

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

A Stochastic Infilling Algorithm for Spatial-Temporal Rainfall Data

A thesis presented in partial fulfillment of the requirements
for the Degree of

Master of Science
in
Statistics

at Massey University, Albany, New Zealand

David Russell Munroe

2005

ABSTRACT

The purpose of this thesis is to develop an infilling algorithm for 24-hour (daily) rainfall data. An infilling algorithm replaces missing data within the historical records with sensible estimates, where any appropriate method (prediction from a fitted model, interpolation between points, or random sampling) could be used to select and/or produce the required estimates. The algorithm developed uses simulation data generated using a stochastic point-process model which has been fitted to historical data. In this thesis, the spatial-temporal Neyman-Scott rectangular pulse model as presented in Cowpertwait et al. (2002) is fitted to data provided by Thames Water from 23 sites in the Thames Valley (UK). The model is shown to fit the data reasonably well; however it fails to fit the proportion of dry sites (which is not used in the fitting process). Nevertheless, simulated data is generated using the model and an infilling algorithm is derived. The algorithm is tested by replacing valid historical data with missing values, infilling these missing values, and then comparing relevant statistics for the two samples. Three algorithms are developed in this thesis, of which the final algorithm maintains the statistical characteristics of the historical data, including the proportion of dry sites, while infilling values that are similar to the known historical record.

Preface

In general, only a sample of the plots produced for any given analysis are included in text within this thesis. This sampling is both for brevity and clarity. Further plots are generally included within the appendices at the end of the thesis.

Note for the spatial data analysis, approximately 700 figures were generated. As only ten plots led to a definite conclusion, only these plots have been included in the Appendices.

Furthermore, for consistency, and to make comparison easier, the same months (January and July) were used to represent any seasonal differences where applicable. However, considerable variation occurs between the seasons and the results for the other months included in the supplementary appendices should not be overlooked.

Acknowledgements

I would like to thank my supervisor, Paul Cowpertwait, for his guidance and stimulating discussions throughout this project. In addition, the use of Bruce Mill's bitmap generating code was most appreciated for the production of the simulation movie.

I would also like to thank my family and friends for their support and prayers during this period. This is much appreciated as without them this work would not have been possible.

Finally, I am grateful to Thames Water for permitting the use of their data.

CONTENTS

1. <i>Introduction</i>	1
1.1 Background	1
1.2 Source data	2
1.2.1 Measurement error	4
1.3 Thesis outline	4
2. <i>Literature review</i>	5
2.1 Stochastic rainfall models	5
2.1.1 Temporal models	5
2.1.2 Spatial-temporal models	8
2.1.3 Applied model	11
2.2 Infilling	11
2.2.1 Missing data assumptions	13
2.2.2 Algorithms	13
2.2.3 Summary	15
3. <i>Methodology</i>	17
3.1 Spatial-temporal NSRP model	17
3.1.1 Assumptions	19
3.1.2 Model notation	20
3.1.3 Mathematical description	22
3.1.4 Sample statistic calculations	24
3.1.5 Model fitting	26

3.1.6	Verification	28
3.2	Data analysis	29
3.2.1	Exploratory plots: time	30
3.2.2	Exploratory plots: spatial	30
3.2.3	Data removal	31
3.2.4	Assumptions	32
3.3	Infilling	33
3.3.1	Algorithm evaluation	33
3.3.2	Notation	35
3.3.3	Best fit least squares	36
3.3.4	Best fit CDF least squares	37
3.3.5	Iterative least squares	38
3.3.6	Further application	38
3.4	Implementation	39
4.	<i>Data Integrity Analysis</i>	41
4.1	Introduction	41
4.1.1	Known issues	42
4.2	Exploratory analysis: temporal	43
4.2.1	TW238097	43
4.2.2	TW238605	44
4.2.3	TW239258	44
4.2.4	TW239320	44
4.2.5	TW239374	47
4.2.6	TW239578	47
4.2.7	TW245176	47
4.2.8	TW246424	48
4.2.9	TW246627	48
4.2.10	TW246847	48

