ABSTRACT

ZHOU, JINGWEN. Calibration of Numerical Model Output Using Nonparametric Spatial Density Functions. (Under the direction of Montserrat Fuentes.)

Studies on the health impacts of climate change routinely use climate model output as future exposure. However, climate model simulations are deterministic, thus resolving the differences between climate model output and the observed data in terms of their entire distributions is of critical interest for air quality management. For health impact analysis, uncertainty quantification is an important component of the risk assessment process and has been limited to the variability due to different emission scenarios or multi-model ensembles in the form of sensitivity analysis. While the uncertainty associated with climate model output is well recognized, it is difficult to assess its magnitude for any single climate model.

Spatial quantile calibration approach is motivated by the need to better characterize the tails of future exposure distributions where the greatest health impacts are likely to occur. Specifically, quantile calibration seeks to model a monotonic transformation function from the entire quantile process of the model output distribution to the quantile process of the observed data distribution jointly. To this end, we use a Bayesian framework to model the quantile processes. The corresponding likelihood approximation is innovatively derived by a mixture density where the central tendency is given by the estimated quantile functions and the tails follow a generalized Pareto distribution. In addition, spatial techniques have been developed to incorporate spatial dependence into the model, which allows us to downscale the gridded climate model output to point-level for projecting exposure over a specific geographical region. By modeling climate model output to reproduce small-scale weather events in a historical period where observations are available, this dissertation not only calibrates future model projections that incorporate a spatial adjustment of the entire distribution of the model output with respect to the distribution of the monitoring data, but also accounts for projection uncertainty in the final health impact estimates.

In each chapter, we present an application for the introduced method. Our approach is able to achieve reduction in the root mean square error (RMSE) compared to the default approach based on the first two moments of the climate model output and the observed data.

In Chapter 1, we motivate this work by introducing and demonstrating the utility of both quantile calibration and spatial adjustment separately.

In Chapter 2, we present the spatial quantile process. We apply a nonlinear monotonic regression approach to the quantile functions taking into account the spatial dependence. We investigate the properties of this spatial quantile process, and present a simulation example to illustrate the utility of this model.

In Chapter 3, we extend the single spatial quantile process to construct the calibration function which maps the climate model quantile process to the quantile functions estimated from the observed data. Subsequently, a spatial adjustment of the entire distribution of the model output with respect to the distribution of the monitoring data was specified to downscale the calibration function and their spatial-quantile processes. As an example, we apply this calibration approach to real ozone data.

In Chapter 4, we outline the modeling steps for risk estimation and how to utilize calibrated model projections to conduct health impact analyses associated with extreme weather conditions. As an example, we apply the methodology to calibrate temperature projections from a regional climate model for the period 2041 to 2050. Accounting for calibration uncertainty, we calculated the number of excess deaths attributed to future temperature for three cities in the US state of Alabama.

© Copyright 2012 by Jingwen Zhou

All Rights Reserved

Calibration of Numerical Model Output Using Nonparametric Spatial Density Functions

by
Jingwen Zhou

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2012

APPROVED BY:

Sujit Ghosh

Brian Reich

Jerry Davis

Frederick Bingham

Montserrat Fuentes
Chair of Advisory Committee

DEDICATION

This thesis is dedicated to my parents who have supported me all the way since the beginning
of my research.

BIOGRAPHY

I was born in Xian, China and studied mathematics and applied mathematics at the University of Science and Technology of Beijing (USTB) in 2000. Following my undergraduate studies, I received a M.S. in statistics from University of Science and Technology of China (USTC), specializing in Bayesian statistics. In year 2007, I entered the Department of Statistics at the North Carolina State University (NCSU).

While at the NCSU, I was taught Experimental Design, Data Mining and Machine Learning, Measure Theory, Advanced Inference, Time Series, Spatial Statistics and Bayesian Statistics. These theoretical classes helped me obtain a deeper understanding of statistical methods development.

In 2009, I was supported as a research assistant, working in support of ongoing research projects regarding the environment science, especially in aspects of field (model output) data collection and analysis. Based on an application for describing the impact of climate change on the public health, I was interested in methods for calibrating the model output when the observations are not available, which turned out to be the topic of my doctoral thesis. I also frequently used geostatistical tools in my work: mapping pollutants distributions, analyzing spatial relationships, and examining spatial-quantile changes over time. I spent three years and completed my doctorate thesis in March 2012.

ACKNOWLEDGEMENTS

It would not have been possible to write this doctoral thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

Above all, I would like to thank my principal supervisor Prof. Fuentes for her help, support and patience, not to mention her advice and unsurpassed knowledge of statistics. The good advice, support and friendship of my second supervisor, Dr. Davis, has been invaluable on both an academic and a personal level, for which I am very grateful. I would like to acknowledge the financial, academic and technical support of the Department of statistics and its staff, particularly in the award that provided the necessary financial support for this research. The library facilities and computer facilities of the University have been indispensable.

I am most grateful to Dr. Reich for providing me with computer files of his work, which have been a valuable and reliable method of checking the published editions and have made referencing quotations from the texts far more efficient. I would like to thank Alison for her kindness, friendship and support. I remember the generosity and encouragement of Dr. Arroway, when I first became interested in NCSU. All her efforts in promoting a stimulating and welcoming academic and social environment will stand as an example to those that succeed them.

I would like to thank my friends Hans Huang for his personal support and great patience at all times. My parents have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice. Last, but by no means least, I would also like to thank my colleagues and friends Brenda, Lan, Lixia and elsewhere for their support and encouragement throughout.

For any errors or inadequacies that may remain in this work, of course, the responsibility is entirely my own.

TABLE OF CONTENTS

List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Motivation	3
1.2 Literature review	6
1.2.1 Quantile functions and density quantile approach	6
1.2.2 Frequentist method for quantile regression	9
1.2.3 Bayesian method for quantile regression	11
1.3 Spatial models	12
1.4 Notation	13
1.5 Outline	14
Chapter 2 Spatial quantile process	16
2.1 Spatial quantile regression	16
2.1.1 Spatial-quantile function	16
2.1.2 Estimation	19
2.2 Spatial-temporal quantile regression	22
2.3 Simulation study	22
2.4 Summary	23
Chapter 3 Spatial quantile calibration	26
3.1 Spatial quantile calibration of deterministic model output	26
3.2 Spatial-temporal quantile calibration	28
3.3 Estimation	29
3.4 Application: calibration of eastern US ozone data	32
3.4.1 Data description	32
3.4.2 Result analysis and discussion	33
Chapter 4 Mortality prediction with calibrated meteorological variables	42
4.1 System calibration and spatial quantile processes	43
4.2 Health Effect Estimation and Impact Analysis	46
4.3 Estimation	48
4.4 Case study of the heat wave	49
4.4.1 Data description	49
4.4.2 Result analysis	50
4.5 Summary	53
Chapter 5 Discussion and Future work	63
References	67

Appendix	70
Appendix A Appendix	71
A.1 Proper posterior distribution	71
A.2 Fitting the Generalized Pareto Distribution to the tails	72

LIST OF TABLES

Table 2.1	(Example 2) Empirical root mean integrated squared error ($\times 100$), with its standard error in parentheses	23
Table 4.1	Posterior medians (95% credible intervals) of the annual excess deaths attributable to high temperature and heat waves across three Alabama cities. For high temperature, the baseline group is defined as days with maximum temperature under 30°C and the at-risk group is defined as days with maximum temperature above the threshold. For heat wave, the at-risk group is defined as days with maximum temperature above the quantile threshold.	61
Table 4.2	Observed and projected average number of high temperature days and heat wave days per year in Birmingham, Alabama. High temperature days occur when maximum daily temperature exceeds the threshold temperature. Heat waves are defined as consecutive days with maximum daily temperature exceeding the quantile threshold. Posterior medians (95% credible intervals) are also given for the projections.	62

LIST OF FIGURES

Figure 1.1	The linear regression (LM) predictor for CMAQ at a selected monitoring site. The left panel shows the corresponding $100\tau^{th}$ quantile curves of the monitoring data and linear-transformed model output, units in parts per billion (ppb); while the right panel is the corresponding quantile-quantile plot.	5
Figure 1.2	Calibration function G_τ with respect to the underlying quantile processes. . . .	8
Figure 1.3	Quantile regression splines can cross each other, so it does not provide a valid quantile function.	10
Figure 2.1	The comparison between the I-splines of order $h=3$ and the integrated Bernstein polynomials of degree 4. The I-splines are defined on $[0, 1]$ and are associated with interior knots 0.2, 0.8. Each I-spline is piecewise cubic, and a monotonic curve is obtained by the linear combination of these $M=5$ splines with non-negative coefficients. The I-spline function is used to ensure the non-crossing quantiles with both flexibility and constrains at different percentiles.	17
Figure 2.2	The probability density function for the generalized Pareto distribution is plotted with respect to the different tail index parameters κ . The parameter ranges $\kappa > 0$, $\kappa = 0$, and $\kappa < 0$ correspond to the short tails, exponential type and long tails, respectively.	21
Figure 2.3	Bayesian nonparametric quantile (BSQ) regression from Example 2. Interior knots are placed at 0.2, 0.8 with $\delta = 0.005$ and $\kappa = 0$. We add a sin function to mimic the temporal trend in reality. We obtain the conditional quantile estimates for $\tau = 0.01, 0.5, 0.9$ and 0.99 . Notice that the classic quantile regression spline generates crossing quantile curves, which does not provide a valid quantile process.	25
Figure 3.1	A process chart for spatial quantile calibration for going from model output to the observations. We calibrate the original model output with the corresponding observations through their underlying spatial-quantile processes. Both of the model output quantile modeling and the adjustment of the entire distribution of model output with respect to the distribution of monitoring data need to be across space.	36
Figure 3.2	Maps of the sample 90^{th} quantile levels of the ozone concentrations for both the CMAQ output and observations; At a randomly selected site (i.e., 59^{th}) where we have $T=153$ observations; we use " *" to represent its location. The histogram, sample quantile and density function of the two data sources are provided to identify the different behave in their distribution.	37
Figure 3.3	Quantile comparison plots. Maps of the 5^{th} , 50^{th} and 95^{th} quantiles for the Bayesian estimated CMAQ model output and monitoring data are plotted. The results show the agreement between the distributions of the CMAQ output and the monitoring data at their median level, but large differences appear for the tails.	38

Figure 3.4	The Bayesian spatial-quantile calibration for CMAQ at a selected monitoring site. The left panel shows the corresponding $100\tau^{th}$ quantile curves of the monitoring data and G_τ -transformed model output, units in parts per billion (ppb); while the right panel is the corresponding quantile-quantile plot.	39
Figure 3.5	The scale predictor for CMAQ at a selected monitoring site. The left panel shows the corresponding $100\tau^{th}$ quantile curves of the monitoring data and scale-transformed model output, units in parts per billion (ppb); while the right panel is the corresponding quantile-quantile plot.	40
Figure 3.6	The 95^{th} quantile for the monitoring data (units in ppb), using both the quantile calibration and linear regression method. We compare the differences between the linear regression and the Bayesian quantile calibration methods in terms of the <i>RMSE</i>	41
Figure 4.1	A process chart for our two-stage estimation. We first obtained the calibration model by comparing the original NARCCAP data with the corresponding observations through their underlying spatial-quantile processes in year 2000. We then fitted the health model to investigate the relationship between heat waves and mortality. Finally, we evaluated the future heat wave excess mortality based on the calibrated NARCCAP maximum temperature from 2041 to 2050.	55
Figure 4.2	Likelihood estimation without specifying a density function a priori. The blue curve is a density plot of simulated data. The likelihood function expressed as a quantile-based central tendency and a generalized Pareto tail are able to characterize the unbounded response variables.	56
Figure 4.3	Quantile-quantile plot of the model output and the observations. The blue curve is an estimated G_τ function.	57
Figure 4.4	Locations of the monitoring sites (blue triangles), centers of the NARCCAP grid cells (red dots) and NMMAPS communities (green circles) within Alabama.	58
Figure 4.5	At three urban communities Birmingham (birm), Mobile (Mobi), and Huntsville (hunt), we plot τ^{th} quantile curves of NARCCAP outputs, observations, our Bayesian calibrations and the simple linear regression (LM) of daily maximum temperature in year 2000, units in Celsius (C).	59
Figure 4.6	τ^{th} quantile curves of the historical NARCCAP data, the corresponding observations, the uncalibrated future model output and the calibrated daily maximum temperature, unit in Celsius. Note that the discrepancies mostly occurred at the lower tail of the distribution.	60
Figure 5.1	The CMAQ and monitoring temporal quantiles at site 4. Under the non-crossing constraints, ozone quantile curves show little trend for both the CMAQ models and the monitoring data.	65
Figure 5.2	Temporal quantile surfaces on site 19 for both the CMAQ data and Observed data.	66

Chapter 1

Introduction

Climate change poses unique challenges to human health. Unlike health threats caused by a particular disease, climate change can lead to potentially harmful health effects. For instance, there are direct health impacts from heat waves, ailments caused by air pollutants such as ozone and many climate-sensitive infectious diseases. The assessment of potential health effects of climate change must include consideration of the capacity to manage new and changing climate conditions.

Projections from climate models provide state-of-the-art quantitative information on future climate. Recently, climate model output has been utilized to quantify the health impacts of various environmental risk factors due to climate change. Using a large number of grid cells, they generate averaged concentrations which have full spatial coverage and high temporal resolution without missing values. However climate model simulations are deterministic, and uncertainty may be present in various modeling stages that attempt to represent the underlying physical processes with numerical models [Knutti, 2008]. Uncertainties about the model outputs should be recognized [Kennedy and O'Hagan, 2001; Paciorek, 2011]. The various sources of such uncertainties are classified as low quality of emissions data, model inadequacy and residual variability [Fuentes and Raftery, 2005; Kennedy and O'Hagan, 2001; Lim et al., 2010; Paciorek, 2011]. On

the other hand, health impact analysis has focused predominantly on the variability that arise from different emission scenarios or multi-model ensembles [Peng et al., 2011]. Nevertheless, the quantification of uncertainty caused by deterministic model output has not been extensively studied.

In this thesis, we do not calibrate the computer model parameters, for instance, as in Kennedy and O’Hagan [2001]. Instead we focus on comparing, evaluating and resolving the differences in the distribution between the model outputs and the observed data. We describe a Bayesian spatial quantile regression model to calibrate climate model output and use daily maximum temperature as the motivating example. Temperature projections have been used extensively in health impacts analysis because of its well-established adverse health effects [Anderson and Bell, 2009; Curriero et al., 2002]. Moreover, temperature can act as a predictor for other environmental processes such as infectious disease transmission [Ogden et al., 2006; Remais et al., 2008], hydrological dynamics for water quality [John and Rose, 2005], or ground-level ozone creation [Bell et al., 2007; Knowlton et al., 2004]. By modeling climate model output to reproduce small-scale weather events in a historical period where observations are available, we want to not only calibrate future model projections, but also incorporate projection uncertainty in the final health impact estimates.

Early evaluation of model performance usually relies on linear least-squares analysis of observations versus model outputs. However, the model outputs and the observations are on different spatial scales; this is referred to as the “change of support” problem. The measurements are made at specific locations in the spatial domain, while modeled concentrations are recorded as averages over grid cells [Eder and Yu, 2006]. Thus the two data sources are not directly comparable. In regard to this problem, efforts have been made in the recent literature to address the spatial “incommensurability” between gridded model averages and point measurements. For instance, Fuentes and Raftery [2005] developed a “fusion” model, defining

the model outputs as the integral over a grid cell of a latent point-level process, which is also related to the observations. Additionally, to achieve computational efficiency, Berrocal et al. [2010] proposed univariate downscaler using a linear regression model with spatially-varying coefficients, thus developing a “spatial-temporal” model to adjust for the potential spatial misalignment.

Another challenge arises from the usual Gaussian assumption in the standard linear regression approach which may underestimate the tail probability for climate variables with skewed distributions [Chang et al., 2010]. For instance, a local heat wave is often defined based on daily temperatures exceeding the 95th percentile of its local summer time climatology. Another example is the U.S. EPA ozone standards which rely on the fourth highest day of the year (99th quantile). Therefore improving the ability to characterize extreme temperature events is of critical importance. As a result, we need a method which allows researchers to adjust for the spatial dependence of the two data sources, and in some sense, to obtain a complete transformation between the model output and monitoring data in terms of their entire distribution.

