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

LIST OF TABLES

LIST OF FIGURES

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).

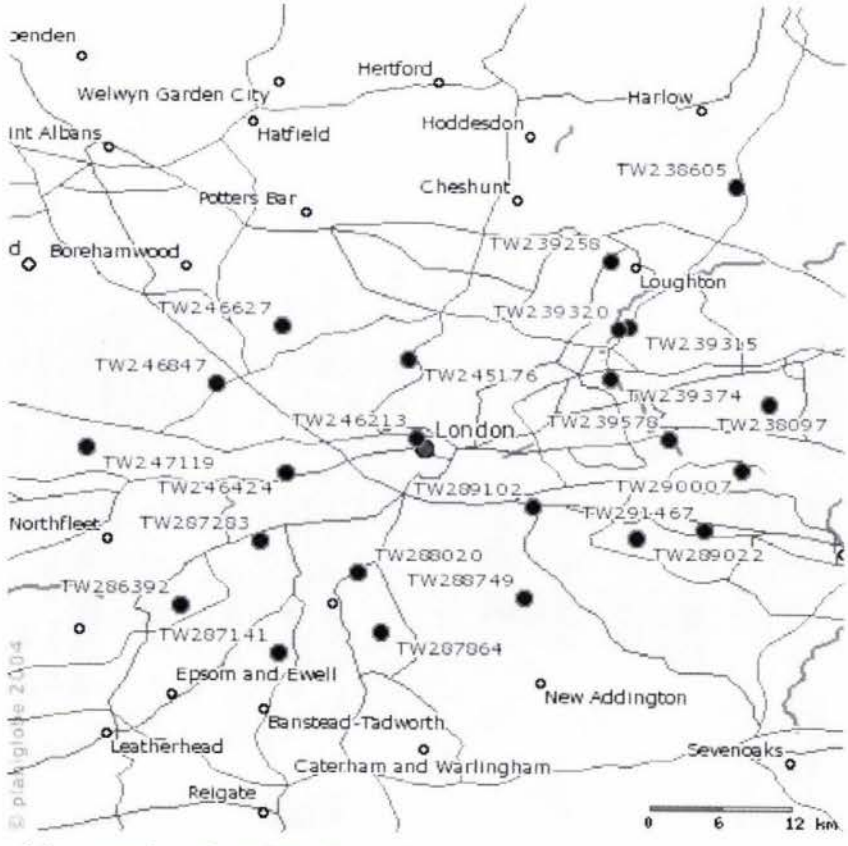
There are two main problems associated with rainfall data. Firstly, the data are sensitive to recording error, especially at fine aggregation levels (see Section 1.2.1). Secondly, the available historical data are generally sparsely populated with valid recordings.

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 Zhiquan 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.

2. LITERATURE REVIEW

> I have not all my facts yet, ... Still, it is an error to argue in front of your data. You find yourself insensibly twisting them round to fit your theories.
>
> **-Sherlock Holmes, The Adventure of Wisteria Lodge**

2.1 Stochastic rainfall models

A number of stochastic models for modelling precipitation over time have been proposed and developed in the literature. These models are generally either temporal or spatial-temporal in nature.

2.1.1 Temporal models

Two temporal models were developed concurrently in the 1960s for modelling rainfall recorded in discrete time steps (ie daily (24-hour) intervals). One family of models used Markov chains to describe the change between wet and dry periods (eg: Feyerherm and Bark, 1964). Rainfall intensity has also been modelled with Markov chains (eg: Stern and Coe, 1984). Another class of models were based on point-processes as suggested by Le Cam (1961). Further developments by (Rodriguez-Iturbe et al., 1987, 1988) provided the theoretical foundation for models with storms arriving in a Poisson process with a cluster of associated rectangular pulses representing rain cells (Olsson and Burlando, 2002).

More recent developments include the modelling of rainfall via raindrop processes (Smith, 1993; Smith and DeVaux, 1994) and constructing models based on the scaling properties of temporal rainfall (see Olsson and Burlando, 2002). The

former raindrop models are likely to be useful at a fine resolution (1 minute or less), whereas the scaling-based rainfall models, while promising, are still being developed (Olsson and Burlando, 2002).

Markov models

Models of rainfall data via Markov chains have been conducted in Stern and Coe (1984); Balaji (1995); Grunwald and Jones (2000); Onof et al. (2002). Most Markov chain models in the literature are first-order (Srikanthan and McMahon, 2001). According to Dobi-Wantuch et al. (2000), the popularity of the first-order, two-state, homogeneous Markov chain model is primarily due to the simplicity of the calculations for generating synthetic series of wet and dry days. Although this model is simple, it usually overestimates the very short dry sequences and underestimates the very long dry sequences (Dobi-Wantuch et al., 2000).

Point process models

As mentioned previously, the models derived by Rodriguez-Iturbe et al. (1987, 1988) formed the basis for the widespread use of rectangular pulse models for modelling rainfall. Within these point process models, each storm arrives in a Poisson process in time. Each storm has a random number of rectangular pulses (rain cells) of random duration and intensity associated with it. (As the rectangular pulse model is the most common model formulation, unless noted otherwise the rain cells are to be assumed to be rectangular pulses.) The amount of rainfall at the site at any given time is the sum of the intensities of the rain cells at that time (see also Section 3.1 and Figure 3.1).

The association of the cells with the storms generally follow either a Bartlett-Lewis (BL) process or a Neyman-Scott (NS) process (Rodriguez-Iturbe et al., 1987; Olsson and Burlando, 2002). These two models differ slightly in their formulation in that the relationship between the cell arrival and storm origin are not the same. In the BL model the times between cell origins are independent whereas in the NS formulation the times between cell origins and the storm origins are independent (Cowpertwait et al., 2002). Furthermore, it has been shown analytically (Cowpertwait, 1998) that these models are statistically equivalent up to their 2^{nd} order properties. The point-process models are less sensitive to errors than the Markov

models as they use summary statistics in the fitting process rather than fitting directly to the historical data.

These two models have been investigated extensively over the past fifteen years. Studies conducted using the BL models include Islam et al. (1990); Kakou (1998); Cameron et al. (2001); Smithers et al. (2002); Skaugen et al. (2003). The NS models have also been frequently applied to various data sets (eg: Rodriguez-Iturbe et al., 1986; Cowpertwait, 1994; Cowpertwait et al., 1996; Cowpertwait and O'Connell, 1997; Cowpertwait, 1998). As with any model specification, these models are limited as they generally fail to meet some characteristic of the historical data (see Maidment, 1993; Onof et al., 2000). For example, the reproduction of wet/dry sequences is often a problem with the Neyman-Scott or Bartlett-Lewis models (Onof et al., 2000). However, these limitations do not imply that the model cannot be used in practice.

Note that within these models it is generally assumed that intensity and duration are independent random variables. However, the dependent cell duration and intensity was investigated for the NS model by Cowpertwait (1994) and, for the BL model by Kakou (1998). Onof et al. (2000) found the dependent duration-intensity BL model improved the reproduction capabilities of the proportion of dry sites when compared with the original BL and NS models.

Other models

Other rainfall models have also be derived such as those based on Cox processes (Ramesh, 1998) and renewal processes (Cowpertwait, 2001; Mohapl, 2002). Note that the renewal process models are applied to real time data rather than discrete data as in the point-process and Markov Chain models discussed previously. Most discrete data are recorded at a 1-hour or higher resolution (or aggregated to this level from a finer resolution), whereas the real time data may be recorded at as fine a resolution as 1-second.

2.1.2 *Spatial-temporal models*

In the late 1980s, spatial-temporal stochastic models were developed and used (eg: Bell, 1987; Cox and Isham, 1988). Since these preliminary formulations, spatial-temporal point process models have been the subject of many studies (eg: Cowpertwait, 1995; Northrop, 1998; Favre and Overney, 1999; Onof et al., 2000; Cowpertwait et al., 2002).

Point process models

The extension of temporal point-process models into the spatial domain was proposed by Cox and Isham (1988). In this model, storms arrive in a spatial-temporal Poisson process where each storm consists of a circular region of rain which moves with a random velocity for a random time after which it disappears. During the storm's lifetime, the intensity of rain over the region remains constant. As mentioned in this paper (Cox and Isham, 1988), the model is highly idealised and does not incorporate known features of rain cell behaviour (for example cell clustering).

This spatial-temporal model was then extended to clustered point processes, once again for the BL model (Northrop, 1998) and for the NS model (eg: Cowpertwait, 1995; Cowpertwait et al., 2002). Comparisons have also been made between the two models (eg: Onof et al., 2000). Note that Northrop (1998) preferred elliptical cells rather than circular cells as the the autocorrelation plots obtained from the radar data (which the model was fitted to) showed elliptical contours.

In order to formulate a clustered point-process model that can be related to the physical process, some assumptions about the process are necessary to make fitting the model easier. Generally, the data are transformed to temporal and spatial stationarity prior to fitting the model (eg Cowpertwait et al., 2002). Furthermore, with all rectangular pulse models the cell intensity is held constant over the cell duration and the cell area.

The models in the literature also specify different assumptions regarding the characteristics of the model. For example, the shape of the rain cell (Onof et al., 2000), number of rain cells types (eg Cowpertwait, 1995), and the hierarchical structure between storm events and rain cell clusters (Northrop, 1998).

The model (Cowpertwait et al., 2002) that is applied within this thesis (see Sections 2.1.3 and 3.1) also specifies the assumptions as listed below.

(a) each rain cell has a randomly generated constant x, y origin

(b) cell intensity, duration, and radius are mutually independent random variables

(c) cell x, y origins are independent within storms

(d) there is only one type of rain cell (convective and stratiform cells are not modelled separately)

Some alternative model formulations, as presented in the literature, are discussed briefly below.

Non-stationary rain cells

The movement of rain cells has been considered previously with the BL type models (see Northrop, 1998; Onof et al., 2000). The correlation functions (temporal and spatial) are reproduced by the model (Northrop, 1998). The main problem with this model is the computational difficulties associated with the fitting procedure.

Spatial storm and cell dependence

As discussed in Northrop (1998), rain cells tend to form in the vicinity of existing rain cells forming larger-scale structures. Therefore, a spatial-temporal model where cells were clustered about the storm origin, spatially and temporally, was proposed (see Northrop, 1998). Due both to the difficulty of specifying a likelihood function in a useful form and the structure of the cell intensities, model fitting via maximum likelihood is not an appropriate method. Rather, a generalised method of moments is preferred (Northrop, 1998).

The structure of storms and clusters within them are explored in De Lannoy et al. (2004). It was observed that rain cells are clustered within a storm area (as expected) and that a simple Poisson process model is adequate to capture the spatial distribution of the cells (De Lannoy et al., 2004).

Multiple cell types

A spatial-temporal Neyman-Scott Rectangular Pulse (NSRP) model containing more than one cell-type was developed by Cowpertwait (1995), however, if used then the equations for the third moment derived by Cowpertwait (1998) can not be used in the model fitting. However, the inclusion of multiple cells is expected to be particular effective over the summer period where local, intense storms are more frequent. Again, as with storm movement, this model is considerably more difficult to fit than the model considered within this thesis.

Multiple cell types at a single site NRSP model have been found to produce realistic output in terms of extreme values and return period (Cowpertwait and O'Connell, 1997), but the fit of this model to the proportion of dry periods was not checked.

Non-constant rain cell intensity

Although the assumption of constant rain cell intensity is physically unjustifiable, this assumption is probably the least important assumption to be corrected. For a model with spatially clustered cells (see Northrop, 1998; Willems, 2001), the assumption of constant cell intensity did not appear to be important. Therefore, random noise could be applied to the simulated records if necessary rather than accounting for non-constant intensity directly.

Other models

In addition to the models previously listed, the following models have also been applied to varying aggregation levels in the spatial-temporal domain.

Conditional models using Markov chains have been applied to daily rainfall data in the spatial-temporal domain (Srikanthan and McMahon, 2001). In addition to daily rainfall amounts other climate factors such as atmospheric circulation patterns may also be used as input into the model as in Hughes et al. (1999).

The Modified Turning Bands (MTB) model, was formulated in Mellor (1996). The main features of frontal rainfall systems (eg: rainbands, clusters of potential

regions, and raincells) were to be reproduced by this model. The model is fitted to radar data in (Mellor and O'Connell, 1996) and future model prospects are covered further in (Mellor and Metcalfe, 1996).

Spatial-temporal models for rainfall have been proposed within a Bayesian framework (see De Oliveira et al., 1997; Sansó and Guenni, 1999; Verlarde et al., 2004) for various aggregation levels.

Willems (2001) developed a spatial rainfall generator for use at small spatial scales (see also Willems and Luyckx, 1999). The generator described the rainfall field by distinguishing between differing levels of rainfall scales (ie rain cells, clusters, and larger areas). Good results were found for both rainfall frequency and for a number of aggregation levels between 20 minutes and 1 day.

2.1.3 Applied model

The results in Cowpertwait et al. (2002) showed that a simple spatial-temporal NS rectangular pulse model was able to maintain regional extremes. This model is mathematically tractable and has been shown to maintain extreme value characteristics. This latter point is often a problem for point process models (see Onof et al., 2000). Therefore, as the extreme value behaviour is preserved, this model is applied within this thesis. The complete mathematical description and fitting method is described in Section 3.1.

2.2 Infilling

In the literature there are a number of papers on estimating missing data for time series, spatial data, or both (eg: Kohn and Ansley, 1986; Carlin et al., 1992; Kong et al., 1994; Mehrotra and Singh, 1998; Venugopal et al., 1999; Johns et al., 2003). The methods are either based on a model or are based on the data directly (Hox, 1999). The methods applied within this thesis are based on a fitted spatial-temporal stochastic model as described in Cowpertwait et al. (2002).

These infilling methods in the literature may either impute (predict a missing record) a single value or generate multiple imputation data sets (see for example

Haining, 2003, p. 157). Obviously, the advantage of multiple imputation is that an estimate of the error associated with the predictions is readily available. Furthermore, the model can be refitted to the multiple imputed data sets and the effect of the imputation on the model can be seen. For this analysis however, only single imputation methods were applied.

As the data used are both spatial and temporal in nature, the infilling methods must also be able to predict in space and time. There are two distinct problems associated with spatial-temporal rainfall data. Firstly, there is the case where a high aggregation level is known (say a 24-hour total) but lower aggregation levels (1-hour totals, 15 minute totals) are not known. In this case, disaggregation algorithms (for example Ormsbee, 1989; Glasbey et al., 1995; Güntner et al., 2001; Kottegoda et al., 2003; Cowpertwait et al., 2004) are applicable. While this is certainly a problem for the data set (see Section 1.2), as 1-hour values were only collected 19 years after the 24-hour values were collected, the focus of this thesis is on the second problem. That is, missing data occurring at the highest available aggregation level - in this case the 24-hour data.

For the Thames Water data, there are two approaches that could be applied. Firstly, all available 24-hour data could be disaggregated using a disaggregation algorithm (see above) and then infilled at the 1-hour level when, if necessary, these 1-hour values can be aggregated up to 24-hour values. Alternatively the data could first be infilled at the higher aggregation level (24-hour) and then the infilled data disaggregated to the lower aggregation levels. Of these, the second method is more intuitive as for precipitation data higher aggregation levels are more highly correlated over larger distances than lower aggregation levels - eg: Section 5.2.2. Furthermore, if overdispersion is an issue at an hourly to daily aggregation level as it is from a daily to monthly aggregation level (see Katz and Parlange, 1998) then it is necessary to infill first *then* disaggregate in order to reduce the effect of this problem. However, this latter option does not make use of any partial data that may be recorded at a finer time scale - for example 12 hours of a 24-hour total may be recorded at the 1-hour level but a missing value recorded at the 24-hour level.

2.2.1 *Missing data assumptions*

Before continuing further, it is important to note that the algorithms described within this section assume that the missing data *mechanism* is ignorable (Haining, 2003). That is, the missing data are spatially and temporally 'missing at random' (MAR). If the data is not MAR but instead there is some form of missing data bias (for example, values of high magnitude are more likely to be missing) then any modelling or imputation method must take this into account if it is to be reliable (Schafer and Graham, 2002; Haining, 2003).

Of these two assumptions (Spatial MAR and temporal MAR) it is generally more important that temporal MAR is maintained and that no data censoring (for example of extreme values) has occurred. The latter is particularly important for flood frequency analysis.

2.2.2 *Algorithms*

Although the data are spatial-temporal, the infilling problem can be reduced to a spatial prediction problem provided enough data are available at a given time point. The justification for this lies in the observation that precipitation data are more highly correlated spatially than temporally over a 'small' area at a 'high' aggregation level (eg: 24-hour or higher for the Thames Water data). It is also obvious that as the area under study increases the necessary aggregation level for this observation to be true will also increase.

From the first order autocorrelation function (see Table 5.1) and the first order cross-correlation function (see Figures B.1 to B.6) for the Thames Water data, it is evident that in general the first order cross-correlation is three times higher than than the first order autocorrelation at the 24-hour level. Thus an infilling algorithm that infills spatially at each discrete time step is applicable and the temporal dimension is not as important as the high cross-correlation is expected to account indirectly for the autocorrelation. However, given the marked difference between the two measurements, it is possible that an algorithm which only infilled spatially will tend to overestimate the first order autocorrelation. For the sake of simplicity, however, the temporal component will be included only if spatial prediction fails to reproduce the required characteristics.

Requirements

Haining (2003) observes that a method estimating values on a spatial surface should satisfy the following four requirements. Firstly, the spatial structure of the surface should be utilised. Secondly, as observed values that are spatially close together duplicate information, regional clusters should be weighted. Thirdly, some estimate of the error of the prediction should be able to be obtained. Finally, the method should, when estimating a known value, predict a value as close as possible to that known value.

These same four requirements also apply directly to any infilling algorithm. In order to fulfill the first requirement, a spatial-temporal NS point process model is fitted to the historical data. The model is expected to capture the spatial (and temporal) characteristics of the rainfall data if the assumptions for the model are satisfied. The second and third requirements can easily be incorporated within the infilling algorithm by weighting site values and imputing missing data multiple times respectively. The last requirement is used to measure the effectiveness of the infilling algorithm within this thesis.

Candidates

Markov Chain Monte Carlo (MCMC) (Dellaportas and Roberts, 2002) is an option for infilling the missing records. While MCMC algorithms are an attractive alternative to model fitting, the disadvantage of this method is the computation time for imputation - especially if the sampling method converges slowly (which is highly probable). Furthermore, the data must be reliable for MCMC to generate useful results. As Section 1.2.1 highlights, the data are subject to high measurement error, especially over winter. Therefore, a model based approach is preferred to MCMC methods.

Furthermore, although interpolation methods based on regression or kriging (see Bennett et al., 1984; Cressie, 1993; Hox, 1999) could be applied to the data to interpolate between points these methods will not be optimal. The reason for this lies in the attributes of the rainfall data, namely, the high positive skewness,

the strictly non-negative values (thus causing residual error to be asymmetric), and high covariance. Note that with the regression and kriging methods, it would be assumed that some form of stochastic error would be added to the predicted values so as to avoid underestimation of the variance (Hox, 1999).

As the model specified in (Cowpertwait et al., 2002) produces sensible characteristics of precipitation data, this model can be used as a foundation for an infilling algorithm. The model, once fitted, can be used to generate a synthetic record of rainfall data for the historical sites. An infilling algorithm can then use these synthetic records to find a sensible prediction for missing values based on the similarity of the valid data within the historical record and a potential synthetic one.

If additional information was available, for example circulation model data, then the inclusion of these records in the infilling method would be desirable (see: Mehrotra and Singh, 1998; Venugopal et al., 1999).

2.2.3 Summary

The infilling algorithms applied within this thesis are constructed from the model-based imputation methods. Firstly, a model is fitted to the valid records. Secondly, this fitted model is used to generate a synthetic record. Lastly, the synthetic record is used for infilling. This sequence of steps could be applied multiple times in order to obtain an estimate of the prediction error (assuming that the algorithm converges).

The first algorithm applied selects values used for infilling from a single row (where a row corresponds to site data collected/simulated at a particular time - for instance: January 1^{st}, 2000 for 24-hour data) of best fit. The best fitting row is found using a least squares fit between any valid historical data and synthetic data records. The second algorithm also selects values from a single row, but transforms the historical and simulated data onto a CDF before applying the least squares function. The third algorithm selects values for infilling from the best fitting *rows* using random uniform variates. A full description of the applied algorithms is given in Sections 3.3 and 5.5.

3. METHODOLOGY

> No matter what the statistical problem may be, it must proceed according to a *plan* ... under all circumstances a definite plan providing for all the detail is an absolute prerequisite.
>
> **-August Meitzen, quoted in Gaither and Cavazos-Gaither (1996)**

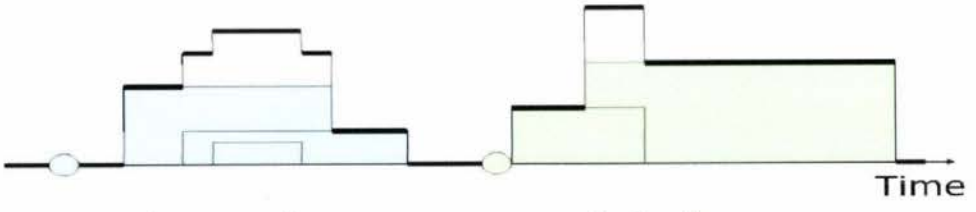
This chapter can be categorised into four broad areas. The first section presents the spatial-temporal Neyman-Scott Rectangular Pulse (NSRP) model along with techniques for fitting and verifying the model. Graphical interpretations of the process underlying the spatial-temporal NSRP model are also presented within this section. The next section describes the methods and plots for validating the rainfall data. The third section briefly outlines the three main algorithms used for infilling the missing data in the historical record for which results are included in Chapter 5. The techniques used to analyse the performance of the infilling algorithms are also discussed within this section. The final section describes the implementation of the algorithms comprising the project.

3.1 Spatial-temporal NSRP model

The model, as presented in Cowpertwait et al. (2002), will be fitted to rainfall data taken from the Thames catchment, thus providing the modelling component required by the infilling algorithms. The model has previously been fitted to precipitation data in the Arno Basin, Italy, where it was shown to preserve the behaviour of extremes (Cowpertwait et al., 2002). Furthermore, it is conceptually

simple and is easy to fit. Note that the simplifications required by this model may impact the *accuracy* of the infilling algorithms; however, these algorithms are generic and can use data generated by any model.

The spatial-temporal NSRP model consists of storms where the arrival times follow a Poisson process. Each storm has a random number of rectangular pulses (or rain cells) associated with it; where the pulse heights correspond to rainfall intensity and the pulse widths correspond to rainfall duration. As overlap of pulses (and storms) is permitted, at any given time, t, the rainfall intensity is the sum of the active pulse intensities at t (Figure 3.1). Each cell is a circular rain disc covering a region (Figure 3.2). As the rainfall data are measured at discrete intervals (1-hour or 24-hour), simulated data are also produced at discrete intervals. In Figure 3.2, each time step can be considered as an xy-plane over the region under study, and the shading of the cells is an indication of the intensity of the rainfall at the (x, y) coordinates.



where:	ovals	are storm arrival points
	rectangles	are rain cells
	thick line	is the rainfall intensity at time t

Figure 3.1: Temporal Neyman-Scott model

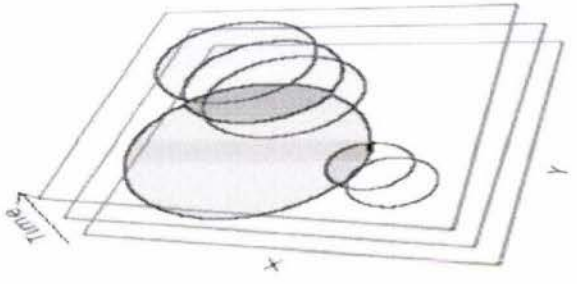


Figure 3.2: Spatial-temporal Neyman-Scott model

3.1.1 Assumptions

In order to ensure the that model parameters are constant, the model is assumed to be stationary in time and space (Cowpertwait et al., 2002). The raw data does not meet this assumption as seasonal patterns exist and site means vary. Thus the data must be transformed to a stationary series or a stationary subset before the model can be fitted. Following Cowpertwait et al. (2002), the data are scaled by the site mean to realise spatial stationarity and, to account for seasonal variation, the model is fitted to each calendar month separately. Twelve estimates are therefore obtained for each of the model parameters.

As with Cowpertwait et al. (2002), the cells are assumed to have zero velocity - that is, they are stationary in space. This assumption ensures the mathematical derivations (Section 3.1.3) are mathematically tractable. Although this seems a simplistic approach for modelling the physical process, the aggregation level of 1-hour produces a model that fits well (Cowpertwait et al., 2002).

It is suggested by Northrop (1998) that provided a model assuming such zero velocity cells is fitted to data with an aggregation level of at least 1-hour, this assumption is not unjustifiable. Although counter-intuitive, this result is derived from the observation that over a time period, storm movement tends to be included within the time intervals. For example, a 15-minute shower at site A may 'move' (within the same 1-hour time interval) to site B and shower for (say) another 15-minutes. However as both 15-minute showers are incorporated within the same 1-hour aggregation time interval (and therefore the total rainfall for that hour), the storm movement does not need to be modelled. It should be obvious therefore, as the aggregation level increases the requirement of modelling cell-movement decreases.

In addition cell origins are assumed to be independent in space, again for the purpose of simplification. A more complex formulation (see Northrop, 1998) is required if cells have either non-zero velocity or are associated spatially with storm origins.

As the primary usage of the model in this thesis is for infilling the historical record, all seasons are grouped together for the same site across multiple years.

If climatic change occurs over the time period requiring infilling, then a model which does not incorporate the changing climate can still be fitted. However, in order to infill the historical record correctly, the infilling algorithm would have to be adjusted to only select candidate points from the synthetic record that are climatically similar. This is particularly important when an infilling algorithm such as the *iterative CDF least squares algorithm*, Section 5.5.4, is applied.

If the model was to be used for other purposes, for example climate prediction, the grouping of all seasons across years would not be correct. Instead, the modelling would have to be adjusted for climate change over the historical period and not just for seasonal effects.

3.1.2 Model notation

Let T_i be the arrival time of the ith storm. The storm arrival times T_i follow a Poisson process with an arrival rate λ per unit time such that $T_{i+1} - T_i$ is exponentially distributed with mean λ^{-1}. The mean number of cells, μ_c, linked with a storm is given by a two-dimensional Poisson process with rate φ per unit area.

Let C_{ij} be the arrival time of the jth cell in the ith storm where $C_{ij} - T_i$ is the displacement between the jth cell arrival time and the ith storm arrival time. Let cell displacement times be independent random variables exponentially distributed with mean β^{-1}. As cell displacement times are dependent on the storm origin and not on the arrival times of adjacent cell origins, cell arrivals C_{ij} form an Neyman-Scott point process.

Cell duration and cell radius are also random variables modelled by independent exponential distributions with means η^{-1} and ϕ^{-1} respectively.

Cell intensity, X, is taken to be distributed as a two-parameter Weibull random variable with parameters α and θ for shape and scale. The probability distribution function for the two-parameter Weibull is given in Equation 3.1. As a result, the mean of X can be derived (Equation 3.2) and the variance of X as Equation 3.3.

$$f(x) = \begin{cases} (\frac{\alpha}{\theta})(\frac{x}{\theta})^{\alpha-1} e^{-(x/\theta)^{\alpha}}, & \theta, \alpha > 0, x \geq 0, \\ 0, & x < 0 \end{cases} \tag{3.1}$$

$$\mu_X = \theta\Gamma(1+\frac{1}{\alpha}) \tag{3.2}$$

$$\sigma_X^2 = \theta^2(\Gamma(1+\frac{2}{\alpha}) - [\Gamma(1+1/\alpha)]^2) \tag{3.3}$$

where $\Gamma(x)$ denotes the gamma function (Equation 3.4).

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx \tag{3.4}$$

As the lowest non-zero record for the 1-hour aggregation is $0.2mm$ for all sites, this is assumed to be the tipping bucket size for the rain gauges in the region (see Tilford et al., 2003, chap. 2). Similarly, the lowest non-zero record for the 24-hour aggregation level is $0.1mm$. It is assumed that the difference between the two aggregation levels is due to storage gauges being used to gather rainfall data at a 24-hour aggregation level rather than tipping buckets. Note also that there are numerous measurements problems associated with tipping buckets (see Section 1.2.1) and the effect of truncation to $0.2mm$ will be significant.

The region spans approximately 50km in diameter with rain gauges distributed as in Figure 1.1. The notation with units for the model can then be specified as shown in Table 3.1. This notation is identical to that as employed by Cowpertwait et al. (2002) as the model and units are the same.

Table 3.1: Model notation

λ^{-1}	mean time ($hours$) between two adjacent storm origins.
β^{-1}	mean time ($hours$) between storm origins and cell starting times.
η^{-1}	mean duration ($hours$) of a cell.
α	shape parameter for cell intensity X.
θ	scale parameter ($mm/hour$) for cell intensity X.
μ_c	mean number of cells per storm per site.
ϕ^{-1}	mean cell radius (km).
φ^{-1}	mean number of cell origins per km^2.

3.1.3 *Mathematical description*

The first and second order properties of the Spatial-Temporal NSRP model have been derived by Cowpertwait (1995), and the third order moment has been derived by Cowpertwait (1998).

For convenience, these equations have been reproduced below:

$$\mu_h = E\{Y_k^{(h)}(\mathbf{x})\} = \lambda \mu_C \mu_X h/\eta. \tag{3.5}$$

where, C is a Poisson random variable (Cowpertwait et al., 2002).

Therefore, in equations (3.8) and (3.9),
$E\{C(C-1)\} = \mu_c^2$ and $E\{C(C-1)(C-2)\} = \mu_c^3$.

$$\begin{aligned} \gamma_{x,\eta,h,L} &= Cov\{Y_k^{(h)}(\mathbf{x}), Y_{k+L}^{(h)}(\mathbf{y})\} \\ &= \gamma_{x,\eta,h,L} - 2\lambda\{1 - P(\phi,d)\}\mu_C E(X^2) A(h,L)/\eta^2, \end{aligned} \tag{3.6}$$

where:

$$\begin{aligned} \gamma_{x,\eta,h,L} &= \gamma_{\eta,\eta,h,L} = \lambda\eta^{-3} A(h,L) \\ &\quad \cdot\{2\mu_C E(X^2) + \mu_X^2 \beta^2 E(C^2 - C)/(\beta^2 - \eta^2)\} \\ &\quad -\lambda\mu_X^2 B(h,L) E(C^2 - C)/\{\beta(\beta^2 - \eta^2)\}; \end{aligned} \tag{3.7}$$

$$\begin{aligned} A(h,0) &= (h\eta + e^{-\eta h} - 1); \\ B(h,0) &= (h\beta + e^{-\beta h} - 1); \\ A(h,L) &= \frac{1}{2}(1 - e^{-\eta h})^2 e^{-\eta h(L-1)}; L = 1, 2, \ldots \\ B(h,L) &= \frac{1}{2}(1 - e^{-\beta h})^2 e^{-\beta h(L-1)}; L = 1, 2, \ldots \end{aligned}$$

As given in Cowpertwait et al. (2002), $P(\phi, d)$ is evaluated by:

$$P(\phi,d) = \frac{2}{\pi}\int_0^{\pi/2}\left(\frac{\phi d}{2\cos y} + 1\right) exp\left(\frac{-\phi d}{2\cos y}\right) dy \tag{3.8}$$

In the fitting procedure, equation (3.8) was evaluated numerically - unlike in Cowpertwait et al. (2002) where an approximation using Simpson's Rule with five ordinates was used. The numerical integration algorithm applied split the intergrand into as many intervals as necessary to reduce the overall error to less than 1×10^{-8}. Thus the calculation of the cross-correlation function (Equation 3.6) is expected to be more accurate, potentially allowing a better fit to be obtained.

The third moment, as derived by Cowpertwait (1998), is:

$$\begin{aligned} E\{Y_k^{(h)}(\mathbf{x}) - \mu_h\}^3 &= 6\lambda\mu_C E(X^3)(\eta h - 2 + \eta h e^{-\eta h} + 2e^{-\eta h})/\eta^4 \\ &\quad +3\lambda\mu_X E(X^2)E\{C(C-1)\}f(\eta,\beta,h) \\ &\quad /\{2\eta^4\beta(\beta^2-\eta^2)^2\} \\ &\quad +\lambda\mu_X^3 E\{C(C-1)(C-2)\}g(\eta,\beta,h) \\ &\quad /\{2\eta^4\beta(\eta^2-\beta^2)(\eta-\beta)(2\beta+\eta)(\beta+2\eta)\} \end{aligned} \tag{3.9}$$

where $f(\eta,\beta,h)$ is:

$$\begin{aligned} f(\eta,\beta,h) &= -2\eta^3\beta^2 e^{-\eta h} - 2\eta^3\beta^2 e^{-\beta h} + \eta^2\beta^3 e^{-2\eta h} + 2\eta^4\beta e^{-\eta h} \\ &\quad +2\eta^4\beta e^{-\beta h} + 2\eta^3\beta^2 e^{-(\eta+\beta)h} - 2\eta^4\beta e^{-(\eta+\beta)h} - 8\eta^3\beta^3 h \\ &\quad +11\eta^2\beta^3 - 2\eta^4\beta + 2\eta^3\beta^2 + 4\eta\beta^5 h + 4\eta^5\beta h - 7\beta^5 \\ &\quad -4\eta^5 + 8\beta^5 e^{-\eta h} - \beta^5 e^{-2\eta h} - 2h\eta^3\beta^3 e^{-\eta h} \\ &\quad -12\eta^2\beta^3 e^{-\eta h} + 2h\eta\beta^5 e^{-\eta h} + 4\eta^5 e^{-\beta h} \end{aligned}$$

and $g(\eta,\beta,h)$ is:

$$\begin{aligned} g(\eta,\beta,h) &= 12\eta^5\beta e^{-\beta h} + 9\eta^4\beta^2 + 12\eta\beta^5 e^{-\eta h} + 9\eta^2\beta^4 \\ &\quad +12\eta^3\beta^3 e^{-(\eta+\beta)h} - \eta^2\beta^4 e^{-2\eta h} - 12\eta^3\beta^3 e^{-\beta h} - 9\eta^5\beta \\ &\quad -9\eta\beta^5 - 3\eta\beta^5 e^{-2\eta h} - \eta^4\beta^2 e^{-2\beta h} - 12\eta^3\beta^3 e^{-\eta h} \\ &\quad +6\eta^5\beta^2 h - 10\beta^4\eta^3 h + 6\beta^5\eta^2 h - 10\beta^3\eta^4 h + 4\beta^6\eta h \\ &\quad -8\beta^2\eta^4 e^{-\beta h} + 4\beta\eta^6 h + 12\beta^3\eta^3 - 8\beta^4\eta^2 e^{-\eta h} - 6\eta^6 \\ &\quad -6\beta^6 - 2\eta^6 e^{-2\beta h} - 2\beta^6 e^{-2\eta h} + 8\eta^6 e^{-\beta h} \\ &\quad +8\beta^6 e^{-\eta h} - 3\beta\eta^5 e^{-2\beta h} \end{aligned}$$

3.1.4 *Sample statistic calculations*

The notation given in Table 3.2 (from Cowpertwait et al., 2002) is used for the variance, coefficient of variation, skewness, autocorrelation (first lag), and cross-correlation respectively. The dependent parameters are also shown.

