# Simulation of Rainfall Events


Eferhonore Efe-Eyefia

School of Mathematics

Cardiff University

A thesis submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy*

2023

## Summary

Rainfall events play an essential role in ecosystems worldwide. While water in good quality and quantity is vital for all life on earth, rainfall can also be the cause of adverse events such as floods, debris movements and droughts. The design of hydrological structures (dams/drainage), urban planning or climate change adaptation needs to take into account such risks. However, as rainfall data are usually only available and measured for a few places of interest, one often relies on simulations of rainfall events that can accurately reproduce realistic rainfall patterns. In order to also capture risks imposed by flash floods (short duration-high intensity events), it is crucial to have simulators with a high-frequency output (e.g. sub-hourly).

It is the purpose of this thesis to build a novel stochastic parsimonious high-frequency rainfall simulator from high-frequency data that can accurately represent key characteristics of rainfall events in the data: duration (D), intensity (I), maximum intensity (M), and volatility (V), collectively referred to as DIMV, as well as temporal patterns of inter-event times.

Therefore, this thesis works with a unique dataset of high-resolution (6-minute) rainfall gauge data from Sunbury, Australia, spanning 36 years from the Australian Bureau of Meteorology. We use a 1-hour minimum inter-event time to extract rainfall events from these data.

First, we analyse the univariate marginal distributions of the above characteristics. Our studies addressed the skewed nature of the DIMV data using log transformations, leading to effective modelling. The skew t was identified as the best fit for duration and volatility, while the generalised extreme value distribution was the best fit for intensity and maximum intensity. We also developed a novel univariate hybrid model, F-Exp-GPD, designed to model rainfall events. By generalising existing hybrid distributions, the F-Exp-GPD showcased versatility, offering a harmonious representation of both bulk and tail behaviour. The model was used to fit duration and intensity to affirm the efficacy of this model, with the GEV-Exp-GPD variant standing out. This knowledge facilitated

sophisticated compound distribution and copula modelling.

To capture the interdependence among the variables, we utilised the vine copula methodology. Among the various vine copula structures, the D-vine copula proved to be the most formidable in representing the dependencies of the DIMV characteristics. This was validated by successful simulations that maintained intricate sample dependencies, drawing a striking resemblance between the copula simulated and observed data.

Lastly, from the fitted models, we developed a flexible rainfall event simulator that also incorporates accurate rainfall temporal intensity patterns using IET data information. It effectively simulates rainfall across long time intervals, reproducing statistical properties of DIMV patterns from the data. This model iterates through specified times, integrates real-world statistical properties by utilising the rain event simulator, and generates detailed rain events, considering consistent and irregular intervals between them. The developed model for simulating an array of rainfall events promises a high degree of authenticity, making it a cornerstone for future hydrological studies, urban planning, and climate change modelling.

## Acknowledgements

My heartfelt appreciation goes to

- Professor Owen Jones, Dr. Kirstin Strokorb, and Professor Michael Singer, my academic supervisors, for their expert guidance and all the time and effort they put into mentoring me and helping me succeed.
- Federal University of Petroleum Resources, Effurun, Delta State, for giving me PhD sponsorship through the Tertiary Education Trust Fund (TETFUND)
- The School of Mathematics, Cardiff University, for providing such a welcoming and encouraging environment for students. I particularly appreciate the PGR team and my PhD colleagues.
- To my family, especially my wife Esther, our children Eguonor and Tega, my parents, Engr. Efe and Mrs. Patience Eyefia, and my siblings for their unwavering support and prayers throughout my years of schooling.
- To my friends, Patrick, Clement, Ifeanyi, and Eghwerido, for your supporting me all along.
- Finally, to God almighty for giving me the grace to complete my PhD and taking me this far in Life.

## Dissemination of work

### Presentations

- 2022: **Hybrid Model for Rainfall Data (conference)**. *Wales Mathematics Colloquium, Gregynog Hall, Newtown, UK.*

- 2020: **Simulation of Rainfall Events**. *GW4 Water Security Alliance Conference, University of Bath, UK.*

- 2020: **Copula Model with Parametric Marginal Distributions (PGR Talk)**. *Cardiff School of Mathematics.*

- 2019: **Simulation of Rainfall Events (SWORDS)**. *Cardiff School of Mathematics.*

## Contents

**Abstract** iii

**Acknowledgements** v

**Dissemination of work** vii

**List of Abbreviations** xx

**1 Introduction** 1

1.1 Introduction . . . . . 1

1.2 Novel contributions of the thesis . . . . . 3

**2 Review of Mathematical Models for Rainfall Events** 5

2.1 Rectangular Pulse Poisson Model (RPPM) . . . . . 5

2.2 Neyman-Scott rectangular pulse Poisson model (NSRPM) . . . . . 6

2.3 Bartlett-Lewis rectangular pulse Poisson model (BLRPM) . . . . . 9

2.4 Modified Bartlett-Lewis Rectangular Pulses Model . . . . . 10

2.5 Modified Neyman-Scott Rectangular Pulses Model . . . . . 12

2.6 A hybrid point rainfall model based on the Jitter process and the Bartlett-Lewis point process . . . . . 14

2.7 A generalized hybrid point rainfall model based on a jitter process and a binary chain model . . . . . 15

2.8 Modified Random Pulse Bartlett-Lewis Stochastic Rainfall Model . . . . 16

2.9 Neyman–Scott Rectangular Pulse Model with Gumbel's Type-II Bivariate Exponential Distribution . . . . . 18

2.10 Bartlett-Lewis Pulse (BLP) Model . . . . . 20

2.11 The Bartlett-Lewis Instantaneous Pulse Rainfall (BLIP) Model . . . . . 21

2.12 A copula-based bivariate frequency analysis: A study on Bartlett-Lewis model . . . . . 23

2.13 Hybrid Exponential GPD Model . . . . . 24
2.14 Doubly Stochastic Point Process Model . . . . . 25
2.15 Enhanced Modeling Through the Random Parameter Bartlett-Lewis Instantaneous Pulse (BLIPR) Model . . . . . 27
2.16 Doubly Stochastic Point Process Exponential Pulse Model . . . . . 28
2.17 A Simple, Flexible and Parsimonious Stochastic Rainfall Model (STORM) 30
2.18 A Simple, Flexible and Parsimonious Stochastic Rainfall Model (STORM-REVISITED) . . . . . 31
2.19 A Censored Approach to Bartlett-Lewis Model . . . . . 33
2.20 A Hybrid Rainfall Model based on the MBLRP and SARIMA Models . . 34
2.21 Copula-based Stochastic Sub-hourly Rainfall Generation Model . . . . . 35
2.22 A Cox Process with State-Dependent Exponential Pulses . . . . . 36
2.23 Stochastic Rainfall Models involving Markov Chain Model . . . . . 38
2.24 Rainfall Event Models involving Spatial Weather Systems . . . . . 43
2.25 A Multi-site Stochastic Weather Generator: The Generative Adversarial Network (GAN) . . . . . 45

**3 Marginal Modelling of Rainfall Events Characteristics 48**

3.1 Introduction . . . . . 48
3.2 Data Pre-processing . . . . . 48
3.3 Marginal Modelling using Parametric Distributions . . . . . 57
3.3.1 Normal Distribution . . . . . 57
3.3.2 Skewed Normal Distribution . . . . . 57
3.3.3 Skew t Distribution . . . . . 58
3.3.4 Generalised Extreme Value (GEV) Distribution . . . . . 58
3.4 Fitting Methodology . . . . . 59
3.4.1 Akaike information criterion (AIC) . . . . . 59
3.4.2 Fitting Results . . . . . 59

**4 Modelling of Extreme Rainfall Events 63**

4.1 Introduction . . . 63
4.2 Peak Over Threshold (POT) Method . . . 64
4.3 Threshold Selection . . . 65
4.3.1 Mean Residual Life Plot . . . 65
4.3.2 Parameter Stability Plot . . . 66
4.4 Parameter Estimation . . . 66
4.5 Return Levels . . . 67
4.6 Results of Univariate Analysis . . . 67
4.7 Multivariate Threshold Model using GPD . . . 81
4.8 Results and Discussion of the Bivariate Analysis . . . 83

**5 Marginal Modelling with Explicit Tail Decay** **87**
5.1 Introduction . . . 87
5.2 The New Hybrid Distribution . . . 89
5.3 Specific Cases of the Hybrid Model . . . 92
5.3.1 St-Exp-GPD . . . 92
5.3.2 Sn-Exp-GPD . . . 93
5.3.3 GEV-Exp-GPD . . . 93
5.4 Maximum Likelihood Estimation of the Parameters of the Hybrid Model 95
5.5 Simulation Study . . . 96
5.6 Application to Rain Events Data and Discussion . . . 98

**6 Dependence Modelling** **103**
6.1 Introduction . . . 103
6.2 Copula . . . 103
6.2.1 Measure of Dependence . . . 105
6.2.2 Tail Dependence . . . 106
6.2.3 Copula family . . . 107
6.2.4 Estimating Bivariate Copula . . . 110
6.2.5 Model Selection for Copulas . . . 111

6.2.6 Results and Discussion . . . 111
6.2.7 Bivariate Copula Simulation . . . 113
6.3 Vine Copula Construction . . . 115
6.3.1 Regular Vine . . . 116
6.3.2 Canonical Vine (C-Vine) . . . 118
6.3.3 Drawable Vine (D-Vine) . . . 119
6.3.4 Selection of R-Vine Model . . . 120
6.3.5 Dependence Results (DIMV) . . . 121

**7 Irregular Pulse Model (Intensity-Duration-Maximum-Volatility) 127**

7.1 Introduction . . . 127
7.2 Simulating Rainfall using IAT . . . 128
7.3 Simulating Rain Depth (Intensity) (given D, I,M, V) . . . 129
7.4 Proposed algorithm for the rainfall event simulator . . . 130
7.5 Implementation in R . . . 132
7.5.1 Handling Single-Step Rainfall Events . . . 132
7.5.2 Addressing Two-Step Events . . . 133
7.5.3 Positioning of Maximum Intensity . . . 134
7.5.4 Enhancing Simulation Accuracy through Random Search . . . 134
7.6 Effects of Parameters . . . 136
7.6.1 Alpha and Beta parameters . . . 136
7.6.2 Tolerance (`tol`) and Maximum Iterations (`max_it`) . . . 137
7.6.3 Recursive Calls (`num_calls`) . . . 139
7.7 Comparison of the simulated event and the original rain event . . . 139
7.8 Assessment of Model Robustness to Extreme Rainfall Scenarios . . . 145
7.8.1 Scaling Factor Choice . . . 145
7.8.2 Results . . . 146
7.9 Simulating of Rainfall Events . . . 149
7.9.1 Proposed Algorithm . . . 149
7.10 Discussion . . . 154

8 **Conclusions** **155**

8.1 Research Summary . . . . . 155

8.2 Further work . . . . . 160

**Bibliography** **162**

**Appendices** **177**

## List of Figures


1.1 Thesis Structure and interdependencies of chapters . . . . . 2

2.1 The Schematic of rectangular pulse Poisson model. For each rainfall event at occurrence time $T_n$, the corresponding pair $U_n = (t_r^{(n)}, i_r^{(n)})$ is defined, where $t_r^{(n)}$ denotes the event's duration and $i_r^{(n)}$ signifies its intensity [1] . 6

2.2 Schematic of the Neyman-Scott rectangular pulse Poisson model [2] . . . 8

2.3 Diagram illustrating the Bartlett-Lewis rectangular pulse model, highlighting that the framework accommodates the superposition of storms and cells [3] . . . . . 9

2.4 Illustration showcasing the Modified Bartlett-Lewis Rectangular Pulse model, where the blue region signifies the duration (represented as width) and the intensity (represented as height) of individual rain cells. The dashed line illustrates the cumulative intensities of all rain cells. [4] . . . . . 11

2.5 A schematic of the BLP model [5] . . . . . 21

2.6 BLIP Model [6]. . . . . 22

2.7 A schematic of the 3-state process with intense rainfall events in State 1 [7] 26

2.8 Schematic description of the DSPP exponential pulse model [8] . . . . . 29

2.9 Schematic flow diagram illustrating key steps in STORM initialization and operation [9] . . . . . 32

2.10 Illustration of the state-dependent initial depth exponential pulse model with a set pulse duration $d$.[10] . . . . . 37

2.11 A flowchart of SDRM-MCREM [11] . . . . . 39

2.12 A framework of MSDRM-MCREM considering two stations as an example [12] . . . . . 41

2.13 A schematic representation of the steps taken in the study [13] . . . . . 45

2.14 The workflow of GAN taking into account extreme rainfall events [14] . . 46

3.1 Rain Data (6-minutes time resolution) . . . . . 49
3.2 Illustration of Rainfall Events . . . . . 52
3.3 Rainfall Events Intensity . . . . . 53
3.4 Rainfall event with largest intensity event (event 923) . . . . . 54
3.5 Rainfall event with the most total rainfall (event 9556) . . . . . 55
3.6 Histogram of DIMV . . . . . 56
3.7 Density plot for DIMV with fitted distributions . . . . . 61

4.1 Rain event duration, intensity and total rainfall with 86, 0.6, 16.5 as selected thresholds for duration, intensity and total intensity respectively . . . . . 68
4.2 Rainfall duration mean residual life plot . . . . . 69
4.3 Parameter estimates against the threshold for rainfall duration . . . . . 70
4.4 Diagnostic plots for rainfall events duration . . . . . 71
4.5 Rainfall intensity mean residual life plot . . . . . 72
4.6 Parameter estimates against the threshold for rainfall intensity . . . . . 73
4.7 Diagnostic plots for rainfall events intensity . . . . . 74
4.8 Total rainfall mean residual life plot . . . . . 76
4.9 Parameter estimates against the threshold for total rainfall . . . . . 77
4.10 Diagnostic plots for rainfall events total intensity . . . . . 79
4.11 Estimates of $\chi(u)$ . . . . . 84
4.12 Estimates of $\bar{\chi}(u)$ . . . . . 84
4.13 Simulated bivariate plot vs actual data . . . . . 86
4.14 Bivariate return level plot for rainfall event duration and intensity . . . . 86

5.1 Hybrid F-Exp-GPD . . . . . 90
5.2 Density plot for log(duration) using the hybrid models . . . . . 99
5.3 Q-Q plot for log(duration) with the hybrid Models . . . . . 99
5.4 Density plot for log(intensity) with the hybrid models . . . . . 101
5.5 Q-Q plot for log(intensity) with hybrid Models . . . . . 101

6.1 Clayton copula rotation with contour plots: left top: 0° rotation ($\tau = 0.5$), right top: 90° rotation ($\tau = -0.5$), left bottom: 180° rotation ($\tau = 0.5$), and right bottom: 270° rotation ($\tau = -0.5$)[15] . . . . . 110
6.2 Upper Triangle: pair plot of copula data (duration and intensity), Diagonal: Marginal histogram of copula data, and Lower Triangle: empirical contour plots of normalized copula data . . . . . 112
6.3 Scatter plot of the rainfall event duration and intensity . . . . . 112
6.4 Scatter plot of observed data vs copula simulated data . . . . . 114
6.5 Six dimensional regular vine tree structure [15] . . . . . 118
6.6 Four-dimensional Canonical Vine [16] . . . . . 119
6.7 Four-dimensional D-Vine [16] . . . . . 120
6.8 Pairs copula plot for DIMV . . . . . 122
6.9 D-Vine Tree Plot for DIMV . . . . . 123
6.10 Pairs plot of copula simulated data (blue) and observed data (red) . . . . 125
6.11 Q-Q plot of copula simulated data and observed data . . . . . 126

7.1 Rainfall events using inter-arrival time (IAT). Where start time $= t_0$, end time $= t_{max}$, intensity $= I$, duration $= D$, and interarrival time $= r_i$ . . . 129
7.2 Simulation Runtime vs Volatility Error across Tolerance Levels and Events: Short Event1 = (I=0.4363636, D=66, M=0.8, V=0.07636364); Long Event1 = (I=0.9428571, D=378, M=6.2, V=1.881905); Short Event2 = (I=0.4, D=18, M=0.4, V=0); Long Event2 = (I=0.1919414, D=1638, M=1.6, V=0.07194139) ) . . . . . 138
7.3 Original Rainfall Events (left) and Simulated Events with observed DIMV (columns 2 to 4) . . . . . 141
7.4 Illustration of how simulated events with given DIMV characteristics look like and resemble realistic events. Original Rainfall Events (left) and Simulated Events with identical DIMV characteristics (columns 2 to 4); different scenarios have been considered: Volatility $V = 0$ (top), High Volatility $V = 25.192$ (middle) and Long Event (Duration $D = 378$ minutes, bottom) 142

7.5 Illustration of how simulated events with given DIMV characteristics look like and resemble realistic events. Original Rainfall Events (left) and Simulated Events with identical DIMV characteristics (columns 2 to 4); different scenarios with long duration . . . . . 143
7.6 Illustration of how simulated events with given DIMV characteristics look like and resemble realistic events. Original Rainfall Events (left) and Simulated Events with identical DIMV characteristics (columns 2 to 4); different scenarios with large total rainfall (DxI) . . . . . 144
7.7 Box plot of intensity for the scaled data and the simulated intensity data obtained using our rainfall event simulator . . . . . 146
7.8 Simulation result for scaled and unscaled data . . . . . 148
7.9 Comparison of Simulated and Observed Rainfall Intensity Characteristics. 150
7.10 Series of Rainfall events for two years period by the Rainfall simulator using Algorithm 4 . . . . . 151
7.11 Series of Rainfall events for a year produced by the Rainfall simulator using Algorithm 4 . . . . . 152
7.12 Series of Rainfall Events at different time resolutions produced by the Rainfall Simulator from Algorithm 4 . . . . . 153

1 Q-Q plot for log(duration) with fitted distributions . . . . . 183
2 Q-Q plot for log(intensity) with fitted distributions . . . . . 184
3 Q-Q plot for log(maximum intensity) with fitted distributions . . . . . 185
4 Q-Q plot for log(volatility) with fitted distributions . . . . . 186
5 Rainfall event with highest total rainfall (Event 9556), start time: 2005-02-02 01:54:00; end time: 2005-02-03 07:48:00; duration: 1794 minutes; total intensity: 168.4mm . . . . . 187
6 Rainfall with largest intensity events . . . . . 188

## List of Tables


3.1 Summary statistics for rain event duration ($D$), intensity ($I$), maximum intensity ($M$), and volatility ($V$) . . . . . 55
3.2 MLE fit for log(duration) using the candidates distributions . . . . . 60
3.3 MLE fit for log(intensity) using the candidates distributions . . . . . 60
3.4 MLE fit for log(maximum intensity) using the candidates distributions . 62
3.5 MLE fit for log(volatility) using the candidates distributions . . . . . 62

4.1 Return Levels with 95% Confidence Intervals of Duration . . . . . 71
4.2 Parameter estimates for rain event duration data . . . . . 71
4.3 Return Levels with 95% Confidence Intervals for Intensity . . . . . 73
4.4 Parameter estimates for rain event intensity data . . . . . 74
4.5 Parameter estimates for rain event total rainfall data . . . . . 78
4.6 Return Levels with 95% Confidence Intervals for Total Rainfall . . . . . 78
4.7 Bivariate Extreme Value Models . . . . . 82
4.8 Bivariate Model Comparison: Parameter Estimates, Deviance, and AIC . 83

5.1 Three Hybrid Models and their Parameters . . . . . 94
5.2 Results of Monte-Carlo simulations for $\mu = 2.5$, $\sigma = 1$, $k = 0.5$, $t_1 = 4.5$, $t_2 = 5$, and $\gamma = 0.2$ . . . . . 97
5.3 MLE fit for log(duration) using the hybrid models . . . . . 98
5.4 MLE fit for log(intensity) using the hybrid models . . . . . 100

6.1 Kendall Tau of different bivariate copula families [17] . . . . . 108
6.2 Tail Dependence of different bivariate copula families ($-$ means undefined) [15] . . . . . 109
6.3 Results for parameter estimates, loglikelihood, $AIC$, $BIC$, $\tau$, $\Lambda_L$, $\Lambda_U$ . . 113
6.4 D-Vine copula with pair-copulas . . . . . 124
6.5 Dependence ($\tau$) table for copula simulated data . . . . . 124

6.6 Dependence ($\tau$) table for observed data . . . . . 124

## List of Abbreviations

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

CDF: Cumulative Distribution Function

GEV: Generalized Extreme Value

GPD: Generalized Pareto Distribution

MLE: Maximum Likelihood Estimation

MSE: Mean Square Error

PDF: Probability Density Function

POT: Peak Over Threshold

EVT: Extreme Value Theory

POT: Peak Over Threshold

# Chapter 1

## Introduction

## 1.1 Introduction

Rainfall is crucial for global ecosystems, providing essential hydration for all life forms. Plants, rivers, and animals heavily rely on it. However, its impacts can be dual-edged. Excessive rain can cause floods and landslides, while insufficient rainfall can lead to droughts, affecting both natural habitats and human communities [18]. The design of hydrological structures (dams/drainage), urban planning, or climate change adaptation must consider such risks. However, to do this, high-resolution rainfall event data is needed. Unfortunately, these data are not available in most cases. Rainfall data are usually only available and measured for a few places of interest. Given the limited availability of rainfall data for specific locations, many turn to rainfall simulation models. These models can accurately replicate realistic rainfall patterns, enabling a more precise quantification of rainfall-associated risks [19, 20].

The advent of stochastic rainfall models such as the Neyman-Scott Rectangular Pulse (NSRP) and Bartlett-Lewis Rectangular Pulse (BLRP) model [21], which blend deterministic processes with elements of randomness, heralded a paradigm shift in rainfall simulations [22]. Unlike their deterministic counterparts, stochastic models can accommodate a broader range of rainfall events and effectively handle uncertainties tied to the complex mechanisms underlying rainfall generation. Nonetheless, challenges persist in selecting and calibrating suitable stochastic models due to the pronounced spatial and temporal variability inherent in rainfall events [23].

A novel rainfall simulation approach incorporates copulas, which provide a flexible methodology for modelling the dependencies between rainfall characteristics such as intensity,

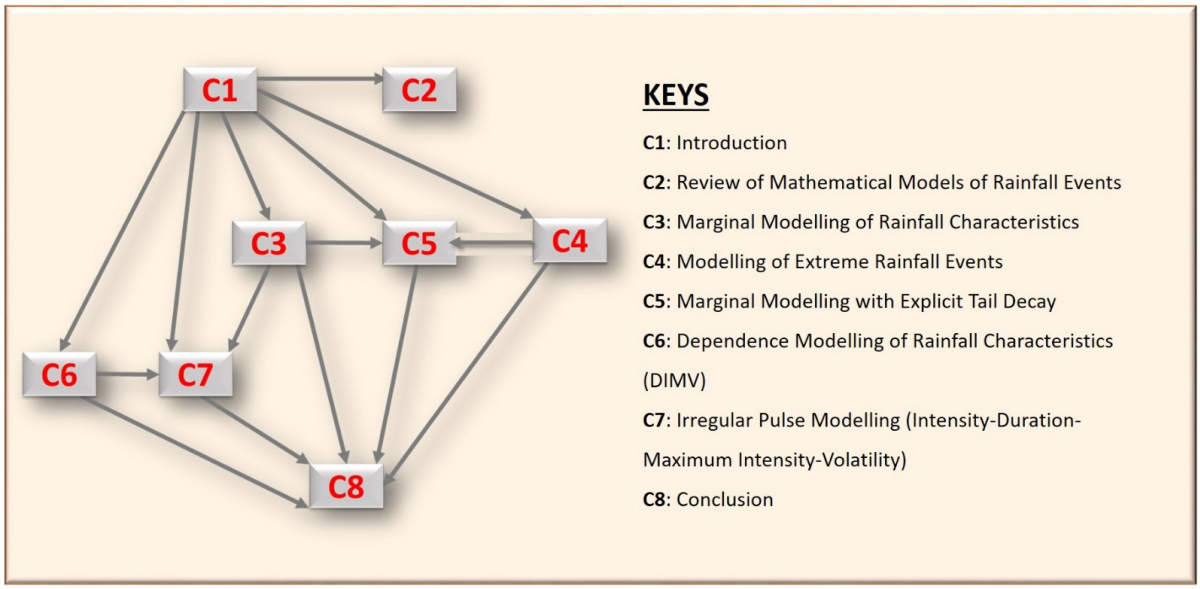


**Figure 1.1:** Thesis Structure and interdependencies of chapters

duration, and inter-arrival time. Copulas allow for the formulation of joint probability distributions, thereby facilitating the creation of more realistic synthetic rainfall data, which is crucial for planning and managing water resources [24].

The overarching goal of this thesis aims to build parsimonious simulation models for rainfall events that can be used to fit high-frequency rainfall (gauge) data and develop a novel stochastic rainfall simulator that can simulate rainfall events reproducing realistic rainfall patterns that reproduce key statistical features of the rainfall events from the data relevant for being informed about risks and therefore relevant for planning purposes.
In order to do so, this thesis pursues the following research agenda:

1. Marginal modelling of duration, intensity, max intensity and volatility.
2. Extreme value analysis of duration and intensity
3. Compound Marginals with Explicit Tail Decay(Extremes)
4. Dependence modelling of duration, intensity, maximum intensity and volatility
5. Simulation of Irregular Pulse Model (intensity-duration-maximum-volatility

Accordingly, the next section elaborates on the structure and contributions of this thesis. Figure 1.1 provides an overview of the interdependencies among chapters.

## 1.2 Novel contributions of the thesis

This section provides a summary of the contributions made in each chapter.

In Chapter 3, we introduce a methodology for defining and extracting rainfall events from a previously untapped 6-minute high-resolution dataset covering 36 years from Sunbury, Australia. We utilized 1-hour minimum interevent time (IET) to differentiate between distinct rainfall events. We discovered and addressed the skewed nature of the DIMV data through log transformation and successfully matched rainfall characteristics to their respective best-fit distributions using the AIC criterion, paving the way for advanced joint modelling of the key characteristics (DIMV).

In Chapter 4, we conducted an explorative extreme value analysis on the univariate variables duration and intensity from the rainfall data from Sunbury, offering insights into potential patterns of extremes. First, a univariate Peaks Over Threshold (POT) model has been fitted. The worst events in terms of flooding would be events where both duration and intensity were extreme. However, the duration and intensity data are e negatively correlated and do not point us to asymptotic dependence between these two variables. We identified the negative bivariate logistic model as the most optimal among tested bivariate models for representing joint extreme duration and intensity data. However, as we can see from a comparison of simulations and original data, even the best bivariate extreme value distribution constitutes a poor fit, and we do not pursue this route any further.

In Chapter 5, we innovate by developing a novel F-Exp-GPD univariate hybrid model for rainfall event duration and intensity. This model improves upon a hybrid distribution model, G-Exp-GPD, by integrating an arbitrary distribution, 'F', for enhanced tailoring to specific datasets. Through the construction of three distinct hybrid distribution models, the GEV-Exp-GPD model emerged as the superior candidate, demonstrating exceptional capability in capturing both the bulk and tail behaviours of intensity and duration data. This advancement addresses the limitations inherent in conventional hybrid

models and significantly expands their scope of applicability, offering a versatile tool for rainfall data analysis.

In Chapter 6, the research pioneers the application of the vine copula approach within our specific context, employing it innovatively to model the complex interdependencies among rainfall event characteristics, denoted as DIMV (Duration, Intensity, Maximum intensity, and Volatility). This chapter introduces a novel methodological framework and establishes the D-vine copula's superiority over alternative vine structures for capturing the nuanced interrelations within the data. The chosen copula model's reliability and effectiveness are rigorously validated through extensive simulations and assessment techniques. This comprehensive validation process effectively bridges the theoretical and empirical realms, offering a robust model that enhances our understanding of rainfall events' dynamics.

Chapter 7 significantly advances rainfall simulation by developing an innovative rainfall event simulator. This tool is distinct in its capacity to simulate detailed rainfall events that integrate both temporal intensity and DIMV patterns in a high-frequency context. The simulation process is grounded in the joint distributional model fitted in Chapter 6, utilizing vine copulas and marginal modelling to generate DIMV patterns for each event. Unlike previous models that focus primarily on duration and intensity, this simulator introduces a deterministic DIMV constraint, underpinning the generation of events with stochastic variability. This approach ensures the simulation of rainfall events that exhibit realistic patterns and faithfully reproduce the statistical characteristics of key attributes. Furthermore, the hybrid methodology that merges temporal intensity with DIMV patterns represents a methodological leap, enabling the simulation of a sequence of rainfall events over a specified time range with unprecedented accuracy and detail. The simulator is a testament to the novel integration of complex statistical modelling with practical simulation techniques, addressing a gap left by prior research and offering a tool with broad applicability and significant potential for advancing our understanding of rainfall phenomena.

# Chapter 2

## Review of Mathematical Models for Rainfall Events

This chapter delves into the literature review of mathematical rainfall simulation models, exploring the advancements and insights gained in point rainfall modelling.

## 2.1 Rectangular Pulse Poisson Model (RPPM)

Rodriguez-Iturbe et al. [1] proposed the rectangular pulse Poisson model. This model is constructed from rectangular pulses associated with a Poisson process. In the rectangular pulse Poisson model, storm occurrences are generated through a Poisson process, where each event is linked to a rainfall duration that is randomly determined and exhibits a constant yet random intensity level. The overall rainfall intensity results from the cumulative contributions of all these storm events. The RPPM can represent a natural process at a fixed level of aggregation as long as such a level is not smaller than the typical duration of a storm event. The model does not account for aggregation and disaggregation of the results, and inferences made from its structure should be confined to the time scale for which it was constructed. The RPPM only perform well at the scale aggregation for which it was constructed. A schematic of the rectangular pulse Poisson model is given in Figure 2.1

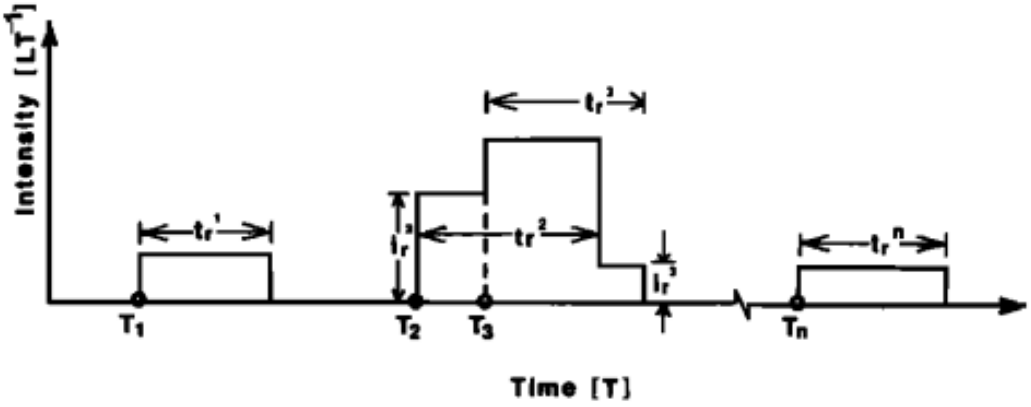


**Figure 2.1:** The Schematic of rectangular pulse Poisson model. For each rainfall event at occurrence time $T_n$, the corresponding pair $U_n = (t_r^{(n)}, i_r^{(n)})$ is defined, where $t_r^{(n)}$ denotes the event's duration and $i_r^{(n)}$ signifies its intensity [1]

Rodriguez-Iturbe et al. [25] employed two special cases of the RPPM where the pulse durations follow the exponential and Pareto distributions to fit the Denver rainfall data (the data comprises hourly precipitation records spanning through May 15 to September 11, 1949 - 1976). The model parameters were computed by fitting the mean, variance, lag-one autocorrelation and the probability of zero rain at a fixed level of aggregation. Results revealed that the rectangular pulse Poisson model based on the exponential durations produced autocorrelation that decayed much too rapidly even at the period used for the fitting and gave a poor fit at other levels of aggregation. Whereas the model based on Pareto durations produced autocorrelation that decays much slower than that of exponential cases, which implies that the Pareto model fits much better at the fixed level of aggregation but also gave a poor fit at other levels of aggregation.

In summary, a significant drawback of the rectangular pulse Poisson models is the inability to aggregate and disaggregate rainfall data.

## 2.2 Neyman-Scott rectangular pulse Poisson model (NSRPM)

Rodriguez-Iturbe et al. in [25] proposed the proposition of two cluster-based models, which include the Neyman-Scott rectangular pulse Poisson model (NSRPM) and the Bartlett-Lewis rectangular pulse Poisson model (BLRPM). In the Neyman-Scott rect-

angular pulse Poisson model, a storm arises in a Poisson process with rate $\lambda$ and each storm is assigned a random number of cells $C\,(C \geq 1)$. The cell origin is not situated at the storm origin, and the distances between cell origins are exponentially distributed with parameter $\beta$. Each cell is a rectangular pulse with duration and depth independent random variables and exponential distribution with parameter $\eta$.In the Neyman-Scott process scenario, the cell positions are specified by a series of independent and identically distributed random variables, representing the time intervals from the storm origin to the birth of the respective cells. Rodriguez-Iturbe et al in [25] gave the second order properties of the aggregated process $Y_i(\tau)$, where $Y_i(\tau)$ is the cumulative rainfall over an interval of length $\tau$:

$$E\left[Y_i(\tau)\right] = \lambda\eta^{-1}\mu_c\mu_x\tau, \tag{2.1}$$

$$\begin{aligned} Var\left[Y_i(\tau)\right] = \lambda\eta^{-3}\left(\eta\tau - 1 + e^{-\eta\tau}\right)\left\{2\mu_c E[X^2] + E\left[C^2 - C\right]\mu_x^2\frac{\beta^2}{\beta^2-\eta^2}\right\} \\ -\lambda\left(\beta\tau - 1 + e^{-\beta\tau}\right)\frac{E\left[C^2-C\right]\mu_x^2}{\beta\left(\beta^2-\eta^2\right)}, \end{aligned} \tag{2.2}$$

$$\begin{aligned} Cov\left[Y_i(\tau), Y_{i+k}(\tau)\right] = \lambda\eta^{-3}\left(1+e^{-\eta\tau}\right)^2 e^{-\eta(k-1)\tau}\left\{\mu_c E[X^2] + \frac{1}{2}\frac{E\left[C^2-C\right]\mu_x^2\beta^2}{\beta^2-\eta^2}\right\} \\ -\lambda\left(1-e^{-\beta\tau}\right)^2 e^{-\beta(k-1)\tau}\left\{\frac{1}{2}\frac{E\left[C^2-C\right]\mu_x^2}{\beta\left(\beta^2-\eta^2\right)}\right\}, \quad k \geq 1, \end{aligned} \tag{2.3}$$

where $\mu_c$ is the mean number $E[C]$ of cells per storm and $X$ is the random variable characterizing the pulse depth or rain cell intensity. The Schematic of the Neyman-Scott rectangular pulse Poisson model is given in Figure 2.2.

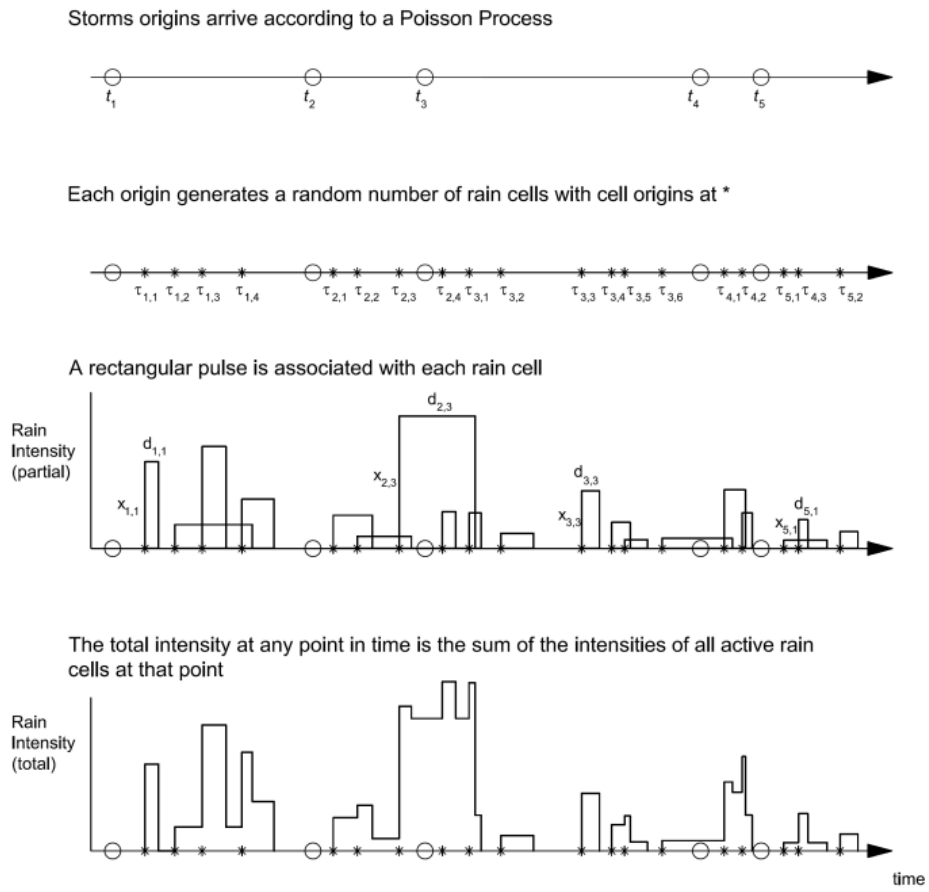


**Figure 2.2:** Schematic of the Neyman-Scott rectangular pulse Poisson model [2]

Calenda and Napolitano in [26] explored how the scale of data aggregation impacts the parameter estimation in classical Neyman-Scott point processes. The authors observed that selecting the data's aggregation scale influences the estimates of the continuous process parameters when determining the parameters through the method of moments. Thus, the motivation to introduce an alternative estimation procedure based on the scale of fluctuation of the observed process. The estimates obtained via the proposed procedure were considerably better than the ones obtained via the procedure employing alternative scales, both in terms of replication of the second-order statistics and extreme values for different aggregation scales, as evident through a Monte Carlo simulation. Conclusively, the authors suggested using the proposed procedure as an adequate alternative to estimating the parameters of the Neyman-Scott processes.

## 2.3 Bartlett-Lewis rectangular pulse Poisson model (BLRPM)

In the Bartlett-Lewis rectangular pulse Poisson model, storm origins occur in a Poisson process with rate $\lambda$ and each origin is followed by a Poisson process of rate $\beta$ of cell origins; after a time, exponentially distributed with rate $\gamma$, the process of cell origins terminates. As stated earlier, the positioning of the cells can be made in several different manners. In the case of Bartlett-Lewis, the intervals between successive cells are independent and identically distributed. Figure 2.3 gives an illustration of the model.

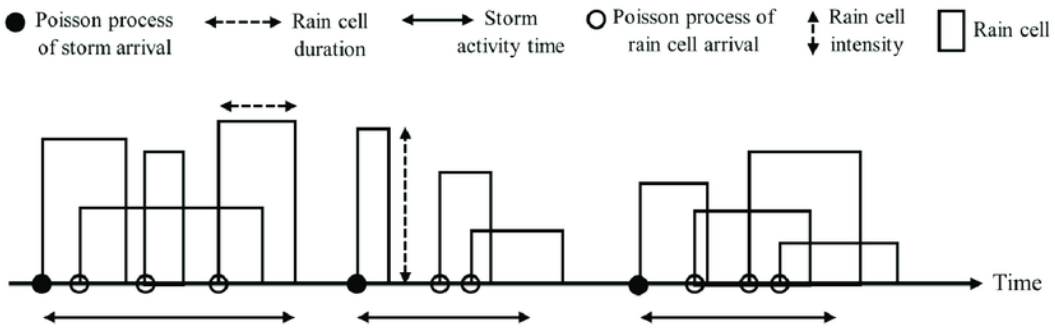


**Figure 2.3:** Diagram illustrating the Bartlett-Lewis rectangular pulse model, highlighting that the framework accommodates the superposition of storms and cells [3]

The second order properties of the aggregated process $Y_i(\tau)$ are given by

$$E\left[Y_i(\tau)\right] = \tau\lambda\eta^{-1}\mu_c\mu_x, \tag{2.4}$$

$$\begin{aligned} Var\left[Y_i(\tau)\right] = 2\lambda\eta^{-1}\mu_c\left\{E[X^2] + \frac{\beta}{\gamma}\mu_x^2\right\}\frac{\tau}{\eta} - 2\lambda\eta^{-1}\mu_c\left\{\mu_x^2 + \frac{\beta\gamma}{\gamma^2-\eta^2}\mu_x^2\right\}\frac{\left(1-e^{-\eta\tau}\right)}{\eta^2} \\ +2\lambda\eta^{-1}\mu_c\mu_x^2\frac{\beta}{(\gamma^2-\eta^2)}\left(1-e^{-\gamma\tau}\right)\frac{\eta}{\gamma^2} \end{aligned} \tag{2.5}$$

$$\begin{aligned} Cov\left[Y_i(\tau), Y_{i+k}(\tau)\right] = \lambda\eta^{-1}\mu_c\left\{E[X^2] + \frac{\beta\gamma}{\gamma^2-\eta^2}\mu_x^2\right\}\left(1-e^{-\eta\tau}\right)^2\frac{e^{-\eta(k-1)\tau}}{\eta^2} \\ -\lambda\eta^{-1}\mu_c\frac{\beta}{(\gamma^2-\eta^2)}\mu_x^2\left(1-e^{-\gamma\tau}\right)^2 e^{-\gamma(k-1)\tau}\frac{\eta}{\gamma^2}, \quad k \geq 1. \end{aligned} \tag{2.6}$$

Rodriguez-Iturbe et al. [25] applied the cluster-based rectangular pulse Poisson models (NSRPM and BLRPM) to the Denver rainfall data. Results revealed that the cluster-based rectangular pulse Poisson models match the historical data well at all aggregation levels. The correlation decay is much slower than the rectangular pulse Poisson models and equally follows the historical correlation structure. The authors remarked that the cluster-based rectangular pulse Poisson models are capable of representing the cumulative rainfall attributes across various time scales, ranging from 1 to 24 hours while maintaining consistent model parameters. More so, the range of temporal scales through which cluster-based rectangular pulse models can aggregate and disaggregate the rainfall process will likely be 1 — 48 hours.

## 2.4 Modified Bartlett-Lewis Rectangular Pulses Model

Rodriguez-Iturbe et al.[25] noted that the classical cluster-based models could preserve the statistical characteristics of rainfall data aggregated at various levels without altering the model parameters. However, there was a clear observation that the probability of zero rain (dry periods) when those periods were above several hours was highly overestimated by both models. The implications for infiltration studies and other hydrologic considerations, such as rainfall runoff transformations, are critical since there can be a significant difference in the runoff output when the period with no rainfall is varied. The classical cluster-based models considered rectangular cells whose stochastic description was invariant throughout the storm events. In other words, the duration of the cells, intensity, and number of cells came from distribution functions whose parameters were the same for all storms.

In an attempt to develop a more flexible model that allows for different structural characteristics among the other storms, Rodriguez-Iturbe et al. [27] developed the modified Bartlett-Lewis rectangular pulses model, which, in addition to accounting for different structural characteristics among the different storms, also is capable of representing a large variety of statistical characteristics of the rainfall process at varying levels of ag-

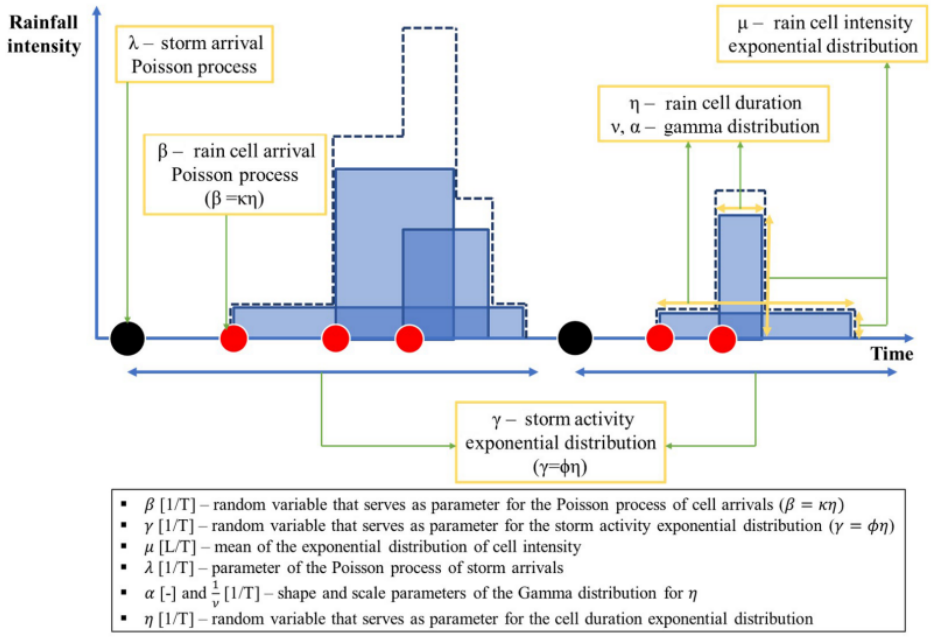


**Figure 2.4:** Illustration showcasing the Modified Bartlett-Lewis Rectangular Pulse model, where the blue region signifies the duration (represented as width) and the intensity (represented as height) of individual rain cells. The dashed line illustrates the cumulative intensities of all rain cells. [4]

gregation including the probability of dry periods. This was achieved by allowing the parameter $\eta$, specifying the duration of cells, to vary randomly between storms. Hence, the authors referred to the modified model as the random $\eta$ model, in contrast to the BLRPM, the fixed $\eta$ model.

To illustrate the usefulness of this model, the authors fitted it to two distinct data sets, namely, the Denver and Boston rainfall data, to assess how well the modified Bartlett-Lewis rectangular pulses model improved on the classical model. The results obtained from the data revealed that the model with random $\eta$ gave a better fit than the model with fixed $\eta$. Onof and Wheater [28] investigated the applicability of the modified Bartlett-Lewis rectangular pulses model using hourly British data and also examined the ability of the model to reproduce seasonality. Khaliq and Cunnane [29] further demonstrated applying the modified Bartlett-Lewis rectangular pulses model to fit two hourly rainfall data sets recorded at Valentia and Shannon Airport, Ireland.

## 2.5 Modified Neyman-Scott Rectangular Pulses Model

Similarly, Entekhabi et al. [30] introduced the modified Neyman-Scott rectangular pulses model using the same approach in [27]. The authors noted that one factor that controls the duration of precipitation and wet and dry runs is the inverse mean cell duration $\eta$. Rather than setting $\eta$ as a constant parameter that dictates the distribution governing the duration of all cells, $\eta$ is now treated as a random variable that varies with each storm. Thus, the duration of the cells from storm $i$ are random quantities governed by an exponential distribution with parameter $\eta_i$. All other assumptions remain the same as in the classical Neyman-Scott rectangular pulses model, and thus $\eta$ is independent of the number of cells and the cell intensities $x$.

Following equation 2.1 - equation 2.3, the second-order properties of the aggregated modified process $Y_i(\tau)$ are obtained as follows:

$$E[Y_i(\tau)] = \mu_x \mu_c \lambda \tau I(1,0), \tag{2.7}$$

$$\begin{aligned} Var[Y_i(\tau)] = [\mu_x \mu_c \lambda \tau I(1,0)]^2 + \left\{2C_1\tau + C_2\beta^{-3}\left(\beta\tau + e^{-\beta\tau} - 1\right) + (\mu_x\mu_c\lambda\tau)^2\right\} I(2,0) \\ -2C_1 I(3,0) - C_2\tau I(4,0) + C_1 I(5,0) + 2C_1 I(3,\tau) - C_2 I(5,\tau), \end{aligned} \tag{2.8}$$

$$\begin{aligned} Cov[Y_i(\tau), Y_{i+k}(\tau)] \;=\; & C_1 I(3,k\tau-\tau) - 2C_1 I(3,k\tau) + C_1 I(3,k\tau+\tau) - \frac{C_2}{2} I(5,k\tau-\tau) \\ & + C_1 I(5,k\tau) - \frac{C_2}{2} I(5,k\tau+\tau) + \frac{C_2}{2}\beta^{-3}\left(1-e^{-\beta\tau}\right)^2 e^{-\beta(k-1)\tau} I(2,0) \\ & + (\mu_x\mu_c\lambda\tau)^2\left[I(2,0) - I^2(1,0)\right] \end{aligned} \tag{2.9}$$

where

$$C_1 = \lambda \mu_c E\left[X^2\right],$$
$$C_2 = \lambda E\left[C^2 - C\right]\mu_x^2 \beta^2$$
$$I(x,y) = E\left[\eta^{-x} e^{-\eta y}\right] = \frac{\Gamma(\alpha - x)}{\Gamma(\alpha)} \theta^{\alpha} (\theta + y)^{x-\alpha}, \quad x > 0, \ y \geq 0.$$

As previously highlighted, the classical Neyman-Scott rectangular pulses model tends to significantly overestimate the ratio of dry intervals relative to the series length, particularly for aggregation periods spanning from several hours to a couple of days. By adopting the same Denver rainfall data, Entekhabi et al. [30] applied both the classical and the modified Neyman-Scott rectangular pulses model to fit the rainfall data and remarked that from the result obtained from the later, even up to 8-day aggregation periods the agreement between the historical probabilities and the modified process probabilities are good. In addition to preserving the mean, variance and lagged autocorrelation, the modified Neyman-Scott rectangular pulses model preserves the dry-wet time structure of point observations of rainfall.

Cowpertwait [31] argued that the empirical distribution, particularly the distribution tail, is likely to be consistently well-fitted, with some high-order properties included in the fitting procedure. This claim instigated the derivation of the third-order aggregated moments of the modified Neyman-Scott rectangular pulses model. New Zealand's National Institute of Water and Atmospheric Research (NIWA) hourly rainfall data was used in the fitting process. The result from the analysis revealed a good fit for the observed extreme values over a range of time scales. On the other hand, a poor fit was evident when the third moment was expunged from the fitting procedure. Cowpertwait [31] finally stressed that the derivation of the third moment function seems well justified, and This function can be helpful in extended simulation studies and in planning hydraulic structures.

## 2.6 A hybrid point rainfall model based on the Jitter process and the Bartlett-Lewis point process

Gyasi-Agyei and Willgoose [32] developed a hybrid point rainfall model by amalgamating the attributes of the jitter process with the well-established Bartlett-Lewis point process. This innovative model, denoted as $\{H(t)\}$, emerges from the interaction between two distinct random processes: $\{A(t)\}$, which serves as a "jitter" process introducing correlated adjustments to refine the model's adherence to the second-order properties of rainfall data, and $\{B(t)\}$, which is dedicated to encapsulating the primary rainfall event characteristics and mean dry probabilities observed in historical data.

The essence of the jitter process $\{A(t)\}$ is encapsulated through its formulation as a lognormal random process, intricately linked with a stationary Gaussian process that meticulously captures the variance, mean, and autocovariance function. These parameters are meticulously derived from the historical data's second-order properties, offering a refined adjustment mechanism to the overall model. Concurrently, the non-randomized Bartlett-Lewis rectangular pulse model adeptly models the rainfall event characteristics—encompassing the average event duration and the probability distributions of dry intervals. This component of the hybrid model stands out for its capacity to accurately simulate the inherent dynamics of rainfall processes, including the distribution and occurrence of dry periods.

By implementing this hybrid modelling approach, the model $\{H(t)\}$ demonstrates exceptional competence in reproducing the historical rainfall data's intricate patterns, including the mean number of events and their durations, with remarkable accuracy. Notably, this model outperforms existing methodologies, such as the modified Bartlett-Lewis model proposed by Rodriguez-Iturbe et al. [27], in capturing the empirical characteristics of rainfall events. This model's efficacy was empirically validated using 15-minute interval rainfall data from Capella, central Queensland, Australia. This application demonstrated that the hybrid model excels at providing thorough and detailed knowledge of rainfall

patterns, backed by its ability to derive explicit event characteristics directly from the model parameters. This innovative approach ensures a robust framework for accurately simulating and analyzing rainfall processes, reflecting a significant advancement in the field of hydrological modelling [32].

## 2.7 A generalized hybrid point rainfall model based on a jitter process and a binary chain model

Gyasi-Agyei and Willgoose [33] generalized the hybrid model due to Gyasi-Agyei and Willgoose [32] by substituting the traditional Bartlett-Lewis model with a binary chain model, yet retaining the autoregressive model employed as a jitter to fix deficiencies in the second order properties of the binary chain. The binary chain consists of a string of two numbers, zero for a dry period and a constant value, $w$, for a wet period. A Markov chain and the Bartlett-Lewis models were used as examples. Lall et al. [34] pointed out that Markov chain models are attractive because of their largely non-parametric nature (i.e. the parameters are derived directly from data), ease of application and interpretability and well-developed literature. However, as the order increases, a recognized limitation of the Markov chain models is their lack of simplicity or parsimony.

As stated earlier, a binary chain model generates a string of two numbers, $Y_i = 0$ for a dry period and a constant value $Y_i = w$ for a wet period, where $Y_i$ is the total amount of rain over a specific time period $i$. The moments of a binary chain were derived analytically as functions of the dry probabilities. For the historical data and a binary chain of the same time scale to have the same mean, $\mu_{Y_i}$ the cumulative rainfall depth over a wet period $w$ must be given by

$$w = \frac{\mu_{Y_i}}{1 - P(i)} \tag{2.10}$$

where $P(i)$ is the probability that an interval $i$ is dry, also written as $P(Y_i = 0)$.

Assuming second-order stationarity, the variance

$$\begin{aligned}\sigma_Y^2 &= E\left[(Y_i-\mu_{Y_i})^2\right] \\ &= P(i)(0-\mu_{Y_i})^2+[1-P(i)](w-\mu_{Y_i})^2 \\ &= (\mu_{Y_i})^2\left\{P(i)+[1-P(i)]\left(\frac{w}{\mu_{Y_i}}-1\right)^2\right\} \\ &= (\mu_{Y_i})^2\frac{P(i)}{[1-P(i)]}\end{aligned} \tag{2.11}$$

The hybrid model based on the classical Bartlett-Lewis and second-order autoregressive models is denoted as BBLAR. Concurrently, the hybrid model of a Markov chain of order 12 and the second-order autoregressive model is represented as MCAR12. For data fitting purposes, the 15-minute point rainfall data reported in Gyasi-Agyei and Willgoose [32] were adapted to compare and evaluate two hybrid models. The authors stressed that while these two models reproduced the aggregated statistics very well, the BBLAR model performed more favourably better than the MCAR12 model because it is parsimonious regarding the number of model parameters.

## 2.8 Modified Random Pulse Bartlett-Lewis Stochastic Rainfall Model

Cameron et al. [35] noted that one of the attractions of pulse-based modelling is that, through the direct simulation of rain cells, the procedure is (intuitively) physically reasonable. Indeed, after a pulse-based model's parameters have been optimised upon a rainfall data series, that model can satisfactorily reproduce many of the properties of that data series (including dry periods). Nevertheless, the efficiency of pulse-based models for extreme rainfall simulation has often been less clear-cut, particularly for extreme rainfalls of short duration (e.g., 1-hour maxima). Although a handful of revised models have been introduced to handle this pitfall, a significant drawback in applying these revised models lies in estimating the model parameters. For example, Cameron et al. [36] examined three stochastic rainfall models (two profile-based models and a gamma version of the

random pulse Bartlett-Lewis model (RPBLGM) using point rain gauge data from three independent sites in the UK. In particular, the RPBLGM provided good simulations for the seasonal extreme rainfall totals of 24 h duration and standard rainfall statistics at each site; the model underestimated the observed seasonal maxima of 1-hour duration. This situation instigated the motivation for Cameron et al. [35] to develop a new version of the random-pulse Bartlett-Lewis model for extreme rainfall simulation. This new model features a generalised Pareto distribution (GPD) to represent the depths of high-intensity rain cells. The GPD is characterised by the distribution function defined as

$$F(x) = 1 - (1 + [\xi (x - \mu)/\sigma])^{-1/\xi}, \qquad \xi \neq 0,$$
$$F(x) = 1 - exp[-(x - \mu)/\sigma], \qquad \xi = 0, \tag{2.12}$$

Where $F(x)$ is a non-exceedance probability, $\xi$ a shape parameter, $\mu$ (the intensity threshold) a location parameter, $x - \mu$ an exceedance (where $x > \mu$), and $\sigma$ is a scale parameter. The GPD was selected due to its flexibility and ability to model peaks over threshold (POT) data in traditional extreme event frequency analysis.

The proposed model's parameter estimation was conducted through a two-stage method, utilising the generalised likelihood uncertainty estimation (GLUE) technique. The first stage estimates the parameters of the random pulse Bartlett-Lewis model (RPBLM) used by Onof and Wheater [28]. In contrast, the second stage holds the GPD threshold, u, fixed. Then, it estimates the two GPD parameters using the generalised likelihood uncertainty estimation (GLUE) technique reported in Beven and Binley [37]. In this technique, it is assumed that, since the GPD parameters are only appropriate to the simulation of extreme rainfalls, they should only have a minimal impact on the standard statistics of the simulated continuous rainfall time series.

An extreme rainfall simulation for a UK site (44 summer half-year data at Elmdon, Birmingham) was investigated to demonstrate the model's efficacy. The result showed that the proposed model is better than older versions of the Bartlett-Lewis model at

reproducing the observed series 1-hour seasonal maxima (SEAMA ) for the summer season at Elmdon. Also, the model was reliable when showing that the 24-hour SEAMAX totals were reliably consistent with the gamma version of the random pulse Bartlett-Lewis model (RPBLGM).

