DISSERTATION

JOINT TAIL MODELING VIA REGULAR VARIATION WITH APPLICATIONS IN CLIMATE AND ENVIRONMENTAL STUDIES

Submitted by

Grant B. Weller

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2013

Doctoral Committee:

Advisor: Dan Cooley

F. Jay Breidt
Donald Estep
Russ Schumacher

ABSTRACT

JOINT TAIL MODELING VIA REGULAR VARIATION
WITH APPLICATIONS IN CLIMATE AND ENVIRONMENTAL STUDIES

This dissertation presents applied, theoretical, and methodological advances in the statistical analysis of multivariate extreme values, employing the underlying mathematical framework of multivariate regular variation. Existing theory is applied in two studies in climatology; these investigations represent novel applications of the regular variation framework in this field. Motivated by applications in environmental studies, a theoretical development in the analysis of extremes is introduced, along with novel statistical methodology.

This work first details a novel study which employs the regular variation modeling framework to study uncertainties in a regional climate model's simulation of extreme precipitation events along the west coast of the United States, with a particular focus on the Pineapple Express (PE), a special type of winter storm. We model the tail dependence in past daily precipitation amounts seen in observational data and output of the regional climate model, and we link atmospheric pressure fields to PE events. The fitted dependence model is utilized as a stochastic simulator of future extreme precipitation events, given output from a future-scenario run of the climate model. The simulator and link to pressure fields are used to quantify the uncertainty in a future simulation of extreme precipitation events from the regional climate model, given boundary conditions from a general circulation model.

A related study investigates two case studies of extreme precipitation from six regional climate models in the North American Regional Climate Change Assessment Program (NARCCAP). We find that simulated winter season daily precipitation along the Pacific coast exhibit tail dependence to extreme events in the observational record. When considering summer season daily precipitation over a central region of the United States, however, we find almost no correspondence between extremes simulated by NARCCAP and those seen

in observations. Furthermore, we discover less consistency among the NARCCAP models in the tail behavior of summer precipitation over this region than that seen in winter precipitation over the west coast region. The analyses in this work indicate that the NARCCAP models are effective at downscaling winter precipitation extremes in the west coast region, but questions remain about their ability to simulate summer-season precipitation extremes in the central region.

A deficiency of existing modeling techniques based on the multivariate regular variation framework is the inability to account for *hidden regular variation*, a feature of many theoretical examples and real data sets. One particular example of this deficiency is the inability to distinguish asymptotic independence from independence in the usual sense. This work develops a novel probabilistic characterization of random vectors possessing hidden regular variation as the sum of independent components. The characterization is shown to be asymptotically valid via a multivariate tail equivalence result, and an example is demonstrated via simulation.

The sum characterization is employed to perform inference for the joint tail of random vectors possessing hidden regular variation. This dissertation develops a likelihood-based estimation procedure, employing a novel version of the Monte Carlo expectation–maximization algorithm which has been modified for tail estimation. The methodology is demonstrated on simulated data and applied to a bivariate series of air pollution data from Leeds, UK. We demonstrate the improvement in tail risk estimates offered by the sum representation over approaches which ignore hidden regular variation in the data.

ACKNOWLEDGEMENTS

I am grateful to many people for their contributions toward the development of my research career and the work presented in this dissertation. I am proud to have Dan Cooley as an advisor, mentor, and friend. His ability to see the bigger picture and provide guidance in the early stages of this research was crucial to its development. Given his past career as a high school instructor, it is no surprise that he is one of the best teachers I have ever had. I am also grateful to him for exposing me to the broader statistical community by funding my travel to a number of research workshops and conferences.

I was fortunate to have been a visiting scientist to the National Center for Atmospheric Research (NCAR) for the past two years, and a number of people in NCAR's Institute for Mathematics Applied to Geosciences (IMAGe) group have contributed to the development of this research. My collaborators Steve Sain, Melissa Bukovsky, and Linda Mearns have taught me more than I ever imagined knowing about climate science and the statistical problems it presents. I have had many stimulating discussions with IMAGe postdocs Will Kleiber, Tammy Greasby, and Matt Heaton. Every conversation I have with Doug Nychka seems to result in a bit of good advice, regardless of the topic. I am forever indebted to Tim Hoar for helping me to understand the NCAR computing systems.

I was also fortunate to be a visitor to the Statistical and Applied Mathematical Sciences Institute (SAMSI) in North Carolina in the fall of 2011, and to participate in SAMSI's program on Uncertainty Quantification. With the help of the working group in which I participated, the foundations of much of this research were developed in my mind during this time. In particular, Robert Wolpert provided useful feedback which helped me understand the broad topic of multivariate extremes. Richard Smith also found some time away from his duties as Director of SAMSI to support discussions and take Dan and me on several runs in the Research Triangle region.

I thank my committee members Jay Breidt, Don Estep, and Russ Schumacher for providing feedback on the work in this dissertation. Useful feedback has also been provided by Sid Resnick and Anja Janßen.

I have greatly enjoyed my five years in Fort Collins and the CSU Department of Statistics. I have learned a great deal from the faculty members, especially Phil Chapman and Jennifer Hoeting, for whom I have served as teaching assistants. It has been a pleasure to know my fellow students, particularly Wade Herndon and Bruce Bugbee, with whom I have shared an office for the duration of my time here. It has also been exciting to see the department change and grow, first under the direction of Jay Breidt, and now under Jean Opsomer.

I acknowledge financial support from National Science Foundation grant DMS-0905315, the Weather and Climate Impacts Assessment Science Program at NCAR, the 2011-2012 SAMSI Program on Uncertainty Quantification, and Environmental Protection Agency grant EPA-STAR RD-83522801-0.

Finally, I could not have reached this point in my career without the support and encouragement of my parents. They instilled the value of hard work in me at a young age and have always encouraged me to pursue my education. I thank them for believing in me and my abilities even when I didn't have full confidence in myself. I may never fully understand how a small-town Minnesota farm boy turned into a Ph.D. statistician, but I know they were instrumental in making it happen.

DEDICATION

to my parents, Jim and Jackie

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years, a number of costly natural disasters and a severe financial crisis in the United States have given way to an increased awareness of extreme events from both scientists and the general public. While these extremes are rare by definition, they typically inflict great human, economic, and societal impacts. Despite the increased scrutiny, in many cases the nature of extreme events, as well as their underlying causes, are not well understood.

Motivated in part by the high cost of these recent catastrophic events, scientists have placed increased focus on the study of extremes. A recent report from the field of climatology (Peterson et al., 2012) directly linked past observed extreme weather events to human-induced climate change. In the field of finance, several works (e.g., Mikosch, 2006) have criticized the use of mathematical models which are unable to adequately describe the likelihood of a severe crisis. Many fields are now exploring the development of mathematical and statistical models which aim to characterize the likelihood and severity of extreme events, as well as the underlying mechanisms which may drive them.

The work in this dissertation centers on the analysis of extreme values. Fundamentally, the aim of an extreme value analysis is to describe the upper (or lower) tail of a probability distribution. In the univariate case, one wishes to describe the upper tail of the distribution of a single quantity, such as daily river flow volume at a single station or daily precipitation at a weather station. In the multivariate setting, one must describe the dependence between extremes of several variables. For example, a financial analyst may wish to estimate the probability of multiple securities in a portfolio simultaneously experiencing losses exceeding

high thresholds. Many extreme value problems also involve extrapolation: one may wish to estimate probabilities of events falling further into the tail than those which have been observed.

From a statistical perspective, the nature of extremes often renders the use of 'ordinary' statistical methods inappropriate for their analysis. For example, the central limit theorem requires only that the second moment of a univariate distribution be finite, and many statistical analyses focus on estimation of means and variances. In 'ordinary' multivariate problems, a typical aim is to study the covariances or correlations between two or more quantities. However, the usual summary measures such as means, variances, and correlations often cannot fully describe the tail of a distribution. The statistical methods described and employed in this dissertation are based on underlying results from probability theory which specifically describe the tail.

Extreme value theory (EVT), the branch of probability and statistics which aims to describe the upper (or lower) tail of a random variable, vector, or process, has roots in the early to middle part of the 20th century. One of the first fields to use EVT extensively was hydrology, which employed it to estimate quantities such as a "100-year flood". EVT has also been employed in fields such as engineering, financial risk analysis, and, more recently, climatology. There are a number of informative books on the subject of EVT, including Resnick (1987), Embrechts et al. (1997), and de Haan and Ferreira (2006), which provide a thorough background on the probability theory underlying the study of extremes. Beirlant et al. (2004) and Coles (2001) provide statistical approaches to the analysis of extremes, focusing on applications. The notation employed in this dissertation generally follows Resnick (2007), which offers a thorough treatment of regular variation, and whose framework is used to develop the methodologies of this work.

1.2 Outline and Links to Publications

This dissertation details applied, theoretical, and methodological advances in the study of extreme values, employing the multivariate regular variation framework. After reviewing classical EVT results, in this chapter we introduce the regular variation framework and provide links to the classical theory. Chapter 2 employs the regular variation theory in a study of the pineapple express, a weather phenomenon which is responsible for extreme precipitation along the Pacific coast of North America. This work further draws a link between large-scale atmospheric processes and extreme precipitation events produced by this phenomenon. Chapter 3 examines a study of extreme precipitation from a suite of deterministic climate models, using the regular variation theory to study the ability of such models to capture extreme precipitation events. These chapters employ existing statistical techniques in a novel examination of extremes in climatology.

Chapter 4 addresses the concept of *hidden regular variation*, which arises in both theoretical examples and real data. When hidden regular variation is present, statistical models for extremes based on the regular variation framework break down, as the limiting measure which characterizes extremal dependence is degenerate on some joint tail regions. We introduce a probabilistic characterization for random vectors possessing hidden regular variation as the sum of independent heavy-tailed components. Asymptotic justification for this characterization is provided by a tail equivalence theorem, and the characterization is demonstrated to exhibit useful finite-sample properties.

Chapter 5 introduces methodology for performing inference for random vectors possessing hidden regular variation based on the characterization introduced in Chapter 4. We introduce a novel modification of the expectation–maximization algorithm for performing maximum likelihood inference for the joint tail of a random vector. The methodology is demonstrated through simulation studies and applied to a bivariate series of air pollution data from Leeds, UK.

Table 1.1: References for published portions of this dissertation, by chapter.

Chapter	Primary Reference
2	Weller, G., Cooley, D., and Sain, S. (2012). An investigation of the pineapple express phenomenon via bivariate extreme value theory. *Environmetrics*, 23(5):420–439.
3	Weller, G., Cooley, D., Sain, S., Bukovsky, M., and Mearns, L. (2013). Two case studies on NARCCAP precipitation extremes. *submitted*.
4	Weller, G. and Cooley, D. (2012). An alternative characterization of hidden regular variation in joint tail modeling. Technical report, Colorado State University Department of Statistics.
5	Weller, G. and Cooley, D. (2013). A sum characterization of hidden regular variation in joint tail modeling with likelihood inference via the Monte Carlo expectation–maximization algorithm. *submitted*.

We conclude in Chapter 6 with a summary and discussion. This chapter also briefly discusses possible future extensions of the work presented in this dissertation.

Portions of this dissertation also appear in jointly authored papers which have been published or submitted for publication. In all cases, the author of this dissertation appears as the first author on these papers. Table 1.1 provides primary references for material in each chapter of the body of this dissertation. References to material which has been either previously published or submitted for publication additionally appear in the introductory section of each chapter, and these papers are cited in the References section.

1.3 Classical Extreme Value Theory

The classical results in EVT arise from the study of maxima, and the univariate and multivariate cases are presented here. Analogous results exist for maxima of stochastic processes; see de Haan (1984) and Smith (1990) for details.

1.3.1 Univariate Case

Classical univariate EVT dates back to the works of Fisher and Tippett (1928) and Gnedenko (1943), which studied the limiting distribution of normalized block maxima. Consider

independent copies $X_1, ..., X_n$ of a random variable X, and define $M_n = \vee_{i=1}^n X_i$. If one can find normalizing sequences $a_n \geq 0$ and $b_n \in \mathbb{R}$ such that

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \longrightarrow G(z) \tag{1}$$

as $n \to \infty$, where G is non-degenerate, then G must be of the form

$$G(z) = \begin{cases} \exp\{-(1+\xi z)^{-1/\xi}\} & \text{if } \xi \neq 0, \\ \exp\{-\exp(-z)\} & \text{if } \xi = 0. \end{cases}$$

The parameter ξ determines the shape of the tail: $\xi < 0$ (reverse Weibull) corresponds to a distribution with a finite upper limit, $\xi = 0$ (Gumbel) corresponds to an exponentially-decaying tail, and $\xi > 0$ (Fréchet) corresponds to a (heavy) tail which decays like a power function on z. Distributions of X for which (1) is satisfied are said to be in the domain of attraction (DOA) of G. For example, the Beta distribution is in the DOA of G with $\xi < 0$, the Gaussian and Gamma distributions are in the DOA of G with $\xi = 0$, and the Student t distribution is in the DOA of G with $\xi > 0$.

In practice, the result (1) can be used to perform inference for block (e.g., annual or seasonal) maximum data. Given m replicates of the maxima of n independent realizations $M_{n,1}, ..., M_{n,m}$, one may assume the limit (1) as an equality. Absorbing the normalizing sequences a_n and b_n into location and scale parameters leads to the three-parameter generalized extreme value (GEV) distribution. The GEV can be fit to the data $\{M_{n,j}\}_{j=1}^m$ via maximum likelihood (Smith, 1985) or probability weighted moments (Hosking et al., 1985), and estimates of high quantiles ("return levels") or small probabilities ("return periods") may be obtained from the fitted model, along with associated uncertainties.

1.3.2 Multivariate Case

In the multivariate setting, classical EVT studies the limiting distribution of the vector of appropriately normalized componentwise maxima. Assume one has independent replicates $\mathbf{X}_1, ..., \mathbf{X}_n$ of a d-dimensional random vector $\mathbf{X}$, and define $\mathbf{M}_n = (\vee_{i=1}^n X_{i,1}, ..., \vee_{i=1}^n X_{i,d})^T$. If there exist normalizing sequences of vectors $\mathbf{a}_n \geq \mathbf{0}$ and $\mathbf{b}_n \in \mathbb{R}^d$ such that

$$\mathbb{P}\left[\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \leq \mathbf{z}\right] \longrightarrow G(\mathbf{z}) \tag{2}$$

(nondegenerate) as $n \to \infty$ (where all operations are taken componentwise), then G is called a multivariate extreme value distribution (MVEVD) with univariate GEV marginals (de Haan and Resnick, 1977; Resnick, 1987). While the marginal distributions of G are GEV and thus fully parameterized, no finite parameterization exists for the dependence structure of the d components. Some parametric subfamilies have been developed, and are often described for the case where G has unit Fréchet marginal distributions; that is, $G_j(z_j) = \exp\{-z_j^{-1}\}$ for $j = 1, ..., d$. Examples are given in Coles (2001); more recently developed models are provided by Cooley et al. (2010) and Ballani and Schlather (2011).

As the marginal distributions of G in (2) are fully parameterized and dependence models are developed under the assumption of common marginals, statistical modeling of multivariate extremes typically involves two steps: marginal estimation and dependence estimation. Given m replicates $\mathbf{M}_{n,1}, ..., \mathbf{M}_{n,m}$ of the vector of componentwise maxima of n independent realizations, a typical approach is to fit a GEV distribution to marginal data $M_{j,n,1}, ..., M_{j,n,m}$ for $j = 1, ..., d$ and use the fitted GEV to transform each margin to follow a unit Fréchet distribution. The dependence structure can then be modeled via an existing parametric form or estimated nonparametrically. Likelihood-based modeling of marginal and dependence effects may also be performed simultaneously; see e.g. Coles and Tawn (1991).

1.4 Multivariate Regular Variation

Multivariate regular variation on cones provides a probabilistic framework for describing the joint upper tail of a random vector and modeling multivariate threshold-exceedance data. An intuitive description is that the joint tail of a regular varying random vector decays like a power function. A decomposition into polar coordinates arises, and tail dependence can be characterized by a limiting angular measure on the unit sphere under a fixed norm. We will focus on the multivariate case; a comprehensive treatment of regular variation of functions in the univariate case can be found in Bingham et al. (1989) and de Haan (1970).

1.4.1 Definition

Consider a d-dimensional random vector $\mathbf{Z}$ taking values in $[\mathbf{0}, \boldsymbol{\infty})$ where $\mathbf{0} = (0, ..., 0)^T$ and $\boldsymbol{\infty}$ is defined analogously. A cone $\mathfrak{C}$ of $\mathbb{R}^d$ is defined such that for any set $A \subseteq \mathfrak{C}$, $tA \subseteq \mathfrak{C}$ for any $t > 0$, and multivariate regular variation considers sets on the cone $\mathfrak{C} = [\mathbf{0}, \boldsymbol{\infty}] \setminus \{\mathbf{0}\}$. Denote by $M_+(\mathfrak{C})$ the space of nonnegative Radon measures on $\mathfrak{C}$. Chapter 6 of Resnick (2007) presents nine equivalent definitions of multivariate regular variation; here we present three which are most relevant to this dissertation.

The first definition formulates multivariate regular variation in terms of the probability of (appropriately normalized) $\mathbf{Z}$ taking values in some set on $\mathfrak{C}$.

Definition 1. *The random vector $\mathbf{Z}$ is regular varying if there exists a function $b(t) \to \infty$ and Radon measure ν on $\mathfrak{C}$ such that*

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{b(t)} \in \cdot\right] \xrightarrow{v} \nu(\cdot)$$

in $M_+(\mathfrak{C})$ as $t \to \infty$, where $\xrightarrow{v}$ denotes vague convergence of measures.

An important idea emerging from Definition 1 is that multivariate regular variation describes the asymptotic distribution of only the upper tail of the random vector $\mathbf{Z}$. Because

the normalizing function $b(t) \to \infty$ and ν measures sets on $\mathfrak{C}$, which excludes the origin $\{\mathbf{0}\}$, regular variation considers only the upper tail.

It follows from Definition 1 that the limit measure has the following homogeneity property: for any set $A \subseteq \mathfrak{C}$,

$$\nu(tA) = t^{-\alpha}\nu(A) \tag{3}$$

for $t > 0$, where $\alpha > 0$ is called the *tail index*. The normalizing function $b(t)$ is itself regular varying of index $1/\alpha$; that is, $b(t) \sim t^{1/\alpha}$ for large t. We denote this by $b(t) \in RV_{1/\alpha}$.

The homogeneity property (3) suggests a transformation to polar coordinates and leads to the second definition of multivariate regular variation. Fix a norm $\|\cdot\|$ on $\mathfrak{C}$, and define 'radial' and 'angular' components $R = \|\mathbf{Z}\|$ and $\mathbf{W} = \|\mathbf{Z}\|^{-1}\mathbf{Z}$. Denote the unit sphere under the chosen norm as $\mathcal{N} = \{\mathbf{z} \in \mathfrak{C} : \|\mathbf{z}\| = 1\}$.

Definition 2. *The random vector* $\mathbf{Z}$ *is regular varying if there exists a function* $b(t) \to \infty$ *and probability measure* H *on* $\mathcal{N}$ *such that for any Borel set* $B \subseteq \mathcal{N}$,

$$t\mathbb{P}\left[\frac{R}{b(t)} > r, \mathbf{W} \in B\right] \xrightarrow{v} cr^{-\alpha}H(B)$$

as $t \to \infty$, *for some* $c \in (0, \infty)$.

Definition 2 states that in the limit, the radial and angular components of the random vector $\mathbf{Z}$ become independent. The probability measure H is called the *angular measure*, and it completely characterizes the limiting tail dependence structure of $\mathbf{Z}$.

Denote by ϵ_x a probability measure placing unit mass at the point x, and denote by $M_p((0, \infty] \times \mathcal{N})$ the space of Radon point measures on the set $(0, \infty] \times \mathcal{N}$. When performing inference from independent samples $\mathbf{Z}_1, ..., \mathbf{Z}_n$ of the random vector $\mathbf{Z}$, the following definition will be useful:

Definition 3. *The random vector* $\mathbf{Z}$ *is regular varying if there exists a sequence* $b_n \to \infty$ *and probability measure* H *on* $\mathcal{N}$ *such that*

$$\sum_{i=1}^{n} \epsilon_{(R_i/b_n, \mathbf{W}_i)} \xrightarrow{d} PRM(\nu_\alpha \times H)$$

in $M_p((0,\infty] \times \mathcal{N})$ *as* $n \to \infty$*, where* ν_α *is a Pareto measure, i.e.,* $\nu_\alpha((r,\infty]) = cr^{-\alpha}$ *for constant* $c \in (0,\infty)$*; and* $PRM(\mu)$ *denotes a Poisson random measure with intensity measure* μ*.*

Definition 3 states that extremes of suitably normalized realizations of $\mathbf{Z}$ converge in distribution to a non-homogeneous Poisson point process on $\mathfrak{C}$ with intensity measure given by the limiting product measure in Definition 2. Here the normalizing sequence can be defined by $b_n = b(n)$, where $b(t)$ is as in Definitions 1 and 2. This limiting Poisson process will provide the probabilistic framework for performing likelihood-based inference for the random vector $\mathbf{Z}$.

Figure 1.1 provides an illustration of regular variation in the two-dimensional case. The left panel of Figure 1.1 shows a realization of $n = 2500$ points from a bivariate regular varying distribution with tail index $\alpha = 1$. Displayed in the right panel is a histogram of angular components values for realizations exceeding a fixed radial component threshold (in terms of the L_1 norm), shown in the left panel. Superimposed on the histogram is the density of the angular measure H from which realizations were simulated.

In the two-dimensional case, the unit sphere $\mathcal{N}$ is one-dimensional, and it is helpful to consider two limiting cases of the form of H. If Z_1 determines Z_2 exactly, H consists of a single point mass on the interior of $\mathcal{N}$, a situation known as *perfect dependence*. In this case, extremes of Z_1 occur in perfect accordance with extremes of Z_2. On the other hand, H may consist of two point masses, one at each end of the unit sphere $\mathcal{N}$. This situation, known as *asymptotic independence*, implies that extreme realizations of Z_1 exhibit no correspondence to extremes of Z_2. Thus in the two dimensional case, H governs the extent to which extreme

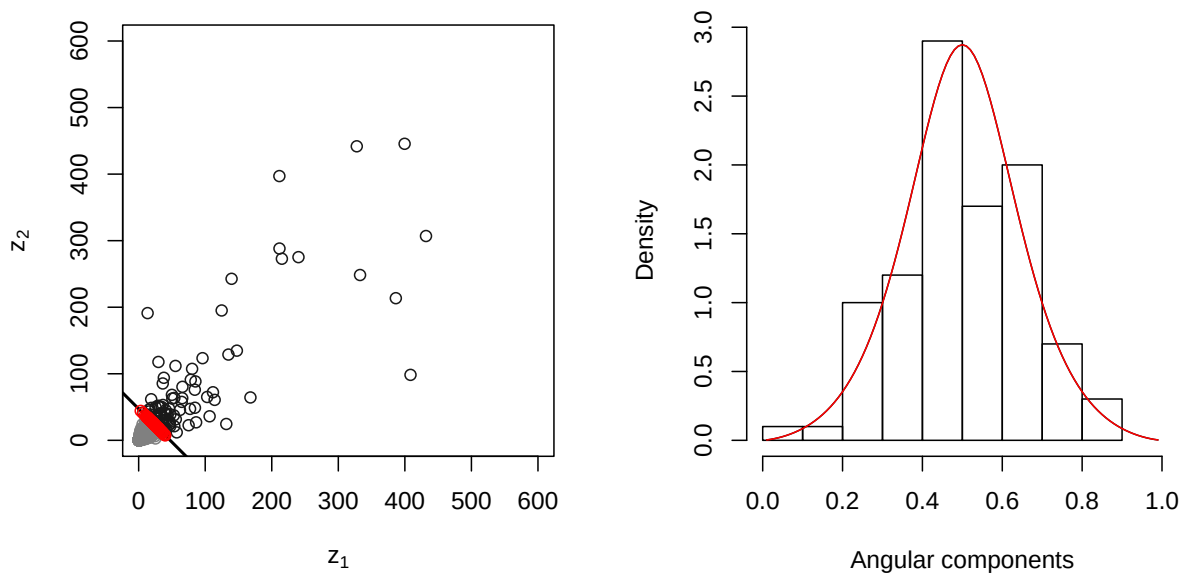

Figure 1.1: A realization of a bivariate regular varying random vector with tail index $\alpha = 1$ (left) and a histogram of angular components for points exceeding the chosen radial component threshold (right).

values of one margin correspond to extreme values of the other margin. We note here that asymptotic independence is not the same as independence in the usual sense; the asymptotic independence case is studied extensively in Chapter 4.

1.4.2 Relationship Between ν and H

An explicit link (Resnick, 2007, Proposition 6.4) exists between the measure ν, which measures sets in Cartesian coordinates, and H, which measures sets on the unit sphere $\mathcal{N}$. Assume $\alpha = 1$, the norm $\|\cdot\| = \|\cdot\|_1$, and consider the set $A = [\mathbf{0}, \mathbf{z}]^c$ for $\mathbf{z} = (z_1, ..., z_d)^T \in \mathfrak{C}$. It follows that

$$\begin{aligned}\nu(A) &= \int_A r^{-2} dr H(d\mathbf{w}) \\ &= \int_{\mathcal{N}} \int_{r=\wedge_{j=1}^d (z_j/w_j)}^{\infty} r^{-2} dr H(d\mathbf{w})\end{aligned}$$

$$= \int_{\mathcal{N}} \vee_{j=1}^{d} (w_j / z_j) H(d\mathbf{w}).$$

Given an angular measure H, one can use the above integration to attain ν. Conversely, Coles and Tawn (1991) provide a method for attaining the angular measure H induced by a given form of ν.

1.4.3 Link to Classical Results

The regular variation framework can be linked to the classical MVEVDs of Section 1.3.2 in the following way (Resnick, 1987, Corollary 5.18). Suppose $\mathbf{Z}$ is multivariate regular varying on $\mathfrak{C}$ with limit measure ν as in Definition 1, and define $\mathbf{M}_n$ to be the vector of componentwise maxima of n independent realizations of $\mathbf{Z}$. Then there exist normalizing sequences of vectors $\mathbf{a}_n \geq \mathbf{0}$ and $\mathbf{b}_n \in \mathbb{R}^d$ such that (2) holds with $G(\mathbf{z}) = \exp\{-\nu([\mathbf{0}, \mathbf{z}]^c)\}$. Furthermore, the marginal distributions of G are GEV with $\xi = 1/\alpha > 0$. That is, the multivariate regular varying $\mathbf{Z}$ is in the domain of attraction of a MVEVD with Fréchet marginal distributions and dependence structure given by the measure ν.

1.5 Statistical Inference for Multivariate Extremes

1.5.1 Threshold Exceedance Modeling

Modeling approaches based on the classical theory of extremes are limited to the study of block maximum data. This approach can be wasteful of data, and in many applications, it is desirable to employ exceedances of a high threshold to estimate the tail of distribution. In the univariate case, a limiting parametric model for exceedances is given by the generalized Pareto distribution (GPD) (Balkema and De Haan, 1974; Pickands, 1975). Statistical modeling of univariate exceedances may also be performed using a related Poisson point process; see Davison and Smith (1990).

In the multivariate setting, the definition of a threshold exceedance is not as obvious as in the univariate case. One approach is to employ a multivariate GPD (Rootzen and Tajvidi, 2006), which leads to the study of data exceeding marginal thresholds. A multivariate threshold model implemented via a censored likelihood approach appears in Smith et al. (1997). The approach employed in this dissertation defines a multivariate threshold exceedance in terms of the norm of a random vector.

1.5.2 Regular Variation Framework in Practice

Definitions 1-3 of multivariate regular variation require that each of the marginal distributions of $\mathbf{Z}$ are univariate regular varying with common tail index α. Typically, it is further assumed that $\mathbf{Z}$ has common marginal distributions. In statistical practice, it is common to apply transformations so that each margin has a common distribution function with tail index $\alpha = 1$. While in theory no dependence structure information is lost by marginal transformations, there are statistical efficiency implications of the choice of transformation; see e.g. Einmahl and Van den Akker (2011). When $\alpha = 1$, a common choice for $\|\cdot\|$ is the L_1 norm, in which case H is a measure on the unit simplex $\mathcal{N} = \{\mathbf{w} \in \mathfrak{C} : w_1 + ... + w_d = 1\}$ (Coles and Tawn, 1991; Ballani and Schlather, 2011; Cooley et al., 2010). Throughout the remainder of this dissertation, we assume the L_1 norm for polar coordinate transformations.

After estimation of marginal effects, attention turns to estimation of the dependence structure, which is determined by the angular measure H. When the marginal distributions of $\mathbf{Z}$ are common, a balance condition is imposed on the measure (Resnick, 1987, Proposition 5.11). Specifically, H must satisfy

$$\int_{\mathcal{N}} w_1 H(d\mathbf{w}) = \int_{\mathcal{N}} w_j H(d\mathbf{w}), \quad j = 2, ..., d. \tag{4}$$

In fact, any probability measure on $\mathcal{N}$ satisfying (4) is a possible angular measure of a regular varying random vector with common marginal distributions (Resnick, 2007, Remark

6.3). Thus, as in the classical MVEVDs, there is no finite parameterization for the class of dependence structures. The work in this dissertation will focus on parametric forms for H; nonparametric estimators have been explored in some low-dimensional problems (Einmahl et al., 2001).

As the only requirement for H is that it is a probability measure on $\mathcal{N}$ satisfying the balance condition (4), in practice it is often advantageous to model H directly, and this is the focus of the work in this dissertation. Alternatively, one could focus on models for the measure ν (Fougères et al., 2009; Einmahl et al., 2012).

Finite-sample estimation of the angular measure assumes that the limit in Definition 2 is an equality for suitably large r; that is, replacing t by n and defining $b_n = b(n)$, one assumes

$$n\mathbb{P}\left[\frac{R}{b_n} > r, \mathbf{W} \in B\right] = cr^{-\alpha}H(B)$$

for $r > r_0$, where r_0 is a suitably high threshold. Further assuming a parametric, continuously differentiable form for H with density $h(\mathbf{w};\boldsymbol{\theta})$ and defining $r_i = \|\mathbf{z}_i\|$ and $\mathbf{w}_i = \|\mathbf{z}_i\|^{-1}\mathbf{z}_i$ allows one to write a likelihood function for the points $\mathbf{z}_1, ..., \mathbf{z}_{N_0}$ for which $r_i > r_0$:

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{z}_1, ..., \mathbf{z}_n) &= \exp\left(-\frac{r_0}{b_n}\right)\left\{\prod_{i=1}^{N_0}(\alpha c r_i^{-1-\alpha}/b_n)h(\mathbf{w}_i;\boldsymbol{\theta})\right\}\Big/ N_0! \\ &\propto \prod_{i=1}^{N_0} h(\mathbf{w}_i;\boldsymbol{\theta}). \end{aligned} \tag{5}$$

Numerical methods may then be used to maximize the Poisson point process likelihood (5) with respect to $\boldsymbol{\theta}$.

CHAPTER 2

AN INVESTIGATION OF THE PINEAPPLE EXPRESS PHENOMENON VIA BIVARIATE EXTREME VALUE THEORY

2.1 Introduction

In this chapter, we examine extreme winter precipitation on the west coast of the United States and southern Canada, with special interest in the phenomenon known informally as "pineapple express" (PE). A PE event is characterized by a narrow stream of moisture, referred to in the meteorological community as an atmospheric river, which extends from near Hawaii to the west coast of the United States or Canada. This phenomenon has been known to cause extreme localized rainfall and flooding along the coast, and can also bring heavy, wet snow to the Sierra Nevada and and Cascade mountain ranges. Because of the subtropical origin of their moisture, PE events are often associated with warm temperatures, which can exacerbate winter flooding. Additionally, precipitation events associated with the PE phenomenon contribute significantly to the water resources of California and other western states (Dettinger et al., 2011).

There is much current interest both in characterizing PE events and in understanding if and how climate models capture and represent such events. In a recent study, Dettinger (2011) analyzed occurrences of atmospheric rivers in past climate reconstructions and an ensemble of simulations of future climate change. Leung and Qian (2009) found that the Weather Research and Forecasting (WRF) regional climate model (RCM) driven by reanalysis was able to reproduce the mean and 95th percentile of precipitation over the western United States, and in addition, accurately captured the spatial extent and intensity of two

major PE events. Here we investigate extreme winter Pacific precipitation and the PE relying on tools developed from statistical extreme value theory.

Much of the recent discussion of climate change has focused on the changes in extreme weather events. There are two primary reasons for this: first, these extreme weather events have the largest financial, environmental, and societal impacts; and second, some recent work (e.g., Holland, 2009) suggests that the bellwether of climate change will be an increase in the frequency and intensity of such events. This leads to several important and very broad questions: 1) How well are climate models able to reproduce extreme temperatures, precipitation, and storms? 2) What are the connections between these extreme events and the atmospheric processes that drive them? and 3) How will climatic change affect the frequency and intensity of extreme events in the future? Our investigation of winter Pacific precipitation and PE events attempts to address these questions with respect to particular atmospheric phenomena.

In terms of aims and statistical methodologies, this investigation differs considerably from previous work describing weather and climate extremes. Here, we perform several bivariate extreme value analyses. Examples of multivariate extremes analyses are few in the climate/atmospheric literature and relatively rare in general. Most multivariate extremes analyses aim to assess the probability of a rare event due to combined effects of several variables. One example is de Haan and de Ronde (1998), who assess flood risk as a result of the combination of wave height and storm surge. Although the statistical tools we use are similar, our aim is quite different. We aim to first describe and model the tail dependence of precipitation amounts from various sources, and second, to link large scale drivers to extreme precipitation. Below we first give a short review of extremes work related to weather and climate studies and then lay out the aims of this chapter in more detail.

2.1.1 Extreme Value Theory in Climate Studies

Because extreme weather events are costly in both human and economic terms, there has long been a need to describe their potential magnitude. Extreme value theory, the branch of probability specifically focused on describing the tail of a distribution, is well-suited for this task. A very nice review and commentary on applications of extreme value theory to temperatures, precipitation, and wind is given by Katz (2010).

Much extremes work aims to describe the tail of a distribution of a single quantity. Such univariate analyses have an extensive history in hydrology, engineering, finance, and climate research. In climate applications, numerous studies have endeavored to describe the tail of temperature or precipitation at a single weather station or climate model grid box. Other analyses in have used such descriptive extremes to identify climate change signals; see Guttorp and Xu (2011) for one recent example.

Although the aim of describing the tail of a distribution is simple, the task is not necessarily easy. Because an extreme value analysis uses a subset of data deemed extreme and extreme data are by definition rare, one is always data limited when describing extremes. This leads to large uncertainty in parameter and high quantile (or return-level) estimates. When one has data from multiple locations, there are methods of borrowing strength across locations to better describe distributions' tails. Such work dates back at least to the flood frequency analysis of Dalrymple (1960), and this methodology was used to produce the official precipitation return-level estimates[1] published by the National Oceanic and Atmospheric Administration (NOAA). The modern practice of regional frequency analysis is well summarized in Hosking and Wallis (1997). Another approach is to fit locations independently and then to spatially smooth fields created from the parameter estimates or of quantities of interest such as return levels (Kharin and Zwiers, 2000). Hierarchical modeling (e.g., Cooley et al., 2007; Sang and Gelfand, 2010) is a recent alternative approach for borrowing strength across locations.

[1] `http://www.nws.noaa.gov/oh/hdsc/currentpf.htm`

A regression approach (Beirlant et al., 2004, Chapter 7) has been used to model extremes of a non-stationary process by letting the parameters of the generalized extreme value (GEV) or generalized Pareto (GPD) distributions to be functions of covariates. To describe how extreme phenomena could be altered due to climate change, researchers (e.g., Kharin and Zwiers, 2005) have allowed the GEV or GPD parameters to be functions of time. Others have modeled an extreme value distribution's parameters as functions of another measurable and generally larger-scale climate variable; for example, Sillman et al. (2011) links the parameters of the GEV distribution describing European monthly minimum temperatures to an atmospheric blocking covariate. Models typically define the GEV or GPD parameters to be simple parametric functions (e.g., linear trends) of time, but more flexible non-parametric approaches have been used (Chavez-Demoulin and Davison, 2005). This regression approach is best suited for linking extreme phenomena to slowly varying covariates (e.g., the El Ñino/Southern Oscillation index (ENSO), North Atlantic Oscillation index (NAO), or other indices), but we believe that linking the parameters of an extreme value distribution to indices on a *daily* scale is generally a poor approach. Fitting either a GEV or GPD implies one has a large number of identically distributed observations from which one can extract either block maxima or threshold exceedances. Given an index that varies slowly (say monthly), one can assume the observations during this month all arise from the same distribution which is a function of that index and hence the parameters of the GEV or GPD can be viewed as functions of the index. However, with an index that varies rapidly (say daily), it is unclear how one could view an observation associated with a specific covariate value as extreme, when generally one has at most a few observations for the value of that covariate. Because we wish to link extremes to a daily index, we choose to employ bivariate extreme value theory rather than the regression approach.

A possible alternative to the bivariate extremes approach we describe in Section 2.2 is to employ a copula model. Copulas (Nelsen, 2006) provide an approach for modeling multivariate data and capturing dependence; such an approach is popular in hydrology.

Aside from the extremal copulas (Joe, 1997), copula models in general are not robust enough to capture tail dependence, and most are not regular varying. Thus, here we choose to model dependence based on the framework of multivariate regular variation.

2.1.2 Bivariate Extremes Investigation of Pacific Winter Precipitation

In this work we have several ambitious goals, each novel to climate extremes studies. Our first aim is to examine the tail dependence between observations and corresponding RCM output. Significant tail dependence would indicate the climate model is getting the extreme precipitation 'right', at least in the sense that the model's extremes correspond with extreme observations, even though the actual precipitation amount produced by the climate model may not be directly interpretable in terms of observations. Finding tail dependence, we then construct a statistical model which relates the extremes of the climate model output to the extreme observations. This model can be viewed as a type of 'statistical generator' of extreme observational precipitation from RCM output.

A second goal of this work is to draw a connection between extreme precipitation events and large-scale atmospheric processes. The connection between the large-scale processes and (often) localized extreme events in climate are, in many cases, not well understood and only beginning to be explored. Much of the work in this area has focused on extreme temperatures (Schubert and Henderson-Sellers, 1997; Sillman et al., 2011), though some studies have focused on extreme precipitation (Mo et al., 1997; Katz, 1999). Extreme precipitation events are short-lived, and our investigation is unique in that we link extreme precipitation to a daily index, rather than to large scale processes whose variation is more meaningful on a monthly or longer time scale. Specifically, we aim to build a daily "PE index" that exhibits extremal dependence with precipitation produced by this phenomenon. By understanding the processes which can lead to a PE event, we can quantify the extent of similarity of process dynamics between future climate model precipitation events, and events known to

produce PE-driven extreme precipitation. We build this index from daily north Pacific mean sea-level pressure (SLP) fields.

Finally, we apply our statistical precipitation generator to a future run of the WRF RCM driven by a general circulation model (GCM). GCM-driven RCM output does not exhibit a daily correspondence to observations, and also exhibits differing tail behavior in terms of precipitation. These obstacles, coupled with the fact that we do not have future observations, creates new challenges for constructing this statistical generator. We examine the extremes of precipitation in our study region from the future climate model run, and use the fitted extremal dependence model to simulate future observations of extreme west coast precipitation.

A brief review of extreme value theory and methods is given in Section 2.2. We discuss the various data products and climate model output sources we utilize in Section 2.3. In Section 2.4, we fit an extremal dependence model to the daily precipitation measurements from the WRF model and observational precipitation data. Section 2.5 details the motivation for the PE index and the method used for building it. The methodology for simulating future extreme precipitation observations is discussed in detail in Section 2.6. We conclude with a summary of results and discussion in Section 2.7. Much of the material in this chapter also appears in Weller et al. (2012).

2.2 Extreme Value Theory Background

In this chapter, we employ the regular variation framework introduced in Chapter 1 to conduct bivariate extreme value analyses. The aim is to characterize the extremal dependence structure between two quantities. The dependence structure can be completely characterized by the angular or spectral measure or equivalently, the Pickands' dependence function (Fougères, 2004). We first provide background on univariate techniques and tail dependence measures which are employed in this chapter.

2.2.1 Univariate Threshold Exceedances

An extreme value analysis does not require knowledge of the entire underlying distribution, and it is standard practice to utilize only those data which are considered to be extreme. The classical theory of extremes developed from the study of block (e.g., annual) maximum data. This can be wasteful of data, however, and an alternative approach is used in this work: we will study and model exceedances of a threshold. According to well-developed probability theory, if a random variable X is in the domain of attraction of an extreme value distribution, and given that the random variable X exceeds a suitably high threshold u, then the exceedance amount should approximately follow a GPD. That is, for $x > u$ and x such that $(1 + \xi(x - u)/\psi_u) > 0$,

$$\mathbb{P}(X > x \mid X > u) \approx \left(1 + \xi \frac{x - u}{\psi_u}\right)^{-1/\xi},$$

where $\psi_u > 0$ depends on the threshold and ξ does not. The parameter ξ characterizes the type of tail behavior: a bounded tail corresponds with $\xi < 0$, a light (exponentially decreasing) tail corresponds with $\xi = 0$, and a heavy (decreasing as a power function) tail corresponds with $\xi > 0$.

2.2.2 Bivariate Extremes and Tail Dependence

For random vectors, dependence in the joint upper tail is characterized differently than dependence in the whole of the joint distribution. While covariances and correlations are useful for describing dependence in many applications, these measures do not capture tail dependence. To describe dependence in the tail of the joint distribution of two random variables, one must first determine whether the pair are asymptotically dependent or asymptotically independent. A pair of random variables Z_1 and Z_2 with a common marginal distribution

are said to be *asymptotically independent* if

$$\lim_{z \to z_+} \mathbb{P}(Z_2 > z \mid Z_1 > z) = 0, \tag{6}$$

where z_+ is the (possibly infinite) right endpoint of the support of the common marginal. The random variables Z_1 and Z_2 are said to be *asmyptotically dependent* if the limit in (6) is greater than 0. It is possible for the dependence in the tail to be quite different than the dependence in the center of the joint distribution. For example, the two components of a bivariate Gaussian distribution with any correlation less than one can be shown to be asymptotically independent. Conversely, it is possible to construct a bivariate distribution with small correlation, but with strong tail dependence. An intuitive description of asymptotic dependence is that the largest values of Z_2 can occur concurrently with the largest values of Z_1.

