NIH Public Access
Author Manuscript
Ann Appl Stat. Author manuscript; available in PMC 2014 February 26.

Published in final edited form as:
Ann Appl Stat. 2013 June 1; 7(2): 739–762. doi:10.1214/13-AOAS628.

Extreme value analysis for evaluating ozone control strategies

Brian Reich[1], **Daniel Cooley**[2], **Kristen Foley**[3], **Sergey Napelenok**[3], and **Benjamin Shaby**[4]
[1]North Carolina State University

[2]Colorado State University

[3]US Environmental Protection Agency

[4]University of California - Berkeley

Abstract

Tropospheric ozone is one of six criteria pollutants regulated by the US EPA, and has been linked to respiratory and cardiovascular endpoints and adverse effects on vegetation and ecosystems. Regional photochemical models have been developed to study the impacts of emission reductions on ozone levels. The standard approach is to run the deterministic model under new emission levels and attribute the change in ozone concentration to the emission control strategy. However, running the deterministic model requires substantial computing time, and this approach does not provide a measure of uncertainty for the change in ozone levels. Recently, a reduced form model (RFM) has been proposed to approximate the complex model as a simple function of a few relevant inputs. In this paper, we develop a new statistical approach to make full use of the RFM to study the effects of various control strategies on the probability and magnitude of extreme ozone events. We fuse the model output with monitoring data to calibrate the RFM by modeling the conditional distribution of monitoring data given the RFM using a combination of flexible semiparametric quantile regression for the center of the distribution where data are abundant and a parametric extreme value distribution for the tail where data are sparse. Selected parameters in the conditional distribution are allowed to vary by the RFM value and the spatial location. Also, due to the simplicity of the RFM, we are able to embed the RFM in our Bayesian hierarchical framework to obtain a full posterior for the model input parameters, and propagate this uncertainty to the estimation of the effects of the control strategies. We use the new framework to evaluate three potential control strategies, and find that reducing mobile-source emissions has a larger impact than reducing point-source emissions or a combination of several emission sources.

Keywords

Bayesian hierarchical modeling; Generalized Pareto distribution; Spatial data analysis; Statistical downscaling

1 Introduction

Due to advances in emissions control technology and regulatory action, air quality has been improving over the last several decades in the United States and Europe. However, areas exist where significant populations are still exposed to elevated levels of tropospheric ozone (O_3). Epidemiological and controlled human exposure studies have shown an association between O_3 exposure and respiratory and cardiovascular endpoints, particularly in sensitive populations (US EPA, 2006). Furthermore, ozone has been linked to a variety of adverse effects on vegetation and ecosystems (US EPA, 2006).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Ozone, together with other compounds, is formed downwind of its two main classes of precursors, volatile organic compounds (VOCs) and nitrogen oxides ($NO_x = NO + NO_2$) in the presence of sunlight. Due to the complexities of the formation processes of secondary pollutants such as ozone and their dependence on various physical and chemical parameters, as well as meteorological conditions, ozone is a highly non-linear function of its inputs, and thus regional three-dimensional Eulerian photochemical models have been developed to track the precursor emissions, transport, and chemical transformations of gases and particles in the troposphere. Such models are used to study the formation and geographical distribution of ozone and also to provide a test bed for possible control strategies. For example, we use the Community Multiscale Air Quality (CMAQ) model (Byun and Schere, 2006) to study changes in ozone levels due to changes in mobile-source NO_x emissions, point-source NO_x emissions, other NO_x emissions, anthropogenic VOCs emissions, and biogenic VOCs emissions. Due to the prevalence of ozone precursor emissions from a wide range of sources, differing chemical reactivities of the various specific chemical compounds that make up VOCs, and the varying cost of different control technologies, the processes of control strategy evaluation is itself complex.

Control strategy evaluation seeks to understand the potential response in air quality levels from various targeted reductions of source pollutants. By studying control strategies, we understand how resources should be allocated to achieve the best results. Control strategies must be studied via atmospheric chemistry models where it is possible to change emissions scenarios. In the context of control strategy evaluation, it is the response of modeled concentration that is desired as a surrogate for the response for actual atmospheric pollutant levels. Such response can be obtained most basically by performing two modeling simulations: one representing current conditions; and one representing conditions under a specific emissions control scenario. The difference in predicted concentrations from the two simulations could then be attributed to the control strategy. Currently, the computational costs of running full regional photochemical models are still nontrivial, making such evaluations potentially costly. To address this issue, various methods have been applied to develop reduced form models (RFMs) that represent pollutant concentrations for a particular episode of interest as a simple function of usually only the regulatory controllable parameters. For example, in the RFM, ozone only may be a function of parameters for the various anthropogenic sources NO_x and VOCs for a particular place and time. One of the various methods for developing RFMs to date is through calculating sensitivity coefficients of a target pollutant to emissions of precursors from controllable sectors. These sensitivities are subsequently used to make adjustments to pollutant concentrations predicted by a base model simulation (Digar et al., 2011; Napelenok et al., 2011).

We propose a new approach for combining an RFM with monitored point-level ozone data to study effects of various control strategies. Our approach is geared towards accurately characterizing extreme ozone events under different scenarios. Extreme ozone is a concern for health effects modeling and regulation. For example, the current EPA regulation of ozone is based on the fourth highest daily eight-hour average ozone (i.e., the maximum eight-hour average ozone concentrations for the day) of the year. Extreme value theory (EVT) (see Coles, 2001, for an overview) provides an asymptotically-justified approach to modeling the tails of wide classes of distributions. Our model employs a parametric form suggested by extreme value theory in the tail, while utilizing a flexible quantile regression approach to model the bulk of the distribution.

CMAQ output and monitor data are not directly comparable as CMAQ output is defined a volume average within a three dimensional grid cell and we wish to make inference about ozone at point locations. Thus we build a downscaler; that is, a statistical model that links gridded numerical model output to point-level observational data (see Berrocal et al., 2010,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and references therein). Most downscaling work employs a Gaussian process framework (e.g, Berrocal et al., 2010), and focuses on modeling the conditional mean of the observations given the numerical model output. For example, Foley et al. (2012) use this approach for ozone control-strategy evaluation. Some recent work has sought to move beyond the parametric Gaussian paradigm. For example, Reich et al. (2011) and Zhou et al. (2012) propose separate models for the quantile process of the model output and the quantile process of the point-located data, which provides a calibration function to link the two sources of data. Mannshardt-Shamseldin et al. (2010) propose an EVT-based downscaling method for extreme precipitation, relating the return levels (i.e., extreme quantiles, a climatological quantity) of the model output to return levels at point-located sites. Unlike the previous downscaling work for ozone which focused on calibration, our focus is prediction. Specifically once our downscaler is constructed to relate monitor data to RFM output under observed emissions, we use it to study the effect of alternative emission scenarios.

Other work which has appeared in the climate literature applies EVT to numerical model output in order to produce maps summarizing the extreme behavior of the studied variable. One approach (e.g., Kharin et al., 2007; Wehner, 2005; Wehner et al., 2010) is to fit extreme value distributions separately to the output at individual grid cells, perhaps employing spatial smoothing procedures after the pointwise fitting to produce the maps. Another approach (e.g., Cooley and Sain, 2010; Schliep et al., 2010) constructs hierarchical Bayesian models that pool information across space by incorporating spatial random effects. These works differ fundamentally in both aim and approach from the present work. These previous studies aim to describe the tail of the (unconditional) distribution and treat the numerical model output as data, whereas we aim to model the conditional distribution of the point-located measurements treating the model output as covariate information.

Our work is somewhat related to recent work which has sought to link extreme behavior to large-scale climatological conditions. Sillman et al. (2011) model the connection between extreme cold temperatures and atmospheric blocking conditions as produced by both reanalysis models and climate models. Maraun et al. (2011) link extreme precipitation to large-scale airflow. Similar to the model we present in Section 3, Sillman et al. (2011) and Maraun et al. (2011) both condition the parameters of the extreme value distribution on the covariate information. Both Sillman et al. (2011) and Maraun et al. (2011) model only the maximum value over a block of time, e.g., monthly or yearly maximum of daily value. In contrast, we use all observations to both flexibly model the center of the distribution and model the upper tail using EVT given any level of the covariate.

Our approach is novel for several reasons. First, the aim of our analysis is to investigate control strategies. The RFM model is simple enough to permit re-evaluation for a new set of input parameters at negligible computing cost. As a result, we are able to embed the RFM inside the Bayesian model to make inference on the input parameters based on the resulting fit of the RFM to the monitor data, as in Kennedy and O'Hagan (2001), Higdon et al. (2004), and Foley et al. (2012) for Gaussian data. Therefore, unlike previous downscaler methods for extremes (Mannshardt-Shamseldin et al., 2010; Reich et al., 2011; Zhou et al., 2012)), we do not estimate the densities of the model and monitor data separately. To make full use of the RFM, we model the conditional distribution of the monitor data given the RFM. Second, our model for the conditional distribution of the monitor data given the RFM has attractive features. EVT tells us that observations which exceed a high threshold are well-approximated by the generalized Pareto (GPD) distribution. Therefore, in contrast to most previous methods for downscaling numerical model output, here we explicitly leverage EVT by specifying that the conditional distribution of monitor data given the RFM has a GPD tail. However, modeling only the extremes is not sufficient here, since even the center of the conditional distribution of the monitor data given an extremely large value of the RFM may

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

in fact be extreme. This requires that we construct a flexible model for the entire conditional distribution which uses EVT to characterize the upper tail. Our overall model is a combination of quantile regression and EVT, employing quantile regression at levels where there is adequate information to fit a flexible model and EVT in the tail which allows one to extrapolate beyond the range of the data. We assume that EVT is appropriate for the upper tail of the conditional distribution given any value of the RFM (high or low) and allow the GPD distribution to vary with the RFM and spatial location. Our work is similar in spirit with a recent study by Bentzien and Friederichs (2012), who use a related strategy for probabilistic quantitative precipitation forecasting. They estimate a conditional mixture with a GPD tail, although without spatial variation in the parameters or a RFM with unknown input parameters. Third, rather than using the standard diagnostic tools to select a threshold, our model estimates the threshold above which EVT becomes appropriate. Prior work on threshold estimation is limited; Frigessi et al. (2003) use a mixture of a parametric light-tailed distribution and a GPD, and Behrens et al. (2004) use a semi-parametric center and GPD tail and estimate the threshold in a Bayesian way.

