Predicting peak daily maximum 8-hour ozone, and linkages to emissions and meteorology, in Southern California using machine learning methods

Ziqi Gao[1], Yifeng Wang[1], Petros Vasilakos[1], Cesunica E. Ivey[2,4], Khanh Do[2,3] Armistead G. Russell[1]

[1] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA

[2] Department of Chemical and Environmental Engineering, University of California, Riverside, Riverside, CA, USA

[3] Center for Environmental Research and Technology, Riverside, CA, USA

[4] Now at Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA

Correspondence to: *Ziqi Gao (zgao71@gatech.edu)*

Abstract. The growing abundance of data is conducive to using numerical methods to relate air quality, meteorology, and emissions to address which factors impact pollutant concentrations. Often, it is the extreme values that are of interest for health and regulatory purposes (e.g., the National Ambient Air Quality Standard for ozone uses the annual, maximum, daily 4th highest, 8-hour average (MDA8) ozone), though such values are the most challenging to predict using empirical models. We developed four different computational models, including the Generalized Additive Model (GAM), the Multivariate Adaptive Regression Splines, the Random Forest, and the Support Vector Regression, to develop observation-based relationships between the 4th highest MDA8 ozone in the South Coast Air Basin and precursor emissions, meteorological factors, and large-scale climate patterns. All models had similar predictive performance, though the GAM showed a relatively higher R^2 value (0.96) with a lower root mean square error and mean bias.

1 Introduction

Tropospheric ozone has proven to be one of the most difficult air pollutants to control, especially in the South Coast Air Basin (SoCAB) of California, which includes the city of Los Angeles and parts of four counties with a 2020 population exceeding 18 million. Exposure to ozone can be harmful to human health, leading to a variety of adverse outcomes, including premature mortality (U.S.EPA, 2020), climate warming and decreased agricultural production (Ainsworth et al., 2012; Hong et al., 2020). Ozone is formed by chemical reactions between volatile organic compounds (VOCs) and nitrogen oxides (NOx) in the presence of sunlight (Seinfeld et al., 2016). In addition to VOC and NOx emissions, meteorology and large-scale climate patterns affect ozone (Aw et al., 2003; Blanchard et al., 2014; Gorai et al., 2015; Kelley et al., 2020; Kleeman, 2008; Lu et al., 2019; Mahmud et al., 2010; McGlynn et al., 2018). As such, the resulting relationships among ozone, emissions, and meteorology are complex and difficult to model accurately. However, the rise of machine learning methods, along with an increasingly long observational record, suggest that observation-based models can be used to understand those relationships.

35   The US Environmental Protection Agency (EPA)'s National Ambient Air Quality Standard (NAAQS) for ozone is based on the annual, maximum, daily 4th highest, 8-hour average (MDA8) ozone observations, which is an extreme statistic, and extreme statistics are often difficult to accurately predict using empirical modeling, though different approaches have been used for various purposes. For example, U.S. EPA adjusted the MDA8 ozone predictions with meteorological observations using Generalized Linear Modeling (GLM) with natural spline smoothing functions in

40   the R program to develop a Generalized Additive Model (GAM) (Camalier et al., 2007; Wells et al., 2021) that meteorologically adjusts ozone trends to help isolate the impact of emissions. The GAM is an extension of the GLM, which was introduced in 1986 (Hastie et al., 1986, 1990). It is more flexible than the GLM due to the smoothing functions on independent variables. Previous studies suggested the GAM was useful to deal with the nonlinear relationship between MDA8 ozone concentrations and meteorological indicators. About 40% to 90% of the variance

45   of the MDA8 ozone concentrations could be explained at different sites with meteorologically-adjusted GAMs (Aldrin et al., 2005; Blanchard et al., 2014; Blanchard et al., 2019; Camalier et al., 2007; Flynn et al., 2021; Gong et al., 2018; Gong et al., 2017; Hu et al., 2021; Huang et al., 2020; Jeong et al., 2020; Ma et al., 2020; McClure et al., 2018; Pearce et al., 2011). GAMs can assess each independent variable's contribution to the dependent variable. The Multivariate Adaptive Regression Splines (MARS) model (Friedman, 1991) has been used to model the nonlinear relationship

50   between ozone concentrations and precursors' concentrations/meteorological factors, including the interactions between the independent indicators (García Nieto et al., 2014; Roy et al., 2018). Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) (Drucker et al., 1996; Rodríguez-Pérez et al., 2017; Smola et al., 2004). Past studies have shown that the SVR model with kernel functions can fit the nonlinear relationships between ozone concentrations and meteorological factors and can obtain accurate predictions (Liu et al., 2017; Luna et al.,

55   2014; Rybarczyk et al., 2018; Sotomayor-Olmedo et al., 2013; Vong et al., 2012). Random Forest (RF) is a machine learning method (Tin Kam, 1995) derived from the traditional decision tree method. Compared to the traditional method, it is more accurate because it contains multiple decision trees. The RF model can be used to fit nonlinear relationships and deal with interaction effects. It can accurately predict ozone concentrations using meteorological variables and emissions and capture about 70% to 95% of the variability in ozone concentrations (Keller et al., 2019;

60   Pernak et al., 2019; Stafoggia et al., 2020; Zhan et al., 2018). However, most prior empirical model applications to simulate peak MDA8 ozone levels were biased low, especially when considering capturing the 4th highest, annual MDA8 ozone concentrations.

In this study, we develop observation-based models (SoCAB-8HR V1.0) using four different methods (GAM, RF, SVR, and MARS) with a broad range of potential independent indicators that impact ozone formation (e.g., precursors

65   emissions, meteorological conditions, large-scale climate events, chemical reactions, seasonal variations, and weekend effects) to predict the annual 4th highest MDA8 ozone in the SoCAB from 1990 to 2019. We assess and compare model performance and their applicability to help understand how emissions and meteorology, independently and combined, impact high ozone levels.

70

2 Methods and Data

2.1 Methods

Brief descriptions of the four methods (GAM, MARS, RF, and SVR) are provided below and they are described in greater detail in the referenced material.

2.1.1 Generalized Additive Model (GAM)

A GAM uses flexible, nonlinear relationships defined between "knots" in the explanatory variables using smoothing functions (Hastie et al., 1986, 1990). The "knot" is the point of the link of two polynomial curves (Wood, 2017). Since the GAM is an additive model, which means each indicator's function adds together to form the model equation, the indicators can have a variety of relationships with the response variable. The general form of the GAM is written as (Hastie et al., 1986, 1990; Wood, 2011, 2017):

$$y = a + \sum_{i=1}^{n} f(x_i) + e$$

where a is the intercept, e is the error term, x refers to each independent indicator, and f means the function applied to the predictors.

There are multiple choices of the functions based on the relationship between each independent and dependent variable, such as splines, linear functions, polynomials, etc. Splines (often cubic) are commonly applied to capture nonlinear relationships. Cubic splines can provide a comparatively more flexible curve than low-order splines. In addition, a cubic spline can avoid overfitting with a smaller curviness and be more effective with less computational time than high-order splines. The basis function of cubic spline is a third-order polynomial equation:

$$y_i = a_i * x^3 + b_i * x^2 + c_i * x + d_i$$

where a, b, c, d are the estimated coefficients of each basis function, the subscript i indicates the number of basis functions (equals to the number of knots). Based on the number of knots, several basis functions are built with different estimated coefficients. Each spline is given by the weighted sum of the basis functions. Three to five knots typically are sufficient in practice, and the knots are evenly distributed based on the percentiles of each indicator (Harrell, 2015). The "mgcv" package in R program was used to build the GAM between the peak MDA8 ozone concentrations and indicators (Hastie, 1991; Hastie et al., 1986, 1990; Wood, 2011, 2017).

2.1.2 Multivariate Adaptive Regression Splines (MARS)

The MARS model is a nonparametric, multivariate, piecewise regression model that can be used to develop the nonlinear relationships between the dependent variable and a set of indicators (Friedman, 1991). Similar to the GAM, linear splines (referred to as "hinge functions") are applied to independent variables in the MARS model. The resulting model is formed by a weighted sum of basis functions. The MARS model can deal with nonlinear relationships and provide a more flexible curve than simple linear regression models and polynomial regression models due to the linear splines between each pair of knots. It is simpler, and the resulting associations between the dependent and indicator variables are easier to interpret than the complex machine learning methods (e.g., random forest and neural network).