Further developments of these proposed evaluation procedures are needed. In this thesis, we are concerned with the discrepancy due to the shape of the distributions, especially the tails. In order to compare the density functions of the numerical models and field data, we estimate the spatial quantile functions for both models and data, and we apply a nonlinear spatial monotonic regression approach to the quantile functions. We use a Bayesian approach for estimating and fitting in order to characterize the uncertainties in the data and statistical models.

1.1 Motivation

To illustrate the necessity of the quantile regression approach, we start by using the standard framework for calibrating numerical models to field data: the least squares regression. Sup-

pose we have two data sets $Y(t, \mathbf{s}_i)$ and $Z(t, \mathbf{B}_{\mathbf{s}_i})$, where $\mathbf{s} = (s_1, s_2)$ is a point measured by EPA monitors using the latitude/longitude coordinates and $\mathbf{B}_{\mathbf{s}}$ the associated model output simulated grid cell in which $\mathbf{s}$ lies, for time $t=1,2,\dots,T$, and location $i=1,2,\dots,n_s$. Based on the comparison of means and variances, we define the linear regression (LM) model at a given location $\mathbf{s}_{i^*}$ as:

$$E(Y(t, \mathbf{s}_{i^*})|Z(t, \mathbf{B}_{\mathbf{s}_{i^*}})) = \beta_{0i^*} + \beta_{1i^*}Z(t, \mathbf{B}_{\mathbf{s}_{i^*}}), \quad (1.1)$$

In other words, the LM model fits the conditional mean of a vector of field data $Y(\cdot, \mathbf{s}_{i^*})$, given the corresponding model outputs $Z(\cdot, \mathbf{B}_{\mathbf{s}_{i^*}})$. It is known that the solution to the ordinary least squares regression is an optimization problem as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_t (Y(t, \mathbf{s}_{i^*}) - \beta_{0i^*} - \beta_{1i^*}Z(t, \mathbf{B}_{\mathbf{s}_{i^*}}))^2. \quad (1.2)$$

In addition, if the normality assumption holds, that is:

$$Y(\cdot, \mathbf{s}_{i^*})|Z(\cdot, \mathbf{B}_{\mathbf{s}_{i^*}}) \sim \mathcal{N}(\beta_{0i^*} + \beta_{1i^*}Z(\cdot, \mathbf{B}_{\mathbf{s}_{i^*}}), \sigma^2 I_T), \quad (1.3)$$

Then the LM estimator $\hat{\boldsymbol{\beta}}$ is normally distributed and reaches the *Cramér – Rao* bound for the model, and thus is optimal in the class of all unbiased estimators. Subsequently, the least squares (LM) estimates $\hat{\beta}_{0i^*}$ and $\hat{\beta}_{1i^*}$ are used to calibrate model output as:

$$\hat{y}(t, \mathbf{s}_{i^*}) = \hat{\beta}_{0i^*} + \hat{\beta}_{1i^*}z(t, \mathbf{B}_{\mathbf{s}_{i^*}}). \quad (1.4)$$

Although the calibrated series $\hat{y}(t, \mathbf{s}_{i^*})$ provide an unbiased estimates of the field data, the p-value = 0.00576 of a Two-sample Kolmogorov-Smirnov test indicates that the LM estimator fails to transform the tails of the model output distribution to the corresponding observed distribution (see Figure 1.1) and capture the non-linear relationship between the quantiles of observed distribution and the quantiles of the model output distribution (see the quantile-quantile plot in Figure 1.1). Thus a more sophisticated approach is needed to provide an

adequate modeling platform when different parts of the observed distribution are suspected to transform from the corresponding parts of the model output distribution at different rates.

Figure 1.1: The linear regression (LM) predictor for CMAQ at a selected monitoring site. The left panel shows the corresponding $100\tau^{th}$ quantile curves of the monitoring data and linear-transformed model output, units in parts per billion (ppb); while the right panel is the corresponding quantile-quantile plot.

1.2 Literature review

1.2.1 Quantile functions and density quantile approach

Heavy tail distributions are important in many scientific and economical studies. Their applications include use in the analysis of the impact of extreme event on air quality, and in the modeling of large insurance claims, among others. In general, it is hard to choose an appropriate distribution for a specific application a priori. To this end it is useful to determine probability distributions by their tail behavior. Specifically, the *density-quantile functions*, introduced by Parzen [1979], was widely used as a measure of tail orderings. Given data $z(1, \mathbf{B}_{s_{i*}}), z(2, \mathbf{B}_{s_{i*}}) \dots, z(T, \mathbf{B}_{s_{i*}})$, we want to find statistical patterns in the data. That is, we seek to model the distribution function $F_Z(z) = P(Z(\cdot, \mathbf{B}_{s_{i*}}) \leq z)$ and the probability density $f_Z(z) = F'_Z(z)$. To develop nonparametric statistical data modeling based on estimating the *quantile functions* (inverse CDF), we define $q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(\tau)$ as:

$$q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(\tau) = F_Z^{-1}(\tau) = \inf\{z : F_Z(z) \geq \tau\}. \quad (1.5)$$

Consequently, $Z(\cdot, \mathbf{B}_{s_{i*}})$ is identically distributed as $q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(U)$, where U is uniformly distributed on $[0,1]$, because:

$$P(q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(U) \leq z) = P(U \leq F_Z(z)) = F_Z(z). \quad (1.6)$$

Such identically distributed properties can be rewritten as:

$$Z(\cdot, \mathbf{B}_{s_{i*}}) \sim q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(U).$$

Suppose F_Z is a continuous distribution function, then $F_Z(q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(\tau)) = \tau$ for all τ . Thus, by taking derivatives, it follows that

$$f_Z(q_{Z(\cdot, \mathbf{B}_{s_{i*}})}(\tau))q'_{Z(\cdot, \mathbf{B}_{s_{i*}})}(\tau) = 1. \quad (1.7)$$

Here $f_Z(q_Z(\cdot, \mathbf{B}_{s_i^*})(\tau))$ is defined to be the density quantile function and model (1.7) implies:

$$f_Z(q_Z(\cdot, \mathbf{B}_{s_i^*})(\tau)) = \frac{1}{q'_{Z(\cdot, \mathbf{B}_{s_i^*})}(\tau)}. \quad (1.8)$$

An examination of the density-quantile functions $f_Z(q_Z(\tau))$ of familiar probability laws indicates that they can be classified according to their limiting behavior as τ tends to 0 or 1. The behavior when $\tau \rightarrow 1$ is described as by Parzen [1979]:

$$f_Z(q_Z(\cdot, \mathbf{B}_{s_i^*})(\tau)) \sim (1 - \tau)^v, v > 0. \quad (1.9)$$

Tail behavior is classified as (1) short tails or limited type ($v < 1$); (2) medium tails or exponential type ($v = 1$); (3) long tails or Cauchy type ($v > 1$). Additionally, Tokdar and Kadane [2011] provide an efficient algorithm to solve $\tau(z)$ of the equation $z = q_Z(\cdot, \mathbf{B}_{s_i^*})(\tau)$ in τ . We make use of the model output that the density quantile function (1.8) lends itself to a likelihood function which could be further used in the Bayesian analysis (see Section 2.1.2).

However, researchers specializing in calibration techniques are specifically interested in investigating the difference across quantiles. Indeed, another important property of quantile functions is how they behave under transformations of random variables. For instance, in order to take the entire distributions of both model outputs and field data into account, let the $100\tau^{th}$ quantile process of the observations be denoted by $q_Y(\tau, \mathbf{s}_i^*)$ and the model output as $q_Z(\tau, \mathbf{B}_{s_i^*})$. Similarly, denote the distribution functions of Y and Z by F_Y and F_Z respectively. Now consider the calibration function G_τ . In fact, $G_\tau(\cdot)$ is a monotonic function of the quantile functions of model output. For instance, given the quantile level τ , we use the quantile functions $q_Y(\cdot, \mathbf{s}_i^*)(\tau)$ and $q_Z(\cdot, \mathbf{B}_{s_i^*})(\tau)$ to construct $G_\tau(\cdot)$ as follows:

$$F_Y(y) = F_Z(G_\tau^{-1}(y)). \quad (1.10)$$

In addition, if F_Z is a strictly increasing continuous distribution, then:

$$q_Y(\tau, \mathbf{s}_{i*}) = G_\tau(q_Z(\tau, \mathbf{B}_{s_{i*}})). \quad (1.11)$$

In general, if the data $z(1, \mathbf{B}_{s_{i*}}), z(2, \mathbf{B}_{s_{i*}}), \dots, z(T, \mathbf{B}_{s_{i*}})$ are assumed to be identically distributed as $Z(\cdot, \mathbf{B}_{s_{i*}})$, and if the quantile function of Z can be transformed to the quantile function of a random variable Y by an increasing continuous transformation G_τ , then the transformed data $G_\tau(z(1, \mathbf{B}_{s_{i*}})), G_\tau(z(2, \mathbf{B}_{s_{i*}})), \dots, G_\tau(z(T, \mathbf{B}_{s_{i*}}))$ are identically distributed as $Y(\cdot, \mathbf{s}_{i*})$:

$$\begin{aligned} P(G_\tau(Z(\tau, \mathbf{B}_{s_{i*}})) \leq q_Y(\tau, \mathbf{s}_{i*})) &= P(G_\tau(Z(\tau, \mathbf{B}_{s_{i*}})) \leq G_\tau(q_Z(\tau, \mathbf{B}_{s_{i*}}))) \\ &= P(Z(\tau, \mathbf{B}_{s_{i*}}) \leq q_Z(\tau, \mathbf{B}_{s_{i*}})) = \tau \end{aligned} \quad (1.12)$$

In summary, in order to characterize a transformation from model outputs to the monitoring data, we are seeking an appropriate function G_τ through their underlying quantile processes (See Figure 1.2). In addition, examining the changes between the model outputs and monitoring quantiles can yield a calibration of their overall distributions. This quantile calibration is used to target a specific part of the distribution of Y , encoded by the corresponding part of the distribution of Z . In particular, interest focuses on a suitable framework to study these changes in the form of quantile processes, where the rate of the change can be different in the tails and other noncentral or central parts of the distributions.

Figure 1.2: Calibration function G_τ with respect to the underlying quantile processes.

1.2.2 Frequentist method for quantile regression

Quantile regression[Koenker, 2005] is an increasingly popular choice to complement the LM model. Modeling of the median is a more robust technique than mean regression when dealing with outlying observations. At each quantile level, we model $q_{Z(t, \mathbf{B}_{s_{i^*}})}(\tau)$ (1.5) as:

$$q_{Z(\cdot, \mathbf{B}_{s_{i^*}})}(\tau) = \beta_1(\tau), \quad (1.13)$$

Where $\beta_1(\tau)$ is the coefficient for the τ^{th} quantile level. In general, the classical τ^{th} sample quantile estimate is obtained by:

$$\begin{aligned} \hat{\beta}_1(\tau) &= \arg \min_{\beta_1 \in R} \sum_{t=1}^T \rho_{\tau}(z(t, \mathbf{B}_{s_{i^*}}) - \beta_1) \\ &= \arg \min_{\beta_1 \in R} \left[(\tau - 1) \sum_{z(t, \mathbf{B}_{s_{i^*}}) < \beta_1} (z(t, \mathbf{B}_{s_{i^*}}) - \beta_1) + \tau \sum_{z(t, \mathbf{B}_{s_{i^*}}) \geq \beta_1} (z(t, \mathbf{B}_{s_{i^*}}) - \beta_1) \right] \end{aligned}$$

Further, let $\mathbf{u}_t = (1, u_{t_2}, \dots, u_{t_J})$ for $t = 1, \dots, T$, correspond to the J-1 degree spline representation of the nonlinear temporal component; we introduce a nonlinear temporal trend in the quantile process as follows:

$$q_{Z(\cdot, \mathbf{B}_{s_{i^*}})}(\tau | \mathbf{u}_t) = \mathbf{u}_t' \boldsymbol{\beta}(\tau), \quad (1.14)$$

Given the distribution function of $Z(t, \mathbf{B}_{s_{i^*}})$, $\boldsymbol{\beta}(\tau)$ can be obtained by solving:

$$\boldsymbol{\beta}(\tau) = \arg \min_{\boldsymbol{\beta} \in R^J} E(\rho_{\tau}(z(t, \mathbf{B}_{s_{i^*}}) - \mathbf{u}_t' \boldsymbol{\beta})), \quad (1.15)$$

Solving the sample analog gives the estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}(\tau) = \arg \min_{\boldsymbol{\beta} \in R^J} \sum_{t=1}^T (\rho_{\tau}(z(t, \mathbf{B}_{s_{i^*}}) - \mathbf{u}_t' \boldsymbol{\beta})), \quad (1.16)$$

However, different quantile levels are analyzed separately. When researchers want to use quantile regression at multiple percentiles, the quantile regression splines can cross each other,

Figure 1.3: Quantile regression splines can cross each other, so it does not provide a valid quantile function.

thus it does not provide a valid quantile function. Consider an example of quantile regression to model the temporal effect on the ozone series in Figure 1.3. The goal is to examine the 8-hour maximum ozone for both model output and monitoring data based on the temporal components $\mathbf{u}_t$. We use these data to fit a quantile regression at the set of percentiles (0.01, 0.05, 0.1, 0.9, 0.99). The fitted curves of the quantile functions give the effects across time. Conditioning on $t = 53$, the 90th percentile of the distribution for model output is even higher than the 99th percentile, indicating an invalid quantile function. Also, because little data are obtained at the upper quantiles (extreme event), it is even more problematic to fit individual quantile curves.

In regard to this problem, He [1997] proposed a location-scale shift method to estimate the multiple quantile curves while ensuring non-crossing. However, as noted by Bondell et al. [2010], the response distribution may be affected by the predictors in a less structured manner, thus a more general development of non-crossing regression quantile is needed. In addition,

Wu and Liu [2009] proposed a stepwise method by fitting the multiple quantile sequentially to ensure non-crossing. Although this method provides a flexible model, the algorithm may depend on the order that the quantiles are fitted. Another computationally efficient approach is linear programming. Bondell et al. [2010] discussed the performance of adding constrained estimators to the typical regression quantiles as follows:

$$\begin{aligned} \hat{\beta}(\tau) = \arg \min_{\beta \in R^J} & \sum_{k=1}^K \sum_{t=1}^T (\rho_{\tau_k}(z(t, \mathbf{B}_{s_i^*}) - \mathbf{u}_t' \beta_{\tau_k})), \\ \text{subject to } & \mathbf{u}_t' \beta_{\tau_k} \geq \mathbf{u}_t' \beta_{\tau_{k-1}}, \quad \text{for all } k = 2, \dots, K \end{aligned} \quad (1.17)$$

Arguably, these multiple quantile fittings in a discrete manner may depend on the number and location of included quantiles. Tokdar and Kadane [2011] proposed a simultaneous quantile fitting approach to describe the full potential of the model: $\{q_{Z(\cdot, \mathbf{B}_{s_i^*})}(\tau | \mathbf{u}_t) = \mathbf{u}_t' \beta(\tau); \text{ for all } 0 \leq \tau \leq 1\}$.

1.2.3 Bayesian method for quantile regression

In general, quantile functions are useful in determining probability distributions by their tail behavior, especially for extreme events. To this end we do not choose a distribution for a specific application a priori, and let the quantile function lead us to a density function. In such a situation there is uncertainty about the distribution; the Bayesian nonparametric methods are useful. However, the non-fully specified likelihood makes a posterior density hard to calculate.

To solve this problem, Lavine [1995] introduced a substitution likelihood approach which depends only on $\boldsymbol{\theta}$, a vector of quantiles of the distribution function F_Z . By splitting the quantile values $\boldsymbol{\theta} = (-\infty = \theta_0, \theta_1, \theta_2, \dots, \theta_K, \theta_{K+1} = \infty)$ into separate bins. Let $\mathbf{H}(\boldsymbol{\theta}) = (H_1(\boldsymbol{\theta}), H_2(\boldsymbol{\theta}), \dots, H_{K+1}(\boldsymbol{\theta}))$ be the number of corresponding data counted within the bins into

which the real line is divided by $\boldsymbol{\theta}$. The substitution likelihood is then defined as:

$$s(\boldsymbol{\theta}) = \begin{pmatrix} T \\ H_1 \ H_2 \ \dots \ H_{K+1} \end{pmatrix} \prod \Delta \mathbf{F}_{\mathbf{Z}_i}^{H_i}, \quad (1.18)$$

where $\Delta \mathbf{F}_{\mathbf{Z}} = (F_Z(\theta_1), F_Z(\theta_2) - F_Z(\theta_1), \dots, 1 - F_Z(\theta_K))$. In 2005, Dunson and Taylor [2005] applied this approximation approach in a Bayesian framework, where the posterior densities are determined by a vector of quantiles and truncated priors. However, multiple quantile fittings in this discrete manner may not satisfy the calibration techniques which are specifically interested in investigating the difference across quantiles (see Formula (1.11)).