Table 3.2: Statistic notation

Statistic	Dependent parameters	Description
ν_h	$\lambda, \beta, \mu_C, \alpha$	coefficient of variation taken over time intervals of width h.
κ_h	$\lambda, \beta, \mu_C, \alpha$	skewness coefficient taken over time intervals of width h.
ρ_h	$\lambda, \beta, \mu_C, \alpha$	first lag temporal autocorrelation taken over time intervals of width h.
$\rho_{x,y,h,L}$	$\lambda, \beta, \mu_C, \alpha, \phi$	lag L correlation of rainfall depths between sites at locations x and y taken over time intervals of width h.

As the functions *coefficient of variation* (CV), *skewness*, and *autocorrelation* are dimensionless as they do not depend on the scale parameter θ (Table 3.2), these functions can be fitted to their equivalent dimensionless sample values (see Cowpertwait et al., 2002). That is, the mean-scaled sample statistics calculated by pooling data across multiple sites as follow in equations 3.11 to 3.16. Note that the notation and equations follow (Cowpertwait et al., 2002) reasonably closely.

Let N be the number of years on record
M be the number of sites in the region
h be a discrete time interval in hours
i be a season for the historical record ($i = 1, 2, \ldots, 12$)
$n(h, i)$ be the number of intervals of width h in season i
j be a year within the historical record ($j = 1, 2, \ldots, N$)
k be an interval within the month ($k = 1, 2, \ldots, n(h, i)$)
l be a site in the region ($l = 1, 2, \ldots, M$)
$x_{ijkl}^{(h)}$ be an observed rainfall depth of interval width h, season i, year j, interval k, and site l

In the equations (3.10 to 3.16) all data are assumed to be valid. In practice, the total number of values (eg: $M \times N \times n(h,k)$) has to be adjusted by subtracting the number of missing records within the time period examined. Similarly, no missing records are included in any summation of $x_{ijkl}^{(\mathbf{h})}$.

In order to calculate the pooled sample statistics, the sample 1-hour mean is first calculated (Equation 3.10).

$$\overline{x}_{il}^{(\mathbf{1})} = \sum_{j=1}^{N} \sum_{k=1}^{n(\mathbf{1},i)} \frac{x_{ijkl}^{(\mathbf{1})}}{Nn(\mathbf{1},i)}, i = 1,\ldots,12; l = 1,\ldots,M \tag{3.10}$$

The mean-scaled statistics for $\hat{\sigma}_{h,i}^2$, $\hat{\gamma}_{h,i}$, $\hat{\nu}_{h,i}$, $\hat{\rho}_{h,i}$, $\hat{\kappa}_{h,i}$, and $\hat{\rho}_{x,y,h,i}$ can then be obtained by pooling all available data across each site and year (Equations 3.11 to 3.16).

$$\hat{\sigma}_{h,i}^2 = \sum_{l=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{n(h,i)} \left[\frac{(x_{ijkl}^{(h)}/\overline{x}_{il}^{(\mathbf{1})} - h)^2}{MNn(h,k)} \right] \tag{3.11}$$

$$\hat{\gamma}_{h,i} = \sum_{l=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{n(h,i)-1} \left[\frac{(x_{ijkl}^{(h)}/\overline{x}_{il}^{(\mathbf{1})} - h) \cdot (x_{ij(k+1)l}^{(h)}/\overline{x}_{il}^{(\mathbf{1})} - h)}{MN(n(h,k)-1)} \right] \tag{3.12}$$

$$\hat{\nu}_{h,i} = \frac{\hat{\sigma}_{h,i}}{h} \tag{3.13}$$

$$\hat{\kappa}_{h,i} = \sum_{l=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{n(h,i)} \left[\frac{(x_{ijkl}^{(h)}/\overline{x}_{il}^{(\mathbf{1})} - h)^3}{\hat{\sigma}_{h,i}^3 MNn(h,k)} \right] \tag{3.14}$$

$$\hat{\rho}_{h,i} = \frac{\hat{\gamma}_{h,i}}{\hat{\sigma}_{h,i}^2} \tag{3.15}$$

$$\hat{\rho}_{x,y,h,i} = \frac{\sum_{j=1}^{N} \sum_{k=1}^{n(h,i)} (x_{ijkx}^{(h)} - \overline{x}_{ix}^{(\mathbf{1})})(x_{ijky}^{(h)} - \overline{x}_{iy}^{(\mathbf{1})})}{\sqrt{\sum_{j=1}^{N} \sum_{k=1}^{n(h,i)} \left(x_{ijkx}^{(h)} - \overline{x}_{ix}^{(\mathbf{1})} \right)^2 \sum_{j=1}^{N} \sum_{k=1}^{n(h,i)} \left(x_{ijky}^{(h)} - \overline{x}_{iy}^{(\mathbf{1})} \right)^2}} \tag{3.16}$$

The pooled sample estimates (ν_h, κ_h, ρ_h, and $\rho_{x,y,h,i}$; where $\rho_{x,y,h,i}$ is the lag 1 cross-correlation between x and y for season i) for the three aggregation levels (h=1,6,24) can be used in the model fitting as follows.

3.1.5 Model fitting

The fitting process is divided into two parts. Firstly, the parameters dependent on the dimensionless statistics ($\lambda, \beta, \eta, \mu_C$, and α) are fitted. The results from this first step are used to obtain an estimate for θ. Finally, ϕ is estimated. Furthermore, by fitting ϕ after the other parameters, the dimensionality of the solution space is reduced and a fit can be obtained more quickly.

Rather than use the raw estimates obtainable from the data, the pooled sample statistics were smoothed using forward stepwise harmonic regression on the first three harmonics as in equation (3.17).

$$\hat{g}_i = c_0 + \sum_{j=1}^{3} \left[c_j \cdot cos\left(\frac{2\pi\, ij}{N}\right) + s_j \cdot sin\left(\frac{2\pi\, ij}{N}\right) \right] + \epsilon_i;\ i = 1, 2, \ldots, N \quad (3.17)$$

where N is the number of seasons, $\hat{g}_i$ is a sample statistic for the ith season, only significant coefficients (c_j, s_j) are included in the final harmonic regression model, and $\hat{f}_i$ is the corresponding predicted value.

The forward selection of the model coefficients minimised the model's AIC (where AIC is the standard Akaike Information Criterion). This method gives estimates for each aggregation level h of the coefficient of variation, $\hat{\nu}$, skewness, $\hat{\kappa}$, and autocorrelation, $\hat{\rho}$, for each month, i, that vary smoothly over the seasons.

This smoothing of the statistics before model fitting is less than ideal as it relies on prior knowledge (see Onof et al., 2000). However, as an alternative procedure is beyond the scope of this work, this approach is retained. It should be noted however that in some cases, for example monsoon climates, this assumption does not hold (as defined above) as the transition between seasons is not smooth.

The model is fitted using smoothed mean-scaled statistics calculated for the 1-hour, 6-hour, and 24-hour aggregation levels to the now dimensionless functions: coefficient of variation, skewness, and lag 1 autocorrelation. A simpler model

using only the 1-hour and 24-hour components is also fitted and the two models compared. The best model is used to simulate data which is then used within the infilling algorithm.

The temporal component of the model is fitted using the mathematical equations given in Section 3.1.3, by minimising equation 3.18 for each month.

$$F_{i,A} = \sum_{k=1}^{n} \left[(1 - \frac{f_{ik}}{\hat{f}_{ik}})^2 + (1 - \frac{\hat{f}_{ik}}{f_{ik}})^2 \right] \tag{3.18}$$

such that f_{ik} is a dimensionless Neyman-Scott function and $\hat{f}_{ik}$ is the corresponding smoothed estimate from the sample data (Equation 3.17).

As the the 1-hour, 6-hour, and 24-hour aggregation levels are used, $\hat{f}_{ik}$ takes the corresponding $\hat{\nu}_1, \hat{\kappa}_1, \hat{\rho}_1, \hat{\nu}_6, \hat{\kappa}_6, \hat{\rho}_6, \hat{\nu}_{24}, \hat{\kappa}_{24}$, and, $\hat{\rho}_{24}$ functions respectively.

Upon the completion of the minimisations of equation 3.18, the scale parameter, θ, can then be estimated for each site, l, month, i, from equation (3.19):

$$\hat{\theta}_{il} = \left[\frac{\bar{x}_{il} \hat{\eta}_i}{\hat{\lambda}_i (\Gamma(1 + \hat{\alpha}_i) \hat{\mu}_{c_i}} \right]^{1/\hat{\alpha}_i}, i = 1, \ldots, 12; l = 1, 2, \ldots, M \tag{3.19}$$

where $\bar{x}_{ik}$ is the 1-hour sample mean calculated for each site-month.

After the temporal components have been fitted (Equations 3.18 and 3.19), the spatial parameter, ϕ, for the cell radius is fitted using the estimates obtained for ν, κ, and ρ in equation (3.20).

$$F_{i,B} = \sum_{x=1}^{M-1} \sum_{y=x+1}^{M} \left[\left(1 - \frac{\rho_{x,y,1,i}}{\hat{\rho}_{x,y,1,i}} \right)^2 + \left(1 - \frac{\hat{\rho}_{x,y,1,i}}{\rho_{x,y,1,i}} \right)^2 + \left(1 - \frac{\rho_{x,y,24,k}}{\hat{\rho}_{x,y,24,k}} \right)^2 + \left(1 - \frac{\hat{\rho}_{x,y,24,i}}{\rho_{x,y,24,i}} \right)^2 \right] \tag{3.20}$$

where M is the number of sites.

It is necessary to fit the 1-hour aggregation level in equation (3.20) for two reasons. Firstly, when a model is used to generate synthetic records, the data are generated at a 1-hour level and then aggregated up to a 24-hour level. This was

done so that disaggregation of the 24-hour records to 1-hour records could be readily computed using the synthetic records. Secondly, as mentioned in Section 3.1.2, the difference between the data collection at the 1-hour and 24-hour aggregation levels (tipping buckets versus rain storage gauges) will have a discernable effect on the data characteristics. Therefore, in spite of the limited availability of the 1-hour data, this aggregation level was included in the fit. The 6-hour cross-correlation is even more limited in value as it is subject to even greater error. As a result, the 6-hour aggregation level was not included in the fitting function (Equation 3.20).

It was assumed throughout this thesis that a single model was adequate to compute both infilling and disaggregation. However, due to the accuracy of the data collection, it may be necessary to fit two models. Where the first model is fitted to the 24-hour aggregation level for infilling, and the second model is fitted to the 1-hour (and 6-hour) aggregation level(s) for disaggregation.

The minimisations of $F_{i,A}$, $\hat{\theta}_{il}$, and $F_{i,B}$ are repeated for each month, i, giving a total of 12 estimates for $\hat{\lambda}_i$, $\hat{\beta}_i$, $\hat{\eta}_i$, $\hat{\alpha}_i$, $\hat{\mu}_{c_i}$, and $\hat{\phi}_i$ along with 12×23 estimates of $\hat{\theta}_{il}$; i=1, 2, ..., 12 and $l = 1, 2, \ldots, 23$.

3.1.6 Verification

In order to verify the model has been fitted correctly it is necessary to show that a record generated using the fitted model has the same statistical properties as the historical data. However, it is insufficient, though necessary, to show that the pooled estimates used to fit the model approximately match between the historical and simulated records. If the simulation algorithm is accurate, the pooled and fitted estimates are expected to match within sampling error. Therefore, the following procedures were used to verify the model is fitting correctly and the simulation algorithm is working.

(a) Plots of fitted statistics: 35-year simulation and historical by season (Section 5.3.1)

(b) Quantile-Quantile plot of pooled site data by season (Section 5.3.1)

(c) Comparison of monthly statistic behaviour: mean, coefficient of variation, skewness, and lag 1 autocorrelation, by month and site (Section 5.3.2)

In addition, the stability of the statistics for the 300 year simulation was investigated (Section 5.3.3). The simulation was repeated 30 times and a boxplot of the fitted statistics was produced for each season. Obviously a repetition of 30 samples is insufficient to deduce the definitive distribution shape for the sample statistics. Due to the large sample used to calculate the statistics, however, the 30 samples are expected to provide a reasonable indication of the fitted statistics' distributions from the Central Limit Theorem.

3.2 Data analysis

As with any investigation, the data must first be shown to be capable of supporting the research hypotheses. That is, they should be proven, as much as possible, to be reliable and to have the characteristics required for the desired model to be fitted. Considerable attention was paid to determining the validity of the data as the constructed model was to be as reliable as possible for the infilling algorithms. Also, of equal importance, if invalid historical data are used as input into an infilling algorithm, then the results will not be as 'accurate' as they could be. Thus the invalid records would introduce a confounding factor into the algorithm analysis.

The data analysis is broken into two sections. The first section deals with the temporal domain and focuses on the internal consistency of the data at each site. This is the primary analysis for determining the validity of the source data. The second section examines the spatial component of the historical records. This latter section is used to confirm whether any discrepancy observed in the temporal analysis is real or whether it is consistent with the behaviour at other sites. If the records are consistent across multiple sites then it is assumed that the data are genuine observations rather than erroneous.

3.2.1 *Exploratory plots: time*

There were four primary plots conducted for each aggregation level in order to validate the historical data temporally for each site.

(a) Data versus time

(b) Monthly summary statistics (mean, median, and maximum) versus time

(c) Boxplot of data split by season

(d) Proportion of dry days versus time

The above plots allow an understanding of the behaviour of the data at each site to be gained. Potential outliers can immediately be seen along with any unusual behaviour at the site. These plots were conducted for the 24-hour record and the 1-hour record. The latter is more difficult to validate due both to the relatively short record length and the inherent variability within the data at this aggregation level.

3.2.2 *Exploratory plots: spatial*

As the data modelled are spatial and temporal, some understanding of the spatial characteristics also needs to be ascertained. To investigate the data fully, a pairwise scatter plot should be completed; however, this would construct 3312 plots per aggregation level ($(23 \times 24)/2 \times 12$). This is an excessive number of plots to view.

Although the 1-hour data should be examined, if the 1-hour data are spurious, then the 24-hour data will also be spurious. Therefore, the 1-hour data can be examined indirectly through the 24-hour totals. The number of examinable plots can therefore be reduced significantly as only the daily data are analysed. Furthermore, a method of sampling these plots is formulated (Figure 3.3) in order to select only those plots that highlight potential issues with the source data.

The summary statistics were used as input into the regression models (Figure 3.3) rather than the raw data to (partially) resolve the following two issues. Firstly, using the raw data would require the use of a General Linear Model (GLM) to

for all seasons (S)
    for all sites $(X; X = 1, 2, \ldots, N-1)$
        for all sites $(Y; Y = X+1, \ldots, N)$ {
            Construct a linear model of the monthly proportion dry $site_X$ versus the monthly proportion dry $site_Y$ for S

            Construct a linear model of the monthly maximum $site_X$ versus the monthly maximum $site_Y$ for S

            If any absolute residual in either of these models is greater than 3, then plot $site_X$ versus $site_Y$, along with $site_X, site_Y$ versus time for the month
        }

Figure 3.3: Scatterplot selection algorithm

satisfy the regression assumptions for the normality of residuals. In theory, the linear model using the maximums should also use a GLM; however, as a relatively high linear correlation is expected between monthly maximums, and as the model is not used for prediction, this requirement was ignored. Secondly, a month where the summary statistics deviated significantly from other sites at the same time period would be of more significance than if a single point differed - especially for the 1-hour record where this is expected to occur frequently.

An x, y scatter plot, a plot of the monthly statistics used to select the site pair (including the fitted regression line), and a 'versus time' plot for any of the problem years causing the selection of this site pair was constructed for each selected site pair. The regression lines are also plotted as the effects of extreme values on the fitted line are also important, especially for the monthly maximums.

3.2.3 Data removal

Once a period of data records had been identified as spurious in either Section 3.2.1 or Section 3.2.2, the data was completely replaced as missing over this period. No attempt was made to make use of the potential information contained within the period at that site - for example monthly totals given on the last day of the month.

The recording of invalid data was assumed to be independent - invalid at random. If the data are dependent (between sites) then it is assumed that the data records are not spurious. While this assumption is dubious, in the absence of any further information, it is the most conservative approach.

When searching for spurious data, it is also important to note that it may be possible to identify data as doubtful or invalid purely because of patterns occurring at multiple sites which are implausibly consistent. For example, erroneous recording due to a data recording procedure, is expected to be relatively consistent at multiple sites. Therefore, particular attention was paid to the boundaries of the currently identified missing periods within each site record - particularly to any consistent behaviour observable at these boundaries.

3.2.4 Assumptions

The main assumptions required by the Spatial-Temporal NSRP model as fitted in this thesis, are approximate temporal and spatial stationarity (Section 3.1.1). Rather than show the source data are approximately stationarity in time and space, however, the data will be assumed to have these characteristics. The model will then be fitted against the historical record and a simulation of 300 years will be used to determine whether the model accurately captures the variation in the historical record. That is, each site is examined separately by season for each monthly sample statistic as follows: mean, coefficient of variation, skewness, and first lag autocorrelation. If the assumption of stationarity is approximately valid, then the monthly statistics should have a similar variation for the model and the historical data. Note that some leniency is required as the record length of the historical data may be inadequate, especially at the 1-hour level, to draw a conclusive match/mismatch.

It is also assumed that missing values in the historical record are missing at random (MAR) - at least in the temporal domain. In particular, extreme events must be recorded accurately as otherwise the process will not be able to be modelled correctly. Although for some spatial processes it is important for the data to be MAR (Haining, 2003) the high cross-correlation for the rainfall data in

this example is expected to reduce the necessity for this assumption to be satisfied. However, if the data were not spatially MAR then this may negatively impact the spatial component of the model fitting especially if the data record is too short.

In any case, the extreme event recording is of more importance than spatial correlation of missing values, for the latter should not have as great an effect provided the length of the valid records is sufficiently long. Unfortunately, the assumption MAR recording of extreme events can not be tested without recourse to external information, and, as a result, is left untested.

3.3 Infilling

The initial algorithm for infilling is based on an ordinary least squares fit and provides the basis from which all subsequent algorithms can be compared. All infilling algorithms, however, rely on 300-year synthetic record generated using a fitted model (eg: Section 5.2) from which sample rows from the simulation can be selected to infill missing data in the historical record. Thus, the infilling algorithms discussed are variants on *model based* imputation methods (see Haining, 2003).

3.3.1 Algorithm evaluation

First a model was fitted to the historical record and a 300-year record was generated. Subsequently, 20% of the valid historical data (stratified by site and season) were replaced with missing values which could then be compared against the historical records. The infilling algorithms then were tested by comparing the true historical values for the sampled out records with the corresponding infilled records. This direct comparison was needed as it is of interest how well the infilling algorithm predicts the occurrence of extreme values.

Although this replacement could have been done prior to the model fitting so that the information incorporated in the 'missing values' was not included in the model, this approach was not used for two reasons. Firstly, it was computationally much faster to apply one model (and therefore generate one 300-year simulation) as a basis for all samples to be infilled taken from the historical record. Secondly,

as the results in Section 4.4.2 show, the effect of erroneous data does not have a significant effect on the pooled statistics *CV*, *skewness*, and *autocorrelation*. A similar result will result for the first order cross-correlation if the replaced records are replaced at random. Therefore, data that are randomly replaced will not have a significant effect on the pooled statistics or, as a result, the fitted model parameters. Thus, a separate model fit per sample was deemed unnecessary.

A set of 100 independent tests are conducted for each algorithm. As the random number generator was set to the same starting seed for the simulator and for the selection of historical random samples, all algorithms used the same source simulation and infilled the same selected points over the 100 samples. Thus a direct comparison can be made on the performance of the algorithms as the effect of random samples and simulations are controlled.

From the 100 samples obtained for each algorithm, the true records and the infilled records are compared directly via intensity plots and quantile-quantile plots. Furthermore, the pooled statistics (Equations 3.13 to 3.16): CV, skewness, autocorrelation (see note below), and cross-correlation used to fit the model are calculated for the historical and infilled records respectively and compared. Lastly, the counts of all dry sites, mixed wet/dry sites, and all wet sites within the historical and infilled record were compared as it was found (Section 5.4.1) that the spatial-temporal NSRP model did not produce the same regional characteristics as the historical record.

Note that the first lag autocorrelation was always calculated by using the previous historical data (even if the previous row was selected for infilling). Similarly, any infilling algorithms making use of the previous and next time points (Section 5.5.6) also revert back to using the historical data, not any infilled data.

The testing of the infilling algorithms, therefore, comprises a harder problem than infilling the missing data in the record. This is obvious as the number of available records that can be used to fit the historical record accurately has been reduced by the removal of 20% of the valid data. Therefore, although the accuracy of the infilling algorithms discussed may not be ideal, the magnitude of the error is expected to be significantly smaller when just the missing records in the historical data are infilled.

3.3.2 Notation

Notation used for the infilling algorithms in Sections 3.3.3 to 3.3.5 is provided in Table 3.3. If the infilling algorithm is to be useful, the seasons must be equivalent between the historical and simulated records. Similarly the order of the l sites must be equivalent between H and S (Table 3.3). However, interpolation between sites is possible using an infilling algorithm if the historical record is set to be missing for this interpolated site l for all H_{ijkl}. For the interpolated site l to be simulated, an estimate of θ_{il} must be specified within the simulation algorithm.

Table 3.3: Infilling notation

Let	H	be the historical record,
	S	be the simulated record,
	i	be a season,
	j	be a year within the historical record,
	J	be a year within the simulated record,
	k	be a day within the season,
	l	be a site in the region,
	M	be the number of sites in the region,
	$n(i)$	be the number of days in season i,
	N	be the number of years in the record (denoted N_j and N_J for the historical and simulated records respectively),
	$F_{il}(x)$	be the cumulative distribution function for season i, site l obtained from the simulated record
	$I(x)$	be an indicator function (Equation 3.21) to show whether the current record is valid or missing

$$I(x) = \begin{cases} 1, & x \text{ is a valid record} \\ 0, & otherwise \end{cases} \tag{3.21}$$

All infilling algorithms can be specified as a function, $G(H_{ijkl}, \ldots)$; where the minimum argument is the record to be infilled. Furthermore, as the infilling algorithms considered make use of a simulated record, the infilling function take the form $G(H_{ijk}, S, i)$; where the algorithms are restricted to use only those records within the ith season to ensure the stationarity of the infilled records is maintained.

A brief outline of the infilling algorithms is given in the following sections.

3.3.3 *Best fit least squares*

The first infilling algorithm searches the synthetic record for a best fit to the non-missing data in the historical data minimising the sum of squares between the observed (H_{ijk}) and the fitted values (S_{iJk}) - Figure 3.4. Note this is a strict definition, the implemented algorithm (see also Section 5.5.2) differs slightly from the specification in Figure 3.4.

BFLS $<-$ function(H_{ijk}, S, i) {
  let $x = H_{ijk}$,
    $S = S_i$
    $SS^{min} = \infty$
  for (J in $1, 2, \ldots, N_J$) {
    for (K in $1, 2, \ldots, n(i)$){
      $SS = \sum_{l=1}^{M} I(x_l) \cdot (x_l - S_{JK})^2$
      if $(SS < SS^{min})$ {
        $SS^{min} = SS$
      }
    }
  }
  for (l in $1, 2, \ldots, M$) {
    if $(I(x_l) == 0)$ {
      $x_l = SS_l^{min}$
    }
  }
}

Figure 3.4: BFLS: infilling algorithm definition

Implemented algorithm

In the implemented algorithm, the record SS^{min} is selected, but on the event of a tie, one of the tied records is chosen with equal probability (in Figure 3.4 the first minimum record encountered is chosen with probability 1). Whenever an exact match is found ($SS^{min} == 0$), the algorithm terminates with a 50% probability. (This potential quick termination was used to speed up the algorithm.) Furthermore, in order to prevent bias toward the first point(s) encountered, whenever a

time point is selected for infilling, the order of selection of values in SS_i is adjusted so that the selected point is placed at the end of the K loop.

Note that all the infilled points come from the the same time point in the simulated record - the row of best fit.

3.3.4 Best fit CDF least squares

The best fit CDF least squares algorithm (Section 5.5.3) uses the simulated record to obtain an approximate CDF of the rainfall data at each of the sites for each month - 23×12 CDFs in all. Rather than fitting the raw data using least squares directly, the CDF was used to transform the data, either historical or simulated, on to a $[0, 1]$ scale according to the CDF for the site being examined (Figure 3.5).

BFCDFLS $<-$ function(H_{ijk}, S, i) {
  let $x = H_{ijk}$,
    $S = S_i$
    $SS^{min} = \infty$
  for (J in $1, 2, \ldots, N_J$) {
    for (K in $1, 2, \ldots, n(i)$){
      $SS = \sum_{l=1}^{M} I(x_l) \cdot (F_{il}(x_l) - F_{il}(S_{JK}))^2$
      if $(SS < SS^{min})$ {
        $SS^{min} = SS$
      }
    }
  }
  for (l in $1, 2, \ldots, M$) {
    if $(I(x_l) == 0)$ {
      $x_l = SS_l^{min}$
    }
  }
}

Figure 3.5: BFCDFLS: infilling algorithm definition

The BFCDFLS algorithm is expected to improve the fitting as:

(a) all fitted records are on the same scale regardless of location,

(b) the transformed values are equally distributed throughout the $[0, 1]$ interval (unlike the raw data where extreme values are sparse), and

(c) as a direct result of (b), rainfall of greater magnitude is fitted less stringently than rainfall of lower magnitude - thus increasing the candidate rows when extreme rainfall is encountered.

The same adjustments made to the BFLS algorithm (Section 3.3.3) were also applied to the implementation of the BFCDFLS algorithm.

3.3.5 Iterative least squares

The iterative sampling method (Figure 3.6), while retaining the use of the CDF fitting method presented in Section 3.3.4, differs greatly from the previous best fit algorithms discussed. Firstly, each missing element is iteratively fitted and replaced, one missing value at a time. Secondly, the fits are biased toward the best fitting row, but, unlike in the best fit algorithms, this bias is not absolute.

Note that the function $uniform$ generates a random uniform variate in the interval (x, y) specified as arguments to the function. Thus, a uniform random selection out of the best 5% of the candidate rows is applied. This ensures that it is possible for infilled records to come from multiple source rows in the simulation record, circumventing the problem observed in Section 5.4.1.

3.3.6 Further application

The primary focus of the infilling is for 24-hour data, however, the same techniques can be used for 1-hour data if valid records are available. For this analysis, however, it is shown in Section 5.4.1 that the fitted model does not maintain the same characteristics as the historical data. Furthermore, as the accuracy of the 1-hour records is questionable (see Tilford et al., 2003), it is likely to be better to model the higher aggregation levels and disaggregate from these records than it is to infill the 1-hour aggregation level directly.

ISCDFLS $<-$ function(H_{ijk}, S, i) {

let $x = H_{ijk}$,
$S = S_i$

for (l in $1, 2, \ldots, M$) {

if $(I(x_l) == 0)$ {

for (J in $1, 2, \ldots, N_J$) {

for (K in $1, 2, \ldots, n(i)$){

$$SS_{JK} = \sum_{l=1}^{M} I(x_l) \cdot (F_{il}(x_l) - F_{il}(S_{JK}))^2$$

}

}

$Sorted_{SS} = sort(SS_{JK})$

$U = uniform(1, 0.05 \times N_J \times n(i))$

$SS^{min} = Sorted_{SS}[U]$

$x_l = SS_l^{min}$

}

}

}

Figure 3.6: ISCDFLS: infilling algorithm definition

3.4 Implementation

One of the side products for this thesis is a computer software package capable of applying the algorithms developed herein to any rainfall data provided it is in an ASCII text format. This program was intended to be able to read in the source data, fit a spatial-temporal NSRP model to the pooled statistics, use the fitted model to generate a synthetic record, then infill any missing data elements in the historical record.

The programming language, C++, was chosen for the application due to its portability, versatility, speed, and object-oriented capabilities. The previous software written by Paul Cowpertwait was written in a combination of *C* and *R* (see `http://www.r-project.org/`) - where *R* was used for the fitting algorithms, plotting routines, and statistical tests. The move to C++ in preference to *C* was mainly to implement a more obvious hierarchical structure to be applied allowing easier maintenance, debugging, and greater reusability of source code.

As portable software was an objective without reliance on external products, the use of *R* was avoided wherever possible. For this analysis, *R* was used for the graphical plots as it was deemed more imperative for the infilling, model fitting, and simulation algorithms to be implemented rather than graphical components.

Various statistical libraries are available for *C* (and C++), but these were not used in preference to development of specialised algorithms which are expected to be more efficient and easier to debug. In addition, no copyright laws are infringed as each algorithm is either implemented from scratch or the algorithms used are released under a public license.

As a stand alone package, this program also requires a number of statistical algorithms to be implemented. Therefore, the following algorithms were coded as part of the program in order to complete the project with sources annotated where appropriate:

(a) Random number generator (Matsumoto, 2004).

(b) Numerical integration (based on Gerald and Wheatley, 1984).

(c) Nelder-Mead simplex algorithm (modified to use parallel searching to reduce the possibility of optimisation failure). Based on Lagarias et al. (1998) and Press et al. (2002).

(d) 1-dimensional optimisation (specialised). (see Press et al., 2002, Chapter 10, p405-406).

(e) Approximation to the Incomplete Gamma function (Toth, 2004).

(f) Standard regression algorithms, harmonic regression, and forward stepwise regression minimising AIC

(g) Two-sample Kolmogorov-Smirnov test and χ^2 test

The random number generator was already available in C code so no major alterations were necessary. The other algorithms were implemented directly in C++ from either a text description or, when available, source code.

4. DATA INTEGRITY ANALYSIS

> If you can't have an experiment, do the best you can with whatever data you can gather, but do be very skeptical of historical data and subject them to all the logical tests you can think of.
>
> **-Robert Hooke, quoted in Gaither and Cavazos-Gaither (1996)**

4.1 Introduction

A model should never be naively fitted to data without the data first checked to see whether the data are reliable. If the data are unreliable then the model estimates will be untrustworthy, regardless of how well the model fits. In addition, any infilling algorithm which makes use of the historical record will also be defective if spurious values are used.

As with any experimental analysis, data are removed very reluctantly, so for any section of the historical record to be removed the data must be obviously defective. Alternatively, if the suspect records span only a short period and thus will not affect the overall distribution behaviour, then these are also removed. Any identifiable untrustworthy records are removed prior to the model fitting.

This chapter is broken into four sections as follows. The first section examines the integrity of the source data in the temporal domain. Within this section, sites are identified as containing temporally inconsistent data. Depending on the severity of the discrepancy, these records are either removed or left for the spatial analysis to confirm as valid. The second exploratory section analyses the spatial

consistency of the source data with particular attention paid to those data segments identified previously as questionable. Any points that can be confirmed as spatially and temporally inconsistent are removed. The next section outlines the changes in the pooled statistics used as input into the model to determine the effect of the spurious records on these statistics. The final section within this chapter briefly summarises the results found in the first three sections.

Within this chapter, as it is only necessary to distinguish between the 1-hour and 24-hour aggregation levels, the terms 'hourly' and 'daily' respectively will be used.

4.1.1 Known issues

A consistent 'problem' was the low amount of data at the end of the available records in 2003 that contributed to a supposed maximum of $0mm$ of rainfall and a proportion dry of 1. These points were not commented on even though they are unusual as this is due to the sequence of available data ceasing around this time.

In all other cases, any sequence with a proportion dry of 0, or where the skewness or variance was 0, was examined for credibility (particularly if it is seen as unusual on any plots). This was important as it was possible that some value had been imputed over the whole month.

While many of the sites had a long record, two of the sites, TW239315 and TW289022, had a much shorter record. After the invalid records were removed (see Section 4.4), site TW239315 (number 4), had an hourly record with less than 15% valid data over the 34-year period (Table 4.1). Also for this site, less than 30% of the valid daily data over the 34-year period was available. Similarly, site TW289022 (number 20) had less than 6% of the possible valid hourly data, and less than 58% of the possible valid daily data (Table 4.1).

For site TW246424, Thames Water specified that there was a period of three years (1997-1999) where missing values were coded as zero (Thames Water, personal communication, September 21, 2004).

4.2 *Exploratory analysis: temporal*

The records of daily data were much longer than those at the hourly level for each site. As these records will be used as input into the infilling algorithms, it was critical to show the daily records were valid. The hourly record was also important as these estimates were used to fit the spatial-temporal NSRP model, but this record was harder to validate due to the shorter recording period (1989 onwards).

Within this section, the time component of the historical data was examined. The rainfall at time u must be consistent with the rainfall at time v, for all $v \neq u$; where particular attention is paid to the immediate surrounding data, but also the seasonal characteristics.

For each site (see also Table 1.1; Figure 1.1 for the site locations), the following four plots per aggregation level were constructed as follows:

(a) Rainfall versus time

(b) Monthly maximum, mean, and median versus time

(c) Boxplot of rainfall versus season

(d) Monthly proportion dry versus time

Of these four plots, (b) and (d) are the most useful for detecting potential invalid points. However, (a) or (c) are important for the detection of any erroneous negative values.

These plots were grouped for each site (Figures 4.1,4.2 and A.1 to A.44); where it is clear that even a cursory examination of these figures reveal that a significant number of the sites (13/23) have potential recording problems. This is a substantial portion of the twenty-three sites provided. Each of the potential issues with the data are discussed in the following sections - partitioned by site.

4.2.1 TW238097

It was apparent from Figures 4.1 and 4.2 that there was an unusual sequence of recorded data in early 1990 (subsequently found to be April) that immediately

followed a period of missing data. From the proportion dry plots (Figures 4.1 and 4.2) a value of 1.0 was relatively unusual, and a maximum rainfall of $0mm$ was also unusual given the subsequent months. Also, as April is the spring rain season, it was expected that there would be at least some rainfall during this period. Therefore, it seemed likely that the period of missing rainfall had not been extended far enough; however the spatial analysis (Section 4.3) must confirm this conclusion.

4.2.2 TW238605

As shown in three plots in Figure A.1, there was an extremely unusual point in the month of July 1993. The actual recording was $92.2mm$ of rainfall, where $50mm$ of it fell in a single hour (Figure A.2). While this seemed extremely unusual, the hourly data that composed this total was not unreasonable so this point was deemed valid.

