Forecasting Surface O_3 in Texas Urban Areas Using Random Forest and Generalized Additive Models

Rick Pernak[1 a], Matthew Alvarado[1], Chantelle Lonsdale[1], Marikate Mountain[1], Jennifer Hegarty[1], and Thomas Nehrkorn[1]

[1] *Atmospheric and Environmental Research, 131 Hartwell Avenue, Lexington, MA 02421.*

Abstract

We developed and evaluated three types of statistical forecast models (quantitative, probabilistic, and classification) for maximum daily 8-hour average of ozone based on meteorological and ozone monitoring data for six Texas urban areas from 2009-2015. The quantitative and probabilistic forecast models were Generalized Additive Models (GAMs), while the classification forecast used the random forest machine learning method. We found that for the quantitative forecast models, five of the eight predictors (day-of-week, day-of-year, water vapor density, wind speed, and previous day ozone measurement) were significant at the $\alpha = 0.001$ level for all urban areas, while the other three had significance that varied with location. The quantitative forecast results for the 2016 ozone season agreed well with associated measurements (R^2 of ~0.70) but tended to under-predict on the days with the highest ozone concentrations. In contrast, the probabilistic forecasting models showed little skill in determining the probability of ozone exceeding policy-relevant thresholds during the 2016 ozone season. The success rate for the random forest classification models typically exceeds 75% and likely would be higher if the training data sets contained more extreme events.

Keywords: Ozone MDA8; Ozone Prediction; Generalized Additive Models; Random Forest

[a] Corresponding author: rpernak@aer.com, 781-761-2317 (phone), 781-761-2299 (fax)

INTRODUCTION

Ozone (O_3) at the surface can have adverse effects on public health. Lippmann (1989) summarized many studies that suggest lungs in people age quicker, lung capacity diminishes, and air flow resistance increases with sustained exposure to ozone. Devlin *et al.* (1997) examined lung inflammation and changes in lung function and found that these afflictions are likely due to elevated ozone and that those with pre-existing respiratory conditions are more prone to experience them. Studies performed on data from the United States, Canada, and Europe have linked air pollution to chronic obstructive pulmonary disease (COPD), increased hospitalization for respiratory illnesses (e.g., asthma), and lower forced expiratory volume (FEV). Anderson *et al.* (1997) confirmed an association with ozone and other pollutants with COPD in six European cities with different climates. Sixteen Canadian cities are the subject of Burnett *et al.* (1997), and they found that, in a population of over 12 million people over more than ten years, even low levels of surface O_3 can lead to an increase in hospitalizations due to respiratory diseases. FEV was found by Brown *et al.* (2008) to decrease a statistically significant amount among young healthy adults after long exposures to ozone in the 60-80 ppb range. Finally, Bell *et al.* (2004) studied data that spanned over a decade for 95 US cities (and 40% of the population) and show that mortality rates increase a small but statistically significant

amount when O_3 levels are raised by just 10 ppb. The EPA provides many more references on the effects of ozone on public health on their website[b] and in the Integrated Science Assessment (2013).

Given the harmful side effects of ozone inhalation, it is imperative that people know when to expect poor air quality days due to elevated ozone and what do in such cases. State and local agencies across the country in conjunction with the EPA and AirNow recognize "Action Days"[c] when ozone and particulate matter concentrations are expected to be high. AirNow provides suggestions for residents of these areas to reduce the amount of pollutants in the air on these Action Days[d], and the EPA provides services like the Air Quality Alert Program so that individuals who want poor air quality alerts can be emailed or notified on their phone regarding a poor air quality day[e]. State and local governments can also issue alerts via radio and television outlets so that individuals can decide to stay in their homes and reduce exposure to pollution. These kinds of precautions require accurate forecasting of air quality, so it is advantageous to have an ability to predict tropospheric O_3

[b] https://www.epa.gov/ozone-pollution-and-your-patients-health/references-ozone-and-your-patients-health

[c] https://www.airnow.gov/index.cfm?action=airnow.actiondays

[d] https://www.airnow.gov/index.cfm?action=resources.whatyoucando

[e] https://www3.epa.gov/region1/airquality/smogalrt.html

54 concentrations so that sufficient precautions can be implemented. O_3 production is dependent on
55 meteorological and photochemical conditions, so knowledge of both is necessary for accurate
56 predictions.

57 Numerical models exist for this purpose, but statistical modeling techniques have been
58 attempted to forecast surface O_3 as well. For example, Thompson *et al.* (2001) reviewed statistical
59 methods for modeling the dependence of surface level O_3 on meteorology and emissions to obtain air
60 quality forecasts, estimate O_3 time trends, and increase understanding of the underlying
61 photochemistry that is responsible for the generation of ozone. These statistical methods were
62 classified as regression (linear, tree-based, non-linear), extreme value, and spatio-temporal modeling,
63 none of which were found to be more appropriate than the others. Thompson *et al.* (2001) suggest that
64 the best type of model to use is dependent on the region being considered and the machine learning
65 technique employed.

66 Most of the literature focuses on forecast model application to urban areas, with perhaps a
67 similar application to rural areas as a baseline. For example, Feister & Balzer (1991) developed
68 nonlinear forecasting models for five (mostly non-urban and non-industrial) stations in Germany using
69 over 300 predictors and found that recent O_3 surface levels and solar irradiance were the most
70 significant but explained no more than 46% of the variance. Camalier *et al.* (2007) determined the

71 effect of meteorology on O_3 in 39 urban areas in the eastern United States using a generalized linear
72 model and found that ozone concentration increases with increasing temperature and decreases with
73 increasing relative humidity. They also show that the temperature dependence is more pronounced in
74 the cities in their study at higher latitudes while humidity is the more significant predictor at lower
75 latitudes. Kgabi & Sehloho (2012) investigated the effect of both emissions and meteorology by
76 monitoring surface O_3 in a rural area (Botsalano Game Reserve near the Botswana-South Africa border)
77 and an industrial area (Marikana, South Africa) in Africa and measuring associated temperature,
78 relative humidity, wind speed, and wind direction. They observed consistently higher ozone
79 concentrations at the industrial location, a negative correlation between O_3 and humidity, and a positive
80 correlation between temperature and O_3. More recently, statistical models have been used to identify
81 unusual conditions, such as wildfires, that may have contributed to ozone formation. For example,
82 Gong *et al.* 2017 developed a Generalized Additive Model (GAM) approach to estimate O_3 in several
83 cities, and then showed that days with measurable fire smoke impact tended to have positive residuals,
84 suggesting the fires where contributing to O3 formation on these days.
85 In addition to the GAM approach, other statistical forecasts for surface O_3 have used artificial
86 neural networks, or ANNs (Luna *et al.*, 2014 and Abdul-Wahab & Al-Alawi, 2002) and random forests.
87 ANNs are non-linear in nature, and the relationships that are discovered are deduced with limited prior

88 knowledge of a given process. They are thus helpful in forecasting ozone for particular locations and
89 provide the ability to model the change in O_3 concentrations due to meteorology and photochemical
90 processes together (by including concentrations of pollutants, specifically primary pollutants like NO_x
91 and NMHC, as predictors in the model). Abdul-Wahab & Al-Alawi (2002) implemented this procedure
92 for Kuwait, while Luna *et al.* (2014) applied it to Rio de Janeiro in Brazil. Both studies produced ozone
93 predictions that yielded > 90% correlation with measurements in training and testing, thus providing
94 models that account for nearly all of the variation in the observations. Abdul-Wahab & Al-Alawi (2002)
95 found that meteorology explains up to 41% of the variance in their models, while variability in the
96 emissions accounted for the rest of the variability in O_3 concentrations. Luna *et al.* (2014) included
97 CO, NO, moisture content, and NO_x as inputs into their ANN model, and showed that the statistical
98 model produced forecasts that are consistent with our knowledge of the chemical processes of ozone
99 production.

100 Random forests are classifying algorithms that utilize a large number of independent,
101 identically distributed (i.i.d.) decision trees with randomly selected predictors and average the results
102 to reduce noise (Breiman, 2001; Hastie *et al.*, 2017). Yu *et al.* (2016) apply this technique to categorize
103 air quality conditions in Shenyang city, China using the Chinese AQI and found that their random
104 forest algorithm (RAQ) was more precise and accurate than other machine learning classification

methods (Naïve Bayes, Logistic, etc.), though their application was not necessarily a forecast of ozone levels since the Chinese AQI is a maximum value of a number of pollutants (SO_2, NO_2, CO, $PM_{2.5}$, PM_{10}, and O_3).

