other vegetation (codes 21, 41, 42, 43, 52, and 71) within 1 km of each monitor.

The normalized difference vegetation index (NDVI) quantifies the density of green vegetation on a scale between -1 and 1 by measuring the visible and near-infrared light reflected at a location via remote sensing. The chlorophyll in healthy vegetation absorbs most of the visible light and reflects much of the near-infrared light to which it is exposed, giving locations with more vegetation a higher NDVI score. NDVI for each monitor location was extracted from the NDVI remote sensing raster surface and included as a covariate.

Nitrogen dioxide (NO_2) was estimated on each day at monitor locations (if available) using the Berkeley High-Resolution (BEHR) NO_2 tropospheric

column density retrieved from NASA's Ozone Monitoring Instrument (OMI) satellite, which has an overpass time of 1:30 local time [48] and a resolution varying between 13 x 24 km to 42 x 162 km.

Predictions of daily total carbon monoxide (CO) and $PM_{2.5}$ and the maximum daily 8-hour average ozone concentration were extracted for each day from the Weather Research and Forecasting with Chemistry (WRF-Chem) 3.2 model. WRF-Chem is a regional chemical transport model that simulates meteorology and behavior of atmospheric gases and aerosols [49, 50]. Appendix A in the supplemental material details the WRF-Chem inputs and options used for our simulations.

2.2. Statistical Analysis

Each observation comprises an outcome, y_i, the log maximum 8-hour average ozone on a given day at a given monitoring location, and a vector of covariates, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})', i = 1, ..., n$, where n is the number of observations, and p is the number of covariates. The vector of outcomes, $\mathbf{y} = (y_1, ..., y_n)'$, and the matrix of covariates, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)'$, together compose the data, $D = \{\mathbf{X}, \mathbf{y}\}$. Ozone observations were log transformed to reduce the impact of heteroscedasticity (non-constant variance across the range of a variable), as data exploration revealed the variance was substantially greater than the mean at high values. The maximum daily 8-hour average ozone from the WRF chemical transport model (WRF-Chem) was also log transformed to have the same scale as the outcome. All other covariates were transformed to have a mean of 0 and variance of 1.

Ten predictive algorithms were trained and evaluated on these data: elastic net regression, generalized additive models (GAM), gradient boosting, k-nearest neighbor regression, lasso regression, linear models, multivariate adaptive regression splines (MARS), neural network, random forest, and support vector machines with a radial basis kernel (SVM). All of the models except for neural networks were fit using models available in version 6.0 of the caret R package [51]. Neural networks were given special consideration on account of their grow-

ing popularity as machine learning prediction tools and especially the recent publication of papers using a neural network with inverse distance weighted convolutional layers to predict ozone and particulate matter [35, 36]. We tested a neural network that mimicked those models, employing inverse distance weighting to create convolutional spatial, temporal and space-time layers using the keras R package [52]. These models were less accurate than a standard feedforward neural network, and so we have reported the results of that network here. Training each prediction model produces a prediction rule $\eta(\mathbf{x}, D_T)$, which is a function of D_T, the data on which it was trained, and a vector of covariates, $\mathbf{x}$, mapping them to a prediction for $y \mid \mathbf{x}$, which is often used as an estimator of $E(y \mid \mathbf{x})$, the conditional expectation of y given $\mathbf{x}$.

The models were tuned, selected, and evaluated using cross-validated estimators of root mean square error (RMSE) and R^2, which are both functions of the mean square error (MSE). The MSE of a prediction rule $\eta(\mathbf{x}, D_T)$, where D_T is the data with which η was trained, may be estimated using a test data set D_W as

$$\hat{MSE}(D_W, \eta(\mathbf{x}, D_T)) = \frac{1}{n_w} \sum_{j \in D_W} (y_j - \eta(\mathbf{x}_j, D_T))^2, \tag{1}$$

where n_w is the number of data points in D_W. If D_W and D_T are disjoint, (i.e., if η was not trained using any part of D_W), then this is an out-of-sample estimator of the MSE. RMSE may be estimated by the square root of $\hat{MSE}$, and R^2 is estimated by

$$\hat{R}^2(D_W, \eta(\mathbf{x}, D_T)) = 1 - \frac{\hat{MSE}(D_W, \eta(\mathbf{x}, D_T))}{n_w^{-1} \sum_{j \in D_W} (y_j - \bar{y}_w)^2}, \tag{2}$$

where $\bar{y}_w = n_w^{-1} \sum_{j \in D_W} y_j$ is the mean outcome in D_W. For ease of notation, the function arguments for $\hat{MSE}$, $\hat{RMSE}$, and $\hat{R}^2$ are hereafter suppressed.

Two different cross-validation (CV) strategies were employed for model evaluation: 10-fold cross-validation and leave-one-location-out (LOLO) cross-validation. For 10-fold CV, the data were randomly partitioned into 10 non-overlapping subsets, each containing one tenth of the data. Each subset served as the test

data for models trained on the other nine tenths of the data, resulting in ten different pairs of training and test sets, with each observation appearing in one test set and the nine training sets not paired with that test set. This yielded 10 estimates of MSE for each model, which were averaged into an overall estimate of MSE, from which the 10-fold CV estimates of RMSE and R^2 were computed.