2.2.3 Measures of Tail Dependence

There has been much work developing measures of dependence in both the asymptotic dependence and independence cases. In the asymptotically dependent case, there are several related dependence metrics (Davis and Resnick, 1993; Schlather and Tawn, 2003; Cooley et al., 2006). In Section 2.4, we use the measure $\chi(u)$ of Coles et al. (1999), where if Z_1 and Z_2 have a common marginal,

$$\chi(u) = \mathbb{P}(Z_2 > u \mid Z_1 > u).$$

Many recent works have aimed to quantify dependence in the asymptotically independent case or to formulate hypothesis tests for determining the category of tail dependence (Coles et al., 1999; Peng, 1999; Draisma et al., 2004; Zhang, 2008; Davis and Mikosch, 2009; Husler and Li, 2009). Ledford and Tawn (1996) provide an early formulation of the different forms of tail dependence of two random variables. They define the joint tail of two Fréchet-distributed

random variables Z_1 and Z_2 to decay as

$$\mathbb{P}(Z_1 > r, Z_2 > r) \sim \mathcal{L}(r) r^{-1/\eta}$$

as $r \to \infty$, where $\eta \in [1/2, 1]$ is a constant, and $\mathcal{L}(r)$ a slowly varying function. The parameter η serves as a measure of the amount of dependence in the asymptotically independent case. The limiting cases are asymptotic dependence ($\eta = 1$) and independence of Z_1 and Z_2 (in the usual sense), which would imply $\eta = 1/2$. Despite all the recent effort, it remains difficult to distinguish between the case of asymptotic dependence at a weak level and the case of relatively strong dependence in the asymptotic independence setting.

2.3 Precipitation Observations and Model Output

Climate models are deterministic tools which simulate long-term interactions between the atmosphere, oceans, and land. At their core, these models are a series of discretized differential equations which model the circulation of the atmosphere and ocean according to the known physics of the Earth system. They produce simulated weather phenomena on a spatial grid and record outputs on a timescale of a few hours. The outputs of these models are numerous; at each grid location and timestep, values of variables such as temperature, precipitation, winds, and pressure are given. These models can be run under current climate conditions or under various future scenarios, for which they provide insight into potential climate changes.

2.3.1 GCMs, RCMs, and Reanalysis Products

General circulation models (GCMs), often referred to as global climate models, simulate the climate over the entire Earth and generally have spatial resolutions of roughly 2.5 degrees latitude/longitude. Analysis of GCM output allows one to study climate on a large scale. Researchers employ GCMs to make statements about global or continental changes

in temperature, precipitation or other quantities. However, as their resolution is too coarse, GCMs are poor tools for making statements about local weather phenomena.

Regional climate models (RCMs) are one approach atmospheric scientists use in order to simulate higher-resolution phenomena and study their impacts. In comparison with GCMs, RCMs simulate weather features at finer spatial scales, on the order of tens of kilometers. Because of the computational cost of their high resolution and extensive parameterizations, these models are run over a subregion of the globe (e.g., over North America). An RCM requires boundary conditions to drive its climate simulations, since weather events at regional scales are driven by synoptic-scale conditions.

One method to drive an RCM is to have the boundary conditions provided by a GCM. RCMs driven by global models allow for examination of regional climate under various (current or future) conditions. They provide dynamical downscaling of large-scale phenomena from a GCM to a more local scale. However, GCMs are not meant to simulate weather on any specific day. Output of an RCM driven by a GCM provides a *distribution* of weather variables over a long period of time, but no correspondence exists between this output and observed weather on a given day.

Alternatively, an RCM can be driven by a reanalysis. A reanalysis product is similar to climate model output, but unlike GCMs, it has observed weather as inputs. Output from RCMs driven by reanalysis has temporal correspondence to past observed weather. For a particular day, we expect weather simulated by a reanalysis-driven RCM to be similar to weather observed on that day. Hence, one can compare RCM output to observations at short (e.g., daily) timescales. However, the RCM output does not exactly replicate observed weather due to variability in the climate system and errors in the reanalysis and regional models. As they require observational data as inputs, reanalysis-driven models do not allow for the study of future climate change scenarios.

2.3.2 WRF Regional Model and NARCCAP

In this work we primarily use output from the Weather Research and Forecasting (WRF) RCM. The WRF model is one of six RCMs included in the North American Regional Climate Change Assessment Program (NARCCAP). NARCCAP is studying regional climate change over North America by driving the six RCMs with four different GCMs, as well as a reanalysis product (Mearns et al., 2009). The NARCCAP models are run at approximately 50 km resolution; more information on NARCCAP can be found at its website[2].

As part of NARCCAP, the WRF model was driven by the NOAA National Center for Environment Prediction (NCEP) Reanalysis-2 (Kanamitsu et al., 2002) for the period 1981-2000. In Section 2.4, we compare output of the WRF model driven by NCEP reanalysis to observed precipitation. The WRF model has also been coupled with two different GCMs to simulate a control period from 1971-2000 and a future period from 2041-2070. The driving GCMs themselves are run for future climate under the Intergovernmental Panel on Climate Change (IPCC) A2 emissions scenario (Nakicenovic et al., 2000). In Section 2.6, we study output of the WRF model forced by the Community Climate System (CCSM) GCM (Collins et al., 2006). The CCSM model is maintained by the National Center for Atmospheric Research (NCAR). The WRF model precipitation output used in this work was downloaded from the Earth System Grid[3] maintained by NCAR in Boulder, CO USA.

2.3.3 Observations

The 'observations' employed in this work are in fact a gridded meteorological data product prepared by the Surface Water Modeling Group at the University of Washington[4]. To construct this product, weather station measurements were interpolated to produce retrospective $1/8^{th}$ degree ($\sim$12 km) gridded values of precipitation over the United States and

[2] `http://www.narccap.ucar.edu/about/index.html`

[3] `http://www.earthsystemgrid.org/project/NARCCAP.html`

[4] `http://www.hydro.washington.edu/SurfaceWaterGroup/Data/gridded/index.html`

parts of southern Canada (Maurer et al., 2002). While the gridded product contains uncertainties due to gaps in weather station locations, its output is more easily compared to climate model output than weather station data, since stations are not uniformly located in space, and their measurements often have missing data. The observational product was accessed from the website[5] of Dr. Edward Maurer at Santa Clara University.

2.4 Investigating and Modeling Tail Dependence between RCM Output and Observations

We begin by examining tail dependence in daily precipitation between reanalysis-driven WRF output and observations, that is, we study the extent to which the largest precipitation events from WRF output correspond to the largest events in the observational record. By its construction, the reanalysis-driven RCM should see the same synoptic-scale conditions as the observational record. When these conditions result in an observed extreme precipitation event, we aim to learn whether the WRF model will produce an extreme event as well. We restrict our attention to precipitation from November plus the winter months (NDJF), as the vast majority of PE events occur within this four-month time frame. Daily precipitation is examined for the years 1981-1999, resulting in a sample of $T_c = 2284$ days.

2.4.1 Study Region and Quantity of Interest

We define our study region to be roughly between 32°N and 53°N latitude, and from 118°W longitude to the west coast. Figure 2.1 shows the spatial extent of the study region from WRF output and observations. With this region we aim to capture extreme winter precipitation events that occur along the coasts of California and the Pacific Northwest, as well as extreme precipitation events in the Sierra Nevada and Cascade mountain ranges.

[5] `http://www.engr.scu.edu/~emaurer/data.shtml`

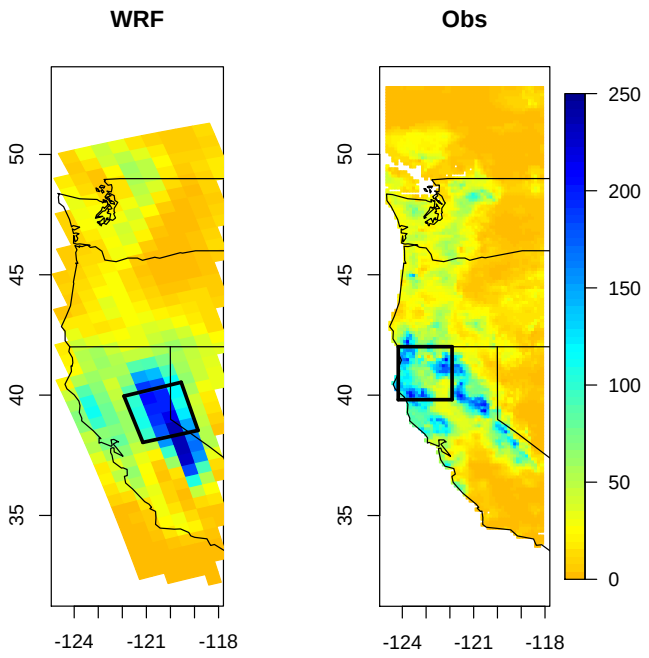

Figure 2.1: Precipitation intensity from WRF model output (left) and observations (right) on January 1, 1997. Footprint captured for this analysis is indicated by black lines (see Section 2.4.1 for details). Both figures are on the scale (mm) indicated by the legend strip.

An exploratory analysis was performed with the purpose of defining a daily precipitation measurement that captures the spatial extent and intensity of PE events. The aim was to find a daily quantity of which the largest realizations corresponded to PE events, as identified by Dettinger (2004). This quantity was chosen to be the maximum total precipitation over an approximately 200×200 km^2 area (4×4 WRF grid cells) within the region on each day. That is, for each day in the study, we find the 4×4 grid cell 'footprint' that has the greatest total precipitation intensity. Let X_t^{NC} be the sum of the daily grid cell precipitation intensities over the maximum footprint from the NCEP reanalysis-driven WRF output for day t ($t = 1, ..., T_c$). We define Y_t^C to be the sum of gridded observations for the maximum 200km $\times$ 200km footprint for day t. We use 'C' in the subscript and superscripts here to denote that we study the 'current' climate. An example of this footprint for both WRF output and observations is shown in Figure 2.1. We extract realizations of (X_t^{NC}, Y_t^C) for each day in the study, without requiring that the two footprints must align spatially on a given day. The primary reason for not requiring spatial alignment is to account for the possibility that the WRF model may be reproducing an extreme precipitation event, but not in the exact location seen in observations. In addition, coordinate systems and spatial resolutions differ between WRF output and the gridded observational product, making exact spatial matching difficult. No systematic bias was found in the latitudes or longitudes of precipitation events simulated by WRF; boxplots of the discrepancies in latitude and longitude between locations of extremes in the two products are provided in Figures 2.2 and 2.3.

A scatterplot of the resulting daily quantities is shown in the left panel of Figure 2.4. As expected, we see strong dependence between the NCEP-driven WRF output and observations in the whole of the data, and the sample correlation is estimated to be 0.74. Our focus here is the joint upper tail, however, and we also see that the largest values of X_t^{NC} appear to have correspondence with the largest values of Y_t^C.

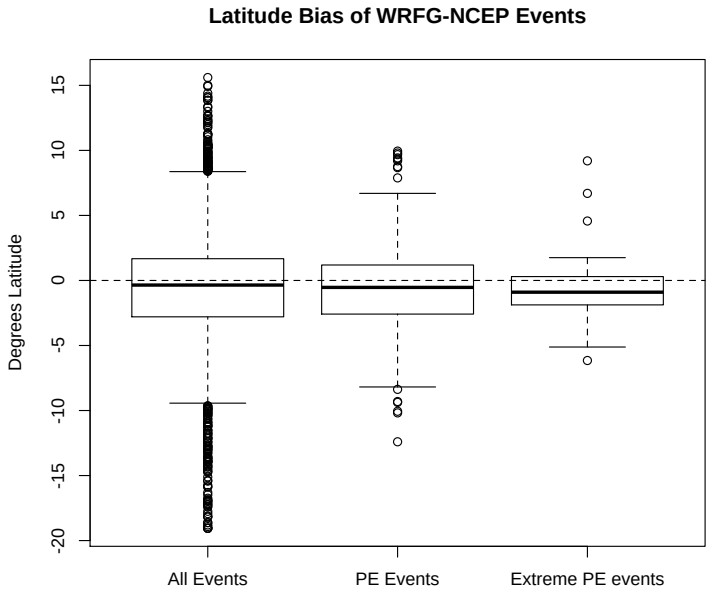

Figure 2.2: Differences in latitudes of center of region from which precipitation quantities X_t^{NC} (WRFG output) and Y_t (observations) were drawn, $t = 1, ..., 2284$. Positive values indicate the RCM simulated an event at higher latitude than seen in observations. PE events are as defined in Section 2.4.4; extreme PE events are largest 130 values of Y_t which correspond to PE days.

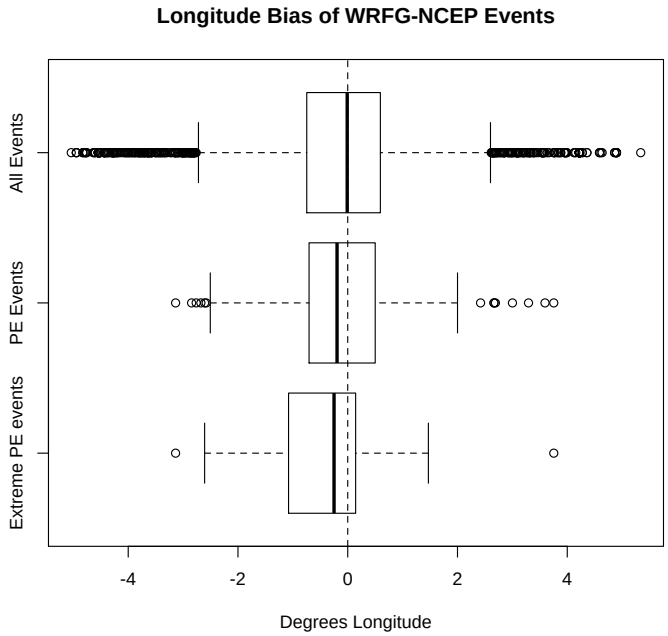

Figure 2.3: Differences in longitudes of center of region from which precipitation quantities X_t^{NC} (WRFG output) and Y_t (observations) were drawn, $t = 1, ..., 2284$. Positive values indicate the RCM simulated an event more eastwardly than seen in observations. PE events are as defined in Section 2.4.4; extreme PE events are largest 130 values of Y_t which correspond to PE days.

Table 2.1: Threshold selected and maximum likelihood GPD parameter estimates with standard errors for precipitation quantities. Marginal thresholds have common exceedance rate 0.057 (130 exceedances).

Margin	$u_\cdot$	$\hat{\psi}_\cdot$ (se)	$\hat{\xi}_\cdot$ (se)
X_t^{NC} (WRF)	1054	288.95 (39.27)	0.0255 (0.104)
Y_t^C (obs)	14240	3895.87 (512.03)	0.0213 (0.099)

2.4.2 Extremal Dependence Estimation

As a preliminary step in examining the upper tail dependence in (X_t^{NC}, Y_t^C), the tails of each marginal distribution are estimated separately. After checking diagnostics to ascertain an appropriate threshold (Coles, 2001, Chapter 4), GPDs are fitted to the largest 130 observations of X_t^{NC} and Y_t^C (corresponding to exceedances of the empirical 0.943 quantile). A summary of maximum likelihood estimates of GPD parameters is given in Table 2.1. It is encouraging and perhaps surprising that estimates of the tail parameter ξ are similar for the RCM output and the observations. It has been hypothesized that climate models could not produce heavy-tailed precipitation. Weller et al. (2013) found that tail behaviors in this precipitation quantity differ between NARCCAP models.

Probability integral transformations $Z_t^{NC} = T^{NC}(X_t^{NC})$ and $Z_t^C = T^C(Y_t^C)$ are applied to each margin using the fitted GPD above the chosen marginal thresholds and the empirical distribution functions below. The transformation results in unit Fréchet margins; that is, Z_t^{NC} and Z_t^C have distribution function $F(z) = \exp\{-z^{-1}\}$. Scatterplots of realizations (x_t^{NC}, y_t^C) and (z_t^{NC}, z_t^C) are shown in Figure 2.4. While it is seen in the left panel that these data exhibit high correlation, the Fréchet scale plot indicates that tail dependence appears to be present as well, as many of the largest points lie on the interior of the first quadrant and not near the axes. Note that the scales for X_t^{NC} and Y_t^C in the left plot differ considerably, showing that the precipitation amounts produced by the RCM are not directly interpretable in terms of observations.

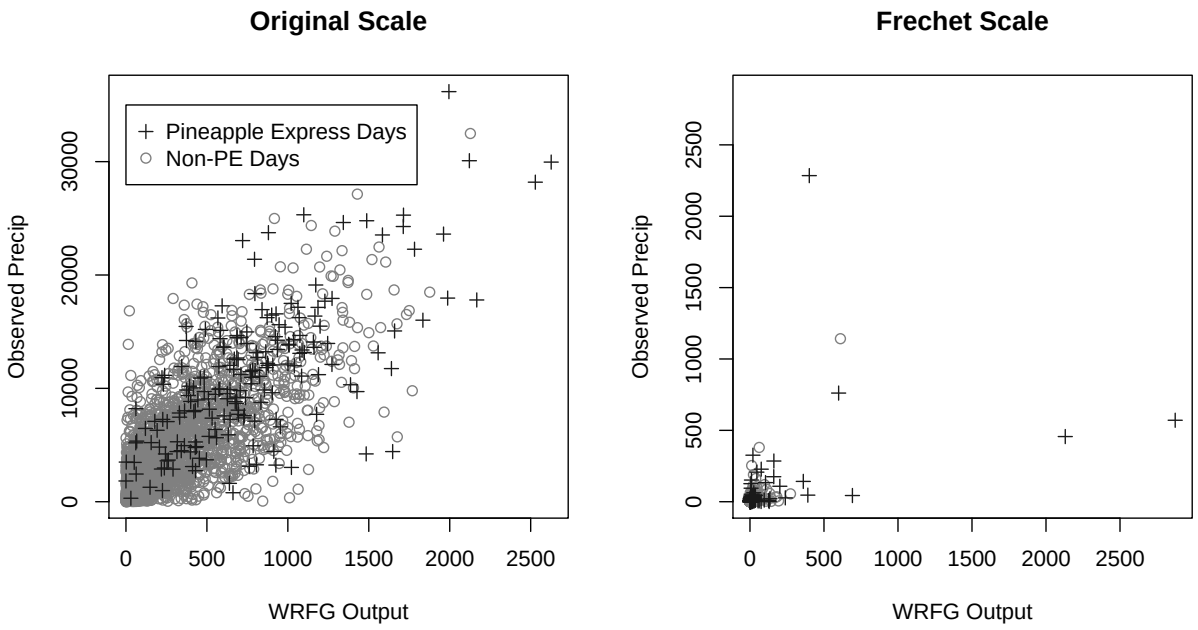

Figure 2.4: Daily precipitation quantity computed from WRFG output driven by NCEP reanalysis and gridded observational data product on original scale (left) and after transformation to Fréchet scale (right). PE events are as identified in Dettinger et al. (2011). See also left panel of Figure 2.5.

To confirm that the WRF output and observations exhibit tail dependence, we use the hypothesis test of Ledford and Tawn (1996). Ledford and Tawn (1996) use an estimate of η in a likelihood ratio test for H_0: $\eta = 1$ against H_1: $\eta < 1$, i.e. asymptotic dependence against asymptotic independence. The resulting p-value is nearly 1, providing little evidence of asymptotic independence of the bivariate pair (Z_t^{NC}, Z_t^C). Furthermore, for q_{95} the 0.95 empirical quantile in each margin, we estimate $\hat{\chi}(q_{95}) = 0.47$, indicating a moderate-to-strong level of tail dependence. That the reanalysis-driven WRF output exhibits this amount of tail dependence means, in a certain sense, that the RCM is reproducing winter extreme precipitation reasonably well. That is, when boundary conditions from the reanalysis are such that the largest precipitation occurs, those conditions are adequately captured by the reanalysis product and the RCM can respond in the correct way–by producing its largest precipitation. This is strong statement about the way the RCM is simulating extreme precipitation events. Not only does the WRF model accurately represent high quantiles, as found by Leung and Qian (2009); it is also producing *correspondence* of its largest precipitation days with those in the observed record.

2.4.3 Models for the Angular Measure

Having found the tails to be asymptotically dependent, we turn our attention to constructing a statistical model relating the tails of the RCM output to the observations. We use an established procedure (Coles and Tawn, 1991; Cooley et al., 2010; Ballani and Schlather, 2011) to fit an angular measure model to (z_t^{NC}, z_t^C). Assuming a parametric model for an angular density, one can write down a corresponding point process likelihood approximation to threshold exceedances given by (5) in Chapter 1, and then estimate the angular density's parameters via numerical maximum likelihood. A transformation to pseudo-polar coordinates is made, as described in Section 2.2. Exceedances are defined in terms of $r_t = z_t^{NC} + z_t^C$, and for these data, we select the threshold $r_0 = 40$, the approximate 0.95 empirical quantile of r_t values. We fit an angular measure corresponding to the logistic model (Gumbel, 1960),

a one-parameter model whose angular density is given by

$$h(w;\alpha) = \frac{1}{2}(\alpha^{-1} - 1)\{w(1-w)\}^{-1-1/\alpha}\{w^{-1/\alpha} + (1-w)^{-1/\alpha}\}^{\alpha-2}.$$

We also fit the Dirichlet angular density (Coles and Tawn, 1991), a more flexible two-parameter model with

$$h(w;\alpha,\beta) = \frac{\alpha\beta\Gamma(\alpha+\beta+1)(\alpha w)^{\alpha-1}\{\beta(1-w)\}^{\beta-1}}{2\Gamma(\alpha)\Gamma(\beta)\{\alpha w + \beta(1-w)\}^{\alpha+\beta+1}}$$

Figure 2.5 provides a closer inspection of the Fréchet-scaled data shown in Figure 2.4, as well as a histogram of angular components $w_t = z_t^{NC}/r_t$ of points for which $r_t > r_0$. For the logistic model, the maximum likelihood procedure yields $\hat{\alpha} = 0.577$, with a standard error of 0.023. The Dirichlet model parameters are estimated as $\hat{\alpha} = 1.203$ and $\hat{\beta} = 0.862$, with standard errors of 0.318 and 0.189, respectively. This suggests slight asymmetry in the dependence structure, which is indicated by the histogram in Figure 2.5. The fitted logistic model and fitted Dirichlet models had AIC statistics of 2162.2 and 2156.1, respectively, suggesting the Dirichlet model as preferable to the logistic. Both dependence model fits are added to the histogram in Figure 2.5.

2.4.4 Pineapple Express Events

As the primary interest in this work is the PE phenomenon, it is of interest to study those (x_t^{NC}, y_t^C) that are associated with PE events. While the narrow band of atmospheric moisture associated with PE events poses difficulty in detecting them, some recent work has been done in this area. Dettinger et al. (2011) applies pattern recognition techniques to NCEP reanalysis fields to classify days that exhibit a PE regime. This results in a list of days which are labeled as PE events. The list first appeared in Dettinger (2004); it was updated in Dettinger et al. (2011). In this work, we include these days as well as subsequent days to be labeled as PE precipitation events. This results in a subset of 186 PE days.

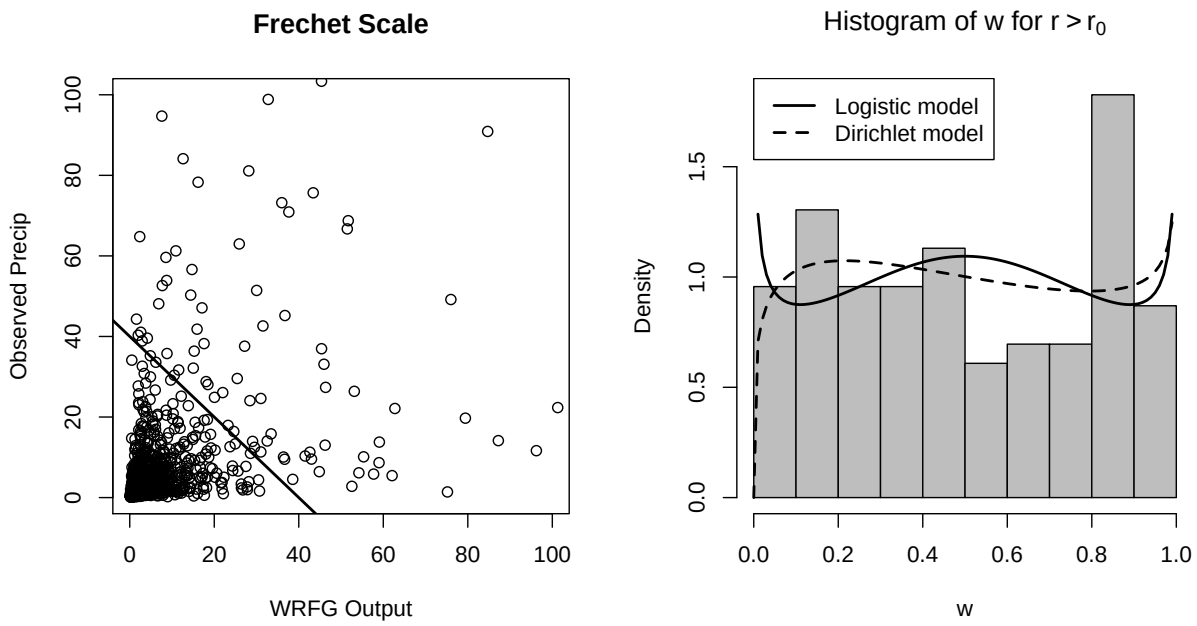

Figure 2.5: Left: WRFG output z_t^{NC} against observations z_t^C on Fréchet scale, with the line $r = r_0$ shown. Right: Distribution of angular component values $w_t = z_t^{NC}/r_t$ for points such that $r_t > r_0$. Estimated densities of $h(w)$ from one-parameter logistic and two-parameter Dirichlet models are added to the histogram. Note the Dirichlet model fit indicates an asymmetric dependence structure.

Figure 2.4 highlights PE days in the scatterplots of (x_t^{NC}, y_t^C) and (z_t^{NC}, z_t^C). The scatterplot indicates that not all PE events produce large precipitation measurements (as defined by our footprints), and not all large precipitation events are associated with the PE phenomenon. However, the tendency for the PE phenomenon to produce extreme precipitation is apparent. While only 8% of all the days in our study are identified as PE days, 49 of the 130 days (37.7%) with the largest observed precipitation measurements Y_t^C are identified as PE events.

The left panel of Figure 2.4 suggests that extreme precipitation events occur in both PE and non-PE conditions. Figure 2.6 shows the average precipitation intensity for the 130 largest observations of Y_t^C, separated into PE days and non-PE days. The figure shows very little difference between mean extreme precipitation patterns for PE and non-PE events. Thus an attempt to identify PE events cannot rely on precipitation patterns only. We turn to synoptic-scale patterns in Section 2.5 to better understand the PE phenomenon.

2.5 Pineapple Express Index

The PE phenomenon has been well-known to meteorologists for many years, and atmospheric rivers have been the subject of several recent publications (e.g., Zhu and Newell, 1994; Dettinger, 2004; Dettinger et al., 2011). The connection between large-scale atmospheric processes and extreme precipitation from PE is not well-known. Much of the work in PE has centered around identification of these events from satellite imagery or reanalysis fields. Many of the methods employed thus far have been used only for pattern recognition (Dettinger et al., 2011). These methods are effective at identifying PE events, but have not produced a direct link to the quantity of precipitation produced by PE. Identification methods also rely on analysis of a combination of several atmospheric fields. Other works have focused on the connections between indices such as the ENSO and Pacific Decadal Oscillation (PDO) modes (Dettinger et al., 2011; Dettinger, 2004). However, these indices

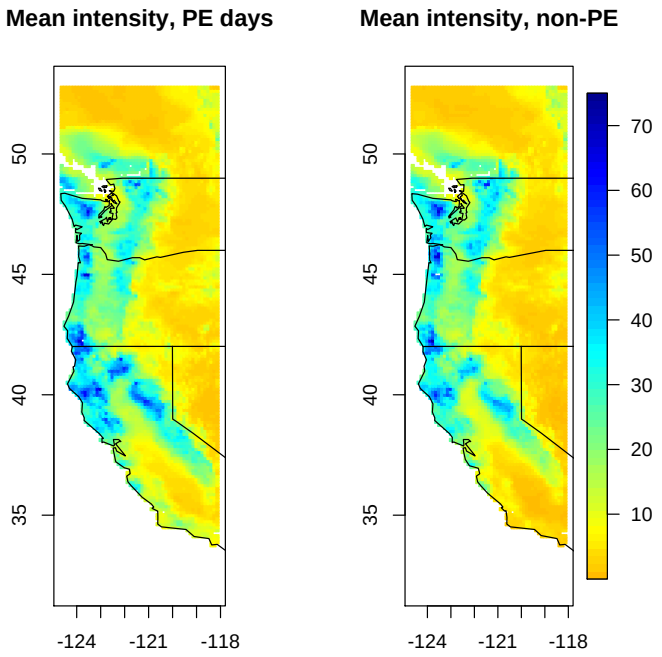

Figure 2.6: Mean observed precipitation intensity for 49 largest PE days (left) and largest 81 non-PE days (right). Figures are on the same scale (mm/day) indicated by the legend strip.

are derived on larger temporal scales; we aim for an index defined the same daily timescale as our precipitation measurements.

2.5.1 Connection to Sea-Level Pressure Fields

As a first step toward connecting synoptic-scale patterns to PE events, we turn our attention to SLP fields, as these are often employed in the atmospheric science literature as indicators for different weather regimes. The link between pressure fields and Pacific coast precipitation has been studied in previous work. Stahl et al. (2006) connected SLP patterns to precipitation and temperature anomalies at weather stations in British Columbia, Canada. Stahl et al. (2006) used principal component analysis on daily mean SLP fields in the north Pacific to classify 13 types of synoptic circulation patterns, and examined temperature and precipitation anomalies associated with each type. However, Stahl et al. (2006) did not

explicitly discuss connection of these circulation patterns to extreme precipitation or the PE phenomenon. While Cannon et al. (2002) examines connections between circulation patterns and surface weather conditions, the analysis is limited to one specific location.

We note here that the origin of the atmospheric river feature of PE events is outside the domain of the NARCCAP simulations. We thus extract mean SLP fields for each day in our study from the NCEP reanalysis product which drives the WRF precipitation output discussed in Section 3. Daily mean SLP fields from NCEP reanalysis were downloaded from the website of the NOAA Earth System Research Laboratory[6] in Boulder, CO USA. Daily average SLP fields have a spatial resolution of 2.5° latitude/longitude, and we examine the region of the north Pacific between 157.5° W and 130° W longitude, and 30° N and 62.5° N latitude (see Figure 2.7). This region is similar to that studied in Stahl et al. (2006). Winter SLP anomalies for each day are calculated by subtracting the NDJF mean daily pressure over the time period 1981-1999. This results in a 280-dimensional vector of daily SLP anomalies. We denote this vector for day t as $\mathbf{M}_t$.

As an exploratory step, we examine SLP anomaly fields of PE events and non-PE events. Figure 2.7 shows the mean anomalies over the domain for the 130 largest observed precipitation events, partitioned into PE and non-PE events. Despite the similar precipitation observations for PE and non-PE identified days shown in Figure 2.6, Figure 2.7 shows that the SLP pattern composite from extreme precipitation days identified as PE by Dettinger et al. (2011) differs dramatically from the composite of non-PE extreme events. It appears that there are at least two different SLP regimes producing extreme precipitation in our study region.

2.5.2 Construction of the PE Index

The PE SLP anomaly field in the left panel of Figure 2.7 is used to build a PE index. Let $\boldsymbol{\mu}_{PE}$ be the normalized vector of mean anomalies from the 49 PE extreme precipitation days.

[6] `http://www.esrl.noaa.gov/psd/`

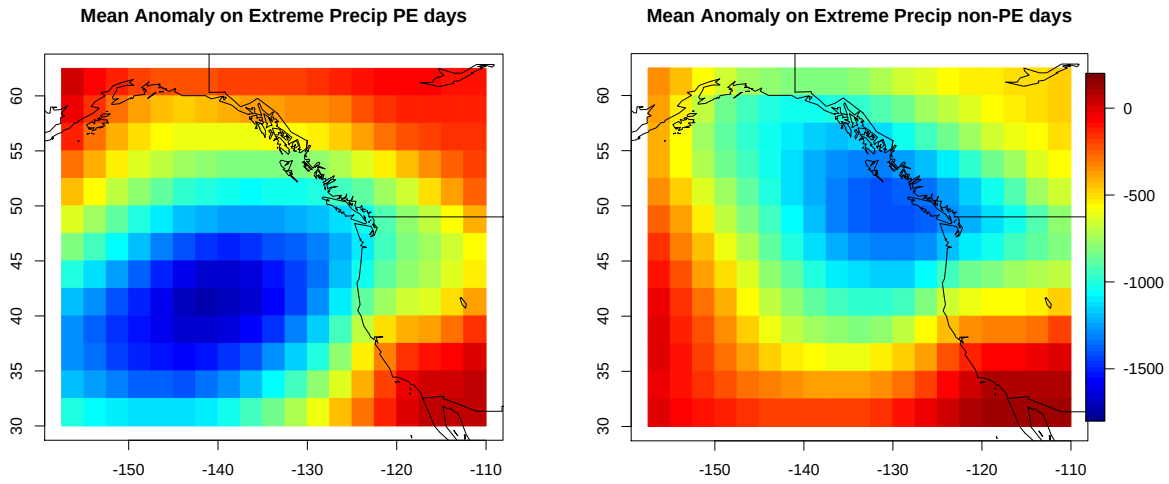

Figure 2.7: Mean SLP anomalies (Pa) for largest 49 PE precipitation days (left) and largest 81 non-PE precipitation days (right).

We define our PE precipitation index for day t as the projection of that day's SLP field onto the mean PE anomaly: $U_t^{PE} = \mathbf{M}_t \cdot \boldsymbol{\mu}_{PE}$. Thus U_t^{PE} is a function of both the "direction" and magnitude of $\mathbf{M}_t$. This index can be thought of as measuring the covariance between a given day's SLP anomaly pattern and the SLP anomaly shown in the left panel of Figure 2.7. The left panel of Figure 2.8 shows the PE index plotted against observed precipitation, for both PE and non-PE events. The index exhibits dependence with our precipitation footprint, and the dependence appears to be slightly stronger for PE events. The sample correlation between U_t^{PE} and $\log(y_t^C)$ is 0.31 for PE days and 0.28 for non-PE days. Additionally, several of the largest PE index values correspond with the largest precipitation values from PE events.

2.5.3 Tail Dependence with Observed Precipitation

In order to examine extremal dependence of U_t^{PE} with observed precipitation Y_t^C, a transformation is made to the Fréchet scale. Here, we make a simplifying assumption of a Gaussian distribution of daily winter SLP anomalies at each of the 280 locations, i.e. $\mathbf{M}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$. This assumption was found to be reasonable through exploratory analysis. It follows that

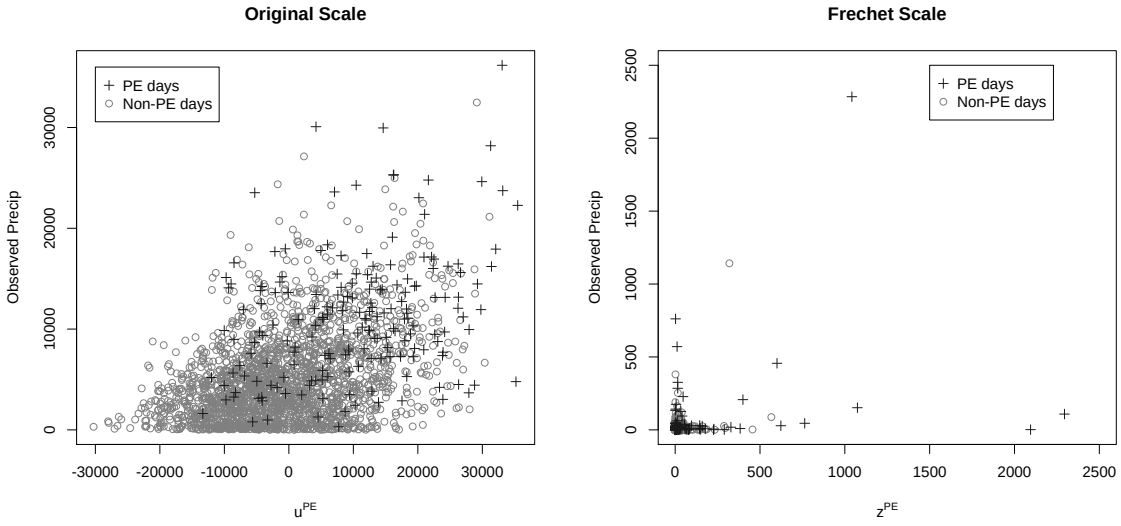

Figure 2.8: PE precipitation index u_t^{PE} plotted against observed precipitation y_t^F. Left: original scale. Right: Fréchet scale.

the index U_t^{PE} is univariate Gaussian. After estimating the variance of U_t^{PE}, a probability integral transform is applied so $Z_t^{PE} = T^{PE}(U_t^{PE})$ follows a unit Fréchet distribution. The right panel of Figure 2.8 shows the scatterplot of points (z_t^{PE}, z_t^C). Although the tail dependence is not as strong as that between Z_t^{NC} and Z_t^C, there are several large points in the interior of the first quadrant, indicating asymptotic dependence of the pair (Z_t^{PE}, Z_t^C).

We again employ the Ledford and Tawn (1996) likelihood ratio test for asymptotic independence against the null hypothesis of dependence. Only observations from days identified as PE events are included in the test. The p-value of the test was found to be approximately 0.8, giving little evidence of asymptotic independence of the PE precipitation index and PE precipitation quantities. Furthermore, for the subset of PE days, $\hat{\chi}(q_{95}) = 0.37$, suggesting moderate-to-weak dependence. Due to the small sample size, caution should be used in interpreting this quantity, as its confidence interval covers both 0 and 1. Although Figure 2.8 shows the tail dependence is not strong, the PE index does give some indication of extreme precipitation on PE days. The same likelihood ratio test found that the PE precipitation index was asymptotically independent of observed precipitation for non-PE days.

The approach to the PE index here is novel in that it examines tail dependence between observed precipitation and an index drawn from synoptic-scale processes. The index employed here is a first attempt and could likely be improved. While previous studies used multiple fields (water vapor, wind, temperatures, etc.) to identify PE events (Dettinger et al., 2011), this work draws upon only SLP fields to define U_t^{PE}. It is clear from Figure 2.8 that there are many days that exhibit a very high value of U_t^{PE}, but do not have extreme precipitation in terms of Y_t^C. Many days that are not identified by Dettinger et al. (2011) as PE events also have high values of U_t^{PE}. Nevertheless, our simple index is clearly correlated with our precipitation quantity, has dependence (both in terms of correlation and tail dependence) which increases for PE events, and exhibits asymptotic dependence with the precipitation observations. Asymptotic dependence is an important feature, as it implies that days with the largest values of the index have a nonzero probability of corresponding to the largest days in terms of observed precipitation. Perhaps most importantly, for a given day from a future run of a climate model, this simple index provides an indication of the similarity between its SLP pattern and a pattern known to be associated with PE-driven extreme precipitation.

2.6 Simulating 21st Century Extreme Precipitation Observations

Previous studies based on both observed data and climate models have suggested that extreme precipitation increases in a warmer climate (Allan and Soden, 2008; Easterling et al., 2000). Leung et al. (2011) found that the WRF model forced by CCSM indicated an increase in the 90^{th} and 95^{th} percentile of west coast precipitation, as well as a 27% increase in the frequency of events associated with atmospheric rivers. The study also found an increase in precipitation intensity over the region from current to future climate. Here we aim to generate future precipitation 'observations' from the precipitation output from the WRF RCM driven by CCSM under the future scenario.

We extract the precipitation quantity defined in Section 2.4 from WRF driven by CCSM for NDJF days in the future period, and denote it for day t ($t = 1, ..., T_f$) by X_t^{CF}. The record for December 2070 was incomplete and thus excluded, resulting in a sample of $T_f = 3569$ days. We denote Y_t^F as the (unobserved) precipitation quantity from observations on day t in the future time period.

Our generator is actually a two-step process. Recall from Section 2.4 that we linked NCEP-driven RCM output to observations. As there is no reanalysis product for future observations, we first use the larger NARCCAP experiment to forecast what the marginal distribution of reanalysis-driven WRF precipitation would look like under the future scenario. We denote this future unobserved NCEP reanalysis-driven WRF precipitation quantity by X_t^{NF}. We estimate the marginal distribution of X_t^{NF}, and then link this distribution to the marginal of CCSM-driven WRF precipitation X_t^{CF} under the future scenario. We then employ the fitted Dirichlet angular measure model from Section 2.4 to simulate Y_t^F, observations of future Pacific coast precipitation from X_t^{CF}, the precipitation quantity simulated by the WRF RCM driven by CCSM, for the future period 2041-2070. We thus allow the marginal distributions to change from current to future, but assume the dependence structure does not change; see Section 2.7 for discussion.

2.6.1 Marginal Distribution of Future Reanalysis-Driven WRF Output Extremes

To project the marginal distribution of WRF output driven by NCEP reanalysis for the period 2041-2070, we borrow information from other RCMs in the NARCCAP project. In addition to the WRF RCM driven by the CCSM global model, we employ output from four other RCMs driven by three other GCMs. Not all RCM-GCM combinations have been run, but those that have been completed are run for the same current and future periods, and under the same future A2 emissions scenario. All RCMs have been forced by the NCEP reanalysis and run for the current climate (1981-2000). See Table 2.2 for a summary of

Table 2.2: RCM-GCM combinations used in Section 2.6.2.

	GCM			
RCM	CCSM	CGCM3	GFDL	NCEP
WRFG	X	X		X
ECP2			X	X
CRCM	X	X		X
MM5I	X			X
RCM3		X	X	X

RCM	Full name	Modeling group
WRFG	Weather Research and Forecasting Model	Pacific Northwest National Lab
ECP2	Experimental Climate Prediction Center	UC San Diego / Scripps
CRCM	Canadian Regional Climate Model	OURANOS / UQAM
MM5I	MM5 - PSU/NCAR mesoscale model	Iowa State University
RCM3	Regional Climate Model version 3	UC Santa Cruz

GCM	Full name	Modeling Group
CCSM	Community Climate System Model	NCAR
CGCM3	Third Generation Coupled Global Climate Model	CCCMA
GFDL	Geophysical Fluid Dynamics Laboratory GCM	GFDL
NCEP	NCEP/DOE AMIP-II Reanalysis	NOAA ESRL

RCM-GCM combinations used for this estimation procedure. Though they are also included in NARCCAP, we exclude the Hadley Centre RCM and GCM, as output from the coupling of these models with other NARCCAP models was not yet available.

