# Mathematical modelling of rainfall in the Murray–Darling Basin

Roslinazairimah Zakaria

MSc (Statistics)

Thesis submitted for the degree of

**Doctor of Philosophy**

**(Mathematics)**



University of
South Australia

School of Mathematics and Statistics
Division of Information Technology
Engineering and the Environment

**October 2011**

# Disclaimer

I declare that this thesis presents work carried out by myself and does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university. To the best of my knowledge it does not contain any materials previously published or written by another person except where due reference is made in the text and all substantive contributions by others to the work presented, including jointly authored publications, is clearly acknowledged.

---

Roslinazairimah Zakaria

October 2011

# Dedication

*To my beloved parents*

*and*

*to my dear husband and children*

*Zulkefli Yusoff*

*Hidayah, Muhammad Naufal, Abdullah Uzair, Ernie Adilah*

# Acknowledgements

It is a pleasure to thank those who made this thesis possible. Firstly, I owe my deepest gratitude to my supervisors, Professor Emeritus Phil Howlett, Associate Professor Dr John Boland and Dr Julia Piantadosi for their countless support, guidance, constructive ideas and patience. Without them this thesis would not have become a reality. Their assistance has been a great help and enabled me to develop an understanding of the subject and ultimately complete this research.

Before I go further, I would like to highlight some important points from my three supervisors that have a great influence on my research. Phil helps me to sharpen my mathematical skills and knowledge. He has given me the courage to learn and use LaTeX and now I prefer LaTeX more than any other package. John is the person who is responsible for giving me the opportunity to start my PhD. He has the skill of simplifying difficult problems so that I can move forward with confidence and instill 'I can do' motto. Julia is my best friend and she supports me in every aspect of my research and study life. I am greatly indebted to all my supervisors for their dedication and willingness to spend time to do proofreading and giving constructive comments for the betterment of the thesis. "Thank you very much Phil, John and Julia". I consider myself lucky to have a good combination of helpful and supportive supervisors.

The thesis was examined by Professor Panagis Moschopoulos (Department of Math-

ematical Sciences, University of Texas at El Paso, Texas) and Professor Bill Whiten (Julius Kruttschnitt Mineral Research Center, The University of Queensland Experimental Mine, Brisbane). I express my special acknowledgement to the examiners for their invaluable comments to improve the thesis.

Special thanks go to Dr Chris Brien who I regard as my special advisor for R language and was always there when I had problems. Also to Associate Professor Dr Andrew Metcalfe (University of Adelaide), for helping me to complete my understanding in some parts of the research. I am also greatly indebted to my former supervisor during my master's degree at Universiti Kebangsaan Malaysia, Professor Abdul Aziz Jemain for developing my interest in this subject and inspiring me to achieve this far.

I would like also to thank Professor Stanley J. Miklavcic (Head, School of Mathematics and Statistics), Ms Maureen Cotton, Ms Anja Van Vliet, Ms Alison Cameron, Mr Richard Rawinski and Ms Janelle Brown, the University of South Australia (UniSA) personnel for their assistance and co-operation during the duration of my study. Appreciation is also extended to Dr Judy Ford and Ms Monica Behrend from Teaching and Learning unit of UniSA for assistance in English and continuous encouragement in writing. My friends – Rosliza, Nizam, Siti Zaharah Ishak, Noor Ajian Mohd Lair (Dr), Mohammed Awang (Dr), Zarina Abu Hassan, Nahid Banihashemi, Andrew Plumridge, Michael Haythorpe (Dr), Hamideh Anjomshoa (Dr), Saba Majeed, Ali Eshragh (Dr), Sara Browning, Indu Wahdawan, Giang Nguyen (Dr), Thao Tran (Dr) and many more, thank you for your unconditional help and it is really very much appreciated.

I would like to thank the Universiti Malaysia Pahang (UMP) and the Ministry of Higher Education Malaysia (MOHE) for sponsoring my studies. My thanks are also extended to the University of South Australia (UniSA) for sponsoring my atten-

dance at various conferences. By attending conferences I have had the opportunity to extend my networking and build up my confidence and understanding in the subject.

My deep appreciation is dedicated to my parents, Zakaria Ali and Fatimah Jusoh, for their endless love since I was born and also to my mother-in-law, Salbiah Hamid. To my sisters, brother and in-laws,"Thank you so much" for looking after my children.

Last, but not least, I am heartily thankful to my husband, Zulkefli Yusoff, for his unconditional love, understanding and endless support. To my children, Hidayah, Muhammad Naufal, Abdullah Uzair and Ernie Adilah, all of you are the most wonderful and precious assets that I have in this world, and I am very grateful for your patience and understanding while I was away to complete my study. I hope my children can achieve more than myself. Our sacrifices are paid off with the completion of this thesis.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the thesis.

# Publications

This work has resulted in the following publications

Roslinazairimah Zakaria, Andrew Metcalfe, Phil Howlett, John Boland and Julia Piantadosi 'Using the bivariate skew-$t$ copula to model bivariate rainfall distribution', ANZIAM Journal, 51, pp. C231–C246, 2010.

Roslinazairimah Zakaria, John Boland, Julia Piantadosi and Phil Howlett 'Spatial rainfall modelling using the sum of four correlated Gamma variables', in Proceeding ICECS: 2010 The 3rd International Conference on Environmental and Computer Science, Kunming, China, 17–19 October, 2010.

# Summary

The Murray–Darling Basin is an important source of life to Australia as it contributes 40% of the national agriculture production. Recently, the Basin has been degraded environmentally and economically by the drought conditions of nearly a decade. Since this type of situation can occur with finite probability, the management of water within the Basin is critical and needs serious attention.

One of the major difficulties in simulating rainfall is the need to accurately represent rainfall accumulations. An accurate simulation of monthly rainfall should also provide an accurate simulation of yearly rainfall by summing the monthly totals. A major problem in this regard is that rainfall distributions for successive months may not be independent. Thus the rainfall accumulation problem must be represented as the summation of dependent random variables. This means I must understand or at least acknowledge that both the marginal (monthly) distributions and the joint (monthly) distribution are important.

For monthly or yearly rainfall many authors have found that a mixed–gamma distribution that allows a finite probability of a zero total will give a good representation. In this study I am mainly concerned with the representation and summation of monthly rainfall totals. Thus I give particular consideration to the problem of summing gamma random variables.

I consider summation of any number of independent gamma random variables with the same shape but with different scale; summation of any number of independent gamma random variables with both different shape and different scale using the method of Moschopoulos [68]; summation of any number of correlated gamma random variables with a common shape using the extended Kotz and Adams [58] method described by Alouni et al. [3] and the generalized correlative coherence of Getz [38]; and the sum of two correlated gamma random variables with the same shape using the McKay distribution. I also show that a proposed model for the sum of two correlated gamma random variables in which I select the parameters to match the observed cumulants is generally not well posed. The models are tested for both temporal variation (the same location with seasonal variation) and spatial variation (different locations at the same time) with varying success. Finally I consider the use of a bivariate skew-$t$ copula to capture the dependence. The advantage of this method is that it should be relatively easy to extend it to more than two variables.

I extend the formulation of the mixed–gamma variables which was used to model rainfall data including zero totals to model the sum of two, three and $n$ independent gamma variables. I apply the sum of mixed–gamma variables for two and three months (December, January, February) for a candidate location, Pooraka South Australia and found a good fit between observed and simulated data using the Kolmogorov–Smirnov test.

I propose the sum of $n$ correlated gamma variables to model a water catchment or to model a seasonal rainfall scenario in which the correlation issues are addressed. Previously, a similar model has been applied successfully in the telecommunication field. I apply the sum of correlated gamma variables with common shape to form a water catchment representation from four adjacent meteorological stations (Hume, Beechworth, Dookie and Rutherglen). For temporal analysis, I apply this model to the sum of four months during winter and summer seasons for Hume station. In

both cases, the model shows a good fit between the observed and simulated data according to the Kolmogorov–Smirnov and Anderson–Darling tests. The model is tractable and can easily be extended to more than four variables. I derive the PDF of the sum of two correlated gamma variables using the McKay distribution and apply the model to find the sum of rainfall totals from two stations (Hume and Beechworth) and demonstrate a good match for the observed and simulated data based on two goodness of fit tests (Kolmogorov–Smirnov and Anderson–Darling).

One of the major difficulties in rainfall modelling is the need to accurately describe rainfall accumulation on several different time scales. As mentioned earlier it is often possible to obtain a good model for monthly rainfall using a gamma distribution. In order to obtain representative yearly rainfall totals from this model it is necessary to allow correlation between the different marginal monthly distributions. The use of a copula-based model for the joint distribution allows us to model correlation and other joint statistics while preserving the marginal gamma distributions. I have used a bivariate skew-$t$ copula with three undetermined parameters (skewness, correlation, degrees of freedom) to model the sum of two pre-determined marginal gamma distributions. The simulation process is refined by first selecting nominal values of the three parameters and then adjusting these values in order to match the upper and lower tail proportions between the observed and simulated data. The Kolmogorov-Smirnov test shows that the model is good for rainfall totals of two correlated monthly totals. The skew-$t$ distribution can be easily extended to more than two correlated random variables and is likely to have practical use in modelling total catchment rainfalls. I believe this is a topic worthy of further research.

The significant contribution of the study is to provide rainfall modelling for correlated variables which is still lacking in many studies as it becomes complicated theoretically and practically. The model that is suggested is a new application in rainfall modelling. The results from this study can be used as an input to a wa-

ter management model such as rainfall–runoff models, catchment runoff models, reservoir management models, flood prediction models, groundwater recharge models or to determine water policy for the Murray–Darling Basin. The generation of synthetic data can accommodate different scenarios by changing the value of the parameters in the model and will also complement data in the situation wherever data is limited.

# Contents

**Disclaimer** i

**Dedication** ii

**Acknowledgements** iii

**Publications** vi

**Summary** vii

**1 Research background** 1

1.1 Motivation . . . . . 1

1.2 Aims . . . . . 2

1.3 Thesis structure . . . . . 3

1.4 Murray–Darling Basin . . . . . 5

1.4.1 A brief history of the Murray–Darling Basin . . . . . 5

1.4.2 The importance of Murray–Darling Basin . . . . . 8

1.4.3 The geography and climatic conditions of Murray–Darling Basin 8

1.4.4 Hydrology of the Murray–Darling Basin . . . . . 10

1.4.5 The proposed Basin Plan . . . . . 13

1.5 Selected meteorological stations in the Murray–Darling Basin . . . . . 14

1.6 Summary . . . . . 19

**2 Literature review of rainfall modelling** **20**

2.1 Rainfall Models . . . . . 21

2.1.1 Modelling rainfall occurrences . . . . . 21

2.1.2 Modelling rainfall amounts . . . . . 22

2.1.3 Modelling rainfall occurrences and amounts . . . . . 24

2.2 Generation of synthetic rainfall data for rainfall simulation model . . 25

2.3 Other rainfall models . . . . . 27

2.4 Knowledge gap in the rainfall modelling . . . . . 28

2.5 Summary . . . . . 28

**3 Mathematical and Statistical Tools** **30**

3.1 The gamma distribution . . . . . 30

3.1.1 Probability density function . . . . . 31

3.1.2 Estimating parameters for the gamma distribution . . . . . 34

3.2 Generation of synthetic positive rainfall totals . . . . . 38

3.3 Generation of synthetic positive rainfall totals from the probability density function . . . . . 40

3.4 Generation of synthetic non-negative rainfall totals . . . . . 42

3.5 Spearman's rank correlation coefficient . . . . . 43

3.6 Generalised correlative coherence . . . . . 44

3.7 Goodness of fit tests . . . . . 46

3.7.1 P-value . . . . . 46

3.7.2 Chi–Square test . . . . . 47

3.7.3 Kolmogorov–Smirnov test . . . . . 48

3.7.4 Anderson–Darling test . . . . . 50

3.8 Summary . . . . . 52

**4 The sum of independent mixed–gamma variables with the same**

**scale parameter** **53**

4.1 Mixed–gamma model . . . . . 54

4.2 Sum of two independent mixed–gamma variables . . . . . 55

4.3 Sum of three independent mixed–gamma variables . . . . . 58

4.4 General formulae . . . . . 63

4.5 Testing formula of mean and variance for the sum of independent mixed–gamma variables . . . . . 63

4.5.1 Testing formula of mean and variance for the sum of two months 66

4.5.2 Testing formula of mean and variance for the sum of three months . . . . . 68

4.5.3 Goodness of fit test . . . . . 69

4.6 Summary . . . . . 70

**5 Modelling the sum of $n$ independent gamma variables for both different shape and scale parameters** **71**

5.1 Rainfall model for the sum of $n$ independent gamma variables . . . . 72

5.1.1 Probability density function of the sum of $n$ independent gamma variables . . . . . 73

5.2 Representation of the sum of independent gamma random variables . 75

5.2.1 An approximate representation of the sum of independent gamma random variables . . . . . 75

5.2.2 An exact representation of the sum of independent gamma random variables – the Moschopoulos method . . . . . 76

5.3 A case study to compare the approximate and exact representations . 78

5.4 Application of the Moschopoulos numerical algorithm to the previous case study . . . . . 81

5.4.1 Results for the temporal analysis . . . . . 89

5.5 A case study for correlated spatial system . . . . . 95

5.5.1 Results for the spatial analysis . . . . . 98

5.6 Summary . . . . . 99

**6 A model for the sum of correlated gamma random variables with common shape 101**

6.1 Probability density function for the sum of $n$ correlated gamma variables 102

6.2 An approximate representation of the sum of dependent gamma random variables . . . . . 104

6.2.1 A case study of an approximate representation for the sum of dependent gamma random variables . . . . . 105

6.3 An exact representation of the sum of dependent gamma random variables with the same shape – the extended Kotz and Adams method 108

6.4 A counter-example . . . . . 112

6.5 A case study for spatial analysis . . . . . 114

6.5.1 Fitting volume of rainfall totals using gamma distribution . . 117

6.5.2 PDF for the sum of volume of rainfall totals to form a catchment 118

6.5.3 Using the generalised correlative coherence to compare the result for the PDF for the sum of volume of rainfall totals to form a catchment . . . . . 124

6.5.4 Marginals for spatial analysis . . . . . 128

6.5.5 Results for spatial analysis . . . . . 129

6.6 A case study for temporal analysis . . . . . 130

6.6.1 Using the gamma distribution to fit individual monthly rainfall totals according to season . . . . . 131

6.6.2 Modelling the sum of monthly rainfall totals according to season 131

6.6.3 Rainfall model for the sum of positive quadruples of correlated gamma variables . . . . . 134

6.6.4 Marginals of positive quadruples . . . . . 135

6.6.5 Sum of positive quadruples of winter season . . . . . 140

6.6.6 Sum of positive quadruples of summer season . . . . . 141

6.6.7 Results for temporal analysis . . . . . 143

6.7 Summary . . . . . 147

**7 PDF for the sum of correlated gamma random variables using the McKay distribution 148**

7.1 Bivariate McKay distribution . . . . . 148

7.2 Sum of two independent gamma variables . . . . . 150

7.3 The bivariate McKay distribution for correlated gamma random variables . . . . . 152

7.3.1 Justification of the PDF formula . . . . . 153

7.4 A PDF for the sum of two correlated gamma random variables with the same shape . . . . . 155

7.5 The marginal probability density function for a bivariate gamma distribution . . . . . 158

7.6 Moments of the bivariate gamma distribution . . . . . 160

7.7 The correlation for a bivariate gamma distribution . . . . . 167

7.8 Distribution for the weighted sum of two correlated gamma variables 167

7.9 List of formulae used . . . . . 171

7.10 A case study for the sum of two correlated gamma random variables . 172

7.10.1 Results . . . . . 174

7.11 Summary . . . . . 178

**8 Using the method of moments to estimate shape and scale parameters 179**

8.1 Moments, moment generating functions and cumulants . . . . . 180

8.2 Moment generating function for the sum of two correlated gamma variables . . . . . 181

8.3 Parameter estimation using the determinant form of the moment generating function . . . . . 183

8.3.1 Solve for the values of $\alpha, \beta_1, \beta_2, \rho$ . . . . . 186

8.4 Parameter estimation using product form of the moment generating function . . . 191
8.5 Alouini distribution example . . . 193
8.6 Result of the analysis of the sum of correlated gamma variables using Alouini model . . . 199
8.7 Result of the marginal analysis for Hume and Beechworth . . . 204
8.8 Summary . . . 206

**9 Bivariate Skew$-t$ Copula 207**

9.1 Copulas . . . 208
9.2 Using skew-$t$ copula to model rainfall totals . . . 210
9.2.1 Univariate Student's $t$ distribution . . . 210
9.2.2 Univariate skew-$t$ distribution . . . 211
9.2.3 Multivariate skew-$t$ distribution . . . 212
9.3 Simulation experiment . . . 212
9.3.1 Fitting data to gamma marginals . . . 213
9.3.2 Generation of simulated data using skew-$t$ distribution . . . 217
9.3.3 Analysis of the simulation process . . . 218
9.3.4 Fitting the bivariate skew$-t$ copula after refining the simulation process . . . 223
9.4 Results . . . 223
9.5 Summary . . . 229

**10 Comparison of rainfall models 230**

10.1 Rainfall models for independent gamma variables . . . 230
10.2 Rainfall models for correlated gamma variables . . . 231
10.3 Methods for parameters estimation . . . 232
10.4 Summary . . . 233

**11 Conclusions and future research** **234**

11.1 Research contributions . . . . . 234

11.1.1 Rainfall modelling . . . . . 235

11.1.2 Generating synthetic rainfall data . . . . . 238

11.2 Recommendations for future research . . . . . 238

**Appendix A River Murray Management** **240**

**Appendix B Rainfall maps and graphs** **242**

**Appendix C R package** **250**

C.1 Data summary . . . . . 251

C.2 Estimating shape and scale parameters for gamma distribution . . . . 254

C.3 Synthetic data generation using gamma model and mixed–gamma model . . . . . 255

C.4 Newton Method . . . . . 257

C.5 Probability density function of the sum of $n$ correlated gamma variables 260

C.6 Probability density function of the sum of correlated gamma variables of McKay distribution . . . . . 264

C.7 Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests . . . 265

C.8 Plotting the PDF and CDF . . . . . 266

C.9 Bivariate skew-$t$ copula . . . . . 268

**Appendix D MATLAB package** **275**

D.1 MATLAB code for solving $C(X^n) = C_n(r)$ . . . . . 276

D.2 MATLAB code for the Alouini density . . . . . 276

**Bibliography** **282**

# List of Figures


1.1 The Murray–Darling Basin (Source:Murray–Darling Basin Commission 2007) . . . . . 7

1.2 Australian Rainfall Distribution for 2007–2010 (Source: Australian Bureau of Meteorology 2010) . . . . . 10

1.3 Selected meteorological stations in the Murray–Darling Basin and Pooraka meteorological station in South Australia . . . . . 15

1.4 Average of monthly rainfall totals (mm) for Hume, Beechworth, Dookie and Rutherglen (1922–2008) . . . . . 18

3.1 Gamma probability density functions for various parameters $\alpha$ and $\beta$ 34

3.2 Gamma cumulative distribution functions for various parameters $\alpha$ and $\beta$ . . . . . 35

3.3 Synthetic Rainfall totals from cumulative distribution function . . . . 39

3.4 Synthetic Rainfall totals from cumulative distribution function . . . . 50

5.1 Plot of Moschopoulos (circles) and approximate gamma (line), PDFs (above) and CDFs (below) . . . . . 82

5.2 Plot of $\delta_k$ values . . . . . 88

5.3 Observed vs fitted PDFs (above) and CDFs (below) for the sum of four independent gamma variables of winter season, Hume station. . . 92

5.4 Plots of observed vs generated (above) and observed vs generated CDFs (below) for the sum of four independent gamma variables of winter season, Hume station. . . . . 93

5.5 Generated vs fitted PDFs (above) and CDFs (below) for the sum of four independent gamma variables of winter season, Hume station. . . 94

5.6 Observed vs fitted PDFs (above) and CDFs (below) for the sum of two correlated gamma variables, Dookie and Rutherglen stations . . . 100

6.1 PDF (above) and CDF (below) for the sum of two correlated gamma variables from two stations (Dookie and Rutherglen) . . . . . 107

6.2 Approximate area of Dookie catchment . . . . . 116

6.3 Values of $\delta_k$ . . . . . 121

6.4 Observed vs fitted PDFs (above) and CDFs (below) for the sum of four correlated gamma variables from four stations (Hume, Beechworth, Dookie and Rutherglen) . . . . . 123

6.5 Observed vs fitted PDFs (above) and CDFs (below) for the sum of four correlated gamma variables from four stations (Hume, Beechworth, Dookie and Rutherglen) using the same correlation values . . . 127

6.6 Plots of the observed versus fitted gamma sum model for Hume, Beechworth, Dookie and Rutherglen . . . . . 129

6.7 Winter season–observed and fitted . . . . . 138

6.8 Summer season–observed and fitted . . . . . 139

6.9 PDF and CDF for observed and fitted gamma sum for winter season 145

6.10 PDF and CDF for observed and fitted gamma sum for summer season 146

7.1 Observed vs fitted PDFs and CDFs for the sum of two correlated gamma variables using the McKay distribution, Hume and Beechworth 176

7.2 Observed vs fitted PDFs of gamma marginals, Hume and Beechworth stations . . . . . 177

8.1 PDF (above) and CDF (below) of the sum of Alouini model . . . . . 198

8.2 Observed vs fitted PDFs (above) and CDFs (below) for the sum of two correlated gamma variables . . . . . 201

8.3 Observed vs generated PDFs (above) and CDFs (below) for the sum of two correlated gamma variables . . . . . 202

8.4 Generated vs fitted PDFs (above) and CDFs (below) for the sum of two correlated gamma variables . . . . . 203

8.5 Observed and fitted PDFs of gamma marginals . . . . . 205

9.1 Bivariate scatter plot of positive pairs for Hume and Beechworth . . . 213

9.2 Scatter plot of bivariate uniform data of Hume and Beechworth . . . 217

9.3 Comparison of the uniformised observed and generated, Hume (above) and Beechworth (below) stations . . . . . 224

9.4 Comparison of the theoretical and empirical copulas . . . . . 227

9.5 CDF of the theoretical and empirical copulas . . . . . 228

B.1 Australian Rainfall Distribution for 1911–1940 (Source: Australian Bureau of Meteorology 2010) . . . . . 243

B.2 Australian Rainfall Distribution for 1941–1970 (Source: Australian Bureau of Meteorology 2010) . . . . . 243

B.3 Australian Rainfall Distribution for 1971–2000 (Source: Australian Bureau of Meteorology 2010) . . . . . 244

B.4 Australia trend in annual total rainfall 1910–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010) . . . . . 244

B.5 Australia trend in annual total rainfall 1960–2009 (mm/10 yrs (Source: Australian Bureau of Meteorology 2010) . . . . . 245

B.6 NSW trend in annual total rainfall 1910–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010) . . . . . 245

B.7 NSW trend in annual total rainfall 1960–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010) . . . . . 246

B.8 Victoria trend in annual total rainfall 1910–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010) . . . . . 246

B.9 Victoria trend in annual total rainfall 1960–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010) . . . . . 247

B.10 Australia annual rainfall 10 years moving average (Source: Australian Bureau of Meteorology 2010) . . . . . 247

B.11 NSW annual rainfall 10 years moving average (Source: Australian Bureau of Meteorology 2010) . . . . . 248

B.12 Victoria annual rainfall 10 years moving average (Source: Australian Bureau of Meteorology 2010) . . . . . 248

B.13 Sydney median annual rainfall (Source: Australian Bureau of Meteorology 2010) . . . . . 249

B.14 Melbourne median annual rainfall (Source: Australian Bureau of Meteorology 2010) . . . . . 249

# List of Tables


1.1 Geographic coordinates and elevations for selected meteorological stations in the Murray–Darling Basin and South Australia . . . . . 16

1.2 The approximate distances between stations (km) . . . . . 16

1.3 Spearman's rank correlation coefficients between stations . . . . . 16

1.4 Monthly statistics for the four stations . . . . . 17

2.1 Summary of rainfall models . . . . . 28

3.1 Lookup table for synthetic rainfall . . . . . 41

4.1 Estimated parameters, mean (mm) and variance for December, January and February using gamma model . . . . . 65

4.2 Comparison of mean and variance of observed data with gamma and mixed–gamma models . . . . . 66

4.3 Estimated parameters, mean and variance for the sum of two observed data using the gamma model . . . . . 67

4.4 Comparison of mean and variance for the sum of two observed data and the sum of two independent mixed–gamma variables . . . . . 68

4.5 Comparison of mean and variance for the sum of three observed data and the sum of three independent mixed–gamma variables . . . . . 69

4.6 P-values of goodness of fit tests for the sum of two and three independent mixed–gamma variables . . . . . 69

5.1 Spearman's rank correlation coefficients between June–September of Hume . . . . . 79

5.2 Values of $\alpha$, $\beta$ and $p_0$ for different months (July, September, August, June) . . . 79

5.3 Goodness of fit tests for Moschopoulos PDF and approximate gamma PDF . . . 81

5.4 Estimated shape and scale parameters for the winter season . . . 83

5.5 Values of $\gamma$ and $\delta$ . . . 86

5.6 Means and variances for the sum of months in winter season of observed and generated rainfall totals . . . 90

5.7 P-values of goodness of fit tests for the sum of observed and generated months in winter season . . . 91

5.8 Estimated parameters, mean, variance and $p_0$ for Dookie and Rutherglen using the gamma model . . . 96

5.9 P-values of goodness of fit tests for the marginal of gamma variables from Dookie and Rutherglen stations . . . 96

5.10 Estimated parameters, mean and variance for the positive pairs using the gamma model . . . 97

5.11 Comparison of mean and variance for the sum of positive pairs of observed data, the sum of two independent gamma variables and using formula . . . 98

5.12 Goodness of fit tests for the sum of positive pairs from of Dookie and Rutherglen stations . . . 98

6.1 P-values of goodness of fit tests for the sum of dependent gamma variables . . . 106

6.2 Approximate area for catchment coverage . . . 116

6.3 Spearman's rank correlation coefficients between the stations . . . 117

6.4 Estimated parameters, means and variances of the observed and formulae from gamma distribution . . . 118

6.5 Re-estimated values of $\beta$ for $\overline{\alpha} = 1.65$ . . . 119

6.6 Statistics for the sum of observed and generated . . . 122

6.7 P-values of goodness of fit tests for the sum of four correlated gamma variables from four stations . . . 122

6.8 Statistics for the sum of observed and generated . . . . . 125
6.9 P-values of goodness of fit tests for the sum of four correlated gamma variables from four stations . . . . . 126
6.10 Means and variance of marginals in spatial analysis . . . . . 128
6.11 Goodness of fit tests for marginals in spatial analysis . . . . . 128
6.12 Descriptive statistics of winter and summer seasons . . . . . 132
6.13 Estimated parameters, observed and estimated means and variances of monthly rainfall . . . . . 133
6.14 Monthly correlations of winter season . . . . . 133
6.15 Monthly correlations of summer season . . . . . 134
6.16 Average $\alpha$ ($\overline{\alpha}$) and re-estimate $\beta$s . . . . . 135
6.17 Winter months–Mean and variance of the observed and generated . . 136
6.18 Summer months–Mean and variance of the observed and generated . 136
6.19 Kolmogorov–Smirnov test for marginal distributions . . . . . 137
6.20 Mean and variance of the observed and generated rainfall totals for winter and summer seasons . . . . . 143
6.21 Kolmogorov–Smirnov test for sum of correlated gamma variables . . . 144

7.1 Shape and scale parameters of monthly rainfall totals in summer months 173
7.2 Re-estimated values of $\beta$ for $\overline{\alpha} = 1.1531$ . . . . . 173
7.3 Statistics of the observed and generated data for monthly rainfall totals 174
7.4 Goodness of fit test of marginals and sum of two correlated gamma variables of McKay distribution . . . . . 175

8.1 Descriptive statistics for positive pairs of Hume and Beechworth (1922–2009) . . . . . 185
8.2 Estimated parameters for positive pairs of Hume and Beechworth . . 185
8.3 Descriptive statistics of the sum of observed, fitted and the Alouini model . . . . . 200
8.4 Goodness of fit tests for the sum of correlated gamma variables . . . 200

8.5 Comparison of mean and variance of the observed and fitted data for nonzero rainfall totals . . . . . 204

8.6 Goodness of fit tests for the marginal analysis of the nonzero rainfall totals . . . . . 204

9.1 The monthly values of shape parameter $\alpha$ and scale parameter $\beta$ . . . 214

9.2 Some values of observed data transformed to uniform data using the CDF of gamma. The Hume data is transformed using $\alpha = 1.5382$ and $\beta = 38.2018$ and for Beechworth the values are $\alpha = 1.7719$ and $\beta = 47.0027$. . . . . 215

9.3 Upper and lower tail proportions and correlation coefficient to measure the degrees of dependence between bivariate data of Hume and Beechworth . . . . . 216

9.4 Variation of skewness ($\gamma$) . . . . . 220

9.5 Variation of degrees of freedom ($\nu$) . . . . . 220

9.6 Variation of model correlation ($\rho$) . . . . . 221

9.7 Refining simulation process . . . . . 222

9.8 Goodness of fit tests of theoretical and empirical copulas for the uniformised observed and generated data . . . . . 225

9.9 Goodness of fit tests of theoretical and empirical copulas . . . . . 228

10.1 Rainfall model comparison . . . . . 232

# Chapter 1

# Research background

## 1.1 Motivation

The need to improve water management is not limited to places where water is scarce. It is a global need which requires policy makers, authorities and governments of countries to work hand in hand to find feasible solutions for solving problems associated with water. Too much or too little water will cause significant problems. The repercussions of mismanagement of water will be felt in agricultural activities, food production, environmental degradation and other related and unrelated problems. Rainfall is a key element of water management in many parts of the world.

Extensive studies of many aspects of water related issues have been conducted throughout the world. Hydrology and climatology are particular areas of study that use rainfall as their input analysis. Hydrology concerns modelling water catchment, rainfall-runoff and storm water management. Climatology focuses on modelling for climate change and weather forecasting. This project is motivated by the need to develop mathematical models related to rainfall modelling that are able to closely

represent or mimic the real situation.

## 1.2 Aims

The focus of this research is to develop mathematical models that can be used for rainfall modelling. I will show how the sum of correlated gamma variables can be used in the modelling of rainfall. I will target modelling of monthly rainfall totals in terms of spatial and temporal analyses. If the amount of rainfall for each station in a catchment is known, then a weighted sum from all the relevant stations could be used to calculate the catchment rainfall and then, in turn, the total storm-water runoff. For temporal analysis, we will be able to generate synthetic seasons and even years of rainfall data for a particular station which give the impact in a localised area.

In this research the interest is in modelling the correlated random variables to form a sum. Whenever the correlated random variables are used, extra parameters will be introduced and the model will become more complex. However, I also study models for the independent random variables whenever the correlation between the variables is weak or insignificant.

I study two types of models which can be used for independent variables and correlated variables, respectively. Firstly, I use the Spearman's rank correlation to calculate the correlation between the variables. The Spearman's rank correlation is used since the rainfall data is usually not normally distributed, refer to [81] and [29]. For the independent random variables, the sum of the independent gamma model and sum of the mixed–gamma model are fitted. The sum of independent gamma model will only fit for strictly positive random variables but the sum of the mixed–gamma model includes zero rainfall totals. The sum of the correlated gamma

model which is used for correlated random variables has two forms. The first form can be extended to more than four correlated random variables easily. However, the form that involves the McKay Bessel function is fitted only to two correlated random variables since it requires extensive algebraic manipulation.

The parameters for all the models used are estimated using the maximum likelihood estimation method. The models developed are applied to generate synthetic rainfall totals. Generating the synthetic rainfall data will fill in the missing data for the nearby station which possesses the same rainfall characteristics since the correlation is strong ($\rho > 0.7$) and also significant (P-value $< 0.05$). We may also be able to simulate the effects of climate change for each of these situations by suitably altering parameters that reflect changes in mean or variance.

## 1.3 Thesis structure

The thesis is presented with chapters in the following sequence:
**Chapter 2** reviews the rainfall modelling available. The rainfall modelling is categorised into modelling rainfall totals, rainfall occurrences and combination of both. We also include the literature for generating synthetic rainfall data.

**Chapter 3** presents the mathematical and statistical tools that are essential in this study such as gamma distribution, formula of Spearman rank correlation, generalised correlative coherence and the goodness of fit tests.

**Chapter 4** discusses the theory of the sum of $n$ independent mixed–gamma variables for the same shape and distinct scale parameters. The derivation of mean and variance formulae for the sum of two and three independent mixed–gamma variables is presented. Consequently, the general formula for the sum of $n$ independent gamma variables is derived. The applications of the model are shown for the sum of two

and three independent mixed–gamma variables.

**Chapter 5** describes the theory of the sum of $n$ independent gamma variables for both different shape and scale parameters. An illustration of the model is shown using the monthly rainfall totals of the winter months from Hume meteorological station.

**Chapter 6** describes the theory of correlated gamma variables with derivation and proof for some of the formulae related to the distribution. It includes application of the model in terms of spatial and temporal analyses. In the spatial it is between stations and in the temporal, between months.

**Chapter 7** studies the theory of the McKay distribution associated with the sum of two independent and two correlated gamma variables. The derivation of the probability density function of two correlated gamma variables is presented. The derivation of moments for bivariate gamma and the distribution of the weighted sum of two correlated gamma variables of McKay function is shown with an application.

**Chapter 8** describes the use of moment generating function method to estimate parameters for the correlated gamma variables. The application in modelling the sum of two correlated stations is presented.

**Chapter 9** describes the use of copula with gamma distributions, the theory of skew-$t$ copula and analysis of tail proportions as an extension of correlation using bivariate skew-$t$ distribution. The simulation technique is used to illustrate the model.

**Chapter 10** compares all the rainfall models that are presented in the study.

**Chapter 11** summarises and concludes the study. The future research topics are proposed.

**Appendices** The appendices include the following:

Appendix A. Part of the speech given by the ex-president of the River Murray Commission regarding the River Murray management.

Appendix B. The rainfall maps and graphs of the Murray–Darling Basin.

Appendix C. The $R$ package and some of the codes are

- data summary
- estimating the shape and scale parameters of the gamma distribution
- solving the Newton method
- generating the synthetic data using the sum of $n$ correlated gamma variables and the statistical analysis
- evaluating the probability density function of the sum of correlated gamma variables of the McKay distribution
- constructing the bivariate skew-$t$ copula.

Appendix D.1. The MATLAB package and code for solving the generalised correlation.

In the next section we present the overview of the Murray–Darling Basin.

## 1.4 Murray–Darling Basin

### 1.4.1 A brief history of the Murray–Darling Basin

Development of the Murray–Darling Basin (MDB) started as early as during the pre-federation era (early 1800s). The rivers were first used to transport people and goods using steam boats, Wigley [102]. Between 1829 and 1830, Captain Charles Sturt and Major Thomas Mitchell discovered and explored the MDB which was

then transformed by European settlement and an extensive irrigation system was constructed, Sinclair [91]. The main issue associated with the MDB management since then is due to its length which covers many states, therefore it needs good cooperation. In 1863, the first meeting between New South Wales, Victoria and South Australia was held in Melbourne to improve water navigability. However, only after forty years of negotiation, the River Murray Commission Agreement was finally signed in 1915 to establish the water sharing policies and improvement of its operation between the states [67].

In 1985, the MDB is officially managed by the Murray–Darling Basin Commission (MDBC) to extend the function of addressing other issues related to land use and environment [67]. The Murray–Darling Basin Commission plays a vital role in co-ordinating the intergovernmental cooperation to develop programs and policies, to handle problems associated with dams operation and to promote mechanisms for natural resource management in the Basin [67]. Recently, the effort to save the MDB extends to non-governmental agencies and other concerned groups such as Venture Innovation Vision Action South Australia (VIVASA) and Global Energy and Water Cycle Experiment (GEWEX). In December 2008, the Murray–Darling Basin Authority (MDBA) is established to replace the Murray–Darling Basin Commission to further improve the management of the Basin especially to tackle issues related to water management within the Basin.

There are many issues associated with the MDB in terms of water management, environment, social and economics. Many research reports are published to address the issues such as in water management ([67], [22], [54] and [66]) and salinity problem [67]. A comprehensive report on the MDB is written by Crabb [27]. Connell [22] explores the political issues surrounding the Murray–Darling Basin. He highlights that although various agreements and regulations have been established, the condition of the Murray–Darling Basin is degrading due to lacking of committed

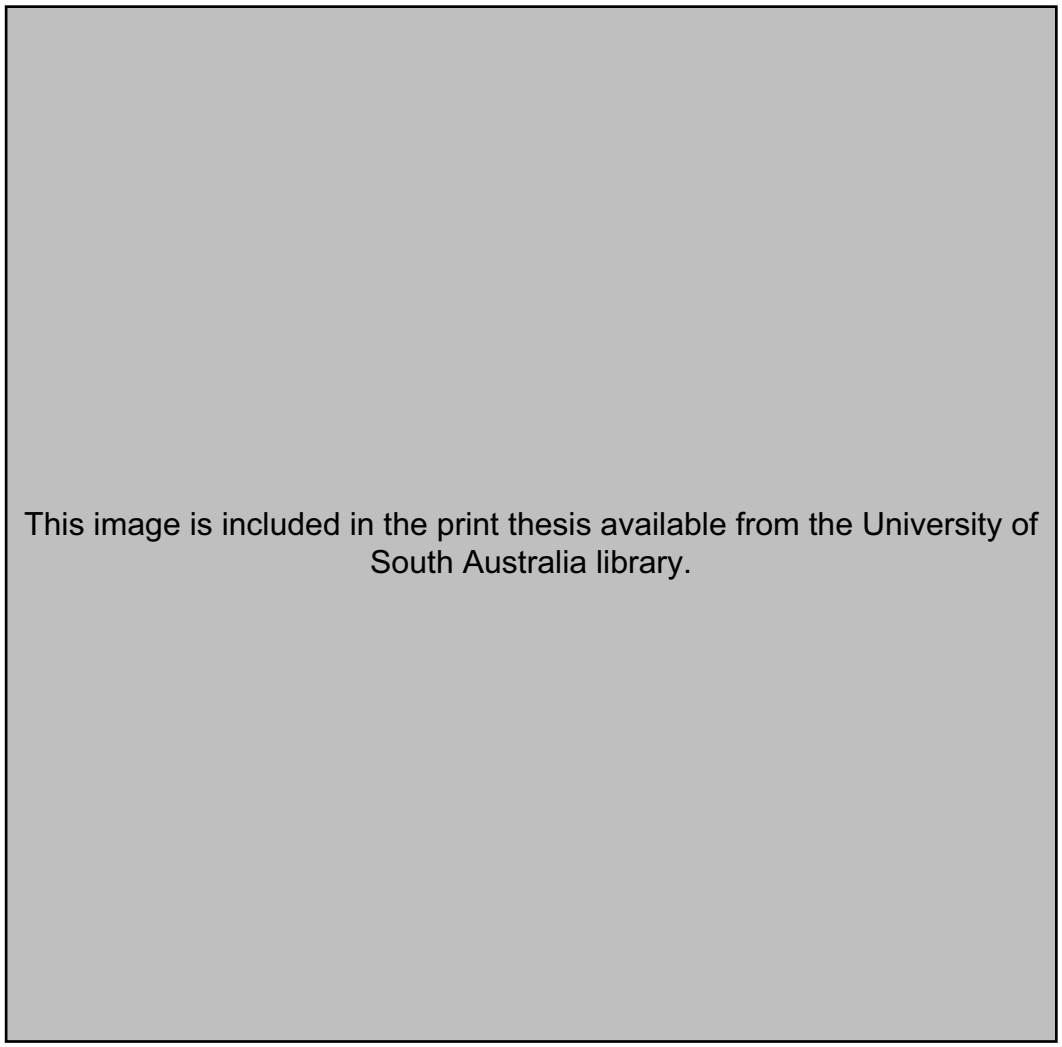



**Figure 1.1:** The Murray–Darling Basin (Source:Murray–Darling Basin Commission 2007)

cooperation between the governments or authorities involved. McMichael [64] also addressed his concern regarding the importance of interdisciplinary cooperation in solving problems within the MDB. An extract of his text is given in Appendix A.

### 1.4.2 The importance of Murray–Darling Basin

The MDB is considered as the 'food bowl' of Australia which represents 40% of Australia agricultural production and provides 3.5% (290,000) employments for Australia, [67, 7]. The agricultural activities (crops and pastures) consume about 70% of water from the Basin. The population across the Basin is about 2 million which is 10% of Australia population and another one million live outside the Basin also depend on the water from the Basin. The natural resources-based industries in the Basin contribute $10 billion per year which is more than 30% of Australia's total output, [101, 79, 39]. For example, in 2001, almost Australia's entire rice crop is grown in the Basin across 176,000 hectares which produce 1.6 million tonnes worth at $348.9 million. Furthermore, the MDB also offers many interesting and spectacular places such as national Parks, State Forests, rivers and wetlands that attract 15 million tourists yearly, [67].

### 1.4.3 The geography and climatic conditions of Murray–Darling Basin

Australia is the driest inhabited continent and the MDB is the most important river system that gives life to most part of Australia. The Murray–Darling Basin is situated at the south-east of Australia and covers an expanse of over 1.1 million square kilometres (one-seventh or 14 % of Australia), [67]. It covers four states (New South Wales, Queensland, Victoria and South Australia) and the Australian Capital Territory (ACT). The percentages of area of the Basin by states are 57% of New South Wales, 25% of Queensland, 12% of Victoria, 6% of South Australia and less than 1% of ACT, [67].

The Murray–Darling Basin is a catchment area of the Murray and Darling rivers

and their other tributaries. The major rivers that contribute to the Basin are the Murray River (2530 km), the Darling River (2740 km) and the Murrumbidgee (1690 km). See Figure 1.1. The other important tributaries are the Goulburn and the Lachlan. The Darling River and its tributaries drain the Darling plain in the north and the Murray and Murrumbidgee and their tributaries drain the Riverine Plain in the south [79, 67, 6].

The MDB is among the largest catchments in the world. However, it is the driest when compared to the Amazon catchment which has 75 times more water flow for every square kilometer than the MDB. The annual flow of the MDB is less than the daily flow for Amazon, Young [109]. Since the Basin covers a vast area, it has a diverse range of climatic scenario and natural environments. Climate in the eastern part is cool and humid with rainforest and sub-tropical in the northeast. The southeast part has a temperate climate while the western plains are dry semi-arid and arid lands. Due to the climate variability, the rainfall patterns vary across the Basin where more than 75% of the Basin area receives less than 600 mm of rainfall annually and the potential evaporation from the Basin is nearly four times the rainfall amount [79, 67]. The rainfall distribution for Australia for the years of 2007–2010 is shown in Figure 1.2. We observe that most areas within the Murray–Darling Basin receive less than 1000 mm rainfall totals over the given years. In brief, these extreme scenarios caused a serious repercussion in the water resource management.

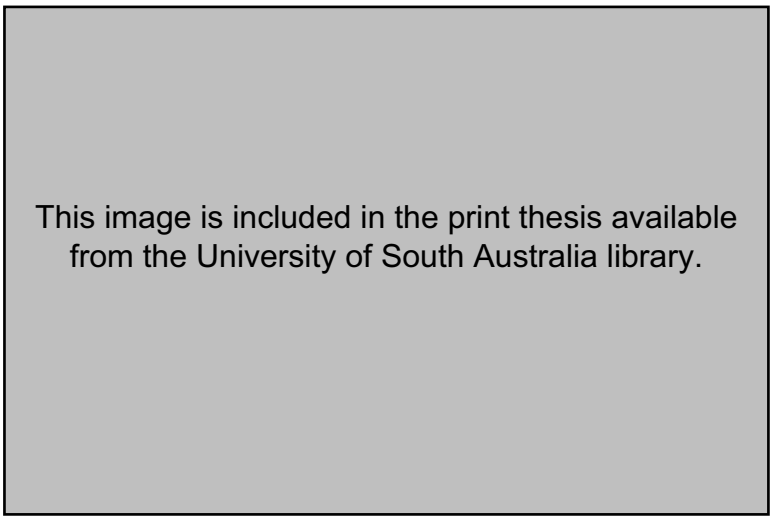



**Figure 1.2:** Australian Rainfall Distribution for 2007–2010 (Source: Australian Bureau of Meteorology 2010)

### 1.4.4 Hydrology of the Murray–Darling Basin

The water storage capacity in the MDB is about 35,000 million litres which is managed by two bodies namely the Murray–Darling Basin Commission (MDBC) and the respective States. The major storages managed by the MDBC are the Dartmouth, Hume, Lake Victoria and Menindee Lakes. These water storages are to be used to store and supply water during wet and dry seasons respectively for domestic, industrial and agricultural activities.

Currently, the volume of water held in many storages within the MDB is insufficient or below the average level. This condition is caused by a severe drought which occurred between 1997 and 2006. The drought is due to low rainfall totals over a long period of time. Potter and Chiew [77] identified three significant drought events throughout the history. The "Federation drought" occurred in early twentieth century (1901-1902), and there was a drought in 1936–1945 and recently the drought in

1997–2006. The current drought event is considered worse than the other drought periods [77, 97, 78]. Potter and Chiew [77] and Potter et al. [78] also reported that the annual mean rainfall between 1997 and 2006 is 16% less than the average over 1895 to 2006 period. The runoff has fallen 39% within the same duration. Refer to [77, 78] and [97] for further details.

The hydrology of the Murray-Darling Basin is driven by the rainfall in South East Australia. It is appropriate to present an objective review of rainfall and rainfall trends in Australia. Our review is based on data that is freely available on the Australian Bureau of Meteorology website. The relevant maps and graphs are shown in Appendix B.

The maps in Figures B.1, B.2 and B.3 show that in broad terms the average annual rainfall has remained much the same. Nevertheless a close inspection reveals a discernible increase in the average over the wet tropical areas of Arnhem Land in the Northern Territory and Cape York Peninsula in Queensland for the period 1971 to 2000. In the same period rainfall has also increased over the dry inland areas of Western Australia.

Figures B.4 and B.5 show the trends in the overall Australian average rainfall for the period 1910-2009 and for the period 1960-2009. The trend for the 100 year period 1910-2009 shows a broadly based increase over all areas except far southwest Western Australia and eastern central Queensland where there has been a significant decrease and parts of southern coastal Australia where there has been no significant change. The 50 year trend for the period 1960-2009 does, however, show a marked decrease in east coastal Queensland and New South Wales, in south east Australia generally and also in west coastal Western Australia. Central and Northern Australia away from the west and east coast show significant increases. For south east Australia the changes are shown in Figures B.6, B.7, B.8 and B.9. One can deduce from these

maps that rainfall increased over the Murray-Darling Basin between 1910 and 1960 especially in the northern regions and decreased by a similar amount between 1960 and 2009.

In general terms the graph of observed yearly rainfall for Australia in Figure B.10 shows an overall increase. The ten-year moving average was approximately 450 mm between 1900 to 1920, dropped to around 400 mm in 1935 and then increased to about 470 mm in 1950. Since then there have been low points of 400 mm in 1961 and 440 mm in 1990 with peaks of 520 mm in 1977 and 510 mm in 2000. The ten-year moving average graph for New South Wales in Figure B.11 shows a steady average of around 500 mm between 1910 and 1930 before falling to a minimum of 430 mm in 1940. Since then the graph peaks at around 600 mm in 1950, 1975 and 1990 with troughs at close to 500 mm in 1960 and 1980 and a steady fall from 1990 to just under 500 mm in 2009. It is clearly difficult to tell whether the current fall is part of the same natural cycles that we have seen over the past 100 years or whether it is the first part of a longer term trend. Similar features are apparent on the graph for Victoria in Figure B.12. The fall from a peak of 700 mm in 1990 to 550 mm in 2009 is more pronounced in Victoria than in New South Wales. The final graphs in Figures B.13 and B.14 show the yearly rainfall records and the median annual rainfall for Sydney and Melbourne respectively. It is fair to say that the recent rainfall in Sydney shows no significant downward trend whereas there has been an extended period of below median rainfall in Melbourne since 1997.

Whether the current situation is part of a long-term downward trend or simply an extended period of drought it is clear that the repercussions have been devastating for both agriculture and the environment. The flow-on effects include a serious downturn in rural industry and rural employment and a potentially significant impact on the Australian economy. The current critical condition of the Murray-Darling Basin highlights the need for special attention to water management, water storage and

water use, management of environmental flows and salinity control.

Van Dijk and Podger [100] discussed various problems associated with the Murray–Darling Basin in water resource management (inequity supply and demand), salinity (land clearing and irrigation drainage), disruptions in river ecosystem (pollution from fertilizer and alga blooms) and climate variability (subtropical, semi-arid and arid). He suggested that a comprehensive model should be developed to incorporate all the problems regarding water management in the Basin. Hence, it is vital to manage the available water resources effectively and efficiently for the future of Australia.

### 1.4.5 The proposed Basin Plan

The Murray–Darling Basin Authority (MDBA) is responsible for preparing the proposed Basin Plan which is based on the Water Act 2007 to further address the issues related to water management in the Basin. The main issue highlighted in the Basin Plan is finding the balance between human needs and restoring sustainable environment [97]. The proposed Basin Plan was released on October, 8th 2010 to the public for discussion.

The proposed Basin Plan also discussed various types of hydrologic modelling including river system and groundwater system. The success of the available model is highly dependent on the availability of a quality data set since some areas have many gaps in the data [97]. Therefore, a hydrologic model which can generate synthetic data sets is needed.

## 1.5 Selected meteorological stations in the Murray–Darling Basin

In this study, we select four meteorological stations located in the south-east region of Australia within the Murray–Darling Basin in the upper Murray river. The stations are Hume Reservoir, Beechworth Composite, Dookie and Rutherglen Research Center, see Figure 1.3 and Table 1.1. The distances and the Spearman's correlation between the stations are given in Table 1.2 and Table 1.3. We use monthly rainfall totals obtained from the Australian Bureau of Meteorology (BOM) for the period 1922–2008. The monthly statistics and plot of monthly rainfall totals averages are given in Table 1.4 and Figure 1.4, respectively.

In general, all four stations have a similar pattern with low average rainfall from December to March and the average rainfall increasing between May and August. Beechworth has the highest monthly rainfall totals averages compared to the other three stations and Rutherglen and Dookie has almost similar rainfall average. We also use daily rainfall totals from Pooraka meteorological station (34.83°S, 138.61°E) in South Australia to illustrate the methodology which will be discussed in Chapter 4, see Figure 1.3.

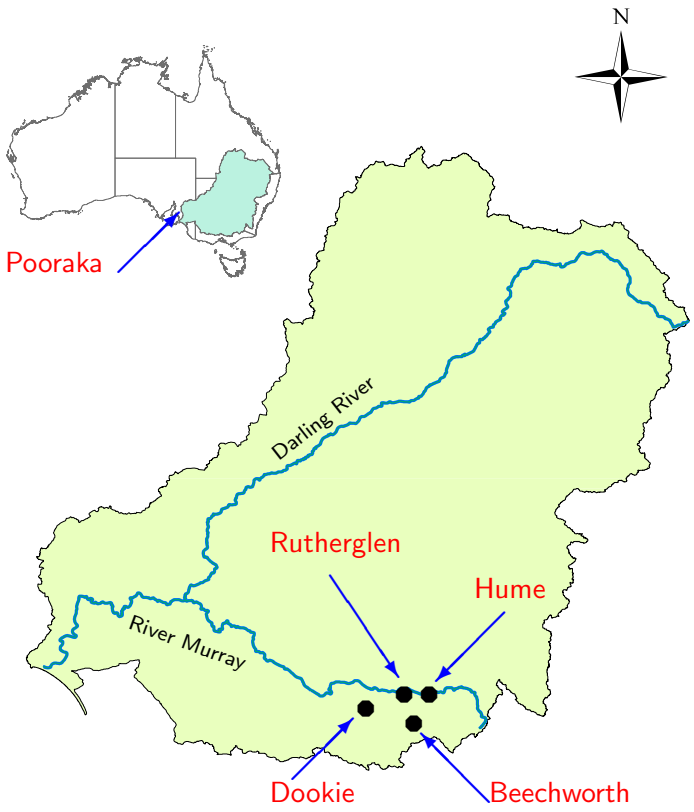



**Figure 1.3:** Selected meteorological stations in the Murray–Darling Basin and Pooraka meteorological station in South Australia

**Table 1.1:** Geographic coordinates and elevations for selected meteorological stations in the Murray–Darling Basin and South Australia

| Station | Abbrev. | State | Latitude (°S) | Longitude (°E) | Elevation (m) |
|---|---|---|---|---|---|
| Hume Reservoir | Hume | NSW | 36.10 | 147.03 | 184 |
| Beechworth Composite | Beec | Vic | 36.70 | 146.71 | 580 |
| Dookie | Dook | Vic | 36.37 | 145.70 | 185 |
| Rutherglen | Ruth | Vic | 36.10 | 146.51 | 175 |
| Pooraka | Pooraka | SA | 34.83 | 138.61 | 21 |

**Table 1.2:** The approximate distances between stations (km)

| | Beechworth | Dookie | Rutherglen |
|---|---|---|---|
| Hume | 72.6 | 123.0 | 46.7 |
| Beechworth | | 97.4 | 69.1 |
| Dookie | | | 78.6 |

**Table 1.3:** Spearman's rank correlation coefficients between stations

| | Beechworth | Dookie | Rutherglen |
|---|---|---|---|
| Hume | 0.90 | 0.84 | 0.91 |
| Beechworth | | 0.87 | 0.91 |
| Dookie | | | 0.88 |

**Table 1.4:** Monthly statistics for the four stations

| | Hume | | Beechworth | | Dookie | | Rutherglen | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Sd | Mean | Sd | Mean | Sd | Mean | Sd |
| Jan | 46.13 | 46.94 | 52.06 | 45.46 | 37.24 | 38.06 | 37.60 | 35.99 |
| Feb | 42.59 | 41.29 | 52.23 | 56.18 | 36.05 | 42.39 | 37.20 | 44.01 |
| Mar | 43.02 | 37.38 | 56.88 | 44.85 | 41.17 | 42.79 | 39.08 | 40.69 |
| Apr | 52.04 | 44.22 | 73.84 | 58.23 | 42.58 | 39.79 | 42.80 | 34.99 |
| May | 58.49 | 42.21 | 92.04 | 63.73 | 49.41 | 32.81 | 50.15 | 37.76 |
| Jun | 67.36 | 40.90 | 106.98 | 62.59 | 52.56 | 31.80 | 55.67 | 34.87 |
| Jul | 77.56 | 39.11 | 119.41 | 58.68 | 57.35 | 27.25 | 63.00 | 35.14 |
| Aug | 76.91 | 37.90 | 117.17 | 51.69 | 57.11 | 31.39 | 61.22 | 30.88 |
| Sep | 61.56 | 32.70 | 92.07 | 43.45 | 49.80 | 27.88 | 53.15 | 28.58 |
| Oct | 68.56 | 45.95 | 96.85 | 54.05 | 53.98 | 37.22 | 57.48 | 35.76 |
| Nov | 51.85 | 36.49 | 69.15 | 43.12 | 40.05 | 31.06 | 45.04 | 31.60 |
| Dec | 51.05 | 44.81 | 66.66 | 51.18 | 38.30 | 35.84 | 45.84 | 40.50 |

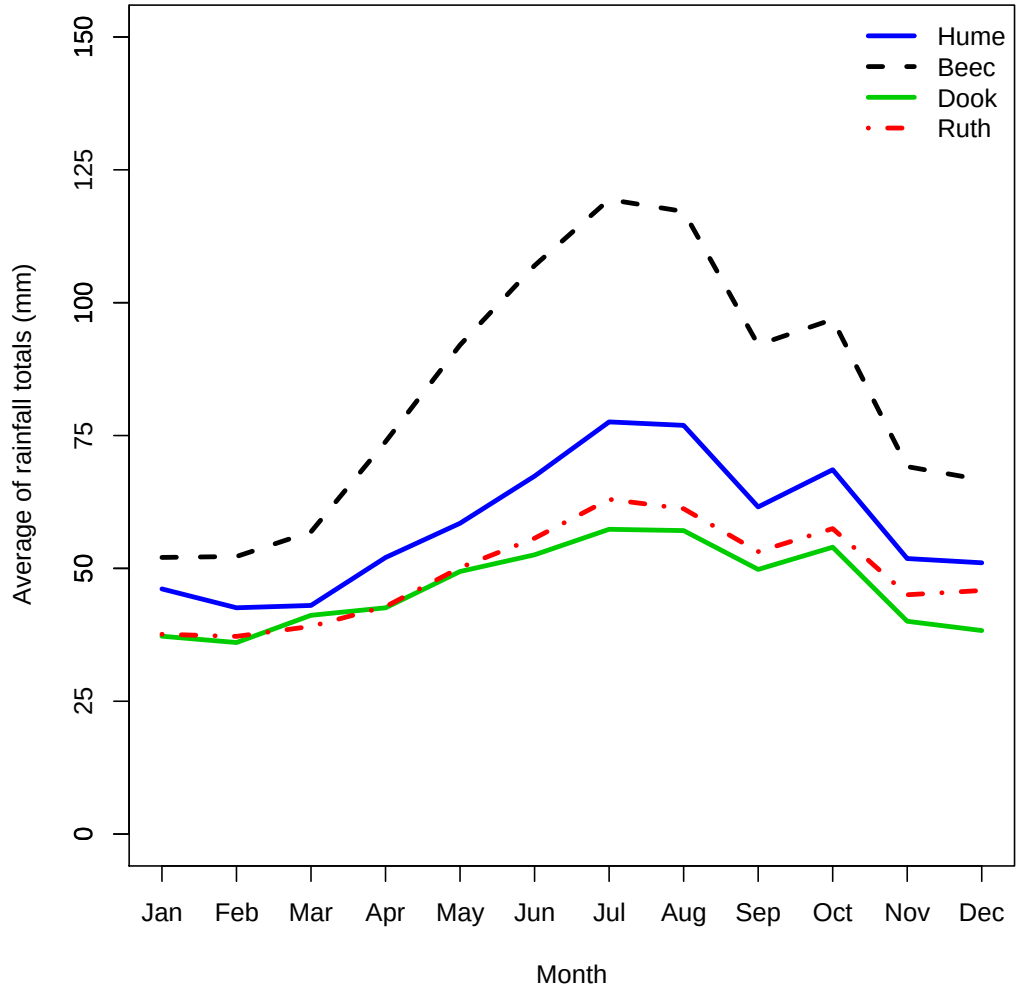



**Figure 1.4:** Average of monthly rainfall totals (mm) for Hume, Beechworth, Dookie and Rutherglen (1922–2008)

## 1.6 Summary

The motivation and aims of this thesis have been presented and the thesis structure outlined. The historical, geographical, the importance and the relevant authorities of the Murray–Darling Basin have been highlighted. Four meteorological stations namely, Hume, Beechworth, Dookie and Rutherglen located in the upper Murray river are selected in this study. In Chapter 2, we will present the literature review in relation to rainfall modelling.

# Chapter 2

# Literature review of rainfall modelling

This chapter gives a synopsis of related literature as outlined in Chapter 1. This literature review will be a basis to acknowledge the latest developments in this field of study. Generally, the differences between the studies in this area include the type of rainfall model (occurrence or amount, or a combination of occurrence and amount) use and type of rainfall data (daily, monthly and yearly) available. Consequently, this will lead to different types of rainfall models to solve different types of problems either in the agriculture or water management sectors. This literature review covers different types of rainfall modelling for precipitation occurrence, precipitation amount or a combination of both and generation of synthetic rainfall data for rainfall simulation models. Precipitation is either in solid (snow and hail) or liquid (rainfall) form.

## 2.1 Rainfall Models

### 2.1.1 Modelling rainfall occurrences

Many studies have applied a Markov model to study rainfall occurrences. According to Roldan and Woolhiser [85], the Markov chain model is superior to an alternating renewal process for modelling daily rainfall occurrence. Stern and Coe [94] stated that a low-order Markov chain model is parsimonious (uses the minimum number of parameters) and hence it provides a better representation. The fitted model can then be used to calculate other quantities such as probability of long wet spells. Furthermore, Stern and Coe [95] state that the Markov model also accommodates seasonality and it offers effective statistical inference procedures for parameter estimation and model selection.

In the context of rainfall modelling, Markov models use an assumption that the probability of rainfall occurrence on a particular day depends only on whether or not it rained on the previous day or days. The first order or simple Markov model assumes that the probability of rainfall on any day depends only on whether it rained on the previous day. According to Chin [19], there is a tendency for rainy days and dry days to cluster and to form sequences which can be described by Markov Chain models of appropriate order corresponding to the order of conditional dependence of the physical phenomena.

As well as having other advantages, the Markov model is considered to be a powerful model for describing the persistency concept. Thus, a Markov model is used to model persistence behaviour of wet and dry spells [12, 62]. A wet spell is defined as consecutive wet days preceded by and followed by at least one dry day. For example, if we define $W$ = wet day when the rainfall amount is over some threshold and $D$ = dry day when the rainfall amount is below the threshold, then the wet spell is

$DW \ldots WD$ and the dry spell will be $WD \ldots DW$. Bekele [12] uses up to a second order Markov model and compares it with a fitted Fourier series up to third order to study the ENSO (El Niño-Southern Oscillation) effect in the rainfall season of Ethiopia. He found that the dry spells are persistent within the minor rainy season. Since dry spells are prominent on most parts of the Spanish Peninsula, Martin-Vide and Gomez [62] apply the length of dry spells to regionalise the area and use up to a tenth order Markov model to fit the dry spells distribution. The study shows that the dry spells (drought) satisfy the Markovian process for the northern part of Spain but not for the central and southern regions.

Some authors have criticised the Markov model. Guttorp [42], for example, shows that the expected length frequency of wet and dry sequences given by a Markov model is always less than the observed sequences. He also states that the Markov model is not suitable for modelling rainfall occurrences at multisites since it involves many parameters. Lall et al. [60] highlight that the Markov model of higher order is not preferable because it is not parsimonious.

To overcome several limitations, Chapman [18] uses and compares several stochastic models (first-order and second-order Markov chains, a truncated negative binomial distribution and a truncated geometric distribution) to model the rainfall occurrence for 22 islands in the Western Pacific (using 18–20 years data sets). He categorises rainfall data into specific classes (neighbouring wet days and single wet days) and the results reveal that data categorisation enhances the model fitting despite additional parameters being required.

### 2.1.2 Modelling rainfall amounts

In general, because the distribution for precipitation amount is skewed to the right, many authors use gamma and exponential distributions as the starting point for

modelling rainfall amount [31, 94]. Todorovic and Woolhiser [98] have analysed the Markov chain-exponential model and binomial-exponential model and claim that both fit reasonably well for modelling total rainfall amount. In contrast, Dunn [30] states that Nick and Lane [71] use the skewed normal distribution, and Zucchini and Adamson [111] use the Weibull distribution and Katz and Parlange [53] use a power transformation to model precipitation amount. It is also known that the gamma distribution can only accommodate strictly positive random variables. Hence, Rosenberg et al. [87] introduced a mixed–gamma distribution to allow a non-zero probability of a zero rainfall total.

Wilks [104] and Woolhiser and Roldan [107] apply the mixed–exponential distribution to model rainfall amounts and Wilks [104] claims that the gamma distribution usually under-estimates the extreme rainfall totals. However, Wu [108] finds that the gamma distribution is the best for modelling the extreme precipitation using the daily rainfall in the Great Miami River watershed in Ohio.

Jamaludin and Jemain [48] compare the normal transform distribution including lognormal, skew normal and mixed–lognormal distributions to model daily rainfall totals at 18 rainfall stations within the Peninsula of Malaysia. They applied different types of goodness of fit tests, namely median of absolute difference, Kolmogorov–Smirnov test, Anderson–Darling test and Cramer-von-Mises to assess the best fit of the models. The mixed-lognormal model is found to be the best to describe the daily rainfall totals. In another study, Jamaludin and Jemain [49] analyse different types of exponential distributions such as mixed–gamma, mixed–Weibull and mixed–exponential to fit the same set of daily rainfall totals. They also apply the similar goodness of fit tests used previously and found that the mixed-Weibull model fits well for most stations. However they highlight that the variation of results of the study are due to geographical, topographical and climatic condition of the stations.

### 2.1.3 Modelling rainfall occurrences and amounts

Generalized linear models (GLM) have been successfully employed to model rainfall occurrence and rainfall amount as discussed above. However, some authors believe that modelling precipitation occurrence and amount separately may not reveal the true feature of the precipitation process or may serve only a particular purpose [17, 2]. The main weakness of GLM is in handling the spatial nature of climate data, [17]. Recently, therefore, many authors have tried to incorporate and accommodate the spatial problem by modelling the rainfall occurrence and amount simultaneously.

Alexandersson [2] uses a compound Poisson-exponential model with two parameters: mean number of independent events (per month) and mean of precipitation amount within the prescribed time interval for each event. He highlights two advantages for using this model: (i) it is flexible enough to handle periods with no precipitation without modification and (ii) parameters can be estimated from the series of monthly precipitation totals. The compound Poisson-exponential can thus be considered practical since it does not require a great deal of calculation.

Dunn [30] shows that the total daily precipitation can be written as the Poisson sum of the gamma random variables which also include days with no precipitation. The Poisson-gamma distribution belongs to the Tweedie family of distributions, [51, 52]. The results prove that the Tweedie family distributions are useful for modelling rainfall on daily and monthly time scales and therefore it is not necessary to have separate models for rainfall occurrence and amount.

Lennox et al. [61] use the Tweedie distribution to generate daily rainfall data as an input for their crop model at Emerald (1900–2002). The Kolmogorov–Smirnov test shows that the observed rainfall distribution is well fitted with the simulated daily rainfall data (combination of 100 simulations). However, the same model fails when it is applied to winter crops (wheat) and this suggests that auto-correlation is

needed for simulated rainfall models.

## 2.2 Generation of synthetic rainfall data for rainfall simulation model

The ground-based meteorological data is often insufficient in terms of length, completeness and spatial coverage. Therefore, statistical models must be used to fill in the missing data and to produce an indefinitely long synthetic weather series. In weather data simulation, precipitation is the most important variable.

In their weather model, Bruhn et al. [14] employ Monte Carlo techniques to simulate daily weather data including precipitation, maximum and minimum temperature, minimum relative humidity and total solar radiation. The results from their study give input for modelling the disease plant management of potato blight. For modelling rainfall occurrence and rainfall totals, Bruhn et al. [14] use the first-order Markov chain and gamma distribution, respectively. A trivariate Normal distribution is then used to simulate temperature data relative to humidity and solar radiation data. To assess the goodness of fit of the generated data, Akaike Information Criteria (AIC) is used for rainfall and Kolmogorov–Smirnov test is used for temperature and humidity. The generated weather data is an appropriate fit to the actual data for Geneva (New York) and Fort Collins (Colorado) in terms of mean response, variability and autocorrelation.

Young [109] uses the multivariate chain to generate both the daily rainfall amount, and the minimum and maximum temperatures concurrently for Tucson and Safford, Arizona. The simulated data is found to fit the actual data for monthly temperature extremes and the median and 90th percentile of monthly rainfall totals. However, the variance and monthly average temperatures and frequency of dry months are

found to be underestimated. Young [109] also highlights that cyclic behaviour in estimating extreme precipitation occurrence should be taken into account for future research.

The need for different time scales of rainfall data from coarse scale (monthly or yearly) to finer scale (daily or hourly) requires reliable and efficient disaggregation or aggregation models. Guenni and Bardossy [41] disaggregate monthly rainfall data to daily rainfall data in two steps: (i) A truncated normal model is used as a reference distribution to initialize the daily rainfall series (ii) the Markov chain Monte Carlo algorithm is used to restructure the series accordingly. The results show that the disaggregation technique applied to the monthly total is successful in preserving the statistical properties for the daily time scale. It is suggested that disaggregation based on the statistical properties is used to provide useful input to regions which are lacking finer time scale data.

Unal et al. [99] compare different methods of generating synthetic annual and monthly rainfall data series. The methods used are the autoregressive (AR) process, the Thomas-Fiering (TF) model, the method of fragments (F) and modified fragments (MF), the two-tier (TT) and wavelet (W) approaches. It is found that the wavelet approach performs well especially in maintaining the statistical characteristics such as the mean of the sequences and correlation between the successive months in the series. Harold [43] and Srikanthan [93] propose a nonparametric model to generate daily rainfall occurrences and amounts on wet days for each Sydney and Melbourne. The nonparametric approach can preserve various time scales rainfall data characteristic as well as maintaining the variability and persistency in the yearly number of wet days despite its complexity.

Piantadosi et al. [76] present a new model to generate daily rainfall data for Parafield in Adelaide, South Australia. A mixed distribution with a non-negative random vari-

able is used for each daily rainfall with either a zero or positive outcome. They note that the monthly standard deviation will be too low if the outcomes are modelled as independent variables. Therefore, generation of correlated daily rainfall totals using a diagonal band copula with a single parameter is suggested. The model maintains the marginal daily distribution as well as means for daily and monthly rainfalls. Furthermore, this method can be applied to higher order lag correlations if more parameters are used.

## 2.3 Other rainfall models

A Poisson-based model is widely used to study a stochastic model for rainfall [26, 23, 25, 82, 83, 24]. Since the Poisson process is two-dimensional in space and time, the model involving the Poisson process is known as the spatial-temporal model. Based on the Poisson process, a stochastic rainfall model at a single site can be represented by a storm event which is associated with a random number of rain cells where each rain cell has independent depth and duration.

Rodriguez-Iturbe et al. [82, 83] introduced a clustered model namely the Neyman-Scott and Bartlett-Lewis model to improve the temporal characteristic of rainfall at different aggregation levels such as from hourly to daily. Using hourly rainfall data from Denver, Colorado, Rodriguez-Iturbe et al. [82] showed that the clustered models gave a better representation on different time scales. However, the probability of zero rainfall for the fitted model was much higher than the observed probability. In [83], they extended the model to improve the discrepancy.

Data which is lacking in terms of length and completeness can cause a problem in modelling hydrological processes. Cowpertwait [23] found the Neyman-Scott model to be particularly useful in hydrology to generate and disaggregate data from daily

to hourly. Using a uniform distribution, the Neyman-Scott model is redefined by Cowpertwait [24] to obtain a parsimonious model in order to accommodate different types of precipitation.

## 2.4 Knowledge gap in the rainfall modelling

Table 2.1 shows a summary of the rainfall models used in the literature and we also highlight the significance of the current study.