The general equation of the MARS model is (Friedman, 1991; Leathwick et al., 2006; MARS; Oduro et al., 2015; Roy
et al., 2018):

$$y = \beta_0 + \sum_{i=2} \beta_i H_i$$

where β_0 is the intercept, H_i shows hinge functions, and β_i is the coefficients of hinge functions. The hinge functions
in the MARS model are pairwise, and the form is

$$(x - k)_+ = \max(x - k, 0)$$

$$(k - x)_+ = \max(k - x, 0)$$

where k is the knot. When applying the MARS model, a two-stage approach is used that includes forward and
backward stages. The forward stage is similar to the forward stepwise regression. At first, the model only includes the
intercept term. Then, the generated pairwise hinge functions are added into the model continuously if they can reduce
the residual error of the model. This process will be terminated when the change of error is small (e.g., less than a
threshold) or the model reaches the defined maximum number of terms. A backward stage is applied to avoid
overfitting and reduce the number of terms, removing terms that do not significantly impact the error (Wikipedia
Contributors, 2022). Generalized cross validation (GCV) is used to find the final MARS model after obtaining multiple
models that have different terms (Friedman, 1991; Friedman et al., 1989; Hastie et al., 1996; Leathwick et al., 2006;
Oduro et al., 2015; Roy et al., 2018). The "earth" package in R was applied to build the relationship between the top
MDA8 ozone concentrations and independent indicators using the MARS model (Friedman et al., 1989; Hastie et al.,
2009; Milborrow, 2021), and this package chose the independent variables, the position of the knots and the interaction
of the terms automatically.

2.1.3 Random Forest Model (RF)

Random forest is a supervised machine learning method that can be used for regression and classification. It is an
ensemble of multiple decision trees. The RF model resolves the limitation of the decision tree that the model can be
overfitting if the depth of the trees is deeper by applying the bagging algorithm. The bagging algorithm effectively
reduces the variance of the model results and makes the RF model quite stable and robust. In regression, the predicted
result of the RF model is the average of the results of all decision trees. The total error of RF is computed by the
average of the error of all the decision trees.

Suppose we build a random forest model which contains m trees (i.e., T_b, $b = 1,2,...,m$) and has a testing point x.
The predicted value of input x would be:

$$\frac{1}{m} \sum_{i=1}^{m} T_i(x)$$

The following steps construct each decision tree in a random forest model. First, randomly select a subset of the
training data set with replacement. Then, at each decision node, randomly select a subset of variables. In order to find
the optimal variable and its corresponding value that can lead to the best fit, we usually define a target function and
compare all variables in the subset to find the variable with the lowest or highest value. Once we find the optimal
variable and corresponding value, we next divide the decision node based on the optimal variable and value. Repeat
the previous step until all decision nodes reach the minimal node size. Finally, for each leaf node, suppose k data

points, $x_1, x_2, \ldots, x_k$, belong to a leaf node, and the corresponding response variables are $y_1, y_2, \ldots, y_k$ respectively. The predicted result of a testing point x which falls into this leaf node should be:

$$\frac{1}{k}\sum_{i=1}^{k} y_i$$

150   We used the "randomForest" package in R software to build the RF model (Liaw et al., 2002). RF models can select interaction terms between the independent variables automatically.

2.1.4   Support Vector Regression (SVR)

The Support Vector Machine (SVM) method is a supervised machine learning approach that is used for classification.
155   The SVR model, which is an extension of the SVM, can be used to describe the nonlinear relationships between the response variable and independent indicators.

Suppose we have a set of indicators $X = \{x_1, x_2, \ldots, x_m\}$ and a set of response variables $Y = \{y_1, y_2, \ldots, y_m\}$. We need to find a hyperplane to minimize error and achieve the best fit, which can be written as $w^T x + b$. We can define the loss function as:

160    $Loss(x_i, y_i) = \max(0, |w^T x_i + b - y_i| - \varepsilon)$

where ε (epsilon) is the margin of error, a user defined variable that can be manipulated to adjust the accuracy of the model. Then, the problem can be written as:

$$min_{w,b} \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} Loss(x_i, y_i)$$

where C is the cost, another user defined variable that determines the tolerance of the model to points outside the
165   bounds set by ε, and m is the number of dataset. To let the loss function result of each training point be 0, we introduced the slack variables. Then, the development of nonlinear relationship between the response variable and indicators can be converted a Lagrangian dual problem (Scholkopf et al., 2001; Smola et al., 2004).

We need to consider the interactions among features sometimes when we build computational models, so we need to map the data into a non-linear feature space. The non-linear feature space increases the dimension of the data space,
170   and consequently the computational complexity grows dramatically. We introduced a kernel function to account for the interactions and reduce the computational complexity. We used the package e1071 in R software to build the SVR model (Chang et al., 2011; Fan et al., 2005).

2.2 Model Evaluation

175   We used the coefficient of determination (R^2), mean bias (MB), and root mean squared error (RMSE) of the observed and predicted peak MDA8 ozone concentrations from 1990 to 2019 to compare the performance of these four models.

$$Mean\ Bias = \frac{\sum_{i=1}^{n}(\hat{x_i} - x_i)}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x_i})^2}{n}}$$

where x_i and $\hat{x}_i$ are the observed and predicted MDA8 ozone concentrations and n is the total number of measurements. In addition, we used 10-fold cross validation (CV) to evaluate the prediction accuracy and stability of these four models. In the 10-fold CV, the dataset is randomly divided into two subsets, in which 90% is used to train the model and 10% is the testing dataset. These two subsets are not overlapped, and this separation process repeats ten times. The averages of the R^2, MB, and RMSE in these ten runs are the final evaluation results of numerical models.

2.3 Study Domain

The SoCAB includes urban and suburban parts of Los Angeles County, Riverside County, San Bernardino County, and all of Orange County. This area historically and still experiences some of the worst air quality in the US, and various air pollutants at multiple sites in the SoCAB do not meet the NAAQS, even with strict regulations leading to significant reductions in pollutant emissions. The poor air quality is because SoCAB is one of the most urbanized and populated regions in the US and is surrounded by mountains on three sides, while the Pacific Oceans lies on the west side. Temperature inversions are formed frequently along the coast due to the warm subsiding air from North Pacific Highs, suppressing vertical mixing. This unique geographical and meteorological environment leads to reduced dilution of air pollutants. In addition, most days in a year are sunny, leading to warm and dry conditions with high solar radiation, exacerbating the formation of photochemically-derived pollutants, such as ozone.

We first focused on the Crestline site to develop the initial regression models to predict the 4th highest MDA8 ozone concentrations in the SoCAB. This site had the annual 4th highest MDA8 ozone concentrations during about 77% of this project's period. The other 23% of the time, the maximum site was close to the Crestline site, such as Glendora, Redlands, and Fontana.

2.4 Data

The daily MDA8 ozone concentrations from 1990 to 2019 in the South Coast Air Basin was retrieved from California Air Resources Board (CARB) archives and EPA AQS pre-generated data files (EPA, 2016). The total number of days of daily MDA8 ozone levels is 10957. We used the top 30 MDA8 ozone days each year to develop the models for the 4th highest MDA8 ozone concentrations to build robust computational models since multiple factors have impacts on the peak MDA8 ozone concentrations. Significant factors may be missed if only the 4th highest MDA8 ozone concentrations are considered, such as day of year, day of the week, meteorological variable impacts as there would only be 30 observations for model training. Furthermore, the size of the 30 years' 4th highest MDA8 ozone dataset is too small to have sufficient statistical power. A small dataset may cause a type II error (failing to identify a statistically significant effect) for some significant features, which would then affect the accuracy of the predictions.

We selected 25 independent indicators, including precursors' emissions, meteorological factors suggested in previous studies (Blanchard et al., 2014; Blanchard et al., 2019; Camalier et al., 2007), Niño 3.4 monthly indices, day of the week and the day of year. A detailed description of all the variables applied to test the final computational models is in Table S1.

215   Estimated NOx and VOC emissions in the SoCAB from 2000 to 2019 were acquired from CARB archives using the
       emissions in 2012 (Cox, 2013). The emissions between 1990 and 2000 were projected with the emissions in 2008 and
       2012 (Cox, 2009, 2013). The detailed calculation is in supplement information (SI).