In Section 2.4 a two-step technique was used to estimate the distribution of the upper tails of the quantities X_t^{NC} and Y_t^C. First a threshold q^* was selected; we chose this to be the 0.943 empirical quantile. A generalized Pareto distribution was then fitted to exceedances of q^*. In order to characterize the upper tail of X_t^{NF}, estimates of q^* as well as the GPD parameters (ψ, ξ) must be obtained. We estimate q^* for X_t^{NF} independently of the estimation of (ψ, ξ), and both are estimated as predictions from linear models.

For each RCM-GCM run available, we compute the daily precipitation quantity X_t as defined in Section 2.4 for current and future runs, and find q^*, the empirical 0.943 quantile of X_t. A summary of these estimates is given in Table 2.3. Denote q^*_{ijr} as the 0.943 quantile

of X_t from RCM i ($i = 1, ..., 5$), coupled with GCM j ($j = 1, ..., 4$), for time period r ($r = 1$ for current, $r = 2$ for future). We write the model

$$q^*_{ijr} = \mu + \alpha_i + \beta_j + \gamma I_{\{r=2\}} + \epsilon_{ijr}, \quad \epsilon_{ijr} \sim N(0, \sigma^2) \tag{7}$$

and estimate the parameters via least squares. Of particular interest is the parameter γ, which estimates the average shift in q^* from current to future climate over the RCM-GCM combinations. We estimate $\hat{\gamma} = 25.0$, but this increase was not found to be statistically significant. This is likely due to the small sample size (21 total climate model runs) and to the fact that RCM runs forced by the GFDL global model exhibit a small decrease in q^* from current to future. A decrease in future scenario precipitation in the region from the GFDL model was also found by Cayan et al. (2008).

From this fitted model we obtain an estimate of q^* for X_t^{NF}, the 0.943 quantile of our precipitation quantity as would be produced by the NCEP-driven WRF model for the future period 2041-2070. This is estimated to be 1081.1, with a 95% confidence interval of $(1005.8, 1156.4)$ based on the least-squares estimation of σ^2. This is an increase of about 27 from the estimate of this quantile for the current run of WRF driven by NCEP given in Table 2.1.

In addition to extracting the quantity q^* for each RCM-GCM-time period combination, we estimate the parameters (ψ, ξ) of the GPD fit to exceedances of this threshold (see Table 2.3). For each run, we retain maximum likelihood estimates of the parameters and their numerically estimated covariance matrix. Defining ψ_{ijr} and ξ_{ijr} to be the parameters of the GPD fit to exceedances of q^* of X_t for RCM i, GCM j, and run r, we write

$$\begin{pmatrix} \psi_{ijr} \\ \xi_{ijr} \end{pmatrix} = \begin{pmatrix} \mu_\psi \\ \mu_\xi \end{pmatrix} + \begin{pmatrix} \alpha_{i\psi} \\ \alpha_{i\xi} \end{pmatrix} + \begin{pmatrix} \beta_{j\psi} \\ \beta_{j\xi} \end{pmatrix} + \begin{pmatrix} \gamma_\psi \\ \gamma_\xi \end{pmatrix} I_{\{r=2\}} + \epsilon_{ijr}, \quad \epsilon_{ijr} \sim N(\mathbf{0}, \Omega_{ijr}), \tag{8}$$

where Ω_{ijr} is the inverse of the numerically approximated information matrix of the estimate of $(\psi_{ijr}, \xi_{ijr})^T$. We estimate this model by generalized least squares.

The model parameters γ_ψ and γ_ξ in (8) provide insight into the change in the GPD scale and shape parameters from current to future climate. Based on estimates from this model we find almost no change in the scale parameter ψ from current to future, as $\hat{\gamma}_\psi = 0.80$, with a standard error of 15.7. We estimate $\hat{\gamma}_\xi = 0.057$, providing evidence of an increase in the shape parameter from current to future climate. The GLS-based 95% confidence interval for this quantity is $(-0.014, 0.129)$, so the evidence is not overwhelming for a change in ξ. Again, this is likely due to the small sample size, and possibly the decrease in precipitation from GFDL-forced RCMs.

It is of note from estimation of the linear model (8) that $\beta_{4\xi}$, the parameter for ξ corresponding to NCEP reanalysis-driven RCM output, was estimated to be 0.150, with a 95% confidence interval of $(0.053, 0.247)$. That is, in terms of the quantity X_t, west coast precipitation as simulated by the NARCCAP RCMs driven by NCEP reanalysis exhibits a significantly heavier tail than the same quantity as simulated by these RCMs driven by the other GCMs. In particular, the reanalysis-driven WRF model produces a much heavier tail of precipitation than the CCSM-driven WRF model in the current climate (see Table 2.3). This is a primary reason for estimating the marginal distribution of X_t^{NF} via (7) and (8), rather than only estimating the marginal of X_t^{CF}. Due to the fundamental differences in tail behavior of our precipitation quantity between reanalysis-driven RCM output and GCM-driven RCM output, we are more comfortable making a transformation to future reanalysis-driven WRF output, rather than using CCSM-driven WRF output directly. The marginal distribution of reanalysis-driven WRF output also has a nice relationship with the distribution of observations, as we show in Section 2.6.2.

We use the fitted linear model (8) to estimate the parameters (ψ, ξ) of the GPD tail of X_t^{NF}, the unobserved reanalysis-driven WRF precipitation output. We estimate $\hat{\psi} = 302.3$, and associated 95% confidence interval of $(242.8, 355.8)$. The tail parameter $\hat{\xi} = 0.069$,

Table 2.3: Empirical 0.943 quantile q^*, and numerical maximum likelihood estimates of GPD parameters (standard errors) for precipitation footprint from each climate model combination used in Section 2.6.2.

RCM-GCM	Current (1981-1999)			Future (2041-2070)		
	q^*	$\hat{\psi}$	$\hat{\xi}$	q^*	$\hat{\psi}$	$\hat{\xi}$
CRCM-CCSM	792	233.7 (31.1)	−0.092 (0.10)	846	275.7 (23.9)	−0.177 (0.05)
CRCM-CGCM	781	187.4 (25.1)	0.018 (0.10)	804	210.4 (21.6)	−0.077 (0.08)
CRCM-NCEP	810	199.5 (23.8)	−0.059 (0.08)	–	–	–
ECP2-GFDL	1117	380.9 (44.0)	−0.101 (0.08)	1091	378.4 (36.5)	−0.109 (0.07)
ECP2-NCEP	1075	315.2 (42.0)	−0.007 (0.10)	–	–	–
MM5I-CCSM	1104	327.9 (34.4)	−0.200 (0.06)	1154	422.5 (39.5)	−0.088 (0.06)
MM5I-NCEP	975	242.9 (32.3)	0.122 (0.10)	–	–	–
RCM3-CGCM3	1090	344.0 (37.6)	−0.125 (0.07)	1119	266.1 (25.8)	−0.054 (0.07)
RCM3-GFDL	1171	349.8 (43.3)	−0.084 (0.09)	1155	421.3 (39.8)	−0.165 (0.06)
RCM3-NCEP	1035	338.8 (43.1)	−0.060 (0.09)	–	–	–
WRFG-CCSM	1158	378.4 (44.0)	−0.219 (0.08)	1234	477.0 (43.8)	−0.123 (0.06)
WRFG-CGCM3	1035	358.1 (44.7)	−0.170 (0.09)	1045	282.3 (28.1)	−0.073 (0.07)
WRFG-NCEP	1053	291.7 (39.4)	0.019 (0.10)	–	–	–

with a 95% confidence interval of $(-0.058, 0.194)$. Although the confidence intervals of these quantities are relatively wide, the point estimates indicate an increase in both the scale and shape parameters from current (see Table 2.1) to future climate.

These estimates of changes in q^*, ψ, and ξ are small, but their effect on changes in probabilities of extreme events is quite large. For example, using the GPD parameters estimated from reanalysis-driven WRF output in the current climate in Table 2.1, the 100-year winter precipitation event (the event which has probability of 0.01 of occurring in any given year) is estimated to have a value of 3105.8. Using the parameters estimated above for future reanalysis-driven WRF output, an event of at least this size has probability of about 1/36.3 of occurring in a given year. The 100-year event in the current climate becomes a 36.3-year event in future climate. Similarly, a 20-year event from the current run is estimated to be a 10.1-year event in the future scenario, which is in striking agreement with estimates reported for the western United States region under the A2 scenario in the IPCC 2012 special report on extremes (IPCC, 2012).

2.6.2 Marginal Distribution of Future Observed Precipitation Extremes

Having estimated the upper tail of the marginal distribution of X_t^{NF}, we can apply the statistical precipitation generator to simulate Fréchet-scaled realizations of future observed precipitation. In order to represent these realizations on their original scale, one needs to know the marginal distribution of Y_t^F, the observational precipitation quantity in future climate. Since the NCEP reanalysis offers a reconstruction of past climate based on observed weather, we turn to the reanalysis-driven WRF output to estimate this marginal distribution.

Letting Y_t^C and Y_t^F be the daily observed precipitation footprint for current and future climate, respectively, we assume

$$Y_t^C \stackrel{d}{=} aX_t^{NC}, \tag{9}$$

that is, the marginal distribution of daily observed precipitation is the same as that of reanalysis-driven precipitation output, up to a scaling constant a. Note that we do not imply daily equivalence here; the relationship is only in the marginal distributions over the current time period. We make a further assumption that this relationship holds for the future climate; that is, $Y_t^F \stackrel{d}{=} aX_t^{NF}$.

The maximum likelihood estimates of the GPD parameters in Table 2.1 provide an estimator of a in (9). Under the assumption (9), one can show that for a given threshold u,

$$[X_t^{NC} \mid X_t^{NC} > u] \sim \text{GPD}(\psi, \xi) \iff [Y_t^C \mid Y_t^C > au] \sim \text{GPD}(a\psi, \xi).$$

Thus an estimate of a is given by the ratio of GPD scale parameters from Table 2.1: $\hat{a} = \hat{\psi}_{Y_t^C}/\hat{\psi}_{X_t^{NC}} = 13.48$. The 95% confidence interval for a based on the delta method is $(8.48, 18.48)$. Figure 2.9 shows the quantile-quantile plot of Y_t^C and $\hat{a}X_t^{NC}$. The plot shows

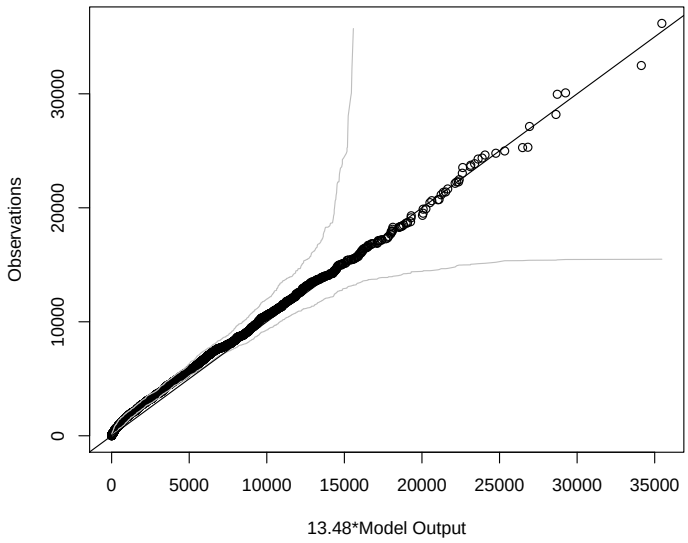

Figure 2.9: Quantile-quantile plot of scaled NCEP-driven WRF output quantity $\hat{a}X_t^{NC}$ and observational quantity Y_t^C, with confidence bands and 45° line added.

the assumption made in (9) is reasonable for the current climate, particularly when examining the upper tails of the two distributions.

2.6.3 Conditional Simulation of Future Extreme Winter Precipitation given WRF-CCSM Output

Having estimated the marginal distributions of the future unobserved quantities X_t^{NF} and Y_t^F, we aim to simulate realizations Y_t^F given X_t^{CF}, the day t precipitation quantity obtained from the future run of the WRF RCM coupled with the CCSM GCM. We apply a probability integral transformation to the realizations of the quantity X_t^{CF} to represent it in terms of X_t^{NF}. For a realization x_t^{CF}, this transformation results in the quantity x_t^{NF}, that we assume would be realized in the NCEP reanalysis-driven WRF model, given the same process dynamics as seen in the CCSM model on day t. Given these realizations of X_t^{NF}, a further transformation to the Fréchet-scaled Z_t^{NF} is made, and we can apply the

fitted Dirichlet model for H to simulate large realizations of Y_t^F, the observed precipitation quantity in future climate.

Definition 3 of multivariate regular variation in Chapter 1 states that the sequence of point processes $N_n = \{\mathbf{Z}_i/b_n\}_{i=1}^n$, where $\mathbf{Z}_i$ are independent copies of $\mathbf{Z} = (Z_1, Z_2)^T$, a regular varying random vector with tail index α, converge on Borel sets of $\mathfrak{C}$ to a nonhomogeneous Poisson process with intensity measure ν. In pseudo-polar coordinates the density of the limiting measure is $\nu(dr \times dw) = r^{-(\alpha+1)} dr H(dw)$ (Coles and Tawn, 1991). Here we work with unit Fréchet margins Z_1 and Z_2 (tail index $\alpha = 1$), and assume a differentiable measure H, so we write $\nu(dr \times dw) = r^{-2} h(w) dw$.

Cooley et al. (2012) showed that the above Poisson process intensity measure in Cartesian coordinates is $\nu(d\mathbf{z}) = \|\mathbf{z}\|^{-3} h(\mathbf{z}\|\mathbf{z}\|^{-1}) d\mathbf{z}$, for $\|\mathbf{z}\|$ large. As this point process convergence is valid on regions bounded away from $\mathbf{0}$, Cooley et al. (2012) derived an approximated conditional density of Z_2 when Z_1 is large: for $z_1 > r^*$,

$$f_{Z_2|Z_1=z_1}(z_2) \approx \frac{(z_1+z_2)^{-3} h(\frac{z_1}{z_1+z_2})}{\int_0^\infty (z_1+s)^{-3} h(\frac{z_1}{z_1+s}) ds}, \tag{10}$$

for r^* sufficiently large.

The aim of Cooley et al. (2012) is to approximate the conditional density of Z_2, given that Z_1 is large. Our objective is the reverse: we aim to simulate values of Z_2 such that $Z_2 > r^*$, given *any* value of Z_1. When $z_1 > r^*$, the approximated conditional density (10) of Cooley et al. (2012) can be used to simulate values of Z_2. When $z_1 \leq r^*$, it follows that for $z_2 > r^*$,

$$\begin{aligned}
f_{Z_2|Z_1=z_1}(z_2) &\approx \mathbb{P}(Z_2 \in (z_2, z_2+dz) \mid Z_1 = z_1) \\
&= \mathbb{P}(Z_2 \in (z_2, z_2+dz) \mid Z_1 = z_1, Z_2 > r^*) \cdot \mathbb{P}(Z_2 > r^* \mid Z_1 = z_1) \\
&\quad + \mathbb{P}(Z_2 \in (z_2, z_2+dz) \mid Z_1 = z_1, Z_2 < r^*) \cdot \mathbb{P}(Z_2 \leq r^* \mid Z_1 = z_1) \\
&= \mathbb{P}(Z_2 \in (z_2, z_2+dz) \mid Z_1 = z_1, Z_2 > r^*) \cdot \mathbb{P}(Z_2 > r^* \mid Z_1 = z_1).
\end{aligned} \tag{11}$$

Recall that for a Poisson process with intensity function λ and sets B_1 and B_2 with finite intensity measure,

$$\mathbb{P}(\mathbf{Z} \in B_1 \mid \mathbf{Z} \in B_2) = \frac{\int_{B_1 \cap B_2} \lambda(d\mathbf{z})}{\int_{B_2} \lambda(d\mathbf{z})}. \tag{12}$$

Define the sets $A = \{(s_1, s_2) : s_1 > 0, s_2 > r^*\}$ and $A' = \{(z_1, s_2) : s_2 > r^*\}$. Recognize that the Poisson process measure $\nu(\cdot)$ has been assumed to be valid on these sets. By applying (12), the first term in (11) is approximately $\nu(d\mathbf{z})|_{\mathbf{z}=(z_1,z_2)}/\nu(A')$. The second term in (11) is not amenable to our Poisson process approximation as the event on which we condition $(Z_1 = z_1 \leq r^*)$ is not sufficiently large, so we apply Bayes' rule:

$$\mathbb{P}(Z_2 > r^* \mid Z_1 = z_1) = \frac{\mathbb{P}(Z_1 = z_1 \mid Z_2 > r^*) \cdot \mathbb{P}(Z_2 > r^*)}{\mathbb{P}(Z_1 = z_1)}.$$

Since $\mathbb{P}(Z_1 = z_1 \mid Z_2 > r^*) = \mathbb{P}(\mathbf{Z} \in A' \mid \mathbf{Z} \in A)$, where $A' \subset A$, (12) gives

$$\mathbb{P}(Z_1 = z_1 \mid Z_2 > r^*) = \frac{\nu(A')}{\nu(A)}.$$

Finally, noting that $\mathbb{P}(Z_1 = z_1) \approx \mathbb{P}(Z_1 \in (z_1, z_1 + dz))$, we obtain

$$f_{Z_2|Z_1=z_1}(z_2) \approx \frac{\nu(d\mathbf{z})|_{\mathbf{z}=(z_1,z_2)}}{\nu(A')} \cdot \frac{\nu(A')}{\nu(A)} \cdot \frac{1 - F_{Z_2}(r^*)}{f_{Z_1}(z_1)}, \tag{13}$$

where

$$\nu(A) = \int_{w=0}^{1} \int_{r=\frac{r^*}{1-w}}^{\infty} r^{-2} dr h(w) dw$$

and

$$\nu(A') = \int_{r^*}^{\infty} (z_1 + s)^{-3} h\left(\frac{z_1}{z_1 + s}\right) ds.$$

To approximate the conditional density of large observational precipitation measurements, Z_t^{NF} (derived via transformation of X_t^{CF}) plays the role of Z_1 above, and $Z_2 = Z_t^F$, Fréchet-scaled daily observational precipitation measurements in future climate. We set $r^* = 40$, which is the threshold chosen for fitting the dependence model in Section 2.4. Here the event $Z_t^F > 40$ corresponds to $Y_t^F > 18070$, which is the 0.979 empirical quantile of current climate observations, but given the parameters estimated in Section 2.6.2, corresponds to the 0.963 quantile of future observed precipitation.

Given the output from the CCSM-driven WRF model on any given day in the future, we approximate (at least the upper tail of) the conditional density of Y_t^F using (13) if $Z_t^{NF} \leq r^*$, and using (10) if $Z_t^{NF} > r^*$. For illustration, we show three events in Figure 2.10. The January 20, 2063 event is the largest event produced by the WRF-CCSM future run, with a precipitation value of $x_t^{CF} = 3343.29$. Let $y^* = \max_t(y_t^C)$, that is, the largest observational footprint from the current climate. For the January 20, 2063 output of the RCM, the approximated conditional density estimates $\mathbb{P}(Y_t^F > y^* \mid X_t^{CF} = x_t^{CF}) = 0.66$. The December 21, 2051 event was the 24^{th} largest in the future WRF run, with a value of $x_t^{CF} = 1819.60$. Given this output, the approximated conditional density estimates $\mathbb{P}(Y_t^F > y^* \mid X_t^{CF} = x_t^{CF}) = 0.008$.

The conditional densities of the two events above were derived via (10). To illustrate (13), we plot the upper tail of the conditional density of Y_t^F on January 11, 2069. The WRF-CCSM precipitation output on this day ($x_t^{CF} = 1181.98$) was quite large; it was the 155^{th} largest of the 3569 days in the future period. But the output from this day was not 'extreme', in the sense that $z_t^{NF} = 23.84 < r^*$. Given the output of this day, we show the upper tail of the approximated conditional density in Figure 2.10, and estimate $\mathbb{P}(Y_t^F > 18070 \mid X_t^{CF} = x_t^{CF}) = 0.15$; that is, the probability that the observed precipitation footprint is 'extreme'. Using the approximation, for this day we also find $\mathbb{P}(Y_t^F > y^* \mid X_t^{CF} = x_t^{CF}) = 0.0001$.

One should recall that climate models are not weather prediction models, but rather climate simulation models. Thus the exact dates attached to the individual events discussed

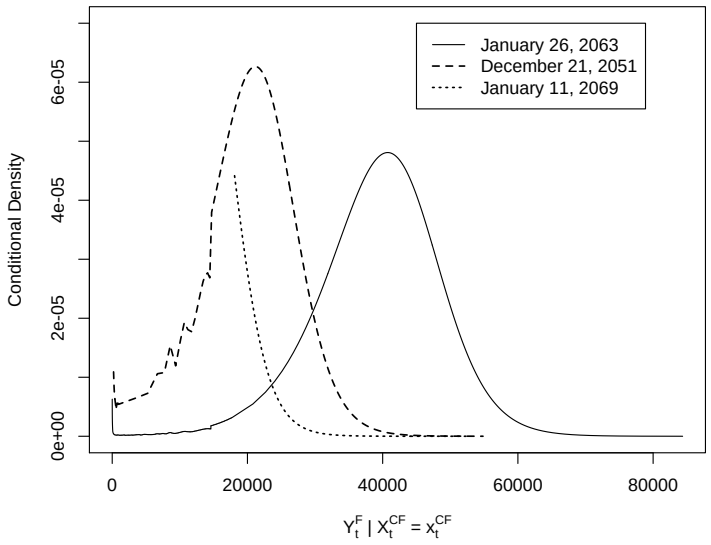

Figure 2.10: Conditional densities of observational precipitation given WRF-CCSM output on labeled day. January 11, 2069 event is approximated via (13). Other two densities are approximated via (10).

above are simply timestamps; we consider their precipitation outputs to be draws from the distribution of climate model-generated precipitation over the time period in which the events occur. A more appropriate interpretation is that if the large-scale dynamics are as the GCM supplies to the RCM on a given day, Figure 2.10 gives the conditional densities of observational precipitation on that day, given the RCM output.

The two components of (11) provide a means for simulating large realizations of Z_2 given small values of Z_1. Given $Z_1 = z_1$, where $z_1 \leq r^*$, the second piece is evaluated and this probability is used to simulate $I_{\{Z_2 > r^*\}}$. Conditional on this event, the first piece is used to draw a realization of Z_2. We employ (10) and (13) to generate future observational precipitation measurements, given CCSM-driven WRF output.

Figure 2.11 shows the future CCSM-driven WRF output X_t^{CF} plotted against one realization of simulated observational footprints Y_t^F, along with marginal histograms. We note that while the CCSM-driven output on the x-axis is estimated to have a bounded tail, the simulated observational footprints exhibit a heavy tail. This is due to the transformation to

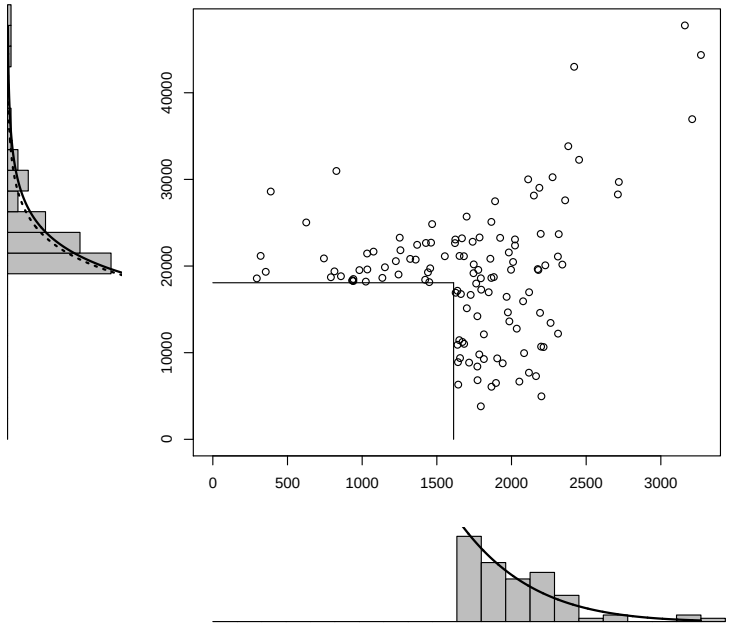

Figure 2.11: Simulated observations of precipitation footprint Y_t^F (y-axis) from future run of WRF-CCSM output X_t^{CF} (x-axis). Marginal tail histograms added, as well as estimated GPD tail densities (solid lines). Dashed line corresponds to estimated GPD tail of observations in current climate (Y_t^C).

future reanalysis-driven output before the simulation. The estimated density of the current observations Y_t^C is also shown to illustrate the estimated change in tail behavior of observed precipitation from current to future climate.

While Figure 2.11 also shows 'non-extreme' precipitation simulated from large values of WRF-CCSM output, the primary interest is in examining the simulated large observational footprints $\{y_t^F : z_t^F > r^*\}$. Applying this statistical generator allows us to examine individual extreme precipitation events from the simulation, and the process dynamics associated with each. Specifically, we study the PE precipitation index (as defined in Section 2.5) of these simulated future observations.

2.6.4 Pineapple Express Index of Future Extreme Precipitation Events

By tying the simulated observations Y_t^F to WRF-CCSM output X_t^{CF} through X_t^{NF}, we can link Y_t^F to large-scale dynamics. We extract daily mean SLP anomalies from the future

run of the CCSM global model that drives the future run of the WRF regional model. The SLP fields were downloaded from the Earth System Grid[7]. The PE index is calculated for future events by projecting future daily anomaly fields onto the mean PE field from Section 2.5.

From the future values of the PE index, we find evidence suggesting an increase in frequency and intensity of PE events in future climate. We find that 8.9% of PE index scores in the future run exceed the 0.95 quantile of PE scores in the current climate. This corroborates the findings of Leung et al. (2011), which suggests an increase in both the frequency of PE events and their total contribution to western United States precipitation. One can also examine $\chi(q_{95})$ of Coles et al. (1999) between the daily PE index and RCM output. In the future scenario $\hat{\chi}(q_{95}) = 0.23$, an increase from the current climate, for which $\hat{\chi}(q_{95}) = 0.18$. This suggests that the tail dependence between the PE index and RCM-produced precipitation increases from current to future, which may be a result of increased precipitation intensity from PE events.

To study the link between the PE index and simulated precipitation extremes in the future, we again assume normality of the anomaly fields, and for ease of interpretation we represent the PE index in terms of z-scores. We apply our statistical precipitation generator to obtain a realization of Y_t^F, $t = 1, ..., T_f$, and repeat the simulation 500 times. Figure 2.12 shows the observed precipitation events simulated from one realization of the generator and their associated PE index scores. It is apparent that many of the simulated extreme events in future climate have very high PE index values. A useful quantity to examine is the proportion of simulated extreme observations (i.e., for which the simulated value $z_t^F > r^*$) which have PE index scores that exceed the 0.95 quantile of PE index scores over the future period. Over 500 simulations, the mean of this quantity was found to be 0.284, with 95% of all values falling between 0.213 and 0.356. The realization shown in Figure 2.12 had

[7] `http://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm.output.html`

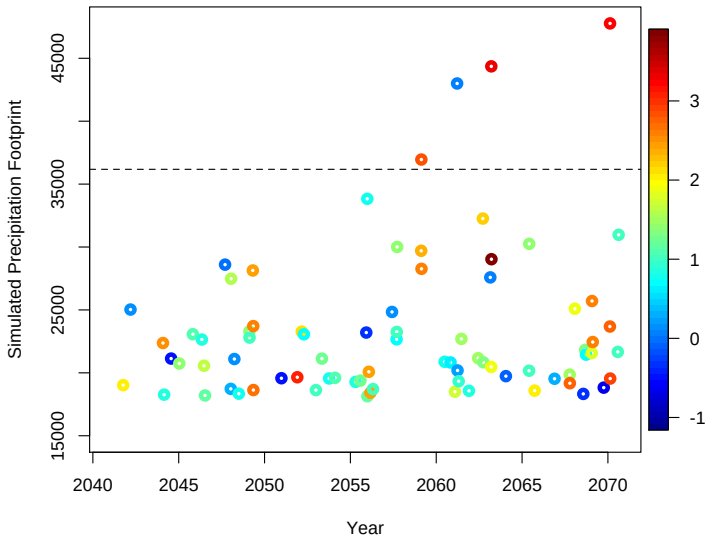

Figure 2.12: Realization of future observed daily extreme precipitation footprints Y_t^F from the generator, plotted against time. Legend strip indicates z-score of the event's associated PE index; horizontal line corresponds to the largest observed event from current climate.

31.6% of observations exceeding this quantile. This is similar to the proportion of extreme observations with high PE index values in the current climate (32.1%).

Figure 2.12 also gives some evidence for non-stationarity of the frequency and intensity of extreme west coast precipitation *within* the future period 2041-2070. The largest 11 precipitation events simulated in this particular realization occur in the years 2055-2070, and 57% of the extreme events simulated occur in the years 2056-2070. Over 500 simulations, the mean of this proportion was 0.571, with 95% of values falling between 0.477 and 0.656. This potential non-stationarity deserves attention in future work.

One can reverse the perspective and examine precipitation simulated from future events which have large PE index values. Figure 2.13 shows the largest 79 days in terms of the future PE index, and the precipitation footprints simulated from these. This number was chosen to correspond with the number of precipitation events that were simulated as extreme (i.e., for which $z_t^F > r^*$) in this realization. In Section 2.5 we found the PE index was tail dependent to observed precipitation in the current climate. If this was true for the future climate, we

would expect some of these large PE index days to exhibit extreme precipitation, and this is indeed the case. In this realization, 18 of the 79 (22.7%) extreme observed precipitation days are also among the 79 largest days in terms of PE index values. Over 500 realizations, an average of 20.3% of the extreme precipitation days were also among the exceedances of the same quantile of the future PE index, with a simulation-based 95% interval of $(0.144, 0.257)$.

The proportion estimated above can be thought of as similar to an estimate of $\chi(q)$ of Coles et al. (1999), for q a very high quantile, between the PE index and observational precipitation. The quantile q changes from one realization to the next based on the number of events for which Z_t^F was simulated to be extreme in that particular realization. The mean of this quantity (0.203) is a significant increase from the current climate scenario, for which the proportion reported above is 0.143. This suggests that the level of tail dependence between the PE index and observed precipitation increases from the current to future scenario. An increase in tail dependence may be the result of PE events producing more intense precipitation in the future, as was found by both Leung et al. (2011) and Dettinger (2011).

Applying a fitted model for tail dependence is a novel approach to generating observed precipitation extremes from climate model output. It allows one to simulate realizations of future observations based on knowledge of how well climate models represent extreme events. The fitted dependence model accounts for uncertainty in what an extreme event, as produced by an RCM driven by a chosen GCM, will look like in observations. Simulations of observations from climate model output can also be used to study the connection between observed extreme precipitation and synoptic-scale processes.

2.7 Summary and Discussion

In Section 2.1.2, we proposed several novel goals which this work aimed to achieve. We have proposed new ways of utilizing bivariate extreme value methods within the context of studying winter precipitation on the Pacific coast and nearby areas. In Section 2.4, we categorized the tail dependence between observed precipitation and that produced by a

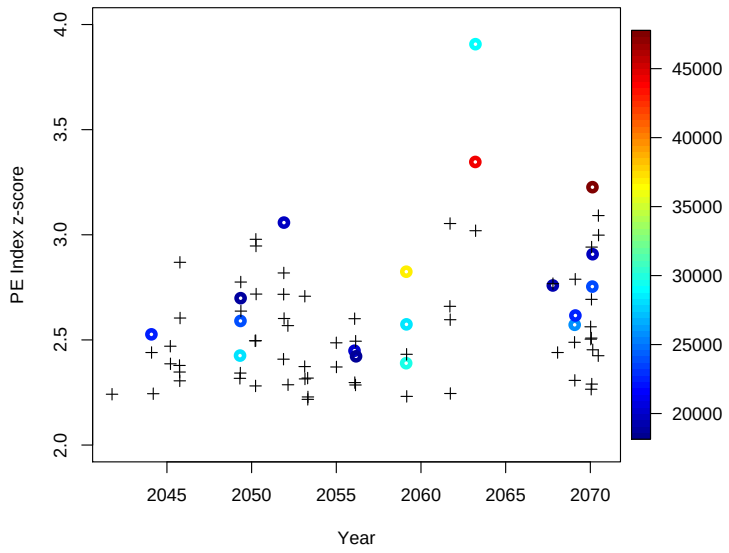

Figure 2.13: Largest 79 PE index z-scores from future climate. Circles indicate simulated observed precipitation y_t^F from that particular day, with the legend strip giving the precipitation value. Crosses indicate the observed precipitation from that day was not simulated to be extreme in this particular realization.

reanalysis-driven RCM as asymptotically dependent, then further modeled the dependence parametrically. This dependence model can be viewed as a statistical simulator of gridded observations from given climate model output. We used SLP fields in Section 2.5 to build a straightforward and easily-defined daily PE precipitation index. We found this index to be asymptotically dependent to observed precipitation. This index links precipitation extremes to large-scale, short-lived process dynamics. Because of its simplicity, the index is can be readily obtained from output from future runs of GCMs. Finally, we extended the statistical generator to RCM output driven by a future GCM run. Doing so required estimation of the marginal distribution of two unobserved quantities: RCM output that would be obtained from a future reanalysis product and the gridded observations. From this estimation, we found evidence of a heavier tail of precipitation in the Pacific region in the future scenario. By relying on the dependence structure estimated in Section 2.4, we can simulate bivariate observations of these two unobserved quantities. Repeated conditional simulation can be

used to account for uncertainties in how RCMs represent these extreme events. We further extended the future precipitation generator by linking to future values of the PE index.

It is worthwhile to review important assumptions made throughout this work. Our procedure for generating future observational precipitation assumes that the parameters of the extremal dependence model fitted in Section 2.4 do not change in future climate; that is, that the future climate model run produces extreme precipitation as 'correctly' as the current climate model run. Also, we have assumed throughout this work that daily precipitation measurements and PE index values are independent and identically distributed. It is known that precipitation from the PE and other storms can persist for several days at a time. Furthermore, SLP patterns associated with PE events develop and evolve over the course of several days, which is in violation of these assumptions. While this likely does not bias estimates, the confidence intervals related to tail dependence parameters will be anticonservative.

The simulation of future extreme precipitation introduced several unique challenges. We turned to precipitation output from other climate models in Section 2.6 in order to estimate the marginal distribution of future reanalysis-driven WRF output. We judged this to be the best approach, due to the fundamental difference in tail behavior between the GCM-driven and reanalysis-driven RCM outputs. Having little information about the future marginal distribution of observed precipitation, we turned to the connection between observations and reanalysis-driven RCM output in current climate, and again assumed this relationship was valid for future precipitation.

We recognize that the work herein represents early attempts at answering difficult questions, and much can be done to extend the work done here. The PE index in particular can almost certainly be improved by incorporating more large-scale information. As the PE typically exhibits itself over the course of several days, an index developed from multiple days is a possible extension. Our link between future precipitation observations and the PE

index could be made more explicit by conditioning on the value of the index as well as the RCM precipitation output.

There are also other avenues for continued work applying multivariate extremes methods to the analysis of climate model output. The methods used in Section 2.4 to examine tail dependence can be used for climate model assessment. It is well known that discrepancies exist in climate model output quantities (Rougier, 2007). Differences in the tail behavior of temperature and precipitation from climate models have been found in other studies (Schliep et al., 2010; Frei et al., 2006). Weller et al. (2013) found that different RCMs forced by the same global reanalysis produce different tail behavior in west coast precipitation output. Examination of tail dependence between climate model output and observations provides a means of studying how well climate models are able to simulate extreme events. Finally, while we applied our statistical generator to the WRF RCM forced by reanalysis and a future scenario run of one GCM, it could easily be applied to other climate models as well.

CHAPTER 3

TWO CASE STUDIES ON NARCCAP PRECIPITATION EXTREMES

3.1 Introduction

In recent years, the literature on climate change has devoted increasing attention to potential changes in the frequency and severity of extreme weather events. Several studies have projected significant changes in extremes due to climate change (e.g., Frei et al., 2006; Karl and Melillo, 2009; Allan and Soden, 2008), and recent efforts have linked some individual extreme events to human-induced warming (Peterson et al., 2012). Because many of the aforementioned studies employed deterministic simulation models to study extreme events, it is important to understand if and how extreme weather events manifest themselves in such models. In some cases, it has been found that climate models are unable to reproduce extreme weather statistics from the past observed record (Wehner, 2013). Additionally, because extreme weather events are often by definition rare, there is great uncertainty in future projections of these events based on both climate model data and past observations.

In this chapter, we examine extreme precipitation produced by the six regional climate models (RCMs) of the North American Regional Climate Change Assessment Program (NARCCAP). Previous studies have examined extreme precipitation from NARCCAP models from different perspectives. Schliep et al. (2010) perform statistical analyses using spatial hierarchical modeling to fit generalized extreme value (GEV) distributions to annual maximum precipitation amounts at each climate model output gridbox. Mailhot et al. (2011) examine future changes in annual maxima of precipitation measurements over Canada from the NARCCAP ensemble. Wehner (2013) used the GEV distribution to compare the be-

havior of past extreme precipitation from NARCCAP models to that seen in observations. Gutowski et al. (2010) examined monthly extreme precipitation events in two different regions of North America, comparing output of the NARCCAP models to observations.

This chapter first describes the marginal behavior of daily extreme precipitation produced by each of the six NARCCAP RCMs, as well as that seen in an observational product. We go on to describe the *tail dependence* in daily precipitation amounts between each RCM and the observational product; that is, we aim to discover the extent to which the most extreme precipitation events produced by each RCM *correspond* to the most extreme events seen in the observations. Studies of marginal extremes of climate variables have an extensive history, and Section 2.9 of Von Storch and Zwiers (2002) gives an overview of appropriate statistical methodologies, all of which analyze a subset of data deemed to be extreme. One common approach is to model seasonal maximum precipitation amounts using the GEV (Wehner, 2013). Alternative methods exist for modeling exceedances over a threshold, and we employ a threshold exceedance approach using the Generalized Pareto distribution (GPD). There are two advantages to using this approach here, rather than the GEV. First, the precipitation record studied here covers a relatively short time period (23 years), and an approach using daily exceedances will offer reduced uncertainty over the GEV, which retains only one data point per year or season. Second, as we wish to compare model output to observations, a threshold exceedance approach allows us to maintain the daily correspondence between these two.

The second and more important aim of this work is to examine the ability of the reanalysis-driven NARCCAP models to reproduce past extreme precipitation events seen in observations. The objective here is to examine the extent of the agreement (on a daily basis) of simulated NARCCAP extreme precipitation with past observed extremes, and this requires multivariate techniques. Typically, statistical dependence measures such as correlation or covariance are quite useful for summarizing dependence in climate and atmospheric data. While these measures are effective at capturing dependence in the center of a mul-

tivariate probability distribution, they do not measure dependence in the joint tail of the distribution. Here, we examine tail dependence through an established probability framework designed specifically for the joint tail of a multivariate distribution (Resnick, 2007). To summarize tail dependence, we employ measures which gauge the extent of correspondence of the largest events in each margin.

To our knowledge, the use of bivariate extreme value methods to compare the extremes of two data sources is novel to the atmospheric sciences literature. A marginal analysis allows one to compare summary statistics from the two sources; Wehner (2013) compared NARCCAP precipitation extremes to observational data using summary statistics obtained from estimated 20-year return values. However, a marginal analysis cannot describe the daily correspondence of extreme behavior. Our analysis is also fundamentally different from conditional approaches (e.g. Sillman et al., 2011; Katz, 2010; Zhang et al., 2010) which model the parameters of the GEV or GPD to be functions of covariates such as time or some large-scale atmospheric variable. A conditional approach is not appropriate here as the quantities we wish to relate are both observed daily. With data such as this, we are not able to extract a subset of extreme values of one variable conditional on a particular value of the other variable. Instead, we treat daily precipitation amounts from NARCCAP output and observations as emerging from a bivariate probability distribution.

We investigate two case studies of extreme precipitation, over different regions and different seasons. The first is a study of winter precipitation on the Pacific coast, which was partially explored in Weller et al. (2012). A portion of that work examined past daily winter-season precipitation over a west coast region of North America in regional climate simulations produced by the Weather Research and Forecasting (WRF) model, one of the NARCCAP RCMs, and an observational product (Maurer et al., 2002). It was found that strong tail dependence in precipitation amounts existed between the WRF model output and observations. We extend that work here by examining tail dependence between the observations and all six RCMs in NARCCAP for winter precipitation in this region. Second,

we investigate summer-season precipitation extremes over the prairie (Corn Belt) region of the US.

These two case studies are chosen due to the differences in the nature of precipitation events between the two regions and seasons. While wintertime precipitation on the west coast is often driven by strong larger-scale systems, central US summer precipitation extremes are typically associated with convective systems forced by processes that are more local and regional in scale. Thus, while past wintertime extreme precipitation events from the NARCCAP models may show correspondence with observations (as Weller et al. (2012) found with WRF), it is not guaranteed that this will be the case for summer precipitation.

The outline of the chapter is as follows: in Section 3.2 we describe the NARCCAP program in more detail and introduce the model output and observational data sources used in this work. Section 3.3 details an examination of the ability of the NARCCAP models to simulate past observed winter precipitation extremes along a west coast region of North America. This section also includes a review of our statistical techniques, which are based in extreme value theory. In Section 3.4 we employ similar techniques to study the models' representations of extreme summer precipitation events over a central US region. We conclude with a summary and discussion in Section 3.5. Much of this chapter is taken from Weller et al. (2013), which has been submitted for publication.

