

# EXTREME VALUE MODELING FOR SPACE-TIME DATA WITH METEOROLOGICAL APPLICATIONS

by

Huiyan Sang

Department of Statistical Sciences  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Dr. Alan E. Gelfand, Supervisor

\_\_\_\_\_  
Dr. Merlise Clyde

\_\_\_\_\_  
Dr. Gabriele C. Hegerl

\_\_\_\_\_  
Dr. Robert L. Wolpert

Dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in the Department of Statistical Sciences  
in the Graduate School of  
Duke University

2008

ABSTRACT

EXTREME VALUE MODELING FOR SPACE-TIME DATA  
WITH METEOROLOGICAL APPLICATIONS

by

Huiyan Sang

Department of Statistical Sciences  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Dr. Alan E. Gelfand, Supervisor

\_\_\_\_\_  
Dr. Merlise Clyde

\_\_\_\_\_  
Dr. Gabriele C. Hegerl

\_\_\_\_\_  
Dr. Robert L. Wolpert

An abstract of a dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in the Department of Statistical Sciences  
in the Graduate School of  
Duke University

2008

Copyright © 2008 by Huiyan Sang  
All rights reserved

# Abstract

Extreme value theory finds wide applications in areas such as environmental science, financial strategy of risk management and biomedical data processing. In this thesis, we present two spatial extreme value studies related to weather and climate events observed in space and in time, one of which motivates a novel methodology in constructing continuous spatial process for extreme values. Motivated by finding multiscale spatial dependence in extreme climate studies, we offer a new Bayesian analysis tool for learning about both large scale spatial dependence and microscale dependence. The last chapter presents a novel application of space-time models to synoptic climatology.

The first investigation is a development of hierarchical modeling approach for explaining a collection of spatially-referenced time series of extreme values. We assume that the observations follow Generalized Extreme Value(GEV) distributions whose locations and scales are jointly spatially dependent where the dependence is captured using multivariate Markov random field models specified through coregionalization. There are various ways to provide appropriate specifications; we consider four choices. The models can be fitted using a Markov Chain Monte Carlo (MCMC) algorithm to enable inference for parameters and to provide spatio-temporal predictions. We fit the models to a set of gridded interpolated precipitation data collected over a 50 year period for the Cape Floristic Region in South Africa, summarizing results for what appears to be the best choice of model.

In chapter 3, we extend the hierarchical modeling approach for explaining a collection of point-referenced time series of extreme values. Annual maxima are still assumed to follow GEV distributions, with parameters  $\mu$ ,  $\sigma$ , and  $\xi$  specified in the latent stage to reflect underlying spatio-temporal structure. Here, we relax the con-

ditionally independence assumption previously imposed in the first stage hierarchical models for annual maxima. Instead, a continuous spatial process model is proposed to account for spatial dependence which is unexplained by the latent spatio-temporal specifications for the GEV parameters. In addition, we offer an approach to make spatial interpolation for extreme values based on this hierarchical models with smoothed residuals across space. A simulation study is illustrated to investigate the model fitting behavior.

Motivated by the findings in extreme climate studies, which is, large scale spatial variations and small scale spatial variations coexist in some extreme climate phenomena, we present a Bayesian spatial modeling approach to make inference of both large scale and small scale spatial dependence in a general spatial setting. In particular, we focus on the investigation of microscale spatial dependence, which is defined as the dependence at distances smaller than the measurement scale of the spatial study. Since microscale spatial variation study often involves data observed at high resolution, we offer several potential solutions to tackle the computational difficulty of 'large n' problem.

In the last chapter, the application we focus on is to synoptic climatology where the goal is to develop an array of atmospheric states to capture a collection of distinct circulations. In particular, Self Organizing Maps (SOMs) are one of the recently used techniques in the meteorology community with regard to developing synoptic weather states. Little discussion about this technique has been found in the statistics literature. We introduce the stochasticity in the form of a space-time process model aiming to illuminate and interpret its performance in the context of application to daily data collection. That is, the observed daily state vectors are viewed as a time series of multivariate process realizations which we try to understand under the dimension reduction achieved by the SOM procedure.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>Acknowledgements</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Extreme Value Theory . . . . .	2
1.1.1 Univariate Extreme Value Theory . . . . .	3
1.1.2 Multivariate Extreme Value Theory . . . . .	6
1.2 Brief Review of Spatial Statistics . . . . .	11
1.2.1 Gaussian Processes and Covariance Functions . . . . .	11
1.2.2 The CAR model . . . . .	14
1.2.3 Coregionalized Models . . . . .	15
1.2.4 Predictive Process . . . . .	17
<b>2 Hierarchical Modeling for Extreme Values Observed over Space and Time</b>	<b>20</b>
2.1 Introduction . . . . .	20
2.2 Exploratory analysis . . . . .	23
2.3 Formal hierarchical space time models . . . . .	26
2.3.1 Modeling the $\mu_{i,t}$ . . . . .	27
2.3.2 The Coregionalized CAR model . . . . .	28
2.3.3 Bayesian implementation . . . . .	30
2.4 Model comparison and results . . . . .	30

2.5	Discussion and future work . . . . .	33
<b>3</b>	<b>A Continuous Spatial Process of Extreme Values</b>	<b>48</b>
3.1	Introduction . . . . .	48
3.2	Review on modeling for spatial extremes . . . . .	52
3.2.1	Hierarchical modeling approach for spatial extremes . . . . .	52
3.2.2	A continuous process for spatial extremes: moving maximum process . . . . .	55
3.3	A smooth spatial process for extreme values . . . . .	57
3.3.1	Copulas with multivariate extreme value distributions . . . . .	57
3.3.2	Transformed Gaussian processes . . . . .	60
3.4	Bayesian modeling approach for spatial extreme values . . . . .	63
3.4.1	A hierarchical modeling approach . . . . .	63
3.4.2	Bayesian implementation . . . . .	67
3.5	A simulation study of GEV model for spatio-temporal point-referenced data . . . . .	70
3.6	Discussion . . . . .	74
<b>4</b>	<b>Bayesian Analysis of Multiscale Spatial Dependence</b>	<b>83</b>
4.1	Introduction . . . . .	83
4.2	Hierarchical modeling for multi-scale spatial data . . . . .	87
4.3	Computations in coarse/fine spatial process model . . . . .	92
4.3.1	”Large n” problem in coarse scale spatial processes . . . . .	92
4.3.2	”Large n” in fine scale spatial processes . . . . .	96
4.4	Generalization to non-Gaussian spatial processes and spatio-temporal processes . . . . .	98
4.5	Simulation examples of prediction approximation . . . . .	101

4.6	Discussion . . . . .	103
<b>5</b>	<b>Interpreting Self Organizing Maps Through Space-time Data Models.</b>	<b>108</b>
5.1	Introduction . . . . .	108
5.2	A Review of SOM theory and implementation . . . . .	111
5.3	The dataset and some exploratory data analysis . . . . .	115
5.4	SOM modeling . . . . .	118
5.4.1	Dimension reduction . . . . .	118
5.4.2	Modeling specifications . . . . .	120
5.5	Model fitting issues . . . . .	124
5.5.1	Fitting models without spatial adjustment . . . . .	124
5.5.2	Fitting models with spatial adjustment . . . . .	125
5.6	Model comparison and model results . . . . .	126
5.7	Discussion . . . . .	129
	<b>Biography</b>	<b>150</b>

# List of Figures

2.1	Map of Cape Floristic Region in South Africa [Image from R. Cowling and D. Richardson, 1995. Fynbos. Fernwood Press]. . . . .	40
2.2	Image plot of the rainfall maxima over the gridded CFR in 1999. . . . .	41
2.3	(a): Histogram of weeks when annual rainfall maxima occurred in CFR; (b): Histogram of weeks when annual rainfall maxima occurred in western CFR; (c): Histogram of weeks when annual rainfall maxima occurred in eastern CFR. . . . .	42
2.4	MLE's of the $\mu_i$ and $\log\sigma_i$ for the exploratory models. . . . .	43
2.5	Scatter plot of the estimated $\mu_i$ and $\log\sigma_i$ for the exploratory models. . . . .	44
2.6	Posterior sample means of scale parameters and local trends obtained from Model C. . . . .	45
2.7	(a): Posterior sample means of the 1/25 return levels in 1998; (b): Posterior ratio of return level $z_{1/25}$ in 1998 over $z_{1/25}$ in 1950 obtained from Model C. . . . .	46
2.8	(a): The lower and upper .025 quantiles of the 1/25 return levels in 1998 obtained from Model C; (b): The lengths of 95% credible intervals of the 1/25 return levels in 1998 obtained from Model C. . . . .	47
3.1	Image plot of the temperature maxima over CFR in 1999; Spatial signal observed . . . . .	77
3.2	The extremal coefficient for the bivariate Gaussian copula extreme value distribution as a function of distance. . . . .	77
3.3	The extremal coefficients for the bivariate Gaussian copula under different correlation coefficients. . . . .	78
3.4	Simulated locations . . . . .	78
3.5	Simulated surfaces . . . . .	79

3.6	Histogram of the range parameter $\phi_z$ and the range parameter $\phi_\mu$ obtained by the smoothed model . . . . .	80
3.7	Histogram of the range parameter $\phi_\mu$ obtained by the nonsmoothed model . . . . .	81
3.8	From left to right, top to bottom: (a) The predictive median surface for extreme values using Model (3.26) (smoothed residuals); (b) The lengths of the 95% predictive intervals for Model (3.26) (smoothed residuals); (c) The predictive median surface for extreme values using Model (3.29) (non-smoothed residuals); (b) The lengths of the 95% predictive intervals for Model (3.29) (non-smoothed residuals) . . . . .	82
4.1	Exponential correlation functions at different scales. . . . .	105
4.2	Histogram of the coarse scale range parameter $\phi_{coarse}$ and the fine scale range parameter $\phi_{small}$ . . . . .	106
4.3	Histogram of single range parameter $\phi_{single}$ . . . . .	107
5.1	Projected locations of SOM nodes in 2 dimensional space. . . . .	131
5.2	Sea level pressure associated with each node. . . . .	132
5.3	Maximum temperature surface associated with each node. . . . .	133
5.4	Mapped coordinates in the 2-dimensional space for each of the 11,315 days. . . . .	134
5.5	Histograms of transition distances for each SOM group. . . . .	135
5.6	Transition probabilities from 1970-2000. . . . .	136

# List of Tables

2.1	DIC, AAPE and $\hat{r}$ for Models (A) - (D). . . . .	38
2.2	Posterior sample means of parameters and the corresponding 95% credible intervals for Model (C). $\beta_{lat}$ , $\beta_{lon}$ and $\beta_{elev}$ are the coefficients of longitude, latitude and elevation; $\xi$ is the shape parameter of GEV distribution; $\sigma_0$ is the center of spatially-varying scale parameters; $\rho$ is the global trend in location parameters for Model (C). . . . .	39
3.1	Posterior sample means of parameters and the corresponding 95% credible intervals for Model (3.26) and Model (3.29). $\beta$ is the intercept. $\sigma$ is the scale parameter of GEV and $\xi$ is the shape parameter of GEV; $\phi_\mu$ is the spatial range parameter in the location parameters and $\phi_z$ is the spatial range parameter in the GEV residuals . . . . .	76
3.2	Performance of Model (3.26) and Model (3.29) using averaged absolute predictive errors (AAPE), the deviance information criterion (DIC), and the empirical coverage probability $\hat{r}$ . . . . .	76
4.1	Parameter estimations obtained by Large/Small scale model . . . . .	104
4.2	Parameter estimations obtained by single process model . . . . .	104
4.3	Performance of Model 4.2 and Model 4.17 using root mean square predictive errors (RMSPE), the deviance information criterion (DIC), and the empirical coverage probability $\hat{r}$ . . . . .	105
5.1	(left): Frequency of occurrence of each node over the entire study period, e.g., 935 <sup>(4)</sup> indicates that the total number of occurrences of node 4 is 935; (right): Frequency of occurrence of each node during 1970-1979 and 1990-1999, e.g., (291, 285) <sup>(4)</sup> indicates that the number of occurrence of node 4 during 1970-1979 is 291, and that number is 285 during 1990-1999. . . . .	137
5.2	(left) Frequency of occurrence of each node during summer; (right) Frequency of occurrence of each node during winter. . . . .	137

5.3	Empirical transition probabilities. Each 4 by 3 sub-table in the following 4 by 3 array shows a set of transition probabilities. The array and sub-tables are arranged in the same way as in Table 5.1. . . . .	138
5.4	Performance of several VAR models using root mean square predictive errors (RMSPE), the deviance information criterion (DIC), and the empirical coverage probability $\hat{r}$ . $\mathbf{A}(tessellation)$ denotes regionally varying $\mathbf{A}$ as described in (5.3), $\mathbf{A}(year)$ denotes annually varying $\mathbf{A}$ as described in (5.4), $\boldsymbol{\eta}(quarter)$ denotes quarterly varying $\mathbf{A}$ which is constant as year within each quarter. $\boldsymbol{\eta}(quarter^*)$ denotes quarterly varying $\mathbf{A}$ which is also changing as year, $\mathbf{A}(tessellation, quarter)$ denotes $\mathbf{A}$ as described in (5.5) and $\boldsymbol{\eta}(spatial)$ denotes spatially varying adjustment as described in (5.6). . . . .	139
5.5	Posterior means and 95% credible intervals of $\mathbf{A}_{tessellation1,year2000}$ and $\boldsymbol{\Sigma}$ . . . . .	140
5.6	Estimated frequency of transition angle from each tessellation towards each of the four directions for the year 1999. . . . .	141
5.7	Estimated mean of the transition matrix in 2000. Each 4 by 3 sub-table in the following 4 by 3 array shows a set of transition probabilities. The array and sub-tables are arranged in the same way as in Table 5.1.	142
5.8	Estimated mean of the standard errors of the transition matrix in 2000.	143

# Acknowledgements

It is a pleasure to acknowledge with gratitude many people who made this thesis possible.

First of all, I would like to express my deep and sincere gratitude to my Ph.D. advisor Alan E. Gelfand, for his enthusiasm, his inspiration, his patience and his great efforts in supervising me. Throughout my Ph.D study, he provided encouragement, excellent teaching and flows of research guidance. He also supported me to visit Europe, South Africa and many US cities for conferences and workshops. More importantly, his enthusiasm for and commitment to research work set a strong model for me to follow.

I would like to give thanks to the members of my committee, Dr. Merlise Clyde, Dr. Robert Wolpert and Dr. Gabi Hegerl. I am also grateful to Chris Leonard, Bruce Hewitson and Anthony Rebelo in South Africa along with Andrew Latimer, Adam Wilson and John Silander, Jr in Connecticut for valuable discussions and advice on meteorology and ecology, and help in the development of the datasets. I am grateful to Sudipto Banerjee and Andrew Finley in Minnesota for the great collaborations in part of my thesis work. I also benefitted a lot from a year-long program at SAMSI. I would like to extend my thanks to Richard Smith and Dipak Dey for organizing such an inspiring working group.

The department of Statistical sciences at Duke university has been very supportive. I feel fortunate to spend four years in such a stimulating and friendly environment. I've had many colleagues within the department whom I wish to thank. Especially thanks to Zhi Ouyang, Joyee Ghosh, Robin Mitra, Liang Zhang for accompanying me through the FYE, the prelim, job search and defense.

I wish to thank my family for their supports. My special gratitude is due to my grandma, my parents, my parents in law, my little brother and my little sister in law. Most of all, I wish to thank my husband, Ping, for his endless support. To them I dedicate this thesis.

# Chapter 1

## Introduction

Increasing evidence shows that changes in climate at large regional scales can be attributed to an anthropogenic forcing (see, e.g., Hegerl et al. (2004), Levitus et al. (2000) and Barnett et al. (2001)). There is considerable attention in literature to the detection of potential changes in mean climate indices. Indeed, compared with mean climate, extreme climate events are likely to have a heavier impact on our whole society. Some extreme cases can yield natural catastrophes, such as major hurricane, severe drought and flood. It therefore becomes an important topic to investigate changes in climate extremes and monitor climate risk.

There are increasing research needs to statistically quantify changes in extreme climate events as well as the associated uncertainties. One of the primary research needs is to develop methods (ideally model based) to analyze and quantify the extreme behavior (both spatial variations and temporal trends). That is, to develop statistical modeling approaches for treating spatio-temporal dependence of extremes, which is especially useful to quantify uncertainties associated with model inferences. Another primary goal of extreme value study is that of prediction. That is, given the observed extreme values, to approximate the predictive distribution of the unobserved value at an unmonitored location (or for a future time). Prediction of extreme values is especially useful to be applied in quantitative risk monitoring of extreme climate events. In addition, prediction for extremes has great potential to be applied in statistical downscaling techniques for extremes.

The focus of this thesis is on addressing these two major research problems in extreme value studies. One essential step is to appropriately fit statistical distribution to describe

the behavior of extremes. At present, the standard approach utilizes extreme value theory, which has many potential advantages to model, detect or project trends in extremes of climate. We follow this path to consider a hierarchical modeling approach for explaining a collection of spatially-referenced time series of extreme values. Block maxima are taken as an extreme index and assumed to follow the GEV distributions conditional on the parameter values. For areal unit extreme values, univariate distributions of extreme values are extended to higher dimensions using latent multivariate Markov random field models specified through coregionalization, which allows the interpretation of high dimensional extreme value analysis including the nature of spatial association and the nature of temporal trend. For point referenced extreme values, latent spatial Gaussian process models are introduced to characterize spatial dependence and temporal trend. By relaxing the assumption of conditional independence in the hierarchical models, we extend hierarchical modeling approach to describe extreme values with a smoothed spatial process, which can be used in spatial interpolation with extremes.

This chapter makes a brief review of current univariate and multivariate extreme value theories in 1.1. In the second section of this chapter, we move to a quick review of spatial models for point-referenced data and areal unit data, making special emphasis on the spatial Gaussian processes, Markov Random Field models and multivariate spatial models.

## 1.1 Extreme Value Theory

Univariate extreme value theory has been well established. The aim of this section is to review the most important results in the classical theory of extremes. Primarily, we summarize the limit theorem for block maxima in subsection 1.1.1.

### 1.1.1 Univariate Extreme Value Theory

Extreme value theory begins with a sequence  $Y_1, Y_2, \dots$  of independent and identically distributed random variables with distribution function  $F(y)$  and, for a given  $n$  aims to find parametric models for  $M_n = \max\{Y_1, \dots, Y_n\}$ .

If the distribution function  $F(y)$  is specified, the exact distribution of  $M_n$  is immediately known. In the absence of such specification, extreme value theory seeks sequences  $\{a_n > 0\}$  and  $\{b_n\}$  such that  $\lim_{n \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq y\right) \equiv G(y)$  for a certain distribution function  $G$ .

**Definition 1.** *If there exist constants  $a_n > 0$ ,  $b_n \in \mathbb{R}$ , such that*

$$\lim_{n \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq y\right) = \lim_{n \rightarrow \infty} F^n(a_n y + b_n) \equiv G(y)$$

*for some non-degenerate distribution function  $G$ , then  $G$  is said to be a extreme value distribution and  $F$  is said to be in the domain of attraction of  $G$ , written  $F \in D(G)$ .*

**Definition 2.** *A distribution  $G$  is said to be max-stable if, for every  $n = 2, 3, \dots$ , there are constants  $a_n > 0$  and  $b_n$  such that*

$$\lim_{n \rightarrow \infty} G^n(a_n y + b_n) \equiv G(y)$$

The connection between the extreme value limit distribution and the max stable distribution is made by the following statement :

**Theorem 3.** *A distribution is max-stable if, and only if, it is a generalized extreme value distribution.*

The detailed proof of the above theorem can be found in Embrechts et al. (1997).

The entire range of possible extreme limit distribution for  $\frac{M_n - b_n}{a_n}$  is summarized by the extremal types theorem due to Fisher and Tippett (1928):

**Theorem 4.** *If there exist sequences of constants  $\{a_n > 0\}$  and  $\{b_n\}$  such that*

$$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \rightarrow G(y) \quad \text{as } n \rightarrow \infty$$

*where  $G$  is a non-degenerate distribution function, then  $G$  has a distribution function of the form*

$$G(y; \mu, \sigma, \xi) = \exp \left\{ - \left[ 1 + \xi \left( \frac{y - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \quad (1.1)$$

*defined on the set  $\{y : 1 + \xi(y - \mu)/\sigma > 0\}$ . This is called the generalized extreme value (GEV) family of distribution.  $\mu \in \mathbb{R}$  is the location parameter,  $\sigma > 0$  is the scale parameter and  $\xi \in \mathbb{R}$  is the shape parameter. There are three subfamilies which can be expressed under the umbrella of the GEV:  $\xi = 0$  corresponds to the Gumbel distribution;  $\xi > 0$  corresponds to the Frechet; and  $\xi < 0$  corresponds to the Weibull distribution.*

The GEV distribution is heavy-tailed and its probability density function decreases at a slow rate when the shape parameter  $\xi$  is positive. On the other hand, the GEV distribution has a bounded upper tail for a negative shape parameter. Formal justification of the Theorem (4) is detailed in Leadbetter et al. (1983).

A useful parameter of interest in many extreme value studies is the return level  $z_p$  associated with return period  $1/p$ . It is derived from the inverse of the GEV cumulative distribution function given parameter values in GEV distribution functions. Defining  $x_p = -\log(1 - p)$ , we have

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - x_p^{-\xi}] & \xi \neq 0 \\ \mu - \sigma \log(x_p) & \xi = 0 \end{cases} \quad (1.2)$$

Evidently, the return level  $z_p$  provides the threshold which is exceeded by the extreme value with probability  $p$ . Equivalently,  $z_p$  can be viewed as a threshold which is such that we expect an exceedance once every  $1/p$  years.

The univariate extreme value theory we reviewed so far is obtained based on a sequence of independent and identically distributed random variables. However, it is usually unrealistic to assume independency for the types of real data to which extreme value models are commonly applied. Coles (2001) discussed a generalization of extreme value theory from a sequence of independent random variables to a stationary series.

**Theorem 5.** *Let  $X_1, X_2, \dots$  be a stationary process and  $X_1^*, X_2^*, \dots$  be a sequence of independent variables with the same marginal distribution. Define  $M_n = \max\{X_1, \dots, X_n\}$  and  $M_n^* = \max\{X_1^*, \dots, X_n^*\}$ . Under suitable regularity conditions,*

$$\{(M_n^* - b_n)/a_n \leq z\} \rightarrow G_1(z) \quad (1.3)$$

*as  $n \rightarrow \infty$  for normalizing sequences  $\{a_n > 0\}$  and  $\{b_n\}$ , where  $G_1$  is a non-degenerate distribution function, if and only if*

$$\{(M_n - b_n)/a_n \leq z\} \rightarrow G_2(z), \quad (1.4)$$

*where*

$$G_2 = G_1^\theta(z) \quad (1.5)$$

*for a constant  $0 < \theta \leq 1$ .*

This theorem implies that maxima of a stationary series converge to the GEV distribution. And the effect of dependence in the stationary series is reflected in the association of their limit distribution  $G_2$  to the limit distribution  $G_1$  which is arisen from the corresponding independent series.

This theorem essentially relaxes the concern of applying the GEV family as a model for block maxima of stationary series, provided that the range of dependence in the stationary series is not too long and the block size is appropriately selected. On one hand, the GEV

distribution is an asymptotic distribution which we use to approximate for maxima of long, finite sequences. The approximation quality is relying on the ‘effective’ length of observations. Therefore, typically, longer range of dependence requires a larger block size to ensure the validity of using GEV family as a limit distribution for block maxima. On the other hand, larger blocks generate fewer block maxima, leading to less sufficient information in extreme value modeling. Ideally, the choice of blocks should balance the trade-off between inference power and asymptotic behavior of the limit distribution.

### 1.1.2 Multivariate Extreme Value Theory

In this subsection, we extend the review from the univariate extreme value theory to the multivariate case. We will give an overview of the various characterizations for multivariate extremes. Again, the focus is on the block maxima.

#### Asymptotic Characterization

Assume that the extreme process is observed at a finite number of locations  $(\mathbf{s}_1, \dots, \mathbf{s}_p)$ . Let  $\mathbf{X}_i$ ,  $i = 1, 2, \dots$  be iid replicates of a random vector  $\mathbf{X} = (X(\mathbf{s}_1), \dots, X(\mathbf{s}_p))^T$ . Suppose the distribution of  $\mathbf{X}_i$  is given by  $F(\mathbf{x}) = Pr(X_i(\mathbf{s}_1) \leq x_1, \dots, X_i(\mathbf{s}_p) \leq x_p)$  for  $i = 1, 2, \dots$ . Let  $\mathbf{M}_n = (\bigvee_{i=1, \dots, n} X_i(\mathbf{s}_1), \dots, \bigvee_{i=1, \dots, n} X_i(\mathbf{s}_p))$  denote the vector of pointwise maxima, where  $\bigvee$  denotes max.

**Definition 6.** *If there exist normalizing sequences  $\mathbf{a}_n > 0$ ,  $\mathbf{b}_n$  such that*

$$\lim_{n \rightarrow \infty} Pr\left(\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \leq \mathbf{x}\right) = \lim_{n \rightarrow \infty} F^n(\mathbf{a}_n \mathbf{x} + \mathbf{b}_n) \equiv \mathbf{G}(\mathbf{x})$$

*for some non-degenerate distribution  $\mathbf{G}$ , the distribution  $\mathbf{G}$  is called a  $D$ -dimensional multivariate max-stable distribution. and  $F$  is said to be in the domain of attraction of  $\mathbf{G}$ , which we write  $F \in D(\mathbf{G})$ .*

These distributions have received considerable attentions in literature (de Haan (1985), Pickands (1981), Resnick (1987) and Resnick (2007)). There are two important facts about the extreme value distribution theory. First, if we fix some  $j \in \{1, \dots, p\}$  and define  $x_{j'} = \infty$  for  $j' \neq j$ , then

$$Pr\left(\frac{\mathbf{M}_{nj} - \mathbf{b}_{nj}}{\mathbf{a}_{nj}} \leq \mathbf{x}\right) \longrightarrow G(\infty, \infty, \dots, x_j, \dots, \infty)$$

Therefore, each marginal distribution of multivariate extremes follows one of those three types of GEV distribution family described in theorem (4). Second, for any given monotone increasing functions  $\{g_j, j = 1, \dots, p\}$ ,  $\max\{g_j(X_{i,j}), \dots, g_j(X_{n,j})\} = g_j(\max\{X_{1,j}, \dots, X_{n,j}\})$ . It is therefore convenient to derive the joint distribution of maxima over  $\{g_j(X_j)\}$  from the joint distribution of  $\mathbf{M}_n$ . One convenient fact about the GEV distribution is that if  $X \sim GEV(\mu, \sigma, \xi)$ , then  $X^* = (1 + \sigma \frac{x - \mu}{\xi})^{-1/\xi}$  follows unit Fréchet distribution, with distribution function  $\exp(-x^{-1})$  for  $x > 0$ . Without loss of generality, it is common to restrict the marginal distribution of multivariate extremes to be a standard Fréchet distribution in multivariate extreme distribution studies.

The limiting distribution for component-wise maxima can be written in Pickand's representation:  $G(\mathbf{x}) = \exp\{-V(\mathbf{x})\}$ , where function  $V(x)$  is called the exponent measure. Define  $\mathcal{S}_p = \{(x_1, \dots, x_p) : x_1 \geq 0, \dots, x_p \geq 0, \sum_{i=1}^p x_i = 1\}$ . Then using the radial-spectral decomposition, we can write :

$$V(x) = \int_{\mathcal{S}_p} \max\left(\frac{w_1}{x_1}, \dots, \frac{w_p}{x_p}\right) dH(w), \quad (1.6)$$

and  $H$  is a spectral measure on the  $p$ -dimensional unit simplex  $\mathcal{S}_p$ . Again, for any  $j \in \{1, \dots, p\}$ ,  $V(\infty, \infty, \dots, j, \dots, \infty) = \int_{\mathcal{S}_p} \max\left(\frac{w_1}{x_1}, \dots, \frac{w_p}{x_p}\right) dH(w) = \frac{1}{x_j} \int_{\mathcal{S}_p} w_j dH(w)$ . To ensure the marginal distributions are standard Fréchet distributions,  $H(w)$  has to satisfy  $\int_{\mathcal{S}_p} w_j dH(w) = 1$  for each  $j$  from 1 to  $p$ .

In fact, any spectral measure  $H$  satisfying the above constraint can yield a valid mul-

tivariate max-stable distribution defined in Definition (6). Therefore, there is no finite parametric model to characterize the entire family of the multivariate max-stable distribution. Nevertheless, several parametric subfamilies have been suggested for lower dimension extreme values, with special emphasis on bivariate case (see the review in Kotz and Nadarajah (2000)). Gumbel and Goldstein (1964) proposed the logistic type multivariate extreme value distribution. The exponent measure function under their model takes the following form:

$$V(x) = \left( \sum_{j=1}^p x_j^{-r} \right)^{1/r} \quad (1.7)$$

where  $r \geq 1$ . Here,  $r$  is a parameter characterizing the dependence between components.  $r = 1$  corresponds the independent case and  $r \rightarrow \infty$  when  $x_1, \dots, x_p$  are complete dependent. The drawback of this logistic type multivariate extreme value distribution is that the dependence structure is only characterized by a single parameter  $r$ . As a result, This method is not flexible enough to capture general dependence structures for large dimension extremes (e.g.,  $p > 2$ ). Tawn (1988) extended the logistic model by introducing the asymmetric logistic extreme value distribution which has more parameters characterizing dependence among extreme values.

$$V(x) = \sum_{c \in C} \left\{ \sum_{i \in c} \left( \frac{\theta_{i,c}}{x_i} \right)^{r_c} \right\}^{1/r_c}, \quad (1.8)$$

where  $C$  is the class of non-empty subsets of  $\{1, \dots, d\}$ ,  $r_c \geq 1$ ,  $\theta_{i,c} = 0$  if  $i \notin c$ ,  $\theta_{i,c} \geq 0$ ,  $\sum_{c \in C} \theta_{i,c} = 1$  for each  $i$ . Now, we introduce the third commonly used parametric multivariate extreme distribution, called negative logistic ( Joe (1999)). Its exponent function has the the following form:

$$V(x) = \sum \frac{1}{x_j} + \sum_{c \in C: |c| \geq 2} (-1)^{|c|} \left\{ \sum_{i \in c} \left( \frac{\theta_{i,c}}{x_i} \right)^{r_c} \right\}^{1/r_c}, \quad (1.9)$$

$r_c \leq 0$ ,  $\theta_{i,c} = 0$  if  $i \notin c$ ,  $\theta_{i,c} \leq 0$ ,  $\sum_{c \in C} (-1)^{|c|} \theta_{i,c} \leq 1$  for each  $i$ . However, both the negative

logistic model and the asymmetric logistic model are not good candidates to model high dimensional extremes because of the over-parametrization.

There are some research in which efforts are concentrated on defining the spectral density  $h$  directly. For example, Coles and Tawn (1991) showed that the tiled Dirichlet density provides a valid spectral density for the multivariate extreme distribution. Suppose  $h^*$  is an arbitrary positive function on  $\mathcal{S}_p$  with  $m_j = \int_{\mathcal{S}_p} u_j h^*(\mathbf{u}) d\mathbf{u} < \infty$ , then define  $h(w) = (\sum m_k w_k)^{-(p+1)} \prod_{j=1}^p m_j h^*(\frac{m_1 w_1}{\sum m_k w_k}, \dots, \frac{m_d w_d}{\sum m_k w_k})$ .  $h$  is the density of positive measure  $H$  satisfying  $\int_{\mathcal{S}_p} u_j dH(\mathbf{u}) = 1$ .

As a special case of this, they considered Dirichlet density  $h^*(\mathbf{u}) \sim Dir(\alpha)$  where  $\alpha = (\alpha_1, \dots, \alpha_p)$ , which leads to

$$h(w) = \prod_{j=1}^p \frac{\alpha_j}{\Gamma(\alpha_j)} \frac{\Gamma(\sum \alpha_j + 1)}{\sum \alpha_j w_j^{p+1}} \prod_{j=1}^p \left( \frac{\alpha_j w_j}{\sum \alpha_k w_k} \right)^{\alpha_j - 1} \quad (1.10)$$

Coles (1993) showed that the Dirichlet model is preferable to the logistic and negative logistic models in a simulated spatial rainfall extremes study. However, a disadvantage of this model is that the integral in equation (1.6) is typically difficult to evaluate. Cooley et al. (2008) constructed a new spectral density which they call the pairwise beta model.  $h_p(w; \alpha, \beta) = K_p(\alpha) \sum_{1 \leq i < j \leq p} h_{i,j}(w_i, w_j; \alpha, \beta_{i,j})$ , where

$$h_{i,j}(w_i, w_j; \alpha, \beta_{i,j}) = (w_i + w_j)^{(p-1)(\alpha-1)} (1 - (w_i + w_j))^{\alpha-1} \times \frac{\Gamma(2\beta_{i,j})}{(\Gamma(\beta_{i,j}))^2} \left( \frac{w_i}{w_i + w_j} \right)^{\beta_{i,j}-1} \left( \frac{w_j}{w_i + w_j} \right)^{\beta_{i,j}-1}$$

$K_p(\alpha)$  is a normalizing constant such that  $\int_{\mathcal{S}_p} w_i h_p(w; \alpha, \beta) dw = 1$ .

## Dependence Measures for the Multivariate Extremes

Several dependence measures have been proposed to quantify the magnitude of dependence for a multivariate max-stable distribution (see, e.g., Joe (1997), Coles et al. (1999) and

Cooley et al. (2006)). First, it is noteworthy that the commonly used correlation coefficient is not a proper candidate to describe extreme dependence. The reason is that the moments of the GEV distribution do not exist under certain cases. For example, the first moment of the GEV distribution is not finite provided that the shape parameter  $\xi > 1$ . Alternative dependence measures to correlation coefficient include Kendall's  $\tau$ , Spearman's  $\rho$  and Schweizer and Wolff's  $\sigma$  (see, e.g., Nelsen (2006)). The above three measures share a common difference from the correlation coefficient in that they are all invariant under any strictly monotone transformation functions. Consider a bivariate random variable  $(X, Y)$ . The definitions of Kendall's  $\tau$ , Spearman's  $\rho$  and Schweizer and Wolff's  $\sigma$  are given as below:

$$\tau_{X,Y} = 4 \iint F(x, y) dF(x, y) - 1$$

Spearman's  $\rho$

$$\rho_{X,Y} = 12 \iint (F(x, y) - F(x)F(y)) dF(x) dF(y)$$

Schweizer and Wolff's  $\sigma$

$$\sigma_{X,Y} = 12 \iint |F(x, y) - F(x)F(y)| dF(x) dF(y)$$

Moreover, several dependence metrics have been developed from the bivariate extreme value distribution theory. The most commonly used one is called the extremal coefficient (Smith (1990) and Schlather and Tawn (2003)). Again, suppose  $(X, Y)$  is a bivariate random vector with standard Fréchet marginal distributions, and the joint distribution of  $(X, Y)$  is  $F(x, y) = \exp\{-V(x, y)\}$ . Then the extremal coefficient is defined as  $\phi = V(1, 1)$ .

$$\Pr(X < r, Y < r) = \exp\{-V(r, r)\} = (\exp(-r^{-1}))^{V(1,1)} = (\Pr(X < r))^{V(1,1)}$$

The pairwise extremal coefficient can be interpreted as the effective number of independent random variables for the bivariate random variables. Hence  $\phi \in [1, 2]$ . And if  $(X, Y)$  are

independent, then  $\phi = 2$ . If  $(X, Y)$  are complete dependent,  $\phi = 1$ . Higher-order extremal coefficients are defined in the similar way as in the bivariate case (see, Schlather and Tawn (2003)).

## 1.2 Brief Review of Spatial Statistics

Spatial data sets are conventionally classified into one of the three basic types: (1) point-referenced data (also referred to geocoded or geostatistical data) where  $Y(\mathbf{s})$  is a random vector at a location  $\mathbf{s} \in \mathbb{R}^r$ , where  $\mathbf{s}$  varies continuously over a fixed region  $D \in \mathbb{R}^r$ ; (2) areal data, where  $D$  is a fixed subset which is partitioned into a finite number of areal units (zip codes, counties, etc.); (3) point pattern data, where the spatial domain  $D$  is random. In subsection 1.2.1 and subsection 1.2.2 we review some basic models for the point-referenced data and the areal unit data respectively. In subsection 1.2.3, we turn our attention to the problem of multivariate spatial modeling. In particular, we focus on the description of the so called coregionalization approach which is extensively applied in the multivariate spatial process models.

### 1.2.1 Gaussian Processes and Covariance Functions

The Gaussian spatial process model is the most common model for the point-referenced data (see, e.g., Banerjee et al. (2004)). Assume a response variable  $Y(\mathbf{s})$  observed at a generic location  $\mathbf{s} \in D \subseteq \mathbb{R}^2$  along with a  $p \times 1$  vector of spatially referenced predictors  $\mathbf{x}(\mathbf{s})$ . Then, model-based geostatistical data analysis typically proceeds from spatial regression models such as,

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (1.11)$$

The residual from the regression is partitioned into a spatial process,  $w(\mathbf{s})$ , capturing

spatial association, and an independent process,  $\epsilon(\mathbf{s})$ , also known as the *nugget* effect, modelling measurement error or micro-scale variation (see, e.g., Chiles and Delfiner (1999)). The  $w(\mathbf{s})$  are often referred to as spatial random effects, providing local adjustment (with structured dependence) to the mean, interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern.

We assume observations  $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^T$  from  $n$  locations and treat the data as a partial realization of a spatial process, modelled through  $w(\mathbf{s})$ . The most common specification for  $w(\mathbf{s}) \sim GP(0, C(\cdot, \cdot))$  is a Gaussian Process, which is a stochastic process specified by its mean and covariance function  $C(\mathbf{s}_i, \mathbf{s}_j)$ , which is defined for pairs of sites  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . Sometimes we specify  $C(\mathbf{s}, \mathbf{s}') = \sigma^2 \rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$  where  $\rho(\cdot; \boldsymbol{\theta})$  is a correlation function and  $\boldsymbol{\theta}$  includes parameters quantifying rate of correlation decay and smoothness of realizations though this limits us to constant process variance. In any event,  $\epsilon(\mathbf{s}) \stackrel{iid}{\sim} N(0, \tau^2)$  for every location  $\mathbf{s}$ .

For a given collection of sites in  $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  the data likelihood is given by  $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$ , with  $\Sigma_{\mathbf{Y}} = C(\boldsymbol{\theta}) + \tau^2 I_n$ , where  $X = [\mathbf{x}^T(\mathbf{s}_i)]_{i=1}^n$  is a matrix of regressors and  $C(\boldsymbol{\theta}) = [C(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$  is the spatial covariance matrix corresponding to  $w(\mathbf{s})$ . A valid covariance function must be positive definite, i.e.,  $\sum_{i=1}^n \sum_{j=1}^n a_i a_j C(\mathbf{s}_i, \mathbf{s}_j)$  for every  $n$  and every collection  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  and every vector  $\mathbf{a}$ .

The process  $\mathbf{Y}$  is said to be a strong stationary spatial process if for any  $n \geq 1$ , any  $\mathbf{h} \in \mathbb{R}^r$  and any set of sites  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ , the distribution of  $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1), \dots, \mathbf{Y}(\mathbf{s}_n))$  is the same as that of  $(\mathbf{Y}(\mathbf{s}_1 + \mathbf{h}), \dots, \mathbf{Y}(\mathbf{s}_n + \mathbf{h}))$ . The process  $\mathbf{Y}$  is said to be a weak stationary spatial process if  $\mu(\mathbf{s}) = \mu$  and  $Cov(\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$  for any  $\mathbf{h} \in \mathbb{R}^r$ . Clearly, for the Gaussian process, weak stationarity implies strong stationarity.