## 2.9 Neyman–Scott Rectangular Pulse Model with Gumbel's Type-II Bivariate Exponential Distribution

Kim and Kavvas [38] argued that except for a few models, several rainfall models assume an independent relation between rain cell intensity and duration to easily derive the temporal covariance structure of the rainfall time series. Even though a few models could consider such dependence, they only applied the Poisson process to rainfall occurrence without considering any clustering feature of rainfall. In particular, Singh and Singh [39], and Bacchi et al. [40] applied Gumbel's Type-I bivariate exponential distribution. Considering that the Gumbel Type-I bivariate exponential distribution always has a negative correlation between the variables of interest, they considered the negative correlation between rainfall intensity and duration. Gumbel in [41] introduced three types of bivariate exponential distribution. The first type is known as the Gumbel Type I bivariate exponential distribution with the joint density function defined as

$$f(x,y) = e^{-(x+y+\theta xy)}\left[(1+\theta x)(1+\theta y)-\theta\right], \qquad 0 \le \theta \le 1, \qquad x,y>0. \tag{2.13}$$

The second type is the Gumbel Type II bivariate exponential distribution, an F-G-M model with exponential marginals. The joint density function is given by

$$f(x,y) = e^{-(x+y)}\left[1+\alpha(2e^{-x}-1)(2e^{-y}-1)\right], \qquad |\alpha|< 1. \tag{2.14}$$

Similarly, Cordova and Rodriguez-Iturbe [42], and Goel et al. [43] applied Downton's bivariate exponential distribution to consider the positive correlation between rainfall intensity and duration. Downton's bivariate exponential distribution is specified by the

joint density function [44].

$$f(x,y) = \frac{1}{1-\rho} exp[-(x+y)/(1-\rho)] I_0\left(\frac{2\sqrt{xy\rho}}{1-\rho}\right), \qquad x,y \geq 0, \tag{2.15}$$

Where $I_0$ is the modified Bessel function of the first kind of order zero.

Kim and Kavvas [38] developed a new stochastic point rainfall model to simultaneously consider both the negative and positive correlation between rain cell intensity and duration. To achieve this, the Gumbel Type-II bivariate distribution was adapted. Additionally, the Neyman–Scott cluster point process was employed to address the clustering characteristic inherent in rainfall processes. The proposed model, thus, accommodates a positive or negative correlation parameter that the historical rainfall time series should determine.

The implementation of the proposed model was demonstrated with data from the rainfall station in Jeonju, utilising 36 years of observational data from July 1961 to 1996. From the application results, the authors concluded that the proposed model could reproduce the historical rainfall time series well when the appropriate correlation between raincell intensity and duration is taken. In addition, they pointed out that the model-generated data with a positive correlation between rain cell intensity and duration is more robust for different parameter sets in the July rainfall time series at the Jeonju rain gauge. They stressed that the proposed model could improve rainfall modelling results and obtain more realistic synthetic rainfall time series.

In an attempt to widen the applicability of this proposed model, Han et al. [45] utilised the model to explore the potential for temporal downscale from hourly rainfall time series to minute rainfall time series. The authors noted that while hourly rainfall data has been observed and considered good quality data in long-term observation data, rainfall data with very short intervals, i.e. data interval of 10 min. or less, is needed to analyse flood events for a small urban drainage catchment. Hence, a method to downscale such well-qualified and quantified hourly rainfall time series into shorter time scales,

such as 10 minutes or less, is required in practical urban drainage system design. For application purposes, the model was simulated with data from July for 35 years, from 1961 to 1995, at the Seoul site of the Korean Meteorological Administration (KMA). The long-term simulation result regenerates the observed data well in terms of statistics. However, problems still existed, such as underestimating maximum rainfall depths and overestimating no-rain probability.

## 2.10 Bartlett-Lewis Pulse (BLP) Model

As mentioned earlier, in the classical Neyman–Scott (NS) and Bartlett–Lewis (BL) processes, rainfall intensity is considered a random variable that stays constant for a rain cell's lifetime, so rain cells are modelled using rectangular profiles. While rectangular profiles are unrealistic in continuous time, they provide a suitable approximation to discrete rainfall series aggregated over time intervals of one hour or more. Cowpertwait et al. [5] developed a stochastic model of rainfall which extends the Bartlett-Lewis (BL) model by adding a level of structure within the rain cells to extend the range of time scales over which it can be applied. More specifically, they replace the constant cell intensity and assume that each rain cell origin initiates a sequence of rainfall pulses that occur in a Poisson process. A schematic of the BLP model is shown in Figure 2.5.

Where Figure 2.5 illustrates several key processes: (a) the initiation of storm events, (b) the lifecycle, encompassing the birth and death, of cells within storm $i$, (c) the pulse sequences within cell $j$ of storm $i$, particularly for cells concluding prior to the storm's end, and (d) the sequence of precipitation pulses within cell $n$ of storm $i$, specifically for cells that concluded after the storm's termination. In the BLP model, storm origins occur in a Poisson process, and every storm possesses a stochastic lifetime wherein the origins of rain cells follow a secondary Poisson process. Moreover, each cell experiences a stochastic lifetime throughout which instantaneous rain depths (or 'pulses') manifest in an additional Poisson process. One motivation for developing the BLP model was to achieve an excellent fit to a series of rainfall depths over a range of time scales, from

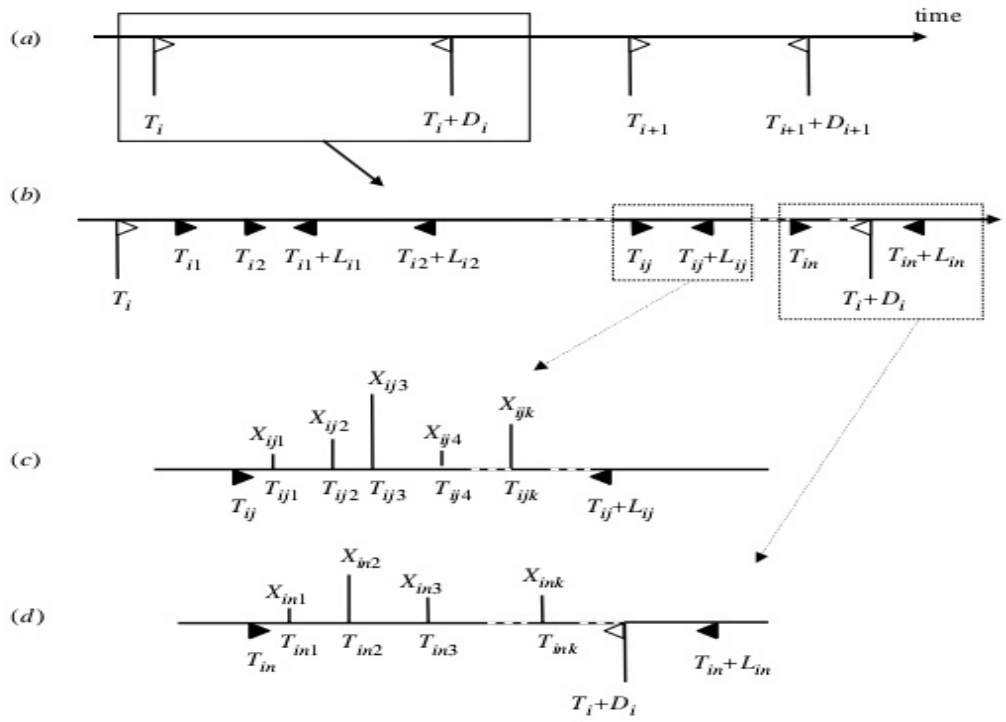


**Figure 2.5:** A schematic of the BLP model [5]

fine resolutions, e.g. 5 minutes, to higher aggregation levels, such as daily. In order to assess the fit of the BLP model at fine resolutions, the authors employed a 60-year record (1945–2004) of rainfall data recorded at a site in Kelburn (near Wellington, New Zealand). The data were based on a digitized pluviograph from a Dine's tilting siphon rain gauge and were aggregated over 5-minute intervals. The fitted properties of the BLP model generally agree well with observed values, indicating that the BLP model could model data for durations starting from 5 minutes and extending longer. This suggests that the BLP model has potential application in many areas, such as the urban drainage catchment studies, which usually require 5 minutes of rainfall series.

## 2.11 The Bartlett-Lewis Instantaneous Pulse Rainfall (BLIP) Model

Cowpertwait et al. [46] studied an extension of the classical Bartlett-Lewis rectangular pulses model, with the rectangular profiles replaced with a Poisson process of instanta-

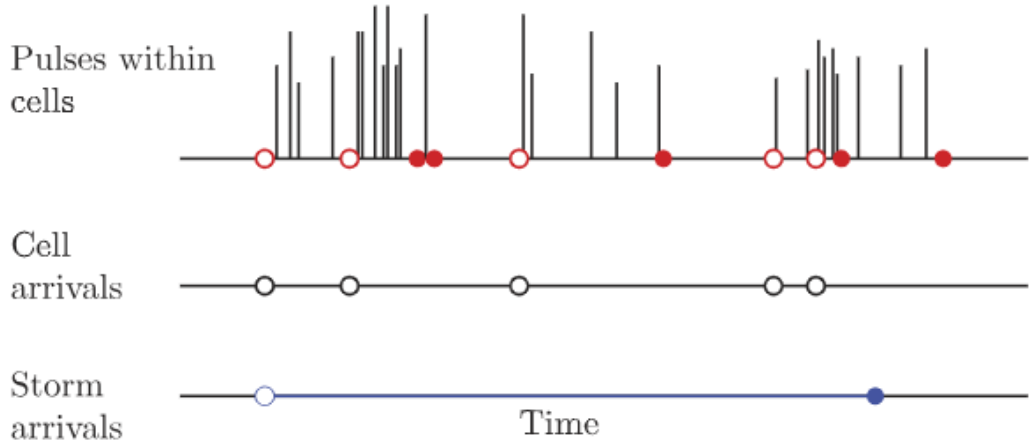


**Figure 2.6:** BLIP Model [6].

neous pulse depths to ensure more realistic rainfall profiles for fine-scale series. The BLIP model initially developed by Cowpertwait et al. [5] is based on three Poisson processes. The first is a Poisson process of storm origins, where each storm has a random (exponential) lifetime. The second is a Poisson process of cell origins that occurs during the storm's lifetime, terminating when the storm finishes. Each cell has a random lifetime that follows an exponential distribution (or terminates when the storm closes, whichever occurs first). During cell lifetimes, a third Poisson process of instantaneous pulses occurs. Figure 2.6 illustrates an individual storm event; the origin and conclusion of storms and cells are marked by unfilled and filled circles, respectively. In this scenario, every cell is characterized by a sequence of instantaneous pulse events. The BLIP model makes no assumptions about a cell origin at the storm origin or a pulse at a cell origin. Consequently, it's possible for both storms and cells to exhibit periods without any rainfall. Typically, recorded rainfall data are presented in a cumulative format, necessitating the aggregation of the BLIP process into a discrete-time series $Y_i^{(d)}$, expressed as:

$$Y_i^{(d)} = \int_{(i-1)d}^{id} X(t)dN(t), \tag{2.16}$$

In the expression, $X(t)$ represents the depth of a pulse occurring at time $t$, while $N(t)$ denotes the counting process for the occurrences of pulses. Referencing Equation 2.16,

the primary and secondary properties are delineated as follows:

$$\mu(d) = E[Y_i^{(d)}] = \lambda \mu_p \mu_X d \tag{2.17}$$

$$var[Y_i^{(d)}] = \lambda \mu_p E(X_{ijk}^2) d + 2A\mu_X^2 \phi(\gamma) + 2[B_1 E(X_{ijk} X_{ijl}) - B_2 \mu_X^2] \phi(\gamma + \eta) \tag{2.18}$$

$$cov[Y_i^{(d)}, Y_{i+k}^{(d)}] = 2A\mu_X^2 \psi(\gamma) + 2[B_1 E(X_{ijk} X_{ijl}) - B_2 \mu_X^2] \psi(\gamma + \eta) \tag{2.19}$$

Where $\lambda$ signifies the frequency of storm initiations, $\beta$ indicates the occurrence rate of cell formations, and $\xi$ represents the rate at which pulses arrive. Moreover, $\gamma^{-1}$ corresponds to the average storm lifetime, $\eta^{-1}$ refers to the average lifespan of cells, and $\theta_1$ denotes the mean depth of pulses.

The model specification due to Cowpertwait et al. [46], alongside the one reported by Cowpertwait et al. [5], was used to fit data consisting of a 60-year rainfall record (1945–2004) of 5-min series taken from a site in Kelburn (near Wellington, New Zealand). The simulation results provided solid evidence confirming that the BLIP model specification by Cowpertwait et al. [46] is more advantageous than the specification proposed by Cowpertwait et al. [5]. Subsequently, the authors emphasize the importance of leveraging this adaptability during model fitting, particularly for practical hydrological studies where understanding the characteristics of the 5-minute series is crucial.

## 2.12 A copula-based bivariate frequency analysis: A study on Bartlett-Lewis model

Vandenberghe et al. [47] stressed that the shortage of long-term rainfall records had given rise to relying on simulated rainfall time series through stochastic point process rainfall models (Bertlett-Lewis and Neyman-Scott models). Evaluating the effectiveness of stochastic point process rainfall models involves examining how well these models replicate extreme rainfall events achieved by conducting an extreme value or frequency analysis. An underestimation of the extremes by these models has been observed in the

literature [See [48, 49, 19]].

Vandenberghe et al. [47] suggested that instead of these univariate methods, multivariate methods for analysing extremes could provide a more powerful tool for assessing the performance of rainfall models. This led to the proposition of a copula-based frequency analysis of storms as a technique used to analyse the variations in the return periods of several different types of actual and simulated storms. In doing this, they examined several storm characteristics; the result showed issues with the model's representation of rainfall's time structure. Subsequently, the bivariate frequency analysis of storms, characterised by their duration and time, was used to highlight the models' miscalculations in the return intervals of the simulated storms. This discrepancy is partly due to significant variations in the marginal distribution functions for storm length and volume, the variation in the relationship between storm length and volume, and a distinct average interval between storms.

In conclusion, incorporating copulas into stochastic rainfall models proved advantageous for capturing the temporal dependence within the rainfall process, encompassing the structure internal to storms.

## 2.13 Hybrid Exponential GPD Model

Li et al. [50] critically examine the existing distributions and their prevalent issues, notably the underestimation of extreme values by nonparametric generators and the numerical and computational challenges of parametric ones. In response, the authors introduce a novel hybrid distribution that combines the strengths of exponential distribution for modelling low to moderate precipitation and the generalized Pareto distribution for extreme events. This merger ensures continuity at the junction, facilitating implicit, unsupervised learning of the threshold for the generalized Pareto component, addressing a significant challenge associated with traditional parametric generators.

$$f(x) = \frac{1}{Z}\left[f_e(x,\nu)I(x \leq \theta) + f_{gp}(x;\kappa,\rho,\theta)I(x > 0)\right], \quad x \geq 0, \quad \nu,\kappa,\rho,\theta > 0 \tag{2.20}$$

Equation 2.20 denotes the PDF of the hybrid model, integrating to 1 over its domain due to the normalization constant $Z$. It combines an exponential component $f_e(x,\nu)$, parameterized by $\nu$, with a Generalized Pareto (GP) component $f_{gp}(x;\kappa,\rho,\theta)$, described by scale $\kappa$, shape $\rho$, and threshold $\theta$ parameters. The PDF is defined for $x \geq 0$, where $\nu,\kappa,\rho,\theta$ are all positive, and utilizes an indicator function $I(\cdot)$ to differentiate between the exponential and GP parts of the model. The authors further substantiate the efficacy of the hybrid model through Monte Carlo simulations and empirical testing using 49 daily precipitation records from Texas. The model's functional simplicity, allowing for easy random number simulations, is a significant advantage for practitioners. While the study provides a strong foundation and promising tool for daily precipitation modelling, the authors rightly acknowledge the need for broader evaluations beyond Texas and emphasize the potential for regional adaptations.

## 2.14 Doubly Stochastic Point Process Model

Ramesh et al. [7] considered a unique way of representing the clustering of rainfall within storms by using doubly stochastic (Cox) models. A doubly stochastic Poisson process (DSPP) originates from a Poisson process when a non-negative stochastic process determines the arrival rate of the process. The authors' first attempt was to describe a class of univariate models, based on a class of doubly stochastic Poisson processes, to analyse the tip times measured by a tipping bucket rain-gauge and then to illustrate the development of the bivariate models to analyse rainfall bursts at two rain gauges. The parameters of the univariate and bivariate models were derived using the MLE approach. Figure 2.7 presents the schematic of the three-state processes where the arrival rate is highest in the State. The line in Figure 2.7 represents a realisation of the underlying Markov chain, and the occurrence times of a realisation of the DSPP are shown as a

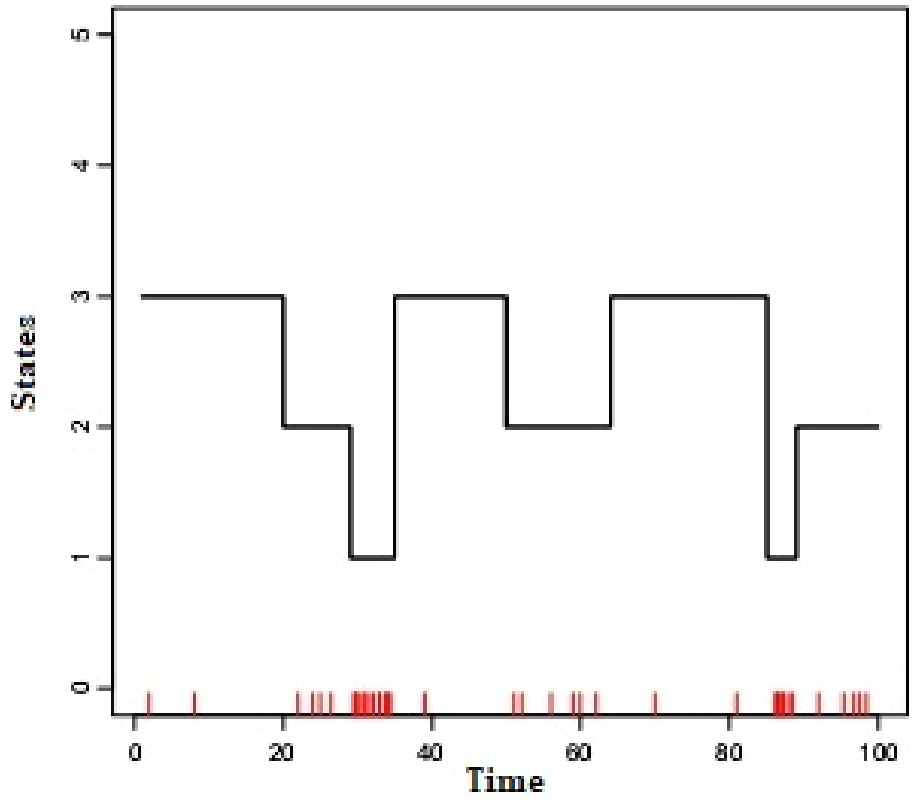


**Figure 2.7:** A schematic of the 3-state process with intense rainfall events in State 1 [7]

comb at the foot. The univariate model was fitted to nearly 13 years (1988 September to 2001 July) of 0.2 mm rainfall bucket tip times data for Heathrow West of London, UK station. In contrast, the bivariate model was fitted to 5 years (1994–1998) of rainfall bucket tip times data for the two stations Mead (Meadow Buildings) and Kite (Kite Lane) from the HYREX rain-gauge network in Somerset, South West England, UK. From the analysis results, the univariate models reproduced many of the rainfall characteristics well at most sub-hourly time scales. On the other hand, the analysis revealed that the simulated data from the fitted bivariate model not only reproduces the properties of rainfall aggregations reasonably well at different time scales for both gauges but also captures the cross-correlations of the rainfall intensities.

## 2.15 Enhanced Modeling Through the Random Parameter Bartlett-Lewis Instantaneous Pulse (BLIPR) Model

Achieving precise modelling for rainfall events over diverse timescales, particularly at the granular level of 5-minute intervals extending up to daily durations, presents a significant challenge. The innovation by Cowpertwait et al. [5] through the introduction of the Bartlett-Lewis Instantaneous Pulse (BLIP) model marked a pivotal shift from the traditional rectangular pulse framework of the classical Bartlett-Lewis model. By employing a Poisson process for instantaneous pulses, the BLIP model advanced the realism of rainfall time series representation at finer scales, allowing for dependent pulse depths within the same cell and independent depths across different cells.

Expanding upon this foundation, Kaczmarska et al. [6] presented the Random Parameter Bartlett-Lewis Instantaneous Pulse (BLIPR) model, an evolution of the Random Parameter Bartlett-Lewis Rectangular Pulse (BLRP) model conceptualized by Rodrigues-Iturbe et al. [27]. This model variation introduces a dynamic rainfall intensity parameter, $\eta$, which adapts according to the cell duration parameter, enabling a more nuanced simulation of rainfall events. The BLIPR model's framework was tested against 69 years of 5-minute resolution rainfall data from Bochum, Germany, demonstrating its superior ability to model rainfall moments, wet/dry spell characteristics, and extreme rainfall events compared to its predecessors, BLRP and BLIP.

The BLIPR model innovates by randomizing the parameter $\eta$, maintaining a constant ratio $\omega = \xi/\eta$ between the pulse arrival rate and the cell duration parameter, thereby enriching the model's flexibility. The computation of the model's moments involves considering it as a composite of independent processes, each characterized by a unique cell duration parameter, $\eta$, and a storm origin rate, $\lambda f(\eta)$, where $f(\eta)$ denotes the density function of $\eta$. This approach necessitates integrating across the spectrum of $\eta$ values to deduce the mean, variance, and third central moment of the aggregated rainfall. Such integration hinges on expectations of functions involving $\eta$, specifically:

$$E\left[\frac{1}{\eta^k}e^{-\eta x}\right] = \nu^{\alpha}\Gamma(\alpha)^{-1}\int_0^{\infty}\eta^{\alpha-1-k}e^{-(\nu+x)\eta}d\eta = \nu^{\alpha}\Gamma(\alpha)^{-1}\Gamma(\alpha-k)(\nu+x)^{\alpha-k} \quad (2.21)$$

Where the condition $\alpha > k$ ensures the integrals' convergence, a critical adjustment from the original Bartlett-Lewis model to avert divergence at zero and accommodate the variance and skewness calculations without inducing unrealistically prolonged rainfall events.

The empirical adaptation of the BLIPR model preserves the allowance for dependent pulse depths within a single cell, a modification poised to correct the underestimation of short-duration extreme rainfall values observed in prior models. This nuanced modelling capability underscores the BLIPR model's advancement in simulating high-intensity rainfall events more accurately, providing a robust framework for understanding and predicting extreme weather patterns. Empirical validation of the BLIPR model utilizing extensive rainfall data underscored its enhanced performance across several metrics. By rigorously comparing the fitted moments, the propensity for wet and dry conditions, and the accuracy in capturing extreme rainfall values, the BLIPR model consistently outmatched its predecessors, offering a more refined tool for detailed rainfall simulation and analysis [5, 6, 27].

## 2.16 Doubly Stochastic Point Process Exponential Pulse Model

Inspired by the favourable outcomes achieved through the utilisation of doubly stochastic Poisson point processes in contrast to Poisson cluster processes as the underlying point process, Ramesh et al. [8] recognised the necessity of associating an exponentially decaying pulse with each point of such a process, particularly when the objective is to replicate the characteristics of fine-scale rainfall. Thus, the authors developed a class of doubly stochastic Poisson process (DSPP) models featuring exponentially decaying

pulses to characterise the probabilistic nature of rainfall measurements obtained from a single rain gauge. The second-order moment characteristics of the rainfall intensity and aggregated rainfall processes were derived. Figure 2.8 provides a schematic description of the pulse process.

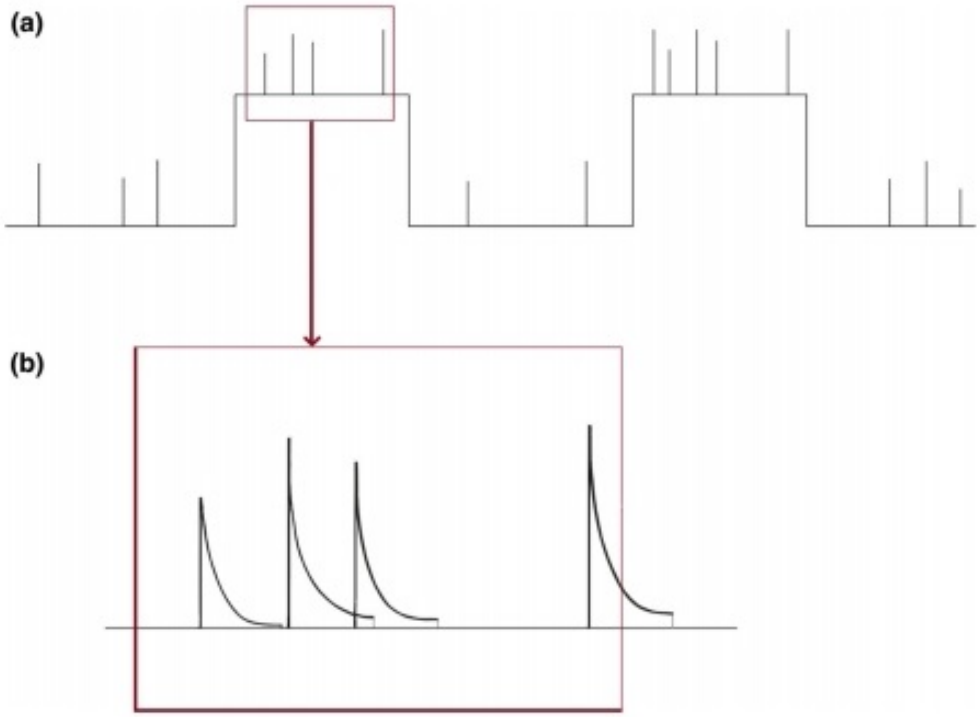


**Figure 2.8:** Schematic description of the DSPP exponential pulse model [8]

Figure 2.8 (a) represents the arrival process of rainfall bursts, depicted through a two-state DSPP, whereas (b) illustrates the pulse process initiated by each burst, persisting for a predetermined duration of $d$. To compare how well the proposed exponentially decaying pulse model works with existing point process models of rainfall, the authors looked at a doubly stochastic rectangular pulse model by Ramesh [51] since the structure of the cell arrivals in both models is the same. An investigation was conducted on a dataset consisting of sub-hourly rainfall data from England spanning 15 years. The proposed model was compared to a double-stochastic rectangular pulse model with the same structure for cell arrivals to see how well it could reproduce the statistical properties of the total rainfall. Both models performed equally well in reproducing the mean rainfall. Nevertheless, the proposed exponential pulse model did better at sub-hourly

aggregations for most other properties considered and even higher aggregations for some. The rectangular pulse model exhibited superior performance solely in the context of lag one autocorrelation at higher levels of aggregation.

## 2.17 A Simple, Flexible and Parsimonious Stochastic Rainfall Model (STORM)

Singer and Michaelides [52] emphasised the need for a simple rainstorm generator that investigates rainfall's spatial and temporal variability in stationary or nonstationary climates (climate change). They noted that most existing rainstorm generators are too complex for simple investigative simulations of convective rainfall under climate change in small basins. In particular, a significant drawback of most of these generators is the reliance on the general circular model (GCM) to characterise climate change. Attempting to develop a simple rainstorm generator, the authors introduced the STOchastic Rainfall Model (STORM) for convective storm simulation. The model uses an empirical-stochastic approach, which involves assembling probability distributions of crucial rainstorm characteristics, followed by Monte Carlo sampling to simulate rainstorms' spatial and temporal variability across a spatial grid. One of the notable aspects of the model is its inherent capacity to forecast the reaction of a watershed to future climate variations. This is achieved by selectively modifying or adjusting pertinent input distributions to account for the anticipated impacts of probable climatic changes.

The researchers utilised the STORM method to evaluate historical climate change impacts on rainfall patterns within a dryland basin located in the Lower Colorado River basin. Furthermore, they explored its imprint on ephemeral channel flow contributions to larger regional rivers, where there are observations of multi-decadal declines in streamflow. Simulation results revealed that STORM produced a corresponding output consistent in magnitude with the historical record of precipitation and runoff for the multi-decadal period of interest. The STORM can provide insights into the probable watershed responses to multi-decadal precipitation changes for research or management applications.

## 2.18 A Simple, Flexible and Parsimonious Stochastic Rainfall Model (STORM-REVISITED)

Singer et al. [9] revisited the STORM (initially reported in previous work) to simulate drainage basin rainfall. Various modifications of the STORM introduced by the authors include;

i) incorporating a randomly sampled probability density function (PDF) of inter-storm periods after each storm event. The incorporation of inter-storm intervals resulted in a modification of the STORM output, transforming it into a time series that accurately represents the real-time conditions occurring at the Earth's surface;

ii) compiling a PDF containing potential evapotranspiration data derived from temperature and relative humidity observations, which are conveniently obtainable across several temporal and spatial scales;

iii) incorporating seasonality in rainfall patterns to facilitate simulations across a specific season, a year, or two seasons with discernible variations in ten precipitation attributes. These attributes encompass unique probability density functions (PDFs) of rainfall during summer in contrast to winter.

STORM is a hybrid empirical-stochastic rainfall simulator tailored for the heuristic generation of high-resolution rainfall across drainage basins, adaptable to specified climate scenarios or varying climate change classes. The deployment of STORM, depicted in Figure 2.9, encompasses initialization and operational phases.

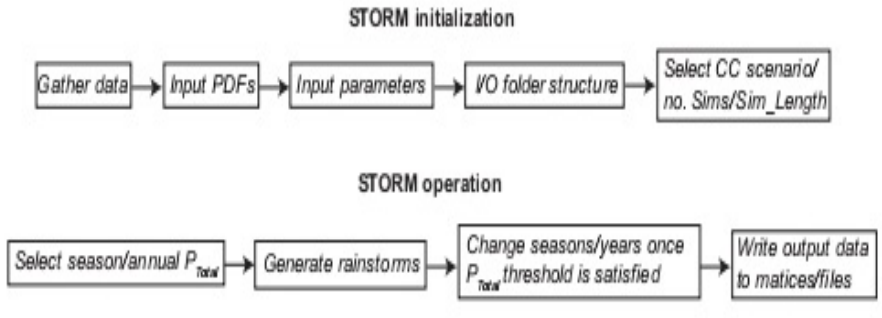


**Figure 2.9:** Schematic flow diagram illustrating key steps in STORM initialization and operation [9]

In Figure 2.9, the initialization phase encompasses the collection of data, formulation of input Probability Density Functions (PDFs), derivation of input parameters, structuring of an input/output (I/O) directory, and the choice of a climate change scenario, in addition to determining the count and duration of simulations. The operational phase involves setting a threshold for seasonal or annual precipitation totals, subsequently generating rainstorms until this threshold is reached, triggering a transition to the next season or year. Subsequently, the outputs are organized into written matrices and files. STORM accommodates one or multiple hydrological seasons, each distinguished by unique rainfall properties.

The authors observed that the enhancements made to STORM have now rendered it an appropriate climate driver for additional watershed response models designed to simulate the hydrological processes occurring between slopes and channels, including surface runoff, infiltration, and streamflow. ( see [53];[54];[55]), groundwater recharge during and after rainfall events (see [56]), and interactions between streamflow and alluvial aquifers (see [57]). The enhancement further allows STORM to be beneficial in water balance models (for instance, Land Surface Models) for evaluating plant water availability through dynamic ecohydrological simulations that model the interactions between plant systems and climate and water utilization (see [58]; [59]; [60]).

## 2.19 A Censored Approach to Bartlett-Lewis Model

Modifications to enhance the precision of fine-scale extreme rainfall predictions at individual sites using mechanistic models have primarily focused on model structure alterations. An initial critique of the standard mechanistic models (BLRP and NSRP) introduced by Rodriguez-Iturbe et al. [25] highlighted the inadequacy of using an exponential distribution for rainfall intensities due to its light-tailed nature. Cross et al. [61] proposed a refined, censored methodology for mechanistic rainfall modelling to improve the estimation of fine-scale extremes by concentrating on the heavier segments of the rainfall series. Their research examined the capability of mechanistic models to act as simulators for detailed design storm events, intending to minimize the influence of smaller magnitude observations on extreme value estimation. This approach involves adjusting the rainfall data such that readings below a specified low threshold are reset to zero while values above this threshold are reduced accordingly. This technique creates a modified time series that emphasizes significant rainfall events, augmenting the proportion of dry intervals and diminishing the magnitude of recorded rainfall amounts.

They emphasized that this method of censored rainfall synthesis is suitable for predicting near-hourly extremes. The exclusion of data below the censoring threshold during model calibration means that the resulting model parameters are scale-specific. These parameters are thus tailored to simulate storm patterns above the threshold, corresponding to the scale of the data used for calibration. The method's efficacy lies in its ability to replicate the more substantial sections of storm patterns, which is crucial for estimating extreme rainfall events. The implementation strategy described by the authors involves four main steps:

1. Choose an appropriate threshold (in mm) for the desired temporal resolution and apply it to the observed rainfall series by setting measurements below the threshold to zero and adjusting values above the threshold by the threshold amount.

2. Adapt the mechanistic rainfall model to the altered data by aggregating the adjusted

series over various time scales and computing the necessary summary statistics for model fitting.

3. Generate synthetic rainfall sequences at the exact resolution as the training data, then extract and record the annual maxima.

4. Reintegrate the threshold adjustment into the simulated annual maxima for comparison with the observed maxima.

This approach was tested on Atherstone in the UK and Bochum in Germany. The traditional Bartlett-Lewis model along with two modified Bartlett–Lewis rectangular pulse models (BLRPR) by Onof and Wheater [62] and (BLIPR) by Kaczmarska et al. [6] were evaluated. All three model variants demonstrated reliable accuracy in estimating sub-hourly rainfall extremes. Nevertheless, the BLIPR model displayed superior performance at both sites for five and 15-minute intervals, especially in accurately forecasting the most extreme observed rainfall events.

## 2.20 A Hybrid Rainfall Model based on the MBLRP and SARIMA Models

Park et al. [4] created a hybrid rainfall model that can recreate different statistical features of observed rainfall on timescales from 1 hour to 1 year. First, The hybrid model employs a seasonal autoregressive integrated moving average (SARIMA) model to produce the monthly rainfall time series. Afterwards, it downscales the generated monthly rainfall time series to the hourly aggregation level using the modified Bartlett-Lewis rectangular pulse (MBLRP) model developed by Rodrigues-Iturbe et al. [27].

The authors emphasised the novelty of the proposed hybrid model in that; (i) The monthly rainfall values are utilised to generate monthly statistics, which are then employed to calibrate the MBLRP model. (ii) The individual monthly rainfall values generated are downscaled using month-specific MBLRP model parameter sets. These pa-

rameter sets capture the intricate correlation structure of different rainfall statistics, including mean, variance, covariance, and proportion of dry periods. This approach is based on Poisson cluster rainfall models, unlike existing composite approaches (see [63] and Paschalis et al. [64]) which showed problems with reproduction, especially at smaller than daily scales. The MBLRP model is cautiously calibrated in this methodology to accurately replicate the sub-daily statistical characteristics of observed rainfall.

The unique methodology employed in the hybrid model allows for precise replication of first- to third-order statistics, the proportion of dry periods across various times ranging from one hour to one year, and the statistical characteristics of monthly maximum values and extreme rainfall events observed in the data. The authors employed hourly rainfall data observed at 34 gauges in the Midwest to the east coast of the continental United States between 1981 and 2010. The fitting results suggested that the hybrid model accurately reproduced the first- to third-order statistics, and the study successfully replicated the intermittent characteristics across several timescales, ranging from hourly to annual. It accurately reproduced the data's statistical patterns of monthly maximum rainfall and extreme values.

## 2.21 Copula-based Stochastic Sub-hourly Rainfall Generation Model

Brigandi and Aronica [20] categorized rainfall stochastic generation models into two primary types: profile-based and pulse-based. The profile-based model concentrates on individual rainfall events, identifying the time between events and utilizing their joint or distinct probability distributions to detail a storm's essential characteristics. On the other hand, the pulse-based model views storm events as isolated incidents occurring randomly over time, with their formation modelled by a Poisson process. Each rainfall event generates a series of rain cells, envisioned as pulses with varying duration and intensity but generally maintaining a consistent intensity during the cell's lifespan.

Pulse-based models are known for their precise depiction of continuous rainfall sequences,

making them valuable for various hydrological applications. However, they require estimating numerous parameters and an extensive record of continuous historical rainfall data. Prior studies, such as those by Cameron et al. [35], highlighted that while pulse-based models effectively replicate the observed timings between rainfall events across different scales, they might not accurately simulate the extreme short-term statistics. Consequently, many researchers ([65]; [66]; [67]) have favored profile-based models for their studies.

Brigandi and Aronica [20] embraced the profile-based approach and developed a method to generate stochastic sub-hourly rainfall at specific locations. Their model's key benefit is its minimal data requirement, needing only a few years of high-resolution rainfall data for calibration, albeit not necessarily continuous. The model stochastically generates rainfall events, modelling their duration and average intensity through a bivariate copula-based framework, while dimensionless mass curves are used to define the event's shape.

10-minute interval rainfall records from two Sicilian sites were used to calibrate this model. The data spanned from 2003 to 2009 at the Monreale station and from 2002 to 2007 at the Palazzolo Acreide station. The model's outcomes validated the effective use of Frank's copula for modelling the interdependence between storm duration and intensity, maintaining the inherent correlation of the variables. Furthermore, the good agreement between the historical data and the model-generated values robustly supports the model's ability to accurately produce extreme rainfall events, affirming its potential to create long sequences of synthetic sub-hourly rainfall that reflect the vital hydrological features of the location.

## 2.22 A Cox Process with State-Dependent Exponential Pulses

Ramesh et al. [10] delved into the application of doubly stochastic Poisson process models (initiated by Ramesh et al. [7]) for simulating actual rainfall observations. The distinctive aspect of their modelling strategy lies in its detailed representation of an unseen

state of the rainfall system, which corresponds to the atmospheric conditions driving the generation of rainfall. The team formulated a series of Cox process models characterized by exponentially diminishing pulses, with the initial pulse depth's distribution relying on the state of the background Markov chain. This innovative model articulates the probabilistic framework of rainfall at individual rain gauges. The researchers evaluated two iterations of the proposed model; the first assumes a fixed duration $d$ for the life of the rain pulses, whereas the second iteration considers the pulse duration $d$ as a stochastic variable.

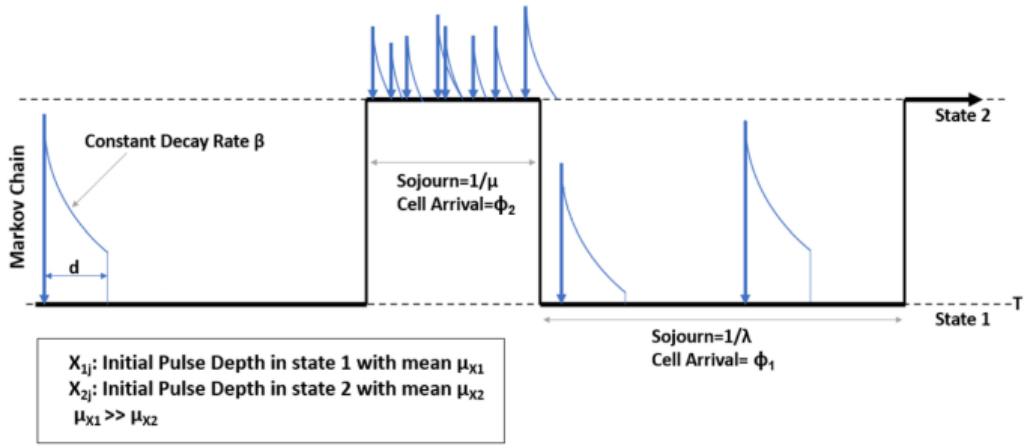


**Figure 2.10:** Illustration of the state-dependent initial depth exponential pulse model with a set pulse duration $d$.[10]

The model captures the arrival dynamics of rain cells at a specific site through a stationary Cox process influenced by a two-state continuous-time Markov chain. This chain distinguishes between low and high-intensity rainfall states, with transition rates and rain cell arrival rates that change accordingly. Each rain cell starts with an initial depth that undergoes exponential decay over time, influenced by the Markov chain's state at the inception of the cell. The average initial depths are distinct between the states. The rain pulses last for a set period, maintaining independence from each other and the cell arrival process. A variable is designated to quantify the rainfall depth at any moment, and the rainfall intensity at any time point is the aggregated result of all active rain pulses. Figure 2.10 displays the model description.

Empirical validation was conducted using two sub-hourly rainfall datasets: a 15-year record from Bracknell, England, sourced from the U.K. Meteorological Office, and a 69-year record from Bochum, Germany. The simulations demonstrated that both models are adept at replicating the second-moment characteristics of rainfall. Notably, the model with variable pulse duration exhibited enhanced congruence between the observed data and the model's outputs. Moreover, the efficacy of this novel model in reflecting the second-moment features of rainfall was compared against two other stochastic models—one with exponential pulses and the other with rectangular pulses. The introduced model proficiently captured the empirical rainfall characteristics and surpassed the comparative performance of the two other models considered in this analysis.

## 2.23 Stochastic Rainfall Models involving Markov Chain Model

By simple definition, a Markov chain is a discrete-time stochastic model defined on a space of states equipped with transition probabilities from one state to another at the next time stage. Here, we revisit some stochastic rainfall models (SRMs) based on the Markov chain model.

Gao et al. [11] developed a stochastic rainfall model termed SDRM-MCREM, integrating a Markov chain model with the stochastic rainfall event model previously investigated in [68]. This model is engineered to produce rainfall time series that maintain the intrinsic characteristics of rainfall events. The SDRM-MCREM initially employs the Markov chain model to create a series of rainfall occurrences, delineating both wet and dry spells. Subsequently, it utilizes the rainfall event model to stochastically generate a sequence of rainfall events aligned with the wet spells identified from the created rainfall occurrence series. A notable aspect of SDRM-MCREM is its dual focus, capturing the statistical nuances of rainfall time series and addressing the distinct features of rainfall events. This includes accounting for the relationship between rainfall depth and duration, the distribution of different classes of rainfall events, and the diversity of temporal rainfall

patterns along with their frequency distribution across various event categories. The innovative stochastic rainfall model SDRM-MCREM schematic representation is depicted in Figure 2.11 and encapsulates three primary modules.

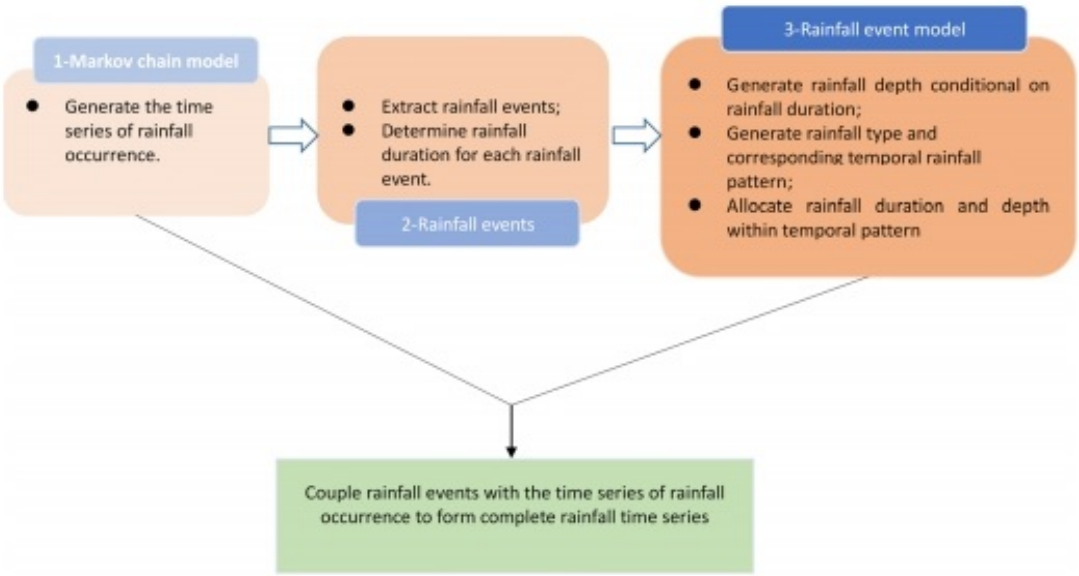


**Figure 2.11:** A flowchart of SDRM-MCREM [11]

From Figure 2.11, the initial module produces the time series of rainfall occurrences. The subsequent module identifies and isolates rainfall events according to a predefined criterion, further determining the duration of each identified rainfall event. The third module is tasked with replicating the attributes of these rainfall events, encompassing both the rainfall depth and the temporal pattern of the rainfall, contingent on the event's duration, modelled in alignment with the statistical patterns observed in actual rainfall data.

The SDRM-MCREM model was implemented in the Qu River basin in East China, and its efficacy was assessed at the catchment scale. The outcomes demonstrated that SDRM-MCREM proficiently replicated a majority of the statistics related to the rainfall time series (such as rainfall percentiles, mean monthly and annual rainfall, variability of rainfall between months, and extreme rainfall occurrences) along with the characteristics associated with rainfall events (involving the distribution trends of wet and dry intervals,

the incidence rate of distinct categories of rainfall events, and the sequential rainfall schemes together with their commonality within different rainfall event categories).

An immediate advancement from the single-site stochastic rainfall model SDRM-MCREM was made by Gao et al. [12], who introduced a multi-site stochastic daily rainfall model, MSDRM-MCREM, which integrates a univariate Markov chain with a multi-site rainfall event model. In MSDRM-MCREM, the univariate Markov chain model is utilized to produce spatially correlated rainfall occurrence series for multiple sites and to identify simulated rainfall events at each station, delineated by consecutive wet periods. Subsequently, the multi-site rainfall event model is employed, employing Vine copulas to create a simulation framework for spatially correlated characteristics of the rainfall events that transpire concurrently at several stations. This includes modelling these concurrent events' rainfall durations, depths, and temporal patterns. The detailed framework of the developed multi-site stochastic daily rainfall model coupling a univariate Markov chain model for multi-site rainfall occurrences (0 or 1 values) and a multi-site rainfall event model using Vine copulas (called MSDRM-MCREM) is shown in Figure 2.12.

In Figure 2.12, the initial phase involves applying a univariate Markov chain model (Breinl et al.[69]) to produce cross-correlated rainfall occurrence sequences across multiple locations. The subsequent phase entails isolating rainfall events by identifying successive wet periods at each station, determining their durations, and categorizing these events into various groups that represent concurrent occurrences at multiple stations (Callau et al.[70]). The third phase employs Vine copulas to flexibly model the dependency structures of multiple variables within each group, specifically combinations of rainfall duration and depth at various sites, and to create multi-site rainfall depths for given durations utilizing conditional Vine copulas. The fourth phase involves generating multi-site rainfall types and temporal patterns based on their frequency probabilities within each group. The final phase, which is the fifth, consists of synthesizing complete rainfall events by amalgamating rainfall depth, duration, and temporal patterns, subsequently reassigning the categorized rainfall events to the respective stations. This is followed by integrating the sequenced rainfall events of each station into its rainfall occurrence series,

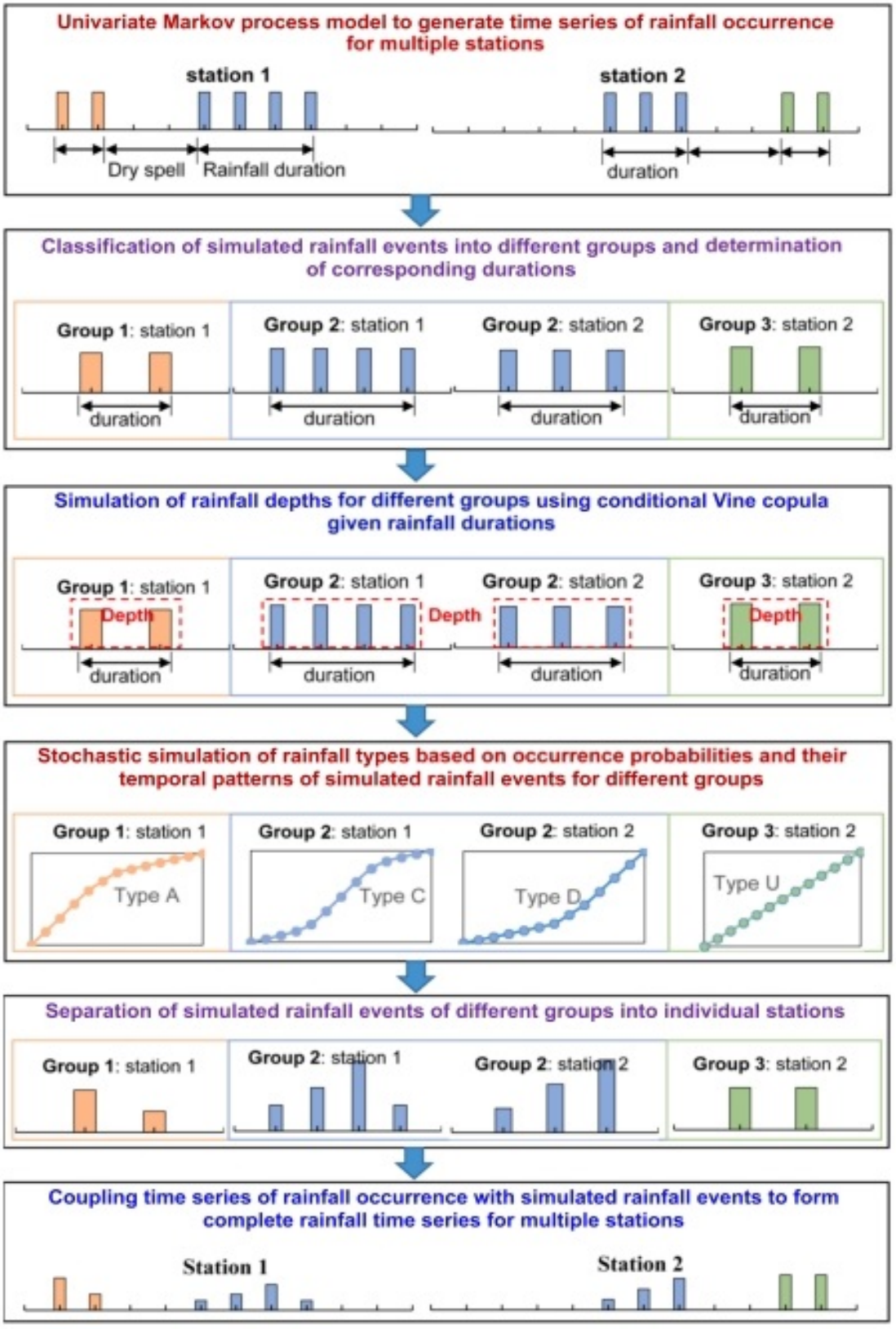


**Figure 2.12:** A framework of MSDRM-MCREM considering two stations as an example [12]

culminating in a comprehensive spatially correlated rainfall series.

The MSDRM-MCREM was deployed in the Changshangang River basin in Zhejiang Province, East China, to assess its capability to mimic rainfall features and spatial correlations. The evaluation covered simulations at two, three, and four-station setups. Analysis outcomes indicated that aside from a tendency to overestimate minor rainfall events, the MSDRM-MCREM proficiently maintains essential statistics of rainfall time series (such as various rainfall percentiles, average monthly rainfall, standard deviations, probabilities, and average counts of wet days), extreme rainfall events (like the exceedance probabilities for annual maximum 1-day, 3-day, and 5-day rainfall totals) as well as characteristics of rainfall events (including cumulative probabilities for wet spells, dry spells, and rainfall depths, along with temporal patterns and likelihoods of distinct rainfall types categorized by event depth) at the individual stations.

Nop et al. [71] proposed a methodology to develop a Markov chain model tailored for rainfall time series in temperate regions, incorporating it into the stochastic dynamic programming framework for optimizing rainwater harvesting (RWH) systems operations. Dynamic programming is notably prevalent in deriving optimal policies for reservoir management. In their model, the Markov chain states represented intervals of rainfall depths over 10-minute periods, with the month-specific transition probabilities defining the Markov chain's behaviour. Frequent dry periods allowed for empirical estimation of transition probabilities from a dry state. Additionally, when a wet state was noted, the subsequent 10-minute rainfall depth was modelled to follow a gamma distribution characterized by two parameters. A distinctive aspect of their approach is the ability to create a multi-state Markov chain model from scarce data sets. They illustrated the Markov chain model's effectiveness in optimizing water resource management and storm-water retention using hypothetical rainwater harvesting system operations scenarios showcased within the stochastic dynamic programming framework.

In a recent exploration, Chauhan et al. [72] employed the extreme value distribu-

tion of frequency analysis along with the Markov Chain model to scrutinize the hydro-meteorological data at the Dakpathar barrage, situated in the Yamuna River Basin, Uttarakhand, India. Their technique scrutinizes persistence and allows for the computation of joint probabilities such as initial and transition probabilities. The investigators underscored the merits of the Markov Chain strategy for rainfall prediction in the study locale. Primarily, it yields a trustworthy forecast of upcoming rainfall trends derived from historical records. Secondly, its construction is straightforward and requires minimal computational capacity, rendering it ideal for environments with limited resources. Notably, in scenarios where data is scant, the stochastic Markov Chain methodology surpasses sophisticated artificial intelligence models like LSTM.

An added benefit of the Markov Chain method is its capacity to produce precise forecasts with scant training data. Additionally, the Markov Chain approach is more interpretable, aiding researchers in better comprehending the systems influencing rainfall patterns. Summarily, the data analysis outcomes indicated that the Markov Chain model had a success rate of 79.17%, suggesting that extended return periods should alert to potential drought and flood risks in the Himalayan region.

## 2.24 Rainfall Event Models involving Spatial Weather Systems

The escalating concerns about intensified heavy rainfall events under the evolving climate scenario highlight a pressing need for societal awareness. As Kundzewicz et al. [73] articulated, extreme precipitation events are precursors to floods and landslides, significantly influencing agriculture, ecosystems, and human settlements. Enhancing our adaptive capabilities and resilience against these natural phenomena necessitates a comprehensive understanding of the alterations in extreme rainfall patterns. The prediction of rainfall events is marred with uncertainties due to the unpredictable nature of weather systems and the intricate interactions between atmospheric dynamics and geographical landscapes. According to the insights from Clark et al. [74], simulating rainfall events aim

to accurately forecast rainfall intensity variations as weather systems navigate diverse geographical terrains. This task is achieved through advanced computational models that leverage real-time atmospheric observations, historical climatological data, and digital terrain models to predict the trajectories and transformations of meteorological systems.

In this context, Chen et al. [75] embarked on an investigation to project the forthcoming variations in intense summer rainfall, employing an advanced, high-resolution climate model tailored for the UK (convention-permitting model). Their analysis focused on anticipated alterations in the intensity, spatial distribution, and duration of rainfall events, alongside their collective impact on hourly extreme rainfall across various spatial dimensions. Their comparative study of past and projected future conditions across three distinct UK regions unveiled a potential increase in the intensity and spatial expanse of heavy rainfall episodes, with projections showing up to a 49.3% expansion in the northwest region.

Peleg et al. [13] highlighted the importance of adapting design storms, which are crucial for evaluating flood risks, to reflect changes in the frequency and intensity patterns of extreme rainfall due to climate change. They developed a spatial quantile mapping (SQM) approach to enhance the use of high-resolution data from convection-permitting models (CPM) in hydrological impact studies of floods. Their approach utilized detailed rainfall simulations from a CPM for various urban regions in Switzerland, simulating extreme weather scenarios consistent with current climate trends using a 2D stochastic model. The research involved adjusting existing design storms to align with future climate projections through SQM, along with two other methods: uniform quantile mapping and an adjustment correlating rainfall with temperature. The methodology comprised steps like extracting spatiotemporal data from radars, generating representative storm patterns, and adjusting these patterns to reflect future climate scenarios, subsequently applying them in flood modelling. A schematic representation of the steps is shown in Figure 2.13.

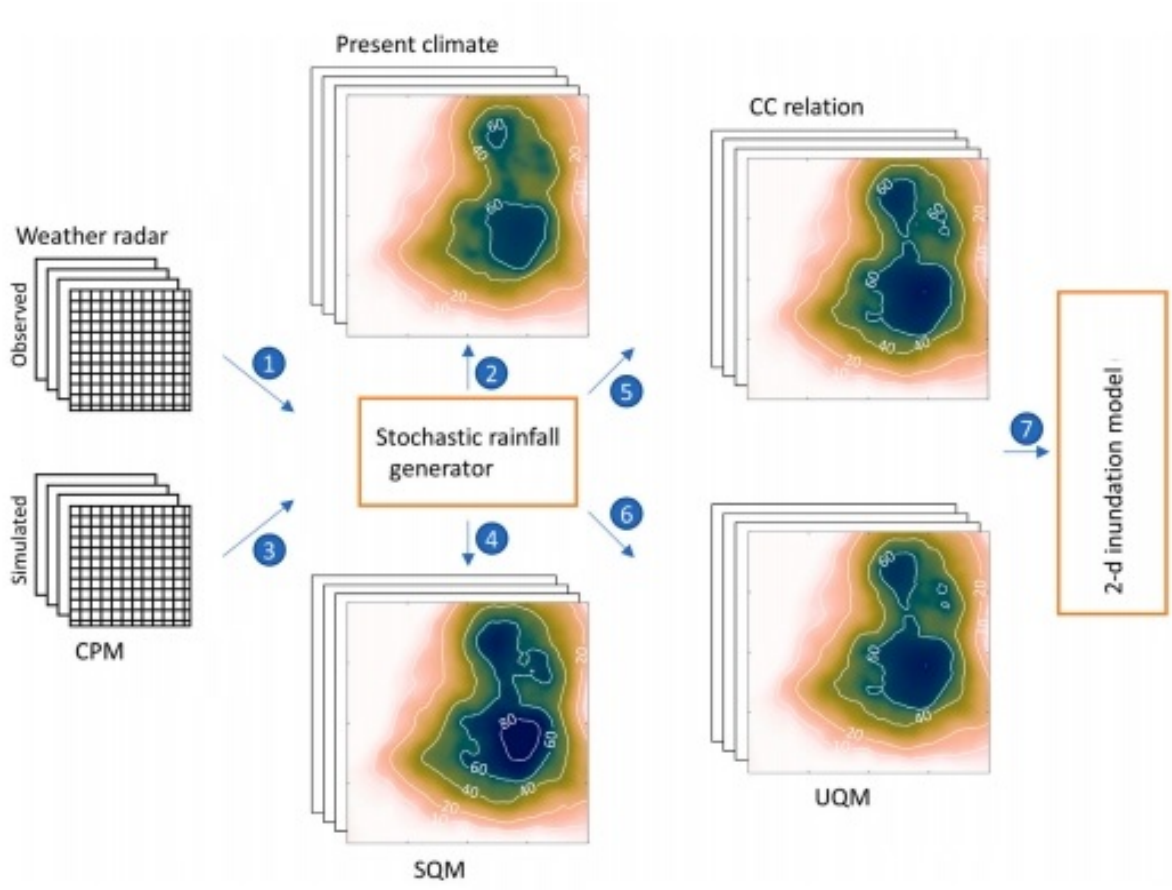


**Figure 2.13:** A schematic representation of the steps taken in the study [13]

Their findings showed distinct variations in flood risks when applying different rainfall adjustment techniques. The SQM method notably impacted flood severity more than the other methods, underscoring the significance of considering spatial rainfall patterns in future hydrological assessments.