Ten-fold cross-validation is widely used for estimating prediction error; however, it is known to be overly optimistic for dependent data [41]. Data recorded by air pollution monitors are expected to exhibit spatial or space-time dependence. To more accurately estimate the downscaling error associated with predicting ozone at an unobserved location, RMSE and R^2 were estimated using LOLO CV, in which a model is trained on data from all but one location, and its prediction error is computed for the observations at the withheld location. This process is repeated with observations at each location serving as the withheld test set once, and the resulting errors are averaged into the LOLO CV estimate of prediction error. Unlike 10-fold CV in which observations are distributed among folds uniformly at random, LOLO CV ensures that no observations from the test location may appear in the training data. This provides a realistic estimate of the downscaling prediction error associated with predicting ozone observations at a new location within the same region as the monitoring data.

Most predictive machine learning algorithms depend upon one or more parameters whose values must be set prior to fitting the model. Algorithm performance can vary greatly depending on these parameter values, and it is often desirable to select values that optimize some criteria in an attempt to improve model performance. The process of choosing values for these parameters is often referred to as tuning and the parameters themselves as tuning parameters (or hyperparameters). In our analysis, most tuning parameter values were selected by comparing the performance of candidate values on 25 bootstrap samples of the data using the caret R package [51]. Parameters for k-nearest neighbors and GAM were specifically tuned for LOLO CV in an attempt to stabilize the LOLO CV prediction error, as these models made extremely poor LOLO predictions using bootstrap-selected tuning parameter values. Appendix C in the supple-

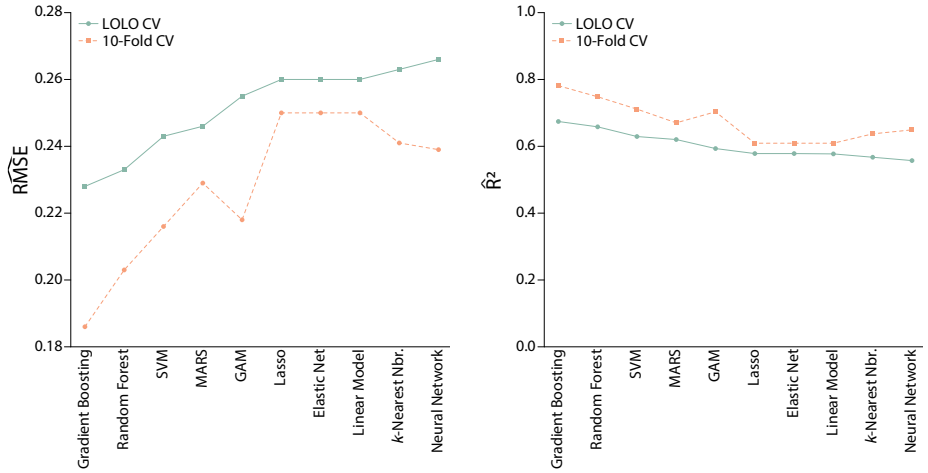
mental material details these tuning procedures and their results. Substantial effort was taken in selecting the number of layers, nodes, activation functions and distance weighting functions for the neural network. The most accurate model was a feedforward neural network with one hidden layer of 21 nodes using a rectified linear unit activation function without the inverse distance weighting functions and convolutional layers employed in previously published models [35, 36].

To investigate the transferability of a model trained on data in one region to another, i.e., its ability to extrapolate rather than downscale, the predictive performance of the two best models trained on the 2008 northern California wildfire period—those two with the lowest LOLO CV estimates of RMSE—was evaluated on data collected during a 2007 wildfire event in southern California. The southern California data consisted of 5,978 daily 8-hour maximum ozone values recorded at 72 monitors between September 1, 2007 and November 28, 2007.

3. Results and Discussion

Figure 2 graphically depicts the cross-validated estimates of RMSE and R^2 for each algorithm using 10-fold CV and LOLO CV. In every case, the 10-fold CV $\hat{RMSE}$ was lower than the LOLO CV $\hat{RMSE}$, and the 10-fold CV $\hat{R}^2$ was higher than the LOLO CV $\hat{R}^2$. Gradient boosting had the lowest 10-fold CV $\hat{RMSE}$ (0.186 log ppm), lowest LOLO CV $\hat{RMSE}$ (0.228 log ppm), highest 10-fold CV $\hat{R}^2$ (0.784), and highest LOLO CV $\hat{R}^2$ (0.677). Random forest placed second in all four categories. The table in Appendix B lists exact values for $\hat{RMSE}$ and $\hat{R}^2$ for each model. These results answer the two primary questions posed by this study, demonstrating that machine learning methods can downscale ozone during a wildfire with reasonable accuracy and identifying gradient boosting and random forest as performing particularly well.

The 10-fold CV estimates of RMSE and R^2 were optimistic compared to those of LOLO CV for all ten models. This over optimism of 10-fold CV is intu-



(a) $R\hat{M}SE$ (b) $\hat{R}^2$

Figure 2: 10-fold and leave-one-location-out cross-validated estimates of RMSE and R^2 for downscaling ozone prediction models.

itive, because of the strong dependence between test and training observations from the same monitor location. The 10-fold CV estimators are also unreliable for model selection. The ordering of model performance is not invariant to the choice of evaluation criterion, which is demonstrated here by the evaluation of the neural network. It is the worst performing model when evaluated by LOLO CV, but is sixth best according to 10-fold CV. The ordering of GAM, MARS and k-nearest neighbors also differ, though the magnitude of those differences is not as substantial.

