Observed and predicted sensitivities of extreme surface ozone to meteorological drivers in three US cities

Miranda J. Fix[a,*], Daniel Cooley[a], Alma Hodzic[b], Eric Gilleland[b], Brook T. Russell[c], William C. Porter[d], Gabriele G. Pfister[b]

[a]*Department of Statistics, Colorado State University*
1877 Campus Delivery, Fort Collins, CO 80523-1877
[b]*National Center for Atmospheric Research*
[c]*Department of Mathematical Sciences, Clemson University*
[d]*Department of Environmental Sciences, University of California, Riverside*

Abstract

We conduct a case study of observed and simulated maximum daily 8-hour average (MDA8) ozone (O_3) in three US cities for summers during 1996-2005. The purpose of this study is to evaluate the ability of a high resolution atmospheric chemistry model to reproduce observed relationships between meteorology and high or extreme O_3. We employ regional coupled chemistry-transport model simulations to make three types of comparisons between simulated and observational data, comparing (1) tails of the O_3 response variable, (2) distributions of meteorological predictor variables, and (3) sensitivities of high and extreme O_3 to meteorological predictors. This last comparison is made using two methods: quantile regression, for the 0.95 quantile of O_3, and tail dependence optimization, which is used to investigate even higher O_3 extremes. Across all three locations, we find substantial differences between simulations and observational data in both meteorology and meteorological sensitivities of high and extreme O_3.

Keywords: surface ozone, meteorological variables, quantile regression, extreme value theory

1. Introduction

Surface ozone (O_3) is one of the major air pollutants associated with adverse health effects. According to the US Environmental Protection Agency (EPA), current scientific evidence supports a causal relationship between short-term exposures to O_3 and respiratory health effects, and a

*Corresponding author
Email address: miranda.fix@colostate.edu (Miranda J. Fix)

likely to be causal association with total mortality (IHME, 2013). The O_3 health effects have been found to be non-linear, and may be especially detrimental at high levels of O_3 (Wilson et al., 2014). In addition, ambient air quality standards for "criteria" pollutants such as O_3 typically impose a penalty for exceeding a high concentration threshold. Thus for both air quality regulation and human health concerns, it is important to understand the conditions leading to the most extreme O_3 levels and to be able to reliably predict these extreme levels under present and future climate via atmospheric chemistry models.

Processes controlling surface O_3 concentrations are relatively well understood (Seinfeld et al., 1998). Surface O_3 is mostly a summertime pollutant produced by photochemical oxidation of volatile organic compounds (VOCs) by hydroxyl radical (OH) in the presence of nitrogen oxides (NO_x) and sunlight. Most efficient losses of surface O_3 include the removal by dry deposition uptake to vegetation, and its photolysis in the presence of water vapor which leads to the formation of OH. It is also well known that O_3 concentrations near the surface are strongly affected by meteorological parameters including (but not limited to) the boundary layer winds (mixing/dispersion), temperature which influences the emissions of biogenic precursors, and cloudiness which influences the radiation fluxes available for photolytic reactions.

Accurate estimation of O_3 sensitivity to individual meteorological variables is challenging due to the complex interdependencies and processes at play. Research conducted across many settings, including both observational and model perturbation studies, suggests that elevated O_3 concentrations are most strongly linked with increases in temperature (Jacob and Winner, 2009; Pearce et al., 2011). Exceptionally high O_3 levels were observed in Europe in August 2003 associated with hot and dry heat-wave conditions (Vautard et al., 2007). In an analysis of covariance performed on observed daily O_3 maxima in Switzerland during the 1992-2002 period, Ordónez et al. (2005) found that temperature and global radiation accounted for most of the meteorological variability in summer O_3 concentration. In a model perturbation study over the eastern US during July 2001, Dawson et al. (2007) found that on average temperature had the largest (positive) effect on maximum daily 8-hour average (MDA8) O_3. Absolute humidity had a smaller but appreciable (negative) impact. Also focusing on the eastern US, Camalier et al. (2007) were able to explain up to 80% of the variability in observed MDA8 O_3 with a generalized linear model. They found regional variability in the prevailing meteorological parameters driving O_3 response, with temper-

ature most dominant in the northeast US and relative humidity playing a more significant role in the southeast US. Transport distance and direction also had strong effects in some areas.

The studies referenced above focus on the average O_3 response. However, meteorological sensitivities at high quantiles of O_3 have been shown to differ from those of the overall median (Baur et al., 2004; Porter et al., 2015). In the present study we focus on high and extreme O_3 levels, thus requiring specialized tools such as quantile regression and extreme value analysis. Quantile regression is beginning to be recognized as a powerful tool in air pollution studies (Zhao et al., 2016). For instance, Otero et al. (2016) applied quantile regression to estimate the meteorological influence on the 0.95 quantile of MDA8 O_3 over Europe during 1998-2012. In summer months, they found that maximum temperature and southerly flow were selected as predictors in over 80% of the models, with relative humidity and surface solar radiation following closely behind. Porter et al. (2015) applied quantile regression to observed daily O_3 levels across the US during 2004-2012, and found maximum temperature to be the dominant driver of 0.95 quantile MDA8 O_3 in the summer. Consistent with the analysis of Camalier et al. (2007), they also found a strong negative relationship of relative humidity with O_3 in many locations, especially in the southern US. For extremely high quantiles, quantile regression suffers from data scarcity and extreme value analysis is needed. Russell et al. (2016b) developed a method to optimize tail dependence between O_3 and a linear combination of meteorological drivers. Russell et al. (2016a) applied this method to a spatial study of extreme summer MDA8 O_3 in the southeast and mid-Atlantic region of the US, and similarly found that air temperature was more important in the northern portion of the region while low humidity was more influential in the southern portion of the region.

Atmospheric chemistry models are essential for making short-term predictions of air quality, as well as projections of future air quality under climate change. Reproducing observed sensitivities of pollutants to meteorology is needed for building confidence in such model projections, but evaluation of model performance is lacking for air quality at high and extreme levels. The goal of this study is to evaluate model skill in reproducing observed relationships between meteorology and O_3 extremes in the US, such as those explored in Porter et al. (2015) and Russell et al. (2016b). We utilize a set of high resolution, regional scale atmospheric chemistry model simulations by Pfister et al. (2014). Although our focus is on the relationship between high/extreme O_3 and meteorological predictors, it is also necessary to examine the marginal distributions of both response and predictor

variables individually. Thus, as illustrated in Figure 1, our study framework includes three types of comparisons between simulated and observational data, comparing (1) the O_3 response variable, (2) the meteorological predictor variables, and (3) the sensitivities of high and extreme O_3 to meteorological predictors. The first two are comparisons of distributions, and for the O_3 response we largely focus on comparing the distributions' tails. The comparison of sensitivities is made using two methods: quantile regression and the tail dependence optimization method developed by Russell et al. (2016b). To our knowledge, this is the first study to apply these statistical methods to O_3 simulated from an atmospheric chemistry model, as well as the first study to compare the meteorological sensitivities of high/extreme O_3 between simulated and observed O_3.