## 2.25 A Multi-site Stochastic Weather Generator: The Generative Adversarial Network (GAN)

Ji et al. [14] highlighted the reliance of conventional multi-site stochastic weather generators (SWGs) on intricate parameterizations, which capture the inherent spatial and temporal sporadicity of meteorological variables and their quantities. This complexity might result in inadequate sampling, failing to accurately represent extreme weather phenomena, such as rainstorms or droughts, and their spatial interconnections. Like-

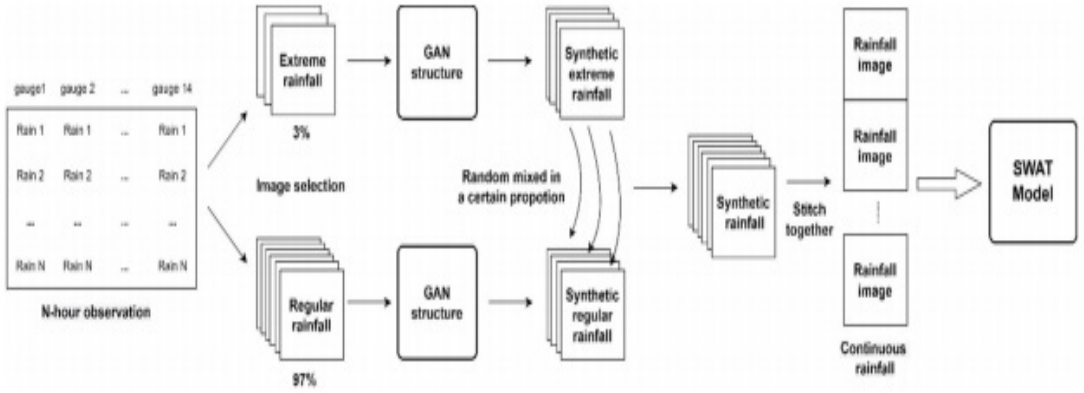


**Figure 2.14:** The workflow of GAN taking into account extreme rainfall events [14]

wise, the authors pointed out that non-parametric methods relying on resampling, like the K-nearest neighbour (KNN) approach, fall short because they cannot generate values beyond the historical data's scope. The authors recommended using advanced deep learning methods like generative adversarial networks (GANs) to avoid setting these complex parameters and ensure that data from different locations are consistent and interact well. These models learn from the data and are more adaptable at creating believable weather patterns than standard statistical approaches.

In their study, the authors specifically examined how well synthetic rainfall data matched actual observations across multiple sites and on an hourly basis, which they described as a two-dimensional approach. They incorporated these data into an hourly calibrated Soil and Water Assessment Tool (SWAT) model, enabling them to simulate continuous hourly flow patterns. This was then used for flood frequency analysis, considering the natural uncertainties in the rainfall input data. Figure 2.14 displays the workflow of a Generative Adversarial Network (GAN) taking into account extreme rainfall events.

The data analysis segment gathered hourly precipitation measurements from 14 stations within the Kelantan River Basin from January 1, 1990, to December 31, 2019, over 30 years, setting 0.1 mm as the minimum threshold for hourly rainfall. They also compiled additional variables like the maximum and minimum temperatures and the observed hourly streamflow data for the SWAT model's calibration, sourcing this information from

the Malaysian Meteorological Department (MMD) and the Department of Irrigation and Drainage (DID) Malaysia. Their findings demonstrated that the effectively trained Generative Adversarial Network (GAN) enhanced the quality of rainfall data, capturing the spatial and temporal patterns of the original data more effectively than merely duplicating its statistical properties.

# Chapter 3

## Marginal Modelling of Rainfall Events Characteristics

## 3.1 Introduction

Rainfall events are summarised by duration, intensity, maximum intensity and volatility. Finding the appropriate marginal distribution for each rain event characteristic is one of the most critical processes in the process of fitting copulas to rainfall characteristics [76]. This chapter aims to analyze the characteristics of rainfall events in Jackson Creek, Sunbury, Victoria, Australia, using gauge rainfall data. The study seeks to fit appropriate marginal distributions to the data.

## 3.2 Data Pre-processing

**Definition 3.2 (Rainfall Event):** Given continuous rainfall at a point with intensity over time, a rainfall event is defined as the period of rain where no rainless gaps exceed the duration of the inter-event time [77]. The inter-event time (IET) is the minimum specified rainless period that must precede the beginning of a new rainfall event [78].

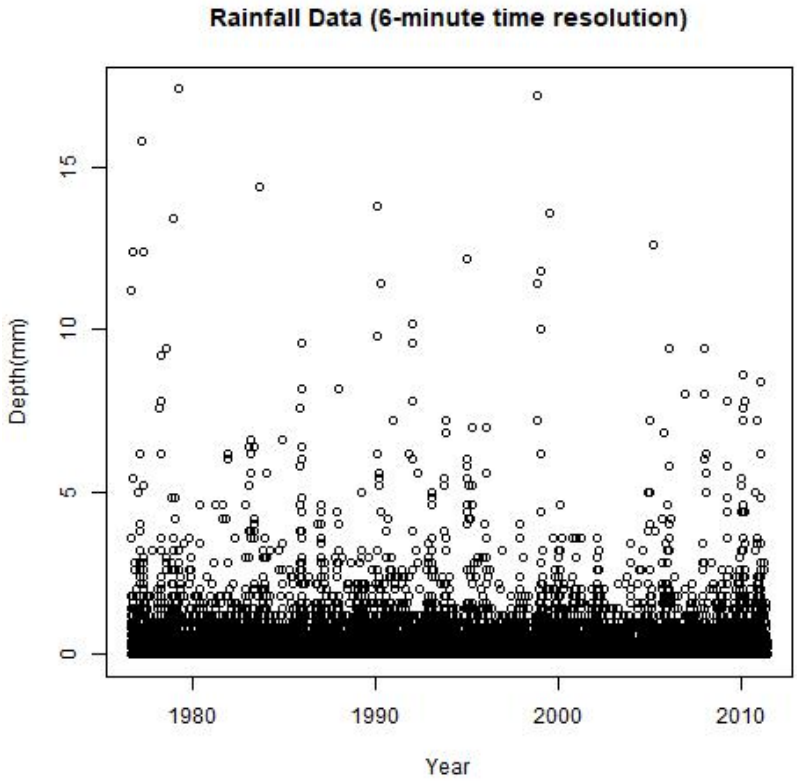


**Figure 3.1:** Rain Data (6-minutes time resolution)

The dataset for this research was obtained from the Meteorological Department of Australia. The rain data is recorded by a tipping bucket rain gauge that has a tip resolution of 6 minutes for the period of 36 years (1976-2011) at Jackson's Creek, Sunbury in Victoria, Australia. To use this data, we carried out data cleaning. First, the data for the last month was not used as they were inconsistent with the previous data and zero readings with odd time stamps were deleted. Victoria, Australia, experiences daylight saving time (DST). To deal with this, the data time was converted to POXISct format using the Lubridate Package in R programming software [79]. Time was also rounded up to the nearest minute and the missing zeros were also filled (that is for the six minutes duration where there was no rainfall, 0mm was recorded). The R code is given in Listing 3.1.

```
#Data Cleaning and Rainfall Extraction
library(lubridate)

rainfall_data <- read.csv("C:/Users/Administrator/Documents/R/
```

```
   SITE_230202_MR_510.csv", header = TRUE, stringsAsFactors =
   FALSE)
all.times <- rainfall_data$measure_date
all.times.POSIX <- dmy_hms(all.times, tz = "Australia/
   Melbourne")
diff.all.times.POSIX <- diff(all.times.POSIX)
diff.all.times.POSIX == 360 #checking if the data had 6mins
   gap
length(diff.all.times.POSIX) #length will be equal to original
    length minus 1
identical(round_date(all.times.POSIX, unit ="seconds"), all.
   times.POSIX)

#View(rainfall_data)
rainfall_data <- rainfall_data[-(111621:111906), ] #deleting
   the last month
rainfall_data <- rainfall_data[-c(107417, 107418, 107445,
   107446), ] #deleting zero readings with odd time stamps
time.POSIX <- dmy_hms(rainfall_data[,3], tz = "Australia/
   Melbourne") #converting to POSIXct format

#Creating 6mins time stamps Method 1
minute.POSIX <- minute(time.POSIX)
w <- which(minute.POSIX %% 6 != 0)
time.POSIX[w]
time.POSIX <- round_date(time.POSIX, unit="minute")
time.POSIX[w]
which(minute(time.POSIX) %% 6 != 0)
all(as.numeric(diff(time.POSIX)) %% 360 == 0)
time.6minutes <- cumsum(c(0, diff(time.POSIX)) / 360) + 1 #
   how many 6 minute intervals have passed since the first ever
    measurement
depth <- rainfall_data[,4]
time <- time.6minutes

# fill in zeros
maxt <- max(time)
newd <- rep(0, maxt)
ti <- 1 #time counter
```

```
idx <- 1 #position in depth vector
newd[1] <- depth[1]
while (ti < maxt) {
  ti <- ti + 1
  if (ti == time[idx+1]) {
    idx <- idx + 1
    newd[ti] <- depth[idx]
  }
}

New_Rainfall_Data <- data.frame(Datetime=time.POSIX[1] + 360*
   (0:(maxt-1)), Time=1:maxt, Depth=newd)

save(New_Rainfall_Data, file = "Clean_Rainfall_Data.RData") #
   data name: New_Rainfall_Data
```

**Listing 3.1:** Data Cleaning and Rainfall Extraction Code

To choose a reasonable inter-event time (IET), we plotted the rainfall distribution for intensity and duration setting different IETs, 30 minutes, 60 minutes, 120 minutes and 360 minutes using the following rainfall depth thresholds: depth $\geq 0mm$, $\geq 0.4mm$, $\geq 0.6mm$ and $\geq 1mm$. Due to the discrete nature of the data, we needed to remove the spike (rainfall depth of $= 0mm$), which were a period without rain, and we were also not interested in small rainfall, thereby deleting small events. The results from the different intensity and duration distributions plots indicated that the best IET to define a rain event using our data is 60 minutes with a rain depth threshold of $1mm$ as it gave us a smooth distribution for intensity and duration. This is consistent with the study by [80]. Figure 3.2 gives an illustration of rainfall events. Here, rainfall event intensity ($I$) is defined as the average amount of rain that falls per unit of the time during the duration of the rain event(mm/mins). The maximum intensity ($M$), often called the peak intensity, is the highest rate of rainfall (depth per unit of time) recorded during a specific time period of a rain event. Volatility (V) captures the variability or fluctuation in these intensities throughout the rainfall event. In this context, the volatility provides insight into the stability or predictability of the rainfall intensity. Specifically, the volatility is

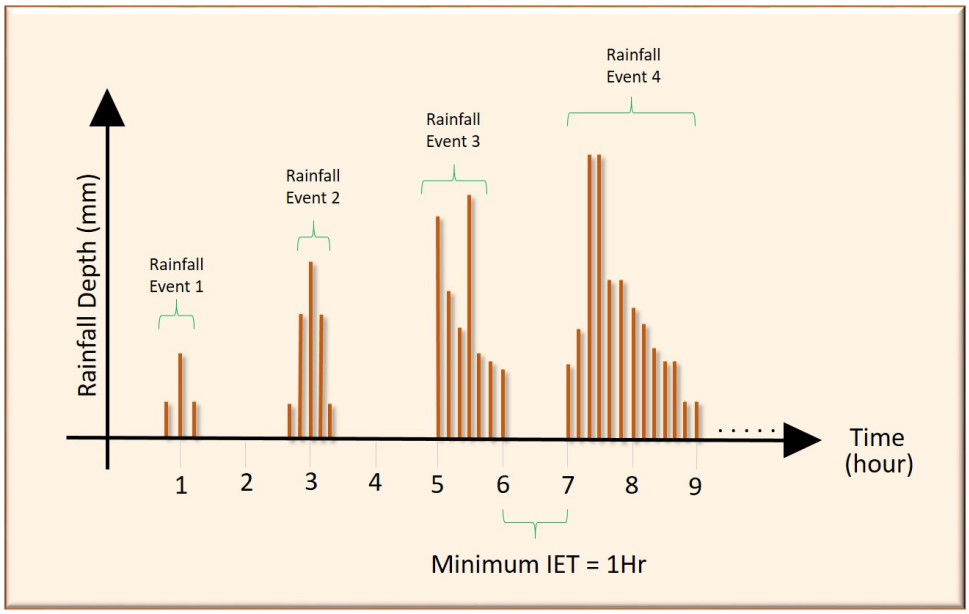


**Figure 3.2:** Illustration of Rainfall Events

defined as:

$$\text{Volatility}(x) = \frac{1}{n}\sum_{i=2}^{n}(x_i - x_{i-1})^2$$

Where $x_i$ denotes the intensity at a specific time point, $x_{i-1}$ is the intensity at the preceding time point, and $n$ is the total number of time points or intensities in the rain event. A higher volatility value indicates larger variations or jumps in intensity, suggesting a more erratic rainfall pattern. Conversely, a lower volatility value implies a steadier and more consistent rainfall intensity over the event's duration ($D$). The summary statistics of the rainfall event duration, intensity, maximum intensity and volatility are given in Table 3.1. See Appendix A.1 for the detailed R code used in data cleaning and rain event extraction.

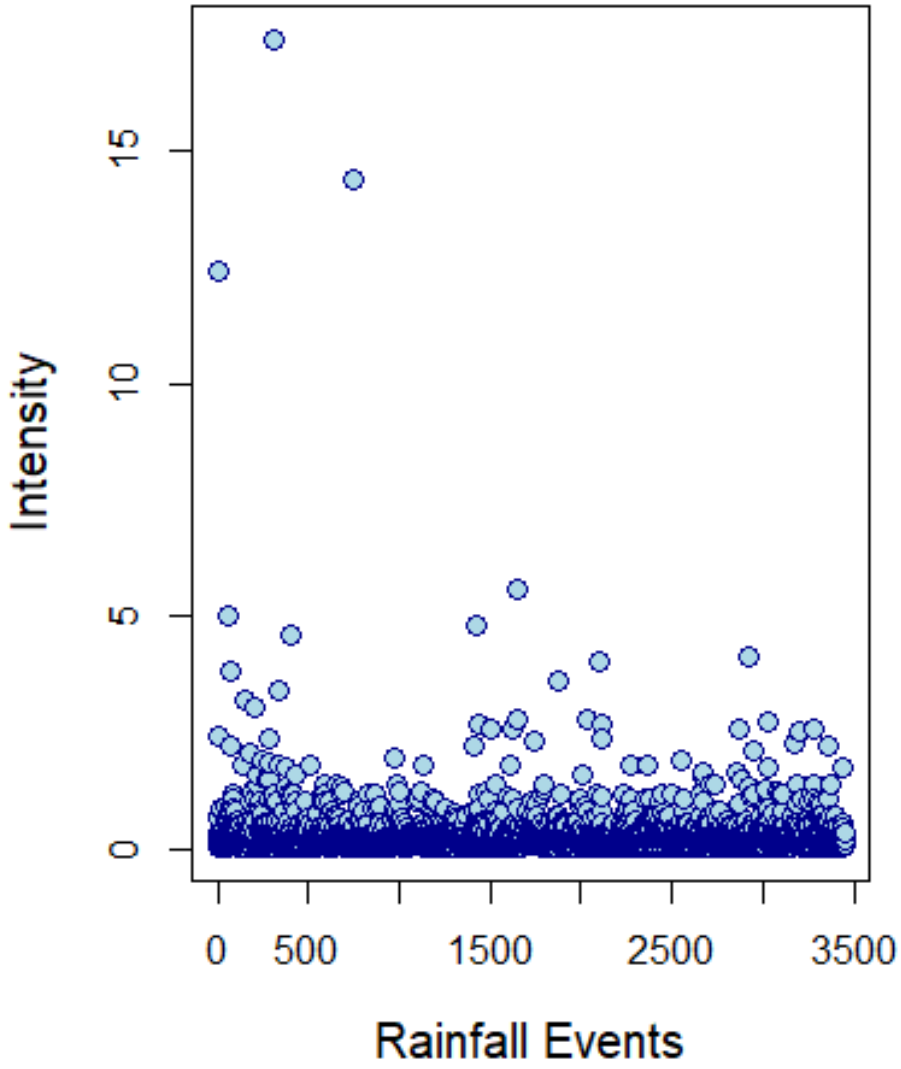


**Figure 3.3:** Rainfall Events Intensity

Figure 3.1 provides a continuous, high-resolution visualization of rainfall intensity over time, detailing how the rainfall depth fluctuates at each recorded time point. This plot is instrumental in revealing the temporal patterns of rainfall, showcasing the duration and intensity fluctuations within and across rainfall events, and allowing observers to discern specific periods of high or low rainfall. Conversely, Figure 3.3, which plots the average depth of each event, abstracts the rainfall data into a summarized form, where each point represents the mean intensity of an individual rainfall event. This comparative plot

simplifies the data, emphasizing the overarching trends and differences in average intensity across events. While Figure 3.1 excels in depicting rainfall's dynamic, time-sensitive nature, offering a granular view of its temporal distribution, Figure 3.3 provides a synthesized, event-centric perspective, focusing on the intensity characteristics of rainfall events and facilitating a straightforward comparison of their average intensities. Figure 3.4 presents the time series for the rainfall event with the largest intensity and adjacent rainfall events. It illustrates the temporal distribution and intensity of the focal event and its neighbouring events. Figure 3.5 depicts the time series of rainfall events surrounding the event characterized by the highest total rainfall, where total rainfall is the product of each event's duration (D) and its average intensity (I). This figure aims to highlight the comparative scale and impact of the event with the maximal aggregate rainfall alongside its immediate temporal neighbours. See Appendix B.1 for the details about event 9556.

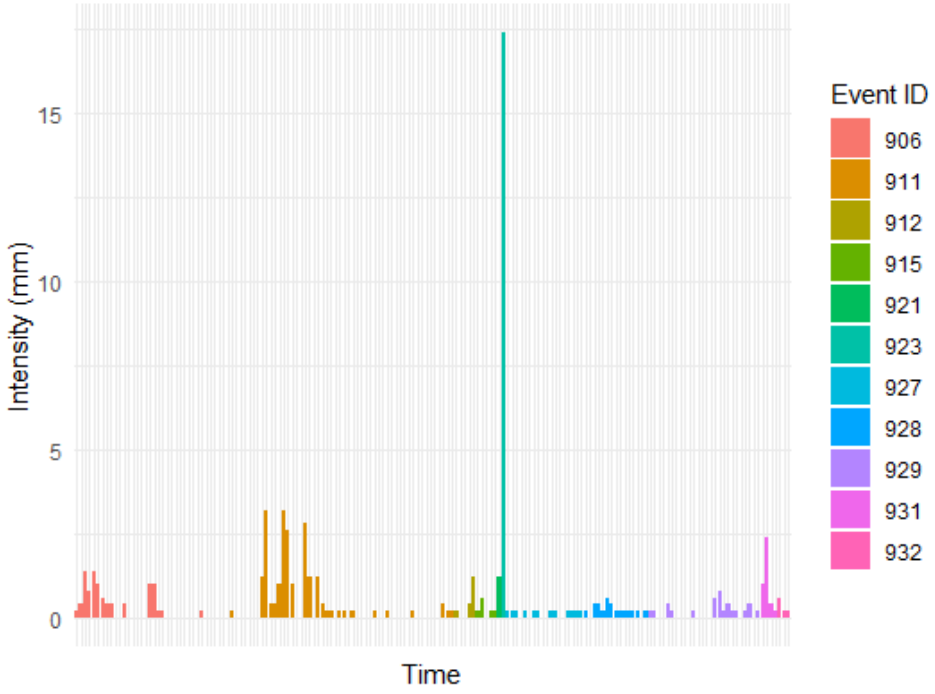


**Figure 3.4:** Rainfall event with largest intensity event (event 923)

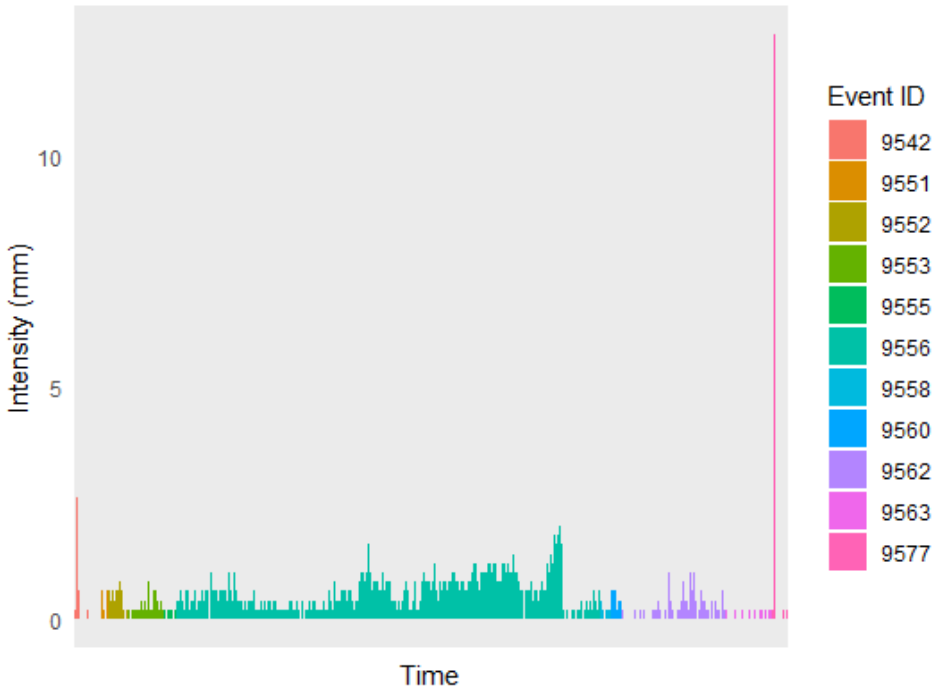


**Figure 3.5:** Rainfall event with the most total rainfall (event 9556)

**Table 3.1:** Summary statistics for rain event duration ($D$), intensity ($I$), maximum intensity ($M$), and volatility ($V$)

| **Rainfall event** | **Mean** | SD | **Min** | **Median** | **Max** | **Skewness** | **Kurtosis** |
|---|---|---|---|---|---|---|---|
| D | 24.34 | 27.29278 | 1.00 | 17.00 | 322.00 | 3.744245 | 24.71962 |
| I | 0.28008 | 0.569851 | 0.03125 | 0.16293 | 17.40000 | 16.70841 | 414.9539 |
| M | 0.9184 | 1.301368 | 0.2 | 0.6 | 17.40 | 5.964268 | 51.66765 |
| V | 0.24350 | 1.307148 | 0 | 0.4 | 28.42182 | 12.40554 | 192.1812 |
| log(D) | 2.754 | 0.9682511 | 0.000 | 2.833 | 5.775 | -0.2823895 | 3.312447 |
| log(I) | -1.714 | 0.8122467 | -3.466 | -1.814 | 2.856 | 0.8648032 | 4.305471 |
| log(M) | -0.4659 | 0.7696502 | -1.6094 | -0.5108 | 2.8565 | 0.8632411 | 4.296694 |
| log(V) | -2.970 | 1.284787 | -9.213 | -3.219 | 3.347 | 1.161323 | 6.731177 |

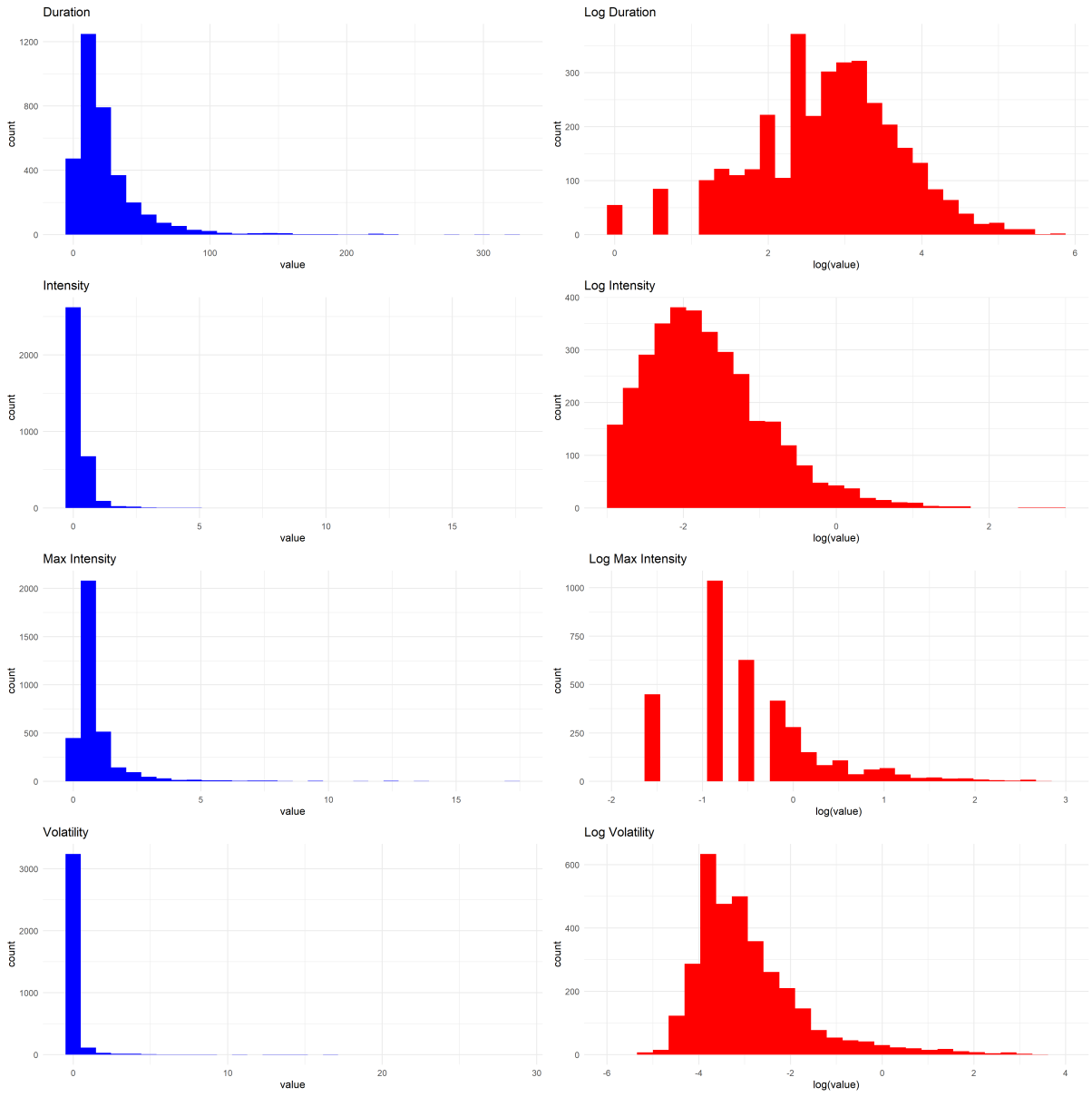


**Figure 3.6:** Histogram of DIMV

## 3.3 Marginal Modelling using Parametric Distributions

Table 3.1 shows that the datasets for $D, I, M, V$ exhibit significant skewness. To achieve a more normalised distribution and potentially enhance the effectiveness of subsequent modelling, we propose the application of a logarithmic transformation. The logarithmic transformation pulled in the tails of DIMV, thereby reducing the positive skewness. This makes the DIMV distribution more symmetric and closer to a normal distribution, as evident in Figure 3.6. Given the nature and characteristics of the data, the candidate distribution functions selected for this study are the Normal, Skew Normal, Skew T, and the Generalized Extreme Value (GEV) distribution. The rationale behind this choice is rooted in their capability to model diverse and skewed datasets, as reflected in their probability density functions provided below.

### 3.3.1 Normal Distribution

The Normal distribution (the Gaussian distribution) is a continuous probability distribution used extensively in statistics and probability theory. The normal distribution's probability density function (PDF) is given by

$$f_{nm}(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2(\frac{x-\mu}{\sigma})^2}. \tag{3.1}$$

Where $\mu$ is the mean (location parameter) and $\sigma$ is the standard deviation (scale parameter).

### 3.3.2 Skewed Normal Distribution

Given a random variable $X$ which has Skew Normal Distribution $X \sim SN(\mu,\sigma,\lambda)$. The pdf is given as

$$f_{sn}(x;\mu,\sigma,\alpha) = \frac{2}{\sigma}\phi(\frac{x-\mu}{\sigma})\Phi(\alpha\frac{x-\mu}{\sigma}), \quad x \in \mathbb{R}. \tag{3.2}$$

Where $\mu \in \mathbb{R}$ is the location parameter, $\sigma \in \mathbb{R}^+$ is the scale parameter and $\alpha \in \mathbb{R}$ is the skewness parameter. The variables $\phi(.)$ and $\Phi(.)$ are the density and cumulative

distribution functions (CDF) of the standard normal distribution, respectively. When $\alpha > 0$ the distribution is right-skewed; when $\alpha < 0$ the distribution is left-skewed; when $\alpha = 0$, the distribution reduces to the standard normal distribution. Compared to the classical normal, the skew-normal offers a more versatile formulation due to introducing a parameter that controls its skewness. [81]

### 3.3.3 Skew t Distribution

The Skew-t distribution, which offers a more flexible formation of the student's t distribution by adding a skewness parameter, whose pdf is given as:

$$f_{st}(x;\mu,\sigma,\alpha,v) = \frac{2}{\sigma} t(\frac{x-\mu}{\sigma};v) T(\alpha \frac{x-\mu}{\sigma} \sqrt{\frac{v+1}{v+Q_x}};v+1), \quad x \in \mathbb{R}. \tag{3.3}$$

The Skew-t distribution is characterized by a location parameter $\mu$, a scale parameter $\sigma$, a shape parameter $v > 0$, and a distinct skewness parameter represented by $\alpha$. Within this framework, $t(\cdot;v)$ and $T(\cdot;v+1)$ denote the density function of the student's t distribution having a degree of freedom $v > 0$ and the cumulative distribution function of the conventional student's t distribution with $v+1$ degrees of freedom, respectively. The expression $Q_x = \sigma^{-2}(x-\mu)^2$ provides a squared deviation measure. It's noteworthy that when $\alpha$ assumes a value of zero, the density in equation (equation 3.3) converges to the standard student's t distribution. As $v$ approaches infinity, this density aligns with the skew-normal distribution [81].

### 3.3.4 Generalised Extreme Value (GEV) Distribution

The foundational theory concerning extreme values in data samples, articulated by Fisher and Tippet [82], paved the way for formulating the extreme value distribution. The GEV class of distributions is regulated by the tail shape parameter, $\xi$. Consequently, the tail index is defined as $\kappa = \xi^{-1}$, which determines the configuration and magnitude of these tails, encompassing three distinct distribution families. Both Jenkinson [83] and von Mises [84] showed that these three distribution forms could be represented using a

unified set of parameters, thus coining the term "generalized extreme value distribution" with the density expressed as:

$$f_{gev}(x;\mu,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}(1+\xi(\frac{x-\mu}{\sigma}))^{1-1/\xi}\exp(-1(1+\xi(\frac{x-\mu}{\sigma}))^{-1/\xi}), & \text{if } \xi \neq 0 \\ \frac{1}{\sigma}e^{-(x-\mu)/\sigma}\exp(e^{-(x-\mu)/\sigma}), & \text{if } \xi = 0 \end{cases} \tag{3.4}$$

In this representation, $\mu$, $\sigma$, and $\xi$ symbolize the location, scale, and shape parameters, respectively. Depending on the value of $\xi$, being less than 0, 0, or greater than 0, the GEV distribution belongs to the Weibull, Gumbel, and Fréchet class respectively [85]."

## 3.4 Fitting Methodology

The candidate distribution functions' density functions were fitted using the maximum likelihood (MLE) technique, which seeks to optimize the log-likelihood function. To identify the most suitable fit for the rain event's duration, intensity, maximum intensity, and volatility, we employed the Akaike information criterion (AIC).

### 3.4.1 Akaike information criterion (AIC)

The AIC serves as a tool for model selection based on the log-likelihood of a given distribution. Typically, the one with the smallest AIC value is considered to provide the best representation of the dataset among competing models. The formula to compute AIC is given as [86]:

$$AIC = -2L + 2k \tag{3.5}$$

In this equation, $L$ represents the log-likelihood of the model, while $k$ indicates the total number of model parameters.

### 3.4.2 Fitting Results

The theoretical distributions discussed in (3.3.1)–(3.3.4) were used to fit the rainfall event duration, intensity, maximum intensity and volatility data. The Nelder-Mead method was

used to obtain the maximum likelihood estimate for the data; the implementation was done with the optim function in R programming software.

**Table 3.2:** MLE fit for log(duration) using the candidates distributions

| **Distributions** | **Normal** | **Skew Normal** | **Skew t** | **GEV** |
|---|---|---|---|---|
| Parameter estimates | $\hat{\mu} = 2.7537070$<br>$\hat{\sigma} = 0.9681107$ | $\hat{\mu} = 2.74537363$<br>$\hat{\sigma} = 0.96809641$<br>$\hat{\alpha} = 0.01046884$ | $\hat{\mu} = 3.464857$<br>$\hat{\sigma} = 1.135427$<br>$\hat{\alpha} = -1.145593$<br>$\hat{v} = 18.412566$ | $\hat{\mu} = 2.411833$<br>$\hat{\sigma} = 1.001275$<br>$\hat{\xi} = -0.290627$ |
| Log-Likelihood | -4783.528 | -4783.528 | **-4755.523** | -4809.622 |
| AIC | 9571.055 | 9573.056 | **9519.046** | 9625.244 |

**Table 3.3:** MLE fit for log(intensity) using the candidates distributions

| **Distributions** | **Normal** | **Skew Normal** | **Skew t** | **GEV** |
|---|---|---|---|---|
| Parameter estimates | $\hat{\mu} = -1.714289$<br>$\hat{\sigma} = 0.812129$ | $\hat{\mu} = -2.692515$<br>$\hat{\sigma} = 1.271552$<br>$\hat{\alpha} = 3.660099$ | $\hat{\mu} = -2.672764$<br>$\hat{\sigma} = 1.226678$<br>$\hat{\alpha} = 3.437406$<br>$\hat{v} = 44.237506$ | $\hat{\mu} = -2.073450$<br>$\hat{\sigma} = 0.670271$<br>$\hat{\xi} = -0.044155$ |
| Log-Likelihood | -4177.406 | -3982.499 | -3981.098 | **-3974.327** |
| AIC | 8358.813 | 7970.997 | 7970.195 | **7954.655** |

The goal here is to identify which of the candidate distributions is the best fit to model duration, intensity, maximum intensity and volatility. For log(duration), the skew t distribution provided the best-fit probability distribution based on the AIC criterion (see Table 3.2), and this is further supported by the density plot in Figure 3.7 and the Q-Q plot in Figure 1. Table 3.3 shows that the GEV distribution has the least AIC value, which indicates that the GEV distribution demonstrates superiority over the normal, skew normal and skew t distribution for log(intensity) (see Figure 2).

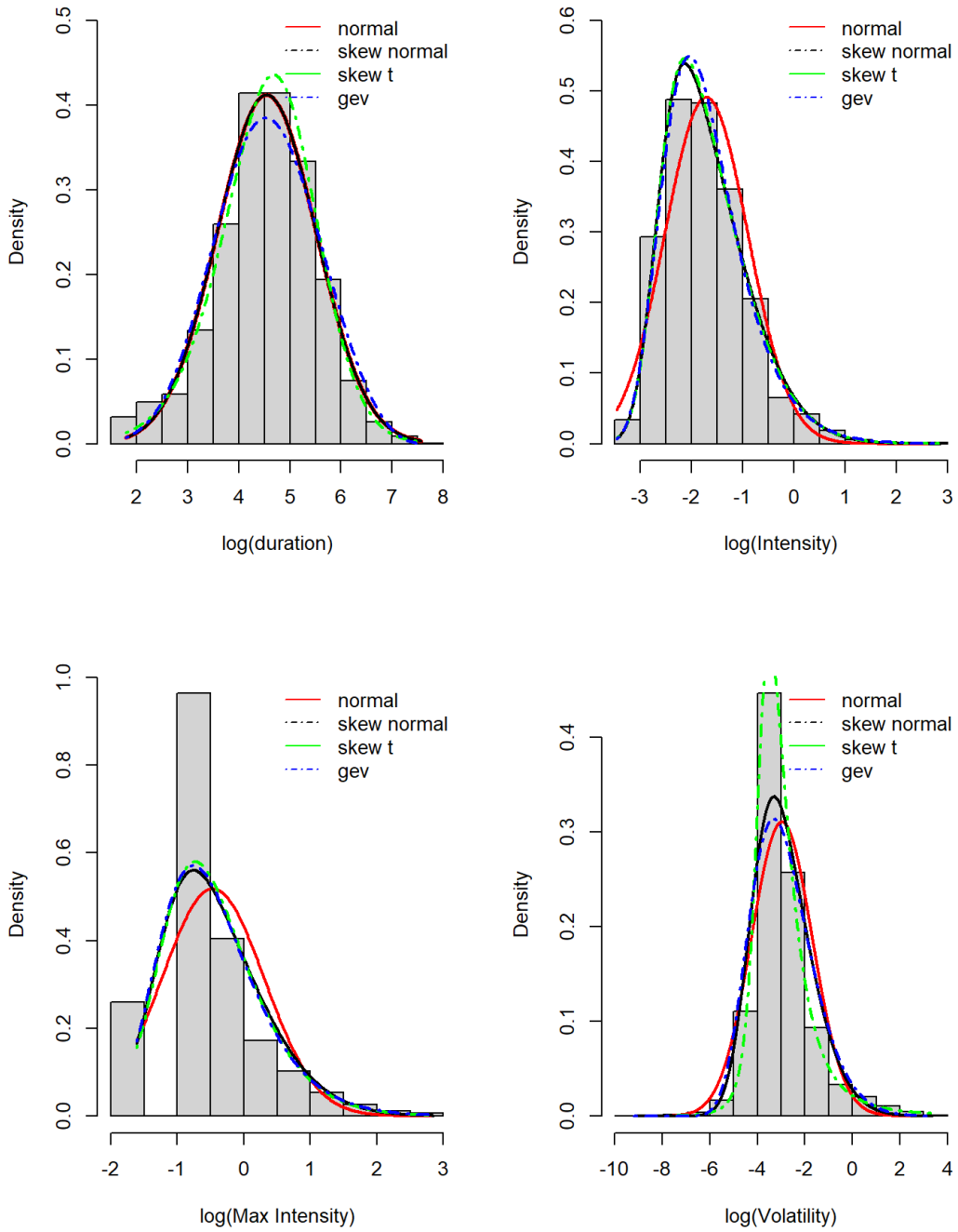


**Figure 3.7:** Density plot for DIMV with fitted distributions

The presence of zero values in the volatility data poses a challenge since the logarithm of zero is undefined. To address this issue, we adopted a specific approach. We identified the smallest non-zero value in the volatility data, denoted as d. This value acts as an upper threshold. Subsequently, all zero values in the dataset were located. Each zero entry

**Table 3.4:** MLE fit for log(maximum intensity) using the candidates distributions

| **Distributions** | **Normal** | **Skew Normal** | **Skew t** | **GEV** |
|---|---|---|---|---|
| Parameter estimates | $\hat{\mu} = -0.46593$<br>$\hat{\sigma} = 0.769539$ | $\hat{\mu} = -1.318797$<br>$\hat{\sigma} = 1.148741$<br>$\hat{\alpha} = 2.728022$ | $\hat{\mu} = -1.251167$<br>$\hat{\sigma} = 1.031159$<br>$\hat{\alpha} = 2.243361$<br>$\hat{v} = 17.856896$ | $\hat{\mu} = -0.800753$<br>$\hat{\sigma} = 0.645630$<br>$\hat{\xi} = -0.060814$ |
| Log-Likelihood | -3991.562 | -3826.192 | -3820.963 | **-3805.216** |
| AIC | 7987.124 | 7658.383 | 7649.926 | **7616.432** |

**Table 3.5:** MLE fit for log(volatility) using the candidates distributions

| **Distributions** | **Normal** | **Skew Normal** | **Skew t** | **GEV** |
|---|---|---|---|---|
| Parameter estimates | $\hat{\mu} = -2.970285$<br>$\hat{\sigma} = 1.284601$ | $\hat{\mu} = -4.245994$<br>$\hat{\sigma} = 1.810555$<br>$\hat{\alpha} = 2.219323$ | $\hat{\mu} = -3.940185$<br>$\hat{\sigma} = 1.015555$<br>$\hat{\alpha} = 1.807691$<br>$\hat{v} = 2.694200$ | $\hat{\mu} = -3.47557$<br>$\hat{\sigma} = 1.18224$<br>$\hat{\xi} = -0.12601$ |
| Log-Likelihood | -5759.384 | -5562.009 | **-5212.376** | -5657.278 |
| AIC | 11522.77 | 11130.02 | **10432.75** | 11320.56 |

is then replaced with a uniformly distributed random number that lies between 0 and the previously determined d. This process ensures that the zero values are substituted with extremely small random numbers, making them suitable for a log transformation. This method makes the transformation feasible and retains the relative scale of the data, minimising any artificial biases that might distort the analysis. Table 3.4 and table 3.5 indicate that log(maximum) and log(volatility) are best fitted by the GEV and skew t distribution, respectively. Identifying these distributions for DIMV sets the foundation for univariate and joint modelling of DIMV needed to model the rainfall characteristics accurately.

# Chapter 4

## Modelling of Extreme Rainfall Events

## 4.1 Introduction

Investigating extreme rainfall is globally crucial, as its severe impacts span from causing ecological havoc, and demolishing infrastructure, to claiming human lives. Thus, numerous facets of human undertakings - from life insurance and civil protection to town planning, regional planning, and civil infrastructure design - derive substantial benefits from this research [87] and extreme value theory (EVT) gives the mathematical background for the modelling of tails of distributions that can be used in extrapolating extreme events beyond the observed data [88], and this technique has been applied to many problems in hydrology [89], in insurance [90], in finance [91], in environment and telecommunications [92], in climate change [93] and in public health [94]. In practice, EVT is a mathematically motivated approach used to predict the likelihood of extreme events that are yet to be observed and characterize their risks to build infrastructures to withstand their impacts, such as extreme events from wind and rainfall. Statistical modelling of extreme rainfall events is crucial to designing and managing hydrological structures such as dams, drainages, and reservoirs [95].

The primary methods are block maxima and peak over threshold (POT). The peak-over-threshold approach utilizes all extreme events that surpass a high threshold, whereas the block maxima method focuses only on the highest value in a block of ordinary observations. Modelling excesses over a specific threshold will be more efficient when dealing with large high-frequency data such as hourly or daily observations [96, 97]. POT modelling has been used in rainfall analysis for studying the distribution of annual and seasonal daily rainfall extremes [95], for modelling extreme flood [98], and for evaluating landslide

episodes [99]. The aim of this chapter is to study extreme rainfall in Sunbury, Victoria, Australia, using the POT methodology with GPD.

## 4.2 Peak Over Threshold (POT) Method

Consider a sequence $X_1, X_2, ...$ where each element represents an independent random variable with the same distribution function, denoted by $F$ and let

$$M_n = \max(X_1, ..., X_n)$$

Let $X$ be an arbitrary term in the sequence $X_i$. Suppose that for large $n$, $M_n$ satisfies the relation [96]:

$$Pr(M_n \leq x) \approx G(x)$$

where

$$G(x) = \exp\left\{ -\left[1+\gamma\left(\frac{x-\mu}{\beta}\right)\right]^{-1/\gamma}\right\} \tag{4.1}$$

for some $\mu, \beta > 0$ and $\gamma$. Then, for a sufficiently large u, the conditional probability distribution of $(X-u)$, given that $X$ exceeds $u$, can be approximated as follows [100]:

$$G_{u,\beta,\gamma}(x) = \begin{cases} 1-\left(1+\gamma\left(\frac{x-u}{\beta_u}\right)\right)^{-1/\gamma}, & \text{if } 1+\gamma\frac{x-u}{\beta_u} > 0, \quad \text{and } \gamma \neq 0, \\ 1-\exp\left(-\frac{x-u}{\beta_u}\right), & \text{if } x-u > 0, \quad \text{and } \gamma = 0. \end{cases} \tag{4.2}$$

The distribution given by equation 4.2 is called generalized Pareto distribution (GPD), where $u$ is the selected threshold, $\beta_u = \beta + \gamma(u-\mu) > 0$ is the scale parameter, and $\gamma \in \mathbf{R}$ is the shape parameter.

- When $\gamma < 0$, the maximum value for the distribution of excesses is determined by the expression $u - \beta_u/\gamma$.
- when $\gamma > 0$, the distribution does not have an upper limit.
- when $\gamma = 0$, the distribution has no upper limits and matches an exponential distribution with an average value of $\beta_u$.

## 4.3 Threshold Selection

The process of selecting a threshold is a crucial aspect in the modelling of extreme occurrences. In the POT method, a very high threshold will generate a few extreme values, while a very low threshold will invalidate the model's assumptions. Two primary guiding tools for threshold selection are (i) Mean residual life plot and (ii) parameter stability estimates.

### 4.3.1 Mean Residual Life Plot

This is a graphical approach performed prior to the parameter estimation. Consider the sequence $X_1, ..., X_n$ i.d.d with the same distribution as $X$ where the excesses above a threshold $u_0$ are following a GPD, then

$$E(X - u_0 | X > u_0) = \frac{\beta_{u_0}}{1 - \gamma}. \tag{4.3}$$

If the GPD is appropriate for excesses above $u_0$, then it should also be a valid model for excesses above $u > u_0$, and

$$E(X - u | X > u) = \frac{\beta_u}{1 - \gamma} = \frac{\beta_{u_0} + \gamma(u - u_0)}{1 - \gamma}. \tag{4.4}$$

That is, for $u > u_0$, the mean excess $E(x - u | x > u)$ is a linear function of $u$. In practice, to select an appropriate threshold, the threshold $u$ is plotted against the empirical mean excesses

$$\frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u); \quad u < x_{\max} \tag{4.5}$$

where $x_{(1)}, ..., x_{(n_u)}$ are the ordered $n_u$ observations exceeding $u$. This plot is known as the mean residual plot or mean excess plot. The threshold $u_0$ above which the plot starts to be approximately linear in $u$ indicates an appropriate threshold at which the GPD gives a valid approximation for the distribution of the excesses [96].

### 4.3.2 Parameter Stability Plot

This method examines the stability of the model's parameter estimates at a range of thresholds. Suppose the GPD gives a valid model for excesses over a threshold $u_0$, for a threshold of $u > u_0$, the excesses should also follow a GPD with the same shape parameter but with shifted scale parameter

$$\beta_u = \beta_{u_0} + \gamma(u - u_0) \tag{4.6}$$

Hence the transformed scale parameter $\beta^* = \beta_u - \gamma u$ remains constant with changes in $u > u_0$. Therefore, $\beta^*$ and $\gamma$ should remain almost constant above $u_0$, if $u_0$ gives an appropriate threshold for excesses to follow the GPD [101]

## 4.4 Parameter Estimation

Given a threshold $u$, the GPD parameters can be estimated using the maximum likelihood (MLE) method. If $y_1, ..., y_k$ are the k threshold exceedances, then, taking the log-likelihood of equation 4.2, we have

for $\gamma \neq 0$

$$\ell(\theta; \beta, \gamma) = -k \log(\beta) - (1 + 1/\gamma) \sum_{i=1}^{k} \log(1 + \gamma y_i/\beta) \tag{4.7}$$

given $(1 + \gamma y_i/\beta) > 0$ for $i = 1, ..., k$; otherwise $\ell(\beta, \gamma) = -\infty$.

For $\gamma = 0$

$$\ell(\theta; \beta) = -k \log \beta - 1/\beta \sum_{i=1}^{k} y_i. \tag{4.8}$$

For the numerical maximization of equation 4.7 and equation 4.8, we can use, for example, the Broyden-Fletcher-Goldfarb-Shannon (BFGS)

## 4.5 Return Levels

The return level plot helps to answer the question of how likely an extreme event will reoccur in the next 20, 40 or 100 years. If a GPD is a valid model for excesses above a threshold $u$ by a variable $Y$. Then, for $y > u$

$$P\{Y > y|Y > u\} = \left[1+\gamma\frac{y-u}{\beta}\right]^{-1/\gamma}. \tag{4.9}$$

Therefore,

$$P\{Y > y\} = \lambda_u\left[1+\gamma\frac{y-u}{\beta}\right]^{-1/\gamma}. \tag{4.10}$$

where $\lambda_u = P(Y > u)$. Then, the level $y_t$ exceeded once for every t observation on average is the answer to

$$\lambda_u\left[1+\gamma\frac{y_t-u}{\beta}\right]^{-1/\gamma} = \frac{1}{t}. \tag{4.11}$$

To obtain

$$y_t = \begin{cases} u+\frac{\beta}{\gamma}[(t\lambda_u)^\gamma - 1], & \text{for} \quad \gamma \neq 0 \\ u+\beta\log(m\lambda_u), & \text{for} \quad \gamma = 0 \end{cases} \tag{4.12}$$

As long as $t$ is large enough to guarantee that $y_t$ exceeds $u$. $y_t$ is the $t$ observation return level. When there are $n_x$ instances noted every year, we can calculate the return level for the t-th observation using $t = M \times n_x$. Following this, the return level for a duration of M years is defined as follows [96]:

$$z_M = u+\frac{\beta}{\gamma}\left[(Mn_x\lambda_u)^\gamma - 1\right], \quad \text{for} \quad \gamma \neq 0. \tag{4.13}$$

$$z_M = u+\beta\log(Mn_x\lambda_u), \quad \text{for} \quad \gamma = 0. \tag{4.14}$$

## 4.6 Results of Univariate Analysis

This section applies the POT method described above to the rain event duration and intensity data. Figure 4.1 shows the rainfall event for the duration, intensity and total

rainfall with the selected threshold method described in Section 4.3. Figure 4.3 displays the plots of the scale and shape parameters over 80 equally separated thresholds from 50 to 180 for the duration data. Additionally, Figure 4.2 gives the mean residual life plot for the event duration data. We used the *extRemes* library in R for the univariate implementation.

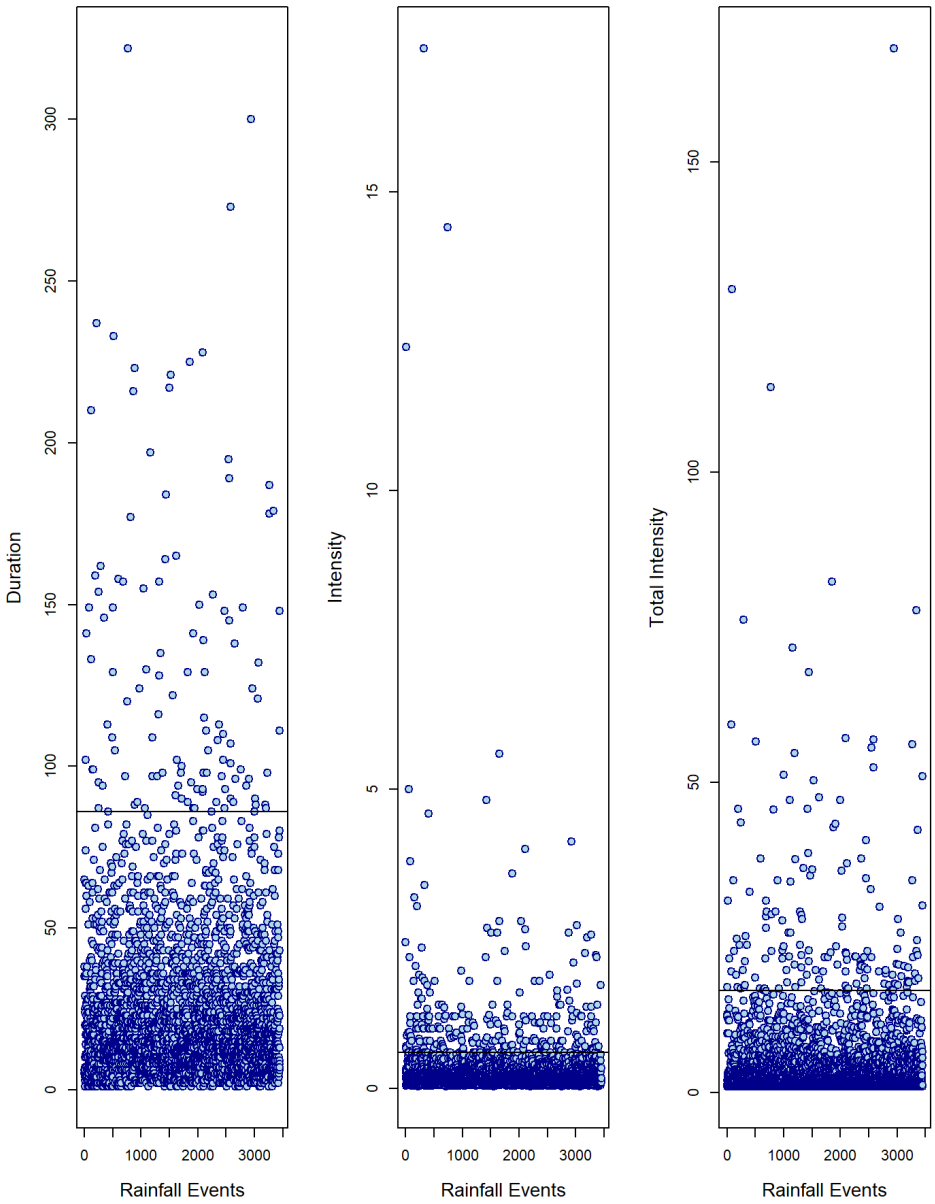


**Figure 4.1:** Rain event duration, intensity and total rainfall with 86, 0.6, 16.5 as selected thresholds for duration, intensity and total intensity respectively

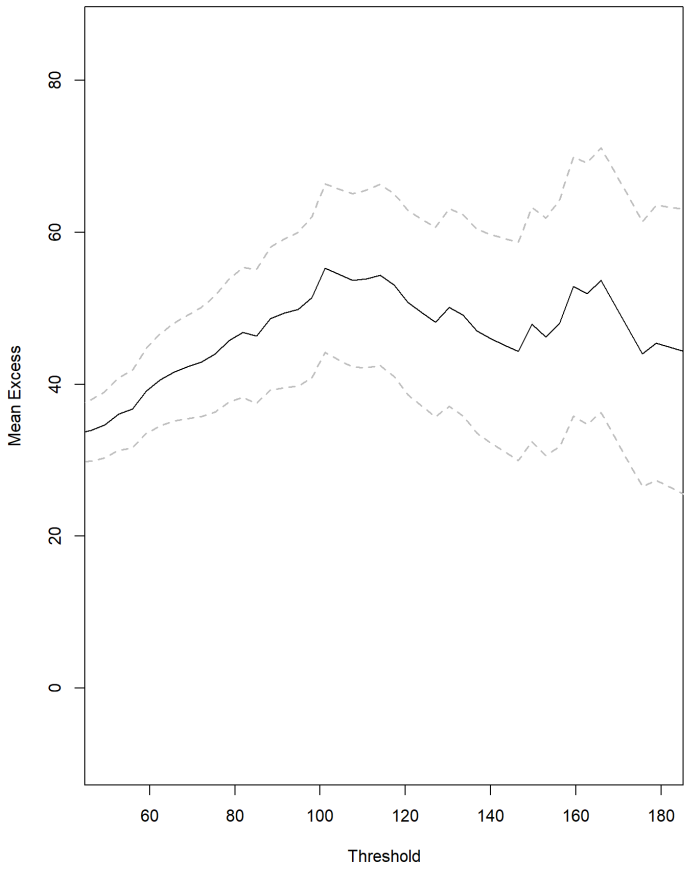


**Figure 4.2:** Rainfall duration mean residual life plot

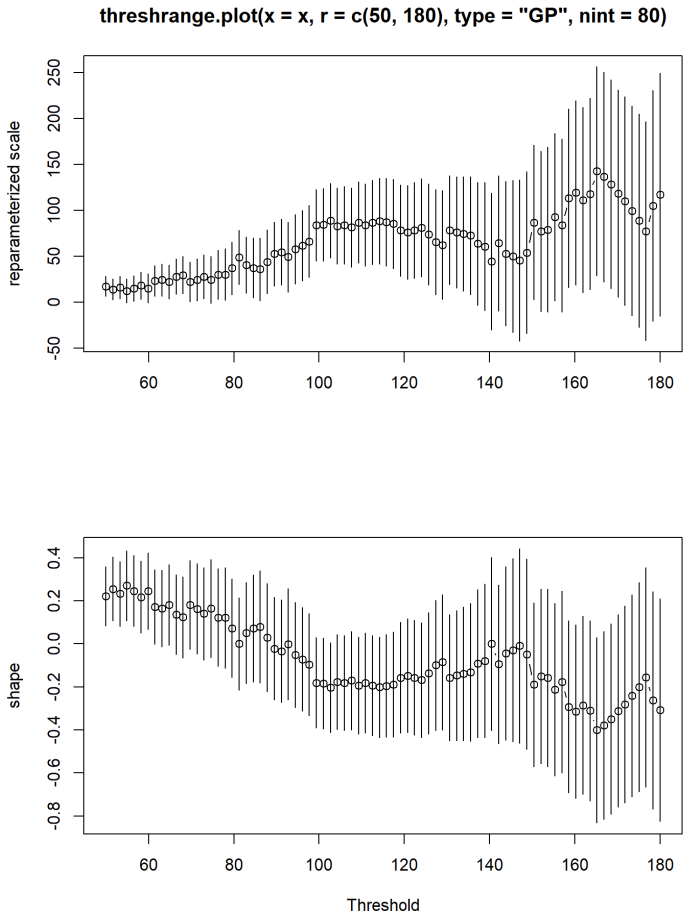


**Figure 4.3:** Parameter estimates against the threshold for rainfall duration

**Table 4.1:** Return Levels with 95% Confidence Intervals of Duration

| Return Level | 95% Lower CI | Estimate | 95% Upper CI |
|---|---|---|---|
| 2-year | 153.3702 | 171.5120 | 189.6537 |
| 20-year | 217.7998 | 295.1206 | 372.4415 |
| 30-year | 221.3957 | 319.0663 | 416.7370 |
| 50-year | 222.2242 | 350.2363 | 478.2485 |
| 70-year | 220.4085 | 371.3962 | 522.3840 |
| 100-year | 216.3361 | 394.3877 | 572.4393 |

**Table 4.2:** Parameter estimates for rain event duration data

| Parameters | Scale | Shape | Log-likelihood value | Threshold |
|---|---|---|---|---|
| Estimates | 43.30305229 | 0.07127485 | -551.7045 | 86 |
| SE | 6.9633174 | 0.1308881 | | |

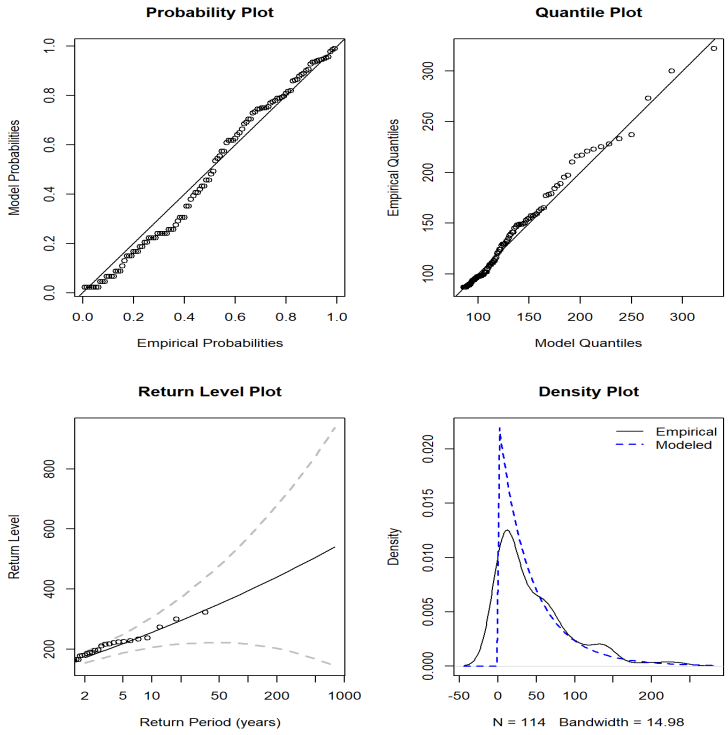


**Figure 4.4:** Diagnostic plots for rainfall events duration

Table 4.2 presents the maximum likelihood parameter estimates for rain event duration data using the POT approach with the GPD. The threshold value used in the analysis is 86. Since we are interested in the annual event return level and the rainfall event data

for duration and intensity is event-based (and not at regular time intervals), we needed to take into account the rainfall event data points per year, and a good solution is to estimate the average number of events per year. The data shows 3450 rainfall events for 36 years, averaging 96 yearly events. Figure 4.4 gives some diagnostic plots for the POT model fitted to the rain event duration data. The density plot is shown in the second column. The Q-Q plot for the fit of GPD to the duration data shows that the GPD model adequately fits the rain event duration data, and the fit is good even at the upper tail. The second-row first column shows the return level plot. From Table 4.1, which shows the return levels with 95% confidence intervals for rainfall event durations ranging from 2 years to 100 years, we see that for a 100-year return period, the estimated return level is 394.3877, with a lower confidence interval of 216.3361 and an upper confidence interval of 572.4393.