The process is said to be isotropic if the covariance function  $C(\mathbf{s}, \mathbf{s}') = C(\|\mathbf{s} - \mathbf{s}'\|)$ , where  $\|\mathbf{s} - \mathbf{s}'\|$  denotes the Euclidean distance between  $\mathbf{s}$  and  $\mathbf{s}'$ . We call a process homogeneous

if the process is intrinsically stationary and isotropic. A homogeneous process implies that the covariance relationship between the values of the process at any two locations can be characterized by a covariance function  $C(\|\mathbf{h}\|)$  which only depends on the length (e.g., Euclidean length) of the separation vector  $\mathbf{h}$ .

Many stationary, isotropic covariance function have been proposed and widely accepted in practice because of their simplicity, interpretability (see Banerjee et al. (2004) for a review). Denoting  $\|\mathbf{h}\|$  by a scalar  $t$ . Below, we briefly review four simple but important parametric forms: the exponential covariance function, the Gaussian covariance function, Spherical and the Matérn covariance function.

- Spherical

$$C(t) = \begin{cases} 0 & \text{if } t \geq \phi \\ \sigma^2(1 - 3t/(2\phi) + (t/\phi)^3/2) & \text{if } 0 < t \leq \phi \\ \sigma^2 + \tau^2 & \text{otherwise} \end{cases} \quad (1.12)$$

The spherical covariance function offers clear interpretations of the nugget parameter  $\tau^2$ , scale parameter  $\sigma$  and range parameters  $\phi$ .

- Exponential covariance function:

$$C(t) = \begin{cases} \sigma^2 \exp(-t/\phi) & \text{if } t > 0 \\ \sigma^2 + \tau^2 & \text{otherwise} \end{cases} \quad (1.13)$$

Exponential covariance function also has a range parameter  $\phi$  which controls the spread of the spatial correlation and is called the effective spatial range (i.e.,  $-\log(0.05)\phi$  is the distance at which the correlation drops to 0.05).

- Gaussian

$$C(t) = \begin{cases} \sigma^2 \exp(-t^2/\phi^2) & \text{if } t > 0 \\ \sigma^2 + \tau^2 & \text{otherwise} \end{cases} \quad (1.14)$$

The Gaussian covariance function yields a very smooth realizations of the spatial process, i.e., mean square derivatives of all orders exist.

- Matérn

$$C(t) = \begin{cases} \frac{\sigma^2}{2^{v-1}\Gamma(v)} (2\sqrt{v}t/\phi)^v K_v(2\sqrt{v}t/\phi) & \text{if } t > 0 \\ \sigma^2 + \tau^2 & \text{otherwise} \end{cases} \quad (1.15)$$

Parameter  $v$  is called the smoothness parameter. Not surprisingly, it is a parameter controlling the smoothness of the realized random field.  $\phi$  is a spatial range parameter. The function  $\Gamma(\cdot)$  is the usual gamma function and  $K_v$  is the modified Bessel function of order  $v$ . Matérn family is perhaps the most widely used covariance function because it flexibly encompasses several class of valid covariance function, including the exponential covariance function ( $v = 1/2$ ) and the Gaussian covariance function ( $v \rightarrow \infty$ )

### 1.2.2 The CAR model

Conditional Autoregressive (CAR) model is a common method in dealing with univariate areal data. Recall the univariate CAR model (see Besag (1974) and Banerjee et al. (2004)). Suppose  $\mathbf{V} = (V_1, V_2, \dots, V_n)'$  is a vector of spatial random effects which is defined at areal sites from 1 to  $n$ . The joint distribution of  $\mathbf{V}$  are defined through the conditional gaussian specifications at each sites as follows.  $V_i | V_j, j \neq i \sim N(\rho \sum_j w_{i,j} / w_{i+} V_j, 1/w_{i+})$  The joint distribution of vector  $\mathbf{V}$  can be expressed as  $V \sim N(0, (D_w - \rho W)^{-1})$  where  $D_W$  is diagonal with  $(D_W)_{i,i} = W_{i,+}$ , which is the summation of the entries in the  $i$ th row of  $W$ .

The matrix  $W$  is referred to as a proximity matrix with entries  $w_{i,j}$  which defines neighborhood information in space. Usually, the entries of  $W$  are normalized to reflect weights in the strength of association among units. For example,  $W$  is commonly pre-specified as binary, i.e.,  $w_{i,j} = 1$  if  $i$  and  $j$  share some common boundary, otherwise  $w_{i,j} = 0$ .

$\rho$  is a so called propriety parameter, which could be viewed as a proportion of  $V_i$  to the weighted average of its neighbors  $\sum_j \rho w_{i,j} V_j / w_{i+}$ .  $\rho = 1$  corresponds to an improper CAR model.  $\rho = 0$  corresponds to the case of independence among  $V_1, V_2, \dots, V_n$ . The introduction of  $\rho$  in CAR model is to ensure the existence of the precision matrix  $\Sigma_v^{-1} = (D_w - \rho W)^{-1}$ . In hierarchical model settings, CAR model is often employed as a prior, which might still yield proper posterior. Therefore,  $\rho$  is set to be 1 in our model. For further discussion, see Banerjee et al. (2004).

### 1.2.3 Coregionalized Models

When we have multivariate spatial data over the same region which have dependence among each other as well as spatial dependence between sites, multivariate spatial model is then desired. There are many statistical investigations of this problem in the literature (see Banerjee et al. (2004)). The crucial objective is to seek flexible, interpretable and computationally tractable models to describe the multivariate spatial data. And the key to the solution usually relies on the way of constructing a valid cross-covariance matrix associated with the multivariate spatial data. Let  $\mathbf{Y}(\mathbf{s})$  denote a  $p \times 1$  vector of random variables at location  $\mathbf{s}$ . The cross-covariance is defined as a  $p \times p$  matrix  $C(\mathbf{s}, \mathbf{s}') \equiv cov(\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}'))$ . A cross-covariance matrix is called valid if for any number of and choices of locations, the resulting  $np \times np$  covariance matrix for  $\mathbf{Y}$  is positive definite. There has been some approaches to construct a valid cross-covariance matrix, including separability, coregionalization, moving averages, and convolution. In this subsection, we primarily focus on the

review of the coregionalization approach.

This approach is through the linear model of coregionalization ( Wackernagel (2003), Matheron (1982), Gelfand and Vounatsou (2003) and Jin et al. (2005)), which is a constructive way to model multivariate spatial variables jointly by specifying their conditional distributions and marginal distributions.

Let  $\mathbf{U}_i = (U_{i,1}, U_{i,2}, \dots, U_{i,n})'$  denotes the spatial component  $i$  in the study region, for  $i = 1, 2, \dots, p$ , where  $p$  is the number of co-regional spatial components. By introducing a transformation lower triangular matrix  $A$  with elements  $a_{i,j}$ , we now could model dependent spatial components  $(\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_p)$  as:

$$\begin{pmatrix} U_{1,i} \\ U_{2,i} \\ \vdots \\ U_{p,i} \end{pmatrix} = A \begin{pmatrix} V_{1,i} \\ V_{2,i} \\ \vdots \\ V_{p,i} \end{pmatrix}$$

Denote  $\mathbf{V}_j = (V_{j,1}, V_{j,2}, V_{j,n})$ , for  $j = 1, \dots, p$ . For each  $\mathbf{V}_j$ , we could separately use univariate CAR models for the areal unit data and use Gaussian process model for the point-referenced data.

Take the multivariate CAR models as an example. The joint distribution of  $(\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_p)$  are constructed as:

$$(\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_p)' \sim N(\mathbf{0}, (A \otimes I_{n \times n})\Gamma^{-1}(A \otimes I_{n \times n})') \quad (1.16)$$

$$\text{where } \Gamma = \begin{pmatrix} D_w - \rho W & 0 & \dots & 0 \\ 0 & D_w - \rho W & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & D_w - \rho W \end{pmatrix}$$

Benefits of working with the conditional form of the model are certainly computational and possibly mechanistic or interpretive.

## 1.2.4 Predictive Process

Both estimation and prediction of Model 1.11 require evaluating the Gaussian likelihood using either Bayesian analysis or Maximum likelihood approach. For  $n$  observations  $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^T$   $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$ , with  $\Sigma_{\mathbf{Y}} = C(\boldsymbol{\theta}) + \tau^2 I_n$ , where  $X = [\mathbf{x}^T(\mathbf{s}_i)]_{i=1}^n$  is a matrix of regressors and  $C(\boldsymbol{\theta}) = [C(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$ . Evidently, evaluating the likelihood involves the inversion of  $n \times n$  matrix  $\Sigma_{\mathbf{y}}$ , which is computationally expensive for big  $n$ , even more so with repeated evaluation as needed in MCMC algorithms.

Recently, Banerjee et al. (2007) proposed a class of models based upon a *predictive process* that operates on a specified lower-dimensional subspace by projecting the original or *parent* process. The lower-dimensional subspace needs to be chosen by the user by selecting a set of “knots”  $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$ , which may or may not form a subset of the entire collection of observed locations  $\mathcal{S}$  (see Finley and Gelfand (2008, submitted) for a discussion on knots selection). The predictive process  $\tilde{w}(\mathbf{s})$  is defined as the “kriging” interpolator

$$\tilde{w}(\mathbf{s}) = E[w(\mathbf{s})|\mathbf{w}^*] = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*, \quad (1.17)$$

where  $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim MVN(\mathbf{0}, C^*(\boldsymbol{\theta}))$  comprises the parent process realization over the knots in  $\mathcal{S}^*$ ,  $C^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$  is the corresponding  $m \times m$  covariance matrix, and  $\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$ .

The predictive process  $\tilde{w}(\mathbf{s}) \sim GP(0, \tilde{C}(\cdot))$  defined in (1.17) has nonstationary covariance function,

$$\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s}'; \boldsymbol{\theta}), \quad (1.18)$$

and is completely specified by the parent covariance function. Realizations of this predictive process are given by  $\tilde{\mathbf{w}} = [\tilde{w}(\mathbf{s}_i)]_{i=1}^n \sim MVN(\mathbf{0}, \mathbf{c}^T(\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\boldsymbol{\theta}))$ , where  $\mathbf{c}^T(\boldsymbol{\theta})$  is the  $n \times m$  matrix whose  $i$ -th row is given by  $\mathbf{c}^T(\mathbf{s}_i; \boldsymbol{\theta})$ . The attractive theoretical properties of

the predictive process including its role as an optimal approximator have been discussed in Banerjee et al. (2007).

The predictive process in (1.17) immediately extends to multivariate Gaussian process settings. For a  $q \times 1$  multivariate Gaussian parent process,  $\mathbf{w}(\mathbf{s})$ , the corresponding predictive process is

$$\tilde{\mathbf{w}}(\mathbf{s}) = Cov(\mathbf{w}(\mathbf{s}), \mathbf{w}^*) Var^{-1}(\mathbf{w}^*) \mathbf{w}^* = \mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta}) \mathcal{C}^{*-1}(\boldsymbol{\theta}) \mathbf{w}^*, \quad (1.19)$$

where  $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}') = Cov(\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}')) = [Cov(w_l(\mathbf{s}), w_m(\mathbf{s}'))]_{l,m=1}^q$  is the *cross-covariance* matrix (see, e.g., Banerjee et al. (2004)),  $\mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}_1^*; \boldsymbol{\theta}), \dots, \Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}_m^*; \boldsymbol{\theta})]$  is  $q \times mq$  and  $\mathcal{C}^*(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$  is the  $mq \times mq$  dispersion matrix of  $\mathbf{w}^* = [\mathbf{w}(\mathbf{s}_i^*)]_{i=1}^m$ . Equation (1.19) shows  $\tilde{\mathbf{w}}(\mathbf{s})$  is a zero mean  $q \times 1$  multivariate predictive process with cross-covariance matrix given by  $\Gamma_{\tilde{\mathbf{w}}}(\mathbf{s}, \mathbf{s}') = \mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta}) \mathcal{C}^{*-1}(\boldsymbol{\theta}) \mathcal{C}(\mathbf{s}'; \boldsymbol{\theta})$ . This is especially important for the applications with multivariate spatial observations, where each location  $\mathbf{s}$  yields observations on  $q$  dependent variables given by a  $q \times 1$  vector  $\mathbf{Y}(\mathbf{s}) = [Y_l(\mathbf{s})]_{l=1}^q$ . For each  $Y_l(\mathbf{s})$ , we also observe a  $p_l \times 1$  vector of regressors  $\mathbf{x}_l(\mathbf{s})$ . Thus, for each location we have  $q$  univariate spatial regression equations which can be combined into the following multivariate regression model:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}^T(\mathbf{s}) \boldsymbol{\beta} + \mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \quad (1.20)$$

where  $\mathbf{X}^T(\mathbf{s})$  is a  $q \times p$  matrix ( $p = \sum_{l=1}^q p_l$ ) having a block diagonal structure with its  $l$ -th diagonal being the  $1 \times p_l$  vector  $\mathbf{x}_l^T(\mathbf{s})$ . Note that  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)^T$  is a  $p \times 1$  vector of regression coefficients with  $\boldsymbol{\beta}_l$  being the  $p_l \times 1$  vector of regression coefficients corresponding to  $\mathbf{x}_l^T(\mathbf{s})$ . Likelihood evaluation from (1.20) that involves  $nq \times nq$  matrices can be reduced to  $mq \times mq$  matrices by simply replacing  $\mathbf{w}(\mathbf{s})$  in (1.20) by  $\tilde{\mathbf{w}}(\mathbf{s})$ .

Further computational gains in computing  $\mathcal{C}^{*-1}(\boldsymbol{\theta})$  can be achieved by adopting ‘‘coregionalization’’ methods (Wackernagel (2003), Gelfand et al. (2004b) and Banerjee et al.

(2007)) that model  $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}') = A(\mathbf{s}) \text{Diag}[\rho_l(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})]_{l=1}^q A^T(\mathbf{s}')$ , where each  $\rho_l(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi})$  is a univariate correlation function satisfying  $\rho_l(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi}) \rightarrow 1$  as  $\mathbf{s} \rightarrow \mathbf{s}'$ . Note that  $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}) = A(\mathbf{s})A^T(\mathbf{s})$ , hence  $A(\mathbf{s}) = \Gamma_{\mathbf{w}}^{1/2}(\mathbf{s}, \mathbf{s})$  can be taken as any square-root of  $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s})$ . Often we assume  $A(\mathbf{s}) = A$  and assign an inverse-Wishart prior on  $AA^T$  with  $A$  a computationally efficient square-root (e.g., Cholesky or spectral). It now easily follows that  $\mathcal{C}^*(\boldsymbol{\theta}) = (I_m \otimes A)\Sigma^*(\boldsymbol{\theta})(I_m \otimes A^T)$ , where  $\Sigma^*(\boldsymbol{\theta})$  is an  $mq \times mq$  matrix partitioned into  $q \times q$  blocks, whose  $(i, j)$ -th block is the diagonal matrix  $\text{Diag}[\rho_l(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{l=1}^q$ . This yields a sparse structure and can be computed efficiently using specialized sparse matrix algorithms. Alternatively, we can write  $\Sigma^*$  as an orthogonally transformed matrix of  $m \times m$  block diagonal matrix,  $P^T[\oplus_{l=1}^q [\rho_l(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m]P$ , where  $\oplus$  is the block diagonal operator and  $P$  is a permutation (hence orthogonal) matrix. Since  $P^{-1} = P^T$ , we need to invert  $q$   $m \times m$  symmetric correlation matrices rather than a single  $qm \times qm$  matrix. Constructing the  $nq \times mq$  matrix  $\tilde{\Sigma} = [\text{Diag}[\rho_l(\mathbf{s}_i, \mathbf{s}_j^*)]_{l=1}^q]_{i,j=1}^{n,m}$ , we further have

$$\text{Var}(\tilde{\mathbf{w}}) = \mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) = (I_n \otimes A)\tilde{\Sigma}(\boldsymbol{\theta})\Sigma^{*-1}(\boldsymbol{\theta})\tilde{\Sigma}^T(\boldsymbol{\theta})(I_m \otimes A^T), \quad (1.21)$$

where the Kronecker structures and sparse matrices render easier computations.

# Chapter 2

## Hierarchical Modeling for Extreme Values Observed over Space and Time

### 2.1 Introduction

Extreme value analysis finds wide application in areas such as environmental science (e.g., Lou Thompson et al. (2001)), financial strategy of risk management (e.g., Dahan and Mendelson (2001)) and biomedical data processing (Roberts (2000)). In this article, we focus on climate extremes, in particular on precipitation events. Our underlying driver is the challenging ecological problem of trying to characterize the effect of changes in climates on the distribution and abundance of species. Particularly, for plants, it is suggested that extreme climate events, such as drought, heavy rainfall and very high or low temperatures, might be significant factors in explaining plant performance. In fact, it is plausible that trends in climate extremes are more informative with regard to survival, growth, reproductivity, etc., than trends in the mean climate.

The motivating data set here is derived from precipitation surfaces in the Cape Floristic Region (CFR) in South Africa. Figure 2.1 shows a map of the CFR in South Africa. The CFR is the smallest but, arguably, the richest of the world's six floristic kingdoms encompassing a region of roughly 90,000  $km^2$  in southwestern South Africa. The region is highly diverse and endemic across spatial scales; it includes about 9000 plant species, 69% of which are found nowhere else. The daily precipitation surfaces we employ arise via interpolation to grid cells at 10km resolution based on records reported by up to 3000 stations across South Africa during the period from 1950-1999. Figure 2.2 displays the

derived surface of annual maxima of daily rainfalls for 1332 grid cells in 1999. In fact, we have fifty such surfaces dating back to 1950. The high resolution grid-aggregated rainfalls are derived from a conditional interpolation technique of Hewitson and Crane (2005) which is especially tailored for precipitation data. The technique uses a distance-decay weighting function, modified by the station-specific relation of a station to its surrounding areas. It is widely used in the meteorology community and we make no attempt here to evaluate its performance; rather, we take the output as our rainfall data.

It is generally accepted that there are, primarily, two distinct patterns generating precipitation in the CFR. There are fronts from the Atlantic Ocean affecting the western part of the region with, typically, more hitting the southwest (Capetown and vicinity), fewer hitting the northwest. There tends to be greater interannual variability in the northwest with most of the rainfall occurring in the two to three winter months. Similarly, most of the rainfall around Capetown occurs in the winter months. As we move east, rainfall tends to arise from Indian Ocean storms and humidity which can come at any time but are more likely in the summer. Finally, in the northeastern part of the region, spring and fall rain occurs and there is little seasonality in the rainfall. Figure 2.3a shows the week of maximum precipitation over all locations and all years, which indicates great spatial and interannual variability in the extreme rainfall occurrence. Figure 2.3b shows this for the western part of the CFR and Figure 2.3c for the eastern part. They support the foregoing description; maximum rainfall is not confined to a specific part of the year. Finally, the climate for a grid cell is expected to depend on altitude so we introduce elevation as a covariate in the modeling. However, we find little effect on precipitation (but would anticipate a stronger effect if we were studying temperature surfaces).

By now there is an enormous literature on the modeling of extremes. At present, the standard approach utilizes Generalized Extreme Value (GEV) distribution families.

Alternatively, daily precipitation exceedances for a given threshold are modelled with Generalized Pareto Distribution (GPD). The book by Coles (2001) provides an introduction and bibliography through 2000. More recent work in the area of rainfall extremes includes DURMAN et al. (2001), Kharin and Zwiers (2005), Huerta and Sansó (2007) and Cooley et al. (2007).

The contribution of this Chapter is to develop models for rainfall that reflect dependence in space and in time. In particular, the GEV is characterized by a location, a scale and a shape parameter. Conceptually, these could all vary in space and time and they could be mutually dependent. Through exploratory data analysis we clarify the nature of this variation and then introduce it appropriately into a hierarchical model. In fact, we offer several such models, fitting them and comparing them. We present interpretation of the data analysis including the nature of spatial association and the nature of trend at grid cell resolution over the fifty year period from 1950-1999. In fact, we also hold out the last year of data to implement forecasting with validation studied through the hold out data. As a result, we extend the recent work of Cooley *et al.* (2007) to introduce a dynamic aspect, in particular, to model the location spatio-temporally, to introduce spatial dependence between scale and location and, through Markov random fields, to handle a considerably larger data set.

The format of the Chapter is as follows. In Section 2.2 we present exploratory analysis to enable us to connect the grid level models in space and time. In Section 2.3 we offer various hierarchical Bayesian models which provide appropriate connections based on the inference about latent parameter structure in Section 2.2. Model implementation, employing MCMC methods and model comparison for the annual rainfall data in the CFR are illustrated in Section 2.4. In Section 2.5 we conclude with a brief summary and suggestions for future work.

## 2.2 Exploratory analysis

As reviewed in the introduction Chapter, The GEV distribution is viewed as an approximate distribution to model the maximum of a sufficiently long sequence of random variables.

$$G(y; \mu, \sigma, \xi) = \exp \left\{ - \left[ 1 + \xi \left( \frac{y - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \quad (2.1)$$

for  $y : 1 + \xi(y - \mu)/\sigma > 0$ . In (1),  $\mu \in \mathbb{R}$  is the location parameter,  $\sigma > 0$  is the scale parameter and  $\xi \in \mathbb{R}$  is the shape parameter.

Here, we will assign daily precipitation into annual blocks and then assume the maxima are conditionally independent (but not identically distributed) across years given their respective, parametrically modelled,  $\mu$ ,  $\sigma$  and  $\xi$ . That is, each maximum is assumed to follow its own GEV with  $(\mu, \sigma, \xi)$ .

A further remark is appropriate here. The annual blocks provide roughly 365 daily precipitation measurements. Many regions in the CFR are essentially deserts, resulting in a very high proportion of “zeroes”. (In some grid cells we see 90% zeroes.) However, it can be claimed that the GEV is still an appropriate model for the maximum in this case. The argument is merely to envision the daily observations as  $Y_t = Y_t^* 1(Y_t^* > 0)$  where  $Y_t^*$  has support  $\mathbb{R}^1$ . Then, if  $\max Y_t^*$  is assumed to follow a GEV, since  $\max Y_t = \max Y_t^*$  with very high probability, we can use a GEV model for  $\max Y_t$ .

We introduce the GEV distribution as a first stage model for annual precipitation maxima, specifying  $\mu$ ,  $\sigma$ , and  $\xi$  at the second stage to reflect underlying spatio-temporal structure. In particular, let  $Y_{i,t}$  denote the annual maximum of daily rainfall at location  $i$  in year  $t$ . We assume the  $Y_{i,t}$  follow a GEV distribution with parameters  $\mu_{i,t}$ ,  $\sigma_{i,t}$  and  $\xi_{i,t}$ , respectively. Again, we assume the  $Y_{i,t}$  are conditionally independent given their  $\mu$ 's,  $\sigma$ 's and  $\xi$ 's. Attention focuses on specification of the model for  $\mu_{i,t}$ ,  $\sigma_{i,t}$  and  $\xi_{i,t}$ .

To facilitate modeling the  $\mu_{i,t}$ ,  $\sigma_{i,t}$  and  $\xi_{i,t}$ , we conduct some exploratory analysis with

two objectives - learning about temporal trend in these parameters and learning about spatial dependence in these parameters. First, it is known that estimation of  $\xi$  is challenging in GEV models (Cooley *et al.*, 2007) and it is unlikely that we will be able to reliably discern space-time pattern in a set of  $\xi_{i,t}$ . Furthermore, model fitting with spatially or temporally structured  $\xi_{i,t}$  will be very difficult. So, we assume  $\xi$  is unknown but constant and focus on learning about  $\mu_{i,t}$  and  $\sigma_{i,t}$ .

Next, we illuminate the need for spatial modeling. At each grid cell, we fitted MLE's for  $\mu_i$  and  $\sigma_i$ , treating the 50 measurements in time as independent, using the S-Plus package *ismev* (Coles, 2001). We then computed customary residuals and studied the dependence in these residuals between pairs of adjacent grid cells. Since these residuals are roughly from a  $\text{GEV}(0, 1, \xi)$ , moments need not exist so it is not appropriate to compute product moment correlations. Instead we computed Spearman rank correlations. In fact, there are a very large number of pairs of adjacent grid cells, hence a very large number of correlations that can be computed. We created a histogram of these correlations (which is not presented in the interest of space) finding that the median of these correlations is roughly .6. Evidently there is strong residual spatial dependence, justifying the spatial random effects models we develop in Section 2.3.

Returning to  $\mu_{i,t}$  and  $\sigma_{i,t}$ , we initially consider the forms

$$\mu_{i,t} = \mu_i + \alpha_{i,\mu}t \tag{2.2}$$

$$\log(\sigma_{i,t}) = \log(\sigma_i) + \alpha_{i,\sigma}t \tag{2.3}$$

In (3) and (4),  $\mu_i$  and  $\log\sigma_i$  can be viewed as baselines at grid cell  $i$ . The  $\alpha_{i,\mu}$  and  $\alpha_{i,\sigma}$  denote the site-level *trend* coefficients for the location and scale parameters. (Nonzero) slopes inform about temporal dependence and the intercepts inform about spatial dependence. In exploratory mode, we assume independence across sites and obtain maximum likelihood estimators for these parameters along with  $\xi$ , modifying open source code for the S-Plus

package `isnev` (Coles (2001)).

In Figure 2.4 we provide two maps showing the  $\mu_i$  and  $\log\sigma_i$ . We clearly see evidence of spatial pattern in both, leading us to model the locations and scales with spatial structure. In fact, using the grid cells themselves with, say usual first order neighbor proximity, we can run Moran's I and Geary's C randomization tests for spatial association (Banerjee *et al.*, 2004) for the  $\mu_i$  and the  $\log\sigma_i$ . For the  $\mu_i$  we obtain  $I = 0.82$  ( $p < 0.01$ ) and  $C = 0.14$  ( $p < 0.01$ ). For the  $\log\sigma_i$  we obtain  $I = 0.74$  ( $p < 0.01$ ) and  $C = 0.23$  ( $p < 0.01$ ). So in both cases, there seems to be evidence of significant spatial dependence. The implication is that extreme rainfall distributions are anticipated to be similar for sites near to each other (though nothing in this analysis says that the day of occurrence of an extreme rainfall at one site coincides with the day of extreme rainfall at an adjacent site).

There is also evidence of temporal trend in the location parameters; roughly 95% of the interval estimates for  $\alpha_{i,\mu}$  are significant. There is much weaker evidence of temporal trend in the scale parameters; far fewer of the  $\alpha_{i,\sigma}$  are significant. Moreover, fitting models with both space and time specification in the scale parameters is difficult and, in any event, we view our modeling as primarily illustrative. Therefore, in the formal space time models proposed in Section 2.3, we assume there is temporal dependence for the  $\mu$ 's but not for the  $\sigma$ 's. Of course, analogous versions of models A - D below could be offered for the  $\log\sigma$ 's.

We might anticipate dependence between  $\mu_i$  and  $\log\sigma_i$ , i.e., often the scale varies with the center. If so, we would employ a bivariate spatial model. The scatter plot in Figure 2.5 for each pair of  $\mu_i$  and  $\log\sigma_i$  shows evidence of high correlation between these two parameters, indeed suggesting the need for modeling dependence between them. In Section 2.3 we build formal multivariate spatial models to account for the findings above.

## 2.3 Formal hierarchical space time models

Let  $Y_{i,t}$  denote the annual maximum precipitation at grid cell  $i$  at time  $t$  with conditionally independent first stage specification given by  $y_{i,t} | (\mu_{i,t}, \sigma_{i,t}, \xi_{i,t}) \sim GEV(\mu_{i,t}, \sigma_{i,t}, \xi_{i,t})$ . The conditional independence assumption is interpreted as interest in smoothing the precipitation surfaces around which the interpolated data is centered rather than smoothing the data surface itself. As a formal assumption, it is defensible in time since the annual maxima at a site likely occur with sufficient time between them to be assumed independent. In space, we would expect small scale dependence in the data at a given time. However, as noted in the Introduction, our observations are assigned to grid cells at 10 km resolution. The exploratory analysis in the previous section motivates modeling residual spatial dependence at this scale. However, with areal unit data at such scale, we have no tools to learn about finer scale dependence.

Following Section 2.2, we assume  $\xi_{i,t} = \xi$  for all  $i$  and  $t$ , and  $\sigma_{i,t} = \sigma_i$  for all  $t$ , where  $\sigma_i$  will be modelled using spatial random effects. Modeling interest therefore focuses on the  $\mu_{i,t}$ . But, in addition, we want the  $\mu_{i,t}$  and  $\sigma_i$  to be co-varying at the same site. This requires the introduction of an association model for a collection of spatially co-varying parameters over a collection of grid cells. Gelfand and Vounatsou (2003) and Jin et al. (2005) developed models for multivariate spatial random effects with areal data. We also do this but adopt the coregionalization method (Gelfand et al. (2004b)) as discussed below to offer a constructive specification of the joint distribution through transformation of simpler conditional and marginal models. The gridding in the precipitation data suggests the use of conditionally autoregressive (CAR) models; they also greatly reduce the computational burden in fitting our high hierarchical models.

### 2.3.1 Modeling the $\mu_{i,t}$

As a result, we propose the specification,  $p(\mu_{i,t}|\boldsymbol{\beta}, W_{i,t}, \tau^2) = N(\mathbf{X}'_i\boldsymbol{\beta} + W_{i,t}, \tau^2)$ . Again,  $\mu_{i,t}$  is the location parameters at location  $i$  at year  $t$ ,  $i = 1, 2, \dots, S$  and  $t = 1, 2, \dots, T$ .  $\mathbf{X}_i$  is the site-specific vector of potential explanatory variables for extreme rainfall, such as geographic coordinates using a trend surface, altitude of the sites and local temperature, with the corresponding coefficient vector  $\boldsymbol{\beta}$ .  $W_{i,t}$  is a spatial-temporal random effect.

Possibilities for modeling  $W_{i,t}$  include:

- An additive form:
  - **Model A:**  $W_{i,t} = \psi_i + \delta_t$ ,  $\delta_t = \phi\delta_{t-1} + \omega_t$ , where  $\omega_t \sim N(0, W_0^2)$  *i.i.d*
- A linear temporal component with spatial random effects:
  - **Model B:**  $W_{i,t} = \psi_i + \rho(t - t_0)$
  - **Model C:**  $W_{i,t} = \psi_i + (\rho + \rho_i)(t - t_0)$
- A multiplicative form in space and time:
  - **Model D:**  $W_{i,t} = \psi_i\delta_t$ ,  $\delta_t = \phi\delta_{t-1} + \omega_t$ , where  $\omega_t \sim N(0, W_0^2)$  *i.i.d*

The additive form in Model A might appear to over-simplify spatial temporal structure. However, the data may not be rich enough to find space-time interaction in the  $\mu_{i,t}$ . Model B and Model C provide evaluations of temporal trends in terms of global and local assessments respectively. The coefficient  $\rho + \rho_i$  in Model C represents the spatial trend in location parameters, where  $\rho$  could be interpreted as the global change level in the CFR per year. Finally, Model D provides a multiplicative representation of  $W_{i,t}$ , similar in spirit to the recent work of Huerta and Sansó (2007). Models A and D yield special cases of a dynamic linear model (West and Harrison (1997)). More general versions can be built in the  $W_{i,t}$ ,

e.g.,  $W_{i,t} = \phi W_{i,t-1} + \eta_{i,t}$  where the  $\eta_{i,t}$  are modeled as independent over  $t$  but dependent across  $i$ . (See, e.g., Banerjee et al. (2004), Chapter 8 in this regard.) We do not follow this path here since, from Section 2.2, we want to model the dependence between location and scale parameters in the GEV model. In models A, B, and D, we do this by specifying  $\log\sigma_i$  and  $\psi_i$  to be dependent. In model C, we specify  $\log\sigma_i, \psi_i$  and  $\rho_i$  to be dependent.

The nugget term  $\tau^2\epsilon_{i,t}$  is introduced for computational convenience. By fixing a small value for variance  $\tau^2$  we enjoy the benefit of a Gaussian framework in sampling the parameters in the second stage specification with conditionally independent  $\mu_{i,t}$ . Of course, the  $Y_{i,t}$ 's continue to be conditionally independent as above. Details of the sampling schemes for MCMC model fitting are provided in the Appendix.

Again, we will have grid cell level scale parameters,  $\sigma_i$ . Following the findings of Section 2.2, we need to jointly model  $\log\sigma_i$  and the  $\psi_i$  (regardless of which of the above models we choose). This is the topic of the next subsection.

### 2.3.2 The Coregionalized CAR model

Recall the univariate CAR model (Besag (1974)), frequently used to model spatial random effects (Banerjee et al. (2004)). Suppose  $\mathbf{V} = (V_1, V_2, \dots, V_S)'$  is a vector of spatial random effects which is defined at areal sites from 1 to  $S$ . The joint distribution of  $\mathbf{V}$  is defined through conditional Gaussian specifications at each site,  $p(V_i|V_j, j \neq i) = N(\sum_j w_{i,j}/w_{i+} V_j, \lambda/w_{i+})$  where  $\lambda$  is a *scale* parameter. The CAR is an improper distribution. We use the usual neighbor-based proximities,  $w_{i,j} = 1$  if  $i$  and  $j$  share any common boundary,  $w_{i,j} = 0$  otherwise.

When we have multivariate spatial random effects over the same region that have dependence among themselves as well as spatial dependence across sites, a multivariate CAR model is then desired. Gelfand and Vounatsou (2003) and Jin et al. (2005) developed mod-

els for multivariate areal data by extending multivariate CAR ideas building on the work of Mardia (1988). Here, we propose a coregionalization model in the spirit of Gelfand et al. (2004b).

Let  $\mathbf{U}_i = (U_{1,i}, U_{2,i}, \dots, U_{P,i})'$  denote the  $P \times 1$  vector of spatial random effects for the study region. Introducing a lower triangular transformation matrix  $\mathbf{A}$  with elements  $a_{i,j}$ , let  $\mathbf{U}_i = \mathbf{A}\mathbf{V}_i$ , where  $\mathbf{V}_i = (V_{1,i}, V_{2,i}, \dots, V_{P,i})'$ . Denote  $\mathbf{V}^j = (V_{j,1}, V_{j,2}, \dots, V_{j,S})$ , for  $j = 1, \dots, P$ . For the  $\mathbf{V}^j$ , we assume independent univariate CAR models as we mentioned above. The CAR models can have scale parameter 1 because the entries in  $\mathbf{A}$  provide the scaling. We further assume that proximity matrix  $W$  remains constant for the  $P$  univariate CAR models. Also, each  $\mathbf{V}^j$  has an improper distribution, hence the distribution of the  $\mathbf{U}^j$  is improper. However, as with customary univariate CAR priors, we introduce centering (on the fly, i.e., after each model-fitting iteration, Besag et al. (1995)). This converts the improper CAR specification to a proper one while still enabling us to take advantage of the convenient full conditional distributions associated with the CAR model. Now, with proper priors, we are ensured that the posterior will be proper.

We work with the log transformation of  $\sigma_i$ , i.e.,  $\sigma_i = \sigma_0 \exp \lambda_i$ .  $\lambda_i$  is now centered at 0. The coregionalization CAR model is used for  $\lambda_i$  along with  $\psi$  in Model A, B and D, i.e.,  $(\lambda_i, \psi_i)' = \mathbf{A}(V_{1,i}, V_{2,i})'$ , where  $\mathbf{A} = \begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{pmatrix}$  and  $\mathbf{V}_1$  and  $\mathbf{V}_2$  are two independent univariate CAR models.

For Model C,  $\rho_i$  in the temporal linear component is spatially varying. So here we extend the multivariate CAR model to incorporate three spatial components, leading to  $(\lambda_i, \psi_i, \rho_i)' = \mathbf{A}(V_{1,i}, V_{2,i}, V_{3,i})'$ , where now  $\mathbf{A}$  is a  $3 \times 3$  lower triangular matrix.

### 2.3.3 Bayesian implementation

Posterior inference for the model parameters is implemented by model fitting with Gibbs samplers (Gelfand and Smith (1990)) and Metropolis Hasting updating (Gelman (2004)). Details of the full conditional distribution for each parameter are given in the Appendix.

In our model, a vague normal prior is assigned to the shape parameter,  $\xi$ . We assign inverse gamma priors for positive value  $\sigma_0$ . Coefficients  $\beta$  and  $\rho$  have normal priors centered at the exploratory estimates with large variances. In Model A and Model D, we follow the dynamic linear model prior setting to update  $\delta_1, \delta_2, \dots, \delta_T$ ,  $\phi$  and  $W_0$ . For the coregionalization matrix  $\mathbf{A}$ , we assign truncated normal priors with positive value support for the diagonal entries, and normal priors for the other entries. Altogether, our priors are very weak suggesting little concern with regard to sensitivity analysis. However, we did implement some sensitivity study, primarily on the various prior uncertainties, which confirmed this. Details are available from the authors.

Under the conditional independence assumptions in the first stage settings,  $\mu_{i,t}$  are sampled independently for each pair of  $(i, t)$ . Given  $\mu_{i,t}$ , we can directly sample  $\beta$ ,  $\rho$ ,  $\mathbf{V}_2$  and off-diagonal entries in  $\mathbf{A}$  by adopting conjugate normal priors. Parameters without closed form full conditional distributions are updated using Metropolis-Hastings.  $\xi$ ,  $\mu_{i,t}$  and  $\mathbf{V}_1$  are sampled by the random walk Metropolis-Hastings with Gaussian proposals. The proposal distributions for  $\sigma_0$  and diagonal entries in  $\mathbf{A}$  are truncated normals centered at the current samples.

## 2.4 Model comparison and results

Again, the data set consists of annual maximum records at 1078 grid cells from 1950 to 1999. We fit the four hierarchical space time models with latitude/longitude and elevation

at the cell centroid entering linearly as covariates. Models are fitted using two parallel chains. Model D takes the longest to run, completing roughly 3 iterations per minute using R code with dual 2.8 GHz Xeon CPUs and 12GB memory. We ran 10000 iterations to collect posterior samples after a burn in period of 2000 iterations, thinning using every fifth iteration. Trace plots of parameters indicate good convergence of the respective marginal distributions.

In the models proposed in Section 4, temporal evolution in the extreme rainfalls is taken into account in the model for  $W_{i,t}$ . More precisely, each of the models enables prediction for any grid cell for any year. Annual maximum rainfalls in year  $T + 1$  could be simply obtained by updating samples from:

$$f(\mathbf{Y}_{T+1}|\mathbf{Y}_1, \dots, \mathbf{Y}_T, \mathbf{X}) = \int f(\mathbf{Y}_{T+1}|\boldsymbol{\theta}, \mathbf{X})f(\boldsymbol{\theta}|\mathbf{Y}_1, \dots, \mathbf{Y}_T, \mathbf{X})d\boldsymbol{\theta} \quad (2.4)$$

In fact, we held out the annual maximum rainfalls in 1999 for validation purposes, in order to compare models in terms of predictive performance. Posterior medians are adopted as the point estimates of the predicted annual maxima in 1999 because of the skewness of the predictive distribution for  $\mathbf{Y}_{T+1}$ . We check the predictive performance by computing the averaged absolute predictive errors (*AAPE*) for each model. Given the true value of  $Y_{i,1999}$  in the hold out data set, *AAPE* is computed by:

$$AAPE = \frac{1}{S} \sum_{i=1}^S |\hat{Y}_{i,1999} - Y_{i,1999}| \quad (2.5)$$

where  $\hat{Y}_{i,1999}$  is the median of the posterior samples  $\{\hat{Y}_{i,1999}^{(b)}\}$  for  $b = 1, 2, \dots, B$  with the averaged absolute deviance (*AAD*) from the  $\hat{Y}_{i,1999}$  estimated by:

$$AAD = \frac{1}{BS} \sum_{b=1}^B \sum_{i=1}^S |\hat{Y}_{i,1999}^{(b)} - \hat{Y}_{i,1999}| \quad (2.6)$$

A second model comparison is to study, for each model, the proportion of the true annual maximum rainfalls in 1999 which lie in the associated estimated 95% credible intervals,

denoted by  $\hat{r}$ , is computed as in Table 2.1. A third model selection criterion which is easily calculated from the posterior samples is the deviance information criterion (DIC) (Spiegelhalter et al. (2002)).

