EXTREME VALUE STATISTICS WITH APPLICATIONS IN HYDROLOGY AND FINANCIAL ENGINEERING

Julia E. Morrison

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

OPERATIONS RESEARCH AND FINANCIAL ENGINEERING

November 2001

© Copyright by Julia Egorova Morrison, 2001. All rights reserved.

Abstract

This work concentrates on features of Extreme Value statistics and bivariate distribution modeling using copulas that are widely applicable to important problems in hydrology and financial engineering. The three-parameter Generalized Extreme Value (GEV) distribution universally applied for modeling extreme processes is employed throughout for data analyses. New MIXed methods for GEV parameter estimation based on constraining the maximum likelihood estimators with different statistics are introduced. It is shown that these methods have good asymptotic behavior and produce better parameter estimates for small samples than do existing maximum likelihood and L-moments methods. The benefits of incorporating additional information into the maximum likelihood parameter estimation methods, such as the value of the second largest flood peak in a given year, are also developed. Annual flood peak data for a sample of 104 drainage basins in the central Appalachians region is analyzed, and the dependence of the estimated GEV distribution parameters on basin morphological properties is addressed.

Based on recent advances in the modeling of bivariate distributions with copulas, a library EVANESCE (Extreme Value ANalysis Employing Statistical Copula Estimation) is developed for the statistical software S-Plus, and the implemented methods are demonstrated on examples in financial engineering. The sensitivity of copula parameter estimates to precision in the description of marginal distributions is studied,

and it is concluded that the copula parameter estimates are robust. The possibility of modeling the joint distribution of flood peaks and flood volumes using copulas is explored in both Annual Maximum Series and Partial Duration Series (PDS) frameworks. A particularly useful copula family for modeling this distribution in the PDS framework is identified, and the use of the new approach is illustrated with examples.

Finally, the scaling behavior of flood peak distributions is examined using a statistical model of the spatio-temporal distribution of rainfall coupled to a hydrological model describing the transformation of rainfall to discharge within a drainage network (the Network Model). It is shown that the scaling behavior of the first two moments and coefficients of variations of the distribution of annual flood peaks can be reproduced using such simulations.

To the memory of my grandparents Boris and Tamara Kucher...

Acknowledgements

I deeply thank my advisors, Professors Jim Smith and René Carmona, for their constant support and advice. Jim Smith encouraged me to discover the aspects of research that were most interesting to me, counseled me the many times I became frustrated, and greatly helped me at all stages of the Princeton graduate school experience. René Carmona inspired me to learn about the theory of copulas and suggested the idea of creating the EVANESCE package. I also thank Professors Jim Smith and René Carmona for reading this dissertation, and providing valuable comments and suggestions for its improvement.

I would like to thank the members of my committee – Professors Çınlar and Rodriguez-Iturbe, for their guidance, and for offering inspiring, challenging courses.

I do not know how to thank Kyle, my husband, for his love, support, understanding, and editing skills. Without him, this thesis would contain far fewer articles.

I owe thanks to Mary Lynn Baeck, who kept rainfall modeling interesting, and bothered herself excessively with the well-being of our computer systems on the students' behalf. I also thank Noelina Hall for her help with all of the paperwork regarding this thesis.

I would like to thank Hande Benson for making the office a fun place to be, putting up with my messy desk, and motivating me to finish up my thesis. Special thanks go to Pavel Diko for listening to several of my interesting and not-so-interesting

research ideas, and to Paula Sturdevant-Rees (now an Assistant Professor at the University of Massachusetts) for helping me during my first years here. I thank all of the first users of the Network Model (Dan Turner-Gillespie, Francesca Giannoni, Naomi Hicks) for discovering bugs, making me feel useful, and ceding the limited computer resources to me on demand. I would also like to thank Aurore Lecanon, who was the first user of EVANESCE, and gave me helpful suggestions for improvements of the package. Thanks to everyone who made Princeton and E-Quad a fun place, especially Radhika Weitijunge, Eduardo Duenez, Teodora Touzharova, Kris Wammer, and Coral Fernandez-Illescas.

Lastly, but not the least, I must thank my parents, whose love and support can be felt from overseas, for encouraging me to attend Princeton in the first place, and my late grandparents for enabling me to love working with mathematics and physics. I owe them a great deal of my success.

Contents

Abstract **iii**

Acknowledgements **vi**

List of Tables **xii**

List of Figures **xv**

1 Introduction **1**

2 Generalized Extreme Value Distribution **6**

2.1 Parameter estimation methods for the GEV distribution 9

2.1.1 Method of Maximum Likelihood 10

2.1.2 Method of L-Moments 12

2.2 Mixed L-Moments – Maximum Likelihood Methods 15

2.2.1 Combination of the MLE and LMOM Methods 15

2.2.2 Comparison of Performance of Different Methods 17

2.3 Extension to PDS Methods 26

2.3.1 GEV Parameter Estimation - PDS Approach 26

2.3.2 The MLE2 Method 28

2.4 Analysis of Flood Peak Data from the Central Appalachian Region . 38

2.5 Asymptotic Properties of MIXed Estimators 55

2.6 Summary of Chapter 2 70

3 EVANESCE **72**

3.1 A Brief Introduction to Copulas 73

3.2 Parametric Copula Classes and Families 77

3.3 Dependence Summaries 87

3.4 Simulation of Random Pairs 94

3.5 Estimation of Copulas 99

3.5.1 Empirical Copulas 99

3.5.2 A Non-Parametric Estimate for Archimedean Copulas 100

3.5.3 A Non-Parametric Estimate for EV Copulas 104

3.5.4 Parametric Copula Estimation 105

3.6 Sensitivity of Parametric Copula Estimates to Errors in the Estimates of Marginal Distributions 107

3.6.1 Comparison of Parametric and Empirical Estimation of the Marginals 108

3.6.2 Comparison of the Separate Marginals Parameter Estimation Procedure and the Full Likelihood Method 114

3.7 Extension of Gumbel Copulas to Include a Third Dimension 117

3.7.1 Model Formulation and Properties 119

3.8 Summary of EVANESCE features 126

4 Financial Engineering Applications 129

4.1 Utility Stocks Indexes 130

4.2 Peaks-Over-Threshold Marginal Estimate 138

4.3 Copula Estimation 140

4.4 Computing Measures of Risk 149

4.5 Prices of Brazilian and Columbian Coffee 156

5 Applications in Hydrology 167

5.1 AMS and PDS approaches revisited 170

5.2 Parameter Estimation 175

5.3 Flood Peaks and Flood Volumes 179

5.4 Flood Peaks at Two Nested Basins 181

5.5 Summary of Chapter 5 188

6 Scaling Properties of Flood Peaks 189

6.1 Storm Model 195

6.2 The Simplified Network Model 206

6.3 Scaling Analyses 209

6.4 Summary of Chapter 6 221

7 Conclusions 224

A The Network Model 231

A.1 Introduction 231

A.1.1 History 231

A.1.2 Description 232

A.2 Network Model Principal Components 235

A.2.1 Green and Ampt Infiltration with Redistribution 236

A.2.2 Flow Directions 245

A.2.3 Water Surface Routing and Infiltration 248

A.2.4 Fast Routing Option 250

A.2.5 Transition to Streams 253

A.3 User Manual 254

A.3.1 Parameter Files 254

A.3.2 Preparing Input Files 256

A.3.3 Model Operation 270

A.3.4 Output 276

A.4 Conclusions 277

Bibliography **278**

List of Tables

2.1 RMSE of parameter estimates for MLE, LMOM, MIX1, and MIX2 methods for selected sample sizes and negative values of k. 20

2.2 Relative RMSE of quantile estimates for MLE, LMOM, MIX1, and MIX2 methods for selected sample sizes and values of k. 21

2.3 Comparison of performance of LMOM and MIX2 methods in parameter and quantile spaces (see text for more details). The true values of the parameters are $(1, 01, -0.3)^T$ and the true value for $Q(.99)$ is 9.9169 23

2.4 RMSE of the estimates of k from LMOM, MLE2, and MLE2-MIX1 methods for selected sample sizes and negative values of k. 36

2.5 RMSE of quantile estimates for LMOM, MLE2, and MLE2-MIX1 methods for selected sample sizes and values of k. 37

2.6 R^2 of linear regression between basin descriptors and MLE2-MIX1 estimates of k 44

2.7 Comparison of the quantiles of the two stochastic models. The Gumbel mixture model is described in the text. 51

3.1 Several parametric families of Extreme Value copulas implemented in the EVANESCE library. 80

3.2 Several parametric families of Archimedean Copulas implemented in the EVANESCE package. Here, $\tilde{u} = -\log(u)$ and $\tilde{v} = -\log(v)$. . . . 81

3.3 Formulas for upper and lower tail indexes for several parametric copula families. 95

3.4 Comparison of the performance of the full likelihood estimation method (Full) to the separate estimation of marginals and copula parameters (Separate). There is little difference between the two methods' parameter estimates. 116

4.1 Comparison of $F(x, y)$ estimates from the data with estimates obtained using different methods. 136

4.2 Parametric copulas that provide a good fit to the utility index data. The Kendall's tau and Spearman's rho estimates for the data are $\tau_{\text{data}} = 0.380$, and $\rho_{\text{data}} = 0.537$. $L_2(\Delta C)$ is the L_2 norm of the difference between the empirical copula and the fitted parametric copula. . 146

5.1 Expected number of flood peaks exceeding certain thresholds in both peak and volume. 183

6.1 Estimated values for the Storm Model parameters. 204

A.1 Typical values of GAR parameter for selected soil textures (from Ogden and Saghafian [1997]). 243

A.2 Parameters for the `RewriteDem` class. 258

A.3 Parameters for the `ComputeFlowDirection` class. 260

A.4 Parameters for the `RewriteStreamData` class. 261

A.5 Parameters for `CreateRectRain` class. 267

A.6 Parameters for the `RewriteSoils` class. 269

A.7 Parameters for the `RewriteImperviousAreas` class. 271

A.8 Parameters for the `NetworkModel.RunModel` class. 273

List of Figures

2.1 Bias **(a)** and RMSE **(b)** of the estimator $\hat{k}$ for a sample size $n = 35$ and four different estimation methods: MLE, LMOM, MIX1, and MIX2. 19

2.2 Bias **(a)** and RMSE **(b)** of the estimator $\hat{k}$ for record length 35 years and four different estimation methods: MLE2, LMOM, MIX1, and MLE2-MIX1. 35

2.3 A map of the central Appalachians region showing the locations of stream gauging stations. Solid circles denote urban basins and empty circles denote rural basins. 39

2.4 A normal QQ-plot of at-site estimates of k for the central Appalachians basins. The lines on the plot correspond to Gaussian approximations of the distributions of LMOM and MIX1 estimates of k. 42

2.5 Dependence of MLE2-MIX1 estimate of k on basin's drainage area. . 43

2.6 Dependence of MLE2-MIX1 estimate of a **(a)** and b **(b)** on basin drainage area. Solid circles denote urban basins, empty circles denote rural basins. 45

2.7 Dependence of MLE2-MIX1 estimate of a **(a)** and b **(b)** on effective area A_{eff}. Solid circles denote urban basins, empty circles denote rural basins. 46

2.8 Envelope curves of flood peaks for the central and southern Appalachians regions. 48

2.9 Distribution of the estimates of k for GEV distribution and the Gumbel mixture model with the parameters given in Table 2.7 and sample size $n = 50$. 52

2.10 Probability density function of the GEV distribution and the Gumbel mixture model with the parameters given in Table 2.7. 53

2.11 A typical quantile-quantile plot for samples generated from the GEV distribution (solid squares) and Gumbel mixture (empty circles) versus quantiles of the GEV distribution with shape parameter $k = -0.42$. . 54

3.1 Contours of density for different copula families with standard Gaussian margins. a) Gaussian, $\delta = 0.707$, $\rho_P = 0.71$ (Pearson correlation coefficient); b) Frank, $\delta = 5.74$, $\rho_P = 0.66$; c) Kimeldorf and Sampson, $\delta = 2$, $\rho_P = 0.69$; d) Gumbel, $\delta = 2$, $\rho_P = 0.70$. 84

3.2 Contours of density for different copula families with standard Gaussian margins. a) BB1, $\theta = 0.42$, $\delta = 1.65$, $\rho_P = 0.71$ (correlation coefficient); b) BB2, $\theta = 0.5$, $\delta = 2.05$, $\rho_P = 0.70$; c) BB3, $\theta = 1.4$, $\delta = 0.85$, $\rho_P = 0.72$; d) BB4, $\theta = 0.7$, $\delta = 0.75$, $\rho_P = 0.72$. 85

3.3 Contours for one-parameter copula families with Gumbel margins and Kendall's tau = 0.5: dashed lines – Gumbel, $\delta = 2$; solid lines – Galambos, $\delta = 1.28$; dotted lines – Hüsler and Reiss, $\delta = 1.8$. 86

3.4 Contours for two-parameter copula families BB1 and BB7 with standard Gaussian margins and Kendall's tau = 0.5: a) BB1, $\theta = 0.8$, $\delta = 1.43$; b) BB1, $\theta = 0.1$, $\delta = 1.90$; c) BB7, $\theta = 1.9$, $\delta = 1.11$; d) BB7, $\theta = 1.43$, $\delta = 1.64$. The upper tail index for the copulas is 0.38 in figures a) and c) and 0.56 in figures b) and d). 88

3.5 Contours of a Gumbel copula (dashed lines) with parameter $\delta = 2.0$ and of an empirical copula (solid lines) from a random sample of size 100 generated from this Gumbel copula. 101

3.6 The estimate of a function $\lambda(s)$ for a sample of size 100 from a BB1 copula (solid curve), and the true function $\lambda(s)$ for this copula (dashed curve). 103

3.7 RMSE in parameter estimates (top) and Kendall's tau estimates (bottom) for the BB1 copula using different estimates of the GP marginals. 112

3.8 RMSE of the Gumbel copula δ parameter estimates using GEV marginals for different sample sizes ($n = 50$ – dotted lines, $n = 100$ – solid lines, and $n = 250$ – dashed lines), and marginal estimation techniques. There is no difference in RMSEs for empirically and parametrically estimated marginals. 113

3.9 Dependence functions of copulas $C^{(\mathrm{E2})}_{3.34,1.5}$ (curve) and $C^{(\mathrm{G2})}_{2.0}$ (circles) . 121

3.10 Contour plots of copulas $C^{(\mathrm{E2})}_{3.34,1.5}$ (dotted lines) and $C^{(\mathrm{G2})}_{2.0}$ (dashed lines). The difference between the two functions is very small. 122

3.11 1000 point simulated from the $C^{(\mathrm{E2})}$ copula with $\delta_1 = 8$ and $\delta_2 = 1.1$ ($\delta_1 \gg \delta_2$). 125

4.1 Time series of indexes reflecting the stock prices of electric (top) and gas (bottom) companies. 131

4.2 Time series of the daily log-returns of two indexes UTI.ELEC (top) and UTI.GAS (bottom). 133

4.3 Dependence between the log-returns of the electric and gas company indexes. 134

4.4 Gaussian quantile-quantile plots for the log-returns. 137

4.5 Plot of estimates of the shape parameter of the GPD distribution as a function of threshold. a) UTI.ELEC, upper tail; b) UTI.ELEC, lower tail; c) UTI.GAS, upper tail; d) UTI.GAS, lower tail. 141

4.6 Fit of the estimated marginal distributions into upper (top) and lower (bottom) data tails. The empirical cdf of the `UTI.ELEC` log-returns is shown with crosses, and that for `UTI.ELEC` is shown with triangles. . 142

4.7 Non-parametric estimate of the $\lambda(t) = \phi(t)/\phi'(t)$ function for Archimedean copulas (solid line), and $\lambda(t)$ for the BB1 (dashed line) and BB7 (dotted line) copulas with parameters given in Table 4.2. 147

4.8 Contour plot showing the fit of the BB1 copula (dashed lines) to the empirical copula of our data (solid lines). 148

4.9 Observed data points (bottom), estimate of the joint density function (middle), and 10,000 simulated points from that density function (top). 150

4.10 Estimates of VaR_q for the utility indexes, obtained using simulations (symbols) and direct integration (lines) for different values of q as functions of the samples size used in the simulations. Bold symbols correspond to the data estimates. 154

4.11 Estimates of the expected shortfall for the utility indexes, obtained using simulations for different values of q as functions of the samples size used in the simulations. For $q = 10^{-5}$, good convergence is not seen even for 1,000,000 simulation points. 155

4.12 Time series of Brazilian and Columbian Coffee prices. 157

4.13 Daily log-returns of Brazilian (x axis) and Columbian (y axis) coffee prices. 158

4.14 Comparison of the dependence of data (a), a sample from a bivariate Gaussian distribution with the same a coefficient of correlation (0.68) and variance as the data. (b), a sample from a bivariate Gaussian distribution with a coefficient of correlation 0.85 and the same marginal variances as the data. (c), a sample from a Gaussian copula with the parameter $\delta = 0.85$ and marginal distributions estimated from data (d). 160

4.15 Coffee data transformed to uniform $(0, 1)$ marginals. Though most of the points suggest a strong dependence structure within the data, there are a number of two dimensional "outliers". 162

4.16 Illustration of how the Tawn copula produces the two dimensional "outliers": Five hundred points $(\tilde{u}, \tilde{v})$ generated from a Tawn copula, as paired maxima of 23 pairs generated from the Gumbel copula and 2 pairs – from the independence copula. If both $\tilde{u}$ and $\tilde{v}$ came from the observation from a Gumbel copula, this observation is represented by an empty square. 165

4.17 Observed data points of coffee price daily log-returns (bottom), estimate of their joint density function (middle), and 10,000 simulated points from that density function (top). 166

5.1 1000 observations generated from a $C^{(\mathrm{PDS})}$, with Galambos dependence function, and parameter $\delta = 2.0$. The copula's singular component is clearly defined. 174

5.2 Simulated 40 years of data for two bivariate PDS models with the same AMS framework parameters and $\lambda = 2.5$ (top) and $\lambda = 2.5$ (bottom) 177

5.3 Maximum flood peaks and flood volumes for Potomac at Point of Rocks plotted against one another. 180

5.4 The non-parametric estimate of the dependence function A (points) and the parametrically estimated dependence function (curve) of $C^{(\mathrm{AMS})}$ for a distribution of flood peaks and volumes for Potomac at Point of Rocks. 182

5.5 Simulated values for a period of 500 years (top), estimate of the joint probability density function (middle), and observed values for 59 years (bottom) of flood peaks and volumes for Potomac at Point of Rocks in the PDS framework. 184

5.6 Values of the observed annual flood peaks (cms) for Potomac at Point of Rocks and South Fork Shenandoah at Front Royal. 186

5.7 Simulated values for a period of 500 years (top), estimate of the joint probability density function (middle), and observed values for 66 years (bottom) of flood peaks at Potomac at Point of Rocks and South Fork Shenandoah at Front Royal in the PDS framework. 187

6.1 CV – Drainage Area relationship for central Appalachian floods. From Smith [1992]. 191

6.2 Partial map of Texas, showing the regions mentioned in the text. The inset corresponds to the area of interest. Triangles represent rain gauge locations. 193

6.3 Measured annual flood peaks for the stream gauging station at Guadalupe River at Comfort. The line represents the median value. 199

6.4 Estimate of the seasonally-varying rate of storm occurrence. Time index 0 corresponds to January 1, and time index 1 to December 31. . 201

6.5 A schematic of the cascade model construction. 204

6.6 The network structure of the Guadalupe river. The bold line shows the path of the water from the marked point in the basin to the outlet. 207

6.7 Log-log plot of first (solid circles) and second (crosses) sample moments of the simulated flood peaks versus drainage area. Empty squares represent estimated values of mean annual flood peak for actual basins in the area of interest. The units of the y axes are $\mathrm{m^3s^{-1}}$ for the first moments and $\mathrm{m^6s^{-2}}$ for the second. The parameters used to obtain this plot were: $v = 1$ m s^{-1}, $K_s = 0.5$ mm h^{-1}, $\varsigma = 0.6$, and $k = 8$. . 212

6.8 CV – Drainage area relationship for the simulated flood peaks. Triangles indicate values obtained with the rainfall simulated with the Storm Model, empty circles indicate values where the rainfall was simulated without the time cascade structure. Squares represent estimated values for CV for actual basins in the area of interest. The parameters of the models are the same as in Figure 6.7. 213

6.9 $\max(f_e(x)/v)$ as a function of drainage area 215

6.10 The envelope curves for the CV-area relationship resulting from simulated rainfall with different spatial scales. (1) $a^{-1/2} = 60$, $c^{-1} = 60$; (2) $a^{-1/2} = 30$, $c^{-1} = 15$; (3) $a^{-1/2} = 15$, $c^{-1} = 15$; (4) $a^{-1/2} = 15$, $c^{-1} = 30$ (km). 218

6.11 The Peano network of order 5. 219

6.12 CVs of annual flood peak simulated in the sub-basins of Peano basin. Parameters of the models are as in Figure 6.7. 222

A.1 The schematic of the Network Model. 233

A.2 Procedures performed at each time step of the Network Model. 237

A.3 Schematic of the GAR algorithm used in the Network Model. For an individual cell, a transition between the states of the code along an arrow occurs when the condition on the arrow is satisfied. 242

A.4 Illustration for the Tarboton method for specification of flow direction. See text for details. 246

A.5 Possible flow paths from cell A. The points that water reaches during one time step are denoted by circles. 249

A.6 Difference between fast and slow routing options. The two methods perform the same for large overland flow velocities. Slow routing procedure results is significant attenuation of the flood peak for small overland flow velocities. 251

A.7 Potential differences between ARC-INFO vector coverage data (left) and the Network Model representation (right) of the same network. The circles denote nodes of the network. In the Network Model streams are numbered from the outlet, and a typical numbering is shown here. 263

A.8 Example of the ARC-generate and attribute files for a river network. 264

Chapter 1

Introduction

Advanced statistical methods can be employed in fields as disparate as Cell Biology and Electrical Engineering to enhance researchers' ability to understand the behavior of the processes they study. In disciplines where large volumes of data are available for analysis, statistical approaches are a necessity, but even in situations where the events of interest are scarce, existing statistical methods can provide a means of extracting useful information from a minimum of data. Extreme value statistics is one of the realms of study that addresses this extrapolation of analytical and stochastic models to ranges beyond those of the finite data samples at hand. Extreme value statistics concerns the patterns and the magnitudes of events that occur with a very low probability, and more importantly, the probability with which those rare events will occur.

In hydrology, the development of stochastic methods for the characterization of flood peaks in drainage basins has both motivated and benefited from the treatment of

classical problems in extreme value theory. Much effort has been expended to record flood peaks and flood volumes for many watersheds and apply statistical methods to these data to predict the probability that floods greater than any given size will occur. In the process of understanding the detailed behavior of already well-diagnosed basins, researchers have come to grips with the difficulty of fitting extreme value distributions to the data, and obtaining useful parameter estimates.

The Generalized Extreme Value (GEV) distribution has been widely used for modeling the distribution of flood peaks in at-site and regional settings [Hosking *et al.*, 1985a; Smith, 1987; Stedinger and Lu, 1995; Rosbjerg and Madsen, 1995; Hosking and Wallis, 1997]. In addition to flood modeling, the GEV distribution is commonly used to model many other natural extreme events [Smith, 1986; Bauer, 1996; Kuchenhoff and Thamerus, 1996; Bruun and Tawn, 1998; Parrett, 1998]. The "extreme events" are often defined to be maximum value of a quantity over a given period of time, such as the maximum annual discharge in a stream, or the maximum volume of a flood. Extreme value theory, in particular the extremal types theorem [Leadbetter *et al.*, 1983], suggests that the distribution of these maxima should be close to one of the extreme value types. The GEV distribution, introduced by Jenkinson [1955], is a three-parameter family that combines all three extreme value types into a single form. Several techniques for GEV parameter estimation, such as the method of L-moments, the method of Maximum Likelihood, and the Method of Moments have been studied. The methods are often pushed past their applicable limit by researchers in need of

more powerful techniques to describe the complicated interdependencies in their data. In Chapter 2 of this dissertation, new methods for fitting the GEV distribution to flood peak data, including the incorporation of more of the available information into the estimation problem, are explored. The precision with which the critical shape parameter of the GEV distribution can be estimated is improved.

One interesting avenue of research in hydrology is how the estimated parameters of the annual flood peak distributions depend on a basin's morphological properties. In Chapter 2, section 2.4 we study the variability of the GEV distribution's estimate for a sample of basins in the central Appalachian region.

A related research direction has focused on the estimation of flood peak distributions for ungauged drainage basins (see Gupta and Waymire [1998]). A point of departure for many studies concerns the role of scale, which for a drainage basin is most often characterized by the drainage area. In hydrology, the scaling theories are two-fold: not only can they be used to infer the distributions of flood peaks for ungauged basins, but also they provide a way to consolidate the observations from different gauged basins in order to improve parameter estimates. Regional analyses of flood peak data, however, suggest that neither of the two available scaling theories (simple-scaling or multi-scaling) should be blindly applied to data, until certain issues involving peculiar behavior of flood peaks are resolved. In Chapter 6, we examine the scaling properties of flood peaks using a coupled stochastic rainfall-runoff model to observe the interactions between the temporal and spatial variations in the rainfall,

and the network structure of the basin. Our main goal is to identify the key factors that can explain the observed scaling behavior of flood peaks.

Financial engineers, in contrast, are interested in describing more precisely the fluctuating behavior of the myriad of financial instruments that mediate business in the modern world. Observation of non-Gaussian character in the tails of daily log-return data distributions has heralded the need for extreme value statistical methods to better approximate the infrequent extreme changes in the data. The parameter estimation techniques developed for the fitting of flood peak distributions may be applied to this problem as well, but the problems that financial engineers face often involve even more complex random variables. Moreover, Monte Carlo simulations appear to be most reliable methods to quantify and control the risks associated with the extreme events.

Recent theoretical advances in the area of multivariate modeling provided an opportunity for researchers to improve their arsenal of tools for modeling dependencies in the data. The modeling of joint distribution functions for different random variables has been enhanced by the development of theories involving copulas, joint distribution functions with uniform $(0, 1)$ marginals. The word *copula*, meaning "a bond" in Latin, was first used by Sklar [1959], when he argued that univariate distributions can be "joined together" by special functions (copulas) to create joint multivariate distribution functions. Many different families and classes of copulas have been introduced. Together they can describe a vast array of data dependencies. The idea behind

copula-based approaches, is that marginals of multivariate distributions can be estimated separately from the "copula part", that determines the dependence structure of the data.

In Chapter 3, several of the properties of copulas are described, and techniques for estimating their parameters and simulating random variables are discussed in detail. Chapter 4 presents several financial engineering applications of techniques involving copulas, such as the analysis of utility stocks indexes and the computation of risk measures, using the EVANESCE S-Plus library, written for this work to simplify the implementation of copulas. Chapter 5 surveys applications of copula theory to hydrological questions, such as the dependence between flood peaks in two different basins or the relationship between the flood peaks and flood volumes for a given basin.

Chapter 7 restates several of the more important results of this work, and notes the many related avenues of further research that could be pursued.

Chapter 2

Stochastic Modeling of Flood Peaks Using The Generalized Extreme Value Distribution

The Generalized Extreme Value (GEV) distribution is widely used for modeling the distribution of flood peaks in at-site and regional analyses [Hosking *et al.*, 1985a; Smith, 1987; Stedinger and Lu, 1995; Rosbjerg and Madsen, 1995; Hosking and Wallis, 1996]. Several parameter estimation procedures for the three-parameter GEV distribution have been developed and extensively studied [Jenkinson, 1969; Prescott and Walden, 1980; Hosking *et al.*, 1985b]. The most commonly used methods are Maximum Likelihood Estimation (MLE) and the method of L-moments (LMOM) [Smith, 1990]. In section 2.1, we discuss the advantages and disadvantages of these existing methods, and in section 2.2, we show how the MLE and LMOM methods can be combined to produce improved GEV parameter estimators. The resulting "mixed" method estimators of the shape parameter have reduced variance compared to the

MLE estimator and reduced bias compared to the LMOM estimator. The behavior of the new estimators is studied for small samples using Monte Carlo simulations, and for large samples using asymptotic techniques. In the applications to flood peaks, the data used for the GEV parameter estimation comprises records of the maximum peaks for each year (this approach is usually referred to as the Annual Maximum Series (AMS) approach).

Partial Duration Series (PDS) [Shane and Lynn, 1964; Todorovich and Zelenhasic, 1970; Karr, 1976] models assume that the arrival times of peaks above a specified threshold form a Poisson process in time, and that the distribution of the peak magnitudes has a particular form. The attraction of this procedure is that additional information to AMS data can be exploited to improve the parameter estimation. There are, for example, many flood records in which the second largest flood peak during a year is larger than the majority of the other annual flood peaks in the record. These observations are relevant to the estimation of extreme flood quantiles. The Generalized Pareto distribution is a common choice for the peak magnitude distribution because it corresponds to a limiting distribution for excesses over a threshold as that threshold is increased, and also produces a GEV distribution of annual flood peaks [Leadbetter *et al.*, 1983; Smith, 1984; Madsen and Rosbjerg, 1997]. Madsen *et al.* [1997a] showed that errors in parameter estimation under the GP/PDS approach can be smaller than those of the GEV/AMS approach, under certain conditions.

In section 2.3, we introduce a Maximum Likelihood Estimation method that incorporates the values of the two largest observations for a given year (MLE2). This method includes additional information from the PDS record with the entire record of the AMS data. This approach is useful because the PDS record is shorter than the AMS record for a majority of stream gauging stations, the threshold value may change over time and is generally not known for many stream gauging records, and the second largest flood peak during a given year should provide most of the additional information contained in the full PDS record. The procedure is developed and tested with Monte Carlo simulations in a PDS framework. The results confirm that additional data can indeed improve parameter estimates.

The GEV distribution has also played an important role in regional flood frequency analyses [Hosking *et al.*, 1985a; Lettenmaier *et al.*, 1987; Chowdhury *et al.*, 1991; Stedinger and Lu, 1995; Hosking and Wallis, 1996]. Simple scaling (or index-flood) theory assumes that appropriately scaled annual flood peaks have the same distribution in a homogeneous geographical region [Gupta and Waymire, 1998], which, for the GEV distribution, means that the shape parameter k and the ratio of scale to location parameters are constant across all basins in the region. In this chapter, we discuss the scaling properties of the parameter estimates for the GEV distribution for a sample of stations in the central Appalachians regions. Further discussion of the different scaling theories of flood peaks can be found in Chapter 6.

2.1 Parameter estimation methods for the GEV distribution

The Generalized Extreme Value (GEV) distribution combines into a single form all three Extreme Value (EV) distributions: Gumbel (EVI, $k = 0$), Frechet (EVII, $k < 0$), and Weibull (EVIII, $k > 0$). The GEV distribution has the following cumulative distribution function (cdf):

$$G(x) = \begin{cases} \exp\left\{-\left(1 - \frac{k(x-b)}{a}\right)^{1/k}\right\} & \text{for } k \neq 0, \\ \exp\left\{-\left(e^{-\frac{(x-b)}{a}}\right)\right\} & \text{for } k = 0. \end{cases} \tag{2.1}$$

It has three parameters: scale $a > 0$, location b, and shape k. Here, $-\infty < x \leq b + a/k$ for $k > 0$, $-\infty < x < \infty$ for $k = 0$, and $b + a/k \leq x < \infty$ for $k < 0$. To simplify our notation, we will write θ for the vector $(a,\ b,\ k)^T$, and $G_\theta(x)$ for the GEV distribution with parameter θ. The corresponding probability density function will be denoted as $g_\theta(x)$.

It has been noted that estimates of the shape parameter k of the GEV distribution for flood peak data are usually negative [Smith, 1987; Madsen *et al.*, 1997a; Martins and Stedinger, 2000], implying that the distribution has heavy tails. In this dissertation, therefore, we will focus on parameter estimation procedures for negative values of k. Because the GEV distribution does not have a third moment when $k < -1/3$, estimators based on the method of moments will not be considered here.

The quantile function of the GEV distribution is given by:

$$Q(p) = \begin{cases} b + \frac{a}{k}\left(1 - (-\log p)^k\right) & k \neq 0, \\ b - a\log(-\log p) & k = 0. \end{cases} \tag{2.2}$$

For a given value of p, quantile estimates are obtained by substituting estimated values of the parameters into the formula above. We are interested in large quantile values: for example, the 100-year-return flood magnitude $Q(.99)$. In subsequent sections, we examine the properties of several quantile estimators $\hat{Q}(p)$, in addition to properties of different parameter estimators $\hat{k}$, $\hat{a}$, and $\hat{b}$.

2.1.1 Method of Maximum Likelihood

The log-likelihood function of a random sample $\{x_1, x_2, \ldots, x_n\}$ of size n is

$$\begin{aligned} \log L(\theta|x) &= -n\log a - \sum_{i=1}^{n}\left\{1 - \frac{k(x_i - b)}{a}\right\}^{1/k} \\ &+ \left(\frac{1}{k} - 1\right)\sum_{i=1}^{n}\log\left\{1 - \frac{k(x_i - b)}{a}\right\}, \end{aligned} \tag{2.3}$$

and the corresponding MLE estimator $\hat{\theta} = (\hat{a},\ \hat{b},\ \hat{k})^T$ is the point at which $\log L(\theta|x)$ attains its maximum. This can also be expressed as the solution to the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \log L(\theta|x) \\ \text{subject to} \quad & k(x_i - b) \leq a \qquad i = 1, \ldots, n \\ & a > 0 \end{aligned} \tag{2.4}$$

The constraints in this problem guarantee that the probability density function of the GEV distribution has to be positive at $\{x_1, x_2, \ldots, x_n\}$.

Jenkinson [1969] first noted the complexity of the MLE method for the GEV distribution. Traditionally, problem (2.4) is solved by setting the partial derivatives of the log-likelihood function (2.3) to zero, and using Newton-Raphson iterations to solve for the parameters. Historically, the lack of computational power necessitated a replacement of the Hessian matrix of (2.3) during the solution of (2.4) by its expected value, either through tables, or explicitly through gamma functions [Jenkinson, 1969; Prescott and Walden, 1980]. Hosking [1985] and MacLeod [1989] implemented the latter approximation in Fortran algorithms. The use of an expected Hessian significantly reduces the computational time required to find a solution to the optimization problem, but also frequently results in non-convergence of the methods when $k < 0$ [Hosking *et al.*, 1985b]. The current availability of optimization solvers (such as LOQO, Vanderbei [1999]) reduces the need for statisticians to worry about optimization techniques and the computational complexities of Maximum Likelihood methods. In the case of the GEV distribution, MLE reduces to a three-dimensional, smooth optimization problem with two inequality constraints, which can be solved in milliseconds – the complexity of the MLE method should no longer be considered a disadvantage. In our study, we used a solver available within the statistical software S-Plus, that employs a quasi-Newton method, to find the MLE and other estimates based on the method of maximum likelihood [Gay, 1983; Bazaraa *et al.*, 1993].

2.1.2 Method of L-Moments

The term L-moment, introduced by Hosking [1986], refers to a linear function of the expected order statistic, implying that L-moments are special cases of L-statistics. The theory behind L-moment estimators has been developed for some time, but was scattered in the statistical literature until the 1980s, when techniques based on L-moments were studied both theoretically and with simulations [Hosking *et al.*, 1985b; Hosking, 1986; Hosking and Wallis, 1987; Hosking, 1990].

For a real-valued random variable X with cumulative distribution function $H(x)$, corresponding quantile function $H^{-1}(p)$, and order statistics $X_{(1:n)} \leq X_{(2:n)} \leq \ldots \leq X_{(n:n)}$ for a random sample of size n, the L-moments of X, λ_r, are defined as

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathbb{E}X_{((r-k):n)} \qquad r = 1, 2, \ldots$$

Using formulas for the expected value of order statistics (see David [1981, p. 33]), one can express λ_r with integrals involving the quantile function H^{-1}. The expressions for the first three L-moments are (see Hosking [1990]):

$$\begin{aligned}
\lambda_1 &= \mathbb{E}X = \int_0^1 H^{-1}(p)dp \\
\lambda_2 &= \frac{1}{2}\mathbb{E}\left[X_{(2:2)} - X_{(1:2)}\right] = \int_0^1 H^{-1}(p)(2p-1)dp \\
\lambda_3 &= \frac{1}{3}\mathbb{E}\left[X_{(3:3)} - 2X_{(2:3)} + X_{(3:3)}\right] = \int_0^1 H^{-1}(p)(6p^2-6p+1)dp \qquad (2.5)
\end{aligned}$$

Hosking [1990] showed that all L-moments of X exist if and only if $\mathbb{E}X < \infty$, and that, if $\mathbb{E}X < \infty$, the distribution of X is fully characterized by its L-moments.

As in the case of conventional moments, parameters of the distribution H can be estimated by estimating L-moments from independent observations and then solving equations (2.5) for the parameters. Several different estimators for L-moments have been proposed (see Hosking *et al.* [1985b]; Hosking and Wallis [1995]; Goel and De [1993]), and the following unbiased estimators are the most frequently used:

$$\hat{\lambda}_r = (-1)^r \sum_{k=0}^{r-1} (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} \alpha_k, \qquad r = 0, 1, 2, \ldots$$

where

$$\alpha_k = n^{-1} \sum_{i=1}^{n} \frac{(n-i)(n-i-1)\ldots(n-i-k)}{(n-1)(n-2)\ldots(n-k)} x_{(i)},$$

and n is the sample size.

L-moments for the GEV distribution can be expressed in terms of the distribution parameters in the following way [Hosking, 1986]:

$$\begin{aligned}
\lambda_1 &= b + \frac{a}{k}\left[1 - \Gamma(1+k)\right] \\
\lambda_2 &= \frac{a}{k}(1 - 2^{-k})\Gamma(1+k) \\
\lambda_3 &= -\frac{a}{k}(1 - 3 \cdot 2^{-k} + 2 \cdot 3^{-k})\Gamma(1+k).
\end{aligned}$$

The LMOM estimator $\hat{k}$ for the shape parameter is then the solution of the equation:

$$\frac{1 - 3^{-\hat{k}}}{1 - 2^{-\hat{k}}} = \frac{\hat{\tau}_3 + 3}{2}, \tag{2.6}$$

and the corresponding LMOM estimators for a and b are

$$\hat{a} = \frac{\hat{\lambda}_2 \hat{k}}{(1 - 2^{-\hat{k}})\Gamma(1 + \hat{k})}, \tag{2.7}$$

$$\hat{b} = \hat{\lambda}_1 - \frac{\hat{a}}{\hat{k}}[1 - \Gamma(1 + \hat{k})], \tag{2.8}$$

where $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\tau}_3 = \hat{\lambda}_3/\hat{\lambda}_2$ are the estimators of the first two L-moments and the L-skewness obtained from the sample [Hosking, 1990]. Equation (2.6) is usually solved using Newton's method or the solution is simply approximated (see Hosking *et al.* [1985b]).

The bias of LMOM estimates of k increases with decreasing k, and is larger than 0.07 when $k = -0.4$ for $n = 30$. LMOM is also capable of producing infeasible estimates (for example, estimates that imply that the lower bound of the distribution is higher than the smallest observation in the sample), although it does this infrequently [Martins and Stedinger, 2000]. The MLE method, in comparison, produces almost unbiased estimates of k, but the variance of these estimates is larger than the variance of LMOM estimates. MLE also produces absurd estimates of k (such as $k < -1$), which leads to very large errors in quantile estimation [Martins and Stedinger, 2000]. This is why, despite the fact that it tends to produce biased estimates, the LMOM method is considered to be preferable to MLE [Hosking *et al.*, 1985b; Madsen *et al.*, 1997a]. MLE-based methods, however, can easily incorporate additional information such as censored data, or a known prior distribution for k [Prescott and Walden, 1983; Martins and Stedinger, 2000]. This property is a great advantage of the MLE methods [Smith, 1990].

In the next section, we present a combination of the LMOM and MLE methods, and show that it provides improved estimates of the shape parameter k for several cases.

2.2 Mixed L-Moments – Maximum Likelihood Methods

2.2.1 Combination of the MLE and LMOM Methods

One way to improve MLE estimates of k is to impose additional constraints on the optimization problem in (2.4). We would like these constraints to be based on the sample rather than on additional assumptions about the process that we have observed. One such constraint could be calibration to the empirical L-moments. For example, we can require the first L-moment of the estimated GEV distribution to be the same as that estimated from the sample. The addition of this constraint to the MLE problem (2.4) produces the first MIXed (MIX1) method. Heuristically, in the MIX1 method we maximize the likelihood function of the sample over a set of distribution parameters, such that the resulting expected value of the distribution is equal to the sample mean.

In the MIX1 method, we maximize the log-likelihood function $L(\theta|x)$ as a function of a and k after substituting b from the L-moment equation (2.8). The MIX1 estimator $\hat{\theta}$ of the parameters of the GEV distribution, then, is the solution to the following

optimization problem:

$$\begin{aligned} \text{maximize} \quad & \log L(\theta|x) \\ \text{subject to} \quad & b = \hat{\lambda}_1 - \frac{a}{k}[1 - \Gamma(1+k)] \\ & k(x_i - b) \leq a \qquad i = 1, \ldots, n \\ & a > 0 \end{aligned} \tag{2.9}$$

In the second MIXed method (MIX2), we maximize the likelihood function $L(\theta|x)$ as a function of k after substituting both b and a from the L-moment equations (2.7 – 2.8). The optimization problem for this method is

$$\begin{aligned} \text{maximize} \quad & \log L(\theta|x) \\ \text{subject to} \quad & b = \hat{\lambda}_1 - \frac{a}{k}[1 - \Gamma(1+k)] \\ & a = \frac{\hat{\lambda}_2 k}{(1 - 2^{-k})\Gamma(1+k)}, \\ & k(x_i - b) \leq a \qquad i = 1, \ldots, n \end{aligned} \tag{2.10}$$

We can also consider MIX2 to be an LMOM method where we maximize the likelihood function to obtain the estimate k, instead of using equation (2.6). From this perspective, we avoid using the estimator for τ_3, which has a large bias if the true value of k is less than -0.2 and the sample size is small (less than 50).

Both MIXed methods are based on the solution of a nonlinear optimization problem with nonlinear constraints involving one or two variables. As was mentioned previously, such problems can be easily solved using quasi-Newton methods. Because the optimization in these methods is performed over fewer variables than MLE,

MIXed methods are faster than MLE. The initial point for the optimization can be taken to be the LMOM estimate, and, if it is infeasible, the value of k can be adjusted to make it feasible.

At the end of this chapter (section 2.5), we show that MIXed methods have desirable asymptotic properties: they are strongly consistent and have an asymptotically jointly Gaussian distribution. The determinant of the asymptotic covariance matrix is larger than the respective determinant of the MLE estimator, but we are interested in studying the performance of the MIXed methods for small sample sizes anyways.

2.2.2 Comparison of Performance of Different Methods

In order to evaluate the performance of the MIXed methods, we conducted a series of simulation experiments. We simulated random samples of different sizes from GEV distributions with values of k ranging from -0.5 to 0.0. For each sample, we estimated the parameters of the GEV distribution using the MLE, LMOM, MIX1, and MIX2 methods. We present the results for the estimators of the shape, location and scale parameters, as well as flood quantiles, although we are mostly interested in the performance of the estimators of the shape parameter. Each simulation experiment was performed using 10,000 replicates of the samples.

The bias and root mean square error (RMSE, defined as $\left(\mathbb{E}\left[(\hat{k}-k)^2\right]\right)^{1/2}$) were computed for MLE, LMOM, MIX1 and MIX2 estimators of k for sample size $n = 35$, and plotted in Figure 2.1. For $k \leq -0.2$, the MIX1 and MIX2 estimators have smaller biases than the LMOM methods and smaller variances than MLE, resulting

in smaller RMSE in the estimation of k that for LMOM and MLE. Neither MIXed method produced any absurd estimates for k. The differences in performance decrease with increasing sample size (Table 2.1). An attractive feature of the MIX1 method is that the RMSE of the estimator $\hat{k}$ is insensitive to the value of k. This is a very desirable property for the estimator if we would like to examine the regional variation in estimates of k. In Section 2.4, we examine the regional distribution of estimates of the shape parameter in the central Appalachian region and examine the dependence of k estimates on basin properties, such as basin area, land use and land cover (LULC), and drainage density.

The RMSE of estimators of the quantile function Q(p) ($\left(Æ\left[\left(\hat{Q}(p) - Q(p)\right)^2\right]\right)^{1/2}$) for LMOM, MLE, MIX1 and MIX2 were computed for $p = 0.99$ (100-year return) and $p = 0.999$ (1000-year return) for different values of n (Table 2.2.2). The ratio (RMSE_Q-MIX2)/(RMSE_Q-LMOM) is between 0.98 and 1.02 for $n = 35$, implying that the two methods perform almost equally well in terms of quantiles. This result seems somewhat unusual in comparison to the estimation results in parameter space, where analogous ratios range from 0.84 when $k = -0.5$ to 0.97 when $k = -0.1$.

The differences between the performance of the MIXed method and LMOM estimators in the parameter and quantile spaces lead to a more detailed analysis of the distribution of the quantile estimators. Surprisingly, the bias of the LMOM estimator of k plays an important role in producing good LMOM quantile estimators. We demonstrate this feature with an example using $n = 30$ and GEV parameter values,

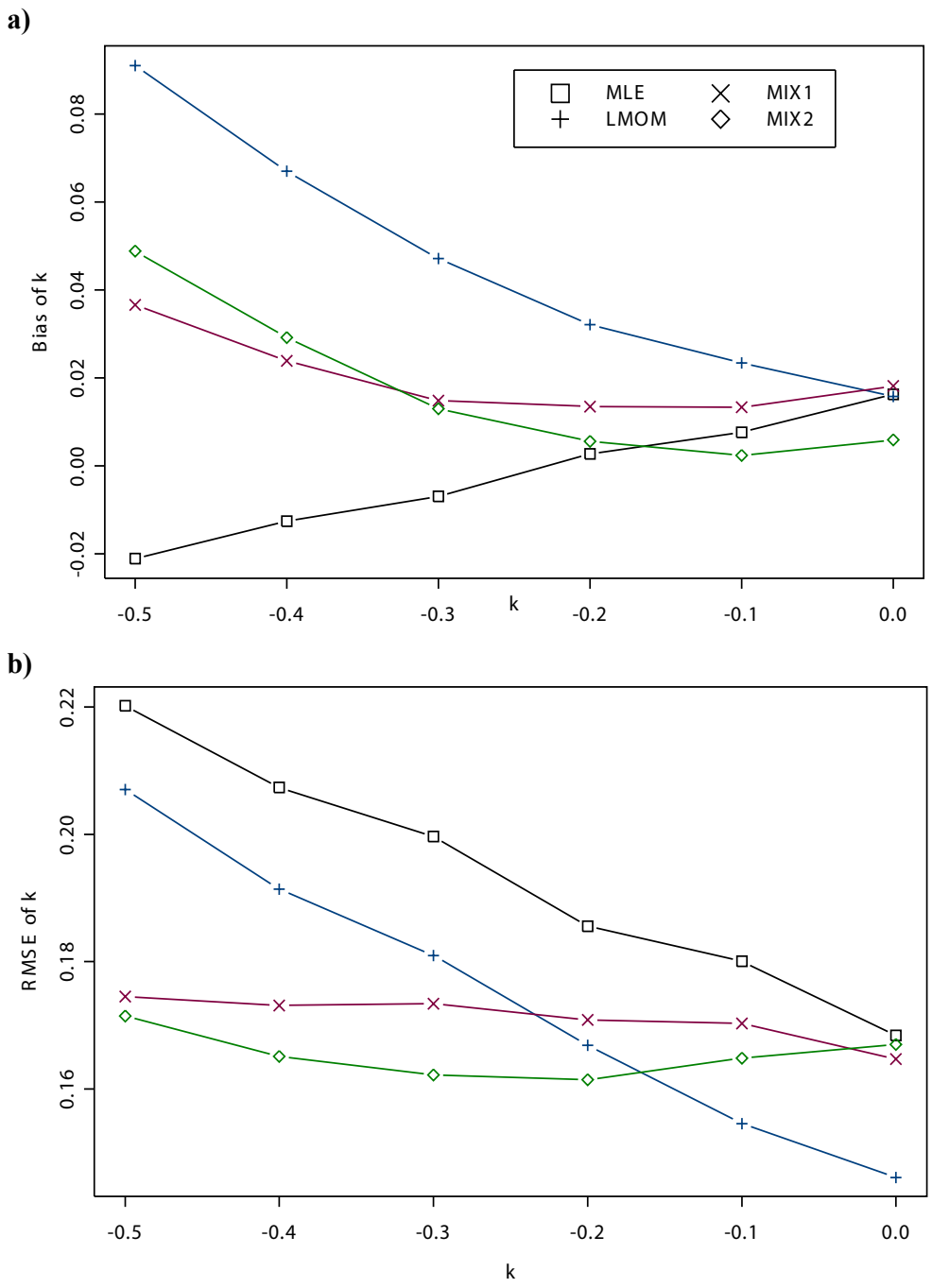

Figure 2.1: Bias **(a)** and RMSE **(b)** of the estimator $\hat{k}$ for a sample size $n = 35$ and four different estimation methods: MLE, LMOM, MIX1, and MIX2.

		$\hat{a}$ RMSE				$\hat{b}$ RMSE				$\hat{k}$ RMSE			
n	k	MLE	LMOM	MIX1	MIX2	MLE	LMOM	MIX1	MIX2	MLE	LMOM	MIX1	MIX2
30	-0.5	0.209	0.263	0.214	0.209	0.221	0.23	0.22	0.221	0.22	0.207	0.175	0.171
30	-0.4	0.194	0.23	0.199	0.194	0.216	0.22	0.217	0.213	0.207	0.191	0.173	0.165
30	-0.3	0.184	0.206	0.189	0.185	0.22	0.219	0.221	0.215	0.200	0.181	0.173	0.162
30	-0.2	0.171	0.185	0.175	0.174	0.212	0.209	0.214	0.207	0.186	0.167	0.171	0.161
30	-0.1	0.165	0.173	0.167	0.169	0.213	0.211	0.215	0.209	0.18	0.155	0.170	0.165
30	0	0.158	0.163	0.159	0.166	0.214	0.209	0.216	0.213	0.168	0.146	0.165	0.167
50	-0.5	0.158	0.2	0.166	0.157	0.164	0.174	0.166	0.167	0.160	0.172	0.138	0.141
50	-0.4	0.150	0.178	0.154	0.149	0.167	0.172	0.168	0.166	0.149	0.155	0.131	0.130
50	-0.3	0.139	0.157	0.142	0.139	0.166	0.167	0.167	0.163	0.139	0.142	0.128	0.123
50	-0.2	0.132	0.144	0.134	0.133	0.164	0.163	0.165	0.161	0.131	0.129	0.124	0.119
50	-0.1	0.126	0.133	0.128	0.129	0.166	0.164	0.167	0.163	0.130	0.122	0.126	0.121
50	0	0.119	0.124	0.12	0.123	0.163	0.161	0.164	0.161	0.119	0.111	0.118	0.117
100	-0.5	0.109	0.141	0.115	0.109	0.115	0.124	0.116	0.119	0.104	0.129	0.0975	0.104
100	-0.4	0.103	0.125	0.107	0.103	0.115	0.120	0.116	0.116	0.0982	0.117	0.0936	0.0964
100	-0.3	0.098	0.112	0.100	0.0986	0.115	0.117	0.116	0.114	0.093	0.104	0.0886	0.0878
100	-0.2	0.0914	0.100	0.0924	0.092	0.114	0.115	0.115	0.114	0.0872	0.0915	0.0845	0.082
100	-0.1	0.0893	0.0967	0.0898	0.0902	0.115	0.115	0.116	0.114	0.081	0.0823	0.08	0.0779
100	0	0.0816	0.0864	0.0818	0.0833	0.111	0.110	0.111	0.110	0.0769	0.0762	0.0767	0.0762

Table 2.1: RMSE of parameter estimates for MLE, LMOM, MIX1, and MIX2 methods for selected sample sizes and negative values of k.

		$\hat{Q}(.99)$ RMSE/$Q(.99)$				$\hat{Q}(.999)$ RMSE/$Q(.999)$			
n	k	MLE	LMOM	MIX1	MIX2	MLE	LMOM	MIX1	MIX2
30	-0.5	1.53	0.555	0.576	0.559	9.12	1.10	1.19	1.14
30	-0.4	1.03	0.53	0.552	0.54	3.63	1.12	1.16	1.14
30	-0.3	0.96	0.499	0.512	0.501	5.44	1.08	1.09	1.07
30	-0.2	0.627	0.436	0.448	0.439	2.04	0.898	0.905	0.884
30	-0.1	0.525	0.369	0.4	0.388	1.7	0.672	0.774	0.73
30	0	0.37	0.304	0.331	0.33	0.767	0.52	0.607	0.596
50	-0.5	0.736	0.482	0.539	0.497	1.87	0.971	1.14	1.01
50	-0.4	0.623	0.450	0.46	0.456	1.86	0.961	0.952	0.957
50	-0.3	0.477	0.392	0.387	0.386	1.05	0.786	0.733	0.741
50	-0.2	0.388	0.343	0.342	0.339	0.77	0.652	0.617	0.616
50	-0.1	0.329	0.288	0.299	0.292	0.624	0.497	0.525	0.507
50	0	0.258	0.235	0.248	0.248	0.443	0.378	0.414	0.413
100	-0.5	0.404	0.366	0.371	0.372	0.79	0.723	0.71	0.72
100	-0.4	0.348	0.338	0.328	0.336	0.658	0.683	0.602	0.638
100	-0.3	0.3	0.293	0.279	0.284	0.541	0.564	0.487	0.508
100	-0.2	0.252	0.241	0.236	0.234	0.435	0.417	0.397	0.392
100	-0.1	0.203	0.194	0.196	0.193	0.333	0.316	0.318	0.313
100	0	0.168	0.163	0.166	0.167	0.263	0.251	0.259	0.261

Table 2.2: Relative RMSE of quantile estimates for MLE, LMOM, MIX1, and MIX2 methods for selected sample sizes and values of k.

$a = 1$, $b = 0$ and $k = -0.3$. The covariance matrix of MIX2 estimators (Table 2.3, based on 10,000 simulation runs) is smaller than that for LMOM estimators for each element. Only the bias of $\hat{a}$ is larger for MIX2 than for LMOM. Overall, the RMSE computed using all three parameters is about 11% smaller for MIX2 than for LMOM.

Under the assumption that the estimators have a jointly Gaussian distribution, we can compute the moments of $\hat{Q}(.99)$ by numerically integrating expression (2.2) with the appropriate Gaussian probability density function (pdf) of the estimators. If we neglect the bias of the estimators and assume that they are centered at (1,0,-0.3)^T, we will find that the MIX2 method performs more precisely, as expected (Table 2.3, center). Note that the bias of the quantile estimator that we obtain under this assumption is large, nearly 12%. Taking into account the biases of the parameter estimators by centering the Gaussian pdf at the estimators' expected values in the computation, we obtain better results for the quantile estimators: biases and variances of the quantile estimators are decreased and the RMSE_Q obtained from LMOM is slightly smaller than that from MIX2. These corrected values agree with the estimates of RMSE_Q obtained from direct calculations of equation(2.2) for the simulation experiments mentioned above (Table 2.3, bottom).

The difference between the theoretically computed and simulated variances and RMSEs can be attributed to non-Gaussian properties of the estimators. Further analysis suggests that the RMSE_Q is most sensitive to $\mathbb{E}\hat{k}$. This is not surprising, because k contributes to (2.2) exponentially. RMSE_Q decreases with increasing $\mathbb{E}\hat{k}$, provided

	LMOM	MIX2
Expected values of the estimators, $\mathbb{E}\left[(\hat{a},\hat{b},\hat{k})^T\right]$	$\begin{pmatrix} 1.002 \\ 0.0142 \\ -0.2529 \end{pmatrix}$	$\begin{pmatrix} 0.9538 \\ -0.0010 \\ -0.2871 \end{pmatrix}$
Covariance matrix of the estimators, B	$\frac{1}{100}\begin{pmatrix} 4.79 & 2.91 & 1.26 \\ 2.91 & 4.26 & 1.02 \\ 1.26 & 1.02 & 3.05 \end{pmatrix}$	$\frac{1}{100}\begin{pmatrix} 4.60 & 2.47 & 1.03 \\ 2.47 & 3.21 & 0.54 \\ 1.03 & 0.54 & 2.62 \end{pmatrix}$
$\det B$	3.2212×10^{-5}	2.0724×10^{-5}
Theoretically computed values for $Q(.99)$, assuming unbiased parameter estimators		
Expected value	11.1817	11.0946
Variance	36.2456	31.9329
RMSE	6.1519	5.7723
Theoretically computed values for $Q(.99)$, accounting for biased parameter estimators		
Expected value	9.7986	10.1838
Variance	26.3093	26.7793
RMSE	5.1305	5.1817
Simulation results		
Expected value	9.7756	10.1427
Variance	24.4626	24.6872
RMSE	4.9475	4.9733

Table 2.3: Comparison of performance of LMOM and MIX2 methods in parameter and quantile spaces (see text for more details). The true values of the parameters are $(1, 01, -0.3)^T$ and the true value for $Q(.99)$ is 9.9169

that the covariance between $\hat{k}$ and $\hat{a}$ is positive. This result agrees with results of Lu and Stedinger [1992], who show that smaller RMSE_Q can be achieved by setting $\hat{k} = 0$ for certain pairs of k and n. In addition, it is clear that decreasing the covariance between $\hat{k}$ and $\hat{a}$ will result in better estimates of the quantiles. Although accurate estimation of the shape parameter k is important for proper characterization of the tail of the flood peak distribution, slight overestimation of k for the LMOM procedure results in smaller values of RMSE_Q. Improvement of the parameter estimators, therefore, does not necessarily mean improvement of the quantile estimators.

We also compared the performance of the four methods for positive values of k. We found that the MIXed methods and MLE performed roughly the same, and LMOM performed slightly better than the other methods.

It is important to note here that we can add associated constraints to the MLE optimization problem, if we have prior knowledge of the underlying physical process, and the estimates of the model parameters will almost certainly be improved. One example of such prior knowledge is including an estimate of a lower bound of the GEV distribution corresponding to $(b + a/k)$. Indeed, absurd estimates of k usually occur in situations where the smallest value in the sample is relatively close to the estimated lower bound. If we know, in addition to our sample $\{x_1, x_2, \ldots, x_n\}$, that a value $x_0 < \min\{x_1, \ldots, x_n\}$ is a value that could possibly be observed, we know $b + a/k \leq x_0$. Adding this lower bound condition to the MLE problem (2.4) will reduce the chance of obtaining absurd estimates, if not eliminate it completely. This

property is used later in the discussion of Partial Duration Series approaches, so we will illustrate it here with an example.

Martins and Stedinger [2000] discuss a sample of size 15 generated from the GEV distribution with parameters $a = 1$, $b = 0$, and $k = -0.2$ for which the MLE estimate of k is less than -2.4. This resulted in an estimate of the .999 quantile on the order of 6×10^6, while the value of the real quantile was only 14.9. The true distribution has a lower bound of -5, while the lowest value in the sample was -0.39. Suppose that we had the additional information that the value $x_0 = -1$ is feasible, i. e., there is a strictly positive probability of obtaining -1 from the underlying process. If we add this information into the problem (2.4), our estimate of k for Martins and Stedinger [2000] sample will be -0.74, resulting in an estimate of the 0.999 quantile of 208. Although the error of this estimate is still quite large, it is almost 30,000 times smaller than without the condition. Simulation experiments show that adding such lower bound condition reduces the overall error in the MLE quantile estimates by a factor of 8 in comparison with the standard MLE method. In applications involving natural events (floods, winds, etc.), it is possible that some useful additional information of this kind is available. For instance, the data might say that there was no flood peak above certain threshold during a certain year, but the value of the maximum flood peak for that year was not recorded. While it is very hard to incorporate such information into the LMOM method, it is very easy to insert it into MLE. We will discuss this approach further in the next section.

To summarize this section, for negative values of k, we conclude that : 1) MIXed methods produce better estimates of the parameters of the GEV distribution than MLE and LMOM; 2) Quantile estimates produced by MIXed methods have RMSE close to that of LMOM; 3) Difference between the performances of the MIXed and LMOM methods in quantile and parameter spaces can be explained by the correlation between parameter estimates, nonlinearity of the quantile function, and the favorable bias of the LMOM estimates; 4) MIXed methods are modifications of the MLE method, that preserve all of the advantages of the MLE method concerning incorporation of additional information; 5) Adding constraints on the lower bound of the distribution to the MLE method reduces the frequency of absurd estimates of k, and improves quantile estimation.

2.3 Extension to PDS Methods

2.3.1 GEV Parameter Estimation - PDS Approach

Partial Duration Series (PDS) models of flood peaks [Shane and Lynn, 1964; Todorovich and Zelenhasic, 1970; Karr, 1976; Lang *et al.*, 1999] (sometimes referred to as the peaks-over-threshold approach) assume that the arrival times of peaks greater than a specified threshold form a Poisson process in time, and that the distribution of peak magnitudes has a particular form. If we assume that the distribution is a Generalized Pareto (GP) distribution (as in Davison and Smith [1990]; Madsen and Rosbjerg [1997]), the annual flood peaks derived from this model have a GEV distribution (with an atom at zero, see Smith [1984]).

Assume that flood peaks above the threshold δ arrive according to a (stationary) Poisson process with rate λ [Karr, 1976], and that the peaks' magnitudes Z_1, $Z_2, \ldots$ are i.i.d. random variables independent of the arrival process, each having a GP distribution with location parameter δ, scale parameter α, and shape parameter κ. The cumulative distribution function of Z_j is then

$$F(x) = \begin{cases} 1 - \left(1 - \kappa\frac{x-\delta}{\alpha}\right)^{1/\kappa} & \kappa \neq 0, \\ 1 - \exp\left\{-\frac{x-\delta}{\alpha}\right\} & \kappa = 0. \end{cases} \tag{2.11}$$

To simplify our notation, we will write η to represent the vector $(\alpha,\ \delta,\ \kappa)^T$, and refer to the distribution function with this particular set of parameters η as $F_\eta(x)$. The corresponding probability density function will be denoted as $f_\eta(x)$. Under our assumptions, the distribution of annual flood peaks (for values greater than δ) is the same as the GEV distribution with parameters

$$\begin{aligned} k &= \kappa, \\ a &= \alpha\lambda^{-\kappa}, \\ b &= \delta + \frac{\alpha}{\kappa}(1 - \lambda^{-\kappa}). \end{aligned} \tag{2.12}$$

After estimating the parameters for the GP/PDS model, then, we can subsequently deduce the appropriate parameters for the GEV model. Madsen and Rosbjerg [1997] showed that the PDS approach can improve MLE and LMOM estimates if $\kappa < 0$ and the arrival rate of flood peaks above threshold is greater than 2 peaks per year.

Although PDS-based methods seem to perform quite well, especially for negative values of κ, there are problems with data availability that must be addressed in

applying the PDS approach to real basins. For a majority of the stream gauges maintained by USGS, the length of the PDS record is shorter than that of the annual flood peak record. In our sample of basins (section 2.4), for example, PDS data were available for only 67% of the period of record of annual flood peak data, on average. Current PDS methods for parameter estimation do not allow us to incorporate annual flood peak data for any years for which PDS data is not available into the estimation procedure. Another problem with the available data is that the value of the threshold level δ may change over time without any documentation of the fact.

With these considerations in mind, we developed a maximum likelihood method based on the magnitudes of the two largest floods each year (MLE2). This method is designed to take into account situations where there are fewer than two peaks for some years in a particular realization of the PDS process. This is not only necessary for the model to be applicable to real data, but also for the mathematical completeness of the model. Methods based on more than one peak per year have been studied previously in application to sea level heights [Smith, 1986; Dupuis, 1997].

2.3.2 The MLE2 Method

Suppose that flood peaks for a given basin obey the GP/PDS process described above. Let X and Y be random variables representing the annual maximum flood peak and the second largest flood peak for a given year, respectively.

Under the GP/PDS model, we can write the probability distribution for X:

$$\mathbb{P}\{X \leq x\} = \begin{cases} e^{-\lambda(1-F_\eta(x))} & \text{for } \; x \geq \delta, \\ e^{-\lambda} & \text{for } \; x < \delta. \end{cases} \tag{2.13}$$

The joint distribution for the two largest peaks in a given year is then:

$$\mathbb{P}\{X \leq x, Y \leq y\} = \begin{cases} e^{-\lambda(1-F_\eta(y))}\left(1 + \lambda F_\eta(x) - \lambda F_\eta(y)\right) & \text{for } \; \delta < y < x, \\ e^{-\lambda}(1 + \lambda F_\eta(x)) & \text{for } \; 0 \leq y < \delta < x, \\ e^{-\lambda} & \text{for } \; 0 \leq y < x < \delta, \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

Under the GEV/AMS approach, we approximate the expression in (2.13) by $G_\theta(x)$ with parameters θ related to those of GP/PDS η through the relationships (2.12). That means that (2.14) will be approximated by

$$\mathbb{P}\{X \leq x, Y \leq y\} = G_\theta(y)[1 + \log G_\theta(x) - \log G_\theta(y)] \qquad \text{for } \; \delta < y < x, \tag{2.15}$$

and the joint probability density function for X and Y is just $g_\theta(y)g_\theta(x)/G_\theta(x)$ for $\delta < y < x$. Using this argument, we can construct a likelihood function for the observations of the two largest peaks per year. Also note that

$$\mathbb{P}\{X < \delta\} = G_\theta(\delta) = e^{-\lambda}. \tag{2.16}$$

Consider a basin with m years of PDS record with threshold δ. Among these m years of PDS record, there are m_0 with no peaks above δ, m_1 years with only one peak above δ (let $z_1, z_2, \ldots, z_{m_1}$ denote the magnitudes of these peaks), and m_2 years with

2 or more peaks (the largest peaks per year will be denoted by $x_1, x_2, \ldots, x_{m_2}$ and the second largest by $y_1, y_2, \ldots, y_{m_2}$). The log-likelihood function of these observations is, then:

$$\begin{aligned}\log \Lambda(\theta|z,x,y) = & - m_0\lambda + m_1(\log(\lambda) - \lambda) + m_2 \log(1 - e^{-\lambda} - \lambda e^{-\lambda}) \\ & + \sum_{i=1}^{m_1} \log(f_\eta(z_i)) \\ & + \sum_{j=1}^{m_2} \left[\log(g_\theta(y_j) + \log(g_\theta(x_j)) - \log(G_\theta(x_j))\right], \end{aligned} \tag{2.17}$$

where the parameters θ, η, and λ are connected through the relationships in (2.12). Here, the first three terms on the right hand side correspond to the probability of having 0, 1, or "2 or more" peaks in a given year, respectively. The term on the second line is the log-likelihood of obtaining the particular values $z_1, z_2, \ldots, z_{m_1}$ of the single flood peaks over threshold that occurred during the m_1 years, and the term on the last line corresponds to the log-likelihood of obtaining the particular pairs of two largest peaks for the m_2 years that we observed two or more peaks. Substituting into (2.17) the derivative of (2.11) for f_η, G_θ and g_θ from (2.1), and using the relationships (2.12), we obtain the following expression for the log-likelihood function of the two maxima:

$$\begin{aligned}\log \Lambda(\theta|z,x,y) &= -m_0 e^{h(\delta)} + m_1(h(\delta) - e^{h(\delta)}) + m_2 \log(1 - \exp(-e^{h(\delta)})(1 + e^{h(\delta)})) \\ &- (m_1 + 2m_2)\log a + (1-k)\sum_{i=1}^{m_1} h(z_i) \\ &+ (1-k)\sum_{j=1}^{m_2}[h(y_j) + h(x_j)] - \sum_{j=1}^{m_2} e^{h(y_j)}, \end{aligned} \tag{2.18}$$

where $h(x) = \frac{1}{k}\log\left(1 - \frac{k(x-b)}{a}\right)$. The time required to evaluate this expression computationally is not much greater than the time necessary to evaluate the standard likelihood function for the MLE method.