In addition, there were two months with a maximum of $0mm$ of rainfall as seen in the hourly plots (Figure A.2). This was found to be a nine-hour period which was a consistent issue with other sites. Therefore, this period (along with similar sequences observed at other sites) was treated as missing.

4.2.3 TW239258

From the boxplot (Figure A.3), there were some outliers, but as they fitted in with the rainfall versus time plot (Figure A.3) they were assumed to be valid.

4.2.4 TW239320

Barring one other site, this site had the most significant problem for the rainfall modelling and infilling algorithms. As shown in the boxplot (Figure A.7), just about every month had very extreme points. Furthermore, the data from where these extreme points came were not consistent. The first period of valid data ($1981 - 1989$), did not have the same characteristics as the latter period of valid data ($1994 - 2003$) - for any of the 'versus time' plots. Also of major concern was the high probability of having almost no wet days particularly over the period $1982 - 1985$.

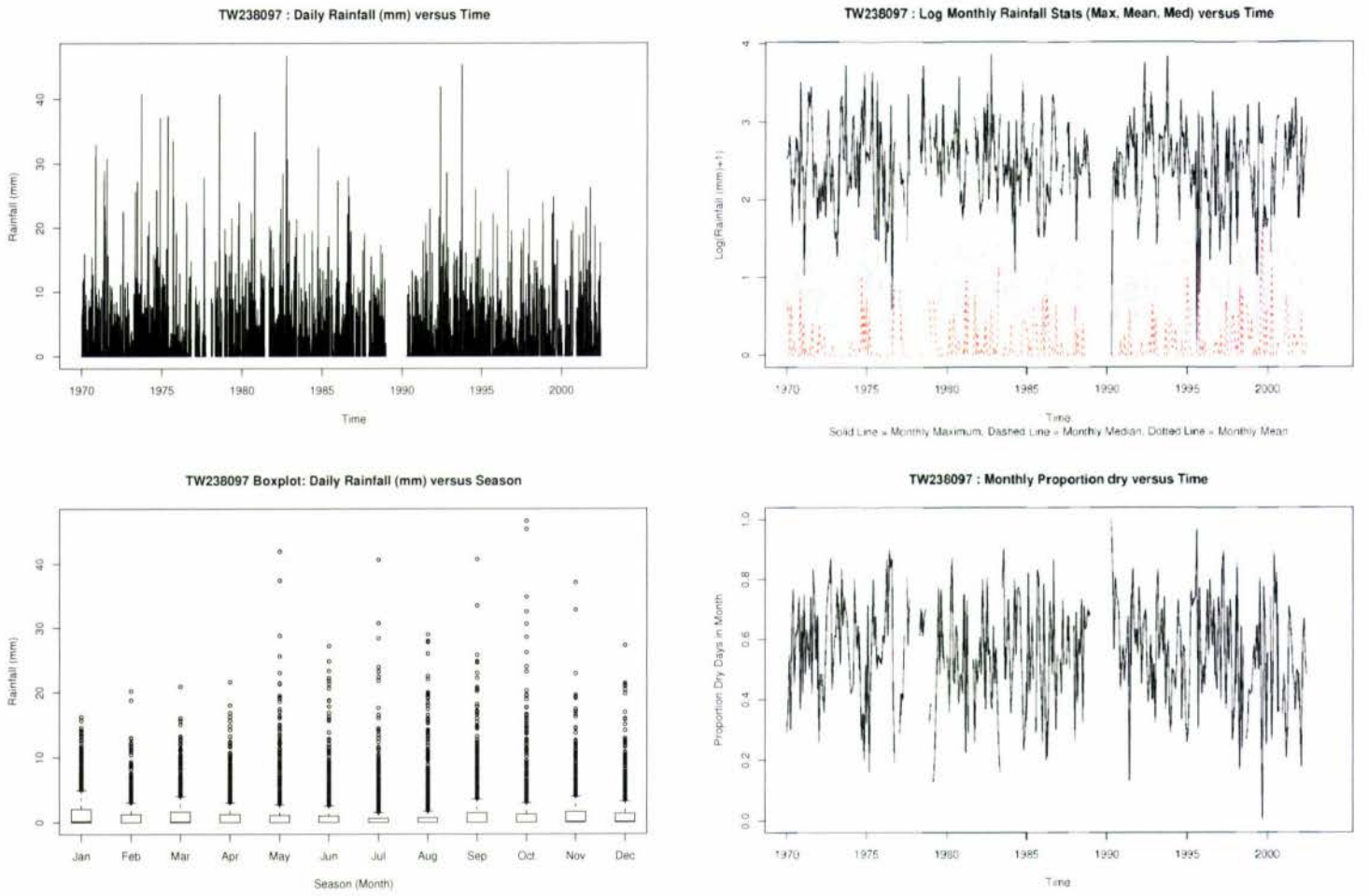


Figure 4.1: Site TW238097 daily plots

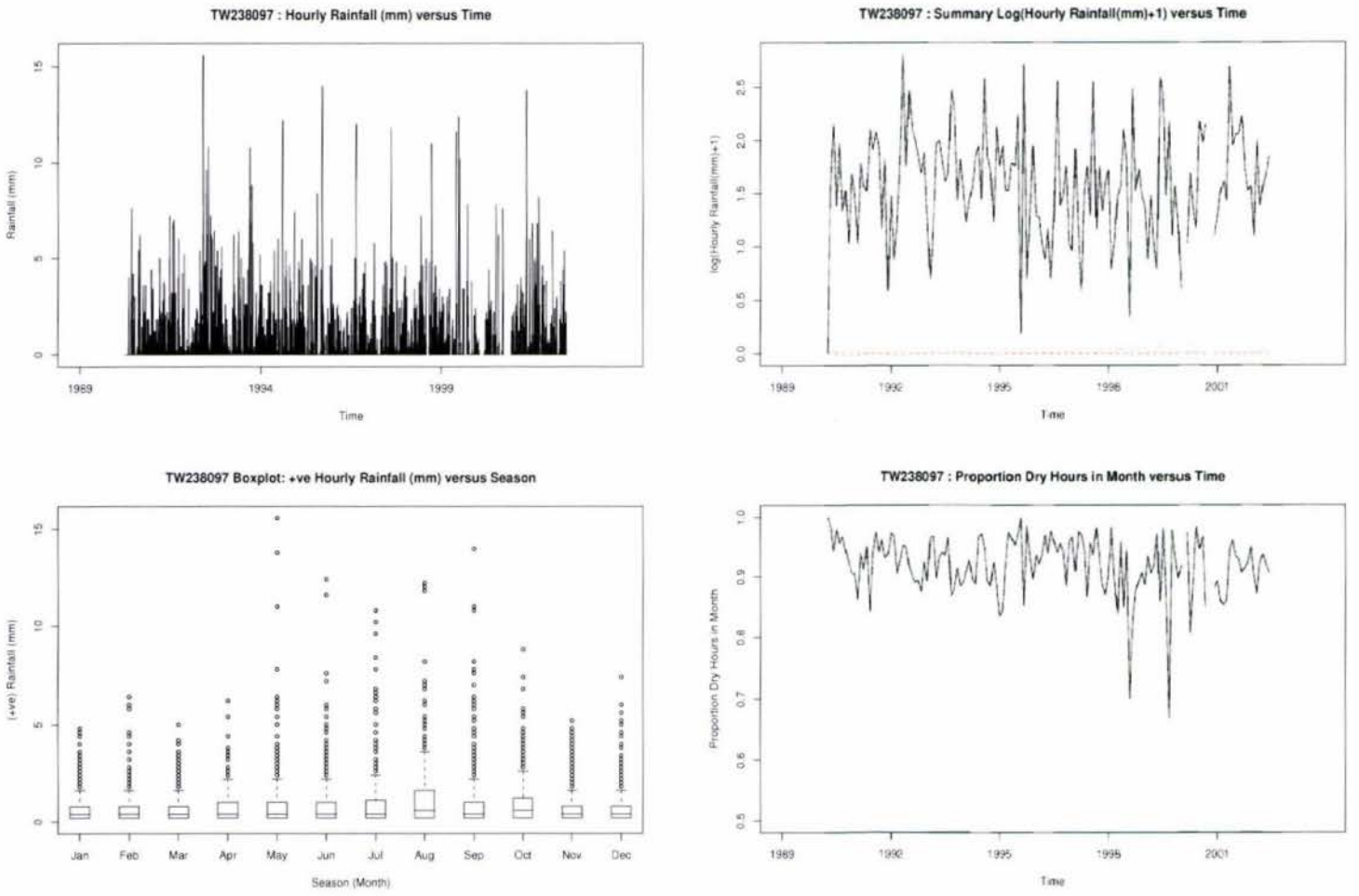


Figure 4.2: Site TW238097 hourly plots

A closer examination of these extreme points showed that they all occurred on the last day of the month where the other days in the month were all $0mm$. The likelihood of this happening for the number of extreme points observed, especially in winter, was so improbable the data were immediately identified as spurious.

It was more probable that the last day contained the monthly total rather than this sequence being a valid record for these months. Therefore, for any month in the period 1981 – 1989 with a strictly positive rainfall value on the last day in the month and zero everywhere else, all values within the month were recoded as missing. Although the value on the last day was most likely to be a monthly total, this information was discarded.

4.2.5 TW239374

There were three months (Figures A.9 and A.10) where a monthly maximum of $0mm$ was observed, of which only the first two were of concern. After scanning through the hourly data, the first unusual month, June 1999, was found to be due to only the first nine hours of data being recorded. This was also the case in July 2001 - the second unusual month in the series. As before (Section 4.2.2), the nine-hour series was treated as missing.

4.2.6 TW239578

The only problem shown in the Figures A.11 and A.12 was that there was a month with no rainfall. This month was found to be August 1995, but as this result was consistent with other sites during the same month, no action was taken either for this site or for any others where this dry period occurred.

4.2.7 TW245176

The only questionable period observed was a problem in the proportion dry versus time plot (Figure A.14), where the next set of records were missing. As for site TW239374, this was found to be due to only the first nine hours being recorded in February 1992 all of which had zero rainfall. Again, the data over this period were recoded as missing.

4.2.8 TW246424

Although this was known already, it was clearly evident from Figures A.17 and A.18, there was a period of data where missing values had been coded as zero (see Section 4.1.1). The zero values over this 3-year period were recoded as missing prior to the spatial analysis being computed as it was known that these were definitely spurious records.

4.2.9 TW246627

There were two potential problems evident in Figures A.19 and A.20. The first was the obvious case of zeroes immediately following a block of missing data - seen in both the hourly and daily records. As this was a consistent problem throughout the dataset, this first block of unlikely data (March 1989) was treated as missing. The second point seen was in the hourly plots (Figure A.20) in July 2001 where only the first nine hourly values in the month were available. Therefore, as consistent with the previous results, this period of nine hours was also recoded as missing.

4.2.10 TW246847

As for site TW246627, July 2001 had only the first nine hourly values recorded - all of which were zero. There were also some extreme levels of rainfall as shown in the two Figures A.21 and A.22; however these seemed consistent with the expected rainfall distribution. In addition, the low maximum occurring in August 1995, evident in Figure A.21, was consistent with the other sites over this period.

4.2.11 TW247119

The Figures A.23 and A.24, corresponding to the daily and hourly plots respectively, illustrated only two areas of concern. The consistent nine hour July 2001 problem was seen in the hourly data along with November 1999 with the same recording pattern. The low maximum in the daily plot corresponded to August 1995 which had been found to be dry with other sites and therefore was of no concern.

4.2.12 TW286392

There were several points of interest in Figures A.25 and A.26. There was a maximum of $0mm$ immediately following a period of missing data. As commented previously, the data may be valid, but, given the problem with zeroes being used to indicate missing values and the consistent problem with boundaries, the data were assumed to be invalid. Another nine point record was found in August 2000, corresponding to the second peak in the proportion dry plot (Figure A.26). The $70mm$ of rainfall was not untoward and was left as correct.

4.2.13 TW287141

As with sites TW238097 and TW246627, the only problem with this site, as revealed in Figures A.27 and A.28, was a block of missing values not extending far enough. The records were flagged as being potentially spurious, but this result was to be confirmed with the spatial analysis (Section 4.3) before any action was taken.

4.2.14 TW287283

The only obviously suspect set of points were those with no wet days in the period January 1991 to March 1991 as shown by the proportion dry plot in Figure A.29. This was confirmed as unusual for the hourly data also in Figure A.29. Therefore, these records were recoded as missing.

4.2.15 TW288020

Although this site had a high proportion dry for March 1993, the maximum hourly rainfall looks believable at the same point (Figures A.33 and A.34). However, one point of concern was that the hourly record terminated with a nine hour sequence before the missing data period begins. By implication, the zeroes recorded immediately prior to this segment were questionable. However, as no observed discrepancy was found in the spatial analysis (Section 4.3), the zero sequence was not identified as spurious and no further action was taken.

4.2.16 TW289022

This site had the usual problem with zeroes being recorded prior to a missing period, as shown in the proportion dry plot (Figures A.37 and A.38). Note also that the boxplot (Figure A.38), showed that December was radically different from the other months, but this seemed to be primarily due to the short record for this site. The daily boxplot (Figure A.37) showed that December was more like January as would be expected.

4.2.17 TW290007

The only point of interest evident in Figures A.41 and A.41 corresponded to August 1995 which was noted previously to be a dry period.

4.2.18 TW291467

Again zeroes recorded after missing periods were a problem for this site (Figure A.43). The dry August 1995 was consistent with other sites and was not an issue. The hourly data also showed May 2001 to be a problem with the proportion dry, which was found to be due to nine hour sequence.

4.3 Exploratory analysis: spatial

Within this section, the spatial consistency of the data was examined. The rainfall at time u site S must be consistent with the rainfall at time u site T; for all $T \neq S$. For the spatial analysis, particular attention was paid to the monthly maximums to determine whether the monthly behaviour was consistent. Also, any periods identified (Section 4.2) as potentially spurious, were examined more closely. Whenever data were replaced as missing over a period, this replacement was always applied to the daily record. The hourly record was only adjusted if data were recorded over this same period.

A complete examination of the spatial consistency of the rainfall data would require, at the bare minimum, pairwise plots for the site matrix for both aggregation levels for each season - a total of 6624 plots as noted previously (Section

3.2.2). As this was infeasible, the algorithm as discussed in Section 3.2.2 was used to select time points that may need further investigation from the daily data only. Note that the obvious problems with site TW239320 and site TW246424, outlined in Section 4.2, were fixed before the selection of plots was conducted.

The plot selection algorithm (Figure 3.3), reduced the number of plots requiring examination from above 3000 down to approximately 700. While this was a large number to view, it was not unmanageable. Upon examination of the plots, it quickly became evident that sites with 'problems' were presented in multiple plots. As a result, the plots were able to be further reduced to the results in Figures 4.3 to A.52.

There was considerable variation between the differing seasons - especially in regard to the monthly maximums between sites. The proportion dry was usually much more consistent and, as a result, should be trusted in preference to the modelling of the maximums - especially over the summer months.

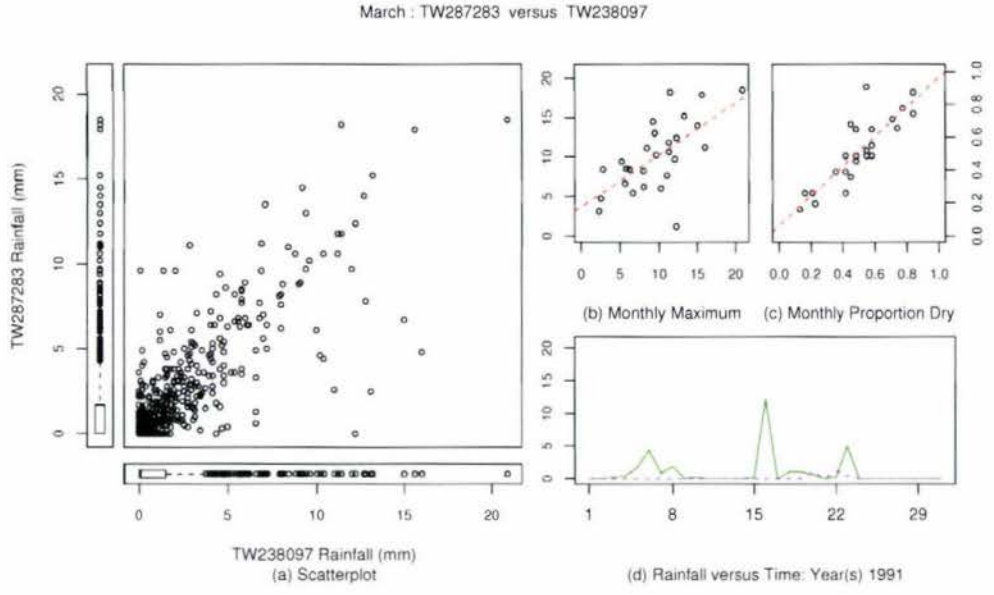
As in Section 4.2, the plots have been broken down and discussed by site. However, a site may not be discussed in this section if the only reason for its identification was due to missing values and/or a short sequence of valid data. In this case, it was assumed that these potentially spurious records had been identified as temporally inconsistent (for example the nine-hour sequence).

4.3.1 TW238097

The data recorded during March and April 1990 (eg: Figure 4.3) were not consistent with the records at this time at other sites. This was consistent with the explanation (Section 4.2.1) that this data should actually be recorded as missing. As the records have been shown to be temporally and spatially inconsistent, the zeroes over this period (March-April 1990) were recoded as missing.

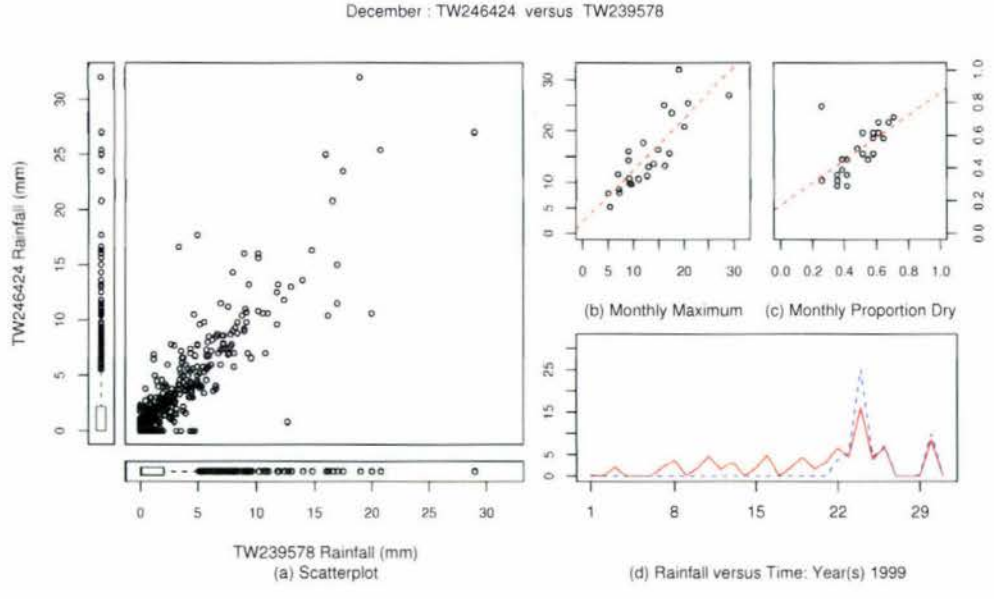
4.3.2 TW246424

Initially data were removed up to December 1999 (Section 4.2.8). However, from Figure 4.4, it was clear that zeroes were still being recorded as missing values up through mid December. Therefore, all zeroes up to the first non-zero record were recoded as missing.



Note: plot (d), Solid line is site TW238097, dashed line is site TW287283

Figure 4.3: Daily data, March, site TW238283 versus site TW238097



Note: plot (d), Solid line is site TW239578, dashed line is site TW246424

Figure 4.4: Daily data, December, site TW246424 versus site TW238578

4.3.3 TW246627

For this site, there were zeroes recorded from March 1989 through to the last week of April 1989. From the plots for April (Figure A.45), this was unlikely from the recordings of the other sites over this period. This concurs with the results shown previously, and the data were treated as missing.

4.3.4 TW247119

The only potential problem shown for this site was in September 1999 (Figure A.46). While this was not detected in the analysis on internal consistency, it seemed unusual that the initial half of this month was consistent with all sites plotted against it and the latter half was dry when the other sites recorded a reasonable level of rainfall. However, as this site was recording a low level of rainfall over this period the zeroes were left as valid.

4.3.5 TW287141

For both January 1990 (and the first week of February 1990), there were only zeroes recorded for this site (Figure A.47). This was both unlikely for the season (mid-winter) and location (England), but was also inconsistent with the rainfall recorded at other sites (Figure A.47). Therefore, the data were recoded as missing over this period.

4.3.6 TW287283

As in Section 4.2.14, the period of interest for this site was January 1991 - March 1991 (eg: Figure A.48). As with site TW287141, this was seen as additional evidence that the data should have been recorded as missing. The zeroes within the period January-March 1991 were replaced with the missing value indicator as the data was both spatially and temporally suspect.

4.3.7 *TW287864*

For this site, November 2000 was selected by the algorithm (Figure A.49). The initial records are $0mm$, but the remainder of the month followed other sites during this month relatively closely. However, as the internal analysis did not detect this period and as it was not an implausible scenario, the data were assumed to be valid.

4.3.8 *TW289022*

According to Figure A.50, November 1990 was unusually dry for this site given the surrounding rainfall at the other sites. As this record was immediately prior to an identified missing period (Section 4.2.16), it was likely that the data during this period should also be identified as missing. Therefore, the data during this month were recoded as missing.

4.3.9 *TW291467*

Two months were identified as spatially inconsistent for this site. The first month, November 1989, was dry for all but one day during the month (Figure A.51). As this month immediately followed a missing period, the first half of the month was recoded as missing, but the results were not completely implausible based on the surrounding site values.

The second period identified, December 2002, was chosen as the monthly maximum was not matched. The results in Figure A.52 indicate that the data for this site probably has been smoothed or interpolated. As the smoothing of one month would not affect either the pooled statistics or an infilling algorithm, the data were left unaltered.

4.4 *Summary statistics*

The summary statistics to be used for the model fitting (1-hour, 6-hour, and 24-hour) were extracted from both the cleaned data and the uncleaned data. These statistics were then plotted against each other (Figures A.53 and A.54).

The scatterplots of the CV, skewness or autocorrelation (Figures A.53) did not distinguish between the aggregation levels. From the plots (Figure A.53), it was evident that the CV, skewness, and autocorrelation were barely affected by the removal of the spurious data. This was probably due both to the standardisation by dividing by the appropriate site-month mean in the calculation of these statistics and the size of the samples - the erroneous data does not comprise a significant component.

For the cross-correlation plots (Figure A.54), it was clear that both aggregation levels were affected by the data cleaning - especially the daily aggregation level. The latter change was particularly due to the removal of the 3-year period of zeroes for site TW246424 and the removal of the incorrect months in site TW239320 (see Section 4.2.4).

4.4.1 Valid data

The percentage of valid data remaining in the historical record for both aggregation levels split by season was shown in Table 4.1. See Table 1.1 for the corresponding Thames Water identification names.

Particularly for the hourly data, the percentage of usable data for each month over the 34-year period that will be infilled (Section 5.5) was very low. The maximum percentage for the hourly data was less than 45% as hourly data were only collected after 1989 whereas the daily record was available since 1970.

The temporal statistics, provided the assumption of homogeneity is satisfied, were not expected to be badly affected by the relative sparseness of the historical record. The spatial statistics, however, will be affected - especially at the hourly level for those sites with less than 20% of valid data within a month.

The incorrect recordings would reduce the effectiveness of any infilling routines, and by implication, any disaggregation algorithms when infilling is computed prior to disaggregation. Therefore, all records identified as spurious are removed from the analysis from this point forward. The only exception to this was the possible smoothing for site TW291467 as this is not expected to affect the performance of the infilling algorithms significantly.

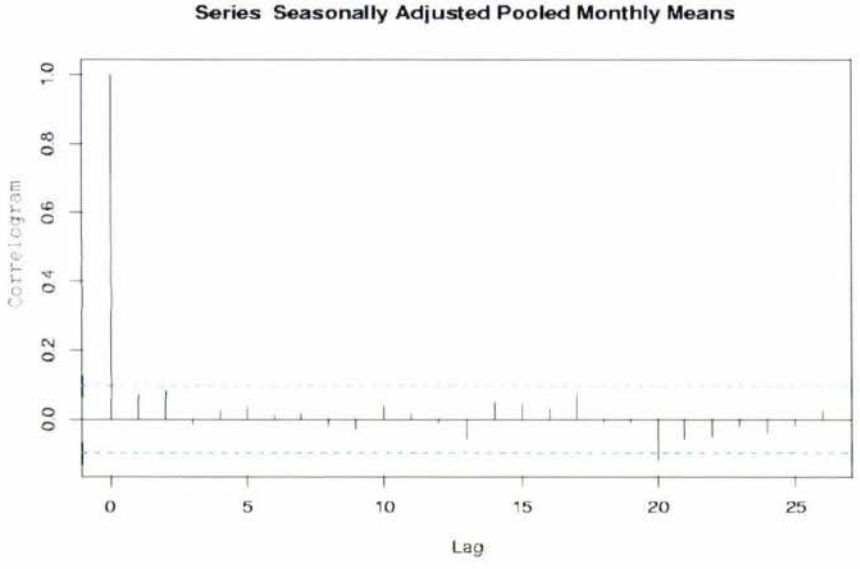
Table 4.1: Percentage of valid data remaining within the historical record

Hourly percentage of valid records by season																							
Month	Site																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.34	0.42	0.41	0.12	0.29	0.35	0.38	0.3	0.26	0.29	0.38	0.38	0.4	0.41	0.38	0.4	0.44	0.18	0.44	0.06	0.44	0.44	0.41
Feb	0.32	0.41	0.41	0.12	0.29	0.35	0.38	0.29	0.24	0.3	0.38	0.4	0.41	0.41	0.41	0.38	0.44	0.17	0.44	0.06	0.44	0.44	0.41
Mar	0.34	0.44	0.41	0.12	0.29	0.35	0.38	0.29	0.24	0.29	0.41	0.41	0.4	0.41	0.41	0.39	0.44	0.14	0.44	0.06	0.44	0.44	0.41
Apr	0.33	0.44	0.41	0.12	0.29	0.35	0.38	0.28	0.24	0.29	0.41	0.41	0.39	0.41	0.39	0.41	0.44	0.12	0.44	0.06	0.44	0.43	0.41
May	0.38	0.44	0.43	0.11	0.29	0.35	0.37	0.28	0.24	0.3	0.4	0.41	0.4	0.44	0.41	0.41	0.44	0.13	0.44	0.06	0.44	0.44	0.38
Jun	0.35	0.42	0.44	0.12	0.29	0.35	0.38	0.32	0.24	0.32	0.41	0.41	0.39	0.42	0.41	0.38	0.44	0.15	0.44	0.06	0.44	0.44	0.4
Jul	0.35	0.43	0.44	0.11	0.29	0.35	0.35	0.32	0.24	0.27	0.35	0.35	0.35	0.38	0.41	0.41	0.44	0.15	0.44	0.06	0.44	0.44	0.41
Aug	0.3	0.43	0.42	0.09	0.29	0.38	0.34	0.32	0.24	0.29	0.38	0.41	0.36	0.4	0.41	0.41	0.44	0.15	0.44	0.06	0.44	0.44	0.41
Sept	0.31	0.43	0.41	0.09	0.31	0.38	0.35	0.29	0.24	0.24	0.38	0.38	0.38	0.44	0.41	0.41	0.44	0.15	0.44	0.06	0.44	0.44	0.41
Oct	0.3	0.38	0.41	0.09	0.3	0.32	0.35	0.24	0.24	0.24	0.36	0.38	0.36	0.41	0.38	0.38	0.41	0.15	0.41	0.06	0.41	0.41	0.38
Nov	0.31	0.38	0.41	0.08	0.29	0.34	0.35	0.26	0.24	0.24	0.38	0.35	0.32	0.39	0.38	0.38	0.41	0.15	0.41	0.03	0.41	0.41	0.39
Dec	0.33	0.37	0.41	0.09	0.28	0.34	0.35	0.26	0.24	0.25	0.38	0.35	0.34	0.4	0.38	0.38	0.41	0.15	0.41	0.03	0.41	0.41	0.41
Daily percentage of valid records by season																							
Month	Site																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.81	0.57	0.71	0.26	0.44	0.76	0.82	0.71	0.76	0.82	0.88	0.94	0.57	0.97	0.65	0.9	0.94	0.62	0.76	0.56	0.97	0.88	0.94
Feb	0.85	0.5	0.71	0.26	0.47	0.76	0.88	0.7	0.8	0.82	0.88	0.93	0.59	0.97	0.67	0.88	0.91	0.73	0.76	0.5	1	0.91	0.85
Mar	0.87	0.62	0.71	0.24	0.47	0.82	0.85	0.71	0.76	0.79	0.91	0.97	0.58	0.97	0.68	0.89	0.94	0.67	0.76	0.53	1	0.91	0.85
Apr	0.86	0.59	0.71	0.26	0.47	0.82	0.84	0.69	0.76	0.79	0.88	0.97	0.57	0.94	0.65	0.9	0.94	0.68	0.76	0.53	1	0.86	0.88
May	0.94	0.55	0.72	0.26	0.47	0.79	0.87	0.69	0.76	0.8	0.87	0.97	0.58	1	0.68	0.91	0.94	0.63	0.74	0.53	1	0.91	0.85
Jun	0.79	0.53	0.73	0.23	0.44	0.82	0.91	0.73	0.76	0.79	0.88	0.97	0.56	0.98	0.68	0.88	0.94	0.65	0.74	0.53	1	0.94	0.84
Jul	0.85	0.58	0.73	0.23	0.53	0.82	0.88	0.73	0.76	0.77	0.82	0.91	0.52	0.94	0.68	0.91	0.94	0.62	0.74	0.56	1	0.94	0.76
Aug	0.82	0.57	0.74	0.21	0.53	0.82	0.83	0.74	0.79	0.79	0.85	0.97	0.54	0.96	0.68	0.88	0.94	0.65	0.74	0.58	1	0.91	0.76
Sept	0.87	0.54	0.74	0.21	0.52	0.85	0.85	0.67	0.76	0.71	0.85	0.93	0.56	1	0.68	0.88	0.94	0.68	0.74	0.56	1	0.91	0.88
Oct	0.77	0.5	0.73	0.21	0.5	0.76	0.88	0.62	0.79	0.74	0.86	0.94	0.53	0.97	0.65	0.85	0.91	0.59	0.71	0.53	0.97	0.85	0.85
Nov	0.75	0.5	0.74	0.2	0.5	0.78	0.88	0.64	0.79	0.74	0.88	0.9	0.5	0.94	0.65	0.85	0.91	0.59	0.71	0.45	0.97	0.88	0.77
Dec	0.8	0.52	0.74	0.21	0.45	0.75	0.88	0.65	0.79	0.75	0.88	0.88	0.51	0.96	0.65	0.85	0.91	0.62	0.7	0.53	0.94	0.85	0.85

4.4.2 *Temporal stationarity*

The assumption of temporal stationarity can be confirmed by plotting the sample autocorrelation of the regional, deseasonalised monthly mean (here the data was deseasonalised by subtracting off the overall mean (by site) for each season pooled by year). Note that the daily data were used to construct the means so that a reasonable record length could be obtained.

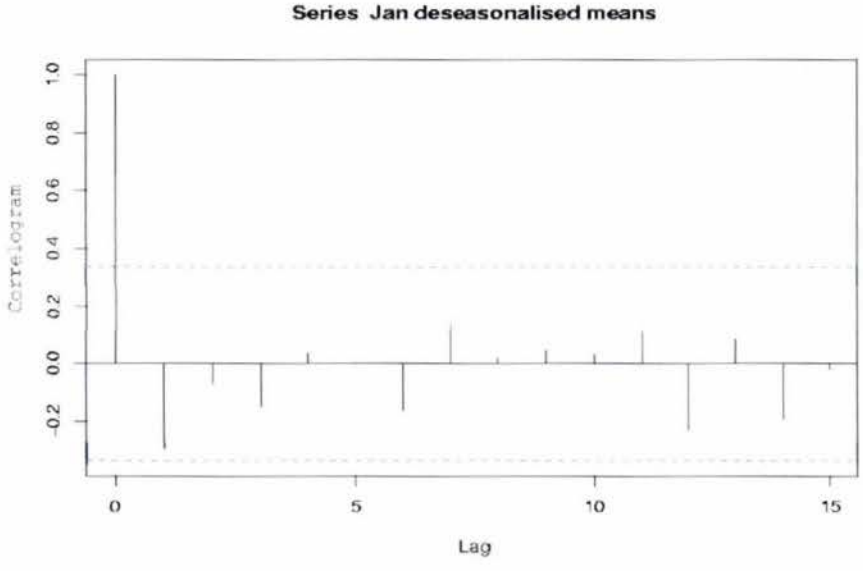
As the assumption of homogeneity has been satisfied, the deseasonalised data can be pooled together without further adjustment and the pooled mean calculated for each season by year. The correlogram (Figure 4.5) shows no significant lags at the seasonal lags $(12, 24)$. Therefore, the seasonal means have been accounted for.



(5% significance level marked)

Figure 4.5: Correlogram: Deseasonalised monthly means

The correlogram, if constructed for a single season (eg: January Figure 4.6) across the 34-year period, also shows no significant lags. Therefore, no discernable climate change could be observed over the period of data collection, and the assumption of temporal stationarity could not be rejected.



(5% significance level marked)

Figure 4.6: Correlogram: January deseasonalised monthly means

5. RESULTS

> Results! Why, man, I have gotten a lot of results. I know several thousand things that won't work.
>
> **-Thomas A. Edison 1847-1931**

5.1 Introduction

This chapter is divided into four divisions. The first section covers the fitting of the spatial-temporal Neyman-Scott Rectangular Pulse model to the historical data. The second section validates that the fitted model is producing the characteristics of the rainfall data correctly. In the third section, a *heuristic*, a general formulation that serves to guide the solution of a problem, for speeding up the infilling algorithms is proposed. See Section 6.5.1 for further applicable heuristics. In the final section, infilling algorithms are developed which maintain the characteristics of the source data.

The standards required for the successful fitting of a spatial-temporal NSRP model are as follows. Firstly, the model must satisfactorily emulate the behaviour of the historical data at the required 1-hour and 24-hour aggregation levels. Secondly, the assumption required by the model of the homogeneity of the region must be supported at least approximately. A model is shown to be fitted which meets both these criteria (see Sections 5.2 and 5.3).