In this work, we developed statistical forecasting models of surface O_3 for six urban areas in Texas (Table 1). We focused on forecasting ozone during eastern Texas's "ozone season" of May through October. We used generalized additive models – GAMs (Wood, 2006) – to describe the potentially non-linear relationship between the measured maximum daily 8-hour average ozone concentrations ($O_{3,MDA8}$), selected meteorological variables from weather forecasts, and measurements of the previous day's surface O_3 levels based on data from 2009-2015. We developed two GAM-based models, a quantitative forecast that predicts the numerical value of $O_{3,MDA8}$ and a probabilistic model that predicts the probability that $O_{3,MDA8}$ will exceed a given threshold. We then used the random forest method to develop a classification forecast of the O_3 air quality index (AQI) – to our knowledge this is the first time this method has been used to forecast surface O_3 concentrations directly. Here we describe our model development and the data used to train and evaluate the models. We then assess the performance of the models in forecasting $O_{3,MDA8}$ during the 2016 O_3 season at the urban areas of interest.

METHODS

Training and Evaluation Data

To develop the quantitative and probabilistic forecast models, we fitted GAMs to calculate the predicted $O_{3,\ MDA8}$ for day D_i in each urban area listed in Table 1 using ozone season data from 2009 to 2015, with the day D_{i-1} $O_{3,\ MDA8}$ calculated from the Texas Commission on Environmental Quality (TCEQ) monitor data as a predictor. The other predictors used in these models are shown in Table 2. $O_{3,\ MDA8}$ values from the Texas Air Monitoring Information System (TAMIS) for D_{i-1} were the only measurements that were used in the forecast models [f]. We also used seven forecasted (D_i) meteorological parameters from the Model Output Statistics (MOS) post-processing of the National Center for Environmental Prediction (NCEP) 12-km North American Model (NAM-12) forecasts[g]. We discuss the selected predictors for each dataset further in this section.

Texas Commission on Environmental Quality Monitor Data

TCEQ provided air quality data from the air quality monitoring network operated by the TCEQ, its grantees, or local agencies whose data are stored in the TAMIS in and near the urban areas listed in

[f] Available at https://doi.org/10.5281/zenodo.2032330

[g] Available at https://doi.org/10.5281/zenodo.1979000

135 Table 1. Historical data for the time period spanning 2009-2015 were provided by TCEQ and used in
136 the training of the forecasting models. We used these data to calculate $O_{3,\,MDA8}$ for a given day over all
137 sites around an urban area with acceptable data. For the 2016 ozone season evaluation period, the
138 operational forecasting models gathered these $O_{3,MDA8}$ data in near-real time (NRT) from the TCEQ
139 website[h], which provided raw hourly measurements of O_3 from all of the monitoring stations in its
140 network. Note that while the training data were subjected to a quality control process by TCEQ, this
141 was not true of the NRT data used in the operational forecasts. These processing differences had a
142 small effect on our forecasts. For example, the mean and standard deviation for the NRT Austin $O_{3,MDA8}$
143 data were 43.17 and 9.68 ppbv, respectively, while the moments of the corresponding quality-
144 controlled distribution were 44.46 and 9.95 ppbv.

145 *North American Mesoscale Numerical Forecasts*

146 NCEP provides daily output from the 12 km (NAM-12) weather forecasting model. For a given
147 model runtime, there are a number of forecast hours and grid types, but the product most relevant for
148 this application is the one produced for the continental United States on a 12-km Lambert-conic
149 conformal grid. Every six hours, model output is written to a single file and archived in the NOAA

[h] https://www.tceq.texas.gov/cgi-bin/compliance/monops/daily_average.pl

150 Operational Model Archive and Distribution System (NOMADS; Rutledge *et al.*, 2006), which is
151 available to the public[i]. Historical data (dataset ds609.0) and forecasts (ds335.0) are archived in the
152 Research Data Archive (RDA; NCEP/NWS/NOAA/ECMWF/UCAR, 2003) at the University
153 Cooperation for Atmospheric Research (UCAR)[j]. Both the NRT and historical data were processed in
154 the same way: grid points closest to the coordinates of the areas of interest (Table 1) were determined,
155 then the temperature, relative humidity, and geopotential height fields were extracted for the 500, 700,
156 850, and 925 mbar pressure levels. As mentioned, there are four forecast times (0000, 0600, 1200, and
157 1800 UTC), but only the 12Z forecast was used in this work, as this was the time found to give the
158 best $O_{3,\ MDA8}$ fit for cities in the eastern US by Camalier *et al.* (2007). This forecast time is at 0700
159 CDT for our urban areas during the ozone season.

160 *Model Output Statistics Forecasts*

161 Model output statistics (MOS) is a technique used by the National Weather Service (NWS) to
162 objectively interpret numerical model output (Carter *et al.*, 1989) and produce site-specific guidance
163 (Glahn *et al.*, 2008). Archived and NRT forecasted parameters that were used in this study include

[i] http://nomads.ncep.noaa.gov/pub/data/nccf/com/nam/prod

[j] https://rda.ucar.edu/

164 temperature, dew point, wind speed, and wind direction. The forecasts are provided every three hours
165 for the subsequent ~2.5 days. All are calculated for the same day (D_i) relative to the model runtime
166 (which is at midnight). MOS forecasts are archived by the Iowa Environmental Mesonet (IEM) at
167 Iowa State University[k] – the archived NAM-12 forecasts were used in the model training for this study.
168 NRT NAM-12 forecasts are provided by NWS[l] and were used for forecasting. Queries were performed
169 for the MOS sites listed in Table 1.

170 *National Air Quality Forecast System*

171 The National Air Quality Forecast System (NAQFC; Pan *et al.*, 2014) is a physical modeling
172 system provided by NOAA that couples the Weather Research and Forecasting Non-hydrostatic
173 Mesoscale Model (WRF-NMM) (Janjic, 2003) and the Community Multiscale Air Quality (CMAQ)
174 (Byun and Schere, 2006) regional chemical transport model. CMAQ is a multi-scale 3D Eulerian
175 chemical transport model that is used to model air quality (e.g., tropospheric O_3 and $PM_{2.5}$) at urban to
176 regional scales. During this study, the CB05 chemical mechanism (Yarwood *et al.*, 2005) and AERO5
177 aerosol module (Carlton *et al.*, 2010) are used in CMAQ v4.7.1 within the NAQFC system. The lateral

[k] https://mesonet.agron.iastate.edu/mos/fe.phtml

[l] http://www.nws.noaa.gov/mdl/synop/products/bullform.met.php#Texas

178 boundary conditions used in the simulation are monthly averaged profiles extracted from GEOS-Chem
179 (Bey et al., 2001) simulation results. The current operational model has a horizontal resolution of 12
180 km. The O_3 forecasts are produced twice daily from the 0600 and 1200 UTC cycles. Forecast time
181 projections extend out to a minimum of 30 hours (0600 run) to a maximum of 48 hours (1200 run).
182 Note that NOAA provided their forecast output for a subset of our testing period, 19-May-2016
183 through 22-Sep-2016.

184 ***Forecasting Models***

185 *Quantitative and Probabilistic Models*

186 In these two procedures, we used the "mgcv" GAM modeling library (version 1.8-23) in R
187 (version 3.2.2) discussed in Wood (2006) to fit the $O_{3,\ MDA8}$ value for day D_i using a number of
188 meteorological and air quality parameters as predictors. The general GAM formulation used in our
189 forecasting models can be written as follows:

$$g(\mu_i) = \beta_o + f_1(x_{i,1}) + f_2(x_{i,2}) + \cdots f_n(x_{i,n}) + f_p(D_i) + D_W \quad (1)$$

190 where $g(\mu_i)$ is the expected value of $O_{3,\ MDA8}$ for day i in the quantitative models and a Boolean value
191 indicating whether the $O_{3,\ MDA8}$ exceeded a given threshold in the probabilistic models. $g(\mu_i)$ is the
192 link function: a log-link is used for the quantitative forecast models and a logistic link is used for the
193 probabilistic forecast models. The $j = 1..n$ meteorological and air quality parameters are denoted with

$x_{i,j}$, with the corresponding $f_j(x_{i,j})$ being an initially unknown smooth function of $x_{i,j}$ made from a cubic-spline basis set. Following Camalier *et al.* (2007), two other predictors are also included: a smooth function $f_p(D_i)$ of the Julian day of the year (D_i) and a factor for the day of the week D_W. As we are only fitting O_3 data during the O_3 season (May-October), $f_p(D_i)$ is built with a non-periodic cubic spline basis.

