An Investigation of Australian Rainfall using Extreme Value Theory

Kate Saunders
Orchid ID: 0000-0002-1436-7802

Submitted in total fulfillment of the requirements of the degree of Doctor of Philosophy

October 19th, 2018

School of Mathematics & Statistics
The University of Melbourne
Abstract

In this thesis, we use extreme value theory to fit statistical models to observations of Australian daily rainfall extremes. We build upon the existing literature by challenging the basic assumptions of these models when applied in a climate setting. The types of applications we consider range from univariate extreme value theory, to spatial extremes with dependence, and finally to an investigation of extremal dependence. The combined application content provides an in-depth investigation into our understanding of Australian rainfall extremes and the risks posed by these extreme events.

We consider how large scale climate drivers, such as El Niño Southern Oscillation (ENSO), can influence the distribution of rainfall extremes. Using observations of daily rainfall from station data, we quantify the magnitude and spatial influence of ENSO on the distribution of seasonal maximum daily rainfall. We contrast these results obtained from an at-station analysis, with those from a simple spatial model, ultimately producing maps of the region of ENSO influence.

We then consider an application where we use max-stable processes to model rainfall extremes in continuous space with dependence. We fit a max-stable process to the annual maximum daily rainfall in South East Queensland and simulate the extreme precipitation field. We quantify the severity of an historical flash flood in this region, showing that the probability of this event was significantly higher given the phase of ENSO.

Finally, we examine variation in the dependence behaviour of daily rainfall extremes. For Australia, a single dependence structure for spatial models of rainfall extremes is unrealistic. This is due to the country size, variations in climate and complexity of topography. In order help account for these variations, we present a regionalisation of Australia. In this regionalisation, locations are grouped according to similar dependence of rainfall extremes.

The overarching goal of this thesis is to improve our understanding of the risks posed by extreme rainfall events in Australia. We achieve this by utilising spatial, statistical models. However, we acknowledge that using extreme value theory for modelling real world applications in practice has challenges. We highlight these practical considerations in our applications, so that other researchers may be aware of the advantages of this kind of modelling, as well as some of its practical limitations.
Declaration

This is to certify that:

i. the thesis comprises only my original work towards the PhD except where indicated in the Preface,

ii. due acknowledgment has been made in the text to all other material used

iii. the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed

\[Signature\]

Kate Saunders
Preface

This thesis is based solely on the research I carried out during my PhD candidature at the School of Mathematics and Statistics, at the University of Melbourne. This research has been supported by the Australian Research Council through the Laureate Fellowship FL130100039, by the Commonwealth Scientific and Industrial Research Organisation through top up funding, and the Australian Centre of Excellence in Mathematical and Statistical Frontiers.

As part of the financial support I have received, I have had the opportunity to travel and to present my research, both at domestic and at international conferences. These opportunities to share my research and network have been invaluable.

I would like to acknowledge the support I received to present at the 13th International Meeting on Statistical Climatology in Canmore in 2016. At this conference, I presented an early version of the work in Chapter 6. It was here that Anthony Davison made some helpful suggestions as to how I might model the non-linearity of my response surface. This work was published in Weather and Climate Extremes (Saunders et al., 2017). I was responsible for the core ideas in this paper and was the primary author, with contribution to the paper content of 90%.

I would also like to acknowledge the support I received to visit the National Centre for Atmospheric Research (NCAR) in Boulder, Colorado in 2016. It was during this visit, that I was fortunate to spend time with Phillipe Naveau while he was there on sabbatical and I began the research that ultimately became Chapter 7. I hope to publish this work shortly.
Acknowledgments

To my three supervisors:

Thank you for your kindness and patience throughout this process. Your combined experience and guidance have been key to making this journey a successful one.

Peter, thank you for always being in my corner. Your unwavering belief in my ability has helped me to grow in confidence. The PhD is as much a personal journey as it is a professional one, and I have had the luxury of taking risks knowing that you were supporting me.

David, thank you for helping me to bridge the interdisciplinary gap. You are the reason I went to the 13th International Meeting on Statistical Climatology and visited the National Center for Atmospheric research, both of which I consider to be turning points in the PhD.

Alec, I'm extremely grateful for time I got to spend with you at CSIRO. Thank you for being my extreme value theory sounding board and for helping me to understand my research as part of the broader discipline.
‘It takes a village . . . ’

I feel blessed to have had such a fantastic network of people who have supported me throughout this process. Unfortunately, I can’t mention you all here, but below are a few special acknowledgments.

Ellen and Tim, I feel very lucky to have shared an office with you both. Your friendship has made this journey all the more enjoyable. Ellen, thank you for being my confidant. Tim, thank you for having a laugh and putting up with all my rot.

To my colleagues at the Climate College, your generosity and support has been overwhelming. Special thanks to Lisa, Mandy, Tash and Alister for both keeping me on track and abetting my procrastination.

Claire. I’m grateful to have shared so much of this journey with you, both here at University and at CSIRO.

Gala, thank you for your honesty and strengthening my voice when it was needed.

Carolyn and Fletcher, and Bretta, you each created a small person while I created this thesis. Thank you for helping me maintain perspective.

Jackson, you sparked my passion for data visualisation, thank you for your R advice and generosity.

Daryl, thank you for taking the time to proofread my thesis. I strongly believe you have made me a better writer, hopefully my reviewers think so too.

To Mum, Dad, and Mackenzie: thank you for your love and support. It has been tumultuous at times throughout this process, and you have gotten me through. I don’t say it enough, but I love you. I’m sorry I can be a ratbag at times.
# Contents

List of Figures ............................................. xv
List of Tables ............................................. xviii

1 Introduction ............................................. 1
   1.1 Motivation ........................................... 1
   1.2 Impacts of Extreme Rainfall and Flood .................. 2
   1.3 Rainfall as a Proxy for Flood ........................... 3
   1.4 Thesis Outline ........................................ 4

2 An Observational Dataset for Extremes .................... 7
   2.1 Introduction .......................................... 7
   2.2 Observational Data ..................................... 9
      2.2.1 Data Summary and Visualisation .................... 9
      2.2.2 Quality Issues .................................. 12
   2.3 Preprocessing the Data ................................ 13
      2.3.1 Reconstruction .................................. 13
      2.3.2 Extreme Observations ............................ 14
      2.3.3 Quality Control ................................. 15
      2.3.4 Homogeneity .................................... 21
   2.4 Limitations of Gridded Data ............................ 22
   2.5 Discussion ........................................... 24

3 An Introduction to Extreme Value Theory ................. 27
   3.1 Introduction .......................................... 27
   3.2 Generalised Extreme Value Distribution ................. 28
      3.2.1 Limiting Distribution for Block Maxima ............ 28
      3.2.2 Three Types ..................................... 30
   3.3 Generalised Pareto Distribution ........................ 31
      3.3.1 Limiting Distribution for Peaks Over Threshold .... 31
      3.3.2 Intuition for the Limit Distribution ............... 32
      3.3.3 Threshold Selection ............................... 32
   3.4 Limit Theorems for Stationary Series ................... 34
      3.4.1 Theorems for Stationary Series .................... 34
## CONTENTS

3.4.2 Block Maxima for Stationary Series .................... 37  
3.4.3 Peaks Over Threshold for Stationary Series ............ 37  
3.5 Extremes of Non-stationary Processes ....................... 38  
3.6 Model fitting .............................................. 39  
3.6.1 Parameter Estimation .................................... 39  
3.6.2 Model Selection ......................................... 41  
3.6.3 Model Diagnostics ....................................... 42  
3.7 Return periods ............................................. 43  
3.7.1 Stationary Series of Maxima in Time .................... 43  
3.7.2 Non-stationary Series of Maxima in Time ............... 44  
3.8 Summary .................................................. 45  

4 Non-stationarity of Seasonal Rainfall Extremes with El Niño Southern Oscillation 47  
4.1 Introduction ................................................. 47  
4.2 Spatial Generalised Extreme Value Distribution ............ 49  
4.3 Forms of ENSO Non-stationarity ............................ 50  
4.4 A Stationary GEV Distribution .............................. 52  
4.5 Symmetric Non-stationarity ................................. 54  
4.5.1 Model Selection for Univariate GEV Distribution ....... 54  
4.5.2 Model Selection for a Spatial GEV ..................... 56  
4.6 Non-stationary forms of the Spatial GEV with ENSO ...... 57  
4.6.1 Preferred from of Non-stationarity ..................... 60  
4.7 Conclusions ................................................ 61  

5 An Introduction to Max-stable Processes 63  
5.1 Introduction ................................................ 63  
5.2 Spatial Statistics .......................................... 64  
5.3 Max-Stable Processes ...................................... 66  
5.3.1 Limiting Process ....................................... 66  
5.3.2 Spectral Representation .................................. 67  
5.3.3 Parametric Forms ....................................... 68  
5.3.4 Finite–Dimensional Distribution ........................ 70  
5.3.5 Extremal Coefficient ..................................... 71  
5.4 Inference ................................................... 72  
5.4.1 Pairwise Likelihood ..................................... 73  
5.4.2 Model Selection ......................................... 75  
5.5 Max-stable Simulation ..................................... 75  
5.5.1 Unconditional Simulation ................................ 76  
5.5.2 Conditional Simulation .................................. 78  
5.5.3 Recent Methods ......................................... 81  
5.6 Alternative Methods ....................................... 81  
5.6.1 Latent variable models .................................. 82  
5.6.2 Copulas .................................................. 83
6  The Spatial Distribution of Rainfall Extremes and the Influence of El Niño Southern Oscillation

6.1 Introduction ................................................. 85
6.2 Data ..................................................... 88
6.3 Methods .................................................... 91
6.4 Results ..................................................... 93
   6.4.1 Model Selection ....................................... 93
   6.4.2 Model Validation ...................................... 95
   6.4.3 Spatial distributions for urban planning applications . 97
   6.4.4 The influence of ENSO on the spatial distribution of rainfall extremes ......................... 100
   6.4.5 Understanding historical flash flooding through conditional simulations .................. 101
6.5 Discussion and conclusions ................................ 104

7  A Regionalisation of Australia for Rainfall Extremes

7.1 Introduction ................................................. 107
7.2 Natural Resource Management Clusters ..................... 110
7.3 Clustering Dissimilarity .................................... 114
   7.3.1 Bivariate dependence ................................. 114
   7.3.2 F-madogram ........................................... 115
7.4 Clustering Methods .......................................... 116
   7.4.1 K-medoids Clustering ................................ 116
   7.4.2 Limitations of K-medoids Clustering ................. 118
   7.4.3 Hierarchical Clustering ............................... 121
   7.4.4 Suitability of Hierarchical Clustering ............... 124
7.5 Clustering Choices .......................................... 125
   7.5.1 Linkage Method ....................................... 125
   7.5.2 Dissimilarity .......................................... 126
7.6 Classification .............................................. 127
7.7 Visualising Dependence .................................... 129
7.8 Practical Considerations .................................. 130
   7.8.1 Missing Dissimilarities ............................... 130
   7.8.2 Model Fitting ......................................... 131
7.9 Regionalisation Results .................................... 132
   7.9.1 Tasmania ............................................... 133
   7.9.2 Southwest Western Australia ......................... 136
   7.9.3 Southeast Australia ................................... 139
   7.9.4 Eastern Australia ..................................... 139
   7.9.5 Northern Australia and Regional Australia ....... 144
   7.9.6 General Comments .................................... 144
7.10 Visualising Dependence .................................... 145
CONTENTS

7.11 Conclusions .................................................................. 153

8 Conclusion .................................................................. 157
  8.1 Major Findings ................................................................. 157
  8.2 Extensions .................................................................. 158
  8.3 Future Research ............................................................... 161
  8.4 Concluding Remarks ....................................................... 164

A Density Based Clustering ....................................................... 165
  A.1 DBSCAN Algorithm .......................................................... 165
  A.2 Density of the Australian Rainfall Network ......................... 166
  A.3 Implications for NRM Clusters ........................................... 167

B Additional Summary Figures for Regionalisation ......................... 169

Bibliography ................................................................ 175
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Plot showing the network density of Australian Rainfall Stations</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Spatial-temporal visualisation of daily rainfall observations</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Plot showing the results of a test that detects untagged Sunday–Monday accumulations</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>Plot showing the results of a test that detects bias in the days that extreme rainfall is observed</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Plot of the three types of distributions that are unified by the GEV distribution</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>An example of a mean residual life plot</td>
<td>34</td>
</tr>
<tr>
<td>3.3</td>
<td>An example of parameter stability plots</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>Plot of the Southern Oscillation Index</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Estimated location parameter from the GEV distribution fitted to annual maximum daily rainfall</td>
<td>53</td>
</tr>
<tr>
<td>4.3</td>
<td>Estimated shape parameter from the GEV distribution fitted to annual maximum daily rainfall</td>
<td>53</td>
</tr>
<tr>
<td>4.4</td>
<td>ENSO influence on the GEV distribution for seasonal daily maxima</td>
<td>55</td>
</tr>
<tr>
<td>4.5</td>
<td>ENSO influence on the spatial GEV model for seasonal daily maxima</td>
<td>57</td>
</tr>
<tr>
<td>4.6</td>
<td>Regions where a non-stationary ENSO influence was detected on the seasonal daily maxima</td>
<td>59</td>
</tr>
<tr>
<td>4.7</td>
<td>Model selection using the TIC for nonstationary spatial GEV models</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>Simulated examples from different max-stable processes</td>
<td>70</td>
</tr>
<tr>
<td>6.1</td>
<td>Elevation map of South East Queensland</td>
<td>88</td>
</tr>
<tr>
<td>6.2</td>
<td>Probability plot of GEV distributions</td>
<td>89</td>
</tr>
<tr>
<td>6.3</td>
<td>Station locations in South East Queensland</td>
<td>90</td>
</tr>
<tr>
<td>6.4</td>
<td>Exploratory data analysis of GEV marginal parameters</td>
<td>93</td>
</tr>
<tr>
<td>6.5</td>
<td>Diagnostic plots showing the marginal distributions</td>
<td>96</td>
</tr>
<tr>
<td>6.6</td>
<td>Diagnostic plots showing the joint distributions</td>
<td>97</td>
</tr>
<tr>
<td>6.7</td>
<td>Diagnostic plots of the extremal coefficient</td>
<td>98</td>
</tr>
<tr>
<td>6.8</td>
<td>Simulations of the annual maximum rainfall field in South East Queensland</td>
<td>99</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.9</td>
<td>Plot showing the tail distributions for the maximum, annual maximum rainfall simulated around the Brisbane CBD</td>
<td>100</td>
</tr>
<tr>
<td>6.10</td>
<td>Plot of the fraction of attributable risk for the distributions given in Figure 6.9</td>
<td>101</td>
</tr>
<tr>
<td>6.11</td>
<td>Example simulation of the annual maximum rainfall field in Toowoomba. The simulation is conditional on observed station data during the 2011 flash flood</td>
<td>103</td>
</tr>
<tr>
<td>6.12</td>
<td>Plot of the simulated total annual maximum rainfall simulated in the Toowoomba region conditional on the 2010-2011 ENSO</td>
<td>104</td>
</tr>
<tr>
<td>6.13</td>
<td>Plot of the fractional of attributable risk for the 2011 flash flood in Toowoomba comparing ENSO conditions</td>
<td>105</td>
</tr>
<tr>
<td>7.1</td>
<td>Plot of the wettest and driest observed annual maximum rainfall</td>
<td>108</td>
</tr>
<tr>
<td>7.2</td>
<td>The NRM super clusters.</td>
<td>111</td>
</tr>
<tr>
<td>7.3</td>
<td>The NRM clusters.</td>
<td>111</td>
</tr>
<tr>
<td>7.4</td>
<td>The NRM sub clusters.</td>
<td>112</td>
</tr>
<tr>
<td>7.5</td>
<td>An example showing that $K$-medoids clustering is sensitive to bounded distances.</td>
<td>119</td>
</tr>
<tr>
<td>7.6</td>
<td>An example showing that $K$-medoids clustering is sensitive to point density.</td>
<td>120</td>
</tr>
<tr>
<td>7.7</td>
<td>The example of Figure 7.5 replicated using hierarchical clustering and the average linkage method.</td>
<td>124</td>
</tr>
<tr>
<td>7.8</td>
<td>The example of Figure 7.6 replicated using hierarchical clustering and the average linkage method.</td>
<td>125</td>
</tr>
<tr>
<td>7.9</td>
<td>Examples of dendrograms generated using different linkage types.</td>
<td>126</td>
</tr>
<tr>
<td>7.10</td>
<td>Comparison of the hierarchical clustering using two different dissimilarities: Euclidean and F-madogram.</td>
<td>127</td>
</tr>
<tr>
<td>7.11</td>
<td>Plot of the F-madogram dissimilarity against the Euclidean distance</td>
<td>131</td>
</tr>
<tr>
<td>7.12</td>
<td>Hierarchical clustering of Tasmania</td>
<td>134</td>
</tr>
<tr>
<td>7.13</td>
<td>Satellite Image of Tasmania</td>
<td>135</td>
</tr>
<tr>
<td>7.14</td>
<td>Hierarchical clustering of Southwest Western Australia</td>
<td>137</td>
</tr>
<tr>
<td>7.15</td>
<td>Satellite Image of Southwest Western Australia</td>
<td>138</td>
</tr>
<tr>
<td>7.16</td>
<td>Hierarchical clustering of Southeast Australia</td>
<td>140</td>
</tr>
<tr>
<td>7.17</td>
<td>Satellite Image of Southeast Australia</td>
<td>141</td>
</tr>
<tr>
<td>7.18</td>
<td>Hierarchical clustering of Eastern Australia</td>
<td>142</td>
</tr>
<tr>
<td>7.19</td>
<td>Satellite Image of Eastern Australia</td>
<td>143</td>
</tr>
<tr>
<td>7.20</td>
<td>Dependence summary of Tasmania for a low cut</td>
<td>146</td>
</tr>
<tr>
<td>7.21</td>
<td>Dependence summary of Tasmania for a high cut</td>
<td>147</td>
</tr>
<tr>
<td>7.22</td>
<td>Dependence summary of Southwest Western Australia for a low cut</td>
<td>148</td>
</tr>
<tr>
<td>7.23</td>
<td>Dependence summary of Southwest Western Australia for a high cut</td>
<td>149</td>
</tr>
<tr>
<td>7.24</td>
<td>Dependence summary of Australia for a low cut</td>
<td>151</td>
</tr>
<tr>
<td>7.25</td>
<td>Dependence summary of Australia for a high cut</td>
<td>152</td>
</tr>
</tbody>
</table>
8.1 Exploratory data analysis of the GEV parameters for rainfall stations in South East Queensland. 163

A.1 Plot of density based clusters for the Australian Rainfall Station Network 166

B.1 Hierarchical clustering of Northern Australia 170
B.2 Satellite Image of Northern Australia 171
B.3 Hierarchical clustering of Regional Australia 172
B.4 Satellite Image of Regional Australia 173
List of Tables

1.1 The ten natural disaster events in Australia that caused the greatest losses in the insurance sector ..................................... 3

2.1 Quality flags within the GHCN-Daily dataset .......................... 9
2.2 Summary of the quality flags used for preprocessing the data ... 14
2.3 Summary of the thresholds used in GHCN-Daily Spatial Consistency Check ................................................................. 17

5.1 The first 10 Bell numbers ..................................................... 73

6.1 Temporal regression forms used for model selection. ............... 95
6.2 Observed daily rainfall totals at My Kynoch and Tamba stations on dates covering the Toowoomba flash flood in 2011 .......................... 102

7.1 Codes for the NRM regions .................................................. 113
Chapter 1

Introduction

1.1 Motivation

Ensuring that mechanisms are in place for speedy financial and social recovery reduces the impacts of extreme rainfall and flood events. These mechanisms improve the resilience of communities, lowering the risk, and reducing the exposure. Methods for reducing potential impacts include mitigation, cost sharing through the wider community, aid and insurance (Lamond and Penning-Rowsell, 2014; McAneney et al., 2016). All of these strategies benefit from accurate assessment of the risks associated with extreme rainfall and flood events.

If assessment of risk is inaccurate, particularly in terms of predictions, warnings and ineffective mitigation strategies, then communities can lose confidence in authorities (Barnes et al., 2007; Keys and Cawood, 2009). This was evident after 2011 Queensland floods, where insurers failed to foresee impacts of riverine flooding (McAneney et al., 2016) and decisions around dam operation conditions resulted in additional downstream flooding (Head, 2014). A Commission of Inquiry was formed to examine the handling of these events and, in part, address public outcry (Head, 2014).

Improved assessment of the risk associated with flood events is possible through advances in statistical research, improvements in climate modelling, and increases in computational capacity. Given this, there is much scope to translate statistical insights about extreme rainfall and flood events into real world outcomes to improve mitigation. This thesis aims to address some of the current statistical shortfall, and create a bridge between highly technical statistical ideas and application areas in climatology and engineering.
1.2 Impacts of Extreme Rainfall and Flood

The need to understand the risk of extreme rainfall and associated flooding is strongly motivated by mitigating the impacts experienced by communities. Common impacts include disruption to essential services, such as electricity, water supply and telephone. Infrastructure damage is also common, with destruction of houses, buildings, roads and bridges. In the worst case scenario, the consequence of these extreme events can be fatalities.

Flooding is a leading cause of natural disaster related fatalities, second in Australia only to heatwaves (Coates et al., 2014). The majority of the recorded flood fatalities in Australia are associated with short duration rainfall events and the primary cause of fatality is the victim attempting to cross a flooded road, bridge or causeway (Haynes et al., 2017).

In addition to fatalities, there are other health impacts due to flooding. These can include non-fatal injuries, infection, increases in the spread of fecal-oral diseases, such as diarrhoea, and exacerbation of pre-existing health conditions (Ahern et al., 2005; Du et al., 2010; Tapsell et al., 2002). The trauma and stress of these events can also manifest in mental health issues, such as anxiety, depression, adjustment disorder and post traumatic stress disorder (Ahern et al., 2005; Mason, Andrews, and Upton, 2010; Tapsell et al., 2002).

From a financial perspective, flooding is the most costly natural disaster worldwide (Miller, Muir-Wood, and Boissonnade, 2008). In Australia, eight of the ten most expensive natural disasters on record are related to events where extreme rainfall, flooding or both was a factor (McAneney et al., 2016). These events are summarised in Table 1.1. The most recent of these events is the 2011 floods in Queensland, where economic losses were estimated to be in excess of 5 billion Australian dollars and 29,000 homes and businesses were inundated (Queensland Floods Commission of Inquiry, 2012).

Other economic impacts include financial losses for businesses. Even if businesses are not directly impacted by flooding, downtime may be experienced due to loss of essential services (Webb, Tierney, and Dahlhamer, 2000). Employees may also be unable to commute to work due to flooding, and flooded roads cause disruption to standard supply and demand chains (Kleindorfer and Saad, 2005; Kousky, 2014). This often includes disruption to food supply chains and consequently food security. Land use strategies often position agricultural land in the floodplain (X. Cai et al., 2015); as a consequence, during flooding food supply chains often need to be rerouted (K. Smith et al., 2015).

---

1We note flood cost figures can vary. For a discussion of methods for estimating flood loss see Kousky, (2014). Variations in estimated cost can be caused by different definitions of what constitutes a flood and different interpretations of the financial interplay of impacts.
1.3. RAINFALL AS A PROXY FOR FLOOD

Table 1.1: This table is taken from (McAneney et al., 2016). The table shows the ten natural disaster events in Australia that caused the greatest losses in the insurance sector. These losses were normalised relative to the period 2014 – 2015. Details of the normalisation method are found in (Crompton and McAneney, 2008). Bolded entries in the table signify natural disaster events where extreme rainfall or flood was a factor.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Year</th>
<th>Event</th>
<th>Millions (AUD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1999</td>
<td>Sydney Hailstorm</td>
<td>4475</td>
</tr>
<tr>
<td>2</td>
<td>1974</td>
<td>Tropical Cyclone Tracy</td>
<td>4178</td>
</tr>
<tr>
<td>3</td>
<td>1989</td>
<td>Newcastle Earthquake</td>
<td>3384</td>
</tr>
<tr>
<td>4</td>
<td>1974</td>
<td>Brisbane Floods</td>
<td>2701</td>
</tr>
<tr>
<td>5</td>
<td>2011</td>
<td>Queensland and Victorian Floods</td>
<td>2506</td>
</tr>
<tr>
<td>6</td>
<td>1983</td>
<td>Ash Wednesday Bushfires</td>
<td>2371</td>
</tr>
<tr>
<td>7</td>
<td>1985</td>
<td>Brisbane Hailstorm</td>
<td>2046</td>
</tr>
<tr>
<td>8</td>
<td>2007</td>
<td>Pasha Bulker East Coast Low Storm</td>
<td>1966</td>
</tr>
<tr>
<td>9</td>
<td>1973</td>
<td>Tropical Cyclone Madge</td>
<td>1520</td>
</tr>
<tr>
<td>10</td>
<td>1990</td>
<td>Sydney Hailstorm</td>
<td>1433</td>
</tr>
</tbody>
</table>

Heavy rainfall also increases the soil moisture, increases erosion, and provides a mechanism for pests to spread, thereby reducing crop productivity (X. Cai et al., 2015).

While this impact summary is by no means comprehensive, it does provide insight into the scale and type of impacts that communities experience due to extreme rainfall and associated flooding. Understanding and assessing the risk of these events is therefore strongly motivated from a socio-economic standpoint.

1.3 Rainfall as a Proxy for Flood

To estimate the potential risk associated with flooding, we can use modelling approaches. These models can be informed by streamflow observations (Haddad et al., 2010; Ishak et al., 2013). However, using streamflow records for statistical modelling has its challenges. For example, in statistics a common modelling assumption is that the observations are identically distributed and this is often not the case for streamflow observations.

Changes in water management practices and river regulation can drastically alter the homogeneity of streamflow measurements (Kundzewicz et al., 2014; Milly et al., 2008; Wheater and Evans, 2009). These changes can be caused by the introduction of storage infrastructure, such as dams or spillways. Streamflow can also be impacted by changes in land use, land cover, irrigation,
INTRODUCTION

and agriculture (Bradshaw et al., 2007; Gordon, Finlayson, and Falkenmark, 2010). In addition, alterations in the upstream water usage impact the homogeneity of downstream measurements. Given these temporal changes, the series of streamflow observations that is reflective of current conditions is often much smaller than the length of the full observational record. For assessing present day flood risk, a shorter series means less data and higher uncertainty.

A further challenge to the use of streamflow observations is that changes in land-use influence the displacement of water. An identical rainfall event in different years can therefore have a vastly different flood footprint. For example, a rainfall event that caused a low impact flood, may be high impact if changes in land-use increase the displacement of water. For these reasons, it can be difficult to interpret and compare flood peak and peak streamflow of previous flood events.

For our particular interest, it is prudent to consider rainfall instead of streamflow. Rainfall is largely independent of the factors that influence inhomogeneity of streamflow measurements, and extreme rainfall is a natural precursor to flood. The risk associated with flooding and the risk from extreme rainfall are therefore highly correlated. Statistical insights about rainfall extremes can later be translated into runoff and streamflow measurements (Ball et al., 2016; Vaze et al., 2011). For these reasons, in this thesis we fit statistical models to daily rainfall extremes. We can use these statistical models to assess the risk of extreme rainfall events and, by proxy, floods.

1.4 Thesis Outline

In this research we are primarily concerned with estimating the probability that the wettest day of the year at a given station exceeds some threshold. This threshold is associated with some level of risk. Understanding this probability means that end users and stakeholders can implement appropriate mitigation strategies and make informed decisions.

The challenge is that extreme events by their nature are scarcely observed, and the sorts of events we are aiming to mitigate against are often yet to be observed. We therefore require a branch of probability theory, aptly named extreme value theory. Extreme value theory helps us to justify extrapolating outside the range of the observed data. We provide the necessary, introductory background to this theory in Chapter 3.

Before reviewing extreme value theory, in Chapter 2, we introduce the data we use for fitting our statistical models. This data is the daily rainfall observations recorded at stations throughout Australia. In this chapter, we probe
the quality of these observations. Understanding the limitations of these observations is fundamental to ensuring the reliability of our statistical inference.

Given our data and application, it is important to remember when reading this thesis that our definition of an extreme event is given in a statistical context and as related to extreme value theory. Therefore, in general, when we are referencing extreme rainfall events we mean events that are well above the 99th percentile. For example, the wettest day of the year is given by an empirical quantile of 0.997, and the wettest day in 10 years has an empirical quantile of 0.9997. In fields such as climate science, extreme can mean above the 90th or 95th percentile depending on the application. Fields such as hydrology and engineering have further variation in the semantics of what extreme means.

In this thesis, we use methods from extreme value theory to improve our understanding of Australian rainfall extremes and how these extremes are influenced by climate. Here climate refers to long term variability within a given region, in contrast to weather, which is related to the short term state of the atmosphere. In Chapter 4, we consider how the large scale climate driver of the El Niño Southern Oscillation (ENSO) influences the probability of extreme rainfall in Australia. We perform a statistical analysis based on observations at single stations, and compare the results with those obtained from simple spatial methods, where dependence between nearby stations is ignored.

However, we know from the physical process of rainfall that we expect nearby locations to experience similar impacts. Given this, spatial models for rainfall extremes with dependence can help to improve our understanding of extreme events and improve mitigation strategies. In Chapter 5, we provide the necessary introduction to more sophisticated statistical models for spatial extremes. This includes max-stable processes, which are the natural extension of univariate extreme value models to models in continuous space with dependence.

An application of max-stable processes to the modelling of rainfall extremes is given in Chapter 6. Here, we highlight how models of spatial extremes with dependence compare with univariate extreme value models. In particular, we use the max-stable process to consider the probability of an historical flash flood, by simulating the extreme precipitation field. We show that a flash flood of this size is more likely to occur during a strong La Niña phase of the large scale climate driver, El Niño Southern Oscillation.

The application given in Chapter 6 demonstrates how models of spatial extremes with dependence can be informative for understanding extreme rain-
fall events and associated flooding. Practically however, there are limitations to using max-stable models for modelling large domains, namely, that the dependence structure of a max-stable process is fixed. Chapter 7 addresses this issue by giving a regionalisation of Australia where regions are defined via similar extremal dependence.

In the final chapter of this thesis, we present an overall summary of the research and discuss its limitations. We also discuss future research directions that lead on from this work.
Chapter 2

An Observational Dataset for Extremes

2.1 Introduction

The statistical analysis of extreme precipitation events has implications for decision makers in a range of fields. Underpinning this statistical analysis is the type and quality of the data used. It is therefore important to understand the limitations of the data, to ensure the interpretation of the statistical results is meaningful.

The data available for extreme rainfall statistics are often limited. To some extent this problem is inherent as extreme events by their nature are infrequently observed. However, the issue is compounded by the lack of observational records that are of high quality and cover a long period.

There is a high quality Australian dataset of daily rainfall observations (Haylock, Nicholls, et al., 2000; Lavery, Kariko, and Nicholls, 1992). The stations in this network have been subject to quality control tests and the corresponding station metadata was examined. Given the quality assurance, the dataset is commonly used in studies of extreme rainfall, (Gallant, Hennessy, and Risbey, 2007; Suppiah and Hennessy, 1998; Westra and Sisson, 2011). However, this dataset includes only 152 stations. The stations are geographically dispersed throughout Australia, and the density of station data is low. Therefore this dataset is unsuitable for fitting a statistical model of rainfall extremes in continuous space with dependence.

The Australian Bureau of Meteorology has upwards of 17,000 observational rainfall stations. These stations have not been subject to quality assurance and therefore are subject to a range of quality issues. This has made it undesirable to use these stations for statistical analyse, as the quality of any resulting
statistical inference would be open to question. However, through bilateral agreements with the Australian Bureau of Meteorology, these stations have been included in the GHCN-Daily dataset (Global Historic Climate Network Daily) (Menne et al., 2012). As a consequence, the station observations have been subject to automated quality assurance as part of the GHCN-Daily station network (Durre, Menne, Gleason, et al., 2010; Durre, Menne, and R. S. Vose, 2008).

The process of automated quality assurance of observations is non-trivial. Some of the problems present within observational data include under reporting small rainfall, rounding observations, recording zero rainfall instead of a missing observation, recording multiple days of rainfall as a single observation, shifts in the dates of the observational record, errors in transcription and errors in digitisation (Lavery et al., 1992; Reek, Doty, and Owen, 1992; Vicente-Serrano et al., 2010; Viney and Bates, 2004).

An important point to make here is that some of these quality issues are irrelevant to the analysis of extremes. As such, it is important not to discard a station record for quality issues unless those issues are directly related to the analysis of extremes. For example, discarding a record for under reporting of small rainfall amounts is unlikely to be related to the quality of observations of annual maxima. GHCN-Daily flags observations according to the related quality issue (for flag types see Table 2.1).

The general structure of the quality assurance for precipitation in GHCN-Daily is as follows:

1. **Basic integrity checks** These include checking for repetition, duplication, and checking that the observations do not exceed world records,

2. **Outlier tests** These use a threshold to identify outliers. Both the raw data and lag one differences are tested

3. **Temporal consistency checks** One set of tests considers internal consistency with other products, such as snow datasets,

4. **Spatial consistency checks** These compare observations at neighbouring stations to determine if a target observation is spatially inconsistent. A date range of ±1 day around the target observations is considered.

For full details refer to (Durre, Menne, Gleason, et al., 2010). The thresholds for each of these tests were set to ensure that the false-positive rate, where true observations are falsely identified as spurious, was less than 20%. These thresholds were set using a semi-automated approach (Durre, Menne, and R. S. Vose, 2008), where some manual inspection of observations was required.
2.2 Observational Data

Table 2.1: Quality flags within the GHCN-Daily dataset (National Centers For Environmental Information with National Oceanic and Atmospheric Administration, 2018).

<table>
<thead>
<tr>
<th>Quality Flag</th>
<th>Quality Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blank</td>
<td>did not fail any quality assurance check</td>
</tr>
<tr>
<td>D</td>
<td>failed duplicate check</td>
</tr>
<tr>
<td>G</td>
<td>failed gap check</td>
</tr>
<tr>
<td>I</td>
<td>failed internal consistency check</td>
</tr>
<tr>
<td>K</td>
<td>failed streak/frequent value check</td>
</tr>
<tr>
<td>L</td>
<td>failed check on length of multiday period</td>
</tr>
<tr>
<td>M</td>
<td>failed megaconsistency check</td>
</tr>
<tr>
<td>N</td>
<td>failed naught check</td>
</tr>
<tr>
<td>O</td>
<td>failed climatological outlier check</td>
</tr>
<tr>
<td>R</td>
<td>failed lagged range check</td>
</tr>
<tr>
<td>S</td>
<td>failed spatial consistency check</td>
</tr>
<tr>
<td>T</td>
<td>failed temporal consistency check</td>
</tr>
<tr>
<td>X</td>
<td>failed bounds check</td>
</tr>
<tr>
<td>Z</td>
<td>flagged as a result of an official Datzilla investigation, where external users report the errors</td>
</tr>
</tbody>
</table>

In this chapter we explore some of the data quality issues that affect the analysis of extreme rainfall and discuss the methods used for preprocessing the extreme observations. We also examine the effectiveness of the quality assurance tests used to flag spurious observations within GHCN-Daily. Finally, we discuss why gridded observational data products are not a suitable alternative for our desired statistical analysis.

2.2 Observational Data

2.2.1 Data Summary and Visualisation

The station IDs in GHCN-Daily are equivalent to those from the Australian Bureau of Meteorology. For GHCN-Daily IDs the first two letters refer to the country code, where ‘AS’ is for Australia. The third letter is a network code to identify the numbering system and the final 8 digits constitute the station number padded with zeroes (National Centers For Environmental Information with National Oceanic and Atmospheric Administration, 2018). The GHCN-Daily station id, ‘ASN00040214’, therefore corresponds to the Australian Bureau of Meteorology station ‘040214’.