A heuristic for speeding up the infilling algorithms is discussed briefly (Section 5.4). Two further heuristics are proposed in Section 6.5.1, but were not implemented within this project as all the results had been collected prior to these latter heuristics being developed. As a result, the time for implementation at this

point was deemed too high compared to the perceived benefits. All heuristics can be used in combination with each other, and it is expected that in future analyses, particularly using more complex models, this would be done. However, due to the nature of the fitted model and the characteristics of the synthetic data produced, the beneficial component of the third heuristic may be outweighed by the management costs associated with the heuristic.

Given the fitted model obtained (Section 5.2), the final section is devoted to development of an infilling algorithm suitable for replacement of missing data in the historical record. The algorithms used are expected to make use of the constructed model via simulation of a synthetic record length of a suitable size. For this analysis, based on the simulation stability investigation (see Section 5.3.3), a model simulation period of 300 years was deemed an adequately long resource for infilling algorithms to use. Many candidate infilling algorithms are considered, but a candidate algorithm based on iterative sampling is shown to be superior to the other algorithms considered.

5.2 *Model fitting*

> The purpose of models is not to fit the data but to sharpen the questions.
>
> **-Samuel Karlin, quoted in Gaither and Cavazos-Gaither (1996)**

5.2.1 Introduction

The first model considered, hereafter referred to as $Model_A$, is fitted using the 1-hour, 6-hour, and 24-hour aggregation levels. The second model, henceforth known as $Model_B$, is constructed using only the 1-hour and 24-hour aggregation levels. These two alternative fits are compared against the historical record and both are shown to satisfactorily emulate the historical data despite the disparate parameter estimates. Furthermore, the monthly variation in the historical record can be reproduced by either model. Lastly, the assumption of homogeneity for the region, required by the spatial-temporal NSRP model, is verified.

5.2.2 *Parameter estimation*

In order to fit the spatial-temporal NSRP model, the sample pooled statistics must be extracted from the historical data at the required aggregation levels. The 1-hour, 6-hour, and 24-hour pooled statistics are necessary in order to compute $Model_A$; for $Model_B$ only the 1-hour and 24-hour statistics are used. As mentioned (Section 3.1.5), the pooled estimates are smoothed using stepwise harmonic regression so the between season changes are less dramatic. Rather than use all harmonics and risk overfitting to sample variation, only the first three harmonics were used (Table 5.1). This overfitting is particularly noticeable at the 1-hour level and for the 24-hour autocorrelation. The variation in the latter can easily be seen to be sampling variation (Figure 5.3 - autocorrelation plot); and a straight line fit is actually the *best* estimate. If all harmonics are used, however, a harmonic regression model using 5th order harmonics is significant which does not seem plausible for the physical process. In any case, the smoothing process does change the estimates of the statistics considerably (especially January), and, as noted previously, this procedure is not an ideal solution.

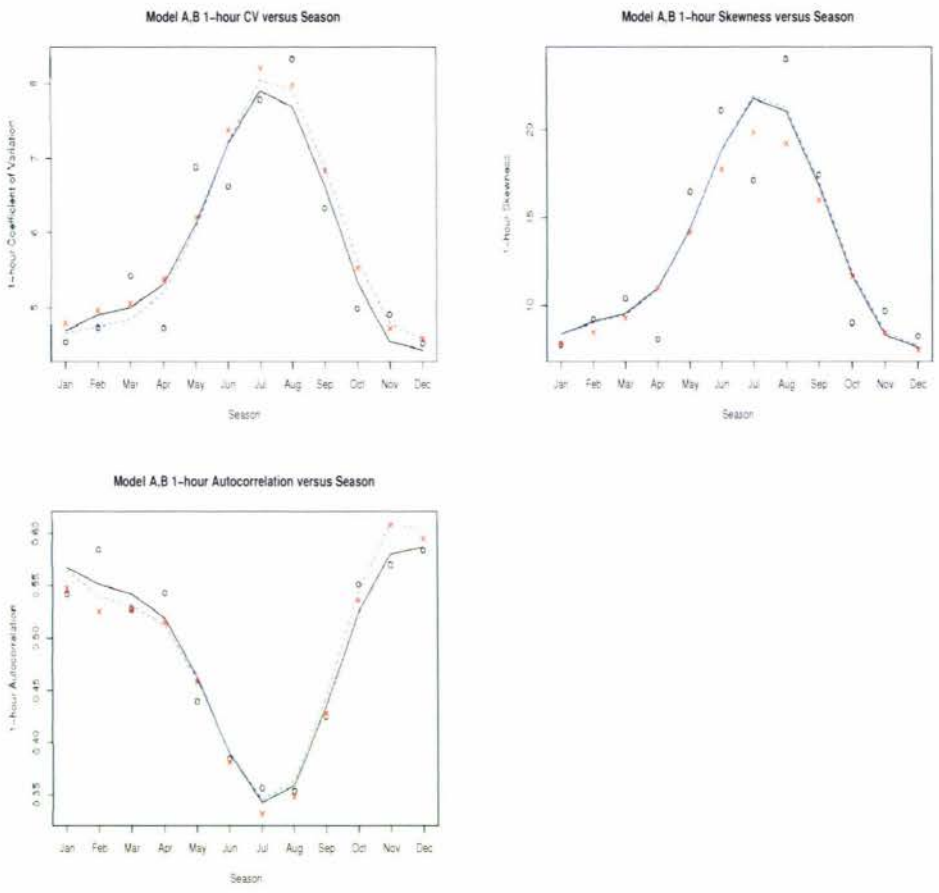
The sample estimates of the 1-hour, 6-hour, and 24-hour coefficient of variation (CV), skewness, and autocorrelation (1st lag), (see Equations 3.13 to 3.15) along with their corresponding smoothed estimates, (Table 5.1) shows that the smoothing may substantially alter the raw estimates. For example, the smoothing has significantly increased the CV and skewness, in January ensuring that this month is more consistent with December and February. Furthermore, the fluctuations of the CV and skewness over the summer months are smoothed to a more reasonable season change (Table 5.1 and Figures 5.1 to 5.3).

The fitting of the 24-hour autocorrelation was retained in the optimisation procedure as a complete fit to each aggregation level was desired. Furthermore, the model fitting function (Equation 3.18) gives equal weighting to all estimates - regardless of their expected accuracy. In this, the procedure is consistent with Cowpertwait et al. (2002) with the exception that exact fitting for the 1-hour aggregation level was not required.

Table 5.1: Historical pooled statistics: raw and smoothed

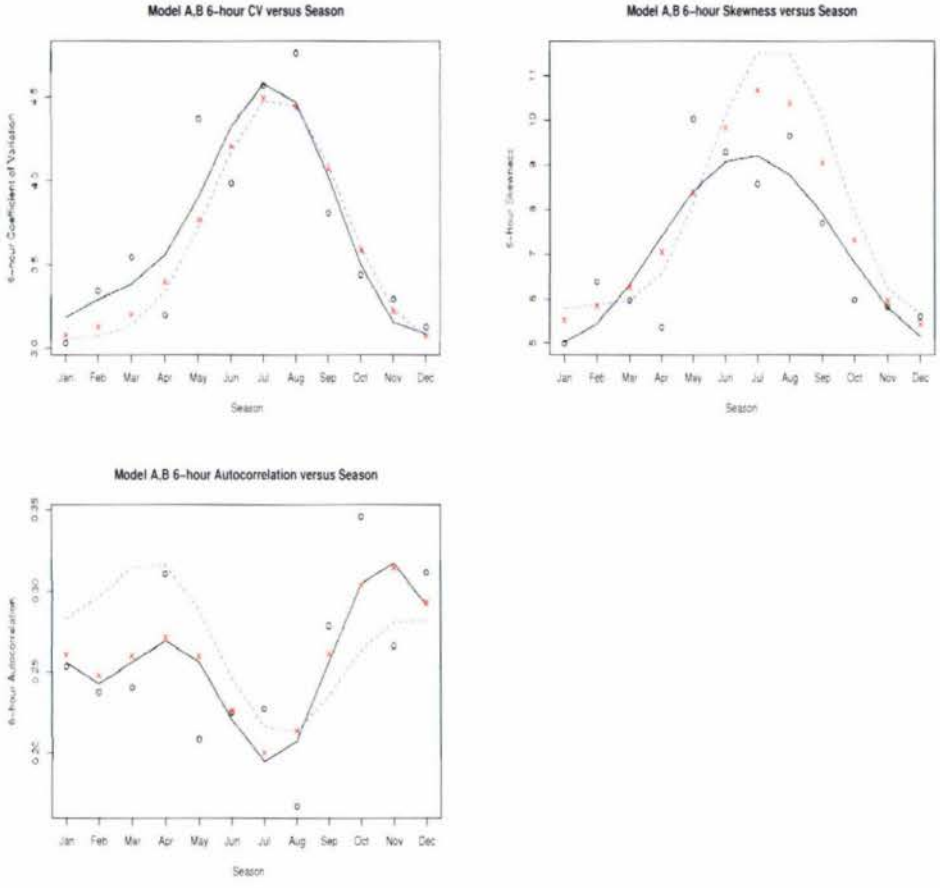
	Aggregation	Raw			Smoothed		
Month	Level (hours)	CV	skew	acor	CV	skew	acor
Jan	1	4.531	7.749	0.542	4.690	8.350	0.568
Feb	1	4.728	9.193	0.585	4.899	9.069	0.551
Mar	1	5.421	10.374	0.528	4.998	9.533	0.542
Apr	1	4.724	8.054	0.543	5.315	10.980	0.519
May	1	6.882	16.465	0.439	6.120	14.386	0.463
Jun	1	6.621	21.103	0.385	7.196	18.837	0.390
Jul	1	7.784	17.126	0.357	7.902	21.779	0.343
Aug	1	8.325	24.017	0.354	7.693	21.061	0.359
Sept	1	6.327	17.450	0.424	6.626	16.874	0.435
Oct	1	4.987	8.996	0.551	5.341	11.704	0.524
Nov	1	4.905	9.679	0.570	4.536	8.298	0.580
Dec	1	4.511	8.234	0.584	4.428	7.569	0.587
Jan	6	3.032	4.986	0.253	3.187	5.009	0.256
Feb	6	3.345	6.377	0.237	3.295	5.436	0.243
Mar	6	3.547	5.962	0.240	3.386	6.310	0.256
Apr	6	3.202	5.355	0.310	3.563	7.397	0.269
May	6	4.369	10.021	0.209	3.906	8.405	0.256
Jun	6	3.986	9.284	0.225	4.324	9.064	0.221
Jul	6	4.568	8.567	0.227	4.576	9.198	0.194
Aug	6	4.758	9.636	0.167	4.468	8.771	0.207
Sept	6	3.810	7.682	0.279	4.029	7.897	0.256
Oct	6	3.443	5.971	0.346	3.504	6.810	0.304
Nov	6	3.298	5.808	0.266	3.160	5.802	0.317
Dec	6	3.132	5.592	0.312	3.091	5.143	0.291
Jan	24	1.784	2.677	0.218	1.917	3.131	0.164
Feb	24	2.047	3.426	0.155	1.921	3.079	0.164
Mar	24	1.949	3.131	0.165	1.983	3.100	0.164
Apr	24	2.068	3.072	0.187	2.124	3.376	0.164
May	24	2.454	4.314	0.124	2.347	4.022	0.164
Jun	24	2.522	4.824	0.181	2.592	4.863	0.164
Jul	24	2.699	5.352	0.137	2.754	5.488	0.164
Aug	24	2.859	5.383	0.120	2.749	5.540	0.164
Sept	24	2.514	5.480	0.178	2.580	5.005	0.164
Oct	24	2.373	3.968	0.178	2.331	4.215	0.164
Nov	24	2.064	3.430	0.113	2.108	3.570	0.164
Dec	24	2.044	3.572	0.211	1.971	3.242	0.164

where CV is the coefficient of variation,
skew is the skewness coefficient, and
acor is the lag 1 autocorrelation



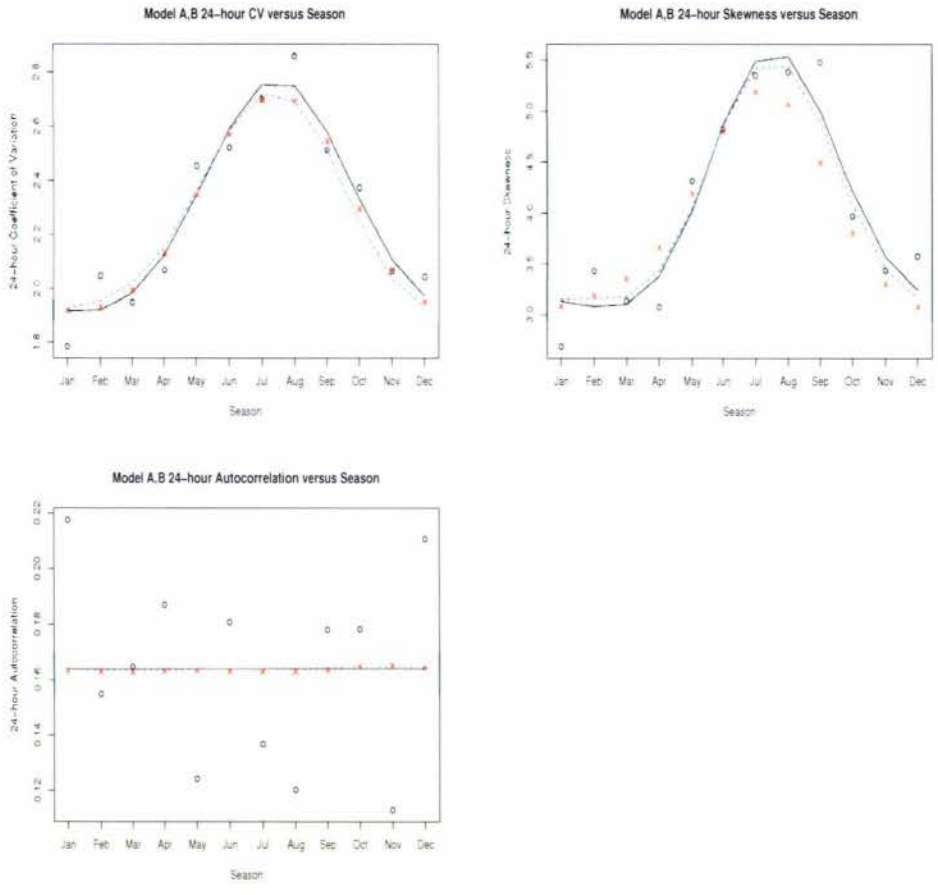
raw = circles; smooth = solid line; $Model_B$ = dashed line; $Model_A$ = cross

Figure 5.1: Model fit: 1-hour aggregation level



raw = circles; smooth = solid line; $Model_B$ = dashed line; $Model_A$ = cross

Figure 5.2: Model fit: 6-hour aggregation level



raw = circles; smooth = solid line; $Model_B$ = dashed line; $Model_A$ = cross

Figure 5.3: Model fit: 24-hour aggregation level

Estimates $Model_A$

The parameters for the spatial-temporal NSRP model for $Model_A$ are given in Tables 5.2 and 5.3. Plots of the raw, smoothed, and fitted statistics, for each of the aggregation levels, 1-hour, 6-hour, and 24-hour, are illustrated as crosses in Figures 5.1 to 5.3. The plots clearly illustrate that, in general, the model is fitting against the sample smoothed statistics quite well. One point point of interest, however, is that the behaviour of the 6-hour skewness is different from the 1-hour and 24-hour skewness. As a least squares fitting function is used, this model attempts to adjust for this behaviour and, as a direct result, under-estimates the 1-hour and 24-hour skewness and coefficient of variation. This under-estimation is particularly evident over the summer months where the discrepancy between the 6-hour aggregation behaviour and the other fitted aggregation levels is the greatest.

Estimates $Model_B$

As the 6-hour aggregation level comparatively relied on less data (compared to the other two aggregation levels), a second model was constructed, $Model_B$, which was fitted solely to the 1-hour and 24-hour aggregation levels. The parameter estimates for this model are presented in Tables 5.4 and 5.5. The estimates for $Model_B$ are very different from those obtained previously for $Model_A$ (Tables 5.2 and 5.3). In particular, $Model_A$'s cells last longer over Spring, but are otherwise are of shorter duration than $Model_B$ (from η), and $Model_A$'s cells are generally more intense (α and θ). The number of rain cells per site, μ_c, is generally higher in $Model_B$ which will have a direct effect on the time taken for simulation. The magnitude of the change between $Model_A$ and $Model_B$ implies that the spatial-temporal NSRP model is not overly robust - at least in terms of the parameter estimates.

Table 5.2: $Model_A$ Monthly parameter estimates

Month	$\hat{\lambda}, h^{-1}$	$\hat{\mu}_C$	$\hat{\beta}, h^{-1}$	$\hat{\eta}, h^{-1}$	$\hat{\alpha}$	$\hat{\phi}, km^{-1}$
Jan	0.0137	5.5211	0.0860	1.2035	1.1156	0.0412
Feb	0.0137	5.7554	0.0827	1.2770	1.0171	0.0401
Mar	0.0125	7.5910	0.0877	1.2983	0.8503	0.0547
Apr	0.0105	11.1894	0.0931	1.3920	0.6943	0.0592
May	0.0087	13.8699	0.0919	1.6724	0.6144	0.0852
Jun	0.0075	12.2686	0.0829	2.0780	0.6149	0.1255
Jul	0.0070	10.2706	0.0745	2.3740	0.6405	0.0956
Aug	0.0069	11.1542	0.0794	2.2896	0.6334	0.1127
Sept	0.0073	15.1041	0.0939	1.8768	0.6032	0.1227
Oct	0.0085	15.9480	0.1027	1.3871	0.6333	0.0630
Nov	0.0108	10.0169	0.1011	1.0666	0.8163	0.0497
Dec	0.0127	6.5589	0.0944	1.0702	1.0606	0.0417

Table 5.3: $Model_A$ Scale parameter estimate $\hat{\theta}_{ik}(mm)$ for each Site-Month

Month	ith Site											
k	1	2	3	4	5	6	7	8	9	10	11	12
Jan	1.11	1.26	1.46	0.98	1.45	1.22	1.09	1.33	1.32	1.22	1.35	1.35
Feb	0.76	1.05	1.04	0.76	0.95	0.90	0.85	1.03	0.88	0.96	1.05	1.00
Mar	0.70	0.70	0.84	0.80	0.81	0.71	0.67	0.83	0.72	0.81	0.81	0.84
Apr	0.52	0.64	0.67	0.58	0.59	0.57	0.50	0.63	0.51	0.62	0.60	0.61
May	0.60	0.58	0.69	0.72	0.67	0.68	0.56	0.67	0.65	0.69	0.69	0.69
Jun	0.97	1.11	1.23	1.08	1.15	1.07	0.95	1.10	1.13	0.98	1.21	1.16
Jul	1.26	1.38	1.61	1.30	1.38	1.17	1.10	1.26	1.35	1.33	1.36	1.36
Aug	1.36	1.02	1.39	0.95	1.28	1.22	1.26	1.44	1.38	1.37	1.40	1.46
Sept	0.84	1.00	0.86	0.84	0.84	0.90	0.86	0.92	0.79	0.90	0.95	0.91
Oct	0.52	0.70	0.73	0.55	0.80	0.55	0.53	0.54	0.58	0.57	0.64	0.65
Nov	0.63	0.70	0.72	0.60	0.71	0.73	0.63	0.77	0.71	0.74	0.78	0.75
Dec	0.79	1.00	1.20	1.01	1.07	1.04	0.85	1.09	0.93	1.11	1.10	1.00
Month	13	14	15	16	17	18	19	20	21	22	23	
Jan	1.25	1.25	1.44	1.21	1.35	1.16	1.36	1.26	1.16	1.04	1.22	
Feb	1.08	0.92	0.95	0.94	0.98	0.81	1.13	0.99	0.88	0.82	0.94	
Mar	0.79	0.77	0.77	0.76	0.77	0.80	0.74	0.87	0.69	0.65	0.66	
Apr	0.61	0.58	0.59	0.59	0.59	0.55	0.63	0.52	0.56	0.51	0.56	
May	0.61	0.59	0.65	0.63	0.64	0.67	0.60	0.68	0.60	0.56	0.64	
Jun	0.94	1.07	1.12	1.16	0.98	1.13	1.16	0.95	1.05	0.91	0.99	
Jul	1.23	1.27	1.49	1.30	1.21	1.49	1.26	1.34	1.22	1.19	1.28	
Aug	1.32	1.36	1.42	1.40	1.26	1.44	1.49	1.28	1.34	1.32	1.24	
Sept	0.86	0.86	0.91	0.94	0.82	0.78	0.94	0.84	0.83	0.83	0.84	
Oct	0.63	0.65	0.81	0.61	0.66	0.61	0.60	0.59	0.59	0.59	0.57	
Nov	0.66	0.74	0.77	0.76	0.69	0.66	0.79	0.77	0.69	0.65	0.73	
Dec	1.12	1.04	1.13	1.06	1.10	1.00	1.12	0.95	1.00	0.84	0.98	

Table 5.4: $Model_B$ Monthly parameter estimates

Month	$\hat{\lambda}, h^{-1}$	$\hat{\mu}_C$	$\hat{\beta}, h^{-1}$	$\hat{\eta}, h^{-1}$	$\hat{\alpha}$	$\hat{\phi}, km^{-1}$
Jan	0.0130	8.2073	0.0935	1.1823	0.8585	0.0436
Feb	0.0120	11.0730	0.1009	1.3453	0.7666	0.0455
Mar	0.0108	14.4238	0.1079	1.4579	0.7103	0.0650
Apr	0.0094	18.1970	0.1091	1.5748	0.6329	0.0683
May	0.0081	19.6841	0.1017	1.7735	0.5662	0.0937
Jun	0.0071	17.6816	0.0898	2.0884	0.5427	0.1339
Jul	0.0067	15.3983	0.0800	2.3246	0.5424	0.0995
Aug	0.0069	14.6760	0.0778	2.1543	0.5455	0.1121
Sept	0.0079	14.2285	0.0823	1.6751	0.5641	0.1102
Oct	0.0098	11.1675	0.0866	1.2246	0.6497	0.0549
Nov	0.0122	7.1262	0.0867	0.9781	0.8591	0.0441
Dec	0.0134	6.3510	0.0882	1.0048	0.9741	0.0399

Table 5.5: $Model_B$ Scale parameter estimate $\hat{\theta}_{ik}(mm)$ for each Site-Month

Month	ith Site											
k	1	2	3	4	5	6	7	8	9	10	11	12
Jan	0.68	0.78	0.90	0.60	0.90	0.75	0.68	0.82	0.82	0.75	0.84	0.83
Feb	0.40	0.56	0.55	0.40	0.50	0.47	0.45	0.54	0.47	0.51	0.56	0.53
Mar	0.42	0.42	0.50	0.48	0.48	0.43	0.40	0.50	0.43	0.49	0.49	0.50
Apr	0.37	0.45	0.47	0.41	0.42	0.40	0.36	0.44	0.36	0.44	0.43	0.44
May	0.43	0.42	0.49	0.52	0.48	0.49	0.40	0.48	0.47	0.50	0.50	0.50
Jun	0.60	0.68	0.76	0.67	0.71	0.66	0.59	0.68	0.70	0.61	0.75	0.71
Jul	0.68	0.75	0.88	0.71	0.75	0.64	0.60	0.69	0.74	0.72	0.74	0.74
Aug	0.79	0.60	0.81	0.55	0.74	0.71	0.73	0.84	0.80	0.80	0.82	0.85
Sept	0.66	0.79	0.69	0.67	0.67	0.71	0.69	0.73	0.63	0.72	0.76	0.73
Oct	0.59	0.79	0.82	0.62	0.91	0.62	0.60	0.61	0.65	0.65	0.72	0.74
Nov	0.74	0.83	0.84	0.70	0.84	0.86	0.74	0.91	0.84	0.87	0.92	0.88
Dec	0.70	0.88	1.06	0.89	0.95	0.92	0.76	0.96	0.82	0.98	0.97	0.89
Month	13	14	15	16	17	18	19	20	21	22	23	
Jan	0.77	0.77	0.89	0.75	0.83	0.72	0.84	0.78	0.72	0.64	0.76	
Feb	0.57	0.49	0.50	0.50	0.52	0.43	0.60	0.52	0.47	0.43	0.50	
Mar	0.47	0.46	0.46	0.45	0.46	0.48	0.44	0.52	0.41	0.39	0.39	
Apr	0.43	0.41	0.42	0.41	0.42	0.39	0.45	0.37	0.40	0.36	0.40	
May	0.44	0.43	0.47	0.45	0.46	0.48	0.43	0.49	0.43	0.40	0.46	
Jun	0.58	0.66	0.69	0.72	0.61	0.70	0.71	0.59	0.65	0.56	0.61	
Jul	0.67	0.69	0.81	0.71	0.66	0.81	0.69	0.73	0.66	0.65	0.70	
Aug	0.77	0.79	0.83	0.82	0.73	0.84	0.87	0.74	0.78	0.77	0.72	
Sept	0.69	0.69	0.73	0.75	0.66	0.62	0.75	0.67	0.66	0.66	0.67	
Oct	0.71	0.74	0.92	0.69	0.75	0.69	0.68	0.66	0.67	0.67	0.64	
Nov	0.78	0.87	0.91	0.90	0.80	0.77	0.92	0.91	0.81	0.77	0.85	
Dec	0.99	0.92	1.00	0.94	0.98	0.89	1.00	0.84	0.88	0.74	0.86	

The seasonal variation in the parameters for both $Model_A$ and $Model_B$ is consistent with the knowledge of the underlying physical process being modelled. For both models, the storm arrival rate, λ, varies in accordance with what is expected - the rate is higher over the winter months and lower over summer (Tables 5.2 and 5.4). The average number of cells per storm is generally higher in $Model_B$ than in $Model_A$, but, again, reflects the variation consistent with the rainfall patterns. That is, the winter months are characterised by a low cell count per storm, but the cells are of longer duration ($1/\eta$), are less intense (α), and cover a wide area ($1/\phi$). The summer months, on the other hand, have a high cell count per storm, but these storms are localised, are of short duration, and cover only a small area. The main difference between the parameters of $Model_A$ and $Model_B$ is summarised by the observation that $Model_B$ models the rainfall using more rain cells of a smaller size than $Model_A$.

The plots for the raw, smoothed, and fitted statistics given for $Model_A$ (Figures 5.1 to 5.3), incorporate $Model_B$ as a dashed line for easy comparison. The fits for the 6-hour statistics are included to determine the effect of removing these estimates from the model fitting procedure even though these statistics were not fitted directly. As expected, the model now fits the smoothed statistics very closely at the 1-hour and 24-hour levels. The 6-hour aggregation level, which was not fitted directly, is within sampling error for the CV, overestimates the autocorrelation over spring, but does not match the historical 6-hour skewness - particularly over summer. For the purpose of this analysis, however, the models were evaluated based on their respective functional performance (see Section 5.3) and the model which produced the better results chosen.

Cross-correlation fit

The cross-correlation (Equation 3.16) versus distance plot was also computed for both models. However, since the results were extremely similar for both $Model_A$ and $Model_B$ for all seasons, only the results for $Model_B$ have been included. As is clear from the plots for January and July (Figure 5.4), the 1-hour cross-correlation is generally fitting reasonably well regardless of distance. There is a tendency to overestimate the cross-correlation at the greater distances at the 1-hour aggregation level.

However, as is clear from both 24-hour cross-correlation versus distance plots, the model is not fitting as well at the greater distances - especially January (Figure 5.4). Rather, the cross-correlation is tapering off too rapidly at the the 24-hour aggregation level. The problem with the under fitting of the 24-hour cross-correlation is a general problem that occurs regardless of season - see Figures B.1 to B.6.

It should be noted, though, in spite of this problem noted with the fitting, the model simulation results also overlaid on these same plots (Figures B.1 to B.6) shows that the model is usually reproducing the same characteristic 'versus distance' behaviour. The main exceptions to this are October and November.

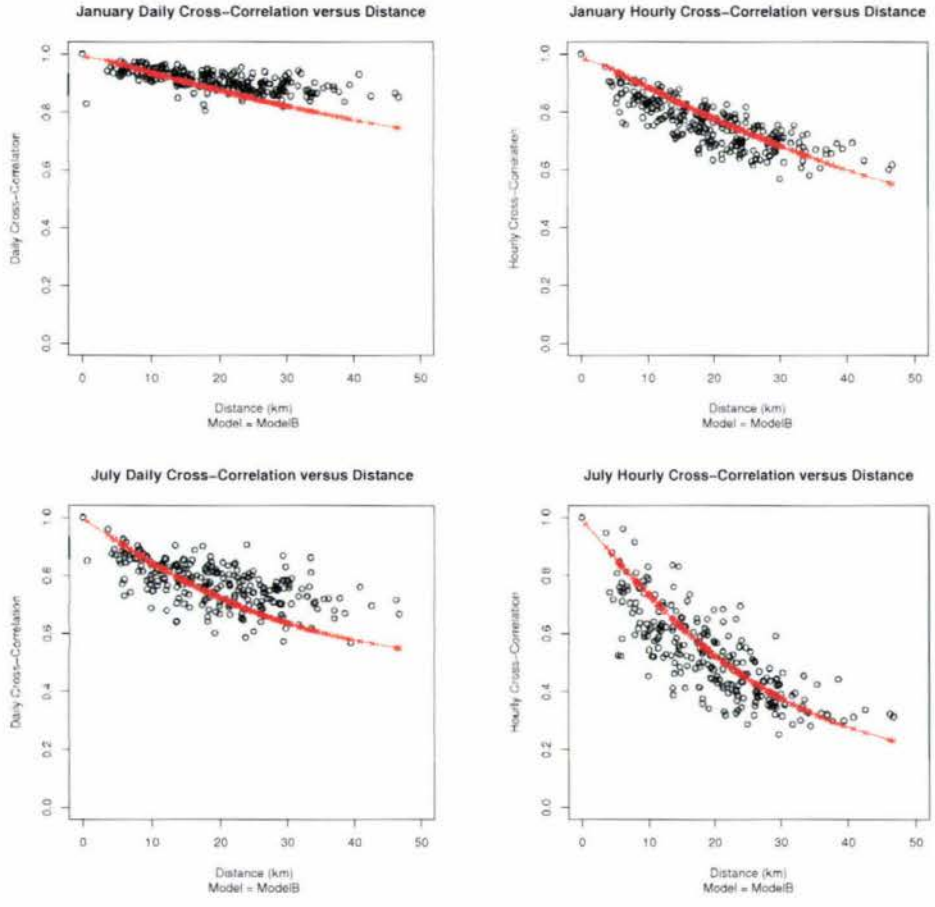


Figure 5.4: $Model_B$ cross-correlation versus distance - January and July
The curved line is the fitted value under the model.

5.3 Model validation

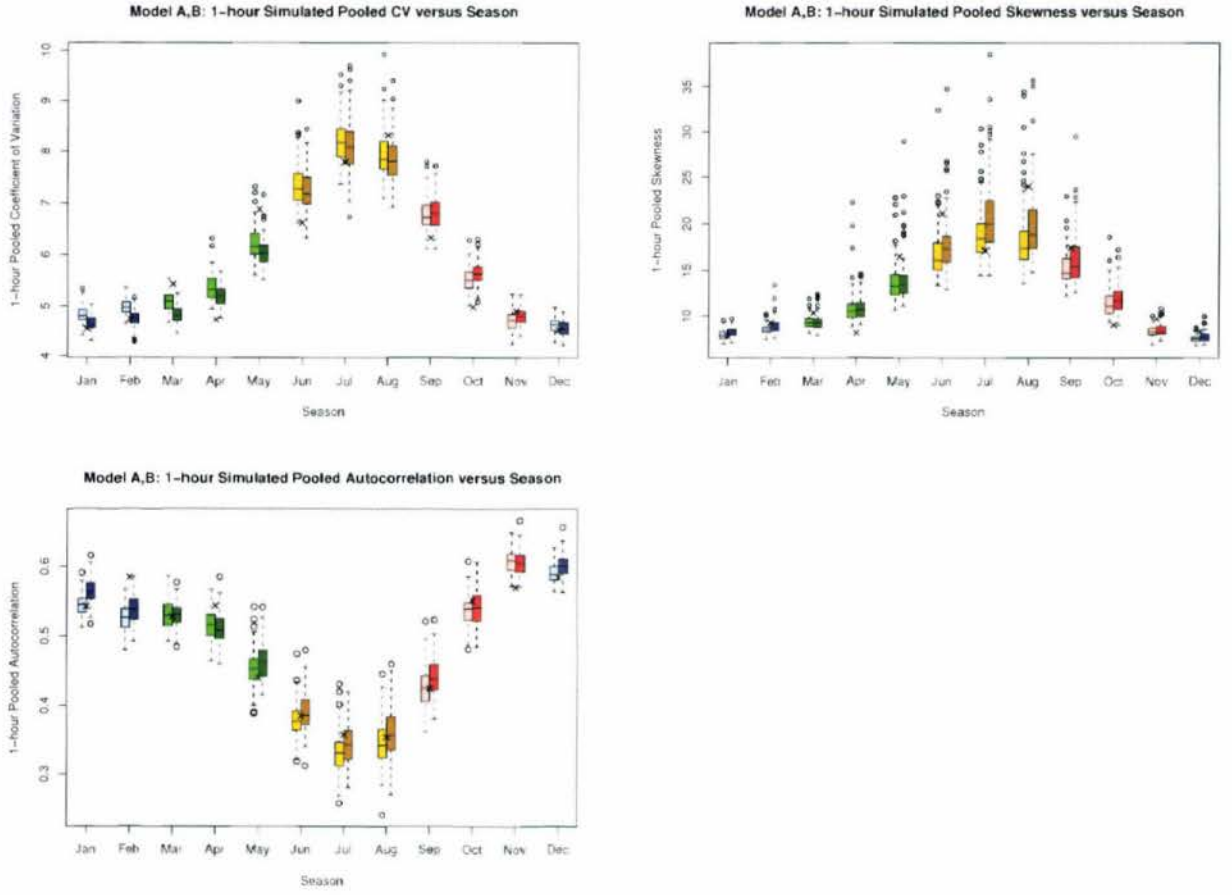
In order to validate the fitted model, a simulation of 300 years was computed for the two models, A and B, with parameters as presented previously. The analysis of this 300 year sample is broken into two portions. The first part examines the overall pooled statistics for both models and compares the distribution of the pooled simulation data with the pooled historical data for each season. The pooling is done by dividing each site's record by the respective site mean for that month. The second section examines the variation of the monthly statistics in order to determine whether the assumption of homogeneity is satisfied and whether the model is producing the correct level of monthly variation.

A valid fitted model may be constructed, but it is necessary to show that a 300-year record is of sufficient length to be used as a basis for infilling. Therefore, the stability of a 300-year simulation by obtaining the pooled statistics from 30 independent samples. It is shown (Section 5.3.3) that there is still considerable variation within a 300-year sample. However, as a 300-year simulation uses $460Mb$ just for the 1-hour record, which is the aggregation level at which any model is simulated, the record length was left unchanged. Due to the problems with the nature of the spatial-temporal NSRP model (see Section 5.4.1), a longer simulation may be of little effective benefit anyway.