The maximum likelihood estimator based on the two largest maxima per year (MLE2) is the (local) maximum of the likelihood function in (2.18). This can be equivalently written as the solution to the following optimization problem:

$$\begin{aligned}
\text{maximize} \quad & \log \Lambda(\theta|x) \\
\text{subject to} \quad & k(z_i - b) \leq a \qquad i = 1, \ldots, m_1 \\
& k(x_j - b) \leq a \\
& k(y_j - b) \leq a, \qquad j = 1, \ldots, m_2 \\
& k(\delta - b) \leq a, \\
& a > 0
\end{aligned} \tag{2.19}$$

The first three conditions correspond to the restriction that the measured peak magnitudes must be feasible values for the varying GP distribution. The last condition is the restriction that the scale parameter is positive. The second-to-last constraint ("the δ constraint") is the condition that the threshold level itself must be a feasible value for the GEV distribution. This constraint is necessary to ensure that the likelihood function can be evaluated. When $k > 0$, since all z_i, x_j, and y_j are greater than δ, the δ constraint never becomes binding on the problem (the upper bound of the GEV distribution is obviously greater than δ); when $k = 0$ the δ constraint

becomes $a \geq 0$, which is less restrictive than the last constraint. For $k < 0$, the δ constraint has an effect, as it requires that the lower bound for annual flood peaks be less than δ. This corresponds to the positive probability of having no peaks above the threshold during a given year. As mentioned in section 2.2, finding an *a priori* expression for a tighter upper bound than the smallest observation in the sample for the lowest value the GEV distribution can produce is valuable for reducing the number of absurd estimates of k.

As with the regular MLE method, we investigated different constraints that can be added to problem (2.19) in order to improve the estimates of the quantiles. The equivalent of the MIX1 method in this case involves adding to problem (2.19) the constraint:

$$b = \frac{\hat{\lambda}_1}{1 - e^{-h_\theta(\delta)}} - \frac{a}{k}\left(1 - \Gamma(1+k)\right), \tag{2.20}$$

where $\hat{\lambda}_1$ is the mean value (estimate of the first L-moment) of all annual maxima above the threshold δ, that is, $\hat{\lambda}_1 = \dfrac{1}{m_1 + m_2}\left[\sum_{i=1}^{m_1} z_i + \sum_{j=1}^{m_2} x_j\right]$. This condition helps to eliminate absurd estimates of k, and improves the estimation. We will refer to this method as MLE2-MIX1.

The procedure that we used for Monte Carlo simulations in order to test the MLE2 and MLE2-MIX1 methods is the same as described in Madsen *et al.* [1997a]. It is designed so that we can compare the performance of the AMS methods to that of the PDS-based methods. It exploits the fact that in GP/PDS model peaks above higher threshold levels from the same process also have a GP distribution with the same

value of k . Specifically, if peaks above the threshold level δ_0 arrive according to a Poisson process with rate λ_0, and have i.i.d. magnitudes independent of the Poisson process with GP distribution with parameters $\eta_0 = (\alpha_0, \delta_0, \kappa)$, peaks above the higher threshold

$$\delta_1 = \begin{cases} \delta_0 + \frac{\alpha_0}{\kappa}\left[1 - \left(\frac{\lambda_1}{\lambda_0}\right)^k\right] & \kappa \neq 0, \\ \delta_0 + \alpha_0 \log\left(\frac{\lambda_0}{\lambda_1}\right) & \kappa = 0. \end{cases} \tag{2.21}$$

arrive according to a Poisson process with rate λ_1 ($\lambda_1 < \lambda_0$), are i.i.d., and have a GP distribution with scale parameter $\alpha_1 = \alpha_0 + \kappa(\delta_1 - \delta_0)$, location parameter δ_1, and the same value of shape parameter κ (see Madsen *et al.* [1997a]).

The simulation procedure for testing can be described as follows:

- Pick parameters for GP/PDS model to produce simulated samples: Choose a large arrival rate λ_0 so that the probability of obtaining zero peaks during any given year is very small. Choose a threshold level δ_0, scale parameter α_0, and number of years in the record m. Choose the value of the shape parameter κ and the arrival rate λ_1 for which we will test the procedure. Compute $\alpha_1 = \alpha_0 + \kappa(\delta_1 - \delta_0)$ and δ_1 from Equation (2.21).

- Generate arrival times from the Poisson process with rate λ_0, and generate flood peak magnitudes from the GP distribution with parameters δ_0, α_0, and κ for each arrival time.

- Extract the PDS sample corresponding to all peaks higher than δ_1 and their arrival times: compute the number of years with no peaks, m_0, the number

of years with 1 peak, m_1, and the number of years with more than one peak, $m_2 = m - m_0 - m_1$. For years with only one peak, record the magnitude of the peak z_i, $i = 1, \ldots, m_1$, and for years with two or more peaks record the magnitudes of the largest and second largest peaks, x_j and y_j, $j = 1, \ldots, m_2$.

- Solve the optimization problem (2.19), and record the MLE2 estimates.
- Extract the AMS sample from the original process by recording the largest peak for each year.
- Use LMOM and MIX1 method to estimate the parameters of the GEV distribution based on the AMS sample.

Monte Carlo simulations were performed according to the procedure above for records 20-100 years long, values of λ_1 from 2 to 7, and values of shape parameter k between -0.5 and 0.0. The results (Figure 2.2) for 35-year records show that incorporating information about the second maxima decreases the estimation error of the shape parameter for. The results obtained for all other sets of the model parameters were similar to this one (Table 2.4), and the decrease in the error was especially significant for very negative values of k. Although the MLE2 method sometimes produces absurd results, it does so far less frequently than MLE, in part due to the "δ condition" in problem (2.19) (see the example in the previous section). The MLE2-MIX1 method did not produce any absurd estimates of k, and its estimates of k have the smallest RMSE. In terms of quantiles, though incorporating second maxima

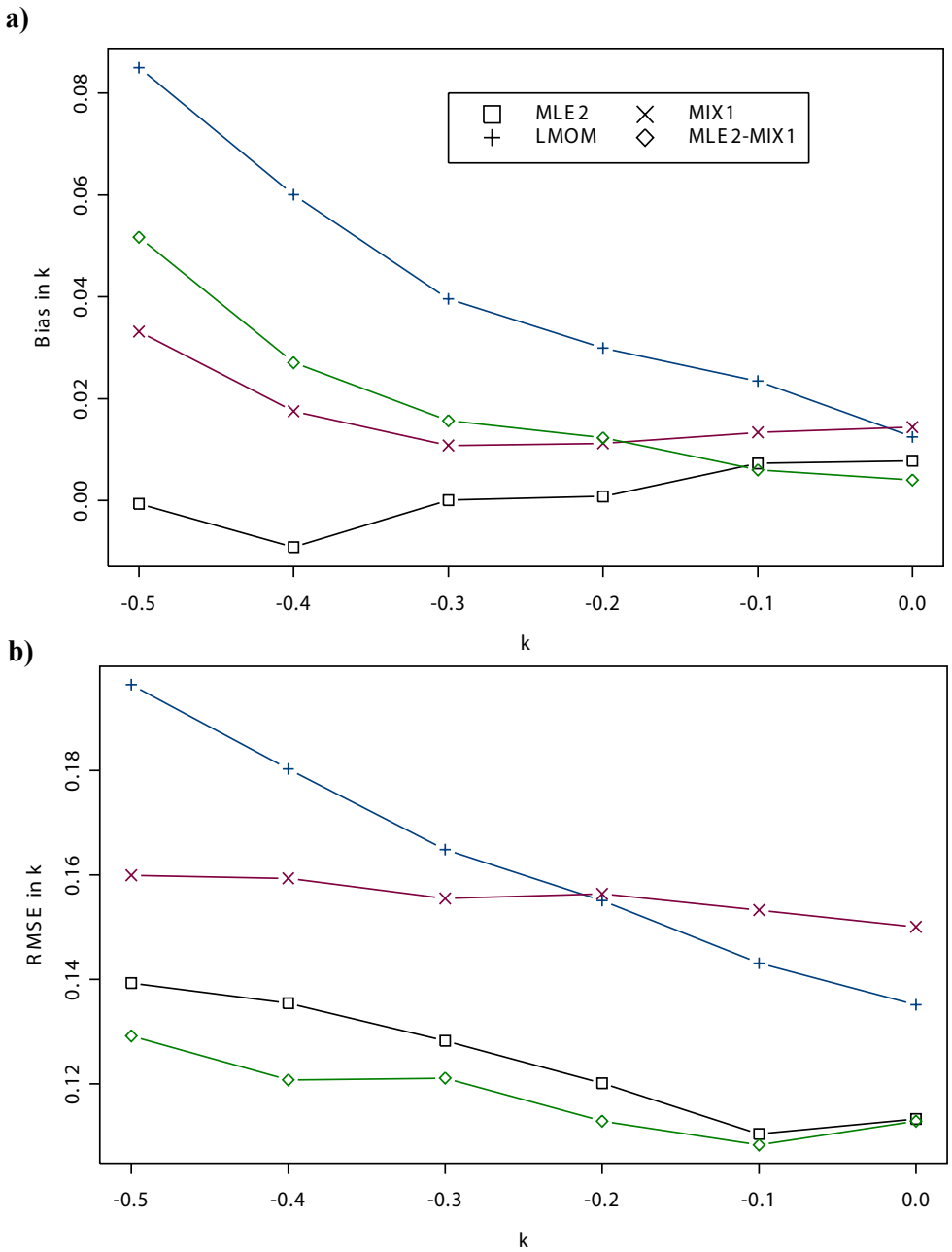

Figure 2.2: Bias **(a)** and RMSE **(b)** of the estimator $\hat{k}$ for record length 35 years and four different estimation methods: MLE2, LMOM, MIX1, and MLE2-MIX1.

		$\hat{k}$ RMSE		
n	k	LMOM	MLE2	MLE2-MIX1
35	-0.5	0.194	0.139	0.129
35	-0.4	0.173	0.135	0.121
35	-0.3	0.170	0.128	0.121
35	-0.2	0.152	0.12	0.113
35	-0.1	0.141	0.110	0.108
35	0	0.136	0.113	0.113
50	-0.5	0.180	0.122	0.118
50	-0.4	0.153	0.109	0.104
50	-0.3	0.151	0.105	0.102
50	-0.2	0.130	0.10	0.0964
50	-0.1	0.114	0.0872	0.0868
50	0	0.111	0.0864	0.086
100	-0.5	0.129	0.0815	0.0847
100	-0.4	0.117	0.078	0.0786
100	-0.3	0.105	0.0704	0.07
100	-0.2	0.0937	0.0685	0.0681
100	-0.1	0.0818	0.0655	0.0655
100	0	0.075	0.0569	0.0572

Table 2.4: RMSE of the estimates of k from LMOM, MLE2, and MLE2-MIX1 methods for selected sample sizes and negative values of k.

		$\hat{Q}(.99)$ RMSE/$Q(.99)$			$\hat{Q}(.999)$ RMSE/$Q(.999)$		
n	k	LMOM	MLE2	MLE2-MIX1	LMOM	MLE2	MLE2-MIX1
35	-0.5	0.482	0.655	0.502	1.02	1.67	1.02
35	-0.4	0.418	0.55	0.412	0.882	1.23	0.798
35	-0.3	0.416	0.452	0.402	1.01	1.03	0.83
35	-0.2	0.304	0.331	0.295	0.642	0.664	0.552
35	-0.1	0.248	0.242	0.243	0.538	0.461	0.45
50	-0.5	0.456	0.536	0.465	0.964	1.16	0.92
50	-0.4	0.38	0.408	0.371	0.807	0.807	0.717
50	-0.3	0.361	0.341	0.332	0.81	0.675	0.638
50	-0.2	0.263	0.264	0.252	0.562	0.502	0.469
50	-0.1	0.202	0.187	0.192	0.395	0.323	0.330
100	-0.5	0.363	0.331	0.375	0.795	0.611	0.736
100	-0.4	0.316	0.274	0.308	0.719	0.494	0.595
100	-0.3	0.259	0.218	0.232	0.599	0.394	0.43
100	-0.2	0.188	0.173	0.173	0.365	0.303	0.301
100	-0.1	0.145	0.143	0.145	0.265	0.245	0.248

Table 2.5: RMSE of quantile estimates for LMOM, MLE2, and MLE2-MIX1 methods for selected sample sizes and values of k.

with MLE2 decreased the RMSE in the quantiles by a factor of 8 in comparison with the standard MLE method, the LMOM/AMS and MIX1/AMS methods still produce better quantile estimates than MLE2 for small record sizes (≤ 50) years, primarily due to the absurd estimates of k. MLE2-MIX1 has the smallest RMSE_Q (see Table 2.5) for $k \leq -0.2$, and $n \leq 50$.

2.4 Analysis of Flood Peak Data from the Central Appalachian Region

GEV parameter estimation procedures were applied to flood peak observations from a sample of 104 USGS stream gauging stations in the central Appalachian region (see Smith [1992] and Hosking and Wallis [1996] for previous analyses of this data set). The gauging station locations are shown in Figure 2.3, on a map of the underlying elevation. These stations were selected because they each recorded at least 30 years of data and their basins are unregulated by dams (see Smith [1992]). The important questions that we would like to address in our analysis are: 1) how variable are at-site estimates of the shape parameter k within the region; 2) how can this variability be explained; and 3) how do the estimates of the three GEV distribution parameters depend on morphological and land cover properties of the drainage basins? To address the third question, basin morphological information was computed for each of the basins from digital elevation data (DEM), and land use - land cover (LULC) data (based on LANDSAT Thematic Mapper images from 1990 - 1992; see Loveland and

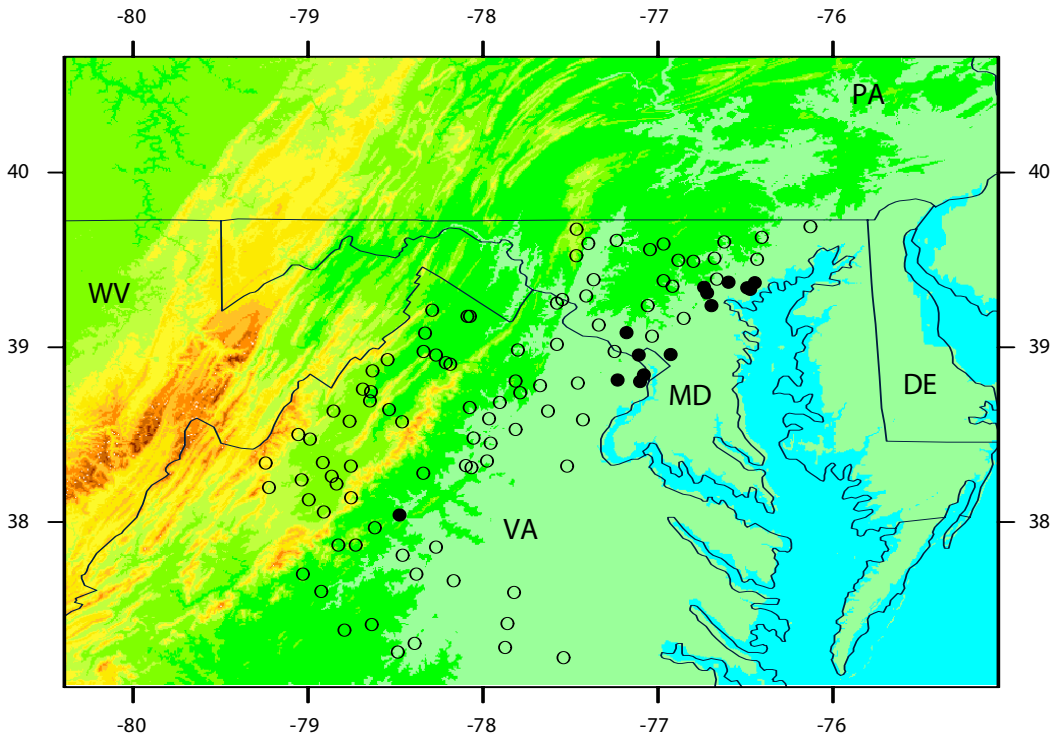

Figure 2.3: A map of the central Appalachians region showing the locations of stream gauging stations. Solid circles denote urban basins and empty circles denote rural basins.

Shaw [1996]). From these sources, we computed the basin drainage area, basin relief, relief ratio, and drainage density.

The MLE2-MIX1 and LMOM methods were used to estimate the parameters of the GEV distribution for each of the basins, and the range of LMOM k estimates for the central Appalachian basins was between -0.74 and 0.02, with a median value of -0.37. The MLE2-MIX1 method, estimates of k ranged from -0.82 to 0.01, with a median value of -0.40. In Hosking and Wallis [1996], the same sample of basins is divided into 5 groups, and regional estimates of k for each group was determined. The values of these estimates ranged from -0.45 to -0.24, which agrees with our distribution of at-site estimates. These values are more negative than what has conventionally been considered physically reasonable [Martins and Stedinger, 2000]. When k is less than $-1/3$, the flood peak distribution has an infinite third moment, and when k is less than $-1/2$, the distribution has infinite variance. Very negative values of k suggest that the distribution of flood peaks has very heavy tails.

Figure 2.4 shows the values of at-site estimates of k obtained using the LMOM and MLE2-MIX1 methods plotted against the corresponding empirical quantiles of the standard Gaussian distribution. The lines on the plot correspond to Gaussian approximations of the distributions of the respective estimates of k for the two methods, assuming the true value of the k is -0.42 and the period of record is 46 (the average number of years of record available for our sample). From this plot, we conclude that the distribution of the estimates of k for our sample is approximately Gaussian, and

the variability of the estimates can be explained by the variability of the estimators used.

The estimates of the shape parameter k for the central Appalachian basins do not exhibit a systematic dependence on basin morphological properties or land cover. The relationship between estimates of k and basin area (Figure 2.5) is representative of the relationship for other basin descriptors. Regression analysis between the estimates of k and the basin descriptors (Table 2.6) produced R^2 values ranging from 0.006 (for the gauge elevation) to 0.04 (for drainage density).

According to index flood theory, values of the scaling parameter a and the location parameter b should exhibit a log-log relationship with drainage area A. For the central Appalachian basins, this property holds (Figure 2.6). A significant contribution to the variability in this scaling relationship is related to land cover properties. It was found that basins with a higher percentage of urbanized land have higher values of the scaling parameter a and b (Figure 2.6). These basins respond as though they have larger effective areas than their true basin drainage area (see Leopold [1968]; Smith *et al.* [2001]). If we define the effective area to be:

$$A_{eff} = A(1 + C \times \text{fraction of impervious cover}),$$

where C is a constant we can find the C that maximizes the R^2 of the linear regression between $\log \hat{a}$ and $\log A_{eff}$. The value of C determined for our sample is 9.5, and the new relationship $\log \hat{a} \sim \log A_{eff}$ is shown in Figure 2.7. Notably, the effect of urban

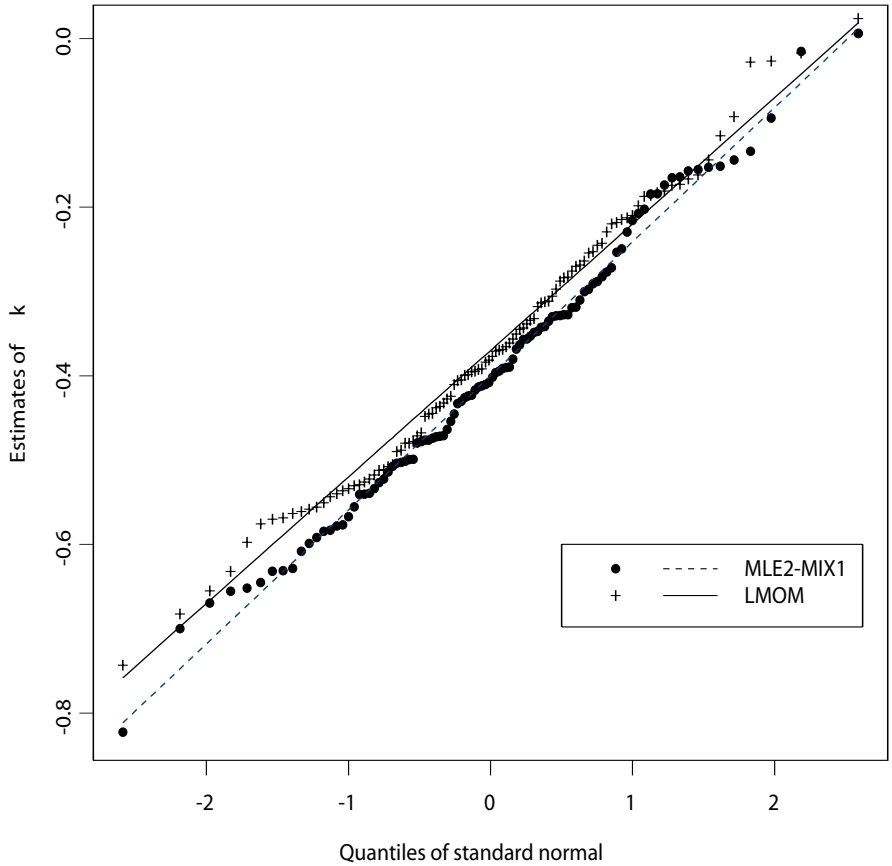

Figure 2.4: A normal QQ-plot of at-site estimates of k for the central Appalachians basins. The lines on the plot correspond to Gaussian approximations of the distributions of LMOM and MIX1 estimates of k.

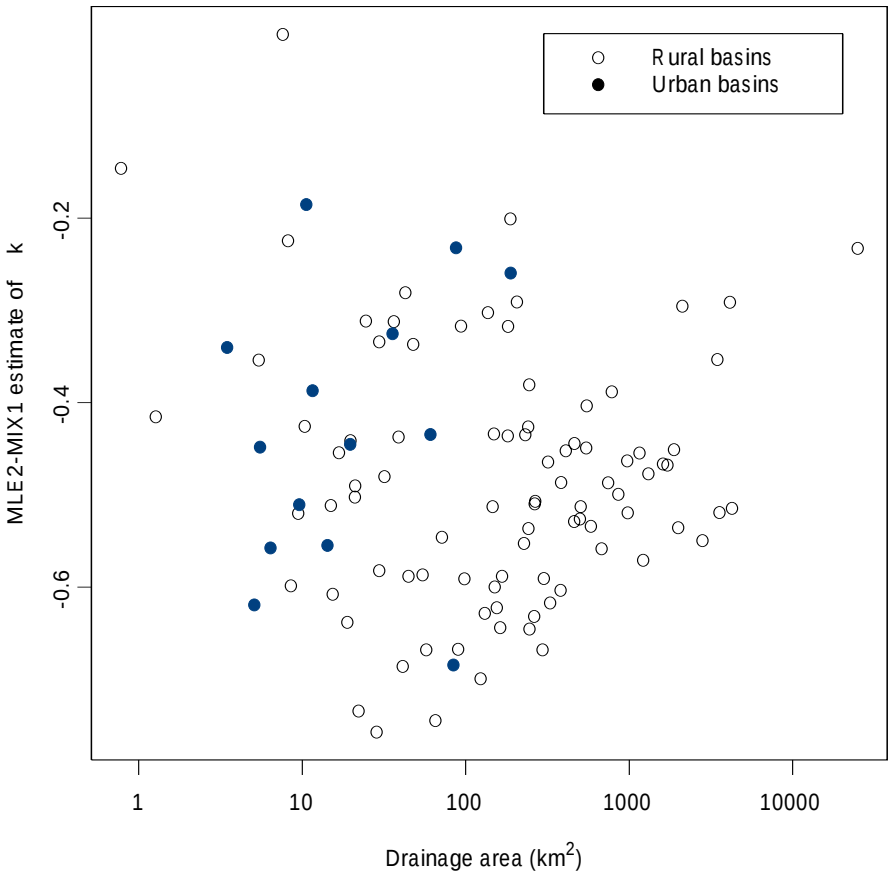

Figure 2.5: Dependence of MLE2-MIX1 estimate of k on basin's drainage area.

Basin Descriptor	R^2
Drainage Area (log)	0.0072
Elevation (log)	0.008
Relief Ratio (log)	0.03
Percent of Forest Cover	0.002
Percent of Urban Cover	0.03
Drainage Density	0.06

Table 2.6: R^2 of linear regression between basin descriptors and MLE2-MIX1 estimates of k

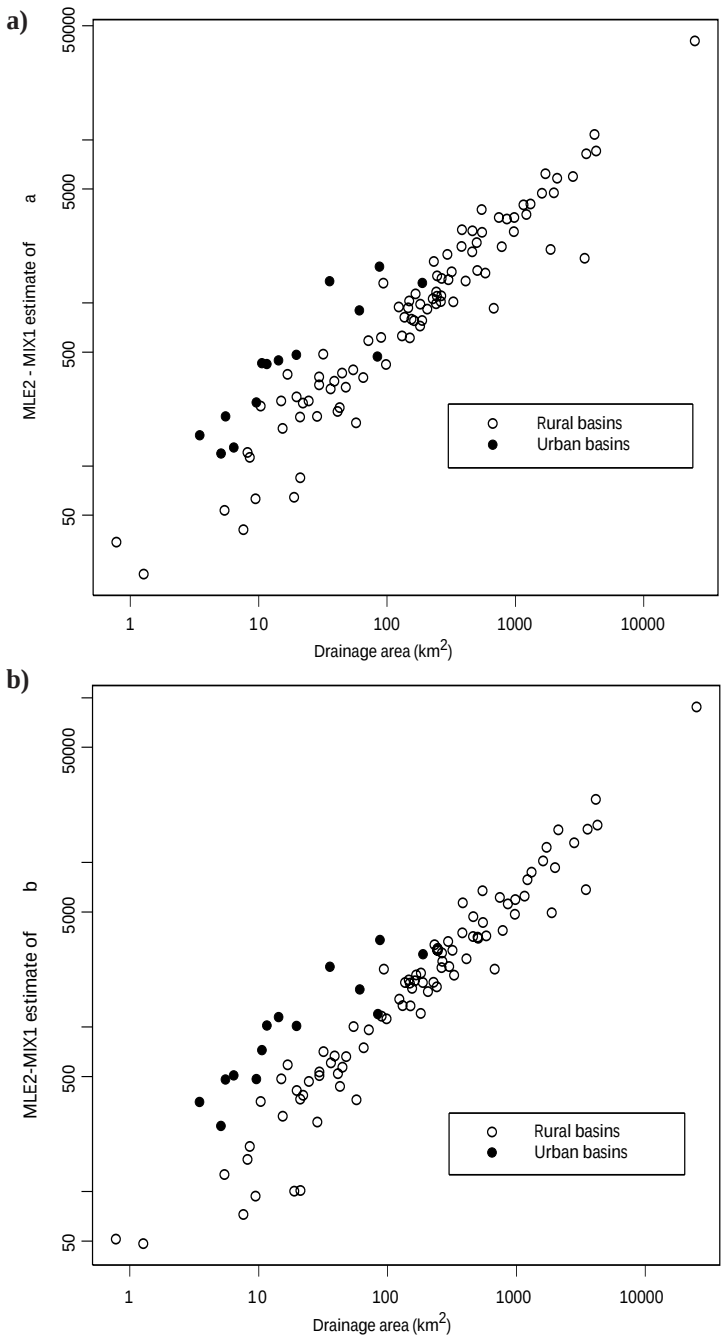

Figure 2.6: Dependence of MLE2-MIX1 estimate of a **(a)** and b **(b)** on basin drainage area. Solid circles denote urban basins, empty circles denote rural basins.

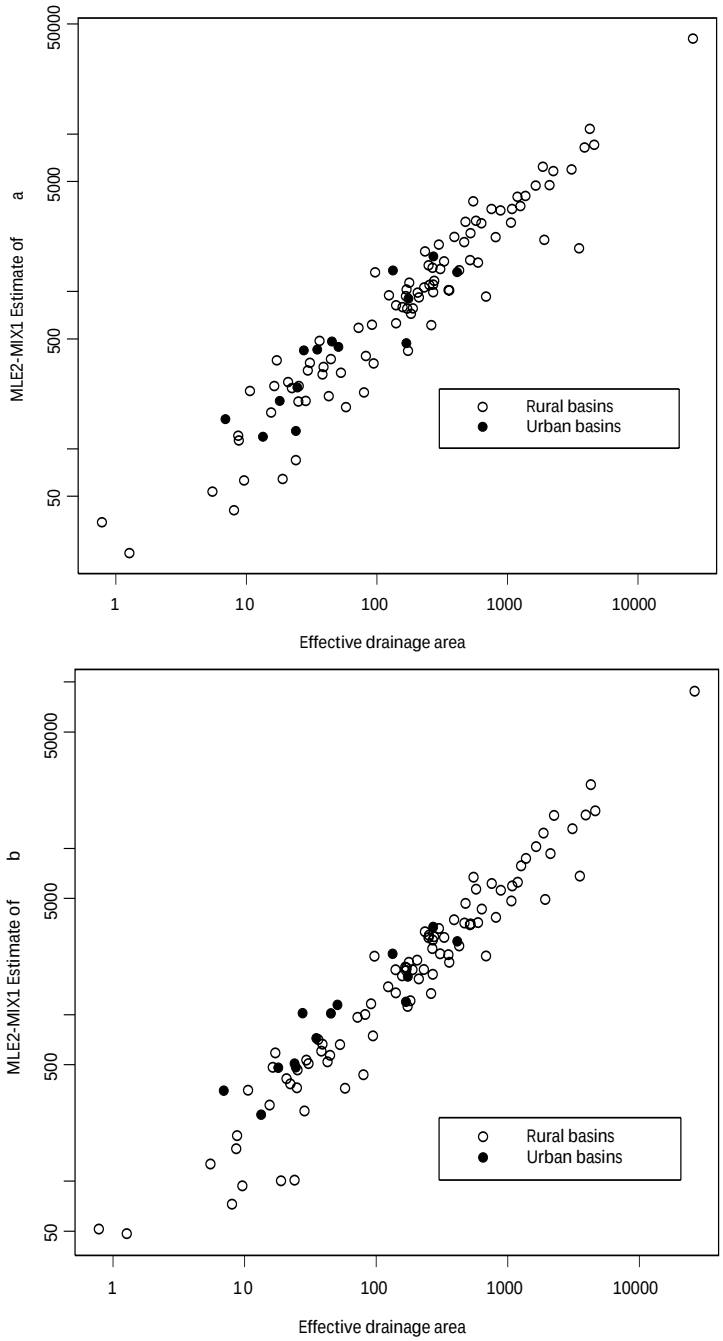

Figure 2.7: Dependence of MLE2-MIX1 estimate of a **(a)** and b **(b)** on effective area A_{eff}. Solid circles denote urban basins, empty circles denote rural basins.

hydrology on GEV parameter estimates is quite important for location and scale, but not as important for the shape parameter k (Figure 2.5).

The "extreme" estimates of the shape parameter k are dependent on the specific flood hydrological properties of the central Appalachian region. A sample of 34 drainage basins from the southern Appalachians yields estimates of k ranging from -0.53 to 0.24, with a median value of -0.11. Selection of the 34 basins (all of which are in North Carolina) was based on identical constraints to those used for the 104 central Appalachian basins. A systematic difference in flood peak distributions between the two regions is that southern Appalachian basins exhibit markedly lower variability than central Appalachian basins, as indicated by the coefficient of variation of annual flood peaks. This contrast in flood peak distributions is linked to contrasts in magnitudes of extreme flood peaks, as illustrated in Figure 2.8 by envelope curves of flood peaks for the two samples. The central Appalachian region has experienced some of the largest unit discharge flood peaks in the United States east of the Rocky Mountains [Smith *et al.*, 1996; Eisenlohr, 1952; Hack and Goodlett, 1960] from summer thunderstorms. The "extreme" estimates of k reflect the influence of extreme flood magnitudes.

A key question to resolve is whether extreme estimates of the GEV tail parameter k necessarily mean that the distribution of flood peaks indeed has heavy tails or whether alternative stochastic models can explain the estimated values of k. One possible scenario for negative estimates of k without thick tails is that the GEV

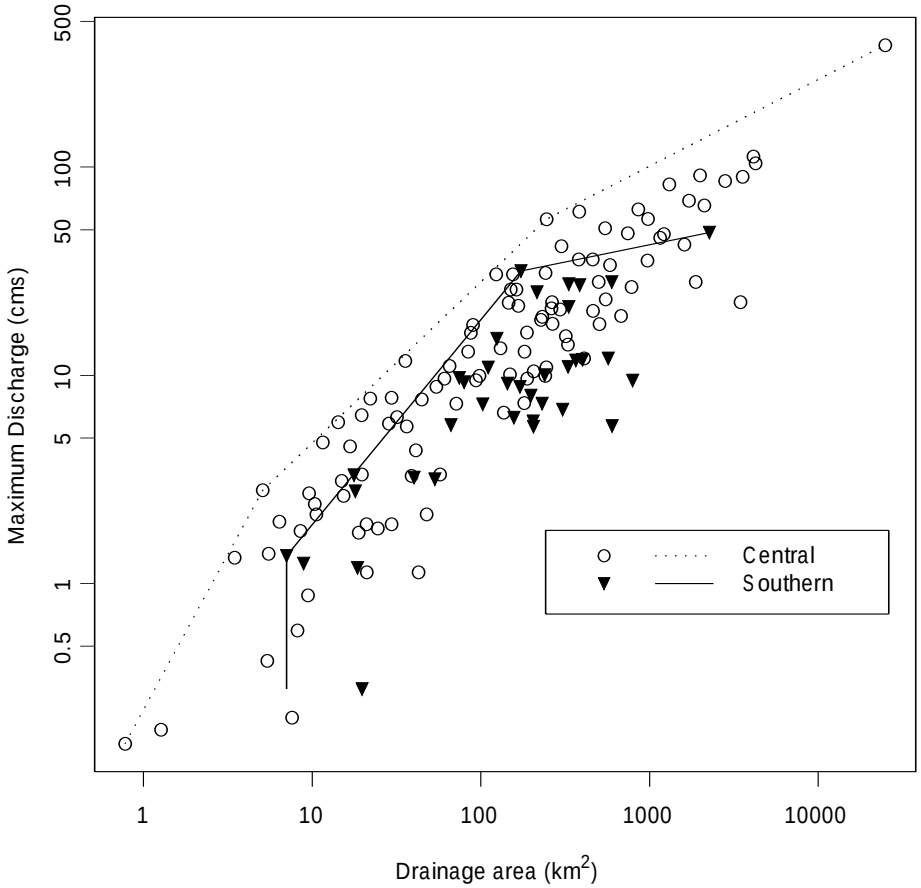

Figure 2.8: Envelope curves of flood peaks for the central and southern Appalachians regions.

distribution is not an appropriate distribution for annual flood peaks. If, for example, we estimate the three parameters of the GEV distribution for samples of size $n = 50$ drawn from an exponential distribution the estimated values of k will be centered at -0.38 for MIX1 estimators and at -0.2 for the LMOM estimators. The exponential distribution is from the Gumbel domain of attraction ($k = 0$), so it does not have thick tails, but can easily be shown to be a poor choice for modeling flood peak distributions. We must determine if there are good alternative models for annual flood peak distributions with thin tails, but large negative estimates of k.

One heuristic explanation for the contrasts in flood distributions between the central and southern Appalachians given above, relies on the influence of a particular type of flood event produced by summer thunderstorms. This notion points to a particular class of alternative models which involve a mixture of different distributions. Let V_1 and V_2 be independent random variables having distributions from the same family, with $\mathbb{E}[V_2] > \mathbb{E}[V_1]$. Again, let X be a random variable representing an annual flood, and suppose that $X = V_1$ with probability p and $X = V_2$ with probability $(1-p)$. V_1 and V_2 can be thought of as flood peaks occurring from different classes of storms; for example, V_2 might represent flood peaks occurring due to tropical storms and V_1 might represent flood peaks due to summer thunderstorms [Sturdevant-Rees *et al.*, 2001; Smith *et al.*, 1996]. If V_1 and V_2 have a GEV distribution with $k = 0$, i.e., both flood populations have a Gumbel distribution, it is possible to obtain estimates of k centered at -0.4 for flood-length samples with the following formulation:

- $V_1 \sim$ GEV with $k_1 = 0$ (Gumbel), $a_1 = \nu_1$, $b_1 = 0$;

  $V_2 \sim$ GEV with $k_2 = 0$, $a_2 = c\nu_1$, $b_1 = 2c\nu_1$, $c > 1$;

Table 2.7 demonstrates how quantiles of this Gumbel mixture model differ from those of the GEV distribution for a particular set of parameters: although $Q(.90)$ is very similar for the two models, $Q(.99)$ for the Gumbel mixture is less than half that for GEV, and $Q(.999)$ for GEV is approximately 5 times larger than that for the Gumbel mixture. This simple example shows the importance of the particular parametric model for quantile estimation. Figure 2.9 demonstrates the distribution of LMOM estimates of k for both models and sample size $n = 50$. This distribution is centered at -0.37 for both models, and the variance of estimates for the Gumbel mixture is *smaller* than that for GEV estimates.

The next question is whether we can distinguish between the two models based on observations of random samples. Figure 2.10 shows the probability density functions for the two models, and they are, unfortunately, very close to one another. Figure 2.11 shows a typical quantile-quantile plot of Gumbel mixture and GEV sample quantiles versus theoretical quantiles of the GEV distribution with shape parameter $k = -0.42$ (the true shape for the GEV model in the simulations). The two samples are so similar that it is impossible to distinguish between the two models solely on the basis of the sample observations. To conclude that central Appalachians flood peaks have thick tails, it is necessary to determine that the GEV rather than the Gumbel mixture distribution is the appropriate model for the flood peaks. Detailed studies of basin

	GEV	Gumbel Mixture
Model Parameters	$a = 1.4$	$\nu_1 = 1$
	$b = 0.2$	$c = 3.3$
	$k = -0.42$	$p = 0.83$
$Q(.90)$	5.31	5.58
$Q(.99)$	18.74	8.23
$Q(.999)$	52.16	10.57

Table 2.7: Comparison of the quantiles of the two stochastic models. The Gumbel mixture model is described in the text.

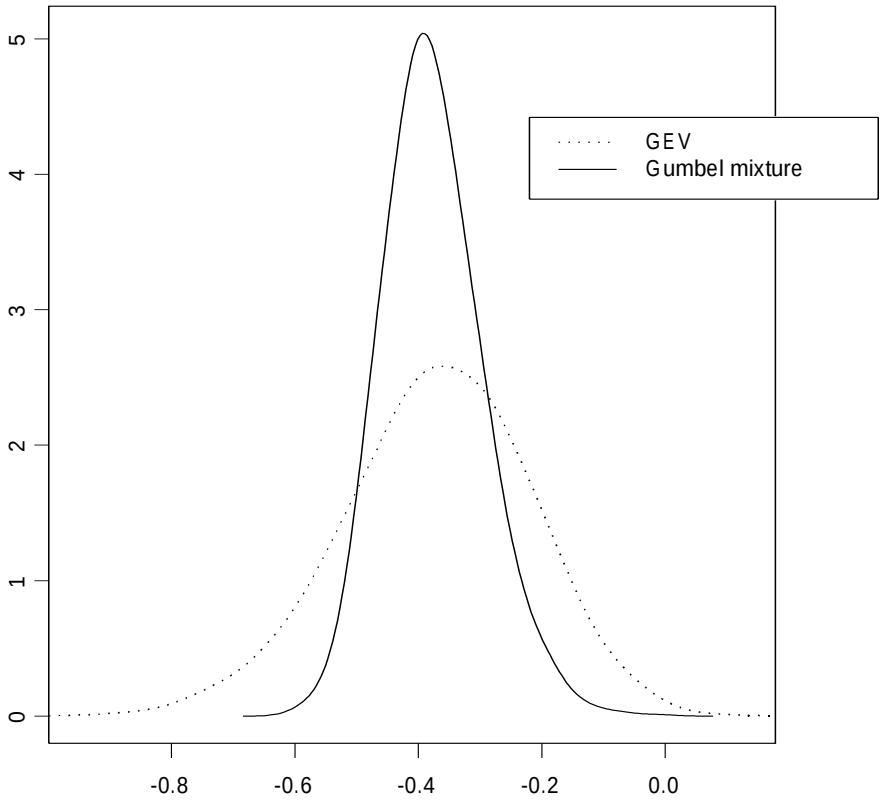

Figure 2.9: Distribution of the estimates of k for GEV distribution and the Gumbel mixture model with the parameters given in Table 2.7 and sample size $n = 50$.

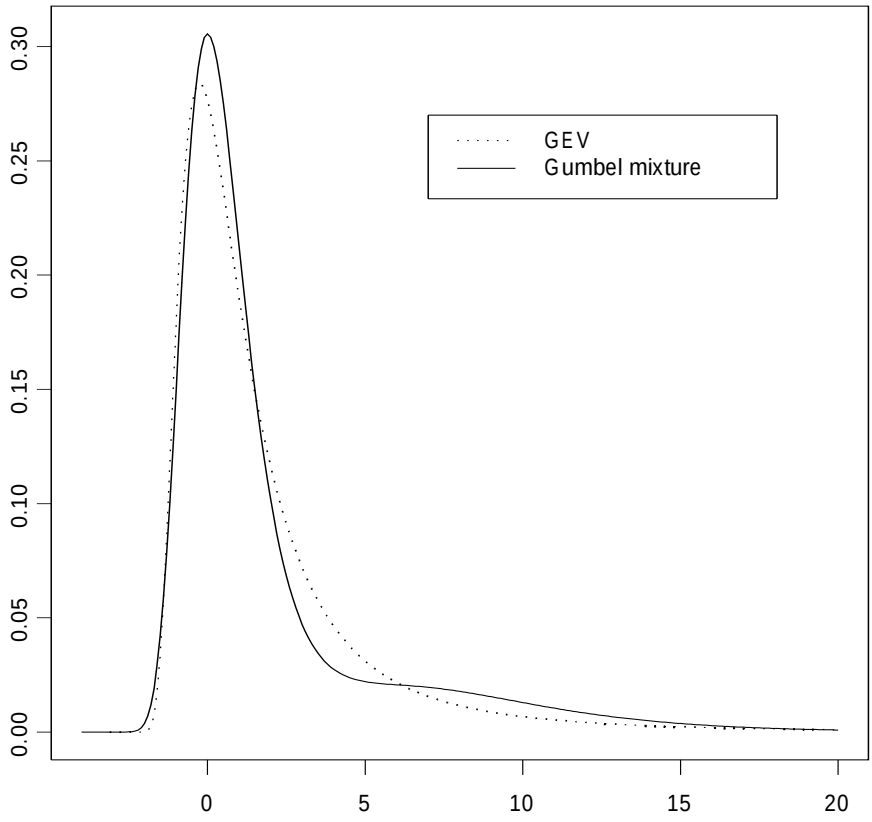

Figure 2.10: Probability density function of the GEV distribution and the Gumbel mixture model with the parameters given in Table 2.7.

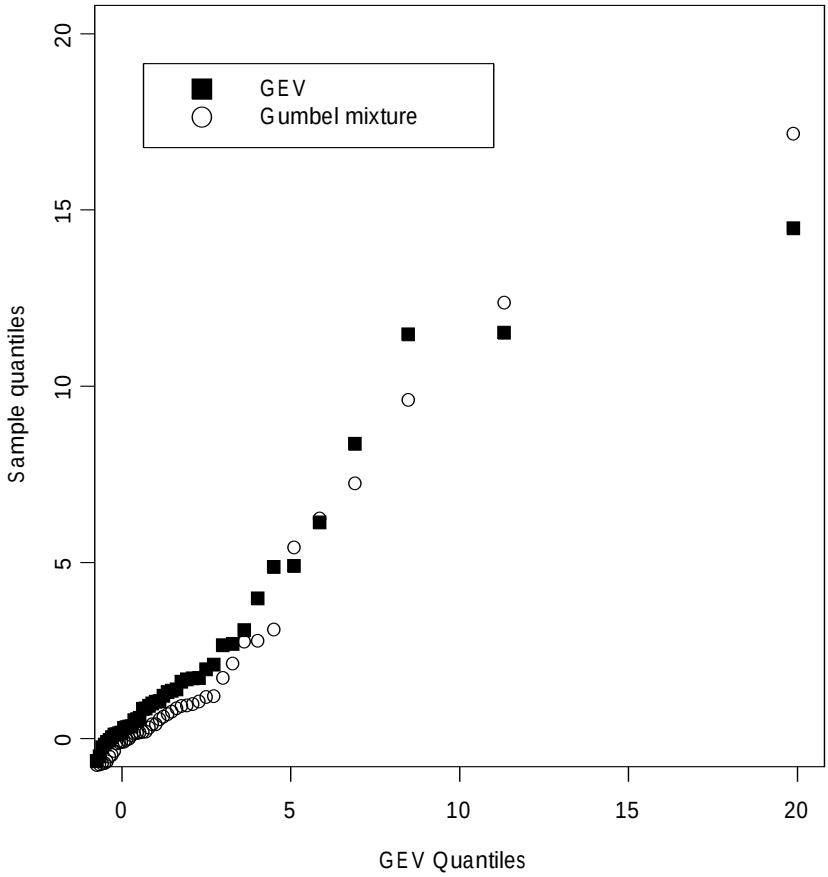

Figure 2.11: A typical quantile-quantile plot for samples generated from the GEV distribution (solid squares) and Gumbel mixture (empty circles) versus quantiles of the GEV distribution with shape parameter $k = -0.42$.

response to different types of extreme rainfall might help to address this problem by providing a physical basis upon which to choose one of the models.

2.5 Asymptotic Properties of MIXed Estimators

In this section, we study the asymptotic properties of the MIXed estimators in the general case; we will consider a Maximum Likelihood method with additional constraints in its optimization problem imposed by U-statistics. We show that such estimators are strongly consistent, under certain regularity conditions, and have asymptotically jointly Gaussian distributions. Since L-moments can be represented with U-statistics [Hosking, 1986], these results can be immediately applied to the MIX1 and MIX2 methods. For simplicity, we will analyze a distribution function depending on only two parameters. The extension of these arguments to higher dimensions is straightforward.

The arguments below can be outlined as follows: we consider a density function $p(x, \theta, \phi)$ of a random variable X, dependent on two parameters θ and ϕ, and a functional $U(\theta, \phi)$ with an unbiased estimator $\hat{U}$ such, that the equation $U(\theta, \phi) = u$ can be solved for ϕ and the solution is $f(\theta, u)$. In the traditional Maximum Likelihood (ML) method, we write the likelihood function $L(\theta, \phi)$ for our sample, and take the ML estimator $\hat{\theta}^{(\mathrm{ML})}$ to be the point at which $L(\theta, \phi)$ attains its maximum. In MIXed methods, we replace ϕ in the likelihood function by $f(\theta, \hat{U})$ and take $\hat{\theta}^{(\mathrm{MIX})}$ to be the point at which the new likelihood function is maximized. In Proposition 1, based on

the Taylor expansion of the new likelihood function, we show that this substitution the estimator $\hat{\theta}^{(\mathrm{MIX})}$ remains strongly consistent. The proof of the Proposition 1 is similar to that of the traditional ML method. Proposition 3 extends this result to show that $(\hat{\theta}^{(\mathrm{MIX})}, U_n)$ have asymptotically jointly Gaussian distribution. Since the MIXed estimator of ϕ is $\hat{\phi}^{(\mathrm{MIX})} = f(\hat{\theta}^{(\mathrm{MIX})}, \hat{U})$ the transformation theorems are then applied to show that $\hat{\phi}^{(\mathrm{MIX})}$ is strongly consistent (Proposition 2), and that $\hat{\theta}^{(\mathrm{MIX})}$ and $\hat{\phi}^{(\mathrm{MIX})}$ have asymptotically jointly Gaussian distribution.

Let $(\Omega, \mathfrak{F}, I\!P_{(\theta_0,\phi_0)})$ be the underlying probability space. Let X be a real valued random variable with distribution μ that is absolutely continuous with respect to the Lebesgue measure with density $p(x, \theta_0, \phi_0)$. Here, θ and ϕ are the parameters of the distribution that we are trying to estimate. The admissible set of (θ, ϕ) is assumed to be a non-degenerate rectangle $\mathbf{\Theta} \times \mathbf{\Phi} \subset I\!R^2$, and $(\theta_0, \phi_0) \in \mathbf{\Theta} \times \mathbf{\Phi}$ denotes the true value of the distribution parameters. Suppose the function $p(x, \theta, \phi)$ satisfies the following regularity conditions:

(R - 1): For each $(\theta, \phi) \in \mathbf{\Theta} \times \mathbf{\Phi}$, all first, second, and third order partial derivatives of $\log p(x, \theta, \phi)$ exist for almost all x. This condition also implies the continuity of the first and second partial derivatives on $\mathbf{\Theta} \times \mathbf{\Phi}$ for almost all x. Hereafter, by "partial derivatives" we refer only to partial derivatives with respect to θ and ϕ ($\frac{\partial}{\partial\theta}$, $\frac{\partial}{\partial\phi}$, $\frac{\partial^2}{\partial\phi^2}$, $\frac{\partial^2}{\partial\phi\partial\theta}$, etc.).

(R - 2): For every $(\theta, \phi) \in \mathbf{\Theta} \times \mathbf{\Phi}$, the absolute value of all first and second partial derivatives of $\log p(x, \theta, \phi)$ are bounded by some integrable function $M_1(x)$, that is

not a function of θ or ϕ. All third partial derivatives of $\log p(x, \theta, \phi)$ are bounded by an integrable function $M_2(x)$, and $\mathbb{E}_{(\theta,\phi)}\left[M_2(x)\right] < \infty$. The last expectation is with respect to a particular probability measure $p\left(x, \theta, \phi\right) dx$.

(R - 3): At the true value (θ_0, ϕ_0), the information matrix

$$\mathcal{I} = \begin{pmatrix} -\mathbb{E}\left[\left.\frac{\partial^2}{\partial^2\theta} \log p\left(x, \theta, \phi\right)\right|_{(\theta_0,\phi_0)}\right] & -\mathbb{E}\left[\left.\frac{\partial^2}{\partial\theta\,\partial\phi} \log p\left(x, \theta, \phi\right)\right|_{(\theta_0,\phi_0)}\right] \\ -\mathbb{E}\left[\left.\frac{\partial^2}{\partial\theta\,\partial\phi} \log p\left(x, \theta, \phi\right)\right|_{(\theta_0,\phi_0)}\right] & -\mathbb{E}\left[\left.\frac{\partial^2}{\partial^2\phi} \log p\left(x, \theta, \phi\right)\right|_{(\theta_0,\phi_0)}\right] \end{pmatrix} \tag{2.22}$$

is finite, nonsingular, and positive definite. The elements of $\mathcal{I}$ will be denoted by $\mathcal{I}_{11}$, $\mathcal{I}_{12}$, and $\mathcal{I}_{22}$.

These assumptions are standard regularity conditions for the maximum likelihood estimation (e.g., Rao [1973], section 5f). Condition (R - 1) ensures that the partial derivative of the likelihood function, $\frac{\partial}{\partial\theta}\sum \log p(X_i, \theta, \phi)$, has a Taylor expansion with respect to θ. Assumption (R - 2) justifies the commutivity of differentiation with respect to either θ or ϕ and integration with respect to x. (R - 2) also ensures that higher-order terms in the Taylor expansion can be made sufficiently small. Assumption (R - 3) ensures that $\left.\frac{\partial}{\partial\theta} \log p(x, \theta, \phi)\right|_{(\theta_0,\phi_0)}$ has a finite variance.

Let $U(\theta, \phi)$ be a functional defined on the family of possible distribution functions of X, such that

$$U(\theta, \phi) = \int \ldots \int h(x_1, \ldots, x_k)\; p\left(x_1, \theta, \phi\right) dx_1 \ldots p\left(x_k, \theta, \phi\right) dx_k,$$

for some function h. That is, given a random sample $X_1, X_2, \ldots, X_n$ of size n, there is a U-statistic:

$$U_n = \binom{n}{k}^{-1} \sum_{(n,k)} h(X_1, \ldots, X_k), \tag{2.23}$$

that provides an unbiased estimator of $U(\theta_0, \phi_0)$ (See Lee [1990] for properties of U-statistics).

We will denote the range of possible values of $U(\theta, \phi)$ for $(\theta, \phi) \in \mathbf{\Theta} \times \mathbf{\Phi}$ as $\mathbf{U} \subset I\!R$. We assume that U_n is a good estimator of $u_0 = U(\theta_0, \phi_0)$, meaning that, as $n \to \infty$,

(U - 1) $U_n \overset{a.s.}{\to} u_0$

and

(U - 2) $U_n - u_0$ converges in distribution to a Gaussian random variable with mean 0 and variance σ_u/n.

Hereafter, $\overset{a.s.}{\to}$ denotes almost sure convergence.

Finally, we assume that, for every $u \in \mathbf{U}$, the equation $U(\theta, \phi) = u$ can be solved for ϕ, and the solution is a "smooth" function of u and θ. More precisely, we assume that there is a function $f : \mathbf{\Theta} \times \mathbf{U} \longrightarrow \mathbf{\Phi}$ satisfying the following conditions:

(F - 1) $f(\theta, u)$ is continuous and bounded, and so are all of its first, second, and third partial derivatives on $\mathbf{\Theta} \times \mathbf{U}$.

(F - 2) $U(\theta, f(\theta, u)) = u$, for all $\theta \in \mathbf{\Theta}$ and $u \in \mathbf{U}$.

Obviously, we have:

$$\phi_0 = f(\theta_0, u_0). \tag{2.24}$$

We also assume that, for a fixed u and x, $p(x, \theta, f(\theta, u))$, is indeed a function of θ. If it is not, it means that $p(x, \theta, \phi)$ can be reparameterized to depend on a single parameter, and there is no need for a two parameter estimation.

Consider the following estimator $(\hat{\theta}_n, \hat{\phi}_n)$ of the parameters (θ, ϕ), based on a sample $X_1, X_2, \ldots, X_n$ of size n:

$$\begin{aligned} \hat{\theta}_n &= \operatorname{argmax} \sum_{i=1}^{n} \log p(X_i, \theta, f(\theta, U_n)), \\ \hat{\phi}_n &= f(\hat{\theta}_n, U_n), \end{aligned} \tag{2.25}$$

where U_n is a given by (2.23).

Proposition 1: Under the assumptions (R - 1) – (F - 2), the optimization problem (2.25) admits a sequence of solutions $\hat{\theta}_n$ such that $\hat{\theta}_n \longrightarrow \theta_0$ almost surely as $n \longrightarrow \infty$.

Let

$$L(\theta, u | X_1, \ldots, X_n) = \sum_{i=1}^{n} \log p(X_i, \theta, f(\theta, u)). \tag{2.26}$$

The estimator $\hat{\theta}_n$ can found by setting the partial derivative of $L(\theta, U_n | X_1, \ldots, X_n)$ with respect to θ to zero, and then solving the obtained equation for θ. That is, $\hat{\theta}_n$ is a solution of

$$\frac{\partial}{\partial \theta} L(\theta, U_n | X_1, \ldots, X_n) = 0. \tag{2.27}$$

We need to show that such a solution exists for almost all samples $X_1, \ldots, X_n$, $L(\theta, U_n | X_1, \ldots, X_n)$ actually attains a local maximum at $\hat{\theta}_n$ (as opposed to a local minimum), and $\hat{\theta}_n$ almost surely converges to θ_0 as $n \longrightarrow \infty$.

To simplify our formulas, we introduce the following notation:

$$
\begin{aligned}
a(x) &= \frac{\partial}{\partial \theta} \log p(x, \theta, \phi) \Big|_{(\theta_0, \phi_0)} \\
b(x) &= \frac{\partial}{\partial \phi} \log p(x, \theta, \phi) \Big|_{(\theta_0, \phi_0)} \\
c(x) &= \frac{\partial^2}{\partial \theta^2} \log p(x, \theta, \phi) \Big|_{(\theta_0, \phi_0)} \\
d(x) &= \frac{\partial^2}{\partial \theta \partial \phi} \log p(x, \theta, \phi) \Big|_{(\theta_0, \phi_0)} \\
g(x) &= \frac{\partial^2}{\partial \phi^2} \log p(x, \theta, \phi) \Big|_{(\theta_0, \phi_0)}
\end{aligned} \tag{2.28}
$$

We will abbreviate linear combinations of second- or third- order derivatives of $\log p(x, \theta, f(\theta, u))$ multiplied by second or third order derivatives of $f(\theta, u)$ evaluated at a point (x, θ', u'), by $r_j(x, \theta', u')$. These types of expressions arise in the the higher-order terms of the Taylor expansion. Conditions (R-2) and (F - 1) ensure that the $r_j(X, \theta', u')$ have finite expectations for any values of θ' and u'. This is the only property of $r_j(X, \theta', u')$ that we will use.

We denote differentiation of $f(\theta, u)$ by indexes. For example, we write $f_{\theta u}(\theta, u)$ instead of $\frac{\partial^2}{\partial \theta \partial u} f(\theta, u)$, and $f_\theta(\theta_0, u_0)$ for $\frac{\partial}{\partial \theta} f(\theta, u)\big|_{(\theta_0, u_0)}$.

Using (R - 1) and (F - 1), we can expand $\frac{\partial}{\partial\theta}\left[\log p(x,\theta,f(\theta,u))\right]$ in the neighborhood of (θ_0, u_0). We have:

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\log p(x,\theta,f(\theta,u))\right] &= a(x) + b(x)f_\theta(\theta_0,u_0)\\ &+ \left[c(x) + 2d(x)f_\theta(\theta_0,u_0) + g(x)\left(f_\theta(\theta_0,u_0)\right)^2 + b(x)f_{\theta\theta}(\theta_0,u_0)\right](\theta-\theta_0)\\ &+ \left[d(x)f_u(\theta_0,u_0) + g(x)f_u(\theta_0,u_0)f_\theta(\theta_0,u_0) + b(x)f_{\theta u}(\theta_0,u_0)\right](u-u_0)\\ &+ r_1(x,\theta',u')(\theta-\theta_0)^2 + r_2(x,\theta',u')(\theta-\theta_0)(u-u_0) + r_3(x,\theta',u')(u-u_0)^2, \quad (2.29)\end{aligned}$$

where $\theta' = \theta_0 + \lambda(\theta-\theta_0)$, and $u' = u + \lambda(u-u_0)$, for some $0 < \lambda < 1$.

Now introduce the following random variables:

$$\begin{aligned}A_n &= \frac{1}{n}\sum_{i=1}^n a(X_i) + b(X_i)f_\theta(\theta_0,u_0)\\ B_n &= \frac{1}{n}\sum_{i=1}^n \left[c(X_i) + 2d(X_i)f_\theta(\theta_0,u_0) + g(X_i)\left(f_\theta(\theta_0,u_0)\right)^2 + b(X_i)f_{\theta\theta}(\theta_0,u_0)\right]\\ C_n &= \frac{1}{n}\sum_{i=1}^n \left[d(X_i)f_u(\theta_0,u_0) + g(X_i)f_u(\theta_0,u_0)f_\theta(\theta_0,u_0) + b(X_i)f_{\theta u}(\theta_0,u_0)\right]\\ R_{j,n} &= \frac{1}{n}\sum_{i=1}^n r_j(X_i,\theta',u'), \qquad j = 1,2,3. \quad (2.30)\end{aligned}$$

Note that all of the random variables above are averages of i.i.d. random variables, and therefore we can apply the strong law of large numbers to determine how these random variables behave as $n \longrightarrow \infty$.

First, note that

$$\mathbb{E}\left[a(X)\right] = \int \frac{\frac{\partial}{\partial\theta}p(x,\theta,\phi)\big|_{(\theta_0,\phi_0)}}{p(x,\theta_0,\phi_0)}p(x,\theta_0,\phi_0)dx = \left[\frac{\partial}{\partial\theta}\int p(x,\theta,\phi)dx\right]\Bigg|_{(\theta_0,\phi_0)} = 0 \quad (2.31)$$

and, similarly,

$$\mathbb{E}\left[b(X)\right] = 0 \tag{2.32}$$

Also, from (2.22),

$$\mathbb{E}\left[c(X)\right] = -\mathcal{I}_{11}, \qquad \mathbb{E}\left[d(X)\right] = -\mathcal{I}_{12}, \qquad \text{and} \qquad \mathbb{E}\left[g(X)\right] = -\mathcal{I}_{22}.$$

Let $\mathbb{E}\left[r_j(X)\right] = \rho_j$, $j = 1, 2, 3$. From conditions (R - 2) and (F - 1), $r_j(X)$ has finite expectation, meaning that there is a constant $K < \infty$, such that

$$\mathbb{E}\left[r_j(X)\right] < K, \qquad j = 1, 2, 3.$$

Applying the strong law of large numbers, we conclude that, as $n \longrightarrow \infty$

$$A_n \xrightarrow{a.s.} 0 \tag{2.33}$$

$$B_n \xrightarrow{a.s.} -\mathcal{I}_{11} - 2\mathcal{I}_{12} f_\theta(\theta_0, u_0) - \mathcal{I}_{22}\left(f_\theta(\theta_0, u_0)\right)^2 = -\beta, \tag{2.34}$$

$$C_n \xrightarrow{a.s.} -\mathcal{I}_{12} f_u(\theta_0, u_0) - \mathcal{I}_{22} f_u(\theta_0, u_0) f_\theta(\theta_0, u_0) = \gamma, \tag{2.35}$$

$$\text{and} \qquad R_{j,n} \xrightarrow{a.s.} \rho_j, \qquad |\rho_j| < K < \infty, \qquad j = 1, 2, 3 \tag{2.36}$$

Note that, since $\mathcal{I}$ is positive definite and (2.34) can be rewritten as a quadratic form, we have

$$0 < \beta < \infty. \tag{2.37}$$

According to conditions (R - 2) and (F - 1), we also find that $|\gamma| < \infty$.

Using the Taylor expansion in (2.29) and the definition of the random variables in (2.30), we can write

$$\frac{\partial}{\partial\theta}L(\theta, U_n|X_1,\ldots,X_n) = A_n + B_n(\theta-\theta_0) + C_n(U_n-u_0)$$
$$+ R_{1,n}(\theta-\theta_0)^2 + R_{2,n}(\theta-\theta_0)(U_n-u_0) + R_{3,n}(U_n-u_0)^2. \quad (2.38)$$

Let $\Omega' \subset \Omega$ be the set of all $\omega \in \Omega$ for which the limits (2.33 – 2.36) and $U_n(\omega) \longrightarrow u_0$ take place and the conditions (R - 1) and (F-1) are satisfied. Note that $I\!P\{\Omega'\} = 1$. Next, fix $\varepsilon > 0$ such that $\varepsilon < \beta/(3K)$ and the points $\theta_1 = \theta_0 - \varepsilon$ and $\theta_2 = \theta_0 + \varepsilon$ are in $\mathbf{\Theta}$, and fix an $\omega \in \Omega'$. Then, there is a $N_\varepsilon(\omega)$ such that, for all $n > N_\varepsilon(\omega)$, we have

$$|U_n(\omega) - u_0| < \max\left(\frac{\beta\varepsilon}{8|\gamma|}, \frac{\varepsilon}{8}\right),$$
$$|A_n| < \frac{\beta\varepsilon}{4},$$
$$|B_n + \beta| < \frac{\beta}{4}.$$

Using equation 2.38, we find

$$\left|\left(\frac{\partial}{\partial\theta}L(\theta,n,\omega)\right)\Big|_{\theta-\varepsilon} - \beta\varepsilon\right| = |A_n(\omega) - (B_n(\omega)+\beta)\varepsilon + C_n(\omega)(U_n(\omega)-u_0)$$
$$+ R_{1,n}(\omega)\varepsilon^2 - R_{2,n}(\omega)\varepsilon(U_n(\omega)-u_0) + R_{3,n}(\omega)(U_n-u_0)^2|$$
$$\leq |A_n(\omega)| + |B_n(\omega)+\beta|\varepsilon + \gamma|U_n(\omega)-u_0|$$
$$+ K(\varepsilon^2 + \varepsilon|U_n(\omega)-u_0| + |U_n(\omega)-u_0|^2)$$
$$< \frac{3\beta\varepsilon}{4}, \quad (2.39)$$

and, similarly,

$$\left| \left(\frac{\partial}{\partial\theta} L(\theta, n, \omega) \right) \Big|_{\theta+\varepsilon} + \beta\varepsilon \right| = |A_n(\omega) + (B_n(\omega) + \beta)\varepsilon + C_n(\omega)(U_n(\omega) - u_0)$$
$$+ R_{1,n}(\omega)\varepsilon^2 - R_{2,n}(\omega)\varepsilon(U_n(\omega) - u_0) + R_{3,n}(\omega)(U_n - u_0)^2|$$
$$< \frac{3\beta\varepsilon}{4}. \quad (2.40)$$

Here, we wrote $L(\theta, U_n(\omega)|X_1(\omega), \ldots, X_n(\omega))$ as $L(\theta, n, \omega)$, since ω determines all $X_1(\omega), \ldots, X_n(\omega)$, and $U_n(\omega)$. Inequalities (2.39) and (2.40) imply that

$$\frac{\partial}{\partial\theta} L(\theta, n, \omega)\Big|_{\theta-\varepsilon} > \frac{\beta\varepsilon}{4} > 0, \qquad \text{and} \qquad \frac{\partial}{\partial\theta} L(\theta, n, \omega)\Big|_{\theta+\varepsilon} < -\frac{\beta\varepsilon}{4} < 0.$$

These inequalities, together with the assumed continuity of $\frac{\partial}{\partial\theta} L(\theta, n, \omega)$, imply that the equation $\frac{\partial}{\partial\theta} L(\theta, n, \omega) = 0$ has at least one root between $\theta_0 - \varepsilon$ and $\theta_0 + \varepsilon$, and, since $\beta > 0$, the function $L(\theta, n, \omega)$ achieves its maximum at one of the roots. For example,

$$\hat{\theta}_{n\varepsilon}(\omega) = \inf \left\{ \theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon] : \frac{\partial}{\partial\theta} L(\theta, n, \omega) = 0 \right\} \quad (2.41)$$

will be one of the solutions of the likelihood equation (2.27). It is easy to see that $\hat{\theta}_{n\varepsilon}$ is a random variable, that is, it is measurable. Indeed, for all $t \geq \theta_0 - \varepsilon$, we have (by the continuity of $\dfrac{\partial}{\partial\theta} L(\theta, n, \omega)$, it is measurable with respect to $\mathfrak{F} \otimes \mathfrak{B}$, where $\mathfrak{B}$ is a Borel σ-algebra):

$$\left\{ \hat{\theta}_{n\varepsilon} > t \right\} = \left\{ \inf_{\substack{\theta_0 - \varepsilon \leq \theta \leq t \\ \theta \text{ rational}}} \frac{\partial}{\partial\theta} L(\theta, n, \omega) > 0 \right\} \bigcup \left\{ \sup_{\substack{\theta_0 - \varepsilon \leq \theta \leq t \\ \theta \text{ rational}}} \frac{\partial}{\partial\theta} L(\theta, n, \omega) < 0 \right\}$$

Thus, $\left\{ \hat{\theta}_{n\varepsilon} > t \right\}$ is a measurable set.