Table 2.1 summarizes the DIC score,  $\hat{r}$ , AAPE and the associated AAD for each predictive posterior samples by each models. Model A wins among the four models under the DIC criterion. However, Model C has the lowest prediction error in terms of AAPE and AAD and is second in DIC. In addition, the  $\hat{r}$  for Model C is quite close to 0.95, as desired. As a result, we summarize results based on Model C.

Table 5.5 provides the posterior means for the parameters and the corresponding 95% credible intervals under Model C. The resulting estimates for the coefficients  $\beta_{lat}$ ,  $\beta_{lon}$  and  $\beta_{elev}$  in the regression part are insignificant; over the CFR we find no linear trend in these variables in explaining extreme rainfalls. This finding agrees with the results in Coles and Tawn (1996) for a different rainfall study. The posterior mean of the  $\xi$  takes value of 0.122 with 95% credible interval (0.112, 0.129), indicating the GEV distributions of annual maximum rainfalls in CFR have heavy upper tails. The estimate of  $\rho$  in Table 5.5 is nearly significant and suggests that, over the whole CFR, annual maximum rainfall is increasing, on average, by 0.015mm per year over the past 50 years.

Model C allows for a further examination of the site-level trends. Figure 2.6 displays the posterior means of the site-level trends and the  $\sigma_i$ 's. The magnitudes of the estimates are represented through the grey scale plot and suggest spatial pattern in the  $\rho_i$ 's and the  $\sigma_i$ 's. Specifically, despite an overall tendency towards increasing extreme rainfalls in the CFR, some small areas show tendency towards decreasing extreme rainfalls.

Return levels for the occurrence of extreme events are of practical interest in climate studies. Given posterior samples of  $\mu_{i,t}$ ,  $\sigma_i$  and  $\xi$  under the model, posterior samples of return levels with return period 25 years are derived by equation 1.2. A map of estimated

$z_{1/25}$  in 1998 is presented in Figure 2.7a. Figure 2.7b compares return level  $z_{1/25,1998}$  and  $z_{1/25,1950}$  obtained from Model C. Though, under our models, the return levels change with time (so they are not useful as actual expected return times), it is informative to see how they vary over time. We see slight temporal changes in  $z_{1/25}$  over the past 50 years. In the southwestern part of the CFR, return levels  $z_{1/25}$  decreased at least 2% from 1950 to 1998, suggesting a decrease in the intensity of heavy rainfalls in CFR over the past 50 years. In contrast, more than half of the regions in CFR have increasing return levels  $z_{1/25}$ , suggesting the shift of the distributions of annual maxima rainfall towards larger heavy rainfalls. Finally, the posterior samples provide a natural way to evaluate the corresponding uncertainty of return level estimates as shown in Figure 2.8. The lower and upper .025 quantiles vary considerably over the region and the lengths of the associated interval estimates do as well.

## 2.5 Discussion and future work

We have presented flexible hierarchical models for the dynamic and spatial change in annual maximum rainfalls collected at more than 1300 areal regions over 50 years. Our hierarchical models are based on the GEV distribution to describe the asymptotic behavior of maxima taken from a time series of daily records. We consider space and time-varying location parameters in the GEV distributions. In addition, we assume spatially varying scale parameters. A coregionalized CAR model is then introduced to capture the underlying spatial dependence between these parameters at site level. Inference for the space-time return level becomes straightforward under the MCMC model fitting approach.

Various extensions for this work are possible. Besides annual maxima of rainfalls, we have interest in similar modeling for annual temperature extremes. A further, and more demanding, challenge will take us to bivariate space-time modeling for precipitation and

temperature extremes (see, e.g., Heffernan and Tawn (2004)).

In this model, the interpolation technique which yields the gridded data is not a point to point interpolation, but rather generates grid cell area integrals. The interpolation technique fits a surface to best estimate the area integral and, thus, inherently suppresses the peaks. Therefore, it is expected that interpolation of extremes will underestimate true extremes over the same area, i.e., the maximum of an average is at most the average of maximums. It is an open challenge to account for this effect in modeling interpolated extreme values.

## Appendix

### Monte Carlo Sampling Procedure

The hierarchical models we proposed are implemented by MCMC algorithm. We draw samples of parameters from their full conditional distribution respectively. We illustrate the MCMC sampling procedures for Model A and C below.

Let  $\boldsymbol{\beta} = (\beta_0, \beta_{lat}, \beta_{lon})$ ,  $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_T)$ ,  $\boldsymbol{\mu} = (\mu_{1,1}, \mu_{1,2}, \dots, \mu_{1,T}, \mu_{2,1}, \dots, \mu_{S,T})'$ ,  $\boldsymbol{\mu}_t = (\mu_{1,t}, \mu_{2,t}, \dots, \mu_{S,t})'$ ,  $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_T)$ ,  $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_T)'$ ,  $\boldsymbol{t} = (1, 2, \dots, T)'$ ,  $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_S)'$ ,  $\boldsymbol{X}_i = (1, Lat_i, Lon_i, Elev_i)'$ ,  $\boldsymbol{W} = (W_{1,1}, W_{1,2}, \dots, W_{1,T}, W_{2,1}, \dots, W_{S,T})'$ . Let  $g(\cdot)$  be the probability density function of the GEV distribution. Let  $N(\cdot)$  denotes the probability density function of normal distribution. Below we use  $\boldsymbol{\theta}_-$  generically to denote parameters other than the target parameter.

#### MCMC for the implementation of Model A

- **Updating  $\boldsymbol{\mu}$ :** Sample  $\mu_{i,t}$  independently via Metropolis-Hasting algorithm for each

i and t.

$$f(\mu_{i,t}|\boldsymbol{\theta}_-) \propto g(Y_{i,t}; \mu_{i,t}, \sigma_i, \xi)N(\mu_{i,t}; \mathbf{X}'_i\boldsymbol{\beta} + W_{i,t}, \tau^2)$$

- **Updating  $\xi$ :** The Metropolis-Hastings step can be applied to sampling  $\xi$  based on its full conditional distribution.

$$f(\xi|\boldsymbol{\theta}_-) \propto \prod_{i=1}^S \prod_{t=1}^T g(Y_{i,t}; \mu_{i,t}, \sigma_i, \xi)N(\xi; \mu_\xi^0, \sigma_\xi^0{}^2)$$

- **Updating  $\boldsymbol{\beta}$ :** Sample  $\boldsymbol{\beta}$  directly from the full conditional distribution, which is

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i\boldsymbol{\beta} + W_{i,t}, \tau^2)N(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta^0, \Sigma_\beta^0) \\ &\sim N(Bb, B) \end{aligned}$$

where  $B^{-1} = \mathbf{X}^T \mathbf{X} + \Sigma_\beta^0$ ,  $b = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{W}) + \Sigma_\beta^0 \boldsymbol{\mu}_\beta^0$

- **Updating  $\sigma_0$ :** Sample  $\sigma_0$  from its full conditional distribution

$$f(\sigma_0|\boldsymbol{\theta}_-) \propto \prod_{i=1}^S \prod_{t=1}^T g(Y_{i,t}; \mu_{i,t}, \sigma_0 \exp(a_{11} V_{1,i}), \xi)IG(\sigma_0; 2, \beta_{\sigma_0}^0)$$

Update  $\boldsymbol{\sigma} = \sigma_0 \exp(a_{11} \mathbf{V}_1)$

- **Updating matrix  $\mathbf{A}$ :**

$\mathbf{A}$  is a  $2 \times 2$  lower triangular matrix. Sample  $a_{11}$  via Metropolis-Hastings step.  $a_{12}$  and  $a_{22}$  are updated from their posterior normal distributions by adopting normal priors.

$$\begin{aligned} f(a_{11}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T g(Y_{i,t}; \mu_{i,t}, \sigma_0 \exp(a_{11} V_{1,i}), \xi)N_+(a_{11}; \mu_{11}^0, \sigma_{11}^0{}^2) \\ f(a_{12}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i\boldsymbol{\beta} + a_{12}V_{1,i} + a_{22}V_{2,i} + \delta_t, \tau^2)N(a_{12}; \mu_{12}^0, \sigma_{12}^0{}^2) \\ f(a_{22}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i\boldsymbol{\beta} + a_{12}V_{1,i} + a_{22}V_{2,i} + \delta_t, \tau^2)N_+(a_{22}; \mu_{22}^0, \sigma_{22}^0{}^2) \end{aligned}$$

- **Updating  $\delta$ :**

Conditional on  $\mu_{i,t}$ ,  $\beta$  and  $\psi$ ,  $\mu_t - \mathbf{X}\beta - \psi = \delta_t \mathbf{1} + \epsilon_t$ ;  $\epsilon_t \sim N(0, \tau^2 \mathbf{I})$ ,  $\delta_t = \phi \delta_{t-1} + \omega_t$ ;  $\omega \sim N(0, W_0)$ . This is a specific dynamic linear model  $DLM(Y; F, G, V, W)$  (West and Harrison (1997); Huerta and Sansó (2007)) with known observation covariance matrix  $\tau^2 \mathbf{I}$ . Sample the state vector  $\delta$  using the Forward Filtering Backward Sampling algorithm (FFBS). Forward in time, sequentially sample  $\delta_t$  from  $\delta_t | (Y_1, \dots, Y_t, W_0, \phi)$  for  $t = 1, 2, \dots, T$ . Backwards in time, smooth the samples of  $\delta_t$  by updating  $\delta_t | (\delta_{t+1}, (y_1, \dots, y_T), W_0, \phi)$  for each time step. The evolution variance  $W_0$  can be directly sampled from an Inverse-Wishart distribution if we adopt  $IG(a_{W_0}^0, \beta_{W_0}^0)$  as priors for  $W_0$ , i.e.,  $f(W_0 | \theta_-) \propto IG(a_{W_0}^0 + \frac{T}{2}, \beta_{W_0}^0 + \frac{\sum_{t=1}^T (\delta_t - \phi \delta_{t-1})^2}{2})$ , where  $\delta_0 = 0$ . Finally, sample  $\phi$  from its conditional distribution where we use normal conjugate prior, i.e.,  $f(\phi | \delta, W_0) \propto \sum_{t=2}^T N(\delta_t; \phi \delta_{t-1}, W_0) N(\phi; \mu_\phi^0, \sigma_\phi^2)$

- **Updating  $V_1$  by Metropolis-Hastings step:** The conditional distribution for  $V_{1,i}$  is :

$$f(V_{1,i} | V_{1,j}, j \neq i, V_{2,i}, \theta_-) \propto \prod_{t=1}^T g(Y_{i,t}; \mu_{i,t}, \sigma_0 \exp(a_{11} V_{1,i}), \xi) f(V_{1,i}; \mu_{v_{1i}}, \sigma_{v_{1i}}^2)$$

- **Updating  $V_2$ :** The full conditional distribution for  $V_2$  is a Gaussian. For each  $i=1, 2, \dots, S$ , update  $V_{2,i}$  directly from

$$f(V_{2,i} | V_{2,j}, j \neq i, V_{1,i}, \theta_-) \propto N(\mu_{i,t}; \mathbf{X}'_i \beta + a_{12} V_{1,i} + a_{22} V_{2,i} + \delta_t, \tau^2) N(\sum_j w_{ij} V_{2,j} / w_{i,+}, w_{i,+})$$

Finally, Update  $W_{i,t} = \psi_i + \delta_t$ .

**MCMC for Model C:** Follow the procedures of implementing Model (A) to update  $\mu$ ,  $\xi$ ,  $\beta$  and  $\sigma_0$ .

- **Updating  $\rho$ :** Sample  $\rho$  from its full conditional distribution

$$f(\rho | \theta_-) \propto \prod_{i=1}^S \prod_{t=1}^t N(\mu_{i,t}; \mathbf{X}'_i \beta + \psi_i + (\rho + \rho_i)(t - t_0), \tau^2) N(\rho; \mu_\rho^0, \sigma_\rho^2)$$

Update  $\delta = \rho(t - t_0)$

- **Updating  $\mathbf{A}$ :** For Model (C),  $\mathbf{A}$  is a  $3 \times 3$  lower triangular matrix. Sample  $a_{11}$  via Metropolis-Hasting step.  $a_{12}$ ,  $a_{22}$ ,  $a_{13}$ ,  $a_{23}$  and  $a_{33}$  are updated from their posterior normal distributions by adopting conjugate normal priors.

$$\begin{aligned}
f(a_{11}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T g(Y_{i,t}; \mu_{i,t}, \sigma_0 \exp(a_{11} V_{1,i}), \xi) N_+(a_{11}; \mu_{11}^0, \sigma_{11}^0{}^2) \\
f(a_{12}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + a_{12} V_{1,i} + a_{22} V_{2,i} + \delta_{i,t}, \tau^2) N(a_{12}; \mu_{12}^0, \sigma_{12}^0{}^2) \\
f(a_{22}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + a_{12} V_{1,i} + a_{22} V_{2,i} + \delta_{i,t}, \tau^2) N_+(a_{22}; \mu_{22}^0, \sigma_{22}^0{}^2) \\
f(a_{13}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + \psi_i + (\rho + \rho_i)(t - t_0), \tau^2) N(a_{13}; \mu_{13}^0, \sigma_{13}^0{}^2) \\
f(a_{23}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + \psi_i + (\rho + \rho_i)(t - t_0), \tau^2) N(a_{23}; \mu_{23}^0, \sigma_{23}^0{}^2) \\
f(a_{33}|\boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + \psi_i + (\rho + \rho_i)(t - t_0), \tau^2) N_+(a_{33}; \mu_{33}^0, \sigma_{33}^0{}^2)
\end{aligned}$$

- For  $i = 1, 2, \dots, S$  and  $t = 1, 2, \dots, T$ , update  $\psi_i = a_{12} \mathbf{V}_{1,i} + a_{22} \mathbf{V}_{2,i}$ ,  $\rho_i = a_{13} \mathbf{V}_{1,i} + a_{23} \mathbf{V}_{2,i} + a_{33} \mathbf{V}_{3,i}$ ,  $\sigma_i = \sigma_0 \exp(a_{11} V_{1,i})$ ,  $W_{i,t} = \psi_i + (\rho + \rho_i)(t - t_0)$ .
- **Updating  $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3$ :** Update  $\mathbf{V}_1$  by Metropolis-Hastings step. The conditional distribution for  $V_{1,i}$  is

$$f(V_{1,i}|V_{1,j}, j \neq i, V_{2,i}, \boldsymbol{\theta}_-) \propto \prod_{t=1}^T g(Y_{i,t}; \mu_{i,t}, \sigma_0 \exp(a_{11} V_{1,i}), \xi) N(V_{1,i}; \mu_{v_{1i}}, \sigma_{v_{1i}}^2)$$

The full conditional distribution for  $\mathbf{V}_2$  is a Gaussian distribution. For each  $i=1,2,\dots,S$ , update  $V_{2,i}$  directly from

$$\begin{aligned}
f(V_{2,i}|V_{2,j}, j \neq i, V_{1,i}, V_{3,i}, \boldsymbol{\theta}_-) &\propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + a_{12} V_{1,i} + a_{22} V_{2,i} \\
&+ (\rho + a_{13} V_{1,i} + a_{23} V_{2,i} + a_{33} V_{3,i})(t - t_0), \tau^2) N\left(\sum_{j=1}^S w_{i,j} V_{2,j} / w_{i,+}, w_{i,+}\right)
\end{aligned}$$

The full conditional distribution for  $\mathbf{V}_3$  is a Gaussian distribution. For each  $i=1,2,\dots,S$ ,

	$\bar{D}$	$p_D$	DIC	AAPE	AAD	$\hat{r}$
Model A	638212	7614	645826	84.5	110.3	0.985
Model B	639835	7332	647167	83.7	104.5	0.944
Model C	638845	7814	646659	77.9	103.3	0.945
Model D	653321	7802	661123	81.7	118.5	0.966

**Table 2.1:** DIC, AAPE and  $\hat{r}$  for Models (A) - (D).

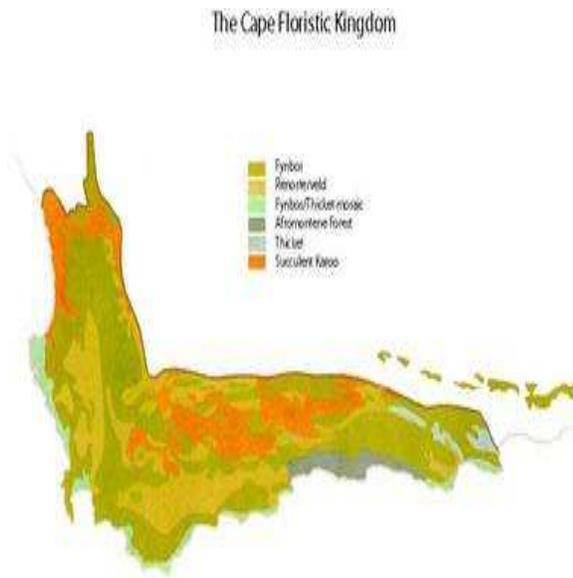
updating  $V_{3,i}$  directly from :

$$f(V_{3,i}|V_{3,j}, j \neq i, V_{1,i}, V_{2,i}, \boldsymbol{\theta}_-) \propto \prod_{i=1}^S \prod_{t=1}^T N(\mu_{i,t}; \mathbf{X}'_i \boldsymbol{\beta} + a_{12}V_{1,i} + a_{22}V_{2,i} \\ + (\rho + a_{13}V_{1,i} + a_{23}V_{2,i} + a_{33}V_{3,i})(t - t_0), \tau^2) N\left(\sum_{j=1}^S w_{i,j}V_{3,j}/w_{i,+}, w_{i,+}\right)$$

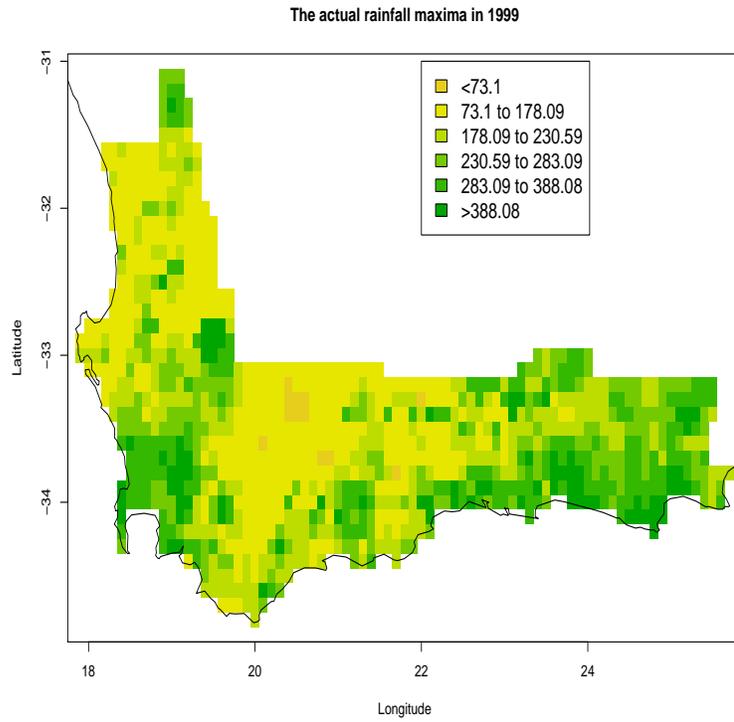
- Follow the updating procedures described in Model (A) to sample the other parameters.

	Mean	95% CI
$\beta_0$	249.4	(236.7, 255.2)
$\beta_{lat}$	-1.64e-03	(-0.054, 0.044)
$\beta_{lon}$	2.03e-03	(-0.015, 0.017)
$\beta_{elev}$	-1.23e-05	(-0.001, 0.001)
$\sigma_0$	90.15	(88.42, 92.23)
$\xi$	0.122	(0.112, 0.129)
$\rho$	0.150	(-0.06, 0.623)

**Table 2.2:** Posterior sample means of parameters and the corresponding 95% credible intervals for Model (C).  $\beta_{lat}$ ,  $\beta_{lon}$  and  $\beta_{elev}$  are the coefficients of longitude, latitude and elevation;  $\xi$  is the shape parameter of GEV distribution;  $\sigma_0$  is the center of spatially-varying scale parameters;  $\rho$  is the global trend in location parameters for Model (C).

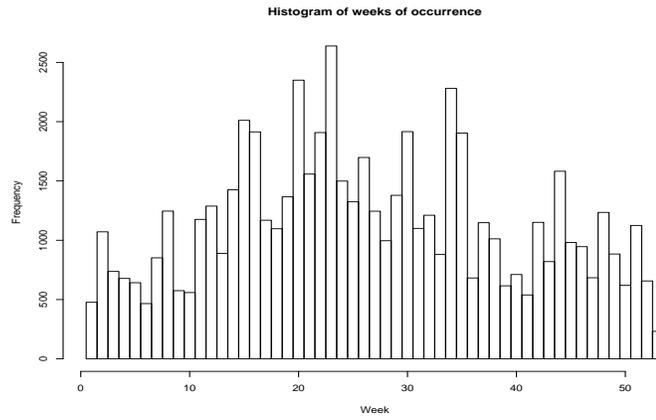


**Figure 2.1:** Map of Cape Floristic Region in South Africa [Image from R. Cowling and D. Richardson, 1995. Fynbos. Fernwood Press].

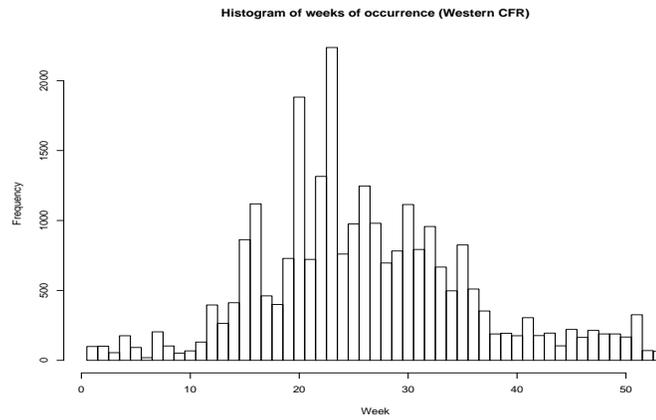


**Figure 2.2:** Image plot of the rainfall maxima over the gridded CFR in 1999.

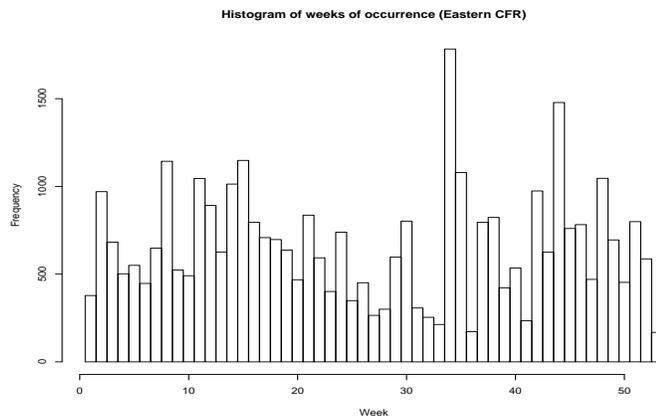
(a)



(b)

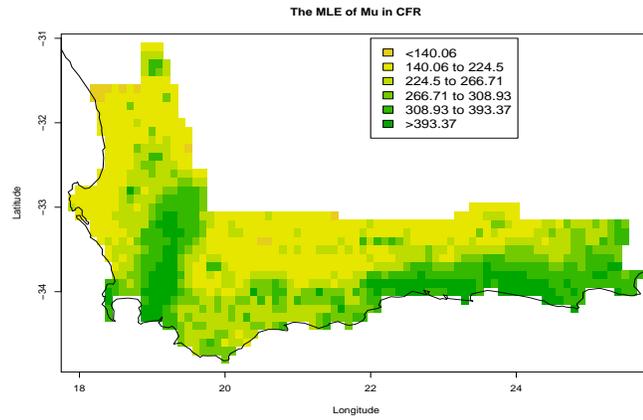


(c)



**Figure 2.3:** (a): Histogram of weeks when annual rainfall maxima occurred in CFR; (b): Histogram of weeks when annual rainfall maxima occurred in western CFR; (c): Histogram of weeks when annual rainfall maxima occurred in eastern CFR.

(a)



(b)

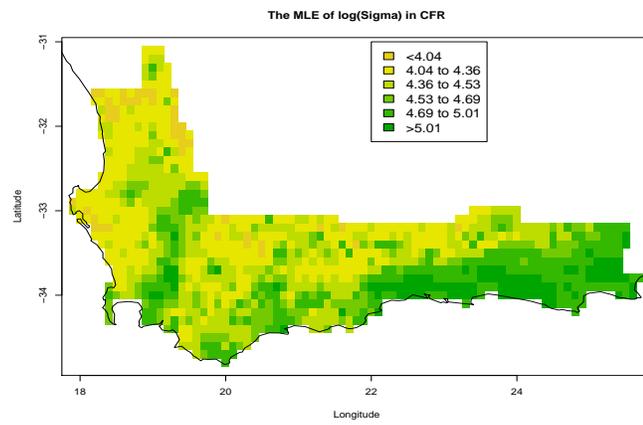
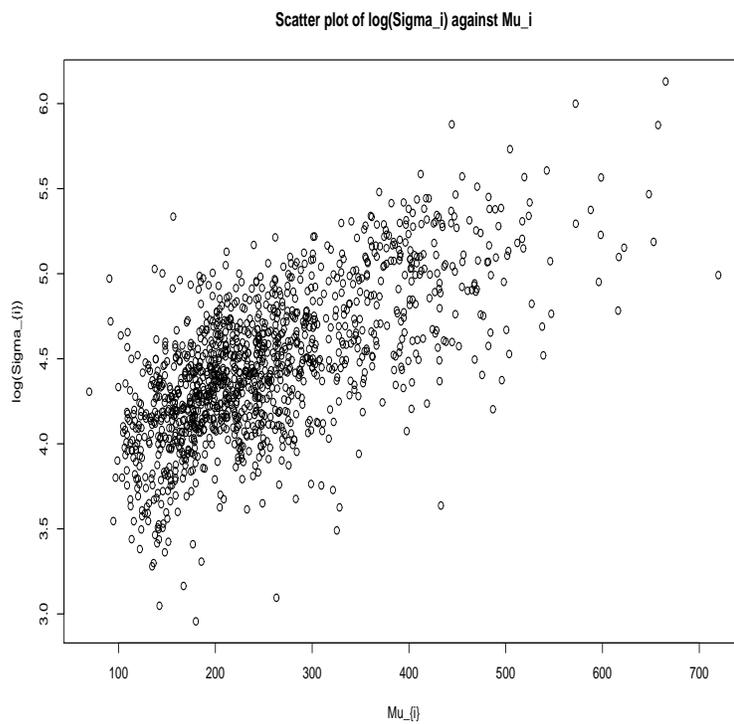
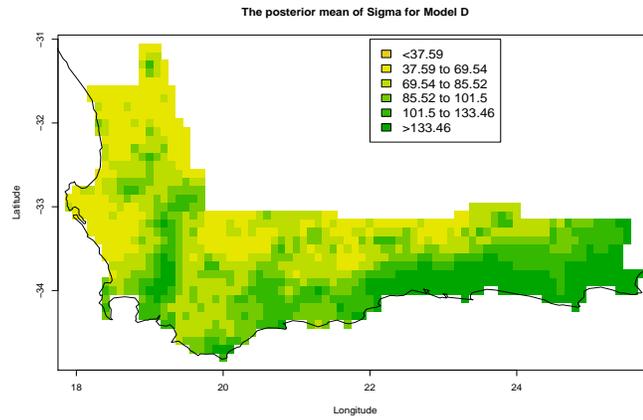


Figure 2.4: MLE's of the  $\mu_i$  and  $\log\sigma_i$  for the exploratory models.

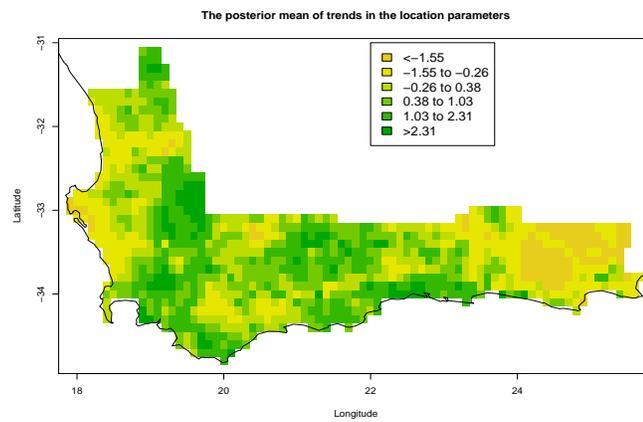


**Figure 2.5:** Scatter plot of the estimated  $\mu_i$  and  $\log\sigma_i$  for the exploratory models.

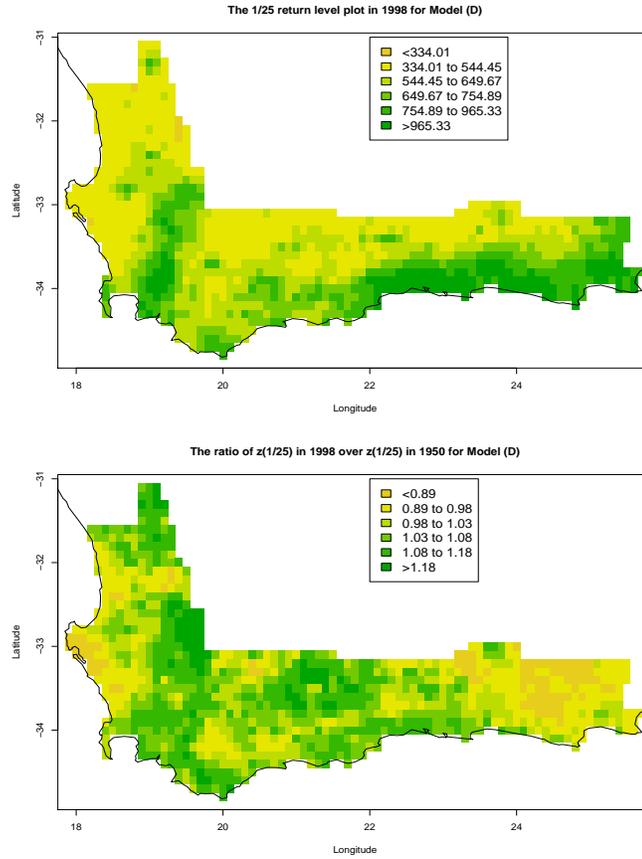
(a)



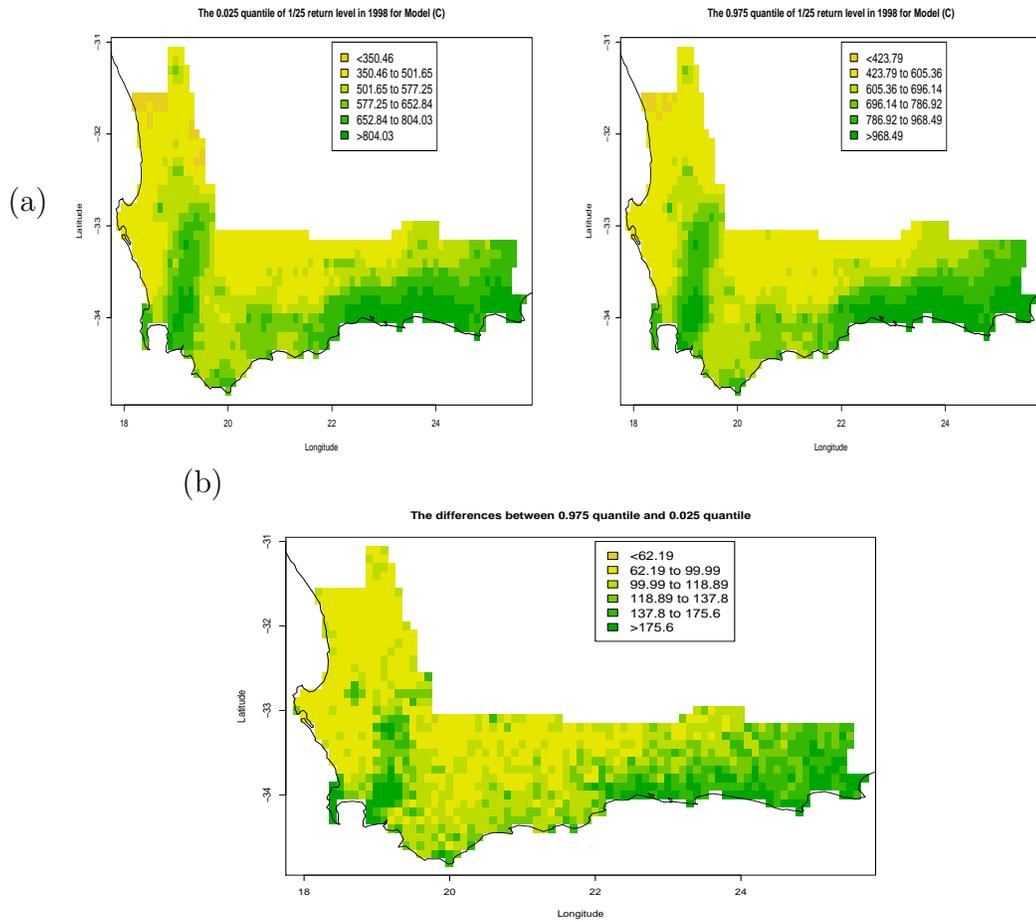
(b)



**Figure 2.6:** Posterior sample means of scale parameters and local trends obtained from Model C.



**Figure 2.7:** (a): Posterior sample means of the 1/25 return levels in 1998; (b): Posterior ratio of return level  $z_{1/25}$  in 1998 over  $z_{1/25}$  in 1950 obtained from Model C.



**Figure 2.8:** (a): The lower and upper .025 quantiles of the 1/25 return levels in 1998 obtained from Model C; (b): The lengths of 95% credible intervals of the 1/25 return levels in 1998 obtained from Model C.

# Chapter 3

## A Continuous Spatial Process of Extreme Values

### 3.1 Introduction

In chapter 2, we proposed a hierarchical modeling approach for explaining a collection of areal unit time series of extreme values. We assume that the observations follow Generalized Extreme Value(GEV) distributions whose locations and scales are jointly spatially dependent where the dependence is captured using multivariate Markov random field models specified through coregionalization. We fit the models to a set of gridded interpolated precipitation data collected over a 50 year period for the Cape Floristic Region in South Africa.

In this chapter, we continue to work on climate extremes with special interests in modeling of extreme temperatures at weather station level. Like heavy rainfalls, high temperatures is suggested to be one of the significant factors to explain plant performance, especially mortality, growth and reproductivity. Various indices of extreme temperatures are certainly available to use, such as the total number of days whose daily high temperatures are above certain threshold, or the number of consecutive days whose daily high temperatures are above certain threshold. We chose the annual maximum daily highest temperature as an index of extreme high temperatures for its clear interpretation and straightforward distribution theory .

In specific, the motivating data are annual maxima of daily highest temperatures derived from daily temperature records reported by 750 weather stations over South Africa

over the period from 1950 to 2000. Figure 3.1 displays the annual maxima of daily highest temperatures for 750 weather stations in 1999. We aim to understand two major scientific problems regarding the observed point-referenced time series of extreme values. The first problem of interest is the characterization of the spatial-temporal variations in the observed high temperatures. For example, we are interested to know changes in extreme temperatures in this region over the study period as well as the spatial scale at which extreme temperature operates across the CFR. As regard to spatial scale, it is not uncommon to find extreme climate events driven by multi spatial scale forcing, say, large regional forcing and small scale local forcing. Therefore, it is of interest to characterize spatial dependence for extreme values with multiple spatial scales. The second problem of interest is to predict the annual maxima of daily highest temperatures over those unmonitored locations in the CFR hence to help assess the risk of having extreme temperatures over the study region. The second problem is often known as the spatial interpolation for extreme values.

By now modeling of spatial extremes has received considerable attentions in the literature. Many existing approaches have been developed following the path of utilizing Extreme Value distribution theory (see Introduction 1.1). Some recent work focusing on spatial (or spatial temporal) characterization of extreme values include Kharin and Zwiers (2005), Cooley et al. (2007) and Sang and Gelfand (2008). One of the challenging issues in spatial extreme value modeling lies in the need for multivariate extreme value techniques in high dimensions. In contrast, most of the multivariate extreme value theory developed up to date only work well for low dimensional extreme values.

Most recently, several papers discuss statistical approaches regarding the problem of spatial interpolation for extreme values (Cooley et al. (2008) and Buishand and Zhou (2007)). In specific, given observed extreme values, the goal is to approximate the predictive distribution of the unobserved value at an unmonitored location (or a future time).

Prediction of extreme values is especially useful in quantitative risk monitoring of extreme climate events. Moreover, prediction methods in extreme climate studies have the potential to be applied in statistical downscaling techniques for extremes. One conventional approach is to build statistical models for the original spatial data without taking block maxima, and then interpolate spatial surface at each time unit. Extreme value surface is derive by taking time block maximum of the interpolated sequence at each location. In fact, this seemingly straightforward interpolation method usually smoothes away true maxima at unknown locations and underestimates the true extreme values. Therefore, we decide to work with the point-referenced annual maxima directly, and consider statistical model-based interpolation technique to obtain the interpolated extreme surface.

In chapter 2, we proposed a hierarchical modeling in which annual maxima follow the GEV distributions independently conditional on the parameters  $\mu$ 's,  $\sigma$ 's, and  $\xi$ 's. Those parameters are specified in the latent stage to reflect underlying spatio-temporal structure. For the point-referenced block maxima, we can use Gaussian processes in the latent parameter specifications where as in Chap 2 we used CAR's. However, there are some issues in directly applying this approach for interpolation of extreme values. In specific, we will obtain a nonsmooth predictive surface with conditional independence assumption in the hierarchical model even if the underlying parameters are specified using smooth processes. This is an undesired restriction especially when we work with smooth response data. In addition, the proposed hierarchical model in Chapter 2 could only characterize large scale spatial dependence in the latent parameter specifications. But in practice, there may still remain unexplained small scale spatial dependence in the extreme data. Therefore, direct use of the hierarchical model could lead to biased interpolation due to the failure in characterizing dependence information at a small scale. A detailed discussion on this issue can be found in 3.2.1.

The contribution of this paper is to extend the hierarchical modeling approach developed in Chapter 2 for explaining a collection of point-referenced time series of extreme values. Following Chapter 2, we assume annual maxima (see, Sang and Gelfand (2008) ) follow GEV distributions, with parameters  $\mu$ ,  $\sigma$ , and  $\xi$  specified in the latent stage to reflect underlying spatio-temporal structure. We relax the conditional independence assumption previously imposed on the first stage hierarchical models for annual maxima conditional on  $\mu$ ,  $\sigma$ , and  $\xi$ . Instead, we propose a continuous spatial process model for extreme values to account for spatial dependence which is unexplained by the latent spatio-temporal specifications for GEV parameters. In addition, we present a way to apply this smoothed spatial process model to the problem of spatial interpolation for extreme values. A simulation study is illustrated to investigate the model fitting behavior.

The format of the chapter is as follows. In Section 3.2 we briefly discuss limitations and possible extensions of the hierarchical modeling approach for spatial extreme values we developed in chapter 2 . We subsequently review some existing theories on continuous spatial process for extreme values in this Section. In Section 3.3, we construct a new continuous spatial process for extreme values through the transformed Gaussian process and discuss its statistical properties from different perspectives. In Section 3.4 we employ this new continuous spatial process for extreme values in the Bayesian hierarchical modeling framework in which we specify latent spatial temporal structures for the GEV parameters to explain large scale spatial dependence and assume a continuous spatial process to explain spatial dependence of extremes at fine scale. Model implementation is illustrated in Section 3.4.2 with special focus on the MCMC method. In section 3.5, we illustrates the proposed methods with a simulated dataset. Finally, Section 3.6 concludes with a brief discussion including future work.