To access the stations for statistical analysis the R package, rnoaa (Chamberlain, 2017) provides an interface to extract the data from the National
Oceanic and Atmospheric Administration (NOAA). Within this package there exists a range of functions for basic data handling that enhance the usability of the data.

Prior to any statistical analysis on this station data an important first step is visualisation. Figure 2.1 shows that the spatial density of stations is highest on the east coast and around the capital cities. Many places within Central Australia do not have any station data spanning these periods. Station temporal density is highest during the visualised period from 1937 – 1990, compared with 1910 – 1936 and 1991 – 2017. This suggests fewer stations were in com-
2.2. OBSERVATIONAL DATA

mission during the early period and stations have since been decommissioned during the later period.

Visualisation also provides insight into the type of process that is being modelled. Consider the following example of multiple days of rainfall in Brisbane, Australia from 22-01-1974 to 30-01-1974, shown in Figure 2.2. This example was selected as it highlights a range of quality issues that concern the statistical analysis of extreme rainfall.

Figure 2.2: Daily rainfall observations in Brisbane, Australia. Solid circles are coloured by the measured precipitation. Empty circles show zero observations. A grey cross indicates missing data. Tagged accumulations, rainfall totals accumulated over multiple days, are shown with an empty circle and a cross. The final day in the accumulation period is coloured by the multiday total and the other days are coloured grey.
2.2.2 Quality Issues

On 26-01-1974 there was an extreme rainfall event in Brisbane, Australia, see Figure 2.2. The majority of stations in this region recorded their annual maximum on this date. However, observations recorded at some stations are spatially inconsistent.

One of the stations has missing data on 26-01-1974. This station is shown with a cross. Observations were recorded on the days prior and missing observations are recorded in the days directly after. This suggests that the reason the observation is missing may be related to the extreme event, such as the station washing away. This missing observation is therefore likely to be a missing maximum.

Also on 26-01-1974 two observations are recorded as 0 mm. One of these was flagged as spatially inconsistent by the GHCN-Daily quality control. The other was not. These zeros are clearly spurious given the surrounding rainfall. The true observations are likely to have been annual maxima for these stations.

There are also three tagged accumulations that cover 26-01-1974. The annual maximum for each of these stations is likely to have been recorded within the accumulation. Automatic identification of individual daily totals from an accumulation is challenging and we do not attempt it here.

Within this example there are also suspected untagged accumulations. Untagged accumulations occur when multiple days of rainfall are recorded as a single day. These types of observations can present as spurious maxima. An example of an untagged accumulation is the annual maximum observation of 240.4 mm recorded on 28-01-1974. On the four days prior to 28-01-1974, 0 mm of rainfall was recorded at the station. However, this period of zero recorded rainfall includes an extreme rainfall event on the 26-01-1974. Additionally, on 28-01-1974 the five closest stations recorded approximately 50 mm of rainfall compared with 240.4 mm. Given the observations at the surrounding stations from 24-01-1974 to 28-01-1974, this observation is clearly spurious and should have been flagged for quality.

Of the 27 stations in this example on 26-01-1974; we suspect that 3 maxima occurred within tagged accumulations, 1 station recorded missing data instead of a maximum and 2 observations were falsely recorded as zero instead of tagged accumulations. If the extremal observations are not treated appropriately during preprocessing, these observations will result in errors in the annual maximum observations. Whilst in general statistical methods will be robust to some errors within the data, care should be taken to identify as many systematic and prevalent sources of error as possible.
2.3 Preprocessing the Data

Given the quality issues highlighted, it is desirable to preprocess the raw observational records. This preprocessing is targeted at enhancing the quality of the extreme observations.

2.3.1 Reconstruction

Reconstruction is used to reduce the amount of missing data. In our context, this decreases the probability that any of the missing observations is an unobserved maximum (Haylock et al., 2000). If any of the extremes identified is reconstructed, then this suggests that the original missing observation was an unobserved maximum. By treating tagged accumulations as missing, reconstruction can also be used to identify whether an extreme observation occurred within a tagged accumulation.

For reconstruction, both absolute and relative methods can be used. Absolute methods are when only the station data is used, whereas relative methods use information from surrounding stations for comparison. Given more information is available to understand the process in relative methods, these methods are favoured where appropriate.

One of the simplest relative methods is nearest neighbour reconstruction. In this method an observation from a neighbouring station with no adjustment is used to infill the missing observation. We consider a station to be a neighbour if the two station records are geographically close and highly correlated. As rainfall relationships are non-linear, to determine whether two stations are highly correlated we use Spearman’s rank correlation coefficient. If the correlation between stations is greater than a threshold of 0.6, the stations are considered to be suitable neighbours for reconstruction. The correlation coefficient is only estimated if there is sufficient data. If more than one station is considered a suitable neighbour for infilling, the station with the highest correlation is used. Note that the closest neighbour in terms of distance is not necessarily the best neighbour for infilling. If a reconstructed observation is extracted as extreme then the station ID is included as a quality flag, see Table 2.2.

More complex relative reconstruction methods are available, but were not necessary given the desired outcomes. These methods include inverse distance weighted approaches and linear regression. However, both of these methods can reduce the frequency of extreme values (Vicente-Serrano et al., 2010), which is undesirable for our purposes.
Table 2.2: This table gives the quality flags associated with preprocessing the data. These flags are additional to those in Table 2.1.

<table>
<thead>
<tr>
<th>Quality Flag</th>
<th>Quality Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>STATION_ID</td>
<td>This observation was reconstructed from the station whose id corresponds to the quality flag.</td>
</tr>
<tr>
<td>AS</td>
<td>This observation is suspected of being a Sunday-Monday untagged accumulation, see Section 2.3.3.</td>
</tr>
<tr>
<td>AE</td>
<td>Untagged accumulations are suspected amongst the extremes at this station, including this observation, see Section 2.3.3.</td>
</tr>
<tr>
<td>A</td>
<td>This observation is suspected of being an untagged accumulation under both the Sunday-Monday test and extremes test that are outlined in Section 2.3.3.</td>
</tr>
</tbody>
</table>

Alternatively, absolute methods include imputing the missing values using only the station observations (Little and Rubin, 2014). Common data imputations include filling by sampling from the data or fitting a distribution and then sampling from that distribution. Imputing via sampling is unlikely to provide additional information about extremes. Also given the final goal is to fit a distribution to the extremes, imputing the extremes by fitting a distribution runs the danger of circular logic. Difficulty also arises if the imputation needs to account for extremal dependence between stations and the influence of climate. Therefore for simplicity, instead of imputing the remaining missing data after nearest neighbour reconstruction, we placed restrictions on the extent of missing data allowed.

2.3.2 Extreme Observations

In general, it is thought that records of longer length with few missing observations are indicative of good quality. This is based on the assumption that the observational practice appears to be consistent through time and is therefore more reliable (Lavery et al., 1992). For extremes, it is recommended that a record not be used if there are too many missing observations, as the true maxima may be unobserved. If the length of the observational record is short, it is also recommended not to use this data, as the tail of the distribution will be poorly estimated. We therefore only considered stations with a minimum of 20 years of rainfall observations. We also restricted our observations to consider only those onward from 1910. This is because prior to 1910, observer practices were not standardised, as the Australian Bureau of Meteorology was yet to be established.

After reconstruction, we extracted the extreme observations from the record. Extremes for a given period were only extracted if the percentage of remaining
2.3. PREPROCESSING THE DATA

missing data in that period was less than 5%. This percentage corresponds to an acceptable probability that an extreme event was unobserved (Haylock et al., 2000). If an extreme observation did not meet the missing data criterion, but another extreme observation occurred on the same day at a neighbouring station that was highly correlated, then the extreme observation was retained regardless of the missing data. After all preprocessing, if fewer than 20 years of the maxima were available, the station record was discarded.

Caution was also taken in dealing with dry stations. Dry stations may fail to meet the assumptions necessary for extreme value modelling; see (eg. S. Coles, 2001) for assumptions. For example, if the seasonal maximum is commonly 0 mm then this violates the assumption that the maximum is a continuous random variable. Also, if the number of nonzero observations per seasonal block is small, then the distribution of the maximum may not be well approximated by a generalised extreme value distribution, where the block size is assumed to be large. In Min, Cai, and Whetton, (2013), a station record was classified as dry if the location parameter of the generalised extreme value distribution was below 10 mm. This study was considering seasonal maxima using gridded data. We found this threshold to be too high for observational data, causing too many stations to be discarded. We suspect this is due to gridded datasets overestimating small rainfall amounts (King, Alexander, and Donat, 2013b). Given the driest areas are in Central and Northern Australia, and these are not our primary areas of interest, we have not excluded stations for being dry. However, we have exercised caution in interpreting results for these regions.

2.3.3 Quality Control

We do not perform our own quality control, but make use of the flags provided within the GHCN-Daily Network, see Table 2.1. However, despite the quality assurance of the GHCN-Daily Network (Durre, Menne, Gleason, et al., 2010), erroneous observations are still present. This was highlighted in the example given around Figure 2.2 in Section 2.2.2. Further testing is therefore required to identify errors amongst extreme observations. In particular, given the presence of untagged accumulations in our example, we question the effectiveness of the spatial consistency test used in GHCN-Daily.

Spatial Consistency Check

The GHCN-Daily spatial consistency check from Durre, Menne, Gleason, et al., (2010) can be performed as follows. Consider an observation that occurs on date \( d \), at a station indexed by zero, \( S_0 \). Let \( S \) be the set of geographically closest stations to station \( S_0 \), where \(|S|\) is restricted at least three and at most seven stations. Also define the three-day window of dates centred on \( d \) as
OBSERVATIONAL DATA

$D = \{d - 1, d, d + 1\}$. For simplicity, whenever we refer to windows centred on given day we are referencing across months and years. For example, if $d = 01-01-2018$, then $D = \{31-12-2017, 01-01-2018, 02-01-2018\}$. Let $N$ be the set of neighbouring observations to $S^d_i$ in space and time

$$N = \{S^d_j | i \in S, j \in D\}. \quad (2.1)$$

To determine whether the observation $S^d_0$ is spatially consistent, it is compared to the neighbouring observations in $N$ using the parameter $\Delta_{obs}$. Conditional on $S^d_0$ lying within the range of the neighbouring observations, $\Delta_{obs}$ is defined as

$$\Delta_{obs} = \begin{cases} \min_{i \in S, j \in D} |S^d_0 - S^d_i|, & \text{for } S^d_0 \notin [\min N, \max N] \\ 0, & \text{for } S^d_0 \in [\min N, \max N] \end{cases} \quad (2.2)$$

Under the test, if $\Delta_{obs}$ exceeds a threshold, then $S^d_0$ is considered spatially inconsistent. To determine the threshold, a baseline is used relative to normal climatology. This baseline is given by $\Delta_{rank}$. The parameter $\Delta_{rank}$ is equivalent to equation (2.2), but expressed via percentiles.

To estimate $\Delta_{rank}$, first define the empirical cumulative distribution function, conditional on nonzero observations, to be

$$F_{S_i,d}(x) = \frac{1}{m} \sum d^{*+14} \sum_{y \in Y} \sum_{k=d^{*-14}} \mathbb{I}(S^{k,y}_i \leq x) \mathbb{I}(S^{k,y}_i > 0). \quad (2.3)$$

Here $d^{*}$ is the day and month to corresponding $d$, $y$ indexes the year, where $Y$ consists of all observational years at station $S_i$, and $k$ indexes observations in a 29-day window centred on the day and month given by $d^{*}$ and year $y$. The normalising constant, $m$, is the total number of non-zero observations in this window across all observational years at station $S_i$. Missing observations are omitted when estimating the empirical quantiles. In Durre, Menne, Gleason, et al., (2010), a restriction was placed such that $m \geq 20$. We consider this number to be too small to produce reliable estimates of the empirical quantiles.

Using equation (2.3), $\Delta_{rank}$ is given by

$$\Delta_{rank} = 100 \min_{i \in S, j \in D} |F_{S_{0},d}(S^d_0) - F_{S_i,d}(S^d_i)|. \quad (2.4)$$

It follows that the threshold for the test, as derived in Durre, Menne, Gleason, et al., (2010), is

$$T = -45.72 \ln \Delta_{rank} + 269.24. \quad (2.5)$$
An observation, $S_0^d$, is flagged as spatially inconsistent if $\Delta_{\text{obs}} \geq T$.

To determine the threshold, Durre, Menne, Gleason, et al., (2010) considered three categories of $\Delta_{\text{rank}}$ values, 0%-5%, 40%-60% and 90%-99%. The threshold in each category was set using the semi-automated testing methods (Durre, Menne, and R. S. Vose, 2008). The threshold function of equation (2.5) was obtained by fitting a logarithmic function to the pairs of $\Delta_{\text{rank}}$ and $T$ determined in semi-automated tests. Table 2.3 gives the corresponding threshold for each of the percentage deciles. When there were insufficient observations to obtain a suitable estimate for $\Delta_{\text{rank}}$, the threshold was set to 269.24 mm, the maximum of equation (2.5). For these cases, only the most egregious spatial inconsistencies were detected.

**Table 2.3: Summary of the thresholds used in GHCN-Daily Spatial Consistency Check.** The thresholds correspond to the percentage deciles calculated using equation (2.5).

<table>
<thead>
<tr>
<th>$\Delta_{\text{rank}}$</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>163.97</td>
</tr>
<tr>
<td>20</td>
<td>132.28</td>
</tr>
<tr>
<td>30</td>
<td>113.74</td>
</tr>
<tr>
<td>40</td>
<td>100.58</td>
</tr>
<tr>
<td>50</td>
<td>90.38</td>
</tr>
<tr>
<td>60</td>
<td>82.05</td>
</tr>
<tr>
<td>70</td>
<td>75.00</td>
</tr>
<tr>
<td>80</td>
<td>68.89</td>
</tr>
<tr>
<td>90</td>
<td>63.51</td>
</tr>
<tr>
<td>100</td>
<td>58.69</td>
</tr>
</tbody>
</table>

**Spatial Consistency Check Limitations**

There are limitations to the spatial inconsistency check outlined in Section 2.3.3. Under this test, untagged accumulations presenting as spurious extremes are difficult to detect. Consider again the example of Figure 2.2.

Only the spurious zero on 26-01-1974 at longitude 153.05 and latitude -27.51 was flagged as spatially inconsistent. For this observation $\Delta_{\text{obs}} = 64$, $\Delta_{\text{rank}} = 95.53$ and the corresponding threshold was $T = 60.78$. This observation fails the test by a margin of less than 4 mm, where for an effective test it should fail with a much larger margin. If the test was only applied to observations on 26-01-1974 instead of the three-day window, then $\Delta_{\text{obs}} = 223$, $\Delta_{\text{rank}} = 99.87$ and the corresponding threshold is $T = 58.75$. This is a much clearer signal that the observation is spatially inconsistent. Given this, we
question the effectiveness of using a three-day window.

It is important also to note that if multiple spurious zeros are present, then the test does not function as intended. If the candidate observation is a spurious zero and any of the neighbours is also zero, then the test passes the candidate observation, as $\Delta_{\text{obs}} = 0$ and $T > 0$. Many of the spurious zeros present in this example pass for this reason. Using the median neighbouring observation to estimate $\Delta_{\text{obs}}$ and $\Delta_{\text{rank}}$ could provide for a more robust test.

These examples of the three-day window and the treatment of zeros, highlight the limitations of this spatial inconsistency check for quality assuring extreme values. It is beyond the scope of this thesis to adapt the test of Durre, Menne, Gleason, et al., (2010) to address the issues outlined. The time needed to inspect flagged observations manually and adapt the thresholds for a desired false positive rate is prohibitive. Instead, we choose to use statistical methods to help identify and eliminate stations where untagged accumulations are prevalent (King, Alexander, and Donat, 2013a; Viney et al., 2004).

Untagged Accumulations

In the following sections, we outline the statistical tests we use to detect untagged accumulations. The tests originate from Viney et al., (2004) and King et al., (2013b), and are complementary. One test examines whether a year of data has untagged accumulations and the other examines whether there are untagged accumulations present amongst the extremes. If either of these tests indicates the presence of untagged accumulations, any extreme observation meeting criterion of the test is flagged, see Table 2.2. If more than 10% of extreme observations had suspected untagged accumulations then the station is discarded.

Sunday–Monday Untagged Accumulations

The statistical test for untagged accumulations in Viney et al., (2004) explores the prevalence of weekend rainfall totals being recorded as a single total on Monday. Let $n$ be the number of observed Sundays in a year and let $p_1$ to be the probability of rain on any given day. Under the assumption that rainfall is equally likely on any day of the year, a binomial distribution can be used to model the number of rainy Sundays, $N_r$. Therefore $N_r \sim \text{Bin}(n, p_1)$ where

$$
\mathbb{P}(N_r = k) = \frac{n!}{k!(n-k)!} p_1^k (1-p_1)^{n-k}, \quad k \in \{1, \ldots, n\} \quad (2.6)
$$

For a given year, let $n_r$ be the number of observed rainy Sundays. If $\mathbb{P}(N_r \leq n_r)$ is below a given $p$-value, then the year is suspected of having systematic Sunday–Monday untagged accumulations. In (Viney et al., 2004),
Monte Carlo simulations were used to determine the level at which just 1 year of 100 years would be falsely discarded for untagged accumulations by no more than 5% of the 152 stations. This condition resulted in a p-value of 0.0008 that was found to be ‘approximately invariant’ with n. We use this p-value in our station testing.

To determine $\Pr(N_r \leq n_r)$, first it is necessary estimate $p_1$ and then $n$. The probability, $p_1$, can estimated empirically by dividing the number of rainy days between Tuesday and Friday inclusive, by the total number of observational days between Tuesday and Friday inclusive. There are underlying assumptions here that observers are more reliable at collecting weekday observations than weekend, and the influence of untagged accumulations during the week is negligible. For simplicity, tagged accumulations were ignored in the estimation of $p_1$.

To obtain the number of observed rainy Sundays, $n_r$, an adjustment is needed to account for Sundays occurring during tagged accumulations. This requires estimating the probability of rainfall on a Sunday, $p_s$, during a tagged accumulation period of $a$ days. Assume that rainfall follows a first order Markov process with two states. Let $p_{01}$ be the probability of a rainy day occurring immediately after a dry day. The probability of rainfall during the $a$ day period is

$$p_a = 1 - (1 - p_1)(1 - p_{01})^{a-1}. \quad (2.7)$$

Estimation of $p_{01}$ is done similarly to $p_1$ using weekday observations, excluding Mondays. The probability of rain on a Sunday, given the Sunday occurs in an accumulation of $a$ days of rainfall, is

$$p_s = \frac{\Pr(\text{rain on Sunday} | \text{rain during a period of } a \text{ days})}{p_a} \quad (2.8)$$

$$= \frac{p_1}{p_a}. \quad (2.9)$$

The expected number of rainy Sundays in the tagged accumulation is then $p_s$ times the number of Sundays in the tagged accumulation. For all tagged accumulations, this adjustment is added to the number of observed rainy Sundays and then rounded to give $n_r$.

An example of the test is given in Figure 2.3 for station Broomehill, Australia. This example is reproduced from the original paper (Viney et al., 2004)\(^1\). For years where the p-value is below the dashed horizontal line at

\(^1\)We note there are subtle differences in our figure to the original paper. Although both graphs identify the presence of gross untagged accumulations and the shape is similar, the year recording the lowest p-values is different. This could be a result of differences in the data.
Figure 2.3: The plot shows the $p$-values from the test for untagged Sunday–Monday accumulations. The test was run at the station BROOMEHILL, ASN00010525, for the years 1910 – 2000.

Methods such as this are only appropriate for stations that have suitable yearly rainfall. Stations with little rainfall or stations with large amounts of missing data should be excluded. Additionally, caution is also needed for stations where the probability of rainfall may be different on a weekday to a weekend. This can occur for stations that are impacted by an urban heat island effect (Earl, Simmonds, and Tapper, 2016).

We acknowledge that some of the test assumptions as related to rainfall are oversimplified. However, given the aim is detect whether untagged accumulations are prevalent throughout a station record, the test serves a useful purpose.

**Extremal Untagged Accumulations**

In addition to Sunday–Monday untagged accumulations, we also test for untagged accumulations that present as spurious extremes. We use the hypothesis test from (King et al., 2013b), to determine if there was systematic bias
in the days that extreme rainfall was observed. We anticipate that extreme events should occur with equal frequency on each day of the week.

To represent the extremes at each station, the four highest rainfall totals for each year were extracted. The distribution of extreme rainfall days was estimated using a basic bootstrap. In total, 1000 bootstrap resamples of the same size as the original data were taken. From these bootstrap resamples the expected number of extremes on a given day was estimated.

If there is no bias between the days when extremes are observed, then the probability of observing an extreme on a given day is $1/7$. Therefore if $X_i$ is a random variable representing the number of extremes, $n_e$, on day $i$, then $X_i$ follows a binomial distribution, $X_i \sim \text{Bin}(n_e, 1/7)$. A two-sided critical region was determined using a 5% confidence level. Untagged accumulations were suspected if on a given day the expected number of extremes from the bootstrap fell below the critical region, and on the following day the expected number fell above the critical region. An example from the original paper, (King et al., 2013b), is given in Figure 2.4. This station failed the test in the original paper, but we have reduced the number of suspect extreme through preprocessing and now it passes. In some respects this is pleasing, but it also hints at the effectiveness of how the test is defined.

**Outlier tests**

Within the analysis we ignore observations flagged under outlier tests in GHCN-Daily. Although it is not uncommon to set a threshold to identify spurious outliers (Durre, Menne, and R. S. Vose, 2008; Vicente-Serrano et al., 2010), threshold setting is prone to systematically eliminating true extreme values. Visual inspection of several outliers flagged in Australian observational records found that these observations were extremes. This was shown by the spatial homogeneity with the surrounding observations and reports in media of extreme events on these days.

**2.3.4 Homogeneity**

In addition to the aforementioned issues of quality control, non-climatic causes of inhomogeneity can result in errors within the precipitation record. Possible causes of these inhomogeneities include changes in exposure of the station, changes in station location and changes to instrumentation (Lavery et al., 1992). As part of the existing quality assurance within GHCN-Daily (Durre, Menne, Gleason, et al., 2010), large jumps were identified using the standard normal homogeneity test (SNHT) (Alexandersson, 1986) on annual totals. Stations with large jumps were eliminated from the dataset. Ideally, homogeneity tests should include tests for changes in both frequency and intensity of
Figure 2.4: The plot shows the expected frequency of rainfall days estimated from 1000 bootstrap resamples. The upper and lower dashed lines give the 95% confidence intervals for a binomial distribution with probability of 1/7. The middle dashed line shows the mean. The test was run at the station ARDROSSAN, ASN00022000 for the years 1910 – 2010. The observed Sunday extremes is below the defined confidence interval. However, as the Monday extremes are within the confidence band the test passes.

Identification of inhomogeneity is a difficult problem given rainfall is a highly non-linear process.

Testing for inhomogeneity of the general record and testing for inhomogeneity amongst extremes are two separate issues. For example, under reporting of small values of rainfall as zero can cause a station to fail the homogeneity test, but the station record is still suitable for the analysis of extremes. There is scope here to make testing for inhomogeneities specific to extremes. This could be achieved by conditioning on observations occurring above a given threshold and applying standard methods.

2.4 Limitations of Gridded Data

Given the quality issues associated with observational records (Lavery et al., 1992; Menne et al., 2012), a natural question is why not use existing gridded data products (M. G. Donat et al., 2013; M. Donat et al., 2013; Jones, Wang, and Fawcett, 2009). However, for statistical modelling of extreme rainfall
there are advantages in using observational station records.

Gridded observational products can be created by interpolating the observations to create a surface, and then converting that surface into a grid. Alternatively, summary statistics are obtained, and then the interpolation is performed. However, the statistical methodology used to interpolate the observational data is subject to model assumptions and uncertainty. For example, a common assumption utilised is normality. This may prove a reasonable assumption for the mean of a distribution, but not for extremes. A comparison of interpolation methods can be found in (Contractor et al., 2015).

Statistical analyse of rainfall extremes performed using these gridded data products is therefore subject two sources of uncertainty, one from the underlying model used to create the dataset and the other from the extremal model. It is therefore important to understand any structural or methodological uncertainties within the gridded data product (Dunn, Donat, and Alexander, 2014). Some interpolation methods may perform well for the bulk of the data but fail to represent extremes accurately (Contractor et al., 2015). Also when the underlying station network is sparse, there will be large uncertainties in the interpolation which can result in spurious trends when the network changes (King et al., 2013b).

A secondary consequence of gridding the data is that extreme rainfall events are generally of a lesser magnitude than for point-located observations (King et al., 2013b). Differences in magnitude will vary with gridded resolution, and the problem is most severe for low resolution grids. For a pointwise analysis of rainfall extremes this poses a direct concern. Data loss also occurs in gridded data products, as multiple observational records are condensed into a single time series. This can improve the signal to noise ratio, but it also can impact the power of a statistical model to detect trends due to information loss (Westra and Sisson, 2011).

A greater issue for our purposes is that gridded data products are not ideal for fitting of statistical models of rainfall extremes in continuous space with dependence. An overview of these types of statistical models can be found in Chapter 5. Statistical models in continuous space with dependence are vital for accurately modelling extreme rainfall, as nearby locations within a region are likely to experience similar impacts. A major concern is that the dependence structure imposed by the underlying statistical model used to create the grid, may not represent the dependence at extreme levels. Additionally, observations at neighbouring stations separated by small distances are needed to accurately resolve extremal dependence, but this information is lost in gridded products. Only a set of discrete distances, restricted by the gridded resolution, are available to estimate the dependence structure.
our application, we therefore give preference to using raw observations.

2.5 Discussion

In this chapter, we identified quality issues present in daily rainfall observations. Some of these issues can impact the statistical analysis of daily extremes including the presence of untagged accumulations, spurious zeros and missing maxima. We discussed methods for detecting these occurrences and addressing related quality issues.

While this chapter provides an important first step towards ensuring the quality of our extreme observations, there is scope for improvement. Of particular concern is the effectiveness of the spatial consistency check (Durre, Menne, Gleason, et al., 2010). In Section 2.3.3, we raised concerns about the use of three-day windows and the representation of zeros within the test. Additionally, we suspect that the threshold may not be set optimally for the desired false positive rate. Intuitively, we would expect different climatic regions to require different thresholds. For an effective test, the acceptable pairwise differences between neighbours should therefore vary based on local climate. The threshold should ideally be set using simulation, then local climatic variations can be appropriately represented within the test.

Also, this chapter is by no means a comprehensive account of all quality issues present. There are other quality issues that impact extreme observations that we did not test for. This was because these issues did not necessarily impact our specific analysis. For example, there are errors in the station metadata. This includes incorrect elevation measurements. Shifts of lag ±1 day are also not uncommon. This is particularly true prior to 1910 and the establishment of the Australian Bureau of Meteorology.

We have also not considered how best to identify when the mechanism behind the missing maxima was systematic. For example, when extreme rainfall events happen, flash flooding can occur and stations may get washed away. It also may not be possible for an observer to get to the station to read the observation. An example of this is shown in Figure 2.2. In these cases, the mechanism behind the missing extreme observation is directly related to the extreme rainfall event. This mechanism is referred to as Missing Not at Random (Little et al., 2014). An adjustment to any statistical analysis is needed for such cases.

Although the methods discussed in this chapter are applicable to all extreme rainfall observations, the approaches to some extent have been tailored to maxima taken over a block of observations, such as annual maxima or sea-
2.5. DISCUSSION

sonal maxima. For the analysis of the highest five-day accumulated totals, Rx5d, or for extreme observations occurring above a high threshold, further work is needed to verify these extreme observational types.

The code needed for preprocessing can be found in the github repository RaingleExtremes https://github.com/katerobsau/RaingleExtremes. This repository also includes relevant documentation for the code and resulting datasets. The development of these datasets have been ongoing throughout the course of this thesis. As such the data used in earlier analyses is slightly different to that hosted online. For example, in 6, we used the raw station data directly from the Bureau of Meteorology, not from GHCN-Daily. In this instance, the data was subject to some preprocessing (Saunders et al., 2017), although stations were not subject to the quality tests of (Durre, Menne, Gleason, et al., 2010). Despite the subtle differences in the data used throughout the thesis, we are confident that the methods and conclusions presented are robust to any differences.
Chapter 3

An Introduction to Extreme Value Theory

3.1 Introduction

Extreme value theory (EVT) is a field within probability and statistics that is used to model extreme events. Applications of EVT exist in the fields of hydrology (Katz, Parlange, and Naveau, 2002), climate change (L. Cheng et al., 2014; Kharin et al., 2013), finance and insurance (Embrechts, Klüppelberg, and Mikosch, 2013; McNeil and Frey, 2000), and public health (Thomas et al., 2016). Modelling extreme events is often motivated by the need to mitigate potential consequences, such as from extreme weather, financial risk or epidemic outbreak.

The statistical challenge in modelling extreme events is that often only a small number of observations are available for model fitting. Also, to mitigate the potential consequences of extreme events, we need to be able to assign a probability to events we are yet to observe. EVT provides the necessary statistical justification to robustly estimate probabilities for events that occur in the far tail of the distribution.

At the core of EVT is the generalised extreme value (GEV) distribution. This distribution is useful for modelling the sample maximum over a block of observations, such as the wettest day of the year. Mathematically, the GEV arises as the non-degenerate limiting distribution of the normalised maximum of a series of independent and identically distributed (iid) random variables, as the sample size tends to infinity. In practice however, the limiting distribution of the GEV is often used to approximate the distribution of finite sample maxima for large sample sizes, such as the wettest day of the year.

By using the known limiting distribution, under appropriate assumptions,
there is a strong statistical justification for extrapolation outside the range of the observed data. In hydrological applications, extrapolation is needed for deciding the design levels for infrastructure, such as drainage, spillways, levee banks and dams (Ball et al., 2016).

The mathematical formulation of GEV begins with an assumption that the maxima are taken over iid observations. For practical applications, such as for environmental processes like rainfall, this assumption is often untrue. For example, rainfall occurring on consecutive days is correlated, so the independence assumption is violated. Additionally, daily rainfall is not identically distributed, as it is subject to seasonal influences. For processes that do not meet the iid assumption, it is still often possible to use EVT provided some further conditions are met or by adjusting the context in which the theory is applied.

In this chapter, we introduce the limiting distribution for block maxima, the GEV, and the limiting distribution for extremal observations occurring above a high threshold, the Generalised Pareto (GP). We provide the foundations needed to fit these extreme value distributions and apply them in practice. This includes outlining when it is possible to extend and adapt the theory to model extremes of processes that are not iid. The theorems and definitions presented in this chapter are adapted from the introductory textbook S. Coles, (2001) unless cited otherwise. For a more rigorous mathematical treatment the reader is referred to Leadbetter, (1983) and De Haan and Ferreira, (2006).

3.2 Generalised Extreme Value Distribution

3.2.1 Limiting Distribution for Block Maxima

The standard formulation of EVT starts with a sequence of iid random variables $X_i$ for which the maximum observation occurring over a fixed block size, $n$, is defined to be

$$M_n = \max\{X_1, \ldots, X_n\}.$$  \hspace{1cm} (3.1)

Let $F$ be the distribution function of $X$ such that $F = \mathbb{P}(X \leq z)$, where $z^F$ is the right endpoint of the support of $F$, such that $z^F = \sup\{z : F(z) < 1\}$. It follows that the distribution function of $M_n$ is

$$\mathbb{P}(M_n \leq z) = \mathbb{P}(X_1 \leq z, \ldots, X_n \leq z) = \mathbb{P}(X \leq z)^n = F(z)^n.$$

In practice, the distribution $F$ is often unknown. Although it is possible to estimate $F$ from observed data, in the tail of the distribution where there are fewer observations, errors can compound for the estimation of $F^n$. This
3.2. GENERALISED EXTREME VALUE DISTRIBUTION

motivates approximating the distribution of $F^n$ by the limiting distribution of $M_n$ for large $n$. As $n \to \infty$ however, $F^n \to 0$ for any $z < z^F$. A degenerate limiting distribution can be avoided, if there exists an appropriate sequence of constants $a_n > 0$ and $b_n \in \mathbb{R}$, such that the sample maxima, $M_n$, can be linearly renormalised (Fisher and Tippett, 1928; Gnedenko, 1943).

**Theorem 3.2.1.** (see S. Coles, (2001) Theorem 3.1.1) If there exist sequences of constants $\{a_n\} > 0$ and $\{b_n\}$ such that

$$
P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \to G(z) \quad \text{as} \quad n \to \infty,
$$

where $G$ is a non-degenerate distribution function, then $G$ is a member of the generalised extreme value (GEV) family

$$
G(z) = \exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]_{+}^{-1/\xi} \right\},
$$

where $[v]_{+} = \max \{0, v\}$, $\mu \in \mathbb{R}$, $\xi \in \mathbb{R}$ and $\sigma \in \mathbb{R}^{+}$.

Formal treatment and proof of the generalised extreme value theorem can be found in De Haan and Ferreira, (2006).

The intuition behind the use of this theorem for sample maxima is analogous to using the central limit theorem (CLT) to approximate the sample mean. For example, the distribution of the finite sample mean, after standardisation, can be approximated by the standard normal limiting distribution using the CLT.

Theorem 3.2.1 seems to imply that we need to know the normalising constants to fit a distribution to $M_n$. In practice, this is not the case. Assuming for large $n$ the limiting distribution is a suitable approximation for the finite sample maxima, then

$$
P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \approx G(z).
$$

Equivalently,

$$
P \left\{ M_n \leq a_n z + b_n \right\} \approx G \left( z \right).
$$

Letting $z^* = a_n z + b_n$,

$$
P \left\{ M_n \leq z^* \right\} \approx G \left( \frac{z^* - b_n}{a_n} \right).
$$

It can easily be shown that the right hand side of (3.7) is a member of the GEV family of distributions. Therefore to fit a distribution to $M_n$ we are not required to find suitable normalising constants, as the distribution of $M_n$ can be directly approximated by a GEV distribution.
3.2.2 Three Types

Three types of distributions are unified within the generalised class of extreme value distributions: Gumbel, Fréchet and Weibull (Fisher et al., 1928; Gnedenko, 1943). The type is determined by the sign of the shape parameter, $\xi$, as the shape parameter controls the tail behaviour of the distribution. An example of each of the types is given in Figure 3.1. As the GEV unifies the three types, there is no need to specify the distribution type prior to fitting.

![Three Types of GEV Distribution](image)

Figure 3.1: Plot of the three types of distributions that are unified by the GEV distribution. The shape parameter for the Gumbel is $\xi = 0$, for Fréchet, $\xi = 0.5$ and for Weibull, $\xi = -0.5$. The location and scale parameters are the same for all distributions, namely $\mu = 0$ and $\sigma = 1$.

**Type I: Fréchet**

For $\xi > 0$, $G(z)$ is type Fréchet and has a heavy tail. This distribution arises as the limiting distribution by Theorem 3.2.1 for the common distributions: Student’s t, Cauchy, Burr or log-gamma (McNeil et al., 2000). Applications of the Fréchet distribution occur for extremes in precipitation (Martins and Stedinger, 2000) and finance (Bradley and Taqqu, 2003). The support of the Fréchet distribution has a finite lower bound of $\mu - \sigma \xi$. 
3.3 Generalised Pareto Distribution

Type II: Weibull

For $\xi < 0$, $G(z)$ is type Weibull and the support has a finite upper bound, $\mu - \frac{\sigma}{\xi}$. By Theorem 3.2.1, the Weibull arises as the limiting distribution for the uniform and beta distribution (McNeil et al., 2000). Applications of extreme value distributions with a finite upper bound occur in extremes in temperature (S. Brown, Caesar, and Ferro, 2008), extremes in sea levels (Oliver et al., 2014) and athletics records (A. G. Stephenson and Tawn, 2013).

