Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.210077

Prediction of High-ozone Events Using GAM, SMOTE, and Tail Dependence Approaches in Texas (2005–2019)

Benjamin Brown-Steiner[1*], Xiong Zhou[1,2], Matthew J. Alvarado[1], Brook T. Russell[3]

[1] Atmospheric and Environmental Research (AER), Lexington, MA 02421, USA
[2] Verisk Insurance Solutions, Buffalo Grove, IL 60089, USA
[3] School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, 29634, USA

ABSTRACT

We test three methods for ozone prediction in the El Paso (ELP) and Houston-Galveston-Brazoria (HGB) regions of Texas from 2005–2019: (1) a Generalized Additive Model (GAMs) approach; (2) a GAM approach with the addition of the Synthetic Minority Over-sampling TEchnique (SMOTE) and (3) a tail dependence modeling approach based in extreme value theory (EVT). We also compare the feature selection capabilities of the tail dependence approach to other feature selection methods. We find that the GAM+SMOTE model outperformed the GAM-only model when predicting ozone values for the root mean square error metric, particularly with regard to the above-threshold ozone values, which may be of particularly useful for extreme ozone event prediction. In addition, we find that the improvement of above-threshold MDA8 O_3 prediction for the GAM+SMOTE method tends to come at the cost of below-threshold prediction, which is particularly important if MDA8 O_3 trends are of interest. We also find that the tail dependence approach is capable of predicting extreme ozone events, but algorithmic stability and configuration complexity can make this approach difficult to operationalize on a broad scale and that the selection of the threshold needs to be carefully considered. Finally, the feature selection via the tail dependence method performs comparably to other forms of machine learning-based feature selection and we find that there are multiple parameter sets that can predict MDA8 O_3 with equal success.

Keywords: GAM, SMOTE, Tail dependence, Ozone prediction, Feature selection

Received: April 3, 2021
Revised: July 20, 2021
Accepted: July 26, 2021

* **Corresponding Author:**
bbrownst@aer.com

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Surface ozone concentrations and their related detrimental health effects (Brunekreef *et al.*, 2002) have been decreasing throughout the United States (U.S.) (Cooper *et al.*, 2012; Fleming *et al.*, 2018) due primarily to the reduction of ozone precursors including nitrogen oxides (NO_x = NO + NO_2) and carbon monoxide (CO) (Granier *et al.*, 2011). These decreasing ozone trends are especially evident in remote or rural regions, while weaker or negligible ozone trends are seen in many urban regions and in the western U.S. due to decreasing emissions of ozone precursors and the growing relative importance of the transport of ozone precursors from Asian countries (Brown-Steiner *et al.*, 2011; Langford *et al.*, 2017; Fleming *et al.*, 2018). While local reductions in ozone precursor emissions are the primary means of reducing local ozone concentrations (Cooper *et al.*, 2012), there are many non-local and thus uncontrollable chemical and meteorological factors that can increase ozone concentrations including upwind sources of ozone precursors such as biomass burning, lightning NO_x emissions, or stratospheric ozone intrusions (Jaffe *et al.*, 2018).

Tropospheric ozone chemistry is notorious for its non-linear dependence on concentrations of NO_x and volatile organic compounds (VOCs) (Lin *et al.*, 1988), which makes controlling high ozone events a challenge. The Environmental Protection Agency (EPA) has prioritized reductions in high

ozone events, and thus has established a framework for the classification of specific extreme ozone events as "exceptional events" that may be exempted from a state's requirement to meet National Ambient Air Quality Standards (NAAQS) (Jaffe *et al.*, 2018; U.S. EPA, 2016a, b). Subsequently, there exists a large body of scientific research focused on characterizing, quantifying, and determining the causes of extreme ozone events (Fleming *et al.*, 2018; Brown-Steiner *et al.*, 2018; Moghani *et al.*, 2018).

Due to the high computational cost and complexity of three-dimensional tropospheric chemistry simulations using chemical tracer models (CTMs), statistical models of ozone and tropospheric chemistry are frequently used to estimate ozone concentrations and trends. These models include traditional statistical techniques such as linear regression models (Shen and Mickley, 2017), land use regression models (Wang *et al.*, 2020), as well as machine learning techniques (Watson *et al.*, 2019). Non-Gaussian statistical representations of ozone distributions have also been successfully used to characterize and understand ozone chemistry including extremal dependence (Phalitnonkiat *et al.*, 2018), generalized extreme value techniques (Quintela-del-Ri'o and Francisco-Fernández, 2011), the generalized Pareto distribution (Rieder *et al.*, 2013), and tail dependence methods (Russell *et al.*, 2016). In addition, generalized additive models (GAMs), which have the capability of representing non-linear relationships between predictand variables and a target predictor variable, have been used to model ozone concentrations (Watson *et al.*, 2019; Davis and Speckman, 1999).

GAMs are a form of linear modeling which allows non-linear functions of individual predictors within a regression framework (Wood, 2017) and are increasingly being used within the atmospheric sciences (Alvarado *et al.*, 2017; Gong *et al.*, 2017). This is similar to standard linear regression techniques, which optimize scalar coefficients (α_k) for each predictor (x_k) for k = 1, ..., p:

$$\hat{y} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p \tag{1}$$

except that the coefficients are replaced with functions that can take on a variety of linear and non-linear smooth functions:

$$\hat{y} = s_0 + s_1(x_1) + s_2(x_2) + \ldots + s_p(x_p) \tag{2}$$

The GAM approach optimizes these smooth functions and allows for predictor-by-predictor variation in the number of degrees of freedom, which can be used to gain additional understanding of the relationship between the predictors and the predictand.

Oversampling and undersampling techniques are widely utilized to balance datasets that have an unequal number of samples above and below a given threshold, especially when the variable of interest is disproportionately represented within either the above- or below-threshold dataset. While these techniques artificially redistribute data distributions, they have been shown to enhance the ability to statistically predict sparsely represented extreme events such as the rapid intensification of tropical cyclones (Yang *et al.*, 2020) and large wildfires (Pérez-Porras *et al.*, 2021). This is particularly useful when the extreme events are of regulatory importance, such as extreme ozone events, where available observations of the event of interest are sparse. Typically, these techniques divide the dataset into an above-threshold and a below-threshold dataset and then add or remove datapoints until there are roughly an equal number of above- and below-threshold data points. The disadvantage of over-sampling, especially for datasets with a small number of observations, is that the resulting dataset does not fill in the full distribution of the target population, but rather uses the existing under-sampled data as representative. This could potentially skew the resultant analysis as the artificial over-sampling makes no assumptions about the underlying population or the sampling procedure.