Besides the approximation methods, The quantile regression approach also influenced early attempts at a Bayesian analysis of (1.14). For the response constructed as $z = \mathbf{u}'_t \boldsymbol{\beta}(\tau) + \epsilon$, ϵ has the asymmetric Laplace density $f_\epsilon(\epsilon) = \text{const} \times \exp(-\epsilon(\tau - \mathbb{I}(\epsilon < 0)))$ [Yu and Moyeed, 2001]. Kozumi and Kobayashi [2011] developed a Gibbs sampling algorithm based on a location-scale mixture presentation of the asymmetric Laplace distribution from a Bayesian point of view. Although this approach achieves the computational efficiency, it analyzes different quantile levels separately.

1.3 Spatial models

In this section we focus on the analysis of spatial models, particularly, using Gaussian processes. In the case of point-level data, the location index $\mathbf{s}$ ($s=1,2,\dots,n_s$) varies continuously over a fixed subset of $\mathcal{R}^d$. Based on the normality assumption at one single location in formula (1.3), let $\mathbf{Y}(\cdot, \mathbf{s})$ be a Gaussian process with mean $\beta_0 + \beta_1 Z(\cdot, \mathbf{B}_{\mathbf{s}})$ and covariance function $c(s, s'; \phi, \sigma^2) = \sigma^2 \rho(s, s'; \phi) + \tau^2 \mathbf{I}(s = s')$ where $\mathbf{I}$ denotes the indicator function (i.e., $\mathbf{I}(s = s') = 1$ if $s = s'$ and 0 otherwise). Denote such a Gaussian process by:

$$\mathbf{Y}(\cdot, \mathbf{s}) \sim \mathcal{GP}(\beta_0 + \beta_1 Z(\cdot, \mathbf{B}_{\mathbf{s}}), \sigma^2 \rho(s, s'; \phi) + \tau^2 \mathbf{I}(s = s')) \quad (1.19)$$

Suppose that the covariance between the random variables at two locations depends on their distance with the exponential association. For instance, the covariance between $Y(\cdot, \mathbf{s})$ and $Y(\cdot, \mathbf{s}')$ is an exponential function of the interpolation distance; in other words, $Cov(Y(\cdot, \mathbf{s}), Y(\cdot, \mathbf{s}')) = \sigma^2 \phi(s, s'; \rho) + \tau^2 \mathbf{I}(s = s') = \sigma^2 \exp(-\frac{\|s-s'\|}{\rho}) + \tau^2 \mathbf{I}(s = s')$, where $\|s - s'\|$ is the distance between locations s and s' , and σ^2 and ϕ are positive parameters. Also, when $s = s'$, or $\|s - s'\| = 0$, we have $Var(Y(\cdot, \mathbf{s})) = \sigma^2 + \tau^2$.

An alternative model for the above Gaussian stationary process model (1.19) is:

$$Y(\cdot, \mathbf{s}) = \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})Z(\cdot, \mathbf{B}_s) + \epsilon(\mathbf{s}), \quad \epsilon(\mathbf{s}) \sim N(0, \tau^2) \quad (1.20)$$

where $\beta_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s}) = \beta_1 + \beta_1(\mathbf{s})$, and $\epsilon(\mathbf{s})$ is a white noise process with nugget variance τ^2 . Then $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ can be interpreted as random spatial adjustments at location $\mathbf{s}$ to the overall additive bias β_0 and the multiplicative bias of model output, respectively. By modeling $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ as bivariate zero-mean spatial Gaussian processes using the coregionalization approach, Berrocal et al. [2010] introduced a downscaling technique based on the point-level model for model output, thus obtaining the spatial interpolation at a new location from the predictive distribution, i.e., sampling $f(Y(\mathbf{s}_0)|Y(\cdot, \mathbf{s}), Z(\cdot, \mathbf{B}_s))$ at location $\mathbf{s}_0$. This downscaling model provides calibration at the local level and allows for a spatial process $Y(\cdot, \mathbf{s})$ with a flexible covariance structure. We notice that there is only the normality assumption, thus it is hard to capture the tail differences when we have the skew distributions.

1.4 Notation

In this section, we introduce the notation used in the following chapters. For the purpose of spatial calibration, we extend quantile calibration model (1.11) to the entire spatial surface. Let $\boldsymbol{\alpha}_\tau$ be a vector of calibration parameters which are assumed to variate across the quantile

level τ . Let $Q_Y(\tau|\mathbf{u}_t, \mathbf{s})$ and $Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s)$ be the column vector formed by vectorizing these n_s EPA observations. By combining the information for all points and grid cells, the ozone calibration model can be expressed as:

$$Q_Y(\tau|\mathbf{u}_t, \mathbf{s}) = \mathbf{G}_{\tau, \mathbf{s}}(Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s)) \quad (1.21)$$

The interpretation of this non-parametric model is that the quantile process of Y is monotonic after an approximate change in the “ τ ” system. Hence, if we take Q_Z as a mapping from a $\mathbb{R}^2 \tau \times t$ system to $\mathbb{R}^3 \tau \times t \times Q_Z$ quantile process system, then $\mathbf{G}_{\tau, \mathbf{s}}$ projects $\tau \times t \times Q_Z$ to the observed $\tau \times t \times Q_Y$ quantile process system. In other words, instead of using the regression methods based on the 2 moments of models and data, we are aimed at calibrating CMAQ and observations through their underlying spatial quantile processes (see Figure 4.1).

1.5 Outline

In summary, we will focus on: (1) How to calibrate the numerical models going beyond the first 2 moments; (2) Analyze all the quantile levels simultaneously; (3) Introduce a spatial adjustment of the entire distribution of deterministic model output with respect to the distribution of monitoring data; (4) Computational efficiency. In this thesis, we transform the random variables to be identically distributed through their spatial-quantile processes. Additionally, we introduce a piecewise polynomial function to ensure the non-crossing quantiles, while remaining the upper/lower tails in isolation from each other. Meanwhile, the quantile setting is generalized to be a mixture likelihood for characterizing the central tendency and tail behavior separately. Finally, we use a Bayesian framework to address this simultaneous spatial-quantile calibration.

The thesis is organized as follows. In Chapter 2, we model the model output quantile process through its quantile functions. In Chapter 3, we develop the calibration procedure and discuss

the Bayesian framework based on the estimated model output quantile processes in Chapter 2, and then adjust the spatio-temporal misalignment in the distributions. We also present an analysis of a spatio-temporal ozone data set over eastern U.S. In Chapter 4, we outline the modeling steps for risk estimation and discuss how to utilize calibrated model projections to conduct health impact analyses associated with high temperature days and heat waves. Finally, we end with some conclusions and remarks in Chapter 5.

Chapter 2

Spatial quantile process

In general, all the points $\mathbf{s}$ falling in the same grid cell $\mathbf{B}_s$ are assigned the same deterministic model output value. However, the model output and the observed data are not comparable due to such different spatial scales. In this chapter, we describe a Bayesian approach to link the spatial process in the model output to a point level process before using it for calibration.

2.1 Spatial quantile regression

2.1.1 Spatial-quantile function

We model the quantile function from the model outputs as follows:

$$Q_Z(\tau|\mathbf{B}_s) = \beta(\tau, \mathbf{B}_s) \quad (2.1)$$

where the parameter functions $\beta(\tau, \mathbf{B}_s)$ are the spatially-varying coefficients for the $100\tau^{th}$ quantile level with $\tau \in [0, 0.01, \dots, 1]$. Because $Q_Z(\tau)$ is nondecreasing in τ given a grid cell $\mathbf{B}_s$, the process $\beta(\tau, \mathbf{B}_s)$ must be constructed as a monotonic function:

$$\beta(\tau, \mathbf{B}_s) = \mathbf{I}(\tau)' \tilde{\boldsymbol{\beta}}_{\mathbf{B}_s} = \beta_0(\mathbf{B}_s) + \sum_{m=1}^M I_m(\tau) \beta_m(\mathbf{B}_s) \quad (2.2)$$

To achieve the monotonic properties, truncated power functions and polynomial basis functions are widely used in the recent literature [Cai and Dunson, 2007; Reich et al., 2011]. For instance, Bernstein basis polynomials $b_m^M(\tau) = \binom{M}{m} \tau^m (1 - \tau)^{M-m}$ reduce the complicated monotonicity constraints to a sequence of simple constraints $\beta_m - \beta_{m-1} \geq 0$, for $m = 2, \dots, M$ [Reich et al., 2011]. Alternatively, we plot the I-splines and integrated Bernstein basis polynomials $Ib_m^M(\tau) = \sum_{k=m+1}^{M+1} \frac{1}{M+1} b_k^{M+1}(\tau)$ for $m = 0, 1, \dots, M$ with constraints $\beta_m \geq 0$ in Figure 2.1. Polynomials do have a limitation: changing the behavior of $\beta(\tau, \mathbf{B}_s)$ near one value (τ_1) has radical implications for its behavior for any other value (τ_2) [Ramsay, 1988]. Thus, when M is small, the polynomial transformation will be dominated by the central portion of the distribution, and might behave unsatisfactorily in the tails. Choosing a large M helps but the computing burden becomes heavy (see Figure 2.1). This poses the problem of how to retain flexibility, while leaving the function constrained at different percentiles as desired.

Figure 2.1: The comparison between the I-splines of order $h=3$ and the integrated Bernstein polynomials of degree 4. The I-splines are defined on $[0, 1]$ and are associated with interior knots 0.2, 0.8. Each I-spline is piecewise cubic, and a monotonic curve is obtained by the linear combination of these $M=5$ splines with nonnegative coefficients. The I-spline function is used to ensure the non-crossing quantiles with both flexibility and constrains at different percentiles.

An *integrated piecewise polynomial spline* achieves these objectives by constructing the ba-

sis function from polynomials joined end-to-end. Particularly if compared with the Bernstein polynomials, the integral of non-negative functions make the monotonicity constraints easier to impose. The piecewise polynomial function is not only defined to ensure the non-crossing quantiles, but also keeps the upper/lower tails in isolation from each other. In this paper, we model the function $\mathbf{I}$ using the monotone spline regression. More specifically, we focus on the integrated splines I_m , or I-splines for the sake of brevity [Lu and Clarkson, 1999; Ramsay, 1988].

For a simple *knot sequence* $\{\gamma_1, \dots, \gamma_{M+h}\}$, M is the number of free parameters that specify the spline function having the specified continuity characteristics, and h is the degree of the piecewise polynomial I_m . For all τ , there exists m such that $\gamma_m \leq \tau < \gamma_{m+1}$. For application to the important case where $h=3$, let: $I_1^* = \frac{(\tau - \gamma_m)}{(\gamma_{m+2} - \gamma_{m+1})}$; $I_2^* = \frac{(\tau - \gamma_{m+1})^2 - (\gamma_{m+3} - \tau)^2}{(\gamma_{m+3} - \gamma_{m+1})(\gamma_{m+2} - \gamma_{m+1})}$; $I_3^* = \frac{(\gamma_{m+3} - \tau)^3}{(\gamma_{m+3} - \gamma_{m+1})(\gamma_{m+3} - \gamma_m)(\gamma_{m+2} - \gamma_{m+1})} - \frac{(\tau - \gamma_m)^3}{(\gamma_{m+3} - \gamma_m)(\gamma_{m+2} - \gamma_m)(\gamma_{m+2} - \gamma_{m+1})}$.

The I-spline I_m will be piecewise cubic, with zero for $\tau < \gamma_m$ and unity for $\tau \geq \gamma_{m+3}$, with the direct expressions:

$$I_m(\tau|\gamma) = \begin{cases} 0, & \text{if } \tau < \gamma_m \\ \frac{(\tau - \gamma_m)^3}{(\gamma_{m+1} - \gamma_m)(\gamma_{m+2} - \gamma_m)(\gamma_{m+3} - \gamma_m)}, & \text{if } \gamma_m \leq \tau < \gamma_{m+1} \\ I_1^* + I_2^* + I_3^*, & \text{if } \gamma_{m+1} \leq \tau < \gamma_{m+2} \\ 1 - \frac{(\gamma_{m+3} - \tau)^3}{(\gamma_{m+3} - \gamma_{m+2})(\gamma_{m+3} - \gamma_{m+1})(\gamma_{m+3} - \gamma_m)}, & \text{if } \gamma_{m+2} \leq \tau < \gamma_{m+3} \\ 1, & \text{if } \tau \geq \gamma_{m+3} \end{cases} \quad (2.3)$$

Because the I-spline is an integral of nonnegative splines, it yields monotone functions $\sum_{m=1}^M I_m(\tau) \beta_m(B_s)$ when combined with nonnegative values of the coefficients $\beta_m(B_s)$ (see Figure 2.1).

To ensure the quantile constraints, we introduce latent unconstrained variables $\beta_m^*(\mathbf{B}_s)$ and

take:

$$\beta_m(\mathbf{B}_s) = \begin{cases} \beta_m^*(\mathbf{B}_s) & \text{if } \beta_m^*(\mathbf{B}_s) \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.4)$$

Therefore a model using $\tilde{\beta}(\mathbf{B}_s)$ induces via (2.2) a quantile process of $Q_Z(\tau|\mathbf{B}_s)$. Without loss of generality, we choose the knots series within $\gamma_1 = 0$ and $\gamma_{M+h} = 1$. The quantile process thus satisfies the boundary conditions:

$$Q_Z(0|\mathbf{B}_s) = \beta_0(\mathbf{B}_s), \quad Q_Z(1|\mathbf{B}_s) = \beta_0(\mathbf{B}_s) + \sum_{m=1}^M \beta_m(\mathbf{B}_s) \quad (2.5)$$

Here, we rescale the model output such that the range of Z over the grid cells belongs to $[0, 1]$. In addition, assume $\beta_m^*(\mathbf{B}_s)$ have prior $\beta_m^*(\mathbf{B}_s) \sim N(\mu_m, \Sigma^m)$, with $\Sigma_{(\mathbf{B}_s, \mathbf{B}'_s)}^m = \sigma_{mB}^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\rho_{mB})$. The full conditional distribution of $\pi(\beta_m(\mathbf{B}_s)|Z)$ are then given by $f(Z|\beta_m(\mathbf{B}_s), \beta_m^*(\mathbf{B}_s)) \pi(\beta_m(\mathbf{B}_s) | \beta_m^*(\mathbf{B}_s)) \pi(\beta_m^*(\mathbf{B}_s))$, and is obtained using the Metropolis-Hastings algorithm for further calibration.