       We included two kinds of meteorological data: surface meteorological data and upper air meteorological data. We
       obtained the surface meteorological data, including temperature, wind speed, and wind direction at Los Angeles

220   International Airport and Barstow-Daggett Airport from National Oceanic and Atmospheric Administration (NOAA)
       archives and CARB archives (Menne et al., 2012a; Menne et al., 2012b). The upper air meteorological data at the
       Miramar site was provided by NOAA and contains geopotential height, temperature, dew point temperature, wind
       speed, and wind direction at 500 and 850 millibars (mb). Using temperature and dew point temperature, we computed
       the RH at 500 and 850 mb with the Clausius-Clapeyron equation (Alduchov et al., 1996; Lawrence, 2005).The height

225   of 500 mb is around 5500 m (NOAA, 2020), and the 850 mb height is about 1500 m, which is close to the boundary
       layer height. The upper air meteorology is related to the synoptic-scale weather and has an impact on the surface
       meteorology (Blanchard et al., 2014; Camalier et al., 2007).

       Past studies have shown there is a relationship between the El Niño–Southern Oscillation (ENSO) events and the
       variability of MDA8 ozone concentrations by affecting the local meteorology (Lu et al., 2019; Oman et al., 2013;

230   Oman et al., 2011; Xu et al., 2017). Niño 3.4 monthly indices were obtained from Climate Prediction Center (CRC)
       to represent ENSO events. To account for the daily variations and weekend effects of MDA8 ozone levels, we included
       day-of-the week and the day of year in the models (Seinfeld et al., 2016).

3   Results
235

3.1 Model Application and Performance

3.1.1 GAM model
       We combined stepwise regression and F values to assess the statistical significance of each independent indicator to
240   refine the model equation to provide the smallest Akaike Information Criterion (AIC) value after excluding the highly
       correlated indicators (Fig. S1) (Pope et al., 1972). However, the stepwise regression may exclude some factors that
       are known to be tied to ozone formation from the final equation, including VOC emissions, the day of year and day
       of the week. We used both statistical indicators and knowledge of important relationships in the final model to avoid
       losing significant factors that affect the peak ozone levels. Furthermore, a limitation of the GAM is that it does not
245   identify interaction terms, so interaction terms were introduced with the spline function manually in the style of s(x1,
       x2).

       We applied cubic splines to the emissions and meteorological variables due to the nonlinear relationship between peak
       MDA8 ozone concentrations and meteorology/emissions. Also, the cubic spline was used for the day of the year to
       add the daily and seasonal variation of the precursors' emissions. In addition, we included the day of the week in
250   factor style to represent the weekend effect and Niño 3.4 monthly indices to show the large-scale climate pattern
       impacts on ozone formation with linear functions. We used the annual top 30 MDA8 ozone concentrations from 1990
       to 2019 on a log scale as the dependent variable at the Crestline site because the ozone concentrations follow a log-

normal distribution (Blanchard et al., 2014; Camalier et al., 2007; EPA, 2020; Henneman et al., 2015; Hogrefe et al., 2000; Rao et al., 1997).

255   The final GAM (GAM-SoCAB-8HR V1.0) included emissions, meteorological factors, large-scale climate index and temporal variables at the Crestline site from 1990 to 2019 (Eq. (1)). The detailed description of each variable is in Table 1 (e: error term):

$\log(MDA8) = a + dayofweek(factor) + dayofyear + s(TMAXBarstow) + s(Mir850RH) + s(AWND) +$

$s(eNOx) + s(eROG) + s(eNOx, TMAXBarstow) + s(eROG, TMAXBarstow) + s(eNOx, eROG) + ENSOmonthly +$

260   e (1)

The correlation (R^2) between independent variables in the final GAM was tested (Fig. 1). The correlation between VOC and NOx emissions was high that close to 1. However, both are the precursors of MDA8 ozone, so these factors were not removed during the model development. Other than emissions, the correlation among all the significant independent variables in the final models is negligible (Fig. 1).

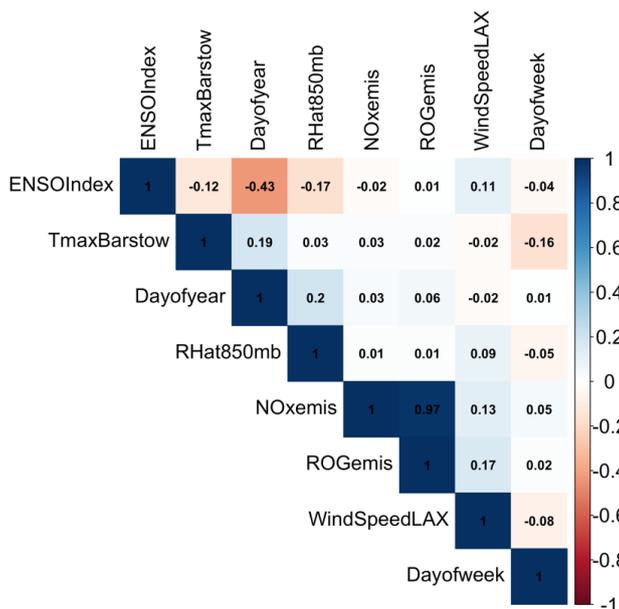
265   84% of the variability of the peak MDA8 ozone concentrations can be explained using this GAM (Fig. 2A). 10-fold validation results show that the R^2 value was 0.85 using the testing dataset, only 0.01 higher than the training dataset. Also, the RMSE of the testing data was only slightly different from that of the training data (Table S4), which indicated that this GAM could predict peak MDA8 ozone concentrations stably. This model had an R^2 value equal to 0.96, and RMSE is 11.1 ppbV for the 4th highest MDA8 ozone predictions from 1990 to 2019 (Fig. 3A).

270

Figure 1: Correlation value between the independent variables (only valid for GAM).

3.1.2 MARS model

275 We used the same dataset as the GAM (GAM-SoCAB-8HR V1.0) to be comparable with the GAM's results. The final model contained six indicators, including the NOx and VOC emissions, the maximum temperature at the Barstow-Daggett Airport, the average wind speed at LAX, Niño 3.4 monthly indices, day of the year, and eleven interaction terms between indicators. Similar to GAM, we applied log function to the ozone concentrations (EPA, 2020; Henneman et al., 2015; Hogrefe et al., 2000; Rao et al., 1997). The final equation of the MARS model (MARS-

280 SoCAB-8HR V1.0) is shown below (Eq. (2)). A detailed description of each variable is in Table 1:

$\log (\text{MDA8}) = 4.59 + (1135.1 - \text{eNOx}) * (-9.57\text{e-}04) + (\text{eNOx} - 1135.1) * (1.55\text{e-}03) + (643.4 - \text{eROG}) * (2.67\text{e-}03) + (\text{TMAXBarstow} - 36.7) * 0.022 + (1135.1 - \text{eNOx}) * (\text{eROG} - 450.7) * (4.29\text{e-}06) + (\text{eNOx} - 1415.8) * (\text{eROG} - 643.4) * (-1.08\text{e-}06) + (\text{eNOx} - 1225) * (\text{TMAXBarstow} - 36.7) * (7.21\text{e-}04) + (1081.5 - \text{eROG}) * (\text{TMAXBarstow} - 36.7) * (-2.18\text{e-}05) + (\text{eROG} - 1081.5) * (\text{TMAXBarstow} - 36.7) * (-3.95\text{e-}04) +$

285 $(\text{eROG} - 643.4) * (2.8 - \text{AWNDLAX}) * (-3.34\text{e-}04) + (\text{eROG} - 643.4) * (\text{AWNDLAX} - 2.8) * (-4.99\text{e-}05) + (26.91 - \text{ENSOmonthly}) * (\text{eROG} - 643.4) * (-3.03\text{e-}04) + (28.3 - \text{ENSOmonthly}) * (\text{eROG} - 1081.5) * (-1.61\text{e-}03) + (28.3 - \text{ENSOmonthly}) * (\text{eROG} - 982.7) * (1.34\text{e-}03) + (235 - \text{dayofyear}) * (\text{TMAXBarstow} - 36.7) * (1.22*\text{e-}04)$ (2)

The R^2 when applied to predict the top 30 MDA8 ozone predictions was 0.83 and showed no overfitting (Table S4).

