Geosci. Model Dev., 15, 9015–9029, 2022
https://doi.org/10.5194/gmd-15-9015-2022

Geoscientific
Model Development
Open Access

Predicting peak daily maximum 8 h ozone and linkages to emissions and meteorology in Southern California using machine learning methods (SoCAB-8HR V1.0)

Ziqi Gao[1], **Yifeng Wang**[1], **Petros Vasilakos**[1], **Cesunica E. Ivey**[2,a], **Khanh Do**[2,3], **and 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 92521, USA
[3]Center for Environmental Research and Technology, University of California, Riverside, Riverside, CA 92521, USA
[a]now at: Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, CA 94720, USA

Correspondence: Ziqi Gao (zgao71@gatech.edu)

Received: 24 May 2022 – Discussion started: 28 July 2022
Revised: 24 November 2022 – Accepted: 2 December 2022 – Published: 16 December 2022

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 fourth highest 8 h 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), multivariate adaptive regression splines, random forest, and support vector regression, to develop observation-based relationships between the fourth 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 (NO_x) in the presence of sunlight (Seinfeld and Pandis, 2016). In addition to VOC and NO_x emissions, meteorology and large-scale climate patterns affect ozone (Aw and Kleeman, 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, suggests that observation-based models can be used to understand those relationships. Since the 17th century statistics have been used to record information about the wealth and population in Europe (Porter, 1981). For example, William Petty, a British scientist and economist, estimated the census data of Ireland through statistics (Banta, 1987). While the application of statistics had been restricted to a few fields until the 19th century, it gradually extended to other areas since then including physics, astronomy and recently air quality (Porter, 1995). At their core, statistical models aim at approximating a relationship between dependent and independent variables, with regression being the most commonly used method, a term that was coined by British statistician Francis Galton back in 1885 when he studied the trend of heights within families (Galton, 1889, 1888; Benirschke, 2004). The method however precedes the name, with the use of regression starting years before the term was introduced, dating back to the beginning of the 19th century with linear regression being applied to questions in astronomy, such as determining orbits of comets, while the least-squares method attributed to Adrien-Marie Legendre and Carl Friedrich Gauss was developed in the early 1800s (Stephen, 1981; Agarwal and Sen, 2014). At the start of the 20th century, some statisticians introduced the idea of nonlinear regression, trying to explain more complex systems (Fisher, 1922). Since then, as computational capacity increased dramatically in the past few decades, regression analysis has been widely used in most scientific fields.

The US Environmental Protection Agency's (EPA) National Ambient Air Quality Standard (NAAQS) for ozone is based on the annual maximum daily fourth highest 8 h 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, the US EPA adjusted the MDA8 ozone predictions with meteorological observations using generalized linear modeling (GLM) with natural spline smoothing functions in 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 and Tibshirani, 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 of the MDA8 ozone concentrations could be explained at different sites with meteorologically adjusted GAMs (Aldrin and Haff, 2005; Blanchard et al., 2014, 2019; Camalier et al., 2007; Flynn et al., 2021; Gong et al., 2018, 2017; Hu et al., 2021; Huang et al., 2020; Jeong et al., 2020; Ma et al., 2020; McClure and Jaffe, 2018; Pearce et al., 2011; Gao et al., 2022). 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 between ozone concentrations and precursors' concentrations/meteorological factors, including the interactions between the independent indicators (García Nieto and Álvarez Antón, 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 and Schölkopf, 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., 2014; Rybarczyk and Zalakeviciute, 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 and Evans, 2019; 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 fourth 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 emissions, meteorological conditions, large-scale climate events, chemical reactions, seasonal variations and weekend effects) to predict the annual fourth 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.

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 and Tibshirani, 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 and Tibshirani, 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 and polynomials. 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 a cubic spline is a third-order polynomial equation:

$$y_i = a_i \cdot x^3 + b_i \cdot x^2 + c_i \cdot x + d_i,$$

where a, b, c and d are the estimated coefficients of each basis function and the subscript i indicates the number of basis functions (equal 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 the R program was used to build the GAM between the peak MDA8 ozone concentrations and indicators (Hastie, 1991; Hastie and Tibshirani, 1990, 1986; 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 as follows (Friedman, 1991; Leathwick et al., 2006; 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 and Silverman, 1989; Hastie and Tibshirani, 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 and Silverman, 1989; Milborrow, 2021; Hastie et al., 2009), 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, $b = 1, 2, \ldots, m$) 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 dataset 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, $x_1x_2, \ldots, x_k$, belong to a leaf node and the corresponding response variables are $y_1y_2, \ldots, y_k$ respectively. The predicted result of a testing point x which falls into this leaf node should be

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

We used the "randomForest" package in the R software to build the RF model (Liaw and Wiener, 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. 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

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

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

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

where C is the cost, another user-defined variable that determines the tolerance of the model to points outside the 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 a nonlinear relationship between the response variable and indicators can be converted a Lagrangian dual problem (Schölkopf and Smola, 2001; Smola and Schölkopf, 2004).

We need to consider the interactions among features sometimes when we build computational models, so we need to map the data into a nonlinear feature space. The nonlinear feature space increases the dimension of the data space, 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 and Lin, 2011; Fan et al., 2005).

2.2 Model evaluation

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.

$$\text{Mean bias} = \frac{\sum_{i=1}^{n} (\hat{x}_i - x_i)}{n},$$

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

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 10 times. The averages of the R^2, MB and RMSE in these 10 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 and 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 Ocean 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 fourth highest MDA8 ozone concentrations in the SoCAB. This site had the annual fourth 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 Air Quality System (AQS) pre-generated data files (CARB, 2020). The total number of days of daily MDA8 ozone levels is 10 957. We used the top 30 MDA8 ozone days each year to develop the models for the fourth 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 fourth highest MDA8 ozone concentrations are considered, such as the day of the year, day of the week and meteorological-variable impacts, as there would only be 30 observations for model training. Furthermore, the size of the 30 years' fourth 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, 2019; Camalier et al., 2007), Niño 3.4 monthly indices, the day of the week and the day of the year. A detailed description of all the variables applied to test the final computational models is in Table S1 in the Supplement.