2 Description of the air quality model and monitor data

2.1 Base CMAQ model

The Community Multiscale Air Quality (CMAQ) model (Byun and Schere, 2006) version 4.7.1 (Foley et al., 2010) is chosen as the regional photochemical transport model used as the base simulation, from which we later construct our RFM. Ozone was simulated hourly with CMAQ in a domain centered on the southeastern United States for an episode between July 1, 2005 and September 30, 2005, with the full month of June 2005 as a spin-up period. Eight-hour average ozone is then computed using the hourly values. Standard model configuration was used with a 12km by 12km horizontal grid spacing and 14 vertical layers from the surface to 100 hPa, and the Statewide Air Pollution Research Center (SAPRC99) gas-phase chemical mechanism (Carter, 2000). Meteorological fields were developed using the fifth generation mesoscale model (MM5) version 3.6.3 (Grell et al., 1994) and chemical emissions based on the 2001 National Emissions Inventory (http://www.epa.gov/ttn/chief/emch/index.html#2001) were processed using the SMOKE processor, version 2.3.2 (http://www.smoke-model.org), augmented with year 2005 specific emissions data for electric generating units equipped with Continuous Emission Monitoring Systems (CEMS), mobile emissions processed by MOBILE 6 (http://www.epa.gov/otaq/m6.htm), and meteorologically adjusted biogenic emissions from the Biogenic Emission Inventory System (BEIS) 3.13 (Schwede et al., 2005).

2.2 Reduced-form CMAQ model

When model runs are computationally intensive and many runs are required for a thorough sensitivity analysis, an approximation to the model output can be used. One such technique is the decoupled direct method in three dimensions (DDM-3D). This gives a reduced-form CMAQ (RF-CMAQ) model described below. Eulerian photochemical models such as CMAQ typically simulate the emissions, transport, and chemistry of gases and particles in the atmosphere by numerically solving the Atmospheric Diffusion Equation (Seinfeld and Pandis, 1998)

$$\frac{\partial C_i}{\partial t} = -\nabla \cdot (\mathbf{u} C_i) + \nabla \cdot (\mathbf{K} \nabla C_i) + R_i + E_i, \quad (1)$$

where $C_i(t, \mathbf{s})$ is the concentration of species $i = 1, 2, ..., N$ at time t and location $\mathbf{s}$ (with notations for space and time dropped for simplicity), $\mathbf{u}$ is fluid velocity, $\mathbf{K}$ is the diffusivity tensor, R_i is the net rate of chemical generation of species, and E_i is the species emissions

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

rate. DDM-3D computes first-order semi-normalized sensitivity coefficients $S_{ij}^{(1)}(t,\mathbf{s})$ to perturbations in an input parameter p_j as

$$S_{ij}^{(1)}=\frac{\partial C_i}{\partial \varepsilon_j}, \quad (2)$$

where ε_j is a scaling variable with a nominal value of 1.0 applied to the unperturbed parameter field, p_j ~as $\varepsilon_j=\frac{p_j}{\tilde{p}_j}$. Differentiating (1) while using the above definitions leads to an analogous equation governing the first-order sensitivity field

$$\frac{\partial S_{ij}}{\partial t}=-\nabla\cdot(\mathbf{u}S_{ij})+\nabla\cdot(\mathbf{K}\nabla S_{ij})+J_iS_j+\tilde{E}_i, \quad (3)$$

where $\tilde{E}_i$ is the unperturbed emission rate, and $\mathbf{J}_i$ is the ith row vector in the Jacobian matrix $\mathbf{J}$ ($J_{ij} = \partial R_i/\partial C_j$) representing the chemical interaction between species as the previously undefined terms.

Calculations for second-order sensitivity coefficients are also possible and are defined as

$$S_{ijk}^{(2)}=\frac{\partial^2 C_i}{\partial \varepsilon_j \partial \varepsilon_k}. \quad (4)$$

For a given species, Taylor series expansion can be used to approximate the concentration $C_i(t, \mathbf{s})$ as a function of perturbations in a set of input parameters of interest using first- and second-order sensitivity coefficients (Cohan et al., 2005). In this application we are interested only in the concentration of a single species, ozone, and therefore the subscript i is dropped. The RF-CMAQ model for d parameters including second-order and cross-sensitivities is

$$C(t,\mathbf{s}|\alpha)=C_0(t,\mathbf{s})+\sum_{j=1}^{d}S_j^{(1)}(t,\mathbf{s})\,\alpha_j+\frac{1}{2}\sum_{j=1}^{d}S_{jj}^{(2)}(t,\mathbf{s})\,\alpha_j^2+0.5\sum_{l\neq j}S_{lj}^{(2)}(t,\mathbf{s})\,\alpha_j\alpha_l \quad (5)$$

where $C(t, \mathbf{s}|\boldsymbol{\alpha})$ is the ozone concentration due to a specific set of perturbations $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_d)$ at time t and location $\mathbf{s}$, $C_0(t, \mathbf{s})$ is unperturbed concentrations from the base CMAQ simulation, and α_j is the perturbation in input parameter p_j. For example $\alpha_j = -0.10$ corresponds to 10% decrease in NO_x emissions compared to the NO_x emissions used for C_0. The sensitivity coefficients produced by DDM-3D vary in space and time, providing a computationally efficient calculation of ozone under different perturbations in emissions inputs through the RFM. For example, in urban centers NO_x emissions frequently act as a sink of ozone resulting in negative sensitivity to sectors involving NO_x emissions (Figure 1c). In this analysis we consider sensitivity to $d = 6$ inputs: mobile-source NO_x emissions (e.g. traffic), point-source NO_x emissions (e.g. power plants), other NO_x emissions (e.g. construction equipment), anthropogenic VOCs emissions (e.g. benzene emitted from fuel combustion by motor vehicles), biogenic VOCs emissions (e.g. limonene emitted from pine trees), and ozone boundary conditions (3-D hourly pollutant concentrations specified at the grid cells surrounding the model domain).

This second-order RFM is used as an emulator for the full CMAQ model. That is, we use (5) to approximate the spatiotemporal output that would result from an evaluation of the full CMAQ model for a new set of perturbations $\boldsymbol{\alpha}$. This RFM has been shown to have normalized mean error within 10% of reevaluating the full model for α perturbations up to -100% (Cohan et al., 2005). An important caveat is that (Cohan et al., 2005) did not address

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

differences between the full CMAQ and RFM for extreme values, which could potentially have greater impact on projections of extremes.

Using the RF-CMAQ model provides substantial computational improvement. A single month-long simulation using the full CMAQ model takes approximately 10 hours using 72 processors on an IBM system x iDataPlex. The DDM-3D model on the same system runs in approximately 10 days but gives us the ability to estimate pollutant concentrations under many different emissions levels. In the paper, the Bayesian framework allows us to evaluate the reduced form model at thousands of different emissions perturbations to create posterior distributions of the ozone concentration across space. In contrast, running the full model just 100 times in order to roughly approximate the uncertainty in the emissions inputs would require more than a month of computational time. Therefore, RF-CMAQ is the only viable way of exploring the effects of control strategies while accounting for uncertainty in the emissions inputs.

2.3 AQS monitor data

Ozone predictions are also paired in time and space with hourly average ozone observations obtained from EPAs Air Quality System (AQS; http://www.epa.gov/ttn/airs/airsaqs/). This analysis focuses on the maximum eight-hour average ozone concentrations per day (MD8 O_3) for July 1 to September 30 in 2005 at 307 monitoring stations in the southeastern US in Figure 1c. The MD8 is the averaging metric of interest, because it is used for determining compliance with the EPA's National Ambient Air Quality Standards for ozone.

3 Statistical Model for extreme ozone

Let $y(t, \mathbf{s})$ be the AQS measurement for day $t \in \{1, ..., n_T\}$ at spatial location s. Our objective is to estimate the conditional distribution of $y(t, \mathbf{s})$ given RF-CMAQ output for the grid cell containing location $\mathbf{s}$, denoted $C(t, \mathbf{s}|\boldsymbol{\alpha})$. As described in Section 2.2, the perturbation vector $\boldsymbol{\alpha}$ is treated as an unknown parameter in the hierarchical model to allow the AQS data to determine the optimal adjustment to the initial emission levels. For notational convenience, we temporarily suppress dependence on t, $\mathbf{s}$, and $\boldsymbol{\alpha}$ and simply describe the model for y given C. We specify a flexible semiparametric model below a threshold μ where data are abundant, and transition to a parametric GPD model above a threshold where data are sparse. The models above and below the threshold are described in Sections 3.1 and 3.2, respectively.

3.1 Parametric EVT model above the threshold

The parametric GPD distribution has three parameters: lower bound μ, scale $\sigma > 0$, and shape ξ. The domain is (μ, ∞) if $\xi > 0$ and $(\mu, \mu - \sigma/\xi)$ if $\xi < 0$, and the density and quantile (inverse CDF) functions are

$$dGPD\,(y|\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\frac{\xi}{\sigma}\,(y-\mu)\right)_{+}^{-1/\xi-1} \quad \text{and} \quad qGPD\,(\tau|\mu,\sigma,\xi) = \mu+\frac{\sigma}{\xi}\left([1-\tau]^{-\xi}-1\right),$$

respectively, where $x_+ = \max\{0, x\}$. In practice, typically a threshold is selected to be the GPD lower bound, μ. Unlike typical extreme value analysis, rather than choosing a threshold, we instead treat it as a parameter in the model-fitting process.

As shown in Figure 1, the distribution of AQS values is highly dependent on the RF-CMAQ output. Therefore, we assume that the semiparametric/parametric threshold depends on C. Since the threshold will likely vary more on the data scale than the percentile scale, we

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

model the threshold and conditional distribution via its quantile function. The conditional quantile function $q(\tau|C)$ satisfies $P[y < q(\tau|C)] = \tau \in [0, 1]$, and therefore the conditional density function is $dq^{-1}(y|C)/dy$. Utilizing the GPD, the full conditional quantile function is

$$q(\tau|C) = \begin{cases} q_o(\tau|C) & \tau \leq T(C) \\ qGPD\left[\frac{\tau - T(C)}{1 - T(C)} | \mu(C), \sigma(C), \xi(C)\right] & \tau > T(C). \end{cases} \quad (6)$$

In this model, $T(C) \in [0, 1]$ is the quantile level that separates the semiparametric quantile function q_0 and the parametric $qGPD$. For τ above $T(C)$, and thus y above $q_0[T(C)|C] = \mu(C)$, the quantile function takes the form of a GPD with lower bound $\mu(C)$, scale $\sigma(C)$, and shape $\xi(C)$.