290 The model also had a high R^2 (0.95), and RMSE equaled 11.2 ppbV when predicting the 4th highest MDA8 ozone concentrations (Fig. 3B).

Multiple tests were performed using the different number of the remaining terms in the output model with 10-fold CV to improve the MARS model performance. The best model was obtained when there were 14 terms maintained in the MARS model (Fig. S2). The performance of the MARS model with 14 terms (R^2=0.83, RMSE= 10.19) was similar

295 to the MARS model with 16 terms (R^2=0.83, RMSE=10.27).

3.1.3 RF model

We first applied the same indicators and dataset as the GAM (GAM-SoCAB-8HR V1.0) in order to compare the results of the above two regression methods. In the base case run, we tried 0-500 trees to find the optimal number of

300 trees. Each tree chose two variables randomly that was equal to one-third of the total number of variables by default. The optimal number of trees was 467 based on the RMSE value (Fig. S3). The majority of the peak 30 and 4th highest MDA8 ozone concentrations can be explained by the RF model (RF-SoCAB-8HR V1.0) (R^2=0.81 and RMSE=10.9 ppbV for top 30 MDA8 ozone, and R^2=0.97 and RMSE=14.0 ppbV for 4th highest MDA8 ozone; Fig. 2C and Fig. 3C). The R^2 and RMSE values of the 10-fold CV results were similar to those using the original RF model, with only

305 a 0.01 difference in R^2 and about a 5% reduction of RMSE value that indicated this RF model had a high prediction accuracy and no overfitting (Table S4).

Two main hyperparameters affect the performance of RF models and can be tuned: the number of trees used in the RF model and the number of random variables in each tree. To improve the model performance further, we created a grid with hyperparameters that the number of indicators considered at each split from 2 to 8, and the number of trees

310 was 1000 to tune the RF. The optimal number of predictors in each tree was 2 due to the lowest out-of-bag (OOB)

error, the same as the default run. Also, the optimal number of trees after model tuning was the same as the default run.

Next, we included all the available indicators in the RF model after excluding the strongly correlated independent variables. Then we removed the statistically insignificant indicators based on the p-value and the variable importance

315 to find the optimal combination of independent variables in the RF model. The final model contained two more variables than the one above: maximum solar radiation and height at 850 mb. The importance of the additional variables was minor and had negligible impacts on the model performance (Fig. 4c). The optimal number of trees was equal to 495. The R^2 and RMSE values for the top 30 MDA8 ozone predictions were similar to those using the RF model with fewer variables, although the mean bias was reduced (Table S3). In addition, the model performance for

320 the 4th highest MDA8 ozone predictions was worse than that using the RF model with the same variables as GAM (Table 2). Therefore, the RF model with the same GAM's variables fit the peak and the annual 4th highest MDA8 ozone concentrations well.

3.1.4 SVR model

325 We first built the SVR model (SVR-SoCAB-8HR V1.0) using the same variables as the built GAM (GAM-SoCAB-8HR V1.0) above with the default setting (the cost was 1, and epsilon was 0.1). We used kernel functions to consider the interactions between the independent indicators. Several kernel functions have been used in machine learning models, including linear kernel, polynomial kernel, radial kernel, etc. In practice, we used the linear kernel for the linear relationship and the radial kernel for the nonlinear relationship. Owing to the nonlinear relationship between

330 the peak MDA8 ozone levels and emissions/meteorology, we applied the radial kernel to the independent variables. The regression method we used was epsilon regression, and the epsilon value is related to the margin tolerance.

The R^2 and RMSE values of the top 30 MDA8 ozone predictions were very similar to the RF model's results, but the MB was larger than that of the RF model (Table S3). Results for predicting the 4th highest MDA8 ozone predictions found that the method did not capture the variability as well as the other methods (R^2 =0.89 and RMSE=14.0 ppbV).

335 The CV results indicated that this SVR model is stable and has no overfitting (Table S4).

Two parameters significantly impact the improvement of predictions and can be defined by users: the value of cost and epsilon. So we ran the SVR model with a hyperparameter grid with the cost value from 1 to 512, and the epsilon from 0 to 1 with an interval of 0.1. The model achieved the best performance when the epsilon was 0.3 and the cost was 1. The predicted top 30 and 4th highest MDA8 ozone concentrations were similar to those using the built SVR

340 with default settings (Table 2, Table S3, and Fig. S5).

We then built the SVR model with all the independent variables we had and removed the insignificant variables using the p-value and variable importance. The optimal SVR model (SVRoptimal-SoCAB-8HR V1.0) contained the variables in the above GAM and height at 850 mb, and maximum solar radiation. The ideal epsilon value was 0.1, and the cost value was 1, the same as the default setting. Though the importance of these two additional variables was

345 close to 0, the model performance of the top 30 days of MDA8 ozone simulations improved slightly (R^2 =0.83 and RMSE=10.4 ppbV) (Fig. 4d and Table S3). However, the 4th highest MDA8 ozone predictions were less accurate compared to the R^2 and RMSE using this SVR model (SVRoptimal-SoCAB-8HR V1.0) and the SVR model (SVR-

SoCAB-8HR V1.0) with the same GAM variables ($R^2 =0.87$ and RMSE=14.3 ppbV for the SVRoptimal-SoCAB-8HR V1.0 and $R^2 =0.89$ and RMSE=14.0 ppbV for the SVR-SoCAB-8HR V1.0) (Table 2).