Estimated NO_x and VOC emissions in the SoCAB from 2000 to 2019 were acquired from CARB archives using the emissions in 2012 (Cox et al., 2013). The emissions between 1990 and 2000 were projected with the emissions in 2008 and 2012 (Cox et al., 2009, 2013). The detailed calculation is in the Supplement.

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 International Airport (LAX) and Barstow-Daggett Airport (Barstow Airport) from National Oceanic and Atmospheric Administration (NOAA) archives and CARB archives (Menne et al., 2012a, b). 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 mb (millibar). Using temperature and dew point temperature, we computed the relative humidity (RH) at 500 and 850 mb with the Clausius–Clapeyron equation (Alduchov and Eskridge, 1996; Lawrence, 2005). The height 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, 2011; Xu et al., 2017). Niño 3.4 monthly indices were obtained from the Climate Prediction Center (CPC) to represent ENSO events. To account for the daily variations and weekend effects of MDA8 ozone levels, we included the day of the week and the day of the year in the models (Seinfeld and Pandis, 2016).

3 Results

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 refine the model equation to provide the smallest Akaike information criterion (AIC) value after excluding the highly correlated indicators (Fig. S1) (Pope and Webster, 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 the year and the 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 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 a 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 (U.S. EPA, 2020; Henneman et al., 2015; Hogrefe et al., 2000; Rao et al., 1997; Blanchard et al., 2014; Camalier et al., 2007).

The final GAM (GAM-SoCAB-8HR V1.0) included emissions, meteorological factors, large-scale climate indices 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):

$$\begin{aligned}\log(\mathrm{MDA8}) = {} & a + \mathrm{dayofweek\ (factor)} + \mathrm{dayofyear} \\ & + s\,(\mathrm{TMAXBarstow}) + s(\mathrm{Mir850RH}) \\ & + s(\mathrm{AWNDLAX}) + s(\mathrm{eNO}_x) + s(\mathrm{eROG}) \\ & + s(\mathrm{eNO}_x, \mathrm{TMAXBarstow}) \\ & + s(\mathrm{eROG,TMAXBarstow}) \\ & + s(\mathrm{eNO}_x, \mathrm{eROG}) + \mathrm{ENSOmonthly} + \mathrm{e}. \qquad (1)\end{aligned}$$

The correlation (R^2) between independent variables in the final GAM was tested (Fig. 1). The correlation between VOC and NO_x emissions was high at close to 1. However, both are the precursors of MDA8 ozone, so these factors were not removed during the model development. Other than emis-

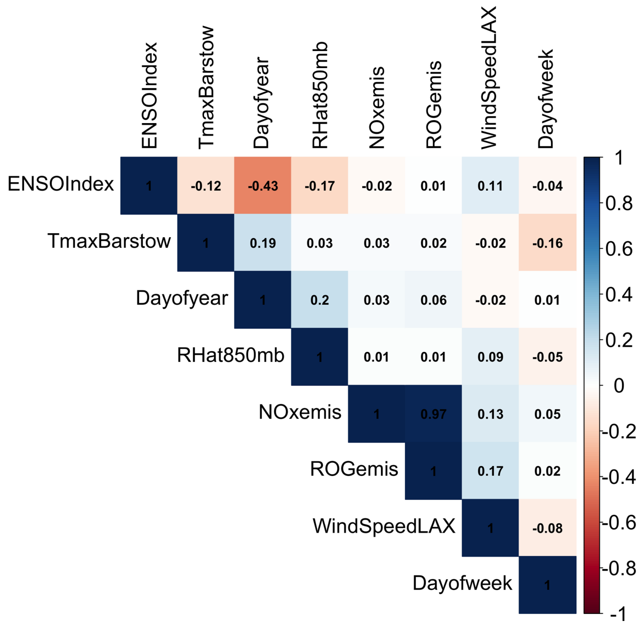

Figure 1. Correlation value between the independent variables (only valid for GAM).

sions, the correlation among all the significant independent variables in the final models is negligible (Fig. 1).

A total of 84 % of the variability of the peak MDA8 ozone concentrations can be explained using this GAM (Fig. 2a). The 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 fourth highest MDA8 ozone predictions from 1990 to 2019 (Fig. 3a).

3.1.2 MARS model

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 NO_x and VOC emissions, the maximum temperature at Barstow-Daggett Airport, the average wind speed at LAX, Niño 3.4 monthly indices, the day of the year and 11 interaction terms between indicators. Similar to GAM, we applied the 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-SoCAB-8HR V1.0) is shown below (Eq. 2). A detailed description of each variable is in Table 1:

$$\begin{aligned}
\log(\mathrm{MDA8}) = {} & 4.59 + (1135.1 - \mathrm{eNO}_x) \\
& \cdot (-9.57 \times 10^{-4}) \\
& + (\mathrm{eNO}_x - 1135.1) \cdot (1.55 \times 10^{-3}) \\
& + (643.4 - \mathrm{eROG}) \cdot (2.67 \times 10^{-3}) \\
& + (\mathrm{TMAXBarstow} - 36.7) \cdot 0.022 \\
& + (1135.1 - \mathrm{eNO}_x) \cdot (\mathrm{eROG} - 450.7) \\
& \cdot (4.29 \times 10^{-6}) + (\mathrm{eNO}_x - 1415.8) \\
& \cdot (\mathrm{eROG} - 643.4) \cdot (-1.08 \times 10^{-6}) \\
& + (\mathrm{eNO}_x - 1225) \cdot (\mathrm{TMAXBarstow} - 36.7) \\
& \cdot (7.21 \times 10^{-4}) + (1081.5 - \mathrm{eROG}) \\
& \cdot (\mathrm{TMAXBarstow} - 36.7) \cdot (-2.18 \times 10^{-5}) \\
& + (\mathrm{eROG} - 1081.5) \\
& \cdot (\mathrm{TMAXBarstow} - 36.7) \cdot (-3.95 \times 10^{-4}) \\
& + (\mathrm{eROG} - 643.4) \cdot (2.8 - \mathrm{AWNDLAX}) \\
& \cdot (-3.34 \times 10^{-4}) + (\mathrm{eROG} - 643.4) \\
& \cdot (\mathrm{AWNDLAX} - 2.8) \cdot (-4.99 \times 10^{-5}) \\
& + (26.91 - \mathrm{ENSOmonthly}) \\
& \cdot (\mathrm{eROG} - 643.4) \cdot (-3.03 \times 10^{-4}) \\
& + (28.3 - \mathrm{ENSOmonthly}) \\
& \cdot (\mathrm{eROG} - 1081.5) \cdot (-1.61 \times 10^{-3}) \\
& + (28.3 - \mathrm{ENSOmonthly}) \\
& \cdot (\mathrm{eROG} - 982.7) \cdot (1.34 \times 10^{-3}) \\
& + (235 - \mathrm{dayofyear}) \\
& \cdot (\mathrm{TMAXBarstow} - 36.7) \cdot (1.22 \times 10^{-4}).
\end{aligned} \tag{2}$$