3.2 NARCCAP Models and Observations

The North American Regional Climate Change Assessment Program (NARCCAP) is an international coordinated effort to investigate uncertainties in high-resolution dynamical simulations of regional climate over North America (Mearns et al., 2009). NARCCAP consists of a suite of six regional climate models (RCMs) run over a common spatial domain at similar resolutions ($\sim$50 km) and over common time periods. Phase I of the experiment involves running each RCM for the period 1979-2004 with boundary conditions provided by the National Center for Environment Prediction (NCEP) – Department of Energy (DOE)

Table 3.1: Acronyms, full names, and major references for RCMs employed in this study.

Acronym	Name	References
CRCM	Canadian RCM	Caya and Laprise (1999)
ECP2	Experimental Climate Prediction Center's version of the Regional Spectral Model	Juang et al. (1997)
HRM3	Third-generation Hadley Centre RCM	Jones et al. (2003)
MM5I	Fifth-generation Pennsylvania State University National Center for Atmospheric Research (NCAR) Mesoscale Model	Grell et al. (1994)
RCM3	International Centre for Theoretical Physics RCM version 3	Giorgi et al. (1993a,b); Pal et al. (2007)
WRFG	Weather Research and Forecasting model	Skamarock et al. (2005)

global reanalysis II product (Kanamitsu et al., 2002). In Phase II, the regional models are run with boundary conditions provided by four different fully coupled global climate models (GCMs) in a fractional factorial design, for the years 1981-2003 (control period) and 2041-2070, under the IPCC A2 scenario (Nakicenovic et al., 2000). As we wish to study past daily correspondence between NARCCAP model output and observations, we only use output from Phase I.

The six RCMs and their major references are listed in Table 3.1. The CRCM and the ECP2 are the only two RCMs that include some form of interior nudging (a push toward the large-scale driving conditions in the interior of the domain). Additional details on NARCCAP and the configuration of the models can be found in Mearns et al. (2012), on the program website[8], and in the provided references.

In this work, we compare precipitation output of each NARCCAP RCM to the gridded observational product produced by Maurer et al. (2002). This product consists of spatially gridded precipitation amounts interpolated from weather station measurements. In constructing this product, the PRISM technique (Daly et al., 1997) was employed to correct for elevation, which is particularly important in mountainous regions. This product is gridded at 1/8° resolution and on a daily temporal scale, with each day defined to begin at midnight local time. We compare daily precipitation output from each of the six NARCCAP models,

[8] `http://www.narccap.ucar.edu`

driven by the NCEP reanalysis, to daily precipitation in the observational product for the years 1981-2003.

A challenge in comparing output from the NARCCAP models and the Maurer et al. (2002) product arises from the differences in spatial resolutions of the two sources. Wehner (2013) noted significant differences over some regions and seasons in precipitation statistics between the Maurer et al. (2002) gridded product and a coarser observational product. We will see in Sections 3.3.1 and 3.4.1 that the observational product appears to capture some small-scale precipitation phenomena in greater detail than the NARCCAP model output. When comparing output from the two sources, we will define quantities of interest which represent output over approximately common spatial areas. After comparing the tail behavior of the marginal distributions of each source, we will study the correspondence of the sources' extreme events. Our investigation of tail dependence will require us to view the different sources on a common scale, achieved via marginal transformation.

3.3 Tail Behavior of Pacific Region Winter Precipitation

We begin with an analysis of daily extreme Pacific region precipitation events seen in the observational record and each of the NARCCAP models forced by the NCEP reanalysis product. We first examine the nature of the upper tails of the probability distribution of precipitation produced by each model. Second, we study the tail dependence in precipitation events between each model and the observational data; the objective here is to determine the extent to which the largest precipitation events in the observed record correspond to the largest events produced by the regional models. We examine daily precipitation occurring in the winter season plus November (NDJF) months for the years 1981-2003, which results in a sample of $T_w = 2765$ days.

3.3.1 Study Region and Quantity of Interest

Our study region is defined as the area between 32°N and 53°N latitude, and from 118°W longitude to the Pacific coast; this is the same region studied in Weller et al. (2012). This region captures extreme precipitation events affecting the coastal regions of California and the Pacific Northwest, as well as heavy precipitation in the mountainous regions of the Cascades and Sierra Nevadas. Figure 3.1 shows the 1981-2003 mean precipitation field over this region from the observational product, as well as the NARCCAP ensemble mean. The figure indicates that, on average, the NARCCAP models are able to reproduce the patterns of spatial variability in winter precipitation over the region. However, their coarse resolution relative to the observational product does not allow them to capture localized areas of heavy precipitation seen in observations, such as over the Olympic peninsula in Washington.

While Figure 3.1 shows time-averaged precipitation, our interest here is daily precipitation output from NARCCAP and the observational product. As in Weller et al. (2012), we extract the maximum total precipitation over an area of approximately 200×200 km^2 (4×4 RCM grid cells, 17×17 grid cells from the observational product). Thus the daily quantity we examine corresponds to the total precipitation amount over the 'footprint' which has the greatest precipitation intensity on a given day. This size footprint was chosen in order to adequately capture the spatial extent and intensity of wintertime extreme precipitation events in the region. Define X_{jt} to be this quantity on day t from RCM j, where $t = 1, ..., T_w$ and $j = 1, ..., 6$. Let Y_t be this quantity extracted from the observational data product on day t. We do not require spatial matching of footprints from model output and observations on a given day.

Figure 3.2 shows scatterplots of the quantity X_j against the observations Y for each regional climate model. We see strong dependence in our precipitation quantity between the observational product and reanalysis-driven model output, and no great discrepancies exist in the scatterplots between the six regional models. Note that the difference in measurement

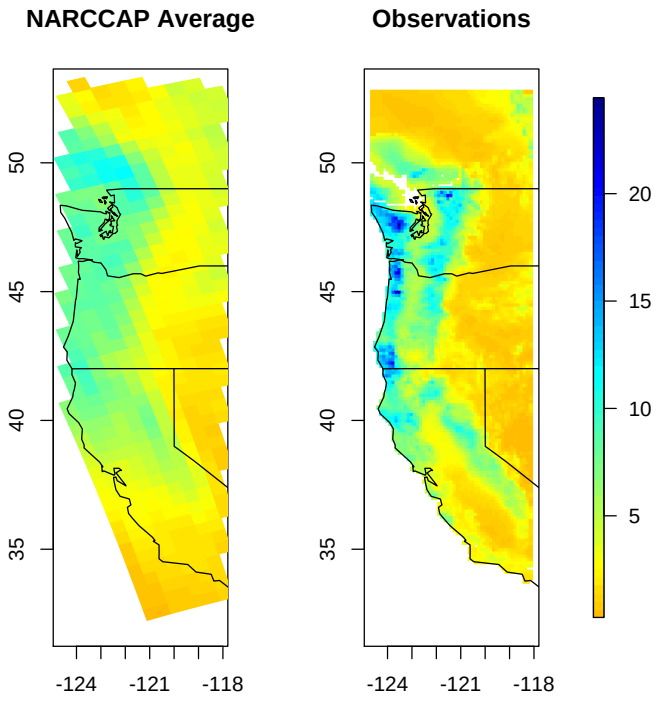

Figure 3.1: Mean NDJF precipitation field over Pacific study region from 1981-2003 from NARCCAP ensemble average (left) and observational product (right). Figures are on a common scale (mm/day) indicated by the legend strip.

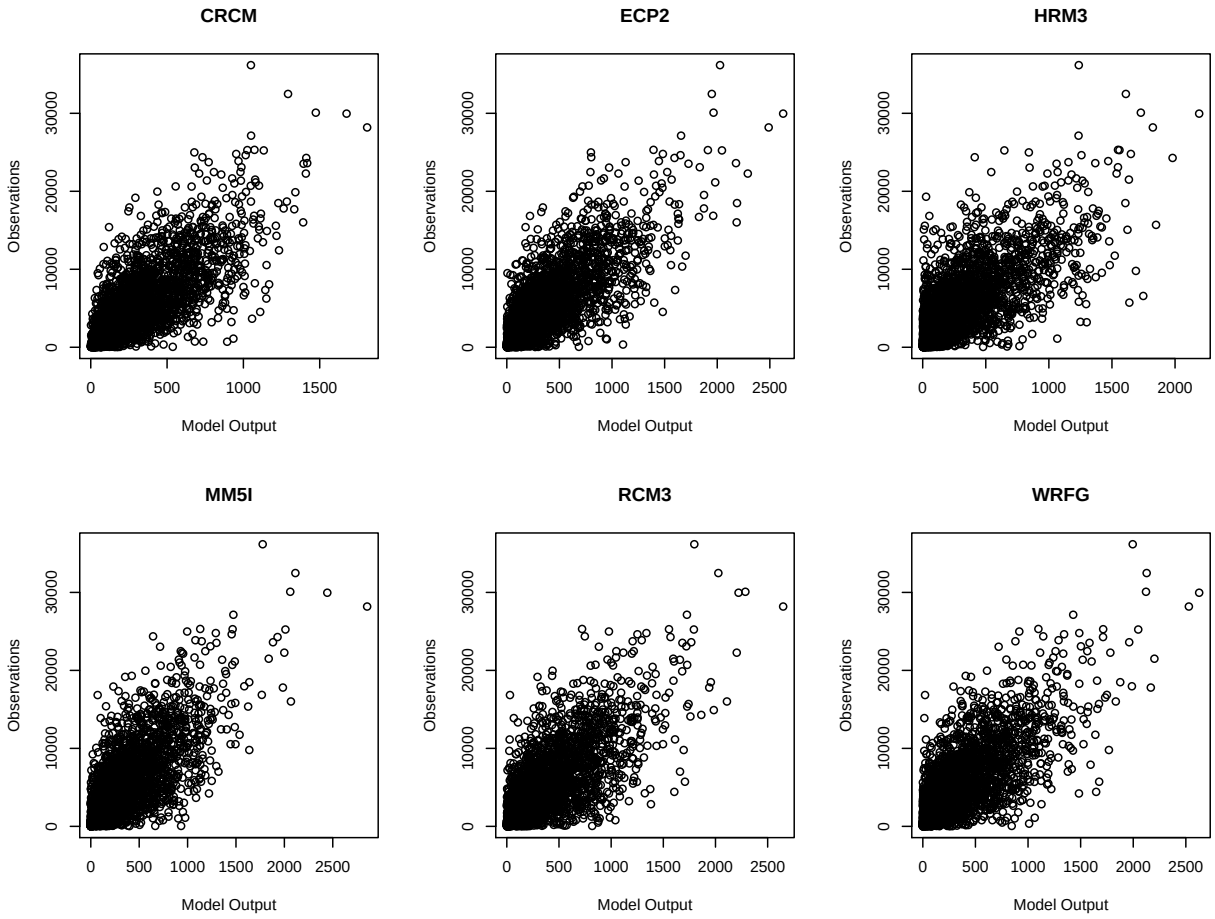

Figure 3.2: Modeled daily winter precipitation quantity X_j plotted against observed precipitation Y, for each regional climate model. Note the scales differ due to differing spatial resolutions of the model output and observational product.

scales between model output and observations is due to the differing resolutions of the NARCCAP models and observational product. We examine the Spearman rank correlation between X_j and Y for $j = 1, ..., 6$. The Spearman correlation is used here as opposed to the usual Pearson correlation, due to its robustness to extreme values. Correlations range from 0.704 (RCM3) to 0.761 (CRCM). While Spearman correlation measures dependence over all levels of the distribution of $(X_j, Y)^T$, our primary aim is to examine tail dependence; we thus proceed to examine the upper tails of the distributions of Y and the X_j for $j = 1, ..., 6$.

3.3.2 Marginal Estimation

We first examine the upper tails of each precipitation quantity X_j and Y. The aim is to study the consistency of the different models' representations of extreme precipitation, as was done via the GEV in Wehner (2013). We fit the GPD to the precipitation quantities extracted from both the observational product and each NARCCAP model. After examining diagnostics to determine an appropriate threshold (Coles, 2001, Chapter 4), the GPD is fit to exceedances of the 0.955 quantile of each X_j and Y, precipitation from the observational product. A summary of chosen thresholds, maximum-likelihood estimated parameter values, and their standard errors is given in Table 3.2. Because of the differences in resolution between the observational product and each regional model, the thresholds u_j are not directly comparable to the threshold chosen for the observations. However, one can see differences in the quantity of precipitation produced from each of the NARCCAP models, with CRCM having the smallest value for its 0.955 quantile, and ECP2 having the largest.

Table 3.2 also shows estimates of ξ, the shape parameter of the GPD, for each NARCCAP RCM and observations. One sees that the precipitation quantity Y derived from the observational data product is estimated to be slightly heavy-tailed, while five of the six NARCCAP models produce bounded-tail estimates for precipitation. Only the MM5I regional model produces a positive estimate of ξ, and this parameter is estimated to be much larger than that estimated from the observational product. The standard errors on estimates of ξ

Table 3.2: Threshold selected and maximum-likelihood parameter estimates (standard errors) from GPDs fit to exceedances of the 0.955 quantile of winter precipitation quantity (125 exceedances) from each source. Also shown are estimated 20 and 50 year return levels (mm) with 95% profile likelihood confidence intervals. Return values and confidence intervals have been normalized to gridbox-level values.

j	Model	u_j	$\hat{\psi}_j$ (se)	$\hat{\xi}_j$ (se)	$\hat{x}_{j,20}$ (CI)	$\hat{x}_{j,50}$ (CI)
1	CRCM	863	172.5 (21.6)	−0.02 (0.09)	102.3 (93.0, 125.7)	111.3 (98.6, 148.0)
2	ECP2	1129	325.9 (43.8)	−0.04 (0.10)	157.4 (140.5, 203.5)	172.5 (149.4, 245.3)
3	HRM3	1032	273.9 (32.3)	−0.13 (0.08)	124.5 (115.6, 145.8)	132.5 (114.2, 161.6)
4	MM5I	1026	246.7 (33.3)	0.11 (0.10)	159.0 (135.0, 222.5)	184.0 (148.3, 293.9)
5	RCM3	1093	325.2 (42.4)	−0.06 (0.10)	151.6 (136.4, 192.4)	165.4 (144.9, 228.7)
6	WRFG	1086	339.8 (43.2)	−0.06 (0.09)	153.8 (138.4, 193.1)	167.7 (147.2, 228.0)
-	(Obs)	14969	3938.5 (554.6)	0.00 (0.11)	116.1 (102.4, 154.8)	128.8 (109.5, 192.1)

reflect the relatively large uncertainty in these estimates, which is inherent in tail estimation problems. Indeed, we cannot conclude that any differences exist in the shape parameter between the different RCMs. Finally, we note that Weller et al. (2012) estimated the tail parameter ξ to be positive for the WRFG model using data from 1981-1999; the analysis here includes daily precipitation measurements through the year 2003.

In Table 3.2, we also show estimates of $x_{j,m}$, the m year return value for our precipitation quantity from regional climate model j, for $m = 20, 50$. For ease of comparison, return levels and associated confidence intervals have been scaled to be gridbox-level values. We see that slight differences in estimates of the tail parameter ξ produce larger differences in the return levels as the return period increases. As an example, consider the HRM3 and MM5I regional models. The thresholds chosen are nearly the same for each, but due to the difference in estimates of ξ, the estimated 20 and 50 year return values are quite different. Also shown are 95% profile likelihood confidence intervals for the return values, which reflect the relatively large uncertainty in these estimates. Large uncertainty is seen even for the 20 year return value, which corresponds to a time frame shorter than that of the NARCCAP simulations studied here.

Due to differences in the study region and the quantity of interest, the 20 year return values reported in Table 3.2 are not directly comparable to results reported in Wehner

(2013). Twenty-year return values were reported by Wehner (2013) for winter-season (DJF) precipitation over the western United States (defined to be west of 100° W longitude), and these values were compared to those obtained from the Maurer et al. (2002) observational data. The 20 year return values in Table 3.2 show some agreement with Wehner (2013) in that five of the six models exhibit larger return values than the observational product. The exception is CRCM, which Wehner (2013) found to have the smallest disagreement with this observational product over the western US in winter.

3.3.3 Tail Dependence

Having examined the marginal behavior of our precipitation quantities X_j and Y from reanalysis-driven NARCCAP models and observations, respectively, we turn attention to the tail dependence between each X_j and Y. When an extreme precipitation event is seen in the observed record on a given day, we expect that, to a certain extent, the synoptic-scale atmospheric conditions which produced that event will be reflected in the NCEP reanalysis for that day. When these boundary conditions are fed into the regional climate model, we aim to discover whether the model is likely to produce an extreme precipitation event.

A first step in examining tail dependence between two study variables is to determine whether they are asymptotically dependent or asymptotically independent. Two random variables Z_1 and Z_2 with common marginal distribution function F are said to be *asymptotically dependent* if

$$\chi = \lim_{z \to z_*} \mathbb{P}(Z_2 > z \mid Z_1 > z) > 0, \tag{14}$$

where z_* is the (possibly infinite) right endpoint of the support of Z_1 and Z_2. Asymptotic independence occurs when $\chi = 0$. Asymptotic dependence implies that the very largest events in one margin exhibit some correspondence to the largest events in the other margin. This is an important feature, and it is different from dependence in the usual sense of corre-

lations and covariances. For example, a bivariate Gaussian distribution with any correlation less than one exhibits asymptotic independence (Sibuya, 1960). In practice, a determination that asymptotic dependence is present must be made from the data.

As an exploratory step, we examine the level of tail dependence between each NARCCAP model and observations via a metric introduced by Coles et al. (1999):

$$\chi_{j,q} = 2 - \frac{\log \mathbb{P}(Y < y_q, X_j < x_{j,q})}{\log \mathbb{P}(Y < y_q)}, \tag{15}$$

where y_q and $x_{j,q}$ are the q quantile levels of Y and X_j, respectively. It can be shown that $\lim_{q \to 1} \chi_{j,q} = \chi_j$, where the subscript j is added to (15) to indicate we are measuring the level of extremal dependence between X_j and Y. Coles et al. (1999) suggest an empirical estimator of $\chi_{j,q}$ for large values of q as a diagnostic for assessing tail dependence.

Figure 3.3 shows a plot of estimated $\chi_{j,q}$ plotted against q for $q \in [0.75, 1)$ and $j = 1, ..., 6$. We see that all six regional models exhibit similar tail dependence with observations, with the $\chi_{j,q}$ estimates fluctuating near 0.5 for large q. Confidence intervals (shown only for CRCM) reveal no significant differences in $\chi_{j,q}$ between the different regional models, and show great uncertainty in these estimates for q close to 1 due to the limited amount of data exceeding high quantiles. The plot provides strong evidence that $\lim_{q \to 1} \chi_{j,q} > 0$ for each j; that is, each bivariate pair $(X_j, Y)^T$ exhibits asymptotic dependence. In other words, some correspondence is seen between the very largest daily precipitation amounts from the observational product and each of the six models.

As seen in the definition of χ in (14), an examination of tail dependence requires that each component of a given random vector has a common marginal distribution. In order to further examine the tail dependence in $(X_j, Y)^T$, we apply probability integral transformations $Z_j = T_j(X_j)$ and $Z_{obs} = T(Y)$ using the fitted GPD above the chosen thresholds, and the empirical distribution below. These transformations result in the Z_j and Z_{obs} having common unit Fréchet marginal distributions, with distribution function $F(z) = \exp\{-z^{-1}\}$, $z > 0$.

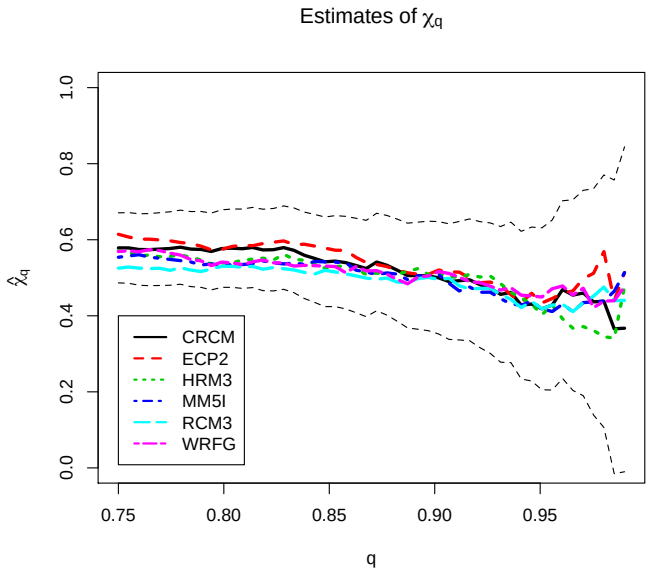

Figure 3.3: Estimates of $\chi_{j,q}$ (for Pacific winter precipitation) plotted against q for six RCMs ($j = 1, ..., 6$), with 95% confidence intervals for CRCM (thin dashed lines) added.

One can think of this transformation as analogous to computing a z-score to standardize a normal variate; however, the transformation here is to a very heavy-tailed distribution. The top left panel of Figure 3.4 shows a scatterplot of transformed precipitation amounts from the WRFG regional model and observations. Notice that the transformation to a heavy-tailed distribution visually magnifies the very largest precipitation amounts, while 'non-extreme' realizations cluster near the origin. Furthermore, many of the largest precipitation events from the climate model correspond to the largest events seen in the observations, which indicates that tail dependence is present. Compare this to the summer precipitation comparison in the lower left panel of the figure (discussed in Section 3.4), in which the very largest events fall near the axes of the scatterplot.

As both Z_j and Z_{obs} are heavy tailed, we assume that $(Z_j, Z_{obs})^T$ is a multivariate regular varying distribution. As described by Definition 2 in Chapter 1, this framework suggests a polar coordinate transformation, and for each bivariate pair $(Z_j, Z_{obs})^T$, we transform to polar coordinates under the L_1 norm by defining $R_j = Z_j + Z_{obs}$ and $W_j = Z_j/R_j$. One

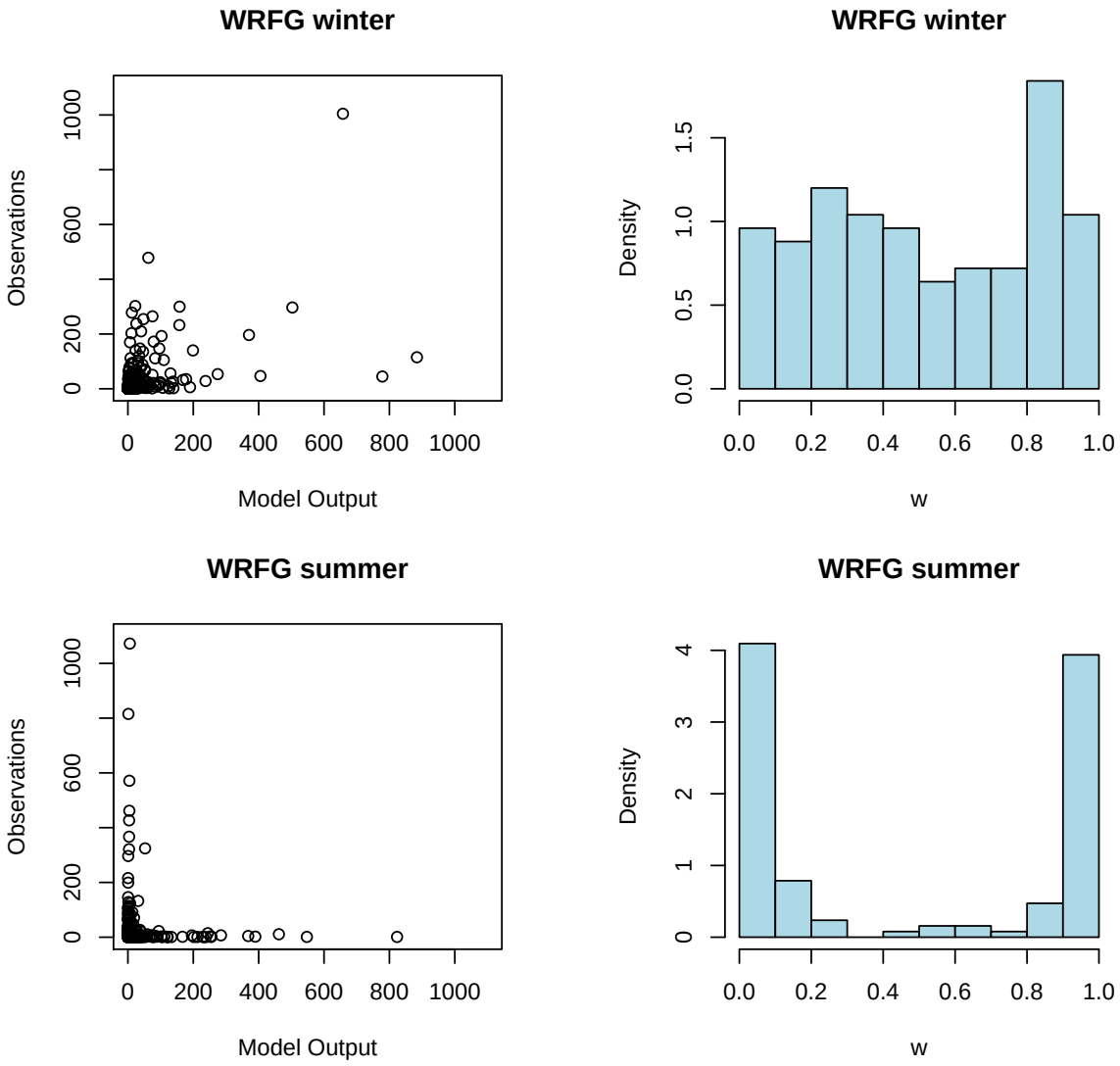

Figure 3.4: Fréchet-transformed winter precipitation amounts in Pacific region from WRFG regional climate model plotted against transformed observations (top left) and histogram of angular components for data points with large radial components (top right). Bottom row shows the same for summer precipitation over the Bukovsky prairie region.

can think of the 'radial' component R_j as measuring the overall size of a realization from $(Z_j, Z_{obs})^T$, and the 'angular' component W_j as measuring the relative contribution of Z_j (the climate model output) to the total size.

In statistical practice, one could fit dependence models for the angular component to realizations with radial component values exceeding r_0, a high empirical quantile. We choose r_{0j} to be the empirical 0.955 quantile of r_j values; thus 125 pairs $(z_{jt}, z_{obs,t})$ are used in each examination of tail dependence. The top right panel of Figure 3.4 shows a histogram of angular component values for these largest 125 radial component values from pairing WRFG output with observations. An angular component value near 0 indicates an extreme precipitation event occurred in the observational data product, but was not produced by the RCM. Conversely, a value near 1 results when the RCM produces an extreme event that was not seen in observations. The histogram shows many of the angular components falling on the interior of the interval $[0, 1]$, indicating relatively strong tail dependence between WRFG output and observations. Similar results are seen for each of the other five RCMs as well; histograms similar to those in Figure 3.4 are given in Figure 3.5.

As the NCEP reanalysis product is derived from past observed weather, the precipitation output of the NARCCAP RCMs forced by the reanalysis exhibits temporal correspondence (on a daily scale) with past observed weather, and we examined precipitation output from each RCM and an observational product. We found strong dependence between the observational quantity and the RCMs, as output from each regional model exhibits high correlation with observations. The upper tails of our precipitation quantity are estimated to have a finite endpoint in 5 of the 6 NARCCAP models, despite the observations being estimated to be slightly heavy-tailed. We further discovered that tail dependence (between each model and observations) is also present: the largest events produced by the RCMs exhibit some *correspondence* with the largest events in the observed record. We quantified the level of tail dependence using an empirical estimator of a tail dependence measure, and examined the dependence visually with histograms of angular component values. The results here indicate

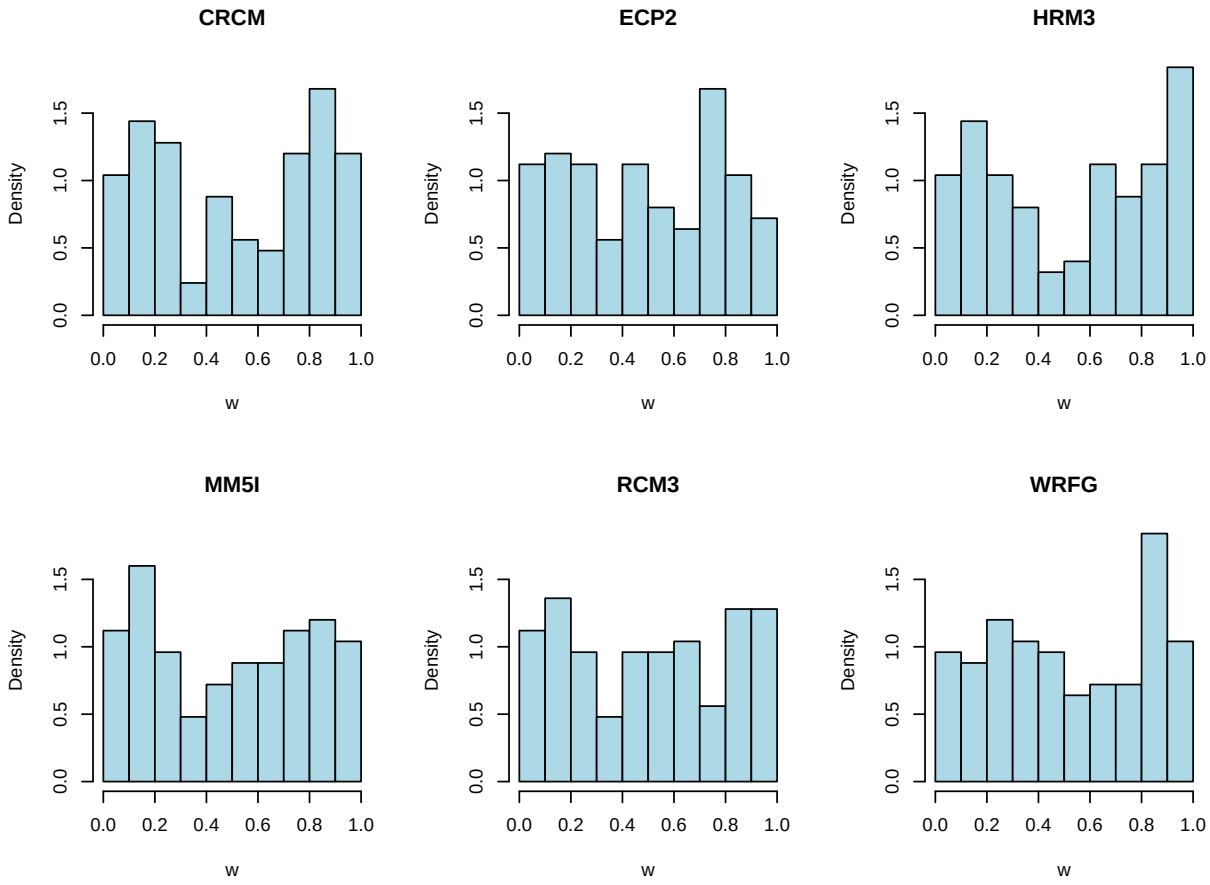

Figure 3.5: Histograms of angular component values for largest 125 values of radial components, for each RCM vs. observation comparison of Pacific region winter precipitation.

that each of the six RCMs are effective at downscaling extreme winter precipitation events in the Pacific region.

3.4 Tail Behavior of Prairie Region Summer Precipitation

In this section, we turn our attention to summer-season (JJA) precipitation over the prairie region defined in Bukovsky (2011). This region was chosen due to its relatively homogeneous terrain, as well as its central location within the NARCCAP domain. While the Pacific region studied in Section 3.3 was near the boundary of the NARCCAP simulations, this region is far from the boundary. Thus it is possible that the boundary conditions will have less influence on regional model precipitation output in this region than in the Pacific region. The prairie region stretches from approximately 38° to 49° N latitude and 86° to 98° W longitude, excluding the Great Lakes. Figure 3.6 shows the mean daily precipitation over this region from the NARCCAP ensemble and gridded observational product.

In addition to the fact that the study region chosen here is much farther from the boundary of the NARCCAP domain than the region studied in Section 3.3, there are other reasons to expect differences in the way the NARCCAP RCMs simulate summer extreme precipitation events in the prairie region. For one, the nature of winter-season extreme precipitation events in the Pacific region is much different than that of summer-season extremes over the region studied here. Winter precipitation over the west coast has been linked to global phenomena such as the El Nĩno/Southern Oscillation (Castello and Shelton, 2004; Ropelewski and Halpert, 1987) and is often associated with strong synoptic-scale features such as atmospheric rivers (Leung and Qian, 2009; Dettinger et al., 2011; Weller et al., 2012). Because these features generally exhibit themselves in large-scale global models (Dettinger, 2011) and coarse-resolution reanalyses, it may not be surprising that the regional models produce these events in similar ways when given boundary conditions from these sources. On the other hand, extreme precipitation events during the summer season in the prairie region are often associated with mesoscale convective systems (Fritsch et al., 1986; Trenberth et al.,

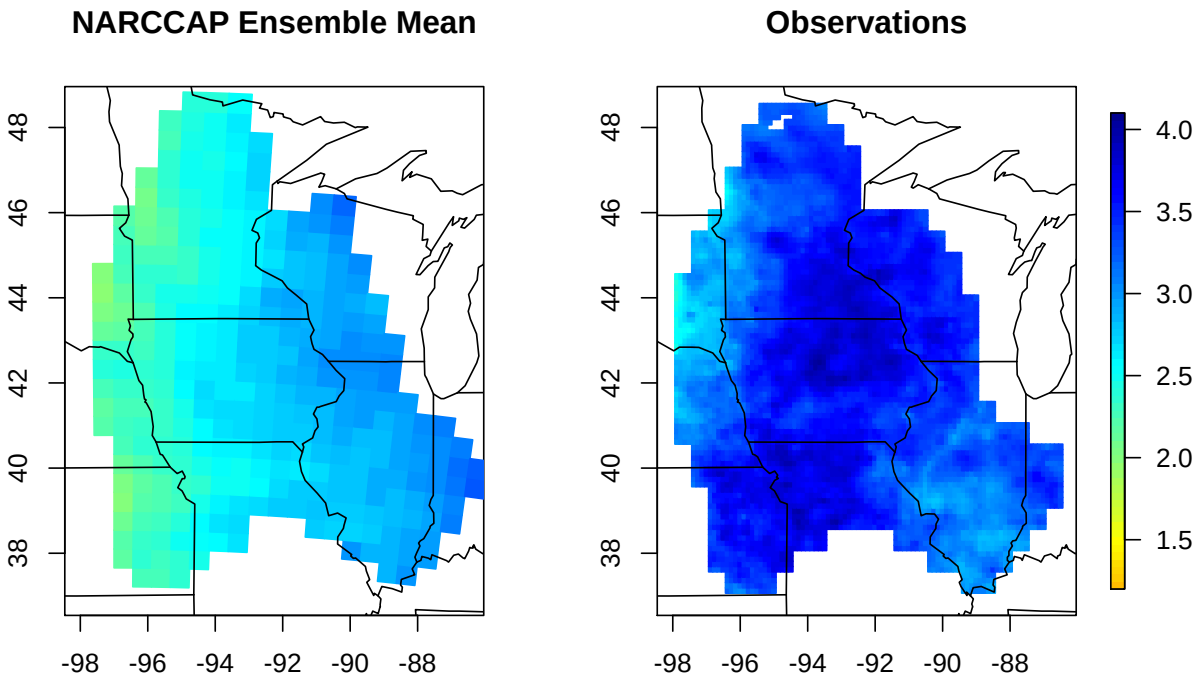

Figure 3.6: Mean JJA precipitation (mm/day) over prairie region from NARCCAP ensemble average (left) and gridded observations (right), 1981-2003. Figures are on a common scale indicated by the legend strip.

2003; Schumacher and Johnson, 2006) which are not well-resolved in coarse-resolution global climate models and even many finer-resolution regional climate model simulations. Convection is not only resolved poorly in many cases, with problems compounded by the use of convective parameterization, but the mesoscale processes that produce favorable conditions for convection are not always well resolved or well simulated either (see, e.g. Anderson et al., 2003; Bukovsky and Karoly, 2011; Davis et al., 2003; Cook et al., 2008). Thus, even given the same boundary conditions, it may not be surprising if each of the six regional models studied here simulate these types of events in their own unique manner.

Some previous examinations of precipitation from NARCCAP models also suggest that we may see differences in extreme summer events over the prairie region. While not directly studying extremes, Sain et al. (2011) used a functional analysis of variance approach to study variability in precipitation fields produced by regional and global model couplings in the NARCCAP experiment. Their work found that summer precipitation fields over the NARCCAP domain exhibited complex interactions between global and regional models, suggesting that different regional models may produce significantly different precipitation fields, even given the same boundary conditions. Schliep et al. (2010) found some differences in marginal behavior of summer-season extreme precipitation among NARCCAP RCMs driven by the NCEP reanalysis product. However, no previous analysis has directly examined the tail dependence in summer precipitation between NARCCAP regional models and observations.

3.4.1 Exploratory Analysis

As a way to begin to understand how both the RCMs and gridded observational product record precipitation over the prairie region, we produce maps of mean precipitation over the study period, as well as from a single day of heavy precipitation over the region. Figure 3.6 displays mean daily amounts from the gridded observational product for the years 1981-2003, as well as the NARCCAP ensemble average over the region over the same time period.

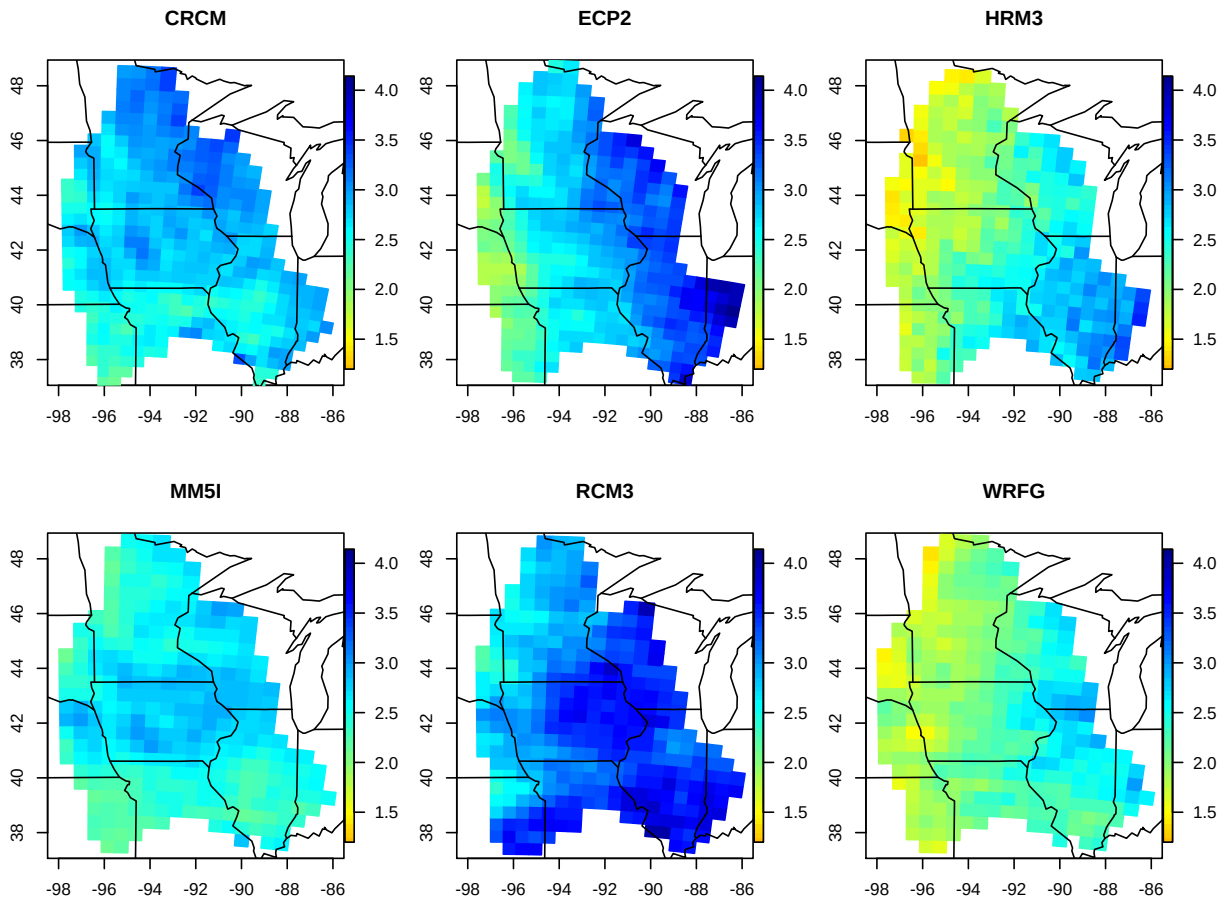

Figure 3.7: Mean JJA precipitation (mm/day) over prairie region from NARCCAP regional climate models driven by NCEP reanalysis, 1981-2003.

The map of observations shows a band of greatest precipitation running from southwest to northeast over the region. The driest part of this region is over the eastern Dakotas and western Minnesota. The average field from the NARCCAP models exhibits a pattern of increasing precipitation from west to east over the region.

In Figure 3.7 we plot the daily mean precipitation over the same region and time period from each of the six NARCCAP RCMs driven by the NCEP reanalysis. In the figure, one sees differences among the six models in both overall precipitation amounts and spatial precipitation patterns. Generally, the RCM3 and ECP2 models appear to be the wettest on average, while HRM3 and WRFG are the driest. While in general there is an increase in precipitation from west to east in each of the models, the spatial patterns also vary signifi-

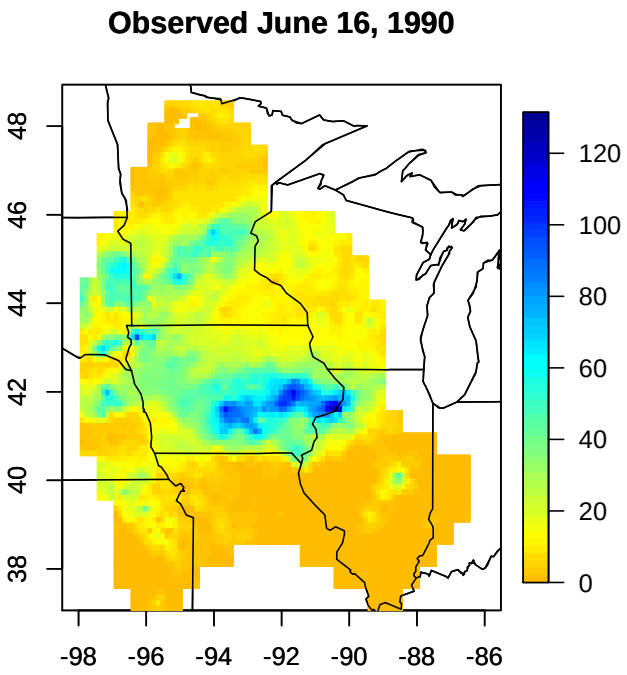

Figure 3.8: Observed precipitation field (mm) over prairie region on June 16, 1990, the day with the second-largest total rainfall over the study region.

cantly between regional models and the observational product. The RCM3 model appears to capture the spatial pattern seen in observations most accurately, with the exception of the southern Illinois region. The HRM3, RCM3, and ECP2 mean fields show fairly dramatic variability over the region, while CRCM and MM5I show less spatial variation.