The difference between the 10-fold and LOLO cross-validated estimates of performance was smaller for relatively inflexible models like lasso, elastic net, and linear regression than for the other models, whose greater flexibility enabled them to better exploit the more highly dependent folds of 10-fold CV. The large difference for k-nearest neighbor regression is due to the strong dependence between observations recorded at the same monitor location. In 10-fold CV, the nearest neighbors of an observation are very likely to be other observations taken at that location. In LOLO CV, no observations taken at the test location

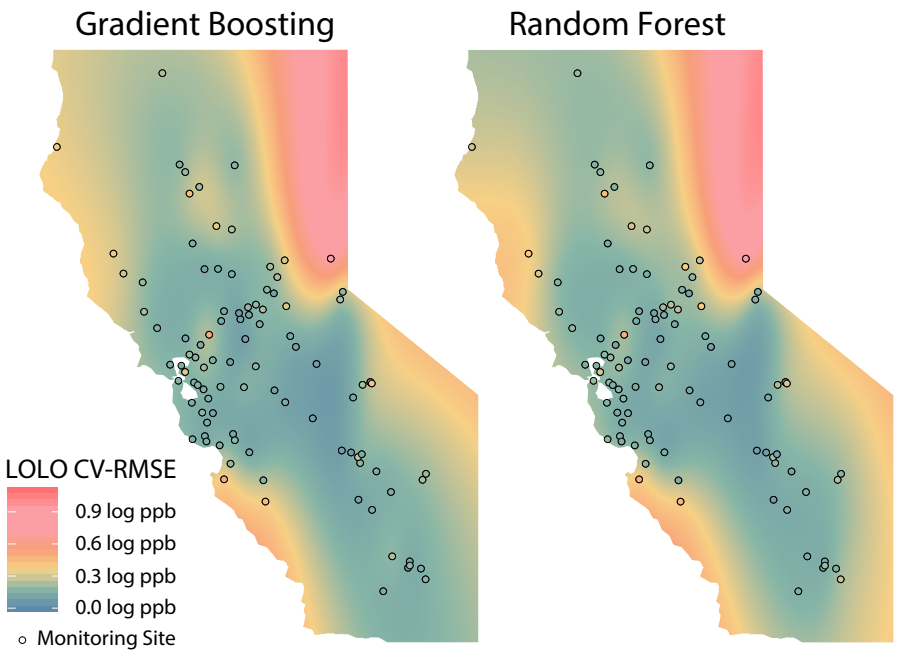


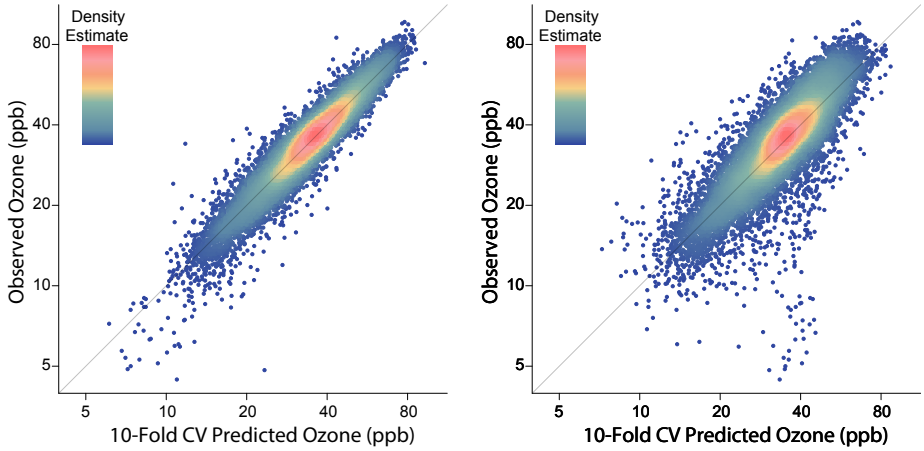
Figure 3: The leave-one-location-out (LOLO) cross-validated estimates of RMSE averaged over the study period (May 6, 2008–September 26, 2008) are plotted at each monitor location for gradient boosting (left) and random forest (right). These average prediction error estimates were then smoothed throughout the study region using a two-dimensional spline-on-sphere smoother to provide a visual estimate of how downscaling predictive performance may vary across the spatial domain.

appear in the training data. It is no surprise that this yields substantially higher estimates of prediction error.

Figure 4 plots 10-fold and LOLO CV predictions against observed daily 8-hour maximum average ozone for gradient boosting and random forest. The CV prediction for each point is the predicted value when it appears in the test fold of the CV procedure. Points that fall on the grey diagonal line are perfectly predicted. The tighter clustering of points around this line in the 10-fold CV plots corresponds to the more accurate predictions made when test set monitor locations are included in the training data.