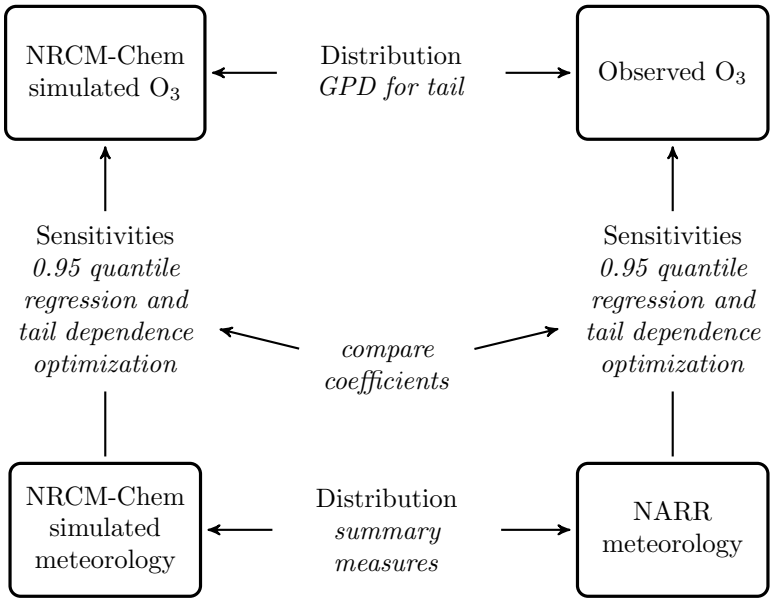

Figure 1: Illustration of the framework used in this study to compare simulated and observational data.

2. Inputs

2.1. Observations and NARR

We analyze surface O_3 measurements from the EPA's air quality system (AQS[1]) for summers (JJA) during the years 1996-2005. For consistency with the EPA's National Ambient Air Quality

[1]https://www.epa.gov/outdoor-air-quality-data

Standards (NAAQS), we extract MDA8 O_3 concentrations for our analysis. Because the statistical methodology is computationally costly, we focus on a case study of three AQS monitoring stations: station 13-121-0055 in Atlanta, station 48-201-0046 in Houston, and station 04-013-3002 in Phoenix. There were at most 5 days of data missing out of 920 days total at each of the stations. These three US cities have historically high levels of O_3, and fall within 8-hour O_3 nonattainment areas as designated by the EPA. Atlanta, Houston, and Phoenix represent a range of regional climates across the southern US, and belong to EPA regions 4, 6, and 9, respectively. However, we do not view this as a comprehensive study of these regions. These stations all reflect urban environments, however exploratory analysis found that a rural station in moderate proximity to Atlanta showed strong correlation to the urban Atlanta station, thus the sensitivities of high and extreme ozone to NARR meteorology would be very similar.

Following Porter et al. (2015) and Russell et al. (2016b), we obtain meteorological variables from the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR) product (Mesinger et al., 2006), which combines model and assimilated observational datasets. NARR is a gridded product with a spatial resolution of 32 km and 8 output fields per day (representing 3-hour means). There is a spatial mismatch between the point-located O_3 observations and the gridded NARR meteorology. We use output from the NARR grid cell whose midpoint is closest to the AQS monitoring station of interest. NARR has been used previously to examine meteorological drivers of observed air pollution (e.g. Tai et al., 2010). In addition, the NARR output is complete and does not need additional quality control.

2.2. NRCM-Chem simulations

We utilize a set of climate simulations conducted by Pfister et al. (2014) using the nested regional climate model with chemistry (NRCM-Chem), which is based on the regional Weather Research and Forecasting model with chemistry (WRF-Chem, version 3.3). WRF-Chem is a fully coupled chemical transport model (Grell et al., 2005), which was run at a high spatial resolution of 12 km providing hourly outputs for the variables that we consider. We extract MDA8 O_3 concentrations from the NRCM-Chem gridpoint closest to each of the AQS stations. Daily meteorological variables (see Section 2.3) are also extracted from the NRCM-Chem simulations at these gridpoints.

We use the present time NRCM-Chem simulations for the 10 summers (1996-2005). Simulations are initialized each April, and we analyze output from June through August to allow for a 2 month

spin-up phase. Meteorological initial conditions (IC) and boundary conditions (BC) driving the NRCM-Chem simulations are provided by a NRCM 36 km domain simulation described in Done et al. (2015). Chemical IC and BC for trace gases and aerosols were taken from a global simulation with the Community Atmosphere Model with Chemistry (CAM-Chem V4) detailed in Lamarque et al. (2011). Each present time NRCM-Chem year uses the same chemical IC and BC based on the CAM-Chem output for the year 2000. More details about the simulations can be found in Pfister et al. (2014).

2.3. Selecting meteorological predictors

To compare the sensitivities to meteorology between observed and simulated O_3, we must choose meteorological predictor variables which are available both in NARR and NRCM-Chem output. Based on results from previous studies, we select five meteorological predictors of interest (see Table 1). These variables represent a subset of those found by Otero et al. (2016), Porter et al. (2015), and Russell et al. (2016b) to be key drivers of high or extreme observed summer O_3. To examine the relationship between meteorology and MDA8 O_3, which is a daily quantity, daily summary measures are chosen for each predictor variable. For consistency between NRCM-Chem and NARR output, which is available as 3-hour means, we first convert the NRCM-Chem output to 3-hour means before taking the daily maximum.

Table 1: Meteorological predictors and corresponding daily summary measures used in the analysis, for both NARR and NRCM-Chem outputs.

Meteorological predictor	**Abbreviation**	**Definition**
Air temperature at 2m	T	Daily maximum
Wind speed at 10m	WS	Daily mean
Relative humidity	RH	Daily mean
Height of the planetary boundary layer	HBL	Daily maximum
Downward shortwave radiation flux	DSR	Daily maximum

3. Statistical methods

3.1. Marginal analysis of extreme O_3

In addition to using standard summary statistics to compare the distributions of MDA8 O_3 between observations and NRCM-Chem simulations, we employ extreme value theory to analyze the tails of these distributions. We use the generalized Pareto distribution (GPD) to model exceedances of a sufficiently high threshold u. This model is asymptotically motivated, as the GPD is the limiting distribution of appropriately normalized threshold excesses (Coles, 2001). It is frequently applied because it offers greater efficiency of data usage over block-maxima approaches, and has been used previously to model the tail behavior of O_3 (e.g. Phalitnonkiat et al., 2016; Rieder et al., 2013). The GPD is parameterized by scale and shape parameters $\sigma_u > 0$ and ξ, and can be defined by

$$Pr(X \leq x \mid X > u) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left[-\left(\frac{x-u}{\sigma_u}\right)\right]_+, & \xi = 0, \end{cases} \tag{1}$$

where $y_+ = \max(y, 0)$. When $\xi < 0$ there is an upper limit such that $u < x < u - \sigma_u/\xi$, i.e. the tail is bounded. $\xi = 0$ and $\xi > 0$ correspond to light and heavy tails, respectively.