**Table 2.1:** Summary of rainfall models

| Model | Properties | Limitation |
|---|---|---|
| Markov model | - model rainfall occurrence (wet/dry spell)<br>- define: {0}=dry, {1}=wet<br>- better than renewal process for modelling occurrence<br>- good for parameter estimation | - higher order not parsimonious<br>- not suitable for multisite<br>- do not include higher order correlation |
| gamma model | - use nonzero rainfall totals | - underestimate extreme rainfall totals |
| generalised linear model | - occurrence and rainfall totals | - incapable to handle spatial data |
| exponential; Markov-chain exponential; binomial-exponential;<br>skewed normal; Weibull; mixed-gamma;<br>mixed-exponential; log normal; skewed normal;<br>mixed-log normal; mixed-weibull; | | - do not include correlated variables |
| Alouini model | - consider independent and dependent variables | - applied in telecommunication field |
| Our model | - extension from Moschopoulos and Alouini models<br>- consider independent and dependent variables<br>- derive new formulae for various means and variances using method of moments<br>- spatial and temporal analysis | - applied in rainfall modelling |

## 2.5 Summary

Various types of rainfall modelling are available in the literature. Different models are used to model rainfall occurrence and rainfall amounts independently and to model both rainfall occurrence and amounts simultaneously. Different types of models are used to generate synthetic rainfall data including Markov models, the

wavelet approach and mixed distributions. Other models based on the Poisson process are also described. Chapter 3 will outline the mathematical and statistical tools required for this study.

# Chapter 3

# Mathematical and Statistical Tools

In this chapter, I describe tools that I need to perform the analysis inherent in this study. The theory of the gamma distribution is discussed in detail with plots for various shape and scale parameters. The gamma distribution is used to fit the monthly rainfall data. The techniques for generating synthetic rainfall totals using a probability density function are also demonstrated. Included is a discussion of Spearman's rank correlation to calculate correlation coefficients for data that do not follow a normal distribution. The generalised correlative coherence is applied to calculate the overall correlation value for more than two series. Finally, there is a discussion of various goodness of fit tests such as Chi–Square, Kolmogorov–Smirnov and Anderson–Darling for assessing the fit between the observed and generated data.

## 3.1 The gamma distribution

The gamma distribution has been used widely in diverse applications such as rainfall modelling [92, 105, 87, 47, 76], insurance losses [35] and wind farm output [13]. It is suitable for data on $(0, \infty)$. Husak et al. [47] review studies of the gamma distri-

bution characteristics for fitting rainfall data in Africa. They highlighted why the gamma distribution is always an initial choice to model rainfall data: it fits positive data, is a good representation of rainfall distribution and is sufficiently flexible to fit various rainfall regimes from exponential to normal shape. The flexibility of the gamma distribution is due to two parameters; shape and scale, see Wilks [103]. The exponential distribution is a special case when the value for the shape parameter is one.

The gamma distribution is also used to generate synthetic rainfall data [92, 105, 87, 76]. Synthetic rainfall data generation is important as a supplement to the historical rainfall data which is always lacking in terms of temporal and spatial extent. A comprehensive review of weather generation models is given by Srikanthan and McMahon [92] and Wilks and Wilby [105]. Piantadosi et al. [76] applied a gamma distribution to generate synthetic monthly rainfall totals and proposed a rainfall generator algorithm (rGEN) to generate synthetic rainfall totals on different timescales (daily, monthly and yearly). Rosenberg et al. [87] generalised the gamma density to a series of associated Laguerre polynomials to generate monthly rainfall totals and showed that the simulated data and the observed data had the same statistical characteristics.

### 3.1.1 Probability density function

Consider a random variable $X$ of a gamma distribution denoted by $X \sim G(\alpha, \beta)$, with the probability density function (PDF) of

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \qquad \text{for } x > 0 \text{ and } \alpha, \beta > 0 \tag{3.1}$$

where $\alpha$ and $\beta$ are the shape and scale parameters, respectively. The incomplete gamma function $\Gamma(\alpha)$ is defined as

$$\Gamma_x(\alpha) = \int_0^x t^{\alpha-1}e^{-t}\,dt.$$

When $\beta = 1$, the gamma density is in the standard form of

$$f_X(x;\alpha,1) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}.$$

Some special types of gamma distributions are the exponential distribution for $\alpha = 1$, the Erlang distribution for $\alpha$ positive integer and the Chi-Square distribution for $\alpha = \nu/2$ and $\beta = 2$ where the parameter $\nu$ denotes the number of degrees of freedom. For the gamma distribution, there are two free parameters, namely, the shape parameter ($\alpha$) and the scale parameter ($\beta$). The shape parameter determines the skewness of the distribution and the scale parameter affects the variance of the distribution, see Fig. 3.1 and refer to [50, 72]. As the value of the shape parameter increases, the shape of the distribution changes from a skewed exponential type to an essentially symmetric bell shape. Since many studies used the gamma distribution to match rainfall totals, this study can be compared with previous studies, see Chapters 4 to 8.

The value of the cumulative distribution function (CDF) is calculated by integrating the PDF and is written as $F(x) = P(X \leq x)$. Consider the gamma CDF denoted by $F(x;\alpha,\beta)$ where

$$F(x;\alpha,\beta) = \int_0^x \frac{\xi^{\alpha-1}e^{-\frac{\xi}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}\,d\xi. \tag{3.2}$$

The derivation is as follows. The mean is

$$\begin{aligned}
\mu = E(x) &= \int_0^\infty x f(x;\alpha,\beta)\, dx \\
&= \int_0^\infty x \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\, dx \\
&= \int_0^\infty \frac{x^{(\alpha+1)-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\, dx \\
&= \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty \frac{x^{(\alpha+1)-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha+1)\beta^{\alpha+1}}\, dx \\
&= \frac{\alpha\Gamma(\alpha)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha} \times 1 \\
&= \alpha\beta
\end{aligned} \tag{3.3}$$

and the variance is

$$\begin{aligned}
\sigma^2 = var(x) &= E(x^2) - [E(x)]^2 \\
&= \int_0^\infty x^2 \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\, dx - (\alpha\beta)^2 \\
&= \int_0^\infty \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\, dx - (\alpha\beta)^2 \\
&= \frac{\Gamma(\alpha+2)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha+2)\beta^{\alpha+2}}\, dx - (\alpha\beta)^2 \\
&= \frac{\alpha(\alpha+1)\Gamma(\alpha)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} \times 1 - (\alpha\beta)^2 \\
&= \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 \\
&= \alpha\beta^2.
\end{aligned}$$

The shape of the graph of gamma density function is skewed (asymmetric) to the right or positively skewed. Some graphs of the PDF and CDF of gamma distribution are shown in Figure 3.1 and Figure 3.2 for various values of $\alpha$ and $\beta$. In the left plots of Figure 3.1, we observe that for a fixed value of $\beta$, the shape of the graph changes according to the values of $\alpha$ from an exponential distribution shape to a

normal distribution shape as $\alpha$ increases. However, the plots on the right show that the shape remains skewed but the peakedness of the plots is decreasing as the value of $\beta$ increases.

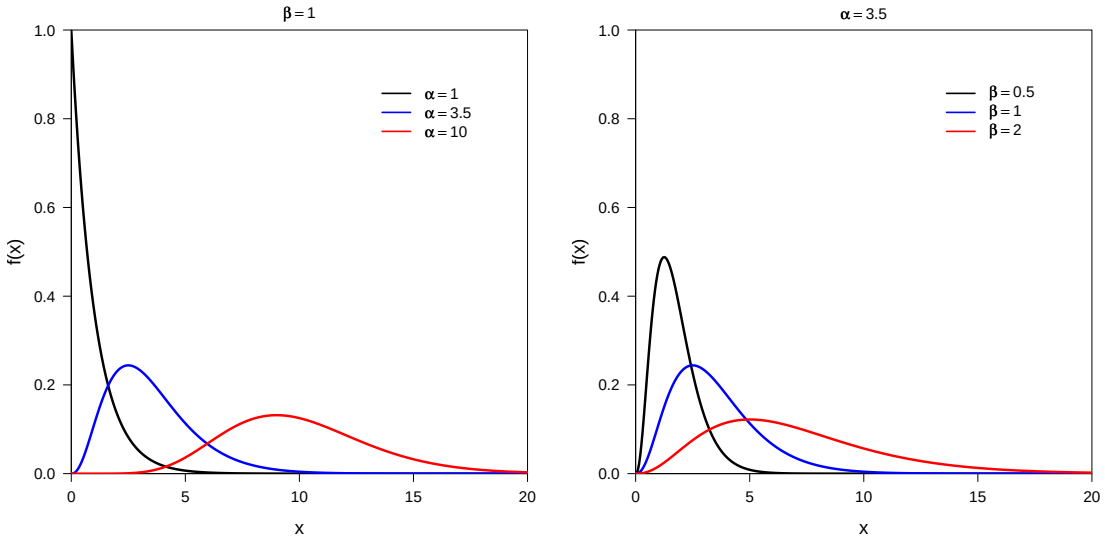



**Figure 3.1:** Gamma probability density functions for various parameters $\alpha$ and $\beta$

### 3.1.2 Estimating parameters for the gamma distribution

In the previous section, we have specified that the gamma distribution with shape and scale parameters is chosen to fit the rainfall data. We will now describe parameter estimation, the process of determining the value of the parameters that will give the best fit to the data. Parameter estimation methods include least-squares; the method of moments; and maximum likelihood estimation (MLE). In most cases we use MLE method to estimate the parameters. However, the method of moments is applied to estimate the parameters and to derive some of the formulae (e.g. 5.4-5.6, 8.1-8.4 and 8.9), refer to Chapters 5 to 8.

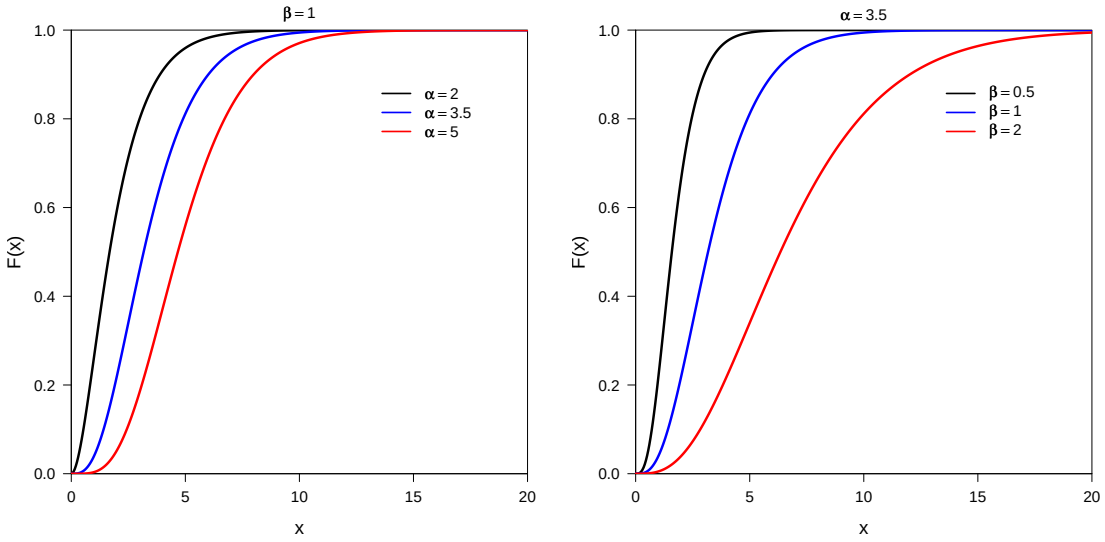



**Figure 3.2:** Gamma cumulative distribution functions for various parameters $\alpha$ and $\beta$

The MLE method was developed by R.A. Fisher in the 1920 s and since then it has been used widely due to sufficiency, consistency and efficiency properties it has [69]. In this study, we will show that the maximum likelihood estimation method automatically matches the mean but not necessarily the variance. The parameters $\alpha$ and $\beta$ are estimated from the observed data using the maximum likelihood method (MLE) via the Newton–Raphson iteration method.

Consider a set of random samples $x_1, x_2, \ldots, x_n$ with the gamma probability density function (3.1). Then the likelihood function ($L$) is the product of the PDFs for each $x_i$ that represents the rainfall distribution given by

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} f(x_i, \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{x_i^{\alpha-1} e^{-\frac{x_i}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \\ &= \frac{\prod_{i=1}^{n} x_i^{\alpha-1} e^{-\frac{\sum x_i}{\beta}}}{[\Gamma(\alpha)\beta^{\alpha}]^n}. \end{aligned}$$

By taking the logarithm of the likelihood function (known as log-likelihood), the product is written in the summation form given by

$$\ln(L) = (\alpha - 1)\sum_{i=1}^{n} \ln(x_i) - \frac{\sum_i^n x_i}{\beta} - n\ln\Gamma(\alpha) - n\alpha\ln\beta.$$

The parameter estimation method that maximises the log-likelihood function will also maximise the likelihood function. To estimate the parameters $\alpha$ and $\beta$, the log-likelihood function is maximised by taking the partial derivatives with respect to each parameter $\alpha$ and $\beta$, respectively and setting both partial derivatives to zero. The resulting log-likelihood equations (likelihood equation) are

$$\begin{aligned} \frac{\partial \ln L}{\partial \alpha} &= \sum_{i=1}^{n} \ln(x_i) - n\Psi(\alpha) - n\ln(\beta) = 0 \\ \frac{\partial \ln L}{\partial \beta} &= \frac{\sum_i^n x_i}{\beta^2} - \frac{n\alpha}{\beta} = 0 \end{aligned} \tag{3.4}$$

where

$$\Psi(\alpha) = \frac{\partial \ln\Gamma(\alpha)}{\partial \alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

is known as the digamma function. The system of equations (3.4) is solved numerically using the Newton–Raphson iteration method. The Newton method or Newton–Raphson method was named after Sir Isaac Newton (1643–1727) and Joseph Raphson (approximately 1648–1715). Sir Isaac Newton first introduced the important form of the method and Raphson modified and improved the method [15]. The Newton–Raphson method is considered efficient since it will often converge quickly if the initial guess supplied is close enough to the root of the equation. Basically, if we want to solve a nonlinear equation of $f(x) = 0$, the Newton–Raphson iteration method is given by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

where $x_k$ is the initial guess of the root of the equation and $x_{k+1}$ is the new value of the root. The Newton–Raphson method can be extended to $n$ nonlinear equations and unknowns (multidimensional). Refer to Komzsik [57] for further details. For example, if we want to solve two nonlinear equations

$$f_1(x_1, x_2) = 0 \quad \text{and} \quad f_2(x_1, x_2) = 0,$$

the system of equations is presented in Taylor's series form of

$$\begin{aligned} f_{1,k+1} &= f_{1,k} + (x_{1,k+1} - x_{1,k})\frac{\partial f_{1,k}}{\partial x_1} + (x_{2,k+1} - x_{2,k})\frac{\partial f_{1,k}}{\partial x_2} \\ f_{2,k+1} &= f_{2,k} + (x_{1,k+1} - x_{1,k})\frac{\partial f_{2,k}}{\partial x_1} + (x_{2,k+1} - x_{2,k})\frac{\partial f_{2,k}}{\partial x_2}. \end{aligned}$$

By rearranging the equations and setting $f_{1,k+1} = 0$ and $f_{2,k+1} = 0$, the matrix form of the system of equations is written as

$$\begin{bmatrix} \frac{\partial f_{1,k}}{\partial x_1} & \frac{\partial f_{1,k}}{\partial x_2} \\ \frac{\partial f_{2,k}}{\partial x_1} & \frac{\partial f_{2,k}}{\partial x_2} \end{bmatrix} \begin{bmatrix} x_{1,k+1} - x_{1,k} \\ x_{2,k+1} - x_{2,k} \end{bmatrix} = - \begin{bmatrix} f_{1,k} \\ f_{2,k} \end{bmatrix}$$

where the matrix of the partial derivatives is known as the Jacobian matrix. Therefore, the Newton–Raphson method is written in matrix form as

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \end{bmatrix} = \begin{bmatrix} x_{1,k} \\ x_{2,k} \end{bmatrix} - J_k^{-1} \begin{bmatrix} f_1(x_{1,k}) \\ f_2(x_{2,k}) \end{bmatrix}$$

where the Jacobian matrix is given by

$$J_k = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}$$

and $J_k^{-1}$ is the inverse of the Jacobian matrix and $k$ is the number of iterations. Therefore for our parameters,

$$\begin{bmatrix} \alpha_{k+1} \\ \beta_{k+1} \end{bmatrix} = \begin{bmatrix} \alpha_k \\ \beta_k \end{bmatrix} - J_k^{-1} \begin{bmatrix} f(\alpha_k) \\ f(\beta_k) \end{bmatrix}$$

where the Jacobian matrix is given by

$$J = \begin{bmatrix} -n\Psi'(\alpha) & -\frac{n}{\beta} \\ -\frac{n}{\beta} & -\frac{2x_i}{\beta^3} + \frac{n\alpha}{\beta^2} \end{bmatrix}$$

and

$$J_k^{-1} = \frac{1}{|J|} \begin{bmatrix} -\frac{2x_i}{\beta^3} + \frac{n\alpha}{\beta^2} & \frac{n}{\beta} \\ \frac{n}{\beta} & -n\Psi'(\alpha) \end{bmatrix}.$$

An alternative solution method to the Newton–Raphson method is to use the second equation of equation 3.4 to eliminate one variable leaving a single equation in one variable to be solved. Also, the second equation of equation 3.4 ensures the fitted gamma distribution matches the data mean. The MLE estimate does not match the data variance, but alternatively $\alpha$ and $\beta$ can be chosen to match both mean and variance.

## 3.2 Generation of synthetic positive rainfall totals

The rainfall data is categorised into two sets, zero totals and positive totals. Zero totals refers to when there is no rain and positive totals refers to when there is more than 0.1 mm of rain. In equation (3.1), the gamma distribution is defined for strictly positive rainfall totals and the CDF is given in equation (3.2). Consider the CDF for a particular value

$$F(x_n) = r_n$$

where $x_n$ is a rainfall amount associated to probability that is $P(X \leq x_n) = r_n$. Therefore, if we have a value of $r_n$, the value of $x_n$ is determined by the inversion

$$x_n = F^{-1}(r_n). \tag{3.5}$$

To generate synthetic positive rainfall totals, first we generate a set of random numbers from the Uniform distribution on [0,1]. We use the code $qgamma(...)$ in the $R$ package for evaluating the rainfall totals, using equation (3.5). The value of the shape and scale parameters are estimated from the observed data using the method described in Section 3.1.2. The process is visualised in Figure 3.3. For example, if the random number is 0.6 and if $\alpha = 1.0767$ and $\beta = 18.7220$ then the rainfall total is 77.0 mm.

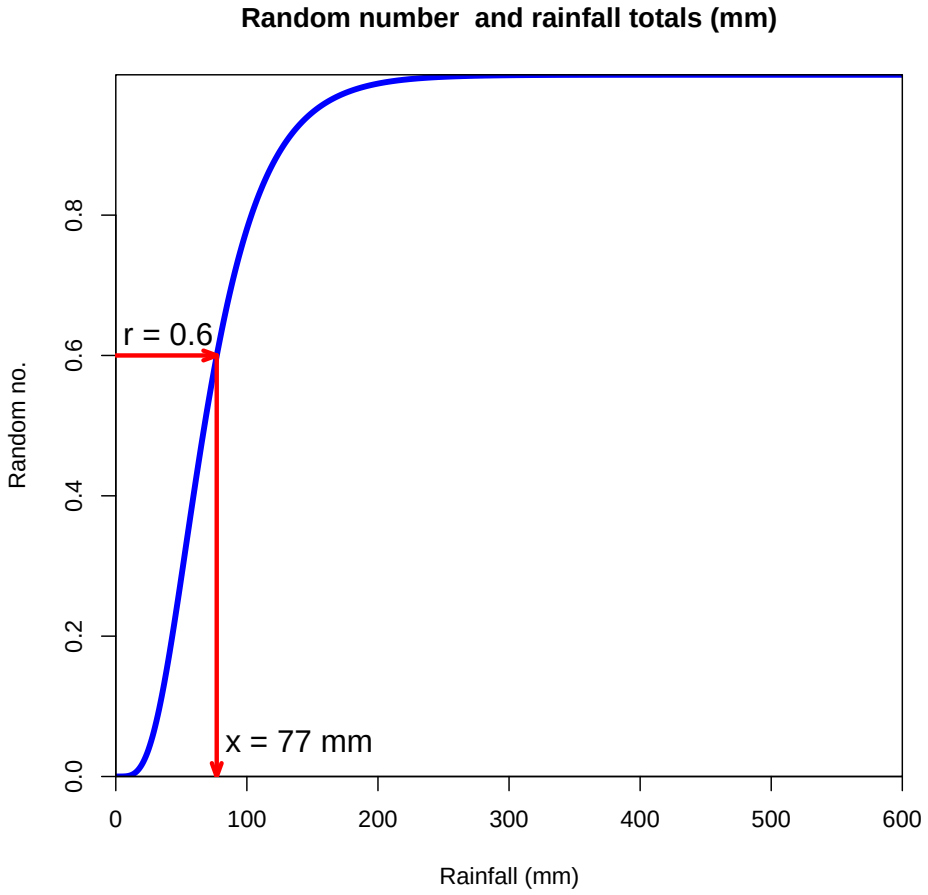



**Figure 3.3:** Synthetic Rainfall totals from cumulative distribution function

## 3.3 Generation of synthetic positive rainfall totals from the probability density function

In many cases, the cumulative distribution function (CDF) is not readily available and difficult to derive. Therefore, we use the probability density function to generate the synthetic data, for which we detail the process. Given a sequence of rainfall totals $0, 1, \ldots, 1000$, we calculate the height of the probability density function for each total. Then, we multiply the height by the interval size to get the area within the interval which is the probability of rainfall being in that interval given by

$$P(X_i = x_i) = f(x_i) \times \Delta x_i$$

where the PDF is $f(x_i)$ and $\Delta x_i$ the interval size. Then the cumulative distribution function is calculated as

$$F(x_i) = P(X \leq x_i).$$

A *lookup* table of the CDF associated with the rainfall totals (number sequences) is constructed as shown in Table 3.1. As an example, if we wish to generate 1000 rainfall totals at random from the CDF, we generate 1000 uniformly distributed random numbers on $[0, 1]$ and use the $'vlookup'$ function in *Excel* to match the random number with the CDF value and obtain the rainfall totals. The value of the sequence number corresponds to rainfall totals. For example, if the maximum value of rainfall totals is 550 mm, we need a set of sequence numbers up to 600 and with 30 intervals of width 20 for each interval. Table 3.1 shows a sample of *lookup* table constructed. In Table 3.1, it is highlighted that for random value of 0.8283 is associated to a rainfall total of 360 mm. The values of parameters we use to evaluate the PDF values are shape parameter $\alpha = 3.2472$ and scale parameter $\beta = 20.9475$.

**Table 3.1:** Lookup table for synthetic rainfall

| PDF | CDF | Sequence | Random number | Generated |
|---|---|---|---|---|
| 0.00E+00 | 0.00E+00 | 0 | 0.9441 | 440 |
| 1.34E-11 | 2.67E-10 | 20 | 0.4999 | 260 |
| 3.11E-08 | 6.23E-07 | 40 | 0.6824 | 300 |
| 1.76E-06 | 3.58E-05 | 60 | 0.6725 | 300 |
| 2.19E-05 | 4.74E-04 | 80 | 0.3268 | 220 |
| 1.20E-04 | 2.88E-03 | 100 | 0.2767 | 220 |
| 3.97E-04 | 1.08E-02 | 120 | 0.8283 | 360 |
| 9.29E-04 | 2.94E-02 | 140 | 0.0494 | 140 |
| 1.70E-03 | 6.34E-02 | 160 | 0.0198 | 120 |
| 2.59E-03 | 1.15E-01 | 180 | 0.9234 | 420 |
| 3.43E-03 | 1.84E-01 | 200 | 0.1196 | 180 |
| 4.09E-03 | 2.65E-01 | 220 | 0.2564 | 200 |
| 4.47E-03 | 3.55E-01 | 240 | 0.8295 | 360 |
| 4.57E-03 | 4.46E-01 | 260 | 0.1810 | 180 |
| 4.41E-03 | 5.34E-01 | 280 | 0.3513 | 220 |
| 4.08E-03 | 6.16E-01 | 300 | 0.3549 | 220 |
| 3.63E-03 | 6.89E-01 | 320 | 0.7340 | 320 |
| 3.14E-03 | 7.51E-01 | 340 | 0.1274 | 180 |
| 2.64E-03 | 8.04E-01 | 360 | 0.7531 | 340 |
| 2.17E-03 | 8.48E-01 | 380 | 0.7651 | 340 |

## 3.4 Generation of synthetic non-negative rainfall totals

The extension of the model to non-negative values which include zero totals is given by Rosenberg et al. [87] where the cumulative distribution function of a mixed–gamma distribution (mixture of probability of zero and positive rainfall totals) is

$$F(x;\alpha,\beta) = p_0 + (1-p_0)\int_0^x f(\xi;\alpha,\beta)\ d\xi \tag{3.6}$$

where

$$\begin{aligned} p_0 &= \text{probability of no rainfall} \\ f(\xi;\alpha,\beta) &= \text{a gamma density with parameters } \alpha,\beta. \end{aligned}$$

The probability of zero rainfall is estimated by

$$p_0 = P(X=0) = \frac{n}{N}$$

where $n$ is the count of zero rainfall totals and $N$ is the total count of the rainfall totals.[1] To generate synthetic rainfall totals $x_n$, we generate a sequence of random numbers $r_n$ in the interval of $(0,1)$ from a Uniform distribution. From the CDF of equation (3.6), we determine the rainfall totals using the following procedure. If

$$r_n < p_0 \quad \text{then} \quad x_n = 0$$

and if

$$r_n \leq p_0 \quad \text{then} \quad x_n = F^{-1}\left(\frac{r_n - p_0}{1-p_0}\right).$$

[1]Strictly speaking we should write $\hat{p_0}$ to denote the estimated value of $p_0$. In this thesis, for reasons of typographical convenience, we have chosen not to make a notational distinction between the parameter and the estimated value.

For example, let $p_0 = 0.6$ and $r_1 = 0.5$ and $r_2 = 0.8$ then the rainfall totals associated with the random numbers are $x_1 = 0\,\text{mm}$ ($r_1 < p_0$) and $x_2 = 60.5\,\text{mm}$ ($r_2 > p_0$), respectively.

## 3.5 Spearman's rank correlation coefficient

Correlation is a measure of the degree of association or relationship between two variables. There are two commonly used formulae–Pearson's correlation and Spearman's rank correlation. Pearson's correlation which is also known as the product-moment correlation, assumes that the two data sets are normally distributed while the Spearman's rank correlation does not. Refer to [81] and [29]. The notation used for Pearson is $r$ and for Spearman is $r_s$.

The formula for the Pearson correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

where $x_i$ and $y_i$ are the observed data from two sets of data and $\overline{x}$ and $\overline{y}$ are the respective means. The test of significance for the correlation is based on the hypothesis test of

$$\begin{aligned} H_0 &: \rho = 0, \quad \text{there is no correlation between the data} \\ H_1 &: \rho \neq 0, \quad \text{there is correlation between the data} \end{aligned}$$

where $\rho$ is the population correlation coefficient, and the test statistic is

$$t_{calc} = r\sqrt{\frac{n-2}{1-r^2}}. \tag{3.7}$$

Here $r$ is the sample correlation and $n$ is the number in each data set. $t_{calc}$ is from the Student-t distribution with $n - 2$ degrees of freedom. It can thus be converted

to a probability (P-value). If the probability is sufficiently small the hypothesis $H_0$ can be rejected to conclude there is correlation between data sets. It is conventional to use a test level of 5% to reject a hypothesis. The 5% value corresponds to there being only 1 change 20 of the P-value being this low due to random variation when $H_0$ is true. Lower values of P are even less likely to have occurred when $H_0$ is true giving stronger evidence that $H_0$ should be rejected.

Since we are dealing with non normal data, we use the Spearman's rank correlation coefficients denoted by $r_{ij}$. The formula of the Spearman's rank correlation coefficient is calculated based on the ranks of each data element. This is a nonparametric measure. Consider two sets of data, $\{X_i\}$ and $\{Y_i\}$, for $i = 1, \ldots, n$. First, we rank both sets of data separately as $R(x_i)$ and $R(y_i)$, for $i = 1, \ldots, n$ from the smallest to the largest. Then, the formula for the Spearman's rank correlation coefficients $r_{ij}$ of $X_i$ and $X_j$ is

$$r_{ij} = 1 - \frac{6 \sum_{k=1}^{n} d_k^2}{n(n^2 - 1)} \tag{3.8}$$

where $d_k = R(x_k) - R(y_k)$ is the difference between the ranked pairs and $n$ is the number of pairs. The same hypothesis for the significance test of correlation is applied with the test statistic as in equation (3.7). Since the rainfall data is usually skewed, the use of Spearman's rank correlation coefficient is more appropriate for determining correlation between rainfall totals at separate locations.

## 3.6 Generalised correlative coherence

In Chapter 6, I will describe a method for finding the distribution of a sum of $n$ positively correlated gamma random variables. To overcome the problem of negative correlation we first use the method described by Getz [38]. This method was used to

construct a measure of overall correlation between more than two series of correlated variables, rather than simply using a set of pairwise correlations. Getz [38] used a generalised correlative coherence and introduced a Shannon–Weaver-type measure for the normalised eigenvalues of the $n$-dimensional correlation matrix. Boland [13] also applied a similar method to estimate the overall correlation for a number of wind farms. The correlative coherence method will rectify the negative correlation problem and only then we apply the method for the sum of $n$ correlated gamma variables.

Firstly, we define matrix $R = [r_{ij}]$ where $r_{ii} = 1$ and $r_{ij} = r \in [0, 1]$ for $i \neq j$ as a correlation matrix and calculate its eigenvalues $(\lambda_i, i = 1, \ldots, n)$ which satisfy $0 \leq \lambda_i \leq n$ and $\sum_{i=1}^{n} \lambda_i = n$, which can also be expressed as $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i/n = 1$. Therefore, a well-defined Shannon–Weaver-type measure of the diversity of the eigenvalues $\lambda_i$ of $R$ is

$$C(X^n) = 1 - \frac{1}{\log(1/n)} \sum_{i=1}^{n} (\lambda_i/n) \log(\lambda_i/n). \tag{3.9}$$

Getz [38] referred to the expression of $C(X^n)$ as a general representation of the coherence concept for periodic signals which are synchronised to the $X_n$ signals. The correlation matrix, $R(r)$ is constructed by considering all off-diagonal entries to be the same value where $r \in [0, 1]$ and the diagonal entries to be 1. Then the eigenvalues of the correlation matrix $R(r)$ are defined as

$$\lambda_1 = 1 + (n-1)r \quad \text{and} \quad \lambda_i = 1 - r, \ i = 2, \ldots, n$$

which gives

$$C_n(r) = \frac{[1 + (n-1)r] \log[1 + (n-1)r] + (n-1)(1-r) \log(1-r)}{n \log n}.$$

The correlative coherence value $r$ is obtained by equating $C(X^n) = C_n(r)$ and solving by Newton's method where $r = C_n^{-1}(C(X^n))$. Note that this approach works since $C$ is positive definite as will be described in Chapter 6.

## 3.7 Goodness of fit tests

We describe various nonparametric goodness of fit tests available to be used in the rainfall model analysis, namely Chi–Square, Kolmogorov–Smirnov and Anderson–Darling. The nonparametric tests are also known as 'distribution-free' where no stringent assumptions are required about the data. These tests are recommended for skewed or non-normal data if the data to be analysed do not satisfy the test assumption for a parametric test [73]. A goodness of fit test assesses how close the empirical distribution is to a specified distribution. Essentially, the test is based on either the probability density function (PDF) or cumulative distribution function (CDF). The Chi–Square test uses the PDF while the Kolmogorov-Smirnov and Anderssen-Darling tests both use the CDF. The Kolmogorov–Smirnov and Anderson–Darling tests are categorised in the class of 'distance test'. The Chi–Square test is applicable for discrete or categorical data and the Kolmogorov–Smirnov and Anderson–Darling tests are for continuous data. However with an extension of the theory of Kolmogorov–Smirnov and Anderson–Darling tests, they are also applicable for grouped or discrete data, [88, 75, 46]. Horn [46] gives a review for Chi–Square and Kolmogorov–Smirnov goodness of fit tests for discrete data with application in a health impairment scale. Since the Kolmogorov–Smirnov test is exact, she suggested that the Kolmogorov–Smirnov test is used in preference to the Chi–Square test.

### 3.7.1 P-value

The P-value refers to the probability of a chosen statistic calculated under some assumption, often referred to as the null hypothesis or $H_0$. A low P-value is unlikely

to have occurred if the null hypothesis $H_0$ is true but can easily occur if the $H_0$ hypothesis is false. The smaller the P-value the less likely it is to have occurred when $H_0$ is true. Often if the P-value is less than 0.05 (the test value is often denoted as $\alpha$) it is considered sufficiently unlikely and the null hypothesis is rejected. An equivalent procedure to testing using a given P-value is to compare the chosen statistic with a critical value obtained from statistical tables (e.g. [84]).

Different statistics measure different properties and can thus give different results at a given test level. This is particularly true where distributions, which have may different aspects, are compared. However the P-values for the different statistics usually indicate a similar conclusion.

### 3.7.2 Chi–Square test

The Chi–Square test or Pearson's Chi–Square was derived by Karl Pearson (1857–1936) in 1900 [20]. Since then, it has been used widely to compare the frequencies or counts of observed data within specified intervals with those expected from the proposed distribution. Cochran [20] studied the historical development and analysed various aspects of the Chi–Square test such as the rule for determining the degrees of freedom. The Chi–Square test is easy to compute and has no restriction in terms of sample size and types of distribution, either univariate or multivariate. However it is limited by the rule that all the counts in each cell must be greater than one and having no more than 20% of the counts with value less than 5. The rule is known as Cochran's Rule [21]. It is recommended that if the cell count is less than 5, adjacent cells should be combined. The binning or grouping process does sometimes affect the result of an analysis as some of the information may be lost. Normally, the grouping occurs at the tails or extreme part of the distributions where the most significant differences between the observed and expected values occur. Hence, the

Chi–Square test will be less powerful than the Kolmogorov–Smirnov and Anderson–Darling tests. The hypothesis of the Chi–Square test for homogeneity is

$H_0$ : No difference between the observed and expected frequencies

$H_1$ : The observed and expected frequencies are different

and the test statistic is

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.10}$$

where $E_i$ and $O_i$ are the expected and observed counts, respectively with $k$ number of classes and $k-1$ degrees of freedom. If the value of the $\chi^2$ statistic is less than the critical value – measure the P-value $> \alpha$, then there is not enough evidence that the observed and expected distributions are significantly different. The table of critical values of $\chi^2$ statistic is in Rohlf and Sokal [84]. In the *R* package, the statistic and P-value of $\chi^2$ is calculated using the code *chisq.test*(...).

### 3.7.3 Kolmogorov–Smirnov test

The Kolmogorov–Smirnov goodness of fit test was introduced by two Russian mathematicians, A. N. Kolmogorov in 1933 and N. V. Smirnov in 1939 [11]. At first, Kolmogorov introduced a test for one sample and later Smirnov introduced a two-sample test. Since the tests were similar, the tests were then called Kolmogorov–Smirnov one-sample test and Kolmogorov–Smirnov two-sample test, respectively. In the Kolmogorov–Smirnov one-sample test, we compare the CDF of observed distribution with the CDF of the hypothesised distribution. For the Kolmogorov–Smirnov two-sample test, we compare two independent samples. The Kolmogorov–Smirnov test is considered to be nonparametric test where no assumption is needed about

the distributions of data [29].

In this study, we apply the Kolmogorov–Smirnov two-sample test to compare cumulative distributions between the grouped data of the observed and generated. Pettitt and Stephens [75] show that the Kolmogorov–Smirnov goodness of fit test is applied successfully for discrete and grouped data. They produce statistical tables for grouped data of sample size up to 30.

The cumulative distribution function of a random variable $X$ is defined as $F(x) = P(X \leq x)$. The hypothesis of the two-sample Kolmogorov–Smirnov test is

$$\begin{aligned} H_0 &: F_1(x) = F_2(x), \text{ two data samples come from the same distribution} \\ H_1 &: F_1(x) \neq F_2(x), \text{ they do not come from the same distribution} \end{aligned}$$

and the test statistic is

$$D_{1,2} = \overset{\max}{x} |F_1(x) - F_2(x)| \tag{3.11}$$

where $F_1(x)$ and $F_2(x)$ are the cumulative distribution functions of the observed and generated data. The value of $D$ determines the absolute maximum difference between the CDF of the observed and generated data as shown in Figure 3.4. The hypothesis is not rejected when the computed $D$ value is less than the critical value or the P-value is greater than the significance level of $\alpha = 0.05$. It indicates that there is no evidence that the two distributions are significantly different from each other. The $D$ and P-value are computed using the $ks.test(...)$ code in $R$-package, see Appendix C.7. The table of critical values of the two-sample Kolmogorov–Smirnov statistic is in Rohlf and Sokal [84].

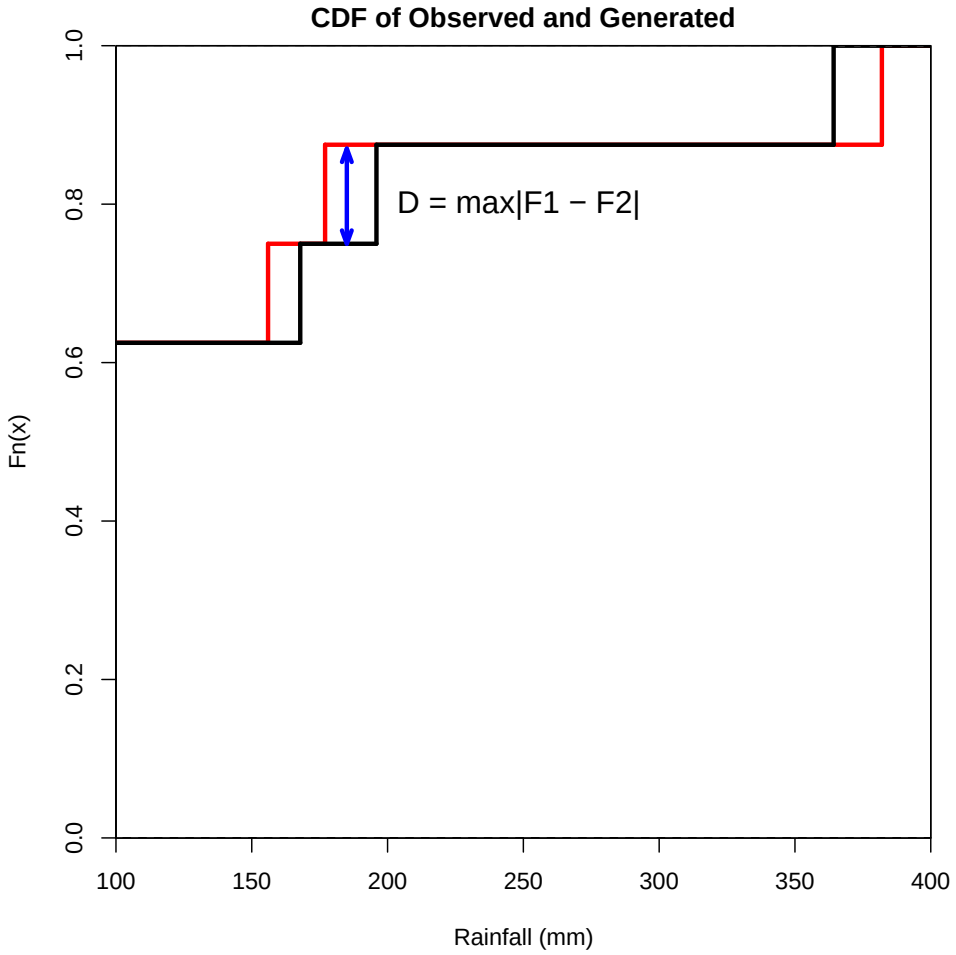



**Figure 3.4:** Synthetic Rainfall totals from cumulative distribution function

### 3.7.4 Anderson–Darling test

Theodore Wilbur Anderson and Donald A. Darling introduced the Anderson–Darling test in 1952 [4]. The Anderson–Darling test is considered better than the Kolmogorov–Smirnov test since it is able to detect the departure in the tails which is suitable for rainfall data [28] and many other applications for which the tail probabilities are important. It is applicable for grouped or ungrouped data. The test statistic of Anderson–Darling test has developed from testing the fit between the sample distribution with the hypothesised distribution [5] to testing the fit between two-sample

from independent distributions [74]. Pettitt [74] introduces a two-sample Anderson–Darling goodness of fit test statistic for two independent continuous distributions which is an extension from the work of Anderson and Darling [5].

Consider testing a null hypothesis that two sets of random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are from the same distributions. The cumulative distribution functions of $x$ and $y$ are represented by $F_n(x)$ and $G_m(x)$, respectively. A two-sample Anderson–Darling test statistic given by Pettitt [74] is expressed in the integral form given by

$$A^2_{nm} = \frac{nm}{N} \int_{-\infty}^{\infty} \frac{[F_n(x) - G_m(x)]^2}{H_N(x)[1 - H_N(x)]} \, dH_N(x) \tag{3.12}$$

where $N = n + m$ is the pooled sample size with cumulative distribution function of $H_N(x)$ defined by

$$H_N(x) = [nF_n(x) + mG_m(x)]/N.$$

The alternative form of equation (3.12) is

$$A^2_{nm} = \frac{1}{nm} \sum_{i=1}^{N-1} \frac{(M_i N - ni)^2}{i(N - i)} \tag{3.13}$$

where $M_i$ is the count of $X$ less than or equal to the $i$th smallest in the pooled sample. The hypothesis is

$$\begin{aligned} H_0 &: F_n(x) = G_m(x), \text{ the samples come from the same distribution} \\ H_1 &: F_n(x) \neq G_m(x), \text{ they do not come from the same distribution.} \end{aligned}$$

The null hypothesis is rejected if the calculated value of the test statistic is much larger than the critical value of $A^2_{nm}$ at $\alpha = 0.05$ significance level. The tabulated critical value of $A^2_{nm}$ can be found in [74] for various sizes of samples including for very large samples. The $R$ code to use the Anderson–Darling test is given in Appendix C.7.

## 3.8 Summary

In this chapter, the relevant mathematical and statistical tools needed throughout the study have been described. Some of the theories discussed are the gamma distribution, parameter estimation method, methodology for synthetic rainfall data generation, correlation measures and goodness of fit tests. In Chapter 4, we will be studying the sum of mixed–gamma variables to model rainfall data including zero totals.

# Chapter 4

# The sum of independent mixed–gamma variables with the same scale parameter

In this chapter I study the distribution for a sum of independent mixed–gamma random variables with the same scale parameter. The mixed–gamma random variables consisting of three parameters, denoted by $G(\alpha, \beta, p_0)$ where $p_0 > 0$ is the probability of a zero rainfall, $\alpha$ is the shape parameter and $\beta$ is the scale parameter. I discuss the fundamental properties and note that the regular two parameter gamma distribution $G(\alpha, \beta) = G(\alpha, \beta, 0)$ is obtained by setting $p_0 = 0$ and does not allow zero rainfall. Rosenberg et al. [87] describes a different type of mixed gamma distribution using Laguerre polynomials. This is different from that used in other modelling such as Piantadosi et al. [76].

I begin by showing that the sum of two independent variables from two parameter gamma distributions with the same scale parameter gives a regular gamma distribution with the same scale parameter and a shape parameter equal to the sum of

the individual shape parameters.

I use the fact that the formula for the sum of the regular independent gamma distributions with the same scale can be applied recursively to find corresponding formulae for the sum of mixed independent gamma distributions with the same scale parameter.

Finally these formulae are applied to a model for total summer rainfall at Pooraka in South Australia. The detail information for Pooraka meteorological station is given in Table 1.1 and Figure 1.3.

## 4.1 Mixed–gamma model

The gamma probability density function (PDF) in equation (3.1) is defined for positive rainfall totals. A mixed–gamma model is defined for zero and positive rainfall totals [86]. The probability of zero rainfall totals is calculated as

$$p_0 = P(X = 0) = \frac{n}{N}$$

where $n$ is the count for zero rainfall totals and $N$ is the total count of the rainfall data. The cumulative distribution function (CDF) for the mixed–gamma model is

$$\begin{aligned} F(x) &= P[X \leq x] \\ &= P[X = 0] + P[0 < X \leq x] \\ &= p_0 + (1 - p_0)F(x; \alpha, \beta) \end{aligned} \tag{4.1}$$

where we have written

$$F(x; \alpha, \beta) = \int_0^x f(\xi; \alpha, \beta)\ d\xi.$$

The mean is denoted by $E[X]$ where

$$\begin{aligned} E[X] &= p_0 \cdot 0 + (1-p_0)E[X] \\ &= (1-p_0)\alpha\beta \end{aligned} \tag{4.2}$$

where $X \sim G(\alpha, \beta)$ and the variance is $V[X] = E[X^2] - (E[X])^2$ where

$$\begin{aligned} E[X^2] &= p_0 \cdot 0^2 + (1-p_0)E[X^2] \\ &= (1-p_0)\alpha(\alpha+1)\beta^2. \end{aligned}$$

Therefore, the variance for a mixed–gamma random variable is

$$\begin{aligned} V[X] &= E[X^2] - (E[X])^2 \\ &= \alpha(\alpha+1)\beta^2(1-p_0) - [\alpha\beta(1-p_0)]^2 \\ &= \alpha^2\beta^2(1-p_0) + \alpha\beta^2(1-p_0) - \alpha^2\beta^2(1-2p_0+p_0^2) \\ &= \alpha^2\beta^2(1-p_0-1+2p_0-p_0^2) + \alpha\beta^2(1-p_0) \\ &= \alpha^2\beta^2(p_0-p_0^2) + \alpha\beta^2(1-p_0). \end{aligned} \tag{4.3}$$

If $X$ is a random variable with mixed–gamma distribution and parameters $p_0$ for the probability of zero, $\alpha$ for the shape and $\beta$ for the scale then we will write $X \sim G(\alpha, \beta, p_0)$. If $p_0 = 0$ then there is no probability of a zero rainfall and we write $G(\alpha, \beta, 0) = G(\alpha, \beta)$.

## 4.2 Sum of two independent mixed–gamma variables

We begin by showing the proof of the sum of two gamma variables with the same scale which is a standard result, refer to [44] of page 148. First, consider two

independent gamma variables $X$ and $Y$ both on $(0, \infty)$, with densities defined and denoted by $X \sim G(\alpha_1, \beta)$ and $Y \sim G(\alpha_2, \beta)$. I use different shape parameters $(\alpha_1, \alpha_2)$ but the same scale parameters $(\beta_1 = \beta_2 = \beta)$. To derive the density for the sum of two independent gamma variables $X$ and $Y$, let $S = X + Y$ denote the sum. Then the cumulative distribution function of the sum is

$$\begin{aligned} F(s) &= \iint_{0 \le x+y \le s} f(x; \alpha_1, \beta) f(y; \alpha_2, \beta)\, dxdy \\ &= \iint_{0 \le x+y \le s} \frac{x^{\alpha_1 - 1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha_1)\ \beta^{\alpha_1}} \frac{y^{\alpha_2 - 1} e^{-\frac{y}{\beta}}}{\Gamma(\alpha_2)\ \beta^{\alpha_2}}\, dxdy. \end{aligned}$$

Using the transformation $u = x + y$ and $v = y$,

$$F(s) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ \beta^{\alpha_1+\alpha_2}} \iint_{0 \le u \le s;\ 0 \le v \le s-u} (u - v)^{\alpha_1 - 1} v^{\alpha_2 - 1} e^{-\frac{u}{\beta}}\, du\, dv.$$

Now substitute $v = uw$ to give

$$\begin{aligned} F(s) &= \int_0^s \left[ \int_0^1 (1 - w)^{\alpha_1 - 1} w^{\alpha_2 - 1}\, dw \right] \frac{u^{\alpha_1+\alpha_2-1} e^{-\frac{u}{\beta}}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ \beta^{\alpha_1+\alpha_2}}\, du \\ &= \int_0^s \left[ \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} \right] \frac{u^{\alpha_1+\alpha_2-1} e^{-\frac{u}{\beta}}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ \beta^{\alpha_1+\alpha_2}}\, du \\ &= \int_0^s \frac{u^{\alpha_1+\alpha_2-1} e^{-\frac{u}{\beta}}}{\Gamma(\alpha_1 + \alpha_2)\ \beta^{\alpha_1+\alpha_2}}\, du \end{aligned}$$

and hence $S$ is a gamma distribution with density $f(s; \alpha_1 + \alpha_2, \beta)$. Thus, if $X \sim G(\alpha_1, \beta)$ and $Y \sim G(\alpha_2, \beta)$ are independent then $X + Y \sim G(\alpha_1 + \alpha_2, \beta)$. If the same shape parameter is used $(\alpha_1 = \alpha_2 = \alpha)$ then the density is $f(s; 2\alpha, \beta)$. Note that the Beta function, $B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1 - y)^{\beta-1}\, dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, has been used to simplify the above integrals.

Now consider two independent mixed–gamma variables $X$ and $Y$ both on $[0, \infty)$, with densities defined and denoted by $X \sim G(\alpha_1, \beta, p_{01})$ and $Y \sim G(\alpha_2, \beta, p_{02})$

where we assume $\beta_1 = \beta_2 = \beta$. To derive the density for the sum of two independent mixed–gamma variables $X, Y$, let $S = X + Y$ denote the sum where

$$\begin{aligned} F(x) &= p_{01} + (1 - p_{01})F(x; \alpha_1, \beta) \\ G(y) &= p_{02} + (1 - p_{02})F(y; \alpha_2, \beta). \end{aligned}$$

Hence the CDF for the sum of two independent mixed–gamma variables is

$$\begin{aligned} H(s) &= P[X + Y \leq s] \\ &= P[X = 0 \text{ and } Y = 0] + P[X = 0 \text{ and } 0 < Y \leq s] \\ &\quad + P[0 < X \leq s \text{ and } Y = 0] + P[0 < X,\ 0 < Y \text{ and } X + Y \leq s] \\ &= p_{01}p_{02} + p_{01}(1 - p_{02})F(s; \alpha_2, \beta) + p_{02}(1 - p_{01})F(s; \alpha_1, \beta) \\ &\quad + (1 - p_{01})(1 - p_{02})F(s; \alpha_1 + \alpha_2, \beta). \end{aligned} \tag{4.4}$$

The expected value (mean) of the sum is calculated using the result obtained in equation (4.2), and is given by

$$\begin{aligned} E[S] &= p_{01}p_{02} \cdot 0 + p_{01}(1 - p_{02})\alpha_2\beta + p_{02}(1 - p_{01})\alpha_1\beta \\ &\quad + (1 - p_{01})(1 - p_{02})(\alpha_1 + \alpha_2)\beta \\ &= \beta[\alpha_1(1 - p_{01}) + \alpha_2(1 - p_{02})]. \end{aligned}$$

Similarly, $E[S^2]$ is calculated using the result from equation (4.3);

$$\begin{aligned} E[S^2] &= p_{01}p_{02} \cdot 0^2 + p_{01}(1 - p_{02})\alpha_2(\alpha_2 + 1)\beta^2 + p_{02}(1 - p_{01})\alpha_1(\alpha_1 + 1)\beta^2 \\ &\quad + (1 - p_{01})(1 - p_{02})(\alpha_1 + \alpha_2)(\alpha_1 + \alpha_2 + 1)\beta^2 \\ &= \beta^2 \left[p_{01}(1 - p_{02})\alpha_2(\alpha_2 + 1) + p_{02}(1 - p_{01})\alpha_1(\alpha_1 + 1)\right. \\ &\quad \left. + (1 - p_{01})(1 - p_{02})(\alpha_1 + \alpha_2)(\alpha_1 + \alpha_2 + 1)\right] \\ &= \beta^2 \left[\alpha_1(\alpha_1 + 1)(1 - p_{01}) + \alpha_2(\alpha_2 + 1)(1 - p_{02})\right. \\ &\quad \left. + 2\alpha_1\alpha_2(1 - p_{01})(1 - p_{02})\right]. \end{aligned}$$

Hence the variance $V[S]$ of the sum of two independent mixed–gamma variables is

$$\begin{aligned} V[S] &= E[S^2] - (E[S])^2 \\ &= \beta^2 \left[\alpha_1(\alpha_1+1)(1-p_{01}) + \alpha_2(\alpha_2+1)(1-p_{02}) + 2\alpha_1\alpha_2(1-p_{01})(1-p_{02})\right] \\ &\quad -\beta^2[\alpha_1(1-p_{01}) + \alpha_2(1-p_{02})]^2 \\ &= \beta^2[\alpha_1(\alpha_1 p_{01}+1)(1-p_{01}) + \alpha_2(\alpha_2 p_{02}+1)(1-p_{02})]. \end{aligned}$$

If $\alpha_1 = \alpha_2 = \alpha$, then

$$E[S] = \alpha\beta[2 - p_{01} - p_{02}] \tag{4.5}$$

and

$$E[S^2] = \alpha^2\beta^2(4 - 3p_{01} - 3p_{02} + 2p_{01}p_{02}) + \alpha\beta^2(2 - p_{01} - p_{02}).$$

Hence

$$V[S] = \alpha^2\beta^2(p_{01} + p_{02} - p_{01}^2 - p_{02}^2) + \alpha\beta^2(2 - p_{01} - p_{02}). \tag{4.6}$$

## 4.3 Sum of three independent mixed–gamma variables

I use the same procedure as described in Section 4.1 and Section 4.2 to derive the PDFs and formulae for mean and variance of the sum of three independent mixed–gamma variables. Consider three independent mixed–gamma variables defined on $(0, \infty)$ denoted by $X \sim G(\alpha_1, \beta)$, $Y \sim G(\alpha_2, \beta)$ and $Z \sim G(\alpha_3, \beta)$ where the shape parameters are distinct and with the same scale parameters ($\beta_1 = \beta_2 = \beta_3 = \beta$). To derive the density for the sum of three independent mixed–gamma variables $X$, $Y$

and $Z$, let $S = X + Y + Z$ denote the sum where the CDF for each mixed–gamma variables are

$$\begin{aligned} F(x) &= p_{01} + (1 - p_{01})F(x; \alpha_1, \beta) \\ G(y) &= p_{02} + (1 - p_{02})F(y; \alpha_2, \beta) \\ H(z) &= p_{03} + (1 - p_{03})F(z; \alpha_3, \beta). \end{aligned}$$

Following from Section 4.2, if $X_1 \sim G(\alpha_1, \beta)$ and $X_2 \sim G(\alpha_2, \beta)$ then we have shown that $X_1 + X_2 \sim G(\alpha_1 + \alpha_2, \beta)$. Therefore, when we have $X_1 + X_2 \sim G(\alpha_1 + \alpha_2, \beta)$ and $X_3 \sim G(\alpha_3, \beta)$ it follows from the previous argument that $(X_1 + X_2) + X_3 \sim G((\alpha_1+\alpha_2)+\alpha_3, \beta)$. Hence the CDF for the sum of three independent mixed–gamma variables is

$$\begin{aligned} I(s) &= P[X = 0 \text{ and } Y = 0 \text{ and } Z = 0] \\ &\quad + P[X = 0 \text{ and } Y = 0 \text{ and } 0 < Z \leq s] \\ &\quad + P[X = 0 \text{ and } 0 < Y \leq s \text{ and } Z = 0] \\ &\quad + P[0 < X \leq s \text{ and } Y = 0 \text{ and } Z = 0] \\ &\quad + P[X = 0 \text{ and } 0 < Y + Z \leq s] \\ &\quad + P[Y = 0 \text{ and } 0 < X + Z \leq s] \\ &\quad + P[Z = 0 \text{ and } 0 < X + Y \leq s] \\ &\quad + P[0 < X,\ 0 < Y,\ 0 < Z \text{ and } X + Y + Z \leq s] \\ &= p_{01}p_{02}p_{03} + p_{01}p_{02}(1 - p_{03})F(s; \alpha_3, \beta) \\ &\quad + p_{01}p_{03}(1 - p_{02})F(s; \alpha_2, \beta) + p_{02}p_{03}(1 - p_{01})F(s; \alpha_1, \beta) \\ &\quad + p_{01}(1 - p_{02})(1 - p_{03})F(s; \alpha_2 + \alpha_3, \beta) \\ &\quad + p_{02}(1 - p_{01})(1 - p_{03})F(s; \alpha_1 + \alpha_3, \beta) \\ &\quad + p_{03}(1 - p_{01})(1 - p_{02})F(s; \alpha_1 + \alpha_2, \beta) \\ &\quad + (1 - p_{01})(1 - p_{02})(1 - p_{03})F(s; \alpha_1 + \alpha_2 + \alpha_3, \beta). \end{aligned} \tag{4.7}$$

The expected value (mean) of the sum is calculated using the result obtained in equation (4.2), and is given by

$$\begin{aligned}
E[S] &= p_{01}p_{02}p_{03} \cdot 0 \\
&\quad + p_{01}p_{02}(1-p_{03})\alpha_3\beta + p_{01}p_{03}(1-p_{02})\alpha_2\beta + p_{02}p_{03}(1-p_{01})\alpha_1\beta \\
&\quad + p_{01}(1-p_{02})(1-p_{03})(\alpha_2+\alpha_3)\beta + p_{02}(1-p_{01})(1-p_{03})(\alpha_1+\alpha_3)\beta \\
&\quad + p_{03}(1-p_{01})(1-p_{02})(\alpha_1+\alpha_2)\beta \\
&\quad + (1-p_{01})(1-p_{02})(1-p_{03})(\alpha_1+\alpha_2+\alpha_3)\beta \\
&= \beta\,[p_{01}p_{02}(1-p_{03})\alpha_3 + p_{01}p_{03}(1-p_{02})\alpha_2 + p_{02}p_{03}(1-p_{01})\alpha_1 \\
&\quad + p_{01}(1-p_{02})(1-p_{03})(\alpha_2+\alpha_3) + p_{02}(1-p_{01})(1-p_{03})(\alpha_1+\alpha_3) \\
&\quad + p_{03}(1-p_{01})(1-p_{02})(\alpha_1+\alpha_2) \\
&\quad + (1-p_{01})(1-p_{02})(1-p_{03})(\alpha_1+\alpha_2+\alpha_3)].
\end{aligned}$$

Similarly, $E[S^2]$ is calculated using the result from equation (4.3);

$$\begin{aligned}
E[S^2] &= p_{01}p_{02}p_{03}\cdot 0^2 \\
&\quad +p_{01}p_{02}(1-p_{03})\alpha_3(\alpha_3+1)\beta^2 + p_{01}p_{03}(1-p_{02})\alpha_2(\alpha_2+1)\beta^2 \\
&\quad +p_{02}p_{03}(1-p_{01})\alpha_1(\alpha_1+1)\beta^2 \\
&\quad +p_{01}(1-p_{02})(1-p_{03})(\alpha_2+\alpha_3)(\alpha_2+\alpha_3+1)\beta^2 \\
&\quad +p_{02}(1-p_{01})(1-p_{03})(\alpha_1+\alpha_3)(\alpha_1+\alpha_3+1)\beta^2 \\
&\quad +p_{03}(1-p_{01})(1-p_{02})(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1)\beta^2 \\
&\quad +(1-p_{01})(1-p_{02})(1-p_{03})(\alpha_1+\alpha_2+\alpha_3)(\alpha_1+\alpha_2+\alpha_3+1)\beta^2 \\
&= \beta^2[p_{01}p_{02}(1-p_{03})\alpha_3(\alpha_3+1) + p_{01}p_{03}(1-p_{02})\alpha_2(\alpha_2+1) \\
&\quad +p_{02}p_{03}(1-p_{01})\alpha_1(\alpha_1+1) \\
&\quad +p_{01}(1-p_{02})(1-p_{03})(\alpha_2+\alpha_3)(\alpha_2+\alpha_3+1) \\
&\quad +p_{02}(1-p_{01})(1-p_{03})(\alpha_1+\alpha_3)(\alpha_1+\alpha_3+1) \\
&\quad +p_{03}(1-p_{01})(1-p_{02})(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1) \\
&\quad +(1-p_{01})(1-p_{02})(1-p_{03})(\alpha_1+\alpha_2+\alpha_3)(\alpha_1+\alpha_2+\alpha_3+1)].
\end{aligned}$$

Therefore, the variance is

$$V[S] = E[S^2] - (E[S])^2.$$

If $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$, then

$$\begin{aligned}
E[S] &= \alpha\beta\,[p_{01}p_{02}(1-p_{03}) + p_{01}p_{03}(1-p_{02}) + p_{02}p_{03}(1-p_{01}) \\
&\quad +2(p_{01}(1-p_{02})(1-p_{03}) + p_{02}(1-p_{01})(1-p_{03}) + p_{03}(1-p_{01})(1-p_{02})) \\
&\quad +3(1-p_{01})(1-p_{02})(1-p_{03})] \\
&= \alpha\beta[3-(p_{01}+p_{02}+p_{03})] \qquad (4.8)
\end{aligned}$$

and

$$\begin{aligned}
E[S^2] = \alpha\beta^2[&(\alpha+1)(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03}-3p_{01}p_{02}p_{03})\\
&+2(2\alpha+1)(p_{01}+p_{02}+p_{03}-2(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03})+3p_{01}p_{02}p_{03})\\
&+3(3\alpha+1)(1-p_{01}-p_{02}-p_{03}+p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03}-p_{01}p_{02}p_{03})]\\
= \alpha\beta^2[&\alpha(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03}-3p_{01}p_{02}p_{03})\\
&+4\alpha(p_{01}+p_{02}+p_{03}-2(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03})+3p_{01}p_{02}p_{03})\\
&+9\alpha(1-p_{01}-p_{02}-p_{03}+p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03}-p_{01}p_{02}p_{03})\\
&+p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03}-3p_{01}p_{02}p_{03}\\
&+2(p_{01}+p_{02}+p_{03}-2(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03})+3p_{01}p_{02}p_{03})\\
&+3(1-p_{01}-p_{02}-p_{03}+p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03}-p_{01}p_{02}p_{03})]\\
= \alpha\beta^2\{&\alpha[9-5(p_{01}+p_{02}+p_{03})+2(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03})]\\
&+3-(p_{01}+p_{02}+p_{03})\}.
\end{aligned}$$

Hence the variance of the sum of three independent mixed–gamma variables is

$$\begin{aligned}
V[S] &= E[S^2]-(E[S])^2\\
&= \alpha^2\beta^2[2(p_{01}p_{02}+p_{01}p_{03}+p_{02}p_{03})-5(p_{01}+p_{02}+p_{03})+9]\\
&\quad +\alpha\beta^2(3-p_{01}-p_{02}-p_{03})\\
&\quad -\alpha^2\beta^2[9-6(p_{01}+p_{02}+p_{03})+(p_{01}+p_{02}+p_{03})^2]\\
&= \alpha^2\beta^2(p_{01}+p_{02}+p_{03}-p_{01}^2-p_{02}^2-p_{03}^2)\\
&\quad +\alpha\beta^2(3-p_{01}-p_{02}-p_{03}). \qquad (4.9)
\end{aligned}$$

## 4.4 General formulae

In this section it may be best to proceed by induction in the following way. In the case of $n = 2$, we have shown that when $X_1 \sim G(\alpha_1, \beta)$ and $X_2 \sim G(\alpha_2, \beta)$ then $X_1 + X_2 \sim G(\alpha_1 + \alpha_2, \beta)$, see Section 4.2. The known result for $n > 2$ can be derived as follows. Assume that if $X_i \sim G(\alpha_i, \beta)$ for all $i = 1, 2, \ldots$ then $X_1 + \cdots + X_n \sim G(\alpha_1 + \cdots + \alpha_n, \beta)$. It then follows from our previous arguments that $(X_1 + \cdots + X_n) + X_{n+1} \sim G((\alpha_1 + \cdots + \alpha_n) + \alpha_{n+1}, \beta)$. Hence the formula is true for all $n$. Now note that $\alpha_1 + \cdots + \alpha_n = n\overline{\alpha}_n$.

Now consider the case of mixed–gamma distributions $X_i \sim G(\alpha_i, \beta, p_{0i})$. Let $S = X_1 + \cdots + X_n$ where $X_i \sim G(\alpha_i, \beta, p_{0i})$. To describe the summation process we need to consider all binary numbers

$$r = (r_1, r_2, \ldots, r_n) \in B_n$$

where $r_i \in \{0, 1\}$. Now it can be seen, after a little thought, that

$$E[S] = \sum_{r \in B_n} \prod_{i|r_i=0} p_{0i} \prod_{j|r_j=1} (1 - p_{0j}) \left[ \sum_{j|r_j=1} \alpha_j \right] \beta$$

and

$$E[S^2] = \sum_{r \in B_n} \prod_{i|r_i=0} p_{0i} \prod_{j|r_j=1} (1 - p_{0j}) \left[ \sum_{j|r_j=1} \alpha_j \right] \left[ \sum_{j|r_j=1} \alpha_j + 1 \right] \beta^2.$$

## 4.5 Testing formula of mean and variance for the sum of independent mixed–gamma variables

We wish to test the formulae of mean and variance for the sum of two and three independent mixed–gamma variables given by formulae (4.5), (4.6), (4.8) and (4.9), respectively. The data used in the analysis is daily rainfall totals from Pooraka

station in South Australia for the period 1901–1990. Before conducting further analysis, we do the marginal analysis for each selected month of December, January and February. The marginal analysis has three parts: 1. Calculate $p_0$ and fit observed non-zero data to a gamma distribution (use MLE method, see Section 3.1.2), 2. fit generated data to gamma distribution and mixed–gamma distribution and 3. compare the mean and variance of observed and generated with the formula. The formulae of mean and variance for the mixed–gamma variables are (4.2) and (4.3), respectively.

In the first part of marginal analysis, each data is fitted to the gamma distribution using formula given in (3.1). The shape and scale parameters $(\alpha_i, \beta_i;\ i = 1, 2, 3)$ are estimated using the MLE method and the probability of zero rainfall totals $(p_{0i};\ i = 1, 2, 3)$ is calculated. The $R$ code is provided in Appendix C.2. Then, we check the values of mean and variance using formulae of $\alpha\beta$ and $\alpha\beta^2$, respectively. Table 4.1 shows that the value of mean of the observed data matches the estimated mean but not the variance.

For part two, using the estimated parameters from part one, the synthetic totals are generated as follows:

- For the gamma model, generate three sets of data using the parameters $(\alpha_i, \beta_i;\ i = 1, 2, 3)$. The procedures for generating synthetic totals for gamma model are described in Section 3.3 and the $R$ code is given in Appendix C.3.
- Similarly, for the mixed–gamma model, we generate three set of data using the parameters $(\alpha_i, \beta_i, p_{0i};\ i = 1, 2, 3)$. The procedure is described in Section 3.4 and the $R$ code is given in Appendix C.3.

Note that, if the sample size for the generated data is large (say 500 years) when compared to the observed sample (say 50 years) then it is not necessary to generate

a large number of samples. Therefore, for each model we will have three sets of data, generated data for December, January and February from gamma model and mixed–gamma model.

Then, the mean and variance are calculated for each data set and compared with the observed data and also with the formula. For the gamma model, the formulae for the mean and variance are given by $\alpha\beta$ and $\alpha\beta^2$, respectively. For the mixed–gamma model, the formulae for the mean and variance are (4.2) and (4.3), respectively. Table 4.2 shows a comparison of the means and variances between the observed and generated data from the two models. The values of the mean from both models are reasonably close to the observed mean, however the values of the variance are sometimes over-estimated or under-estimated of the observed values.

**Table 4.1:** Estimated parameters, mean (mm) and variance for December, January and February using gamma model

| | Estimated parameters | | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|---|
| | $p_0$ | $\alpha$ | $\beta$ | $\alpha\beta$ | observed | $\alpha\beta^2$ | observed |
| Dec | 0.0222 | 1.4954 | 17.9893 | 26.9011 | 26.9011 | 483.9318 | 401.8061 |
| Jan | 0.0667 | 1.2598 | 16.7341 | 21.0821 | 21.0821 | 352.7911 | 393.9547 |
| Feb | 0.0778 | 0.8829 | 24.3492 | 21.4988 | 21.4988 | 523.4788 | 556.5196 |

**Table 4.2:** Comparison of mean and variance of observed data with gamma and mixed–gamma models

| | | Observed | Gamma | | Mixed–gamma | |
|---|---|---|---|---|---|---|
| | | | generated | formula | generated | formula |
| Mean (mm) | Dec | 26.90 | 25.92 | 26.90 | 25.02 | 26.30 |
| | Jan | 21.08 | 21.66 | 21.08 | 19.44 | 19.68 |
| | Feb | 21.50 | 22.41 | 21.50 | 20.33 | 19.83 |
| Variance | Dec | 401.81 | 428.51 | 483.93 | 499.14 | 488.90 |
| | Jan | 393.95 | 362.18 | 352.79 | 364.34 | 356.93 |
| | Feb | 556.52 | 583.47 | 523.48 | 483.36 | 515.92 |

In the following section, I will show how to test the formula of mean and variance for the sum of two months and three months using the sum of mixed–gamma model.

### 4.5.1 Testing formula of mean and variance for the sum of two months

I test the application of the CDF of the sum of two independent mixed–gamma variables given by formula (4.4) on the months of December and January, which have similar scale values ($\beta$), using rainfall data from Pooraka station in South Australia for the period 1901–1990. Firstly, construct the sum of two rainfall totals by adding the corresponding rainfall data for December and January. Then calculate the mean, variance and $p_0$ for the sum and also check the mean and variance using the formulae of mean and variance from the gamma model. The parameters ($\alpha$ and $\beta$) are estimated by MLE method. The results are shown in Table 4.3.

**Table 4.3:** Estimated parameters, mean and variance for the sum of two observed data using the gamma model

| | Estimated parameters | | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|---|
| | $p_0$ | $\alpha$ | $\beta$ | $\alpha\beta$ | observed | $\alpha\beta^2$ | observed |
| Sum | 0.00 | 2.90 | 15.87 | 45.98 | 45.98 | 667.36 | 729.56 |

The generated sum of rainfall data is formed from the sum of two individual generated set from the mixed–gamma model. In each generated set, we use the average of shape parameters ($\overline{\alpha} = (\alpha_1 + \alpha_2)/2 = 1.3776$), the average of scale parameters ($\overline{\beta} = (\beta_1 + \beta_2)/2 = 17.3617$) and the respective probability of zero rainfall ($p_{0i}, i = 1, 2$), refer Table 4.1. Then we take the corresponding sum of the two generated data.

In the next step, we calculate and compare the mean and variance of the sum of the observed data with the sum of the generated data of the mixed–gamma. Also, we compare the mean and variance of the generated sum obtained from the formulae given by (4.5) and (4.6), respectively. Table 4.4 shows a comparison of the mean and variance between the observed sum and the sum of two independent mixed–gamma variables. The value of generated mean (45.28 mm) and using formula (45.71 mm) is close to the observed mean (45.98 mm). The observed variance is lower possibly due to correlation between the experimental months data, that have not been included in the mixed gamma models.