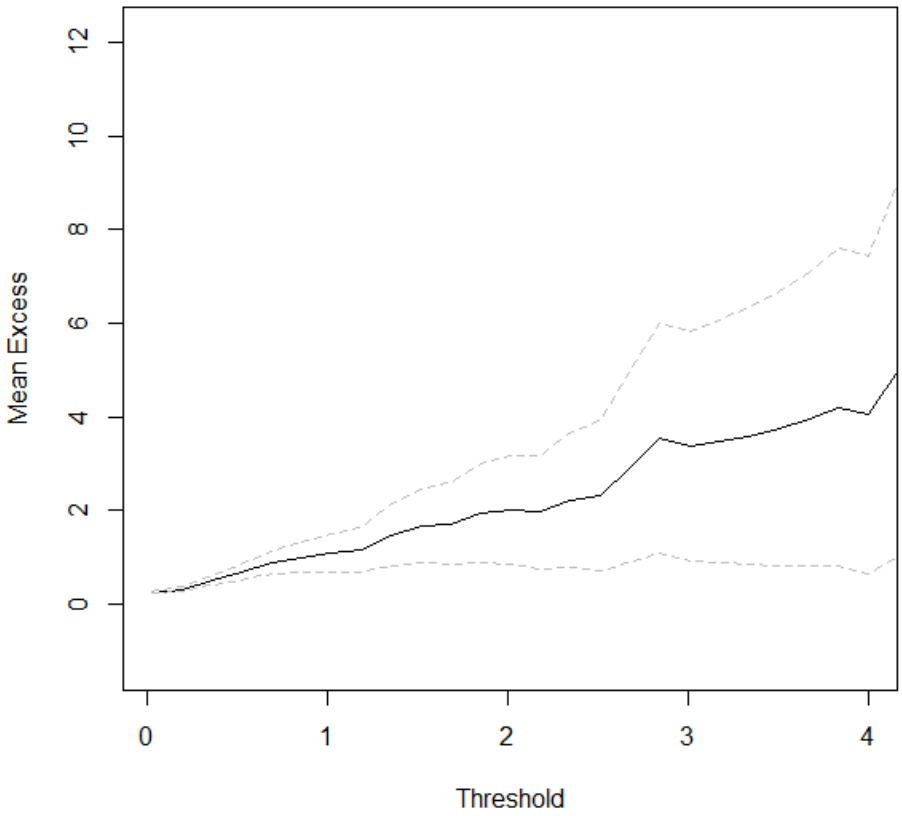


**Figure 4.5:** Rainfall intensity mean residual life plot

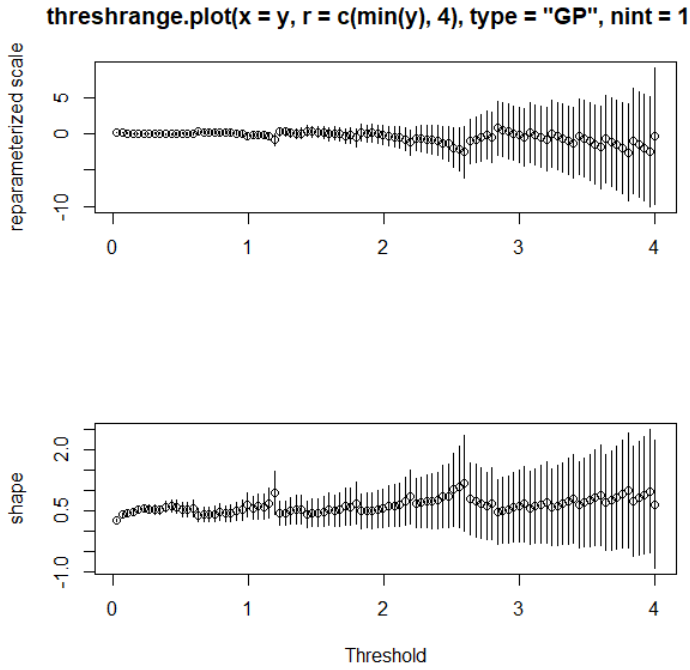


**Figure 4.6:** Parameter estimates against the threshold for rainfall intensity

**Table 4.3:** Return Levels with 95% Confidence Intervals for Intensity

| Return Level | 95% Lower CI | Estimate | 95% Upper CI |
|---|---|---|---|
| 2-year | 2.420946 | 2.890559 | 3.360171 |
| 20-year | 4.794556 | 7.610885 | 10.427214 |
| 30-year | 5.207918 | 8.902540 | 12.597162 |
| 50-year | 5.685060 | 10.808672 | 15.932284 |
| 70-year | 5.956408 | 12.260239 | 18.564070 |
| 100-year | 6.189639 | 13.994018 | 21.798398 |

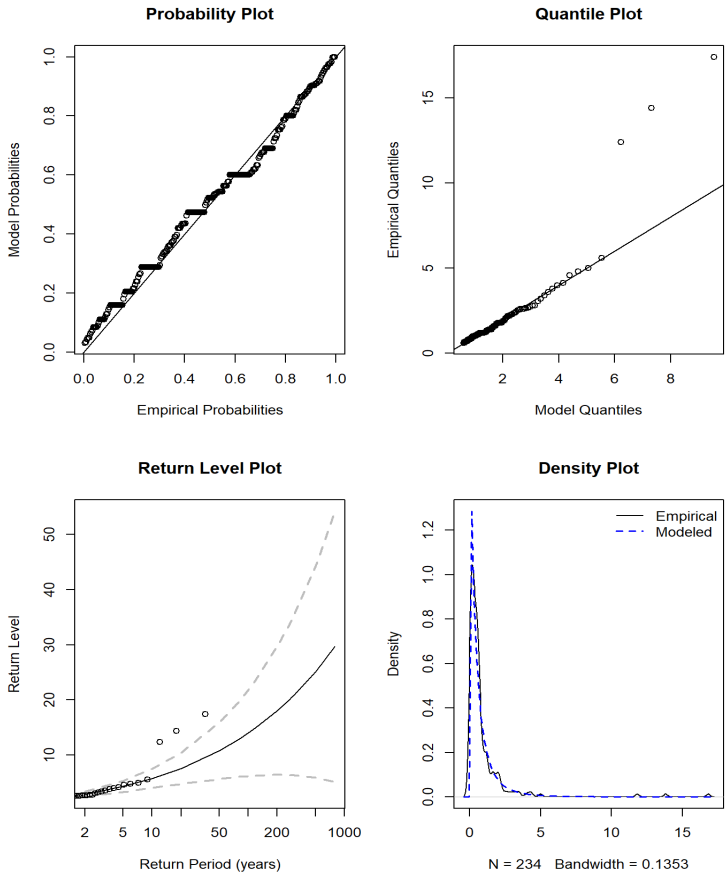


**Figure 4.7:** Diagnostic plots for rainfall events intensity

**Table 4.4:** Parameter estimates for rain event intensity data

| Parameters | Scale | Shape | Log-likelihood value | Threshold |
|---|---|---|---|---|
| Estimates | 0.5555366 | 0.3441744 | -176.9867 | 0.6 |
| SE | 0.05487924 | 0.07740442 | | |

Analysing rain event intensity data through the Peaks Over Threshold (POT) model provides critical insights and challenging standard modelling approaches. The estimated shape parameter, represented as $\gamma$, exhibits a positive value indicative of a heavy-tailed distribution. This suggests a likelihood of extreme rainfall events within the dataset, as shown in Table 4.4. Although the density plot initially indicated a good fit, closer examination of the quantile plot, as shown in Figure 4.7, revealed significant limitations. The model struggles to account for three distinct extreme outliers accurately, highlighting a notable gap in its predictive capability. This issue is particularly evident in the return

level plot, an essential tool for assessing the extremity and frequency of significant rainfall events. The extreme empirical observations are expected to be encapsulated within the model's confidence intervals, yet these outliers are conspicuously positioned outside, revealing a pronounced misfit.

Interestingly, the identified outliers correspond to brief events, encapsulating single-time-point (short-duration event) occurrences rather than sustained periods of heavy rainfall. For instance, the most extreme intensity observed—around 17.4mm within just six minutes—though notably high represents an isolated incident rather than a prolonged heavy downpour. See Appendix B.2 for the details of these events. This specificity challenges the model's relevance for predicting flood-inducing events, as the sheer intensity doesn't translate to substantial accumulated rainfall. This scenario accentuates the necessity of pivoting our analytical focus towards total rainfall, which integrates both duration and intensity. Such a shift is pivotal for a more comprehensive understanding and forecasting of extreme rainfall events, advocating for a nuanced approach that transcends mere intensity modelling. For a holistic extreme value analysis, the combined metric of total rainfall (duration multiplied by intensity) might offer a more reliable predictor, potentially sidelining the anomalies presented by short-duration extremes.

The current results highlight the necessity for refining our modelling approach or adopting more sophisticated models capable of effectively representing the extreme upper tail of the rainfall intensity distribution. Such improvements are crucial to increase the model's predictive precision and ensure it effectively captures the essential characteristics of extreme rainfall events.

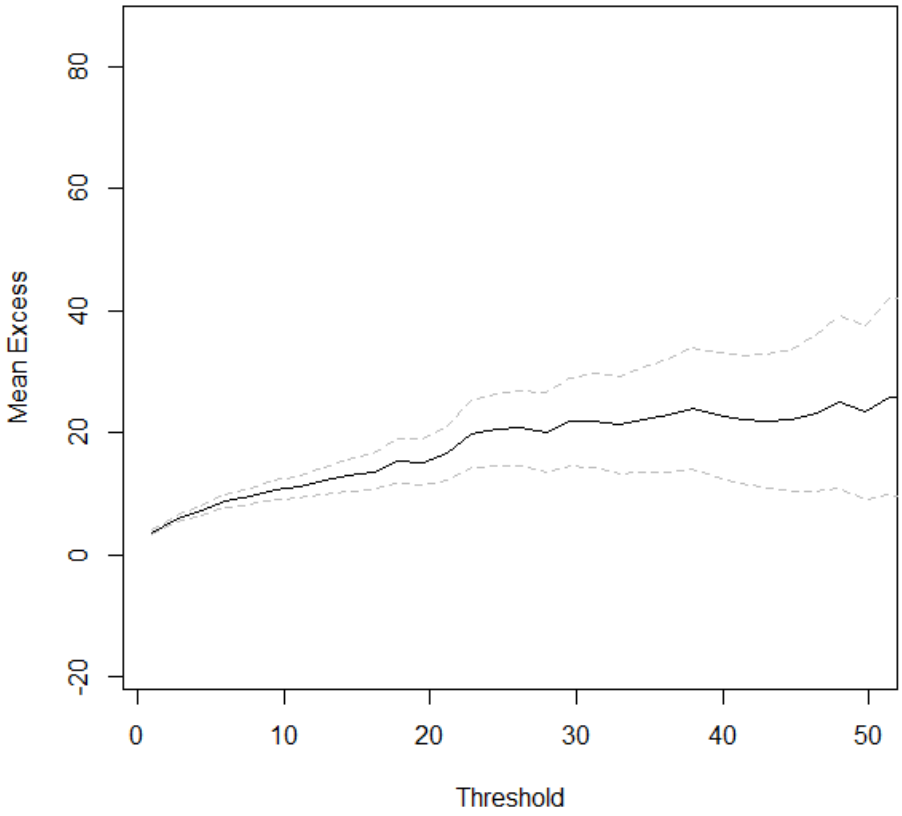


**Figure 4.8:** Total rainfall mean residual life plot

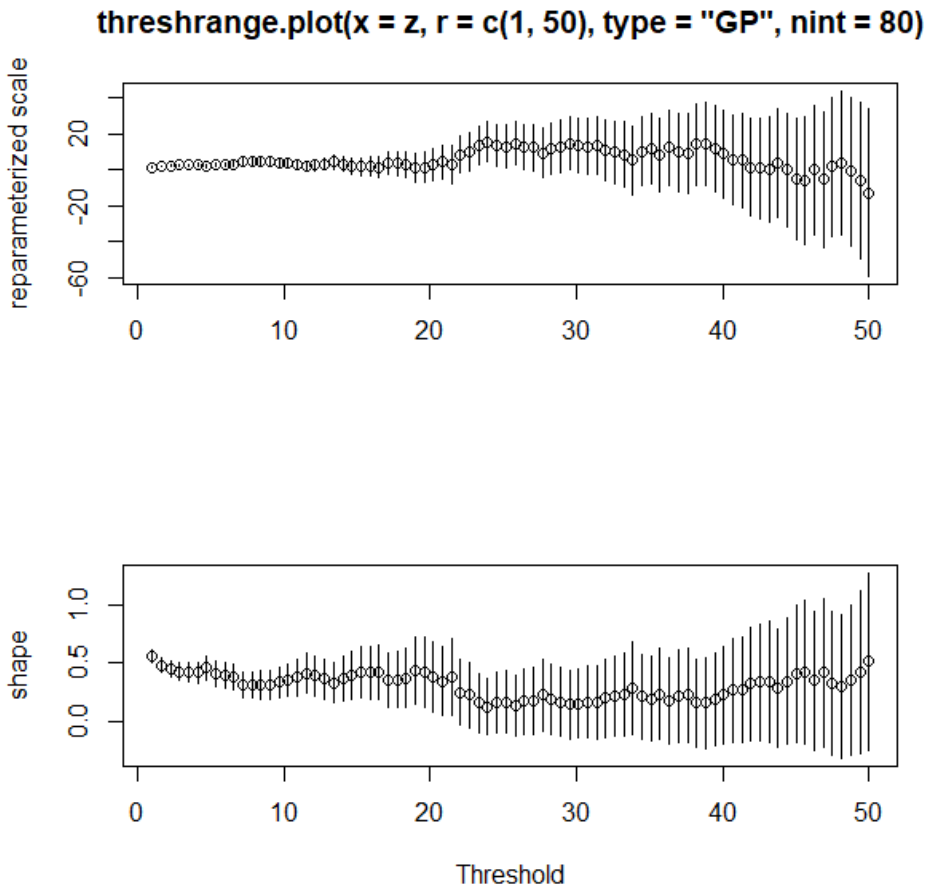


**Figure 4.9:** Parameter estimates against the threshold for total rainfall

**Table 4.5:** Parameter estimates for rain event total rainfall data

| Parameters | Scale | Shape | Log-likelihood value | Threshold |
|---|---|---|---|---|
| Estimates | 8.4486 | 0.4220 | -576.0708 | 16.5 |
| SE | 1.1966 | 0.1224 | | |

**Table 4.6:** Return Levels with 95% Confidence Intervals for Total Rainfall

| Return Level | 95% Lower CI | Estimate | 95% Upper CI |
|---|---|---|---|
| 2-year | 39.62265 | 47.11816 | 54.61366 |
| 20-year | 66.34175 | 130.28748 | 194.23321 |
| 30-year | 67.26749 | 155.25832 | 243.24915 |
| 50-year | 64.45847 | 193.45527 | 322.45208 |
| 70-year | 59.23890 | 223.50723 | 387.77555 |
| 100-year | 49.77344 | 260.38449 | 470.99553 |

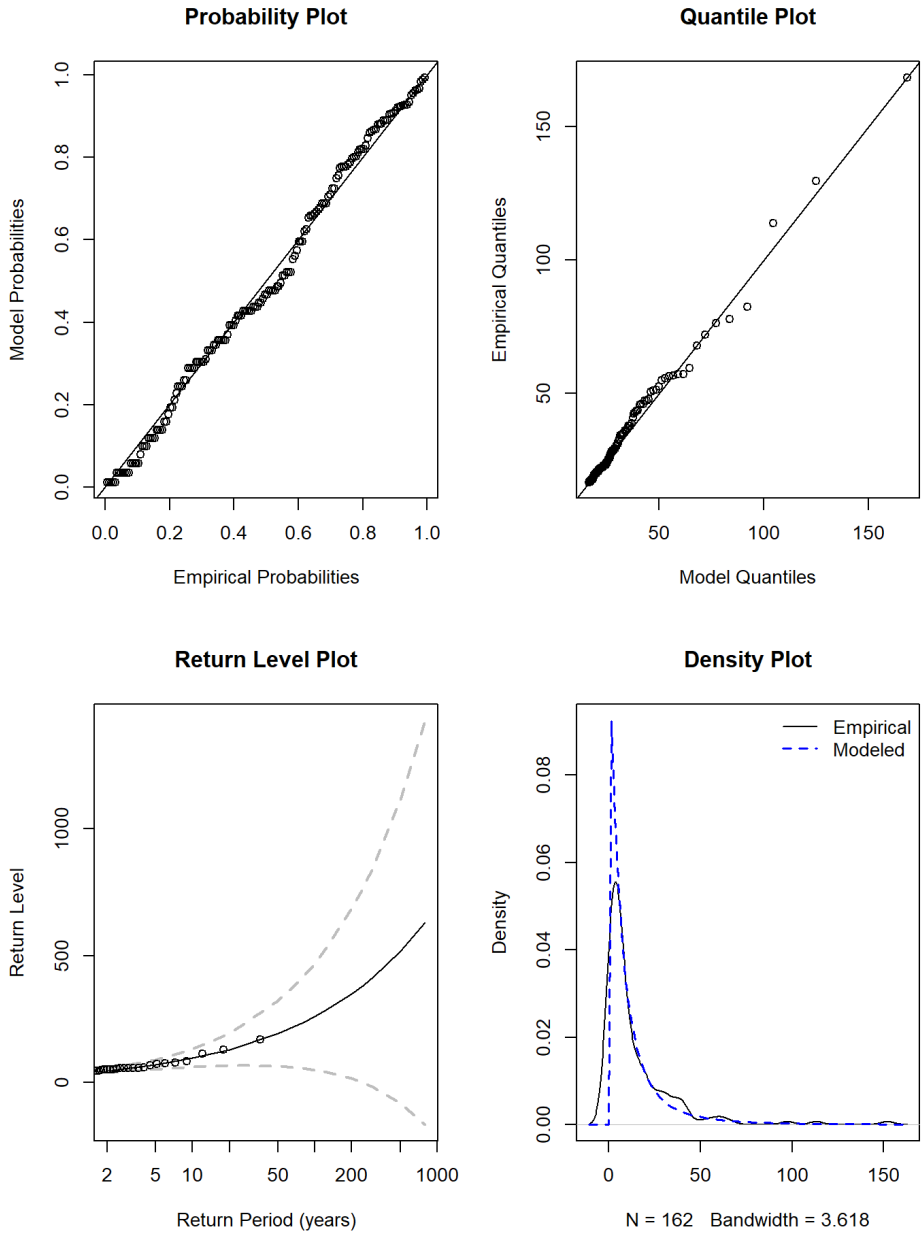


**Figure 4.10:** Diagnostic plots for rainfall events total intensity

The exploration of total rainfall data through applying the POT approach, in conjunction with the Generalized Pareto Distribution (GPD) model, demonstrates compelling outcomes. The efficacy of this modelling strategy is substantiated by the diagnostic evaluations presented in Figure 4.10, highlighting the model's proficiency in accurately capturing the data's extreme value characteristics. The GPD model's adaptation to the total rainfall data is notably successful, accurately encapsulating the data distribution's intricacies, especially the extreme values that are crucial in extreme value analysis. The

shape parameter, $\gamma$, is crucial for understanding the tail behaviour of the distribution. The estimated value 0.4220, as shown in the parameter estimates table (Table 4.5), indicates a heavy-tailed distribution. This positive shape parameter suggests that the distribution has a significant propensity for extreme total rainfall events, a vital consideration for hydrological risk assessment and infrastructure planning.

The quantile plot, an integral part of the diagnostic figures, visually represents the fit between the empirical data and the theoretical model. The excellent alignment observed in the quantile plot, particularly the inclusion of the most extreme rainfall event, confirms the GPD model's capability to model the tail of the total rainfall distribution accurately. This alignment is a statistical success and a practical affirmation of the model's reliability in predicting extreme events. The return level plot further reinforces the model's adequacy. The estimated return levels from the GPD model are found to be well-aligned with the empirical data, all lying within the model-derived confidence intervals. This unity is essential as it underlines the model's effectiveness in providing reliable estimates for extreme total rainfall levels, which are crucial for designing resilient water management and flood prevention strategies.

As substantiated by the diagnostic plots, the positive shape parameter and the good fit of the model collectively suggest that the POT method with the GPD model is well-suited for modelling the extremes of total rainfall data. The model's ability to capture the highest rainfall total events and align the return level estimates with the empirical data within the confidence intervals enhances our confidence in its predictive power and utility in extreme value analysis. The compelling evidence of the GPD model's good fit, especially its success in encapsulating the extremes, underscores its potential as a reliable tool in extreme rainfall analysis. This leads to a nuanced understanding of the rainfall patterns, enabling accurate risk assessments and informed decision-making for managing the potential impacts of extreme rainfall events. Therefore, the current findings solidify the methodological framework for total rainfall analysis, advocating its continued use and further exploration in hydrological modelling and extreme event prediction.

## 4.7 Multivariate Threshold Model using GPD

Let's consider two sequences of vectors, $\{X_i\}$ and $\{Y_i\}$, that are independently and uniformly distributed (iid), following the distribution function denoted as $F(x,y)$. We then define the vector of their componentwise maxima as $M_n = (M_{x,n}, M_{y,n})$ where $M_{x,n} = \max_{i=1,\ldots,n}\{X_i\}$ and $M_{y,n} = \max_{i=1,\ldots,n}\{Y_i\}$. As $n$ approaches infinity, the limiting behaviour of this vector is represented by $G_*(x,y)$. Where $G_*$ is a distribution function that is not degenerate, $G_*$ takes the form [102]:

$$G_*(x,y) = \exp[-V(x,y)]; \quad x > 0, \quad y > 0 \tag{4.15}$$

where

$$V(x,y) = 2\int_0^1 \max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right) dH(\omega) \tag{4.16}$$

and $H$ represents a distribution function that lies within the [0,1] range and fulfils the requirement of the mean constraint

$$\int_0^1 \omega\, dH(\omega) = \frac{1}{2} \tag{4.17}$$

The six bivariate extreme value models (Table 4.7) in the POT R package was applied in this study. The logistic (log), asymmetric logistic (alog), negative logistic (blog), asymmetric negative logistic (anlog), mixed (mix), and asymmetric mixed (amix). The model of bivariate threshold excess is designed to estimate the joint distribution $F(x,y)$ in areas where $x > u_x$ and $y > u_y$, given sufficiently large $u_x$ and $u_y$. When appropriate thresholds are in place, each of the marginal distributions of F can be approximated using a univariate GPD [103]. The marginals $x$ and $y$ are transformed as [102]:

$$x^* = -\left\{\log\left(1 - \zeta_{u_x}\left[1 + \gamma_x\left(\frac{x-u_x}{\beta_u}\right)\right]\right)^{-1/\gamma_x}\right\} \tag{4.18}$$

and

$$y^* = -\left\{\log\left(1 - \zeta_{u_y}\left[1 + \gamma_y\left(\frac{x-u_y}{\beta_u}\right)\right]\right)^{-1/\gamma_y}\right\} \tag{4.19}$$

$(x^*, y^*)$ are Fréchet transformed variables for $X > u_x$, and $Y > u_y$ while $\zeta_u$ gives the exceedance rate.

$$F(x,y) \approx G_*(x,y) = \exp[-V(x^*,y^*)]; \quad x > u_x, \quad y > u_y \tag{4.20}$$

**Table 4.7:** Bivariate Extreme Value Models

| **Model** | $V(y_1, y_2)$ [104] | **Independence** | **Dependence** |
|---|---|---|---|
| log | $\left(y_1^{-1/\alpha} + y_2^{-1/\alpha}\right)^{\alpha}, \quad 0 < \alpha \leq 1$ | $\alpha = 1$ | $\alpha \to 0$ |
| alog | $\frac{1-t_1}{y_1} + \frac{1-t_2}{y_2} + \left[\left(\frac{y_1}{t_1}\right)^{-\frac{1}{\alpha}} + \left(\frac{y_2}{t_2}\right)^{-\frac{1}{\alpha}}\right]^{\alpha}$, $0 < \alpha \leq 1, 0 \leq t_1, t_2 \leq 1$ | $\alpha = 1$ $t_1 = 0$ or $t_2 = 0$ | $\alpha \to 0$ |
| nlog | $\frac{1}{y_1} + \frac{1}{y_2} - \left(y_1^{\alpha} + y_2^{\alpha}\right)^{-\frac{1}{\alpha}}, \quad \alpha > 0$ | $\alpha \to 0$ | $\alpha \to +\infty$ |
| anlog | $\frac{1}{y_1} + \frac{1}{y_2} - \left[\left(\frac{y_1}{t_1}\right)^{\alpha} + \left(\frac{y_2}{t_2}\right)^{\alpha}\right]^{-\frac{1}{\alpha}}, \quad \alpha > 0$ $0 < t_1, t_2 \leq 1$ | $\alpha \to 0$ | $\alpha \to +\infty$ |
| mix | $\frac{1}{y_1} + \frac{1}{y_2} - \frac{\alpha}{y_1+y_2}, \quad 0 < \alpha \leq 1$ | $\alpha = 0$ | |
| amix | $\frac{1}{y_1} + \frac{1}{y_2} - \frac{(\alpha+t)y_1 + (\alpha+2t)y_2}{(y_1+y_2)^2}$ $\alpha \geq 0, \quad \alpha + 2t \leq 1, \quad \alpha + 3t \geq 0$ | $\alpha = t = 0$ | |

The measure of dependence $\chi$ for bivariate variables is employed to quantify extreme dependence. Assuming $W_1$ and $W_2$ are random variables and suppose $F_1(W_1)$ and $F_2(W_2)$ are transforms of $W_1$ and $W_2$ using $F$, $\chi$ is defined as [105]:

$$\chi = \lim_{u \to 1} Pr(F_2(W_2) > u | F_1(W_1) > u) \tag{4.21}$$

for $0 < u < 1$,

$$\chi(u) = 2 - \frac{\log Pr(F_1(W_1) < u, F_2(W_2) < u)}{\log Pr(F_1(W_1) < u)} \tag{4.22}$$

$$= 2 - \frac{\log Pr(F_1(W_1) < u, F_2(W_2) < u)}{\log(u)} \tag{4.23}$$

then

$$\chi = \lim_{u \to 1} \chi(u).$$

For asymptotically independent variables, $\chi(u) = 0$ for all u in (0,1), and for totally dependent variables, $\chi(u) = 1$ for all u in (0,1).

## 4.8 Results and Discussion of the Bivariate Analysis

The bivariate GPD models described above were applied to the rain event duration and intensity data. The thresholds used during the univariate extreme value analysis for rain event duration and intensity were adopted. The parameter estimates were obtained by numerical optimization of the loglikelihood function using the POT Package in R [104].

**Table 4.8:** Bivariate Model Comparison: Parameter Estimates, Deviance, and AIC

| Parameters | log | alog | nlog | anlog | mix | amix |
|---|---|---|---|---|---|---|
| $\beta_x$ | 39.332 | 39.373 | 43.306 | 39.3756 | 39.36 | 39.27280 |
| $\gamma_x$ | 0.040 | 0.125 | 0.071 | 0.1167 | 0.0915 | 0.14775 |
| $\beta_y$ | 0.439 | 0.454 | 0.556 | 0.5144 | 0.541 | 0.50007 |
| $\gamma_y$ | 0.316 | 0.300 | 0.344 | 0.3399 | 0.374 | 0.30261 |
| $t$ | | | | | | -0.0369 |
| $t_1$ | | 0.796 | | 0.5095 | | |
| $t_2$ | | 0.858 | | 0.7487 | | |
| $\alpha$ | 0.999 | 0.999 | 0.02844 | 0.2009 | 0.0005 | 0.1111 |
| Deviance | 4182.183 | 4179.283 | 4170.093 | 4178.074 | 4170.915 | 4183.251 |
| AIC | 4192.183 | 4193.283 | 4180.093 | 4192.074 | 4180.915 | 4195.251 |
| $\chi$ | 0.001 | 0.001 | 0 | 0.02 | 0 | 0.028 |

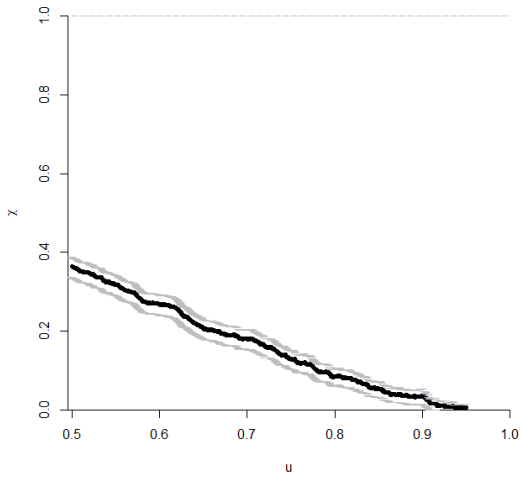


**Figure 4.11:** Estimates of $\chi(u)$

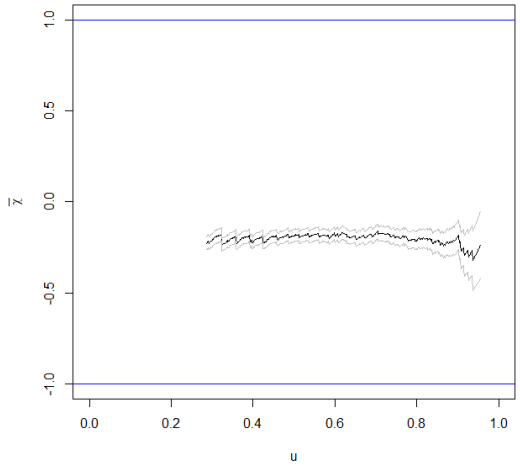


**Figure 4.12:** Estimates of $\bar{\chi}(u)$

From the bivariate result in Table 4.8, the negative logistic provides the best fit for the joint extremes of rainfall event duration and intensity by possessing the least AIC (4180.093) and deviance value (4170.093), and the bivariate extreme value model using the negative logistic distribution gave a $\chi = 0$ which indicates extremal independence between rainfall event duration and intensity. Using the negative logistic bivariate GPD for the duration and intensity of the rain event, we generated 1000 sample points, as shown in Figure 4.13. Figure 4.14 gives bivariate return level plot for rainfall event duration and intensity. The bivariate negative logistic (nlog) model is proficient at examining positive dependencies between extreme values in two variables. It effectively captures scenarios where such extremes are positively correlated. Despite its strengths, the model lacks a detailed parametric structure for inter-marginal dependencies. It restricts its ability to predict scenarios where an extreme event in one variable would correspond to a non-extreme event in the other [103]. Consequently, its effectiveness is reduced in situations requiring the analysis of negative or inverse correlations between extreme events, especially when these extreme values are not concurrently occurring. The dependence parameter ($\alpha$) within the model is tailored to highlight positive extremal dependencies, lacking the functionality to address negative dependencies. This limitation confines the model's use in applications necessitating forecasting negative extremal dependencies. Here, we

identified the negative bivariate logistic model as the most optimal among tested bivariate models for representing joint extreme duration and intensity data. However, as we can see from comparing simulations and original data in Figure 4.13, even the best bivariate extreme value distribution constitutes a poor fit, and we do not pursue this route any further.

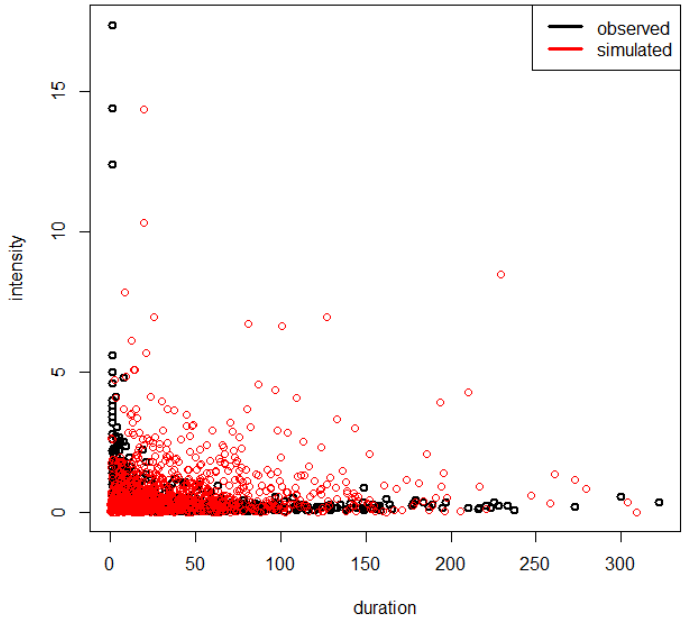


**Figure 4.13:** Simulated bivariate plot vs actual data

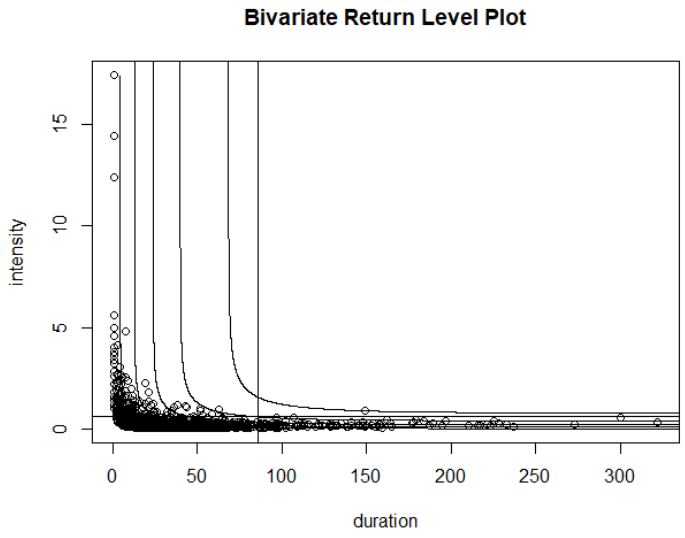


**Figure 4.14:** Bivariate return level plot for rainfall event duration and intensity

# Chapter 5

## Marginal Modelling with Explicit Tail Decay

## 5.1 Introduction

Modelling real-life processes involving extreme events, like rainfall, may pose challenges because of their unforeseeable characteristics. However, we can fit a GPD to the tail of the data using a suitable threshold. This approach allows us to model extreme events accurately. However, how do we simultaneously model the rest of the distribution? This question is crucial since most of the data is in the non-extreme region. Finding a straightforward closed-form probability distribution suitable for fitting such events might not always be feasible. Given this challenge, several statistical techniques have been proposed for modelling this data type, like the kernel density estimation approach [106] and the non-parametric Bayesian framework [107, 108] and mixture models.

To model this data type, we are modelling a 1-dimensional distribution using different parametric forms for different domain segments. The parametric modelling process typically necessitates employing a parametric hybrid model. In this structure, each hybrid component is uniquely tailored to manage a distinct data segment, enhancing the model's effectiveness and precision. Examples of such hybrid models defined and studied in the literature include the composite lognormal-Pareto model [109], the hybrid Pareto model [110], the hybrid exponential distribution [50], Normex - mixed normal [111], the lognormal-Burr model [112], the gaussian-exponential-GPD [113], right-truncated composite lognormal-Pareto distribution [114] and the three-Part composite Pareto [115].

If the data exhibits a heavy-tailed nature, the tail component can be accurately modelled using the GPD [116]. Simultaneously, an alternative distribution may handle the bulk

of the data. In the case of right-tailed asymmetric data sets, a distribution that aligns with the bulk of the data is selected and combined with the GPD, with an intermediate distribution serving as a bridge between the two. The three-component hybrid model extends the foundational two-part composite Pareto framework initially presented by Cooray and Ananda [109]. The choice of a three-component model over a simpler two-component model is motivated by the requirement for a more detailed representation of the data. This additional component offers a refined modelling capacity, particularly for datasets exhibiting intermediate behaviours or transitional states, which two-component models do not adequately capture. Such a detailed approach is essential for applications where overlooking these intermediate behaviours could result in significant modelling biases or inaccuracies, where the mid-range dynamics are as crucial as the tails [117]. This chosen distribution for the main innovation is expected to represent the data's mean characteristics. The three distributions that comprise the three-component hybrid system are each given a separate scaling. As a result, the three-component hybrid model would contain two junction points. If unnecessary elements are added, the space between two consecutive junction points naturally approaches zero [113].

This chapter presents the development of a hybrid model designed for fitting rainfall event data (heavy-tailed) by extending and generalising G-Exp-GPD, the work of Debbabi et al. [113]. Here, the exponential distribution's incorporation as the intermediate component to connect the mean and asymptotic behaviours of the data is pivotal. The exponential distribution offers a mathematically tractable option that ensures continuity and differentiability in the model. It bridges the gap between the central data bulk and the extreme tails, providing a smooth transition that is essential for the overall cohesiveness and effectiveness of the model. The choice of this distribution is strategic, leveraging its simplicity and analytical tractability to model the moderately extreme events or the tail of the central data distribution, thereby enriching the model's capacity to accurately represent the entire data spectrum. We introduce an F-Exp-GPD model where 'F' is optimally fitted to the considered data. This approach enhances the model's fit and deepens our understanding of the fundamental processes that generate the observed

data, contributing significantly to the field rainfall event simulation.

## 5.2 The New Hybrid Distribution

Suppose we have non-negative data with a heavy right tail. The objective is to employ a piecewise model to fit the data. Let $f(x;\theta)$ denote a valid density function with parameter vector $\theta$. It's assumed that $f(x;\theta)$ is continuous, and its first derivative exists. Furthermore, consider $e(x;\lambda)$ as the density function of the exponential distribution with the rate parameter $\lambda$ defined as:

$$e(x;\lambda) = \lambda e^{-\lambda x}; \quad \lambda > 0, \quad x > 0 \tag{5.1}$$

The cdf corresponding to the density in Equation 5.1 can be represented as:

$$E(x;\lambda) = 1 - e^{-\lambda x}; \quad \lambda > 0, \quad x > 0 \tag{5.2}$$

The cdf in Equation 5.2 has a corresponding inverse (quantile function) given by:

$$Q_1(u) = -\frac{1}{\lambda}\log(1-u); \quad \lambda > 0, \quad 0 < u < 1 \tag{5.3}$$

Let $g(x;\gamma,\beta)$ be the density function of the generalized Pareto distribution (GPD), with $\gamma$ and $\beta$ as the shape and scale parameters respectively ($\beta > 0$), and it can be expressed as:

$$g(x;\gamma,\beta) = \begin{cases} \frac{1}{\beta}\left(1+\frac{\gamma}{\beta}x\right)^{-1-\frac{1}{\gamma}} & \text{if } \gamma \neq 0 \\ \frac{1}{\beta}e^{-\frac{x}{\beta}}, & \text{if } \gamma = 0 \end{cases} \tag{5.4}$$

For all $x \in D(\gamma,\beta)$. Let $D(\gamma,\beta)$ represent the domain of $g$, then:

$$D(\gamma,\beta) = \begin{cases} [0,\infty) & \text{if } \gamma \geq 0 \\ \left[0, -\frac{\beta}{\gamma}\right] & \text{if } \gamma < 0 \end{cases}$$

The cdf corresponding to the density in Equation 5.4 can be articulated as:

$$G(x;\gamma,\beta) = \begin{cases} 1-\left(1+\frac{\gamma}{\beta}x\right)^{-\frac{1}{\gamma}} & \text{if } \gamma \neq 0 \\ 1-e^{-\frac{x}{\beta}} & \text{if } \gamma = 0 \end{cases} \tag{5.5}$$

The inverse of the cdf in Equation 5.5 can be formulated as:

$$Q_2(u) = \begin{cases} \frac{\beta}{\gamma}[(1-u)^{-\gamma}-1] & \text{if } \gamma \neq 0 \\ -\beta\log(1-u) & \text{if } \gamma = 0 \end{cases} \tag{5.6}$$

where $0 < u < 1$.

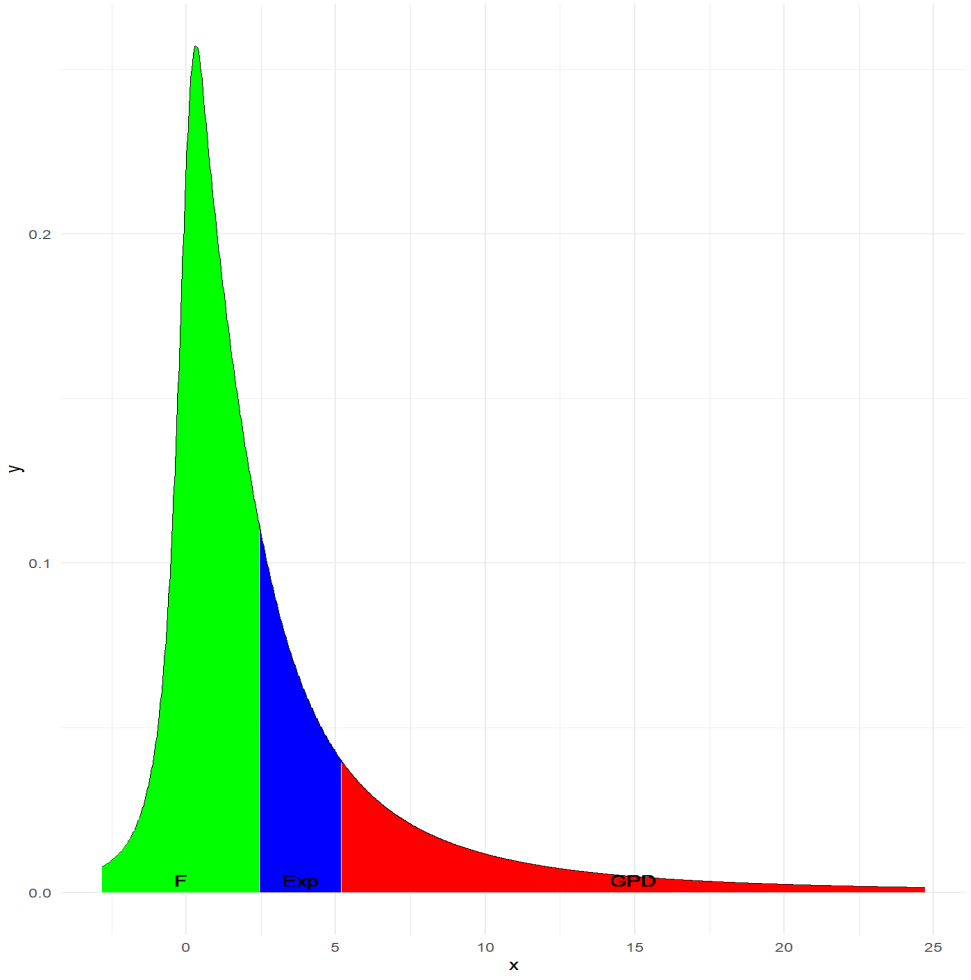


**Figure 5.1:** Hybrid F-Exp-GPD

Let $f(x;\theta)$ be the density used to model the bulk, with CDF $F(x;\theta)$. We define a three-component hybrid F-Exponential-GPD model by the pdf of the form

$$h(x;\theta,\lambda,\gamma,\beta,t_1,t_2,r_1,r_2,r_3) = \begin{cases} r_1 f(x;\theta) & \text{if } -\infty < x \leq t_1 \\ r_2 e(x;\lambda) & \text{if } t_1 < x \leq t_2 \\ r_3 g(x-t_2;\gamma,\beta) & \text{if } t_2 < x < \infty \end{cases} \tag{5.7}$$

where $r_1, r_2$, and $r_3$ denote the weight corresponding to each density part, $h$. An example of the new hybrid density is illustrated using Figure 5.1

Firstly, the model presumes that $\mathbf{h}$ is non-negative and functions as a valid density. Thus integrating $\mathbf{h}(x;\theta)$ over all $x$ yields one. This translates to $r_1F(t_1;\theta)+r_2(E(t_2)-E(t_1))+r_3=1$. Secondly, the model assumes that $\mathbf{h}$ is continuous and differentiable at the two junction points or thresholds $t_1$ and $t_2$. The cdf corresponding to the hybrid density, $\mathbf{h}$ can be expressed as

$$H(x;\Theta) = \begin{cases} r_1 F(x;\theta) & \text{if } -\infty < x < t_1 \\ r_1 F(t_1;\theta) + r_2(e^{-\lambda t_1} - e^{-\lambda x}) & \text{if } t_1 < x \leq t_2 \\ 1 - r_3(1+\frac{\gamma}{\beta}(x-t_2))^{-1/\gamma} & \text{if } t_2 < x < \infty \end{cases} \tag{5.8}$$

The quantile function is derived in Appendix C.1 and given by

$$H^{-1}(u;\Theta) = \left\{ \begin{array}{ll} F^{-1}(\frac{u}{r_1};\theta) & \text{if } u \leq u_1 = r_1F(t_1;\theta) \\ \frac{1}{\lambda}\log\left[\frac{r_2}{u_1-u+r_2e^{-\lambda t_1}}\right] & \text{if } u_1 \leq u \leq u_2 = 1-r_3 \\ \frac{\beta}{\gamma}\left[(1-\frac{u-u_2}{r_3})^{\gamma}-1\right]+t_2 & \text{if } u \geq u_2 \end{array} \right\} \tag{5.9}$$

Matching the cdf and pdf and matching their derivatives we have

$$\begin{cases} r_1F(t_1;\theta) = r_2e(t_1;\lambda); & r_1F'(t_1;\theta) = r_2e'(t_1;\lambda) \\ r_2e(t_2;\lambda) = r_3g(x-t_2;\gamma,\beta); & r_2e'(t_2;\lambda) = r_3g'(x-t_2;\gamma,\beta) \end{cases} \tag{5.10}$$

Using the model assumptions and resolving Equation 5.10 we have

$$\begin{cases} \lambda = -\frac{f'(t_1;\theta)}{f(t_1;\theta)}; & r_1 = r_2 \dfrac{e(t_1;\lambda)}{f(t_1;\theta)} \\ \beta = \frac{\gamma+1}{\lambda}; & r_2 = \left[(\lambda\beta - 1)e^{-\lambda t_2} + (1 + \lambda \dfrac{F(t_1;\theta)}{f(t_1;\theta)})e^{-\lambda t_1}\right] \\ r_3 = \beta r_2 e(t_2;\lambda) & \end{cases} \tag{5.11}$$

Thus $\Theta = (\theta, t_1, t_2, \gamma)$ is the number of free parameters to be estimated.

## 5.3 Specific Cases of the Hybrid Model

### 5.3.1 St-Exp-GPD

Suppose $f(x;\theta)$ is the density function of the Skew t distribution with $\theta = (\mu, \sigma, \alpha, v)$ where $\mu, \sigma, \alpha$ and $v$ are location, scale, shape and degrees of freedom parameters respectively. The pdf of the Skew t can be written as

$$f(x;\mu,\sigma,\alpha,v) = \frac{2}{\omega} t(z;v) T(\alpha z \sqrt{\frac{v+1}{v+z^2}}; v+1) \qquad x \in \mathbf{R} \tag{5.12}$$

where $z = \frac{x-\mu}{\sigma}$. $t(\cdot;v)$ and $T(\cdot;v+1)$ are the pdf and cdf of the student t distributions respectively. Also $F(x;\mu,\sigma,\alpha,v)$ and $F^{-1}(P;\mu,\sigma,\alpha,v)$ are the cdf and the quantile function of the Skew-t distribution respectively. Now, we have that

$$f'(u_1;\mu,\sigma,\alpha,v) = \frac{dz}{dx}\frac{df}{dz} \tag{5.13}$$

where $\frac{dz}{dx} = \frac{1}{\omega}$ and $\frac{df}{dz} = \frac{2}{\omega}\left[v\frac{du}{dz} + u\frac{dv}{dz}\right]$

Therefore

$$\frac{df}{dz} = \frac{2}{\omega}\left[(\frac{-z(v+1)}{v})\left(1+\frac{z^2}{v}\right)^{-1} t(z,v)T(p;v+1) + t(z,v)t(p;v+1)\alpha(v+1)^{\frac{1}{2}} \frac{v}{(v+z^2)^{\frac{3}{2}}}\right] \tag{5.14}$$

Then

$$\frac{df}{dx} = \frac{2}{\omega^2} t(z,v) \left[ \left( \frac{-z(v+1)}{v} \right) \left( 1 + \frac{z^2}{v} \right)^{-1} T(p; v+1) + t(p; v+1) \alpha (v+1)^{\frac{1}{2}} \frac{v}{(v+z^2)^{\frac{3}{2}}} \right] \tag{5.15}$$

Appendix C.2 gives the r functions for St-Exp-GPD.

### 5.3.2 Sn-Exp-GPD

Suppose $f(x;\theta)$ is the density function of the Skew-Normal distribution with $\theta = (\mu, \sigma, \alpha)$, where $\mu, \sigma$ and $\alpha$ are location, scale and shape parameters respectively, we have that

$$f(x; \mu, \sigma, \alpha) = \frac{2}{\sigma} \phi(\frac{x-\mu}{\sigma}) \Phi[\alpha(\frac{x-\mu}{\sigma})] \quad \alpha \in \mathbf{R};\ x \in \mathbf{R},\ \sigma > 0,\ x \geq \mu \tag{5.16}$$

Where $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal distribution's pdf and cdf, respectively. Also, $F(x; \mu, \sigma, \alpha)$ and $F^{-1}(u; \mu, \sigma, \alpha)$ are the cdf and quantile function of the Skew-Normal distribution respectively.

We realise that

$$\begin{aligned} f'(t_1; \mu, \sigma, \alpha) &= \frac{2}{\sigma} \left\{ \phi' \left( \frac{t_1 - \mu}{\sigma} \right) \Phi \left[ \alpha \left( \frac{t_1 - \mu}{\sigma} \right) \right] + \phi \left( \frac{t_1 - \mu}{\sigma} \right) \Phi' \left[ \alpha \left( \frac{t_1 - \mu}{\sigma} \right) \right] \right\} \\ &= \frac{2}{\sigma} \left\{ \frac{\alpha}{\sigma} \phi \left[ \alpha \left( \frac{t_1 - \mu}{\sigma} \right) \right] \phi \left( \frac{t_1 - \mu}{\sigma} \right) - \left( \frac{t_1 - \mu}{\sigma^2} \right) \phi \left( \frac{t_1 - \mu}{\sigma} \right) \Phi \left[ \alpha \left( \frac{t_1 - \mu}{\sigma} \right) \right] \right\} \\ &= \frac{2\phi \left( \frac{t_1 - \mu}{\sigma} \right)}{\sigma^2} \left\{ \alpha \phi \left[ \alpha \left( \frac{t_1 - \mu}{\sigma} \right) \right] - \left( \frac{t_1 - \mu}{\sigma^2} \right) \Phi \left[ \alpha \left( \frac{t_1 - \mu}{\sigma} \right) \right] \right\} \end{aligned} \tag{5.17}$$

Appendix C.3 provides the r functions for Sn-Exp-GPD.

### 5.3.3 GEV-Exp-GPD

Let $f(x;\theta)$ be the density function of the GEV distribution with $\theta = (\mu, \sigma, k)$ where $\mu, \sigma$ and $k$ are location, scale and shape parameters respectively. The pdf of the GEV distribution can be written as

$$f(x; \mu, \sigma, k) = \frac{1}{\sigma} [v(x)]^{k+1} e^{-v(x)}; \quad \mu \in \mathbf{R},\ \sigma > 0,\ k \in \mathbf{R},\ x \in D(\mu, \sigma, k) \tag{5.18}$$

where

$$v(x) = \left[1 + k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/k} \tag{5.19}$$

and

$$D(\mu,\sigma,k) = \begin{cases} [\mu - \frac{\sigma}{k}, \infty) & \text{when } k > 0 \\ (-\infty, \infty) & \text{when } k = 0 \\ [-\infty, \mu - \frac{\sigma}{k}] & \text{when } k < 0 \end{cases} \tag{5.20}$$

Also, let $F(x;\mu,\sigma,k)$ and $F^{-1}(P;\mu,\sigma,k)$ be the cdf and quantile function of the GEV distribution respectively.

Then

$$\begin{aligned} f'(x;\mu,\sigma,k) &= \frac{1}{\sigma}\left\{ -v'(x)\Big[v(x)\Big]^{k+1} e^{-v(x)} + (k+1)v'(x)\Big[v(x)\Big]^{k} e^{-v(x)} \right\} \\ &= \frac{1}{\sigma}\left\{ (k+1)v'(x)\Big[v(x)\Big]^{k+1}\Big[v(x)\Big]^{-1} e^{-v(x)} - v'(x)\Big[v(x)\Big]^{k+1} e^{-v(x)} \right\} \\ &= v'(x)\frac{1}{\sigma}\Big[v(x)\Big]^{k+1} e^{-v(x)}\left\{ (k+1)[v(x)]^{-1} - 1 \right\} \\ &= v'(x) f(x;\mu,\sigma,k)\left\{ (k+1)[v(x)]^{-1} - 1 \right\} \end{aligned} \tag{5.21}$$

where

$$v'(x) = -\frac{1}{\sigma}\left[1 + k(\frac{x-\mu}{\sigma})\right]^{-1/k-1} \tag{5.22}$$

Appendix C.4 gives the r functions for GEV-Exp-GPD.

**Table 5.1:** Three Hybrid Models and their Parameters

| Parameters | ST-Exp-GPD | SN-Exp-GPD | GEV-Exp-GPD |
|---|---|---|---|
| $f$<br>Exp<br>GPD | $\mu,\sigma,\alpha,v$<br>$\lambda$<br>$\gamma,\beta$ | $\mu,\sigma,\alpha$<br>$\lambda$<br>$\gamma,\beta$ | $\mu,\sigma,k,$<br>$\lambda$<br>$\gamma,\beta$ |
| Threshold<br>Weights | $t_1,t_2$<br>$r_1,r_2,r_3$ | $t_1,t_2$<br>$r_1,r_2,r_3$ | $t_1,t_2$<br>$r_1,r_2,r_3$ |
| Total number of parameters | 12 | 11 | 11 |
| Number of free parameters | 7 | 6 | 6 |

The ST-Exp-GPD, SN-Exp-GPD, and GEV-Exp-GPD models possess 7, 6, and 6 degrees

of freedom respectively. The corresponding sets of free parameters for each model are $\theta = [\mu, \sigma, \alpha, v, t_1, t_2, \gamma]$ for ST-Exp-GPD, $\theta = [\mu, \sigma, \alpha, t_1, t_2, \gamma]$ for SN-Exp-GPD, and $\theta = [\mu, \sigma, k, t_1, t_2, \gamma]$ for GEV-Exp-GPD.