The estimator $\hat{\theta}_{n\varepsilon}(\omega)$ depends on the choice of ε, and we need to construct an estimator $\hat{\theta}_n(\omega)$ independent of ε. To this end, we replace ε by a sequence $\varepsilon_m = 1/m$, $m = 1, 2, \ldots$. Then, without loss of generality, we may assume that

$$N_1(\omega) \leq N_{1/2}(\omega) \leq N_{1/3}(\omega) \leq \ldots$$

and for $N_{1/m}(\omega) \leq n < N_{1/(m+1)}(\omega)$, we set

$$\hat{\theta}_n(\omega) = \hat{\theta}_{n,1/m}(\omega).$$

For $n < N_1(\omega)$ or for $\omega \notin \Omega'$, we can set $\hat{\theta}_n(\omega) = 0$.

Under this construction, $\hat{\theta}_n$ is the desired solution to the equation (2.27), and converges to θ_0 almost surely as $n \longrightarrow \infty$. This completes the proof of Proposition 1.

Our construction of $\hat{\theta}_n$ ensures that it is a local maximum of $L(\theta, U_n | X_1, \ldots X_n)$, but we did not consider whether it is actually a global maximum. As in the case of maximum likelihood, additional regularity conditions have to be imposed on the density function $p(x, \theta, \phi)$ to show that $\hat{\theta}_n$ is a global maximum. Doing so is beyond the scope of this section, but the same arguments as for the MLE estimators, e.g. [Prakasa Rao, 1987, section 3.6], can be readily applied to our case, so the global properties for MIXed estimators should be similar to that of the MLE estimators.

Proposition 2: Under assumptions (R - 1) – (F - 2), the estimator $\hat{\phi}_n$ defined in (2.25) converges to ϕ_0 almost surely as $n \longrightarrow \infty$.

This directly follows from Proposition 1, the almost sure convergence of U_n, and the continuity and boundedness of the derivatives of f. See Rao [1973], section 6a.2 for the appropriate theorems.

Propositions 1 and 2 imply strong consistency of MIXed-type estimators. We now consider their asymptotic distribution and show that it is jointly Gaussian.

Proposition 3: Under assumptions (R - 1) – (F - 2), the estimator $\hat{\theta}_n$, defined as a solution to (2.27), and U_n have an asymptotically jointly Gaussian distribution with mean $(\theta_0, u_0)^T$ and covariance matrix $\frac{1}{n}\mathcal{B}$, where

$$\mathcal{B} = \frac{1}{\beta}\begin{pmatrix} 1 + \frac{2\sigma_{au}\gamma + \gamma^2\sigma_u}{\beta} & \sigma_{au} + \gamma\sigma_u \\ \sigma_{au} + \gamma\sigma_u & \beta\sigma_u \end{pmatrix}, \tag{2.42}$$

σ_u is the asymptotic variance of U_n, β is defined in (2.34), γ is defined in (2.35), and σ_{au} is given by (2.43).

A_n, defined in (2.30), is an average of i.i.d. random variables $a(X)+b(X)f_\theta(\theta_0, u_0)$. By (2.31) and (2.32),

$$\mathbb{E}\left[a(X) + b(X)f_\theta(\theta_0, u_0)\right] = 0$$

To compute the variance of $a(X) + b(X)f_\theta(\theta_0, u_0)$, note that

$$\begin{aligned}
\mathbb{E}\left[(a(X))^2\right] &= \int \frac{\left(\frac{\partial}{\partial\theta}p(x,\theta,\phi)\big|_{(\theta_0,\phi_0)}\right)^2}{p^2(x,\theta_0,\phi_0)} p(x,\theta_0,\phi_0)dx \\
&= \underbrace{\int \frac{\frac{\partial^2}{\partial^2\theta}p(x,\theta,\phi)\Big|_{(\theta_0,\phi_0)}}{p(x,\theta_0,\phi_0)} p(x,\theta_0,\phi_0)dx}_{=0} - \int \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}p(x,\theta,\phi)}{p(x,\theta_0,\phi_0)}\right)\Bigg|_{(\theta_0,\phi_0)} p(x,\theta_0,\phi_0)dx \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(X,\theta,\phi)\Big|_{(\theta_0,\phi_0)}\right] = \mathcal{I}_{11}
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[(b(X))^2\right] = \mathcal{I}_{22} \qquad \text{and} \qquad \mathbb{E}\left[a(X)b(X)\right] = \mathcal{I}_{12}.$$

This implies that

$$\mathrm{Var}(a(X) + b(X)f_\theta(\theta_0, u_0)) = \left[\mathcal{I}_{11} + (f_\theta(\theta_0, u_0))^2\,\mathcal{I}_{22} + 2f_\theta(\theta_0, u_0)\mathcal{I}_{12}\right] = \beta < \infty.$$

Thus, by the central limit theorem, A_n is asymptotically a Gaussian random variable with mean zero and variance β/n. A_n, as defined in (2.30), can actually be viewed as a U-statistic with first order kernel

$$h_a(x) = a(x) + b(x)f_\theta(\theta_0, u_0).$$

Therefore, according to the theory of U-statistics, the random variables A_n and $U_n - u_0$ have a jointly Gaussian distribution with covariance

$$\sigma_{au} = \frac{k}{n}\,\mathrm{Cov}\left(h_a(X_1), h(X_1, \ldots X_k)\right), \tag{2.43}$$

where k is the degree of U_n in (2.23) [Lee, 1990, p. 76]. Thus, the asymptotic distribution of the vector $\sqrt{n}(A_n, U_n - u_0)^T$ is a bivariate Gaussian with mean vector $(0, 0)^T$ and covariance matrix

$$\mathcal{A} = \begin{pmatrix} \beta & \sigma_{au} \\ \sigma_{au} & \sigma_u \end{pmatrix}.$$

The vector $(\hat{\theta}_n, U_n)^T$ satisfies the following system of equations, written in matrix form (see (2.38)):

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} A_n \\ U_n - u_0 \end{pmatrix} + \begin{pmatrix} B_n + R_1(\hat{\theta}_n - \theta_0) & C_n + R_2(\hat{\theta}_n - \theta_0) + R_3(U_n - u_0) \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ U_n - u_0 \end{pmatrix}. \tag{2.44}$$

By equations (2.34 – 2.36) and the strong consistency of the estimators,

$$-\mathcal{J}_n = \begin{pmatrix} B_n + R_1(\hat{\theta} - \theta_0) & C_n + R_2(\hat{\theta} - \theta_0) + R_3(U_n - u_0) \\ 0 & -1 \end{pmatrix} \xrightarrow{a.s.} \begin{pmatrix} -\beta & \gamma \\ 0 & -1 \end{pmatrix} = -\mathcal{J} \tag{2.45}$$

as $n \longrightarrow \infty$. Notice that matrix $\mathcal{J}$ is nonsingular, so matrix $\mathcal{J}_n$ is almost surely nonsingular for large n. Employing (2.44), this yields

$$\begin{pmatrix} \hat{\theta} - \theta_0 \\ U_n - u_0 \end{pmatrix} = \mathcal{J}_n^{-1} \begin{pmatrix} A_n \\ U_n - u_0 \end{pmatrix}, \tag{2.46}$$

which can be rewritten as

$$\sqrt{n}\mathcal{J} \begin{pmatrix} \hat{\theta} - \theta_0 \\ U_n - u_0 \end{pmatrix} - \sqrt{n} \begin{pmatrix} A_n \\ U_n - u_0 \end{pmatrix} = \sqrt{n}(\mathcal{J}\mathcal{J}_n^{-1} - 1) \begin{pmatrix} A_n \\ U_n - u_0 \end{pmatrix}. \tag{2.47}$$

From (2.45), we see that $\mathcal{J}\mathcal{J}_n^{-1} - 1 \xrightarrow{a.s.} 0$ as $n \longrightarrow \infty$. Equation (2.47), then, implies that vectors $\sqrt{n}\mathcal{J}(\hat{\theta}_n - \theta_0, U_n - u_0)^T$ and $\sqrt{n}(A_n, U_n - u_0)^T$ have the same limiting distribution (see e.g. Rao [1973], p.122). Since the asymptotic distribution of $(A_n, U_n - u_0)^T$ is a bivariate Gaussian with mean vector $(0, 0)^T$ and covariance matrix $\frac{1}{n}\mathcal{A}$, the asymptotic distribution of $(\hat{\theta}_n, U_n)^T$ is also a bivariate Gaussian with mean vector $(\theta_0, u_0)^T$ and covariance matrix

$$\frac{1}{n}\mathcal{B} = \frac{1}{n}\mathcal{J}^{-1}\mathcal{A}(\mathcal{J}^{-1})^T.$$

Noting that $\mathcal{J}^{-1} = \dfrac{1}{\beta^2}\begin{pmatrix} 1 & \gamma \\ 0 & \beta \end{pmatrix}$, we see that $\mathcal{B}$ is exactly the same as in (2.42). This completes the proof of proposition 3.

Proposition 4: Under assumptions (R - 1) – (F - 2), the estimators $\hat{\theta}_n$ and $\hat{\phi}_n$ have asymptotically jointly Gaussian distribution, with mean $(\theta_0, \phi_0)^T$ and covariance matrix $\frac{1}{n}\mathcal{C}$, where

$$\mathcal{C} = \frac{1}{\beta}\begin{pmatrix} 1 & 0 \\ f_\theta(\theta_0, u_0) & f_u(\theta_0, u_0) \end{pmatrix}\begin{pmatrix} 1 + \frac{2\sigma_{au}\gamma + \gamma^2\sigma_u}{\beta} & \sigma_{au} + \gamma\sigma_u \\ \sigma_{au} + \gamma\sigma_u & \beta\sigma_u \end{pmatrix}\begin{pmatrix} 1 & f_\theta(\theta_0, u_0) \\ 0 & f_u(\theta_0, u_0) \end{pmatrix}$$

The statement follows directly from Proposition 3 and the continuity and boundedness of the derivatives of f (see Rao [1973], p. 388).

The results established above, can be easily extended to higher dimensions. Note from the form of asymptotic covariance matrix $\mathcal{C}$ that $\det \mathcal{C} \geq \frac{1}{\beta} = -\det(\mathcal{I}^{-1})$, meaning that the overall asymptotic variance of the MIXed estimators is greater than that of the MLE, as expected. Also, we were able to prove the asymptotic properties of the MIXed estimator under conditions, where both MLE and U-statistic estimators exist. This implies that MIXed methods underperform MLE methods for large samples which we saw in section 2.2.

To conclude, we have established that MIXed parameter estimates possess desirable asymptotic properties, under certain regularity conditions: they are strongly consistent, and have an asymptotically jointly Gaussian distribution. The regularity conditions correspond to the regularity conditions for the MLE method. Smith [1985] shows that the GEV density function satisfies this conditions for $k \leq 1/2$.

2.6 Summary of Chapter 2

New methods (MIX1 and MIX2) of parameter estimation for the GEV distribution are introduced based on a combination of the MLE and LMOM methods. These procedures can be viewed as LMOM-constrained MLE methods. MIX1 and MIX2 estimators are strongly consistent and have asymptotically jointly Gaussian distributions (section 2.5). The performance of the MIX1 and MIX2 methods has been studied using simulations for small samples. MIX1 and MIX2 do not produce absurd estimates of k, and the RMSE of these estimators are smaller than that of LMOM and MLE for $k < -0.2$ for small sample sizes ($n \leq 50$). The new methods are based on the maximum likelihood principle and can easily incorporate additional information.

The MIX method estimators are extended to incorporate PDS information. The MLE2 method which employs both AMS and PDS records, is introduced and its performance is studied using simulations. We show that, for $k \leq -0.1$, the MLE2-MIX1 method results in smaller values of RMSE for quantiles.

These techniques were applied to the sample of 104 basins in the central Appalachians, and no dependence of the shape parameter on basin descriptors was detected. We conclude that the intrinsic variability of the parameter estimates correspond to the variability of the estimation techniques and observe that basins with a higher percentage of urban area have larger values of the scale and location parameters. The estimated values of the shape parameter k were found to be more negative than previously considered physically realistic, implying that the distribution has thick

tails. The estimates of k for central Appalachian watersheds are shown to differ from those of southern Appalachian watersheds, and that difference is compared to the contrasting properties of extreme floods. The negative values found for k might be explained by a stochastic model with tails thinner than those of the GEV distribution with very negative k.

The methods described in this chapter are implemented in the EVANESCE ("Extreme Value ANalysis Employing Statistical Copula Estimation") package for S-Plus, which is described in the next chapter.

Chapter 3

Modeling Joint Distributions with Copulas: EVANESCE

Multivariate statistical models have attracted attention recently, due to the increasing need for stochastic models that address sophisticated relationships between multiple related datasets and the increases in computing power that have made complicated computational problems more tractable [Joe, 1997]. A number of important theoretical constructions and techniques for modeling the interdependencies of random variables have recently been established, and we have written a library for bivariate, trivariate, and extreme value analysis for the statistical package S-Plus (by Insightful) that simplifies the application of these advanced concepts for data analysis. This library is named EVANESCE (Extreme Value ANalysis Employing Statistical Copula Estimation). It provides a comprehensive suite of ready-to-use functions for data analysis using copulas.

In this chapter, we describe the theory and methods for modeling bivariate distributions with copulas, focusing on the methods that allow direct, practical application. Section 3.8 summarizes the basic capabilities of EVANESCE, and Chapters 4 and 5 illustrate the use of copulas for data analysis.

A two dimensional copula is essentially a distribution function of two random variables with marginals uniform on $(0, 1)$. The theory indicates that, by mapping two random variables of interest to uniform $(0, 1)$ random variables, we can use the properties of copulas to infer information about their mutual dependence. A full introduction to this concept is presented below.

3.1 A Brief Introduction to Copulas

Suppose we are interested in modeling the stochastic behavior of two random variables X and Y based on a set of n independent observations of the couple (X, Y), say $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Furthermore, suppose that we have estimated their respective marginal distributions, $G_1(x)$ and $G_2(y)$, using standard statistical techniques. We are then interested to know whether X and Y are independent, how we test whether they are independent, and how we describe their dependence if they are not independent.

One of the best ways to describe the dependence between two random variables is to estimate their joint distribution function

$$F(x, y) = \mathbb{P}\{X \leq x, Y \leq y\},$$

since any other measure of the dependence is ultimately some function of this joint distribution. Tests of hypotheses concerning the independence of X and Y reduce to tests of whether $F(x, y)$ is a product of the respective marginal distributions.

The random variables $G_1(X)$ and $G_2(Y)$ are uniformly distributed between 0 and 1 (see Nelsen [1999], p. 11), and their joint distribution is

$$C(u, v) = I\!P\{G_1(X) \leq u, G_2(Y) \leq v\} = F(G_1^{-1}(u), G_2^{-1}(v)), \tag{3.1}$$

where G_1^{-1} and G_2^{-1} are the quantile functions of X and Y respectively ($G_1^{-1}(q) = \inf\{x \in I\!R : G(x) > q\}$). If we can estimate the function C, we will then be able to compute F by setting

$$F(x, y) = C(G_1(x), G_2(y)). \tag{3.2}$$

Under this construction, C is a distribution function of two random variables with uniform $(0, 1)$ margins. Such functions are called *two-dimensional copulas*, and have been extensively studied in the past few years (Nelsen [1999] and Joe [1997] are only two of several recently-published textbooks that provide a detailed review of the subject).

It is not difficult to prove that C is a two-dimensional copula if and only if it is a function $C(u, v) : \mathbf{I}^2 \mapsto \mathbf{I}$ ($\mathbf{I} = [0, 1]$) that satisfies the following two conditions:

- For every u and v in $\mathbf{I}$:

  $$C(u, 0) = C(0, v) = 0, \qquad C(u, 1) = u, \qquad \text{and} \qquad C(1, v) = v;$$

- For every $u_1 \leq u_2$, and $v_1 \leq v_2$, and $u_1, u_2, v_1, v_2 \in \mathbf{I}$, the following inequality holds:

$$C(u_1, v_1) - C(u_2, v_1) - C(u_1, v_2) + C(u_2, v_2) \geq 0$$

If C is considered to be a distribution function of two random variables U and V, the first condition ensures that U and V have uniform marginal distributions. The second condition, often referred to as the rectangular inequality, simply requires that C is a valid distribution function, i.e. $\mathbb{P}\{u_1 \leq U \leq u_2, v_1 \leq V \leq v_2\} \geq 0$.

Of course, the notion of copula extends into higher dimensions: n-dimensional copulas are joint distribution functions of n random variable with uniform $(0, 1)$ marginals. For the next few sections, we will focus on two-dimensional copulas and their application and discuss the extension of the methods to higher dimensions in section 3.7.

Formula (3.1) shows that, if X and Y are continuous random variables, the function C satisfying (3.1) is unique. If X and Y are not continuous random variables, C is uniquely determined on $\operatorname{Range} G_1 \times \operatorname{Range} G_2$ (this result is known as Sklar's theorem [Sklar, 1959], see also Nelsen [1999], p. 15). In order to estimate F, we can transform the observations of X and Y, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, into observations of $U = G_1(X)$ and $V = G_2(Y)$, $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)$, where $u_i = G_1(x_i)$ and $v_i = G_2(y_i)$, $i = 1, 2, \ldots, n$, and C may then be estimated as the joint distribution of U and V. Then we can use formula (3.2) to produce an estimate of F. The function C is a copula, and there have been a few non-parametric methods de-

veloped for its estimation [Deheuvels, 1978; Genest and Rivest, 1993; Genest *et al.*, 1995; Capéraà *et al.*, 1997a]. One can also assume that the copula has a particular parametric form and estimate its parameters using, for example, the method of maximum likelihood. The various non-parametric techniques that are implemented in EVANESCE are summarized in section 3.5, and section 3.2 lists several widely used families of parametric copulas, all of which are implemented in EVANESCE.

Other issues that might be of interest in practical applications of bivariate statistics include measuring the association between two random variables and generating random pairs from a particular copula. In section 3.3, we discuss different measures between the association of the random variables, and explain why the most common measure of association, the coefficient of correlation, might not be the best. Alternative measures that are implemented in EVANESCE, namely Kendall's tau and Spearman's rho, are discussed in that section, along with the concept of tail dependence. Algorithms suggested by Genest and Mackay [1986a] and Capéraà *et al.* [2000] for the generation of random pairs from certain parametric copulas that are implemented in EVANESCE are discussed in Section 3.4.

Alternatives to copula-based models include multivariate threshold exceedances and Markov chain models [Tawn, 1988, 1990; Joe *et al.*, 1992; Smith, 1994; Smith *et al.*, 1997]. These models provide estimates of the tails of joint distribution functions of two random variables on the basis of certain theorems in multivariate extreme value theory, but do not always allow direct simulation of the random variables. In addition,

they are not as generally applicable as copula-based models [Joe, 1997]. For these reasons, bivariate threshold exceedances models are not included in EVANESCE, and are not discussed here.

3.2 Parametric Copula Classes and Families

In this section, we describe the parametric copula families implemented in EVANESCE. EVANESCE users can use these families to construct joint cumulative and probability density functions, generate random variables, compute Kendall's tau, Spearman's rho, and the tail index parameter, or use the maximum likelihood method to estimate parameters of any of these copulas.

- One of the most frequently used copulas (especially for financial modeling) is the copula of a bivariate Gaussian distribution with correlation δ. It is defined by

$$\begin{aligned} C(u,v) &= \int_{-\infty}^{\Phi^{-1}(u)} dx \int_{-\infty}^{\Phi^{-1}(v)} dy \, \frac{1}{2\pi\sqrt{1-\delta^2}} \exp\left\{-\frac{x^2-2\delta xy+y^2}{2(1-\delta^2)}\right\} \\ &= \Phi_\delta\left(\Phi^{-1}(u), \Phi^{-1}(v)\right), \end{aligned} \tag{3.3}$$

where Φ^{-1} is the quantile function of the standard univariate Gaussian distribution, and Φ_δ is the joint cumulative distribution function of a standard bivariate Gaussian distribution with correlation coefficient δ ($0 \leq \delta \leq 1$). Since this copula is a very familiar object to most researchers (especially when used with Gaussian margins), it has been incorporated in a number of applications simply because it was the only tool available for quite some time [Embrechts *et al.*, 2000b]. In fact, J. P. Morgan's

RiskMetrics [1995] has been using this copula for portfolio risk management by Monte Carlo simulations long before it was related that one were dealing with copulas.

• An important class of copulas is the Extreme Value (EV) class. A copula is said to be an EV copula if for all $t > 0$ the scaling property

$$C(u^t, v^t) = (C(u, v))^t$$

holds for all $(u, v) \in \mathbf{I}^2$. Such copulas are max-stable, meaning that, if $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. random pairs from an EV copula C and $M_n = \max\{X_1, X_2, \ldots, X_n\}$, $N_n = \max\{Y_1, Y_2, \ldots, Y_n\}$, a copula associated with the random pair (M_n, N_n) is also C. It can be shown [Joe, 1997, p. 175] that EV copulas can be represented in the form:

$$C(u, v) = \exp\left\{\log(uv)\, A\left(\frac{\log(u)}{\log(uv)}\right)\right\},$$

where $A : [0, 1] \to [\frac{1}{2}, 1]$ is a convex function such that $\max(t, 1 - t) \leq A(t) \leq 1$ for all $t \in [0, 1]$. The function $A(t)$ is called the *dependence function*. As in the case of univariate random variables, it can be shown that the limiting copula for the sequence $\{(a_n + b_n M_n, c_n + d_n N_n)\}$ is an EV copula, if the sequence $\{(a_n + b_n M_n, c_n + d_n N_n)\}$ converges weakly in distribution for some sequences of numbers a_n, b_n, c_n, and d_n, under certain regularity conditions [Galambos, 1987].

• The Gumbel copula [Gumbel, 1960], probably the next-best-known copula to the Gaussian copula [Nelsen, 1999], has the following form:

$$C(u, v) = \exp\left\{-\left[(-\log u)^\delta + (-\log v)^\delta\right]^{1/\delta}\right\}, \tag{3.4}$$

where $\delta \geq 1$ is a parameter. It has a rather simple form, and it is easy to generate random samples from this copula [Johnson, 1987, p. 197]. It is an EV copula, and its dependence function is

$$A(t) = (t^\delta + (1-t)^\delta)^{1/\delta}.$$

Dependence functions for four other EV copula families that are implemented in EVANESCE are listed in Table 3.2. Several of the copulas in this chapter (and in the EVANESCE library) are taken from Joe's [1997] textbook, and we retain their abbreviated names from the textbook, e.g., BB5, BB2, etc.

- As can be seen from equation (3.4), the Gumbel copula can be written in the following form:

$$C(u,v) = \phi^{-1}[\phi(u) + \phi(v)], \tag{3.5}$$

for $\phi(t) = (-\log t)^\delta$. Copulas that can be written in this form are called Archimedean copulas, and the function ϕ is called the *Archimedean generator* [Nelsen, 1999, chapter 4]. In order for the function C in (3.5) to be a valid copula, the function $\phi(t) : \mathbf{I} \mapsto I\!R^+$ has to be continuous, strictly decreasing, convex, and satisfy $\phi(1) = 0$. $\phi(0)$ is defined as $\lim_{t\to 0^+} \phi(t)$, and $\phi^{-1}(z) = 0$ for all $z > \phi(0)$, if $\phi(0) < \infty$ [Nelsen, 1999, p. 91].

We have implemented seven Archimedean copulas in EVANESCE. Five of these are two-parameter families that correspond to the BB1 – BB3 and BB6 – BB7 families from Joe [1997]. The Archimedean copulas implemented and their generators are listed in Table 3.6.

Name	Description
Galambos [1975]	$C(u,v) = uv \exp\left\{\left[(-\log u)^{-\delta} + (-\log v)^{-\delta}\right]^{-1/\delta}\right\}$, $0 \le \delta < \infty$. The dependence function for this copula is $A(t) = 1 - (t^{-\delta} + (1-t)^{-\delta})^{-1/\delta}$
Hüsler and Reiss [1989]	$C(u,v) = \exp\left\{-\tilde{u}\ \Phi\left[\frac{1}{\delta} + \frac{1}{2}\delta \log\left(\frac{\tilde{u}}{\tilde{v}}\right)\right] - \tilde{v}\ \Phi\left[\frac{1}{\delta} + \frac{1}{2}\delta \log\left(\frac{\tilde{v}}{\tilde{u}}\right)\right]\right\}$, where $0 \le \delta < \infty$, $\tilde{u} = -\log u$, $\tilde{v} = -\log v$, and Φ is the cdf of a standard Gaussian. The dependence function for this copula is $A(t) = t\ \Phi\left[\delta^{-1} + \frac{1}{2}\delta \log\left(\frac{t}{1-t}\right)\right] + (1-t)\ \Phi\left[\delta^{-1} - \frac{1}{2}\delta \log\left(\frac{t}{1-t}\right)\right]$.
Tawn [1988]	This is an asymmetric extension of the Gumbel copula. The dependence function for this copula is $A(t) = 1 - \beta + (\beta - \alpha) + \{\alpha^r t^r + \beta^r (1-t)^r\}^{1/r}$, $0 \le \alpha,\ \beta \le 1,\ \ 1 \le r < \infty$.
BB5 copula	$C(u,v) = \exp\left\{-\left[\tilde{u}^\theta + \tilde{v}^\theta - \left(\tilde{u}^{-\theta\delta} + \tilde{v}^{-\theta\delta}\right)^{-1/\delta}\right]^{1/\theta}\right\}$, $\delta > 0$, $\theta \ge 1$, $\tilde{u} = -\log v$, $\tilde{u} = -\log v$. This is a two-parameter extension of the Gumbel copula. The dependence function for this copula is $A(t) = \left[t^\theta + (1-t)^\theta - \left(t^{-\delta\theta} + (1-t)^{-\delta\theta}\right)^{-1/\delta}\right]^{1/\theta}$.

Table 3.1: Several parametric families of Extreme Value copulas implemented in the EVANESCE library.

Abbreviation	Generator, $\phi(t)$	$C(u,v)$	Parameters
Frank [1979]	$-\log \dfrac{e^{-\delta t}-1}{e^{-\delta}-1}$	$C(u,v) = -\delta^{-1}\log\left([\eta-(1-e^{-\delta u})(1-e^{-\delta v})]/\eta\right),$	$\delta > 0$
Kimeldorf and Sampson [1975]	$t^{-\delta}-1$	$C(u,v) = (u^{-\delta}+v^{-\delta}-1)^{-1/\delta}$	$\delta \geq 0$
Joe [1993]	$-\log\left(1-(1-t)^{\theta}\right)$	$C(u,v) = 1-\left((1-u)^{\delta}+(1-v)^{\delta}-(1-u)^{\delta}(1-v)^{\delta}\right)^{1/\delta}$	$\delta \geq 1$
BB1	$\left(t^{-\theta}-1\right)^{\delta}$	$C(u,v) = \left\{1+\left[(u^{-\theta}-1)^{\delta}+(v^{-\theta}-1)^{\delta}\right]^{1/\delta}\right\}^{-1/\theta}$	$\theta > 0, \delta \geq 1$
BB2	$e^{\delta(t^{-\theta}-1)}-1$	$C(u,v) = \left[1+\delta^{-1}\log\left(e^{\delta(u^{-\theta})}+e^{\delta(v^{-\theta})}-1\right)\right]^{1/\theta}$	$\theta > 0, \delta > 0$
BB3	$\exp\left\{\delta\ (-\log t)^{\theta}\right\}-1$	$C(u,v) = \exp\left\{-\left[\delta^{-1}\log\left(e^{\delta\tilde{u}^{\theta}}+e^{\delta\tilde{v}^{\theta}}-1\right)\right]^{1/\theta}\right\}$	$\theta \geq 1, \delta > 0$
BB6	$\left[-\log\left(1-(1-t)^{\theta}\right)\right]^{\delta}$	$C(u,v) = 1-\left(1-\exp\left\{-\left[\left(-\log(1-(1-u)^{\theta})\right)\right)^{\delta} + \left(-\log(1-(1-v)^{\theta})\right)^{\delta}\right]^{\frac{1}{\delta}}\right\}\right)^{\frac{1}{\theta}}$	$\theta \geq 1,\ \delta \geq 1$
BB7	$\left(1-(1-t)^{\theta}\right)^{-\delta}-1$	$C(u,v) = 1-\left(1-\left[\left(1-(1-u)^{\theta}\right)^{-\delta} + \left(1-(1-v)^{\theta}\right)^{-\delta}-1\right]^{-1/\delta}\right)^{\frac{1}{\theta}}$	$\theta \geq 1,\ \delta > 0$

Table 3.2: Several parametric families of Archimedean Copulas implemented in the EVANESCE package. Here, $\tilde{u} = -\log(u)$ and $\tilde{v} = -\log(v)$.

• Capéraà *et al.* [2000] combined the EV and Archimedean copula classes into a single class called *Archimax* copulas. The Archimax copulas are copulas which can be represented in the following form:

$$C(u,v) = \phi^{-1}\left[(\phi(u)+\phi(v))\,A\left(\frac{\phi(u)}{\phi(u)+\phi(v)}\right)\right], \tag{3.6}$$

where $A(t)$ is a valid dependence function and ϕ is a valid Archimedean generator. Archimax copulas reduce to Archimedean copulas for $A(t) = 1$, and to EV copulas for $\phi(t) = -\log(t)$. Capéraà *et al.* [2000] proved that (3.6) is a valid copula, for any combination of valid functions $\phi(t)$ and $A(t)$. For example, one might take

$$\phi(t) = t^{-\theta} - 1 \qquad \text{and} \qquad A(t) = 1 - (t^{-\delta} + (1-t)^{-\delta})^{-1/\delta}$$

and obtain the following copula:

$$C(u,v) = \left(x^{-\theta} + u^{-\theta} - 1 - \left[\left(u^{-\theta}-1\right)^{-\delta} + \left(v^{-\theta}-1\right)^{-\delta}\right]^{-\frac{1}{\delta}}\right)^{-\frac{1}{\theta}}.$$

This copula family is denoted as BB4 in Joe [1997], and is also implemented in EVANESCE.

• Some Archimedean copulas can also be represented an the Archimax form with dependence functions not equal to 1. For example, the BB1 copula

$$C(u,v) = \left\{1 + \left[(u^{-\theta}-1)^{\delta} + (v^{-\theta}-1)^{\delta}\right]^{1/\delta}\right\}^{-1/\theta},$$

which is an Archimedean copula with a two-parameter generator (Table 3.6), can be represented as an Archimax copula with the Gumbel dependence function

$$A(t) = (t^{\delta} + (1-t)^{\delta})^{1/\delta} \qquad \text{and} \qquad \phi(t) = t^{-\theta} - 1$$

where $\theta > 0$ and $\delta \geq 1$. This representation is important, since Capéraà *et al.* [2000] demonstrated how to determine the maximum attractors of these distributions. This representation is also convenient from a purely computational standpoint, especially for the generation of random variables (see section 3.4).

To get a feeling of how copula families differ from one another and what type of dependencies they can describe, we plotted contours of joint density for several bivariate distributions with different copulas and standard Gaussian margins (Figures 3.1 and 3.2). The parameters were chosen so that all the copulas have the same Kendall's tau $\tau = 0.5$ (see (3.7) for the definition of the Kendall's tau). The Pearson correlation coefficients of these distributions range from 0.66 to 0.72 (see figure captions), which is inconsequential. Note how different resulting density contours are.

EV copulas are very similar to each other, especially if one considers one parameter families. The contour plots of density of the three one-parameter EV copulas with Gumbel margins is shown in Figure 3.3. This is one of the reasons that it is very hard to discern the best copula from these families to describe the observed data for similar values of Kendall's tau [Cistinot *et al.*, 2000]. Fortunately, it is usually not necessary to do this, since a bivariate model based on any of these copula, will produce a very similar dependence between random variables to that of the best copula that could be chosen. The same is true for the two-parameter extension of the Gumbel copula (BB5): for a fixed Kendall's tau, the copula density it is very insensitive changes for in the parameters.

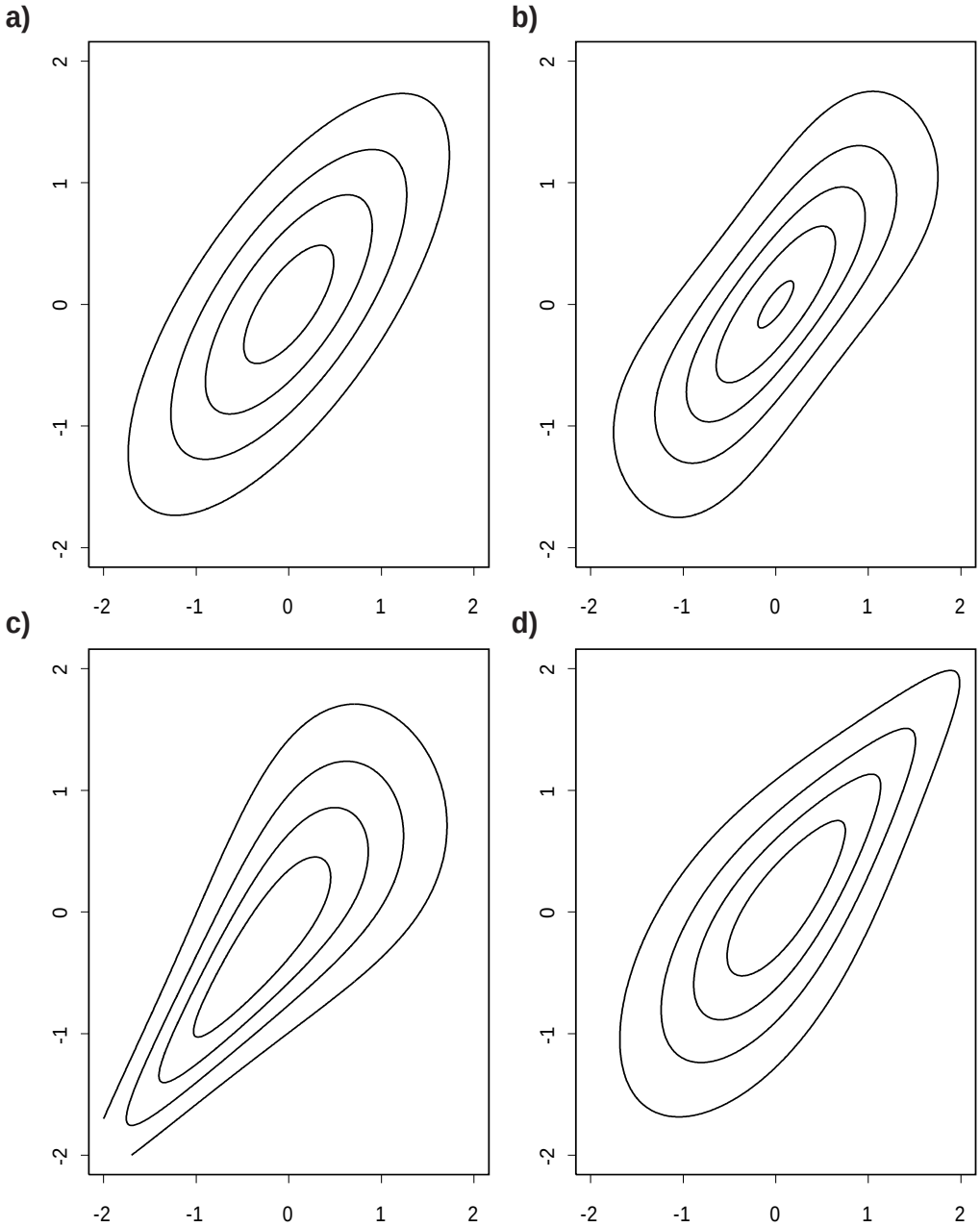

Figure 3.1: Contours of density for different copula families with standard Gaussian margins. a) Gaussian, $\delta = 0.707$, $\rho_P = 0.71$ (Pearson correlation coefficient); b) Frank, $\delta = 5.74$, $\rho_P = 0.66$; c) Kimeldorf and Sampson, $\delta = 2$, $\rho_P = 0.69$; d) Gumbel, $\delta = 2$, $\rho_P = 0.70$.

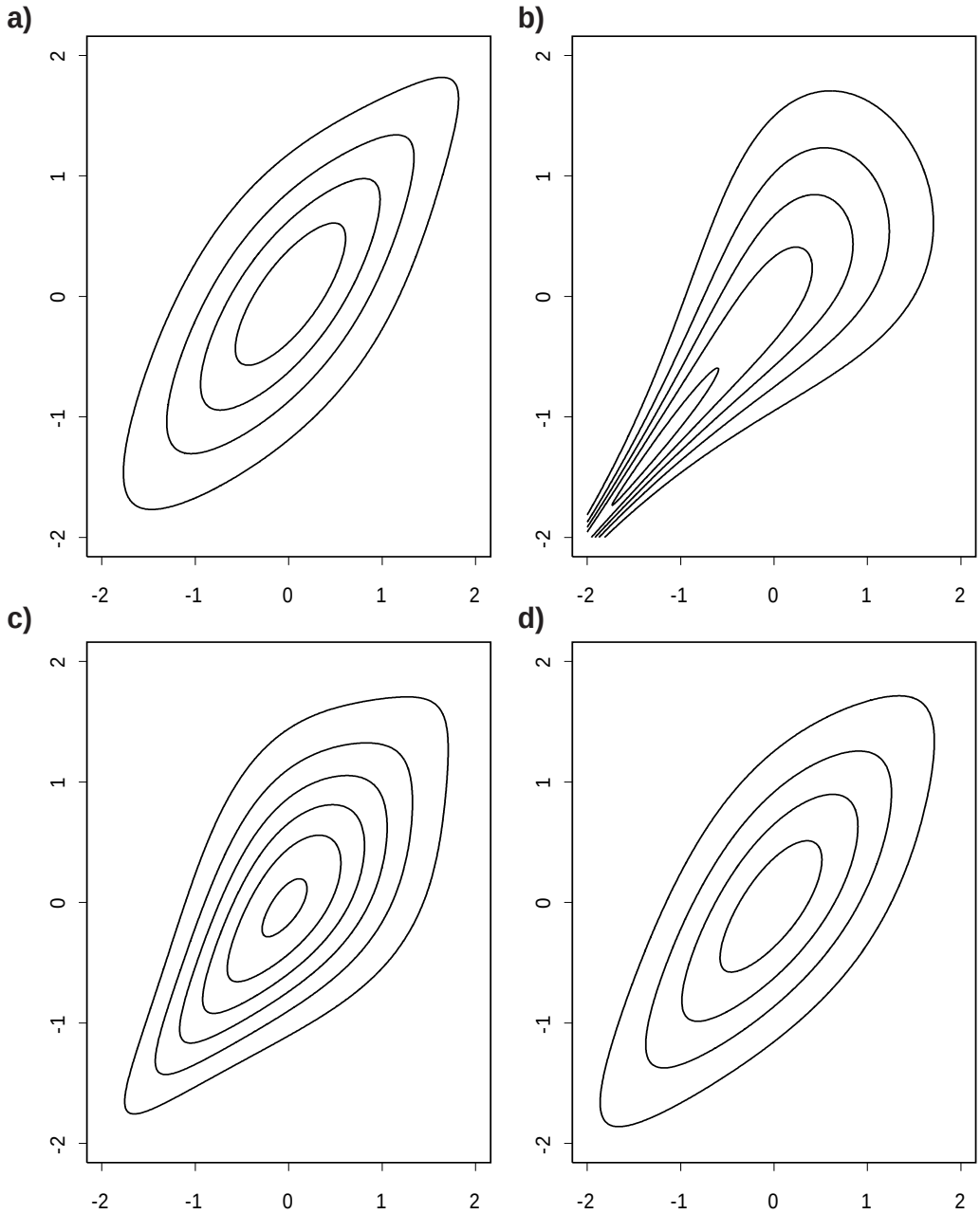

Figure 3.2: Contours of density for different copula families with standard Gaussian margins. a) BB1, $\theta = 0.42$, $\delta = 1.65$, $\rho_P = 0.71$ (correlation coefficient); b) BB2, $\theta = 0.5$, $\delta = 2.05$, $\rho_P = 0.70$; c) BB3, $\theta = 1.4$, $\delta = 0.85$, $\rho_P = 0.72$; d) BB4, $\theta = 0.7$, $\delta = 0.75$, $\rho_P = 0.72$.

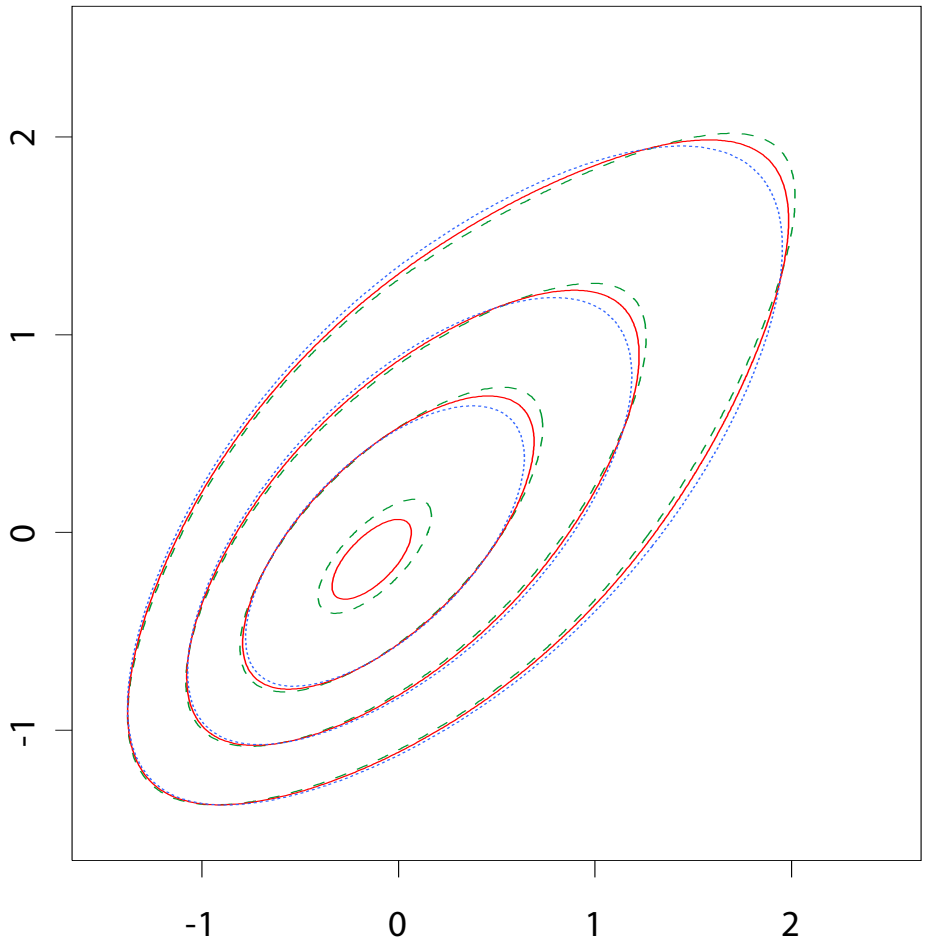

Figure 3.3: Contours for one-parameter copula families with Gumbel margins and Kendall's tau = 0.5: dashed lines – Gumbel, $\delta = 2$; solid lines – Galambos, $\delta = 1.28$; dotted lines – Hüsler and Reiss, $\delta = 1.8$.

Unlike EV copulas, Archimedean copulas can describe a variety of dependencies, even for a fixed Kendall's tau (compare Figures 3.1 b), c), 3.2 a), b), and c)). The two-parameter Archimedean families described here can also describe quite different dependencies at fixed Kendall's tau. Figure 3.4 shows density contours for BB1 and BB7 copulas with standard Gaussian margins.

3.3 Dependence Summaries

A bivariate copula provides complete information regarding the dependence between its two marginal random variables, but sometimes we would like to condense this information into one number.

A dependence measure can be defined as a functional on a space $\mathcal{F}$ of joint distribution functions, or alternatively defined as a function of two random variables [Embrechts *et al.*, 2000b]. We will use the former definition in order to avoid notational ambiguity arising with the latter. Technically speaking, any functional $d : \mathcal{F} \mapsto I\!R$ can be considered a dependence measure, but the following properties are considered desirable for such measures [see Hutchinson and Lai, 1990]:

1. $-1 \leq d(F) \leq 1$ for all $F \in \mathcal{F}$ (this normalization is customary; it is motivated by the properties of the Pearson correlation coefficient).

2. Symmetry: $F_1(x, y) = F_2(y, x) \quad \forall (x, y) \in I\!R^2 \implies d(F_1) = d(F_2)$

3. Invariance under increasing transformations: if $F, G \in \mathcal{F}$, and $F(x, y) = G(g_1(x), g_2(y))$ for some strictly increasing functions g_1 and g_2, then $d(F) =$

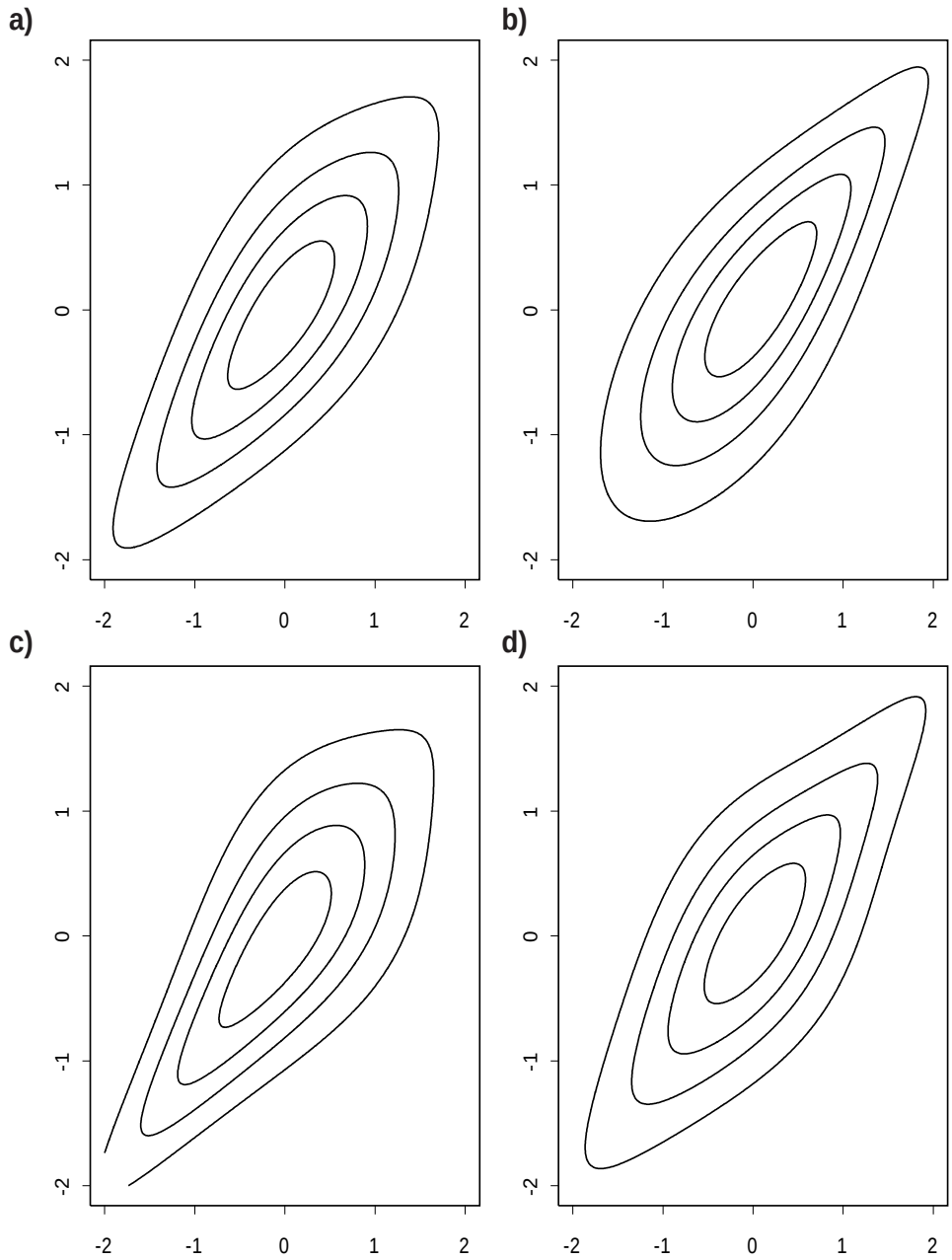

Figure 3.4: Contours for two-parameter copula families BB1 and BB7 with standard Gaussian margins and Kendall's tau = 0.5: a) BB1, $\theta = 0.8$, $\delta = 1.43$; b) BB1, $\theta = 0.1$, $\delta = 1.90$; c) BB7, $\theta = 1.9$, $\delta = 1.11$; d) BB7, $\theta = 1.43$, $\delta = 1.64$. The upper tail index for the copulas is 0.38 in figures a) and c) and 0.56 in figures b) and d).

$d(G)$. Additionally, if g_1 is strictly increasing and g_2 is strictly decreasing, then $d(F) = -d(G)$.

4. For X and Y as the two random variables with joint distribution F:

   a) $d(F) = 1 \iff X = g(Y)$ and $d(F) = -1 \iff X = -g(Y)$ for some strictly increasing function g.

   b) $d(F) = 0 \iff X$ and Y are independent.

It is a well-known result that there is no measure for bivariate distributions that can satisfy all of the properties above, (see, for example, Hutchinson and Lai [1990]), so we must sacrifice at least one property when defining a dependence measure. The usual victim is property 4 b): it is difficult to require that a zero dependence measure implies the independence of two random variables for all distributions, but this property usually holds for a number of the distributions of interest, regardless.

- The traditional measure of the dependence between two random variables X and Y is the Pearson correlation coefficient:

$$\rho_P = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

It can be easily seen that ρ_P satisfies properties 1 and 2 from the above, but not properties 3 and 4 [Joe, 1997, p. 32]. Another drawback of ρ_P is that the random variables are required to have a finite second moment for ρ_P to exist. This is not always true for distributions with heavy tails of crucial importance in hydrology and financial engineering.

Two very common measures that satisfy all of the properties except 4 b) are Kendall's tau and Spearman's rho.

- Let F be a continuous bivariate cumulative distribution function (cdf), and let (X_1, Y_1) and (X_2, Y_2) be two independent pairs of random variables from this distribution. The *Kendall's tau* for the distribution F is defined as

$$\tau = I\!P\left\{(X_1 - X_2)(Y_1 - Y_2) > 0\right\} - I\!P\left\{(X_1 - X_2)(Y_1 - Y_2) < 0\right\}$$

If C is the copula associated with F, then

$$\tau = 4\int CdC - 1 = 4\int_0^1\int_0^1 C(u,v)c(u,v)dudv - 1, \tag{3.7}$$

where $c(u,v)$ is the copula density. Kendall's tau can be viewed as a functional on the space of copulas, and is the same for all distributions associated with the same copula. (A copula C is said to be associated with the distribution F if C and F are related by equation (3.2).).

In general, it is impossible to express the value of Kendall's tau as a function of an individual copula's parameters. However, simple formulas are available for Gaussian copulas ($\tau_{\text{Gauss}} = \frac{2}{\pi}\arcsin\delta$), Kimeldorf and Sampson copulas ($\tau_{\text{KimSamp}} = \frac{\delta}{\delta+2}$), and Gumbel copulas ($\tau_{Gumbel} = \frac{\delta-1}{\delta}$). For Archimedean, EV, and Archimax copulas, the double integration in (3.7) can be reduced to a one-dimensional integral; for an Archimedean copula with generator $\phi(t)$,

$$\tau_{\text{Archm}} = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt,$$

(see Genest and Mackay [1986b]), and for an EV copula with dependence function $A(t)$,

$$\tau_{\mathrm{EV}} = \int_0^1 \frac{t(1-t)A''(t)}{A(t)} dt$$

(see Ghoudi *et al.* [1998]). For Archimax copulas, we also have

$$\tau_{\mathrm{Archimax}} = \tau_{\mathrm{EV}} + (1 - \tau_{\mathrm{EV}})\tau_{\mathrm{Archm}},$$

where τ_{EV} and τ_{Archm} are computed using the formulas above with the dependence function and the Archimedean generator of the Archimax copula [Capéraà *et al.*, 2000].

The EVANESCE library contains functions for computing Kendall's tau for all parametric families introduced in section 3.2. The library is designed so that the call `Kendalls.tau(copula)` in the least-computationally-intensive way will result in the computation of Kendall's tau for this specific parametric copula.

• For a pair of random variables (X, Y) with joint distribution F and marginal distributions G_1 and G_2, *Spearman's rho* is defined as the correlation between $G_1(X)$ and $G_2(Y)$. Since $G_1(X)$ and $G_2(Y)$ are uniform $(0, 1)$ random variables, $\mathbb{E}[G_1(X)] = \mathbb{E}[G_2(Y)] = \frac{1}{2}$ and $\mathrm{Var}(G_1(X)) = \mathrm{Var}(G_2(Y)) = \frac{1}{12}$. For a copula C associated with X and Y, then,

$$\rho_S = 12 \int_0^1 \int_0^1 u\, v\, dC(u, v) - 3 = 12 \int_0^1 \int_0^1 C(u, v)\, du\, dv - 3.$$

Though there is no simplification of this formula for the Archimedean copulas, it has been shown for EV copulas that

$$\rho_S = 12 \int_0^1 \frac{dt}{(A(t)+1)^2} - 3,$$

(see Capéraà *et al.* [1997a]). EVANESCE can compute Spearman's rho for all parametric copulas described in section 3.2.

For three random variables of interest X, Y, and Z, one important question to answer is whether variable X has greater dependence on Y or Z, and how we can best make the distinction. The underlying idea is that one can define different orderings on the space of all copulas, and ask whether $C_{(X,Y)} \prec C_{(X,Z)}$ or $C_{(X,Y)} \succ C_{(X,Z)}$ under any particular ordering (see Joe [1997], section 2.2, for several dependence orderings). The most common ordering is the concordant ordering:

$$C_1 \prec_c C_2 \Longleftrightarrow C_1(u,v) \leq C_2(u,v) \qquad \forall\,(u,v) \in \mathbf{I}^2.$$

Unfortunately, it is not always convenient to use copulas for this sort of analyses. Moreover, this ordering is not total and it can very easily occur that $C_1(u_1,v_1) < C_2(u_1,v_1)$ for one point $(u_1,v_1) \in \mathbf{I}^2$, and $C_1(u_2,v_2) > C_2(u_2,v_2)$ for another point (u_2,v_2). Sometimes, we can compare copulas by mapping the area where $C_1(u_2,v_2) > C_2(u_2,v_2)$ [Cistinot *et al.*, 2000] (this can be done automatically in EVANESCE), but it is more convenient to use another ordering. Let (U_1,V_1) and (U_2,V_2) be two pairs of random variables from copulas C_1 and C_2, respectively. Define $W_1 = C_1(U_1,V_1)$, and $W_2 = C_2(U_2,V_2)$ and let $K_1(w)$ and $K_2(w)$ denote the cumulative distribution

functions of W_1 and W_2, respectively. Capéraà *et al.* [1997b] showed that one can define a copula ordering by

$$C_1 \prec_K C_2 \Longleftrightarrow K_1(w) \geq K_2(w) \quad \forall w \in [0,1].$$

This ordering is neither stronger nor weaker than the concordance ordering $\prec_c$. As in the case of concordance ordering, $C_1 \prec_K C_2 \Longrightarrow \tau_{C_1} \leq \tau_{C_2}$, where τ_{C_i} denotes the Kendall's tau of copula C_i, $i = 1, 2$. The ordering $\prec_K$ has some very convenient properties for Archimedean, EV, and Archimax copulas. Let C_A denote an EV copula with dependence function A, C_ϕ denote an Archimedean copula with generator ϕ, and $C_{\phi,A}$ denote an Archimax copula with generator ϕ and dependence function A. The following properties hold (see Capéraà *et al.* [1997b, 2000]):

$$C_{A_1} \prec_c C_{A_2} \Longrightarrow \tau_{C_{A_1}} \leq \tau_{C_{A_2}} \Longleftrightarrow C_{\phi,A_1} \prec_K C_{\phi,A_2} \Longleftrightarrow C_{A_1} \prec_K C_{A_2};$$

$$\tau_{C_{\phi_1}} \leq \tau_{C_{\phi_2}} \Longleftarrow C_{\phi_1} \prec_K C_{\phi_2} \Longleftrightarrow C_{\phi_1,A} \prec_K C_{\phi_2,A} \Longrightarrow C_{\phi_1} \prec_c C_{\phi_2}$$

We will only use the $\prec_K$ ordering in this dissertation, so hereafter we omit the K index.

The last quantity of interest to define is the tail index. The copula's upper tail index λ_U is defined as

$$\lambda_U = \lim_{\alpha \longrightarrow 1-} \frac{1 - 2\alpha + C(\alpha, \alpha)}{1 - \alpha}, \tag{3.8}$$

whenever the limit exists. If the distribution of a pair (U, V) is given by a copula with upper tail index λ_U, then

$$\lambda_U = \lim_{\alpha \longrightarrow 1-} I\!P \left\{ U > \alpha | V > \alpha \right\}. \tag{3.9}$$

This measure is very important for risk management, as it specifies how much the tails of the marginal distributions are correlated [Embrechts *et al.*, 2000b]. Similarly, the lower tail index is defined as

$$\lambda_L = \lim_{\alpha \longrightarrow 0+} \frac{C(\alpha, \alpha)}{\alpha}.$$

EVANESCE includes functions that compute upper and lower tail indexes for each parametric copula. For EV copula with generator $A(t)$, note that we will always have $\lambda_L = 0$ and $\lambda_U = 2 - 2A(1/2)$. The upper and lower tail indexes for the Gaussian copula are zero. Formulas for upper and lower tail indexes for several parametric families from section 3.2 are listed in Table 3.3.

The ability to compute different measures of dependence becomes very useful in providing a feel for the dependence between several random variables. In section 4.1, for example, we will see how computing these measures will help us to choose between the possible parametric copulas for describing a given set of data.

3.4 Simulation of Random Pairs

The ability to perform Monte Carlo simulations from a given (or estimated) bivariate distribution is a powerful tool in financial engineering and statistical hydrology. EVANESCE contains the function `rcopula` that returns simulated observations from a given parametric bivariate copula. Used with marginal quantile functions, it allows the user to simulate random observations from many different bivariate distributions. In this section, we describe the simulation algorithms implemented in EVANESCE.

Copula family	Lower tail index	Upper tail index
BB1	$2^{-\frac{1}{\delta\theta}}$	$2-2^{1/\delta}$
BB2	1	0
BB3	$2^{-1/\delta}$ if $\theta=1$ 1 if $\theta>1$	$2-2^{1/\theta}$
BB4	$\left(2-2^{-1/\delta}\right)^{-1/\theta}$	$2^{-1/\delta}$
BB6	0	$2-2^{-\frac{1}{\delta\theta}}$
BB7	$2^{-1/\delta}$	$2-2^{1/\theta}$

Table 3.3: Formulas for upper and lower tail indexes for several parametric copula families.

One standard method for simulating a random observation (u, v) from a bivariate copula C involves using the inverse of the conditional distribution $I\!P\{V \leq v|U = u\} = C_{V|U} = \frac{\partial C(u,v)}{\partial v}$. Of all the families listed in section 3.2, the inverse of $C_{V|U}$ exists in a closed form only for the Frank copula and the Kimeldorf and Sampson copula, and the inverse method is therefore used in EVANESCE for only these two bivariate copulas. It is also employed in simulations from three-dimensional copulas (see section 3.7), however.

Genest and Mackay [1986a] introduced the following algorithm for generating pairs (u, v) of random observations from an Archimedean copula with an Archimedean generator ϕ:

- Generate two independent random numbers u and t from the uniform $(0, 1)$ distribution.
- Set $w = \phi'^{(-1)}\left(\dfrac{\phi'(u)}{t}\right)$, where $\phi'(t) = \frac{d}{dt}\phi(t)$.
- Set $v = \phi^{-1}\left[\phi(w) - \phi(u)\right]$.
- Take (u, v) to be the random observation.

Another algorithm for generating random pairs from an Archimedean copula is based on the following theorem (from Genest and Rivest [1993]):

Let U and V be uniform random variables with a joint distribution given by an Archimedean copula C with generator ϕ. Let $S = \dfrac{\phi(U)}{\phi(U) + \phi(V)}$ and $T = C(U, V)$.

Then, S and T are independent random variables, S is uniformly distributed on $(0,1)$, and the distribution of T is given by

$$K(t) = I\!P\{T \leq t\} = t - \frac{\phi(t)}{\phi'(t)},$$

where $\phi'(t) = \frac{d}{dt}\phi(t)$.

It immediately follows from this theorem that the following algorithm generates an observation (u, v) from an Archimedean copula.

- Draw a number r from a distribution uniform on $(0,1)$. Set $t = K^{-1}(r)$ (K^{-1} denotes the inverse of K).
- Draw another number s from a distribution uniform on $(0,1)$, independently of r.
- Set $u = \phi^{-1}[s\phi(t)]$ and $v = \phi^{-1}[(1-s)\phi(t)]$.

The Genest and Mackay [1986a] algorithm uses the inverse of ϕ', and the algorithm above uses the inverse of K. For some parametric Archimedean copulas in section 3.2, neither function is invertible in closed form, and we employed the Newton method for computationally inverting them. The points at which the inverse is to be computed are sorted, and the value of the inverse at each point is used as the initial guess for the next point. The number of iterations performed to find a solution is thusly minimized. The second algorithm can also be used to simulate random observations from empirically estimated copulas (see section 3.5).

Capéraà *et al.* [2000] suggested an algorithm for generating random pairs from Archimax copulas. Let C be an Archimax copula of with Archimedean generator ϕ and dependence function A, and (U, V) be a pair of random variables joint cumulative distribution function C. Under the assumption that A has a continuous second derivative, it can be shown that the function H defined by:

$$H(z) = z - z(1-z)\frac{A'(z)}{A(z)} \qquad 0 \leq z \leq 1.$$

is the cumulative distribution function of a random variable Z of the form $Z = \frac{\phi(U)}{\phi(U)+\phi(V)}$, and that, if $W = C(U, V)$, the conditional distribution of W given Z is

$$\mathbb{P}\{W \leq w | Z = z\} = p(z)w + (1 - p(z))K_\phi(w),$$

where

$$p(z) = \frac{z(1-z)A''(z)}{H'(z)A(z)}$$

and $K_\phi(t) = \left(1 - \frac{\phi(t)}{\phi'(t)}\right)$. Note that K_ϕ is the distribution of $\phi^{-1}[\phi(R) + \phi(S)]$, where (R, S) is a random pair from an Archimedean copula with the generator ϕ. The following algorithm will then generate an observation (u, v) from an Archimedean copula C [Capéraà *et al.*, 2000]:

- Generate z from the distribution H. Notice that, since A'' is continuous, H' is bounded on $[0, 1]$ and the rejection method can be used for this purpose.

- Generate t from the uniform $(0, 1)$ distribution. If $t \leq p(z)$, draw w from a uniform distribution, otherwise, generate a pair of numbers (r, s) from the Archimedean copula with generator ϕ, and set $w = \phi^{-1}[\phi(r) + \phi(s)]$.

- Set $u = \phi^{-1}\left[\frac{z\phi(w)}{A(z)}\right]$ and $v = \phi^{-1}\left[\frac{(1-z)\phi(w)}{A(z)}\right]$.

Recall that $\phi(t) = -\log t$ for EV copulas, and the Archimedean copula with this generator is simply an independence copula (C(u,v) = uv). For EV copulas, then, r and s can be independently drawn from the uniform distribution.

All of the algorithms above are implemented in EVANESCE. The library is designed so that a call `rcopula(copula,n)` automatically employs the most efficient algorithm for generating random observations from the `copula`.

3.5 Estimation of Copulas

Again, starting with a sample of n independent observations of random pairs $(u_1, v_1), (u_2, v_2), \ldots (u_n, v_n)$ from the distribution $C(u, v)$ with uniform marginals, we discuss how to estimate C.

3.5.1 Empirical Copulas

Similarly to the univariate case, Deheuvels [1978] introduced non-parametric empirical estimate for copulas called *empirical copulas*. Let $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$ and $v_{(1)} \leq v_{(2)} \leq \ldots \leq v_{(n)}$ be the order statistics of the univariate samples. The empirical copula $\hat{C}_{\text{emp}}$ is then defined at the points $\left(\frac{i}{n}, \frac{j}{n}\right)$ by the formula:

$$\hat{C}_{\text{emp}}\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{u_k \leq u_{(i)}, v_k \leq v_{(j)}\}}, \qquad i, j = 1, 2, \ldots, n \tag{3.10}$$

One can also define an *empirical copula frequency* $\hat{c}_{\text{emp}}$ as

$$\hat{c}_{\text{emp}}\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{1}{n} \cdot \begin{cases} 1 & \text{if } (u_{(i)}, v_{(j)}) \text{ is an element of the sample,} \\ 0 & \text{otherwise.} \end{cases}$$

The empirical copula frequency can be considered an analog of empirical copula density, since

$$\hat{C}_{\text{emp}}\left(\frac{i}{n}, \frac{j}{n}\right) = \sum_{k=1}^{i} \sum_{m=1}^{j} \hat{c}_{\text{emp}}\left(\frac{k}{n}, \frac{m}{n}\right)$$

Deheuvels [1978] showed that the sequence of empirical copulas converges to the true copula of the data as the sample size increases.

Empirical copulas can be a useful tool for exploratory data analyses. Figure 3.5, shows contours of a Gumbel copula with parameter $\delta = 2.0$ and of an empirical copula from a random sample of size 100 generated from this Gumbel copula. If we did not know the true copula of our sample, we could have inferred from this plot that the Gumbel copula was a good model for our data. One can also construct non-parametric tests for independence based on the empirical copulas. Details about these tests can be found in Deheuvels [1981a,b].