5.3.1 Fitted statistics

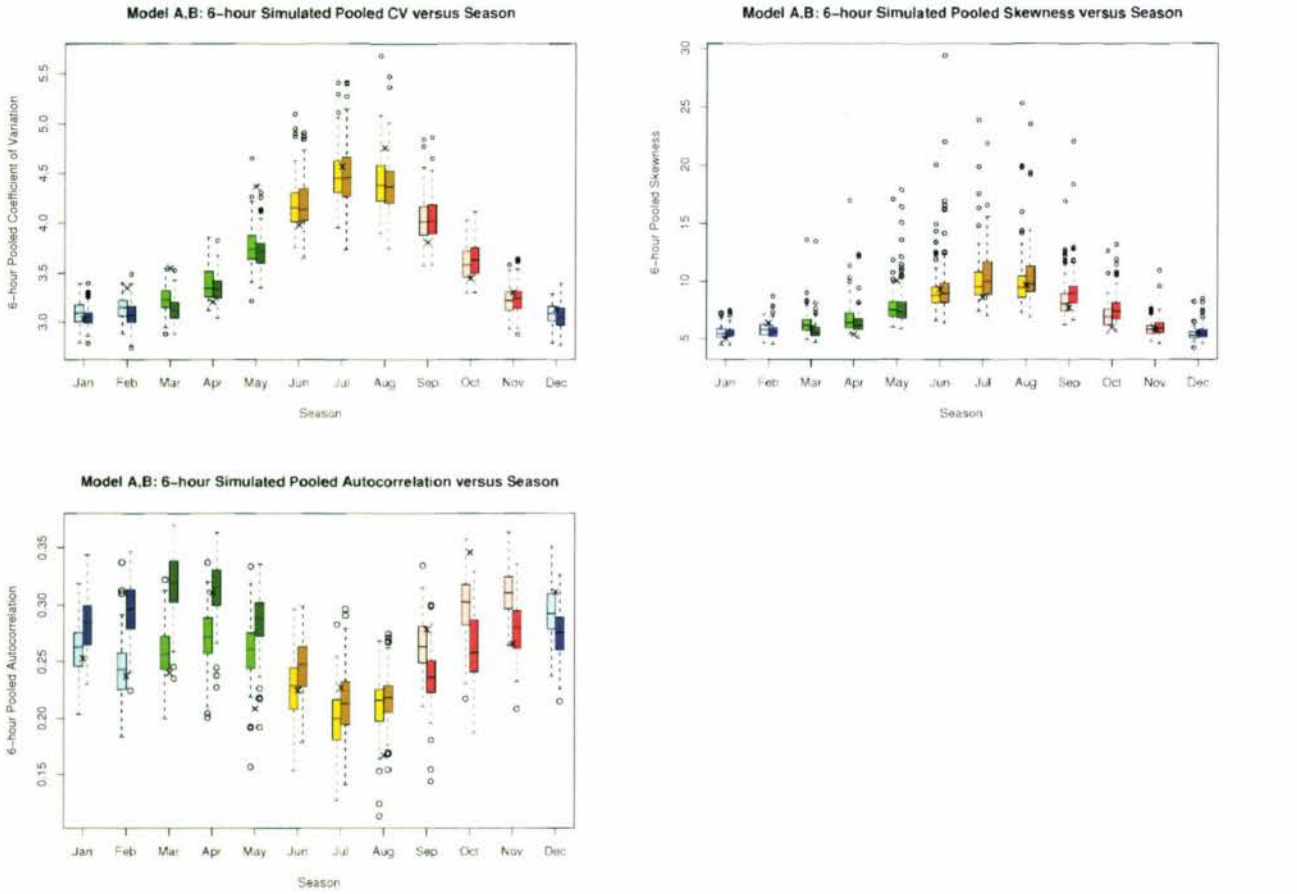
The plots of the pooled statistics (Figures 5.5 to 5.7) are based on a synthetic record generated using the models, A and B, for a 35-year period. These plots were constructed to determine the probability of generating the observed pooled statistics as seen in the historical data which also spanned an approximate 35 year period. Therefore, each model was simulated for 35 years 100 times and the resulting pooled statistics recorded.

The pooled statistics of *coefficient of variation*, *skewness*, and lag 1 *autocorrelation*, were plotted for each aggregation level, but instead of a collection of points, a boxplot for each statistic was given based on the 100 samples (Figures 5.5 to 5.7). The exact historical estimates are also marked on the plots.



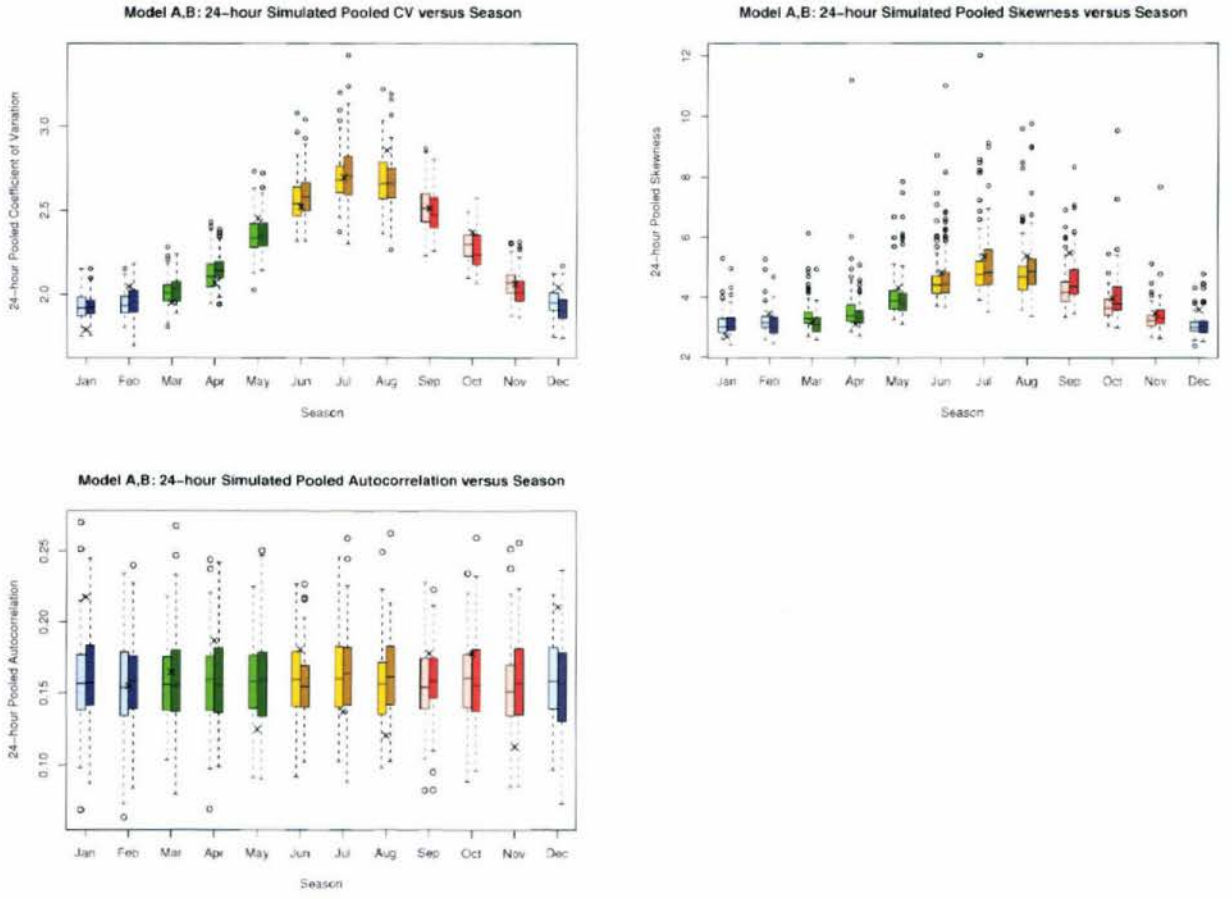
cross=historical estimate; light (left) shaded boxes=$Model_A$, dark (right) shaded boxes=$Model_B$

Figure 5.5: 35 year simulation: 1-hour aggregation level



cross=historical estimate; light (left) shaded boxes=$Model_A$, dark (right) shaded boxes=$Model_B$

Figure 5.6: 35 year simulation: 6-hour aggregation level



cross=historical estimate: light (left) shaded boxes=$Model_A$, dark (right) shaded boxes=$Model_B$

Figure 5.7: 35 year simulation: 24-hour aggregation level

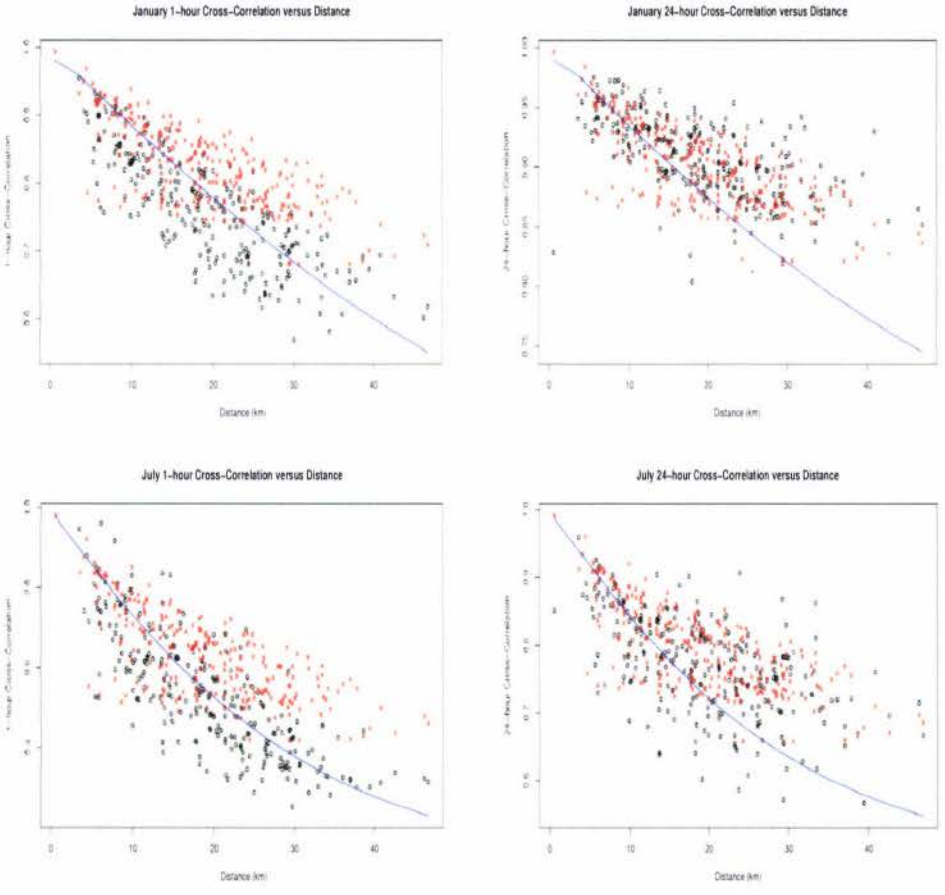
Note that, as consistent with the fitting procedure results (Figures 5.1 to 5.3), the CV and skewness are generally slightly underestimated. As with $Model_A$ the statistics tend to be underestimated for $Model_B$, however, it is evident that $Model_B$ tends to be closer to the true values at the 1-hour and 24-hour aggregation levels (Figures 5.5 to 5.7). $Model_A$, unsurprisingly, fits more closely to the 6-hour historical estimates (Figure 5.6). From the simulation results, $Model_B$ is marginally superior to $Model_A$ as the pooled statistics from the simulation are generally closer to the historical values seen in the Thames Valley dataset.

In general, the historical estimates are within the range of the distribution of the sampled simulations for all the statistics. This is not entirely unexpected, as the same underlying process is attempted to be modelled by both spatial-temporal NSRP parameterisations. However, some months, January for example, are modelled differently at every aggregation level. The different estimates for January, are a direct result of the smoothing algorithm correcting these lower values. This could be seen as either a problem with the smoothing algorithm or just that the sample of historical data are unusual. Given the estimates for December and February and that underestimation of snowfall is a known problem for rain gauges (see Maidment, 1993), the latter seems more likely.

Spatial analysis

Based on the spatial component fitting (Section 5.2.2), it was expected that the cross-correlation analysis would show a similar result. The results for January and July are shown in Figure 5.8. That is, the discrepancy at the 24-hour aggregation level between the fitted and historical cross-correlations would increase as the distance increased. Surprisingly, the synthetic record analysed (one 300-year sample) only tended to underestimate the 24-hour cross-correlation for September through November at the 24-hour level. All other months and aggregation levels either fitted or slightly overestimated the cross-correlation (Figures B.1 to B.6).

The primary result of interest, especially for the months September-November, is the difference between the cross-correlation values at the 1-hour and 24-hour aggregation levels. For these months, compared to the historical record, the 1-hour record cross-correlation is significantly overestimated while the 24-hour record is significantly underestimated.



The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure 5.8: $Model_B$ cross-correlation 1-hour and 24-hour for January and July

Regional distribution analysis

The models, A and B, were simulated for 300 years and the distribution of the values compared to the historical values at the respective 1-hour and 24-hour aggregation levels. The results (synthetic and historical) were split into months and standardised (by dividing each site record by its respective site mean) so that the whole region could be pooled together for comparison purposes. Although it may be beneficial to compare each site independently, the limited availability of the historical record at some of the sites would more than negate the benefit of doing so. Therefore, the results were pooled so that the effectiveness of the models producing the correct distribution as a whole could be examined.

The quantile-quantile plots (Figures 5.9 and 5.10) have been produced for each model at the 1-hour and 24-hour aggregation level for January and July respectively. Taking into account the differing scales for the plots, neither model fits January well - though $Model_A$ is closer to the historical quantiles, but $Model_B$ fits better for July (particularly at the 24-hour level). For the other months (Figures B.7 to B.16), generally $Model_A$ is closer to the tail of the distribution than $Model_B$. $Model_B$ is still preferred as infilling algorithms can avoid fitting the extreme points if necessary. However, if the simulated data (eg: from $Model_A$) does not contain extreme points then obviously no infilling algorithm that uses this data solely can infill with extreme values. The difference between the two models is really only observable in the tail of the distribution.

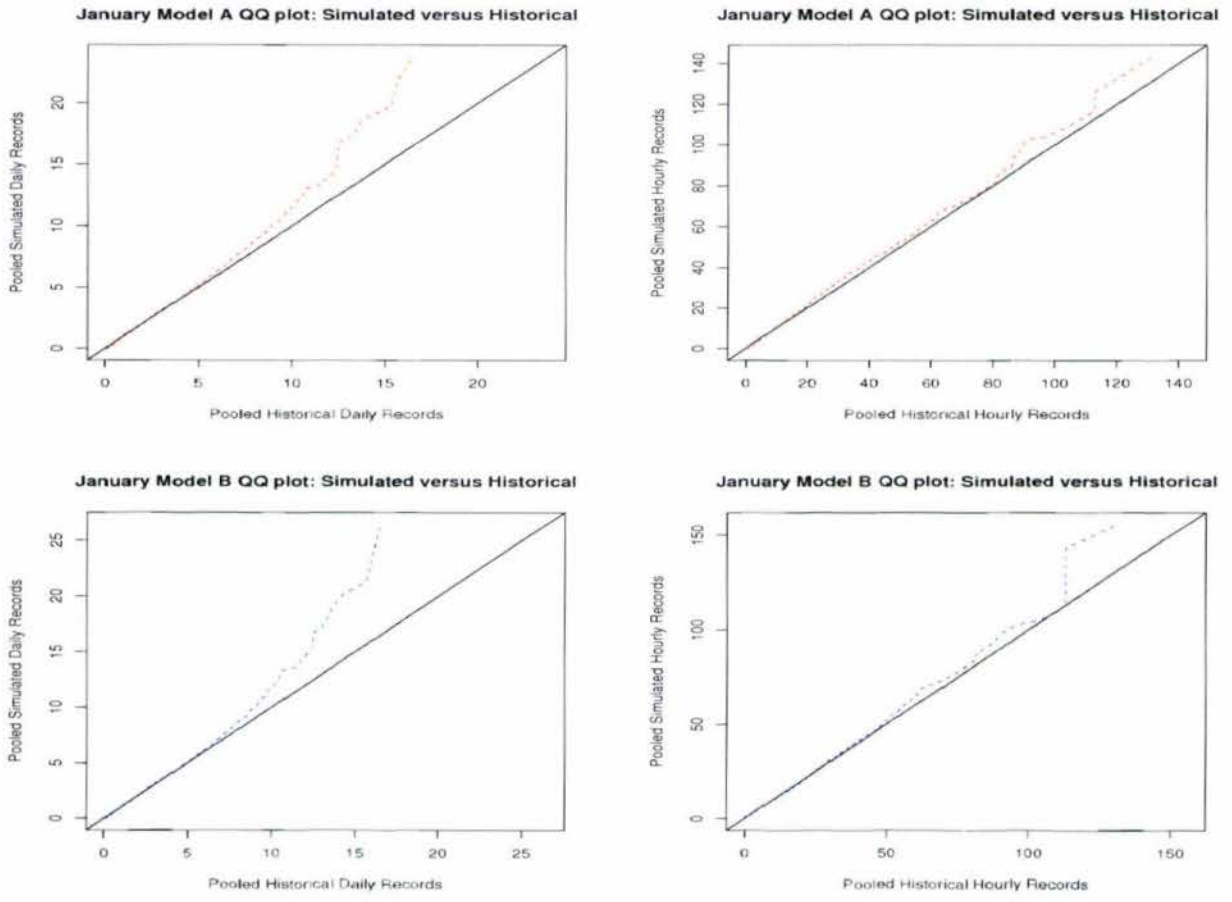


Figure 5.9: Quantile-Quantile plots: January $Model_A, Model_B$

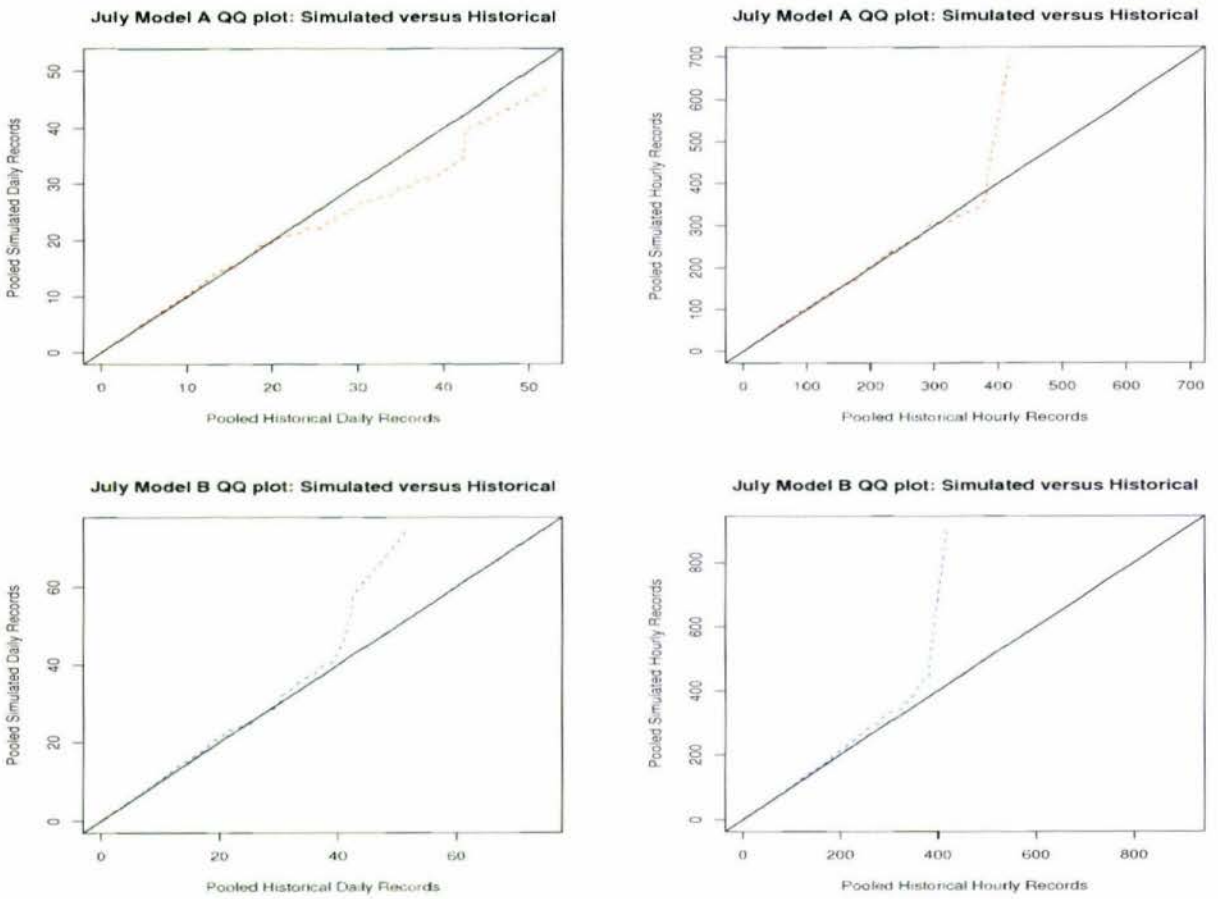


Figure 5.10: Quantile-Quantile plots: July $Model_A, Model_B$

Regional proportion of dry sites

As the proportion of dry sites was not used in the fitting procedure, this was deemed a useful criteria for determining whether the model is fitting the data well. The results for the 1-hour and 24-hour proportions by season, aggregation level, and model are presented in Table 5.6.

Table 5.6: 1-hour and 24-hour: regional proportion dry by season

	24-hour aggregation			1-hour aggregation		
Month	Historical	$Model_A$	$Model_B$	Historical	$Model_A$	$Model_B$
Jan	0.447	0.555	0.544	0.898	0.918	0.906
Feb	0.532	0.558	0.573	0.903	0.923	0.911
Mar	0.518	0.571	0.599	0.926	0.920	0.911
Apr	0.561	0.610	0.630	0.908	0.922	0.915
May	0.594	0.639	0.678	0.939	0.926	0.928
Jun	0.610	0.687	0.682	0.928	0.944	0.935
Jul	0.656	0.692	0.701	0.947	0.951	0.948
Aug	0.654	0.692	0.694	0.948	0.948	0.946
Sept	0.590	0.685	0.651	0.915	0.934	0.930
Oct	0.544	0.654	0.618	0.905	0.917	0.919
Nov	0.473	0.610	0.564	0.898	0.906	0.908
Dec	0.472	0.574	0.548	0.890	0.912	0.907

From table 5.6, both $Model_A$ and $Model_B$ are not producing the same proportion dry as the historical records over the region for the 24-hour aggregation level. At the 1-hour level, the $1-2\%$ difference is statistically significant, given the sample size it encompasses. At either aggregation level, however, $Model_B$ usually gives a better fit.

The discrepancy between the observed proportion dry and the historical proportion dry is an issue commonly found with the NSRP models (see Onof et al. (2000), Cowpertwait et al. (1996)) at least in the temporal domain. The mismatch of the proportion dry at the 24-hour level may cause issues when the model is used for infilling. However, as the infilling algorithms use any available valid data, this is not expected to cause a major problem provided enough data are available. However, as with Cowpertwait et al. (1996) it does not follow that the model is not useful, but rather the model limitations must be taken into account when the model is applied.

Conclusion

As it is clear that $Model_A$ and $Model_B$ are producing similar output (with the exception of the 6-hour aggregation level), a decision was made to retain the model fitted using the simpler method, $Model_B$, rather than $Model_A$. Therefore, all subsequent model analysis and results within this thesis relate to $Model_B$ solely.

As noted, there are some potential problems with the lack of fitting at the 24-hour cross-correlation level as well as with the proportion dry being overestimated. Provided there is enough valid data at a given time point, however, it is likely that the fitted NSRP model can still be used for infilling (see Section 5.5).

5.3.2 Monthly statistics

In this section, the model assumptions of approximate stationarity, both in time and in space, are checked against the historical record. The unpooled monthly sample statistics: mean, CV, skewness, and autocorrelation, are computed for each site, month, and year. The monthly variation of the sample statistics was compared between the historical record, 24-hour and 1-hour, and a 300-year model simulation record for the equivalent aggregation levels. Obviously, if the model, for which stationarity is assumed, can produce the same sample characteristics as seen in the historical record, then the assumption is valid - at least at the level required by the model.

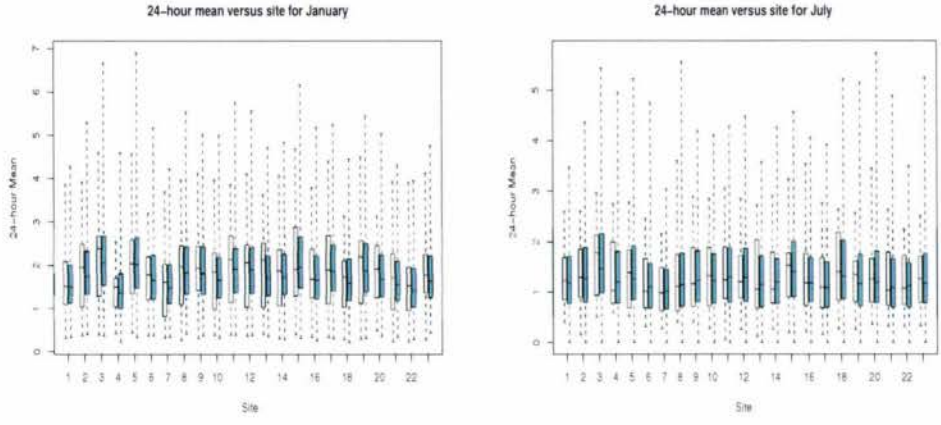
Each statistic listed is examined in a separate subsection, where the analysis is split into 24-hour and 1-hour levels and by season. This gives a total of 48 plots for the 24-hour level and 36 plots for the 1-hour aggregation level (as an analysis of the means is not necessary). The full list of plots is given in Appendix C Figures C.1 to C.24, but the results for January and July are given in text.

In order to compare the monthly distributions obtained from the historical and simulation records a two-sample Kolmogorov-Smirnov (K-S) test was computed by site and season. This test is of limited value due to equal weighting being placed on the historical data values regardless of the number of data points contributing to that value, but, nevertheless, is able to give some indication of the likelihood

of the distributions matching. A table is produced for each statistic examined (Tables 5.7 to 5.10) of the P-Values for each of the Kolmogorov-Smirnov tests.

Monthly means

For the two seasons included (January and July) it is clear that the distributions of the means are similar for all sites for the synthetic and the historical record (Figure 5.11). Furthermore, the results for the K-S tests for the monthly means (Table 5.7) are indicative that, particularly for the 24-hour data, the distributions of the monthly means can not be rejected from coming from the same distribution (at the 5% significance level). These results are as expected as the mean is matched exactly via θ in the model fitting.



White = Historical; Shaded = $Model_B$

Figure 5.11: Monthly 24-hour means: historical versus 300 year simulation - Jan,July

Monthly coefficient of variation

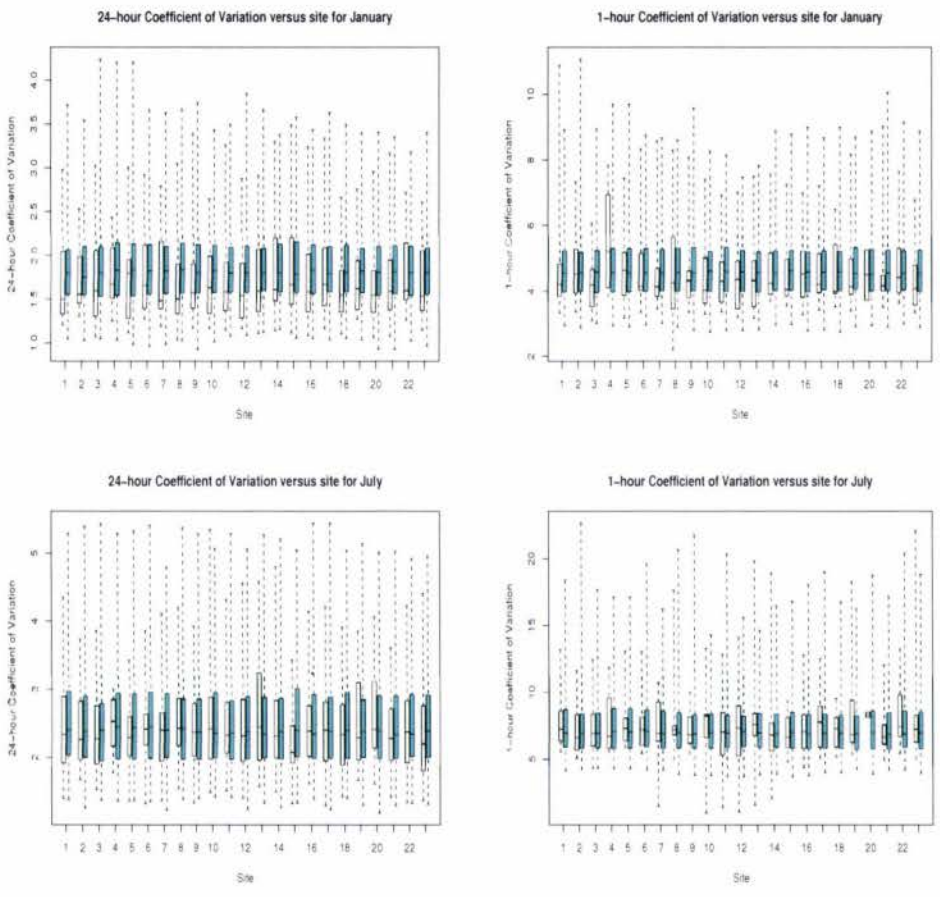
From the boxplots of the monthly variation for January (Figure 5.12), it is clear that for the 24-hour record, the CV tends to be overestimated. At the 1-hour level, this would also seem to be the case. The P-Values obtained from the corresponding two-sample K-S tests (Table 5.8), only the season of January shows up as having

Table 5.7: Kolmogorov-Smirnov test p-values: monthly mean simulated versus historical

	24-hour comparison by season-site											
Month	1	2	3	4	5	6	7	8	9	10	11	12
Jan	0.84	0.67	0.39	0.86	0.95	0.67	0.58	0.63	0.79	0.65	0.45	0.50
Feb	0.16	0.28	0.33	0.96	0.25	0.10	0.09	0.55	0.71	0.44	0.12	0.37
Mar	0.62	0.35	0.48	0.95	0.71	0.39	0.38	0.48	0.44	0.73	0.16	0.33
Apr	0.33	0.36	0.39	0.74	0.33	0.29	0.56	0.28	0.59	0.65	0.50	0.68
May	0.39	0.47	0.73	0.28	0.91	0.22	0.32	0.44	0.58	0.69	0.43	0.23
Jun	0.49	0.66	0.30	0.49	0.13	0.27	0.13	0.69	0.28	0.16	0.59	0.25
Jul	0.58	0.79	0.26	0.99	0.98	0.93	0.75	0.93	0.84	0.76	0.82	0.97
Aug	0.80	0.70	0.97	0.97	0.77	0.64	0.26	0.21	0.84	0.67	0.70	0.57
Sept	0.28	0.86	0.04	0.95	0.75	0.38	0.41	0.23	0.29	0.11	0.22	0.24
Oct	0.11	0.62	0.79	0.67	0.48	0.11	0.24	0.24	0.23	0.11	0.15	0.41
Nov	0.14	0.40	0.06	0.99	0.37	0.56	0.45	0.38	0.42	0.15	0.12	0.14
Dec	0.47	0.52	0.54	0.90	0.36	0.56	0.53	0.23	0.65	0.83	0.93	0.62
Month	13	14	15	16	17	18	19	20	21	22	23	
Jan	0.36	0.32	0.45	0.66	0.38	0.84	0.49	0.78	0.53	0.90	0.76	
Feb	0.44	0.38	0.66	0.45	0.74	0.79	0.74	0.97	0.23	0.27	0.31	
Mar	0.26	0.37	0.71	0.35	0.64	0.46	0.50	0.84	0.40	0.31	0.22	
Apr	0.79	0.35	0.31	0.42	0.11	0.78	0.38	0.61	0.46	0.39	0.22	
May	0.50	0.63	0.85	0.36	0.39	0.11	0.61	0.32	0.28	0.15	0.73	
Jun	0.19	0.68	0.39	0.15	0.12	0.34	0.65	0.53	0.32	0.54	0.28	
Jul	0.70	0.74	0.50	0.98	0.94	0.38	0.70	1.00	0.55	0.83	0.91	
Aug	0.81	0.45	0.51	0.83	0.86	0.75	0.89	0.79	0.68	0.84	0.66	
Sept	0.08	0.10	0.41	0.09	0.12	0.20	0.07	0.39	0.11	0.12	0.56	
Oct	0.50	0.26	0.76	0.34	0.47	0.40	0.55	0.59	0.16	0.56	0.26	
Nov	0.62	0.11	0.08	0.18	0.28	0.51	0.14	0.27	0.52	0.30	0.12	
Dec	0.53	0.59	0.39	0.62	0.60	0.75	0.69	0.74	0.81	0.25	0.43	

NB: 1-hour mean not included as this is the same as 24-hour mean / 24.

any major issues - and then only at the 24-hour level. This overestimation of January fits with what was observed previously (Section 5.2.2) where it was noted that the smoothing had significantly altered the estimates for the pooled statistics. Furthermore in the previous section (Section 5.3.1) it was seen that, again, January was not modelled well at any aggregation level for any statistic when compared to the historical estimates.