3.2 Semiparametric quantile regression below the threshold

We use the model of Reich (2012) for the quantile function below the threshold, q_0. We assume that

$$q_0(\tau|C) = \beta(C) + \sum_{l=1}^{L} B_l(\tau)\theta_l(C). \quad (7)$$

The quantile function is the sum of an overall location term $\beta(C)$ and a linear combination of known basis functions B_l with unknown coefficients $\theta_l(C)$ which determine the shape of the quantile function given C. For the choice of basis functions below, $\beta(C)$ is the median. Although this model is quite flexible, it is centered on the heteroskedastic Gaussian model in that if $\theta_1(C) = ... = \theta_L(C)$, then the quantile function reduces to the Gaussian quantile function with mean $\beta(C)$ and standard deviation $\theta_1(C)$.

To lead to a valid statistical model, q_0 must be increasing in τ for all C. To do this, we define $B_1(\tau) = \Phi^{-1}(\tau)$ if $L = 1$, where Φ^{-1} is the standard normal quantile function. In this case, the model below the threshold is Gaussian with mean $\beta(C)$ and standard deviation $\theta(C)$. For non-Gaussian data we generalize by allowing $L \geq 2$, and specifying basis functions

$$B_l(\tau) = \begin{cases} \Phi^{-1}(\kappa_l) - \Phi^{-1}(\kappa_{l+1}) & \tau < \kappa_l \\ \Phi^{-1}(\tau) - \Phi^{-1}(\kappa_{l+1}) & \kappa_l \leq \tau < \kappa_{l+1} \\ 0 & \kappa_{l+1} \leq \tau, \end{cases} \quad (8)$$

for l with $\kappa_l < 0.5$ and

$$B_l(\tau) = \begin{cases} 0 & \tau < \kappa_l \\ \Phi^{-1}(\tau) - \Phi^{-1}(\kappa_l) & \kappa_l \leq \tau < \kappa_{l+1} \\ \Phi^{-1}(\kappa_{l+1}) - \Phi^{-1}(\kappa_l) & \kappa_{l+1} \leq \tau, \end{cases} \quad (9)$$

for l such that $\kappa_l \geq 0.5$, where $0 = \kappa_1 < ... < \kappa_{L+1} = 1$ be a grid of equally-spaced knots covering [0, 1]. Then the quantile function is increasing if and only if $\theta_l(C) > 0$ for all l and C. We only consider even L in which case $B_l(0.5) = 0$ for all l, and the median is $q_0(0.5|C) = \beta(C)$. Also, if $\theta_1(C) = ... = \theta_L(C) = \theta(C)$, then for $\tau < \mu(C)$, $q(\tau|C) = \beta(C) + \theta(C)\Phi^{-1}(\tau)$, and the density below the threshold is Gaussian with mean $\beta(C)$ and standard deviation $\theta(C)$. It is also possible to use other basis function B_l. For example, taking B_l to be the gamma or log-normal distribution function would ensure a lower bound $q_0(0|C) = 0$.

These basis functions also permit a closed-form expression for the conditional density

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$\begin{aligned} p(y|C) &= I[y<\mu(C)]\sum_{l=1}^{L} I[q_0(\kappa_l|C)\le y<q_0(\kappa_{l+1}|C)]\,\mathbf{N}\left[y|a_l(C),\theta_l(C)^2\right] \\ &+ I[y\ge\mu(C)][1-T(C)]\,dGPD(y|\mu(C),\sigma(C),\xi(C)) \end{aligned} \quad (10)$$

where $\mathbf{N}(\cdot|a, b^2)$ denotes the density of a normal with mean a and standard deviation b, and $a_l(C) = q(\kappa_{l+1}|C) - \theta_l(C)\Phi^{-1}(\kappa_{l+1})$ if $\kappa_l < 0.5$ and $a_l(C) = q(\kappa_l|C) - \theta_l(C)\Phi^{-1}(\kappa_l)$ if $\kappa_l \ge 0.5$. Therefore, the density is multiply-split normal with breakpoints (and points of discontinuity) $q(\kappa_l|C)$ and $\mu(C)$. However, our primary interest, the quantile function, is a continuous function.

3.3 Modeling dependence on RF-CMAQ

The conditional density varies with C via $\beta(C)$, $\theta_l(C)$, $T(C)$, $\sigma(C)$, and $\xi(C)$. These parameters could be allowed to have a complex dependence on C to capture subtle features of the conditional distribution. For example, one could use a Gaussian process defined over C. For simplicity, we assume that after a suitable transformation each parameter is an order–M polynomial expansion of C. That is, $\beta(C) = \mathbf{X}\mathbf{a}^{(\beta)}$, $\log[\theta_l(C)] = \mathbf{X}\mathbf{a}^{(\theta_l)}$, $\log[\sigma(C)] = \mathbf{X}\mathbf{a}^{(\sigma)}$, $\xi(C) = \mathbf{X}\mathbf{a}^{(\xi)}$, where $\mathbf{X} = (1, \bar{C}, ..., \bar{C}^M)$ and $\bar{C} = (C - 50)/15$ is the standardized CMAQ output (where 50 and 15 are the approximate mean and standard deviation, respectively). Note that this polynomial model contains the linear regression model as a special case, and that in this case the intercept can account for systematic bias between the AQS data and the RF-CMAQ predictions. Higher-order polynomials or spline basis expansion would allow for more complex relationships between RF-CMAQ and the AQS data.

The semi-parametric/parametric threshold $T(C)$ must be modeled so that $T(C)$ is confined to [0, 1] for all C. In our analysis, we intend for $T(C)$ to be an extreme quantile to theoretically justify the GPD fit, so we restrict $T(C) \in [l, u]$ where l and u are unknown parameters with $l \sim \text{Uniform}(0.8,1.0)$ and $u|l \sim \text{Uniform}(l, 1.0)$. The variability of $T(C)$ within (l, u) is modeled using the logistic link

$$T(C) = l\frac{\exp[d(C)]}{1+\exp[d(C)]} + u\frac{1}{1+\exp[d(C)]}. \quad (11)$$

As with the other parameters, d, and thus T, varies with C as $d(C) = \mathbf{X}\mathbf{a}^{(d)}$.

3.4 Spatiotemporal modeling

There are two potential sources of spatial and temporal dependence in the data: spatial variation in the conditional distribution of AQS given RF-CMAQ, and residual spatiotemporal association in the observations given the conditional distribution. To account for spatial variation in the conditional distribution, we allow the parameters that define the semiparametric model below the threshold, $\mathbf{a}^{(\beta)}$ and $\mathbf{a}^{(\theta_l)}$, as well as the GPD scale $\mathbf{a}^{(\sigma)}$ to vary by spatial location. These processes are then smoothed by Gaussian process priors. For example, denote $\mathbf{a}^{(\beta)}$ at location $\mathbf{s}$ as $\mathbf{a}^{(\beta)}(\mathbf{s}) = \left[a_0^{(\beta)}(\mathbf{s}), \ldots a_M^{(\beta)}(\mathbf{s})\right]^T$, and thus the spatially-varying coefficients $\beta(x, \mathbf{s}) = \mathbf{X}\mathbf{a}^{(\beta)}(\mathbf{s})$. Then $a_j^{(\beta)}(\mathbf{s})$ has a Gaussian process prior with mean $\bar{a}_j^{(\beta)}$, variance $\tau_j^{(\beta)}$, and exponential spatial correlation $\text{Cor}\left[a_j^{(\beta)}(\mathbf{s}), a_j^{(\beta)}(\mathbf{s}')\right] = \exp\left(-\|\mathbf{s}-\mathbf{s}'\|/\rho\right)$. The hyperparameters have priors $\bar{a}_j^{(k)} \sim \mathbf{N}\left(0, c_1^2\right)$ and $\tau_j^{(k)} \sim \text{Gamma}(c_2, c_3)$ for $k \in \{\beta, \theta_1, ..., \theta_L, \sigma\}$. To borrow strength

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

across processes we assume a common spatial range ρ, which is reasonable since all of these spatially-varying parameters represent the change in response distribution across locations.

We find that while the threshold parameters $\mathbf{a}^{(d)}$ and GPD shape parameters $\mathbf{a}^{(\xi)}$ are well-identified when allowed to vary by the value of RF-CMAQ, they are poorly-identified when allowed to vary by RF-CMAQ and spatial location. This is not surprising since they are by definition related only to the tail of the distribution, and thus there are only a few relevant observations at each location. These parameters are thus held constant for all locations with priors $a_j^{(k)} \sim \mathbf{N}\left(0, c_1^2\right)$ for $k \in \{d, \xi\}$. We note that although the threshold is fixed at a constant quantile level across space, the actual threshold on the ozone scale is $\mu[C(t, \mathbf{s}|\boldsymbol{\alpha})] = q_o\{T[C(t, \mathbf{s}|\boldsymbol{\alpha})]|C(t, \mathbf{s}|\boldsymbol{\alpha}), \mathbf{s}\}$, which does vary spatially. Also, since the interpretation of the GPD scale is dependent on the threshold, it seems a reasonable approach to not allow the parameters dictating the threshold to vary spatially, and assume that the spatially varying scale parameter can account for spatial variation. Combining these specifications gives the final quantile model fit to the ozone data in Section 5

$$q\left(\tau|C,\mathbf{s}\right)=\begin{cases} q_o\left[\tau|C\left(t,\mathbf{s}|\alpha\right),\mathbf{s}\right] & \tau\leq T\left[C\left(t,\mathbf{s}|\alpha\right)\right] \\ qGPD\left\{\frac{\tau-T[C(t,\mathbf{s}|\alpha)]}{1-T[C(t,\mathbf{s}|\alpha)]}|,\mu\left[C\left(t,\mathbf{s}|\alpha\right)\right],\sigma\left[C\left(t,\mathbf{s}|\alpha\right),\mathbf{s}\right],\xi\left[C\left(t,\mathbf{s}|\alpha\right)\right]\right\} & \tau>T\left[C\left(t,\mathbf{s}|\alpha\right)\right]. \end{cases} \quad (12)$$

where $q_0\left[\tau|C\left(t,\mathbf{s}|\alpha\right),\mathbf{s}\right]=\beta\left[C\left(t,\mathbf{s}|\alpha\right),\mathbf{s}\right]+\sum_{l=1}^{L}B_l\left(\tau\right)\theta_l\left[C\left(t,\mathbf{s}|\alpha\right),\mathbf{s}\right]$.