To maintain a consistent approach among our analyses, for each series we choose our threshold, u, such that approximately 5% of the O_3 values exceed it. Standard diagnostics such as the mean residual life plot (Coles, 2001) confirm that this threshold appears to be high enough that the limiting GPD is a good approximation for the exceedance distribution, while at the same time this threshold retains a reasonable number of exceedances for the analysis. As a result of emissions controls, concentrations of surface O_3 have been decreasing over much of the US in recent years (Lefohn et al., 2008). We see this downward trend in observed O_3 at the Atlanta and Houston stations (see Figure 2), and account for this non-stationarity by setting a linearly-varying threshold in time, u_y, via 0.95 quantile regression (Koenker and Bassett Jr, 1978). The quantile regression coefficient for year is significantly less than zero at Atlanta and Houston (point estimates and standard errors are given in Table 2). Because the NRCM-Chem simulations use anthropogenic emission inputs from the year 2000 for the entire time period, we do not observe the same downward trend as in the observations, and thus employ a constant threshold u for simulated O_3 which is the empirical 0.95 quantile over the entire series at a given location. Given the threshold estimate, GPD

parameters are estimated by maximum likelihood, and standard errors are obtained via standard likelihood-based procedures. These standard error estimates do not take into account threshold uncertainty.

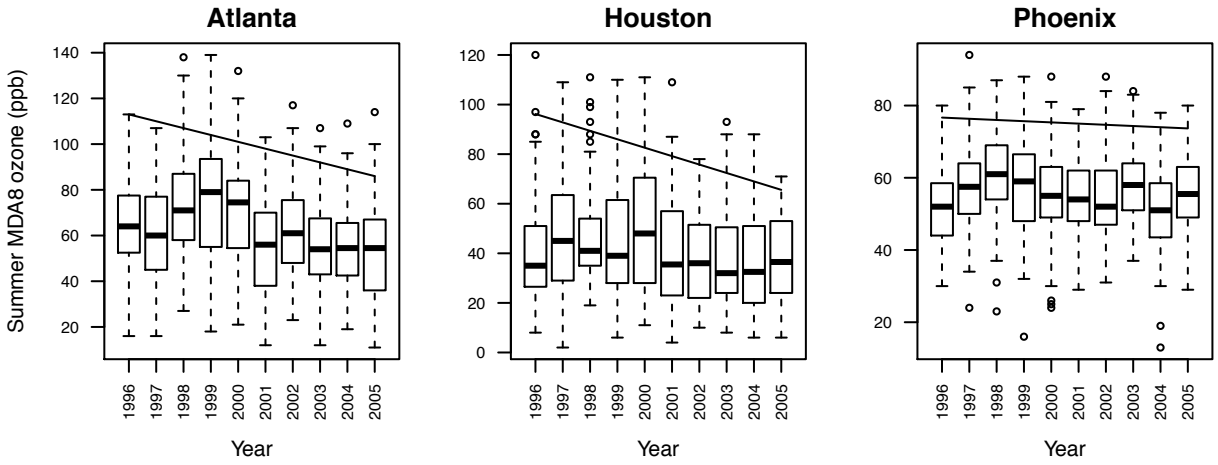

Figure 2: Distribution of observed summer (JJA) MDA8 O_3 by year at three AQS monitoring stations. The 0.95 quantile regression line represents the linearly-varying threshold in time used for the marginal analysis of extreme O_3. The trend in year is significantly less than zero for Atlanta and Houston stations, but not for the Phoenix station.

The usual likelihood formed by the product of GPD densities assumes independence of threshold excesses. However, initial examination of the O_3 series reveals short-term temporal dependence in the exceedances – if O_3 concentration exceeds the threshold today, it is more likely to exceed the threshold tomorrow compared to if it did not exceed today. Fitting the GPD to all exceedances using the usual likelihood in the presence of such serial correlation would result in underestimated standard errors. We avoid this issue by declustering the excesses prior to model fitting. We use the intervals method proposed by Ferro and Segers (2003) to estimate run length, and then apply runs declustering (Leadbetter et al., 1989) with clusters restricted to occur within the same year. Once a cluster is identified, it is replaced with the cluster maximum. The GPD is fit to the declustered series, with parameters computed via numerical maximum likelihood estimation. Sample sizes after declustering are given in Table 2. Analyses are done using the `extRemes` package (Gilleland and Katz, 2016) in `R` (R Core Team, 2015). Using the fitted GPD, we can estimate high quantiles of the O_3 distributions. In this study we report estimates of the 0.99 quantile, with confidence intervals obtained by profile likelihood to account for asymmetry in the likelihood surface.

3.2. Relating high and extreme O_3 to meteorological drivers

We use two methods to examine the sensitivities of high or extreme O_3 to the selected meteorological predictors: quantile regression and the tail dependence optimization method developed by Russell et al. (2016b). The two frameworks are described below. In both approaches, we fit statistical models relating (a) NRCM-Chem O_3 to NRCM-Chem meteorology and (b) observed O_3 to NARR meteorology. The fitted models include the five meteorological predictor variables found in Table 1 for both NRCM-Chem and NARR, allowing us to compare the estimated model coefficients which represent the sensitivities of the O_3 response to the meteorological drivers.

3.2.1. Quantile regression

In contrast to ordinary least squares regression, which models the linear relationship between one or more predictor variables $\mathbf{X}$ and the conditional *mean* of a response variable Y given $\mathbf{X} = \mathbf{x}$, quantile regression (Koenker and Bassett Jr, 1978) extends the regression model to conditional *quantiles* of the response Y given $\mathbf{X} = \mathbf{x}$. For $\tau \in (0, 1)$, we define the τth *conditional quantile* of Y by

$$Q_{Y|\mathbf{X}}(\tau) = \inf\{y : Pr(Y \leq y|\mathbf{X} = \mathbf{x}) \geq \tau\}.$$

Our model assumes a linear relationship between the conditional quantile and the p predictors, i.e.

$$Q_{Y|\mathbf{x}}(\tau) = \mathbf{x}^T\boldsymbol{\alpha}(\tau) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p. \tag{2}$$

The coefficients $\boldsymbol{\alpha}(\tau) = (\alpha_0, \alpha_1, \ldots, \alpha_p)$ of the linear conditional quantile function can be estimated by solving

$$\hat{\boldsymbol{\alpha}}(\tau) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\alpha}), \tag{3}$$

where $\rho_\tau(\cdot)$ represents the check function $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ and $\mathbb{I}(\cdot)$ is the indicator function. In this study, $\tau = 0.95$ because we are interested in a high level of ozone.

Because the distributions of O_3 and meteorology may differ between NRCM-Chem simulations and observational products (see Sections 4.1 and 4.2), we center and scale both the O_3 response and each of the meteorological predictors so as to be able to compare the estimated coefficients between the two analyses. We also center the year variable so that the intercept is at the year 2000. We implement quantile regression using the `quantreg` package (Koenker, 2016) in `R`, with standard errors obtained by paired bootstrap. Specifically, we fit a model for the conditional 0.95

quantile with all five meteorological main effects. Note that these quantile regression models are different from the quantile regression used for threshold estimation in Section 3.1, which included only year as a predictor to account for non-stationarity in the tail.

3.2.2. Tail dependence optimization

Quantile regression is not well-suited to modeling extremely high quantiles for which there may be inadequate data above the desired quantile for quantile regression estimation methods to succeed. To understand the meteorological variables associated with the highest O_3 levels, we apply the method developed by Russell et al. (2016b) to find the linear combination of a set of meteorological predictors which has the strongest tail dependence with the O_3 response. This approach is based on multivariate (in this case, bivariate) regular variation, which is a framework used for characterizing multivariate extremes.