White = Historical; Shaded = $Model_B$

Figure 5.12: Monthly CV: historical versus 300 year simulation - Jan,July

Table 5.8: Kolmogorov-Smirnov test p-values: monthly CV simulated versus historical

	1-hour comparison by season-site																					
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.66	0.94	0.27	0.42	0.90	0.53	0.58	0.13	0.74	0.53	0.29	0.34	0.25	0.69	0.16	0.74	0.60	0.33	0.40	0.97	0.26	0.74	0.31
Feb	0.79	0.41	0.13	0.99	0.04	0.68	0.85	0.56	0.96	0.44	0.62	0.09	0.28	0.88	0.62	0.67	0.86	0.56	0.24	0.26	0.98	0.90	0.91
Mar	0.09	0.11	0.12	0.27	0.28	0.14	0.06	0.13	0.09	0.62	0.43	0.66	0.47	0.22	0.13	0.35	0.11	0.25	0.13	0.59	0.05	0.01	0.01
Apr	0.60	0.44	0.58	0.92	0.34	0.66	0.63	0.51	0.68	0.29	0.25	0.12	0.11	0.20	0.55	0.26	0.26	0.96	0.37	0.74	0.77	0.75	0.49
May	0.09	0.48	0.33	0.09	0.43	0.26	0.01	0.12	0.73	0.55	0.04	0.09	0.96	0.22	0.18	0.68	0.38	0.09	0.13	0.00	0.13	0.03	0.33
Jun	0.35	0.29	0.58	0.63	0.28	0.44	0.47	0.21	0.39	0.45	0.64	0.51	0.55	0.81	0.81	0.80	0.62	0.07	0.60	0.46	0.73	0.75	0.62
Jul	0.99	0.92	0.94	0.97	0.60	0.97	0.51	0.49	0.83	0.18	0.61	0.43	0.45	0.55	0.48	0.77	0.60	0.98	0.77	0.20	0.49	0.49	0.98
Aug	0.53	0.61	0.30	0.98	0.29	0.55	0.79	0.87	0.88	0.61	0.50	0.96	0.57	0.89	0.19	0.27	0.37	0.88	0.23	0.55	0.53	0.04	0.05
Sept	0.30	0.05	0.54	0.26	0.14	0.26	0.41	0.43	0.60	0.36	0.27	0.55	0.11	0.33	0.10	0.86	0.05	0.12	0.02	0.01	0.19	0.44	0.09
Oct	0.98	0.22	0.11	0.30	0.05	0.12	0.24	0.90	0.17	0.39	0.05	0.03	0.04	0.62	0.10	0.01	0.16	0.95	0.03	0.93	0.12	0.09	0.10
Nov	0.75	0.22	0.32	0.13	0.93	0.90	0.83	0.47	0.76	0.55	0.73	0.75	0.70	0.49	0.53	0.68	0.31	0.47	0.70	0.06	0.94	0.25	0.63
Dec	0.76	0.16	0.10	0.85	0.09	0.36	0.25	0.61	0.65	0.10	0.63	0.41	0.81	0.52	0.27	0.15	0.74	0.41	0.25	0.57	0.66	0.44	0.55
	24-hour comparison by season-site																						
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.00	0.05	0.08	0.86	0.19	0.43	0.00	0.00	0.03	0.08	0.03	0.01	0.24	0.04	0.38	0.01	0.33	0.03	0.22	0.11	0.06	0.07	0.01
Feb	0.57	0.66	1.00	0.59	0.87	0.64	0.60	0.67	0.47	0.73	0.65	0.74	0.42	0.34	0.83	0.90	0.59	0.10	0.34	0.74	0.29	0.58	0.64
Mar	0.19	0.24	0.03	0.10	0.35	0.73	0.62	0.79	0.74	0.21	0.03	0.01	0.20	0.09	0.12	0.42	0.21	0.04	0.55	0.26	0.55	0.65	0.36
Apr	0.15	0.21	0.14	0.10	0.29	0.05	0.01	0.30	0.13	0.31	0.19	0.30	0.38	0.18	0.15	0.19	0.31	0.20	0.30	0.28	0.16	0.17	0.21
May	0.47	0.35	0.02	0.86	0.00	0.29	0.79	0.55	0.36	0.26	0.06	0.05	0.06	0.20	0.19	0.19	0.14	0.71	0.91	0.85	0.61	0.79	0.51
Jun	0.29	0.16	0.08	0.53	0.37	0.09	0.65	0.97	0.22	0.90	0.93	0.56	0.58	0.96	0.56	0.94	0.78	0.78	0.92	0.63	0.69	0.69	0.93
Jul	0.76	0.85	0.56	0.60	0.42	0.53	0.73	0.96	0.94	0.58	0.97	0.61	0.96	0.92	0.08	0.99	0.85	0.54	0.72	0.95	0.28	0.71	0.22
Aug	0.70	0.29	0.47	0.54	0.58	0.12	0.47	0.08	0.01	0.03	0.77	0.77	0.44	0.56	0.23	0.25	0.95	0.63	0.48	0.87	0.26	0.80	0.69
Sept	0.11	0.05	0.74	0.75	0.03	0.16	0.11	0.54	0.94	0.35	0.37	0.21	0.55	0.40	0.23	0.65	0.17	0.86	0.16	0.60	0.59	0.18	0.02
Oct	0.09	0.95	0.79	1.00	0.03	0.72	0.91	0.47	0.45	0.66	0.83	0.78	0.50	0.75	0.15	0.38	0.46	0.44	0.19	0.91	0.71	0.44	0.39
Nov	0.36	0.21	0.18	0.66	0.69	0.97	0.20	0.86	0.44	0.82	0.42	0.51	0.75	0.31	0.49	0.79	0.69	0.55	0.89	0.99	0.29	0.53	0.44
Dec	0.14	0.89	0.12	0.65	0.46	0.73	0.25	0.30	0.07	0.57	0.98	0.77	0.47	0.51	0.99	0.74	0.15	0.39	0.69	0.33	0.38	0.13	0.31

Monthly skewness

The skewness generally seems to be matching satisfactorily for all the sites (Figure 5.13 and Table 5.9). Note that site number 20, (TW289022), has a low number of months at an 1-hour level - thus the apparent discrepancy in the plot. As expected, the K-S test does not pick this sample (July site 20) up as significantly different from the model.

The results in Table 5.9 from the K-S tests, indicate that, in general, there is no difference between the monthly variation in skewness between the simulated and historical records. However, in December the number of significantly different sites at the 24-hour aggregation level seems a little too high (Figure B.16), and similarly for October at the 1-hour aggregation level (Figure B.14). From these plots (Figures B.14 and B.16), it can be seen that the synthetic 1-hour skewness for October is generally higher than the historical skewness and the 24-hour skewness for December is generally slightly lower than for the historical skewness.

Monthly autocorrelation

Given the variation of the autocorrelation, particularly in the 24-hour record, this was deemed the least useful out of these three analyses. The boxplots (Figure 5.14) show that there is considerable variation within the historical data which is not being captured by the simulated record - particularly at the 1-hour aggregation level. For the most part, the 24-hour aggregation level seems to be matching (at January and July at least).

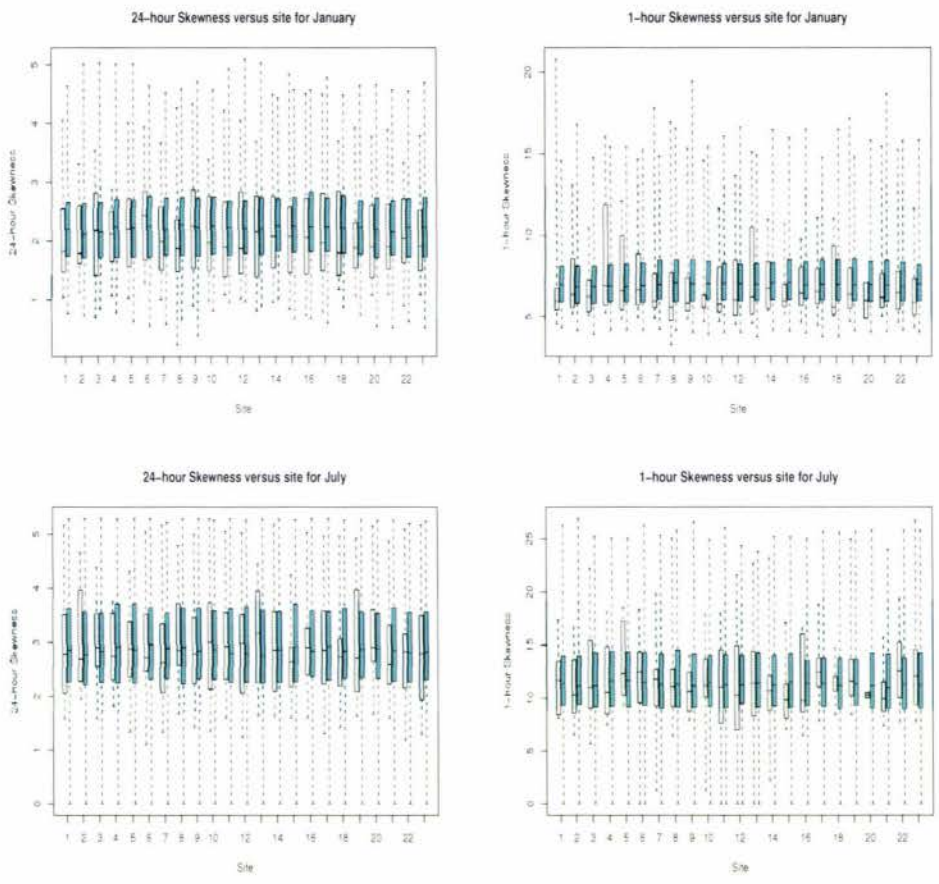
The results for the Kolmogorov-Smirnov tests (Table 5.10) indicate that the 1-hour correlation is generally matched - including for January observed previously. However, the 24-hour autocorrelation for the months June through November tend to have far too many significantly different sites within the respective months. This problem may result from the assumption that the variation seen in the 24-hour autocorrelation is constant across all seasons being incorrect (Section 5.2.2).

Table 5.9: Kolmogorov-Smirnov test p-values: monthly skewness simulated versus historical

	1-hour comparison by season-site																					
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.02	0.79	0.40	0.92	0.43	0.90	0.13	0.10	0.11	0.03	0.14	0.25	0.25	0.81	0.26	0.72	1.00	0.11	0.58	0.71	0.42	0.71	0.12
Feb	0.63	0.71	0.83	0.59	0.07	0.64	0.01	0.44	0.05	0.74	0.45	0.04	0.87	0.55	0.09	0.47	0.92	0.20	0.19	0.43	0.99	0.44	0.62
Mar	0.25	0.09	0.23	0.29	0.34	0.16	0.16	0.19	0.52	0.53	0.52	0.94	0.32	0.45	0.62	0.42	0.12	0.33	0.16	0.81	0.22	0.04	0.04
Apr	0.21	0.37	0.20	0.01	0.80	0.49	0.57	0.16	0.66	0.03	0.04	0.01	0.02	0.11	0.84	0.09	0.10	0.44	0.28	0.43	0.08	1.00	0.57
May	0.41	0.46	0.52	0.68	0.31	0.19	0.20	0.63	0.85	0.16	0.18	0.56	0.69	0.33	0.16	0.31	0.94	0.73	0.83	0.04	0.08	0.17	0.48
Jun	1.00	0.14	0.18	0.20	0.54	0.58	0.45	0.60	0.26	0.61	0.88	0.27	0.25	0.52	0.50	0.56	0.03	0.23	0.56	0.71	0.96	0.94	0.53
Jul	0.63	0.49	0.68	0.96	0.63	0.97	0.99	0.62	0.62	0.97	0.32	0.25	0.67	0.45	0.08	0.21	0.37	0.98	0.90	0.38	0.14	0.78	0.94
Aug	0.67	0.51	0.21	0.62	0.40	0.48	0.89	0.97	0.43	0.82	0.98	0.93	0.61	0.05	0.49	0.58	0.46	0.92	0.70	0.79	0.87	0.21	0.31
Sept	0.89	0.21	0.34	0.59	0.14	0.79	0.72	0.52	0.59	0.37	0.44	0.77	0.68	0.25	0.94	0.56	0.87	0.01	0.29	0.01	0.44	0.62	0.10
Oct	0.55	0.11	0.07	0.67	0.02	0.01	0.08	0.72	0.14	0.32	0.02	0.01	0.05	0.25	0.10	0.00	0.19	0.42	0.07	0.75	0.02	0.02	0.08
Nov	0.78	0.58	0.33	0.09	0.93	0.37	0.74	0.66	0.79	0.05	0.50	0.17	0.96	0.75	0.89	0.23	0.17	0.20	0.27	0.08	0.92	0.14	0.41
Dec	0.29	0.24	0.30	0.95	0.79	0.24	0.11	0.86	0.18	0.83	0.22	0.99	0.68	0.31	0.52	0.98	0.44	0.84	0.29	0.35	0.43	0.08	0.47
	24-hour comparison by season-site																						
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.12	0.15	0.16	0.71	0.86	0.64	0.17	0.12	0.33	0.12	0.15	0.15	0.51	0.57	0.33	0.26	0.65	0.06	0.38	0.30	0.08	0.77	0.24
Feb	0.44	0.72	0.40	0.25	0.28	0.75	0.40	0.70	0.50	0.80	0.40	0.85	0.34	0.09	0.03	0.42	0.14	0.10	0.46	0.56	0.19	0.43	0.15
Mar	0.38	0.12	0.98	0.88	0.69	0.56	0.49	0.53	0.67	0.81	0.81	0.55	0.70	0.79	0.41	0.79	0.16	0.61	0.21	0.50	0.95	0.25	0.26
Apr	0.31	0.55	0.18	0.35	0.59	0.16	0.07	0.08	0.15	0.66	0.17	0.47	0.17	0.15	0.38	0.09	0.45	0.24	0.47	0.33	0.03	0.14	0.62
May	0.85	0.71	0.08	0.79	0.01	0.71	0.56	0.20	0.97	0.82	0.62	0.08	0.14	0.63	0.58	0.23	0.41	0.27	0.86	0.06	0.05	0.24	0.54
Jun	0.76	0.85	0.38	0.16	0.90	0.84	0.84	0.98	0.69	0.97	0.65	1.00	0.34	0.58	0.67	0.93	0.65	0.96	0.65	0.72	0.44	0.66	0.91
Jul	0.39	0.72	0.57	0.82	0.85	0.77	0.28	0.73	0.82	0.99	0.93	0.56	0.50	0.67	0.06	0.43	0.97	0.38	0.65	0.42	0.27	0.15	0.24
Aug	0.94	0.31	0.65	0.32	0.88	0.49	0.68	0.00	0.01	0.13	0.18	0.10	0.25	0.11	0.15	0.30	0.54	0.72	0.28	0.14	0.07	0.87	0.90
Sept	0.72	0.26	0.08	0.64	0.08	0.89	0.59	0.59	0.14	0.37	0.72	0.38	0.31	0.14	0.83	0.05	0.46	0.62	0.18	0.49	0.13	0.25	0.87
Oct	0.25	0.93	0.66	0.09	0.11	0.71	0.94	0.55	0.40	0.94	0.87	0.67	0.47	0.74	0.09	0.24	0.53	0.88	0.70	0.74	0.47	0.61	0.74
Nov	0.85	0.33	0.77	0.29	0.67	0.87	0.45	0.78	0.28	0.69	0.45	0.77	0.85	0.94	0.80	0.80	0.80	0.44	0.57	0.67	0.35	0.93	0.84
Dec	0.04	0.60	0.36	0.37	0.33	0.01	0.06	0.01	0.00	0.23	0.48	0.07	0.22	0.18	0.01	0.19	0.00	0.00	0.10	0.18	0.02	0.02	0.06

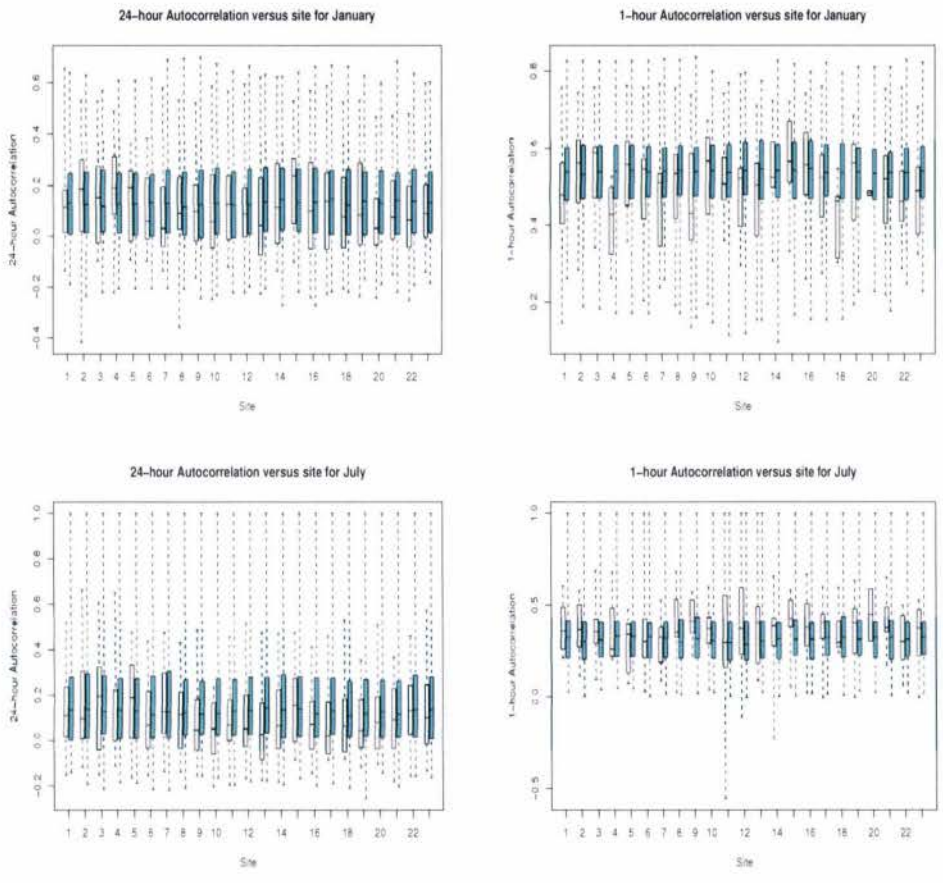
Table 5.10: Kolmogorov-Smirnov test p-values: monthly autocorrelation simulated versus historical

	1-hour comparison by season-site																						
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.15	0.90	0.16	0.12	0.97	0.81	0.07	0.56	0.14	0.82	0.23	0.41	0.57	0.74	0.53	0.82	0.77	0.03	0.37	0.22	0.71	0.06	0.17
Feb	0.48	0.18	0.03	0.15	0.07	0.73	0.22	0.01	0.24	0.06	0.03	0.02	0.10	0.01	0.10	0.99	0.06	0.12	0.01	0.12	0.07	0.05	0.07
Mar	0.45	0.95	0.49	0.50	0.53	0.78	0.56	0.54	0.94	0.42	0.44	0.91	0.30	0.19	0.36	0.03	0.11	0.38	0.16	0.40	0.71	0.25	0.22
Apr	0.10	0.77	1.00	0.02	0.83	0.12	0.51	0.38	0.20	0.00	0.02	0.03	0.01	0.43	0.72	0.15	0.19	0.00	0.48	0.12	0.83	0.21	0.36
May	0.69	0.33	0.81	0.75	0.13	0.29	0.32	0.12	0.01	0.38	0.10	0.26	0.24	0.71	0.37	0.57	0.62	0.13	0.21	0.45	0.26	0.29	0.20
Jun	0.29	0.10	0.10	0.01	0.30	0.22	0.92	0.15	0.33	0.21	0.43	0.88	0.15	0.18	0.95	0.06	0.09	0.12	0.58	0.93	0.44	0.03	0.31
Jul	0.56	0.12	0.24	0.73	0.74	0.77	0.73	0.24	0.22	0.71	0.29	0.36	0.21	0.68	0.02	0.35	0.22	0.88	0.32	0.77	0.03	0.80	0.62
Aug	0.50	0.10	0.62	0.29	0.58	0.67	0.47	0.89	0.28	0.10	0.63	0.77	0.43	0.01	0.75	0.12	0.37	0.81	0.58	0.88	0.53	0.51	0.11
Sept	0.29	0.99	0.02	0.08	0.40	0.65	0.83	0.33	0.73	0.91	0.21	0.57	0.33	0.91	0.66	0.82	0.12	0.64	0.62	0.01	0.08	0.86	0.47
Oct	0.02	0.21	0.46	0.87	0.45	0.11	0.62	0.28	0.29	0.53	0.30	0.65	0.64	0.48	1.00	0.54	0.66	0.14	0.95	0.99	0.90	0.75	0.36
Nov	0.25	0.09	0.64	0.06	0.44	0.11	0.10	0.72	0.23	0.94	0.99	0.95	0.58	0.05	0.86	0.03	0.23	0.78	0.23	0.01	0.05	0.04	0.11
Dec	0.34	0.40	0.56	0.95	0.25	0.98	0.09	0.54	0.94	0.23	0.17	0.60	0.46	0.78	0.12	0.99	0.73	0.14	0.06	0.76	0.76	0.40	0.75
	24-hour comparison by season-site																						
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Jan	0.22	0.46	0.44	0.46	0.79	0.56	0.15	0.99	0.60	0.24	0.98	0.41	0.14	0.64	0.12	0.50	0.40	0.52	0.26	0.08	0.45	0.19	0.45
Feb	0.64	0.07	0.14	0.97	0.29	0.09	0.29	0.48	0.49	0.17	0.22	0.70	0.18	0.17	0.14	0.62	0.37	0.90	0.18	0.52	0.17	0.09	0.24
Mar	0.93	0.88	0.95	0.10	0.70	0.64	0.92	0.81	0.63	0.54	0.70	0.99	0.20	0.79	0.97	0.78	0.31	0.31	0.62	0.77	0.44	0.92	0.45
Apr	0.29	0.91	0.93	0.72	0.26	0.92	0.36	0.57	0.54	0.62	0.45	0.81	0.46	0.45	0.72	0.89	0.49	0.53	0.29	0.48	0.32	0.14	0.42
May	0.02	0.38	0.34	0.28	0.06	0.14	0.04	0.27	0.11	0.28	0.33	0.31	0.72	0.60	0.11	0.49	0.74	0.72	0.88	0.70	0.04	0.66	0.42
Jun	0.05	0.13	0.04	0.18	0.66	0.08	0.04	0.06	0.86	0.76	0.04	0.03	0.10	0.07	0.02	0.08	0.02	0.74	0.09	0.90	0.05	0.16	0.03
Jul	0.60	0.58	0.24	0.51	0.51	0.38	0.85	0.47	0.17	0.01	0.44	0.37	0.13	0.15	0.87	0.26	0.00	0.22	0.20	0.14	0.40	0.99	0.93
Aug	0.03	0.04	0.70	0.43	0.28	0.00	0.01	0.19	0.01	0.00	0.19	0.12	0.48	0.09	0.15	0.12	0.26	0.00	0.03	0.60	0.01	0.00	0.00
Sept	0.93	0.55	0.11	0.09	0.87	0.89	0.58	0.65	0.35	0.69	0.68	0.54	0.81	0.46	0.58	0.38	0.24	0.82	0.43	0.28	0.89	0.73	0.56
Oct	0.01	0.14	0.38	0.12	0.19	0.18	0.00	0.03	0.15	0.20	0.17	0.21	0.26	0.47	0.39	0.25	0.48	0.12	0.77	0.65	0.15	0.00	0.76
Nov	0.01	0.08	0.02	0.85	0.03	0.01	0.00	0.01	0.01	0.01	0.03	0.03	0.21	0.00	0.03	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00
Dec	0.22	0.78	0.58	0.23	0.78	0.80	0.74	0.34	0.60	0.52	0.77	1.00	0.88	0.19	0.46	0.21	0.12	0.26	0.40	0.96	0.60	0.65	0.79



White = Historical; Shaded = $Model_B$

Figure 5.13: Monthly skewness: historical versus 300 year simulation - Jan,Jul



White = Historical; Shaded = $Model_B$

Figure 5.14: Monthly autocorrelation: historical versus 300 year simulation

Homogeneity of region

The consistency between the statistical variation across the sites between the historical record and the simulated data for the CV, skewness, and 1st lag autocorrelation, shows that the assumption of homogeneity in the region has been satisfied. Furthermore, with the exception of where the data record is short, there is no site that has a different distribution of the statistics from the other sites within the historical data. The transformation to spatial stationarity by dividing each site-month by its respective mean (Section 3.1.1) has been successful.

5.3.3 Stability

The purpose of the stability analysis is to determine the stability of a 300 year period as a synthetic record of this length was to be used for the infilling of the historical records. If a 300 year period is not sufficiently stable, then, for useful infilling, a longer record may be necessary. Obviously, a shorter record is desirable as it is less demanding on physical resources and, in this example, the maximum record length able to be handled comfortably is approximately 400 years. For 400 years, the memory usage is around $650Mb$ for the simulated rainfall records alone.

Therefore, $Model_B$ was simulated for 300 years 30 times and the pooled statistics recorded after each run. As before, the results are split by aggregation level (Figures 5.15 to 5.17). The historical and smoothed estimates are overlaid on the plot as indicative measures only as the increased sample size is expected to reduce the variation to centre round the fitted values not the historical estimates. For an analysis of the probability of observing the historical record given the model see Section 5.3.1. While 30 simulations are hardly enough to get an accurate picture of the sampling variation for the pooled statistics, they are adequate, given the large sample size (300 months), to give an approximate distribution.

The CV for the 1-hour and 24-hour values (Figure 5.15) shows that, for both these aggregation levels, the pooled CV is well centered. There is slightly more variation over the summer months compared to the winter months - particularly at the 1-hour aggregation level.

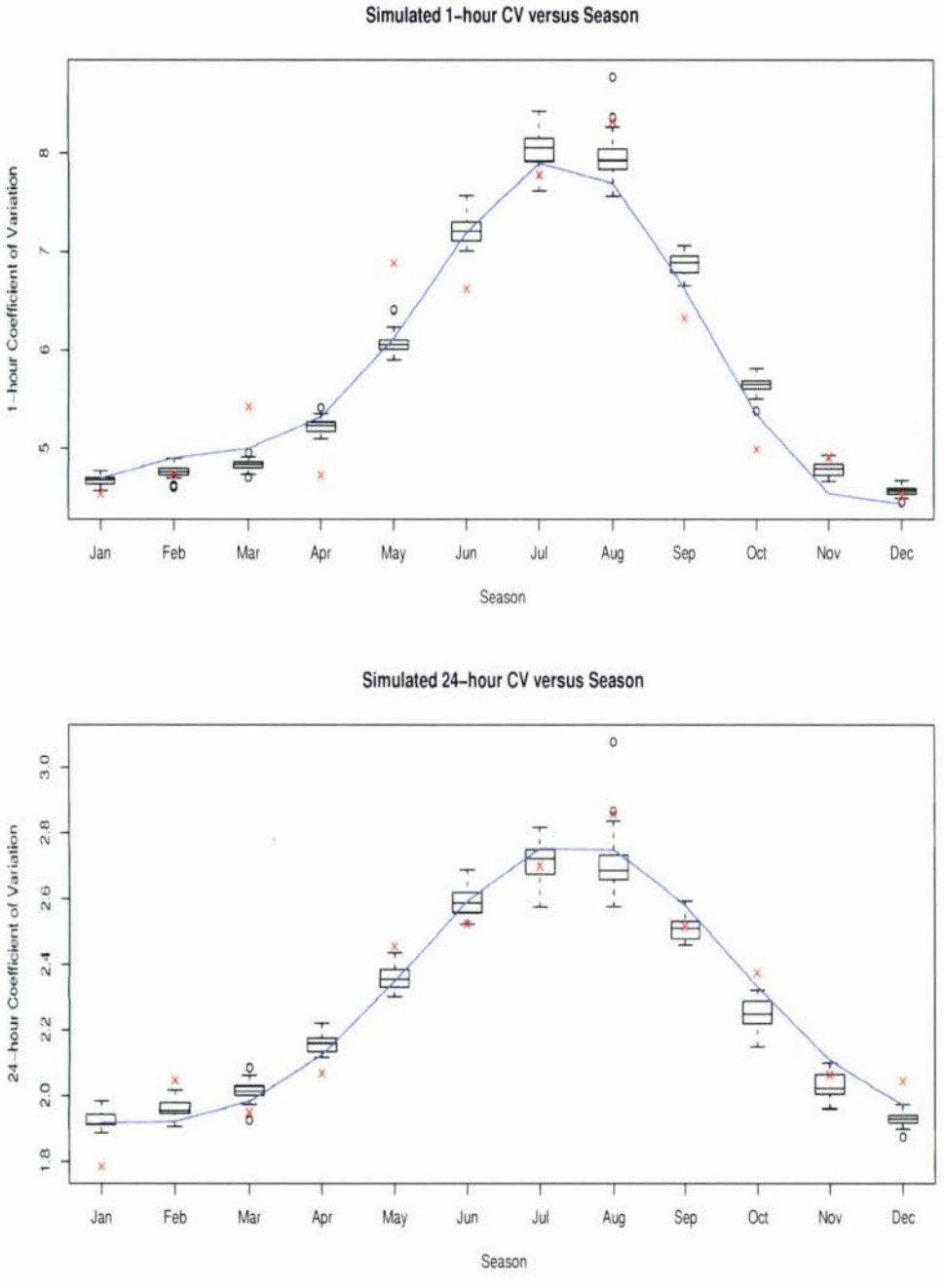
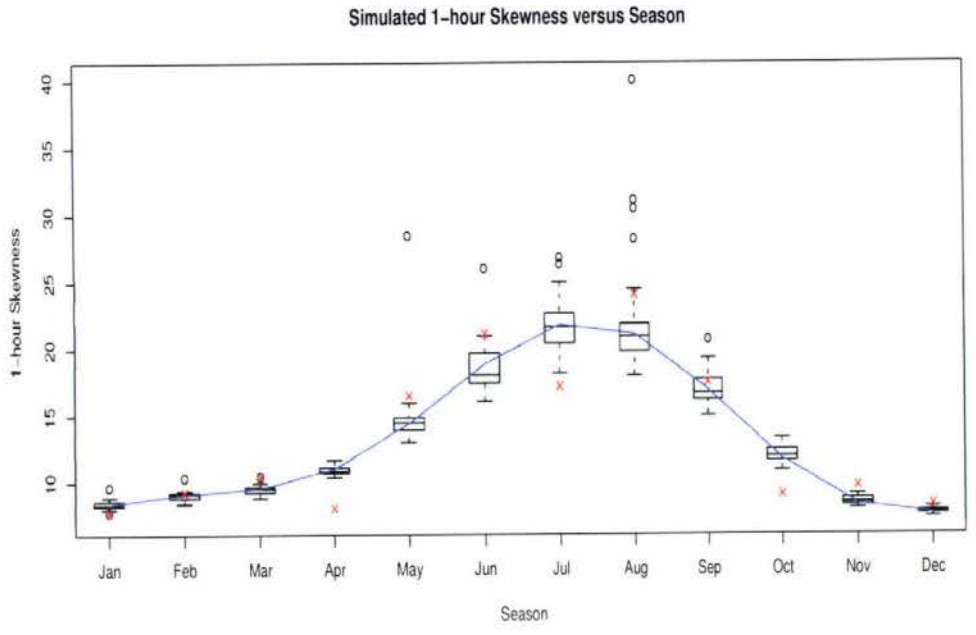


Figure 5.15: Stability of 300 year sample - pooled CV

It is evident from the skewness plots (Figure 5.16) that, while the simulations are relatively centered around the smoothed value, the potential for 'outliers' for an overall pooled statistic, even for a 300 year simulation, is quite high. Granted, this is only a sample size of 30 simulations, but a pooled skewness of 16 at the 24-hour level and 40 at the 1-hour level when the expected value is half that, is a considerable difference given that the statistic is based on 23 sites over approximately 30 days for 300 years (for 24 hours at the 1-hour aggregation level). This *extreme* fitting is a feature of the independent processes that forms the simulation algorithm.

The autocorrelation plots (Figure 5.17) shows no problems with the centrality of the fitting. From the 24-hour autocorrelation, the sample historical values, are actually included within the 300 year variation for 6 of the months. Thus the assumption that the variation seen in the 24-hour autocorrelation was due to sampling variation alone is vindicated.

Based on the results (Figures 5.15 to 5.17), the simulations are sufficiently stable to be used for infilling. As with any sampling with independent processes, there is a possibility of extreme fitting occurring relative to the central fit, but, with a sample size of 300 months per model parameterisation, it is clear that the tendency to the exact fit is quite high (even for skewness). Obviously, with a larger sample size, this tendency will be even more apparent. For the purpose of the analysis, however, a 300 year sample size is adequate to use as a source for infilling. This is important as hardware limitations (particularly memory space) become influential once the simulation size is increased.