**Table 4.4:** Comparison of mean and variance for the sum of two observed data and the sum of two independent mixed–gamma variables

| | Observed | Mixed–gamma | |
|---|---|---|---|
| | | generated | formula |
| Mean (mm) | 45.98 | 45.28 | 45.71 |
| Variance | 667.36 | 847.98 | 841.61 |

### 4.5.2 Testing formula of mean and variance for the sum of three months

Again, I use the same rainfall data from Pooraka, South Australia, for the period of 1901–1990 to test the formulae of mean and variance for the sum of three months. The months selected for analysis are December, January and February. A similar procedure is followed as described in Section 4.5.1. Now the sum of rainfall data is constructed by adding three corresponding months. We use average $\alpha$ ($\overline{\alpha} = (\alpha_1 + \alpha_2 + \alpha_3)/3 = 1.2127$), average $\beta$ ($\overline{\beta} = (\beta_1 + \beta_2 + \beta_3)/3 = 19.6909$) and the corresponding probability of zero rainfall ($p_{0i}; i = 1, 2, 3$) to generate three sets of synthetic data, refer Table 4.1. Table 4.5 presents a comparison of the mean and variance for the observed data, generated data and the formula from sum of three independent mixed–gamma variables. We obtain a similar result as before that the value of generated mean (68.50 mm) and using formula (67.66 mm) are close to the observed mean (65.81 mm). Again, the variance of the data appears lower due to the same reason mentioned in the previous section.

### 4.5.3 Goodness of fit test

The means and variances of the observed sum and the sum of independent mixed–gamma variables using generated data and formulae are compared. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are used to assess the fit between the observed sum and the sum of independent mixed–gamma variables. Table 4.6 gives the P-values for the sum of two and three independent mixed–gamma variables. In both cases, P-values $> 0.05$, hence the observed sum is not significantly different from the generated sum of independent mixed–gamma model at 5% significance level.

**Table 4.5:** Comparison of mean and variance for the sum of three observed data and the sum of three independent mixed–gamma variables

| | Observed | Mixed-gamma | |
|---|---|---|---|
| | | generated | formula |
| Mean (mm) | 65.81 | 68.50 | 67.66 |
| Variance | 1237.53 | 1510.01 | 1421.03 |

**Table 4.6:** P-values of goodness of fit tests for the sum of two and three independent mixed–gamma variables

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| Sum of two | 0.4649 | 0.4230 |
| Sum of three | 0.9750 | 0.6205 |

## 4.6 Summary

The probability density function of the sum of two, three and $n$ mixed–gamma variables are derived for the simplest cases when the scale parameters are common. We also obtained the formulae of the mean and variance for each of the cases considered. Comparisons of the estimated values of the mean and variance from our simple model with both the estimated values from the fitted distribution and the observed values are presented. The general form of the formulae is also derived. The results demonstrate that the formulae derived will give a good estimate for the mean and the goodness of fit test confirmed a good fit between the cumulative distributions of the observed sum and the generated sum. This methodology can be applied to find the sum of more than three independent mixed–gamma variables. In the following Chapter, I will explore the Moschopoulos [68] method for application in rainfall modelling.

# Chapter 5

# Modelling the sum of $n$ independent gamma variables for both different shape and scale parameters

In this chapter I use a more complicated method first proposed by Moschopoulos [68] to model the sum of $n$ independent gamma random variables with both different shape and scale parameters. This formulation will be used whenever the correlation between the separate data sets is not significant. Firstly I state the result derived by Moschopoulos [68] and give details of the formulae used in the numerical computation. Secondly I derive a simple approximate representation of the proposed sum by a single gamma random variable. The approximation is obtained by selecting the parameters so that the first and second cumulants agree. I then explain the Moschopoulos [68] representation of the sum by a series of gamma random variables each having the same scale parameter. After this I consider an example in which

I model the winter rainfall at the Hume meteorological station using both the elementary approximate representation and the more sophisticated Moschopoulos [68] model. This is an example of a temporal model.

I present a new elementary derivation of the Moschopoulos model to develop an intuitive understanding of the important ideas. I apply the simple approach directly to a selected case study and then confirm and extend the results using the numerical algorithm suggested by Moschopoulos. Surprisingly, in this case study, there is no discernible difference in the graphs of the approximate and exact probability density functions for the sum. Finally I consider the sum of monthly rainfall at two different stations within the same region. Since the rainfall totals are strongly correlated it is no surprise to find that neither model produces an acceptable result.

## 5.1 Rainfall model for the sum of $n$ independent gamma variables

The basic theory of the sum of $n$ independent gamma variables was developed by Mathai [63] in modelling dam storage capacity. Mathai [63] discusses three forms of the sum of $n$ independent gamma variables which depend on different values of shape parameter $\alpha_i$ and scale parameter $\beta_i > 0$, for $i = 1, \ldots, n$. In the first case he uses moment generating functions to show that when the values of $\alpha_i$ are integers, a finite sum of gamma densities is formed. In the second case when $\alpha_1 = \cdots = \alpha_n$ the density is expressed in zonal polynomials. Zonal polynomials are useful in the study of noncentral multivariate distributions which are defined as eigenfunctions of an expectation. The reader is referred to Takemura [96] for further details of the concept and usage of zonal polynomials. In the third case Mathai [63] shows that in general, when the values for shape and scale are all different, the PDF for the sum

will involve the hypergeometric function.

Moschopoulos [68] extended the theory by introducing a single recursive form of gamma-series representation. Before I conduct any analysis, the correlations between the variables need to be checked to show that the correlations are not significant. In this case, the variables can be modelled independently. I choose Spearman's correlation as described in Section 3.5 because, in general, rainfall totals have a skewed distribution. See [81] and [29] for more information. I will use the Moschopoulos [68] method to model seasonal rainfall at a meteorological station in the Murray-Darling Basin. I first represent the monthly rainfalls in the wet season by independent gamma distributions and then use the Moschopoulos [68] method to model the total of the monthly rainfalls during the wet season. I choose a gamma distribution to fit the individual monthly totals. The method has been described earlier in Section 3.1.

### 5.1.1 Probability density function of the sum of $n$ independent gamma variables

I use the single gamma series suggested by Moschopoulos [68] for easy computation of the sum of $n$ independent gamma variables. Following [68] I consider a set of $n$ independent gamma variables, $\{X_i\}_{i=1}^{n}$ with parameters $\alpha_i$ and $\beta_i$ written as $X_i \sim G(\alpha_i, \beta_i)$ where the PDF of $X_i$ is given by

$$f_i(x_i; \alpha_i, \beta_i) = \frac{x_i^{\alpha_i - 1} e^{-\frac{x_i}{\beta_i}}}{\Gamma(\alpha_i)\beta_i^{\alpha_i}} \qquad \text{for } x_i > 0 \text{ and } \alpha_i, \beta_i > 0. \tag{5.1}$$

If we write $Y = \sum_{i=1}^{n} X_i$ then according to [68] the PDF of the sum of $n$ independent gamma variables is

$$f_Y(y) = C \sum_{k=0}^{\infty} \frac{\delta_k \; y^{p+k-1} \exp\left(-y/\beta_1\right)}{\beta_1^{p+k} \; \Gamma(p+k)} \tag{5.2}$$

where $\beta_1 = \min_i\{\beta_i\}$, $C = \prod_{i=2}^{n} \left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i}$ and $p = \sum_{i=1}^{n} \alpha_i > 0$. The coefficients of $\delta_k$ are defined using

$$\begin{cases} \delta_0 & = 1 \\ \delta_{k+1} & = \frac{1}{k+1} \sum_{i=1}^{k+1} i \; \gamma_i \; \delta_{k+1-i}, \quad k = 0, 1, 2, \ldots \end{cases}$$

where

$$\gamma_i = \sum_{j=1}^{n} \frac{\alpha_j (1 - \beta_1/\beta_j)^i}{i} \quad \text{for } i = 1, 2, \ldots$$

For convenience, we rewrite the recursive relationship for the coefficients $\delta_1, \delta_2, \ldots$ in the form $\delta_0 = 1$ and

$$\delta_k = \frac{1}{k} \sum_{i=1}^{k} i \; \gamma_i \; \delta_{k-i}, \quad k = 1, 2, \ldots$$

therefore,

$$\begin{aligned} \delta_k &= \frac{1}{k} \sum_{i=1}^{k} i \left[ \sum_{j=1}^{n} \frac{\alpha_j (1 - \beta_1/\beta_j)^i}{i} \right] \delta_{k-i}, \quad k = 1, 2, \ldots \\ &= \frac{1}{k} \sum_{i=1}^{k} \left[ \sum_{j=1}^{n} \alpha_j (1 - \beta_1/\beta_j)^i \right] \delta_{k-i}. \end{aligned}$$

In the next section I present a new approximate representation for the sum by a single gamma random variable and also present an elementary derivation of the the exact Moschopoulos [68] representation.

## 5.2 Representation of the sum of independent gamma random variables

In this section I propose a simple approximate representation for the sum of independent gamma random variables and also present an elementary derivation of the Moschopoulos [68] representation.

### 5.2.1 An approximate representation of the sum of independent gamma random variables

Let $X_i \sim G(\alpha_i, \beta_i)$ for $i = 1, 2, \ldots, n$ be independent gamma random variables and let $Y = X_1 + \cdots + X_n$. The moment generating function for $Y$ is given by

$$m_Y(t) = \prod_{i=1}^{n} (1 - \beta_i t)^{-\alpha_i}$$

where we assume for convenience that $0 < \beta_1 \leq \cdots \leq \beta_n$. It follows that the cumulant generating function $\kappa_Y(t) = \log_e m_Y(t)$ is given by

$$\kappa_Y(t) = (-1) \sum_{i=1}^{n} \alpha_i \log_e(1 - \beta_i t)$$

and hence

$$\kappa_Y(0) = 0, \quad \kappa'_Y(0) = \sum_{i=1}^{n} \alpha_i \beta_i, \quad \text{and} \quad \kappa''_Y(0) = \sum_{i=1}^{n} \alpha_i \beta_i^2.$$

If we wished to approximate $Y$ by a gamma random variable $Z$ then we could choose $Z$ so that the first two cumulants are the same. That is $\kappa'_Z(0) = \kappa'_Y(0)$ and $\kappa''_Z(0) = \kappa''_Y(0)$. This gives us two equations in two unknowns

$$\alpha\beta = \alpha_1\beta_1 + \cdots + \alpha_n\beta_n \tag{5.3}$$

and

$$\alpha\beta^2 = \alpha_1\beta_1^2 + \cdots + \alpha_n\beta_n^2. \tag{5.4}$$

It follows that

$$\alpha = \frac{(\alpha_1\beta_1 + \cdots + \alpha_n\beta_n)^2}{\alpha_1\beta_1^2 + \cdots + \alpha_n\beta_n^2} \tag{5.5}$$

and

$$\beta = \frac{\alpha_1\beta_1^2 + \cdots + \alpha_n\beta_n^2}{\alpha_1\beta_1 + \cdots + \alpha_n\beta_n}. \tag{5.6}$$

Thus we approximate $Y$ by $Z \sim G(\alpha, \beta)$. Note that mean and variance formulae of gamma random variables are well known and independent of the distribution.

### 5.2.2 An exact representation of the sum of independent gamma random variables – the Moschopoulos method

The idea is to write the moment generating function

$$m_Y(t) = \prod_{i=1}^{n}(1 - \beta_i t)^{-\alpha_i}$$

as a function of

$$\tau = (1 - \beta_1 t)^{-1}$$

and then to expand as a series in $\tau$. We have

$$t = \frac{1}{\beta_1}\left(1 - \frac{1}{\tau}\right)$$

from which it follows by elementary algebra that

$$(1 - \beta_i t)^{-1} = \tau \cdot \frac{\beta_1}{\beta_i} \cdot \left[1 - \left(1 - \frac{\beta_i}{\beta_1}\right)\tau\right]^{-1}.$$

Hence we see that

$$\begin{aligned} m_Y(t) &= \tau^{\alpha_1} \prod_{i=2}^{n} \left\{ \tau^{\alpha_i} \left( \frac{\beta_1}{\beta_i} \right)^{\alpha_i} \left[ 1 - \left( 1 - \frac{\beta_1}{\beta_i} \right) \tau \right]^{-\alpha_i} \right\} \\ &= \tau^p \prod_{i=2}^{n} \left( \frac{\beta_1}{\beta_i} \right)^{\alpha_i} \prod_{j=2}^{n} \Bigg[ 1 + \alpha_j \left( 1 - \frac{\beta_1}{\beta_j} \right) \tau \\ &\qquad + \frac{\alpha_j(\alpha_j+1)}{2!} \left( 1 - \frac{\beta_1}{\beta_j} \right)^2 \tau^2 + \cdots \Bigg] \end{aligned}$$

where we have written $p = \alpha_1 + \cdots + \alpha_n$ for convenience and where the convergence of the binomial series expansion is justified for $|\beta_i t| < 1 \Rightarrow 0 < \tau < 1$. Furthermore, since $\tau = (1 - \beta_1 t)^{-1}$, the entire right hand side is an infinite sum of terms corresponding to the moment generating functions of gamma distributions all with scale parameter $\beta_1$ and with shape parameters $p, p+1, p+2, \ldots$ respectively. Hence we can invert the sum term by term to obtain an expression for the PDF. We have

$$\begin{aligned} f_Y(y) &= \prod_{i=2}^{n} \left( \frac{\beta_1}{\beta_i} \right)^{\alpha_i} \Bigg[ f_{p,\beta_1}(y) \; + \; \sum_{j=2}^{n} \alpha_j \left( 1 - \frac{\beta_1}{\beta_j} \right) f_{p+1,\beta_1}(y) \\ &\quad + \Bigg\{ \sum_{j=2}^{n} \sum_{k=2}^{n} \alpha_j \alpha_k \left( 1 - \frac{\beta_1}{\beta_j} \right) \left( 1 - \frac{\beta_1}{\beta_k} \right) \\ &\qquad + \sum_{j=2}^{n} \alpha_j(\alpha_j+1) \left( 1 - \frac{\beta_1}{\beta_j} \right)^2 \Bigg\} f_{p+2,\beta_1}(y) \; + \; \cdots \Bigg] \end{aligned}$$

where we have used the notation $f_{\alpha,\beta}(\cdot)$ to denote the PDF of a gamma distribution with parameters $\alpha$ and $\beta$. We refer to the paper by Moschopoulos [68] for more details about efficient numerical calculations. Note that integration term by term

yields

$$\begin{aligned}
&\int_0^\infty f_Y(y)dy \\
&= \prod_{i=2}^{n}\left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i}\left[1 + \sum_{j=2}^{n}\alpha_j\left(1-\frac{\beta_1}{\beta_j}\right)\right. \\
&\qquad + \left\{\sum_{j=2}^{n}\sum_{k=2}^{n}\alpha_j\alpha_k\left(1-\frac{\beta_1}{\beta_j}\right)\left(1-\frac{\beta_1}{\beta_k}\right)\right. \\
&\qquad \left.\left. + \sum_{j=2}^{n}\alpha_j(\alpha_j+1)\left(1-\frac{\beta_1}{\beta_j}\right)^2\right\} + \cdots\right] \\
&= \prod_{i=2}^{n}\left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i}\prod_{j=2}^{n}\left[1+\alpha_j\left(1-\frac{\beta_1}{\beta_j}\right) + \frac{\alpha_j(\alpha_j+1)}{2!}\left(1-\frac{\beta_1}{\beta_j}\right)^2+\cdots\right] \\
&= \prod_{i=2}^{n}\left\{\left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i}\left[1-\left(1-\frac{\beta_1}{\beta_i}\right)\right]^{-\alpha_i}\right\} \\
&= 1.
\end{aligned}$$

Thus we have verified retrospectively that $f_Y(\cdot)$ is indeed a valid PDF.

## 5.3 A case study to compare the approximate and exact representations

We choose monthly rainfall totals from the Hume meteorological station and consider the total rainfall for the months June–September inclusive from 1922 to 2009. Firstly, we calculate the pairwise Spearman's correlation between the months using the hypothesis test for correlation as described in Section 3.5, refer to Table 5.1. We find that the P-value $> 0.05$ for every pair of months and conclude that the monthly correlations are not significant. Therefore we assume there is no correlation between any two winter months. In this case we have $n = 4$. The parameters $(\alpha, \beta)$ for the

gamma distribution and the probability of zero rainfall totals ($p_0$), representing the monthly rainfalls for June, July, August and September are given in Table 5.2[1]. The analysis for marginals of gamma for observed and generated are done before further analyses are conducted. The results of the marginals analysis are presented in Section 6.6.7, also refer to Tables 6.20 and 8.4 and Figure 6.10.

**Table 5.1:** Spearman's rank correlation coefficients between June–September of Hume

| | Jul | Aug | Sep |
|---|---|---|---|
| Jun | 0.12 | 0.13 | -0.01 |
| Jul | | 0.13 | 0.24 |
| Aug | | | 0.18 |

**Table 5.2:** Values of $\alpha$, $\beta$ and $p_0$ for different months (July, September, August, June)

| i | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\alpha$ | 3.91 | 3.12 | 3.23 | 2.72 |
| $\beta$ | 19.68 | 19.87 | 23.51 | 24.97 |
| $p_0$ | 0.00 | 0.00 | 0.00 | 0.00 |

Let $\beta_1$ be the minimum value of $\beta$, thus we have

$$\frac{\beta_1}{\beta_2} = 0.9904, \quad \frac{\beta_1}{\beta_3} = 0.8371, \quad \frac{\beta_1}{\beta_4} = 0.7881$$

and

$$p = \alpha_1 + \cdots + \alpha_4 = 12.98.$$

[1]Estimation of monthly parameters has been discussed in detail previously and suffice it to say here that the values given in Table 5.2 were found to provide a good fit.

The Maclaurin series in the moment generating function take the general form

$$\begin{aligned} s_j(\tau) &= 1 + \alpha_j \left(1 - \frac{\beta_1}{\beta_j}\right) \tau + \frac{\alpha_j(\alpha_j+1)}{2!} \left(1 - \frac{\beta_1}{\beta_j}\right)^2 \tau^2 \\ &\quad + \frac{\alpha_j(\alpha_j+1)(\alpha_j+2)}{3!} \left(1 - \frac{\beta_1}{\beta_j}\right)^3 \tau^3 + \cdots \end{aligned}$$

and so in this case we have

$$\begin{aligned} s_2(\tau) &= 1 + \ 0.029834\tau + \ 0.000588\tau^2 + \ 0.000010\tau^3 + \cdots \\ s_3(\tau) &= 1 + \ 0.526197\tau + \ 0.181303\tau^2 + \ 0.051491\tau^3 + \ 0.013065\tau^4 \\ &\quad + 0.003078\tau^5 + \ 0.000688\tau^6 + \cdots \\ s_4(\tau) &= 1 + \ 0.576243\tau + \ 0.227068\tau^2 + \ 0.075686\tau^3 + \ 0.022929\tau^4 \\ &\quad + 0.006529\tau^5 + \ 0.001780\tau^6 + \cdots \end{aligned}$$

where all coefficients are calculated to six decimal place accuracy and the non-zero coefficients are shown up to the coefficient of the sixth power. We also have

$$C = \left(\frac{\beta_1}{\beta_2}\right)^{\alpha_2} \left(\frac{\beta_1}{\beta_3}\right)^{\alpha_3} \left(\frac{\beta_1}{\beta_4}\right)^{\alpha_4} = 0.285960.$$

Hence the distribution for the random variable $Y = X_1 + \cdots + X_4$ is given by

$$\begin{aligned} f_Y &= 0.285960 \left\{ f_{12.98,19.68} + 1.132274 f_{13.98,19.68} + 0.745067 f_{14.98,19.68} \right. \\ &\quad + 0.373021 f_{15.98,19.68} + 0.157564 f_{16.98,19.68} + 0.059203 f_{17.98,19.68} \\ &\quad \left. + 0.020416 f_{18.98,19.68} + \cdots \right\}. \end{aligned}$$

Note that if we truncate the expression at this stage then we obtain

$$\begin{aligned} \int_0^\infty f_Y(y)dy &= 0.285960 \{1 + 1.132274 + 0.745067 + 0.373021 \\ &\qquad + 0.157564 + 0.059203 + 0.020416 + \cdots \} \\ &\approx 0.997300 \end{aligned}$$

where we have used $f_{\alpha,\beta}$ to denote the PDF for the standard two parameter gamma distribution $G(\alpha, \beta)$. The truncated Moschopoulos PDF and CDF are sketched in Figure 5.1. I also use formulae (5.5) and (5.6) to calculate $\alpha = 12.8425$ and $\beta = 22.0206$ for the approximate representation of the sum by a single gamma random variable $Z \sim G(\alpha, \beta)$. The graphs of the corresponding PDF and CDF are also shown in Figure 5.1. It can be seen that there is no discernible difference at this scale between the two representations. Furthermore, based on the goodness of fit tests of Kolmogorov–Smirnov and Anderson–Darling tests, there is no significant difference between the observed and the theoretical distributions at 5% significance level. See Table 5.3 for details.

**Table 5.3:** Goodness of fit tests for Moschopoulos PDF and approximate gamma PDF

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| P-value | 0.2244 | 0.1606 |

## 5.4 Application of the Moschopoulos numerical algorithm to the previous case study

We use the data of the previous case study to demonstrate the recommended Moschopoulos numerical algorithm. We show explicitly that the algorithm reproduces the results that we obtained in the previous section by elementary methods. In the previous section, we have checked that there are no correlation between any two winter months and hence we can apply the Moschopoulos method. Each month is modelled independently using equation (5.1) and the value of the shape and scale parameters are estimated using the MLE method which is described in Section 3.1.2. The estimated parameters for each month are given in Table 5.4. The observed mean

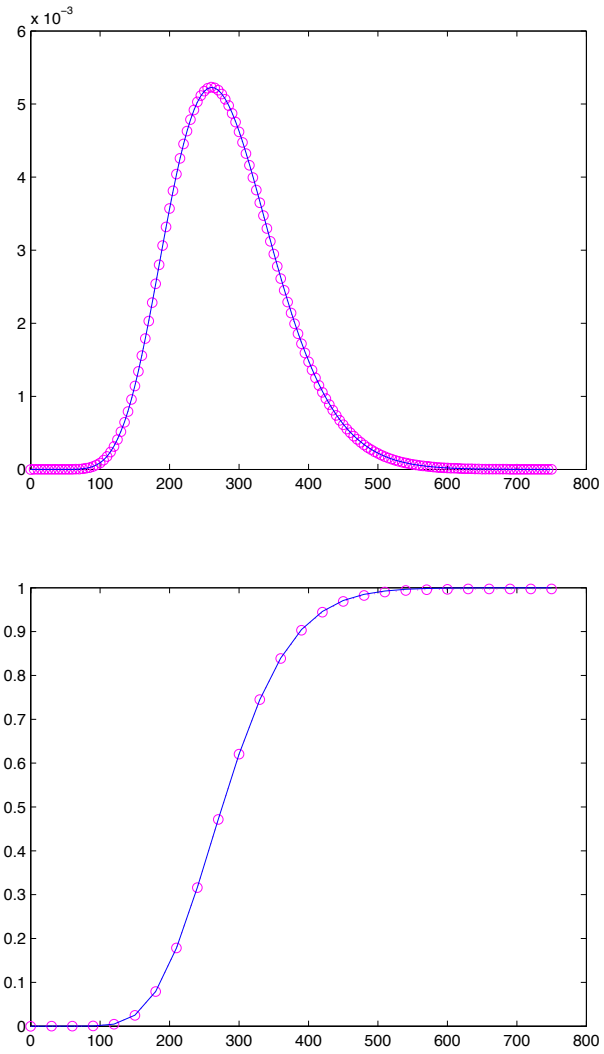



**Figure 5.1:** Plot of Moschopoulos (circles) and approximate gamma (line), PDFs (above) and CDFs (below)

and variance and the theoretical mean and variance from the fitted distributions are also shown. Note that the method of maximum likelihood gives parameters estimates for the model which match the observed mean exactly but do not match the observed variance. If the fit is good then the variances will match approximately.

**Table 5.4:** Estimated shape and scale parameters for the winter season

| | Estimated parameters | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | Observed | $\alpha\beta$ | Observed | $\alpha\beta^2$ |
| Jun | 2.72 | 24.97 | 68.02 | 68.02 | 1582.69 | 1698.39 |
| Jul | 3.91 | 19.68 | 77.03 | 77.03 | 1545.23 | 1515.94 |
| Aug | 3.23 | 23.51 | 75.92 | 75.92 | 1398.71 | 1784.82 |
| Sept | 3.12 | 19.87 | 62.00 | 62.00 | 1072.84 | 1231.95 |

Now we use the Moschopoulos method with $n = 4$ to model the total winter rainfall. Using the PDF of the sum of independent gamma variables given in equation (5.2), we generate synthetic rainfall totals. The detailed procedure on how to generate synthetic rainfall totals whenever the CDF is not available is given in Section 3.4. The CDF is easily obtained directly by numerical integration or by using one of the many available packages for evaluation of the incomplete gamma function. The incomplete gamma function is defined as

$$F(x, \alpha) = 1/\Gamma(\alpha) \int_0^x \exp(-\xi)\ \xi^{\alpha-1}\ d\xi$$

and we note that the substitution $t = x/\beta$ gives the CDF of the gamma density $f_{\alpha,\beta}$. In the numerical procedure, we use the values of $\alpha$ and $\beta$ in Table 5.4 to obtain $p = \sum_{i=1}^{4} \alpha_i = 12.98$ and $\beta_1 = 19.68$ where $\beta_1$ is the minimum from the values of $\beta_i$. Therefore, we define $\beta_2 = 19.87$, $\beta_3 = 23.51$ and $\beta_4 = 24.97$. For convenience, the formula for the PDF is divided into two terms: the product term ($C$) and the summation term ($S$). Consider the product term and substitute the parameters values given above. We obtain

$$\frac{\beta_1}{\beta_2} = 0.9904, \quad \frac{\beta_1}{\beta_3} = 0.8371 \quad \text{and} \quad \frac{\beta_1}{\beta_4} = 0.7881$$

and

$$\begin{aligned} C = \prod_{i=2}^{4} \left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i} &= \left(\frac{\beta_1}{\beta_2}\right)^{\alpha_2} \left(\frac{\beta_1}{\beta_3}\right)^{\alpha_3} \left(\frac{\beta_1}{\beta_4}\right)^{\alpha_4} \\ &= 0.2859603571. \end{aligned}$$

The summation term is

$$S = \sum_{k=0}^{\infty} \frac{\delta_k \; y^{p+k-1} \exp\left(-y/\beta_1\right)}{\beta_1^{p+k} \; \Gamma(p+k)}$$

where the values of $\delta_k$ for $k = 1, 2, \ldots$ are calculated using the recursive formula

$$\delta_k = \frac{1}{k} \sum_{i=1}^{k} i \; \gamma_i \; \delta_{k-i}$$

where $\gamma_i = \sum_{j=1}^{n} \frac{\alpha_j (1-\beta_1/\beta_j)^i}{i}$ for $i = 1, 2, \ldots$. Given that $\delta_0 = 1$ we calculate for

$$\begin{aligned} k = 1, \quad \delta_1 &= \gamma_1 \; \delta_0 \\ k = 2, \quad \delta_2 &= \frac{1}{2}(\gamma_1 \; \delta_1 + 2\gamma_2 \; \delta_0) \\ k = 3, \quad \delta_3 &= \frac{1}{3}(\gamma_1 \; \delta_2 + 2\gamma_2 \; \delta_1 + 3\gamma_3 \; \delta_0) \\ k = 4, \quad \delta_4 &= \frac{1}{4}(\gamma_1 \; \delta_3 + 2\gamma_2 \; \delta_2 + 3\gamma_3 \; \delta_1 + 4\gamma_4 \; \delta_0) \\ k = 5, \quad \delta_4 &= \frac{1}{5}(\gamma_1 \; \delta_4 + 2\gamma_2 \; \delta_3 + 3\gamma_3 \; \delta_2 + 4\gamma_4 \; \delta_1 + 5\gamma_5 \; \delta_0) \end{aligned}$$

and so on. The values of $\gamma_i$ and $\delta_k$ for $i, k = 1, 2, \ldots, 20$ are shown in Table 5.5. Care must be used when evaluating the terms in the summation. Although both Kotz and Adams [58] (with a similar series) and Moschopoulos [68] prove that the sum converges uniformly and establish bounds on the size of the error we note that the individual terms $f_{\alpha,\beta}$ involve products of very large and very small terms. This can lead to unwanted numerical errors when evaluating the individual terms. Therefore,

instead of evaluating directly, we write the PDF as

$$\begin{aligned} f_Y(y) &= \prod_{i=1}^{2} \left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i} \sum_{k=0}^{\infty} \exp\left(\ln\left[\frac{\delta_k \; y^{p+k-1} \exp\left(-y/\beta_1\right)}{\beta_1^{p+k} \; \Gamma(p+k)}\right]\right) \\ &= \prod_{i=1}^{2} \left(\frac{\beta_1}{\beta_i}\right)^{\alpha_i} \sum_{k=0}^{\infty} \exp\Bigg(\ln\delta_k + (p+k-1)\ln y - y/\beta_1 \\ &\qquad -(p+k)\ln\beta_1 - \ln\Gamma(p+k)\Bigg). \end{aligned} \tag{5.7}$$

**Table 5.5:** Values of $\gamma$ and $\delta$

| $k$ | $\gamma$ | $\delta$ |
|---|---|---|
| 0 | N/A | 1.0000000000 |
| 1 | 1.1322747755 | 1.1322747755 |
| 2 | 0.1040436959 | 0.7450667795 |
| 3 | 0.0132769371 | 0.3730210789 |
| 4 | 0.0019385648 | 0.1575637609 |
| 5 | 0.0003062833 | 0.0592028877 |
| 6 | 0.0000510495 | 0.0204160705 |
| 7 | 0.0000088479 | 0.0065934465 |
| 8 | 0.0000015800 | 0.0020220049 |
| 9 | 0.0000002888 | 0.0005947055 |
| 10 | 0.0000000538 | 0.0001689985 |
| 11 | 0.0000000102 | 0.0000466636 |
| 12 | 0.0000000019 | 0.0000125749 |
| 13 | 0.0000000004 | 0.0000033188 |
| 14 | 0.0000000001 | 0.0000008603 |
| 15 | 0.0000000000 | 0.0000000220 |
| 16 | 0.0000000000 | 0.0000000055 |
| 17 | 0.0000000000 | 0.0000000014 |
| 18 | 0.0000000000 | 0.0000000003 |
| 19 | 0.0000000000 | 0.0000000001 |
| 20 | 0.0000000000 | 0.0000000000 |
| 21 | 0.0000000000 | 0.0000000000 |

The PDF is given by

$$\begin{aligned} f_Y(y) &= 0.2859604\{f_{p,\beta_1} + 1.1322747755 f_{p+1,\beta_1} + 0.7450667795 f_{p+2,\beta_1} \\ &\quad +0.3730210789 f_{p+3,\beta_1} + 0.1575637609 f_{p+4,\beta_1} + 0.0592028877 f_{p+5,\beta_1} \\ &\quad +0.0204160705 f_{p+6,\beta_1} + 0.0065934465 f_{p+7,\beta_1} + 0.0020220049 f_{p+8,\beta_1} \\ &\quad +0.0005947055 f_{p+9,\beta_1} + 0.0001689985 f_{p+10,\beta_1} + 0.0000466636 f_{p+11,\beta_1} \\ &\quad +0.0000125749 f_{p+12,\beta_1} + 0.0000033188 f_{p+13,\beta_1} + 0.0000008603 f_{p+14,\beta_1} \\ &\quad +0.0000000220 f_{p+15,\beta_1} + 0.0000000055 f_{p+16,\beta_1} + 0.0000000014 f_{p+17,\beta_1} \\ &\quad +0.0000000003 f_{p+18,\beta_1} + 0.0000000001 f_{p+19,\beta_1} + 0.0000000000 f_{p+20,\beta_1} \\ &\quad +0.0000000000 f_{p+21,\beta_1}\} \end{aligned}$$

and the total probability $F_Y(\infty) \approx C(\sigma_0 + \sigma_1 + \cdots + \sigma_{21})$ is given by

$$\begin{aligned} \int_0^\infty f_Y(y) &\approx 0.2859604\{1 + 1.1322750000 + 0.7450667795 \\ &\quad +0.3730210789 + 0.1575637609 + 0.0592028877 \\ &\quad +0.0204160705 + 0.0065934465 + 0.0020220049 \\ &\quad +0.0005947055 + 0.0001689985 + 0.0000466636 \\ &\quad +0.0000125749 + 0.0000033188 + 0.0000008603 \\ &\quad +0.0000000220 + 0.0000000055 + 0.0000000014 \\ &\quad +0.0000000003 + 0.0000000001 + 0.0000000000 \\ &\quad +0.0000000000\} \\ &\approx 0.999999999995746. \end{aligned}$$

All numbers were computed in MATLAB to machine accuracy. Figure 5.2 shows the plot of the distribution of $\delta_k$ values. It is observed that the values of $\delta_k$ decrease quickly to zero as $k$ increases. Since the CDF for $Y$ is not directly available it must be calculated from the PDF by numerical integration. Thus we proceed as suggested in

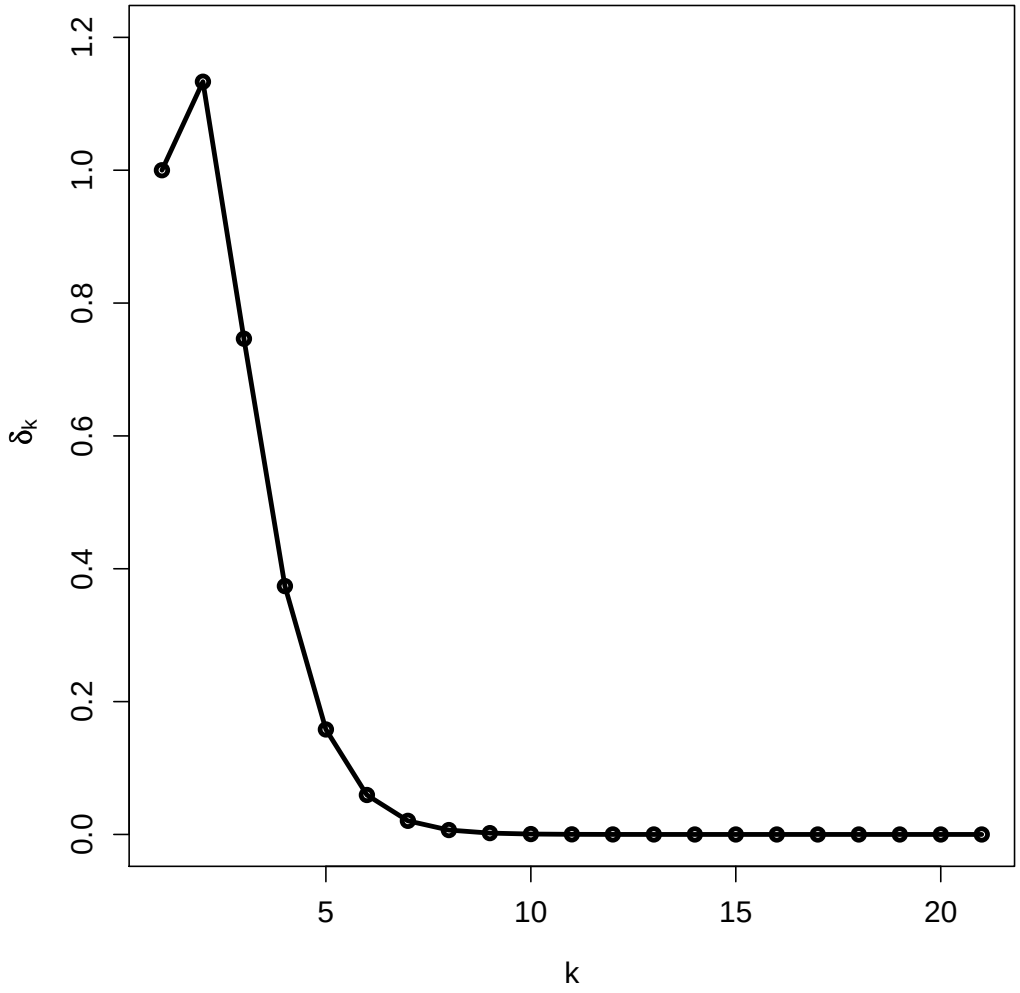



**Figure 5.2:** Plot of $\delta_k$ values

Section 3.3. In principle one could select a sufficiently fine equally spaced subdivision $0 < r_1 < r_2 < \cdots < r_m$ where $r_m$ exceeds the maximum observed total and use numerical integration to calculate $F(y) = P[Y < r_j]$ at each subdivision point $r_j$. Now one can construct a look-up table for the tabulated values of $F(y)$. The look-up table of the CDF associated with the rainfall totals is used to match the uniformly distributed random numbers that are generated on the interval $[0, 1]$ with

the rainfall associated with the corresponding cumulative probability. It would be relatively easy to use Simpson's rule or some adaptive numerical integration scheme if one wished rather than an elementary rule such as

$$F(r_j) = \sum_{i=1}^{j} f_Y(r_i)\Delta r_i$$

with $r_i = i$ and $\Delta r_i = 1$. In practice the generation of synthetic data does not require painstaking accuracy in the numerical integration. One should be sure, however, that there is no bias introduced by the numerical calculations.

The means and variances for the sum of monthly rainfalls in the winter season for the observed and fitted rainfall totals are given in Table 5.6. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are used to assess the fit between the observed and fitted rainfall totals. Refer to Sections 3.7.3 and 3.7.4 for further details about the goodness of fit test. I also generate 5000 random numbers from the PDF of the sum of four independent gamma variables using $\{\alpha_1 = 2.72, \alpha_2 = 3.91, \alpha_3 = 3.23, \alpha_4 = 3.12\}$ and $\{\beta_1 = 24.97, \beta_2 = 19.68, \beta_3 = 23.51, \beta_4 = 19.87\}$ and compare the mean and variance with the observed data. The objective of generating simulated distributions is to compare how good the model can resemble the historical data.

### 5.4.1 Results for the temporal analysis

I use graphical and statistical tests to evaluate the goodness of fit between the observed, generated and fitted data. The statistics of the observed and generated data of rainfall totals are given in Table 5.6. The size of generated data we use is much larger than the observed data (generated 700 years and observed 87 years). Note that, if the generated data is of the same size of the observed, then there is a need to generate repeated samples. Figure 5.3 shows plots of PDFs and CDFs of

the observed and fitted data from the sum of four independent gamma variables.

**Table 5.6:** Means and variances for the sum of months in winter season of observed and generated rainfall totals

| Data | Mean (mm) | Variance |
|---|---|---|
| Observed | 282.98 | 7436.37 |
| Generated | 301.21 | 6650.38 |

The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are used to compare the distribution of observed and generated data. Using the Kolmogorov–Smirnov test, the P-values $> 0.05$. However, for the Anderson–Darling test, the P-values $< 0.05$. Refer to Table 5.7. Therefore, the observed and generated data are not significantly different at the 5% significance level based on the Kolmogorov–Smirnov test. The plots of histogram and CDF of the observed and generated support the results. See Figure 5.4. The plots of the generated and fitted PDF and CDF are also compared in Figure 5.5. The slight bias shown in the Figure 5.5 is possibly because some correlation exist.

Because the Moschopoulos and approximate gamma fits were so close in this example, we expect a similar conclusion with synthetic data generated from the approximate gamma distribution for the sum of four gamma variables. In conclusion, the sum of $n$ independent gamma model is suitable for modelling the sum of monthly rainfall totals between the months in the same season whenever the correlation between the months is insignificant.

**Table 5.7:** P-values of goodness of fit tests for the sum of observed and generated months in winter season

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| P-values | 0.0887 | 0.0269 |

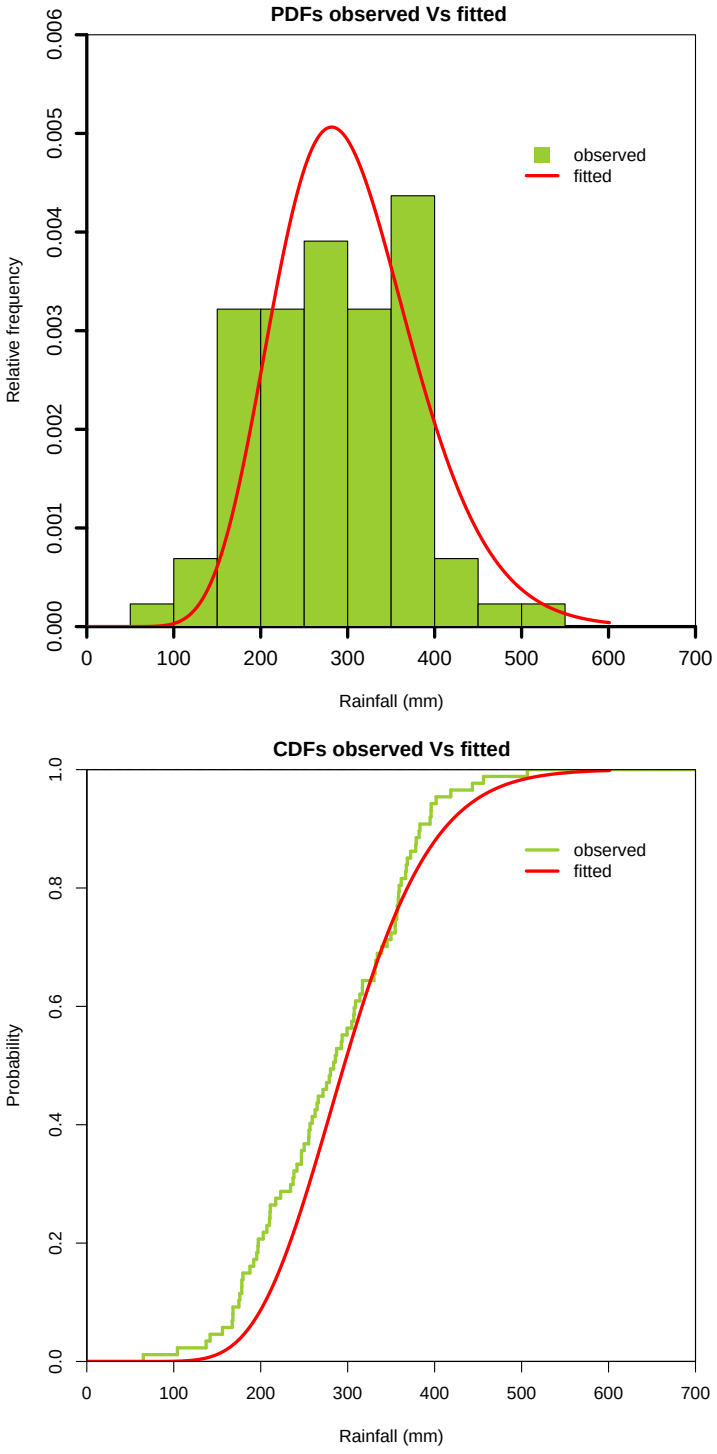



**Figure 5.3:** Observed vs fitted PDFs (above) and CDFs (below) for the sum of four independent gamma variables of winter season, Hume station.

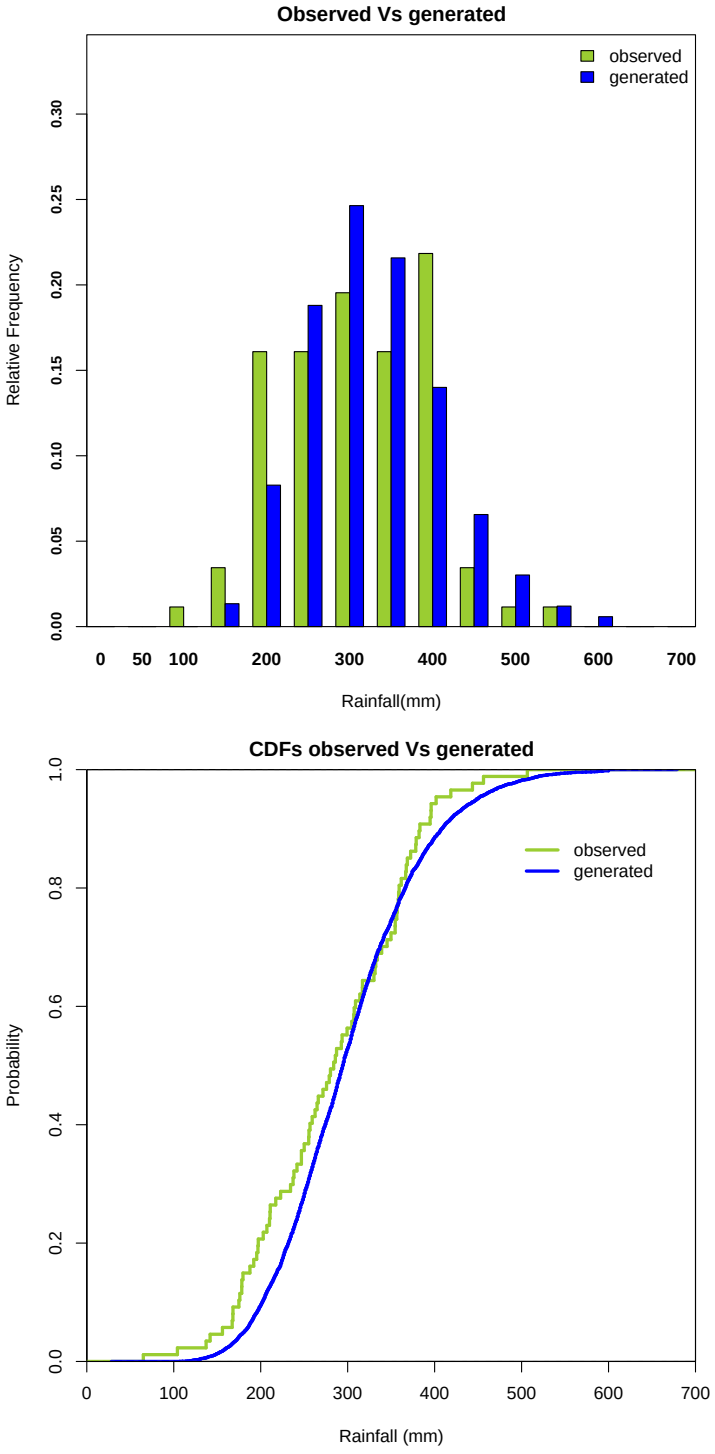



**Figure 5.4:** Plots of observed vs generated (above) and observed vs generated CDFs (below) for the sum of four independent gamma variables of winter season, Hume station.

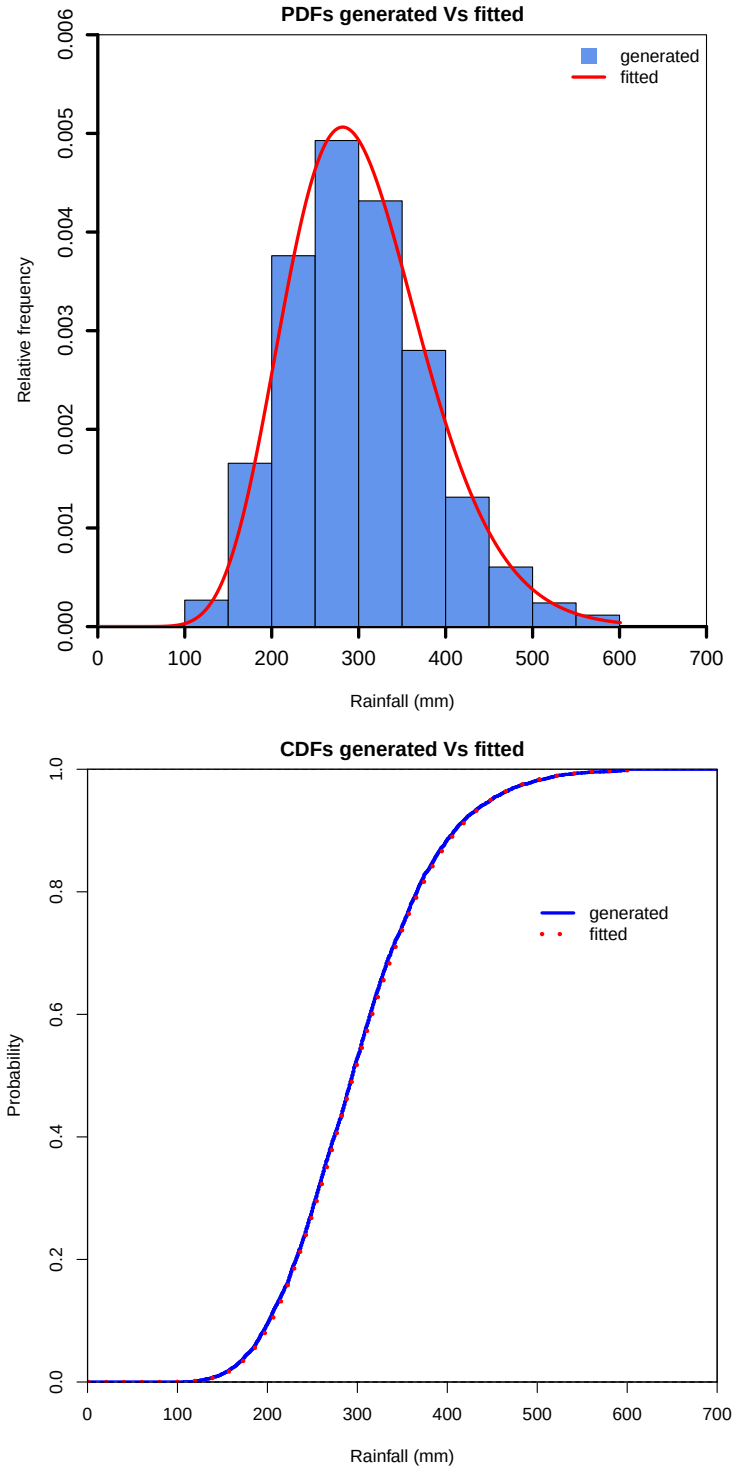



**Figure 5.5:** Generated vs fitted PDFs (above) and CDFs (below) for the sum of four independent gamma variables of winter season, Hume station.

## 5.5 A case study for correlated spatial system

In this section I wish to use the model of the sum of $n$ independent gamma variables to model two stations where the correlation between them is significant. So far, I know that the sum of $n$ independent gamma variables is valid for independent variables. However, I would like to investigate the model of the sum of $n$ independent gamma variables when it is fitted to the sum of correlated gamma variables. I follow the same procedure as described in Section 5.1 to model the sum of two stations within the Murray–Darling Basin.

I select two stations, Dookie and Rutherglen which are about 80 km apart. Refer to Section 1.5 for more information for both stations. The rainfall data from 1922 to 2009 is used in the analysis. Then, each data set is fitted to gamma model where the parameters are estimated using the MLE method. The estimated mean and variance are compared using the formulae of mean ($\alpha\beta$) and variance ($\alpha\beta^2$) from the gamma model. The probability of zero rainfall totals ($p_0$) is also calculated. Refer to Table 5.8 for details. I calculate the Spearman correlation between the stations where the correlation is $r = 0.88$ and is found to be significant at $\alpha = 0.05$ significance level. The goodness of fit tests are conducted to assess the fit between the marginals of the observed and generated gamma variables. See Table 5.9. Because the correlation is high we expect that a model which assumes independence is not likely to be suitable.

**Table 5.8:** Estimated parameters, mean, variance and $p_0$ for Dookie and Rutherglen using the gamma model

| | Estimated | | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $p_0$ | Observed | $\alpha\beta$ | Observed | $\alpha\beta^2$ |
| Dookie | 1.4593 | 32.1950 | 0.0146 | 46.9849 | 46.9835 | 1270.724 | 1512.635 |
| Rutherglen | 1.4864 | 33.3959 | 0.0125 | 49.6408 | 49.6408 | 1353.992 | 1657.801 |

**Table 5.9:** P-values of goodness of fit tests for the marginal of gamma variables from Dookie and Rutherglen stations

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| Dookie | 0.3244 | 0.0393 |
| Rutherglen | 0.5692 | 0.2225 |

Now, the main part is to model the sum of two stations. I construct a set of positive pairs, that is when both stations have rain and calculate the sum of the corresponding pairs. Again use the gamma model to fit each marginal distribution from the set of positive pairs and estimate the shape $(\alpha_1, \alpha_2)$ and scale $(\beta_1, \beta_2)$ parameters. Note that pairs where one or more components are zero must be omitted. Thus we are restricting our estimation to occasions where both marginal totals are non-zero.

To generate the synthetic sum, we use formula (5.2). We have estimated values for $\alpha$ and $\beta$, refer to Table 5.10 and calculate the following values:

$$\beta_1 = 31.0885, \quad C = \left(\frac{\beta_1}{\beta_2}\right)^{\alpha_2} = 0.9610, \quad p = \sum \alpha_i = 3.0928,$$

**Table 5.10:** Estimated parameters, mean and variance for the positive pairs using the gamma model

| | Estimated parameters | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | Observed | $\alpha\beta$ | Observed | $\alpha\beta^2$ |
| Dookie | 1.5218 | 31.0885 | 47.3102 | 47.3102 | 1265.7240 | 1470.8050 |
| Rutherglen | 1.5710 | 31.8867 | 50.0949 | 50.0949 | 1345.2100 | 1597.3600 |

and

$$f(y) = C \sum_{k=0}^{\infty} \frac{\delta_k y^{p+k-1} \exp(-y/\beta_1)}{\Gamma(p+k)\beta_1^{p+k}}$$

and $F(y) = \int_0^y f_Y(w)dw$ is the CDF. The total probability is $F(\infty) = 1$ and is written as

$$F(y) = \int_0^{\infty} f_Y(y)\ dy \approx 0.9998867$$

where the values of $\delta_k$ for $k = 1, 2, \ldots$ are calculated using the recursive formula

$$\delta_k = \frac{1}{k} \sum_{i=1}^{k} i\ \gamma_i\ \delta_{k-i}$$

and where $\gamma_i = \sum_{j=1}^{n} \frac{\alpha_j(1-\beta_1/\beta_j)^i}{i}$ for $i = 1, 2, \ldots$

The method of calculation follows the same procedure as that used in the previous example. Table 5.11 shows the values of the means and variances of the observed data, the fitted data from the sum of two independent gamma variables and using the formulae (mean $=\alpha\beta$, variance$=\alpha\beta^2$). We apply the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests to analyse the results.

**Table 5.11:** Comparison of mean and variance for the sum of positive pairs of observed data, the sum of two independent gamma variables and using formula

| | Observed | Formula | Generated |
|---|---|---|---|
| Mean (mm) | 97.41 | 97.41 | 125.77 |
| Variance | 4815.84 | 5569.01 | 3985.72 |

**Table 5.12:** Goodness of fit tests for the sum of positive pairs from of Dookie and Rutherglen stations

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| P-value | 0.0000 | 0.0000 |

### 5.5.1 Results for the spatial analysis

The values of mean and variance calculated from the independent gamma variables are found to be higher than the values estimated from the observed data and by using formulae. To assess the fit between the observed sum and the generated sum from the model, we apply Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests. The Kolmogorov–Smirnov goodness of fit test is based on the difference between the cumulative distributions of the tested distributions whereas the Anderson–Darling test is based on formula (3.13). The hypothesis of the test is that the two distributions come from the same distributions. The results are presented in Table 5.12. For both analyses, the P-values $< 0.05$ which imply that there is a significant difference between the observed and the fitted data from the model. This is easy to see from the graphs in Figure 5.6. Therefore, I have shown that whenever the two variables are correlated, the model needs to incorporate the correlation pa-

rameter. Hence, the sum of positive pairs independent gamma variables cannot be applied for correlated variables.

## 5.6 Summary

In this chapter we have considered the application of an elementary model for approximate representation of a sum of independent gamma random variables and a more complicated exact representation proposed by Moschopoulos. I have shown that the Moschopoulos model for the sum of independent gamma random variables can be used to model seasonal rainfall as a sum of monthly rainfalls provided the independence assumption is reasonable. Although the fitted graph is statistically acceptable it is apparent from the relevant diagrams that the fitted data is slightly biased. This may be due to small correlations. I also show that the model does not work well if the random variables are correlated. In the next Chapter I will investigate models that allow correlation. Although our model for the temporal analysis is statistically acceptable and our model for the spatial analysis is not, this is all to do with the breakdown of the independence assumption and should not preclude a spatial analysis where independence has been established.

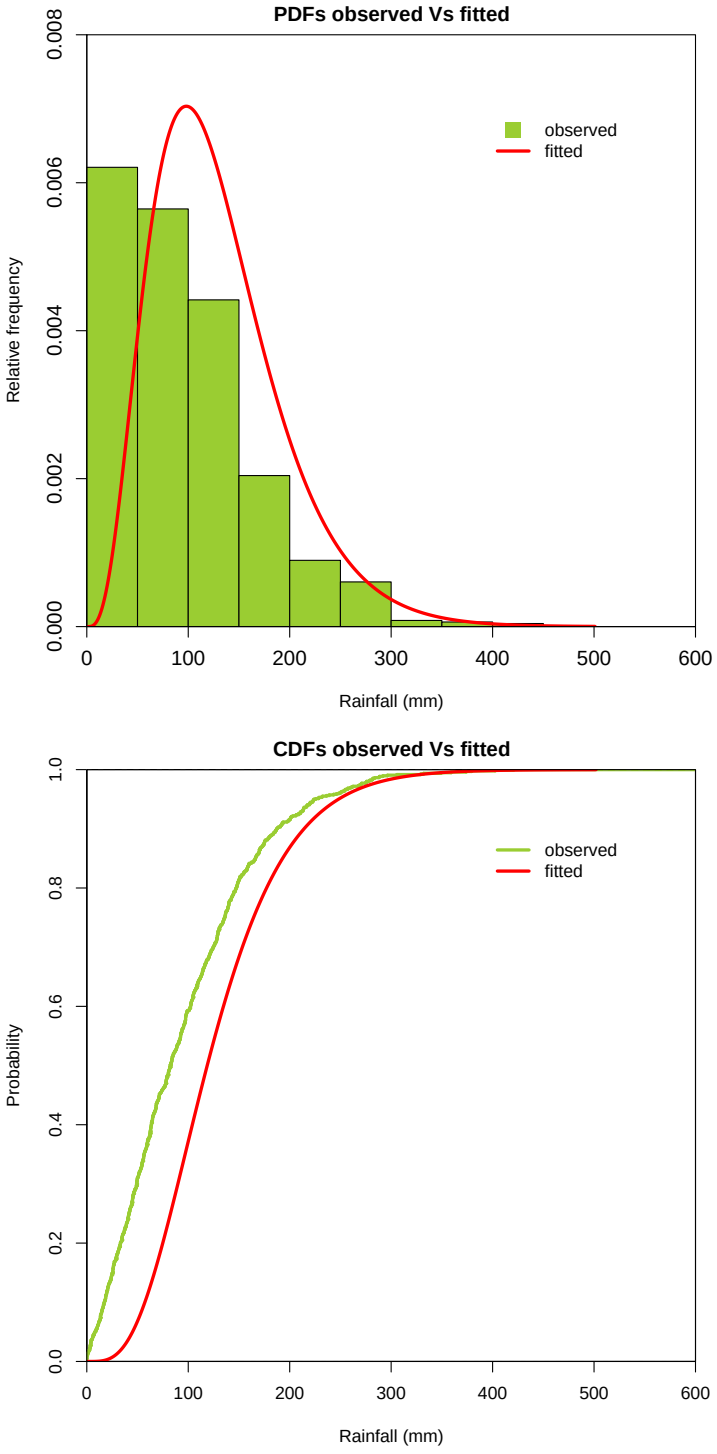



**Figure 5.6:** Observed vs fitted PDFs (above) and CDFs (below) for the sum of two correlated gamma variables, Dookie and Rutherglen stations

# Chapter 6

# A model for the sum of correlated gamma random variables with common shape

In Chapter 5 we found a probability density function (PDF) for the sum of independent gamma random variables. In order to find a PDF for the sum of correlated gamma functions, it is difficult to get an exact formula, so a simple formula is used to approximate a sum of general gamma functions, so the sum is restricted to a common shape parameter. Thus we consider $Y = \sum_{i=1}^{n} X_i$ where $X_i \sim G(\alpha, \beta_i)$ for each $i = 1, 2, \ldots, n$. The new model is more complex because the correlation adds additional parameters. We use the resultant distribution to develop new models for both temporal and spatial analysis of rainfall. In the former we consider total rainfall for four consecutive months at a single station and in the latter we model the total monthly rainfall at four stations within the same region.

## 6.1 Probability density function for the sum of $n$ correlated gamma variables

The correlated gamma variables is employed whenever the correlations between the variables are significant. Kotz and Adams [58] derived a distribution of a sum for identically correlated gamma variables. They compared two forms of distribution for the sum, approximate and exact under the assumption that the sum is also a gamma variable. Moschopoulos [68] extended the technique developed by Kotz and Adams [58] but he considered to the sum of $n$ independent gamma variables with both different shape ($\alpha$) and scale ($\beta$) parameters. Alouini et al. [3] extended the work of Moschopoulos [68] from the sum of $n$ independent gamma variables with both different shape ($\alpha$) and scale ($\beta$) parameters to the sum of $n$ correlated gamma variables with a common shape ($\alpha$) but distinct scale ($\beta$) parameters. There are some difficulties that may arise with the Alouini et al. [3] method which we will comment on later.

Consider a set of $n$ correlated gamma variables, $\{X_i\}_{i=1}^n$ with parameters $\alpha$ and $\beta_i$ written as $X_i \sim G(\alpha, \beta_i)$ where the probability density function (PDF) of $\{X_i\}_{i=1}^n$ is given by

$$f_i(x_i; \alpha, \beta_i) = \frac{x_i^{\alpha-1} e^{-\frac{x_i}{\beta_i}}}{\Gamma(\alpha)\beta_i^{\alpha}} \tag{6.1}$$

for $x_i > 0$ and $\alpha, \beta_i > 0$. Since we are dealing with non normal data, we use the Spearman's rank correlation coefficients $\rho_{ij}$ of $X_i$ and $X_j$ with the formula of

$$\rho_{ij} = 1 - \frac{6 \sum_{k=1}^{n} d_{ijk}^2}{n(n^2 - 1)}$$

where the difference between the ranked pairs $R(x_{i,k})$ and $R(x_{j,k})$ is denoted by $d_{ijk} = R(x_{i,k}) - R(x_{j,k})$ and $n$ is the number of pairs. Refer to Chapter 3.5 for

further detail.

Define $Y = \sum_{i=1}^{n} X_i$ as the sum of $n$ correlated gamma variables with the PDF [3] given by

$$f_Y(y) = \prod_{i=1}^{n} \left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha} \sum_{k=0}^{\infty} \frac{\delta_k \; y^{n\alpha+k-1} \exp\left(-y/\lambda_1\right)}{\lambda_1^{n\alpha+k} \; \Gamma(n\alpha+k)} \tag{6.2}$$

where $\{\lambda_i\}_{i=1}^{n}$ are the eigenvalues of the matrix $A = DC$ where we assume that $A$ is positive definite and where the minimum eigenvalue is denoted by $\lambda_1$. The matrices $C$ and $D$ are defined by

$$D = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{bmatrix}_{n\times n}, \quad C = \begin{bmatrix} 1 & \sqrt{\rho_{12}} & \cdots & \sqrt{\rho_{1n}} \\ \sqrt{\rho_{21}} & 1 & \cdots & \sqrt{\rho_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\rho_{n1}} & \sqrt{\rho_{n2}} & \cdots & 1 \end{bmatrix}_{n\times n}.$$

Note that the application of the Alouini et al. [3] method depends on knowing that $A$ is positive definite and this condition must be tested by calculating the eigenvalues of $A$ for each potential application. After reading the paper by Alouini et al. [3] it would be reasonable to assume that the matrix $C$ is guaranteed to be positive definite. The authors do not indicate any limitations to their proposed analysis, but I will show later that this is not necessarily true. The coefficients of $\delta_k$ are defined by

$$\delta_{k+1} = \frac{\alpha}{k+1} \sum_{i=1}^{k+1} \left[\sum_{j=1}^{n} \left(1 - \frac{\lambda_1}{\lambda_j}\right)^{i}\right] \delta_{k+1-i} \tag{6.3}$$

for $k = 0, 1, 2 \ldots$ where $\delta_0 = 1$. This method works well and $\delta_k$ converges rapidly to zero if the terms $\left(1 - \frac{\lambda_1}{\lambda_j}\right)$ are all small (say less than 0.7). However we shall see later in our case study that $\lambda_1 \approx 1.78$ and $\lambda_4 \approx 174.81$. Thus, in our case study, convergence is appallingly slow.

In Section 6.3 we present some theory behind the work of Kotz and Adams [58] and extend to the sum of correlated gamma variables with the same shape parameters. Before we do that however we will consider a much more elementary approximation.

## 6.2 An approximate representation of the sum of dependent gamma random variables

Let $X_i \sim G(\alpha_i, \beta_i)$ for $i = 1, 2, \ldots, n$ be dependent gamma random variables and let $Y = X_1 + \cdots + X_n$. Suppose that

$$\rho_{ij} = \frac{E[(x_i - \alpha_i\beta_i)(x_j - \alpha_j\beta_j)]}{\sqrt{\alpha_i\alpha_j} \cdot \beta_i\beta_j}$$

for $i < j$ and each $i, j = 1, 2, \ldots, n$. We note that

$$\begin{aligned} E[Y] &= \sum_{i=1}^{n} E[X_i] \\ &= \sum_{i=1}^{n} \alpha_i\beta_i \end{aligned}$$

and

$$\begin{aligned} & E[\{Y - (\alpha_1\beta_1 + \cdots + \alpha_n\beta_n)\}^2] \\ = & \sum_{i=1}^{n} E[(X_i - \alpha_i\beta_i)^2] + \sum_{i=1}^{n}\sum_{j=1}^{i-1} 2E[(X_i - \alpha_i\beta_i)(X_j - \alpha_j\beta_j)] \\ = & \sum_{i=1}^{n} \alpha_i\beta_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{i-1} 2\rho_{ij}\sqrt{\alpha_i\alpha_j} \cdot \beta_i\beta_j. \end{aligned}$$

If we wished to approximate $Y$ by a gamma random variable $Z$ then we could choose $Z$ so that the mean and variance were equal. Thus we would solve the equations

$$\alpha\beta = \sum_{i=1}^{n} \alpha_i\beta_i \tag{6.4}$$

and

$$\alpha\beta^2 = \sum_{i=1}^{n} \alpha_i \beta_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} 2\rho_{ij} \sqrt{\alpha_i \alpha_j} \cdot \beta_i \beta_j \tag{6.5}$$

to obtain

$$\alpha = \frac{\left(\sum_{i=1}^{n} \alpha_i \beta_i\right)^2}{\sum_{i=1}^{n} \alpha_i \beta_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} 2\rho_{ij} \sqrt{\alpha_i \alpha_j} \cdot \beta_i \beta_j} \tag{6.6}$$

and

$$\beta = \frac{\sum_{i=1}^{n} \alpha_i \beta_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} 2\rho_{ij} \sqrt{\alpha_i \alpha_j} \cdot \beta_i \beta_j}{\sum_{i=1}^{n} \alpha_i \beta_i}. \tag{6.7}$$

Hence we approximate $Y$ by $Z \sim G(\alpha, \beta)$. This is a direct extension of the approximation method used earlier for the sum of independent gamma random variables.

### 6.2.1 A case study of an approximate representation for the sum of dependent gamma random variables

We choose Dookie and Rutherglen meteorological stations to illustrate the approximate form of the sum of dependent gamma random variables where in this case $n = 2$. In Chapter 5.5.1, we know that correlation between Dookie and Rutherglen is $\rho = 0.88$ and that it is significant at 5% significance level. The methodology is set up for describing the relationship between the two series through the correlation coefficient. Therefore, each set of data is fitted to the gamma distribution and the parameter values estimated are

$$\alpha_1 = 1.52, \quad \beta_1 = 31.09, \quad \alpha_2 = 1.57 \quad \text{and} \quad \beta_2 = 31.89.$$

For $n = 2$, the formulae of (6.4) and (6.5) become

$$\alpha\beta = \alpha_1\beta_1 + \alpha_2\beta_2$$

and

$$\alpha\beta^2 = \alpha_1\beta_1^2 + \alpha_2\beta_2^2 + 2\rho_{21}\sqrt{\alpha_2\alpha_1}\beta_1\beta_2.$$

Now, we calculate

$$\alpha = \frac{(\alpha_1\beta_1 + \alpha_2\beta_2)^2}{\alpha_1\beta_1^2 + \alpha_2\beta_2^2 + 2\rho_{21}\sqrt{\alpha_2\alpha_1}\beta_1\beta_2} = 1.644017$$

and

$$\beta = \frac{\alpha_1\beta_1^2 + \alpha_2\beta_2^2 + 2\rho_{21}\sqrt{\alpha 2\alpha_1}\beta_1\beta_2}{\alpha_1\beta_1 + \alpha_2\beta_2} = 59.19896.$$

The plots of the PDF and CDF of the approximate form of the sum of dependent gamma variables and the sum of the observed data are given in Figure 6.1. The results from the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests shown in Table 6.1 also validate the graphical analysis. The P-values for both tests are not significant at 5% significance level. The results indicate that both the observed and fitted data are not significantly different.