Random Forest O_3 Classification Forecasts

We also used the random forest approach to develop a third forecasting model. Bosch *et al.* (2007) and Hastie *et al.* (2017) provide detailed documentation on how decision trees and their associated random forests operate[m], which we will only summarize. Given a set of data and parameters that will be used in event classification, a set of True-False tests are performed (tree nodes), and the end result is a number of end nodes (leafs, where the tests stop) that yield a percentage of each class that each leaf contains. The set of parameters that defines a class is the one that yields the highest probability of a given class. Extending this methodology to a number of trees – each of which contains a random subset of the training data and performs node tests on a random subset of the given predictors – is what is known as a random forest, and it reduces the risk of over-fitting. For the forest, the probability of a given class given a set of parameters is averaged over the ensemble of trees, and class designation is again defined by a maximum probability (this time, over the ensemble average). For random forest implementation in this study, we used the same parameters that were selected for the quantitative and probabilistic forecasting models (Table 2) and the color-coded classification scheme for O_3 air quality indices (AQI) in

[m] Also see slides 25-36 in http://www.robots.ox.ac.uk/~az/lectures/ml/lect4.pdf

Table 3.

Predictor Selection

We performed several sensitivity studies that assessed model performance for different combinations of predictors with other data sources. The sensitivity studies included:

- A base model with the parameters listed in Table 2
- Replacing the $O_{3,\ MDA8}$ derived from TCEQ measurements with the EPA AirNow Air Quality Index (AQI)
- Utilizing a stability parameter – ΔT_{NCEP} – defined as the difference between the NCEP-predicted temperature at 925 mbar, T_{925}, and the temperature at 850 mbar, T_{850}
- Using relative humidity or dew point temperature as a proxy for water vapor density
- Replacing MOS and NCEP forecasts with TCEQ current conditions
- Using HYSPLIT 24-hour back trajectory distances and bearings in addition to the TCEQ, MOS, and NCEP parameters
- Using HYSPLIT-deduced synoptic types instead of the bearings and distances

Ultimately, each sensitivity study produced statistics – best-fit line slope and intercept, Pearson *r* correlation coefficient, mean and standard deviation of residuals – that were generally worse than the model described in Table 2 when validated with data from the evaluation period.

231 ΔT_{NCEP}, the difference in NCEP-forecasted temperature difference between 925 and 850 mbar,
232 was only significant to the 5% level for Austin and San Antonio, thus with this predictor over-fitting
233 of the data was a potential danger. In addition, by omitting this parameter, NCEP forecast gathering
234 was no longer necessary. An added advantage with this improvement was that the forecaster became
235 much more efficient – NCEP processing took approximately forty seconds per city, so excluding this
236 data set saved four minutes of runtime per forecast.

237 The moisture-related MOS forecasted parameters H_R (relative humidity) and T_{DP} (dew point
238 temperature) were attempted before water vapor density because of their usage in Camalier *et al.* (2007)
239 and they were usually significant to at least the 1% level. The significance level was dependent on the
240 urban site, particularly when T_{DP} and H_R were used together. Fit separately, T_{DP} is a significant
241 predictor at the 0.1% level for all sites except DFW, while H_R is significant to the highest level (0.1%)
242 for all locations. We also investigated how an absolute measure of moisture could affect the
243 predictions by transforming these variables into water vapor density (ρ_{WV}). Like H_R, ρ_{WV} is significant
244 to the highest level (0.1%) for all locations. All of these significances are summarized in Supplement
245 Table 1. Water vapor density better reflects the impact water vapor can have on the chemical
246 production of ozone than relative humidity or dew point temperature (as chemical reactions depend on
247 concentrations in units of molecules cm^{-3}), though, and consequently is the preferred moisture

predictor. H_R and T_{DP} were thus removed from our model and replaced with an average water vapor density that is a function of temperature (in ºC) and relative humidity. ρ_{WV} is computed every hour between 0600 UTC and 21 hours later, and then an average ρ_{WV} is found.

RESULTS AND DISCUSSION

To assess the performance of the forecast models, we compared the observed and predicted $O_{3,MDA8}$ for the 2009-2015 training dataset and for the 2016 ozone season. Assessing accuracy of the models in the training period is important when, as an example, determining whether models should be applied in an operational environment to compute NRT predictions of O_3. If the residuals do not follow a normal distribution, the fitting needs to be refined. To quantify model accuracy completely, though, one must calculate the O_3 level predictions then compare them with associated measurements.

Evaluating the Statistical Fit

Quantitative Forecasts

Table 4 presents the R^2 correlation coefficient for the quantitative forecast models for each urban area. The models typically can explain over 70% of the variance ($R^2 > 0.7$), with the best fits found for Austin/Round Rock and the worst for Tyler/Longview/Marshall (the only model that does not explain at least 70% of the variance). Examination of the quantitative forecast model residuals shows that the fits to the training data set are generally good, with a normal distribution of residuals,

little trend in the residuals with the magnitude of the predicted or fitted values, and a linear relationship between the measured and fitted $O_{3,\ MDA8}$. An example is provided in Figure 1 for the Austin/Round Rock quantitative forecast model, which plots the quantile-quantile (Q-Q) comparison, residuals as a function of the log of predictor value, residual distribution, and modeled $O_{3,\ MDA8}$ as a function of measured $O_{3,\ MDA8}$. Additionally, Figure 2 exhibits the functional dependence of the $O_{3,\ MDA8}$ predictions with each predictor. As expected, the forecasted $O_{3,\ MDA8}$ increases as the previous day's $O_{3,\ MDA8}$ increases and as the forecasted afternoon temperature increases, but decreases as water vapor density and wind speed increase.

Supplement Table 1 shows the estimated significance of each predictor in the quantitative forecast models. We see that five of the eight predictors are significant for all urban areas to the α = 0.001 level. The remaining predictors are forecasted afternoon temperature, forecasted daily temperature difference, and average forecasted wind direction. The significance of each depends on the city, but even in most of these cases the predictors are significant to at least the $\alpha = 0.05$ (afternoon temperature) or $\alpha = 0.10$ (diurnal temperature difference and wind direction) level. Wind direction at TLM is the one exception. *p*-values for separate runs with T_{DP}, $H_{R,}$ T_{DP} and H_R together, ΔT_{NCEP}, and ρ_{WV} demonstrate that no significance is lost when omitting ΔT_{NCEP} and T_{DP}, and that H_R and ρ_{WV} (both of which are derived from the MOS T_{DP} and surface temperatures) are equally significant (i.e., to the

282 $\alpha = 0.001$ level) in predicting ozone levels. In selecting a moisture predictor, we decided to use ρ_{WV},
283 which yielded a substantially higher correlation coefficient for Austin, Beaumont/Port Arther, and
284 Houston.

285 *Probabilistic Forecasts*

286 We trained probabilistic forecast models that determine the odds of predicted $O_{3,\,MDA8}$ being
287 greater than or equal to four thresholds – 55 ppb, 71 ppb, 86 ppb, and 105 ppb – given the MOS
288 forecasts and the previous day's $O_{3,\,MDA8}$. However, most urban areas did not have sufficient historical
289 data to fit GAMs for the upper two thresholds, and none had enough data to fit a 105 ppb model. Thus,
290 most urban areas only have 55 ppb and 71 ppb probabilistic forecasting models, while DFW and HGB
291 also have an 86 ppb model.

292 Supplement Table 2 shows the R^2 and percent variance explained for the probabilistic forecast
293 models. In general, the explained variance was relatively low for the probabilistic models, with values
294 between 45 and 60%. Most, but not all, predictors were significant to at least the $\alpha = 0.10$ level, as
295 shown in Supplement Table 3. There is much more variance in predictor significance with the
296 probabilistic models – the p value changes with not only urban area but threshold level – compared to
297 the quantitative models, whose predictors were typically significant to the $\alpha = 0.001$ level.