## 5.4 Maximum Likelihood Estimation of the Parameters of the Hybrid Model

The parameter estimation process for a parametric family of distributions, through the maximum likelihood method, encompasses maximising the loglikelihood function. This maximization pertains to the distribution parameters, contingent on a random independent sample of size $n$ derived from the distribution above. The maximum likelihood estimator (MLE) converges to the true parameter as the sample size increases, and the MLE is asymptotically efficient. Considering a density function, denoted as $h(x;\Theta)$, that contains an unknown parameter vector $\Theta$, along with rainfall event data, expressed as $x_1, x_2, \ldots, x_n$ and the loglikelihood function can thus be defined as:

$$L = \sum_{i=1}^{n} \ln\left(h(x_i;\Theta)\right). \tag{5.23}$$

Suppose $\Theta$ is the unknown parameter vector; the associated score function is given by

$$U(\Theta) = \frac{\partial L}{\partial \Theta_i},$$

The partial differentiation of the loglikelihood function with respect to the $i^{th}$ parameter in the vector is represented by the expression $\frac{\partial L}{\partial \Theta_i}$. Solving the system of equations $U(\Theta) = 0$ gives us the maximum likelihood estimate of $\Theta$. When these systems of equations do not have a straightforward analytical solution, numerical methods can be employed to find the solutions. For this purpose, the *maxLik* library in R was utilized to implement the numerical solution process.

## 5.5 Simulation Study

In this section, we conduct a simulation study to evaluate the effectiveness and precision of the hybrid distribution's Maximum Likelihood Estimates (MLEs). We generate 100 samples, each with sizes of n = 1000, 10000, and 20000, across various parameter values, utilizing the quantile function of the Hybrid distribution. For each generated sample, we derive the MLEs, which are then employed to calculate values for the specified metrics using R (Here $\Theta = (\mu, \sigma, k, t_1, t_2, \gamma)$).

(i) Mean (ME) $= \frac{1}{N}\sum_{i=1}^{N}\hat{\Theta}$

(ii) Average Bias (AB) $= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Theta} - \Theta\right)$

(iii) Root Mean Square Error (RMSE) $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Theta} - \Theta\right)^2}$

(iv) The average width (AW) of 95% confidence intervals for the parameters, $\Theta$.

**Table 5.2:** Results of Monte-Carlo simulations for $\mu = 2.5$, $\sigma = 1$, $k = 0.5$, $t_1 = 4.5$, $t_2 = 5$, and $\gamma = 0.2$

| Parameters | | $10^3$ | $10^4$ | $2 \times 10^4$ |
|---|---|---|---|---|
| $\mu = 2.5$ | $ME_{\hat{\mu}}$ | 2.5221 | 2.5039 | 2.5029 |
| | $AB_{\hat{\mu}}$ | 0.0221 | 0.0039 | 0.0029 |
| | $RSME_{\hat{\mu}}$ | 0.0614 | 0.0217 | 0.0146 |
| | $AW_{\hat{\mu}}$ | 0.2392 | 0.0808 | 0.0608 |
| $\sigma = 1$ | $ME_{\hat{\sigma}}$ | 1.0404 | 1.0083 | 1.0042 |
| | $AB_{\hat{\sigma}}$ | 0.0403 | 0.0083 | 0.0042 |
| | $RSME_{\hat{\sigma}}$ | 0.0971 | 0.0368 | 0.0226 |
| | $AW_{\hat{\sigma}}$ | 0.4079 | 0.1359 | 0.1022 |
| $k = 0.5$ | $ME_{\hat{k}}$ | 0.5305 | 0.5074 | 0.5031 |
| | $AB_{\hat{k}}$ | 0.0305 | 0.0074 | 0.0031 |
| | $RSME_{\hat{k}}$ | 0.0826 | 0.0323 | 0.0197 |
| | $AW_{\hat{k}}$ | 0.3681 | 0.1228 | 0.0914 |
| $t_1 = 4.5$ | $ME_{\hat{t_1}}$ | 4.2629 | 4.4720 | 4.5210 |
| | $AB_{\hat{t_1}}$ | -0.2371 | -0.0280 | 0.0210 |
| | $RSME_{\hat{t_1}}$ | 0.7947 | 0.6542 | 0.5925 |
| | $AW_{\hat{t_1}}$ | 4.3107 | 2.8296 | 2.4143 |
| $t_2 = 5$ | $ME_{\hat{t_2}}$ | 5.6904 | 5.2867 | 5.0365 |
| | $AB_{\hat{t_2}}$ | 0.6904 | 0.2867 | 0.0364 |
| | $RSME_{\hat{t_2}}$ | 1.5287 | 1.0783 | 0.6306 |
| | $AW_{\hat{t_2}}$ | 5.6960 | 3.5688 | 1.9631 |
| $\gamma = 0.2$ | $ME_{\hat{\gamma}}$ | 0.2303 | 0.2092 | 0.2072 |
| | $AB_{\hat{\gamma}}$ | 0.0303 | 0.0092 | 0.0072 |
| | $RSME_{\hat{\gamma}}$ | 0.1081 | 0.0327 | 0.0221 |
| | $AW_{\hat{\gamma}}$ | 0.4246 | 0.1196 | 0.0879 |

Table 5.2 presents the values of four metrics: mean (ME), average bias (AB), root mean squared error (RMSE), and average width (AW) for the parameters across varying sample sizes of n = $10^3$, $10^4$, and $2 \times 10^4$. Our observations from the results reveal that the average bias and RMSE tend to decrease with the increment in sample size. Furthermore, as the sample size increases, the average width of these confidence intervals shows a decreasing trend. Consequently, the MLEs and their asymptotic properties can effectively be employed for parameter estimation and confidence interval construction, particularly for reasonable sample sizes.

## 5.6 Application to Rain Events Data and Discussion

To show the usefulness of the newly proposed hybrid models, we employed the three constructed distributions: St-Exp-Gpd, Sn-Exp-Gpd, and Gev-Exp-Gpd. For comparative analysis, the G-Exp-GPD model by Debbabi et al. [113] was also incorporated to fit the duration and intensity of rainfall events. The Maximum Likelihood Estimation (MLE) technique estimated these distributions' parameters. The computational execution was conducted in the R environment. Herein, the loglikelihood function was optimised by applying the Nelder-Mead method. The parameter estimates, loglikelihood, and AIC values of all the fitted hybrid distribution models are given in Table 5.3 and Table 5.4 for log(duration) and log(intensity). Figure 5.2 and Figure 5.4 provides a graph of all the hybrid model densities alongside the histogram of log(duration) and log(intensity). Figure 5.3 and Figure 5.5 give the QQ plots for the fitted hybrid models for log(duration) and log(intensity) data, respectively.

**Table 5.3:** MLE fit for log(duration) using the hybrid models

| Distributions | ST-Exp-GPD | SN-Exp-GPD | GEV-Exp-GPD | G-Exp-GPD |
|---|---|---|---|---|
| Parameter estimates | $\mu = 3.3223$<br>$\sigma = 1.0062$<br>$\alpha = -0.8496$<br>$v = 9.3070$<br>$t_1 = 4.4579$<br>$t_2 = 5.5032$<br>$\gamma = -0.6029$ | $\mu = 3.6434621$<br>$\sigma = 1.3199$<br>$\alpha = -1.8017$<br>$t_1 = 4.3066$<br>$t_2 = 5.5735$<br>$\gamma = -0.6551$ | $\mu = 2.3854$<br>$\sigma = 0.8999$<br>$k = -0.5771$<br>$t_1 = 3.5858$<br>$t_2 = 4.7943$<br>$\gamma = -0.3316$ | $\mu = 2.7547$<br>$\sigma = 0.9400$<br>$t_2 = 5.3753$<br>$\gamma = 0.0508$ |
| LogLikelihood | -4754.263 | -4760.153 | -4750.769 | -4784.212 |
| AIC | 9522.526 | 9532.306 | 9513.538 | 9576.424 |

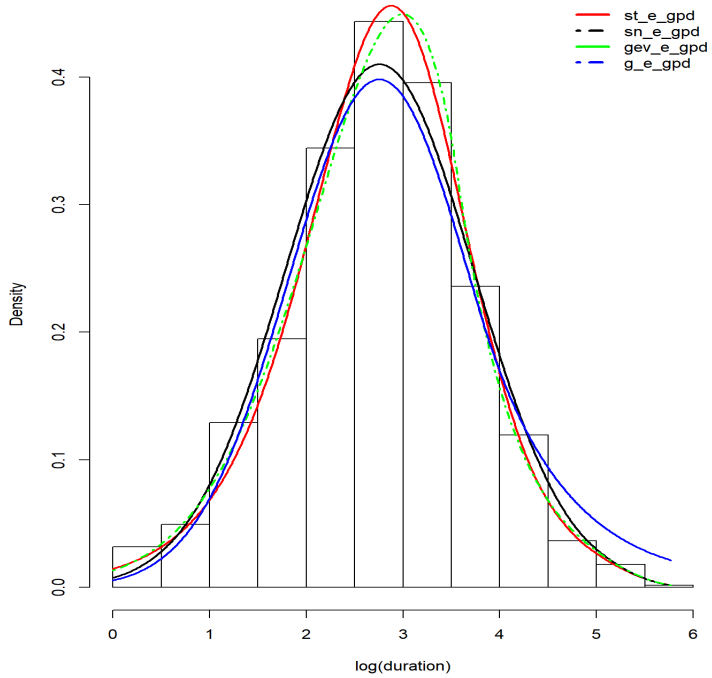


**Figure 5.2:** Density plot for log(duration) using the hybrid models

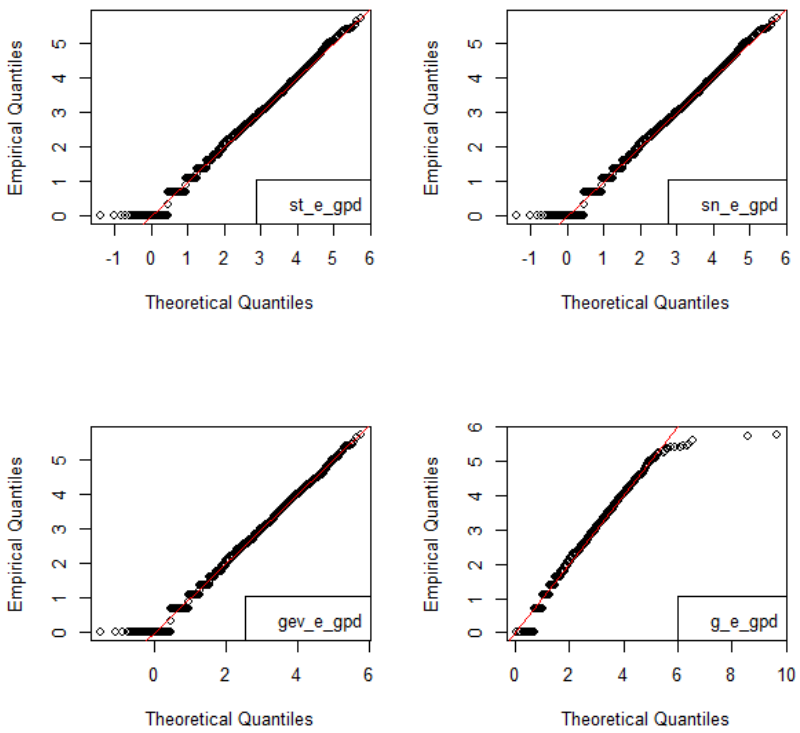


**Figure 5.3:** Q-Q plot for log(duration) with the hybrid Models

The results indicate that the GEV-Exp-GPD model is the most suitable among the hybrid models for describing the duration and intensity of rainfall events. The observed lowest AIC and highest loglikelihood values (see Table 5.3 and Table 5.4) are also better than the AIC values for the skew t distribution for log(duration) (see Table 3.2) and GEV distribution for log(intensity) (see Table 3.3) in Chapter 3. This validates the GEV-Exp-GPD's efficacy in capturing the data's underlying distribution. The model's versatility in accommodating different distributional characteristics highlights its potential for other hydrological applications where the bulk and tail behaviour must be accurately represented. The presence of zero or near-zero values in log-transformed rain event durations results from including very short events, such as those lasting only 1.00 (6 mins). Despite their brevity, these values are critical for a detailed analysis of rainfall patterns, representing the lower end of event durations. By including these short events, our model comprehensively captures the full range of rainfall dynamics, emphasizing the significance of even the briefest rain events.

**Table 5.4:** MLE fit for log(intensity) using the hybrid models

| Distributions | ST-Exp-GPD | SN-Exp-GPD | GEV-Exp-GPD | G-Exp-GPD |
|---|---|---|---|---|
| Parameter estimates | $\mu = -2.6181$<br>$\sigma = 1.1143$<br>$\alpha = 2.8803$<br>$v = 11.7954$<br>$t_1 = 0.1753$<br>$t_2 = 1.3350$<br>$\gamma = -0.1439$ | $\mu = -2.6667$<br>$\sigma = 1.2231$<br>$\alpha = 3.3404$<br>$t_1 = 0.015$<br>$t_2 = 1.5235$<br>$\gamma = -0.0443$ | $\mu = -2.0738$<br>$\sigma = 0.6697$<br>$k = -0.0444$<br>$t_1 = 1.5445$<br>$t_2 = 1.8309$<br>$\gamma = -0.0176$ | $\mu = -1.731$<br>$\sigma = 0.7804$<br>$t_2 = 0.6712$<br>$\gamma = 0.3773$ |
| LogLikelihood | -3989.025 | -3978.132 | -3970.288 | -4123.839 |
| AIC | 7992.05 | 7966.447 | 7952.576 | 8255.678 |

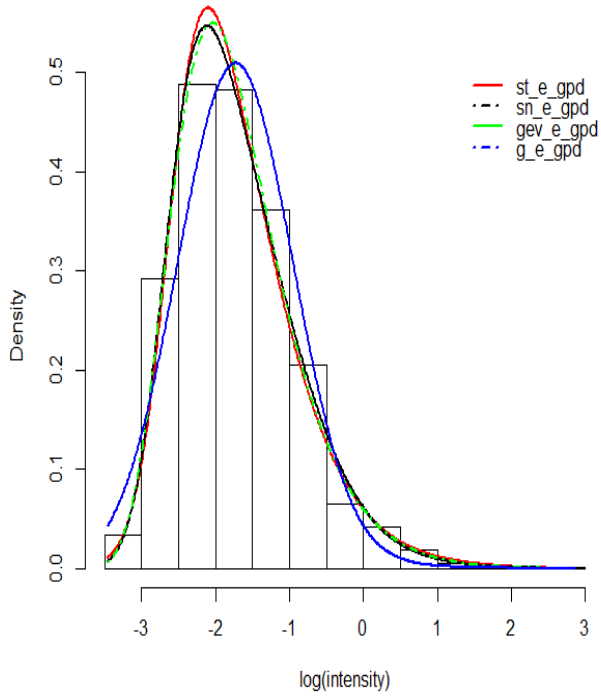


**Figure 5.4:** Density plot for log(intensity) with the hybrid models

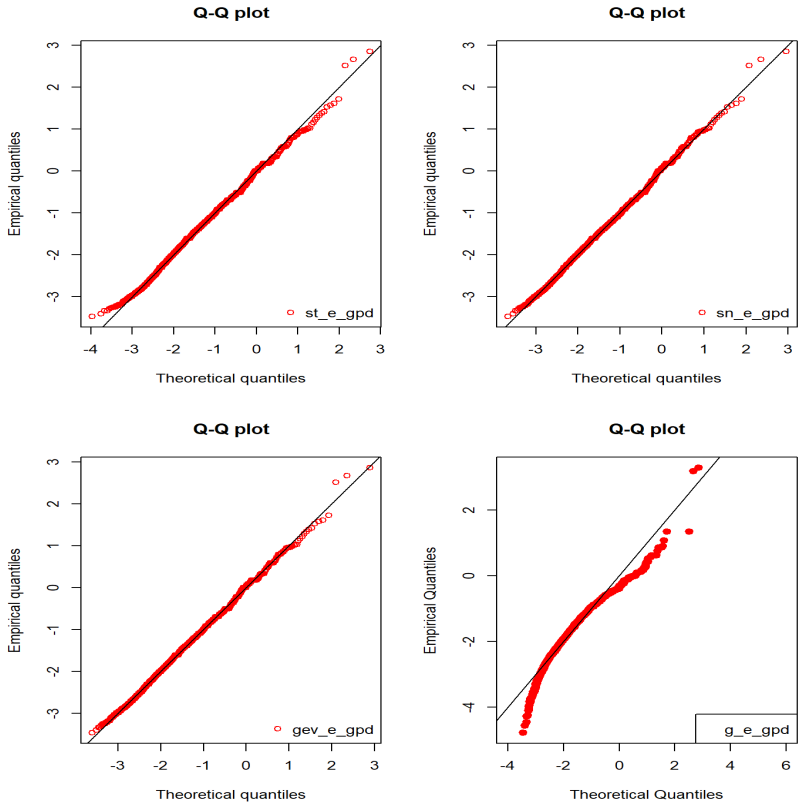


**Figure 5.5:** Q-Q plot for log(intensity) with hybrid Models

Furthermore, the Skew t-Exp-GPD model is also a good fit for the rainfall event duration, as evident from the density and QQ plots. From Figure 5.3 and Figure 5.5, the QQ plots reveal that our hybrid models adequately fit the bulk and tail of the rainfall duration and intensity. From the results of $\gamma$ in Table 5.3 and Table 5.4, we can see that our rainfall duration and intensity have a light tail with $\gamma < 0$. The outcomes of this research exhibit promising implications for modelling rainfall event duration and intensity using the generalized F-Exp-GPD framework. By introducing an arbitrary distribution F, where F is the best fit marginal distribution for the dataset, this study enriches the flexibility of the hybrid modelling approach. This approach considers the Gaussian and GPD to represent the bulk and tail behaviour and adapts to the dataset's unique characteristics through the F distribution.

# Chapter 6

## Dependence Modelling

## 6.1 Introduction

Rainfall events are summarized by duration, intensity, maximum intensity and volatility. The dependence modelling of these variables is essential to understanding the relationship between the different rainfall properties and their impact on various hydrological processes. The multivariate Gaussian distribution is one of the most popular models used. Multivariate Gaussian distributions are rarely adequate for summarising multivariate data because the univariate margins may be skewed or heavily tailed, and the joint distribution may exhibit tail dependency or asymmetries stronger than Gaussian dependence [118]. From the data, we can see that there is no apparent multivariate distribution to model the dependence of the data; therefore, we are proposing the copula approach as it can be used to model complex dependencies [118] and has been applied in multivariate modelling of hydrological variables [119].

## 6.2 Copula

Copulas are flexible functions that join two or more univariate distribution functions to create a multivariate distribution for modelling the dependence structure of variables independent of their marginal distributions. Sklar's theorem is the foundation for the development and applications of copula theory. Given the m-dimensional random vector $Y = (y_1, ..., y_m)$ with univariate marginal distributions $F_i(y_i)$ for $i = 1, ..., m$ Sklar theorem states that the joint distribution can be defined as [120]:

$$F(y_1, ..., y_m) = C(F_1(y_1), ..., F_m(y_m)) \tag{6.1}$$

Where $C(.)$ is a m-dimensional copula that converts the marginal uniform distributions to their multivariate CDF defined in $[0,1]^m \longrightarrow [0,1]$.

**Definition 6.2 (Copula)**[121]: A m-dimensional copula, denoted as $C$, is a function that fulfils the following requirements:

- Dom $C = I^m = [0,1]^m$
- $C$ has margins $C_m$ satisfying $C_m(u) = C(1,...,1,u,1,...,1) = u \ \forall u \in I$

When $m$ is two, we have a bivariate or 2-dimensional copula C with domain $I^2$ given the following terms.

- $\forall \ u, w$ in $I$

$$C(u,0) = 0 = C(0,w) \tag{6.2}$$

and

$$C(u,1) = u \quad \text{and} \quad C(1,w) = w; \tag{6.3}$$

- $\forall \ u_1, u_2, w_1, w_2$ in $I \ni u_1 \leq u_2$ and $w_1 \leq w_2$

$$C(u_2,w_2) - C(u_2,w_1) - C(u_1,w_2) + C(u_1,w_1) \geq 0 \tag{6.4}$$

Numerous bivariate copula families exist, each with distinct properties such as tail dependence and asymmetric behaviour. Transitioning from bivariate to multivariate copulas is complex due to the need for precise yet adaptable models. Historically, only elliptical (containing the Gaussian and t-copula) and Archimedean families (including Clayton and Gumbel copulas) were considered multivariate, but they had symmetry and tail dependency limitations. Vine copulas overcome these limits by building a multivariate model solely from bivariate copulas.

If variables $X$ and $Y$ represent rainfall events log(duration) and log(intensity) respectively, with $H_X(x)$ and $H_Y(y)$ being their respective distribution function, then the joint pdf of

$X$ and $Y$ can be expressed as [121]:

$$H_{X,Y}(x,y) = C\Big(H_X(x), H_Y(y); \theta\Big) = C(u,w;\theta) \tag{6.5}$$

The parameter $\theta$ is the variable linked to the copula function, $u = H_X(x)$ and $w = H_Y(y)$.

### 6.2.1 Measure of Dependence

In this section, we discuss the measures of dependence in copulas, notably Kendall's tau and Spearman's rho.

**Definition 6.2.1 (Concordance) [24]**: if $(X_i, Y_i)^T$ and $(X_j, Y_j)^T$ are two pairs of observations of the continuous random variable $(X,Y)^T$ then $(X_i, Y_i)^T$ and $(X_j, Y_j)^T$ are concordant and discordant when $(X_i - X_j)(Y_i - Y_j) > 0$ and $(X_i - X_j)(Y_i - Y_j) < 0$ respectively.

**Definition 6.2.2 (Kendall Tau) [122]:** Given the random vector $(X,Y)^T$ the Kendall tau is defined by the expression:

$$\tau(X,Y) = \mathbf{Pr}\{(X_i - X_j)(Y_i - Y_j) > 0\} - \mathbf{Pr}\{(X_i - X_j)(Y_i - Y_j) < 0\} \tag{6.6}$$

**Theorem 6.2.1 [122]:** Given the continuous random variables $X$ and $Y$ with marginal distributions F and G, then the Kendall tau is given as

$$\tau(X,Y) = 4 \int \int_{I^2} C_2(u,w)\, dC_1(u,w) - 1 \tag{6.7}$$

$$\tau(X,Y) = 4\mathbf{E}(C(U,W)) - 1 \tag{6.8}$$

*Proof*: See [121]

The theoretical Kendall's tau of the different bivariate copulas used in this study is listed in Table 6.1.

**Definition 6.2.3 (Spearman's rho) [122]:** Given the random vector $(X,Y)^T$ the Spear-

man's rho is defined by the expression

$$\rho s(X,Y) = 3(\mathbf{Pr}\{(X_i - X_j)(Y_i - Y_j) > 0\} - \mathbf{Pr}\{(X_i - X_j)(Y_i - Y_j) < 0\}) \tag{6.9}$$

**Theorem 6.2.2 [121]:** Given the continuous random variables $X$ and $Y$ with marginal distributions F and G and joint distribution functions $H_1$ and $H_2$, respectively, If $C$ denote their copula, then the Spearman's rho is given as

$$\rho s(X,Y) = 12 \int \int_{I^2} u, w \, dC(u,w) - 3 = 12 \int \int_{I^2} C(u,w) \, du \, dw - 3 \tag{6.10}$$

Let $U = F(X)$ and $W = G(Y)$, then

$$\rho s(X,Y) = 12 \int \int_{I^2} u, w \, dC(u,w) - 3 = 12\mathbf{E}(UW) - 3 \tag{6.11}$$

$$= \frac{\mathbf{E}(UW) - 1/4}{1/12} = \frac{\mathbf{E}(UW) - \mathbf{E}(U)\mathbf{E}(W)}{\sqrt{Var(U)}\sqrt{Var(W)}} \tag{6.12}$$

### 6.2.2 Tail Dependence

A measure of dependence strength in a multivariate distribution's joint lower or upper tail is known as tail dependence. It represents a metric that describes the extent of extreme dependence between two random variables [123]. The coefficient of tail dependency is a conditional probability that ranges from 0 to 1.

Lower tail dependence for a bivariate copula [17]:

$$\Lambda_L = \lim_{q \searrow 0^+} Pr\{X \leq F_X^{-1}(q) | Y \leq G_Y^{-1}(q)\} = \lim_{q \searrow 0^+} \frac{C(q,q)}{q} \tag{6.13}$$

If $\Lambda_L \in (0,1]$, and $\Lambda_L = 0$, then the copula, $C$ has lower tail dependence and lower tail independence respectively.

Upper tail dependence for a bivariate copula:

$$\Lambda_U = \lim_{q \nearrow 1^-} Pr\{X > F_X^{-1}(q) | Y > G_Y^{-1}(q)\} = \lim_{q \nearrow 1^-} \frac{1 - 2q + C(q,q)}{1 - q} \tag{6.14}$$

If $\Lambda_U \in (0,1]$, and $\Lambda_U = 0$, then the copula, $C$ has upper tail dependence and upper tail independence respectively. Where $q$ is the quantile of $X$ and $Y$. The upper and lower tail dependence of the copulas used in research is given in Table 6.2.

### 6.2.3 Copula family

This study utilized different copulas that can be used to capture negative and positive dependence: the Normal, t, Frank, Clayton, Gumbel, Joe, and Tawn Copula. The density functions of these bivariate copulas are as follows:

- Normal Copula [124]: for $\theta \in (-1,1)$

$$C(u,w;\theta) = \frac{1}{\sqrt{1-\theta^2}} \exp\left(\frac{2\theta\Phi^{-1}(u)\Phi^{-1}(w) - \theta^2(\Phi^{-1}(u)^2 + \Phi^{-1}(w)^2)}{2(1-\theta^2)}\right) \tag{6.15}$$

  where $\Phi^{-1}(\cdot)$ corresponds to the inverse of the standard normal distribution function

- t Copula [125]: for $\theta \in (-1,1)$ and $v \in (2,\infty)$

$$C(u,w;\theta,v) = \frac{\Gamma((v+2)/2)}{\Gamma(v/2)\pi v\sqrt{1-\theta^2}}\left(1 + \frac{x^2 - 2\theta xy + y^2}{v}\right)^{\frac{-(v+2)}{2}} \tag{6.16}$$

  where $x = t_v^{-1}(u)$, $y = t_v^{-1}(w)$ and $t_v$ is the Student's t distribution with $v$ degree of freedom

- Clayton Copula [119]: for $\theta \in (0,\infty)$

$$C(u,w;\theta) = (u^{-\theta} + w^{-\theta} - 1)^{-1/\theta} \tag{6.17}$$

- Frank Copula [119]: for $\theta \in (-\infty,\infty), \theta \neq 0$

$$C(u,w;\theta) = \frac{-1}{\theta}\ln\left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta w} - 1)}{e^{-\theta} - 1}\right] \tag{6.18}$$

- Joe Copula [126]: for $\theta \in (1,\infty)$

$$C(u,w;\theta) = 1 - [(1-u)^\theta + (1-w)^\theta - (1-u)^\theta(1-w)^\theta]^{1/\theta} \tag{6.19}$$

- Gumbel Copula [127]: for $\theta \in [1, \infty]$

$$C(u, w; \theta) = \exp\Big\{ -[(-\ln u)^{\theta} + (-\ln w)^{\theta})]^{1/\theta} \Big\} \tag{6.20}$$

- Tawn Copula [128, 17]: for $\theta \in [1, \infty)$ and $t \in [0, 1]$

$$C(u, w; \theta) = (u, w) \exp\{A(\varphi)\} \tag{6.21}$$

with $\varphi = \frac{\log(u)}{\log(uw)}$ and $A(\cdot)$ is the Tawn copula Pickand function given as:

$$A(t) = (1-\psi_2)(1-t) + (1-\psi_1)t + \Big[ (\psi_1(1-t))^{\theta} + (\psi_2 t)^{\theta} \Big]^{1/\theta} \tag{6.22}$$

with $0 \leq \psi_1, \psi_2 \leq 1$.

The Tawn copula is a Gumbel copula with two extra asymmetry parameters, $\psi_1$ and $\psi_2$. We obtain the Gumbel copula when both of these parameters equal 1 [128]. The Type 1 and Type 2 Tawn copulas, which refer to $\psi_2 = 1$ and $\psi_1 = 1$ [17]. The normal and t copulas belong to the class of elliptical copulas. The Frank, Clayton, Joe, and Gumbel copulas belong to the Archimedean copulas, while the Tawn copula is an extreme value copula (see [123] for definition of copula classes).

**Table 6.1:** Kendall Tau of different bivariate copula families [17]

| **Copula Families** | **Kendall Tau** | **Range of $\tau$** |
|---|---|---|
| Normal | $\frac{2}{\pi} \arcsin \theta$ | $[-1, 1]$ |
| t | $\frac{2}{\pi} \arcsin \theta$ | $[-1, 1]$ |
| Clayton | $\frac{\theta}{\theta+2}$ | $[0, 1]$ |
| Frank | $1 - \frac{4}{\theta} + 4\frac{D(\theta)}{\theta}$ with $D(\theta) = \int_0^{\theta} \frac{x/\theta}{\exp(x)-1}\, dx$ (Debye function) [129] | $[-1, 1]$ |
| Joe | $1 + \frac{4}{\theta^2} \int_0^1 x \log(x)(1-x)^{2(1-\theta)/\theta}\, dx$ | $[0, 1]$ |
| Gumbel | $1 - \frac{1}{\theta}$ | $[0, 1]$ |
| Tawn Type 1 | $\int_0^1 \frac{t(1-t)A''(t)}{A(t)}$ with $A(t) = (1-\psi_1)t + [(\psi_1(1-t))^{\theta} + t^{\theta}]^{1/\theta}$ | $[0, 1]$ |

**Table 6.2:** Tail Dependence of different bivariate copula families (− means undefined) [15]

| **Copula Families** | **Lower Tail Dependence** | **Upper Tail Dependence** |
|---|---|---|
| Normal | - | - |
| t | $2t_{v+1}\left(-\sqrt{v+1}\sqrt{\frac{1-\theta}{1+\theta}}\right)$ | $2t_{v+1}\left(-\sqrt{v+1}\sqrt{\frac{1-\theta}{1+\theta}}\right)$ |
| Clayton | $2^{-1/\theta}$ | - |
| Frank | - | - |
| Joe | - | $2-2^{1/\theta}$ |
| Gumbel | - | $2-2^{1/\theta}$ |
| Tawn | - | $(\psi_1+\psi_2)-(\psi_1^\theta+\psi_2^\theta)^{1/\theta}$ |

If $C(\cdot,\cdot)$ is a bivariate copula, rotating it will cause the tail dependency to move to one of the four vertices of the unit square, and the resulting versions may be derived using the formulas below [130, 15]:

$$C_{r90^\circ}(u,w) = C(1-w,u) \tag{6.23}$$

$$C_{r180^\circ}(u,w) = C(1-u,1-w) \tag{6.24}$$

$$C_{r270^\circ}(u,w) = C(w,1-u) \tag{6.25}$$

Figure 6.1 gives an example of copula rotation.

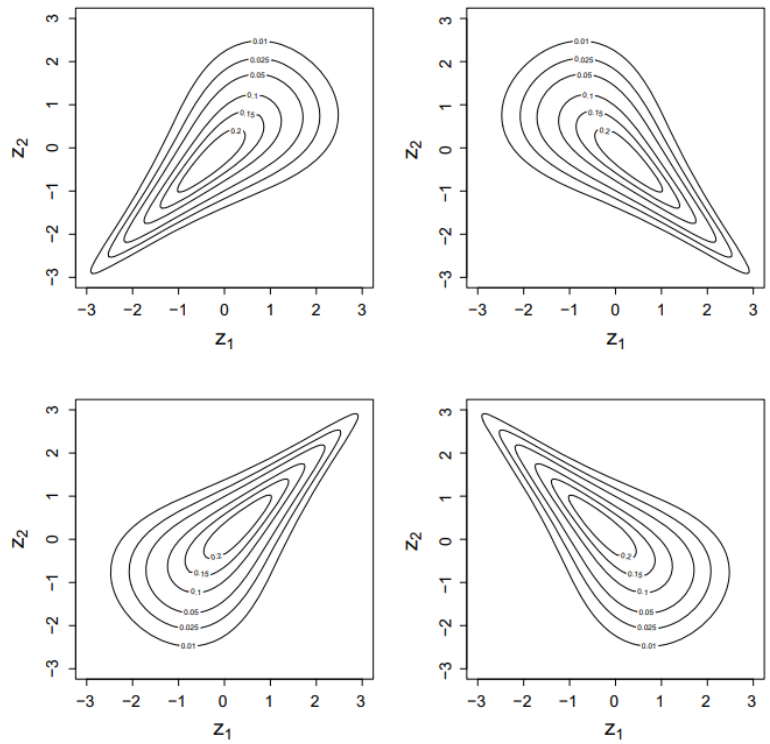


**Figure 6.1:** Clayton copula rotation with contour plots: left top: 0° rotation ($\tau = 0.5$), right top: 90° rotation ($\tau = -0.5$), left bottom: 180° rotation ($\tau = 0.5$), and right bottom: 270° rotation ($\tau = -0.5$)[15]

### 6.2.4 Estimating Bivariate Copula

Before estimating the copula parameters, the appropriate distributions for the marginals (duration and intensity) were selected. Given $X$ and $Y$ with their corresponding cdf as $H(x;\beta_1)$ and $H(y;\beta_2)$ and pdf as $h(x;\beta_1)$ and $h(y;\beta_2)$ where $\beta_1$ and $\beta_2$ are the vector parameters. Then the copula-based pdf is given as

$$H(x,y;\beta_1,\beta_2;\theta) = C(H_X(x;\beta_1), H_Y(y;\beta_2);\theta) \tag{6.26}$$

where $\theta$ is the vector of parameters of the copula function. The loglikelihood function of the copula model is given by

$$l(\theta,\beta_1,\beta_2) = \sum_{i=1}^{n} \log h(x_i,y_i;\beta_1,\beta_2,\theta) \tag{6.27}$$

where $n$ is the number of data. Using numerical optimisation, the copula parameters are obtained by maximizing Equation 6.27. The algorithm requires starting values, which were calculated by inverting Kendall's $\tau$. Inverting Kendall's tau in copula modelling involves finding the variable value corresponding to a given probability under a copula distribution. In cases with greater than one parameter, the process becomes more complex as the parameters influence the dependency structure, requiring numerical optimization to determine the variable value that satisfies the desired probability.

### 6.2.5 Model Selection for Copulas

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to select an appropriate copula model. Given the transformed data $u_i$ and $w_i$, $i = 1, ..., n$. The AIC and BIC for a bivariate copula $C(u, w; \theta)$ is given as [17]:

$$AIC = -2\sum_{i=1}^{N} \log\Big[C(u_i, w_i; \theta)\Big] + 2k \tag{6.28}$$

$$BIC = -2\sum_{i=1}^{N} \log\Big[C(u_i, w_i; \theta)\Big] + \log(N)k \tag{6.29}$$

Where $k$ is the number of copula parameters, and $N$ represents the total amount of data. The copula model with the lowest AIC and BIC values is the most appropriate among the copula models considered for the joint distribution of rainfall event duration and intensity. N represents the total amount of data.

### 6.2.6 Results and Discussion

The methodology described above for copula parameter estimation, model selection, and Kendall's tau measure of dependence was applied to the rainfall event duration and intensity data using the different considered copula models and the result is given Table 6.3. As seen in the copula pairs plot (Figure 6.2), duration and intensity exhibit negative dependence $\tau = -0.32$.

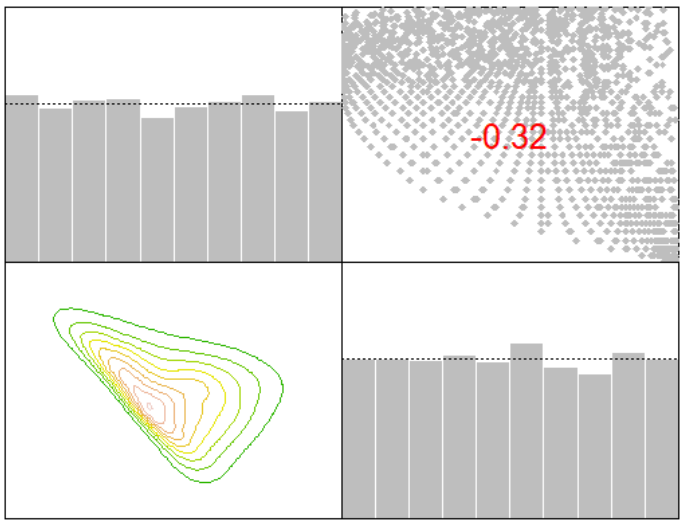


**Figure 6.2:** Upper Triangle: pair plot of copula data (duration and intensity), Diagonal: Marginal histogram of copula data, and Lower Triangle: empirical contour plots of normalized copula data

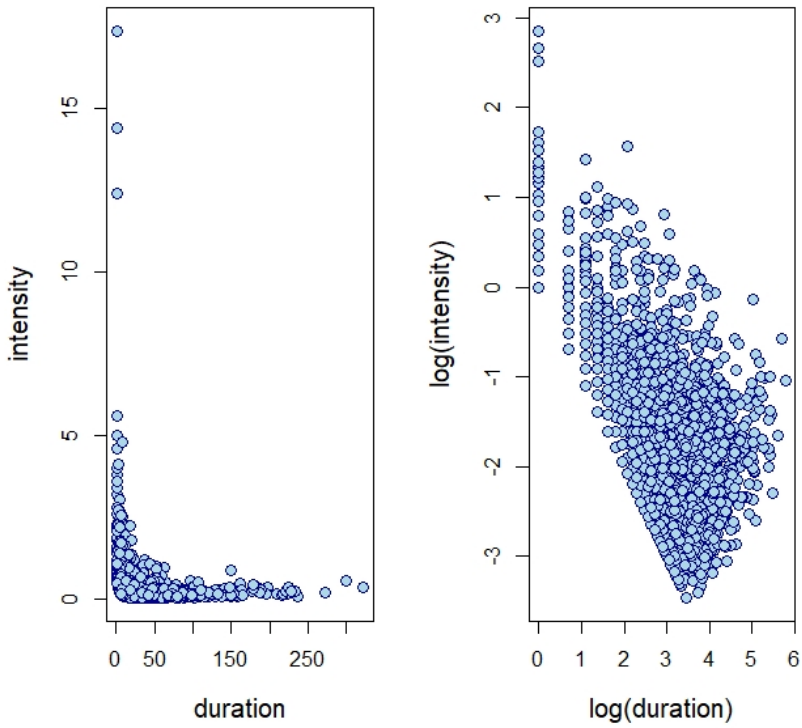


**Figure 6.3:** Scatter plot of the rainfall event duration and intensity

**Table 6.3:** Results for parameter estimates, loglikelihood, $AIC$, $BIC$, $\tau$, $\Lambda_L$, $\Lambda_U$

| Copula | Parameter(s) | loglikelihood | AIC | BIC | $\tau$ | $\Lambda_L$ | $\Lambda_U$ |
|---|---|---|---|---|---|---|---|
| Normal | $\theta_1$=-0.47 | 416.71 | -831.43 | -825.28 | -0.31 | 0 | 0 |
| t | $\theta_1$=-0.47 | 422.36 | -840.71 | -828.42 | -0.31 | 0 | 0 |
| | $\theta_2$=19.67 | | | | | | |
| Rotated 270° Clayton | $\theta = -1$ | 662.66 | -1323.31 | -1317.17 | -0.33 | 0 | 0 |
| Frank | $\theta = -2.99$ | 385.27 | -768.53 | -762.39 | -0.31 | 0 | 0 |
| Rotated 90° Joe | $\theta = -1.83$ | 657.8 | -1313.59 | -1307.45 | -0.32 | 0 | 0 |
| Rotated 90° Gumbel | $\theta = -1.49$ | 579.02 | -1156.03 | -1149.89 | -0.33 | 0 | 0 |
| **Rotated 90° Tawn Type 1** | $\theta_1$=-4.69 | **869.63** | **-1735.25** | **-1722.96** | -0.29 | 0 | 0 |
| | $\theta_2$=0.32 | | | | | | |

Various copula models were considered to capture the joint distribution of duration and intensity. Our evaluation revolves around different bivariate copulas, including Normal, t, Rotated 270° Clayton, Frank, Rotated 90° Joe, Rotated 90° Gumbel and Rotated 90° Tawn Type 1 copula. The performance of these copulas is compared based on log-likelihood, AIC, BIC, $\tau$, and tail dependence parameters, $\Lambda_L$ and $\Lambda_U$. From the result in Table 6.3, the Rotated 90° Tawn Type 1 copula exhibits the most desirable characteristics. This copula has the highest log-likelihood value of 869.35, surpassing all other tested copulas. Furthermore, this copula presents the lowest AIC of -1735.25. The parameters of the Rotated 90° Tawn Type 1 copula, $\theta_1 = -4.69$ and $\theta_2 = 0.32$, and the Kendall's tau is -0.29 for the Rotated 90° Tawn Type 1 copula. This value suggests a weak negative association between duration and intensity, consistent with our empirical observations. As for the tail dependence, none of the evaluated copulas exhibits lower or upper tail dependence, as evidenced by $\Lambda_L$ and $\Lambda_U$ equal to 0 across all models. This lack of tail dependence indicates no increased likelihood of extreme events coinciding in the distribution's lower or upper tails. Based on our evaluation criteria, the Rotated 90° Tawn Type 1 copula best represent the joint distribution of duration and intensity.

### 6.2.7 Bivariate Copula Simulation

This section explores data simulation using the suggested copula model and compares the simulated and observed data correlations. We used samples from the cumulative distributions to conduct the simulations. Generating $(u, w)$ from the tawn type 1 copula,

then applying the marginal transformations, i.e. skew t for log(duration) and GEV for log(intensity). See Appendix D.1 for the R code.

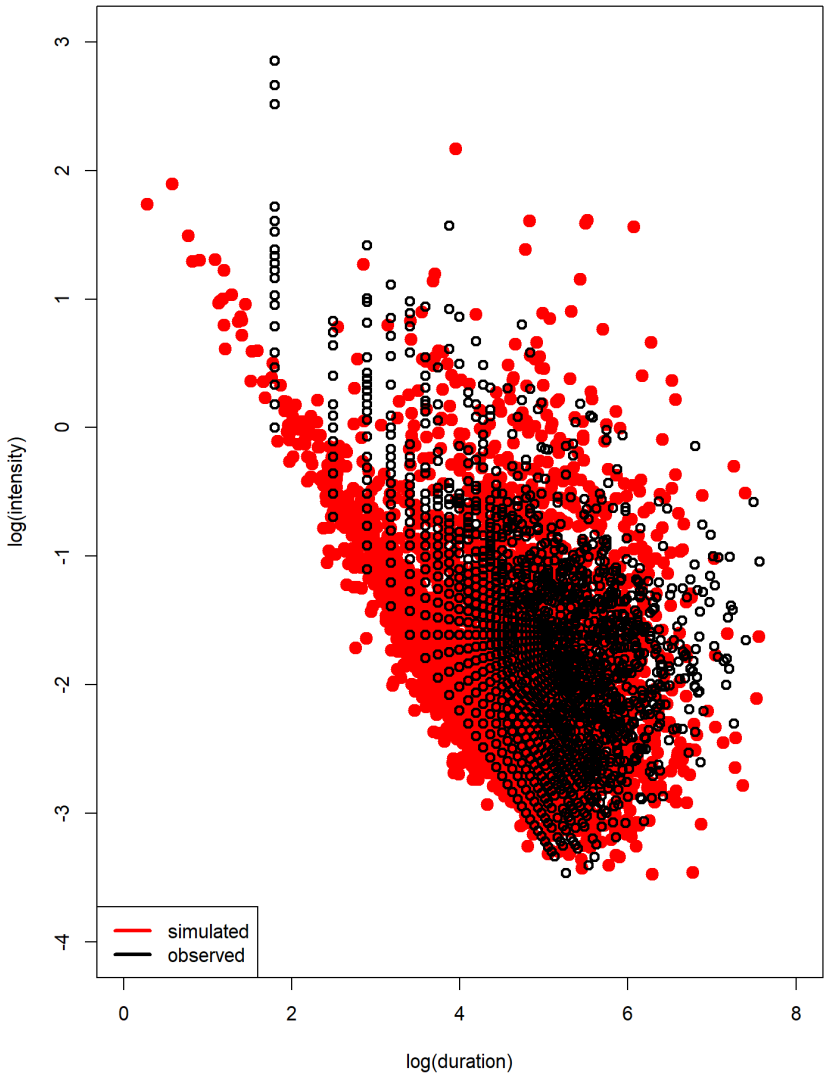


**Figure 6.4:** Scatter plot of observed data vs copula simulated data

Figure 6.4 shows the scatter plot of the observed data versus simulated data from the proposed copula model. The figure shows that the simulated data compares favourably with the observed data. We can see that both the simulated and original data exhibit similar dependence patterns.

## 6.3 Vine Copula Construction

Vine copula models aim to provide a method for creating multivariate copulas solely from bivariate copulas, which are highly flexible and capable of describing varieties of complex dependencies [131]. The modelling approach involves decomposing a multivariate density by utilizing a sequence of pair copulae. This methodology is applied to both the original variables and their respective conditional and unconditional distribution functions. [132]. Joe [133] provided the first pair copula structure in terms of distribution functions, whereas Bedford and Cooke [134, 135] produced constructs in terms of densities. Given the joint density function $f(y_1, ..., y_m)$ of a vector $Y = (Y_1, ..., Y_m)$ of random variables. This joint density function can be rewritten as

$$f(y_1, ..., y_m) = f(y_m) \cdot f(y_{m-1}|y_m) \cdot f(y_{m-2}|y_{m-1}, y_m) \cdots f(y_1|y_2, ..., y_m) \tag{6.30}$$

Equation 6.30 shows that the variables' dependency structure and probability distributions are implicitly included in the joint distribution function. Applying Sklar theorem to the density $f(y_1, ..., y_m)$ we have

$$f(y_1, ..., y_m) = c_{1...m}\{F_1(y_1), ..., F_m(y_m)\} \cdot f_1(y_1) \cdots f_m(y_m) \tag{6.31}$$

where $c_{1...m}$ is some m-variate copula density. Equation 6.31 can be simplified in a bivariate case as

$$f(y_1, y_2) = c_{12}\{F_1(y_1), F_2(y_2)\} \cdot f_1(y_1) \cdot f_2(y_2)$$

Here $c_{1,2}(\cdot, \cdot)$ is the proper pair-copula density for the two transformed variables $F_1(y_1)$ and $F_2(y_2)$. According to [132], the conditional probability distribution function in equation (3) can be decomposed into the proper pair-copula $(c_{y\nu_j|\nu_{-j}})$ and a conditional marginal density using the following formula

$$f(y|\nu) = c_{y\nu_j|\nu_{-j}}\{F(y|\nu_{-j}), F(\nu_j|\nu_{-j})\} \cdot f(y|\nu_{-j}) \tag{6.32}$$

where $\nu$ is a m-dimensional vector; $\nu_j$ is an arbitrary chosen component from vector $\nu$ and $\nu_{-j}$ is vector $\nu$, where $\nu_j$ is not included. Using Equation 6.32 the conditional density for two random variables $Y_1$ and $Y_2$ can be written as

$$f(y_1|y_2) = c_{12}\{F_1(y_1), F_2(y_2)\} \cdot f_1(y_1)$$

for three random variables $Y_1$, $Y_2$ and $Y_3$ we have

$$\begin{aligned} f(y_1|y_2, y_3) &= c_{12|3}\{F(y_1|y_3), F(y_2|y_3)\} \cdot f(y_1|y_3) \\ &= c_{12|3}\{F(y_1|y_3), F(y_2|y_3)\} \cdot c_{13}\{F_1(y_1), F_3(y_3)\} \cdot f_1(y_1) \end{aligned}$$

Pair-copula construction requires marginal conditional distributions like $F(y|\nu)$ and [133] demonstrated that, for every $j$,

$$F(y|\nu) = \frac{\partial c_{y\nu_j|\nu_{-j}}\{F(y|\nu_{-j}), F(\nu_j|\nu_{-j})\}}{\partial F(\nu_j|\nu_{-j})} \tag{6.33}$$

where $C_{y\nu_j|\nu_{-j}}$ is a bivariate copula function. When $\nu$ is univariate Equation 6.33 becomes

$$F(y|\nu) = \frac{\partial c_{y\nu}\{F(y), F(\nu)\}}{\partial F(\nu)} \tag{6.34}$$

When $y$ and $\nu$ are uniform, i.e., $f(y) = f(\nu) = 1, F(y) = y$ and $F(\nu) = \nu$ the function $h(y, \nu, \Theta)$ gives the conditional distribution [132].

$$h(y, \nu, \Theta) = F(y|\nu) = \frac{\partial C_{y,\nu}(y, \nu, \Theta)}{\partial \nu} \tag{6.35}$$

where $\nu$ denotes the conditioning variable and $\Theta$ is the copula's set of parameters for the joint distribution of $y$ and $\nu$.

### 6.3.1 Regular Vine

The regular vine structure was developed by [134] and [135] as a suitable graphical tool to model high-dimensional dependencies. A typical vine structure comprises linked trees,

where the edges of one tree become the nodes of the next tree.

**Definition 6.3.1** [136]: A $m$-dimensional regular vine copula $V$ is a pair-copula construction with m variables consisting of linked trees $T_1, ..., T_{m-1}$ where $N_j$ and $E_j$ are the nodes and edges respectively in tree $T_j$ satisfying:

1. $T_1$ consist nodes $N_1 = \{1, ..., m\}$ and edges $E_1$

2. For $j = 2, ..., m-1$, $T_j$ is with nodes $N_j = E_{j-1}$, i.e., the nodes in $T_j$ are the edges in $T_{j-1}$

3. Two edges in $T_j$ can only be joined as nodes in $T_{j+1}$ by an edge, if they share common node in $T_j$ (Proximity condition)

Each edge $e$ in $E_j$ is connected to a bivariate copula $C_{g(e),k(e)|D(e)}$ to construct a regular vine with nodes $N = \{N_1, ..., N_{m-1}\}$ and edges $E = \{E_1, ..., E_{m-1}\}$. Here, nodes $g(e)$ and $k(e)$ represent the nodes subject to conditioning, with $D(e)$ serving as the set that performs the conditioning. The combined set $\{g(e), k(e), D(e)\}$ acts as the set of constraints. The density of the R-Vine copula is given as [137]:

$$f(y_1, ..., y_m) = \Big[\prod_{k=1}^{m} f_k(y_k)\Big] \times \Big[\prod_{j=1}^{m-1} \prod_{e \in E_j} C_{g(e),k(e)|D(e)}(F(y_{g(e)}|y_{D(e)}), F(y_{k(e)}|y_{D(e)}))\Big] \tag{6.36}$$

on the right-hand side of Equation 6.36, the right factor is a result of $m(m-1)/2$ bivariate copula densities. The log-likelihood function of the regular vine copula with parameter $\Theta_{RV}$ and $E_1, ..., E_{m-1}$ is written as [138]:

$$\ell_{RV}(\Theta_{RV}|\nu) = \sum_{k=1}^{N} \sum_{j=1}^{m-1} \sum_{e \in E_j} \log\Big[C_{g(e),k(e)|D(e)}\Big(F(u_{j,g(e)}|u_{j,D(e)})|\Theta_{g(e),k(e)|D(e)}\Big)\Big] \tag{6.37}$$

where $u_j = (u_j, 1, ..., u_j, m)' \in [0,1]^m, j = 1, ..., N$. $C_{g(e),k(e)|D(e)}$ is a bivariate copula density with parameter $\Theta_{g(e),k(e)|D(e)}$ and edge $e$

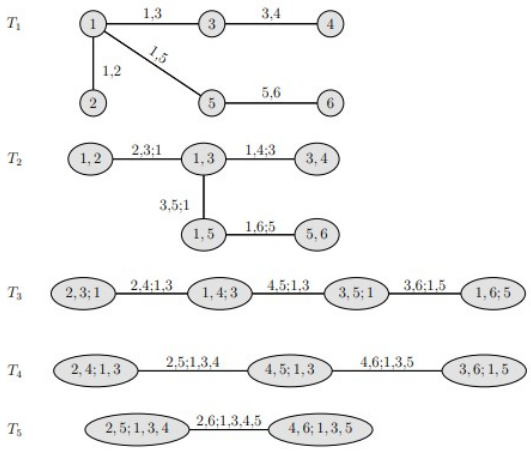


**Figure 6.5:** Six dimensional regular vine tree structure [15]

Explaining complete union, conditioned, and conditioning set, we make use of the edge $e = \{\{1,2\},\{1,5\}\}$ of $T_1$ in Figure 6.5. The complete union in $T_1$ included in $e$ is $\{1,2,5\}$. The conditioning set and conditioned set are $D(e) = \{1\}$ and $\{2,5\}$ respectively. Special cases of regular vine copula are canonical (C-Vine) and drawable (D-Vine) copula. For more details and properties of R-Vine copulas see [137] and [15]

### 6.3.2 Canonical Vine (C-Vine)

**Definition 6.3.2**: A regular vine tree structure $V = (T_1, ..., T_{m-1})$ is called a canonical vine (C-Vine) if in each tree $T_j$ there exists one unique node $n$ such that it has degree $m-j$. i.e., this unique node is connected to $m-j$ edges, and it is called the root node of the tree $T_j$ [15]. The m-dimensional density of a C-Vine is given by [132]:

$$\prod_{k=1}^{m} f(y_k) \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{j,j+1|1,...,j-1}\{F(y_j|y_1,...,y_{j-1}), F(y_{j+i}|y_1,...,y_{j-1})\} \tag{6.38}$$

The log-likelihood function of the C-Vine is given by

$$\sum_{j=1}^{m-1} \sum_{i=1}^{m-j} \sum_{t=1}^{T} \log\Big[c_{j,j+1|1,...,j-1}\{F(y_{j,t}|y_{1,t},...,y_{j-1,t}), F(y_{j+i,t}|y_{1,t},...,y_{j-1,t})\}\Big] \tag{6.39}$$

Equation 6.33 and Equation 6.35 are used to calculated the conditional distributions $F(y_{j,t}|y_{1,t},...,y_{j-1,t})$ and $F(y_{j+i,t}|y_{1,t},...,y_{j-1,t})$. If the C-Vine has m variables, then we

have m!/2 different canonical vines on m nodes [132]. Employing the canonical vine could be beneficial if a specific variable is identified as the main driver of interactions within the data. In such scenarios, this variable could be positioned at the core of the C-Vine [136].

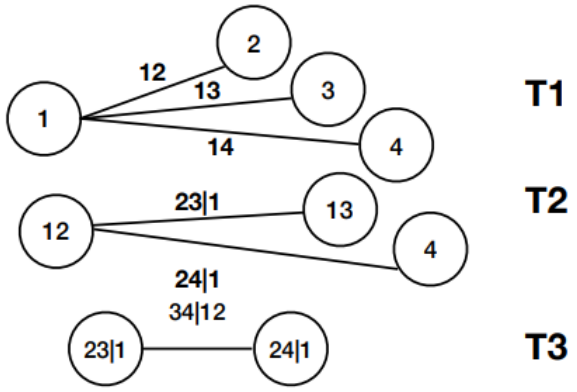


**Figure 6.6:** Four-dimensional Canonical Vine [16]

The four-dimensional C-Vine in Figure 6.6 is expressed as

$$\begin{aligned} f(y_1,y_2,y_3,y_4) &= f_1(y_1)\cdot f_2(y_2)\cdot f_3(y_3)\cdot f_4(y_4) \\ &\quad \cdot c_{12}\{F_1(y_1),F_2(y_2)\}\cdot c_{13}\{F_1(y_1),F_3(y_3)\}\cdot c_{14}\{F_1(y_1),F_4(y_4)\} \\ &\quad c_{23|1}\{F(y_2|y_1),F(y_3|y_1)\}\cdot c_{24|1}\{F(y_2|y_1),F(y_4|y_1)\} \\ &\quad \cdot c_{34|12}\{F(y_3|y_1,y_2),F(y_4|y_1,y_2)\} \end{aligned} \tag{6.40}$$

### 6.3.3 Drawable Vine (D-Vine)

**Definition 6.3.3**: A regular vine tree structure $V=(T_1,...,T_{m-1})$ is called a drawable vine (D-Vine) if each tree $T_j$ has a degree less or equal to two. i.e. no node is connected to more than two edges. The decomposition is determined by the $m(m-1)/2$ edges and the marginal densities of each variable. In $T_{j+1}$, edges from $T_j$ transform into nodes. If they have a mutual node in $T_j$, they are connected by an edge in $T_{j+1}$. The m-dimensional

density of a D-Vine is given by [132]:

$$\prod_{k=1}^{m} f(y_k) \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{i,i+j|i+1,\ldots,i+j-1}\{F(y_i|y_{i+1},\ldots,y_{i+j-1}), F(y_{i+j}|y_{i+1},\ldots,y_{i+j-1})\} \quad (6.41)$$

The log-likelihood function of the D-Vine is given by

$$\sum_{j=1}^{m-1} \sum_{i=1}^{m-j} \sum_{t=1}^{T} \log \left[ c_{i,i+j|i+1,\ldots,i+j-1}\{F(y_{i,t}|y_{i+1,t},\ldots,y_{i+j-1,t}), F(y_{i+j,t}|y_{i+1,t},\ldots,y_{i+j-1,t})\} \right] \quad (6.42)$$

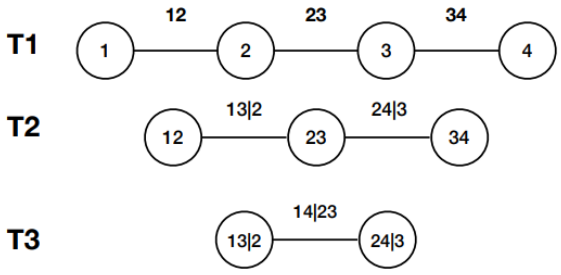


**Figure 6.7:** Four-dimensional D-Vine [16]

Figure 6.7 shows a four-dimensional D-Vine structure. The m-dimensional D-Vine has $m-1$ trees. Tree $T_j$ has $(m+1-j)$ nodes and $(m-j)$ edges. Each edge corresponds to a pair-copula, e.g., in tree 2 edge 24|3 corresponds to the pair-copula $C_{24|3}$. The four-dimensional D-Vine is expressed as

$$\begin{aligned} f(y_1,y_2,y_3,y_4) \;=\; & f_1(y_1)\cdot f_2(y_2)\cdot f_3(y_3)\cdot f_4(y_4) \\ & \cdot c_{12}\{F_1(y_1),F_2(y_2)\}\cdot c_{23}\{F_2(y_2),F_3(y_3)\}\cdot c_{34}\{F_3(y_3),F_4(y_4)\} \\ & c_{13|2}\{F(y_1|y_2),F(y_3|y_2)\}\cdot c_{24|3}\{F(y_2|y_3),F(y_4|y_3)\} \\ & \cdot c_{14|23}\{F(y_1|y_2,y_3),F(y_4|y_2,y_3)\} \end{aligned} \quad (6.43)$$

### 6.3.4 Selection of R-Vine Model

This section describes a systematic procedure to specify an R-Vine copula model, leveraging the empirical Kendall's tau. Given a set i.i.d random vectors, the method identifies the optimal spanning tree by maximizing the sum of absolute empirical Kendall's taus.