The R^2 when applied to predict the top 30 MDA8 ozone predictions was 0.83 and showed no overfitting (Table S4). The model also had a high R^2 (0.95), and RMSE equaled 11.2 ppbv when predicting the fourth 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 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 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 top 30 and fourth 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 the top 30 MDA8 ozone concentrations and $R^2 = 0.97$ and RMSE = 14.0 ppbv for the fourth highest MDA8 ozone concentration; Figs. 2c and 3c). The R^2 and RMSE values of the 10-fold CV results were similar to those using the original RF model, with only a 0.01 difference in R^2 and about a 5 % reduction of the 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 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 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 the fourth 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 fourth highest MDA8 ozone concentrations well.

3.1.4 SVR model

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 the 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 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 fourth 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). 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 fourth highest MDA8 ozone concentrations were similar to those using the built SVR with default settings (Tables 2 and S3, 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 close to 0, the model performance of simulations of the top 30 MDA8 ozone days improved slightly ($R^2 = 0.83$ and RMSE = 10.4 ppbv) (Fig. 4d and Table S3). However, the fourth 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).

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 fourth highest MDA8 ozone predictions using all these four models. The statistical results of simulations of the top 30 MDA8 ozone days 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

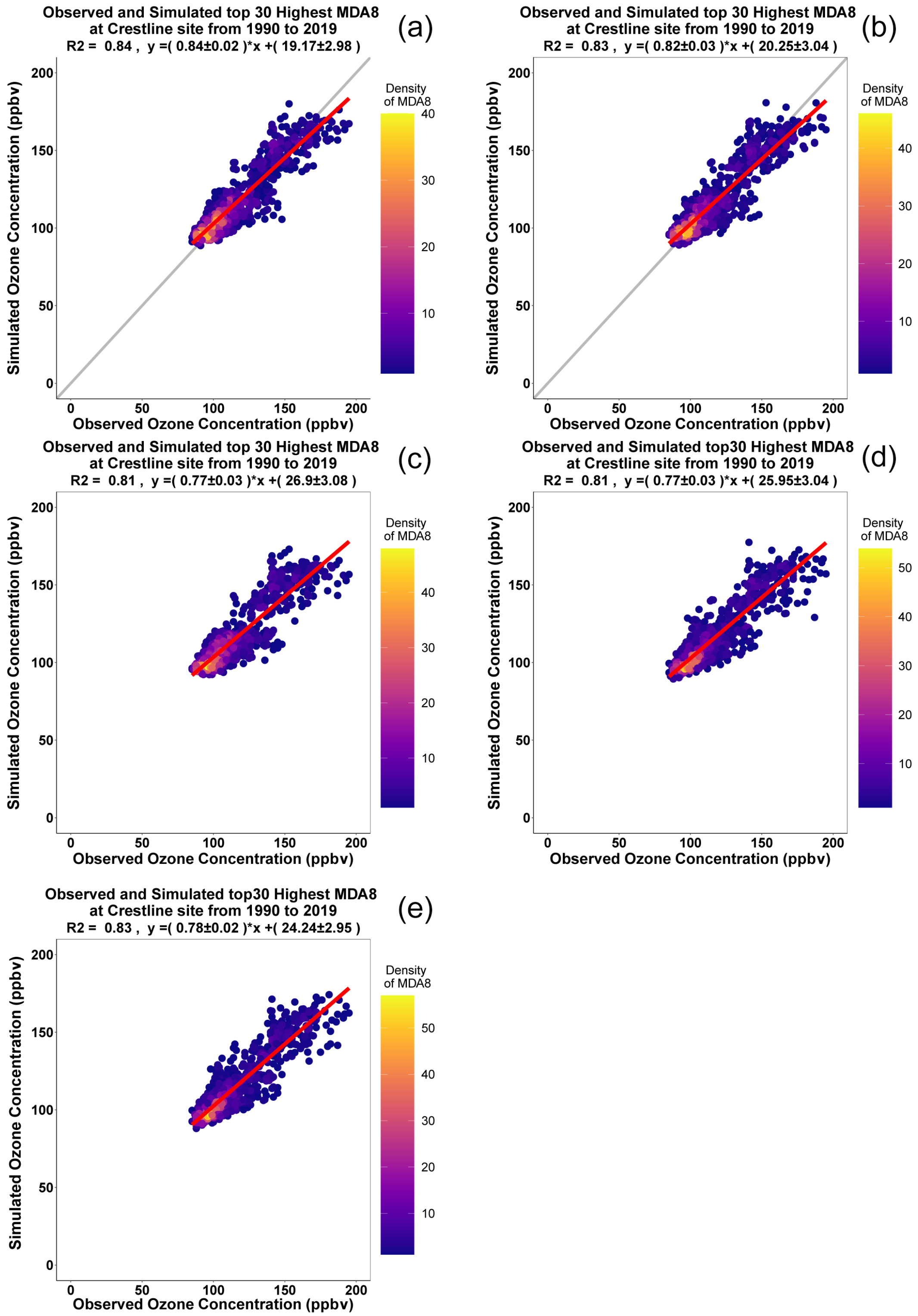

Figure 2. Comparison between the top 30 observed and predicted MDA8 ozone concentrations using the GAM-SoCAB-8HR V1.0 model **(a)**, MARS-SoCAB-8HR V1.0 model **(b)**, RF-SoCAB-8HR V1.0 model **(c)**, SVR-SoCAB-8HR V1.0 model **(d)** and the SVRoptimal-SoCAB-8HR V1.0 model **(e)**.