As the primary interest here is in daily extreme precipitation events, we also examine the precipitation pattern on June 16, 1990, the day which saw the second-largest total precipitation amount over the study region in the observed record. Figure 3.8 shows the precipitation pattern over the region on this day from the gridded observational product. The figure indicates several very intense precipitation pockets over eastern Iowa, which caused severe flash flooding over the region (Barnes and Eash, 1994). Notice that the most intense precipitation amounts are highly concentrated over an area that would be covered by only a few RCM gridboxes.

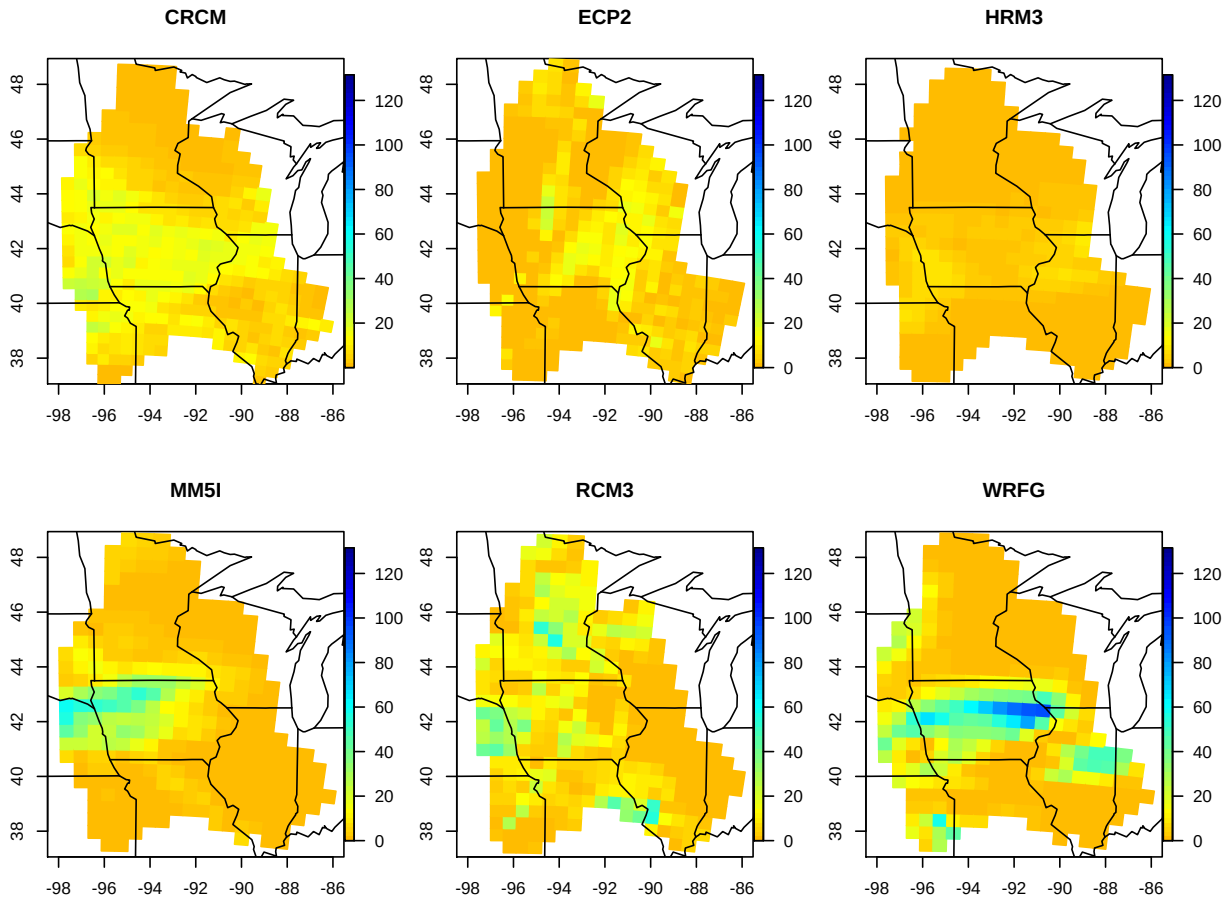

Figure 3.9: Precipitation fields (mm) for June 16, 1990 over the prairie region from each NARCCAP regional climate model driven by NCEP reanalysis.

In Figure 3.9 we plot the spatial precipitation fields for June 16, 1990 from each of the six NARCCAP models driven by the NCEP reanalysis. Generally speaking, Figure 3.9 shows that the regional models do not capture the extreme precipitation event very well. It appears that WRFG captures the event most adequately, simulating large precipitation amounts over eastern Iowa. The MM5I and RCM3 models do exhibit significant precipitation amounts; however, they fail to adequately capture the location of the observed rainfall event. The other three models do not exhibit large precipitation events; in fact, the HRM3 model indicates nearly zero rainfall over most of the region on this day. It is possible that some of the processes forcing this large precipitation event are not captured well in the NCEP reanalysis boundary conditions or translating well into the RCM domains. On the other hand, it is quite likely that the resolution of the NARCCAP regional models is still too coarse to capture such an event. In Section 3.4.2, we study the ability of each regional model to capture summer extreme precipitation events over the prairie region with a formal statistical analysis.

3.4.2 Analysis of Summer Precipitation Extremes

We now carry out an analysis similar to that in Section 3.3, focusing on summer precipitation extremes over the prairie region of North America. The aim here is to learn whether the largest summer precipitation events over the prairie region in the regional models correspond with the largest events in the observed record. We study daily precipitation from the summer months (JJA) for the years 1981-2003, for a total sample size of $T_s = 2116$ days.

In the prairie region, extreme precipitation events in summer are often associated with relatively small-scale convective systems; thus, following the exploratory analysis above, here we choose to examine a smaller 'footprint' than was chosen in Section 3.3.1. For each day $t = 1, ..., T_s$, we extract the sum of precipitation values over the 2×2 climate model grid box area (9×9 from the observational product) for which this sum is greatest over the prairie region. This quantity on day t from each regional climate model is denoted by X^s_{jt}

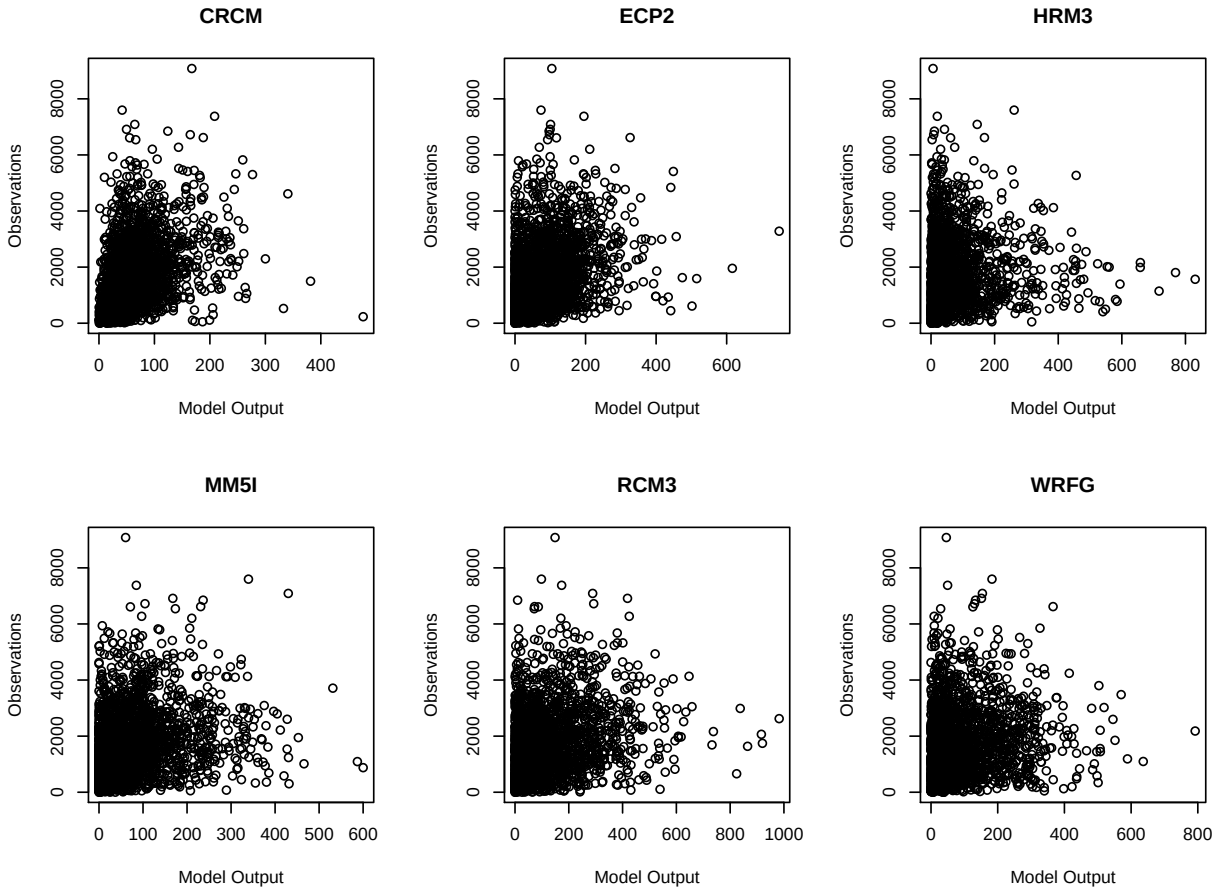

Figure 3.10: Modeled daily precipitation quantity X_j^s plotted against observed precipitation Y^s, for each regional climate model. Note the different scales due to differing spatial resolutions of the data sources.

for $j = 1, ..., 6$ and by Y_t^s from the observational product. A 4×4 grid box footprint was also examined, with similar results.

Figure 3.10 shows the resulting daily quantities from each regional model plotted against observations. Note that the dependence between output of each regional climate model and observations appears much weaker than in winter precipitation, shown in Figure 3.2. The visual differences between each of the six scatterplots are also more apparent than in Figure 3.2, suggesting larger variation between the regional models in summer precipitation over the prairie than seen in winter precipitation over the west coast. Spearman correlations between output of each regional climate model and observations are found to range from

0.162 (HRM3) to 0.408 (CRCM). Despite the positive correlations, it appears from Figure 3.10 that the largest events produced by the RCMs do not show much correspondence to the largest events from the observational product. Our primary interest is the joint upper tail of each bivariate distribution $(X_j^s, Y^s)^T$, and we proceed to examine both marginal and joint tail behavior.

As in Section 3.3.2, we fit a GPD to the exceedances of a high quantile in each margin separately. After checking diagnostics, we choose the 0.94 quantile, resulting in 127 exceedances used for each estimation procedure. A summary of maximum likelihood parameter estimates and scaled estimates of 20 and 50 year return values is shown in Table 3.3. We see greater variability between the six regional models in terms of marginal tail behavior of our computed precipitation quantity; tail parameter estimates range from -0.12 (MM5I) to 0.15 (WRFG). Given that X_j and Y were computed using relatively small spatial areas, it is perhaps surprising that the observational precipitation quantity is estimated to have a bounded tail. The effects of the varying parameter estimates are most clearly seen in extrapolation: the point estimates of the 50 year summer precipitation return value from each RCM cover a wide range. Due to the relatively large uncertainty in its estimation, we cannot conclude that significant differences in the tail parameter ξ exist between the six RCMs; however, less consistency among the RCMs is seen in these estimates for prairie region summer precipitation than was observed for Pacific winter precipitation.

Wehner (2013) found significant wet biases (compared to this observational product) in the 20 year return values for the eastern US (defined to be east of 100° W longitude) for all NARCCAP models except CRCM. The results in Table 3.3 are in agreement with these findings over the prairie region. The estimated 20 year return values from five of the NARCCAP models are significantly higher than the value estimated from the observational product. The exception is CRCM, which is in approximate agreement with this observational product.

Table 3.3: Threshold selected and maximum-likelihood parameter estimates (standard errors) from GPDs fit to exceedances of the 0.94 quantile of defined summer precipitation quantity (127 exceedances) from each source. Also shown are estimated 20 and 50 year return levels (mm) with 95% profile likelihood confidence intervals. Return values and confidence intervals have been normalized to gridbox-level values.

j	Model	u_j	$\hat{\psi}_j$ (se)	$\hat{\xi}_j$ (se)	$\hat{x}_{j,20}$	$\hat{x}_{j,50}$
1	CRCM	153	51.1(5.9)	−0.03(0.07)	94.2(84.2, 116.8)	104.2(91.2, 139.0)
2	ECP2	220	72.7(9.3)	0.06(0.09)	153.0(130.8, 202.2)	175.3(144.1, 267.2)
3	HRM3	230	142.7(18.1)	−0.10(0.09)	191.9(169.1, 249.8)	211.6(181.9, 299.7)
4	MM5I	237	85.9(10.0)	−0.12(0.08)	136.8(124.9, 163.9)	147.5(132.7, 187.2)
5	RCM3	364	108.2(14.5)	0.07(0.10)	240.5(204.6, 331.8)	275.4(223.7, 429.4)
6	WRFG	280	67.8(10.1)	0.15(0.12)	184.6(151.4, 280.0)	217.5(166.8, 391.7)
-	(Obs)	3939	964.3(119.1)	−0.05(0.09)	99.0(89.8, 121.8)	107.7(95.4, 142.8)

In terms of the marginal distribution of summer precipitation over the prairie region, we see varying tail behaviors produced by the six NARCCAP regional climate models. The observational product produces a distribution which is estimated to have a finite upper endpoint. This is also seen in three of the NARCCAP models (CRCM, HRM3, and MM5I), while the other three produce heavy-tailed estimates for precipitation. We also see relatively large variability in the empirical 0.94 quantile level of precipitation from each model, suggesting that some models are generally wetter than others (see also Figure 3.7). We now turn to an examination tail dependence; that is, we aim to learn if the largest summer precipitation events observed in the prairie region exhibit any daily correspondence to the largest events produced by the NARCCAP RCMs.

As an exploration of the level of tail dependence in summer precipitation amounts between climate model output and observations, we again employ the Coles et al. (1999) estimator of χ as defined in (14). Figure 3.11 shows empirical estimates of $\chi_{j,q}$ in (15) for $q \in [0.75, 0.97)$ and $j = 1, ..., 6$. In contrast to Figure 3.3, which shows estimates of $\chi_{j,q}$ near 0.5 for large q, Figure 3.11 shows this tail dependence metric decreasing toward zero for all six regional models. This provides strong evidence that asymptotic independence is present here; that is, there is little to no correspondence between the largest summer precipitation events over the region produced by the RCMs, and those seen in observations. We

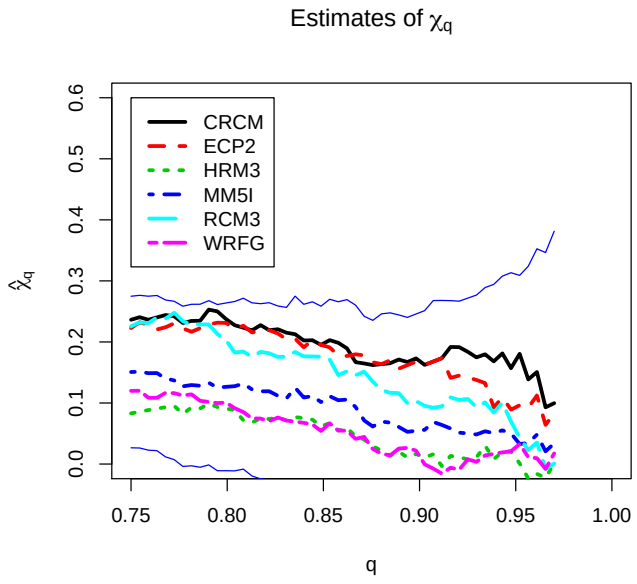

Figure 3.11: Estimates of $\chi_{j,q}$ for summer precipitation plotted against q for six RCMs ($j = 1, ..., 6$). 95% confidence bands (thin solid lines) added for MM5I model.

also add 95% confidence bands for the MM5I comparison with observations to the plot to illustrate the uncertainty in these estimates. Note that all six estimates of χ_q are within these confidence bands for all $q \in [0.75, 0.97)$.

One can further see evidence for asymptotic independence by applying transformations to marginal distributions so that each is unit Fréchet, and examine the angular component values for those points with large radial components (see Section 3.3.3). The lower left panel of Figure 3.4 shows a scatterplot of Fréchet-transformed summer precipitation amounts over the region from the WRFG RCM plotted against observations. Note that nearly all points with large values in either margin are near the axes of the plot. Compare this with winter precipitation in the top left panel, which contains many large points on the interior of the quadrant. The lower right panel of Figure 3.4 shows the corresponding plot of angular components for the largest 127 sample values in terms of the radial component, for the WRFG-observation comparison. In contrast to the upper right panel, the histogram here is heavily U-shaped, indicating a lack of correspondence of large precipitation events in

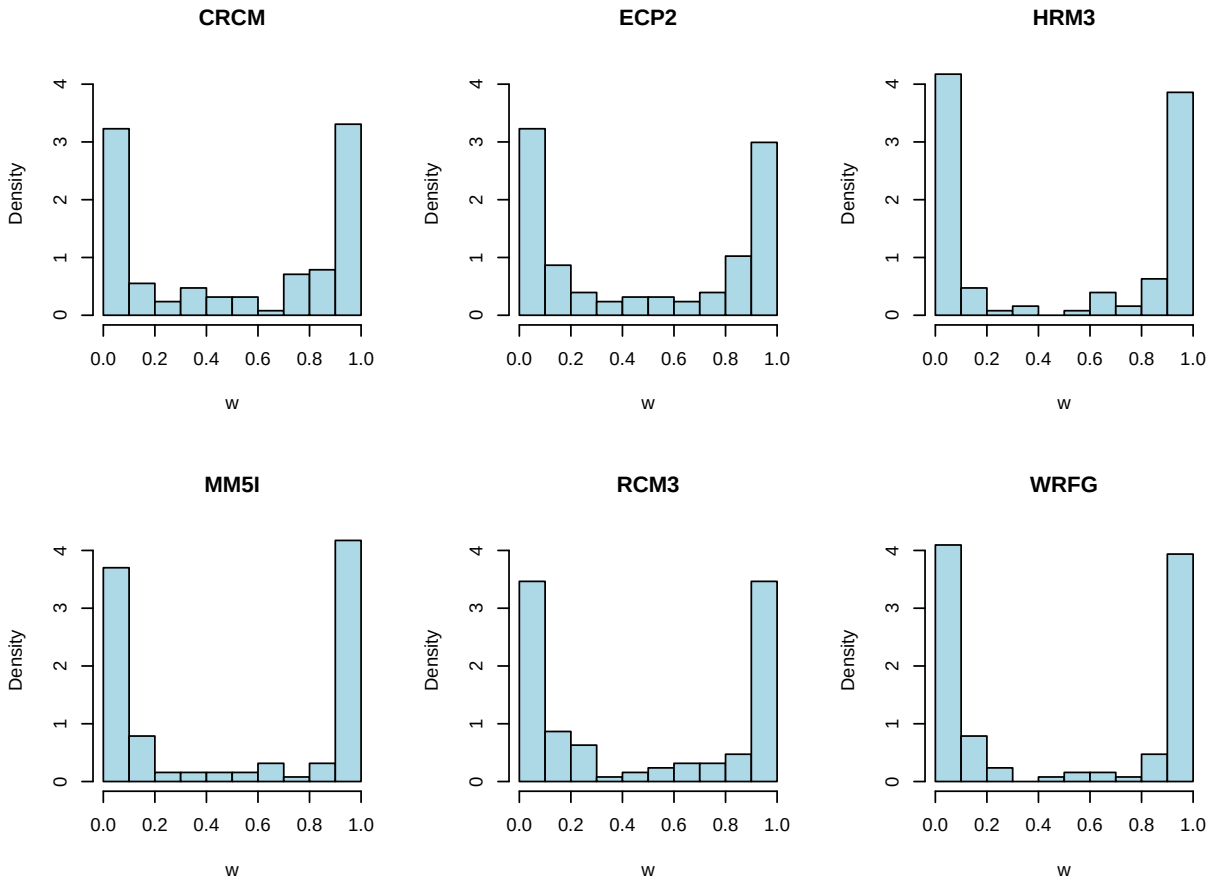

Figure 3.12: Histograms of angular component values for largest 127 values of radial components, for each RCM vs. observation comparison of prairie region summer precipitation.

observations with large precipitation events from WRFG. Similar results are seen for each of the six NARCCAP models, indicating asymptotic independence of summer precipitation amounts produced by the NARCCAP models and those seen in observations. Histograms of angular components for all six RCM-observation comparisons are shown in Figure 3.12.

While it was found that prairie region summer precipitation amounts seen in observations are asymptotically independent of those amounts produced by the NARCCAP models, the empirical Spearman correlations between each model and observations suggest that there is some positive dependence in each bivariate pair $(X_j^s, Y^s)^T$. From an extremes perspective, a conclusion of asymptotic independence drawn from data does not tell the whole story

about dependence at observable, sub-asymptotic levels. For example, a bivariate Gaussian distribution with correlation $\rho = 0.9$ possesses asymptotic independence, but still exhibits tail dependence at finite levels. For practical application, it is useful to address the strength of this 'second-order' dependence in the distribution's tail.

Coles et al. (1999) introduce a metric to quantify the strength of this second-order dependence. For a bivariate random vector $(X_1, X_2)^T$ with common marginal distributions, they define

$$\bar{\chi}_q = \frac{2 \log \mathbb{P}(X_1 > x_q)}{\log \mathbb{P}(X_1 > x_q, X_2 > x_q)} - 1,$$

where x_q is the q quantile level of X_1 and X_2. For $q \in [0, 1]$, $\bar{\chi}_q \in (-1, 1]$ serves as a measure of the level of tail dependence in the asymptotic independence setting. Specifically of interest is the limiting behavior, and Coles et al. (1999) define

$$\bar{\chi} = \lim_{q \to 1} \bar{\chi}_q.$$

As an example, if $(X_1, X_2)^T$ follow a bivariate Gaussian distribution with correlation $\rho \in (-1, 1)$, it follows that $\bar{\chi} = \rho$.

Here we apply empirical estimates of $\bar{\chi}_q$ to our samples from $(X_j^s, Y^s)^T$ for $j = 1, ..., 6$. Figure 3.13 shows these estimates plotted against $q \in [0.75, 0.97)$. With the exception of the CRCM and ECP2 regional models, the estimates of $\bar{\chi}$ are all near zero. For perspective, a situation in which X_j^s and Y^s were exactly independent would correspond to $\bar{\chi} = 0$. Figure 3.13 suggests, then, for four of the six NARCCAP models, the extreme summer precipitation events in the prairie region produced by the RCMs occur nearly independently of the observed extremes. Some sub-extremal dependence is seen between observations and the CRCM and ECP2 models, although it is rather weak. That the CRCM and ECP2 models exhibit stronger dependence to observations than the other four models is not surprising in

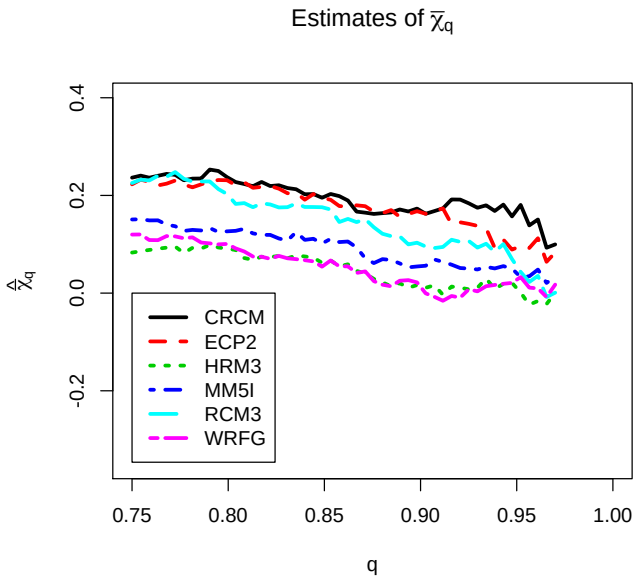

Figure 3.13: Estimates of $\bar{\chi}_{j,q}$ for summer precipitation plotted against q for six RCMs $(j = 1, ..., 6)$.

light of the fact that these two models employ nudging techniques that force the RCM to more closely follow the reanalysis product in the interior of their domains and not just at their boundaries.

In contrast to the examination of Pacific region winter precipitation in Section 3.3, we have found that extreme summer precipitation events over the prairie region produced by the NARCCAP RCMs forced by the NCEP reanalysis show little to no correspondence with such events seen in the observed record. This may have a number of causes, including the still-too-coarse resolution of the NARCCAP RCMs for these convective events and some of their driving processes, errors related to the use of convective parameterization, potential biases in the NCEP reanalysis being inherited by RCMs, resolved-scale forcing for the events not occurring over the oceans where the RCM boundaries are in the NCEP reanalysis, resolved-scale forcing from the NCEP reanalysis not translating well to the center of the large RCM domains, and errors in the timing and/or placement of events in the RCMs that would place

them later and not within the daily totals as defined in the observations and/or not within the prairie region.

3.5 Summary and Discussion

This study examines two case studies on the ability of NARCCAP RCMs, driven by the NCEP reanalysis, to reproduce past observed extreme precipitation events. The NARCCAP models are able to simulate past observed winter precipitation events over the west coast region of North America reasonably well, and we studied the tail dependence between daily RCM output precipitation through techniques from statistical extreme value theory. The results lend confidence to the idea that the RCMs, when provided boundary conditions that are conducive to an extreme precipitation event over this region in winter, can downscale the synoptic-scale conditions appropriately and exhibit this event in their precipitation output. When such conditions are present in a future-scenario run of a global model, the RCM can then produce the extreme event in its simulation from the boundary conditions it is given. The nature of winter west coast precipitation, which is often driven by synoptic-scale processes and orographic features, lends itself to adequate simulation by the RCMs.

The results for summer precipitation over the prairie region in the central US are quite different. Greater differences in marginal tail behavior of precipitation are seen between the six NARCCAP models, as compared to simulated west coast winter precipitation. Additionally, tail dependence in daily precipitation between model output and observations is virtually non-existent for prairie region summer precipitation. Further analysis indicates that most of the NARCCAP RCMs produce daily precipitation extremes occurring nearly independently of extreme precipitation events in the observed record. This is likely due to the nature of summer-season extreme precipitation events over the prairie region: most are the result of regional to local-scale convective storms. It may be that the forcing for these storms are not captured in the NCEP reanalysis boundary conditions for the NARCCAP RCMs, and/or that the RCMs themselves may not be able to simulate these events ade-

quately. When attempting to project summer precipitation extremes over this region from future-scenario climate model runs, this deficiency should be acknowledged.

Despite studying different quantities and employing different statistical techniques, the results of marginal analyses conducted here are consistent with the findings of Wehner (2013). Twenty year return values for winter west coast precipitation as seen in five of six NARCCAP models were greater than that estimated from the observational product. The same is seen for summer precipitation return values in the prairie region, with the CRCM model as the exception in each case. While the marginal analyses result in comparison of return values from each source, the novelty of this work is the examination of tail dependence, which examines daily correspondence of extremes from NARCCAP output and observations. Here a stark contrast is found between the two regions and seasons: the models are able to reproduce observed winter extreme precipitation events on the west coast, but not summer precipitation extremes in the prairie region.

While the work here provides an exploration of the ability of NARCCAP RCMs to simulate extreme precipitation events over two different regions in different seasons, the increased emphasis on potential changes in extreme weather events under climate change motivates further examination of the ability of RCMs to simulate such events. Future studies may apply the techniques used here to study precipitation extremes over different regions and further examine uncertainties in climate model representations of extreme precipitation events.

CHAPTER 4

A SUM CHARACTERIZATION OF HIDDEN REGULAR VARIATION

4.1 Introduction

As described in Chapter 1, fundamental to the multivariate regular variation framework is a polar coordinate decomposition which describes joint tail behavior in terms of a radial component and an angular component, which become independent as the radial component becomes large. Dependence in the joint tail of the random vector is described by a probability measure governing the angular component. Over the past 15 years, it has been recognized that modeling approaches based on this framework may not adequately describe some tail dependence structures. As a canonical example, the bivariate Gaussian distribution $(X_1, X_2)^T$ with correlation $\rho < 1$ can be shown (Sibuya, 1960) to be *asymptotically independent*; that is,

$$\lim_{x \to \infty} \mathbb{P}(X_2 > x \mid X_1 > x) = 0.$$

In this case, a modeling approach based on the first-order limiting measure under the regular variation framework does not capture tail dependence induced by the correlation ρ (Ledford and Tawn, 1996). The fundamental shortcoming is that the first-order limit measure is degenerate on some joint tail regions, thus masking possible dependence structure at finite levels.

The concept of hidden regular variation (Resnick, 2002) offers a mathematical structure for describing sub-asymptotic tail dependence. Hidden regular variation is essentially a second-order formulation of regular variation on regions where the first-order limit is degen-

erate. More treatment is given in Maulik and Resnick (2004), Heffernan and Resnick (2007), and Mitra and Resnick (2010). More recently, de Haan and Zhou (2011) offered characterizations of hidden regular variation based on an alternative polar coordinate transformation.

In this chapter, we offer a characterization of a random vector with hidden regular variation as the sum of independent first- and second-order components. We demonstrate that our characterization exhibits useful finite-sample properties, thus lending itself to inference. This characterization is asymptotically justified via the concept of multivariate tail equivalence (Maulik and Resnick, 2004). We demonstrate the characterization through simulation.

The remainder of this chapter is structured as follows: in Section 4.2 we provide background on the concept of hidden regular variation. Through two examples, Section 4.3 illustrates possible hidden angular measure structures. Previous characterizations of hidden regular variation which have appeared in the literature are reviewed in Section 4.4. In Section 4.5 we describe the construction of our characterization and show that it is tail equivalent to a random vector with hidden regular variation. Section 4.6 demonstrates simulation from our representation for a random vector with Gaussian dependence structure and compares it to other representations. We conclude in Section 4.7 with a summary and discussion. Portions of this chapter also appear in a Colorado State University Department of Statistics technical report (Weller and Cooley, 2012), as well as in Weller and Cooley (2013), which has been submitted for publication.

4.2 Hidden Regular Variation

Recall Definition 1 of multivariate regular variation in Chapter 1: we say that a random vector $\mathbf{Z}$ taking values in a subset of $[0, \infty)^d$ is regular varying with finite limiting measure

$\nu \neq 0$ if there exists a function $b(t) \to \infty$ such that

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{b(t)} \in \cdot\right] \xrightarrow{v} \nu(\cdot) \tag{16}$$

in $M_+(\mathfrak{C})$ as $t \to \infty$.

Definition 2 introduces the polar coordinate transformation. Fix a norm $\|\cdot\|$ on $\mathfrak{C}$, and consider the unit sphere $\mathcal{N} = \{\mathbf{z} \in \mathfrak{C} : \|\mathbf{z}\| = 1\}$. Define the bijective transformation $T : \mathfrak{C} \to (0, \infty] \times \mathcal{N}$ via $T(\mathbf{z}) = (\|\mathbf{z}\|, \mathbf{z}\|\mathbf{z}\|^{-1}) = (r, \mathbf{w})$. The measure ν can then be expressed in terms of the new coordinate system via $\nu = \nu_\alpha \times H$, where ν_α is a Pareto measure and H is a non-negative measure on $\mathcal{N}$. The measure H is called the *angular measure*. By appropriate choice of normalizing function $b(t)$, H can be made to be a probability measure on $\mathcal{N}$.

It is possible that the limiting measure ν in (16) places zero mass on pie slice-shaped regions $\{\mathbf{z} \in \mathfrak{C} : \mathbf{z}\|\mathbf{z}\|^{-1} \in B \subset \mathcal{N}\}$ of the cone $\mathfrak{C}$. In such cases, the normalizing function $b(t)$ obliterates any finer structure of the random variable on such regions, if such a finer structure exists. The angular measure H thus places zero mass on corresponding regions of the unit sphere $\mathcal{N}$. This prompted Resnick (2002) to formulate the concept of hidden regular variation.

Definition 4. *Consider a subcone $\mathfrak{C}_0 \subset \mathfrak{C}$ with $\nu(\mathfrak{C}_0) = 0$. A random vector $\mathbf{Z}$ is said to possess hidden regular variation if, in addition to* (16)*, there exists a non-decreasing function $b_0(t) \to \infty$ with $b(t)/b_0(t) \to \infty$ and nonnegative Radon measure ν_0 such that*

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{b_0(t)} \in \cdot\right] \xrightarrow{v} \nu_0(\cdot) \tag{17}$$

as $t \to \infty$ in $M_+(\mathfrak{C}_0)$.

The measure ν_0 is homogeneous with tail index α_0; that is, for any measurable set $A \subset \mathfrak{C}_0$, $\nu_0(tA) = t^{-\alpha_0}\nu_0(A)$. A polar decomposition of ν_0 arises; the measure can be written as a product of Pareto measure ν_{α_0} and positive Radon measure H_0 on $\mathcal{N}_0 = \mathcal{N} \cap \mathfrak{C}_0$. The function

$b_0(t) \in RV_{1/\alpha_0}$, with $\alpha_0 \geq \alpha$; thus, $\mathbf{Z}$ has a lighter tail on $\mathfrak{C}_0$ than on $\mathfrak{C}$. As $\mathcal{N}_0$ may not be a relatively compact set of $\mathcal{N}$, H_0 (called the *hidden angular measure*) may be either finite or infinite; see Section 4.3.

In the bivariate case when asymptotic independence is present, the first-order limit measure ν concentrates fully on the axes of the cone $\mathfrak{C}$. If hidden regular variation exists, it can be formulated on the subcone $\mathfrak{C}_0 = (0, \infty]^2$: the first quadrant with the axes removed (Resnick, 2002).

Figure 4.1 provides an illustration of hidden regular variation in the two dimensional case. This figure plots realizations of sizes $n = 1000, 2500, 5000$ of a bivariate Gaussian random vector with correlation $\rho = 0.75$, after marginal transformation to unit Fréchet and normalization by n. Also shown is a histogram of angular component values for normalized points exceeding a fixed radial component threshold. Notice as the sample size increases, the histograms become more heavily U-shaped, and the angular measure H degenerates to point masses on the endpoints of $\mathcal{N}$, indicating asymptotic independence. However, for any finite sample of n realizations of this random vector, tail dependence exists, as it is induced by the correlation ρ. In this example, the shortcoming of the regular variation framework is the inability to distinguish asymptotic independence from independence in the usual sense (e.g., $\rho = 0$). Hidden regular variation captures tail dependence induced by ρ; we explore this example more thoroughly in Section 4.6.

4.3 Finite and Infinite Hidden Measures

Definition 2 of multivariate regular variation in Chapter 1 guarantees finiteness of the limit measure ν, as the associated angular measure H can be made to be a probability measure. The subcone $\mathfrak{C}_0$ on which hidden regular variation exists need not be a relatively compact subset of $\mathfrak{C}$, and thus while the hidden measure ν_0 is Radon (assigning finite measure to relatively compact sets), it may be infinite on the entire subcone $\mathfrak{C}_0$. The corresponding hidden angular measure H_0 may thus be infinite on $\mathcal{N}_0 \subset \mathcal{N}$.

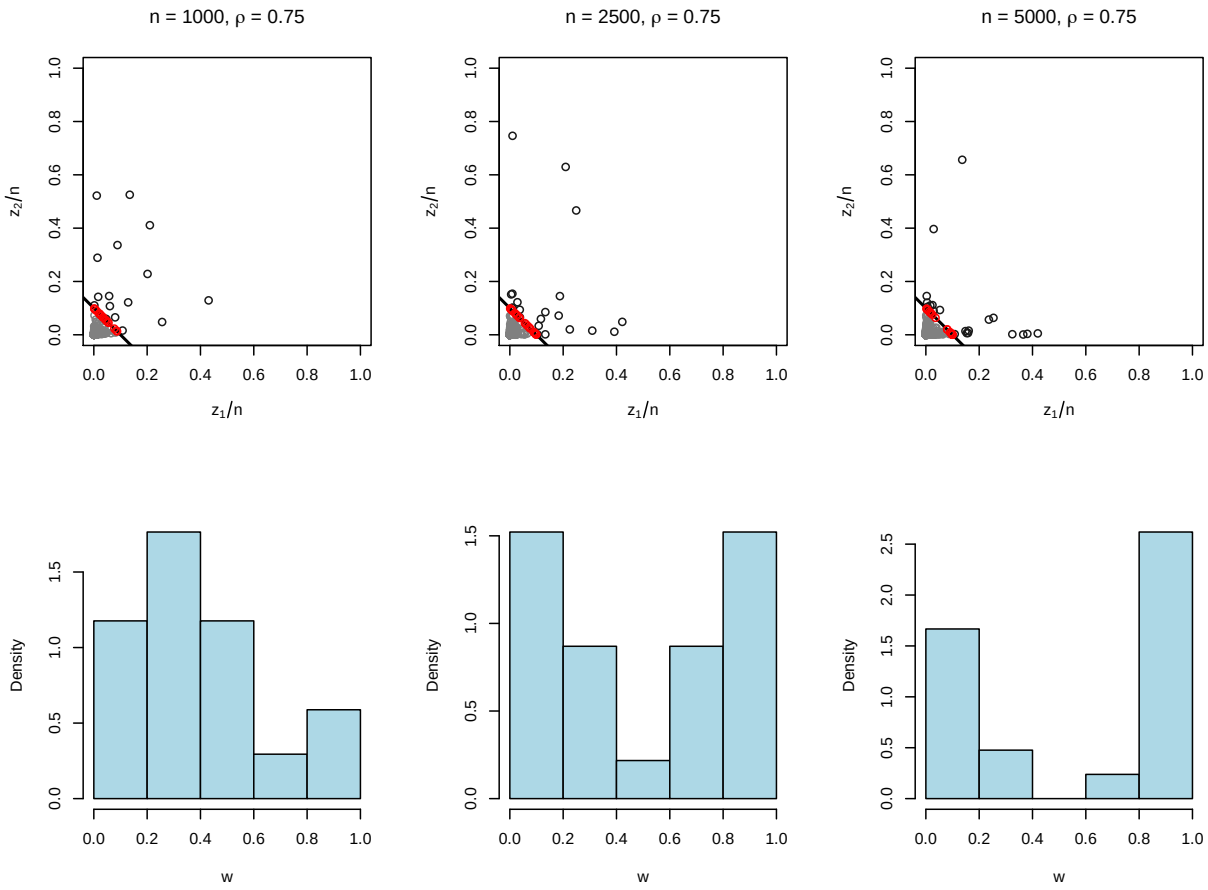

Figure 4.1: Realizations of a bivariate Gaussian random vector with correlation $\rho = 0.75$, plotted after transformation to Fréchet scale and normalization by n. Bottom row gives histograms of angular component values for realizations with radial component values exceeding $r_0 = 0.15$.

4.3.1 Examples

We present two examples of hidden regular variation: one with finite hidden measure and one for which the hidden measure is infinite. More examples are provided in Resnick (2002), Maulik and Resnick (2004), and Mitra and Resnick (2010).

Example 4.1. Let $\mathbf{X} \in \mathbb{R}^3$ have distribution function $F(\mathbf{x}) = \exp\{-(x_1^{-1/\beta} + x_2^{-1/\beta} + x_3^{-1/\beta})^\beta\}$, for $\beta \in (0,1)$; that is, $\mathbf{X}$ has unit Fréchet marginal distributions and trivariate logistic dependence (Gumbel, 1960). Let $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$, where $Y_j, j = 1,2,3$ are independent Pareto random variables with distribution function $F(y) = y^{-1}$ for $y > 1$. Let B be a Bernoulli(0.5) random variable and define

$$\mathbf{Z} = B\mathbf{X}^{1/\alpha_0} + (1-B)\mathbf{Y}$$

for $\alpha_0 \in (1,2)$.

It is straightforward to show that

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{t} \in \cdot\right] \xrightarrow{v} \nu(\cdot)$$

in $M_+(\mathfrak{C})$ as $t \to \infty$, where ν places all mass on the axes $\mathbf{L} = \{\mathbf{z} \in \mathfrak{C} : z_{(2)} = 0\}$, where $z_{(2)}$ is the second-largest component of $\mathbf{z}$. Consider the subcone $\mathfrak{C}_0 = \mathfrak{C} \setminus \mathbf{L}$. One can show that in $M_+(\mathfrak{C}_0)$,

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{t^{1/\alpha_0}} \in \cdot\right] \xrightarrow{v} \nu_0(\cdot),$$

where the form of ν_0 is given by

$$\nu_0([\mathbf{0}, \mathbf{z}]^c \cap \mathfrak{C}_0) = (z_1^{-\alpha_0/\beta} + z_2^{-\alpha_0/\beta} + z_3^{-\alpha_0/\beta})^\beta.$$

Thus $\mathbf{Z}$ is regular varying on $\mathfrak{C}$ with tail index $\alpha = 1$ and exhibits hidden regular variation on $\mathfrak{C}_0$ with tail index α_0. The finiteness of the hidden angular measure ν_0 is seen in the fact that it can be extended to the axes of the cone. Specifically, for this example $\nu_0([\mathbf{0}, \mathbf{z}]^c \cap \mathfrak{C}_0) = \nu_0([\mathbf{0}, \mathbf{z}]^c)$ because the logistic angular dependence structure concentrates fully on the interior $\{\mathbf{z} \in \mathfrak{C} : z_1, z_2, z_3 > 0\}$ of the cone $\mathfrak{C}$ (Coles and Tawn, 1991).