4.2.11	TW247119	48
4.2.12	TW286392	49
4.2.13	TW287141	49
4.2.14	TW287283	49
4.2.15	TW288020	49
4.2.16	TW289022	50
4.2.17	TW290007	50
4.2.18	TW291467	50
4.3	Exploratory analysis: spatial	50
4.3.1	TW238097	51
4.3.2	TW246424	51
4.3.3	TW246627	53
4.3.4	TW247119	53
4.3.5	TW287141	53
4.3.6	TW287283	53
4.3.7	TW287864	54
4.3.8	TW289022	54
4.3.9	TW291467	54
4.4	Summary statistics	54
4.4.1	Valid data	55
4.4.2	Temporal stationarity	57
5.	<i>Results</i>	59
5.1	Introduction	59
5.2	Model fitting	60
5.2.1	Introduction	60
5.2.2	Parameter estimation	61
5.3	Model validation	71
5.3.1	Fitted statistics	71

5.3.2	Monthly statistics	81
5.3.3	Stability	91
5.3.4	Summary	96
5.4	Fitting algorithm heuristics	96
5.4.1	Partitioning of wet/dry days	96
5.5	Infilling	102
5.5.1	Introduction	102
5.5.2	Best fit least squares	104
5.5.3	Best fit CDF least squares	114
5.5.4	Iterative sampling CDF least squares	124
5.5.5	Comparison of algorithms	133
5.5.6	Other infilling derivations	134
6.	<i>Conclusions</i>	139
6.1	Data analysis	139
6.2	Model	139
6.2.1	Issues	140
6.2.2	Simulation movie	141
6.3	Infilling	142
6.3.1	Best fit algorithms	143
6.3.2	Iterative sampling algorithms	144
6.4	Conclusions	145
6.5	Future research	146
6.5.1	Internal algorithms	146
6.5.2	General improvements	150
	<i>Bibliography</i>	154
A.	<i>Data Integrity Analysis: Plots and Tables</i>	163
A.1	Temporal plots	163

A.2 Spatial plots	208
B. Model Fitting: Plots	215
C. Model Validation: Plots	231
D. Infilling plots	255

LIST OF TABLES

1.1	Site location coordinates: Easting, Northing, and Altitude . . .	3
3.1	Model notation	21
3.2	Statistic notation	24
3.3	Infilling notation	35
4.1	Percentage of valid data remaining within the historical record	56
5.1	Historical pooled statistics: raw and smoothed	62
5.2	<i>Model_A</i> Monthly parameter estimates	67
5.3	<i>Model_A</i> Scale parameter estimate $\hat{\theta}_{ik}(mm)$ for each Site-Month	67
5.4	<i>Model_B</i> Monthly parameter estimates	68
5.5	<i>Model_B</i> Scale parameter estimate $\hat{\theta}_{ik}(mm)$ for each Site-Month	68
5.6	1-hour and 24-hour: regional proportion dry by season	80
5.7	Kolmogorov-Smirnov test p-values: monthly mean simulated versus historical	83
5.8	Kolmogorov-Smirnov test p-values: monthly CV simulated versus historical	85
5.9	Kolmogorov-Smirnov test p-values: monthly skewness simu- lated versus historical	87
5.10	Kolmogorov-Smirnov test p-values: monthly autocorrelation simulated versus historical	88
5.11	Partitioning on wet sites: 24-hour aggregation level	99
5.12	24-hour historical and simulated: Dry, Some Dry, and Wet . .	101

5.13	ISCDF, BFLS, BFCDF: proportion test on the overestimation of historical, H , statistics	135
5.14	ISCDF, BFLS, BFCDF: P-Values for KS test	136

LIST OF FIGURES

1.1	Map of site locations in the Thames catchment	2
3.1	Temporal Neyman-Scott model	18
3.2	Spatial-temporal Neyman-Scott model	18
3.3	Scatterplot selection algorithm	31
3.4	BFLS: infilling algorithm definition	36
3.5	BFCDFLS: infilling algorithm definition	37
3.6	ISCDFLS: infilling algorithm definition	39
4.1	Site TW238097 daily plots	45
4.2	Site TW238097 hourly plots	46
4.3	Daily data, March, site TW238283 versus site TW238097 . . .	52
4.4	Daily data, December, site TW246424 versus site TW238578 .	52
4.5	Correlogram: Deseasonalised monthly means	57
4.6	Correlogram: January deseasonalised monthly means	58
5.1	Model fit: 1-hour aggregation level	63
5.2	Model fit: 6-hour aggregation level	64
5.3	Model fit: 24-hour aggregation level	65
5.4	$Model_B$ cross-correlation Jan,July	70
5.5	35 year simulation: 1-hour aggregation level	72
5.6	35 year simulation: 6-hour aggregation level	73
5.7	35 year simulation: 24-hour aggregation level	74
5.8	Simulation cross-correlation Jan,July	76