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

Variable	Abbreviation	Unit
Day of the week (factor, from Monday to Sunday)	dayofweek	None
Day of the 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	$m\,s^{-1}$
Daily RH at 850 mb	Mir850RH	%
Monthly Niño 3.4 indices	ENSOmonthly	°C
Annual averaged NO_x emissions	eNO_x	$t\,d^{-1}$
Annual averaged VOC emissions	eROG	$t\,d^{-1}$

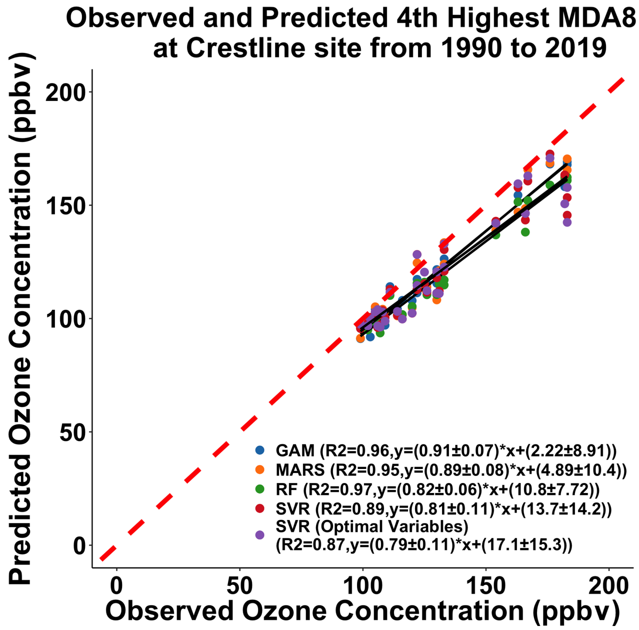

Figure 3. Comparison between the fourth highest observed and predicted MDA8 ozone concentrations using the GAM model built for the top 30 MDA8 ozone days at the Crestline site using the GAM-SoCAB-8HR V1.0 model (blue), MARS-SoCAB-8HR V1.0 model (orange), RF-SoCAB-8HR V1.0 model (green), SVR-SoCAB-8HR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

Table 2. Summary of statistical results of the fourth MDA8 ozone predictions using four methods at the 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[a]	−12.5	0.97	14.0
RF model[b]	−12.7	0.95	14.5
SVR model[a]	−10.6	0.89	14.0
SVR model[a]+ tune	−10.3	0.89	13.7
SVR model[b]	−10.4	0.87	14.3

[a,b] 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.

of the top 30 MDA8 ozone predictions based on CV results (Table S4).

In addition, these four numerical methods can capture the fourth highest MDA8 ozone variations well. The RF model using the same variables as the built GAM (RF-SoCAB-8HR V1.0) of the fourth 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 with an MB of −9.28 ppbv and RMSE of 11.16 ppbv. In comparison to the 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 fourth highest MDA8 ozone predictions. The predicted fourth highest MDA8 ozone levels with RF and SVR using the optimal variable combination had a lower R^2 and higher MB and 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 fourth 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).

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

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

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

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

3.2.2 Two-step method

The R^2 values of the fourth 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 fourth highest MDA8 ozone predictions. The steps are shown below:

1. Predict the top 30 MDA8 ozone concentrations from 1990 to 2019 using the models built in Sect. 3.1.

2. Extract the annual predicted fourth highest MDA8 ozone concentrations based on the date of the observations.

3. Apply the regression equation derived using the observations and the predictions in step 2 to the fourth maximum value in each year's top 30 MDA8 ozone predictions (as the response variable) to get the updated fourth highest MDA8 ozone predictions.

4. Use the regression equation from step 3 with the updated predictions to get the improved fourth highest 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.

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 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 the original and 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 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, reduced solar radiation and affected photolysis reactions.

3.3 Limitations

There are several limitations in the comparisons among these four models. Some significant factors to the fourth 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 and solar irradiance) and other large-scale climate indices (e.g., Atlantic Multidecadal Oscillation (AMO) and tropical Pacific sea surface temperature anomalies (TROP)), 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