The procedure of Russell et al. (2016b) aims to optimize a metric of tail dependence γ. Because the regular variation framework requires heavy-tailed marginals, the procedure requires transformation of both the response and predictor functional. Let Y_t be the random variable representing the response at time t, and let $X_{t,i}$ be value of the ith predictor at time t, for $i = 1, \ldots, k$. First we transform the response to be approximately unit Fréchet by letting $Y_t^{**} = G^{-1}[\hat{F}_Y(Y_t)]$ where G is the unit Fréchet distribution function and $\hat{F}_Y$ is the estimated marginal distribution of Y_t. Next we apply a two-step transformation procedure to the predictors. In the first step, each predictor is transformed to the standard Gaussian scale using $X_{t,i}^* = \Phi^{-1}[\hat{F}_{X_i}(X_{t,i})]$ where Φ is the standard Gaussian distribution function. We consider linear combinations of the form $\boldsymbol{X_t^{*\prime}\beta} = \beta_1 X_{t,1}^* + \cdots + \beta_k X_{t,k}^*$, where $\mathbf{X_t^*} = (X_{t,1}^*, \ldots, X_{t,k}^*)$ and $\boldsymbol{\beta}$ is constrained such that $\boldsymbol{\beta'}\mathrm{Cov}(\mathbf{X_t^*})\boldsymbol{\beta} = 1$ to ensure identifiability. In the second step, this linear combination is transformed to be approximately unit Fréchet using $X_t^{**}(\boldsymbol{\beta}) = G^{-1}[\Phi(\boldsymbol{X_t^{*\prime}\beta})]$.

Our modeling framework assumes the random vector $(X_t^{**}(\boldsymbol{\beta}), Y_t^{**})$ is bivariate regularly varying, and we seek the vector of coefficients $\tilde{\boldsymbol{\beta}}$ whose linear combination has the highest degree of tail dependence with the response. As $\gamma = 0$ corresponds to perfect asymptotic dependence, while $\gamma = 1$ corresponds to asymptotic independence (Russell et al., 2016b), we find

$$\tilde{\boldsymbol{\beta}} = \underset{\{\boldsymbol{\beta} \in \mathbb{R}^k : \boldsymbol{\beta'}\mathrm{Cov}(\mathbf{X_t^*})\boldsymbol{\beta} = 1\}}{\arg\min} \hat{\gamma}(\boldsymbol{\beta}),$$

where the estimator

$$\hat{\gamma}(\boldsymbol{\beta}) = \frac{\sum_{t=1}^{n} \delta(x_t^{**}(\boldsymbol{\beta}) + y_t^{**}) \frac{|x_t^{**}(\beta) - y_t^{**}|}{x_t^{**}(\beta) + y_t^{**}}}{\sum_{t=1}^{n} \delta(x_t^{**}(\boldsymbol{\beta}) + y_t^{**})}, \tag{4}$$

and $\delta : \mathbb{R}_+ \to [0, 1]$ is a non-decreasing weighting function. More details can be found in Russell et al. (2016b). Russell (2015) found that tail dependence optimization outperformed regression approaches, including quantile regression and logistic regression, as well as other extreme value approaches in terms of concordance in the upper tail. One disadvantage, however, is the large uncertainty in parameter estimates inherent to this and other extreme value methods.

We obtain 95% confidence intervals for parameter estimates using paired bootstrap and the percentile method (Givens and Hoeting, 2012). Model comparison can be achieved via cross-validation. Specifically, we use 10-fold cross-validation, in which the data is partitioned into 10 subsets. For each fold, the optimization is done on the training set (90% of the data) and $\hat{\gamma}$ is calculated for the test set (the remaining 10% of the data). The cross-validation score $\hat{\gamma}_{CV}$ is then the average over all 10 test sets.

4. Results

4.1. Comparing tails of O_3 response

Having implemented the procedure described in Section 3.1, Figure 3 compares the distribution of summer MDA8 O_3 between observations and NRCM-Chem simulations at our three study locations. In each panel between boxplots, the estimated 0.99 quantile for the year 2000 is shown with the corresponding 95% profile likelihood confidence interval. The 0.99 quantile roughly corresponds to the annual 4th highest MDA8, which forms the basis of the NAAQS for O_3. These extreme quantile estimates are made using the GPD fit to each series. The fitted GPD parameters are given in Table 2. In Atlanta and Phoenix, we see relatively good correspondence between observations and NRCM-Chem simulations, and 0.99 quantile estimates are not significantly different. In Houston, there is a noticeable difference in the upper tail, and the 0.99 quantile estimate is significantly lower for simulated O_3. This result is consistent with the tendency of regional air quality models to underpredict the high O_3 events, as found by Im et al. (2015) for example.

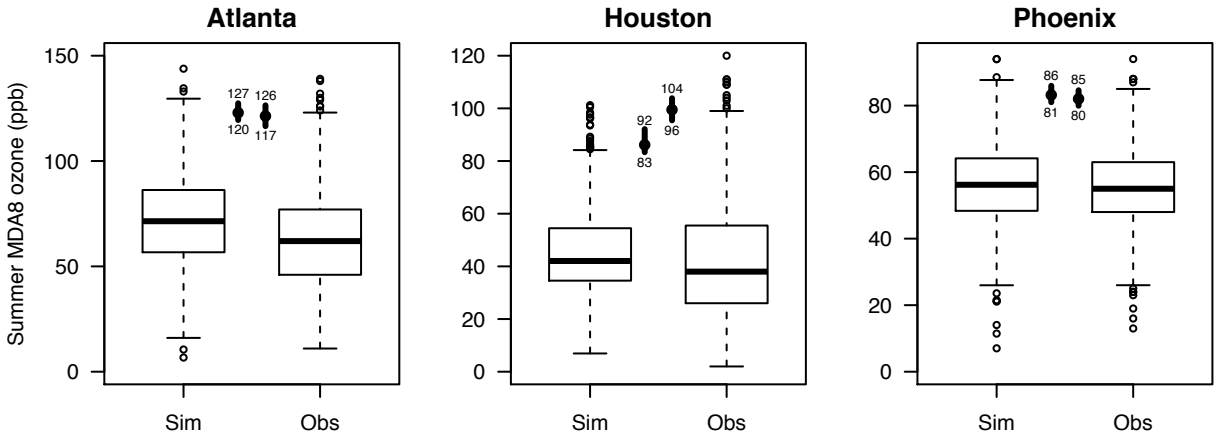

Figure 3: Boxplots of summer MDA8 O_3 during the years 1996-2005 from NRCM-Chem simulations (Sim) and AQS observations (Obs) at the three study locations. In each panel between the boxplots is the 0.99 quantile for the year 2000 estimated by fitting a GPD to threshold exceedances of simulations (left) and observations (right). Upper and lower limits are given for the corresponding 95% profile likelihood confidence intervals of each quantile estimate.