3.5.2 A Non-Parametric Estimate for Archimedean Copulas

The following non-parametric estimate for Archimedean copulas was proposed by Genest and Rivest [1993]. The Archimedean generator $\phi(t)$ can be written as

$$\phi(t) = \exp\left\{\int_{t_0}^{t} \frac{1}{\lambda(s)}\, ds\right\}, \tag{3.11}$$

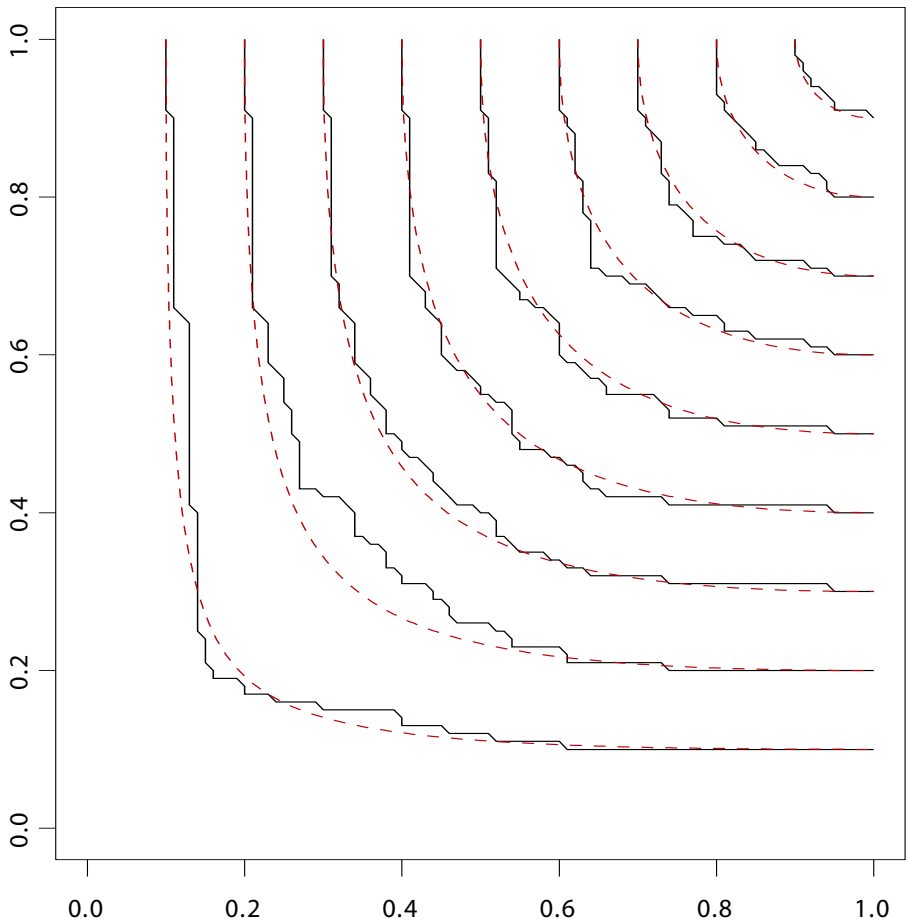

Figure 3.5: Contours of a Gumbel copula (dashed lines) with parameter $\delta = 2.0$ and of an empirical copula (solid lines) from a random sample of size 100 generated from this Gumbel copula.

where $\lambda(s) = \frac{\phi(s)}{\phi'(s)} = s - K(s)$ and t_0 is an arbitrary constant. Genest and Rivest's [1993] non-parametric procedure consists of using a kernel estimator for $K(s)$, using (3.11) to estimate ϕ (numerical integration is required), and then using (3.6) to obtain the actual estimate of C. To compute the kernel estimate for $K(s)$, one computes $z_i = \frac{1}{n-1} \cdot \text{Card}\ \{(u_j, v_j) : u_j < u_i, v_j < v_i\}$. For a given kernel $k(s)$ (which integrates to 1), the estimator of $K(s)$ is then

$$\hat{K}(s) = \frac{1}{n} \sum_{i=1}^{n} k(s - z_i). \tag{3.12}$$

Genest and Rivest [1993] show that this estimator is consistent under certain regularity conditions. A nice property of this estimator is that it can also be applied directly to raw data points before the marginal transformation. Also, if the true copula of (U, V) is not an Archimedean copula, this estimator (asymptotically) approaches a projection of the true copula of (U, V) onto the class of Archimedean copulas.

Figure 3.6 illustrates the estimate of a function $\lambda(s)$ for a sample of size 100 from a BB1 copula (solid curve), and the true function $\lambda(s)$ for this copula (dashed curve). It is important to note, that the non-parametric estimator $\hat{K}$ constructs a copula from which we can generate random observations, using the algorithm on page 97. This procedure is implemented in EVANESCE (function `rcopula.archm.est`), but it is much slower than the procedures that generate random observations from parametrically defined copulas.

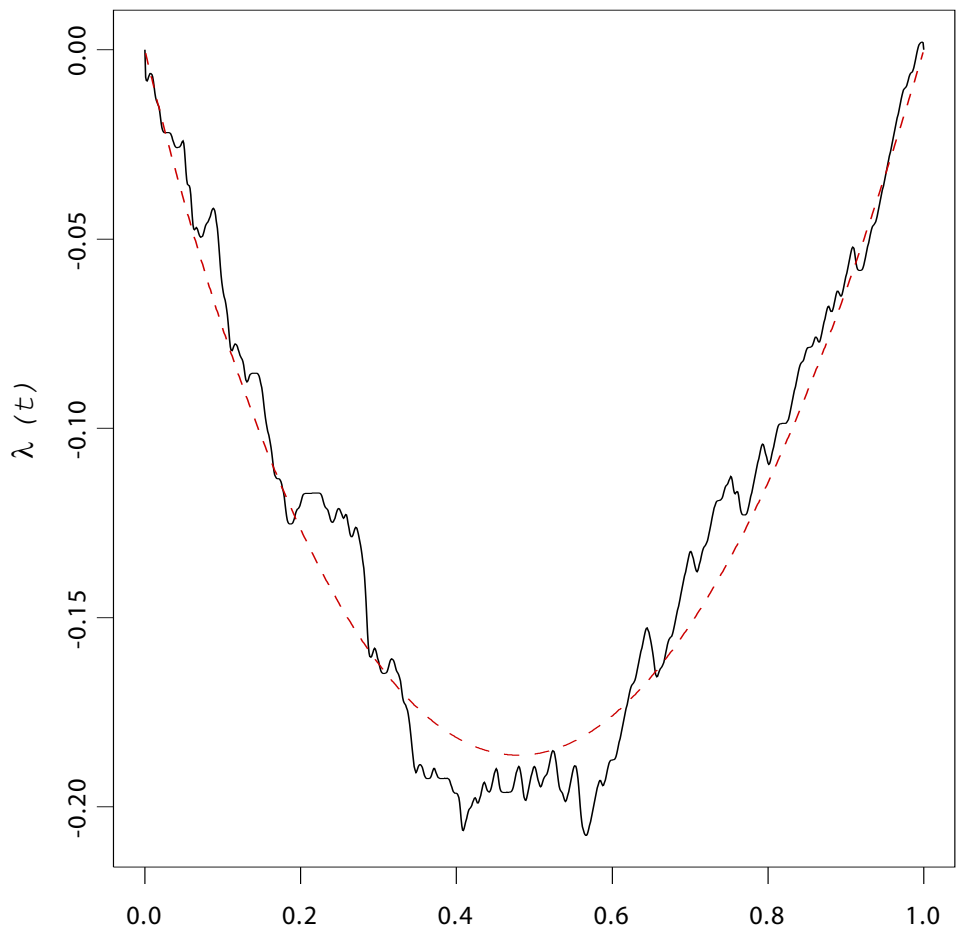

Figure 3.6: The estimate of a function $\lambda(s)$ for a sample of size 100 from a BB1 copula (solid curve), and the true function $\lambda(s)$ for this copula (dashed curve).

3.5.3 A Non-Parametric Estimate for EV Copulas

Capéraà *et al.* [1997a] introduced a non-parametric estimate for the dependence function of EV copulas. It can be computed using the following algorithm:

- Compute $z_i = \dfrac{\log(u_i)}{\log(u_i v_i)}$, $i = 1, 2, \ldots, n$. Let $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$ denote ordered z_i's.

- Compute
$$q_i = \left[\prod_{k=1}^{i} \frac{z_{(k)}}{1 - z_{(k)}}\right]^{\frac{1}{n}} \qquad i = 1, 2, \ldots, n$$

- The estimator of $A(t)$ is given by the following formula
$$\hat{A}(t) = \begin{cases} (1-t)q_n^{1-p(t)} & \text{if} \quad 0 \le t \le z_{(1)}, \\ t^{i/n}(1-t)^{1-i/n} q_n^{1-p(t)} q_i^{-1} & \text{if} \quad z_{(i)} < t \le z_{(i+1)} \qquad i = 1, 2, \ldots, n-1, \\ t q_n^{-p(t)} & \text{if} \quad z_{(n)} < t \le 1, \end{cases}$$
where $p(t)$ is a weighting function. $p(t)$ must to be bounded on $[0, 1]$ and satisfy $p(0) = 1 - p(1) = 1$. One can choose several expressions for $p(t)$, the simplest one being $p(t) = 1 - t$.

- The copula estimate is then
$$\hat{C}_{\text{Cap}}(u, v) = \exp\left\{\log(uv)\hat{A}\left(\frac{\log u}{\log(uv)}\right)\right\}$$

Capéraà *et al.* [1997a] showed that, though the estimator $\hat{A}(t)$ is biased for finite samples, it is asymptotically unbiased and strongly consistent. It is also easy to see that the estimate of A is a valid dependence function. Capéraà *et al.* [1997a] compared

the performance of this estimator using simulations for Gumbel and Tawn copulas on small and medium samples to two other estimators proposed by Pickands [1981] and Deheuvels [1991], and concluded that this estimator $\hat{A}(t)$ performed better than the other two, which is why we chose to implement only this estimator in EVANESCE.

3.5.4 Parametric Copula Estimation

The major advantage of the non-parametric estimators is that we do not have to make any assumptions about the true copula structure, or only minimal assumptions that it belongs to a certain class. Unfortunately, simulation of random samples using these estimates is not practical, since it requires a consistent definition of the inverse of an empirical copula's conditional distribution function. One could instead use the Archimedean estimator described above, but the procedure is not particularly efficient. As a result, parametric estimation of copulas are more desirable.

For a parametric estimation, methods of maximum likelihood are the first choice [Joe, 1997]. The interesting question that arises is how to choose the best parametric copula for our data.

One method for doing this is based on Pearson goodness-of-fit statistics [D'Agostino and Stephens, 1986, p. 66]: a) divide $\mathbf{I}^2$ into M cells; b) compute from a sample of size n the number of observations N_j falling in each cell E_j for $j = 1, 2, \ldots M$; c) compute the probability of obtaining an observation in each cell $p_j = I\!P\{(X, Y) \text{ is in cell } E_j\}$ for $j = 1, 2, \ldots M$, for the parametric copula after its parameters have been estimated using, say, maximum likelihood method; d) compute $\mathcal{X}^2 = \sum_{j=1}^{N}(N_j - np_j)^2/(np_j)$;

e) claim that $\mathcal{X}^2$ has a χ^2 distribution with $M - \nu - 1$ degrees of freedom, where ν is the number of parameters of the parametric copula that we estimated. Based on this test, we can reject or not reject the hypothesis that data comes from a particular parametric family of copulas. Genest and Rivest [1993] also suggested that this test be used to choose the "best" copula of the several copulas, i.e., that we should choose the parametric copula with the smallest value of $\mathcal{X}^2$.

Non-parametric estimates of Kendall's tau and Spearman's rho can also help to evaluate the copulas' fit: After estimating parameters for a parametric copula using the method of maximum likelihood, we can compare the implied Kendall's tau and Spearman's rho of the estimated copula with that of the sample. The estimate of Kendall's tau for a sample of size n is the number of the sample's concordant pairs minus the number of discordant pairs divided by the total number of pairs $\binom{n}{2}$. (A pair (u_i, v_i) and (u_j, v_j) of the sample is called concordant if either $u_i < u_j$ and $v_i < v_j$ or $u_i > u_j$ and $v_i > v_j$. It is called discordant if either $u_i < u_j$ and $v_i > v_j$ or $u_i > u_j$ and $v_i < v_j$.) The estimate of the Spearman's rho for a sample of size n can be computed from an empirical copula $\hat{C}_n\left(\frac{i}{n}, \frac{j}{n}\right)$ by the following formula:

$$\hat{\rho}_S = \frac{12}{n^2 - 1} \sum_{j=1}^{n} \sum_{i=1}^{n} \left[\hat{C}_n\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{i}{n} \cdot \frac{j}{n} \right]$$

EVANESCE implements functions for computing both Spearman's rho and Kendall's tau from a sample.

Similarly, empirical estimates of the upper and lower tail index might be useful for comparison between copulas. Recall that the definition of the upper tail index

($\lambda_U = \lim_{\alpha \longrightarrow 1-} \frac{1-2\alpha+C(\alpha,\alpha)}{1-\alpha}$) involves taking a limit as $\alpha \longrightarrow 1-$. For empirical copulas the estimate of λ_U involves creating a plot of $\frac{1-2\alpha+\hat{C}(\alpha,\alpha)}{1-\alpha}$ for some empirical estimate of copula $\hat{C}$ and visually observing possible convergence (see Coles *et al.* [1999]). EVANESCE implements functions for producing such plots for both upper and lower tails.

Non-parametric estimates of $\lambda(s)$ and $A(t)$ can be used in choosing among different parametric copula families with estimated parameters (see Genest and Rivest [1993] and Capéraà *et al.* [1997a]). We will use these non-parametric estimates in exploratory analysis of data (sections 4.1 and 4.5).

3.6 Sensitivity of Parametric Copula Estimates to Errors in the Estimates of Marginal Distributions

The methods for copula estimation described in the previous section assume that the respective marginal distributions $G_1(x)$ and $G_2(y)$ of X and Y are known. In practice, the marginals of X and Y must be estimated from the data. As shown in chapter 2, errors in the quantile estimates can be very large for distributions with heavy tails, implying that there can be large errors in the cdf estimates. To avoid problems with parametric cdf estimates, Genest *et al.* [1995] proposed using empirical estimates of marginal cdfs for copula estimation rather than estimated parametric marginals. Specifically, instead of assuming that the marginal distributions G_1 and G_2 have a particular parametric form and estimating the appropriate parameters,

set $\hat{G}_1(x_i) = (r_i - 0.5)/n$, where r_i is the rank of the observation x_i in the sample $\{x_1, \ldots, x_n\}$ and n is the sample size, and use a similar expression for $\hat{G}_2$. In section 3.6.1, we explore how much the errors in the copula parameter estimates are different for the two methods of marginal cdf estimation. We address this issue for small- to medium-size samples using Monte Carlo simulations, and focus on cases where X and Y have heavy tails. If we find the sensitivity of copula estimates to imprecision in the marginal estimates to be small for heavy-tailed distributions, it indicates that there is an even smaller sensitivity for distributions with bounded or Gaussian tails, since estimation of the cdfs for these distributions can be performed with much better precision than for heavy-tailed distributions.

Another interesting problem, is to determine how much the copula estimates are affected by estimating the marginals parameters before the copula parameters instead of determining all the parameters at once by maximizing the full likelihood (as suggested by Cook and Johnson [1981]). This problem is studied in section 3.6.2 for bivariate distributions with GEV marginals and Gumbel copula.

3.6.1 Comparison of Parametric and Empirical Estimation of the Marginals

To compare the errors made in copula parameter estimation for the two different marginal estimation procedures, we will generate sample data from a known bivariate distribution with Generalized Pareto (GP) marginals, and perform both estimation procedures. The performance of the copula parameter estimation technique depends

on the Kendall's tau of the data's true copula [Capéraà *et al.*, 1997a], so we will also compare the behavior of the estimates at different values of Kendall's tau. We have performed the following simulation experiment:

1. Choose a parametric copula family, C, and a set of parameters for C corresponding to a predetermined value of Kendall's tau. Choose a sample size n and a shape parameter k.

2. Generate a sample of n observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ from a bivariate distribution associated with the copula C and GP marginals with identical scale parameters $\alpha = 1$, location parameters $\delta = 0$, and shape parameters k.

3. Perform the transformations $u_i = G_1(x_i)$, $v_i = G_2(y_i)$, $i = 1, 2, \ldots, n$, where G_1 and G_2 are the true marginal distributions used in step 2. Based on the observations $\{(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)\}$, estimate the parameters of copula C using the maximum likelihood method (In this and the following step, only the true copula family was fitted to the data).

4. , Estimate the parameters of the marginal GP distributions of X and Y, using the MLE method. Perform the transformations $u_i^{(\mathrm{GP})} = \hat{G}_1(x_i)$, $v_i^{(\mathrm{GP})} = \hat{G}_2(y_i)$, $i = 1, 2, \ldots, n$, where $\hat{G}_1$ and $\hat{G}_2$ are the estimated GP marginal distributions. Based on the observations $\{(u_1^{(\mathrm{GP})}, v_1^{(\mathrm{GP})}), (u_2^{(\mathrm{GP})}, v_2^{(\mathrm{GP})}), \ldots, (u_n^{(\mathrm{GP})}, v_n^{(\mathrm{GP})})\}$, estimate the parameters of copula C using MLE.

5. Let $\hat{H}_1(x)$ and $\hat{H}_2(y)$ be the empirical estimates of the marginals. Perform the transformations $u_i^{(\text{emp})} = \hat{H}_1(x_i)$, $v_i^{(\text{emp})} = \hat{H}_2(y_i)$, $i = 1, 2, \ldots, n$. Based on the observations $\{(u_1^{(\text{emp})}, v_1^{(\text{emp})}), (u_2^{(\text{emp})}, v_2^{(\text{emp})}), \ldots, (u_n^{(\text{emp})}, v_n^{(\text{emp})})\}$, estimate the parameters of copula C using MLE.

6. Record the estimates of the copula parameters obtained in steps 3, 4, and 5 for later comparison.

The above procedure was repeated 1,000 times for each choice of copula, k, and n.

The results of the copula parameter estimation may be characterized in several different ways. One could compute some norm (e.g. L_1, L_2, or L_∞) between the estimated copula function and the true copula, compute the distance in parameter space between the vector of estimated parameters and the true parameter vector, or compare the Kendall's tau, Spearman's rho, or tail indexes of the fitted and true copulas [Genest *et al.*, 1995; Genest and Mackay, 1986b; Capéraà *et al.*, 1997a, 2000]. We choose to compare the RMSEs of the parameter estimates (computed in the two-dimensional parameter space for the two-parameter copulas) and the RMSEs of the Kendall's tau estimates.

We have conducted simulations with a variety of parametric copulas, sample sizes, and values of k. Figure 3.7 provides an illustration of these result for $n = 250$, $k = -0.3$, and the BB1 parametric copula, with different sets of parameters corresponding to different values of Kendall's tau. One definitive result was observed for all the simulation experiments: the estimation errors in step 3 were lower than those in step

4, which, in turn, were lower that the errors obtained in step 5, but the difference between the errors obtained in steps 4 and 5 was not large. Overall, the errors in copula estimation with true margins constitute between 77% and 97% of the error in copula estimation with estimated parametric marginals, and between 65% and 93% of the error obtained when using empirical marginals.

- Our results indicate that imperfect specification of the marginal distributions does not significantly increase the copula estimation error.

Even more interesting results were obtained for EV copulas with GEV marginals. The RMSEs of the Gumbel copula parameter estimates in Figure 3.8 are plotted against different values of the marginals shape parameter k. For a given copula, the errors in the copula parameter estimates when using the true marginals were about 80% of those when using estimated marginals, but there was almost no difference between the errors in the estimated copula parameters in the cases were parametrically estimated or empirical marginals where used to transform data for the copula parameter estimation. The error in the estimates when using true or empirical marginals should not depend on k, and the small variations that we observe are an artifact of the simulation experiments. From Figure 3.8, we note that the RMSEs for the estimates using parametrically estimated marginals also do not depend on the value of k.

- Overall, the implication of this analysis is that parametric methods do such a poor job of estimating the GEV cdf that we can replace the parametrically estimated

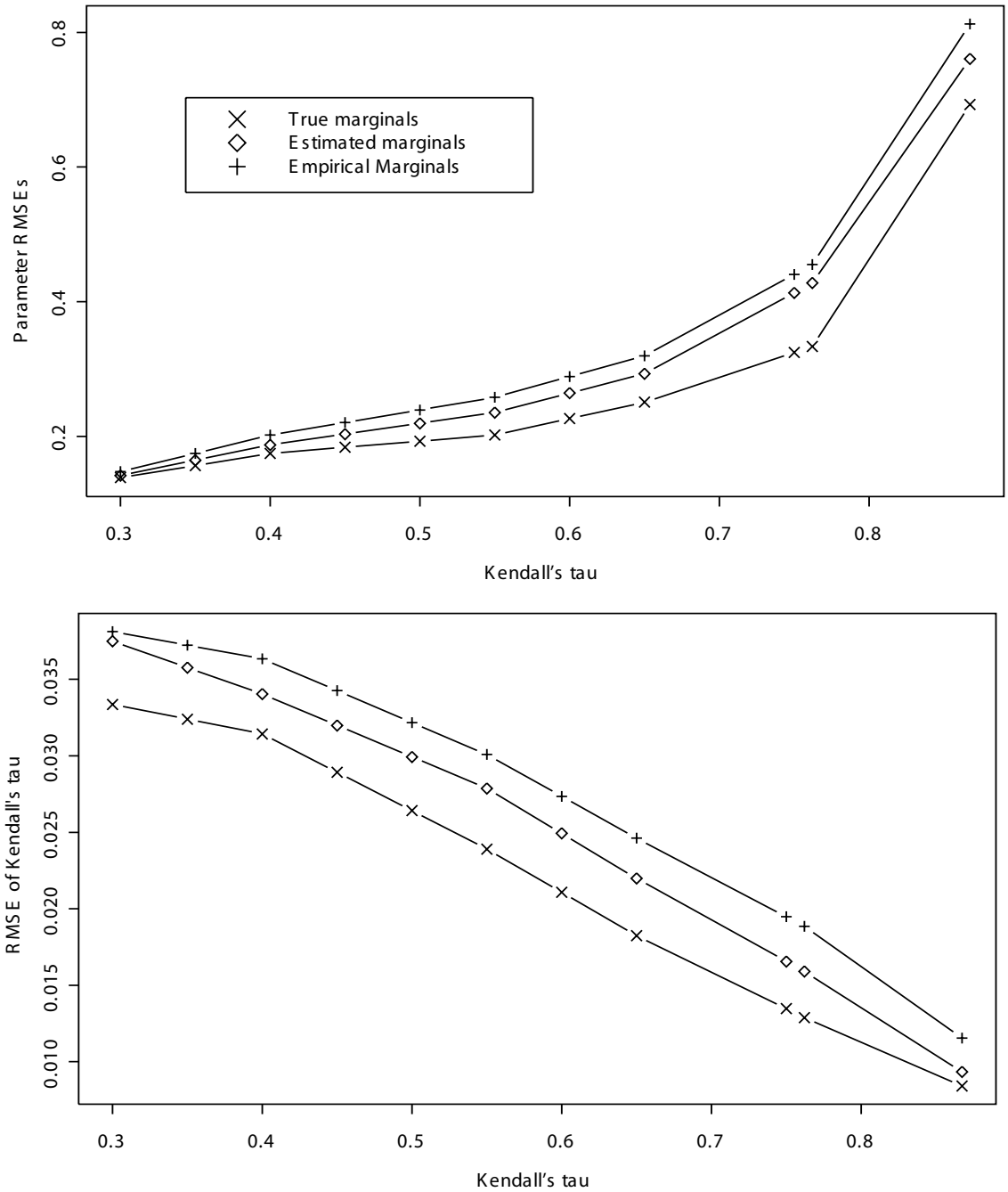

Figure 3.7: RMSE in parameter estimates (top) and Kendall's tau estimates (bottom) for the BB1 copula using different estimates of the GP marginals.

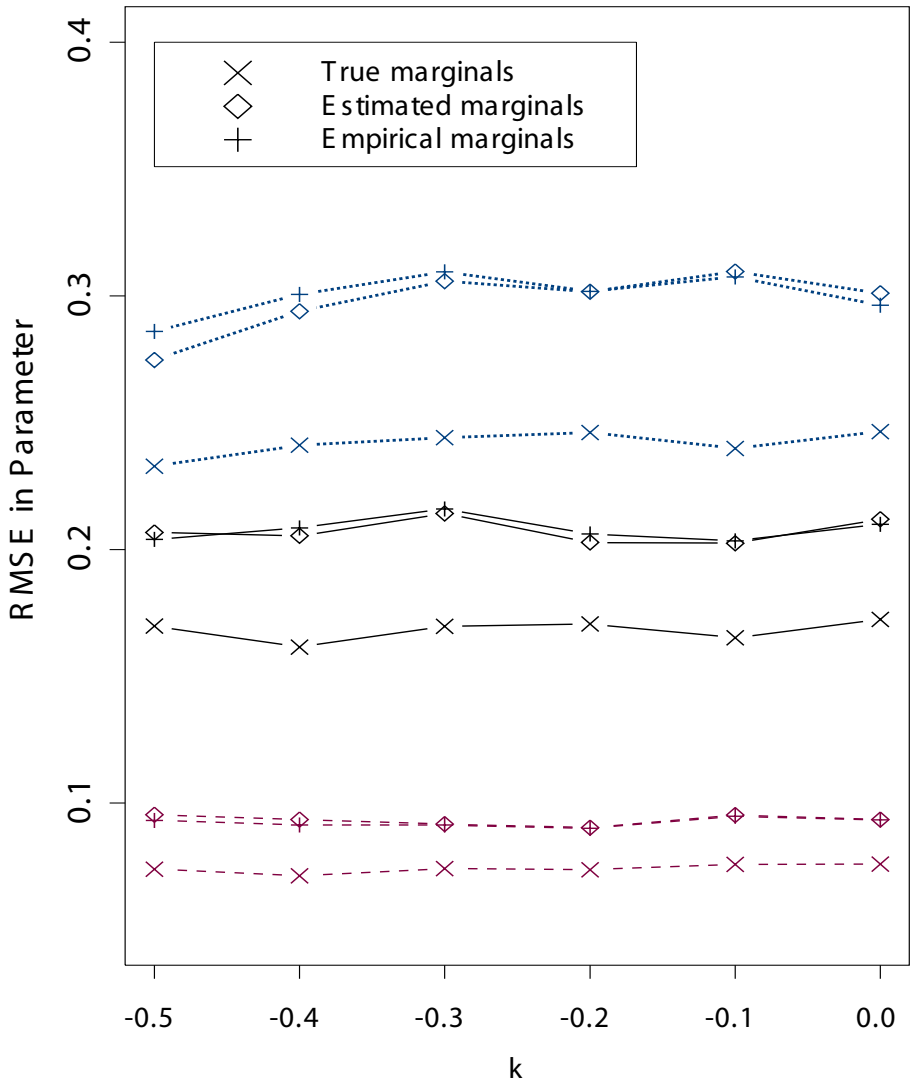

Figure 3.8: RMSE of the Gumbel copula δ parameter estimates using GEV marginals for different sample sizes ($n = 50$ – dotted lines, $n = 100$ – solid lines, and $n = 250$ – dashed lines), and marginal estimation techniques. There is no difference in RMSEs for empirically and parametrically estimated marginals.

margins by empirical estimates, without negatively impacting our copula parameter estimates.

Our analysis supports the suggestion of Genest *et al.* [1995] that we replace GEV marginal estimates by the empirical estimates.

3.6.2 Comparison of the Separate Marginals Parameter Estimation Procedure and the Full Likelihood Method

The performance of the full likelihood method was tested on samples generated from a bivariate distribution with a Gumbel copula and GEV marginals. The distribution has seven parameters: three parameters for each marginal, and one parameter for the Gumbel copula.

To estimate the parameters, we could first estimate the marginals (using the MIX1 method, for instance), and then use the transformed data to estimate the copula, or we could write the total likelihood function of this bivariate distribution involving all 7 parameters, and estimate all parameters together from the original data. The first approach (which we have been using) is less difficult, but the second, full likelihood approach might produce better parameter estimates. We compared the errors in the parameter estimates from both approaches for sample sizes between 50 and 500 using Monte Carlo simulations.

For the full likelihood estimation, we a solver available within S-Plus was used to find the point in seven dimensions where the full likelihood function achieves its max-

imum. When performing maximization in seven dimensions, one might encounter serious difficulties if the likelihood function is not convex. In our simulations experiments, this occurred for fewer than 1% of the samples, and these samples were excluded from the analysis. It is worth noting that in most of the cases when the seven-dimensional likelihood did not converge, the MIX1 method for one of the marginals did not converge, either.

The parameter estimation error was defined for our comparison as the Euclidean distance between the vector of estimated parameters and the true parameter vector in the seven-dimensional parameter space. Table 3.4 reports the RMSEs of the parameter estimates for selected sample sizes and values of k. It also provides RMSEs of the Gumbel copula parameter δ estimates. It is evident from Table 3.4 that there is very little difference between the results of the two methods. Similar results were also obtained using Gaussian marginals.

- For a Gumbel copula with GEV marginals, it appears that there is no significant benefit to using the full likelihood method to estimate copula and marginals parameters together, rather than separately estimating the marginals parameters and the copula parameters.

In summary, we have observed that

- The method for the estimation of the marginal (either parametric or empirical) parameters has little impact on the error in the copula parameter estimates.

Sample size	Shape	RMSE of 7 parameters		RMSE of δ	
n	k	Separate	Full	Separate	Full
50	0	0.452	0.443	0.301	0.305
50	-0.1	0.453	0.446	0.31	0.315
50	-0.2	0.464	0.462	0.302	0.319
50	-0.3	0.468	0.474	0.306	0.325
50	-0.4	0.469	0.474	0.294	0.313
50	-0.5	0.464	0.475	0.275	0.294
100	0	0.312	0.304	0.212	0.216
100	-0.1	0.309	0.301	0.202	0.204
100	-0.2	0.313	0.306	0.203	0.205
100	-0.3	0.327	0.320	0.214	0.218
100	-0.4	0.333	0.331	0.205	0.212
100	-0.5	0.333	0.331	0.207	0.212
500	0	0.135	0.130	0.094	0.092
500	-0.1	0.138	0.134	0.095	0.095
500	-0.2	0.139	0.134	0.09	0.09
500	-0.3	0.142	0.137	0.092	0.091
500	-0.4	0.154	0.144	0.093	0.089
500	-0.5	0.159	0.148	0.095	0.092

Table 3.4: Comparison of the performance of the full likelihood estimation method (Full) to the separate estimation of marginals and copula parameters (Separate). There is little difference between the two methods' parameter estimates.

- Estimation of the marginals of the bivariate distribution separately from the copula performs approximately as well as using full likelihood method (for distributions with Gumbel copula and GEV or Gaussian marginals).

3.7 Extension of Gumbel Copulas to Include a Third Dimension

We observe a set of n independent observations, $\{(u_1, v_1, w_1), (u_2, v_2, w_2), \ldots, (u_n, v_n, w_n)\}$, of three random variables U, V, and W, each having a uniform distribution between 0 and 1. We are interested in constructing a model for the joint distribution of U, V, and W, fit such a model to the observations, and later simulate random samples from it. Among other things, the model should also be able to adequately describe the bivariate marginal distributions for (U, V), (V, W), and (W, U). Assume that U, V, and W correspond to maxima of some other random variables, so that it is reasonable to model their joint distribution with an EV copula. Moreover, we will focus on extensions of the Gumbel copula, which is both an EV and Archimedean copula. It should be noted here that the Hüsler and Reiss copula does allow extension to a third dimension with different bivariate margins, but simulation of random observations from this copula can be computationally inefficient because of its complicated algebraic form. The Gumbel copula, on the other hand, has a simple algebraic form, but its direct extension to a third dimension requires that two of the three bivariate margins are the same. We introduce a different extension of the Gumbel copula that can describe a wide range of dependencies.

We denote the two-dimensional Gumbel copula from (3.4) as $C^{(\mathrm{G2})}(u,v)$. The generalization of the Gumbel copula to three dimensions is [Joe, 1997, p. 157]

$$C^{(\mathrm{G3})}(u,v,w) = \exp\left\{-\left(\left[(-\log u)^{\beta} + (-\log v)^{\beta}\right]^{\alpha/\beta} + (-\log w)^{\alpha}\right)^{1/\alpha}\right\}, \quad (3.13)$$

where $\beta \geq \alpha \geq 1$ are two parameters. To refer to Gumbel copulas with a specific set of parameters, we will write $C_{\delta}^{(\mathrm{G2})}$ for two-dimensional copula and $C_{\alpha,\beta}^{(\mathrm{G3})}$ for three-dimensional copula.

Assume that the dependence of U, V, and W is described by copula $C_{\alpha,\beta}^{(\mathrm{G3})}$. The bivariate distribution of U and V is then given by the two-dimensional Gumbel copula with parameter β, $C_{\beta}^{(\mathrm{G2})}$, and the bivariate distributions of V and W, and W and U are both described by a Gumbel copula with parameter α, $C_{\alpha}^{(\mathrm{G2})}$. Since $\beta \geq \alpha$, the dependence between U and V is larger than that between either V and W or W and U. This does not constitute a problem, however, since we can always reassign our model random variable names to reflect the dependence we observe in the data. A serious problem is created by the fact that two out of the three bivariate margins of $C_{\alpha,\beta}^{(\mathrm{G3})}$ are the same. It is likely that a real world data that we want to model will not possess this property. For example, in application to flood peaks, the degree of dependence between annual flood peak data might be a function of distances between stream gauging station, so if the three distance are different, all three bivariate margins should be also different. It is also worth noting that the three dimensional extension of the Galambos copula has this same property (see [Joe, 1997, p. 186]). This is the

reason that we are interested in finding alternative extensions of the bivariate Gumbel copula to three dimensions.

3.7.1 Model Formulation and Properties

Here, we introduce a three-parameter modification of (3.13) that allows three different bivariate EV margins, study the dependence properties of the new copula. To begin, let's consider the following two-parameter bivariate copula

$$C^{(\mathrm{E2})}(x,y) = \exp\left\{-\frac{1}{2}\left(\left[(-\log x)^{\delta_1} + (-\log y)^{\delta_1}\right]^{1/\delta_1} + \left[(-\log x)^{\delta_2} + (-\log y)^{\delta_2}\right]^{1/\delta_2}\right)\right\}, \tag{3.14}$$

where $\delta_1 \geq \delta_2 \geq 1$ are the two parameters. Certainly, if $\delta_1 = \delta_2$, copula $C^{(\mathrm{E2})}$ becomes the Gumbel copula $C^{(\mathrm{G2})}_{\delta_1}$. If $\delta_1 \neq \delta_2$, $C^{(\mathrm{E2})}$ is still an EV copula, with dependence function $A^{(\mathrm{E2})}(t) = \frac{1}{2}A^{(\mathrm{G2})}_{\delta_1}(t) + \frac{1}{2}A^{(\mathrm{G2})}_{\delta_2}(t)$, where $A^{(\mathrm{G2})}_{\delta}(t)$ represents the dependence function of $C^{(\mathrm{G2})}_{\delta}$, a two dimensional Gumbel copula with parameter δ.

From the properties of copulas ordering (see page 92) we see that

$$C^{(\mathrm{G2})}_{\delta_1} \prec C^{(\mathrm{E2})}_{\delta_1,\delta_2} \prec C^{(\mathrm{G2})}_{\delta_2},$$

and

$$C^{(\mathrm{E2})}_{\alpha,\delta} \prec C^{(\mathrm{E2})}_{\beta,\delta}, \qquad \text{and} \qquad C^{(\mathrm{E2})}_{\delta,\alpha} \prec C^{(\mathrm{E2})}_{\delta,\beta}, \tag{3.15}$$

for $\alpha \leq \beta$. This means that the copula $C^{(\mathrm{E2})}$ can describe a wide range of dependencies. Hereafter, we will write $C^{(\mathrm{E2})}_{\delta_1,\delta_2}$ to refer to a copula $C^{(\mathrm{E2})}$ with specific parameters δ_1 and δ_2.

To illustrate that the copula $C^{(\mathrm{E2})}$ can be very close to a Gumbel copula for $\delta_1 \neq \delta_2$, we plot the dependence functions of $C^{(\mathrm{E2})}_{3.34,1.5}$ and $C^{(\mathrm{G2})}_{2.0}$ in Figure 3.9. Figure 3.10 shows the contour levels of these two copulas. It very difficult to distinguish between the two copulas. The L_1 norm of the difference between functions $A^{(\mathrm{E2})}_{3.34,1.5}$ and $A^{(\mathrm{G2})}_{2.0}$ is less than 0.005, which is less than a quarter of the L_1 norm of the error of the estimators of A for samples of size 100 [Capéraà *et al.*, 1997a]. In contrast, Kendall's τ for $C^{(\mathrm{G2})}_{3.24}$ is 0.7, while it is 0.33 for $C^{(\mathrm{G2})}_{1.5}$, so by changing just one of the two parameters of $C^{(\mathrm{E2})}_{3.34,1.5}$, we can create copulas with values of Kendall's τ ranging from 0.33 to 0.7.

Another property of the copula $C^{(\mathrm{E2})}$ is that similar dependencies can be described with different sets of parameters. Though this feature is not desirable at all when talking about modeling random variables in two dimensions, we will see later how it is important for three dimensional modeling.

Copula $C^{(\mathrm{E2})}$ has a straightforward interpretation: If (U_1, V_1) and (U_2, V_2) are two pairs of uniform random variables, having Gumbel copulas with parameters δ_1 and δ_2, respectively, (U_1, V_1) is independent of (U_2, V_2), and $\tilde{U} = (\max\{U_1, U_2\})^2$ and $\tilde{V} = (\max\{V_1, V_2\})^2$, the random variables $\tilde{U}$ and $\tilde{V}$ have copula $C^{(\mathrm{E2})}_{\delta_1,\delta_2}$. This property allows us to easily simulate random samples from copula $C^{(\mathrm{E2})}$, if we can already simulate random samples from the Gumbel copula.

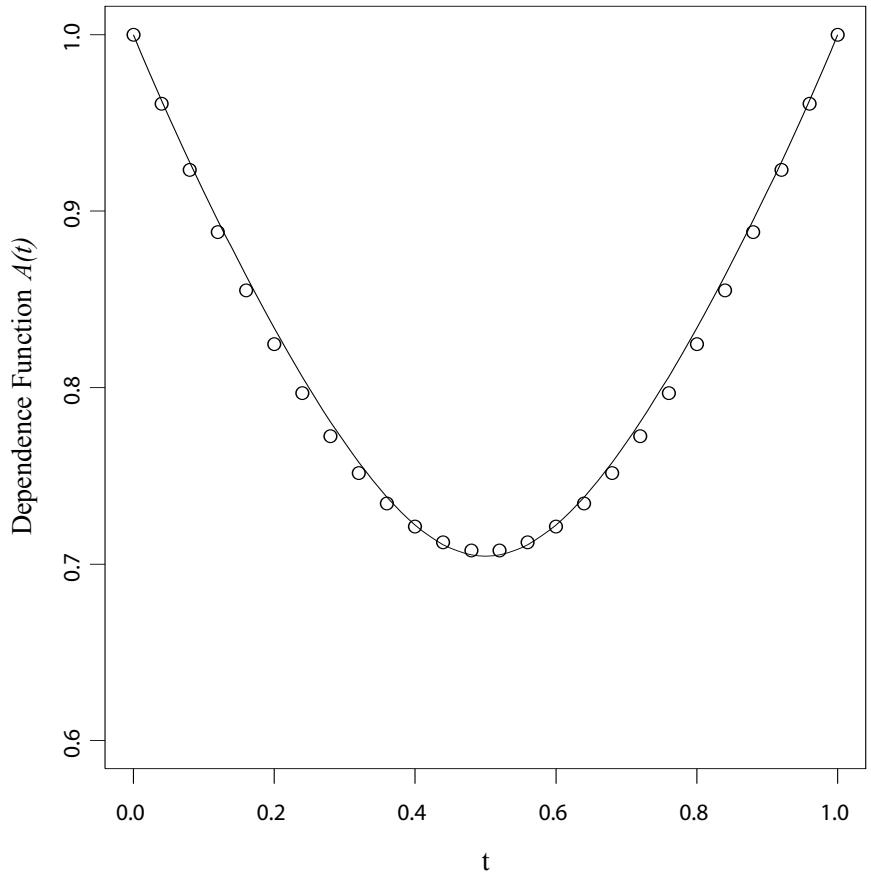

Figure 3.9: Dependence functions of copulas $C^{(\mathrm{E2})}_{3.34,1.5}$ (curve) and $C^{(\mathrm{G2})}_{2.0}$ (circles)

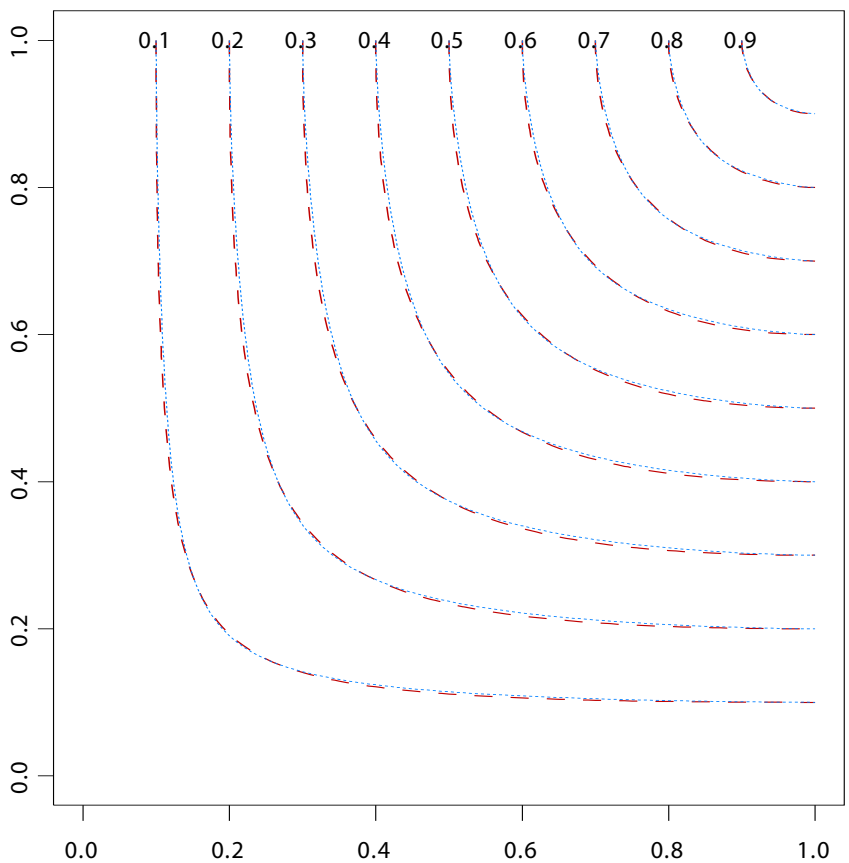

Figure 3.10: Contour plots of copulas $C^{(\mathrm{E2})}_{3.34,1.5}$ (dotted lines) and $C^{(\mathrm{G2})}_{2.0}$ (dashed lines). The difference between the two functions is very small.

Now consider a three dimensional copula $C^{(\mathrm{E3})}$ that has bivariate margins $C^{(\mathrm{E2})}$. We will show that this copula can describe different marginal dependencies in all three bivariate margins.

Copula $C^{(\mathrm{E3})}$ has the following form:

$$C^{(\mathrm{E3})}(u,v,w) = \exp\left\{-\frac{1}{2}\left(\left[(-\log u)^{\beta} + (-\log v)^{\beta}\right]^{\alpha/\beta} + (-\log w)^{\alpha}\right)^{1/\alpha}\right.$$
$$\left. -\frac{1}{2}\left(\left[(-\log w)^{\gamma} + (-\log u)^{\gamma}\right]^{\alpha/\gamma} + (-\log v)^{\alpha}\right)^{1/\alpha}\right\}, \quad (3.16)$$

where $\beta \geq \alpha \geq 1$ and $\gamma \geq \alpha$ are the three parameters. When $\beta = \gamma = \alpha$, this copula is equivalent to the three dimensional Gumbel copula with equal parameters. $C^{(\mathrm{E3})}$ has the following bivariate margins: $(U,V) \sim C^{(\mathrm{E2})}_{\beta,\alpha} = C_{(U,V)}$, $(V,W) \sim C^{(\mathrm{E2})}_{\alpha,\alpha} = C^{(\mathrm{G2})}_{\alpha} = C_{(V,W)}$, and $(W,U) \sim C^{(\mathrm{E2})}_{\gamma,\alpha} = C_{(W,U)}$. Also, since $\alpha \leq \gamma$, and $\alpha \leq \beta$, we have $C_{(V,W)} \prec C_{(U,V)}$ and $C_{(V,W)} \prec C_{(W,U)}$. Depending on which parameter β or γ is greater, we will have either $C_{(U,V)} \prec C_{(W,U)}$ or $C_{(U,V)} \succ C_{(W,U)}$. In any case, the dependence ordering of bivariate margins is precisely specified by the copula parameters. The parameters also have a straightforward interpretation: α determines the smallest dependence between the two of the three random variables, β specifies how much more dependence there is between U and V than between V and W, and γ acts similarly for W and U.

Since $C^{(\mathrm{E2})}$ can describe a wide range of dependencies, the same thing is true for $C^{(\mathrm{E3})}$. It is important to note that this copula will be close to the three dimensional Gumbel copula only when the difference between the three bivariate marginals is

not very large, for instance, if the differences in the marginals' Kendall's taus are not more than 0.5. This is due to the fact that copula $C^{(\mathrm{E2})}$ is very different from the Gumbel copula if $\delta_1 \gg \delta_2$ in (3.14), especially in the patterns generated by the sample observations (Figure 3.11). In fact, these patterns become similar to the ones observed for the Tawn copula in section 4.5. If using the Gumbel copula for modeling a trivariate distribution is desirable, and the dependence between one pair (V, W) is much less than that of the other two pairs, no models based on extensions of the Gumbel copula or the Archimedean copula would work well, since those models require two identical bivariate marginals have smaller dependence than the third bivariate margin. In such a case, it might be easier to see whether an assumption of the independence of (V, W) is appropriate. If two of the three bivariate marginals imply a very small dependence, which occurs if one of the three variables does not significantly depend on the other two, one might want to use a regular $C^{(\mathrm{G3})}$ copula. Copula $C^{(\mathrm{E3})}$ might be a good choice if all of the bivariate marginals have a similar but not exactly equal dependence, for example, when the values of Kendall's tau for the bivariate marginals are 0.27, 0.35 and 5.0.

The relationship of copula $C^{(\mathrm{E3})}$ to the three dimensional Gumbel copula is simple: if (U_1, V_1, W_1) and (U_2, V_2, W_2) are two triplets of uniform random variables having $C^{(\mathrm{G3})}$ copulas with parameters (β, α) and (γ, α), respectively, (U_1, V_1, W_1) is independent of (U_2, V_2, W_2), and $\tilde{U} = (\max\{U_1, V_2\})^2$, $\tilde{V} = (\max\{V_1, W_2\})^2$, and $\tilde{W} = (\max\{W_1, U_2\})$, the random variables $\tilde{U}, \tilde{V}$, and $\tilde{W}$ have the copula $C^{(\mathrm{E3})}_{\alpha,\beta,\gamma}$.

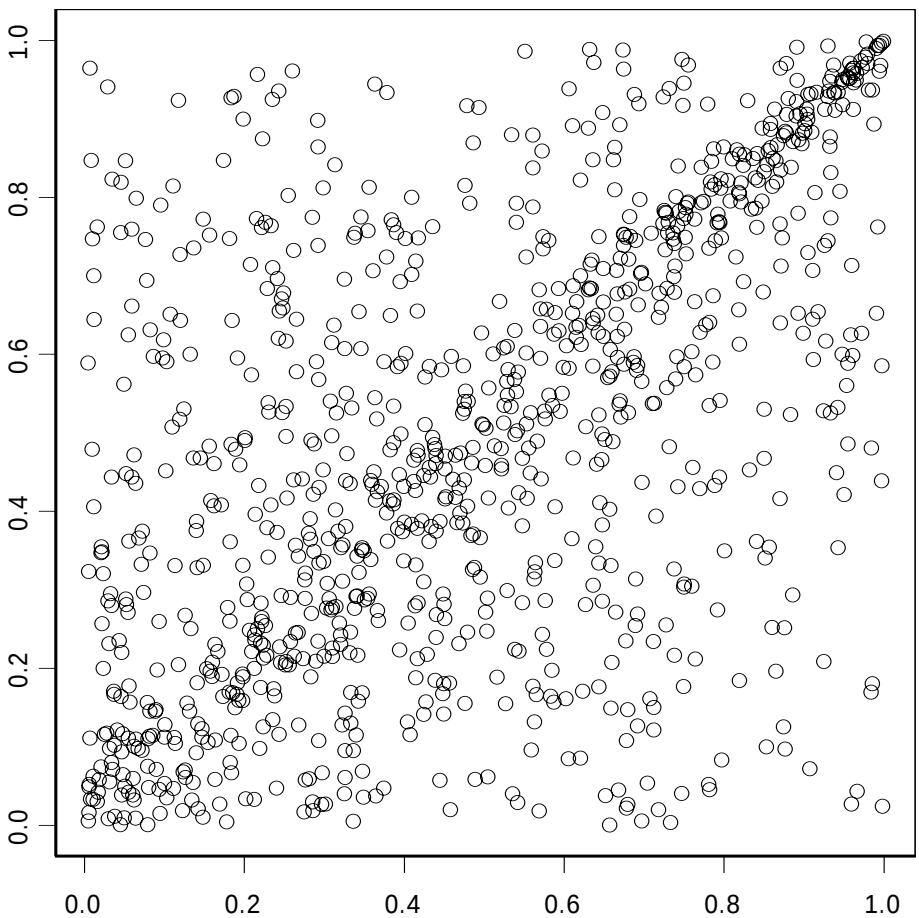

Figure 3.11: 1000 point simulated from the $C^{(\mathrm{E2})}$ copula with $\delta_1 = 8$ and $\delta_2 = 1.1$ ($\delta_1 \gg \delta_2$).

This property allows us to easily simulate random samples from the copula $C^{(\mathrm{E3})}$, if we can already simulate random samples from the three-dimensional Gumbel copula.

To simulate random variables from the trivariate Gumbel copula, we implemented a procedure that numerically computes the inverse of the conditional distribution $C^{(\mathrm{G3})}(w|u,v)$:

$$C^{(\mathrm{G3})}_{\alpha,\beta}(w|u,v) = \frac{\frac{\partial^2}{\partial u \partial v} C^{(\mathrm{G3})}_{\alpha,\beta}(u,v,w)}{C^{(\mathrm{G2})}_{\beta}(u,v)}.$$

This conditional distribution has a complex algebraic form, and we do not express it here. After simulating a pair (U,V) from copula $C^{(\mathrm{G2})}_{\beta}$, we then set $W = {C^{(\mathrm{G3})}_{\alpha,\beta}}^{[-1]}(S|U,V)$, where S is a uniform $(0,1)$ random variable independent of U and V.

We also implemented a maximum likelihood estimation procedure that estimates the parameters of both $C^{(\mathrm{G3})}$ and $C^{(\mathrm{E3})}$ by maximizing the likelihood function in three dimensions. Parameters of the $C^{(\mathrm{E3})}$ copula can be estimated either using this procedure or by fitting each bivariate marginal distribution independently. Indeed, by fitting $C^{(\mathrm{G2})}$ to observations of V and W, we can easily obtain an estimate of α. We can, then, estimate β by fitting $C^{(\mathrm{E2})}_{\beta,\alpha}$ to observations of U and V, and then estimate γ by fitting $C^{(\mathrm{E2})}_{\gamma,\alpha}$ to observations of W and U. Simulation experiments show that the second approach generally provides better parameter estimates.

3.8 Summary of EVANESCE features

The capabilities of EVANESCE for bivariate data analysis include:

- Functions for evaluating copula distributions and densities for 17 parametric copula families.
- Generation of random observations from these copulas.
- Analytical computation of Kendall's tau and Spearman's rho for the parametric families.
- Construction of empirical copulas from data points.
- Computation of empirical Kendall's tau and Spearman's rho for data.
- Maximum Likelihood parameter estimation procedure for all parametric copula families available with the library.
- Additional non-parametric copula estimators due to Capéraà *et al.* [1997a] and Genest and Rivest [1993].
- Visualization functions for plotting copula contour levels, densities, etc.

The library harnesses the object-oriented capabilities of S-Plus, and is designed in such a way that adding new copula families is very easy. For example, to add a new EV copula family, one must simply provide a formula for its dependence function and its two derivatives. All pertinent methods from this library will then be available for the new copula.

The library also includes a set a functions useful in univariate extreme value analysis. Not only does it implement all of the methods described in the previous

chapter, but also provides a semi-parametric estimator for marginal distributions based on the Peaks-Over-Threshold model described in section 4.2.

Furthermore, as explained in section 3.7, we implemented an extension of a Gumbel copula to the third dimension with relevant parameter estimation and random sample generation for three dimensional models. EVANESCE also has several useful functions specifically designed for financial engineers, such as an analytic computation of the risk measure VaR and and a function returning holidays and closings of the New York Stock exchange.

The EVANESCE library is written in the S-Plus scripting language. Since any `for` or `while` loops are evaluated very slowly in S-Plus (see Venables and Ripley [1997, p. 158]), the entire library is written without them. When using a `for` statement was inevitable, it was placed in an external C code, and only one call from S-Plus to a C code is made when needed.

The next two chapters of this dissertation illustrate applications of the methods described in this chapter to analyses of several data sets of interest to financial engineers and hydrologists.

Chapter 4

Financial Engineering Applications of Tails and Copula Estimations

In this chapter, we illustrate how methods based on copula estimation can be used to enhance financial engineering data analysis.

Before proceeding, it is important to draw attention to the different parameterizations of the Generalized Extreme Value (2.1) and Generalized Pareto (2.11) distributions in financial engineering and hydrology applications. In financial applications, it is common to replace k in (2.1) and (2.11) with $-\xi$, so that *positive* values of the shape parameter ξ correspond to heavy tails in the distribution [Embrechts *et al.*, 2000a; McNeil and Saladin, 1997]. For example, the following form of the GP cumulative distribution function

$$F_{\alpha,\delta,\xi}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-\delta}{\alpha}\right)^{-1/\xi} & \xi \neq 0, \\ 1 - \exp\left\{-\frac{x-\delta}{\alpha}\right\} & \xi = 0, \end{cases} \tag{4.1}$$

is traditional in the financial engineering applications. One of the options of the EVANESCE library is to set a boolean parameter `SHAPE.XI` to `TRUE` it the latter parameterization should be used. Since this chapter presents examples from financial engineering, we will change k to $-\xi$ and use the GP distribution in form (4.1) for the duration of this chapter. We consider only distributions with heavy tails in this dissertation, and changes of notation from chapter to chapter should not be a cause for confusion.

4.1 Utility Stocks Indexes

The following example illustrates that the assumption that two samples have a bivariate Gaussian distribution does not necessarily result in a good estimate of the joint distribution function, even for financial data. We have at our disposal the daily values of indexes reflecting the stock prices of electric and gas companies from January 1, 1992 to May 30, 2000 [Sales, 2001]. Figure 4.1 shows the time series of each index. Obviously these series are not stationary. To make the sequences stationary, it is customary to consider the difference between the logarithms of a stock index (log-returns) on two consecutive days. We will denote a random variable representing the log of a daily return for the electric index as X (dataset `UTI.ELEC`), and a similar random variable for the gas index as Y (dataset `UTI.GAS`). Figure 4.2 shows the time sequence of log-returns for each data set, and Figure 4.3 shows the log-returns for the gas and electric indexes plotted against one another. Some correlation between the

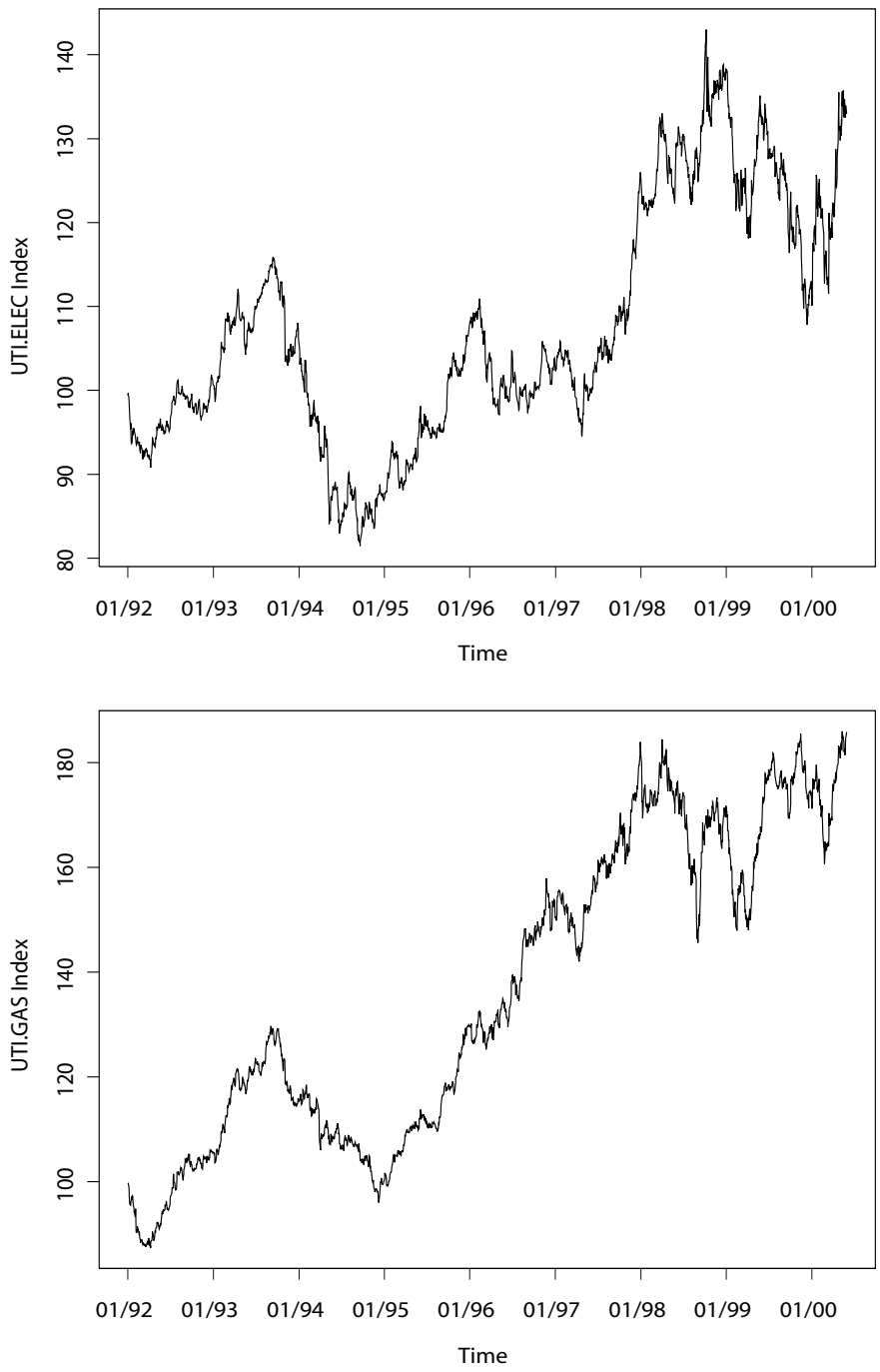

Figure 4.1: Time series of indexes reflecting the stock prices of electric (top) and gas (bottom) companies.

log-returns is seen, which is not surprising, as both indexes represent utility companies. We are interested in estimating the joint probability distribution of X and Y, meaning that we want to find probabilities of the form

$$F(x, y) = I\!P\{X \leq x, Y \leq y\}$$

for any x and y. The tails are usually the most challenging part of the distribution to work with, and we will be estimating the above probability when it is either small (≤ 0.1) or large (≥ 0.9). In particular, we intend to estimate this probability "beyond the data", i. e., when it is smaller than $1/n$ (or larger than $1 - 1/n$), where n is the number of data points in the sample.

The marginals of $F(x, y)$ will be denoted as $G_1(x)$ and $G_2(y)$. Assuming that $F(x, y)$ is continuous, $F(x, y) = C(G_1^{-1}(x), G_2^{-1}(y))$ for some copula $C(u, v)$. In order to estimate the function F, we will estimate the functions G_1, G_2, and the the copula C.

According to the Samuelson's approach in which stocks are modeled as a geometric Brownian motion, it is reasonable to assume that the log-returns have a jointly Gaussian distribution with mean vector a and variance matrix b. The estimated values for our dataset are

$$a = 10^{-4} \begin{pmatrix} 1.34 \\ 2.92 \end{pmatrix} \qquad \text{and} \qquad b = 10^{-5} \begin{pmatrix} 5.78 & 2.85 \\ 2.85 & 2.92 \end{pmatrix}.$$

If we were to work with the assumption that X and Y have a jointly Gaussian bivariate distribution we would use the estimator $\hat{F}_{\text{Gaussn.}}(x, y)$ given by the jointly Gaussian

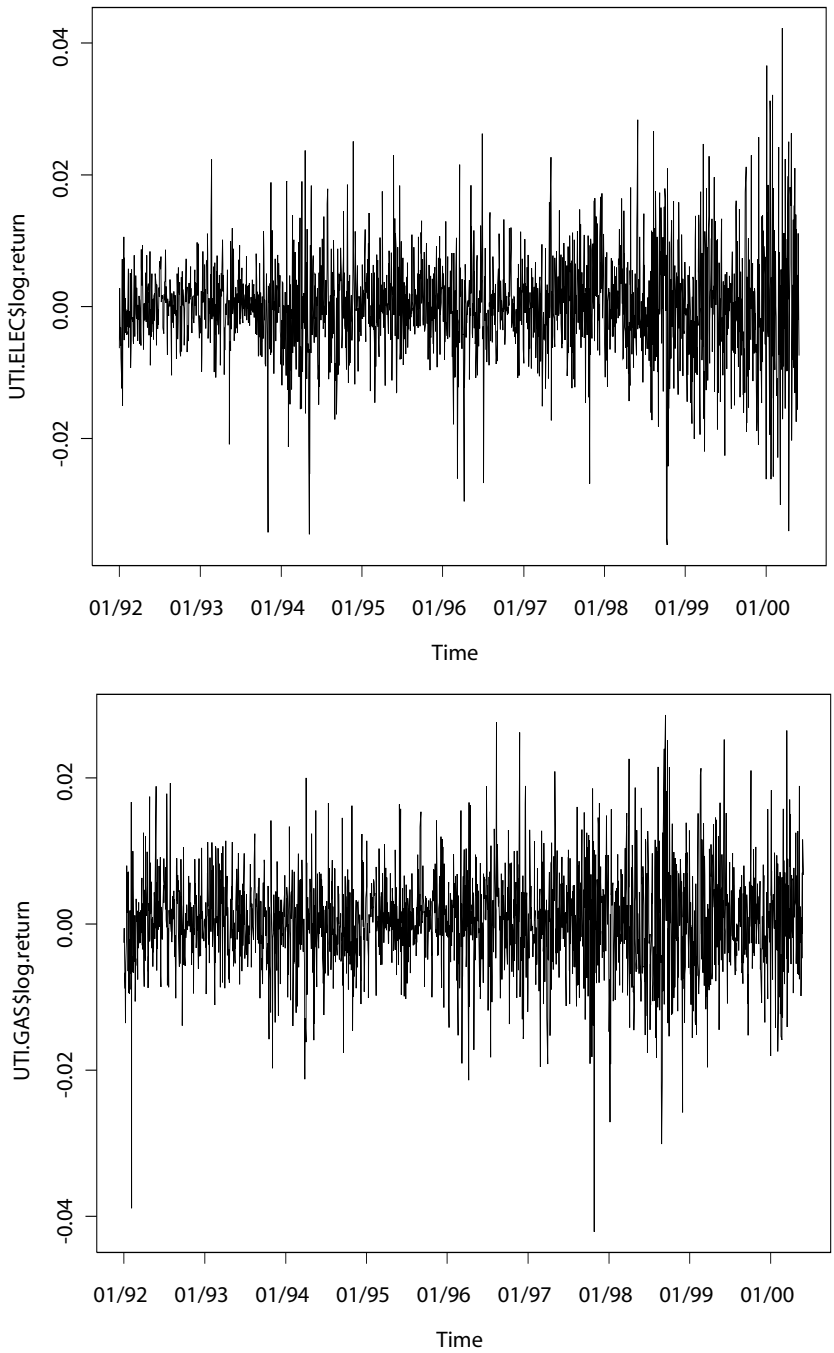

Figure 4.2: Time series of the daily log-returns of two indexes UTI.ELEC (top) and UTI.GAS (bottom).

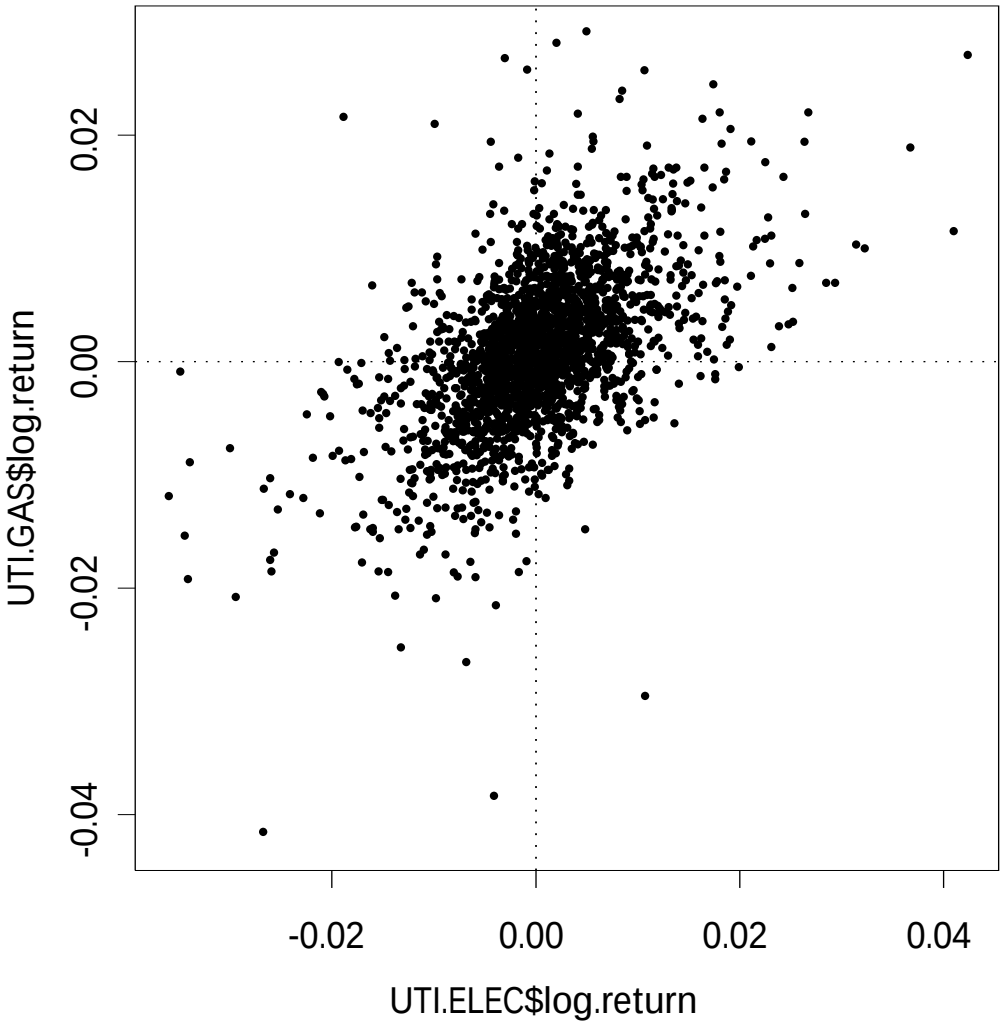

Figure 4.3: Dependence between the log-returns of the electric and gas company indexes.

distribution with parameters a and b given above. We can also estimate $F(x, y)$ from the data as

$$\hat{F}_{\text{data est.}}(x, y) = \frac{\text{Card}\{i : x_i \leq x, y_i \leq y\}}{n},$$

and compare it to $\hat{F}_{\text{Gaussn.}}(x, y)$ (Table 4.1). To understand why the estimates from the jointly Gaussian distribution disagree with the data, first observe from Table 4.1 that the jointly Gaussian distribution estimate of $F(-0.015, -0.015)$ is 0.0031, but the estimate from the data is 0.0071. The sample size n of our data is $n = 2125$, and the data contains 15 points where both coordinates less than -0.015. If $\mathbb{P}\{X \leq -0.015, Y \leq -0.015\}$ was 0.0031, the probability that we would observe 15 points with both coordinates less than -0.015 in a sample of size 2125 is less than 0.003. If we repeat this comparison at other points, we arrive at the same conclusion: the tails of the bivariate jointly Gaussian distribution are not heavy enough to describe our data. To see that with standard univariate tools, we may compare the tails of the marginal distributions of X and Y to the tails of the univariate Gaussian distribution separately. Quantile-quantile plots (QQ-plots) for the log-returns against quantiles of a Gaussian distribution are shown in Figure 4.4. Notice that these plots deviate strongly from straight lines, which indicates the non-Gaussian nature of both series of log-returns.