Even after accounting for spatial variation in the conditional distributions, there is spatial and temporal dependence in the residuals due to day-to-day variation in ozone. We account for this dependence with a Gaussian copula (Nelsen, 1999). The copula is defined by a latent Gaussian process $z(t, \mathbf{s})$ with mean zero, variance one, and a spatiotemporal correlation function. Then $U(t, \mathbf{s}) = \Phi[z(t, \mathbf{s})] \sim \text{Unif}(0,1)$, and the latent process is related to the response as $y(t, \mathbf{s}) = q[U(t, \mathbf{s})|C(t, \mathbf{s}|\boldsymbol{\alpha}), \mathbf{s}]$. While the Gaussian copula induces some spatiotemporal dependence, it is well known that the Gaussian copula gives asymptotic independence. That is, assuming the same marginal distribution for $y(t, s)$ and $y(t', \mathbf{s}')$, then $\lim_{u\to U} P\,[y(t, s) > u|y(t', \mathbf{s}') > u] = 0$, where U is the upper bound of $y(t, \mathbf{s})$, which implies that the Gaussian copula is equivalent to ignoring dependence between very extreme events. Therefore the Gaussian copula may not be ideal for extreme data in all settings. However, our exploratory analysis in Section 5 suggests that there is little extremal dependence in the residuals and therefore that this model fits the AQS data well after accounting for RF-CMAQ output. In other cases, copulas with asymptotic dependence may be desirable. Examples of copulas with asymptotic dependence include the t-copula (Nelsen, 1999) or a non-parametric copula (Fuentes et al., 2012). Another possibility is to specifically target extremal dependence (for example, Davison and Smith, 1990; Chavez-Demoulin and Davison, 2012; Eastoe and Tawn, 2012).

4 Computational approach to evaluating control strategies

The computing used for the ozone data analysis has two main steps: we first analyze the AQS data to estimate the parameters in the conditional distribution of AQS given RF-CMAQ, and then generate replications of the summer ozone process under different control strategies. These two steps are described separately in the subsections below.

4.1 Parameter estimation

Because of the size of the dataset, we fit this model in two stages. We first estimate $\boldsymbol{\alpha}$ and the $\mathbf{a}^{(k)}$ parameters for $k \in \{\beta, \theta_l, \sigma, \xi, d\}$ in one model fit assuming the observations are independent conditioned on these parameters. In the second stage, we estimate the copula

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

parameters conditioned on the first-stage parameter estimates. Assuming independence of the observations conditioned on $\Theta(\mathbf{s}) = \{\boldsymbol{\alpha}, \mathbf{a}^{(\beta)}(\mathbf{s}), \mathbf{a}^{(\theta)}(\mathbf{s}), \mathbf{a}^{(\sigma)}(\mathbf{s}), \mathbf{a}^{(\xi)}, \mathbf{a}^{(d)}, l, u\}$, the likelihood is simply the product of terms of the form of (10). Denote $p[y(\mathbf{s}, t)|\Theta(\mathbf{s})]$ as the density of $y(\mathbf{s}, t)$. None of the parameters in the likelihood have conjugate full conditionals, so we use Metropolis-Hastings sampling. Metropolis-Hastings sampling proceeds by specifying initial values for all parameters, and then updating the parameters one-at-a-time, conditioned on the current value of all other parameters. For example, to update α_j, we draw candidate $\alpha_j^{can} \sim \mathbf{N}\left(\alpha_j^{cur}, c^2\right)$, where α_j^{cur} is the current value and the standard deviation c is a tuning parameter. With probability R, α_j is set to α_j^{can}, and α_j is set to α_j^{cur} otherwise, where

$$R = \min\left\{\frac{\prod_{i,t} p\left[y\left(\mathbf{s}_i, t\right)|\Theta(\mathbf{s}_i)^{can}\right] p\left(\alpha_j^{can}\right)}{\prod_{i,t} p\left[y\left(\mathbf{s}_i, t\right)|\Theta(\mathbf{s}_i)^{cur}\right] p\left(\alpha_j^{cur}\right)}, 1\right\},$$

$\Theta(\mathbf{s})^{can}$ includes α_j^{can}, $\Theta(\mathbf{s})^{cur}$ includes α_j^{cur}, and $p(\alpha)$ is the Gaussian prior. Evaluating $p[y(\mathbf{s}, t)|\Theta(\mathbf{s})]$ requires first computing RF-CMAQ given perturbation parameters $\boldsymbol{\alpha}$, $C(\mathbf{s}, t|\boldsymbol{\alpha})$, which is trivial for RF-CMAQ model following (5). All parameters are updated similarly with Gaussian candidate distributions tuned to give acceptance rates near 0.4. We generate 25,000 samples from the posterior and discard the first 10,000 as burn-in. Convergence is monitored using trace plots of several representative parameters.

To estimate the copula parameters, we compute the estimated Gaussian-transformed residuals $z(t, \mathbf{s}) = \Phi^{-1}\{\hat{q}^{-1}[y(t, s)|C(t, \mathbf{s}), \mathbf{s}]\}$, where $\hat{q}$ is the quantile function evaluated at the posterior mean of all model parameters. There is of course spatial dependence in the residuals, and sampling with spatial dependence would be crucial for statistics defined over the spatial domain, for example, total precipitation in a watershed. However, our interest is in projecting the change in ozone distribution at each site, and not for a collection of sites simultaneously. Therefore, assessment of spatial dependence in the predictions is not a concern and we assume the residuals are independent over space for computational convenience. We then fit a first-order autoregressive model for the temporal dependence, $\mathrm{Cor}[z(t, \mathbf{s}), z(t', \mathbf{s}')] = \exp(-|t - t'|/\varphi)I(\mathbf{s} = \mathbf{s}')$, with autocorrelation parameters held constant over space and time. We fix φ to match the sample correlation of subsequent residuals at the same location.

4.2 Generating samples from the posterior predictive distribution

To evaluate the effects of control strategies on the likelihood and magnitude of extreme ozone events, we generate several replications of summer ozone at each spatial location. The control strategies correspond to reductions of emissions in various sectors, and are parameterized in terms of the RF-CMAQ model inputs $\boldsymbol{\alpha}$. In the RF-CMAQ model, α_j represents a $100\alpha_j$% change in the initial estimate of the emissions in sector j. Therefore, the posterior of $\boldsymbol{\alpha}$ represents the calibrated emissions for the base case based on fitting to the AQS data. To simulate RFM output that corresponds to an additional change of $100\eta_j$% after the calibration via α_j, we use $\alpha_j^* = (1+\alpha_j) * (1+\eta_j) - 1$ as inputs to the RFM. By simulating data for different values of $\boldsymbol{\rho} = (\eta_1, ..., \eta_p)$, we simulate ozone data under different control strategies. These control strategies assume a uniform reduction across the entire region.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

For each control strategy we generate R = 10, 000 replicates of the summer ozone at each CMAQ grid cell. Due to the computational burden, the grid cells are thinned by removing every other column and row. For each replicate we randomly sample one of the posterior draws for the parameters in the conditional distribution $\{\mathbf{a}^{(\beta)}, \mathbf{a}^{(\theta)}, \mathbf{a}^{(d)}, \mathbf{a}^{(\sigma)}, \mathbf{a}^{(\xi)}, \boldsymbol{\alpha}\}$. At each iteration all spatially-varying parameters are interpolated from the AQS stations used for model-fitting to the grid cell locations by sampling from their posterior predictive distribution. We then compute the RF-CMAQ model corresponding to input $\alpha^* = \left(\alpha_1^*, \ldots, \alpha_p^*\right)$, and the conditional distribution of $y(t, \mathbf{s})$ given $C(t, \mathbf{s}|\boldsymbol{\alpha}^*)$ and $\{\mathbf{a}^{(\beta)}, \mathbf{a}^{(\theta)}, \mathbf{a}^{(d)}, \mathbf{a}^{(\sigma)}, \mathbf{a}^{(\xi)}\}$. We generate the responses for each simulated year for location $\mathbf{s}$ by generating $\mathbf{z}(\mathbf{s}) = [z(1, \mathbf{s}), ..., z(n_T, \mathbf{s})]^T$ from a multivariate normal model with mean zero and covariance $\mathrm{Cov}[z(t, \mathbf{s}), z(t', \mathbf{s})] = \exp(-\|t - t'\|/\varphi)$ and transforming to $y(t, \mathbf{s}) = q(\Phi\{z(t, \mathbf{s})\}|C(t, \mathbf{s}|\boldsymbol{\alpha}^*, \mathbf{s}))$ so that $y(t, \mathbf{s})$ has the quantile function in (12).

This method of simulation accounts for uncertainty in the AQS value given RFM output, and uncertainty in the parameters in the conditional distribution of the AQS value given the RFM output. It also partially accounts for randomness in the RF-CMAQ output by marginalizing over the posterior of $\boldsymbol{\alpha}$. However, there are many additional inputs to RF-CMAQ that are taken as fixed, and so not all the randomness in RF-CMAQ from year to year is accounted for by this approach. Ideally we would have a larger sample of RF-CMAQ output to better represent the sampling distribution of RF-CMAQ from year to year. Therefore, the results should be interpreted cautiously as pertaining to the changes in the ozone distribution for this particular simulated year, which may suppress some variability for an arbitrary future year.

5 Constructing the downscaler between RFM output and AQS data

To display the results of fitting the conditional distribution of AQS given RF-CMAQ, we first compare several models based on test set prediction in Section 5.1. We then illustrate the fitted distribution of our final model in Section 5.2.

5.1 Model comparisons

We compare several models by varying the number of basis functions in the semiparametric quantile process, L, the order of the polynomial for RF-CMAQ in the conditional distribution, M, and with and without (i.e., $T(C) = 1$) the GPD tail. For comparison, we also include the non-statistical forecast by simply taking the base CMAQ output $C_0(\mathbf{s}, t)$ as the prediction. For all fits, we use uninformative priors $c_1 = 100$, $c_2 = c_3 = 0.1$, and $\log(\rho) \sim \mathbf{N}(0, 10)$. To compare these models, we randomly (across space and time) split the data equally into training (n=13,645) and testing data sets (n=13,645). We fit each model to the training set, calculate the posterior mean of all model parameters, and then compute the predictive distribution for each test set observation.