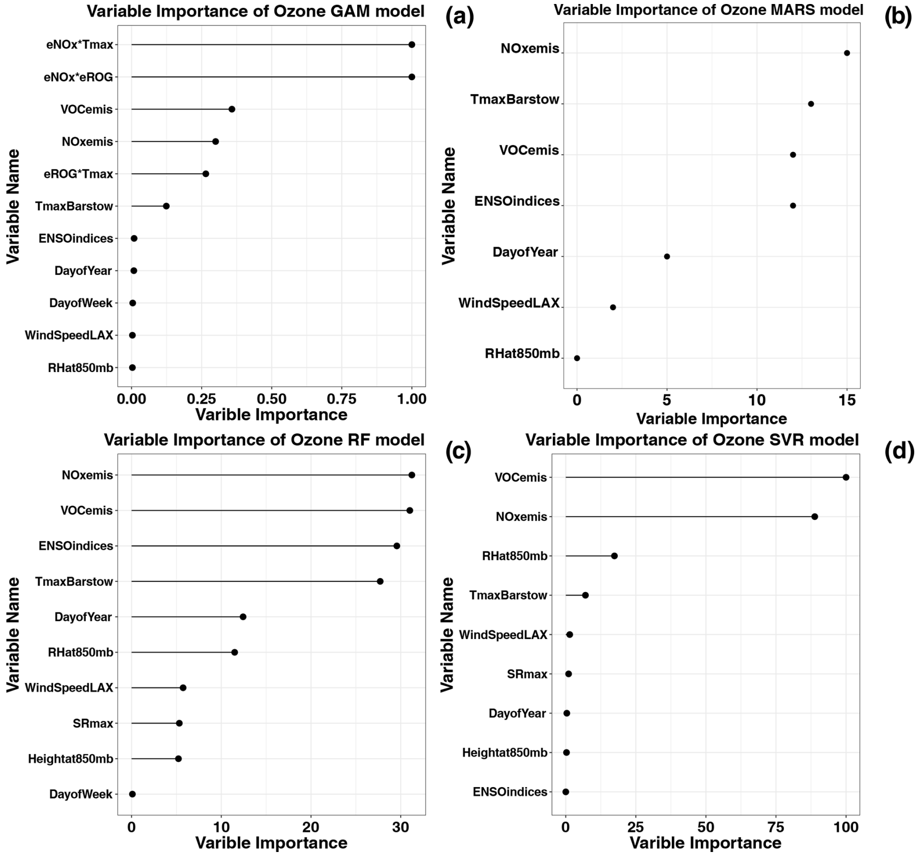

Figure 4. Variable importance for simulations of the top 30 MDA8 ozone days using the GAM-SoCAB-8HR V1.0 model (**a**), MARS-SoCAB-8HR V1.0 model (**b**), RF-SoCAB-8HR V1.0 model (**c**) and SVR-SoCAB-8HR V1.0 model (**d**). The variable importance for each model is calculated with different methods (see text). SR: solar radiation.

developed for the Crestline site, these models may not offer the same performance with the peak ozone levels at other sites in the SoCAB. Given that Crestline is downwind of Los Angeles, which then is bordered by the Pacific Ocean, the models using SoCAB emissions capture the upwind conditions. In other regions, such models could be expanded to include both local emissions and upwind states' emissions. 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 et al., 2014, 2019; Camalier et al., 2007; García Nieto and Álvarez Antón, 2014; Gong et al., 2018, 2017; Jeong et al., 2020; Jin et al., 2013; Ling et al., 2013; Liu et al., 2013; Lu and Turco, 1996; Lu et al., 2019; Luna et al., 2014; Ma et al., 2020; McClure and Jaffe, 2018; Sun et al., 2019). This is similar to the variable-importance results in this study. In addition, although the GAM and MARS model outperform the RF and SVR model in this study, the machine learning methods (e.g., RF, neural network and SVR) may potentially offer better performance than the 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

This study compared four observation-based approaches to predict the fourth 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 variations of the top 30 and fourth highest MDA8 ozone concentrations ($R^2 = 0.81$–0.84 for the top 30 MDA8 ozone concentrations and $R^2 = 0.89$–0.97 for the fourth highest MDA8 ozone concentrations). Among the top 30 MDA8 ozone models, the GAM (GAM-SoCAB-8HR V1.0) achieved the highest R^2 (0.84) 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 fourth 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 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 dataset 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 for indicating the importance of the emissions impact on peak ozone levels. Maximum temperature presented relatively high importance among all the meteorological variables.

Code and data availability. Current versions of the models are available from the project website at https://doi.org/10.5281/zenodo.6892066 (Gao, 2022a) under the Creative Commons Attribution 4.0 International license. The dataset used to develop the model is available from the project website at https://doi.org/10.5281/zenodo.6892062 (Gao, 2022b) under the Creative Commons Attribution 4.0 International license.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-15-9015-2022-supplement.

Author contributions. PV, CEI and AGR conceived the research. ZG and AGR designed and performed the 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.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We would like to thank Charles L. Blanchard for his comments that helped us significantly improve the manuscript.

Financial support. This research has been supported by the South Coast Air Quality Management District (grant no. 20058), a generous donation from Howard T. Tellepsen, NASA's Health and Air Quality Applied Sciences Team (HAQAST) program (grant no. 80NSSC21K0506), the Phillips 66 Company, and the US Environmental Protection Agency (EPA; grant no. 84024601).

Review statement. This paper was edited by Volker Grewe and reviewed by William Stockwell and one anonymous referee.

References

Agarwal, R. and Sen, S.: Creators of Mathematical and Computational Sciences, https://doi.org/10.1007/978-3-319-10870-4, 2014.

Ainsworth, E. A., Yendrek, C. R., Sitch, S., Collins, W. J., and Emberson, L. D.: The Effects of Tropospheric Ozone on Net Primary Productivity and Implications for Climate Change, Annu. Rev. Plant Biol., 63, 637–661, https://doi.org/10.1146/annurev-arplant-042110-103829, 2012.

Aldrin, M. and Haff, I.: Generalised additive modelling of air pollution, traffic volume and meteorology, Atmos. Environ., 39, 2145–2155, https://doi.org/10.1016/j.atmosenv.2004.12.020, 2005.

Alduchov, O. A. and Eskridge, R. E.: Improved Magnus Form Approximation of Saturation Vapor Pressure, J. Appl. Meteorol., 35, 601–609, https://doi.org/10.1175/1520-0450(1996)035<0601:imfaos>2.0.co;2, 1996.

Aw, J. and Kleeman, M. J.: Evaluating the first-order effect of intraannual temperature variability on urban air pollution, J. Geophys. Res., 108, 4365, https://doi.org/10.1029/2002jd002688, 2003.

Banta, J. E.: Sir William Petty: Modern epidemiologist (1623–1687), J. Commun. He., 12, 185–198, https://doi.org/10.1007/bf01323480, 1987.

Benirschke, K.: Francis Galton: Pioneer of Heredity and Biometry, J. Heredity, 95, 273–273, https://doi.org/10.1093/jhered/esh039, 2004.

Blanchard, C. L., Hidy, G. M., and Tanenbaum, S.: Ozone in the southeastern United States: An observation-based model using measurements from the SEARCH network, Atmos. Environ., 88, 192–200, https://doi.org/10.1016/j.atmosenv.2014.02.006, 2014.

Blanchard, C. L., Shaw, S. L., Edgerton, E. S., and Schwab, J. J.: Emission influences on air pollutant concentrations in New York State: I. ozone, Atmos. Environ. X, 3, 100033, https://doi.org/10.1016/j.aeaoa.2019.100033, 2019.

Camalier, L., Cox, W., and Dolwick, P.: The effects of meteorology on ozone in urban areas and their use in assessing ozone trends, Atmos. Environ., 41, 7127–7137, https://doi.org/10.1016/j.atmosenv.2007.04.061, 2007.