350

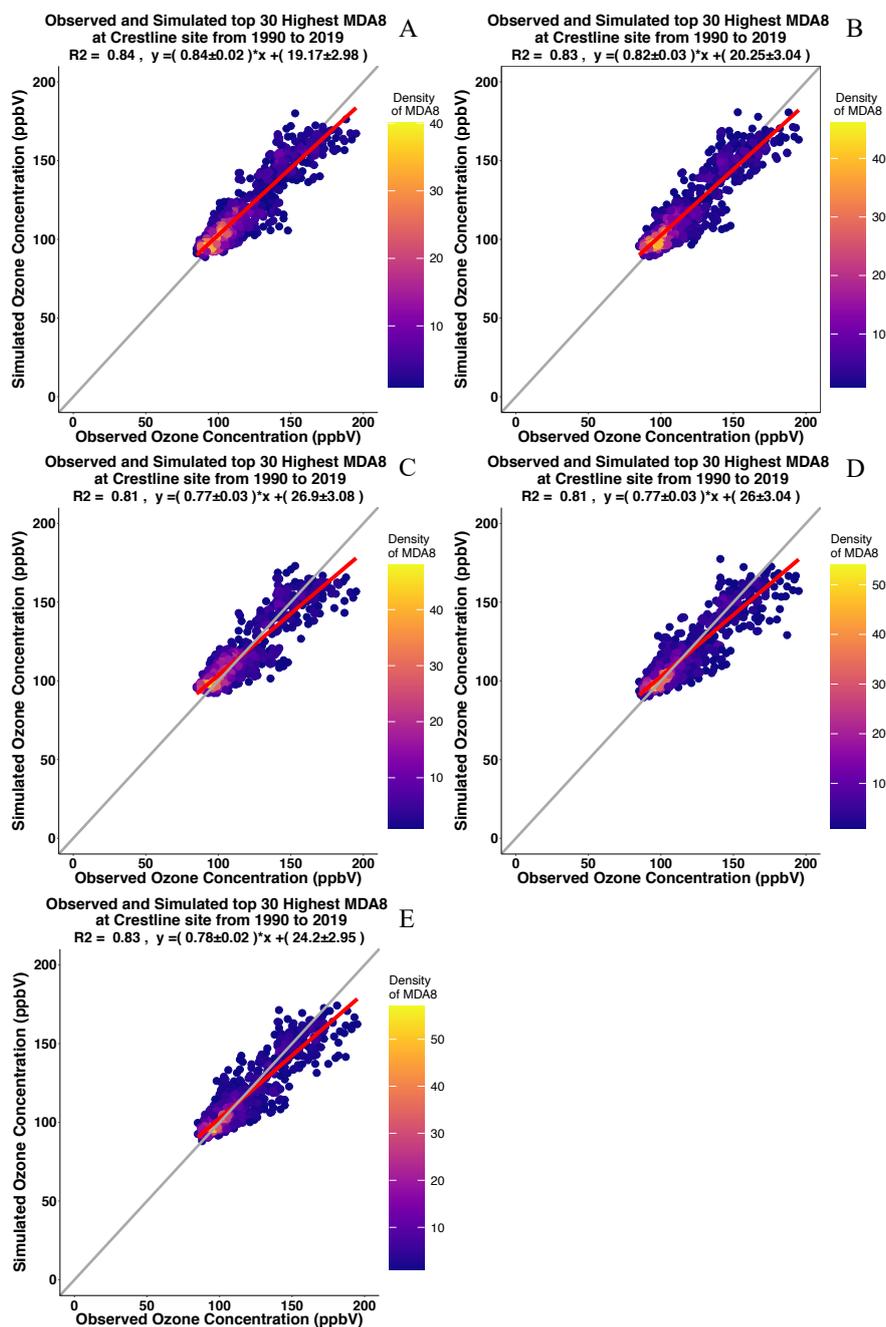



Figure 2: Comparison between the top 30 observed and predicted MDA8 ozone concentrations using the GAM-SoCAB-8HR V1.0 model (A), MARS-SoCAB-8HR V1.0 model (B), RF-SoCAB-8HR V1.0 model (C), SVR-SoCAB-8HR V1.0 model (D) and the SVRoptimal-SoCAB-8HR V1.0 model (E).

355

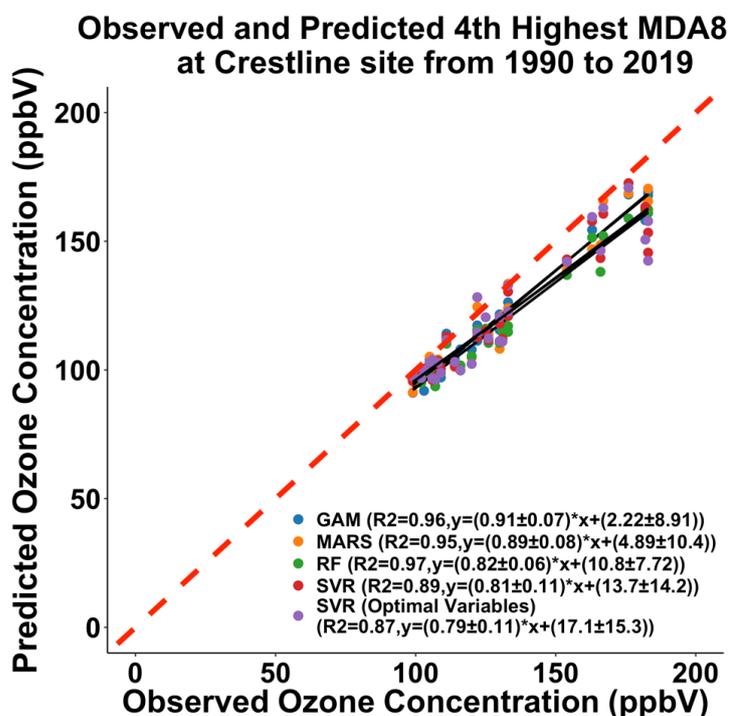



Figure 3: Comparison between the 4th highest observed and predicted MDA8 ozone concentrations using the GAM model built for the top 30 MDA8 ozone days at the Crestline site using GAM-SoCAB-8HR V1.0 model (blue), MARS-SoCAB-8HR V1.0 model (orange), RF-SoCAB-8HR V1.0 model (green), SVR-SoCAB-8HR V1.0 model (red) and SVRoptimal-SoCAB-8HR V1.0 model (purple).

360

Table 1: **Predictors used in the GAM and MARS model equations.**

Variable	Abbreviation	Unit
Day of the week (factor, from Mon to Sun)	dayofweek	None
Day of year (from 1 to 365/366)	dayofyear	None
Daily maximum surface temperature at the Barstow Airport site	TMAXBarstow	°C
Daily average wind speed at the LAX site	AWNDLAX	m/s
Daily RH at 850 mb	Mir850RH	%
Monthly Niño 3.4 indices	ENSOmonthly	°C
Annual averaged NOx emissions	eNOx	Tons/day

Annual averaged VOC emissions	eROG	Tons/day

365 **Table 2: Summary of statistical results of the 4th MDA8 ozone predictions using four methods at Crestline site.**

Method	Mean Bias (ppbV)	R^2	RMSE (ppbV)
GAM	-9.71	0.96	11.1
MARS model	-9.28	0.95	11.2
RF model[1]	-12.5	0.97	14.0
RF model[2]	-12.7	0.95	14.5
SVR model[1]	-10.6	0.89	14.0
SVR model[1]+tune	-10.3	0.89	13.7
SVR model[2]	-10.4	0.87	14.3

1 and 2: RF/ SVR model with the same variables as the GAM (GAM-SoCAB-8HR V1.0) and RF/ SVR model with
the optimal combination of the indicators.

370

3.2 Comparisons among the nonlinear methods

3.2.1 Statistical results and computational time (efficiency)

We compared the R^2, MB, and RMSE of the peak and 4th highest MDA8 ozone predictions using all these four models.
375 The statistical results of top 30 days MDA8 ozone simulations showed that all these four methods explain most of the
variability of observations, especially the GAM (Table S3). The GAM (GAM-SoCAB-8HR V1.0) had the lowest MB
and RMSE and the highest R^2 for the top 30 MDA8 ozone simulations among all these four methods. Also, the GAM
(GAM-SoCAB-8HR V1.0) showed the best stability of the top 30 MDA8 ozone predictions based on CV results (Table
S4).

380 In addition, these four numerical methods can capture the 4th highest MDA8 ozone variations well. The RF model
using the same variables as the built GAM (RF-SoCAB-8HR V1.0) of the 4th highest MDA8 ozone predictions with
an R^2 of 0.97, MB of -12.53 ppbV, and RMSE of 14.02 ppbV showed a lower model performance when compared to
the GAM whose R^2 equaled 0.96, MB was -9.71 ppbV, and RMSE was 11.07 ppbV. The R^2 for the MARS model
(MARS-SoCAB-8HR V1.0) equaled 0.95, MB was -9.28 ppbV, and RMSE was 11.16 ppbV. In comparison to the
385 performance of the GAM, the MARS model had a better MB value, but a worse R^2 and RMSE value. The SVR model
(SVR-SoCAB-8HR V1.0) showed the highest MB and RMSE value and lowest R^2 among all these four methods,
implying that the SVR model predictions gave the highest variations and lowest prediction accuracy. In general, all
these four methods showed a similar performance to the 4th highest MDA8 ozone predictions. The predicted 4th highest
MDA8 ozone levels with RF and SVR using the optimal variable combination had a lower R^2 and higher MB and
390 RMSE than those using the same variables as the GAM. Therefore, the variables used in the GAM (GAM-SoCAB-
8HR V1.0) were the best combination to build the models for peak ozone levels.

The statistical results and computational time need to be considered together to compare the model performance of all the models, especially for a large size dataset. There were no significant differences among these four methods in terms of the top 30 and the annual 4th highest ozone predictions. The GAM was marginally better compared to the other three models outside of cost-effectiveness. The computational requirements for each model in this work is small due to the small dataset size (Table 3). If computational time is a key factor, the MARS model can be a good choice for a larger dataset (Table 3).

395

Table 3: Summary of the computational time of each model.

400

Method	Computational time (s)
GAM	14
MARS model	0.04
RF model	1.2
SVR model	4.9

3.2.2 Two-step method

The R^2 values of the 4th highest MDA8 ozone predictions using these four regression methods were similar and agreed with the observations, but the RMSE and MB values were larger than desired. In order to reduce the bias, we applied a two-step method using the least squares method to the 4th highest MDA8 ozone predictions. The steps are shown

405

below:

1. Predicted the top 30 MDA8 ozone concentrations from 1990 to 2019 using the models built in section 3.1.
2. Extracted the annual predicted 4th highest MDA8 ozone concentrations based on the date of the observations

410

3. Applied the regression equation derived using the observations and the predictions in step 2 to the 4th maximum value in each year's top 30 MDA8 ozone predictions (as the response variable) to get the updated 4th highest MDA8 ozone predictions
4. Used the regression equation from step 3 with the updated predictions to get the improved 4th highest

415

MDA8 ozone predictions

The mean bias of the improved predictions was removed, the R^2 value was increased, and the RMSE was reduced. After applying the two-step method, the GAM showed the best model performance among all the models with the highest R^2 and lowest RMSE value. The performance of the MARS model and RF model were almost the same. The improved SVR model results were still the least accurate due to the lowest R^2 and highest MB and RMSE.