Example 4.2. (Resnick, 2002) Define $\mathbf{Z} = (Z_1, Z_2)^T$ where $Z_j, j = 1, 2$ are independent standard Pareto random variables. Then in $M_+(\mathfrak{C})$

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{t} \in \cdot\right] \xrightarrow{v} \nu(\cdot),$$

where ν concentrates on the axes; that is, $\mathbf{Z}$ exhibits asymptotic independence. On $\mathfrak{C}_0 = (0, \infty]^2$, it suffices to consider sets of the form $(\mathbf{z}, \boldsymbol{\infty}]$ for $\mathbf{z} = (z_1, z_2)^T \in \mathfrak{C}_0$, and we have

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{t^{1/2}} \in (\mathbf{z}, \boldsymbol{\infty}]\right] \longrightarrow (z_1 z_2)^{-1}$$

as $t \to \infty$. The hidden measure in this case cannot be extended to the axes, however: consider the set $A = \{\mathbf{z} \in \mathfrak{C}_0 : \|\mathbf{z}\| > 1\}$. For any $\delta > 0$,

$$\nu_0(A) \geq \nu_0([(2, \delta), \boldsymbol{\infty}]) = (2\delta)^{-1} \to \infty$$

as $\delta \to 0$. Despite being finite on any relatively compact set of $\mathfrak{C}_0$, the hidden measure ν_0 is infinite on $\mathfrak{C}_0$ because it diverges near the axes of the cone.

4.3.2 Alternative Coordinate Transformations when H_0 is Infinite

An infinite ν_0 on $\mathfrak{C}_0$ admits an infinite hidden angular measure; that is, $H_0(\mathcal{N}_0) = \infty$. Several authors have proposed alternative coordinate transformations to represent infinite hidden measures. We examine two such alternatives here, focusing on the bivariate case; generalizations to higher dimensions are made in the corresponding references.

Mitra and Resnick (2010) propose a transformation $\tilde{T} : \mathfrak{C}_0 \mapsto (0, \infty] \times \tilde{\mathcal{N}}$, where $\tilde{\mathcal{N}} = \{\mathbf{z} \in \mathfrak{C}_0 : z_1 \wedge z_2 = 1\}$. This transformation is given by

$$\tilde{T}(\mathbf{z}) = \left(z_1 \wedge z_2, \frac{\mathbf{z}}{z_1 \wedge z_2} \right) = (\tilde{r}, \tilde{\mathbf{w}}).$$

Mitra and Resnick (2010) show that the hidden measure ν_0 decomposes into the product of a Pareto measure on $(0, \infty]$ and finite measure on $\tilde{\mathcal{N}}$.

We demonstrate the Mitra and Resnick (2010) transformation on Example 4.2 above. The density of the limit measure ν_0 in Cartesian coordinates is

$$\nu_0(d\mathbf{z}) = (z_1 z_2)^{-2}.$$

Notice that this density is symmetric about the ray given by $\{\mathbf{z} \in \mathfrak{C}_0 : z_1 = z_2\}$, so first consider the case $z_1 > z_2$. The coordinate decomposition is given by $(\tilde{r}, \tilde{w}) = (z_2, z_1/z_2)$. Applying a simple change-of-variable argument to this case and the case $z_2 > z_1$ gives

$$\nu_0(d\tilde{r} \times d\tilde{w}) = 2\tilde{r}^{-3} \times \frac{1}{2} \left(\tilde{w}^{-2} I_{\{\tilde{w} = \frac{z_1}{z_2}\}} + \tilde{w}^{-2} I_{\{\tilde{w} = \frac{z_2}{z_1}\}} \right).$$

Hence, under this transformation ν_0 decomposes into the product of a Pareto measure and a mixture of Pareto measures.

A different transformation is examined by de Haan and Zhou (2011). These authors consider a two-dimensional random vector $\mathbf{Z}$ which is regular varying of index $\alpha = 1$ and exhibits hidden regular variation of index $\alpha_0 \in (1, 2)$. de Haan and Zhou (2011) first normalize the random vector by $\mathbf{Z}^{\alpha_0}$ and then define the transformation $T^* : \mathfrak{C}_0 \mapsto (0, \infty) \times (0, 1)$ via

$$T^*(z_1, z_2) = \left(\{z_1^{-1} + z_2^{-1}\}^{-1}, \frac{\{z_1^{-1} + z_2^{-1}\}^{-1}}{z_1} \right) = (r^*, w^*).$$

The limiting measure of the normalized $\mathbf{Z}^{\alpha_0}$ then decomposes into the product of a Pareto measure of tail index 1 on $(0,\infty)$ and finite measure H^* on $(0,1)$. An example is presented in Section 4.4.3.

4.4 Previous Characterizations of Hidden Regular Variation

Consider a random vector $\mathbf{Z}$ with support on $[0,\infty)^d$ which is multivariate regular varying on $\mathfrak{C}$ with limit measure ν as in (16). Further assume that $\mathbf{Z}$ exhibits hidden regular variation of index α_0 on a subcone $\mathfrak{C}_0 \subset \mathfrak{C}$ as in (17). Two previous works have developed probabilistic characterizations of the joint tail of $\mathbf{Z}$. Maulik and Resnick (2004) introduce a mixture characterization when the hidden measure ν_0 is finite. In the two-dimensional asymptotic independence case, de Haan and Zhou (2011) construct a characterization when the tail index of the hidden measure $\alpha_0 \in (1,2)$. Before presenting these characterizations, we review the concept of multivariate tail equivalence.

4.4.1 Multivariate Tail Equivalence

Multivariate tail equivalence was introduced in Maulik and Resnick (2004).

Definition 5. *Consider random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *taking values in* $[\mathbf{0},\boldsymbol{\infty})$ *with distribution functions* F *and* G*, respectively. The random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *are said to be tail equivalent on the cone* $\mathfrak{C}^* \subseteq \mathfrak{C}$ *if there exists a scaling function* $b^*(t) \to \infty$ *such that*

$$t\mathbb{P}\left[\frac{\mathbf{X}}{b^*(t)} \in \cdot\right] \xrightarrow{v} \nu_*(\cdot) \quad \text{and} \quad t\mathbb{P}\left[\frac{\mathbf{Y}}{b^*(t)} \in \cdot\right] \xrightarrow{v} c\nu_*(\cdot) \tag{18}$$

as $t \to \infty$ *in* $M_+(\mathfrak{C}^*)$*, for some constant* $c \in (0,\infty)$ *and measure* ν_* *on* $\mathfrak{C}^*$*.*

The definition (18) implies that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ have the same asymptotic tail properties on $\mathfrak{C}^*$, up to a scaling constant. Following Maulik and Resnick (2004) we

write

$$\mathbf{X} \overset{\text{te}(\mathfrak{C}^*)}{\sim} \mathbf{Y}.$$

Maulik and Resnick (2004) and de Haan and Zhou (2011) develop characterizations which are tail equivalent to $\mathbf{Z}$ on both $\mathfrak{C}$ and $\mathfrak{C}_0$ in special cases.

4.4.2 Mixture Characterization when ν_0 is Finite

When the hidden measure ν_0 is finite, Maulik and Resnick (2004) introduce a mixture characterization of $\mathbf{Z}$. Without loss of generality, assume $b(t)$ in (16) is such that H is a probability measure on $\mathcal{N}$. Define the random vector $\mathbf{X} = R\mathbf{W}$, where $\mathbb{P}(R > r) = 1/b^{\leftarrow}(r), r > 1$ and $\mathbf{W}$ is drawn from the probability distribution H. As $H(\mathcal{N}_0) = 0$, $\mathbf{X}$ has support only on $\mathfrak{C} \setminus \mathfrak{C}_0$.

Because the hidden measure is finite, there exists a random vector $\mathbf{U}$ defined on the same probability space and independent of $\mathbf{X}$ which is regular varying on $\mathfrak{C}_0$ with limit measure ν_0; that is,

$$t\mathbb{P}\left[\frac{\mathbf{U}}{b_0(t)} \in \cdot\right] \xrightarrow{v} \nu_0(\cdot)$$

as $t \to \infty$ in $M_+(\mathfrak{C}_0)$. Letting B be a Bernoulli(0.5) random variable, Maulik and Resnick (2004) define $\mathbf{Z}_{mix} = B\mathbf{X} + (1 - B)\mathbf{U}$ and show that $\mathbf{Z}_{mix}$ is tail equivalent to $\mathbf{Z}$ on both $\mathfrak{C}$ and $\mathfrak{C}_0$.

While it satisfies the tail equivalence requirements, $\mathbf{Z}_{mix}$ may exhibit different finite-sample behavior than $\mathbf{Z}$. As an example, asymptotic independence in $d = 2$ implies that ν concentrates on the axes of the positive quadrant in $\mathbb{R}^2$. Thus, a realization from $\mathbf{Z}_{mix}$ would consist of points falling exactly on the axes. However, this may not be the case for finite-sample realizations of $\mathbf{Z}$; see Section 4.6.2.

4.4.3 Two-dimensional Max-linear Combination

In the case where $\mathbf{Z} \in \mathbb{R}^2$ is regular varying with tail index $\alpha = 1$ and exhibits asymptotic independence and hidden regular variation with tail index $\alpha_0 \in (1, 2)$, de Haan and Zhou (2011) employ the transformation discussed in Section 4.3.2 to construct a tail equivalent representation to $\mathbf{Z}$. In this case, the hidden angular measure H^* under the transformation T^* is guaranteed to be finite. Denote the total mass of the measure H^* as

$$D = \int_{(0,1)} H^*(dw)$$

and let R^* be such that $\mathbb{P}(R^* > r) = D/r$ for $r \geq D$. Consider Θ independent of R^* with $\mathbb{P}(\Theta \leq \theta) = D^{-1} \int_0^\theta H^*(dw)$. Since $\alpha_0 \in (1, 2)$ there exists a constant $\beta \in (1, 1/\alpha_0)$. Define $\mathbf{X} = (X_1, X_2)^T$ via

$$X_1 = \frac{R^*}{\Theta} \wedge (R^*)^\beta, \quad X_2 = \frac{R^*}{1-\Theta} \wedge (R^*)^\beta.$$

Independently of $\mathbf{X}$, define $W_j, j = 1, 2$ independent with $\mathbb{P}(W_j > x) = 1/x$ for $x \geq 1$. Construct $\mathbf{Z}^* = (Z_1^*, Z_2^*)^T$ via

$$Z_1^* = W_1 \vee X_1^{1/\alpha_0}, \quad Z_2^* = W_2 \vee X_2^{1/\alpha_0}.$$

de Haan and Zhou (2011) show that $\mathbf{Z}^*$ is tail equivalent to $\mathbf{Z}$ on both $\mathfrak{C}$ and $\mathfrak{C}_0$. However, finite-sample behavior of $\mathbf{Z}^*$ may differ considerably from that of $\mathbf{Z}$; we explore this in Section 4.6.2.

4.5 Regular Varying Sum Characterization

Consider a random vector $\mathbf{Z}$ satisfying (16) and (17). We develop a characterization for such a random vector as the sum of independent regular varying components. The

components of the sum are constructed in a manner similar to the components of the mixture characterization of Maulik and Resnick (2004); however, our construction does not require that the hidden angular measure be finite. We prove tail equivalence to $\mathbf{Z}$ on both $\mathfrak{C}$ and $\mathfrak{C}_0$.

4.5.1 Construction

Define a random vector $\mathbf{Y} = R\mathbf{W}$ taking values in $[0, \infty)^d$, where $\mathbb{P}(R > r) \sim 1/b^{\leftarrow}(r)$ as $r \to \infty$ and $\mathbf{W}$ is drawn from $H(\cdot)$, where $b(t)$ is as in (16) and H is the angular measure corresponding to ν in (16). As $\nu(\mathfrak{C}_0) = 0$, $H(\mathcal{N} \cap \mathfrak{C}_0) = 0$. Assume that the quantities R and $\mathbf{W}$ are independent. It follows that in $M_+(\mathfrak{C})$

$$t\mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in \cdot\right] \xrightarrow{v} \nu(\cdot)$$

as $t \to \infty$ (Maulik and Resnick, 2004). The limited support of $\mathbf{Y}$ on only $\mathfrak{C} \setminus \mathfrak{C}_0$ is a key feature of this construction; $\mathbf{Y}$ has no hidden regular variation on $\mathfrak{C}_0$.

Now consider a random vector $\mathbf{V} \in [0, \infty)^d$ defined on the same probability space and independent of R and $\mathbf{W}$ which satisfies the hidden regular variation condition on $\mathfrak{C}_0$ with tail index α_0. Specifically, assume

$$t\mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in \cdot\right] \xrightarrow{v} \nu_0(\cdot) \tag{19}$$

in $M_+(\mathfrak{C}_0)$, with ν_0 as in (17); that is, $\mathbf{V}$ has the same tail behavior as $\mathbf{Z}$ on $\mathfrak{C}_0$.

When the measure ν_0 is finite, $\mathbf{V}$ can be constructed in a similar manner to $\mathbf{Y}$, with support only on $\mathfrak{C}_0$. Assumption (19) does not restrict the support of $\mathbf{V}$; instead, assume

that on the full cone $\mathfrak{C}$,

$$\mathbb{P}(\|\mathbf{V}\| > r) \sim cr^{-\alpha^*} \text{ as } r \to \infty,$$

for some $c > 0$, where the tail index α^* satisfies

$$\alpha^* > \alpha \vee (\alpha_0 - \alpha). \tag{20}$$

We will see that the purpose of assumption (20) is twofold: the condition $\alpha^* > \alpha$ is needed to obtain tail equivalence on $\mathfrak{C}$, while obtaining tail equivalence on $\mathfrak{C}_0$ will require $\alpha^* > \alpha_0 - \alpha$.

Maulik and Resnick (2004) show that mixtures of $\mathbf{Y}$ and $\mathbf{V}$ are tail equivalent to $\mathbf{Z}$ on both $\mathfrak{C}$ and $\mathfrak{C}_0$. In practice, it may be more natural to represent $\mathbf{Z}$ as a sum of the random vectors $\mathbf{Y}$ and $\mathbf{V}$. Next we show

$$\mathbf{Z} \overset{\text{te}(\mathfrak{C})}{\sim} \mathbf{Y} + \mathbf{V} \qquad \text{and} \tag{21}$$

$$\mathbf{Z} \overset{\text{te}(\mathfrak{C}_0)}{\sim} \mathbf{Y} + \mathbf{V}, \tag{22}$$

with scaling constant $c = 1$ in (18). The result (21) follows from Jessen and Mikosch (2006); we review the proof below. Following a similar argument, we prove (22).

4.5.2 Tail Equivalence on $\mathfrak{C}$

With $\mathbf{Y}$ and $\mathbf{V}$ defined above, we adapt Lemma 3.12 of Jessen and Mikosch (2006) to show tail equivalence on the full cone $\mathfrak{C}$. Consider a relatively compact rectangle $A \in \mathfrak{C}$; that is, A is bounded away from $\mathbf{0}$. This class of sets A generates vague convergence in $\mathfrak{C}$ (Resnick, 2007, Lemma 6.1); thus it is sufficient to show

$$\lim_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A\right] = \lim_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Z}}{b(t)} \in A\right] = \nu(A).$$

Assume without loss of generality that $A = [\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in \mathfrak{C} : \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$. For small $\epsilon > 0$, define $\mathbf{a}^{-\epsilon} = (\max\{0, a_1 - \epsilon\}, ..., \max\{0, a_d - \epsilon\})^T$, and define $\mathbf{b}^{-\epsilon}$ analogously. Define the rectangles $A^{-\epsilon} = [\mathbf{a}^{-\epsilon}, \mathbf{b}]$ and $A^{\epsilon} = [\mathbf{a}, \mathbf{b}^{-\epsilon}]$. For small ϵ, the rectangles A^{ϵ} and $A^{-\epsilon}$ are relatively compact in $\mathfrak{C}$, and $A^{\epsilon} \subset A \subset A^{-\epsilon}$. Note that $\nu(\partial A) = 0$ (Jessen and Mikosch, 2006); there is no mass on the boundary of A.

For small $\epsilon > 0$ and fixed $t > 0$,

$$\begin{aligned}\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A\right] &= \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A, \frac{\|\mathbf{V}\|}{b(t)} > \epsilon\right] + \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A, \frac{\|\mathbf{V}\|}{b(t)} \leq \epsilon\right] \\ &\leq \mathbb{P}\left[\|\mathbf{V}\| > b(t)\epsilon\right] + \mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in A^{-\epsilon}\right].\end{aligned}$$

Thus

$$\begin{aligned}\limsup_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A\right] &\leq \limsup_{t\to\infty} t\mathbb{P}\left[\|\mathbf{V}\| > b(t)\epsilon\right] + \limsup_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in A^{-\epsilon}\right] \\ &= \lim_{t\to\infty} t^{1-\alpha^*/\alpha}\epsilon^{-\alpha^*} + \limsup_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in A^{-\epsilon}\right] \\ &= \nu(A^{-\epsilon}) \searrow \nu(A) \text{ as } \epsilon \to 0,\end{aligned}$$

since $\alpha^* > \alpha$ by assumption. For the lower bound, recognize

$$\begin{aligned}\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A\right] &\geq \mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in A^{\epsilon}, \frac{\|\mathbf{V}\|}{b(t)} \leq \epsilon\right] \\ &\geq \mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in A^{\epsilon}\right] - \mathbb{P}\left[\|\mathbf{V}\| > b(t)\epsilon\right],\end{aligned}$$

and so

$$\begin{aligned}\liminf_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in A\right] &\geq \liminf_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}}{b(t)} \in A^{\epsilon}\right] - \liminf_{t\to\infty} t\mathbb{P}\left[\|\mathbf{V}\| > b(t)\epsilon\right] \\ &= \nu(A^{\epsilon}) \nearrow \nu(A) \text{ as } \epsilon \to 0.\end{aligned}$$

Collecting the upper and lower bounds, and using the fact that A is a ν-continuity set, we achieve the desired result

$$t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b(t)} \in \cdot\right] \xrightarrow{v} \nu(\cdot)$$

in $M_+(\mathfrak{C})$.

4.5.3 Tail Equivalence on $\mathfrak{C}_0$

For (22) to hold, it is sufficient to show the following result:

Theorem 1. *For* $\mathbf{Y}$, $\mathbf{V}$, $b_0(t)$, *and* ν_0 *defined as above,*

$$t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in \cdot\right] \xrightarrow{v} \nu_0(\cdot) \tag{23}$$

as $t \to \infty$ *in* $M_+(\mathfrak{C}_0)$. *That is,* $\mathbf{Y}+\mathbf{V} \overset{te(\mathfrak{C}_0)}{\sim} \mathbf{Z}$.

Proof. It suffices to consider any rectangle A_0 which is relatively compact in $\mathfrak{C}_0$, and show that

$$\lim_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0\right] = \nu_0(A_0).$$

Without loss of generality assume $A_0 = [\mathbf{c}, \mathbf{d}] = \{\mathbf{x} \in \mathfrak{C}_0 : \mathbf{c} \le \mathbf{x} \le \mathbf{d}\}$. For small $\epsilon > 0$, define the rectangles $A_0^{-\epsilon} = [\mathbf{c}^{-\epsilon}, \mathbf{d}]$ and $A_0^{\epsilon} = [\mathbf{c}, \mathbf{d}^{-\epsilon}]$, with $\mathbf{c}^{-\epsilon} = (c_1 - \epsilon, ..., c_d - \epsilon)^T$ and $\mathbf{d}^{-\epsilon} = (d_1 - \epsilon, ..., d_d - \epsilon)^T$. For small ϵ, A_0^{ϵ} and $A_0^{-\epsilon}$ are relatively compact in $\mathfrak{C}_0$, $A_0^{\epsilon} \subset A_0 \subset A_0^{-\epsilon}$, and $\nu_0(\partial A_0) = 0$.

Recognize that for small $\epsilon > 0$ and fixed $t > 0$,

$$\begin{aligned}\mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in A_0^{\epsilon}\right] &= \mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in A_0^{\epsilon}, \frac{\|\mathbf{Y}\|}{b_0(t)} \le \epsilon\right] + \mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in A_0^{\epsilon}, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right] \\ &\le \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0\right] + \mathbb{P}\left[\frac{\|\mathbf{V}\|}{b_0(t)} \ge \|\mathbf{c}\|, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right].\end{aligned}$$

Thus by definition of $\mathbf{Y}$ and $\mathbf{V}$ and independence,

$$\begin{aligned}
\liminf_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0\right] &\geq \liminf_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in A_0^{\epsilon}\right] \\
&\quad - \liminf_{t\to\infty} t\mathbb{P}\left[\frac{\|\mathbf{V}\|}{b_0(t)} \geq \|\mathbf{c}\|\right] \mathbb{P}\left[\frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right] \\
&= \nu_0(A_0^{\epsilon}) - \lim_{t\to\infty} t(t^{-\alpha^*/\alpha_0}\|\mathbf{c}\|^{-\alpha^*})(t^{-\alpha/\alpha_0}\epsilon^{-\alpha}) \\
&= \nu_0(A_0^{\epsilon}) - \lim_{t\to\infty} t^{1-(\alpha^*+\alpha)/\alpha_0}\|\mathbf{c}\|^{-\alpha^*}\epsilon^{-\alpha} \\
&= \nu_0(A_0^{\epsilon}) \nearrow \nu_0(A_0) \text{ as } \epsilon \to 0,
\end{aligned}$$

since A_0 is a ν_0-continuity set. Here we have used the assumption $\alpha^* > \alpha_0 - \alpha$.

For the upper bound, we employ the fact that $\nu(\mathfrak{C}_0) = 0$. For fixed t,

$$\begin{aligned}
\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0\right] &= \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0, \frac{\|\mathbf{Y}\|}{b_0(t)} \leq \epsilon\right] + \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right] \\
&= I + II.
\end{aligned}$$

Notice that I is bounded above by

$$\mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in A_0^{-\epsilon}\right].$$

Recalling that by construction $\mathbb{P}[\mathbf{Y}/b_0(t) \in A^{-\epsilon}] = 0$,

$$\begin{aligned}
II &= \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon, \frac{\mathbf{V}}{b_0(t)} \in A_0, \frac{\mathbf{Y}}{b_0(t)} \notin A_0^{-\epsilon}\right] \\
&\quad + \mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon, \frac{\mathbf{V}}{b_0(t)} \notin A_0, \frac{\mathbf{Y}}{b_0(t)} \notin A_0^{-\epsilon}\right] \\
&\leq \mathbb{P}\left[\frac{\|\mathbf{V}\|}{b_0(t)} \geq \|\mathbf{c}\|, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right] + \mathbb{P}\left[\frac{\vee_{i=1}^d V_i}{b_0(t)} > \epsilon, \frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right].
\end{aligned}$$

Then

$$\begin{aligned}
\limsup_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{Y}+\mathbf{V}}{b_0(t)} \in A_0\right] &\leq \limsup_{t\to\infty} t\mathbb{P}\left[\frac{\mathbf{V}}{b_0(t)} \in A_0^{-\epsilon}\right] \\
&\quad + \limsup_{t\to\infty} t\mathbb{P}\left[\frac{\|\mathbf{V}\|}{b_0(t)} \geq \|\mathbf{c}\|\right] \mathbb{P}\left[\frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right] \\
&\quad + \limsup_{t\to\infty} t\mathbb{P}\left[\frac{\vee_{i=1}^d V_i}{b_0(t)} > \epsilon\right] \mathbb{P}\left[\frac{\|\mathbf{Y}\|}{b_0(t)} > \epsilon\right] \\
&= \nu_0(A_0^{-\epsilon}) + \lim_{t\to\infty} t(t^{-\alpha^*/\alpha_0}\|\mathbf{c}\|^{-\alpha^*})(t^{-\alpha/\alpha_0}\epsilon^{-\alpha}) \\
&\quad + \lim_{t\to\infty} t(t^{-\alpha^*/\alpha_0}\epsilon^{-\alpha^*})(t^{-\alpha/\alpha_0}\epsilon^{-\alpha}) \\
&= \nu_0(A_0^{-\epsilon}) \searrow \nu_0(A_0) \text{ as } \epsilon \to 0,
\end{aligned}$$

by independence and ν_0-continuity of A_0, and again following from $\alpha^* > \alpha_0 - \alpha$.

Finally, putting together the upper and lower bounds yields the desired result (23).

■

Remark 1. *Heuristically, the scaled random vector* $(\mathbf{Y}+\mathbf{V})/b_0(t)$ *can only land in* $\mathfrak{C}_0$ *when* $\|\mathbf{Y}\|$ *is small and* $\|\mathbf{V}\|$ *is large. Suitably normalized large values of* $\mathbf{Y}$ *will converge to points outside of* $\mathfrak{C}_0$*, and by independence, the probability of* $\mathbf{Y}$ *and* $\mathbf{V}$ *being simultaneously large is asymptotically negligible.*

Remark 2. *The proof relies on* $\mathbf{Y}$ *being constructed in such a way that* $\mathbb{P}[\mathbf{Y} \in \mathfrak{C}_0] = 0$*. Such a condition gives convergence to the measure* ν_0 *on* $\mathfrak{C}_0$*. The result may not hold in general if* $\mathbf{Y}$ *has angular measure* H *only in the limit and exhibits hidden regular variation on* $\mathfrak{C}_0$*. We do not impose such additional conditions on the support of* $\mathbf{V}$*.*

Remark 3. *Assumption* (20) *imposes two constraints on the behavior of* $\mathbf{V}$ *on* $\mathfrak{C}$*, and Figure 4.2 gives a plot of valid values of* (α_0, α^*) *for* $\alpha = 1$*. The first, requiring* $\alpha^* > \alpha$*, is needed to obtain convergence of the properly normalized sum to the required limit measure* ν *on* $\mathfrak{C}$*. This assumption eliminates the possibility of taking* $\mathbf{V}$ *to be* $\mathbf{Z}$ *itself, and poses additional difficulty in the case where* ν_0 *is infinite on* $\mathfrak{C}_0$*.*

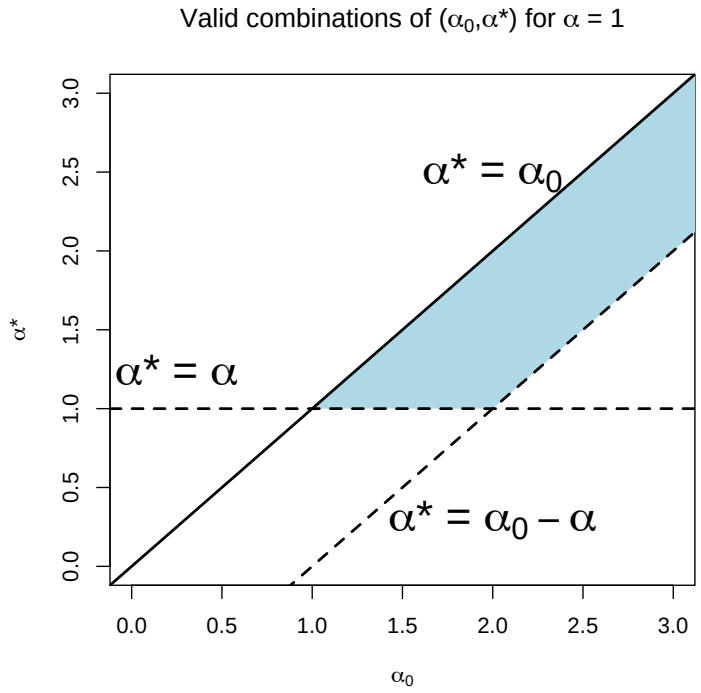

Figure 4.2: Valid choices of α^* for different values of α_0 (blue shading) when $\alpha = 1$.

The second assumption imposed by (20)*, namely that* $\alpha^* > \alpha_0 - \alpha$*, is necessary to obtain convergence on* $\mathfrak{C}_0$*. Essentially, if* $\mathbf{V}$ *is much heavier-tailed on* $\mathfrak{C}$ *than on* $\mathfrak{C}_0$*, convergence is not obtained on* $\mathfrak{C}_0$*. As an example, consider a random vector in dimension* $d = 3$ *which is regular varying on* $\mathfrak{C}$ *with tail index* $\alpha = 1$*. To represent hidden regular variation of tail index* $\alpha_0 = 5/2$ *on the full open subcone* $\mathfrak{C}_0 = \{\mathbf{z} \in \mathfrak{C} : z_1 \wedge z_2 \wedge z_3 > 0\}$*, one would need to choose* $\mathbf{V}$ *to have tail decay on* $\mathfrak{C}$ *which is lighter than that corresponding to* $\alpha^* = 3/2$*.*

On the other hand, the case when this condition does not hold may not be of interest in applications, as hidden measure tail dependence implied by such cases is weaker than a scenario of exact independence. One example in dimension $d = 2$ *is a distribution with unit Fréchet margins and Gaussian dependence with negative correlation. When the hidden angular measure is finite, one can always choose* $\alpha^* = \alpha_0$ *and restrict the support of* $\mathbf{V}$ *to* $\mathfrak{C}_0$*.*

4.6 Tail Equivalent Representations to the Bivariate Gaussian Example

In this section, we consider the bivariate random vector $\mathbf{Z} = (Z_1, Z_2)^T$, where $Z_j = -1/\log \Phi(X_j)$, $j = 1, 2$ and $(X_1, X_2)^T$ follows a bivariate Gaussian distribution with unit marginal variances and correlation $\rho \in [0, 1)$. Here $\Phi(\cdot)$ is the standard Gaussian distribution function. Sibuya (1960) showed that asymptotic independence holds; indeed, we can find $b(t) \in RV_1$ such that (16) holds with limit measure $\nu = \nu_1 \times H$, where H consists of point masses at the endpoints $\mathcal{N} \cap \{\mathbf{z} \in \mathfrak{C} : z_1 \wedge z_2 = 0\}$. If $b(t) = 2t$, H is a probability measure with point masses of $1/2$ at $w = 0$ and $w = 1$.

An exploration of the hidden regular variation of $\mathbf{Z}$ was provided by Ledford and Tawn (1996, 1997). Ledford and Tawn (1996) formulate this in terms of the joint survivor function $\bar{F}(z_1, z_2) = \mathbb{P}[Z_1 > z_1, Z_2 > z_2]$. Ledford and Tawn (1997) show

$$\bar{F}(z_1, z_2) \approx (z_1 z_2)^{-1/(1+\rho)} \mathcal{L}(z_1, z_2; \rho)(1 + O[1/\log\{\min(z_1, z_2)\}])$$

for large $z_1, z_2 > 0$, where $\mathcal{L}(z_1, z_2)$ is a slowly varying function with (Ledford and Tawn, 1996)

$$\mathcal{L}(t, t; \rho) = (1 + \rho)^{3/2}(1 - \rho)^{-1/2}(4\pi \log t)^{-\rho/(1+\rho)}. \tag{24}$$

Let $\eta = (1 + \rho)/2$. Ledford and Tawn (1996) refer to η as the coefficient of tail dependence. Consider a set $(\mathbf{z}, \boldsymbol{\infty}]$ for $\mathbf{z} = (z_1, z_2)^T$ with $z_1, z_2 > 0$. One can show

$$t\mathbb{P}\left[\frac{\mathbf{Z}}{b_0(t)} \in (\mathbf{z}, \boldsymbol{\infty}]\right] \longrightarrow (z_1 z_2)^{-1/2\eta} = \nu_0\left((\mathbf{z}, \boldsymbol{\infty}]\right) \tag{25}$$

as $t \to \infty$, where the function $b_0(t) = 2U^{\leftarrow}(t)$, with

$$U(t) = \frac{(2t)^{1/\eta}}{\mathcal{L}(2t, 2t)},$$

for $\mathcal{L}$ given by (24).

It is easily shown for sets of the form $A(r, B) = \{\mathbf{z} \in \mathfrak{C}_0 : \|\mathbf{z}\| > r, \mathbf{z}\|\mathbf{z}\|^{-1} \in B\}$ that

$$\nu_0(A(r, B)) = r^{-1/\eta} H_0(B), \tag{26}$$

where B is a Borel set of $\mathcal{N}_0 = (0, 1)$ and

$$H_0(dw) = \frac{1}{4\eta}\{w(1-w)\}^{-1-1/2\eta}; \tag{27}$$

see, e.g., Beirlant et al., 2004, Chapter 9. Note that $H_0(\mathcal{N}_0) = \int_{(0,1)} H_0(dw) = +\infty$, thus the hidden measure ν_0 is infinite on $\mathfrak{C}_0$.

The fact that the hidden angular measure is infinite makes this a challenging example and poses difficulty in finite-sample simulation of the joint tail of $\mathbf{Z}$. Because the hidden measure diverges near the endpoints of $\mathcal{N}_0$, one always encounters difficulty near the axes of $\mathfrak{C}$. We first offer a modification of the $\mathbf{Y} + \mathbf{V}$ representation introduced above, and then compare it to previous approaches described in Section 4.4.

4.6.1 Simulation from Sum Representation

Because the hidden measure with density $H_0(dw)$ given by (27) is infinite on $(0, 1)$, one cannot simulate from it directly. Furthermore, in this situation we are unaware of a random vector $\mathbf{V}$ which satisfies both assumptions (19) and (20). As an alternative, we propose an approximation to $H_0(dw)$ by restricting the subcone $\mathfrak{C}_0$ to $\mathfrak{C}_0^\epsilon = \{\mathbf{z} \in \mathfrak{C}_0 : z_1\|\mathbf{z}\|^{-1} \in \mathcal{N}_0^\epsilon\}$, where $\mathcal{N}_0^\epsilon = [\epsilon, 1-\epsilon]$ for some $\epsilon \in (0, 1/2)$. The density (27) can then be made to be a

probability density on $\mathcal{N}_0^\epsilon$ via $H_0^\epsilon(dw) = H_0(dw)/H_0(\mathcal{N}_0^\epsilon)$ for $w \in \mathcal{N}_0^\epsilon$. One can then simulate realizations from H_0^ϵ via an accept-reject algorithm or other sampling method.

We proceed to simulate realizations of $\mathbf{Z}_{sum} = \mathbf{Y} + \mathbf{V}$, which is tail equivalent to $\mathbf{Z}$ on $\mathfrak{C}$ and $\mathfrak{C}_0^\epsilon$. Define $\mathbf{Y}$ as follows: let R follow a Pareto distribution with $\mathbb{P}(R > r) = 2/r$ for $r \geq 2$. Draw a Bernoulli(1/2) random variable W independently of R, and let $Y_1 = RW$, $Y_2 = R(1-W)$. For a fixed sample size n, draw R_0 independently of $\mathbf{Y} = (Y_1, Y_2)^T$ with R_0 such that

$$\mathbb{P}(R_0 > x) = \begin{cases} d_{\epsilon,n} x^{-1/\eta} & \text{if } x > (d_{\epsilon,n})^\eta \\ 1 & \text{otherwise,} \end{cases}$$

where

$$d_{\epsilon,n} = \{2U^{\leftarrow}(n)\}^{1/\eta} \left\{ \frac{H_0(\mathcal{N}_0^\epsilon)}{n} \right\}.$$

Draw n independent realizations of W_0 from the density $H_0^\epsilon(dw)$ independently of $\mathbf{Y}$ and R_0. Define $\mathbf{V} = (V_1, V_2)^T$ via

$$V_1 = R_0 W_0, \quad V_2 = R_0(1 - W_0).$$

Then for any set $A(r, B) = \{\mathbf{z} \in \mathfrak{C}_0^\epsilon : \|\mathbf{z}\| > r, z_1 \|\mathbf{z}\|^{-1} \in B\}$ with B a Borel set of $\mathcal{N}_0^\epsilon$,

$$\begin{aligned} n\mathbb{P}\left[\frac{\mathbf{V}}{b_0(n)} \in A(r,B) \right] &= n\mathbb{P}\left[\frac{R_0}{2U^{\leftarrow}(n)} > r, W_0 \in B \right] \\ &= n\mathbb{P}\left[R_0 > 2rU^{\leftarrow}(n) \right] \mathbb{P}[W_0 \in B] \\ &= n\left[d_{\epsilon,n}(2rU^{\leftarrow}(n))^{-1/\eta} \right] \frac{H_0(B)}{H_0(\mathcal{N}_0^\epsilon)} \\ &= r^{-1/\eta} H_0(B) \end{aligned}$$

for $r > \{H_0(\mathcal{N}_0^\epsilon)/n\}^\eta$, which is precisely the decomposition of ν_0 in (26).

When examining the limiting measure of a set in the full subcone $\mathfrak{C}_0$ which is not completely contained in $\mathfrak{C}_0^\epsilon$, a bias is induced by the choice of ϵ. To see this, extend the restricted hidden measure by setting $H_0^\epsilon\{(0,\epsilon)\} = H_0^\epsilon\{(1-\epsilon,1)\} = 0$, and consider a set $A = (\mathbf{z}, \boldsymbol{\infty}]$ for $z_1, z_2 > 0$. Note that one can choose n and ϵ such that $\mathbf{z} \in \mathfrak{C}_0^\epsilon$ and $z_1 + z_2 > \{H_0(\mathcal{N}_0^\epsilon)/n\}^\eta$, and in this case we have

$$\begin{aligned}
n\mathbb{P}\left[\frac{\mathbf{V}}{b_0(n)} \in A\right] &= n\int_0^1 \int_{\frac{b_0(n)z_1}{w} \vee \frac{b_0(n)z_2}{1-w}}^\infty \eta^{-1} d_{\epsilon,n} r^{-(1+1/\eta)} dr H_0^\epsilon(dw) \\
&= \int_0^1 \left\{\frac{w}{z_1} \bigwedge \frac{1-w}{z_2}\right\}^{1/\eta} H_0(\mathcal{N}_0^\epsilon) H_0^\epsilon(dw) \\
&= \int_\epsilon^{1-\epsilon} \left\{\frac{w}{z_1} \bigwedge \frac{1-w}{z_2}\right\}^{1/\eta} H_0(dw) \\
&= (z_1 z_2)^{-1/2\eta} - \frac{1}{2}\left(\frac{\epsilon}{1-\epsilon}\right)^{1/2\eta}\left(z_1^{-1/\eta} + z_2^{-1/\eta}\right) \\
&= \nu_0(A) - B(\epsilon, \mathbf{z}), \qquad (28)
\end{aligned}$$

where the bias term $B(\epsilon, \mathbf{z})$ can be made arbitrarily small via choice of ϵ.

4.6.2 Comparison to Other Representations

Figure 4.3 shows $n = 2500$ simulated realizations of $\mathbf{Z}$ with correlation $\rho = 0.5$, as well as simulated points from four tail equivalent representations. We display a simulation from $\mathbf{Z}_{sum}$ for $\epsilon = 0.01, 0.1$, as well as $\mathbf{Z}^*$ of de Haan and Zhou (2011), and a mixture $\mathbf{Z}_{mix}$ of $\mathbf{Y}$ and $\mathbf{V}$ for $\epsilon = 0.1$. We also display a realization of $\mathbf{Y}$, the limiting first-order component, in the lower middle panel of Figure 4.3. The dashed lines imposed on plots of realizations from each of the tail equivalent representations correspond to the restriction of $\mathfrak{C}_0$ on which the hidden regular varying component can be simulated; for $\mathbf{Z}_{sum}$, this corresponds to $\mathfrak{C}_0^\epsilon$. The de Haan and Zhou (2011) representation $\mathbf{Z}^*$ was constructed as described in Section 4.4; for this example the measure H^* is proportional to a Beta$(1/2, 1/2)$ distribution. This construction also requires a choice of fixed constant $\beta \in (1, 1/\eta)$, which restricts the simulation of the 'hidden'

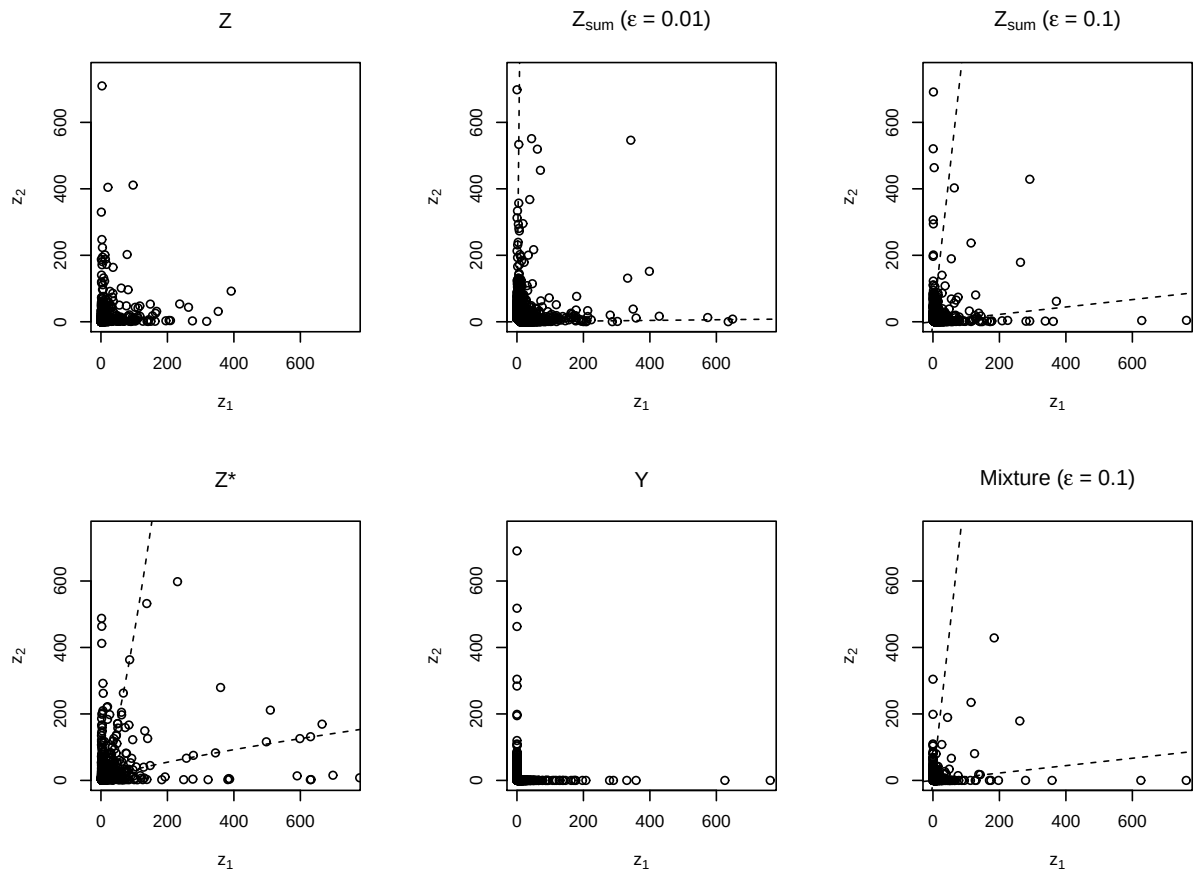

Figure 4.3: Simulation of $n = 2500$ points from (left to right, top to bottom) $\mathbf{Z}$ with $\rho = 0.5$ and its tail equivalent representations: $\mathbf{Z}_{sum}$ for $\epsilon = 0.01, 0.1$; $\mathbf{Z}^*$ of de Haan and Zhou (2011); $\mathbf{Y}$; and a mixture of $\mathbf{Y}$ and $\mathbf{V}$ with $\epsilon = 0.1$. Boundaries of set for which hidden components are simulated are given by dashed lines.

component to a subset of $\mathfrak{C}_0$ given by $\{(z_1, z_2) : z_1^{1/\beta} \leq z_2 \leq z_1^{\beta}\}$. Here $\beta = 1/\eta - 0.01 \approx 1.32$ was selected.