Models are compared in terms of their fit to the upper tail of the distribution using Brier scores for exceedances and quantile scores for extreme quantiles (see, e.g., Gneiting and Raftery, 2007). The quantile score for quantile level τ is $2\{I[y < q(\hat{\tau})] - \tau\}(q - y)$, where y is the test set AQS value and $q(\hat{\tau})$ is its estimated τ^{th} quantile. The Brier score for evaluating accuracy of predicting exceedance of threshold c is $[e(c) - P(c)]^2$, where $e(c) = I(y > c)$ is the indicator that the test set AQS value exceeds c and $P(c)$ is the predicted probability of an exceedance. For the non-statistical predictions, we take $q(\hat{\tau}) = C_0$ and $P(c) = I(C_0 > c)$. We compare models using several extreme values of τ and c, and average these values over all observations in the test set. For both quantile and Brier scores, small values are preferred.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1 gives the results. The non-stochastic bias-adjusted base CMAQ fit ("SLR"), i.e., $a(\mathbf{s}) + b(\mathbf{s})C_0(\mathbf{s}, t)$ where $a(\mathbf{s})$ and $b(\mathbf{s})$ are fit using separate linear regressions at each location, has the highest scores, verifying the need for statistical calibration. We also fit simple linear quantile regression ("QR", using the quantreg package in R) and logistic regression ("LR") with base CMAQ as a linear predictor separate by site (for very extreme quantiles and threshold these methods had some convergence problems, and we simply carried forward the estimates from the next lowest quantile or threshold). Although these models do not provide a means to generate ozone under different scenarios, they do provide improved fit compared to linear regression.

The quantile scores clearly show the value of the GPD tail model. Although the score values are hard to interpret, the scores for the 0.99 and 0.995 quantiles are universally lower than those for the model without the GPD tail; the scores for the statistical models for the 0.99 quantile are 0.521, 0.518, 0.520, and 0.515 for the models without GPD tail, compared to 0.507, 0.498, 0.505, and 0.502 for the model with GPD tail. Exceedence prediction for high thresholds is not only affected by the tail of the distribution, but also the center. For example, Figure 1 shows that most exceedences of 80ppb occur when CMAQ is large, and that 80ppb is in the center of the conditional distribution for large CMAQ. Therefore, the GPD tail is not the most influential factor for Brier score, but rather the most important factor is accurate modeling of the relationship between AQS and RF-CMAQ via the degree of the polynomial in the model parameters, M. The Brier score for 80ppb is 2.105, 2.108, 2.122, and 2.150 for the linear models with $M = 1$, compared to 1.975, 1.951, 1.978, and 1.973 for the quadratic models with $M = 2$. More complex models for the RF-CMAQ predictors such as higher-order polynomials (that is, $M > 2$) or spline fits are also possible. We fit the model with GPD tails and $L = 1$ with $M = 3$ and found a slight improvement for moderate quantile levels but poor performance for the extremes, likely due to overfitting. Therefore, we conclude a second-order polynomial is sufficient for these data. Non-Gaussian modeling ($L = 4$) of the distribution below the threshold does not appear to improve model fit compared to the Gaussian model ($L = 1$) for these data. Therefore, although other models are fairly similar, the best model in terms of both the quantile scores and Brier scores has $L = 1$, $M = 2$, and GPD tail. This model is Gaussian below the threshold, and all the model parameters are quadratic in RFM. The results below are from the data analysis using this model on the complete dataset.

5.2 Summary of the final model

For the final model with $L = 1$ and $M = 2$, RF-CMAQ input parameters α_j are all negative with high probability, suggesting that all emissions inputs used in the base simulation C_0 are too high. Their 95% posterior intervals are (-0.17, -0.05) for mobile source NO_x, (-0.22, -0.11) for point source NO_x, (-0.27, -0.10) for other NO_x, (-0.74, -0.59) for anthropogenic VOC emissions, (-0.24, -0.17) for biogenic VOC emissions, and (-0.10, -0.07) for ozone boundary conditions.

Figure 2 summarizes the GPD fit to the tail of the conditional distribution. The threshold $T[C(t, \mathbf{s}|\boldsymbol{\alpha})]$ in Figure 2a, which depends on l, u, and $d(C)$, varies between the 0.80 and 0.95. The threshold is lower, and thus the GPD fits a larger portion of the tail, for moderate to high RF-CMAQ values 50-80ppb. The GPD shape $\xi[C(t, \mathbf{s}|\boldsymbol{\alpha})]$ in Figure 2b is near zero for low RF-CMAQ values, negative for moderate RF-CMAQ, and positive for large RF-CMAQ. This generally agrees with the sample density estimates in Figure 1b, which have heavier tails for low and high values of RF-CMAQ. Unlike the threshold and shape, the GPD scale $\sigma[C(t, \mathbf{s}|\boldsymbol{\alpha}), \mathbf{s}]$ varies spatially (Figures 2c and 2d). The GPD scale is larger in the south and the Chesapeake Bay area. Also, the scale is generally larger for RF-CMAQ equal 50ppb than 80ppb. Note that this does not imply a lighter tail for extreme RF-CMAQ, since

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the GPD shape parameter (Figure 2b) is higher for large RF-CMAQ values compared to moderate RF-CMAQ values.

To determine the form of residual dependence, we compute the estimated Gaussian-transformed residuals $z(t, \mathbf{s})$. Figure 3a plots the residuals for each consecutive pair of observations at the same station, $[z(t–1, \mathbf{s}), z(t, \mathbf{s})]$. The sample correlation is 0.26 (p-value for the test of correlation is <0.001). We note that much of the autocorrelation in ozone is captured by RF-CMAQ model, and thus the correlation in the residuals is lower than the correlation in the raw ozone values. To test for extremal dependence, Figure 3b plots the residuals transformed to have unit Fréchet margins to emphasize dependence in the tails. The unit Fréchet distribution function is $P(Z < c) = \exp(–1/c)$, therefore if $z(t, \mathbf{s})$ is standard normal then $z_F(t, \mathbf{s}) = –1/\log\{\Phi[z(t, \mathbf{s})]\}$ is unit Fréchet. The pairs $[z_F(t – 1, \mathbf{s}), z_F(t, \mathbf{s})]$ show no asymptotic dependence since for all pairs with one extremely large value the other member of the pair is near zero. Therefore, we use a Gaussian copula for predictive purposes.

6 The distribution of extremes under various control strategies

To determine the local effects on extreme ozone events of several control strategies, we sample R replicates of summer ozone at each grid cell from the predictive distribution, as described in Section 4.2. We compare four control strategies:

S0: the base case with no change in emissions, $\boldsymbol{\rho} = (0, 0, 0, 0, 0, 0)$

S1: a 50% reduction in mobile-source NO_x, $\boldsymbol{\rho} = (–0.5, 0, 0, 0, 0, 0)$

S2: a 50% reduction in point-source NO_x, $\boldsymbol{\rho} = (0, –0.5, 0, 0, 0, 0)$

S3: a 15% reduction in mobile, point, and other-source NO_x, $\boldsymbol{\rho} = (–0.15, –0.15, –0.15, 0, 0, 0)$.

These emission reductions were selected to give roughly a spatial-average of 3ppb decrease in the base CMAQ $C_0(\mathbf{s}, t)$. These reductions are in line with reductions often considered by regulators and air quality managers. For example, the 2008 EPA Regulatory Impact Analysis (http://www.epa.gov/ttn/ecas/regdata/RIAs/452_R_08_003.pdf), which considers reductions of 30% to 90% for both VOC and NOx. We display the predictive distribution by computing various summary statistics for each replication, e.g., $y_4^{(r)}(\mathbf{s})$, the fourth largest value of $\{y(1, \mathbf{s}), ..., y(n_T, \mathbf{s})\}$ for replication r, and plotting its mean, $\sum_{r=1}^{R} y_4^{(r)}(\mathbf{s})/R$, and proportion above 75ppb, $\sum_{r=1}^{R} I\left[y_4^{(r)}(\mathbf{s}) > 75\right]/R$.

To illustrate the effects modeling the tail as GPD rather than Gaussian, Figure 4 plots the average yearly maximum and fourth-highest day for the base case S0 with the final model with $L = 1$, $M = 2$, and GPD tails and the fit without the GPD tail, i.e., a Gaussian model with mean $\beta(\mathbf{x}, \mathbf{s})$ and standard deviation $\theta_1(\mathbf{x}, \mathbf{s})$. The two models differ by 3-5ppb in many locations, which is a meaningful difference for regulatory purposes. The probability that the yearly maximum and fourth highest day are larger using the GPD model is 0.8-0.9 in eastern North Carolina and the Chesapeake Bay area (Figures 4c and 4f). Although these probabilities are not definitive, we note that they are computed using separate samples from the residual distribution ($\mathbf{z}(\mathbf{s})$ in Section 4.2), and therefore these probabilities represent an almost complete separation of the predictive distributions under these two models.

Figure 5 compares the projection of the 4th highest day of the year under the four scenarios. Compared to the base case, the 50% reduction in mobile source NO_x has the largest effects in the area surrounding Atlanta (Figure 5a). The reduction is as large as 6ppb to the east of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Atlanta. In the center of the city, however, this control strategy gives virtually no reduction. It is well known that high NO released in high-traffic areas destroys ozone near the source, and therefore reducing the mobile-source emissions does not reduce ozone near the source, but rather downwind where NO concentrations are lower. This can be seen in the map of the sensitivities in Figure 1c, which is negative in Atlanta's center, but positive in its southeastern suburbs. The reduction in ozone for the point-source NO_x scenario is generally smaller and is more uniform across space (Figure 5b). The reduction is 2-5ppb for most of the region with exceptions of smaller reductions in Atlanta and Northern Virginia. Comparing the reductions corresponding to the mobile-source and point-source control strategies (Figure 5d), we find a larger reduction for the mobile-source strategy in most of the spatial domain, with exceptions in Kentucky, West Virginia, and central Atlanta. The third control strategy of reducing all the NO_x emissions by 15% shows a similar spatial pattern to the mobile-source strategy, but with generally smaller reductions.

In addition to comparing changes in the mean of the fourth highest day of the year, the predictive distributions can also be used to study the probability that the fourth highest day exceeds the current standard of 75ppb. Figure 6 shows that many areas have a substantial reduction in the exceedance probability. For example, the Birmingham and Raleigh areas go from near 1.0 in the base case to 0.6-0.8 under the mobile-source control strategy. However, areas with the highest ozone level, Atlanta and Chesapeake Bay, have exceedance probability near one in all cases. Comparing the mobile-source and point-source strategies (Figure 6f), the exceedance probability is lower under the mobile-source control strategy than the point-source strategy for 74% of the grid cells. Overall, the mobile-sources reduction strategy appears to be the most effective.