An alternative, and the method used in this project, is the Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla *et al.*, 2002). For values under the specified threshold, the SMOTE algorithm randomly removes datapoints. For values above the specified threshold, the SMOTE algorithm inserts additional datapoints at interpolated positions between the existing datapoints. This method has advantages over other oversampling techniques as the points added by the SMOTE algorithm better approximate new datapoints which are essentially "drawn" from the same underlying distribution, and thus is more representative, of the population of datapoints being

sampled (Chawla *et al.*, 2002). While SMOTE does produce a synthetic dataset, the redistribution of the ozone data to over-emphasize the extreme ozone events, and the increase in predictive capability this redistribution produces, is worth the artificiality of the final synthetic dataset.

The tail dependence approach was developed by Russell *et al.* (2016) to "explore which combinations of meteorological conditions are associated with extreme ground level ozone conditions." This method is built within the framework provided by extreme value theory (EVT), and therefore has the advantage of being theoretically justified. Russell *et al.* (2016) use their method to attempt to identify the linear combination of meteorological covariates that achieved the highest degree of asymptotic dependence with the ozone response. Informally, two variables are termed asymptotically dependent if the probability that they are at extreme levels simultaneously is non-zero. The predictand, ozone, is treated as the response, while the set of *n* meteorological variables are treated as the covariates.

As is often done in multivariate extremes, the method of Russell *et al.* (2016) incorporates marginal transformations in order to identify the optimal set of regression parameters (details can be found in Russell *et al.* (2016)). The resulting optimization equation is similar to that used in linear regression (Eq. (1)) with one primary difference: the model optimization is performed in this transformed parameter space. Working in this transformed space has numerous computation and theoretical advantages; however, some information is lost as a result. For this reason, the optimized beta parameters should not be used directly for prediction without accounting for this potential loss of information.

This work applies these three statistical approaches to surface ozone and meteorological data within Texas and has two motivating science questions:

(1) Can a combination of GAM and SMOTE modeling improve the prediction of extreme ozone events compared to GAM-only techniques?
(2) Can the feature selection method developed by Russel *et al.* (2016) be used in combination with GAM+SMOTE to further improve the prediction of extreme ozone events?

In this work we present an analysis of 15 years (2005–2019) of surface ozone data from the El Paso (ELP) and Houston-Galveston-Brazoria (HGB) regions of Texas with a focus on fitting a GAM to the extreme ozone data using: (1) a standard GAM approach; (2) a GAM approach coupled with a SMOTE algorithm; and (3) a tail dependence approach developed initially by Russell *et al.* (2016). In the next section (Section 2) we detail the datasets used and the methods. In Section 3 we show our results, while in Section 4 we draw conclusions.

2 METHODS

2.1 Texas Ozone and Meteorological Data

The raw ozone and meteorological data used in this work comes from Texas Commission on Environmental Quality (TCEQ) surface station observations located within El Paso (ELP) and the Houston-Galveston-Brazoria (HBG) regions from January 1, 2005 to December 31, 2019. The raw data includes hourly observations of ozone and twelve meteorological variables. We use the 12 raw meteorological variables to derive 36 meteorological variables temporally aggregated at daily, morning, or afternoon levels. However, not all meteorological variables are available at all sites, so in this study for site-by-site prediction we use all derived meteorological variables which ranged from 20 to 35, while for aggregated area-level prediction, we use only 20 derived meteorological variables used in this study that are available at all sites. The maximum daily 8-hour ozone (MDA8 O_3) was calculated, and due to the persistence of surface-level ozone, we also included the previous day's MDA8 O_3 within the final predictor data set. Supplemental Table S1 summarizes these 12 hourly meteorological variables and the 36 derived variables. We additionally filtered the ELP and HGB data to include only sites that had observations that spanned the full 15 year time frame (2005–2019). This included all 6 sites for the ELP region and only 18 of the 58 sites for the HGB region. We further selected the 6 sites in the HGB region with the most complete temporal coverage, so that the final datasets used in this analysis include 6 sites from the ELP region and 6 sites from HGB region. Individual sites were labelled by concatenating their state, county, and site codes, while the parameter codes for the meteorological variables are copied from the raw datasets. For all data processing and prediction procedures in the following sections, timesteps with missing values

were dropped from the predictor data set. Finally, we subdivided the following analysis into annual data (January–December) and the ozone season (May–October).

2.2 GAM-Only Ozone Prediction

We applied a standard GAM approach for a baseline assessment of ozone prediction, using meteorological variables constant among sites, as well as the previous day's MDA8 O_3 value. Gong *et al.* (2017) successfully applied GAM to Houston and other cities using meteorological parameters at both the surface level and at 500 mb, as well as information derived from Hysplit back trajectory simulations. In this work, we include meteorological parameters form surface stations only. Here we describe an overview of our GAM formulation, while additional information and sensitivity tests are described in the Supplemental Information. For the simple ozone prediction with the GAM model, we began by including all 20 common meteorological features among the 6 sites in both ELP and HGB, plus the previous day's ozone, such that the GAM optimization function has the form:

$$E(\text{MDA8 } O_3) = f_0(\text{Previous\_MDA8}) + f_1(\text{max\_ws}) + f_2(\text{avg\_ws}) + f_3(\text{morning\_ws}) + f_4(\text{afternoon\_ws}) + f_5(\text{max\_pwg}) + f_6(\text{avg\_pwg}) + f_7(\text{max\_sdwd}) + f_8(\text{avg\_sdwd}) + f_9(\text{daily\_max\_T}) + f_{10}(\text{daily\_min\_T}) + f_{11}(\text{diurnal\_T}) + f_{12}(\text{daily\_mean\_T}) + f_{13}(\text{morning\_mean\_T}) + f_{14}(\text{afternoon\_mean\_T}) + f_{15}(\text{avg\_wind\_u}) + f_{16}(\text{avg\_wind\_v}) + f_{17}(\text{morning\_wind\_u}) + f_{18}(\text{morning\_wind\_v}) + f_{19}(\text{afternoon\_wind\_u}) + f_{20}(\text{afternoon\_wind\_v}) \quad (3)$$