298 Examining the model fits shows some interesting differences between the urban areas. For
299 example, the left panel of Figure 3 shows results for HGB for a threshold of 71 ppb. Only the
300 dependence of $O_{3,\ MDA8}$ on T_{2100}, and W_S are shown, with all other variables held constant at their mean
301 values. These $O_{3,\ MDA8}$ forecasts are primarily a function of forecasted afternoon temperature (T_{2100}),
302 showing a strong increase in the probabilities between 23 ºC to 35 ºC. However, the results for SA in
303 the right panel of Figure 3 show that in this area $O_{3,\ MDA8}$ and forecasted W_S are more important in
304 predicting when the $O_{3,\ MDA8}$ will exceed 71 ppb.

305 *O_3 Classification Forecasts*

306 Accuracy in classification models can be computed by utilizing the associated confusion matrix,
307 which is a table of what the forecaster predicted compared to what was observed (Kuhn, 2008). The
308 diagonal elements of the matrix are successful classifications, and accuracy is given by the ratio of the
309 number of successes to the total number of events. The sample size for the training period is 1280, and
310 the success rate for each site during the training set is given in Table 5. Success rates for our
311 classification models range between 65% and 85%, depending on the urban area of interest.

Forecasting Performance

Quantitative Forecasts

The quantitative models were used to forecast the entire 2016 evaluation period and then validated with the observations from TCEQ monitoring stations. Statistics to specify the forecast-measurement agreement were then calculated – these included the coefficients of the linear ordinary least-squares (OLS) fit (slope, a, and intercept, b), Pearson linear correlation coefficient (r), and the moments of the forecast-data residuals (mean bias μ and standard deviation σ). These statistics are provided in Table 6. While not a one-to-one correspondence with the observations, the models do exhibit some predictive power given that the correlation coefficient for all cities is on the order of 0.7 or higher.

We emphasize the slopes of the fits, all of which are less than unity, suggesting that the higher $O_{3,\ MDA8}$ measurements are underestimated and the lower measurements are overestimated in the models. This behavior is an artifact of the regression model – expectation values will be biased toward the mean because the mean minimizes the loss error in the fit to the training data. However, there are a few reasons this kind of disagreement might be expected. As we alluded to in the TCEQ monitor data training set section, NRT and training data were not subject to identical quality control protocol, and the training data are provided as floating point numbers while the NRT data are given in integers.

While these two quality control procedures will at most only account for a couple of ppb error in a given data point, they can contribute to some of the noise that is exhibited in the validation statistics (e.g., correlation coefficient $r < 1$).

Another possible and probably more significant factor is a change in emissions, which is not accounted for in our models. With our training set, we assume that year to year changes in emissions and other variables are small enough throughout the training period such that the year was not used as a predictor. Furthermore, we assume that the emissions during 2016 testing period resemble those from the 7-year training period. While the former is mostly true, the latter is perhaps not an accurate assumption. In Table 7, the spread of the evaluation period is much smaller than the spread of any of the training data years. When comparing 2016 to the training sample, its spread has a *z*-score of 3.6, so it is clear that 2016 was anomalous. This behavior does not explain our test period error, but it does highlight a limitation of this type of forecasting.

Finally, one must consider the complexity of the model when contemplating what the test error is relative to the training error. An overly complex model with respect to degrees of freedom will yield very low error in the training period while exhibiting substantial error in the testing period; this is because the probability of overfitting the data increases with increasing degrees of freedom (DOF). Though Table 2 indicates only 8 fitted parameters, each one of the predictors has a corresponding

346 effective degrees of freedom (EDOF), which is calculated in the cubic spline of the associated
347 smoothed function for the predictor (these smoothed functions are fitted in the GAM, see Figure 2).
348 For example, in the Austin quantitative model, there are 32.017 EDOF when considering the
349 smoothing of all of the predictors. Test error as a function of degrees of freedom can be characterized
350 as high-bias/low variance at small DOF and low-bias/high variance at large DOF (Hastie *et al.*, 2017).
351 The function is minimized at some DOF, and it is possible that our models are beyond this test error
352 minimization point.

353 To further assess the quality of the quantitative models, we compared the performance of the
354 quantitative statistical model with the WRF/CMAQ based NAQFC numerical forecast for our six
355 urban areas. NOAA provided us with the output from NAQFS numerical forecasts for the six Texas
356 urban areas, but only for a subset of the evaluation period (19-May-2016 through 22-September-2016)
357 used for the statistics in Table 6. Evaluation statistics for these WRF/CMAQ forecasts are given in
358 Table 8 (labeled "NOAA") along with the statistics for the quantitative statistical forecasts (labeled
359 "GAM") in the same time period (19-May-2016 through 22-September-2016). Our statistical models
360 have a consistently higher correlation coefficient with the measured maximum $O_{3,\,MDA8}$ than the NOAA
361 numerical forecasts, and the standard deviation of the residuals is much smaller in our statistical models,
362 both of which suggest that our statistical models are able to capture some of the variability in $O_{3,\,MDA8}$

363 that the NOAA numerical forecasts do not. Furthermore, our forecasts show a smaller absolute mean
364 bias for all urban areas except Dallas/Fort Worth. This may be because the NOAA forecasts use the
365 CMAQ model, which is most commonly used to investigate high O_3 events in similar large urban areas
366 that are in non-attainment of the National Ambient Air Quality Standards (NAAQS), and thus may
367 tend to overpredict O_3 in less polluted environments.

368 *Probabilistic Forecasts*

369 To assess our probabilistic model performance, we attempted to use the reliability and Relative
370 Operating Characteristic (ROC) area statistics, which assess how well the predicted probabilities of an
371 event correspond to their observed frequencies and how well the model discriminates between events
372 and non-events, respectively. For the purpose of our study, an event is defined as the $O_{3,\,MDA8}$ exceeding
373 a given threshold for a given urban area on a given day. The reliability and ROC area for the models
374 are given in Supplement Table 4. Perfect scores for reliability are zero and for ROC area are unity, so
375 our models ostensibly exhibit substantial skill in their event forecasting. However, it is difficult to
376 ascertain anything conclusive given the small number of extreme events that occurred during the
377 evaluation period. Our test period consists of a modest sample of 184 days, but for probabilistic
378 validation this sample is broken down even further – first into five probability bins for both statistics
379 (where most of the forecasts are contained in the first bin), then subsequently into events and non-

events (most of which are non-events) in the ROC metric. The reliability and ROC area are thus calculated at small sample sizes. Additionally, given the R^2 in the fits to the training data (Supplement Table 2), there is little reason to believe that the probabilistic models will be as successful as implied by Supplement Table 4 when applied to a larger evaluation sample.

O_3 Classification Forecasts

Evaluation of classification models is routinely done using the confusion matrix utilized to assess the performance of the training set. In our study, diagonal elements of the matrix are considered successful classifications or true positives, elements above the diagonal are false negatives (i.e., a poor air quality event was missed or the severity was under-predicted by the forecaster), and elements below the diagonal are false positives (i.e., the forecaster incorrectly predicted poor air quality). The confusion matrix for each model (i.e., each urban area) is presented in Supplement Table 5 through Supplement Table 10. Total model success rate is defined as the ratio of true positives to all events, which is effectively the sum of the diagonal matrices divided by the total number of elements in the confusion matrices. Likewise, false alarm rate is defined as the ratio of false positives to all events, and miss rate is defined as the ratio of false negatives to all events. These rates are given in Table 9.

Overall, the O_3 classification models perform well, but the tables in the supplemental data show that the more extreme the events become, the less accurate the forecasting classifier is – most of the

success is due to the "Green" days, and the areas that do have more "Orange" days (Dallas and Houston) are significantly less accurate. This again is likely due to the lack of extreme events on which to train the models and is the random forest equivalent of the underestimation of high $O_{3,\ MDA8}$ days that the quantitative forecasts exhibit. The classification models can thus be considered binary air quality forecasters – they can successfully predict either "Good" or "Bad" air quality days based on EPA AQI and $O_{3,\ MDA8}$ specifications[n].