2.1.2 Estimation

In this section we obtain the likelihood function of the model output using the quantile function. Given the probability of the threshold exceedance δ , we have $q_Z(\tau) = \sum_{m=1}^M I_m(\tau) \beta_m + \beta_0$ be the underlying quantile process of $z_1, \dots, z_T$ given a grid cell $\mathbf{B}_s$. Suppose $z_{(1)} \dots, z_{(l-1)} < \beta_0 + \sum_m I_m(\delta) \beta_m \leq z_{(l)} \dots \leq z_{(u)} \leq \beta_0 + \sum_m I_m(1 - \delta) \beta_m < z_{(u+1)}, \dots, z_{(T)}$, where $q_Z(\tau_{z_{(l)}}) = z_{(l)}$, $q_Z(\tau_{z_{(u)}}) = z_{(u)}$. Then the probability density function of $Z \in [q_Z(\tau_{z_{(l)}}), q_Z(\tau_{z_{(u)}})]$ is obtained using $q_Z(\tau)$:

$$f_Z(q_Z(\tau_{z_{(i)}})) = f_Z(z_{(i)}) = \frac{1}{\frac{\partial q_Z(\tau)}{\partial \tau}} \Big|_{\tau=\tau_{z_{(i)}}}, \text{ if } q_Z(\tau_{z_{(l)}}) \leq z \leq q_Z(\tau_{z_{(u)}}) \quad (2.6)$$

where τ_{z_t} solves $z_t = q_Z(\tau_{z_t})$ in τ , for $t = 1, 2, \dots, T$. Let $\tau_1, \tau_2, \dots, \tau_K$ denote $K = 101$ equally spaced quantile levels in $[0, 1]$, a solution τ_{z_t} to $q_Z(\tau_{z_t}|Z) - z_t = 0$ can be efficiently obtained using Newton's Recursion method [Tokdar and Kadane, 2011]:

$$\tau_{z_t}^{(k+1)} = \tau_{z_t}^{(k)} - \frac{q_Z(\tau_{z_t}^{(k)}) - z_t}{\frac{\partial}{\partial \tau} q_Z(\tau_{z_t}^{(k)})}, \quad (2.7)$$

where $\tau_{z_t}^{(0)}$ is the initial value defined as the lower bound of the estimated quantile interval containing z , i.e., $\tau_{z_t}^{(0)} = \sum_{k=1}^K \tau_k \mathbb{I}(\tau_k \leq z < \tau_{k+1})$. However, the output data which are outside of $[q_Z(\tau_{z_{(l)}}), q_Z(\tau_{z_{(u)}})]$ cannot be obtained directly by the quantile functions; how to specify the density at *extreme* values has been studied in the recent literature. A widely used approach to characterize such tail behavior is through the generalized Pareto distribution whose density function is [Hosking and Wallis, 1987]:

$$\begin{aligned} \varphi_Z(z) &= \omega(1 - \kappa\omega z)^{1/\kappa-1}, & \kappa \neq 0 \\ &= \omega \exp(-\omega z), & \kappa = 0. \end{aligned} \quad (2.8)$$

the range of z is $0 \leq z < \infty$ for $\kappa \leq 0$ and $0 \leq z \leq 1/(\omega\kappa)$ for $\kappa > 0$. Consequently, quantiles of the generalized Pareto distribution are given by:

$$\begin{aligned} q_Z(\tau) &= (1 - (1 - \tau)^\kappa) / (\omega\kappa), & \kappa \neq 0 \\ &= -\log(1 - \tau) / \omega, & \kappa = 0. \end{aligned} \quad (2.9)$$

where the parameter ranges $\kappa > 0$, $\kappa = 0$, and $\kappa < 0$ correspond to the short tails, exponential type and long tails, respectively. Note that the exponential distribution is a special case of a generalized Pareto distribution with $\kappa = 0$ (see Figure 2.2). Therefore, in practice, the generalized Pareto distribution is used in which some robustness is required against heavier tailed or lighter tailed distributions.

Figure 2.2: The probability density function for the generalized Pareto distribution is plotted with respect to the different tail index parameters κ . The parameter ranges $\kappa > 0$, $\kappa = 0$, and $\kappa < 0$ correspond to the short tails, exponential type and long tails, respectively.

Subsequently, given the *tail index parameter* κ , the likelihood function of Z is defined by a mixture distribution with its central part (2.6) and tails (4.6), such that:

$$f_Z(z) = \begin{cases} C_L \varphi_Z \left(q_Z(\tau_{z(l)}) - z; \omega_L \right), & \text{if } z < q_Z(\tau_{z(l)}) \\ \frac{1}{\frac{\partial q_Z(\tau)}{\partial \tau} \big|_{\tau=\tau(z)}}, & \text{if } q_Z(\tau_{z(l)}) \leq z \leq q_Z(\tau_{z(u)}) \\ C_U \varphi_Z \left(z - q_Z(\tau_{z(u)}); \omega_U \right), & \text{if } z > q_Z(\tau_{z(u)}) \end{cases} \quad (2.10)$$

where $C_L = \tau_{z(l)}$, $C_U = 1 - \tau_{z(u)}$, $\omega_L = \frac{1}{\frac{\partial q_Z(\tau)}{\partial \tau} \big|_{\tau=\tau(z_{(l)})}} \frac{1}{C_L}$, and $\omega_U = \frac{1}{\frac{\partial q_Z(\tau)}{\partial \tau} \big|_{\tau=\tau(z_{(u)})}} \frac{1}{C_U}$. To evaluate the likelihood, we use Markov chain Monte Carlo to sample from and summarize the posterior distribution of the parameters. Let $\pi(\beta|Z)$ be the posterior distribution of β ; it will have a proper distribution even with an improper prior (Appendix 1).

2.2 Spatial-temporal quantile regression

If we denote time with t , $t=1,2,\dots,T$, $\mathbf{u}_t=(u_{t1}, u_{t2}, \dots, u_{tJ})'$. $u_{t1} \equiv 1$ and u_{tj} is the B-spline of t with $df=J-1$, $j=2,\dots,J$. Then $Q_Y(\tau|\mathbf{u}_t, \mathbf{s})$ denotes the τ^{th} quantiles process of the observed daily 8-hour maximum ozone concentration at $\mathbf{s}$ and time t , while $Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s)$ is the τ^{th} model output quantile level for grid cell $\mathbf{B}_s$ given time t . Again, we relate the 12 km^2 model output grid cell $\mathbf{B}_s$ to each monitoring site $\mathbf{s}$.

We start by using quantile functions that vary with $\mathbf{B}_s$, $\mathbf{u}_t$ and τ for model output; thus, they give a density regression model where the temporal trend is allowed to affect the shape of the model output distribution. This means that:

$$Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s) = \mathbf{u}_t' \beta_{0, \mathbf{B}_s} + \beta_{\mathbf{B}_s}(\tau) = \sum_{j=1}^J u_{tj} \beta_{0j}(\mathbf{B}_s) + \sum_{m=1}^M I_m(\tau) \beta_m(\mathbf{B}_s), \quad (2.11)$$

To specify monotonic constraints for $Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s)$ with the temporal component $\mathbf{u}_t$, the non-negativity of $\beta_{\mathbf{B}_s}(\tau)$ is required. More specifically, we introduce latent unconstrained variables $\beta_m^*(\mathbf{B}_s)$ and take constraints as (2.4) in section 2.1.1.

2.3 Simulation study

For nonparametric quantile regression, the proposed Bayesian spatial quantile method (BSQ) is compared with the frequentist quantile regression method. In each example, $L=100$ data sets are simulated. The data are given by:

$$z_{(t_i, \mathbf{s}_i)} = f(t_i, \mathbf{s}_i) + g(t_i, \mathbf{s}_i) \epsilon_i \quad (2.12)$$

for the mean function f and variance function g . The time t_i has a $U(0,1)$ distribution, and $\epsilon_i \sim N(0, 1)$ with sample size $n = 100$. Examples are given as:

- Example: Temporal quantile: $f(t_i, \mathbf{s}_i) = 0.5 + 2t_i + \sin(2\pi t_i - 0.5)$, and $g(t_i, \mathbf{s}_i)=1$.

Model performance is evaluated in terms of the empirical root mean integrated squared error $RMISE = [T^{-1} \sum_{t=1}^T (\hat{q}_z(\tau|u_t) - q_z(\tau|u_t))^2]^{1/2}$ for $\tau = 0.01, 0.5, 0.9$ and 0.99 . $\hat{q}_z(\tau|u_t)$ is the estimated function and $q_z(\tau|u_t)$ is the real function. The interior knots at $(0.2, 0.8)$ provide a smaller $RMISE$ at the tails. In Figure 2.3, we plot a time series of the simulated data and its underlying $100\tau^{th}$ curve as the real process. The quantile spline regression captures most of the variations in the data but demonstrates a crossing problem. Our BSQ approach yields an overall trend with a wide band.

Table 2.1: (Example 2) Empirical root mean integrated squared error ($\times 100$), with its standard error in parentheses

Method	Knots	δ	$\tau = 0.01$	$\tau = 0.50$	$\tau = 0.90$	$\tau = 0.99$
BSQ	(0.2, 0.8)	0.05	7.61(0.034)	5.11(0.014)	5.44(0.014)	7.07(0.030)
BSQ	(0.3, 0.7)	0.005	7.84(0.029)	5.57(0.015)	6.77(0.018)	7.95(0.031)
Frequentist	-	-	15.80(0.038)	6.14(0.017)	8.51(0.030)	15.50(0.042)

2.4 Summary

In this chapter, we introduced a Bayesian framework for a simultaneous analysis of quantile functions and quantile regression models. Because the frequentist method analyzes quantiles separately, it is hard to provide a valid quantile function when taking the temporal effect into account. Thus, constructing the quantile functions through a continuous quantile process is essential to attain the true quantile framework, but there is a lack of computational efficiency due to the associated monotonic constraints. Also, to obtain the non-fully specified likelihood, we present a Bayesian density model based on the quantile functions. Tokdar and Kadane [2011] investigated the quantile regression setting with a bounded response variable, and we

extend it to deal with the unbounded variables. The resulting formulation leads to a mixture distribution defined by both the $1 - \tau_{(z_{(l)})} - \tau_{(z_{(u)})}$ percentile of the data and a generalized Pareto distribution for the tails. The simulation analysis presented in section 2.3 indicates the advantages of a simultaneous quantile function analysis over individual fits.

Figure 2.3: Bayesian nonparametric quantile (BSQ) regression from Example 2. Interior knots are placed at 0.2, 0.8 with $\delta = 0.005$ and $\kappa = 0$. We add a sin function to mimic the temporal trend in reality. We obtain the conditional quantile estimates for $\tau = 0.01, 0.5, 0.9$ and 0.99 . Notice that the classic quantile regression spline generates crossing quantile curves, which does not provide a valid quantile process.

Chapter 3

Spatial quantile calibration

As we discussed in Chapter 1, when linear regression is used for a data transformation problem, there might be a strong desire to move beyond the most elementary type of transformation. In this chapter, we consider only the model grid cells with a monitoring location, which allows us to identify a calibration function G_τ to map the climate model quantile process to the quantile functions estimated by the observed monitoring data. Subsequently, a spatial adjustment of the entire distribution of the model output with respect to the distribution of the monitoring data is specified to downscale the calibration function and their spatial-quantile processes (See Figure 4.1) . Finally, we incorporate a smoothed temporal trend in the calibration model to handle the spatio-temporal calibration refer to the entire distribution.

3.1 Spatial quantile calibration of deterministic model output

For the purpose of calibrating the spatial-quantile process, we make use of a monotonically increasing map $\eta_s(\tau, \tilde{\beta}_{B_s})$ as:

$$\eta_s(\tau, \hat{\tilde{\beta}}_{B_s}) = Q_Z(\tau | B_s; \hat{\tilde{\beta}}_{B_s}) \quad (3.1)$$

where $\hat{\tilde{\beta}}_{B_s}$ are sampled from the posterior distribution $\pi(\tilde{\beta}_{B_s}|Z)$. Thus we have the observed quantiles of Y as follows:

$$Q_Y(\tau|\mathbf{s}) = \mathbf{I} \left(\eta_{\mathbf{s}}(\tau, \hat{\tilde{\beta}}_{B_s}) \right)' \tilde{\boldsymbol{\alpha}}(\mathbf{s}) = \alpha_0(\mathbf{s}) + \sum_{m=1}^M I_m \left(\eta_{\mathbf{s}}(\tau, \hat{\tilde{\beta}}_{B_s}) \right) \alpha_m(\mathbf{s}) \quad (3.2)$$

where $\tilde{\boldsymbol{\alpha}}(\mathbf{s})$ are spatially-varying coefficients. Similar to equation (2.4), we introduce a latent unconstrained variable $\alpha_m^*(\mathbf{s})$ to ensure the quantile constraints:

$$\alpha_m(\mathbf{s}) = \begin{cases} \alpha_m^*(\mathbf{s}) & \text{if } \alpha_m^*(\mathbf{s}) \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.3)$$

$\alpha_m^*(\mathbf{s})$ are modeled as multivariate mean-zero Gaussian spatial processes with boundary conditions:

$$Q_Y(0|\mathbf{s}) = \alpha_0(\mathbf{s}), \quad Q_Y(1|\mathbf{s}) = \alpha_0(\mathbf{s}) + \sum_{m=1}^M \alpha_m(\mathbf{s}) \quad (3.4)$$

However, strict bounds on Y may not be known a priori. To ensure that the posterior has a proper distribution (see Appendix 1), we express the likelihood function as a mixture distribution similar to (4.5), i.e., for $\kappa = 0$:

$$\begin{aligned} f_Y(y) &= \left(C_L \exp^{-\omega_L(\alpha_0 - y)} \right) \mathbb{I}(y < \alpha_0) \\ &\times \left(C_U \exp^{-\omega_U(y - (\alpha_0 + \sum \alpha_m))} \right) \mathbb{I}(y > \alpha_0 + \sum \alpha_m) \\ &\times (f_Y^*(y|\mathbf{s})) \mathbb{I}(\alpha_0 \leq y \leq \alpha_0 + \sum \alpha_m) \end{aligned} \quad (3.5)$$

where ω_L and ω_U are positive rate parameters; C_L and C_U provide the probability of the left and right truncations. Additionally, $f_Y^*(y|\mathbf{s})$ is the density function derived from both the model output and observed quantile functions, and its computing algorithm is provided in Section 3.3. The resulting likelihood has an exponential decay once the estimated quantile boundaries do not

include certain observed values. Also, we assume that there exist $(M+1)$ independent Gaussian processes $\alpha_0(\mathbf{s}), \alpha_1(\mathbf{s}), \dots, \alpha_M(\mathbf{s})$ such that, $\text{Cov}(\alpha_m(\mathbf{s}), \alpha_m(\mathbf{s}')) = \sigma_{\mathbf{ms}}^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\rho_{\mathbf{ms}})$ where $\rho_{\mathbf{ms}}$ is the spatial decay parameter for the Gaussian process $\alpha_m(\mathbf{s})$, $m = 0, 1, \dots, M$.

3.2 Spatial-temporal quantile calibration

The calibration model in section 3 can be extended to accommodate data collected over space and time. When we want to calibrate the model output with respect to the monitoring data for multiple years, it is necessary to take into account the effect of the temporal trend on the distribution. If we denote time with t , $t = 1, 2, \dots, T$, $\mathbf{u}_t = (u_{t1}, u_{t2}, \dots, u_{tJ})'$. Let $u_{t1} = 1$ and u_{tj} be the B-spline of t with $\text{df} = J-1$, $j = 2, \dots, J$. Then $Q_Y(\tau|\mathbf{u}_t, \mathbf{s})$ denotes the τ^{th} quantile process of the observed daily 8-hour maximum ozone concentrations at $\mathbf{s}$ and time t , while $Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s)$ is the τ^{th} model output quantile levels for grid cell $\mathbf{B}_s$ given time t . Again, we relate the model grid cell $\mathbf{B}_s$ to each monitoring site $\mathbf{s}$.

We start by using quantile functions that vary with $\mathbf{B}_s$, $\mathbf{u}_t$ and τ for model output. They give us a density regression model where the temporal trend is allowed to affect the shape of the model output distribution. This means that:

$$Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s) = \mathbf{u}_t' \boldsymbol{\beta}_{0, \mathbf{B}_s} + \beta_{\mathbf{B}_s}(\tau) = \sum_{j=1}^J u_{tj} \beta_{0j}(\mathbf{B}_s) + \sum_{m=1}^M I_m(\tau) \beta_m(\mathbf{B}_s) \quad (3.6)$$

To specify monotonic constraints for $Q_Z(\tau|\mathbf{u}_t, \mathbf{B}_s)$ with the temporal component $\mathbf{u}_t$, the non-negativity of $\beta_{\mathbf{B}_s}(\tau)$ is required. More specifically, we introduce latent unconstrained variables $\beta_m^*(\mathbf{B}_s)$ and take constraints as (2.4) in section 2.1.1.