Type III: Gumbel

As $\xi \to 0$, the Gumbel type is attained as the limit of equation (3.4):

$$\lim_{\xi \to 0} G(z) = \exp \left\{ - \exp \left[ - \left( \frac{z - \mu}{\sigma} \right) \right] \right\}. \quad (3.8)$$

The tail of the Gumbel distribution decays exponentially. By Theorem 3.2.1, the Gumbel arises as the limiting distribution for extremes of the normal, exponential, gamma and lognormal distributions (McNeil et al., 2000).

3.3 Generalised Pareto Distribution

3.3.1 Limiting Distribution for Peaks Over Threshold

An alternative to adopting the block maxima approach of the GEV distribution is to define a sufficiently high threshold, above which the observations are assumed to be extreme. Then if two or more extreme events occur in the same block, these extremal observations can be included in the modelling, as opposed to only retaining the block maximum. The distribution of exceedances above a threshold can be obtained through a conditional argument,

$$P \{ X > u + y \mid X > u \} = \frac{1 - F(y + u)}{1 - F(u)}, \quad y > 0. \quad (3.9)$$

If $F$ is known, then it would be simple to obtain the conditional probability in equation (3.9). However, in practice $F$ is often unknown, and therefore the GEV distribution is used to approximate the distribution of extremes in the tail.

**Theorem 3.3.1.** (see S. Coles, (2001) Theorem 4.1) Assume the distribution of the sample maxima, $M_n$, can be approximated with a GEV distribution by Theorem 3.2.1, such that

$$P \{ M_n < z \} \approx G(z), \quad (3.10)$$

where

$$G(z) = \exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]_{+}^{-1/\xi} \right\}, \quad (3.11)$$
THE INTRODUCTION

with $[v]_+ = \max\{0, v\}$, $\mu \in \mathbb{R}$, $\xi \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. Then for large enough $u$, the distribution function of $Y = X - u$ conditional on $X > u$ is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-1/\xi},$$

(3.12)

where $y > 0$ and

$$\sigma_u = \sigma + \xi(u - \mu).$$

(3.13)

The parameters of the Generalised Pareto (GP) distribution are uniquely defined by those of the GEV distribution.

### 3.3.2 Intuition for the Limit Distribution

The following provides the intuition underlining why the GP distribution is used for modelling exceedances occurring above a sufficiently high threshold. Formal treatment, and proof that the GP is also the limiting distribution as $u$ increases, can be found in Leadbetter, (1983).

According to Theorem 3.2.1, for large $n$, we have

$$F^n(z) \approx \exp\left\{-\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

(3.14)

or equivalently

$$n \log F(z) \approx -\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}. 

(3.15)

For large $z$, using a Taylor Series expansion, $\log F(z) \sim -[1 - F(z)]$. It therefore follows

$$1 - F(z) \approx \frac{1}{n} \left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}. 

(3.16)

By evaluating $1 - F(z)$ at $z = u + y$ and $z = u$, the required result follows:

$$P\{X > u + y | X > u\} \approx \left[\frac{1 + \xi (u + y - \mu) / \sigma}{1 + \xi (u - \mu) / \sigma}\right]^{-1/\xi},$$

(3.17)

$$= \left(1 + \frac{\xi y}{\sigma_u}\right)^{-1/\xi}. 

(3.18)

### 3.3.3 Threshold Selection

There is subjectivity associated with fitting a GP distribution, such as setting a sufficiently high threshold prior to fitting. The two standard methods for threshold selection include (i) mean residual life plots, which can be performed prior to parameter estimation, and (ii) examining parameter stability based on fitting at a range of thresholds.
Mean Residual Life Plots

The first method, mean residual life plots, uses the mean exceedance above the threshold $u$, defined for $\xi < 1$ as

$$
E(Y) = E(X - u \mid X > u) = \frac{\sigma_u}{1 - \xi}.
$$

(3.19)

Assuming $u_0$ is an appropriate threshold, then for $u > u_0$ by equation (3.13),

$$
\sigma_u = \sigma_{u_0} + \xi (u - u_0)
$$

(3.20)

and

$$
E(X - u \mid X > u) = \frac{\sigma_{u_0} + \xi (u - u_0)}{1 - \xi}.
$$

(3.21)

The mean residual above $u_0$ is a linear function of $u$. Therefore to identify an appropriate threshold, we can plot the threshold, $u$, against the empirically estimated mean residual

$$
\frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u),
$$

(3.22)

where $u < x_{\text{max}}$, $n_u$ is the number of observations exceeding the threshold and $x_{(i)}$ are the order statistics above $u$. When the plot changes linearly above a given $u$, this indicates an appropriate threshold. An example of such a plot is given in Figure 3.2. The 95th and 99th quantiles are marked with red dotted lines. The plot is observed to begin to change linearly above approximately 0.3, suggesting an appropriate threshold is equal to or greater than $u = 0.3$. This is above the 95th quantile. Notice that the estimator for the mean residual deteriorates for a small number of exceedances. This occurs in Figure 3.2 for thresholds above $u = 1.2$.

Parameter Stability Plots

Alternatively, we can use the stability of parameter estimates to provide an indication for an appropriate threshold. As the threshold, $u$, increases, $\xi$ remains constant. Additionally, by rearranging (3.20) such that

$$
\sigma_u - \xi u = \sigma_{u_0} + \xi u_0
$$

(3.23)

then the transformed scale parameter, $\sigma^* = \sigma_u - \xi u$, remains constant with changes in threshold. An example is given in Figure 3.3. These plots also indicate an appropriate threshold is attained above $u = 0.3$. Similarly to mean residual life plots, at high thresholds where there are few exceedances the parameter estimates deteriorate.
Figure 3.2: The mean residual life plot for the example dataset in the extRemes R package (Gilleland and Katz, 2016) of Fort Collins precipitation. The gray dashed lines give the 95% confidence intervals and the red dotted lines give the 95th and 99th quantiles for reference.

Challenge of Threshold Selection

In practice both of these methods can be difficult to interpret and are therefore prone to ambiguity. Threshold selection also leads to a variance–bias trade off. The higher the threshold the better approximation to the limiting distribution, reducing bias, but the less data available for fitting, increasing the variance. Efforts have been made to automate threshold selection (Ferro and Segers, 2003; Fukutome, Liniger, and Sîveges, 2015), however no single solution exists to the problem and manual checking is often required.

To avoid the problem of threshold selection, the threshold is often set arbitrarily at the 95th or 99th quantile, under the assumption that this is sufficiently high. However, we can see from Figure 3.2 that this is not always a good assumption.

3.4 Limit Theorems for Stationary Series

3.4.1 Theorems for Stationary Series

One of the assumptions underlying extreme value theory is that the sequence of random variables that the maximum is taken over are independent. However, rainfall and many other environmental processes are not independent. To apply the limit theory of Theorem 3.2.1 to processes such as rainfall, the independence assumption needs to be relaxed, and an extension of the theorem
3.4. LIMIT THEOREMS FOR STATIONARY SERIES

Figure 3.3: The parameter stability plots corresponding to the example in Figure 3.2.

is needed for stationary processes.

Definition 3.4.1. A stochastic process \( \{X_i\}_{i \geq 0} \) is said to be stationary if, given any sequence of integers, \( \{i_1, i_2, \ldots, i_k\} \) and any integer \( m \), the joint distribution of \( (X_{i_1}, X_{i_2}, \ldots, X_{i_k}) \) and \( (X_{i_1+m}, X_{i_2+m}, \ldots, X_{i_k+m}) \) are identical.

The extension to stationary processes requires restrictions on the extent of long range dependence within the process. Therefore a condition is desired such that \( X_i > u \) and \( X_j > u \) will be ‘approximately’ independent if their separation, \(|i - j| = m\), is large, and the threshold, \( u \), is high. Intuitively this is true of processes like rainfall where it is expected that the occurrence of extreme rainfall today is independent of the extreme rainfall in the proceeding months.

To formalise mathematically the condition that restricts the long range dependence within a stationary series, the following definition for \( D(u_n) \) is introduced.

Definition 3.4.2. (Leadbetter, 1983) The \( D(u_n) \) condition holds if for any integers

\[
1 \leq 1 < i_1 < \cdots < i_p < j_1 < \cdots < j_p \leq n
\]  \hspace{1cm} (3.24)
for which $j_p - i_p > 1$, we have

\[
\left| \mathbb{P}\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{i_q} \leq u_n, \ldots, X_{i_r} \leq u_n\} 
- \mathbb{P}\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\}\mathbb{P}\{X_{j_1} \leq u_n, \ldots, X_{j_p} \leq u_n\} \right| \leq \alpha_{n,l},
\]

where $\alpha_{n,l} \to 0$ as $n \to \infty$ for some sequence $l_n = o(n)$.

For an independent sequence of variables, Definition 3.4.2 is satisfied trivially by any sequence $u_n$.

For a stationary sequence of variables, the $D(u_n)$ condition need only be satisfied by one sequence of thresholds $u_n$ that increase with $n$. In that case $\alpha_{n,l}$ may be non-zero, however, the difference in probabilities will be sufficiently small to ensure no effect on the extremal limit laws. Therefore for stationary series that satisfy Definition 3.4.2, the limit distribution of the sample maxima will converge to the same family of limit distributions as the sample maxima of an independent series, provided appropriate normalising constants exist.

**Theorem 3.4.3.** (see S. Coles, (2001) Theorem 5.1) Let $\{X_i\}_{i \geq 1}$ be a stationary process, and $M_n = \max\{X_1, \ldots, X_n\}$. If $\{a_n > 0\}$ and $\{b_n\}$ are sequences of normalising constants such that

\[
\mathbb{P}\left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \to G(z),
\]

where $G$ is a non-degenerate distribution function, and the $D(u_n)$ condition is satisfied with $u_n = a_n z + b_n$ for every real $z$, then $G$ is a member of the generalised extreme value family of distributions.

The caveat is that when we fit an extreme value distribution to a stationary series compared with an independent series, the parameter estimates are affected by short range dependence within the series. For example, for stationary series with short range dependence it is common for extrema to occur in clusters. A relationship exists between stationary series, $\{X_i\}_{i \geq 1}$, and independent series, $\{X_i^*\}_{i \geq 1}$, with the same marginal distribution that quantifies the extent of clustering at extreme levels.

**Theorem 3.4.4.** (see S. Coles, (2001) Theorem 5.2) Let $\{X_i\}_{i \geq 1}$ be a stationary process and $\{X_i^*\}_{i \geq 1}$ a sequence of independent variables with the same marginal distribution. Define $M_n = \max\{X_1, \ldots, X_n\}$ and $M_n^* = \max\{X_1^*, \ldots, X_n^*\}$. Under suitable regularity conditions,

\[
\text{Pr}\left\{ \frac{M_n^* - b_n}{a_n} \leq z \right\} \to G_1(z),
\]

(3.27)
as \( n \to \infty \) for normalising sequences \( \{a_n > 0\} \) and \( \{b_n\} \), where \( G_1 \) is a non-degenerate distribution function, if and only if

\[
Pr\left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \to G_2(z),
\]

where

\[
G_2(z) = G_1^{\theta}(z)
\]

for a constant \( \theta \) such that \( 0 < \theta \leq 1 \).

In Theorem 3.4.4, the constant \( \theta \) is typically called the extremal index (Leadbetter, 1983). The reciprocal of the extremal index can be interpreted as the limiting mean cluster size. For example, \( \theta = 0.5 \) indicates a limiting mean cluster size of 2 extreme observations. For independent series, \( \theta = 1 \), however the converse is not true. A stationary series can have a value of \( \theta = 1 \) corresponding to negligible dependence at asymptotically high thresholds.

### 3.4.2 Block Maxima for Stationary Series

For environmental processes like rainfall, because the independence assumption of Theorem 3.2.1 is violated, it is unclear whether the GEV is a suitable limit distribution. Theorem 3.4.3 however, provides the necessary justification so that the sample maxima of stationary series with limited long range dependence may reasonably be approximated using the GEV distribution.

In practice there is little difference between fitting a GEV to an independent series compared with a stationary series. The parameter estimates will be different, although this does not impose any additional difficulties for model inference. Caution is needed however to ensure the appropriateness of the approximation by a limit distribution. For stationary series with high levels of dependence, the effective number of independent observations reduces from \( n \) to \( n\theta \) thereby affecting the quality of the approximation.

### 3.4.3 Peaks Over Threshold for Stationary Series

In fitting a GP distribution, in contrast to a GEV distribution, the stationary case needs to be considered differently to the independent case. For a stationary series with short range dependence, clusters of extremal observations can occur above the threshold. In this instance, extremal observations may not be independent and this affects model inference and parameter estimation. Therefore the standard modelling approach for peaks over threshold needs to be adapted for dependent extremes.
It is possible to identify clusters of extremal observations prior to fitting a GP distribution. The cluster maxima can then be retained to achieve independent extrema while discarding the other observations within the cluster. This practice is referred to as declustering. Two common methods of declustering are runs (S. Coles, 2001) and the interval method (Ferro et al., 2003). Parameter estimates, however, are sensitive to declustering and this provides an additional challenge to threshold selection when fitting a GP.

### 3.5 Extremes of Non-stationary Processes

The limit theorems for the GEV and GP do not apply to non-stationary processes (see Definition 3.4.1). This is because an underlying assumption is violated: the sequence of random variables, \( \{X_i\}_{i=1}^{n} \), over which the maximum is taken is not identically distributed. In Section 3.4, the assumption of independence needed for the limit theorems was relaxed to include processes with limited long-range dependence, see Definition 3.4.2. No such extension exists in generality for processes which are not identically distributed.

Many environmental processes, such as temperature and rainfall, are non-stationary. These processes can be influenced by long term change, such as anthropogenic forcings due to climate change, or by natural variability, such as seasonality or large scale climate drivers. However, there are many common approaches used to circumvent non-stationarity (Albert M.G Klein Tank, 2009). In using these approaches, we can still operate within the statistical modelling framework to apply extreme value theory to non-stationary processes.

Time-varying parameters can be included within the GEV or GP distributions to model the non-stationarity. This assumes that the extremal distribution is gradually changing in time, such as from a shifting or scaling. For example if the GEV distribution is linearly shifted in time, then an appropriate form for the location parameter is

\[
\mu(t) = \mu_0 + \mu_1 t, \tag{3.30}
\]

where \( t \) is the index of the block. If the GEV distribution is both linearly shifted and scaled then

\[
\begin{align*}
\mu(t) &= \mu_0 + \mu_1 t, \tag{3.31} \\
\log(\sigma(t)) &= \sigma_0 + \sigma_1 t. \tag{3.32}
\end{align*}
\]

Here the log of the shape parameter is taken to ensure that this parameter remains positive despite being allowed to vary in time. In a univariate setting, given the difficulty of accurately estimating the shape parameter, this...
3.6 MODEL FITTING

Parameter is normally not varied in time. Examples of studies that use time-varying parameters for environmental applications include (Dupuis, 2014; Kyśelý, Picek, and Beranová, 2010; Min et al., 2013; Oesting and Stein, 2017; Sun, Renard, et al., 2015).

An alternative parameteric approach, is to address the non-stationarity by preprocessing the data prior to fitting. This may include fitting a time series model (Dupuis, 2012) or transforming the series, such as by using a Box–Cox transform (Eastoe and Tawn, 2009). Extreme value analysis can then be performed on the stationary residuals or transformed data. A time-varying threshold can be also used for the GP distribution (Eastoe et al., 2009). The drawback of this approach is that two sources of statistical uncertainty are created: one from the preprocessed model and one from the extremal model.

It some instances, it is also possible to stratify the maxima in such a way that stationarity may reasonably be assumed within each block. Extreme value analysis can then be applied to each set of stationary blocks. Examples of this method include stratifying in time, whether by month (Oesting and Stein, 2017), season (Min et al., 2013) or splitting the data into earlier and later blocks (Jakob and Walland, 2016).

3.6 Model fitting

3.6.1 Parameter Estimation

Maximum likelihood estimation (MLE) is commonly used to estimate the parameters of extreme value distributions. Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed observations from the probability density, $f(x; \theta_0)$, with parameter set, $\theta_0$. The maximum likelihood estimator, $\hat{\theta}$, is obtained by maximising the likelihood function

$$\mathcal{L}(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta). \quad (3.33)$$

In practice however, it is usually more convenient to maximise the log of the likelihood

$$\ell(\theta; x) = \sum_{i=1}^{n} \log f(x_i; \theta). \quad (3.34)$$

The maximum likelihood estimators are the same for both equations (3.33) and (3.34), due to the monotonocity of the log function. Maximising the log-likelihood rather than the likelihood helps to avoid issues with arithmetic underflow.
To obtain maximum likelihood estimates for the parameters of the GEV distribution, \( \vartheta = (\mu, \sigma, \xi) \), the GEV density can be substituted into equation (3.34). The GEV density, \( g(z) \), is given by

\[
g(z) = \frac{1}{\sigma} t^{\xi+1}(z) \exp[-t(z)],
\]

where

\[
t(z) = \begin{cases} 
1 + \xi \left( \frac{z - \mu}{\sigma} \right)^{1/\xi} & \xi \neq 0, \\
\exp \left( \frac{z - \mu}{\sigma} \right) & \xi = 0.
\end{cases}
\]

(3.35)

Parameter estimates are obtained numerically through maximising equation (3.34), as no analytical solution exists to the optimisation.

Similarly, maximum likelihood estimates for the parameters of the GP distribution, \( \vartheta = (\sigma_u, \xi) \), can be obtained by substituting the GP density into the log-likelihood equation (3.34). The density of the GP is

\[
h(y) = \begin{cases} 
\frac{1}{\sigma_u} \left( 1 + \frac{\xi y}{\sigma_u} \right)^{-1/\xi - 1}, & \xi \neq 0, \\
\frac{1}{\sigma_u} \exp \left( -\frac{y}{\sigma_u} \right) & \xi = 0.
\end{cases}
\]

(3.37)

It is important to understand the uncertainty of the maximum likelihood estimates. As the sample size, \( n \), approaches infinity, the distribution of the maximum likelihood estimates is asymptotically normal under appropriate regularity conditions. This is a desirable property as standard errors and confidence intervals can be easily calculated for the estimates.

However, when using MLE to estimate GEV parameters, the regularity conditions needed for asymptotic normality of estimators are not always valid. This is a consequence of the support of the GEV distribution, where \( \mu - \frac{\sigma}{\xi} \) is a finite upper endpoint when \( \xi < 0 \) and a finite lower endpoint when \( \xi > 0 \). R. L. Smith, (1985) showed that

- for \( \xi \geq -0.5 \) the MLE estimates are asymptotically normal
- for \( -1 < \xi < -0.5 \) the MLE estimates exists, but may not have the standard asymptotic properties
- and for \( \xi \leq -1 \) the MLE may not exist.

As we are discussing hydrological random variables in this thesis, we are mostly concerned with distributions that are heavy tailed, where \( \xi > 0 \). This means
that our MLEs will behave according to the standard asymptotic properties for MLE.

There are alternatives to maximum likelihood estimation for estimating GEV parameters. Parameter estimates and confidence intervals can also be obtained using profile likelihood (e.g., S. Coles, 2001). Probability weighted moments (PWM) (Greenwood et al., 1979; Hosking, Wallis, and Wood, 1985) or equivalently L-moments (Hosking and Wallis, 2005) are common in hydrological studies. For small sample sizes, PWM is known to be more robust numerically and more efficient than MLE (Hosking, Wallis, and Wood, 1985). However, PWM estimators are not straightforward to apply in a non-stationary setting. In contrast, time-varying parameters can be easily included in MLE. Bayesian methods are also common (S. G. Coles and Powell, 1996; A. Stephenson and Tawn, 2004) and are capable of dealing with non-stationary models. When the MLE does not exist, parameter estimation can be performed using maximum product spacing (R. Cheng and Amin, 1983; Ranneby, 1984).

### 3.6.2 Model Selection

Standard techniques for model selection, such as the log-likelihood ratio test or information criteria, can be applied to extreme value models.

A log-likelihood ratio test can be used to decide between two nested models. Two models, $M_1$ and $M_2$, are nested if the parameter set of one model is a subset of the other, $\vartheta_1 \subset \vartheta_2$. Let $k$ be the difference in the number of parameters between the two models. The deviance statistic is then defined as

$$
D = 2\{\ell(\vartheta_2) - \ell(\vartheta_1)\},
$$

where $\ell(\vartheta)$ is the evaluated log-likelihood (3.34). The deviance statistic, $D$ is approximately $\chi^2$ distributed with $k$ degrees of freedom, $D \sim \chi^2_k$. For hypothesis testing, a critical region with probability $\alpha$ can be determined using the quantiles of the $\chi^2_k$ distribution. If the deviance statistic falls outside the defined critical region, the simpler model $M_1$ is rejected under the null hypothesis in favour of $M_2$. This is referred to as the log-likelihood ratio test.

An approach based on information criteria (IC) can be used for non-nested models, and may also be preferred when the number of models for selection is large. The most common approaches are the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz et al., 1978). For a model with $p$ parameters fitted using $m$ observations, these IC are

$$
AIC(p) = -2\ell(p) + 2p \quad \text{and} \quad BIC(p) = -2\ell(p) + p \ln m.
$$

(3.39)
The preferred model minimises the selected information criterion. Both AIC and BIC penalize the model performance relative to the number of parameters used.

3.6.3 Model Diagnostics

The performance of the fitted model can be assessed using standard diagnostic plots, such as a quantile–quantile or probability plot. For non-stationary models this requires a transformation using the estimated parameters.

Consider the following non-stationary GEV model

\[ Z_t \sim GEV(\hat{\mu}(t), \hat{\sigma}(t), \hat{\xi}(t)). \]  

(3.40)

Without loss of generality, we can transform \( Z_t \) to a standard Gumbel distribution, \( GEV(0,1,0) \), using estimated parameters evaluated at \( t \)

\[ Z_t^* = \frac{1}{\hat{\xi}(t)} \log \left\{ 1 + \hat{\xi}(t) \left( \frac{Z_t - \hat{\mu}(t)}{\hat{\sigma}(t)} \right) \right\}. \]  

(3.41)

In the case of a non-stationary GP model, it is convenient to transform the GP distribution to a standard exponential distribution, which is a special case of the GP as \( \xi \to 0 \). Consider

\[ Y_t \sim GP(\hat{\sigma}(t), \hat{\xi}(t)), \]  

(3.42)

where the threshold may be time-varying, denoted \( u_t \). Indexing the observations, \( y_t \) relative to their respective thresholds, the transformation is given by

\[ Y_t^* = \frac{1}{\hat{\xi}(t)} \log \left\{ 1 + \hat{\xi}(t) \left( \frac{Y_t - \hat{u}_t}{\hat{\sigma}(t)} \right) \right\}. \]  

(3.43)

The probability plot is invariant to the choice of the reference distribution in both cases. However, this is not true for the quantile–quantile plot. Therefore selecting a reference distribution where \( \xi \to 0 \) provides an appropriate choice given the three types of the GEV family.

Although diagnostic plots are useful, when there are multiple models to be fitted it is convenient to automate checking the goodness of fit. A Kolmogorov–Smirnov test can be used to test whether the fitted distribution is appropriate given the sample (Stephens, 1970). When applying the test in this way, a critical region relative to the number of observations can be determined using a block bootstrap (X. Zhang et al., 2010).
3.7 Return periods

The last concept we would like to introduce in this chapter is the return period. The return period is an important measure of risk within extreme value analysis, as it provides an estimate of how frequently an extreme event is expected to occur. For this reason, in hydrological applications, return periods are used to set the design standards for mitigating structures. This helps to ensure the structure’s successful operation for the desired lifespan.

Consider the following example. Let $p$ be the probability of exceeding a level, $z_p$, in a given year, and let $T$ be a random variable describing the number of years until the first exceedance above $z_p$.

<table>
<thead>
<tr>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>$t-1$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exceedance</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>...</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Probability</td>
<td>$1-p$</td>
<td>$1-p$</td>
<td>$1-p$</td>
<td>...</td>
<td>$1-p$</td>
<td>$p$.</td>
</tr>
</tbody>
</table>

The waiting time until the first exceedance above $z_p$ can be modelled as a geometric random variable, where the geometric density is

$$P(T = t) = (1 - p)^{t-1} p, \quad t = 1, 2, \ldots.$$  

Here $p$ is the probability of success and $t$ is the number of trials. The expected value of $T$ is

$$E(T) = \sum_{t=1}^{\infty} t (1 - p)^{t-1} p = \frac{1}{p}.$$  

It follows that the return level, $z_p$, is expected to be exceeded on average every $1/p$ years, where $1/p$ is the return period. The level $z_p$ is therefore often referred to as a “1 in 1/p” year event. For example, if the level $z_{0.01}$ is exceeded then this is referred to as a 1 in 100 year event.

3.7.1 Stationary Series of Maxima in Time

For the GEV distribution and in a stationary context, the interpretation of a return period is straight-forward. For a given tail probability the return level, $z_p$, can be calculated directly from the quantile function of the GEV distribution

$$G^{-1}(1 - p) = z_p = \begin{cases} 
\mu - \frac{\xi}{\xi} \left( 1 - [-\log(1-p)]^{-\xi} \right) & \xi \neq 0, \\
\mu - \sigma \log [-\log(1-p)] & \xi = 0.
\end{cases} \quad (3.44)$$

For the GP distribution the interpretation is more complicated as one also needs to consider the frequency of events occurring above a threshold. Using the conditional argument for exceedances above a threshold, equation
(3.9), and substituting $\zeta_u = \mathbb{P}(X > u)$ into the equation (3.12) for the GP distribution, we have

$$
P(X > x) = \zeta_u \left(1 + \frac{\xi(x-u)}{\sigma_u}\right)^{-1/\xi},$$  
(3.45)

where $x = u + y$ and $y > 0$. Letting $\mathbb{P}(X > x_m) = m$, and inverting equation (3.45) we obtain the quantile function

$$Q(1-m) = x_m = \begin{cases} 
  u + \frac{\sigma}{\xi} \left(\frac{\zeta_u}{m}\right)^{\xi} - 1 & \xi \neq 0, \\
  u + \sigma \log \left(\frac{\zeta_u}{m}\right) & \xi = 0.
\end{cases}$$  
(3.46)

It follows that $x_m$ is the return level exceeded on average every $\frac{1}{m}$ observations. To adapt this to an annual context, if $n_y$ is the average number of exceedances per year and $N$ is the number of years, then the annual return level can be obtained by evaluating equation (3.46) at $\frac{1}{m} = Nn_y$. Empirical estimators arise naturally for both $\zeta_u$ and $n_y$.

For the GP and stationary series, the return levels may need to be adjusted to account for dependence. This is done using the extremal coefficient, $\theta$, of Theorem 3.4.4. As aforementioned, $\theta^{-1}$ is the expected limiting cluster size of observations above the threshold. We can therefore account for dependence by replacing $\zeta_u/m$ in equation (3.46) with $\theta \zeta_u/m$. Empirical estimators also exist for $\theta$ (Ferro et al., 2003).

### 3.7.2 Non-stationary Series of Maxima in Time

The formulation of the return period presented thus far relies on the probability of an exceedence remaining constant in time. In a non-stationary context this is clearly untrue. The purpose of the return period is often to quantify risk, so that the cost of adaptation or mitigation can be appropriately weighed. The problem therefore needs to be reformulated so that the risk can still be quantified and communicated in a non-stationary setting.

In an engineering context, instead of assuming the probability of an exceedence in each year is equal, it is prudent to consider the problem in terms of an exceedence occurring in a given design lifetime (Rootzén and Katz, 2013). However, this requires knowledge of how the extremal distribution will change in future. In a simple framework, where the time-varying behaviour is linear, the return period framework can be easily adapted for design lifetimes, as exemplified in Salas and Obeysekera, (2013).

For other time-varying covariates, future behaviour maybe unknown and this makes understanding future risk difficult. This is particularly relevant
for applications in climate. Projections based on numerical weather predictions can often be used to estimate the future exceedance probabilities. For example, if the temporal covariate was temperature, then projected changes in temperature could be used. However, for temporal covariates that represent more chaotic systems, future knowledge of a covariate may not be available. In these instances, a conditional framework may serve as the best way to communicate risk. For further discussion on risk and return levels under non-stationarity the reader is referred to Cooley, (2013) and L. Cheng et al., (2014).

3.8 Summary

In this chapter, we introduced the extreme value distributions of the GEV and GP. We provided the foundations necessary to fit these distributions and apply them in practice. Practical considerations were also outlined so that these limit distributions can be used for applications involving environmental data. Such considerations included the difficulty of threshold selection, how to handle non-stationarity, and how to handle dependent data. In the following chapter, we demonstrate how to apply this theory to an application of rainfall extremes in Australia. In subsequent chapters, we extend this theory in order to model spatial rainfall extremes.
Chapter 4

Non-stationarity of Seasonal Rainfall Extremes with El Niño Southern Oscillation

4.1 Introduction

To mitigate the potential impact of extreme rainfall events and to make informed decisions, we require an understanding of the likelihood of an extreme rainfall occurrence. The return period gives an estimate how frequently a given event is expected to occur on average, for example a one in 100 year event (see Section 3.7). As such, the return period is a measure of risk commonly used to understand probability of extreme events (Jakob, 2013; Rootzén et al., 2013). However, when return periods are estimated the climate is often assumed to be stationary (Salas et al., 2013). In reality, this is often not the case.

For rainfall extremes, non-stationarity may take the form of increasing or decreasing trends due to climate change (L. V. Alexander and Arblaster, 2017; Jakob and Walland, 2016). Non-stationarity may also be driven by natural variability, such as seasonal influences or climate drivers like the El Niño Southern Oscillation (ENSO) (King et al., 2013a; Min et al., 2013). To ensure potential risks are not underestimated it is therefore important both to detect and to quantify the impact of non-stationarity on the distribution of rainfall extremes. In the context of hydrological extremes, Milly et al., (2008) went as far as to state that ‘Stationarity is dead’.

Detecting whether non-stationarity is present within rainfall extremes can be difficult. The challenge arises in part due to the temporal length of the observational records, which are often too short to detect statistical significance reliably (Albert M.G Klein Tank, 2009; Westra and Sisson, 2011). Additionally, as rainfall is a highly stochastic phenomenon, with a high signal to noise
ratio, it can be difficult to identify smoothly varying changes in the distribution with time.

There are many common univariate methods for detecting non-stationarity. Methods free of distributional assumptions include the non-parametric rank tests of Mann–Kendall and Spearman’s Rho (Yue, Pilon, and Cavadias, 2002). These methods are useful for detection, but they do not quantify the influence of non-stationarity on the distribution. In contrast, although parametric methods require distributional assumptions, we can quantify the effect of non-stationarity on the distribution. For generalised extreme value (GEV) distributions, we can model the non-stationarity through the inclusion of time-varying coefficients within the parameters of the distribution (Section 3.5). Model selection methods can then be used to determine whether the inclusion of time-varying parameters is significant (see Section 3.6.2).

However, where spatial information is available, performing a univariate analysis to detect non-stationarity in extremes does not make optimal use of the data. Spatial methods benefit by pooling multiple observational records, such as from nearby stations, to increase the amount of information available for parameter estimation (Albert M.G Klein Tank, 2009; Hosking and Wallis, 2005). Westra and Sisson, (2011) showed that in the absence of longer temporal records, utilising spatial information improved the ability to detect whether non-stationarity was statistically significant.

In a spatial setting we could model the dependence between extremes (see Chapter 5). However, as we are only interested in the time-varying nature of the marginal parameters, this brings added mathematical and computational complication that is unnecessary. Instead, we opt to use the simplest extension of the GEV distribution to include a spatial dimension, the spatial GEV (Buishand, 1991). In the spatial GEV, the univariate marginal distributions at each station are assumed to be GEV distributions. The marginal parameters can also be assumed to be smoothly varying and these parameters can be modelled as a function of spatial covariates. The simplicity of this modelling approach comes from the fact that no dependence structure is specified, and instead observations at nearby stations are assumed to be independent.

We acknowledge assuming independence between neighbouring stations may not be a true physical representation of rainfall extremes. Despite this, the parameter estimates obtained through standard maximum likelihood estimation are unbiased and asymptotically normal (Varin, Reid, and Firth, 2011). The key difference is that to ensure the variance of the parameter estimates is not underestimated, a sandwich estimator is used (Godambe, 1960). The spatial GEV has been used for applications in rainfall extremes in studies including A. G. Stephenson, Saunders, and Tafakori, (2018), Westra and
In this chapter, we consider the ability of both univariate extreme value models and spatial GEV models to detect non-stationarity. The non-stationarity we are considering arises from natural variability driven by the El Niño Southern Oscillation (ENSO). ENSO is known to be one of the main drivers of rainfall variability in Australia (Risbey et al., 2009). We are interested in identifying and quantifying where ENSO affects the distribution of seasonal daily maxima in Australia. This will improve our understanding of non-stationarity and the risk posed by extreme rainfall events. In our modelling approach, we grid the domain at a 1 degree resolution and fit a spatial GEV to the stations in each gridded cell. This improves our ability to detect and model temporal non-stationarity by using spatial information, while also being able to use simple models to represent the underlying topography in each cell. Additionally, by performing our analysis on a grid, we can contrast our results directly to those produced using gridded products (Min et al., 2013; Risbey et al., 2009).

### 4.2 Spatial Generalised Extreme Value Distribution

In this section we introduce the spatial GEV. For each station, \( k \), assume the distribution of seasonal maxima is well approximated by a GEV distribution

\[
Z_k \sim GEV(\mu_k, \sigma_k, \xi_k). \tag{4.1}
\]

Therefore for the spatial GEV, if there are \( k \) stations, then there are \( 3k \) parameters to fit. This number of parameters can grow large relative to the data available for parameter estimation. Given this, we assume that the marginal parameters of the GEV distribution, \( \mu \), \( \sigma \) and \( \xi \), are smoothly varying in space and time. We can then express the marginal parameters as simple linear functions of geographic and temporal covariates

\[
\mu(x, t) = X_\mu \beta_\mu + T_\mu \beta_\alpha \tag{4.2}
\]
\[
\sigma(x, t) = X_\sigma \beta_\sigma + T_\sigma \beta_\gamma \tag{4.3}
\]
\[
\xi(x, t) = X_\xi \beta_\xi + T_\xi \beta_\delta, \tag{4.4}
\]

where for \( \mu \), \( X_\mu \) is the design matrix of geographic covariates, \( \beta_\mu \) are the related geographic coefficients, \( T_\mu \) is the design matrix of temporal covariates and \( \beta_\alpha \) are the related temporal coefficients. The equations for \( \sigma \) and \( \xi \) are analogously defined. The spatial GEV then can be written as

\[
Z(x, t) \sim GEV(\mu(x, t), \sigma(x, t), \xi(x, t)). \tag{4.5}
\]
To fit the parameters of a spatial GEV we maximise the log-likelihood

\[ \ell(\theta) = \sum_{k \in K} \sum_{n=1}^{N} \log f(x_{kn}, t_{kn}; \theta), \]

where \( f \) is the probability density for the GEV distribution, \( N \) is the number of observations at a given station and \( K \) is the set of all stations. Here the univariate marginal distributions of rainfall extremes at locations, \( x^{k_1} \) and \( x^{k_2} \) are assumed to be independent.