This optimization is essential for capturing the intricate dependencies between variables. A suitable copula is selected for every edge within this spanning tree using the smallest AIC, and its parameters are estimated. The algorithm then extends its approach to conditional settings, ensuring a comprehensive model fit across all variable interactions. Utilizing this structured method ensures a robust and nuanced understanding of the dependencies within the data.

**Algorithm 1:** Choosing an R-Vine model using Kendall's tau [137]

**Input:** Data points $(y_{l1}, \ldots, y_{ln})$, where $l = 1, \ldots, N$ (independent and identically distributed random vectors)

**Output:** Specification of an R-Vine copula.

1 Determine the empirical Kendall's tau, $\hat{\tau}_{i,j}$ for each unique pair of variables $\{k, l\}$ where $1 \leq i < j \leq n$.

2 Choose the spanning tree that optimizes the absolute value of the $\tau$ (weight) $\max \sum_{e=\{i,j\} \text{in spanning tree}} |\hat{\tau}_{i,j}|$

3 Select a copula for every edge $i, j$ within the chosen spanning tree and determine its associated parameter(s). Then Transform $\hat{F}_{i|j}(y_{li}|y_{lj})$ and $\hat{F}_{j|i}(y_{lj}|y_{li})$, $l = 1, .., N$, utilizing the estimated copula model $\hat{C}_{ij}$.

4 **for** $k = 2$ **to** $d - 1$ **do**

5 Compute the $\hat{\tau}_{i,j|D}$, for every pair of variables $\{i, j|D\}$ eligible to be in the tree $T_i$, specifically those pairs that meet the proximity criteria.

6 Choose the spanning tree from these edges that optimizes the cumulative absolute value of empirical Kendall's taus. $\max \sum_{e=\{i,j|D\} \text{in spanning tree}} |\hat{\tau}_{i,j|D}|$

7 Select a conditional copula and determine its associated parameters for every edge, $\{j, k|D\}$ within the chosen spanning tree. Then Transform $\hat{F}_{i|j \cup D}(y_{li}|y_{lj}, y_{lD})$ and $\hat{F}_{j|i \cup D}(y_{lj}|y_{li}, y_{lD})$, $l = 1, .., N$, utilizing the estimated copula model $\hat{C}_{ij}$.

8 **end**

### 6.3.5 Dependence Results (DIMV)

Applying the Vine copula methodology described above and implementing Algorithm 1 using the *RVineStructureSelect* function (VineCopula package) in R. The D-Vine Copula was selected as the best fit Vine Copula structure for the dependence modelling of duration, intensity, maximum intensity and volatility (DIMV), having the least AIC value compared to the R-Vine and C-Vine and the results are given in Table 6.4. The copula

pairs plot is given in Figure 6.8 and the D-Vine tree structure for DIMV is given in Figure 6.9. The R code for the implementation is given in Appendix D.2

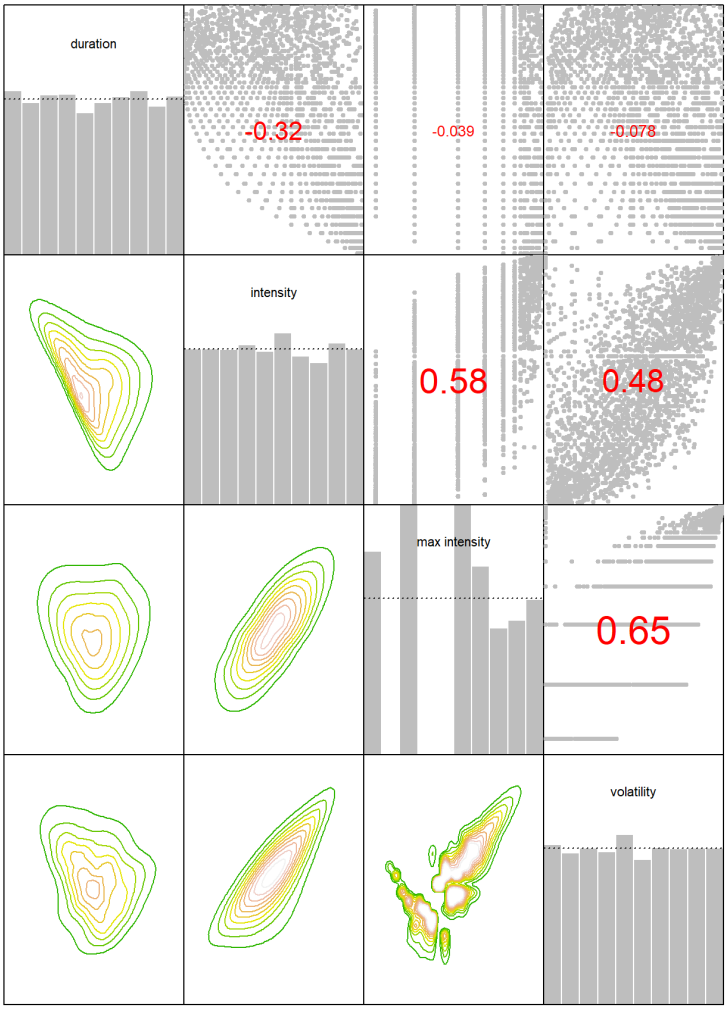


**Figure 6.8:** Pairs copula plot for DIMV

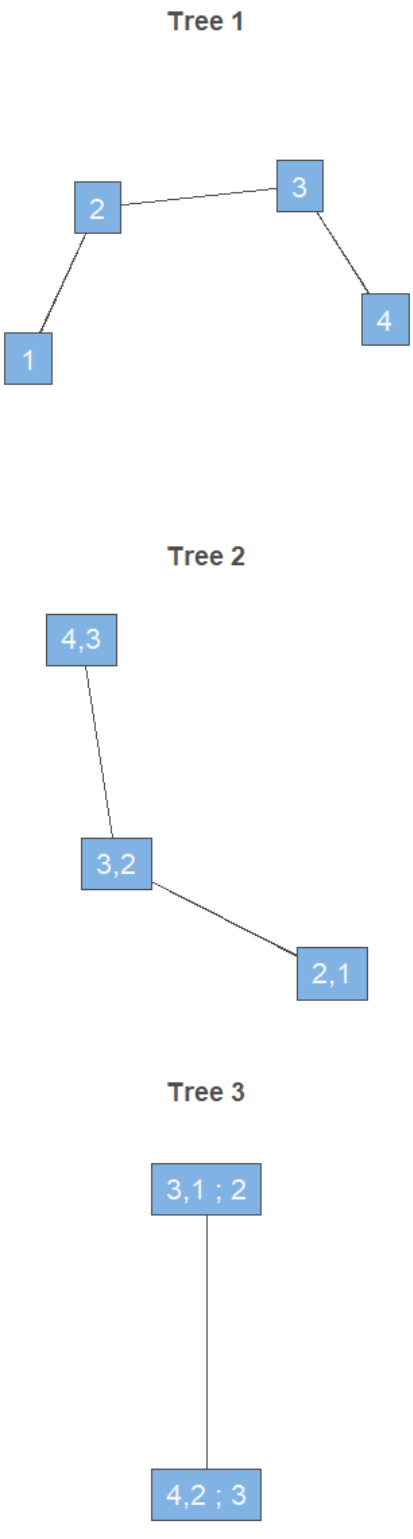


**Figure 6.9:** D-Vine Tree Plot for DIMV

**Table 6.4:** D-Vine copula with pair-copulas

| Tree | Edge | Pair-Copula | $\theta_1$ | $\theta_2$ | $\tau$ | $\lambda_L$ | $\lambda_U$ |
|---|---|---|---|---|---|---|---|
| 1 | 2,1 | Rotated Tawn type 2 270° | -4.69 | 0.32 | -0.29 | - | - |
| 1 | 3,2 | Gaussian | 0.76 | - | 0.55 | - | - |
| 1 | 4,3 | Tawn type 2 | 3.17 | 0.85 | 0.6 | - | 0.69 |
| 2 | 3,1;2 | Rotated BB1 180° | 0.10 | 1.42 | 0.33 | 0.37 | 0.01 |
| 2 | 4,2;3 | Tawn type 2 | 2.15 | 0.28 | 0.21 | - | 0.25 |
| 3 | 4,1;3,2 | Rotated BB8 270° | -3.11 | -0.54 | -0.20 | - | - |

1 ↔ Duration; 2 ↔ Intensity; 3 ↔ Maximum Intensity; 4 ↔ Volatility

The Table 6.4 presents the resulting D-Vine copula structure derived from the earlier discussed algorithm. An interesting observation from the table is the presence of multiple types of copulas, reflecting the complex interplay of dependencies among variables. The negative dependence in specific pairs indicates that as one variable decreases, the other tends to increase, and vice versa. Meanwhile, the positive tau values represent direct correlations.

**Table 6.5:** Dependence ($\tau$) table for copula simulated data

| | Duration | Intensity | Max Intensity | Volatility |
|---|---|---|---|---|
| Duration | 1.00000000 | -0.3527273 | -0.09252525 | -0.1321212 |
| Intensity | -0.35272727 | 1.0000000 | 0.56040404 | 0.3866667 |
| Max Intensity | -0.09252525 | 0.5604040 | 1.00000000 | 0.5959596 |
| Volatility | -0.13212121 | 0.3866667 | 0.59595960 | 1.0000000 |

**Table 6.6:** Dependence ($\tau$) table for observed data

| | Duration | Intensity | Max Intensity | Volatility |
|---|---|---|---|---|
| Duration | 1.00000000 | -0.3244918 | -0.03858472 | -0.07826699 |
| Intensity | -0.32449177 | 1.0000000 | 0.58241328 | 0.48054569 |
| Max Intensity | -0.03858472 | 0.5824133 | 1.00000000 | 0.64976489 |
| Volatility | -0.07826699 | 0.4805457 | 0.64976489 | 1.00000000 |

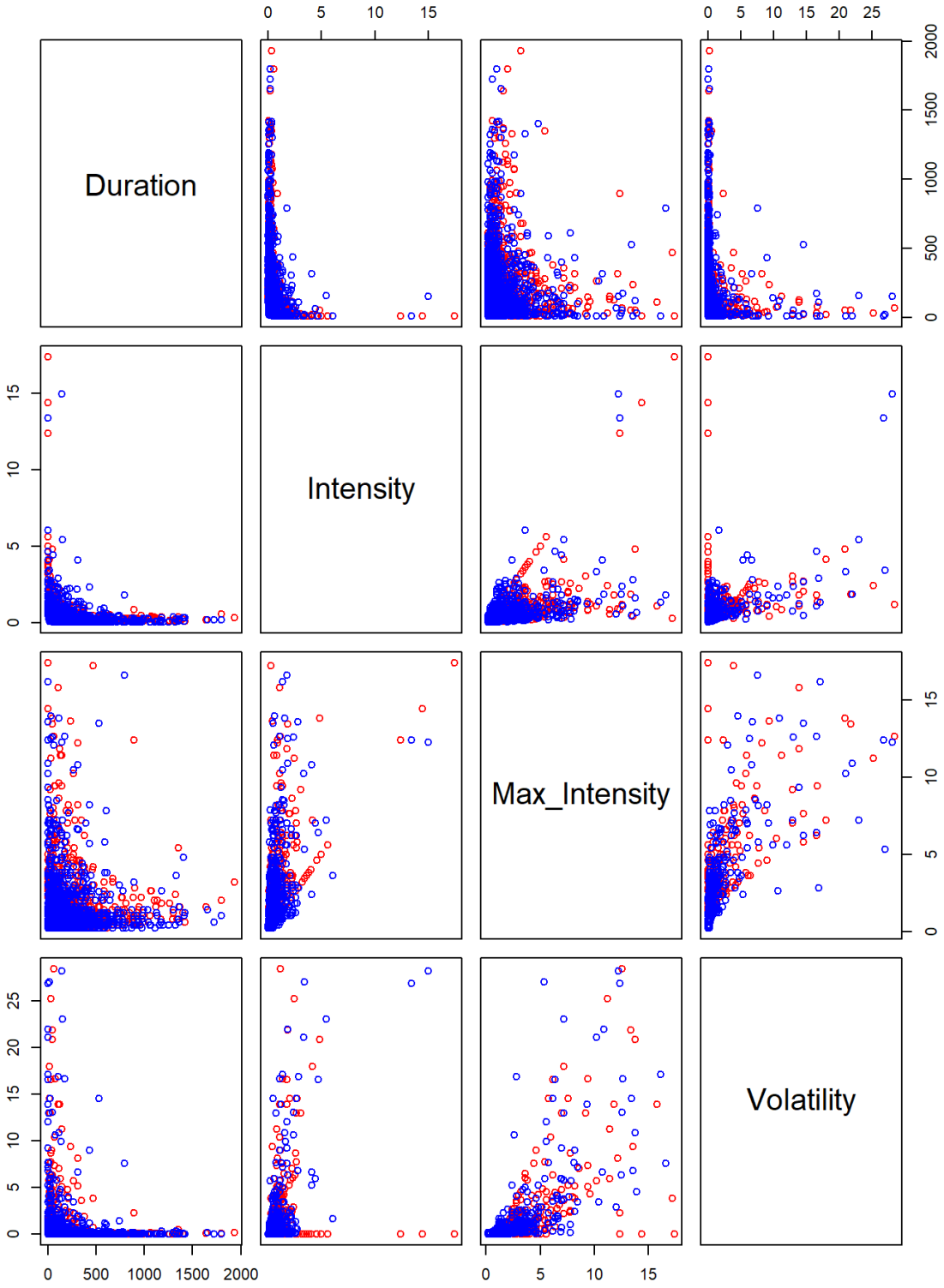


**Figure 6.10:** Pairs plot of copula simulated data (blue) and observed data (red)

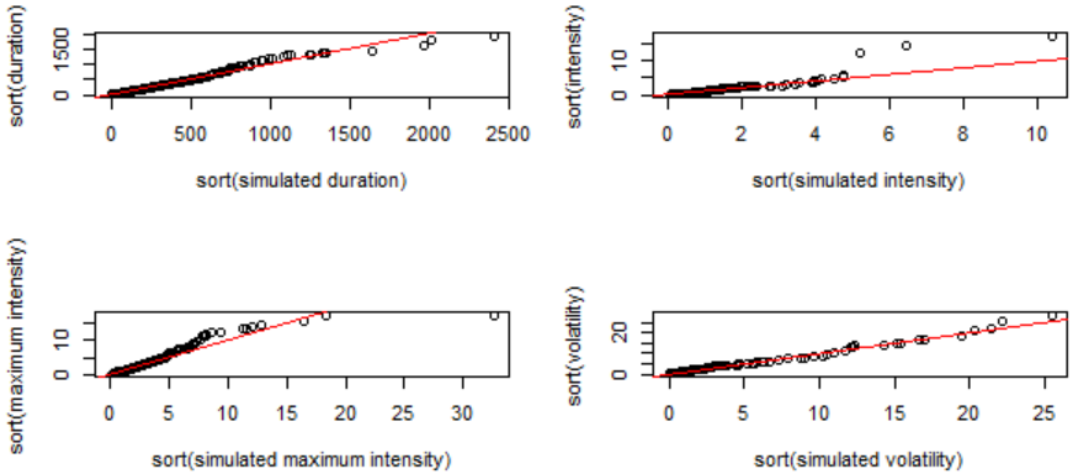


**Figure 6.11:** Q-Q plot of copula simulated data and observed data

To study the accuracy of the fitted vine copula model. The quadruple (DIMV) was simulated using the selected vine copula, and the appropriate transformations were carried out as defined in Chapter 3 for each variable. We calculated Kendall's tau for each pair of D-Vine copula simulated sample variables to see if the proposed vine copula can keep the sample dependencies (DIMV) among the rain event characteristics. We compared the results with Kendall's tau for each pair of variables of the observed DIMV. The results are given in Table 6.5 and Table 6.6. From the result, the D-Vine copula can preserve the sample dependencies, as shown in Figure 6.8. Figure 6.11 shows that the Q-Q plot of the copula simulated data resembles that of the observed data, giving a promising alignment between theoretical expectations and empirical observation.

# Chapter 7

## Irregular Pulse Model (Intensity-Duration-Maximum-Volatility)

## 7.1 Introduction

Stochastic generation models can be broadly categorized into profile-based and pulse-based. Profile-based models concentrate on individual rainfall events, typically defining them through inter-event time and leveraging joint or individual statistical distributions to delineate the main storm features. This accumulated rainfall is splited into distinct depth values at defined time steps. On the other hand, pulse-based models view rain events as events distributed randomly over time, following a Poisson distribution. Each storm is seen as a cluster of rain cells, with each cell being a pulse with a random length and consistent intensity. These cells are distributed over time based on models like Neyman-Scott or Bartlett-Lewis [20]. Due to their proficiency in mirroring ongoing rainfall sequences, they have numerous applications in hydrological studies [4]. However, their implementation requires estimating numerous parameters and a substantial historical rainfall dataset in a continuous format. Cameron et al. [36] pointed out that while these models adeptly mimic observed gaps between rainfalls across various scales, they struggle with accuracy when simulating extreme statistics at shorter durations. These models can not reproduce the long-term rainfall event data needed in reality [139]. As a result, many experts have proposed the use of profile-based models. This chapter presents a novel stochastic rainfall event simulator that reproduces observed rainfall events using its variables: duration, intensity, maximum intensity and volatility.

## 7.2 Simulating Rainfall using IAT

Suppose $X_i$ is the inter-arrival time (IAT) between $i^{th}$ and $(i+1)^{th}$ rainfall event. Then we model $X_i \sim 1 + E_i$, where $E_i \sim$ Exponential distribution with rate parameter $\lambda_i$. $\lambda_i = \lambda(t_i)$ where $t_i$ is the end time of the $i^{th}$ event. For rv's $X_1, ..., X_n$ with joint density $f(x_1, ..., x_n)$ then, $\log L = \log f(x_1, ..., x_n)$. If $X_i$ are independent, we have,

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_i(x_i) \tag{7.1}$$

$$\log L(\theta; x) = \sum_{i=1}^{n} \log f_i(\theta; x_i) \tag{7.2}$$

Since $X_i - 1$ follows an exponential distribution with rate parameter $\lambda_i$

$$f(x_i - 1; \lambda_\theta(t_i)) = \lambda(t_i) e^{-\lambda(t_i)(x_i - 1)} \tag{7.3}$$

$$\log f(x_i - 1; \lambda_\theta(t_i)) = \log \lambda(t_i) e^{-\lambda(t_i)(x_i - 1)} \tag{7.4}$$

$$\log f(x_i - 1; \lambda_\theta(t_i)) = \log \lambda(t_i) - \lambda(t_i)(x_i - 1) \tag{7.5}$$

$$l(\lambda_\theta(t_i); x_i - 1) = \sum_{i=1}^{n} \Big[ \log \lambda(t_i) - \lambda(t_i)(x_i - 1) \Big] \tag{7.6}$$

$$l(\lambda_\theta(t_i); x_i - 1) = n \log \lambda(t_i) - \lambda(t_i) \sum_{i=1}^{n} (x_i - 1) \tag{7.7}$$

We explore several ways for a parametric model for $\lambda$. In terms of notation, we will summarise the involved parameter on $\theta$.

$$\lambda_\theta(t) = \alpha_1 \sin\left(2\pi \frac{t}{365}\right) + \beta_1 \cos\left(2\pi \frac{t}{365}\right) + \alpha_2 \sin\left(4\pi \frac{t}{365}\right) + \beta_2 \cos\left(4\pi \frac{t}{365}\right) \tag{7.8}$$

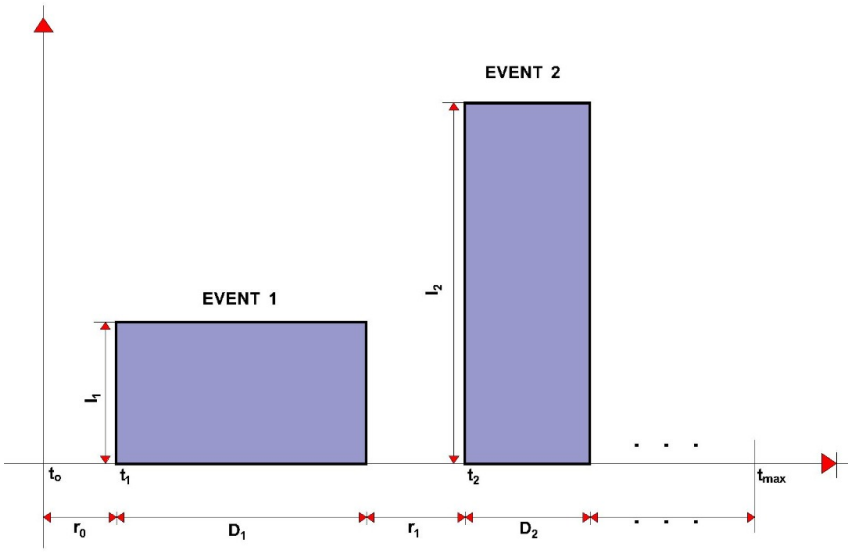


**Figure 7.1:** Rainfall events using inter-arrival time (IAT). Where start time $= t_0$, end time $= t_{max}$, intensity $= I$, duration $= D$, and interarrival time $= r_i$

## 7.3 Simulating Rain Depth (Intensity) (given D, I,M, V)

Given intensity, duration, max intensity, volatility $(I, D, M, V)$, $h =$ time step, the goal is to create a rainfall event as given by a sequence $(x_i), i = 1, ..., n$, where $n = \lceil \frac{d}{h} \rceil$ which meets the following constraints,

$$I = \frac{1}{n} \sum_i x_i \tag{7.9}$$

$$M = \max_{1 \leq i \leq N} x_i \tag{7.10}$$

$$V = \frac{1}{n} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 \tag{7.11}$$

$$x_i \geq 0 \tag{7.12}$$

at least approximately with some numerical precision. Therefore, we propose the following algorithm given by the following pseudo-code

## 7.4 Proposed algorithm for the rainfall event simulator

**Algorithm 2:** Simulating a Rainfall Event (Pseudocode) - Part 1

**Input:** intensity (I), duration (d), maximum intensity (M), volatility (V), stepsize, $tol$, $\alpha$, $\beta$, maximum iteration (max_it), $j$, num_calls

**Output:** sequence of rain depths $(x_i), i = 1, ..., n$ where $n = \lceil \frac{d}{\text{stepsize}} \rceil$

```
Set x = array of size n
if num_calls ≥ 2 then
    return (x, "failure")
end
if n = 1 then
    if |M − n × I| ≥ tol then
        return (x, "raineventsim: inconsistent I, D, M, V, n == 1")
    end
    else
        x[1] ← M
        return (x, "success")
    end
end
if n = 2 then
    if random number < 0.5 then
        x ← (M, 2I − M)
    end
    else
        x ← (2I − M, M)
    end
end
if Vol(x) ≈ V and min(x) ≥ 0 then
    return (x, "success")
end
else
    return (x, "raineventsim: inconsistent I, D, M, V, n == 2")
end
```

**Algorithm 3:** Simulating a Rainfall Event (Pseudocode) - Part 2

```
if j = −1 then
    j ← ⌈random number from Beta(α, β) × n⌉
end
x[j] ← M
for i in 1, ..., n but not j do
    x[i] ← (n×I−M)/(n−1)
end
if min(x) < 0 or max(x) > M then
    return (x, "inconsistent I, D, M, V, n ≥ 3")
end
num_reps ← 0
fail_count ← 0
while fail_count < max_it do
    V_err ← |Vol(x) − V|
    if V_err ≤ tol then
        return (x, "success")
    end
    Pick two random indices h and i excluding j
    Compute perturbations for the values at these indices
    if either perturbation improves V_err then
        Update x with better perturbation
        Reset fail_count
    end
    else
        Increment fail_count
    end
end
if j ≠ 1 and j ≠ n then
    j ← random choice from {1, n}
end
else
    j ← random choice from {2, ..., n − 1}
end
Call raineventsim with new j and incremented num_calls
return (x, "raineventsim: maximum iterations exceeded")
```

## 7.5 Implementation in R

Building on the algorithmic foundation in Sections 7.3 and 7.4, this section delves into the actual R implementation of the rainfall event simulator. Each snippet of R code corresponds to critical segments of the proposed algorithm, elucidating how theoretical constructs translate into practical, executable code.

### 7.5.1 Handling Single-Step Rainfall Events

Single-step events represent the simplest form of rainfall events, where the event duration is equal to the step size. In such cases, the simulator directly assigns the maximum intensity ($M$) to the event, ensuring consistency with the specified average intensity ($I$) within a defined tolerance ($tol$).

```r
# Handle case where duration consists of only one step
if (n == 1) {
  # Check if maximum intensity matches expected average intensity within tolerance
  if (abs(M - (n * I)) >= tol) {
    # Return failure state if parameters are inconsistent
    return(list(x = x, state = "raineventsim: inconsistent I, D, M, V, n == 1"))
  } else {
    # Assign maximum intensity to the event and return success
    x[1] = M
    return(list(x = x, state = "success"))
  }
}
```

**Listing 7.1:** Single-Step Events

The code handles the scenario where the rainfall event duration consists of only a single time step. This is a particular case in the simulation process. The function checks whether the maximum intensity ($M$) matches the expected average intensity ($I$), given that there is only one step in the event. If the absolute difference between M and the product of the number of steps ($n$) and average intensity ($I$) is within a specified tolerance ($tol$), it is

considered a successful match, and the single intensity value ($M$) is assigned to the event. Otherwise, it indicates an inconsistency between the input parameters for a single-step event.

### 7.5.2 Addressing Two-Step Events

In scenarios where the event spans two steps, the simulator employs a randomized approach to distribute intensity, ensuring the overall average intensity aligns with the input while meeting volatility constraints.

```
# Handle case for events with two steps
if (n == 2) {
  # Randomly decide the order of intensities to match the
   average intensity requirement
  x <- ifelse(runif(1) < 0.5, c(M, 2 * I - M), c(2 * I - M, M)
   )
  # Validate if adjusted intensities result in correct
   volatility within tolerance
  if (abs(Vol(x) - V) <= tol && min(x) >= 0) {
    # Return success if conditions are met
    return(list(x = x, state = "success"))
  } else {
    # Return failure state if conditions are not met
    return(list(x = x, state = "raineventsim: inconsistent I,
   D, M, V, n == 2"))
  }
}
```

**Listing 7.2:** Two-Step Events

In the unique scenario of two-step rainfall events, this code segment addresses the challenge of appropriately allocating the event's total intensity across both steps. The aim is to ensure that the computed average intensity accurately reflects the predefined input value for average intensity ($I$). To achieve this, the algorithm introduces a measure of randomness in selecting the placement of the maximum intensity value ($M$) in the initial or concluding step. Subsequently, it recalibrates the intensity of the remaining step to preserve the overall average intensity as specified. This process includes a meticulous

verification step to confirm that the modified intensity values sum up correctly while maintaining all intensities above zero and aligning the event's volatility with the specified tolerance range. When these conditions are satisfied, the configuration qualifies as successfully executed.

### 7.5.3 Positioning of Maximum Intensity

The position of the maximum intensity within the event is a pivotal aspect, influenced by the beta distribution parameters, alpha and beta, fostering variability in simulation outcomes.

```
# Determine the position of maximum intensity if not
   predefined
if (j == -1) {
  # Position is determined based on a beta distribution for
   variability
  j <- ceiling(rbeta(1, alpha, beta) * n)
}
# Assign maximum intensity to its position
x[j] <- M
```

**Listing 7.3:** Maximum Intensity Position

The variable $j$ determines the maximum intensity ($M$) positioning within the event. If $j$ is not predefined (i.e., -1), its position is decided stochastically based on a beta distribution characterized by parameters $\alpha$ and $\beta$. This probabilistic approach for determining $M$'s position introduces variability in the simulation, allowing for exploring diverse rainfall event patterns where the peak intensity can occur at different points within the event duration.

### 7.5.4 Enhancing Simulation Accuracy through Random Search

For events exceeding two steps, a random search algorithm fine-tunes intensity values across the event's duration, aiming for volatility that mirrors the input value as closely as possible.

```r
# Implement random search to minimize discrepancy in target volatility
while (fail_count < max_it) {
  # Calculate the current error in volatility
  V_err <- abs(Vol(x) - V)
  # Check if the current configuration meets the tolerance criteria
  if (V_err <= tol) return(list(x = x, state = "success"))

  # Randomly select two positions to adjust, excluding the position of M
  hi <- sample((1:n)[-j], 2)
  h <- hi[1]
  i <- hi[2]
  # Determine the maximum possible adjustment without violating constraints
  eps_max <- min(x[h], M - x[h], x[i], M - x[i])
  # Apply adjustment
  eps <- runif(1, -eps_max, eps_max)

  # Create two potential new configurations
  y <- x; y[h] <- y[h] - eps; y[i] <- y[i] + eps
  z <- x; z[h] <- z[h] + eps; z[i] <- z[i] - eps

  # Calculate new errors
  V_err_y <- abs(Vol(y) - V)
  V_err_z <- abs(Vol(z) - V)

  # Select the configuration that reduces the error most
  if (V_err_y < V_err || V_err_z < V_err) {
    fail_count <- 0  # Reset failure count on improvement
    x <- (V_err_y < V_err_z) ? y : z
  } else {
    # Increment failure count if no improvement
    fail_count <- fail_count + 1
  }
}
```

**Listing 7.4:** Random Search for Better Volatility Matching

This part of the code implements a random search algorithm to adjust the intensities within the event better to match the target volatility ($V$). By iteratively making small adjustments to a pair of randomly selected intensities, the algorithm seeks to minimize the discrepancy between the calculated volatility of the current event configuration and the target volatility. This process involves evaluating potential adjustments ($eps$) that would bring the event's volatility closer to the desired value, ensuring the event remains realistic by keeping intensities positive and within bounds.

The function is a sophisticated tool for simulating intricate rainfall events over discrete time intervals. Accepting inputs such as average intensity ($I$), event duration ($D$), maximum intensity ($M$), and volatility ($V$), along with optional parameters to control the simulation's precision, the function offers a hybrid approach that melds deterministic constraints with stochastic variability. At its core, the function is conditioned to handle different event lengths: for short durations (when $n = 1$ or 2), specific logic is applied, while for longer durations ($n \geq 3$), a random search algorithm is invoked to satisfy given constraints. This search algorithm dynamically adjusts the intensities of two random intervals to approximate the desired volatility. Suppose adjustments don't align with the target criteria. In that case, the algorithm iteratively refines its solution, using recursive calls to ensure the optimal placement of the maximum rainfall intensity within the simulated event. This systematic logic flow, combined with its balanced approach of deterministic adjustments and stochastic considerations, makes it a quintessential tool for hydrologists and climatologists seeking precision and realism in their rainfall simulations. The R code for the implementation is given in Appendix E.1.

## 7.6 Effects of Parameters

### 7.6.1 Alpha and Beta parameters

The $\alpha$ and $\beta$ parameters are pivotal in the simulator's approach to determining the position of maximum intensity ($M$) within a rainfall event. They influence the probability

distribution from which the position ($j$) is drawn, utilizing the Beta distribution as a foundational tool. The choice of $\alpha$ and $\beta$ values introduces a layer of flexibility, allowing the simulation to represent a wide range of natural rainfall patterns, from those with early peaks to those with late peaks in intensity.

(i) Setting $\alpha$ and $\beta$ to very high values and maintaining them constant effectively reduces the variability in the placement of $j$, making the simulation less random and more predictable. In extreme cases where $\alpha$ and $\beta$ are exceedingly high, this could approximate "fixing" the position of $j$ as the Beta distribution's variance diminishes, concentrating the probability mass around a specific point within the event's duration.

(ii) Setting both $\alpha$ and $\beta$ to 1 in the Beta distribution for determining the position of maximum intensity ($j$) in a rainfall event converts the distribution to a Uniform one, equalizing the likelihood of $j$'s placement across the event. While this maintains randomness, it removes the nuanced control over $j$'s placement offered by the Beta distribution with varied $\alpha$ and $\beta$ values, which can more closely mimic natural rainfall patterns by allowing for skewed or specific placements of peak intensities.

Therefore, to capture the variability and complexity of natural rainfall events accurately, it's crucial to avoid setting the $\alpha$ and $\beta$ parameters of the Beta distribution to excessively high values or to the uniformity of 1, ensuring a balanced approach that leverages the distribution's ability to offer nuanced randomness and precise control over peak intensity placement.

### 7.6.2 Tolerance (`tol`) and Maximum Iterations (`max_it`)

Tolerance (`tol`) and maximum iterations (`max_it`) control the simulator's precision and computational effort. The tolerance level sets the acceptable deviation from desired parameters, such as volatility, ensuring that simulated events closely match specified conditions. The maximum iterations parameter caps the number of attempts to adjust an event to meet these criteria, balancing accuracy with computational efficiency.

We undertake a computational experiment to investigate the impact of varying simulation tolerance (`tol`) levels on the runtime and volatility error (`V_err`) across a spectrum of rainfall events characterized by differing DIMV. For each specified event, the simulation is executed across multiple tolerance levels, ranging from relatively coarse (0.1) to extremely fine (0.00001). The tolerance parameter is crucial as it dictates the permissible deviation from the target volatility (`V`). The core of the analysis lies in measuring two primary outcomes for each simulation run: the runtime, indicative of computational efficiency, and the volatility error (`V_err`), reflective of the simulation's accuracy in achieving the desired event volatility. The runtime is measured in milliseconds, quantitatively measuring the simulator's performance efficiency, while `V_err` offers insight into the precision of the simulated rainfall event relative to the specified volatility target.

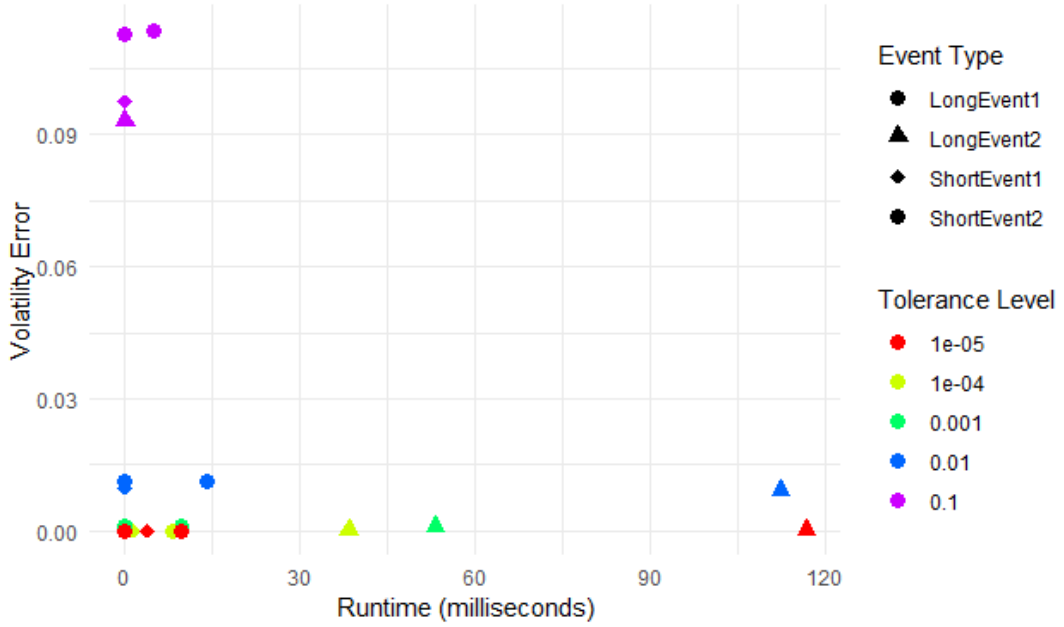


**Figure 7.2:** Simulation Runtime vs Volatility Error across Tolerance Levels and Events: Short Event1 = (I=0.4363636, D=66, M=0.8, V=0.07636364); Long Event1 = (I=0.9428571, D=378, M=6.2, V=1.881905); Short Event2 = (I=0.4, D=18, M=0.4, V=0); Long Event2 = (I=0.1919414, D=1638, M=1.6, V=0.07194139) )

From the results presented in Figure 7.2, simulations with higher tolerance levels, such as 0.1 and 0.01, exhibited notably higher volatility errors. This observation suggests that while a coarser tolerance may reduce computational complexity, it does so at the

expense of precision in achieving the target volatility (V). Conversely, lower tolerance levels, ranging from 0.001 to 0.00001, consistently resulted in a negligible volatility error, approaching zero. This delineates the efficacy of fine-grained tolerance settings in enhancing the simulator's precision, ensuring the simulated rainfall events closely align with the specified volatility parameters. Furthermore, examining the runtime across various event characteristics revealed that longer-duration events typically necessitated extended computational times. This trend highlights the inherent computational demands of simulating more complex or prolonged rainfall events due to the increased number of iterations required to converge on the specified event parameters within the defined tolerance levels.

### 7.6.3 Recursive Calls (`num_calls`)

The `num_calls` parameter manages the depth of recursive adjustments when optimizing event characteristics. It prevents infinite loops during the simulation process, ensuring the algorithm converges to a solution within a reasonable timeframe.

The interplay of these parameters within the rainfall event simulator allows for the generation of diverse and realistic rainfall events. By adjusting these parameters, we can tailor simulations to specific scenarios or investigate the impacts of different rainfall characteristics. This flexibility makes the simulator a valuable tool in practical applications in climatology and hydrology.

## 7.7 Comparison of the simulated event and the original rain event

To check for applicability, the simulator was tested using our 36 years of high-resolution rainfall events (DIMV), and the simulator reproduced realistic rainfall intensities for each case and its main characteristics. Figure 7.3 compares the actual event and the simulated events; from the plot, it can be seen that the simulator shows very satisfying results. We also explored different scenarios: (I) When the volatility is high with a short duration,

the maximum intensity often lies at the border. The simulator effectively reproduced this, reinforcing the simulator's adaptability to diverse event structures. (ii) When V=0, the depth is equally distributed along the duration, which is also reproduced by the simulator; this affirms its robustness in capturing and representing even linear rainfall patterns. (iii) event with long duration. In many of these events, the maximum intensity (M) tends to gravitate towards the central region, but this is more difficult to reproduce as the distribution of M has different possibilities. While the simulator demonstrated competence in approximating these central intensities, it is conceivable that more intricate elements influencing the placement of M were partially encapsulated. This opens the door for further refinement, where potential external features not yet integrated into our model could be explored for a more holistic simulation approach. Figure 7.4 gives the plot of the different scenarios considered. Our simulator is a potent tool for simulating realistic rainfall patterns, which will be helpful in the design of hydrological structures and urban planning.

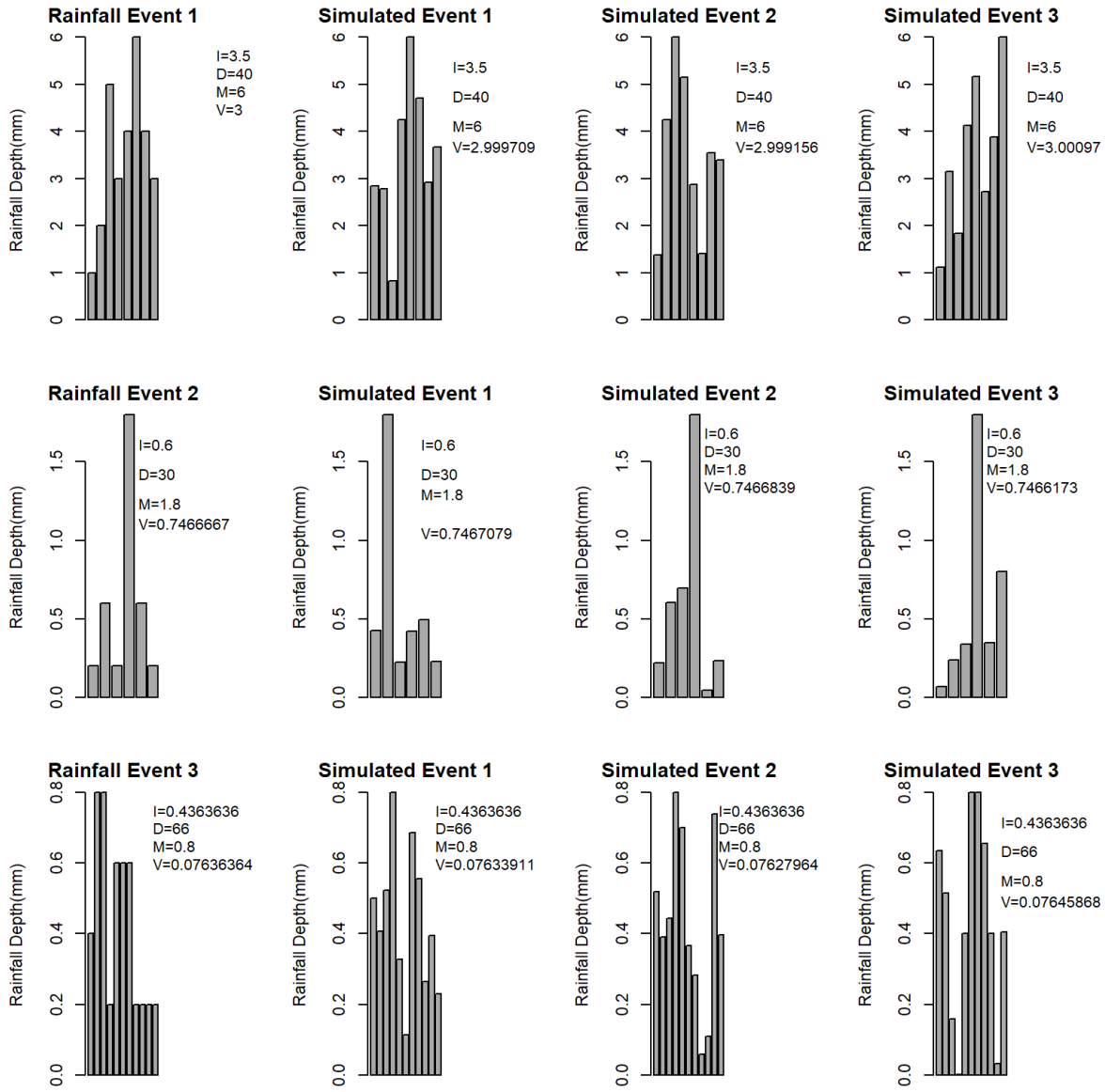


**Figure 7.3:** Original Rainfall Events (left) and Simulated Events with observed DIMV (columns 2 to 4)

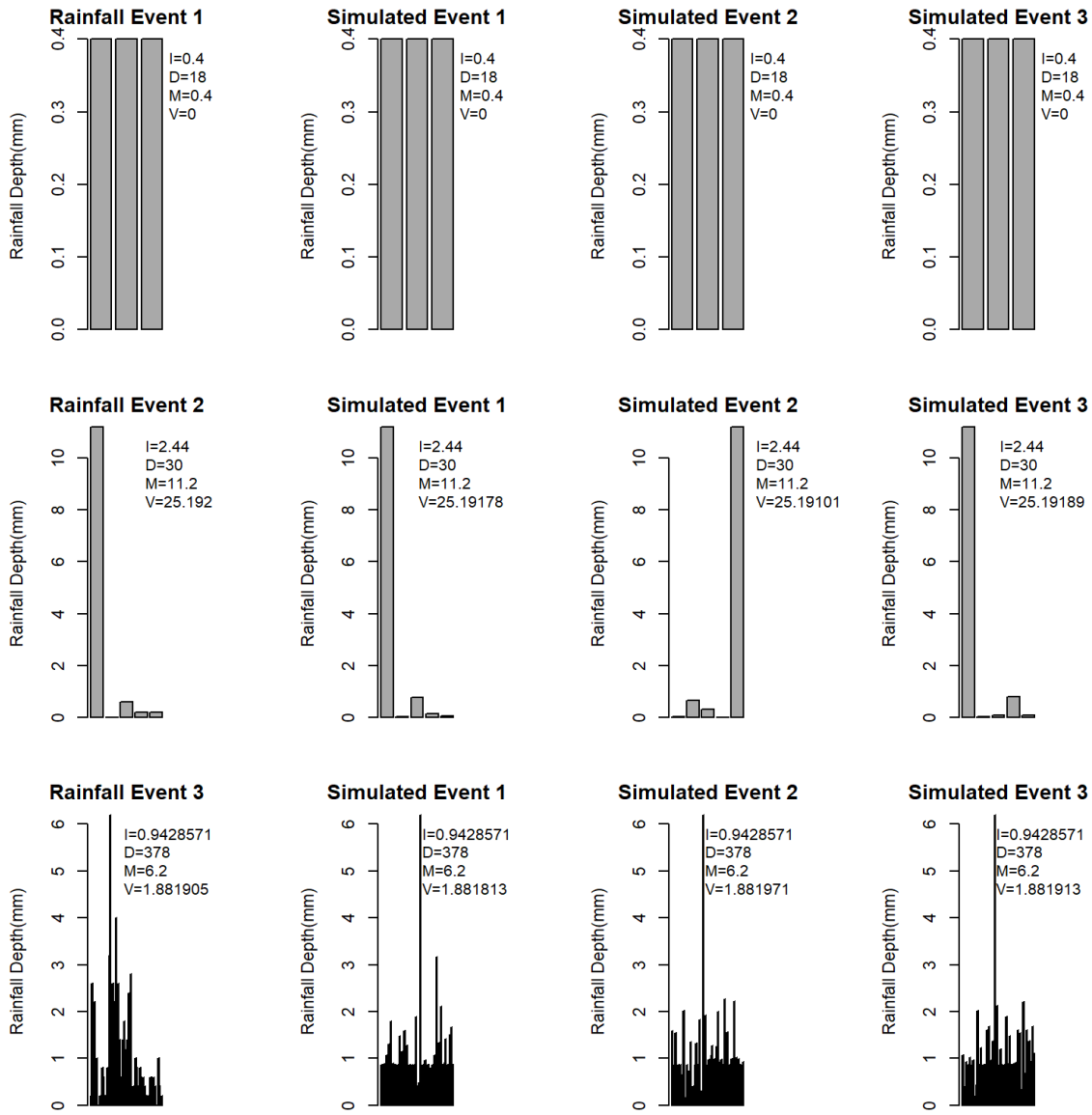


**Figure 7.4:** Illustration of how simulated events with given DIMV characteristics look like and resemble realistic events. Original Rainfall Events (left) and Simulated Events with identical DIMV characteristics (columns 2 to 4); different scenarios have been considered: Volatility $V = 0$ (top), High Volatility $V = 25.192$ (middle) and Long Event (Duration $D = 378$ minutes, bottom)

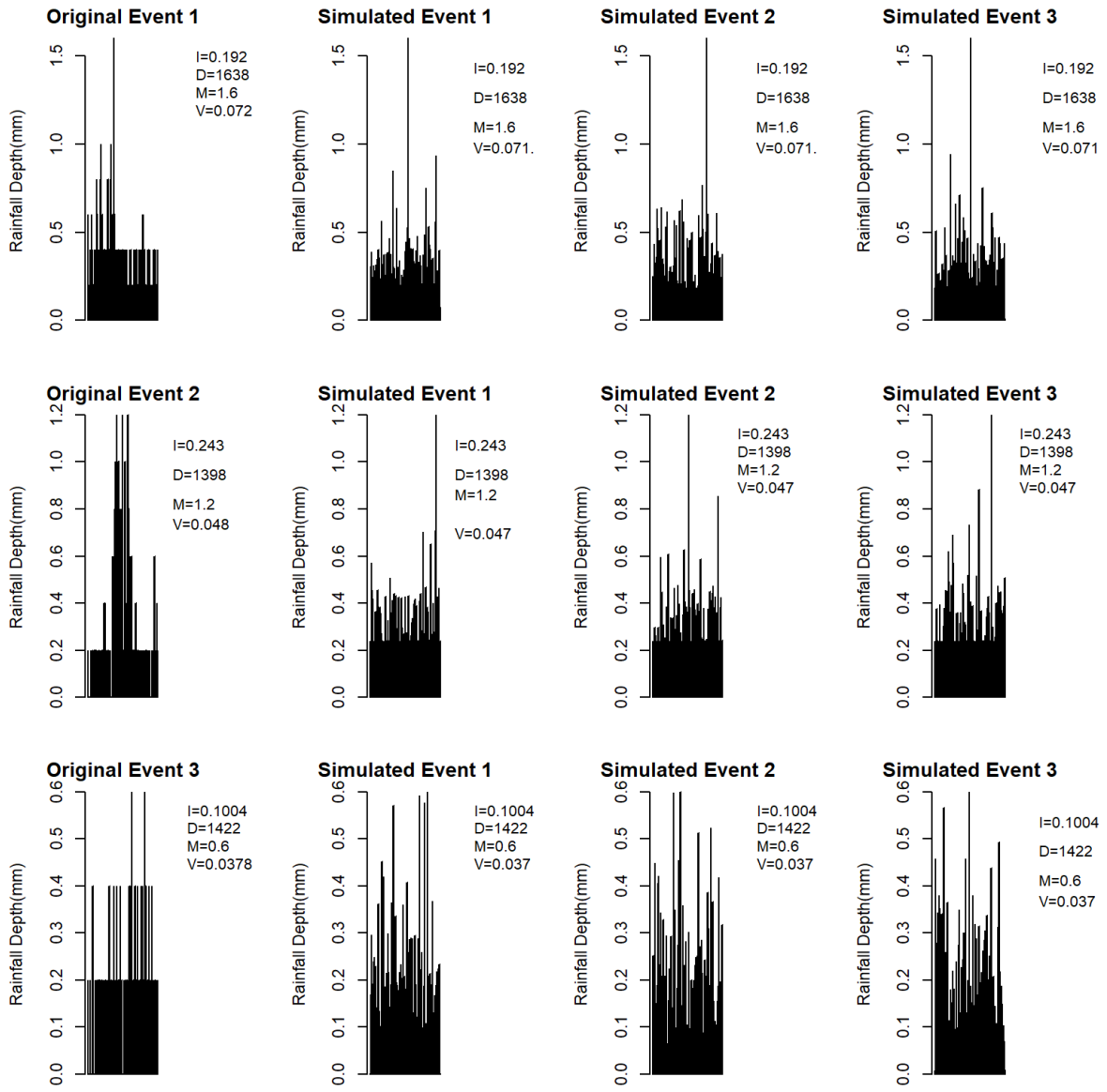


**Figure 7.5:** Illustration of how simulated events with given DIMV characteristics look like and resemble realistic events. Original Rainfall Events (left) and Simulated Events with identical DIMV characteristics (columns 2 to 4); different scenarios with long duration

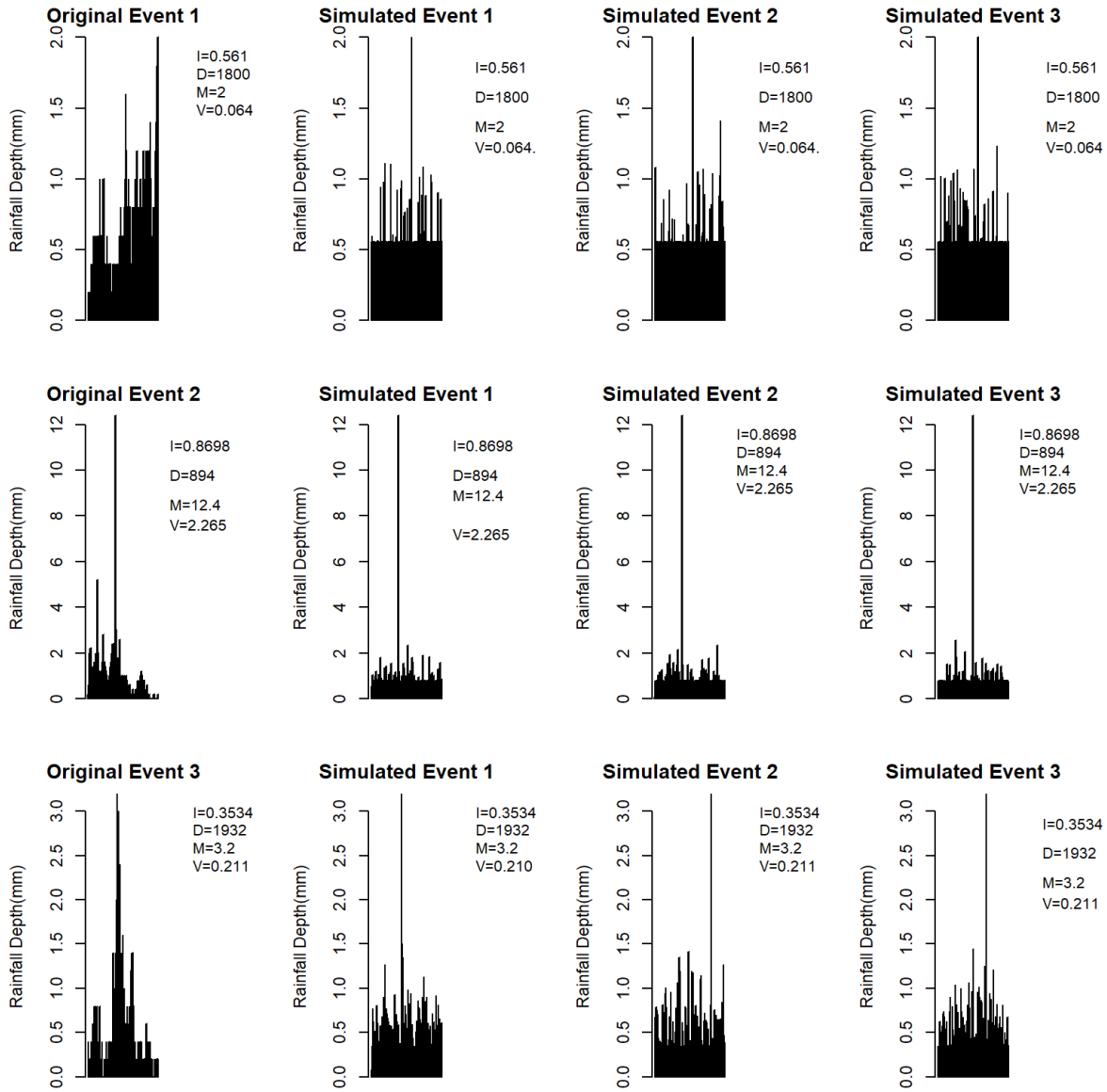


**Figure 7.6:** Illustration of how simulated events with given DIMV characteristics look like and resemble realistic events. Original Rainfall Events (left) and Simulated Events with identical DIMV characteristics (columns 2 to 4); different scenarios with large total rainfall (DxI)

## 7.8 Assessment of Model Robustness to Extreme Rainfall Scenarios

This section systematically evaluates our rainfall simulation model's resilience and accuracy in the face of hypothesized extreme weather conditions. Such an assessment is crucial for demonstrating the model's reliability for future climate variability and hydrological forecasting.

### 7.8.1 Scaling Factor Choice

The assessment of the model's robustness was tested using the last year's data. The selection of scaling factors for our analysis was meticulously considered to ensure a meaningful augmentation of rainfall data parameters. Specifically, average intensity (I) and maximum intensity (M) were scaled by a factor of 2, effectively doubling their values. In parallel, duration (D) and volatility (V) increased by 1.5. This strategic approach aimed to create conditions that simulate plausible extreme weather scenarios potentially arising from climate change. The underlying rationale for these choices was grounded in observed data trends and a concerted effort to balance realism and the imperative to test the model under challenging conditions rigorously.

Our methodology applied the aforementioned scaling factors to the observed rainfall event data last year. Following this modification, the simulation model was deployed to generate rainfall events based on the original (unscaled) and modified (scaled) datasets.

```
run_simulation <- function(df) {
  x_sim <- list()
  errors <- 0
  for (i in 1:nrow(df)) {
    c_I <- df$I[i]
    c_D <- df$D[i]
    c_M <- df$M[i]
    c_V <- df$V[i]

```

```
    x <- raineventsim2(I=c_I, D=c_D, M=c_M, V=c_V, max_it =
  100)
    if (x$state != "success") {
      errors <- errors + 1
      x_sim[[i]] <- NA
      cat(i, x$state, "\n")
    } else {
      x_sim[[i]] <- x$x
    }
  }
  list(simulated_data = x_sim, errors = errors)
}
```

**Listing 7.5:** Simulation code for unscaled and scaled data

### 7.8.2 Results

Visualization of the simulation outcomes, specifically through box plots in Figure 7.7, alongside Welch Two Sample t-tests conducted to compare the average intensities between scaled simulations and their corresponding original data, furnished valuable insights.