The next question to address is whether the source of the disagreement with the jointly Gaussian distribution is the marginals of our data (see Figure 4.4), the copula associated with our data, or both.

	$\mathbb{P}\{X \leq x, Y \leq y\}$				$\mathbb{P}\{X > x, Y > y\}$			
x	-0.005	-0.015	-0.02	-0.01	0.005	0.015	0.02	0.01
y	-0.005	-0.015	-0.01	-0.025	0.005	0.015	0.01	0.025
Data Estimates	0.0979	0.0071	0.0066	0.0009	0.0941	0.0075	0.0061	0.0009
Gaussian approx.	0.1155	0.0031	0.0022	0.0001	0.1292	0.0039	0.0026	0.0001
$\hat{F}_{emp}$	0.0974	0.0069	0.0066	0.0009	0.0939	0.0071	0.0057	0.0005
Non-Gauss. margins, Gauss. copula	0.1854	0.0190	0.0106	0.0021	0.2012	0.0213	0.0131	0.0019
Fitted copula	0.0936	0.0080	0.0072	0.0020	0.0938	0.0081	0.0077	0.0014
Bootstrap	0.0929	0.0073	0.0069	0.0017	0.0919	0.0072	0.0068	0.0011

Table 4.1: Comparison of $F(x, y)$ estimates from the data with estimates obtained using different methods.

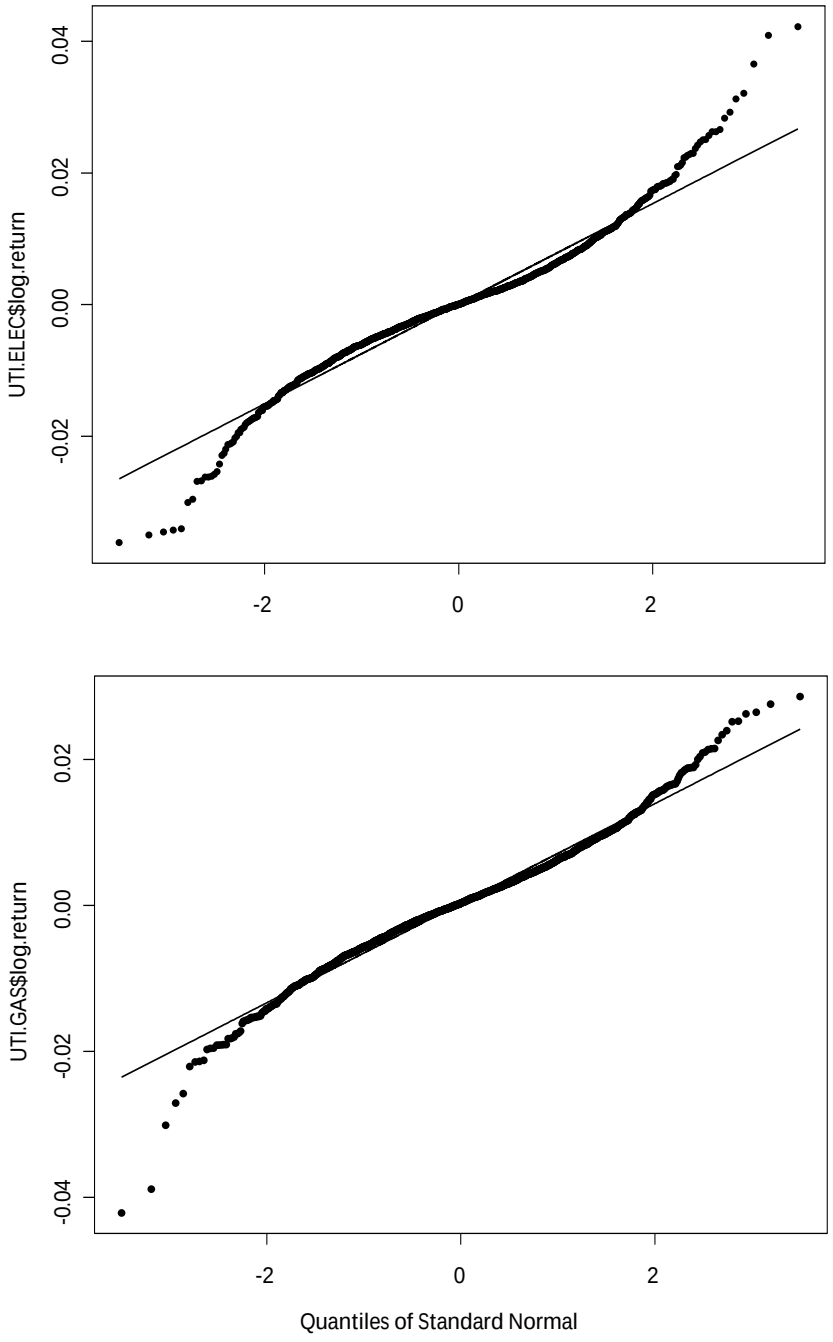

Figure 4.4: Gaussian quantile-quantile plots for the log-returns.

4.2 Peaks-Over-Threshold Marginal Estimate

To proceed, we estimate the marginal distributions of X and Y separately. EVANESCE implements a semi-parametric estimate of the cumulative distribution and corresponding quantile functions based on the Peaks-Over-Threshold method. This method is based on the idea that, if $G_1(x)$ is in the domain of attraction of the Extreme Value distribution with shape parameter ξ_1, the distribution of excesses over a high threshold u can be closely approximated by a Generalized Pareto Distribution (GPD) with the same shape parameter (ξ_1):

$$\mathbb{P}\left\{X \leq s+u | X > u\right\} = \frac{G_1(u+s) - G_1(u)}{1 - G_1(u)} \sim F_{a_1,u,\xi_1}(s),$$

where F_{a_1,u,k_1} is a GPD distribution function with scale parameter a_1, location parameter u, and shape parameter ξ_1 (see (4.1)). In this example, we are interested in modeling both the upper and lower tails of G_1, so we also approximate $\mathbb{P}\left\{X \geq -t-v | X < -v\right\}$ by $F_{a_2,v,\xi_2}(t)$. Given an upper threshold u and a lower threshold $-v$, the parameters of the tail GPD distributions are estimated from the points in the sample that are either less than $-v$ or greater than u, $G_1(x)$ is approximated by an empirical cdf for $-v < x \leq u$ and a appropriate GPDs for $x > u$ and $x \leq -v$:

$$\hat{G}_1(x) = \begin{cases} \frac{N_2}{n}\left(1 + \frac{\hat{\xi}_2(x+v)}{\hat{a}_2}\right)^{-1/\hat{\xi}_2} & \text{if } x \leq -v, \\ \frac{i-0.5}{n} & \text{if } x_{(i)} \leq x < x_{(i+1)} \text{ and } -v < x \leq u, \\ 1 - \frac{N_1}{n}\left(1 + \frac{\hat{\xi}_1(x-u)}{\hat{a}_1}\right)^{-1/\hat{\xi}_1} & \text{if } y > u, \end{cases} \tag{4.2}$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is the order statistics, $N_1 = \max\{i : x_{(i)} \leq u\}$, and $N_2 = \min\{i : x_{(i)} > -v\}$. Not only is this estimate implemented in EVANESCE function `cdf.estimate.2tails`, but also the function has an optional argument that toggles on linear interpolation of $\hat{G}(x)$ for $-v < x \leq u$. Estimates of ξ_1, a_1, ξ_2, and a_2 are calculated using either the method of L-moments or MLE (see Hosking and Wallis [1987] for a comparison of these estimators for the GPD).

EVANESCE also implements single-tail versions of the above estimator for cases where we want to approximate only one tail of the marginal distribution of X (for example, when $X \geq 0$).

In order to estimate the two-tailed cumulative distribution function of X, we must determine the upper and lower thresholds that we will use in (4.2) to estimate tails. To that end, we can use the function `shape` written by Alexander McNeil [McNeil and Saladin, 1997], which plots an estimate of the shape parameter versus the value of the upper threshold (the top axis on the plots), to choose the threshold value that we want to use. The general rule in this case is that we want the shape plot to be rather flat in the vicinity of a good threshold, but we don't want the number of exceedances above the threshold to be too low (indicated on the bottom axis of the plot). Since we want to estimate both the upper and lower tails, we produced (Figure 4.5) the shape plots for $\{x_1, x_2, \ldots, x_n\}$, which lets us to determine the appropriate threshold of the upper tail, and the shape plots for $\{-x_1, -x_2, \ldots, -x_n\}$ to determine the threshold for the lower tail. We choose a 0.005 threshold for the upper tail and

a -0.005 threshold for the lower tail for both X and Y. A detailed review of the methods for choosing the best threshold can be found in Lang *et al.* [1999], but it is believed that there is no good algorithm that can be generally applied for doing so [Embrechts *et al.*, 2000a].

The estimates of the shape parameter for the tails of both electric and gas indexes are around zero. This means that both G_1 and G_2 are in the domain of attraction of a Gumbel distribution, but this does not necessarily mean that they have Gaussian tails; the exponential distribution, for example, is also in the domain of attraction of the Gumbel distribution. Quantile-quantile plots of the tail quantiles against exponential quantiles appear to be quite linear, so we adopt the hypothesis that the tails of G_1 and G_2 have an exponential decay. Figure 4.6 shows fits of the estimated marginal distributions to the data tails.

Using the marginal estimates, we construct paired observations of (U, V), where $U = G_1(X)$ and $V = G_2(Y)$. We then set $u_i = \hat{G}_1(x_i)$, and $v_i = \hat{G}_2(y_i)$, $i = 1, 2, \ldots, n$.

4.3 Copula Estimation

From the transformed observations $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)$, we can estimate the copula of U and V (and therefore of X and Y) empirically (using (3.10)), and estimate F (the joint distribution function of X and Y) with the empirical estimates $\hat{C}_{\text{emp}}$, $\hat{G}_1$

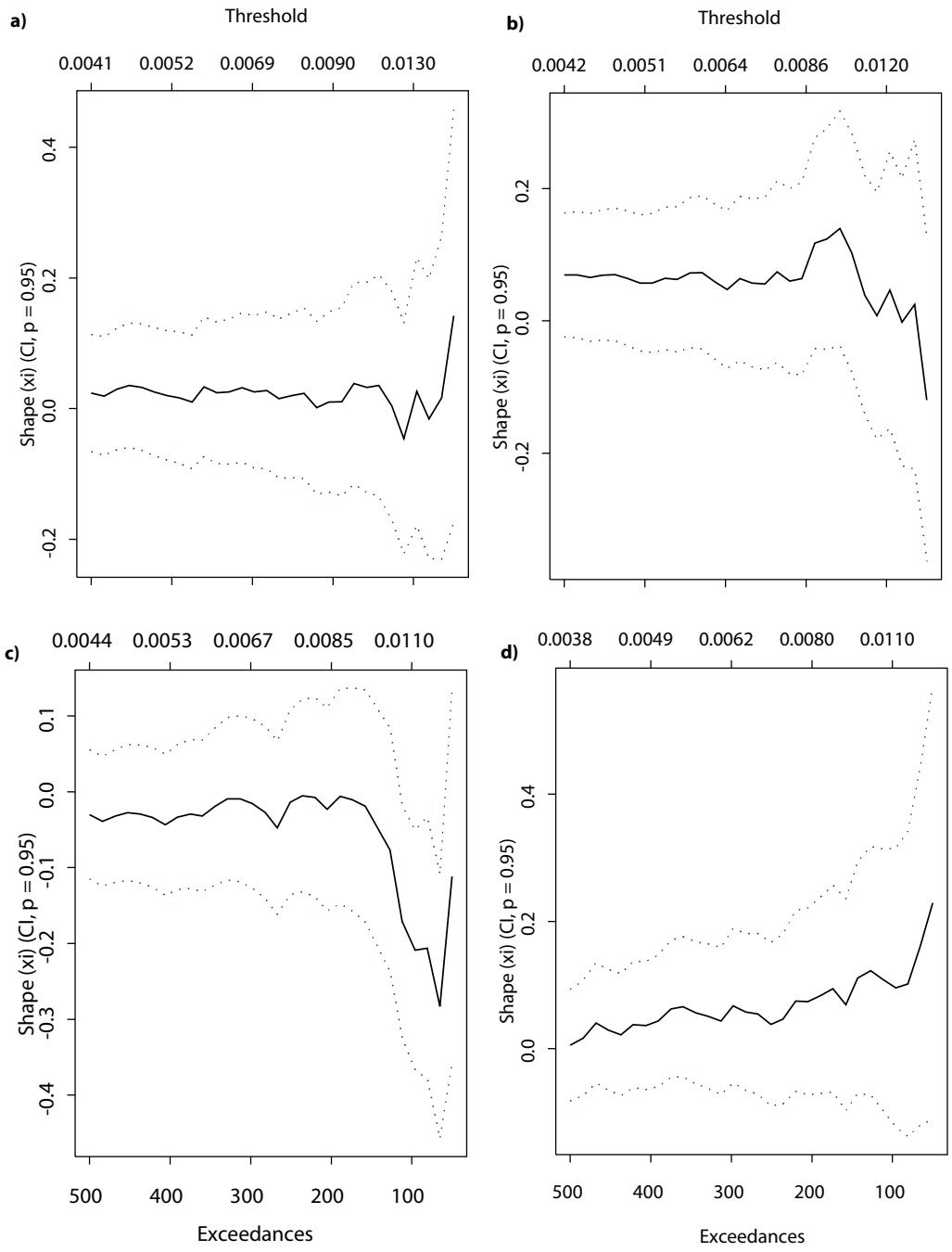

Figure 4.5: Plot of estimates of the shape parameter of the GPD distribution as a function of threshold. a) UTI.ELEC, upper tail; b) UTI.ELEC, lower tail; c) UTI.GAS, upper tail; d) UTI.GAS, lower tail.

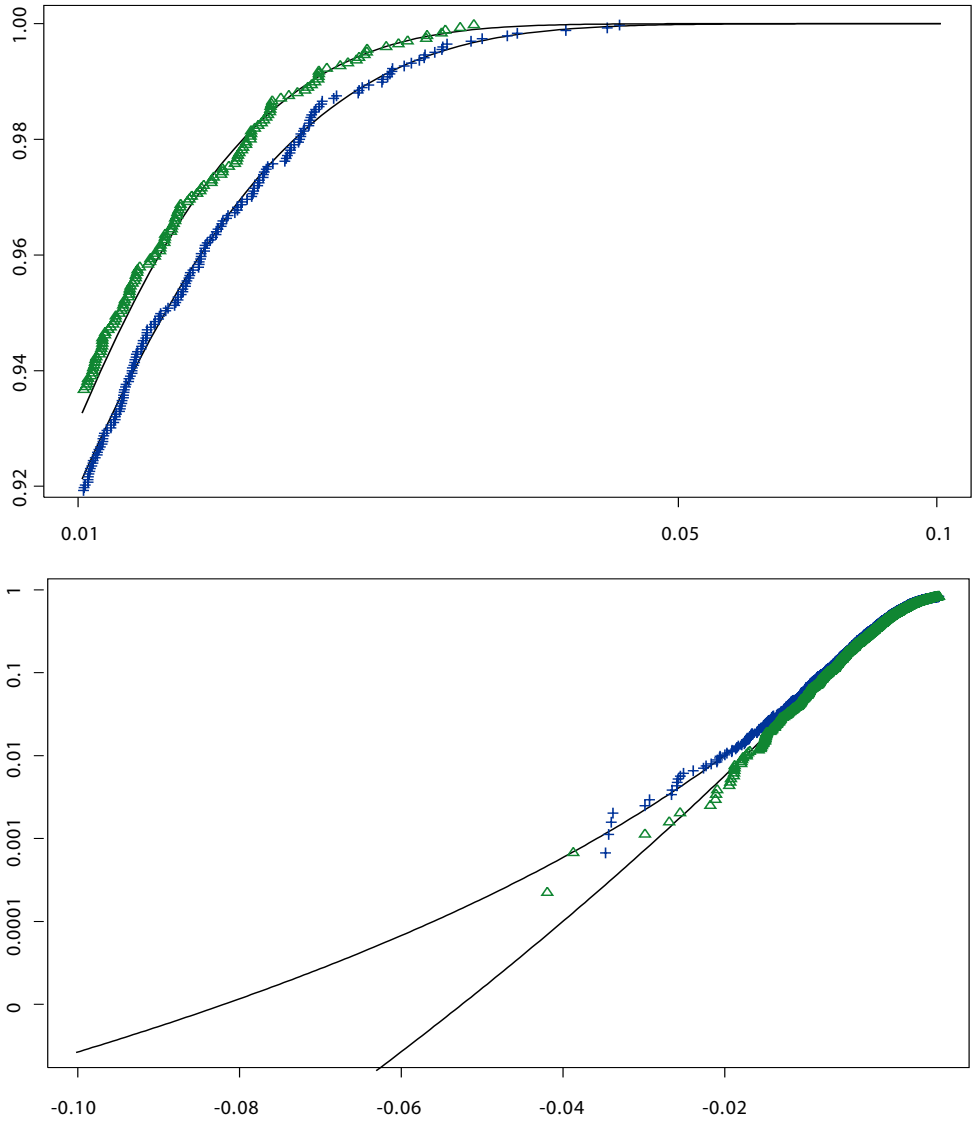

Figure 4.6: Fit of the estimated marginal distributions into upper (top) and lower (bottom) data tails. The empirical cdf of the `UTI.ELEC` log-returns is shown with crosses, and that for `UTI.ELEC` is shown with triangles.

and $\hat{G}_2$ using the formula:

$$\hat{F}_{emp}(x, y) = \hat{C}_{emp}(\hat{G}_1(x), \hat{G}_2(y))$$

This formula is implemented in EVANESCE by the function called `jointcdf.empest.2tails`. Values of the $\hat{F}_{emp}(x, y)$ for our data at selected points are reported in Table 4.1. We see that the agreement with the data estimate of F is better than for the Gaussian approximation.

Unfortunately, this non-parametric estimation does not allow us to generate random variables with a joint distribution similar to X and Y. The ability to produce simulated variables is necessary for the use of bootstrapping techniques and risk analysis. In particular, we will later see how the Monte Carlo method might help us to evaluate the Value-at-Risk (VaR) of a linear portfolio. Using the non-parametric estimation of F also abandons any attempt to characterize the tail index of F.

Our next step, then, is fitting different families of parametric copulas into the data. We estimated parameters for all 15 parametric copulas described in section 3.2 for the data. Based on a Pearson χ^2 goodness of fit test (see section 3.5), we had to reject most of the parametric families at the 0.0001 significance level. The rejected families included all EV copulas, Kimeldorf and Sampson, Joe, and BB6 copula. The Frank copula was rejected at the 0.05 significance level. This result should not be surprising, since none of these copulas have lower-tail dependence. From observing Figure 4.3 one might suspect that neither the upper tails, nor the lower tails are independent.

It is interesting to determine whether the Gaussian copula is a good approximation to the data copula. If it is, it will mean that, the joint distribution of X and Y is simply a monotonically transformed bivariate Gaussian, even though the marginals of X and Y are not Gaussian. The MLE estimate of the parameter δ for the Gaussian copula is 0.558, which is decent, since the correlation coefficient computed directly from the original data is 0.549. Table 4.1 also reports estimates of F based on a Gaussian copula with semi-parametric estimated margins, which are much higher than the values implied by the data. Gaussian copula have a zero upper and lower tail dependence parameter, but the convergence in the tail index expression in (3.8) and (3.9) is very slow [Borlot *et al.*, 2000], which explains the fact that we could not reject the Gaussian copula immediately. In order to reproduce the data tail, Gaussian copula must allocate too much of the probability density to the tails, resulting in a significant overestimation of the probabilities of interest.

Estimated parameters for the four remaining parametric copula families and their implied values of Kendall's tau, Spearman's rho, and the upper and lower tail indexes are reported in Table 4.2. Also reported is the L_2 norm of the difference between the empirically estimated copula and the fitted parametric copulas. The BB1 and BB4 families result in approximately the same estimates of all dependence measures. Kendall's tau and Spearman's rho for the data ($\tau_{\text{data}} = 0.380$, and $\rho_{\text{data}} = 0.537$) are very close to those implied by all of the fitted copula, with the BB7 copula giving the worse values. The lower tail index for the BB3 copula is always 1, unless there

is no upper tail dependence, which doesn't seem likely for our data. Since the BB3 copula cannot represent any arbitrary dependence in the lower tail, we exclude this copula from consideration. It is evident from Table 4.2 that the fitted copulas BB1 and BB4 are effectively the same, and are somewhat different from BB7 estimate. Since both BB1 and BB7 copulas are Archimedean copulas, we plotted the Genest and Rivest [1993] non-parametric estimate for the function $\lambda(t) = \phi(t)/\phi'(t)$, along with the $\lambda(t)$ for the fitted BB1 and BB7 copulas ($\lambda(t)$ is virtually the same for BB1 and BB4 copulas) in Figure 4.7. It can be seen from this figure that $\lambda(t)$ for the non-parametric estimate is closer to that of the BB1 copula than to the $\lambda(t)$ of the BB7 copula. In addition, both Kendall's tau and Spearman's rho for the BB7 copula are farther from those estimated from the data than for the BB1 copula. Accordingly, we decided to use the BB1 copula family to model our data. The results below would not change significantly should we have chosen the BB7 copula (even for risk measures), however, as all three of these copulas describe the data sufficiently well to model it. Figure 4.8 shows the contour plots of the empirical copula and the fitted BB1 copula, which are very close.

The estimate of F using a fitted parametric copula $\hat{C}_{\text{param}}$ and the estimated marginal quantile functions is then

$$\hat{F}_{\text{param}} = \hat{C}_{\text{param}}(\hat{G}_1(x), \hat{G}_2(y)). \tag{4.3}$$

The values of $\hat{F}_{\text{param}}$ for selected points are also reported in Table 4.1. These estimates are in good agreement with the data.

Copula	Estimates		Kendall's	Spearman's	Tail Index		$L_2(\Delta C)$
	θ	δ	tau	rho	Lower	Upper	$\times 10^6$
BB1	0.433	1.302	0.368	0.522	0.292	0.297	7.582
BB3	1.261	0.482	0.365	0.516	1.000	0.267	13.234
BB4	0.436	0.559	0.367	0.522	0.292	0.289	8.402
BB7	1.370	0.699	0.353	0.500	0.371	0.341	22.342

Table 4.2: Parametric copulas that provide a good fit to the utility index data. The Kendall's tau and Spearman's rho estimates for the data are $\tau_{\text{data}} = 0.380$, and $\rho_{\text{data}} = 0.537$. $L_2(\Delta C)$ is the L_2 norm of the difference between the empirical copula and the fitted parametric copula.

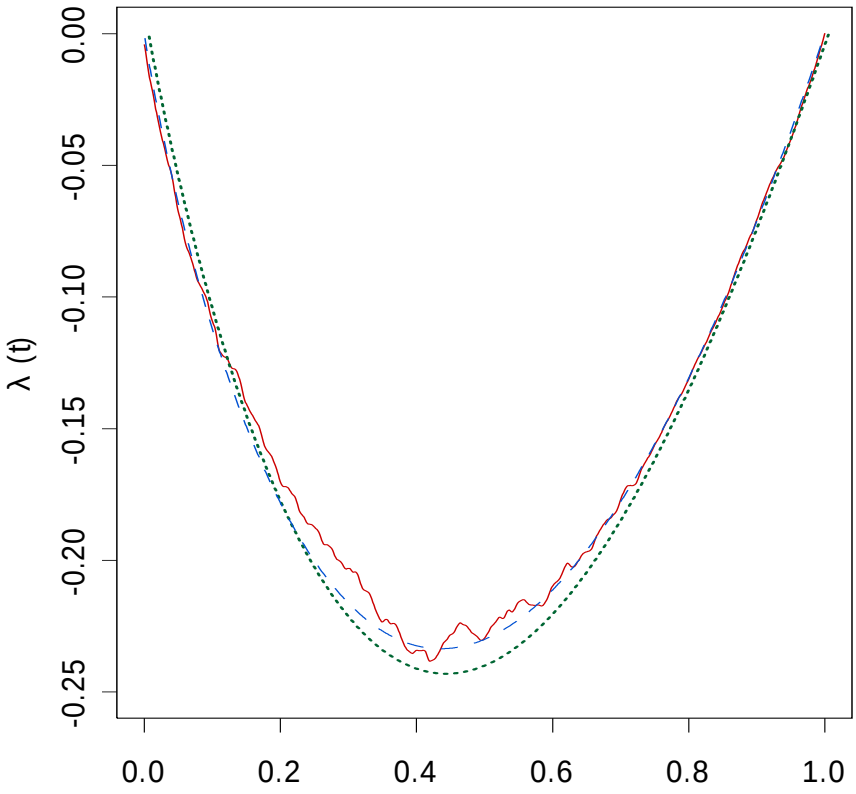

Figure 4.7: Non-parametric estimate of the $\lambda(t) = \phi(t)/\phi'(t)$ function for Archimedean copulas (solid line), and $\lambda(t)$ for the BB1 (dashed line) and BB7 (dotted line) copulas with parameters given in Table 4.2.

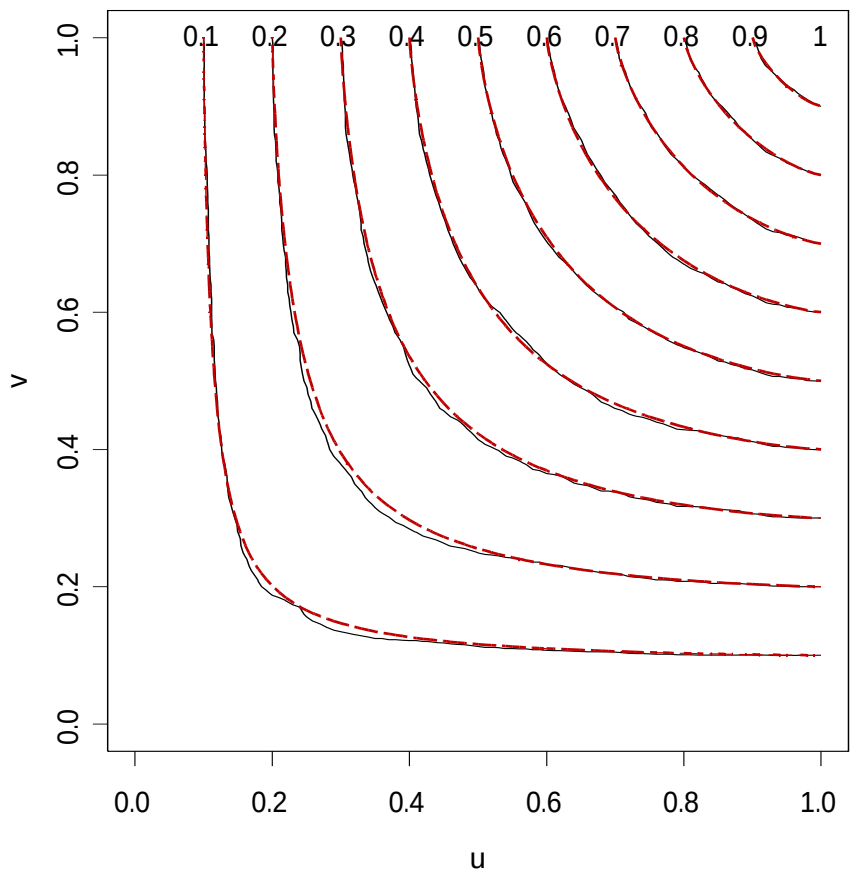

Figure 4.8: Contour plot showing the fit of the BB1 copula (dashed lines) to the empirical copula of our data (solid lines).

We can now simulate random pairs $(\tilde{X}, \tilde{Y})$ from the estimated joint distribution $\hat{F}_{\text{param}}$ by generating a random pair $(\tilde{U}, \tilde{V})$ from the estimated BB1 copula and setting $\tilde{X} = G_1^{-1}(\tilde{U})$ and $\tilde{Y} = G_2^{-1}(\tilde{V})$. (As seen from the form of (4.2), the inverse of the marginal estimates can be readily computed.) Figure 4.9 shows a plot of $10,000$ points generated in this way, together with the observed points, and the estimate of the joint probability density function of X and Y computed by differentiating (4.3) and using the kernel density estimator for the derivative of the second line in the rhs of (4.2). The distribution of simulated points is close to that of the observed X and Y (there are approximately 5 times more simulated points in Figure 4.9 than observed points). Table 4.1 also contains estimates of F computed from the fitted copula's simulated samples (based on 50,000 simulated samples). Since these estimates (the "Bootstrap" values in Table 4.1) match the data very well for points not in the tail, the estimation of the tail probabilities can be considered good.

4.4 Computing Measures of Risk

In applications of this type of analysis to risk management, several measures of risk can be computed either directly from the estimated values or through Monte Carlo simulations. We will limit our illustration to the two measures of risk for the daily loss of a linear portfolio: Value-at-Risk (VaR_q) and expected shortfall ($\mathbb{E}[\Theta_q]$). The initial value V_0 of a portfolio comprised of b_1 shares of the stock index of electric

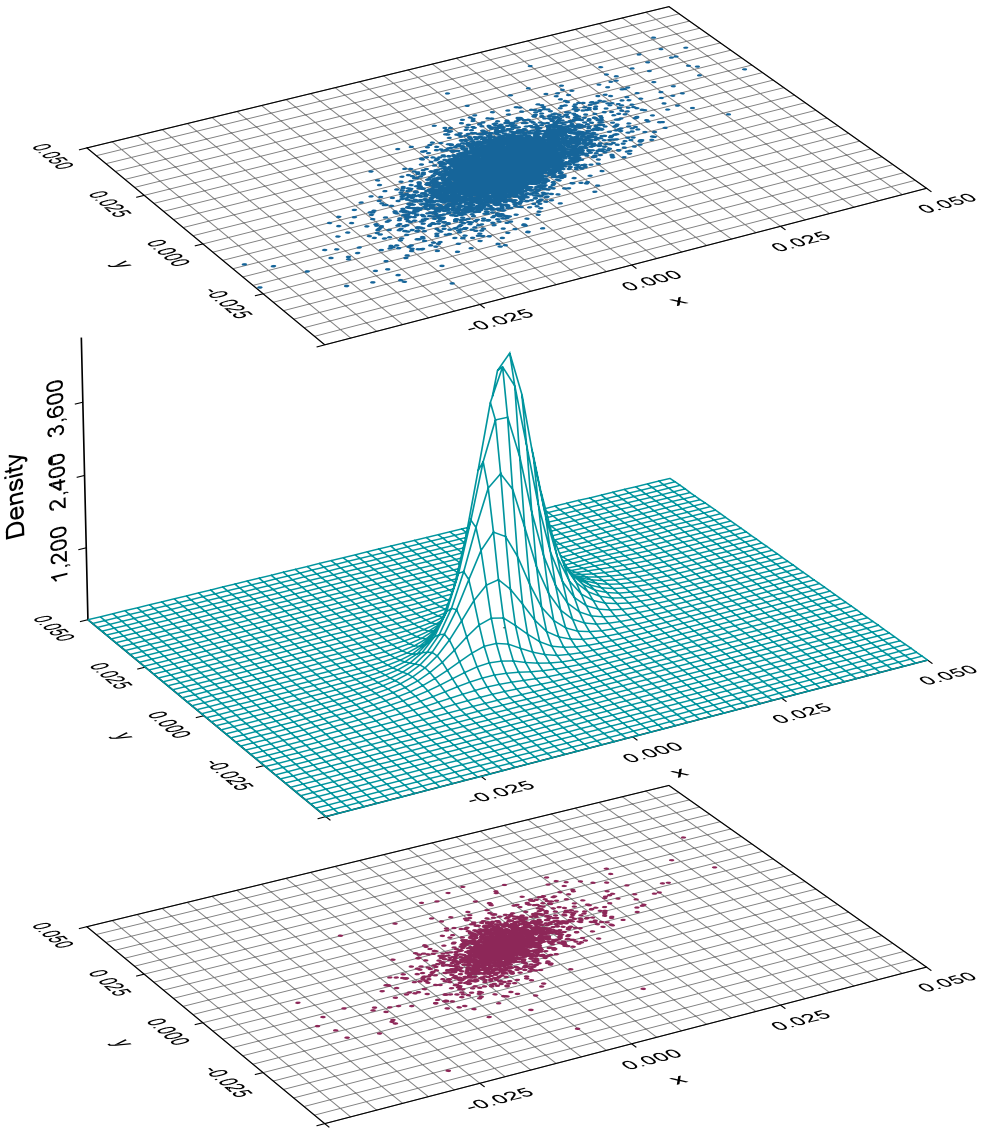

Figure 4.9: Observed data points (bottom), estimate of the joint density function (middle), and 10,000 simulated points from that density function (top).

companies S_1 and b_2 shares of the stock index of gas companies S_2 is given by

$$V_0 = b_1 S_1 + b_2 S_2.$$

At the end of the day, the stock index of electric companies will be $S_1 e^X$, and the stock index of gas companies will be $S_2 e^Y$, where X and Y represent the daily log-returns. The value of the portfolio at the end of the day is

$$V = b_1 S_1 \, e^X + b_2 S_2 \, e^Y,$$

and the log return of the portfolio is

$$R = \log\left(\frac{V}{V_0}\right) = \log\left(\frac{b_1 S_1}{b_1 S_1 + b_2 S_2} \, e^X + \frac{b_2 S_2}{b_1 S_1 + b_2 S_2} \, e^Y\right) = \log\left(\lambda_1 \, e^X + \lambda_2 \, e^Y\right).$$

Here, λ_1 and λ_2 are the fractions of the initial portfolio value invested in the electric and gas company indexes, respectively. These quantities are assumed to be deterministic in this model.

For a given level q, the value at risk VaR_q is defined as the $q \cdot 100$th percentile of the loss distribution. In this example, negative values of R are considered losses, so VaR_q is the solution to

$$q = I\!P\left\{-R \geq VaR_q\right\}. \tag{4.4}$$

Though this measure of risk is very popular (in fact, company regulations often require managers to control the VaR of their exposures), it involves only the probability of the loss and not the size of the loss, and it also does not satisfy some of the desirable properties of a risk measure (for example, it might unduly discourage diversification

[Artzner *et al.*, 1999]). For these reasons, we also consider the expected short fall, $\mathbb{E}[\Theta_q]$:

$$\mathbb{E}[\Theta_q] = \mathbb{E}[-R| - R > VaR_q]. \tag{4.5}$$

Again, let $G_1(x)$ and $G_2(y)$ be marginal distributions of X and Y, let $C(u,v)$ be a copula associated with X and Y, and let $c(u,v)$ be the copula density

$$c(u,v) = \frac{\partial^2}{\partial u \partial v} C(u,v).$$

The joint probability density function of X and Y is then

$$f(x,y) = c(G_1(x), G_2(y))\; g_1(x)\; g_2(y),$$

where $g_1(x) = \frac{d}{dx} G_1(x)$ is the probability density function of X, and $g_2(y) = \frac{d}{dy} G_2(y)$ is the probability density function of Y. Then, for any $r \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\{-R \geq r\} &= \mathbb{P}\left\{\log\left(\lambda_1\, e^X + \lambda_2\, e^Y\right) \leq e^{-r}\right\} \\
&= \int_{-\infty}^{-r-\log\lambda_1} dx \int_{-\infty}^{\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2\, e^x)} dy\; c(G_1(x), G_2(y))\; g_1(x)\; g_2(y) \\
&= \int_{-\infty}^{G_1(-r-\log\lambda_1)} du \int_{-\infty}^{G_2(\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2\, e^{G_1^{-1}(u)}))} dv\; c(u,v) \\
&= \int_{-\infty}^{G_1(-r-\log\lambda_1)} du\; \frac{\partial}{\partial u} C(u,v)\Big|_{v=G_2(\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2\, e^{G_1^{-1}(u)}))}
\end{aligned}$$

Generally speaking, it is not possible to compute the above integral analytically, but EVANESCE includes a function (`fbar.exp.portf`) that calculates it numerically for a given parametric copula and estimated marginals, and the function `VaR.exp.portf` uses `fbar.exp.portf` to solve equation (4.4) for a given q. This function is virtually

instantaneous if the marginals have only moderately heavy tails and the values of q are sufficiently large, since we must simply compute a one-dimensional integral. If the marginal distributions have very heavy tails or q is very small, the integration routine's precision must be increased, but the computation time is still a matter of seconds. The EVANESCE library also includes a function `VaR.lin.portf` that computes VaR for models with $R = \lambda_1 X + \lambda_2 Y$.

Figure 4.10 shows estimates of VaR computed using simulations and theoretical estimates for $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$, with q taking values between 10^{-2} and 10^{-5}. Also shown are estimates obtained directly from the data. It would appear that the model does a good job of estimating VaR_q. Moreover, this plot gives us an idea of the sample size that should be used in our simulations to estimate VaR and other derived risk measures.

The expected short fall can be calculated either by directly computing the expectation in (4.5), or through Monte Carlo simulations. S-Plus has a fast routine for computing one-dimensional integrals, `integrate`, that we used in the calculation of VaR. The `integrate` routine exploits S-Plus' capability to evaluate functions simultaneously at several points. The two-dimensional version of `integrate` doesn't benefit from this capability in the second dimension, which slows the computation. Overall, it take more time to compute the expectation in (4.5) with a relative tolerance of only 10^{-3} than it does to simulate 1,000,000 pairs from the model (see Venables and Ripley [1997], section 4.10 for a detailed discussion of computational issues that arise when

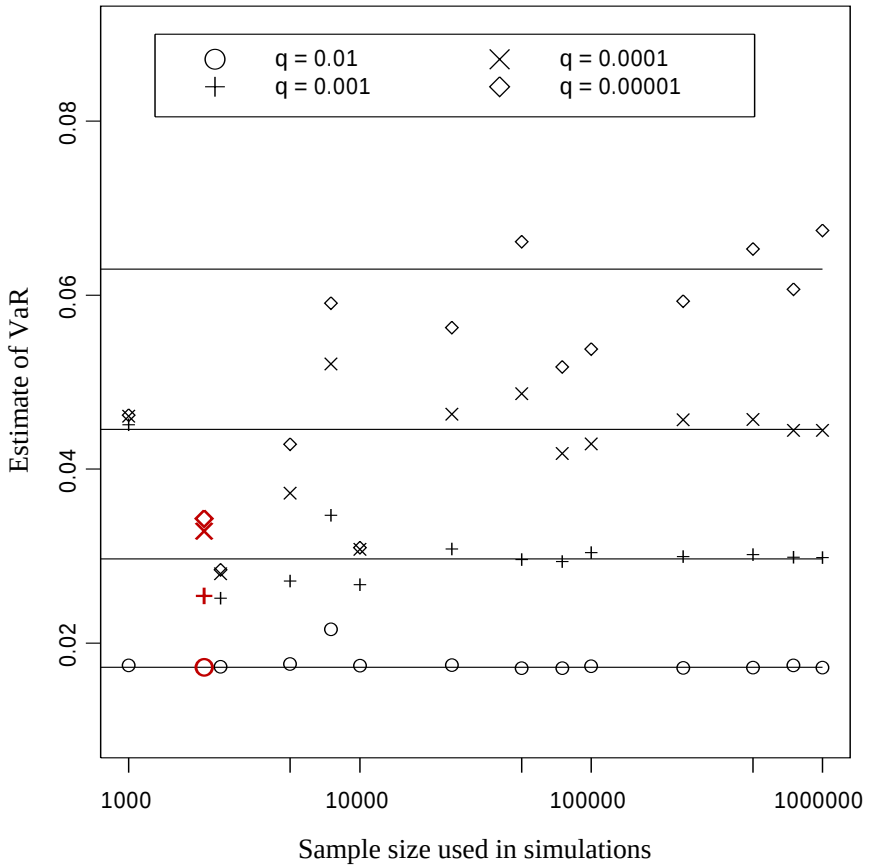

Figure 4.10: Estimates of VaR_q for the utility indexes, obtained using simulations (symbols) and direct integration (lines) for different values of q as functions of the samples size used in the simulations. Bold symbols correspond to the data estimates.

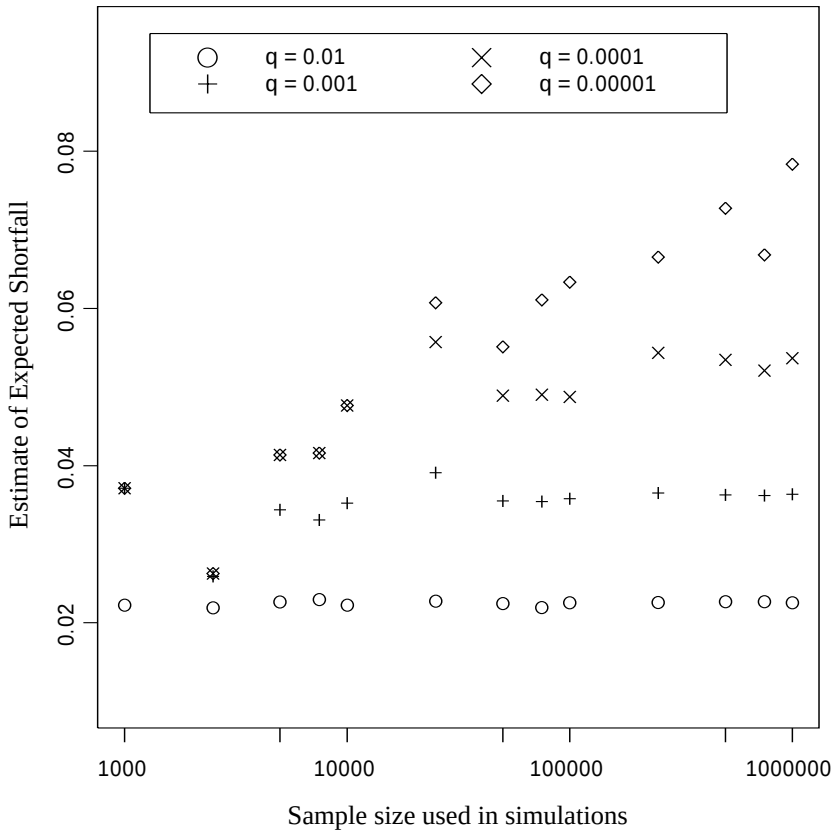

Figure 4.11: Estimates of the expected shortfall for the utility indexes, obtained using simulations for different values of q as functions of the samples size used in the simulations. For $q = 10^{-5}$, good convergence is not seen even for 1,000,000 simulation points.

using S-Plus for statistical modeling). Accordingly, we recommend using simulations to compute the expected short fall. Figure 4.11 shows estimates of expected short fall for $\lambda_1 = 0.5$ and several values of q as functions of the sample size used. This plot suggest that simulation of 1,000,000 points is not enough to estimate the expected short fall for $q = 10^{-5}$. Simulation of 5 and 10 million points suggests an estimate of 0.78 for $\mathbb{E}[\Theta_{10^{-5}}]$.

To conclude, we have demonstrated how the daily time series of the utility stock indexes can be easily analyzed using tools available in the EVANESCE package. Our analysis suggests that the daily log-returns of the indexes of both electric and gas companies have exponential tails. Neither a bivariate Gaussian distribution nor a bivariate Gaussian copula with estimated margins appeared to satisfactorily model our data. There is dependence in both upper and lower tails between the two series, and we can model this dependence with copula. Based on our model, we estimated *VaR* and expected short fall for a linear portfolio.

4.5 Prices of Brazilian and Columbian Coffee

Another interesting illustration of the capabilities of EVANESCE involves the analysis of a time series of Brazilian and Columbian coffee prices, which are plotted in Figure 4.12. Again, we are interested in modeling the dependence of the daily log-returns of the prices. Figure 4.13 shows the price log-returns plotted against one other. We

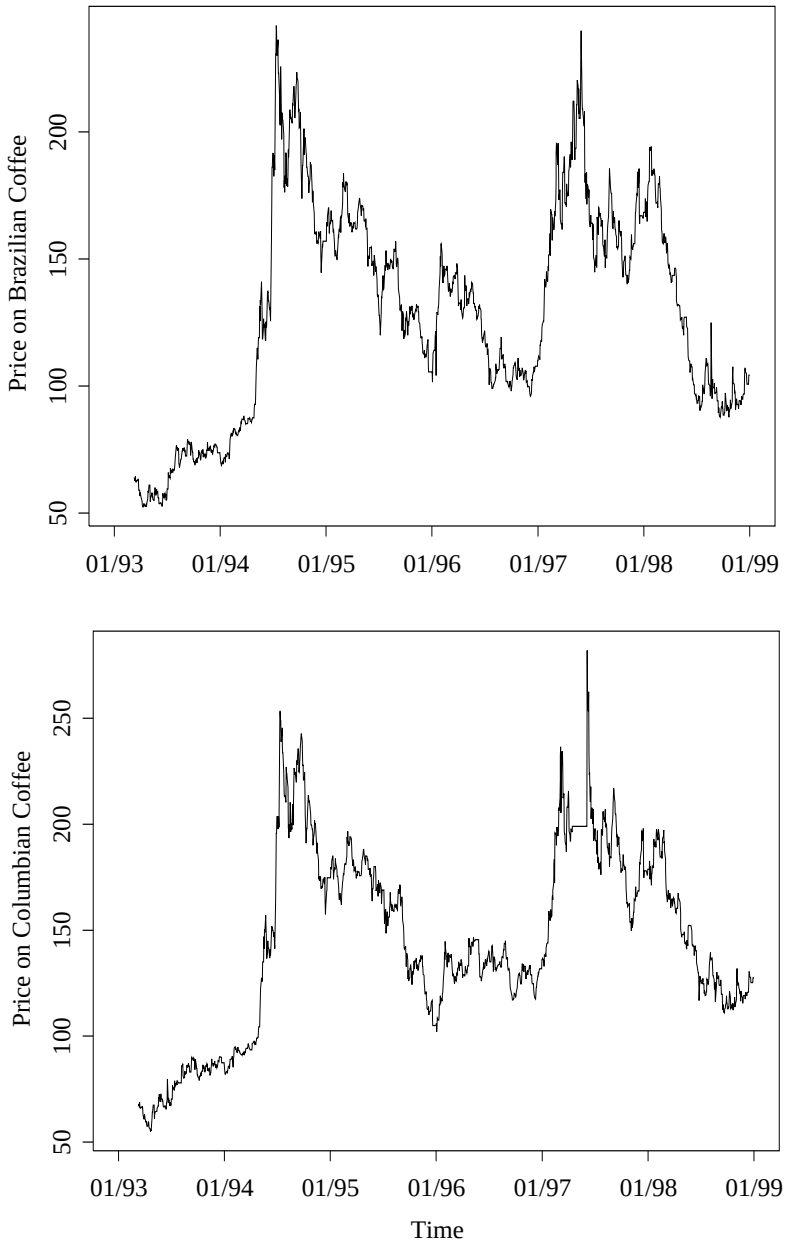

Figure 4.12: Time series of Brazilian and Columbian Coffee prices.

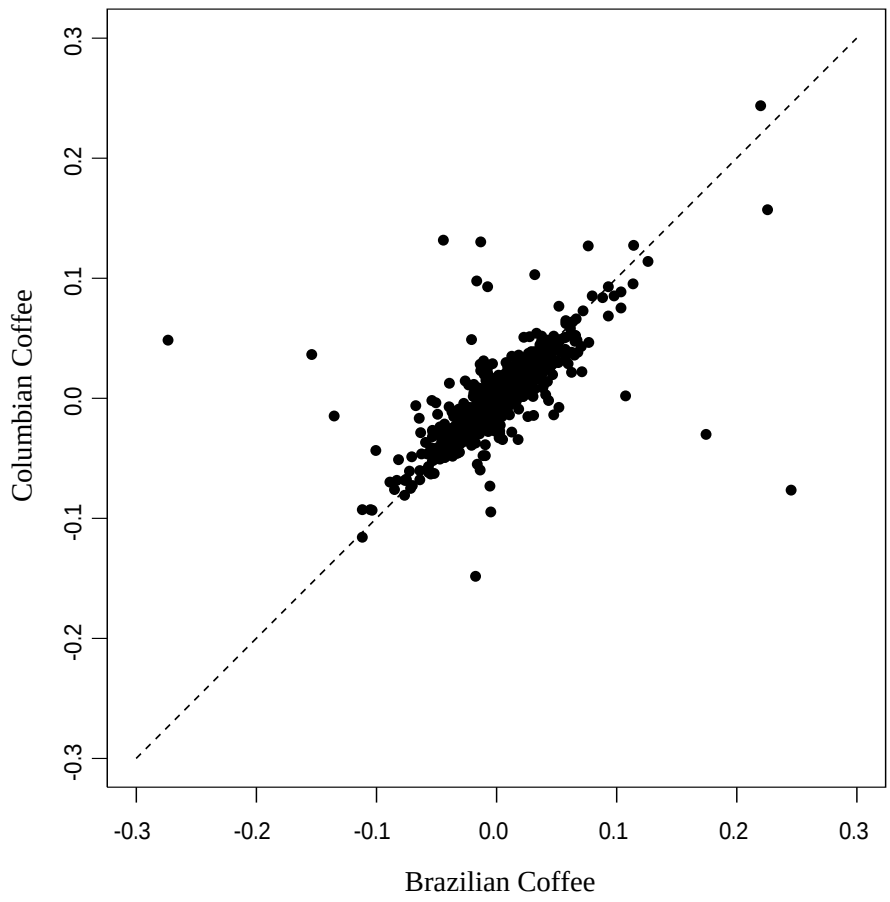

Figure 4.13: Daily log-returns of Brazilian (x axis) and Columbian (y axis) coffee prices.

denote the daily log-return of the Brazilian price by X, and the log-return of the Columbian price by Y.

From Figure 4.13, we notice several important characteristics of the data:

1. Most of the data points in the figure are very close to the $x = y$ line, implying a strong dependence between X and Y. This dependence was expected, since both quantities represent the daily change in price for the same commodity in the neighboring countries.

2. In spite of the strong dependence, there are some two dimensional "outliers", that is, points that are very far from the majority of the data. This phenomenon is observed on smaller scales, as well: in Figure 4.14 a), we plot the enlarged central portion of the data. Figure 4.14 b) shows a similar picture for a sample from a bivariate Gaussian distribution with the same covariance matrix as the data. This suggests that, with some probability, there are days for which the change in one of the prices is much greater than for the other price. We also plot samples generated from different distributions based on Gaussian copula in Figure 4.14. Comparison of plots 4.14 a) and 4.14 d) shows that the peculiar behavior of the data can not be explained solely by a non-Gaussian character in the marginal distributions.

It was noted previously that changes in commodities tend to have heavy tails [Lye, 1991]. Estimates of the tail shape parameter for our data support this observation:

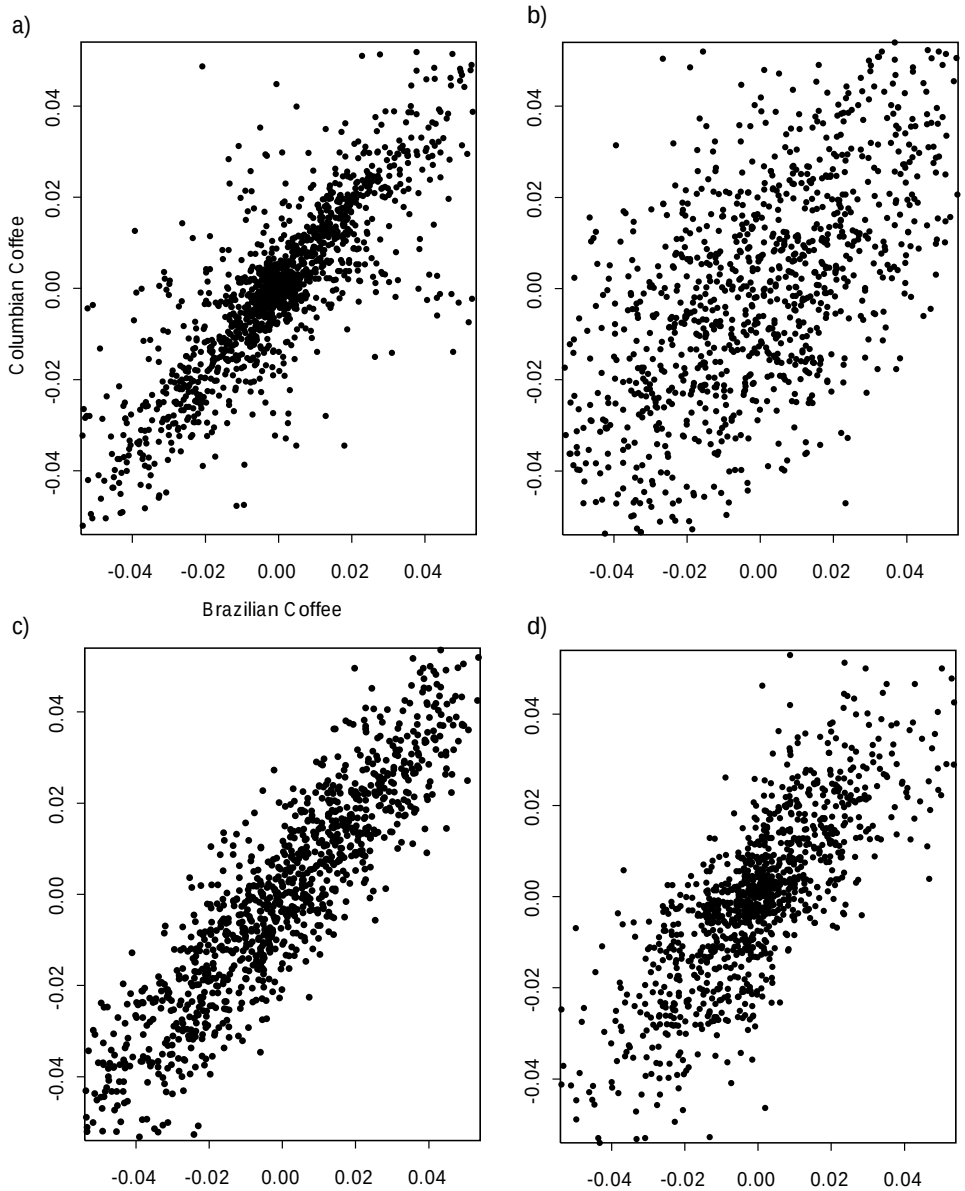

Figure 4.14: Comparison of the dependence of data (a), a sample from a bivariate Gaussian distribution with the same a coefficient of correlation (0.68) and variance as the data. (b), a sample from a bivariate Gaussian distribution with a coefficient of correlation 0.85 and the same marginal variances as the data. (c), a sample from a Gaussian copula with the parameter $\delta = 0.85$ and marginal distributions estimated from data (d).

the estimates were about 0.2 for the lower and upper tails, for both X and Y. After estimating marginal distributions using (4.2), we transformed our data to observations of uniform $(0,1)$ random variables (Figure 4.15). Most of the points lie very close to $u = v$, but some points seem wildly located, as though they were generated from an independent copula. Will we be able to model such a dependence?

From the parametric copula families described in section 3.2, only the Tawn family was not rejected as an acceptable model for our data based on a χ^2 test. The maximum likelihood parameter estimates for the Tawn copula were $\alpha = 0.926$, $\beta = 0.921$, and $r = 5.113$. Since $\alpha \approx \beta$, asymmetry in the Tawn copula is not what makes it describe this data better than the other copulas (all the other copulas described in section 3.2 are symmetric). It is interesting to determine what exactly about the Tawn copula allows it to describe dependencies such as this one well.

It is first helpful to note that the independence (or product) copula $C_{\text{ind}}(u,v) = uv$ can be represented as an EV copula with dependence function $A_{\text{ind}}(t) = 1$ for all $t \in [0,1]$. Also, if (U_j, V_j) have a distribution given by a copula $C_j(u,v)$, $j = 1, 2, \ldots, m$, the pairs (U_j, V_j) are independent, and we set $\tilde{U} = (\max\{U_1, U_2, \ldots, U_m\})^m$, and $\tilde{V} = (\max\{V_1, V_2, \ldots, V_m\})^m$, $\tilde{U}$ and $\tilde{V}$ are then uniformly distributed on $(0,1)$ and have copula $\tilde{C}(u,v) = \prod_{j=1}^{m} C_j(u^{1/m}, v^{1/m})$. If all C_js, $j = 1, 2, \ldots, m$, are EV copulas with dependence functions $A_j(t)$, then $\tilde{C}$ is also an EV copula with dependence function $\tilde{A}(t) = \dfrac{1}{m}\sum_{j=1}^{m} A_j(t)$.

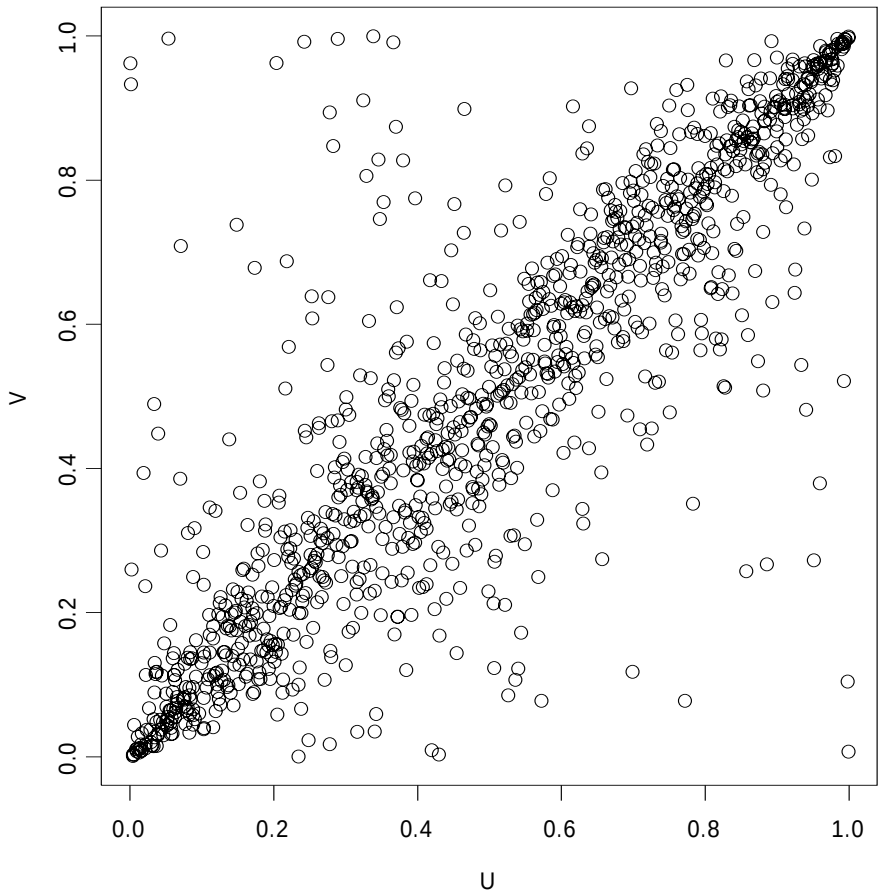

Figure 4.15: Coffee data transformed to uniform $(0, 1)$ marginals. Though most of the points suggest a strong dependence structure within the data, there are a number of two dimensional "outliers".

For $\alpha = \beta = \dfrac{k}{m}$, the dependence function for the Tawn copula can be written as

$$A(t) = \frac{m-k}{m} + \frac{k}{m}\left(t^r + (1-t)^r\right)^{1/r} = \frac{1}{m}\sum_{j=1}^{m} A_j(t),$$

where

$$A_j(t) = \begin{cases} \left(t^r + (1-t)^r\right)^{1/r} & \text{if } j \le k \\ 1 & \text{if } j > k \end{cases}.$$

A pair with a symmetric Tawn copula has the same distribution as paired maxima drawn from k independent random pairs generated by the Gumbel copula with parameter r and $m-k$ random pairs generated from the independence copula. According to the estimate $\alpha = 0.92$, we should take $m = 25$ and $k = 23$. We generated random observations from the estimated Tawn copula using this property, i. e., in order to obtain an observation $(\tilde{u}, \tilde{v})$, we generate 23 observations $\{(u_1, v_1), \ldots (u_{23}, v_{23})\}$, and 2 observations from an independence copula, (u_{24}, v_{24}) and (u_{25}, v_{25}), and set $\tilde{u} = (\max\{u_1, \ldots, u_{25}\})^{25}$ and $\tilde{v} = (\max\{v_1, \ldots, v_{25}\})^{25})$; let j_u denote the index of the maximum u_i, and j_v denote the index of the maximum v_i. Five hundred points, generated in this way, are plotted in Figure 4.16, in which points with both j_u and j_v less than or equal to 23 (points, for which both coordinates came from the Gumbel copula) are represented by empty squares, and the rest of the points are represented by solid circles. It is evident from Figure 4.16 that the "outliers" arise as maxima of the independent variables. This is why the symmetric Tawn copula can reproduce the two-dimensional structure seen in Figure 4.15. Similar dependence structures can

be achieved by setting (U, V) equal to a pair (U_1, V_1) from some copula C_1 with probability p and to a pair (U_2, V_2) from copula C_2 with probability $1 - p$, where copulas C_1 and C_2 have very different dependence structures (e.g. C_1 is a Gaussian copula with parameter 0.9, and C_2 is an independent copula).

Similarly to the example about utility indexes, we can employ the Tawn copula with the estimated parameters and the estimates of the marginal quantile functions to simulate observations from an estimated joint distribution of X and Y. Figure 4.17 shows 10,000 points simulated in this way, and the observed data points. We see that the simulated observations reproduce the characteristic "outliers".

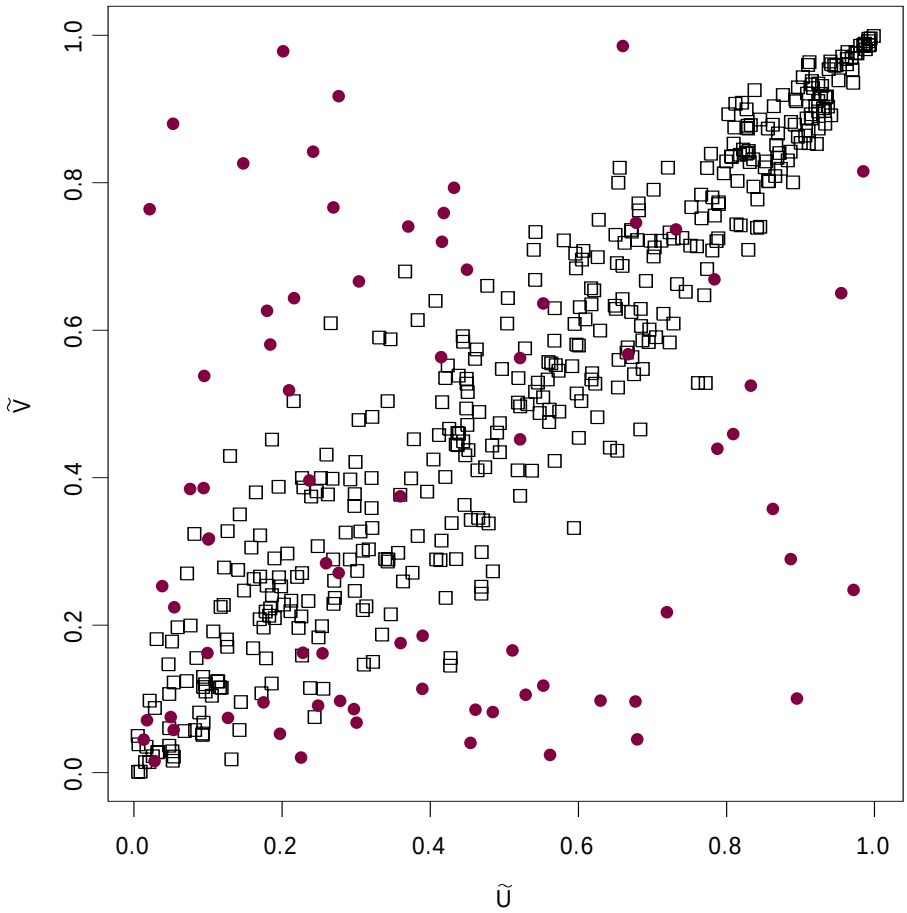

Figure 4.16: Illustration of how the Tawn copula produces the two dimensional "outliers": Five hundred points $(\tilde{u}, \tilde{v})$ generated from a Tawn copula, as paired maxima of 23 pairs generated from the Gumbel copula and 2 pairs – from the independence copula. If both $\tilde{u}$ and $\tilde{v}$ came from the observation from a Gumbel copula, this observation is represented by an empty square.