7 Discussion

In this paper, we propose a new framework for downscaling extremes. We propose to model the conditional distribution of the monitor data given the RFM as a combination of quantile regression and extreme value modeling, using generalized Pareto tails. Using a fully-Bayesian analysis, we propagate many sources of uncertainty through to the final estimate of the effect of each control strategy. Using this approach, we evaluate three control strategies related to reduction in NO_x. We find that reducing mobile-sources NO_x has the largest impact of the strategies considered, especially in suburban Atlanta. However, the probability of non-compliance with the EPA regulation remains near one for the Atlanta area for all control strategies.

Although our modeling framework is quite flexible, it has several limitations. First we have not considered residual spatial correlation in the model fitting stage. Reich (2012) does include residual spatiotemporal dependence for spatial quantile regression. Although estimating residual spatial dependence in not the primary focus in this work, failing to account for this may cause underestimation of the uncertainty of model parameters. Also, although the quantile function is modeled as increasing for each value of the RFM, the quantile function is not forced to be increasing in RFM for each quantile level, as one would naturally expect. Including this prior belief may be possible by adding further restrictions to the polynomial coefficients in the conditional distribution, and would certainly improve the fit for small and moderate data sets.

Also, we note that our analysis only considers global emission control strategies that assume a uniform reduction in emissions across the entire spatial domain. An extension would be to calculate sensitivities to local changes, to study the effects of emission reduction in one grid cell on neighboring cells. This would add another spatial aspect to the model, and provide a more comprehensive sensitivity analysis.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Acknowledgments

The authors with to thank the editor and anonymous referees for their thoughtful comments which greatly improved the manuscript. This research was partially supported by NIH grant R01-ES-014843-02 (Reich), EPA-STAR award R835228 (Reich and Cooley), and NSF grant 1107046 (Reich). The United States Environmental Protection Agency through its Office of Research and Development funded and managed the research described here (Foley and Napelenok). It has been subjected to Agencys administrative review and approved for publication.

References

Behrens C, Lopes H, Gamerman D. Bayesian analysis of extreme events with threshold estimation. Statistical Modelling. 2004; 4:227.

Bentzien S, Friederichs P. Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution nwp model cosmo-de. Weather and Forecasting. 2012; 27:998–1002.

Berrocal VJ, Gelfand AE, Holland DM. A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. 2010; 15:176–197.

Byun D, Schere KL. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. Applied Mechanics Reviews. 2006; 59:51–77.

Carter, WPL. Implementation of the SAPRC-99 chemical mechanism into the models-3 framework.. Report to the United States Environmental Protection Agency. 2000. online available at: http://www.cert.ucr.edu/carter/absts.htm#s99mod3

Chavez-Demoulin V, Davison A. Modeling time series extremes. REVSTAT. 2012; 10:109–133.

Cohan DS, Hakami A, Hu YT, Russell AG. Nonlinear response of ozone to emissions: Source apportionment and sensitivity analysis. Environmental Science and Technology. 2005; 39:6739–6748. [PubMed: 16190234]

Coles, S. An introduction to statistical modeling of extreme values. Springer-Verlag London Ltd; London: 2001. Springer Series in Statistics

Cooley D, Sain SR. Spatial hierarchical modeling of precipitation extremes from a regional climate model. J. Agric. Biol. Environ. Stat. 2010; 15:381–402.

Davison A, Smith R. Models for exceedances over high thresholds (with discussion). Journal of the Royal Statistical Society: Series B. 1990; 52:393–442.

Digar A, Cohan D, Cox D, Byeong-Uk K, Boylan J. Likelihood of achieving air quality targets under model uncertainties. Environ. Sci. Technol. 2011; 45:189–196.

Eastoe E, Tawn J. Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. Biometrika. 2012; 99:43–55.

Foley KM, Reich BJ, Napelenok SL. Bayesian analysis of a reduced-form air quality model. Environmental Science & Technology. 2012 Accepted.

Foley KM, Roselle SJ, Appel KW, Bhave PV, Pleim JE, Otte TL, Mathur R, Sarwar G, Young JO, Gilliam RC, Nolte CG, Kelly JT, Gilliland AB, Bash JO. Incremental testing of the community multiscale air quality (CMAQ) modeling system version 4.7. Geoscientific Model Development. 2010; 3:204–226.

Frigessi A, Haug O, Rue H. A dynamic mixture model for unsupervised tail estimation without threshold selection. Extremes. 2003; 5:219–236.

Fuentes M, Henry J, Reich B. Nonparametric spatial models for extremes: Applications to extreme temperature data. Extremes. 2012 Accepted.

Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 2007; 102:359–378.

Grell, GA.; Dudhia, AJ.; Stauffer, DR. A description of the fifth-generation pennstate/ncar mesoscale model (mm5).. NCAR Technical Note NCAR/TN-398+STR. 1994. online available at: http://www.mmm.ucar.edu/mm5/doc1.html

Higdon D, Kennedy M, Cavendish J, Cafeo J, Ryne R. Combining field data and computer simulation for calibration and prediction. SIAM Journal on Scientific Computing. 2004; 26:448–466.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Kennedy M, O'Hagan A. Bayesian calibration of computer models (with discussion). Journal of the Royal Statistical Society: Series B. 2001; 63:425–464.

Kharin V, Zwiers F, Zhang X, Hegerl G. Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. J. Climate. 2007; 20:1419–1444.

Mannshardt-Shamseldin EC, Smith RL, Sain SR, Mearns LO, Cooley D. Downscaling extremes: a comparison of extreme value distributions in point-source and gridded precipitation data. Ann. Appl. Stat. 2010; 4:484–502.

Maraun D, Osborn T, Rust H. The influence of synoptic airflow on UK daily precipitation extremes. Part I: Observed spatio-temporal relationships. Climate Dynamics. 2011; 36:261–275.

Napelenok S, Foley K, Kang D, Mathur R, Pierce T, Rao S. Dynamic evaluation of regional air quality model's response to emissions reduction in the presence of uncertain emission inventories. Atmospheric Environment. 2011; 45:4091–4098.

Nelsen, R. An introduction to copulas. Springer-Verlag; New York: 1999.

Reich B. Spatiotemporal quantile regression for detecting distributional changes in environmental processes. Journal of the Royal Statistical Society: Series C. 2012; 64:535–553.

Reich B, Fuentes M, Dunson D. Bayesian spatial quantile regression. Journal of the American Statistical Association. 2011; 106:6–20.

Schliep EM, Cooley D, Sain SR, Hoeting JA. A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. Extremes. 2010; 13:219–239.

Schwede, D.; Pouliot, GA.; Pierce, T. Changes to the biogenic emissions inventory system version 3 (BEIS3).. In Proceedings of the 4th CMAS Models-3 Users Conference; Chapel Hill, NC, 2628. September 2005; 2005.

Seinfeld, J.; Pandis, S. Atmospheric Chemistry and Physics. Wiley & Sons, Inc; New York: 1998.

Sillman J, Croci-Maspoli M, Kallache M, Katz RW. Extreme cold winter temperatures in Europe under the influence of North Atlantic atmospheric blocking. J. Climate. 2011; 24:5899–5913.

US Environmental Protection Agency (US EPA). Air Quality Criteria for Ozone and Related Photochemical Oxidants (Final). U.S. EPA; Washington, DC: 2006. Available on the Internet at http://cfpub.epa.gov/ncea/CFM/recordisplay.cfm?deid=149923

Wehner M. Changes in daily precipitation and surface air temperature extremes in the IPCC AR4 models. US CLIVAR Variations. 2005; 3:5–9.

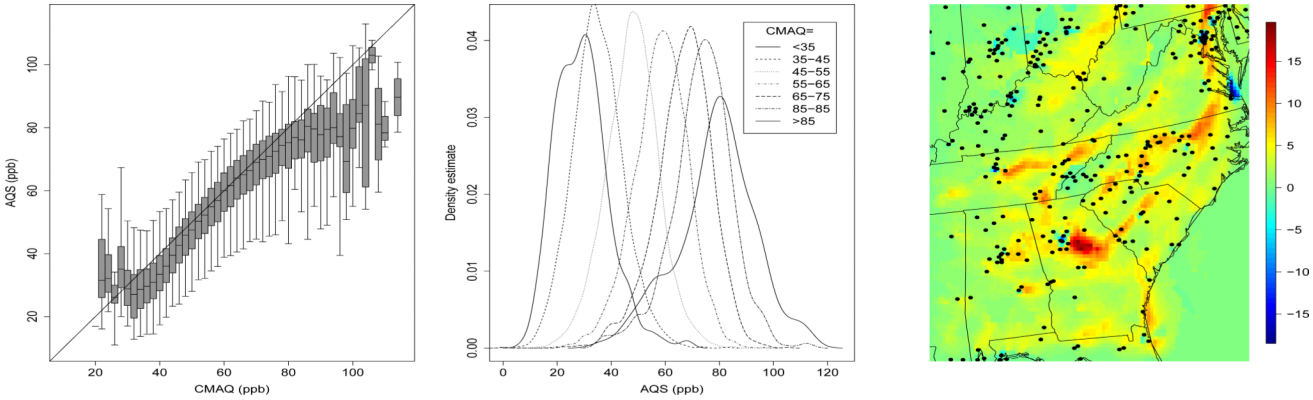
Wehner M, Smith R, Bala G, Duffy P. The effect of horizontal resolution on simulation of very extreme US precipitation events in a global atmosphere model. Climate Dynamics. 2010; 34:241–247.