## 3.2 Review on modeling for spatial extremes

There have been a few literature focusing on the statistical characterization of dependence structures of extreme values observed in space. In Section 3.2.1, we review a hierarchical modeling approach (Sang and Gelfand (2008)) for explaining a set of gridded interpolated precipitation time series of extreme values. The primary focus of that paper is to explain the extreme climate change and the spatial pattern associated with spatially referenced heavy rainfalls. In section 3.2.2, we review a paper by de Haan and Pereira (2006) in which the primary goal is that of spatial prediction for extreme rainfall event. The way they model spatial extreme values is through an introduction of a new stationary max-stable process, which they called the moving maximum process.

### 3.2.1 Hierarchical modeling approach for spatial extremes

In Chapter 2, we introduced the GEV distribution as a first stage model for annual maxima (see, Sang and Gelfand (2008)), specifying  $\mu$ ,  $\sigma$ , and  $\xi$  at the second stage to reflect underlying spatio-temporal structure. In particular, we let  $Y_{i,t}$  denote the annual maximum of daily rainfall at location  $i$  in year  $t$ . Then we assumed the  $Y_{i,t}$  follow a GEV distribution with parameters  $\mu_{i,t}$ ,  $\sigma_{i,t}$  and  $\xi_{i,t}$ , respectively. We assumed the  $Y_{i,t}$  are conditionally independent given their  $\mu$ 's,  $\sigma$ 's and  $\xi$ 's. Attention had been focused on specification of the model for  $\mu_{i,t}$ ,  $\sigma_{i,t}$  and  $\xi_{i,t}$ .

We now turn our attention to the modeling of annual maxima observed at point level. Let  $Y(\mathbf{s}, t)$  denote the annual maximum of daily highest temperature at location  $s$  in year  $t$ . With annual maxima as the response data. It is natural to keep the assumption that  $Y(\mathbf{s}, t)$  follows a GEV distribution with parameters  $\mu(\mathbf{s}, t)$ ,  $\sigma(\mathbf{s}, t)$  and  $\xi(\mathbf{s}, t)$ , respectively. Following Chapter 2, we would specify the latent stage models for  $\mu$ 's,  $\sigma$ 's and  $\xi$ 's. If we follow the method described in Chapter 2, assumption of conditional independence has to

be imposed in the first stage for the response data. However, in practice, one might question the plausibility of this conditional independence assumption. Assumption of independence in time may be plausible since taking a year as block size for taking maximum may lead to roughly weakly independent annual maximum observations at a particular site. However, in space, despite the fact that large scale spatial dependence may have been accounted for in the latent parameter specifications, there may still remain unexplained small scale spatial dependence in the extreme data. In chapter 2, since we only have areal unit data at 10km scale, modeled as areal unit data since that is the interpretation of the interpolation. We do not have tools to study finer scale dependence. In this chapter, we are studying the point-referenced annual maximum temperatures, which is likely to involve small scale dependence in space.

There is another issue associated with this conditional independence assumption if we consider to apply it on spatial interpolation of extreme values. Conditional on the parameters in GEV distribution, suppose we assume  $Y(\mathbf{s}, t)$  independently follows a GEV distribution as follows:

$$Y(\mathbf{s}, t) | (\mu(\mathbf{s}, t), \sigma(\mathbf{s}, t), \xi(\mathbf{s}, t)) \sim GEV(\mu(\mathbf{s}, t), \sigma(\mathbf{s}, t), \xi(\mathbf{s}, t))$$

Given a new site  $\mathbf{s}_0$  and year  $t$ , we are interested in knowing the posterior distribution of  $Y(\mathbf{s}_0, t)$  conditional on all the observed annual maxima. The predictive distribution of  $Y(\mathbf{s}_0, t)$  is given by:

$$(Y(\mathbf{s}_0, t) | \mathbf{Y}) \sim \int P(Y(\mathbf{s}_0, t) | \mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0, t), \xi(\mathbf{s}, t)) \quad (3.1)$$

$$P((\mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0, t), \xi(\mathbf{s}, t)) | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \Omega) \quad (3.2)$$

$$P(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \Omega | \mathbf{Y}) d\boldsymbol{\mu} d\boldsymbol{\sigma} d\boldsymbol{\xi} d\Omega \quad (3.3)$$

where  $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}) = (\{\mu(\mathbf{s}, t)\}, \{\sigma(\mathbf{s}, t)\}, \{\xi(\mathbf{s}, t)\}, s = 1, \dots, n; t = 1, \dots, T)$ ,  $\Omega$  denotes all the other parameters except for  $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}$  in the model. Equation (3.3) suggests that  $Y(\mathbf{s}_0, t)$

should be sampled from its predictive distribution by composition. We first obtain the posterior samples  $(\mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0, t), \boldsymbol{\xi}(\mathbf{s}_0, t))$  conditional on the posterior draws of  $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \Omega$ . In the next step,  $Y(\mathbf{s}_0, t)$  are drawn independently given samples of  $(\mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0, t), \boldsymbol{\xi}(\mathbf{s}_0, t))$ .

Assuming continuous spatial-temporal processes for the latent parameters  $\boldsymbol{\mu}, \boldsymbol{\sigma}$  and  $\boldsymbol{\xi}$ , we could obtain smooth surface for each model parameter. However, as suggested in (3.3), the conditional independence assumption imposed in the first stage will result in a nonsmooth response surface even with smooth surfaces for the GEV parameters. In practice, many climate events such as temperature are suggested to be smooth across space.

We elaborate the above argument by expressing the model in a different form. Considering the marginal transformation of the GEV distribution, the first stage of the hierarchical model in Chapter 2 can be written as:

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \frac{\sigma(\mathbf{s}, t)}{\xi(\mathbf{s}, t)}(z(\mathbf{s}, t)^{\xi(\mathbf{s}, t)} - 1) \quad (3.4)$$

where  $z(\mathbf{s}, t)$  follows a standard Fréchet distribution, i.e.,  $GEV(-1, 1, 1)$  (see the details in 1.1.2). We may view  $z(\mathbf{s}, t)$  as the 'standardized residual' in the first stage GEV model. It is noteworthy that the conditional independence assumption we made previously is equivalent to making the *i.i.d* assumption for each  $z(\mathbf{s}, t)$ , i.e.,

$$z(\mathbf{s}, t) \sim \text{standard Fréchet } i.i.d \quad (3.5)$$

Again, (3.5) clearly reveals that even if the surface for each model parameter is smooth, we will obtain a nonsmooth predictive surface with conditional independence assumption in the hierarchical model. In this regard, we would like to remedy the hierarchical modeling approach by relaxing the conditional independence assumption in some fashion.

### 3.2.2 A continuous process for spatial extremes: moving maximum process

de Haan and Pereira (2006) proposed a stationary max-stable process as a model of the dependence structure in two dimensional spatial problems. They have applied this max-stable process to predict the distributions of extreme rainfalls across Netherland.

They started with independent replicated of a stochastic process with continuous sample paths  $\{x_n(t)\}_{t \in \mathbb{R}}$  for  $n = 1, 2, \dots$ . Suppose that this process is in the domain of attraction of a max-stable process, that is, there are sequences of continuous functions  $a_n > 0$  and  $b_n$  such that as  $n \rightarrow \infty$ ,  $\{\frac{\max_{1 \leq i \leq n} X_i(t) - b_n(t)}{a_n(t)}\} \rightarrow \{\tilde{\eta}(t)\}_{t \in \mathbb{R}}$  in C-space. Then the limit process  $\tilde{\eta}(t)_{t \in \mathbb{R}}$  is a max-stable process (see 1.1.2 for definitions). Without loss of generality, they worked with simple max-stable process whose marginal distribution functions are all standard Fréchet.

The proposed max stable process is described as follows:

**Definition 7.** : A mapping  $\Phi$  from  $L_1^+$  (the non-negative integrable functions on  $\mathbb{R}$ ) to  $L_1^+$  is called a piston if for  $h \in L_1^+$ ,

$$\Phi(h(t)) = r(t)h(H(t))$$

with  $H$  is a one-to-one measurable mapping from  $\mathbb{R}$  to  $\mathbb{R}$  and  $r$  is a positive measurable function, such that for every  $h \in L_1^+$ ,

$$\int_{\mathbb{R}} \Phi(h(t))dt = \int_{\mathbb{R}} h(t)dt \tag{3.6}$$

**Theorem 8.** Let  $\{(Z_i, T_i)\}$  be a realization of a Poisson process on  $(0, +\infty] \times \mathbb{R}$  with mean measure  $(dr/r^2) \times d\lambda$  ( $\lambda$  is the Lebesgue measure). If the stochastic process  $\{\eta(s)\}_{s \in \mathbb{R}}$  is simple max-stable, strictly stationary and continuous a.s., then there is a function  $h \in L_1^+$

with  $\int_{\mathbb{R}} h(t)dt = 1$  and a continuous group of pistons  $\{\Phi_s\}_{s \in \mathbb{R}}$  (continuous, i.e.  $\Phi_{s_n}(h(t)) \rightarrow \Phi_s(h(t))$  as  $s_n \rightarrow s$  for almost all  $t \in \mathbb{R}$  with

$$\int_{\mathbb{R}} \sup_{s \in I} \Phi_s(h(t)) dt < \infty$$

for each compact interval  $I$ , such that

$$\{\eta(s)\}_{s \in \mathbb{R}} = \{\max_{i \geq 1} Z_i \Phi_s(h(T_i))\}_{s \in \mathbb{R}}$$

They proved that, conversely, every stochastic process of the form exhibited at the right hand side of the above equation with the stated conditions is simple max-stable, strictly stationary and a.s. continuous.

A special case is obtained by setting  $\Phi_s(h(t)) := h(t - s)$ , where  $h$  is a continuous probability density. Examples have been discussed in (de Haan and Pereira (2006)). Examples for  $h$  include normal density, double exponential and t-density. Following the simple max-stable process defined in (??), de Haan and Pereira (2006) derived an explicit form of the bivariate distribution function for any two realizations of the process at given two points. In climate study, however, the number of observed locations is typically hundreds or more. Unfortunately, the integration in the distribution function of realizations of their proposed spatial process is intractable hence places major obstacle for us to derive the likelihood function for high-dimensional ( $n > 2$ ) realizations. Recall the review on multivariate extreme value theory in 1.1.2. We find most of the multivariate extreme value theory developed up to date are only applicable to low dimensional problems. For example, logistic type of multivariate extreme value distribution (see Coles and Tawn (1991)) has a too restrictive parametric form to capture general dependence structures for large dimension extremes. Some nonparametric models such as the tilted Dirichlet model (Coles (1993)) require integration over the spectral function, which is typically infeasible especially for high dimensions.

### 3.3 A smooth spatial process for extreme values

In this section, we provide an alternative simple and intuitive continuous spatial process of extreme values through the transformation of a Gaussian process. Again, without loss of generality, we restrict ourselves to discussing the standard spatial process of extreme values with Fréchet marginal distributions.

#### 3.3.1 Copulas with multivariate extreme value distributions

Before formally introducing the transformed Gaussian process, we begin with an introduction of a constructed multivariate extreme value distribution family using Gaussian copulas. In fact, the development of transformed Gaussian process is largely motivated by the Gaussian copulas idea.

As discussed in the preceding section, the main difficulty of the multivariate extreme value distribution lies in the lack of an efficient way to describe dependence structures for high dimensional extreme variables. A good news is that the univariate extreme value distribution theory has been well studied, which enables us to concentrate on the dependence in extreme values. We now seek to find a strategy to effectively reflect dependence structures in a multivariate distribution for high dimensional extremes in some fashion.

One available way to construct the multivariate distributions is through copulas, which have received increasing attentions and applications in the past two decades. (see, Nelsen (2006) for a review). We will briefly describe some fundamental definitions and theorems about copulas below.

**Definition 9.** *A  $p$ -dimensional copula is the distribution function of a random vector in  $\mathbb{R}^p$  defined on the  $p$ -dimensional unit cube  $[0, 1]^p$  and with uniform marginal distribution. An equivalent definition is:  $C : [0, 1]^p \rightarrow [0, 1]$  is an  $p$ -dimensional copula if:*

- $C(\mathbf{u}) = 0$  whenever  $\mathbf{u} \in [0, 1]^p$  has at least one component equal to 0;
- $C(\mathbf{u}) = u_i$  whenever  $\mathbf{u} \in [0, 1]^p$  has all the components equal to 1 except the  $i$ -th one, which is equal to  $u_i$ ;
- For all  $(a_1, \dots, a_p), (b_1, \dots, b_p) \in [0, 1]^p$  with  $a_i < b_i$ , we have

$$\sum_{i_1=1}^2 \dots \sum_{i_p=1}^2 (-1)^{i_1 + \dots + i_p} C(u_{1,i_1}, \dots, u_{p,i_p}) \geq 0 \quad (3.7)$$

where  $u_{j,1} = a_j$  and  $u_{j,2} = b_j$  for all  $j \in \{1, \dots, p\}$

Sklar (1959) proved a fundamental theorem which provides the representation of any given multivariate distribution using copulas.

**Theorem 10.** *Let  $H$  be a  $p$ -dimensional distribution function with margins  $F_1, \dots, F_p$ . Then there exists a copula  $C$  such that for all  $x \in \bar{\mathbb{R}}^p$ , where  $\bar{\mathbb{R}}^p = \mathbb{R} \cup \{\pm\infty\}$ , such that*

$$H(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p)) \quad (3.8)$$

If the marginal distributions are all continuous, then the copula function  $C$  is unique. Otherwise, the copula  $C$  is unique on the range of values of the marginal distributions. Conversely, for a copula  $C$  and continuous margins  $F_1, \dots, F_p$ , the function  $H$  define by (3.8) is a  $p$ -dimensional distribution function with margins  $F_1, \dots, F_p$ .

There are some commonly used parametric copulas. One example is the Gaussian Copula, which is constructed from Gaussian distribution via Sklar's theorem. Consider a bivariate random vector distributed as standard bivariate Gaussian distribution with correlation  $\rho$ . The Gaussian copula function is as follows:  $C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$

where the  $u, v \in [0, 1]$ ,  $\Phi$  denotes the standard normal cumulative distribution function and  $\Phi_\rho$  denotes the cumulative distribution function of the standard bivariate Gaussian distribution with correlation  $\rho$ . Differentiating this yields  $c_\rho(u, v) = \frac{\phi_\rho(\Phi_X^{-1}(u), \Phi_Y^{-1}(v))}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))}$

where  $\phi_\rho(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} [x^2 + y^2 - 2\rho xy]\right)$  is the density function for the bivariate normal variate with correlation coefficient  $\rho$ ,  $\phi$  is the density of the  $N(0, 1)$  distribution (the marginal density).

Let  $C_N$  denote the Gaussian copula. Consider a bivariate random vector  $(X, Y)$  with marginal GEV distributions  $G_x$  and  $G_y$ . Following theorem (3.8), it is easy to construct a bivariate extreme value distribution with the Gaussian copula with the following distribution function  $G(x, y) = C_N(G_x(x), G_y(y))$ . Let  $(X, Y) = (G_x^{-1}\Phi(X'), G_y^{-1}\Phi(Y'))$ , where  $G_x^{-1}$  and  $G_y^{-1}$  are the inverse marginal distribution functions for  $X$  and  $Y$ . Then the distribution function of  $(X, Y)$  is given by  $H(X, Y) = C_N(\Phi(X'), \Phi(Y'))$ . And the marginal distributions of  $X$  and  $Y$  remain to be  $G_x$  and  $G_y$ .

Now we can study the dependence structures defined through the Gaussian copula. As we discussed in the Introduction 1.1.2, several dependence metrics are available to measure the bivariate dependence. Since the constructed bivariate extreme value distribution has an explicit form, it is convenient to compute those proposed dependence measures. In specific, for a given bivariate Gaussian Copula with correlation coefficient  $\{\rho \geq 0\}$ , we can analytically solve the extremal coefficient for the constructed bivariate extreme distribution  $G(x, y; \rho)$ . Remembering that the extremal coefficient is defined as  $\phi = Pr(X < 1, Y < 1)$  (see, 1.1.2), we obtain

$$\phi = Pr(X < 1, Y < 1, \rho) \tag{3.9}$$

$$= \Phi_\rho(\Phi^{-1}G_x(1), \Phi^{-1}G_y(1); \rho) \tag{3.10}$$

where  $\Phi_\rho$  is the cumulative distribution function for a standard bivariate Gaussian distribution with correlation coefficient  $\rho$  and  $\Phi^{-1}$  is the inverse CDF of  $N(0, 1)$ . Let  $H(\rho) = \Phi_\rho(X' < \Phi^{-1}G_x(1), Y' < \Phi^{-1}G_y(1); \rho)$ . In Figure 3.3 we plot the extremal coefficients of the constructed bivariate extreme distribution with standard Fréchet marginals for a set of  $\{\rho_i \geq 0\}$ . It is not surprising to find a strictly monotone decreasing relationship between  $\phi$

and  $\rho$ . When  $\rho = 1$ ,  $\phi = 0$  which corresponds to the complete dependence case. When  $\rho = 0$ ,  $\phi = 1$  corresponding to the independence between  $X$  and  $Y$ .

### 3.3.2 Transformed Gaussian processes

Consider a standard spatial Gaussian process  $z^*(\mathbf{s})$  with mean centered at  $\mathbf{0}$  and variance at 1. We constructed a transformed Gaussian process based on  $z^*(\mathbf{s})$  as follows:

$$z(\mathbf{s}) = G^{-1}\Phi(z^*(\mathbf{s})) \quad (3.11)$$

where  $\Phi$  is the distribution function of a standard normal distribution, and  $G$  is the distribution function of a standard Fréchet distribution.

Suppose we observe extreme values at a set of sites  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ . The realizations of  $z^*(\mathbf{s})$  at  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  follow a multivariate normal distribution which is determined by the given covariance function of  $z^*(\mathbf{s})$ . Denote  $\mathbf{z} = (G^{-1}\Phi(z^*(\mathbf{s}_1)), \dots, G^{-1}\Phi(z^*(\mathbf{s}_n)))$  and  $\mathbf{z}^* = (z^*(\mathbf{s}_1), \dots, z^*(\mathbf{s}_n))$ .

Given the correlation function  $\rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$  of the standard Gaussian process  $z^*(\mathbf{s})$ , we can derive the Gaussian copula  $C_{\mathbf{z}^*}$  for the distribution function of  $\mathbf{z}^*$  as follows:

$$C_{\mathbf{z}^*}(u_1, \dots, u_n) = F_{\mathbf{z}^*}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)) \quad (3.12)$$

where  $(u_1, \dots, u_n) \in [0, 1]^n$  and  $F$  is the multivariate distribution function of  $\text{MVN}(\mathbf{0}, \Sigma)$  with  $\Sigma = [\rho(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$ .

Let  $F(\mathbf{z})$  denote the multivariate distribution of  $\mathbf{z}$ , then

$F(\mathbf{z}) = C_{\mathbf{z}^*}(\Phi^{-1}G(z_1), \dots, \Phi^{-1}G(z_n))$ , where  $\mathbf{z} = (z_1, \dots, z_n) \in \mathbb{R}^n$ . It can be proved that the marginal GEV distributions for each  $z(\mathbf{s}_i)$  with joint distribution  $F(\mathbf{z})$  is still a standard Fréchet, i.e.,  $\text{GEV}(-1, 1, 1)$ . In this regard, we construct a multivariate GEV distribution in which the marginal distribution of each variable follows a standard Fréchet distribution and the dependence structure is completely determined by the Gaussian copula  $C_{\mathbf{z}^*}$ .

(3.11) clearly reveals that the Gaussian process  $z^*(\mathbf{s})$  introduces the spatial dependence into our model. In fact, the transformed Gaussian process we proposed in (3.11) is essentially a one to one monotone mapping from a Gaussian process to an extreme value process. There has been a rich amount of effort focusing on the development of Gaussian spatial process models (see Banerjee et al. (2004)). We can conveniently embed the transformed Gaussian process into various existing Gaussian spatial processes models, hence share several advantages of working under Gaussian framework. For example, we can immediately derive several nice statistical properties of the transformed Gaussian process as follows:

- Joint, marginal and conditional distributions of extreme values  $\mathbf{z}$  are all immediately obtained from the standard distribution theory once the mean and covariance structure of Gaussian processes have been specified.
- There are plentiful candidates of covariance structures established for Gaussian spatial processes, which provide a large pool for us to flexibly specify dependence structures for spatial extreme observations.
- Many established efficient computational algorithms could be utilized by transforming back to Gaussian processes.
- Transformation in (3.11) retains the stationarity property. If the Gaussian process  $z^*(\mathbf{s})$  is a stationary spatial process with a valid covariance function  $\rho(\mathbf{s} - \mathbf{s}', \boldsymbol{\theta})$ , then  $z^*(\mathbf{s})$  is also strongly stationary, which implies that  $z(\mathbf{s})$  is a strongly stationary process, i.e., the distribution of  $(z(\mathbf{s}_1), \dots, z(\mathbf{s}_n))$  is the same as that of  $(\mathbf{z}(\mathbf{s}_1 + \mathbf{h}), \dots, \mathbf{z}(\mathbf{s}_n + \mathbf{h}))$  for any  $\mathbf{h} \in \mathbb{R}^2$ . And all the dependence metrics described in Section 1.1.2, including Kendall's  $\tau$ , Spearman's  $\rho$ , Schweizer and Wolff's  $\sigma$  and the extremal coefficient, only depend on  $\mathbf{s} - \mathbf{s}'$ . To illustrate, we derive the extremal

coefficient between  $z(\mathbf{s})$  and  $z(\mathbf{s}')$  as follows:

$$\phi(z(\mathbf{s}), z(\mathbf{s}')) = \Phi_\rho(\Phi^{-1}G(1), \Phi^{-1}G(1); \rho(\mathbf{s} - \mathbf{s}', \boldsymbol{\theta})) \quad (3.13)$$

where  $\Phi_\rho$  is the bivariate normal distribution function with mean  $\mathbf{0}$  and variance-covariance matrix  $\begin{pmatrix} 1 & \rho(\mathbf{s} - \mathbf{s}', \boldsymbol{\theta}) \\ \rho(\mathbf{s} - \mathbf{s}', \boldsymbol{\theta}) & 1 \end{pmatrix}$ .

- Most importantly, the transformed Gaussian approach is not limited to the application on extreme value analysis. In theory, for any Non-Gaussian spatial response data set which are assumed to have marginal distribution function  $G$  (or more generally,  $\{G_{\mathbf{s}}\}$  at each location  $\mathbf{s}$ ),  $G^{-1}\Phi(\mathbf{s})$  described in (3.11) provides a valid spatial process with a consistent marginal distribution with the one we initially assumed for the given response data. Therefore, we could envision a general extension of this method to a broad range of Non-Gaussian spatial context besides extreme values. However, in some cases, function  $G$  may involve a complicated form (e.g., intractable integration) such that  $G^{-1}$  is intractable and difficult to evaluate, which could yield the failure of using (3.11).

Despite of those appealing properties of transformed Gaussian approach we list above, we should admit that this copula-based approach suffers from a number of drawbacks. First, the multivariate extreme value process developed using transformed Gaussian is adopted as an asymptotic model for spatial maxima. But the dependence of multivariate extremes in our model is constructed in a way that has little connection with the asymptotic properties of standard multivariate extreme value theory (see 1.1.2). The second aspect is that after marginal transformations, the constructed multivariate extreme value distribution is essentially determined by the specified covariance function of the Gaussian process, which is not able to reasonably explain dependence properties for a general class of multivariate extremes. In specific, the transformed Gaussian process could only characterize higher or-

der dependence through the bivariate dependence specifications, which is not necessarily the case for multivariate extremes.

## 3.4 Bayesian modeling approach for spatial extreme values

In Section 3.3, we construct a continuous process for spatial extremes with standard Fréchet marginal distributions. However, in practice, the marginal behaviors for spatial extreme values are typically unknown to us and in need of modeling as well. In this section, we aim to develop a hierarchical modeling approach in which spatial dependence in extreme values is captured through latent parameter specification as well as the transformed Gaussian processes modeling for 'standardized residuals'. We offer a MCMC approach for the implementation of the proposed hierarchical models in the end of this section.

### 3.4.1 A hierarchical modeling approach

We introduce the GEV distribution as a first stage model for annual maxima at point level. Notice that  $Y(s, t)$  denotes the annual maxima at location  $s$  at time  $t$ . The first stage of the hierarchical model is written as follows:

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \frac{\sigma(\mathbf{s}, t)}{\xi(\mathbf{s}, t)}(z(\mathbf{s}, t)^{\xi(\mathbf{s}, t)} - 1) \quad (3.14)$$

For the second stage model, model specifications for  $\mu_{i,t}$ ,  $\sigma_{i,t}$  and  $\xi_{i,t}$  have to be taken into careful consideration. In chapter 2, we first conducted some exploratory analysis to learn about temporal trend in these parameters and learn about spatial dependence in these parameters. We then followed those exploratory results as guidelines in making formal model specifications. Here, we can follow the similar path to conduct some type of

exploratory analysis and obtain GEV parameter estimations based on independent GEV models (see 2.2). We find that the annual maximum temperatures at weather station exhibit similar spatial temporal pattern in GEV parameters as we observed in the annual maximum rainfall data. In specific, we assume there is spatial temporal dependence for the  $\mu$ 's and  $\sigma$ 's, the temporal dependence for the  $\mu$ 's but not for the  $\sigma$ 's,  $\xi$  is unknown but constant across the study region. Following those assumptions, we illustrate some examples of modeling specifications at the second stage explaining extreme values observed at point level. First, we can propose the specification  $p(\mu(\mathbf{s}, t)|\boldsymbol{\beta}, W(\mathbf{s}, t), \tau^2) = N(\mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + W(\mathbf{s}, t), \tau^2)$ .  $\mathbf{X}(\mathbf{s})$  is the site-specific vector of potential explanatory variables. Here, we are using Gaussian processes and assuming  $W(\mathbf{s}, t)$  is a spatial-temporal random effect where as in Chap 2 we used CAR's.

- An additive form:

- **Model A:**  $W(\mathbf{s}, t) = \psi(\mathbf{s}) + \delta_t$ ,  $\delta_t = \phi\delta_{t-1} + \omega_t$ , where  $\omega_t \sim N(0, W_0^2)$  *i.i.d*

- A linear temporal component with spatial random effects:

- **Model B:**  $W(\mathbf{s}, t) = \psi(\mathbf{s}) + \rho(t - t_0)$

- **Model C:**  $W(\mathbf{s}, t) = \psi(\mathbf{s}) + (\rho + \rho(\mathbf{s}))(t - t_0)$

- A multiplicative form in space and time:

- **Model D:**  $W(\mathbf{s}, t) = \psi(\mathbf{s})\delta_t$ ,  $\delta_t = \phi\delta_{t-1} + \omega_t$ , where  $\omega_t \sim N(0, W_0^2)$  *i.i.d*

$\psi(\mathbf{s})$  can be interpreted as the spatial component in the location parameter, commonly modelled as Gaussian process.  $\rho(\mathbf{s})$  can be interpreted as the site level trends in the location parameters, and usually modelled as Gaussian process as well. And  $\delta_t$  can be interpreted the temporal component. See Chapter 2 for the detail interpretations about these 5 proposed specification.

For the weather station extreme data, we could also check the association between the estimated location parameters and the estimated scale parameters in the exploratory data analysis. For the annual maxima of temperatures in the CFR, we found the correlation is not as strong as that of annual maxima of rainfalls in the CFR. We may treat the locations and the scales to be independent Gaussian processes. But in a general context, if we observe the existence of dependence between the location parameter and the scale parameters, we will need to jointly model  $\log\sigma(\mathbf{s})$  and the  $\psi(\mathbf{s})$ . We could borrow the concept of coregionalization once again in the same fashion as we model multivariate spatial areal data. Let  $\mathbf{U}(\mathbf{s}) = (U_1(\mathbf{s}), U_2(\mathbf{s}), \dots, U_p(\mathbf{s}))'$  denote the  $P \times 1$  vector of spatial random effects for the study region. Introducing a lower triangular transformation matrix  $\mathbf{A}$  with elements  $a_{i,j}$ , let  $\mathbf{U}(\mathbf{s}) = \mathbf{A}\mathbf{V}(\mathbf{s})$ , where  $\mathbf{V}(\mathbf{s}) = (V_1(\mathbf{s}), V_2(\mathbf{s}), \dots, V_p(\mathbf{s}))'$ . For the  $\mathbf{V}_j(\mathbf{s})$ , we assume independent spatial Gaussian process models with a specified covariance function  $C$ . The scale parameter in covariance function  $C$  is set to be 1 since the entries in  $\mathbf{A}$  already provide the scaling. Take Model (1) as an example, let  $\log(\sigma(\mathbf{s})) = \sigma_0 \exp(\lambda(\mathbf{s}))$ . the coregionalization spatial Gaussian model is used to jointly model  $\lambda(\mathbf{s})$  and  $\psi(\mathbf{s})$ .  $(\lambda(\mathbf{s}), \psi(\mathbf{s}))' = \mathbf{A}(V_1(\mathbf{s}), V_2(\mathbf{s}))'$ , where  $\mathbf{A} = \begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{pmatrix}$  where  $\mathbf{V}_1(\mathbf{s})$  and  $\mathbf{V}_2(\mathbf{s})$  are two independent Gaussian process models.

Next, we model the 'standardized residuals' in (3.14) using the transformed Gaussian process model:

$$z(\mathbf{s}, t) = G^{-1}\Phi(z^*(\mathbf{s}, t)) \quad (3.15)$$

In the spirit of Section 3.3,  $z^*(\mathbf{s}, t)$  is modelled using a spatiotemporal Gaussian process from a rich class of specifications. There has been considerable research work on the specifications of spatial temporal Gaussian process (see Gelfand et al. (2004a) and Banerjee et al. (2004)). As an illustration, we consider the following specification for  $z^*(\mathbf{s}, t)$ . For any given year  $t$

in  $\{1, \dots, T\}$ , let  $z_t^*(\mathbf{s}) \sim GP(\mathbf{0}, C(\cdot, \boldsymbol{\theta}))$  in which the covariance scale is 1. Moreover, we assume  $z_t^*(\mathbf{s})$  and  $z_{t'}^*(\mathbf{s})$  are two independent Gaussian processes when  $t \neq t'$ . It is plausible to make this assumption of temporal independence since annual block size may be long enough to yield independent annual maximum observations.

Under this hierarchical modeling settings, the marginal distribution of the annual maximum observations at a particular site is still  $GEV(\mu_{\mathbf{s},t}, \sigma_{\mathbf{s}}, \xi)$ . By introducing Model (3.15), we are able to learn the dependence structures in the residuals. And, more importantly, if the underlying parameters in GEV are specified using continuous processes and the explanatory variables are also continuous in space, the hierarchical modeling we proposed leads to a smooth predictive surface of extreme values which is desired for many extreme climate events with smooth surfaces.

Given a new location  $\mathbf{s}_0$  at year  $t$ , the posterior predictive distribution of  $Y(\mathbf{s}_0, t)$  conditional on all the observed annual maxima is given by:

$$(Y(\mathbf{s}_0, t) | \mathbf{Y}) \sim \int P(Y(\mathbf{s}_0, t) | \mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0), \xi, z_t^*(\mathbf{s}_0)) \quad (3.16)$$

$$P((\mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0), z_t^*(\mathbf{s}_0)) | \boldsymbol{\mu}, \boldsymbol{\sigma}, \xi, \mathbf{z}^*, \Omega, \mathbf{X}(\mathbf{s}_0)) \quad (3.17)$$

$$P(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi, \mathbf{z}^*, \Omega | \mathbf{Y}, \mathbf{X}, \mathbf{X}(\mathbf{s}_0)) d\boldsymbol{\mu} d\boldsymbol{\sigma} d\xi d\mathbf{z}^* d\Omega \quad (3.18)$$

where  $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = (\{\mu(\mathbf{s}, t), s = 1, \dots, n; t = 1, \dots, T\}, \{\sigma(\mathbf{s}), s = 1, \dots, n\}, \mathbf{z}^* = \{\mathbf{z}^*(\mathbf{s}, t), s = 1, \dots, n; t = 1, \dots, T\}$ ,  $\Omega$  denotes all the other parameters except for  $\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi, \mathbf{z}^*$  in the model. Under Bayesian modeling framework, we could draw the posterior samples from the predictive distribution for  $Y(\mathbf{s}_0, t)$  by composition. We first draw posterior samples of  $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi, \mathbf{z}^*, \Omega)$  using MCMC algorithms. Next, conditional on the posterior draws of  $\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi, \mathbf{z}^*, \Omega$ , we then draw samples of  $(\mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0), z_t^*(\mathbf{s}_0))$ . Recall the conventional prediction procedure in Gaussian processes (see Banerjee et al. (2004)). Explicit forms of conditional posterior distributions for  $(\mu(\mathbf{s}_0, t), \sigma(\mathbf{s}_0))$  and  $z_t^*(\mathbf{s}_0)$  help us to draw samples straightforwardly within the Gaussian framework. The final step of spatial interpolation

now becomes straightforward:  $Y(\mathbf{s}_0, t) = \mu(\mathbf{s}_0, t) + \frac{\sigma(\mathbf{s})}{\xi}(z(\mathbf{s}_0, t)^\xi - 1)$ .

### 3.4.2 Bayesian implementation

Given the complexity of the model, we employ MCMC methods, the most popular Bayesian computing tools, to implement the proposed models in (3.4.1). We begin with prior specifications for the parameters. Customarily, we set  $\boldsymbol{\beta} \sim MVN(\boldsymbol{\mu}_\beta, \Sigma_\beta)$ . In our model, a vague normal prior is assigned to the shape parameter,  $\xi$ . We assign inverse gamma priors for positive value  $\sigma_0$ . Also recall that transformation matrix  $A$  itself is unknown and needs to be stochastically specified. We model  $AA^T$  with an Inverse-Wishart prior. In general the spatial range parameters are weakly identifiable and prior specifications become somewhat delicate. Reasonably informative priors are needed for satisfactory MCMC behavior and the priors for the range parameters are set relative to the size of their domains. Noteworthy, spatial random effects exist at least in both the location parameters and transformed standardized residuals, without any constraints, which may cause some identifiability issues regarding the inference about these spatial ranges. Therefore, we impose a restriction as follows:  $\phi_\mu > \phi_z$ , where  $\phi_\mu$  and  $\phi_z$  are the spatial range parameters in the location parameter model, the scale parameter model and the standardized residuals model respectively. In practice, it is plausible to assume that large scale behavior of a particular extreme phenomenon is captured by the location parameters. And small scale dependence in space is captured by the residuals.

Posterior inference for the model parameters is completed using Gibbs samplers (Gelfand and Smith (1990)) and Metropolis Hasting updating (Gelman (2004)).

Given  $\mu(\mathbf{s}, t)$ , we directly sample  $\boldsymbol{\beta}$ ,  $\rho$ ,  $\mathbf{V}_2$  and off-diagonal entries in  $\mathbf{A}$  by adopting conjugate normal priors. Parameters without closed form full conditional distributions are updated using Metropolis-Hastings.  $\xi$  is sampled by the random walk Metropolis-Hastings

with Gaussian proposals. The proposal distributions for  $\sigma_0$  and diagonal entries in  $\mathbf{A}$  are truncated normals centered at the current samples. The conditional distribution of the latent spatial component  $\{\boldsymbol{\lambda}\}$  and  $\{\mathbf{z}\}$  follow a high dimensional non-Gaussian distribution, which is computationally challenging and extremely inefficient to follow the metropolis hasting path because of the correlations among individual spatial variables for  $\{\lambda(\mathbf{s})\}$  and  $\{z(\mathbf{s})\}$ . In the subsequent subsection, we introduce a gradient based algorithm to help the mixing of MCMC for drawing sample from those two high-dimensional latent variables.

Bayesian posterior computation is easy if we only focus on the Gaussian process model because the integration of the Gaussian process out of the model has a nice analytical solution (see, Banerjee et al. (2004)). However, in our situation, Gaussian processes models are embedded in the latent stage of the hierarchical model which has non-Gaussian first stage model. As a result, implementation of extreme value models in (3.14) essentially requires sampling values for the latent Gaussian random effects, which, unfortunately, cannot be integrated out of the model analytically. And the large number of spatial observations inevitably leads to high dimensional and highly correlated realizations of the latent Gaussian process, which typically yields poor convergence and slow mixing of MCMC when we sample from those latent Gaussian random effects.

Here, we can employ a version of the predictive process model described in Banerjee et al. (2007) (see Introduction 1.2.4 as well) to reduce the dimension. The latent Gaussian processes in the hierarchical models are now replaced with their induced predictive processes correspondingly. Briefly, we consider a set of “knots”  $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$  which forms a subset of the study region in 2 dimensional space. Take  $z_t^*(\mathbf{s})$  as an example, the Gaussian process with covariance function  $C_z(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_z)$  would yield  $\dot{z}_t^* = [\mathbf{z}_t^*(\mathbf{s}_i^*)]_{i=1}^m$  as its realizations over  $\mathcal{S}^*$ . Then the induced predictive process model is defined as:

$$\tilde{z}_t^*(\mathbf{s}) = \mathbf{c}_z^T(\mathbf{s}; \boldsymbol{\theta}_z) C_z^{*-1}(\boldsymbol{\theta}_z) \dot{z}_t^*, \quad (3.19)$$

where  $\mathbf{c}_z^*(\mathbf{s}; \boldsymbol{\theta}_z) = [C_z(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta}_z)]_{j=1}^m$ . and  $C_z^*(\boldsymbol{\theta}) = [C_z(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i=1, \dots, m; j=1, \dots, m}$ . In fact, we only need to work with  $\dot{z}_t^*$  which is in a relatively low dimension determined by the number of knots, say, 100. Therefore, predictive processes models can effectively reduce computational burdens of drawing samples from latent Gaussian random effects.

However, it is still a computationally challenging task to sample from a high dimensional (say, dimension  $> 50$ ) latent Gaussian random effects model. Here, we employ a gradient information based algorithm hoping to help the mixing of MCMC. The algorithm is called Metropolis-adjusted Langevin Algorithm (MALA), also known as Langevin-Hastings, which uses the gradient of the posterior distribution in making proposal distributions (Christensen and Waagepetersen (2002); Robert and Casella (2004)). We take the MALA sampling method for  $\dot{z}_t^*$  as an example to illustrate this algorithm.

Conditional on the other parameters, the posterior conditional distribution of  $\dot{z}_t^*(\mathbf{s}_1), \dots, \dot{z}_t^*(\mathbf{s}_m)$  is given by:

$$\pi(\dot{z}_t^* | \mathbf{Y}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\theta}_z) = f(Y | \dot{z}_t^*, \bar{\boldsymbol{\Omega}}) G(\dot{z}_t^* | \boldsymbol{\theta}_z) \quad (3.20)$$

where  $\bar{\boldsymbol{\Omega}}$  denotes all the other parameters except for  $\dot{z}_t^*$ ,  $f(Y | \dot{z}_t^*, \bar{\boldsymbol{\Omega}}, \boldsymbol{\theta}_z)$  is the likelihood function of  $\dot{z}_t^*$  and  $G(\dot{z}_t^* | \boldsymbol{\theta}_z)$  is the Gaussian process prior for  $\dot{z}_t^*$ .

In the MCMC algorithm, at iteration  $k$ , conditional on the samples we obtained for  $\dot{z}_t^{*(k-1)}$  and  $\bar{\boldsymbol{\Omega}}^{(k-1)}$  at the previous step. The Langevin proposal distribution for  $\dot{z}^*$  is

$$\dot{z}^* \sim N(\dot{z}^{*(k-1)} + \frac{v_{\dot{z}^*}}{2} \nabla(\dot{z}^*), v_{\dot{z}^*}^2 I) \quad (3.21)$$

where

$$\nabla(\dot{z}^*) = \frac{\partial}{\partial \dot{z}^*} \lg \Pi(\dot{z}^* | \mathbf{Y}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\theta}_z) \quad (3.22)$$

$v_{\dot{z}^*}$  is the proposal variance which controls the acceptance ratio. There are some studies to allow for an adjustable  $v_{\dot{z}^*}$  along with sample iterations.