For hydrological applications assuming independence of nearby locations may prove to be a poor modelling assumption. If this is the case, instead of the standard maximum likelihood equation, the likelihood of equation (4.6) takes the form of a misspecified model (White, 1982). A consequence of a misspecified likelihood is that the estimated confidence intervals will be underestimated. To adjust for this, the asymptotic covariance matrix can be estimated using a sandwich estimator (Huber, 1967)

\[ \Sigma = H^{-1}VH^{-1}, \]

where \( H = -\mathbb{E}[\ell''(\theta)] \) is the negative second derivative of the log likelihood, equation (4.6), and \( V = \text{Var}[\ell'(\theta)] \) is variance of the derivative of the log likelihood. The estimator for \( H \) is the Hessian matrix from the log likelihood maximisation evaluated at the estimated parameters, while \( V \) is empirically estimated (Varin, Reid, et al., 2011). If \( H = V \), \( \Sigma \) reduces to the inverse of the Fisher information matrix.

These mathematical concepts are extended in Chapter 5 where we introduce models for spatial extremes with dependence.

### 4.3 Forms of ENSO Non-stationarity

ENSO has three different phases, El Niño, Neutral and La Niña. During La Niña phases there is an increase in mean and total rainfall, particularly in Eastern Australia (Risbey et al., 2009). In contrast, a decrease is observed during El Niño phases. To model the non-stationarity due to ENSO, we used the Southern Oscillation Index (SOI), Figure 4.1.

The SOI is a measure for the strength of ENSO, given by the normalised pressure difference between Tahiti and Darwin. Sustained monthly values of the SOI that are large and negative characterise El Niño phases, whereas sustained large and positive SOI values characterise La Niña phases. We model the GEV parameters as time-varying functions using the average seasonal SOI as a covariate.
4.3. FORMS OF ENSO NON-STATIONARITY

The forms of non-stationarity we consider include:

\[
\begin{align*}
\text{Model 1: } & \text{GEV}(\mu, \sigma, \xi), & (4.8) \\
\text{Model 2: } & \text{GEV}(h(\mu, \text{SOI}), \sigma, \xi), & (4.9) \\
\text{Model 3: } & \text{GEV}(h(\mu, \text{SOI}), h(\sigma, \text{SOI}) \xi), & (4.10)
\end{align*}
\]

where depending on the model, \( \mu \) and \( \sigma \) are modelled as time-varying parameters according to the function, \( h \). For simplicity, we are assuming if the parameters are time-varying, then the time-varying function is the same for both \( \mu \) and \( \sigma \). We considered the following forms for the function \( h \):

\[
\begin{align*}
\text{Symmetric: } h(p, \text{SOI}) &= p_0 + p_1 \text{SOI} & (4.11) \\
\text{Asymmetric: } h(p, \text{SOI}) &= p_0 + p_1 \text{SOI} + p_2 [\text{SOI}]^+ & (4.12) \\
\text{One-sided asymmetric: } h(p, \text{SOI}) &= p_0 + p_1 [\text{SOI}]^+ & (4.13)
\end{align*}
\]

where \([\text{SOI}]^+ = \max(\text{SOI}, 0)\). These non-stationary forms were chosen specifically to target the dichotomy between the ENSO phases, El Niño and La Niña. We also considered a nonlinear form for non-stationarity

\[
\text{Nonlinear: } h(p, \text{SOI}) = p_0 + p_1 \text{SOI}^2. & (4.14)
\]

However, we found this form resulted in overfitting and we will therefore not discuss it further.
For univariate model selection, a log-likelihood ratio test (see equation (3.38)) was used for selection between the nested models of equations (4.8), (4.9) and (4.10). The Akaike Information Criterion (see equation (3.39)) was then used to decide between the different forms of non-stationarity given by equations (4.11), (4.12) and (4.13). In general, the log-likelihood ratio test is preferred where possible as it has an associated hypothesis test. However, given we are required to evaluate between a large number of models, we have chosen to use a mixture of these two model selection methods.

For model selection involving spatial GEV models, we use an information criterion for misspecified likelihoods, the Takeuchi Information Criterion (TIC) (Varin and Vidoni, 2005). The TIC is an adjusted version of the AIC and is defined by

\[
TIC = -2\ell(\hat{\theta}) + 2\text{tr}(VH^{-1}),
\]

where \(\ell(\hat{\theta})\) is the log-likelihood maximised by the fitted parameters.

### 4.4 A Stationary GEV Distribution

Prior to considering a non-stationary model, we fit a stationary GEV distribution to the observed annual maximum at each rainfall station. The fitted GEV parameters for location and shape are shown in Figures 4.2 and 4.3 respectively. We do not show the visualisation for the scale parameter as the figure was visually similar to Figure 4.2, except for the expected differences in parameter range.

From visual inspection, we observed that the estimated location and scale parameters vary smoothly in space relative to the local topography and climate. The estimated shape parameter however is highly variable spatially and subject to high levels of uncertainty. Given the challenge of reliably estimating this parameter, the addition of nearby spatial information would help to reduce uncertainty.
4.4. A STATIONARY GEV DISTRIBUTION

Figure 4.2: Estimated location parameter from the GEV distribution fitted to annual maximum daily rainfall.

Figure 4.3: Estimated shape parameter from the GEV distribution fitted to annual maximum daily rainfall.
4.5 Symmetric Non-stationarity

Below we investigate whether models with symmetric non-stationarity in the location parameter are preferred to stationary models (see Section 4.3). This form of non-stationarity is commonly used in the literature (Min et al., 2013; Risbey et al., 2009), so it serves as a benchmark to which we can compare our results.

4.5.1 Model Selection for Univariate GEV Distribution

For the univariate GEV distribution, Figure 4.4 shows where non-stationarity in the location parameter was found to be significant at a 5% level. The plots in the right column show the stations for which non-stationarity was significant, compared with those in left column where the stationary GEV distribution was preferred.

In the right column, we observe that there is a non-stationary influence of ENSO predominantly in Eastern Australia during the seasons of Winter (JJA), Spring (SON) and Summer (DJF), with a negligible influence for Autumn (MAM). However, while we are capturing the broad spatial influences we expect (Min et al., 2013; Risbey et al., 2009), we notice that model selection does not partition our domain into definitive regions where non-stationarity is significant. This is shown by the overlapping of significant and non-significant regions between the left and right columns.

It is reasonable to assume that the influence of ENSO would vary smoothly in space, so at nearby stations we expect similar models to be selected. The differences in model selection at nearby stations could be attributed to the observations at these stations having different record lengths and record periods. However, it is more likely that in a univariate setting we are not able to detect reliably if non-stationarity is present.

We did consider using a univariate peaks-over-threshold approach to improve our ability to detect non-stationarity. However, we experienced similar difficulty detecting non-stationarity in the scale parameter of the GP distribution with ENSO. This problem was further compounded by the selection of threshold. At lower thresholds there is more data available, however the approximation to the limit distribution is less sound. Given this, in the following section we consider the results obtained from using a spatial GEV.
4.5. SYMMETRIC NON-STATIONARITY

Figure 4.4: Stations where symmetric non-stationarity in the location parameter of the GEV distribution was found to be significant at a 5% level (right column). Stations where the stationary model was preferred (left column). Here the strip labels are given by the monthly acronyms corresponding to the season. For example, DJF, corresponds to the Summer season of December, January and February. The boolean labels correspond to if significance was detected.
4.5.2 Model Selection for a Spatial GEV

Given the difficulty of detecting non-stationarity in a univariate setting, we consider the same form of non-stationarity using a spatial GEV. For fitting the spatial GEV, we assume that the form of the temporal covariates is constant across each gridded cell. We also assume a constant location, scale and shape parameter across each cell. This greatly reduces the complexity of the problem. In effect, we are therefore using the TIC (equation (4.15)) to determine the preferred form of non-stationarity.

Figure 4.5 shows where the symmetric form of non-stationarity was preferred using the TIC. If the symmetric model was preferred, a cross is shown in the lower left corner of the cell. The shading of each of the one degree cells shows the magnitude of the coefficient for the SOI from the fitted spatial GEV model.

This visualisation mirrors the patterns in which ENSO was found to have a non-stationary influence on the univariate GEV distribution in 4.4. However in contrast, the pattern of response shown in Figure 4.5 is smoothly varying. In Eastern Australia, from Winter through to Summer, there is a shift in the marginal GEV distribution right with SOI. There is also a shift in the distribution with ENSO in Northern Australia from Spring to Autumn. These results are similar to those of Risbey et al., (2009), where the correlation was considered between SOI and the mean seasonal rainfall. Our results are also consistent with those given in (Min et al., 2013).
4.6 Non-stationary forms of the Spatial GEV with ENSO

Given we are able to replicate what we observe for the symmetric form of non-stationarity existing in the literature, we now consider other possible forms of non-stationarity with ENSO. In Figure 4.6 we plot whether a non-stationary spatial GEV model was preferred to the stationary spatial GEV model. It is difficult to compare non-stationary forms by considering the magnitude of a single parameter as we did in Figure 4.4, so in Figure 4.6 we have shaded each gridded cell by the signs of the temporal covariates. There are clear regions in Australia where ENSO has a non-stationary influence on the distribution of seasonal extremes. We observe these regions to be broadly consistent across all the different forms of non-stationarity considered.
For both the symmetric and one-sided asymmetric forms of non-stationarity, models with a time-varying coefficient in the location parameter shift the distribution to the right, and models with a time-varying coefficient in the scale parameter increase the spread of the distribution. For the asymmetric model, the signs of the temporal covariates are both negative and positive, so it is less clear how ENSO influences the distribution in this model. Examining the fitted parameters closer, the signs of the coefficients in equation (4.12) are generally opposing. Given the sign and magnitude of these fitted parameters, in conjunction with our physical understanding of ENSO, we believe this fitted model is not realistically reflecting the physical process. We are therefore suspicious of overfitting, particularly in areas of less data.

We wish to note that many cells in Northern Australia only contain one or two stations. As such, the results in these regions are less robust than for gridded cells with many stations. In particular, if there is only one station in the cell then the information criterion becomes a much weaker form of model selection compared with the log-likelihood ratio test we used to produce the univariate results of Section 4.5.1. We decided to include these cells regardless, given the spatial cohesion observed in Figure 4.5. We also wish to note that many stations in Western Australia, particularly those recording for Autumn (MAM), have periods of little rainfall. This makes the approximation to the GEV limit distribution poor. We are therefore also cautious of the non-stationary result in Western Australia during this period.

These results have helped us to understand in which regions of Australia there is a non-stationary influence of ENSO. The question is now, which non-stationary form is preferred.
4.6. NON-STATIONARY FORMS OF THE SPATIAL GEV WITH ENSO

Figure 4.6: Regions where a non-stationary ENSO influence was detected on the seasonal daily maxima. The rows give the form of non-stationarity considered, abbreviated for convenience, with the numeric model referring to the model type in Section 4.3. The columns give the austral season. Each cell is coloured by the signs of the temporal covariates. Only cells in which a non-stationary spatial GEV model was preferred to a stationary model under the TIC are shown.
4.6.1 Preferred from of Non-stationarity

For the reduced set of suitable non-stationary forms, symmetric and one-sided asymmetric, we show which models were preferred under the TIC in Figure 4.7. In some places, and in some seasons, the type of non-stationarity is clear. For example, the symmetric model with a time-varying location parameter is preferred in parts of Eastern Australia and Southeast Australia for Winter and Spring seasons. The symmetric model with a time-varying location and scale parameter is preferred in mid latitude parts of Eastern Australia for Winter and Spring, and in Southeast Australia in Summer.

![Map of Australia showing preferred non-stationary models](image)

Figure 4.7: The preferred from of non-stationarity as determined using the TIC. Non-stationary models are abbreviated in the legend, but correspond to the models and non-stationary forms given in Section 4.3.

In some parts of Australia, particularly Northern and Central Australia, the preferred form of non-stationarity is less clear. We partially attribute this to that fact that many of the grid cells in these regions have only a few sta-
4.7. CONCLUSIONS

As with the stations, and that the recording period of these stations cover different years. The influence of ENSO varies annually (Risbey et al., 2009). Therefore different recording years capture different ENSO phases and consequently we will have different parameter estimates if we use data over different recording periods.

This suggests determining non-stationarity from only a handful of stations in a single gridded cell in these regions is not appropriate. By extension, this is also true for these regions in gridded datasets. It is common to mask areas of central Australia in gridded products due to lack of data. This may also be necessary for regions in far Northern Australia when considering applications of extremes.

4.7 Conclusions

We have shown that ENSO has a non-stationary influence on the distribution of seasonal daily rainfall extremes in Australia. To reliably detect this non-stationary influence, we gridded the domain, pooled the station data within each one degree gridded cell and fitted a spatial GEV. We then used model selection to determine which form of the non-stationarity with ENSO was preferred. Using a spatial model was necessary as we found model selection to produce contradictory results in a univariate setting.

In adopting this spatial approach, we further developed our understanding of where and in which season ENSO affects the distribution of seasonal rainfall extremes. We considered a range of different non-stationarity forms to model the influence of ENSO on the marginal GEV distributions. We found that different non-stationary forms were preferred in different parts of Australia and for different seasons. Although, we note that when we compared where, and for which seasons, non-stationarity was detected under the different models, we found the results to be broadly consistent. For example, a non-stationary influence of ENSO was detected across much of Eastern Australia, particularly in Winter and Spring for all models. For effective mitigation and accurate risk assessment, it is therefore clear that quantifying any non-stationarity due to ENSO is necessary.

We note our ability to decide between non-stationary forms depended on the number of stations in each gridded cell and the period of record of these stations. Ultimately the forms of non-stationarity we were able to reliably consider using this approach where very simple, with linear time-varying coefficients. As such, using spatial models to consider larger regions and including geographic covariates is needed to truly understand the influence on ENSO. We leave this extension for future work. In the following chapter, we introduce other spatial models for extremes that we could potentially also use for this
Detecting non-stationarity

Extending upon this research, we are also interested in considering the relationship between ENSO and other climate drivers, and how these other drivers affect non-stationarity. Two key drivers of interest are the Indian Ocean Dipole (IOD) (Risbey et al., 2009) and the Interdecadal Pacific Oscillation (IPO) (Power et al., 2006). This was explored for limited forms of non-stationarity by (Min et al., 2013).

In terms of modelling decisions, we used the SOI as measure for the strength of ENSO. However, the SOI is a single index and ENSO is a complex, non-linear process. Therefore there is the incentive to experiment with other measures for ENSO strength in order to fully understand non-stationarity. Also, in using the raw maxima and a spatial GEV instead of a gridded product, we did not reduce observations in each cell to a single aggregate time series. In future, we would be interested in comparing the magnitude of our fitted time-varying parameters to those obtained from gridded products. This is of particular interest given we know gridded products often underestimate extremes (Contractor et al., 2015; King et al., 2013b).
Chapter 5

An Introduction to Max-stable Processes

5.1 Introduction

For locations that are geographically close, it can be expected that these locations are likely to experience similar impacts from extreme rainfall. Therefore to assess risks from extreme rainfall accurately, it is important to understand the impact of extreme rainfall in a region, not just a single location. For example, if extreme rainfall occurs at one location, what is the probability of an event of equivalent size occurring at nearby locations? Or alternatively, what is the probability of extreme rainfall occurring at a location without a station? To answer these kinds of questions, max-stable processes provide an ideal modelling tool.

Max-stable processes can be thought of as a natural extension of univariate extreme value theory, and they are suitable for use in modelling extremes in continuous space with dependence. This class of models arises from asymptotic arguments that provide mathematical justification for extrapolation into the joint-tail of the distribution. Additionally, these processes have the desirable property that the univariate marginal distributions are GEV distributions, and the multivariate marginal distributions are multivariate extreme value distributions. As such, max-stable processes have been used to model rainfall extremes in numerous studies (Bechler, Bel, and Vrac, 2015; Davison, S. Padoan, Ribatet, et al., 2012; Oesting and Stein, 2017; Saunders et al., 2017; R. L. Smith, 1990).

Max-stable processes can be used to quantify the risk to a region from a given extreme rainfall event (De Fondeville and Davison, 2016; Ribatet, 2013). Let $Z(x)$ be the amount of rainfall at location $x$, where $x \in \mathcal{X} \subset \mathbb{R}^2$. An event we may be concerned with is when the total rainfall in the region exceeds a
critical level, \( z_{\text{crit}} \),
\[
\mathbb{P} \left( \int_{\mathcal{X}} Z(x) \, dx > z_{\text{crit}} \right), \quad z_{\text{crit}} \geq 0. \tag{5.1}
\]
In practice, this situation can arise when the total rainfall in a catchment exceeds a given level and riverine flooding ensues. Equation (5.1) therefore provides an important measure of risk. Alternatively, consider the probability of the largest pointwise rainfall in a region exceeding a critical level,
\[
\mathbb{P} \left( \sup_{x \in \mathcal{X}} Z(x) > z_{\text{crit}} \right), \quad z_{\text{crit}} \geq 0. \tag{5.2}
\]
This type of risk is indicative of flash flooding, which is highly localised.

Both these risk measures, of aggregate and pointwise rainfall, can be evaluated using a max-stable process, albeit often numerically. Max-stable processes therefore provide a powerful modelling tool that can be used to understand the risks from extreme rainfall.

In this chapter we introduce the theory necessary for using max-stable processes to model spatial extremes. This includes introducing the spectral representation of max-stable processes, giving the common parametric forms, describing how to fit the models and outlining simulation methods. A brief comparison to alternative methods for modelling spatial extremes is also provided. For further detail on the content presented in this chapter the reader is referred to the textbook (Dey, Jiang, and Yan, 2015) or the review papers (Davison and Huser, 2015; Davison, S. Padoan, et al., 2012).

5.2 Spatial Statistics

Before introducing max-stable processes, we require some basic terminology from spatial statistics (for reference see Gelfand et al., (2010)). Let \( \{Y(x) : x \in \mathcal{X} \subset \mathbb{R}^d \} \) be a spatial stochastic process, such that for any sequence of spatial locations, \( \{x_i \in \mathcal{X} : 1 \leq i \leq k \} \), the finite-dimensional distribution function is given by
\[
F_{Y(x_1), \ldots, Y(x_k)}(y_1, \ldots, y_k) = \mathbb{P}(Y(x_1) \leq y_1, \ldots, Y(x_k) \leq y_k). \tag{5.3}
\]
The process, \( Y(x) \), is strictly stationary, if for all lag vectors, \( h \in \mathbb{R}^d \),
\[
F(y_1, \ldots, y_k; x_1, \ldots, x_k) \overset{d}{=} F(y_1, \ldots, y_k; x_1 + h, \ldots, x_k + h). \tag{5.4}
\]
For a strictly stationary Gaussian process, \( \{W(x) : x \in \mathbb{R}^d \} \), the following properties hold:
\[
\mathbb{E}[W(x + h) - W(x)] = 0, \tag{5.5}
\]
5.2. SPATIAL STATISTICS

and

\[
\text{Cov}(W(x + h), W(x)) = \text{Cov}(W(0), W(h)) = C(h),
\]

where the function \( C(h) \) is known as the covariance function. It follows that \( C(0) = \text{Var}(W(h)) \). Any spatial stochastic process satisfying equations (5.5) and (5.6) is called weakly stationary or second order stationary. Gaussian processes that are second order stationary are therefore also strictly stationary.

An alternative to specifying the covariance is to use a variogram (Matheron, 1971). The variogram is defined as

\[
2\gamma(h) = \text{Var}(Y(x + h) - Y(x))
\]

If the spatial stochastic process, \( Y(x) \), of equation (5.7) is second order stationary, then

\[
2\gamma(h) = \mathbb{E} [Y(x + h) - Y(x)]^2,
\]

and the variogram can be written in terms of the covariance function

\[
\gamma(h) = C(0) - C(h).
\]

It follows that for second order stationary processes the correlation function is

\[
\rho(h) = \frac{\text{Cov}(Y(x), Y(x + h))}{\sqrt{\text{Var}(Y(x))\text{Var}(Y(x + h))}} = \frac{C(h)}{C(0)},
\]

Common parametric forms for \( \rho(h) \), where \( h = x_i - x_j \), and \( x_i, x_j \in \mathcal{X} \) include:

- **Powered Exponential**: \( \rho(h) = \exp \left[ - \left( \frac{||h||}{r} \right)^{s} \right], \quad r > 0, 0 < s \leq 2 \)

  \[
  \text{(5.11)}
  \]

- **Cauchy**: \( \rho(h) = \left[ 1 + \left( \frac{||h||}{r} \right)^2 \right]^{-s}, \quad r > 0, 0 < s \)

  \[
  \text{(5.12)}
  \]

- **Whittle–Matérn**: \( \rho(h) = \frac{2^{1-s}}{\Gamma(s)} \left( \frac{||h||}{s} \right)^r K_r \left( \frac{||h||}{s} \right), \quad r > 0, s > 0. \)

  \[
  \text{(5.13)}
  \]

Here \( r \) is the range parameter, \( s \) is a smoothness parameter and \( K_r \) is the modified Bessel function of order \( r \) (e.g. Davison and Gholamrezaee, 2011). The parameter \( r \) indicates the distance limit of correlation for our process. For these correlation functions \( \rho(0) = 1. \) However, it is possible to model
the correlation function with \( \rho(0) = 1 - \varepsilon \), where \( \varepsilon \) represents a nugget effect arising from small-scale variability, such as natural variation or measurement error.

If the correlation process depends only on the distance, \( \| h \| \), not the direction between \( x_i \) and \( x_j \), then the correlation function is isotropic. One possible approach to model anisotropy is to replace \( h \) in the correlation function with \( (h^T A h) \) where \( A \) is a positive definite matrix with unit determinant (eg. Davison, S. Padoan, et al., 2012).

There are some cases where the covariance function of a spatial stochastic process does not exist. In this instance it can be convenient to consider intrinsically stationary processes. Intrinsically stationary processes, are those where the spatial increments are second order stationary. That is the process \( \{ Y(x + h) - Y(x) : x \in X \} \) is stationary for all lag vectors \( h \in \mathbb{R}^d \). For these cases the variogram is still defined but may be unbounded. An example of a process that is intrinsically stationary, but not second order stationary is one-dimensional Brownian motion, which has variogram \( \gamma(h) = |h| \).

This spatial statistical background provides the foundations necessary to introduce max-stable processes.

5.3 Max-Stable Processes

The intuition that underlies our use of a max-stable process to model spatial extremes is similar to the asymptotic argument that prompts the use of the GEV distribution. Let \( \{ Y_i \}_{i \geq 1} \) be identical copies of a stochastic process that maps \( X \) to \( \mathbb{R} \), and define the partial maxima process as \( \{ \max_{i=1}^n Y_i(x) : x \in X \} \). It is possible to estimate the distribution of the partial maxima process, but for small tail probabilities the quality of this estimate is likely to be poor. This motivates our approximating the partial maxima process by its limit process for large \( n \). Under appropriate conditions, the limit process is a max-stable process.

5.3.1 Limiting Process

The definition of a max-stable process is as follows:

**Definition 5.3.1.** Let \( \{ Z_i \}_{i \geq 1} \) be a sequence of independent copies of a stochastic process \( \{ Z(x) : x \in X \subset \mathbb{R}^d \} \). The process \( Z(x) \) is max-stable, if there exist normalising functions, \( \{ a_n(x) \} \in \mathbb{R}^+ \) and \( \{ b_n(x) \} \in \mathbb{R} \), such that

\[
Z(x) \overset{d}{=} \lim_{n \to \infty} \frac{\max_{i=1}^n Z_i(x) - b_n(x)}{a_n(x)}, \quad x \in X.
\]
5.3. MAX-STABLE PROCESSES

If the limit process for the partial maxima process exists and is non-degenerate, then it is a max-stable process.

**Theorem 5.3.2.** (see Dey et al., (2015) Theorem 9.2.2 and cf. to De Haan, (1984)) Let \( \{Y_i\}_{i \geq 1} \) be a sequence of independent copies of a stochastic process \( \{Y(x) : x \in \mathcal{X}\} \) with continuous sample paths. If there exist continuous normalising functions, \( \{a_n(x)\} \in \mathbb{R}^+ \) and \( \{b_n(x)\} \in \mathbb{R} \), such that

\[
\lim_{n \to \infty} \frac{\max_{i=1}^n Y_i(x) - b_n(x)}{a_n(x)} = Z(x), \quad x \in \mathcal{X}
\]

and \( Z(x) \) is non-degenerate, then the process \( Z(x) \) is a max-stable process.

If \( \mathcal{X} \) is a single point, then by Theorem 3.2.1 the max-stable process reduces to the GEV distribution. The univariate marginals of a max-stable process therefore belong to the GEV family. The finite-dimensional marginal distributions are multivariate extreme value distributions.

If the univariate marginals follow a standard Fréchet distribution, \( \text{GEV}(1,1,1) \),

\[
P[Z(x) \leq z] = \exp\left(\frac{-1}{z}\right), \quad z > 0, \tag{5.14}
\]

then the max-stable process is called a simple max-stable process. Without loss of generality, the univariate marginals of a max-stable process can be transformed to standard Fréchet marginals using

\[
\tilde{\xi}_x = \left[1 + \frac{\xi_x(z_x - \mu_x)}{\sigma_x}\right]^{-1/\xi_x}, \quad \tag{5.15}
\]

where \( [c]_+ = \max(c,0) \). Mathematically it is convenient to use simple max-stable processes, as the resulting mathematical arguments are more elegantly specified. For the remainder of this chapter we will assume that the max-stable process is a simple max-stable process unless otherwise specified.

### 5.3.2 Spectral Representation

Neither Definition 5.3.1 or Theorem 5.3.2 provide a specific form for the max-stable process; for this a spectral representation is required.

**Theorem 5.3.3.** (see Dey et al., (2015) Theorem 9.2.4 and cf. to De Haan, (1984) and Penrose, (1992)) Any non-degenerate simple max-stable process \( \{Z(x) : x \in \mathcal{X}\} \) defined on a compact set \( \mathcal{X} \subset \mathbb{R}^d \), \( d \geq 1 \), with continuous sample paths satisfies

\[
Z(x) \overset{d}{=} \max_{i \geq 1} \xi_i f_i(x), \quad x \in \mathcal{X}, \tag{5.16}
\]

where

\[
\xi_i = \frac{\max_{x \in \mathcal{X}} \xi_i(x)}{\max_{x \in \mathcal{X}} \xi_i(x)}, \quad f_i(x) = \frac{\xi_i(x)}{\sum_{j \geq 1} \xi_j(x)},
\]

\( \xi_i(x) \) is the \( i \)-th largest value of \( \xi(x) \) for each \( x \in \mathcal{X} \), and \( f_i(x) \) is the \( i \)-th largest value of \( f(x) \) for each \( x \in \mathcal{X} \).
where \( \{(\zeta_i, f_i) : i \geq 1\} \) are the points of a Poisson process on \((0, \infty) \times \mathcal{C}\) with intensity \(\zeta^{-2} d\zeta \nu(df)\) for some locally finite measure \(\nu\) defined on the space \(\mathcal{C}\) of non-negative continuous functions on \(\mathcal{X}\) such that

\[
\int_{\mathcal{C}} f(x) \nu(df) = 1, \quad x \in \mathcal{X}.
\]  

(5.17)

If \(\nu\) is a probability measure, the spectral representation of Theorem 5.3.3 can be simplified to

\[
Z(x) \overset{d}{=} \max_{i \geq 1} \zeta_i Y_i(x), \quad x \in \mathcal{X},
\]

(5.18)

where \(\{\zeta_i : i \geq 1\}\) are points of a Poisson process on \((0, \infty)\) with intensity \(\zeta^{-2} d\zeta\), and \(Y_1, Y_2, \ldots\) constitute a sequence of independent copies of a non-negative stochastic process \(\{Y(x) : x \in \mathcal{X}\}\) with continuous sample paths such that the \(E\{Y(x)\} = 1\) for all \(x \in \mathcal{X}\). For the applications of extreme rainfall considered in this thesis the spectral representation of equation (5.18) is appropriate.

A motivation for the spectral representation follows from the decomposition of the max-stable process into a radial component and a 'pseudo' angular component. The radial component is determined by the magnitude of points in the Poisson process \(\zeta_i\) and the angle is determined by the stochastic process \(Y_i(x)\).

Decomposition of the spectral representation naturally lends itself to an interpretation in terms of rainfall-storms, as proposed by R. L. Smith, (1990). The points of the Poisson process may be interpreted as storm magnitudes or intensities. The stochastic process, \(Y(x)\) then defines the spatial extent of the storm over the space \(\mathcal{X}\), colloquially the storm shape. The amount of rainfall from the \(i^{th}\) storm at location \(x\) is therefore \(\zeta_i Y_i(x)\). The max-stable process then arises by taking the pointwise maxima over all storms. The intensity measure for the Poisson process is such that most storms will contribute only small amounts of rainfall and only relatively few storms will contribute extreme amounts of rainfall. In practice, this theoretical rainfall-storm interpretation lacks physical meaning. However, it is useful in terms of garnering our intuition for max-stable processes.

5.3.3 Parametric Forms

The spectral representation of equation (5.18) requires choosing a suitable process for \(Y_i(x)\). Common spectral representations include R. L. Smith, (1990), Schlather, (2002), Brown–Resnick (B. M. Brown and Resnick, 1977; Kabluchko, Schlather, and De Haan, 2009) and the extremal-\(t\) (Opitz, 2013).
For these processes, the forms $Y_i(x)$ as are listed below (eg. Dey et al., 2015).

For the Smith Process
\[ Y_i(x) = W(x - U_i; 0, \Sigma), \quad x \in \mathcal{X}, \] 
where $\mathcal{X} \in \mathbb{R}^d$, $W(\cdot; 0, \Sigma)$ is the $d$-dimensional Gaussian density, with mean zero and covariance matrix $\Sigma$, and $U_i$ are points of a homogeneous Poisson Process defined on $\mathbb{R}^d$.

For the Schlather Process
\[ Y_i(x) = \sqrt{2\pi} \max\{0, W_i(x)\}, \quad x \in \mathcal{X}, \] 
where $\{W_i(x) : x \in \mathcal{X}\}$ are independent copies of a stationary standard Gaussian process with correlation function $\rho(h)$. The scaling factor of $\sqrt{2\pi}$ therefore ensures that $E[Y_i(x)] = 1$.

For the Brown–Resnick Process
\[ Y_i(x) = \exp\{W_i(x) - \gamma(x)\}, \quad x \in \mathcal{X}, \] 
where $\{W_i(x) : x \in \mathcal{X}\}$ are independent copies of a zero mean Gaussian process with stationary increments and semi-variogram $\gamma(h) = \frac{1}{2} \text{Var}\{W(x + h) - W(x)\}$, with $W(0) = 0$ almost surely.

For the extremal-$t$ process
\[ Y_i(x) = g(\nu) \max\{0, W_i(x)\}^\nu, \quad x \in \mathcal{X}, \] 
where $\nu \geq 1$, $g(\nu) = \sqrt{\pi} 2^{-(\nu-2)/2} \Gamma((\nu-1)/2)^{-1}$, $\Gamma$ is the gamma function and $\{W_i(x) : x \in \mathcal{X}\}$ are independent copies of a stationary standard Gaussian process with correlation function $\rho(h)$.

The Smith process, or Gaussian extreme value process, was the first max-stable process to appear in the literature and to be used to model rainfall extremes. This max-stable process provided the rainfall-storm interpretation. The Smith process arises as a special case of the Brown–Resnick process and the Schlather process arises a special case of the extremal-$t$ process.

Simulated examples from Smith, Schlather, Brown–Resnick and Extremal-$t$ process is given in Figure 5.1. These realisations were obtained by fitting the processes to daily annual maxima data from South East Queensland, Australia and then simulating. The fitting and simulations were performed using the R packages SpatialExtremes (Ribatet, 2015) and RandomFields (Schlather, ...
This figure serves to highlight the differences in the dependence structures between each of the max-stable processes. We observe that the extremal precipitation field produced by the Smith process is too smooth to reliably capture the physical behaviour of extreme rainfall in this region. Therefore although the Smith process is historically significant, the practical use of this model is limited due to the smoothness of the rainfall fields produced (Dey et al., 2015).

![Example Simulations](image)

Figure 5.1: Simulations from each of the max-stable processes given in Section 5.3.3

### 5.3.4 Finite–Dimensional Distribution

The finite dimensional extreme value distribution can be derived from the spectral representation of equation (5.16). Let \( \Phi = \{ (\zeta_i, Y_i) \}_{i \geq 1} \) be a Poisson process, where \( \zeta_i \) and \( Y_i \) are as defined in Section 5.3.2. For finite-dimensional vectors, \( x \subset \mathcal{X}^k \) and \( z \subset (0, \infty)^k \), where \( k \geq 1 \), the finite \( k \)-dimensional
distribution is then
\[ P[Z(x) \leq z] = P[\mathbb{H}(\zeta, Y) \in \Phi : \zeta Y(x_j) > z_j \text{ for some } j \in \{1, \ldots, k\}] \] (5.23)
\[ = \exp \left[ - \int_0^\infty P \left( \zeta > \min_{j=1}^k z_j \frac{Y(x_j)}{z_j} \right) \zeta^{-2} d\zeta \right]. \] (5.24)

If the univariate marginal distributions of the max-stable process are unit Fréchet, the \(k\)-dimensional distribution can then be written as
\[ P[Z(x) \leq z] = \exp \{-V(z)\}, \] (5.25)
where \(V(z)\) is the exponent measure (Resnick, 1987) and
\[ V(z) = \mathbb{E} \left[ k \max_{j=1}^k \frac{Y(x_j)}{z_j} \right]. \] (5.26)

Here \(V\) is an homogeneous function of order 1,
\[ V(tz_1, \ldots, tz_k) = t^{-1}V(z_1, \ldots, z_k) = t^{-1}V(z), \quad t > 0. \] (5.27)

First order homogeneity is necessary for the max-stability of the process. Additionally, given the process is a simple max-stable process,
\[ \lim_{z_j \to \infty, \forall j \neq i} V(z_1, \ldots, z_k) = \frac{1}{z_i} \mathbb{E}[Y(x_i)] = \frac{1}{z_i}. \] (5.28)

5.3.5 Extremal Coefficient

To understand the dependence within a max-stable process, summary measures for extremal dependence are useful. To this end, we introduce the extremal coefficient.

The extremal coefficient, \(\theta(x)\) arises as a special case of equation (5.25), where \(z_i = z\) for all \(i = 1, \ldots, k\),
\[ V(z1) = V(z, \ldots, z) = \frac{V(1, \ldots, 1)}{z} = \frac{\theta(x)}{z}, \] (5.29)
and
\[ \theta(x) = \mathbb{E} \left[ \max_{j=1}^k Y(x_j) \right]. \] (5.30)

Given that the radial and angular components of the spectral representation (5.16) are separable, the extremal coefficient provides a useful indication of the partial dependence between maxima within the max-stable process, independent of the radial component.
MAX-STABLE PROCESSES

For spatial applications it is often convenient to express the extremal coefficient in terms of the distance between two locations. For all \( x_1 \in \mathcal{X} \) and \( x_2 \in \mathcal{X} \), and with \( h = x_2 - x_1 \), the extremal coefficient is

\[
\theta(h) = -\log [\mathbb{P}(Z(x_1) \leq z, Z(x_2) \leq z)].
\]

The function \( \theta(h) \) takes values in the interval \([1, 2]\), with the ends of the interval corresponding respectively to complete dependence and independence between \( Z(x_1) \) and \( Z(x_2) \).