**Table 6.1:** P-values of goodness of fit tests for the sum of dependent gamma variables

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| P-value | 0.3559 | 0.1244 |

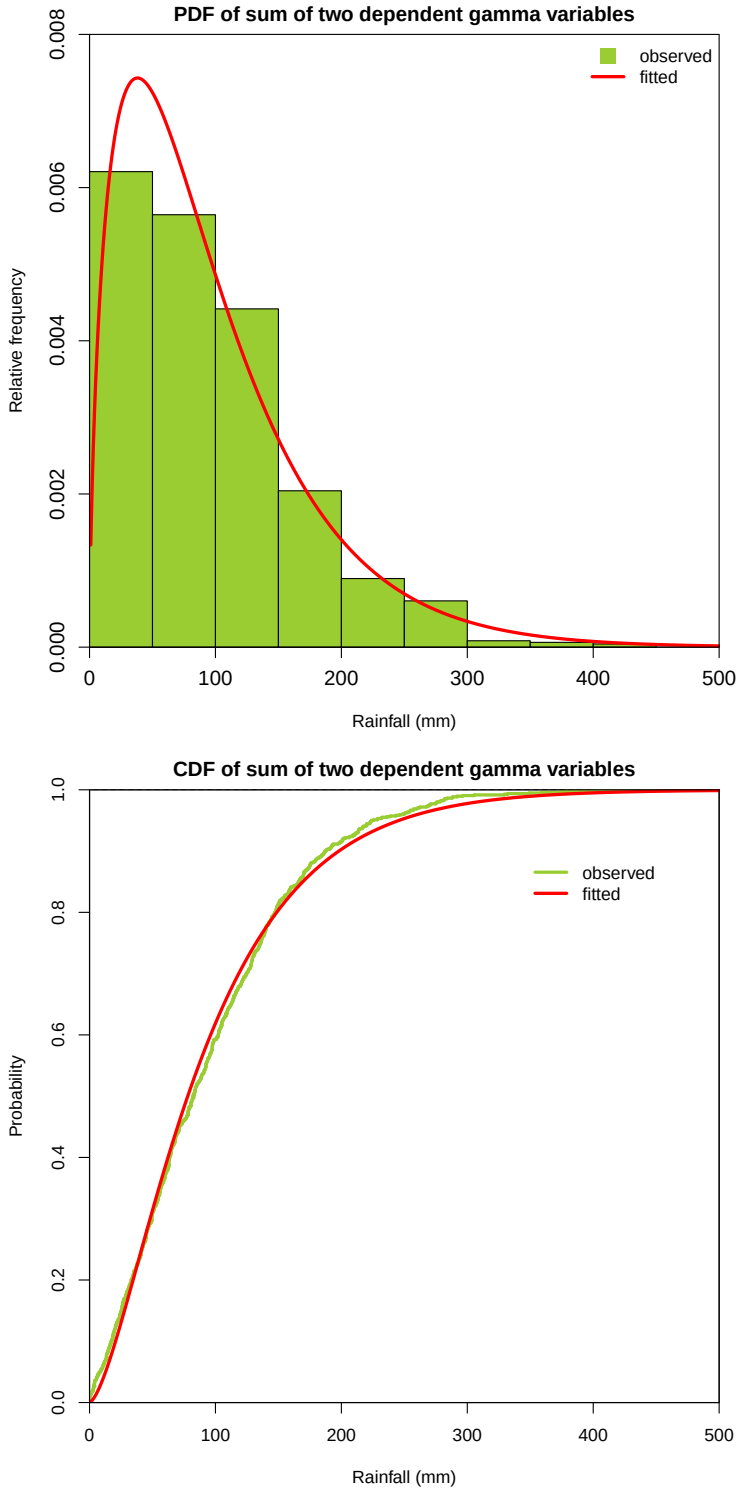



**Figure 6.1:** PDF (above) and CDF (below) for the sum of two correlated gamma variables from two stations (Dookie and Rutherglen)

## 6.3 An exact representation of the sum of dependent gamma random variables with the same shape – the extended Kotz and Adams method

We propose a moment generating function $\varphi : (-\delta, \delta)^n \to \mathbb{R}$ in the form

$$\varphi(t_1, t_2, \ldots, t_n) = [\det(I - TBS)]^{-\alpha}.$$

That is, we propose jointly distributed real valued random variables $X_1, X_2, \ldots, X_n$ such that

$$E[e^{t_1X_1+t_2X_2+\cdots+t_nX_n}] = [\det(I - TBS)]^{-\alpha}$$

where

$$T = \begin{bmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_n \end{bmatrix}, \quad B = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{bmatrix}$$

and

$$S = \begin{bmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{12} & 1 & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1n} & s_{2n} & \cdots & 1 \end{bmatrix}$$

where $|s_{ij}| \leq 1$ for all $1 \leq i < j \leq n$. We note immediately that

$$\begin{aligned} E[e^{t_iX_i}] &= \varphi(0, \ldots, 0, t_i, 0, \ldots, 0) \\ &= (1 - t_i\beta_i)^{-\alpha} \end{aligned}$$

and hence the marginal random variable $X_i \sim G(\alpha, \beta_i)$ is a gamma random variable with shape parameter $\alpha$ and scale parameter $\beta_i$. Note that the joint distribution

of $(X_i, X_j)$ where $i, j$ is described by the characteristic function $\psi_{ij} : (-\delta, \delta)^2 \to \mathbb{R}$ defined by

$$\begin{aligned}
\psi(t_i, t_j) &= E[e^{t_i X_+ t_j X_j}] \\
&= \varphi(0, \ldots, 0, t_i, 0, \ldots, 0, t_j, 0, \ldots, 0) \\
&= \left\{ \det \begin{bmatrix} 1 - t_i\beta_i & -t_i\beta_i s_{ij} \\ -t_j\beta_j s_{ij} & 1 - t_j\beta_j \end{bmatrix} \right\}^{-\alpha} \\
&= \left\{\beta_i\beta_j(1 - s_{ij}^2)t_i t_j - (\beta_i t_i + \beta_j t_j) + 1\right\}^{-\alpha}.
\end{aligned}$$

It follows that

$$\frac{\partial \psi}{\partial t_i}(t_i, t_j) = (-\alpha)\left\{\beta_i\beta_j(1 - s_{ij}^2)t_i t_j - (\beta_i t_i + \beta_j t_j) + 1\right\}^{-\alpha-1}\left(\beta_i\beta_j(1 - s_{ij}^2)t_j - \beta_i\right)$$

and hence that

$$\begin{aligned}
\frac{\partial^2 \psi}{\partial t_i \partial t_j}(t_i, t_j) &= \alpha(\alpha+1)\left\{\beta_i\beta_j(1 - \rho_{ij})t_i t_j - (\beta_i t_i + \beta_j t_j) + 1\right\}^{-\alpha-2} \\
&\qquad \times \left\{\beta_i\beta_j(1 - s_{ij}^2)t_j - \beta_i\right\}\left\{\beta_i\beta_j(1 - s_{ij}^2)t_i - \beta_j\right\} \\
&\quad -\alpha\left\{\beta_i\beta_j(1 - s_{ij}^2)t_i t_j - (\beta_i t_i + \beta_j t_j) + 1\right\}^{-\alpha-1}\beta_i\beta_j(1 - s_{ij}^2)
\end{aligned}$$

from which it follows that

$$\begin{aligned}
E[X_i X_j] &= \frac{\partial^2 \psi}{\partial t_i \partial t_j}(0, 0) \\
&= \alpha(\alpha+1)\beta_i\beta_j - \alpha\beta_i\beta_j(1 - s_{ij}^2) \\
&= \alpha^2\beta_i\beta_j + \alpha\beta_i\beta_j s_{ij}^2.
\end{aligned}$$

Therefore $E[(X_i - \alpha\beta_i)(X_j - \alpha\beta_j)] = \alpha\beta_i\beta_j s_{ij}^2$ and hence the variables $(X_i, X_j)$ are correlated with correlation coefficient

$$\begin{aligned}\rho_{ij} &= \frac{E[(X_i - \alpha\beta_i)(X_j - \alpha\beta_j)]}{\sqrt{E[(X_i - \alpha\beta_i)^2]E[(X_j - \alpha\beta_j)^2]}} \\ &= \frac{E[(X_i - \alpha\beta_i)(X_j - \alpha\beta_j)]}{\alpha\beta_i\beta_j} \\ &= s_{ij}^2.\end{aligned}$$

Thus we must assume $\rho_{ij} \geq 0$ and that the matrix $C$ is given by

$$C = \begin{bmatrix} 1 & \sqrt{\rho_{12}} & \cdots & \sqrt{\rho_{1n}} \\ \sqrt{\rho_{12}} & 1 & \cdots & \sqrt{\rho_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\rho_{1n}} & \sqrt{\rho_{2n}} & \cdots & 1 \end{bmatrix}.$$

Suppose we define the random variable $Y = X_1 + X_2 + \cdots + X_n$. Then

$$\begin{aligned}E[e^{tY}] &= \varphi(t, t, \ldots, t) \\ &= [\det(I - tBC)]^{-\alpha} \\ &= [\det(I - tA)]^{-\alpha}\end{aligned}$$

where we have written $A = BC$ for convenience. If $A$ is positive definite then it follows that

$$\begin{aligned}m_Y(t) &= E[e^{tY}] \\ &= \prod_{i=1}^{n}(1 - \lambda_i t)^{-\alpha}\end{aligned}$$

where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n$ are the eigenvalues of $A$. We note that Alouini et al. [3] apparently assume that $C$ is positive definite. We show later that this need

not be true and that $C$ may be indefinite. If $C$ does turn out to be positive definite then

$$x^T C x \geq 0 \quad \Leftrightarrow \quad (B^{1/2}y)^T C (B^{1/2}y) \geq 0 \quad \Leftrightarrow \quad y^T B^{1/2} C B^{1/2} y \geq 0$$

for all $x, y \in \mathbb{R}^n$ and so $B^{1/2}CB^{1/2}$ is also positive definite. Since

$$B^{1/2}CB^{1/2}v = \lambda v \quad \Leftrightarrow \quad BCu = \lambda u$$

where $u = B^{1/2}v$ it follows that $\lambda$ is an eigenvalue of $BC$ if and only if it is an eigenvalue of $B^{1/2}CB^{1/2}$. Thus $A = BC$ is also positive definite. In general one would need to calculate the eigenvalues of $A$ in each potential application. In those cases where $A$ is positive definite the moment generating function has the same form as the moment generating function obtained by Moschopoulos [68] in the case of summation of independent gamma random variables it follows that we can use the same inversion technique to obtain a series of single gamma functions for the PDF.

In the following sections, we will present two case studies to demonstrate the application of the model. Before we do that we will present a counter-example to show that the matrix $S$ need not be positive definite.

**Note** Alouini et al. [3] assume that $\rho_{ii} = 1$ and that $0 \leq \rho_{ij} \leq 1$ and then their analysis implies that $C$ is a positive definite matrix. This implication is not justified as we shall show in the next section.

## 6.4 A counter-example

Let $\{Z_k\}_{k=1}^3$ be independent real-valued random variables with mean 0 and variance 1. Choose unit length vectors

$$\boldsymbol{w}_k = \begin{bmatrix} w_{1k} \\ w_{2k} \\ w_{3k} \end{bmatrix}$$

for each $k = 1, 2, 3$ and define new random variables $\{X_i\}_{i=1}^3$ by setting

$$X_i = \sum_{k=1}^{3} w_{ik} Z_k$$

for each $i = 1, 2, 3$. Now

$$\begin{aligned} E[X_i X_j] &= E\left[\sum_{k=1}^{3} w_{ik} Z_k \sum_{\ell=1}^{3} w_{j\ell} Z_\ell\right] \\ &= \sum_{k=1}^{3}\sum_{\ell=1}^{3} w_{ik} w_{j\ell} E[Z_k Z_\ell] \\ &= \sum_{k=1}^{3} w_{ik} w_{jk} \\ &= \langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle \\ &= \cos\theta_{ij} \end{aligned}$$

where we suppose $\cos\theta_{ij} \geq 0$ for all $i < j$. If we define a vector-valued random variable

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

then

$$\begin{aligned} R &= E[X_i X_j^T] \\ &= \begin{bmatrix} X_1^2 & X_1X_2 & X_1X_3 \\ X_2X_1 & X_2^2 & X_2X_3 \\ X_3X_1 & X_3X_2 & X_3^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \cos\theta_{12} & \cos\theta_{13} \\ \cos\theta_{12} & 1 & \cos\theta_{23} \\ \cos\theta_{13} & \cos\theta_{23} & 1 \end{bmatrix}. \end{aligned}$$

Since $\boldsymbol{p}^T R \boldsymbol{p} = \boldsymbol{p}^T E[X_i X_j^T]\boldsymbol{p} = E[(\boldsymbol{p}^T X)^2] \geq 0$ for all $\boldsymbol{p} \in \mathbb{R}^3$ it follows that the matrix $R = [\rho_{ij}]$ is non-negative. If we assume that $|\theta_{ij}| \leq \pi/2$ for all $i, j$ then $\cos\theta_{ij} \geq 0$ and we can define

$$\begin{aligned} C &= [\sqrt{\rho_{ij}}\,] \\ &= \begin{bmatrix} 1 & \sqrt{\cos\theta_{12}} & \sqrt{\cos\theta_{13}} \\ \sqrt{\cos\theta_{12}} & 1 & \sqrt{\cos\theta_{23}} \\ \sqrt{\cos\theta_{13}} & \sqrt{\cos\theta_{23}} & 1 \end{bmatrix}. \end{aligned}$$

We will show with a simple counter-example that $C$ is not necessarily non-negative definite. If we let $\theta_{12} = 0.1$, $\theta_{13} = 0.2$ and $\theta_{23} = 0.3$ then according to MATLAB

$$R = \begin{bmatrix} 1.000000000000000 & 0.995004165278026 & 0.980066577841242 \\ 0.995004165278026 & 1.000000000000000 & 0.955336489125606 \\ 0.980066577841242 & 0.955336489125606 & 1.000000000000000 \end{bmatrix}$$

with eigenvalues

$$\mathrm{eig}(R) = \begin{bmatrix} 0.000000000000000 \\ 0.046304160638511 \\ 2.953695839361488 \end{bmatrix}.$$

On the other hand

$$C = \begin{bmatrix} 1.000000000000000 & 0.997498955026032 & 0.989983119977933 \\ 0.997498955026032 & 1.000000000000000 & 0.977413161935937 \\ 0.989983119977933 & 0.977413161935937 & 1.000000000000000 \end{bmatrix}$$

with eigenvalues

$$\text{eig}(C) = \begin{bmatrix} -0.000033060657551 \\ 0.023413116937492 \\ 2.976619943720060 \end{bmatrix}.$$

Hence, in this case, the matrix $C$ is indefinite. The selected angles for the matrices are feasible. In fact the sum of the two smaller angles equals the largest angle so this represents an instance where the random variables are co-planar. The point of the example is to show that there are feasible $R$ matrices for which the $C$ matrices are not non-negative. We conclude this section by observing that the Alouini et al. [3] method requires $C$ to be non-negative definite and the implication that the matrix $C$ defined above is a *positive definite matrix* is misleading. In fact if one wishes to apply the Alouini et al. [3] method then it is necessary to check the eigenvalues of $C$ for each particular case study to check that $C$ is indeed non-negative.

## 6.5 A case study for spatial analysis

This case study is aimed to show if we know the statistical parameters for several stations within a particular catchment then we may be able to use a weighted sum to determine a rainfall model for the entire local catchment. We conduct a spatial analysis for the sum of rainfall volumes from four selected meteorological stations within the same region using the monthly rainfall data.

We wish to model a volume of total catchment from four meteorological stations

using the sum of four correlated gamma variables given by formula (6.2). The chosen meteorological stations are located in the upper Murray river of the Murray–Darling Basin, namely Hume Reservoir, Beechworth Composite, Dookie and Rutherglen Research. The detail information of the meteorological stations location and their statistical summary of the means and variances for monthly rainfall totals are given in Chapter 1.5. We use monthly rainfall data based on completeness and the years of data selected are between 1922 and 2008.

For convenience, the size of surface area of Hume Dam is used as the area of Hume catchment. The catchment area ($km^2$) for each station in Victoria is approximated using the interactive map software from the Victorian Government land services and spatial information website, see (http://www.land.vic.gov.au). On the website, choose the '*interactive map*' window then write the area name we want to estimate and use the '*measure area*' in the tool bar to approximate the chosen area. An example on how to use the interactive map is shown for Dookie, see Figure 6.2. The approximated area for the stations is in Table 6.2. The volume of a catchment is calculated using the formula

$$\text{Volume}\,(m^3) = 10^3 \times \text{area}\,(km^2) \times \text{rainfall}\,(\text{mm})$$

where the unit of volume is in cubic metre ($m^3$). Normally, the unit for the area is given in square kilometre ($km^2$) and the unit for rainfall totals is in millimetres (mm). Therefore, we convert the unit of volume into cubic metre. The Spearman's rank correlation between the data from four stations is also calculated. We find that there is a significant correlation at 1% significant level with correlation coefficients greater than 0.8 (see Table 6.3).

In principal, we will use the following steps to do the analyses:

- Fit each data set to the gamma distribution. Estimate the shape and scale parameters for the gamma distribution using maximum likelihood estimation

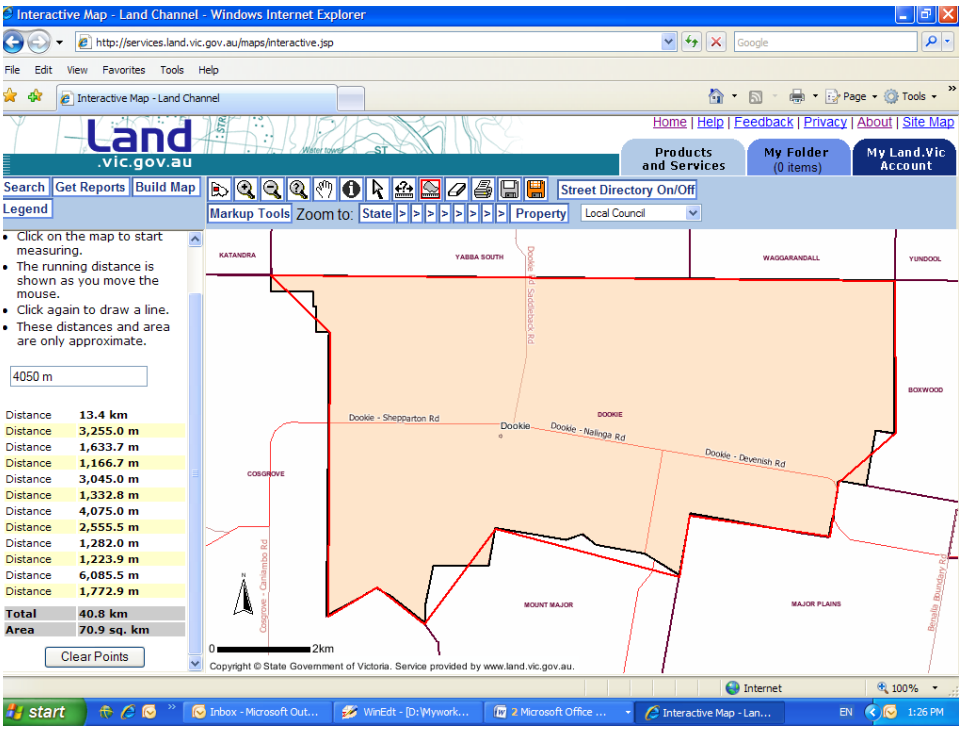



**Figure 6.2:** Approximate area of Dookie catchment

**Table 6.2:** Approximate area for catchment coverage

| Station | Area ($km^2$) |
|---|---|
| Hume | 180 |
| Beechworth | 157 |
| Dookie | 70 |
| Rutherglen | 125 |

method. Calculate the mean and variance using the formulae, $\alpha\beta$ and $\alpha\beta^2$ respectively. Compare the mean and variance from the formulae with the observed values.

- Consider the PDF of the model to generate synthetic data. Identify and cal-

**Table 6.3:** Spearman's rank correlation coefficients between the stations

| | Beechworth | Dookie | Rutherglen |
|---|---|---|---|
| Hume | 0.90 | 0.84 | 0.91 |
| Beechworth | | 0.87 | 0.91 |
| Dookie | | | 0.88 |

culate the relevant parameters required for the model. Compare the statistics of the observed and generated data (mean and variance). Plot the PDF and CDF for both observed and generated data. Use the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests to assess the fit between the observed and generated data statistically.

### 6.5.1 Fitting volume of rainfall totals using gamma distribution

We choose a gamma distribution to fit the individual volumes of monthly rainfall with shape and scale parameters estimated by the maximum likelihood estimation (MLE) method. The gamma distribution is used based on the reasons describe in Section 3.1. The probability density function for a gamma random variables is given in equation (3.1). The MLE method is applied to estimate the parameters, see Section 3.1.2. The values of $\alpha$ and $\beta$ associated with the volume are calculated and shown in Table 6.4.

**Table 6.4:** Estimated parameters, means and variances of the observed and formulae from gamma distribution

| | Parameters | | Mean ($m^3$) | | Variance | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | observed | $\alpha\beta$ | observed | $\alpha\beta^2$ |
| Hume | 1.67 | 6419000 | 10693000 | 10693000 | 573922000 | 686335000 |
| Beechworth | 1.84 | 3266000 | 6017000 | 6017000 | 165422000 | 196544000 |
| Dookie | 1.53 | 4863000 | 7435000 | 7435000 | 311788000 | 361600000 |
| Rutherglen | 1.58 | 3968000 | 6268000 | 6268000 | 210024000 | 248702000 |

### 6.5.2 PDF for the sum of volume of rainfall totals to form a catchment

To use the PDF for the sum of four correlated gamma variables in formula (6.2), we need a single $\alpha$ with distinct values of $\beta$ and the correlation parameter ($\rho_{ij}$). One might ask whether it is possible to choose $\bar{\alpha}$ and $\beta_1, \ldots, \beta_4$ so that the overall mean matches exactly. We show later (see Chapter 8) that such a procedure may not provide a satisfactory resolution. The goodness of fit to the individual data will effect the quality of the calculated estimate for the sum. Similarly after adjusting the distributions to have equal alpha values, which only occur later in Section 6.5.4. For convenience and since the values of $\alpha$ given in Table 6.4 are not dissimilar, the average of $\alpha$ is used for the PDF. Therefore, the average of $\alpha$ denoted by $\overline{\alpha} = (\alpha_1 + \cdots + \alpha_4)/4 = 1.65$ is calculated. Then using the $\overline{\alpha}$ fixed, the values of $\beta_i$ using $X_i \sim G(\overline{\alpha}, \beta_i)$ for each $i = 1, \ldots, 4$ are re-estimated using the maximum likelihood estimation method, see Table 6.5. The $R$ code is provided for the calculation of the $\beta_i$, refer to Appendix C.4. Now substitute the new values for the parameters into the PDF of the sum of four correlated gamma variables. Since the volume of

**Table 6.5:** Re-estimated values of $\beta$ for $\overline{\alpha} = 1.65$

| | Hume | Beechworth | Dookie | Rutherglen |
|---|---|---|---|---|
| $\beta_i$ | 6464000 | 3586000 | 4495000 | 3789000 |

four stations is significantly correlated, we apply the sum of four correlated gamma variables (see Section 6.1). Let $Y = X_1 + \cdots + X_4$ denote the sum of four stations. The matrix of the correlation coefficients $(\rho_{ij})$, matrix $C$ and $D$ are constructed as follows

$$\rho = \begin{bmatrix} 1 & 0.89 & 0.83 & 0.91 \\ & 1 & 0.86 & 0.91 \\ & & 1 & 0.88 \\ & & & 1 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 64.64 & 0 & 0 & 0 \\ & 35.86 & 0 & 0 \\ & & 44.95 & 0 \\ & & & 37.89 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0.95 & 0.91 & 0.95 \\ & 1 & 0.93 & 0.95 \\ & & 1 & 0.94 \\ & & & 1 \end{bmatrix}.$$

Note that the matrices $\rho$, $D$ and $C$ are symmetric. Therefore we calculate the eigenvalues of matrix $C$ which are given by

$$\{0.0418, 0.0500, 0.0931, 3.8151\}.$$

Hence, matrix $C$ is positive definite. Since $A = DC$ we have

$$A = \begin{bmatrix} 64.64 & 61.41 & 58.82 & 61.41 \\ 34.07 & 35.86 & 33.35 & 34.07 \\ 40.90 & 41.80 & 44.95 & 42.25 \\ 36.00 & 36.00 & 35.62 & 37.89 \end{bmatrix}$$

and the eigenvalues of matrix $A$ are

$$\lambda_i = \{174.81, 4.78, 1.97, 1.78\}$$

where the minimum eigenvalue is $\lambda_1 = 1.78$. Then define $\lambda_2 = 1.97$, $\lambda_3 = 4.78$ and $\lambda_4 = 174.81$. Consider the product term and substitute the parameters values given above. We obtain

$$\frac{\lambda_1}{\lambda_2} = 0.9035533, \quad \frac{\lambda_1}{\lambda_3} = 0.3723849 \quad \text{and} \quad \frac{\lambda_1}{\lambda_4} = 0.01018248$$

where

$$C = \left(\frac{\lambda_1}{\lambda_2}\right)^{\alpha} \left(\frac{\lambda_1}{\lambda_3}\right)^{\alpha} \left(\frac{\lambda_1}{\lambda_4}\right)^{\alpha} = 0.003426097.$$

The formula of the probability density function of the sum of four correlated gamma variables is comprised of two terms: the product term and the summation term given by

$$f_Y(y) = \prod_{i=2}^{4} \left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha} \sum_{k=0}^{\infty} \frac{\delta_k \; y^{4\alpha+k-1} \exp\left(-y/\lambda_1\right)}{\lambda_1^{4\alpha+k} \; \Gamma(4\alpha+k)}$$

where $\alpha = 1.6541$ and the values of $\delta_k$ for $k = 1, 2, \ldots$ are calculated using the formula

$$\delta_k = \frac{\alpha}{k} \sum_{i=1}^{k} \left[\sum_{j=1}^{4} \left(1 - \frac{\lambda_1}{\lambda_j}\right)^{i}\right] \delta_{k-i}.$$

Note that $(1 - \lambda_1/\lambda_4) \approx 0.99$ which means the convergence of the coefficient $\delta_k$ will be extremely slow. In this case it is necessary to use 1000 terms to obtain a reasonable approximation. Values of $\delta_k$ calculated are plotted in Figure 6.3. Since individual terms in this summation comprise products of both very large and very

small numbers we avoid unwanted numerical errors by expressing the density as

$$\begin{aligned} f_Y(y) &= \prod_{i=1}^{4} \left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha} \exp\left(\log\left[\sum_{k=0}^{\infty} \frac{\delta_k \; y^{4\alpha+k-1} \exp\left(-y/\lambda_1\right)}{\lambda_1^{4\alpha+k} \; \Gamma(4\alpha+k)}\right]\right) \\ &= \prod_{i=1}^{4} \left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha} \sum_{k=0}^{\infty} \exp\Big( \log \delta_k + (4\alpha+k-1)\log y - y/\lambda_1 \\ &\qquad -(4\alpha+k)\log\lambda_1 - \log\Gamma(4\alpha+k)\Big). \end{aligned}$$

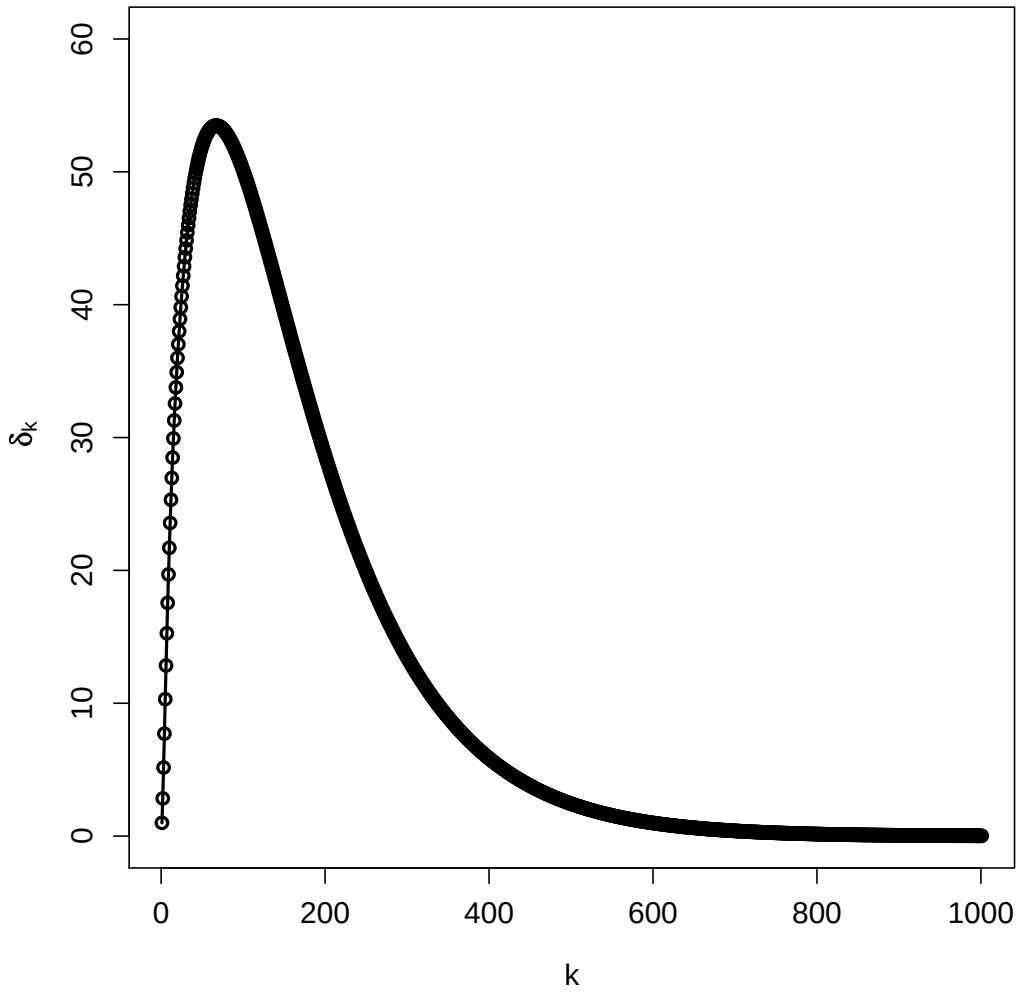



**Figure 6.3:** Values of $\delta_k$

The numerical integration can be used to obtain the CDF and construct the look-up table to determine the associated rainfall sums. The detail procedure of generating the synthetic sum is described in Section 3.3. To get the sum from the observed data, we calculate the corresponding sum from four stations. Then, we compare the statistics of the sum between the observed and generated sum. Refer to Table 6.10. We use Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests to assess the fit between the distributions of the sum of observed and generated rainfalls. Figure 6.4 presents the PDFs and CDFs for both sums, observed and fitted respectively.

**Table 6.6:** Statistics for the sum of observed and generated

| | Mean ($m^3$) | Variance | Median | IQR |
|---|---|---|---|---|
| Observed | 30328000 | 4189403000 | 26496000 | 28724000 |
| Generated | 30959000 | 5253858000 | 25100000 | 27800000 |

**Table 6.7:** P-values of goodness of fit tests for the sum of four correlated gamma variables from four stations

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| Observed vs. Generated | 0.3411 | 0.1018 |

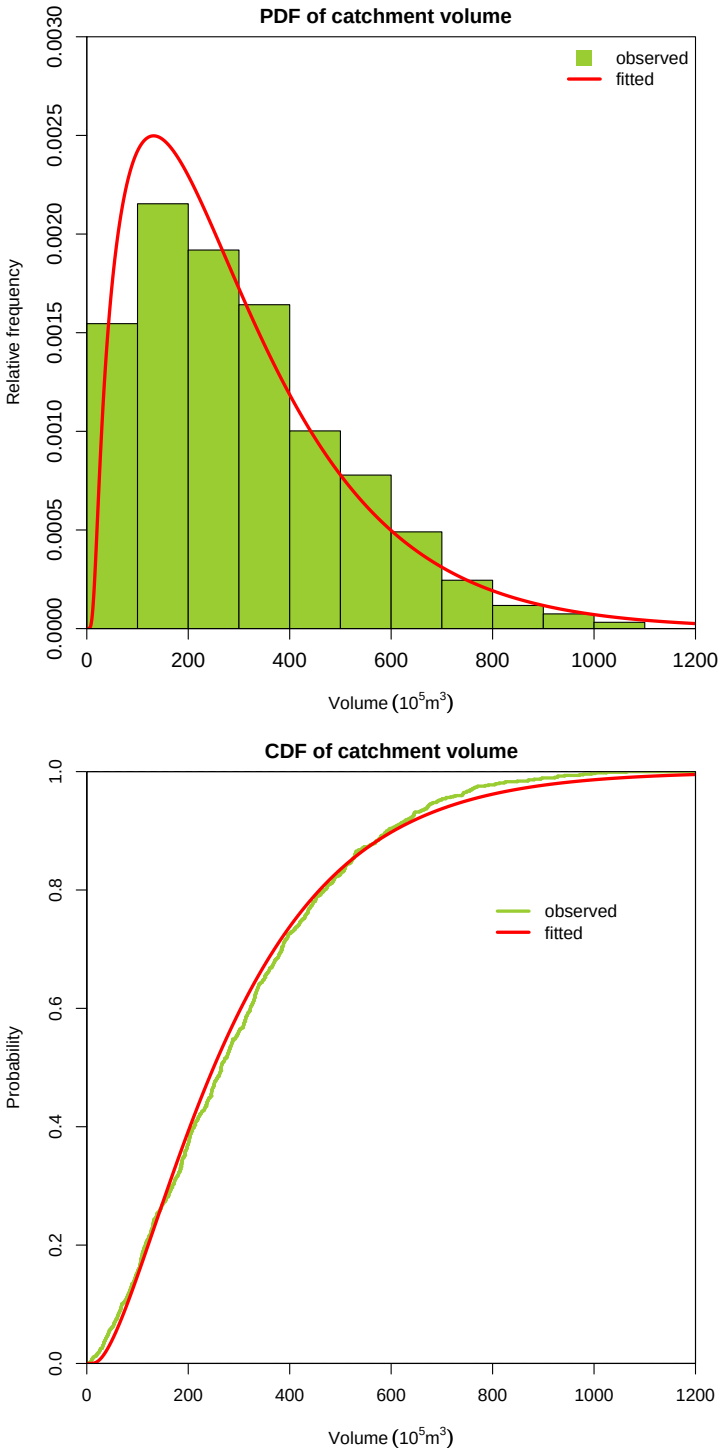



**Figure 6.4:** Observed vs fitted PDFs (above) and CDFs (below) for the sum of four correlated gamma variables from four stations (Hume, Beechworth, Dookie and Rutherglen)

### 6.5.3 Using the generalised correlative coherence to compare the result for the PDF for the sum of volume of rainfall totals to form a catchment

To apply Alouini et al. [3] method, matrix $C$ is constructed from the square root of the correlation coefficients. Hence, the correlation values must be all positive values to guarantee that matrix $C$ is positive definite. If some of the correlation values are negatives, we can still use the Alouini et al. [3] method by first calculating the generalised correlation suggested by Getz [38].

In this section, using the similar data as in Section 6.5.2, we wish to compare the results using the direct correlation values with the generalised correlation values. Note that, the values of correlation for the data are all positives. If the generalised correlation can give results which is as good as the results in Section 6.5.2, then the generalised correlation is an advantage.

We follow the same process as in Section 6.5.2 and the formulae used in this section can be found in Section 6.1. The correlation matrix is based on Getz [38] in which we have described in Section 3.6. The correlation matrices of correlation, $C$ and $D$ are given by

$$\rho = \begin{bmatrix} 1 & 0.8812 & 0.8812 & 0.8812 \\ & 1 & 0.8812 & 0.8812 \\ & & 1 & 0.8812 \\ & & & 1 \end{bmatrix}$$

$$D = \begin{bmatrix} 64.64 & 0 & 0 & 0 \\ & 35.86 & 0 & 0 \\ & & 44.95 & 0 \\ & & & 37.89 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0.9387 & 0.9387 & 0.9387 \\ & 1 & 0.9387 & 0.9387 \\ & & 1 & 0.9387 \\ & & & 1 \end{bmatrix}.$$

Note that the matrices $\rho$, $D$ and $C$ are symmetric. The eigenvalues of matrix $C$ are $\{3.8161, 0.0613, 0.0613, 0.0613\}$. Hence, matrix $C$ is positive definite. Since $A = DC$ we have

$$A = \begin{bmatrix} 64.64 & 60.68 & 60.68 & 60.68 \\ 33.66 & 35.86 & 33.66 & 33.66 \\ 42.19 & 42.19 & 44.95 & 42.19 \\ 35.57 & 35.57 & 35.57 & 37.89 \end{bmatrix}$$

and the eigenvalues of matrix $A$ are

$$\lambda_i = \{175.09, 3.45, 2.55, 2.25\}$$

where the minimum eigenvalue is $\lambda_1 = 2.25$. Then define $\lambda_2 = 2.55$, $\lambda_3 = 3.45$ and $\lambda_4 = 175.09$. The descriptive statistics of the sum of observed and generated using the generalised correlation is given in Table 6.8. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are conducted to assess the fit between the distributions of the sum of observed and generated catchment volumes. Refer to Table 6.9. There is not much different between Figures 6.5 and 6.4. Therefore, the use of generalised correlation is sufficient.

**Table 6.8:** Statistics for the sum of observed and generated

| | Mean ($m^3$) | Std | Median | IQR |
|---|---|---|---|---|
| Observed | 30328000 | 20468031 | 26496000 | 28724000 |
| Generated | 31011000 | 22971000 | 25100000 | 27825000 |

**Table 6.9:** P-values of goodness of fit tests for the sum of four correlated gamma variables from four stations

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| Observed vs. Generated | 0.3628 | 0.1014 |

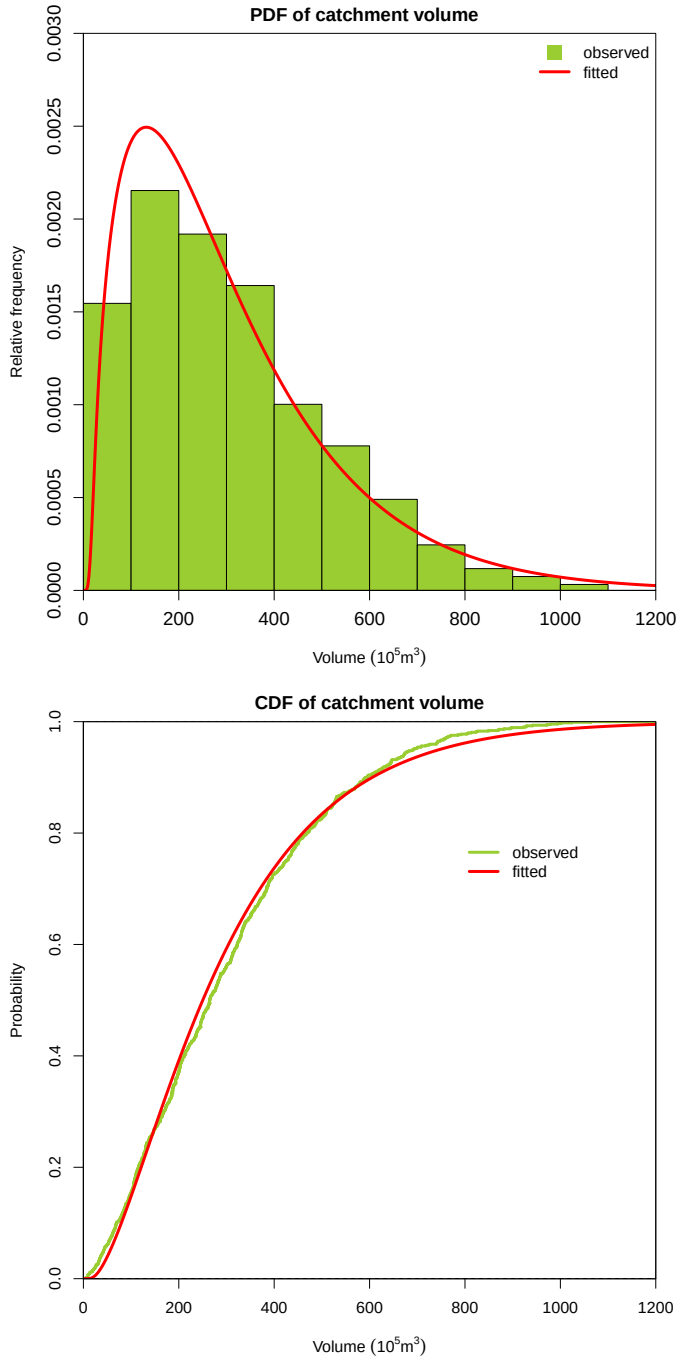



**Figure 6.5:** Observed vs fitted PDFs (above) and CDFs (below) for the sum of four correlated gamma variables from four stations (Hume, Beechworth, Dookie and Rutherglen) using the same correlation values

### 6.5.4 Marginals for spatial analysis

We generate four sets of simulated data using the average $\alpha$ ($\overline{\alpha}$) with the respective re-estimated values of $\beta_i$ for $i = 1, \ldots, 4$. Refer to Table 6.5. Then, each set of simulated data is compared to each set of observed data. The statistics of the marginals of the observed and simulated are given in Table 6.10. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are used to compare each marginal statistically, refer to Table 6.11. Figures 6.6 supports the graphical aspect of the analyses.

**Table 6.10:** Means and variance of marginals in spatial analysis

| | Mean ($m^3$) | | Variance | |
|---|---|---|---|---|
| | observed | generated | observed | generated |
| Hume | 10693000 | 10876000 | 573958000 | 640800000 |
| Beechworth | 5932000 | 6050000 | 160801000 | 218276000 |
| Dookie | 7435000 | 7664000 | 311811000 | 377979000 |
| Rutherglen | 6268000 | 6174000 | 210039000 | 222973000 |

**Table 6.11:** Goodness of fit tests for marginals in spatial analysis

| | Hume | Beechworth | Dookie | Rutherglen |
|---|---|---|---|---|
| Kolmogorov–Smirnov | 0.8769 | 0.6463 | 0.6933 | 0.1530 |
| Anderson–Darling | 0.5028 | 0.2770 | 0.3353 | 0.1108 |

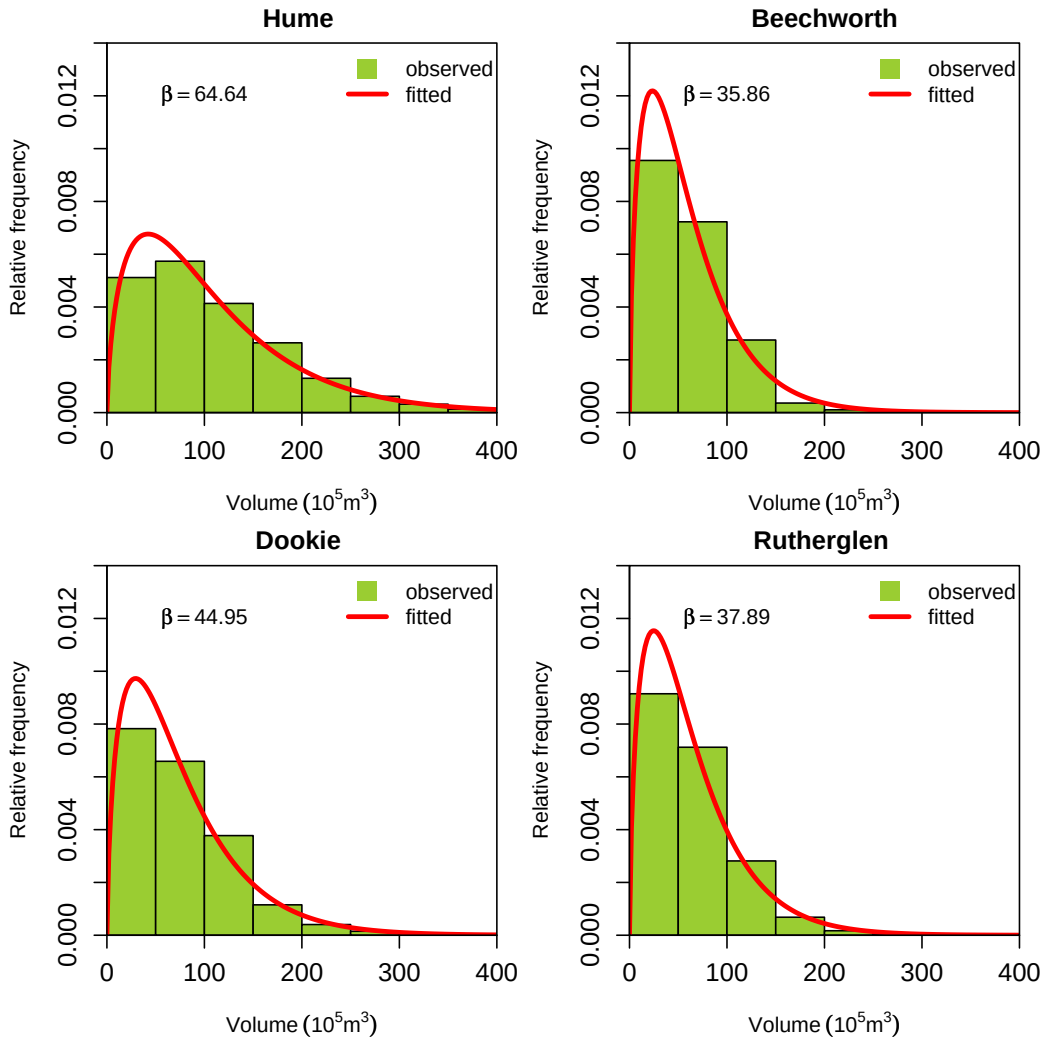



**Figure 6.6:** Plots of the observed versus fitted gamma sum model for Hume, Beechworth, Dookie and Rutherglen

### 6.5.5 Results for spatial analysis

We use graphical and statistical tests to evaluate the goodness of fit between the observed and generated data. The analyses are divided into two parts–analysis for the sum and analysis for the marginals.

In analysis for the sum of four correlated gamma variables, the P-value $>$ 0.05

for both goodness of fit tests, Kolmogorov–Smirnov and Anderson–Darling. These shows that the generated sum is not significantly different from the observed sum at 5% significance level, see Table 6.7.

In the marginal analysis, we compare each single set of data between the observed and the generated data and found that the P-values are all greater than 0.05, refer to Table 6.11. Hence the marginal for each observed and simulated data is not significantly different at 5% significance level.

Figures 6.6 and 6.4 show the plots of histogram and their fitted curves of the PDFs and their respective CDFs. Graphically, they also demonstrate the same results. In conclusion, the sum of $n$ correlated gamma model is suitable for modelling the sum of rainfall volume from nearby stations within the same region which has significant correlation to obtain the total volume of a catchment.[1]

## 6.6 A case study for temporal analysis

In this section, we wish to apply the same model of the sum of $n$ correlated gamma variables, for the temporal analysis, we wish to demonstrate if by knowing the statistical characteristic of a single month, we are able to determine a distribution of a seasonal rainfall. This is in the framework of the investigation of the characteristic of aggregated correlated random variables from the knowledge of the same variable of a higher frequency. We choose Hume meteorological station and use monthly rainfall data from 1922–2009. The analysis is done in two stages: Stage 1, according to summer and winter seasons, respectively; and Stage 2, for a single month in

[1]Note that simple matching of the mean and variance could also have been used. The primary purpose of this section has been to illustrate another possible methodology. There are many comparisons that could be made in discussing the relative merits of the different methods. We will address some of these issues later.

each season. In both stages, we compared the statistics of the observed data and generated data and confirm the results using the Kolmogorov–Smirnov goodness of fit test.

### 6.6.1 Using the gamma distribution to fit individual monthly rainfall totals according to season

To illustrate the method, two seasons, summer (December–March) and winter (June–September) from Hume station are used in the analysis. Care must be taken when we consider the summer season so that the data matches December from year $k$ with January–March from year $k+1$. Therefore, to avoid this problem, we arrange the data so it starts from May 1922 to April 1923 and continues the same order until May 2009 to April 2010. If we do not carefully arrange the data according to correct sequence, the correlation values calculated will be incorrect. The descriptive statistics for the seasons are given in Table 6.12. The marginals of the monthly rainfall totals are fitted to the gamma distribution. The methodology used is similar to Section 6.5.1. The estimated shape and scale parameters and the observed and estimated means and variances, respectively are given in Table 6.13.

### 6.6.2 Modelling the sum of monthly rainfall totals according to season

To model the sum of monthly rainfall totals according to season, we again use the probability density function of the sum of $n$ correlated gamma variables and follow a similar procedure as described in Section 6.1. Since the monthly rainfall totals we use are not normally distributed, we apply the Spearman's rank correlation (non-parametric correlation) method to calculate the monthly correlations and found that

**Table 6.12:** Descriptive statistics of winter and summer seasons

| | Mean (mm) | Variance | Median (mm) | IQR (mm) |
|---|---|---|---|---|
| | | Winter | | |
| Jun | 68.02 | 1582.69 | 64.70 | 58.65 |
| Jul | 77.03 | 1545.23 | 71.40 | 51.05 |
| Aug | 75.92 | 1398.71 | 76.00 | 48.60 |
| Sep | 62.00 | 1072.84 | 57.20 | 45.90 |
| | | Summer | | |
| Dec | 48.44 | 1941.85 | 37.10 | 50.95 |
| Jan | 44.78 | 2179.19 | 24.70 | 50.50 |
| Feb | 41.67 | 1665.17 | 33.70 | 52.20 |
| Mar | 43.12 | 1336.30 | 33.40 | 47.45 |

the P-value $> 0.05$ for every pair of months which show that the monthly correlations are not significant, see Section 5.3. However, there are small correlations between the months. The monthly correlations are shown in Table 6.14 and Table 6.15 for winter and summer, respectively.

To apply the model of sum of $n$ correlated gamma variables, the matrix $C$ has to be positive definite. A matrix is said to be positive definite if all its eigenvalues are positive. To obtain the matrix $C$, we calculate the square root of the corresponding correlation values. If there is negative correlation value, the construction of matrix $C$ is not possible since the square root of the negative value is undefined. Therefore, we have to apply the method of Getz [38] described in Section 3.6 to rectify the negative correlation values. In the method, a generalised correlative coherence gives an overall correlation between more than the two series of correlated variables, refer

**Table 6.13:** Estimated parameters, observed and estimated means and variances of monthly rainfall

| | Parameter | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | observed | $\alpha\beta$ | observed | $\alpha\beta^2$ |
| | | | Winter | | | |
| Jun | 2.72 | 24.97 | 68.02 | 68.02 | 1582.69 | 1698.39 |
| Jul | 3.91 | 19.68 | 77.03 | 77.03 | 1545.23 | 1515.94 |
| Aug | 3.23 | 23.51 | 75.92 | 75.92 | 1398.71 | 1784.82 |
| Sep | 3.12 | 19.87 | 62.00 | 62.00 | 1072.84 | 1231.95 |
| | | | Summer | | | |
| Dec | 1.03 | 48.34 | 49.69 | 49.69 | 2018.91 | 2402.17 |
| Jan | 0.95 | 48.81 | 46.33 | 46.33 | 2257.00 | 2261.48 |
| Feb | 1.06 | 41.91 | 44.21 | 44.21 | 1654.22 | 1852.85 |
| Mar | 1.18 | 36.49 | 43.04 | 43.04 | 1279.65 | 1570.32 |

to [38] and [13]. Once the negative correlation values are removed, then the model of the sum of $n$ correlated gamma variables is applied. The numerical illustration of the generalised correlative coherence is given in Section 6.6.5.

**Table 6.14:** Monthly correlations of winter season

| | Jul | Aug | Sep |
|---|---|---|---|
| Jun | 0.12 | 0.13 | -0.01 |
| Jul | | 0.13 | 0.24 |
| Aug | | | 0.18 |

**Table 6.15:** Monthly correlations of summer season

| | Jan | Feb | Mar |
|---|---|---|---|
| Dec | 0.29 | 0.09 | -0.14 |
| Jan | | 0.26 | -0.05 |
| Feb | | | 0.32 |

### 6.6.3 Rainfall model for the sum of positive quadruples of correlated gamma variables

For illustration, we fit a sum of positive quadruples monthly rainfall data in winter and summer seasons selected before, to the sum of $n$ correlated gamma distribution described in [3].

We divide our analysis into two sections according to winter season and summer season. The sum of $n$ correlated gamma variables uses a common $\alpha$ but different values of $\beta$. For each season the values of $\alpha$ are not dissimilar (see Table 6.13), so we take an averaged values of $\alpha$ ($\overline{\alpha}$) for each season and use $\overline{\alpha}$ to re-estimate the values of $\beta$ (see Table 6.16). We are encouraged by the fact that in the previous case study the averaged values of $\alpha$ appeared to give good results. The re-estimated values of $\beta$ are obtained by the Newton–Raphson iteration method using a fixed value of $\alpha$. The methodology has been described in Section 3.1.2 and the $R$ code is provided in Appendix C.4.

**Table 6.16:** Average $\alpha$ ($\overline{\alpha}$) and re-estimate $\beta$ s

| Season | Parameters | | | | |
|---|---|---|---|---|---|
| | $\overline{\alpha}$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
| Winter | 3.2472 | 20.9475 | 23.7234 | 23.3793 | 19.0948 |
| Summer | 1.0529 | 47.1937 | 44.0027 | 41.992 | 40.8778 |

### 6.6.4 Marginals of positive quadruples

We begin with winter months. For each month, we generate 1000 random numbers from the gamma distribution using the respective parameters in Table 6.16. For example, the synthetic data for June is generated using $\overline{\alpha} = 3.2472$ and $\beta_1 = 20.9475$. The statistical summary of the observed and generated data of the winter season are shown in Table 6.17. The estimated mean and variance are calculated using the formula $\alpha\beta$ and $\alpha\beta^2$, respectively. Similarly, for the summer season, we generate 1000 data from the gamma distribution using the respective parameters in Table 6.16. The statistical summary of the observed data and generated data of the summer months are shown in Table 6.18. Section 6.6.7 discuss the results of the analysis.

#### Analysis of marginals of gamma for observed and generated data

Before further analyses are conducted, we assess the fit of the marginals for both seasons, winter and summer. We compare the marginals of winter season (June, July, August, September) and summer season (December, January, February, March) to the marginals of the generated data, respectively. The value of the respective shape and scale parameters to generate synthetic data are given in Table 6.16.

From the Kolmogorov–Smirnov test we obtain P-value $> 0.05$ for both analyses–analysis of marginals of winter season and summer season (see Table 6.19). We conclude that the distributions of the observed and the generated data are not significantly different from one another at 5% significant level. The histogram of observed data and the fitted curve support the results we have for the Kolmogorov–Smirnov test (see Figure 6.7 and Figure 6.8).

**Table 6.17:** Winter months–Mean and variance of the observed and generated

| | Mean (mm) | | Variance | |
|---|---|---|---|---|
| | observed | estimated | observed | estimated |
| Jun | 68.02 | 68.02 | 1582.45 | 1424.86 |
| Jul | 77.03 | 77.03 | 1545.28 | 1827.52 |
| Aug | 75.92 | 75.92 | 1398.76 | 1774.89 |
| Sep | 62.00 | 62.00 | 1072.56 | 1183.97 |

**Table 6.18:** Summer months–Mean and variance of the observed and generated

| | Mean (mm) | | Variance | |
|---|---|---|---|---|
| | observed | estimated | observed | estimated |
| Dec | 49.69 | 49.69 | 2018.70 | 2345.07 |
| Jan | 46.33 | 46.33 | 2257.20 | 2038.67 |
| Feb | 44.21 | 44.21 | 1654.05 | 1856.61 |
| Mar | 43.04 | 43.04 | 1279.49 | 1759.39 |

**Table 6.19:** Kolmogorov–Smirnov test for marginal distributions

| | | | | |
|---|---|---|---|---|
| | | Winter | | |
| | Jun | Jul | Aug | Sep |
| P-value | 0.8164 | 0.8855 | 0.8570 | 0.9380 |
| | | Summer | | |
| | Dec | Jan | Feb | Mar |
| P-value | 0.8434 | 0.3736 | 0.7626 | 0.2865 |

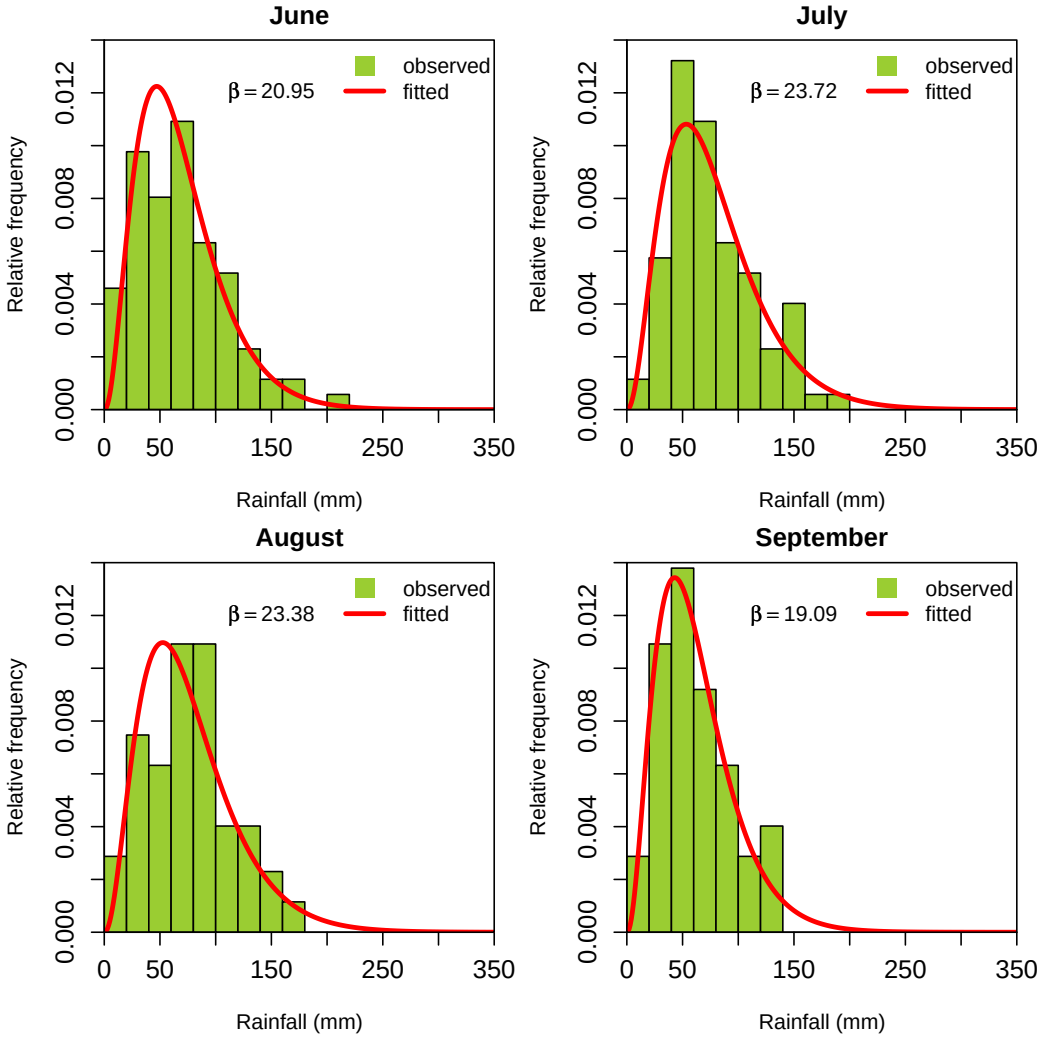



**Figure 6.7:** Winter season–observed and fitted

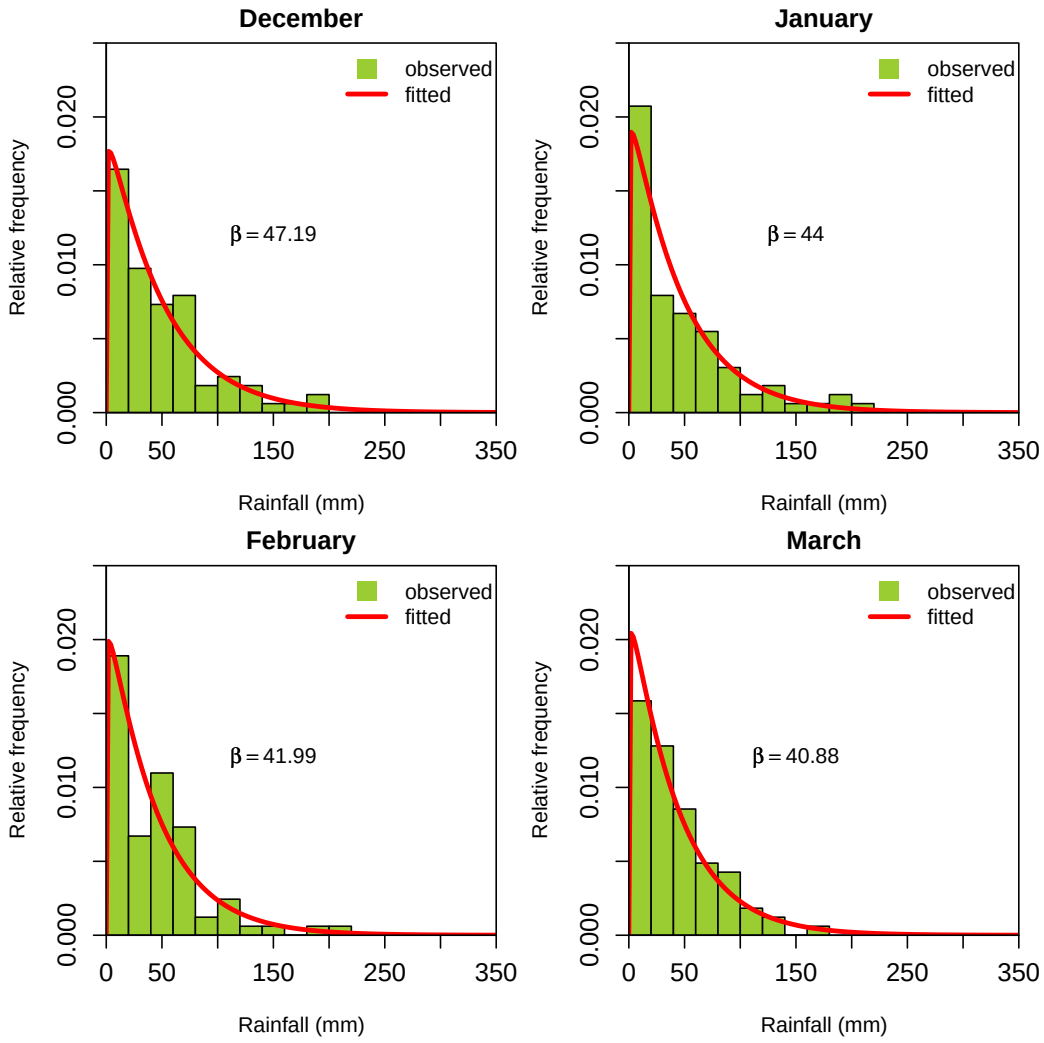



**Figure 6.8:** Summer season–observed and fitted

In the following section, we use equation 6.2 to generate two sets of data of sum of positive quadruples according to winter and summer seasons.

### 6.6.5 Sum of positive quadruples of winter season

To begin with, we consider the monthly correlation matrix

$$R = \begin{bmatrix} 1 & 0.12 & 0.13 & -0.01 \\ & 1 & 0.13 & 0.24 \\ & & 1 & 0.18 \\ & & & 1 \end{bmatrix}.$$

Since one of the correlation values is negative, we apply the method described in Section 6.6.2. We estimate the value of $r = 0.16$ by solving $C(X^n) = C_n(r)$ using the Newton's method and the MATLAB code is given in Appendix D.1. Now we construct a new correlation matrix ($\rho$) and continue using the model of the sum of $n$ correlated gamma distribution. Since we have four months, then $y = x_1+x_2+x_3+x_4$ and the correlation matrix ($\rho$), matrix $C$ and $D$ are constructed as follows

$$\rho = \begin{bmatrix} 1 & 0.16 & 0.16 & 0.16 \\ & 1 & 0.16 & 0.16 \\ & & 1 & 0.16 \\ & & & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0.39 & 0.39 & 0.39 \\ & 1 & 0.39 & 0.39 \\ & & 1 & 0.39 \\ & & & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 20.95 & 0 & 0 & 0 \\ & 23.72 & 0 & 0 \\ & & 23.38 & 0 \\ & & & 19.09 \end{bmatrix}.$$

Matrix $A$ is the product of matrix $D$ and $C$ given by

$$A = \begin{bmatrix} 20.95 & 8.17 & 8.17 & 8.17 \\ 9.25 & 23.72 & 9.25 & 9.25 \\ 9.12 & 9.12 & 23.38 & 9.12 \\ 7.45 & 7.45 & 7.45 & 19.09 \end{bmatrix}$$

and the eigenvalues of matrix $A$ are

$$\lambda_i = \{47.41, 14.37, 13.34, 12.02\}$$

for $i = 1, \ldots, 4$ with the minimum eigenvalue is $\lambda_1 = 12.02$. Now we use the probability density function of sum of $n$ correlated gamma variables to generate a synthetic rainfall data. Firstly, we select a sequence of numbers $0, 2, \ldots, 1000$ and calculate the value of the probability for each number using the probability density function of sum of $n$ correlated gamma variables. Then, we calculate the cumulative density function (CDF) and we construct a lookup table of CDF associated with the number sequences. We generate one thousand uniform numbers on $[0, 1]$ and use the $'vlookup'$ function in *Excel* to match the random number with the CDF value and obtain the sequence number associated with it. The value of the sequence number is corresponded to rainfall totals. For example, the random value of 0.9441 is associated to 440 sequence number which gives a rainfall totals of 440 mm. We also develop a code in *R* package which has a similar function as $'vlookup'$, refer to Appendix C.5. We analyse the goodness of fit between the observed data and generated data using the Kolmogorov–Smirnov test, then present the plots of histogram and fitted curve. The results are discussed in Section 6.6.7.

### 6.6.6 Sum of positive quadruples of summer season

We apply similar procedures as in Section 6.6.5 to analyse the sum of positive quadruples of summer season. Since the monthly correlation have some negative elements, we use [38] as before to solve the problem of non-positive definite. The

monthly correlation is

$$R = \begin{bmatrix} 1 & 0.28 & 0.06 & -0.12 \\ & 1 & 0.22 & -0.08 \\ & & 1 & 0.34 \\ & & & 1 \end{bmatrix}.$$

We solve the value of $r = 0.23$ by solving $C(X^n) = C_n(r)$ using the Newton's method. A new correlation matrix ($\rho$) is

$$\rho = \begin{bmatrix} 1 & 0.23 & 0.23 & 0.23 \\ & 1 & 0.23 & 0.23 \\ & & 1 & 0.23 \\ & & & 1 \end{bmatrix}$$

and the matrices $C$ and $D$ are

$$C = \begin{bmatrix} 1 & 0.48 & 0.48 & 0.48 \\ & 1 & 0.48 & 0.48 \\ & & 1 & 0.48 \\ & & & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 47.19 & 0 & 0 & 0 \\ & 44.00 & 0 & 0 \\ & & 41.99 & 0 \\ & & & 40.88 \end{bmatrix}.$$

Matrix $A$ is the product of matrix $D$ and $C$,

$$A = \begin{bmatrix} 47.19 & 22.65 & 22.65 & 22.65 \\ 21.12 & 44.00 & 21.12 & 21.12 \\ 20.16 & 20.16 & 41.99 & 20.16 \\ 19.62 & 19.62 & 19.62 & 40.88 \end{bmatrix}$$

and the eigenvalues of matrix $A$ are $\lambda_i = \{106.26, 23.92, 22.38, 21.49\}$ for $i = 1, \ldots, 4$. Therefore, the minimum eigenvalue is $\lambda_1 = 21.49$. Using the same procedure as described in Section 3.3, we generate synthetic rainfall totals. Then, we compare the observed data and the generated data using Kolmogorov–Smirnov goodness of fit test and we also compare the plots of histogram and fitted curve of both data. The results are discussed in Section 6.6.7.

### 6.6.7 Results for temporal analysis

The results show the analyses of the sum of four correlated gamma variables for observed data and generated data according to two seasons, winter and summer.

We compare the distribution of the relative frequencies of sum of observed positive quadruples with the data generated using the probability density function of the sum of four correlated gamma variables for winter and summer seasons. We conduct a Kolmogorov–Smirnov goodness of fit test and support with the graphical representation (histogram and fitted curve) of relative frequencies of the observed data versus the generated data.

From the Kolmogorov–Smirnov test we obtain P-value $> 0.05$ for both analyses–analysis of sum of four correlated gamma variables of winter season and summer season (see Table 8.4). We conclude that the distribution of the observed data and the generated data are not significantly different from one another at 5% significant level. The histogram of observed data and the fitted curve support the results we have for the Kolmogorov–Smirnov test (see Figure 6.10).