5.9	Quantile-Quantile plots: January <i>Model_A</i> , <i>Model_B</i>	78
5.10	Quantile-Quantile plots: July <i>Model_A</i> , <i>Model_B</i>	79
5.11	Monthly 24-hour means: historical versus 300 year simulation - Jan,July	82
5.12	Monthly CV: historical versus 300 year simulation - Jan,July .	84
5.13	Monthly skewness: historical versus 300 year simulation - Jan,Jul	89
5.14	Monthly autocorrelation: historical versus 300 year simulation	90
5.15	Stability of 300 year sample - pooled CV	92
5.16	Stability of 300 year sample - pooled skewness	94
5.17	Stability of 300 year sample - pooled autocorrelation	95
5.18	Example historical record with wet/dry indicators	97
5.19	BFLS: Algorithm	105
5.20	BFLS intensity: Jan,July	107
5.21	BFLS QQ regional: Jan,July	108
5.22	BFLS χ^2 tests: infilled versus historical	109
5.23	BFLS pooled	111
5.24	BFLS Pooled QQ plots	112
5.25	BFLS cross-correlation: Jan,July	113
5.26	BFCDF: Algorithm	116
5.27	BFCDF intensity: Jan,July	119
5.28	BFCDF QQ regional: Jun,Sept	120
5.29	BFCDF pooled	121
5.30	BFCDF Pooled QQ plots	122
5.31	BFCDF cross-correlation: Jan,July	123
5.32	ISCDF: Algorithm	125
5.33	ISCDF χ^2 tests: infilled versus historical	126
5.34	ISCDF intensity: Jan,July	127
5.35	ISCDF QQ regional: Jan,July	128

5.36	ISCDF pooled	130
5.37	ISCDF Pooled QQ plots	131
5.38	ISCDF cross-correlation: Jan,July	132
A.1	Site TW238605 daily plots	164
A.2	Site TW238605 hourly plots	165
A.3	Site TW239258 daily plots	166
A.4	Site TW239258 hourly plots	167
A.5	Site TW239315 daily plots	168
A.6	Site TW239315 hourly plots	169
A.7	Site TW239320 daily plots	170
A.8	Site TW239320 hourly plots	171
A.9	Site TW239374 daily plots	172
A.10	Site TW239374 hourly plots	173
A.11	Site TW239578 daily plots	174
A.12	Site TW239578 hourly plots	175
A.13	Site TW245176 daily plots	176
A.14	Site TW245176 hourly plots	177
A.15	Site TW246213 daily plots	178
A.16	Site TW246213 hourly plots	179
A.17	Site TW246424 daily plots	180
A.18	Site TW246424 hourly plots	181
A.19	Site TW246627 daily plots	182
A.20	Site TW246627 hourly plots	183
A.21	Site TW246847 daily plots	184
A.22	Site TW246847 hourly plots	185
A.23	Site TW247119 daily plots	186
A.24	Site TW247119 hourly plots	187
A.25	Site TW286392 daily plots	188
A.26	Site TW286392 hourly plots	189

A.27	Site TW287141 daily plots	190
A.28	Site TW287141 hourly plots	191
A.29	Site TW287283 daily plots	192
A.30	Site TW287283 hourly plots	193
A.31	Site TW287864 daily plots	194
A.32	Site TW287864 hourly plots	195
A.33	Site TW288020 daily plots	196
A.34	Site TW288020 hourly plots	197
A.35	Site TW288749 daily plots	198
A.36	Site TW288749 hourly plots	199
A.37	Site TW289022 daily plots	200
A.38	Site TW289022 hourly plots	201
A.39	Site TW289102 daily plots	202
A.40	Site TW289102 hourly plots	203
A.41	Site TW290007 daily plots	204
A.42	Site TW290007 hourly plots	205
A.43	Site TW291467 daily plots	206
A.44	Site TW291467 hourly plots	207
A.45	Daily data, April, site TW287283 versus site TW246627 . . .	209
A.46	Daily data, September, site TW287874 versus site TW247119	209
A.47	Daily data, January, site TW287874 versus site TW287141 . .	210
A.48	Daily data, February, site TW290007 versus site TW287283 .	210
A.49	Daily data, February, site TW288749 versus site TW287864 .	211
A.50	Daily data, November, site TW289022 versus site TW287283	211
A.51	Daily data, November, site TW291467 versus site TW290007	212
A.52	Daily data, December, site TW291467 versus site TW290007 .	212
A.53	Pooled statistics: cleaned data versus uncleaned data (CV, skewness, autocorrelation)	213