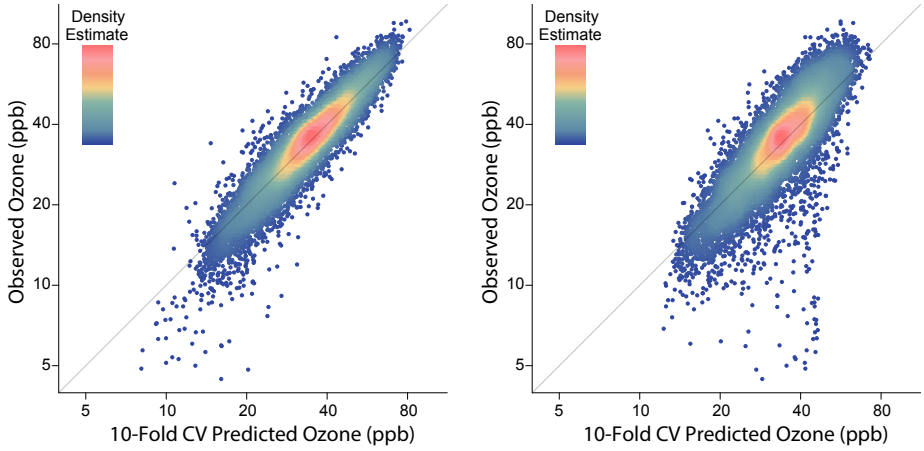
The neural network performed less well than in previous applications for predicting ozone and particulate matter over a spatial grid across the continental United States [35, 36]. The recent popularity of convolutional neural networks is largely due to their performance on image-processing problems. Gridded spatial (or space-time) data bear a much greater resemblance to image data than do the point process monitor data upon which they were evaluated here. It is also possible that alternate specifications of the network architecture could improve performance, but developing such a model goes well beyond the scope of this comparison, which is limited to readily available algorithms that do not require substantial expertise in model specification or implementation. At least in this context neural networks do not succeed as an automatically regularized statistical learning tool (i.e., as a black box), but their performance in other studies suggests they may work well as a highly specialized tool designed using domain knowledge specific to a particular application.

The magnitude of the difference between the LOLO and 10-fold CV estimates of prediction error has meaningful consequences for estimating exposures for subsequent epidemiological analyses. Downscaled exposure is often used as the covariate of interest in analyses seeking to infer the health consequences of air pollution without accounting for prediction uncertainty. The more realistic estimates of prediction error provided by LOLO CV offer better insight into whether it is reasonable to ignore this uncertainty. This may motivate improvements to epidemiological models to account for exposure measurement



(a) Gradient Boosting 10-Fold CV

(b) Gradient Boosting LOLO CV



(c) Random Forest 10-Fold CV

(d) Random Forest LOLO CV

Figure 4: 10-fold and leave-one-location-out (LOLO) cross-validated gradient boosting and random forest predictions plotted against observed daily 8-hour maximum average ozone on the log scale.

error.

The increased accuracy of LOLO CV comes at a computational cost. When the number of locations exceeds 10, LOLO CV is more computationally expensive than 10-fold CV. In this analysis, there are 100 monitor locations and therefore 100 folds in LOLO CV, corresponding to approximately 10 times the computational burden of 10-fold CV. Grouping monitor locations into folds is an appealing strategy to alleviate this burden [43], however, it estimates prediction error over a different spatial resolution than LOLO CV, and will result in overly conservative estimates.

The top two models, gradient boosting and random forest, are both ensembles of tree-based models that provide very flexible mean structures. Their excellence suggests that the mean structure characterizing the relationship between covariates and ozone likely includes interactions, non-linearities and possibly discontinuities. These results do not demonstrate that the underlying chemical processes by which ozone forms are similarly complicated, but that seems likely. The 10-fold CV estimates of RMSE and R^2 were similar to, although slightly lower than, those reported in a similar analysis of machine learning exposure models for $PM_{2.5}$ during the same wildfire time period [22]. Tree-based ensembles were also the best performing models in that study, suggesting that algorithms with flexible mean structures can produce useful exposure models for ozone and $PM_{2.5}$ during wildfire events. Traditional exposure models have focused on modeling the dependence between observations, while employing a simple mean structure. The machine learning models evaluated here assume independent observations, but offer much greater flexibility in modeling the mean. This approach is expected to provide more accurate predictions distant from the observations on which the model was trained than methods that rely upon the dependence between observations. Combining the flexible mean structure of tree-based ensembles with the dependence structures of traditional spatial statistics models is a promising avenue for future work.

An interesting related analysis would be assessing the effect wildfires have on ozone formation. Doing so would require additional numerical simulations

from the WRF-Chem model that exclude wildfire emissions from its inputs. With the data currently available to us, it is impossible to disentangle the effect of wildfires from the other inputs of the WRF-Chem model. This inferential problem is also quite different from the downscaling task which is the focus of this study and may require an entirely different modeling approach. A model that provides excellent downscaling predictions may not be useful for drawing scientific inference.

Another interesting question is whether ozone formation was NOx-limited or VOC-limited. During a wildfire the chemical regime is primarily determined by the amount and conditions of the wildfire fuel, leading to rapid changes in NOx and VOC sensitivity from day to day and even within the course of a day. Understanding this would require a fully separate chemical analysis of air quality conditions in northern California that is beyond both the scope of this paper and the scope of our data, as we lack data on VOC concentrations. One previous study, however, examined the changes that occurred in atmospheric chemistry when wildfire plumes interacted with urban pollution during these fires [53].

We also lack information on chlorofluorocarbon (CFC) emissions, which break down ozone and thus influence ozone concentrations. If data on these compounds were available during the study period, we could include them as covariates in an attempt to improve predictions. The absence of data on VOC and CFC emissions does not invalidate the downscaling enterprise. Downscaling models are constructed to combine the available information, whether from observational processes or complex computational models like WRF-Chem, into accurate predictions within a particular domain. They are not scientific models and do not necessarily imply anything about the chemical or physical mechanisms by which pollutants form and move. Using flexible models strictly for downscaling also protects us from biases in the WRF-Chem output. We need not validate the accuracy of the WRF-Chem simulations; we simply rely on the models to learn the relationship between this output and observed ozone concentrations.