Alternative Predictors

Our forecasting scheme predicts the ozone for a given day, based on meteorology forecasts for the day and ozone measurements the day before the models are run. However, for forecasting purposes, it may be more practical to use an earlier ozone measurement or predict levels on a more intermediate scale, since $O_{3,\ MDA8}$ calculations are not known until shortly before our forecast. To address issue, we replaced $O_{3,\ MDA8}$ from day D_{i-1} in our models with the $O_{3,\ MDA8}$ of day D_{i-2}, and we predicted $O_{3,\ MDA8}$ for both day D_i and day D_{i+1}.

Ozone levels are dependent on photochemistry, so it is also important to investigate the significance of solar radiation in our predictions. We accomplished this by utilizing the "CLD" field

[n] https://www.tceq.texas.gov/cgi-bin/compliance/monops/ozone_summary.pl#interpret

412 in the NAM MOS forecasts, which are classifications of total sky cover for a given hour. The categories
413 are given in Supplement Table 11. In training and forecasting, we assign a numerical value to the
414 categories as shown in Supplement Table 11, then average over the same time period that was used for
415 the W_S, W_D, and ρ_{WV} daily averages.

416 One limitation that exists with these new parameters is that the MOS forecasts only extend 6-
417 72 hours from the NAM MOS runtime. This forecast range inhibits any of our models from predicting
418 past day D_{i+2}. Additionally, the training data that we collected do not have forecasts for 2100 UTC on
419 day D_{i+2}, and this time is required for T_{2100} and the predictors that are daily averages. Consequently,
420 our new forecasts only use the new predictors for a day D_i and D_{i+1} forecast.

421 As we did for our quantitative models, we compared the predictions with these three new
422 models with the observations by computing the linear coefficients of the best-fit lines, linear
423 correlation coefficient, and mean and standard deviation of the differences for each model and
424 observation set (i.e., for each of the 3 new models and 6 cities). Results are given in Supplementary
425 Tables 12-14. When compared with Table 6, it is clear that using older ozone observations in the
426 models and attempting to predict the levels further out is less precise than the baseline model using the
427 predictors given in Table 2. Supplementary Table 12 provides statistics for a model where the clouds
428 predictor is added to the baseline model and the $O_{3,\,MDA8}$ predictor was modified to be for day D_{i-2}. The

altered model yields smaller slopes and correlation coefficients but larger best-fit intercepts, biases, and spread compared to the baseline. A model with the same predictors (clouds and $O_{3,\ MDA8}$ for day D_{i-2}) but whose target is the $O_{3,\ MDA8}$ for day D_{i+1} was also developed, and its agreement with the data exhibits similar trends as the previous model. Supplementary Table 13 illustrates that the statistics further degrade as the prediction becomes more intermediate rather than short-term.

One more model we trained was the baseline model (Table 2) with only the clouds predictor added to isolate the effect of including cloud forecasts. Statistics are provided in Supplementary Table 14. Again, aggregate agreement worsens, but only slightly compared to the baseline. It is worth noting that the slope and bias for San Antonio improve, as does the bias for Dallas, though for both of these sites the spread in the model-data differences increases.

CONCLUSIONS AND FUTURE WORK

For six urban areas in eastern Texas, we built three types of O_3 forecasting models that predict the maximum daily 8-hour averaged O_3 ($O_{3,\ MDA8}$) for a given day based on meteorological forecasts for the same day and the $O_{3,\ MDA8}$ from the previous day. The three forecast model types were a quantitative forecast of $O_{3,\ MDA8}$ (based on a GAM with a log-link function), a probabilistic forecast that predicted the probability that $O_{3,\ MDA8}$ was above a given threshold (based on a GAM with a logit function), and a random forest classification model for predicting the Air Quality Index classifications.

446 We performed a number of sensitivity studies to find the best set of predictors to use in all three model
447 types (and for all six urban areas). These parameters included MOS forecasts of afternoon high
448 temperature; MOS-forecasted daily temperature difference; daily-averaged MOS wind speed, wind
449 direction, and water vapor density; day-of-week; day-of-year; and $O_{3,\ MDA8}$. The models were then
450 tested by forecasting the 2016 ozone season (May through October).

451 The quantitative forecasts explain 69% or more of the variance in each urban area except for
452 TLM. Five of the eight predictors were significant at the $\alpha = 0.001$ level for all urban areas, while the
453 other three (afternoon temperature, diurnal temperature difference, and wind direction) had
454 significance that varied with location. The probabilistic forecasting models showed little skill in
455 determining the probability of ozone exceeding policy-relevant thresholds during the 2016 ozone
456 season. The ozone classification forecast using the random forest technique correctly predicted the O_3
457 AQI classification 67% of the time in HGB, 77% of the time in DFW, and over 86% of the time
458 elsewhere. False negatives (where the severity of a poor air quality event is under-predicted) were
459 generally more common than false positives (where air quality is predicted to be worse than observed).

460 While the quantitative and classification forecast models show skill in forecasting O_3 and can
461 thus be used to assist with issuance of public statements and advisories regarding air quality, there are
462 remaining issues that need to be addressed in future work. The quantitative model forecasts trend to

the mean and have less variance than the observations, resulting in slopes of the best-fit line to the model and data that are between $0.5 < a < 0.75$ rather than near unity. Additionally, our classification scheme can best be described as binary because it correctly identifies good (green) and poor air quality days while not successfully distinguishing between different degrees of elevated O_3 (yellow, orange, red, maroon) concentration. The reason for this characteristic is likely the small number of extreme events in the training set.

Future work will focus on expanding the models to work in areas outside of Texas and improving the forecast model performance. Funding limited the scope of this project to only urban areas in Texas. Extending the forecasts to other states and perhaps countries should be possible, but more work is needed to determine how the methods in this paper will perform outside of Texas. We intend on addressing performance with techniques such as weighting the aforementioned extreme events more strongly in the model training or applying neural networks instead of generalized additive models.

ACKNOWLEDGMENTS

This study was funded by TCEQ Contract No. 582-15-50414 to AER, and we would also like to thank E. Gribbin of TCEQ for providing meteorological and air quality data for our training and forecast evaluation. We thank Pius Lee, Jeff McQueen, and Daniel Tong of NOAA for providing us

480 with their numerical O_3 forecasts that we used in our validation. Finally, we thank the UCAR Data
481 Archive support staff, specifically Douglas Schuster and Chi-Fan Shih, for greatly facilitating the
482 collection of historical NCEP NAM-12 model output. Lastly, we express our gratitude to the
483 anonymous reviewers that enriched the content of this publication.

484 # REFERENCES

485 Abdul-Wahab, S. A., & Al-Alawi, S. M. (2002). Assessment and prediction of tropospheric ozone
486 concentration levels using artificial neural networks. *Environ. Modell. Software*, 219-228.

487 Anderson, H., Spix, C., Medina, S., Schouten, J., Castellsague, J., Rossi, G., . . . Katsouyanni, K.
488 (1997). Air pollution and daily admissions for chronic obstructive pulmonary disease in 6
489 European cities: results from the APHEA project. *Eur Respir J.*, 1064-71.

490 Bell, M., McDermott, A., Zeger, S., Samet, J., & Dominici, F. (2004). Ozone and short-term mortality
491 in 95 US urban communities, 1987-2000. *JAMA*, 2372-8.

492 Bey, I. D., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., . . . Schultz, M. G.
493 (2001). Global modeling of tropospheric chemistry with assimilated meteorology: Model
494 description and evaluation. *J. Geophys. Res*, 23073-23095.

495 Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

496 Brown, J., Bateson, T., & McDonnell, W. (2008). Effects of exposure to 0.06 ppm ozone on FEV1 in
497 humans: a secondary analysis of existing data. *Environ Health Perspect*, 1023-6.
498 Burnett, R., Brook, J., Yung, W., Dales, R., & Krewski, D. (1997). Association between ozone and
499 hospitalization for respiratory diseases in 16 Canadian cities. *Environ Res*, 24-31.
500 Byun, D., & Schere, K. L. (2006). Review of the governing equations, computational algorithms, and
501 other components of the models-3 Community Multiscale Air Quality (CMAQ) system.
502 *Applied Mechanics Reviews*, 51-77.
503 Camalier, L., Cox, W., & Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and
504 their use in assessing ozone trends. *Atmos. Environ.*, 7127-7137.
505 Carlton, A. G., Bhave, P. V., Napelenok, S. L., Edney, E. D., Sarwar, G., Pinder, R. W., . . . Houyoux,
506 M. (2010). Model representation of secondary organic aerosol in CMAQv4.7. *Environ. Sci.*
507 *Technol.*, 8553-8560.
508 Carter, G., Dallavalle, J., & Glahn, H. (1989). Statistical Forecasts Based on the National
509 Meteorological Center's Numerical Weather Prediction System. *Weather and Forecasting*,
510 401-412.
511 Devlin, R., Raub, J., & Folinsbee, L. (1997). Health effects of ozone. *Science and Medicine*, 8.