In order to construct the quantile functions of Y based on the model output process, we first consider the predictive model output spatial-quantile processes $\eta_{\mathbf{u}_t, \mathbf{s}}(\tau)$ as monotonically

increasing maps from $[0,1]$ onto itself given any location $\mathbf{s}$:

$$\eta_{\mathbf{u}_t, \mathbf{s}}(\tau, \hat{\beta}_{0, \mathbf{B}_s}, \hat{\beta}_{\mathbf{B}_s}) = Q_Z(\tau | \mathbf{u}_t, \mathbf{B}_s; \hat{\beta}_{0, \mathbf{B}_s}, \hat{\beta}_{\mathbf{B}_s}) \quad (3.7)$$

where $\hat{\beta}_{0, \mathbf{B}_s} \propto \pi(\beta_{0, \mathbf{B}_s} | Z)$ and $\hat{\beta}_{\mathbf{B}_s} \propto \pi(\beta_{\mathbf{B}_s} | Z)$. Then we have the quantiles of observed data Y as follows:

$$\begin{aligned} Q_Y(\tau | \mathbf{u}_t, \mathbf{s}) &= \mathbf{u}_t' \alpha_{0, \mathbf{s}} + \alpha_{\mathbf{s}}(\eta_{\mathbf{u}_t, \mathbf{s}}(\tau, \hat{\beta}_{0, \mathbf{B}_s}, \hat{\beta}_{\mathbf{B}_s})) \\ &= \sum_j^J u_{tj} \alpha_{0j}(\mathbf{s}) + \sum_m^M I_m(\eta_{\mathbf{u}_t, \mathbf{s}}(\tau, \hat{\beta}_{0, \mathbf{B}_s}, \hat{\beta}_{\mathbf{B}_s})) \alpha_m(\mathbf{s}) \end{aligned} \quad (3.8)$$

We define the monotonic spatially-variant $\alpha_m(\mathbf{s})$ as the following latent variables:

$$\alpha_m(\mathbf{s}) = \begin{cases} \alpha_m^*(\mathbf{s}) & \text{if } \alpha_m^*(\mathbf{s}) \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad m = 1, \dots, M \quad (3.9)$$

as in section 3. We assume that there exist zero-mean Gaussian processes $\alpha_m(\mathbf{s})$ such that, $\text{Cov}(\alpha_m(\mathbf{s}), \alpha_m(\mathbf{s}')) = \sigma_m^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\rho_m \mathbf{s})$ and $\rho_m \mathbf{s}$ is the spatial decay parameter for the Gaussian process $\alpha_m(\mathbf{s})$. The different temporal trends between model output and the observed quantile process are then adjusted through the calibration parameters $\alpha_0(\mathbf{s}), \alpha_1(\mathbf{s}), \dots, \alpha_m(\mathbf{s})$.

3.3 Estimation

In this section, we focus on how to obtain the likelihood for the monitoring data based only on its quantile process $Q_Y(\tau | \mathbf{s}) = \mathbf{I}(\eta_{\mathbf{s}}(\tau, \hat{\beta}_{\mathbf{B}_s}))' \tilde{\alpha}(\mathbf{s})$ and the model output predictive quantile $\eta_{\mathbf{s}}(\tau, \hat{\beta}_{\mathbf{B}_s})$. Suppose the constraints (2.4) and (3.3) are satisfied, then $\tau \rightarrow Q_Y(\tau | \mathbf{s})$ is monotonically increasing. Hence, the process (4.5) uniquely determines an unconditional sampling

density for Y in the form [Tokdar and Kadane, 2011]:

$$f_Y(y|\mathbf{s}) = \frac{1}{\frac{\partial}{\partial \tau} Q_Y(\tau|\mathbf{s})} \Big|_{\tau=\tau_{Z,\mathbf{s}}(y)} \quad (3.10)$$

where $\tau_{Z,\mathbf{s}}(y)$ is the solution to $y = Q_Y(\tau|\mathbf{s})$ in τ . We apply the truncated likelihood (3.5) to approximate the density function:

$$\begin{aligned} f_Y(y|\mathbf{s}) &= \left(C_L \exp^{-\omega_L(\alpha_0 - y)} \right) \mathbb{I}(y < \alpha_0) \\ &\times \left(C_U \exp^{-\omega_U(y - (\alpha_0 + \sum \alpha_m))} \right) \mathbb{I}(y > \alpha_0 + \sum \alpha_m) \\ &\times \left(\frac{1}{\frac{\partial}{\partial \tau} Q_Y(\tau|\mathbf{s})} \Big|_{\tau=\tau_{Z,\mathbf{s}}(y)} \right) \mathbb{I}(\alpha_0 \leq y \leq \alpha_0 + \sum \alpha_m) \end{aligned} \quad (3.11)$$

when $\alpha_0 \leq y \leq \alpha_0 + \sum \alpha_m$, the partial log-likelihood function of $f_Y^*(y|\mathbf{s})$, over the monotonicity restrictions of $(\eta_{\mathbf{s}}, \boldsymbol{\alpha}(\mathbf{s}))$ is defined as:

$$\begin{aligned} \sum_i \log f_Y^*(y_i|\mathbf{s}) &= - \sum_i \log \frac{\partial}{\partial \tau} Q_Y(\tau|\mathbf{s}) \Big|_{\tau=\tau_{Z,\mathbf{s}}(y_i)} \\ &= - \sum_i \log \frac{\partial Q_Y(\tau|\mathbf{s})}{\partial \eta_{\mathbf{s}}(\tau, \hat{\boldsymbol{\beta}}_{\mathbf{B}_{\mathbf{s}}})} \cdot \frac{\partial \eta_{\mathbf{s}}(\tau, \hat{\boldsymbol{\beta}}_{\mathbf{B}_{\mathbf{s}}})}{\partial \tau} \Big|_{\tau=\tau_{Z,\mathbf{s}}(y_i)} \end{aligned} \quad (3.12)$$

where $\tau_{Z,\mathbf{s}}(y_i)$ is the solution to $y_i = Q_Y(\tau|\mathbf{s})$, $i = 1, 2, \dots, n$. A solution $\tau_{Z,\mathbf{s}}(y)$ to $Q_Y(\tau|\mathbf{s}) - y = 0$ can be efficiently obtained using Newton's Recursion method:

$$\tau_{Z,\mathbf{s}}^{(k+1)}(y) = \tau_{\mathbf{s}}^{(k)}(y) - \frac{Q_Y(\tau_{\mathbf{s}}^{(k)}(y)|\mathbf{s}) - y}{\frac{\partial}{\partial \tau} Q_Y(\tau_{\mathbf{s}}^{(k)}(y)|\mathbf{s})}, \quad (3.13)$$

where $\tau_{Z,s}^{(0)}$ is an initial value in $[0, 1]$. The evaluations of $Q_Y(\tau|\mathbf{s})$ and $\frac{\partial}{\partial \tau} Q_Y(\tau|\mathbf{s})$ at various values of $\tau \in [0, 1]$ can be done by:

$$\begin{aligned} \frac{\partial}{\partial \tau} Q_Y(\tau|\mathbf{s}) &= \frac{\partial}{\partial \eta_{\mathbf{s}}} Q_Y(\tau|\mathbf{s}) \cdot \frac{\partial}{\partial \tau} \eta_{\mathbf{s}} \\ &= \left(\sum_{m=1}^M \frac{\partial}{\partial \eta_{\mathbf{s}}} I_m(\eta_{\mathbf{s}}(\tau_{Z,s}(y), \hat{\beta}_{\mathbf{B}_s})) \alpha_m(\mathbf{s}) \right) \cdot \left(\sum_{m=1}^M \frac{\partial}{\partial \tau} I_m(\tau_{Z,s}(y)) \hat{\beta}_m(\mathbf{s}) \right) \end{aligned}$$

To simplify the notation, let $D_1^* = \frac{3}{(\gamma_{m+2}-\gamma_{m+1})}$; $D_2^* = \frac{-3(\gamma_{m+3}-\eta)^2}{(\gamma_{m+3}-\gamma_{m+1})(\gamma_{m+3}-\gamma_m)(\gamma_{m+2}-\gamma_{m+1})}$ + $\frac{-3(\eta-\gamma_m)^2}{(\gamma_{m+3}-\gamma_m)(\gamma_{m+2}-\gamma_m)(\gamma_{m+2}-\gamma_{m+1})}$. Then the derivative of the I-spline, $\frac{\partial}{\partial \eta} I_m(\eta(\cdot))$ consists of quadratic segments as follows:

$$\frac{\partial}{\partial \eta} I_m(\eta|\gamma) = \begin{cases} 0, & \text{if } \eta < \gamma_m \\ \frac{3(\eta-\gamma_m)^2}{(\gamma_{m+1}-\gamma_m)(\gamma_{m+2}-\gamma_m)(\gamma_{m+3}-\gamma_m)}, & \text{if } \gamma_m \leq \eta < \gamma_{m+1} \\ D_1^* + D_2^*, & \text{if } \gamma_{m+1} \leq \eta < \gamma_{m+2} \\ \frac{3(\gamma_{m+3}-\eta)^2}{(\gamma_{m+3}-\gamma_{m+2})(\gamma_{m+3}-\gamma_{m+1})(\gamma_{m+3}-\gamma_m)}, & \text{if } \gamma_{m+2} \leq \eta < \gamma_{m+3} \\ 0, & \text{if } \eta \geq \gamma_{m+3} \end{cases} \quad (3.14)$$

The steps given in equations (3.12) and (3.14) provide a computationally feasible algorithm and can efficiently be ran in a PC or laptop. This algorithm depends on the number of observations, the number of grid points and the power of a computer. On a laptop computer with a 2.0 GHz Intel Core 2 Duo processor and 4 GB memory, it takes several days for calibrating the eastern U.S. ozone data using R, with $n_s = 68$, $T = 153$ and $K = 101$.

3.4 Application: calibration of eastern US ozone data

3.4.1 Data description

We use maximum daily 8-hour average ozone concentrations in parts per billion (ppb) from $n_s = 68$ sites covering the eastern U.S. from May, 1st, 2002 to September, 30th 2002 (T=153), which were obtained from the EPA Air Quality System (AQS) and can be acquired from the following website: <http://www.epa.gov/ttn/airs/airsaqs/index.htm>.

Another source of data is the 2002 base-run simulations from the Community Multiscale Air Quality (CMAQ) model. CMAQ is a multi-pollutant, multi-scale air quality model that uses state-of-the science techniques for simulating all atmospheric and land processes that affect the transport, transformation, and deposition of atmospheric pollutants and their precursors on both regional and urban scales. It is designed as a modeling tool for handling all the major pollutant issues based on a whole atmosphere approach. In this study, four annual (2002 to 2005) CMAQ model runs were completed over the eastern U.S. using a 12 km by 12 km horizontal grid. We use the ozone monitoring stations as the spatial unit and extract climate data from the grid cell containing the ozone monitoring station. Additional information and a complete technical description of the CMAQ model are given by Byun and Schere [2006].

The range of the CMAQ forecast data is quite similar to the range of the ground level ozone monitoring data. To compare the CMAQ forecasts with the observed monitoring data, we plot the sample quantile levels for the 90th percentile for our data sets over U.S. in Figure 4.4. Specifically, we extract data from a randomly selected site (the 59th site is marked on the map as *), and investigate the histogram, sample quantile and density function of both observed and CMAQ data on this site. The observed ozone data have a heavier tail than the CMAQ data. Also, modeled ozone data agree quite well with the observations at its 50th percentile, but present an overall lower 90th percentile level over our study region. This implies that there

are unknown discrepancies in the CMAQ forecasts and an appropriate calibration is needed.

3.4.2 Result analysis and discussion

To compare spatial surfaces and distributions between the observed data and the CMAQ output, we choose two data sources in the eastern U.S. We use the Metropolis-Hastings algorithm for updating $\pi(\tilde{\beta}, \rho_{m_B}, \sigma_{m_B}^2 | Z)$, $\pi(\tilde{\alpha}, \rho_{m_s}, \sigma_{m_s}^2 | Y, Z)$ individually. The likelihood is calculated by the likelihood approximation approach of $Q_Y(\tau | \mathbf{s})$ on a grid of 101 equally-spaced $\tau_k \in [0, 1]$. The I-splines have interior knots at (0.2, 0.8).

The estimated CMAQ quantiles and the calibration for the monitoring data are plotted in Figure 3.3. Both of the two spatial-quantile processes are obtained by our Bayesian algorithm. At $\tau = 0.05, 0.5$, and 0.95 , the empirical root mean integrated squared error is calculated as:

$$RMISE = [n_s^{-1} \sum_{i=1}^{n_s} (\hat{Q}_Z(\tau | \mathbf{B}_{s_i}) - \hat{Q}_Y(\tau | \mathbf{s}_i))^2]^{1/2}$$

The *RMISE* at the 50th quantile is equal to 7.13, while the value is 13.17 for the 5th percentile and 15.46 for the 95th percentile, respectively. The results show the agreement between the distributions of the CMAQ output and the monitoring data at their median level, but show large differences for the tails. Also, from the contour plot, we conclude that the CMAQ data are smoother than the observed spatial structure, indicating that the physically based numerical models cannot capture both the extreme values and spatial correlations that are in the monitoring data.

Due to these differences, it is critical to calibrate the CMAQ data considering its spatial-quantile structure. Based on the estimated CMAQ-monitoring calibration model, a nonlinear transformation is made to the CMAQ data using $\mathbf{G}_{\tau, \mathbf{s}}(z(t, \mathbf{B}_s), \hat{\alpha}) = \hat{\alpha}_0 + \sum_m^M I_m(z(t, \mathbf{B}_s)) \hat{\alpha}_m$,

where $\hat{\hat{\alpha}}$ are the posterior estimations. Then we rescale $\mathbf{G}_{\tau, \mathbf{s}}(z(t, \mathbf{B}_{\mathbf{s}}), \hat{\hat{\alpha}})$ to its original range (see Figure 3.4). We calculate $\hat{Q}_Y(\tau|\mathbf{s}_i)$ (the estimated quantiles of the monitoring data), $\tilde{Q}_{S_Z}(\tau|\mathbf{s}_i)$ (the quantiles of the Bayesian calibrated data), $\tilde{Q}_{L_Z}(\tau|\mathbf{s}_i)$ (the quantiles from the linear regression model) and $\tilde{Q}_{L_S}(\tau|\mathbf{s}_i)$ (the quantiles from the scale model), at $\tau \in [0.01, 0.97]$ and location $\mathbf{s}_i$. The scale predictor is defined as:

$$y(t, \hat{\mathbf{s}}_i) = \overline{y(\cdot, \mathbf{s}_i)} + \hat{\alpha} * (z(t, \mathbf{B}_{\mathbf{s}_i}) - \overline{z(\cdot, \mathbf{B}_{\mathbf{s}_i})})$$

where the scalar $\hat{\alpha} = \sqrt{\text{Var}(y(t, \mathbf{s}_i) - \overline{y(\cdot, \mathbf{s}_i)}) / \text{Var}(z(t, \mathbf{B}_{\mathbf{s}_i}) - \overline{z(\cdot, \mathbf{B}_{\mathbf{s}_i})})}$.

The root mean squared error $RMSE(\hat{Q}, \tilde{q}|\mathbf{s}_i) = [K^{-1} \sum_{k=1}^K (\hat{Q}(\tau_k, \mathbf{s}_i) - \tilde{q}(\tau_k, \mathbf{s}_i))^2]^{1/2}$ is calculated for both the linear regression method and our Bayesian approach at each location $\mathbf{s}_i$. Figure 3.6 shows maps of the above quantiles when $\tau = 0.95$, and we use the difference root mean squared error:

$$DRMSE = n_s^{-1} \sum_{i=1}^{n_s} \frac{RMSE(\hat{Q}_Y(\tau|\mathbf{s}_i), \tilde{Q}_{S_Z}(\tau|\mathbf{s}_i)) - RMSE(\hat{Q}_Y(\tau|\mathbf{s}_i), \tilde{Q}_{L_Z}(\tau|\mathbf{s}_i))}{RMSE(\hat{Q}_Y(\tau|\mathbf{s}_i), \tilde{Q}_{L_Z}(\tau|\mathbf{s}_i))}$$

The DRMSE between the linear regression method and the quantile calibration method range from -77% to 66%, and is -50.23% on average. The results show that 57 out of 68 (83.8%) sites have a reduced RMSE using the Bayesian calibration method (i.e., Figure 3.4). The mean DRMSE between the scale method and the quantile calibration method is -20.6%. The p-value of a Two-sample Kolmogorov-Smirnov test between the observed data and the scale calibration is $2.45e^{-11}$ (Figure 3.5), while the p-value is equal to 0.454 for the Bayesian calibration. Thus we do not reject the null hypothesis that the quantile calibrated series has the same distribution as the observed data. We also obtain the Kolmogorov-Smirnov tests for Bayesian calibration Vs. LM calibration (p-value: 0.0037) and Bayesian calibration Vs. Scale calibration (p-value: $8.29e-08$).

In summary, the spatial-quantile calibration provides a platform to transform the different parts of the model output distribution to the corresponding parts of the observed distribution at different rates. As we expected, the performance of the calibrated CMAQ model data is consistent with the performance of the monitoring data in terms of the quantile level τ .

Figure 3.1: A process chart for spatial quantile calibration for going from model output to the observations. We calibrate the original model output with the corresponding observations through their underlying spatial-quantile processes. Both of the model output quantile modeling and the adjustment of the entire distribution of model output with respect to the distribution of monitoring data need to be across space.