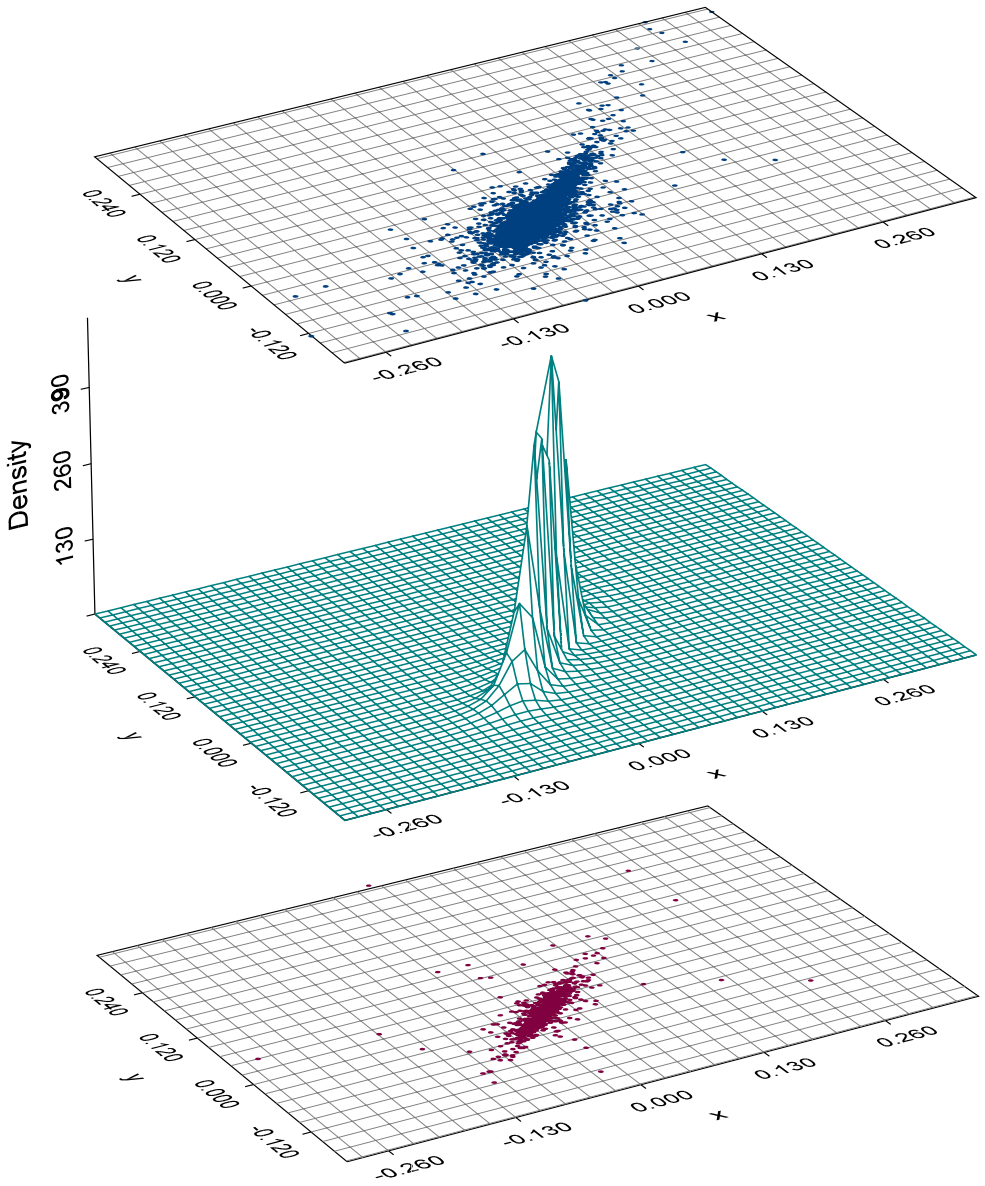

Figure 4.17: Observed data points of coffee price daily log-returns (bottom), estimate of their joint density function (middle), and 10,000 simulated points from that density function (top).

Chapter 5

Applications of Copula Modeling in Hydrology

Although it is customary to characterize extreme flood events by the magnitude of the peak discharge, the total impact from a flood may be associated with other variables, such as the flood volume and duration [Adamson *et al.*, 1999; Yue *et al.*, 1999; Sturdevant-Rees *et al.*, 2001; Smith *et al.*, 2001]. Hydrologic planning and design problems would greatly benefit from the stochastic models for modeling the joint distribution of flood peak and volume [Adamson *et al.*, 1999]. Such models should allow easy random variable generation, since stochastically generated hydrologic data are frequently used by water authorities for long-term planning of water resource development projects, including problems of reservoir design [Vodel and Stedinger, 1988; Simonovic, 2000].

The problem of inferring the bivariate distribution of flood peaks and volumes has been addressed earlier in the literature [e. g. Sackl and Bergmann, 1987; Adamson

et al., 1999; Yue *et al.*, 1999; Kelly and Krstanovic, 1997]. Some of the proposed models are based on a Partial Duration Series (PDS) approach (see section 2.3.1) and the assumption that, for a particular event, flood peaks and volumes have a bivariate Gaussian distribution [Sackl and Bergmann, 1987; Goel *et al.*, 1998]. Krstanovic and Singh [1987] used the principle of maximum entropy to characterize the bivariate distribution of flood peaks and volumes, under assumptions that the marginals of this distribution are Gaussian or exponential (see also Joe [1997] section 4.7.2, for an explanation of the maximum entropy principle). Correia [1987], considered a joint distribution of flood peaks and flood durations in the PDS framework under the assumption that both variables have an exponential distribution, but the conditional distribution of flood peaks given the flood duration is Gaussian. Kelly and Krstanovic [1997] introduced a bivariate meta-Gaussian model for analysis in the Annual Maximum Series (AMS) framework. Two recent papers attempted to avoid any assumption of Gaussian distributions and suggested bivariate models for the annual maximum peak and corresponding flood volume that can be essentially reduced to a Gumbel copula with Gumbel [Yue *et al.*, 1999] and lognormal [Yue, 2000a] marginal distributions. A related paper pursued a similar approach with rainfall peaks and storm-totals [Yue, 2000b], only logistic marginal distributions were proposed. Adamson *et al.* [1999] proposed yet another model based on a Gibbs sampler.

There have been no attempts to apply the theory of copulas to modeling the joint distribution of flood peaks and flood volumes in either the AMS or PDS framework,

and this is one of the primary topics of this chapter. Our definition of the AMS framework for the bivariate distribution is quite different from the traditional one [e.g. Adamson *et al.*, 1999; Yue *et al.*, 1999], where the joint distribution of annual maximum flood peaks and the *corresponding* volumes to these particular flood peaks is considered. In order to apply multivariate extreme value theory to this problem, one should define "flood events" and then consider the annual maximum flood peaks and the annual maximum flood volumes due to these events. Under this definition, we can argue that annual maximum flood peaks and flood volumes correspond to a bivariate extreme event, since each of the variables is a maximum of a collection of random variables that might be considered independent. In the same way that we consider the Generalized Extreme Value (GEV) distribution to be a starting point for modeling univariate distributions of extreme events, bivariate extreme value distributions can be used as a starting point for modeling bivariate extreme events [Tawn, 1990]. Bivariate Generalized Extreme Value (BGEV) distributions are distributions with GEV marginals and EV copulas [Joe, 1997, p. 172]. It seems reasonable, then, to assume that annual maximum flood peaks and volumes have a BGEV distribution.

In application, however, one is not that interested in the AMS model as we define it, but rather in the joint distribution of flood peaks and corresponding volumes. For this reason, we revisit the PDS approach, and consider the joint distribution of flood peaks and volumes in the PDS framework. We derive the form of a copula associated with flood peaks and volumes in the PDS framework, under the assumption that the

joint distribution of annual maxima of flood peaks and volumes is BGEV. In a later section, we illustrate this approach with examples.

5.1 AMS and PDS approaches revisited

We would like to revisit the issue of modeling extreme flood events using the Annual Maximum Series (AMS) and Partial Duration Series (PDS) framework with an additional random variable in our analysis: the total volume of water associated with the flood. This volume is usually defined as the total amount of water that passed the stream gauging station in some time window including the time of the flood peak. USGS maintains historical records of the mean daily discharge measurements for most of its stream gauging stations, so the flood volumes can be easily computed from this data.

In the PDS framework, we assume that "flood events" arrive according to a (stationary) Poisson process with (annual) rate λ, and that the floods' magnitudes $Z_1, Z_2, \ldots$ are i.i.d. random variables independent of the arrival process, each having a GP distribution F_{η_1} with parameter vector $\eta_1 = (a_1, \delta_1, \kappa_1)$ (see section 2.3 and equation (2.11)). Each flood peak Z_i is associated with a flood volume V_i, $i = 1, 2, \ldots$. The V_is are also assumed to be i.i.d. random variables, independent of the arrival process, each having a GP distribution F_{η_2} with parameter vector $\eta_2 = (a_2, \delta_2, \kappa_2)$. The pairs (Z_i, V_i) are assumed to be independent, and the joint distribution of Z_i and V_i is associated with a copula $C^{(\mathrm{PDS})}$ that is the same for all i's, $i = 1, 2, \ldots$.

In the AMS framework, we consider random variables X and Y representing the annual maximum flood peak and annual maximum flood volume, respectively. As discussed above, it is natural to assume that X and Y have a joint BGEV distribution, meaning that X and Y have marginal GEV distributions G_{θ_1} and G_{θ_2} with parameter vectors $\theta_1 = (a_1, b_1, k_1)$ and $\theta_2 = (a_2, b_2, k_2)$, respectively, and that the copula associated with X and Y is an EV copula

$$C^{(\mathrm{AMS})}(u, v) = \exp\{\log(uv)\, A(\log u / \log(uv))\}, \tag{5.1}$$

for some dependence function $A(t)$.

We know that in the univariate case, the two frameworks imply the same distribution of annual flood peaks if and only if θ_1 and η_1 are connected through the relationships in (2.12) (of course, this is true only for flood peaks above the threshold δ_1, see section 2.3.1). For the bivariate case, it can be assumed that the joint annual maximum distribution implied by the two models will be also the same for a certain form of $C^{(\mathrm{PDS})}$. The following proposition summarizes this result:

Proposition. Let $\tilde{X}$ and $\tilde{Y}$ represent annual flood peak and volume under the PDS framework. The joint distribution of $\tilde{X}$ and $\tilde{Y}$ for $x \geq \delta_1$ and $y \geq \delta_2$

$$\mathbb{P}\left\{\tilde{X} \leq x, \tilde{Y} \leq y\right\} = \tilde{H}(x, y), \tag{5.2}$$

is equal to the BGEV distribution with marginals G_{θ_1}, G_{θ_2}, and copula $C^{(\mathrm{AMS})}$ if and only if the parameters θ_1 and θ_2 are connected with η_1, and η_2, respectively,

through the relationships in (2.12), and $C^{(\mathrm{PDS})}$ is an Archimax copula with generator $\phi(t) = 1 - t$ and the same dependence function $A(t)$ as the $C^{(\mathrm{AMS})}$ copula .

Since the results for the marginals are well known (e. g., Smith [1984]; Madsen and Rosbjerg [1997]), we need only show that the copula associated with $\tilde{X}$ and $\tilde{Y}$ is an EV copula with dependence function $A(t)$ if and only if

$$C^{(\mathrm{PDS})}(u,v) = \max\left(1 - \left[(2-u-v)\,A\left(\frac{1-u}{2-u-v}\right)\right], 0\right) \tag{5.3}$$

(recall that, by the convention for Archimedean generators, $\phi(t) = 1 - t$ implies $\phi^{-1}(z) = 0$, for $z \geq 1$). To this end, note that, for $x \geq \delta_1$ and $y \geq \delta_2$,

$$I\!P\left\{\tilde{X} \leq x, \tilde{Y} \leq y\right\} = \exp\{-\lambda\left(1 - C^{(\mathrm{PDS})}(F_{\eta_1}(x), F_{\eta_2}(y))\right)\}, \tag{5.4}$$

so the copula $\tilde{C}$ associated with $\tilde{X}$ and $\tilde{Y}$, is

$$\tilde{C}(\tilde{u},\tilde{v}) = \exp\{-\lambda\left[1 - C^{(\mathrm{PDS})}\left(F_{\eta_1}(G_{\theta_1}^{-1}(\tilde{u})), F_{\eta_2}(G_{\theta_2}^{-1}(\tilde{v}))\right)\right]\}, \tag{5.5}$$

where $G_{\theta_1}^{-1}$ and $G_{\theta_2}^{-1}$ are the quantile functions of the GEV distribution (see equation (2.2)). Assuming that η_1 and θ_1 are connected through equations (2.12), we have

$$\begin{aligned} F_{\eta_1}(G_{\theta_1}^{-1}(u)) &= 1 - \left[1 - \frac{\kappa}{\alpha}\left(b + \frac{a}{k}\left(1 - (-\log u)^k\right) - \delta\right)\right]^{1/\kappa} \\ &= \begin{cases} 1 + \frac{\log u}{\lambda} & \text{if } u \geq e^{-\lambda} \\ 0 & \text{otherwize} \end{cases} \end{aligned}$$

A similar expression holds for $F_{\eta_2}(G_{\theta_2}^{-1}(u))$. Substituting these expressions into (5.5) (note that (5.5) was derived under the condition that $x \geq \delta_1$ and $y \geq \delta_2$, which

implies $\tilde{u} \geq e^{-\lambda}$ and $\tilde{v} \geq e^{-\lambda}$), we get

$$\tilde{C}(\tilde{u}, \tilde{v}) = \exp\left\{-\lambda\left[1 - C^{(\mathrm{PDS})}\left(1 - \frac{-\log \tilde{u}}{\lambda}, 1 - \frac{-\log \tilde{v}}{\lambda}\right)\right]\right\}. \tag{5.6}$$

We now set the $\tilde{C}(\tilde{u}, \tilde{v})$ in (5.6) equal to $C^{(\mathrm{AMS})}(\tilde{u}, \tilde{v}) = \exp\{\log(\tilde{u}\tilde{v})\, A(\log \tilde{u}/\log(\tilde{u}\tilde{v}))$ to express the copula in EV form. For $\tilde{u} \geq e^{-\lambda}$ and $\tilde{v} \geq e^{-\lambda}$, i.e., for years with at least one event, and making the substitutions $u = 1 - (-\log \tilde{u})/\lambda$ and $v = 1 - (-\log \tilde{v})/\lambda$, we obtain the following equation for $C^{(\mathrm{PDS})}$:

$$-\lambda\left(1 - C^{(\mathrm{PDS})}(u, v)\right) = -\lambda(2 - u - v)A\left(\frac{1 - u}{2 - u - v}\right).$$

This equation, together with the condition that $C^{(\mathrm{PDS})}(u, v)$ is a valid copula, implies that $C^{(\mathrm{PDS})}(u, v)$ has the form given in (5.3). The converse of this statement is easy to see by substituting the expression for $C^{(\mathrm{PDS})}(u, v)$ in (5.3) into (5.6). This concludes the proof of the proposition.

Copula $C^{(\mathrm{PDS})}$ is an Archimax copula with a dependence function A determined by the AMS model, and the Archimedean generator of the Frechet lower limit copula $C_l(u, v) = \max(u + v - 1, 0)$ (see, e.g., Nelsen [1999], section 2.2). This observation is very interesting, since it implies that $C^{(\mathrm{PDS})}$ has the least dependence structure of all Archimax copulas with generator A under the ordering $\prec_K$. Figure 5.1 shows a sample generated from this copula. Note that this copula has a singular component, and there is no points produced in the area labeled A on the plot. This has a clear interpretation: for any given "event", at least one of the variables (flood peak or volume) must be of significant magnitude.

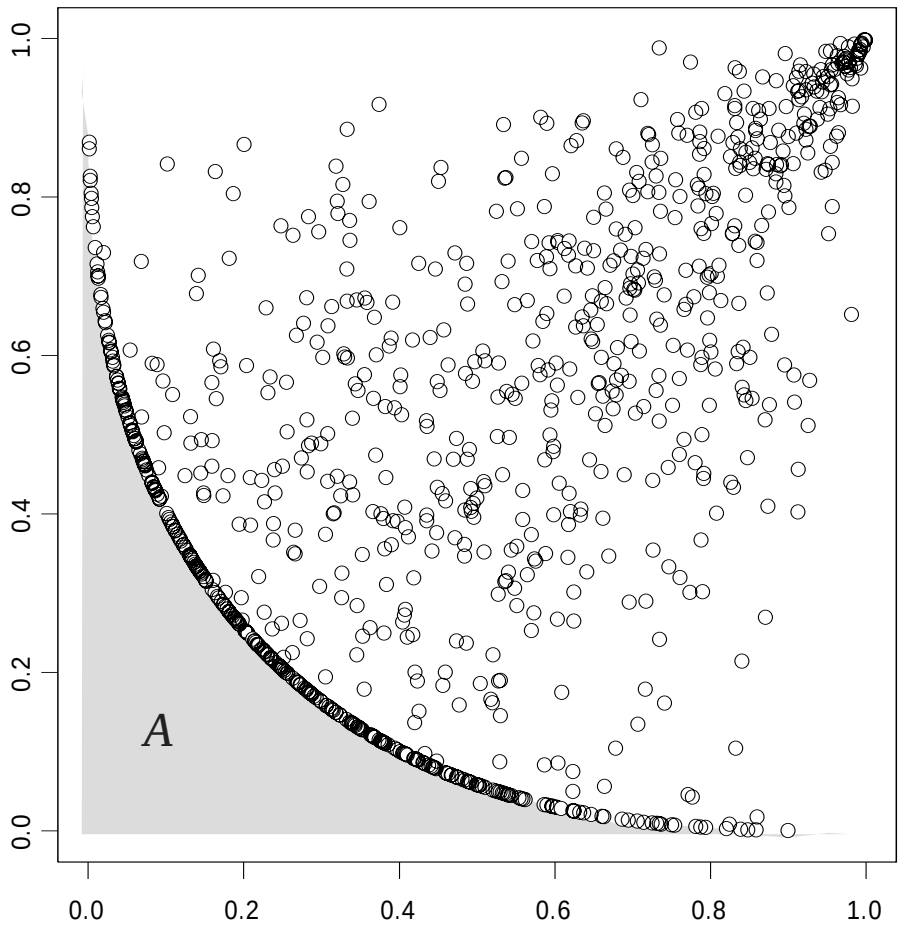

Figure 5.1: 1000 observations generated from a $C^{(\mathrm{PDS})}$, with Galambos dependence function, and parameter $\delta = 2.0$. The copula's singular component is clearly defined.

The AMS copula $C^{(\mathrm{AMS})}$ determines the $C^{(\mathrm{PDS})}$ copula we will use for modeling the joint distribution of flood peaks and flood volumes in the PDS framework. Consider two particular choices for copula $C^{(\mathrm{AMS})}$: the Gumbel and Galambos copulas (see section 3.2). Each of these copulas has only one parameter, and a relatively simple algebraic form. If $C^{(\mathrm{AMS})}$ is a Galambos copula with parameter $\delta \geq 0$, the copula $C^{(\mathrm{PDS})}$ reduces to

$$C^{(\mathrm{PDS})}(u,v) = \max\left(u + v + \left[(1-u)^{-\delta} + (1-v)^{-\delta}\right]^{-1/\delta} - 1, 0\right),$$

and if it is a Gumbel copula with parameter $\delta \geq 1$, the copula $C^{(\mathrm{PDS})}$ becomes

$$C^{(\mathrm{PDS})}(u,v) = \max\left(1 - (u^{\delta} + v^{\delta})^{1/\delta}, 0\right).$$

The latter form is that of an Archimedean copula, and this copula is briefly discussed in Chapter 4 of Nelsen [1999].

5.2 Parameter Estimation

Before estimating the parameters of this model, one must choose the parametric form of the dependence function A. We used Gumbel and Galambos dependence functions (see section 3.2) for this purpose, since they have the simplest form and only one parameter each. As we saw in section 3.2, there is not much difference between EV copulas from different families at a fixed Kendall's tau.

Assuming that a one-parameter copula is used for $C^{(\mathrm{AMS})}$, the AMS model has seven parameters in all: three parameters for each GEV distribution and one copula

parameter. As in section 3.6, one can estimate parameters of this model by either using the full likelihood method, or by separately estimating the margins and then estimating the copula parameters using the transformed data. Moreover, simulation experiments suggest that the error produced by either method is approximately the same. Separate estimation of the marginals is more convenient, since we can use an AMS, PDS, or even MLE2 approach to estimate the marginals parameters. In order to estimate the copula parameter, we can extract maximum flood peaks and maximum flood volumes for each year from available datasets, apply the marginal transformation to the extracted data, and estimate a copula parameter using the maximum likelihood method.

In order to simulate data for a PDS approach, we must specify λ – the rate of peak arrivals in the Poisson process. It is easy to see that the PDS model is set up such that any value of λ will result in a valid model for maximum flood peaks and volumes. Different values of λ imply different values of the threshold over which we will observe flood peaks and volumes. Figure 5.2 illustrates this point with simulated observations for 40 years of data for two bivariate PDS models with the same AMS framework parameters and different values of λ. The distribution of points in the shaded areas is the same for both models. The threshold δ_1 over which we observe flood peaks satisfies $G_{\theta_1}(\delta_1) = e^{-\lambda}$ (see equation (2.16)), and, similarly, the threshold δ_2 over which we observe flood volumes satisfies $G_{\theta_2}(\delta_2) = e^{-\lambda}$. Note that not all

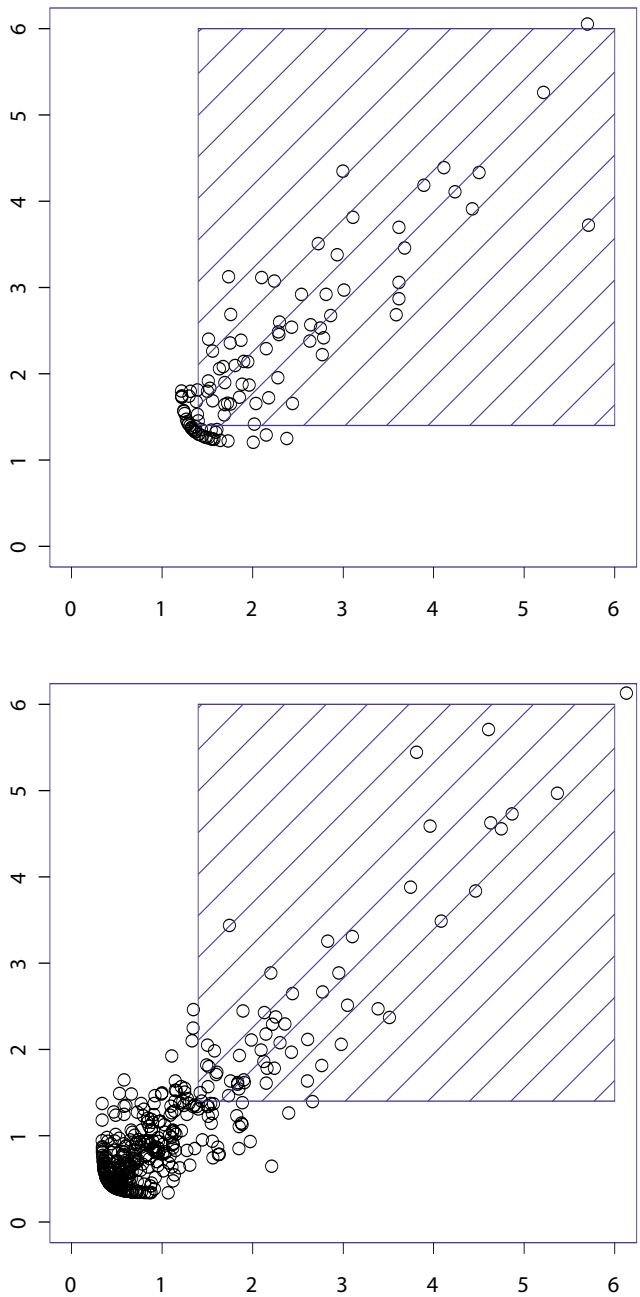

Figure 5.2: Simulated 40 years of data for two bivariate PDS models with the same AMS framework parameters and $\lambda = 2.5$ (top) and $\lambda = 2.5$ (bottom)

values of flood peaks and flood volumes above the specified thresholds are admissible, because of the structure of $C^{(\mathrm{PDS})}$.

In the examples below, we compared simulated observations from the described model with actual data. We chose λ to correspond to the threshold for which we had records of measured peaks. Generally, the problem of choosing the "best" lambda is as complicated as choosing the "best" threshold in the univariate PDS model, and it is beyond the scope of this discussion.

The $C^{(\mathrm{PDS})}$ copula can be used not only for modeling joint distributions of peaks and volumes for flood events, but also for modeling flood peaks in adjacent or nested basins. Recall that the historical PDS records for the USGS stream gauging stations include only peaks over a predetermined threshold. Therefore, if a particular storm caused a peak above the threshold in one of the basins but not in another, the second peak would not be recorded and be unavailable for any parameter estimation procedure. This fact does not cause a problem for the approach suggested here, since parameter estimation is performed in the AMS framework for which data is regularly available.

Obviously, the proposed simultaneous approach in the PDS and AMS frameworks can be applied to modeling other extreme events, such as sea level heights, winds, rainfall, etc.

5.3 Flood Peaks and Flood Volumes

We analyzed the flood peak and flood volume data for the Potomac at Point of Rocks stream gauging station, a nearly 10,000 mi^2 - sized watershed. This is the largest basin in the sample of basins considered in Chapter 2, and it has the longest period of record available.

PDS flood peak data from 1931 to 1997, AMS data from 1894 to present, and mean daily discharge data from 1921 to present is available for this station. We used the AMS data for the period from 1921 to 1997 to compute estimates for the model's parameters, and then used the PDS data for the period from 1931 to 1997 to test the performance of the model. To compute flood volumes from mean daily discharge data, we need to define a time window. Based on hydrological analyses for the Appalachian region, Sturdevant-Rees *et al.* [2001] suggested a window beginning two days before the flood peak and ending three days after the flood peak. Using this window, we extracted a PDS series of flood volumes corresponding to the PDS flood peak data, and independently computed the maximum flood volumes for each year. Figure 5.3 shows the maximum flood peaks and flood volumes for this station plotted against one another. It is customary to divide flood volumes by the basin drainage area, and express the volume (sometimes referred to as runoff) in millimeters.

The parameters of the marginal distributions were estimated using the MLE2-MIX1 method for flood peaks (the estimates obtained are $a_1 = 1212$, $b_1 = 2461$, $k_1 = -0.26$) and the MIX1 method for flood volumes ($a_2 = 10.8$, $b_2 = 24.5$, $k_2 =$

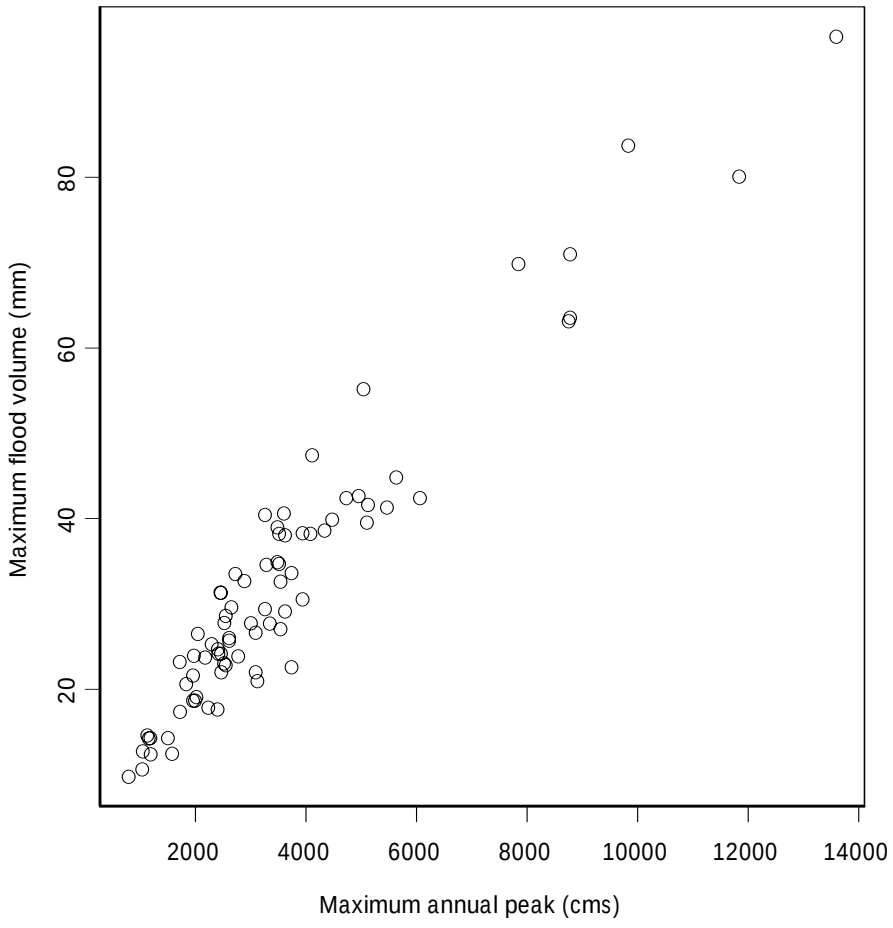

Figure 5.3: Maximum flood peaks and flood volumes for Potomac at Point of Rocks plotted against one another.

-0.14). Following the estimation procedure, the observations in Figure 5.3 were transformed to uniform $(0, 1)$ marginals, and the Galambos copula was fitted to the observations. The estimated parameter for the Galambos copula was determined to be 3.85. Figure 5.4 shows the non-parametric estimate of the dependence function A and the parametrically estimated dependence function. The two estimates are very close to one another.

In order to compare the agreement of the estimated PDS model with the observed data, we computed flood volumes associated with the measured peaks-over-threshold for the Potomac river basins. For peak and volume thresholds x and y, we computed the expected number of flood peaks exceeding threshold x and the *corresponding* flood volume exceeding the threshold y for a given year. We then compared these values with those of the observed data. Table 5.1 provides this comparison for selected values of x and y. We see that the model agrees well with the observations. Figure 5.5 shows simulated values of flood peaks and volumes in the PDS framework for a period of 500 years (black diamonds), and blue circles mark the observations (for 59 years). The simulated observations seem to follow a process close to the observed one.

5.4 Flood Peaks at Two Nested Basins

As mentioned previously, the model described above can be used to model the joint distribution of flood peaks in two adjacent or nested basins. To illustrate this ap-

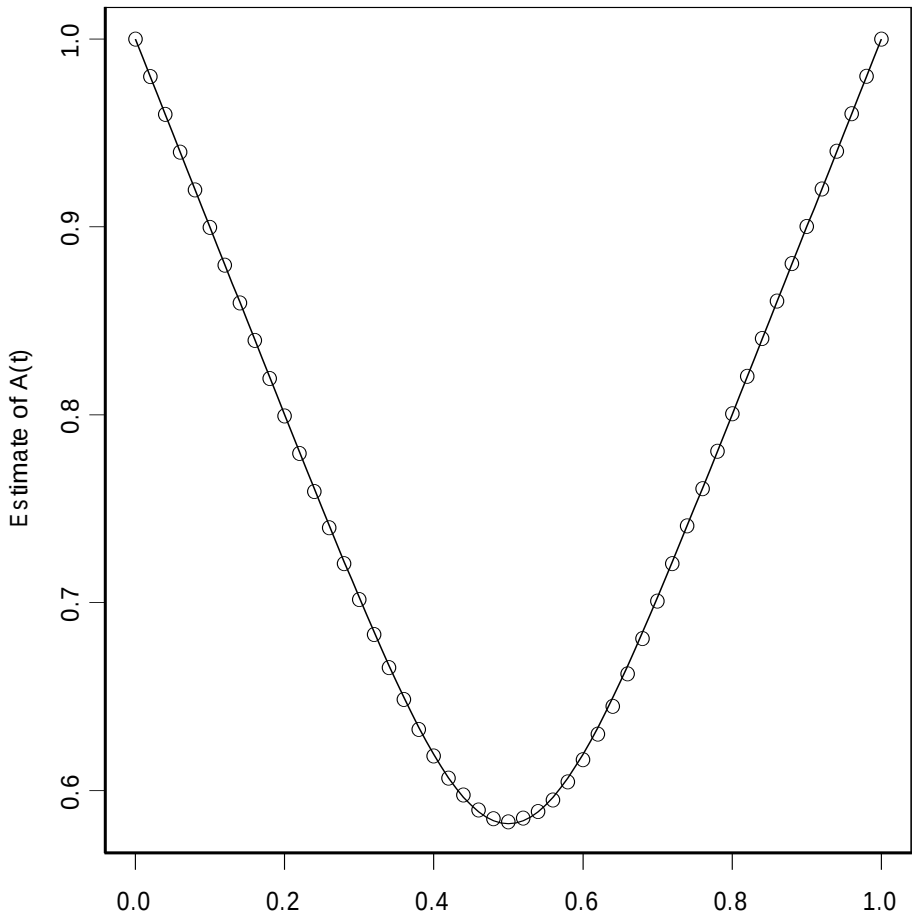

Figure 5.4: The non-parametric estimate of the dependence function A (points) and the parametrically estimated dependence function (curve) of $C^{(\mathrm{AMS})}$ for a distribution of flood peaks and volumes for Potomac at Point of Rocks.

Peak threshold (cms)	6000	5000	4000	2500
Volume Threshold (mm)	40	30	40	20
Observed	0.134	0.203	0.220	1.25
Modeled	0.127	0.221	0.252	1.23

Table 5.1: Expected number of flood peaks exceeding certain thresholds in both peak and volume.

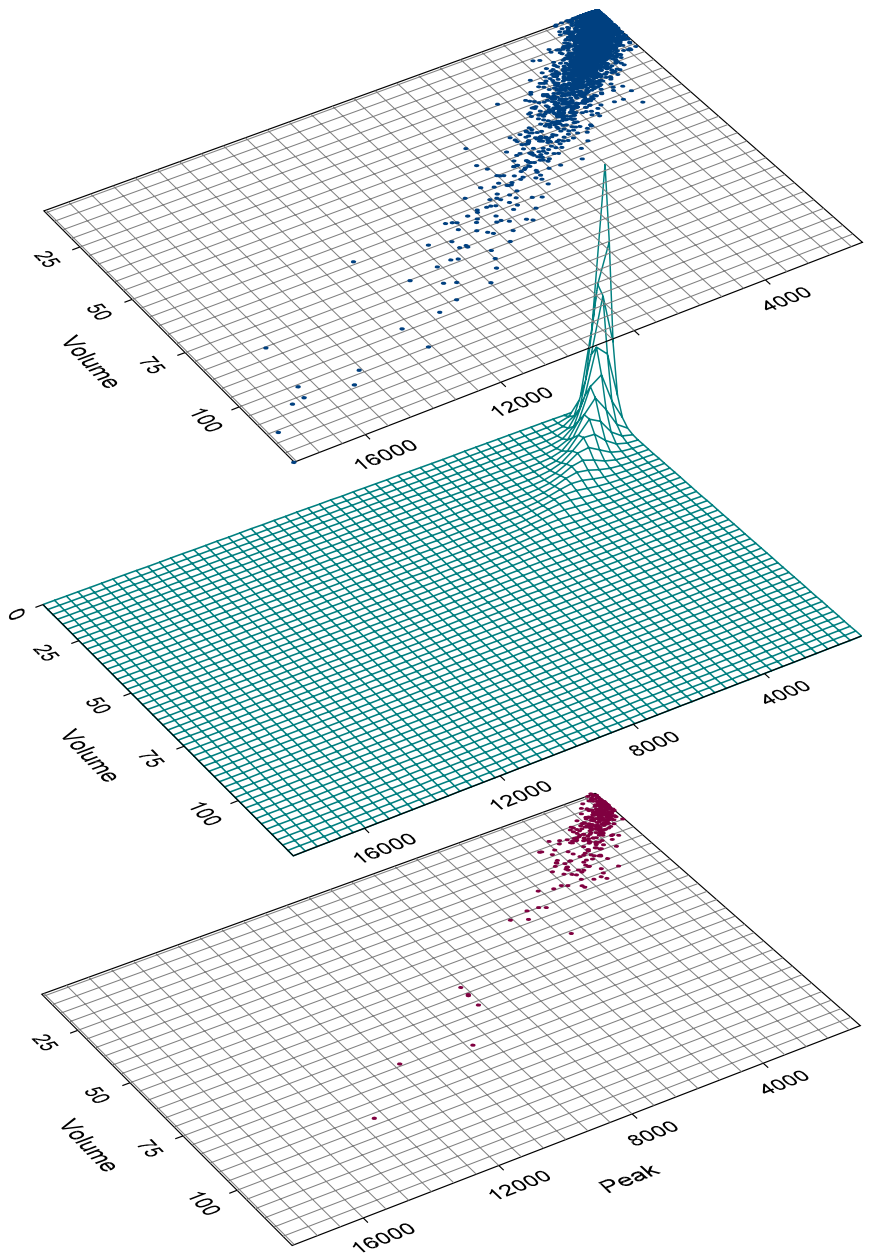

Figure 5.5: Simulated values for a period of 500 years (top), estimate of the joint probability density function (middle), and observed values for 59 years (bottom) of flood peaks and volumes for Potomac at Point of Rocks in the PDS framework.

proach, we consider the following two basins: Potomac at Point of Rocks, and South Fork Shenandoah at Front Royal, a basin with drainage area 1642 mi^2 that is upstream of Potomac of Point of Rocks [Sturdevant-Rees *et al.*, 2001] and also included in the sample of basins discussed in Chapter 2.

Following the prescription, we estimate the marginal GEV distributions of annual flood peaks for the two basins in the AMS framework. The estimated parameters for the Potomac at Point of Rocks basin are $a_1 = 1212$, $b_1 = 2461$, $k_1 = -0.26$ (as above), and are $a_2 = 300$, $b_2 = 470$, $k_2 = -0.33$ for the South Fork Shenandoah at Front Royal. Figure 5.6 shows the values of the observed annual flood peaks for the two basins plotted against one another. Using the estimates of the marginal distributions, we transform this data to observations of two uniform $(0, 1)$ random variables, and estimate their copula. Again, we used the Galambos copula. The parameter estimated for the Galambos copula was 1.65.

We can now compare our model to the observations of peaks-over-thresholds for the two basins in the PDS record (the threshold that is used by USGS for Potomac at Point of Rocks station is 990 cms and is 243 cms for the South Fork Shenandoah at Front Royal stream gauge). The measured data suggests that an individual storm causes a peak-over-[corresponding]-threshold in both basins with probability 0.83, and a peak-over-threshold at South Fork Shenandoah but not at Potomac at Point of Rocks with probability 0.052. The corresponding values from our PDS model are 0.87 and 0.043, which provides a very good agreement. Figure 5.7 shows the observed

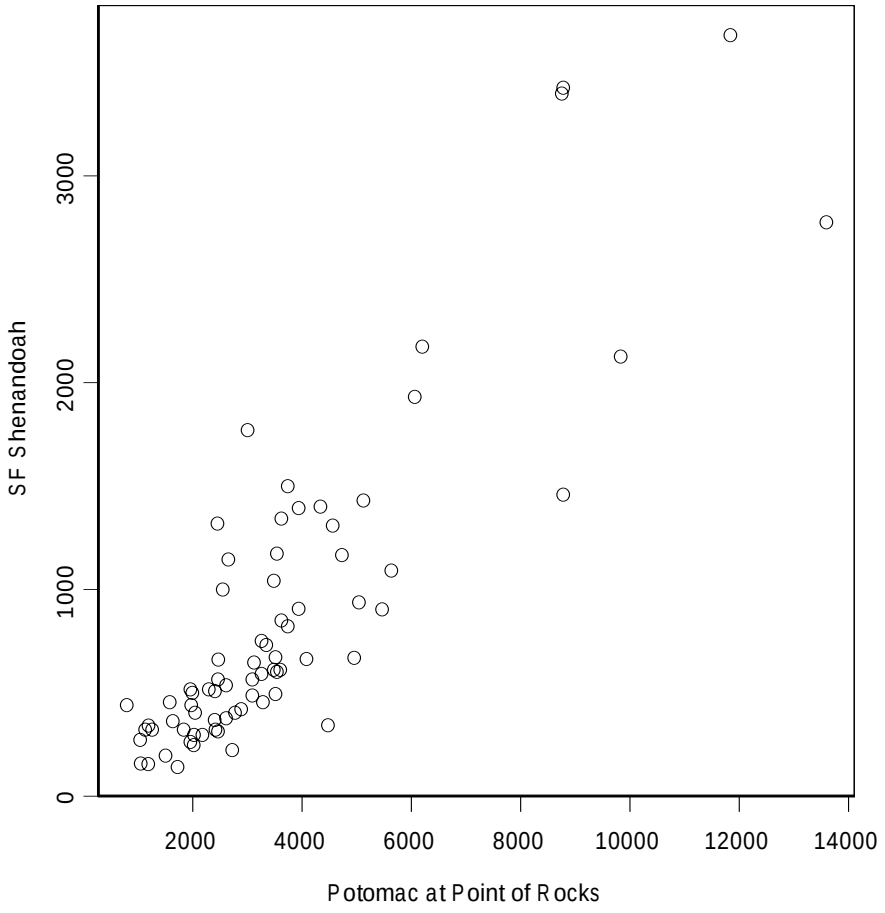

Figure 5.6: Values of the observed annual flood peaks (cms) for Potomac at Point of Rocks and South Fork Shenandoah at Front Royal.

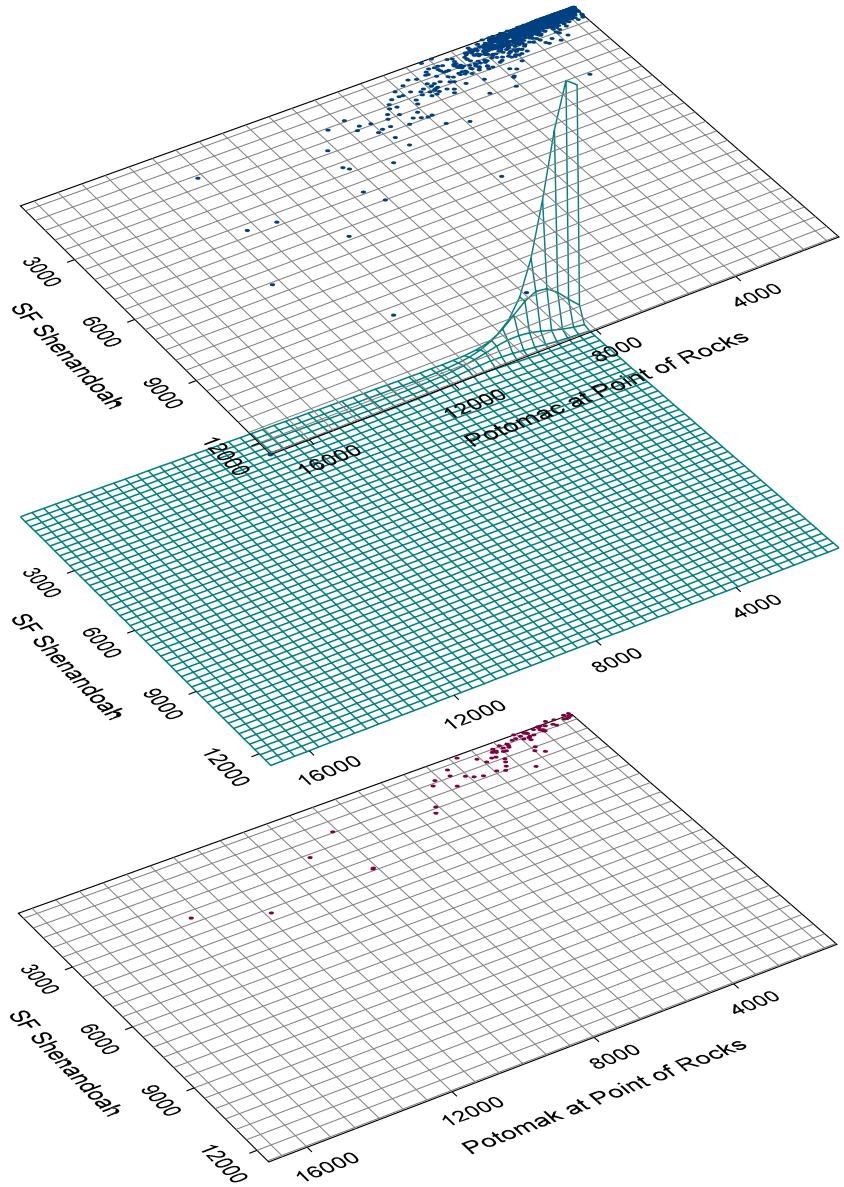

Figure 5.7: Simulated values for a period of 500 years (top), estimate of the joint probability density function (middle), and observed values for 66 years (bottom) of flood peaks at Potomac at Point of Rocks and South Fork Shenandoah at Front Royal in the PDS framework.

flood peaks for the two basins (red) and the simulated ones (empty) using the PDS approach.

5.5 Summary of Chapter 5

There are two main conclusions from the results reported in this chapter:

- We have derived the general form for a copula in the PDS framework, which implies the BGEV distribution in the AMS framework for GP marginals.
- This approach is illustrated by examples of modeling the joint distribution of flood peaks and flood volumes, and of flood peaks in two nested basins.

Chapter 6

Scaling Properties of Flood Peaks

In Chapter 2, we briefly discussed the scaling properties of flood peaks in the central Appalachians. The scaling theory of flood peaks is important for creating methods to assess flood peak distributions for ungauged drainage basins (see Gupta and Waymire [1998]). Two major theories for the areal dependence of the annual flood peak distribution have been developed: the simple-scaling (index-flood) theory, and the multiscaling theory [Gupta and Waymire, 1990, 1998; Smith, 1992].

To illustrate these two theories, let X_A be a random variable denoting the magnitude of the annual flood peak in a basin with drainage area A. The theory of simple scaling suggests that there is a positive function g such that for any two basins in a particular region with drainage areas A_1 and A_2, the random variables X_{A_1} and $g(A_1/A_2)X_{A_2}$ have the same distribution. Multiscaling theory allows the function g to be random and statistically independent of X_A. Both theories lead to the conclusion

that all moments $\mathbb{E}\left[(X_A)^m\right]$ of X_A exhibit log-log linearity with the drainage basin area A. Under the assumptions of simple-scaling theory, the slopes of these linear relations change linearly with the order of the moment, m. In addition, the coefficient of variation (CV), defined as the ratio of the standard deviation of X_A to the expected value of X_A, should not change with basin area [Gupta and Waymire, 1990, 1998]. In the multiscaling framework, however, the CV decreases with increasing basin area [Gupta and Waymire, 1990, 1998].

Both theories were tested in Smith [1992] on annual flood peak data for a sample of stream gauging stations described in Chapter 2 in the central Appalachians (see Figure 2.3 for locations of these stream gauging stations). It was shown that the logarithms of the moments of X_A do vary linearly with $\log A$, although the CVs exhibit a very peculiar dependence on basin area (A): CV increases with increasing A for basins with drainage areas smaller than some critical area A_c, and decreases with increasing A for basins larger than A_c. This effect is shown in Figure 6.1. For the central Appalachians region, the critical area A_c was determined to be approximately 50 km^2. Similar results have been obtained subsequently by other investigators for a diverse collection of study sites in the United States and abroad [Gupta *et al.*, 1994; Bloschl and Sivapalan, 1997].

In section 2.4, we fit the GEV distribution to flood peak data from each station, and studied the scaling behavior of the ensuing parameter estimates. We saw that the estimates of the scaling parameter a and location parameter b of the GEV distribution

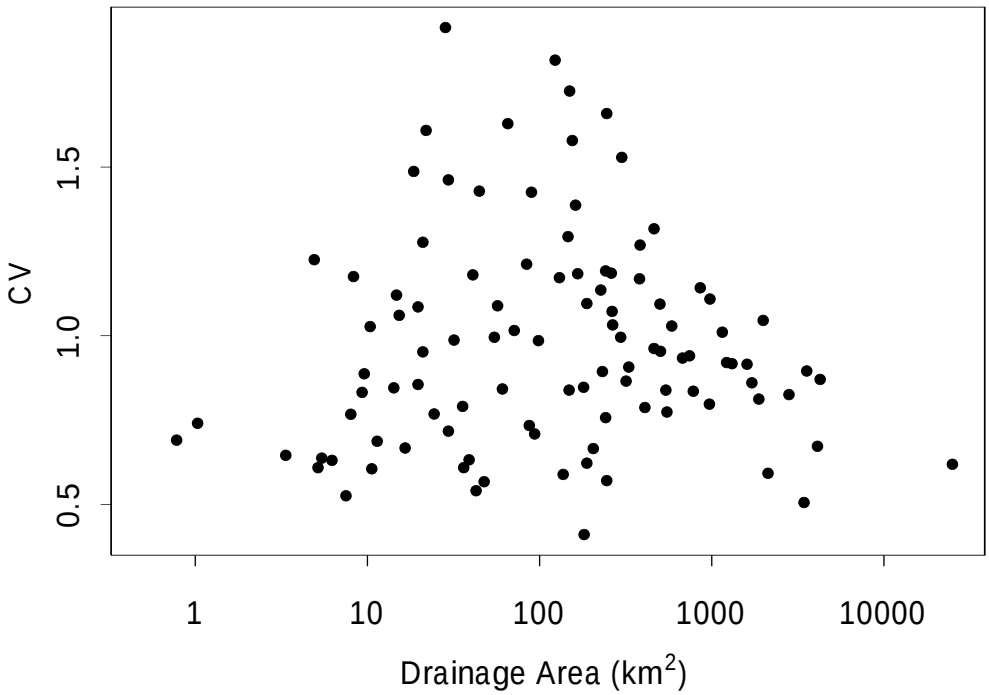

Figure 6.1: CV – Drainage Area relationship for central Appalachian floods. From Smith [1992].

do exhibit log-log linear relationships with basin area (Figure 2.6), and that the estimates of the shape parameter k did not exhibit any systematic behavior with basin area (page 41). In this chapter, we focus on the scaling behavior of moments and CVs estimated from the data, without any assumptions about the distribution of annual flood peaks.

It has been suggested that the spatial variability of rainfall is the predominant factor in the decrease of CV at large scales while the structure of the drainage network controls the behavior of CV on smaller scales [Gupta and Dawdy, 1995; Gupta *et al.*, 1994; Robinson and Sivapalan, 1995]. We will explore the dominant factors influencing the behavior of CV by studying the interaction between the spatial and temporal properties of rainfall and the network structure of river basins. The principal tools used in this study are a statistical model of rainfall, which produces a rainfall field that varies both in space and time, and a simple, network-based drainage model for converting rainfall fields to a discharge at the drainage basin outlet. The study region is a $\sim$5000 km^2 area in the Edwards Plateau, Texas (see Figure 6.2 for location). The Edwards Plateau is an interesting study site because it has experienced some of the largest floods in the continental US [Costa, 1987].

Our model assumes that flood peaks are due to storms with exceptionally heavy rainfall ("extreme rainfall events"), and we are interested in computing the peak discharge generated by these heavy storms. Randomly sampled river basins are drawn from the population of river basins in our area of study, and checked to ensure that

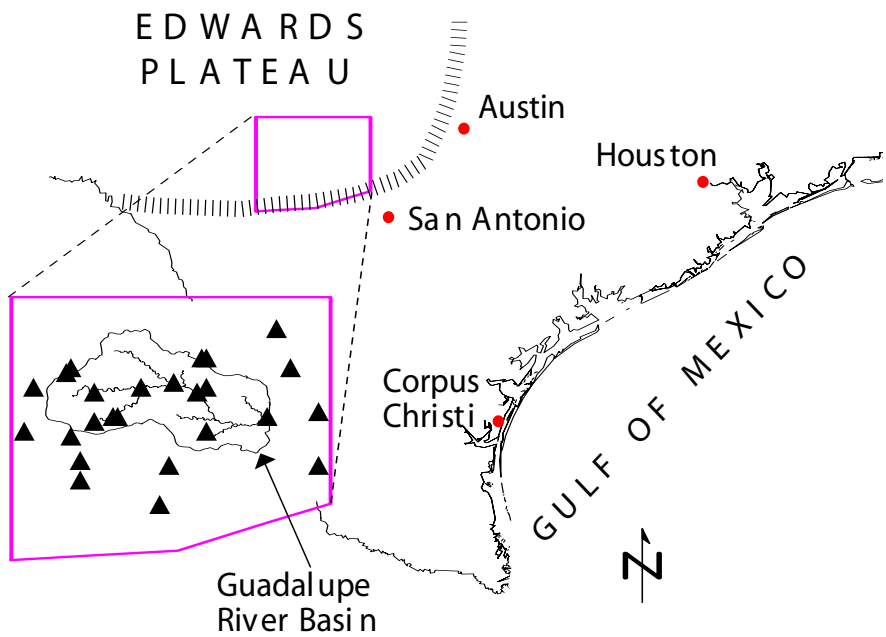

Figure 6.2: Partial map of Texas, showing the regions mentioned in the text. The inset corresponds to the area of interest. Triangles represent rain gauge locations.

they are not nested inside one another (that is, none of them is a sub-basin of any other). We can then assume that runoff generated at the outlet of one basin does not influence the runoff from any other basin. Additionally, comparison of the results from different random samples of basins will assure us that the results are not an artifact of the basin choice.

Because the rainfall field is a complicated stochastic process, its interaction with the river network makes direct theoretical computation of the flood peak distribution difficult. Instead, numerical simulations are used to infer the properties of the distribution. To simulate the rainfall field, the models of Smith and Karr [1990] and Over and Gupta [1996] were combined, leading to the model described in section 6.1 (see Harris *et al.* [1996]; Onof *et al.* [1996]; Over and Gupta [1996]; Perica and Foufoula-Georgiou [1996]; Rodriguez-Iturbe *et al.* [1987]; Smith and Karr [1990]; Waymire *et al.* [1984] for additional discussion of space-time rainfall models). To compute the discharge at basin outlets, a simplified version of the Network Model was used which depends only on the structure of the network (i.e., two different basins with an identical network structure will respond the same to the same rainfall field). See [Rodriguez-Iturbe and Rinaldo, 1998, chapter 7] for a detailed discussion of network models. Using this model, the maximum discharge for each basin is determined for each simulated storm, allowing analysis of flood peak moments and CVs. The complete version of the Network Model, which also includes surface water routing and

the Green and Ampt with redistribution model for water infiltration, is described in appendix A.

6.1 Storm Model

Simulating an extreme rainfall event over an area of $\sim$5000 km^2 is a complicated task. Accordingly, there are a number of spatio-temporal stochastic models for extreme rainfall that have been developed in the literature (e.g. Northrop [1998]; Cox and Isham [1998]; Rodriguez-Iturbe *et al.* [1987]; Waymire *et al.* [1984]). To determine which model will work the best for us, we note that the only extended record of data available in our area of interest (for the purpose of parameter estimation) is the daily rainfall accumulation measured by 23 rain gauges, which limits the number of models for which we would be able to estimate parameters. Additionally, there is no information available on the temporal structure of the rainfall for each event, because the data includes only daily accumulation totals, so we modeled the spatial accumulations and the temporal structure independently. For modeling the spatial structure of the rainfall, we used a model developed by Smith and Karr [1990] with a relatively small number of parameters to ease parameter estimation. The use of more complicated models (e.g. Rodriguez-Iturbe *et al.* [1987]; Northrop [1998]) would be practical and desirable if additional rain gauge or radar data were available for the parameter estimation. For modeling the temporal structure, we used a cascade model [Over and Gupta, 1996].

To simulate the extreme rainfall totals, the statistical model described in Smith and Karr [1990] was used, with an additional assumption of spatial homogeneity in the region. This model can be described as follows:

- The arrival times of extreme rainfall events form a Poisson process N on the time axis with seasonally-varying mean $\lambda(t)$. Because only the annual flood peaks are of interest, only the annual frequency of extreme rainfall events Λ is utilized in the simulations.

- A rain field is constructed as a collection of distinct storm cells. The centers of these cells form a Poisson process M (independent of N) on the plane with spatial rate a.

- For a particular storm, the total rainfall S_c in the center of each cell is independent of N and M, and has an exponential distribution with parameter b.

- The storm-total rainfall associated with each cell is assumed to be distributed symmetrically about the center of the cell with spread function

$$h(r) = e^{-2c^2r^2},$$

  and a magnitude equal to S_c at $r = 0$.

Suppose that, for the ith storm, there are M storms cells with coordinates $Y_1, Y_2, \ldots, Y_M$ and center accumulations $S_{c,1}, S_{c,2}, \ldots, S_{c,M}$, respectively. Under the

assumptions above, the storm-total rainfall $S_i(x)$ from the storm at a point x on the plane is

$$S_i(x) = \sum_{j=1}^{M} S_{c,j} h(\|x - Y_j\|). \tag{6.1}$$

Moreover, note that the points $\{(Y_1, S_{c,1}), (Y_2, S_{c,2}), \ldots, (Y_M, S_{c,M})\}$ form a Poisson process on $I\!R^2 \times I\!R_+$ with intensity measure $abe^{-bs}\, dyds$. This property is used to derive the moment parameter estimators for the model.

Smith and Karr [1990] computed the moments for this field, giving the mean μ and the variance σ^2 for the amount of storm-total rainfall at all locations x (the "storm totals") as

$$\mu = \frac{\pi}{2} \cdot \frac{a}{c^2 b}$$

$$\sigma^2 = \frac{\pi}{2} \cdot \frac{a}{c^2 b^2},$$

and the correlation coefficient ρ for the storm totals between two locations separated by a distance r as

$$\rho(r) = e^{-c^2 r^2}.$$

The mean, variance, and correlation coefficient do not depend on x because of the assumption of homogeneity in the region.

This model has 4 parameters to be estimated: the rate of storm occurrence $\lambda(t)$, the mean number a of storm cells per unit area for a particular event, the mean total rainfall b^{-1} at the center of the cell, and the decay length c^{-1} of the spread function $h(r)$. To find characteristic parameters for the model from the Edwards Plateau study

region, daily-total rainfall accumulations from 23 rain gauges were used. This data spans a period of 48 years (from 1948 to 1996), and, though there were gaps in the data sequence for most of the gauges, there were on average 9 gauges reporting each day (see Figure 6.2 for the gauge locations).

An extreme rainfall event (storm) is defined as any period of three days with a total rain accumulation of more than 130 mm at one or more rain gauges [Smith and Karr, 1990]. This definition certainly includes the extreme rainfall events we are studying, since the accumulation measured for flood peaks above the median annual flood peak value for our basins all satisfy this criterion. We use the same definition of storms as Smith and Karr [1990] to allow comparison of the model parameter estimates computed for the central Appalachians region and Edwards Plateau. The storm definition was tested by comparing the annual flood peak data from the four stream gauging stations and the rain gauge data in the area of interest. The measured annual flood peak data for the majority of basins describes peaks that are orders of magnitude smaller than for the large floods we are interested in, as illustrated in Figure 6.3 for the stream gauging station at Guadalupe River at Comfort. In our framework, the small peaks correspond to the fact that there is a strictly positive probability of no extreme rainfall event occurring during a given year. We computed the total rainfall accumulation reported by the rain gauges for the three day period ending the day of the annual flood peak for each year. When we then exclude the

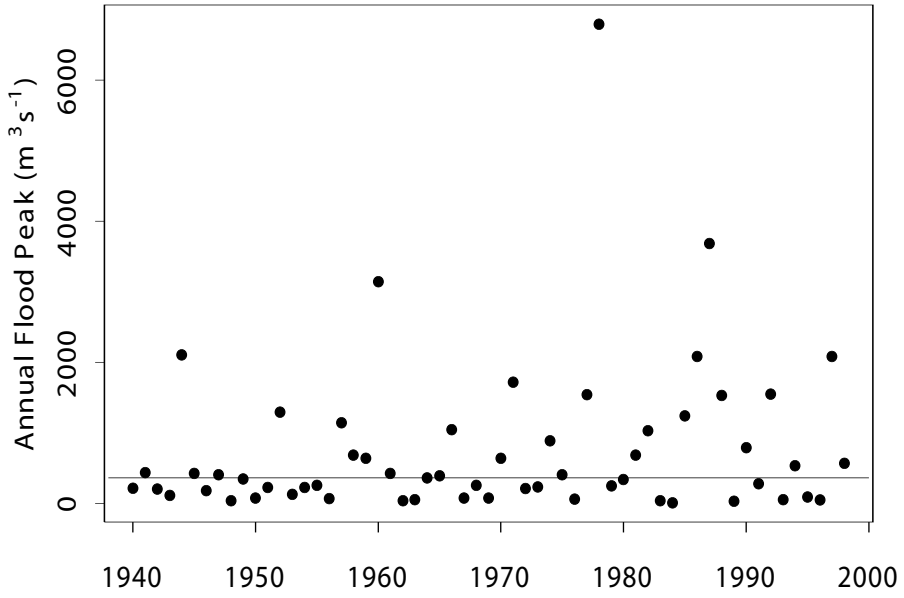

Figure 6.3: Measured annual flood peaks for the stream gauging station at Guadalupe River at Comfort. The line represents the median value.

peaks below the median flood peak values for each of the four basins, the minimum value of the recorded storm total accumulations is 125 mm.

We denote the total number of gauges in the region as ν, and the total number of years of observation as n. The storm arrival times are denoted T_i, and the total number of storms detected during the n years is N. For the ith storm, the total accumulation of rainfall S_{ij} for the jth gauge ($j = 1, \ldots, \nu$) is computed. The estimators derived in Smith and Karr [1990] can then be used to find approximate values for the parameters in our model.

In our notation, the estimator for the annual frequency Λ is

$$\hat{\Lambda} = \frac{N}{n}.$$

The estimated value for Λ from our data set is 1.53 storms/year. To conveniently express the seasonally-varying rate $\lambda(t)$, one can think of a year as a time interval from 0 to 1, where 0 corresponds to January 1 and 1 to December 31. If the arrival times T_i are expressed in years, then the mantissa of the arrival time $\{T_i\}$ indicates the time of year of the event. Then, for the seasonally-varying rate $\lambda(t)$, the window estimator

$$\hat{\lambda}(t) = \frac{1}{n}\frac{1}{\Delta t}\sum_{i=1}^{N} 1_{(t-\Delta t/2, t+\Delta t/2)}(\{T_i\}),$$

is used, where Δt is the width of the time window. The estimate for $\lambda(t)$, using a 5 month window [Smith and Karr, 1990] is shown in Figure 6.4. The seasonally-varying intensity varies from 0.6 storms per year in January to 2.65 in August.

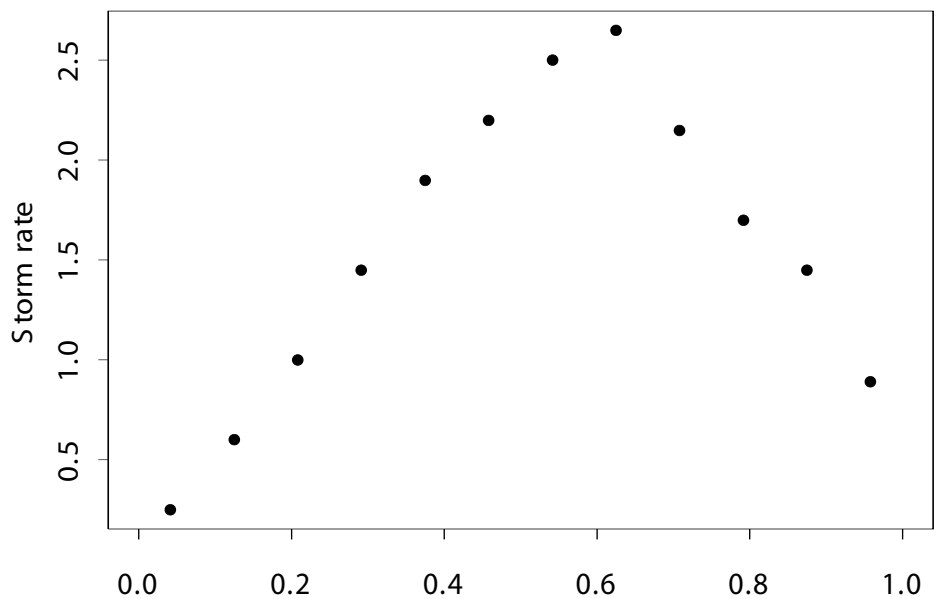

Figure 6.4: Estimate of the seasonally-varying rate of storm occurrence. Time index 0 corresponds to January 1, and time index 1 to December 31.

In order to estimate the spread function parameter c, estimates of pairwise correlation coefficients ρ of the storm totals between gauges are plotted against the distance between the gauges. Each estimate $\hat{\rho}$ was computed as a sample correlation coefficient for exactly 20 observations per pair of gauges. Pairs of rain gauges for which simultaneous data exists during fewer than 20 storms were excluded from the analyses to eliminate uncertainties in the parameter estimates. A least squares estimator for c is used: $\hat{c}$ is the number that minimizes

$$\sum \left(\hat{\rho}(r) - e^{-c^2 r^2}\right)^2,$$

giving an estimate of 30 km for c^{-1}.

The estimators for the parameters a and b of the model are obtained from the moment equations:

$$\hat{b} = \frac{\hat{\mu}}{\hat{\sigma^2}},$$
$$\hat{a} = \frac{2}{\pi}\frac{\hat{\mu}^2}{\hat{\sigma^2}}\hat{c}^2.$$

Estimates of μ and σ are readily computed from the data, giving estimates of 50 mm for b^{-1} and 0.0015 storm cells km^{-2} for a. The mean number of storm cells in the region of interest (area $\sim$ 5000 km^2) during an extreme rainfall event is then 7.5.

It is interesting to compare our model parameter estimates from the Edwards Plateau with those for the Appalachians as obtained by Smith and Karr [1990]. Both sets of estimates are listed in Table 6.1. We note that this region of Texas has approximately three times as many heavy storms per year as the Appalachian region.

Although each storm has fewer storm cells, on average, than do those in the Appalachians, each cell has higher rain accumulation, on average. Each comparison in the table agrees with the qualitative understanding of the climatic differences between the Edwards Plateau and the central Appalachians [Caracena and Fritch, 1983; Smith *et al.*, 1996].

Although analysis of daily rainfall accumulation is useful for estimation of storm totals, it does not provide information about the temporal evolution of the storm. To represent the temporal structure of the rainfall, a cascade model [Over and Gupta, 1996] was used. The cascade model used is outlined as follows: For each storm cell with total rainfall S_c at its center, we assign a lifetime τ such that the average rain rate $\bar{Z}$ at the center of the cell during the period τ is $\bar{Z} = S_c/\tau$. Next we divide the time interval $(0, \tau)$ into two halves. We assign a rain rate $W_{1,1}\bar{Z}$ to the interval $(0, \tau/2]$, and $W_{1,2}\bar{Z}$ to the interval $(\tau/2, \tau)$, where $W_{i,j}$ are independent random variables having the same distribution as a chosen positive random variable W. The expected amount of rain in every interval should be conserved, necessitating the condition $\mathbb{E}W = 1$. The next step is to divide each subinterval (each half) into two new intervals, giving four subintervals with length $\tau/4$. Again, we take independent positive random variables $W_{2,1}$, $W_{2,2}$, $W_{2,3}$, and $W_{2,4}$ with the same distribution as W, and assign rain rates $W_{2,1}W_{1,1}\bar{Z}$ to the first subinterval, $W_{2,2}W_{1,1}\bar{Z}$ to the second, $W_{2,3}W_{1,2}\bar{Z}$ to the third, and $W_{2,4}W_{1,2}\bar{Z}$ to the fourth. Continuing in this fashion, we will have 2^k

Parameter	Edwards Plateau (present study)	Central Appalachians Smith and Karr [1990]
Λ (storms/year)	1.53	0.45
a (cells per km^2)	0.0015	0.0027
b^{-1} (mm)	50	15 – 40
c^{-1} (km)	30	34

Table 6.1: Estimated values for the Storm Model parameters.