The extremal coefficient helps to highlight the differences between the common max-stable processes presented in Section 5.3.3. For the Schlather model, independence cannot be achieved and the maximum value that \( \theta(h) \) can attain is approximately 1.7 in the limit as the distance between \( x_1 \) and \( x_2 \) grows large. For the Brown–Resnick model, as this distance approaches \( \infty \), \( \theta(h) \) approaches 2 and independence is achieved. The introduction of the parameter \( \nu \) allows the extremal-\( t \) process to model a wider range of dependence behaviour. In the extremal-\( t \) process as the distance between \( x_2 \) and \( x_1 \) approaches \( 1 \), \( \theta(h) \) approaches \( 2^{t+1}(\nu + 1) \), where \( t_{\nu+1} \) is the distribution function for the univariate \( t \)-distribution with \( \nu + 1 \) degrees of freedom. Independence can therefore be achieved in the limit as \( \nu \) grows large.

5.4 Inference

Although max-stable processes have existed in the literature for decades (R. L. Smith, 1990), historically the challenge to their implementation for practical applications has been that the intractability of the likelihood. Consider the density of the finite-dimensional multivariate extreme value distribution. This is given by the derivative of equation (5.25)

\[
f(z) = \exp \{-V(z)\} \sum_{\tau \in \mathcal{P}_k} w(\tau),
\]

where

\[
w(\tau) = (-1)^{|\tau|} \prod_{j=1}^{|\tau|} \frac{\partial^{\tau_j}}{\partial z^{\tau_j}} V(z)
\]

and \( \mathcal{P}_k \) is the set of all partitions of the set \( \mathbf{x} = (x_1, \ldots, x_k) \). \( \tau = (\tau_1, \ldots, \tau_l) \) denotes a specific partition of size \( l \), and \( \partial^{\tau_j}/\partial z^{\tau_j} \) denotes the mixed partial derivative with respect to the \( j \)th element of that partition \( \tau \). For further detail refer to (Dombry, Éyi-Minko, and Ribatet, 2012). The number of possible partitions of the set \( \mathbf{x} \), \( |\mathcal{P}_k| \), corresponds to the Bell numbers. The first ten Bell numbers are given in Table 5.1. As \( k \) increases, a combinatorial explosion occurs for the number of partitions. Given this, computational optimisation of the likelihood is infeasible for large \( k \).
Table 5.1: The first 10 Bell numbers, $B_k$, where $k$ gives the size of the set and $B_k$ gives the number of possible partitions for a set of that size.

<table>
<thead>
<tr>
<th>$B_0$</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
<th>$B_7$</th>
<th>$B_8$</th>
<th>$B_9$</th>
<th>$B_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>15</td>
<td>52</td>
<td>203</td>
<td>877</td>
<td>4140</td>
<td>21147</td>
<td>115975</td>
</tr>
</tbody>
</table>

5.4.1 Pairwise Likelihood

Given the infeasibility of full likelihood inference for max-stable processes, statistical inference is commonly performed using composite likelihood (S. A. Padoan, Ribatet, and Sisson, 2010). Compared with maximum likelihood methods, the use of composite likelihood (Varin, Reid, et al., 2011) avoids the need for the full density by using the bivariate density and summing over pairs of observations. Therefore in this instance, composite likelihood may also be referred to as pairwise likelihood. The log of the composite likelihood equation is given by

$$
\ell(\vartheta; z) = \sum_{i<j} \sum_{n=1}^{N_{ij}} \log f(x_i^{(n)}, x_j^{(n)}; \vartheta),
$$

(5.34)

where $f$ is the bivariate extreme value density, $N_{ij}$ is the number of shared observations at $x_i$ and $x_j$, and $\vartheta$ is the parameter vector to be estimated by maximising equation (5.34). The bivariate distributions for the common max-stable processes given in Section 5.3.3 can be found in (Dey et al., 2015).

In contrast to standard maximum likelihood, the composite likelihood takes the form of a misspecified likelihood (White, 1982). This is because the sum is taken over the bivariate distribution, and observational pairs at different pairs of locations are not independent.

The properties of composite likelihood estimators, under the appropriate regularity conditions, are similar to those of maximum likelihood estimators: the estimating equation are unbiased and the estimators are asymptotically normal. However, the variance matrix of the composite likelihood is estimated by the inverse the sandwich information matrix, also known as the Godambe information matrix (Godambe, 1960),

$$
G(\vartheta) = H(\vartheta)K(\vartheta)^{-1}H(\vartheta),
$$

(5.35)

where the parameters estimated using composite likelihood, $\hat{\vartheta}$, are normally distributed,

$$
\hat{\vartheta} \sim N(\vartheta, G(\vartheta)^{-1}).
$$

(5.36)

In equation (5.35), $H(\vartheta)$ and $K(\vartheta)$ are analogues of the observed information matrix and covariance matrix of the score vector (eg. Varin, Reid, et al.,
If the composite likelihood were a true likelihood, then \( H(\vartheta) = K(\vartheta) \), and \( G(\vartheta) \) would be equal to the Fisher information matrix. When this is not the case, using the sandwich information matrix ensures that confidence intervals are adjusted so the variability of composite likelihood estimators is not underestimated.

Although the mathematical arguments presented thus far have been in terms of a simple max-stable process, composite likelihood can be used to fit max-stable processes with GEV marginals, \( Z(x) \sim \text{GEV}(\mu(x), \sigma(x), \xi(x)) \). Without loss of generality, the GEV marginal distributions can be transformed to standard Fréchet marginal distributions using the transform, \( t \), given in equation (5.15). Let \( \{Z(x): x \in X\} \) be a max-stable process with marginals, \( Z(x) \sim \text{GEV}(\mu(x), \sigma(x), \xi(x)) \), and let \( \{\tilde{Z}(x): x \in X\} \) be the transformed simple max-stable process with marginals \( \tilde{Z}(x) \sim \text{GEV}(1, 1, 1) \). The form of the composite log likelihood is then given by the decomposition

\[
\ell(\vartheta; z) = \ell(\vartheta; t(\tilde{z})) + \sum_{i<j} \sum_{n=1}^{N_{ij}} \log |J\{z(x_i)\} J\{z(x_j)\}|,
\]

where the additional summation term is from the Jacobian matrix, \( J \), associated with the transformation of marginal distributions.

In using composite likelihood, compared with standard maximum likelihood, there is a trade-off between computational efficiency and the number of pairs selected. Given this, it can be useful to include weights in the composite likelihood equation (5.34). One popular method for weight selection omits pairs that are separated by large distances (Bevilacqua et al., 2012). The maxima of these pairs is assumed to be approximately independent given the large separation, and as such little information can be contributed to the likelihood that can improve the estimation of the dependence parameters. It is also possible to use other multivariate densities, instead of the bivariate density, in the composite likelihood equation (5.34). However, for trivariate densities, the efficiency gain versus computational cost has been found to be small (Genton, Ma, and Sang, 2011; Huser and Davison, 2013).

Inference using the full likelihood in equation (5.32) is possible, based on work on conditional max-stable distributions by (Dombry, Eyi-Minko, et al., 2013; Dombry, Éyi-Minko, et al., 2012). However, given the combinatorial explosion that occurs with the number of partitions, and the limitations imposed by current technology, full likelihood inference is limited to a density, \( f(x) \), where the length of \( x \) is small. In the method presented by Castruccio, Huser, and Genton, (2016), this dimension was of order 12 to 13. In Dombry, Genton, et al., (2017), for the Brown-Resnick process a dimension of 20 was successfully fitted. We require higher dimensions for the applications con-
5.5. MAX-STABLE SIMULATION

sidered in this thesis. Given this, we use composite likelihood for parameter estimation.

5.4.2 Model Selection

For model selection, an analogue to the AIC in equation (3.39) exists for composite likelihood (Varin, 2008; Varin and Vidoni, 2005). Let \( \theta \) be the parameter set; then the information criterion is given by

\[
TIC(\hat{\theta}) = -2\ell(\hat{\theta}) - \text{dim}(\hat{\theta}),
\]

where \( \text{dim}(\hat{\theta}) = \text{tr}\{H(\hat{\theta})K(\hat{\theta})^{-1}\} \). This information criterion is a version of the Takeuchi Information Criterion (TIC) (Takeuchi, 1976) using a penalty term appropriate for composite likelihood. If the penalty term is changed to \( n\text{dim}(\theta) \), where \( n \) is the number of observations, then the Information Criterion is an analogue of the BIC in equation (3.39).

As the composite likelihood takes the form of a misspecified likelihood, the likelihood ratio statistic is not chi-square distributed. Therefore, unlike univariate methods discussed in Chapter 3, a hypothesis test similar to equation (3.38) cannot be directly applied for selection between nested models. Methodology proposed in (Chandler and Bate, 2007; Rotnitzky and Jewell, 1990) can be used to adjust the distribution of the composite likelihood ratio statistic. This allows for comparison of nested composite likelihood models using an analysis of variance.

5.5 Max-stable Simulation

Simulation is a key tool when analysing max-stable processes. Simulation provides a means by which the intractable distribution of random variables can be estimated. These random variables may include the risk measures mentioned in Section 5.1, such as the largest rainfall in a region or the total maximum rainfall in a region. Simulation from max-stable processes however, is not necessarily straightforward. A max-stable process is obtained in the limit as maximum over a series of independent replicates of a stochastic process approaches infinity, and it is not possible to simulate infinitely many processes. However, in some instances exact simulation is possible. Alternatively, it is possible to simulate from the spectral representation until a stopping criterion is met. This ensures only a finite number of iterations is required. However, as with many simulation methods a balance is often struck between simulation accuracy and computational efficiency. In this section, a brief introduction is provided to max-stable process simulation. An overview of the methods introduced in this chapter can be found in the textbook, Dey et al., (2015).
5.5.1 Unconditional Simulation

Schlather, (2002) introduced a method for the exact simulation of max-stable processes, under appropriate conditions. This method provides the foundation for many modern simulation methods. Consider the spectral representation of equation (5.18), where $\nu$ is a probability measure. The inversion of the inhomogeneous Poisson process, $\{\zeta_i\}_{i \geq 1}$ on $(0, \infty)$ with measure $\zeta^{-2}d\zeta$, is a homogeneous Poisson process, $\{\zeta_i^{-1}\}_{i \geq 1}$ with rate, $\lambda = 1$. The interarrival times, $\zeta_1^{-1}, \zeta_2^{-1} - \zeta_1^{-1}, \zeta_3^{-1} - \zeta_2^{-1} \ldots$ are therefore independent and identically distributed according to a standard exponential distribution. Let $E_j$ be the $j$th interarrival time, where $E_j \sim \text{Exp}(1)$ and $j \geq 1$. The spectral representation of (5.18) can therefore be rewritten as

$$Z(x) \overset{d}{=} \max_{i \geq 1} \frac{Y_i(x)}{\sum_{j=1}^{i} E_j}, \quad x \in \mathcal{X}. \tag{5.39}$$

It is possible to simulate exactly from the max-stable process in the case where $Y(x)$ is bounded above almost surely by a constant $C$,

$$\sup_{x \in \mathcal{X}} Y_i(x) \leq C \quad \text{a.s.} \tag{5.40}$$

This follows because the factors, $\left(\sum_{j=1}^{i} E_j\right)^{-1}$ are monotonically decreasing, so there must exist a $k$ such that any terms after $Y_k(x) \left(\sum_{j=1}^{k} E_j\right)^{-1}$ cannot contribute to the the pointwise maxima of the spectral representation,

$$\frac{C}{\sum_{j=1}^{k} E_j} \leq \inf_{x \in \mathcal{X}} \left(\max_{i \geq 1} \frac{Y_i(x)}{\sum_{j=1}^{i} E_j}\right). \tag{5.41}$$

Therefore only finitely many realisations of the spectral representation are required for exact simulation.

**Theorem 5.5.1.** (From Dey et al., (2015) Proposition 10.2.1 and cf. to Schlather, (2002) Theorem 4)

Let $\{E_j\}_{j \geq 1}$ be i.i.d. standard exponential random variables, and independently of $\{E_j\}_{j \geq 1}$, let $\{Y_i\}_{i \geq 1}$ be independent copies of a stochastic process bounded above a.s. by a constant $C > 0$. Then

$$\max_{i=1}^{N} \frac{Y_i(x)}{\sum_{j=1}^{i} E_j} = \max_{i \geq 1} \frac{Y_i(x)}{\sum_{j=1}^{i} E_j}, \quad x \in \mathcal{X}, \tag{5.42}$$

where

$$N = \min \left\{ k : \frac{C}{\sum_{j=1}^{k+1} E_j} \leq \inf_{x \in \mathcal{X}} \left(\max_{i=1}^{k} \frac{Y_i(x)}{\sum_{j=1}^{i} E_j}\right) \right\}, \tag{5.43}$$

and it follows that the process $Z(\cdot)$ given by equation (5.18) satisfies

$$Z(\cdot) \overset{d}{=} \max_{i=1}^{N} \left\{ \left(\sum_{j=1}^{i} E_j\right)^{-1} Y_i(\cdot) \right\}. $$
5.5. MAX-STABLE SIMULATION

In practice, the existence of an upper bound (5.40) may not be satisfied for a given \( Y(x) \). However, an approximation can be made by choosing \( C \) such that \( \mathbb{P}(\sup_{x \in \mathcal{X}} Y(x) > C) \) is acceptably small. This results in a compromise between simulation accuracy and computational time, as the larger the value of \( C \) the longer the simulation takes to converge.

Additionally, the spectral representation of a max-stable process is not unique. It is possible to choose a different but stochastically equivalent representation that improves the efficiency of the simulation algorithm (see Corollary 9.4.5 in De Haan and Ferreira, (2006)). A class of transformations for the spectral representation is given in Oesting, Schlather, and Zhou, (2013), often referred to as the normalized spectral representation.

For some max-stable processes, for example the Smith process (equation (5.19)), the simulation methods described in the above section are not suitable. These processes are known as mixed moving maxima processes. The parametric form of the spectral representation (see equation 5.18) for mixed moving maxima processes is

\[
Z(x) = \zeta_i Y_i(x - U_i), \quad x \in \mathcal{X}.
\]  

(5.44)

Here \( \{(\zeta_i, U_i)\}_{i \geq 1} \) are points of a Poisson process on \((0, \infty) \times \mathbb{R}^d\) with intensity measure \( \zeta^{-2} d\zeta \times ds \) and \( \{Y_i(x)\}_{i \geq 1} \) is a sequence of independent copies of a non-negative stochastic process \( \{Y(x) : x \in \mathcal{X}\} \) with continuous sample paths such that the \( \mathbb{E}\{Y(x)\} = 1 \) for all \( x \in \mathcal{X} \). The points \( \{U_i\}_{i \geq 1} \) can be regarded as random translations of the function \( Y(x) \). However, the result of these random translations is that an adjustment is needed to Theorem 5.5.2.

Consider if the function \( Y(x) \) had compact support, such that

\[
\mathbb{P}\left[Y(x) = 0, \forall x \in \mathbb{R}^d \setminus b(o, R)\right] = 1,
\]  

(5.45)

where \( R > 0, b(o, R) \) is a \( d \)-dimensional ball around the origin with radius \( R \). The product \( \zeta_i Y_i(x - U_i) \) cannot contribute to the pointwise maxima of (5.44) if

\[
U_i \notin \mathcal{X} \oplus b(o, R),
\]

where \( A \oplus B = \{a + b : a \in A, b \in B\} \) for \( A, B \in \mathbb{R}^d \). Therefore instead of considering all points of the Poisson process \( \{(\zeta_i, U_i)\}_{i \geq 1} \), it is sufficient to consider the max-stable process \( Z(x) \) restricted to the set of Poisson points in \((0, \infty) \times (\mathcal{X}, \mathbb{R}^d \oplus b(o, R))\). Care needs to be taken to ensure that the measure on the restricted Poisson process is the same. Therefore it is important to rescale the points by \(|\mathcal{X} \oplus b(o, R)|\), where \(|\cdot|\) is the Lebesgue measure of the set. Then it is possible to simulate the mixed moving maxima process using only finitely many realisations of the spectral representation.

**Theorem 5.5.2.** (From Dey et al., (2015) and cf. Schlather, (2002) Theorem 4)
Let \( \{Y_i\}_{i \geq 1} \) be independent copies of a stationary stochastic process that satisfies equation (5.45) for some \( R > 0 \) and

\[
\sup_{x \in \mathbb{R}^d} Y(x) \leq C \quad \text{a.s.,} \tag{5.46}
\]

where \( C > 0 \) is a constant. Independently of \( \{Y_i\}_{i \geq 1} \), let \( \{E_j\}_{j \geq 1} \) be i.i.d. standard exponential random variables and let \( \{U_i\}_{i \geq 1} \) be an independent sequence of uniformly distributed random variables on \( \mathcal{X} \oplus b(o, R) \), where \( b(o, R) \) is a \( d \)-dimensional ball around the origin with radius \( R \). Then the mixed moving maxima process of equation (5.44), \( \{Z(x) : x \in \mathcal{X}\} \), satisfies

\[
Z(\cdot) \overset{d}{=} \max_{i=1}^{N} \frac{\mathcal{X} \oplus b(o, R)}{\sum_{j=1}^{k+1} E_j} Y_i(\cdot - U_i) \tag{5.47}
\]

for an a.s finite number, \( N \), defined by

\[
N = \min \left\{ k : \frac{C}{\sum_{j=1}^{k} E_j} \leq \inf_{x \in \mathcal{X}} \left( \frac{k}{\sum_{j=1}^{k} E_j} F_i(x - U_i) \right) \right\}. \tag{5.48}
\]

It is possible to select an \( R \) and \( C \) such that \( \mathbb{P}(\sup_{x \in \mathcal{X}} Y(x) > C) \) is acceptably small or so that the function \( Y \) is negligibly small outside \( b(o, R) \). This can improve the speed and may be necessary if equation (5.45) is not satisfied. Again, this creates a trade-off between computational efficiency and simulation accuracy. Examples of unconditional simulations are visualised in Figure 5.1.

### 5.5.2 Conditional Simulation

In addition to the simulation methods presented so far, it can also be useful to simulate from the max-stable process conditional on fixed information. For example, it is useful to understand the probability of extreme events occurring in a region, given that an extreme event has been observed at a nearby location. In order to answer such questions, conditional simulations are needed. This is due to the intractability of finite-dimensional distributions in higher dimensions (see Section 5.3.4).

Conditional simulation arises from the mathematics of point processes and their conditional distributions. The spectral representation in equation (5.18) can be rewritten as a Poisson point process, \( \Phi = \{\phi_i\}_{i \geq 1} \), where \( \phi_i(x) = \zeta_i Y_i(x) \). For a Borel set \( A \subset \mathcal{X} \), the intensity measure of the point process, \( \Phi \), defined on \( \mathbb{R}^d \) is given by

\[
\Lambda_x(A) = \int_0^\infty \mathbb{P}[\zeta Y \in A] \zeta^{-2} d\zeta, \tag{5.49}
\]
5.5. MAX-STABLE SIMULATION

(eg. Dey et al., 2015).

The point process \( \Phi \) is defined to be regular if the intensity measure \( \Lambda_x \) has an intensity function \( \lambda_x \) such that \( \Lambda_x(dx) = \lambda_x(z)dz \), \( \forall x \in X \). Let \( x = \{x_i\}_{1 \leq i \leq k} \) be a set of conditioning points and \( s = \{s_j\}_{1 \leq j \leq m} \) be the unconstrained set of points at which the max-stable process is to be simulated. If \( \Phi \) is regular, it follows that the conditional intensity function is given by

\[
\lambda_{s|x,z}(u) = \frac{\lambda_{s,x}(u,z)}{\lambda_x(z)},
\]

(5.50)

where \( (s,x) \in X^{m+k} \); \( u \in (0, \infty)^m \) and \( z \in (0, \infty)^k \). Conditional simulation of \( Z(s) \) is therefore subject to the constraint

\[
Z(x_1) = z_1, Z(x_2) = z_2, \ldots, Z(x_k) = z_k.
\]

(5.51)

The following three definitions introduce the terminology needed for conditional simulations.

**Definition 5.5.3. (Extremal Functions)** There exists almost surely a unique function \( \phi \in \Phi \), such that for a point \( x_j \), \( \max_{i \geq 1} \phi_i(x_j) = \phi(x_j) = z_j \). This unique function is called an extremal function and is denoted by \( \varphi_{x_i}^+ \). The set of all extremal functions associated with the conditioning points, \( \{x_i\}_{1 \leq i \leq k} \), is known as the extremal point process, \( \Phi^+ = \{\varphi_{x_i}^+\}_{1 \leq i \leq k} \).

It is important to note that elements of \( \Phi^+ \) are not necessarily unique. It is possible for \( \varphi_{x_i}^+ = \varphi_{x_i}^+ \) where \( x_{i_1} \neq x_{i_2} \). Given this, the conditioning points can be partitioned into sets of points that share the same extremal function. For example, if the conditioning points are \( \{x_1, x_2, x_3\} \) and the associated extremal functions are \( \{\varphi_{x_1}^+, \varphi_{x_2}^+, \varphi_{x_3}^+\} \), if \( \varphi_{x_1}^+ = \varphi_{x_2}^+ \) and \( \varphi_{x_1}^+ \neq \varphi_{x_3}^+ \), then the partition induced by the extremal functions is \( \{\{x_1, x_2\}, \{x_3\}\} \). This partition is known as a hitting scenario (A. Stephenson and Tawn, 2005; Y. Wang and Stoew, 2011).

**Definition 5.5.4. (Hitting Scenario)** The hitting scenario, \( \tau = (\tau_1, \tau_2, \ldots, \tau_l) \), is a partition of the conditioning points \( \{x_i\}_{1 \leq i \leq k} \). Any two conditioning points, \( x_{i_1} \) and \( x_{i_2} \) with \( i_1 \neq i_2 \), are part of the same set, \( \{x_{i_1}, x_{i_2}\} \subset \tau_j \), if the extremal function associated with those points is equal, \( \varphi_{x_{i_1}}^+ = \varphi_{x_{i_2}}^+ \).

The final definition necessary for conditional simulation relates to the elements of \( \Phi \) that are not extremal functions.

**Definition 5.5.5. (Subextremal Functions)** A function \( \phi \in \Phi \) is a subextremal function if \( \phi(x_i) < z_i \) for all \( 1 \leq i \leq k \). The set of all subextremal functions forms a subextremal point process, \( \Phi^- \).
The point process $\Phi$ can therefore be decomposed into its extremal and subextremal functions, such that $\Phi = \Phi^+ \cup \Phi^-$. 

The following theorem can now be introduced for conditional simulation. 

**Theorem 5.5.6.** (Dombry, Ėyi-Minko, et al., 2012) Suppose that the point process $\Phi$ is regular and let $(x, s) \in \mathcal{X}^{k+m}$. For $\tau = (\tau_1, \ldots, \tau_\ell) \in \mathcal{P}_k$ and $j = 1, \ldots, \ell$, define $I_j = \{i : x_i \in \tau_j\}$, $x_{\tau_j} = (x_i)_{i \in I_j}$, $z_{\tau_j} = (z_i)_{i \in I_j}$, $x_{\tau_j} = (x_i)_{i \notin I_j}$ and $z_{\tau_j} = (z_i)_{i \notin I_j}$. Consider the following three-step procedure. 

**Step 1:** Draw a random partition $\tau^* \in \mathcal{P}_k$ with conditional distribution

$$\pi_x(z, \tau) = \mathbb{P}\{\tau^* = \tau \mid Z(x) = z\}, \quad (5.52)$$

$$= \frac{1}{C(x, z)} \prod_{j=1}^{\tau} \lambda_{x_{\tau_j}}(z_{\tau_j}) \int_{\{u_j < z_{\tau_j}\}} \lambda_{x_{\tau_j}}(u_j) du_j, \quad (5.53)$$

where the normalizing constant is

$$C(x, z) = \sum_{\tau \in \mathcal{P}_k} \prod_{j=1}^{\tau} \lambda_{x_{\tau_j}}(z_{\tau_j}) \int_{\{u_j < z_{\tau_j}\}} \lambda_{x_{\tau_j}}(u_j) du_j \quad (5.54)$$

**Step 2:** Given $\tau = (\tau_1, \ldots, \tau_\ell)$, draw $\ell$ independent random vectors $\varphi^j_+(s), \ldots, \varphi^j_+(s)$ from the distribution

$$\mathbb{P}\{\varphi^j_+(s) \in dv \mid Z(x) = z, \tau^* = \tau\} = \frac{1}{C_j} \left\{ \int [u < z_{\tau_j}] \lambda_{(s,x_{\tau_j})}(v, u) du \right\} dv, \quad (5.55)$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function and

$$C_j = \int [u < z_{\tau_j}] \lambda_{(s,x_{\tau_j})}(v, u) dv, \quad (5.56)$$

and define the random vector $Z^+(s) = \max_{j=1, \ldots, \ell} \varphi^j_+(s)$.

**Step 3:** Independently draw a realization of a Poisson point process $\{\zeta_i\}_{i \geq 1}$ on $(0, \infty)$ with intensity $\zeta^{-2} d\zeta$ and $\{Y_i\}_{i \geq 1}$ independent copies of $Y$, and define the random vector $Z^-(s) = \max_{i \geq 1} \zeta_i Y_i(s) \mathbb{I}_{\{\zeta_i < z\}}$. 

Then the random vector $\tilde{Z}(s) = \max\{Z^+(s), Z^-(s)\}$ has the conditional distribution of $Z(s)$ given $Z(x) = z$. 

80 MAX-STABLE PROCESSES
5.6. ALTERNATIVE METHODS

The corresponding conditional cumulative distribution function is

\[
P \{ Z(s) \leq a \mid Z(x) = z \} = \left\{ \sum_{\tau \in \mathcal{P}_k} \pi_x(z, \tau) \prod_{j=1}^{\lvert \tau \rvert} F_{\tau,j}(a) \right\} \frac{P \{ Z(s) \leq a, Z(x) \leq z \}}{P \{ Z(x) \leq z \}},
\]

where

\[
F_{\tau,j}(a) = P \{ \varphi^+_{j}(s) \leq a \mid Z(x) = z, \tau^* = \tau \},
\]

\[
= \frac{\int_{\{u \leq z_{\tau^*}^{(j)}, v < a\}} \lambda(s, x_{\tau^*}^{(j)}|x_{\tau^*}^{(j)}, z_{\tau^*}^{(j)}(v, u)) \, du \, dv}{\int_{\{u \leq z_{\tau^*}^{(j)}\}} \lambda_x(z_{\tau^*}^{(j)}(u)) \, du}.
\]

The essence of steps 1–3 is: draw a partition, simulate the extremal functions and then simulate the subextremal functions. Here, step 3 is relatively simple and can be performed by unconditional simulation from a relevant max-stable process, and rejecting all simulations that do not satisfy the conditioning constraint. However, there is some nuance to sampling the random partition of \( \tau^* \in \mathcal{P}_k \) in step 1.

We saw in Section (5.3.4) that as the size of the set of conditional points grows large, a combinatorial explosion occurs for the number of possible partitions, \( \mathcal{P}_k \), making the normalising constant of step 1 all but impossible to calculate. Consequently, a Gibbs sampler (Geman and Geman, 1984) is often used to draw a random partition of Step 1 when \( |x| \) is too large. For further detail on the derivation of Theorem 5.5.6, and its computational implementation, the reader is referred to the papers (Dombry, Eyi-Minko, et al., 2013; Dombry, Éyi-Minko, et al., 2012).

5.5.3 Recent Methods

The simulation methods detailed in this chapter are at the core of the simulation methods implemented in the SpatialExtremes package in R (Ribatet, 2013, 2015) and used in this thesis. However, the literature on max-stable process simulation is currently developing rapidly. New methods for exact simulation of max-stable processes are available (Dieker and Mikosch, 2015; Dombry, Engelke, and Oesting, 2016; Oesting, Schlather, Zhou, et al., 2018). Depending on the type of max-stable process used and the application, these methods offer improvements both in simulation accuracy and simulation efficiency.

5.6 Alternative Methods

Max-stable processes are not the only statistical framework within which the dependence of spatial extremes can be modelled. Other approaches include
latent variable models and copula based approaches. These methods are each subject to different modelling assumptions. Depending upon the application and the required inference, each of these methods has different strengths and weaknesses.

5.6.1 Latent variable models

For latent variable models, the marginal distributions of a stochastic process, $Y(x)$, conditional on an unobserved latent process, are assumed to follow an extreme value distribution. If the extreme value distribution is a GEV distribution then

$$Y(x)|\{\mu(x), \sigma(x), \xi(x)\} \sim GEV(\mu(x), \sigma(x), \xi(x)), \ x \in \mathcal{X}. \quad (5.60)$$

It follows that for $x_1 \neq x_2$, $Y(x_1)$ is conditionally independent of $Y(x_2)$. In practice, the latent process, $\{\mu(x), \sigma(x), \xi(x)\}$, is often taken to be a Gaussian process. A simple example of a possible latent process is

$$\mu(x) = \beta^T \mu X + \epsilon_\mu(x, \alpha_\mu), \quad (5.61)$$
$$\sigma(x) = \beta^T \sigma X + \epsilon_\sigma(x, \alpha_\sigma), \quad (5.62)$$
$$\xi(x) = \beta^T \xi X + \epsilon_\xi(x, \alpha_\xi), \quad (5.63)$$

where $\beta$ are regression parameters and $\epsilon$ is a Gaussian process with an appropriate covariance function having parameters $\alpha$, (for common covariance functions see Section 5.3.3). Here $\epsilon_\mu$, $\epsilon_\sigma$ and $\epsilon_\xi$ are assumed to be independent, although this assumption can be relaxed (Sang and Gelfand, 2010).

Fitting latent variable models is normally done in a Bayesian setting, as the likelihood of the model involves intractable integrals. This requires us to specify appropriate priors and solve for the posterior distributions using Markov chain Monte Carlo (MCMC) (for more details on the MCMC algorithms see Davison, S. Padoan, et al., (2012)). Examples of hierarchical models with latent variables for environmental applications can be found in (Casson and Coles, 1999; Turkman, Turkman, and Pereira, 2010), with (Cooley, Nychka, and Naveau, 2007; Sang and Gelfand, 2009) demonstrating applications to extreme rainfall data.

Latent variable models are ideal for accurately modelling the univariate extreme value parameters in space, particularly when the response surface is complex. However, there are two main drawbacks to this method. First, the process $Y(x)$ is not a multivariate extreme value distribution. Second, the assumption of conditional independence given the latent process may not be realistic (Davison, S. Padoan, et al., 2012). Therefore these models are less well suited for applications if it is important to model the dependence explicitly.
5.6. ALTERNATIVE METHODS

5.6.2 Copulas

An alternative method for modelling the dependence between spatial extremes involves using copulas. Let \( Y_1, \ldots, Y_k \) be a set of random variables where \( F_i \) is the univariate marginal distribution for \( Y_i \). By Sklar’s Theorem (e.g. Nelsen, 2007), the joint distribution \( F(Y_1, \ldots, Y_k) \) can then be written in terms of a copula, \( C \),

\[
F(y_1, \ldots y_k) = C(F_1(y_1), \ldots, F_k(y_k)), \tag{5.64}
\]

where \( C(\cdot) \) is a \( k \)-dimensional function defined on \([0, 1]^k\), with \( k \geq 1 \). The copula, \( C \), is unique if the univariate margins of \( F \) are absolutely continuous. Additionally, if the univariate margins are continuous and strictly increasing then

\[
F(y_1, \ldots y_k) = C(u_1, \ldots, u_k), \tag{5.65}
\]

\[
C(u_1, \ldots, u_k) = F(F_1^{-1}(u_1), \ldots, F_k^{-1}(u_k)), \tag{5.66}
\]

where \( U_i \) is a uniform random variable such that \( U_i = F_i(Y_i) \) and \( Y_i = F_i^{-1}(U_i) \). Common copulas include elliptical copulas, such as the Gaussian or the student-t (e.g. Nelsen, 2007). Examples of their use for modelling rainfall extremes can be found in (Sang et al., 2010; Sun, Thyer, et al., 2014). Asymptotically however these copulas are not max-stable and therefore may fail to accurately capture the spatial dependence of extremes.

An exception can be made for extreme value copulas, which are max-stable (Gudendorf and Segers, 2010). Let \( X_i = (X_{i,1}, \ldots, X_{i,k}) \), where \( i = 1, \ldots, n \), be a sequence of iid random vectors with joint distribution function, \( F \), and continuous univariate marginals, \( F_1, \ldots, F_k \). By equation (5.64), a copula exists associated with the distribution \( F \). Let this copula be defined as \( C_F(\cdot) \). Further define \( M_n = (M_{n,1}, \ldots, M_{n,k}) \), where \( M_{n,j} = \max_{i=1}^n X_{i,j} \). It follows that the distribution of \( M_n \) is \( F^n \) and the distribution of \( M_{n,j} \) is \( F_j^n \). The copula \( C_n \) associated with the distribution for \( M_n \) therefore satisfies

\[
C_n(u_1, \ldots, u_k) = C_F(u_1^{1/n}, \ldots, u_k^{1/n})^n, \tag{5.67}
\]

for all \( u_1, \ldots, u_k \in [0, 1]^k \). As \( n \) approaches infinity, the copula \( C_n \) approaches an extreme value copula.

**Definition 5.6.1.** (Gudendorf et al., 2010) A copula \( C(\cdot) \) is an extreme-value copula if there exists a copula \( C_F(\cdot) \) such that for all \( u_1, \ldots, u_k \in [0, 1]^k \)

\[
C_F(u_1^{1/n}, \ldots, u_k^{1/n})^n \to C(u_1, \ldots, u_k), \quad \text{as} \quad n \to \infty. \tag{5.68}
\]

The Hülsler–Reiss copula is the limiting case for the bivariate Gaussian copula (Hüsler and Reiss, 1989) and the extremal-t copula is the limiting case
of the bivariate student-\(t\) copula (Nikoloulopoulos, Joe, and Li, 2009). A copula is an extreme value copula if and only if it is max-stable. An extreme value copula can also be related to the exponent measure of equation (5.25) by

\[
C(u_1, \ldots, u_k) = \exp \left\{ -V \left( \frac{-1}{\log u_1}, \ldots, \frac{-1}{\log u_k} \right) \right\}.
\] (5.69)

Applications of extremal copulas to modelling extreme rainfall are described further in (Davison, S. Padoan, et al., 2012; Reich and Shaby, 2012).

Copulas provide a possible alternative method for modelling the dependence of spatial extremes. Note however, we note that copulas do not provide full specification of the extremal process across continuous space, as is the case with max-stable processes.

5.7 Conclusions

This chapter provides the background necessary to model extremes in continuous space with dependence using max-stable processes. The content covered includes the spectral representation of max-stable processes, common parametric forms, methods for fitting, methods for model selection and methods for simulation. This content also provides the background needed to understand the applications of rainfall extremes presented in Chapters 6 and 7.
Chapter 6

The spatial distribution of rainfall extremes and the influence of El Niño Southern Oscillation

In the previous chapter, we provided some background concerning max-stable processes. In this chapter, we present an application of this theory for modelling rainfall extremes. The content and analyses presented are from the peer reviewed journal publication, Saunders et al., (2017). In adapting this paper to a chapter, we have tried to minimise any overlap with content in previous chapters, while still ensuring the continuity of the ideas presented.

6.1 Introduction

Extreme rainfall can have major societal impacts, including flash flooding, crop destruction, loss of life and infrastructure damage (see Section 1.2). To mitigate these potential consequences, an understanding is needed of the impact of extreme rainfall over a region, not just a single point location. In this chapter, we demonstrate that statistical models of the extreme precipitation field can be used to understand the risks and inform mitigation strategies. Moreover, we use our modelling to explore how the large scale climate drivers, such as El Niño Southern Oscillation (ENSO), can influence the spatial-distribution of extreme rainfall and consequently also the risk.