Figure 3.2: Maps of the sample 90th quantile levels of the ozone concentrations for both the CMAQ output and observations; At a randomly selected site (i.e., 59th) where we have T=153 observations; we use " *" to represent its location. The histogram, sample quantile and density function of the two data sources are provided to identify the different behavior in their distribution.

Figure 3.3: Quantile comparison plots. Maps of the 5th, 50th and 95th quantiles for the Bayesian estimated CMAQ model output and monitoring data are plotted. The results show the agreement between the distributions of the CMAQ output and the monitoring data at their median level, but large differences appear for the tails.

Figure 3.4: The Bayesian spatial-quantile calibration for CMAQ at a selected monitoring site. The left panel shows the corresponding $100\tau^{th}$ quantile curves of the monitoring data and G_τ -transformed model output, units in parts per billion (ppb); while the right panel is the corresponding quantile-quantile plot.

Figure 3.5: The scale predictor for CMAQ at a selected monitoring site. The left panel shows the corresponding $100\tau^{th}$ quantile curves of the monitoring data and scale-transformed model output, units in parts per billion (ppb); while the right panel is the corresponding quantile-quantile plot.

Figure 3.6: The 95th quantile for the monitoring data (units in ppb), using both the quantile calibration and linear regression method. We compare the differences between the linear regression and the Bayesian quantile calibration methods in terms of the *RMSE*.

Chapter 4

Mortality prediction with calibrated meteorological variables

In this Chapter, we evaluated the future heat wave excess mortality based on the calibrated NARCCAP data from 2041 to 2050 (see Figure 4.1). Our ability to quantify the health impacts of future climate has significant implications in guiding policies towards environmental sustainability and in protecting public health. Uncertainty quantification for climate projections is an important component in the risk assessment process. To the authors's knowledge, this is the first health impact analyses to utilize a calibration approach that directly characterizes the uncertainty in the quantiles of future weather. This is particularly useful for examining heat-related impacts where extreme events such as heat waves have been associated with considerable health risks. The main contribution of this paper is to present a statistical framework aimed to not only provide more accurate projections, but to also allow uncertainty propagation in the health impact calculations. We believe the approach described here represents a crucial step towards enhancing the applicability and relevance of the results from climate change and health studies.

In Section 4.1, we review a Bayesian approach to model the spatial-quantile processes of

the climate model output. In Section 4.2, we outline the modeling steps for risk estimation and how to utilize calibrated model projections to conduct health impact analysis associated with high temperature days and heat waves. We provide the estimation algorithm in section 4.3. In Section 4.4.1, we present the climate model projections and health data for a case study in the US state of Alabama. The southeast region of the US was selected because it has been shown to experience the greatest increase in 20-year return values of daily maximum temperature in the U.S. [Kharin and Zwiers, 2000]. In section 4.4 we present results of the calibration and estimates of the number of attributable deaths due to future temperature extremes. We also compared our approach to calibration using linear regression methods in a cross-validation study. Finally, we end with some conclusions and remarks in Section 4.5.

4.1 System calibration and spatial quantile processes

In this section, we will review the spatial-quantile calibration method for obtaining the transformation (calibration) function G_τ in (1.11) across space. Suppose we have two data sets $Y(t, \mathbf{s}_i)$ and $Z(t, \mathbf{B}_{\mathbf{s}_i})$, where $\mathbf{s}_i = (s_{i1}, s_{i2})$ are latitude/longitude coordinates of a temperature monitoring location and $\mathbf{B}_{\mathbf{s}_i}$ is the associated 50×50 km grid cell in which $\mathbf{s}_i$ lies. At each location $i=1,2,\dots,n_s$, we have time $t = 1, 2, \dots, T$ replications. In order to calibrate the underlying spatial-quantile processes $Q_Z(\tau|\mathbf{B}_s)$ and $Q_Y(\tau|\mathbf{s})$, we extend the transformation function G_τ to be space-dependent, and we denote it as $\mathbf{G}_{\tau,s}$. Then we have:

$$Q_Y(\tau|\mathbf{s}) = \mathbf{G}_{\tau,s}(Q_Z(\tau|\mathbf{B}_s)) \quad (4.1)$$

Without loss of generality, we rescale both the NARCCAP data (in year 2000 and 2041-2050) and observations (year 2000) such that the range of Z and Y is within $[0, 1]$.

We first model the spatially-varying quantile function of the NARCCAP model outputs $Q_Z(\tau|\mathbf{B}_s)$ for $\tau \in [0, 0.01, \dots, 1]$. The monotonic properties for the quantile functions are

achieved through the use of the *integrated piecewise polynomial spline* (or I-splines) [Lu and Clarkson, 1999; Ramsay, 1988]. The I-splines construct basis functions using piecewise polynomials, such that upper and lower tails are estimated independently from each other. For a knot sequence $\{\gamma_1, \dots, \gamma_{M+h}\}$, M is the number of free parameters that determine the spline function's continuity characteristics, and h is the degree of the piecewise polynomial I_m (For example, see the piecewise cubic I-spline as an example in Zhou et al. [2011]). Subsequently, the process $Q_Z(\tau|\mathbf{B}_s)$ is defined as:

$$Q_Z(\tau|\mathbf{B}_s) = \mathbf{I}(\tau)' \tilde{\boldsymbol{\beta}}_{\mathbf{B}_s} = \beta_0(\mathbf{B}_s) + \sum_{m=1}^M I_m(\tau) \beta_m(\mathbf{B}_s) \quad (4.2)$$

where $\mathbf{I}(\tau)'$ denotes the corresponding I-spline basis functions. Because the I-spline is an integral of nonnegative splines, it yields monotone functions $\sum_{m=1}^M I_m(\tau) \beta_m(\mathbf{B}_s)$ when combined with nonnegative values of the coefficients $\beta_m(\mathbf{B}_s)$. To incorporate spatial dependence, we assign the following prior distribution for the spline coefficients: $\beta_m(\mathbf{B}_s) \sim N(\mu_m, \Sigma^m)$, with $\Sigma_{(\mathbf{B}_s, \mathbf{B}_{s'})}^m = \sigma_{mB}^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\rho_{mB})$.

We then calibrate the NARCCAP spatial-quantile process with the observations in (4.1) by modeling the observed quantiles of Y as follows:

$$Q_Y(\tau|\mathbf{s}) = \mathbf{I}(\eta_s(\tau, \hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s}))' \tilde{\boldsymbol{\alpha}}(\mathbf{s}) = \alpha_0(\mathbf{s}) + \sum_{m=1}^M I_m(\eta_s(\tau, \hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s})) \alpha_m(\mathbf{s}) \quad (4.3)$$

where $\eta_s(\tau, \hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s})$ is a monotonic function defined as:

$$\eta_s(\tau, \hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s}) = Q_Z(\tau|\mathbf{B}_s; \hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s}) \quad (4.4)$$

Therefore, we first model quantile functions of the model output $\eta_s(\tau, \hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s})$ as a function of τ , then we model the quantile values $Q_Y(\tau|\mathbf{s})$ to obtain the calibration function $\mathbf{G}_{\tau, \mathbf{s}}$ between the two quantile processes (see Formula (4.1)). $\hat{\tilde{\boldsymbol{\beta}}}_{\mathbf{B}_s}$ are sampled from the posterior distribution

$\pi(\beta_m(\mathbf{B}_s)|Z)$. Note that we mainly focus on $\eta_s(\tau, \tilde{\beta}_{\mathbf{B}_s}) \in [0, 1]$ because Z is standardized to be between 0 and 1. Therefore, for non-negative $\alpha_m(\mathbf{s})$ values we obtain a monotonic mapping between the model output and observed quantile functions. Again, we assume $\alpha_m(\mathbf{s})$ are spatially-varying coefficients modeled as multivariate mean-zero Gaussian spatial processes $\alpha_0(\mathbf{s}), \alpha_1(\mathbf{s}), \dots, \alpha_M(\mathbf{s})$, such that, $\text{Cov}(\alpha_m(\mathbf{s}), \alpha_m(\mathbf{s}')) = \sigma_{ms}^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\rho_{ms})$ and ρ_{ms} is the spatial decay parameter for the Gaussian process $\alpha_m(\mathbf{s})$, $m = 0, 1, \dots, M$.

We extend the work of Tokdar and Kadane [2011] for a bounded response variable to deal with an unbounded variable. Given a quantile function $q_Z(\tau)$ we define its density as

$$f_Z(z_{\mathbf{B}_s}) = \begin{cases} C_{\mathbf{B}_s L} \cdot \varphi_Z(Q_Z(\tau_{z(l, \mathbf{B}_s)}|\mathbf{B}_s) - z_{\mathbf{B}_s}; \omega_{\mathbf{B}_s L}) & \text{if } z_{\mathbf{B}_s} < Q_Z(\tau_{z(l, \mathbf{B}_s)}|\mathbf{B}_s) \\ \frac{1}{\frac{\partial Q_Z(\tau|\mathbf{B}_s)}{\partial \tau}} \Big|_{\tau=\tau_{z_{\mathbf{B}_s}}} & \text{if } Q_Z(\tau_{z(l, \mathbf{B}_s)}|\mathbf{B}_s) \leq z_{\mathbf{B}_s} \leq Q_Z(\tau_{z(u, \mathbf{B}_s)}|\mathbf{B}_s) \\ C_{\mathbf{B}_s U} \cdot \varphi_Z(z_{\mathbf{B}_s} - Q_Z(\tau_{z(u, \mathbf{B}_s)}|\mathbf{B}_s); \omega_{\mathbf{B}_s U}) & \text{if } z_{\mathbf{B}_s} > Q_Z(\tau_{z(u, \mathbf{B}_s)}|\mathbf{B}_s). \end{cases} \quad (4.5)$$

where $z_{\mathbf{B}_s}$ is the model output at the grid cell $\mathbf{B}_s$; $z_{(l, \mathbf{B}_s)}$ and $z_{(u, \mathbf{B}_s)}$ are lower and upper cut points that satisfy $Q_Z(\tau_{z(l, \mathbf{B}_s)}|\mathbf{B}_s) = z_{(l, \mathbf{B}_s)}$ and $Q_Z(\tau_{z(u, \mathbf{B}_s)}|\mathbf{B}_s) = z_{(u, \mathbf{B}_s)}$, respectively. The Newton's Recursion method [Tokdar and Kadane, 2011] is used to approximate τ_z . Therefore, equation (4.5) allows us to evaluate the likelihood given a threshold δ . For example, if $\tau_{z(l)} = \delta$, $\tau_{z(u)} = 1 - \delta$ and $\delta = 0.1$, then we evaluate the central 80% density using the equation (4.2), and the 10% upper/lower tail's density using the generalized Pareto distribution (4.6).

The sample quantiles are bounded by the sample minimum and maximum. In order to model the tail behaviors of the distribution better, we describe a Bayesian mixture approach to recover the density function. It is based on (1) the quantile functions estimated from the data for characterizing the central tendency; and (2) a parametric density function φ_Z for characterizing the tails. The excess over a threshold $Q_Z(\tau_{z(u, \mathbf{B}_s)}|\mathbf{B}_s)$, for instance, is described by φ_Z . In our application, we treat φ_Z as the density function of the generalized Pareto distribution (GPD)

[Hosking and Wallis, 1987]:

$$\varphi_Z(z_{\mathbf{B}_s}^*) = \begin{cases} \omega_{\mathbf{B}_s} (1 - \kappa_{\mathbf{B}_s} \omega_{\mathbf{B}_s} z_{\mathbf{B}_s}^*)^{1/\kappa_{\mathbf{B}_s}-1} & \text{if } \kappa_{\mathbf{B}_s} \neq 0 \\ \omega_{\mathbf{B}_s} \exp(-\omega_{\mathbf{B}_s} z_{\mathbf{B}_s}^*) & \text{if } \kappa_{\mathbf{B}_s} = 0. \end{cases} \quad (4.6)$$

To ensure a continuous density function, the constants in equation (4.5) are given by:

$$\begin{aligned} C_{\mathbf{B}_s L} &= \tau_{z(l, \mathbf{B}_s)} & , & & \omega_{\mathbf{B}_s L} &= \frac{1}{C_{\mathbf{B}_s L}} \left[\frac{\partial Q_Z(\tau | \mathbf{B}_s)}{\partial \tau} \Big|_{\tau=\tau_{z(l, \mathbf{B}_s)}} \right]^{-1} \\ C_{\mathbf{B}_s U} &= 1 - \tau_{z(u, \mathbf{B}_s)} & , & & \omega_{\mathbf{B}_s U} &= \frac{1}{C_{\mathbf{B}_s U}} \left[\frac{\partial Q_Z(\tau | \mathbf{B}_s)}{\partial \tau} \Big|_{\tau=\tau_{z(u, \mathbf{B}_s)}} \right]^{-1} \end{aligned}$$

Because $C_{\mathbf{B}_s L} \times \omega_{\mathbf{B}_s L}$ and $C_{\mathbf{B}_s U} \times \omega_{\mathbf{B}_s U}$ are the density evaluated at the point where the generalized Pareto distribution joins the quantile derived distribution (See Figure 4.2). As a result, $\omega_{\mathbf{B}_s L}$ and $\omega_{\mathbf{B}_s U}$ ensure that equation (4.5) is continuous. Finally, technical details of estimating the parameter $\kappa_{\mathbf{B}_s}$ are provided in the Appendix.

4.2 Health Effect Estimation and Impact Analysis

Relative change in the rate of mortality associated with variation in daily maximum temperature was estimated via Poisson regression

$$\log E(y_t^c) = ns(\text{temp}_t^c, 3) + \text{confounders},$$

where y_t^c is the number of deaths on day t in city c . We modeled the effects of temperature as a non-linear function using natural cubic splines with 3 degrees of freedom denoted by $ns(\text{temp}_t^c, 3)$. We have conducted sensitivity analyses with respect to the different degrees of freedom, and have found that the relative risk estimates and the expected number of excess deaths attributable to the heat waves are robust to the type of smooth functions.

Following Anderson and Bell [2011], we also included the following confounders in the model: (1) indicators for day of the week; (2) indicators for city; (3) indicators of age group (under 65, 65 to 75, above 75); (4) city and age-group interactions; (5) smooth function of calendar date $ns(t, 10 \times 3)$; and (6) smooth functions of dew-point temperature for the current day and previous three-day averages using natural cubic splines with 2 degrees of freedom. In our analysis, we did not find significant interactions between the effects of temperature and age-groups. We also assumed the effects of weather are identical across the three cities.

Because the effect of temperature is non-linear, we used the metric by Peng et al. [2011] to summarize the overall relative risk associated with high temperature and heat waves. The relative risks (RR) were calculated separately for each city as follows

$$\hat{RR}^c = \frac{\frac{1}{n_{\text{risk}}^c} \sum_t \exp\{\hat{f}(\text{temp}_t^c)\} \times I_{\text{risk},t}^c}{\frac{1}{n_{\text{ref}}^c} \sum_t \exp\{\hat{f}(\text{temp}_t^c)\} \times I_{\text{ref},t}^c} \quad (4.7)$$

where $I_{\text{risk},t}^c$ and $I_{\text{ref},t}^c$ are indicators for whether the daily maximum temperature on day t represents an *at-risk* day or a *reference* day. Similarly n_{risk}^c and n_{ref}^c denote the total number of at-risk and reference days in county c . Therefore the above metric describes the ratio of average attributable mortality between the at-risk and reference days.

To estimate the overall effect of high temperature at different thresholds, we set the reference temperature to be less than 30°C, the median value between 1991 to 2000. We then calculated the RR associated with daily maximum temperatures exceeding values 32.5°C, 34°C, 35.5°C, or 37°C (approximately the 90th, 95th, 97.5th and 99th percentile, respectively). There exists no universal definition of heat waves and various metrics have been proposed [Huth et al., 2000; Karl and Knight, 1997]. We used a similar definition from a recent national population study to estimate the additional adverse effect of heat waves by creating indicators for heat wave days [Anderson and Bell, 2011]. Specifically, we identified heat waves as a period of ≥ 2 consecu-

tive days with daily maximum temperature higher than a city-specific quantile thresholds. We examined heat waves defined based on the 90th, 95th, 97.5th, and 99th quantiles. For the heat wave RR, reference days were defined as all non-heat wave days.