Zhou J, Fuentes M, Davis J. Calibration of numerical model output using nonparametric spatial density functions. J. Agric. Biol. Environ. Stat. 2012; 16:531–553.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



(a) AQS by CMAQ (b) AQS density (c) RF-CMAQ sensitivities

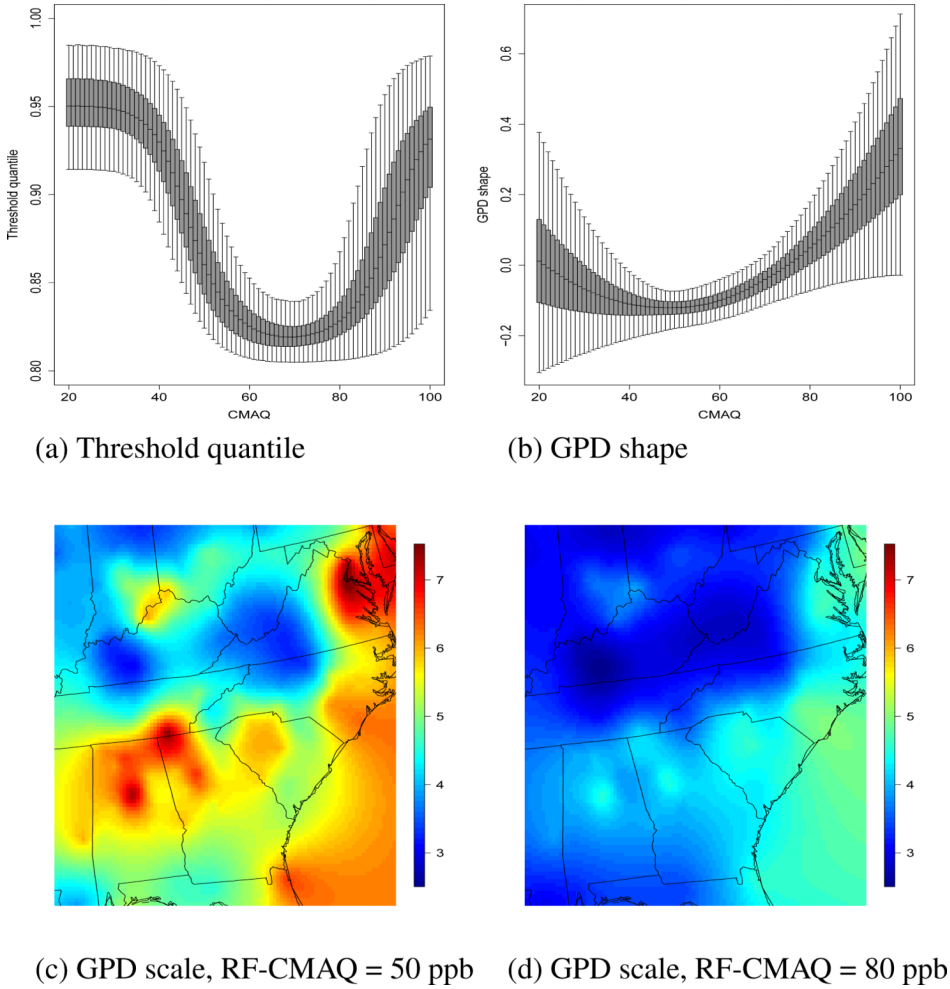
Figure 1.

Plot of the CMAQ output versus AQS monitor data pooled over all sites (Panel (a)), the kernel-smoothed density estimate of the AQS density by binned CMAQ (Panel (b)), and the RF-CMAQ sensitivity $S_1^{(1)}(1, \mathbf{s})$ for one day for mobile source NO_x (Panel (c); points are AQS monitor locations).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



(a) Threshold quantile (b) GPD shape

(c) GPD scale, RF-CMAQ = 50 ppb (d) GPD scale, RF-CMAQ = 80 ppb

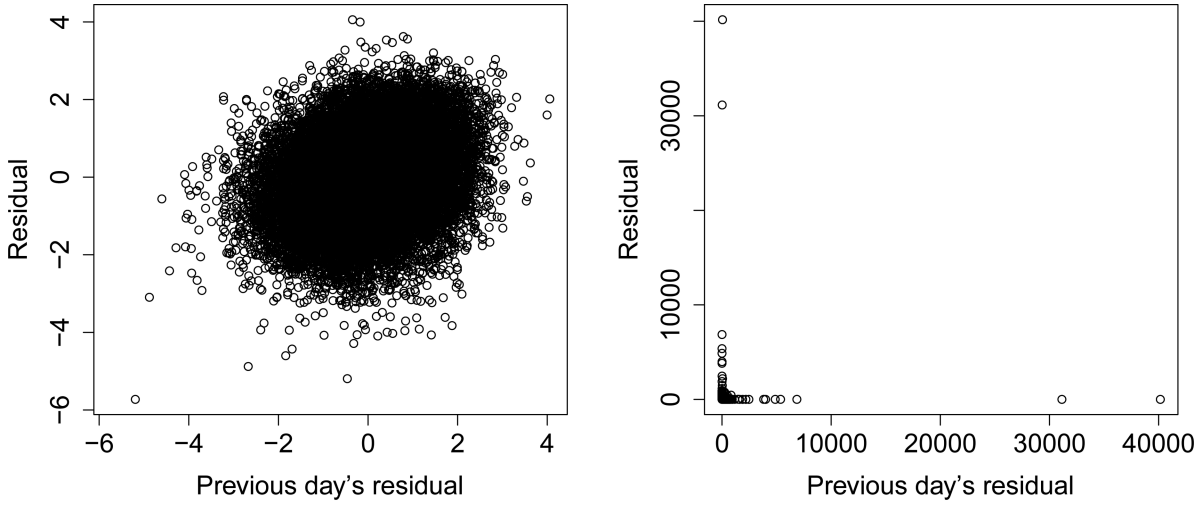
Figure 2.

Posterior of the threshold $T(C)$, GPD shape $\xi(C)$, and GPD scale $\sigma(C, \mathbf{s})$ by RF-CMAQ output. In Panels (a) and (b), the horizontal line in each boxplot gives the median, interquartile range, and 95% interval. Panels (c) and (d) plot the posterior mean.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



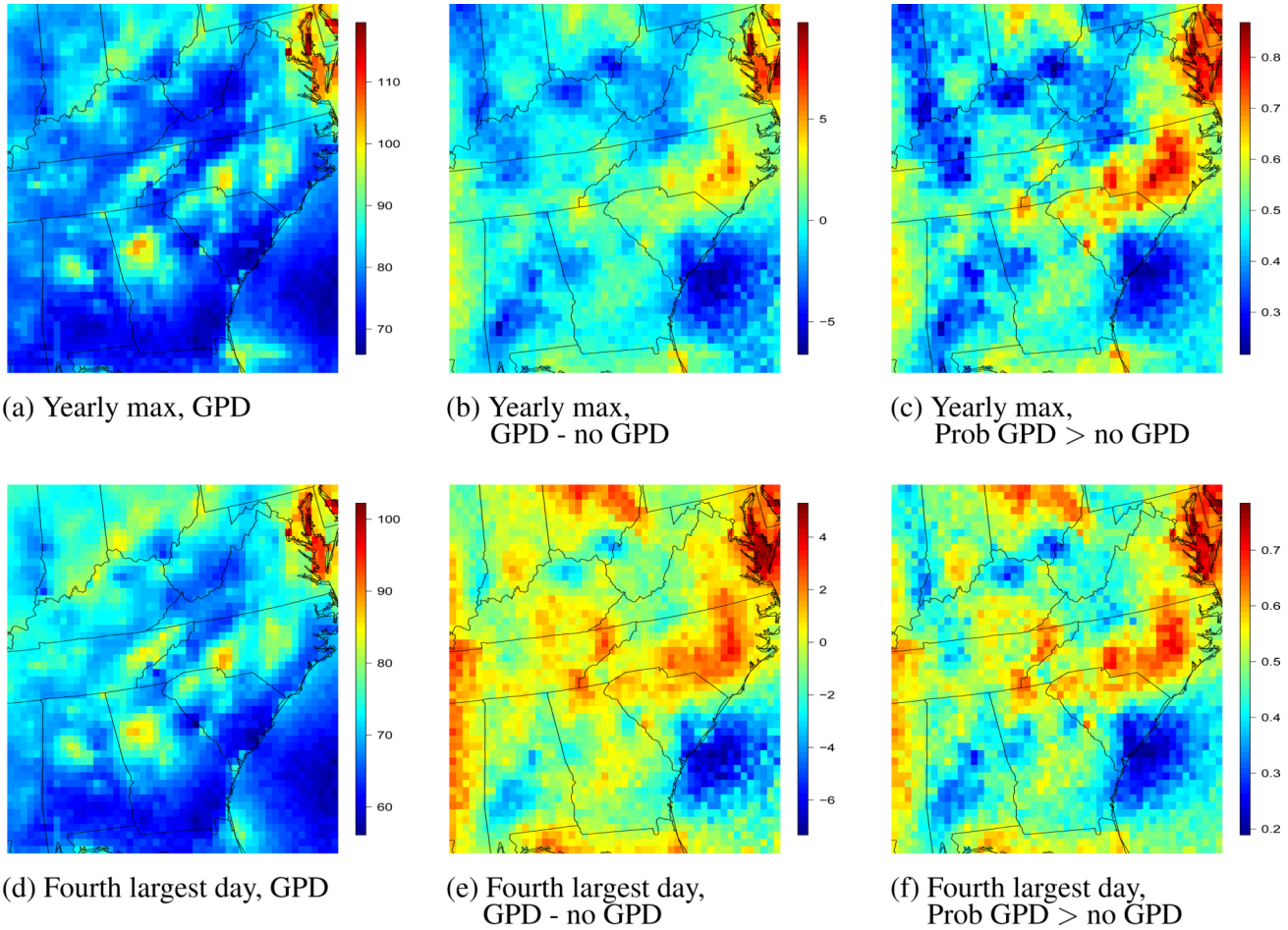
(a) Gaussian marginals (b) Fréchet marginals

Figure 3.
Plots of the residuals for consecutive days at the same locations, transformed to (a) standard normal and (b) unit Fréchet marginals.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