where E(MDA8 O_3) represents the predictand ($\mathbf{y_t}$) for which we are fitting the GAM model. The GAM model for each region includes the six sites within each region with the most complete temporal coverage. We also conduct GAM runs for the six individual sites both for the ozone season (May–October) and the entire year (January–December). Note that we also include the previous day's ozone within the predictor variables, which is a common practice when using available observations for ozone prediction (see Oufdou *et al.* (2021) and references therein), as ozone has high day-to-day persistence. If we exclude the previous day's ozone, the performance of these models drops by roughly half (data not shown). We did not include any predictor variables with a temporal trend (such as NO_x emissions) as the MDA8 O_3 exhibited little trend over this period (for ELP, MDA8 O_3 was 44.8 ppb averaged over 2005–2007 and 45.4 ppb averaged over 2017–2019; for HGB, MDA8 O_3 did decrease from 40.7 ppb to 36.9 ppb over the same averaging periods, which we did not feel was a strong enough trend to make any adjustments. Plots of the full time series for ELP and HGB, along with plots of residuals, can be seen in the Supplemental Figs. S3, S4, and S5. The *gam* function in R (via the *mgcv* package (Wood, 2019)) was used for this project and details of the parameter selection and sensitivity testing can be found in the Supplemental Information. We randomly distributed the data for all time periods into either the training (70%) or testing (30%) datasets, both for the annual (January–December) and O_3 season (May–October) datasets.

2.3 GAM+SMOTE Ozone Prediction

The combination of GAM regression and SMOTE dataset redistribution takes the dataset from the GAM-only modeling and selects a threshold value to separate the above- and below-threshold MDA8 O_3 data samples. The SMOTE algorithm is then applied to the training and testing data such that there is a roughly equal number of above- and below-threshold datasets. The example in Fig. 1 applies the SMOTE algorithm with a 60 ppb threshold to a two-dimensional (ozone and temperature) dataset, but SMOTE can be similarly applied to higher dimensional datasets, such as the multidimensional meteorological dataset in this work. The oversampling of the underrepresented event (here, extreme ozone events) occurs by linearly interpolating new points between existing points within the higher dimensional data distribution. Table 1 compares the absolute (70 ppb) theshold to relative (90^{th} percentile) thresholds for our data. The selection of a threshold of 70 ppb, which matches the ozone National Ambient Air Quality Standard (NAAQS) and thus is a natural choice for a threshold for all sites, results in a very skewed distribution of data with few above-threshold samples. Meanwhile, a 90^{th} percentile threshold, which may result in a less skewed distribution for each site, can result in difficulties when comparing sites across regions. Further

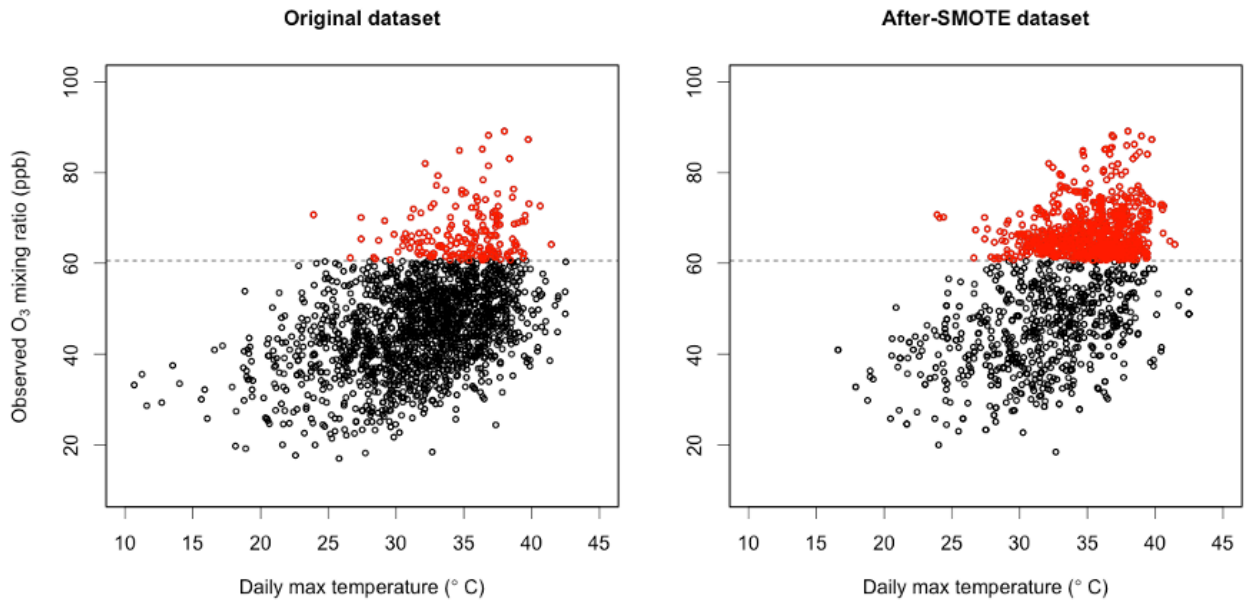

Fig. 1. A comparison of the original (left) and post-SMOTE (right) ozone season (May–October, 2005–2019) MDA8 O_3 distribution versus daily maximum temperature for the ELP region. Red points are in excess of a 60 ppb threshold.

Table 1. A comparison of the above- and below-threshold distribution of MDA8 O_3 values for the ozone season (May–October) for both the ELP and HGB regions using an absolute (70 ppb) and relative (90th percentile) threshold selection for both the GAM-Only and GAM+SMOTE approaches. Data are divided into 70% training and 30% testing datasets.

			Original		SMOTE	
			Training	Testing	Training	Testing
ELP All Sites	70 ppb	Below	10,381	4,492	1,071	121
		Above	357	110	1,071	121
	90th percentile	Below	9,662	4,144	3,228	495
		Above	1,076	458	3,228	503
HGB All Sites	70 ppb	Below	9,371	4,031	1,644	242
		Above	548	221	1,644	243
	90th percentile	Below	8,901	3,852	3,054	440
		Above	1,018	400	3,054	440

sensitivity tests can be found in the Supplemental Information. For the rest of this work we selected the 90th percentile threshold for each site and each region. Finally, a standard GAM regression is trained using the post-SMOTE datasets. Details on specific parameter selection and sensitivity tests can be found in the Supplemental Information.