**Table 6.20:** Mean and variance of the observed and generated rainfall totals for winter and summer seasons

| | Mean (mm) | | Variance | |
|---|---|---|---|---|
| | observed | estimated | observed | estimated |
| Winter | 282.98 | 296.29 | 7435.61 | 9502.35 |
| Summer | 183.27 | 204.45 | 9729.85 | 13691.34 |

**Table 6.21:** Kolmogorov–Smirnov test for sum of correlated gamma variables

| | Winter | Summer |
|---|---|---|
| P-value | 0.5330 | 0.1460 |

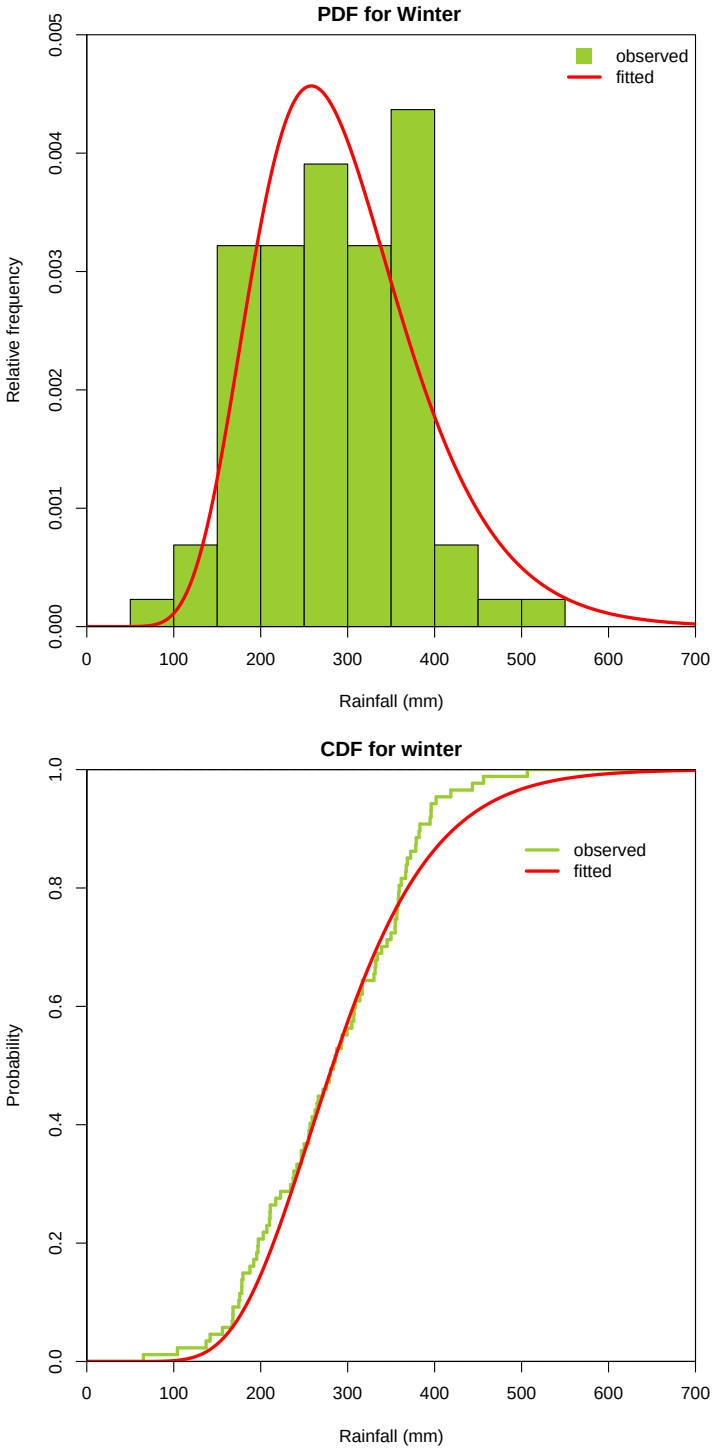



**Figure 6.9:** PDF and CDF for observed and fitted gamma sum for winter season

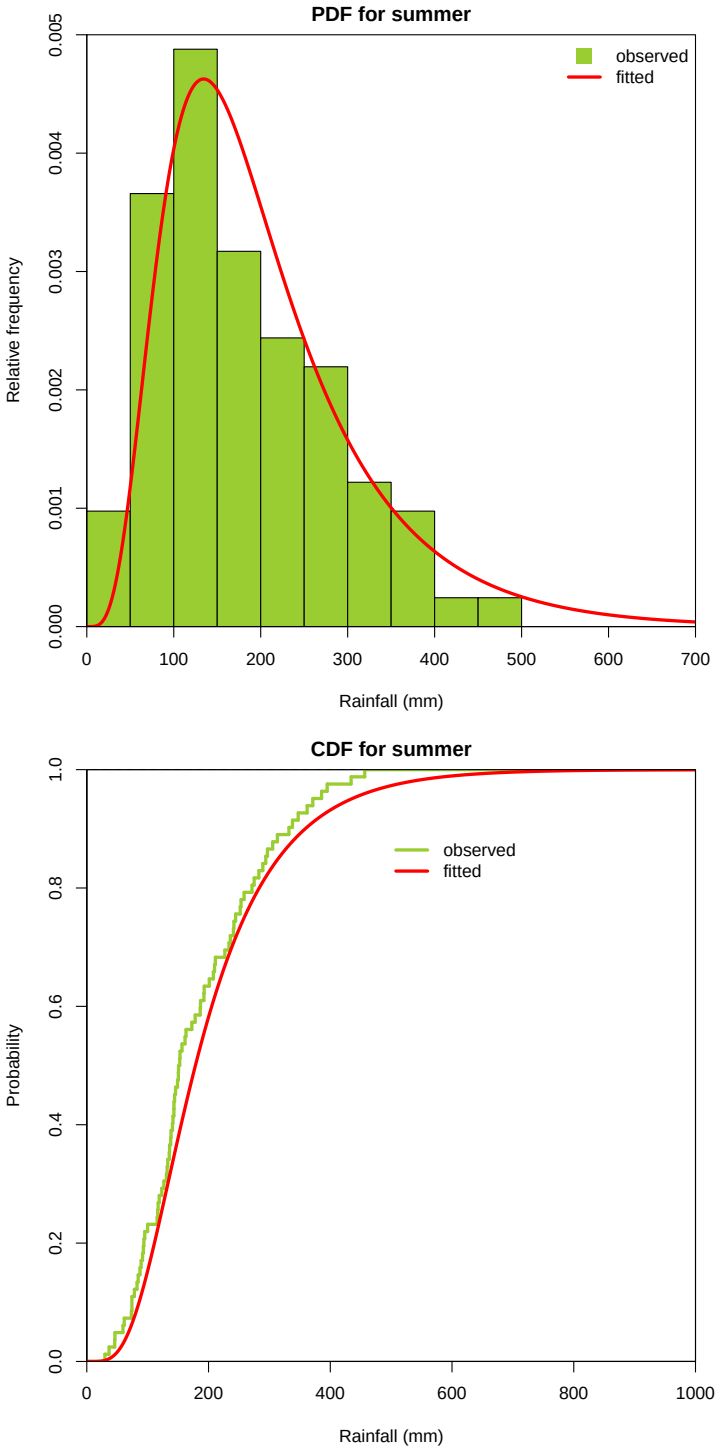



**Figure 6.10:** PDF and CDF for observed and fitted gamma sum for summer season

## 6.7 Summary

We have conducted a spatial and temporal analyses for monthly rainfall totals using the sum of $n$ correlated gamma variables. For the spatial analysis, we have fitted the sum of the rainfall totals from four stations (Hume, Beechworth, Dookie and Rutherglen) which can represents a local water catchment. The results of the spatial analysis can be extended to the study of the total rainfall-runoff for a certain rainfall region. In the temporal analysis, we consider Hume station and model the sum of rainfall totals for winter and summer seasons. From the illustration, we have shown that the goodness of fit for marginal analyses were good based on the Kolmogorov–Smirnov goodness of fit test. The Kolmogorov–Smirnov goodness of fit test also demonstrate a similar results for the sum of positive quadruples of correlated gamma marginals. In general, the sum of $n$ correlated gamma variables can easily be extended to more than four random variables. We also highlighted the need to check the eigenvalues of $C$ when applying the Alouini et al. [3] method and the potential for very slow convergence of the Alouini et al. [3] method when one or more eigenvalues of $A$ are very large by comparison with the smallest eigenvalue. In conclusion, the model of the sum of $n$ correlated gamma variables is suitable in modelling rainfall totals spatially and temporally for random variables which have significant correlation. In the next section we will describe another form of PDF for the sum of correlated gamma random variables using the McKay distribution.

# Chapter 7

# PDF for the sum of correlated gamma random variables using the McKay distribution

In this chapter we will show a derivation of a formula for the PDF of a sum of two correlated gamma random variables using the McKay distribution introduced by Holm and Alouini [45]. We will also verify that the proposed function satisfies the necessary conditions for a PDF. Using the McKay distribution, we derive some new moments formulae for the bivariate gamma distribution and introduce the weighted sum of two correlated gamma variables.

## 7.1 Bivariate McKay distribution

The bivariate McKay distribution is considered tractable in terms of moment generating function and cumulants [45]. There are two forms for the McKay distribution.

We will focus on the form used for the bivariate gamma distribution. Holm and Alouini [45] derive general formulae for the probability density function (PDF) of the sum and difference of two correlated gamma variables in the form of the bivariate McKay distribution. They define the sum form as type I McKay distribution and the difference form as type II McKay distribution. We consider type I McKay distribution for the sum of gamma variables. We will show that the type I McKay distribution is suitable for modelling rainfall totals.

Consider a set of $N$ independent gamma variables, $\{X_i\}_{i=1}^N$ with parameters $\alpha$ and $\beta_i$ written as $X_i \sim G(\alpha, \beta_i)$ where the PDF of $\{X_i\}_{i=1}^N$ is given by equation (6.1). The type I McKay distribution is first defined for the sum of independent gamma variables where $\Sigma = \sum_{i=1}^n X_i$ and the probability density function (PDF) is given by

$$f_\Sigma(\sigma) = \frac{\sqrt{\pi}}{\Gamma(a+\frac{1}{2})} \frac{(c^2-1)^{a+\frac{1}{2}}}{2^a b^{a+1}} \ \sigma^a \ \exp\left(-\frac{\sigma c}{b}\right) \ I_a\left(\frac{\sigma}{b}\right) \tag{7.1}$$

where $a, b, c$ are real parameters with $a > -1/2$, $b > 0$ and $c > 1$. The gamma function and modified Bessel function of the first kind of order $a$ are denoted by $\Gamma(\cdot)$ and $I_a(\cdot)$, respectively. The parameters $a, b$ and $c$ can be expressed in terms of the parameters $\alpha$ and $\beta$ for the sum of two independent gamma variables given in type I McKay distribution as

$$a = \alpha - \frac{1}{2}, \quad b = \frac{2\beta_1\beta_2}{|\beta_1 - \beta_2|} \quad \text{and} \quad c = \frac{\beta_1 + \beta_2}{|\beta_1 - \beta_2|}.$$

We refer the reader to [45] and [40] for further details. In the next section we derive the properties of a random variable $S = X_1 + X_2$ defined as the sum of independent gamma random variables.

## 7.2 Sum of two independent gamma variables

In this section we derive the formula of the PDF for the sum of two independent gamma variables for the case where $n = 2$. Let $X_i \sim G(\alpha, \beta_i)$ where $\alpha, \beta_i > 0$ for $i = 1, 2$. The shape parameter $\alpha$ is the same for each distribution but the scale parameters $(\beta_1, \beta_2)$ are different. Then the marginal PDF's are given by

$$f_i(x_i) = \frac{x_i^{\alpha-1} \exp(-\frac{x_i}{\beta_i})}{\Gamma(\alpha)\beta_i^{\alpha}} \quad \text{for } x_i > 0;\ \alpha, \beta_i > 0. \tag{7.2}$$

Consider $\Sigma = X_1 + X_2$ be the sum of two independent variables, then the PDF of the sum of two independent gamma variables denoted by $f_\Sigma(\sigma)$ is

$$f_\Sigma = \int_0^\sigma f_1(\sigma - x) f_2(x) dx \tag{7.3}$$

using the convolution. To obtain the expression of the PDF of $\sigma$ in terms of parameters $\alpha$ and $\beta_i$, substitute formula (7.2) into the integral in (7.3) and it becomes

$$\begin{aligned} f_\Sigma(\sigma) &= \int_0^\sigma \frac{(\sigma - x)^{\alpha-1} \ \exp\left[-\frac{(\sigma-x)}{\beta_1}\right]}{\Gamma(\alpha)\beta_1^{\alpha}} \ \frac{x^{\alpha-1} \ \exp(-\frac{x}{\beta_2})}{\Gamma(\alpha)\beta_2^{\alpha}} \ dx \\ &= \frac{\exp(-\frac{\sigma}{\beta_1})}{[\Gamma(\alpha)]^2(\beta_1\beta_2)^{\alpha}} \int_0^\sigma [x(\sigma - x)]^{\alpha-1} \ \exp\left[-x\left(\frac{1}{\beta_2} - \frac{1}{\beta_1}\right)\right] \ dx \\ &= \frac{\exp(-\frac{\sigma}{\beta_1})}{[\Gamma(\alpha)]^2(\beta_1\beta_2)^{\alpha}} \int_0^\sigma [x(\sigma - x)]^{\alpha-1} \ \exp\left[-x\left(\frac{\beta_1 - \beta_2}{\beta_1\beta_2}\right)\right] \ dx. \end{aligned}$$

We transform the PDF of $\sigma$ in the following way. We apply the standard formula

$$\int_0^{2u} (2ux - x^2)^{\nu-1} e^{-\mu x} dx = \sqrt{\pi} \left(\frac{2u}{\mu}\right)^{\nu-\frac{1}{2}} e^{\mu u} \Gamma(\nu) I_{\nu-\frac{1}{2}}(\mu u)$$

given as equation (3.383–2) on page 365 of the book by Gradshteyn and Ryzik [40] with $\sigma = 2u$, $\alpha = \nu$ and $\mu = \frac{\beta_1 - \beta_2}{\beta_1\beta_2}$ and where $I_{\nu-\frac{1}{2}}(\cdot)$ is the modified Bessel function

of the first kind of order $\nu - \frac{1}{2}$. It follows that

$$\begin{aligned} f_\Sigma(\sigma) &= \frac{\exp(-\frac{\sigma}{\beta_1})}{[\Gamma(\alpha)]^2(\beta_1\beta_2)^\alpha} \int_0^{2u} (2ux - x^2)^{\nu-1}\ e^{-x\mu}\ dx \\ &= \frac{\exp(-\frac{\sigma}{\beta_1})}{[\Gamma(\alpha)]^2(\beta_1\beta_2)^\alpha}\ \sqrt{\pi}\Big(\frac{2u}{\mu}\Big)^{\nu-\frac{1}{2}} e^{u\mu}\Gamma(\nu)\ I_{\nu-\frac{1}{2}}(u\mu). \end{aligned}$$

Now if we substitute back $2u = \sigma,\ \nu = \alpha$ and $\mu = \frac{\beta_1-\beta_2}{\beta_1\beta_2}$ and simplify, we obtain

$$\begin{aligned} &= \frac{\exp(-\frac{\sigma}{\beta_1})\sqrt{\pi}}{[\Gamma(\alpha)]^2(\beta_1\beta_2)^\alpha}\Big(\frac{\sigma}{\frac{\beta_1-\beta_2}{\beta_1\beta_2}}\Big)^{\alpha-\frac{1}{2}} \exp\left[-\frac{\sigma}{2}\left(\frac{\beta_1-\beta_2}{\beta_1\beta_2}\right)\right]\Gamma(\alpha)\ I_{\alpha-\frac{1}{2}}\Big(\frac{\sigma}{2}\frac{\beta_1-\beta_2}{\beta_1\beta_2}\Big) \\ &= \frac{\sqrt{\pi}}{\Gamma(\alpha)(\beta_1\beta_2)^\alpha}\ \Big(\frac{\sigma\beta_1\beta_2}{\beta_1-\beta_2}\Big)^{\alpha-\frac{1}{2}} \exp\left[-\sigma\left(\frac{\beta_1-\beta_2}{2\beta_1\beta_2}+\frac{1}{\beta_1}\right)\right] I_{\alpha-\frac{1}{2}}\Big(\sigma\frac{\beta_1-\beta_2}{2\beta_1\beta_2}\Big) \\ &= \frac{\sqrt{\pi}}{\Gamma(\alpha)(\beta_1\beta_2)^\alpha}\ \Big(\frac{\sigma\beta_1\beta_2}{\beta_1-\beta_2}\Big)^{\alpha-\frac{1}{2}} \exp\left[-\sigma\left(\frac{\beta_1+\beta_2}{2\beta_1\beta_2}\right)\right]\ I_{\alpha-\frac{1}{2}}\ \Big(\sigma\frac{\beta_1-\beta_2}{2\beta_1\beta_2}\Big). \end{aligned}$$

Substitute $a = \alpha - \frac{1}{2},\ b = \frac{2\beta_1\beta_2}{|\beta_1-\beta_2|}$ and $c = \frac{\beta_1+\beta_2}{|\beta_1-\beta_2|}$ and simplify. We obtain

$$\begin{aligned} f_\Sigma(\sigma) &= \frac{\sqrt{\pi}}{\Gamma(a+\frac{1}{2})(\beta_1\beta_2)^{a+\frac{1}{2}}}\ \Big(\sigma\frac{\beta_1\beta_2}{\beta_1-\beta_2}\Big)^a \exp\Big(-\frac{\sigma c}{b}\Big)\ I_a\Big(\frac{\sigma}{b}\Big) \\ &= \frac{\sqrt{\pi}}{\Gamma(a+\frac{1}{2})}\frac{1}{(\beta_1\beta_2)^{\frac{1}{2}}}\ \Big(\frac{1}{\beta_1-\beta_2}\Big)^a\ \sigma^a\ \exp\Big(-\frac{\sigma c}{b}\Big)\ I_a\Big(\frac{\sigma}{b}\Big) \\ &= \frac{\sqrt{\pi}}{\Gamma(a+\frac{1}{2})}\frac{(c^2-1)^{a+\frac{1}{2}}}{2^a b^{a+1}}\ \sigma^a\ \exp\Big(-\frac{\sigma c}{b}\Big)\ I_a\Big(\frac{\sigma}{b}\Big) \end{aligned}$$

where

$$\begin{aligned}
\frac{(c^2-1)^{a+\frac{1}{2}}}{2^a b^{a+1}} &= \frac{\left[\left(\frac{\beta_1+\beta_2}{\beta_1-\beta_2}\right)^2-1\right]^{a+\frac{1}{2}}}{2^a\left(\frac{2\beta_1\beta_2}{\beta_1-\beta_2}\right)^{a+1}} \\
&= \frac{\left[(\beta_1+\beta_2)^2-(\beta_1-\beta_2)^2\right]^{a+\frac{1}{2}}}{(\beta_1-\beta_2)^{2a+1}}\ \frac{(\beta_1-\beta_2)^{a+1}}{2^a(2\beta_1\beta_2)^{a+1}} \\
&= \frac{\left[4\beta_1\beta_2)\right]^{a+\frac{1}{2}}}{(\beta_1-\beta_2)^{2a+1}}\ \frac{(\beta_1-\beta_2)^{a+1}}{2^a(2\beta_1\beta_2)^{a+1}} \\
&= \frac{2^{2a+1}(\beta_1\beta_2)^{a+\frac{1}{2}}}{2^{2a+1}(\beta_1\beta_2)^{a+1}(\beta_1-\beta_2)^a} \\
&= \frac{1}{(\beta_1\beta_2)^{\frac{1}{2}}}\frac{1}{(\beta_1-\beta_2)^a}.
\end{aligned}$$

Therefore, the PDF of the sum of two independent gamma variables is in the form of a type I McKay distribution as described in equation (7.1).

## 7.3 The bivariate McKay distribution for correlated gamma random variables

If $X_1$ and $X_2$ are gamma variables which are correlated, the joint distribution of two correlated gamma variables is called a bivariate gamma distribution. The joint PDF of the bivariate gamma distribution of $X_1$ and $X_2$ with common shape $\alpha$ and respective scales $\beta_1$ and $\beta_2$, is denoted by Holm and Alouini [45] as

$$X_1, X_2 \sim G(\alpha, \beta_1, \beta_2, \rho)$$

where

$$f_{X_1X_2}(x_1,x_2) = \frac{(x_1x_2)^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}} \exp\Big(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\Big) \times I_{\alpha-1}\Big(\frac{2\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\Big) \tag{7.4}$$

for $x > 0$ where

$$\begin{aligned} \alpha &> 0 \quad \text{both have same shape parameters} \\ \beta_1 &> 0,\ \beta_2 > 0 \quad \text{distinct scale parameters} \\ \rho &= \frac{Cov(X_1,X_2)}{\sqrt{Var(X_1)Var(X_2)}},\ 0 \le \rho \le 1 \\ I_{\alpha-1}(.) &- \text{ modified Bessel function of the first kind of order } (\alpha-1). \end{aligned}$$

Note that the monthly rainfall totals we have used are not normally distributed and hence we will use the Spearman's rank correlation for the correlation parameter $\rho$. For additional comments refer to Section 3.5. In the following section we show that the proposed function is a legitimate PDF.

### 7.3.1 Justification of the PDF formula

By integration, we can show that the integral of joint PDF is equal to one using a special transformation. Let $X_1$ and $X_2$ be random variables such that $X_1 \sim G(\alpha,\beta_1)$ and $X_2 \sim G(\alpha,\beta_2)$, $x_1, x_2 > 0$. The proposed joint density $f(x_1,x_2;\alpha,\beta_1,\beta_2,\rho)$ for

a joint gamma distribution given by Holm and Alouini [45] is

$$\begin{aligned} f(x_1,x_2) &= \frac{(x_1x_2)^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}}\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right) \\ &\quad \times I_{\alpha-1}\left(\frac{2\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right) \\ &= \frac{(x_1x_2)^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}}\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right) \\ &\quad \times \sum_{k=0}^{\infty}\frac{1}{k!\Gamma(\alpha+k)}\left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)^{\alpha+2k-1} \end{aligned}$$

on the region $[0,\infty)^2 = [0,\infty)\times[0,\infty)$. The modified Bessel function of the first kind used in the above formula is defined by the power series

$$I_\nu(x) = \sum_{k=0}^{\infty}\frac{\left(\frac{x}{2}\right)^{\nu+2k}}{k!\Gamma(\nu+k+1)}$$

where $\nu = \alpha-1$ and $x = \frac{2\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}$. We need to show that

$$\iint_{[0,\infty)^2} f_{X_1X_2}(x_1,x_2)\ dx_1dx_2 = 1.$$

Consider the integral

$$\begin{aligned} I &= \iint_{[0,\infty)^2} f_{X_1X_2}(x_1,x_2)\ dx_1dx_2 \\ &= \iint_{[0,\infty)^2}\frac{(x_1x_2)^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}}\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right) \\ &\quad \times\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(\alpha+k)}\left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)^{\alpha+2k-1}\ dx_1dx_2. \end{aligned}$$

If we make the transformation $x_j = \beta_j z_j$ and $\frac{dx_j}{dz_j} = \beta_j$ for each $j = 1, 2$ also on $[0, \infty)^2$, we obtain

$$\begin{aligned}
I &= \iint_{[0,\infty)^2} \frac{(z_1 z_2)^{\frac{\alpha-1}{2}} (\beta_1\beta_2)^{\frac{\alpha-1}{2}} \beta_1\beta_2}{\Gamma(\alpha)(1-\rho)\rho^{\frac{\alpha-1}{2}} (\beta_1\beta_2)^{\frac{\alpha+1}{2}}} \exp\left(-\frac{z_1+z_2}{1-\rho}\right) \\
&\quad \times \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)} \left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{z_1 z_2}\right)^{\alpha+2k-1} dz_1 dz_2 \\
&= \iint_{[0,\infty)^2} \frac{(z_1 z_2)^{\alpha-1+k}}{\Gamma(\alpha)(1-\rho)^{\alpha}} \exp\left(-\frac{z_1+z_2}{1-\rho}\right) \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \, dz_1 dz_2 \\
&= \frac{1}{\Gamma(\alpha)(1-\rho)^{\alpha}} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \\
&= \qquad \times \iint_{[0,\infty)^2} (z_1 z_2)^{\alpha-1+k} \exp\left(-\frac{z_1+z_2}{1-\rho}\right) \, dz_1 dz_2.
\end{aligned}$$

The transformation $z_j = (1-\rho)\zeta_j$ for each $j = 1, 2$ on the right hand side gives

$$\begin{aligned}
I &= \frac{(1-\rho)^{\alpha}}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\,\Gamma(\alpha+k)} \int_{[0,\infty)} \zeta_2^{\alpha+k-1} e^{-\zeta_2} \left[\int_{[0,\infty)} \zeta_1^{\alpha+k-1} e^{-\zeta_1} d\zeta_1\right] d\zeta_2 \\
&= (1-\rho)^{\alpha} \sum_{k=0}^{\infty} \frac{(\alpha+k-1)(\alpha+k-2)\cdots(\alpha+1)\alpha}{k!} \rho^k \\
&= (1-\rho)^{\alpha}(1-\rho)^{-\alpha} \\
&= 1
\end{aligned}$$

by using the binomial expansion for $(1-\rho)^{-\alpha}$. Thus the proposed function $f_{X_1,X_2}(x_1, x_2)$ is a suitable probability density.

## 7.4 A PDF for the sum of two correlated gamma random variables with the same shape

The type I McKay distribution is also suitable to be used to define the sum of two correlated gamma variables [45]. The full derivation is not shown in [45]. Because

the derivation is quite complicated we have chosen to present it here in order to convey some additional appreciation of the necessary analysis.

Let $X_1, X_2 \sim G(\alpha, \beta_1, \beta_2, \rho)$ and consider the sum $\Sigma = X_1 + X_2$ follows type I McKay distribution as in equation (7.1) for $c > 1$, $b > 0$ and $0 \leq \rho < 1$. The PDF of the sum of two correlated gamma variables is given by

$$f_\Sigma(\sigma) = \frac{\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha \Gamma(\alpha)} \Big(\frac{\sigma}{2\widetilde{\beta}}\Big)^{\alpha-\frac{1}{2}} \; e^{-\frac{\sigma(\beta_1+\beta_2)}{2\beta_1\beta_2(1-\rho)}} \; I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta}) \tag{7.5}$$

with

$$\begin{aligned} a &= \alpha - \frac{1}{2} \\ b &= \frac{2\beta_1\beta_2(1-\rho)}{\sqrt{(\beta_1+\beta_2)^2 - 4\beta_1\beta_2(1-\rho)}} = \frac{2\beta_1\beta_2(1-\rho)}{\sqrt{(\beta_1-\beta_2)^2 + 4\beta_1\beta_2\rho}} \\ c &= \frac{\beta_1+\beta_2}{\sqrt{(\beta_1+\beta_2)^2 - 4\beta_1\beta_2(1-\rho)}} = \frac{\beta_1+\beta_2}{\sqrt{(\beta_1-\beta_2)^2 + 4\beta_1\beta_2\rho}} \end{aligned}$$

where $\frac{c}{b} = \frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)}$ and $\widetilde{\beta}^2 = \frac{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}{4\beta_1^2\beta_2^2(1-\rho)^2}$ then $\widetilde{\beta} = \frac{\sqrt{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}}{2\beta_1\beta_2(1-\rho)} = \frac{1}{b}$. The term $\frac{(c^2-1)^\alpha}{2^{\alpha-\frac{1}{2}}b^{\alpha+\frac{1}{2}}} \; \sigma^{\alpha-\frac{1}{2}}$ can be expressed in terms of the parameters $\alpha$, $\beta_i$ and $\rho$ as

follows:

$$\begin{aligned}
\frac{(c^2-1)^\alpha}{2^{\alpha-\frac{1}{2}}b^{\alpha+\frac{1}{2}}}\ \sigma^{\alpha-\frac{1}{2}} &= \frac{\left[\left(\frac{\beta_1+\beta_2}{\sqrt{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}}\right)^2-1\right]^\alpha}{2^{\alpha-\frac{1}{2}}\left[\frac{2\beta_1\beta_2(1-\rho)}{\sqrt{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}}\right]^{\alpha+\frac{1}{2}}}\ \sigma^{\alpha-\frac{1}{2}} \\
&= \left(\frac{\sigma}{2}\right)^{\alpha-\frac{1}{2}}\left[\frac{\beta_1^2+2\beta_1\beta_2+\beta_2^2-[(\beta_1-\beta_2)^2+4\beta_1\beta_2]}{(\beta_1-\beta_2)^2+4\beta_1\beta_2}\right]^\alpha \\
&\qquad \times\left[\frac{\sqrt{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}}{2\beta_1\beta_2(1-\rho)}\right]^{\alpha+\frac{1}{2}} \\
&= \left(\frac{\sigma}{2}\right)^{\alpha-\frac{1}{2}}\left[\frac{\beta_1^2+2\beta_1\beta_2+\beta_2^2-[\beta_1^2-2\beta_1\beta_2+\beta_2^2+4\beta_1\beta_2]}{(\beta_1-\beta_2)^2+4\beta_1\beta_2}\right]^\alpha \\
&\qquad \times\widetilde{\beta}^{\alpha+\frac{1}{2}} \\
&= \left(\frac{\sigma}{2}\right)^{\alpha-\frac{1}{2}}\left[\frac{1}{\widetilde{\beta}^2}\ \frac{1}{\beta_1\beta_2(1-\rho)}\right]^\alpha\ \widetilde{\beta}^{\alpha+\frac{1}{2}} \\
&= \left(\frac{\sigma}{2}\right)^{\alpha-\frac{1}{2}}\ \widetilde{\beta}^{-2\alpha}\ \left[\frac{1}{\beta_1\beta_2(1-\rho)}\right]^\alpha\ \widetilde{\beta}^{\alpha+\frac{1}{2}} \\
&= \left(\frac{\sigma}{2\widetilde{\beta}}\right)^{\alpha-\frac{1}{2}}\ \left[\frac{1}{\beta_1\beta_2(1-\rho)}\right]^\alpha.
\end{aligned}$$

Hence, the PDF of $\sigma$ is written as

$$\begin{aligned}
f_\Sigma(\sigma) &= \frac{\sqrt{\pi}}{\Gamma(\alpha)}\frac{(c^2-1)^\alpha}{2^{\alpha-\frac{1}{2}}b^{\alpha+\frac{1}{2}}}\ \sigma^{\alpha-\frac{1}{2}}\ \exp\left[-\frac{\sigma(\beta_1+\beta_2)}{2\beta_1\beta_2(1-\rho)}\right]\ I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta}) \\
&= \frac{\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{\sigma}{2\widetilde{\beta}}\right)^{\alpha-\frac{1}{2}}\ \left[\frac{1}{\beta_1\beta_2(1-\rho)}\right]^\alpha\ e^{-\frac{\sigma(\beta_1+\beta_2)}{2\beta_1\beta_2(1-\rho)}}\ I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta}).
\end{aligned}$$

Now, if $R=\sqrt{\Sigma}=\sqrt{X_1+X_2}$, then $R^2=\Sigma=X_1+X_2$ and $\frac{\partial\Sigma}{\partial R}=2R$. The PDF of $r$ is given by

$$\begin{aligned}
f_R(r) &= \frac{\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{r^2}{2\widetilde{\beta}}\right)^{\alpha-\frac{1}{2}}\ \left[\frac{1}{\beta_1\beta_2(1-\rho)}\right]^\alpha\ e^{-r^2(\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)})}\ I_{\alpha-\frac{1}{2}}(r^2\widetilde{\beta})\ \frac{\partial\sigma}{\partial r} \\
&= \frac{\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha\Gamma(\alpha)}\left(\frac{r^2}{2\widetilde{\beta}}\right)^{\alpha-\frac{1}{2}}\ e^{-r^2(\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)})}\ I_{\alpha-\frac{1}{2}}(r^2\widetilde{\beta})\ 2r \\
&= \frac{2r\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha\Gamma(\alpha)}\left(\frac{r^2}{2\widetilde{\beta}}\right)^{\alpha-\frac{1}{2}}\ e^{-r^2(\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)})}\ I_{\alpha-\frac{1}{2}}(r^2\widetilde{\beta}).
\end{aligned}$$

We can also show that $f_R(r)$ can be written as $f_\Sigma(\sigma)$ using a standard transformation of variables as shown below. Let $R = \sqrt{X_1 + X_2}$, then

$$\begin{aligned} f_R(r) &= \frac{2r\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha\Gamma(\alpha)}\Big(\frac{r^2}{2\widetilde{\beta}}\Big)^{\alpha-\frac{1}{2}} \exp\left[-r^2\,\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)}\right] \\ &\quad \times I_{\alpha-\frac{1}{2}}(r^2\widetilde{\beta}) \end{aligned}$$

where $\widetilde{\beta}^2 = \frac{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}{4\beta_1^2\beta_2^2(1-\rho)^2}$. Now, if $R^2 = \Sigma = X_1+X_2$, then $R = \sqrt{\Sigma}$ and $\frac{\partial R}{\partial \Sigma} = \frac{1}{2\sqrt{\sigma}}$. Therefore, the PDF of $\sigma$ is

$$\begin{aligned} f_\Sigma(\sigma) &= \frac{2\sqrt{\sigma}\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha\Gamma(\alpha)}\Big(\frac{\sigma}{2\widetilde{\beta}}\Big)^{\alpha-\frac{1}{2}} e^{-\sigma(\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)})}\, I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta})\,\frac{\partial r}{\partial\sigma} \\ &= \frac{2\sqrt{\sigma}\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha\Gamma(\alpha)}\Big(\frac{\sigma}{2\widetilde{\beta}}\Big)^{\alpha-\frac{1}{2}} e^{-\sigma(\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)})}\, I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta})\,\frac{1}{2\sqrt{\sigma}} \\ &= \frac{\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha\Gamma(\alpha)}\Big(\frac{\sigma}{2\widetilde{\beta}}\Big)^{\alpha-\frac{1}{2}} e^{-\sigma(\frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)})}\, I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta}). \end{aligned}$$

So far, we have seen that both the PDFs of the sum of two independent gamma variables and sum of two correlated gamma variables can be expressed in the form of a type 1 McKay distribution. Note that the latter expression reduces to the former when $\rho = 0$. In the next section, we want to show that the marginal distributions for the bivariate gamma distribution are indeed gamma distributions in the form $X_i \sim G(\alpha, \beta_i)$ for each $i = 1, 2$.

## 7.5 The marginal probability density function for a bivariate gamma distribution

In this section, we show that the marginal distribution of the joint PDF is the univariate gamma distribution. To obtain the marginal distribution, integrate formula (7.4) with respect to a selected random variable. The Bessel function in the

formula (7.4) is expressed in the series form for easy computation introduced by [40]. Let the marginal probability density function for a bivariate gamma distribution given by

$$f_{X_1}(x_1) = \int_{[0,\infty)} f_{X_1X_2}(x_1, x_2)\ dx_2$$

for all $x_1 \in [0, \infty)$ where $f_{X_1X_2}(x_1, x_2)$ is the joint PDF of the bivariate gamma distribution of $X_1$ and $X_2$ with common shape $\alpha$ and respective scales $\beta_1$ and $\beta_2$ is given in (7.4). Hence the marginal probability density function for a bivariate gamma distribution is

$$\begin{aligned} f_{X_1}(x_1) &= \frac{x_1^{\alpha-1}}{\Gamma(\alpha)(1-\rho)^{\alpha}(\beta_1\beta_2)^{\alpha}} \exp\left(-\frac{\frac{x_1}{\beta_1}}{1-\rho}\right) \sum_{k=0}^{\infty} \frac{\left(\frac{x_1\rho}{\beta_1\beta_2}\right)^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \\ &\quad \times \int_{[0,\infty)} x_2^{\alpha+k-1} \exp\left(-\frac{\frac{x_2}{\beta_2}}{1-\rho}\right)\ dx_2 \end{aligned}$$

with the modified Bessel function of the first kind is expressed in series form of

$$I_{\alpha-1}\left(\frac{2\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)} \left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)^{\alpha+2k-1}$$

and if we make the transformation $x_2 = z_2\beta_2,\ dx_2 = \beta_2 dz_2$ on the right hand side this becomes

$$\begin{aligned} f_{X_1}(x_1) &= \frac{x_1^{\alpha-1}}{\Gamma(\alpha)(1-\rho)^{\alpha}\beta_1^{\alpha}} \exp\left(-\frac{\frac{x_1}{\beta_1}}{1-\rho}\right) \sum_{k=0}^{\infty} \frac{x_1^k\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}\beta_1^k} \\ &\quad \times \int_{[0,\infty)} z_2^{\alpha+k-1} \exp\left(-\frac{z_2}{1-\rho}\right)\ dz_2 \end{aligned}$$

and if we make the transformation $z_2 = (1-\rho)\zeta_2,\ dz_2 = (1-\rho)d\zeta_2$ on the right hand side this becomes

$$\begin{aligned} f_{X_1}(x_1) &= \frac{x_1^{\alpha-1}}{\Gamma(\alpha)\beta_1^{\alpha}} \exp\left(-\frac{\frac{x_1}{\beta_1}}{1-\rho}\right) \sum_{k=0}^{\infty} \frac{x_1^k \rho^k}{k!\Gamma(\alpha+k)(1-\rho)^k \beta_1^k} \\ &\qquad \times \int_{[0,\infty)} \zeta_2^{\alpha+k-1} \exp(-\zeta_2)\ d\zeta_2 \\ &= \frac{x_1^{\alpha-1}}{\Gamma(\alpha)\beta_1^{\alpha}} \exp\left(-\frac{x_1}{\beta_1}\right)(1-\rho) \sum_{k=0}^{\infty} \frac{x_1^k \rho^k}{k!(1-\rho)^k} \\ &= \frac{x_1^{\alpha-1}}{\Gamma(\alpha)\beta_1^{\alpha}} \exp\left(-\frac{\frac{x_1}{\beta_1}}{1-\rho}\right) \exp\left(\frac{x_1\rho}{(1-\rho)\beta_1}\right) \\ &= \frac{x_1^{\alpha-1}}{\Gamma(\alpha)\beta_1^{\alpha}} \exp\left(-\frac{x_1}{\beta_1}\right). \end{aligned}$$

Therefore $X_1 \sim G(\alpha, \beta_1)$. A similar argument shows that $X_2 \sim G(\alpha, \beta_2)$.

## 7.6 Moments of the bivariate gamma distribution

In this section, we will show how to derive expressions for various moments of bivariate gamma distribution. The mean of $X_1$ is given by

$$\begin{aligned} E[X_1] &= \int\int_{(0,\infty)^2} x_1 f_{X_1,X_2}(x_1,x_2)\ dx_1 dx_2 \\ &= \int_{[0,\infty)} x_1 f_{X_1}(x_1)\ dx_1 \\ &= \int_{[0,\infty)} x_1 \frac{x_1^{\alpha-1} e^{\frac{-x_1}{\beta_1}}}{\beta_1^{\alpha}\Gamma(\alpha)}\ dx_1 \\ &= \frac{\beta_1^{\alpha+1}\Gamma(\alpha+1)}{\beta_1^{\alpha}\Gamma(\alpha)} \int_{[0,\infty)} \frac{x_1^{(\alpha+1)-1} e^{\frac{-x_1}{\beta_1}}}{\beta_1^{\alpha+1}\Gamma(\alpha+1)}\ dx_1 \\ &= \alpha\beta_1 \end{aligned}$$

and similarly the mean of $X_2$ is given by

$$E[X_2] = \alpha\beta_2.$$

The expression for the second moments are derived as follows:

$$\begin{aligned}
E[X_1^2] &= \int\int_{(0,\infty)^2} x_1^2 f_{X_1,X_2}(x_1,x_2)\ dx_1 dx_2 \\
&= \int_{[0,\infty)} x_1^2\ f_{X_1}(x_1)\ dx_1 \\
&= \int_{[0,\infty)} x_1^2 \frac{x_1^{\alpha-1} e^{\frac{-x_1}{\beta_1}}}{\beta_1^\alpha \Gamma(\alpha)}\ dx_1 \\
&= \frac{\beta_1^{\alpha+2}\Gamma(\alpha+2)}{\beta_1^\alpha\Gamma(\alpha)} \int_{[0,\infty)} \frac{x_1^{(\alpha+2)-1} e^{\frac{-x_1}{\beta_1}}}{\beta_1^{\alpha+2}\Gamma(\alpha+2)}\ dx_1 \\
&= \alpha(\alpha+1)\beta_1^2
\end{aligned}$$

and similarly for

$$E[X_2^2] = \alpha(\alpha+1)\beta_2^2.$$

The moment for the product is

$$\begin{aligned}
E[X_1X_2] &= \int\int_{(0,\infty)^2} x_1x_2 f_{X_1,X_2}(x_1,x_2)\ dx_1dx_2 \\
&= \frac{1}{\Gamma(\alpha)(1-\rho)^\alpha(\beta_1\beta_2)^\alpha} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}(\beta_1\beta_2)^k} \\
&\quad \times \iint_{[0,\infty)^2} (x_1x_2)^{\alpha+k}\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right)\ dx_1dx_2
\end{aligned}$$

where the modified Bessel function of the first kind is expressed in series form of

$$I_{\alpha-1}\left(\frac{2\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)} \left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)^{\alpha+2k-1}.$$

If we make the transformation $x_j = \beta_j z_j$ and $\frac{dx_j}{dz_j} = \beta_j$ for each $j = 1, 2$ also on $[0, \infty)^2$, we obtain

$$\begin{aligned} E[X_1 X_2] &= \frac{\beta_1 \beta_2}{\Gamma(\alpha)(1-\rho)^\alpha} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \\ &\quad \times \iint_{[0,\infty)^2} (z_1 z_2)^{\alpha+k} \exp\left(-\frac{z_1+z_2}{1-\rho}\right) dz_1 dz_2 \end{aligned}$$

and if we make the transformation $z_j = (1-\rho)\zeta_j$ and $dz_j = (1-\rho)d\zeta_j$ for each $j = 1, 2$ on the right hand side this becomes

$$\begin{aligned} E[X_1 X_2] &= \frac{\beta_1 \beta_2 (1-\rho)^{\alpha+2}}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)} \\ &\quad \times \iint_{[0,\infty)^2} (\zeta_1 \zeta_2)^{(\alpha+k+1)-1} \exp(-\zeta_1 - \zeta_2)\, d\zeta_1 d\zeta_2 \\ &= \beta_1 \beta_2 (1-\rho)^{\alpha+2} \frac{1}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{\rho^k [\Gamma(\alpha+k+1)]^2}{k!\Gamma(\alpha+k)} \end{aligned}$$

where the summation can be simplified as

$$\begin{aligned} &\frac{1}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{[\Gamma(\alpha+k+1)]^2\ \rho^k}{k!\Gamma(\alpha+k)} \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{[(\alpha+k)(\alpha+k-1)\cdots\alpha\Gamma(\alpha)]^2\ \rho^k}{k!(\alpha+k-1)\cdots\alpha\Gamma(\alpha)} \\ &= \sum_{k=0}^{\infty} \frac{(\alpha+k)[(\alpha+k)(\alpha+k-1)\cdots\alpha]\ \rho^k}{k!} \\ &= \alpha \sum_{k=0}^{\infty} \frac{(\alpha+k)(\alpha+k-1)\cdots\alpha\ \rho^k}{k!} + \sum_{k=1}^{\infty} \frac{(\alpha+k)(\alpha+k-1)\cdots\alpha\ \rho^k}{(k-1)!} \\ &= \alpha^2 \left[1 + \frac{\rho(\alpha+1)}{1!} + \frac{\rho^2(\alpha+1)(\alpha+2)}{2!} + \frac{\rho^3(\alpha+1)(\alpha+2)(\alpha+3)}{3!} + \cdots\right] \\ &\quad + \alpha(\alpha+1)\rho \left[1 + \frac{\rho(\alpha+2)}{1!} + \frac{\rho^2(\alpha+2)(\alpha+3)}{2!} + \cdots\right] \\ &= \alpha^2(1-\rho)^{-(\alpha+1)} + \alpha(\alpha+1)\rho(1-\rho)^{-(\alpha+2)}. \end{aligned}$$

Therefore,

$$\begin{aligned} E[X_1X_2] &= \beta_1\beta_2(1-\rho)^{\alpha+2}\left[\frac{\alpha^2}{(1-\rho)^{\alpha+1}}+\frac{\alpha(\alpha+1)\rho}{(1-\rho)^{\alpha+2}}\right] \\ &= \alpha\beta_1\beta_2(\alpha+\rho). \end{aligned}$$

We use a similar technique to obtain a general formula of the moments $E[X_1^{r_1}X_2^{r_2}]$ for all $r_1, r_2 \geq 0$.

$$\begin{aligned} & E[X_1^{r_1}X_2^{r_2}] \\ = & \iint_{[0,\infty)^2} x_1^{r_1}x_2^{r_2} f_{X_1X_2}(x_1,x_2)\, dx_1dx_2 \\ = & \iint_{[0,\infty)^2} x_1^{r_1}x_2^{r_2}\frac{(x_1x_2)^{\alpha-1}}{\Gamma(\alpha)(1-\rho)^{\alpha}(\beta_1\beta_2)^{\alpha}}\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right) \\ & \times\sum_{k=0}^{\infty}\frac{(x_1x_2)^k\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}(\beta_1\beta_2)^k}\, dx_1dx_2 \\ = & \frac{1}{\Gamma(\alpha)(1-\rho)^{\alpha}(\beta_1\beta_2)^{\alpha}}\sum_{k=0}^{\infty}\frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}(\beta_1\beta_2)^k} \\ & \times\iint_{[0,\infty)^2} x_1^{r_1}x_2^{r_2}(x_1x_2)^{\alpha-1}(x_1x_2)^k\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right)\ \ dx_1dx_2 \end{aligned}$$

where $f_{X_1X_2}(x_1,x_2)$ is the PDF of the bivariate McKay distribution for correlated gamma variables as in formula (7.4). For easy computation, the modified Bessel function is expressed in series form given by

$$I_{\alpha-1}\left(\frac{2\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)=\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(\alpha+k)}\left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)^{\alpha+2k-1}$$

If we make the transformation $x_j = \beta_j z_j$ and $dx_j = \beta_j dz_j$ for each $j = 1, 2$ on the right hand side over $[0, \infty)^2$, this becomes

$$\begin{aligned}
& E[X_1^{r_1} X_2^{r_2}] \\
= \; & \frac{1}{\Gamma(\alpha)(1-\rho)^\alpha (\beta_1\beta_2)^\alpha} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}(\beta_1\beta_2)^k} \\
& \times \int_{[0,\infty)} (\beta_1 z_1)^{(r_1+k+\alpha)-1} \exp\left(-\frac{z_1}{1-\rho}\right) \beta_1 dz_1 \\
& \times \int_{[0,\infty)} (\beta_2 z_2)^{(r_2+k+\alpha)-1} \exp\left(-\frac{z_2}{1-\rho}\right) \beta_2 dz_2 \\
= \; & \frac{\beta_1^{r_1}\beta_2^{r_2}}{\Gamma(\alpha)(1-\rho)^\alpha} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \\
& \times \int_{[0,\infty)} z_1^{(r_1+k+\alpha)-1} \exp\left(-\frac{z_1}{1-\rho}\right) dz_1 \\
& \times \int_{[0,\infty)} z_2^{(r_2+k+\alpha)-1} \exp\left(-\frac{z_2}{1-\rho}\right) dz_2
\end{aligned}$$

and if we make the transformation $z_j = (1-\rho)\zeta_j$ and $dz_j = (1-\rho)d\zeta_j$ for each $j = 1, 2$ on the right hand side this becomes

$$\begin{aligned}
& E[X_1^{r_1} X_2^{r_2}] \\
= \; & \frac{\beta_1^{r_1}\beta_2^{r_2}}{\Gamma(\alpha)(1-\rho)^\alpha} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \\
& \times \int_{[0,\infty)} z_1^{(r_1+k+\alpha)-1} \exp\left(-\frac{z_1}{1-\rho}\right) dz_1 \\
& \times \int_{[0,\infty)} z_2^{(r_2+k+\alpha)-1} \exp\left(-\frac{z_2}{1-\rho}\right) dz_2
\end{aligned}$$

and if we make the transformation $z_j = (1-\rho)\zeta_j$ and $\frac{dz_j}{d\zeta_j} = (1-\rho)$ for each $j = 1, 2$ on the right hand side this becomes

$$
\begin{aligned}
& E[X_1^{r_1} X_2^{r_2}] \\
= \; & \frac{\beta_1^{r_1}\beta_2^{r_2}}{\Gamma(\alpha)(1-\rho)^{\alpha}} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)(1-\rho)^{2k}} \\
& \times \int_{[0,\infty)} [(1-\rho)\zeta_1]^{(r_1+k+\alpha)-1} \exp(-\zeta_1)\ (1-\rho)d\zeta_1 \\
& \times \int_{[0,\infty)} [(1-\rho)d\zeta_2]^{(r_2+k+\alpha)-1}\ \exp(-\zeta_2)\ (1-\rho)d\zeta_2 \\
= \; & \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)} \\
& \times \int_{[0,\infty)} \zeta_1^{(r_1+k+\alpha)-1} \exp(-\zeta_1) \left[ \int_{[0,\infty)} \zeta_2^{(r_2+k+\alpha)-1}\ \exp(-\zeta_2)\ d\zeta_2 \right]\ d\zeta_1 \\
= \; & \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{\rho^k}{k!\Gamma(\alpha+k)}\ \Gamma(r_1+k+\alpha)\Gamma(r_2+k+\alpha) \\
& \times \int_{[0,\infty)} \frac{\zeta_1^{(r_1+k+\alpha)-1}}{\Gamma(r_1+k+\alpha)} \exp(-\zeta_1) \left[ \int_{[0,\infty)} \frac{\zeta_2^{(r_2+k+\alpha)-1}}{\Gamma(r_2+k+\alpha)}\ \exp(-\zeta_2)\ d\zeta_2 \right]\ d\zeta_1 \\
= \; & \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{\rho^k\ \Gamma(r_1+k+\alpha)\Gamma(r_2+k+\alpha)}{k!\Gamma(\alpha+k)}.
\end{aligned}
$$

Therefore, the required moments of correlated gamma variables is

$$
\begin{aligned}
&E[X_1^{r_1}X_2^{r_2}] \\
&\quad = \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}}{\Gamma(\alpha)}\sum_{k=0}^{\infty}\frac{\rho^k\ \Gamma(r_1+k+\alpha)\Gamma(r_2+k+\alpha)}{k!\Gamma(\alpha+k)} \\
&\quad = \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}}{\Gamma(\alpha)} \\
&\qquad \times\left[\frac{\Gamma(r_1+\alpha)\Gamma(r_2+\alpha)}{\Gamma(\alpha)}\right. \\
&\qquad\quad +\frac{\rho(r_1+\alpha)(r_2+\alpha)\Gamma(r_1+\alpha)\Gamma(r_2+\alpha)}{1!\alpha\Gamma(\alpha)} \\
&\qquad\quad \left.+\frac{\rho^2(r_1+\alpha)(r_1+\alpha+1)(r_2+\alpha)(r_2+\alpha+1)\Gamma(r_1+\alpha)\Gamma(r_2+\alpha)}{2!\alpha(\alpha+1)\Gamma(\alpha)}+\cdots\right] \\
&\quad = \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}}{\Gamma(\alpha)}\ \frac{\Gamma(r_1+\alpha)\Gamma(r_2+\alpha)}{\Gamma(\alpha)} \\
&\qquad \times\left[1+\frac{\rho(r_1+\alpha)(r_2+\alpha)}{1!\alpha}\right. \\
&\qquad\quad \left.+\frac{\rho^2(r_1+\alpha)(r_1+\alpha+1)(r_2+\alpha)(r_2+\alpha+1)}{2!\alpha(\alpha+1)}+\cdots\right] \\
&\quad = \frac{\beta_1^{r_1}\beta_2^{r_2}(1-\rho)^{\alpha+r_1+r_2}\Gamma(r_1+\alpha)\Gamma(r_2+\alpha)}{[\Gamma(\alpha)]^2}\ {}_2F_1(r_1+\alpha,r_2+\alpha;\alpha,\rho)
\end{aligned}
$$

where ${}_2F_1(.)$ define a confluent hypergeometric series given in [1], refer to formula (15.1.1) of page 556. If we consider a special case of $r_1 = r_2 = 2$ and follow the same process described above, then we have

$$
\begin{aligned}
E(X_1^2X_2^2) &= (\beta_1\beta_2)^2(1-\rho)^{\alpha+4}\frac{\alpha(\alpha+1)}{(1-\rho)^{\alpha+4}}(\alpha^2+\alpha+4\alpha\rho+\rho+2\rho^2) \\
&= (\beta_1\beta_2)^2\alpha(\alpha+1)(\alpha^2+\alpha+4\alpha\rho+\rho+2\rho^2).
\end{aligned}
$$

Hence, the variance of the product of $X_1$ and $X_2$ is written and simplified as

$$
\begin{aligned}
Var[X_1X_2] &= E[X_1^2X_2^2]-(E[X_1X_2])^2 \\
&= (\beta_1\beta_2)^2\alpha(\alpha+1)(\alpha^2+\alpha+4\alpha\rho+\rho+2\rho^2)-[\alpha\beta_1\beta_2(\alpha+\rho)]^2 \\
&= (\beta_1\beta_2)^2\alpha(2\alpha^2+\alpha+2\alpha^2\rho+8\alpha\rho+4\rho+\alpha\rho^2+2\rho^2).
\end{aligned}
$$

## 7.7 The correlation for a bivariate gamma distribution

In this section we want to show that

$$\frac{E[(X_1-\mu_1)(X_2-\mu_2)]}{\sqrt{E[(X_1-\mu_1)^2]E[(X_2-\mu_2)^2]}} = \frac{E[X_1X_2]-\mu_1\mu_2}{\sigma_1\sigma_2} = \rho$$

where $(X_1, X_2) \sim G(\alpha, \beta_1, \beta_2, \rho)$ and $\sigma_i^2 = \alpha\beta_i^2$, $i = 1, 2$ are means and variances respectively. The expected value of $E(X_1X_2)$ is given by

$$\begin{aligned} E[X_1X_2] &= \beta_1\beta_2(1-\rho)^{\alpha+2}\left[\frac{\alpha^2}{(1-\rho)^{\alpha+1}} + \frac{\alpha(\alpha+1)\rho}{(1-\rho)^{\alpha+2}}\right] \\ &= \alpha\beta_1\beta_2(\alpha+\rho) \end{aligned}$$

and it follows that

$$\begin{aligned} \frac{E[(X_1-\mu_1)(X_2-\mu_2)]}{\sigma_1\ \sigma_2} &= \frac{E[X_1X_2]-\mu_1\mu_2}{\sigma_1\ \sigma_2} \\ &= \frac{\alpha\beta_1\beta_2(\alpha+\rho)-(\alpha\beta_1)(\alpha\beta_2)}{\sqrt{\alpha\beta_1^2\alpha\beta_2^2}} \\ &= \rho. \end{aligned}$$

Hence $\rho$ is the correlation.

## 7.8 Distribution for the weighted sum of two correlated gamma variables

Consider the joint probability density function $f(x_1, x_2)$

$$\begin{aligned} f_{X_1X_2}(x_1,x_2) &= \frac{(x_1x_2)^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}}\ \exp\left(-\frac{\frac{x_1}{\beta_1}+\frac{x_2}{\beta_2}}{1-\rho}\right) \\ &\quad \times\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(\alpha+k)}\left(\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{x_1x_2}{\beta_1\beta_2}}\right)^{\alpha+2k-1}. \end{aligned}$$

Let $Z = \frac{\theta X_1}{\beta_1} + \frac{(1-\theta)X_2}{\beta_2}$ be the weighted sum of two gamma variables. By letting $u = \frac{\theta x_1}{\beta_1}$ and $z - u = \frac{(1-\theta)x_2}{\beta_2}$ then $\frac{dx_1}{du} = \frac{\beta_1}{\theta}$ and $\frac{dx_2}{d(u-z)} = \frac{\beta_2}{1-\theta}$. Therefore, the density $f_Z(z)$ for Z is given by

$$
\begin{aligned}
&\int_0^z f\left(\frac{\beta_1 u}{\theta}, \frac{\beta_2(z-u)}{1-\theta}\right)\, du \\
&= \int_0^z \frac{\left[\frac{\beta_1 u}{\theta}\frac{\beta_2(z-u)}{1-\theta})\right]^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}}\ \exp\left(-\frac{\frac{\frac{\beta_1 u}{\theta}}{\beta_1} + \frac{\frac{\beta_2(z-u)}{1-\theta}}{\beta_2}}{1-\rho}\right) \\
&\qquad \times \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)}\left[\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{\frac{\beta_1 u}{\theta}\frac{\beta_2(z-u)}{1-\theta}}{\beta_1\beta_2}}\right]^{\alpha+2k-1} \frac{\beta_1\beta_2}{\theta(1-\theta)}\, du \\
&= \int_0^z \frac{\left[\left(\frac{u}{\theta}\right)\left(\frac{z-u}{1-\theta}\right)\right]^{\frac{\alpha-1}{2}}(\beta_1\beta_2)^{\frac{\alpha-1}{2}}}{\Gamma(\alpha)(\beta_1\beta_2)^{\frac{\alpha+1}{2}}(1-\rho)\rho^{\frac{\alpha-1}{2}}}\ \exp\left(-\frac{\frac{u}{\theta}+\frac{z-u}{1-\theta}}{1-\rho}\right) \\
&\qquad \times \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)}\left[\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{u(z-u)}{\theta(1-\theta)}}\right]^{\alpha+2k-1} \frac{\beta_1\beta_2}{\theta(1-\theta)}\, du \\
&= \int_0^z \frac{[u(z-u)]^{\alpha-1}}{\Gamma(\alpha)[(1-\rho)\theta(1-\theta)]^{\alpha}} \exp\left(-\frac{\frac{u}{\theta}+\frac{z-u}{1-\theta}}{1-\rho}\right) \\
&\qquad \times \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)}\left[\frac{\sqrt{\rho}}{1-\rho}\sqrt{\frac{u(z-u)}{\theta(1-\theta)}}\right]^{2k} du \\
&= \frac{1}{\Gamma(\alpha)[(1-\rho)\theta(1-\theta)]^{\alpha}} \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)}\left[\frac{\sqrt{\rho}}{(1-\rho)\sqrt{\theta(1-\theta)}}\right]^{2k} \\
&\qquad \times \int_0^z \exp\left(-\frac{\frac{u}{\theta}+\frac{z-u}{1-\theta}}{1-\rho}\right)[u(z-u)]^{\alpha-1}\, du.
\end{aligned}
$$

Consider the integral,

$$
I_1 = \int_0^z \exp\left(-\frac{\frac{u}{\theta}+\frac{z-u}{1-\theta}}{1-\rho}\right)[u(z-u)]^{\alpha-1}\, du
$$

by letting $v = \frac{u}{z}$ and $du = zdv$,

$$\begin{aligned} I_1 &= \int_0^1 \exp\left(-\frac{\frac{vz}{\theta} + \frac{z-vz}{1-\theta}}{1-\rho}\right)[vz(z-vz)]^{\alpha-1} z\, dv \\ &= z^{2\alpha-1}\int_0^1 \exp\left[\frac{v(2z\theta - z) - z\theta}{(1-\rho)\theta(1-\theta)}\right] v^{\alpha-1}(1-v)^{\alpha-1}dv \\ &= z^{2\alpha-1}\exp\left[\frac{-z\theta}{(1-\rho)\theta(1-\theta)}\right] \\ &\quad \times \int_0^1 \exp\left[\frac{vz(2\theta-1)}{(1-\rho)\theta(1-\theta)}\right] v^{\alpha-1}(1-v)^{\alpha-1}dv. \end{aligned}$$

If $C = \frac{2\theta-1}{(1-\rho)\theta(1-\theta)}$, the integral

$$\begin{aligned} I_2 &= \int_0^1 e^{(Cz)v} v^{\alpha-1}(1-v)^{\alpha-1}dv \\ &= \sqrt{\pi}\left(\frac{1}{Cz}\right)^{\alpha-1/2} \exp\left(\frac{Cz}{2}\right)\, I_{\alpha-1/2}\,\left(\frac{Cz}{2}\right)\,\Gamma(\alpha). \end{aligned}$$

Therefore,

$$\begin{aligned} I_1 &= z^{2\alpha-1}\exp\left[\frac{-z\theta}{(1-\rho)\theta(1-\theta)}\right]\int_0^1 \exp\left[\frac{v(2z\theta-z)}{(1-\rho)\theta(1-\theta)}\right] v^{\alpha-1}(1-v)^{\alpha-1}dv \\ &= z^{2\alpha-1}\exp\left[\frac{-z\theta}{(1-\rho)\theta(1-\theta)}\right]\sqrt{\pi}\left(\frac{1}{Cz}\right)^{\alpha-1/2}\exp\left(\frac{Cz}{2}\right) \\ &\quad \times I_{\alpha-1/2}\,\left(\frac{Cz}{2}\right)\Gamma(\alpha) \\ &= \sqrt{\pi}\,\Gamma(\alpha)\,\left(\frac{z}{C}\right)^{\alpha-1/2}\exp\left[\frac{-z}{2(1-\rho)\theta(1-\theta)}\right]\, I_{\alpha-1/2}\,\left(\frac{Cz}{2}\right). \end{aligned}$$

Hence,

$$\begin{aligned}
&\int_0^z f\left(\frac{\beta_1 u}{\theta}, \frac{\beta_2(z-u)}{1-\theta}\right)\, du \\
&= \frac{1}{\Gamma(\alpha)[(1-\rho)\theta(1-\theta)]^{\alpha}} \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)} \left[\frac{\sqrt{\rho}}{(1-\rho)\sqrt{\theta(1-\theta)}}\right]^{2k} \\
&\quad \times\sqrt{\pi}\ \Gamma(\alpha)\ \left(\frac{z}{C}\right)^{\alpha-1/2} \exp\left[\frac{-z}{2(1-\rho)\theta(1-\theta)}\right]\ I_{\alpha-1/2}\ \left(\frac{Cz}{2}\right) \\
&= \frac{\sqrt{\pi}}{\sqrt{(1-\rho)\theta(1-\theta)}}\ \left(\frac{z}{2\theta-1}\right)^{\alpha-1/2}\ \exp\left[\frac{-z}{2(1-\rho)\theta(1-\theta)}\right] \\
&\quad \times\ I_{\alpha-1/2}\ \left(\frac{Cz}{2}\right)\ \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)} \left[\frac{\sqrt{\rho}}{2(1-\rho)\sqrt{\theta(1-\theta)}}\right]^{2k} \\
&= \frac{\sqrt{\pi}}{\Gamma(\alpha)\sqrt{(1-\rho)\theta(1-\theta)}}\ \left(\frac{z}{2\theta-1}\right)^{\alpha-1/2}\ \exp\left[\frac{-z}{2(1-\rho)\theta(1-\theta)}\right] \\
&\quad \times\ I_{\alpha-1/2}\ \left(\frac{Cz}{2}\right)\ F_1(\alpha; r^2).
\end{aligned}$$

Note that we have written $r = \frac{\sqrt{\rho}}{(1-\rho)\sqrt{\theta(1-\theta)}}$ in the previous expression and that

$$\begin{aligned}
&\sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k)}\ r^{2k} \\
&= \frac{1}{\Gamma(\alpha)} + \frac{1}{1!\Gamma(\alpha+1)} r^2 + \frac{1}{2!\Gamma(\alpha+2)} r^4 + \cdots \\
&= \frac{1}{\Gamma(\alpha)} + \frac{1}{1!\alpha\Gamma(\alpha)} r^2 + \frac{1}{2!\alpha(\alpha+1)\Gamma(\alpha)} r^4 + \cdots \\
&= \frac{1}{\Gamma(\alpha)} \left[1 + \frac{1}{1!\alpha} r^2 + \frac{1}{2!\alpha(\alpha+1)} (r^2)^2 + \cdots\right] \\
&= \frac{1}{\Gamma(\alpha)}\ F_1(\alpha; r^2)
\end{aligned}$$

where $F_1(\alpha; r^2)$ is a confluent hypergeometric series. In general, the confluent hypergeometric series is written as

$$\begin{aligned} {}_1F_1(a;b;z) &= 1 + \frac{a}{b}z + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + \cdots \\ &= \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k}\frac{z^k}{k!} \end{aligned}$$

where $(a)_k$ and $(b)_k$ are known as Pochhammer symbols. The Pochhammer symbol of $(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)}$ and $(a)_0 = 1$. Refer to formula (15.1.1) of page 556 in [1].

## 7.9 List of formulae used

- Series representation for modified Bessel function of the first kind, equation (8.445) on page 971 [40]

$$I_\nu(x) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\nu + k + 1)} \left(\frac{x}{2}\right)^{\nu+2k}$$

- Binomial series

$$(1-x)^{-r} = 1 + rx + \frac{r(r+1)x^2}{2!} + \frac{r(r+1)(r+2)x^3}{3!} + \cdots$$

- Duplication (or doubling) formula for gamma function (Legendre duplication formula), equation (8.335–1) on page 946 [40]

$$\frac{\Gamma(2x)}{\Gamma(x)} = \frac{2^{2x-1}\Gamma(x+1/2)}{\Gamma(1/2)} = \frac{2^{2x-1}\Gamma(x+1/2)}{\sqrt{\pi}}.$$

- The formula for (3.388–1) on page 368 of [40] is

$$\int_0^{2u} (2ux - x^2)^{\nu-1}\ e^{-\mu x}\ dx = \sqrt{\pi}\left(\frac{2u}{\mu}\right)^{\nu-1/2}\ e^{-u\mu}\ \Gamma(\nu)\ I_{\nu-1/2}(u\mu)$$

where $u > 0,\ \nu > 0$ and the formula for (3.383–2) on page 365 of [40] is

$$\int_0^u x^{\mu-1}(u-x)^{\mu-1}\ e^{\beta x}\ dx = \sqrt{\pi}\left(\frac{u}{\beta}\right)^{u-1/2} \exp\left(\frac{\beta u}{2}\right)\Gamma(\mu)\ I_{\mu-1/2}\left(\frac{\beta u}{2}\right).$$

Combining the integration formulae of (3.388–1) and (3.383–2) will give

$$\int_0^z x^{\nu-1}(z-x)^{\nu-1}\ e^{\mu x}\ dx = \sqrt{\pi}\left(\frac{z}{|\mu|}\right)^{\nu-1/2}\ e^{\frac{z\mu}{2}}\ \Gamma(\nu)\ I_{\nu-1/2}\left(\frac{z|\mu|}{2}\right)$$

for $v > 0$ and $z > 0$.

## 7.10 A case study for the sum of two correlated gamma random variables

To illustrate the use of the PDF of the sum of two correlated gamma variables of the McKay distribution given in formula (C.1), we use monthly rainfall totals from Hume and Beechworth meteorological stations. The data used is from 1922 to 2007 for the months of December to February during the summer season. The analysis is based on the two stations not the monthly analysis. We consider the positive pairs set of data from two stations for the analysis. The Spearman correlation between the positive pairs is significant at 5% significance level with correlation value of $\rho = 0.88$. We use $R$ package to do data manipulation. Refer to Appendix C.

Firstly, the data from each month is fitted to the gamma distribution using the method described in Section 3.1. The shape and scale parameters are estimated using the maximum likelihood method in Section 3.1.2 and the mean and variance of the observed and estimated distributions are calculated as shown in Table 7.1.

**Table 7.1:** Shape and scale parameters of monthly rainfall totals in summer months

| Data | Parameter | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | observed | $\alpha\beta$ | observed | $\alpha\beta^2$ |
| Hume | 1.0132 | 46.1847 | 46.7950 | 46.7950 | 1983.793 | 2161.2120 |
| Beechworth | 1.2930 | 44.3058 | 57.2866 | 57.2866 | 2649.387 | 2538.1300 |

The PDF of the sum of correlated gamma variables of the McKay distribution requires a single value of $\alpha$ with distinct values of $\beta$ and the correlation parameter ($\rho_{ij}$). For convenience, we use the average values of $\alpha$ denoted by $\overline{\alpha} = (\alpha_1+\alpha_2)/2 = 1.1531$ in the PDF. Using the fixed value of $\overline{\alpha}$, re-estimate the values of $\beta_1$ and $\beta_2$ using $X_i \sim G(\overline{\alpha}, \beta_i)$ for each $i = 1, 2$, see in Table 7.2.

**Table 7.2:** Re-estimated values of $\beta$ for $\overline{\alpha} = 1.1531$

| | Hume | Beechworth |
|---|---|---|
| $\beta_i$ | 40.5820 | 49.6806 |

Since the two stations is significantly correlated, we apply the sum of two correlated gamma variables of the McKay distribution. The sum of two stations is denoted by $Y = X_1 + X_2$. Then a synthetic rainfall totals is generated using the PDF of the sum of correlated gamma variables of the McKay distribution for parameters $\overline{\alpha}$, $\beta_1$, $\beta_2$ and $\rho_{ij}$. The $R$ code for evaluating the PDF of the McKay distribution is given in Appendix C.6. To obtain the CDF, we construct a look-up table to determine the associated rainfall sums. The detail procedure of simulating the generated sum is described in Section 3.3.

Both the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are used to assess the fit between the observed and generated data. Plots of the PDFs and CDFs are also presented to support the results.

### 7.10.1 Results

We analyse the statistics of the observed monthly rainfall totals and generated data for the marginals and the sum of the corresponding monthly rainfall. The value of means and variances of the observed and generated are shown in Table 7.3.

The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are applied to test the null hypothesis that the observed and generated data are not significantly different. Table 7.4 shows the P-value for the analyses. Since the P-value $> 0.05$ for all cases, therefore, the observed and generated data are not significantly different from one another at 5% significance level. These results are also supported by the plot of histogram and a fitted curve, see Figures 7.1 and 7.2.