512 Feister, U., & Balzer, K. (1991). Surface ozone and meteorological predictors on a subregional scale.
513 *Atmos. Environ.*, 1781-1970.
514 Glahn, B., Gilbert, K., Cosgrove, R., Ruth, D., & Sheets, K. (2008). The Gridding of MOS. *Weather*
515 *and Forecasting*, 520-529.
516 Gong, X., Kaulfus, A., Nair, U., & Jaffe, D. (2017). Quantifying O3 Impacts in Urban Areas Due to
517 Wildfires Using a Generalized Additive Model. *Environmental Science & Technology*, 13216-
518 13223.
519 Janjic, Z. I. (2003). A nonhydrostatic model based on a new approach. *Meteorol. Atmos. Phys*, 271-
520 285.
521 Kgabi, N. A., & Sehloho, R. M. (2012). Tropospheric ozone concentrations and meteorological
522 parameters. *Global Journal of Science Frontier Research*, 11-21.
523 Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical*
524 *Software*, 1-26.
525 Lippmann, M. (1989). Health Effects of Ozone: A Critical Review. *JAPCA*, 672-695.
526 Logan, J. (1985). Tropospheric Ozone: Seasonal Behavior, Trends, and Anthropogenic Influence. *J.*
527 *Geophys. Res.*, 10463-10482.

528 Luna, A. S., Paredes, M. L., de Oliveira, G. C., & Correa, S. M. (2014). Predicion of ozone
529 concentration in tropospheric levels using artificial neural networks and support vector
530 machine at Rio de Janeiro, Brazil. *Atmos. Environ.*, 98-104.
531 National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of
532 Commerce, European Centre for Medium-Range Weather Forecasts, and Unidata/University
533 Corporation for Atmospheric Research. (2016, April 26). *Historical Unidata Internet Data*
534 *Distribution (IDD) Gridded Model Data. 2003, updated daily*. Retrieved from Research Data
535 Archive at the National Center for Atmospheric Research, Computational and Information
536 Systems Laboratory: http://rda.ucar.edu/datasets/ds335.0/
537 Pan, L., Tong, D., Lee, P., Kim, H.-C., & Chai, T. (2014). Assessment of NOx and O3 forecasting
538 performances in the U.S. National Air Quality Forecasting Capability before and after the 2012
539 major emissions updates. *Atmos. Environ.*, 610-619.
540 Rutledge, G., Alpert, J., & Ebisuzaki, W. (2006). NOMADS: A Climate and Weather Model Archive
541 at the National Oceanic and Atmospheric Administration. *Bulletin for the American*
542 *Meteorological Society*, 327-341.
543 Thompson, M., Reynolds, J., Cox, L. H., Guttorp, P., & Sampson, P. D. (2001). A review of statistical
544 methods for meteorological adjustment of tropospheric ozone. *Atmos. Environ.*, 617-630.

U.S. Environmental Protection Agency. (2013). *U.S. EPA. Integrated Science Assessment (ISA) of Ozone and Related Photochemical Oxidants.* Washington, DC, DC, USA: EPA/600/R-10/076F.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R.* Boca Raton, FL: Chapman and Hall/CRC.

Yarwood, G. (2005). *Updates to the carbon bond chemical mechanism: CB05*. Retrieved from http://www.camx.com/publ/pdfs/cb05_final_report_120805.pdf

Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. A. (2016). RAQ -- A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. *Sensors*.

555 # TABLES

556 Table 1. TCEQ and MOS sites corresponding to the six urban areas investigated in this report.

TCEQ Area Abbreviation	MOS Site Abbreviation	MOS Site Location (latitude, longitude)	Associated Cities
ARR	KAUS	30.321, -90.760	Austin/Round Rock
BPA	KBPT	29.951, -94.021	Beaumont/Port Arthur
DFW	KDFW	32.898, -97.019	Dallas/Fort Worth
HGB	KIAH	29.980, -95.360	Houston/Galveston/Brazoria
SA	KSAT	29.544, -98.484	San Antonio
TLM	KGGG	32.385, -94.712	Tyler/Longview/Marshall

557

558 Table 2. Predictors used in the GAM-based forecasts (quantitative and probabilistic) and random forest

559 classification forecast.

Predictor	Units
Day-of-week (D_w)	None
Day-of-year (D_i)	None
MOS D_i Forecast Afternoon (2100 UTC) Temperature (T_{2100})	ºC
MOS D_i Forecast Diurnal Temperature Difference (ΔT)	ºC
MOS D_i Forecast Daily Average water vapor density (ρ_{wv})	g/m^3
MOS D_i Forecast Daily Average Wind Speed (W_S)	$m\ s^{-1}$
MOS D_i Forecast Daily Average Wind Direction (W_D)	degrees clockwise from North
TAMIS O_3 *MDA8* for D_{i-1} ($O_{3,\ MDA8}$)	ppbv

560

561

Table 3. $O_{3,\ MDA8}$ classes used in the classification model. The ranges are defined by the EPA and summarized by the TCEQ at https://www.tceq.texas.gov/cgi-bin/compliance/monops/ozone_summary.pl#interpret.

Color Class	$O_{3,\ MDA8}$ **Range (ppb)**
Green	$O_{3,\ MDA8} < 55$
Yellow	$55 \leq O_{3,\ MDA8} < 71$
Orange	$71 \leq O_{3,\ MDA8} < 86$
Red	$86 \leq O_{3,\ MDA8} < 105$
Maroon	$105 < O_{3,\ MDA8}$

566 Table 4. R^2 and standard deviation (σ) of the residuals for the quantitative model training.

Urban Area	R^2	**σ of Residuals (ppbv)**
ARR	0.755	6.61
BPA	0.726	8.36
DFW	0.716	8.57
HGB	0.734	9.77
SA	0.734	7.31
TLM	0.698	7.72

567

568

569 Table 5. Success rate for the classification models executed on the training data set.

City	Success Rate
Austin	84.4%
Beaumont	80.3%
Dallas	66.3%
Houston	68.4%
San Antonio	76.3%
Tyler	77.3%

570

571

572 Table 6. Best-fit line slope (a) and intercept (b) and Pearson correlation coefficient (r) of the $O_{3,\ MDA8}$
573 forecasted by the quantitative models as a function of the $O_{3,\ MDA8}$ calculated from the TAMIS O_3
574 measurements. The mean (μ) and standard deviation (σ) of the model-data residuals are also provided.

Urban Area	*a*	*b*	*r*	*μ (ppbv)*	*σ (ppbv)*
ARR	0.70	11.70	0.81	-1.04	5.74
BPA	0.65	11.98	0.77	-2.14	7.80
DFW	0.70	16.74	0.80	0.90	7.54
HGB	0.53	19.59	0.69	-4.71	9.98
SA	0.74	12.93	0.80	2.04	6.42
TLM	0.64	17.12	0.77	2.27	7.02

575

576 Table 7. The mean (μ) and standard deviation (σ) of the observed O_3 MDA8 distributions for the
577 entire Austin training data set, each of the individual years in the training data set, and the evaluation
578 data set (2016).

Year	$O_{3,\,MDA8}$ μ (ppb)	$O_{3,\,MDA8}$ σ (ppb)
2009-2015	46.47	13.56
2009	45.09	13.17
2010	43.45	14.65
2011	52.16	12.4
2012	46.48	13.27
2013	44.81	13.00
2014	45.64	12.01
2015	47.65	14.62
2016	43.17	9.68

579

580

Table 8. Best-fit line slope (a) and intercept (b) and Pearson correlation coefficient (r) of the $O_{3,\ MDA8}$ forecasted by both the NOAA numerical models (*i.e.*, the WRF/CMAQ National Air Quality Forecast Capability results for our six urban areas, labeled "NOAA" in the table) and the quantitative statistical models (labeled "GAM" in the table) as a function of the $O_{3,\ MDA8}$ calculated from the TAMIS O_3 measurements. The mean (μ) and standard deviation (σ) of the model-data residuals is also provided. Note: the time period for these statistics is 19-May-2016 through 22-Sep-2016.