420

Table 4: Summary of statistical results of the 4th MDA8 ozone predictions after applying the two-step method using four methods at Crestline site.

425

Method	Mean Bias (ppbV)	R^2	RMSE (ppbV)
GAM	0	0.98	3.85
MARS model	0	0.97	4.54
RF model	0	0.97	4.55
SVR model	0	0.90	8.75

3.2.3 Relative importance of the independent variables

There are multiple methods to determine the importance of each independent variable of computational models, but the differences among all the methods are negligible. The algorithms used to calculate the variable importance of the

430 GAM, the RF model, and the SVR model are similar, based on the differences between the simulations using the original dataset and the dataset with one indicator's value randomly permutated. If the change of the simulations is significant, then that indicator is important, or vice versa. The variable importance shown is 1-r (r: the Pearson correlation coefficient between the simulations using the original and random permutation datasets) when the GAM is used. The RF and SVR model used the change of mean square error between the simulations using original and

435 random permutation datasets. The MARS model computed the variable importance by adding the indicator into the model and evaluating the error changes by GCV.

The precursors' emissions routinely are the most important indicators among all the variables in these four models that indicate that the emissions have more impact on the peak MDA8 ozone formation than the meteorology in the SoCAB. The maximum temperature is quite significant among all the meteorological factors. The GAM and the MARS model

440 also included the interaction terms between the emissions and maximum temperature, which capture more variability of the peak MDA8 ozone concentrations. The maximum temperature is related to solar radiation, which has an influence on the rate of photolysis reactions. RH at 850 mb showed relatively high importance in the RF and SVR models. It had a negative correlation with peak MDA8 ozone concentrations because of its relationship with precipitation and cloud cover, and, in consequence, reduce solar radiation and affect photolysis reactions.

445

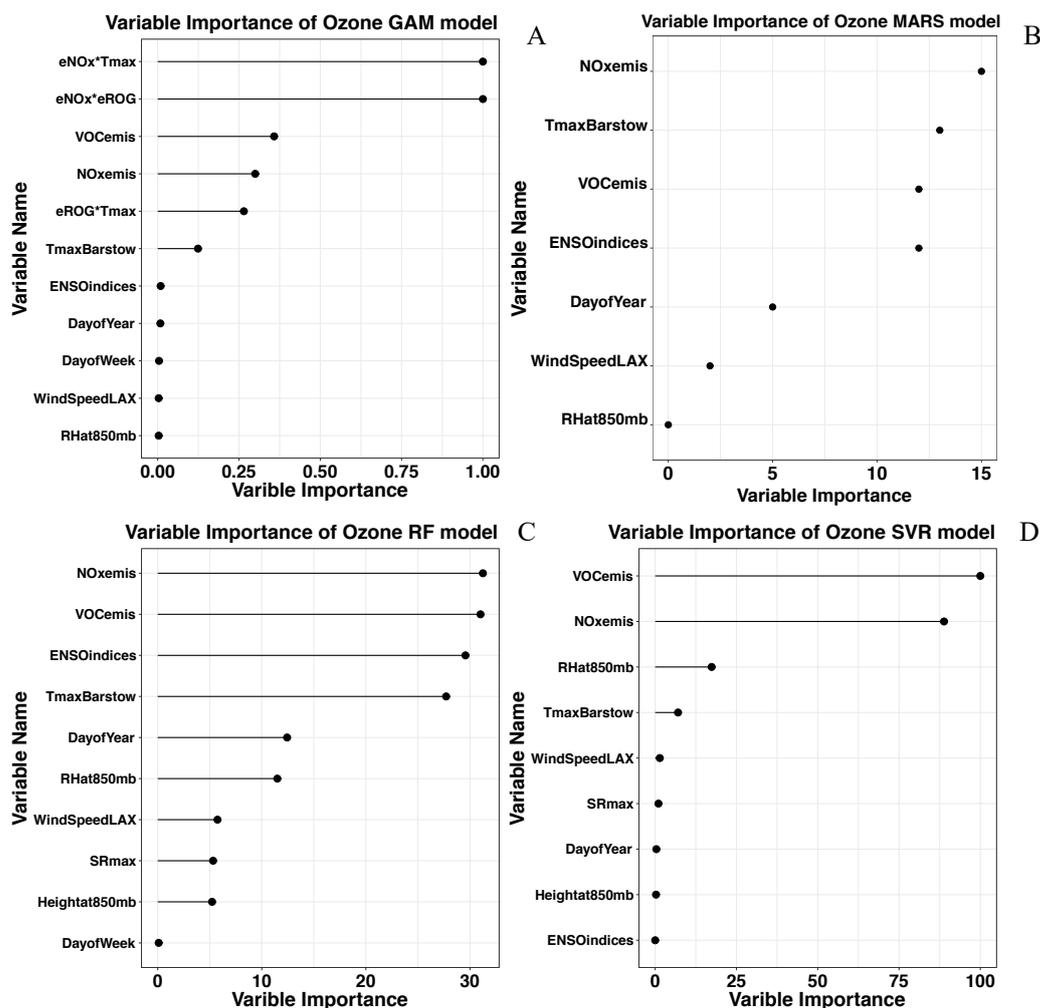
450

455

Figure 4: Variable importance for the top 30 days MDA8 ozone simulations using the GAM-SoCAB-8HR V1.0 model (A), MARS-SoCAB-8HR V1.0 model (B), RF-SoCAB-8HR V1.0 model (C) and the SVR-SoCAB-8HR V1.0 model (D). The variable importance for each model is calculated with different methods (see text).

460 **3.3 Limitations**

There are several limitations in the comparisons among these four models. Some significant factors to the 4th highest MDA8 ozone concentrations may be excluded from the models due to the relatively small dataset, that can consequently affect the prediction accuracy of the models; adding more available meteorological factors (e.g., cloud coverage, planetary boundary layer, surface wind direction, solar irradiance, etc.) and other large-scale climate indices

465 (e.g., Atlantic Multi-decal Oscillation (AMO), tropical Pacific sea surface temperature anomalies (TROP), etc.), we can expect an improvement in the model performance albeit at the cost of performance. Second, the running time of these four models with a small dataset does not show any significant differences. The computational time, however,

will be a key criterion if the dataset is quite large and may affect the best model choice. Third, since we included the local meteorological variables in the model equations and the models were developed for the Crestline site, these

470 models may not offer the same performance with the peak ozone levels at other sites in the SoCAB. However, applying the structure of the built model equations at the Crestline site to other regions worldwide with the local emissions and meteorological data around those areas, the models are equipped to capture the 4th highest MDA8 ozone concentrations well. Previous studies showed the emissions, maximum temperature, RH, wind speed, wind direction, and large-scale climate patterns have impacts on the daily MDA8 ozone concentrations in different regions in the world (Blanchard

475 et al., 2014; Blanchard et al., 2019; Camalier et al., 2007; García Nieto et al., 2014; Gong et al., 2018; Gong et al., 2017; Jeong et al., 2020; Jin et al., 2013; Ling et al., 2013; Liu et al., 2013; Lu et al., 1996; Lu et al., 2019; Luna et al., 2014; Ma et al., 2020; McClure et al., 2018; Sun et al., 2019). This is similar to the variable importance results in this study. In addition, although the GAM and the MARS model outperform RF and SVR models in this study, the machine learning methods (e.g., RF, neural network, and SVR) may potentially offer better performance than the

480 GAM and the MARS model with a significantly larger dataset. Finally, the models were not developed to predict daily MDA8 ozone concentrations because they are trained using the highest 30 MDA8 ozone levels of each year. The relationships between inputs and predicted ozone are very different at lower ozone levels.