The representation $\mathbf{Z}_{sum}$ is tail equivalent to $\mathbf{Z}$ on $\mathfrak{C}$ and $\mathfrak{C}_0^{\epsilon}$, and its realizations are comparable to those from $\mathbf{Z}$ in tail regions both near the axes and on the interior of the positive quadrant. The primary difference between $\mathbf{Z}_{sum}$ simulations with $\epsilon = 0.01$ and $\epsilon = 0.1$ is the number of points near the axes of the cone $\mathfrak{C}$. As ϵ decreases, we see more large points with angular components near 0 and 1 in finite samples. This is due to the increase in scale parameter of R_0 induced by smaller ϵ; see Section 4.6.3. The restriction induced by choice of β in the construction of $\mathbf{Z}^*$ is seen in finite-sample simulations from this

representation. The first-order component $\mathbf{Y}$ does not capture tail behavior on the interior of the cone $\mathfrak{C}$. The mixture representation of Maulik and Resnick (2004) results in points which fall exactly on each axis, which does not occur in finite samples from $\mathbf{Z}$.

In Table 4.1, we provide a comparison of $\mathbf{Z}$ and each of the tail equivalent representations displayed in Figure 4.3 by examining the empirical average number of points in specific sets over 250 simulations of $n = 2500$ points from each for $\rho = 0.5$ and $\epsilon = 0.01, 0.1$. We display mean number of points in the set $(\mathbf{z}, \boldsymbol{\infty}]$, where $\mathbf{z} = (z, z)^T$ for various z. Also shown are mean number of realizations for which $z_2 > z$ for a range of z. Simulation-based 95% intervals for each quantity are also displayed.

For $\epsilon = 0.01$, the convergence of $\mathbf{Z}_{sum}$ to the limiting measure is slow for regions that are near the axes of the cone $\mathfrak{C}$, as demonstrated by examination of the empirical average number of points with $z_2 > z$. We examine this in detail in Section 4.6.3. For sets of the form $A = (\mathbf{z}, \boldsymbol{\infty}]$, choosing ϵ small results in near-unbiased approximation of $\nu_0(A)$ by $\mathbf{Z}_{sum}$. Choosing ϵ slightly larger results in faster convergence to the limiting measure in terms of marginal distributions, but results in greater bias for sets on $\mathfrak{C}_0$; see (28). The de Haan and Zhou (2011) representation $\mathbf{Z}^*$ results in more points than expected in $\mathbf{Z}$ in most regions of the cone $\mathfrak{C}$. This is likely due to the slowly varying function $\mathcal{L}$, which is ignored by this representation. The first-order approximation $\mathbf{Y}$ fails to capture any of the distribution of $\mathbf{Z}$ on $\mathfrak{C}_0$. The mixture characterization exhibits approximately half as many points in each set as seen in $\mathbf{Z}$ due to the 'thinning' induced by mixing $\mathbf{Y}$ and $\mathbf{V}$.

4.6.3 Choice of ϵ

Figure 4.3 and Table 4.1 show that realizations of $\mathbf{Z}_{sum}$ result in more large observations near the axes of the cone $\mathfrak{C}$ than seen in realizations of $\mathbf{Z}$. This is not surprising when one considers the limiting marginal measure of $\mathbf{V}$:

Table 4.1: Summary statistics from 250 simulations of $n = 2500$ points from $\mathbf{Z}$ with $\rho = 0.5$ and its tail equivalent representations. Figures reported are empirical means and simulation-based 95% intervals.

	Number of points in the set $(\mathbf{z}, \boldsymbol{\infty}]$			Number of points with $Z_2 > z$		
z	100	250	500	500	1000	2000
$\mathbf{Z}$	3.20 $(1, 6)$	0.90 $(0, 3)$	0.36 $(0, 2)$	4.80 $(2, 8)$	2.39 $(0, 6)$	1.29 $(0, 3)$
$\mathbf{Z}_{sum}$ $(\epsilon = 0.01)$	3.97 $(1, 7)$	1.08 $(0, 3)$	0.36 $(0, 2)$	9.39 $(5, 14)$	4.19 $(2, 8)$	1.89 $(0, 4)$
$\mathbf{Z}_{sum}$ $(\epsilon = 0.1)$	2.89 $(1, 6)$	0.84 $(0, 3)$	0.3 $(0, 2)$	6.03 $(2, 10)$	2.86 $(0, 6)$	1.42 $(0, 4)$
$\mathbf{Z}^*$	11.29 $(7, 16)$	3.21 $(1, 6)$	1.14 $(0, 3)$	8.48 $(5, 14)$	3.97 $(1, 7)$	1.93 $(0, 4)$
$\mathbf{Z}_{mix}$ $(\epsilon = 0.1)$	1.46 $(0, 4)$	0.46 $(0, 2)$	0.18 $(0, 1)$	3.00 $(1, 6)$	1.42 $(0, 4)$	0.70 $(0, 2)$
$\mathbf{Y}$	0	0	0	5.00 $(2, 9)$	2.37 $(0, 5)$	1.16 $(0, 3)$

$$
\begin{aligned}
n\mathbb{P}\left[\frac{V_1}{b_0(n)} > z_1\right] &= \int_0^1 \left(\frac{w}{z_1}\right)^{1/\eta} H_0(\mathcal{N}_0^\epsilon) H_0^\epsilon(dw) \\
&= \int_\epsilon^{1-\epsilon} \left(\frac{w}{z_1}\right)^{1/\eta} H_0(dw) \\
&= z_1^{-1/\eta} \int_\epsilon^{1-\epsilon} w^{1/\eta} H_0(dw) \\
&= \frac{1}{2} z_1^{-1/\eta} \left\{ \left(\frac{\epsilon}{1-\epsilon}\right)^{-1/2\eta} - \left(\frac{\epsilon}{1-\epsilon}\right)^{1/2\eta} \right\}. \qquad (29)
\end{aligned}
$$

For very large z_1, (29) is negligible compared to the heavier-tailed component Y_1 of $\hat{Z}_1$, which has limit measure z_1^{-1}. However, for small ϵ the scaling factor in (29) is quite large, and plays a significant role in finite samples. This difficulty can be alleviated by choosing a slightly larger ϵ, which will reduce the magnitude of the scaling factor in (29).

The result of choosing a larger value for ϵ is an increase in the bias term in (28). That is, for sets in $\mathfrak{C}_0$ for which smaller ϵ results in greater coverage by $\mathfrak{C}_0^\epsilon$, a larger ϵ increases the rate of convergence to the limiting measure, but also decreases the accuracy of the approximation to the limiting measure of such a set. Thus the choice of ϵ involves a trade-off between the marginal behavior of $\mathbf{Z}_{sum}$ and the size of the restricted subcone $\mathfrak{C}_0^\epsilon$. This tradeoff will reappear in Chapter 5 when we examine estimation from this representation.

While the infinite hidden angular measure of a Fréchet-marginal random vector with Gaussian dependence poses difficulty in simulation, our sum representation of $\mathbf{Z}$ in terms of

independent $\mathbf{Y}$ and $\mathbf{V}$ provides several advantages over previous approaches. We are able to capture not only the first-order limit on the whole cone $\mathfrak{C}$, but also the hidden regular varying component on the subcone $\mathfrak{C}_0^\epsilon$. We can choose ϵ such that the restricted subcone $\mathfrak{C}_0^\epsilon$ becomes arbitrarily close to $\mathfrak{C}_0$. Choosing ϵ involves a tradeoff between bias in the limiting measure of sets not fully contained in $\mathfrak{C}_0^\epsilon$, and the level at which the limiting measure is a useful approximation for finite samples.

4.7 Summary and Discussion

This chapter presents a new representation of a multivariate regular varying random vector with hidden regular variation, in terms of a sum of independent regular varying components. We have shown our representation to be asymptotically justified via the concept of multivariate tail equivalence. An illustration of simulation from our characterization was provided using the bivariate Gaussian as an example. The infinite hidden measure of this example introduced difficulty in simulation; however, one can still simulate the lighter-tailed component $\mathbf{V}$ on a restricted subcone. Our sum representation shares features with real data in applications and provides an intuitive model for the joint tail of a random vector.

While the examples of the $\mathbf{Y}+\mathbf{V}$ representation presented here are limited to the bivariate case, this construction is able to accommodate more general cases of hidden regular variation which may arise. In general, the support of the angular measure of the first-order component $\mathbf{Y}$ may be any subset of the unit sphere $\mathcal{N}$, and the hidden regular varying component $\mathbf{V}$ can be chosen to have the appropriate hidden angular measure on $\mathcal{N}_0$. For example, this representation might be used to study hidden regular variation in extreme-value factor models (Einmahl et al., 2012) with discrete spectral measures. In higher dimensions, more complex hidden regular variation structures are possible; some examples are given in Mitra and Resnick (2010) and de Haan and Zhou (2011).

CHAPTER 5

LIKELIHOOD INFERENCE FOR HIDDEN REGULAR VARIATION VIA THE MONTE CARLO EXPECTATION–MAXIMIZATION ALGORITHM

5.1 Introduction

In this chapter, we turn our attention to inference for random vectors possessing hidden regular variation. Working from the multivariate regular variation framework introduced in Chapter 1, we focus on estimation of hidden regular variation structures discussed in Chapter 4. Specifically, we leverage the sum characterization introduced in Chapter 4 and tail equivalence result provided by Theorem 1 to perform parametric maximum likelihood inference.

Though the term "hidden regular variation" was first introduced to the literature by Resnick (2002), methodology for its estimation dates back to at least the work of Ledford and Tawn (1996), which considered the structure of the joint tail $\mathbb{P}(X > x, Y > y)$ of a random vector $(X, Y)^T$ with unit Fréchet marginal distributions. This work formulated η, the coefficient of tail dependence described in Chapter 4. This parameter serves as a measure of tail dependence in the asymptotic independence setting, and Ledford and Tawn (1996) provided methodology to distinguish asymptotic independence from independence in the usual sense.

Following Ledford and Tawn (1996), many papers have offered approaches for estimation of hidden regular variation. Ledford and Tawn (1997) and Ramos and Ledford (2009) focus specifically on modeling bivariate joint tails in the case that the first-order limit fails to capture tail dependence. Heffernan and Tawn (2004) offer a conditional approach, while Coles et al. (1999) examine measures of dependence in the asymptotic independence setting.

Draisma et al. (2004) and Peng (1999) offer other approaches for estimating tail dependence in the presence of asymptotic independence. Several of these previous approaches are reviewed in this chapter.

Assuming hidden regular variation which is governed by a parametric form, in this chapter we develop inference methods for these parameters. Specifically, we propose an estimation scheme from the sum characterization of Chapter 4 based on the Monte Carlo expectation–maximization (MCEM) algorithm (Wei and Tanner, 1990). The MCEM algorithm is an alternative to deterministic expectation–maximization when the E step is not available in closed form. Here, we present a modification of MCEM for tail estimation from the proposed sum representation which employs likelihoods for the first- and second-order components which are assumed to be valid above fixed thresholds. The estimation scheme is demonstrated on several examples through simulation studies.

The proposed methodology is applied to a bivariate series of air pollution measurements at Leeds city centre, UK. We examine the tail dependence in daily maximum levels of two pollutants: nitrogen dioxide and sulfur dioxide. The bivariate series appears to exhibit hidden regular variation, and we show that the proposed methodology results in improved estimation of risk set probabilities over approaches which ignore the hidden regular variation.

The remainder of this chapter is structured as follows: in Section 5.2 we review selected previous hidden regular variation estimation approaches in the literature. We detail an estimation scheme based on MCEM in Section 5.3. In Section 5.4 we demonstrate the proposed methodology on simulated data in two studies. An application of the methodology to the Leeds air pollution data is discussed in Section 5.5. We conclude in Section 5.6 with a summary and discussion. A portion of this chapter also appears in Weller and Cooley (2013), which has been submitted for publication.

5.2 Existing Hidden Regular Variation Estimation Methods

We provide a brief review of selected previous statistical approaches to estimation of hidden regular variation. Estimation approaches which are most relevant to the work in this dissertation are presented.

5.2.1 Survivor Function Methods

Ledford and Tawn (1996, 1997) provided a framework for estimation of hidden regular variation in a special case. Focusing on a bivariate random vector $(Z_1, Z_2)^T$ with unit Fréchet marginal distributions, these authors write a model for $\bar{F}(z_1, z_2) = \mathbb{P}(Z_1 > z_1, Z_2 > z_2)$. They provide the following model, which captures a broad class of distributions with Fréchet marginal distributions:

$$\bar{F}(z_1, z_2) \approx \mathcal{L}(z_1, z_2)(z_1 z_2)^{-1/2\eta} \tag{30}$$

for a constant $\eta \in (0, 1]$, and $\mathcal{L}$ a bivariate slowly varying function with

$$\lim_{t\to\infty} \frac{\mathcal{L}(tz_1, tz_2)}{\mathcal{L}(t, t)} = g(z_1, z_2) = g_*\{z_1/(z_1 + z_2)\} = g_*(w). \tag{31}$$

The parameter η describes the rate of decay of the joint survivor function: $\eta = 1$ corresponds to asymptotic dependence, $\eta = 1/2$ is termed 'near independence', and $\eta < 1/2$ can be thought of as negative dependence. In the hidden regular variation framework described in Chapter 4, the hidden tail index α_0 is related to η via $\alpha_0 = 1/\eta$.

Ledford and Tawn (1996) provided an approach for estimation of η. These authors reduced the estimation problem to one dimension by defining the 'structure variable' $T = \min(Z_1, Z_2)$. Given independent replicates $T_1, ..., T_n$, Ledford and Tawn (1996) estimated η as the shape parameter from a generalized Pareto distribution fit to structure variable

exceedances of a high threshold. A generalization to higher dimensions, from an applied probability perspective, is given by Mitra and Resnick (2010).

Using the model (30), Ledford and Tawn (1997) develop likelihood-based inference methods. Specifically, they employ a censored likelihood similar to that of Smith et al. (1997) to estimate η and parameters of the function g_* in (31). The methodology was limited to the bivariate case, and it was later pointed out that the models introduced by Ledford and Tawn (1997) can lead to improper joint survivor functions (Ramos and Ledford, 2011).

The work of Ramos and Ledford (2009) provided a unified point process framework for the model (30) of Ledford and Tawn (1997). Considering a bivariate random vector $(Z_1, Z_2)^T$ with unit Fréchet marginal distributions, these authors provide models for the distribution of

$$(S, T)^T = \lim_{u \to \infty} \{(Z_1/u, Z_2/u)^T \mid (Z_1 > u, Z_2 > u)\}.$$

The authors introduce a polar coordinate decomposition and provide necessary conditions on the resulting angular measure in order to achieve a proper joint tail model. Though not presented through the hidden regular variation framework directly, the parametric angular measure model introduced by Ramos and Ledford (2009) for $(S, T)^T$ can be thought of as the hidden angular measure of $(Z_1, Z_2)^T$.

A common theme of the estimation approaches of Ledford and Tawn (1997) and Ramos and Ledford (2009) is the focus on models for the bivariate region $\{(z_1, z_2) : z_1 > u, z_2 > u\}$ for some large $u > 0$; the proposed models are not able to simultaneously describe regions of the form $\{(z_1, z_2) : z_1 \vee z_2 > u, z_1 \wedge z_2 \leq u\}$. Ramos and Ledford (2011) also note that their proposed model is unable to accommodate the situation where $g_*(w)$ in (31) is constant over w; an example of this case is the bivariate Gaussian example discussed in Section 4.6 of Chapter 4. Additionally, greater complexity is encountered when extending the methodology of Ramos and Ledford (2009) into dimension $d > 2$.

5.2.2 Conditional Approach

An alternative to studying the joint survivor function $\bar{F}$ (in the bivariate setting) is to focus on the conditional distribution $[Z_2 \mid Z_1 = z_1]$. Starting with a d-dimensional random vector with Gumbel-distributed marginals, Heffernan and Tawn (2004) study the conditional distribution of the remaining $d-1$ components given that the jth component is large ($j = 1, ..., d$). These authors develop likelihood-based inference methods for this conditional distribution. Further elaboration of this approach and connections to hidden regular variation are made in Heffernan and Resnick (2007).

5.2.3 Estimators of Dependence in Asymptotic Independence

In addition to estimation of the parameter η of the Ledford and Tawn (1997) model (30), other measures of dependence in the asymptotic independence setting, as well as various estimators of these measures, have been proposed. Section 3.4.2 of Chapter 3 introduced the metric $\bar{\chi}$ of Coles et al. (1999), which also offers an estimator of this parameter. In the bivariate case, this measure can be related to η via $\bar{\chi} = 2\eta - 1$. Peng (1999) introduces a consistent and asymptotically normal estimator of η. Draisma et al. (2004) generalize the approach of Peng (1999) and examine other estimators of η.

5.3 Likelihood Inference via Sum Characterization

We now develop an inference method for hidden regular variation based on the sum characterization introduced in Section 4.5 of Chapter 4. Given independent realizations $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$ from a random vector $\mathbf{Z}$ assumed to be multivariate regular varying on $\mathfrak{C}$ and possessing hidden regular variation on a subcone $\mathfrak{C}_0 \subset \mathfrak{C}$, we leverage the tail equivalence results (21) and (22) of Chapter 4 to perform maximum likelihood inference for the joint tail of $\mathbf{Z}$ via the MCEM algorithm.

5.3.1 The Expectation–Maximization Algorithm

Assume we observe n independent realizations of $\mathbf{Z}$, which exhibits hidden regular variation on a subcone $\mathfrak{C}_0$. We assume there exists a $\mathbf{Y} + \mathbf{V}$ which is tail equivalent on both $\mathfrak{C}$ and $\mathfrak{C}_0$. As the corresponding $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ and $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ are unobserved, we employ the expectation–maximization (EM) algorithm (Dempster et al., 1977), which is widely used to perform inference in such cases. However, EM has not heretofore been employed in tail estimation problems, and an adaptation is necessary here as we assume only tail equivalence of $\mathbf{Z}$ and $\mathbf{Y} + \mathbf{V}$ rather than exact equality.

Classical EM assumes we observe realizations from $\mathbf{Z}$ but wish to make inference on a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ governing the 'complete' $(\mathbf{Z}, \mathbf{Y}, \mathbf{V})$. Given some fixed value of the parameter vector $\boldsymbol{\theta}^{(k)}$ and temporarily assuming $\mathbf{Z} = \mathbf{Y} + \mathbf{V}$, the log-likelihood function for observed $\mathbf{z}$ can be written

$$\begin{aligned} \log f(\mathbf{z}; \boldsymbol{\theta}) &= \int \log f(\mathbf{z}, \mathbf{y}, \mathbf{v}; \boldsymbol{\theta}) f(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) d\mathbf{y} d\mathbf{v} \\ &\quad - \int \log f(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}) f(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) d\mathbf{y} d\mathbf{v} \\ &= Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}). \end{aligned} \tag{32}$$

In standard problems, it can be shown that $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ attains its maximum at $\boldsymbol{\theta}^{(k)}$ (Wu, 1983), and the algorithm iterates on $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ to achieve the optimum $\hat{\boldsymbol{\theta}}$, which corresponds to the maximum likelihood estimate. In the EM literature, $f(\mathbf{z}, \mathbf{y}, \mathbf{v}; \boldsymbol{\theta})$ is referred to as the complete data likelihood, while $f(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)})$ is the conditional likelihood.

5.3.2 Expectation–Maximization for Tail Estimation

The results (21) and (22) imply that inference performed on $\mathbf{Z}$ based on a model for $\mathbf{Y} + \mathbf{V}$ is only valid for the tails of $\mathbf{Y}$ and $\mathbf{V}$, and cannot be assumed to be appropriate for the entire distribution. In the following text, we introduce probability distributions for

$\mathbf{Y}$ and $\mathbf{V}$ components which exist on the entire positive orthant $[\mathbf{0}, \boldsymbol{\infty})$ and which give tail equivalence to the random vector of interest $\mathbf{Z}$. We define a likelihood corresponding to limiting Poisson point process forms arising from regular variation convergences (16) and (17) in Section 4.2 of Chapter 4 (see also Definition 3 of Chapter 1). As in typical threshold exceedance methods for extremes, these are taken to be exact likelihoods for exceedances over high thresholds, defined in terms of the norm of the components $\mathbf{Y}$ and $\mathbf{V}$. The likelihood is defined such that it does not depend on $\boldsymbol{\theta}$, the parameter of interest, below these thresholds. At the E step of the algorithm, the expectation of this likelihood is taken with respect to the distribution of $[\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}]$ implied by the chosen $\mathbf{Y}$ and $\mathbf{V}$ distributions over their entire supports. The maximization at the M step will then be taken only over the tails of $\mathbf{Y}$ and $\mathbf{V}$. Below we introduce the components of (32) in this modified setup and show that it obtains the maximum likelihood estimate of $\boldsymbol{\theta}$.

The E step of the algorithm computes the function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$, the expectation of the complete data log-likelihood. The regular variation conditions (16) and (17) imply that points of suitably normalized realizations of $\mathbf{Z}$ will converge on $\mathfrak{C}$ and $\mathfrak{C}_0$ to inhomogeneous Poisson point processes with intensity measures ν and ν_0, respectively (Resnick, 2007, 2002). Equivalent convergences for $\mathbf{Y}$ and $\mathbf{V}$ follow directly from their definitions, and the likelihood employed here arises from these limiting Poisson processes.

We assume parametric, continuously differentiable forms for the angular measure H and the hidden angular measure H_0 introduced in Chapter 4, with associated densities h and h_0. Let $\boldsymbol{\theta}$ be a parameter vector governing the tails of $\mathbf{Y}$ and $\mathbf{V}$. Fix thresholds $r^*_{\mathbf{Y}}$ and $r^*_{\mathbf{V}}$. Define radial and angular components of $\mathbf{Y}$ as $r_{\mathbf{Y}} = \|\mathbf{Y}\|$ and $\mathbf{w}_{\mathbf{Y}} = \|\mathbf{Y}\|^{-1}\mathbf{Y}$, where $\|\cdot\|$ is the L_1 norm, and define radial and angular components analogously for $\mathbf{V}$.

Assume we were to observe n iid realizations $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ and $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$. Following, e.g., (Coles and Tawn, 1991; Cooley et al., 2010; Ballani and Schlather, 2011), we assume the limiting Poisson point process models following from (16) and (17) are valid for realizations of $\mathbf{Y}$ and $\mathbf{V}$ with norms exceeding the thresholds $r^*_{\mathbf{Y}}$ and $r^*_{\mathbf{V}}$, respectively. The log-likelihood

for 'large' realizations of $\mathbf{Y}$ can be written

$$\ell_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{y}_1, ..., \mathbf{y}_n) = \sum_{i=1}^{n} \log \left\{ \alpha \left(\frac{r_{\mathbf{y}_i}}{b(n)} \right)^{-1-\alpha} h(\mathbf{w}_{\mathbf{y}_i}; \boldsymbol{\theta}) \right\} I_{\{r_{\mathbf{y}_i} > r^*_{\mathbf{Y}}\}} - \left(\frac{r^*_{\mathbf{Y}}}{b(n)} \right)^{-\alpha} - \log \left\{ \left(\sum_{i=1}^{n} I_{\{r_{\mathbf{y}_i} > r^*_{\mathbf{Y}}\}} \right) ! \right\}, \tag{33}$$

where $I_{\{\cdot\}}$ is the indicator function. For the observed realizations of $\mathbf{V}$, the log-likelihood is

$$\ell_{\mathbf{V}}(\boldsymbol{\theta}; \mathbf{v}_1, ..., \mathbf{v}_n) = \sum_{i=1}^{n} \log \left\{ \alpha_0 \left(\frac{r_{\mathbf{v}_i}}{b_0(n)} \right)^{-1-\alpha_0} h_0(\mathbf{w}_{\mathbf{v}_i}; \boldsymbol{\theta}) \right\} I_{\{r_{\mathbf{v}_i} > r^*_{\mathbf{V}}\}} - \left(\frac{r^*_{\mathbf{V}}}{b_0(n)} \right)^{-\alpha_0} - \log \left\{ \left(\sum_{i=1}^{n} I_{\{r_{\mathbf{v}_i} > r^*_{\mathbf{V}}\}} \right) ! \right\}. \tag{34}$$

In practice, α is often assumed known due to choice of marginal distributions, while α_0, the tail index of $\mathbf{V}$, is likely to be a parameter of interest.

The Poisson point process likelihoods (33) and (34) admit parametric forms of the unconditional densities of $\mathbf{Y}$ and $\mathbf{V}$, given that these components exceed their respective radial component thresholds $r^*_{\mathbf{Y}}$ and $r^*_{\mathbf{V}}$. Denote these densities by $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ and $f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta})$. Consider densities $g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ and $g_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta})$ each defined on $[\mathbf{0}, \boldsymbol{\infty})$, which are tail equivalent to f; that is,

$$g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) \cong f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) \quad \text{for} \quad \|\mathbf{y}\| > r^*_{\mathbf{Y}}$$
$$g_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}) \cong f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}) \quad \text{for} \quad \|\mathbf{v}\| > r^*_{\mathbf{V}}.$$

A way to think of the above tail equivalence conditions is that 'large' realizations from g are approximately equivalent in law to realizations of f. Extend the complete likelihood below the chosen thresholds by defining

$$\ell_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{y}) = \log g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) \quad \text{for} \quad \|\mathbf{y}\| \leq r^*_{\mathbf{Y}}$$

$$\ell_{\mathbf{V}}(\boldsymbol{\theta}; \mathbf{v}) = \log g_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}^{(k)}) \quad \text{for} \quad \|\mathbf{v}\| \leq r^*_{\mathbf{V}},$$

noting these are functions of $\boldsymbol{\theta}^{(k)}$ but not of $\boldsymbol{\theta}$. By independence, the complete log-likelihood for $(\mathbf{Z}, \mathbf{Y}, \mathbf{V})$ is

$$\ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{y}, \mathbf{v}) = \ell_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{y}) + \ell_{\mathbf{V}}(\boldsymbol{\theta}; \mathbf{v}) + \log I_{\{\mathbf{z}_i = \mathbf{y}_i + \mathbf{v}_i, i=1,\ldots,n\}}. \tag{35}$$

At the M step of the algorithm, we maximize the expected value of this likelihood with respect to $\boldsymbol{\theta}$.

The conditional distribution with respect to which we take the expectation of the complete log-likelihood at the E step is only assumed to be tail equivalent to the true distributions of $\mathbf{Y}$ and $\mathbf{V}$. Specifically, given a value of the parameter $\boldsymbol{\theta}^{(k)}$, we construct the conditional distribution

$$\begin{aligned} g(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) &= \frac{g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) g_{\mathbf{V}}(\mathbf{z} - \mathbf{y}; \boldsymbol{\theta}^{(k)})}{\int g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) g_{\mathbf{V}}(\mathbf{z} - \mathbf{y}; \boldsymbol{\theta}^{(k)}) d\mathbf{y}} \\ &\propto g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) g_{\mathbf{V}}(\mathbf{z} - \mathbf{y}; \boldsymbol{\theta}^{(k)}), \end{aligned} \tag{36}$$

where the densities g are given above. While this conditional density is defined on the entire supports of $\mathbf{Y}$ and $\mathbf{V}$, its crucial property is its tail equivalence to the densities of interest $f_{\mathbf{Y}}$ and $f_{\mathbf{V}}$. Define the objective function

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \int \ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v}, \mathbf{z}) g(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) d\mathbf{y} d\mathbf{v}. \tag{37}$$

In standard EM problems, the maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained because the function $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ in (32) attains its maximum at $\boldsymbol{\theta}^{(k)}$. In the modified setup here, it remains to show that this property holds. Let the densities $f_{\mathbf{Y}}$, $f_{\mathbf{V}}$, $g_{\mathbf{Y}}$, $g_{\mathbf{V}}$ be as defined

above. The complete likelihood function of $(\mathbf{z}, \mathbf{y}, \mathbf{v})$ is

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v}, \mathbf{z}) = I_{\{\mathbf{y}+\mathbf{v}=\mathbf{z}\}} \times \big\{ & f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}) I_{\{\|\mathbf{y}\|>r^*_{\mathbf{Y}}, \|\mathbf{v}\|>r^*_{\mathbf{V}}\}} \\ & + f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) g_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}^{(k)}) I_{\{\|\mathbf{y}\|>r^*_{\mathbf{Y}}, \|\mathbf{v}\|\le r^*_{\mathbf{V}}\}} \\ & + g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}) I_{\{\|\mathbf{y}\|\le r^*_{\mathbf{Y}}, \|\mathbf{v}\|>r^*_{\mathbf{V}}\}} \\ & + \ g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) g_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}^{(k)}) I_{\{\|\mathbf{y}\|\le r^*_{\mathbf{Y}}, \|\mathbf{v}\|\le r^*_{\mathbf{V}}\}} \big\}. \end{aligned}$$

Define the conditional likelihood $L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v} \mid \mathbf{z}) = L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v}, \mathbf{z}) / L(\boldsymbol{\theta}; \mathbf{z})$, where

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{z}) = & \int f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) f_{\mathbf{V}}(\mathbf{z}-\mathbf{y}; \boldsymbol{\theta}) I_{\{\|\mathbf{y}\|>r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|>r^*_{\mathbf{V}}\}} d\mathbf{y} \\ & + \int f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) g_{\mathbf{V}}(\mathbf{z}-\mathbf{y}; \boldsymbol{\theta}^{(k)}) I_{\{\|\mathbf{y}\|>r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|\le r^*_{\mathbf{V}}\}} d\mathbf{y} \\ & + \int g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) f_{\mathbf{V}}(\mathbf{z}-\mathbf{y}; \boldsymbol{\theta}) I_{\{\|\mathbf{y}\|\le r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|>r^*_{\mathbf{V}}\}} d\mathbf{y} \\ & + \int g_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}^{(k)}) g_{\mathbf{V}}(\mathbf{z}-\mathbf{y}; \boldsymbol{\theta}^{(k)}) I_{\{\|\mathbf{y}\|\le r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|\le r^*_{\mathbf{V}}\}} d\mathbf{y}. \end{aligned} \tag{38}$$

Define

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \int \log\{L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v} | \mathbf{z})\} g(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) d\mathbf{y} d\mathbf{v},$$

and for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, consider

$$\begin{aligned} H(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) &= \int \log \left\{ \frac{L(\boldsymbol{\theta}^{(k)}; \mathbf{y}, \mathbf{v} \mid \mathbf{z})}{L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v} \mid \mathbf{z})} \right\} g(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) d\mathbf{y} d\mathbf{v} \\ &\ge -\log \left\{ \int \frac{L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v} \mid \mathbf{z})}{L(\boldsymbol{\theta}^{(k)}; \mathbf{y}, \mathbf{v} \mid \mathbf{z})} g(\mathbf{y}, \mathbf{v} \mid \mathbf{z}; \boldsymbol{\theta}^{(k)}) d\mathbf{y} d\mathbf{v} \right\}, \end{aligned}$$

which follows from Jensen's inequality. Assuming equality of f and g above thresholds, the integral can be written

$$\frac{L(\boldsymbol{\theta}^{(k)}; \mathbf{z})}{L(\boldsymbol{\theta}; \mathbf{z}) g(\mathbf{z}; \boldsymbol{\theta}^{(k)})} \times \left\{ \int f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) f_{\mathbf{V}}(\mathbf{z}-\mathbf{y}; \boldsymbol{\theta}) I_{\{\|\mathbf{y}\|>r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|>r^*_{\mathbf{V}}\}} d\mathbf{y} \right.$$

$$+ \int f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) g_{\mathbf{V}}(\mathbf{z}-\mathbf{y};\boldsymbol{\theta}^{(k)}) I_{\{\|\mathbf{y}\|>r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|\le r^*_{\mathbf{V}}\}} d\mathbf{y}$$
$$+ \int g_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}^{(k)}) f_{\mathbf{V}}(\mathbf{z}-\mathbf{y};\boldsymbol{\theta}) I_{\{\|\mathbf{y}\|\le r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|> r^*_{\mathbf{V}}\}} d\mathbf{y}$$
$$+ \int g_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}^{(k)}) g_{\mathbf{V}}(\mathbf{z}-\mathbf{y};\boldsymbol{\theta}^{(k)}) I_{\{\|\mathbf{y}\|\le r^*_{\mathbf{Y}}, \|\mathbf{z}-\mathbf{y}\|\le r^*_{\mathbf{V}}\}} d\mathbf{y} \Bigg\}.$$

It follows from (38) that $L(\boldsymbol{\theta}^{(k)};\mathbf{z}) = g(\mathbf{z};\boldsymbol{\theta}^{(k)})$, and the sum of the four integrals above is $L(\boldsymbol{\theta};\mathbf{z})$ by definition. Thus $H(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) \ge 0$.

5.3.3 Implementation via Monte Carlo Sampling

In the modified expectation–maximization algorithm proposed here, the integral (37) is generally not available in closed form. In such cases, Wei and Tanner (1990) propose the use of Monte Carlo integration to approximate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$. Instead of computing Q directly, at the E step of the algorithm we construct

$$\hat{Q}_m(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \frac{1}{m} \sum_{j=1}^{m} \ell(\boldsymbol{\theta};\mathbf{z},\mathbf{y}_j,\mathbf{v}_j), \tag{39}$$

where the $\{(\mathbf{y}_j,\mathbf{v}_j)\}$ are independent draws from $g(\mathbf{y},\mathbf{v} \mid \mathbf{z},\boldsymbol{\theta}^{(k)})$ and the likelihood, defined in (35), employs only large simulated realizations of $\mathbf{Y}$ and $\mathbf{V}$. At the M step, we find

$$\boldsymbol{\theta}^{(k+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \hat{Q}_m(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$$

and return to the E step. The cycle is repeated until convergence. The estimated $\hat{Q}_m$ is not guaranteed to increase from each iteration to the next; however, the procedure still converges to the maximum likelihood value (Booth and Hobert, 1999).

The choice of the Monte Carlo sample size m in (39) involves a trade-off between the precision of the approximation of Q and the computational burden. Booth and Hobert (1999) proposed an automated strategy which uses small values of m at the start of the algorithm and increases m as the algorithm nears convergence. Specifically, they provide a formula for

the Monte Carlo error of the estimated parameter values, and increase the value of m by m/r if a $(1-\alpha)100\%$ confidence interval for $\boldsymbol{\theta}^{(k+1)}$ contains $\boldsymbol{\theta}^{(k)}$. Suggested values are $r = 3$ and $\alpha = 0.25$.

A number of stopping strategies exist for both deterministic and Monte Carlo EM (Givens and Hoeting, 2005). A common strategy is to stop the algorithm when

$$\max_i \left(\frac{|\theta_i^{(k+1)} - \theta_i^{(k)}|}{|\theta_i^{(k)}| + \delta_1} \right) < \delta_2 \tag{40}$$

for predetermined constants δ_1 and δ_2. For the Monte Carlo version, Booth and Hobert (1999) suggest using $\delta_2 = 0.002$ or 0.005. In this work, we will employ the automated strategy of Booth and Hobert (1999) using the above stopping criterion and strategy for increasing m.

We employ Louis' method (Louis, 1982) for uncertainty estimation about estimates of $\boldsymbol{\theta}$. This approach uses the 'missing information principle' to rewrite the negative Hessian of the observed log-likelihood as $\hat{I}_{\mathbf{Z}}(\boldsymbol{\theta}) = \hat{I}_{\mathbf{Z},\mathbf{Y},\mathbf{V}}(\boldsymbol{\theta}) - \hat{I}_{\mathbf{Y},\mathbf{V}|\mathbf{Z}}(\boldsymbol{\theta})$, where

$$\hat{I}_{\mathbf{Z},\mathbf{Y},\mathbf{V}}(\boldsymbol{\theta}) = -Q''(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}} \quad \text{and} \quad \hat{I}_{\mathbf{Y},\mathbf{V}|\mathbf{Z}}(\boldsymbol{\theta}) = \text{var}\left\{ \frac{d \log L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{v} \mid \mathbf{z})}{d\boldsymbol{\theta}} \right\}.$$

After we determine our algorithm has converged, we use the numerically estimated Hessian of $-\hat{Q}_m$ to estimate $I_{\mathbf{Z},\mathbf{Y},\mathbf{V}}(\hat{\boldsymbol{\theta}})$. Using the m simulated 'complete' datasets, we then use the sample variance of the

$$\left\{ \frac{d \log L(\boldsymbol{\theta}; \mathbf{y}_j, \mathbf{v}_j \mid \mathbf{z})}{d\boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}$$

as an estimate of $I_{\mathbf{Y},\mathbf{V}|\mathbf{Z}}(\hat{\boldsymbol{\theta}})$. Confidence intervals for parameters are constructed via a normal approximation.

5.4 Simulation Studies

5.4.1 Finite Hidden Measure Example

We begin by applying the proposed methodology of Section 5.3 to simulated data of dimension $d = 2$ which exhibit asymptotic independence and hidden regular variation with finite hidden angular measure. We generate n independent realizations of $\mathbf{Z}_{sim} = \mathbf{Y}_{sim} + \mathbf{V}_{sim}$, where $\mathbf{Y}_{sim} = [RW, R(1-W)]^T$ and $\mathbf{V}_{sim} = [R_0W_0, R_0(1-W_0)]^T$, with

$$F_R(r) = 2/r,\ r > 1 \qquad W \sim Bernoulli(0.5)$$
$$F_{R_0}(r) = r^{-1/\eta},\ r > 1 \qquad W_0 \sim H_0(\cdot),$$

all mutually independent, with H_0 the integrated measure density associated with the bivariate logistic dependence model (Gumbel, 1960). The angular density of this dependence model is given by

$$h_0(w;\beta) = \frac{1}{2}\left(\frac{1}{\beta} - 1\right)\{w(1-w)\}^{-1-1/\beta}\{w^{-1/\beta} + (1-w)^{-1/\beta}\}^{\beta-2},$$

for $\beta \in (0,1)$. As $\beta \to 1$, h_0 degenerates to point masses at $w = 0$ and $w = 1$, while the limiting case $\beta \to 0$ corresponds to a single point mass at $w = 1/2$. We fix $\eta = 0.75$ and assume it is known, and we aim to estimate β via the proposed estimation procedure.

While the full density of $\mathbf{Z}_{sim}$ could be written as a convolution in this case, we aim to study the effects of misspecification of the model for non-extreme realizations of $\mathbf{Y}_{sim}$ and $\mathbf{V}_{sim}$. Although the true radial component densities of $\mathbf{Y}_{sim}$ and $\mathbf{V}_{sim}$ follow Pareto distributions, here we let $g_{\mathbf{Y}_{mod}}$ and $g_{\mathbf{V}_{mod}}$ be densities associated with the same angular component distributions as in $\mathbf{Y}_{sim}$ and $\mathbf{V}_{sim}$, but Fréchet-distributed radial components $\|\mathbf{Y}_{mod}\|$ and $\|\mathbf{V}_{mod}\|$ with scale parameters $2, 1$ and shape parameters $1, 1/\eta$, respectively. These densities differ from true densities of $\mathbf{Y}_{sim}$ and $\mathbf{V}_{sim}$ for small $\|\mathbf{y}\|$ and $\|\mathbf{v}\|$ but rapidly converge to these as the magnitude grows, despite having differing supports.

To illustrate the performance of the proposed algorithm over different parameter values, threshold settings, and sample sizes, we perform the estimation scheme on 500 replications of n realizations of $\mathbf{Z}_{sim}$, with three different settings:

1. $n = 2500$, $\beta = 0.5$, $r^*_{\mathbf{Y}} = 18.98$, $r^*_{\mathbf{V}} = 5.41$
2. $n = 5000$, $\beta = 0.7$, $r^*_{\mathbf{Y}} = 38.99$, $r^*_{\mathbf{V}} = 9.28$
3. $n = 10000$, $\beta = 0.25$, $r^*_{\mathbf{Y}} = 199.00$, $r^*_{\mathbf{V}} = 31.50$

The dependence parameters chosen signify moderate, weak, and strong 'hidden' tail dependence, respectively, while the thresholds chosen correspond to the 0.9, 0.95, and 0.99 theoretical quantiles of the imposed Fréchet distributions of radial components. In each case, we choose an initial value of $m = 250$ for the number of Monte Carlo replications, and implement the scheme of Booth and Hobert (1999) for increasing m as described in Section 5.3.3 with $\alpha = 0.25$ and $r = 3$. We determine the algorithm has converged when criterion (40) is met for three successive iterations, with $\delta_1 = 0.001$ and $\delta_2 = 0.002$. Initial values $\beta^{(0)} \sim U$, with U following a uniform distribution centered at β and of widths 0.5, 0.4, and 0.3 for settings 1, 2, and 3, respectively. Simulations were performed using R on the Lynx computing system at the National Center for Atmospheric Research in Boulder, CO USA.

Table 5.1 shows mean parameter estimates, their root mean square errors, coverage rates of 95% confidence intervals constructed via a normal approximation, and the median number of iterations needed to obtain convergence for each simulation scenario. In each case, the algorithm converged relatively quickly, with median number of iterations of 20, 12, and 18, respectively. We note the bias in the estimates of β from the EM procedure due to the misspecification of the model, which is largest in setting 3, corresponding to strong tail dependence in the $\mathbf{V}_{sim}$ component. Further examination found that this bias is most severe when β is close to 0 or 1; that is, near the limiting hidden dependence cases. This bias was reduced by choosing a higher threshold; however, in small samples relatively low thresholds must be chosen to reduce uncertainty. Confidence intervals constructed via Louis'

Table 5.1: Mean parameter estimates, root mean square errors, 95% confidence interval coverage rates, and median number of iterations for proposed estimation procedure applied to 500 repetitions of simulated data.