The Hasting ratio is given by:

$$\frac{\exp(-\frac{1}{2v_{\dot{z}^*}^2}(\dot{z}^{*(k-1)} - (\dot{z}^* + \frac{v^2}{2} \nabla(\dot{z}^*)))^T(\dot{z}^{*(k-1)} - (\dot{z}^* + \frac{v^2}{2} \nabla(\dot{z}^*))))}{\exp(-\frac{1}{2v_{\dot{z}^*}^2}(\dot{z}^* - (\dot{z}^{*(k-1)} + \frac{v^2}{2} \nabla(\dot{z}^{*(k-1)})))^T(\dot{z}^* - (\dot{z}^{*(k-1)} + \frac{v^2}{2} \nabla(\dot{z}^{*(k-1)}))))} \quad (3.23)$$

One drawback of this derivative-based method is that one cannot always obtain a closed form for the derivative of the process values. See (Christensen et al. (2003) and Stramer and Tweedie (1999)) for more discussions on this issue.

### 3.5 A simulation study of GEV model for spatio-temporal point-referenced data

In this section we present a simulation study designed to examine the performance of the transformed Gaussian process model relative to the non-smoothed spatial GEV models. For simplicity, we do not consider extremes in both space and time. We only work on spatial extreme values in this study. Nevertheless, the spirit of this simulation design and model fitting can be generalized to the case of spatial-temporal extremes. The set up of the study is as follows. A set of 1200 locations is sampled over a  $[0, 10] \times [0, 10]$  rectangle, in which 300 points are sampled uniformly and 900 points are sampled at a regular lattice as shown in Figure 3.4. Let  $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))$  denote a response vector, where  $n = 85$ .  $\mathbf{Y}$  is obtained from the following GEV model:

$$y(\mathbf{s}) = \mu(\mathbf{s}) + \frac{\sigma}{\xi}(z(\mathbf{s})^\xi - 1) \quad (3.24)$$

$$\mu(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + w_\mu(\mathbf{s}) \quad (3.25)$$

$$z(\mathbf{s}) = G^{-1}\Phi(z^*(\mathbf{s})) \quad (3.26)$$

where  $\mu(\mathbf{s})$  is the location parameter at location  $s$ ,  $\sigma$  is the scale parameter and  $\xi$  is the shape parameter. We admit here that the assumptions of having an constant  $\sigma$  and

constant  $\xi$  over space may not be plausible in real data analysis. Nevertheless, the major aim of doing this simulation is to study the dependence structures in GEV 'residuals'. Moreover, model fitting with spatially structured  $\sigma(\mathbf{s})$  or  $\xi(\mathbf{s})$  may not be feasible here. Therefore, margins of the simulated extreme values are specified with Fréchet distributions with constant scale parameters and shape parameters over the study domain.  $\mu(\mathbf{s})$  is partitioned into a covariate term and spatial random effects  $w_\mu(\mathbf{s})$ .  $X(\mathbf{s})$  are the explanatory covariates at location  $\mathbf{s}$  with  $\boldsymbol{\beta}$  as its corresponding coefficient. Realizations of  $w_\mu(\mathbf{s})$  are obtained from a Gaussian process with exponential covariance function  $\sigma_\mu^2 \rho(\phi_\mu)$  with range parameter  $\phi_\mu = 4$  and scale parameter  $\sigma_\mu = 1$ . We set  $\beta = 10$ ,  $\xi = 0.5$ ,  $\sigma = 3$ . In this illustrative simulation, we only include a constant intercept in the covariate part, i.e.,  $X(\mathbf{s}) = 1$ .  $z(\mathbf{s})$  is sampled from a transformed standard Gaussian process. Marginally,  $z(\mathbf{s}, t) \sim GEV(-1, 1, 1)$ . Let  $\mathbf{z}^* = (z_1^*, \dots, z_n^*)$ .  $\mathbf{z}^*$  is generated as realizations of a Gaussian process with exponential correlation function  $\rho(\phi_z)$  and standard spatial scale parameter at observed location set  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ , where  $\phi_z = 1.4$  is the range parameter in the residuals.  $z(\mathbf{s})$  is then immediately obtained using (3.11). The simulated surface of  $Y(\mathbf{s})$  is shown in Figure 3.5.

We hold out 900 locations at regular lattice for validation purpose and fit Model (3.26) using the rest of the simulated extreme value at those 300 uniformly sampled locations. We also fit the following model using the same data set to make comparison with (3.26):

$$y(\mathbf{s}) = \mu(\mathbf{s}) + \frac{\sigma}{\xi} (z(\mathbf{s})^\xi - 1) \quad (3.27)$$

$$\mu(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + w_\mu(\mathbf{s}) \quad (3.28)$$

$$z(\mathbf{s}) \sim GEV(-1, 1, 1) \quad i.i.d \quad (3.29)$$

The model settings in Model (3.29) are similar with those of (3.26) in terms of the specifications for  $\{\mu(\mathbf{s})\}$ ,  $\sigma$  and  $\xi$ . The only difference lies in the distinct modeling approaches for the standardized residuals  $\{z(\mathbf{s})\}$ . In Model (3.29),  $\{z(\mathbf{s})\}$  is assumed to be independent

among different sites. The reason of making comparisons between Model (3.26) and (3.29) is to investigate whether the spatial signals in the standardized residuals can be successfully detected by fitting the Model with smoothed residuals, or in contrast, the spatial signals in the standardized residuals are difficult to be detected and hence mistreated as noises.

Prior distributions are assigned to model parameters to complete the Bayesian specification. We briefly list the prior specifications for a selected list of parameters and hyperparameters.

$$\begin{aligned}\pi(\xi) &\sim N(\mu_\xi^0, v_\xi^0) \\ \pi(\sigma_0) &\sim IG(2, b) \\ \pi(\beta) &\sim N(\mu_\beta^0, \sigma_\beta^0) \\ \pi(\sigma_\mu^2) &\sim IG(2, b_\mu)\end{aligned}$$

Assuming exponential covariance functions for  $w(\mathbf{s})$  and  $z(\mathbf{s})$  in the hierarchical model, we seek to make inference of the range parameters  $\phi_z$  and  $\phi_\mu$  respectively. We assume a uniform prior for  $\phi_\mu$  at  $[0, 5.8]$ . Conditional on  $\phi_\mu$ , the prior distribution for  $\phi_z$  is assumed to be  $U[0, \phi_\mu]$  which reflects our belief on the ratio of the spatial scale in  $z^*(\mathbf{s})$  relative to the spatial scale in  $w_\mu(\mathbf{s})$ . Because of computational concern, we discretize the prior distribution for each of them, i.e.,  $\phi_\mu$  and  $\phi_z$  only take values on finite discrete sets with equal weights, respectively.

Posterior inference for the model parameters is implemented by model fitting with Gibbs samplers and Metropolis Hasting. In (3.25), we introduce i.i.d white noise  $\epsilon(\mathbf{s}) \sim N(0, \tau^2)$  for computational convenience. By fixing a small value for variance  $\tau^2$  we enjoy the benefit of a Gaussian framework in sampling the parameters in the second stage specification with conditionally independent  $\mu_{\mathbf{s}}$ . Conditional on  $\{\mu(\mathbf{s})\}$  for each  $s$  and some other parameter, the posterior conditional distributions for  $\beta$ ,  $\rho$  and  $\{w_\mu(\mathbf{s})\}$  are all Gaussian distributions

when we adopt conjugate Gaussian priors.  $\sigma$  and  $\xi$  have to be drawn from their conditional distributions via Metropolis Hasting steps respectively.  $\mathbf{z}^*$  is the most difficult random vector to make inference via MCMC. We adopt the gradient-based MALA method to draw samples from its posterior multivariate non-Gaussian distribution. For each model, we ran 2000000 iterations to collect posterior samples after a burn in period of 100000 iterations, thinning using every fifth iteration. Trace plots of parameters indicate good convergence of the respective marginal distributions. The long length of burn in period is mainly caused by the slow mixing of  $\mathbf{z}^*$ . As a result of that, the smoothed GEV model (3.26) required longer time to complete the MCMC sampling than the non smoothed GEV model (3.29).

The subsequent results focus on the parameter estimates. Table 3.1 displays the posterior means for the parameters and the corresponding 95% credible intervals under Model (3.26) and Model (3.29) respectively. All 95% credible intervals cover the true parameter values for the Model we simulated from. In particular, the point estimation for  $\phi_z = 1.5064$  and  $\phi_\mu = 2.6658$ . We display the histograms of the posterior samples of  $\phi_z$  and  $\phi_\mu$  in Figure 3.6. In Model (3.29), we only have spatial random effects which are specified in the location parameters. The point estimation of  $\phi_\mu$  under this model is 2.8069. Figure 3.7 shows the histogram of  $\phi_\mu$  we obtain by Model (3.29).

Recalling that 900 locations were held out for validation purpose, we now judge the performance of these models based on: Deviance Information Criterion (DIC) (Spiegelhalter et al. (2002)); ability to recover the *true* parameter values; prediction of a holdout set of 900 locations. In specific, adopting the posterior medians as the point estimates of predicted maxima at new locations, we plot the predicted values for all hold out sites in Figure 3.8. In addition, we check the performance of prediction by computing the averaged absolute predictive errors (*AAPE*) of each model. Given the true value of  $Y_s$  in hold out data set,

$AAPE$  is given by:

$$AAPE = \frac{1}{900} \sum_{i=1}^{900} |\hat{Y}(\mathbf{s}) - Y(\mathbf{s})| \quad (3.30)$$

Where  $\hat{Y}(\mathbf{s})$  is the median of the posterior samples  $(\hat{Y}(\mathbf{s})^{(t)})$  from its posterior predictive distribution.

Table 3.2 provides the DIC model fit statistics for Model (3.26) and Model (3.29). The first column, labeled DIC, provides a relative measure of goodness of model fit, with lower values indicating better fit (see., e.g., Spiegelhalter et al. (2002) for a full explanation of the criterion). Model (3.26) has a slightly smaller DIC than Model (3.29), which indicates the better fit using model with smoothed residuals. And the prediction performance for Model (3.26) behaves better than of Model (3.29) based on the AAPE score. Notice that extreme values with positive shape parameters are heavy tailed. We display the predictive median surface for the extreme values as well as the associated surface for the lengths of the 95% predictive intervals in Figure 3.8 for (3.26) and (3.29), respectively. Encouragingly, the predictive surface for the extreme values based on Model (3.26) reasonably recovers the true surface for the simulated extreme values. It also successfully captures the local spatial patterns in the simulated extreme values. In contrast, the predictive median surface we obtained from Model (3.29) could only reflect a large scale spatial dependence and fails to capture the local peaks. In addition, Model (3.29), on average, has a larger length of 95% predictive interval compared with Model (3.26).

## 3.6 Discussion

We have presented a continuous spatial process of extreme values, which we call the transformation of a Gaussian process. The joint distribution function of realizations of the transformed Gaussian process at any given set of locations can be expressed explicitly. We

apply the proposed transformed Gaussian processes in the hierarchical models for explaining the dynamic and spatial change in annual maximum observations. The first stage of our hierarchical models follows GEV distribution to describe the asymptotic distributions of maxima taken from a time series of daily records. The second stage of our hierarchical models specifies variations in time and large scale dependence in space for the parameters in the GEV distributions. We model the "residuals" in the GEV models using the transformed Gaussian processes to capture the small scale dependence in space. In addition, we offer an approach to make spatial interpolation for extreme values based on this hierarchical models with smoothed residuals across space.

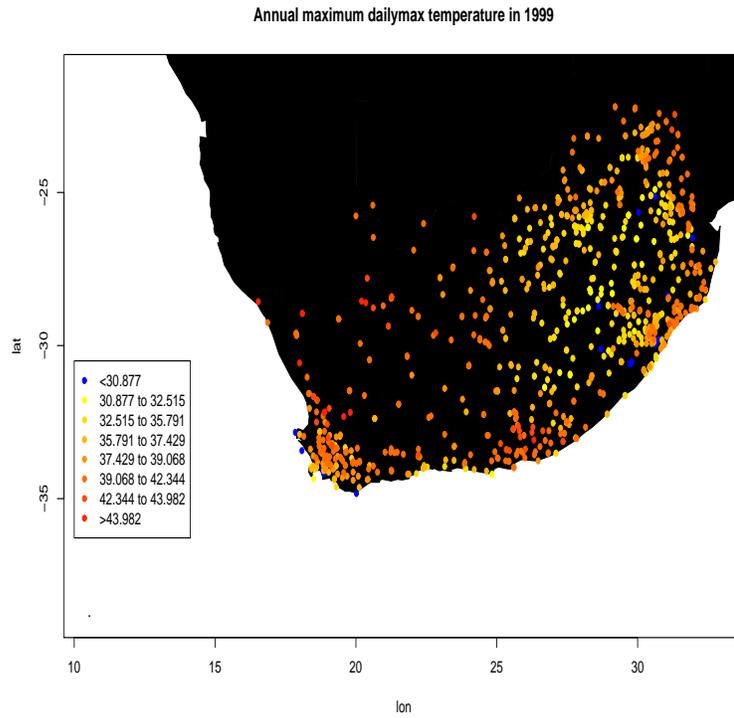
Various future extensions for this work are possible. Besides simulation studies, we are aiming to apply this modeling approach to our motivating dataset, annual temperature extremes in the CFR. We also have interests to using this approach as spatial interpolation tool to produce risk maps of extreme values. For example, the risk of having annual maximum of daily highest temperatures greater than  $40^\circ$ . Furthermore, it is important to develop some easily implemented diagnostic tools in the exploratory stage of extreme value analysis, which could help us to detect multiscale spatial patterns in extreme values. Finally, we are seeking to find computationally more efficient algorithm to implement these complex hierarchical models with high dimensional non-Gaussian first stage models.

	True	Model 3.26 (Smoothed residuals)		Model 3.29 (Non-smoothed residuals)	
		Mean	95% CI	Mean	95% CI
$\beta$	10	9.7684	(9.1094, 10.4622)	9.7146	(9.3444, 10.5167)
$\sigma_0$	3	3.0229	(2.6740, 3.3006)	2.9209	(2.7218, 3.1725)
$\xi$	0.5	0.4253	(0.3522, 0.5493)	0.4162	(0.3619, 0.5331)
$\phi_\mu$	4	2.6658	(0.4, 5.6)	2.8069	(0.4, 5.6)
$\phi_z$	1.4	1.5074	(0.2, 3.6)		

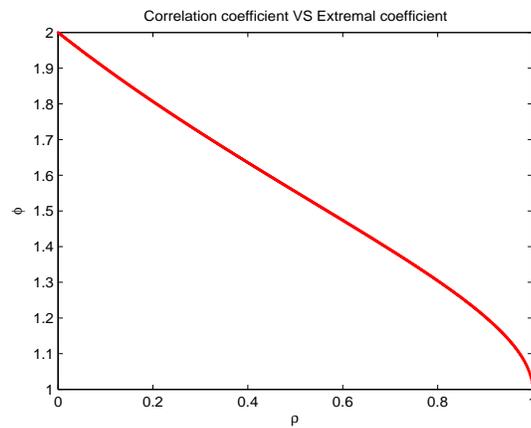
**Table 3.1:** Posterior sample means of parameters and the corresponding 95% credible intervals for Model (3.26) and Model (3.29).  $\beta$  is the intercept.  $\sigma$  is the scale parameter of GEV and  $\xi$  is the shape parameter of GEV;  $\phi_\mu$  is the spatial range parameter in the location parameters and  $\phi_z$  is the spatial range parameter in the GEV residuals

	DIC	AAPE	$\hat{r}$
Smoothed GEV	322.27	2.6474	0.951
Nonsmoothed GEV	325.24	3.8549	0.956

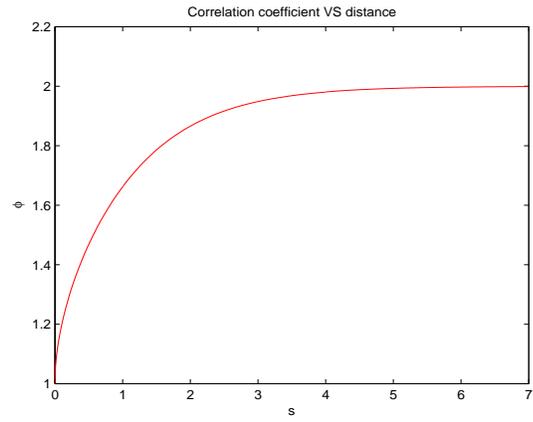
**Table 3.2:** Performance of Model (3.26) and Model (3.29) using averaged absolute predictive errors (AAPE), the deviance information criterion (DIC), and the empirical coverage probability  $\hat{r}$ .



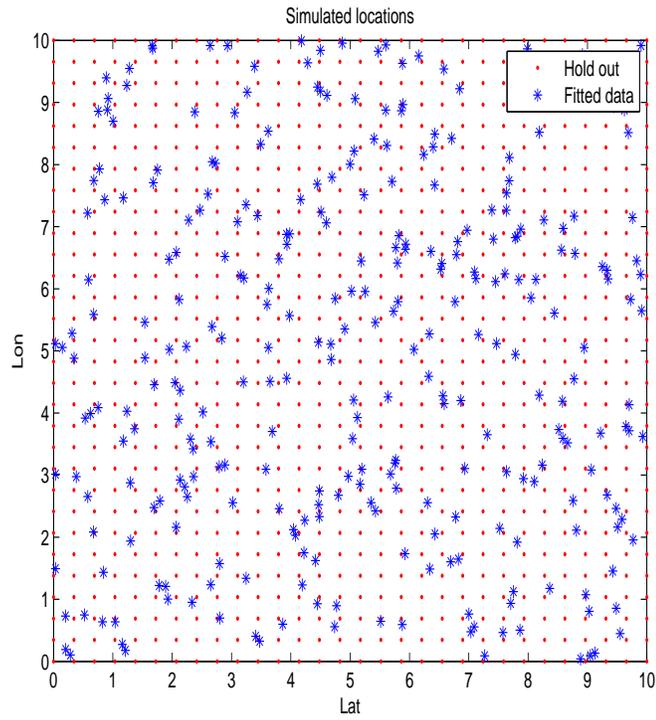
**Figure 3.1:** Image plot of the temperature maxima over CFR in 1999; Spatial signal observed



**Figure 3.2:** The extremal coefficient for the bivariate Gaussian copula extreme value distribution as a function of distance.



**Figure 3.3:** The extremal coefficients for the bivariate Gaussian copula under different correlation coefficients.



**Figure 3.4:** Simulated locations

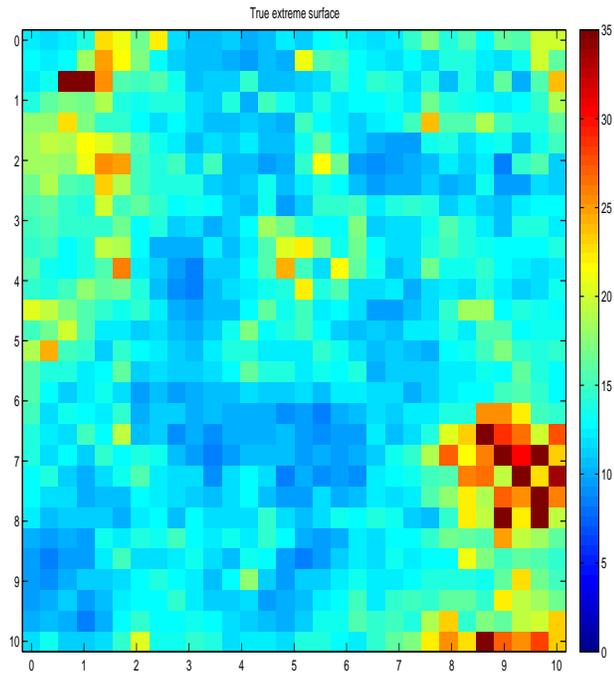
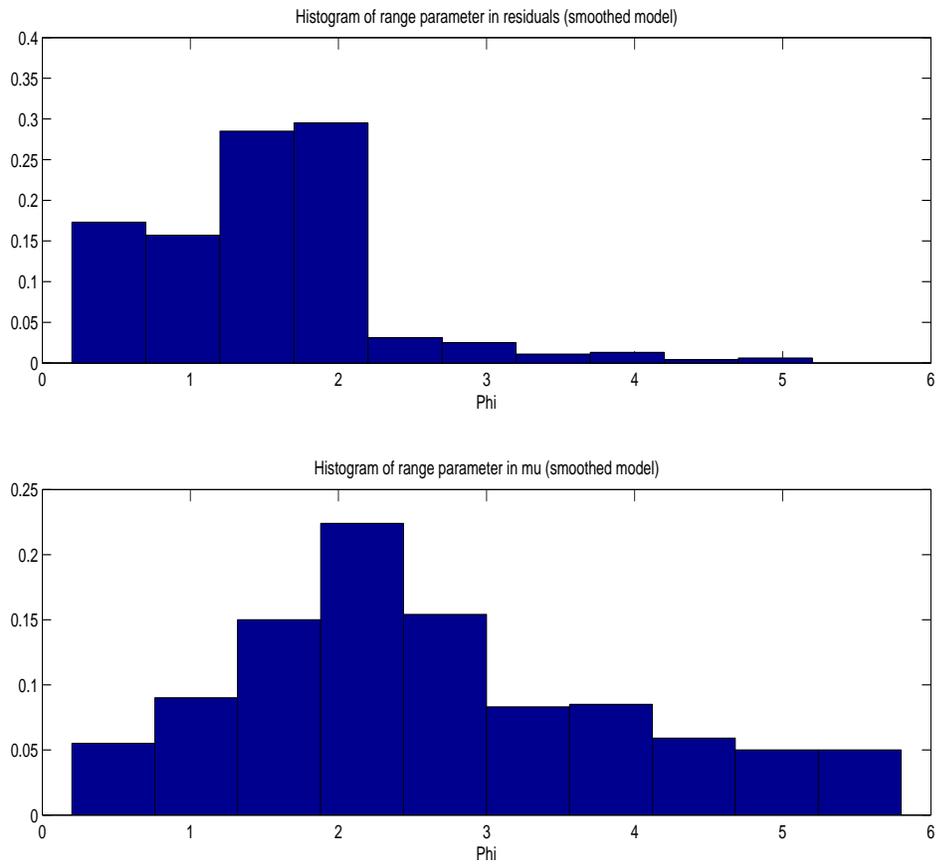
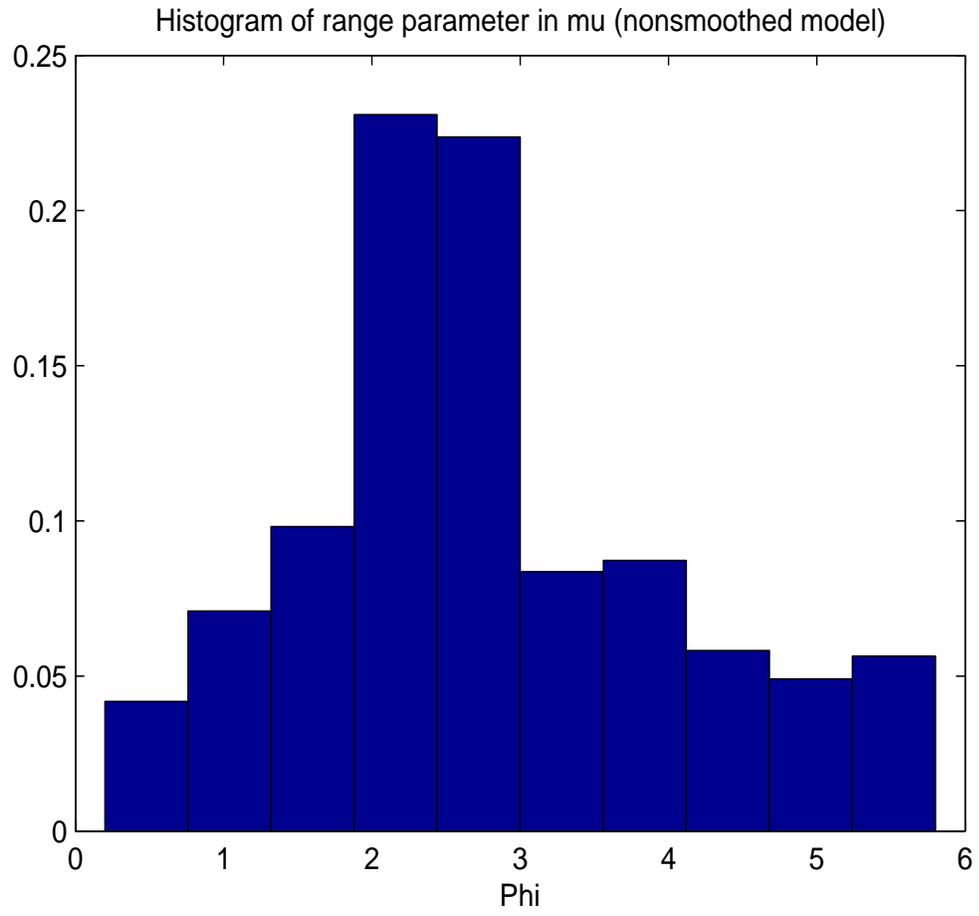


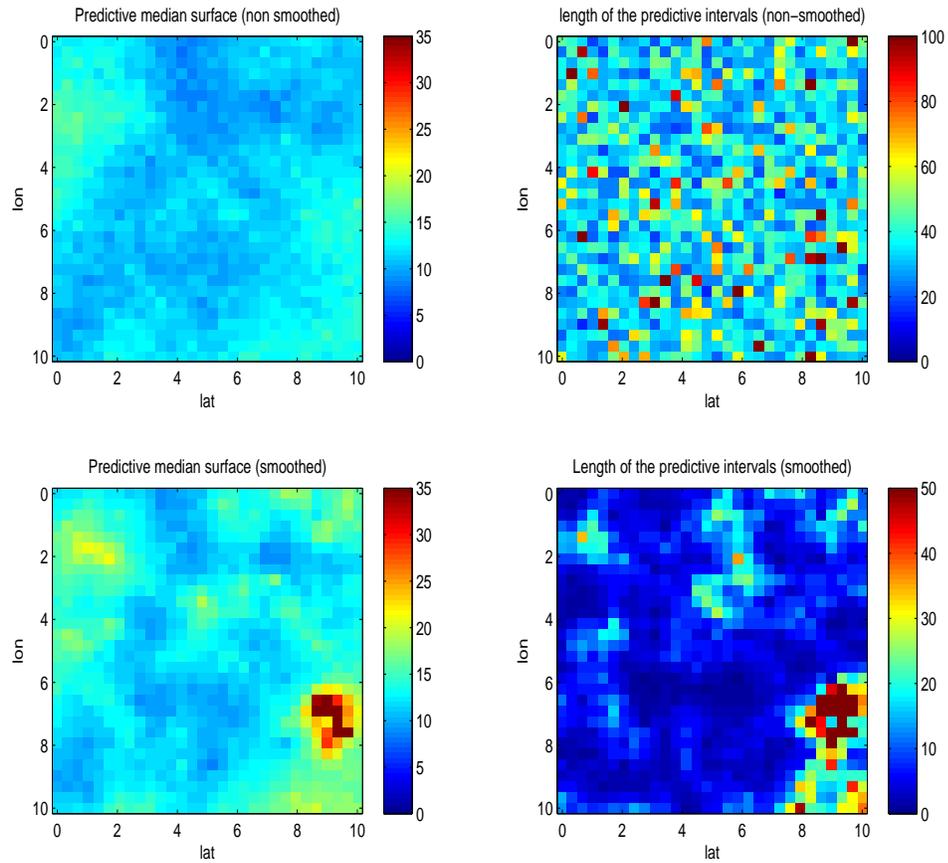
Figure 3.5: Simulated surfaces



**Figure 3.6:** Histogram of the range parameter  $\phi_z$  and the range parameter  $\phi_\mu$  obtained by the smoothed model



**Figure 3.7:** Histogram of the range parameter  $\phi_\mu$  obtained by the nonsmoothed model



**Figure 3.8:** From left to right, top to bottom: (a) The predictive median surface for extreme values using Model (3.26) (smoothed residuals); (b) The lengths of the 95% predictive intervals for Model (3.26) (smoothed residuals); (c) The predictive median surface for extreme values using Model (3.29) (non-smoothed residuals); (d) The lengths of the 95% predictive intervals for Model (3.29) (non-smoothed residuals)

# Chapter 4

## Bayesian Analysis of Multiscale Spatial Dependence

### 4.1 Introduction

Spatial statistical models are gaining increasing amount of interests in the past few decades partly thanks to the recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS). Given a collection of large spatial dataset and the corresponding accurate geocoding of locations, researchers and resource analysts are typically interested in modeling how variables are associated both within and across locations. Specific interest lies in learning the spatial structures which usually correspond to physical features of the environment or to intrinsic characteristics of environmental processes and phenomena.

One of the important aspects of spatial statistical models is to detect and identify spatial scales. The concept of spatial scale is fundamental to geography (see e.g., Marceau and Hay (1999), Atkinson and Tate (2000) Dungan et al. (2002) and Quattrochi and Goodchild (1997)). It has been broadly applied in geographic study, but, with various meanings. Bian (1997) summarized the definition of scale from four perspectives. The first connotation is that of traditional *cartographic scale* or *map scale*; The second connotation is that of *geographic scale*, referring to the spatial extent of a study region. For instance, the development of ozone surfaces for the Atlanta metropolitan area and the North American continent would be considered small and large geographical scale studies, respectively. The third connotation of scale is *measurement scale*, referring to the sampling interval used in a study. In the case of regular lattice measurements, it often refers to data resolution. In

the case of point measurements, it is often to refer to the mean distance to the nearest neighbors for each point or the minimum distance to the nearest neighbors for each point. And the fourth connotation is *operational scale*, referring to the spatial extent at which a particular phenomenon operates in the environment. For instance, a phenomenon is called to have a fine operational scale if its surface is dominated by fine scale variations. And operational scale and spatial complexity are usually inversely related.

Considerable amount of attentions in statistical literature have been focused on the inference of operational scales by building models to examine the collected spatial measurements (see, e.g., Cressie (1993) and Banerjee et al. (2004)). One commonly used strategy is to represent operational scale in the parameterization of process models. For example, a range parameter is defined in the commonly used exponential covariance function (or exponential covariance variogram) used in point-level spatial Gaussian process specifications. The range is intended to quantify the effective operational range associated with a particular spatial dataset.

In reality, statistical inference for the observational scale of a particular environmental phenomenon is considered to rely on the measurement scale in the geographical study (Cao and Lam (1997)). For example, to study the abundance pattern of invasive species from one county to its neighboring counties, it is suggested that one should monitor the population change of invasive species at a county level. One may possibly observe different abundance patterns working at different spatial scales. Like the abundance pattern for invasive species, in reality, many environmental processes and patterns varies with scale (Quattrochi and Goodchild (1997)). And those geographical phenomena are regarded as scale-dependent. Depending on the observation scale, scale-dependent processes that appear homogeneous at a fine scale may become heterogeneous at a large scale, and parameters and factors that are important at one scale may become trivial at another. However, in practice, the geo-

graphical data are often collected only at a single pre-specified measurement scale which is subject to observer's choice. Based on the collected single measurement scale spatial dataset, most existing statistical methods only assume a spatial model with a single operational scale. Those methods may lead to two important issues regarding the learning of true operational scale of the spatial phenomenon. The first issue is that if the spatial phenomena are scale dependent, say, with a spatial pattern at a state level along with a spatial pattern at a county level, a single scale spatial model will fail to capture the two distinct spatial patterns. There have been little research work on studying multiscale spatial dependence (see, e.g., Banerjee and Finley (2007)). The second issue is that if a single scale of dependence operates, the measurement scale of the study may not be fine enough to discover the actual operational scale of the geographical phenomenon (e.g, when the operation scale is significantly smaller than the measurement scale), then a single scale spatial model may not be able to capture any useful spatial signals. In particular, this second issue is sometimes referred to as the microscale problem in statistical literature (see, e.g., Cressie (1993)). Cressie (1993) provided a definition of the micro scale variation in the context of stationary spatial process: an intrinsically stationary process is called microscale variation if its variogram range exists and is smaller than the minimum interlocation distances. Little work has been done to address these two issues. In fact, Cressie (1993) addressed that: " In the absence of information beyond the  $n$  data and their spatial locations, the behavior of the microscale component in the decomposition is, by definition, unobtainable. But knowing this behavior is important (for both predictions and mean squared prediction errors) when a predictor location  $s_0$  is close to an observation location  $s_i$ ." However, the definition of microscale spatial process given by Cressie actually encourages a Bayesian approach in which priors on microscale spatial range could be appropriately specified to reflect its association with the observational scale as well as practitioners' opinions.

The contribution of this paper is to propose a spatial point process model that aims to account for both the large scale spatial dependence and the fine scale spatial dependence given the spatial dataset collected at a single measurement scale. We offer an Bayesian method to detect and distinguish the multi-scale spatial dependence structures by effectively incorporating weak prior information for each of the spatial scales, which connects the notion of coarse and fine scale dependence relative to the scale at which data are collected. In particular, our model is able to tackle the problem associated with microscale spatial dependence when the given measurement scale in the spatial data is larger than the true micro spatial scale. Moreover, spatial study with multiscale spatial dependence often involves large number of observations. The reason is that, by definition, the scale at which data are collected should not be too coarse relative to the fine spatial operational scale, which often suggests spatial study at a relatively high resolution. Otherwise, local spatial variations may be indistinguishable from noise. In addition, spatial data with coarse scale of dependence are typically collected at a relatively large global geographical scale. We will also provide several solutions to the computational problems involved in fitting these spatial models.

The remainder of this Chapter evolves as follows. Section 4.2 reviews the conventional spatial point-referenced process, and then introduces our proposed spatial point-referenced process with coarse scale and fine scale spatial components. Section 4.3 describes the computational strategies to deal with coarse scale and fine scale structures. Section 4.5 illustrates the proposed methods with a simulated dataset. Finally, Section 4.6 concludes with a brief discussion including future work.

## 4.2 Hierarchical modeling for multi-scale spatial data

In the introduction section, we provide a brief review of the univariate spatial Gaussian processes. Assume we observe a response or dependent variable  $Y(\mathbf{s})$  at a generic location  $\mathbf{s} \in D \subseteq R^2$  along with a  $p \times 1$  vector of spatially referenced predictors  $\mathbf{x}(\mathbf{s})$ . Then, model-based geostatistical data analysis typically proceeds from spatial regression models such as,

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (4.1)$$

The residual from the regression is partitioned into a spatial process,  $w(\mathbf{s})$ , capturing spatial association, and an independent process,  $\epsilon(\mathbf{s})$ , also known as the *nugget* effect, modelling measurement error(see, e.g., Chiles and Delfiner (1999)). The  $w(\mathbf{s})$  are often referred to as spatial random effects, providing local adjustment (with structured dependence) to the mean, interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern. The most common specification for  $w(\mathbf{s}) \sim GP(0, C(\cdot, \cdot))$  is a zero-centered Gaussian Process determined by a valid covariance function  $C(\mathbf{s}_i, \mathbf{s}_j)$  defined for pairs of sites  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . Sometimes we specify  $C(\mathbf{s}, \mathbf{s}') = \sigma^2 \rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$  where  $\rho(\cdot, \cdot; \boldsymbol{\theta})$  is a correlation function and  $\boldsymbol{\theta}$  includes parameters quantifying rate of correlation range and smoothness of realizations though this limits us to constant process variance.  $\epsilon(\mathbf{s})$  is often assumed to follow a normal distribution with variance  $\tau^2$  i.i.d for every location  $\mathbf{s}$ . In practice, it is a simple but reasonable assumption to model measurement errors.

In some situations,  $\epsilon(\mathbf{s})$  is viewed as microscale variability. As defined in the introduction, this refers to the variability at distances smaller than the smallest interlocation distance in the data. Therefore,  $\epsilon(\mathbf{s})$  is a spatial process which has a very fine operational scale compared with its measurement scale (see, Cressie (1993)). And the spatial depen-

dence of  $\epsilon(\mathbf{s})$  decays extremely rapidly. Clearly, the  $N(0, \tau^2)$  assumption for  $\epsilon(\mathbf{s})$  in (4.1) suggests lack of inference power to discern between measurement error and micro-scale variation. However, not much work have been done to investigate the micro-scale spatial variability.

For a given spatial dataset which can be explained by several unmeasured or unobserved covariates with spatial patterns at different scales, we seek to develop modeling approach which could capture and distinguish each of these spatial patterns. To illustrate, we consider a spatial Gaussian model with spatial variations with spatial dependence at two different scales, simply called "coarse scale variation" and "fine scale variation".

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w_{coarse}(\mathbf{s}) + w_{fine}(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (4.2)$$

The residual from the regression is now partitioned into three components: a coarse scale spatial process,  $w_{coarse}(\mathbf{s})$ , capturing coarse spatial association; a fine scale spatial process,  $w_{fine}(\mathbf{s})$ , capturing fine spatial association, and an independent or pure error process,  $\epsilon(\mathbf{s})$ , modelling measurement errors. Again,  $w_{coarse}(\mathbf{s})$  can be interpreted as the unobserved covariates with large operational scale spatial pattern. And  $w_{fine}(\mathbf{s})$  can be interpreted as a surrogate for the unobserved covariates with fine operational scale spatial pattern.

We consider  $w_{sum}(\mathbf{s}) = w_{coarse}(\mathbf{s}) + w_{fine}(\mathbf{s})$ , which is viewed as the overall spatial variation at all scales. Then for any two locations  $\mathbf{s}$  and  $\mathbf{s}'$ , the covariance  $w(\mathbf{s})$  and  $w(\mathbf{s}')$  is given as follows:

$$C_{sum}(\mathbf{s}, \mathbf{s}') = C_{coarse}(\mathbf{s}, \mathbf{s}') + C_{fine}(\mathbf{s}, \mathbf{s}') \quad (4.3)$$

First, suppose we model  $w_{coarse}(\mathbf{s})$  and  $w_{fine}(\mathbf{s})$  through isotropic stationary Gaussian process models. Assume  $w_{coarse}(\mathbf{s})$  is a zero-centered Gaussian Process with covariance function  $C_{coarse}(\mathbf{s}, \mathbf{s}')$  and  $w_{fine}(\mathbf{s})$  is a zero-centered Gaussian Process with covariance function  $C_{fine}(\mathbf{s}, \mathbf{s}')$ . And we assume  $w_{fine}(\mathbf{s})$  and  $w_{coarse}(\mathbf{s})$  are independent. In appli-

cations, we often specify  $C(\mathbf{s}, \mathbf{s}') = \sigma^2 \rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$  where  $\rho(\cdot; \boldsymbol{\theta})$  is a correlation function and  $\boldsymbol{\theta}$  usually includes a decay and a smoothness parameters. Common choices for correlation functions include exponential, Gaussian, spherical and Matérn (see, introduction 1.2.1) correlation functions, all of which induce valid stationary isotropic Gaussian processes. The covariance function of  $w_{sum}(\mathbf{s})$  is then given by:

$$C_{sum}(\mathbf{s}, \mathbf{s}') = \sigma_{coarse}^2 \rho_{coarse}(\|\mathbf{s} - \mathbf{s}'\|; \boldsymbol{\theta}_{coarse}) + \sigma_{fine}^2 \rho_{fine}(\|\mathbf{s} - \mathbf{s}'\|; \boldsymbol{\theta}_{fine}) \quad (4.4)$$

Take the exponential correlation function as an example, Figure 4.1 is presented to compare the overall spatial correlation function of  $w(\mathbf{s})$  with the coarse scale spatial correlation function and the fine scale correlation function. For simplicity, we assume  $\sigma_{coarse} = \sigma_{fine}$ ,  $\phi_{coarse} = 10$  and  $\phi_{fine} = 1$ . Therefore,  $(\rho_{coarse} + \rho_{fine})/2$  is the correlation function for  $w_{sum}(\mathbf{s})$ . Clearly, when distance between two sites is large, the spatial correlation is primarily determined by the coarse scale spatial process. When distance between two sites is small, the single coarse scale spatial correlation function fails to capture the overall correlation pattern because of the existence of the fine scale spatial correlation. In many spatial statistical models, the fine scale pattern is often overshadowed and ignored because of the existence of coarse scale spatial variation.