The development of spatial-temporal statistical models for extreme rainfall applications is a growing area of research (De Fondeville et al., 2016; Thibaud, Mutzner, and Davison, 2013; Zheng et al., 2015). However, within the climate and engineering literature the analysis still frequently
has a univariate focus (Jakob and Walland, 2016; Westra, L. V. Alexander, and Zwiers, 2013; Yilmaz, Imteaz, and Perera, 2017), in which spatial relationships and dependence are ignored. For improved urban planning and to ensure our communities are resilient against extreme rainfall, the analysis needs to shift away from univariate extremes. Max-stable processes provide the natural extension from univariate extreme value theory to dependent extremes in continuous space. Examples of their application for modelling precipitation extremes include summer maximal daily rainfall in Switzerland (Davison, S. Padoan, et al., 2012), autumnal maximal daily rainfall in France (Bechler et al., 2015) and drought in Rwanda, Africa (Oesting and Stein, 2017).

An advantage of using a max-stable model is that the dependence of spatial extremes is captured, and consequently it is possible to produce accurate simulations of an extreme rainfall field. Simulations are a key tool, providing aggregate measures of the extreme rainfall field, such as the total extreme rainfall or maximum rainfall in a region, are intractable analytically. If the primary concern were modelling the marginal distributions, there are simpler spatial models that we could use; examples include spatial GEV models (Buishand, 1991), kriging as applied to quantiles (Aryal et al., 2009) and latent variable models (Cooley, Nychka, et al., 2007) (see Section 5.6.1). These models leverage space to improve upon marginal parameter estimates and have more flexible methods for fitting. However, a consequence of this flexibility is that simplifying assumptions are made at the cost of capturing spatial dependence. Another alternative is provided by copula models (Sang et al., 2010; Sun, Thyer, et al., 2014) (see Section 5.6.2). These models are better at capturing spatial dependence than models discussed above, but they lack the max-stable property and consequent asymptotic justification. In this chapter, we show that max-stable models, despite their perceived complexity compared with univariate models or other spatial models, provide a useful tool for climate scientists and engineers to analyse rainfall extremes.

Max-stable processes can also be used to understand how climate driver-scan affect the extreme precipitation field. They have been used to examine the influence of ENSO on rainfall extremes in California, USA (Shang, Yan, and Zhang, 2011, 2015), Poyang Lake basin, China (Q. Zhang et al., 2014) and Eastern Australia (Westra and Sisson, 2011). Utilising spatial models to detect a significant influence of ENSO on the distribution of rainfall extremes is often needed, as in a univariate setting the temporal length of the observational record is often insufficient (Westra and Sisson, 2011).

In Australia, there has been little research into quantifying the influence of ENSO on the distribution of rainfall extremes (Kiem, Franks, and Kuczera, 2003; Min et al., 2013; Sun, Thyer, et al., 2014). Falsely assuming a stationary
climate or underestimating the influence of ENSO may therefore result in inadequate design standards, particularly for a strong La Niña phase. The La Niña phase of ENSO is known to be associated with cooler temperatures and wetter conditions along the Eastern Australian coast (Risbey et al., 2009). This is in contrast to the El Niño phase of ENSO which is associated with hotter and drier conditions. In Eastern Australia, the influence of ENSO on rainfall totals and extremes is also suspected to be asymmetric; in the sense that La Niña has a greater impact on the distribution of rainfall extremes than El Niño (W. Cai, Van Rensch, et al., 2010; King et al., 2013a; Sun, Thyer, et al., 2014).

Motivated by wanting to understand how ENSO can affect the extreme precipitation field, we have selected a case study region of South East Queensland (SEQ) (Figure 6.1). Historically the SEQ region has been severely impacted by extreme rainfall and flooding. For example in January 2011, there was a record flash flood in Toowoomba and the nearby Lockyer Valley, the entire SEQ region was affected by flooding, a total of 33 lives were lost, and the insurance bill was of the order of billions of dollars (Queensland Floods Commission of Inquiry, 2012). ENSO was also in a strong La Niña phase during this period. Given the socio-economic impacts of the 2011 flooding, questions were raised about the effectiveness of disaster mitigation strategies and the region has received much attention in the literature (Brodie, 2013; W. Cai and Rensch, 2012; King, Lewis, et al., 2013; Sun, Thyer, et al., 2014; Van den Honert and McAneney, 2011).

In this chapter, we show how max-stable models can be used to improve our understanding of the probability of extreme rainfall and consequently urban planning. This includes highlighting the need for mitigation strategies to account for the influence of ENSO on the distribution of extreme rainfall. Using the fraction of attributable risk (Stone and Allen, 2005), we quantify the effect of ENSO on the spatial distribution of rainfall extremes in the case study site by comparing spatial distributions under different ENSO conditions. We also show that the univariate distribution does not provide a suitable approximation for the spatial distribution with dependence, even for small domains.

The chapter is structured as follows. Section 6.2 describes the data sources used in this study. Section 6.3 introduces the necessary extreme value theory, including the max-stable process definition and brief introduction to simulation. Section 6.4 gives the results from the fitted model, including model validation and simulations from our fitted max-stable process.
6.2 Data

Observed Annual Maxima

In South East Queensland there are 957 Bureau of Meteorology stations. Of these, 196 had 50 years or more of daily observations and records that were at least 90% complete. The annual maximum rainfall was taken for each of these 196 stations for the period 1910–2014, over 12-month blocks from July to June the following year. The year 2010 therefore refers to the period July 2010 – June 2011. Adopting a non-calendar block ensures that the SEQ wet season is not split in two and that we capture the peak in the strength of ENSO; this peak generally occurs at the end of the calendar year (Rasmusson and Carpenter, 1982).

As in Chapter 2, observed annual maxima were treated as missing if more than 5% of the year’s observational record was missing. This reduces the probability that the true maximum was unobserved. However, the annual maximum was retained regardless of the missing data, if the date of the observation was the same as the date of the annual maximum at a neighbouring station with similar rainfall patterns. This allowed for a stricter condition compared with other studies (eg. Haylock et al., 2000) on the extent of allowable missing data, without unnecessarily discarding true observed maxima. After examining records for missing data, stations were discarded for records that did not have a minimum of 45 years of observed annual maxima.
Prior to model fitting the marginal distributions were tested to ensure a generalised extreme value (GEV) distribution was appropriate. One check was performed by visualising the observed annual maxima using a probability plot (Figure 6.2). A Kolmogorov–Smirnov test was also performed, where the critical value for each station was determined through bootstrap resampling (see Q. Zhang et al., (2014)).

Figure 6.2: Probability plot of 50 randomly selected stations and their fitted GEV distributions. These stations passed quality assurance testing.

Quality Control

We use the Bureau of Meteorology’s raw station data instead of 152 high quality stations in the Bureau of Meteorology’s Climate Change Network (Lavery et al., 1992). Although data from the Climate Change Network are commonly used in studies of extreme rainfall (Gallant et al., 2007; Suppiah et al., 1998), they were not suited to our application. There are only five of the 152 stations in our region of interest. These are insufficient to accurately estimate the dependence parameters of the model. Also, the locations of these stations are not representative of the region’s complex topography. Additionally, the quality assurance of the stations in the Climate Change Network was not specific to the analysis of extremes (see Chapter 2 for details).

Given these concerns we performed our own quality control. For the 196 stations of appropriate length and completeness, reconstruction was performed to help reduce the amount of missing data and quantiles of annual maxima were compared with observations at neighbouring stations (see Vicente-
Serrano et al., 2010 for details). Frequency based statistical tests from Viney et al., 2004 and King et al., 2013b were also performed to identify untagged accumulations. In total, 15 of the 196 stations were considered unsuitable, as more than 5% of the observed annual maxima failed the quality control. The locations of the 174 stations considered of suitable length and quality are shown in Figure 6.3. Refer to Chapter 2 for details on reconstruction methods and tests for untagged accumulations.

![Station Locations](image-url)

**Figure 6.3:** Station locations in South East Queensland. Large cities are marked for reference.

### Covariate Data

To consider the influence of ENSO on the distribution of extremes, the Southern Oscillation Index (SOI) was included as a covariate. The SOI is calculated using the normalised pressure difference between Tahiti and Darwin and can be used as a measure of the strength of ENSO. Sustained SOI values above 8 indicate a La Niña phase and sustained SOI values below −8 indicate an El Niño phase (Bureau of Meteorology, 2012). Despite the fact there are three phases of ENSO; El Niño, Neutral and La Niña, in our
modelling, the asymmetry was modelled by splitting the SOI index into its positive and negative values. This is common in many studies (W. Cai, Van Rensch, et al., 2010; King et al., 2013a).

We obtained monthly data for the SOI was obtained from the Australian Bureau of Meteorology. For max-stable models temporal covariates are often aggregated over each block period. For this application, the average annual SOI value was taken over the July to June 12-month period.

The spatial covariates that we considered were longitude, latitude, elevation, distance to the coast and average annual total rainfall. We obtained longitude, latitude and elevation were obtained from station meta data. The distance to the coast was calculated using the coastal outline stored in the freely available R package, Oz (Venables and Hornik, 2014). The covariate of average annual total rainfall was calculated from station records. At locations without stations, the average annual total rainfall was calculated using the Australian Water Availability project (Jones et al., 2009). This is a gridded data product, and we used the version of the grid with a 5 km resolution.

6.3 Methods

Max-Stable Processes

For our application of extreme rainfall, we used max-stable processes. These processes were introduced in Chapter 5. Recall from Section 5.3 and Definition 5.3.1 that a process $Z(x)$ is a max-stable process, if there exist normalising sequences, $\{a_n(x) > 0\}$ and $\{b_n(x)\} \in \mathbb{R}$, such that

$$Z(x) \overset{d}{=} \lim_{n \to \infty} \max_{i=1}^{n} \frac{Z_i(x) - b_n(x)}{a_n(x)}, \quad x \in \mathcal{X} \subset \mathbb{R}^2. \quad (6.1)$$

The univariate marginals of a max-stable process belong to the generalised extreme value (GEV) family for which

$$\mathbb{P}(Z(x) \leq z) = \exp\left\{-\left[1 + \xi(x)\left(\frac{z - \mu(x)}{\sigma(x)}\right)\right]^{-1/\xi(x)}\right\}, \quad (6.2)$$

where $[a]_+ = \max(a, 0)$ for a real number $a$, the location parameter is $-\infty < \mu < \infty$, the scale parameter is $\sigma > 0$ and the shape parameter is $-\infty < \xi < \infty$. We refer to this distribution as GEV($\mu, \sigma, \xi$). Details about the GEV distribution can be found in Theorem 3.2.1 of Section 3.2.

Parametric representation of the max-stable processes is given using a spectral representation (see Section 5.3.2 and equation (5.18)). The form of
the spectral representation is determined by choosing a suitable form for the underlying stochastic process (see Section 5.3.3). Common spectral representations include R. L. Smith, (1990), Schlather, (2002), Brown–Resnick (B. M. Brown et al., 1977; Kabluchko et al., 2009) and the extremal-$t$ (Opitz, 2013). The Schlather process is a special case of the extremal-$t$ process and the Smith process is a special case of the Brown–Resnick process. To decide between the extremal-$t$ and the Brown–Resnick we used model selection (see Section 5.4.2). The extremal-$t$ was selected. The form for the extremal-$t$ process is given in equation (5.22). The dependence of the extremal-$t$ is determined by a correlation function $\rho$, for which we use a Whittle–Matérn (equation 5.13), and an additional parameter for the degrees of freedom.

Fitting

To estimate the parameters of the max-stable model, we used composite likelihood methods, which have been described in Section 5.4.1. We used the R package, SpatialExtremes (Ribatet, 2015), to perform the optimisation of the computation likelihood. To ensure the parsimony of our fitted model, the marginal parameters are expressed as linear functions of space and time, also known as fitting a response surface. The GEV marginal parameters of the response surface and dependence parameters were estimated simultaneously, see equation (5.37). To determine appropriate covariates for the response surface, model selection was performed using forward selection and the Takeuchi Information Criterion (TIC) (equation (5.38)). Model selection was also used to decide to fix the nugget at zero.

Simulation

Simulation is an important tool when using max-stable processes in extreme rainfall applications. Simulation allows us to estimate spatial distributions that would otherwise be intractable, for example the distribution of the total of the annual maximum rainfall field. However, the complication in simulating from a max-stable process, is the maximum is defined over an infinite number of replicates of the underlying process. Fortunately, if we simulate from the spectral representation in a suitable order until a stopping criterion is met, only a finite number of iterations is required. Details of simulation methods are given in Section 5.5.

For our application, we also use conditional simulation methods to simulate realisations of extreme rainfall fields conditional on observed station data. Conditional simulation from a max-stable process can be done using a three-step procedure (Dombry, Eyi-Minko, et al., 2013). Refer to Section 5.5.2 for further details.
6.4 Results

6.4.1 Model Selection

We performed an exploratory data analysis in which univariate GEV(μ, σ, ξ) distributions were fitted to the station data, ignoring ENSO. The GEV parameters were then plotted as functions of space. In using the denser network of stations, nonlinearity was clearly identified in the response of the location and scale parameter as a function of the distance of the station to the coast (Figure 6.4). There was also heteroskedasticity present, reflecting an increase in the variance of parameter estimates closer to the coast. On closer inspection of the topography (Figure 6.1), the nonlinearity was attributed to interactions between the coastline and the inland orography, while the heteroskedasticity was due to elevation changes in the coastal topography.

![Exploratory data analysis](image)

Figure 6.4: Exploratory data analysis of GEV marginal parameters. The plot shows the nonlinear relation between the univariate location parameter, μ, and the distance from the station to the coast. Points are coloured by station elevation.

To express the marginal parameters as simple linear functions of space, the average yearly rainfall recorded at each station was included as a covariate. The inclusion of this covariate manages the nonlinearity and heteroskedasticity in a similar way to the inclusion of a normalisation term. Using aggregated historical information about the station is a technique commonly employed in hydrological studies (Cooley, Nychka, et al., 2007).
Model fitting was performed using composite likelihood. A trade-off exists between goodness-of-fit and computational complexity when using composite likelihood methods. This is because the sum is taken over observational pairs and as the number of pairs grows large, fitting becomes computationally intensive. For model fitting we used a random sample of 58 stations, one third of all stations available.\textsuperscript{1} Pairs with fewer than 30 years of common observations were omitted to ensure an adequate amount of overlapping observational years. The resulting number of pairs provided a suitable balance between goodness-of-fit and computational time needed for model selection. The sampled stations and their associated pairs also provide good coverage of the domain and the range of possible distances between pairs of stations. The remaining stations were then used in model validation (see Section 6.4.2).

For the response surface the following geographic covariates were considered: latitude, elevation, distance to the coast, average annual total rainfall and their associated interaction terms. Longitude was not considered as it was highly correlated with the distance to the coast over this domain. Four different forms of SOI responses for temporal covariates were also considered: no response, symmetric, asymmetric and one-sided asymmetric (Sun, Renard, et al., 2015). The temporal regression forms are given in Table 6.1. The form of the asymmetric model presented here avoids the discontinuity at zero present in many studies of the asymmetry of ENSO (W. Cai, Van Rensch, et al., 2010; King et al., 2013a). For example

$$\mu(\text{SOI}) = \beta_{\mu_1} \text{SOI} \cdot I(\text{SOI} > 0) + \beta_{\mu_2} \text{SOI} \cdot I(\text{SOI} \leq 0), \quad (6.3)$$

is discontinuous at zero. We considered the inclusion of temporal covariates in two cases: in the first, just for the location parameter, and in the second, for both the location and scale parameters.

For the response surface, geographic covariates were determined first using forward selection, and then the temporal covariates were considered. The final form of the response surface determined through model selection was

\textsuperscript{1}Another reason for partitioning the data was that we intended to perform a cross-validation, where we repeated the fitting and compared the uncertainty of parameter estimates. Due to time constraints, this is left for future work.
Table 6.1: Temporal regression forms used for model selection. The subscript $t$ refers to temporal coefficients.

<table>
<thead>
<tr>
<th>Symmetric</th>
<th>Asymmetric</th>
<th>One sided asymmetric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu(\text{SOI}) = \beta_{\mu, \text{SOI}}$</td>
<td>$\mu(\text{SOI}) = \beta_{\mu_1, \text{SOI}} \cdot I(\text{SOI} &gt; 0) + \beta_{\mu_2, \text{SOI}}$</td>
<td>$\mu(\text{SOI}) = \beta_{\mu, \text{SOI}} \cdot I(\text{SOI} &gt; 0)$</td>
</tr>
<tr>
<td>$\sigma(\text{SOI}) = \beta_{\sigma, \text{SOI}}$</td>
<td>$\sigma(\text{SOI}) = \beta_{\sigma_1, \text{SOI}} \cdot I(\text{SOI} &gt; 0) + \beta_{\sigma_2, \text{SOI}}$</td>
<td>$\sigma(\text{SOI}) = \beta_{\sigma, \text{SOI}} \cdot I(\text{SOI} &gt; 0)$</td>
</tr>
</tbody>
</table>

\[
\mu(\text{Dist}, \text{Total}, \text{SOI}) = \beta_{\mu_0} + \beta_{\mu_1, \text{Dist}} + \beta_{\mu_2, \text{Total}} + \beta_{\mu_3, \text{Dist} \cdot \text{Total}} + \beta_{\mu_1, \text{SOI} \cdot I(\text{SOI} > 0)}, \tag{6.4}
\]

\[
\sigma(\text{Dist}, \text{Total}, \text{SOI}) = \beta_{\sigma_0} + \beta_{\sigma_1, \text{Dist}} + \beta_{\sigma_2, \text{Total}} + \beta_{\sigma_3, \text{Dist} \cdot \text{Total}} + \beta_{\sigma_1, \text{SOI} \cdot I(\text{SOI} > 0)}, \tag{6.5}
\]

\[
\xi = \beta_{\xi_0}, \tag{6.6}
\]

where Dist is the distance of the station to the coast in km, Total is the average annual total rainfall (mm) and SOI is the annually averaged Southern Oscillation Index. Model selection revealed that a constant shape parameter across the domain provided a better fit. A one-sided asymmetric model for both location and scale parameters best represented the influence of ENSO on rainfall extremes.

### 6.4.2 Model Validation

Stations omitted from model fitting were used to produce standard diagnostic plots. For marginal diagnostic plots, quantile plots with bootstrapped confidence intervals were produced, for example in Figure 6.5. To plot all observed annual maxima on the same axes, the annual maxima were transformed from $\text{GEV}(\mu(s, t), \sigma(s, t), \xi)$ to $\text{GEV}(\mu(s), \sigma(s), \xi)$ using the
estimated temporal coefficients. Observations are shaded by the sign of the SOI for the corresponding year.

![Figure 6.5: Marginal diagnostic plots showing observed and simulated quantiles. The 95% simulated confidence intervals are shaded.](image)

For diagnostic plots of the joint distribution, the observed maxima at pairs of stations were plotted against the maxima simulated from our fitted model, as in Figure 6.6. We obtained 95% confidence intervals from simulation. To compare the joint distributions on the same scale we used Gumbel marginals, GEV(0,1,0). The joint diagnostic plots of Figure 6.6 are provided for station pairs at a range of distances, showing that the fitted statistical model resolves rainfall behaviour for small, medium and large distances.

We also produced diagnostic plots for dependence using the extremal coefficient, $\theta(h)$, which is a measure of partial dependence (for details on the extremal coefficient see Section 5.3.5). The extremal coefficients estimated from the fitted parameters of the max-stable model were compared to the empirical estimates from the stations omitted from fitting, as shown in Figure 6.7. Empirical estimates were obtained from the F-madogram, an analogue of the semi-varioagram for extremes (Cooley, Naveau, and Poncet, 2006). Al-
6.4. RESULTS

Figure 6.6: Diagnostic plots of observed and simulated maxima at pairs of sites. The 95% simulated bootstrapped confidence intervals are shaded. The $h$ value on each plot gives the distance between the two stations.

though the theoretical range of the extremal coefficient is $1 \leq \theta(h) \leq 2$, it is possible for the empirical estimate to take greater than 2. The plot shows that the fitted model does well at capturing the dependence of stations in the region, particularly given that the comparison is with stations omitted from fitting.

6.4.3 Spatial distributions for urban planning applications

As the model selection revealed, the distribution of extreme rainfall in South East Queensland is conditional on the state of ENSO. To motivate the need for spatial models with dependence as compared with univariate models, in this section we present simulated results conditional on a negative SOI value. A negative SOI value is reflective of the state of the model in an El Niño phase. To understand the impact of ENSO on the spatial distributions of rainfall extremes, in sections 6.4.4 and 6.4.5 we contrast these simulated results with simulations conditional on a strong La Niña phase.

Simulations of the annual maximum rainfall field from the fitted max-
Figure 6.7: Model validation plot of the extremal coefficient. The points are the empirical estimates from the stations omitted during fitting and the line shows extremal coefficients estimated from the fitted parametric model. The distance is given in kilometres.

stable process are depicted in Figure 6.8. These simulations were performed on a $100 \times 100$ resolution grid. These realisations of the process highlight random differences from year to year in annual maximum rainfall amounts and in the patterns of spatial dependence. Simulations such as those in Figure 6.8 are currently underutilised for practical applications, such as disaster mitigation, rainfall inputs to flood models and to supplement physical climate models.

To give an example of how simulations from a max-stable process can be used for disaster mitigation, we simulated $10^4$ realisations of the annual maxima rainfall field within a 5 km radius of the Brisbane central business district. Brisbane is the state capital of Queensland, (153.03, -27.47). These simulations were performed using 276 gridded points to resolve the 5 km radius domain. From the simulations, we estimated the empirical cumulative distribution of the maximum rainfall observed across the domain (Figure 6.9). To demonstrate the effect of spatial dependence we considered the maximum annual rainfall simulated in the domain at radiuses of 1 km, 2.5 km and 5 km. This was compared with the univariate marginal distribution for Brisbane from our fitted max-stable process (Figure 6.9).

We empirically estimated the exceedance level, $z$, for maximum rainfall
6.4. RESULTS

Figure 6.8: Simulations of the annual maximum rainfall field conditional on a negative SOI.

above the 95$^{th}$ quantile

\[ P[Z(x) \geq z] = 0.95. \]  \hspace{1cm} (6.7)

For a 1 km, 2.5 km and 5 km radius the exceedance levels are 210 mm, 230 mm and 251 mm respectively. This is in contrast to 188 mm for the marginal distribution at Brisbane, or equivalently 0 km radius.

The differences in the distributions of Figure 6.9 highlight the fact that spatial distributions with dependence are needed to fully understand the probability of extreme rainfall events and to mitigate their potential impacts. Consider for example the application of urban drainage. If we used the marginal distribution at Brisbane to estimate drainage capacities needed in the CBD, then these drainage capacities would be insufficient for the annual maximum rainfall event observed in the domain. Underestimation of this event could re-
result in drain failure and urban flooding. For example, if we use the marginal distribution at Brisbane to approximate a 1 km radius domain then we underestimate the $95^{th}$ quantile by 10%. For a 5 km radius, the percentage underestimated increases to 25%.

6.4.4 The influence of ENSO on the spatial distribution of rainfall extremes

To investigate the impact of ENSO on our simulated process, we conditioned our simulations on two different SOI values; one a negative SOI value and the other a SOI value of 18. The negative SOI value reflects a year under El Niño conditions. The SOI value of 18 reflects a strong La Niña year, chosen to be the same magnitude as the SOI value for the non-calendar year, 2010–2011. To compare the distribution of a random variable $X$ conditional on different values of SOI, we used the fraction of attributable risk (FAR)

$$\text{FAR}(x) := 1 - \frac{\Pr(X > x \mid SOI \leq 0)}{\Pr(X > x \mid SOI = 18)}. \quad (6.8)$$
If there is no difference in the distributions of $X$ conditional on different SOI values then the FAR value is zero, reflecting no increased risk. In contrast, if the ratio of tail probabilities becomes small and positive then the FAR value will increase, reflecting a much higher risk conditional on an SOI value of 18, compared with a negative SOI value.

Figure 6.10 gives the FAR plots for the distributions described in Figure 6.9. Observe that during a year in which the average SOI value is 18, the probability of experiencing a given extreme rainfall event is much higher compared with negative SOI year. For example, a maximum rainfall exceeding the 95th quantile in the 1 km radius domain is 65.3% more likely under the strong La Niña conditions. Similarly, for the other radii of 2.5 km and 5 km, the FAR values are 67.2% and 67.8% respectively. As the probability of exceedance decreases and fewer simulations occur above a given exceedance level, the variability in the FAR estimator increases.

![Maximum rainfall (mm) simulated in Brisbane CBD](image)

Figure 6.10: The fraction of attributable risk for the distributions in Figure 6.9 for a negative SOI value compared with an SOI value of 18.

### 6.4.5 Understanding historical flash flooding through conditional simulations

In this section, we demonstrate further how simulation can be used as a tool for disaster mitigation. Conditional simulations can be used by disaster
planners to better understand extreme events in hindsight, by examining the effectiveness of the mitigation strategy and the response to the disaster. Using this information, the response to any similar events in future can be improved.

We used conditional simulations to investigate the extreme rainfall and consequent flash-flooding in Toowoomba, \((151.95, -27.56)\), on 11th of January 2011. This event occurred during a strong La Niña year with an average SOI of 18. To understand the severity of this flash flood we answered two questions: (i) If we experience a similar magnitude La Niña year again, how likely are we to observe this event; and (ii) How likely is this event to occur in an El Niño or a neutral phase of ENSO?

Conditional simulations were performed on a \(20 \times 20\) grid with cells of approximately 1 km that was centred on Toowoomba. Within the gridded domain there are five stations with data suitable for use as conditional observations: 180.8 mm at Withcott, 82.4 mm at Moyala, 123.4 mm at Toowoomba Airport, 149.6 mm at Middle Ridge and 143.2 mm at Mt Kynoch. For each of these stations, the annual maximum was observed on January 11th 2011. The dates of the maxima are important, as conditional observations occurring concurrently indicate that the observations are all part of the same extremal function (see Definition 5.5.3).

Tamba station also had an observation for the period of interest, however it was aggregated over 4 days. This aggregated total was within 1 mm of the aggregated total at Mt Kynoch over the same period. Given the similarity in the observational records (see Table 6.2), we assumed the observations at Tamba were equal to those at Mt Kynoch. Otherwise the northern boundary of our domain was unconstrained, leading to greater variance in simulated results.

Table 6.2: Similarity of rainfall observations at Mt Kynoch and Tamba stations. A four-day aggregation period of 178.8 mm was observed at Tamba station. The rainfall total for the same four-day aggregation at Mt Kynoch was 179.6 mm.

<table>
<thead>
<tr>
<th>Date</th>
<th>Mt Kynoch</th>
<th>Tamba</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-01-11</td>
<td>104.6</td>
<td>104.2</td>
</tr>
<tr>
<td>11-01-11</td>
<td>143.2</td>
<td>+</td>
</tr>
<tr>
<td>12-01-11</td>
<td>36.4</td>
<td>+</td>
</tr>
<tr>
<td>13-01-11</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>14-01-11</td>
<td>0</td>
<td>178.8</td>
</tr>
</tbody>
</table>

To understand the severity of the 2011 Toowoomba flash flood we
simulated $2 \times 10^4$ realisations of the annual maximum rainfall field. Of these simulations, $10^4$ were conditional on the 2010–2011 station data and $10^4$ realisations were without the conditional station data. All simulations were conditional on an SOI value of 18, approximately the same SOI magnitude for the 2010–2011 non-calendar year. By comparing simulations with and without the conditional station data, we can better understand the probability of witnessing an event of this magnitude again given the SOI value of 18 that year. An example of a conditional simulation for the Toowoomba Region is given in Figure 6.11. Locations of conditional stations are as marked.

Figure 6.11: An example of a simulation from our fitted max-stable process conditional on the observations: 180.8 mm at Withcott, 82.4 mm at Moyala, 123.4 mm at Toowoomba Airport, 149.6 mm at Middle Ridge and 143.2 mm at both Mt Kynoch and Tamba. The conditional SOI value was 18.

We use the total annual maximum rainfall to occur in the region as a proxy for the severity of the extreme rainfall event. Frequency histograms for simulations both with and without the conditional station data are given in Figure 6.12. The median total annual maximum rainfall simulated conditional on the station data was 60646 mm. In a year with an average
SOI value of 18, this amount of rainfall would be exceeded with probability 0.134, corresponding to a 1 in 7.5 chance of occurrence.

![Frequency histograms of total annual maximum rainfall simulated in the Toowoomba region. Simulations are conditional on an SOI value of 18 to reflect the 2010–2011 conditions. Simulations conditional on observed station data for the 2010–2011 year are shown in purple and simulations without the conditional station data are shown in orange.](image)

To understand the probability of this flash flood event occurring in an alternate phase of ENSO, we again use the fraction of attributable risk, equation (6.8). We calculated our FAR values for the total annual maximum rainfall using simulations occurring under conditions where the SOI value is negative compared with an SOI value of 18 (Figure 6.13). We found that the probability of observing the median event from our simulations conditional on station data is approximately 82% more likely in a strong SOI year of 18 compared with an El Niño year.

### 6.5 Discussion and conclusions

In this chapter, we have shown how we could better use max-stable processes for applications in extreme rainfall and disaster mitigation. A particular focus of this was demonstrating how the El Niño Southern Oscillation can affect the distribution of spatial extremes. To convey this point we fitted a max-stable process to the daily annual maximum rainfall in a case-study region of South East Queensland (SEQ). Using max-stable process simulation we answered...
6.5. DISCUSSION AND CONCLUSIONS

Figure 6.13: Fraction of attributable risk comparing negative SOI values and an SOI value of 18 for the distribution of total annual maximum rainfall in Toowoomba.

questions about the spatial distribution of extremes; such questions would otherwise have been intractable without accurately capturing the dependence. We also emphasised that the univariate distribution of rainfall extremes was not a suitable proxy for the spatial distribution with dependence, even for small domains of 1 km radius.

Our model selection revealed that the distribution of extreme rainfall in SEQ was conditional on the state of the ENSO. Using the SOI as an indicator for ENSO strength, we found for positive SOI values, as the SOI value increased the marginal GEV distributions were shifted right and there was increased variability. In contrast, no significant change was detected in the marginal parameters for negative SOI values. This confirms a one-sided asymmetric response of extreme rainfall to ENSO also observed by W. Cai, Van Rensch, et al., (2010), King et al., (2013a), and Sun, Thyer, et al., (2014). The La Niña phase of ENSO therefore has a far greater impact on the distribution of extreme rainfall in SEQ than the El Niño phase. As the state of ENSO can often be predicted months ahead of time (Dake et al., 2004), this raises the possibility of conditioning on the predicted state of ENSO to improve the estimation of the probability of extreme rainfall events in the coming year.

In addition to quantifying the magnitude of the shift and scale in the marginal distributions with ENSO, we sought to examine the influence of
ENSO on spatial extremes. We compared our simulations of the annual maximum rainfall field conditional on El Niño conditions to simulations conditional on strong La Niña conditions. To quantify the difference in distributions we used the fraction of attributable risk. Our results highlight that mitigation strategies need to account for the influence of climate drivers such as ENSO, otherwise our design standards will be insufficient for strong La Niña years. For example, we found the flash flood in Toowoomba, Queensland of 2011, was 82% more likely under the 2011 strong La Niña compared with El Niño conditions.

For mathematical and computational ease, we assumed an isotropic dependence structure. In coastal areas, an anisotropic dependence structure could improve the ability of our fitted model to reproduce extreme rainfall events. However, given the complexity of the domain, a dependence structure that varies in space may be required. This problem is complex and research in adapting methods from spatial statistics to extreme value models is ongoing (Risser and Calder, 2015).

It should be noted that simulation uncertainty and uncertainty in parameter estimates are not the same. To incorporate parameter uncertainty one possibility could be to use an ensemble of models, where the parameters are sampled from the asymptotic distribution of the maximum composite likelihood estimator.

Ideally we would like to make better use of all available data. This includes using all available stations for fitting. Other uses of the data could include adopting a peaks over threshold approach to extremes compared with block maxima. The literature on spatial approaches using peaks over threshold is still developing, and does have the potential to be used for this application in the future (De Fondeville et al., 2016; Thibaud et al., 2013). Repeating the analysis on sub-daily scales would also be interesting given the change in the intensity, frequency and duration relationship.
Chapter 7

A Regionalisation of Australia for Rainfall Extremes

7.1 Introduction

The socio-economic impacts of extreme rainfall and associated flooding can occur on a scale that covers hundreds of kilometres. For example, the 2011 floods in Australia affected an area the size of France and Germany (Queensland Floods Commission of Inquiry, 2012). Flooding on this scale is not unprecedented in Australia, with other examples including the flooding in Eastern Australia in 2017 caused by Cyclone Debbie and the 1974 Queensland floods. This is further demonstrated in Figure 7.1, where we plot the years in which station records their wettest and driest, observed annual maximum. This figure clearly shows that record daily totals are positively correlated over large geographical scales.

The scale of these historical instances establishes the need for statistical models of rainfall extremes that are effective across large spatial distances. Australia however, is one of the largest countries by area, with a varied climate and complex topography. Attempting to capture this complexity accurately in a single, parsimonious statistical model is impractical.

Temporally, rainfall is highly variable within each season, year and decade. Inter-annual rainfall variability, particularly of heavy rainfall days, is influenced by tropical cyclones (Hopkins and Holland, 1997) and east coast lows (Pepler, Coutts-Smith, and Timbal, 2014). Variability on a decadal scale is impacted by large scale climate drivers including, El Niño Southern Oscillation (ENSO), Interdecadal Pacific Oscillation (IPO), Southern Annaul Mode (SAM), Madden–Julian Oscillation (MJO) and Indian Ocean Dipole (IOD) (Power et al., 2006; Risbey et al., 2009).
Figure 7.1: Plot of the stations that record their wettest (blue) and driest (red) observed annual maximum rainfall in a given year. Stations shown have a minimum of 20 years of observational data, but note that observational periods do vary between stations.

Additionally, rainfall is influenced by the local topography. The mountain range that spans the Eastern Australian coastline, the Great Dividing Range, has a profound impact on Australian precipitation due to orographic lift. This is demonstrated in Chapter 6, where complex interactions between extreme rainfall and orography were evident even on a small geographical scale. However, it is unrealistic to attempt to build a model based on detailed geographic covariates, climate covariates and their nuanced interactions over the whole continent.

An added obstacle to the creation of an Australia-wide statistical model,
is that current methods for extremes can be inflexible and were not designed for the desired scale. A prime example is max-stable processes, see Chapter 5. The dependence structure of these processes is often fixed for computational and mathematical simplicity.

For small domains a fixed dependence structure may prove to be a fair modelling assumption. For a domain the size of Australia this assumption is poor. Caution therefore needs to be employed when grouping stations and assuming a single dependence structure. If the dependence structure used is incorrect this affects the accuracy and performance of the model. Assuming an incorrect dependence structure may prove to be a worse modelling assumption than assuming independence between stations (Zheng et al., 2015). Another common assumption is that the dependence structure is isotropic. This may not be a valid assumption depending on the region. For coastal regions in Australia, we expect the dependence to exhibit anisotropy, and for the direction of the anisotropy to vary relative to the direction of the coastline.

We have chosen to simplify the complexity of the problem in order to develop an understanding of rainfall extremes on an Australia-wide scale. In this chapter, we consider only the dependence of Australian rainfall extremes, by transforming the GEV marginal distributions to standard Fréchet distributions. We then create a regionalisation of Australia, where regions are identified based on the dependence of rainfall extremes. We investigate whether it is reasonable to assume a single dependence structure across each of the regions identified by fitting max-stable processes to model local rainfall extremes. The added bonus of this approach is that we can parallelise the model fitting. The identification of homogeneous regions of extremal dependence greatly improves the parsimony of our modelling.