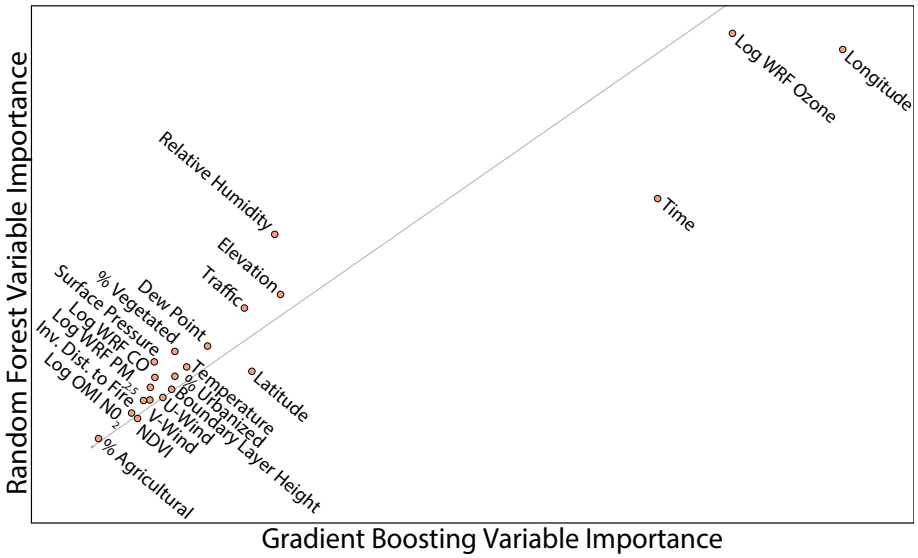


Figure 5: Pairwise normalized gradient boosting and random forest variable importance scores for models trained on the full data. The light grey line denotes equal importance in the two models.

Figure 5 plots covariate importance scores for random forest and gradient boosting models fit to the full data. Random forest variable importance was calculated as the mean decrease in residual sum of squares resulting from splitting on that covariate averaged across all the trees of the forest. Variable importance for gradient boosting was calculated by permuting that covariate's values and computing the average difference in MSE between predictions made with permuted and un-permuted values [54]. The variable importance scores for the two models were normalized to sum to one for ease of comparison. Most covariates are close to the grey diagonal, which indicates equal importance in the two models. Longitude, WRF-Chem ozone and time were the three most important covariates for both models. WRF-Chem is constructed to estimate atmospheric ozone and so it is not surprising that it has a high importance score. Longitude, latitude and time can proxy for unobserved factors, but also calibrate the effect of other covariates similar to space- or time-varying coefficient models. This calibration may be especially important for the numerical outputs of the WRF-

Chem model. Longitude is likely particularly important because it can be used to index many of the significant geographical features of northern California that run approximately North-South, including the coast, San Joaquin Valley, coastal and Sierra Nevada mountain ranges. These geographical features may be associated with important, unobserved information that is not captured by the other covariates including VOC and CFC concentrations.

Neither model when trained on the northern California 2008 wildfire data accurately predicted ozone exposure in southern California in 2007. The predictions from both models had negative $\hat{R}^2$, indicating that their predictions were less accurate (i.e., had higher estimated MSE) than the sample mean of the southern California ozone monitors, which by definition has an $\hat{R}^2$ of 0. In fairness to gradient boosting and random forest, in an out-of-domain prediction problem, the sample mean is unknown, and therefore cannot be used as a prediction rule. When downscaling gradient boosting and random forest models were fit to the 2007 southern California wildfire data, they had LOLO CV errors comparable to the LOLO CV errors reported above. These models accurately downscaled ozone observations during the southern California wildfire (just as they did for the northern California data), but they did not extrapolate well outside of the domain in which they were trained.

This is not surprising and illustrates the proper interpretation of our modeling efforts, which is statistical downscaling (i.e., interpolating) within the observed space-time domain. The substantial decrease in predictive accuracy between within-domain (i.e., downscaling) and out-of-domain (i.e., extrapolating) predictive performance suggests that the relationships between covariates and ozone exposure differ in space and time and demonstrates the dangers of using a downscaling model for extrapolation. Within the observed space-time domain, space and time can proxy for unobserved spatially-indexed covariates in flexible models like gradient boosting and random forest, improving downscaling predictions, but impeding straightforward extrapolation to different space-time domains where the relationship between these unmeasured covariates and space-time may be different. One referee suggested that the difference in ozone pre-