4   Conclusions

485 This study compared four observation-based approaches to predict the 4th highest MDA8 ozone concentrations as a function of emissions, meteorological factors, and large-scale climate patterns. The statistical results showed that these four models with estimated emissions and observed meteorological factors can explain most of the top 30/ 4th highest MDA8 ozone concentrations' variations (R^2=0.81-0.84 for top 30 MDA8 ozone and R^2=0.89-0.97 for 4th high MDA8 ozone). Among the top 30 MDA8 ozone models, the GAM (GAM-SoCAB-8HR V1.0) achieved the highest R^2 (0.84)

490 and lowest RMSE value (9.74 ppbV), and the SVR (SVR-SoCAB-8HR V1.0) and RF (RF-SoCAB-8HR V1.0) achieved a lower R^2 value (0.81) and a higher RMSE value (10.9 ppbV). So, in terms of the top 30 highest MDA8 ozone predictions, there was little difference among these four models. These models showed a better performance for predicting the 4th highest MDA8 ozone predictions than the peak ozone level. All models had a high R^2 value (close to or higher than 0.9), but after considering RMSE and MB values, the GAM and the MARS model described the

495 dataset better and provided a significantly better prediction accuracy as compared to the RF and SVR models. Although the computational time of each model was small for the data set employed here, the MARS model required the least. The order of the variable importance of the factors of each model was similar. The precursors' emissions were the most significant factors that indicated the importance of the emissions impact on peak ozone levels. Maximum temperature presented relatively high importance among all the meteorological variables.

500

Competing interests. The authors declare that they have no conflict of interest.

Code and data availability. Code and data availability. Current version of the models are available from the project website: https://zenodo.org/record/6892066#.Ytw00uzMI8M   under   the   Creative   Commons   Attribution   4.0

505   International license. The dataset used to develop the model is available from the project website: https://zenodo.org/record/6892062#.Ytw1FezMI8M under the Creative Commons Attribution 4.0 International license.

Author contribution. PV, CI and AGR gave the research conception, ZG and AGR designed and performed the
510   research, ZG and KD collected the observed data, ZG analyzed the data, built the models and interpreted the results. All authors contributed to editing the manuscript

Acknowledgements. This study was supported, in part, by the South Coast Air Quality Management District (#20058), a generous donation from Howard T. Tellespsen, NASA's Health and Air Quality Applied Sciences Team (HAQAST)
515   program (#80NSSC21K0506) and the U.S. Environmental Protection Agency (EPA) under grant #84024601.We would like to thank Dr. Charles L. Blanchard for his comments that helped us significantly improve the manuscript.

References

Ainsworth, E. A., Yendrek, C. R., Sitch, S., Collins, W. J., & Emberson, L. D. (2012). The Effects of Tropospheric
520   Ozone on Net Primary Productivity and Implications for Climate Change. *Annual Review of Plant Biology, 63*(1), 637-661.
Aldrin, M., & Haff, I. (2005). Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmospheric Environment, 39*(11), 2145-2155.
Alduchov, O. A., & Eskridge, R. E. (1996). Improved Magnus Form Approximation of Saturation Vapor Pressure.
525   *Journal of Applied Meteorology, 35*(4), 601-609.
Aw, J., & Kleeman, M. J. (2003). Evaluating the first-order effect of intraannual temperature variability on urban air pollution. *Journal of Geophysical Research, 108*(D12).
Blanchard, C. L., Hidy, G. M., & Tanenbaum, S. (2014). Ozone in the southeastern United States: An observation-based model using measurements from the SEARCH network. *88*, 192-200.
530   Blanchard, C. L., Shaw, S. L., Edgerton, E. S., & Schwab, J. J. (2019). Emission influences on air pollutant concentrations in New York State: I. ozone. *Atmospheric Environment: X, 3*, 100033.
Camalier, L., Cox, W., & Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment, 41*(33), 7127-7137.
Chang, C.-C., & Lin, C.-J. (2011). LIBSVM. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1-27.
535   Cox, P. (2009). The California Almanac of Emissions and Air Quality - 2009 edition.
Cox, P. (2013). The California Almanac of Emissions and Air Quality - 2013 edition.
Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). *Support vector regression machines*. Paper presented at the Proceedings of the 9th International Conference on Neural Information Processing Systems, Denver, Colorado.
540   EPA, U. S. (2016). Pre-Generated Data Files. Last access May 23, 2020. Retrieved from: https://aqs.epa.gov/aqsweb/airdata/download_files.html.
EPA, U. S. (2020). Trends in Ozone Adjusted for Weather Conditions. Retrieved from https://www.epa.gov/air-trends/trends-ozone-adjusted-weather-conditions
Fan, R.-E., Chen, P.-H., & Lin, C.-J. (2005). Working Set Selection Using Second Order Information for Training
545   Support Vector Machines. *J. Mach. Learn. Res., 6*, 1889–1918.
Flynn, M. T., Mattson, E. J., Jaffe, D. A., & Gratz, L. E. (2021). Spatial patterns in summertime surface ozone in the Southern Front Range of the U.S. Rocky Mountains. *Elementa: Science of the Anthropocene, 9*(1).
Friedman, J. H., & Silverman, B. W. (1989). Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics, 31*(1), 3-21.
550   Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics, 19*(1), 1-67.

García Nieto, P. J., & Álvarez Antón, J. C. (2014). Nonlinear air quality modeling using multivariate adaptive regression splines in Gijón urban area (Northern Spain) at local scale. *Applied Mathematics and Computation, 235*, 50-65.

Gong, X., Hong, S., & Jaffe, D. A. (2018). Ozone in China: Spatial Distribution and Leading Meteorological Factors Controlling O3 in 16 Chinese Cities. *Aerosol and Air Quality Research, 18*(9), 2287-2300.

Gong, X., Kaulfus, A., Nair, U., & Jaffe, D. A. (2017). Quantifying O3 Impacts in Urban Areas Due to Wildfires Using a Generalized Additive Model. *Environmental Science & Technology, 51*(22), 13216-13223.

Gorai, A. K., Tuluri, F., Tchounwou, P. B., & Ambinakudige, S. (2015). Influence of local meteorology and NO2 conditions on ground-level ozone concentrations in the eastern part of Texas, USA. *Air Quality, Atmosphere & Health, 8*(1), 81-96.

Harrell, F. E. (2015). General Aspects of Fitting Regression Models. In (pp. 13-44): Springer International Publishing.

Hastie, T. (1991). *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie: Wadsworth \& Brooks/Cole.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science, 1*(3), 297-310.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman & Hall/CRC.

Hastie, T., & Tibshirani, R. (1996). Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 155-176.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction. 2nd ed.* New York: Springer.

Henneman, L. R. F., Holmes, H. A., Mulholland, J. A., & Russell, A. G. (2015). Meteorological detrending of primary and secondary pollutant concentrations: Method application and evaluation using long-term (2000–2012) data in Atlanta. *Atmospheric Environment, 119*, 201-210.

Hogrefe, C., Rao, S. T., Zurbenko, I. G., & Porter, P. S. (2000). Interpreting the Information in Ozone Observations and Model Predictions Relevant to Regulatory Policies in the Eastern United States. *Bulletin of the American Meteorological Society, 81*(9), 2083-2106.

Hong, C., Mueller, N. D., Burney, J. A., Zhang, Y., AghaKouchak, A., Moore, F. C., Qin, Y., Tong, D., & Davis, S. J. (2020). Impacts of ozone and climate change on yields of perennial crops in California. *Nature Food, 1*(3), 166-172.

Hu, C., Kang, P., Jaffe, D. A., Li, C., Zhang, X., Wu, K., & Zhou, M. (2021). Understanding the impact of meteorology on ozone in 334 cities of China. *Atmospheric Environment, 248*, 118221.

Huang, X. G., Shao, T. J., Zhao, J. B., Cao, J. J., & Lü, X. H. (2020). [Influencing Factors of Ozone Concentration in Xi'an Based on Generalized Additive Models]. *Huan Jing Ke Xue, 41*(4), 1535-1543.

Jeong, Y., Lee, H. W., & Jeon, W. (2020). Regional Differences of Primary Meteorological Factors Impacting O3 Variability in South Korea. *Atmosphere, 11*(1), 74.

Jin, L., Loisy, A., & Brown, N. J. (2013). Role of meteorological processes in ozone responses to emission controls in California's San Joaquin Valley. *Journal of Geophysical Research: Atmospheres, 118*(14), 8010-8022.