Regionalisations are common in flood frequency analysis and studies of hydrological extremes. Examples of different approaches to regionalisation based on extreme rainfall are given in Asadi, Engelke, and Davison, (2018), Carreau, Naveau, and Neppel, (2016), and Hosking and Wallis, (2005). For Australia a regionalisation specific to rainfall extremes does not exist. However, there are regionalisations formed using topography and mean climate (CSIRO and Bureau of Meteorology, 2015; Stern, De Hoedt, and Ernst, 2000).

Our regionalisation method extends the clustering method presented in Bernard et al., (2013). In their approach, clustering is done using $K$-medoids (Kaufman and Rousseeuw, 1990) and a dissimilarity measure that is based on bivariate extremal dependence and derived from the F-madogram (Cooley, Naveau, et al., 2006). Such dissimilarity measures, dissimilarities for short, differ from distances as they do not need to satisfy the triangle inequality.
Using the F-madogram to create the dissimilarity measure is powerful. No information about climate or topography is required to form spatially homogeneous clusters. Additionally, we are free from distributional assumptions as this dissimilarity can be estimated non-parametrically from raw maxima.

However, unsupervised learning algorithms should not be applied without proper consideration. We found that $K$-medoids produced spurious clustering when used with the F-madogram dissimilarity if there are too few clusters. This spurious clustering can be further exacerbated by spatial density of the station network. We highlight this in two toy examples. Given this, we found $K$-medoids to be ill-suited for clustering the Australian rainfall stations. To address this, instead of $K$-medoids, we use hierarchical clustering with the F-madogram dissimilarity. This ensures the clustering we obtain is well informed by extremal dependence. To convert the clustering to a regionalisation, we perform a final classification step.

### 7.2 Natural Resource Management Clusters

The concept of sub-dividing a region based on climate and topography is not new. For example, under the Köppen classification system (Köppen, 1931), Australia can be broadly split into six major classification groups (Stern et al., 2000). These groups comprise equatorial, tropical, subtropical, desert, grasslands and temperate. A sub-classification of these broad categories, specific to Australian climate and topography, is given by Stern et al., (2000). This sub-classification, in conjunction with stakeholder engagement, was used to produce the Natural Resource Management (NRM) clusters (CSIRO et al., 2015).

There exist three different NRM regionalisations, shown in Figures 7.4, 7.3 and 7.2. These three different regionalisations were designed to enable stakeholders to assess the impact of climate change on a range of different regional scales. Note that some of these regions encompass islands and their surrounding water boundaries. The region codes that accompany each of the NRM Figures are given in Table 7.1.

We have used the NRM super-clusters to help inform our regionalisation based on rainfall extremes. As part of this process, we make some interesting anecdotal observations regarding the network density and location of the NRM boundaries (Appendix A).
7.2. NATURAL RESOURCE MANAGEMENT CLUSTERS

Figure 7.2: The NRM super clusters.

Figure 7.3: The NRM clusters.
Figure 7.4: The NRM sub clusters.
7.2. **NATURAL RESOURCE MANAGEMENT CLUSTERS**

<table>
<thead>
<tr>
<th>code</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CS Central Slopes</td>
</tr>
<tr>
<td>2</td>
<td>EC East Coast</td>
</tr>
<tr>
<td>3</td>
<td>MB Murray Basin</td>
</tr>
<tr>
<td>4</td>
<td>MN Monsoonal North</td>
</tr>
<tr>
<td>5</td>
<td>R Rangelands</td>
</tr>
<tr>
<td>6</td>
<td>SS Southern Slopes</td>
</tr>
<tr>
<td>7</td>
<td>SSWF Southern and South-Western Flatlands</td>
</tr>
<tr>
<td>8</td>
<td>WT Wet Tropics</td>
</tr>
<tr>
<td>9</td>
<td>NA Northern Australia</td>
</tr>
<tr>
<td>10</td>
<td>EA Eastern Australia</td>
</tr>
<tr>
<td>11</td>
<td>SA Southern Australia</td>
</tr>
<tr>
<td>12</td>
<td>RN Rangelands (North)</td>
</tr>
<tr>
<td>13</td>
<td>SSVW Southern Slopes (Vic West)</td>
</tr>
<tr>
<td>14</td>
<td>SSTE Southern Slopes (Tas East)</td>
</tr>
<tr>
<td>15</td>
<td>SSTW Southern Slopes (Tas West)</td>
</tr>
<tr>
<td>16</td>
<td>ECN East Coast (North)</td>
</tr>
<tr>
<td>17</td>
<td>RS Rangelands (South)</td>
</tr>
<tr>
<td>18</td>
<td>MNE Monsoonal North (East)</td>
</tr>
<tr>
<td>19</td>
<td>MNW Monsoonal North (West)</td>
</tr>
<tr>
<td>20</td>
<td>ECS East Coast (South)</td>
</tr>
<tr>
<td>21</td>
<td>SSWFW Southern and South Western Flatlands (West)</td>
</tr>
<tr>
<td>22</td>
<td>SSWFE Southern and South Western Flatlands (East)</td>
</tr>
<tr>
<td>23</td>
<td>SSVE Southern Slopes (Vic/NSW East)</td>
</tr>
</tbody>
</table>

Table 7.1: Codes for the NRM regions. These codes are based on climate conditions, biophysical factors and expected patterns under climate change. Observe that the sub-classifications are derivatives of the more general classifications, for example, ECS and ECN, are derivatives of EC.
7.3 Clustering Dissimilarity

In this section, we outline the mathematical background needed to obtain the dissimilarity that is used for clustering.

7.3.1 Bivariate dependence

The dissimilarity we use is based on the extremal dependence between pairs of stations. We first introduced extremal dependence in Section 5.3.5 in the context of multivariate extreme value distributions and max-stable processes. Here we discuss bivariate extreme value distributions.

Let $S$ be the set of stations we wish to cluster. Define $M_i$ to be the random variable for the annual maximum daily rainfall at station $x_i \in S$, where $i = 1, \ldots, n$. For any pair of stations, $x_i$ and $x_j$, assume that the bivariate distribution of $(M_i, M_j)$ is well approximated by a bivariate extreme value distribution

$$P(M_i \leq z_i, M_j \leq z_j) = \exp \left\{ -V_{ij} \left( \frac{-1}{\ln F_i(z_i)}, \frac{-1}{\ln F_j(z_j)} \right) \right\}, \quad (7.1)$$

where $F_i(z_i) = P(M_i \leq z_i)$, $V_{ij}(a, b)$ is given by

$$V_{ij}(a, b) = 2 \int_0^1 \max \left( \frac{w}{a}, \frac{1-w}{b} \right) dH_{ij}(w), \quad (7.2)$$

and $H_{ij}$ is any distribution function on $[0, 1]$ with expectation equal to 0.5 (eg. S. Coles, 2001). For examples of functional forms for $H_{ij}$ see S. Coles, (2001) and for forms of the bivariate extreme value distribution associated with common max-stable processes see Dey et al., (2015). Equations (7.1) and (7.2) are consistent with the more general definition of a finite-dimensional extreme value distribution presented in Section 5.3.4.

For clustering, we are interested in the special case that $F_i(z_i) = F_j(z_j)$. In this case, the bivariate extreme value distribution can be characterised in terms of the extremal coefficient (see equation (5.30)) repeated below for the reader’s convenience, for all $x_i \in S$ and $x_j \in S$ where $h = x_j - x_i$,

$$\theta(h) = V_{ij}(1, 1),$$

$$\theta(h) = -z \log \left[ P(M_i \leq z, M_j \leq z) \right]. \quad (7.3)$$

(7.4)

The value of $\theta(h)$ provides an indication of the partial dependence between the maxima at the two locations $x_i$ and $x_j$. The range of $\theta(h)$ is $[1, 2]$, where the lower bound of the interval corresponds to dependence of $M_i$ and $M_j$, and the upper bound conversely indicates independence. For clustering, the dissimilarity used is a function of $\theta(h)$ and is derived from the F-madogram.
7.3. CLUSTERING DISSIMILARITY

7.3.2 F-madogram

The F-madogram (Cooley, Naveau, et al., 2006) links ideas of standard dependence in spatial statistics and dependence in extreme value theory. In spatial statistics a variogram (see equation (5.7)) is commonly used to interpret the dependence between two locations in a stochastic process. However, for extremes the variogram is often undefined, as the distributions can be heavy-tailed and the variance is not finite. In contrast to the variogram, the F-madogram is always defined. This follows as \( F_i(M) \) is mapped to the interval \([0, 1]\).

The F-madogram is the mean absolute difference (MAD) between two distribution functions

\[
d(x_i, x_j) = \frac{1}{2} \mathbb{E} \left[ |F_i(M_i) - F_j(M_j)| \right],
\]

where \( x_i \in \mathcal{S} \) and \( x_j \in \mathcal{S} \). The F-madogram dissimilarity links to ideas in extreme value theory as equation (7.5) can be written as a function of the extremal coefficient, \( \theta(h) \), equation (5.31),

\[
d(x_i, x_j) = \frac{\theta(h) - 1}{2(\theta(h) + 1)}.
\]

Equation (7.5) follows using the identity \( 2 \max(a, b) = a + b - |a - b| \), where \( a, b \in \mathbb{R} \). Using the F-madogram dissimilarity to cluster the stations means that the resulting clusters have an interpretation in terms of partial dependence of extremes.

We estimate the F-madogram dissimilarities non-parametrically to avoid distributional assumptions and model fitting. Let \( \mathcal{Y}_{ij} \) be the set of years when both stations \( x_i \) and \( x_j \) have annual maximum observations; then

\[
\widehat{d}(x_i, x_j) = \frac{1}{2|\mathcal{Y}_{ij}|} \sum_{y \in \mathcal{Y}_{ij}} \left| \hat{F}_i(M^y_i) - \hat{F}_j(M^y_j) \right|,
\]

where \( \hat{F}_i(z) \) is the empirical cumulative distribution function at \( x_i \)

\[
\hat{F}_i(z) = \frac{1}{|\mathcal{Y}_i|} \sum_{y \in \mathcal{Y}_i} 1_{(M^y_i < z)},
\]

and \( \mathcal{Y}_i \) is the number of annual maximum observations at \( x_i \). Note that \( \mathcal{Y}_i \) and \( \mathcal{Y}_{ij} \) may differ depending on missing observations.

Non-parametric estimation is fast and uses only the raw station observations. This makes clustering using the F-madogram particularly powerful, as
no external information about climate or topography is needed to identify regions of similar extremal dependence.

Further to the work of Bernard et al., (2013), we make a small adjustment to the dissimilarity used in clustering. As the range of $\theta(h)$ is $[1, 2]$, the theoretical range of $d(x_i, x_j)$ is $[0, \frac{1}{6}]$. However, the empirical estimator $\hat{d}(x_i, x_j)$ can take values outside of this range. Given this we use the estimator

$$
\hat{d}^*(x_i, x_j) = \begin{cases} 
\hat{d}(x_i, x_j), & 0 \leq \hat{d}(x_i, x_j) \leq \frac{1}{6}, \\
\frac{1}{6}, & \hat{d}(x_i, x_j) > \frac{1}{6}.
\end{cases}
$$

(7.9)

7.4 Clustering Methods

In this section, we introduce two clustering methods, $K$-medoids clustering and hierarchical clustering. Both of these methods can be used to partition the stations into groups based on extremal dependence. However, the suitability of these methods depends on the application. We create examples that are based on our desired clustering behaviour to demonstrate the difference between the methods.

7.4.1 $K$-medoids Clustering

The clustering method $K$-medoids (eg. Hastie, Tibshirani, and Friedman, 2009) can be used to partition the stations. Let $n = |S|$ and $K$ be the number of clusters. The parameter $K$ has to be prespecified and induces a partition of $S$ into clusters $C_k$, such that $S = \bigcup_{k=1}^K C_k$ where $1 \leq K \leq n$ and $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. For each cluster, $C_k$, the medoid of that cluster, $m_k$, is the station that minimises the sum of intrACLuster dissimilarities

$$
m_k = \text{argmin}_{x_i \in C_k} \sum_{x_j \in C_k} d^*(x_i, x_j).
$$

(7.10)

To determine the clusters and medoids associated with a given value of $K$, we use the greedy algorithm, partitioning around medoids (PAM) (Kaufman et al., 1990), which is as follows.

1. Randomly select an initial set of $K$ stations. These are the initial medoids.

2. Assign each station, $x_i$, to its closest medoid, $m_k$, based on the F-madogram dissimilarity of equation (7.9). This determines the initial clustering of the stations.

3. For each cluster, update the medoid according to equation (7.10)
4. Repeat steps 2–4 until the medoids no longer change.

Like many clustering algorithms, PAM converges to a local minimum, but not necessarily the global minimum. It is therefore advisable to repeat PAM with different initialisations to help ensure performance consistency in the performance of the algorithm.

As the number of stations increases, and the number of clusters grows large, so too does the computational complexity of PAM. If limits of computational capacity are encountered, Clustering LARge Applications (CLARA) (Rousseeuw and Kaufman, 1990) can be used in conjunction with PAM. CLARA uses sampling to reduce the computational complexity and determine the medoids. Provided that the subsets sampled are representative of the original data, CLARA performs well.

**Number of Clusters**

As the number of clusters needs to be prespecified in K-medoids, we must choose a suitable value for $K$. This can be done using expert knowledge, such as using prior knowledge about climate or topography. Alternatively there is a range of statistical techniques for determining the optimal number of clusters. The R package, NbClust, implements 30 different cluster selection methods (Charrad et al., 2014) and gives a summary of which methods selected which number of clusters. Examples of the methods implemented include the commonly used gap statistic (Tibshirani, Walther, and Hastie, 2001) and the silhouette coefficient (Rousseeuw, 1987).

The silhouette coefficient is a measure for cluster cohesion and is given by the ratio of intraclass dissimilarity to interclass dissimilarity,

$$s_i(k) = 1 - \frac{d^*(m_k, x_i)}{d^*(m_{k'}, x_i)},$$  \hspace{1cm} (7.11)

where $x_i \in C_k$ and $m_{k'}$ is the closest medoid to $x_i$ excluding $m_k$. If the station $x_i$ is well clustered, then $s_i(k)$ should be close to 1, as $d^*(m_k, x_i)$ is small relative to $d^*(m_{k'}, x_i)$. In contrast, if $s_i(k)$ is close to zero, the station $x_i$ could reasonably be clustered in $C_k$ or $C_{k'}$. In this case the clustering of $x_i$ is not informative.

An overall metric of how well stations are clustered is given by the average silhouette coefficient,

$$\bar{s}(K) = \frac{1}{n} \sum_{k=1}^{K} \sum_{x_i \in C_k} s_i(k).$$  \hspace{1cm} (7.12)
To determine the number of clusters, we can plot $s(K)$ against $K$. The optimal number of clusters, $K^*$, can be determined when the addition of a new cluster does not greatly improve the average silhouette coefficient. Bernard et al., (2013) used this method to select the number of clusters.

**Choice of $K$-Medoids**

There is a strong motivation to use $K$-medoids clustering over the closely related $K$-means (eg. Hastie et al., 2009). In $K$-means, instead of medoids, centroids are used as the representative objects in each cluster. The clustering is then determined by finding the partition of $\mathcal{S}$, $\tau_\mathcal{S}$, that minimises

$$\tau_\mathcal{S} = \arg\min_{\tau_\mathcal{S}} \sum_{k=1}^{K} \sum_{x_i \in C_k} d^*(x_i, \mu_k)^2,$$

(7.13)

where $\mu_k$ is the mean of the points in $\mathcal{S}$. As a consequence, the centroid of each cluster does not need to be a station and the centroids are not robust to outliers. For extremes this is undesirable.

Additionally, the centroid in $K$-means is determined by taking an average. An average of a set of normally distributed random variables, will still be normally distributed. However, when taking the average of extremes, there is no similar natural interpretation. The average of extremes may poorly represent the underlying process. It is therefore preferred for the representative object in the cluster to be one of the clustered points, as is the case with a medoid. It was for these reasons, that Bernard et al., (2013) preferred $K$-medoids clustering for this application involving extremes.

### 7.4.2 Limitations of $K$-medoids Clustering

In Bernard et al., (2013), the application region was France. As the Australian station network is different, we must decide whether the F-madogram dissimilarity and the unsupervised learning method of $K$-medoids is still suitable. Given this we highlight some features of the clustering that are not unexpected, but need to be understood for our goal of identifying homogeneous regions.

**Bounded Range**

The F-madogram dissimilarity has an upper bound of 1/6. This creates fundamentally different clustering behaviour under $K$-medoids compared with a dissimilarity that does not have a bounded range. Consider the illustrative example given in Figure 7.5. In this example we created two distinct groups of points. These groups are clearly distinguishable by the human eye. We used Euclidean distance for clustering, but restricted the
maximum distance to 2. This restricted distance mimics the bounded range property of the F-madogram dissimilarity.

![Capped Distance Example of K-Medoids](image)

Figure 7.5: An example of $K$-medoids clustering. The distance used was Euclidean, but the maximum distance was restricted to 2. The points enclosed by the dotted circle, are points that are within a distance of 2 away from the medoid.

We observe that the medoids are in the centre of the two groups of points. We expected this under the optimisation. However, we note that the clustering did not recover the two original groups. This is because points falling outside of the dotted circular lines are a distance of more than 2 away from both medoids. Therefore under the optimisation, these points can be assigned to either medoid.

We note that the points in the right group that are a distance more than 2 away from both medoids were assigned to the left cluster. This was simply a consequence of basic ordering in the PAM algorithm, with the purple cluster ID assigned as 1 and the orange cluster ID assigned as 2. This means if a group of stations are independent of all medoids, they can appear to be grouped meaningfully even though they are not.

**Network Density**

Changes in network density also impact our natural intuition of how the clustering should work. Consider another illustrative example, using the same distance, where there are 1000 points generated in the left group, compared
with 100 points in the right group, Figure 7.6. The clustering results are shown for \( K = 4 \) and \( K = 5 \). Only when \( K = 5 \) is a medoid assigned to the group of fewer points. The optimisation favours the creation of medoids in regions where there is a higher density of points. We desire a clustering method that is capable of identifying these two groups as two separate clusters.

Figure 7.6: Examples of \( K \)-medoids clustering. The distance used was Euclidean, but the maximum distance was restricted to 2.

**Implications for the Australian Network**

Applying \( K \)-medoids clustering with a F-madogram dissimilarity to the Australian station network is unlikely to result in homogeneous clusters based
on extremal dependence. This is clear from the examples in Figure 7.5 and Figure 7.6.

Stations in inland Australia are sparsely located. With few neighbours the extremes at these stations are likely to be very weakly dependent, or completely independent, of extremes at other stations. Sparsely located stations are therefore likely to be clustered poorly, similarly to Figure 7.5.

In addition, the spatial density of stations in Eastern Australia is much higher than in Central and Northern Australia. Medoids are more likely to be created in regions where the density of points is higher, for example in Figure 7.6. We would therefore need to question whether clusters created along the Eastern Australian coast reflect features of extremal dependence, or are simply a function of station density.

Favouring the creation of medoids in denser regions forces stations in regions where the dependence is weaker to be assigned poorly. Given this, it is not appropriate to cluster regions of different densities. This poses a challenge, as the density of stations is determined in F-madogram space, not Euclidean space.

Finally, it is often not possible to obtain an ‘optimal’ value for $K^*$ by viewing the plot of $K$ against the average silhouette coefficient. The range of dependence for rainfall extremes is small, generally less than 30 km. Given this range, variable network density, and how medoids are generated, plots of the average silhouette coefficient against the number of clusters need not be monotonic. Therefore these plots are very difficult to interpret meaningfully given the interplay of these effects. For our application, local maxima in the plot were found to produce clustering that was spurious in contrast to the intended purpose of finding an optimal balance between intracluster and extraccluster dissimilarities. An example of this is Tasmania, where the plot of the average silhouette coefficient identifies two clusters, but this is an insufficient number of medoids to cover the domain.

In these examples, we have demonstrated that using $K$-medoids and an F-madogram dissimilarity can produce clustering that is contrary to our intuition. Many researchers are adopting machine learning methods for their perceived simplicity. However, due consideration needs to be given to methodological assumptions in order to prevent invalid interpretation of results.

### 7.4.3 Hierarchical Clustering

An alternative unsupervised learning algorithm to $K$-medoids is hierarchical clustering (eg. Hastie et al., 2009). This clustering method is less susceptible
to the limitations of $K$-medoids demonstrated in Section 7.4.2. In hierarchical clustering, there is also no need to specify the number of clusters, $K$, prior to clustering.

Core to the hierarchical clustering method is the dendrogram. The dendrogram is a tree-like structure that can be used to visualise the partitioning of the points into clusters. Each single point forms a leaf in the tree. When these leaves are connected they form branches, where a branch in the most trivial sense can be a single leaf. Branches are connected according to a linkage criterion, where the linkage criterion defines the dissimilarity between two branches. Examples of dendrograms can be seen in Figure 7.9.

The dendrogram naturally partitions points into clusters. A cut across the tree at a height $h$, induces a partitioning of the points, where the points in each branch are assigned to the same the cluster. Equivalently, we can specify the number of clusters, $K$. The tree is then cut at the height that will induce $K$ clusters.

To determine an appropriate height to cut the tree it is useful visualise the dendrogram. The height at which the branches are fused determines the strength of association between two clusters. For two branches joined at the bottom of the tree, this suggests the points in these branches are strongly associated. Branches joined at the top of the tree suggest a much weaker association between groups of points. The height of the cut should be made with reference to the desired strength of association between the clusters and the separability of those clusters within the tree.

The dendrogram can be generated in two ways, namely, via agglomerative or divisive approaches. When hierarchical clustering is agglomerative, initially each point is in its own cluster. Branches are successively connected until all points are in the same cluster. In contrast, when the approach is divisive, all points initially belong to the same cluster and the algorithm then successively splits clusters into branches until each branch is a single leaf.

We prefer the agglomerative approach for our application. In general, the agglomerative approach is more popular (Hastie et al., 2009). The divisive approach runs the risk of clustering points that do not belong together in the initial state. This can result in a failure to generate meaningful clusters.

**Algorithm**

The algorithm to generate the agglomerative dendrogram is given in the following steps.
7.4. CLUSTERING METHODS

1. Merge the branches of the clusters with the smallest dissimilarity.
2. Update the dissimilarities relative to the new cluster.
3. Repeat steps 1–3, until all points are combined in a single cluster.

Linkage Methods

To generate the dendrogram we require a linkage criterion to determine how branches should be fused. Selecting an appropriate linkage method for the desired application is essential in order to produce a meaningful clustering. In this section, we introduce a range of standard linkages between two clusters, \( C_k \) and \( C_{k'} \), and define the dissimilarity, \( d_{lw}(C_k, C_{k'}) \).

**Single (Nearest Neighbour):** This dissimilarity is given by the minimum pairwise distance

\[
d_{lw}(C_k, C_{k'}) = \min \{d^*(x_k, x_{k'}) : x_k \in C_k, x_{k'} \in C_{k'} \}.
\]

(7.14)

**Complete (Furthest Neighbour):** This dissimilarity is given by the maximum pairwise distance

\[
d_{lw}(C_k, C_{k'}) = \max \{d^*(x_k, x_{k'}) : x_k \in C_k, x_{k'} \in C_{k'} \}.
\]

(7.15)

**Group Average:** This method is also known as the unweighted pair group method with arithmetic mean (UPGMA) and the dissimilarity is given by

\[
d_{lw}(C_k, C_{k'}) = \frac{1}{|C_k| |C_{k'}|} \sum_{x_k \in C_k} \sum_{x_{k'} \in C_{k'}} d^*(x_k, x_{k'}).
\]

(7.16)

**McQuitty:** This method is also known as the weighted pair group method with arithmetic mean (WPGMA). Let clusters \( C_i \) and \( C_j \) be combined to form cluster \( C_k \), the dissimilarity is given by

\[
d_{lw}(C_k, C_{k'}) = \frac{1}{2} \left( \frac{1}{|C_i|} \sum_{x_i \in C_i} \sum_{x_{k'} \in C_{k'}} d^*(x_i, x_{k'}) + \frac{1}{|C_j|} \sum_{x_j \in C_j} \sum_{x_{k'} \in C_{k'}} d^*(x_j, x_{k'}) \right).
\]

(7.17)

There are many linkage types present in the literature. Other common linkages include centroid, median and Ward. These methods are different from those presented, as the dissimilarity between branches is defined relative to a cluster centre. Within Euclidean space the cluster centres have a simple interpretation. However, it is less clear how the linkages generalise to F-madogram space, and whether coded implementations of these linkages utilise Euclidean assumptions. As such we do not consider these linkages here.
For further details on linkage methods refer to Müllner, (2011), Murtagh, (1983), and Murtagh and Legendre, (2014). All of the linkages mentioned in this section are implemented in the core R package, stats (R Core Team, 2017).

7.4.4 Suitability of Hierarchical Clustering

We have shown that $K$-medoids clustering performed with an F-madogram dissimilarity can produce undesirable clustering results. Hierarchical clustering can be used to avoid the issues identified. To demonstrate this, we reproduced the two examples given in Section 7.4.2 using hierarchical clustering with an average linkage, see Figures 7.7 and 7.8. We observe that hierarchical clustering with an average linkage successfully recovers the two original groups in both examples.

![Capped Distance Example](image)

Figure 7.7: The example of Figure 7.5 replicated using hierarchical clustering and the average linkage method.

For our application, hierarchical clustering is the preferred method to $K$-medoids. The spatial density of the Australian station network is highly variable. We therefore require a method that is not sensitive to station density and that is not affected by the bounded range of the F-madogram dissimilarity.
7.5 Clustering Choices

7.5.1 Linkage Method

To ensure we use an appropriate linkage method for our application, we applied a range of standard linkages to the stations in Tasmania, see Section 7.4.3. The resulting dendrograms are shown in Figure 7.9. Of the linkage types considered, not all produced suitable dendrograms.

Despite the single linkage having a nice physical interpretation, the dendrogram exhibits chaining. Chaining occurs when branches are connected sequentially and this means that the partitioning of the points is not informative. The single linkage is therefore not appropriate for our application.

Complete linkage also failed to produce appropriate clusters. This is a consequence of the choice of F-madogram dissimilarity. The F-madogram has a maximum value of 1/6, therefore all branches are connected as soon as the maximum distance is reached. This causes crowding, where compact clusters are generated. In our case, these clusters are too small to provide an appropriate partition. However, we can use the number of clusters produced by a cut height of 1/6 to help inform the maximum number of clusters for other linkages.

The dendrograms generated using the linkage methods of average and Mc-
Figure 7.9: Examples of dendrograms generated using different linkage types. Dendrograms are generated from stations in Tasmania using an F-madogram dissimilarity.

Quitty are appropriate for partitioning our stations. For the Tasmanian station network, the two trees are morphologically similar.

7.5.2 Dissimilarity

Unsupervised learning algorithms produce vastly different clusters depending on the dissimilarity used. Consider the two examples of hierarchical clustering as applied to stations in Southwest Western Australia, Figure 7.10. In these examples, one used Euclidean distance and the other used the F-madogram dissimilarity.

Of particular note, is that the clusters produced using the F-madogram dissimilarity are largely homogeneous in Euclidean space. This is remarkable as only the observed annual maximum rainfall at each of these stations was used to estimate the dissimilarity. No external information of climate or topography was used. This speaks to the power of this method for identifying regions of similar extremal dependence.
7.6 Classification

We wish to emphasise that the formation of geographically homogeneous sub-regions is done well under hierarchical clustering. However, the partition induced by hierarchical clustering is created using the F-madogram and

The F-madogram dissimilarity also produces clusters of different shapes. This is in contrast to the clusters formed using Euclidean distance, which are roughly equal in size and shape. Of particular note is that for the F-madogram dissimilarity, we observe clusters which are single points. Given the cut height was 0.14, this means these single points more than 0.14 away on average from stations in the nearby clusters. This type of clustering behaviour is desirable for our application. We do not want to group stations that do not have similar dependence.

Figure 7.10: Comparison of the hierarchical clustering using two different dissimilarities: Euclidean and F-madogram. The number of clusters was chosen by considering both dendrograms. The cut heights shown in the sub-figure labels correspond to this number of clusters.
we require this partition to be in Euclidean space. An additional classification step is therefore required to convert the clusters formed using the F-madogram dissimilarity back into Euclidean space. We can then create a regionalisation, by identifying regional boundaries and classifying locations without stations. Classification also helps to identify stations whose clustering is not geographically consistent. This can occur when a station is at a high elevation relative to surrounding stations, or if the quality of the data is poor.

To partition the domain in Euclidean space, we first grid the domain and then use a weighted $k$-nearest neighbour classifier ($wk$-NN) (Dudani, 1976) to classify the grid points. We chose $wk$-NN as it is non-parametric, it is based on minimal assumptions, and it can form non-linear boundaries. This is in keeping with our earlier modelling choices. For a prespecified $k$ value, the classification algorithm of $wk$-NN works as follows. Let $x$ be the point we wish to classify and let $c_i$ be the cluster label assigned to station $x_i$.

1. Identify the $k + 1$ nearest stations (neighbours) to $x$ in Euclidean space.
2. Standardise the distances of the closest $k$ neighbours to $x$ using the distance to the $k + 1$ neighbour
   \[ d_s(x, x_i) = \frac{\|x - x_i\|^2}{\|x - x_{k+1}\|^2}. \]  
   (7.18)
3. Transform the normalised distances using a kernel function to determine the weights. We have chosen an inverse weighted kernel
   \[ w_i = \frac{1}{|d_s(x, x_i)|}. \]  
   (7.19)
4. Classify $x$ according to the weighted majority of $k$ nearest neighbours,
   \[ c = \arg\max_{c \in \mathcal{C}} \left( \sum_{i=1}^{k} w_i I(c_i = c) \right), \]  
   (7.20)
   where $\mathcal{C}$ is the set of classifications assigned to the $k$-nearest neighbours.

To perform the classification we used the R package, kknn, (Schliep and Hechenbichler, 2016). Further details about the package are given in (Hechenbichler and Schliep, 2004). We were cautious not to classify locations where there was insufficient information about nearby rainfall. We therefore constrained our classification by only classifying points that were within 0.5 degrees of a station.

There is a variance bias trade-off when selecting the number of neighbours, $k$. We needed to select $k$, such that erroneous points do not impact the
7.7. VISUALISING DEPENDENCE

classification, and smaller clusters of only a few stations were not engulfed by a larger cluster and its label. We found the value, \( k = 15 \) to provide a suitable balance between our competing objectives.

We note that in standard \( k \)-nearest neighbour classification (\( k \)-NN) (see Hastie et al., 2009), points are classified similarly to the majority of their \( k \) nearest neighbours without using weights. Given that we are translating insights from F-madogram space to Euclidean space, and the relationship between these two distances is non-linear a weighted classifier is more appropriate for this application.

7.7 Visualising Dependence

To ensure that the partitioning of Australia is an appropriate regional summary, we can fit max-stable models to the stations in each region. The max-stable model we used is the Smith model (see Section 5.3.2 and equation (5.19)). The Smith model is a canonical example of a max-stable process and we have chosen to use it in this instance for its simplicity. The dependence structure of this process is Gaussian, so we can visualise the dependence structure in two-dimensional space using ellipses. The direction and size of these ellipses has a natural interpretation in terms of anisotropy and the range of the dependence.

We centre the ellipses on the median longitude and median latitude of all suitable stations that can be used for fitting. We denote this centre, \( \mathbf{m}_k \). The ellipse is given by the parameterisation

\[
\mathbf{x} = r(\cos \theta, \sin \theta)\mathbf{M} + \mathbf{m}_k, \tag{7.21}
\]

where \( \mathbf{M} \) is obtained from the Choleski decomposition of the covariance matrix, \( \Sigma = \mathbf{M}'\mathbf{M} \). The value of \( r \) can be chosen to correspond to a desired level curve.

The probability of a point, \( \mathbf{x} \), lying within distance, \( r \), of the mean of the bivariate normal distribution is given by the Chi-squared distribution with two degrees of freedom

\[
P(||\mathbf{x} - \mu|| < r) = 1 - \exp \left( -\frac{r^2}{2} \right). \tag{7.22}
\]

We chose \( r \) to correspond to the 1\% level curve, for which \( r \approx 3 \).

In general, if the partition is a good representative summary for Australia, then we expect that the ellipses not to overlap. If the ellipses were to overlap, this could indicate that points in the intersection could reasonably have been
assigned to either cluster, or that we may have used too many clusters. If we have too few clusters, then we should be able to add in further ellipses to summarise dependence of these regions.

7.8 Practical Considerations

7.8.1 Missing Dissimilarities

If two stations have recording periods that do not overlap due to missing data, or if the number of overlapping years is small, the F-madogram dissimilarity cannot be estimated via equation (7.7). We require all pairwise dissimilarities for stations used in clustering. To maximise the stations available, particularly in sparse regions, we interpolated missing dissimilarities so that the stations could still be clustered.

At large Euclidean distances, we expect the maximum rainfall observed at pairs of stations to be independent. Given this, if the Euclidean distance between stations exceeds 1.5 degrees, any missing F-madogram dissimilarity is set to 1/6. This is a reasonable assumption and greatly reduces the missing dissimilarities. If the F-madogram dissimilarity is missing and the Euclidean distance between stations is 0, then the F-madogram dissimilarity is set to 0. This situation occurs commonly if the station has been renamed.

For the remaining missing dissimilarities we used a crude estimator. For all pairs where the Euclidean distance was less than 1.5 degrees, we fit a linear model to the log-transform of the Euclidean distance. A plot of a fitted model is given in Figure 7.11. These predictions do not approximate local dependence well, but they serve as a reasonable approximation of overall dependence.
7.8. PRACTICAL CONSIDERATIONS

We observe from Figure 7.11, that as $x \to 0$, the function approaches $-\infty$ and the fitted model can take negative values. We therefore took the estimator to be the maximum of the predicted F-madogram dissimilarity and zero.

There is a trade-off when interpolating missing dissimilarities. The higher the required minimum number of observations at each station, the fewer stations we have for clustering. The fewer years of maxima at a station, the more missing dissimilarities that need to be interpolated. If there are only a few missing values relative to the bulk of the distances, then the clustering is robust to infilling.

7.8.2 Model Fitting

To fit the Smith model and plot ellipses to summarise dependence, we used composite likelihood, equation (5.34). In composite likelihood, instead of using the full density, the optimisation is performed by summing over the bivariate density (see Section 5.4.1). Pairs of stations therefore need to have annual maximum observations in common years. In many instances this is not the case, and weights need to be set to zero in the likelihood to exclude the pair. Note we also modelled the dependence separately to the marginal distributions, transforming our data to Fréchet marginal distributions, equation (5.15).
For fitting, we carefully considered which stations were suitable for use in modelling. For example, it is possible for two disjoint regions to have the same classification. In this case, using all stations for fitting is unlikely to yield a meaningful model and the disjoint sub-regions should to be modelled separately.

7.9 Regionalisation Results

To choose an appropriate cut height we created visual summaries of different possible regionalisations (see Figures 7.12, 7.14, 7.16, 7.18). This helps us to understand how a cut height in the dendrogram induces geographic separation. Regional summaries are given for each of the NRM super clusters (see Section 7.2). Considering the NRM super clusters helps us to understand the regional implications of our modelling decisions and how these insights can be translated to an Australia wide-scale. As the NRM super cluster of Southern Australia is geographically separated, we further split this cluster into three: Tasmania (TAS), Southeast Australia (SEA) and Southwest Western Australia (SWWA).