Suppose we are working with the isotropic Gaussian process models. Parameters of interests in Model (4.2) are the range parameters  $\phi_{coarse}$  and  $\phi_{fine}$  as well as the spatial scale parameters for the coarse scale variation and the fine scale variation. Maximum likelihood methods are one of the conventional ways to make inference for the model parameters. For a given collection of sites in  $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  the data likelihood is given by  $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$ , with  $\Sigma_{\mathbf{Y}} = C_{sum}(\boldsymbol{\theta}) + \tau^2 I_n$ , where  $X = [\mathbf{x}^T(\mathbf{s}_i)]_{i=1}^n$  is a matrix of regressors and  $C_{sum}(\boldsymbol{\theta}) = [C_{coarse}(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}_{coarse})]_{i,j=1}^n + [C_{fine}(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}_{fine})]_{i,j=1}^n$  is the spatial covariance matrix associated with  $w_{sum}(\mathbf{s})$ . However, there are some identifiability problems with the parameter estimations in the covariance function. Even if we consider the

single spatial process model, the range parameter and the scale parameter are weakly identified in the likelihood. Zhang (2004) shows that in model-based geostatistics with Gaussian observations, one cannot distinguish between two Matérn correlation functions with probability 1 no matter how much sample data are observed in fixed region. Consequently, not all parameters are consistently estimable. Noticing the exchangeability between  $(\sigma_{fine}, \phi_{fine})$  and  $(\sigma_{coarse}, \phi_{coarse})$  indicated by equation (4.3). The parameters in coarse scale covariance and the parameters in fine scale covariance function are not identifiable without any imposed constraints.

We follow a Bayesian inference path, which is another commonly used method to make inference of the model parameters. We introduce hierarchical models, and assign prior (hyperprior) distributions to the model parameters (hyperparameters) and inference proceeds by sampling from the posterior distribution of the parameters (see, e.g., Banerjee et al. (2004)). A nice advantage of using Bayesian inference is that we can incorporate the prior knowledge about the spatial scales when into model settings. Again, suppose we are working with the isotropic Gaussian process models. Taking exponential correlation function as an example, the fact that coarse scale range parameter  $\phi_{coarse}$  is undoubtedly greater than the fine scale range parameter  $\phi_{fine}$  gives us some natural guidance about the prior assumption for  $\phi_{coarse}$  and  $\phi_{fine}$ . In specific, one possibility is to assume

$$\phi_{coarse} = K\phi_{fine} \tag{4.5}$$

where  $K$  is a ratio in  $[c_1, c_2]$  reflecting the modeler's (or practitioner's) prior belief on the difference between coarse operational scale (coarse spatial range) and fine operational scale (fine spatial range) for a particular geographical phenomenon of interests.  $c_1$  should be set greater than 1 in order to ensure  $\phi_{coarse} > \phi_{fine}$ . The constraint in (4.5) actually offers a reparameterization in Model (4.2). That is, we can replace the parameters  $(\phi_{coarse}, \phi_{fine})$  with the new parameters  $(K, \phi_{coarse})$ . Following this reparameterization, we assign prior

distributions on  $\phi_{coarse}$  and  $K$ , e.g.,  $K \sim U(c_1, c_2)$ ,  $\phi_{coarse} \sim U(0, \max_{i,j}(d_{i,j})/2)$ . Given posterior samples on  $K$  and  $\phi_{coarse}$ , inference of  $\phi_{fine}$  is immediately obtained by simply computing  $\phi_{coarse}/K$ .

Prediction at an arbitrary site  $\mathbf{s}_0$  is straightforward under Bayesian inference framework. Generically denoting the set of all model parameters by  $\boldsymbol{\theta}$ , model estimation computes  $[\Omega|\mathbf{Y}]$ . We seek to find the predictive distribution  $[Y(\mathbf{s}_0)|\mathbf{Y}]$  at a new site  $\mathbf{s}_0$ . This predictive distribution can be sampled by composition, drawing  $Y^{(l)}(\mathbf{s}_0) \sim [Y(\mathbf{s}_0)|\Omega^{(l)}, \mathbf{Y}]$  for each  $\Omega^{(l)}, l = 1, \dots, L$ , where  $\Omega^{(l)}$  is the  $l_{th}$  posterior samples. This is especially convenient for Gaussian likelihoods (such as (4.1)) with  $C_{sum}(\boldsymbol{\theta}) = \sigma_{coarse}^2 \rho_{coarse}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) + \sigma_{fine}^2 \rho_{fine}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$  since  $[Y(\mathbf{s}_0)|\Omega^{(l)}, \mathbf{Y}]$  is itself Gaussian with

$$E[Y(\mathbf{s}_0)|\Omega, \mathbf{Y}] = \mathbf{x}^T(\mathbf{s}_0)\boldsymbol{\beta} + \mathbf{h}^T(\mathbf{s}_0) (C_{sum}(\boldsymbol{\theta}) + \tau^2 I_n)^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \text{ and} \quad (4.6)$$

$$Var[Y(\mathbf{s}_0)|\Omega, \mathbf{Y}] = \left( \sigma_{coarse}^2 + \sigma_{fine}^2 - \mathbf{h}^T(\mathbf{s}_0) (C_{sum}(\boldsymbol{\theta}) + \tau^2 I_n)^{-1} \mathbf{h}(\mathbf{s}_0) \right) + \tau^2. \quad (4.7)$$

where  $\mathbf{h}(\mathbf{s}_0) = [C_{sum}(\mathbf{s}_0, \mathbf{s}_i; \boldsymbol{\theta})]_{i=1}^N$

We should remark that so far discussions on coarse/fine scale spatial processes model in this paper have been focusing on isotropic stationary Gaussian processes. In fact, the idea of employing Bayesian inference to reflect prior information on multi scale spatial dependence could be flexibly extended to a general spatial context. In fact, despite the fact that many spatial models do not necessarily have a clear notion of 'range parameter' as we see in isotropic stationary covariance functions, spatial scales are often reflected in those spatial models in some fashion. For example, below we will introduce two knots based spatial processes which are non-stationary to tackle computational problems with large spatial dataset. But as an illustration, we will also briefly explain how to reflect prior information in those cases.

## 4.3 Computations in coarse/fine spatial process model

In practice, the fine scale dependence problems receive attention when the data collection is at very high resolution so that it becomes feasible to investigate the scale of dependence. In addition, spatial data with coarse scale of dependence are typically collected at a relatively large global geographical scale. Therefore, geostatistical studies with multiple scales of dependence typically involve large data sets. Suppose we have a set of locations of size  $n$ , such fitting involves evaluation of the Gaussian likelihood with the  $n \times n$  matrix  $\Sigma_{\mathbf{y}}^{-1}$ . Markov chain Monte Carlo methods we described above is an effective approach for fitting such models when  $n$  is not large (say,  $n < 1000$ ). But when  $n$  is large, matrix computation complexity increases as  $O(n^3)$  in the number of locations  $n$ . Within an MCMC implementation, these computations occur at every iteration of the MCMC algorithm, rendering computational infeasibility for large datasets. This problem has been referred to as the “big  $n$  problem” in the spatial statistics literature. Below, we offer three approaches to tackle this problem: the predictive processes (Banerjee et al. (2007)) and kernel convolution(see, e.g.,Higdon (2002)), covariance tapering (Kaufman (2006)). In particular, we discuss the effect of operational scale on our choices among those different strategies for dealing with large ‘n’ spatial models.

### 4.3.1 ”Large n” problem in coarse scale spatial processes

The predictive processes method (see Banerjee et al. (2007)) and the kernel convolution method adopt the similar path in that both of these two methods replace the original process  $w(\mathbf{s})$  with an approximation  $\tilde{w}(\mathbf{s})$  that represents the realizations in a lower-dimensional subspace. Below, we briefly introduce these two methods and show how to employ them

into our models with coarse and fine scale spatial variations.

First, we briefly summarize the kernel mixing method as below. Kernel mixing method was mainly developed in Higdon (1998) Higdon et al. (1999) and Higdon (2002). It provides an attractive approach of introducing nonstationary spatial association while retaining clear interpretation and permitting analytic calculation. A particular kernel mixing model is in the form of a finite sum approximation. Consider a set of “knots”  $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$  which forms a subset of the study region in 2 dimensional space. Define

$$w(\mathbf{s}) = \sum_{m=1}^M k(\mathbf{s} - \mathbf{s}_m^*) z(\mathbf{s}_m^*) \quad (4.8)$$

Assume  $z(\mathbf{s}_m^*)$  has a standard normal distribution *i.i.d* for each  $m$  from 1 to  $M$ . It directly follows that:

$$Var(w(\mathbf{s})) = \sigma^2 \sum_{m=1}^M k^2(\mathbf{s} - \mathbf{s}_m^*) \quad (4.9)$$

$$Cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \sum_{m=1}^M k(\mathbf{s} - \mathbf{s}_m^*) k(\mathbf{s}' - \mathbf{s}_m^*) \quad (4.10)$$

It is noteworthy that Equation (4.10) clearly indicates the nonstationarity property constructed for  $w(s)$ .

An attractive feature of this finite sum approximation of a kernel mixing model is dimension reduction. As shown in (4.8), the set of  $\{z(\mathbf{s}_m^*)\}, m = 1, \dots, M$  serve as latent base variables, typically with the total number of points being smaller than the sample size  $n$ . Using this kernel mixing model (4.10), we can represent every given random variable  $w(\mathbf{s})$  as a linear combination of the set  $\{z(\mathbf{s}_m^*)\}, m = 1, \dots, M$ . Let  $K = [k_{i,m}, i = 1, \dots, n; m = 1, \dots, M]$ , where  $k_{i,m} = k(\mathbf{s}_i, \mathbf{s}_m^*)$ . In the case of a white noise assumption for each  $z(\mathbf{s}_m^*)$ , the variance-covariance function of  $\{w(\mathbf{s}), \mathbf{s} = \mathbf{s}_1, \dots, \mathbf{s}_n\}$  can be written as  $\Sigma = K^T K$ . This suggests that no matter how large  $n$  is, we are essentially breaking a  $n \times n$  matrix  $\Sigma$  down to a product of  $n \times M$  matrix  $K$  and its transpose.

This kernel mixing modeling approach can be embedded into our Model (4.2) easily. First we select a set of knots  $\{z(\mathbf{s}_m^*)\}, m = 1, \dots, M$  which covers the study domain. Then we model coarse operational scale spatial component as follows:  $w_{coarse}(\mathbf{s}) = \sum_{m=1}^M k(\mathbf{s} - \mathbf{s}_m^*)z(\mathbf{s}_m^*)$ .

Predictive process model is another approximated process  $\tilde{w}(\mathbf{s})$  that represents the realizations of  $w(\mathbf{s})$  in a lower-dimensional subspace. Banerjee et al. (2007) develop predictive process models for spatial and spatio-temporal data. The idea is that every spatial (or spatio-temporal) process could induces a predictive process, which projects the underlying original process onto a subspace generated by a set of realizations of the original process at a specified set of locations. We consider a set of ‘‘knots’’  $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$  which forms a subset of the study region in 2 dimensional space. The bivariate Gaussian process above would yield  $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim MVN(\mathbf{0}, C^*(\boldsymbol{\theta}))$  as its realizations over  $\mathcal{S}^*$ , where  $C^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$  is the corresponding  $m \times m$  covariance matrix. The predictive process model is defined as

$$\tilde{w}(\mathbf{s}) = E[w(\mathbf{s})|\mathbf{w}^*] = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*, \quad (4.11)$$

where  $\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$ . In fact,  $\tilde{w}(\mathbf{s})$  is a Gaussian Process with covariance function  $\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^{*T}(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}^*(\mathbf{s}'; \boldsymbol{\theta})$  where  $\mathbf{c}^*(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}_0, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$ . The realization of  $\tilde{w}(\mathbf{s})$  on any collection of sites are the interpolated predictions conditional upon the realization of  $w(\mathbf{s})$  over  $\mathcal{S}^*$ . To work with this process we only need to work with  $\mathbf{w}_1^*, \mathbf{w}_2^*$  and the associated pair of  $m \times m$  correlation matrices.

This predictive process approach provides a nice alternative to model the coarse operational scale spatial component. Given a selected set of knots  $\{z(\mathbf{s}_m^*)\}, m = 1, \dots, M$ , we model coarse operational scale spatial component following , i.e.,  $w_{coarse}(\mathbf{s}) = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*$ .

As knot-based methods, the performance of the kernel mixing method and the predictive processes method are highly relying on knots selection. For spatial process with relatively

large scale compared with the measurement scale, both of these two methods only require a small number of knots to effectively represent the realizations of the original processes in a lower-dimensional subspace hence achieve efficient computations. Therefore, we could employ either one of them to model the coarse scale dependence in (4.2). In particular, the predictive process method can accommodate a wide range of spatial stochastic processes (see Banerjee et al. (2007) for details), which provides great flexibility to characterize the dependence structures at coarse scale.

However, for spatial process with relatively fine scale, both of these two methods now require a relatively denser set of knots in order to preserve more complete information about the fine scale spatial pattern. For the predictive processes, assuming for the moment we work on the regular lattice based knots, Banerjee et al. (2007) admitted that the performance of the approximation depends on the size of the range relative to the spacing of the grid for the knots. Their study shows that the approximation is weak at short distances, and is even worse for processes with short observational spatial ranges. In this regard, the predictive process with a limited number of knots will not be able to make inference of dependence for pairs of sites that are very close (relative to the spacing of the grid) to each other. Therefore, when there is fine scale spatial variation, a dense set of knots has to be employed to ensure the good behavior of the predictive process. In this regard, this knot based method will lose all its computational advantage when modeling fine scale spatial variation since it essentially involves the intense matrix computations in a high dimension determined by the knot size. Return to the kernel mixing method which is also a knot-based method dimension reduction method, we realize a denser set of knots is also required to preserve more complete information about the fine scale spatial pattern, which inevitably increase the computational burden mainly due to high dimensional matrix product. But an encouraging news is that the basis variables  $\{z(\mathbf{s}_m^*)\}, m = 1, \dots, M$  are

mutually independent, which could still help us to avoid high dimensional matrix inversion for likelihood evaluation in large 'n' problem. Therefore, the kernel mixing method may still be applicable to modeling the fine operational scale spatial component  $w_{fine}(\mathbf{s})$ .

Prior settings for spatial scales in those two non-stationary spatial processes are straightforward. As discussed above, the choice of knots itself reflects the modeler's prior belief on the spatial range in the given spatial dataset. Typically, denser set of knots have to be placed in order to explore finer scale spatial dependence. Besides that, for the kernel mixing method, we have a kernel function  $K$  often involving a parameter  $\theta$  which controls the spatial range. Therefore, we might be able to place restriction on priors for  $\theta$ 's at different scales. In the predictive process method, there is an original process which may also contain some parameters controlling the spatial range. We could set priors for them in the similar spirit as we discussed before.

### 4.3.2 "Large n" in fine scale spatial processes

In section 4.3.1, we have concluded that dimension reduction based approaches such as kernel mixing method and the predictive processes method may not be optimal choices to model the fine operational scale spatial component. In this section, we consider two possible approaches to tackle the large n problem in the presence of fine scale spatial variations. Return to the spherical covariance function described in the introduction 1.2.1. This particular covariance function has the following form:

$$C(t) = \begin{cases} 0 & \text{if } t \geq 1/\phi \\ \sigma^2(1 - 3t/2\phi + (t/\phi)^3/2) & \text{if } 0 < t \leq 1/\phi \\ \sigma^2 + \tau^2 & \text{otherwise} \end{cases} \quad (4.12)$$

where  $t$  denotes the distance between any two locations. The spherical covariance function offers clear interpretations of the nugget parameter  $\tau^2$ , scale parameter  $\sigma$  and range param-

eters  $\phi$ . In particular, it is natural to set an upper bound for the range parameters  $\phi$ , which is typically a small value for the fine operational scale spatial process. Therefore, assuming a spherical covariance function, the variance covariance matrix  $\Sigma$  of  $w(\mathbf{s})$  with a small range is a sparse matrix. There are some efficient computations to compute the inverse of a sparse matrix and the determinant of it. Therefore, it may facilitate the computation by assuming a spherical covariance function for  $w_{fine}(\mathbf{s})$  in Model (4.2). However, the obvious drawback of using the spherical covariance function is that we would be restricted to work on this particular covariance structure, which is not flexible enough to cover more general covariance structures.

Kaufman (2006) proposed a covariance tapering method which provides two sparse approximations for the covariance matrices. Her method starts with the Matérn class of covariance functions denoted as  $K_0(x; \theta)$ , which is clearly a more general and flexible spatial covariance structures than the spherical covariance. Then she considers a tapering function  $K_{taper}(x; \gamma)$ , an isotropic correlation function which is identically zero whenever  $x \geq \gamma$ .  $\gamma$  can be viewed as the effective range for the spatial phenomenon being studied. The tapered covariance function is defined as

$$K_a(x; \theta, \gamma)K_{taper}(x; \gamma), x > 0 \tag{4.13}$$

From (4.13), we can derive that the tapered covariance matrix is denoted as  $\Sigma(\theta) \circ T(\gamma)$ , where  $T(\gamma)_{i,j} = K_{taper}(\|\mathbf{s}_i - \mathbf{s}_j\|)$ . The 'o' notation refers to the element-wise matrix product, also called the Schur product. The tapered covariance matrix is still a valid covariance matrix since the Schur product of two covariance matrices is still a positive definite matrix. By assigning  $\gamma$  a small value, we can constrain the range of the spatial process to be small as desired. Again, the resulting covariance matrices is sparse, hence can be manipulated using efficient sparse matrix algorithms. Therefore, the covariance tapering method provides us an computational efficient yet flexible way to model the fine

scale spatial component  $w_{fine}(\mathbf{s})$  in Model (4.2).

Finally, considering the limited range that fine scale dependence could operate and hence the limited information about the fine scale dependence, we think it is needed and reasonable to assume that the spatial process is isotropic stationary at fine scale in order to effectively learn the fine scale dependence. But large scale spatial dependence can certainly be specified in a more flexible way since it is not necessary to have a stationary spatial dependence structure at large scale.

## 4.4 Generalization to non-Gaussian spatial processes and spatio-temporal processes

In this section, we consider the potential generalization of coarse/fine scale spatial process model to cases where we work with Non-Gaussian spatial processes and spatio-temporal processes.

Many existing Non-Gaussian spatial processes typically follow the path to make use of generalized linear models in the context of spatial data. (see, e.g., Diggle et al. (1998), Kammann and Wand (2003) and in Banerjee et al. (2004)). Commonly used non-Gaussian first stage settings include: (i) binary response at locations, modeled using logit or probit regression, and (ii) count data at locations modeled using Poisson regression. The basic idea is to make the assumption that the mean of response  $Y(\mathbf{s})$  is linear on a transformed scale, i.e.,  $\eta(\mathbf{s}) \equiv g(E(Y(\mathbf{s}))) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s})$  where  $g(\cdot)$  is a suitable link function, and  $w(\mathbf{s})$  is a spatial (Gaussian) process.

We have another path to follow for any given Non-Gaussian continuous process which has an explicit form of inverse marginal distribution function, denoted as  $G^{-1}$ . The idea is to make the assumption that the Non-Gaussian process, denoted as  $v(\mathbf{s})$ , is a transformed

Gaussian process  $w(\mathbf{s})$ , i.e.  $v(\mathbf{s}) = G^{-1}\Phi(w(\mathbf{s}))$ , where  $\Phi$  is the marginal distribution function of  $w(\mathbf{s})$ .

With the Gaussian first stage, we can replace the single spatial random effects  $w(\mathbf{s})$  with  $w_{coarse}(\mathbf{s}) + w_{fine}(\mathbf{s})$  in the context of multi scale spatial associations. With say a binary or Poisson first stage to model the Non-Gaussian response data which exhibit spatial dependence at different scales, we could now replace  $w(\mathbf{s})$  in the latent stage with  $w_{coarse}(\mathbf{s}) + w_{fine}(\mathbf{s})$ , where  $w_{coarse}(\mathbf{s})$  and  $w_{fine}(\mathbf{s})$  are typically two independent spatial Gaussian processes. With a Non-Gaussian continuous process with inverse marginal distribution  $G^{-1}$ , we could also assume  $v(\mathbf{s}) = G^{-1}\Phi(w_{coarse}(\mathbf{s}) + w_{fine}(\mathbf{s}))$ . In some situations, we might flexibly combine these two approach. We will explain this by reviewing the smoothed GEV model in Chapter 3.

In Chapter 3, we essentially consider a hybrid version of those two non-Gaussian modeling strategies. We model the extreme value spatial process through a hierarchical modeling which has the GEV first stage model. We have discussed the multi spatial dependence structures exhibited in some extreme climate events (e.g., annual maxima of rainfalls and annual maxima of temperatures). Following the discussion, we have generalized the single scale spatial process model to the model which accounts for spatial dependence at two scales. The Gaussian spatial process  $w_{coarse}(\mathbf{s})$  with large spatial range is built in the latent stage model for the location parameters in GEV, aiming to capture the large scale spatial pattern in extreme climates. The fine scale dependence is introduced through the transformation of a Gaussian process  $w_{fine}^{fine}(\mathbf{s})$  with small scale range, i.e.,  $\sigma * (G^l - 1)\Phi(w_{fine}(\mathbf{s}))$ .

Again, those methods we employ to tackle large 'n' problems in Gaussian version are still applicable to modeling the latent fine scale spatial process and coarse scale spatial process.

We could also extend the Gaussian version of model (4.2) to various spatio-temporal

contexts in which multiscale spatial dependence structures operate. A general formulation of spatio-temporal process model is

$$Y(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)\boldsymbol{\beta} + w(\mathbf{s}, t) + \epsilon(\mathbf{s}, t) \quad (4.14)$$

for  $\mathbf{s} \in D$  and  $t \in [0, T]$ . In (4.14)  $\mathbf{x}(\mathbf{s}, t)$  are local space-time covariate vectors,  $\boldsymbol{\beta}$  is an associated coefficient vector and  $\epsilon$ 's are pure error terms.

In many cases, we only observe discretized time, say,  $t = 1, 2, \dots, T$ . Now, we would rewrite the response as  $Y_t(\mathbf{s})$  and the random effects as  $w_t(\mathbf{s})$ . Now, we would rewrite the response as  $Y_t(\mathbf{s})$  and the random effects as  $w_t(\mathbf{s})$ .  $w_t(\mathbf{s})$  now arises as a time series of spatial processes. Use of coarse/fine scale spatial processes model for  $w_t(\mathbf{s})$ , i.e., assuming  $w_t(\mathbf{s}) = w_{coarse,t}(\mathbf{s}) + w_{fine,t}(\mathbf{s})$ , enables us to investigate coarse and fine scale spatial dependence in a dynamic manner.

For continuous data, we replaced the spatial random effects,  $w(\mathbf{s})$ , with space-time random effects,  $w(\mathbf{s}, t)$  which come from a Gaussian process with valid covariance function  $cov(w(\mathbf{s}, t), w(\mathbf{s}', t')) \equiv C(\mathbf{s}, \mathbf{s}'; t, t')$ . There have been some literature regarding valid space-time covariance functions (Banerjee et al. (2004)). In particular, a frequently used choice is the separable form:

$$cov(w(\mathbf{s}, t), w(\mathbf{s}', t')) = C_1(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_1)C_2(t, t'; \boldsymbol{\theta}_2) \quad (4.15)$$

where  $C_1$  is a valid two-dimensional covariance function and  $C_2$  is a valid one-dimensional covariance function. We may assume the separated spatial covariance function  $C_1$  should characterize both coarse and fine spatial dependence, i.e.  $C_1(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_1) = C_{coarse}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_{coarse}) + C_{fine}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_{fine})$ . Furthermore, if investigation of long term temporal dependence and short term temporal dependence is of research interest, we could analogously assume  $C_2(t, t'; \boldsymbol{\theta}_2) = C_{long}(t, t'; \boldsymbol{\theta}_{long}) + C_{short}(t, t'; \boldsymbol{\theta}_{short})$ . Bayesian analysis of coarse/fine spatial dependence and long/short temporal dependence can be proceeded in a similar fashion as

discussed in the preceding sections

## 4.5 Simulation examples of prediction approximation

In this section, we simulated a response vector  $\mathbf{Y}(\mathbf{s})$  for each of 10000 irregularly scattered locations over a  $[0, 10] \times [0, 10]$  rectangle using Model (4.2). The covariate variables  $X$  are randomly generated from  $[0, 1]$  and the covariance coefficient  $\boldsymbol{\beta} = (1, 1, 1)'$ . Spatial association was assumed to follow the exponential correlation function, with fine scale range  $\phi_{fine}$  and coarse scale range  $\phi_{coarse}$ . For simplicity, assume the scale parameters for both coarse and fine scale spatial random effect are 1, i.e.,  $\sigma_{coarse} = \sigma_{fine} = 1$ . And true fine scale range parameter is set to be  $\phi_{fine} = 0.24/3$ , and the true coarse scale range parameter is set to be  $\phi_{coarse} = 1$ .

Our focus is on the fine scale spatial dependence, in which the true operational scale is smaller than the measurement scale. Notice that the minimum interlocation distance is defined as a measure for the measurement scale for the spatial process we simulated, and the operational scale is characterized by the range parameter of the covariance function we assumed for the spatial random effects. Therefore, we keep 1000 points by thinning out the original data such that the minimum interlocation distance for the set of remaining sites is close to 0.25.

Given these data, we considered two spatial models to fit them. The first model is

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + W_{coarse}(\mathbf{s}) + W_{fine}(\mathbf{s}) + \epsilon(\mathbf{s}) \quad (4.16)$$

$W_{fine}(\mathbf{s})$  and  $W_{coarse}(\mathbf{s})$  are modeled as spatial gaussian processes with exponential covariance functions.

The second model we consider is a single spatial point process model as follows:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + w_{single}(\mathbf{s}) + \epsilon(\mathbf{s}) \quad (4.17)$$

where  $w_{single}(\mathbf{s})$  is a Gaussian process with exponential covariance function and  $\epsilon(\mathbf{s})$  is nugget with variance  $\tau^2$ . The idea of making comparison between Model (4.2) and Model (4.17) is to see if data can distinguish the fine scale dependence from the pure errors.

Prior distributions are assigned to model parameters to complete the Bayesian specification. A normal prior with large variance was assigned to each intercept parameter  $\boldsymbol{\beta}$ . The coarse scale spatial range parameter in the exponential correlation function  $\phi_{coarse}$  was assumed to have  $U(0.08, 3)$  as prior distribution. And  $K$ , the ratio of the coarse scale range to the fine scale range, was assigned a flat prior in  $(1, \infty]$ . For each model, we ran 10000 iterations to collect posterior samples after a burn in period of 5000 iterations, thinning using every fifth iteration. Trace plots of parameters indicate good convergence of the respective marginal distributions. We saved the remaining 1,000 samples from each chain for posterior inference. We compare the performance of these models based on the following criteria: Deviance Information Criterion (DIC) (Spiegelhalter et al. (2002)); estimation accuracy of the *true* parameter values; prediction of a holdout set of 9,000 locations.

Table 4.1 and Table 4.2 display the parameter estimates under Model (4.16) and Model (4.17) respectively. In Figure 4.2, we plot histograms of the effective ranges based on the posterior samples for  $3\phi_{coarse}$  and  $3\phi_{fine}$  under Model 4.2. Similarly, Figure 4.3 shows the histogram of the posterior samples for  $3\phi_{single}$  under Model 4.3. As shown in these two figures, the estimated mean of the effective range  $3\phi_{single} = 2.5$ , which is between that of  $3\phi_{coarse} = 3$  and that of  $3\phi_{fine} = 0.25$ . Noteworthy, the histogram of  $3\phi_{fine}$  in Figure 4.2 clearly suggests that the fine scale range parameter is different from 0. This finding provide evidence that Model 4.2 is able to distinguish micro scale spatial variations from measurement errors when it is present and we have enough data.

Table 4.3 provides the DIC model fit statistics for Model (4.16) and Model (4.17). The first column, labelled DIC, provides a relative measure of goodness of model fit, with lower values indicating better fit (see., e.g., Spiegelhalter et al. (2002) for a full explanation of the criterion). Model (4.16) clearly has a smaller DIC score than Model (4.17), which indicates the better fit using model with coarse and fine scale components. The subsequent results focus on the parameter estimates and prediction performance for these two models. We now turn our attention to prediction of the holdout set. We compare models in terms of prediction performance at hold out set. For each location  $\mathbf{s}_0$  in the hold out set, samples of  $Y(\mathbf{s}_0)$  are drawn from the posterior predictive distribution. Posterior means are adopted as the point estimates of the predicted positions for each new location. The root mean square predictive errors (*RMSPE*) are computed to assess the predictive performance for each model. Model (4.16) also wins over Model 4.17 based on RMSPE. The third column of Table 4.3, called the coverage probability labelled  $\hat{r}$ , is the proportion of times the true observations lie in their associated 95% predictive interval under each model.

## 4.6 Discussion

Little work has been done to study the so called microscale spatial process. In this paper, we bring up this problem from a statistical point of view. We have presented flexible spatial point-level process models to account for both coarse and fine scale spatial patterns. In particular, we follow the Bayesian inference approach to take advantage of the prior information about the spatial operational scales, which effectively help us to detect both coarse and fine operational scales appeared within some spatial data. Various methods have also been introduced to efficiently reduce the computational cost associated with the modeling of spatial data set with large number of locations.

Various extensions for this work are possible. We have discussed a few possibilities in

Parameter	True Value	Posterior Mean	Lower CI (0.025%)	Upper CI (0.975%)
$\beta_0$	1	0.9786	-0.0308	1.8556
$\beta_1$	-1	-0.8888	-1.3924	-0.4297
$\beta_2$	1	0.7231	0.2358	1.1801
$\sigma_2$	1	0.9249	0.7625	1.1035

**Table 4.1:** Parameter estimations obtained by Large/Small scale model

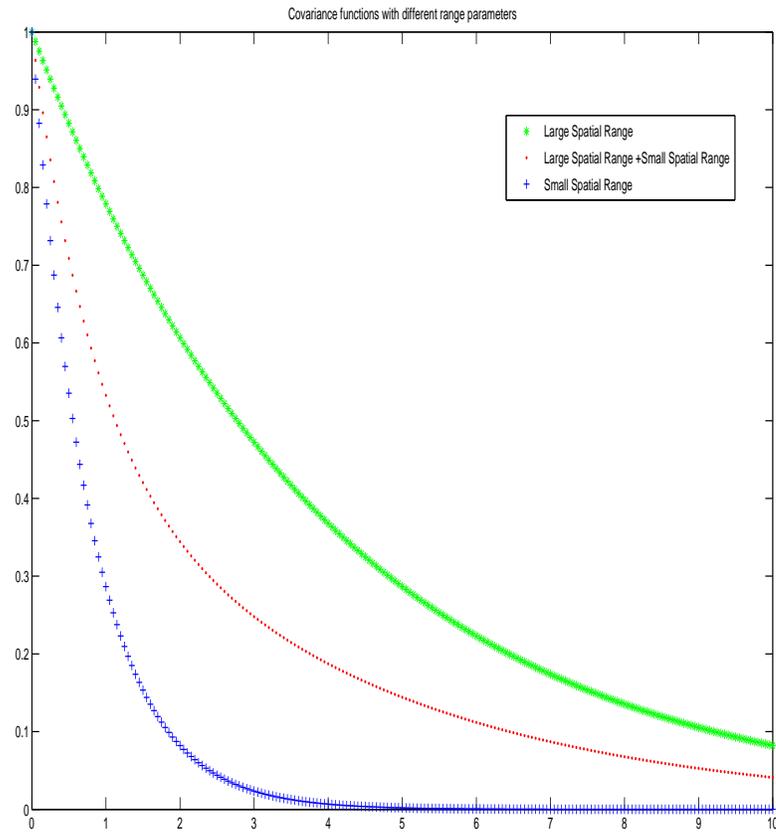
Parameter	True Value	Posterior Mean	Lower CI (0.025%)	Upper CI (0.975%)
$\beta_0$	1	1.0447	0.7195	1.3191
$\beta_1$	-1	-0.7982	-1.0888	-0.5435
$\beta_2$	1	0.6419	0.3939	0.9434
$\sigma_2$	1	1.1234	0.6422	1.7629
$\tau_2$	0.02	0.5277	0.0591	1.2555

**Table 4.2:** Parameter estimations obtained by single process model

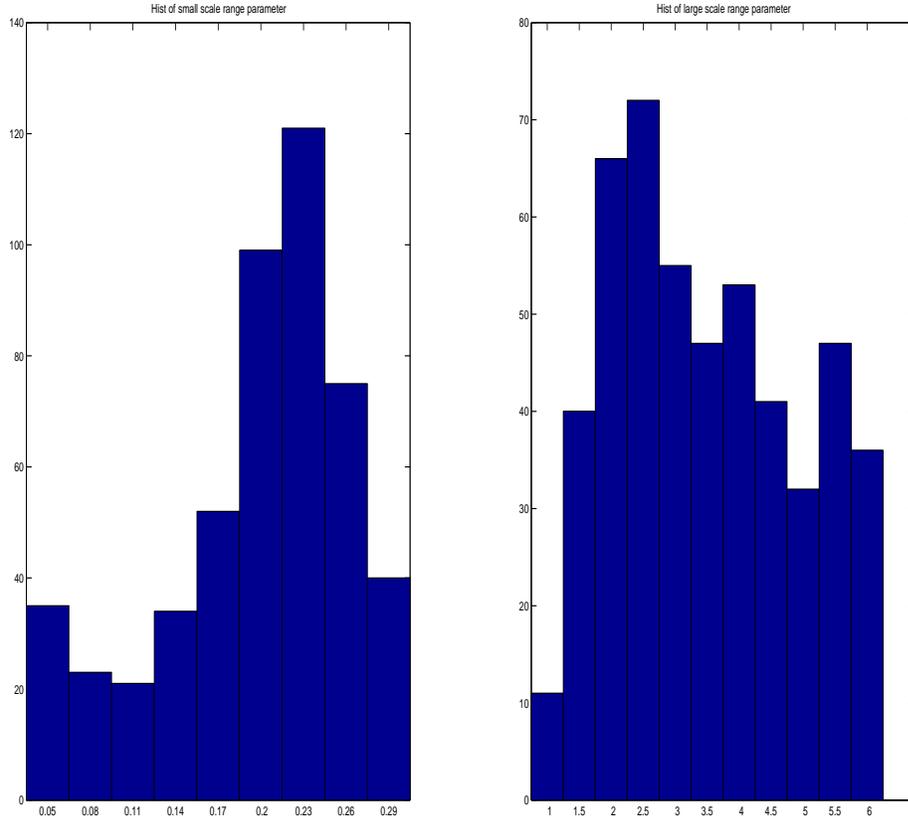
Section 4.4. With non-Gaussian models in the first stage settings, we can embed Model 4.2 in the latent stage to study the coarse and fine spatial variations for non-Gaussian response spatial data. We can also extend Model 4.2 to various spatio-temporal contexts in which Model 4.2 can be introduced to study micro spatial temporal scales.

	DIC	RMSPE	$\hat{r}$
Large/Small scale	1091.1	0.8699	0.956
Single scale	1219.4	0.9097	0.934

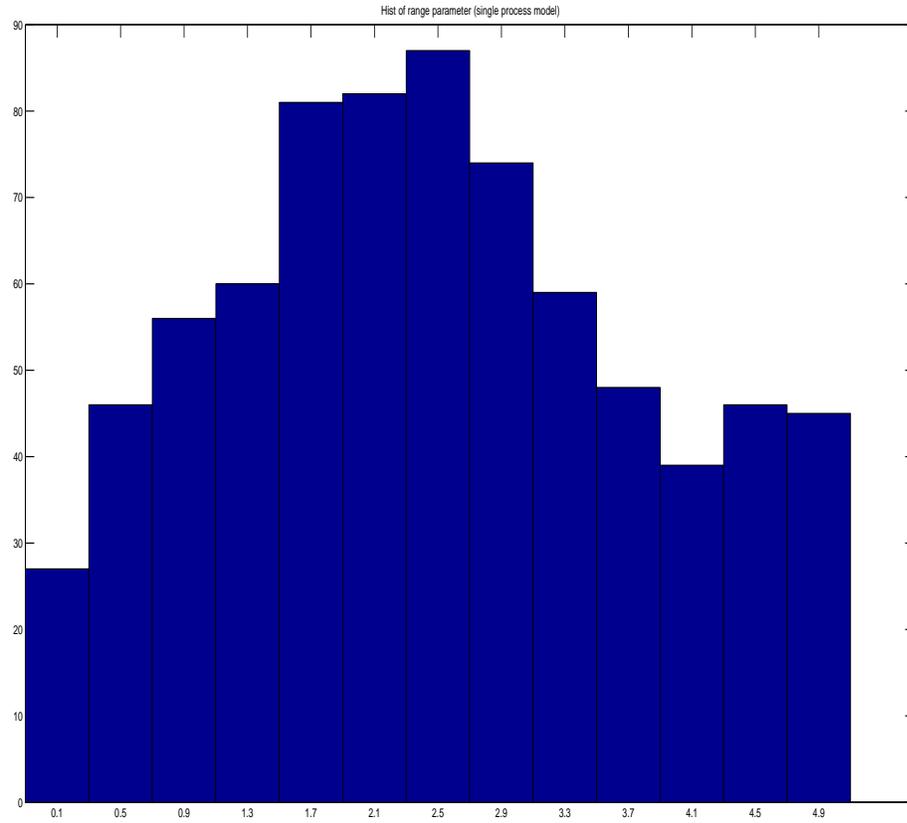
**Table 4.3:** Performance of Model 4.2 and Model 4.17 using root mean square predictive errors (RMSPE), the deviance information criterion (DIC), and the empirical coverage probability  $\hat{r}$ .



**Figure 4.1:** Exponential correlation functions at different scales.



**Figure 4.2:** Histogram of the coarse scale range parameter  $\phi_{coarse}$  and the fine scale range parameter  $\phi_{small}$



**Figure 4.3:** Histogram of single range parameter  $\phi_{single}$

# Chapter 5

## Interpreting Self Organizing Maps Through Space-time Data Models.

### 5.1 Introduction

Self organizing maps (SOM's) are a technique that has been used with high dimensional data vectors to develop an archetypal set of states (nodes) that span, in some sense, the high dimensional space. First developed by Kohonen (1995), the technique has subsequently found application to automatic speech recognition, analysis of electrical signals from the brain, data visualization, and meteorology. See, e.g., Ferrandez et al. (1997), Tamayo et al. (1999), Kaski (1997) and Crane and Hewitson (2003), respectively.

The SOM approach is essentially a neural network model that implements a nonlinear projection from a high-dimensional input space to a low-dimensional array of neurons. In the process, it also becomes a clustering technique, assigning to any vector in the high dimensional data space the node/neuron (reference vector) to which it is closest (using, say, Euclidean distance) in the data space. The number of nodes is thus equal to the number of clusters. However, the primary use for the SOM is as a representation technique, i.e., finding a set of nodes which representatively *span* the high dimensional space. These nodes are typically displayed using maps to enable visualization of the continuum of the data space. Hence, the approach should not be viewed as an “optimal” clustering technique; in particular, in application it is expected to produce roughly equal cluster sizes.