2.4 Tail Dependence

The tail dependence procedure employed in this work seeks to find the linear combination of covariates that have the highest possible degree of asymptotic dependence with the ozone response (see Russell *et al.*, 2016; Fix *et al.*, 2018). Methods based in extreme value theory (EVT), such as this one, are critical to consider as they offer theoretically justified approaches for modeling the far upper tail. As the objective of Russell *et al.* (2016) was to identify the meteorological conditions that were associated with extreme ozone conditions, and not ozone prediction directly, Russell *et al.* (2016) did not consider methods for direct ozone prediction. As is commonplace in multivariate EVT, marginal transformations are made (see Russell *et al.*, 2016 for details). The transformation employed here is a *lossy transformation*. As such, we were unable to make direct ozone predictions without a loss of information. However, if we use the regression parameter estimates to make ozone predictions, understanding that there is some loss of information inherent in the procedure, ozone prediction is possible.

We divide this task into two parts. First, we use the Russell *et al.* (2016) procedure to estimate the regressions parameters that best fit the TCEQ-supplied metrological covariates to the highest MDA8 O_3 values and use these parameter estimates as a feature selection procedure (see Supplemental Table S2). We then use in the GAM+SMOTE analysis to evaluate the selected covariates prediction performance. Second, we attempt to replace the lossy rank transformation with a lossless transformation in order to made ozone predictions directly.

In order to assess a model's performance and as a means of protecting against overfitting, Russell *et al.* (2016) suggest a 10-fold cross-validation (CV) procedure. CV is a commonly utilized non-parametric procedure for assessing a model's out of sample predictive ability. Data are first randomly divided into 10 (approximately) equally sized partitions. At each step, exactly one partition is held out, and the model is fit on the remaining nine. The resulting parameter estimates are used to predict the values of the observations in the held-out partition, and the sum of the squared errors (SSE) is calculated for each of these observations. This is repeated for each partition, and the CV score is the average of the 10 resulting SSEs. When comparing two models, a higher CV score corresponds with a worse ability to model MDA8 O_3 for new data; therefore, a smaller CV score is preferred.

For analysis based on this approach, we utilize all available covariates and focus on the ozone season (May–October) for all years (2005–2019) available from the six selected sites in the ELP and HGB regions. As the tail dependence procedure consists of several steps and contains a number of decisions that could potentially impact the final results, we perform a series of sensitivity tests which we summarize in the Supplemental Information. Additionally, in the Supplemental Information, we compare predictions using the lossy rank transformation to a less lossy transformation procedure, and also compare these prediction results to the previous GAM+SMOTE process.

2.5 Metrics for Comparison

We use the R-squared value (R^2) and the root mean square error (RMSE) to compare the different time series of observations and model predictions. Two RMSE metrics are included in the following tables and analysis: one that used all available ozone data (RMSE_All) and one that only used the above threshold data above 60 ppb (RMSE_Highest). In addition, we primarily use a single summary statistic: the True Positive Rate (TPR). See Supplemental Table S1 for sensitivity tests on the 80th or 90th percentile selection for the TPR metric. The TPR quantifies the actual number of positive datapoints successfully predicted by the algorithm based on the true positives (TPs) and the false negatives (FNs), and is defined as: TPR = TP/(TP + FN) expressed as a percentage.

3 RESULTS AND DISCUSSION

3.1 GAM-Only and GAM+SMOTE Ozone Prediction Results

Fig. 2 plots the GAM-only and GAM+SMOTE predictions of MDA8 O_3 using the 20 meteorological covariates compared to the observed values in the testing datasets. For GAM-only, the regression model fit will be dominated by lower MDA8 O_3 values which results in significant biases when predicting at the higher MDA8 O_3 levels (Fig. 2). In both regions, the highest MDA8 O_3 values are systematically underpredicted by the GAM regression (indicated by digression from the one-to-one line (red)), while the lower and mid-range MDA8 O_3 values are generally well-predicted. R^2 and RMSE statistics are shown in Table 2.

In contrast, the highest MDA8 O_3 values are less underpredicted with the GAM+SMOTE model compared with GAM-only model. A summary of evaluation metric results for a GAM-only and GAM+SMOTE comparison for the six sites analyzed in both the ELP and HGB regions are in Table 2. Overall, compared to the GAM-only model the GAM+SMOTE model: (1) did not substantially change the R^2 values; (2) had varying impacts on the RMSE when all test samples are included (RMSE_All); and (3) consistently reduced the RMSE for above-threshold testing samples (RMSE_Highest) at the expense of the lowest ozone values. From this we conclude that the SMOTE procedure improves the ability of the GAM-only regression to predict the above-threshold ozone values, but this improvement comes at the expense of the below-threshold values. We find that the GAM+SMOTE method consistently outperforms the GAM-only method when looking at the