**Table 7.3:** Statistics of the observed and generated data for monthly rainfall totals

| | Mean (mm) | | Variance | |
|---|---|---|---|---|
| | observed | generated | observed | generated |
| Hume | 46.79 | 48.04 | 1983.81 | 2043.04 |
| Beechworth | 57.29 | 59.91 | 2649.16 | 3110.29 |
| Sum of two stations | 104.08 | 102.51 | 8443.20 | 8344.77 |

**Table 7.4:** Goodness of fit test of marginals and sum of two correlated gamma variables of McKay distribution

| | Hume | Beechworth | Sum of two stations |
|---|---|---|---|
| Kolmogorov–Smirnov | 0.6515 | 0.1857 | 0.7866 |
| Anderson–Darling | 0.4756 | 0.4171 | 0.6129 |

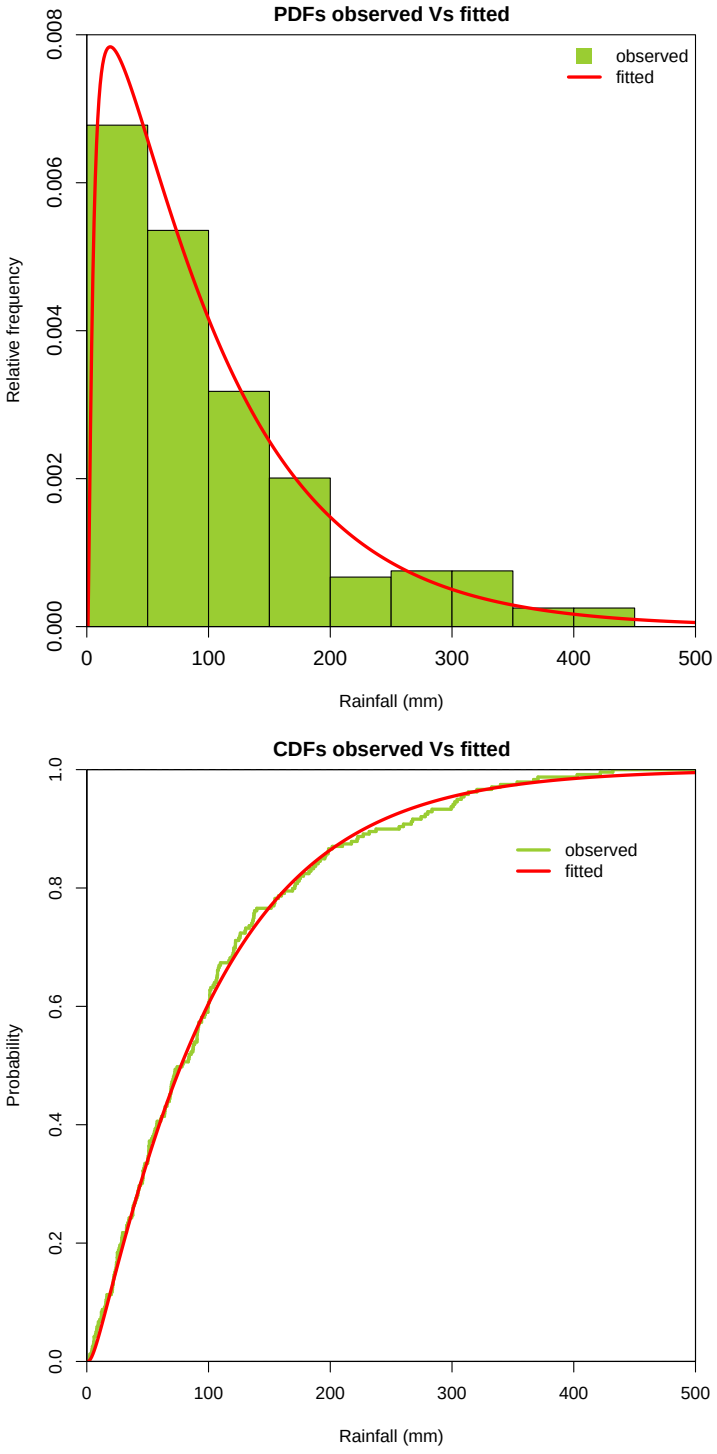



**Figure 7.1:** Observed vs fitted PDFs and CDFs for the sum of two correlated gamma variables using the McKay distribution, Hume and Beechworth

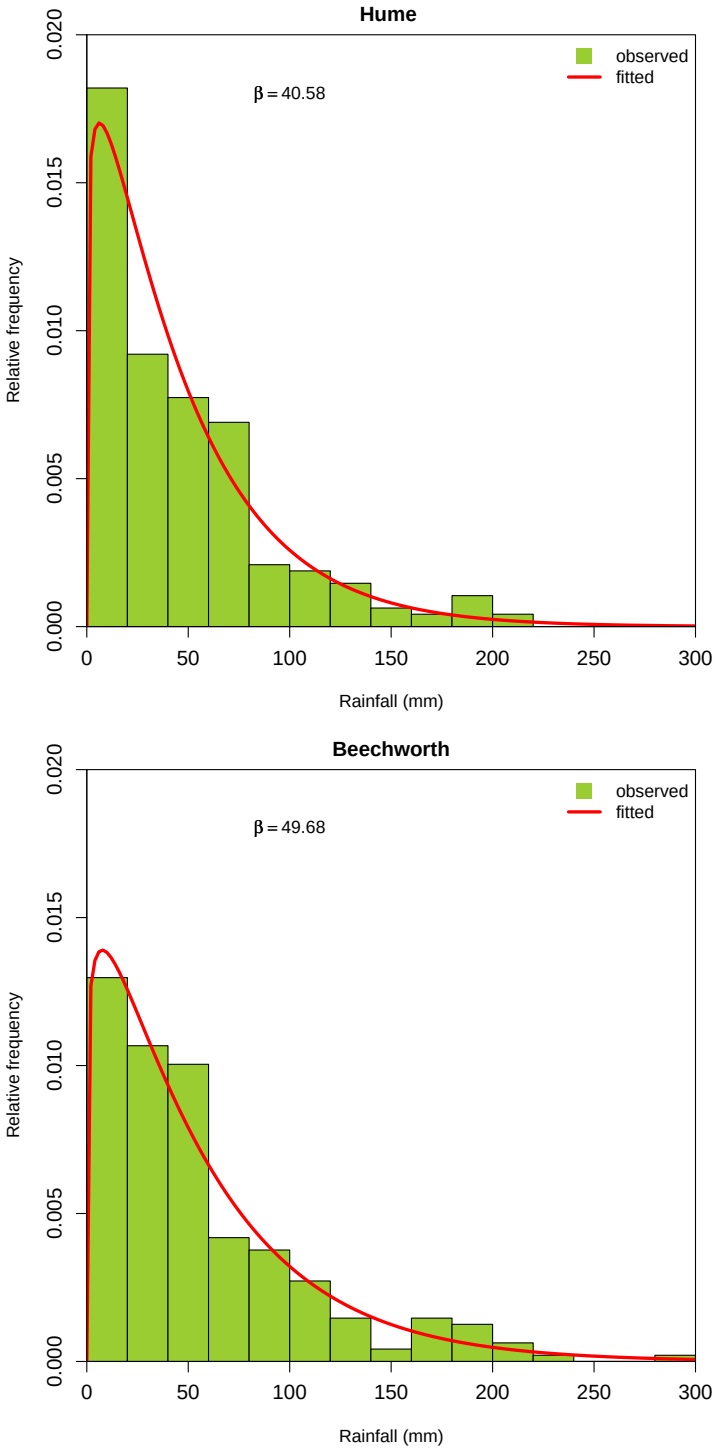



**Figure 7.2:** Observed vs fitted PDFs of gamma marginals, Hume and Beechworth stations

## 7.11 Summary

We have shown the proofs of the PDF of the sum of two independent gamma variables, PDF of the sum of two correlated gamma variables, joint PDF of two correlated gamma variables, the marginal PDF of bivariate gamma distribution and the formula for correlation. We derived the PDF of the weighted sum of two correlated gamma variables which involve hypergeometric function. Modelling monthly rainfall data using the sum of two correlated gamma variables is also demonstrated. In the next section, we will show the use of moment generating function in parameter estimation for the sum of two correlated gamma variables.

# Chapter 8

# Using the method of moments to estimate shape and scale parameters

In the previous chapters we have used the maximum likelihood method to estimate shape and scale parameters for the sums of both independent and correlated gamma variables. Suppose we wished to sum $n$ correlated gamma variables where we assume that all distributions have the same shape parameter by matching observed moments for the sum. Once again we use the model proposed by Alouini et al. [3] that we described in Chapter 6. We would have to evaluate the shape parameter $\alpha$, the scale parameters $\beta_1, \ldots, \beta_n$ and the correlations $\rho_{ij}$ for $1 \leq i < j \leq n$. Thus it would seem to be natural to match the $(n+1) + n(n-1)/2$ least order moments. In this chapter we will use a detailed study of this method for the case $n = 2$ to show that the proposed procedure is satisfactory. In the case $n = 2$ we have four unknowns and so we will use the first four moments to form four nonlinear equations. We will use data from two meteorological stations (Hume and Beechworth, 1922–2009) to

demonstrate the methodology.

## 8.1 Moments, moment generating functions and cumulants

First, we review the theory of moments and introduce the moment generating function (MGF) and the cumulant generating function (CGF). We know that $E[x^k]$ is defined as the $k^{th}$ moment of $x$. The MGF of a random variable $x$ is denoted by $M_X(t) = E[e^{tx}]$ which is defined by

$$E[X^k] = \frac{d^k}{dt^k} M_X(t)|_{t=0} \quad \text{for} \quad k = 1, 2, \ldots$$

Therefore, we generate the moments of $X$ from the MGF. In other words, the $n^{th}$ moment is equal to the $n^{th}$ derivative of $M_X(t)$ evaluated at $t = 0$. Some examples of moments are

$$\begin{aligned} E[X] &= \frac{d}{dt} M_X(t)|_{t=0} = \mu \\ E[X^2] &= \frac{d^2}{dt^2} M_X(t)|_{t=0} \end{aligned}$$

where the variance is calculated as

$$V[X] = E[X^2] - (E[X])^2 .$$

We know that moments are used to measure the distribution properties such as mean and variance. We may also use the cumulant generating function denoted by $\kappa_n$ of a random variable X which is defined as the logarithm of the moment generating function to derive an equivalent set of descriptive statistics. If the cumulant generating function

$$g(t) = \log(M_X(t)) = \log(E[e^{tX}])$$

is analytic in some region $|t| < r$ then since $g(0) = 0$ we can write

$$g(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}.$$

Cumulants are closely related to moments. By evaluating the derivatives at $t = 0$ we obtain

$$\begin{aligned}
\kappa_1 &= g'(0) = \mu \\
\kappa_2 &= g''(0) = \sigma^2 \\
\kappa_3 &= g'''(0) \\
\kappa_4 &= g''''(0) \\
&\vdots \\
\kappa_n &= g^n(0).
\end{aligned}$$

Note that the first cumulant $\kappa_1$ defines the mean and the second cumulant $\kappa_2$ defines the variance. We can use cumulants to express skewness $\Gamma_1$ and kurtosis $\Gamma_2$ using the formulae

$$\Gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{g^3(0)}{(\sigma^2)^{3/2}} \quad \text{and} \quad \Gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{g^4(0)}{(\sigma^2)^2}.$$

The reader is referred to [16, 55, 59] for further details.

## 8.2 Moment generating function for the sum of two correlated gamma variables

Consider a set of gamma random variables $\{X_i\}_{i=1}^2$ with respective density functions

$$f(x; \alpha_i, \beta_i) = \frac{x^{\alpha_i - 1} \exp(-x/\beta_i)}{\Gamma(\alpha_i)\beta_i^{\alpha_i}}$$

where $x > 0$ and $\alpha_i, \beta_i > 0$ are the shape and scale parameters, respectively. Mathai [63] uses the moment generating function to obtain various density functions with different combinations of parameters, such as both shape and scale parameters are integers. He also studies density function having the common shape and common scale parameters. Moschopoulos [68] extends this work for the sum of $n$ independent gamma variables with both different shape and scale parameters and obtained a single gamma-series form with coefficients calculated recursively. He uses some algebraic transformations to derive a new form of the moment generating function which has more practical use. By employing the same technique as Moschopoulos [68], Alouini et al. [3] introduce a probability density function for the sum of $n$ correlated gamma random variables all having the same shape to model the wireless communication system.

For the sum of $n$ correlated gamma variables, we follow Alouini et al. [3] and assume that $X_i \sim G(\alpha, \beta_i)$ for all $i = 1, 2, \ldots, n$ are gamma random variables which all have the same shape parameter. We write $Y = \sum_{i=1}^{n} X_i$ and note that the moment generating function $M_Y(t) = E[e^{tY}]$ is given by

$$M_Y(t) = |I - tDC|^{-\alpha}$$

where $|.|$ is the determinant operator, $I$ is the $n \times n$ identity matrix, the matrix $D$ is a diagonal matrix of $\{\beta_i\}_{i=1}^{n}$ given by

$$D = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{bmatrix}_{n \times n}$$

and $C$ is defined by

$$C = \begin{bmatrix} 1 & \sqrt{\rho_{12}} & \cdots & \sqrt{\rho_{1n}} \\ \sqrt{\rho_{21}} & 1 & \cdots & \sqrt{\rho_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\rho_{n1}} & \sqrt{\rho_{N2}} & \cdots & 1 \end{bmatrix}_{n \times n}.$$

We note once again that the Alouini et al. [3] method requires $C$ to be positive definite. Since this is not guaranteed, the eigenvalues of $C$ must be calculated to check the viability of every potential case study.[1] The eigenvalues of the matrix $DC$ are denoted by $\{\lambda_i\}_{i=1}^{n}$. The moment generating function can also be expressed in terms of the eigenvalues as

$$M_Y(t) = \prod_{i=1}^{n} (1 - t\lambda_i)^{-\alpha}.$$

Hence, the moment generating function has two forms, the determinant form and product form.

## 8.3 Parameter estimation using the determinant form of the moment generating function

Consider rainfall totals from two stations with a common shape parameter ($\alpha$), two scale parameters ($\beta_1, \beta_2$) and correlation coefficient $\rho$. We use the Spearman's rank correlation coefficient since our data is skewed. Refer to [81] and [29]. The determinant form of the moment generating function is $M_Y(t) = |I - tDC|^{-\alpha}$ where

[1] Recall that in Chapter 6.4 we displayed a counter-example to show that the matrix $C$ may have one or more negative eigenvalues.

the matrices are

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & \sqrt{\rho_{12}} \\ \sqrt{\rho_{21}} & 1 \end{bmatrix}.$$

The product $A = DC$ is given by

$$DC = \begin{bmatrix} \beta_1 & \beta_1\sqrt{\rho} \\ \beta_2\sqrt{\rho} & \beta_2 \end{bmatrix}$$

and hence

$$I - tDC = \begin{bmatrix} 1 - t\beta_1 & -t\beta_1\sqrt{\rho} \\ -t\beta_2\sqrt{\rho} & 1 - t\beta_2 \end{bmatrix}$$

and

$$\begin{aligned} |I - tDC|^{-\alpha} &= [(1 - t\beta_1)(1 - t\beta_2) - (t\beta_1\sqrt{\rho})(t\beta_2\sqrt{\rho})]^{-\alpha} \\ &= [t^2\beta_1\beta_2(1 - \rho) - t(\beta_1 + \beta_2) + 1]^{-\alpha}. \end{aligned}$$

We want to estimate four parameters, namely $\alpha, \beta_1, \beta_2$ and $\rho$ and we want to solve a system of nonlinear equations with four unknowns. We set up the equations for mean ($\mu$), variance ($\sigma^2$), skewness ($\Gamma_1$) and kurtosis ($\Gamma_2$) using the cumulants and equate these values with the values calculated from the observed data. Refer to Table 8.1. The maximum likelihood method is used to fit a single gamma distribution to the observed sums. The estimated parameters are shown in Table 8.2. The gamma distribution will preserve the mean value of the sum but is unlikely to match the other statistics (variance, skewness and kurtosis). Note that the variance of this single gamma distribution does not match the observed variance.

**Table 8.1:** Descriptive statistics for positive pairs of Hume and Beechworth (1922–2009)

| | Mean (mm) | Variance | Skewness | Kurtosis |
|---|---|---|---|---|
| Hume | 58.64 | 1789.51 | 0.95 | 0.64 |
| Beechworth | 83.72 | 3315.24 | 0.87 | 0.49 |
| Sum | 142.36 | 9335.69 | 0.81 | 0.25 |

**Table 8.2:** Estimated parameters for positive pairs of Hume and Beechworth

| Data | Parameter | | Mean (mm) | | Variance | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | observed | $\alpha\beta$ | observed | $\alpha\beta^2$ |
| Hume | 1.55 | 37.91 | 58.64 | 58.64 | 1789.51 | 2223.09 |
| Beechworth | 1.72 | 48.67 | 83.72 | 83.72 | 3315.24 | 4047.59 |
| Sum | 1.76 | 80.95 | 142.36 | 142.36 | 9335.69 | 11524.52 |

We know the moment generating function for the sum of two correlated gamma variables is given by

$$M_Z(t) = |I - tDC|^{-\alpha} = [t^2\beta_1\beta_2(1-\rho) - t(\beta_1+\beta_2) + 1]^{-\alpha}$$

where $Z = X_1 + X_2$. It follows that

$$g(t) = \log(M_Z(t)) = -\alpha \log(\beta_1\beta_2(1-\rho)t^2 - (\beta_1+\beta_2)t + 1)$$

and hence that

$$\begin{aligned}
\kappa_1 &= g'(0) = \alpha(\beta_1 + \beta_2) = \mu \\
\kappa_2 &= g''(0) = \alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2) = \sigma^2 \\
\kappa_3 &= g'''(0) = 2\alpha(\beta_1 + \beta_2)(\beta_1^2 + \beta_2^2 - \beta_1\beta_2 + 3\beta_1\beta_2\rho) \\
\kappa_4 &= g''''(0) = 6\alpha\left[(\beta_1 + \beta_2)^2(\beta_1^2 + \beta_2^2 - 2\beta_1\beta_2 + 4\beta_1\beta_2\rho) + 2\beta_1^2\beta_2^2(1-\rho)^2\right].
\end{aligned}$$

Following from Section 8.1, the formula of skewness and kurtosis are given by

Skewness

$$\Gamma_1 = \frac{2\alpha(\beta_1 + \beta_2)(3\beta_1\beta_2\rho - \beta_1\beta_2 + \beta_1^2 + \beta_2^2)}{[\alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2)]^{3/2}}$$

and

Kurtosis

$$\Gamma_2 = \frac{6\alpha(\beta_1 + \beta_2)^2(4\beta_1\beta_2\rho - 2\beta_1\beta_2 + \beta_1^2 + \beta_2^2) + 12\alpha\beta_1^2\beta_2^2(1-\rho)^2}{[\alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2)]^2}.$$

### 8.3.1 Solve for the values of $\alpha, \beta_1, \beta_2, \rho$

We wish to calculate the parameters $\alpha$, $\beta_1$, $\beta_2$ and $\rho$. We attempt to solve the moment equations (8.1)–(8.4) for the first four cumulants in a general form. The moment equations are given by

$$\alpha(\beta_1 + \beta_2) = \mu \tag{8.1}$$

$$\alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2) = \sigma^2 \tag{8.2}$$

$$\frac{2\alpha(\beta_1 + \beta_2)(3\beta_1\beta_2\rho - \beta_1\beta_2 + \beta_1^2 + \beta_2^2)}{[\alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2)]^{3/2}} = \Gamma_1 \tag{8.3}$$

$$\frac{6\alpha(\beta_1 + \beta_2)^2(4\beta_1\beta_2\rho - 2\beta_1\beta_2 + \beta_1^2 + \beta_2^2) + 12\alpha\beta_1^2\beta_2^2(1-\rho)^2}{[\alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2)]^2} = \Gamma_2 \tag{8.4}$$

where we have used $\mu = \kappa_1$, $\sigma^2 = \kappa_2$, $\Gamma_1 = \kappa_3/\sigma^3$ and $\Gamma_2 = \kappa_4/\sigma^4$. Consider equation (8.1), we have

$$\alpha = \frac{\mu}{\beta_1 + \beta_2} \tag{8.5}$$

and eliminate $\alpha$ from the remaining equations to reduce the system to three equations in three unknowns given by

$$\mu(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2) = \sigma^2(\beta_1 + \beta_2) \tag{8.6}$$

$$2\mu(3\beta_1\beta_2\rho - \beta_1\beta_2 + \beta_1^2 + \beta_2^2) = \Gamma_1\sigma^3 \tag{8.7}$$

$$\begin{aligned} 6\mu(\beta_1 + \beta_2)(4\beta_1\beta_2\rho - 2\beta_1\beta_2 + \beta_1^2 + \beta_2^2) & \\ +12\frac{\mu}{(\beta_1 + \beta_2)}\beta_1^2\beta_2^2(1-\rho)^2 &= \Gamma_2\sigma^4. \end{aligned} \tag{8.8}$$

From equation (8.6), $\rho$ is expressed as

$$\rho = \frac{\sigma^2(\beta_1 + \beta_2) - \mu(\beta_1^2 + \beta_2^2)}{2\mu\beta_1\beta_2}$$

which it is convenient to rewrite as

$$\rho = \frac{\sigma^2(\beta_1 + \beta_2) - \mu(\beta_1 + \beta_2)^2}{2\mu\beta_1\beta_2} + 1. \tag{8.9}$$

Hence we can also eliminate $\rho$ from the remaining equations. The equation (8.7) becomes

$$2\mu\left[3\left\{\frac{\sigma^2(\beta_1 + \beta_2) - \mu(\beta_1 + \beta_2)^2}{2\mu} + \beta_1\beta_2\right\} - 3\beta_1\beta_2 + (\beta_1 + \beta_2)^2\right] = \Gamma_1\sigma^3$$

which reduces to a quadratic equation for $(\beta_1 + \beta_2)$ given by

$$(\beta_1 + \beta_2)^2 - \frac{3\sigma^2}{\mu}(\beta_1 + \beta_2) + \frac{\Gamma_1\sigma^3}{\mu} = 0. \tag{8.10}$$

After substitution for $\rho$, the equation (8.8) becomes

$$\begin{aligned} 6\mu(\beta_1 + \beta_2)^2\left[4\left\{\frac{\sigma^2(\beta_1 + \beta_2) - \mu(\beta_1 + \beta_2)^2}{2\mu} + \beta_1\beta_2\right\} - 4\beta_1\beta_2 + (\beta_1 + \beta_2)^2\right] & \\ +12\mu\left\{\frac{\mu(\beta_1 + \beta_2)^2 - \sigma^2(\beta_1 + \beta_2)}{2\mu}\right\}^2 &= \Gamma_2\sigma^4(\beta_1 + \beta_2) \end{aligned}$$

which, after minor simplification, gives

$$6\mu(\beta_1+\beta_2)^2\left[\frac{2\sigma^2}{\mu}(\beta_1+\beta_2)-(\beta_1+\beta_2)^2\right] \\ +12\mu\left\{\frac{\mu(\beta_1+\beta_2)^2-\sigma^2(\beta_1+\beta_2)}{2\mu}\right\}^2 = \Gamma_2\sigma^4(\beta_1+\beta_2).$$

So the equation (8.8) has also become an equation for $(\beta_1+\beta_2)$. Indeed the only possible satisfactory outcome now is that this equation will be the same as the previous equation. On further simplification we obtain

$$(\beta_1+\beta_2)^4-\frac{2\sigma^2}{\mu}(\beta_1+\beta_2)^3-\left(\frac{\sigma^2}{\mu}\right)^2(\beta_1+\beta_2)^2+\frac{\Gamma_2\sigma^4}{3\mu}(\beta_1+\beta_2)=0. \quad (8.11)$$

We need to normalise this equation in such a way that we can judge the significance of any error when we substitute for $\beta_1+\beta_2$. Hence, we divide through equation (8.11) by $\frac{3\mu}{\Gamma_2\sigma^4(\beta_1+\beta_2)}$ so that it makes the individual terms approximately equal to 1. Then, we rewrite equation (8.11) as

$$\frac{3\mu}{\Gamma_2\sigma^4}(\beta_1+\beta_2)^3-\frac{6}{\Gamma_2\sigma^2}(\beta_1+\beta_2)^2-\frac{3}{\Gamma_2\mu}(\beta_1+\beta_2)+1=0. \quad (8.12)$$

The correct interpretation of our result is that there must be a deterministic relationship between the first four cumulants. We can find the relationship by writing down the theoretical solutions to equation (8.10) and substituting these solutions into equation (8.12). Thus we substitute the solution

$$\beta_1+\beta_2=\frac{3\sigma^2}{2\mu}\pm\sqrt{\left(\frac{3\sigma^2}{2\mu}\right)^2-\frac{\Gamma_1\sigma^3}{\mu}}$$

to equation (8.10) into equation (8.12). If the observed cumulants do not satisfy this equation then our distribution cannot be written in the proposed theoretical form. Of course we can never expect observed values to be precisely equal to the theoretical values. In practice we must make a value judgement about whether the observed values approximately satisfy the required deterministic relationship.

Now, using the values of mean, variance, skewness and kurtosis calculated from the observed data, substitute into the RHS of equations (8.1)–(8.4) where

$$\mu = 142.36; \quad \sigma^2 = 9335.69; \quad \Gamma_1 = 0.81; \quad \Gamma_2 = 0.25.$$

Substitute the values of $\mu, \sigma^2, \Gamma_1$ and $\Gamma_2$ into the solution of equation (8.10) as follows

$$\begin{aligned} \beta_1 + \beta_2 &= \left(\frac{3(9335.69)}{2(142.36)}\right) \pm \sqrt{\left(\frac{3(9335.69)}{2(142.36)}\right)^2 - \frac{(0.81)(9335.69)^{\frac{3}{2}}}{142.36}} \\ &= 98.37 \pm 67.14 \\ &= 30.96, 165.78 \end{aligned}$$

Then use the values of $\beta_1 + \beta_2 = 30.96$ or $\beta_1 + \beta_2 = 165.78$ to satisfy equation (8.12).

When $\beta_1 + \beta_2 = 30.96$, the LHS of equation (8.12) becomes

$$\begin{aligned} LHS &= \frac{3(142.36)(30.96)^3}{(0.25)(9335.69)^2} - \frac{6(30.96)^2}{(0.25)(9335.69)} - \frac{3(30.96)}{(0.25)(142.36)} + 1 \\ &= -3.492196 \\ &\neq 0. \end{aligned}$$

Similarly, when $\beta_1 + \beta_2 = 165.78$, the LHS of equation (8.12) becomes

$$\begin{aligned} LHS &= \frac{3(142.36)(165.78)^3}{(0.25)(9335.69)^2} - \frac{6(165.78)^2}{(0.25)(9335.69)} - \frac{3(165.78)}{(0.25)(142.36)} + 1 \\ &= 5.677482 \\ &\neq 0. \end{aligned}$$

The error is "small" if the error is much less in magnitude than one but it is "large" if the error is much bigger in magnitude than one. Hence the equation is not satisfied. Since the consistency equation is not satisfied we conclude that our model cannot be fitted to this data. Thus, unfortunately, the system of equations is theoretically inconsistent. Since the equations are inconsistent it is not possible to obtain a

"numerical solution". Note that the value $\alpha = 1.76$ obtained when we attempted to fit a gamma distribution directly to the sum does not match the values we obtained by attempting to solve the (inconsistent) moment equations. This value is much less than the value $142.36/30.96 \approx 4.60$ and much greater than $142.36/165.78 \approx 0.86$.

The unsatisfactory nature of our attempted solution is further confirmed when we look more closely at the calculated values for $\beta_1$ and $\beta_2$. Rearrange equations (8.6) and (8.7) for $\beta_1^2 + \beta_2^2$ and equate both expression and solve for $\beta_1\beta_2$

$$\begin{aligned} \frac{\sigma^2(\beta_1+\beta_2) - 2\mu\beta_1\beta_2\rho}{\mu} &= \frac{\Gamma_1\sigma^3 + 2\mu\beta_1\beta_2(1-3\rho)}{2\mu} \\ \beta_1\beta_2 &= \frac{\Gamma_1\sigma^3 - 2\sigma^2(\beta_1+\beta_2)}{2\mu(\rho-1)}. \end{aligned}$$

For example, calculate the value for $\beta_1\beta_2$ by substituting

$$\mu = 142.36,\ \sigma^2 = 9335.69,\ \Gamma_1 = 0.81,\ \rho = 0.51,\ \beta_1 + \beta_2 = 165.78$$

will give $\beta_1\beta_2 = 16949.69$ and $\beta_1^2 + \beta_2^2 = -6417.16$ which is clearly impossible.

**Method 2**

In method 2, we want to estimate $\alpha$ and $\rho$ from equations (8.1) and (8.2). First, the values of $\beta_1$ and $\beta_2$ are estimated using the maximum likelihood method as described in Chapter 3.1.2. Refer to Table 8.2, we have

$$\begin{aligned} Hume: \beta_1 &= 37.91, \quad \alpha_1 = 1.55 \\ Beec: \beta_2 &= 48.67, \quad \alpha_2 = 1.72 \end{aligned}$$

Use equation (8.1) to calculate $\alpha$

$$\begin{aligned} \alpha &= \frac{142.36}{\beta_1+\beta_2} \\ &= \frac{142.36}{37.91+48.67} \\ &= 1.6443 \end{aligned}$$

Then, use equation (8.2)

$$\alpha(2\beta_1\beta_2\rho + \beta_1^2 + \beta_2^2) = 9335.69$$

to express $\rho$ in terms of $\beta_1$ and $\beta_2$. Hence deduce that

$$\begin{aligned}\rho &= \frac{9335.69(\beta_1+\beta_2) - 142.36(\beta_1+\beta_2)^2}{2(142.36)\beta_1\beta_2} + 1 \\ &= \frac{9335.69(37.91+48.67) - 142.36(37.91+48.67)^2}{284.72(37.91)(48.67)} + 1 \\ &= 0.51\end{aligned}$$

The value of $\alpha = 1.6443$ is found to be quite reasonable. However the value of $\rho = 0.51$ is quite small compared to the Spearman correlation calculated from the observed data which is 0.90. Thus we conclude that this method is also unsatisfactory.

## 8.4 Parameter estimation using product form of the moment generating function

Using the product form of the moment generating function, we again consider rainfall totals from two stations ($n = 2$) with a common shape parameter ($\alpha_1 = \alpha_2 = \alpha$) and two scale parameters ($\beta_1, \beta_2$). The product form of the moment generating function is $M_Y(t) = \prod_{i=1}^{2}(1 - t\lambda_i)^{-\alpha}$, see [3].

Firstly, we study the formula where we let $\lambda_1 = \min_i(\lambda_i)$ and transform,

$$(1 - t\lambda_i) = (1 - t\lambda_1)\left(\frac{\lambda_i}{\lambda_1}\right)\left(1 - \frac{1 - \frac{\lambda_1}{\lambda_i}}{1 - t\lambda_1}\right).$$

For further details, we refer the reader to [68]. Then, we have

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n}(1-t\lambda_i)^{-\alpha} \\ &= (1-t\lambda_1)^{-\alpha}\prod_{i=2}^{n}\left[(1-t\lambda_1)\left(\frac{\lambda_i}{\lambda_1}\right)\left(1-\frac{1-\frac{\lambda_1}{\lambda_i}}{1-t\lambda_1}\right)\right]^{-\alpha}. \end{aligned}$$

Taking the natural logarithm gives

$$\begin{aligned} \ln M_Y(t) &= -n\alpha\ln(1-t\lambda_1)+n\alpha\ln\lambda_1-\alpha\sum_{i=2}^{n}\ln\lambda_i-\alpha\sum_{i=2}^{n}\ln\left(1-\frac{1-\frac{\lambda_1}{\lambda_i}}{1-t\lambda_1}\right) \\ &= -n\alpha\ln(1-t\lambda_1)+n\alpha\ln\lambda_1-\alpha\sum_{i=2}^{n}\ln\lambda_i+\sum_{k=1}^{\infty}\Gamma_k(1-t\lambda_1)^{-k} \end{aligned}$$

where

$$\Gamma_k = \alpha\sum_{i=2}^{n}(1-\lambda_1/\lambda_i)^k/k$$

for each $k\in\mathbb{N}$ and where we have used the expansion

$$\ln(1-x) = (-1)\sum_{k=1}^{\infty}x^k/k.$$

In the case $n=2$ this becomes

$$\ln M_Y(t) = -2\alpha\ln(1-t\lambda_1)+2\alpha\ln\lambda_1-\alpha\lambda_2+\sum_{k=1}^{\infty}\Gamma_k(1-t\lambda_1)^{-k}$$

where

$$\Gamma_k = \alpha(1-\lambda_1/\lambda_2)^k/k$$

for $k\in\mathbb{N}$. To calculate $\lambda_1$ and $\lambda_2$ we solve $\det(\lambda I - A)=0$ where

$$A = \begin{bmatrix} \beta_1 & \beta_1\sqrt{\rho} \\ \beta_2\sqrt{\rho} & \beta_2 \end{bmatrix}.$$

The solution is

$$\lambda = \frac{1}{2}\left[(\beta_1+\beta_2)\pm\sqrt{4\beta_1\beta_2\rho+(\beta_1-\beta_2)^2}\right].$$

Now we could use the expression $g(t) = \ln M_Y(t)$ obtained above to calculate the derivatives $g'(0)$, $g''(0)$, $g'''(0)$ and $g''''(0)$ and hence find the first four cumulants. Of course this simply reproduces the same expressions as those calculated directly from the determinant.

## 8.5 Alouini distribution example

For illustration we use the Alouini distribution [3] given by (6.2) for the sum of two correlated gamma variables in the case where $n = 2$, $\alpha = 1.6433$, $\lambda_1 = 12.32$, $\lambda_2 = 74.26$ and $\rho = 0.51$. The density for the Alouini distribution in this special case is given by

$$f(y) = r^{\alpha} \sum_{k=0}^{\infty} \delta_k \cdot \frac{y^{2\alpha+k-1} \exp(-y/\lambda_1)}{\lambda_1^{2\alpha+k} \Gamma(2\alpha+k)}$$

where $r = \lambda_1/\lambda_2$, $\delta_0 = 1$ and

$$\delta_k = \frac{\alpha}{k} \sum_{i=1}^{k} (1-r)^i \delta_{k-i}$$

for each $k \in \mathbb{N}$. We have the following results.

**Lemma 1.** *For each $k \in \mathbb{N}$ we have*

$$\delta_k = \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{k!} \, (1-r)^k.$$

**Proof** The proof is by induction. First of all we note that

$$\begin{aligned}
&\Delta \frac{(\alpha+j-1)(\alpha+j-2)\cdots(\alpha+1)}{(j-1)!} \\
&\quad = \frac{(\alpha+j)(\alpha+j-1)\cdots(\alpha+1)}{j!} - \frac{(\alpha+j-1)(\alpha+j-2)\cdots(\alpha+1)}{(j-1)!} \\
&\quad = \frac{(\alpha+j-1)(\alpha+j-2)\cdots(\alpha+1)}{j!} \, [(\alpha+j)-j] \\
&\quad = \frac{(\alpha+j-1)(\alpha+j-2)\cdots\alpha}{j!}.
\end{aligned}$$

It follows that

$$\sum_{j=0}^{m} \frac{(\alpha+j-1)(\alpha+j-2)\cdots\alpha}{j!} = \frac{(\alpha+m)(\alpha+m-1)\cdots(\alpha+1)}{m!}$$

for each $m \in \mathbb{N}$. We note that $\delta_1 = \alpha(1-r)$. Thus the result is true for $k = 1$. If we suppose the result is true for all $k \leq m$ then by the inductive hypothesis

$$\begin{aligned}
\delta_{m+1} &= \frac{\alpha}{m+1}\sum_{i=1}^{m+1}(1-r)^i\delta_{m+1-i} \\
&= \frac{\alpha}{m+1}\sum_{i=1}^{m+1}(1-r)^i\,\frac{(\alpha+m-i)(\alpha+m-i-1)\cdots\alpha}{(m+1-i)!}\,(1-r)^{m+1-i} \\
&= \frac{\alpha}{m+1}\sum_{j=0}^{m}\frac{(\alpha+j-1)(\alpha+j-2)\cdots\alpha}{j!}\cdot(1-r)^{m+1} \\
&= \frac{(\alpha+m)(\alpha+m-1)\cdots(\alpha+1)\alpha}{(m+1)!}\,(1-r)^{m+1}
\end{aligned}$$

as required. □

**Lemma 2.** *The mean of the Alouini distribution is given by*

$$\mu = \alpha(\lambda_1+\lambda_2) = 142.36.$$

**Proof** We have

$$\begin{aligned}
\mu &= r^\alpha\sum_{k=0}^{\infty}\frac{\delta_k}{\lambda_1^{2\alpha+k}\Gamma(2\alpha+k)}\int_0^\infty y^{2\alpha+k}\exp(-y/\lambda_1)dy \\
&= r^\alpha\lambda_1\sum_{k=0}^{\infty}(2\alpha+k)\delta_k.
\end{aligned}$$

We note that

$$\begin{aligned}
\sum_{k=0}^{\infty}\delta_k &= 1+\sum_{k=1}^{\infty}\frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{k!}(1-r)^k \\
&= (1-(1-r))^{-\alpha} \\
&= r^{-\alpha} \qquad (8.13)
\end{aligned}$$

and

$$\begin{aligned}\sum_{k=0}^{\infty} k\delta_k &= \sum_{k=1}^{\infty} k \cdot \frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{k!}(1-r)^k \\ &= \alpha(1-r)\sum_{\ell=0}^{\infty}\frac{(\alpha+\ell)(\alpha+\ell-1)\cdots(\alpha+1)}{\ell!}(1-r)^{\ell} \\ &= \alpha(1-r)\left(1-(1-r)\right)^{-(\alpha+1)} \\ &= \alpha(1-r)r^{-(\alpha+1)}. \end{aligned} \tag{8.14}$$

By substituting back into the original sum we find

$$\begin{aligned}\mu &= r^{\alpha}\lambda_1\left[2\alpha r^{-\alpha} + \alpha(1-r)r^{-(\alpha+1)}\right] \\ &= 2\alpha\lambda_1 + \alpha\lambda_1(1-\lambda_1/\lambda_2)\cdot\frac{\lambda_2}{\lambda_1} \\ &= \alpha(\lambda_1+\lambda_2)\end{aligned}$$

as required. □

**Lemma 3.** *The variance of the Alouini distribution is given by*

$$\sigma^2 = \alpha(\lambda_1^2+\lambda_2^2) = 9335.69.$$

**Proof** We have

$$\begin{aligned}
\sigma^2 &= r^\alpha \sum_{k=0}^{\infty} \frac{\delta_k}{\lambda_1^{2\alpha+k}\Gamma(2\alpha+k)} \int_0^\infty [y-\alpha(\lambda_1+\lambda_2)]^2 \, y^{2\alpha+k-1} \exp(-y/\lambda_1) dy \\
&= r^\alpha \sum_{k=0}^{\infty} \frac{\delta_k}{\lambda_1^{2\alpha+k}\Gamma(2\alpha+k)} \\
&\qquad \int_0^\infty \left[y^2 - 2\alpha(\lambda_1+\lambda_2)y + \alpha^2(\lambda_1+\lambda_2)^2\right] y^{2\alpha+k-1} \exp(-y/\lambda_1) dy \\
&= r^\alpha \sum_{k=0}^{\infty} \left\{\lambda_1^2(2\alpha+k+1)(2\alpha+k) - 2\alpha(\lambda_1+\lambda_2)\lambda_1(2\alpha+k) \right. \\
&\qquad \left. +\alpha^2(\lambda_1+\lambda_2)^2\right\} \delta_k \\
&= r^\alpha \sum_{k=0}^{\infty} \left\{k(k-1)\lambda_1^2 + [2(\alpha+1)\lambda_1^2 - 2\alpha\lambda_1\lambda_2]k + \alpha^2(\lambda_1-\lambda_2)^2 \right. \\
&\qquad \left. +2\alpha\lambda_1^2\right\} \delta_k.
\end{aligned}$$

We note that

$$\begin{aligned}
\sum_{k=0}^{\infty} k(k-1)\delta_k &= \sum_{k=2}^{\infty} k(k-1) \cdot \frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{k!}(1-r)^k \\
&= \alpha(\alpha+1)(1-r)^2 \sum_{\ell=0}^{\infty} \frac{(\alpha+\ell)(\alpha+\ell-1)\cdots(\alpha+2)}{\ell!} (1-r)^\ell \\
&= \alpha(\alpha+1)(1-r)^2 \left(1-(1-r)\right)^{-(\alpha+2)} \\
&= \alpha(\alpha+1)(1-r)^2 r^{-(\alpha+2)}. \qquad (8.15)
\end{aligned}$$

By recalling previous summation formulae (8.13) and (8.14) and the new formula (8.15) and by substituting back into the original sum we obtain

$$\begin{aligned}
\sigma^2 &= \alpha(\alpha+1)\lambda_1^2(1-r)^2 r^{-2} + \left[2(\alpha+1)\lambda_1^2 - 2\alpha\lambda_1\lambda_2\right]\alpha(1-r)r^{-1} \\
&\qquad + \left[\alpha^2(\lambda_1-\lambda_2)^2 + 2\alpha\lambda_1^2\right] \\
&= \alpha(\lambda_1^2+\lambda_2^2)
\end{aligned}$$

as required. □

The Alouini distribution is an infinite linear sum of gamma distributions with different shape parameters $(2\alpha + k)$ for each $k = 0, 1, 2, \ldots$ but with the same scale parameter $\lambda_1$. If we define a basic gamma density function in the form

$$\varphi_s(x) = \frac{x^{s-1}\exp(-x)}{\Gamma(s)}$$

for each $s > 0$ then the Alouini density can be written as

$$f(y) = r^{\alpha}\sum_{k=0}^{\infty}\delta_k\frac{\varphi_{2\alpha+k}(y/\lambda_1)}{\lambda_1}.$$

We set $n = 2$ and use the results obtained from Method 2 of Section 8.3.1. The values of the parameters are $\alpha = 1.6443$, $\beta_1 = 37.91$, $\beta_2 = 48.67$ and $\rho = 0.51$. We have

$$D = \begin{bmatrix} 37.91 & 0 \\ 0 & 48.67 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0.71 \\ 0.71 & 1 \end{bmatrix}$$

and

$$A = DC = \begin{bmatrix} 37.91 & 26.92 \\ 34.56 & 48.67 \end{bmatrix}$$

It follows that the eigenvalues of $A$ are $\lambda_1 = 12.32$ and $\lambda_2 = 74.26$. For convenience we calculate $r = \lambda_1/\lambda_2 = 0.165903582$ and hence find $r^{\alpha} = 0.05223829$. We used the first 50 terms of this series in MATLAB to sketch the approximate Alouini probability density function and the approximate Alouini cumulative distribution function. The density and CDF are shown in Figure 8.1.

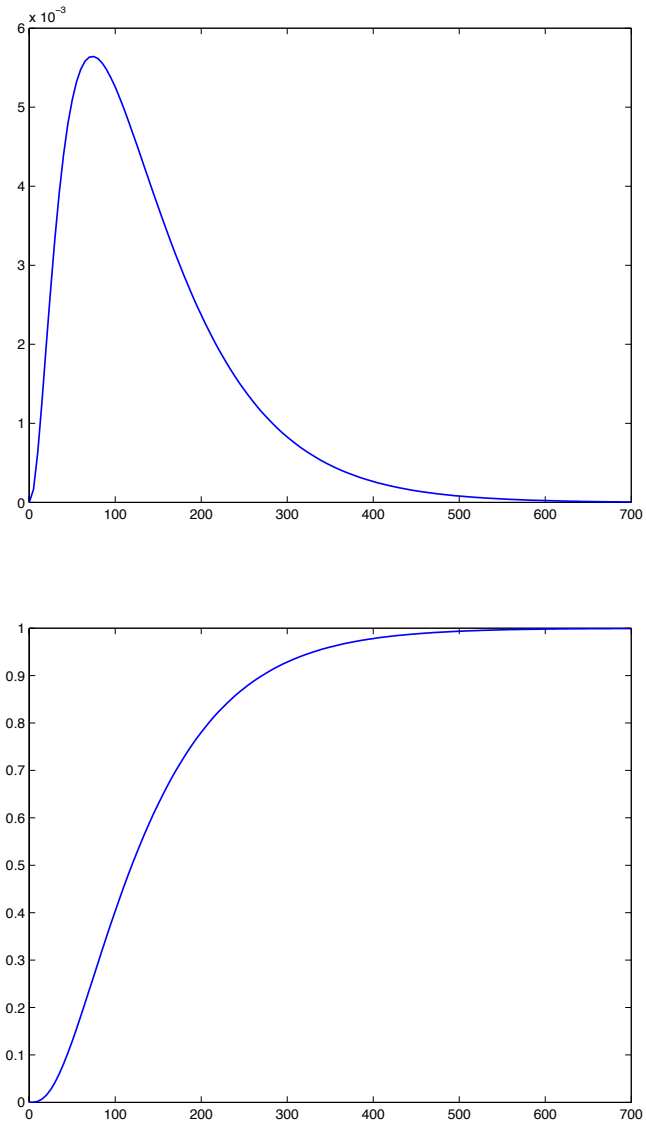



**Figure 8.1:** PDF (above) and CDF (below) of the sum of Alouini model

Hence, use the density given by equation (6.2) with the above parameter values to generate the simulated data for the rainfall totals using the Alouini model. The fitted Alouini data is formed using $\alpha = 1.644292, \lambda_1 = 12.22792$ and $\lambda_2 = 74.35130$.

The theoretical mean and variance are calculated as follows:

$$\begin{aligned} \text{Mean} &: \alpha(\lambda_1 + \lambda_2) = 1.644292(12.22792 + 74.35130) = 142.36 \\ \text{Variance} &: \alpha(\lambda_1^2 + \lambda_2^2) = 1.644292(12.22792^2 + 74.35130^2) = 9335.69 \end{aligned}$$

Then we compare the fit between the sum of the observed, fitted and generated rainfall totals using the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests. We also analyse the marginals of each observed and generated rainfall totals and assess the fit between the observed marginal with the generated marginal using $\alpha = 1.6443$, $\beta_1 = 37.91$ for Hume and $\alpha = 1.6443$, $\beta_1 = 37.91$ for Beechworth using the gamma distribution. Therefore, the analysis of results is divided into two parts– analysis of the sum of the two correlated random variables generated from the fitted Alouini distribution shown above and analysis of their marginals.

## 8.6 Result of the analysis of the sum of correlated gamma variables using Alouini model

Table 8.3 shows the means and variances of the observed sum and theoretical values matched. The generated data consisted of a sample of size 10000. The simulations are repeated for ten times and the average of the simulated means and variances are taken. The mean and variance of the generated data suggest that the mean and variance of the generated data are consistent with the theoretical values shown in the Table 8.3. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit test is based on the null hypothesis that the two distributions of observed and generated data for the sum of two correlated gamma variables are the same. Then the probability (P-value) is calculated and compared with the value of significance level. If the P-value calculated is greater than the significance level then the there

is not enough evidence to reject the null hypothesis. The P-value $< 0.05$ for both goodness of tests and the results suggest that the observed sum and generated sum are significantly different from one another. See Table 8.4. The plots of the analysis are presented in Figures 8.2, 8.3 and 8.4. It is found that the plots for both PDF and CDF of the generated and fitted matched perfectly as expected.

**Table 8.3:** Descriptive statistics of the sum of observed, fitted and the Alouini model

| Data | Observed | Formula | Generated |
|---|---|---|---|
| Mean | 142.36 | 142.36 | 144.38 |
| Variance | 9335.69 | 9335.69 | 9369.01 |

**Table 8.4:** Goodness of fit tests for the sum of correlated gamma variables

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| Observed Vs generated | 0.00007 | 0.00204 |

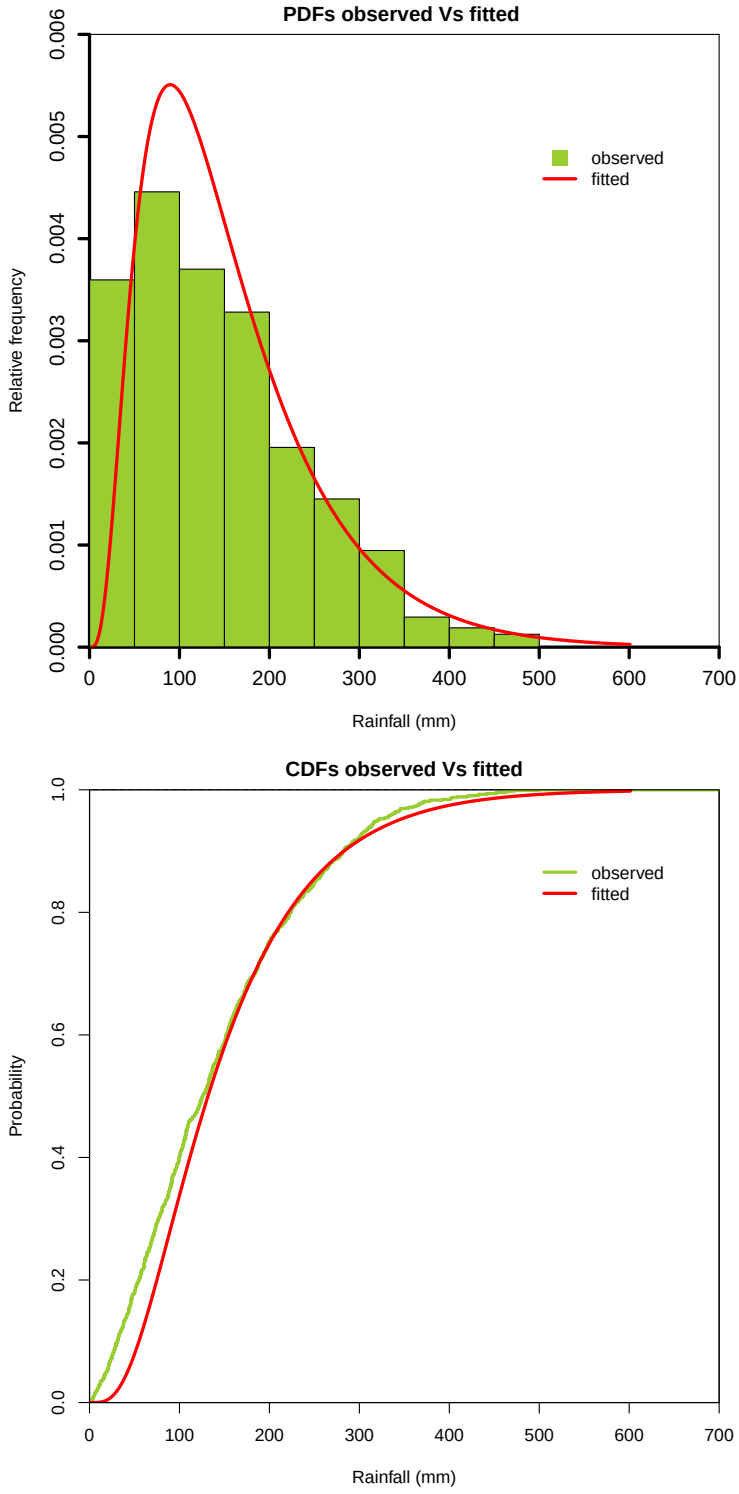



**Figure 8.2:** Observed vs fitted PDFs (above) and CDFs (below) for the sum of two correlated gamma variables

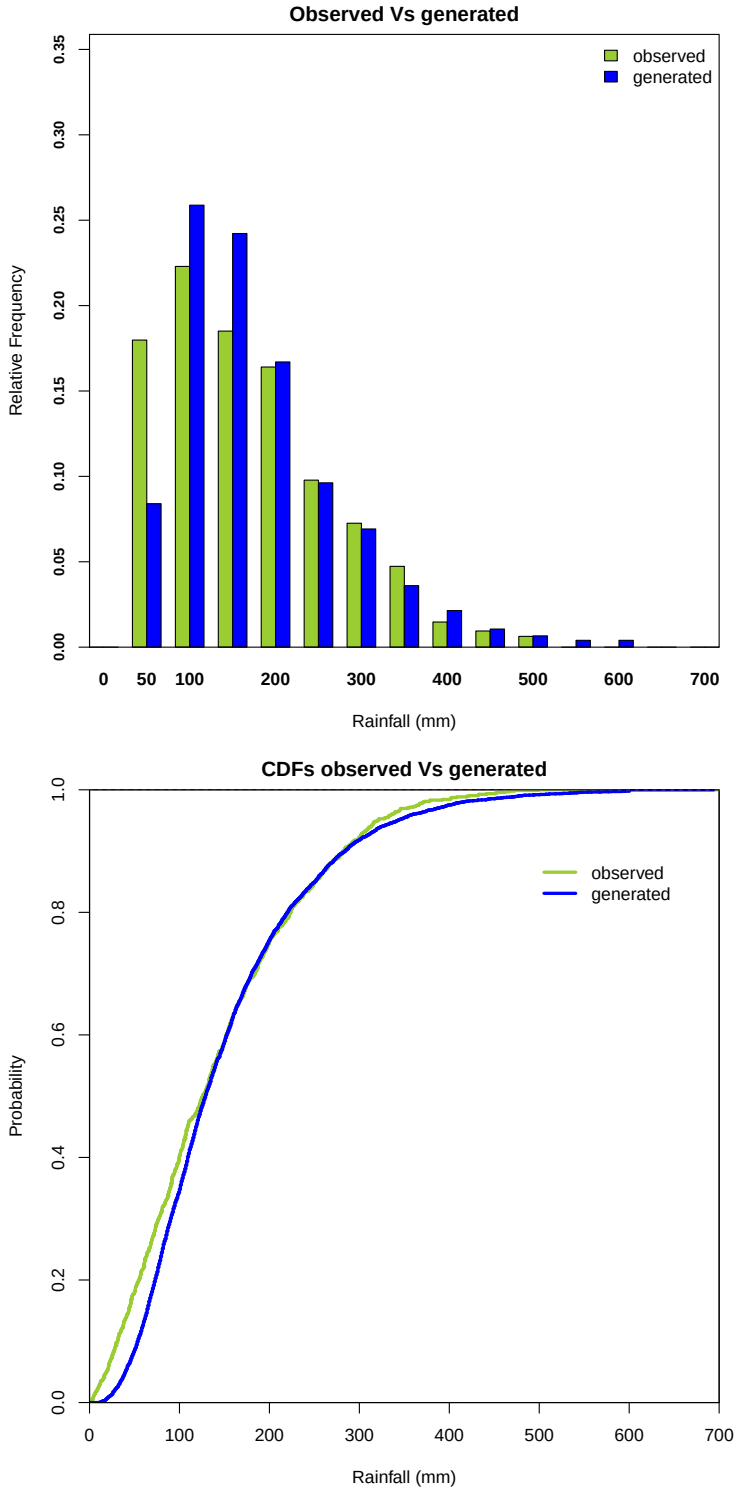



**Figure 8.3:** Observed vs generated PDFs (above) and CDFs (below) for the sum of two correlated gamma variables

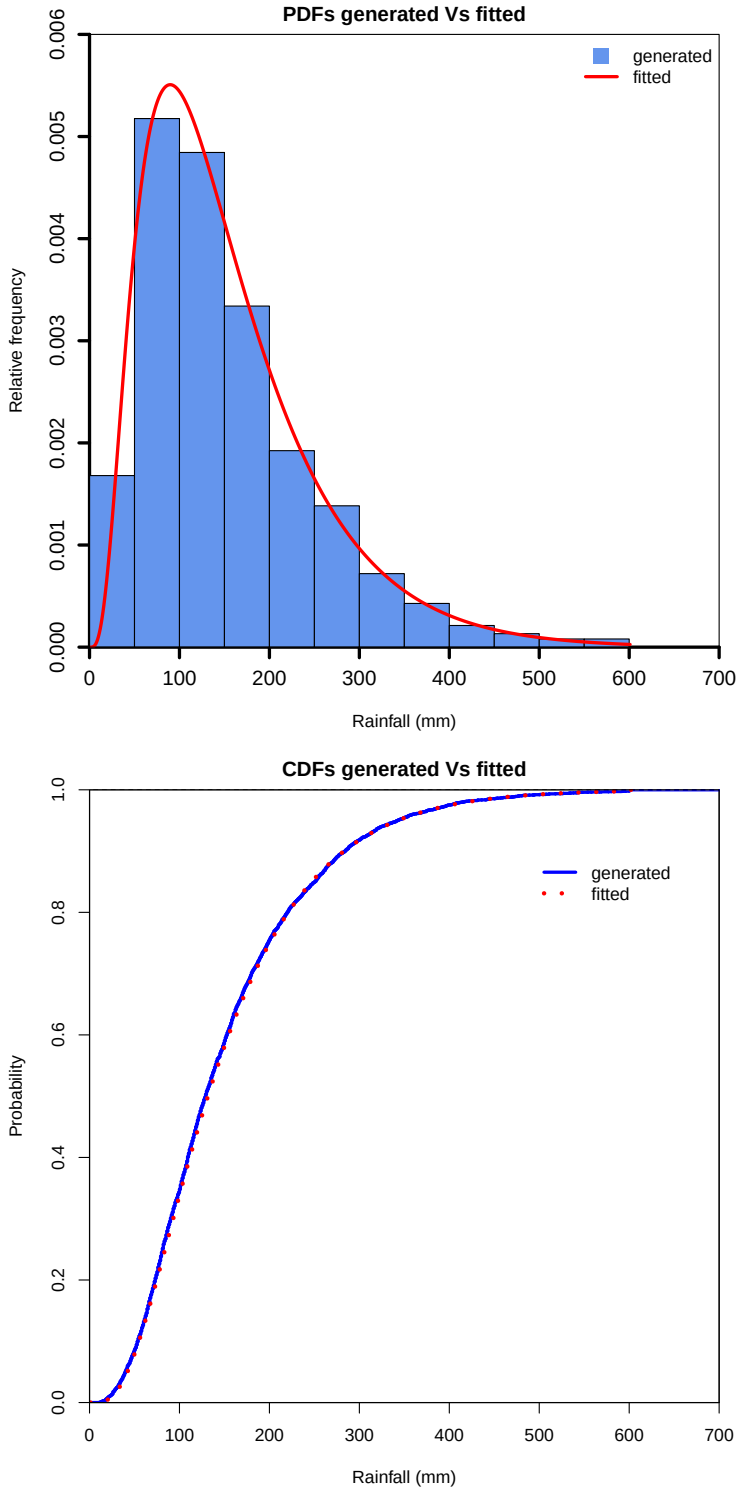



**Figure 8.4:** Generated vs fitted PDFs (above) and CDFs (below) for the sum of two correlated gamma variables

## 8.7 Result of the marginal analysis for Hume and Beechworth

The marginal analysis of means and variances of nonzero rainfall totals for Hume and Beechworth are given in Table 8.5. The goodness of fit tests show that the P-value $> 0.05$ for both Hume and Beechworth marginals. The results imply that the observed and fitted data are not significantly different from one another at 5% significant level. See Table 8.6. The graphical display of histogram and a fitted curve support the results (see Figures 8.1 and 8.5).

**Table 8.5:** Comparison of mean and variance of the observed and fitted data for nonzero rainfall totals

| | Hume | | Beechworth | |
|---|---|---|---|---|
| | observed | fitted | observed | fitted |
| mean | 58.64 | 61.59 | 83.72 | 78.78 |
| variance | 1789.51 | 2297.83 | 3315.24 | 3742.62 |

**Table 8.6:** Goodness of fit tests for the marginal analysis of the nonzero rainfall totals

| | Hume | Beechworth |
|---|---|---|
| Kolmogorov–Smirnov | 0.8655 | 0.7613 |
| Anderson–Darling | 0.5345 | 0.5046 |

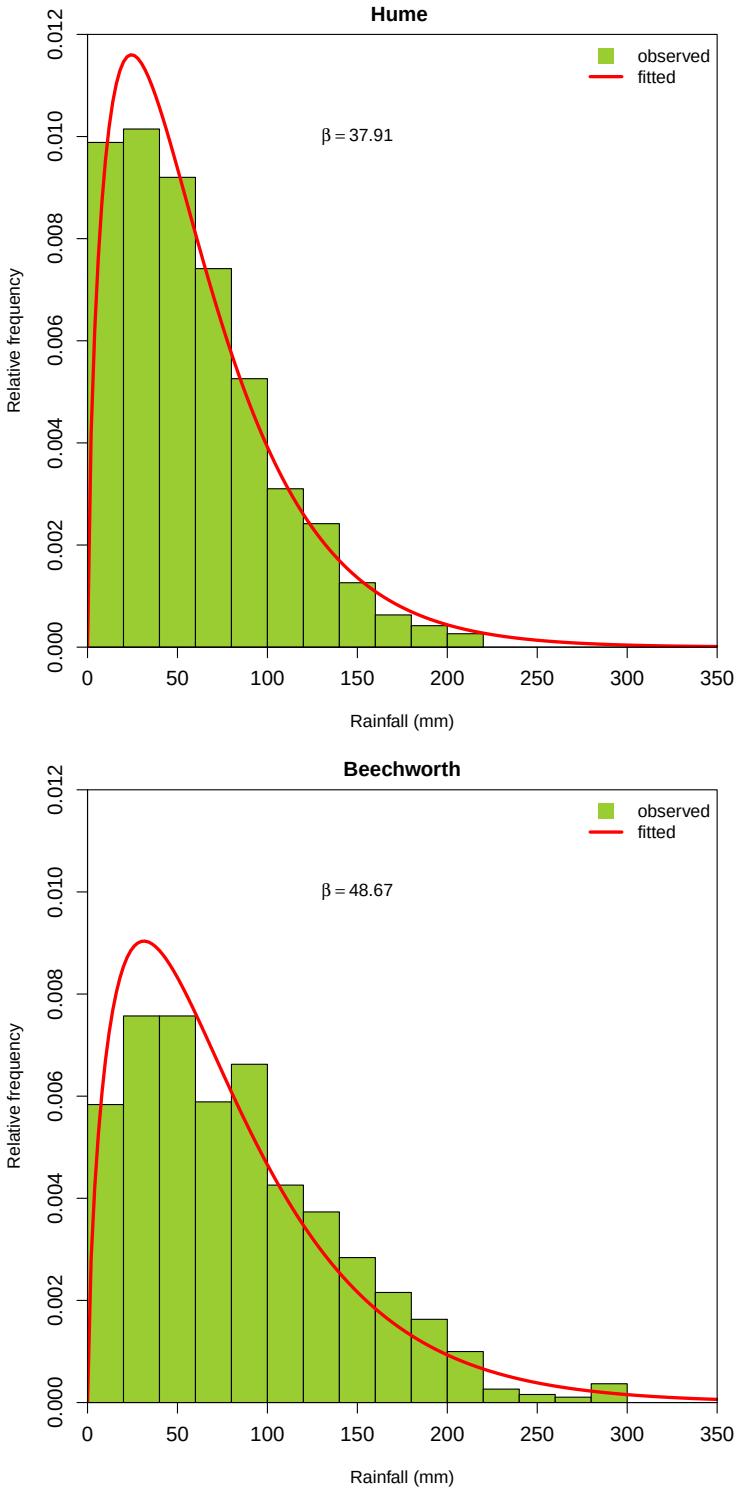



**Figure 8.5:** Observed and fitted PDFs of gamma marginals

## 8.8 Summary

We present two ways to estimate the shape and scale parameters for the sum of two correlated gamma variables using the moment generating method. The first model cannot be used because the equations are clearly inconsistent with the observed values. However, the second model gives a value for $\rho$ that is not consistent with the observations but can nevertheless be used to construct a legitimate density using the Alouini et al.[3] model. We found the discrepancy between the observed data and the generated data from Alouini et al.[3] model. The reason for the discrepancy is that the skewness and the kurtosis of the generated data do not match the observed data even though the mean and variance are closed. In conclusion, the moment generating method is also suitable for estimating the shape and scale parameters for the sum of two correlated gamma variables if the other descriptive values such as skewness and kurtosis can be matched. The marginal analysis between the observed and fitted data from Hume and Beechworth using the parameters estimated from the moment generating function show a good fit for both Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests. Hence, the moment generating function can be used to estimate the parameters.

# Chapter 9

# Bivariate Skew$-t$ Copula

In previous chapters we described the use of independent and correlated gamma variables in modelling the sum of rainfall totals in terms of spatial and temporal synthesis. The spatial synthesis refers to modelling the sum of individual rainfall totals over a collection of stations in the same water catchment whereas the temporal synthesis refers to the sum of monthly totals at a particular station to obtain a seasonal rainfall total. In this chapter, we use the bivariate skew-$t$ distribution to model rainfall totals from two selected meteorological stations in the Murray–Darling Basin namely Hume Reservoir and Beechworth Composite. In order to construct a suitable joint distribution, we model the individual totals using appropriate gamma distributions and use a bivariate skew-$t$ distribution to construct an appropriate copula. We investigate the characteristics of the bivariate skew-$t$ distribution and show how the parameters can be adjusted to generate simulated data which matches the observed data. Note that some parts of this chapter have been published in ANZIAM Journal (51, pp. C231–C246, 2010).

## 9.1 Copulas

Copula-based models have been used widely and successfully in the field of financial and insurance studies. For more information refer to [32, 65]. Recently the method has become more popular in hydrological and meteorological applications. See, for instance, the references [34, 37, 89, 90]. The copula method, which generates a general structure and can accommodate extreme events, has been used to model the interdependency. Copulas are functions that join multivariate distributions to their known one-dimensional marginal distribution functions. Copulas are discussed in the book by Nelsen [70]. The copula concept is based on Sklar's theorem developed by A. Sklar in 1959. A $p$-dimensional *copula* where $p \geq 2$, is a continuous, $p$-increasing, grounded probability distribution $C : [0,1]^p \mapsto [0,1]$ on the unit $p$-dimensional hyper-cube with uniform marginal probability distributions. That is

1. $C(u_1 + h_1, \ldots, u_p + h_p) \to C(u_1, \ldots, u_p)$ as $\|h\| = \sqrt{h_1^2 + \cdots h_p^2} \to 0$;

2. $C(u_1, \ldots, u_p) \leq C(u_1 + h_1, \ldots, u_p + h_p)$ when $h_j \geq 0$ for all $j = 1, 2, \ldots, p$;

3. $C(0, \ldots, 0) = 0$ and $C(1, \ldots, 1) = 1$; and

4. $\int_{[0,1]^{p-1}} C(u_1, \ldots, u_p)\, du_1 \ldots du_{k-1} du_{k+1} \ldots du_p = 1$ for each $k = 1, \ldots, p$ and each $u_i \in [0,1]^p$.

The copula $C$ defines a CDF on the unit $p$-dimensional hypercube $[0,1]^p$ for a $p$-dimensional random variable $(U_1, \ldots, U_p)$ with uniform marginal distributions. We write $U_k \sim U[0,1]$ for each $k = 1, 2, \ldots, p$.

We can use a copula to construct a joint distribution with known marginals. If

$C : [0,1]^p \to [0,1]$ is a $p$-dimensional copula and if $F_k : [0,\infty) \to [0,1]$ for each $k = 1, 2, \ldots, p$ are known 1-dimensional CDF's and we define $F : [0,\infty)^p \to [0,1]$ by the formula

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$$

then $F$ is a joint CDF for the $p$-dimensional random variable $(X_1, X_2, \ldots, X_p)$ on $(0,\infty)^p$ with marginal CDF defined by $F_k$ for each $k = 1, 2, \ldots, p$. That is

$$P[X_1 \leq x_1, \ldots, X_p \leq x_p] = F(x_1, \ldots, x_p)$$

and

$$P[X_k \leq x_k] = F(x_k)$$

for each $k = 1, 2, \ldots, p$. Note that $U_k = F_k(X_k)$ for each $k = 1, 2, \ldots, p$.

Some examples of copula types are the Gaussian copula and the Student's $t$ copula ($t$-copula). The Gaussian copula is based on a multivariate normal distribution and does not model tail dependence, whereas the $t$-copula is based on the multivariate Student's $t$ distribution and does accommodate tail dependence. The tail dependence is used to define the degree of dependence in the lower or upper tails of a bivariate distribution. This concept is widely applied in finance to model dependence of loss events with various assets [32].

Kollo and Pettere [56] show the application of the multivariate skew-$t$ copula using simulation techniques since the expression of the probability density function and distribution of the skew-$t$ are complex. They use human measurement data to study linear correlation of height, waist and bust circumference measures from primary school students in Riga, Latvia. The gamma and lognormal distributions were used to estimate the marginal distribution for each data sets. The results of the study demonstrated that the gamma distribution gave the best fit for height and waist circumference whereas lognormal is best for bust circumference.

In this study, the asymmetric $t$-copula, also known as skew-$t$, is employed to analyse asymmetric tail dependence using monthly rainfall totals. The skew-$t$ distribution is considered robust as it includes both skewness and tail dependence structure and allows us to model correlations.

## 9.2 Using skew-$t$ copula to model rainfall totals

We discuss the theory of the development of our rainfall model. We choose a skew-$t$ copula based on the reasons mentioned in Section 9.1. In this section, we present the theory of univariate Student's $t$ distribution which is the basis of the univariate skew-$t$ followed by the univariate and multivariate skew-$t$ distributions.

### 9.2.1 Univariate Student's $t$ distribution

The Student's $t$ distribution was introduced by W. S. Gosset in 1908 [110] in the article titled "The probable error of a mean". Recently, Zabell [110] writes a review of the impact of Gosset's work in statistical theory and practice after a century. The skew-$t$ distribution is constructed based on the Student's $t$ distribution.

A random variable X has a Student's $t$ distribution with $\nu$ degrees of freedom denoted by $X \sim t(\nu)$ with the probability density function given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

where $x \in (-\infty, \infty)$ and $\Gamma(\cdot)$ defines a gamma function. The Student's $t$ PDF can also be written in the Beta function form of

$$f(x) = \frac{1}{\sqrt{\nu}B(1/2,\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

where $B(\cdot)$ is the Beta function defined by

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt = B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

with $\sqrt{\pi} = \Gamma(1/2)$. The cumulative distribution function of the Student's $t$ is

$$F(x) = \frac{1}{2} + x\ \Gamma\left(\frac{\nu+1}{2}\right)\ \frac{{}_2F_1(1/2,\ (\nu+1)/2;\ 3/2;\ -x^2/\nu)}{\sqrt{\nu\pi}\ \Gamma(\nu/2)}$$

where ${}_2F_1(\cdot)$ is the confluent hypergeometric function given in [1] by formula (15.1.1) on page 556. The general form of the confluent hypergeometric series is

$$\begin{aligned} {}_1F_1(a;b;z) &= 1 + \frac{a}{b}z + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + \cdots \\ &= \Sigma_{k=0}^{\infty}\frac{(a)_k}{(b)_k}\frac{z^k}{k!} \end{aligned}$$

where $(a)_k$ and $(b)_k$ are known as Pochhammer symbols. Note that the Student's $t$ distribution has the bell shape distribution and only depends on the number of degrees of freedom ($\nu$) with mean zero and variance one.

### 9.2.2 Univariate skew-$t$ distribution

A random variable X has a skew-$t$ distribution, $X \sim ST(\mu,\omega,\gamma,\nu)$ if it has a probability density function given by

$$f_{ST}(x;\mu,\omega,\gamma,\nu) = \frac{2}{\omega}\ t(z;\nu)\ T\left\{\gamma z\sqrt{\frac{\nu+1}{\nu+z^2}};\ \nu+1\right\}, \quad x \in \mathbb{R}$$

where $z = (x-\mu)/\omega$ is a scaled variable, $t(\cdot)$ and $T(\cdot)$ are the PDF and CDF respectively for the univariate Student's $t$-distribution with $\nu+1$ degrees of freedom, $\mu \in \mathbb{R}$ is a location parameter, $\omega \in (0,\infty)$ is a scale parameter and $\gamma \in \mathbb{R}$ is a shape parameter and where the distribution has $\nu$ degrees of freedom. Skewness and tail

proportions are controlled by parameters $\gamma$ and $\nu$, respectively. When $\gamma = 0$, the standard $t$ distribution is obtained and when $\nu \to \infty$, the skew normal distribution is formed. However when both $\gamma = 0$ and $\nu \to \infty$, the normal distribution is recovered. We refer the reader to [10] for further details.

### 9.2.3 Multivariate skew-$t$ distribution

A higher dimensional skew-$t$ distribution to handle multivariate data is written as $MST(\mu, \mathbf{\Omega}, \gamma, \nu)$ which has a probability density function of $x \in \mathbb{R}^d$,

$$f_{MST}(x; \mu, \mathbf{\Omega}, \gamma, \nu) = 2t_d(x - \mu; \mathbf{\Omega}, \nu)\ T\left\{\gamma^T \mathbf{\Omega}^{-1}(x - \mu)\sqrt{\frac{\nu + d}{Q(x) + \nu}};\ \nu + d\right\}$$

where $Q(x) = (x - \mu)^T \mathbf{\Omega}^{-1}(x - \mu)$. The $d$ - dimensional Student's $t$ distribution with zero location ($\mu = 0$), $\mathbf{\Omega}$ scale matrix and $\nu$ degrees of freedom is defined as

$$t_d(x; \mathbf{\Omega}, \nu) = \frac{\Gamma[(\nu + d)/2]}{\sqrt{|\mathbf{\Omega}|}\ (\nu\pi)^{d/2}\ \Gamma(\nu/2)}\left(1 + \frac{Q(x)}{\nu}\right)^{-\frac{\nu+d}{2}}, \quad x \in \mathbb{R}^d$$

and $T(\cdot)$ is the cumulative distribution function of Student's $t$ distribution with $(\nu + d)$ degrees of freedom. We refer the reader to [9] for further details.

## 9.3 Simulation experiment

As an illustration for the simulation experiment, we use a bivariate skew-$t$ distribution. The procedure is divided into three parts: Part 1 – fitting data to gamma marginals, Part 2 – generating simulated data using the skew-$t$ distribution and Part 3 – comparing the analysis from Part 1 and Part 2. The $R$ codes for the data analysis using the skew-$t$ distribution is given in Appendix C.9.

### 9.3.1 Fitting data to gamma marginals

We select two sets of monthly rainfall totals from 1928 to 1985 for two adjacent meteorological stations in the Murray–Darling Basin, Hume and Beechworth. The stations are 73 km apart with lag zero cross-correlation of 0.88.

We categorise the rainfall data into four difference cases – (wet, wet), (wet, dry), (dry, wet) and (dry, dry). For illustration, we consider only the case (wet, wet) when both stations have rain. Therefore, we have pairs of positive numbers. The scatter plot between the positive pairs of Hume and Beechworth is shown in Figure 9.1. The bivariate distribution of the data demonstrate the tail dependence.