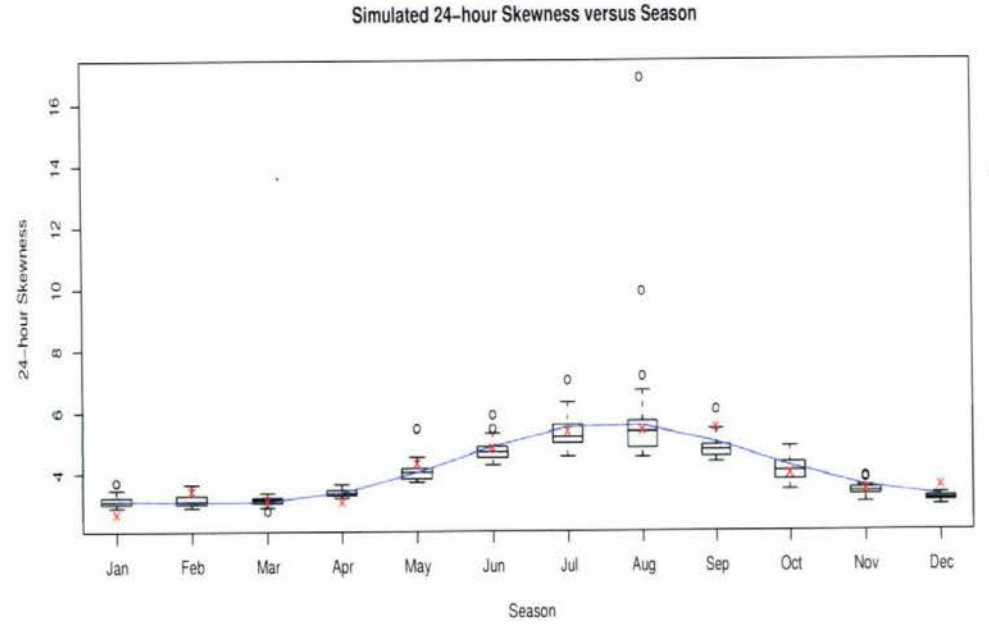


Figure 5.16: Stability of 300 year sample - pooled skewness

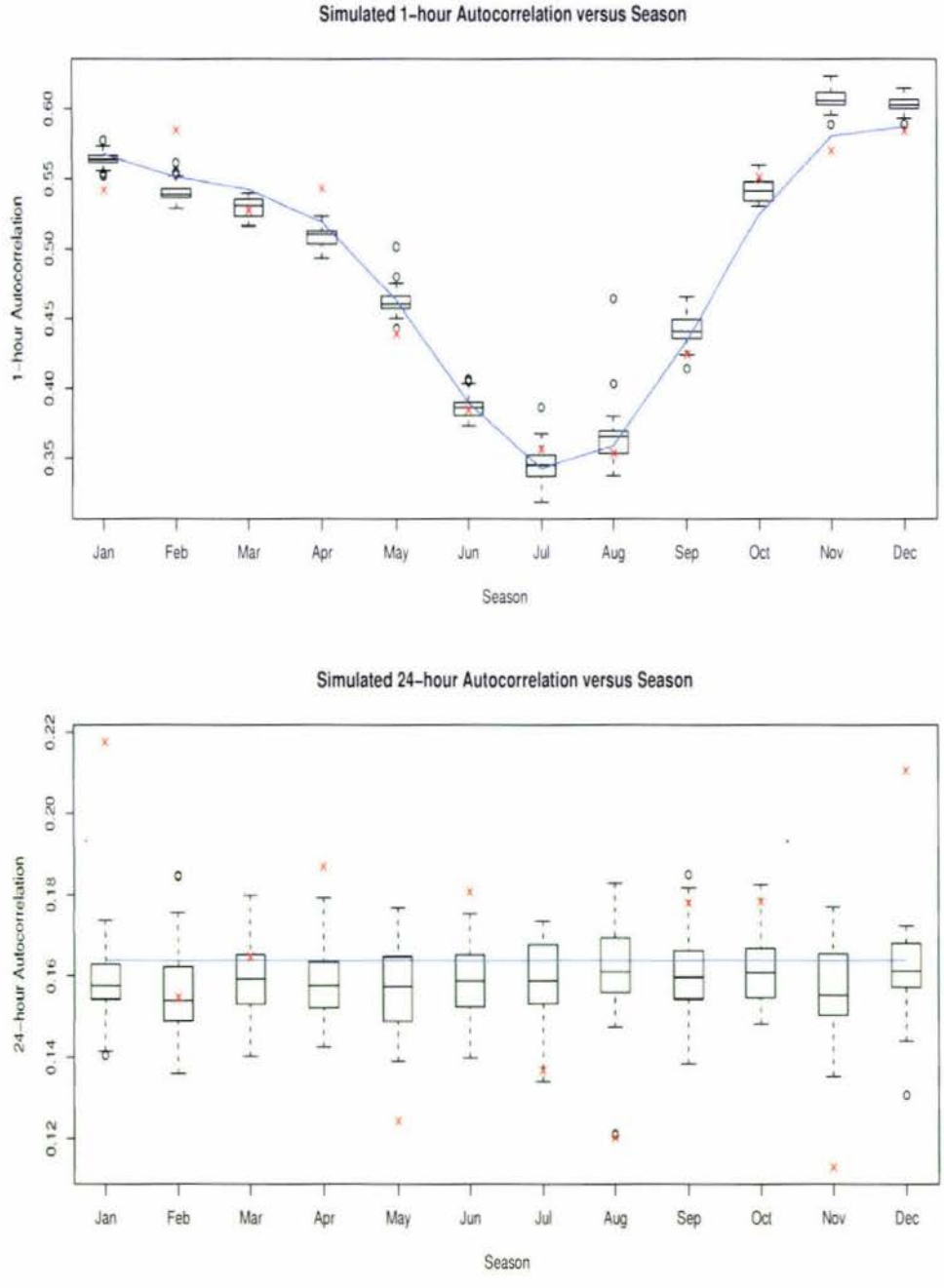


Figure 5.17: Stability of 300 year sample - pooled autocorrelation

5.3.4 Summary

Either model, A or B, is suitable for simulating rainfall data (Section 5.3.1). The cross-correlation (Section 5.3.1) is generally overestimated at higher distances, and this is expected to have an impact on infilling accuracy. Furthermore, the variation in the sample estimates, particularly skewness, is considerable for either model - even over a 300 year sampling period. Thus, it is expected that any difference in the infilling results is most likely to be a result of simulation variation rather than a beneficial model change. As $Model_B$ is a simpler model, this is the model that is used to produce the simulation data for the infilling algorithms (Section 5.5).

5.4 Fitting algorithm heuristics

In the original algorithm development, a fit was generated for every available simulated record in order to select the best possible fit (Figure 5.19). However, this is not the most efficient way of selecting candidate rows when valid historical data are present. Heuristics making further use of the data can restrict the search for a good fit to directions where this is more likely. Therefore, such a heuristic for reducing the search time was investigated.

During the implementation of the infilling algorithms, the first heuristic proposed used partitioning of the simulation data based on the number of wet days in the record. This heuristic resulted in a significant speed increase and no further developments were analysed at that point. However, as the project was being written up, two additional heuristics were proposed and are discussed in Section 6.5.1.

5.4.1 Partitioning of wet/dry days

When a historical data point is present, it was observed that the following information can instantly be obtained - whether it is wet ($+ve$) or dry (0) at the available site. Also, which sites have non-zero rainfall can also be obtained, however, for the model of the region this was found to not be overly useful information as there are few records with a mixture of wet and dry sites (see Section 5.4.1). For a large

simulation (say 1000 years), however, this may be a useful further partitioning option - especially if the model is more complex (see Section 2.1.2).

Let W be the total number of wet sites at time t,
D be the total number of dry sites at time t,
w_o be the number of observed wet sites,
d_o be the number of observed dry sites, and
N be the number of sites in the region.

Then, given w_o and d_o, it is evident that the W is bounded by (5.1) and (5.2) and D is bounded by (5.3) and (5.4).

$$\max(W) = N - d_o \tag{5.1}$$

$$\min(W) = w_o \tag{5.2}$$

$$\max(D) = N - w_o \tag{5.3}$$

$$\min(D) = d_o \tag{5.4}$$

To illustrate, let Figure 5.18 represent a sample historical record at some time point for 15 sites. In this example, the number of observed wet records (w_o) is 3, the number of observed dry records (d_o) is 5. Therefore, as $N = 15$, the maximum possible W is $(15 - 5 = 10)$ and the minimum possible W is 3 (the number of observed wet sites). Similarly, the maximum possible D is $(15 - 3 = 12)$ and the minimum possible D is 5 (the number of observed dry sites). Thus, for this example, the search for potential records to use for infilling would be restricted to those synthetic records where at least 3 sites are wet and at most 10 sites are wet.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Indicator	1	-	0	0	-	1	1	-	-	0	-	0	-	-	0

where: 1 represents a wet site
0 represents a dry site
- represents a missing record

Figure 5.18: Example historical record with wet/dry indicators

In order to make use of this heuristic effectively, the simulated records were stratified by the total number of wet sites in the region. For any record requiring infilling where historical data were available, w_o and d_o were obtained. The search for the a suitable record was then restricted to the sample within in the stratification levels: $N - d_o, N - d_o - 1, \ldots, d_o$. The order was important as the vast number of records in the simulation were either either all Wet or all Dry but, as it was more common that there was at least one dry site, preference was given to starting at the maximum number of wet sites. This does not affect the output of the algorithm, but did increase the speed of each algorithm examined by a factor of at least 4. The investigation was deemed to be successful and investigation time was reallocated back into algorithm development.

Analysis of partitions of wet/dry regions

An example partition with the number of records split by season is shown in Table 5.11. Note that one record has been removed from January and similarly with December so that the surrounding data (next and previous) is always available should this be used by the fitting algorithm. This was to prevent out of bounds errors should next or previous time points be considered in the fitting algorithm.

The results (Table 5.11) clearly highlight the benefit obtainable as soon as information is available about the magnitude of rainfall at sites within the region. In fact, just by knowing one site's value the searching can be reduced by approximately 30% if the site is dry or approximately 50% if the site is wet. Furthermore, the proportion of records where there is one dry site and one wet site is less than 15% during winter/spring and less than 20% over summer / autumn (see also Table 5.12).

The next question to answer is whether the simulated record, given above, matches the regional characteristics of the historical data. The historical 24-hour data were examined, and the records counted where more than one valid record was available and the results split into the categories: all dry, some dry, and all wet. The resulting table over the 35 year period of historical data is shown in Table 5.12 as proportions of all dry, some dry, and all wet - along with the corresponding proportions from the synthetic data (300-year record) for easy comparison.

Table 5.11: Example partitioning on number of wet sites: 24-hour aggregation level

Number of Wet Sites	Proportion in each category by season											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
0	0.480	0.531	0.540	0.573	0.614	0.596	0.622	0.611	0.561	0.555	0.499	0.480
1	0.010	0.009	0.010	0.009	0.012	0.016	0.016	0.014	0.016	0.010	0.012	0.011
2	0.008	0.006	0.007	0.008	0.009	0.011	0.010	0.012	0.012	0.007	0.008	0.008
3	0.007	0.005	0.006	0.005	0.006	0.009	0.009	0.008	0.010	0.006	0.006	0.005
4	0.005	0.004	0.003	0.005	0.005	0.007	0.007	0.007	0.008	0.005	0.003	0.004
5	0.004	0.003	0.005	0.003	0.004	0.007	0.006	0.007	0.008	0.006	0.005	0.004
6	0.005	0.005	0.005	0.004	0.005	0.006	0.007	0.007	0.005	0.005	0.004	0.005
7	0.004	0.004	0.005	0.003	0.004	0.007	0.005	0.007	0.006	0.005	0.006	0.005
8	0.005	0.004	0.005	0.005	0.006	0.008	0.005	0.005	0.007	0.006	0.006	0.007
9	0.005	0.004	0.003	0.004	0.004	0.006	0.005	0.006	0.006	0.004	0.004	0.005
10	0.004	0.005	0.006	0.005	0.005	0.006	0.005	0.006	0.006	0.004	0.004	0.005
11	0.006	0.005	0.005	0.004	0.004	0.006	0.005	0.005	0.006	0.005	0.004	0.007
12	0.005	0.004	0.003	0.003	0.004	0.005	0.005	0.004	0.006	0.005	0.005	0.006
13	0.004	0.006	0.005	0.006	0.004	0.005	0.005	0.006	0.008	0.006	0.005	0.004
14	0.005	0.004	0.004	0.004	0.005	0.005	0.006	0.005	0.006	0.004	0.005	0.006
15	0.006	0.005	0.005	0.004	0.005	0.005	0.005	0.005	0.007	0.004	0.006	0.005
16	0.006	0.004	0.005	0.004	0.005	0.006	0.004	0.006	0.006	0.004	0.006	0.005
17	0.005	0.003	0.004	0.006	0.006	0.007	0.005	0.006	0.007	0.004	0.006	0.006
18	0.006	0.005	0.004	0.005	0.006	0.006	0.008	0.007	0.006	0.006	0.005	0.008
19	0.006	0.004	0.005	0.005	0.005	0.007	0.007	0.008	0.008	0.007	0.006	0.007
20	0.008	0.008	0.007	0.005	0.006	0.010	0.010	0.007	0.009	0.006	0.008	0.007
21	0.011	0.008	0.006	0.008	0.008	0.010	0.010	0.011	0.013	0.009	0.009	0.011
22	0.014	0.012	0.013	0.012	0.013	0.019	0.018	0.022	0.018	0.014	0.015	0.016
23	0.383	0.375	0.339	0.310	0.256	0.231	0.217	0.216	0.256	0.314	0.364	0.374
Total	9299	8273	9300	9000	9300	9000	9300	9300	9000	9300	9000	9299

The proportions underscore a major difference between the simulated and historical records (Table 5.12) - a result which was not found in the previous chapter as a comparison of the proportion of dry records between the historical and simulated data was not examined. The simulated 24-hour records do not have the right mixture of dry and wet regions - they are practically dichotomous as all dry or all wet, whereas the historical records are obviously not so. The χ^2 statistic for comparing the counts, one for each season, has a probability of occurrence of 0 for all of them. Note also, because of the incompleteness of the data record, the counts for all wet and all dry may be accentuated beyond their true proportion. This would only further increase the discrepancy seen between the historical and simulated record.

As the model fits the data (Sections 5.2 and 5.3), the cutoff point of $0.1mm$ for a wet site was changed to determine the effect of this classification on the mismatch between the synthetic and historical record (Table 5.12). In order to obtain a measure of the difference between the two datasets (simulated and historical), a χ^2 test was bootstrapped 1000 times based on a sample size of 1000 records (1000 records is equivalent to 35 months of values recorded at a 24-hour aggregation level). (The counts in each record were set based on the proportions observed in Table 5.12). The bootstrapped χ^2 test produced results with a median p-value of less than 0.002 for the $1.0mm$ cutoff level. The other two levels were even more significantly different with a *maximum* p-value observed of 1×10^{-14}.

The results from this analysis, however, give further incentive to move away from the usage of the 'best fit' approach in infilling (Section 5.5). The reason for this is obvious, as any best fit algorithm requires the model to be accurately representing the underlying process. As seen above, the model does not emulate the true process behaviour adequately in the regional mix of wet and dry sites. Therefore, it is logical to deduce that any infilling algorithm making use of this model 'as is' is bound to be less effective than an algorithm that is able to account, in some way, for the potential discrepancy of the fitted model.

Table 5.12: 24-hour historical and simulated: Dry, Some Dry, and Wet

		dry cutoff = 0.05mm				dry cutoff = 0.1mm				dry cutoff = 1.0mm			
Month	Data	D	D+W	W	$P\chi^2$	D	D+W	W	$P\chi^2$	D	D+W	W	$P\chi^2$
Jan	H	0.172	0.479	0.349		0.172	0.479	0.349		0.479	0.277	0.244	
	S	0.471	0.128	0.401	7.1e-76	0.48	0.138	0.382	9.2e-74	0.568	0.199	0.232	2.0e-05
Feb	H	0.261	0.46	0.278		0.261	0.46	0.278		0.543	0.29	0.168	
	S	0.507	0.106	0.387	4.1e-70	0.518	0.115	0.366	3.2e-67	0.609	0.177	0.214	5.2e-09
Mar	H	0.277	0.441	0.282		0.277	0.442	0.281		0.522	0.305	0.174	
	S	0.531	0.115	0.354	1.1e-61	0.54	0.121	0.339	4.3e-60	0.616	0.181	0.203	3.6e-10
Apr	H	0.328	0.428	0.243		0.328	0.428	0.243		0.553	0.291	0.156	
	S	0.561	0.115	0.324	5.7e-56	0.573	0.118	0.31	4.1e-55	0.651	0.15	0.2	1.7e-13
May	H	0.356	0.42	0.224		0.356	0.42	0.224		0.546	0.306	0.147	
	S	0.605	0.127	0.268	3.9e-50	0.614	0.131	0.256	5.7e-49	0.677	0.158	0.165	2.0e-14
Jun	H	0.372	0.425	0.203		0.372	0.425	0.203		0.565	0.31	0.125	
	S	0.587	0.168	0.244	6.7e-37	0.596	0.173	0.231	4.7e-36	0.659	0.21	0.13	1.0e-06
Jul	H	0.398	0.425	0.177		0.398	0.425	0.177		0.605	0.289	0.105	
	S	0.612	0.156	0.232	1.8e-39	0.622	0.16	0.217	2.2e-38	0.692	0.196	0.112	3.4e-06
Aug	H	0.415	0.422	0.163		0.415	0.422	0.163		0.61	0.287	0.103	
	S	0.6	0.168	0.232	1.9e-34	0.611	0.174	0.216	4.0e-33	0.674	0.217	0.109	9.6e-04
Sep	H	0.375	0.389	0.236		0.375	0.389	0.236		0.573	0.273	0.155	
	S	0.551	0.178	0.271	7.1e-26	0.561	0.183	0.256	4.6e-25	0.629	0.226	0.146	1.5e-02
Oct	H	0.27	0.46	0.271		0.27	0.46	0.271		0.536	0.283	0.181	
	S	0.545	0.126	0.329	1.1e-63	0.555	0.131	0.314	2.7e-63	0.625	0.177	0.198	8.0e-08
Nov	H	0.176	0.517	0.307		0.176	0.517	0.307		0.506	0.287	0.207	
	S	0.489	0.131	0.38	1.7e-84	0.499	0.138	0.364	1.6e-83	0.576	0.193	0.231	3.4e-06
Dec	H	0.176	0.524	0.3		0.176	0.524	0.3		0.505	0.286	0.208	
	S	0.473	0.134	0.393	1.9e-83	0.48	0.146	0.374	2.1e-79	0.562	0.202	0.235	3.6e-05

where Data represents either historical (H) or simulated (S) records
$P\chi^2$ is the median p-value of the bootstrapped χ^2 statistic
D is proportion of times that all sites were dry
D+W is proportion of times that some sites were dry
W is the proportion of times that all sites were wet

5.5 *Infilling*

Man can learn nothing unless he proceeds from the known to the unknown.

- Claude Bernard (1813 - 1878)

5.5.1 Introduction

In this section, infilling algorithms for estimating the missing values in the historical record are presented and discussed. The algorithms are ordered such that each algorithm is a logical extension of the previous one. Furthermore, the results discussed are not intended to provide a definitive list of algorithms examined, but rather illustrate any improvements made. Attempted derivations that did not improve the results are briefly discussed in Section 5.5.6.

The algorithms herein make use of a simulation of a spatial-temporal NSRP model, $Model_B$, as fitted (Section 5.2). The model has been shown to fit the data (Section 5.3.1) and satisfies the assumptions of the model fitting (Section 5.3.2). The fitted model provides a specification of the whole region including the missing data provided the unobserved data follows a similar pattern to the observable values. Therefore, the fitted model is likely to be more beneficial than (for example) infilling with either stochastic linear regression or stochastic kriging.

The observed difference between the spatial characteristics of the spread of mixed wet and dry days (Section 5.4.1) is a potential problem for the infilling algorithms. However, as historical values are taken into account whenever available, it is possible that this problem may be negated by the selection of similar time points to the current mixture of wet/dry sites. Therefore, the assumption of the spatial-temporal NSRP model being representative enough for infilling is not unreasonable, and the fitted model, $Model_B$, can used to interpolate between points - both spatially and temporally.

The emphasis within this chapter is on the infilling results for the 24-hour record, but technically the algorithm could be applied to any aggregation level where there is data - for example the 1-hour record. However, prediction of 24-hour values is substantially easier than prediction of the 1-hour records as the variability between the rainfall values is significantly less than at the 1-hour aggregation level. This is inherently apparent as variability within discrete time steps is less evident as the aggregation level increases.

As the fitted model generates zero velocity rain cells (Section 3.1.1), the expected result is that any algorithm making use of this model for infilling will be more accurate at a high aggregation level (eg: 24-hour) as this assumption is more readily satisfied. Furthermore, while the results derived are acceptable at a 24-hour level, it is evident that there are still problems with the algorithm that will only be exacerbated at a lower aggregation level (Section 5.4.1). Therefore, no attempt is made to use the current infilling algorithms as methods for infilling the 1-hour record.

Requirements

In order to demonstrate that an infilling algorithm is acceptable for missing data, the algorithm must first accurately predict data which are known. This prediction is made more difficult in that missing values abound in the 24-hour historical record (see Table 4.1). As a result, it is likely that infilling solely the missing data is easier as more information about rainfall in the region is available. Nevertheless, if a method cannot accurately predict what is already known, then it cannot be trusted to sensibly predict that which is unknown.

A stratified random sample of 20% of the valid historical data for each site and season, was infilled and the results compared with the known records. This sampling is repeated 100 times so that a reasonable evaluation of the algorithms behaviour can be made - especially with regard to flaws in the methodology.

The algorithms are analysed based on three criteria as discussed in Section 3.3. Firstly, the raw infilled and historical records are compared both directly and via quantile-quantile plots. Secondly, the mix of wet and dry sites in the region are

checked for equivalence as this was an area in which the spatial-temporal NSRP model was not matching correctly (Section 5.4.1). Finally, the pooled statistics and the distribution of these statistics across the 100 samples are examined. Note that these plots and/or tests may not be produced if it is already known that the algorithm will not produce the required results.

5.5.2 Best fit least squares

A simple algorithm to select a suitable best fitting row in the simulation record is to minimise the sum of squares between the observed historical data and the simulation records (Figure 5.19). If any values in the historical data are missing then they are imputed with the values from the best fitting row. As the algorithm is fitted to the 24-hour record, only the points from the current time point are likely to be needed within the fitting function.

Analysis

The analysis of the best fit least squares (BFLS) infilling algorithm is broken down into two sections. The first section analyses the accuracy of the infilled data as it relates to known historical data both directly (intensity plots), from a distributional perspective (quantile-quantile plots), and spatially (proportion of dry, mixed wet/dry, and wet sites) in the *infilled* region. Although high accuracy of the algorithm is desirable, the distribution of the infilled data must match the distribution of the historical data for any inference to be reliable. Therefore, in the second section, the pooled statistics are compared between the infilled and historical estimates. Again, from the perspective of inference, the behaviour of the statistics must match between the infilled and historical records. Furthermore, the distribution of the pooled statistics should be equivalent between the two datasets if the infilling algorithm is optimal.

for all seasons, s
{
    let H be the historical records during s,
    S be the 300-month simulation record for s,
    t be the historical time,
    i be the simulation time,
    j be a site in the region, and
    n be the total number of sites in the region
    for all rows in H containing missing data
    {
        if (the entire row is missing)
        {
            randomly select a row, i, from S
        } else {
            set S = $h(S)$;
                where h is some heuristical function (Section 5.4)
            minimise
                $SS_i = \sum_{j=1}^{n} I_j * (H_{t,j} - S_{i,j})^2$

                where I_j is 1 if $H_{t,j}$ is a valid point and 0 otherwise
            select the record with index, i, given by $min(SS_i)$ [$*$]
        }
        for all missing records at H_t set:
            $H_{t,j} = S_{i,j}$; where j is the site index
    }
}
[$*$] If a fit of 0 is found then the algorithm terminates with this index, i, with 50% probability and does not continue to examine other fits.

Figure 5.19: Best fit least squares: algorithm

Infilling accuracy

An intensity plot of error versus historical estimates was constructed for each season based on the 100 samples obtained from the infilling algorithm runs. The results for January and July are shown in Figure 5.20. Note that the lower diagonal line corresponds to the lower error bound - that is error given the predicted value was $0mm$. For the purpose of clarity and interpretability the error plotted is restricted to $(-50mm, 50mm)$ and the historical range restricted to $(0mm, 70mm)$. These plot dimensions are maintained regardless of season (and infilling algorithm) so comparisons between seasons and algorithms can easily be made.

As expected, the dry rainfall values are well predicted as their occurrence is very high. The ability to predict wet $(+ve)$ amounts of rainfall deteriorates as the magnitude increases, however, the rate of deterioration varies considerably with the season. The winter months (Figures D.1 and D.6) indicate that, while far from perfect, the algorithm performs reasonably well for the first $20mm$ then begins to underestimate the total rainfall with increasing regularity. The summer months (Figures D.3 to D.4) illustrate the problems with the variation of rainfall intensity and the difficulty of predicting when intense precipitation occurs.

Note also that the characteristics of the BFLS algorithm are such that there is little scatter in the selection of infilled points (Figure 5.20). This is due to the selection of the exact best fit from the simulation data. Some scatter is present, however, as when a fit of 0 is found the algorithm terminates with 50% probability at that point (Figure 5.19).

While infilling accuracy is important, for the purpose of extreme value frequency analysis, it is even more imperative that the distribution of the rainfall between the historical values and the infilled values correspond. Therefore, quantile-quantile plots (eg: Figure 5.21) of each season have been plotted comparing the distribution of the infilled rainfall and the true historical records based on the 100 samples of the infilling runs. Where, rather than plotting 100 lines on the plot, the median of the infilled quantiles for the 100 samples was plotted against the corresponding historical quantile along with a vertical error bar corresponding to the interquartile range. The maximum and minimum quantiles observed are also plotted as dots.

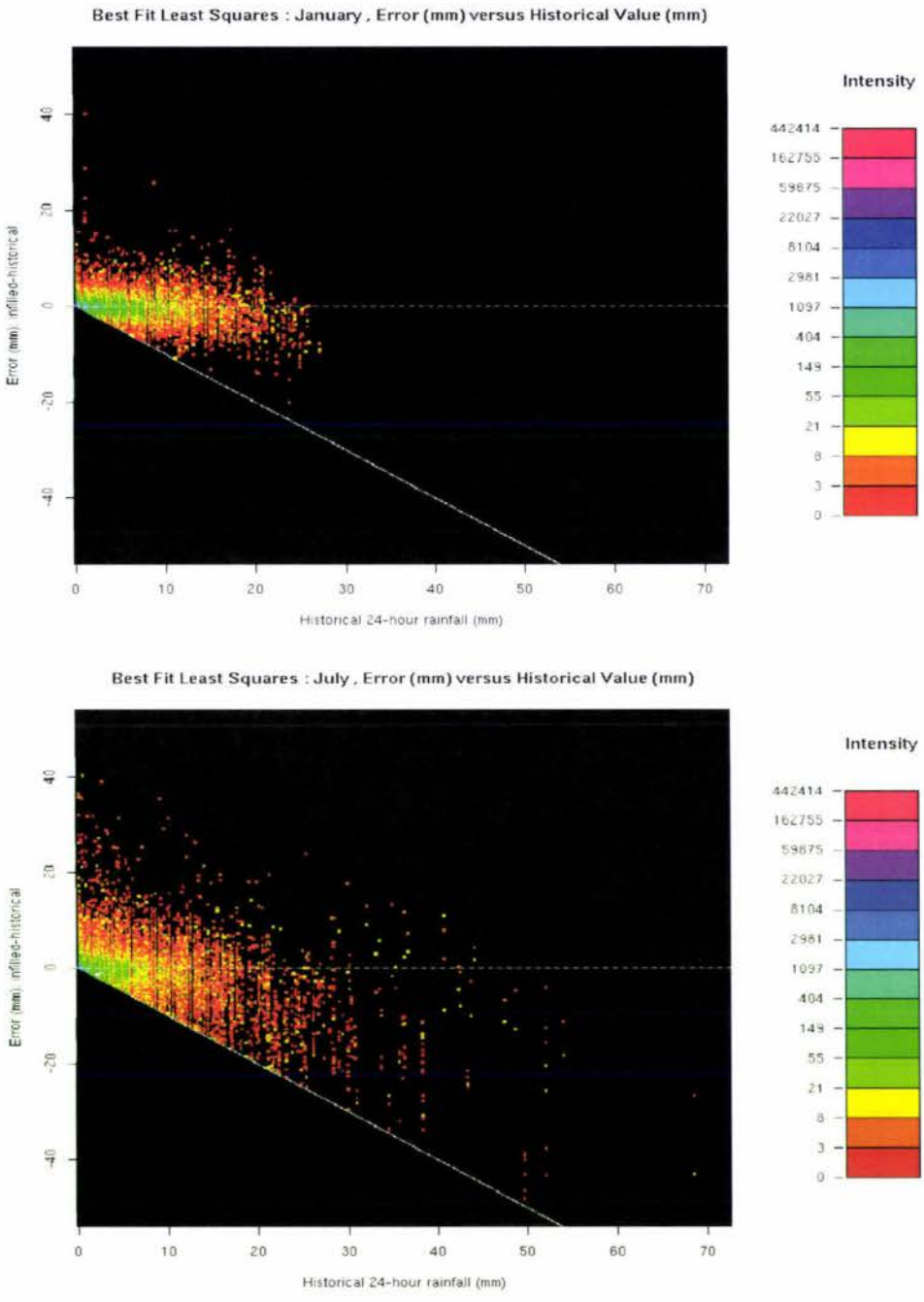
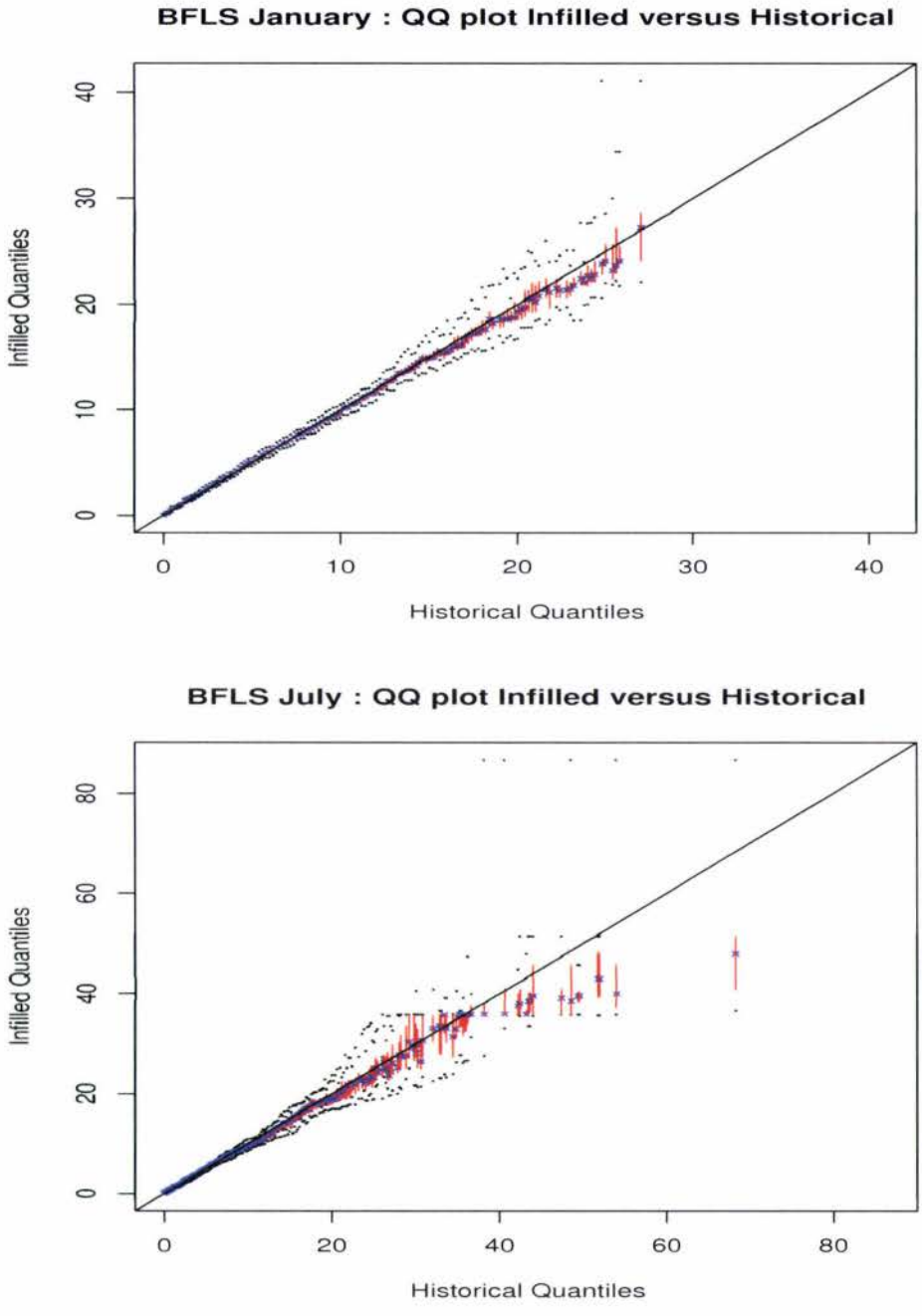


Figure 5.20: BFLS intensity plots: January and July

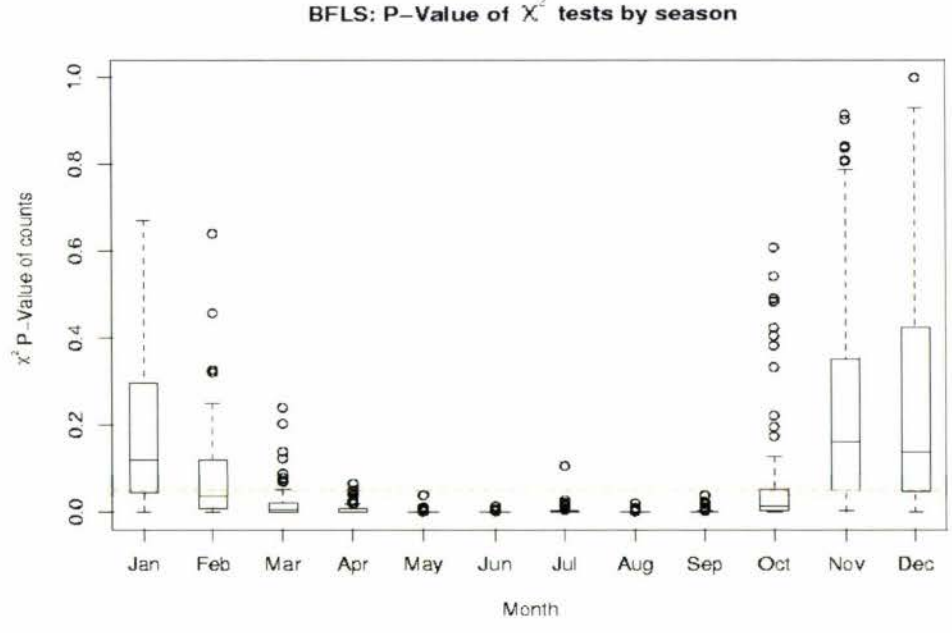


cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure 5.21: BFLS regional QQ plots: January and July

The plots for January and July (Figure 5.21) indicate that overall the distribution of the infilled and historical correspond reasonably well. There is a tendency, for both months, for the quantiles of the historical data to be underestimated at the tail end of the distribution - that is, the infilling is avoiding fitting extreme points. The other months (Figures D.7 to D.12), especially the summer months, also show a similar problem to some degree.

The results of the χ^2 tests for the record counts of all dry, some dry, and all wet for the infilled and historical records are shown in Figure 5.22. The tests show that only the winter months (November through February) do not always reject (at the 5% threshold) the null hypothesis that the proportion of counts in the historical and infilled record are the same. For all other months, nearly all of the test results were significant at the 5% level (marked on the plot). Clearly, the best fit least squares algorithm has maintained the characteristics of the simulation in regard to the excessive infilling of either all wet or all dry rows.



5% significance level = dashed line

Figure 5.22: BFLS χ^2 tests: infilled versus historical

Pooled statistics

The previous section highlighted the problem with the infilling algorithm underestimating the historical records as the true historical record increased. The overall impact on the pooled statistics is not immediately obvious, as the error of the infilling algorithm is bounded by the minimum amount of rainfall being 0mm. That is, overestimation of the true value is unbounded, but underestimation is limited by the true value. However, as likelihood of extreme values is consistently underestimated, this underestimation will have a significant effect on the skewness and CV. As the cross-correlation was maintained between the historical and the simulated records (Section 5.3.1), the cross-correlation is expected to match regardless of the accuracy of the infilling algorithm due to the missing data being replaced from a single row.

The pooled statistics used to fit the data (Section 5.2.2) were calculated both for the true historical data and for the infilled data. These statistics were then plotted in a scatter plot (Figure 5.23). As suspected, the underestimation of the historical data has resulted in a general underestimation of the CV and skewness. The autocorrelation, on the other hand, is usually slightly overestimated. The first point is of more import and indicates the extent of the problem with the infilling algorithm. Not only is the algorithm not sufficiently accurate, but the pooled statistics are not maintained between the historical and infilled records.