	Urban Area	*a*	*b*	*r*	*μ (ppbv)*	*σ (ppbv)*
NOAA	**ARR**	0.42	30.62	0.32	6.04	7.71
	BPA	0.55	28.36	0.34	11.35	10.3
	DFW	0.36	35.47	0.38	0.62	10.3
	HGB	0.31	40.00	0.19	5.86	12.4
	SA	0.38	33.25	0.34	8.15	7.95
	TLM	0.58	28.64	0.42	11.69	9.14
GAM	**ARR**	0.62	14.12	0.79	-1.59	5.34
	BPA	0.59	13.72	0.72	-1.82	7.91
	DFW	0.70	17.49	0.80	1.14	7.80
	HGB	0.44	23.09	0.64	-4.48	10.23
	SA	0.72	12.67	0.79	1.59	5.84
	TLM	0.65	16.86	0.80	2.56	6.74

589 Table 9. Classification forecast success, false alarm, and miss rates for all models.

Urban Area	*Success Rate*	*False Alarm Rate*	*Miss Rate*
Austin	87.01%	4.54%	8.44%
Beaumont	86.36%	3.25%	10.39%
Dallas	77.27%	7.14%	15.58%
Houston	66.88%	1.95%	31.17%
San Antonio	89.12%	2.72%	8.16%
Tyler	90.26%	2.60%	7.14%

590

591

TABLE TITLES

Table 1. TCEQ and MOS sites corresponding to the six urban areas investigated in this report.

Table 2. Predictors used in the GAM-based forecasts (quantitative and probabilistic) and random forest classification forecast.

Table 3. $O_{3,\ MDA8}$ classes used in the classification model. The ranges are defined by the EPA and summarized by the TCEQ at https://www.tceq.texas.gov/cgi-bin/compliance/monops/ozone_summary.pl#interpret.

Table 4. R^2 and standard deviation (σ) of the residuals for the quantitative model training.

Table 5. Success rate for the classification models executed on the training data set.

Table 6. Best-fit line slope (a) and intercept (b) and Pearson correlation coefficient (r) of the $O_{3,\ MDA8}$ forecasted by the quantitative models as a function of the $O_{3,\ MDA8}$ calculated from the TAMIS O_3 measurements. The mean (μ) and standard deviation (σ) of the model-data residuals are also provided.

Table 7. The mean (μ) and standard deviation (σ) of the observed O_3 MDA8 distributions for the entire Austin training data set, each of the individual years in the training data set, and the evaluation data set (2016).

Table 8. Best-fit line slope (a) and intercept (b) and Pearson correlation coefficient (r) of the $O_{3,\ MDA8}$ forecasted by both the NOAA numerical models (*i.e.*, the WRF/CMAQ National Air Quality Forecast

609 Capability results for our six urban areas, labeled "NOAA" in the table) and the quantitative statistical
610 models (labeled "GAM" in the table) as a function of the $O_{3,\ MDA8}$ calculated from the TAMIS O_3
611 measurements. The mean (μ) and standard deviation (σ) of the model-data residuals is also provided.
612 Note: the time period for these statistics is 19-May-2016 through 22-Sep-2016.
613 Table 9. Classification forecast success, false alarm, and miss rates for all models.
614

FIGURES

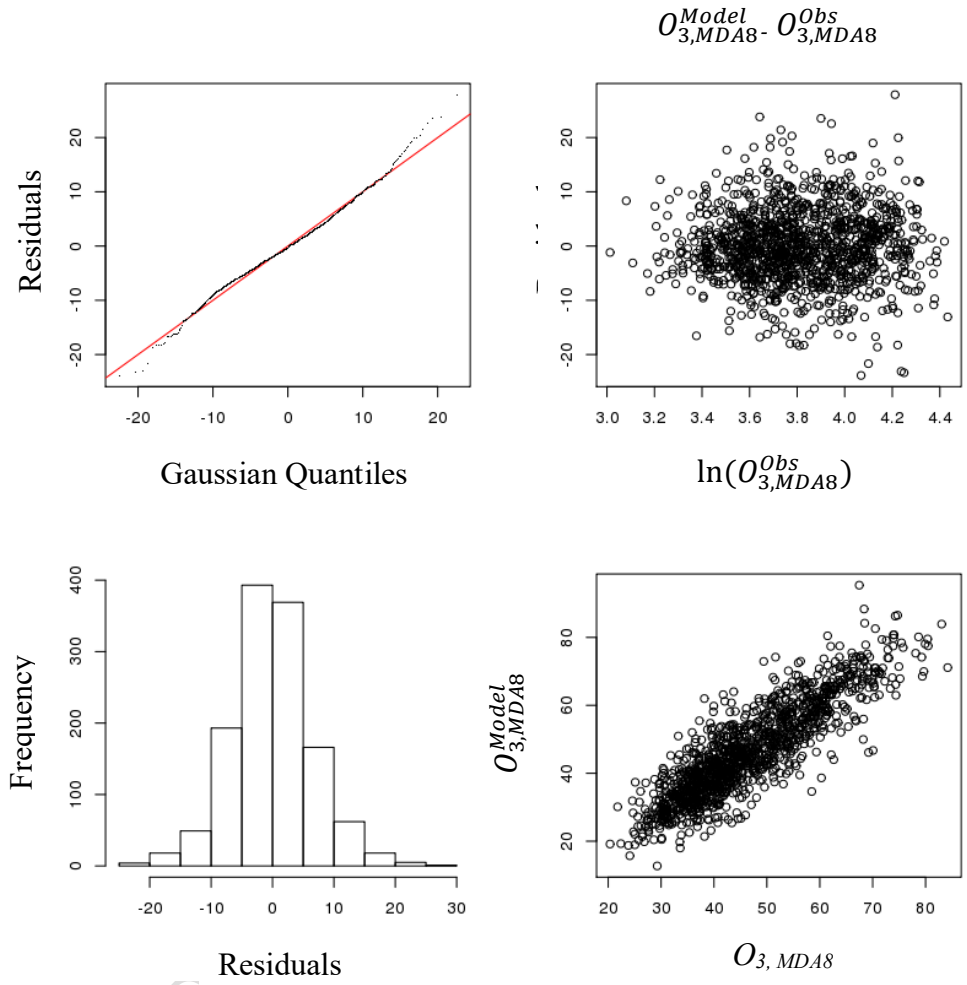


Figure 1. Training period residual analysis for the Austin/Round Rock quantitative model. The top-left panel is a Q-Q plot which implies a normal distribution if the data points exhibit a y=x relationship (represented by the red line). The top-right panel plots the data-model residuals as a function of the log of the measurements. A residual distribution is presented in the bottom-left panel. Finally, model predictions as a function of their associated measurements are plotted in the bottom-right panel. For ARR, the plots show that the quantitative model produces residuals that are normally distributed.