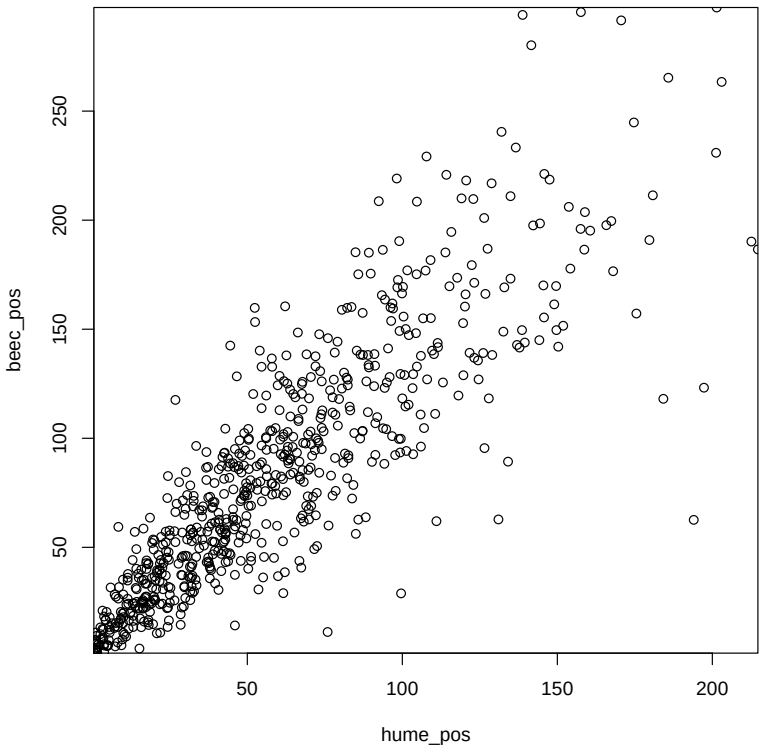



**Figure 9.1:** Bivariate scatter plot of positive pairs for Hume and Beechworth

The individual sets of rainfall data are fitted separately using gamma marginals. The gamma distribution is chosen as it is suitable to model continuous variables, like rainfall totals, that are always positive and have a skewed distribution. Furthermore, it is also flexible as it involves two parameters, scale and shape. The probability density function for the gamma distribution is given by formula (3.1) where the shape parameter $\alpha$ and the scale parameter $\beta$ for each month are estimated using the maximum likelihood method. Refer in Table 9.1. These sets of monthly rainfall totals between Hume and Beechworth are then transformed into bivariate uniform data using the fitted cumulative gamma density function. Some examples of data transformation are given in Table 9.2.

**Table 9.1:** The monthly values of shape parameter $\alpha$ and scale parameter $\beta$

| | Hume | | Beec | |
|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| Jan | 0.90 | 48.39 | 1.22 | 42.89 |
| Feb | 1.01 | 44.26 | 1.07 | 49.11 |
| Mar | 1.24 | 37.34 | 1.40 | 44.38 |
| Apr | 1.85 | 29.93 | 1.87 | 42.97 |
| May | 1.85 | 33.50 | 2.06 | 46.07 |
| Jun | 2.45 | 26.54 | 2.74 | 36.75 |
| Jul | 3.67 | 20.26 | 4.44 | 25.28 |
| Aug | 3.02 | 25.73 | 3.46 | 33.80 |
| Sep | 3.06 | 20.14 | 3.75 | 24.76 |
| Oct | 2.33 | 31.57 | 3.24 | 31.11 |
| Nov | 1.86 | 27.38 | 2.12 | 32.24 |
| Dec | 0.96 | 50.32 | 1.52 | 40.21 |

**Table 9.2:** Some values of observed data transformed to uniform data using the CDF of gamma. The Hume data is transformed using $\alpha = 1.5382$ and $\beta = 38.2018$ and for Beechworth the values are $\alpha = 1.7719$ and $\beta = 47.0027$.

| | Hume | | Beechworth | |
|---|---|---|---|---|
| | observed | uniform | observed | uniform |
| 1 | 49.7 | 0.5290 | 54.1 | 0.3915 |
| 2 | 141.6 | 0.9367 | 280.2 | 0.9876 |
| 3 | 119.7 | 0.8953 | 128.9 | 0.8080 |
| 4 | 27.2 | 0.2869 | 70.0 | 0.5125 |
| 5 | 56.0 | 0.5845 | 100.5 | 0.6940 |
| 6 | 58.1 | 0.6017 | 79.3 | 0.5750 |
| 7 | 59.2 | 0.6105 | 83.2 | 0.5993 |
| 8 | 22.8 | 0.2333 | 30.8 | 0.1925 |
| 9 | 28.0 | 0.2966 | 79.9 | 0.5788 |
| 10 | 69.9 | 0.6878 | 118.3 | 0.7709 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 682 | 46.6 | 0.4997 | 128.4 | 0.8064 |
| 683 | 49.6 | 0.5281 | 79.4 | 0.5756 |
| 684 | 84.2 | 0.7698 | 78.6 | 0.5705 |
| 685 | 68.2 | 0.676481 | 80.0 | 0.5794 |
| 686 | 48.6 | 0.518756 | 58.0 | 0.4227 |
| 687 | 160.6 | 0.959301 | 195.2 | 0.9399 |
| 688 | 31.6 | 0.3391 | 25.8 | 0.1499 |
| 689 | 94.8 | 0.817333 | 104.2 | 0.7116 |
| 690 | 62.6 | 0.636675 | 138.0 | 0.8355 |
| 691 | 67.6 | 0.672406 | 120.4 | 0.7787 |

The correlation and the tail proportions of the uniform data for Hume and Beechworth are calculated (refer Table 9.3). The lower and upper 10% tail proportions are

obtained by subsetting the number of pairs for Hume and Beechworth within those proportions. These tail proportions also describe the amount of dependence in the tail of a bivariate distribution. Figure 9.2 shows the scatter plot of uniform data for Hume and Beechworth. The bottom-left and top-right corners of the Figure 9.2 are the 10% lower and upper tail proportions, respectively. In the following section, we want to generate simulated data using the bivariate skew-$t$ distribution.

**Table 9.3:** Upper and lower tail proportions and correlation coefficient to measure the degrees of dependence between bivariate data of Hume and Beechworth

| | Tail proportions | | Correlation (r) |
|---|---|---|---|
| | lower 10% | upper 10% | |
| Observed | 0.081 | 0.056 | 0.893 |

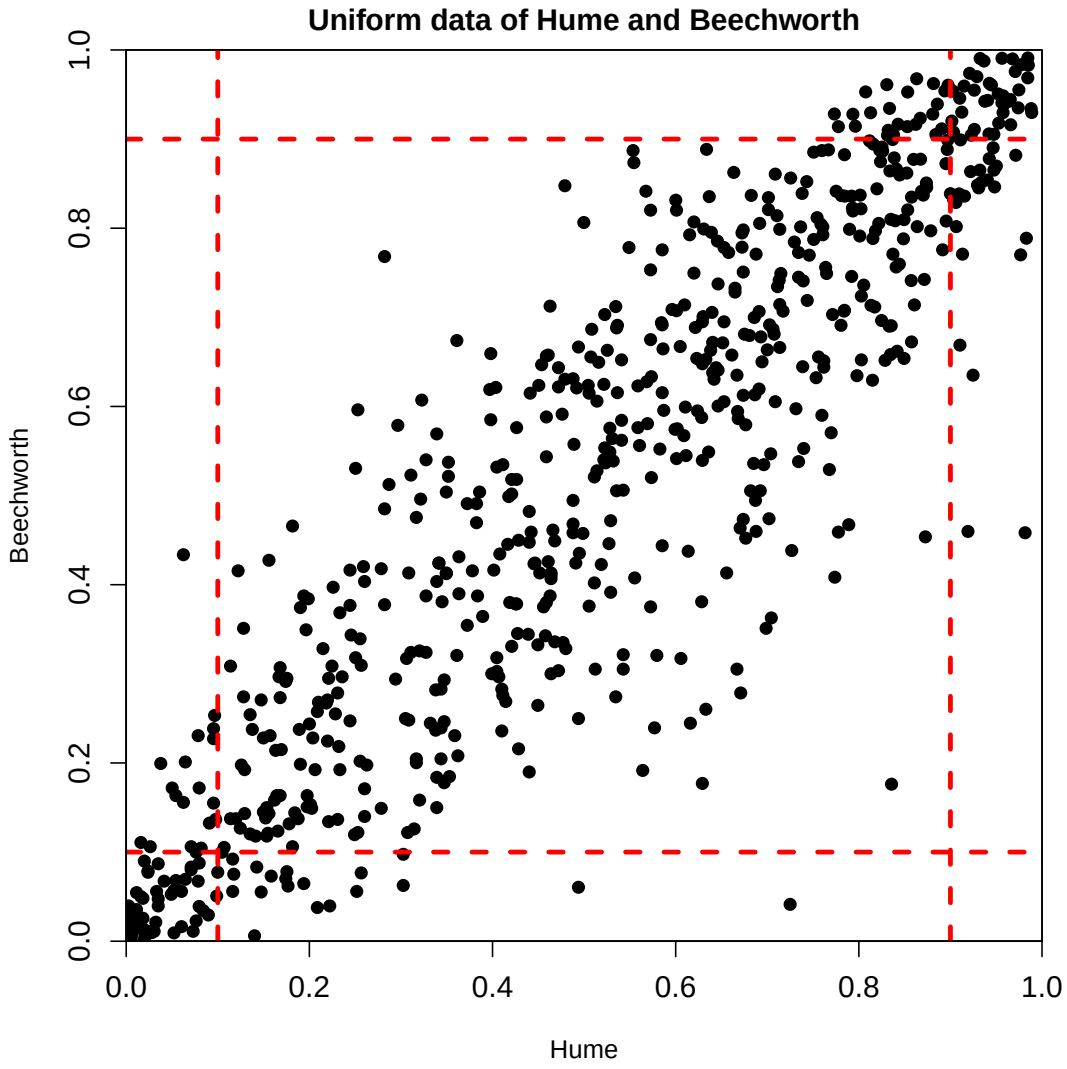



**Figure 9.2:** Scatter plot of bivariate uniform data of Hume and Beechworth

### 9.3.2 Generation of simulated data using skew-$t$ distribution

For simulation purposes, we use code from the '*sn*' package in *R* developed by Azzalini [8]. The '*sn*' stands for skew normal and originally the codes were developed for the skew normal distribution. Then, Azzalini [8] developed and included the

codes for the skew-$t$ distribution in the same package. The codes used are probability density function, distribution function and random number generation for the skew-$t$ distribution (univariate: dst, pst, rst; multivariate: pmst, dmst, rmst). The freely available codes can be found from the $R$ website (http://cran.r-project.org).

We generate $10^5$ simulated data values for the bivariate skew-$t$ using random number generation of multivariate skew-$t$, $\text{rmst}(n, \mu_i, \mathbf{\Omega}, \gamma_i, \nu_i)$ for $i = 1, 2$. The mean and covariance are set to $\mu_1 = \mu_2 = 0$, $\omega_{11} = \omega_{22} = 1$ and $\omega_{12} = \omega_{21} = 0$. These values can be chosen arbitrarily as the effects are removed when we transform the data to uniform. Therefore, there are three free parameters namely; skewness ($\gamma$), correlation ($\rho$) and degrees of freedom ($\nu$). By varying the free parameters, we would like to investigate the effect on correlation ($r$) and tail proportions of the lower and upper 10% of the simulated data. The main objective of this exercise is to match the correlation coefficient of the observed data with that of the simulated data. The set of random numbers generated by the bivariate skew-$t$ distribution can be between $-\infty$ and $+\infty$. Each set of bivariate skew-$t$ data generated is transformed to uniform data using marginals of univariate skew-$t$ distribution, $\text{pst}(n, \mu = 0, \omega = 1, \gamma, \nu)$. Then, the correlation ($r$) and the tail proportions (lower 10% and upper 10%) for the simulated uniform data are calculated.

### 9.3.3 Analysis of the simulation process

The analysis of the simulation process is divided into two parts as follows.

**(i) The effect of parameter variation on correlation and tail proportions of the simulated data**

In each simulation, we generate $10^5$ simulated data values. Firstly, the values of parameters are varied systematically and we try various combinations such as in-

creasing the skewness value but with correlation and degrees of freedom parameters are fixed. Then we do the same by increasing the degrees of freedom and fix correlation and skewness parameters. In each iteration, the value of simulated correlation and the tail proportions are noted. Based on the simulated results we finally decide that we should focus on finding the best values for the simulated parameters by varying the values of skewness ($\gamma$) from 0 to 2, degrees of freedom ($\nu$) from 3.1 to infinity and model correlation ($\rho$) from 0.8 to 0.99. It is found that varying the degrees of freedom has only a slight effect on the simulation results. Refer to Table 9.5. Therefore, we concentrate on variation of the skewness and correlation parameters. If we increase the skewness, it will decrease the correlation of the uniformised simulated data (see Table 9.4).

In order to match the correlation of the uniformised simulated data with the uniformised observed data, the correlation ($\rho$) chosen for the simulated data needs to be as high as possible where $0.9 < \rho < 1$. Note that, $\rho$ is the model correlation which we use to generate synthetic data with correlation $r$ that matches the observed correlation. In Table 9.4, we observe that as the value of skewness increases, the estimated values for the correlation decreases. However, there is little effect on the value of the estimated correlation when the degrees of freedom varies (see Table 9.5). The correlation for simulated data increases as the correlation ($\rho$) increases (see Table 9.6).

**Table 9.4:** Variation of skewness ($\gamma$)

| Model parameters | | | Simulation results | | |
|---|---|---|---|---|---|
| $\rho$ | $\gamma$ | $\nu$ | $r$ | lower 10% | upper 10% |
| 0.89 | 0.0 | 5 | 0.8721 | 0.0696 | 0.0696 |
| | 0.5 | | 0.8273 | 0.0594 | 0.0679 |
| | 1.0 | | 0.7699 | 0.0471 | 0.0670 |
| | 1.5 | | 0.7262 | 0.0368 | 0.0665 |
| | 2.0 | | 0.7091 | 0.0292 | 0.0668 |

**Table 9.5:** Variation of degrees of freedom ($\nu$)

| Model parameters | | | Simulation results | | |
|---|---|---|---|---|---|
| $\rho$ | $\gamma$ | $\nu$ | $r$ | lower 10% | upper 10% |
| 0.89 | 1.5 | 3.1 | 0.7389 | 0.0381 | 0.0669 |
| | | 3.5 | 0.7252 | 0.0358 | 0.0700 |
| | | 4 | 0.7304 | 0.0370 | 0.0678 |
| | | 5 | 0.7299 | 0.0371 | 0.0669 |
| | | 10 | 0.7281 | 0.0362 | 0.0643 |
| | | Inf | 0.7270 | 0.0359 | 0.0627 |

**Table 9.6:** Variation of model correlation ($\rho$)

| Model parameters | | | Simulation results | | |
|---|---|---|---|---|---|
| $\rho$ | $\gamma$ | $\nu$ | $r$ | lower 10% | upper 10% |
| 0.80 | 1.5 | 5 | 0.556 | 0.023 | 0.056 |
| 0.85 | | | 0.651 | 0.029 | 0.066 |
| 0.90 | | | 0.752 | 0.039 | 0.069 |
| 0.95 | | | 0.867 | 0.055 | 0.077 |
| 0.98 | | | 0.945 | 0.069 | 0.086 |
| 0.99 | | | 0.971 | 0.078 | 0.090 |

### (ii) Refining the simulation result

We would like to match the statistics of the uniformised simulated data with the uniformised observed data as closely as possible. Based on the analysis of simulation results in the previous section, the simulation results can be refined further by repeating the same process by carefully choosing the right combination of the parameters. Therefore, the simulation process is reiterated to refine the simulation results with a sample size of $10^4$. Each simulation is repeated five times and the mean of the results are calculated. The set of parameters that best match the observed statistics are $\rho = 0.965, \gamma_1 = \gamma_2 = -1.95$ and $\nu = 5$ which give the simulated correlation $r = 0.892$ and tail proportions of 0.082 and 0.055, for lower 10% and upper 10% respectively. Refer to Table 9.7. In comparison, the statistics of the observed data are correlation 0.893 and with tail proportions of 0.081 and 0.056, for lower 10% and upper 10%, respectively.

**Table 9.7:** Refining simulation process

| Model parameters | | | Simulation results | | |
|---|---|---|---|---|---|
| $\rho$ | $\gamma$ | $\nu$ | $r$ | lower 10% | upper 10% |
| 0.90 | 2.0 | 5 | 0.733 | 0.031 | 0.065 |
| | -0.1 | 5 | 0.881 | 0.069 | 0.070 |
| | -2.0 | 5 | 0.732 | 0.069 | 0.032 |
| 0.91 | -0.1 | 5 | 0.894 | 0.074 | 0.071 |
| | -1.0 | 5 | 0.802 | 0.068 | 0.050 |
| | -1.5 | 5 | 0.767 | 0.070 | 0.041 |
| 0.92 | -0.1 | 5 | 0.906 | 0.075 | 0.074 |
| | -0.3 | 5 | 0.891 | 0.073 | 0.067 |
| | -1.5 | 5 | 0.790 | 0.071 | 0.045 |
| 0.93 | -0.1 | 5 | 0.914 | 0.079 | 0.073 |
| | -0.4 | 5 | 0.895 | 0.074 | 0.069 |
| | -1.5 | 5 | 0.809 | 0.073 | 0.043 |
| 0.94 | -0.1 | 5 | 0.928 | 0.082 | 0.078 |
| | -0.6 | 5 | 0.891 | 0.073 | 0.065 |
| | -1.5 | 5 | 0.842 | 0.075 | 0.051 |
| 0.95 | -0.3 | 5 | 0.932 | 0.080 | 0.079 |
| | -0.8 | 5 | 0.898 | 0.075 | 0.065 |
| | -2.0 | 5 | 0.857 | 0.082 | 0.049 |
| | -2.1 | 5 | 0.849 | 0.077 | 0.046 |
| 0.96 | -0.3 | 5 | 0.946 | 0.082 | 0.082 |
| | -1.3 | 5 | 0.895 | 0.076 | 0.064 |
| | -2.0 | 5 | 0.880 | 0.079 | 0.053 |
| 0.965 | -1.0 | 5 | 0.917 | 0.084 | 0.070 |
| | -1.5 | 5 | 0.904 | 0.080 | 0.061 |
| | -1.8 | 5 | 0.895 | 0.081 | 0.054 |
| | -1.9 | 5 | 0.892 | 0.082 | 0.053 |
| | -1.95 | 5 | 0.892 | 0.082 | 0.055 |
| observed values | | | 0.893 | 0.081 | 0.056 |

### 9.3.4 Fitting the bivariate skew$-t$ copula after refining the simulation process

Using copula, the observed data is transformed to uniform data on [0,1] and it is almost impossible to revert to the original data. Therefore, the goodness of fit test is used to measure the fit between the uniformised observed data and the uniformised theoretical data. In this section we use the bivariate skew$-t$ copula to generate bivariate uniform rainfall totals and compare with the observed bivariate uniform rainfall totals. After refining the simulation process, we generate 10000 data using parameters value of $\rho = 0.9965$ (correlation), $\nu = 5$ (degrees of freedom), $\xi = 0$ (mean of zero) and $\gamma = -1.95$ (skewness). Then, both the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are applied to assess the fit.

## 9.4 Results

The discussion of the results are divided into two parts: 1. Analysis of the uniformised rainfall totals and the generated data, 2. Analysis of empirical copula versus theoretical copula.

**Analysis of the uniformised rainfall totals and the generated data**

The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are applied to test the null hypothesis that the uniformised observed and generated data are not significantly different at 5% significance level. Table 9.8 shows that the P-values $> 0.05$ for both tests, therefore, the uniformised observed and generated data are not significantly different from one another at 5% significance level. These results are supported by the graphical representation shown in Figure 9.3.

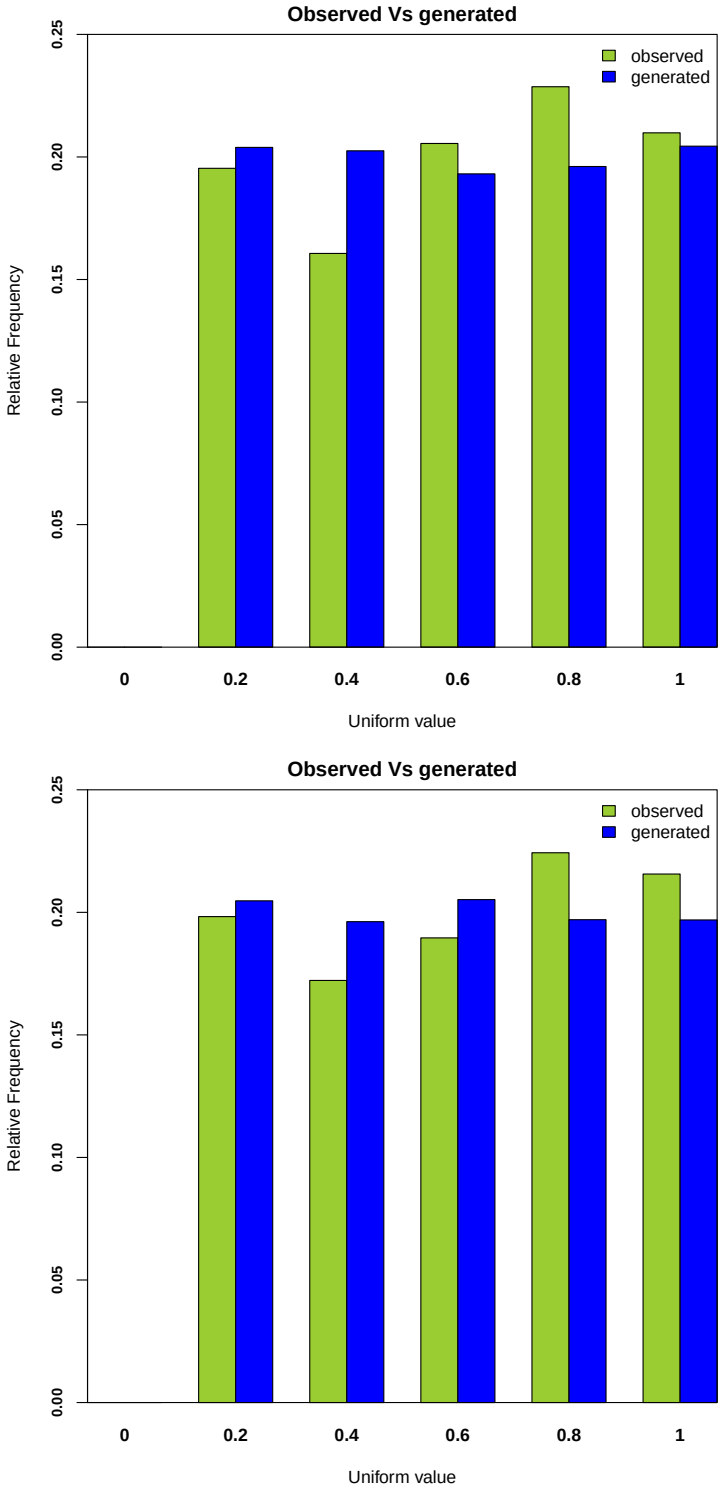



**Figure 9.3:** Comparison of the uniformised observed and generated, Hume (above) and Beechworth (below) stations

**Table 9.8:** Goodness of fit tests of theoretical and empirical copulas for the uniformised observed and generated data

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| Hume | 0.0542 | 0.0677 |
| Beechworth | 0.0810 | 0.1073 |

### Analysis of empirical copula versus theoretical copula

We use a graphical method and goodness of fit tests to evaluate the fit between the observed and the generated data. A scatter plot of empirical copula versus theoretical copula is presented. A similar graphical representation is also used by [106] and [36]. The empirical copula ($C_n$) is constructed using the observed data which is transformed to uniform data on [0,1] using the gamma marginals as described in Section 9.3.1. Then the uniform data is ranked and used to calculate $C_n$. For the theoretical copula ($\widetilde{C}$), we generate $10^4$ bivariate data values from the bivariate skew-$t$ model which is then transformed to uniform using the univariate skew-$t$ for each marginal. The generated uniform data is also ranked and used to calculate $\widetilde{C}$. The empirical copula is

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} I\left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v\right)$$

where $n$ is the sample size and $R_i$ and $S_i$ are the rank for each marginal. In our example we only deal with bivariate data, however the formula can easily be extended to multivariate. Wong et al. [106] apply the trivariate copulas to model duration, peak intensity and average intensity of the drought risk analysis using rainfall data from two stations on the east and west of the Great Dividing Range of Australia. $I(A)$ defines the indicator variable for a logical expression A, where $I(A) = 1$ if A

is true or $I(A) = 0$ if A is false. The axes are rescaled on $[0, 1]^2$ and a set of points within that configuration is calculated. For example, the point (0.025, 0.023) on the scatter plot of theoretical versus empirical copula (see Figure 9.4) is calculated from

$$(1/691) \sum_{i=1}^{691} I \left( \frac{R_i}{691 + 1} \leq 0.2, \frac{S_i}{691 + 1} \leq 0.3 \right)$$

and

$$(1/10000) \sum_{i=1}^{10000} I \left( \frac{R_i}{10000 + 1} \leq 0.2, \frac{S_i}{10000 + 1} \leq 0.3 \right).$$

The line $y = x$ is inserted and the fitness of the empirical and theoretical copula is based on the closeness of the points to this line. However, the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests are used to compare the distributions of the empirical and theoretical copula, see Table 9.9. The P-values are found to be greater than 0.05 significance level for both tests, therefore we conclude that the two distributions are not significantly different at 5% significance level. Figure 9.5 also support these results.

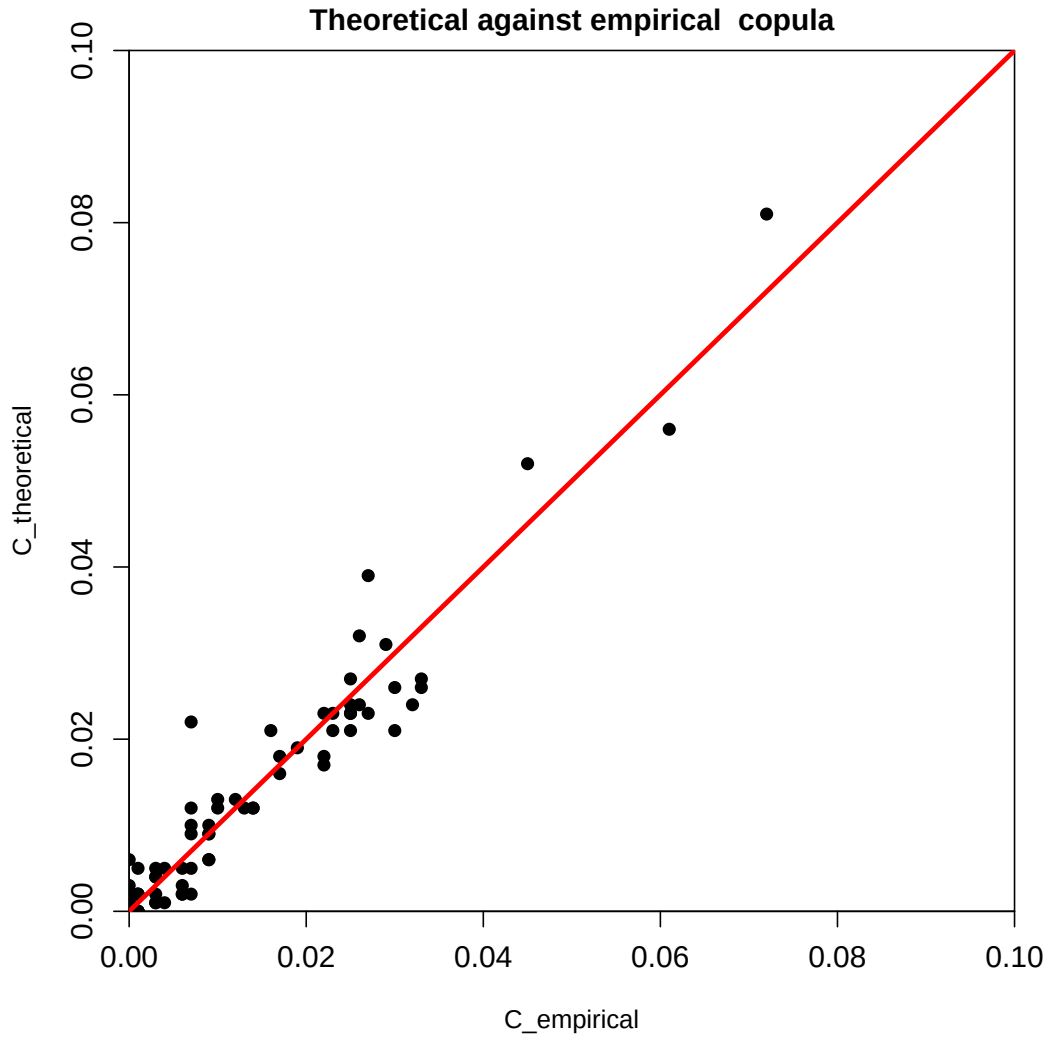



**Figure 9.4:** Comparison of the theoretical and empirical copulas

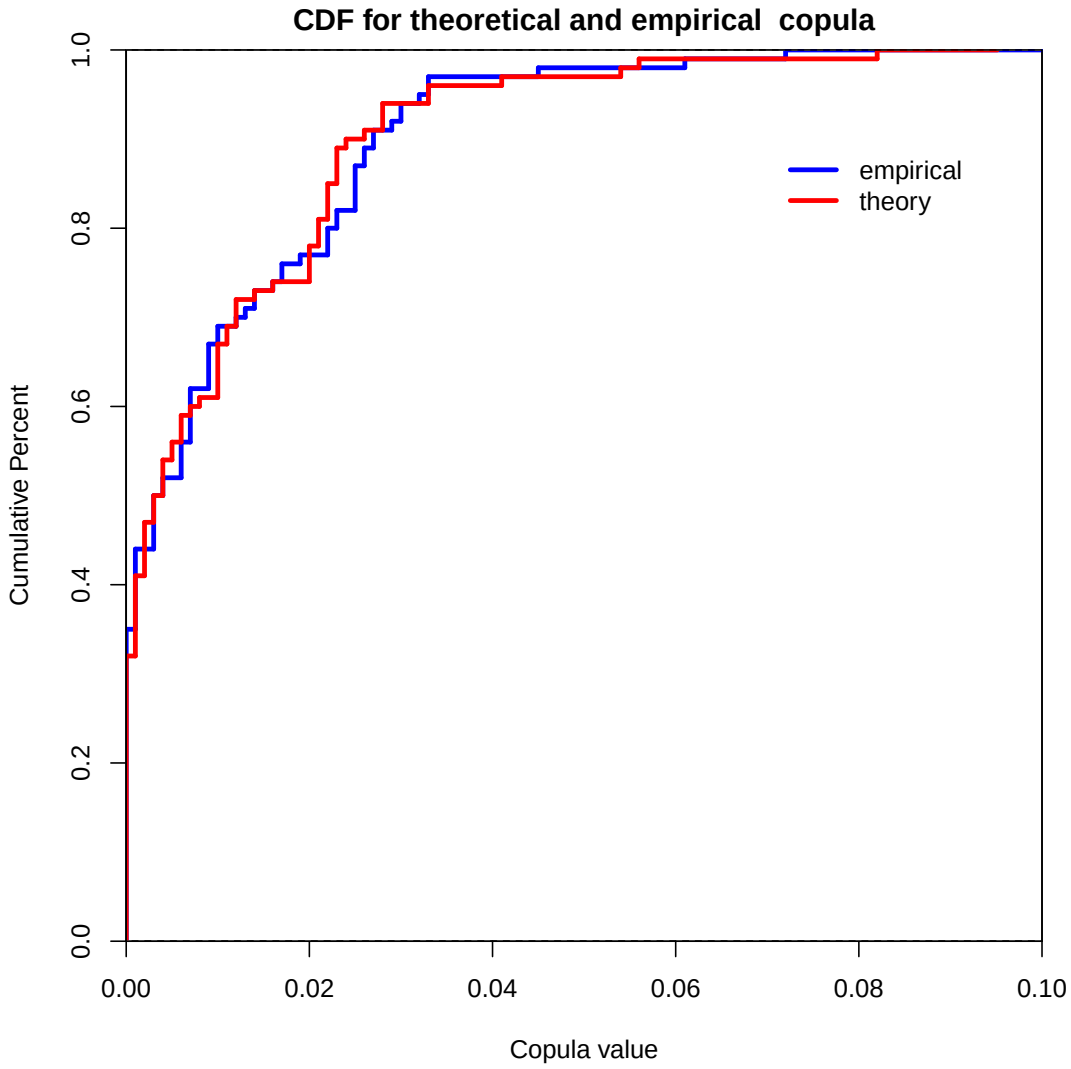



**Figure 9.5:** CDF of the theoretical and empirical copulas

**Table 9.9:** Goodness of fit tests of theoretical and empirical copulas

| | Kolmogorov–Smirnov | Anderson–Darling |
|---|---|---|
| P-value | 0.9938 | 0.6077 |

## 9.5 Summary

The bivariate skew-$t$ distribution is presented and the application of the theory is also examined using monthly rainfall data from two sites in the Murray–Darling Basin. Each marginal distribution of the observed data set is transformed to a uniform distribution on [0,1] using the gamma distribution. The simulated data are generated using the bivariate skew-$t$ distribution and each marginal distribution of the generated data set is transformed to a uniform distribution using the univariate skew-$t$ distribution.

A simulation approach is used to estimate the parameters of the skew-$t$ distribution. A refining process is carried out to match the statistics between the observed and simulated data which is based on their correlation coefficient and tail proportions. Matching correlation coefficients is one way to ensure that inter-dependence is included in the model. Using the best parameter values from the refining process, a data is generated from the bivariate skew-$t$ distribution and then compare the fit with the observed data. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests have shown a good result which was also supported by the graphical representation. A plot of empirical copula versus theoretical copula is presented for a goodness of fit test coupled with the Kolmogorov–Smirnov and Anderson–Darling tests. The results obtained demonstrate that the skew-$t$ copula is suitable for modelling monthly rainfall totals for correlated stations.

The copula methods extend to more than two variables and give the freedom to choose the correlation and marginal distributions. In fact the ability to manipulate the tail dependencies extends the inter-variable relations beyond simple correlation. This appears to be suitable for future work.

# Chapter 10

# Comparison of rainfall models

In this chapter we will compare the strengths and weaknesses of the models presented in the previous chapters. The rainfall models discussed are divided into two types, namely for meteorological stations which are weakly correlated and those which are strongly correlated. We also compare the methods for estimating the parameters such as the maximum likelihood method and moment generating function method.

## 10.1 Rainfall models for independent gamma variables

The models which are fitted for independent gamma variables are sum of independent mixed-gamma model with common scale parameters, sum of $n$ independent gamma variables with distinct shape and scale parameters and the sum of two independent gamma variables of Type I McKay distribution. The first two models demonstrate a good fit based on the Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests. The sum of independent mixed-gamma model with common

scale parameters is suitable when the analysis includes the zero value for rainfall. However, the sum of $n$ independent gamma variables with distinct shape and scale parameters is used for non-zero rainfall whenever the correlation between the data is insignificant. The model is easily applied to more than two independent gamma variables. The sum of two independent gamma variables of Type I McKay distribution is also fitted for non-zero rainfall. However, it is not easily extended to more than two because it is algebraically complex. These models are discussed in Chapter 4, Chapter 5 and Chapter 7, respectively. Hence, the sum of $n$ independent gamma variables with distinct shape and scale parameters is recommended to be used for independent cases of nonzero rainfall totals.

## 10.2 Rainfall models for correlated gamma variables

The models which are suitable to accommodate the correlation parameter are the sum of correlated gamma with common shape introduced by Alouini et al. [3], the sum of correlated gamma variables using the McKay distribution and the method using the skew-$t$ distribution. The best model between these three models is the sum of correlated gammas with common shape parameter because it is easily extended to more than four stations. We have shown that it fits well for four stations which are strongly correlated and also suitable for temporal analysis. For temporal analysis, we fit the model using the sum from four consecutive months within the same season. The sum of correlated gamma variables using the McKay distribution also gives a good fit for two correlated stations. However, it is algebraically complex to be extended for more than two correlated gamma variables. We show that the skew-$t$ distribution is used to match the tail proportion for two correlated gamma

variables. It is a promising model since the two variables are combined using any choice of marginal distributions. However, more studies are needed to explore the application for the sum of more than two correlated gamma variables. These models are discussed in Chapter 6, Chapter 7 and Chapter 9, respectively. For correlated cases, the sum of correlated gammas with common shape parameter is recommended.

The summary of the model comparison is given in Table 10.1 and the recommended models are in italics.

**Table 10.1:** Rainfall model comparison

| Rainfall model | Types of variable | Types of data | Parameters |
|---|---|---|---|
| Sum of mixed–gamma variables | independent | $\{0,1\}^a$ | Common $\beta$ |
| *Sum of n gamma variables* | independent | $\{1\}^b$ | Distinct $\alpha$, $\beta$ |
| Sum of two gamma variables of Type I McKay distribution | independent | $\{1\}$ | Distinct $\alpha$, $\beta$ |
| *Sum of correlated gamma variables* | dependent | $\{1\}$ | Common $\alpha$ |
| Sum of correlated gamma variables of McKay distribution | dependent | $\{1\}$ | Common $\alpha$ |
| Skew-$t$ | dependent | $\{1\}$ | |

[a](no rain, rain)
[b](rain)

## 10.3 Methods for parameters estimation

We examined two ways to estimate the parameters, namely maximum likelihood method and moment generating method. The maximum likelihood method is found

to be the best for estimating the parameters. The moment generating method can be used, however the results obtained do not give a good fit. More work is needed.

## 10.4 Summary

We have discussed the strengths and weaknesses of the rainfall models used for independent and dependent gamma variables. The choice of rainfall model used depends on the purpose of the study and the type of data available. So far, the maximum likelihood method is considered the best to estimate the parameters.

# Chapter 11

# Conclusions and future research

Rainfall modelling is important to supplement research in the hydrological and climatological fields of study. This chapter summarises the rainfall models studied. There are two types of rainfall modelling presented in the study, rainfall models for independent variables and rainfall model for correlated variables. We will highlight the advantages and disadvantages of the rainfall models in the following sections. We also demonstrated the methodology to generate synthetic rainfall data using each rainfall model discussed. For future research, the results of the study may be used in the rainfall–runoff modelling.

## 11.1 Research contributions

We discussed two types of rainfall modelling, for independent variables and correlated variables. The rainfall models for independent variables are the sum of independent mixed–gamma model, the sum of $n$ independent gamma variables and the sum of independent gamma variables of Type I McKay distribution. The rainfall models for correlated variables are the sum of correlated gamma variables, the sum

of correlated McKay distribution and the bivariate skew-$t$ distribution.

### 11.1.1 Rainfall modelling

The following models were used. In each case we summarise the findings.

**Rainfall models for independent variables**

- In **Chapter 4**, we discuss the independent mixed–gamma model introduced by Rosenberg et al. [87] to model rainfall totals including zero and non-zero rainfall totals. We extend this work to model the sum of two and three of monthly rainfall totals where the monthly rainfall totals are modelled by independent gamma random variables with same scale parameter but different shape parameters. We derive the formulae of mean and variance for each cases and also the general form of the formula. For illustration, the model is modified for both the same shape and the same scale parameters. This model is applied using rainfall totals from Pooraka, South Australia. We compare the mean and variance between the observed and generated rainfall totals and find that the means matched but the variances are underestimated. The distribution of the generated and the observed data is assessed using the Kolmogorov–Smirnov goodness of fit test. The P-value $> 0.05$ for both tests of sum of two and sum of three show that the observed data fits the model.

- In **Chapter 5**, we study the formulation of the sum of $n$ independent gamma variables with both different shape and scale parameters introduced by Moschopoulos [68]. Based on the Moschopoulos [68] model, we derive an approximate form of the sum of independent gamma variables. The parameters for the model are estimated using the cumulant generating function. The model was tested on

the total of monthly rainfalls from June to September inclusive at Hume and was found to provide a very good fit. The model was tested on correlated spatial data for monthly rainfall totals at nearby stations Dookie and Rutherglen and was found, as expected, to be unsatisfactory. The Kolmogorov–Smirnov and Darling–Anderson tests were used in each case to test the goodness of fit between the observed and simulated data sets.

- In **Chapter 7**, we go through the algebraic formulation of the sum of two independent gamma variables of Type I McKay distribution proposed by Holm and Alouini [45] to be able to understand the model of the sum of correlated gamma variables of the McKay distribution. We derive the PDF of the sum of two independent gamma variables with same shape and different scale parameters.

**Rainfall models for correlated variables**

- In **Chapter 6**, we describe the sum of correlated gamma variables with common shape parameter proposed by Alouini et al. [3] to model wireless communication system. We use the sum of correlated gamma variables with common shape to model rainfall totals. We highlight two limitations in the model (1) the matrix $C$ is not necessarily positive definite and (2) the values of the correlation parameter have to be positive. We apply the sum of four correlated gamma variables to model the sum of monthly rainfall totals from the nearby meteorological stations (Hume, Beechworth, Dookie and Rutherglen) and the sum of monthly totals from the same season (winter and summer) at Hume meteorological station. In the spatial example to represent the local water catchment, we apply the model directly since the matrix $C$ is positive definite and the correlation values are positive. However, we also test using

the generalised correlation and both results demonstrate a good fit based on the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests. In the temporal example, we have to use the generalised correlation [38] since some of the correlation values are negative and in order for the matrix $C$ to be positive definite. Again, the result is found to provide a very good fit using the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests.

- In **Chapter 7**, we discuss the sum of two correlated gamma variables of McKay distribution studied by Holm and Alouini [45]. We justify the PDF of the model and derive the various moments of the bivariate gamma distribution. The model is tested to find the sum of rainfall totals from two meteorological stations Hume and Beechworth during the summer months and the model is found to provide an excellent fit using the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests.

- In **Chapter 8**, we show how to estimate parameters using moments method. We use the first four cumulants to estimate the parameters of $\alpha, \beta_1, \beta_2$ and $\rho$. Based on the Alouini et al. [3], we proposed the formulae of mean and variance for the sum of two correlated gamma variables in terms of the eigenvalues. It is found that the observed mean and variance are matched with the formulae of mean and variance. The model is applied to the sum of monthly rainfall totals from Hume and Beechworth meteorological stations.

- In **Chapter 9**, we study the copula-based model which is normally used in financial application. We investigate the characteristics of the bivariate skew-$t$ distribution. Since, the distribution of rainfall totals is normally skewed, we

find that the bivariate skew-$t$ distribution is suitable to model the sum of two correlated random variables. In the methodology, we use gamma marginal to transform the data to uniform data and use the bivariate skew-$t$ distribution to correlate the data. Simulation technique is used to match the tail proportions of the observed and generated data. The Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests show that the model gives a good fit. It is possible to extend to more than two variables as described by Wong et al. [106]. Therefore, more research on using this method to model multivariate joint rainfall distributions is suggested.

### 11.1.2 Generating synthetic rainfall data

In this study, we fit various rainfall models for independent and correlated variables. The rainfall model that can match the statistical characteristics of the observed data will be used to generate the synthetic data. In most cases, we generate synthetic data and compare their statistical characteristic with the observed data. The fit of the model is also assessed using the Kolmogorov–Smirnov and Anderson–Darling goodness of fit tests. Therefore, a rainfall model that can generate synthetic rainfall data is useful whenever the data available is limited in terms of length, completeness and spatial coverage. I have demonstrated methods to generate synthetic data from the probability density function or cumulative distribution function in chapters 4, 5, 6, 7 and 8 (no synthetic rainfall data in 9).

## 11.2 Recommendations for future research

For future research, the rainfall models of independent and correlated variables can be extended or applied as follows:

- The development of models that allow different scale parameters to be used would be a valuable contribution. The method proposed by Kotz and Adams [58] and the modifications suggested by Moschopoulos [68] and by Alouini et al. [3] would probably be the best starting point.

- The bivariate skew-$t$ copula is used to model the sum of two correlated gamma variables. More research using the copula-based model is needed since many research studies highlight the practical use of the copula concept for multivariate data where the freedom of choosing the marginal distribution is really interesting. This study is limited to the sum of two correlated random variables.

- The synthetic data generation is able to generate synthetic rainfall totals which has a similar characteristic of the observed data. Some models required certain amount of data for the model to be valid. Therefore, the methodology is very useful to supplement data in the area where data availability is limited. Furthermore, it can also be used to present different rainfall scenarios by changing the value of the parameters in the model.

- The methodology described in this study can be applied to other rainfall basins anywhere in the world.

# Appendix A

# River Murray Management

*"The River Murray Valley is in a very real sense, a commons and the people of the Valley indeed, of Australia as a whole are the commoners. Only by mutual restraint mutually agreed upon as being in the common interest, can the resources of the Valley be maintained in perpetuity and then, only if we have better knowledge of the inter-relationships between the various parts of the system and if we are to continually modify our management of the system in the light of that knowledge..."*

*"That there is a need for co-operation in dealing with some aspects of the River Murray Valley, was recognised over 60 years ago, when the River Murray Waters Agreement was made and the River Murray Commission was formed..."*

*"Unfortunately, there has been very little co-ordination between the States and indeed very little interdisciplinary co-operation between individual agencies within individual States in other fields. Since then each agency appears intent on pursuing its own undoubtedly well meant, but often ill-advised objectives..."*

*"Surely, if it is not too much to ask of individuals, our Governments and their Departments and agencies, that they set aside their individual interests for a moment*

*in pursuit of a common good and so as to avoid (in the Murray) the 'Tragedy of the Common' – Hardin (1968)".*

The ex-president of the River Murray Commission delivered his speech in the Dinner Address for the Dept. of Dec. and Dev. NSW Govt. [64].

# Appendix B

# Rainfall maps and graphs

- 30 years of Australia rainfall distribution from 1911–2000: Figures B.1 – B.3;
- Australia, New South Wales and Victoria trend in annual total rainfall from 1910–2009 and 1960–2009(mm/10 yrs): Figures B.4 – B.9;
- Australia, New South Wales and Victoria annual rainfall 10 years moving average from 1961–2009(mm/10 yrs): Figures B.10 – B.12;
- Median monthly rainfall Sydney and Melbourne from 1961–2009(mm/10 yrs): Figures B.13 – B.14;

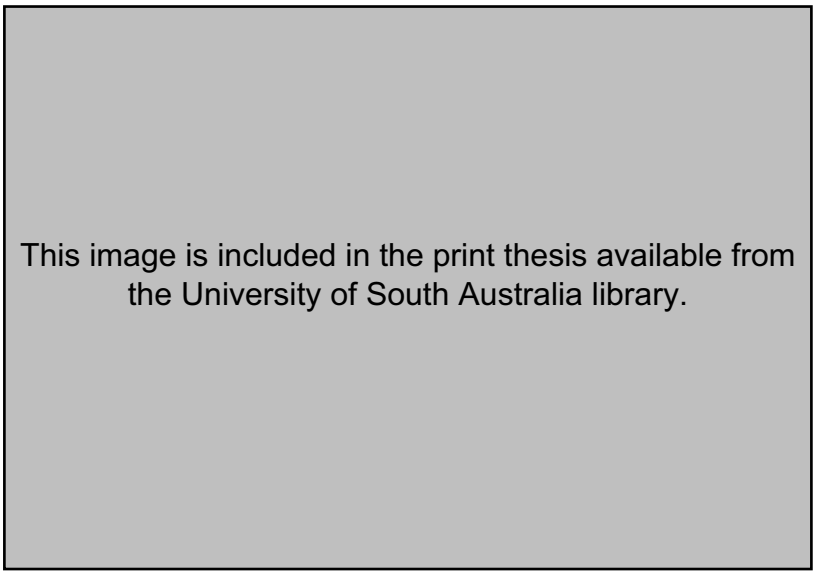



**Figure B.1:** Australian Rainfall Distribution for 1911–1940 (Source: Australian Bureau of Meteorology 2010)

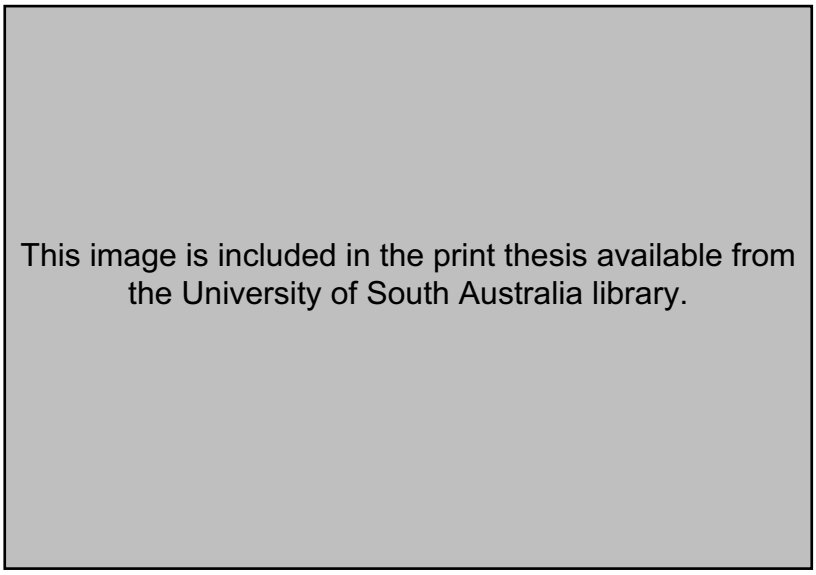



**Figure B.2:** Australian Rainfall Distribution for 1941–1970 (Source: Australian Bureau of Meteorology 2010)

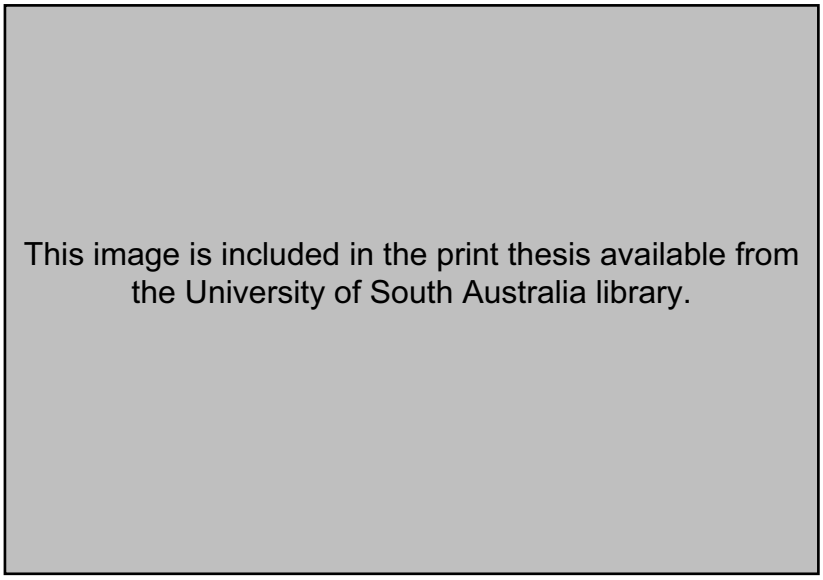



**Figure B.3:** Australian Rainfall Distribution for 1971–2000 (Source: Australian Bureau of Meteorology 2010)

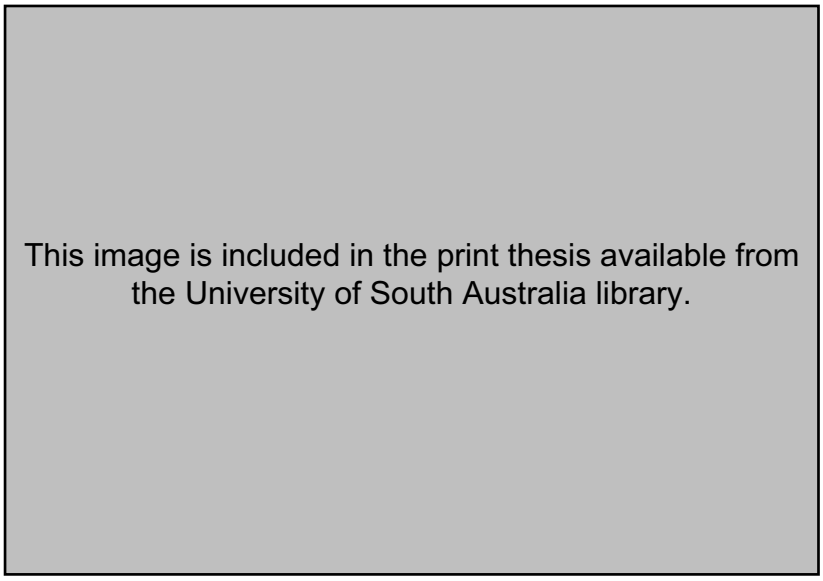



**Figure B.4:** Australia trend in annual total rainfall 1910–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010)

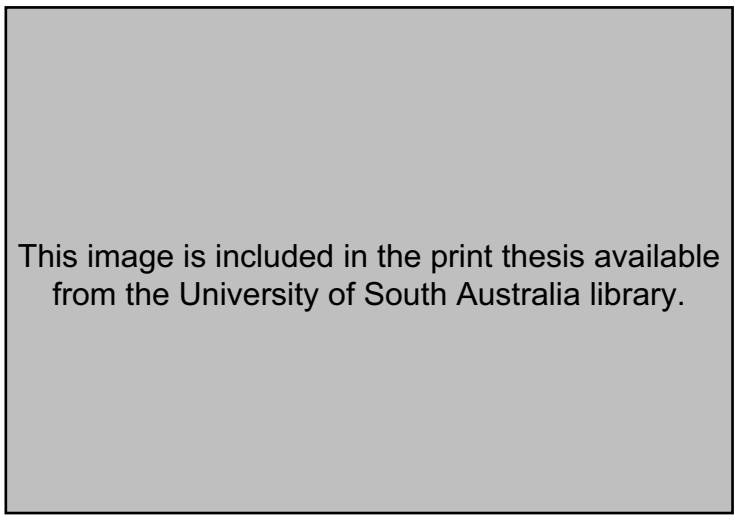



**Figure B.5:** Australia trend in annual total rainfall 1960–2009 (mm/10 yrs (Source: Australian Bureau of Meteorology 2010)

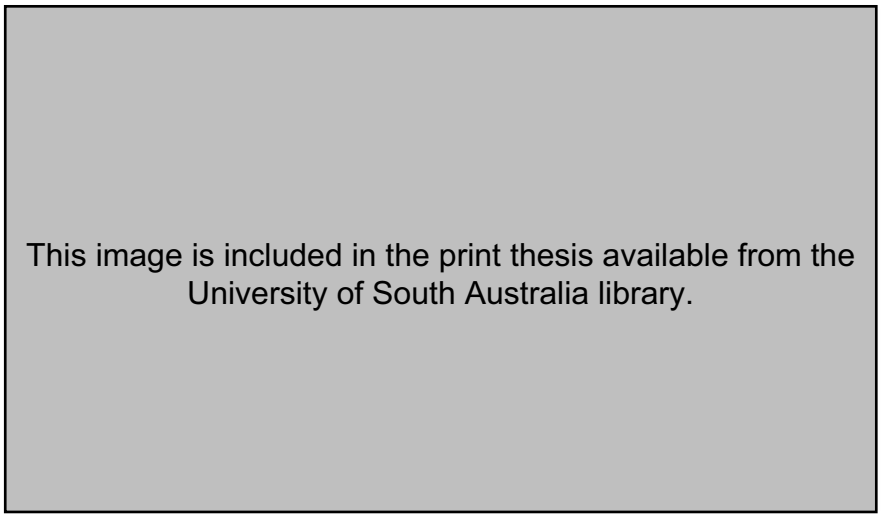



**Figure B.6:** NSW trend in annual total rainfall 1910–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010)

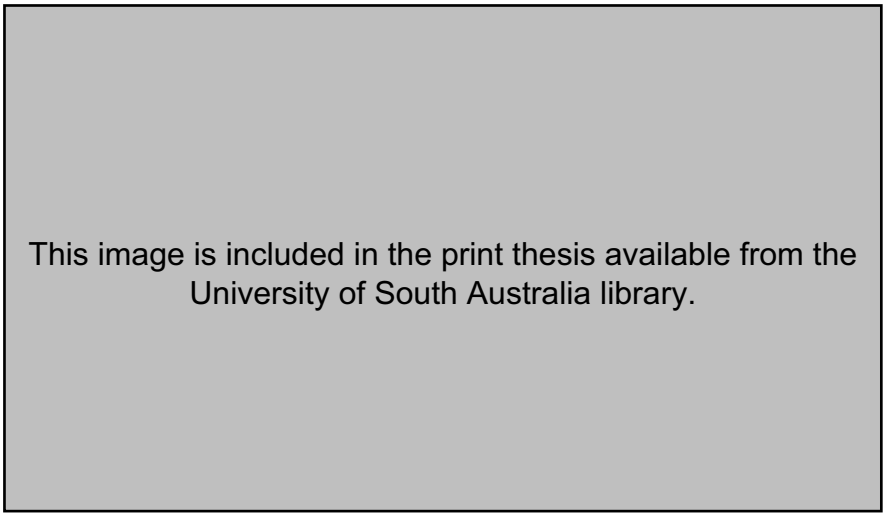



**Figure B.7:** NSW trend in annual total rainfall 1960–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010)

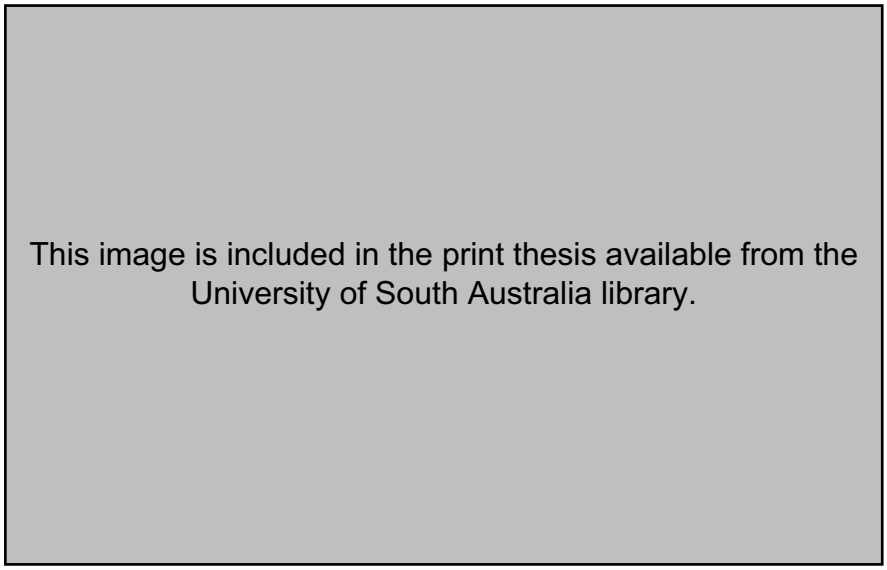



**Figure B.8:** Victoria trend in annual total rainfall 1910–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010)

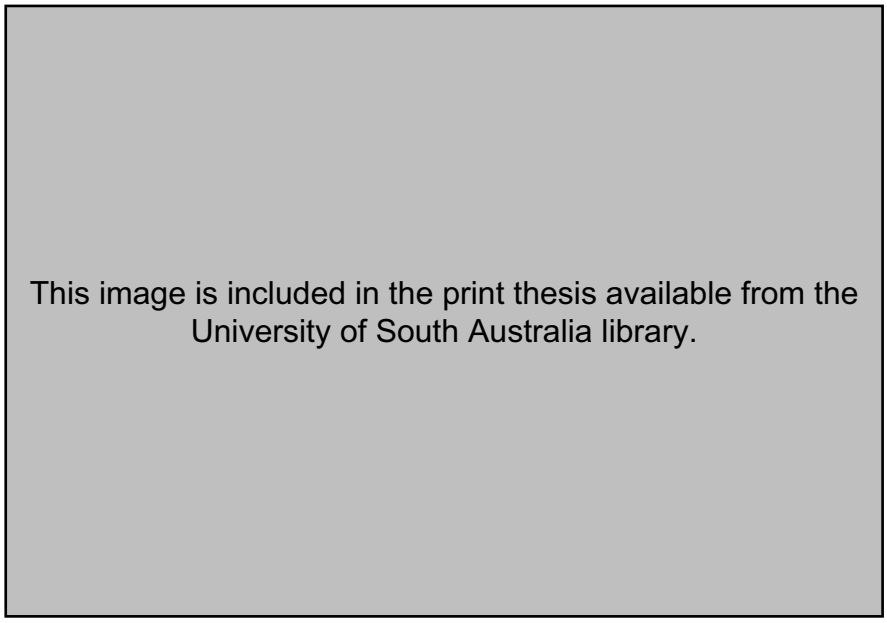



**Figure B.9:** Victoria trend in annual total rainfall 1960–2009 (mm/10 yrs) (Source: Australian Bureau of Meteorology 2010)

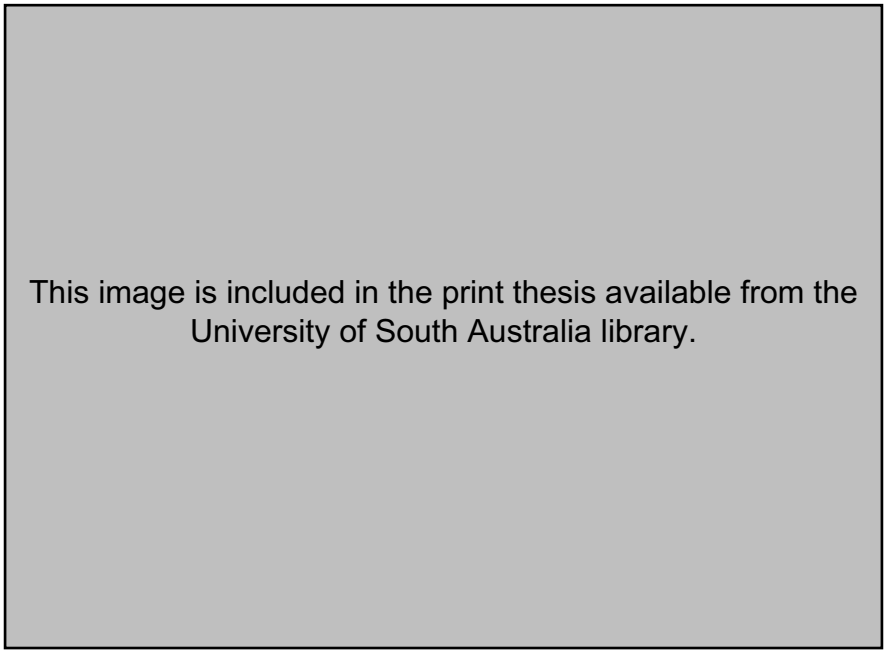



**Figure B.10:** Australia annual rainfall 10 years moving average (Source: Australian Bureau of Meteorology 2010)

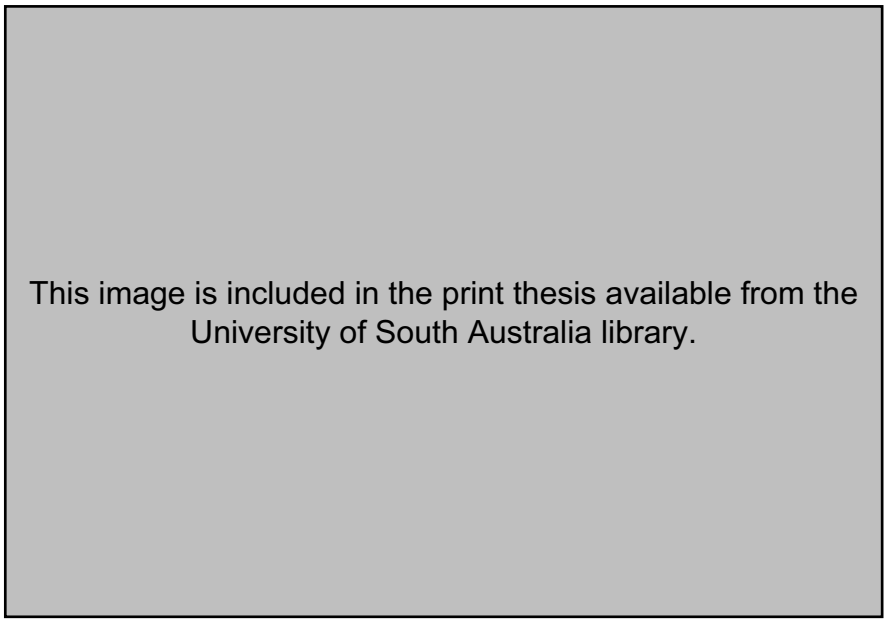



**Figure B.11:** NSW annual rainfall 10 years moving average (Source: Australian Bureau of Meteorology 2010)

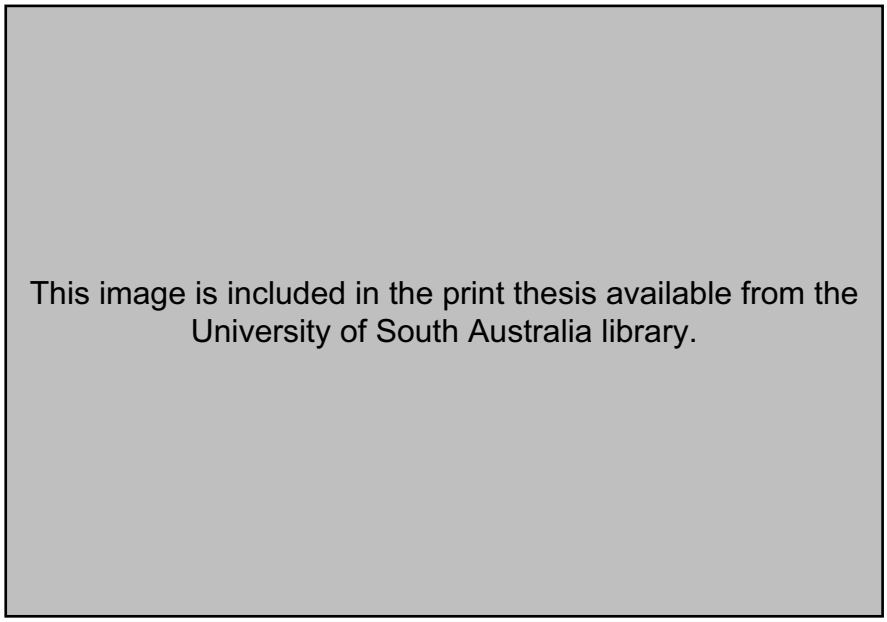



**Figure B.12:** Victoria annual rainfall 10 years moving average (Source: Australian Bureau of Meteorology 2010)

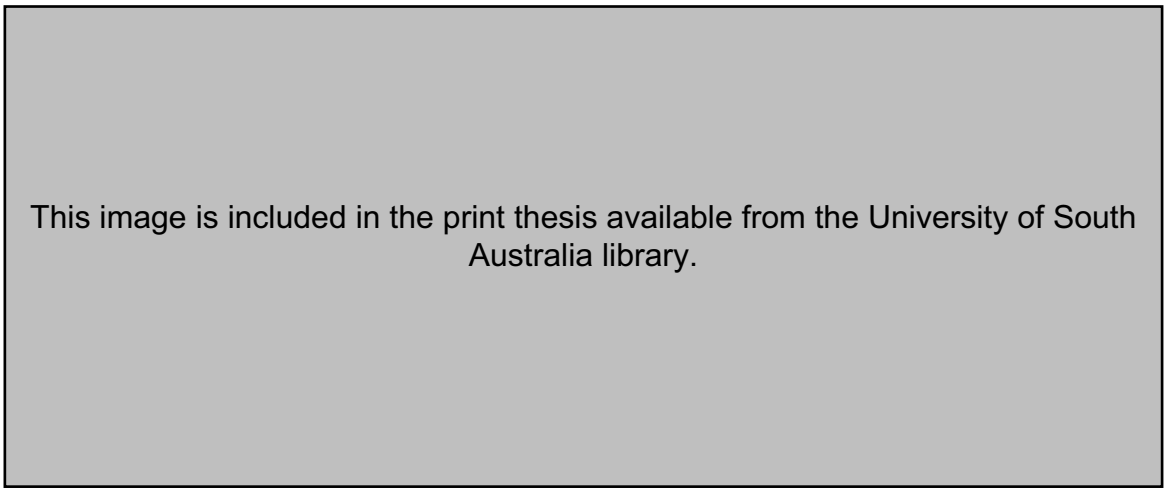



**Figure B.13:** Sydney median annual rainfall (Source: Australian Bureau of Meteorology 2010)

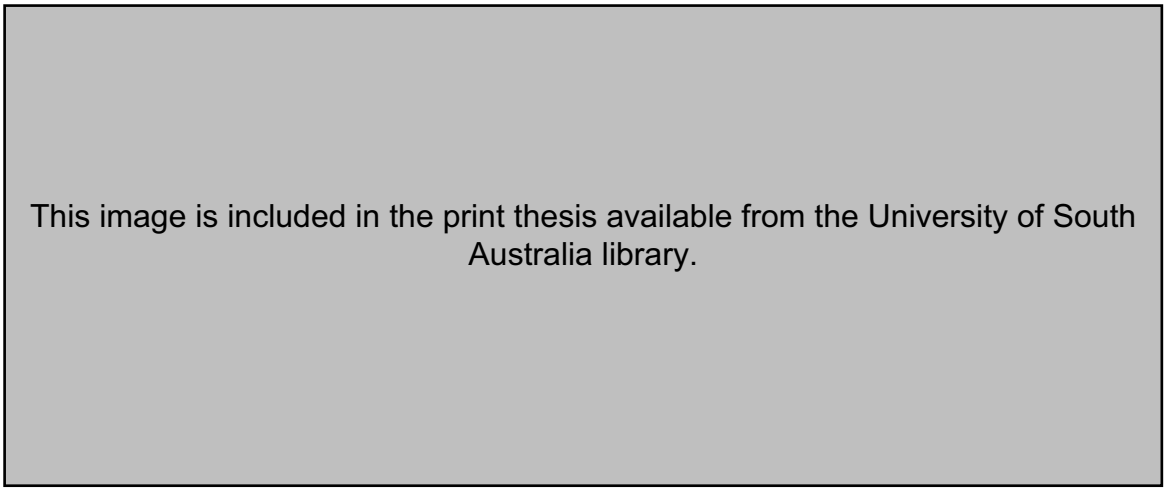



**Figure B.14:** Melbourne median annual rainfall (Source: Australian Bureau of Meteorology 2010)

# Appendix C

# R package



$R$ is a programming language and software environment for statistical computing and graphics. It is a free software and it is based on *open source*. The language is similar to $S$ language but it is more dynamic and highly extensible. The codes or functions developed in $R$ come from contributed packages. A package contains a collection of files and directories (folders). Ross Ihaka and Robert Gentleman from the University of Auckland were responsible in creating the $R$ language and now it is managed and developed by the *R Development Core Team* [80]. The version that I use is $Tinn - R$ 2.3.7.1. [33].