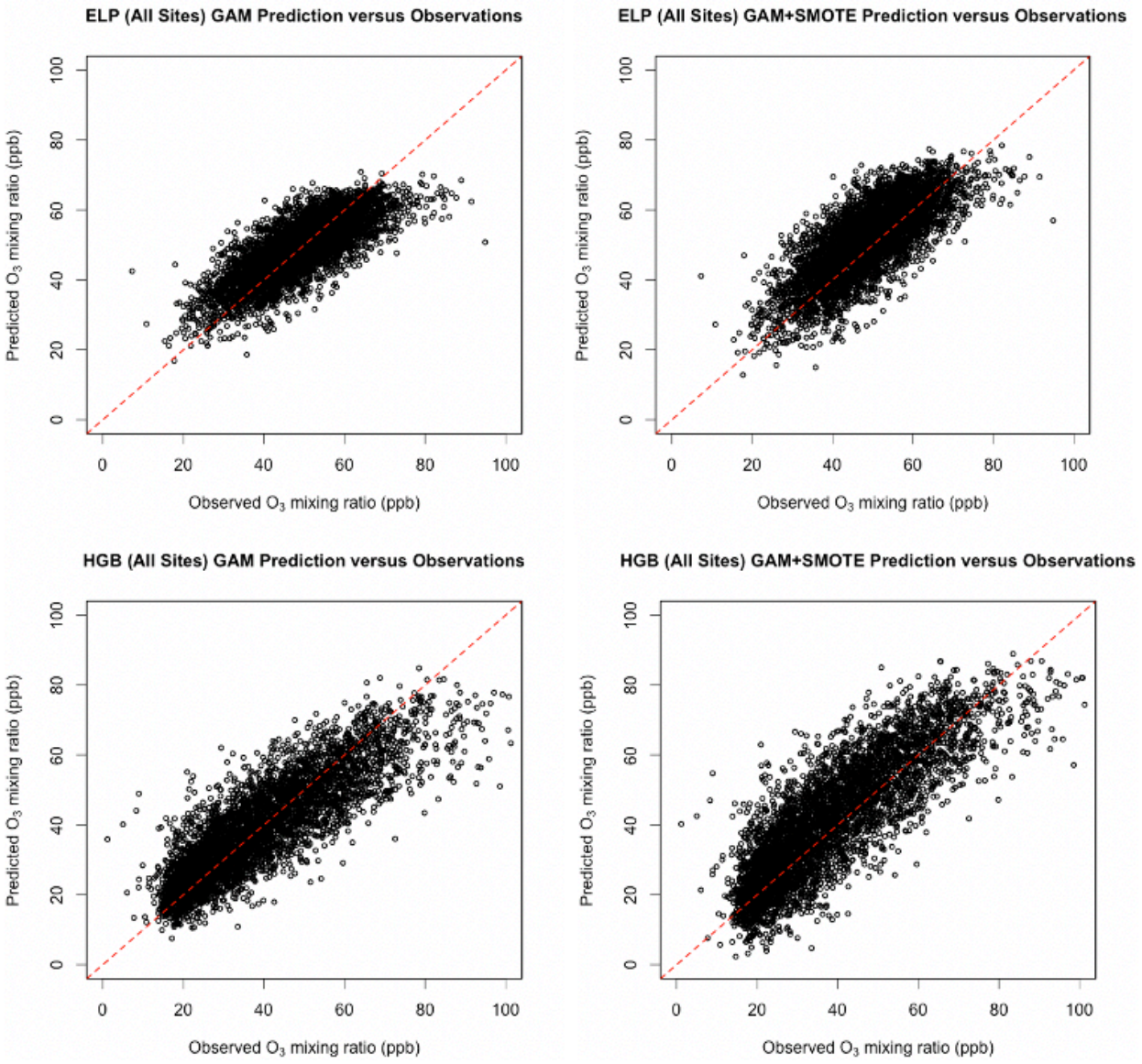

Fig. 2. Predicted versus observed MDA8 O_3 values for ELP (top) and HGB (bottom) regions using the GAM-Only (left) and GAM+SMOTE (right) approaches. The dotted red line is the 1:1 line.

RMSE values, but not the R^2 values. In addition, the GAM-only method has lower RMSE_all values than RMSE_highest values, and the GAM+SMOTE method improves both RMSE scores with the largest improvements made for the RMSE_highest scores. Finally, both models perform better using the R^2 metric over the HGB region than the ELP region, while the RMSE metric outperforms the R^2 metric over the ELP region compared to the HGB region. This is likely due to different chemical and meteorological conditions within each region.

In Fig. 3 we plot the observed MDA8 O_3 values within the ELP training datasets (black) compared to the GAM-only (blue) and GAM+SMOTE (red) regressions for above-threshold events. The GAM-only predictions consistently underpredicts the observed MDA8 O_3 values, while the GAM+SMOTE more consistently predicts MDA8 O_3 values. However, neither the GAM-only or the GAM+SMOTE regressions are capable of correctly predicting the highest MDA8 O_3 values. This is likely because the highest MDA8 O_3 values are the result of complex, non-liner meteorological and chemical conditions, which statistical models have trouble predicting.

3.2 Feature Selection Results

We test different parameter sets via different feature selection approaches as indicated in

Table 2. Summary table for the ELP and HGB sites (each individual site and the regional average) comparing GAM-Only and GAM+SMOTE R^2 and RMSE metrics. NOTE: RMSE_All includes all data points, while RMSE_Highest includes only those above the 90th percentile.

Area	Site	GAM-Only			GAM+SMOTE		
		R^2	RMSE_All	RMSE_Highest	R^2	RMSE_All	RMSE_Highest
El Paso	481410029	0.69	8.61	10.63	0.62	8.34	8.4
	481410037	0.73	8.1	9.02	0.67	8.2	7.66
	481410044	0.67	9.79	11.35	0.44	11.5	9.11
	481410055	0.72	8.75	13.48	0.60	9.17	8.27
	481410057	0.65	8.39	11.25	0.63	7.82	8.94
	481410058	0.72	8.66	10.23	0.66	7.47	9.02
	All sites	0.60	6.95	10.76	0.57	8.17	7.52
Houston	480390618	0.79	10.67	11.94	0.78	10.3	10.51
	480391004	0.80	12.66	15.11	0.81	11.1	12.33
	480391016	0.81	9.8	11.72	0.81	9.11	9.59
	480390056	0.74	12	13.8	0.76	10.52	12.7
	480390024	0.83	9.4	10.68	0.81	8.95	9.01
	480390026	0.77	10.8	11.8	0.73	11.7	10.7
	All sites	0.73	8.93	11.82	0.70	10.37	10.08

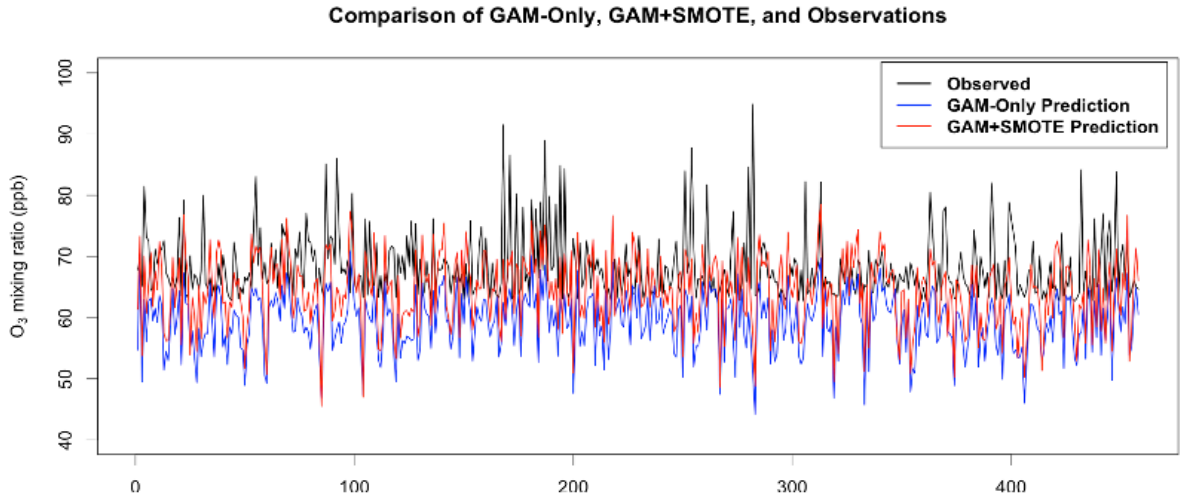

Fig. 3. Ozone season time series for the ELP region for observed (black), GAM-Only (blue), and GAM+SMOTE (red) approaches. GAM-Only and GAM+SMOTE plots compared to observations are also plotted to better see the distribution of points in comparison to the observed values.