Table 2: GPD parameter estimates (standard errors in parentheses) for simulated (Sim) and observed (Obs) summer MDA8 O_3 at the three study locations. σ_u is the scale parameter, ξ is the shape parameter, u is the threshold, and n_{exc} is the number of exceedances after declustering. For Sim, u is set to the empirical 0.95 quantile. For Obs, u_y is a linearly-varying threshold in time with 0.95 quantile regression coefficients α_0 and α_1, where the intercept α_0 represents the threshold for the year 2000. Standard errors for threshold parameters are obtained via bootstrapping.

		Atlanta	**Houston**	**Phoenix**
σ_u	**Sim**	11.88 (2.50)	7.84 (2.56)	6.36 (1.70)
	Obs	23.91 (5.20)	17.09 (3.89)	6.03 (1.62)
ξ	**Sim**	-0.26 (0.13)	0.00 (0.30)	-0.19 (0.20)
	Obs	-0.66 (0.17)	-0.52 (0.18)	-0.20 (0.21)
u	**Sim**	109.74 (1.80)	76.62 (1.07)	76.11 (0.72)
	Obs α_0	101.00 (1.74)	82.60 (1.75)	75.33 (0.81)
	Obs α_1	-3.00 (0.56)	-3.40 (0.61)	-0.33 (0.34)
n_{exc}	**Sim**	34	31	32
	Obs	32	37	32

4.2. Comparing meteorological predictors

As in Section 4.1, we compare the *distributions* of meteorological variables. We do this because the NRCM-Chem simulations are not driven by reanalysis. We find that the distributions of the

selected meteorological predictors differ considerably between NRCM-Chem and NARR output. NRCM-Chem tends to underestimate daily maximum air temperature and daily mean relative humidity, and exhibits much larger variability in these predictors than seen in the NARR product (Figure 4 top, center). In Atlanta, for example, the summer median for relative humidity according to NRCM-Chem is 55% compared to 74% based on NARR. In Phoenix, the summer median for relative humidity is 16% in NRCM-Chem vs. 23% in NARR, however NRCM-Chem records a summer maximum of 79% daily mean relative humidity compared to NARR which has a maximum value of 52%. NRCM-Chem also tends to underestimate daily maximum height of the planetary boundary layer compared to the NARR product (Figure 4 bottom).

The discrepancy between the meteorology in NRCM-Chem and NARR is not explained by their difference in spatial resolution. We explored taking the average of the nine NRCM-Chem grid cells surrounding each location, to obtain 36 km resolution similar to NARR's 32 km resolution. However, the NRCM-Chem simulations are so strongly correlated between neighboring grid cells that the results are extremely similar to what is shown in Figure 4.

4.3. Comparing relationships between O_3 and meteorology

4.3.1. Quantile regression

The left column of Figure 5 presents the estimated coefficients of the fitted 0.95 quantile regression models at each location for two analyses. The first analysis (in triangles) relates NRCM-Chem O_3 to NRCM-Chem meteorology, and the second (in circles) relates observed O_3 to NARR meteorology. Year is included as a predictor for observed O_3, to account for the downward trend observed in Figure 2. As expected, we find a significant negative year trend for observed O_3 in both Atlanta and Houston that is not present in the NRCM-Chem simulations (not shown). Coefficients are shown for the five meteorological predictors included as main effects in the full model. Some coefficients are not significantly different from zero, as indicated by the 95% confidence interval intersecting zero. We explored using backwards stepwise selection to remove nonsignificant predictors, however we found that in all cases the full model had the best (lowest) Akaike information criterion (AIC) value, so we report results for this model.

For both analyses across the three study locations, in most cases we see that daily mean wind speed (WS) and relative humidity (RH) have negative effects on the 0.95 quantile of MDA8 O_3. In Atlanta and Phoenix, daily maximum air temperature (T) has a positive effect. (The negative

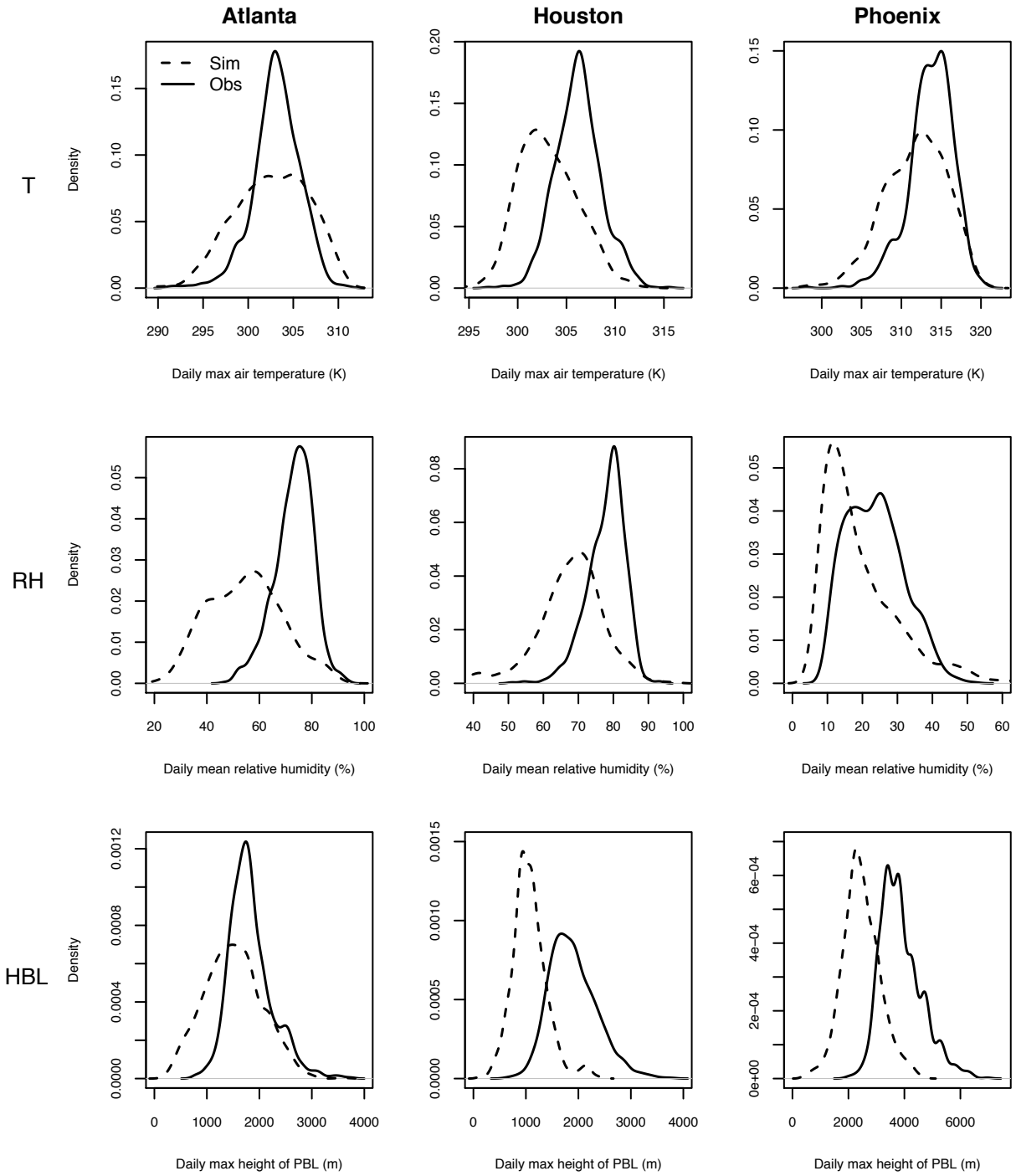

Figure 4: Kernel density plots of NRCM-Chem simulated (Sim, dashed lines) and NARR (Obs, solid lines) daily maximum air temperature at 2m (top row), daily mean relative humidity (center row), and daily maximum height of the planetary boundary layer (bottom row) for summers (JJA) during 1996-2005 at the three study locations.

coefficient for T in Houston is evidence of multicollinearity, as a quantile regression model including only T results in a positive coefficient for T.) When the daily maximum height of the planetary boundary layer (HBL) is significant, it appears to have a positive effect. Daily maximum downward shortwave radiation flux (DSR) does not have a significant effect in any of the fitted models.