# C.1 Data summary

This code gives a descriptive statistical summary of monthly rainfall totals for Hume Reservoir (NSW), Beechworth Composite (Vic), Dookie (Vic) and Rutherglen Research (Vic) for the period of 1922–2008.

```
## Analysis for 4 stations
## NSW - Hume Reservoir
## Vic - Beechworth Composite, Dookie, Rutherglen Research
## Data type: Monthly rainfall
## Years selected : 1922-1978, 1980-1986, 1988-1994, 1996-1998,
##                  2002-2005, 2007-2008
## Missing years : 1979, 1987, 1995, 1999, 2000, 2001, 2006
## Total number of years - 80 years
## Total number of data - 80 * 12 = 960

## Data
hume <- read.csv('D:/Mywork/UniSA_Roslina/PhD/2010_Research PhD/
data/hume2.csv',header=FALSE)
dook <- read.csv('D:/Mywork/UniSA_Roslina/PhD/2010_Research PhD/
data/dookie2.csv',header=FALSE)
beec <- read.csv('D:/Mywork/UniSA_Roslina/PhD/2010_Research PhD/
data/beechworth2.csv',header=FALSE)
ruth <- read.csv('D:/Mywork/UniSA_Roslina/PhD/2010_Research PhD/
data/rutherglen2.csv',header=FALSE)

## Package
library(psych)   ## need for 'describe'
```

```
## Data summary
dat.summary <- function(dt)
{  mn   <- colMeans(dt)
   vr   <- diag(var(dt))
   sd   <- sqrt(vr)
   p0   <- colSums(dt == 0)/nrow(dt)
   comb <- rbind(dt,mn, sd,p0)
   #comb <- rbind(dt,mn, vr, sd,p0)
   yr   <- rowSums(comb)
   all  <- cbind(comb,yr)
   rownames(all) <- c(1922:1978, 1980:1986,1988:1994,1996:1998,
                      2002:2005,2007:2008,"mean","sd","p0")
   colnames(all) <- c("Jan","Feb","Mar","Apr","May","Jun","Jul",
                      "Aug","Sep","Oct","Nov","Dec","Annual")
   round(all,2)
}

d1 <- dat.summary(hume)[81:83,]
d2 <- dat.summary(beec)[81:83,]
d3 <- dat.summary(dook)[81:83,]
d4 <- dat.summary(ruth)[81:83,]
rbind(d1,d2,d3,d4)

## overall summary
describe(hume);describe(beec);describe(dook);describe(ruth)

all_dt <- cbind(c(t(hume)),c(t(beec)),c(t(dook)),c(t(ruth)))

## Pearson correlation (assume data normally distributed)
```

```
correl <- cor(all_dt)
rownames(correl) <- c("hume","beec","dook","ruth")
colnames(correl) <- c("hume","beec","dook","ruth")
round(correl,2)

## Spearman's Rank correlation (assume data not normally distributed)
rho <- rcorr(all_dt, type="spearman") [[1]]; round(rho, 2)

## ******************************************************** ##
## plot of average monthly rainfall totals
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0))
plot(colMeans(hume), xaxt = "n",xlab="Month",
ylab="Average of rainfall totals (mm)",
main="Average of monthly rainfall totals (mm) ",
ylim=c(0,150),type="l",col="blue",lwd=3)
lines(colMeans(beec),col="black",lwd=3,lty=2)
lines(colMeans(dook),col="green3",lwd=3,type="l")
lines(colMeans(ruth),col="red",lwd=3,lty=4)
legend("topright", bty="n",c("Hume","Beec","Dook","Ruth"),
lwd=c(3,3,3,3,3), lty =c(1,2,1,4),col= c("blue","black","green3","red"))
axis(1,at=1:12,labels=month.abb)
axis(2,  at=25*0:150)
```

## C.2 Estimating shape and scale parameters for gamma distribution

This code is used to fit and estimate values of shape ($\alpha$) and scale ($\beta$) for the gamma distribution. The gamma distribution is fitted for nonzero rainfall totals and given by

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \qquad \text{for } x > 0 \text{ and } \alpha, \beta > 0.$$

```
library(MASS)  ## need for fitdistr(...)

alpha.beta.multicol <- function(dt)
{ cl    <- ncol(dt)
  tol   <- 1E-6
  alp   <- vector(length=cl, mode="numeric")
  bt    <- vector(length=cl, mode="numeric")
  mn    <- vector(length=cl, mode="numeric")
  vr    <- vector(length=cl, mode="numeric")
  p0    <- vector(length=cl, mode="numeric")
  alp_bt   <- vector(length=cl, mode="numeric")
  alp_bt2  <- vector(length=cl, mode="numeric")

  for (i in 1:cl)
  { xi  <- dt[,i]
    ai  <- xi [xi > tol]          ## extract only positive data
  alp[i] <-
  fitdistr(ai,dgamma,start=list(shape=10,scale=2),lower=0.01)[[c(1,1)]]
  bt[i]  <-
```

```
  fitdistr(ai,dgamma,start=list(shape=10,scale=2),lower=0.01)[[c(1,2)]]

    mn[i]       <- mean(ai)         ## calc mean
    alp_bt[i]   <- alp[i]*bt[i]     ## calc mean from formula
    vr[i]       <- var(ai)          ## calc var
    alp_bt2[i]  <- alp[i]*bt[i]^2   ## calc var from formula
    p0 [i]      <- sum (xi == 0)/length (xi)
  }
  abmvrp0 <- cbind(p0=p0,alp=alp,bt=bt,mn=mn,alp_bt=alp_bt,
  vr=vr,alp_bt2=alp_bt2)
  rownames(abmvrp0) <- c(1:cl)
  mn_all <- colMeans(abmvrp0)
  rbind(abmvrp0,mn_all)
}
```

## C.3 Synthetic data generation using gamma model and mixed–gamma model

This code is used to generate synthetic data using the specified shape ($\alpha$) and scale ($\beta$) parameters for the gamma distribution and using shape ($\alpha$), scale ($\beta$) and probability of zero ($p_0$) for the mixed–gamma distribution.

```
## Gamma model
## rgamma(n, shape= alpha, scale =beta)
rgamma(1000, shape=alp_dec, scale = bt_dec)

## Mixed-gamma model
prob_amt <- function(n,itr,p0,alpha,beta)
```

```
{  rand <- rand.no(n,itr)
   y <- rand
   for (i in 1:itr)
  { ind   <- rand[,i] < p0
    y[ind,i]  <- 0
    y[!ind,i] <- qgamma((rand[!ind,i]-p0)/(1-p0),alpha,scale=beta)
  }
  #list(prob=rand, rain.amt=y)
  #cbind(list(prob=rand) $prob[,1],list(rain.amt=y) $rain.amt[,1] )
  #c(list(prob=rand) ,list(rain.amt=y) )
  #list(rain.amt=y)
  data.frame(rand.no = as.vector(as.matrix(rand)),
               rain.amt = as.vector(as.matrix(y)))
}
```

## C.4 Newton Method

This code below is used to support the other codes that I develop to solve the Newton iteration problem. It was written by Dr Rolf Turner, currently leader of the Quantitative Team in the Starpath Project, University of Auckland, New Zealand. The code of *newt*() was written while he was a Professor in the Department of Mathematics and Statistics at the University of New Brunswick, Fredericton, New Brunswick, Canada.

```
newt <- function(fn,start,...,eps.p = 1e-08,eps.v = NULL,
maxit = 50,verb = FALSE)
{ p.o <- start
  itno <- 1
  repeat {
    fj <- fn(p.o,...)
    v <- fj$fval
    t1 <- if(is.null(eps.v)) NULL else sum(abs(v))
    J <- as.matrix(fj$jacobian)
    if(qr(J)$rank < ncol(J)) {
        cat("Singular Jacobian.\n")
        rslt <- if(is.null(eps.v)) NA else if(t1 < eps.v) p.o
        else NA
        break
    }
    else {
        p.n <- p.o - solve(J) %*% v
        t2 <- max(abs(p.n - p.o))
        if(verb) {
            tmp <- format(round(c(p.o,p.n,v,t2,t1),6))
```

```
            np <- length(v)
            v1 <- paste(tmp[1:np],collapse = "  ")
            v2 <- paste(tmp[(np + 1):(2 * np)],collapse = "  ")
            v3 <- paste(tmp[(2 * np + 1):(3 * np)],collapse = "  ")
            v4 <- tmp[3 * np + 1]
            v5 <- tmp[3 * np + 2]
            cat("\nIteration  : ",itno,"\n",sep = "")
            cat("Old par    : ",v1,"\n",sep = "")
            cat("New par    : ",v2,"\n",sep = "")
            cat("Test ch.par: ",v4,"\n",sep = "")
            cat("Fn. vals.  : ",v3,"\n",sep = "")
            if(!is.null(t1))
              cat("Test f.val: ",v5,"\n",sep = "")
        }
        if((!is.null(t1) && t1 < eps.v) | t2 < eps.p) {
            rslt <- p.n
            break
        }
        itno <- itno + 1
        if(itno > maxit) {
            cat("Newton's method failed to converge in\n")
            cat(maxit,"iterations.\n")
            rslt <- NA
            break
        }
        p.o <- p.n
    }
}
as.vector(rslt)
```

```
}

newton.alpha_fix <- function(pars,x,r)
{  alp   <- 1.6542       ## fixed parameter
   bt    <- pars[1]
   r     <- length(x)
   sm1   <- sum(log(x))
   sm2   <- sum(x)
## Loglikelihood function
   fn1   <- (sm2/pars[1]^2)-(r*alp/pars[1])
   fval  <- fn1
## jacobian matrix
   jc11  <- (-2)*sm2/pars[1]^3 + r*alp/pars[1]^2
   jacobian <- jc11
## output
   list(fval=fval,jacobian=jacobian)
}

## Fix alpha and re-estimate beta
yypos <- stn_pos
x <- yypos[,1]   #Hume
bt1 <- newt(newton.alpha_fix,0.05,x=x,r=length(x))
```

## C.5 Probability density function of the sum of $n$ correlated gamma variables

This code is used to calculate the PDF value of the sum of $n$ correlated gamma variables. Define $Y = \sum_{i=1}^{n} X_i$ as the sum of $n$ correlated gamma variables with the PDF [3] given by

$$f_Y(y) = \prod_{i=1}^{n} \left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha} \sum_{k=0}^{\infty} \frac{\delta_k \; y^{n\alpha+k-1} \exp\left(-y/\lambda_1\right)}{\lambda_1^{n\alpha+k} \; \Gamma(n\alpha+k)}$$

where $\{\lambda_i\}_{i=1}^{n}$ are the eigenvalues of the matrix $A = DC$ where we assume that $A$ is positive definite and denote $\lambda_1$ as the minimum eigenvalue.

```
## ------------------------------------------------------------ ##
## Pdf of Sum of correlated gamma variables - Alouini (2001)    ##
## ------------------------------------------------------------ ##

alp <- 2.0670; bt1 <- 24.5549; bt2 <- 44.3067
betas <- round(c(bt1,bt2),2)
mt.beta <- diag(betas);mt.beta
rho <- rcorr(stn_pos,type="spearman")[[1]]; round(rho,4)
sqrt_rho <- round(sqrt(rho),2); sqrt_rho
C   <- sqrt_rho; C
D   <-  mt.beta; D
A   <-  D %*% C; A    ## calculate eigen values
eigen(A); min(eigen(A)$values)

eign_val <- c(67.283666, 1.576334)
alp <- 2.0670 ;  lambda <- eign_val
lambda_m <- min(lambda)
```

```
t <- 400; t1 <- t+1

#first get the sum over the eigenvalues for a particular power i
lambda_sum <- function(lambda,lambda_m,i)
{(1-lambda_m/lambda)^i}

term <- function(i,lambda,lambda_m)
{sum(sapply(lambda,lambda_sum , lambda_m, i))}

## to get the values of sum(j=1:N, (1 - lambda_m/lambda_j)^i )
lambda_sum <- sapply(1:t, term,  lambda, lambda_m); lambda_sum

## Initialise delta to 1
delta <- rep(1, t1)

## code for delta_calc
sm <- lambda_sum; sm

delta_calc <- function(k, sm, alp)
{ get("delta",envir=parent.frame(1))
  delta[k+1] <- alp/k*sum(unlist(sapply(1:k, function(i, k, delta, sm)
  {sm[i]*delta[k - i + 1]}, k, delta, sm)))
  assign("delta", delta, envir=parent.frame(1), inherits=TRUE)
  invisible(delta[k+1])
}
invisible(sapply(1:t, delta_calc, sm, alp)); delta
plot(delta,main=expression(paste("Values of "*delta)),
xlab ="k",ylab="Delta_k")
```

```
## product term
prod_term <- function(lambda,alp)
{ n       <- length(lambda)
  lambda_m <- min(lambda)
  p       <- (lambda_m^n/prod(lambda))^alp
  p
}

prod_term(lambda,alp)

prod_term(lambda,alp)*sum(delta)

## compute kth term in the sum
pdf_term <- function(k, dt, alp, lambda_m, delta)
{  exp(log(delta[k])+(4*alp+k-1)*log(dt)-dt/lambda_m
            - (4*alp+k)*log(lambda_m)-lgamma(4*alp+k))}

## compute kth sum of the pdf_term
pdf_sum1 <- function(dt, alp, lambda_m, delta)
{sum(unlist(sapply(1:t, pdf_term, dt, alp, lambda_m, delta)))}

## ---------------------------------- ##
## Generation of synthetic data      ##
## ---------------------------------- ##
datobs <- stn_pos_sum          ## observed data
matrix(datobs,ncol=1)
max(datobs); min(datobs)

seq_size <- 1    ;  n <- seq_size      ## interval of data
```

```
len_dt   <- 1000 ;  m <- len_dt       ## length of data
seq_dt <- seq(0,m,n); length(seq_dt)

## use pdf of sum of correlated data to generate synthetic data
pdf_formula.2 <-
prod_term(lambda,alp)*sapply(seq_dt, pdf_sum1,alp,lambda_m,delta)
gen <- pdf_formula.2
sum(gen*seq_size)

## to obtain the generated rainfall from the pdf
sq <- seq_dt
rel_gen <- pdf_formula.2*seq_size
cdf <- c(cumsum(rel_gen),1)
cbind(rel_gen,seq_dt,cumsum(rel_gen))

set.seed(4455)
rand <- runif(2000)

## similar to lookup table, in Excel 'vlookup'
dd2 <- cbind(rand,gen1 = sq[findInterval(rand,cdf)],
             gen2 = sq[findInterval(rand,cdf)+1])
head(dd2,20); tail(dd2,20)
dd  <- dd2[,2]
matrix(dd,ncol=1)

describe(datobs);IQR(datobs)  ## observed sum
describe(dd);IQR(dd)          ## generated sum
```

## C.6 Probability density function of the sum of correlated gamma variables of McKay distribution

This code is used to calculate the PDF value of the sum of correlated gamma variables of McKay distribution. Let $X_1, X_2 \sim G(\alpha, \beta_1, \beta_2, \rho)$ and consider the sum $\Sigma = X_1 + X_2$ follows type I McKay distribution as in equation (7.1) for $c > 1$, $b > 0$ and $0 \leq \rho < 1$. The PDF of the sum of two correlated gamma variables is given by

$$f_\Sigma(\sigma) = \frac{\sqrt{\pi}}{[\beta_1\beta_2(1-\rho)]^\alpha \Gamma(\alpha)} \Big(\frac{\sigma}{2\widetilde{\beta}}\Big)^{\alpha-\frac{1}{2}} \; e^{-\frac{\sigma(\beta_1+\beta_2)}{2\beta_1\beta_2(1-\rho)}} \; I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta})$$

with

$$\begin{aligned} a &= \alpha - \frac{1}{2} \\ b &= \frac{2\beta_1\beta_2(1-\rho)}{\sqrt{(\beta_1+\beta_2)^2 - 4\beta_1\beta_2(1-\rho)}} = \frac{2\beta_1\beta_2(1-\rho)}{\sqrt{(\beta_1-\beta_2)^2 + 4\beta_1\beta_2\rho}} \\ c &= \frac{\beta_1+\beta_2}{\sqrt{(\beta_1+\beta_2)^2 - 4\beta_1\beta_2(1-\rho)}} = \frac{\beta_1+\beta_2}{\sqrt{(\beta_1-\beta_2)^2 + 4\beta_1\beta_2\rho}} \end{aligned}$$

where $\frac{c}{b} = \frac{\beta_1+\beta_2}{2\beta_1\beta_2(1-\rho)}$ and $\widetilde{\beta}^2 = \frac{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}{4\beta_1^2\beta_2^2(1-\rho)^2}$ then $\widetilde{\beta} = \frac{\sqrt{(\beta_1-\beta_2)^2+4\beta_1\beta_2\rho}}{2\beta_1\beta_2(1-\rho)} = \frac{1}{b}$.

```
## ------------------------------------------------------------ ##
## Pdf of Sum of correlated gamma variables of McKay           ##
## ------------------------------------------------------------ ##


pdf_mcKay <- function(dt,alp, bt1,bt2,rho)
{  a       <- alp - 0.5
   btilda  <- sqrt((bt1-bt2)^2+4*bt1*bt2*rho)/(2*bt1*bt2*(1-rho))
   balp    <- (bt1*bt2*(1-rho))^alp
```

```
   b1b2rho <- (bt1+bt2)/(2*bt1*bt2*(1-rho))

   n       <- length(dt)
   bexp    <- vector(length=n, mode="numeric")
   bess    <- vector(length=n, mode="numeric")
   pdfmcKay <- vector(length=n, mode="numeric")

   for (i in 1:n)
   { bexp[i]     <- exp(-dt[i]*b1b2rho)
     bess[i]     <- besselI(dt[i]*btilda,a)
     pdfmcKay[i] <- (sqrt(pi)/(balp*gamma(alp)))*
                    (dt[i]/(2*btilda))^a*bexp[i]*bess[i]
   }
   pdfmcKay
}
```

## C.7 Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests

The Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests are considered nonparametric test or also known as 'distribution-free' test. It is normally used to assess the fit between the empirical distribution and the specified distribution whenever the data is skewed or non-normal data. The data that is being assessed can be discrete, continuous or grouped.

```
## Kolmogorov_Smirnov Goodness of fit test
ks.test(observe,generated)$p.value
```

```
## Anderson-Darling Goodness of fit test
library(adk)
x3 <- list(observe,generated)
adk.test(x3)
```

## C.8 Plotting the PDF and CDF

This code is used to plot the probability density function and cumulative distribution function for the observed, generated and fitted data.

```
#########################################################
## plots of observed vs. fitted PDF                    ##
#########################################################
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
hist(datobs,prob=TRUE, main ="PDFs observed Vs fitted ",col="yellowgreen", cex.axis=1.2,
xlab="Rainfall (mm)", ylab="Relative frequency", ylim= c(0,.006), xlim=c(0,700),nclass=12,lwd=3)
lines(rel_gen_dd, lwd=3,col="red")
legend(500, 0.005, legend = c("observed","fitted"),
       col = c("yellowgreen","red"), pch=c(15,NA), lty = c(0, 1),
       lwd=c(0,3),bty="n", pt.cex=2)
box()
#########################################################
## plots of generated vs. fitted PDF           ##
#########################################################
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
hist(dd,prob=TRUE, main ="PDFs generated Vs fitted",col="cornflowerblue", cex.axis=1.2,
xlab="Rainfall (mm)", ylab="Relative frequency", ylim= c(0,.006), xlim=c(0,700),nclass=12,lwd=3)
lines(rel_gen_dd, lwd=3,col="red")
legend("topright", legend = c("generated","fitted"),
       col = c("cornflowerblue","red"), pch=c(15,NA), lty = c(0, 1),
       lwd=c(0,3),bty="n", pt.cex=2)
box()
#########################################################
## plots of observed density vs. generated PDF
#########################################################
## To draw side by side histogram
```

```
bin_size <- 50  ;  n <- bin_size     ## interval of data
len_dt   <- 700 ;  m <- len_dt       ## length of data

seq_dt <- seq(0,m,n); length(seq_dt)

histo_obs  <- hist(datobs,breaks=seq_dt,freq=F)
histo_obs$breaks
histo_obs$counts

histo_gen  <- hist(dd,breaks=seq_dt,freq=F)
histo_gen$breaks
histo_gen$counts

## extract counts from histogram and calculate the probability in
## each specified interval. Also check the length of the interval is
## the same with the length of the fitted model.

rel_obs <- c(0,histo_obs$counts)/length(datobs)## relative frequency for observe data
sum(rel_obs)
rel_gen <- c(0,histo_gen$counts)/length(dd)## relative frequency for generated data
sum(rel_gen)

hist_jun <- cbind(rel_obs,rel_gen)

## -- Plots of histogram and curve for observed and fitted gamma marginals -- ##

## side by side barplot (histogram)
sq.n   <- as.character(seq_dt)  ## to label x-axis
max.relgen <- max(rel_gen)+.1

par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
rownames(hist_jun) <- sq.n
barplot(t(hist_jun), col= c("yellowgreen","blue"), beside=TRUE,
ylim=c(0,max.relgen), main="Observed Vs generated",
xlab="Rainfall (mm)",cex.axis=0.8,font.axis=2,ylab="Relative Frequency")
legend("topright", c("observed","generated"), fill= c("yellowgreen","blue"),bty="n")
box()
########################################
## CDF plots of observed and fitted  ##
########################################
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
```

```
plot(ecdf(datobs),xlim=c(0,700),verticals=TRUE, horizontals=TRUE,
pch=45,do.p=FALSE, col.hor="yellowgreen", col.vert="yellowgreen",lwd=3,
main="CDFs observed Vs fitted",
xlab="Rainfall (mm)",ylab="Probability")
lines(cdf, col="red",lwd=3)
legend(500,.9, legend = c("observed","fitted"), col = c("yellowgreen","red"),
pch=c(NA,NA), lty = c(1, 1),lwd=c(3,3),bty="n")
box()
########################################
## CDF plots of observed and generated  ##
########################################
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
plot(ecdf(datobs),xlim=c(0,700),verticals=TRUE, horizontals=TRUE,
pch=45,do.p=FALSE, col.hor="yellowgreen", col.vert="yellowgreen",lwd=3,
main="CDFs observed Vs generated",
xlab="Rainfall (mm)",ylab="Probability")
lines(ecdf(dd),verticals=TRUE, horizontals=TRUE,
pch=20,do.p=FALSE, col.hor="blue", col.vert="blue",lwd=3)
legend(500, 0.9, legend = c("observed","generated"), col = c("yellowgreen","blue"),
pch=c(NA,NA), lty = c(1, 1),lwd=c(3,3),bty="n")
box()
#####################################
## CDF plots of generated and fitted ##
#####################################
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
plot(ecdf(dd),xlim=c(0,700),verticals=TRUE, horizontals=TRUE,
pch=45,do.p=FALSE, col.hor="blue", col.vert="blue",lwd=3,
main="CDFs generated Vs fitted",
xlab="Rainfall (mm)",ylab="Probability")
lines(cdf, lty=3, col="red",lwd=4)
legend(500,.9, legend = c("generated","fitted"), col = c("blue","red"),
pch=c(NA,NA), lty = c(1, 3),lwd=c(3,4),bty="n")
box()
```

## C.9 Bivariate skew-$t$ copula

This code is used to transform data to uniform and to construct empirical and theoretical copulas.

```
## Meteorology Station: Hume and Beechworth
## Data used: Monthly rainfall (1928 to 1985 = 58 years)
## ---------------------------------------------------------- ##

##  Ranked the observed positive data
rk_hume <- rank(hume_pos);
hume_pos_rk <- cbind(hume_pos, rk_hume)
head(hume_pos_rk); tail(hume_pos_rk)

rk_beec <- rank(beec_pos);
beec_pos_rk <- cbind(beec_pos, rk_beec)
head(beec_pos_rk); tail(beec_pos_rk)

## --------------   Data transformation   ------------------- ##
## transform each positive pairs into uniform data
## using gamma marginals (pgamma)
## ---------------------------------------------------------- ##
alp_hume <- alpha.beta1(hume_pos)[1]
beta_hume <- alpha.beta1(hume_pos)[2]

hume_uni <-
transform(hume_pos,newhumeuni=pgamma(hume_pos,shape= alp_hume,
                        scale=beta_hume))
head(hume_uni,10); tail(hume_uni,10)
cbind(hume_pos,hume_uni)

alp_beec <- alpha.beta1(beec_pos)[1]
beta_beec <- alpha.beta1(beec_pos)[2]

beec_uni  <-
transform(beec_pos,newbeecuni=pgamma(beec_pos,shape= alp_beec,
                        scale=beta_beec))
head(beec_uni,10); tail(beec_uni,10)

## uniform data
obs_uni <- cbind(hume_uni[,2],beec_uni[,2])
head(obs_uni); tail(obs_uni)

## ---------------------------------------------------------- ##
##  Alternative method
## ---------------------------------------------------------- ##
x <- cbind(hume_pos, beec_pos)
```

```
b <- c(5,10)

## estimate alpha and beta using mle method
loglik.marg <-
function(b,x) sum(dgamma(x,shape = b[1],scale = b[2],log = TRUE))

ctrl <- list(fnscale = -1)

b1hat <- optim(c(1,5), fn = loglik.marg, x = dat[, 1], control = ctrl)$par

b2hat <- optim(c(1,5), fn = loglik.marg, x = dat[, 2], control = ctrl)$par

rbind(b1hat,b2hat) ## estimate alpha and beta using mle method

## ------------------------------------------------------------------ ##
b1hat <- c(alp_hume,beta_hume) ; b2hat <- c(alp_beec,beta_beec)
rbind(b1hat,b2hat)
dat <- cbind(hume_pos,beec_pos)

## transform observed data to uniform data using pgamma
udat <- cbind(pgamma(dat[, 1], shape = b1hat[1], scale = b1hat[2]),
              pgamma(dat[, 2], shape = b2hat[1], scale = b2hat[2]))
head(cbind(hume_pos,udat[,1])); tail(cbind(hume_pos,udat[,1]))
head(cbind(beec_pos,udat[,2])); tail(cbind(beec_pos,udat[,2]))

## ------------------------------------------------------------------ ##
##                    EMPIRICAL COPULA
## ------------------------------------------------------------------ ##
## This code is used to construct the empirical (historical)
## copula values (C_n).  The empirical copula is constructed based on
## the uniformised observed data The observed data are transformed
## to uniform data, in this case using  gamma marginals.
## Then the uniformised data is ranked.

R_i <- rank(obs_uni[,1])
S_i <- rank(obs_uni[,2])
rankg <- cbind(R_i,S_i)
head(cbind(R_i,obs_uni[,1]))
tail(cbind(R_i,obs_uni[,1]))
head(cbind(S_i,obs_uni[,2]))
tail(cbind(S_i,obs_uni[,2]))
```

```
dat_uni_obs  <- cbind(obs_uni[,1],obs_uni[,2])
dat_rank_obs <- cbind(R_i,S_i)

## rk  - ranked data
## uni - uniform data
install.packages(repos=NULL,pkgs="c:\\Tinn-R\\Rlab_2.9.0.zip")
library(Rlab)

par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
plot(dat_uni_obs[,1], dat_uni_obs[,2],
main ="Uniform data of Hume and Beechworth",pch=19,
xlab="C_empirical", ylab="C_theoretical",xaxs='i', yaxs='i',
xlim=c(0,1),ylim=c(0,1),cex.axis=1.2)
abline(0.1, 0, col="red", lty=2, lwd=3)
abline(0.9, 0, col="red", lty=2, lwd=3)
abline(0, 1, col="red", lty=1, lwd=2)
xline(0.1, lty=2, col="red", lwd=3)
xline(0.9, lty=2, col="red", lwd=3)
box()

## ------------------------------------------------------------------##
## Theoretical copula
## ------------------------------------------------------------------##
## Using Azzalini code
## generate the bivariate skew-t and transformed to univariate skew-t
library(psych)   ## need this package to find skewness
library(mnormt)
library(mvtnorm) ## sn package need this
library(sn)      ## need this package to use rmst and pmst

uni_biskewt2 <- function(n,r,a)
{ # setting parameters
  #r1  <- r2  # correlation
  #a1  <- a2  # skewness
  df  <- 5    # degrees of freedom

  xi    <- c(0,0)                       ## mean
  Omega <- matrix(c(1,r,r,1),nrow=2)    ## var-cov
  #alpha <- c(a1,a2)                     ## skewness
  alpha <- c(a,a)

  # set.seed(1122)
```

```
  # rmst - to generate a bivariate skewed t
  rand     <- rmst(n,xi, Omega, alpha, df)[1:n,1:2]
  unisk1   <- vector("numeric",length=n)
  unisk2   <- vector("numeric",length=n)

  for (i in 1:n)
  { # msn.marginal - to estimate the values of alpha
    #                from the marginals
    marg1 <- msn.marginal(xi,Omega,alpha,1)$alpha
    marg2 <- msn.marginal(xi,Omega,alpha,2)$alpha

    # pst - to transform each column of bivariate skew-t
    #       to uniform
    unisk1[i] <- pst(rand[i,1],0,1,marg1, df)
    unisk2[i] <- pst(rand[i,2],0,1,marg2, df) }

    # output
    # cbind(rand[,1],unisk1,rand[,2],unisk2)
    cbind(unisk1,unisk2)
}

uni_theory <- round(uni_biskewt2(10000,0.965,-1.95),3)
head(uni_theory); tail(uni_theory)

## ------------------------------------------------------------##
## Ranking the uniform data from theoretical (formula)
R_i_theo <- rank(uni_theory[,1])
S_i_theo <- rank(uni_theory[,2])

head(cbind(R_i_theo,uni_theory[,1]))
tail(cbind(R_i_theo,uni_theory[,1]))
head(cbind(S_i_theo,uni_theory[,2]))
tail(cbind(S_i_theo,uni_theory[,2]))

dat_uni_theo  <- cbind(uni_theory[,1],uni_theory[,2])
dat_rank_theo <- cbind(R_i_theo,S_i_theo)

head(dat_uni_theo); tail(dat_uni_theo)
head(dat_rank_theo); tail(dat_rank_theo)

## rk  - ranked data
## uni - uniform data
```

```
## ------------------------------------------------------------##
## dt1 = ranking pairs of positive observed data
## dt2 = ranking pairs of positive generated data
A_star     <- function(dt1,dt2)
{ ## empirical copula
  n        <- nrow(dt1)
  cdt.1a   <- trunc(dt1[,1]/(n+1)*10)
  cdt.1b   <- trunc(dt1[,2]/(n+1)*10)
  Astar.1a <- table(cdt.1a,cdt.1b)/n
  Astar.1b <- matrix(Astar.1a, nrow=10)

  ## theoretical copula
  m        <- nrow(dt2)
  cdt.2a   <- trunc(dt2[,1]/(m+1)*10)
  cdt.2b   <- trunc(dt2[,2]/(m+1)*10)
  Astar.2a <- table(cdt.2a,cdt.2b)/m
  Astar.2b <- matrix(Astar.2a, nrow=10)

  round(cbind(c(Astar.1b),c(Astar.2b)),3)
}

dat_rank_theo <- cbind(R_i_theo,S_i_theo)
dat_rank <- cbind(R_i,S_i)

AA <- A_star(dat_rank, dat_rank_theo);
colnames(AA) <- c("empirical","theory")
head(AA); tail(AA)

## checking the sum of each empirical and theoretical must be close to 1
sum(AA[,1]);sum(AA[,2])

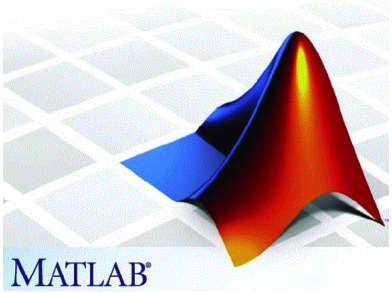
## Scatter plot of theory and empirical copula
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
plot(AA[,1], AA[,2], main ="Theoretical against empirical  copula",
pch=19,xlab="C_empirical", ylab="C_theoretical",xaxs='i', yaxs='i',
xlim=c(0,0.1),ylim=c(0,0.1),cex.axis=1.2)
abline(0, 1, col="red",lwd=3)  ## fitted line
box()

## Kolmogorov-Smirnov test - Goodness of fit test
ks.test(AA[,1],AA[,2])
```

```
## Cumulative density plots
par(mar=c(4,4,2,1.2),oma=c(0,0,0,0),xaxs="i", yaxs="i")
plot(ecdf(AA[,1]), verticals= TRUE,do.p = FALSE, col.h="blue",col.v="blue",
lwd=3,main ="CDF for theoretical and empirical  copula",
xlab="Rainfall (mm)", ylab="Cumulative Percent", xlim=c(0,.1),
xaxs ='i', yaxs ='i')
lines(ecdf(AA[,2]),verticals= TRUE,do.p = FALSE,col.h="red",col.v="red",
lwd=3)
legend(.07,.9,c("empirical","theory"),col=c("blue","red"),lwd=c(3,3),bty="n")
box()
```

# Appendix D

# MATLAB package




MATLAB is an abbreviation of MATrix LABoratory. It was first introduced by Cleve Moler in the mid 1970s as a teaching tool for computing matrices and solving system of linear equations. In 1980, Moler and his two colleagues commercialized the product and now it is a licensed programming language from The MathWorks, Natick, MA (www.mathworks.com). Since then it has been developed greatly to cover vast application including graphical presentation, data analysis and numerical computation. The MATLAB programming is considered a high-level computing language and can be integrated with other languages and application such as $R$ language and $LaTeX$. MATLAB also capable for performing complex computation more efficiently than traditional programming languages such as C, C++, and Fortran.

## D.1 MATLAB code for solving $C(X^n) = C_n(r)$

This code is used to calculate the generalised correlation value, refer to Getz [38].

```
function F = cxncnr(r)
n = 4;
% terms
t1 = (1+(n-1)*r)*log((1+(n-1)*r));
t2 = (n-1)*(1-r)*log(1-r);

%f = term - cxn
f = (t1+t2)/(n*log(n)) - 0.05011007

F = [f];
% r0 = [0.5] ; r = fsolve(@cxncnr,r0)
```

## D.2 MATLAB code for the Alouini density

This code is used to sketch a graph of the Alouini probability density function and cumulative distribution function in a particular case.

```
%% Alouini PDF

alpha = 1.6443;
lambda1 = 12.1462;
lambda2 = 74.4338;
r = lambda1/lambda2;
c = r^(alpha);
delta0 = 1;
delta1 = alpha*(1-r);
delta2 = (alpha+1)*(1-r)*delta1/2;
delta3 = (alpha+2)*(1-r)*delta2/3;
```

```
delta4 = (alpha+3)*(1-r)*delta3/4;
delta5 = (alpha+4)*(1-r)*delta4/5;
delta6 = (alpha+5)*(1-r)*delta5/6;
delta7 = (alpha+6)*(1-r)*delta6/7;
delta8 = (alpha+7)*(1-r)*delta7/8;
delta9 = (alpha+8)*(1-r)*delta8/9;
delta10 = (alpha+9)*(1-r)*delta9/10;
delta11 = (alpha+10)*(1-r)*delta10/11;
delta12 = (alpha+11)*(1-r)*delta11/12;
delta13 = (alpha+12)*(1-r)*delta12/13;
delta14 = (alpha+13)*(1-r)*delta13/14;
delta15 = (alpha+14)*(1-r)*delta14/15;
delta16 = (alpha+15)*(1-r)*delta15/16;
delta17 = (alpha+16)*(1-r)*delta16/17;
delta18 = (alpha+17)*(1-r)*delta17/18;
delta19 = (alpha+18)*(1-r)*delta18/19;
delta20 = (alpha+19)*(1-r)*delta19/20;
delta21 = (alpha+20)*(1-r)*delta20/21;
delta22 = (alpha+21)*(1-r)*delta21/22;
delta23 = (alpha+22)*(1-r)*delta22/23;
delta24 = (alpha+23)*(1-r)*delta23/24;
delta25 = (alpha+24)*(1-r)*delta24/25;
delta26 = (alpha+25)*(1-r)*delta25/26;
delta27 = (alpha+26)*(1-r)*delta26/27;
delta28 = (alpha+27)*(1-r)*delta27/28;
delta29 = (alpha+28)*(1-r)*delta28/29;
delta30 = (alpha+29)*(1-r)*delta29/30;
delta31 = (alpha+30)*(1-r)*delta30/31;
delta32 = (alpha+31)*(1-r)*delta31/32;
delta33 = (alpha+32)*(1-r)*delta32/33;
delta34 = (alpha+33)*(1-r)*delta33/34;
delta35 = (alpha+34)*(1-r)*delta34/35;
delta36 = (alpha+35)*(1-r)*delta35/36;
delta37 = (alpha+36)*(1-r)*delta36/37;
delta38 = (alpha+37)*(1-r)*delta37/38;
delta39 = (alpha+38)*(1-r)*delta38/39;
delta40 = (alpha+39)*(1-r)*delta39/40;
delta41 = (alpha+40)*(1-r)*delta40/41;
delta42 = (alpha+41)*(1-r)*delta41/42;
delta43 = (alpha+42)*(1-r)*delta42/43;
delta44 = (alpha+43)*(1-r)*delta43/44;
delta45 = (alpha+44)*(1-r)*delta44/45;
```

```
delta46 = (alpha+45)*(1-r)*delta45/46;
delta47 = (alpha+46)*(1-r)*delta46/47;
delta48 = (alpha+47)*(1-r)*delta47/48;
delta49 = (alpha+48)*(1-r)*delta48/49;
x = 0:5:700;
y0 = c.*delta0.*x.^(2*alpha-1).*exp(-x./lambda1)./lambda1^(2*alpha)./gamma(2*alpha);
y1 = c.*delta1.*x.^(2*alpha).*exp(-x./lambda1)./lambda1^(2*alpha+1)./gamma(2*alpha+1);
y2 = c.*delta2.*x.^(2*alpha+1).*exp(-x./lambda1)./lambda1^(2*alpha+2)./gamma(2*alpha+2);
y3 = c.*delta3.*x.^(2*alpha+2).*exp(-x./lambda1)./lambda1^(2*alpha+3)./gamma(2*alpha+3);
y4 = c.*delta4.*x.^(2*alpha+3).*exp(-x./lambda1)./lambda1^(2*alpha+4)./gamma(2*alpha+4);
y5 = c.*delta5.*x.^(2*alpha+4).*exp(-x./lambda1)./lambda1^(2*alpha+5)./gamma(2*alpha+5);
y6 = c.*delta6.*x.^(2*alpha+5).*exp(-x./lambda1)./lambda1^(2*alpha+6)./gamma(2*alpha+6);
y7 = c.*delta7.*x.^(2*alpha+6).*exp(-x./lambda1)./lambda1^(2*alpha+7)./gamma(2*alpha+7);
y8 = c.*delta8.*x.^(2*alpha+7).*exp(-x./lambda1)./lambda1^(2*alpha+8)./gamma(2*alpha+8);
y9 = c.*delta9.*x.^(2*alpha+8).*exp(-x./lambda1)./lambda1^(2*alpha+9)./gamma(2*alpha+9);
y10 = c.*delta10.*x.^(2*alpha+9).*exp(-x./lambda1)./lambda1^(2*alpha+10)./gamma(2*alpha+10);
y11 = c.*delta11.*x.^(2*alpha+10).*exp(-x./lambda1)./lambda1^(2*alpha+11)./gamma(2*alpha+11);
y12 = c.*delta12.*x.^(2*alpha+11).*exp(-x./lambda1)./lambda1^(2*alpha+12)./gamma(2*alpha+12);
y13 = c.*delta13.*x.^(2*alpha+12).*exp(-x./lambda1)./lambda1^(2*alpha+13)./gamma(2*alpha+13);
y14 = c.*delta14.*x.^(2*alpha+13).*exp(-x./lambda1)./lambda1^(2*alpha+14)./gamma(2*alpha+14);
y15 = c.*delta15.*x.^(2*alpha+14).*exp(-x./lambda1)./lambda1^(2*alpha+15)./gamma(2*alpha+15);
y16 = c.*delta16.*x.^(2*alpha+15).*exp(-x./lambda1)./lambda1^(2*alpha+16)./gamma(2*alpha+16);
y17 = c.*delta17.*x.^(2*alpha+16).*exp(-x./lambda1)./lambda1^(2*alpha+17)./gamma(2*alpha+17);
y18 = c.*delta18.*x.^(2*alpha+17).*exp(-x./lambda1)./lambda1^(2*alpha+18)./gamma(2*alpha+18);
y19 = c.*delta19.*x.^(2*alpha+18).*exp(-x./lambda1)./lambda1^(2*alpha+19)./gamma(2*alpha+19);
y20 = c.*delta20.*x.^(2*alpha+19).*exp(-x./lambda1)./lambda1^(2*alpha+20)./gamma(2*alpha+20);
y21 = c.*delta21.*x.^(2*alpha+20).*exp(-x./lambda1)./lambda1^(2*alpha+21)./gamma(2*alpha+21);
y22 = c.*delta22.*x.^(2*alpha+21).*exp(-x./lambda1)./lambda1^(2*alpha+22)./gamma(2*alpha+22);
y23 = c.*delta23.*x.^(2*alpha+22).*exp(-x./lambda1)./lambda1^(2*alpha+23)./gamma(2*alpha+23);
y24 = c.*delta24.*x.^(2*alpha+23).*exp(-x./lambda1)./lambda1^(2*alpha+24)./gamma(2*alpha+24);
y25 = c.*delta25.*x.^(2*alpha+24).*exp(-x./lambda1)./lambda1^(2*alpha+25)./gamma(2*alpha+25);
y26 = c.*delta26.*x.^(2*alpha+25).*exp(-x./lambda1)./lambda1^(2*alpha+26)./gamma(2*alpha+26);
y27 = c.*delta27.*x.^(2*alpha+26).*exp(-x./lambda1)./lambda1^(2*alpha+27)./gamma(2*alpha+27);
y28 = c.*delta28.*x.^(2*alpha+27).*exp(-x./lambda1)./lambda1^(2*alpha+28)./gamma(2*alpha+28);
y29 = c.*delta29.*x.^(2*alpha+28).*exp(-x./lambda1)./lambda1^(2*alpha+29)./gamma(2*alpha+29);
y30 = c.*delta30.*x.^(2*alpha+29).*exp(-x./lambda1)./lambda1^(2*alpha+30)./gamma(2*alpha+30);
y31 = c.*delta31.*x.^(2*alpha+30).*exp(-x./lambda1)./lambda1^(2*alpha+31)./gamma(2*alpha+31);
y32 = c.*delta32.*x.^(2*alpha+31).*exp(-x./lambda1)./lambda1^(2*alpha+32)./gamma(2*alpha+32);
y33 = c.*delta33.*x.^(2*alpha+32).*exp(-x./lambda1)./lambda1^(2*alpha+33)./gamma(2*alpha+33);
y34 = c.*delta34.*x.^(2*alpha+33).*exp(-x./lambda1)./lambda1^(2*alpha+34)./gamma(2*alpha+34);
y35 = c.*delta35.*x.^(2*alpha+34).*exp(-x./lambda1)./lambda1^(2*alpha+35)./gamma(2*alpha+35);
y36 = c.*delta36.*x.^(2*alpha+35).*exp(-x./lambda1)./lambda1^(2*alpha+36)./gamma(2*alpha+36);
```

```
y37 = c.*delta37.*x.^(2*alpha+36).*exp(-x./lambda1)./lambda1^(2*alpha+37)./gamma(2*alpha+37);
y38 = c.*delta38.*x.^(2*alpha+37).*exp(-x./lambda1)./lambda1^(2*alpha+38)./gamma(2*alpha+38);
y39 = c.*delta39.*x.^(2*alpha+38).*exp(-x./lambda1)./lambda1^(2*alpha+39)./gamma(2*alpha+39);
y40 = c.*delta40.*x.^(2*alpha+39).*exp(-x./lambda1)./lambda1^(2*alpha+40)./gamma(2*alpha+40);
y41 = c.*delta41.*x.^(2*alpha+40).*exp(-x./lambda1)./lambda1^(2*alpha+41)./gamma(2*alpha+41);
y42 = c.*delta42.*x.^(2*alpha+41).*exp(-x./lambda1)./lambda1^(2*alpha+42)./gamma(2*alpha+42);
y43 = c.*delta43.*x.^(2*alpha+42).*exp(-x./lambda1)./lambda1^(2*alpha+43)./gamma(2*alpha+43);
y44 = c.*delta44.*x.^(2*alpha+43).*exp(-x./lambda1)./lambda1^(2*alpha+44)./gamma(2*alpha+44);
y45 = c.*delta45.*x.^(2*alpha+44).*exp(-x./lambda1)./lambda1^(2*alpha+45)./gamma(2*alpha+45);
y46 = c.*delta46.*x.^(2*alpha+45).*exp(-x./lambda1)./lambda1^(2*alpha+46)./gamma(2*alpha+46);
y47 = c.*delta47.*x.^(2*alpha+46).*exp(-x./lambda1)./lambda1^(2*alpha+47)./gamma(2*alpha+47);
y48 = c.*delta48.*x.^(2*alpha+47).*exp(-x./lambda1)./lambda1^(2*alpha+48)./gamma(2*alpha+48);
y49 = c.*delta49.*x.^(2*alpha+48).*exp(-x./lambda1)./lambda1^(2*alpha+49)./gamma(2*alpha+49);
y = y0 + y1 + y2 + y3 + y4 + y5 + y6 + y7 + y8 + y9 + ...
+ y10 + y11 + y12 + y13 + y14 + y15 + y16 + y17 + y18 + y19 + ...
+ y20 + y21 + y22 + y23 + y24 + y25 + y26 + y27 + y28 + y29 + ...
+ y30 + y31 + y32 + y33 + y34 + y35 + y36 + y37 + y38 + y39 + ...
+ y40 + y41 + y42 + y43 + y44 + y45 + y46 + y47 + y48 + y49;
plot(x,y,'Color','blue','LineWidth',1.2)


%% Alouini CDF
alpha = 1.6443;
lambda1 = 12.1462;
lambda2 = 74.4338;
r = lambda1/lambda2;
c = r^(alpha);
delta0 = 1;
delta1 = alpha*(1-r);
delta2 = (alpha+1)*(1-r)*delta1/2;
delta3 = (alpha+2)*(1-r)*delta2/3;
delta4 = (alpha+3)*(1-r)*delta3/4;
delta5 = (alpha+4)*(1-r)*delta4/5;
delta6 = (alpha+5)*(1-r)*delta5/6;
delta7 = (alpha+6)*(1-r)*delta6/7;
delta8 = (alpha+7)*(1-r)*delta7/8;
delta9 = (alpha+8)*(1-r)*delta8/9;
delta10 = (alpha+9)*(1-r)*delta9/10;
delta11 = (alpha+10)*(1-r)*delta10/11;
delta12 = (alpha+11)*(1-r)*delta11/12;
delta13 = (alpha+12)*(1-r)*delta12/13;
delta14 = (alpha+13)*(1-r)*delta13/14;
```

```
delta15 = (alpha+14)*(1-r)*delta14/15;
delta16 = (alpha+15)*(1-r)*delta15/16;
delta17 = (alpha+16)*(1-r)*delta16/17;
delta18 = (alpha+17)*(1-r)*delta17/18;
delta19 = (alpha+18)*(1-r)*delta18/19;
delta20 = (alpha+19)*(1-r)*delta19/20;
delta21 = (alpha+20)*(1-r)*delta20/21;
delta22 = (alpha+21)*(1-r)*delta21/22;
delta23 = (alpha+22)*(1-r)*delta22/23;
delta24 = (alpha+23)*(1-r)*delta23/24;
delta25 = (alpha+24)*(1-r)*delta24/25;
delta26 = (alpha+25)*(1-r)*delta25/26;
delta27 = (alpha+26)*(1-r)*delta26/27;
delta28 = (alpha+27)*(1-r)*delta27/28;
delta29 = (alpha+28)*(1-r)*delta28/29;
delta30 = (alpha+29)*(1-r)*delta29/30;
delta31 = (alpha+30)*(1-r)*delta30/31;
delta32 = (alpha+31)*(1-r)*delta31/32;
delta33 = (alpha+32)*(1-r)*delta32/33;
delta34 = (alpha+33)*(1-r)*delta33/34;
delta35 = (alpha+34)*(1-r)*delta34/35;
delta36 = (alpha+35)*(1-r)*delta35/36;
delta37 = (alpha+36)*(1-r)*delta36/37;
delta38 = (alpha+37)*(1-r)*delta37/38;
delta39 = (alpha+38)*(1-r)*delta38/39;
delta40 = (alpha+39)*(1-r)*delta39/40;
delta41 = (alpha+40)*(1-r)*delta40/41;
delta42 = (alpha+41)*(1-r)*delta41/42;
delta43 = (alpha+42)*(1-r)*delta42/43;
delta44 = (alpha+43)*(1-r)*delta43/44;
delta45 = (alpha+44)*(1-r)*delta44/45;
delta46 = (alpha+45)*(1-r)*delta45/46;
delta47 = (alpha+46)*(1-r)*delta46/47;
delta48 = (alpha+47)*(1-r)*delta47/48;
delta49 = (alpha+48)*(1-r)*delta48/49;
x = 0:5:700;
z0 = c*delta0*gammainc(x/lambda1,2*alpha);
z1 = c*delta1*gammainc(x/lambda1,2*alpha+1);
z2 = c*delta2*gammainc(x/lambda1,2*alpha+2);
z3 = c*delta3*gammainc(x/lambda1,2*alpha+3);
z4 = c*delta4*gammainc(x/lambda1,2*alpha+4);
z5 = c*delta5*gammainc(x/lambda1,2*alpha+5);
```

```
z6 = c*delta6*gammainc(x/lambda1,2*alpha+6);
z7 = c*delta7*gammainc(x/lambda1,2*alpha+7);
z8 = c*delta8*gammainc(x/lambda1,2*alpha+8);
z9 = c*delta9*gammainc(x/lambda1,2*alpha+9);
z10 = c*delta10*gammainc(x/lambda1,2*alpha+10);
z11 = c*delta11*gammainc(x/lambda1,2*alpha+11);
z12 = c*delta12*gammainc(x/lambda1,2*alpha+12);
z13 = c*delta13*gammainc(x/lambda1,2*alpha+13);
z14 = c*delta14*gammainc(x/lambda1,2*alpha+14);
z15 = c*delta15*gammainc(x/lambda1,2*alpha+15);
z16 = c*delta16*gammainc(x/lambda1,2*alpha+16);
z17 = c*delta17*gammainc(x/lambda1,2*alpha+17);
z18 = c*delta18*gammainc(x/lambda1,2*alpha+18);
z19 = c*delta19*gammainc(x/lambda1,2*alpha+19);
z20 = c*delta20*gammainc(x/lambda1,2*alpha+20);
z21 = c*delta21*gammainc(x/lambda1,2*alpha+21);
z22 = c*delta22*gammainc(x/lambda1,2*alpha+22);
z23 = c*delta23*gammainc(x/lambda1,2*alpha+23);
z24 = c*delta24*gammainc(x/lambda1,2*alpha+24);
z25 = c*delta25*gammainc(x/lambda1,2*alpha+25);
z26 = c*delta26*gammainc(x/lambda1,2*alpha+26);
z27 = c*delta27*gammainc(x/lambda1,2*alpha+27);
z28 = c*delta28*gammainc(x/lambda1,2*alpha+28);
z29 = c*delta29*gammainc(x/lambda1,2*alpha+29);
z30 = c*delta30*gammainc(x/lambda1,2*alpha+30);
z31 = c*delta31*gammainc(x/lambda1,2*alpha+31);
z32 = c*delta32*gammainc(x/lambda1,2*alpha+32);
z33 = c*delta33*gammainc(x/lambda1,2*alpha+33);
z34 = c*delta34*gammainc(x/lambda1,2*alpha+34);
z35 = c*delta35*gammainc(x/lambda1,2*alpha+35);
z36 = c*delta36*gammainc(x/lambda1,2*alpha+36);
z37 = c*delta37*gammainc(x/lambda1,2*alpha+37);
z38 = c*delta38*gammainc(x/lambda1,2*alpha+38);
z39 = c*delta39*gammainc(x/lambda1,2*alpha+39);
z40 = c*delta40*gammainc(x/lambda1,2*alpha+40);
z41 = c*delta41*gammainc(x/lambda1,2*alpha+41);
z42 = c*delta42*gammainc(x/lambda1,2*alpha+42);
z43 = c*delta43*gammainc(x/lambda1,2*alpha+43);
z44 = c*delta44*gammainc(x/lambda1,2*alpha+44);
z45 = c*delta45*gammainc(x/lambda1,2*alpha+45);
z46 = c*delta46*gammainc(x/lambda1,2*alpha+46);
z47 = c*delta47*gammainc(x/lambda1,2*alpha+47);
```

```
z48 = c*delta48*gammainc(x/lambda1,2*alpha+48);
z49 = c*delta49*gammainc(x/lambda1,2*alpha+49);
z = z0 + z1 + z2 + z3 + z4 + z5 + z6 + z7 + z8 + z9 + ...
z10 + z11 + z12 + z13 + z14 + z15 + z16 + z17 + z18 + z19 + ...
z20 + z21 + z22 + z23 + z24 + z25 + z26 + z27 + z28 + z29 + ...
z30 + z31 + z32 + z33 + z34 + z35 + z36 + z37 + z38 + z39 + ...
z40 + z41 + z42 + z43 + z44 + z45 + z46 + z47 + z48 + z49;
plot(x,z,'Color','blue','LineWidth',1.2)
```

# Bibliography


[1] M. Abramowitz and I.E. Stegun. *Handbook of Mathematical Functions*. Dover Publications, U.S. Dept. Commerce, 9th edition, 1972.

[2] H. Alexandersson. A simple stochastic model of the precipitation process. *Journal of Climate and Applied Meteorology*, 24(12):1285–1295, 1985.

[3] M.S. Alouini, A. Abdi, and M. Kaveh. Sum of gamma variates and performance of wireless communication systems over Nakagami-fading channels. *IEEE Trans. Veh. Technol.*, 50(6):1471–1480, 2001.

[4] T.W. Anderson and D.A. Darling. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23, 1952.

[5] T.W. Anderson and D.A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[6] ASISTM. Australian School Innovation in Science, Technology and Mathematics. http://www.murrayusers.sa.gov.au, 2007. [Online; accessed 24-Oct-2007].

[7] Australian Bureau of Statistics. Water and the Murray–Darling Basin - A Statistical Profile (4610.0.55.007), 2000-01 to 2005-06. `http://www.abs.gov.au/ausstats`, 2008. [Online; accessed 10-Jan-2010].

[8] A. Azzalini. *R package* **sn***: The skew-normal and skew-t distributions (version 0.4-12)*. Università di Padova, Italia, 2009. [Online; accessed 13-Apr-2010].

[9] A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B*, 65(2):367–389, 2003.

[10] A. Azzalini and M.G. Genton. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76(1):106–129, 2008.

[11] R.H. Bartels, S.D. Horn, A.M. Liebetrau, and W.L. Harris. A computational investigation of Conover's Kolmogorov-Smirnov test for discrete distributions. *Journal of Statistical Computation and Simulation*, 7(2):151–162, 1978.

[12] E. Bekele. Markov chain modelling and ENSO influences on the rainfall seasons of Ethiopia. pages 25–35, 2002.

[13] J.W. Boland. Windfarm output variability in South Australia. In *14th International Conference on Applied Simulation and Modelling*, 15-17 July, Benalmadena, Spain, 2005.

[14] J.A. Bruhn, W.E. Fry, and G.W. Fick. Simulation of daily weather data using theoretical probability distributions. *Journal of Applied Meteorology*, 19:1029–1036, 1980.

[15] F. Cajori. Historical note on the Newton-Raphson method of approximation. *The American Mathematical Monthly*, 18(2):29–32, 1911.

[16] I.M. Chakravarti, R.G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics*, volume I. John Wiley and Sons, New York, 1967.

[17] R.E. Chandler. On the use of generalized linear models for interpreting climate variability. *Environmetrics*, 16:699–715, 2005.

[18] T.G. Chapman. Stochastic models for daily rainfall in the western pacific. *Mathematics and Computers in Simulation*, 43(3-6):351–358, 1997.

[19] E.H. Chin. Modelling daily precipitation occurrence process in markov chain. *Water Resources Research*, 13(6):949–956, 1977.

[20] W.G. Cochran. The $\chi^2$ test of goodness of fit. *Annals of Mathematical Statistics*, 3:315–345, Dec. 1952.

[21] W.G. Cochran. Some methods for strengthening the common $\chi^2$ tests. *Biometrics*, 10(4):417–451, Dec. 1954.

[22] D. Connell. *Water politics in the Murray–Darling Basin*. Federation Press, Annandale, N.S.W., 2007.

[23] P.S.P. Cowpertwait. A generalized spatial-temporal model of rainfall based on a clustered point process. *Proc. R. Soc. Lond. A*, 450(1938):163–175, 1995.

[24] P.S.P. Cowpertwait. A Neyman-Scott model with continuous distributions of storm types. *ANZIAM J. 51 (EMAC2009)*, pages C97–C108, 2010.

[25] P.S.P. Cowpertwait, V. Isham, and C. Onof. Point process models of rainfall: Developments for fine-scale structure. *Proc. R. Soc. A*, 463:2569–2587, 2007.

[26] D.R. Cox and V. Isham. A simple spatial-temporal model of rainfall. *Proc. R. Soc. Lond. A*, 415(1849):317–328, 1988.

[27] P. Crabb. *Murray-Darling Basin Resources*. Murray-Darling Basin Commission, Canberra, A.C.T., 1997.

[28] R.B. D'Agostino and M.A. Stephens, editors. *Goodness-of-fit techniques*. 1986.

[29] W.W. Daniel. *Applied Nonparametric Statistics*. PWS-Kent Publishing Company, Boston, 2nd edition, 1990.

[30] P.K. Dunn. Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24(10):1231–1239, 2004.

[31] P.R. Edelsten. A stochastic model of the weather at Hurley in south-east England. *Meteorol. Mag.*, 105:206–214, 1976.

[32] P. Embrechts, F. Lindskog, and A. Mc Neil. *Modelling Dependence with Copulas and Applications to Risk Management In: Handbook of Heavy Tailed Distributions in Finance*, chapter 8, pages 329–384. Elsevier, 2003.

[33] J.C. Faria. *Resources of Tinn-R GUI/Editor for R Environment.* UESC, Ilheus, Bahia, Brasil, 2009.

[34] A.-C. Favre, S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée. Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40(W01101), 2004.

[35] E. Furman. On a multivariate gamma distribution. *Statistics and Probability Letters*, 78:2353–2360, 2008.

[36] C. Genest and A.-C. Favre. Everything you always wanted to know about copula modelling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007.

[37] C. Genest, A.-C. Favre, J. Béliveau, and C. Jacques. Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resources Research*, 43(W09401), 2007.

[38] W.M. Getz. Correlative coherence analysis: Variation from intrinsic and extrinsic sources in competing populations. *Theoretical Population Biology*, 64 (1):89–99, 2003.

[39] GEWEX. Gewex Energy and Water Cycle Experiment. http://www.gewex.org, 2007. [Online; accessed 24-Oct-2007].

[40] I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series and Products*. Academic Press, San Diego, CA, 6th edition, 2000.

[41] L. Guenni and A. Bardossy. A two steps disaggregation method for highly seasonal monthly rainfall. *Stochastic Environmental Research and Risk Assessment*, 16:188–206, 2002.

[42] P. Guttorp. *Stochastic modelling for scientific data*. Chapman and Hall, London, 1995.

[43] T.I. Harold, A. Sharma, and S.J. Sheather. A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resource Research*, 39(12):1343, 2003.

[44] P.G. Hoel, S.C. Port, and C.J. Stone. *Introduction to Probability Theory*. Houghton Mifflin, Boston., 1971.

[45] H. Holm and M.-S. Alouini. Sum and difference of two squared correlated Nakagami variates in connection with the McKay distribution. *IEEE Transactions On Communications*, 52(8):1367–1376, 2004.

[46] S.D. Horn. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, 33(1):237–247, 1977.

[47] G.J. Husak, J. Michaelsen, and C. Funk. Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, 27:935–944, 2007.

[48] S. Jamaludin and A.A. Jemain. Fitting daily rainfall amount in Malaysia using the normal transform distribution. *Journal of Applied Sciences*, 7(14): 1880–1886, 2007.

[49] S. Jamaludin and A.A. Jemain. Fitting daily rainfall amount in Peninsular Malaysia using several types of exponential distributions. *Journal of Applied Sciences Research*, 3(10):1027–1036, 2007.

[50] S. Jamaludin and A.A. Jemain. Fitting the statistical distribution for daily rainfall in Peninsular Malaysia based on AIC criterion. *Journal of Applied Sciences Research*, 4(12):1846–1851, 2008.

[51] B. Jorgensen. Exponential dispersion model (with discussion). *Journal of Royal Statistics Series B*, 49:127–162, 1987.

[52] B. Jorgensen. *The Theory of Dispersion Models*. Chapman and Hall, London, 1997.

[53] R.W. Katz and M.B. Parlange. Generalization of chain-dependent processes: applications to hourly precipitation. *Water Resource Research*, 31:1331–1341, 1995.

[54] M. Kirby, R. Evans, G. Walker, C. Richard, J. Coram, S. Khan, Z. Paydar, M. Mainuddin, N. McKenzie, and S. Ryan. The Shared Water Resources of the Murray–Darling Basin, 2006. [Online; accessed 20-July-2010].

[55] K. Koehler. *Moment Generating Function in Encyclopedia of Biostatistics*. John Wiley and Sons Ltd., New York, 2005.

[56] T. Kollo and G. Pettere. Parameter estimation and application of the multivariate skew $t$-copula. In P. Jaworski, F. Durante, W. Härdle, and T. Rychlik, editors, *Copula Theory and Its Applications, Proceedings of the Workshop, Warsaw, 25–26 Sept 2009*, pages 289–298, Dordrecht, 2010. Springer.

[57] L. Komzsik. *Approximation Techniques for Engineers.* CRC Press, United States of America, 2007.

[58] S. Kotz and J.W. Adams. Distribution of sum of identically distributed exponentially correlated gamma-variables. *The Annals of Mathematical Statistics*, 35(1):277–283, 1964.

[59] S. Kotz, N.L. (eds.) Johnson, and C.B. Read. *The Encyclopedia of Statistical Sciences. Supplement volume.* John Wiley and Sons Ltd., New York, 1989.

[60] U. Lall, B. Rajagopalan, and D.G. Tarboton. A nonparametric wet/dry spell model for resampling daily precipitation. *Water Resour. Res.*, 32(9):2803–2823, 1996.

[61] S.M. Lennox, P.K. Dunn, B.D. Power, and P. Devoil. A statistical distribution for modelling rainfall with promising applications in crop science. In *4th International Crop Science Congress*, 26 September-1 October 2004, Brisbane, Australia, 2004.

[62] J. Martin-Vide and L. Gomez. Regionalization of Peninsular Spain based on the length of dry spells. *Int. J. Climatol.*, 19:537–555, 1999.

[63] A.M. Mathai. Storage capacity of a dam with gamma type inputs. *Ann. Inst. Statist. Math.–Part A*, 34(3):591–597, 1982.

[64] D.F. McMichael. In: The pollution problems of the River Murray. Dinner Address for Dept. of Dec. and Dev., 1973. (Aust. Govt. Representative to the

River Murray Commission 1972–1974, President of the River Murray Commission 1973–75).

[65] A. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press, Princeton, NJ, 2005.

[66] MDBC. *Integrated catchment management in the Murray–Darling Basin: 2001–2010*. Murray Darling Basin Ministerial Council, Canberra, A.C.T., 2000.

[67] MDBC. Murray–Darling Basin Commission. http://www.mdbc.gov.au, 2007. [Online; accessed 10-Jan-2010].

[68] P.G. Moschopoulos. The distribution of the sum of independent gamma random variables. *Ann. Inst. Statist. Math.–Part A*, 37:541–544, 1985.

[69] I.J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, 2003.

[70] R.B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2nd edition, 2006.

[71] A.D. Nick and L.J. Lane. *Weather generator. In Water Erosion Prediction Project*. National Erosion Research Laboratory, West Lafayette, 2.1–2.19 edition, 1989.

[72] T.J. Osborn and M. Hulme. Evidence for trends in heavy rainfall events over the UK. *Phil. Trans. R. Soc. Lond.*, 360(1796):1313–1325, 2002.

[73] M.A. Pett. *Nonparametric Statistics for Health Care Research: Statistics for small samples and unusual distributions*. Sage Publications, Thousand Oaks CA, 1997.

[74] A.N. Pettitt. A two-sample Anderson-Darling rank statistic. *Biometrika*, 63 (1):161–168, 1976. doi: 10.1093/biomet/63.1.161.

[75] A.N. Pettitt and M.A. Stephens. The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2):205–210, 1977.

[76] J. Piantadosi, J.W. Boland, and P.G. Howlett. Simulation of rainfall totals on various time scales–daily, monthly and yearly. *Environmental Modeling and Assessment*, 14(4):431–438, 2009.

[77] N.J. Potter and F.H.S. Chiew. Statistical characterisation and attribution of recent rainfall and runoff in the Murray–Darling Basin. In *18th World IMACS / MODSIM Congress*, 13-17 July, Cairns, Australia, 2009.

[78] N.J. Potter, F.H.S. Chiew, and A.J. Frost. An assessment of the severity of recent reductions in rainfall and runoff in the Murray–Darling Basin. *Journal of Hydrology*, 381:52–64, 2010.

[79] A. Prasad and S. Khan. Murray–Darling Basin dialogue on water and climate. In *River Symposium*, 6 September 2002, Brisbane, Australia, 2002.

[80] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[81] D.G. Rees. *Essential Statistics*. Chapman and Hall, United States of America, 4th edition, 2001.

[82] I. Rodriguez-Iturbe, D. Cox, and V. Isham. Some models for rainfall based on stochastic point processes. *Proc. R. Soc. A*, 410:269–288, 1987.

[83] I. Rodriguez-Iturbe, D. Cox, and V. Isham. A point process model for rainfall: further developments. *Proc. R. Soc. A*, 417:283–298, 1988.

[84] F.J. Rohlf and R.R. Sokal. *Statistical tables*. Freeman, New York, 3rd edition, 1995.

[85] J. Roldan and D.A. Woolhiser. Stochastic daily precipitation models: A comparison of occurrence processes. *Water Resource Research*, 18(5):1451–1459, 1982.

[86] K. Rosenberg. *Stochastic modelling of rainfall and generation of synthetic rainfall data at Mawson Lakes*. PhD thesis, University of South Australia, 2004.

[87] K. Rosenberg, J.W. Boland, and P.G. Howlett. Simulation of monthly rainfall totals. *ANZIAM Journal*, 46(E):E85–E104, 2004.

[88] F.W. Scholz and M.A. Stephens. K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

[89] C. Scholzel and P. Friederichs. Multivariate non-normally distributed random variables in climate research introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15(5):761–772, 2008.

[90] F. Serinaldi. Copula-based mixed models for bivariate rainfall data: An empirical study in regression perspective. *Stochastic Environmental Research and Risk Assessment*, 23(5):677–693, 2009.

[91] P. Sinclair. *The Murray. A river and it's people*. Melbourne University Press, Melbourne, 2001.

[92] R. Srikanthan and T.A. McMahon. Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, 5(4): 653–670, 2001.

[93] R. Srikanthan, T.I. Harold, A. Sharma, and T.A. McMahon. Comparison of two approaches for generation of daily rainfall data. *Stochastic Environmental Research and Risk Assessment*, 19:215–226, 2005.

[94] R.D. Stern and R. Coe. Fitting models to daily rainfall data. *Journal of Applied Meteorology*, 21:1024–1031, 1982.

[95] R.D. Stern and R. Coe. A model fitting analysis of daily rainfall data. *Journal of Royal Statistics Society A*, 147(1):1–34, 1984.

[96] A. Takemura. *Zonal Polynomials*. Inst. of Math. Statistics Lecture Notes–Monograph Series vol. 4, Hayward, California, 1984.

[97] MDBA. Guide to the proposed Basin Plan: Overview, MurrayDarling Basin Authority, Canberra. http://thebasinplan.mdba.gov.au/, 2010. [Online; accessed 2-Nov-2010].

[98] P. Todorovic and D.A. Woolhiser. A stochastic model of n-day precipitation. *Journal of Applied Meteorology*, 14:17–24, 1975.

[99] N.E. Unal, H. Aksoy, and T. Akar. Annual and monthly rainfall data generation schemes. *Stochastic Environmental Research and Risk Assessment*, 18: 245–257, 2004.

[100] A. Van Dijk and G. Podger. Towards integrated modelling in the Murray–Darling Basin. In *MODSIM*, 2005, Melbourne, Australia, 2005.

[101] VIVASA. Venture Innovation Vision Action South Australia. http://www.savethemurray.com, 2007. [Online; accessed 24-Oct-2007].

[102] S. Wigley. South Australia's Majestic Murray. http://www.abs.gov.au/ausstats, 1999. [Online; accessed 10-Jan-2010].

[103] D.S. Wilks. Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of Climate*, 3(12):1495–1501, 1990.

[104] D.S. Wilks. Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, 93 (3):153–169, 1999.

[105] D.S. Wilks and R.L. Wilby. The weather generation game: A review of stochastic weather models. *Progress in Physical Geography*, 23(3):329–357, 1999.

[106] G. Wong, M.F. Lambert, M. Leonard, and A. Metcalfe. Drought analysis using trivariate copulas conditional on climatic states. *Journal of Hydrologic Engineering*, 15(2):129–141, 2010.

[107] D.A. Woolhiser and J. Roldan. Seasonal and regional variability of parameters for stochastic dailly precipitation models: South Dakota, USA. *Water Resources Research*, 22(6):965–978, 1982.

[108] S.Y. Wu. Projecting extreme precipitation for Great Miami river watershed, Ohio. In *AWRA 2010 Spring Speciality Conference, Orlando, FL, March 29-31, 2010*, pages 1–7, American Water Resources Association, 2010.

[109] K.C. Young. A multivariate chain model for simulating climatic parameters from daily data. *J. App. Meteorol.*, 33:661–671, 1994.

[110] S.L. Zabell. On student's 1908 article. *Journal of the American Statistical Association*, 103(481):1–7, 2008.

[111] W. Zucchini and P.T. Adamson. The occurrence and severity droughts in South Africa. *Water Research Commission Report no. 91/1/84*, 1984.