Supplemental Table S2 including: (1) the lowest cross-validation (CV) scores, a backward selection method using the random forest increase in mean squared error (%IncMSE) metric, a completely random selection of features (7 for ELP and 6 for HGB), and a random selection of meteorological features selected within each raw meteorological sub-group. The full list of selected parameters can be found in Supplemental Table S1 and a complete description of the different feature selection approaches can be found in Supplemental Table S2. We tested the feature selection approaches for both the ELP and HGB regions using both an individual and the regional average. Supplemental Table S7 elucidates many of the differences between the regions, models, and feature selection methods. First, all the feature selection methods outperform (via higher R^2 and lower RMSE values) the fully random selection of features. Second, among the other feature selection methods, there is no individual method that clearly outperforms any of the others.

From these differences we find that, while feature selection methods outperform random feature selection, none of the methods tested here outperform the others, which indicates that there are many sets of predictand variables that perform equally well. In most cases, other than

the previous day's ozone, some derived version of wind speed, temperature, and relative humidity were used in the highest performing models. Also, it is likely that many of the observed meteorological variables are highly correlated, so including multiple variables derived from the same raw observations does not provide any additional performance in the model results. Thus, while it seems like there is no clear way of identifying a single best-performing combination of features, feature selection procedures are capable of identifying a set of top-performing but functionally equivalent models. Finally, additional variables, such as model output or observations at different altitudes, such as the variables used in the GAM modeling of Gong *et al.* (2017) is likely to improve the modeling results compared to the surface station data included in this study.

3.3 Tail Dependence Ozone Prediction Results

We found that using the tail dependence optimized beta parameters to make ozone predictions using the top-performing set of parameters (Supplemental Table S7) resulted in successful prediction of MDA8 O_3 time series for the ELP region (Fig. 4). Supplemental Table S7 compares the top performing tail dependence feature selection sets (using the CV metric) to alternative feature selection approaches. Compared to the GAM+SMOTE approach (Table 2), the tail dependence approach produced comparable summary metrics. For the full year, the tail dependence has an R^2 value of 0.54 with a RMSE_All value of 8.41 ppb and a RMSE_Highest value of 11.02. This is comparable to the GAM+SMOTE performance via the R^2 metric (R^2 = 0.62, Table 2), comparable for the RMSE_All metric (8.34 ppb, Table 2), and moderately worse than for the RMSE_Highest metric (8.40 ppb, Table 2). However, the tail prediction results can vary from site to site. For instance, Supplemental Fig. S2 plots the tail dependence prediction results for a site within the HGB region, which shows less predictive success than the example site for the ELP region (Fig. 4), likely due to data loss during the rank transformation procedure. This further highlights challenges associated with the trail dependence approach.

4 CONCLUSIONS

In this project we have shown that statistical approaches for ozone prediction can successfully forecast ozone time series, although the predictive capabilities can vary by site and season. In particular, the combination of GAM and SMOTE techniques together can be used to predict extreme ozone events, although this success comes at the expense of the non-extreme ozone prediction capabilities. Thus, for air quality applications in which exceedances of NAAQS are of primary importance, the GAM+SMOTE approach can be straightforwardly utilized to make extreme ozone predictions based on meteorological and previous-day ozone observations. In these cases, the artificial redistribution of the datasets used for prediction can be determined to be worthwhile if the result is an improved predicative capability for the extreme event in question, as we have shown here for extreme ozone events. In addition, more technical statistical approaches, such as the tail dependence approach of Russell *et al.* (2016), can also be utilized or employed to make ozone predictions. However, more care is required in the initialization, tuning, and optimization on a per-site or per-region basis to ensure algorithmic stability. In this work we found that the tail dependence method performed comparably to the GAM+SMOTE method, and thus we could not recommend the use of the tail dependence method for similar ozone extreme event prediction. Nonetheless, once a tuned and optimized tail dependence approach has been initialized for a given region or site, it may be used operationally without much need for further tuning or optimization, and future work may find cases in which the tail dependence method outperforms other statistical approaches.

We have also demonstrated that feature selection for ozone prediction is a consistent challenge with no clear optimum. The number of meteorological features that might be included in a statistical prediction approach, along with the differing forms or normalization and algorithmic parameter selection, are large. We have shown that feature selection using the tail dependence approach performs equally as well as other machine learning (ML)-based and expertise-based feature selection approaches, and that all of these approaches outperform a randomized feature selection approach. Additionally, we have shown that ozone prediction capabilities are greatly enhanced with the inclusion of the previous day's ozone concentration, which highlights the