(a) Yearly max, GPD

(b) Yearly max, GPD - no GPD

(c) Yearly max, Prob GPD > no GPD

(d) Fourth largest day, GPD

(e) Fourth largest day, GPD - no GPD

(f) Fourth largest day, Prob GPD > no GPD

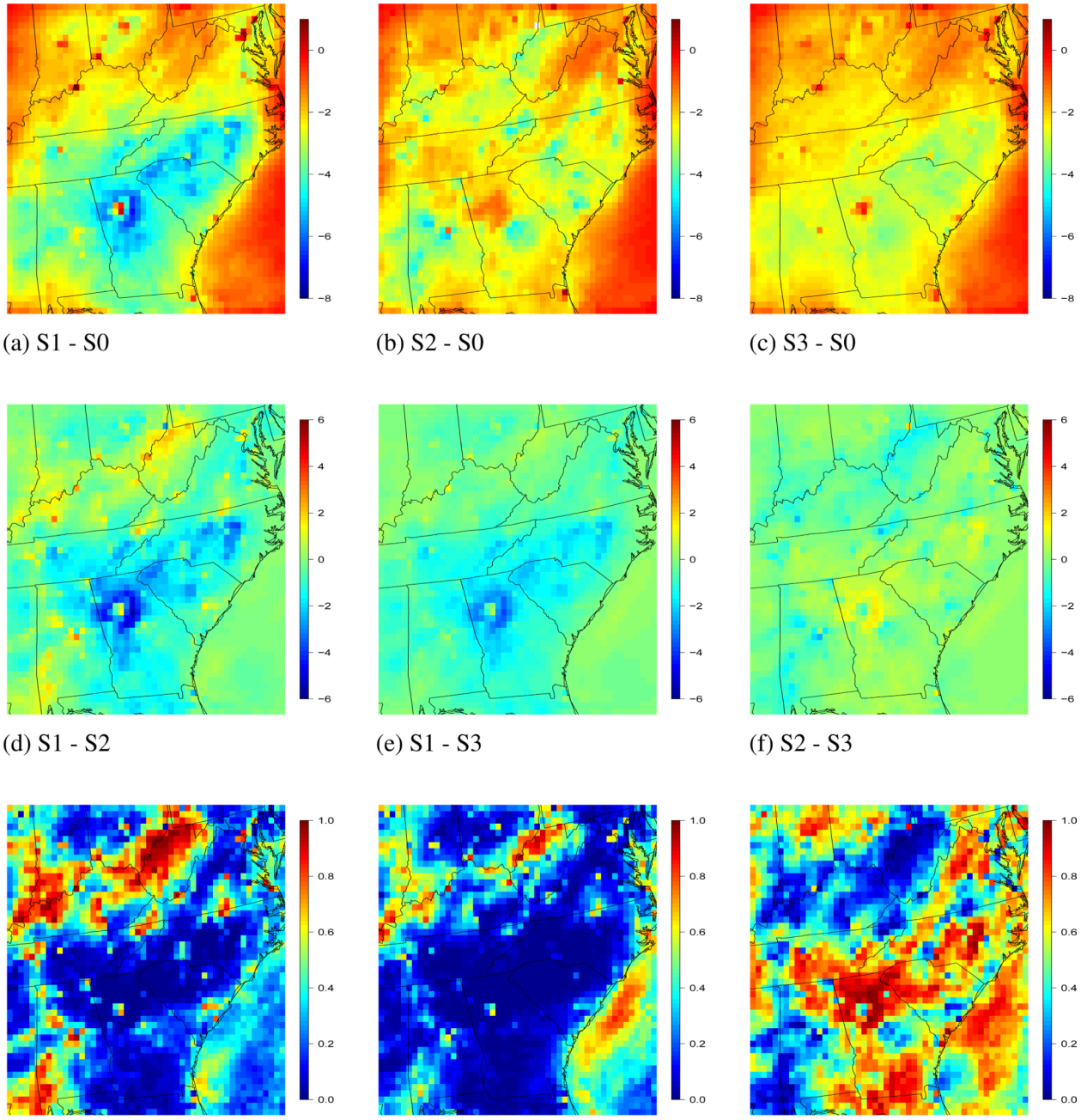
Figure 4.

Comparison of the predictive distribution for the yearly maximum and yearly 4th highest value (ppb) for the base case with ("GPD") and without ("noGPD") Generalized Pareto tails. Plotted are the posterior mean for the GPD model (Panels (a) and (d)), the posterior mean difference between GPD and noGPD (Panels (b) and (e)), and the posterior probability that GPD gives a larger value than noGPD (Panels (c) and (f)).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



(a) S1 - S0 (b) S2 - S0 (c) S3 - S0

(d) S1 - S2 (e) S1 - S3 (f) S2 - S3

(g) Prob S1 > S2 (h) Prob S1 > S3 (i) Prob S2 > S3

Figure 5.

Difference in the posterior predictive mean of the fourth highest day of the year for the several pairs of scenarios (Panels (a)-(f); parts per billion) and probability that the fourth highest day of the year is larger for one scenario than another (Panels (g)-(i)). The probabilities of a reduction from S0 are near one for all other scenarios and thus not shown.

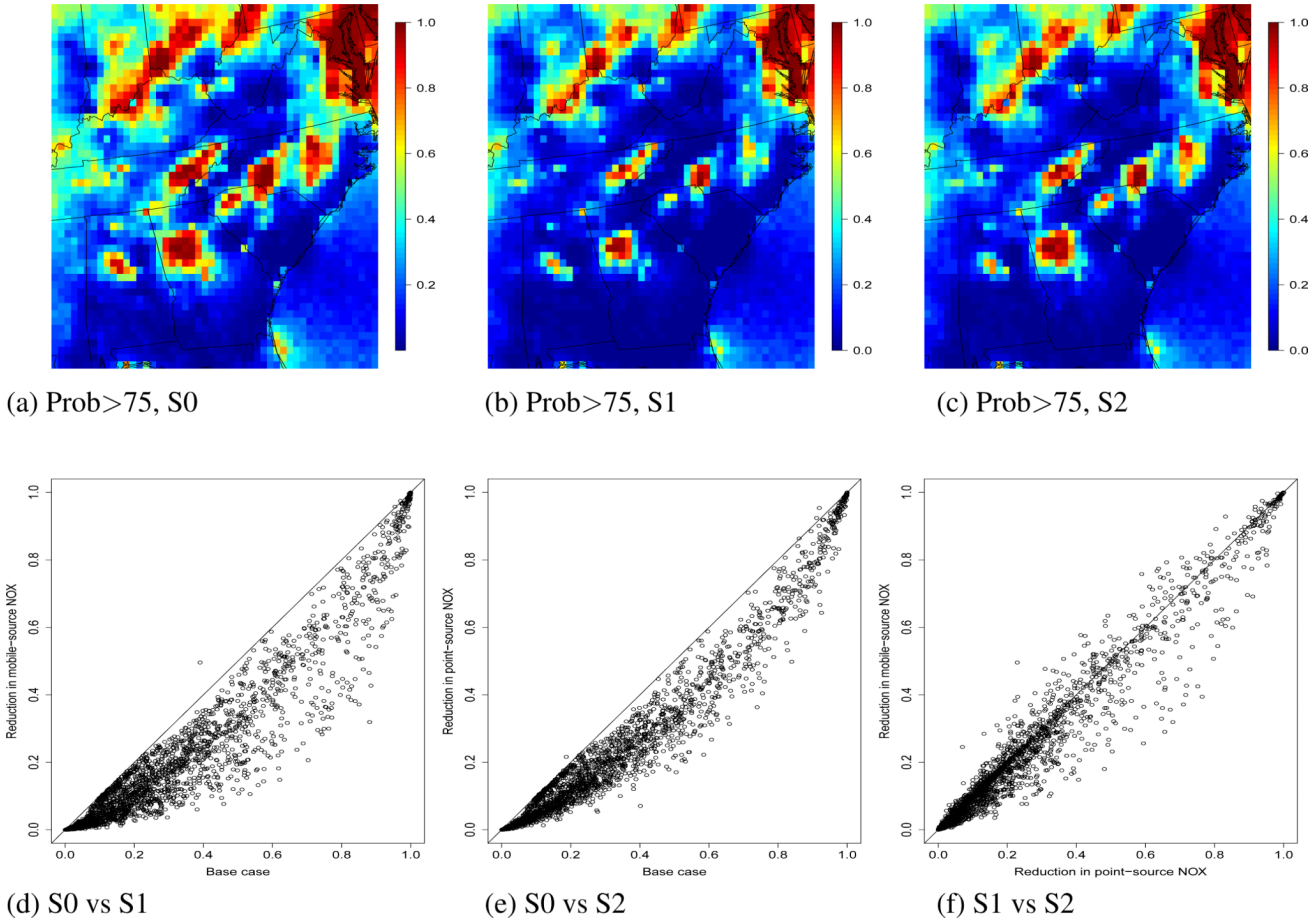


Figure 6.

Comparison of the predictive probability of greater than 75ppb for the yearly 4th highest value under three control strategies. Each point in Panels (d)-(f) corresponds to one of the grid cells in Panels (a)-(c).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

Quantile and Brier scores for various models. Models vary by the number of basis func tions in the quantile process (L), the degree of polynomial expansion of the RF-CMAQ predictors (M), and whether the upper tail is (GPD) or is not (NoGPD) a generalized Pareto distribution. "SLR", "QR", and "LR" are simple linear regression, quantile regression, and logistic regression, respectively, with linear predictor $a(\mathbf{s})+b(\mathbf{s})C_0(\mathbf{s}, t)$, where $a(\mathbf{s})$ and $b(\mathbf{s})$ are estimated separately by site and $C_0(\mathbf{s}, t)$ is the full CMAQ output. The lowest value (including ties) for each criteria are in bold.

(a) Quantile scores (ppb)

			$L = 1$				$L = 4$			
Quantile			$M = 1$		$M = 2$		$M = 1$		$M = 2$	
level	**SLR**	**QR**	**NoGPD**	**GPD**	**NoGPD**	**GPD**	**NoGPD**	**GPD**	**NoGPD**	**GPD**
0.750	7.73	5.51	5.37	5.35	5.33	**5.30**	5.35	5.35	5.31	5.32
0.950	7.66	2.02	1.85	1.84	1.84	**1.81**	1.84	1.84	1.83	1.83
0.990	7.66	0.768	0.521	0.507	0.518	**0.498**	0.520	0.505	0.515	0.502
0.995	7.64	0.600	0.302	0.287	0.299	**0.284**	0.301	0.287	0.299	0.289

(b) Brier scores (multiplied by 100)

			$L = 1$				$L = 4$			
			$M = 1$		$M = 2$		$M = 1$		$M = 2$	
Threshold	**SLR**	**LR**	**NoGPD**	**GPD**	**NoGPD**	**GPD**	**NoGPD**	**GPD**	**NoGPD**	**GPD**
70	9.38	6.88	6.18	6.17	6.13	**6.11**	6.19	6.21	6.12	6.14
75	5.91	4.35	3.89	3.89	3.79	**3.76**	3.90	3.94	3.80	3.81
80	3.15	2.43	2.11	2.11	1.98	**1.95**	2.12	2.15	1.98	1.97
85	1.554	1.061	0.985	0.999	0.866	**0.852**	0.997	1.022	0.859	**0.852**
90	0.791	0.458	0.427	0.440	0.350	0.344	0.430	0.445	0.343	**0.341**
95	0.418	0.277	0.225	0.229	0.184	**0.182**	0.226	0.229	**0.182**	**0.182**
100	0.198	0.127	0.098	0.098	0.078	**0.077**	0.098	0.098	0.079	0.078