At each location, we find differences in the fitted full model for NRCM-Chem simulated vs. observed O_3. These differences are not consistent across study locations. In Atlanta, T has a significant (positive) effect on observed O_3, however it is borderline nonsignificant for NRCM-Chem simulations. WS has the strongest (negative) effect on simulated O_3, but not a significant effect on observed O_3. In Houston, in contrast, WS has a significantly more negative effect on observed O_3. Unlike at other locations, there is a similar (significant, positive) effect of HBL on both simulations and observations in Houston. In Phoenix, the largest difference is in relative humidity: RH has a significant negative effect on simulated O_3, but a significant positive effect on observed O_3, conditional on the other predictors. T has a significant effect for simulations but not for observations, while HBL is significant for observations but not for simulations. In contrast to the other two locations, there is a similar negative effect of WS on both simulated and observed O_3 in Phoenix.

4.3.2. Tail dependence optimization

Unlike quantile regression, where we directly model the effect of year for observed O_3, the tail dependence optimization method of Russell et al. (2016b) requires stationary data. To account for non-stationarity in observed O_3, we transform the response variable by using 0.95 quantile regression to obtain the linearly-varying threshold in time as in Section 3.1. We then fit a gamma distribution to observations below and a GPD to observations above this year-varying 0.95 quantile, as explained in Russell et al. (2016b). This detrended response is then transformed to unit Fréchet as required by the method.

The right column of Figure 5 presents parameter estimates with bootstrap confidence intervals (based on 1000 bootstrap replicates) for tail dependence optimization applied to the two analyses at each of the three study locations. Similar to quantile regression, across all locations and analyses we see that T tends to have a positive relationship with extreme O_3, while WS tends to have a negative relationship. When RH is found to be significant, it has a negative relationship with extreme O_3. These three predictors have significant effects in at least some cases, while confidence

intervals for HBL and DSR cover zero in all cases. Therefore, in addition to the full model with all five meteorological variables, we also fit a model with only T, WS, and RH as predictors (see Figure 5 results in gray). In all cases this improves (lowers) the cross-validation score $\hat{\gamma}_{CV}$.

Parameters obtained by tail dependence optimization are less straightforward to interpret than those obtained by quantile regression. For a given model fit, we can compare relative magnitudes and signs of the estimated parameters. Some differences are evident between model fits for simulated vs. observed extreme O_3. In Atlanta, as was the case for quantile regression, there appears to be a more negative effect of WS on simulated than observed O_3. In Houston, RH has the strongest effect for simulated O_3, while T and WS appear to have stronger effects on observed O_3. In almost all cases, the point estimate for RH differs in sign between analyses for simulated and observed O_3, although the bootstrap confidence intervals are too wide to conclude any significant difference. In all three locations, DSR is estimated to have a negative effect on simulated O_3 and a positive effect on observed O_3, though again we are not able to conclude a significant difference.

5. Summary and Discussion

In this case study of summer surface O_3 in three US cities, we employ a set of high resolution NRCM-Chem simulations to make three types of comparisons between simulated and observational data, comparing (1) tails of the O_3 response, (2) distributions of meteorological predictor variables, and (3) sensitivities of high and extreme O_3 to meteorological predictors. This last comparison is made using both quantile regression, for the 0.95 quantile of O_3, and the tail dependence method of Russell et al. (2016b), which is used to investigate even higher O_3 extremes. To our knowledge, ours is the first study to apply quantile regression and tail dependence optimization to O_3 simulated from an atmospheric chemistry model. Additionally, this is the first study to compare the meteorological sensitivities of high/extreme O_3 between simulations and observational data.

Results from comparing the distributions of the O_3 response variable show that NCRM-Chem represents O_3 adequately overall, but underestimates extreme quantiles of O_3 in Houston. Results from comparing the distributions of meteorological predictors show clear discrepancies between the meteorology produced by NARR and that found in the NRCM-Chem simulations at all three locations. We recognize that NARR, being a reanalysis product, will not exactly match weather station data. There are further questions about NARR, for example the diagnostic parameter HBL

is likely too high in NARR, e.g. a comparison with the MERRA reanalysis found that NARR is more than 500m higher over the western US (McGrath-Spangler and Denning, 2012). However, we see surprising meteorology produced by NRCM-Chem, for instance the very low RH levels in Atlanta. These drier model conditions could increase the lifetime of O_3, as lower water vapor leads to reduced loss of O_3. The bias in the NRCM-Chem meteorology could be at least partially due to known SST errors in the model runs (Pfister et al., 2014).

Comparisons of the sensitivities of high and extreme O_3 to meteorological drivers also show clear differences between simulations and observational data. These differences are not consistent across the three study locations. For both high and extreme O_3 in Atlanta, simulations significantly overpredict the strength of the (negative) effect of WS. For the 0.95 quantile in Houston, we see the opposite, in that simulations significantly underpredict the effect of WS. In Phoenix, the quantile regression coefficient for RH is negative for simulated O_3, but positive for observed O_3. We also see a sign difference in the Phoenix point estimates for RH from the tail dependence method, however the confidence intervals are too large to conclude significance.

Differences in the sensitivities of observed vs. simulated O_3 could be driven by differences in how meteorological variables interact with O_3 formation and removal processes. For example, previous studies of average MDA8 O_3 over the eastern US have found that air quality models underpredict the strength of the effects of T and RH (Davis et al., 2011; Rasmussen et al., 2012). In Atlanta, we similarly find that our 0.95 quantile regression coefficients underestimate the effects of T and RH, though not significantly. Kavassalis and Murphy (2017) suggest that such a discrepancy may result from the lack of vapor pressure deficit-dependent dry deposition in the chemical transport model. Differences in Houston may be attributed to the difficulty in representing coastal dynamics such as recirculation patterns (e.g., Russo et al., 2016). Coastal areas often show a diurnal cycle in wind patterns, which in some cases can lead to either stagnancy, or the recirculation of polluted air away and then back to the original location. Poorly representing these coastal wind patterns, and how the observed recirculation or stagnancy affect O_3 levels, could explain the discrepancies in sensitivities for both WS and RH.