Setting	β	$\hat{\beta}_{MCEM}$	RMSE	Coverage	k_{med}
1	0.50	0.520	0.031	94.4%	20
2	0.70	0.670	0.033	77.6%	12
3	0.25	0.288	0.062	70.2%	18

method were somewhat anticonservative in all cases, with coverage rates decreasing as the bias increases.

5.4.2 Infinite Hidden Measure

We now illustrate the procedure on the bivariate Gaussian example described in Section 4.6 of Chapter 4; our aim here is to estimate η, which describes the hidden regular variation in this case. This is a special case of the more general model (30) introduced by Ledford and Tawn (1997), for which the hidden regular variation is specified by condition (25) in Chapter 4, with $b_0(t)$ a function which is regular varying of order $\eta \in (1/2, 1]$. For the Gaussian example, the function $g_*(w)$ in (31) is identically one for all $w \in (0, 1)$, and Ramos and Ledford (2011) refer to the general case where $g_*(w)$ is constant as the *ray independence* case. Ramos and Ledford (2011) note that their modeling approach was unable to accommodate this situation when $\eta \in (1/2, 1]$. Here, we directly specify this model for the $\mathbf{V}$ component of our sum representation on the restricted subcone $\mathfrak{C}_0^\epsilon$.

We employ $\mathbf{Z}_{sum}$ defined in Section 4.6 of Chapter 4, with slight modifications. In order to more closely match realizations of the bivariate Gaussian dependence structure with Fréchet marginals, we choose the radial components of $\mathbf{Y}$ and $\mathbf{V}$ to be Fréchet-distributed, rather than exactly Pareto. In addition, we drop the slowly varying function $\mathcal{L}$ and define the scale parameter of the distribution of $\|\mathbf{V}\|$, $d_{\epsilon,n} = d_\epsilon = H_0(\mathcal{N}_0^\epsilon)$. The resulting $\mathbf{Z}_{sum}$ is tail equivalent on $\mathfrak{C}$ and $\mathfrak{C}_0^\epsilon$ to $\mathbf{Z}$, a random vector with Fréchet marginals and Gaussian dependence structure. We employ the Poisson point process likelihoods (33) and (34) and

conditional density (36) in the EM context to estimate η. Note that in this setup, $\eta = 1/\alpha_0$ is a parameter of the model for both the radial component and the angular component of $\mathbf{V}$.

In addition to the choice of thresholds $r^*_{\mathbf{Y}}$ and $r^*_{\mathbf{V}}$, in this case one must also choose ϵ, which restricts the support of $\mathbf{V}$ to a compact subcone. As described in Section 4.6.3 of Chapter 4, the choice of ϵ involves a trade-off between bias in the resulting limit measure, and the threshold level at which the sum representation is a useful approximation to the random vector $\mathbf{Z}$. In estimation problems, the latter is of critical importance, as one has fewer observations upon which to make inference as the threshold increases. To illustrate this, we examine a range of values for $r^*_{\mathbf{V}}$ and ϵ. In each case, we initialize the algorithm at $\eta^{(0)} = 0.75$; it was found that the algorithm displayed no sensitivity to starting values. We begin with the Monte Carlo sample size $m = 100$ and increase as needed via the Booth and Hobert (1999) procedure.

Table 5.2 shows results of our estimation procedure applied to replicates of $n = 10000$ realizations from the bivariate normal dependence structure with correlations $\rho = 0.2, 0.5, 0.9$. The values in each cell are mean estimates of η, root mean square errors, and coverage rates of 95% confidence intervals. Values were computed from 200 replications, and replicates across different cells were independent. Clearly, mean estimates of η are dependent on the choice of ϵ, with estimates of η increasing as we increase ϵ. This dependence appears to be reduced as the chosen threshold is increased. In general, increasing the threshold $r^*_{\mathbf{V}}$ appears to reduce the bias in estimation of η. For any chosen ϵ, confidence interval coverage rates improve as $r^*_{\mathbf{V}}$ is increased. Coverage rates for the largest threshold setting when $\rho = 0.2$ were conservative, perhaps because the estimation procedure employs very few data points, and the normal approximation may not be valid.

The results in Table 5.2 illustrate the tradeoff involved in the choice of ϵ and $r^*_{\mathbf{V}}$. While it is preferable to choose a very high threshold and small ϵ, in finite-sample estimation problems the analyst is forced to choose a lower threshold to reduce uncertainty in the estimation prcedure. The scaling factor in (29) must then be reduced by increasing ϵ in order to obtain

Table 5.2: Summary of results of estimation procedure applied to 200 replicates of $n = 10000$ data points from a bivariate random vector with unit Fréchet margins and bivariate Gaussian dependence structure with correlations $\rho = 0.2, 0.5, 0.8$. Values in each cell are (top to bottom) mean estimate of η, root mean square error, and 95% confidence interval coverage based on a normal approximation.

	$\eta = 0.6$ ($\rho = 0.2$)				$\eta = 0.75$ ($\rho = 0.5$)				$\eta = 0.9$ ($\rho = 0.8$)			
$\epsilon \setminus r^*_{\mathbf{V}}$	22	45	100	200	22	45	100	200	22	45	100	200
	0.532	0.565	0.578	0.597	0.671	0.704	0.721	0.725	0.880	0.896	0.900	0.895
0.185	0.069	0.039	0.033	0.025	0.080	0.050	0.035	0.038	0.028	0.020	0.023	0.031
	0.5%	43%	86.5%	99%	0%	26%	80.5%	88.5%	82.5%	94%	96.5%	93%
	0.542	0.571	0.584	0.596	0.688	0.713	0.725	0.729	0.906	0.912	0.917	0.909
0.20	0.060	0.033	0.032	0.027	0.063	0.041	0.033	0.036	0.022	0.025	0.031	0.030
	0.5%	62%	89.5%	99%	2.5%	48.5%	82%	92%	96%	91.5%	91%	96.5%
	0.552	0.576	0.591	0.603	0.705	0.727	0.736	0.737	0.930	0.933	0.929	0.920
0.215	0.049	0.029	0.027	0.028	0.047	0.030	0.027	0.032	0.036	0.039	0.040	0.037
	6%	73.5%	95.5%	99%	24%	72.5%	92%	94.5%	79%	76%	79.5%	91.5%
	0.565	0.587	0.592	0.608	0.724	0.735	0.742	0.742	0.957	0.954	0.940	0.931
0.23	0.037	0.022	0.027	0.030	0.031	0.025	0.027	0.034	0.060	0.059	0.049	0.044
	30%	87.5%	96%	98%	68.5%	85%	91.5%	94.5%	34.5%	43%	73%	85.5%

a model $\mathbf{Y} + \mathbf{V}$ which is a useful approximation at the chosen threshold level. This difficulty is not unique to the methodology proposed here; the estimation method of Ledford and Tawn (1996) found bias in estimation of η as well. In practice, a range of ϵ and $r^*_{\mathbf{V}}$ values can be examined, and the simulation results from the Gaussian example in Table 5.2 can be used as a guide for their selection.

5.5 Application: Air Pollution Data

5.5.1 Exploratory Analysis

We apply the proposed MCEM methodology to data on air pollution measurements from a monitoring station at Leeds city centre, UK. Heffernan and Tawn (2004) analyze the tail behavior of five pollutants for the years 1994-1998; here, we restrict our analysis to nitrogen dioxide and sulfur dioxide, denoted by the bivariate pair (NO_2, SO_2). We examine daily data from the winter months (November - February) from January 1, 1994 to December 11,

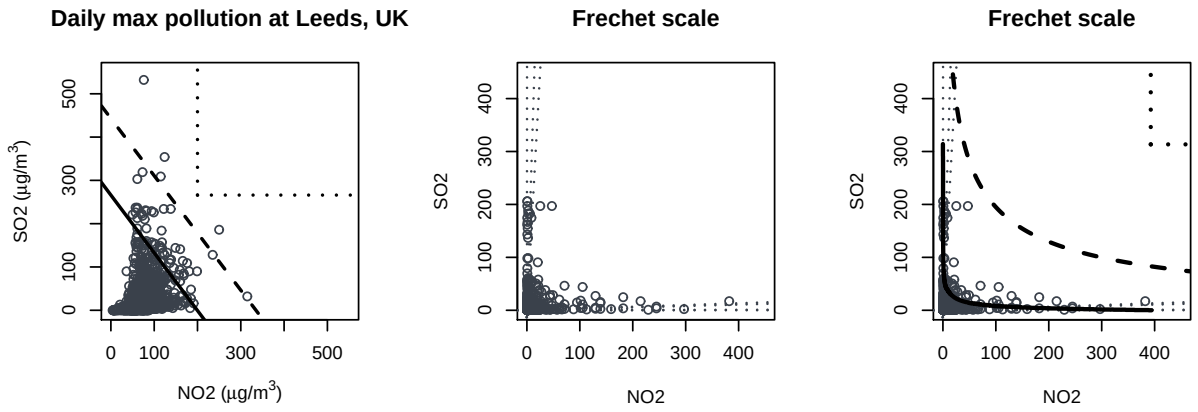

Figure 5.1: Left: daily maximum air pollutant measurements for winter months 1994-2012 at Leeds city centre. Center and right: after transformation to Fréchet scale. Dashed lines are drawn to points outside the plotting window. Left and right plots show boundaries of risk regions A_1 (solid), A_2 (dashes), and A_3 (dots).

2012. Days for which the record was incomplete were excluded; this resulted in a sample of $n = 1988$ observations. The data exhibit some temporal dependence due to short-term persistence of weather patterns; we do not attempt to account for this dependence in this analysis.

A plot of the bivariate data is given in the left panel of Figure 5.1. There is clear dependence in the data, as we estimate the Spearman correlation to be 0.42; furthermore, this dependence appears to extend somewhat into the joint tail. Asymptotic independence is demonstrated, as the very largest values of NO_2 and SO_2 do not tend to occur simultaneously. A conclusion of asymptotic independence of this pair was also reached by Heffernan and Tawn (2004).

5.5.2 Marginal Estimation

A first step in describing tail dependence is to account for marginal effects, and we accomplish this by fitting a generalized Pareto distribution (GPD) to the tails of each margin separately. After examining diagnostics (Coles, 2001, Chapter 4), we select thresholds corresponding to the 0.93 empirical quantile of each margin, and fit the GPD via maximum

Table 5.3: Threshold selected and maximum-likelihood parameter estimates (standard errors) from generalized Pareto distributions fit to exceedances of the 0.93 quantiles of pollution measurements.

Margin	Threshold	$\hat{\psi}$ (se)	$\hat{\xi}$ (se)
NO_2	109	24.07 (3.03)	0.09 (0.09)
SO_2	80	51.82 (6.63)	0.10 (0.10)

likelihood. A summary of parameter estimates is given in Table 5.3. The shape parameter for SO_2 is similar to that obtained by Heffernan and Tawn (2004) using data from 1994-1998, while for NO_2 our shape parameter estimate is larger than that reported in Heffernan and Tawn (2004) (-0.03).

5.5.3 Dependence Modeling

To model tail dependence, we apply probability integral transformations to the data, using the fitted GPD above the selected thresholds, and the empirical distribution function below (Coles and Tawn, 1991), so that each margin follows a unit Fréchet distribution. A plot of the transformed data is given in the middle panel of Figure 5.1. Note that the largest values in each margin fall near the axes of the plot, providing further evidence of asymptotic independence of the two quantities. We proceed assuming this bivariate random vector, which we denote by $\mathbf{Z}$, is regular varying, and we explore the following three approaches to modeling the tail dependence in these data:

1. Assume asymptotic dependence is present; that is, the resulting first-order limiting measure ν has mass on the full interior of $\mathfrak{C}$. We fit the bivariate logistic model of Gumbel (1960) to model the tail dependence.

2. Assume asymptotic independence holds, and disregard any hidden regular variation.

3. Assume asymptotic independence and hidden regular variation with an infinite hidden measure of the form (25); estimate η through the ϵ-restricted sum representation introduced in Section 4.6 of Chapter 4.

We assess the three approaches through their estimation of three risk set probabilities. Define the region A_1 by the set of (NO_2, SO_2) such that $SO_2 + (4/3)NO_2 > 266$. When $SO_2 = 0$, this region corresponds to a level of NO_2 in Index Band 4 (of 10) or higher on the air quality scale defined by the UK Department for Environment, Food and Rural Affairs[9]. Similarly, when $NO_2 = 0$, the region A_1 corresponds to SO_2 levels falling in Index Band 4 or higher. At these marginal levels, recommendations are issued to at-risk groups to reduce strenuous outdoor activity. In addition to capturing hazardous levels of the individual pollutants, this region also captures events with large observations of both pollutants. Studies in the medical and phytology literatures have found enhanced negative effects of exposure to a large combination of these two pollutants on both plants (Wellburn et al., 2006) and at-risk humans (Devalia et al., 1994). We also examine a more extreme region A_2, for which $SO_2 + (443/335)NO_2 > 443$. This region has the same interpretation as A_1, replacing Index Band 4 by Index Band 6. As a goal of an extreme value analysis is often extrapolation, we examine a third risk set in which no data were observed in the time period studied. Define the region $A_3 = \{(NO_2, SO_2) : NO_2 > 200, SO_2 > 266\}$. This region corresponds to levels of both pollutants falling in Index Band 4 or higher simultaneously. The boundaries of the three regions are shown on the original scale, and after transformation to the Fréchet scale, in Figure 5.1. There were 58 and 5 observations in the regions A_1 and A_2, respectively, among the $n = 1988$ observations in the study period. Finally, we note that the methods introduced by Ledford and Tawn (1997) and Ramos and Ledford (2009) are unable to estimate probabilities of sets A_1 and A_2, as they provide models only for regions of the form $(\mathbf{z}, \boldsymbol{\infty}]$. We first describe details of the fitting procedure for each of the three approaches; estimates of the risk set probabilities obtained from each are then compared.

To implement modeling approach 1, we employ the techniques of previous works (Coles and Tawn, 1991; Cooley et al., 2010; Ballani and Schlather, 2011) and transform to radial and angular components under the L_1 norm. We then take the regular variation limit condition

[9] `http://uk-air.defra.gov.uk/air-pollution/daqi`

(16) as an equality for points exceeding a high radial component threshold, and estimate the parameter β of the bivariate logistic angular measure via numerical maximization of a Poisson point process likelihood. Through exploratory analysis, it was found that estimates $\hat{\beta} \to 1$ as the threshold increases, indicating asymptotic independence. Nonetheless, we select a threshold corresponding to the 0.9 empirical quantile of radial component values, and we estimate $\hat{\beta} = 0.713$, indicating relatively weak extremal dependence in these data. We compute an associated 95% confidence interval for β of $(0.685, 0.742)$. The estimates $\hat{\mathbb{P}}(\mathbf{Z} \in A_j), j = 1, 2, 3$ can be computed from the fitted model via numerical integration.

As modeling approach 2 assumes asymptotic independence and no hidden regular variation, no dependence estimation procedure is needed. The probabilities of the risk sets A_1 and A_2 correspond to the sum of marginal probabilities which can be computed directly. The asymptotic independence assumption implies $\mathbb{P}(\mathbf{Z} \in A_3) = 0$.

As described in Section 5.4.2, implementation of approach 3 involves selection of both $r^*_{\mathbf{V}}$ and ϵ. While the relatively small sample size ($n = 1988$) here does not allow us to choose $r^*_{\mathbf{V}}$ as large as examined in the simulation study in Section 5.4.2, a range of choices for $r^*_{\mathbf{V}}$ and ϵ were examined. To maintain some comparability with modeling approach 1, we select $r^*_{\mathbf{V}} = 7.5$, which results in approximately 10% of observations being used in the estimation of η. We select $\epsilon = 0.3$; while this choice is somewhat *ad hoc*, it was chosen based on guidance provided by the simulation study in Section 5.4.2. We initialize the algorithm at $\eta^{(0)} = 0.75$ and $m = 100$. The algorithm converged after 6 iterations to an estimate of $\hat{\eta} = 0.748$. We employ Louis' method to compute a 95% confidence interval of $(0.645, 0.851)$ for this parameter. Risk set probabilities are estimated via simulation.

5.5.4 Results

Table 5.4 displays the estimates of $\mathbb{P}(\mathbf{Z} \in A_j), j = 1, 2, 3$ for each of the three modeling approaches, as well as the one-sided p-value of the observed data, assuming the estimated probability is the true probability. The reported p-values were computed via the binomial

Table 5.4: Risk set probability estimates and associated one-sided p-values of observed $\mathbf{Z} \in A_j, j = 1, 2, 3$, for modeling approaches 1-3 described in Section 5.5.

Model	$\hat{\mathbb{P}}(\mathbf{Z} \in A_1)$	p-val	$\hat{\mathbb{P}}(\mathbf{Z} \in A_2)$	p-val	$\hat{\mathbb{P}}(\mathbf{Z} \in A_3)$	p-val
1	0.0297	0.480	0.0044	0.132	0.0010	0.130
2	0.0120	8.17×10^{-5}	0.0002	0.009	0	1
3	0.0261	0.210	0.0018	0.274	0.0002	0.704
(empirical)	0.0292	–	0.0025	–	0	–

distribution; for example, approach 2 estimates $\hat{\mathbb{P}}(\mathbf{Z} \in A_1) = 0.0120$, implying an expected 23.86 observations in a sample of $n = 1988$. The p-value reported is the binomial probability of observing 58 or more of the 1988 realizations in the set A_1, if the true probability is 0.0120. We see that approach 1, assuming asymptotic dependence, estimates $\mathbb{P}(\mathbf{Z} \in A_1)$ quite well. This approach overestimates the probability of the more extreme set A_2, although the overestimation is not severe. It estimates $\mathbb{P}(\mathbf{Z} \in A_3)$ to be 0.0010; however, zero of the 1988 observations fall in this set, which has a p-value of 0.13 if this were the true probability. Approach 2 suffers from the opposite problem: it underestimates the risk of sets A_1 and A_2, particularly the less extreme set A_1. Approach 2 assumes that $\mathbb{P}(\mathbf{Z} \in A_3) = 0$; while no observations were seen in the set A_3, this assumption may be unreasonable.

Modeling approach 3 based on the proposed $\mathbf{Y} + \mathbf{V}$ performs best, as it accounts for the hidden regular variation in the presence of asymptotic independence. The estimates of $\mathbb{P}(\mathbf{Z} \in A_1)$ and $\mathbb{P}(\mathbf{Z} \in A_2)$ are quite plausible given the observed data. Furthermore, this model estimates the event $\mathbf{Z} \in A_3$ to occur about once every 5000 observations, in which case the probability that zero such events would be observed in a sample of this size is 0.70.

The three approaches to risk set estimation described here illustrate the advantages of the $\mathbf{Y} + \mathbf{V}$ representation for estimating hidden regular variation. An approach which assumes asymptotic dependence when asymptotic independence is actually present will overestimate probabilities of extreme sets. On the other hand, a simple modeling approach assuming asymptotic independence will tend to underestimate tail risk sets at finite levels if hidden regular variation is present. The representation presented here allows flexibility to account for

hidden regular variation in the presence of asymptotic independence and provides reasonable estimates of risk set probabilities at high but observable levels, as well as extrapolation further into the joint tail.

5.6 Summary and Discussion

This chapter presents novel methodology for tail estimation in the presence of hidden regular variation, employing the sum characterization of hidden regular variation introduced in Chapter 4. The likelihood-based estimation scheme proposed here is a novel version of the MCEM algorithm which has been modified for tail estimation. The conditional distribution from which we sample for the E step of the algorithm needs only to be tail equivalent to the tail dependence structure which we wish to estimate. The likelihood which we maximize at the M step is a Poisson point process likelihood which follows from the regular variation conditions (16) and (17) in Chapter 4.

The examples in this chapter demonstrate the ability of our method to estimate tail dependence in the case of asymptotic independence. Using the proposed estimation strategy, we are able to estimate probabilities of general tail risk sets on the cone $\mathfrak{C}$. This approach differs from the joint tail estimation methods proposed by Ledford and Tawn (1997) and Ramos and Ledford (2009), which focus on estimation of probabilities of risk sets of the form $(\mathbf{z}, \boldsymbol{\infty}]$. In the bivariate infinite hidden measure case, we are able to obtain the ray independent hidden angular measure discussed in Section 4.6 of Chapter 4 on the restricted subcone $\mathfrak{C}_0^{\epsilon}$; the model of Ramos and Ledford (2009) is unable to accommodate this case, which includes the bivariate Gaussian tail dependence structure.

Through the application to air pollution data from Leeds, UK, we demonstrated the ability of the $\mathbf{Y} + \mathbf{V}$ representation to estimate risk sets at both observable levels as well as further extrapolation into the joint tail. In estimating these risk set probabilities, we are able to model the presence of both asymptotic independence and hidden regular variation.

The sum representation accounts for both the limiting regular variation structure and the 'residual tail dependence' in the data.

Finite-sample estimation of η (or $\alpha_0 = 1/\eta$) remains an inherently difficult problem. The estimation scheme using the ϵ-restricted infinite hidden angular measure representation presented in Section 4.6 is useful in that it employs both the radial and angular component of $\mathbf{V}$ to estimate η, as opposed to the Hill estimator (Hill, 1975) and the 'structure variable' approach of Ledford and Tawn (1996), which only use the radial component for estimation.

The focus of this work was estimation of parameters of the hidden regular variation represented by $\mathbf{V}$. In general, the EM approach employed here could also be used to estimate parameters of $\mathbf{Y}$, or of both $\mathbf{Y}$ and $\mathbf{V}$. Increasing complexity is encountered in estimating hidden regular variation in higher-dimensional problems (Mitra and Resnick, 2010), and implementing the methodology in such problems would likely require simplifying assumptions on the supports of $\mathbf{Y}$ and $\mathbf{V}$.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This dissertation presents applied, theoretical, and methodological advances in the statistical modeling of extreme values, employing multivariate regular variation as an underlying mathematical framework. This framework provides a probabilistic characterization of the joint tail of a random vector, is useful for describing multivariate threshold exceedances, and can be linked to classical results from extreme value theory.

Chapter 2 extended existing methodology based on the regular variation framework in a novel study in climate science. Focusing on a particular atmospheric phenomena, we modeled the tail dependence in daily precipitation amounts as produced by a deterministic climate simulation model and seen in observational data. For the first time, a connection was drawn between large-scale atmospheric processes and localized extreme precipitation events, and the fitted tail dependence model was employed to study uncertainties in future extreme precipitation events produced by a regional climate model, given output of a general circulation model which provides its boundary conditions.

Further exploration of regional climate simulation models' representation of extreme precipitation events was provided in Chapter 3. Focusing on a suite of six particular regional climate models, it was found that these models are able to adequately reproduce past observed extreme precipitation events over a Pacific coast region in winter. When examining summer precipitation over a central region of North America, however, it was found that the climate models produced poor representations of past observed precipitation extremes. Employing the regular variation framework, we were able to assess the ability of these deterministic models to simulate these weather extremes.

Chapter 4 introduced a new probabilistic characterization of hidden regular variation. This characterization is defined in arbitrary dimensions and for arbitrary hidden regular variation structures. It was proven via a tail equivalence theorem that this characterization satisfies the necessary asymptotic properties on all joint tail regions. Through simulation of a canonical example, we demonstrated that this characterization also possesses more realistic finite-sample properties than previous characterizations.

The finite-sample properties of the sum characterization make it amenable for inference, and novel methodology for inference from hidden regular variation structures was presented in Chapter 5. The expectation–maximization (EM) algorithm, a classical method in the field of Statistics, was employed for the first time in a tail estimation problem. Using a modified EM setup and Monte Carlo sampling, we are able to perform maximum likelihood inference for parameters governing hidden regular variation. The method was applied to air pollution data, with improved results over alternative tail modeling approaches.

As the climate science literature devotes increasing attention to the study of extremes, the work presented in Chapters 2 and 3 offers opportunities for further use of the regular variation framework to study multivariate extremes in climate and to explore connections between climate extremes and underlying atmospheric processes. For one example, regular variation might be employed to examine possible teleconnections: links between climate extremes at long distances. Multivariate extremes techniques might also be employed to study extremal dependence across space in climate.

The theoretical and methodological developments in Chapters 4 and 5 offer new advances in the statistical modeling of multivariate tails. However, it remains difficult to distinguish a weak level of asymptotic dependence from asymptotic independence with hidden regular variation in finite samples. The estimation procedure introduced in Chapter 5 may lead to the development of likelihood-based model selection criteria for extremes. Ramos and Ledford (2009) introduce a likelihood-based test of asymptotic independence; these methods might be incorporated into the estimation procedure presented in Chapter 5.

While the applications presented in this dissertation have focused on climate and environmental studies, the methodology employed here could be more broadly applied. The characterization introduced in Chapter 4 might be used to study hidden regular variation in returns of financial instruments. The methodology may be extended to time series or spatial settings as well.

Exciting challenges remain in the study of multivariate (and spatial) extremes. The growing availability of data offers more opportunities to study extremes and presents more challenging problems which demand the development of new statistical methodologies. Much work remains to be done.

REFERENCES

Allan, R. and Soden, B. (2008). Atmospheric warming and the amplification of precipitation extremes. *Science*, 321(5895):1481.

Anderson, C., Arritt, R., Pan, Z., Takle, E., Gutowski Jr, W., Otieno, F., da Silva, R., Caya, D., Christensen, J., Lüthi, D., et al. (2003). Hydrological processes in regional climate model simulations of the central United States flood of June-July 1993. *Journal of Hydrometeorology*, 4(3):584–598.

Balkema, A. and De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Ballani, F. and Schlather, M. (2011). A construction principle for multivariate extreme value distributions. *Biometrika*, 98(3):633–645.

Barnes, K. and Eash, D. (1994). Flood of June 17, 1990, in the Clear Creek basin, east-central Iowa. Technical report, US Geological Survey.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D. D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, New York.

Bingham, N., Goldie, C., and Teugels, J. (1989). *Regular variation*, volume 27. Cambridge Univ Pr.

Booth, J. and Hobert, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

Bukovsky, M. (2011). Masks for the Bukovsky regionalization of North America. Regional Integrated Sciences Collective, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO. Downloaded 2012-08-21.

Bukovsky, M. and Karoly, D. (2011). A regional modeling study of climate change impacts on warm-season precipitation in the central United States. *Journal of Climate*, 24(7):1985–2002.

Cannon, A., Whitfield, P., and Lord, E. (2002). Synoptic map-pattern classification using recursive partitioning and principal component analysis. *Monthly Weather Review*, 130(5):1187–1206.

Castello, A. and Shelton, M. (2004). Winter precipitation on the US Pacific coast and El Niño–Southern Oscillation events. *International Journal of Climatology*, 24(4):481–497.

Caya, D. and Laprise, R. (1999). A semi-implicit semi-Lagrangian regional climate model: The Canadian RCM. *Monthly Weather Review*, 127(3):341–362.

Cayan, D., Maurer, E., Dettinger, M., Tyree, M., and Hayhoe, K. (2008). Climate change scenarios for the California region. *Climatic Change*, 87(Suppl 1):21–42.

Chavez-Demoulin, V. and Davison, A. (2005). Generalized additive models for sample extremes. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 54(1):207–222.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analysis. *Extremes*, 2(4):339–365.

Coles, S. and Tawn, J. (1991). Modeling multivariate extreme events. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 53(2):377–92.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag London Ltd., London.

Collins, W., Bitz, C., Blackmon, M., Bonan, G., Bretherton, C., Carton, J., Chang, P., Doney, S., Hack, J., Henderson, T., et al. (2006). The community climate system model version 3 (CCSM3). *Journal of Climate*, 19(11):2122–2143.

Cook, K., Vizy, E., Launer, Z., and Patricola, C. (2008). Springtime intensification of the great plains low-level jet and midwest precipitation in GCM simulations of the 21st century. *Journal of Climate*, 21(23):6321–6340.

Cooley, D., Davis, R., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

Cooley, D., Davis, R., Naveau, P., and Gif-sur Yvette, F. (2012). Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *Annals of Applied Statistics*, 6(4):1406–1429.

Cooley, D., Naveau, P., and Poncet, P. (2006). Variograms for spatial max-stable random fields. In Bertail, P., Doukhan, P., and Soulier, P., editors, *Dependence in Probability and Statistics*, Springer Lecture Notes in Statistics. Springer, New York.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840.

Dalrymple, T. (1960). Flood frequency analyses. Water supply paper 1543-a, U.S. Geological Survey, Reston, VA.

Daly, C., Taylor, G., and Gibson, W. (1997). The PRISM approach to mapping precipitation and temperature. In *Proceedings, 10th AMS Conference on Applied Climatology*, pages 20–23.

Davis, C., Manning, K., Carbone, R., Trier, S., and Tuttle, J. (2003). Coherence of warm-season continental rainfall in numerical weather prediction models. *Monthly Weather Review*, 131(11):2667–2679.

Davis, R. and Mikosch, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4):977–1009.

Davis, R. and Resnick, S. (1993). Prediction of stationary max-stable processes. *Ann. of Applied Prob*, 3(2):497–525.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(3):393–442.

de Haan, L. (1970). *On regular variation and its application to the weak convergence of sample extremes*. Mathematisch Centrum.

de Haan, L. (1984). A spectral representation for max-stable processes. *Annals of Probability*, 12(4):1194–2004.

de Haan, L. and de Ronde, J. (1998). Sea and wind: Multivariate extremes at work. *Extremes*, 1(1):7–45.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

de Haan, L. and Resnick, S. (1977). Limit theory for multivariate sample extremes. *Probability Theory and Related Fields*, 40(4):317–337.

de Haan, L. and Zhou, C. (2011). Extreme residual dependence for random vectors and processes. *Advances in Applied Probability*, 43(1):217–242.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38.

Dettinger, M. (2004). Fifty-two years of pineapple-express storms across the west coast of North America. *US Geological Survey, Scripps Institution of Oceanography for the California Energy Commission, PIER Project Rep. CEC-500-2005-004*.

Dettinger, M. (2011). Climate change, atmospheric rivers, and floods in California–a multimodel analysis of storm frequency and magnitude changes. *JAWRA Journal of the American Water Resources Association*, 47(3):514–523.

Dettinger, M., Ralph, F., Das, T., Neiman, P., and Cayan, D. (2011). Atmospheric rivers, floods and the water resources of California. *Water*, 3(2):445–478.

Devalia, J., Rusznak, C., Herdman, M., Trigg, C., Davies, R., and Tarraf, H. (1994). Effect of nitrogen dioxide and sulphur dioxide on airway response of mild asthmatic patients to allergen inhalation. *The Lancet*, 344(8938):1668–1671.

Draisma, G., Drees, H., Ferreira, A., and De Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli*, 10(2):251–280.

Easterling, D., Meehl, G., Parmesan, C., S.A., C., Karl, T., and Mearns, L. (2000). Climate extremes: Observations, modeling, and impacts. *Science*, 289(5487):2068–2074.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *Annals of Statistics*, 29(5):1401–1423.

Einmahl, J., Krajina, A., and Segers, J. (2012). An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764–1793.

Einmahl, J. and Van den Akker, R. (2011). Superefficient estimation of the marginals by exploiting knowledge on the copula. *Journal of Multivariate Analysis*, 102(9):1315–1319.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, volume 33 of *Applications of Mathematics*. Springer-Verlag, Berlin.

Fisher, R. and Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Fougères, A., Nolan, J., and Rootzén, H. (2009). Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36(1):42–59.

Fougères, A. L. (2004). Multivariate extremes. In B. Finkenstadt and H. Rootzen, editor, *Extreme Values in Finance, Telecommunications and the Environment*, pages 373–388. Chapman and Hall CRC Press, London.

Frei, C., Scholl, R., Fukutome, S., Schmidli, J., and Vidale, P. L. (2006). Future change of precipitation extremes in Europe: Intercomparison of scenarios from regional climate models. *Journal of Geophysical Research*, 111(D6):D06105.

Fritsch, J., Kane, R., and Chelius, C. (1986). The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. *Journal of Climate and Applied Meteorology*, 25(10):1333–1345.

Giorgi, F., Marinucci, M., and Bates, G. (1993a). Development of a second-generation regional climate model (RegCM2). Part I: Boundary-layer and radiative transfer processes. *Monthly Weather Review*, 121(10):2794–2813.

Giorgi, F., Marinucci, M., Bates, G., and De Canio, G. (1993b). Development of a second-generation regional climate model (RegCM2). II: Convective processes and assimilation of lateral boundary conditions. *Monthly Weather Review*, 121(10):2814–2832.

Givens, G. and Hoeting, J. (2005). *Computational statistics*. John Wiley & Sons, New York.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453.

Grell, G., Dudhia, J., Stauffer, D., et al. (1994). A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). Technical report, Mesoscale and Microscale Meteorology Division, National Center for Atmospheric Research.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9:171–173.

Gutowski, W., Arritt, R., Kawazoe, S., Flory, D., Takle, E., Biner, S., Caya, D., Jones, R., Laprise, R., Leung, L., et al. (2010). Regional extreme monthly precipitation simulated by NARCCAP RCMs. *Journal of Hydrometeorology*, 11(6):1373–1379.

Guttorp, P. and Xu, J. (2011). Climate change, trends in extremes, and model assessment for a long temperature time series from Sweden. *Environmetrics*, 22(3):456–463.

Heffernan, J. and Resnick, S. (2007). Limit laws for random vectors with an extreme component. *The Annals of Applied Probability*, 17(2):537–571.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Holland, G. (2009). Climate change and extreme weather. In *IOP Conference Series: Earth and Environmental Science*, volume 6, page 092007. IOP Publishing.

Hosking, J., Wallis, J., and Wood, E. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted-moments. *Technometrics*, 27(3):251–261.

Hosking, J. R. M. and Wallis, J. R. (1997). *Regional Frequency Analysis: An approach based on L-Moments*. Cambridge, University Press, Cambridge., U.K.

Husler, J. and Li, D. (2009). Testing asymptotic independence in bivariate extremes. *Journal of Statistical Planning and Inference*, 139(3):990–998.

IPCC (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, Cambridge, UK, and New York, NY, USA. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor, and P.M. Midgley (eds.)].

Jessen, A. and Mikosch, T. (2006). Regularly varying functions. *Publications de l'Institut Mathématique*, 80(94):171–192.

Joe, H. (1997). *Multivariate models and dependence concepts*, volume 73. Chapman & Hall/CRC.

Jones, R., Hassell, D., Hudson, D., Wilson, S., Jenkins, G., and Mitchell, J. (2003). Workbook on generating high resolution climate change scenarios using PRECIS. *National Communications Support Unit Workbook*.

Juang, H., Hong, S., and Kanamitsu, M. (1997). The NCEP regional spectral model: An update. *Bulletin of the American Meteorological Society*, 78(10):2125–2143.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S., Hnilo, J., Fiorino, M., and Potter, G. (2002). NCEP-DOE AMIP-II reanalysis (r-2). *Bulletin of the American Meteorological Society*, 83(11):1631–1644.

Karl, T. and Melillo, J. (2009). *Global climate change impacts in the United States*. Cambridge Univ Pr.

Katz, R. (1999). Extreme value theory for precipitation: Sensitivity analysis for climate change. *Advances in Water Resources*, 23(2):133–139.

Katz, R. (2010). Statistics of extremes in climate change. *Climatic Change*, 100(1):71–76.

Kharin, V. and Zwiers, F. (2005). Estimating extremes in transient climate change simulations. *Journal of Climate*, 18(8):1156–1173.

Kharin, V. V. and Zwiers, F. W. (2000). Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. *Journal of Climate*, 13(21):3760–3788.

Ledford, A. and Tawn, J. (1997). Modelling depedence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):475–499.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Leung, L., Correia, J., and Qian, Y. (2011). Regional climate change scenarios for North America simulated by WRF driven by two global climate models. In *Proceedings of the Ninety-first American Meteorological Society Annual Meeting*, Seattle, WA USA.

Leung, L. and Qian, Y. (2009). Atmospheric rivers induced heavy precipitation and flooding in the western US simulated by the WRF regional climate model. *Geophysical Research Letters*, 36(3):L03820.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44(2):226–233.

Mailhot, A., Beauregard, I., Talbot, G., Caya, D., and Biner, S. (2011). Future changes in intense precipitation over canada assessed from multi-model NARCCAP ensemble simulations. *International Journal of Climatology*, 32(8):1151–1163.

Maulik, K. and Resnick, S. (2004). Characterizations and examples of hidden regular variation. *Extremes*, 7(1):31–67.

Maurer, E., Wood, A., Adam, J., Lettenmaier, D., and Nijssen, B. (2002). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of Climate*, 15(22):3237–3251.

Mearns, L., Arritt, R., Biner, S., Bukovsky, M., McGinnis, S., Sain, S., Caya, D., Correia Jr, J., Flory, D., Gutowski, W., et al. (2012). The North American regional climate change assessment program: Overview of phase I results. *Bulletin of the American Meteorological Society*, 93(9):1337–1362.

Mearns, L., Gutowski, W., Jones, R., Leung, R., McGinnis, S., Nunes, A., and Qian, Y. (2009). A regional climate change assessment program for North America. *Eos Trans. AGU*, 90(36):311.

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9(1):3–20.

Mitra, A. and Resnick, S. (2010). Hidden regular variation: Detection and estimation. *Arxiv preprint arXiv:1001.5058*.

Mo, K., Paegle, J., and Higgins, R. (1997). Atmospheric processes associated with summer floods and droughts in the central United States. *Journal of Climate*, 10(12):3028–3046.

Nakicenovic, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., Gaffin, S., Gregory, K., Grubler, A., Jung, T., Kram, T., et al. (2000). Special report on emissions scenarios: a special report of Working Group III of the Intergovernmental Panel on Climate Change. Technical report, Pacific Northwest National Laboratory, Richland, WA (US), Environmental Molecular Sciences Laboratory (US).

Nelsen, R. (2006). *An Introduction to Copulas*. Springer Verlag.

Pal, J., Giorgi, F., Bi, X., Elguindi, N., Solmon, F., Rauscher, S., Gao, X., Francisco, R., Zakey, A., Winter, J., et al. (2007). Regional climate modeling for the developing world: The ICTP RegCM3 and RegCNET. *Bulletin of the American Meteorological Society*, 88(9):1395–1409.

Peng, L. (1999). Estimation of the coefficient of tail dependence in bivariate extremes. *Statistics and Probability Letters*, 43(4):399–409.

Peterson, T., Stott, P., and Herring, S. (2012). Explaining extreme events of 2011 from a climate perspective. *Bulletin of the American Meteorological Society*, 93(7):1041–1067.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):219–241.

Ramos, A. and Ledford, A. (2011). An alternative point process framework for modeling multivariate extreme values. *Communications in Statistics – Theory and Methods*, 40(12):2205–2224.

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer-Verlag, New York.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5(4):303–336.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

Rootzen, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930.

Ropelewski, C. and Halpert, M. (1987). Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, 115(8):1606–1626.

Rougier, J. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, 81(3):247–264.

Sain, S., Nychka, D., and Mearns, L. (2011). Functional ANOVA and regional climate experiments: a statistical analysis of dynamic downscaling. *Environmetrics*, 22(6):700–711.

Sang, H. and Gelfand, A. E. (2010). Continuous spatial process models for spatial extreme values. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(1):49–65.

Schlather, M. and Tawn, J. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156.

Schliep, E., Cooley, D., Sain, S., and Hoeting, J. (2010). A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, 13(2):219–239.

Schubert, S. and Henderson-Sellers, A. (1997). A statistical model to downscale local daily temperature extremes from synoptic-scale atmospheric circulation patterns in the Australian region. *Climate Dynamics*, 13(3):223–234.

Schumacher, R. S. and Johnson, R. H. (2006). Characteristics of US extreme rain events during 1999-2003. *Weather and Forecasting*, 21(1):69–85.

Sibuya, M. (1960). Bivariate extremal distribution. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210.

Sillman, J., Croci-Maspoli, M., Kallache, M., and Katz, R. (2011). Extreme cold winter temperatures in Europe under the influence of north Atlantic atmospheric blocking. *Journal of Climate*, 24(22):5899–5913.

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Wang, W., and Powers, J. (2005). A description of the advanced research WRF version 2. Technical report, DTIC Document.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1990). Max-stable processes and spatial extremes. *Unpublished manuscript*.

Smith, R. L., Tawn, J. A., and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Stahl, K., Moore, R., and Mckendry, I. (2006). The role of synoptic-scale circulation in the linkage between large-scale ocean–atmosphere indices and winter surface climate in British Columbia, Canada. *International Journal of Climatology*, 26(4):541–560.

Trenberth, K., Dai, A., Rasmussen, R., and Parsons, D. (2003). The changing character of precipitation. *Bulletin of the American Meteorological Society*, 84(9):1205–1218.

Von Storch, H. and Zwiers, F. (2002). *Statistical Analysis in Climate Research*. Cambridge University Press.

Wehner, M. (2013). Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. *Climate Dynamics*, 40(1–2):59–80.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Wellburn, A., Higginson, C., Robinson, D., and Walmsley, C. (2006). Biochemical explanations of more than additive inhibitory effects of low atmospheric levels of sulphur dioxide plus nitrogen dioxide upon plants. *New phytologist*, 88(2):223–237.

Weller, G. and Cooley, D. (2012). An alternative characterization of hidden regular variation in joint tail modeling. Technical report, Colorado State University Department of Statistics.

Weller, G. and Cooley, D. (2013). A sum characterization of hidden regular variation in joint tail modeling with likelihood inference via the Monte Carlo expectation–maximization algorithm. *submitted*.

Weller, G., Cooley, D., and Sain, S. (2012). An investigation of the pineapple express phenomenon via bivariate extreme value theory. *Environmetrics*, 23(5):420–439.

Weller, G., Cooley, D., Sain, S., Bukovsky, M., and Mearns, L. (2013). Two case studies on NARCCAP precipitation extremes. *submitted*.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.

Zhang, X., Wang, J., Zwiers, F., and Groisman, P. (2010). The influence of large-scale climate variability on winter maximum daily precipitation over North America. *Journal of Climate*, 23(11):2902–2915.

Zhang, Z. (2008). Quotient correlation: A sample based alternative to Pearsons correlation. *The Annals of Statistics*, 36(2):1007–1030.

Zhu, Y. and Newell, R. (1994). Atmospheric rivers and bombs. *Geophysical Research Letters*, 21(18):1999–2002.