CARB: Air Quality and Meteorological Information System (AQMIS), https://www.arb.ca.gov/aqmis2/aqdselect.php, last access: 27 May 2020.

Chang, C.-C. and Lin, C.-J.: LIBSVM, ACM Transactions on Intelligent Systems and Technology, 2, 1–27, https://doi.org/10.1145/1961189.1961199, 2011.

Cox, P., Delao, A., Komorniczak, A., and Weller, R.: The California Almanac of Emissions and Air Quality – 2009 edition, Planning and Technical Support Division California Air Resources Board, 2009.

Cox, P., Delao, A., and Komorniczak, A.: The California Almanac of Emissions and Air Quality – 2013 edition, Air Quality Planning and Science Division California Air Resources Board, 2013.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V.: Support vector regression machines, Proceedings of the

9th International Conference on Neural Information Processing Systems, Denver, Colorado, 1996.

Fan, R.-E., Chen, P.-H., and Lin, C.-J.: Working Set Selection Using Second Order Information for Training Support Vector Machines, J. Mach. Learn. Res., 6, 1889–1918, 2005.

Fisher, R. A.: On the mathematical foundations of theoretical statistics, Philos. T. Roy. Soc. Lond. A, 222, 309–368, https://doi.org/10.1098/rsta.1922.0009, 1922.

Flynn, M. T., Mattson, E. J., Jaffe, D. A., and Gratz, L. E.: Spatial patterns in summertime surface ozone in the Southern Front Range of the U.S. Rocky Mountains, Elementa: Science of the Anthropocene, 9, 00104, https://doi.org/10.1525/elementa.2020.00104, 2021.

Friedman, J. H.: Multivariate Adaptive Regression Splines, Ann. Stat., 19, 1–67, https://doi.org/10.1214/aos/1176347963, 1991.

Friedman, J. H. and Silverman, B. W.: Flexible Parsimonious Smoothing and Additive Modeling, Technometrics, 31, 3–21, https://doi.org/10.2307/1270359, 1989.

Galton, F.: Co-Relations and Their Measurement, Chiefly from Anthropometric Data, P. Roy. Soc. Lond., 45, 135–145, 1888.

Galton, F. S.: Natural inheritance, Macmillan, London, https://doi.org/10.5962/bhl.title.32181, 1889.

Gao, Z.: Predicting peak daily maximum 8-hour ozone, and linkages to emissions and meteorology, in Southern California using machine learning methods, Zenodo [code], https://doi.org/10.5281/zenodo.6892066, 2022a.

Gao, Z.: Predicting peak daily maximum 8-hour ozone, and linkages to emissions and meteorology, in Southern California using machine learning methods, Zenodo [data set], https://doi.org/10.5281/zenodo.6892062, 2022b.

Gao, Z., Ivey, C. E., Blanchard, C. L., Do, K., Lee, S.-M., and Russell, A. G.: Separating emissions and meteorological impacts on peak ozone concentrations in Southern California using generalized additive modeling, Environ. Pollut., 307, 119503, https://doi.org/10.1016/j.envpol.2022.119503, 2022.

García Nieto, P. J. and Álvarez Antón, J. C.: Nonlinear air quality modeling using multivariate adaptive regression splines in Gijón urban area (Northern Spain) at local scale, Appl. Math. Comput., 235, 50–65, https://doi.org/10.1016/j.amc.2014.02.096, 2014.

Gong, X., Kaulfus, A., Nair, U., and Jaffe, D. A.: Quantifying O3 Impacts in Urban Areas Due to Wildfires Using a Generalized Additive Model, Environ. Sci. Technol., 51, 13216–13223, https://doi.org/10.1021/acs.est.7b03130, 2017.

Gong, X., Hong, S., and Jaffe, D. A.: Ozone in China: Spatial Distribution and Leading Meteorological Factors Controlling O_3 in 16 Chinese Cities, Aerosol Air Qual. Res., 18, 2287–2300, https://doi.org/10.4209/aaqr.2017.10.0368, 2018.

Gorai, A. K., Tuluri, F., Tchounwou, P. B., and Ambinakudige, S.: Influence of local meteorology and NO_2 conditions on ground-level ozone concentrations in the eastern part of Texas, USA, Air Quality, Atmos. He., 8, 81–96, https://doi.org/10.1007/s11869-014-0276-5, 2015.

Harrell, F. E.: General Aspects of Fitting Regression Models, Springer International Publishing, 13–44, https://doi.org/10.1007/978-3-319-19425-7_2, 2015.

Hastie, T.: Generalized additive models, Chapter 7, in: Statistical Models in S, edited by: Chambers, J. M. and Hastie, T. J., Wadsworth & Brooks/Cole, 1991.

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Stat. Sci., 1, 297–310, https://doi.org/10.1214/ss/1177013604, 1986.

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

Hastie, T. and Tibshirani, R.: Discriminant Analysis by Gaussian Mixtures, J. Roy. Stat. Soc. B, 58, 155–176, 1996.

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

Henneman, L. R. F., Holmes, H. A., Mulholland, J. A., and Russell, A. G.: Meteorological detrending of primary and secondary pollutant concentrations: Method application and evaluation using long-term (2000–2012) data in Atlanta, Atmos. Environ., 119, 201–210, https://doi.org/10.1016/j.atmosenv.2015.08.007, 2015.

Hogrefe, C., Rao, S. T., Zurbenko, I. G., and Porter, P. S.: Interpreting the Information in Ozone Observations and Model Predictions Relevant to Regulatory Policies in the Eastern United States, B. Am. Meteorol. Soc., 81, 2083–2106, https://doi.org/10.1175/1520-0477(2000)081<2083:itiioo>2.3.co;2, 2000.

Hong, C., Mueller, N. D., Burney, J. A., Zhang, Y., AghaKouchak, A., Moore, F. C., Qin, Y., Tong, D., and Davis, S. J.: Impacts of ozone and climate change on yields of perennial crops in California, Nature Food, 1, 166–172, https://doi.org/10.1038/s43016-020-0043-8, 2020.

Hu, C., Kang, P., Jaffe, D. A., Li, C., Zhang, X., Wu, K., and Zhou, M.: Understanding the impact of meteorology on ozone in 334 cities of China, Atmos. Environ., 248, 118221, https://doi.org/10.1016/j.atmosenv.2021.118221, 2021.