A.54	Pooled statistics: cleaned data versus uncleaned data (Cross-Correlation)	214
B.1	$Model_B$ cross-correlation versus distance - January, February	215
B.2	$Model_B$ cross-correlation versus distance - March, April . . .	216
B.3	$Model_B$ cross-correlation versus distance - May, June	217
B.4	$Model_B$ cross-correlation versus distance - July, August	218
B.5	$Model_B$ cross-correlation versus distance - September, October	219
B.6	$Model_B$ cross-correlation versus distance - November, December	220
B.7	Quantile-Quantile plots: February $Model_A, Model_B$	221
B.8	Quantile-Quantile plots: March $Model_A, Model_B$	222
B.9	Quantile-Quantile plots: April $Model_A, Model_B$	223
B.10	Quantile-Quantile plots: May $Model_A, Model_B$	224
B.11	Quantile-Quantile plots: June $Model_A, Model_B$	225
B.12	Quantile-Quantile plots: August $Model_A, Model_B$	226
B.13	Quantile-Quantile plots: September $Model_A, Model_B$	227
B.14	Quantile-Quantile plots: October $Model_A, Model_B$	228
B.15	Quantile-Quantile plots: November $Model_A, Model_B$	229
B.16	Quantile-Quantile plots: December $Model_A, Model_B$	230
C.1	Monthly Means by site - January and February	231
C.2	Monthly Means by site - March and April	232
C.3	Monthly Means by site - May and June	233
C.4	Monthly Means by site - July and August	234
C.5	Monthly Means by site - September and October	235
C.6	Monthly Means by site - November and December	236
C.7	Monthly Coefficient of Variation by site - January and February	237
C.8	Monthly Coefficient of Variation by site - March and April . .	238
C.9	Monthly Coefficient of Variation by site - May and June . . .	239
C.10	Monthly Coefficient of Variation by site - July and August . .	240

C.11 Monthly Coefficient of Variation by site - September and October	241
C.12 Monthly Coefficient of Variation by site - November and December	242
C.13 Monthly Skewness by site - January and February	243
C.14 Monthly Skewness by site - March and April	244
C.15 Monthly Skewness by site - May and June	245
C.16 Monthly Skewness by site - July and August	246
C.17 Monthly Skewness by site - September and October	247
C.18 Monthly Skewness by site - November and December	248
C.19 Monthly Autocorrelation by site - January and February	249
C.20 Monthly Autocorrelation by site - March and April	250
C.21 Monthly Autocorrelation by site - May and June	251
C.22 Monthly Autocorrelation by site - July and August	252
C.23 Monthly Autocorrelation by site - September and October	253
C.24 Monthly Autocorrelation by site - November and December	254
D.1 BFLS Intensity: Jan.Feb	256
D.2 BFLS Intensity: Mar.Apr	257
D.3 BFLS Intensity: Jan.Feb	258
D.4 BFLS Intensity: Jul.Aug	259
D.5 BFLS Intensity: Sept.Oct	260
D.6 BFLS Intensity: Nov.Dec	261
D.7 BFLS QQ Regional: Jan.Feb	262
D.8 BFLS QQ Regional: Mar.Apr	263
D.9 BFLS QQ Regional: May.Jun	264
D.10 BFLS QQ Regional: Jul.Aug	265
D.11 BFLS QQ Regional: Sept.Oct	266
D.12 BFLS QQ Regional: Nov.Dec	267
D.13 BFLS cross-correlation: Jan.Feb	268