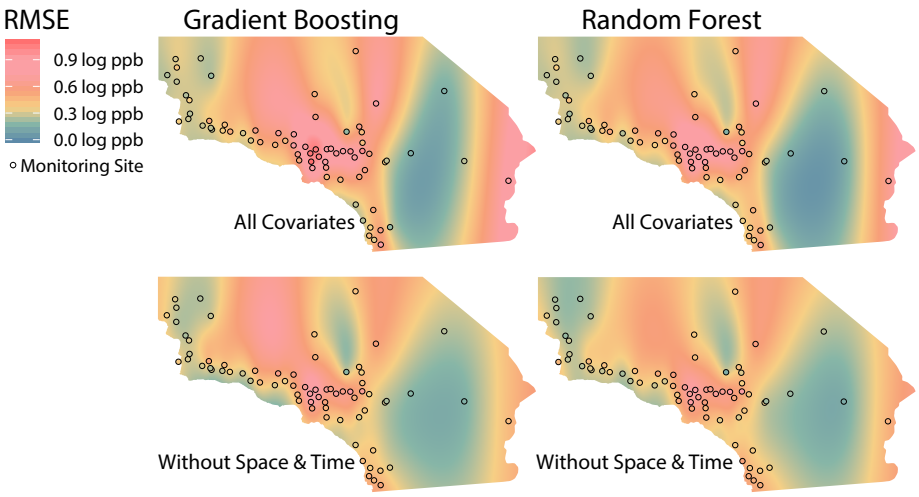


Figure 6: The extrapolation error made by gradient boosting and random forest models trained on the 2008 data and predicting ozone during the 2007 southern California fire. The mean RMSE at each location is smoothed throughout the study region using a two-dimensional spline-on-sphere smoother.

cursors between regions, specifically whether ozone formation is NOx-limited or VOC-limited, is likely one such unobserved, spatially-indexed covariate hindering spatial extrapolation.

As a further check, we repeated this extrapolating procedure excluding latitude, longitude and time from the covariates. The domains of the other covariates were comparable between the two data sets, with the exception of a few low values for surface pressure in the southern California data. Excluding spatial and temporal covariates improved predictive accuracy, reducing LOLO CV $\hat{RMSE}$ from 0.587 to 0.499 for gradient boosting and 0.544 to 0.462 for random forest. As expected, space-time covariates improve downscaling predictions but worsen extrapolating predictions, because the unobserved information indexed by these covariates differs in other regions and at different times.

The scale on which prediction is performed is also important and application specific. In this analysis we performed prediction on the log scale to normalize the variance, but also because it balances large and small errors allowing

accurate predictions to be made at both large and small concentrations. We believe this is a natural scale for the intended subsequent applications of our predictions in epidemiological analyses of pollution health effects. Prediction on the untransformed, original scale would more heavily weight predictions at high ozone concentrations. This may be useful for some applications, but we prefer to balance predictive accuracy at low and high concentrations for our application.

The comparison performed here demonstrates that machine learning prediction algorithms, especially ensembles of tree models like gradient boosting and random forest, can accurately downscale ozone concentrations during wildfire events. We believe they would downscale ozone similarly well in the absence of wildfire events. The models we consider here, however, did not accurately extrapolate beyond the space-time domain on which they were trained. This analysis also demonstrates that the choice of evaluation metric is critical to understanding predictive performance. Metrics that ignore the dependent structure of the data, including k-fold CV, are overly optimistic and unreliable for model selection. LOLO CV is a superior alternative that accounts of the spatial dependence of the data in evaluating model predictive performance, resulting in more reliable estimates of predictive performance.

4. Acknowledgments

This work was funded in part by a grant from the Bureau of Land Management [grant number L14AC00173].

References

[1] World Health Organization, Health effects of particulate matter. Policy implications for countries in eastern Europe, Caucasus and central Asia. Copenhagen: WHO Regional Office for Europe; 2013.

[2] D. B. Menzel, Ozone: an overview of its toxicity in man and animals, J. Toxicol. Environ. Health.

[3] M. L. Bell, A. McDermott, S. L. Zeger, J. M. Samet, F. Dominici, Ozone and short-term mortality in 95 us urban communities, 1987-2000, JAMA 292 (19) (2004) 2,372–2,378.

[4] T. Stocker, D. Qin, G. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. Midgley, Fifth assessment report of the intergovernmental panel on climate change, The Phys. Sci. Basis.

[5] K. R. Smith, M. Jerrett, H. R. Anderson, R. T. Burnett, V. Stone, R. Derwent, R. W. Atkinson, A. Cohen, S. B. Shonkoff, D. Krewski, C. A. Pope, Public health benefits of strategies to reduce greenhouse-gas emissions: health implications of short-lived greenhouse pollutants, The Lancet 374 (9707) (2009) 2091–2103.

[6] R. A. Silva, J. J. West, J.-F. Lamarque, D. T. Shindell, W. J. Collins, G. Faluvegi, G. A. Folberth, L. W. Horowitz, T. Nagashima, V. Naik, S. T. Rumbold, Future global mortality from changes in air pollution attributable to climate change, Nat. Clim. Change 7 (9) (2017) 647.

[7] D. Fowler, M. Amann, F. Anderson, M. Ashmore, P. Cox, M. Depledge, D. Derwent, P. Grennfelt, N. Hewitt, O. Hov, M. Jenkin, Ground-level ozone in the 21st century: future trends, impacts and policy implications, R. Soc. Sci. Policy Rep. 15 (08).

[8] National Research Council, Estimating mortality risk reduction and economic benefits from controlling ozone air pollution, National Academies Press, 2008.

[9] S. Sousa, M. Alvim-Ferraz, F. Martins, Health effects of ozone focusing on childhood asthma: what is now known–a review from an epidemiological point of view, Chemosphere 90 (7) (2013) 2051–2058.

[10] M. Guarnieri, J. R. Balmes, Outdoor air pollution and asthma, The Lancet 383 (9928) (2014) 1581–1592.

[11] J. Y. Lee, S.-B. Lee, G.-N. Bae, A review of the association between air pollutant exposure and allergic diseases in children, Atmos. Pollut. Res. 5 (4) (2014) 616–629.