Huang, X. G., Shao, T. J., Zhao, J. B., Cao, J. J., and Lü, X. H.: Influencing Factors of Ozone Concentration in Xi'an Based on Generalized Additive Models, Huan Jing Ke Xue, 41, 1535–1543, https://doi.org/10.13227/j.hjkx.201906067, 2020.

Jeong, Y., Lee, H. W., and Jeon, W.: Regional Differences of Primary Meteorological Factors Impacting O_3 Variability in South Korea, Atmosphere, 11, 74, https://doi.org/10.3390/atmos11010074, 2020.

Jin, L., Loisy, A., and Brown, N. J.: Role of meteorological processes in ozone responses to emission controls in California's San Joaquin Valley, J. Geophys. Res.-Atmos., 118, 8010–8022, https://doi.org/10.1002/jgrd.50559, 2013.

Keller, C. A. and Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, Geosci. Model Dev., 12, 1209–1225, https://doi.org/10.5194/gmd-12-1209-2019, 2019.

Kelley, M. C., Brown, M. M., Fedler, C. B., and Ardon-Dryer, K.: Long Term Measurements of $PM_{2.5}$ Concentrations in Lubbock, Texas, Aerosol Air Qual. Res., 20, 1306–1318, https://doi.org/10.4209/aaqr.2019.09.0469, 2020.

Kleeman, M. J.: A preliminary assessment of the sensitivity of air quality in California to global change, Clim. Change, 87, 273–292, https://doi.org/10.1007/s10584-007-9351-3, 2008.

Lawrence, M. G.: The Relationship between Relative Humidity and the Dewpoint Temperature in Moist Air: A Simple Conversion and Applications, B. Am. Meteorol. Soc., 86, 225–234, https://doi.org/10.1175/BAMS-86-2-225, 2005.

Leathwick, J. R., Elith, J., and Hastie, T.: Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions, Ecol. Model., 199, 188–196, https://doi.org/10.1016/j.ecolmodel.2006.05.022, 2006.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, 2002.

Ling, Z. H., Guo, H., Zheng, J. Y., Louie, P. K. K., Cheng, H. R., Jiang, F., Cheung, K., Wong, L. C., and Feng, X. Q.: Establishing a conceptual model for photochemical ozone pollution in subtropical Hong Kong, Atmos. Environ., 76, 208–220, https://doi.org/10.1016/j.atmosenv.2012.09.051, 2013.

Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K., and Tsao, C.-C.: Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang, PLOS ONE, 12, e0179763, https://doi.org/10.1371/journal.pone.0179763, 2017.

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., and Ma, W. J.: The short-term effect of ambient ozone on mortality is modified by temperature in Guangzhou, China, Atmos. Environ., 76, 59–67, https://doi.org/10.1016/j.atmosenv.2012.07.011, 2013.

Lu, R. and Turco, R. P.: Ozone distributions over the los angeles basin: Three-dimensional simulations with the smog model, Atmos. Environ., 30, 4155–4176, https://doi.org/10.1016/1352-2310(96)00153-7, 1996.

Lu, X., Zhang, L., and Shen, L.: Meteorology and Climate Influences on Tropospheric Ozone: a Review of Natural Sources, Chemistry, and Transport Patterns, Current Pollut. Rep., 5, 238–260, https://doi.org/10.1007/s40726-019-00118-3, 2019.

Luna, A. S., Paredes, M. L. L., De Oliveira, G. C. G., and Corrêa, S. M.: Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil, Atmos. Environ., 98, 98–104, https://doi.org/10.1016/j.atmosenv.2014.08.060, 2014.

Ma, Y., Ma, B., Jiao, H., Zhang, Y., Xin, J., and Yu, Z.: An analysis of the effects of weather and air pollution on tropospheric ozone using a generalized additive model in Western China: Lanzhou, Gansu, Atmos. Environ., 224, 117342, https://doi.org/10.1016/j.atmosenv.2020.117342, 2020.

Mahmud, A., Hixson, M., Hu, J., Zhao, Z., Chen, S.-H., and Kleeman, M. J.: Climate impact on airborne particulate matter concentrations in California using seven year analysis periods, Atmos. Chem. Phys., 10, 11097–11114, https://doi.org/10.5194/acp-10-11097-2010, 2010.

McClure, C. D. and Jaffe, D. A.: Investigation of high ozone events due to wildfire smoke in an urban area, Atmos. Environ., 194, 146–157, https://doi.org/10.1016/j.atmosenv.2018.09.021, 2018.

McGlynn, D., Mao, H., Sive, B., and Sharac, T.: Understanding Long-Term Variations in Surface Ozone in United States (U.S.) National Parks, Atmosphere, 9, 125, https://doi.org/10.3390/atmos9040125, 2018.

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B. E., and Houston, T. G.: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3, https://doi.org/10.7289/V5D21VHZ, 2012a.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, J. Atmos. Ocean. Tech., 29, 897–910, https://doi.org/10.1175/jtech-d-11-00103.1, 2012b.

Milborrow, S.: Regression Splines, FIM Forecast Model: 500 mb Wind Speed and 500 mb Height Contours – Real-time: https://CRAN.R-project.org/package=earth (last access: 13 November 2021), R program [code], 2021.

NOAA: FIM Forecast Model: 500 mb Wind Speed and 500 mb Height Contours – Real-time, https://sos.noaa.gov/datasets/fim-forecast-model-500mb-wind-speed-and-500mb-height-contours-real-time/ (last access: 23 May 2021), 2020.

Oduro, S. D., Metia, S., Duc, H., Hong, G., and Ha, Q. P.: Multivariate adaptive regression splines models for vehicular emission prediction, Visualization in Engineering, 3, 13, https://doi.org/10.1186/s40327-015-0024-4, 2015.

Oman, L. D., Ziemke, J. R., Douglass, A. R., Waugh, D. W., Lang, C., Rodriguez, J. M., and Nielsen, J. E.: The response of tropical tropospheric ozone to ENSO, Geophys. Res. Lett., 38, 13706, https://doi.org/10.1029/2011GL047865, 2011.