An important finding of this study is that the distribution of simulated O_3 matches observed O_3 quite well at two out of the three locations, despite rather large differences – and in some cases even sign reversal – in the meteorological sensitivities. It is possible that the differences in modeled

and observed sensitivities are superficial, and that the underlying mechanisms leading to extreme O_3 formation and loss are still being represented, even if attribution is not identical between model and observation. It may be that the linearity assumption inherent to both quantile regression and tail dependence optimization methods is too simple to capture the complex relationship between O_3 and meteorology. In addition, models with multiple predictors face issues of collinearity in the predictors which increase the uncertainty. Future work could relax linearity assumptions or investigate interaction effects between predictor variables.

However, our finding raises a concern for modelers that the O_3 distributions are matching up well for the wrong reasons, due to parameter tuning within the model. Modeled chemistry related to O_3 formation, for example, has been steadily evolving and improving, but some of the improvements actually worsen agreement with observations because other processes are not included yet, or else have been misrepresented (see, e.g., Porter et al., 2017; Sherwen et al., 2016). If we seek modeling tools that can adapt to changing emissions and climatology, it is important to not only capture the current pollutant distribution, but also the relationships between the pollutant and its drivers. Our results suggest that, even in the locations where O_3 seems to be fairly well represented, NCRM-Chem may not be accurately representing the mechanisms behind O_3 formation or loss. Correctly describing current levels of O_3, while failing to capture the key mechanisms responsible, implies that our predictions will be unable to adapt to a changing climate. The poor agreement of meteorological sensitivities may evidence a need for mechanism improvement, either in terms of chemistry or physical dynamics.

We have proposed and applied a framework for comparing the meteorological sensitivities of high/extreme O_3 between observed data and simulated output. While this study analyzes only one atmospheric chemistry model, our methodology could be applied to any pairs of observational and simulated O_3-meteorology data. Despite having only 10 years of data which is a very short record for an extreme value analysis, we find important differences between the observed and simulated O_3, the driving meteorology, and the sensitivities linking these. However, there are large uncertainties in parameter estimates, as evidenced by the wide confidence intervals in Figure 5. Such uncertainty is inherent to extremes approaches which focus on the most extreme values and thus use only a small subset of the data. In addition, this case study was a detailed analysis of only a few urban locations. Future work could consider aggregating results or conducting a spatial

analysis over a larger region, as borrowing strength across locations could reduce uncertainties in parameter estimates (Russell et al., 2016a).

Acknowledgements

This work was supported by the US Environmental Protection Agency's Science to Achieve Results (EPA-STAR) program. Although this article was developed under grant/cooperative agreement RD-83522801-0, it has not been formally reviewed by the EPA and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. The NARR data for this study are from the Research Data Archive, which is maintained by the Computational and Information Systems Laboratory at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. We also thank two anonymous reviewers for helpful comments on an earlier draft of this article.

Baur, D., Saisana, M., Schulze, N., 2004. Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach. Atmospheric Environment 38 (28), 4689–4699.

Camalier, L., Cox, W., Dolwick, P., 2007. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. Atmospheric Environment 41 (33), 7127–7137.

Coles, S., 2001. An introduction to statistical modeling of extreme values. Springer-Verlag.

Davis, J., Cox, W., Reff, A., Dolwick, P., 2011. A comparison of CMAQ-based and observation-based statistical models relating ozone to meteorological parameters. Atmospheric Environment 45 (20), 3481–3487.

Dawson, J. P., Adams, P. J., Pandis, S. N., 2007. Sensitivity of ozone to summertime climate in the eastern usa: A modeling case study. Atmospheric Environment 41 (7), 1494–1511.

Done, J. M., Holland, G. J., Bruyère, C. L., Leung, L. R., Suzuki-Parker, A., 2015. Modeling high-impact weather and climate: lessons from a tropical cyclone perspective. Climatic Change 129 (3-4), 381–395.

Ferro, C. A., Segers, J., 2003. Inference for clusters of extreme values. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 (2), 545–556.

Gilleland, E., Katz, R. W., 2016. extRemes 2.0: An extreme value analysis package in R. Journal of Statistical Software 72 (8), 1–39.

Givens, G. H., Hoeting, J. A., 2012. Computational Statistics. Vol. 710. John Wiley & Sons.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., Eder, B., 2005. Fully coupled "online" chemistry within the WRF model. Atmospheric Environment 39 (37), 6957–6975.

IHME, 2013. Global burden of disease study 2010 results by cause 1990-2010 - country level. Tech. rep., Institute for Health Metrics and Evaluation, Seattle, United States.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., et al., 2015. Evaluation of operational on-line-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I: Ozone. Atmospheric Environment 115, 404–420.

Jacob, D. J., Winner, D. A., 2009. Effect of climate change on air quality. Atmospheric Environment 43 (1), 51–63.

Kavassalis, S., Murphy, J. G., 2017. Understanding ozone-meteorology correlations: a role for dry deposition. Geophysical Research Letters 44 (6), 2922–2931.

Koenker, R., 2016. quantreg: Quantile Regression.

Koenker, R., Bassett Jr, G., 1978. Regression quantiles. Econometrica: Journal of the Econometric Society, 33–50.

Lamarque, J.-F., Kyle, G. P., Meinshausen, M., Riahi, K., Smith, S. J., van Vuuren, D. P., Conley, A. J., Vitt, F., 2011. Global and regional evolution of short-lived radiatively-active gases and aerosols in the representative concentration pathways. Climatic Change 109 (1-2), 191–212.

Leadbetter, M., Weissman, I., De Haan, L., Rootzén, H., 1989. On clustering of high values in statistically stationary series. Proc. 4th Int. Meet. Statistical Climatology 16, 217–222.

Lefohn, A. S., Shadwick, D., Oltmans, S. J., 2008. Characterizing long-term changes in surface ozone levels in the United States (1980–2005). Atmospheric Environment 42 (35), 8252–8262.

McGrath-Spangler, E. L., Denning, A. S., 2012. Estimates of North American summertime planetary boundary layer depths derived from space-borne lidar. Journal of Geophysical Research: Atmospheres 117 (D15).

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E. H., et al., 2006. North American regional reanalysis. Bulletin of the American Meteorological Society 87 (3), 343–360.

Ordónez, C., Mathis, H., Furger, M., Henne, S., Hüglin, C., Staehelin, J., Prévôt, A., 2005. Changes of daily surface ozone maxima in Switzerland in all seasons from 1992 to 2002 and discussion of summer 2003. Atmospheric Chemistry and Physics 5 (5), 1187–1203.

Otero, N., Sillmann, J., Schnell, J. L., Rust, H. W., Butler, T., 2016. Synoptic and meteorological drivers of extreme ozone concentrations over Europe. Environmental Research Letters 11 (2).

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

Pfister, G., Walters, S., Lamarque, J.-F., Fast, J., Barth, M., Wong, J., Done, J., Holland, G., Bruyère, C., 2014. Projections of future summertime ozone over the US. Journal of Geophysical Research: Atmospheres 119 (9), 5559–5582.

Phalitnonkiat, P., Sun, W., Grigoriu, M. D., Hess, P., Samorodnitsky, G., 2016. Extreme ozone events: Tail behavior of the surface ozone distribution over the US. Atmospheric Environment 128, 134–146.

Porter, W., Heald, C., Cooley, D., Russell, B., 2015. Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression. Atmospheric Chemistry and Physics 15 (18), 10349–10366.