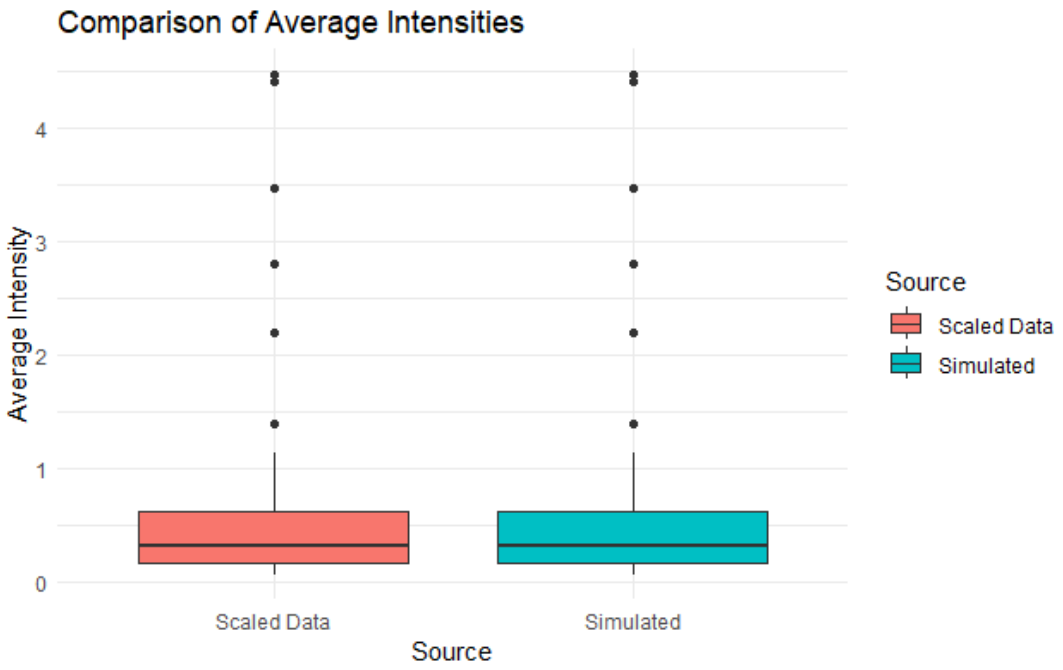


**Figure 7.7:** Box plot of intensity for the scaled data and the simulated intensity data obtained using our rainfall event simulator

```
t_test_result <- t.test(average_intensities_simulated, df_
    scaled$I)
print(t_test_result)

# Output
  Welch Two Sample t-test

data:  average_intensities_simulated and df_scaled$I
t = 0, df = 190, p-value = 1
alternative hypothesis: true difference in means is not equal
    to 0
95 percent confidence interval:
 -0.2200866  0.2200866
sample estimates:
mean of x mean of y
0.5601614 0.5601614
```

**Listing 7.6:** Welch Two Sample t-test results

The t-test result reinforces the model's accuracy, as it shows no significant difference in mean intensities between the simulated events and the scaled data, further validating the model's efficacy under simulated extreme conditions. The consistency between simulated and scaled intensities, as evidenced by the t-test and error-free simulation runs, compellingly illustrates the model's efficacy under extreme conditions. Successfully reproducing scaled intensities—a proxy for extreme weather scenarios—without deviation from statistical norms or introducing errors attests to the model's precision and reliability. This finding validates the model's utility in current climatic conditions and bolsters confidence in its application to future climate variability studies and hydrological forecasting endeavours.

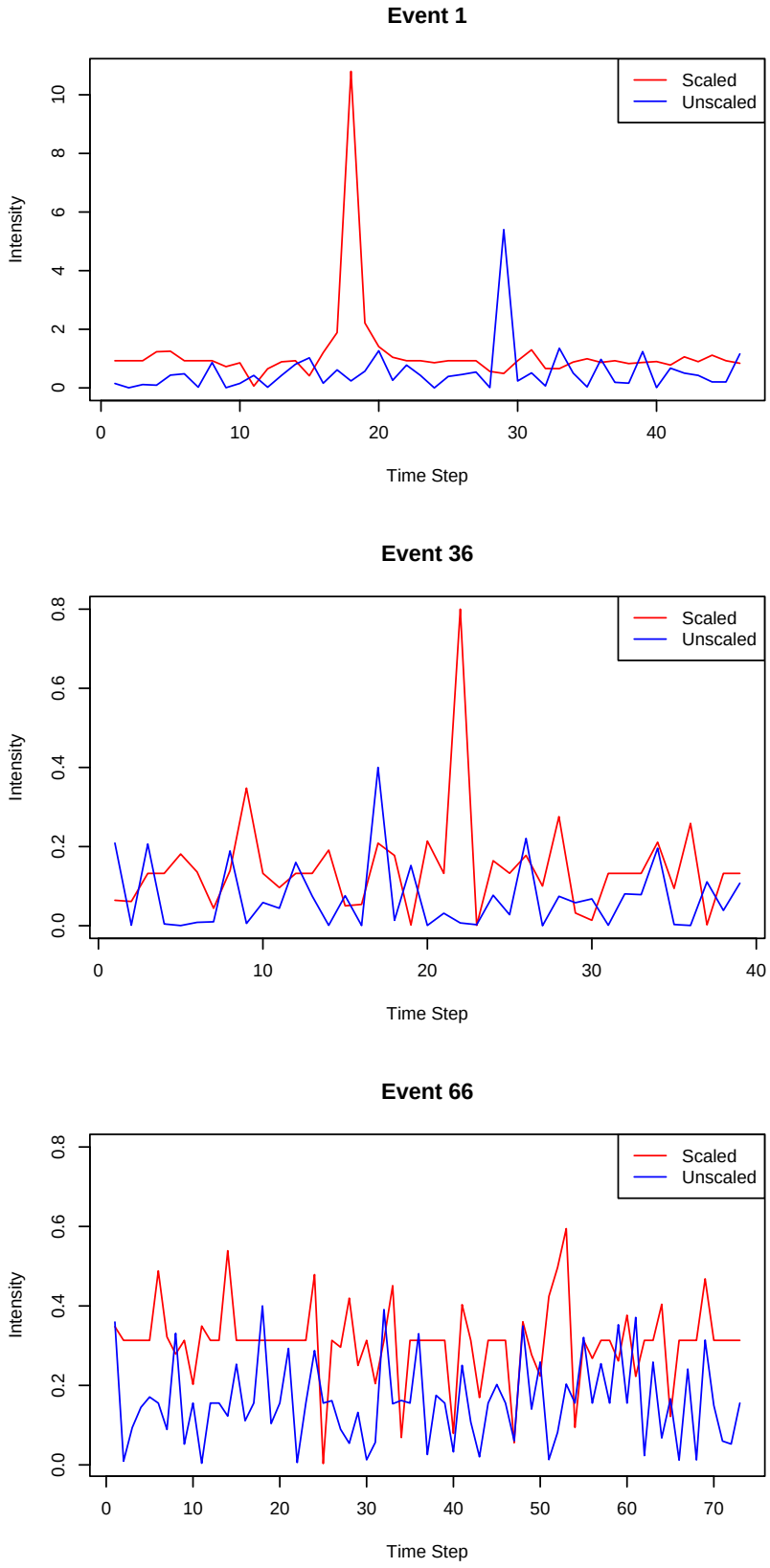


**Figure 7.8:** Simulation result for scaled and unscaled data

## 7.9 Simulating of Rainfall Events

In this section, we introduce an algorithm designed to reproduce a sequence of sub-hourly rain events, considering their inherent complexities and diverse temporal patterns. The proposed algorithm is given below:

### 7.9.1 Proposed Algorithm

**Algorithm 4:** Simulating a Sequence of Rainfall Events (Pseudocode)

**Input:** start time $t_1$, end time $t_{max}$, seasonal rate function $\lambda$, rate constant $c$, joint distribution model, time step $h$, raineventsim with turning parameters $\Theta$

**Output:** List of events $X^i = (x_1^i, ..., x_{n(i-j)}^i)$, event start time $s_i$, and event end time $t_i = n_{(i)}h + s_i$

Set $t_1 = h\lfloor t_1/h \rfloor$
Set $i = 1$
**while** $t_i < t_{max}$ **do**
    Generate duration $D_i$, intensity $I_i$, max intensity $M_i$, and volatility $V_i$ from joint distribution model
    Update $\bar{D}_i = n_i * h$
    Generate $r_i = c + e_i$, where $e_i$ is drawn from the exponential distribution with rate parameter $\lambda(t_i + \bar{D}_i)$
    Update $r_i = \lceil r_i/h \rceil h$
    $t_{i+1} = t_i + \bar{D}_i + r_i$
    $i = i + 1$
    $n_i = \lceil \bar{D}_i/h \rceil h$
    Using $\bar{D}_i, I_i, M_i, V_i$ generate list of rain events $X^i$ from raineventsim with $\Theta$
**end**
**Return** list of rain events $X^i$, start time $s_i$ and end time $t_i$

This algorithm facilitates the simulation of sub-hourly rain events within a specified timeframe, using inputs like a seasonal rate function $\lambda$, a joint distribution model, a time step $h$, and the rain event simulator. It derives values for $D_i$, $I_i$, $M_i$, and $V_i$ from the D-vine copula joint distribution model. After updating $\bar{D}_i$, the algorithm calculates a random interval $r_i$ influenced by $\lambda$, ensuring a realistic temporal spacing between rain events. This interval, along with the duration $\bar{D}_i$, is used to predict the next event's start

time $t_{i+1}$. The number $n_i$ is then computed and utilized with other derived parameters to generate a series of rain events $X^i$ using the simulator. This sophisticated approach allows for detailed and accurate simulations of rainfall patterns. The R code for the implementation is given in Appendix E.2

Also, to assess our rainfall event simulator's robustness, we conducted simulations to replicate observed daily rainfall data characteristics. We calculated the mean and standard deviation of rainfall intensities for each simulation set and compared these metrics against the observed data.

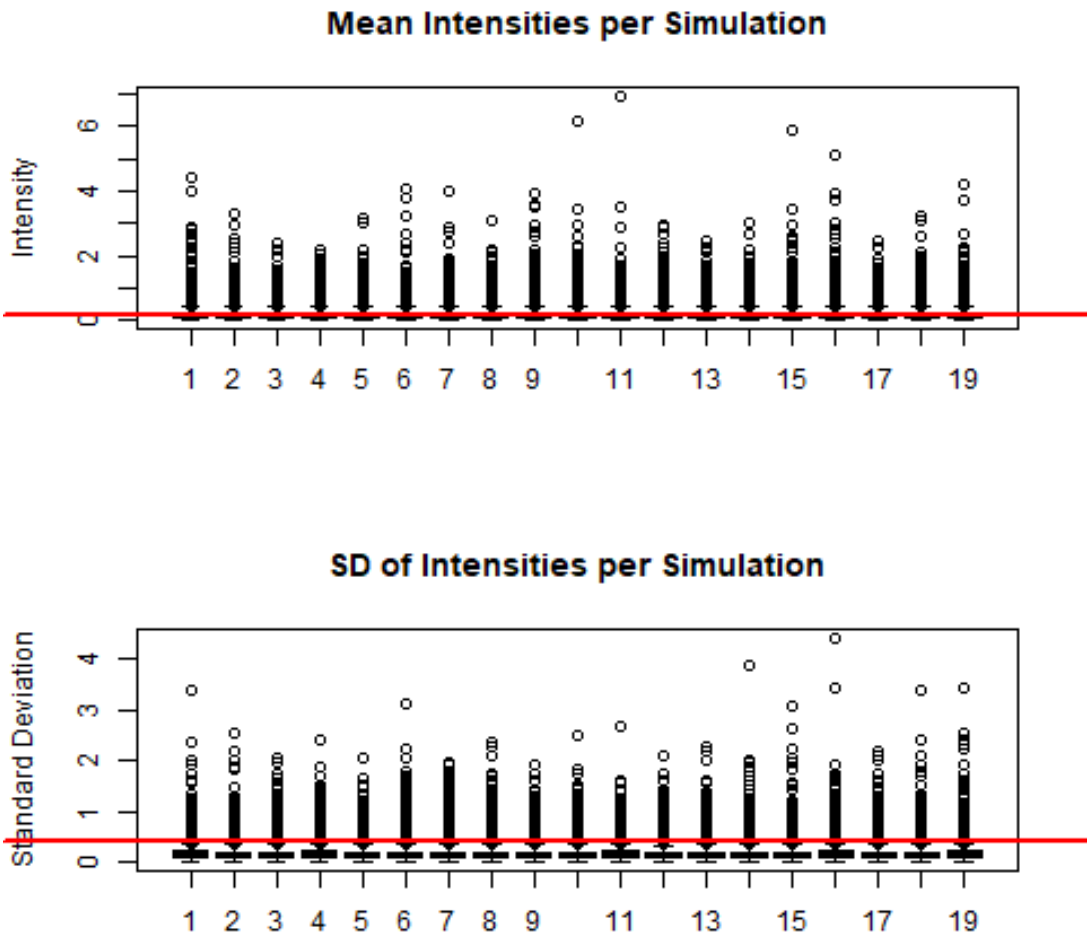


**Figure 7.9:** Comparison of Simulated and Observed Rainfall Intensity Characteristics.

The boxplots in Figure 7.9 illustrate the distribution of mean intensities and standard

deviations across 19 sets of simulated daily rainfall events, with red lines indicating the mean (0.1928) and standard deviation (0.4115092) of the observed rainfall data. The alignment of these reference lines within the simulated data demonstrates the simulator's capability to replicate the variability and intensity distribution of real-world rainfall events accurately. This indicates the simulator's reliability for hydrological and climatological studies, especially in contexts lacking observational data, confirming its suitability for generating realistic rainfall event sequences for analytical purposes.

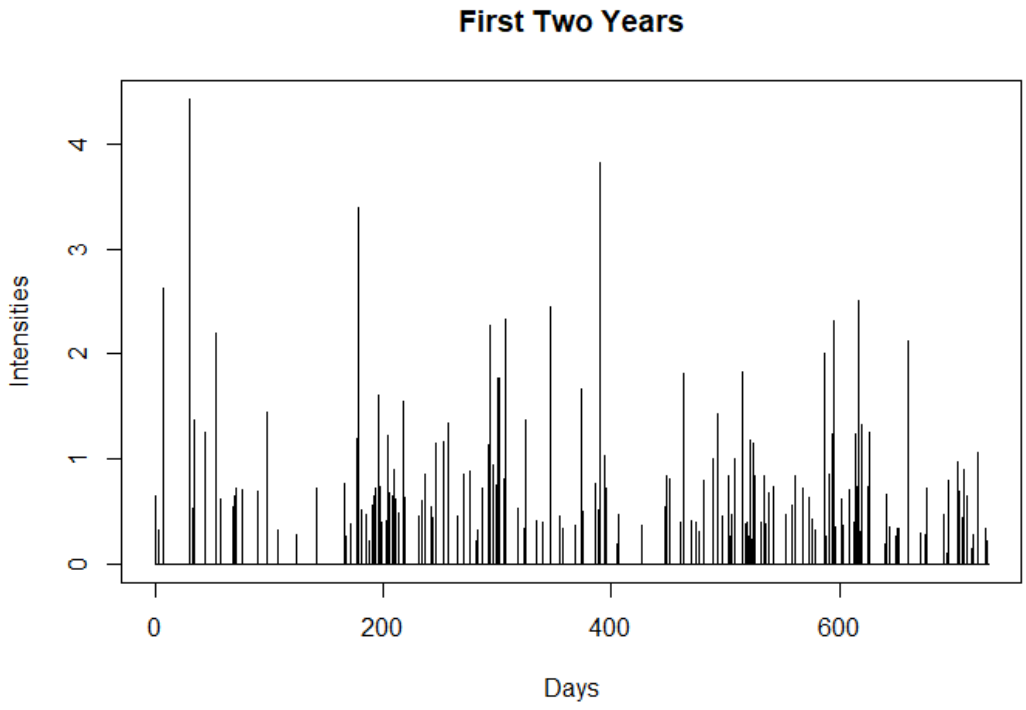


**Figure 7.10:** Series of Rainfall events for two years period by the Rainfall simulator using Algorithm 4

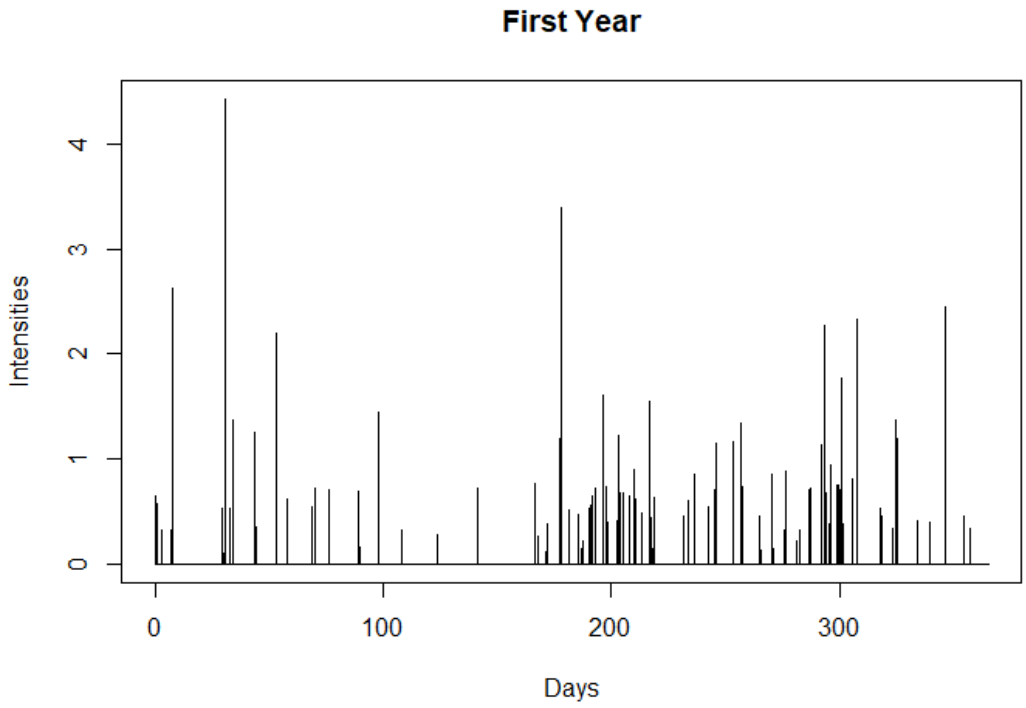


**Figure 7.11:** Series of Rainfall events for a year produced by the Rainfall simulator using Algorithm 4

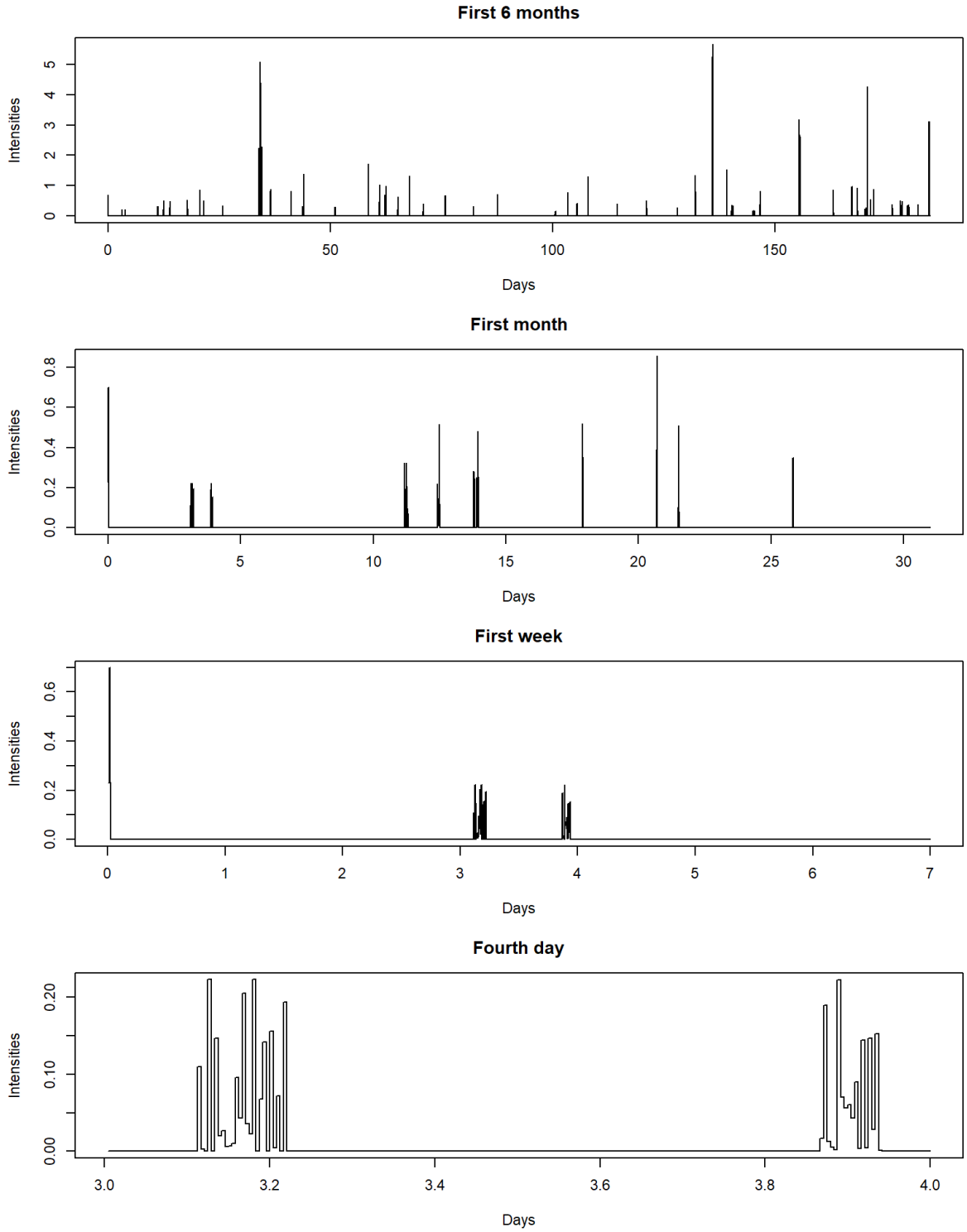


**Figure 7.12:** Series of Rainfall Events at different time resolutions produced by the Rainfall Simulator from Algorithm 4

## 7.10 Discussion

The rainfall event simulator generates values for DIMV using a joint distribution model calibrated from observed data. This approach ensures that the simulated events authentically reflect the statistical characteristics of real-world rainfall occurrences. Furthermore, the starting time of subsequent events is computed by integrating the seasonal rate function (representing inter-event time) with the exponential distribution. This method introduces a temporal pattern to the events that account for both regular and irregular intervals between rainfall occurrences, ensuring that the results closely resemble actual events. We simulated a 6-month sequence of rainfall events at different time resolutions to show the model's (Algorithm 4) efficacy in consistently generating sequences of rainfall events that align with the hydrological characteristics of the specified region. The results are given in Figure 7.10, Figure 7.11 and Figure 7.12. The model's parsimony is evident, given that its only parameters are those of D, I, M, and V, the seasonal rate parameters, and the tuning parameters of the rainfall simulator. The strategic choice of scaling factors, grounded in data-driven insights, played a pivotal role in this validation process. By carefully calibrating the extent of modification to the rainfall data parameters, the model was effectively challenged yet demonstrated its capacity to produce realistic and accurate simulations. This is particularly commendable given the intricate nature of rainfall events. This model provides a good representation and forecasting tool for rainfall patterns needed to design and plan hydrological structures. This balance between fidelity to observed data trends and the imposition of extreme conditions facilitated a robust evaluation of the simulation model, underscoring its potential as a predictive tool in hydrology and climate science.

# Chapter 8

## Conclusions

This chapter offers a comprehensive summary of the research and findings outlined in this thesis. Finally, some future research directions on rainfall event modelling are also outlined in this chapter.

## 8.1 Research Summary

Rainfall is crucial for our environment, but extreme rainfall events lead to problems like floods and droughts. When planning hydrological structures, it's essential to consider these risks. Unfortunately, the high-resolution rainfall data needed to model these risks is not available in most cases, and we often use rainfall simulators that can accurately reproduce realistic rainfall patterns. This research aims to build a stochastic parsimonious high-frequency rainfall simulator from high-frequency data that can accurately represent critical characteristics of rainfall events and temporal patterns of inter-event times.

Chapter 1 described the importance of the simulation of rainfall events and outlined the research aim and agenda. It also described the structure of the thesis and the novel contribution. Chapter 2 review of mathematical rainfall simulation models.

In Chapter 3, we defined and extracted rainfall events from a novel dataset of 6-minute high-resolution rainfall gauge data spanning 36 years from Sunbury, Australia, using rainfall characteristics: duration (D), intensity(I), maximum intensity (M), and volatility (V)(collectively referred to as DIMV) using 1-hour minimum interevent time (IET). DIMV were found to exhibit skewness in their distributions. A log transformation was applied to address the skewed nature of the DIMV data and effectively model them, and a log transformation was applied, i.e. (log(DIMV)). To study the behaviour of the data, we fitted appropriate marginal distributions to each of the rainfall characteristics, as this

also plays a critical role in joint (copula) modelling. Based on the AIC Criterion, the skew t distribution provided the most suitable fit for duration and volatility. On the other hand, the intensity and maximum intensity were appropriately fitted by the Generalized Extreme Value (GEV) distribution. Identifying these distributions sets the foundation for compound distribution and copula modelling.

Chapter 4, the extreme value analysis conducted on Sunbury, Victoria, Australia's rainfall events, intricately examines three pivotal variables: duration, intensity, and total rainfall. Through the adept application of univariate and bivariate threshold methodologies underpinned by the GPD, nuanced insights into the behaviour of these variables under extreme conditions have been garnered.The univariate POT approach, specifically tailored for this analysis, has proficiently modelled the exceedances for rainfall duration and total rainfall data, demonstrating a commendable fit as evidenced by the diagnostic plots. This robust modelling provides valuable predictive insights, notably enabling the accurate estimation of significant return levels, such as the 100-year return level for duration and total rainfall. The GPD model's efficacy in capturing these variables' tail behaviour and extremes underscores its utility in providing reliable forecasts for these crucial aspects of rainfall events.

Conversely, the intensity data presented unique challenges. While the GPD model generally showed a good fit in initial evaluations, a closer examination revealed its limitations in accurately modelling the most extreme outliers in the intensity data. This shortfall indicates a significant area for further research and methodological refinement, especially in enhancing the model's capacity to encapsulate all aspects of intensity data's extremal behaviour. The duration and intensity data manifest a negative correlation. This observation does not suggest an asymptotic dependence between these two variables. Among the array of extreme bivariate models assessed, the negative bivariate logistic model emerged as the superior fit, having the smallest AIC value. When we compare our simulated data to the observed data, we find that even the best bivariate extreme value distribution provides a poor fit as this model is designed for positive dependence, making

it impractical. We do not pursue any further analysis on bivariate extremes. The study validates the POT method and GPD model's effectiveness in capturing the extremes of rainfall duration and total rainfall, which is essential for water resource management and climate adaptation strategies. It also highlights the challenges in modelling rainfall intensity extremes, emphasizing the need for improved extreme value modelling techniques. Additionally, the research exposes the complexities of bivariate relationships in extreme values, suggesting a reevaluation of current models towards more accurate representations of rainfall events. These insights are pivotal in advancing the understanding of extreme rainfall dynamics and stress the necessity for ongoing methodological advancements to boost the precision and scope of extreme value analyses in hydrological contexts.

In Chapter 5, we developed a flexible univariate hybrid model F-Exp-GPD for modelling rainfall events duration and intensity by generalising the existing hybrid distribution: G-Exp-GPD by Debbabi et al. [113] by incorporating an arbitrary distribution F, where F is the best-fit distribution for the data set under consideration. The novel hybrid model F-Exp-GPD achieves a balanced representation of the bulk and tail behaviour. Using the F-Exp-GPD hybrid, three unique hybrid distribution models were formed: Skew t-Exp-GPD, Skew Normal-Exp-GPD and GEV-Exp-GPD. To demonstrate the efficacy in accurately characterising rainfall duration and intensity. The parameter estimation through MLE, implemented using the Maxlik package and the Nelder-Mead method in R, ensures precise fitting and robust results. The GEV-Exp-GPD model emerges as the most suitable, evidenced by its lowest AIC value which is also better than the AIC values for the skew t distribution for log(duration) (see Table 3.2) and GEV distribution for log(intensity) (see Table 3.3) in Chapter 3. Thereby, capturing the bulk and tail behaviour of the rainfall events duration and intensity data. The skew-t-Exp-GPD can also be used to model rainfall events duration. This research underscores the potential of the F-Exp-GPD approach for hydrological applications.

In Chapter 6, the dependence modelling and simulation of rainfall event characteristics - duration (D), intensity (I), maximum intensity (M), and volatility (V) were performed

using the vine copula approach. The focus was on determining the best copula model to represent the interdependence among these characteristics. Out of the vine copula structures, namely the R-vine, C-vine, and D-vine, The D-vine copula was identified as the best model because of its lowest AIC value, signifying a superior fit to the data. The robustness of the D-Vine copula model was further analysed by simulating the quadruple (DIMV) using the D-Vine copula. The appropriate transformations were implemented using the best marginal distribution for each characteristic: Skew t for both duration and volatility and GEV distribution for intensity and maximum intensity to make these simulated results comparable to the observed data. The results revealed that the proposed copula model effectively maintained the sample dependencies among the rain event characteristics. The copula simulated and observed data show a substantial similarity. The D-vine copula effectively bridges theoretical expectations with real-world data, proving its accuracy in modelling the complex dependencies of rainfall event characteristics. Both statistical and graphical assessments back its reliability. The D-vine copula offers a robust method for understanding and simulating rainfall events data dependencies.

In Chapter 7, we developed a flexible rainfall event simulator (raineventsim) for the detailed simulation of rainfall events, emphasising temporal intensity patterns. The function simulates accurate and realistic rainfall patterns across discrete time intervals by acknowledging crucial parameters such as average intensity, event duration, maximum intensity, and volatility. Its hybrid approach, incorporating deterministic constraints and stochastic variability via a random search method, ensures that simulated events represent actual rainfall events and are adaptable to specific requirements. Furthermore, the function's recursive structure and ability to adjust based on the provided constraints attest to its robustness. This tool is helpful for hydrologists, climatologists, and researchers seeking to model or analyse rainfall patterns with precision and realism. Next, we developed a model for simulating a sequence of rainfall events over a specified time range, which follows the steps: (1) Iterating from the start time $t_1$ to the end time $t_{max}$ (2) Drawing values for DIMV from the joint distribution model calibrated based on the observed data which add authenticity to the simulated events and assures that the generated DIMV

data mirrors the statistical properties of real-world rainfall events (3) Computing the next event's start time through the integration of the season rate function (inter-event time) and the exponential distribution, which introduces a realistic temporal pattern to the events and respects both consistent and irregular intervals between rain events (4) Generating a sequence of rain events where each event's internal structure is also captured making the simulated event rich in detail. This stochastic model is invaluable for hydrological studies, urban planning and climate change modelling, where high-resolution rainfall event data might be scarce or insufficient.

The methodology developed in this thesis for simulating point rainfall events exhibits a unique combination of strengths and areas for improvement when considered in the context of spatial simulations. Among its most notable strengths is the innovative approach to encapsulating the intricacies of rainfall dynamics through a high-resolution simulator. This approach leverages vine copulas and a hybrid model to accurately capture the dependencies among key rainfall event characteristics—duration, intensity, maximum intensity, and volatility—offering detailed insights into rainfall patterns with implications for hydrological forecasting and climate change studies.

However, while detailed and precise, the model's focus on point rainfall events introduces limitations in its applicability to broader spatial analyses. The calibration and validation of the model are demonstrated within a specific geographical and climatic context, raising questions about its generalizability across diverse environments. This limitation suggests a potential area for further research to adapt and validate the model for various geographic locations and extend its capabilities to encompass spatial rainfall distribution, thereby enhancing its utility for regional-scale hydrological modelling and planning.

Furthermore, the model's computational demands, particularly when employing copula-based methodologies for capturing complex dependencies among rainfall event characteristics, present challenges for scalability and real-time applications. Addressing these computational efficiency aspects, alongside expanding the model to simulate spatial rainfall

variability more effectively, represents critical pathways for future work. This evolution would bolster the model's robustness and applicability and contribute significantly to the broader field of hydrological sciences. Specifically, adapting to and minimising the effects of climatic unpredictability and change.

## 8.2 Further work

- The study of the hybrid model in Chapter 5 opens several avenues for future research. Further refinements to the F-Exp-GPD model could also be explored, incorporating other distribution types or additional parameters to enhance its ability to capture complex data patterns. The model could also be tested on different environmental data types to assess its applicability beyond rainfall. Additionally, investigating more sophisticated linking functions or different fitting methods might offer additional improvements in model performance. Lastly, incorporating this hybrid model into broader weather forecasting and climate change models could be an exciting and impactful direction for future work.

- Chapter 6 presents significant potential for future exploration. Applying the copula methodology to model the joint distribution of rainfall characteristics in other regions of Australia would provide a comprehensive assessment of its versatility and robustness across varied climates and geographical conditions. Since this study utilized a parametric method, future work could consider exploring non-parametric approaches for modelling the joint distribution of rainfall event characteristics. Non-parametric methods offer a flexible alternative advantageous in cases where the underlying data distribution is complex or unknown.

- Given the advancements achieved in Chapter 7, several intriguing avenues for future research manifest themselves. Firstly, while the developed simulator exhibits significant precision in modelling rainfall patterns, incorporating spatial variability into the tool could further enhance its realism. This spatial extension would allow researchers to investigate the interconnectedness of rainfall patterns across

varying geographical terrains. Additionally, with the ever-evolving nature of climatic patterns due to global changes, integrating real-time data or climate change projections could keep the simulator's outputs continually relevant. The current model's adeptness in capturing temporal patterns and internal event structures offers a firm foundation for extending it to simulate more complex events, such as simultaneous rain and wind, potentially of paramount importance for urban planners and architects. Machine learning techniques could be amalgamated into the model, facilitating it to continually learn and adapt from newly generated data. Lastly, given the emphasis on hydrological studies and urban planning, it would be pertinent to integrate the simulator's outputs with hydraulic and hydrodynamic models, ensuring a holistic approach to water and infrastructure management. Such synergistic interplay between different models could revolutionize the predictability and mitigation strategies associated with flooding and related natural calamities.

## Bibliography


1. I. Rodriguez-Iturbe, V. K. Gupta, and E. Waymire. Scale considerations in the modeling of temporal rainfall. *Water Resources Research*, 20(11):1611–1619, 1984.

2. D.J. Peres and A. Cancelliere. Derivation and evaluation of landslide-triggering thresholds by a monte carlo approach. *Hydrology and Earth System Sciences*, 18(12):4913–4931, 2014.

3. M.A. Islam, B. Yu, and N. Cartwright. Bartlett–lewis model calibrated with satellite-derived precipitation data to estimate daily peak 15 min rainfall intensity. *Atmosphere*, 14(6):985, 2023.

4. J. Park, C. Onof, and D. Kim. A hybrid stochastic rainfall model that reproduces some important rainfall characteristics at hourly to yearly timescales. *Hydrology and Earth System Sciences*, 23(2):989–1014, 2019.

5. P. Cowpertwait, V. Isham, and C. Onof. Point process models of rainfall: developments for fine-scale structure. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2086):2569–2587, 2007.

6. J. Kaczmarska, V. Isham, and C. Onof. Point process models for fine-resolution rainfall. *Hydrological Sciences Journal*, 59(11):1972–1991, 2014.

7. N. I. Ramesh, C. Onof, and D. Xie. Doubly stochastic poisson process models for precipitation at fine time-scales. *Advances in water resources*, 45:58–64, 2012.

8. N. I. Ramesh, A. P. Garthwaite, and C. Onof. A doubly stochastic rainfall model with exponentially decaying pulses. *Stochastic Environmental Research and Risk Assessment*, 32(6):1645–1664, 2017.

9. M. B. Singer, K. Michaelides, and D. E. Hobley. Storm 1.0: a simple, flexible, and parsimonious stochastic rainfall generator for simulating climate and climate change. *Geoscientific Model Development*, 11(9):3713–3726, 2018.

10. N. I. Ramesh, G. Rode, and C. Onof. A cox process with state-dependent exponential pulses to model rainfall. *Water Resources Management*, pages 1–17, 2022.

11. C. Gao, M. J. Booij, and Y. P. Xu. Development and hydrometeorological evaluation of a new stochastic daily rainfall model: Coupling markov chain with rainfall event model. *Journal of hydrology*, 589:125337, 2020.

12. C. Gao, X. Guan, M. J. Booij, Y. Meng, and Y. P. Xu. A new framework for a multi-site stochastic daily rainfall model: Coupling a univariate markov chain model with a multi-site rainfall event model. *Journal of Hydrology*, 598:126478, 2021.

13. N. Peleg, N. Ban, M. J. Gibson, A. S. Chen, A. Paschalis, P. Burlando, and J. P. Leitão. Mapping storm spatial profiles for flood impact assessments. *Advances in Water Resources*, 166:104258, 2022.

14. H. K. Ji, M. Mirzaei, S. H. Lai, A. Dehghani, and A. Dehghani. Implementing generative adversarial network (gan) as a data-driven multi-site stochastic weather generator for flood frequency estimation. *Environmental Modelling & Software*, 172:105896, 2024.

15. C. Czado. *Analyzing dependent data with vine copulas: A practical guide with R.* Springer, 2019.

16. Q. Song, J. Liu, and S. Sriboonchitta. Risk measurement of stock markets in brics, g7, and g20: Vine copulas versus factor copulas. *Mathematics*, 7(3):274, 2019.

17. T. Nagler., U. Schepsmeier, J. Stoeber, E. C. Brechmann, G. Graeler, T. Erhardt, C. Almeida, A. Min, C. Czado, and M. Hofmann. Statistical Inference of Vine Copulas. *R Package*, version:2.4.5, 2023.

18. P. J. Ward, M. C. de Ruiter, J. Mård, K. Schröter, A. Van Loon, T. Veldkamp, N. von Uexkull, N. Wanders, A. AghaKouchak, K. Arnbjerg-Nielsen, and L. Capewell. The need to integrate flood and drought disaster risk reduction strategies. *Water Security*, 11:100070, 2020.

19. C. Onof, R. E. Chandler, A. Kakou, P. Northrop, H. S. Wheater, and V. Isham. Rainfall modelling using poisson-cluster processes: a review of developments. *Stochastic Environmental Research and Risk Assessment*, 14:384–411, 2000.

20. G. Brigandì and G. T. Aronica. Generation of sub-hourly rainfall events through a point stochastic rainfall model. *Geosciences*, 9(5):226, 2019.

21. I. Rodriguez-Iturbe, B. F. De Power, and J.B. Valdes. Rectangular pulses point process models for rainfall: analysis of empirical data. *Journal of Geophysical Research: Atmospheres*, 92(D8):9645–9656, 1987.

22. D.S. Cameron, K. J. Beven, J. Tawn, S. Blazkova, and P. Naden. Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). *Journal of Hydrology*, 219(3-4):169–187, 1999.

23. P. S. Cowpertwait. A generalized point process model for rainfall. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 447(1929):23–37, 1994.

24. C. Genest and A. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

25. I. Rodriguez-Iturbe, D. R. Cox, and V. Isham. Some models for rainfall based on stochastic point processes. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 410(1839):269–288, 1987.

26. G. Calenda and F. Napolitano. Parameter estimation of neyman–scott processes for temporal point rainfall simulation. *Journal of Hydrology*, 225(1-2):45–66, 1999.

27. I. Rodriguez-Iturbe, D. R. Cox, and V. Isham. A point process model for rainfall: further developments. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 417(1853):283–298, 1988.

28. C. Onof and H. S. Wheater. Modelling of british rainfall using a random parameter bartlett-lewis rectangular pulse model. *Journal of Hydrology*, 149(1-4):67–95, 1993.

29. M.N. Khaliq and C. Cunnane. Modelling point rainfall occurrences with the modified bartlett-lewis rectangular pulses model. *Journal of Hydrology*, 180(1-4):109–138, 1996.

30. D. Entekhabi, I. Rodriguez-Iturbe, and P. S. Eagleson. Probabilistic representation of the temporal rainfall process by a modified neyman-scott rectangular pulses model: Parameter estimation and validation. *Water Resources Research*, 25(2):295–302, 1989.

31. P.S.P Cowpertwait. A poisson-cluster model of rainfall: some high-order moments and extreme values. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):885–898, 1998.

32. Y. Gyasi-Agyei and G. R. Willgoose. A hybrid model for point rainfall modeling. *Water Resources Research*, 33(7):1699–1706, 1997.

33. Y. Gyasi-Agyei and G.R. Willgoose. Generalisation of a hybrid model for point rainfall. *Journal of Hydrology*, 219(3-4):218–224, 1999.

34. U. Lall, B. Rajagopalan, and D. G. Tarboton. A nonparametric wet/dry spell model for resampling daily precipitation. *Water resources research*, 32(9):2803–2823, 1996.

35. D. Cameron, K. Beven, and J. Tawn. Modelling extreme rainfalls using a modified random pulse bartlett–lewis stochastic rainfall model (with uncertainty). *Advances in water resources*, 24(2):203–211, 2000.

36. D Cameron, K Beven, and J Tawn. An evaluation of three stochastic rainfall models. *Journal of Hydrology*, 228(1-2):130–149, 2000.

37. K. Beven and A. Binley. The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6(3):279–298, 1992.

38. S. Kim and M. L. Kavvas. Stochastic point rainfall modeling for correlated rain cell intensity and duration. *Journal of Hydrologic Engineering*, 11(1):29–36, 2006.

39. K. Singh and V.P. Singh. Derivation of bivariate probability density functions with exponential marginals. *Stochastic Hydrology and Hydraulics*, 5(1):55–68, 1991.

40. B. Bacchi, G. Becciu, and N. T. Kottegoda. Bivariate exponential model applied to intensities and durations of extreme rainfall. *Journal of hydrology*, 155(1-2):225–236, 1994.

41. E. J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.

42. J. R. Córdova and I. Rodríguez-Iturbe. On the probabilistic structure of storm surface runoff. *Water Resources Research*, 21(5):755–763, 1985.

43. N. K. Goel, R.S. Kurothe, B.S. Mathur, and R.M. Vogel. A derived flood frequency distribution for correlated rainfall intensity and duration. *Journal of Hydrology*, 228(1-2):56–67, 2000.

44. F Downton. Bivariate exponential distributions in reliability theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(3):408–417, 1970.

45. S. Han, H. S. Shin, and S. Kim. Temporal downscale for hourly rainfall time series using correlated neyman-scott rectangular pulse point rainfall model. *KSCE Journal of Civil Engineering*, 13:463–469, 2009.

46. P.S.P. Cowpertwait, G. Xie, V. Isham, C. Onof, and D.C.I. Walsh. A fine-scale point process model of rainfall with dependent pulse depths within cells. *Hydrological sciences journal*, 56(7):1110–1117, 2011.

47. N. E. C. Vandenberghe, S.and Verhoest, C. Onof, and B. De Baets. A comparative copula-based bivariate frequency analysis of observed and simulated storm events: A case study on bartlett-lewis modeled rainfall. *Water Resources Research*, 47(7), 2011.

48. T. Velghe, P. A. Troch, F.P. De Troch, and J. Van de Velde. Evaluation of cluster-based rectangular pulses point process models for rainfall. *Water Resources Research*, 30(10):2847–2857, 1994.

49. N. Verhoest, P. A. Troch, and F. P. De Troch. On the applicability of bartlett–lewis rectangular pulses models in the modeling of design storms at a point. *Journal of hydrology*, 202(1-4):108–120, 1997.

50. C. Li, V. P. Singh, and A. K. Mishra. Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water resources research*, 48(3), 2012.

51. N.I. Ramesh. Temporal modelling of short-term rainfall using cox processes. *Environmetrics: The official journal of the International Environmetrics Society*, 9(6):629–643, 1998.

52. M.B. Singer and K. Michaelides. Deciphering the expression of climate change within the lower colorado river basin by stochastic simulation of convective rainfall. *Environmental Research Letters*, 12(10):104011, 2017.

53. K. Michaelides and J. Wainwright. Modelling the effects of hillslope–channel coupling on catchment hydrological response. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, 27(13):1441–1457, 2002.

54. K. Michaelides and M. D. Wilson. Uncertainty in predicted runoff due to patterns of spatially variable infiltration. *Water Resources Research*, 43(2), 2007.

55. K. Michaelides and J. Wainwright. Internal testing of a numerical model of hillslope–channel coupling using laboratory flume experiments. *Hydrological Processes: An International Journal*, 22(13):2274–2291, 2008.

56. K. Beven and J. Freer. A dynamic topmodel. *Hydrological processes*, 15(10):1993–2011, 2001.

57. C. M. Evans, D. G. Dritschel, and M. B. Singer. Modeling subsurface hydrology in floodplains. *Water Resources Research*, 54(3):1428–1459, 2018.

58. K. K. Caylor, P. D'Odorico, and I. Rodriguez-Iturbe. On the ecohydrology of structurally heterogeneous semiarid landscapes. *Water Resources Research*, 42(7), 2006.

59. F. Laio, P. D'Odorico, and L. Ridolfi. An analytical model to relate the vertical root distribution to climate and soil properties. *Geophysical Research Letters*, 33(18), 2006.

60. P. D'Odorico, K. Caylor, G. S. Okin, and T. M. Scanlon. On soil moisture–vegetation feedbacks and their possible effects on the dynamics of dryland ecosystems. *Journal of Geophysical Research: Biogeosciences*, 112(G4), 2007.

61. D. Cross, C. Onof, H. Winter, and P. Bernardara. Censored rainfall modelling for estimation of fine-scale extremes. *Hydrology and Earth System Sciences*, 22(1):727–756, 2018.

62. C. Onof and H. S Wheater. Improvements to the modelling of british rainfall using a modified random parameter bartlett-lewis rectangular pulse model. *Journal of Hydrology*, 157(1-4):177–195, 1994.

63. S. Fatichi, V. Y. Ivanov, and E. Caporali. Simulation of future climate scenarios with a weather generator. *Advances in Water resources*, 34(4):448–467, 2011.

64. A. Paschalis, P. Molnar, S. Fatichi, and P. Burlando. On temporal stochastic modeling of precipitation, nesting models across scales. *Advances in water resources*, 63:152–166, 2014.

65. S. Blazkov and K. Beven. Flood frequency prediction for data limited catchments in the czech republic using a stochastic rainfall model and topmodel. *Journal of Hydrology*, 195(1-4):256–278, 1997.

66. A. Mendoza-Resendiz, M. Arganis-Juarez, R. Dominguez-Mora, and B. Echavarria. Method for generating spatial and temporal synthetic hourly rainfall in the valley of mexico. *Atmospheric research*, 132:411–422, 2013.

67. B. Bonaccorso, G. Brigandì, and G. T. Aronica. Combining regional rainfall frequency analysis and rainfall-runoff modelling to derive frequency distributions of peak flows in ungauged basins: a proposal for sicily region (italy). *Advances in Geosciences*, 44:15–22, 2017.

68. C. Gao, Y. P. Xu, Q. Zhu, Z. Bai, and L. Liu. Stochastic generation of daily rainfall events: A single-site rainfall model with copula-based joint simulation of rainfall characteristics and classification and simulation of rainfall patterns. *Journal of Hydrology*, 564:41–58, 2018.

69. K. Breinl, T. Turkington, and M. Stowasser. Stochastic generation of multi-site daily precipitation for applications in risk management. *Journal of Hydrology*, 498:23–35, 2013.

70. A. C. Callau Poduje and U. Haberlandt. Spatio-temporal synthesis of continuous precipitation series using vine copulas. *Water*, 10(7):862, 2018.

71. C. Nop, R. M. Fadhil, and K. Unami. A multi-state markov chain model for rainfall to be used in optimal operation of rainwater harvesting systems. *Journal of Cleaner Production*, 285:124912, 2021.

72. P. Chauhan, M. E. Akiner, R. Shaw, and K. Sain. Forecast future disasters using hydro-meteorological datasets in the yamuna river basin, western himalaya: Using markov chain and lstm approaches. *Artificial Intelligence in Geosciences*, page 100069, 2024.

73. Z. W. Kundzewicz, S. Kanae, S. I. Seneviratne, J. Handmer, N. Nicholls, P. Peduzzi, R. Mechler, L. M. Bouwer, N. Arnell, and K. Mach. Flood risk and climate change: global and regional perspectives. *Hydrological Sciences Journal*, 59(1):1–28, 2014.

74. M. P. Clark, A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay. Framework for understanding structural errors (fuse): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44(12), 2008.

75. Y. Chen, A. Paschalis, E. Kendon, D. Kim, and C. Onof. Changing spatial structure of summer heavy rainfall, using convection-permitting ensemble. *Geophysical Research Letters*, 48(3):e2020GL090903, 2021.

76. U. F. Abdul Rauf and P. Zeephongsekul. Copula based analysis of rainfall severity and duration: a case study. *Theoretical and applied climatology*, 115(1-2):153–166, 2014.

77. M. L. Larsen and J. B. Teves. Identifying individual rain events with a dense disdrometer network. *Advances in Meteorology*, 2015, 2015.

78. D. Dunkerley. Identifying individual rain events from pluviograph records: a review with analysis of data from an australian dryland site. *Hydrological Processes*, 22(26):5024–5036, 2008.

79. G. Grolemund and H. Wickham. Dates and times made easy with lubridate. *Journal of statistical software*, 40(3), 2011.

80. I. Molina-Sanchis, R. Lázaro, E. Arnau-Rosalén, and A. Calvo-Cases. Rainfall timing and runoff: The influence of the criterion for rain event separation. *J. Hydrol. Hydromech*, 64:226–236, 2016.

81. A. Azzalini and R. B. Arellano-Valle. Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *Journal of Statistical Planning and Inference*, 143(2):419–433, 2013.

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

83. A. F. Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly journal of the Royal Meteorological Society*, 81(348):158–171, 1955.

84. R. Mises von. "la distribution de la plus grande de n valeurs, "in selected papers of richard von mises, vol. 2. *American Mathematical Society*, pages 271–294, 1936.

85. S. Markose and A. Alentorn. The generalized extreme value distribution, implied tail index, and option pricing. *The Journal of derivatives*, 18(3):35–60, 2011.

86. H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

87. F. J. Acero, J. A. García, and M. C. Gallego. Peaks-over-threshold study of trends in extreme rainfall over the iberian peninsula. *Journal of climate*, 24(4):1089–1105, 2011.

88. R. de Fondeville and Anthony C. Davison. Functional peaks over threshold analysis. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 84(4):1392–1422, 2022.

89. R. P. Towe, J. A. Tawn, R. Lamb, and C. G. Sherlock. Model-based inference of conditional extreme value distributions with hydrological applications. *Environmetrics*, 30(8):e2575, 2019.

90. J. Beirlant, G. Maribe, and A. Verster. Penalized bias reduction in extreme value estimation for censored Pareto-type data, and long-tailed insurance applications. *Insurance: Mathematics and Economics*, 78:114–122, 2018.

91. M. Gilli and E. K. Ellezi. An Application of Extreme Value Theory for Measuring Financial Risk. *Computational Economics*, 27:207–228, 2006.

92. B. Finkenstadt and H. Rootzen. *Extreme values in finance, telecommunications, and the environment*. Chapman & Hall/CRC, 2003.

93. L. Cheng, A. AghaKouchak, E. Gilleland, and R. W. Katz. Non-stationary extreme value analysis in a changing climate. *Climatic Change*, 127(2):353–369, 2014.

94. M. Thomas, M. Lemaitre, M. L. Wilson, C. Viboud, Y. Yordanov, H. Wackernagel, and F. Carrat. Applications of Extreme Value Theory in Public Health. *PLOS ONE*, 11(7):e0159312, 2016.

95. G. Mascaro. On the distributions of annual and seasonal daily rainfall extremes in central arizona and their spatial variability. *Journal of Hydrology*, 559:266–281, 2018.

96. S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

97. J. Mohamed and M. B. Adam. Modeling of magnitude and frequency of extreme rainfall in somalia. *Modeling Earth Systems and Environment*, 8(3):4277–4294, 2022.

98. R. W. Katz, M. B Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in Water Resources*, 25:1287–1304, 2002.

99. A. Kiriliouk, H. Rootzén, J. Segers, and J. L. Wadsworth. Peaks over thresholds modeling with multivariate generalized pareto distributions. *Technometrics*, 61(1):123–135, 2019.

100. X. Zhao, Z. Zhang, W. Cheng, and P. Zhang. A new parameter estimator for the generalized pareto distribution under the peaks over threshold framework. *Mathematics*, 7(5):406, 2019.

101. Carl Scarrott and Anna Macdonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1):33–60, 2012.

102. Q. C. Chukwudum and S. Nadarajah. Bivariate extreme value analysis of rainfall and temperature in Nigeria. *Environmental Modeling & Assessment*, 27(2):343–362, 2022.

103. A. Borsos. Application of bivariate extreme value models to describe the joint behavior of temporal and speed related surrogate measures of safety. *Accident Analysis & Prevention*, 159:106274, 2021.

104. M. A Ribatet and D. Christophe. POT: Generalized Pareto Distribution and Peaks Over Threshold (R package version 1.1-7. 2019.

105. S.G. Coles, J. Heffernan, and J.A. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2:339 – 365, 1999.

106. J. Kim and C. D. Scott. Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(1):2529–2565, 2012.

107. S. G. Walker, P. Damien, P. W. Laud, and A. F. Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):485–527, 1999.

108. P. Orbanz and Y. W. Teh. Bayesian nonparametric models. *Encyclopedia of machine learning*, 1:81–89, 2010.

109. K. Cooray and M. M. Ananda. Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian Actuarial Journal*, 2005(5):321–334, 2005.

110. J. Carreau and Y. Bengio. A hybrid pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12:53–76, 2009.

111. M. Kratz. Normex, a new method for evaluating the distribution of aggregated heavy tailed risks: application to risk measures. *Extremes*, 17:661–691, 2014.

112. S. Nadarajah and S. A. Bakar. New composite models for the danish fire insurance data. *Scandinavian Actuarial Journal*, 2014(2):180–187, 2014.

113. N. Debbabi, M. Kratz, and M. Mboup. A self-calibrating method for heavy tailed data modeling: Application in neuroscience and finance. *SSRN Journal*, 2016.

114. E. K. KARA. A study on modeling of lifetime with right-truncated composite lognormal-pareto distribution: Actuarial premium calculations. *Gazi University Journal of Science*, pages 1–1, 2021.

115. M. H. A. Majid, K. Ibrahim, and N. Masseran. Three-part composite pareto modelling for income distribution in malaysia. *Mathematics*, 11(13):2899, 2023.

116. J. Pickands. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3:119–131, 1975.

117. C. Kleiber and S. Kotz. *Statistical size distributions in economics and actuarial sciences*. John Wiley & Sons, 2003.

118. C. Czado. Pair-copula constructions of multivariate copulas. In *Copula Theory and Its Applications*, pages 93–109, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

119. D. Dupuis. Using copulas in hydrology: Benefits, cautions, and issues. *Journal of Hydrologic Engineering*, 12(4):381–393, 2007.

120. A. Sklar. Fonctions de Répartition â n Dimensions et Leurs Marges. *Publications de l'Institut de Statistique de L'Universitè de Paris*, 8:229–231, 1959.

121. R. B. Nelsen. *An introduction to copulas*. Springer, 2006.

122. P. Embrechts, F. Lindskog, and A. Mcneil. Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, page 329–384, 2003.

123. H. Joe. *Dependence modeling with copulas*. CRC Press/Taylor & Francis, 2015.

124. C. Meyer. The bivariate normal copula. *Communications in Statistics - Theory and Methods*, 42(13):2402–2422, 2013.

125. M. Sadegh, E. Ragno, and A. AghaKouchak. Multivariate copula analysis toolbox (mvcat): Describing dependence and underlying uncertainty using a bayesian framework. *Water resources research*, 53(6):5166–5183, 2017.

126. M. S. Ismail and N. Masseran. Modeling the characteristics of unhealthy air pollution events using bivariate copulas. *Symmetry (Basel)*, 15(4):907–, 2023.

127. X. Yang, Z. Chen, and M. Qin. Joint probability analysis of streamflow and sediment load based on hybrid copula. *Environmental science and pollution research international*, 30(16):46489–46502, 2023.

128. G. Taillon, K. Onishi, T. Mineshima, and K. Miyagawa. Statistical analysis of cavitation erosion impacts in a vibratory apparatus with copulas. *IOP Conference Series: Earth and Environmental Science*, 240(6):62035–, 2019.

129. M.A. Boateng, A.Y. Omari-Sasu, R.K. Avuglah, and N.K. Frempong. A mixture of clayton, gumbel, and frank copulas: A complete dependence model. *Journal of Probability and Statistics*, 2022, 2022.

130. N. Klein, T. Kneib, G. Marra, and R. Radice. Bayesian mixed binary-continuous copula regression with an application to childhood undernutrition. In *Flexible Bayesian Regression Modelling*, pages 121–152. Elsevier, 2020.

131. C. Czado and T. Nagler. Vine copula based modeling. *Annual Review of Statistics and Its Application*, 9(1):453–477, 2022.

132. K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2009.

133. H. Joe. Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. In *Ruschendorf, L., Schweizer, B., Taylor, M.D. (Eds.), Distributions with Fixed Marginals and Related Topics*, 1996.

134. T. Bedford and R. M. Cooke. Monte carlo simulation of vine dependent random variables for applications in uncertainty analysis. In *Proceedings of ESREL2001*, Turin, Italy, 2001.

135. T. Bedford and R. M. Cooke. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031 – 1068, 2002.

136. K. Aas. Pair-copula constructions for financial applications: A review. *Econometrics*, 4(4), 2016.

137. J. Dißmann, E.C. Brechmann, C. Czado, and D. Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69, 2013.

138. T. M. Nazeri, Y. Ramezani, C. de Michele, and R. Mirabbasi. Multivariate analysis of rainfall and its deficiency signatures using vine copulas. *International Journal of Climatology*, 42(4):2005–2018, 2022.