Finally, we quantified current and future health impacts of temperature by calculating the expected number of excess deaths attributable to high temperature days as

$$\sum_c N_{\text{ref}}^c \times (RR^c - 1) \times M_{\text{risk}}^c$$

where N_{ref}^c is the expected daily mortality in city c on a reference day, and M_{risk}^c is the number of at-risk days over the 10-year period of 1991 to 2000, or 2041 to 2050. We estimated N_c by the mean daily mortality across all days under 30°C for each city.

4.3 Estimation

Estimation was carried out in a Bayesian framework via Markov chain Monte Carlo (MCMC) in two stages (see Figure 4.1). In the first stage, we estimated the calibration parameters using the Metropolis-Hastings algorithm with data in year 2000. To sample from $\pi(\tilde{\beta}, \rho_{m_B}, \sigma_{m_B}^2 | Z)$ and $\pi(\tilde{\alpha}, \rho_{ms}, \sigma_{ms}^2 | Y, Z)$ sequentially, the associated density function was approximated from the quantile function $Q_Y(\tau | Z, \mathbf{s})$ using a grid of 101 equally-spaced quantile values $\tau_k \in [0, 1]$ and I-splines with interior knots at (0.3, 0.7). We then obtained posterior predictive samples of the calibrated future NARCCAP data based on the nonlinear transformation function $\mathbf{G}_{\tau, \mathbf{s}}$,

$$\mathbf{G}_{\tau, \mathbf{s}}(z(t, \mathbf{B}_s), \hat{\alpha}) = \hat{\alpha}_0 + \sum_m^M I_m(z(t, \mathbf{B}_s)) \hat{\alpha}_m, \quad (4.8)$$

where $\hat{\alpha}$ are posterior samples from the MCMC. Finally, $\mathbf{G}_{\tau, \mathbf{s}}(z(t, \mathbf{B}_s), \hat{\alpha})$ was re-scaled to its original range (see Figure 4.3).

In the second stage, we fitted the health model in a separate MCMC run. Posterior distributions for the relative risks were obtained from the posterior samples of the temperature-mortality dose-response function $ns(\text{temp}_t^c, 3)$. We combined the uncertainty in both climate projection and risk estimation in calculating the number of excess deaths (ED) as follows. For the future period, the number of at-risk days M_{risk}^c was calculated for each posterior time series of the calibrated NARCCAP daily maximum temperature. We then took an *exposure simulation* approach Gryparis et al. [2009] where for each posterior sample of the relative risk, a realization of M_{risk}^c was randomly drawn from its posterior samples. We then pooled the posterior samples of ED and calculated its posterior median and 95% credible intervals.

By modeling the calibration and health models separately, we broke the feedback between the mortality data and the estimated exposures. This *directional* Bayesian approach [Gelman, 2004] not only reduces computational burden, but it also avoids the potential unintuitive assumption that health data could provide information on future exposure, for example in a causal pathway. Another benefit of fitting the exposure and health model separately is that often different metrics of exposure or sets of confounders are examined in a sensitivity analysis. For example, in our analysis different definitions of heat wave and extreme temperature were defined, and fitting a joint model repeatedly is computationally expensive.

4.4 Case study of the heat wave

4.4.1 Data description

Future climate projections and health impact analyses were conducted in the state of Alabama, USA for the period 2041 to 2050. We restricted the analysis to the months of May to September, a total of 153 days. Future daily maximum temperatures were obtained from the North

American Regional Climate Change Assessment Program (NARCCAP). NARCCAP is an international program to assess uncertainties in regional climate projections using different combinations of regional climate models and general circulation models. We utilized results from the Canadian Regional Climate Model (CRCM) [Caya et al., 1995] using boundary conditions from the third version of the Coupled Global Climate Model (CGCM3) [Scinocca et al., 2008]. NARCCAP provides gridded output with a 50×50km spatial resolution generated under the A2 emissions scenario of the Intergovernmental Panel on Climate Change. The A2 scenario projects large population increases, high carbon dioxide emissions, weak environmental concerns, and regionally oriented economic growth with slower and more heterogeneous technological changes.

We estimated the short-term effects of high temperatures on daily mortality for three urban communities in Alabama (Birmingham, Mobile, and Huntsville). Time series of daily maximum temperature, dew-point temperature, and total non-accidental deaths aggregated across the county were obtained from the National Mortality, Morbidity, and Air Pollution Study (NMMAPS) [Peng and Wealty, 2004] for the period 1991 to 2000. To perform output calibration, we also obtained NARCCAP data for the year 2000. Observed daily maximum temperature for 13 monitors in Alabama were obtained from the National Oceanic Atmospheric Administration’s National Climatic Data Center.

4.4.2 Result analysis

We first compared the uncalibrated NARCCAP output and the observations in terms of their entire distribution for year 2000. At $\tau = 0.05, 0.5$, and 0.95 , we calculated the empirical root mean integrated squared error as:

$$RMISE = \left[n_s^{-1} \sum_{i=1}^{n_s} (\hat{Q}_Z(\tau | \mathbf{B}_{s_i}) - \hat{Q}_Y(\tau | \mathbf{s}_i))^2 \right]^{1/2}$$

For the uncalibrated model output, the *RMISE* at the 5th quantile was equal to 3.93 (1.35 for the 50th percentile, and 1.72 for the 95th percentile). After calibration, the corresponding *RMISE* at $\tau = 0.05$, 0.5, and 0.95 were 0.66, 0.37, and 0.32, respectively..

The improvement in RMISE can be visually depicted in Figure 4.5. We plot the estimated quantile functions for the observed temperature at the three cities, as well as the quantile functions estimated with the uncalibrated NARCCAP model outputs and the calibrated outputs using our Bayesian algorithm. We also considered a naive calibration approach using linear regression (LM), which demonstrates the large discrepancies between the distributions of the NARCCAP output and the observed data at their lower tails. Moreover, Figure 4.5 highlights the ineffectiveness of the linear model to characterize the tail behaviors. Across monitors, we also found that the NARCCAP data present a different spatial pattern from the observed spatial structure, indicating that the process-based numerical models are biased toward “lower tails” and may not capture the spatial correlations existed in the monitoring data.

Because our objective is to perform calibration of future climate variables, we conducted a 5-fold cross-validation study to examine the performance of out-of-sample predictions. At each site, we first randomly split the original daily maximum temperature data ($T = 153$) into 5 subsamples. Then we retained a single subsample as the validation set for testing our spatial-quantile calibration, and the remaining 4 subsamples were used as training data. This process was repeated 5 times, with each of the 5 subsamples used once as the validation data. We calculated $\hat{Q}_Y(\tau|\mathbf{s}_i)$ (the estimated quantiles of the observations), $\tilde{Q}_{S_Z}(\tau|\mathbf{s}_i)$ (the averaged 5-fold quantiles of the Bayesian calibrated data), and $\tilde{Q}_{L_Z}(\tau|\mathbf{s}_i)$ (the quantiles from the linear regression model), at $\tau \in [0.01, 0.97]$ and location $\mathbf{s}_i$. The root mean squared error $RMSE(\hat{Q}, \tilde{q}|\mathbf{s}_i) = [K^{-1} \sum_{k=1}^K (\hat{Q}(\tau_k, \mathbf{s}_i) - \tilde{q}(\tau_k, \mathbf{s}_i))^2]^{1/2}$ is calculated for both the linear regression method and our Bayesian approach at each location $\mathbf{s}_i$. To compare the predictive

performance of different methods, we used the difference root mean squared error defined as

$$DRMSE = n_s^{-1} \sum_{i=1}^{n_s} \frac{RMSE(\hat{Q}_Y(\tau|\mathbf{s}_i), \tilde{Q}_{S_Z}(\tau|\mathbf{s}_i)) - RMSE(\hat{Q}_Y(\tau|\mathbf{s}_i), \tilde{Q}_{L_Z}(\tau|\mathbf{s}_i))}{RMSE(\hat{Q}_Y(\tau|\mathbf{s}_i), \tilde{Q}_{L_Z}(\tau|\mathbf{s}_i))}.$$

The DRMSE between the linear regression method and the quantile calibration method range from -81.6% to -55.3% across monitor. The average DRMSE over monitors was -68.7%, indicating a 68.7% decrease in RMSE when quantile regression was used compared to linear regression. In addition, the quantile regression achieves a 74.9% reduction in RMSE compared to the uncalibrated NARCCAP data. As a result, successfully calibrating the entire distribution of model outputs would be necessary for the model-based projections in the future.

Future projections of model outputs during the period 2041 to 2050 were calibrated similarly as in year 2000. We first scaled the NARCCAP data between 0 and 1, which allows an implementation of the transformation function $\mathbf{G}_{\tau, \mathbf{s}}$ in (4.8), as well as the associated posterior samples to be developed in the future. Figure 4.6 provides the integrated quantile curves for the uncalibrated and calibrated NARCCAP data over the decade 2041 to 2050. For comparison purposes, we added the historical values to the same graph at the three targeted cities. In the US state of Alabama, raw climate model outputs suggest that future warming at the 95th percentile was equal to 5.47°C (3.04°C for the 50th percentile and a decreasing 1.34°C for the 5th quantile). However, after calibration, the increased maximum temperature at $\tau = 0.05, 0.5$, and 0.95 were 1.73°C, 2.90°C and 4.27°C, respectively. Note that the largest discrepancies mostly occurred at the lower tail of the distribution.

Table 4.1 gives the estimated number of excess deaths due to high temperature and heat waves for the historical period (1991-2000) and the future period (2041 - 2050). The estimates are presented as annual averages across the three study cities in Alabama. The health impact of future temperatures were based on calibrated climate model output where the 95% credible intervals reflect both uncertainty in the temperature exposure-response relationship, and the

uncertainty in model output. The average number of deaths on days with maximum temperature under 30 °C was approximately 29.8 per day (16.2 for Birmingham, 8.6 for Mobile, and 4.9 for Huntsville). Similarly average mortality rates were observed on non-heat wave days defined with the 95% quantile threshold. Assuming similar future baseline mortality rates, we found a considerable increase mortality attributed to high temperature. This is dominated by increases in the number of days with extreme temperature. For example, Table 4.2 gives the observed and projected average number of high temperature days and heat wave days per year in Birmingham Alabama. Similar patterns were observed in the other two cities.

4.5 Summary

Uncertainty in climate projections can arise from various sources. We described a statistical calibration approach that models the distributional discrepancy between model outputs and historical observations. This allows us to calibrate future projections, as well as propagating its uncertainty in a health impact analysis. This differs from previous studies where future exposures from climate model outputs are assumed to be deterministic. We chose daily maximum temperature as the motivating example; however the proposed approach can be applied to other weather variables such as precipitation, solar radiation, or cloud cover that may be associated with adverse health outcomes through various pathways. In our study region of Alabama, we found the disagreement between observed and modeled temperatures to be the greatest at the lower tail of the distribution. Consequently, the estimated health impacts due to future extreme temperatures were similar between the calibrated and the original model projections. This result may vary across locations, especially if other weather variables or climate models are examined, and should be systematically explored in future analysis.

Our health impact calculations also suffer from many common limitations in estimating future disease burden due to climate change. For example, we did not account for the

expected changes in population structure, behaviors, size, or other factors that may influence the underlying health status of the population. Moreover, studies have shown that the temperature-mortality relationships exhibit heterogeneity across different NMMAPS regions [Anderson and Bell, 2011; Curriero et al., 2002]. While the attributable deaths calculation was not based on future population size, the IPCC scenario used for climate modeling includes population growth. We also recognize that our findings may be dependent on the chosen climate model and emission scenario. It is straightforward to conduct similar analysis with calibrated projections from different climate models following a recent exemplary effort by Peng et al. [2011]. Nonetheless, we report a significant increase in mortality attributable to future temperature extremes even in a region that is characterized by hot and humid summer seasons.

Figure 4.1: A process chart for our two-stage estimation. We first obtained the calibration model by comparing the original NARCCAP data with the corresponding observations through their underlying spatial-quantile processes in year 2000. We then fitted the health model to investigate the relationship between heat waves and mortality. Finally, we evaluated the future heat wave excess mortality based on the calibrated NARCCAP maximum temperature from 2041 to 2050.

Figure 4.2: Likelihood estimation without specifying a density function a priori. The blue curve is a density plot of simulated data. The likelihood function expressed as a quantile-based central tendency and a generalized Pareto tail are able to characterize the unbounded response variables.

Figure 4.3: Quantile-quantile plot of the model output and the observations. The blue curve is an estimated G_τ function.

Figure 4.4: Locations of the monitoring sites (blue triangles), centers of the NARCCAP grid cells (red dots) and NMMAPS communities (green circles) within Alabama.

Figure 4.5: At three urban communities Birmingham (birm), Mobile (Mobi), and Huntsville (hunt), we plot τ^{th} quantile curves of NARCCAP outputs, observations, our Bayesian calibrations and the simple linear regression (LM) of daily maximum temperature in year 2000, units in Celsius (C).

Figure 4.6: τ^{th} quantile curves of the historical NARCCAP data, the corresponding observations, the uncalibrated future model output and the calibrated daily maximum temperature, unit in Celsius. Note that the discrepancies mostly occurred at the lower tail of the distribution.

Table 4.1: Posterior medians (95% credible intervals) of the annual excess deaths attributable to high temperature and heat waves across three Alabama cities. For high temperature, the baseline group is defined as days with maximum temperature under 30°C and the at-risk group is defined as days with maximum temperature above the threshold. For heat wave, the at-risk group is defined as days with maximum temperature above the quantile threshold.

<u>High Temperature</u>			<u>Heat Wave</u>		
Threshold	1991-2000	2041-2050	Threshold	1991-2000	2041-2050
32.5 °C	48 (12, 88)	111 (29, 206)	90 th percentile	0.24 (-0.17, 0.74)	1.8 (-1.2, 5.5)
34.0 °C	31 (11, 50)	135 (48, 216)	95 th percentile	0.29 (0.04, 0.60)	3.9 (0.8, 7.9)
35.5 °C	20 (6, 29)	134 (44, 203)	97.5 th percentile	0.17 (0.10, 0.24)	4.8 (2.8, 7.1)
37.0 °C	7 (2, 11)	130 (38, 199)	99 th percentile	0.09 (0.06, 0.12)	5.2 (3.6, 6.9)

Table 4.2: Observed and projected average number of high temperature days and heat wave days per year in Birmingham, Alabama. High temperature days occur when maximum daily temperature exceeds the threshold temperature. Heat waves are defined as consecutive days with maximum daily temperature exceeding the quantile threshold. Posterior medians (95% credible intervals) are also given for the projections.

<u>High Temperature</u>			<u>Heat Wave</u>		
Threshold	1991-2000	2041-2050	Threshold	1991-2000	2041-2050
32.5 °C	45.9	106 (98, 113)	90 th percentile	11.1	75 (67, 86)
34.0 °C	22.7	89 (82, 97)	95 th percentile	5.2	61 (53, 71)
35.5 °C	12.5	72 (64, 82)	97.5 th percentile	2.0	49 (42, 59)
37.0 °C	4.1	54 (46, 64)	99 th percentile	0.9	44 (36, 52)

Chapter 5

Discussion and Future work

We propose a Bayesian spatial quantile calibration model for adjusting the behavior between model outputs and monitoring data. Particularly, we focus on calibrating the extreme values. Thus, instead of using the default approach based on the first two moments of the models and data, we calibrated the two data sources through their underlying quantile processes. We investigated two quantile processes: (1) estimated spatial-quantiles for the model output; (2) the predicted monitoring quantiles based on the model output calibrations. We conclude that the model output and monitoring data are similar around their median values, but present large differences at the upper and lower tails over the eastern US. The transformation between the model output and the observed quantile process is then applied to model output data, resulting in a calibrated series whose spatial and quantile structure is consistent with the monitoring data.

Due to the different spatial scales of the deterministic model output and the observations, we assume that both the model output and observed quantile processes have a spatial structure with exponential decay parameters. This assumption is made to obtain computing efficiency. More complicated spatial processes, i.e., a conditional autoregressive (CAR) model for the gridded model output data, and spatial linear coregionalization models for calibrating spatial quantiles, will be considered in future work.

Also, temporal components, known to be an important factor for ozone trend, play less of a role when taking both quantile and spatial structure into account (see Figure 5.1). Another approach is to consider the smoothing spline as a covariate, then evaluate its effect on the conditional distributions (see Figure 5.2 for the individual quantile surfaces for both the model output and monitoring data at a specific site); however, the quantile calibrations, as a transformations of one quantile process to another simultaneously, require a valid quantile process with the non-crossing and monotonic constraints. An efficient way to calibrate this type of spatial-temporal-quantile surface simultaneously is another avenue for future work.