D.14	BFLS cross-correlation: Mar, Apr	269
D.15	BFLS cross-correlation: May, Jun	270
D.16	BFLS cross-correlation: Jul, Aug	271
D.17	BFLS cross-correlation: Sept, Oct	272
D.18	BFLS cross-correlation: Nov, Dec	273
D.19	BFCDF cross-correlation: Jan, Feb	274
D.20	BFCDF cross-correlation: Mar, Apr	275
D.21	BFCDF cross-correlation: May, Jun	276
D.22	BFCDF cross-correlation: Jul, Aug	277
D.23	BFCDF cross-correlation: Sept, Oct	278
D.24	BFCDF cross-correlation: Nov, Dec	279
D.25	ISCDF Intensity: Jan, Feb	280
D.26	ISCDF Intensity: Mar, Apr	281
D.27	ISCDF Intensity: Jan, Feb	282
D.28	ISCDF Intensity: Jul, Aug	283
D.29	ISCDF Intensity: Sept, Oct	284
D.30	ISCDF Intensity: Nov, Dec	285
D.31	ISCDF QQ Regional: Jan, Feb	286
D.32	ISCDF QQ Regional: Mar, Apr	287
D.33	ISCDF QQ Regional: May, Jun	288
D.34	ISCDF QQ Regional: Jul, Aug	289
D.35	ISCDF QQ Regional: Sept, Oct	290
D.36	ISCDF QQ Regional: Nov, Dec	291
D.37	ISCDF cross-correlation: Jan, Feb	292
D.38	ISCDF cross-correlation: Mar, Apr	293
D.39	ISCDF cross-correlation: May, Jun	294
D.40	ISCDF cross-correlation: Jul, Aug	295
D.41	ISCDF cross-correlation: Sept, Oct	296
D.42	ISCDF cross-correlation: Nov, Dec	297

1. INTRODUCTION

The Lord will open the heavens, the storehouse of His bounty, to send rain on your land in season ...

Deuteronomy 28:12a NIV

1.1 Background

Accurate modelling of rainfall is critical to the successful design of effective urban drainage and stormwater systems. In order to build such systems, a long historical record is necessary so that the likelihood of extreme events and their relative location can be estimated. However, records of sufficient length and fine resolution are not available. As a result, considerable attention over the last two decades has been placed on developing a model suitable for simulating rainfall.

A fitted model, however, is only as accurate as the source data that the model is based on. Furthermore, any hydrodynamical model (eg: for surface runoff or flood frequency analysis) is heavily dependent on the rainfall modelling component as any inadequacy is directly incorporated into the pipe flow models (Mark and Hosner, 2002).

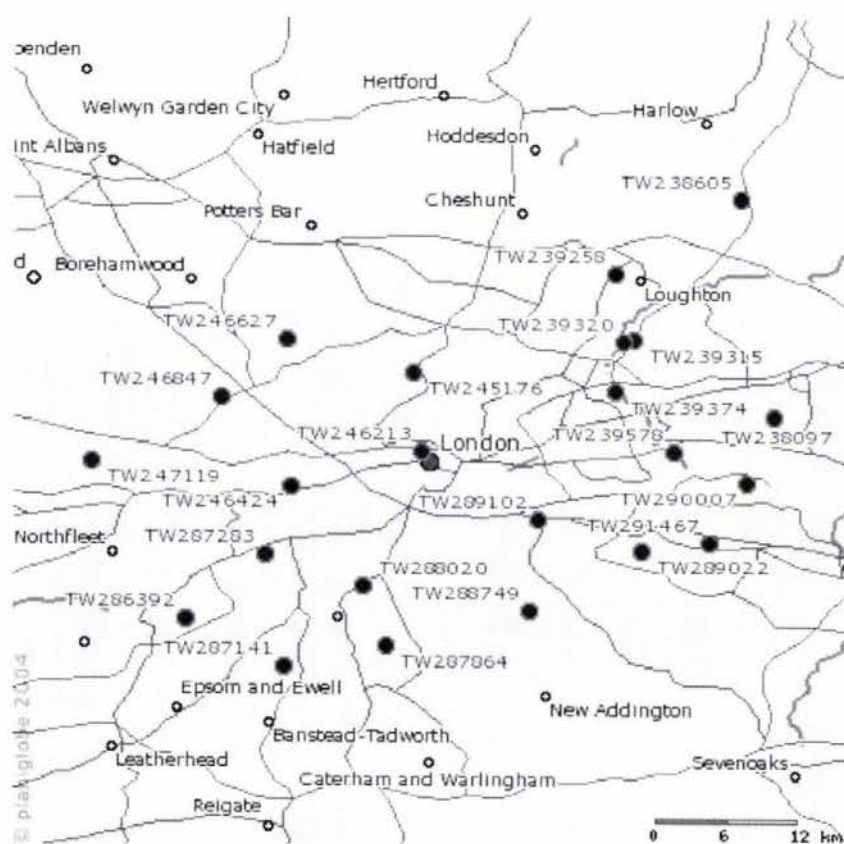
There are two main problems associated with rainfall data. Firstly, the data are sensitive to recording error, especially at fine aggregation levels (see Section 1.2.1). Secondly, the available historical data are generally sparsely populated with valid recordings.