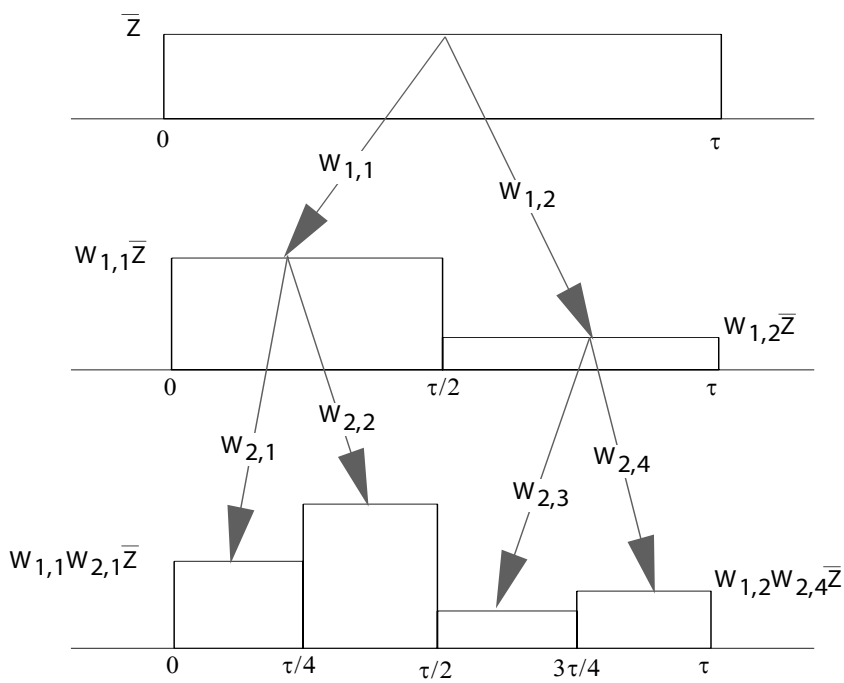

Figure 6.5: A schematic of the cascade model construction.

intervals at the kth stage, each with a length $\tau/2^k$ and rain rate $Z_c(t) = \bar{Z}\prod_{i=1}^{k} W_{i,j}$. An illustration of this construction is pictured in Figure 6.5.

The distribution in time of the storm total for a given storm cell is assumed to be independent of that for all other cells. Because storm cells represent different storm systems, it is natural to assume that they will evolve independently in time. In general, the temporal evolution of a storm cell might depend on changing meteorological conditions and interactions with other storm cells, but this is at a finer degree of detail than present in the spatial model of rainfall described above and the runoff model described below.

For our analyses the storm cell lifetime, τ, is taken to be characteristic of all the storms cells in the simulated storms and assumed to be a Gaussian random variable with mean 72 hours and standard deviation 3.0 hours (since we used 3-day rain gauge accumulations to obtain estimates of storm-totals, the long storm lifetime is required for the discharges to have reasonable values). k, the parameter corresponding to the number of cascades used, was varied from 4 to 10, giving a range of rainfall patterns averaged over a period of a few hours to patterns averaged every minute. The random variable W is taken to have a lognormal distribution [Over and Gupta, 1996], i.e., $W = e^{\gamma+\varsigma U}$, where U is a standard Gaussian random variable, and γ and ς are parameters of the distribution. The condition $\mathbb{E}W = 1$ then corresponds to the condition $\gamma = -\varsigma^2/2$. The parameter ς in our simulations was varied from 0.4 to 0.8.

To summarize, we simulate the rainfield from a storm as follows: M storm cell centers Y_j are simulated as a realization of a Poisson process on the plane with spatial rate a, and the storm length τ is taken from the Gaussian distribution. Each cell is assigned a total rain accumulation $S_{c,j}$ at its center obtained from an exponential distribution with parameter b. A cascade model is used for distributing $S_{c,j}$ over the interval $(0, \tau)$ to form the rain rate function $Z_{c,j}(t)$. A spread function $h(r)$ is used to describe the proportion of $S_{c,j}$ that falls at a point a distance r from the center of cell j. The rainfield $Z(x,t)$ at a point x is then

$$Z(x,t) = \sum_{j=1}^{M} h(\|x - Y_j\|) Z_{c,j}(t).$$

6.2 The Simplified Network Model

In the drainage network runoff model (the Network Model), the discharge at the outlet of the basin is realized as an extension of the inverse GIUH function (geomorphological instantaneous unit hydrograph, defined as the basin's response to instantaneous rainfall with a unit magnitude; [Rodriguez-Iturbe and Rinaldo, 1998, p. 477]). For this study, we used the simplified version of the Network Model described below in order to isolate the network properties of the river basin. A description of the full version of the Network Model can be found in Appenndix A.

We represent the river network as a binary tree with n links. We will denote an individual link by e. Each link has a certain length l_e, an upper vertex $\bar{e}$, and a lower vertex $\underline{e}$. We know the position and geometry of every link, and the way that links

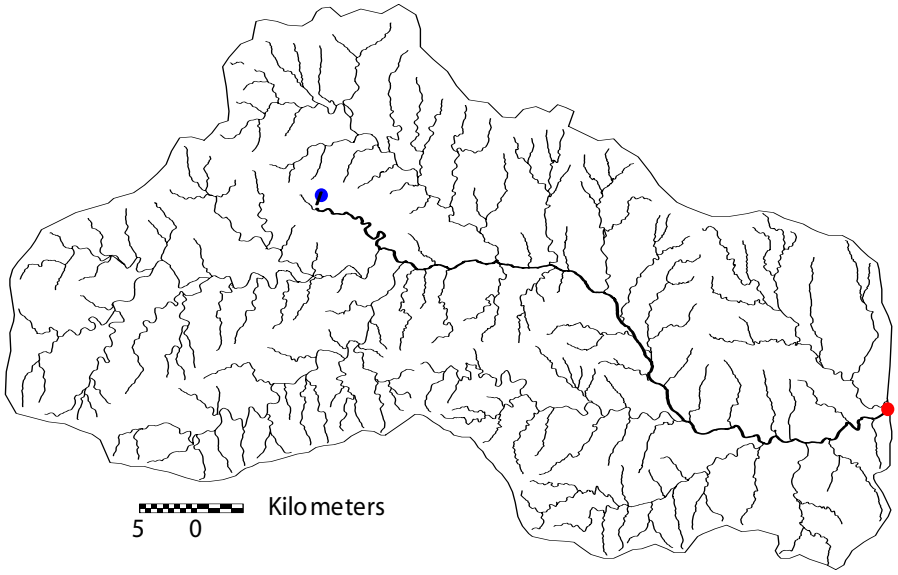

Figure 6.6: The network structure of the Guadalupe river. The bold line shows the path of the water from the marked point in the basin to the outlet.

are connected (i.e., we know the structure of a network similar to the one represented in Figure 6.6).

We make the assumption that water flows to the nearest channel following the path of steepest descent on the underlying elevation, from every point in the basin. After reaching the channel, the water follows the river network to the outlet. Under these assumptions, we can compute the travel distance $f_{\underline{e}}(x)$ to the outlet $\underline{e}$ for each point x in the basin, and, assuming a constant flow velocity v both over the land and in the channels, the discharge $Q_{\underline{e}}$ at the outlet is given by:

$$Q_{\underline{e}}(t) = \int_{\mathcal{B}} R\left(x, t - \frac{f_{\underline{e}}(x)}{v}\right) dx, \tag{6.2}$$

where $\mathcal{B}$ denotes that the integral is taken over the basin and $R(x,t)$ denotes the runoff generated at location x at time t. Note that if $R(x,t)$ is nonzero only for a certain time period $(0,T)$, then the net discharge at the outlet will be nonzero at most during the period $(0,T_0)$, with $T_0 = T + \max f_{\underline{e}}(x)/v$.

To approximate the effects of infiltration, the runoff field $R(x,t)$ is computed from the rainfall field $Z(x,t)$ by

$$R(x,t) = (Z(x,t) - K_s)^+, \tag{6.3}$$

where K_s is a parameter similar to the saturated hydraulic conductivity, and $y^+ = \max(0,y)$. We assume that K_s is the same throughout the basin. This representation of the infiltration process is reasonable for extreme rainfall events [Ogden and Saghafian, 1997].

This model provides a useful tool for studying drainage basin response to extreme rainfall. It has only two parameters: the average surface flow velocity v, and the value of K_s. We can estimate the values of these parameters from reconstructed hydrographs for extreme events for which we have both good rainfall measurements and discharge measurements at one or more nodes. The assumption of constant velocity in the channels for extreme events has been discussed previously [Gupta and Waymire, 1998]. Analyses of measured velocity in different channels suggests that the velocity is effectively the same if the respective resultant discharge levels in the channels have the same probability of occurrence [Rodriguez-Iturbe and Rinaldo, 1998, p. 14].

We have implemented this model for several extreme events in Texas and other areas [Smith *et al.*, 2000, 2001], and achieved acceptable reconstruction of discharge hydrographs. For the purposes of our simulations, we varied the flow velocity from 0.6 to 3.0 m s^{-1}, and K_s from 0 to 20 mm h^{-1}.

6.3 Scaling Analyses

We simulated flood peaks for 100 non-nested sub-basins of the Guadalupe river basin in the Edwards Plateau for a period of 50 years (see Figure 6.2 for the location of the Guadalupe river basin). For the the river network structure, we used EPA RF3 river reach data (Figure 6.6), which corresponds to the blue lines on USGS topographical maps. Rainfall was simulated according to the Storm Model described in section 6.1,

and the ensuing runoff for each sub-basin was computed using the Network Runoff Model described in section 6.2.

For each sub-basin, the annual flood peaks were identified as the maximum computed discharge from the simulated storms in that year. Since extreme rainfall events form a Poisson process on the time axis, there is a nonzero probability of obtaining zero events in a given year. For years with zero events, the annual flood peak was taken to be zero. As was noted earlier, this is justified by the fact that the measured annual flood peak data for most basins includes peaks that are orders of magnitude smaller than the large flood peaks of interest (Figure 6.3). The computations were repeated for different values of the Storm Model time cascade parameters, K_s value, and channel velocity, and different catchment sampling (choice of sub-basins). The dependence of the flood peaks behavior on the values of the Network Model parameters was not statistically significant.

The first and second sample moments of the simulated flood peak data set for one set of sampled sub-basins are plotted against sub-basin area in Figure 6.7 (for $v = 1$ m s^{-1}, $K_s = 0.5$ mm h^{-1}, $\varsigma = 0.6$, and $k = 8$). The logarithms of the moments clearly exhibit a linear relationship with the logarithm of the sub-basin drainage area. The slope of the linear regression for the sample means is 0.72, which is consistent with typical slopes for the observed data [Gupta and Waymire, 1998; Robinson and Sivapalan, 1997; Morrison and Smith, 2001b]. Means of annual flood peaks computed from measurements at several stream gauging stations in our area of interest are also

shown in Figure 6.7 (squares). There are no small gauged basins in the area of interest with a long (> 30 years) record of maximum annual flood peaks.

The dependence of the coefficients of variation on the sub-basin drainage area is shown in Figure 6.8 (triangles). The overall shape of this dependence is similar to that obtained from an empirical study of actual flood peak data (Figure 6.1); there tends to be an increase of CV with increasing drainage basin area for catchments with areas smaller than $A_c = 10$ km^2, and a decrease of CV with increasing drainage basin area for larger catchments. Similar results were obtained for all input parameters used in the simulation models. It is interesting to note that the CVs computed from annual flood peak data from several stream gauging stations in our area of interest is of the same order of magnitude (Figure 6.1, squares) as the simulated results. The inference that the behavior of CVs for annual flood peaks can be largely explained in terms of the spatial and temporal properties of the rainfall and the structure of the river networks is supported by the observation that only these three elements were included in our simulations, which reproduced the expected CV behavior

To gain further understand the factors that influence the behavior of CVs, the simulation experiments were repeated with a stationary rainfall field, i.e., the time cascade construction was removed from the Storm Model. This simulation was performed in an attempt to isolate the effects of the spatial properties of the rainfall and the network on CV. CVs obtained without the cascade structure are also shown in Figure 6.8 (empty circles). They were computed with the same rainfall field spatial

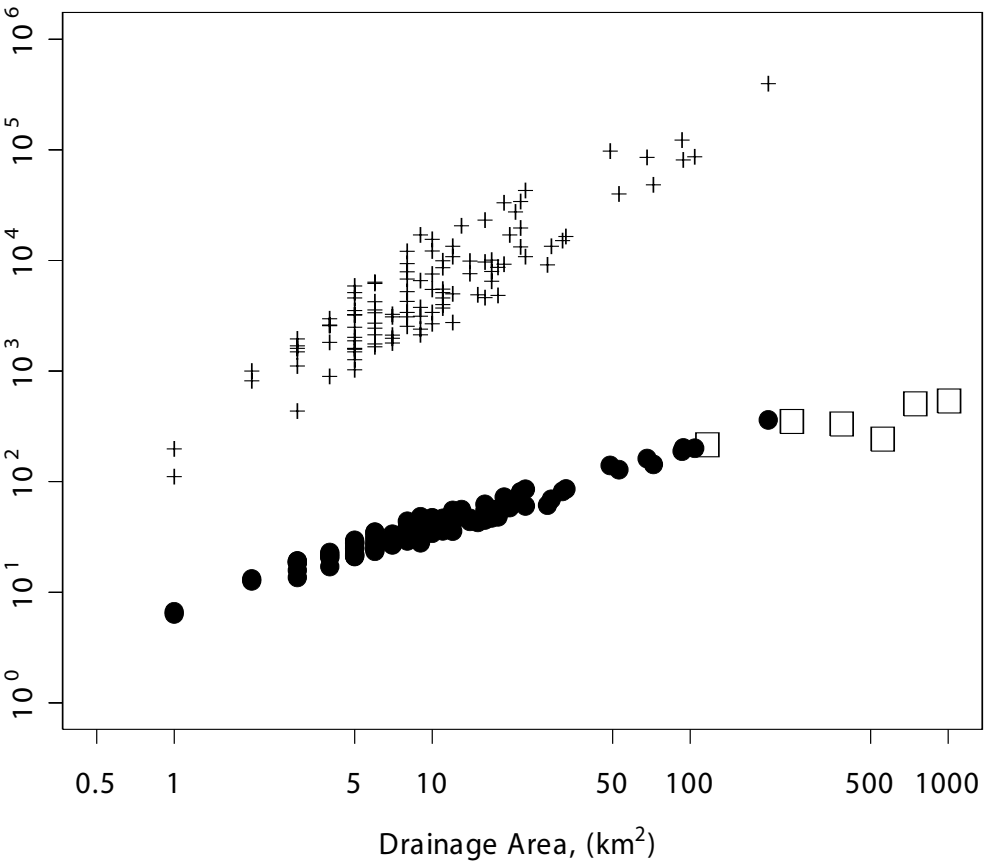

Figure 6.7: Log-log plot of first (solid circles) and second (crosses) sample moments of the simulated flood peaks versus drainage area. Empty squares represent estimated values of mean annual flood peak for actual basins in the area of interest. The units of the y axes are m^3s^{-1} for the first moments and m^6s^{-2} for the second. The parameters used to obtain this plot were: $v = 1$ m s^{-1}, $K_s = 0.5$ mm h^{-1}, $\varsigma = 0.6$, and $k = 8$.

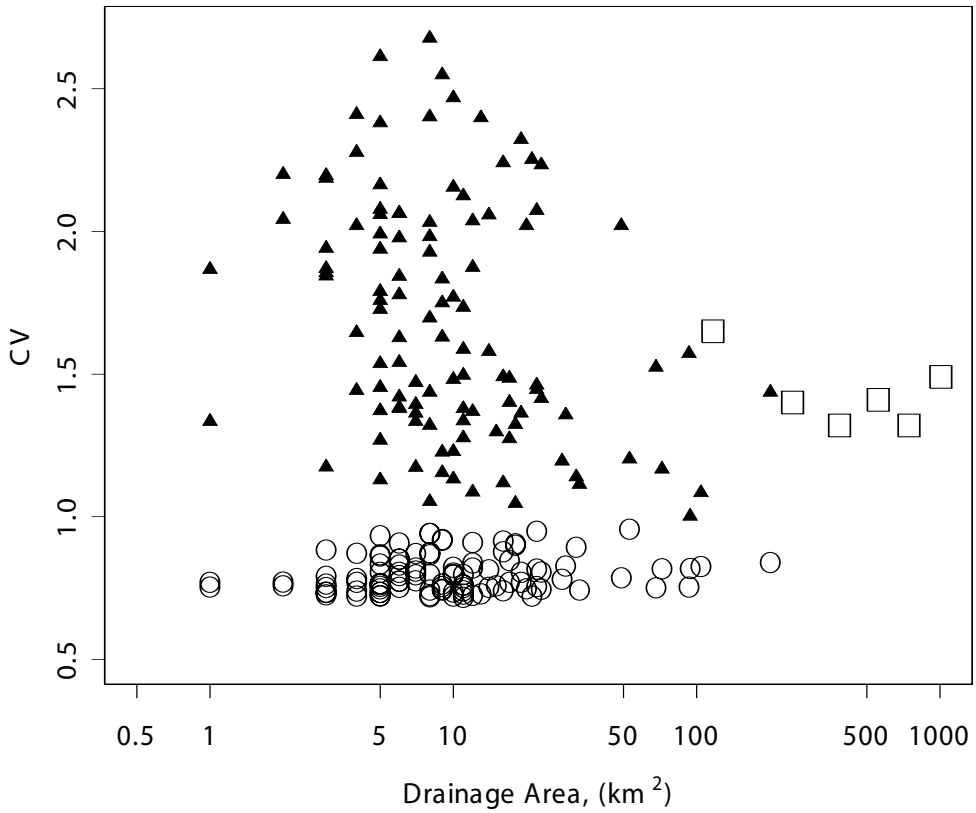

Figure 6.8: CV – Drainage area relationship for the simulated flood peaks. Triangles indicate values obtained with the rainfall simulated with the Storm Model, empty circles indicate values where the rainfall was simulated without the time cascade structure. Squares represent estimated values for CV for actual basins in the area of interest. The parameters of the models are the same as in Figure 6.7.

structure as the CVs obtained with the cascade structure (triangles) for each simulated event. The areal dependence of the prior analyses is no longer present. Because similar results were obtained for a wide range of simulation parameters, we infer that the temporal structure of the simulated rainfall is important for producing the small area behavior of CVs in our results.

There are multiple explanations for the change in low-area scaling in the context of our model. Because small basins have much shorter response times than larger basins, their instantaneous discharges depend on the rainfall during a shorter period of time than do larger basins' discharges. For the runoff calculated using the Network Model, this effect is readily apparent, since the discharge $Q_{\underline{e}}(t)$, for each link $\underline{e}$ at time t, is calculated from the rainfall during the time interval $[t - \max(f_{\underline{e}}(x)/v),\ t]$. Moreover, the form of Equation (6.2) is similar to that of a resampling and averaging of the rainfall. A plot of $\max(f_{\underline{e}}(x)/v)$ as a function of basin drainage area is shown in Figure 6.9 for the sub-basins of the Guadalupe river (assuming $v = 1.0$ m/s). We see that the largest transit time scales are roughly ~ 1 hour for small basins and 30 hours for large basins. In the simulation experiments, the rainrates generated by the Storm Model generally changed every 10 minutes, so small catchments experienced little averaging of the temporal structure. This plays a role in determining the importance of the rainfall temporal structure for small basin CVs, and especially for the discharges calculated using the Network Model.

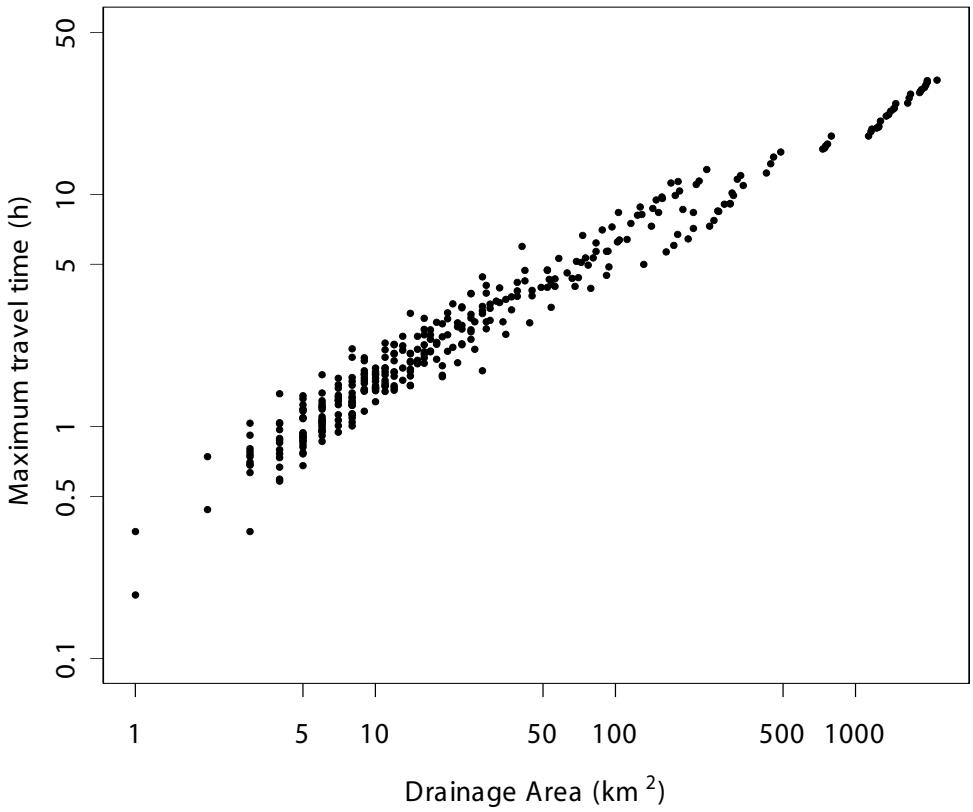

Figure 6.9: In the Network Runoff Model, the time scale of a basin response is $\max(f_e(x)/v)$. Values of $\max(f_{\underline{e}}(x)/v)$ are plotted as a function of drainage area for $v = 1\text{m s}^{-1}$.

The spatial structure of the simulated rainfall might also be suspected as the explanation for the low-area dependence. The two parameters that determine the spatial scale of the Storm Model are a and c. According to the estimates from Table 6.1, the average distance between storm cells is $a^{-1/2} = 26$ km and the characteristic scale of a single storm cell is $c^{-1} = 30$ km. These scales are somewhat larger than the resolution observable by the rain gauge network (the average distance between a particular rain gauge and its nearest neighbor was about 12 km), so the rain gauge network adequately captures the spatial structure considered. However, the error in the estimation of the parameters a and c is large (see Bloschl [1999]), and the CV results are dependent on the estimated characteristic spatial scale of the simulated rainfall.

To address the issue of the uncertainty in determining the spatial scale of the rainfall patterns, we simulated rainfall with several different values of a and c. The results are presented in Figure 6.10, which shows the envelope curves for four different simulation experiments with unique values of the parameters a and c. The shapes of these curves are all similar to that of the real CVs (Figure 6.1)). While the particular estimates of the spatial structure might be imprecise, the scaling behavior of the simulated CVs does not change significantly.

Many authors have stressed the importance of the network structure of the basin in determining the behavior of CVs for small catchments [Gupta and Waymire, 1990; Gupta and Dawdy, 1995; Robinson and Sivapalan, 1995]. In this study, we used

a network taken from an actual basin, so realistic basin network properties were employed. In order to further understand the importance of employing actual river networks, though, we tested the same rainfall model with the Peano network.

The Peano network is an idealized model of a channel network [Marani *et al.*, 1991; Gupta and Waymire, 1998; Gupta *et al.*, 1996]. An example of a Peano network and a description of its basic properties can be found in Rodriguez-Iturbe and Rinaldo [1998, p. 123], and a Peano network of order 5 is illustrated in Figure 6.11. Under the assumption of constant rainfall and constant water velocity in streams, the peak discharge in a Peano network exhibits simple scaling [Gupta and Waymire, 1998]. The scaling of flood peaks in a Peano basin with the spatial cascade model for rainfall has been studied both analytically and computationally by Gupta *et al.* [1996], and it has been concluded that the peaks exhibit multiscaling in this case.

We studied the response of a Peano basin to rainfall generated by the Storm Model through simulations, and discharges at the end of each link were computed using the Network Model. The area of the entire Peano basin was taken to be approximately the same as the area of the Guadalupe River at Comfort basin (2190 km^2), and the order of the Peano basin was taken to be 5 (Figure 6.11) so the drainage density is approximately the same as for the Guadalupe river basin. Figure 6.12 shows the flood peak CVs obtained with the same set of model parameters as used for Figures 6.7 and 6.8. It is clearly seen that CVs decrease with increasing drainage area for all scales. Although it is difficult to study the distribution of flood peaks with our

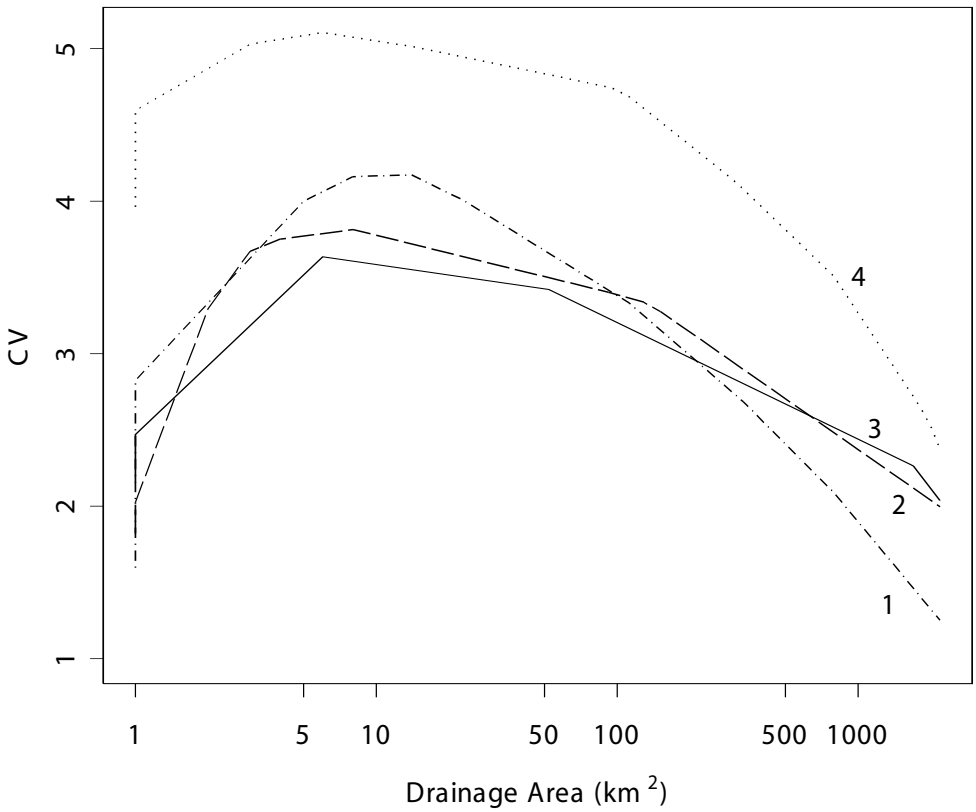

Figure 6.10: The envelope curves for the CV-area relationship resulting from simulated rainfall with different spatial scales. (1) $a^{-1/2} = 60$, $c^{-1} = 60$; (2) $a^{-1/2} = 30$, $c^{-1} = 15$; (3) $a^{-1/2} = 15$, $c^{-1} = 15$; (4) $a^{-1/2} = 15$, $c^{-1} = 30$ (km).

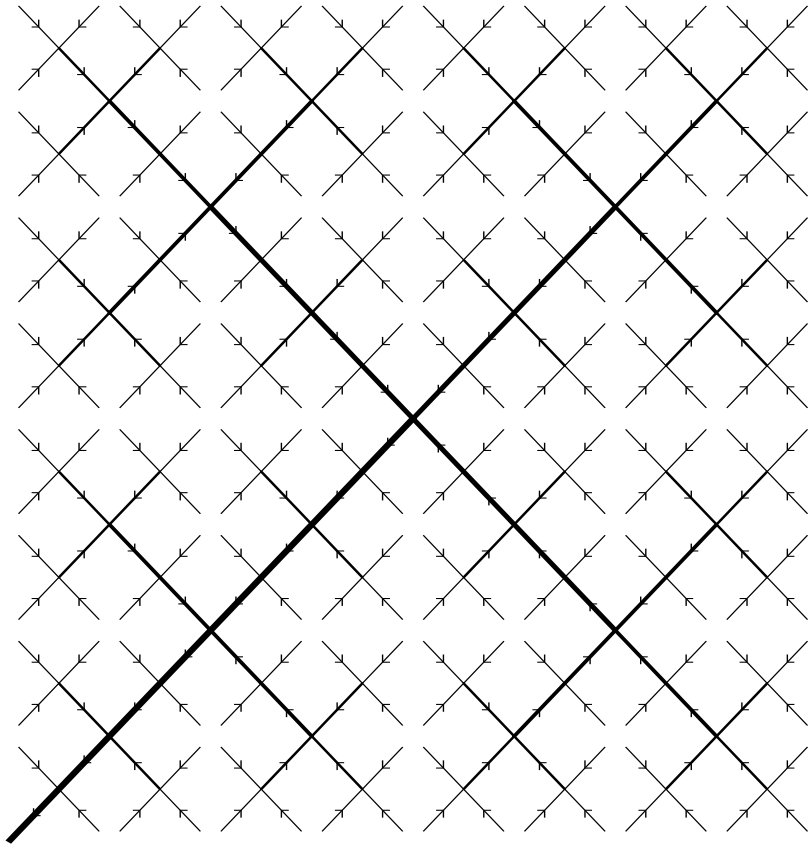

Figure 6.11: The Peano network of order 5.

rainfall model even in the Peano basin, this simulation experiment supports the hypothesis of multiscaling. The simulation experiment with a Peano basin shows that the network structure of the drainage basin is an important factor for determining the scaling behavior of flood peaks CVs, which agrees with previous findings [Gupta and Waymire, 1990; Gupta and Dawdy, 1995; Robinson and Sivapalan, 1995].

Finally, we observe that our results provide a starting point from which to further test additional influences on the behavior of CVs. Much research has been performed to determine the features in the basins' morphology, dynamic response, and rainfall that produce the greatest effects on the behavior of flood peaks [Robinson and Sivapalan, 1997; Bloschl and Sivapalan, 1997; Gupta and Dawdy, 1995]. It has been suggested that the basin response to rainfall on a small scale is largely controlled by the spatial variability of basin's morphological properties [Wood *et al.*, 1988], which are therefore critical for the scaling of annual flood peaks. It has also been hypothesized that the transfer of water from hill slopes to channels may be the key process defining the characteristic spatial scales of regional hydrology [Robinson *et al.*, 1995]. Another hypothesis that nonlinearities in runoff generation contribute to the increase in CVs for small scales [Bloschl and Sivapalan, 1997] is supported by the effect of the temporal structure of the model rainfall on the small scale behavior of CVs that was observed in this study. This study did not directly model any of the processes mentioned above. That our results reflect the observed behavior of CVs does not rule out any of these additional influences, but merely demonstrates that the overall

response of the real system can be mimicked by the fundamental response of a much less complex system that we can tractably analyze.

It is further encouraging to note that many of the additional processes not included in our results are quickly coming within the reach of computational models due to continuing increases in computing power, while they would have been prohibitively computationally intensive only two years ago.

6.4 Summary of Chapter 6

There are three primary results from this chapter:

- The simulation of rainfall over an area coupled with the Network Model provides a useful tool for studying flood peaks. Estimation of most of the parameters for such simulations can be done based on analysis of daily-accumulation rain gauge data.
- The areal scaling properties of flood peaks (log-log linearity of moments vs. basin drainage area and the peculiar behavior of the coefficients of variation) can be explained by the spatial and temporal structure of the rainfall and the spatial structure of the stream network.
- There is evidence that the interaction between the temporal structure of the rainfall and the network structure of the basin might be important to the scaling

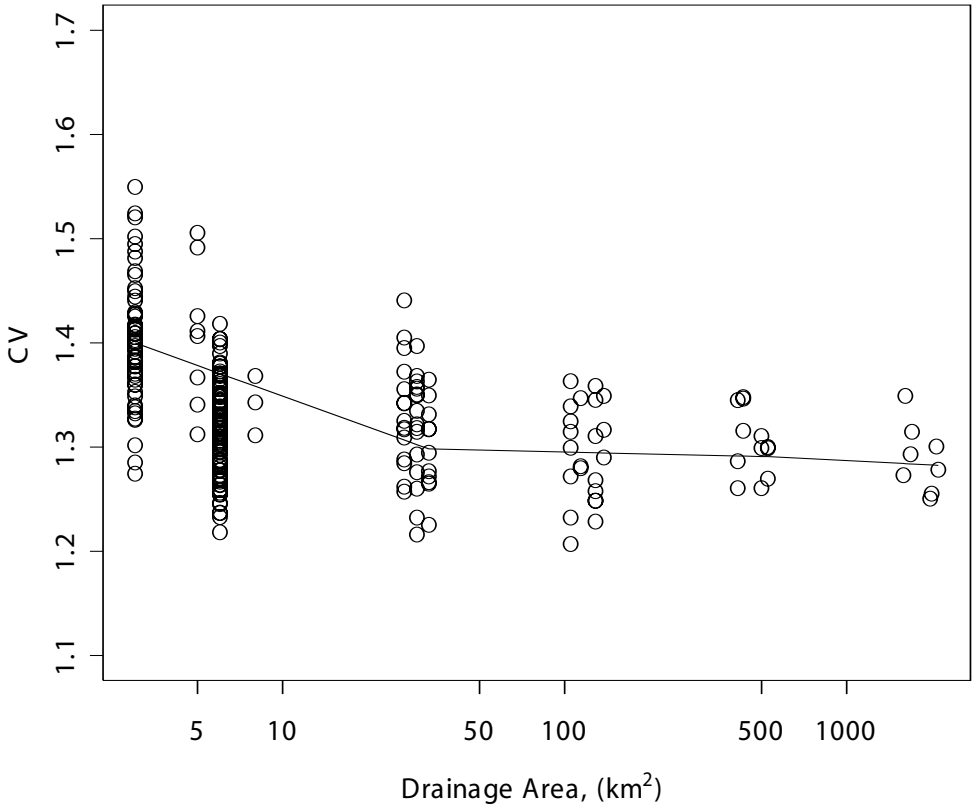

Figure 6.12: CVs of annual flood peak simulated in the sub-basins of Peano basin. Parameters of the models are as in Figure 6.7.

behavior of the flood peak coefficients of variation for basins with a sufficiently small area.

Chapter 7

Conclusions and Future Research

The Generalized Extreme Value distribution (GEV) has been widely used for modeling extreme events, such as flood, winds, and rainfall, and estimation of its parameters is an interesting problem that has been tackled by a number of authors. In this dissertation, we have developed methods to improve estimates of the GEV's shape parameter, since it is the parameter that determines the thickness of the distribution's tail. We have introduced MIXed parameter estimation methods based on a combination of the standard Maximum Likelihood and method of L-moments, and shown, that the new methods have smaller parameter estimates RMSE for small sample sizes and certain values of k (≤ -0.2). We have seen that the MIXed methods perform roughly as well in quantile estimation as the method of L-moments, and showed that the bias of the L-moments estimates is favorable and necessary for good quantile estimates from this method.

One of the advantages of the MIXed methods is that they retain the flexibility of of maximum likelihood methods towards incorporating additional information, such as a known prior for the distribution parameters, or a positive density at some point. Such information, can significantly improve the quantile estimates of the distribution. We argued that MLE can be conditioned by any "good" U-statistics and produce asymptotically consistent and jointly Gaussian estimates. It would be interesting to investigate whether conditioning not by L-moments but by other statistics can produce better estimates than the new MIXed methods. It is possible that different constraints on the MLE problem will produce the optimal parameter estimates at different finite samples sizes.

Incorporating additional information into statistical parameter estimation is probably one of the best ways to improve the estimation procedure. For this reason we extended the MIXed estimators to incorporate PDS information, for application to the modeling of flood peaks. We have studied the performance of the MLE2 method which employs both AMS and PDS records. One of the benefits of this method is that it uses the longest period of record available for a stream gauging station, and still includes information additional to the AMS record.

Under the assumption that annual flood peaks for a given basin have a GEV distribution, the proposed methods were used to estimate the GEV distribution's parameters for a sample of basins in the central Appalachians region, and the relationship between morphological basin descriptors and parameter estimates was investigated.

The estimates of the shape parameter k did not exhibit significant dependence on any of the basin descriptors considered. Moreover, the variability of the shape parameter k corresponds to the variability of the estimators used. As was expected, the logarithms of the scale and location parameter estimates exhibit a linear dependence with the logarithm of the basin area. It was argued that basins with higher percentage of impervious cover should be considered as though they have larger effective contributing area. This finding is related to the premises of scaling theories that argue that the *scale* in hydrology may be defined not just by a basin area, but as some function of the basin morphological descriptors (see Gupta and Waymire [1998]). How this function can be determined and what basin descriptors provide the most significant contribution to that function are subjects for further study. Our analysis suggests that besides drainage area, the percentage of urbanized cover makes an appreciable contribution to the basin scale definition. The slope of a basin has also been suggested to be important for determining the basin scale parameter [Madsen *et al.*, 1997b]. We did not find any significant dependence of the GEV estimates on the basin slope, but one could use different definitions for the basin slope. It could be the case that the slope has to couple with another parameter that we did not investigate (that was not well represented in our data) for the parameters to exhibit a marked dependence on the slope. It is also possible that our basin sample was not simply representative enough in terms of slope, so that we could not detect the influence of the basin slope on the parameter estimates. More data analyses are needed in this area.

In Chapters 3 – 5, we explained and illustrated some techniques for modeling bivariate distributions with copulas. We briefly discussed how errors in the estimation of bivariate distribution marginals affect the statistical inference of copulas. From Monte Carlo simulations, it appeared that the increase in the error of copula parameter estimates arising from the use of nonideal marginals is relatively small. Our simulation results with Gumbel copulas and GEV marginals suggests that parametrically estimated marginals work as well as the empirical ones for sample sizes ≤ 250. Are these results an artifact of the particular copulas and marginals that we have considered, and is there a theoretical explanation for these observation? These type of questions are very important for developing robust statistical techniques involving copula estimation.

An interesting question not considered in this dissertation is how confidence intervals for copulas and bivariate distributions can be constructed. Though this issue has been raised in the literature (e.g Genest and Rivest [1993]; Capéraà *et al.* [1997a]; Joe [1997]), the methods considered involved the values of the asymptotic covariance matrix of the estimators. For the univariate GEV distribution, the asymptotic covariance matrix very poorly describes the behavior of parameter estimators for small samples (see Prescott and Walden [1983]; Lu and Stedinger [1992]), so it is unlikely that variances of parameter estimates derived from asymptotic analysis techniques are applicable to bivariate GEV distributions, at the very least. Extensive simulation

studies in this area are needed in order to understand the behavior of the different copula estimators. Such simulations can be readily carried out using EVANESCE.

We have illustrated how copula-based models provide good fits to bivariate distributions of variables of interest to financial engineers. We demonstrated how using the functions available with EVANESCE simplifies estimations of marginal distributions, parametric and non-parametric copula estimation, comparison of how well different parametric copulas fit the data, and the use of simulations to replicate the available data sample. Such routines are quite valuable for risk analysis. However, financial engineers are very interested in modeling multivariate random variables, with the typical numbers of dimensions of interest from 50 to 100. Though a number of copula families can be generalized to n dimensions, such generalizations, apart from the Gaussian copula, are of little practical use. Even though one can estimate parameters for n dimensional copulas, there are too few methods developed for simulating random observations from these copulas without inverting $n-1$ conditional distribution functions. The copulas for which such algorithms exist (e.g., the Clayton copula, [Clayton, 1978]), have too few parameters to fully describe the multi-dimensional structure of the data (see Johnson [1987]. For example, the Clayton copula has only one parameter, so all its bivariate marginals are the same, and the multi-dimensional structure is fixed). It is improbable that multi-dimensional copula-based modeling will be widely used in practice as an alternative to using multivariate Gaussian cop-

ula, unless either radically new simulation techniques are developed or the increase in computing power makes the currently available techniques practical.

In Chapter 5, we discussed possible approaches to copula-based modeling in hydrology. We argued that for modeling the joint distribution of flood peaks and flood volumes in a PDS framework, one should use a particular type of copula with an implied bivariate distribution of annual flood peaks and volumes that is a bivariate GEV distribution. This BGEV distribution can be considered to be a starting point for the modeling of bivariate extremes. The new bivariate PDS approach is used to model the joint distribution of flood peaks and volumes for Potomac at Point of Rocks, and the joint distribution of flood peaks for Potomac at Point of Rocks and South Fork Shenandoah River at Front Royal, its subbasin.

Theoretically, the PDS copula introduced in Chapter 5 allows extensions into higher dimensions (see Joe [1997], Chapter 6, for construction of multivariate EV copulas). The applications of such extensions would be the modeling of the joint distribution of flood peaks in several closely situated basins. However, since the PDS copula has a singular component, the simulation of random observations from this copula even in three dimensions constitutes an unsolved problem. Special methods have to be designed for such simulations.

In Chapter 6, we discussed the scaling properties of flood peaks. Using a stochastic model for extreme rainfall events (i.e., those that could potentially cause the annual flood peak in a basin) to provide input for a rainfall-runoff model, we simu-

lated discharges in all of river's streams. We call this approach a coupled stochastic rainfall-runoff model. The results of this work were rather promising: we were able to reproduce the peculiar scaling properties of the moments and CVs of the flood peaks. Moreover, the range of the simulated flood peaks agreed with that of those measured by USGS, and the slopes of the mean annual flood peaks agreed with those noted in the literature. The results of this study show that appropriately designed simulations of space-time rainfall can be very useful tool for studying the distribution of annual flood peaks, so further development of coupled stochastic rainfall-runoff models is an interesting avenue of research. After the simulation study described in Chapter 6 was concluded [Morrison and Smith, 2001a], the Network Model underwent a series of improvements (see Appendix A), and its modeling capabilities were extensively studied [Smith *et al.*, 2000, 2001; Sturdevant-Rees *et al.*, 2001; Zhang *et al.*, 2001; Turner-Gillespie, 2001]. In our study, the spatial structure of the rainfall was rather primitive, and the temporal structure was assumed rather than estimated. Coupling the Network Model with one of the more sophisticated space-time rainfall models (e.g. Rodriguez-Iturbe *et al.* [1987]; Northrop [1998]) might allow reproduction of the observed distribution of annual flood peaks for gauged basins, and this distribution for ungauged basins could be inferred from the model.

Appendix A

The Network Model

A.1 Introduction

A.1.1 History

The development of the Network Model began in 1998 with a set of routines for the computation of a number of river network descriptors, including the width function and Geomorphological Instantaneous Unit Hydrograph (GIUH) function (see Rodriguez-Iturbe and Rinaldo [1998], p. 477). Discharge at an outlet was computed as a convolution of the GIUH inverse function with a rainfall field (see section 6.2). The Model was significantly revised in 1999 with the addition of DEM (Digital Elevation Model) routing, Green and Ampt infiltration with Redistribution, impervious land cover, rainfall interpolation in space and time, and the definition of Network-Model-specific data storage formats. The current version (3.0) is enhanced with a fast routing option, better display capabilities, simplified input parameters and the ability to read files in standard GIS formats (ARC-INFO GRID and vector coverages).

Simulation of retention depth was added to the model by Daniel Turner-Gillespie [Turner-Gillespie, 2001].

A.1.2 Description

The Network Model is a distributed raster (grid-based) hydrologic model which simulates the hydrologic response of a watershed to a given rainfall field, under the assumption that all water velocities are explicitly specified. The user is required to specify the rainfall field, DEM, river network, and soil maps as inputs. The program then computes the amount of infiltration and generated runoff, and routes the water on the DEM surface and through the river network (Figure A.1).

Rainfall can be either explicitly input from radar observations, assumed to be constant, or simulated using a specified space-time stochastic rainfall model. Rainfall infiltration can be modeled with a Green and Ampt with Redistribution (GAR) infiltration method [Ogden and Saghafian, 1997] or using a simpler equation (6.3). Information about impervious land cover may be added to make the computed infiltration more realistic in an urban environment. Runoff is determined as the difference between the rainfall and the infiltrated water, and is routed on a surface specified by the underlying DEM. The flow direction is determined using the enhanced Tarboton's $D\infty$ method [Tarboton, 1997] and the specified flow velocity. The transition of the runoff from the surface into the streams may be modeled using one of the two included empirical rules. Stream routing is performed using the specified velocity and known network geometry. The discharge is computed at each node in the network,

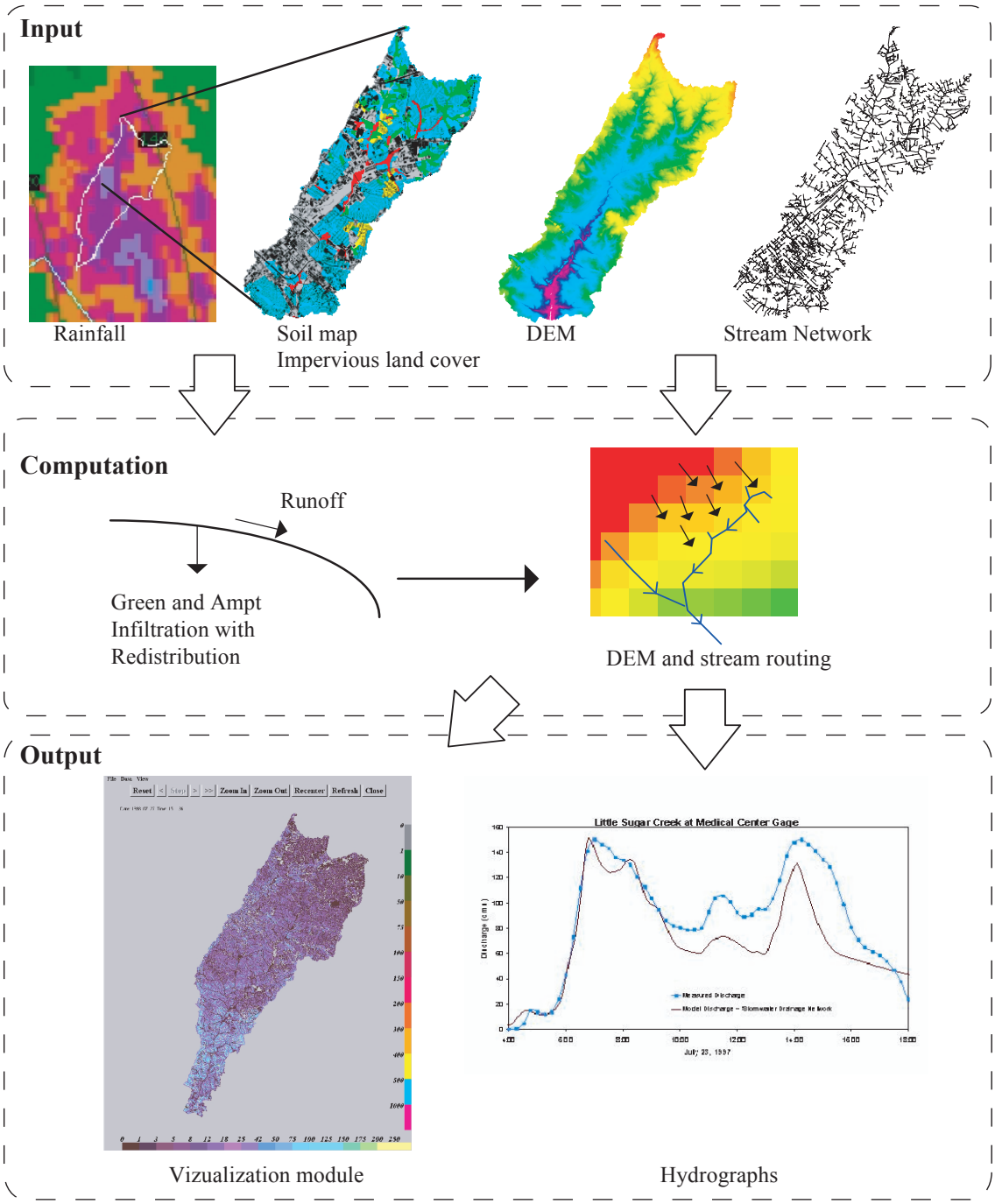

Figure A.1: The schematic of the Network Model.

and may be output for further analyses. Optionally, the output can be viewed with model visualization module (Figure A.1).

The major distinction between the Network Model and other raster-based distributed hydrologic models is that the water velocities must be specified explicitly (either constant or a power function of the local slope and/or water discharge), and not determined as the solution of a partial differential equation (PDE), as done in CASC2D [Ogden and Julien, 1993], for example. Though the solutions of PDEs provide better approximations of the physical processes involved in runoff routing than explicit formulas in theory, the PDEs themselves involve a number of generally unknown parameters. For example, the open channel flow equation that is often used to route water in streams requires knowledge of the channel cross-section profile everywhere in the network [Fred, 1993]. Generally, this information is not available for any particular watershed. Specifying the channel flow velocity explicitly, thusly, may provide as good a model of the flow as guessing the channel profile at every point. For a particular watershed, the computational time for the Network Model is much less than that of PDE-based models, and allows fast calibration. It is also free of numerical problems in the water routing, unlike its PDE-based analogues.

The Network Model is written in java. This language was chosen primarily for its platform-independent features, which are important for the Model visualization module. The model greatly exploits the object-oriented nature of java, allowing a user to specify which methods he or she prefers to use in different components

of the model. For example, a user may choose to substitute radar observations of rainfall with rainfall generated using a specified space-time stochastic rainfall model by redefining one object, and, independently, choose between different infiltration methods by changing another object. The model is distributed with an open source code, allowing users to substitute their own alternative methods at every step of computations. The full documentation of the java classes comprising the Network Model is distributed with the Model.

The Network Model has been successfully implemented in a number of hydrological studies (see Smith *et al.* [2000, 2001]; Sturdevant-Rees *et al.* [2001]; Zhang *et al.* [2001]; Turner-Gillespie [2001]).

In the following sections, we describe the principles and computations involved in several of the Model's components in a greater detail, and provide a User Manual.

A.2 Network Model Principal Components

The Network Model is grid-based, and the main input into the model is the DEM (Digital Elevation Model) of a watershed. The pixels of the DEM are often referred to as *grid cells*, or just *cells*. It is assumed that all DEM cells are square and have the same size.

In the Network Model, time is discretized into equal steps, such that points in time can be represented as $t_0, t_1, t_2, \ldots$, where $t_{j+1} - t_j = s$ is the constant time step. Figure A.2 is a schematic for the computations performed during one time step in the

model. As seen from this schematic the following procedures are are performed at each time step: (1) rainfall computation (either spatial/temporal interpolation of the rainfall determined from radar observations or generation of rainfall using a stochastic space-time rainfall model), (2) infiltration of the rainfall into the soil, (3) computation of runoff, (4) routing of the runoff over the DEM, (5) computation of how much water transfers into the streams, and (6) routing of the water in the streams.

If rainfall from radar observations is used, it is linearly interpolated in time. In space, it is interpolated either using the nearest neighbor approach (i.e. the rainfall at a particular cell in the DEM is assumed to be equal to that observed at the nearest radar bin center), or by linear interpolation in space.

We discuss the particular methods used to compute infiltration and DEM routing in a greater detail below.

A.2.1 Green and Ampt Infiltration with Redistribution

The optional Green and Ampt infiltration with Redistribution (GAR) routine in the Network Model is based upon the procedure suggested by Ogden and Saghafian [1997] for modeling infiltration. The modifications to this procedure in this implementation mainly address numerical problems stemming from fluctuations in rainfall rates.

In the Network Model, we consider the rainfall $r(x, t_j)$ at time t_j to be applied during the period $(t_j - s/2, t_j + s/2)$, and we are interested in the runoff $u(x, t_j)$ generated from each grid cell x during that period of time. The GAR model is

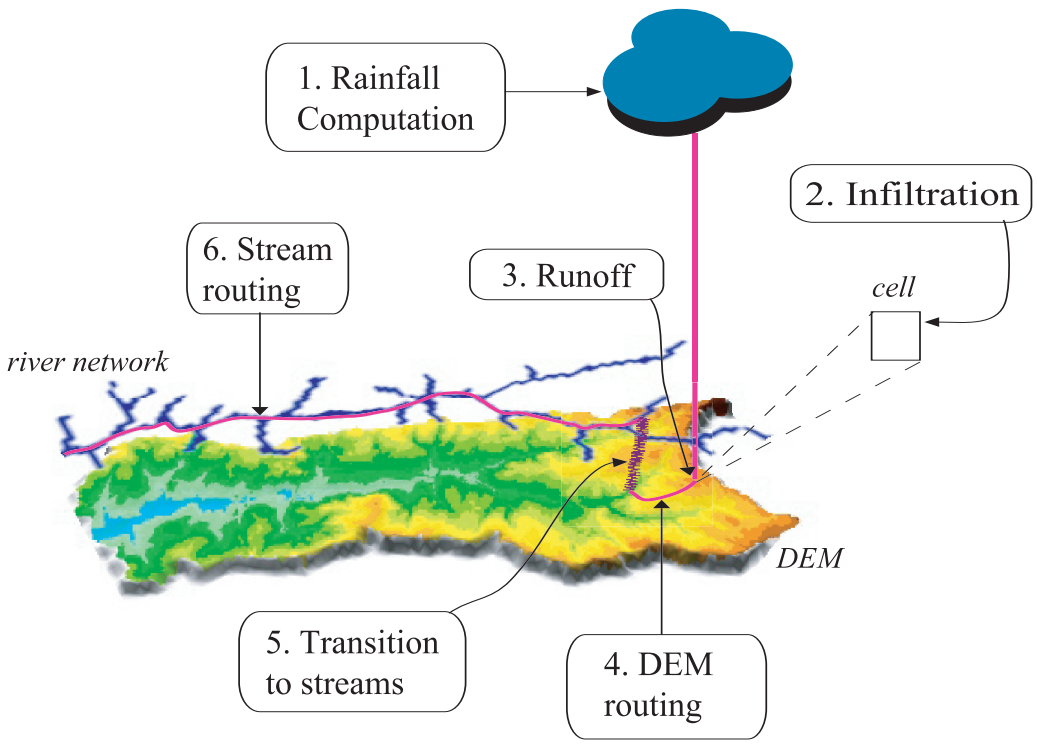

Figure A.2: Procedures performed at each time step of the Network Model.

used to calculate the infiltration $\Delta F(x, t_j)$, and the runoff $u(x, t_j)$ is computed as $u(x, t_j) = r(x, t_j)s - \Delta F(x, t_j)$. The infiltration is computed separately for each cell, so we fix x and write $r(t_j)$ instead of $r(x, t_j)$, $u(t_j)$ instead of $u(x, t_j)$, etc.

Assume the soil has saturated hydraulic conductivity K_s, porosity θ_e, residual water content θ_r, initial soil moisture θ_i, pore distribution index λ, and GA wetting-front capillary pressure parameter H_c. Let $\theta(t)$ be the instantaneous soil moisture, and $Z(t)$ be the depth of profile. $F(t) = Z(t)(\theta(t) - \theta_i)$ is the total amount of water in the profile, and we are going to determine $\Delta F(t_j) = F(t_j + s/2) - F(t_j - s/2)$.

The rules for computing runoff using the GAR model are as follows:

1. If the rainfall rate $r(t)$ is greater than K_s, the original GA equation for a single ponding event is used to determine the potential infiltration $f_p(t)$:

$$f_p(t) = K_s \left(\frac{H_c(\theta_e - \theta_i)}{F(t)} + 1 \right). \tag{A.1}$$

   If $r(t) < f_p(t)$, the infiltration rate is equal to the rainfall rate, and $\Delta F(t) = r(t)s$. If $r(t) \geq f_p(t)$, the infiltration rate is $f_p(t)$ (from equation (A.1)). Equation (A.1) can be solved numerically by noting that $f_p(t) = \frac{d}{dt}F(t)$. In the Network Model, there is a parameter that controls the precision with which equation (A.1) is solved.

2. If the rainfall rate $r(t)$ is less than K_s for the rectangular soil moisture profile approximation, Ogden and Saghafian [1997] provide the following system of

differential equations for $Z(t)$ and $\theta(t)$:

$$(\theta - \theta_i)\frac{dZ}{dt} = \frac{K_s G(\theta_i, \theta)}{Z} + K(\theta),$$

$$\frac{d\theta}{dt} = \frac{1}{Z}\left[r - K(\theta_i) - \left(\frac{K_s G(\theta_i, \theta)}{Z} + K(\theta)\right)\right], \tag{A.2}$$

where $G(\theta_i, \theta)$ is the capillary drive term

$$G(\theta_i, \theta) = H_c \left(\frac{\Theta^{3+1/\lambda} - \Theta_i^{3+1/\lambda}}{1 - \Theta_i^{3+1/\lambda}}\right),$$

$K(\theta)$ is the hydraulic conductivity of soil at the corresponding moisture θ

$$K(\theta) = K_s \Theta^{(2+3\lambda)},$$

and Θ is the relative saturation

$$\Theta = \frac{\theta - \theta_r}{\theta_e - \theta_r}$$

The system of equations (A.2) is solved to update the values of the soil moisture θ and the depth of the profile Z. It is assumed that all the rainfall is infiltrated into the soil, so $\Delta F(t) = r(t)s$.

3. If the rainrate is smaller than K_s for a certain period of time, and the rainfall eventually increases above the K_s level, Ogden and Saghafian [1997], suggest creating a second profile. The procedure provided for this purpose is described below. Note that $\theta^{(1)}$ and $\theta^{(2)}$ denote the water content of the first and second profiles, $Z^{(1)}$ and $Z^{(2)}$ denote the depth of the profiles, and $F^{(1)}$ and $F^{(2)}$ denote the total amount of water in the profiles.

- Initialization (performed at the first time step where the second profile is needed): set both $\theta^{(1)} = \theta$, and $\theta^{(2)} = \theta$. For the initial soil water content, use $\theta_i^{(1)} = \theta_i$ for the first profile and $\theta_i^{(2)} = \theta^{(1)}$ for the second. In this way, the initial soil water content for the second profile varies with time. Set $Z^{(1)} = Z$, $Z^{(2)} = 0$, $F^{(1)} = F$, and $F^{(2)} = 0$.

- Assume that subsequent rainfall fills only the second profile, and use equation (A.1) to compute the infiltration. Use system of equations (A.2) to update the values of $\theta^{(2)}$ and $Z^{(2)}$, concurrently.

- Merge the profiles when $Z^{(1)} = Z^{(2)}$ into a single profile with $\theta = \theta_e$, $Z = Z^{(1)} = Z^{(2)}$, and $F = F^{(1)} + F^{(2)}$.

- If the rainrate drops below K_s before $Z^{(1)} = Z^{(2)}$, merge the profiles anyways. If the second profile is saturated at this time, set the new depth of the profile to

$$Z = \frac{(\theta^{(1)} - \theta_i)Z^{(1)} + (\theta_e - \theta^{(1)})Z^{(2)}}{(\theta_e - \theta_i)}, \tag{A.3}$$

  total amount of water in the new profile to $F = F^{(1)} + F^{(2)}$, and, correspondingly, $\theta = \theta_i + F/Z$.

The GAR model is very sensitive to the time it takes for the soil to become saturated, which is called the ponding time. The ponding time can be determined either by solving system of equations (A.2) until $\theta \geq \theta_e$, or by computing f_p until it is lower than the rainrate. Ogden and Saghafian [1997] point out that the solution

of system of equations (A.2) usually overestimates the ponding time, so the latter approach is preferred. For the purposes of the Network Model, we are not concerned with the value of soil moisture before the first ponding event, since all of the rain is infiltrated. So, by definition, our ponding occurs when $r(t) > f_p(t) \geq K_s$, and the model then starts using equation (A.1) to compute the infiltration. This assures us that we will not solve the system (A.2) before ponding occurs, so the total amount of infiltrated water will be greater than zero by the time the rainrate drops below K_s and we must solve system (A.2).

Small values of Z might cause numerical instabilities in the solution of the system, however. To avoid this problem, a parameter Z_{thresh} is introduced, and system (A.2) is solved only if $Z > Z_{thresh}$. If $Z \leq Z_{thresh}$, the soil moisture is approximated with a formula suggested by Brooks and Corey [1964]:

$$\theta = \theta_i + (\theta_e - \theta_i) \left(\frac{r}{f_p} \right)^{1/(3+2/\lambda)} . \tag{A.4}$$

It is assumed that all water is infiltrated, in this case. Z_{thresh} is a user-specified parameter. Numerical studies with fluctuating rainrates suggest that resultant runoff is not sensitive to this value, as long as it is reasonably low. We suggest values between 5 and 20 mm.

The solution of system (A.2) can also encounter numerical problems at small values of $(\theta - \theta_i)$, resulting in very high values for Z. This problem was addressed by introducing a small threshold ε ($\varepsilon = 10^{-3}$) such that if $\theta - \theta_i < \varepsilon$, we set $\theta = \theta_i$. If two profiles are being updated and $\theta^{(1)} - \theta_i < \varepsilon$ in the first profile, the first profile

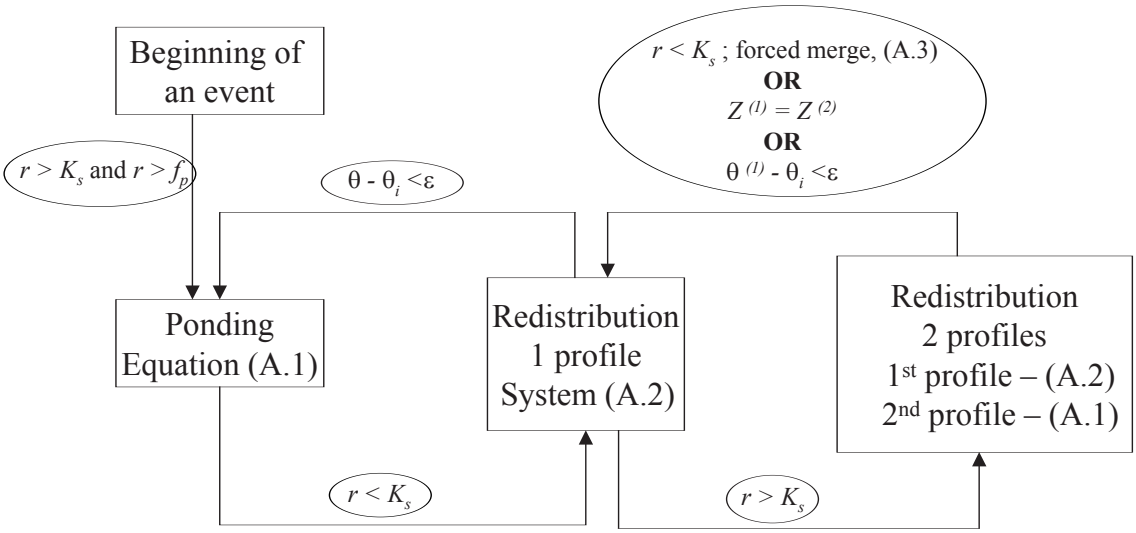

Figure A.3: Schematic of the GAR algorithm used in the Network Model. For an individual cell, a transition between the states of the code along an arrow occurs when the condition on the arrow is satisfied.

Soil texture	Effective porosity θ_e	Residual water cont. θ_r	Pore dist. index λ	K_s, mm h^{-1}	H_c, mm
Sand	0.417	0.02	0.694	235.6	96.2
Loamy sand	0.401	0.035	0.553	59.8	119.6
Sandy loam	0.412	0.041	0.378	21.8	215.3
Loam	0.434	0.027	0.252	13.2	175.0
Silty loam	0.486	0.015	0.234	6.8	320.6
Sandy clay loam	0.330	0.068	0.319	3.0	424.3
Clay loam	0.390	0.075	0.242	2.0	408.9
Silty clay loam	0.432	0.040	0.177	2.0	538.3
Sandy clay	0.321	0.109	0.223	1.2	466.5
Silty clay	0.423	0.056	0.150	1.0	577.7
Clay	0.385	0.090	0.165	0.6	622.5

Table A.1: Typical values of GAR parameter for selected soil textures (from Ogden and Saghafian [1997]).

is eliminated and the second profile is assumed to propagate into soil with initial moisture θ_i.

If the rainfall rate fluctuates around the value of K_s, it is possible that the second profile will not be saturated before the two profiles are forced to merge. We account for this situation in our algorithm; if two profiles are being updated and the rainrate falls below K_s before ponding occurs in the second profile, we estimate the soil moisture in the second profile *during the previous time step* (all variables are taken at time t_{i-1}) using formula (A.4) as

$$\theta^{(2)} = \theta^{(1)} + (\theta_e - \theta^{(1)}) \left(\frac{r}{f_p} \right)^{1/(3+2/\lambda)} .$$

We then merge the profiles and use this estimate of $\theta^{(2)}$ instead of θ_e in equation (A.3) to determine Z.

Figure A.3 summarizes the algorithm described above. GAR computation requires six soil parameters (K_s, θ_e, θ_r, θ_i, λ, and H_c; see Table A.1, Ogden and Saghafian [1997]; Rawls *et al.* [1993] and references therein for typical values of these parameters). Since GAR computation involves numerical solution of systems of differential equations, the user has the ability to control the number of iterations the GAR routine performs each time step in the Network Model (parameter `num_steps_otp`, see Table A.8). It is suggested that this number be increased if the Network Model time step is increased. Numerical studies indicate that the GAR solution is stable if s/`num_steps_otp` is between 30 and 120 seconds.

A.2.2 Flow Directions

The Network Model supports multiple flow paths from a particular grid cell to the outlet. This is achieved by allowing water from a grid cell to flow into any of the eight cells surrounding it (see Figure A.4). What fraction of the water in a grid cell flows into each of its eight neighbors is determined in the "model preparation" stage using the algorithm described below. This algorithm is based on Tarboton's $D\infty$ method for computing flow directions [Tarboton, 1997]. In the Network Model, we added rules for determining the flow fractions when dealing with flat areas and ties. For convenience, we consider a grid pixel O, and label its neighbors as a, b, $\ldots$, h, starting from the pixel directly to the right from O and moving counter-clockwise (Figure A.4). Prior to computing the flow fractions, the DEM is corrected to eliminate depressions using methods suggested by Jenson and Domingue [1988], and it is thusly assured, that each pixel O has at least one neighbor with equal or lower elevation. Only the outlet pixel in the basin, does not have this property, and the runoff from that pixel is directly added to the to the discharge.