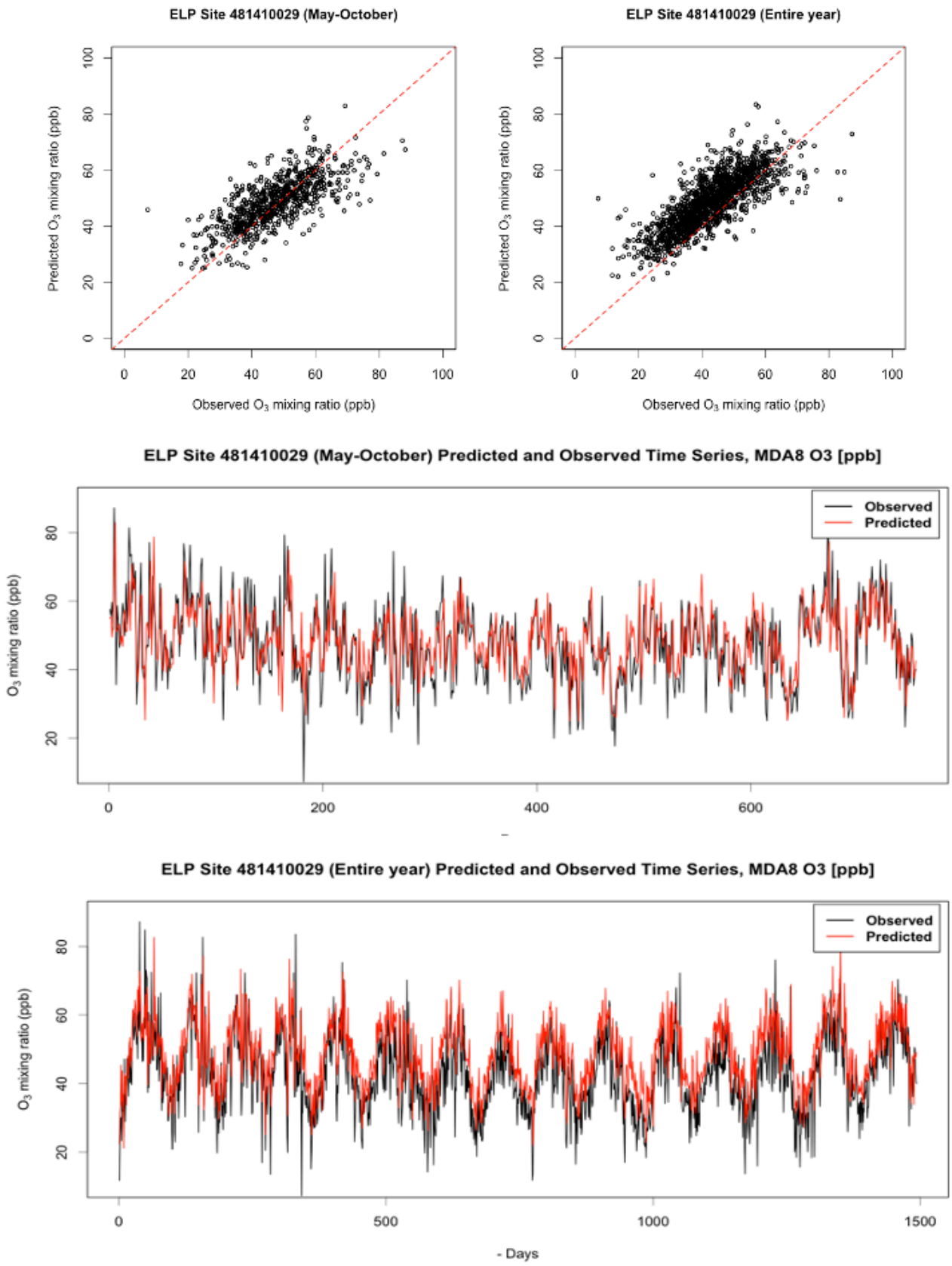

Fig. 4. Predicted versus Observed plots for one ELP site for the ozone season and the full annual cycle (top row) along with the full observed (black) and predicted (red) times series for the ozone season (middle row) and full annual time period (bottom) for the tail-dependence prediction. A plot for a site within the HGB region can be found in the Supplemental Information (Supplemental Fig. S2).

regional and persistent nature of ozone concentrations at the surface. Model performance without the inclusion of the previous day's ozone decreases the performance by roughly half. Meteorology-only prediction of ozone is unlikely to enable high-performing ozone prediction even with a large number of selected features. This also highlights the need for chemical tracer models which are capable of directly simulating the non-linear complex chemistry that ultimately controls extreme ozone events, especially when project objectives include predicting and understanding the highest ozone events. Based on our results, it seems unlikely that statistical or even ML-based approaches to ozone prediction will be capable of predicting these extreme events as well as chemical tracer modeling efforts, especially if there are trends in the chemical, meteorological, and climatological conditions over time, as ML-based approaches have trouble predicting out-of-sample cases under these changing conditions resulting from these trends.

Big datasets like the TCEQ-supplied datasets utilized in this report have a large number of observed variables, and any statistical modeling approach using large datasets needs to have a feature selection process in order to remain computationally manageable. In these cases, concerns regarding model overfitting and variable independence need to be considered. In addition, there are non-meteorological variables that can have significant impacts on MDA8 O_3 such as local sources of emission of ozone -precursor gases and large-scale transportation of ozone and ozone -precursors, none of which are included in this analysis.

We recommend that future work includes an expanded feature selection process and the addition of additional non-meteorological datasets. We found significant site-to-site and region-to-region variability in the magnitude of high-ozone events, such that a selection of an absolute threshold for determine high- and low-ozone events is not practical. Additionally, other statistical performance metrics beyond R^2 values and RSME used in this work could further quantify site-to-site and regional-scale performance. We recommend future work to characterize the site-to-site and region-to-region variability of both ozone and meteorology, as this would better enable air quality managers to understand and characterize the air quality as a heterogeneous and highly variable system.

ACKNOWLEDGMENTS

B.B.S., X.Z., and M.J.A. received support from the Texas Commission on Environmental Quality (TCEQ) Contract No. 582-19-90498, Work Order No. 582-20-12403-005.

DISCLAIMER

A portion of the data or other information in this document was prepared under a contract from the Texas Commission on Environmental Quality (TCEQ). The contents of this document do not necessarily reflect the views and policies of the TCEQ, nor does the TCEQ endorse any trade names or recommend the use of any commercial products mentioned in this document.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.210077

REFERENCES

Alvarado, M., Londsdale, C., Mountain, M., Hegarty, J. (2017). Investigating the Impact of Meteorology on O_3 and $PM_{2.5}$ Trends, Background Levels, and NAAQS Exceedances. Contract Report TCEQ Work Order No. 582-15-54118-01. Atmospheric and Environmental Research (AER) Inc. https://www.tceq.texas.gov/assets/public/implementation/air/am/contracts/reports/da/5821554118FY1501-20150831-aer-MeteorologyAndO3PMTrends.pdf

Brown-Steiner, B., Hess, P. (2011). Asian influence on surface ozone in the United States: A comparison of chemistry, seasonality, and transport mechanisms. J. Geophys. Res. 116,