The majority of historical source data available are collected at a 24-hour (daily) resolution. To a lesser extent, 1-hour records are also obtainable. However,

for drainage purposes, it is necessary to have a much finer timescale (for example 1-5 minute resolution). This presents two major hurdles to be overcome before a realistic parametric model can be produced. The data must be fully populated and made available at a useful resolution.

1.2 Source data

The data used in this project were collected from rain gauges from twenty-three sites in the Thames catchment from 1970 to 2003. A map of the site locations is shown in Figure 1.1. The site names along with corresponding site numbers are listed in Table 1.1 along with their corresponding Easting/Northing grid coordinates.



Map produced online from <http://www.planiglobe.com/>

Figure 1.1: Map of site locations in the Thames catchment

Table 1.1: Site location coordinates: Easting, Northing, and Altitude

Thames ID	Site number	Easting (0.1)km	Northing (0.1)km	Altitude (m)
TW238097	1	5499	1863	16
TW238605	2	5476	2048	75
TW239258	3	5412	1981	115
TW239315	4	5423	1926	15
TW239320	5	5418	1923	17
TW239374	6	5415	1882	8
TW239578	7	5447	1830	2
TW245176	8	5308	1894	33
TW246213	9	5314	1828	25
TW246424	10	5246	1795	21
TW246627	11	5241	1920	78
TW246847	12	5208	1870	42
TW247119	13	5141	1815	23
TW286392	14	5194	1682	12
TW287141	15	5247	1641	47
TW287283	16	5234	1737	56
TW287864	17	5299	1661	35
TW288020	18	5286	1712	40
TW288749	19	5375	1692	33
TW289022	20	5433	1745	75
TW289102	21	5377	1771	5
TW290007	22	5486	1805	8
TW291467	23	5468	1754	50

In general, a Thames Water site name is preferred to a site number, particularly when analysing the historical data (Chapter 4). However, if it is not necessary to be able to immediately identify a particular site, then the sites are referred to by their corresponding site number.

Two aggregation levels were available for use: a 1-hour record and a 24-hour (daily) record. The 24-hour record was substantially longer and covered the years 1970 – 2003. The 1-hour record was only available from 1989 – 2003. All sites had some data, however, for some sites (eg: TW239315, TW289022) the percentage of valid records after the data were cleaned was quite low (see Section 4.4.1).

1.2.1 Measurement error

There are a number of devices for recording rainfall measurements. The oldest method is to use a collection of rain gauges, however, more recent developments enable collection of data via radar networks or satellite sensors (Maidment, 1993).

The data used within this project were collected from rain gauges (Section 1.2). Tipping bucket rain gauges, used by the Environment Agency in the United Kingdom (see Tilford et al., 2003, chap. 2) for real-time monitoring of rainfall, are affected by a variety of environmental conditions. The primary sources of measurement error associated with this collection method are well known and include wind speed, the height of the gauge above the ground, and snowfall (Tilford et al., 2003; Maidment, 1993). Note that the errors associated with measuring snowfall are usually larger than rainfall (Maidment, 1993).

1.3 Thesis outline

Techniques for fully populating the historical record at the 24-hour aggregation level will be derived within this thesis. The constructed algorithms will use a synthetic record generated using a spatial-temporal point process model of rainfall (see Section 3.1). Once a technique for infilling the historical record at a 24-hour level is developed, methods for *disaggregation* from this fully populated record can be applied (for example Zhiqian and Eltahir, 1994; Glasbey et al., 1995; Güntner et al., 2001; Kottegoda et al., 2003; Cowpertwait et al., 2004). Where *disaggregation* describes an algorithm to generate data at a finer resolution (eg: 1-hour) than the available data (eg: 24-hour). Note that generally the total amount of rainfall at the same site and time interval (24-hour) is expected to match exactly for the available and disaggregated records.

The remainder of this thesis is organised as follows. The next chapter (Chapter 2) presents a review of the literature for both rainfall models and infilling algorithms. In Chapter 3, the model is mathematically described along with the fitting technique. Furthermore, the methods for cleaning the historical data are presented, the infilling algorithms are proposed, and the implementation of the algorithms is discussed. The results of the data analysis and cleaning is covered in Chapter 4. In Chapter 5, the results of the model fitting, validation, and infilling algorithms are presented. Finally, in Chapter 6 the conclusions and directions for future research for the model and infilling are derived.