In the summary plots provided, the strip label gives the cut height and the corresponding number of clusters. Note that the number of regions visible may not be the same as the number of clusters given in the strip label. This may be because regions can be small, consisting of a single point, or because small clusters of stations can be subsumed into a larger cluster under the classification. The colouring of the visual summary is done simply to help distinguish between different regions and to indicate the formation of new regions between sub-figures.

In each summary, the first few sub-figures were generated using low cut heights. As a consequence these cut heights generally produce too many regions. However, these figures serve to help us understand which regions are strongly dependent. We can then develop an understanding of the finer regional details and knowledge of which smaller regions were amalgamated at higher cut heights. In contrast, the last few sub-figures are at a higher cut heights and generally show too few regions. These sub-figures highlight the most important features that serve to create regional boundaries.

For the each of the larger regions we considered, we also provide an example of regionalisation overlaid on a satellite image. This helps us to contextualise the regionalisation relative to orography and vegetation features, such as grasslands and dense forest. We note the projection system of our region boundaries and the satellite images are different, but this difference in projection is negligible and does not alter the interpretation. In each of
these satellite figures, we also include the station locations. This provides additional information about the formation of region boundaries and how these boundaries were influenced by station locations.

7.9.1 Tasmania

Regionalisations of Tasmania for different cut heights are given in Figure 7.12. The regionalisation for the cut height that was closest to 0.115 was overlaid on the satellite image in Figure 7.13. Both of these figures contribute to determining a suitable regionalisation for Tasmania.

At high cut heights, the Central Highlands is identified as an important feature that helps to regulate extreme rainfall. The Central Highlands is a mountain range located in Central Tasmania. The regionalisation of Tasmania is therefore strongly influenced by orographic features.

For the middle row of summary figures, we observe small regions in Central Tasmania. This area is where the Central Highlands meets an area of lower elevation that extends into Southeast Tasmania. Given the size of these regions and the cut height, we should exercise caution in assuming a single dependence structure here.

For lower cut heights we observe more regions in coastal areas than at higher cut heights. This is true along the coast of Northern Tasmania and on the Eastern Coast of Tasmania around Hobart, (147.37, -42.88). For example near Hobart we observe two regions paralleling the coast. At higher cut heights these regions are amalgamated into a larger region. We demonstrate in the next Section, where we show the results from our fitted max-stable models, that the dependence of these coastal regions and inland regions should be treated differently.

Western Tasmania and the Southern tip of Tasmania are densely forested and mainly National Parks. In these areas, there are only a few stations. Despite this, smaller regions are formed in the West that consist of only a few stations and this shows that dependence is different in these areas. We should therefore consider regions in the West to be different to regions in the East, and respect the boundary created by the Central Highlands.

Under the NRM sub-clusters, Tasmania is split East–West by the Central Highlands orography (see Figure 7.4). Our regionalisations suggest that further sub-classification is required. An East–West split oversimplifies the complexity of the Tasmanian domain. This is particularly evident at low cut heights. This view is consistent with Grose et al., (2010), who advocated that
Figure 7.12: Hierarchical clustering of Tasmania with an average linkage.
7.9. REGIONALISATION RESULTS

Figure 7.13: Satellite Image of Tasmania. The overlaid regionalisation is for the cut height closest to 0.115.
Tasmania be split into several regions, not just two.

Based on these summary plots, a cut height of approximately 0.115 to 0.133 appears to produce the best regionalisation. For this range of cut heights, we capture the most important geographic and climatic features of Tasmania. Depending on the application, an appropriate cut height can be selected in this range so as to reflect the required level of detail.

### 7.9.2 Southwest Western Australia

Regionalisations of Southwest Western Australia for different cut heights are given in Figure 7.14. The regionalisation corresponding to the cut height closest to 0.133 was overlaid on a satellite image, shown in Figure 7.15.

For Southwest Western Australia, we observe that the predominant orientation of inland clusters is Northwest to Southeast. This could be attributed to Western fronts, which approach from the Southwest and move Northeast (Risbey et al., 2009).

At lower cut heights we observe that the shape and scale of the inland clusters is similar, and at higher cut heights these regions are combined. Given this, we explore whether it is reasonable to assume similar dependence of rainfall extremes in inland Southwest Western Australia with our fitted models.

We observe a demarcation between the shape and orientation of regions in coastal areas compared with inland areas. In coastal areas the regions are orientated parallel to the coastline. The elevation of these areas is also lower. It is intuitive that extreme rainfall in coastal regions should be coastally driven, and the regionalisation reflects this.

Of note is the form of the regionalisations near the coordinate, (115, -30). The stations in this region are weakly dependent in relation to other stations, and are therefore clustered poorly in terms of geographic homogeneity. We are unaware of a climatic or topographic reason for why this is the case.

In Southwest Western Australia, there is far less topographic variability compared with Tasmania. As such, the range of cut heights that capture the important features is much smaller, with cut heights close to 0.133 producing the best regionalisations.
Figure 7.14: Hierarchical clustering of Southwest Western Australia with average linkage.
Figure 7.15: Satellite Image of Southwest Western Australia. The overlaid regionalisation is for a cut height closest to 0.133.
7.9.3 Southeast Australia

Regionalisations of Southeast Australia for different cut heights are given in Figure 7.16. In Figure 7.17, the regionalisation corresponding to the cut height closest to 0.115 is shown overlaid on a satellite image.

Regions in Southeast Australia are strongly affected by the Great Dividing Range, as shown by regional boundaries at highest cut heights. The Great Dividing Range is a mountain range that spans the entire coast of Eastern Australia. We generally observe that regions are formed either side of the range. At lower cut heights, when there is stronger dependence within regions, we observe more nuanced effects in the location of region boundaries, particularly around the range and in coastal areas.

Of note is that there are unique coastal features in Southeast Australia. For example, there is a bay around Melbourne (144.96, -37.81). For low and medium cut heights, regions encircle the bay and the geographic separation of stations on either side of the bay is respected. There are also three peninsulas near Adelaide (138.60, -34.93). The regions created around Adelaide also respect the geographic separation of stations on each of these peninsulas for low and medium cut heights.

It is worth noting that there is a large region inland, (142.5, -36), that is most obvious at high cut heights, but also persists for a range of medium cut heights. Stations in this region therefore experience extreme rainfall events of similar sizes. From our visualisation of other regions, we observe that in general the scale of dependence of the observed annual maximum rainfall is much smaller. This region is inland and of low elevation. Other than this, it is not immediately clear whether there exists a climatic or topographic reason for a region to be of this size.

For this visualisation, suitable cut heights range from approximately 0.115 to 0.133. At the lower end of this range, much finer detail can be observed given the orography and in coastal areas. Therefore it depends on the application, as to which cut height is best suited.

7.9.4 Eastern Australia

Regionalisations of Eastern Australia for different cut heights are given in Figure 7.18. The regionalisation corresponding to a cut height nearest 0.115 was overlaid on a satellite image, shown in Figure 7.19.

Similar to Southeast Australia, we observe the influence of the Great Dividing Range on the location of regional boundaries. Regions are formed
Figure 7.16: Hierarchical clustering of Southeast Australia with average linkage.
7.9. REGIONALISATION RESULTS

Figure 7.17: Satellite Image of Southeast Australia. The overlaid regionalisation corresponds to the cut height that is closest to 0.115.
Figure 7.18: Hierarchical clustering of Eastern Australia with average linkage.
Figure 7.19: Satellite Image of Eastern Australia. The overlaid regionalisation is for the cut height closest to 0.115.
either side of the Great Dividing Range. The general orientation of regions on the coastal side of the range is observed to parallel both the coastline and the range.

We note that the size of the regions on the coastal side of the range, particularly at lower cut heights, is generally larger than for clusters further inland. Further inland clusters appear more circular and of a similar size.

For this visualisation, suitable cut heights range from approximately 0.115 to 0.133. Similarly to Southeast Australia, at the lower cut heights, much finer detail can be observed given the orography and along the coast.

7.9.5 Northern Australia and Regional Australia

In Northern Australia and regional Australia, many stations are sparsely located and clustered individually. As such, the clustering and resulting regionalisation is only meaningful where there are sufficient stations, such as in far North Queensland, parts of inland Queensland, and Northwest Western Australia. Given this, the summary plots for Northern Australia and Regional Australia are given in Appendix B. As the stations are sparsely located in these regions, generally using a cut height greater than 0.133 is more appropriate.

7.9.6 General Comments

Hierarchical clustering with an appropriate linkage is robust to variations in station density. We showed this in the examples given in Section 7.4.4. However, we observe that for each of the different NRM clusters there is no single cut height that is suitable for all regions. Cut heights in the range of 0.115 to 0.133 seem broadly appropriate. The top end of this interval provided the best height for all of the NRM regions considered.

This range of cut heights makes intuitive sense. Stations with a dissimilarity above 0.133 are very weakly dependent even when separated by large geographical distances, Figure 7.11. Therefore for cut heights above 0.133, we group stations that do not belong together. From Figure 7.11, we also observe at small Euclidean distances the graph changes rapidly. Therefore for cut heights that are less than 0.115, too fine a partition tends to be produced. Therefore the dissimilarities in between 0.115 to 0.133 provided a suitable range, capturing the change from very strong dependence to very weak dependence.

We also observe based on the satellite imagery that the location of regional boundaries is affected by station locations. There are clearly regions where the
boundary of the region is stretched over orographic detail due to an absence of station data. A clear example of this is in Northern Tasmania, see Figure 7.13. In this example all the stations are located in lower elevation regions and there are no stations in the higher elevation regions. As a consequence, although our regionalisation respects the distinction between regions of different elevations, the boundary formed encroaches on the true geographic boundary. Results therefore need to be interpreted judiciously.

7.10 Visualising Dependence

An important part of this research is understanding whether it is suitable to group a set of stations and assume a single dependence structure. To this end, we considered the regionalisations produced by the cut heights, 0.115 and 0.133, and we fitted Smith models in each region. To understand the uncertainty of the estimated dependence parameters, we repeated the model fitting by sub-sampling stations in each region. Where appropriate we took 30 samples of 25 stations. If the number of stations was too few, but there were still sufficient stations for fitting, we used a jackknife sample, which is a leave one out approach to sampling.

We note that the optimisation of the likelihood may not always converge. When this occurs, generally optimisation algorithms will produce an error warning. A greater concern is when the optimisation converges to a local maximum instead of a global maximum, and a warning is not provided. For example, this can occur at the boundary of the parameter space.

We can identify some of the instances when the global maximum has not been reached. For example we might also be suspicious of convergence if the ratio between the two elliptical axes is large or if the parameter estimates are vastly different to the fitted values from the other samples. In these instances, we repeated the model fitting with starting values to help improve the performance of the optimisation.

We note that the number of ellipses shown in each region may differ depending on whether the optimisation converges for each sample. Additionally, we were cautious not to bias the results by discarding model fits that looked suspicious. Therefore, after providing starting values, we retained all fitted models that appear to have converged.

In the following sections, we consider the regions of Tasmania and Southwest Western Australia in detail. We then provide a plot showing the ellipses for all of Australia at both cut heights. Note that each elliptical curve corresponds to one sample of stations and one fitted model. We can be confident
that a single dependence for a region is a reasonable assumption if the size and direction of all the elliptical curves is similar. If the elliptical curves have different sizes or directions this is an indication that either the models failed to converge, or the stations sampled have different dependence, or both. Given the ellipses represent level curves 1% level curves, if the regionalisation provides a good summary of dependence with the domain, then the ellipses will be tightly packed with minimal overlap.

**Dependence in Tasmania**

The fitted elliptical curves for the regionalisations of Tasmania at cut heights of 0.115 and 0.133 are given in Figures 7.20 and 7.21. At the higher cut height regions contain more stations, therefore there is more variability in the stations sampled and within the fitted elliptical curves.

At the lower cut height, Figure 7.20, we observe that orientation and the range of the elliptical curves seems reasonable in terms of our expectation of dependence. In general the ellipses partition the domain well, with minimal overlap. For regions where the ellipses do overlap, this suggests that it is not possible to consider the dependence of extreme rainfall in these regions independently of one another.
7.10. VISUALISING DEPENDENCE

We observe that for regions in Northern Tasmania, the range and orientation of ellipses is similar at both cut heights. This is not the case for Eastern Tasmania. At the higher cut height, there is no consensus in the structure of the fitted ellipses on the Eastern Coast. Comparing this with Figure 7.20, it is clear that we cannot assume a common dependence structure for these larger regions and there is a good reason for the elliptical curves to be so different at the higher cut height.

Dependence in Southwest Western Australia

The fitted elliptical curves for the regionalisations of Southwest Western Australia, at cut heights of 0.115 and 0.133, are given in Figures 7.22 and 7.23. We observe that at the higher cut height the regionalisation partitions the dependence in this region well. At the lower cut height, the size and orientation of the fitted ellipses is similar; however, the elliptical curves overlap. This suggests that this cut height is too low, and many of these regions should be modelled together given the similarity of their fitted dependence structures.

Examining some of the finer detail in Southwest Western Australia, we make the following observations. Around the coordinate (116.75, -30), there are sufficient stations for fitting, but at both cut heights the optimisation did not converge. Further investigation is required here. At the high cut height,
Figure 7.22: The fitted Smith models for regions in Southwest Western Australia. The regionalisation is given by the cut height closest to 0.115.
Figure 7.23: The fitted Smith models for regions in Southwest Western Australia. The regionalisation is given by a cut height closest to 0.133.
there are also small groups of stations that are not covered by ellipses. At the lower cut heights, some of these smaller dependence structures are resolved. However, there are also regions where the dependence appears highly variable and a max-stable model would be inappropriate, such as the Southern tip of Western Australia. Around Perth (116, -32), we also observe no consensus in the fitted models at either cut height. Rottnest Island is located off the coast of Perth and this may contribute to the uncertainty of the dependence in this region.

**General Comments on Dependence**

The types of observations we have made for Tasmania and Southwest Western Australia, are broadly similar for Eastern Australia, Southeast Australia and Regional Australia. In Northern Australia, there is insufficient station data except in a few regions for this type of analysis. The plots of elliptical curves for all Australia, at cut heights of 0.115 and 0.133, are shown in Figures 7.24 and 7.25. We make two core observations here. Firstly, despite the station density, there is a failure to fit models along parts of the Great Dividing Range in Eastern Australian and Southeast Australia. This suggests a single dependence structure in regions with orography is not a good assumption. We also observe that the ellipses are too large and clearly spurious for inland regions where the station data is sparsest, and the climate is dry. An example of this are some of the ellipses in Northwest Western Australia. In these regions, we have to question whether a GEV distribution is suitable to model the data at all. We would caution interpreting anything meaningful from the ellipses in these drier areas, particularly those of large size.
The cut height is approximately 0.115 for the Smith model.
The cut height is approximately 0.133 for each of the NRM super-clusters. The dependence of the Smith model is given by the cut height closest to 0.133.
7.11 Conclusions

We have created a regionalisation based on the dependence of rainfall extremes using hierarchical clustering and the F-madogram dissimilarity. The regions created are broadly consistent with our understanding of climate and topographic features (Risbey et al., 2009; Stern et al., 2000). This is impressive, as only the observed, daily annual maxima were used to estimate the dissimilarity.

Climate scientists, hydrologists and other researchers can use our regionalisations for inference about the behaviour of rainfall extremes in Australia. Depending on the application, the cut height can be altered to resolve different levels of regional detail. The NRM clusters (CSIRO et al., 2015) were in part designed for this purpose; however the NRM regions were not designed specifically for rainfall or for extremes. We therefore advocate using our regionalisation for extreme rainfall applications.

The regions identified improve our understanding of local dependence of rainfall extremes. By fitting max-stable models, we were able to quantify the range of dependence and direction of anisotropy for each region. We observed that there are many and varied dependence structures for rainfall extremes throughout Australia. We also observe that the range of dependence, as shown by the elliptical curves, is generally small and of the order of tens of kilometres, not hundreds.

In addition to presenting the regionalisation, we highlighted key methodological considerations when using the F-madogram dissimilarity for unsupervised learning. The F-madogram dissimilarity can produce spurious clustering, depending on the underlying station network and the clustering method used. This motivated using hierarchical clustering instead of $K$-medoids for our application.

We discussed some standard linkage methods used in hierarchical clustering; however there are many other linkages available in the literature. Examples include minimax, a radius based linkage (Bien and Tibshirani, 2011) and geoClust, a geoconstrained linkage (Chavent et al., 2017). We leave examining other linkages and considering their suitability for future research.

Choosing the linkage was one of many modelling decisions that contributed to the form of the regionalisation. For example, we are also required to select a cut height and the number of nearest neighbours used in classification. These choices were largely informed by visualisation, cluster cohesion and the coverage of the elliptical level curves. In future work, we would like to design a metric that allows for cross-validation of these chosen parameters.
We also acknowledge that there are other max-stable processes that we could have used to model the dependence of rainfall extremes (see Section 5.3.3). However, the computational implementation of other max-stable processes in the SpatialExtremes package (Ribatet, 2015) does not include anisotropy, and we have shown anisotropy to be important. In future work, we could use an alternative package for fitting that included anisotropy, for example spatialTailDep (Einmahl et al., 2016; Kiriliouk and Segers, 2014). However, as we only required the max-stable model to help visualise the dependence and to validate the partitioning of the domain, the Smith model was appropriate for our application. If more detailed cluster level inference or simulation is required, we would recommend using a different max-stable process.

The future directions of this research are many and exciting. We have produced a body of code to create regionalisations, see the package clusterExtremes hosted here https://github.com/katerobsau/clusterExtremes. The code can therefore be applied to consider entirely different regions. By using different maxima, such as seasonal maxima, we can create different regionalisations of Australia. This would further our understanding of how the dependence of extremes may vary temporally, in addition to spatially.

Developing this regionalisation has helped to improve our understanding of the validity of the assumptions we make when modelling dependence. By sampling station and fitting max-stable models, we determined in which regions it is appropriate to group stations and assume a single dependence structure. Conversely, we also have an understanding of which stations should not be modelled together, and which stations can be modelled independently of one another.

There is scope to use this understanding of dependence, to inform simulation of adjacent regions with different dependence behaviour. Where appropriate, we can simulate one region, conditionally on a boundary of another. The conditional simulations would be simplified by the fact that the partition would be known and inherited from the simulations in the first region (see Section 5.5.2). In addition, we can use known information about climate and topography to help inform the order in which the regions are simulated. This process would be highly parallelisable. Caution would be needed to ensure any numerical errors do not propagate. We have already experimented with conditional simulation based on this regionalisation to good effect.

We hope to use our regionalisations to develop statistical models on both regional and local scales for Australian rainfall extremes. When we started
this research, this goal was aspirational. However, given the knowledge we have generated about dependence of rainfall extremes in Australia, this is now a very tangible direction for future research.
Chapter 8

Conclusions

8.1 Major Findings

This thesis improves upon our current understanding of the probability of extreme rainfall in Australia and the associated risks. For each of the applications presented, we utilise different methodology from extreme value theory. This methodology builds in complexity throughout the thesis. Each modelling approach, implicit with its assumptions and uncertainties, therefore provides a different perspective on the challenge of modelling rainfall extremes. Below we outline the contribution of each chapter to this field of research.

We began by discussing the quality issues of daily rainfall observations, and how these issues can affect statistical analyses of rainfall extremes (see Chapter 2). All practitioners who use observational records are forced to make decisions around how to handle these quality issues; however, there are no standard methods for preprocessing. This makes the use of observational data prohibitive for some researchers. We created a dataset of observed annual and seasonal maxima in Australia (hosted here https://github.com/katerobsau/RaingleExtremes). This dataset identifies some of the more insidious quality issues that affect extremes. We have therefore helped to remove an obstacle that previously prevented researchers from using raw observational data for the analysis of Australian rainfall extremes.

We then examined the validity of an assumption that is commonly used in applications involving extreme value theory, namely stationarity (see Chapter 4). Large scale climate drivers are known to influence the distribution of Australian rainfall (Risbey et al., 2009). We quantified the magnitude of the influence of the large scale driver, El Niño Southern Oscillation (ENSO), on seasonal maxima in Australia and identified the geographic range of
influence. To improve on our ability to detect whether non-stationarity was statistically significant or not, we fitted a spatial Generalised Extreme Value (GEV) distribution to station data instead of a univariate GEV distribution. We found our results, which used observational data, are broadly consistent with studies that use gridded products (Min et al., 2013).

Extending these ideas, we used max-stable processes to understand the influence of ENSO on the spatial distribution of rainfall extremes in a case-study region of Southeast Queensland, Australia (see Chapter 6). As part of this application, we estimated the probability of an extreme rainfall event exceeding the amount received during the Toowoomba flash flood in 2011. Our ability to estimate this probability reliably was improved by accounting for dependence and by simulating the extreme precipitation field, both of which were possible using max-stable processes. We considered non-stationarity in this modelling by including covariates for ENSO in the marginal parameters, with the form of these covariates determined through model selection. We found an event exceeding that of the 2011 flash flood was 85% more likely during La Nina phase of ENSO, compared with a neutral phase. This has serious implications for decision makers in disaster mitigation.

This application showed that max-stable processes are a powerful statistical tool that can be used to generate insights about the probability of extreme rainfall; however, when using these processes we commonly assume that the dependence is fixed across the modelling domain. In Chapter 7, we explored the validity of this assumption and showed that dependence of annual maximum rainfall observations is highly variable and highly localised throughout Australia. To identify which stations are suitable to be grouped and modelled under the same dependence structure, we created a regionalisation. The regions formed are based on the bivariate extremal dependence between the observed annual maxima at pairs of stations. No information about climate or topography was used to create the regionalisation. Impressively though, even without this information the resulting regions are spatially homogeneous and respect climatic and topographic features that need to be considered when modelling extreme rainfall. This work has improved our understanding of local rainfall extremes and can be used by other researchers to inform their applications involving extreme rainfall.

8.2 Extensions

In the following sections, we outline some extensions to the body of work presented in this thesis. We begin by providing some general comments and then provide chapter specific extensions.
8.2. EXTENSIONS

General Comments
In Chapter 3, we gave an introduction to extreme value theory. We explained how to use this theory for statistical modelling of extreme rainfall and discussed modelling assumptions. Here we consider some of the subtleties regarding these assumptions.

In our modelling, we approximated the distribution of the finite sample maxima by a limit distribution. In general, for larger block sizes we have greater confidence in the suitability of this approximation. In future work, it may be worth considering whether there is any advantage to come from considering pre-asymptotic behaviour and using a penultimate distribution instead of the limit distribution (Gomes and Guillou, 2015).

We also assumed that the parameters of the limit distributions vary smoothly in time. This allowed us to model non-stationarity by introducing temporal covariates in the parameters of the distribution. This use of covariates was targeted at modelling temporal changes in intensity of the sample maxima. However, changes in frequency and in intensity of rainfall extremes may be different, and therefore should be modelled differently. In future work, it may be worth considering whether the number of wet days per year is varying, and how this may affect the suitability of our modelling assumptions.

It is also important to remember that rainfall is a physical process and it is therefore subject to physical constraints. For hydrologists, for a given period and under certain meteorological conditions, there is a theoretical limit on the peak maximum precipitation (PMP). However, for distributions with an infinite support, we would assign a non-zero probability for rainfall exceeding this theoretical physical limit. In future work we would like to examine how the limit theory for extremes relates to the theoretical physical limit, and whether adaption is needed within our statistical modelling.

Extreme value theory is a powerful modelling tool for applications of climate and hydrology. However, it is prudent to remind ourselves given the modelling assumptions that ‘All models are wrong but some are useful’ - George Box.

Chapter 2: Observation Quality
In Chapter 2, we took steps to identify the core quality issues that affect the statistical analysis of rainfall extremes. However, there are many other additional quality issues that need assessment and targeting. Some of these issues include shifts in observational dates, errors in elevation metadata and
errors in record homogeneity. Additionally, we are also concerned with the effectiveness of existing quality checks. These include the outlier check and the spatial consistency check implemented within GHCN-Daily (Durre, Menne, Gleason, et al., 2010) (see Section 2.2.2). The test we used for detecting bias in the days that extreme events were observed could be extended to a Chi-square test for categorical data. In future our intention is to update these quality assurance methods; however as we do so, it is essential that any scientific claims that have been built upon earlier versions of the data are reproducible. Embracing methods from reproducible research, such as time-stamps, version control and appropriate documentation, will help meet these needs.

Chapter 4: Seasonal ENSO Influence

In Chapter 4, we discussed detecting and quantifying non-stationarity of seasonal rainfall extremes due to ENSO. The approach we adopted, first gridded the domain and then fit a spatial GEV distribution to the stations in each gridded cell. This modelling decision was made to ensure model parsimony, and so we could compare our results to studies using gridded products. In future work, we do not need to constrain our fitted model to consider gridded cells, and instead can fit models over much larger domains. This will further improve our ability to detect non-stationarity. We are also interested in examining the interplay between climate drivers, and how this affects seasonal non-stationarity.

Chapter 6: Max-stable Process Application

In Chapter 6 we considered an application of max-stable processes. Initially, we performed some exploratory analyses to check whether dependence might vary across our modelling domain. However, after producing our regionalisation in Chapter 7, there is now scope to revisit our original assumptions with new insights about dependence and model fitting. In revisiting the application, we would aim to improve our understanding of parameter uncertainty, whether it be of dependence or of marginal parameters. This would in turn improve our understanding of how uncertainty influences the conclusions of this chapter.

Chapter 7: Regionalisation

In Chapter 7, we presented a regionalisation of Australia based on the extremal dependence of observed, daily, annual maximum rainfall. As part of this research, we established a body of code that can be used to repeat the analysis using different maxima.

In using different maxima, we can begin to understand how dependence of rainfall extremes may vary temporally, as well as spatially. We would be
particularly interested in the differences between regionalisations and fitted models if we were to compare the following types of maxima:

- seasonal maxima, including comparing Summer, Autumn, Winter and Spring; and wet and dry seasons;
- different aggregation periods, for example sub-daily, daily, and 5-day aggregated totals;
- stratifying the maxima to consider possible anthropogenic influences, for example we could stratify maxima prior to 1960 and after 1960;
- stratifying the maxima by anomalous years of a given climate driver, for example to understand the influence of ENSO we could stratify maxima by positive SOI years and compare with negative SOI years;
- combinations of the above.

There is also scope to determine whether gridded data products are accurately capturing dependence of rainfall extremes. We could compare the regionalisations produced by the station data to regionalisations from gridded model products. The gridded products might include observational products, such as AWAP (Jones et al., 2009) or HADeX2 (M. Donat et al., 2013), or model based products, such as CMIP6 (Eyring et al., 2016). Understanding dependence in these data products is vital to assessing the extent to which climate models are accurately capturing physical processes.

8.3 Future Research

In the following sections, we outline some future research directions.

Means versus extremes

We have explored how to model Australian rainfall extremes; however, as part of this research we have not discussed whether there is any relationship between the mean of the distribution and the extremes. This decision was made in part to minimise modelling assumptions, and to obtain an understanding of extremes separate to the mean. However, if there is a relationship between mean and extremes it could be exploited for modelling, for example in Oliver et al., (2014). As such, it would be useful to know where in Australia changes in the mean rainfall are reflective of changes in the extreme rainfall, and conversely, where changes in the extremes are different from those of the mean. Some studies have considered this (CSIRO et al., 2015); however the relationship between mean and extremes is often quantified using simple statistics, such as a correlation or ordinary least squares (L. V. Alexander, Hope, et al., 2007). We would like to probe this relationship further. Perhaps exploring
how the spatial dependence between rainfall events that occur in the body of the distribution differs from events that occur in the far tail.

Peaks over threshold

The applications of extremes presented in this thesis have primarily utilised spatial-statistical methods. The standard formulation of these methods for extremes begins by taking the maximum over a block of observations, and does not require the date of the occurrence of extreme events. Depending on the application, utilising spatial-temporal statistical models for extremes is a natural direction for future work. Possible extensions to the spatial methods range from regional peaks-over-threshold approaches without dependence to more advanced space-time max-stable processes; examples of these methods can be found in (De Fondeville et al., 2016; Embrechts, Koch, and Robert, 2016; Roth et al., 2012; Thibaud et al., 2013). We hope to use these methods in future to further improve our understanding of the risk of extreme rainfall events.

Sub-daily Rainfall

The statistical inference in this thesis has made use of observed daily rainfall; however understanding the risk on events on a sub-daily time-scale is also of vital importance for mitigating the impacts of extreme rainfall events. This is particularly true and relevant for short-duration rainfall events that can result in flash flooding.

In many cities, radar data can be used to provide high-resolution sub-daily observations. Unfortunately in Australia, radar data has yet to be quality assured. This is problematic for extremes, given extremes are often poorly captured by radar. There is potential to use radar data for the statistical analyses of rainfall extremes in future after quality assurance. However, we suspect there will be interesting statistical challenges for fitting extreme value models given the high spatial-temporal resolution of radar data. In future work we hope to meet these challenges and utilise sub-daily observations in statistical analyses.

Record Homogeneity

As part of the research stemming from this thesis, we have considered how to assess homogeneity of climate records. Homogeneity tests often falsely assume normality (Alexandersson, 1986). Research is needed to assess the statistical implications of such an assumption, with the impacts of such assumptions most prevalent for daily observations (X. L. Wang, Chen, et al., 2010). Additionally, inhomogeneities in a record may not be spurious, but can be attributed to
climatic reasons. For optimal use of the data, we therefore need to separate spurious inhomogeneity from natural variation due to climate.

Improving Model Parsimony

In Chapter 6, we considered how to best model the marginal parameters of a GEV distribution in a max-stable model. We observed that there is a balance between capturing the complexity of the topography and creating a parsimonious model. We believe there is a possibility to improve upon the parsimony of our fitted model. As part of the exploratory data analysis, we identified a linear relationship between the location parameter and the scale parameter of the GEV distribution in this region, Figure 8.1. There is scope to exploit this relationship by fixing the coefficient of variation, \( c_v \), so that for a station \( x \)

\[
\sigma_x = \frac{1}{c_v} \mu_x. \tag{8.1}
\]

For a discussion on fixing the coefficient of variation for hydrological applications see Hosking and Wallis, (2005).

Figure 8.1: An exploratory data analysis of the marginal parameters of a GEV distribution. Fitted parameters are from stations in South East Queensland. The plot shows the linear relationship between the location and scale parameters of the GEV distribution.
However, we must consider the implications of modelling the parameters using this linear relationship. In particular, how does combining the two parameters affect the overall model uncertainty, and is it appropriate to assume dependence between the location and scale parameters. A discussion of a similar application that considers the parameters of the GP distribution can be found in Ribereau, Naveau, and Guillou, (2011).

**Clustering Methods and Dependence**

In Chapter 7, we discussed the selection of an appropriate unsupervised learning method for our clustering application. We considered well-known and simple clustering methods, including hierarchical clustering, $K$-medoids and $K$-means. However, there are many other clustering methods in the literature. Of particular interest are fuzzy clustering methods, where instead of each point being assigned a single class, each point has a discrete probability distribution. Examples of such methods include fuzzy-C clustering (Bezdek, Ehrlich, and Full, 1984) and T-distributed stochastic neighbour embedding (t-SNE) (Maaten and Hinton, 2008). These methods could be useful for understanding whether the dependence of extremes is smoothly varying in space and time and how variable dependence could be modelled. Creating a statistical model for extremes with a variable dependence structure potentially could help us to model the physical behaviour of rainfall extremes more realistically.

**8.4 Concluding Remarks**

This thesis has addressed gaps within the literature by helping to meet some of the current statistical shortfall when modelling rainfall extremes in Australia. The research presented creates a bridge between highly technical statistical ideas and application areas in climatology and engineering. We hope to pursue some of the possible research directions outlined in the above two sections in future work.
Appendix A

Density Based Clustering

In Chapter 7, we raised concerns about the interpretability of clusters due to variations in network density. To understand how the network density might influence our clustering, we used the unsupervised learning algorithm, Density Based Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). We observe a relationship between the clusters we form using density based clustering and the NRM clusters.

A.1 DBSCAN Algorithm

DBSCAN is a density based clustering algorithm that aims to identify noise and clusters of irregular shape. The algorithm has two parameters, a distance, $\varepsilon$, and a minimum number of points, $m$. The number of clusters therefore does not need to be prespecified, as these two parameters specify the clusters based on the point density.

For DBSCAN each point is classified as belonging to one of three point types:

- a core point, if there are at least $m$ other points within distance $\varepsilon$;
- a boundary point, if there are fewer than $m$ points within distance $\varepsilon$, but there is at least one point within distance $\varepsilon$ that is a core point; or
- a noise point, if it is neither a core point nor a boundary point.

Clusters are formed by grouping core points that have intersecting $\varepsilon$ neighbourhoods. Boundary points are clustered with their neighbouring core points. If a boundary point has two or more neighbouring core points that belong to different clusters, then the boundary point is randomly assigned to one of the core points and its associated cluster. Noise points are not clustered. This algorithm is implemented in the R package, dbscan (Hahsler
A.2 Density of the Australian Rainfall Network

To understand the density of the network of Australian rainfall stations, we applied DBSCAN using 4 different values of $\epsilon$. The minimum number of points was set to 4 to prevent long highways being included in connected regions. The results are shown in Figure A.1.

Figure A.1: Plot of DBSCAN clusters formed using a minimum of 4 points. The clustering is given for 4 different $\epsilon$ values. In each of these sub-figures, the dashed line gives the boundaries of the NRM super clusters.

Under DBSCAN, the regions of highest connectivity are close to the major
cities. Noise points, indicating stations with weaker dependence, were primarily located in regions of Central Australia and Northern Australia. With \( \epsilon = 0.3 \), we identified three core regions: Eastern Australia, Southwest Western Australia and Tasmania. The connectivity of stations in these regions is much higher relative to Central and Northern Australia.

\section*{A.3 Implications for NRM Clusters}

The core regions we have identified using DBSCAN are coincident with boundaries of the NRM clusters. For \( \epsilon = 0.3 \) and a minimum of 4 points, we recover the combined boundaries of Eastern Australia and Southern Australia from the NRM super clusters (see Figure 7.2 and 7.1). Additionally, for \( \epsilon = 0.2 \) and a minimum of 4 points, we recover some of the more obscure features of the boundary of the Western Tropics in the NRM clusters (see Figure 7.3).

The coincidence of the NRM clusters and the density based clusters has important implications for researchers in the climate space. This suggests that the creation of the NRM boundaries may not be driven by climatic or topographic features, but instead constrained by the physical locations of rainfall stations. Using the NRM regions to infer information about the state of the climate is therefore at risk of bias from the underlying network, and as such, any resulting inference must be interpreted with caution.
Appendix B

Additional Summary Figures for Regionalisation
Figure B.1: Hierarchical clustering of Northern Australia with average linkage.
Figure B.2: Satellite Image of Northern Australia. The overlaid regionalisation is for a cut height near 0.133.
172  ADDITIONAL SUMMARY FIGURES FOR REGIONALISATION

Figure B.3: Hierarchical clustering of Regional Australia with average linkage.
Figure B.4: Satellite Image of Regional Australia. The overlaid regionalisation is for a cut height near 0.133.
Bibliography


Hechenbichler, K. and K. Schliep (2004). “Weighted k-nearest-neighbor techniques and ordinal classification”. In:


BIBLIOGRAPHY


Köppen, W. P. (1931). “Grundriss der klimakunde”. In:


Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Saunders, Kate

Title:
An investigation of Australian rainfall using extreme value theory

Date:
2018

Persistent Link:
http://hdl.handle.net/11343/220318

File Description:
Complete Thesis: An Investigation of Australian Rainfall using Extreme Value Theory

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.