A SOM algorithm is usually implemented in three stages. First, a specified number of nodes is selected and the values of the components for each node are initialized, typically

selecting random values. Second, iterative training is performed where the nodes are adjusted in response to a set of training vectors so that the nodes approximately minimized an integrated distance criterion. The last stage of the SOM technique is visualization where each node's reference vector is projected in some fashion to a lower dimensional space and plotted as a map (perhaps several maps). Customary projection creates a set of neurons in two-dimensional space which arise as a deformation of a regular lattice. For a given training set, the frequency of occurrence of each node can be calculated as well as the average error at each node, the latter interpreted as a measure of coherence around the node. With regard to implementation, the number of nodes is arbitrary. However, presumably more nodes are needed to capture a more variable high dimensional space or a greater level of detail in distinguishing nodes.

In any event, it is not our contribution here to criticize the SOM approach or to compare it with other clustering procedures. Rather, in practice, the procedure is implemented in a purely algorithmic manner, ignoring any spatial or temporal structure which may be anticipated in the training set. Our contribution is to attempt to incorporate structural dependence, through the introduction of stochasticity in the form of a space-time process model. As a result, we hope to illuminate and interpret the performance of the SOM procedure in the context of application to daily data collection. That is, the observed daily state vectors are viewed as a time series of multivariate spatial process realizations. Working with the original high dimensional data renders formal modeling infeasible. Instead, we try to achieve this understanding through the dimension reduction achieved by the SOM procedure.

The application we focus on here is to synoptic climatology as introduced by Hewitson and Crane (2002) where the goal is to develop an array of atmospheric states to capture a collection of distinct circulations. There has been some literatures on estimating

synoptic states with the purpose of downscaling climate models. For example, Hughes et al. (1999) and Bellone et al. (2000) propose nonhomogeneous hidden Markov models (NHMM) which relates precipitation occurrences and amounts at multiple rain gauge stations to broad-scale atmospheric circulation patterns. Specifically, let  $\mathbf{R}_t = \{R_t^1, \dots, R_t^n\}$  be a multivariate random vector giving rainfall amounts at a network of  $n$  sites, let  $S_t$  be the hidden (latent) synoptic weather state at time  $t$  and  $\mathbf{X}_t$  the vector of atmospheric measures at time  $t$ . The NHMM is defined by two modeling stages: (1) a rain fall occurrence process:  $P(\mathbf{R}_t | S_{1:T}, \mathbf{R}_{t-1}, \mathbf{X}_{1:T}) = P(\mathbf{R}_t | S_t)$ ; and (2) a state process model:  $P(S_t | S_{1:t-1}, \mathbf{X}_{1:T}) = P(S_t | S_{t-1}, \mathbf{X}_t)$ . In particular, both papers assume that occurrences and precipitation amounts at each rain gauge are conditionally independent given the current weather state, i.e., all the spatial dependence between rain gauges is induced by the synoptic weather state.

In this Chapter, we work with daily weather data observed in the form of 11 variables measured for each of 77 grid cells yielding an  $847 \times 1$  vector for each day. We have such daily vectors for a period of 31 years (11,315 days). Twelve SOM nodes have been obtained by the meteorologists to represent the space of these data vectors. Fuller detail is provided in Section 5.3. We also note that a broader view of the use for a SOM in climatology is for inference at longer than daily time scales.

The format of the Chapter is as follows. In Section 5.2 we provide a brief review of the SOM theory and implementation. In Section 5.3 we detail of the motivating dataset and some exploratory data analysis. In Section 5.4 we present a collection of models to investigate. Section 5.5 addresses model fitting issues while Section 5.6 considers model selection and results. Section 5.7 offers some summary discussion.

## 5.2 A Review of SOM theory and implementation

A self-organizing map (SOM) is a neural network model and algorithm that implements a nonlinear projection from a high-dimensional space of input vectors to a low-dimensional array of neurons. That is, input vectors are assigned to nodes (or neurons). Nodes have two positions, one in the high-dimensional space, say in a subset of  $\mathbb{R}^d$ , one in the low-dimensional visualization space, typically taken to be a deformation of a regular lattice in two-dimensional space. For a given set of nodes  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$  in the high-dimensional space, an array index taking values in  $\{j = 1, 2, \dots, M\}$  is defined, for each  $\mathbf{x} \in \mathbb{R}^d$ , as  $c(\mathbf{x}, \{\mathbf{w}_m\}) = j$  if  $d(\mathbf{x}, \mathbf{w}_j) = \min_m d(\mathbf{x}, \mathbf{w}_m)$  for some distance  $d$  (usually Euclidean). The theoretical objective of the SOM is to minimize, over all choices of  $\{\mathbf{w}_m, m = 1, 2, \dots, M\}$ ,  $\int g(d(\mathbf{x}, \mathbf{w}_{c(\mathbf{x}, \{\mathbf{w}_m\})}))p(\mathbf{x})d\mathbf{x}$  where  $g(\cdot)$  is a monotone function and  $p(\mathbf{x})$  is the density function for the random input vectors in  $\mathbb{R}^d$ . Solution to this vector quantization problem is generally intractable. We note that if we confine  $\mathbf{x}$  to a bounded rectangular subset of  $\mathbb{R}^d$  and if  $p(\mathbf{x})$  is assumed uniform over this subset then, at the optimal  $\{\mathbf{w}_m\}$ ,  $c$  will be equally likely to take on each of its  $M$  possible values. Hence, with a sample of  $\mathbf{x}$ 's from this uniform distribution, we expect equal numbers of the  $\mathbf{x}$ 's to be assigned to each of the index values, i.e., to each of the nodes. A special version which seeks to minimize  $\int \sum_m h(\mathbf{w}_{c(\mathbf{x}, \{\mathbf{w}_m\})}, \mathbf{w}_m)g(d(\mathbf{x}, \mathbf{w}_m))p(\mathbf{x})d\mathbf{x}$  is customarily used. It lacks a closed form solution but an approximate solution can be obtained iteratively using stochastic approximation (see ?) as we describe below.

We now offer a bit more detail on the nature of a SOM algorithm. In practice, the SOM procedure consists of three stages. Let  $\{\mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$  denote the input training vectors. In our case,  $d = 847$  reflecting the daily 847-element climate records from 1970 to 2000. SOMs seek to “optimally” place a specified number of nodes,  $M$ , again denoted by  $\mathbf{w}_m \in \mathbb{R}^d, m = 1, 2, \dots, M$ . In the SOMs literature (and, as the default in the publicly

available software package cited below) the suggested number of nodes is  $5\sqrt{n}$ .<sup>1</sup> We describe two versions of the iterative training algorithm procedure of the SOM technique as follows:

- Initialization stage: given  $M$ , the node vectors are initialized with random values.
- Iterative training(Version 1):
  - At step  $t$ , randomly choose an input vector  $\mathbf{x}^{(t)}$  from the training set  $\{\mathbf{x}_i\}$  for  $i = 1, \dots, n$ .
  - Compute the distance (e.g., Euclidean) between  $\mathbf{x}^{(t)}$  and each of the node vectors  $\mathbf{w}_m$ . Identify the winning node  $\mathbf{w}_{c(\mathbf{x}^{(t)})}$  whose node vector is most similar to the input vector, i.e.,  $\|\mathbf{w}_{c(\mathbf{x}^{(t)})} - \mathbf{x}^{(t)}\| \leq \|\mathbf{w}_m - \mathbf{x}^{(t)}\|$  for  $m \in \{1, 2, \dots, M\}$ .
  - Every node has its vector adjusted according to the following equation:
 
$$\mathbf{w}_m^{(t+1)} = \mathbf{w}_m^{(t)} + \alpha(t)K(m, c(\mathbf{x}^{(t)}))(\mathbf{x}^{(t)} - \mathbf{w}_m^{(t)}),$$
 where  $K(m, c(\mathbf{x}^{(t)}))$  is called the neighborhood function, and  $\alpha(t)$  is called learning rate which is usually a decreasing function of step  $t$ . One example of  $K(m, c(\mathbf{x}^{(t)}))$  is the Gaussian kernel  $K(m, c(\mathbf{x}^{(t)})) = \exp - \{\|\mathbf{w}_m^{(t)} - \mathbf{w}_{c(\mathbf{x}^{(t)})}\|^2 / 2\sigma^2\}$ . A simpler choice is a so-called ‘bubble’ function, i.e., a uniform over the neighborhood (Voronoi tessellation) of  $\mathbf{w}_{c(\mathbf{x}^{(t)})}$ , zero elsewhere.

Usually, the SOM training is performed in two phases. In the first phase, relatively large initial learning rate is used in the first phase and small learning rate is used in the second phase. This updating suggests that nodes close to the winner node as well as the winner itself, update their vectors closer to  $\mathbf{x}^{(t)}$

---

<sup>1</sup>In our application, with  $n \approx 10,000$  this would suggest roughly  $M=500$  nodes. However, climatologists categorize far fewer types of circulation patterns; for our South African data, they conclude that  $M = 12$  is adequate.

in the input data space. Vectors associated with far away output nodes do not change significantly.

- Repeat the above steps until the nodes converge. (Convergence is vaguely defined and is usually taken as the default in the software.)

- Iterative training(Version 2):

- At step  $t$ , for each input vector  $\mathbf{x}_i$  for  $i = 1, \dots, n$ , compute the distance (e.g., Euclidean) between  $\mathbf{x}_i$  and each of the node vectors  $\mathbf{w}_m^{(t)}$ . Identify the winning node  $c(i)$  whose node vector is most similar to the input vector, i.e.,  $|\mathbf{w}_{c(i)}^{(t)} - \mathbf{x}_i| \leq |\mathbf{w}_m^{(t)} - \mathbf{x}_i|$  for  $m \in \{1, 2, \dots, M\}$ .

- Every node has its vector adjusted according to the following equation:

$$\mathbf{w}_m^{(t+1)} = \frac{\sum_{i=1:n} h_{m,c(i)}(t)\mathbf{x}_i}{\sum_{i=1:n} h_{m,c(i)}(t)},$$

where  $h_{m,c(i)}(t)$  is the neighborhood function around the winning node  $c(i)$ . One example is  $h_{m,c(i)}(t) = \alpha(t)K(m, c(i))(t)$ , where, again,  $K(m, c(i))$  is the Gaussian neighborhood kernel  $K(m, c(i)) = \exp - \{ \|\mathbf{w}_m^{(t)} - \mathbf{w}_{c(i)}^{(t)}\|^2 / 2\sigma^2 \}$ . Here,  $\alpha(t)$  is called the learning rate and is usually a decreasing function of step  $t$ .

Again, the SOM training is performed in two phases. In the first phase, relatively large initial learning rate is used in the first phase and small learning rate is used in the second phase. In this updating, the contribution (weight) of a particular training vector to each node only depends the distance between the corresponding winning node of this training vector and each of the other nodes.  $h_{m,c(i)}$  can be viewed as a smoothing function such that nodes close to the winner node as well as the winner itself update their vectors closer to the training vector in the input data space.

- Repeat the above steps until the nodes converge.

- The final stage seeks to achieve visualization. When the number of nodes is large, visualization is most easily presented in two dimensions beginning with either a rectangular or hexagonal lattice of nodes. The iterative updating of the nodes eventually leads to a distortion of the lattice. (See Figure 5.4 and related discussion below.) An approach which is incorporated into the standard SOMs software (cited below) is the Sammon mapping scheme (Sammon, 1969). Note that the goal here is “clustering” components of the vectors to achieve a two dimensional representation, not clustering of the training vectors. In order to equalize the contributions of each of the components in the high-dimensional vectors with regard to classification, centering and scaling is recommended as pre-processing of the data. Returning to the Sammon projection method, the basic idea is to arrange all the nodes on a 2-dimensional plane in such a way, that the distances between the nodes in a 2-dimensional space resemble the distances in the original vector space as defined by some metric as faithfully as possible. Given the distance matrix  $D$  with element  $d(i, j)$  being the distance between node  $i$  and node  $j$  according to some metric (e.g., Euclidean distance), our goal is to find  $\mathbf{O}_m$  in  $\mathbb{R}^2$  for each node  $m$  for  $m = 1, \dots, M$  to minimize an error function  $E$  defined by the following cost function: 
$$E = \frac{1}{\sum_i \sum_{j>i} d_{i,j}} \sum_i \sum_{j>i} \frac{(d_{i,j} - \|\mathbf{O}_i - \mathbf{O}_j\|)^2}{d_{i,j}}$$
 Note that the  $O$ 's need a ‘center’ to locate them. Also, the projections can be implemented at each iteration to see stability, hence convergence, as well as to assess interpretation.

A software package for implementing SOMs is available (<http://www.cis.hut.fi/research/som-research>). For more detailed explanation of the SOM procedure see the references and guidelines at this publicly available software website.

As a final comment on visualization, in our application below,

## 5.3 The dataset and some exploratory data analysis

The weather in a local region is conditional on the nature of the synoptic state of the atmosphere. Relating the synoptic scale characteristics to local scale responses requires the reduction of a large number of variables into a smaller set of data that still, in some sense, represent the original data. This goal motivated the use of the SOM technology. In this application, we use daily multivariate weather data over a specified time period to produce generalized weather circulations. These are then easily visualized as an array of archetypal synoptic circulations that span the continuum of the data. In so doing, daily synoptic atmospheric data are categorized into a prescribed number of archetypal synoptic (circulation) modes characteristic of a specified time period. South African weather systems have been categorized into six to eight main "types" of circulation (Tyson and Preston-Whyte (2000)). On testing various SOM sizes, a 12-node SOM was selected which was deemed to adequately represent all the expected synoptic types.

The SOM was trained using gridded ( $2.5^\circ \times 2.5^\circ$ ), daily mean atmospheric fields constructed from six-hourly National Center for Environmental Prediction / National Center for Atmospheric Research (NCEP/NCAR) global reanalysis data (see Kalnay et al. (1996)). Data were extracted for a domain with  $11 \times 7$  grid cells over southern Africa whose latitudinal and longitudinal extent ( $25^\circ\text{S}$  to  $40^\circ\text{S}$ ;  $10^\circ\text{E}$  to  $34.5^\circ\text{E}$ ) captures synoptic circulation patterns from the sub-tropics to the mid-latitudes. The following 11 variables were chosen as training data: mean sea level pressure, 500 hectopascal (hPa) geopotential height level, relative and specific humidities at the surface and at 700hPa, daily maximum temperature at the surface, U- and V- wind components at the surface and at 700hPa. Each of these variables was first standardized using the mean and the standard deviation of its corre-

sponding  $11 \times 7$  cell time series. These standardized variables were then used to create an 847-element vector ( $11 \times 11 \times 7$ ) which described the daily atmospheric state. The time period from 1970 to 2000 was used which resulted in 11,315 daily records that were used to train the SOM. (The eight extra days in the included leap years were not included in the analysis for computational purposes but these would not significantly alter the results.) Each climate variable fields were standardized to preserve the local gradients in each field.

The twelve resulting SOM nodes are labelled with their locations in two-dimensional space in Figure 5.1. This figure is intended to suggest that nodes near to each other are associated with somewhat similar synoptic states and that transition in SOM nodes is most likely to be to a neighboring node.

To clarify the visualization, the SOM of sea level pressure (SLP) is presented in Figure 5.2. It is used to assess the characteristic surface circulation associated with each node as it most clearly demonstrates the general synoptic circulation as well as associated regional weather patterns. The SOM of maximum temperatures is also presented (Figure 5.3) to assess temperature patterns associated with each node.

Elaborating further, we briefly detail the features of the synoptic types captured by the 12 SOM nodes. The majority (80 %) of summer days map to nodes 5, 6, 8 and 9 and to a lesser degree nodes 3 and 12. These nodes are associated with characteristic summer circulation features. On the right hand side of the SOM, a sub-tropical low pressure system is situated over the northern part of the domain which bring rainfall to the interior of the country. In nodes 8, 9, and 12 a high pressure system is located at relatively high latitudes to the south of the domain which push frontal system southward and result in dry summers over the western parts of the country and introduce moisture to the eastern and central parts. In node 5, a linkage between the tropical low and mid-latitude circulation forms a tropical-temperate trough which results in rainfall over a large part of the interior of the

country. Highest maximum temperatures reach their greatest southerly extent and broadest spatial distribution over the region in these nodes. The majority of winter days (over 70%) map to nodes on the left side of the SOM (nodes 1,4,7,10). To the south of the country, these nodes are associated with the west-east progression of mid-latitude cyclones (cold fronts) across the south of the country which bring rainfall to the south and south-western parts of the country and very cold temperatures, especially over the interior. Over the interior of the country, the sub-tropical low has moved northward and is replaced by a high pressure system which dominates the circulation resulting in cold, dry conditions. In nodes 7 and 10 a high pressure system brings cold, polar air into the country once the cold front has moved past. Highest maximum temperatures are situated at much lower latitudes in these nodes and cooler maxima evident over much of the region. A typical winter synoptic sequence would be a progression from node 1 to node 4 to node 7 to node 10 over the period of about 2-3 days. Most spring days map to nodes 3, 10, 11, 12 and most autumn days map to nodes 1, 3 and 12. These nodes represent both summer and winter circulations expected in a transitional season.

Undertaking some preliminary exploratory data analysis, a first investigation is to explore the frequencies of occurrence of each node, hence of the synoptic climate systems. Table 5.1(left) provides a histogram showing the frequencies of daily observations mapped to each node over the entire study period, from which we observe fairly evenly distributed percentage frequency of occurrence for each synoptic node and no particular archetype is found dominant over the study period. This is an anticipated result of the SOM algorithm as clarified in Section 5.2. As a crude look to infer temporal behavior of the synoptic climate states, Table 5.1(right) compares the histogram of frequency of occurrence during 1970 to 1979 along with that during 1990 to 1999. Some evidence of temporal shifting in the distribution of incidences over the study area is seen. For example, node 1, which represents

strong low-pressure systems, occurs more frequently during 1990 to 1999 compared with the 1970's. Table 5.2(left) shows the frequency distribution of occurrence for each node in the summer season (December, January, February). It is clear that the climate archetype which corresponds to nodes 5, 6, 8, and 9 dominates during summer period. As shown in Table 5.2(right), the dominant climate archetypes transfer to another type of circulation in winter (June, July, August). Now, we see high frequency of SOM nodes 1, 4, 7, and 10. We may also use SOM arrays to examine short term (e.g., daily) temporal evolution of synoptic events. The frequencies of daily transitions from each node to other nodes are calculated and shown in Table 5.3, which reveals a somewhat clockwise cyclic evolution (with regard to Figure 5.1) of the weather systems. For example, SOM node group 9 displays preferential transition to group 6 while SOM node group 6 tends to most prefer transition to group 3. In Section 5.4, we elaborate this analysis by introducing formal time series modeling to interpret the SOM arrays.

## 5.4 SOM modeling

### 5.4.1 Dimension reduction

The daily climate reference data consist of an  $847 \times 1$  vector for each day within a 31-year period, which raises methodological and computational challenges when we attempt to interpret them in high dimension. In fact, since the  $847 \times 1$  vector arises as an 11 dimensional vector at each of 77 grid units, it is clear that we are monitoring a multivariate space-time data process. We do not seek to model this process directly, a very challenging task to develop and, likely, infeasible to fit. Rather, we seek to understand this process in terms of the SOM nodes that have been created. We will take advantage of the dimension reduction provided by the SOM procedure to, instead, model the induced collection of

two-dimensional locations across time. As we remarked earlier, the SOM algorithm ignores time and space in creating the nodes. By introducing a space-time process model, we seek to enhance behavioral interpretation for the set of SOM nodes. The result of the SOM algorithm yields, in our case, twelve nodes, each with an associated two-dimensional location (Figure 5.1). We now seek to “project” the 11,315 daily state vectors onto this space of locations. Many schemes are available to accomplish this; there is no “best” one. We choose to map daily high dimensional reference data onto 2-dimensional surface using high dimensional pairwise distances along with the 2-dim coordinates of the SOM nodes. For each daily state vector, the Euclidean distances between it and each of the nodes are calculated in high dimensional space. Then, for each daily vector, nodes are ranked by their distances to the vector. We next introduce a greedy space searching technique that maps each daily vector onto a 2-dimensional surface. To be specific, a 2-dimensional point within a finite bounded region is selected as the projection if the ranks of the Euclidean distances in two dimensions to each node using this point agree with the ranks between the vector and the nodes in high dimensional space. Evidently, there may be no point in 2-dimensional space which satisfies this condition so we seek agreement in ranking starting from smallest distance. Also, for a given high-dimensional point, such mapping may yield multiple mapped positions that provide the same extent of agreement in terms of rank distance agreement. In that situation, we averaged the coordinates of the multiple positions to ensure the uniqueness of the mapped position. Such an algorithm is easy to code and easily handles 11,315 points in 847 dimensional space. In Section 5.4.2, we primarily focus on modeling the 2-dimensional coordinates derived using this method. Other projection approaches utilizing alternative, possibly global, optimization criteria are certainly available though they may be difficult to implement. However, the modeling approach we develop in Section 5.4.2 can be applied to the results of any projection strategy.

## 5.4.2 Modeling specifications

The projection method described in Section 5.4.1 is performed on the daily referenced data from 1970-2000 to yield 2- dimensional mapped coordinates which are plotted in Figure 5.4 along with the coordinates of the SOM nodes. We can see that the two dimensional space is naturally partitioned into 12 tessellations, each attached to a SOM node. Of course, Figure 5.4 gives no information regarding the temporal sequence of the points.

However, let  $\mathbf{s}_t = (x_t, y_t)'$  denote the coordinates in 2-dims for day  $t$ ,  $t = 1, \dots, T$ . Before beginning the time series analysis, it is natural to examine the autocorrelation in this bivariate time series. We ran a standard vector autoregression software package for lags 1 up to 50. The plot (not shown in the interest of space) finds an adjusted AIC value of 7.88 for the AR(1) model, 7.81 for the AR(2) model and reaches its minimum at 7.76 for the AR(24) model. So, while there may be some evidence of longer range dependence in the series, the relative decrease in the model choice criterion is very small; AR(1) models may be good enough. Moreover, with interest in studying transition probabilities, in the sequel we work exclusively with AR(1) specifications. Under 12 nodes this still yields 144 transition probabilities. For the AR(2), we arrive at 1728 transition probabilities, too many to estimate well and to display.

We start the analysis with a bivariate random walk Gaussian model

$$\mathbf{s}_{t+1} = \mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}$$

where  $\boldsymbol{\epsilon}_t$  follows a bivariate Gaussian distribution centered at 0 with a  $2 \times 2$  covariance matrix  $\boldsymbol{\Sigma}$ . Under this model, the conditional Gaussian probability density functions of the coordinates at each time step are completely determined by the coordinates at the previous time step along with the covariance matrices. For us, the bivariate random walk model plays the role of a *straw man*. If the SOM nodes effectively capture synoptic weather states, there should be some structure to the daily transitions in the  $\mathbf{s}_t$ 's. In other words,

the algorithm yielding the SOM nodes is applied to spatially-referenced explanatory climate variables observed over time and therefore we would expect behavior with a more mechanistic description than purely random movement of the daily states in 2-dimensions. In this regard, denote  $\mathbf{Y}' = (\mathbf{s}_2, \dots, \mathbf{s}_T)$ ,  $\mathbf{X}' = (\mathbf{s}_1, \dots, \mathbf{s}_{T-1})$ . Then, the conditional maximum likelihood estimator (MLE) of  $\Sigma$  is:

$$\Sigma = (\mathbf{Y} - \mathbf{X})'(\mathbf{Y} - \mathbf{X})/(T - 1).$$

A very general form of bivariate time series model is the following:

$$\mathbf{s}_{t+1} = \mathbf{A}(\mathbf{s}_t, t)\mathbf{s}_t + \boldsymbol{\eta}(\mathbf{s}_t, t) + \boldsymbol{\epsilon}_{t+1} \quad (5.1)$$

where  $\mathbf{A}(\mathbf{s}_t, t)$  is a  $2 \times 2$  unknown matrix containing autoregression coefficients that are allowed to vary in space and time, the values of which are specified by location and time at the preceding step.  $\boldsymbol{\eta}(\mathbf{s}_t, t)$  enters (5.1) as an adjustment to the autoregressive component and can also be specified as a function of the preceding location and time. Again, error terms  $\boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_T$  are independently identically distributed  $N(\mathbf{0}, \Sigma)$  representing unstructured noise or pure error in the model. The model in (5.1) provides a very flexible specification in the form of a locally affine transition model. In fact, it is also very challenging to fit. We are convinced that, even with more than 11,000 days of data, the data can not support or identify such a general model; we can not achieve a well-behaved MCMC fitting algorithm. Hence, we turn to some model simplifications. We begin with specifications on  $\mathbf{A}(\mathbf{s}_t, t)$

- The first is a constant transformation matrix model  $\mathbf{A}(\mathbf{s}_t, t) = \mathbf{A}$  yielding

$$\mathbf{s}_{t+1} = \mathbf{A}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1} \quad (5.2)$$

This is a simple case of vector autoregressive (VAR) models, which have been widely used in multiple time series analysis (see, e.g., Sims (1972) and Enders (2003)).

- We next consider a spatially-varying transformation matrix. It is most convenient to assign a distinct transformation matrix to each of the tessellations induced by

the SOM nodes as described above. Let  $\mathbf{A}_l$  be the transformation matrix when  $\mathbf{s}_t \in \mathbf{\Lambda}_l$ , where  $\mathbf{\Lambda}_l$  is tessellation  $l$ , for  $l = 1, 2, \dots, L$ . Let  $Z_l$  be a binary indicator, i.e.,  $Z_l(\mathbf{s}_t) = 1$  if  $\mathbf{s}_t \in \mathbf{\Lambda}_l$ , and 0 otherwise. Then,

$$\mathbf{A}(\mathbf{s}_t, t) = \sum_{l=1}^L \mathbf{A}_l Z_l(\mathbf{s}_t) \quad (5.3)$$

This specification allows us to study regional change in the linear transformation.

- One of the important questions we seek to address in our climate study is whether there is a change in circulation patterns. That is, we assume the same collection of synoptic states continues to operate over time. However, temporal change would be manifested by a change in incidence rates of the states and thus would be modeled using time varying transition matrices. Let  $\mathbf{B}_m$  be the transformation matrix when  $t \in \mathbf{\Gamma}_m$ , where  $\mathbf{\Gamma}_m$  are disjoint time blocks, i.e.,  $\cup_{m=1}^M \mathbf{\Gamma}_m = \{1, \dots, T\}$ . Let  $V_m(t) = 1$  if  $t \in \mathbf{\Gamma}_m$ , and 0 otherwise. Now,

$$\mathbf{A}(\mathbf{s}_t, t) = \sum_{m=1}^M \mathbf{B}_m V_m(t) \quad (5.4)$$

Expression (5.4) enables us to study temporally varying linear transformation over suitable time scales, e.g., months, quarters or years.

- A spatially and temporally varying structure can be extended from the specifications described above in the form:

$$\mathbf{A}(\mathbf{s}_t, t) = \sum_{l,m} \mathbf{D}_{m,l} Z_l(\mathbf{s}_t) V_m(t) \quad (5.5)$$

Special cases of (5.5) include *separable* forms in space and time, e.g.,  $D_{m,l} = B_m A_l$  or  $D_{m,l} = A_l B_m$ . The former provides spatial transition followed by temporal transition, the latter vice versa.

The modeling in (5.3), (5.4), and (5.5) works at aggregated spatial and temporal scale. Similarly, we could add spatially and temporally aggregated intercepts. In particular, we

could introduce  $\boldsymbol{\eta}(\mathbf{s}_t, t) = \boldsymbol{\eta}$ ,  $\boldsymbol{\eta}(\mathbf{s}_t, t) = \sum_{l=1}^L \boldsymbol{\eta}_l Z_l(\mathbf{s}_t)$ , or  $\boldsymbol{\eta}(\mathbf{s}_t, t) = \sum_{m=1}^M \boldsymbol{\eta}_m V_m(t)$ . However, we view the role of the  $\boldsymbol{\eta}(\mathbf{s}_t, t)$ 's as introducing point level refinement to aggregated level affine transformations. We do so by introducing a bivariate spatial Gaussian realization intended to provide spatially dependent adjustments to the linear transformation specification. The adjustment at time  $t$  is  $\boldsymbol{\eta}(\mathbf{s}_t)$  yielding the model

$$\mathbf{s}_{t+1} = \mathbf{A}(\mathbf{s}_t, t)\mathbf{s}_t + \boldsymbol{\eta}(\mathbf{s}_t) + \boldsymbol{\epsilon}_{t+1} \quad (5.6)$$

We propose a coregionalization model for the bivariate Gaussian process realization in the spirit of Gelfand et al. (2004b). Let  $\mathbf{w}(\mathbf{s}) = (w_1(\mathbf{s}), w_2(\mathbf{s}))'$ , where  $w_1(\mathbf{s})$  and  $w_2(\mathbf{s})$  are uncorrelated spatial processes each with zero mean and unit variance. Coregionalization constructs a bivariate spatial process by linear transformation of these two independent univariate processes, i.e.,  $\boldsymbol{\eta}(\mathbf{s}) = (\eta_1(\mathbf{s}), \eta_2(\mathbf{s}))' = \mathbf{Q}(w_1(\mathbf{s}), w_2(\mathbf{s}))'$ , where  $\mathbf{Q}$  is a  $2 \times 2$  unknown coregionalization matrix and can be taken as lower triangular without loss of generality, i.e.,  $\mathbf{Q} = \begin{pmatrix} q_{11} & 0 \\ q_{12} & q_{22} \end{pmatrix}$ . An unusual aspect of the employment of this bivariate specification is that it provides a spatial surface to smooth all locations in the region while the observations are, in fact, a time series of locations. In other words, this bivariate spatial process is created for observed locations at multiple time points rather than multiple locations observed at the same time. The process realization reflects the spatial variation unexplained by the autoregressive component, regardless of the specific times of the observations.

The model in (5.6) is now completely specified. However, recall that we work with 11,315 days hence 11,315 locations in total. The joint distribution of the collection of 11,315  $\boldsymbol{\eta}(\mathbf{s})$ 's introduces an  $11,315 \times 11,315$  covariance matrix. To handle this dimension, we employ a version of the predictive process model described in Banerjee et al. (2007). Briefly, we consider a set of ‘‘knots’’  $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$  which forms a subset of the study

region in 2 dimensional space. The bivariate Gaussian process above would yield  $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim MVN(\mathbf{0}, C^*(\boldsymbol{\theta}))$  as its realizations over  $\mathcal{S}^*$ , where  $C^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$  is the corresponding  $m \times m$  covariance matrix. The predictive process model is defined as

$$\tilde{w}(\mathbf{s}) = E[w(\mathbf{s})|\mathbf{w}^*] = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*, \quad (5.7)$$

where  $\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$ . In fact,  $\tilde{w}(\mathbf{s})$  is a Gaussian Process with covariance function  $\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^{*T}(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}^*(\mathbf{s}'; \boldsymbol{\theta})$  where  $\mathbf{c}^*(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}_0, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$ . The realization of  $\tilde{w}(\mathbf{s})$  on any collection of sites are the interpolated predictions conditional upon the realization of  $w(\mathbf{s})$  over  $\mathcal{S}^*$ . To work with this process we only need to work with  $\mathbf{w}_1^*$ ,  $\mathbf{w}_2^*$  and the associated pair of  $m \times m$  correlation matrices.

## 5.5 Model fitting issues

VAR models are well-discussed in the literature (see, e.g., Lütkepohl (1993) and Zivot and Wang (2006)). Analysis within the Bayesian paradigm is presented in, e.g., Sims and Zha (1998), Sun and Ni (2004). We employ MCMC to fit the various submodels of (5.6) described in the previous section. In fact, we first discuss the computational issues in fitting the proposed models without spatial adjustment. Then we turn to issues in fitting the models with such adjustment.

### 5.5.1 Fitting models without spatial adjustment

For those proposed models without spatial adjustment, we illustrate the MCMC fitting procedure for model:  $\mathbf{s}_{t+1} = \sum_{l=1}^L \mathbf{A}_l Z_l(\mathbf{s}_t) \mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}$ .

Denote  $\mathbf{Z}(\mathbf{s}_t) = (Z_1(\mathbf{s}_t), \dots, Z_L(\mathbf{s}_t))$ ,  $\mathbf{x}_t = (\mathbf{Z}(\mathbf{s}_t) \otimes \mathbf{s}_t')$ ,  $\mathbf{Y}' = (\mathbf{s}_2, \dots, \mathbf{s}_T)$ ,  $\mathbf{X}' = (\mathbf{x}_1, \dots, \mathbf{x}_{T-1})$ ,  $\boldsymbol{\epsilon}' = (\boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_T)$ ,  $\boldsymbol{\Phi}' = (\mathbf{A}'_1, \dots, \mathbf{A}'_L)$ . Here  $\mathbf{Y}$  and  $\boldsymbol{\epsilon}$  are  $(T-1) \times 2$  matrices,  $\boldsymbol{\Phi}$  is a  $2L \times 2$  matrix of unknown transformation parameters,  $\mathbf{x}_t$  is a  $1 \times 2L$  row vector and  $\mathbf{X}$  is a  $(T-1) \times 2L$

matrix of observations. Then the VAR model can be written as

$$\mathbf{Y} = \mathbf{X}\Phi + \epsilon$$

The MLE's of  $\Phi$  and  $\Sigma$  are obtained by maximizing:

$$\begin{aligned} L(\Phi, \Sigma) &= \frac{1}{|\Sigma|^{(T-1)/2}} \exp\left\{-\frac{1}{2} \sum_{t=1}^{T-1} (\mathbf{s}_{t+1} - \mathbf{x}_t \Phi) \Sigma^{-1} (\mathbf{s}_{t+1} - \mathbf{x}_t \Phi)'\right\} \\ &= \frac{1}{|\Sigma|^{(T-1)/2}} \text{etr}\left\{-\frac{1}{2} (\mathbf{Y} - \mathbf{X}\Phi) \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\Phi)'\right\} \end{aligned}$$

We obtain MLE's of  $\Phi$  and  $\Sigma$  as  $\hat{\Phi}_M = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$  and  $\hat{\Sigma}_M = (\mathbf{Y} - \mathbf{X}\hat{\Phi}_M)'(\mathbf{Y} - \mathbf{X}\hat{\Phi}_M)/(T-1)$ .

Bayesian model fitting is completed by assigning prior distributions on the unknown parameters of interest. Denote  $\phi = \text{vec}(\Phi)$ , we assign  $\phi$  with a flat prior. We consider a noninformative Jeffreys prior for  $\Sigma$ , which, in our case, is  $\pi(\Sigma) \propto \frac{1}{|\Sigma|^{3/2}}$ .

Given  $\Sigma$ , we can directly sample  $\phi$  from its conditional distribution given by  $\pi(\phi|\Sigma, \mathbf{Y}) \sim MVN(\text{vec}(\hat{\Phi}_M), \Sigma \otimes (\mathbf{X}'\mathbf{X})^{-1})$ . Conditional on  $\phi$ ,  $\Sigma$  is updated using an Inverse Wishart  $((T-1)\hat{\Sigma}(\hat{\Phi}_M), T-2L)$ .

## 5.5.2 Fitting models with spatial adjustment

For the models with spatial adjustment, for convenience, we adopt an exponential correlation function for each of the two parent processes, hence bringing in two decay parameters  $\theta_1$  and  $\theta_2$ .<sup>2</sup> A uniform prior is assigned for each of  $\theta_1$  and  $\theta_2$  and updated using Metropolis steps. For the coregionalization matrix  $\mathbf{Q}$ , we assign truncated normal priors with positive support for the diagonal entries, and a normal prior for the off-diagonal entry. The entries in  $\mathbf{Q}$  are updated from normal distributions conditional on the other parameters. Samples

---

<sup>2</sup>We would not anticipate sensitivity in the bivariate predicted  $\eta$  surfaces to the choice of correlation function.

of  $\mathbf{W}_1^{*(b)}$  and  $\mathbf{W}_2^{*(b)}$  are generated in blocks from their multivariate normal posterior distributions, which in turn yields samples of  $\tilde{w}(\mathbf{s})^{(b)} = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}^{(b)})C^{*-1}(\boldsymbol{\theta}^{(b)})\mathbf{w}^{*(b)}$ . Substantial gains in computational efficiency are achieved by working with  $\mathbf{W}^*$  at a relatively small number of knots.

Each of the proposed models enables one day ahead prediction, i.e., the posterior predictive distribution of location at time  $t+1$  given location say  $\tilde{\mathbf{s}}$  at time  $t$ . This is implemented by composition; a posterior draw of the parameters in whatever version of (5.6) we fit, setting  $\mathbf{s}_t = \tilde{\mathbf{s}}$ , enables a predictive draw for the location at time  $t+1$ . Posterior samples,  $\tilde{\mathbf{s}}^{(b)}$  enable a density estimate for the transition distribution at any time and given any location. In addition, these models enable inference about the “transition distance” and the “transition angle” in 2-dimensional space.

In fact, again using posterior predictive samples, these models allow us to induce inference for categorical analysis at tessellation level. The probability of transitioning from  $\tilde{\mathbf{s}}$  to tessellation  $l$  can be straightforwardly estimated as well as the  $12 \times 12$  transition matrix from SOM node to SOM node (we omit details). This model-based estimate can be compared with the empirical estimate presented in Section 5.3 (Figure 5.3). Evidently, we can learn about the movement of the daily state vectors at any spatial scale in 2-dimensions. Working at the scale of the tessellations enables us to inform about circulation among synoptic weather states defined by the SOM nodes.

## 5.6 Model comparison and model results

Given the various possible model specifications detailed in Section 5.4, our first analysis goal would appear to be model comparison. We consider three criteria. First, we compare models in terms of one step ahead prediction performance. For each observation  $\mathbf{s}_t$  at  $t$ ,

samples of  $\mathbf{s}_{t+1}$  are drawn from the posterior predictive distribution as described above. Posterior means are adopted as the point estimates of the predicted positions for each of  $t = 2$  to  $t = T$ . The root mean square predictive errors (*RMSPE*) are computed to assess the predictive performance for each model

$$RM\hat{SPE} = \sqrt{\frac{1}{T-1} \sum_{i=2}^T \|\hat{\mathbf{s}}_t - \mathbf{s}_t\|^2} \quad (5.8)$$

where  $\hat{\mathbf{s}}_t$  is the mean of the posterior samples  $\{\hat{\mathbf{s}}_t^{(b)}\}$  for  $b = 1, 2, \dots, B$ .

A second comparison among models is to study the proportion of times the true locations lie in their associated 95% predictive interval under each model. This coverage proportion is denoted by  $\hat{r}$ . A third model selection criterion which is easily calculated from the posterior samples is the deviance information criterion (DIC) (Spiegelhalter et al. (2002)). DIC is a generalization of the AIC and BIC criteria and is defined as  $DIC = p_D + \bar{D}$  where  $p_D$  is a penalty for dimension and  $\bar{D}$  is a goodness of fit measure. Smaller values of DIC correspond to preferred models.

Table 5.4 summarizes the *RMSPE*, *DIC* score, and  $\hat{r}$  for a collection of models as indicated. (When time is included it is either blocked quarterly or annually. Based upon model fitting to more than 11,000 data points and using a large number of posterior samples (10,000), we are comfortable with the number of significant digits provided. First, all of the proposed autoregressive models are apparently superior to the random walk model in terms of predictive performance and DIC scores. Second, disappointingly, the models including spatial adjustment show no evidence of improving performance on data fitting and prediction; there appears to be little spatial dependence left in the autoregression residuals. Further disappointment emerges in the similarity of performance of these models; we can do better than a random walk model but can not find any interesting spatial or temporal structure.