Oman, L. D., Douglass, A. R., Ziemke, J. R., Rodriguez, J. M., Waugh, D. W., and Nielsen, J. E.: The ozone response to ENSO in Aura satellite measurements and a chemistry-climate simulation, J. Geophys. Res.-Atmos., 118, 965–976, https://doi.org/10.1029/2012jd018546, 2013.

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., and Tapper, N. J.: Quantifying the influence of local meteorology on air quality using generalized additive models, Atmos. Environ., 45, 1328–1336, https://doi.org/10.1016/j.atmosenv.2010.11.051, 2011.

Pernak, R., Alvarado, M., Lonsdale, C., Mountain, M., Hegarty, J., and Nehrkorn, T.: Forecasting Surface O_3 in Texas Urban Areas Using Random Forest and Generalized Additive Models, Aerosol Air Qual. Res., 9, 2815–2826, https://doi.org/10.4209/aaqr.2018.12.0464, 2019.

Pope, P. T. and Webster, J. T.: The Use of an F-Statistic in Stepwise Regression Procedures, Technometrics, 14, 327–340, 1972.

Porter, T. M.: Social Interests and Statistical Theory: Statistics in Britain, 1865–1930, Science, 214, 784–784, https://doi.org/10.1126/science.214.4522.784.a, 1981.

Porter, T. M.: Trust in Numbers, Princeton University Press, ISBN 9780691208411, 1995.

Rao, S. T., Zurbenko, I. G., Neagu, R., Porter, P. S., Ku, J. Y., and Henry, R. F.: Space and Time Scales in Ambient Ozone Data, B. Am. Meteorol. Soc., 78, 2153–2166, https://doi.org/10.1175/1520-0477(1997)078<2153:satsia>2.0.co;2, 1997.

Rodríguez-Pérez, R., Vogt, M., and Bajorath, J.: Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction, ACS Omega, 2, 6371–6379, https://doi.org/10.1021/acsomega.7b01079, 2017.

Roy, S. S., Pratyush, C., and Barna, C.: Predicting Ozone Layer Concentration Using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree, Springer International Publishing, 140–152, https://doi.org/10.1007/978-3-319-62524-9_11, 2018.

Rybarczyk, Y. and Zalakeviciute, R.: Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review, Appl. Sci., 8, 2570, https://doi.org/10.3390/app8122570, 2018.

Schölkopf, B. and Smola, A. J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, ISBN 9780262536578, 2001.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, Wiley, ISBN 978-1-118-94740-1, 2016.

Smola, A. J. and Schölkopf, B.: A tutorial on support vector regression, Stat. Comput., 14, 199–222, https://doi.org/10.1023/b:stco.0000035301.49549.88, 2004.

Sotomayor-Olmedo, A., Aceves-Fernández, M. A., Gorrostieta-Hurtado, E., Pedraza-Ortega, C., Ramos-Arreguín, J. M., and Vargas-Soto, J. E.: Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach, International Journal of Intelligence Science, 03, 126–135, https://doi.org/10.4236/ijis.2013.33014, 2013.

Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., De Hoogh, K., Kloog, I., Davoli, M., Michelozzi, P., and Bellander, T.: A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden, Atmosphere, 11, 239, https://doi.org/10.3390/atmos11030239, 2020.

Stephen, M. S.: Gauss and the Invention of Least Squares, Ann. Stat., 9, 465–474, https://doi.org/10.1214/aos/1176345451, 1981.

Sun, L., Xue, L., Wang, Y., Li, L., Lin, J., Ni, R., Yan, Y., Chen, L., Li, J., Zhang, Q., and Wang, W.: Impacts of meteorology and emissions on summertime surface ozone increases over central eastern China between 2003 and 2015, Atmos. Chem. Phys., 19, 1455–1469, https://doi.org/10.5194/acp-19-1455-2019, 2019.

Tin Kam, H.: Random decision forests, Proceedings of 3rd International Conference on Document Analysis and Recognition, 14–16 August 1995, 271, 278–282, https://doi.org/10.1109/ICDAR.1995.598994, 1995.

U.S. EPA: Trends in Ozone Adjusted for Weather Conditions, https://www.epa.gov/air-trends/trends-ozone-adjusted-weather-conditions (last access: 13 November 2021), U.S. EPA, 2016.

U.S. EPA: Integrated Science Assessment (ISA) for Ozone and Related Photochemical Oxidants (Final Report, Apr 2020), U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-20/012, 2020.

Vong, C.-M., Ip, W.-F., Wong, P.-K., and Yang, J.-Y.: Short-Term Prediction of Air Pollution in Macau Using Support Vector Machines, J. Control Sci. Eng., 2012, 1–11, https://doi.org/10.1155/2012/518032, 2012.

Wells, B., Dolwick, P., Eder, B., Evangelista, M., Foley, K., Mannshardt, E., Misenis, C., and Weishampel, A.: Improved estimation of trends in U.S. ozone concentrations adjusted for interannual variability in meteorological conditions, Atmos. Environ., 248, 118234, https://doi.org/10.1016/j.atmosenv.2021.118234, 2021.

Wikipedia Contributors: Multivariate adaptive regression spline, https://en.wikipedia.org/w/index.php?title=Multivariate_adaptive_regression_spline&oldid=1083057440, last access: 19 April 2022.

Wood, S. N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, J. Roy. Statist. Soc. B, 73, 3–36, https://doi.org/10.1111/j.1467-9868.2010.00749.x, 2011.

Wood, S. N.: Generalized Additive Models: An Introduction with R, 2nd edn., Chapman and Hall/CRC, ISBN 9781315370279, https://doi.org/10.1201/9781315370279, 2017.

Xu, L., Yu, J.-Y., Schnell, J. L., and Prather, M. J.: The Seasonality and Geographic Dependence of ENSO Impacts on US Surface Ozone Variability, Geophys. Res. Lett., 44, 3420–3428, https://doi.org/10.1002/2017gl073044, 2017.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., and Di, B.: Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment, Environ. Pollut., 233, 464–473, https://doi.org/10.1016/j.envpol.2017.10.029, 2018.