139. M. Marani. On the correlation structure of continuous and discrete point rainfall. *Water Resources Research*, 39(5), 2003.

# Appendix A: R code for Rainfall Events Extraction

## Appendix A.1: Data Cleaning and Rainfall Extraction

```
#Data Cleaning and Rainfall Extraction
library(lubridate)

rainfall_data <- read.csv("C:/Users/Administrator/
Documents/R/SITE_230202_MR_510.csv", header = TRUE, stringsAsFactors = FALSE)
rainfall_data<-read.csv(file.choose(),header = TRUE)
all.times <- rainfall_data$measure_date
all.times.POSIX <- dmy_hms(all.times, tz = "Australia/Melbourne")

diff.all.times.POSIX <- diff(all.times.POSIX)

diff.all.times.POSIX==360 #checking if the data had 6mins gap
#
length(diff.all.times.POSIX) #length will be equal to original length minus 1

png(file="C:/Users/Administrator/Documents/rainfall plot")
plot(rainfall_data$measure_value ~ all.times.POSIX,
main="Rainfall Data", ylab = "Depth(mm)" ,xlab = "Year",)
dev.off()

identical(round_date(all.times.POSIX, unit ="seconds"),all.times.POSIX)



#View(rainfall_data)
rainfall_data <- rainfall_data[-(111621:111906), ] #deleting the last month
rainfall_data <- rainfall_data[-c(107417, 107418, 107445, 107446), ]
#deleting zero readings with odd time stamps
time.POSIX <- dmy_hms(rainfall_data[,3], tz = "Australia/Melbourne")
#converting to POSIXct format

#Creating 6mins time stamps Method 1
minute.POSIX <- minute(time.POSIX)
w <- which(minute.POSIX %% 6 != 0)
```

```
time.POSIX[w]
time.POSIX <- round_date(time.POSIX, unit="minute")
time.POSIX[w]
which(minute(time.POSIX) %% 6 != 0)
all(as.numeric(diff(time.POSIX)) %% 360 == 0)

# year(time.POSIX)
# month(time.POSIX, label=T)
# month(time.POSIX)

time.6minutes <- cumsum(c(0,diff(time.POSIX)) / 360) + 1
## how many 6 minute intervals have passed since the first ever measurement
## time.POSIX[which(time.6minutes==42)]

depth<-rainfall_data[,4]
time<-time.6minutes

# fill in zeros
maxt <- max(time)
newd <- rep(0, maxt)
ti <- 1 #time counter
idx <- 1 #position in depth vector
newd[1] <- depth[1]
while (ti < maxt) {
  ti <- ti + 1
  if (ti == time[idx+1]) {
    idx <- idx + 1
    newd[ti] <- depth[idx]
  }
}

New_Rainfall_Data<-data.frame(Datetime=time.POSIX[1]+360*
(0:(maxt-1)), Time=1:maxt, Depth=newd)

save(New_Rainfall_Data, file = "Clean_Rainfall_Data.RData")
#data name:New_Rainfall_Data

#Load the clean data
load(file="Clean_Rainfall_Data.RData")
```

```
IET <- 10
#since it is 6 mins time resolution

newd <- New_Rainfall_Data$Depth

  # newd vector of rainfall depth every 6min
  # IET measured in units of 6min
  # returns a matrix where each row is a rainfall event
  # cols give durations depths intensities starttimes
  durations <- c()
  depths <- c()
  starttimes <- c()
  n_events <- 0
  ti <- 1
  ti_max <- length(newd)
  event_flag <- FALSE
  event_zeros <- 0
  event_durtn <- 0
  event_depth <- 0

  while (ti <= ti_max) {
    if (event_flag) { # in an event
      if (newd[ti] == 0) {
        event_zeros <- event_zeros + 1
        event_durtn <- event_durtn + 1
        if (event_zeros == IET) { # rain event finishes
          durations[n_events] <- event_durtn - IET
          depths[n_events] <- event_depth
          event_flag <- FALSE
          cat("event", n_events, "duration", event_durtn - IET,
          "depth", event_depth, "\n")
        }
      } else {
        event_zeros <- 0
        event_durtn <- event_durtn + 1
        event_depth <- event_depth + newd[ti]
      }
    } else { # between events
```

```
      if (newd[ti] > 0) {
        n_events <- n_events + 1
        event_flag <- TRUE
        event_zeros <- 0
        event_durtn <- 1
        event_depth <- newd[ti]
        starttimes[n_events] <- ti
      }
    }
    ti <- ti + 1
  }
  #return(cbind(durations, depths, depths/durations, starttimes))


events <- list()
for (i in 1:n_events) {
  idx <- starttimes[i]:(starttimes[i]+durations[i]-1)
  events[[i]] <- New_Rainfall_Data[idx,]
}

events
save(events, file = "events_rainfall.RData")


#Load the rainfall events data
load(file="events_rainfall.RData")
library(lubridate)

#Extraction of relevant summaries
#durations; total depth, total intensity, volatility and  max intensity
n_events<- length(events)
t_duration <- rep(NA, n_events)
t_depth <- rep(NA, n_events)
event_rows <- rep(NA, n_events)
t_intensity <- rep(NA, n_events)# Total intensity
idx_intensity <- rep(NA, n_events)
volatility<- rep(NA, n_events)
mod_volatility<- rep(NA, n_events)
max_intensity <- rep(NA, n_events)
```

```
#Start Time, End Time,
start_time <- c()
end_time <- c()

n_depth_zeros <-rep()

# Function to calculate the volatility
vol <- function(x) {
  sum(diff(x)^2)/length(x)
}
# # Function to calculate the modified volatility
# mod_vol <- function(x) {
#   sum(diff(x + runif(length(x),0,.1/6))^2)/length(x)
# }

indx<-setNames( rep(c('summer', 'autumn', 'winter', 'spring'),each=3), c(12,1:11))

for (i in 1:n_events ){
  event_rows[i] <- (nrow(events[[I]]))
  # Previous duration used before conversion
  t_duration[i] <-(event_rows[[i]]*6)
  # Multiplied by 6 to give total number of mins
  t_depth[i] <- sum(events[[i]]$Depth)
  t_intensity[i] <- (t_depth[i])/t_duration[i]
  idx_intensity <- events[[i]]$Depth/6 # Indexed intensities
  max_intensity[i] <- max(idx_intensity)
  volatility[i] <- vol(idx_intensity)
  #mod_volatility[i] <- mod_vol(idx_intensity)
  start_time[i] <- events[[i]]$Datetime[[1]]
  end_time[i] <- events[[i]]$Datetime[[1]] + t_duration[i]*60
}

start_time <- as_datetime(start_time, tz = "Australia/Melbourne")
end_time<-as_datetime(end_time, tz = "Australia/Melbourne")
months<- month(start_time)
seasons <-unname(indx[as.character(months)])

R_Summary<-data.frame(Duration=t_duration,Depth=t_depth,Intensity=t_intensity,
```

```
Max_intensity=max_intensity,Volatility=volatility,
start_time=start_time,end_time=end_time,months=months,seasons=seasons)

View(R_Summary)
save(R_Summary, file = "Summary_Events.RData")

#Load the rainfall events data
load(file="Summary_Events.RData")

#Removing small events where total depth is lower than 1mm
idx<-R_Summary[,2] >=1
idx
Summary_Events_needed <- data.frame(R_Summary[idx,])
names(Summary_Events_needed)<- c("Duration","Depth","Intensity",
"Max_Intensity", "Volatility", "Start_Time", "End_Time","Months", "Seasons")
Summary_Events_needed

save(Summary_Events_needed, file = "Rainfall_Events_Summary.RData")

View(Summary_Events_needed)
```

## Appendix A.2: Q-Q Plot for DIMV using fitted distributions

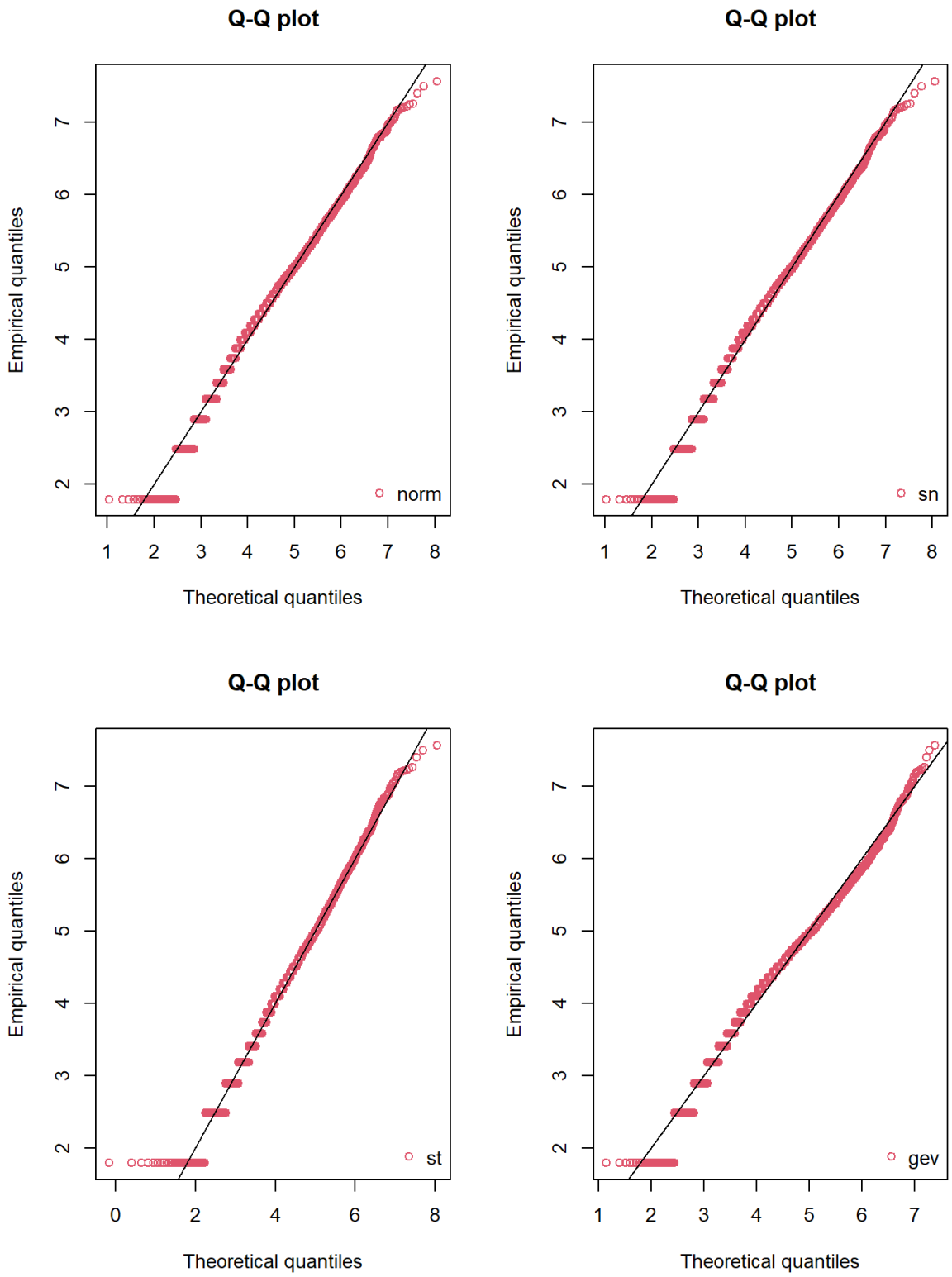


**Figure 1:** Q-Q plot for log(duration) with fitted distributions

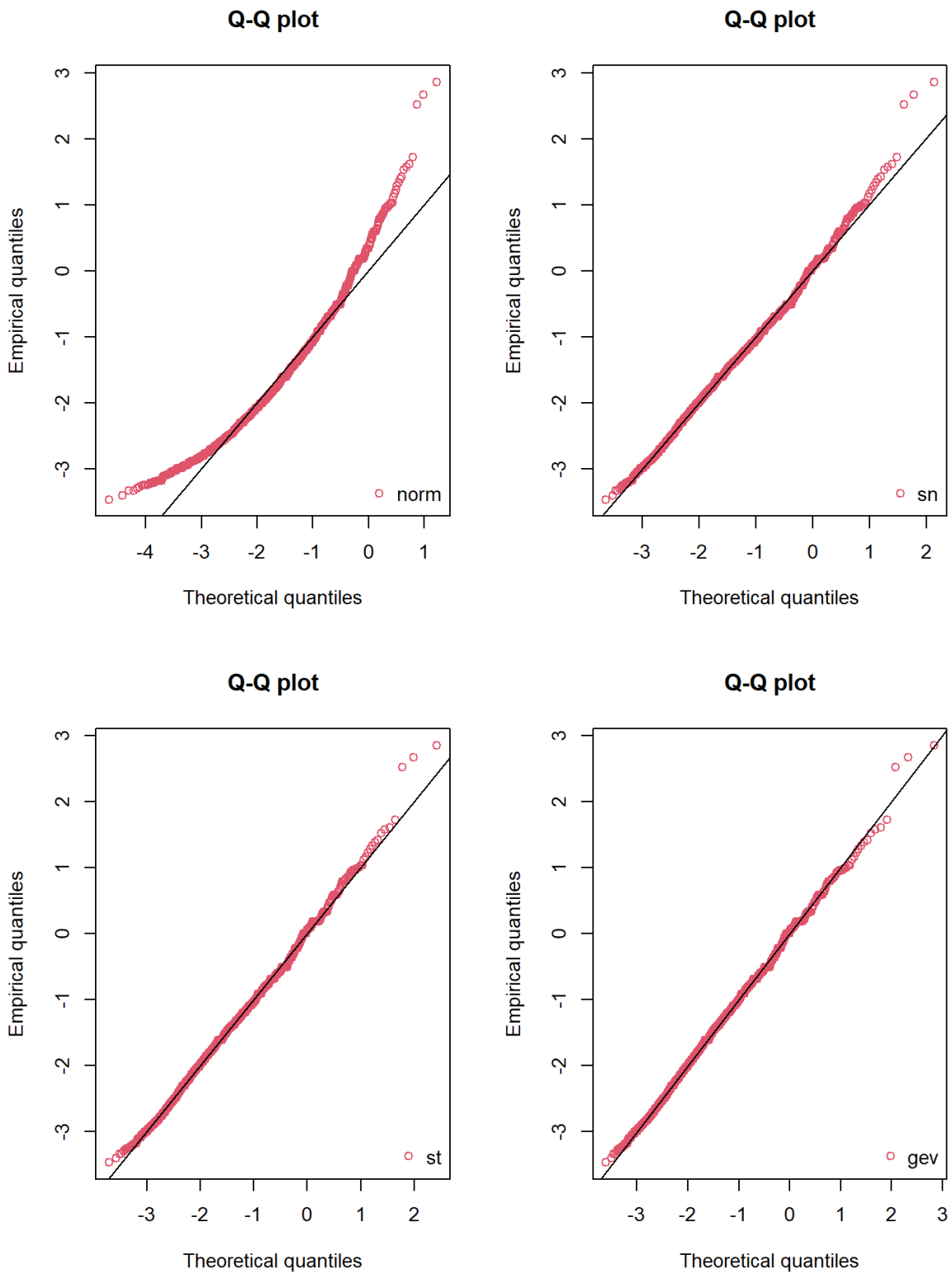


**Figure 2:** Q-Q plot for log(intensity) with fitted distributions

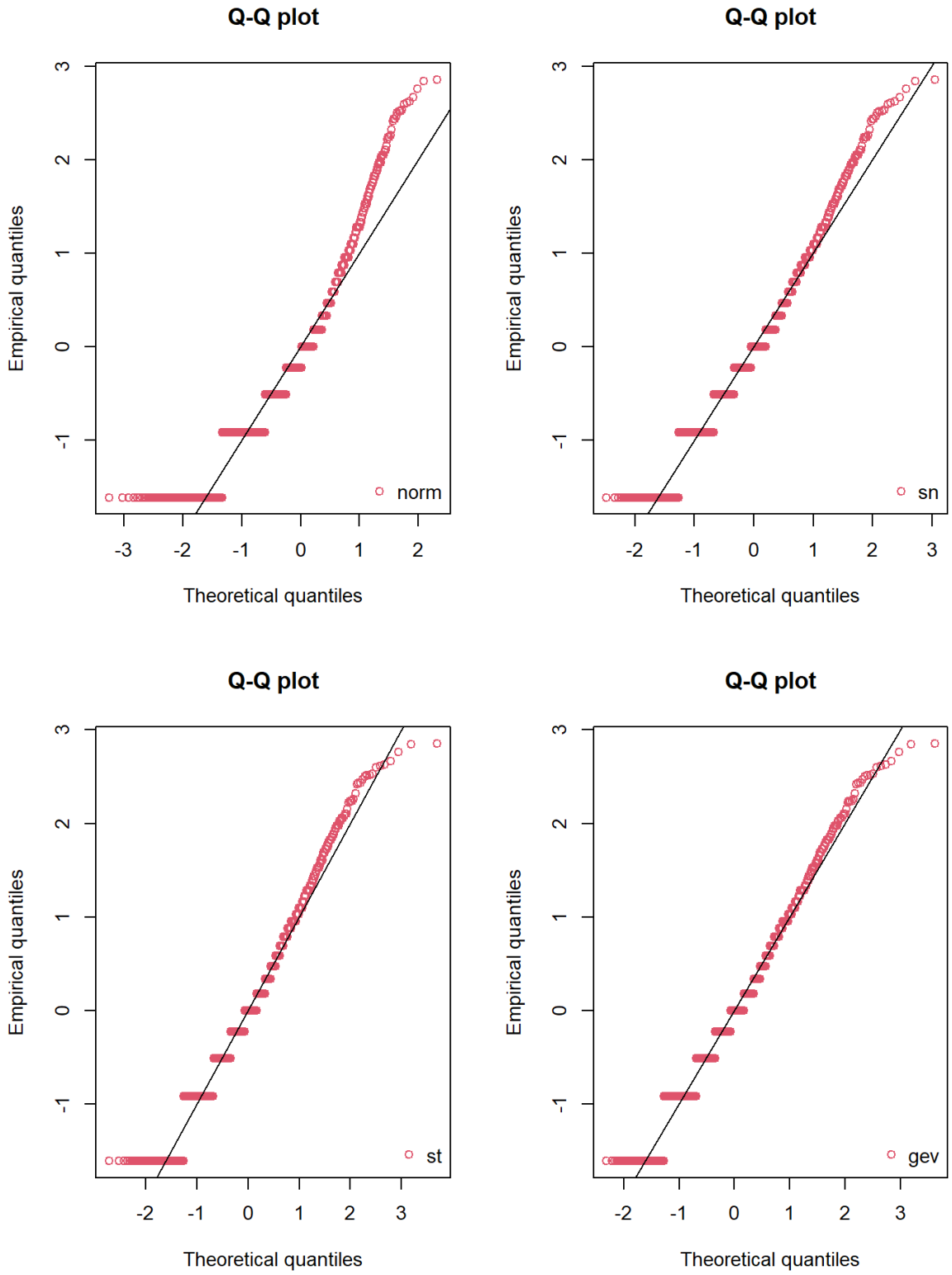


**Figure 3:** Q-Q plot for log(maximum intensity) with fitted distributions

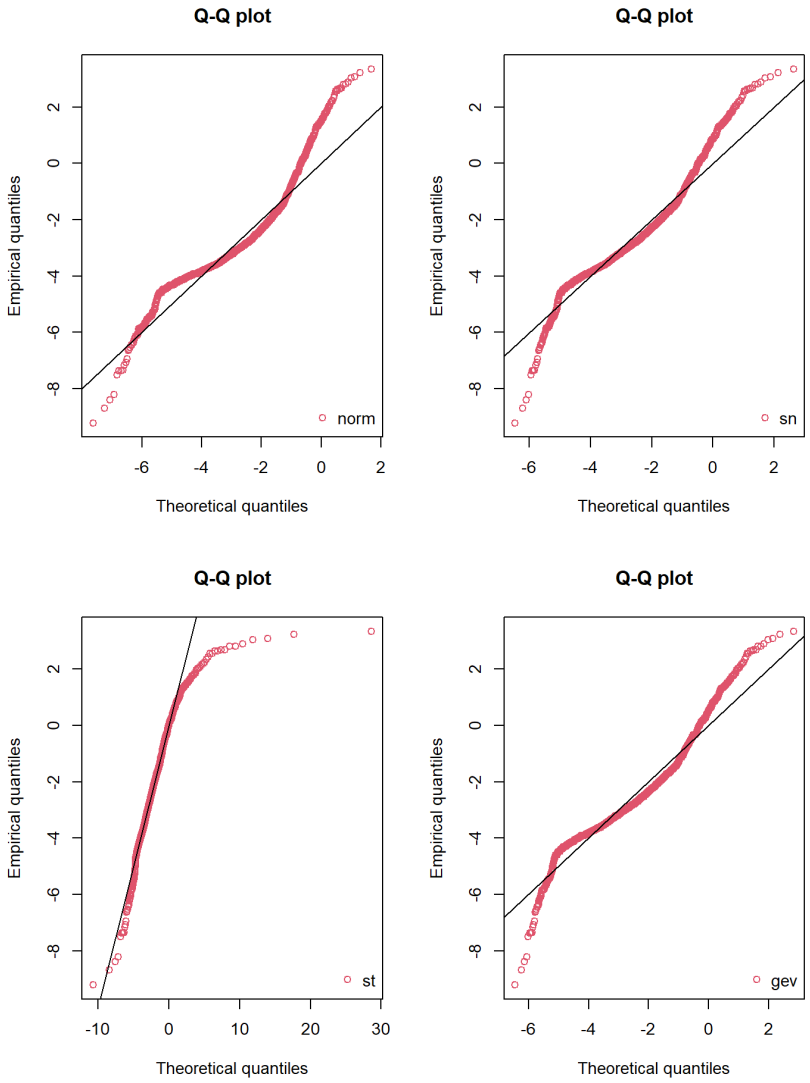


**Figure 4:** Q-Q plot for log(volatility) with fitted distributions

## Appendix B.1: Rainfall event with highest total rainfall (Event 9556)

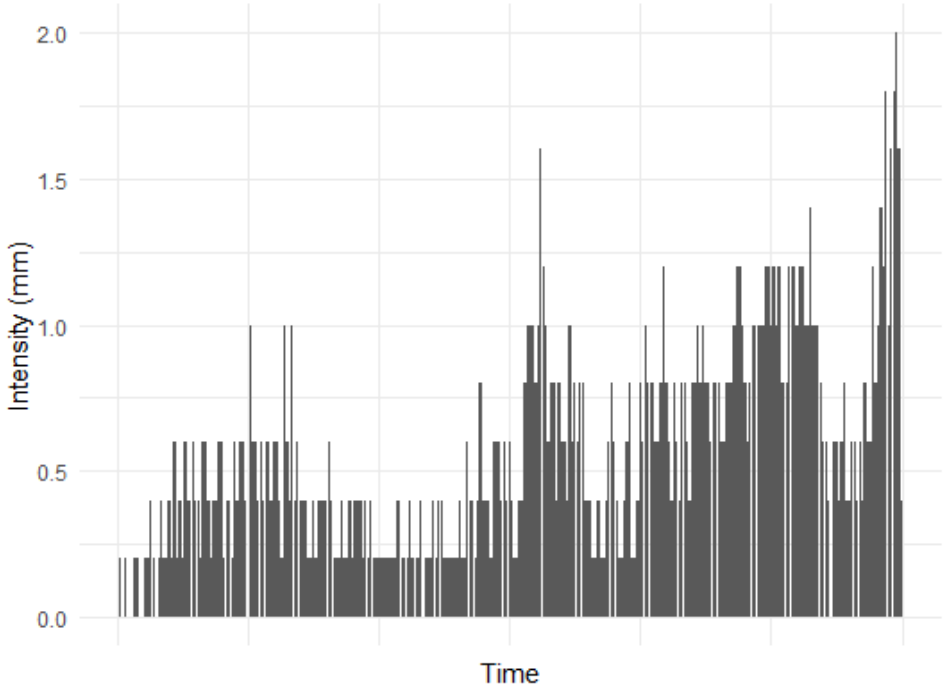


**Figure 5:** Rainfall event with highest total rainfall (Event 9556), start time: 2005-02-02 01:54:00; end time: 2005-02-03 07:48:00; duration: 1794 minutes; total intensity: 168.4mm

## Appendix B.2: Rainfall with largest intensity events

```
Original Event Number: 923 - Average Intensity: 17.4 mm/interval
Original Event Number: 2268 - Average Intensity: 14.4 mm/interval
Original Event Number: 22 - Average Intensity: 12.4 mm/interval
> events[[923]]
                  Datetime   Time Depth
228916 1979-03-18 08:36:00 228916  17.4
> events[[2268]]
                  Datetime   Time Depth
616063 1983-08-17 11:18:00 616063  14.4
> events[[22]]
                 Datetime Time Depth
4289 1976-08-24 09:54:00 4289  12.4
```

**Figure 6:** Rainfall with largest intensity events

## Appendix C: F-Exp-GPD

### C.1: Quantile Function of F-Exp-GPD

Given the cdf of the hybrid density, $h$ can be expressed as

$$H(x;\theta) = \begin{cases} r_1 F(x;\theta) & \text{if } -\infty < x < t_1 \\ r_1 F(t_1;\theta) + r_2(e^{-\lambda t_1} - e^{-\lambda x}) & \text{if } t_1 < x \leq t_2 \\ 1 - r_3(1 + \frac{\gamma}{\beta}(x - t_2))^{-1/\gamma} & \text{if } t_2 < x < \infty \end{cases} \tag{1}$$

Solving for the quantile function by

when $x \leq t_1$

Set $u = r_1 F(x,\theta)$

$$\begin{aligned} F(x,\theta) &= \frac{u}{r_1} \\ x &= F^{-1}\left(\frac{u}{r_1}\right) \end{aligned} \tag{2}$$

when $t_1 \leq x \leq t_2$

set $u = r_1 F(t_1,\theta) + r_2(e^{-\gamma t_1} - e^{-\gamma x})$; Let $u_1 = r_1 F(t_1,\theta)$

Then $u = u_1 + r_2(e^{-\gamma t_1} - e^{-\gamma x})$

$$\begin{aligned} \frac{u - u_1}{r_2} &= e^{-\gamma t_1} - e^{-\gamma x} \\ -e^{-\gamma x} &= \frac{u - u_1}{r_2} \\ e^{-\gamma x} &= \frac{u_1 - u + r_2 e^{-\gamma t_1}}{r_2} \\ -\gamma x &= \log\left(\frac{u_1 - u + r_2 e^{-\gamma t_1}}{r_2}\right) \\ x &= -\frac{1}{\gamma}\log\left(\frac{u_1 - u + r_2 e^{-\gamma t_1}}{r_2}\right) \\ &= \gamma^{-1}\log\left(\frac{r_2}{u_1 - u + r_2 e^{-\gamma t_1}}\right) \end{aligned} \tag{3}$$

when $x \geq t_2$

$$
\begin{aligned}
u &= 1 - r_3\Big(1+\frac{\gamma}{\beta}(x-t_2)\Big)^{-1/\gamma} \\
u-1 &= -r_3\Big(1+\frac{\gamma}{\beta}(x-t_2)\Big)^{-1/\gamma} \\
1-u &= r_3\Big(1+\frac{\gamma}{\beta}(x-t_2)\Big)^{-1/\gamma} \\
\frac{1-u}{r_3} &= \Big(1+\frac{\gamma}{\beta}(x-t_2)\Big)^{-1/\gamma} \\
(\frac{1-u}{r_3})^{-\gamma} &= 1+\frac{\gamma}{\beta}(x-t_2) \\
\Big(\frac{1-u}{r_3}\Big)^{-\gamma} - 1 &= \frac{\gamma}{\beta}(x-t_2) \\
\frac{\beta}{\gamma}\Big(\frac{1-u}{r_3}\Big)^{-\gamma} - 1 &= x-t_2 \\
x &= \Big[\frac{\beta}{\gamma}\Big(\frac{1-u}{r_3}\Big)^{-\gamma} - 1\Big] + t_2 \\
&= \Big[\frac{\beta}{\gamma}\Big(\frac{1-u-u_2}{r_3}\Big)^{-\gamma} - 1\Big] + t_2
\end{aligned}
\tag{4}
$$

$u_2 = 1 - r_3$

Then, the quantile function is given by

$$
\mathbf{H}^{-1}(u;\theta) = \begin{cases}
F^{-1}(\frac{u}{r_1};\theta) & \text{if } u \leq u_1 = r_1 F(t_1;\theta) \\
\frac{1}{\lambda}\log\Big[\frac{r_2}{u_1-u+r_2 e^{-\lambda t_1}}\Big] & \text{if } u_1 \leq u \leq u_2 = 1-r_3 \\
\frac{\beta}{\gamma}\Big[(1-\frac{u-u_2}{r_3})^{\gamma}-1\Big]+t_2 & \text{if } u \geq u_2
\end{cases}
\tag{5}
$$

## Appendix C.2: Functions for St-Exp-GPD

```
#=============================================================================
#=============================================================================
library(extraDistr)# contains some univariate distributions
library(sn)# contains the skew families
library(EnvStats)# contains some univariate distribution
#=============================================================================
######################## SKEW T-EXP-GPD hybrid model ############################
#=============================================================================
#derivative of the stew t density
#u1=threshold 1, u2=threshold 2,mu=mu,sigma=sigma, a=alpha,v=v
dstprime<-function(u1,mu,sigma,a,v){
  z=(u1-mu)/sigma
  p=a*z*((v+1)/(v+z^2))^(0.5)
  ti=dt(x=z,df=v)
  Ti=pt(q=p,df=v+1)
  tiprime=(-z*(v+1)/v)*((1+(z^2/v))^(-1)*ti)*Ti
  Tiprime=ti*a*((v+1)^(0.5))*(v/(v+z^2)^(3/2))*dt(p,df=v+1)
  Ttprime=(2/(sigma^2))*(tiprime+Tiprime)
  return(Ttprime)
}
#=============================================================================
#=============================================================================
#st_e_gpd pdf
dst_e_gpd<-function(x,u1,mu,sigma,a,v,u2,xi){
  d1=rep(0,length(x))
  lambda= -dstprime(u1,mu,sigma,a,v)/dst(u1,mu,sigma,a,v)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*pst(u1,mu,sigma,a,v)/dst(u1,mu,sigma,a,v)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dst(u1,mu,sigma,a,v)
  r3=r2*beta1*dexp(u2,lambda)
  d1[which(x<=u1)]<-r1*dst(x[which(x<=u1)],mu,sigma,a,v)
  d1[which(x<=u2 & x>u1)]<-r2*dexp(x[which(x<=u2 & x>u1)],lambda)
  d1[which(x>u2)]<-r3*dgpd(x[which(x>u2)]-u2,0,beta1,xi)
  return(d1)
}
#=============================================================================
```

```
#st_e_gpd cdf
pst_e_gpd<-function(q,u1,mu,sigma,a,v,u2,xi){
  p1=rep(0,length(q))
  lambda= -dstprime(u1,mu,sigma,a,v)/dst(u1,mu,sigma,a,v)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*pst(u1,mu,sigma,a,v)/dst(u1,mu,sigma,a,v)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dst(u1,mu,sigma,a,v)
  r3=r2*beta1*dexp(u2,lambda)
  p1[which(q<=u1)]<-r1*pst(q[which(q<=u1)],mu,sigma,a,v)
  p1[which(q<=u2 & q>=u1)]<- (r1*pst(u1,mu,sigma,a,v)+
                                r2*(pexp(q[which(q<=u2 & q>=u1)],lambda)
                                -pexp(u1,lambda)) )
  p1[which(q>=u2)]<- (r1*pst(u1,mu,sigma,a,v)+
                        r2*(pexp(u2,lambda)-pexp(u1,lambda))+
                        r3*pgpd(q[which(q>=u2)],u2,beta1,xi) )
  return(p1)
}
#=============================================================================
#st_e_gpd quantile
qst_e_gpd<-function(p,u1,mu,sigma,a,v,u2,xi){
  q1=rep(0,length(p))
  lambda= -dstprime(u1,mu,sigma,a,v)/dst(u1,mu,sigma,a,v)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*pst(u1,mu,sigma,a,v)/dst(u1,mu,sigma,a,v)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dst(u1,mu,sigma,a,v)
  r3=r2*beta1*dexp(u2,lambda)
  a1=r1*pst(u1,mu,sigma,a,v)
  b1=1-r3
  q1[which(p<=a1)]=qst(p[which(p<=a1)]/r1,mu,sigma,a,v)
  q1[which(p<=b1 & p>=a1)]=qexp(((p[which(p<=b1 & p>=a1)]-a1)/r2)+
                                  pexp(u1,lambda),lambda)
  q1[which(p>=b1)]=qgpd((p[which(p>=b1)]-b1)/r3,0,beta1,xi)+u2
  return(q1)
}
```

## Appendix C.3: R Functions for Sn-Exp-GPD

```
#============================================================================
#============================================================================
library(extraDistr)# contains some univariate distributions
library(sn)# contains the skew families
library(EnvStats)# contains some univariate distribution
#============================================================================
############################################################################
#################### SKEW NORMAL-EXP-GPD hybrid model ######################
############################################################################
#============================================================================
#derivative of the skew normal density
#u1=u1,mu=mu,sigma=sigma, a=alpha
dsnprime<-function(u1,mu,sigma,a){
  z=(u1-mu)/sigma
  s=a*z
  snprime= (2*dnorm(z,0,1)/sigma^2)*(a*dnorm(s,0,1)-z*pnorm(s,0,1))
  return(snprime)
}
#============================================================================
#sn_e_gpd pdf
dsn_e_gpd<-function(x,u1,mu,sigma,a,u2,xi){
  d2=rep(0,length(x))
  lambda= -dsnprime(u1,mu,sigma,a)/dsn(u1,mu,sigma,a)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*psn(u1,mu,sigma,a)/dsn(u1,mu,sigma,a)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dsn(u1,mu,sigma,a)
  r3=r2*beta1*dexp(u2,lambda)
  d2[which(x<=u1)]<-r1*dsn(x[which(x<=u1)],mu,sigma,a)
  d2[which(x<=u2 & x>u1)]<-r2*dexp(x[which(x<=u2 & x>u1)],lambda)
  d2[which(x>u2)]<-r3*dgpd(x[which(x>u2)]-u2,0,beta1,xi)
  return(d2)
}
#============================================================================
#sn_e_gpd cdf
psn_e_gpd<-function(q,u1,mu,sigma,a,u2,xi){
  p2=rep(0,length(q))
```

```
  lambda= -dsnprime(u1,mu,sigma,a)/dsn(u1,mu,sigma,a)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*psn(u1,mu,sigma,a)/dsn(u1,mu,sigma,a)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dsn(u1,mu,sigma,a)
  r3=r2*beta1*dexp(u2,lambda)
  p2[which(q<=u1)]<-r1*psn(q[which(q<=u1)],mu,sigma,a)
  p2[which(q<=u2 & q>=u1)]<- (r1*psn(u1,mu,sigma,a)+
                                r2*(pexp(q[which(q<=u2 & q>=u1)],lambda)
                                - pexp(u1,lambda)))
  p2[which(q>=u2)]<- (r1*psn(u1,mu,sigma,a)+
                        r2*(pexp(u2,lambda)-pexp(u1,lambda))+
                        r3*pgpd(q[which(q>=u2)],u2,beta1,xi))
  return(p2)

}
#=============================================================================
#sn_e_gpd quantile function
qsn_e_gpd<-function(p,u1,mu,sigma,a,u2,xi){
  q2=rep(0,length(p))
  lambda= -dsnprime(u1,mu,sigma,a)/dsn(u1,mu,sigma,a)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*psn(u1,mu,sigma,a)/dsn(u1,mu,sigma,a)))*
  exp(-lambda*u1)

        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dsn(u1,mu,sigma,a)
  r3=r2*beta1*dexp(u2,lambda)
  a1=r1*psn(u1,mu,sigma,a)
  b1=1-r3
  q2[which(p<=a1)]=qsn(p[which(p<=a1)]/r1,mu,sigma,a)
  q2[which(p<=b1 & p>=a1)]=qexp(((p[which(p<=b1 & p>=a1)]-a1)/r2)+
  pexp(u1,lambda),lambda)
  q2[which(p>=b1)]=qgpd((p[which(p>=b1)]-b1)/r3,0,beta1,xi)+u2
  return(q2)
}
```

## Appendix C.4: R Functions for GEV-Exp-GPD

```
#============================================================================
#============================================================================
library(extraDistr)# contains some univariate distributions
library(sn)# contains the skew families
library(EnvStats)# contains some univariate distribution
library(stats)
#============================================================================
#################### GEV-EXP-GPD hybrid model ################################
#============================================================================
#derivative of the GEV density
dgevprime<-function(u1,mu,sigma,k){
  z=(u1-mu)/sigma
  v<-(1+k*z)^(-1/k)
  vprime<-(-1/sigma)*(1+k*z)^((-1/k)-1)
  gevprime= vprime*dgev(u1,mu,sigma,k)*((k+1)*(v)^(-1)-1)
  return(gevprime)
}
#============================================================================
#gev_e_gpd pdf
dgev_e_gpd<-function(x,u1,mu,sigma,k,u2,xi){
  d3=rep(0,length(x))
  lambda= -dgevprime(u1,mu,sigma,k)/dgev(u1,mu,sigma,k)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*pgev(u1,mu,sigma,k)/dgev(u1,mu,sigma,k)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dgev(u1,mu,sigma,k)
  r3=r2*beta1*dexp(u2,lambda)
  d3[which(x<=u1)]<-r1*dgev(x[which(x<=u1)],mu,sigma,k)
  d3[which(x<=u2 & x>u1)]<-r2*dexp(x[which(x<=u2 & x>u1)],lambda)
  d3[which(x>u2)]<-r3*dgpd(x[which(x>u2)]-u2,0,beta1,xi)
  return(d3)
}
#============================================================================
#gev_e_gpd cdf
pgev_e_gpd<-function(q,u1,mu,sigma,k,u2,xi){
  p3=rep(0,length(q))
  lambda= -dgevprime(u1,mu,sigma,k)/dgev(u1,mu,sigma,k)
```

```
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*pgev(u1,mu,sigma,k)/dgev(u1,mu,sigma,k)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dgev(u1,mu,sigma,k)
  r3=r2*beta1*dexp(u2,lambda)
  p3[which(q<=u1)]<-r1*pgev(q[which(q<=u1)],mu,sigma,k)
  p3[which(q<=u2 & q>=u1)]<- (r1*pgev(u1,mu,sigma,k)+
                                r2*(pexp(q[which(q<=u2 & q>=u1)],lambda)
                                - pexp(u1,lambda)) )
  p3[which(q>=u2)]<-( r1*pgev(u1,mu,sigma,k)+
                         r2*(pexp(u2,lambda)-pexp(u1,lambda))+
                         r3*pgpd(q[which(q>=u2)],u2,beta1,xi) )
  return(p3)
}
#=============================================================================
#gev_e_gpd quantile
qgev_e_gpd<-function(p,u1,mu,sigma,k,u2,xi){
  q3=rep(0,length(p))
  lambda= -dgevprime(u1,mu,sigma,k)/dgev(u1,mu,sigma,k)
  beta1= (xi+1)/lambda
  r2=1/((1+(lambda*pgev(u1,mu,sigma,k)/dgev(u1,mu,sigma,k)))*exp(-lambda*u1)
        +(lambda*beta1-1)*exp(-lambda*u2))
  r1=r2*dexp(u1,lambda)/dgev(u1,mu,sigma,k)
  r3=r2*beta1*dexp(u2,lambda)
  a1=r1*pgev(u1,mu,sigma,k)
  b1=1-r3
  q3[which(p<=a1)]=qgev(p[which(p<=a1)]/r1,mu,sigma,k)
  q3[which(p<=b1 & p>=a1)]=qexp(((p[which(p<=b1 & p>=a1)]-a1)/r2)+
                                   pexp(u1,lambda),lambda)
  q3[which(p>=b1)]=qgpd((p[which(p>=b1)]-b1)/r3,0,beta1,xi)+u2
  return(q3)
}
```

# Appendix D: R Functions for Copula Simulation

## Appendix D.1: Bivariate Copula Simulation

```
rm(list = ls())
#load the VineCopula package
library(VineCopula)

##Loading Event Summary Data
load("C:/Users/c1834516/OneDrive - Cardiff University/
     Documents/Research Data and Code/Simulation/Rain_Events_OS.RData")
View(Rain_Events_OS)

#Selecting DIMV
data=Rain_Events_OS[, c("Duration","Intensity",
                        "Max_Intensity","Volatility")]

#Selecting D & I then taking the log in order to
#effectively model the data
ndata<-cbind(log(data[,1]),log(data[,2]))

#Converting the data to Univariate U[0,1]
udata<-pobs(cbind(ndata[,1],ndata[,2]))

#Selecting the Best fitted Copula using the AIC criteria
cop<-BiCopSelect(udata[,1],udata[,2])
cop
#Bivariate copula: Rotated Tawn type 1 90 degrees
#(par = -4.69, par2 = 0.32, tau = -0.29)

# Simulate from the proposed Copula
set.seed(1234)  # For reproducibility
n <- 3450       # Number of samples
simData <- BiCopSim(n, family= 124, par=-4.69, par2=0.32)

#Loading the needed library for the transformation
library(fitdistrplus)
```

```
library(sn)
library(VGAM)


#Converting simData from U[0,1] to original scale
#Using skew t dist for log(duration)
D <- (qst(simData[,1], xi = 5.256585, omega = 1.135632,
          alpha = -1.145267, nu = 18.396395))

#Using GEV for log(intensity)
I <- (qgev(simData[,2], location = -2.07344987,
           scale = 0.67027053, shape = -0.04415461))

# Visualization for Comparism of Simulated data and Observed Data
png(filename = "Observed data vs Copula Simulated.png",
    height=24,width=18,units="cm",res=200)
plot(D,I, main = '', xlab = "log(duration)", ylab = "log(intensity)",
     col='darkgray',ylim=c(-4,3), xlim=c(0,8),
     lwd=4, pch = 19, cex = 0.5)
points(ndata[,1],ndata[,2], col="black")
legend('bottomleft', c('simulated','observed'), lwd=3,
       col= c('darkgray','black'))
dev.off()

#Table 6.3 formation
#Normal Copula
Est1<-BiCopEst(udata[,1],udata[,2], family=1, method = "mle")
summary(Est1)

#t Copula
Est2<-BiCopEst(udata[,1],udata[,2], family=2, method = "mle")
summary(Est2)

#Rotated 270 degree Clayton Copula
Est3<-BiCopEst(udata[,1],udata[,2], family=33, method = "mle")
summary(Est3)

#Frank Copula
Est4<-BiCopEst(udata[,1],udata[,2], family=5, method = "mle")
```

```
summary(Est4)

#Rotated 90 degree Joe Copula
Est5<-BiCopEst(udata[,1],udata[,2], family=26, method = "mle")
summary(Est5)

#Rotated 90 Tawn type 1 copula
Est6<-BiCopEst(udata[,1],udata[,2], family=124, method = "mle")
summary(Est6)
```

## Appendix D.2: Vine Copula Simulation for DIMV

```
## Loading Appropriate Copula Package
library(VineCopula)

##Compute Pseudo-observations for Copula Inference
udata <- pobs(log(data))

#Selecting Appropriate Tree Structure
rvm=RVineStructureSelect(data=udata)

#Copula pair selection multivariate
pair_cop=RVineCopSelect(data=udata, familyset = NA,
                        Matrix =rvm$Matrix, selectioncrit = "AIC",
                        indeptest = FALSE, level = 0.05)



#Simulating from the multivariate copula
RVM =RVineMatrix(Matrix = pair_cop$Matrix, family = pair_cop$family,
                 par= pair_cop$par,
                 par2 = pair_cop$par2, names = c("D", "I", "M", "V"))

DIMV_sim<-function(){
# Initialize a variable to keep track of the condition
condition_met <- FALSE

# Run the simulation until the condition is met
while (!condition_met) {
  # Simulating from the multivariate copula
  simdata <- RVineSim(1, RVM)
  # Check dimensions of simdata
  #print(dim(simdata))

  # Converting to initial scale

  # Convert simdata[1] to original scale using skew #t distribution
  Duration <- exp(qst(simdata[1], xi = 5.256585,
                      omega = 1.135632, alpha = -1.145267, nu = 18.396395))
```

```
  # Convert simdata[2] to original scale using GEV #distribution
  Intensity <- exp(qgev(simdata[2],
                        location = -2.07344987, scale = 0.67027053,
                        shape = -0.04415461))

  # Convert simdata[3] to original scale using GEV #distribution
  Max_Intensity <- exp(qgev(simdata[3],
                            location = -0.80075293, scale = 0.64562950,
                            shape = -0.06081438))

  # Convert simdata[4] to original scale using  #skew t distribution
  Volatility <- exp(qst(simdata[4], xi = -3.9269208,
                        omega = 0.9956432, alpha = 1.7540797,
                        nu = 2.5856904))

  # Check if the conditions are met
  #(simdata$Max_Intensity >= (simdata$Intensity/simdata$Duration)) and
  #(Volatility >= 0)
  if (Max_Intensity <= Intensity * Duration &&
      Max_Intensity >= Intensity && Volatility >= 0) {
    condition_met <- TRUE
  }
}
return(c(Duration, Intensity, Max_Intensity, Volatility))
}

DIMV <- matrix(nrow=3450, ncol=4)
for (I in 1:3450) DIMV[I,] <- DIMV_sim()

# Visualization for Comparison of Simulated Data and Observed Data
png(filename = "Q-Q plot of observed data vs Copula Simulated.png",
    height=24,width=18,units="cm",res=200)
par(mfrow=c(4,2))
plot(sort(DIMV[,1]), sort(data[,1]), xlab="sort(simulated duration)",
     ylab="sort(duration)")
abline(a=0,b=1,col="red")
plot(sort(DIMV[,2]), sort(data[,2]), xlab="sort(simulated intensity)",
     ylab="sort(intensity)")
```

```
abline(a=0,b=1,col="red")
plot(sort(DIMV[,3]), sort(data[,3]),
     xlab="sort(simulated maximum intensity)",
     ylab="sort(maximum intensity)")
abline(a=0,b=1,col="red")
plot(sort(DIMV[,4]), sort(data[,4]), xlab="sort(simulated volatility)",
     ylab="sort(volatility)")
abline(a=0,b=1,col="red")
dev.off()
```

# Appendix E: R Functions for Rain Events Simulation

## Appendix E.1: R Code for Simulating Simulating a Rain Event

```
rm(list=ls())

# Function for Vol(x)
Vol <- function(x) {
  sum(diff(x)^2)/length(x)
}



raineventsim2 <- function(I, D, M, V, stepsize = 6, tol = 0.001,
                          alpha = 2, beta = 2, max_it = 100,
                          j = -1, num_calls = 0) {
  # comments: what are the inputs, what is the output, what is the method

  n <- ceiling(D/stepsize)
  x <- rep(NA, n)

  if (num_calls >=2) return(list(x = x, state = "failure"))

  # case n == 1
  if (n == 1) {
    if (abs(M - (n * I)) >= tol) {
      return(list(x = x, state = "raineventsim: inconsistent I, D, M, V, n == 1"))
    } else {
      x[1] = M
      return(list(x = x, state = "success"))
    }
  }

  if (n == 2) {
    if (runif(1) < 0.5) {
      x <- c(M, 2*I - M)
    } else {
      x <- c(2*I - M, M)
```

```
  }
  if (abs(Vol(x) - V) <= tol && min(x) >= 0) {
    return(list(x = x, state = "success"))
  } else {
    return(list(x = x, state = "raineventsim:
    inconsistent I, D, M, V, n == 2"))
  }
}
# # choose case when v==0
# if (V==0){
# x<-rep(M,length(x))
# return(list(x=x, state= "success"))
# }

# choose location of max value
if (j == -1) {
    j <- ceiling(rbeta(1, alpha, beta) * n)
}

x[j] <- M

# initial solution satisfying I, D, M constraints
for (i in (1:n)[-j]) {
  x[i] <- ((n * I) - M) / (n - 1)
}
# check min max value constraints
if (min(x) < 0 || max(x) > M + sqrt(.Machine$double.eps)) {
  return(list(x = x, state = paste0("raineventsim: inconsistent I = ", I,
  ", D = ", D, ", M = ", M, ", V =", V, ", n >= 3")))
}

# ============================================================
# random search for better volatility, assuming n >= 3
num_reps <- 0
fail_count <- 0
while (fail_count < max_it) {
  num_reps <- num_reps + 1
  V_err <- abs(Vol(x) - V)
  if (V_err <= tol) {
```

```
    return(list(x = x, state = "success"))
  }
  hi <- sample((1:n)[-j], 2)
  h <- hi[1]
  i <- hi[2]
  eps_max <- min(x[h], M - x[h], x[i], M - x[i])
  eps <- runif(1, -eps_max, eps_max)
  y <- x
  z <- x
  y[h] <- y[h] - eps
  y[i] <- y[i] + eps
  z[h] <- z[h] + eps
  z[i] <- z[i] - eps
  V_err_y <- abs(Vol(y) - V)
  V_err_z <- abs(Vol(z) - V)
  if (V_err_y < V_err || V_err_z < V_err) {
    fail_count <- 0
    if (V_err_y < V_err_z) {
      x <- y
      V_err <- V_err_y
    } else {
      x <- z
      V_err <- V_err_z
    }
  } else {
    fail_count <- fail_count + 1
  }
}

#print(Vol(x) - V)

if (!(j == 1 || j == n)) {
  j <- sample(c(1, n), 1)
  return(raineventsim2(I, D, M, V, stepsize, tol, alpha, beta, max_it, j = j,
  num_calls = num_calls + 1))
} else {
  j <- sample(2:(n-1), 1)
  return(raineventsim2(I, D, M, V, stepsize, tol, alpha, beta, max_it, j = j,
  num_calls = num_calls + 1))
```

```
  }
  return(list(x = x, state = "raineventsim: maximum iterations exceeded",
  num_reps = num_reps))
}
```

## Appendix E.2: R Code for Testing Rainevent Simulator using observed DIMV

```
rm(list = ls())
#Load the rainfall events data
load(file="Rain_Events_OS.RData") #Rain_Events_OS
View(Rain_Events_OS)

source("raineventsim.R")#script for rain simulation

#input data frame
df<-data.frame(I=Rain_Events_OS$Intensity,D=Rain_Events_OS$Duration,
               M=Rain_Events_OS$Max_Intensity,V=Rain_Events_OS$Volatility)

#loop over each row of the dataframe
x_sim <- list()
errors <-0
for (i in 1:nrow(df)) {
  #get the current values of I,D,M,V
  c_I<-df$I[i]
  c_D<-df$D[i]
  c_M<-df$M[i]
  c_V<-df$V[i]

  #perform the simulation
  x <- raineventsim2(I=c_I, D=c_D, M=c_M, V=c_V, max_it = 100)
  if (x$state != "success") {
    errors <- errors + 1
    x_sim[[i]] <- NA
    cat(i, x$state, "\n")
  } else {
    x_sim[[i]] <- x$x
  }
}
errors
#[1] 0
```

## Appendix E.3: R Code for Simulating a Sequence of Rainfall Events

```
## Simulating list of Rainfall Events
rm(list = ls())



#loading Auxilliary Functions
#script for generating time, IAT & \lambda
source("Auxilliary function for time.R")
source("Vine Copula Model.R") #script for generating D,I,M,V
source("raineventsim.R") #script for rain simulation

start <- as_datetime(0)
end <- start + dyears(1) + dminutes(0)
#time.df <- time.sequence.information(start,end)
# head(time.df)
# tail(time.df)




load(file="Rain_Events_OS.RData") #Rain_Events_OS
# #Simulation function for Rainfall Events using IAT
lambda_hat_hours = 1/(tapply(Rain_Events_OS$IAT,
                             Rain_Events_OS$Months, mean, na.rm=TRUE) - 1)
lambda_hat_hours



SimRF_List <- function(start_time, end_time, rainsim = raineventsim3, stepsize = 6) {
  # Simulating Rainfall Events using IAT
  # stepsize in minutes
  # start_time and end_time as POSIX objects (count time in seconds)
  sim_time <- start_time

  X <- list()
  s <- c()
  t <- c()

  event_counter <- 0
```

```
  while (sim_time < end_time) {
    print(sim_time)

    event_counter <- event_counter + 1

    # current event
    result <- rainsim()
    D <- length(result)*stepsize*60
    s[event_counter] <- sim_time
    t[event_counter] <- sim_time + D
    X[[event_counter]] <- result

    # time for next event
    sim_time <- sim_time + D # time previous event finishes
    m <- month(sim_time)
    #print(m)
    r <- 3600 + rexp(1, (lambda_hat_hours[m] / 3600))
    #print(r)
    r <- ceiling(r/(stepsize*60)) * stepsize * 60  # update r
    sim_time <- sim_time + r
  }
  return(list(X = X, s = s, t = t))
}

Event_list <- SimRF_List(start_time = start, end_time = end)
View(Event_list)

all_intensities <- rep(0, (as.numeric(end) - as.numeric(start))/60/6)
for (i in 1:length(Event_list$X)) {
  j <- (Event_list$s[i] - as.numeric(start))/60/6
  k <- (Event_list$t[i] - as.numeric(start))/60/6
  print(c(k - j, length(Event_list$X[[i]])))
  all_intensities[(j+1):k] <- Event_list$X[[i]]
}

plot((1:1000)/240, all_intensities[1:1000], type="s") # time in days

plot((1:2000)/240, all_intensities[1:2000], type="s") # time in days
```

## Appendix E.5: source("Auxilliary function for time.R") in E.4

```
rm(list = ls())

#function for start and end time

## auxiliary function
library(lubridate)

time.sequence.information <- function(start,end){
  require(lubridate)
  times  <- seq(start, end, by="1 min")
  ## derived quantities with lubridate functions
  year   <- year(times)
  ymonth <- month(times)     ## month in the year  (starting with 1)
  yweek  <- week(times)      ## week in the year   (starting with 1)
  yday   <- yday(times)      ## day in the year    (starting with 1)
  dhour  <- hour(times)      ## hour in the day    (starting with 0)
  hmin   <- minute(times)    ## minute in the hour (starting with 0)
  ## derived quantities beyond lubridate functions
  dmin   <- dhour * 60 + hmin         ## minute in the day
  ymin   <- (yday-1) * 24 * 60 + dmin ## minute in the year
  ## create data frame storing this information
  times.df <- data.frame(time=times, year=year,
                         ymonth=ymonth, yweek=yweek, yday=yday, ymin=ymin,
                         dhour=dhour, dmin=dmin,
                         hmin=hmin)
  stopifnot(as.integer(difftime(end,start,units="mins"))+1==dim(times.df)[1])
  return(times.df)
}


#Load the rainfall events data
#load("C:/Users/c1834516/OneDrive - Cardiff University/Documents
/Research Data and Code/Simulation/Rain_Events_OS.RData") #Rain_Events_OS
load(file="Rain_Events_OS.RData") #Rain_Events_OS
#View(Rain_Events_OS)

# #Simulation function for Rainfall Events using IAT
```

```
lambda_hat_hours = 1/(tapply(Rain_Events_OS$IAT,
                             Rain_Events_OS$Months, mean, na.rm=TRUE) - 1)
lambda_hat_hours
```

## Appendix E.6: source("Vine Copula Model.R") in E.4

```
# Clear the environment
rm(list = ls())

############################################################
## Loading Appropriate Copula Package
library(VineCopula)
library(sn)
library(VGAM)

# Matrix for D-Vine tree structure
Matrix <- c(1, 4, 3, 2, 0, 2, 4, 3, 0, 0, 3, 4, 0, 0, 0, 4)
Matrix <- matrix(Matrix, 4, 4)

# Selected Copula families
family <- c(0, 40, 17, 234, 0, 0, 204, 1, 0, 0, 0, 204, 0, 0, 0, 0)
family <- matrix(family, 4, 4)

# Selected families Parameter 1
par1 <- c(0, -3.10604485, 0.09710537, -4.68930339, 0, 0, 2.1508557, 0.7618961,
            0, 0, 0, 3.16914, 0, 0, 0, 0)
par1 <- matrix(par1, 4, 4)

# Selected families Parameter 2
par2 <- c(0, -0.5414058, 1.4199392, 0.3232532,
0, 0, 0.2789502, 0, 0, 0, 0, 0.8501189, 0, 0, 0, 0)
par2 <- matrix(par2, 4, 4)

# Define RVineMatrix Object
RVM <- RVineMatrix(Matrix = Matrix, family = family, par = par1, par2 = par2,
                     names = c("D", "I", "M", "V"))



# Function for simulation
sim_data1 <- function(n) {
  # Create an empty matrix to store the results
  result <- matrix(NA, nrow = n, ncol = 4)
```

```
  for (i in 1:n) {
    simdata <- RVineSim(1, RVM)

    # Converting to initial scale
    # Convert simdata[1] to original scale using skew t distribution
    Duration <- exp(qst(simdata[1], xi = 5.256585, omega = 1.135632,
    alpha = -1.145267, nu = 18.396395))

    # Convert simdata[2] to original scale using GEV distribution
    Intensity <- exp(qgev(simdata[2], location = -2.07344987,
    scale = 0.67027053, shape = -0.04415461))

    # Convert simdata[3] to original scale using GEV distribution
    Max_Intensity <- exp(qgev(simdata[3], location = -0.80075293,
    scale = 0.64562950, shape = -0.06081438))

    # Convert simdata[4] to original scale using skew t distribution
    Volatility <- exp(qst(simdata[4], xi = -3.9269208,
    omega = 0.9956432, alpha = 1.7540797, nu = 2.5856904))

    # Store the results in the result matrix
    result[i, ] <- c(Duration, Intensity, Max_Intensity, Volatility)
  }

  # Convert the matrix to a data frame and return
  return(result)
}
```

## Appendix E.7: source("raineventsim.R") in E.4

```
#Same as E2 with the addition below
raineventsim3 <- function() {
  state <- "failure"
  while (state != "success") {
    theta <- sim_data1(1)
    out <- raineventsim2(theta[2], theta[1], theta[3], theta[4])
    state <- out$state
    #print(state)
  }
  return(out$x)
}
```