We offer several thoughts in this regard. Perhaps the projection to a two dimensional

space which yields our bivariate time series has removed the interesting structure. In particular, the spatially varying covariate information associated with the 11 climate variables was used to create the projected locations in two dimensions; it is not available to explain the bivariate time series. Moreover, the spatial dependence that is induced in the two-dimensional space may have little to do with the spatial structure in the original 11 space-time processes. Finally, climatologists would assert that the SOM which was created is not intended for short term weather prediction; in capturing climate states, the SOM might be more appropriate for assessing regional climate change over a longer temporal scale (see Section 5.6). So, while our modeling goal here was to learn about spatial and temporal structure in the created SOM (and, what follows below indicates that there is still a story to tell), to learn about the space time structure in the original daily data a different dimension-reduction strategy might be more appropriate.

In any event, Model (9), which has transformation matrix  $A$  specified as tessellation and year, has the lowest prediction error in terms of  $RMSPE$ .<sup>3</sup> In addition, the  $\hat{r}$  for Model (9) is quite close to 0.95, as desired. As a result, we summarize results based on Model (9). Table 5.5 provides the posterior means for several parameters of interest and the corresponding 95% credible intervals under Model (9). We notice that estimates for the elements in  $\Sigma$  take large values, suggesting substantial unexplained variances in the SOM array. Again, this reflects our lack of covariate information but also comments upon the utility of SOM's for one-step ahead prediction of weather states.

Following the discussion in Section 5.5, we provide some illustrations of the induced categorical analysis at the scale of the tessellations. Table 5.6 displays the estimated frequency of transition angle for each of the 12 tessellations in the year 2000 and clearly demonstrates the cyclic nature of weather systems in the study region. For example, SOM

---

<sup>3</sup>We might speculate that annual blocking captures an El Nino effect.

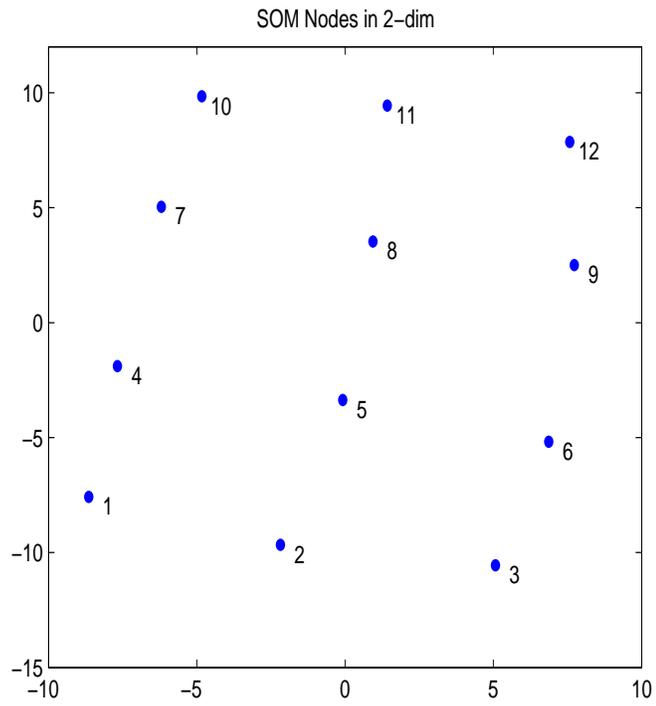
node group 12 prefers trajectory pathways towards SOM node group 11. Figure 5.5 shows the histograms of transition distance for each of the 12 tessellations in the year 2000, which suggests possible regional heterogeneity in the distribution of transition distance. In fact, synoptic weather in South Africa may display node-specific magnitude in volatility. For instance, SOM node group 2 which is expected in a transitional season, on average, tends to have higher transition distances than SOM node group 5 associated with characteristic summer circulation features. And this phenomenon might reveal high climate volatility associated with SOM node group 2 and relatively low climate volatility associated with SOM node group 5. Model (9) enables us to make inference of the transition matrices as well as the corresponding estimated errors year by year. The estimated transition matrix and the associated standard error are shown in Table 5.7 and Table 5.8 for the year 2000, which again support the findings on the clockwise cyclic evolution of the weather systems. The trajectories of transition probabilities can be aligned in each row for 31 consecutive years, from which we can examine the temporal behavior of the transition probabilities in synoptic climate states. As an illustration, Figure 5.6 plots the trajectories of three selected transition probabilities. For the general synoptic state associated with SOM node 1, the persistence probability reached above .35 in 1986 and then dropped below .15 in the subsequent year. The transition probability from SOM node group 11 to SOM node group 12 roughly remained at a stationary level during the 1970's then dramatically fell below .15 in 1983. It appears to be evolving with a more volatile trajectory since the late 1980's, finally reaching a peak which is above .35 in 2000.

## 5.7 Discussion

The use of SOM's has made considerable inroads in the meteorology community with regard to developing synoptic weather states to describe the collection of available weather

patterns across a region. Since the SOM technology represents high dimensional vectors in two dimensional space, we considered vector AR models to try to better understand the temporal evolution of SOM nodes. We have demonstrated that, while these SOM's may adequately span the high dimensional space of daily weather data vectors, they reveal little interesting spatial or temporal structure with regard to forecasting weather states.

Finally, we offer a potentially useful remark. The inability of the SOM to predict short term temporal evolution of these states does not imply that the SOM will not be useful for projection of future climate. If we assume that the SOM nodes describe regional weather well and that the same weather states continue to operate in the future, we may be able to forecast climate change in the form of a less uniform incidence of the different states than we currently see (Tables 5.1 and 5.2). Then, characterization of the SOM nodes as in Section 5.2 along with extraction of further node level descriptive features such as chance of rainfall, rainfall range, average temperature and temperature extremes will inform about the nature of the change. Such an investigation requires regional climate models at suitable spatial resolution that provide scenarios at the daily time scale and is beyond the scope of the current work.



**Figure 5.1:** Projected locations of SOM nodes in 2 dimensional space.

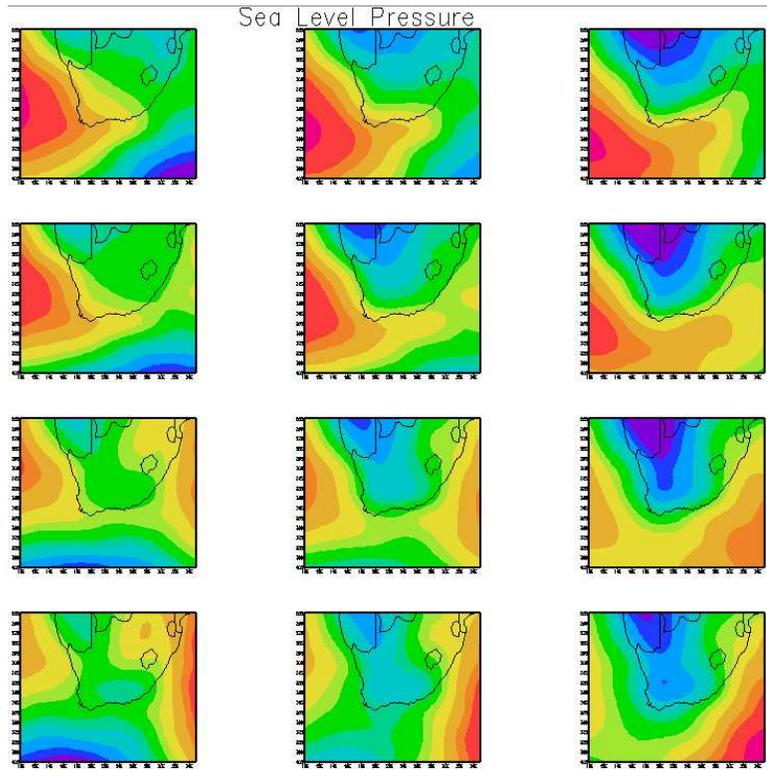


Figure 5.2: Sea level pressure associated with each node.

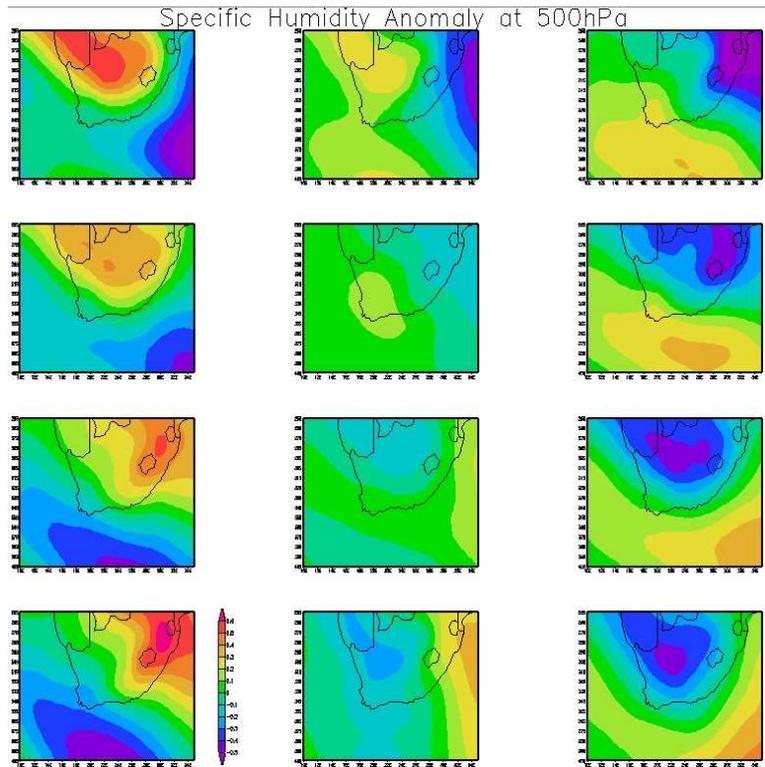
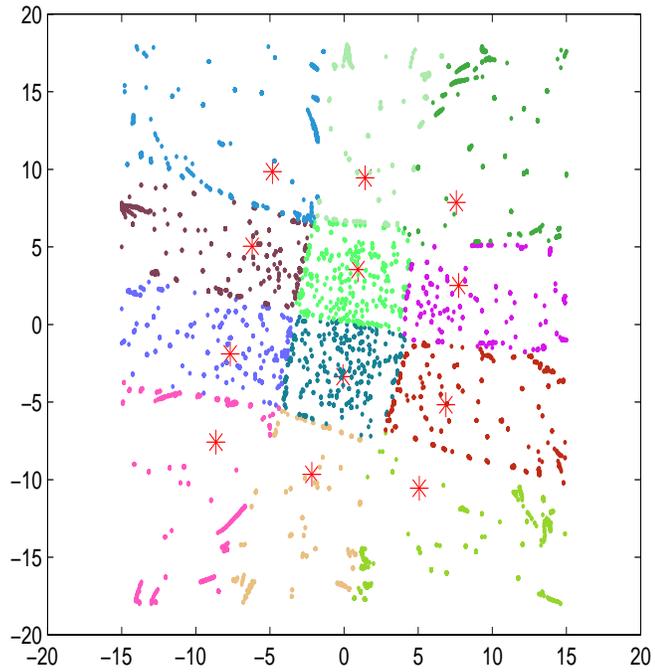


Figure 5.3: Maximum temperature surface associated with each node.



**Figure 5.4:** Mapped coordinates in the 2-dimensional space for each of the 11,315 days.

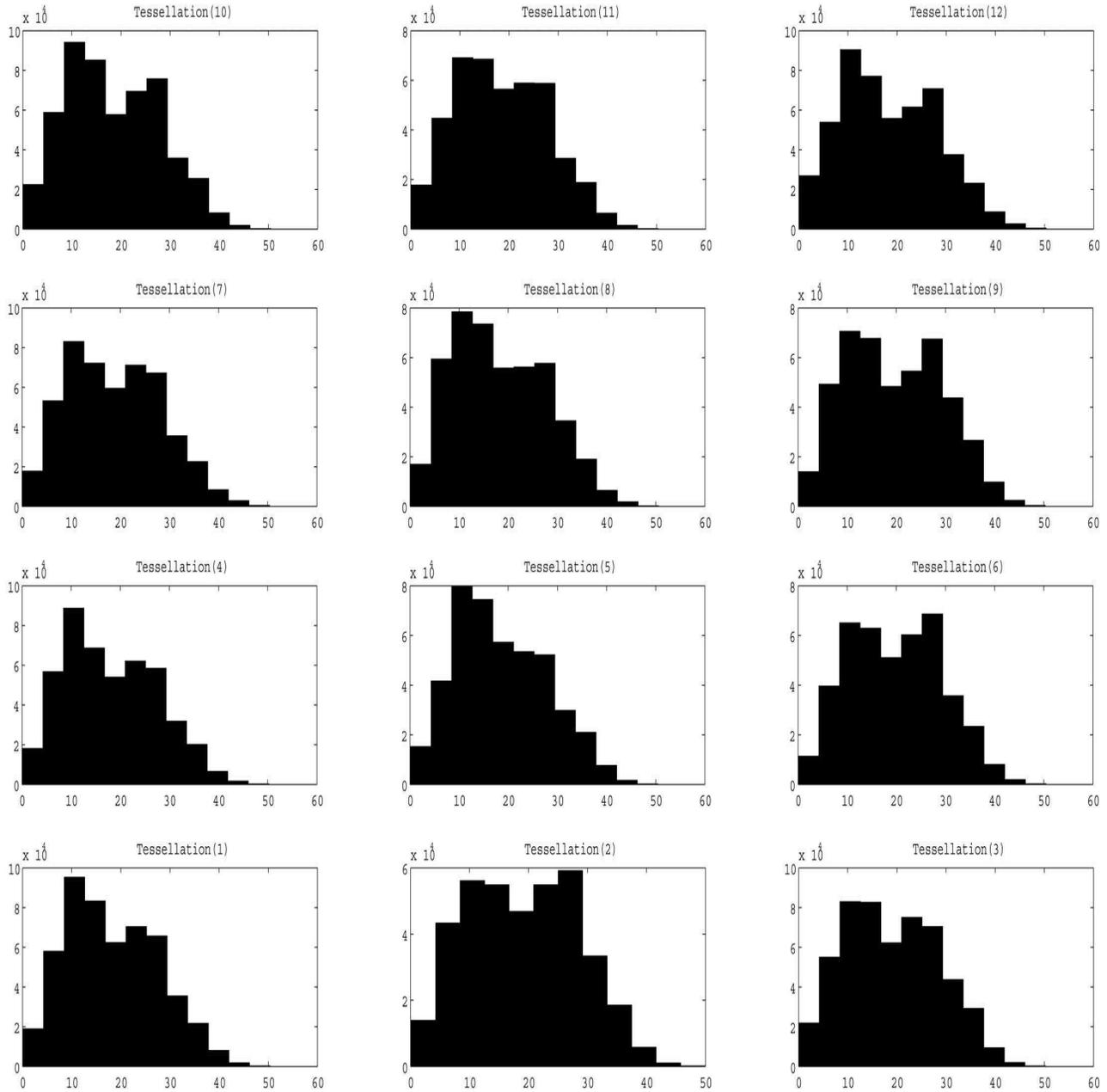
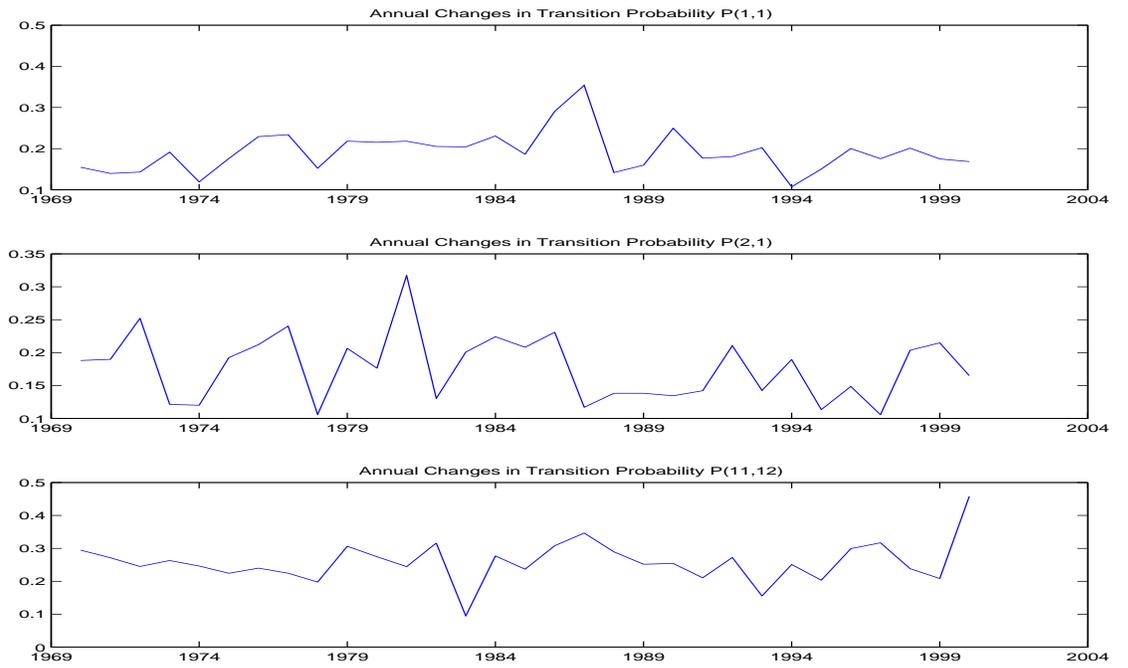


Figure 5.5: Histograms of transition distances for each SOM group.



**Figure 5.6:** Transition probabilities from 1970-2000.

1070 <sup>(10)</sup>	859 <sup>(11)</sup>	1017 <sup>(12)</sup>	(340, 335) <sup>(10)</sup>	(320, 263) <sup>(11)</sup>	(336, 319) <sup>(12)</sup>
989 <sup>(7)</sup>	920 <sup>(8)</sup>	910 <sup>(9)</sup>	(291, 327) <sup>(7)</sup>	(308, 274) <sup>(8)</sup>	(341, 269) <sup>(9)</sup>
935 <sup>(4)</sup>	869 <sup>(5)</sup>	857 <sup>(6)</sup>	(291, 285) <sup>(4)</sup>	(278, 274) <sup>(5)</sup>	(290, 296) <sup>(6)</sup>
1043 <sup>(1)</sup>	776 <sup>(2)</sup>	1070 <sup>(3)</sup>	(283, 347) <sup>(1)</sup>	(242, 273) <sup>(2)</sup>	(330, 388) <sup>(3)</sup>

**Table 5.1:** (left): Frequency of occurrence of each node over the entire study period, e.g., 935<sup>(4)</sup> indicates that the total number of occurrences of node 4 is 935; (right): Frequency of occurrence of each node during 1970-1979 and 1990-1999, e.g., (291, 285)<sup>(4)</sup> indicates that the number of occurrence of node 4 during 1970-1979 is 291, and that number is 285 during 1990-1999.

63 <sup>(10)</sup>	149 <sup>(11)</sup>	256 <sup>(12)</sup>	503 <sup>(10)</sup>	198 <sup>(11)</sup>	159 <sup>(12)</sup>
86 <sup>(7)</sup>	389 <sup>(8)</sup>	416 <sup>(9)</sup>	463 <sup>(7)</sup>	81 <sup>(8)</sup>	56 <sup>(9)</sup>
71 <sup>(4)</sup>	402 <sup>(5)</sup>	401 <sup>(6)</sup>	494 <sup>(4)</sup>	59 <sup>(5)</sup>	42 <sup>(6)</sup>
56 <sup>(1)</sup>	189 <sup>(2)</sup>	312 <sup>(3)</sup>	530 <sup>(1)</sup>	143 <sup>(2)</sup>	124 <sup>(3)</sup>

**Table 5.2:** (left) Frequency of occurrence of each node during summer; (right) Frequency of occurrence of each node during winter.

**Table 5.3:** Empirical transition probabilities. Each 4 by 3 sub-table in the following 4 by 3 array shows a set of transition probabilities. The array and sub-tables are arranged in the same way as in Table 5.1.

0.244	0.180	0.107	0.051	0.148	0.278	0.007	0.025	0.246
0.144	0.064	0.032	0.036	0.097	0.134	0.008	0.029	0.204
0.091	0.035	0.018	0.024	0.063	0.058	0.002	0.026	0.184
0.043	0.024	0.019	0.015	0.027	0.069	0.009	0.045	0.216
0.166	0.104	0.045	0.040	0.071	0.114	0.002	0.012	0.076
0.227	0.080	0.022	0.043	0.201	0.170	0.005	0.042	0.211
0.190	0.038	0.013	0.035	0.136	0.086	0.016	0.120	0.254
0.079	0.022	0.013	0.033	0.038	0.034	0.015	0.066	0.180
0.163	0.071	0.020	0.045	0.077	0.059	0.008	0.028	0.064
0.207	0.066	0.007	0.061	0.183	0.086	0.013	0.056	0.074
0.224	0.034	0.003	0.089	0.166	0.072	0.021	0.135	0.183
0.174	0.016	0.014	0.087	0.047	0.028	0.049	0.133	0.236
0.212	0.043	0.007	0.106	0.088	0.026	0.050	0.061	0.041
0.186	0.039	0.007	0.072	0.107	0.017	0.018	0.040	0.018
0.170	0.012	0.004	0.086	0.115	0.012	0.030	0.081	0.039
0.283	0.025	0.013	0.228	0.116	0.028	0.093	0.259	0.269

Model Specification	RMSPE	DIC ( $\times 10^5$ )	$\hat{r}$
Model 0: $\mathbf{s}_{t+1} = \mathbf{s}_t + \boldsymbol{\epsilon}$	11.419	1.6403	0.929
Model 1: $\mathbf{s}_{t+1} = \mathbf{A}\mathbf{s}_t + \boldsymbol{\epsilon}$	9.226	1.5353	0.944
Model 2: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{tessellation})\mathbf{s}_t + \boldsymbol{\epsilon}$	9.125	1.5317	0.943
Model 3: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{tessellation})\mathbf{s}_t + \boldsymbol{\eta}(\textit{quarter}) + \boldsymbol{\epsilon}$	8.943	1.5238	0.939
Model 4: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{quarter})\mathbf{s}_t + \boldsymbol{\epsilon}$	9.215	1.5337	0.942
Model 5: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{quarter})\mathbf{s}_t + \boldsymbol{\eta} + \boldsymbol{\epsilon}$	9.205	1.5337	0.941
Model 6: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{quarter})\mathbf{s}_t + \boldsymbol{\eta}(\textit{year}) + \boldsymbol{\epsilon}$	9.190	1.5297	0.947
Model 7: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{quarter}^*)\mathbf{s}_t + \boldsymbol{\epsilon}$	9.088	1.5378	0.948
Model 8: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{year})\mathbf{s}_t + \boldsymbol{\epsilon}$	9.224	1.5361	0.944
Model 9: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{tessellation}, \textit{year})\mathbf{s}_t + \boldsymbol{\epsilon}$	8.777	1.5407	0.956
Model 10: $\mathbf{s}_{t+1} = \mathbf{A}(\textit{tessellation}, \textit{quarter})\mathbf{s}_t + \boldsymbol{\epsilon}$	8.920	1.5247	0.940
Model 11: $\mathbf{s}_{t+1} = \mathbf{A}\mathbf{s}_t + \boldsymbol{\eta}(\textit{spatial}) + \boldsymbol{\epsilon}$	9.214	1.5421	0.934

**Table 5.4:** Performance of several VAR models using root mean square predictive errors (RMSPE), the deviance information criterion (DIC), and the empirical coverage probability  $\hat{r}$ .  $\mathbf{A}(\textit{tessellation})$  denotes regionally varying  $\mathbf{A}$  as described in (5.3),  $\mathbf{A}(\textit{year})$  denotes annually varying  $\mathbf{A}$  as described in (5.4),  $\boldsymbol{\eta}(\textit{quarter})$  denotes quarterly varying  $\mathbf{A}$  which is constant as year within each quarter.  $\boldsymbol{\eta}(\textit{quarter}^*)$  denotes quarterly varying  $\mathbf{A}$  which is also changing as year,  $\mathbf{A}(\textit{tessellation}, \textit{quarter})$  denotes  $\mathbf{A}$  as described in (5.5) and  $\boldsymbol{\eta}(\textit{spatial})$  denotes spatially varying adjustment as described in (5.6).

	Mean	0.025%	0.975%	SE
$a_{11}$	0.604	0.151	1.033	0.237
$a_{12}$	0.168	-0.220	0.618	0.226
$a_{21}$	-0.199	-0.739	0.409	0.286
$a_{22}$	-0.028	-0.678	0.503	0.289
$\Sigma_{11}$	40.394	39.212	41.457	0.566
$\Sigma_{12}$	17.572	16.544	18.648	0.531
$\Sigma_{22}$	68.441	66.429	70.297	0.961

**Table 5.5:** Posterior means and 95% credible intervals of  $\mathbf{A}_{tessellation1,year2000}$  and  $\Sigma$ .

Starting tessellation	$-45^{\circ} \dashrightarrow 45^{\circ}$	$45^{\circ} \dashrightarrow 135^{\circ}$	$135^{\circ} \dashrightarrow 225^{\circ}$	$225^{\circ} \dashrightarrow 315^{\circ}$
1	0.145	0.722	0.074	0.059
2	0.047	0.765	0.147	0.041
3	0.015	0.458	0.459	0.067
4	0.383	0.357	0.070	0.190
5	0.218	0.474	0.137	0.171
6	0.071	0.277	0.344	0.308
7	0.302	0.248	0.109	0.341
8	0.335	0.189	0.095	0.381
9	0.091	0.142	0.268	0.500
10	0.391	0.092	0.033	0.484
11	0.412	0.179	0.062	0.348
12	0.068	0.031	0.107	0.795

**Table 5.6:** Estimated frequency of transition angle from each tessellation towards each of the four directions for the year 1999.

**Table 5.7:** Estimated mean of the transition matrix in 2000. Each 4 by 3 sub-table in the following 4 by 3 array shows a set of transition probabilities. The array and sub-tables are arranged in the same way as in Table 5.1.

0.144	0.169	0.150	0.042	0.205	0.458	0.003	0.022	0.162
0.081	0.100	0.081	0.017	0.060	0.104	0.003	0.036	0.231
0.059	0.074	0.048	0.011	0.035	0.040	0.005	0.048	0.249
0.030	0.037	0.026	0.004	0.011	0.013	0.004	0.041	0.194
0.177	0.108	0.040	0.027	0.085	0.159	0.008	0.035	0.106
0.161	0.087	0.025	0.031	0.100	0.146	0.012	0.063	0.161
0.148	0.073	0.016	0.035	0.107	0.119	0.017	0.091	0.194
0.103	0.049	0.012	0.029	0.077	0.085	0.019	0.094	0.200
0.139	0.106	0.041	0.057	0.128	0.130	0.008	0.027	0.060
0.132	0.095	0.031	0.055	0.119	0.094	0.013	0.053	0.112
0.133	0.092	0.024	0.057	0.115	0.070	0.019	0.086	0.177
0.107	0.076	0.023	0.043	0.081	0.052	0.024	0.124	0.297
0.187	0.064	0.013	0.089	0.082	0.041	0.014	0.024	0.025
0.197	0.055	0.008	0.116	0.091	0.033	0.028	0.054	0.053
0.203	0.050	0.006	0.145	0.097	0.026	0.049	0.103	0.098
0.168	0.042	0.005	0.165	0.091	0.024	0.085	0.227	0.240

0.034	0.036	0.043	0.028	0.064	0.098	0.002	0.009	0.037
0.017	0.014	0.020	0.010	0.018	0.034	0.002	0.010	0.026
0.016	0.016	0.018	0.007	0.014	0.019	0.003	0.012	0.023
0.013	0.015	0.014	0.004	0.007	0.009	0.003	0.013	0.040
0.036	0.024	0.014	0.009	0.021	0.040	0.004	0.011	0.030
0.019	0.012	0.009	0.009	0.015	0.023	0.005	0.012	0.025
0.022	0.013	0.007	0.009	0.016	0.023	0.006	0.014	0.023
0.026	0.014	0.006	0.009	0.019	0.025	0.007	0.024	0.043
0.038	0.030	0.017	0.019	0.029	0.036	0.004	0.008	0.017
0.024	0.016	0.013	0.014	0.013	0.021	0.005	0.011	0.023
0.027	0.017	0.011	0.015	0.016	0.017	0.006	0.014	0.025
0.037	0.026	0.014	0.013	0.019	0.016	0.008	0.026	0.045
0.047	0.021	0.006	0.033	0.024	0.017	0.008	0.010	0.010
0.022	0.013	0.004	0.026	0.016	0.013	0.010	0.011	0.015
0.025	0.013	0.003	0.025	0.018	0.010	0.013	0.013	0.020
0.047	0.015	0.003	0.051	0.029	0.011	0.024	0.036	0.037

**Table 5.8:** Estimated mean of the standard errors of the transition matrix in 2000.

# Bibliography

- ATKINSON, P. and TATE, N. (2000). Spatial Scale Problems and Geostatistical Solutions: A Review. *The Professional Geographer*, **52** 607–623.
- BANERJEE, S., CARLIN, B. and GELFAND, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall.
- BANERJEE, S. and FINLEY, A. (2007). Bayesian multi-resolution modeling for spatially replicated data sets with application to forest biomass data. *Journal of Statistical Planning and Inference*, **137** 3193–3205.
- BANERJEE, S., GELFAND, A., FINLEY, A. and SANG, H. (2007). Gaussian Predictive Process Models for Large Spatial Datasets. *Journal of the Royal Statistical Society, Series B*, **submitted**.
- BARNETT, T., PIERCE, D. and SCHNUR, R. (2001). Detection of Anthropogenic Climate Change in the World's Oceans.
- BELLONE, E., HUGHES, J. and GUTTORP, P. (2000). A hidden Markov model for down-scaling synoptic atmospheric patterns to precipitation amounts. *Clim. Res.*, **15** 1–12.
- BESAG, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36** 192–236.
- BESAG, J., GREEN, P., HIGDON, D. and MENGENSEN, K. (1995). Bayesian Computation and Stochastic Systems. *Statistical Science*, **10** 3–41.
- BIAN, L. (1997). Multiscale nature of spatial data in scaling up environmental models. *Scale in remote sensing and GIS* 13–26.
- BUISHAND, D. H. L., T.A. and ZHOU, C. (2007). On spatial extremes: with application to a rainfall problem. *Annals of Statistics*.
- CAO, C. and LAM, N. (1997). Understanding the Scale and Resolution Effects in Remote Sensing and GIS. *Scale in Remote Sensing and GIS*.
- CHILES, J. and DELFINER, P. (1999). *Geostatistics: modeling spatial uncertainty*. Wiley.
- CHRISTENSEN, O., ROBERTS, G. and LUND, M. (2003). Robust MCMC for spatial GLMMs. *Preprints in Mathematical Sciences*, **23**.
- CHRISTENSEN, O. and WAAGEPETERSEN, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, **58** 280–286.
- COLES, S. (1993). Regional Modelling of Extreme Storms via Max-Stable Processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55** 797–816.
- COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

- COLES, S., HEFFERNAN, J. and TAWN, J. (1999). Dependence Measures for Extreme Value Analyses. *Extremes*, **2** 339–365.
- COLES, S. and TAWN, J. (1991). Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53** 377–392.
- COLES, S. and TAWN, J. (1996). A Bayesian Analysis of Extreme Rainfall Data. *Applied Statistics*, **45** 463–478.
- COOLEY, D., NAVEAU, P. and DAVIS, R. (2008). Dependence and Spatial Prediction in Max-Stable Random Fields. *University of Colorado*.
- COOLEY, D., NAVEAU, P. and PONCET, P. (2006). Variograms for spatial max-stable random fields. *Statistics for Dependent Data: STATDEP2005, Springer Lecture Notes in Statistics*. Springer, London.
- COOLEY, D., NYCHKA, D. and NAVEAU, P. (2007). Bayesian Spatial Modeling of Extreme Precipitation Return Levels. *Journal of The American Statistical Association*.
- CRANE, R. and HEWITSON, B. (2003). Clustering and upscaling of station precipitation records to regional patterns using self-organizing maps (SOMs). *Clim Res*, **25** 95–107.
- CRESSIE, N. (1993). Statistics for Spatial Data, revised edition. *Wiley Interscience*, **605** 10158–0012.
- DAHAN, E. and MENDELSON, H. (2001). An Extreme-Value Model of Concept Testing. *Management Science*, **47** 102–116.
- DE HAAN, L. (1985). *Extremes in higher dimensions: the model and some statistics*. Erasmus University.
- DE HAAN, L. and PEREIRA, T. (2006). Spatial extremes: Models for the stationary case. *Ann. Statist*, **34** 146–168.
- DIGGLE, P., TAWN, J. and MOYEED, R. (1998). Model-Based Geostatistics. *Applied Statistics*, **47** 299–350.
- DUNGAN, J., PERRY, J., DALE, M., LEGENDRE, P., CITRON-POUSTY, S., FORTIN, M., JAKOMULSKA, A., MIRITI, M. and ROSENBERG, M. (2002). A balanced view of scale in spatial statistical analysis. *Ecography*, **25** 626–640.
- DURMAN, C., GREGORY, J., HASSELL, D., JONES, R. and MURPHY, J. (2001). A comparison of extreme European daily precipitation simulated by a global and a regional climate model for present and future climates. *Quarterly Journal of the Royal Meteorological Society*, **127** 1005–1015.
- EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.
- ENDERS, W. (2003). *Applied econometric time series, 2nd Edition*. John Wiley and Sons.

- FERRANDEZ, J. M., DEL VALLE, D., RODELLAR, V. and GOMEZ, P. (1997). An automatic speech recognition system using time-delays self-organizing maps with physiological parametric extraction. *Acoustical Society of America Journal*, **102** 3165–+.
- FINLEY, S. H. B. S., A.O. and GELFAND, A. (2008, submitted). Improving the Performance of Predictive Process Modeling for Large Datasets. *Computational Statistics and Data Analysis*.
- FISHER, R. and TIPPETT, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24** 190.
- GELFAND, A., ECKER, M., KNIGHT, J. and SIRMANS, C. (2004a). The Dynamics of Location in Home Price. *The Journal of Real Estate Finance and Economics*, **29** 149–166.
- GELFAND, A., SCHMIDT, A., BANERJEE, S. and SIRMANS, C. (2004b). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, **13** 263–312.
- GELFAND, A. and SMITH, A. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85** 398–409.
- GELFAND, A. and VOUNATSOU, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4** 11–15.
- GELMAN, A. (2004). *Bayesian Data Analysis*. CRC Press.
- GUMBEL, E. and GOLDSTEIN, N. (1964). Analysis of Empirical Bivariate Extremal Distributions. *Journal of the American Statistical Association*, **59** 794–816.
- HEFFERNAN, J. and TAWN, J. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, **66** 497–530.
- HEGERL, G., ZWIERS, F., STOTT, P. and KHARIN, V. (2004). Detectability of Anthropogenic Changes in Annual Temperature and Precipitation Extremes. *Journal of Climate*, **17** 3683–3700.
- HEWITSON, B. and CRANE, R. (2002). Self-organizing maps: applications to synoptic climatology. *Climate Research*, **22** 13–26.
- HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, **5** 173–190.
- HIGDON, D. (2002). Space and space-time modeling using process convolutions. *Quantitative methods for current environmental issues* 37–56.
- HIGDON, D., SWALL, J. and KERN, J. (1999). Non-stationary spatial modeling. *Bayesian Statistics*, **6** 761–768.

- HUERTA, G. and SANSÓ, B. (2007). Time-varying models for extreme values. *Environmental and Ecological Statistics*, **14** 285–299.
- HUGHES, J., GUTTORP, P. and CHARLES, S. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Applied Statistics*, **48** 15–30.
- JIN, X., CARLIN, B. and BANERJEE, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, **61** 950–961.
- JOE, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall/CRC.
- KALNAY, E., KANAMITSU, M., KISTLER, R., COLLINS, W., DEAVEN, D., GANDIN, L., IREDELL, M., SAHA, S., WHITE, G., WOOLLEN, J. ET AL. (1996). The NCEP/NCAR reanalysis 40-year project. *Bull. Am. Meteorol. Soc.*, **77** 437–471.
- KAMMANN, E. and WAND, M. (2003). Geoadditive models. *Journal of the Royal Statistical Society Series C(Applied Statistics)*, **52** 1–18.
- KASKI, S. (1997). Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series*, **82** 57.
- KAUFMAN, C. (2006). *Covariance Tapering for Likelihood Based Estimation in Large Spatial Datasets*. Ph.D. thesis, PhD thesis, Carnegie Mellon University.
- KHARIN, V. and ZWIERS, F. (2005). Estimating Extremes in Transient Climate Change Simulations. *Journal of Climate*, **18** 1156–1173.
- KOHONEN, T. (1995). Self-Organising Maps. *Springer Series in Information Sciences*, **30**.
- KOTZ, S. and NADARAJAH, S. (2000). *Extreme Value Distributions: Theory and Applications*. World Scientific.
- LEADBETTER, M., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag New York.
- LEVITUS, S., ANTONOV, J., BOYER, T. and STEPHENS, C. (2000). Warming of the World Ocean. *Science*, **287** 2225.
- LOU THOMPSON, M., REYNOLDS, J., COX, L., GUTTORP, P. and SAMPSON, P. (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, **35** 617–630.
- LÜTKEPOHL, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer.
- MARCEAU, D. and HAY, G. (1999). Remote sensing contributions to the scale issue. *Canadian Journal of Remote Sensing*, **25** 357–366.
- MATHERON, G. (1982). Pour une analyse krigéante des données régionalisées. *Publication N-732, Centre de Géostatistique, Fontainebleau, France*.

- NELSEN, R. (2006). *An Introduction to Copulas*. Springer.
- PICKANDS, J. (1981). Multivariate extreme value distributions. *Proc. 43rd Sess. Int. Statist. Inst.*, **49** 31.
- QUATTROCHI, D. and GOODCHILD, M. (1997). *Scale in Remote Sensing and GIS*. CRC Press.
- RESNICK, S. (1987). Extreme Values, Regular Variation, and Point Processes. *New York*.
- RESNICK, S. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer.
- ROBERT, C. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*. Springer.
- ROBERTS, S. (2000). Extreme value statistics for novelty detection in biomedical dataprocessing. *Science, Measurement and Technology, IEE Proceedings-*, **147** 363–367.
- SANG, H. and GELFAND, A. (2008). *Hierarchical Modeling for Extreme Values Observed over Space and Time*. Environmental and Ecological Statistics.
- SCHLATHER, M. and TAWN, J. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, **90** 139–156.
- SIMS, C. (1972). Money, Income, and Causality. *The American Economic Review*, **62** 540–552.
- SIMS, C. and ZHA, T. (1998). Bayesian Methods for Dynamic Multivariate Models. *International Economic Review*, **39** 949–968.
- SMITH, R. (1990). Max-stable processes and spatial extremes. *Unpublished manuscript*.
- SPIEGELHALTER, D., BEST, N., CARLIN, B. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64** 583–639.
- STRAMER, O. and TWEEDIE, R. (1999). Langevin-Type Models II: Self-Targeting Candidates for MCMC Algorithms\*. *Methodology and Computing in Applied Probability*, **1** 307–328.
- SUN, D. and NI, S. (2004). Bayesian analysis of vector-autoregressive models with noninformative priors. *Journal of Statistical Planning and Inference*, **121** 291–309.
- TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. and GOLUB, T. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, **96** 2907–2912.
- TAWN, J. (1988). Bivariate Extreme Value Theory: Models and Estimation. *Biometrika*, **75** 397–415.

- TYSON, P. and PRESTON-WHYTE, R. (2000). *The Weather and Climate of Southern Africa*. Oxford University Press.
- WACKERNAGEL, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer.
- WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.
- ZHANG, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, **99** 250–262.
- ZIVOT, E. and WANG, J. (2006). *Modeling Financial Time Series with S-plus*. Springer.

# Biography

Huiyan Sang is a Ph.D. student of Statistical Science at Duke University in Durham, NC. She received the B.S. degree in mathematics and applied mathematics (2004) from Peking University in Beijing. Her research interests include spatial statistics, modeling extreme values and complex modeling for environmental data.