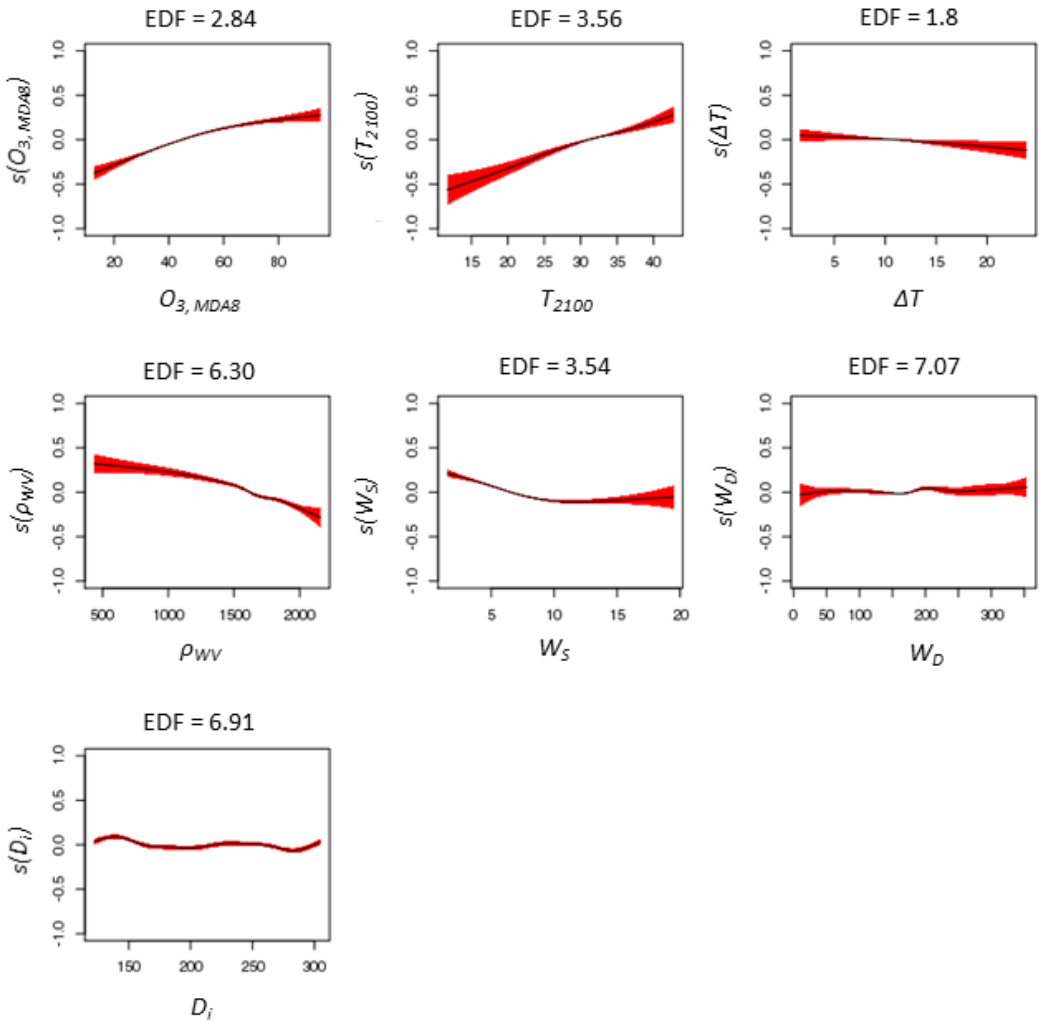
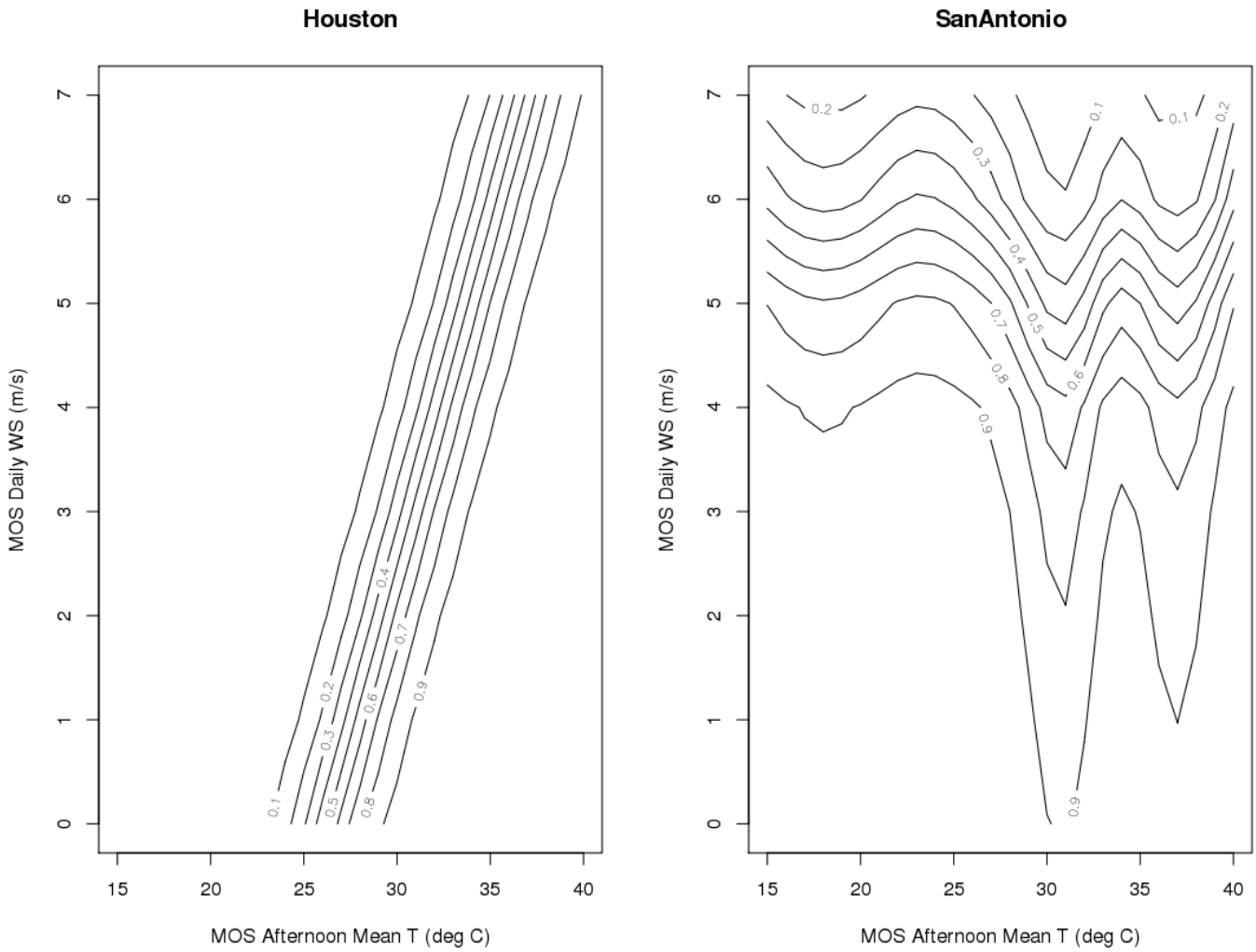


Figure 2. Effects (i.e., weights) of each predictor as a function of predictor value, as represented by smoothed curves determined in the GAM fitting. Effective degrees of freedom (EDF) for each of the smooth functions is given in the title of each panel. 95% confidence levels are displayed as the red regions around the black curves. As an example, one interpretation of the T_{2100} plot is that its effect on the GAM response (predicted O_3) is estimated by a smooth curve with 3.56 degrees of freedom (Wood, 2006).

633



634

635 Figure 3. Plot of the probabilities of forecasted $O_{3,\,MDA}$ being ≥ 71 ppb in Houston-Galveston-Brazoria
636 (HGB) and San Antonio (SA) when the $O_{3,\,MDA}$ from the previous day is 70 ppb. The *x*-axis is the MOS
637 NAM-12 forecast of afternoon mean temperature for a given day while the *y*-axis is the MOS NAM-
638 12 forecast of daily averaged wind speed for a given day. All other model predictors are held constant
639 at their mean values.

640

FIGURE CAPTIONS

Figure 1. Training period residual analysis for the Austin/Round Rock quantitative model. The top-left panel is a Q-Q plot which implies a normal distribution if the data points exhibit a y=x relationship (represented by the red line). The top-right panel plots the data-model residuals as a function of the log of the measurements. A residual distribution is presented in the bottom-left panel. Finally, model predictions as a function of their associated measurements are plotted in the bottom-right panel. For ARR, the plots show that the quantitative model produces residuals that are normally distributed.

Figure 2. Effects (i.e., weights) of each predictor as a function of predictor value, as represented by smoothed curves determined in the GAM fitting. Effective degrees of freedom (EDF) for each of the smooth functions is given in the title of each panel. 95% confidence levels are displayed as the red regions around the black curves. As an example, one interpretation of the T_{2100} plot is that its effect on the GAM response (predicted O_3) is estimated by a smooth curve with 3.56 degrees of freedom (Wood, 2006).

Figure 3. Plot of the probabilities of forecasted $O_{3,\,MDA}$ being $\geq$ 71 ppb in Houston-Galveston-Brazoria (HGB) and San Antonio (SA) when the $O_{3,\,MDA}$ from the previous day is 70 ppb. The *x*-axis is the MOS NAM-12 forecast of afternoon mean temperature for a given day while the *y*-axis is the MOS NAM-

12 forecast of daily averaged wind speed for a given day. All other model predictors are held constant at their mean values.