D17309. https://doi.org/10.1029/2011JD015846
Brown-Steiner, B., Selin, N. E, Prinn, R. G., Monier, E., Tilmes, S., Emmons, L., Garcia-Menendez, F. (2018). Maximizing ozone signals among chemical, meteorological, and climatological variability. Atmos. Chem. Phys. 18, 8373–8388. https://doi.org/10.5194/acp-18-8373-2018
Brunekreef, B., Holgate, S.T. (2002). Air pollution and health. Lancet 360, 1233–1242. https://doi.org/10.1016/s0140-6736(02)11274-8
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953
Cooper, O.R., Gao, R.S., Tarasic, D., Leblanc, T., Sweeney, C. (2012). Long-term ozone trends at rural ozone monitoring sites across the United States, 1990–2010. J. Geophys. Res. 117, D22307. https://doi.org/10.1029/2012JD018261
Davis, J.M., Speckman, P. (1999). A model for predicting maximum and 8 h average ozone in Houston. Atmos. Environ. 33, 2487–2500. https://doi.org/10.1016/S1352-2310(98)00320-3
Fix, M.J., Cooley, D., Hodzic, A., Gilleland, E., Russell, B.T., Porter, W.C., Pfister, G.G. (2018). Observed and predicted sensitivities of extreme surface ozone to meteorological drivers in three US cities. Atmos. Environ. 176, 292–300. https://doi.org/10.1016/j.atmosenv.2017.12.036
Fleming, Z.L., Doherty, R.M., von Schneidemesser, E., Malley, C.S., Cooper, O.R., Pinto, J.P., Colette, A., Xu, X., Simpson, D., Schultz, M.G., Lefohn, A.S., Hamad, S., Moolla, R., Solberg, S., Feng, Z. (2018). Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health. Elem. Sci. Anth. 6, 12. https://doi.org/10.1525/elementa.273
Gong, X., Kaulfus, A., Nair, U., Jaffe, D.A. (2017). Quantifying O_3 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
Granier, C., Bessagnet, B., Bond, T., D'Angiola, A., Denier van der Gon, H., Frost, G.J., Heil, A., Kaiser, J.W., Kinne, S., Klimont, Z., Kloster, S., Lamarque, J.F., Liousse, C., Masui, T., Meleux, F., Mieville, A., Ohara, T., Raut, J.C., Riahi, K., Schultz, M.G., *et al.* (2011). Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. Clim. Change 109, 163–190. https://doi.org/10.1007/s10584-011-0154-1
Jaffe, D.A., Cooper, O.R., Fiore, A.M., Henderson, B.H., Tonnesen, G.S., Russell, A.G., Henze, D.K., Langford, A.O., Lin, M., Moore, T., Helmig, D., Goldstein, A. (2018). Scientific assessment of background ozone over the U.S.: Implications for air quality management. Elem. Sci. Anth. 6, 56. https://doi.org/10.1525/elementa.309
Langford, A.O., Alvarez, R.J., Brioude, J., Fine, R., Gustin, M.S., Lin, M.Y., Marchbanks, R.D., Pierce, R.B., Sandberg, S.P., Senff, C.J., Weickmann, A.M., Williams, E.J. (2017). Entrainment of stratospheric air and Asian pollution by the convective boundary layer in the southwestern U.S.: Entrainment of Stratospheric Air. J. Geophys. Res. 122, 1312–1337. https://doi.org/10.1002/2016JD025987
Lin, X, Trainer, M., Liu, S.C. (1988). On the nonlinearity of tropospheric ozone. J. Geophys. Res. 93, 15879–15888. https://doi.org/10.1029/JD093iD12p15879
Moghani, M., Archer, C.L., Mirzakhalili, A. (2018). The importance of transport to ozone pollution in the U.S. Mid-Atlantic. Atmos. Environ. 191, 420–431. https://doi.org/10.1016/j.atmosenv.2018.08.005
Oufdou, H., Bellanger, L., Bergam, A., Khomsi, K. (2021). Forecasting daily of surface ozone concentration in the Grand Casablanca region using parametric and nonparametric statistical models. Atmosphere, 12, 666. https://doi.org/10.3390/atmos12060666
Pérez-Porras, F.J., Triviño-Tarradas, P., Cima-Rodríguez, C., Meroño-de-Larriva, J.E., García-Ferrer, A., Mesas-Carrascosa, F.J. (2021). Machine learning methods and synthetic data generation to predict large wildfires. Sensors 21, 3694. https://doi.org/10.3390/s21113694
Phalitnonkiat, P., Hess, P.G.M., Grigoriu, M.D., Samorodnitsky, G., Sun, W., Beaudry, E., Tilmes, S., Deushi, M., Josse, B., Plummer, D., Sudo, K. (2018). Extremal dependence between temperature and ozone over the continental US, Atmos. Chem. Phys. 18, 11927–11948. https://doi.org/10.5194/acp-18-11927-2018
Quintela-del-Ri'o, A., Francisco-Fernández, M. (2011). Nonparametric functional data estimation applied to ozone data: Prediction and extreme value analysis. Chemosphere 82, 800–808. https://doi.org/10.1016/j.chemosphere.2010.11.025
Rieder, H.E., Fiore, A.M., Polvani, L.M., Lamarque, J.F., Fang, Y. (2013). Changes in the frequency

and return level of high ozone pollution events over the eastern United States following emission controls. Environ. Res. Lett. 8, 014012. https://doi.org/10.1088/1748-9326/8/1/014012

Russell, B.T., Cooley, D.S., Porter, W.C., Reich, B.J., Heald, C.L. (2016). Data mining to investigate the meteorological drivers for extreme ground level ozone events. Ann. Appl. Stat. 10, 1673–1698. https://doi.org/10.1214/16-AOAS954

Shen, L., Mickley, L.J. (2017). Seasonal prediction of US summertime ozone using statistical analysis of large scale climate patterns. Proc. Natl. Acad. Sci. U.S.A. 114, 2491. https://doi.org/10.1073/pnas.1610708114

U.S. Environmental Protection Agency (U.S. EPA). (2016a) Guidance on the Preparation of Exceptional Events Demonstrations for Wildfire Events that May Influence Ozone Concentrations. Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards. https://www.epa.gov/sites/production/files/2016-09/documents/exceptional_events_guidance_9-16-16_final.pdf (accessed 8 December 8 2020).

U.S. Environmental Protection Agency (U.S. EPA). (2016b). Treatment of Data Influenced by Exceptional Events. 81 FR, 68216, October 3, 2016. https://www.epa.gov/sites/production/files/2018-10/documents/exceptional_events_rule_revisions_2060-as02_final.pdf (accessed 8 December 8 2020).

Wang, J., Cohan, D.S., Xu, H. (2020). Spatiotemporal ozone pollution LUR models: Suitable statistical algorithms and time scales for a megacity scale. Atmos. Environ. 237, 117671. https://doi.org/10.1016/j.atmosenv.2020.117671

Watson, G.L., Telesca, D., Reid, C.E., Pfister, G.G., Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events, Environ. Pollut. 254, 112792–112792. https://doi.org/10.1016/j.envpol.2019.06.088

Wood, S. (2017). Generalized Additive Models: An Introduction with R, 2nd Edition. Chapman and Hall/CRC.

Wood, S. (2019). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Version 1.8-31. Package for the R statistical computing language. https://cran.r-project.org/package=mgcv

Yang, Q., Lee, C.Y., Tippett, M.K. (2020). A long short-term memory model for global rapid intensification prediction. Weather Forecasting 35, 4, 1203–1220. https://doi.org/10.1175/WAF-D-19-0199.1