Porter, W. C., Safieddine, S. A., Heald, C. L., 2017. Impact of aromatics and monoterpenes on simulated tropospheric ozone and total OH reactivity. Atmospheric Environment 169, 250–257.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, D., Fiore, A., Naik, V., Horowitz, L., McGinnis, S., Schultz, M., 2012. Surface ozone-temperature relationships in the eastern US: A monthly climatology for evaluating chemistry-climate models. Atmospheric Environment 47, 142–153.

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. Environmental Research Letters 8 (1), 014012.

Russell, B. T., 2015. Understanding extreme behavior by optimizing tail dependence with application to ground level ozone via data mining and spatial modeling. Ph.D. thesis, Colorado State University. Libraries.

Russell, B. T., Cooley, D. S., Porter, W. C., Heald, C. L., 2016a. Modeling the spatial behavior of the meteorological drivers' effects on extreme ozone. Environmetrics 27 (6), 334–344.

Russell, B. T., Cooley, D. S., Porter, W. C., Reich, B. J., Heald, C. L., 2016b. Data mining to investigate the meteorological drivers for extreme ground level ozone events. The Annals of Applied Statistics 10 (3), 1673–1698.

Russo, A., Gouveia, C., Levy, I., Dayan, U., Jerez, S., Mendes, M., Trigo, R., 2016. Coastal recirculation potential affecting air pollutants in Portugal: The role of circulation weather types. Atmospheric Environment 135 (9–19).

Seinfeld, J. H., Pandis, S. N., Noone, K., 1998. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley & Sons, Inc.

Sherwen, T., Schmidt, J. A., Evans, M. J., Carpenter, L. J., Großmann, K., Eastham, S. D., Jacob, D. J., Dix, B., Koenig, T. K., Sinreich, R., et al., 2016. Global impacts of tropospheric halogens (Cl, Br, I) on oxidants and composition in GEOS-Chem. Atmospheric Chemistry and Physics 16 (18), 12239–12271.

Tai, A. P., Mickley, L. J., Jacob, D. J., 2010. Correlations between fine particulate matter (PM 2.5) and meteorological variables in the United States: Implications for the sensitivity of PM 2.5 to climate change. Atmospheric Environment 44 (32), 3976–3984.

Vautard, R., Beekmann, M., Desplat, J., Hodzic, A., Morel, S., 2007. Air quality in Europe during the summer of 2003 as a prototype of air quality in a warmer climate. Comptes Rendus Geoscience 339 (11), 747–763.

Wilson, A., Rappold, A. G., Neas, L. M., Reich, B. J., et al., 2014. Modeling the effect of temperature on ozone-related mortality. The Annals of Applied Statistics 8 (3), 1728–1749.

Zhao, W., Fan, S., Guo, H., Gao, B., Sun, J., Chen, L., 2016. Assessing the impact of local meteorological variables on surface ozone in Hong Kong during 2000–2015 using quantile and multiple line regression models. Atmospheric Environment 144, 182–193.

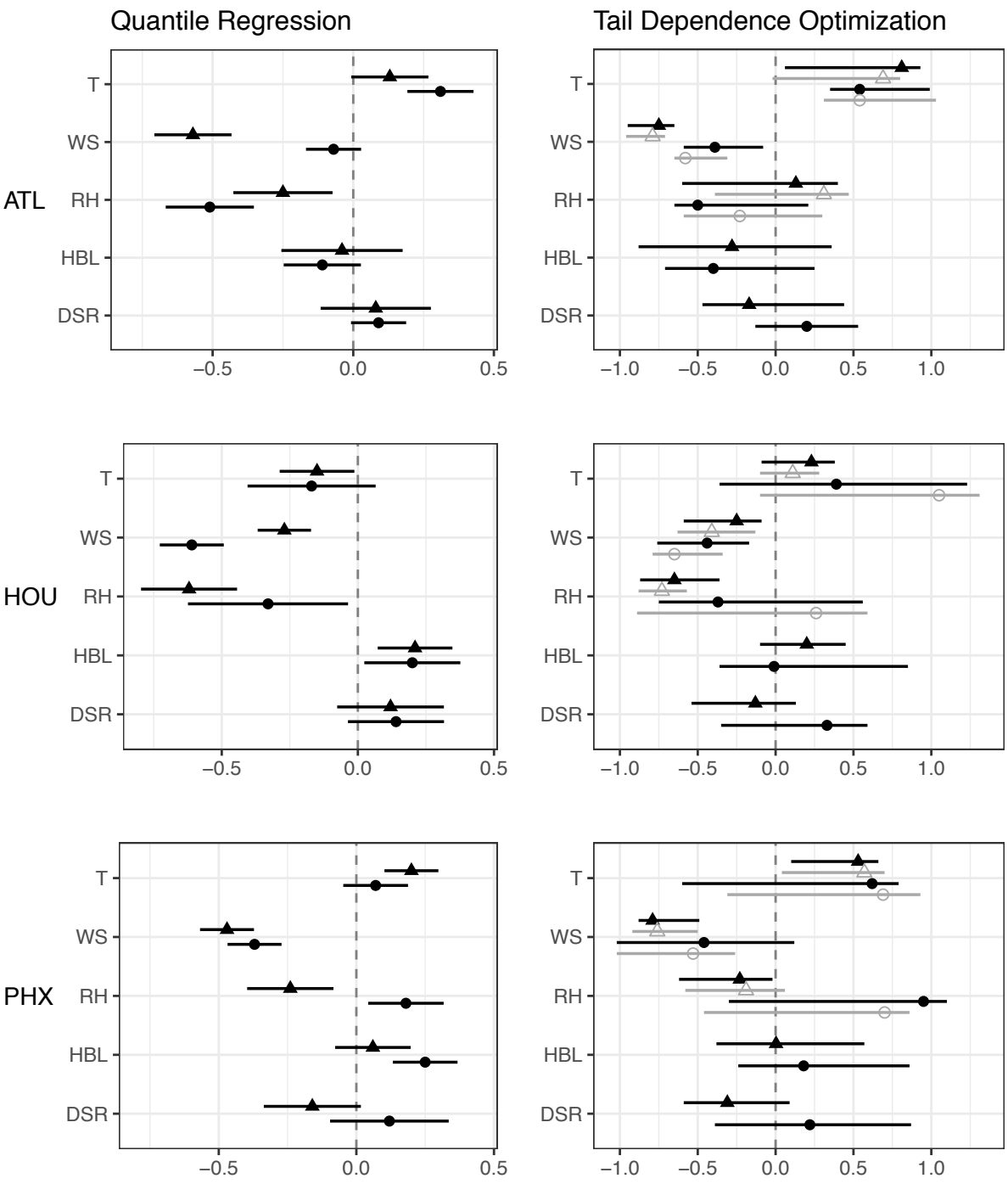

Figure 5: Parameter estimates with 95% confidence intervals for 0.95 quantile regression (left column) and tail dependence optimization (right column) at Atlanta (top), Houston (center), and Phoenix (bottom). Triangles indicate estimates from models relating NRCM-Chem O_3 to NRCM-Chem meteorology, while circles indicate estimates from models relating observed O_3 to NARR meteorology. Estimates in gray correspond to a reduced tail dependence model including only T, WS, and RH as predictors.