Tarboton's $D\infty$ can be briefly described as follows: for pixel O, determine the slope vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_8$ of eight triangular facets formed by O and its eight neighbors (see Figure A.4, facets are numbered starting with the facet formed by pixels O, a, and b, and moving counter-clockwise; the facet's slope is defined as the vector pointing in the direction of the steepest descent of the facet. For facet 1, for example, the absolute value of the slope vector $\mathbf{s}_1$ is $|\mathbf{s}_1| = \sqrt{(e_a - e_O)^2 + (e_b - e_a)^2}/d$, where

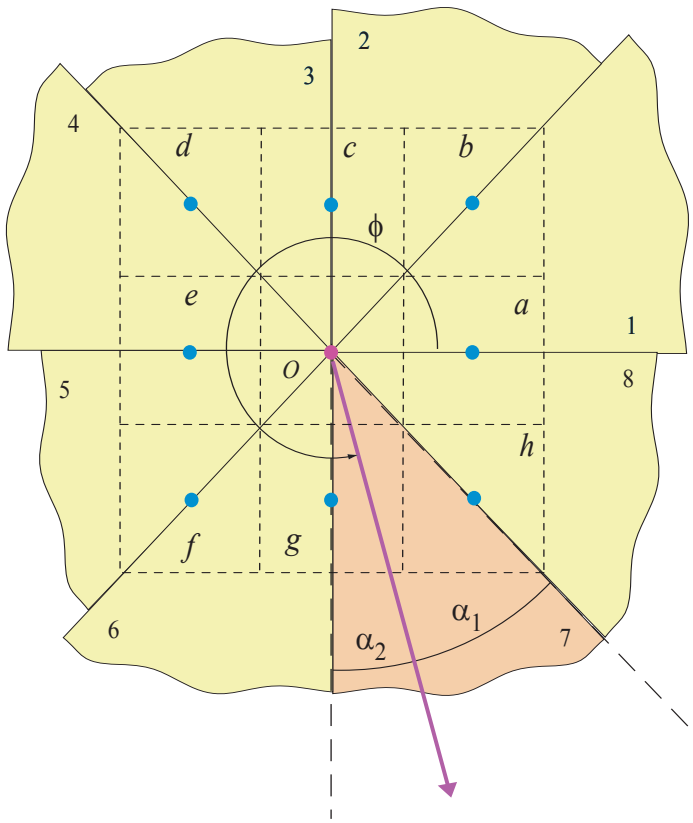

Figure A.4: Illustration for the Tarboton method for specification of flow direction. See text for details.

e_O, e_a, and e_b are the elevations of cells O, a, and b, and d is the DEM cell size. The angle of this slope vector is $\phi_1 = \arctan((e_a - e_O)/(e_b - e_a))$. If ϕ_1 is not in the interval $[0, \pi/4]$, we replace $\mathbf{s}_1$ with the projection of $\mathbf{s}_1$ onto the closest edge of the facet. If this projection is negative, i. e. the vector points away from the facet, we say that $\mathbf{s}_1$ points upward. Otherwise we say that that $\mathbf{s}_1$ points downward. See Tarboton [1997] for details). Of the downward-pointing facet slope vectors, the slope vector with the largest magnitude (thick arrow in Figure A.4) is taken to be the flow direction ϕ of cell O. ϕ is assumed to be a continuous quantity in the range $[0, 2\pi)$. α_1 and α_2 are calculated as differences between ϕ and the directions to the centers of the two neighboring pixels defining the facet (here g and h, respectively). It is assumed, that $\alpha_1/(\alpha_1 + \alpha_2)$ of the water from cell O flows into cell g and $\alpha_2/(\alpha_1 + \alpha_2)$ flows into cell h (these numbers are called *flow fractions*).

Tarboton's flow fractions are used in the Network Model if the flow direction ϕ can be determined. This is possible if at least one of $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_8$ is pointing downward, and the maximum of $|\mathbf{s}_1|, |\mathbf{s}_2|, \ldots, |\mathbf{s}_8|$ is unique (no ties). Here is how flow fractions are computed if ϕ can not be determined:

1. In case of ties: identify all facets with the largest magnitudes of the slope vectors; for each of these facets, determine flow fractions into the pixels defining the facet. Then, for each neighboring cell, we add up the flow fractions going into that cell, and divide it by the number of facets that are tied.

2. In case of flat areas: If none of the $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_8$ are pointing downward, but at least one of the neighboring pixels has the same elevation as O, pixel O and its neighbor are considered to be a part of a flat area. After all pixels $\mathcal{F}$ that are part of the same flat area are identified, the following algorithm is applied to assign flow fractions:

   - Determine the collection of pixels, $\mathcal{G} \subset \mathcal{F}$, within the flat area for which the flow fractions are already assigned.
   - For each pixel in $\mathcal{F}$ but not in $\mathcal{G}$ find neighboring pixels in $\mathcal{G}$, and assign equal flow fractions in the directions of these neighboring pixels. If for a particular pixel no neighboring pixels in $\mathcal{G}$ are found, the flow fractions for this pixel can not be determined at this step.

   This procedure is repeated until flow fraction for all pixel in $\mathcal{F}$ are assigned.

A.2.3 Water Surface Routing and Infiltration

After runoff is generated in a grid cell during a time step, the water is routed on the surface of the DEM until it reaches a grid cell which is defined to be in a stream. Suppose the runoff in a cell A is generated at time t_j, and the overland velocity of the water is constant and equal to v. Figure A.5 shows possible flow paths from a cell A, and the distances $v \cdot s$ that the water will travel on these flow paths during one time step. The water that has not reached streams (in the paths 1 and 2), remains in the corresponding cells (cells B and C) until the next time DEM routing

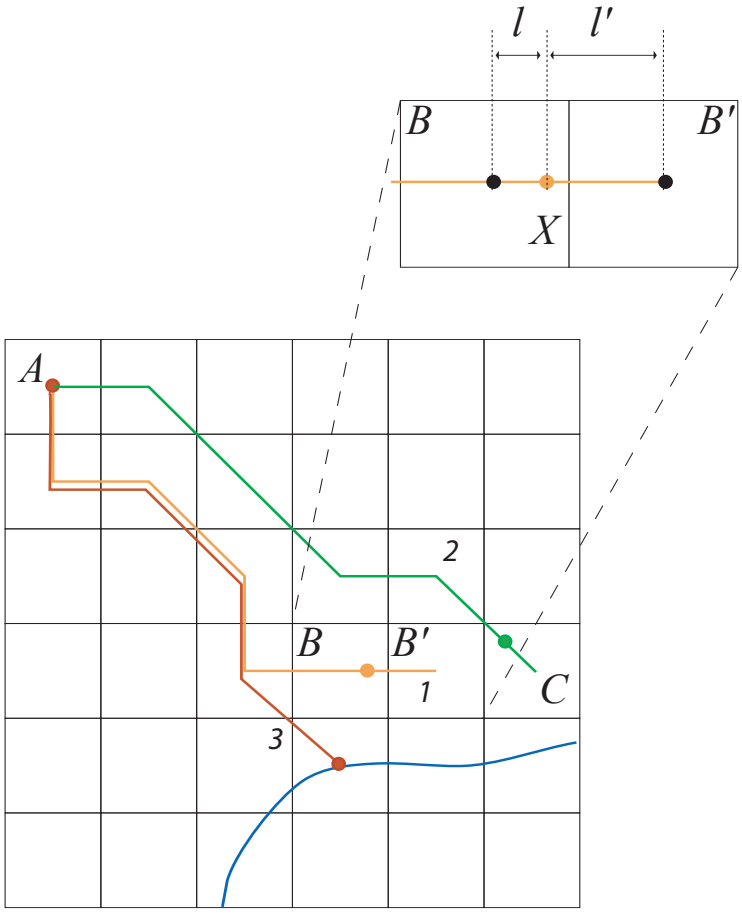

Figure A.5: Possible flow paths from cell A. The points that water reaches during one time step are denoted by circles.

step (this is called *runon*). Since there is an infiltration step between the two DEM routing steps (see schematic A.2), the water can be allowed to infiltrate during the infiltration step in cells B and C or it can be assumed that this water does not contribute to the infiltration in these cells. The latter assumption can be justified by the argument that surface water aggregates into small streams that are narrower than grid cells, and there is no infiltration of water in streams. The Network Model parameter `infiltrate_runon` must be set to `false` to use this option (Table A.8).

A.2.4 Fast Routing Option

In the example using Figure A.5, the water from cell A reaches cells B and C during one time step. In raster-based hydrologic models, we keep track of the water height in only in the centers of grid cells. The water from on the flow path 1 in Figure A.5 is divided between cells B and B', using an inverse distance rule: if l and l' are the distances between the point X (the point reached by the water in one time step) and the centers of cells B and B' respectively, $l'/(l + l')$ of water in the flow path 1 is assumed to stay in cell B and $l/(l + l')$ is assumed to flow into cell B'.

This rule provides good approximation to the surface water flow provided that $vs/d \gg 1$, where d is the DEM cell size. Figure A.6 shows theoretical and modeled contribution to discharge hydrograph at node S of the river network from a constant runoff generated in cell A (assuming $d = 30$ m, and $v = 2.0$ m s^{-1}, and $s = 1$ min). We see that the approximation is as good as we can hope for, given that we model the discharge only at discrete time periods. If we instead assume that $v = 0.2$, we will be

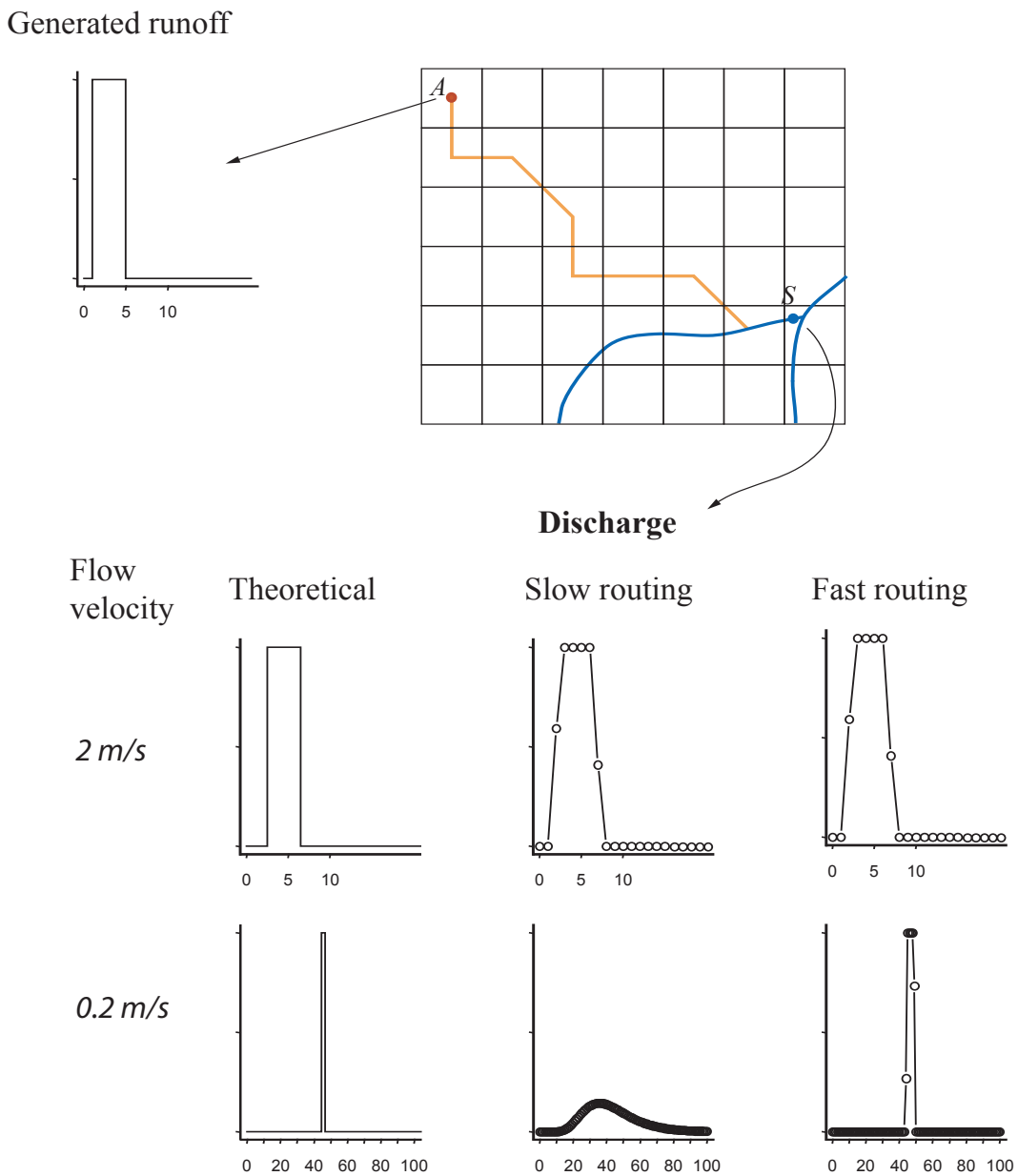

Figure A.6: Difference between fast and slow routing options. The two methods perform the same for large overland flow velocities. Slow routing procedure results is significant attenuation of the flood peak for small overland flow velocities.

distributing water between two neighboring cells at each time step, and Figure A.6 also shows the theoretical and modeled discharges in the stream for this case. We see that the modeled response is attenuated in comparison with the theoretical one, though its center of mass is in the correct place.

In order to correct this error, a user may increase the time step s for a given overland flow velocity v, so that $vs/d \gg 1$ still holds, but this technique may result in a poor representation of rainfall (indeed, if the time step is longer than the period between radar scans, the observed rainfall will be averaged in time rather than interpolated). An alternative way to correct the error is to use the *fast routing* option. With this option, the theoretical water travel time for each flow path from cell A to a stream node S is computed initially, and is continually used to calculate the corresponding discharge at node S without computing the water heights in the DEM cells that the water passes. This option can be used only if the overland and channel flow velocity do not depend on the surface water height or current value of the discharge in the stream (e.g. constant velocity, or a power function of the local slope), and the infiltration of the runon should be turned off (section A.2.3). Since the surface water heights are not computed with this option, display of the analysis using Network Model visualization module is not possible. The output hydrographs for specified arcs, however, may be written into an ASCII file. The `fast_routing` parameter in the main parameter file should be set to `true` if this option is desired (see Table A.8)

The computation time with the fast routing option is about five times less than without it (for an average basin). This is achieved by computing the lengths of all possible flow paths for all DEM cells in the initialization stage which significantly reduces the number of operations during the "running" stage (for a particular DEM cell, flow paths with less than 1% of water flowing on them are disregarded, and total of the flow fractions on the other flow paths is normalized to one). The most time consuming procedure with this option is the infiltration procedure, especially when the redistribution of water in soil moisture profiles for many cells is performed for a long period of time (the rainfall falls below K_s for many cells in the basin). The total amount of computer memory occupied for the fast routing option this larger than for regular operation, since the information about the flow paths has to be kept in memory for faster computations.

A.2.5 Transition to Streams

For transitions of water to streams DEM grid cells that streams pass through are specially marked by the Network Model (as *in-stream cells*), and the point in the stream that is the closest to the center of each in-stream cell is found. All water from the in-stream cells that contributes to streams enters the streams at these points.

Two empirical rules are available to determine how much water from a particular DEM cell enters streams at a particular time step: (1) A fraction f of water from each cell will enter the stream or (2) all water above a height h will enter the stream. If the parameter `partial_discharge` in the main parameter file is set to true, method (1)

is used, and a parameter `fraction_to_stream` should be set to the desired value of f ($0 < f \leq 1$). Otherwise, method (2) is used and the parameter `critical_height` should be set to the desired critical height h in mm (Table A.8).

A.3 User Manual

The primary purpose of the Network Model is to simulate the discharge at catchment's outlet and all nodes of the river network. In order to perform these simulations, the user must prepare the required input files and the main parameter file, and then run the `NetworkModel.RunModel` java class. The Network Model also contains a visualization module, which is briefly described at the end of this section.

A.3.1 Parameter Files

Most of the Network Model's executables require specification of input parameters. All necessary parameters should be stored in a parameter file. The name of the parameter file is passed as the only command line argument to the executable.

Parameter files are regular text files. Parameter values can be specified thusly

```
parameter_name : <parameter_value>
```

Parameters can be either of integer, floating point, date/time, true/false, or string type. Parameters can appear in a file in any order, but each parameter file has to end with `END` at the beginning of a line. All information after the `END` is ignored. `#` is the a commenting character, and will comment out all of the information following it to the end of a line.

Some parameters have default values, and they are specified in appropriate tables below. If the program doesn't find a parameter's name in a file, the default value is used. All parameter names that are not expected by the program are ignored. If the program can't find a required parameter, or interpret a parameter value, or if the parameter file seems to contain mistakes, an error message is printed to the standard output and a `ParameterFileException` is thrown into the main body of the program.

If a parameter specifies a file name, the full path should be included in the name. If a file name specified for the output file is the same as a file already existing in the system, the Network Model will overwrite it without prompting the user. If the Network Model cannot find a particular file, an error message including the name of a `java.io.IOExeption` is printed to the standard output. Users are encouraged to learn more about the meanings of different `java.io.IOExeptions` if they are not familiar with this topic (e. g. see Horstmann and Cornell [2000]).

Sample parameter files are included in the distribution of the Network Model. This user manual will explain the required structure of parameter files in great detail.

In order to run a java class on Unix/Linux, simply type a java command and a class name after it. For example,

```
% java NetworkModel.FilePreparation.RewriteDem rewrite_dem.param
```

runs the `NetworkModel.FilePreparation.RewriteDem` class with a parameter file called `rewrite_dem.param`. If the file `rewrite_dem.param` is in another directory, the

full file name has to be provided. On a Windows system, the codes can be run similarly using the DOS prompt, or using any available java development software. If the name of the parameter file is not provided, the program will try to find a parameter file with a `.param` extension and a name similar to the name of the class (e. g. `RewriteDem` looks for a `rewrite_dem.param` file, `RunModel` looks for a `run_model.param` file, etc.). This option is provided for users of java development products such as Microsoft J++, that make it inconvenient for users to run java classes with arguments.

A.3.2 Preparing Input Files

The Network Model requires inputs to be stored in its own binary format. This format simulates the way java stores its objects in memory, and, as a result, is read by java much faster than ordinary binary or ASCII files. The Network Model has a set of modules that allow users to rewrite ASCII and standard GIS files into the Network Model format. These modules are included in the `NetworkModel.FilePreparation` package and their usage is described in this section. For exact, byte-for-byte, explanation of these formats, one should refer to the source codes.

The standard inputs to the model include the rainfall radar observations, DEM, river network geometry, and soils map. Additionally, a map of impervious areas may be included. All GIS and coordinate data must be in the same projection. The Network Model was tested only for the UTM projection (Universal Transverse Mercator, see Snyder [1987]), but other projections are allowed. It is required, however, that

X coordinates increase in the eastward direction and Y coordinates increase in the northward direction.

DEM and Flow Direction Files

The java class `NetworkModel.FilePreparation.RewriteDem` can be used to transform DEM files to the Network Model binary format. Table A.2 summarizes the parameters required for this class. The elevation of grid cells outside the basin should be set to zero.

The initial DEM can be either in ASCII or Arc-INFO GRID file format. The ASCII format has 4 header lines (any information in the header is ignored by the code) followed by a table of pixel X-coordinate, Y-coordinate, and elevation values. The number of columns and rows in the rectangular DEM grid must be specified in the parameter file. The number of lines in the ASCII file should be equal to (the number of rows in the grid) $\times$ (the number of columns) + 4. If the DEM to be converted is in the Arc-INFO GRID format the dimensions of the GRID file need not be specified in the parameter file. Please note in Table A.2 the different parameter names for the two initial DEM formats. When referencing the location of the outlet in the DEM, please keep in mind that the columns and rows are numbered from zero. It is also important that the name of the output DEM file has a `.bin` extension.

Flow directions for the DEM can be computed using the `NetworkModel.FilePreparation.ComputeFlowDirection` class. The parameters required for this program are summarized in Table A.3. This class computes the flow

Parameter Name	Type	Value/Description	Notes
`dem_from_ascii`	*boolean*	`true`, if input DEM is in ASCII format `false`, otherwise	Default value is `true`
`dem_ascii_file`	*string*	Name of input ASCII file	Required if `dem_from_ascii` is `true`
`dem_GRID_file`	*string*	Name of input GRID file	Required if `dem_from_ascii` is `false`
`number_of_rows`	*integer*	Number of rows in the input DEM grid	Required if `dem_from_ascii` is `true`
`number_of_columns`	*integer*	Number of columns in the input DEM grid	Required if `dem_from_ascii` is `true`
`outlet_row`	*integer*	Outlet's row number	Required; the first row is row number 0
`outlet_column`	*integer*	Outlet's column number	Required; the first column is column number 0
`dem_bin_file`	*string*	Name of output binary file	Required; must end in `.bin`

Table A.2: Parameters for the `NetworkModel.FilePreparation.RewriteDem` class.

fractions, as described in section A.2.2, and stores them in the Network-Model-specific binary format. This program also computes the flow accumulation for each pixel to check the self-consistency of the assigned flow fractions (the flow accumulation for a particular cell is defined as the number of cells from which water flows into this cell). Optionally, the flow accumulation data may be written into an ASCII file that can be imported by ERDAS Imagine.

River (Network) File

The data describing the river network geometry for the Network Model is stored in a binary file with a `.riv` extension. It can be created from an ARC-INFO vector file or two ASCII data files, and the java class `NetworkModel.FilePreparation.RewriteStreamData` maybe used for this purpose (see table A.4). It is important that the input data is indeed a network; i. e. it does not contain any loops, and all arcs are ultimately connected to the outlet. The network is constructed with the outlet node as the starting point, so incorrect specification of the outlet node will misrepresent the network in the Model. For better discharge simulation results, it is recommended that the outlet node lie in the outlet DEM pixel.

The Network Model defines nodes in the network as points where two or more streams intersect to form a new stream, and arcs as links between the nodes. These definitions are different from the definitions of nodes and arcs in ARC-INFO vector files, where a node can connect just two arcs. The Network Model, therefore, examines the input network, and removes superfluous nodes as necessary. All remaining arcs

Parameter Name	Type	Value/Description	Notes
`dem_bin_file`	*string*	Name of the DEM `.bin` file	Required
`flow_direction_file`	*string*	Name of output binary flow direction file	Required
`create_fa_file`	*boolean*	`true`, if the user wishes to output an ASCII flow accumulation file; `false`, otherwise	Default value is `false`
`name_of_ascii_file`	*string*	Name of the output ASCII flow accumulation file	Required if `create_fa_file` is `true`
`name_of_img_file`	*string*	Name of the output ERDAS Imagine `.img` file for use in ERDAS Imagine	Required if `create_fa_file` is `true`. On a Windows platform, use `\\` instead of `\` in the filename

Table A.3: Parameters for the `NetworkModel.FilePreparation.ComputeFlowDirection` class.

Parameter Name	Type	Value/Description	Notes
`network_from_ascii`	*boolean*	`true`, if input stream data is in ASCII format; `false`, otherwise	Default value is `true`
`vector_file_name`	*string*	Name of the ARC-INFO vector coverage	Required if `network_from_ascii` is `false`
`input_arcgen_file`	*string*	Name of input `.arcgen` file	Required if `network_from_ascii` is `true`
`arcgen_format`	*integer*	`2`, if the coordinates in the `.arcgen` file are in decimal format; `1` if they are in scientific format using `D` instead of the usual `e`	Required if `network_from_ascii` is `true`
`input_attributes_file`	*string*	Name of input attributes file	Required if `network_from_ascii` is `true`
`number_of_lines`	*integer*	Number of arcs in the network (lines in the attributes file)	Required if `network_from_ascii` is `true`
`outlet_node_number`	*integer*	The number of the node corresponding to the outlet in the original network	Required
`output_river_file`	*string*	Name of the output river file. Extension `.riv` is required.	Required
`create_arcgen_output`	*boolean*	`true` if the user wishes to export the network for external software	Default value is `false`
`output_arcgen_file`	*string*	Name of the optional output `.arcgen` file.	Required if `create_arcgen_output` is `true`
`dem_cell_size`	*double*	Grid cell size of the DEM	Required

Table A.4: Parameters for the `NetworkModel.FilePreparation.RewriteStreamData` class.

are then assigned labels, so that it is easy for the user to identify them. Figure A.7 illustrates how the network in the input ARC-INFO vector file and the representation of this network in the Network Model may be different. Since the network is closely associated with the DEM, additional breakpoints are added to the arcs to ensure that the distances between adjacent break points are smaller than the DEM grid cell size. It is recommended that the user choose the option to create an output ARC-generate file describing the exact network used in the simulations.

Stream input in ASCII format requires two files: a file describing the geometry of each arc (an `.arcgen` file), and a file describing the connectivity between the arcs (an attribute file). The `.arcgen` file is in standard ARC-generate format: it has entries for each arc, each terminated with `END`, and the file is terminated with another `END`. Each arc entry contains a header line (any information there is ignored), followed by two columns with the X and Y coordinates of the arc breakpoints. The attribute file contains a row for each arc which lists the number of the arc's starting node, the number of the arc's end node, the length of the arc, and two extra numbers that can be set to help identify the arc in the `.arcgen` file. If two arcs are connected, it is required that they have a node number in common. The fourth column must have a unique entry for each individual arc. It is also assumed that the order in which the arcs' information appears in both files is the same. Figure A.8 demonstrates the `.arcgen` and attribute files for a sample network.

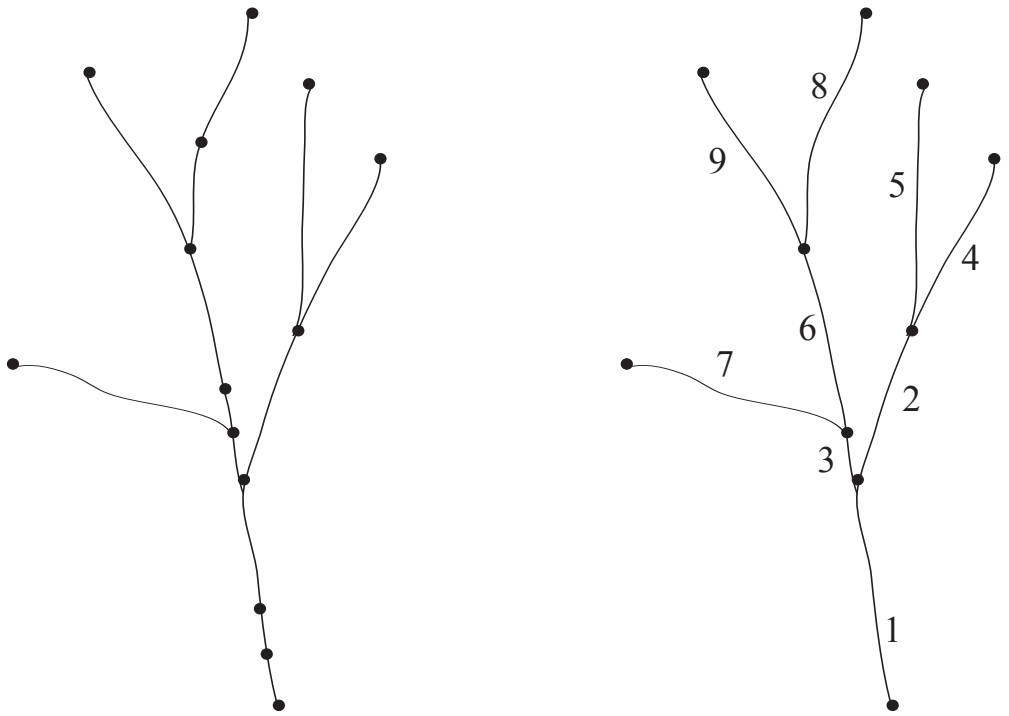

Figure A.7: Potential differences between ARC-INFO vector coverage data (left) and the Network Model representation (right) of the same network. The circles denote nodes of the network. In the Network Model streams are numbered from the outlet, and a typical numbering is shown here.

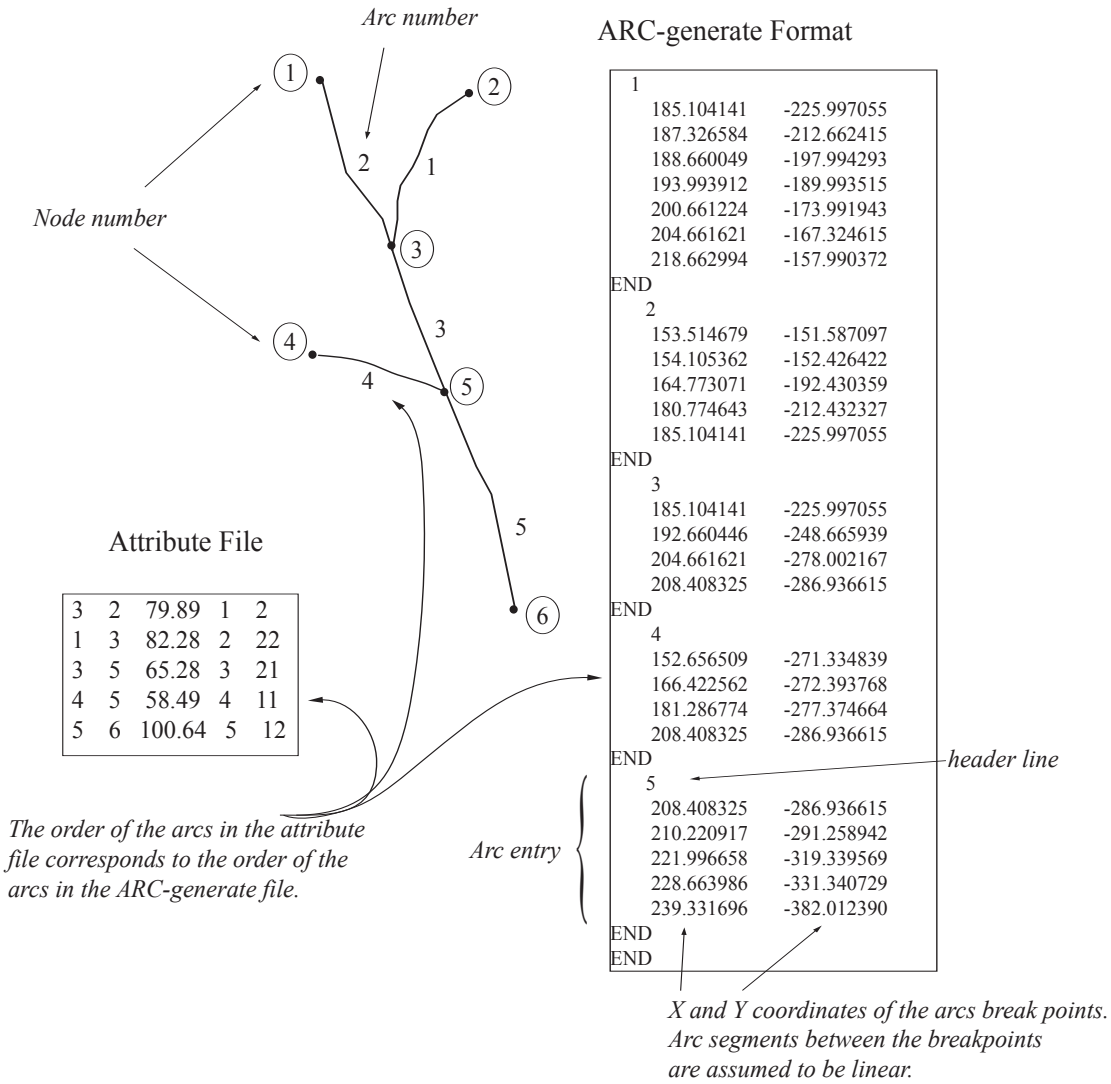

Figure A.8: Example of the ARC-generate and attribute files for a river network.

Rainfall Map

For each event, radar observations of rainfall are stored in special rainfall binary files. There are two types of rainfall binary files used in the Network Model: `.rbin` and `.rcell`. `.rcell` files assume that radar bins form an array of unrestricted size. Therefore, it can cover an area of any shape, for example, a basin. The coordinates of the bins are not included in the file, and must be supplied separately. This format is very convenient for viewing rainfall structures with the `Viewer`. The `.rbin` format assumes that radar bins form a rectangular grid (though not necessarily square). The information from the radar bins is stored in a matrix, and the spatial structure of the data is preserved. `.rbin` files are the ones used in computations, since they allow easy interpolation of the rainfall in space. The coordinates of the bins are included in the `.rbin` file and need not be supplied later.

In order to create `.rcell` or `.rbin` files one needs an $n \times m$ bin ASCII rainfall map. This map is a table where each row contains the observed rainrates during one time period, and each column contains the observations of a particular radar bin. Rainrates are assumed to be in units of mm h^{-1}. An ASCII file with the time periods (the ASCII time file) and a file with the X and Y coordinates of the $n \times m$ radar bins are also needed. It is assumed that the rainrate observed in the bin with coordinates on the ith line in the coordinate file, is provided in the ith column of the rainfall file. n and m may be very large (typical values are ~ 100), and the program will extract the observations for the area covering the basin from the map. It is important that the

coordinates of the bins are in the same projection as the basin DEM (the coordinate calculator in ERDAS Imagine can be used to convert the latitude and longitude of the bins to the appropriate projection).

Time periods in the ASCII time file correspond to the lengths of the radar scans. Each time period should be specified on a separate line, in the format `Day Hour Minute Second`, delimited by spaces. `Day` refers to the day of the month. Month and year of an event are provided separately, among the parameters (see Table A.5). During the program's execution, the "Time of the beginning of the event" is printed to the standard output. This time can be also found in the model's output hydrographs files. In the Network Model, the time is measured in hours from the "beginning of the event".

The java class `NetworkModel.FilePreparation.CreateRectRain` creates both an `.rbin` file for a rectangular area completely covering the basin and an `.rcell` file for just the bins with centers inside the basin boundaries. The parameters for this class are summarized in Table A.5.

Soil and Impervious Cover Maps

The soil map is one of the optional inputs into the Network Model. To use this option, users must create an integer raster map of soil types, with the grid close to a grid of the input DEM. Each soil type should be assigned a unique number between 1 and the total number of soil types (`num_soils`), and the value at each pixel in the soil map assigned to reflect the type of soil there. The soil map can then be converted to the Net-

Parameter Name	Type	Value/Description	Notes
`n`	*integer*	Number of bins in the X direction of the input map	Default value is 100
`m`	*integer*	Number of bins in the Y direction of the input map	Default value is 100
`number_of_time_periods`	*integer*	Number of scans in the input radar observations	Required
`positions_of_bins`	*string*	Name of file with X and Y positions of bins, containing $n \times m$ lines.	Required
`input_time_file`	*string*	Name of ASCII file with radar scans' time periods, containing `number_of_time_periods` lines.	Required
`nXm_rain_map`	*string*	Filename of ASCII table of rainrates observed by radar, containing `number_of_time_periods` rows and $n \times m$ columns	Required
`dem_bin_file`	*string*	Name of the DEM `.bin` file	Required
`output_file`	*string*	A header that is used for output files name	Required
`rainfall_bias`	*double*	A number by which radar observations are multiplied	Default value is 1.0
`date_begin`	*data*	Year, Month and Date of the beginning of the event separated by spaces	Required

Table A.5: Parameters for `NetworkModel.FilePreparation.CreateRectRain` class.

work Model format using the java class `NetworkModel.FilePreparation.RewriteSoils` (see Table A.6 for parameters). The soil map input file can be either in ARC-INFO GRID format or ASCII format, similar to that of the DEM.

Users also must prepare a `soil.table` file. This file should contain at least `num_soils` lines; only the first `num_soils` lines are read by the Model, and users may add comments at the bottom of the file. Each line in the `soil.table` file corresponds to one type of soil in the soil map and contains eight fields in the following order:

1. Soil Number;
2. Textural Classification: a string with no spaces;
3. Porosity, θ_e;
4. Residual water content, θ_r;
5. Pore distribution index, λ;
6. Hydraulic conductivity at saturation, K_s (mm h^{-1});
7. GA wetting-front capillary pressure parameter, H_c (mm);
8. Initial soil moisture, θ_i.

For each soil, the expression $\theta_r < \theta_i < \theta_e$ should hold. Table A.1 provides typical parameter values for different soil textures, as suggested in Ogden and Saghafian

Parameter Name	Type	Value/Description	Notes
`soils_from_ascii`	*boolean*	`true`, if input soil map is in ASCII format `false`, otherwise	Default value is `true`
`soil_bin_file`	*string*	Name of the output soil binary file	Required
`dem_bin_file`	*string*	Name of the DEM `.bin` file	Required
`soil_ascii_file`	*string*	Name of input soil map in ASCII format	Required if `soils_from_ascii` is `true`
`soil_grid_file`	*string*	Name of input soil map in GRID format	Required if `soils_from_ascii` is `false`
`uniform_soil`	*boolean*	If `true` all parameters except `soil_bin_file` are ignored, and the file with only one soil texture is created	Default is `false`

Table A.6: Parameters for the `NetworkModel.FilePreparation.RewriteSoils` java class.

[1997]. It is customary to use the wilting point as a guess for the initial soil moisture if no other information is available.

Instead of using a soil map, one may make an assumption that soils are uniform in the watershed. In this case, parameter `uniform_soil : true` should be added to the parameter file. All other parameters will be ignored by the program, except the `soil_bin_file`. Soil table with only one line corresponding to the basin soil texture should still be created.

In an urban environment, accounting for impervious cover such as roads, buildings, and parking lots is important for accurate simulation of discharges. The Network Model allows users to specify what fraction of each grid cell has impervious cover. This is accomplish with a raster file with the same grid as the DEM and pixel values corresponding to these fractions (all values should be between zero and one). The java class `NetworkModel.FilePreparation.RewriteImperviousAreas` can then be used to transform this data into the Network-Model-specific format (see Table A.7). The impervious cover input files can be either in ARC-INFO GRID or ASCII format, similar to the DEM initial file.

A.3.3 Model Operation

To run the model, one must create a computation parameter file that contains the locations of the input files described above, values of the Model parameters to use in the simulations, as well as specification of the type and destination of the output. Acceptable parameters values in this computation parameter file are summarized in

Parameter Name	Type	Value/Description	Notes
`imperv_area_from_ascii`	*boolean*	`true`, if input impervious cover fraction map is in ASCII format; `false`, otherwise	Default value is `true`
`imperv_area_bin_file`	*string*	Name of the output impervious cover binary file	Required
`dem_bin_file`	*string*	Name of the DEM `.bin` file	Required
`imperv_area_ascii_file`	*string*	Name of input impervious cover in ASCII format	Required if `imperv_area_from_ascii` is `true`
`imperv_area_grid_file`	*string*	Name of input impervious cover in GRID format	Required if `imperv_area_from_ascii` is `false`

Table A.7: Parameters for the `NetworkModel.FilePreparation.RewriteImperviousAreas` java class.

Table A.8. To perform computation, the user must run a `NetworkModel.RunModel` java class with this parameter file as a command-line argument, or save this parameter file as `run_model.param` in the main Network Model directory, and run the `NetworkModel.RunModel` class without any arguments.

Instead of specifying rainfall from radar observations, one can apply a constant rainfall to the basin. This feature is useful for determining the GIUH of a particular basin. To do this, one should use the following three parameters, instead of the `rainfall_file` parameter:

```
rainfall_from_file : false
constant_rainfall : true
rainrate_value : 10
number_time_steps : 5
```

where `rainrate_value` specifies the rainrate of the constant rainfall that should be applied to the basin in units of mm h^{-1} (10 mm h^{-1} in this example), and `number_time_periods` is the number of time steps over which the rainfall should be applied. (For example, if a time step is 180 seconds, and `number_time_periods` is 5, the constant rainfall will be applied for 900 seconds, or 15 minutes.)

The rainfall can be also generated using a stochastic rainfall model. Currently the storm model described in section 6.1 is implemented. We refer users to a sample parameter file included in the Network Model distribution for a particular set of parameters.

Currently, two different options for specifying velocities are implemented in the Network Model: (1) Constant velocity and (2) Power function of local slope, and

Parameter Name	Type	Value/Description	Notes
`dem_bin_file`	*string*	Name of the DEM `.bin` file	Required
`flow_direction_file`	*string*	Name of the binary flow direction file	Required
`streams_available`	*boolean*	`true` if the network data is available. If `false` the DEM routing is performed	Default value is `false`
`network_river_file`	*string*	Name of the network `.riv` file	Required if `streams_available` is `true`
`rainfall_file`	*string*	Name of the binary `.rain` file	Required
`use_Green_Ampt`	*boolean*	`true` if GAR routine should be used; `false` if equation (6.3) should be used	Default value is `true`
`soil_texture_file`	*string*	Name of the binary file with soil texture codes	Required
`num_texture`	*integer*	Number of different soil textures	Required
`soil_table_file`	*string*	Name of the soil table file	Required
`num_steps_otp`	*integer*	Number of iterations to perform in the GAR routine for each step in the main module	Required if `use_Green_Ampt` is `true`
`GAR_theta_threshold`	*float*	The threshold value Z_{tresh} (see page 241)	In mm; default value is 10.0

Table A.8: Parameters for the `NetworkModel.RunModel` class.

Parameter Name	Type	Value/Description	Notes
`impervious_area_available`	*boolean*	`true` if the impervious area data is available for the area	Default value is `false`
`impervious_area_file`	*string*	Name of the binary file with fractions of impervious area	Required if impervious area map is used
`time_step`	*float*	The time step to use during the computations	In seconds; required
`velocity_over_land`	*See text*	Overland flow velocity	Required
`velocity_in_channel`	*See text*	Channel flow velocity	Required
`event_duration`	*float*	Period of time for which to simulate the discharge	In hours; required.
`nearest_neighbor`	*boolean*	Use nearest neighbor rule for rainfall interpolation	Default is `true`
`infiltrate_runon`	*boolean*	`true` if runon from a cell can contribute to the infiltration in other cells (see section A.2.3)	Default is `false`
`fast_routing`	*boolean*	`true` if fast routing option should be used (see section A.2.4)	Default is `false`

Table A.8(continued): Parameters for the `NetworkModel.RunModel` class.

Parameter Name	Type	Value/Description	Notes
`write_hydrographs`	*boolean*	`true` if user wishes to output discharge hydrographs into text files	Default is `true`
`write_only_outlet`	*boolean*	Specifies whether only hydrograph for the outlet arc should be written.	Default is `true`
`arcs_to_write`	*string*	Specifies IDs of the arcs for which hydrographs should be output. The arc number list is delimited by commas	Required only if `write_hydrographs` is `false` and `write_only_outlet` is `false`
`hydrograph_output_file`	*string*	Name of the hydrograph output file	Required if `write_hydrographs` is `true`
`write_results`	*boolean*	`true` if user wishes to output water surface maps for future display in the visualization model.	Default is `false`
`output_directory`	*string*	Name of the directory to write output surface maps	Required if `write_results` is `true`
`output_file_header`	*string*	The beginning of the output file name for the water surface maps.	Required if `write_results` is `true`
`output_frequency`	*integer*	The frequency of output maps	In time steps

Table A.8(continued): Parameters for the `NetworkModel.RunModel` class.

water height (or current discharge). Constant velocity can be specified by typing a keyword `CONSTANT` followed by the value in m s^{-1}. For example

`velocity_in_channel: CONSTANT 1.2` sets constant channel velocity of 1.2 m s^{-1} in all streams, and a similar expression can be used for the overland flow velocity. The power law for the overland velocity assumes that $v = aS^b h^c$, where S is the local slope (in m/m) and h is the current water height (in mm). The power function option for the overland flow velocity is specified using the keyword `POWER_LAW` followed by values of a, b, and c. For example,

`velocity_over_land: POWER_LAW 0.3 2.1 0.1` The formula for the channel velocity is similar: $v = aS^b Q^c$, where Q is the current discharge (in m^3 s^{-1}). The Model has been thoroughly tested only for constant velocities.

A.3.4 Output

The output of the Network Model can be viewed in two different ways: (1) the model can write output hydrographs for a set of specified arcs in the network and/or (2) the model can write a series of binary files that can be viewed using the visualization module.

The output hydrographs are ASCII files. The header of these files contains information on selected input parameters. The body of the file contains $k + 2$ columns, where k is the number of arcs for which the hydrograph should be saved. The first column of the file contains the absolute time, the next contains the time measured

in hours from the beginning of the event, and the following k columns contain the values of modeled discharge for each specified arc.

The visualization module can be started by running the java class `NetworkModel.-Viewer` (no parameter file is required). The user then can use menus and buttons to view the Network Model inputs (rainfall, DEM, soil map, network) and outputs (water surface height, hydrographs for each arc). The user interface is straightforward, and a help menu is available.

A.4 Conclusions

The Network Model provides a set of tools for modeling the discharge at basin outlets due to rainfall. In this Appendix, we have described the basic principles of the Network Model, its options, and how to set up computations using the Network Model.

Bibliography

Adamson, P. T., A. V. Metcalfe, and B. Parmentier, Bivariate extreme value distributions: An application of the Gibbs sampler to the analysis of floods, *Water Resources Research*, **35**(9), 2825–2832, 1999.

Artzner, P., F. Delbaen, J. Eber, and D. Heath, Coherent measures of risk, *Mathematical Finance*, **9**(3), 203–228, 1999.

Bauer, E., Characteristic frequency distributions of remotely sensed in situ and modelled wind speeds, *International Journal of Climatology*, **16**(10), 1087–1102, 1996.

Bazaraa, M. S., H. D. Sherall, and C. M. Shetty, editors, *Nonlinear Programming. Theory and Algorithms*, New York: John Wiley & Sons, Inc., 1993.

Bloschl, G., Scaling issues in snow hydrology, *Hydrological Processes*, **13**, 2149–2175, 1999.

Bloschl, G. and M. Sivapalan, Process controls on regional flood frequency: Coefficient of variation and basin scale, *Water Resources Research*, **33**(12), 2967–2980, 1997.

Borlot, P., S. Coles, and J. Tawn, The multivariate gaussian tail model: an application to oceanographic data, *Applied Statistics*, **49**, 31–49, 2000.

Brooks, R. H. and A. T. Corey, Hydraulic properties of porous media, Hydrology Paper 3, Colorado State University, Fort Collins, CO, 1964.

Bruun, J. T. and J. A. Tawn, Comparison of approaches for estimating the probability of coastal flooding, *Journal of The Royal Statistical Society. Series C – Applied Statistics*, **47**, 405–423, 1998.

Capéraà, P., A.-L. Fourgères, and C. Genest, A nonparametric estimation procedure for bivariate extreme value copulas, *Biometrika*, **84**, 567–577, 1997a.

Capéraà, P., A.-L. Fourgères, and C. Genest, A stochastic ordering based on a decomposition of kendall's tau, in *Distributions with Given Marginals and Moment Problems*, edited by V. Beneš and J. Štěpán, 81–86, Dordrecht: Kluwer Academic, 1997b.

Capéraà, P., A.-L. Fourgères, and C. Genest, Bivariate distributions with given extreme value attractor, *Journal of Multivariate Analysis*, **72**, 30–49, 2000.

Caracena, F. and J. M. Fritch, Focusing mechanisms in the Texas Hill Country flash floods of 1978, *Monthly Weather Review*, **111**, 2319–2332, 1983.

Chowdhury, J. U., J. R. Stedinger, and L.-H. Lu, Goodness-of-fit tests for regional generalized extreme value flood distributions, *Water Resources Research*, **27**(7), 1765–1776, 1991.

Cistinot, A., T. Roncalli, and J. Teïletche, Revisiting the dependence between financial markets with copulas, working paper, 2000.

Clayton, D. G., A model for association in bivariate life tables and its application in epidemiological studies of familial tendendency in chronic disease incidence, *Biometrika*, **65**(1), 141–151, 1978.

Coles, S., J. Currie, and J. Tawn, Dependence measures for the extreme value analyses, working paper, 1999.

Cook, R. D. and M. E. Johnson, A family of distributions for modelling nonelliptically symmetric multivariate data, *Journal of Royal Statistical Society. Series B*, **43**, 210–218, 1981.

Correia, F. N., Multivariate partial duration series in flood risk analysis, in *Hydrologic Frequency Modeling*, edited by V. P. Singh, 541–554, Dordrecht: Reidel, 1987.

Costa, J. E., Hydraulics and basin morphometry of the largest flash floods in the conterminous United States, *Journal of Hydrology*, **93**, 313–338, 1987.

Cox, D. R. and V. S. Isham, Stochastic spatial-temporal models for rain, in *Stochastic Methods in Hydrology*, edited by O. E. Barnoff-Nielsen, V. K. Gupta, V. Perez-Abreu, and E. Waymire, New York: World Scientific, 1998.

D'Agostino, R. B. and M. A. Stephens, editors, *Goodness-of-Fit Techniques*, New York: Marcel Dekker, Inc., 1986.

David, H. A., editor, *Order Statistics*, New York: Willey, 1981.

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

Deheuvels, P., Caractérisation complète des lois extrêmes multivariées et de la convergence des types extêmes, *Publ. Inst. Statist. Univ. Paris*, **23**, 1–36, 1978.

Deheuvels, P., A Kolmogorov-Smirnov type test for independence and multivariate samples, *Rev. Roumaine Meth. Pures Appl.*, **26**, 213–226, 1981a.

Deheuvels, P., A non parametric test independence, *Publ. Inst. Statist. Univ. Paris*, **26**, 29–50, 1981b.

Deheuvels, P., On limiting behavior of the Pickands estimator for bivariate extreme-value distributions, *Statistics & Probability Letters*, **12**, 429–439, 1991.

Dupuis, D. J., Extreme Value Theory based on the r largest annual events: a robust approach, *Journal of Hydrology*, **200**, 295–306, 1997.

Eisenlohr, W. S., Floods of July 12, 1942 in North-central Pennsylvania, *U.S. Geological Survey Water-Supply Paper*, **1134-B**, 1952.

Embrechts, P., C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, New York: Springer-Verlag, 2000a.

Embrechts, P., A. McNeil, and D. Straumann, Correlation and dependence in risk management: Properties and pitfalls, in *Risk management: value at risk and be-*

yond, edited by M. Dempster and H. K. Moffatt, Cambridge University Press, 2000b.

Frank, M. J., On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$, *Aequationes Mathematicae*, **19**, 194–226, 1979.

Fred, D. L., Flow routing, in *Handbook of Hydrology*, edited by D. R. Maidment, chapter 10, New York: McGraw-Hill, 1993.

Galambos, J., Order statistics of samples from multivariate distributions, *Journal of American Statistical Association*, **70**, 674–680, 1975.

Galambos, J., *The Asymptotic Theory of Extreme Order Statistics*, Malabar, FL.: Kreiger Publishing Co., 1987.

Gay, D. M., Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach, *ACM Transactions on Mathematical Software*, **9**, 503–524, 1983.

Genest, C., K. Ghoudi, and L.-P. Rivest, A semiparametric procedure of dependence parameters in multivariate families of distributions, *Biometrika*, **82**(3), 543–552, 1995.

Genest, C. and R. J. Mackay, Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données, *Canadian Journal of Statistics*, **14**, 145–159, 1986a.

Genest, C. and R. J. Mackay, The joy of copulas: Bivariate distributions with uniform marginals, *American Statistician*, **40**, 280–283, 1986b.

Genest, C. and L.-P. Rivest, Statistical inference procedures for bivariate Archimedean copulas, *Journal of American Statistical Association*, **88**(423), 1034–1043, 1993.

Ghoudi, K., A. Khoudraji, and L.-P. Rivest, Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles, *Canadian Journal of Statististics*, **26**, 187–197, 1998.

Goel, N. K. and M. De, Development of unbiased plotting position formula for general extreme value distributions, *Stochastic Hydrology and Hydraulics*, **7**(1), 1993.

Goel, N. K., S. M. Seth, and S. Chandra, Multivariate modeling of flood flows, *ASCE, Journal of Hydraulics Engineering*, **124**(2), 146–155, 1998.

Gumbel, E. J., Distributions des valeurs extrêmes en plusiers dimensions, *Publ. Inst. Statist. Univ. Paris*, **9**, 171–173, 1960.

Gupta, V. K., S. Castro, and T. M. Over, On scaling exponents of spatial peak flows from rainfall and river network geometry, *Journal of Hydrology*, **187**, 81–104, 1996.

Gupta, V. K. and D. R. Dawdy, Physical interpretations of regional variations in the scaling exponents of flood quantiles, *Hydrological Processes*, **9**, 347–361, 1995.

Gupta, V. K., O. J. Messa, and D. R. Dawdy, Multiscaling theory of flood peaks: Regional quantile analysis, *Water Resources Research*, **30**(12), 3405–3421, 1994.

Gupta, V. K. and E. Waymire, Multiscaling properties of spatial rainfall and river flow distributions, *Journal of Geophysical Research*, **95**(D3), 1999–2009, 1990.

Gupta, V. K. and E. Waymire, Spatial variability and scale invariance in hydrologic regionalization, in *Scale Dependence and Scale Invariance in Hydrology*, Cambridge University Press, 1998.

Hack, J. T. and J. C. Goodlett, Geomorphology and forest ecology of a mountain region in the Central Appalachians, *U.S. Geological Survey Professional Paper*, **347**, 1960.

Harris, D., M. Menabde, A. Seed, and G. Austin, Multifractal characterization of rain fields with a strong orographic influence, *Journal of Geophysical Research*, **101**(D21), 26405–26414, 1996.

Horstmann, C. and G. Cornell, *Core Java 2, Volume 1: Fundamentals*, Prentice Hall/Sun Microsystems Press, 2000.

Hosking, J. R. M., Algorithm AS 215. Maximum-Likelihood estimation of the parameters of the Generalized Extreme Value distribution, *Applied Statistics*, **34**, 3001–310, 1985.

Hosking, J. R. M., The theory of probability weighted moments, Research Report RC12210, IBM Research, Yorktown Heights, NY, 1986.

Hosking, J. R. M., L-moments: Analysis and estimation of distributions using linear combinations of order statistics, *Journal of Royal Statistical Society*, **52**(1), 105–124, 1990.

Hosking, J. R. M. and J. R. Wallis, Parameter and quantile estimation for the generalized pareto distribution, *Technometrics*, **29**(3), 339–349, 1987.

Hosking, J. R. M. and J. R. Wallis, A comparison of unbiased and plotting-position estimators of l-moments, *Water Resources Research*, **31**(8), 2019–2025, 1995.

Hosking, J. R. M. and J. R. Wallis, Regional frequency analysis of floods in central Appalachia, Research Report RC20349, IBM Research, Yorktown Heights, NY, 1996.

Hosking, J. R. M. and J. R. Wallis, *Regional Frequency Analysis*, Cambridge: Cambridge University Press, 1997.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood, An appraisal of the regional flood frequency procedure in the UK Flood Studies Report, *Hydrological Sciences Journal*, **30**(1), 85–109, 1985a.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood, Estimation of the generalized extreme value distribution by the Method of Probability-Weighted Moments, *Technometrics*, **27**(3), 251–261, 1985b.

Hüsler, J. and R.-D. Reiss, Maxima of normal random vectors: beween independence and complete dependence, *Statist. Probab. Lett.*, **7**, 283–286, 1989.

Hutchinson, T. P. and C. D. Lai, *Continuous Bivariate Distributions, Emphasizing Applications*, Adelaide: Rumsby Scientic Publishing, 1990.

Jenkinson, A. F., The frequency distribution of the maximum (or minimum) of meteorological elements., *Quarterly Journal of the Royal Meteorological Society*, **81**, 158, 1955.

Jenkinson, A. F., Statistics of extremes, in *Estimation of Maximum Floods, WMO 233, TP 126, Tech. Note 98*, 183–228, Geneva, Switzerland: World Meteorological Office, 1969.

Jenson, S. C. and J. O. Domingue, Extracting topographic structure from digital elevation data for geographic information system analysis, *Photogrammetric Engineering and Remote Sensing*, **54**(11), 1593 – 1600, 1988.

Joe, H., Pametric families of multivariate distributions with given margins, *Journal of Multivariate Analysis*, **46**, 262–282, 1993.

Joe, H., *Multivariate Models and Dependence Concepts*, London: Chapman & Hall, 1997.

Joe, H., R. L. Smith, and I. Weissman, Bivariate threshold methods for extremes, *Journal of Royal Statistical Society. Series B*, 171–183, 1992.

Johnson, M. E., *Multivariate statistical simulation*, New York: Wiley, 1987.

Karr, A. F., Two extreme value processes arising in hydrology, *Journal of Applied Probability*, **13**, 190–194, 1976.

Kelly, K. S. and P. F. Krstanovic, A bivariate meta-Gaussian density for use in hydrology, *Stochastic Hydrology and Hydraulics*, **11**, 17–31, 1997.

Kimeldorf, G. and A. R. Sampson, Uniform representations of bivariate distributions, *Communications in Statistics*, **4**, 617–627, 1975.

Krstanovic, P. F. and V. P. Singh, A multivariate stochastic flood analysis using entropy, in *Hydrologic Frequency Modeling*, edited by V. P. Singh, 515–539, Dordrecht: Reidel, 1987.

Kuchenhoff, H. and M. Thamerus, Extreme value analysis of Munich air pollution data, *Environmental and Ecological Statistics*, **3**(2), 127–141, 1996.

Lang, M., T. B. M. J. Ouarda, and B. Bobee, Towards operational guidelines for over-threshold modeling, *Journal of Hydrology*, **225**, 103–117, 1999.

Leadbetter, M. R., G. Lindren, and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*, New York: Springer-Verlang, 1983.

Lee, A. J., *U-Statistics. Theory and Practice.*, New York: Marcel Dekker, 1990.

Leopold, L. B., Hydrology for urban land planning – a guidebook on the hydrologic effects of urban land use, Geological survey circular 554, U.S. Geological Survey, 1968.

Lettenmaier, D. P., J. R. Wallis, and E. F. Wood, Effect of regional heterogeneity on flood frequency estimation, *Water Resources Research*, **23**(2), 313–323, 1987.

Loveland, T. R. and D. M. Shaw, Multi-resolution land characterization: building collaborative partnerships, in *GAP Analysis: A Landscape Approach to Biodiversity Planning*, edited by Scott, *et al*, 83–90, American Society for Photogrammetry and Remote Sensing, 1996.

Lu, L.-H. and J. R. Stedinger, Variance of two- and three- GEV/PWM quantile estimators: formulae, confidence intervals, and a comparison, *Journal of Hydrology*, **138**, 247–267, 1992.

Lye, J. K., Implications of uncorrelatedness and independence in linear-regression, *Communications In Statistics – Theory And Methods*, **20**(8), 2407–2430, 1991.

MacLeod, A. J., R76-A a remark on algorithm AS 215. Maximum-Likelihood estimation of the parameters of the GEV distribution, *Applied Statistics*, **38**, 198–199, 1989.

Madsen, H., P. F. Rasmussen, and D. Rosbjerg, Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events. 1. At-site modeling, *Water Resources Research*, **33**(4), 746 – 757, 1997a.

Madsen, H., P. F. Rasmussen, and D. Rosbjerg, Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events. 2. Regional modeling, *Water Resources Research*, **33**(4), 759 – 770, 1997b.

Madsen, H. and D. Rosbjerg, The partial duration series method in regional index-flood modeling, *Water Resources Research*, **33**(4), 737–746, 1997.

Marani, A., R. Rigon, and A. Rinaldo, A note on fractal channel structure, *Water Resources Research*, **27**, 3041–3049, 1991.

Martins, E. S. and J. R. Stedinger, Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, *Water Resources Research*, **36**(3), 737–744, 2000.

McNeil, A. J. and T. Saladin, The peaks over thresholds method for estimating high quantiles of loss distributions, in *Proceedings of 28th International ASTIN Colloquium*, 1997.

Morrison, J. E. and J. A. Smith, Scaling properties of flood peaks, to appear in *Extremes*, 2001a.

Morrison, J. E. and J. A. Smith, Stochastic modeling of flood peaks using generalized extreme value distribution, submitted to *Water Resourses Research*, 2001b.

Nelsen, R. B., *An Introduction to Copulas*, New York: Springer-Verlag, 1999.

Northrop, F., A clustered spatial-temporal model of rainfall, *Proceedings of the Royal Society of London. Series A – Mathematical Physical and Engineering Sciences*, **454**, 1875–1888, 1998.

Ogden, F. L. and P. Y. Julien, Runoff sensitivity to temporal and spatial rainfall variability at runoff plane and small basin scales, *Water Resources Research*, **29**(8), 2589–2597, 1993.

Ogden, F. L. and B. Saghafian, Green and Ampt infiltration with redistribution, *Journal of Irrigation and Drainage Engineering – ASCE*, **123**(5), 386–393, 1997.

Onof, C., P. Northrop, H. S. Wheater, and V. Isham, Spatiotemporal storm structure and scaling property analysis for modelling, *Journal of Geophysical Research*, **101**(D21), 26415–26425, 1996.

Over, T. M. and V. K. Gupta, A space-time theory of mesoscale rainfall using random cascades, *Journal of Geophysical Research*, **101**(D21), 26319–26331, 1996.

Parrett, C., Regionalization of precipitation characteristics in Montana using L-moments, *General Design and Roadside Safety Features*, **1647**, 43–52, 1998.

Perica, S. and E. Foufoula-Georgiou, Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling description, *Journal of Geophysical Research*, **101**(D21), 26347–26361, 1996.

Pickands, J., Multivariate extreme value distributions, *Bulletin of the International Statistical Institute*, 229–231, 1981.

Prakasa Rao, B. L. S., *Asymptotic Theory of Statistical Inference*, New York: Wiley, 1987.

Prescott, P. and A. T. Walden, Maximum likelihood estimation of the parameters of the generalized extreme-value distribution, *Biometrika*, **67**(3), 723–724, 1980.

Prescott, P. and A. T. Walden, Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples, *Journal of Statistical Computation and Simulation*, **16**, 241–250, 1983.

Rao, C. R., *Linear Statistical Inference and Its Applications*, New York: Wiley, 1973.

Rawls, W. J., L. R. Ahuja, D. L. Brankensiek, and A. Shirmohammadi, Infiltration and soil water movement, in *Handbook of Hydrology*, edited by D. R. Maidment, chapter 5, New York: McGraw-Hill, 1993.

RiskMetrics, *RiskMetrics Technical Document*, J.P. Morgan, 3rd ed., 1995.

Robinson, J. S. and M. Sivapalan, Catchment scale runoff generation model by aggregation and similarity analysis, in *Scale Issues in Hydrologic Modeling*, edited by J. Kalma and M. Sivapalan, 311–330, New York: Wiley, 1995.

Robinson, J. S. and M. Sivapalan, An investigation into the physical causes of scaling and heterogeneity of regional flood frequency, *Water Resources Research*, **33**(5), 1045–1059, 1997.

Robinson, J. S., M. Sivapalan, and J. Snell, On the relative roles of hillslope processes, channel routing, and network geomorphology in the hydrologic response of natural catchments, *Water Resources Research*, **31**, 3089–3101, 1995.

Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, Some models for rainfall based on stochastic point processes, *Proceedings of the Royal Society of London. Series A – Mathematical Physical and Engineering Sciences*, **410**, 269–288, 1987.

Rodriguez-Iturbe, I. and A. Rinaldo, *Fractal River Basins: Chance and Self-Organization*, Cambridge University Press, 1998.

Rosbjerg, D. and H. Madsen, Uncertainty measures of regional flood frequency estimators, *Journal of Hydrology*, **167**, 209–224, 1995.

Sackl, B. and H. Bergmann, A bivariate flood model and its application, in *Hydrologic Frequency Modeling*, edited by V. P. Singh, 571–582, Dordrecht: Reidel, 1987.

Sales, M., *Index Modelling and Calibration*, Ph.D. thesis, Princeton University, 2001.

Shane, R. M. and W. R. Lynn, Mathematical model for flood risk evaluation, *Journal of Hydrologic Division – ASCE*, **90**(HY6), 1–20, 1964.

Simonovic, S. P., Tools for water management – one view of the future, *Water International*, **25**(1), 76–88, 2000.

Sklar, A., Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231, 1959.

Smith, J. A., Estimating the upper tail of flood frequency distributions, *Water Resources Research*, **23**(8), 1657–1666, 1987.

Smith, J. A., Representation of basin scale in flood peak distributions, *Water Resources Research*, **28**(11), 2993–2999, 1992.

Smith, J. A., M. L. Baeck, J. E. Morrison, and P. Sturdevant-Rees, Catastrophic rainfall and flooding in Texas, *Journal of Hydrometeorology*, **1**, 5–25, 2000.

Smith, J. A., M. L. Baeck, J. E. Morrison, P. Sturdevant-Rees, D. Turner-Gillespie, and P. Bates, The regional hydrology of extreme floods in an urban environment, submitted to *Journal of Hydrometeorology*, 2001.

Smith, J. A., M. L. Baeck, M. Steiner, and A. Miller, Catastrophic rainfall from an upslope thunderstorm in the central Appalachians: The Rapidan storm of June 27, 1995, *Water Resources Research*, **32**(10), 3099–3113, 1996.

Smith, J. A. and A. F. Karr, A statistical model of extreme storm rainfall, *Journal of Geophysical Research*, **95**(D3), 2083–2092, 1990.

Smith, R. L., Threshold methods in statistics, in *Statistical Extremes and Applications*, edited by J. Tiago de Olivera, volume 131 of *NATO ASI Series*, 621–638, Reidel, 1984.

Smith, R. L., Maximum-Likelihood estimation in class of nonregular cases., *Biometrika*, **72**, 67–90, 1985.

Smith, R. L., Extreme Value Theory based on the r largest annual events, *Journal of Hydrology*, **86**, 27–43, 1986.

Smith, R. L., Extreme value theory, in *Handbook of Applicable Mathematics, Supplement*, edited by W. Ledermann, 437–472, Wiley, 1990.

Smith, R. L., Multivariate threshold methods, in *Extreme Value Theory & Applications*, edited by J. Galambos, J. Lechner, and E. Simiu, 225–248, Dordrecht: Kluwer, 1994.

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

Snyder, J. P., *Map Projections – A working manual*, U.S. Geological Survey, Washington, DC, Professional Paper 1453, 1987.

Stedinger, J. R. and L.-H. Lu, Appraisal of regional and index flood quantile estimators, *Stochastic Hydrology and Hydraulics*, **9**(1), 49–75, 1995.

Sturdevant-Rees, P., J. A. Smith, M. L. Baeck, and J. E. Morrison, Tropical storms and the flood hydrology of the central Appalachians, *Water Resources Research*, in press, 2001.

Tarboton, D. G., A new method for the determination of flow directions and upslope areas in grid digital elevation models, *Water Resources Research*, **33**(2), 309–319, 1997.

Tawn, J. A., Bivariate extreme value theory: Models and estimation, *Biometrika*, **75**, 397–415, 1988.

Tawn, J. A., Modelling multivariate extreme value disstribution, *Biometrika*, **77**, 245–253, 1990.

Todorovich, P. and E. Zelenhasic, A stochastic model for flood peaks, *Water Resources Research*, **6**(6), 1641–1648, 1970.

Turner-Gillespie, D., *Urban Drainage System and Extreme Flood Response*, Master's thesis, Princeton University, 2001.

Vanderbei, R. J., LOQO user's manual – Version 3.10, *Optimization Methods and Software*, **11**(2), 485–514, 1999.

Venables, W. N. and B. D. Ripley, *Modern Applied Statistics with S-Plus*, New York: Springer, 1997.

Vodel, R. and J. R. Stedinger, The value of stochastic streamflow models in overyear reservoir design applications, *Water Resources Research*, **24**(9), 1483–1490, 1988.

Waymire, E., V. K. Gupta, and I. Rodriguez-Iturbe, A spectral theory of rainfall intensity at the meso-β scale, *Water Resources Research*, **20**(10), 1453–1465, 1984.

Wood, E. F., M. Sivapalan, K. Beven, and L. Band, Effects of spatial variability and scale with implications to hydrologic modeling, *Journal of Hydrology*, **102**, 29–47, 1988.

Yue, S., The bivariate lognormal distribution to model a multivariate flood episode, *Hydrological Processes*, **14**, 2575–2588, 2000a.

Yue, S., The Gumbel logistic model for representing a multivariate storm event, *Advances in Water Resources*, **24**, 179–185, 2000b.

Yue, S., T. B. M. J. Ouarda, B. Bobee, P. Legendre, and P. Bruneau, The Gumbel mixed model for flood frequency analysis, *Journal of Hydrology*, **226**, 88–100, 1999.

Zhang, Y., J. A. Smith, and M. L. Baeck, The hydrology and hydrometeorology of extreme floods in the Great Plains of eastern Nebraska, *Advances in Water Resources*, in press, 2001.