[12] M. C. Turner, M. Jerrett, C. A. Pope III, D. Krewski, S. M. Gapstur, W. R. Diver, B. S. Beckerman, J. D. Marshall, J. Su, D. L. Crouse, R. T. Burnett, Long-term ozone exposure and mortality in a large prospective study, Am. J. Respir. Crit. Care Med. 193 (10) (2016) 1,134–1,142.

[13] D. L. Crouse, P. A. Peters, P. Hystad, J. R. Brook, A. van Donkelaar, R. V. Martin, P. J. Villeneuve, M. Jerrett, M. S. Goldberg, C. A. Pope III, M. Brauer, R. D. Brook, A. Robichaud, R. Menard, R. T. Burnett, Ambient $pm_{2.5}$, o_3, and no_2 exposures and associations with mortality over 16 years of follow-up in the canadian census health and environment cohort (canchec), Environ. Health Perspect. 123 (11) (2015) 1180.

[14] B.-F. Hwang, Y.-H. Chen, Y.-T. Lin, X.-T. Wu, Y. L. Lee, Relationship between exposure to fine particulates and ozone and reduced lung function in children, Environ. Res. 137 (2015) 382–390.

[15] R. Barr, E. Hoffman, J. Madrigano, C. Aaron, P. Sampson, L. Sheppard, S. Vedal, J. Kaufman, M. Wang, Long-term exposure to ozone and accelerated progression of percent emphysema and decline in lung function: The mesa air and lung studies, Medicine 1 (2) (2016) 3.

[16] M. Val Martin, R. Honrath, R. C. Owen, G. Pfister, P. Fialho, F. Barata, Significant enhancements of nitrogen oxides, black carbon, and ozone in the north atlantic lower free troposphere resulting from north american boreal wildfires, J. Geophys. Res. D: Atmos. 111 (D23).

[17] D. A. Jaffe, N. L. Wigder, Ozone production from wildfires: A critical review, Atmos. Environ. 51 (2012) 1–10.

[18] J. Wu, A. M. Winer, R. J. Delfino, Exposure assessment of particulate

matter air pollution before, during, and after the 2003 southern california wildfires, Atmos. Environ. 40 (18) (2006) 3333–3348.

[19] U. Confalonieri, B. Menne, R. Akhtar, K. L. Ebi, M. Hauengue, R. S. Kovats, B. Revich, A. Woodward, Human health (2007), in: M. Parry, O. Canziani, J. Palutikof (Eds.), Climate change 2007: impacts, adaptation and vulnerability, Vol. 4, Cambridge University Press Cambridge, Cambridge, 2007.

[20] J. M. Azevedo, F. L. Gonçalves, M. de Fátima Andrade, Long-range ozone transport and its impact on respiratory and cardiovascular health in the north of portugal, Int. J. Biometeorol. 55 (2) (2011) 187–202.

[21] R. Tham, B. Erbas, M. Akram, M. Dennekamp, M. J. Abramson, The impact of smoke on respiratory hospital outcomes during the 2002–2003 bushfire season, victoria, australia, Respirology 14 (1) (2009) 69–75.

[22] C. E. Reid, M. Jerrett, M. L. Petersen, G. G. Pfister, P. E. Morefield, I. B. Tager, S. M. Raffuse, J. R. Balmes, Spatiotemporal prediction of fine particulate matter during the 2008 northern california wildfires using machine learning, Environ. Sci. Technol. 49 (6) (2015) 3887–3896.

[23] M. L. Stein, Interpolation of spatial data: some theory for kriging, Springer Science & Business Media, 2012.

[24] D. J. Briggs, S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. Van Reeuwijk, K. Smallbone, A. Van Der Veen, Mapping urban air pollution using gis: a regression-based approach, Int. J. Geog. Inf. Sci. 11 (7) (1997) 699–718.

[25] A. E. Gelfand, H.-J. Kim, C. Sirmans, S. Banerjee, Spatial modeling with spatially varying coefficient processes, J. Am. Stat. Assoc. 98 (462) (2003) 387–396.

[26] C. Robert, Machine learning, a probabilistic perspective, Taylor & Francis, 2014.

[27] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 161–168.

[28] A. J. Ferreira, M. A. Figueiredo, Boosting algorithms: A review of methods, theory, and applications, in: Ensemble machine learning, Springer, 2012, pp. 35–85.

[29] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[30] Y. Liu, C. J. Paciorek, P. Koutrakis, Estimating regional spatial and temporal variability of $pm_{2.5}$ concentrations using satellite data, meteorology, and land use information, Environ. Health Perspect. 117 (6) (2009) 886.

[31] W.-Z. Lu, W.-J. Wang, Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends, Chemosphere 59 (5) (2005) 693–701.

[32] H. Weizhen, L. Zhengqiang, Z. Yuhuan, X. Hua, Z. Ying, L. Kaitao, L. Donghui, W. Peng, M. Yan, Using support vector regression to predict pm_{10} and $pm_{2.5}$, in: IOP Conference Series: Earth and Environmental Science, Vol. 17, IOP Publishing, 2014, p. 012268.

[33] Y. Zhan, Y. Luo, X. Deng, H. Chen, M. L. Grieneisen, X. Shen, L. Zhu, M. Zhang, Spatiotemporal prediction of continuous daily $pm_{2.5}$ concentrations across china using a spatially explicit machine learning algorithm, Atmos. Environ. 155 (2017) 129–139.