Figure 5.1: The CMAQ and monitoring temporal quantiles at site 4. Under the non-crossing constraints, ozone quantile curves show little trend for both the CMAQ models and the monitoring data.

Figure 5.2: Temporal quantile surfaces on site 19 for both the CMAQ data and Observed data.

References

- G. B. Anderson and M. L. Bell. Weather-related mortality: a study of how heat, cold, and heat waves affect mortality in the united states. *Epidemiology*, 20(2):205–213, 2009.
- G. B. Anderson and M. L. Bell. Heat waves in the united states: Mortality risk during heat waves and effect modification by heat wave characteristics in 43 u.s. communities. *Environ Health Perspect*, 119(2):210–218, 2011. doi: 10.1289/ehp.1002313.
- M. L. Bell, R. Goldberg, C. Hogrefe, P. L. Kinney, K. Knowlton, B. Lynn, J. Rosenthal, C. Rosenzweig, and J. A. Patz. Climate change, ambient ozone, and health in 59 us cities. *Climatic Change*, 82:61–76, 2007.
- V. J. Berrocal, A. E. Gelfand, and D. M. Holland. A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15: 176–197, 2010. ISSN 1085-7117.
- H. D. Bondell, B. J. Reich, and H. Wang. Non-crossing quantil regression curve estimation. *Biometrika*, 97(4):825–838, 2010.
- D. Byun and K. L. Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system. *Applied Mechanics Reviews*, 59:51–77, 2006.
- B. Cai and D. B. Dunson. Bayesian multivariate isotonic regression splines. *Journal of the American Statistical Association*, 102(480):1158–1171, 2007.
- D. Caya, R. Laprise, M. Gigur and G. Bergeron, J. P. Blanchet, B. J. Stocks, and G. J. Boer. Description of the canadian regional climate model. *Water Air Soil Poll.*, 1:477–482, 1995.
- H. H. Chang, J. Zhou, and M. Fuentes. Impact of climate change on ambient ozone level and mortality in southeastern united states. *International Journal of Environmental Research and Public Health*, 7(7):2866–2880, 2010.
- F. C. Curriero, K. S. Heiner, J. M. Samet, S. L. Zeger, L. Strug, and J. A. Patz. Temperature and mortality in 11 cities of the eastern united states. *American Journal of Epidemiology*, 155(1):81–87, 2002.
- D. B. Dunson and J. A. Taylor. Approximate bayesian inference for quantiles. *Journal of Nonparametric Statistics*, 17(3):385 – 400, 2005.
- B. Eder and S. Yu. A performance evaluation of the 2004 release of models-3 cmaq. atmospheric environment, 2006.
- M. Fuentes and A. E. Raftery. Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45, 2005. ISSN 1541-0420.

- A. Gelman. Parameterization and bayesian modelling. *Journal of the American Statistical Association*, 99:537–545, 2004.
- A. Gryparis, C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10:258–274, 2009.
- Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):pp. 186–192, 1997. ISSN 00031305.
- J. R. M. Hosking and J. R. Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349, 1987. ISSN 00401706.
- R. Huth, J. Kysely, and L. Pokorna. A gcm simulation of heat waves, dry spells, and their relationships to circulation. *Climatic Change*, 46:29–60, 2000.
- D. E. John and J. B. Rose. Review of factors affecting microbial survival in groundwater. *Environ. Sci. Technol.*, 39(19):7345–56, 2005.
- T. R. Karl and R. W. Knight. The 1995 chicago heat wave: how likely is a recurrence? *Bulletin of the American Meteorological Society*, 78:1107–1119, 1997.
- M.C. Kennedy and A. O’Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. ISSN 1467-9868. doi: 10.1111/1467-9868.00294.
- VV. Kharin and F. W. Zwiers. Changes in extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean gcm. *J. Clim.*, 13:3760–3788, 2000.
- K. Knowlton, J. E. Rosenthal, C. Hogrefe, B. Lynn, S. Gaffin, R. Goldberg, C. Rosenzweig, J. Ku, and P. L. Kinney. Assessing ozone-related health impacts under a change climate. *Environ Health Perspec.*, 112:1557–1563, 2004.
- R. Knutti. Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Soceity A*, 366:4647–4664, 2008.
- R. Koenker. *Quantile Regression*. Econometric Society Monograph Series, Cambridge University Press, 2005.
- H. Kozumi and G. Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81:1 – 14, 2011. ISSN 0094-9655. doi: 10.1080/00949655.2010.496117.
- M. Lavine. On an approximate likelihood for quantiles. *Biometrika*, 82:220–222, 1995.
- C. Y. Lim, M. Stein, J. K. Ching, and R. Tang. Statistical properties of differences between low and high resolution cmaq runs with matched initial and boundary conditions. *Environmental Modelling and Software*, (25(1)):158–169, 2010.

- Z. Q. Lu and D. B. Clarkson. Monotone spline and multidimensional scaling. *Proceedings of American Statistical Association, Section of Statistical Computing*, pages 185–190, 1999. URL <http://www.reocities.com/zqjlu/asa2.pdf>.
- N. H. Ogden, A. Maarouf, I. K. Barker, M. Bigas-Poulin, M. G. Lindsay, C. J. Callaghan, F. Ramay, D. Waltner-Toews, and D. F. Charrn. Climate change and the potential for range expansion of the lyme disease vector *Ixodes scapularis* in Canada. *International Journal of Parasitology*, 36(1):63–70, 2006.
- C.J. Paciorek. Combining spatial information sources while accounting for systematic errors in proxies. *Journal of the Royal Statistical Society, In press*, 2011.
- E. Parzen. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365):pp. 105–121, 1979. ISSN 01621459.
- R. D. Peng and L. J. Wealty. The nmmapsdata package. *R News*, 4:10–14, 2004.
- R. D. Peng, J. F. Bobb, C. Tebaldi, L. McDaniel, M. L. Bell, and F. Dominici. Toward a quantitative estimate of future heat wave mortality under global climate change. *Environ Health Perspect*, 119(5):701–706, 2011. doi: 10.1289/ehp.1002430.
- J. O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4):425–441, 1988.
- B. J. Reich, M. Fuentes, and D. B. Dunson. Bayesian spatial quantile regression. *Journal of the American Statistical Association*, 106(493):6–20, 2011.
- J. S. Remais, S. Liang, and R. C. Spear. Coupling hydrologic and infectious disease models to explain regional differences in schistosomiasis transmission in southwestern China. *Environ. Sci. Technol.*, 42(7):2643–2649, 2008.
- J.F. Scinocca, N.A. McFarlane, M. Lazare, J. Li, and D. Plummer. The CCCMA third generation AGCM and its extension into the middle atmosphere. *Atmos. Chem. Phys.*, 8:7055–7074, 2008.
- S.T. Tokdar and J.B. Kadane. Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis*, 6(4):1–22, 2011.
- Y. Wu and Y. Liu. Stepwise multiple quantile regression estimation. *Statistics and Its Interface*, 2:299–310, 2009.
- K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4): 437 – 447, 2001. ISSN 0167-7152.
- J. Zhou, F. Montes, and J. Davis. Calibration of numerical model output using nonparametric spatial density functions. *Journal of Agricultural, Biological, and Environmental Statistics*, 16:531–553, 2011. ISSN 1085-7117.

APPENDIX

Appendix A

Appendix

A.1 Proper posterior distribution

If the likelihood is given by formula (3.5) and $p(\boldsymbol{\alpha}) \propto 1$, then the posterior distribution of $\boldsymbol{\alpha}$, $\pi(\boldsymbol{\alpha}|Y)$, will have a proper distribution. In other words:

$$0 < \int \pi(\boldsymbol{\alpha}|Y) d\boldsymbol{\alpha} < \infty \quad (\text{A.1})$$

proof: Suppose $y_{(1)} \leq y_{(2)} \dots \leq y_{(T)}$, and $\omega_L, \omega_U, C_L, C_U$ are positive numbers. Without loss of generality, let $\delta = 0$ and $\kappa = 0$. We first consider two extreme situations:

(1) $y_t < \alpha_0$, for all $t=1, 2, \dots, T$. Hence, we have $y_{(T)} < \alpha_0$ and:

$$\begin{aligned} \int \pi(\boldsymbol{\alpha}|Y) d\boldsymbol{\alpha} &= \int \prod_{t=1}^T f_Y(y_t | (\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \propto \int_{\{\alpha_0 > y_{(T)}\}} C_L \omega_L \exp\left\{-\sum_t \omega_L (\alpha_0 - y_t)\right\} d\boldsymbol{\alpha} \\ &\propto \int_{\{\alpha_0 > y_{(T)}\}} C_L \omega_L \exp\{-T\omega_L (\alpha_0 - \bar{y})\} d\boldsymbol{\alpha} \\ &\propto \frac{C_L}{T} \exp\{-T\omega_L (y_{(T)} - \bar{y})\} \\ &\in (0, \infty) \end{aligned} \quad (\text{A.2})$$

(2) Another situation is: $y_t > \alpha_0 + \sum \alpha_m$, for all y_t , $t=1, 2, \dots, T$. As a result, we have $y_{(1)} > \alpha_0 + \sum \alpha_m$ and:

$$\begin{aligned}
\int \pi(\boldsymbol{\alpha}|Y) d\boldsymbol{\alpha} &= \int \prod_{t=1}^n f_Y(y_t | (\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \\
&\propto \int_{\{\alpha_0 + \sum_m \alpha_m < y_{(1)}\}} C_U \omega_U \exp\left\{-\sum_t \omega_U (y_t - (\alpha_0 + \sum_m \alpha_m))\right\} d\boldsymbol{\alpha} \\
&\propto \int_{\{\alpha_0 + \sum_m \alpha_m < y_{(1)}\}} C_U \omega_U \exp\left\{-T\omega_U (\bar{y} - (\alpha_0 + \sum_m \alpha_m))\right\} d\boldsymbol{\alpha} \\
&\propto \frac{C_U}{T} \exp\{-T\omega_U (\bar{y} - y_{(1)})\} \\
&\in (0, \infty)
\end{aligned} \tag{A.3}$$

In general, suppose $y_{(1)} \dots, y_{(l-1)} < \alpha_0 \leq y_{(l)} \dots \leq y_{(u)} \leq \alpha_0 + \sum_m \alpha_m < y_{(u+1)} \dots, y_{(T)}$, then we have:

$$\begin{aligned}
\int \pi(\boldsymbol{\alpha}|Y) d\boldsymbol{\alpha} &\propto \frac{C_L}{(l-1)} \exp\left\{-\omega_L((l-1)y_{(l-1)} - \sum_{t=1}^{l-1} y_{(t)})\right\} \\
&\quad \times \frac{C_U}{(T-u)} \exp\left\{-\omega_U\left(\sum_{t=u+1}^T y_{(t)} - (T-u)y_{(u+1)}\right)\right\} \\
&\quad \times \int_{t=l}^u \left\{ \frac{1}{\frac{\partial}{\partial \tau} Q_Y(\tau)} \Big|_{\tau=\tau(y_{(t)})} \right\} d\boldsymbol{\alpha} \\
&\in (0, \infty)
\end{aligned} \tag{A.4}$$

The statement is proved.

A.2 Fitting the Generalized Pareto Distribution to the tails

When $\kappa = 0$, the GPD distribution defined as (4.6) reduces to an exponential distribution with mean $1/\omega$. Now we concentrate on the more difficult case where $\kappa \neq 0$. The idea is to make full use of the information contained in both central tendency and tails. Note that all the discussions below are aimed at analyzing upper tails, and it can be similarly applied on the

lower tails.

If the likelihood is given by (4.5), we first calculate the derivative given μ_U as

$$\frac{\partial f_Z(z)}{\partial z} \Big|_{z=\mu_U} = C_U \omega_U \left(\frac{1}{\kappa} - 1 \right) (-\kappa \omega_U) (1 - \kappa \omega_U (z - \mu_U)) \Big|_{z=\mu_U} = C_U \omega_U^2 (\kappa - 1) \quad (\text{A.5})$$

Because $\frac{\partial f_Z(z)}{\partial z} \Big|_{z=\mu_U} \leq 0$, thus $\kappa \leq 1$.

Based on the likelihood, we transform the overall order statistics $z_{(u+1)}, z_{(u+2)}, \dots, z_{(T)}$ to the partial order statistics on the upper tail as: $z_{(1;q)}^*, z_{(2;q)}^*, \dots, z_{(q;q)}^*$, where $z_{(i;q)}^* = z_{(u+i)} - \mu_U$, $i = 1, 2, \dots, q$, and $q = T - u$. Such order statistics are evaluated by first equating the CDF at the observed order statistics to their corresponding percentile values:

$$F(z_{(i;q)}^*; \kappa) = p_{i;q} \quad (\text{A.6})$$

where $p_{i;q} = \frac{1}{q}$ if $q \neq 1$. Substituting the corresponding CDF of (4.6) to (A.6) we have:

$$\frac{1}{\kappa} \ln(1 - \kappa \omega_U z_{(i;q)}^*) = -\Omega_i \quad (\text{A.7})$$

where $\Omega_i = -\ln(1 - p_{i;q}) > 0$. Consider the following function of κ :

$$h(\kappa) = \ln(1 - \kappa \omega_U z_{(i;q)}^*) + \kappa \Omega_i \quad (\text{A.8})$$

which is defined in $(-\infty, \min(1, \frac{1}{\omega_U z_{(i;q)}^*})]$. Then (A.7) is equal to:

$$h(\kappa) = 0 \quad (\text{A.9})$$

Subsequently, we use the resulting equation (A.9) as a basis for obtaining initial parameters for κ . By taking the derivative of $h(\kappa)$, we have:

$$\frac{\partial h(\kappa)}{\partial \kappa} = -\frac{\omega_U z_{(i;q)}^*}{1 - \kappa \omega_U z_{(i;q)}^*} + \Omega_i \quad (\text{A.10})$$

Let $\frac{\partial h(\kappa)}{\kappa} = 0$, we have $\kappa^* = \frac{1}{\omega_U z_{(i:q)}^*} - \frac{1}{\Omega_i}$. In general, the function $h(\kappa)$ is increasing when $\kappa \in (-\infty, \kappa^*)$ and decreasing when $\kappa \in (\kappa^*, \frac{1}{\omega_U z_{(i:q)}^*})$. Considering κ 's constraints, $h(\kappa)$ has the following properties:

1. $\kappa^* = 0$ if $\Omega_i = \omega_U z_{(i:q)}^*$. Therefore the function $h(\kappa)$ is increasing when $\kappa \in (-\infty, 0)$ and decreasing when $\kappa \in (0, \min(1, \frac{1}{\omega_U z_{(i:q)}^*}))$. Additionally, the solution for (A.9) is $\hat{\kappa} = 0$.
2. $\kappa^* > 0$ if $d\Omega_i = \omega_U z_{(i:q)}^*$, where $0 < d < 1$. Then $\kappa^* = (\frac{1}{d} - 1)\frac{1}{\Omega_i} > 0$ and $h(\kappa^*) > 0$. The solutions of (A.9) are: (1) $\kappa^* \geq 1 \Rightarrow \hat{\kappa} = 0$; (2) for $\kappa^* < 1$ and $\frac{1}{d\Omega_i} > 1$, if $h(1) < 0$, we use the bisection method to determine the solution in $(\kappa^*, 1)$, otherwise $\kappa = 0$; (3) if $\frac{1}{d\Omega_i} < 1$, we use the bisection method to determine the solution in $(\kappa^*, \frac{1}{d\Omega_i})$.
3. $\kappa^* < 0$ if $d\Omega_i = \omega_U z_{(i:q)}^*$, where $d > 1$. Given a negative number of large magnitude Δ , the solutions of (A.9) are: (1) $\kappa^* < \Delta \Rightarrow \hat{\kappa} = 0$; (2) $\kappa^* > \Delta$ and $h(\Delta) < 0$, we use the bisection method to obtain the solutions in (Δ, κ^*) .

These estimates of $\hat{\kappa}$ are computed for $z_{(i:q)}^*$, then combined in a suitable way to obtain final estimates (i.e., use the median of each of the foregoing set of estimators to obtain a corresponding overall estimators of κ). Subsequently, if $\tau > \tau_{z(u)}$, the τ^{th} quantile for the model output Z is obtained by the $\frac{\tau - \tau_{z(u)}}{1 - \tau_{z(u)}}$ th quantile of the GPD distribution with $\hat{\kappa}$.