Keller, C. A., & Evans, M. J. (2019). Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. *Geoscientific Model Development, 12*(3), 1209-1225.

Kelley, M. C., Brown, M. M., Fedler, C. B., & Ardon-Dryer, K. (2020). Long Term Measurements of PM2.5 Concentrations in Lubbock, Texas. *Aerosol and Air Quality Research*.

Kleeman, M. J. (2008). A preliminary assessment of the sensitivity of air quality in California to global change. *Climatic Change, 87*(S1), 273-292.

Lawrence, M. G. (2005). The Relationship between Relative Humidity and the Dewpoint Temperature in Moist Air: A Simple Conversion and Applications. *Bulletin of the American Meteorological Society, 86*(2), 225-234.

Leathwick, J. R., Elith, J., & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling, 199*(2), 188-196.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News 2(3)*, 18--22.

Ling, Z. H., Guo, H., Zheng, J. Y., Louie, P. K. K., Cheng, H. R., Jiang, F., Cheung, K., Wong, L. C., & Feng, X. Q. (2013). Establishing a conceptual model for photochemical ozone pollution in subtropical Hong Kong. *Atmospheric Environment, 76*, 208-220.

Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K., & Tsao, C.-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PLOS ONE, 12*(7), e0179763.

605   Liu, T., Li, T. T., Zhang, Y. H., Xu, Y. J., Lao, X. Q., Rutherford, S., Chu, C., Luo, Y., Zhu, Q., Xu, X. J., Xie, H. Y., Liu, Z. R., & Ma, W. J. (2013). The short-term effect of ambient ozone on mortality is modified by temperature in Guangzhou, China. *Atmospheric Environment, 76*, 59-67.

Lu, R., & Turco, R. P. (1996). Ozone distributions over the los angeles basin: Three-dimensional simulations with the smog model. *Atmospheric Environment, 30*(24), 4155-4176.

610   Lu, X., Zhang, L., & Shen, L. (2019). Meteorology and Climate Influences on Tropospheric Ozone: a Review of Natural Sources, Chemistry, and Transport Patterns. *Current Pollution Reports*.

Luna, A. S., Paredes, M. L. L., De Oliveira, G. C. G., & Corrêa, S. M. (2014). Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. *Atmospheric Environment, 98*, 98-104.

615   Ma, Y., Ma, B., Jiao, H., Zhang, Y., Xin, J., & Yu, Z. (2020). An analysis of the effects of weather and air pollution on tropospheric ozone using a generalized additive model in Western China: Lanzhou, Gansu. *Atmospheric Environment, 224*, 117342.

Mahmud, A., Hixson, M., Hu, J., Zhao, Z., Chen, S. H., & Kleeman, M. J. (2010). Climate impact on airborne particulate matter concentrations in California using seven year analysis periods. *Atmospheric Chemistry and*
620   *Physics, 10*(22), 11097-11114.

MARS, W. a. o. **Multivariate adaptive regression spline**. Retrieved from https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_spline

McClure, C. D., & Jaffe, D. A. (2018). Investigation of high ozone events due to wildfire smoke in an urban area. *Atmospheric Environment, 194*, 146-157.

625   McGlynn, D., Mao, H., Sive, B., & Sharac, T. (2018). Understanding Long-Term Variations in Surface Ozone in United States (U.S.) National Parks. *Atmosphere, 9*(4), 125.

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012a). Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. NOAA National Climatic Data Center. doi: 10.7289/V5D21VHZ Retrieved May 27, 2020

630   Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012b). An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology, 29*(7), 897-910.

Milborrow, S. (2021). Regression Splines. R package version 5.3.1. Retrieved from https://CRAN.R-project.org/package=earth

NOAA. (2020). FIM Forecast Model: 500mb Wind Speed and 500mb Height Contours - Real-time. Retrieved from
635   https://sos.noaa.gov/datasets/fim-forecast-model-500mb-wind-speed-and-500mb-height-contours-real-time/

Oduro, S. D., Metia, S., Duc, H., Hong, G., & Ha, Q. P. (2015). Multivariate adaptive regression splines models for vehicular emission prediction. *Visualization in Engineering, 3*(1).

Oman, L. D., Douglass, A. R., Ziemke, J. R., Rodriguez, J. M., Waugh, D. W., & Nielsen, J. E. (2013). The ozone response to ENSO in Aura satellite measurements and a chemistry-climate simulation. *118*(2), 965-976.

640   Oman, L. D., Ziemke, J. R., Douglass, A. R., Waugh, D. W., Lang, C., Rodriguez, J. M., & Nielsen, J. E. (2011). The response of tropical tropospheric ozone to ENSO. *Geophysical Research Letters, 38*(13).

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., & Tapper, N. J. (2011). Quantifying the influence of local meteorology on air quality using generalized additive models. *Atmospheric Environment, 45*(6), 1328-1336.

Pernak, R., Alvarado, M., Lonsdale, C., Mountain, M., Hegarty, J., & Nehrkorn, T. (2019). Forecasting Surface O3 in
645   Texas Urban Areas Using Random Forest and Generalized Additive Models. *Aerosol and Air Quality Research, 9*(12), 2815-2826.

Pope, P. T., & Webster, J. T. (1972). The Use of an F-Statistic in Stepwise Regression Procedures. *Technometrics, 14*, 327-340.

Rao, S. T., Zurbenko, I. G., Neagu, R., Porter, P. S., Ku, J. Y., & Henry, R. F. (1997). Space and Time Scales in
650   Ambient Ozone Data. *Bulletin of the American Meteorological Society, 78*(10), 2153-2166.

Rodríguez-Pérez, R., Vogt, M., & Bajorath, J. (2017). Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction. *ACS Omega, 2*(10), 6371-6379.

Roy, S. S., Pratyush, C., & Barna, C. (2018). Predicting Ozone Layer Concentration Using Multivariate Adaptive
655   Regression Splines, Random Forest and Classification and Regression Tree. In (pp. 140-152): Springer International Publishing.

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Applied Sciences, 8*(12), 2570.

Scholkopf, B., & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization,*
660   *and Beyond*: MIT Press.

Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*: Wiley.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222.

Sotomayor-Olmedo, A., Aceves-Fernández, M. A., Gorrostieta-Hurtado, E., Pedraza-Ortega, C., Ramos-Arreguín, J. M., & Vargas-Soto, J. E. (2013). Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach. *International Journal of Intelligence Science, 03*(03), 126-135.

Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., De Hoogh, K., Kloog, I., Davoli, M., Michelozzi, P., & Bellander, T. (2020). A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden. *Atmosphere, 11*(3), 239.

Sun, L., Xue, L., Wang, Y., Li, L., Lin, J., Ni, R., Yan, Y., Chen, L., Li, J., Zhang, Q., & Wang, W. (2019). Impacts of meteorology and emissions on summertime surface ozone increases over central eastern China between 2003 and 2015. *Atmospheric Chemistry and Physics, 19*(3), 1455-1469.

Tin Kam, H. (1995, 14-16 Aug. 1995). *Random decision forests.* Paper presented at the Proceedings of 3rd International Conference on Document Analysis and Recognition.

Vong, C.-M., Ip, W.-F., Wong, P.-K., & Yang, J.-Y. (2012). Short-Term Prediction of Air Pollution in Macau Using Support Vector Machines. *Journal of Control Science and Engineering, 2012*, 1-11.

Wells, B., Dolwick, P., Eder, B., Evangelista, M., Foley, K., Mannshardt, E., Misenis, C., & Weishampel, A. (2021). Improved estimation of trends in U.S. ozone concentrations adjusted for interannual variability in meteorological conditions. *Atmospheric Environment, 248*, 118234.

Wikipedia contributors. (2022, April 16). Multivariate adaptive regression spline. In Wikipedia, The Free Encyclopedia. Retrieved 17:46, April 19, 2022, from https://en.wikipedia.org/w/index.php?title=Multivariate_adaptive_regression_spline&oldid=1083057440

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(1), 3-36.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2 edition ed.). Chapman and Hall/CRC.

Xu, L., Yu, J.-Y., Schnell, J. L., & Prather, M. J. (2017). The Seasonality and Geographic Dependence of ENSO Impacts on US Surface Ozone Variability.

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