[34] B. S. Beckerman, M. Jerrett, M. Serre, R. V. Martin, S.-J. Lee, A. Van Donkelaar, Z. Ross, J. Su, R. T. Burnett, A hybrid approach to estimating national scale spatiotemporal variability of $pm_{2.5}$ in the contiguous united states, Environ. Sci. Technol. 47 (13) (2013) 7,233–7,241.

[35] Q. Di, I. Kloog, P. Koutrakis, A. Lyapustin, Y. Wang, J. Schwartz, Assessing $pm_{2.5}$ exposures with high spatiotemporal resolution across the continental united states, Environ. Sci. Technol. 50 (9) (2016) 4,712–4,721.

[36] Q. Di, S. Rowland, P. Koutrakis, J. Schwartz, A hybrid model for spatially and temporally resolved ozone exposures in the continental united states, J. Air Waste Manage. Assoc. 67 (1) (2017) 39–52.

[37] X. Hu, J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, Y. Liu, Estimating $pm_{2.5}$ concentrations in the conterminous united states using the random forest approach, Environ. Sci. Technol. 51 (12) (2017) 6,936–6,944.

[38] Y. Zhan, Y. Luo, X. Deng, M. L. Grieneisen, M. Zhang, B. Di, Spatiotemporal prediction of daily ambient ozone levels across china using random forest for human exposure assessment, Environmental Pollution 233 (2018) 464–473.

[39] C. Brokamp, R. Jandarov, M. Hossain, P. Ryan, Predicting daily urban fine particulate matter concentrations using a random forest model, Environmental science & technology 52 (7) (2018) 4173–4179.

[40] Y. Xu, H. C. Ho, M. S. Wong, C. Deng, Y. Shi, T.-C. Chan, A. Knudby, Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level pm2. 5, Environmental Pollution 242 (2018) 1417–1426.

[41] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (8) (2017) 913–929.

[42] M. Lee, I. Kloog, A. Chudnovsky, A. Lyapustin, Y. Wang, S. Melly, B. Coull, P. Koutrakis, J. Schwartz, Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the southeastern us 2003–2011, J. Exposure Sci. Environ. Epidemiol. 26 (4) (2016) 377.

[43] Y. Zhan, Y. Luo, X. Deng, K. Zhang, M. Zhang, M. L. Grieneisen, B. Di, Satellite-based estimates of daily no2 exposure in china using hybrid random forest and spatiotemporal kriging model, Environmental science & technology 52 (7) (2018) 4180–4189.

[44] A. C. Just, R. O. Wright, J. Schwartz, B. A. Coull, A. A. Baccarelli, M. M. Tellez-Rojo, E. Moody, Y. Wang, A. Lyapustin, I. Kloog, Using high-resolution satellite aerosol optical depth to estimate daily pm2. 5 geographical distribution in mexico city, Environmental science & technology 49 (14) (2015) 8576–8584.

[45] U.S. Department of Commerce National Oceanic and Atmospheric Administration, Rapid update cycle (2018).
URL `https://ruc.noaa.gov/`

[46] C. Wiedinmyer, S. Akagi, R. J. Yokelson, L. Emmons, J. Al-Saadi, J. Orlando, A. Soja, The fire inventory from ncar (finn): a high resolution global model to estimate the emissions from open burning, Geosci. Model Dev. 4 (3) (2011) 625.

[47] J. Fry, G. Z. Xian, S. Jin, J. Dewitz, C. G. Homer, L. Yang, C. A. Barnes, N. D. Herold, J. D. Wickham, Completion of the 2006 national land cover database for the conterminous united states, Photogramm. Eng. Remote Sens. 77 (9) (2011) 858–864.

[48] P. F. Levelt, G. H. van den Oord, M. R. Dobber, A. Malkki, H. Visser, J. de Vries, P. Stammes, J. O. Lundell, H. Saari, The ozone monitoring instrument, IEEE Trans. Geosci. Remote Sens. 44 (5) (2006) 1093–1101.

[49] J. G. Powers, J. B. Klemp, W. C. Skamarock, C. A. Davis, J. Dudhia, D. O. Gill, J. L. Coen, D. J. Gochis, R. Ahmadov, S. E. Peckham, G. A. Grell, The weather research and forecasting model: Overview, system efforts, and future directions, Bull. Am. Meteorol. Soc. 98 (8) (2017) 1717–1737.

[50] G. Pfister, D. Parrish, H. Worden, L. Emmons, D. Edwards, C. Wiedinmyer, G. Diskin, G. Huey, S. Oltmans, V. Thouret, A. Weinheimer, Characterizing summertime chemical boundary conditions for airmasses entering the us west coast, Atmos. Chem. Phys. 11 (4) (2011) 1769–1790.

[51] M. Kuhn, C. 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, T. Hunt, caret: Classification and Regression Training, r package version 6.0-80 (2018).
URL `https://CRAN.R-project.org/package=caret`

[52] J. Allaire, F. Chollet, keras: R Interface to 'Keras', r package version 2.2.0.9001.
URL `https://keras.rstudio.com`

[53] H. Singh, C. Cai, A. Kaduwela, A. Weinheimer, A. Wisthaler, Interactions of fire emissions and urban pollution over california: Ozone formation and air quality simulations, Atmospheric environment 56 (2012) 45–51.

[54] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.