The distributions of the samples - infilled and historical - are not equivalent (Figure 5.24). This result was expected given the problem with the continual underestimation of the raw data and the pooled statistics. The cross-correlation distributions (Figure 5.25 and CD: Figures D.13 to D.18) shows that generally the cross-correlation is matched reasonably well across the seasons, however, there is a tendency to underestimate the low cross-correlations.

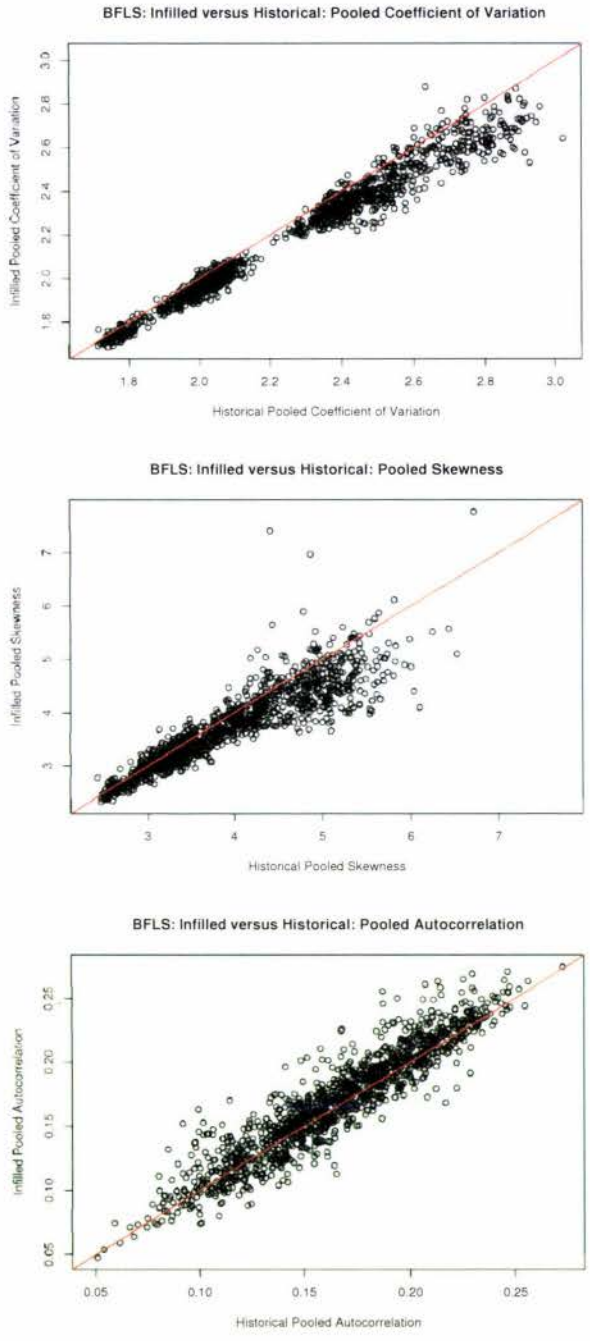


Figure 5.23: BFLS pooled statistics: CV, skew, and acor

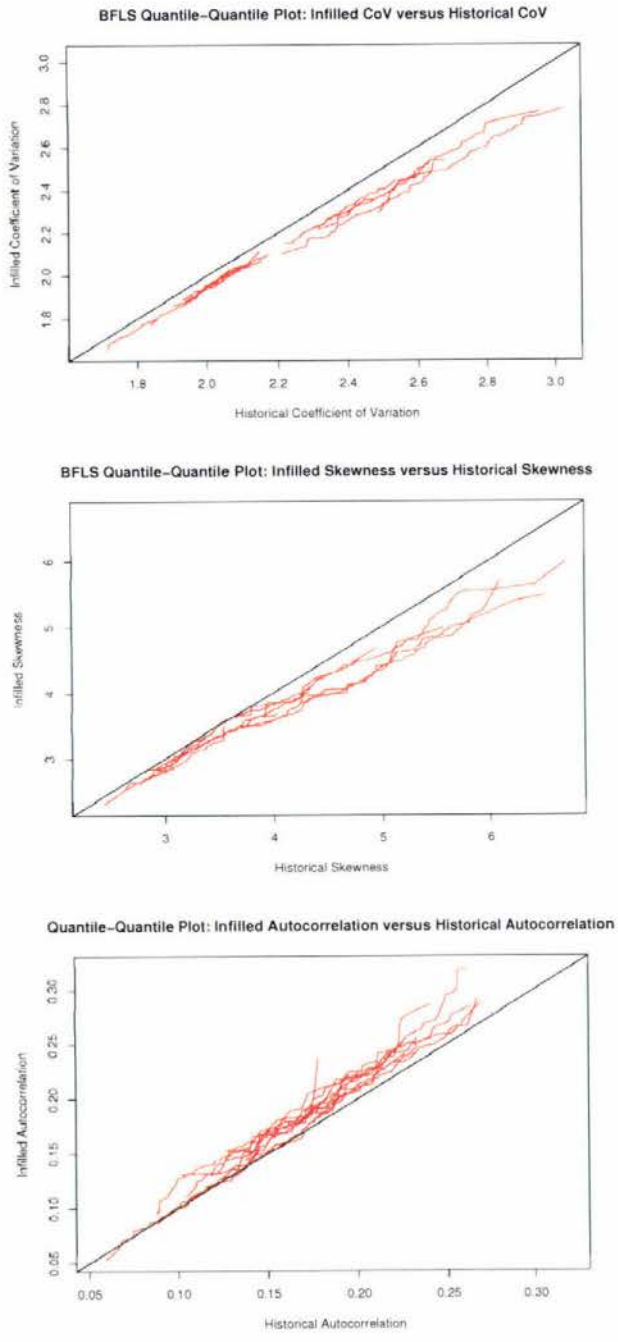


Figure 5.24: BFLS pooled statistics QQ plots: CV, skew, and acor

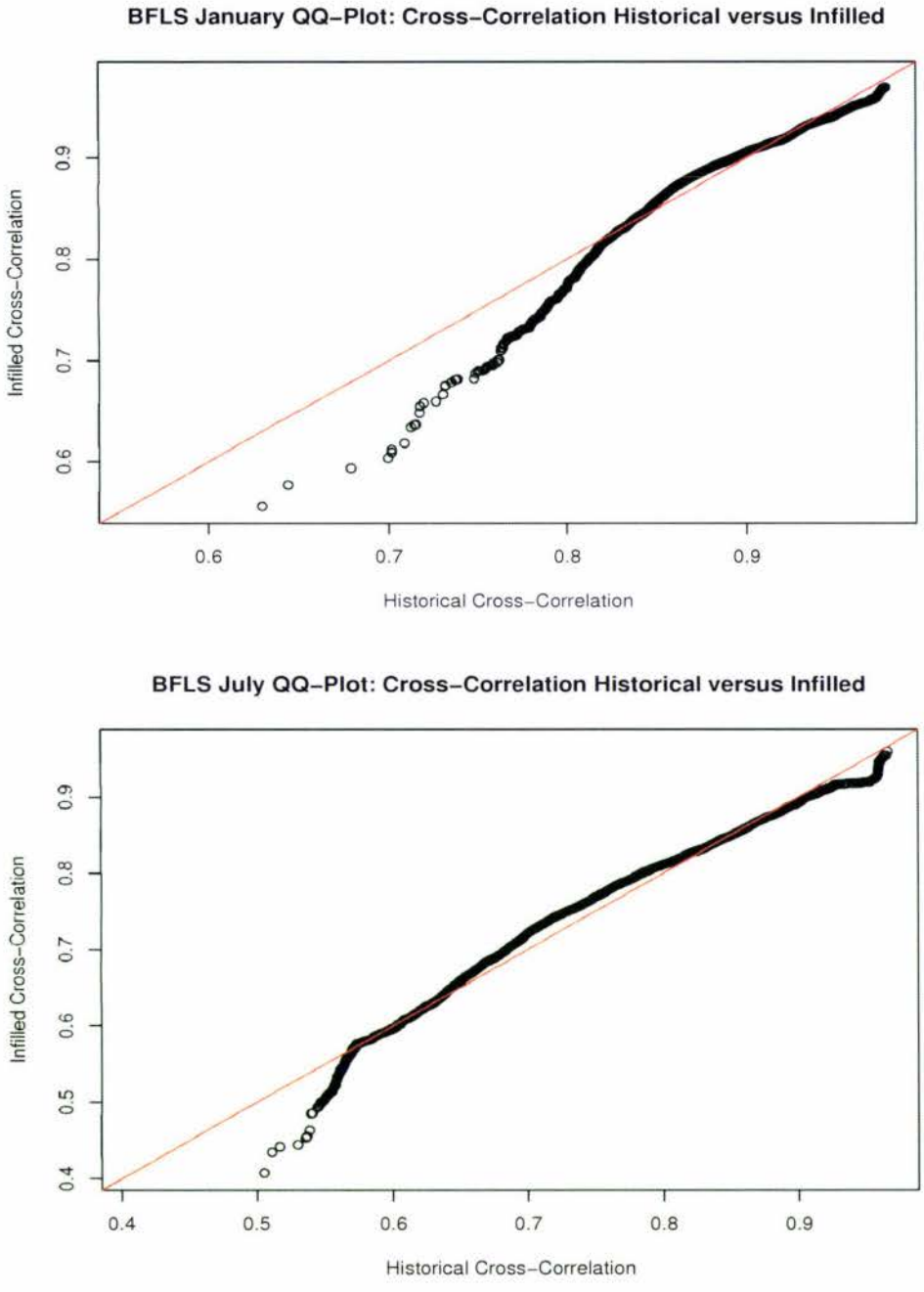


Figure 5.25: BFLS cross-correlation QQ plots: January and July

Summary

The best fit least squares algorithm has been shown to be deficient both for infilling accuracy and for the maintenance of the pooled statistics over the region being infilled. The historical records are usually underestimated especially as the true values increases. This latter point is independent of season although the apparent severity of the problem may vary.

It is known that the model is not producing the correct characteristics for the rainfall distribution across the region (Section 5.4.1). Furthermore, the results in Section 5.5.2, have shown that the infilling algorithm, as it uses the row of best fit, also is deficient in this regard. However, the problem demonstrated in Section 5.5.2 with the underestimation of the tail of the historical quantiles is easier to address by adjusting the least squares matching. Therefore, this is the first issue addressed (Section 5.5.3), whereas the incorrect regional characteristics are postponed till Section 5.5.4.

5.5.3 Best fit CDF least squares

The next fitting algorithm presented only differs from the BFLS algorithm in that, instead of using the raw data as input into the least squares minimisation, a transformed value is used. In particular, the $CDF(X_i)$, where the CDF is calculated from the simulated record for each site, i, for each month. Thus all data inputted into the least squares minimisation are transformed to an interval of $[0, 1]$ regardless of the scale of the raw rainfall data. This is the only difference implemented within this thesis between this algorithm and the previous best fit algorithm (Section 5.5.2). Henceforth, this algorithm is abbreviated to BFCDF where, if not given, it is assumed that this is fitted by least squares.

This algorithm was derived as an alternative to the least squares fitting as it was noted that one disadvantage of the ordinary least squares fitting on the raw data is that the distribution of the data is expected to be roughly symmetrical about the point being fitted. Additionally, and more importantly, extreme distributions with ($x >= 0$) require weighting to ensure that more weight is given to the correctness

of the smaller values than the larger ones. These two issues, if left unresolved, bias the fitting toward less extreme points. This bias is seen through the problems that have been seen with the algorithm - particularly the under fitting of skewness and variance. Furthermore, the frequency of infilled extreme points will not be correct and, as such, the algorithm will fail to achieve its purpose - to accurately infill missing records so that decisions can be made regarding the occurrence and distribution of extreme rainfall events.

One such function satisfying at least the requirement of equal weighting regardless of magnitude is the $CDF(X)$. The CDF takes all possible inputs of X and transforms them on to the scale $[0, 1]$ where elements within the interval are uniformly distributed. The transformed values are symmetric about 0.5 with variance $1/12$.

In addition, the difference between $CDF(X) - CDF(Y)$ where X is the historical data and Y is the simulated data is also symmetrical with the probability function (Equation 5.5) from Weisstein (2003b).

$$\begin{aligned} P_{X_1 - X_2}(u) &= \int_0^1 \int_0^1 \delta((x - y) - u) dx\, dy \\ &= 1 - u + 2u\, H(-u), \end{aligned} \tag{5.5}$$

where X_1, X_2 distributed Uniform $[0, 1]$, δ is a delta function, and $H(x)$ is a Heaviside step function (Equation 5.6 as obtained from Weisstein (2003a)). This distribution is expected to have a mean of 0 and a variance of $1/6$.

$$H(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x = 0 \\ 1, & x > 0 \end{cases} \tag{5.6}$$

In this case, the probability of observing a discrepancy is given by Equation 5.5. Therefore, a further infilling heuristic would be to cease evaluation of a record if the probability of an observed discrepancy between the two records was too large. This could either be a strict condition (eg: records with a probability of observance of less than 0.025 are discounted), or could be applied with some probability, p.

If the fitting is computed via the CDF, then the second point raised is resolved, and the asymmetry of the extreme value distribution is then correctly handled. The simulated records can be used to obtain the sample CDF as an estimate of the population CDF - provided the length of the simulation is sufficient. In this example, the simulation period of 300 months is assumed to be adequate for an approximation to the population CDF.

for all seasons, s
{
  let H be the historical records during s,
  S be the 300-month simulation record for s,
  t be the historical time,
  i be the simulation time,
  j be a site in the region, and
  n be the total number of sites in the region
  for all rows in H containing missing data
  {
    if (the entire row is missing)
    {
      randomly select a row, i, from S
    } else {
      set S = $h(S)$;
        where h is some heuristical function (Section 5.4)
      minimise
        $SS_i = \sum_{j=1}^{n} I_j * (CDF(H_{t,j}) - CDF(S_{i,j}))^2$

        where I_j is 1 if $H_{t,j}$ is a valid point and 0 otherwise
      select the record with index, i, given by $min(SS_i)$ [*]
    }
    for all missing records at H_t set:
      $H_{t,j} = S_{i,j}$; where j is the site index
  }
}
[*] If a fit of 0 is found then the algorithm terminates with this index, i, with 50% probability and does not continue to examine other fits.

Figure 5.26: Best fit CDF least squares: algorithm

It should be noted that the $CDF(X_i)$ was calculated for each site-month independently and, just as importantly, the calculations were stored in $0.1mm$ increments as this was the historical, 24-hour data's granularity. This storage method was used for two reasons.

Firstly, as the historical data does not distinguish between records of a granularity of less than $0.1mm$, realistically, the infilling algorithm should not either when selecting records from the simulated data. If such records are distinguished between, then the selection algorithm automatically becomes biased toward values that are close to the $0.1mm$ boundaries ($0.0mm$, $0.1mm$, $0.2mm$, ...). Since a CDF is applied, truncation was applied to the simulated records rather than rounding.

Secondly, the $0.1mm$ increments was convenient for precalculation of the CDF values. By precalculating the $CDF(X_i)$ for each site for all values of X_i; ($X = 0.0, 0.1, ..., \max(X_i)$) and storing these results in a look up table, there is no heavy penalty for using this method. Furthermore, it is not strictly necessary to compute the $CDF(X_i)$ up to $\max(X_i)$, although this was done for this implementation. Instead, some storage space could be saved by only computing the $CDF(X_i)$ up to some specified maximum - for example 0.999. As the algorithm was to run as quickly as possible, it was decided that it was better to 'waste' space so that the look-up table could be used directly without having to checking whether the X_i's were within the boundaries of the table. The amount of memory that could be saved, say $200Kb$ for 23 sites, is miniscule compared with the amount of memory used by the simulation record.

Analysis

The main change in the fitting results from the BFLS algorithm results (Section 5.5.2) is that fitting via the CDF allows more extreme predictions to be fitted more readily. As a result, it is expected that this algorithm will match up the quantiles better at the tail end of the distribution and the dropping away observed in the BFLS algorithm will be fixed. Note the problem seen with the incorrect

regional partitioning will not be fixed as all missing values are infilled from the same simulated record. Therefore, the results from the χ^2 test are not presented within this section as they are very similar to the BFLS output seen previously.

Infilling accuracy

The intensity plots of the error of the algorithm are, in general, little different from the results for the BFLS algorithm. The plots January and July (Figure 5.27) are nearly identical to those for the BFLS algorithm (Figure 5.20). As such, only the results for January and July have been included.

The distributional plots for the analysis are also quite similar (and therefore have not been included), with the exception of June and September (Figure 5.28). For June especially, it is evident that the BFCDF algorithm is definitely including extreme fits more often as the median of the quantiles is closer to the expected historical quantile.

Pooled statistics

The scatter plots for the pooled statistics (CV, skewness, and autocorrelation) are shown in Figure 5.29. The algorithm is still underestimating the pooled statistics, but it is also evident that an improvement has been made from the BFLS algorithm (Figure 5.23). The quantile-quantile plots of the distributions of the statistics (Figures 5.30) show a significant improvement over the previous plots (Figure 5.24) - once the changes of scale has been taken into account. As predicted, the most affected statistic is the skewness where the extreme overestimation is for June.

The quantile-quantile plots for the cross-correlation for January and July (Figure 5.31), shows a slight improvement for both these seasons over the BFLS plots (see also Figures D.19 to D.24). However, the quantile-quantile plots are still far from ideal. The spread of the cross-correlation is significantly more in the infilled record for the lower tails. This would be accounted for if 0's were fitted too readily within the infilled data.

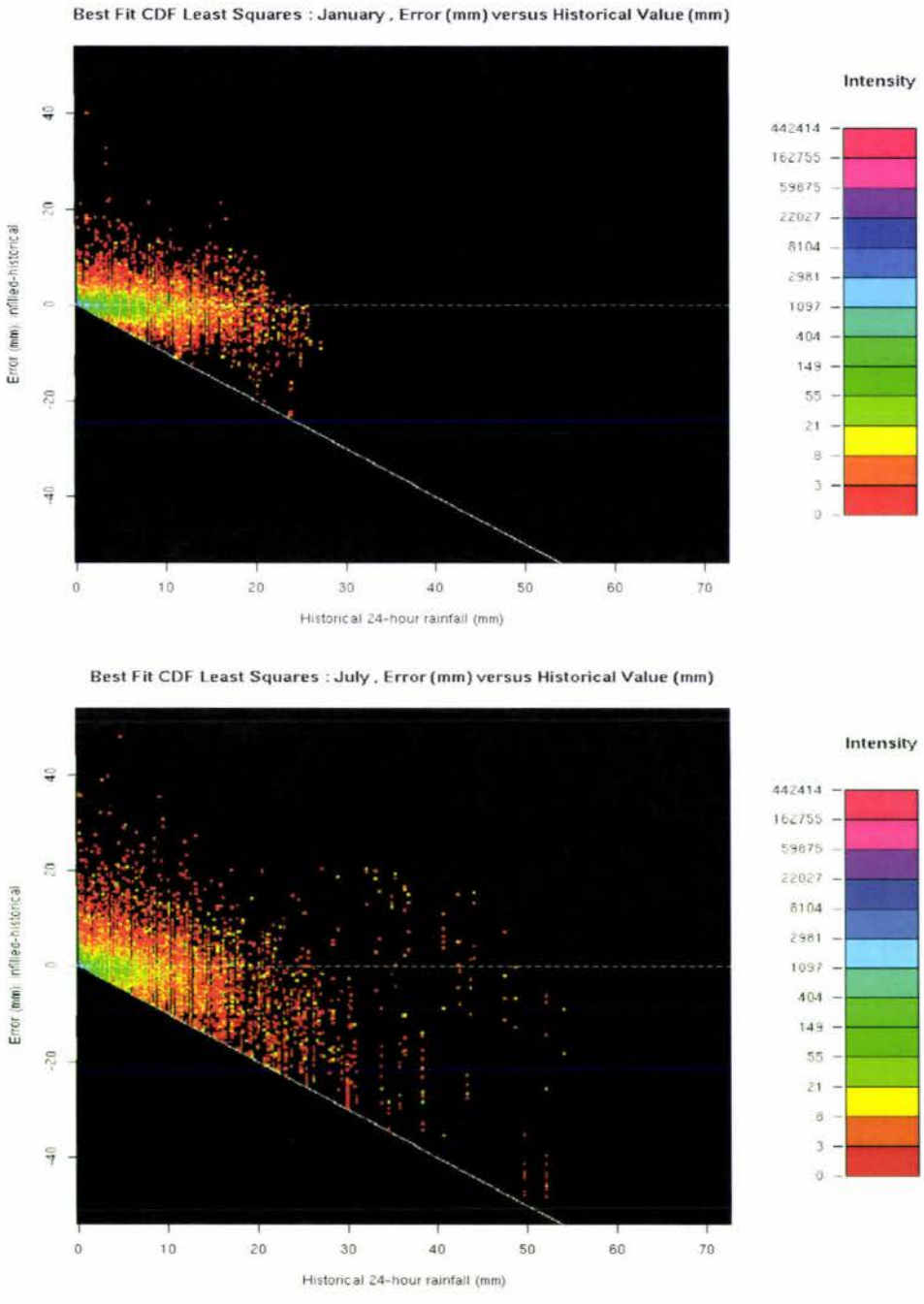
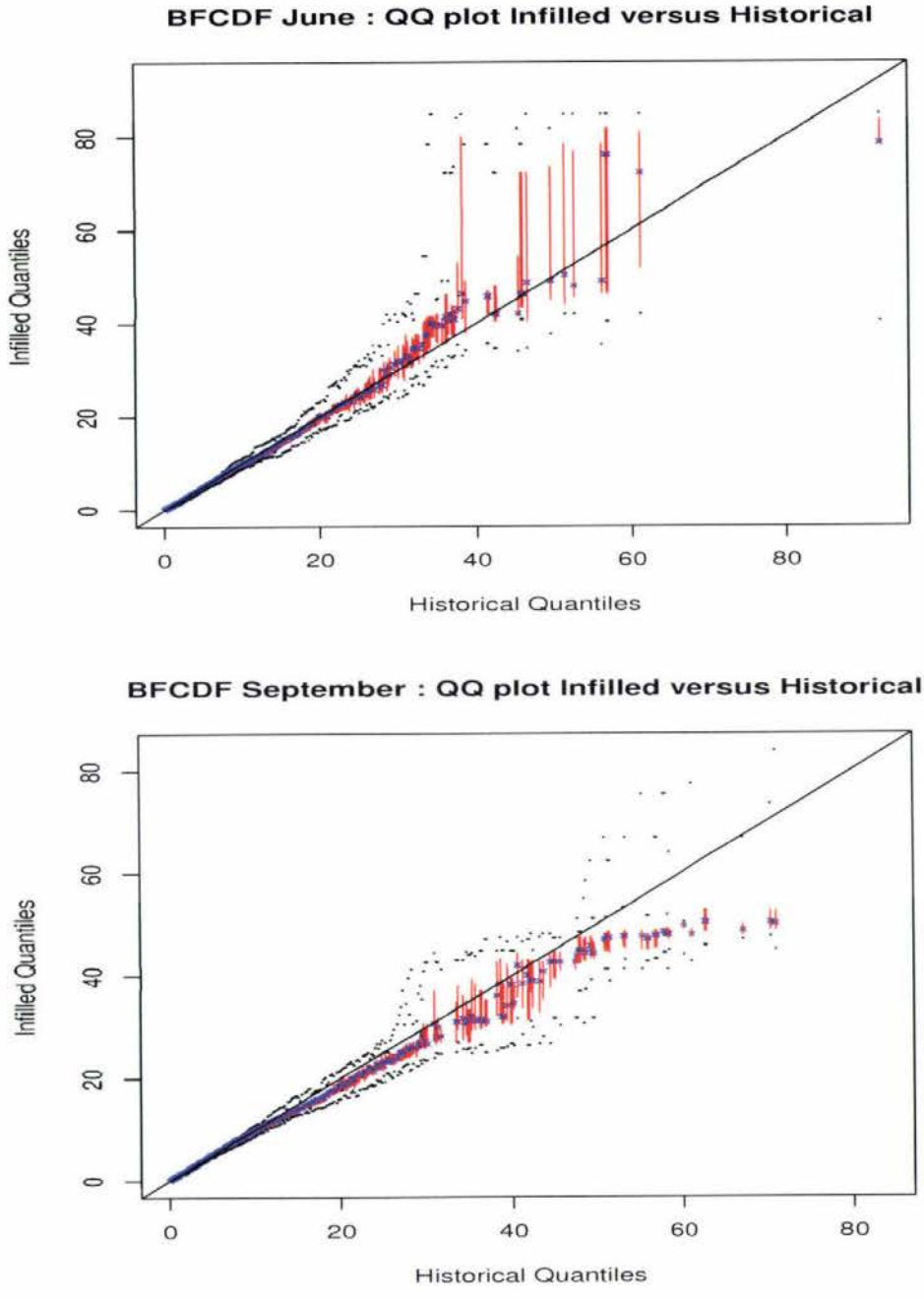


Figure 5.27: BFCDF intensity plots: January and July



cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure 5.28: BFCDF regional QQ plots: June and September

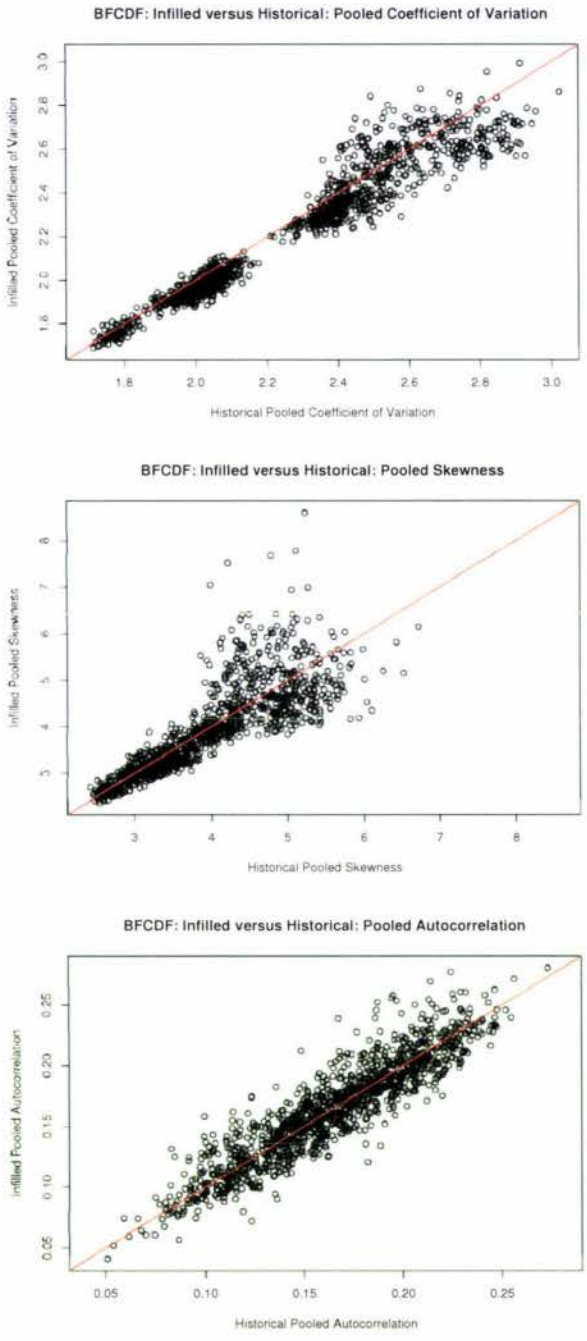


Figure 5.29: BFCDF pooled statistics: CV, skew, and acor

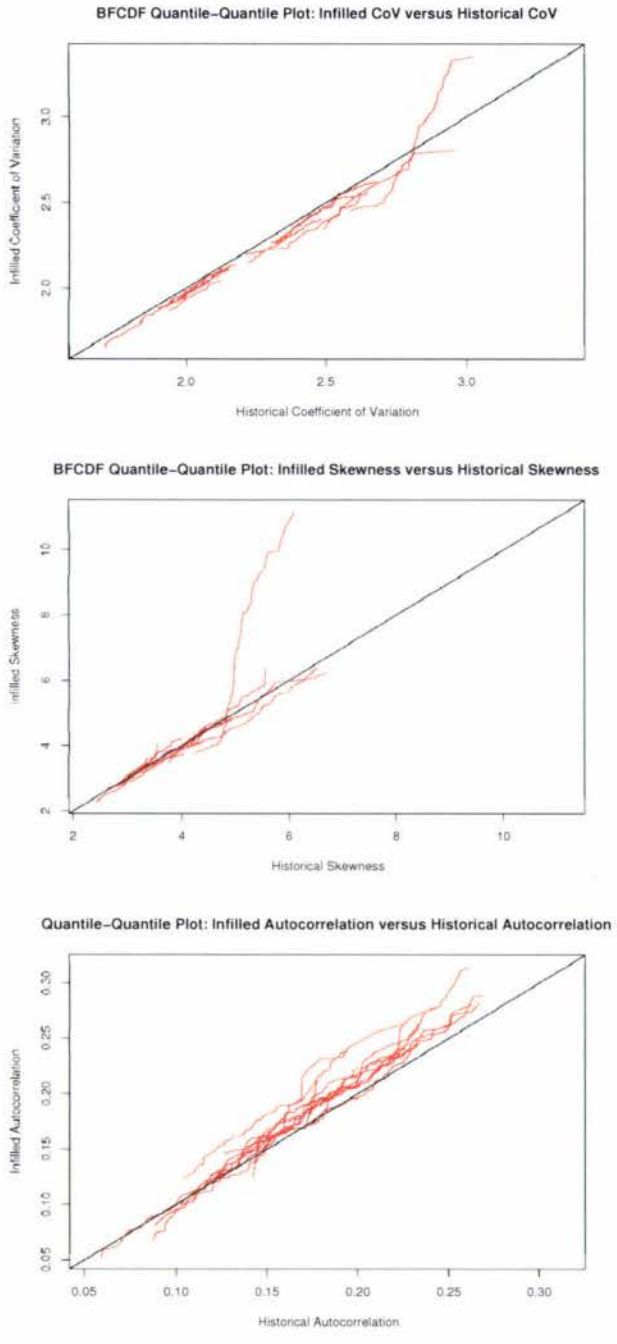


Figure 5.30: BFCDF pooled statistics QQ plots: CV, skew, and acor

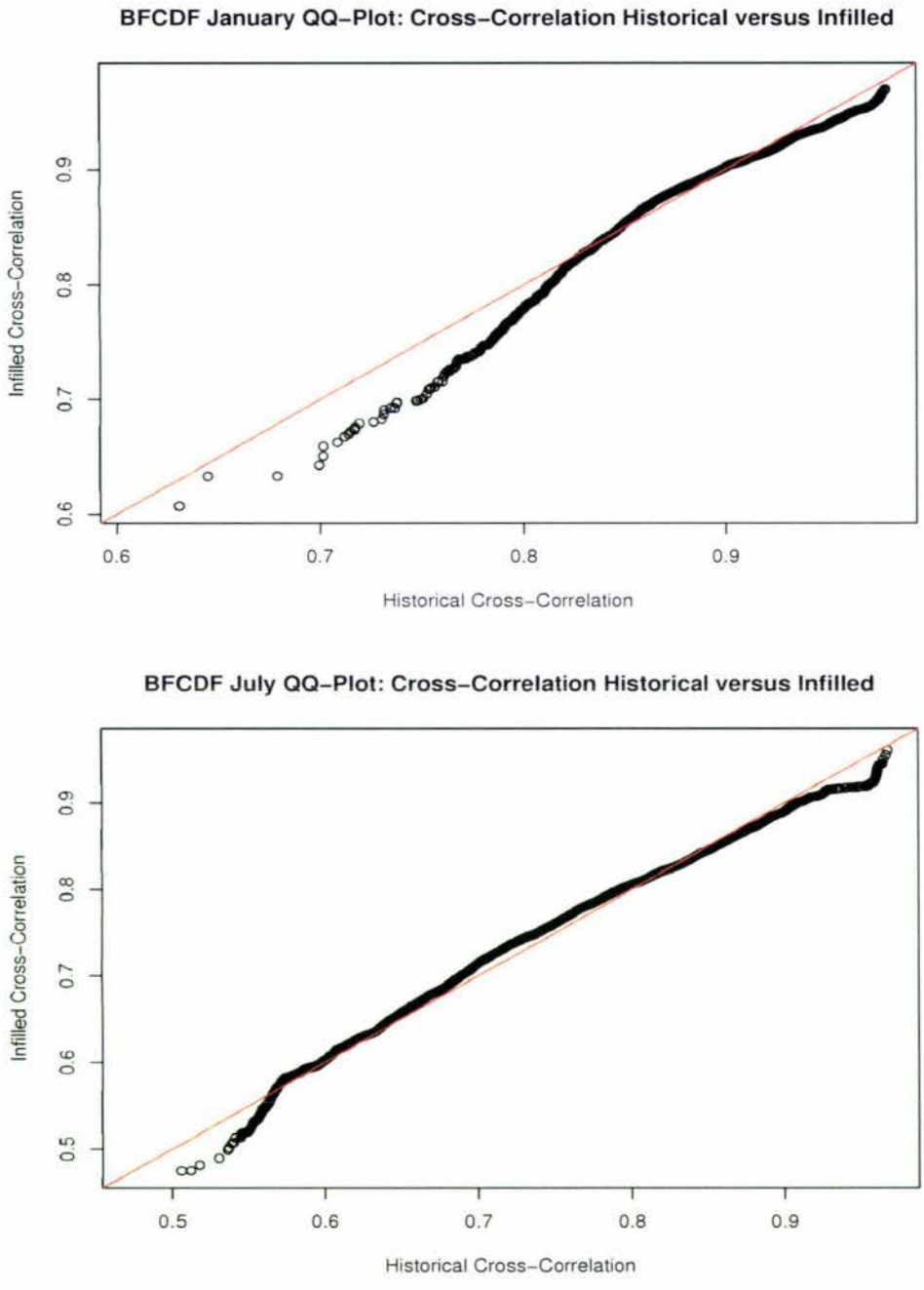


Figure 5.31: BFCDF cross-correlation QQ plots: January and July

Summary

Overall, the use of the $CDF(X_i)$ within the fitting procedure has not improved the algorithm as much as expected. There are still significant problems with under estimation of the pooled statistics, and the regional analysis of the infilled and historical data have generally not been improved. However, the use of the CDF has resulted in a more frequent inclusion of *extreme* fits - especially noted with June. This is beneficial as the algorithm must be able to infill using these values, therefore, the use of the CDF is retained in the next infilling algorithm analysed.

5.5.4 Iterative sampling CDF least squares

This last algorithm discussed uses **single imputation** from the synthetic record by sampling values for infilling from a set of best fitting rows (eg: the best 5%). The algorithm is not to be confused with multiple imputation methods which may also be referred to as iterative methods. The term *iterative* within this context refers to the iterative replacing of missing data at the sites at the same historical time point **not** iterative replacing of the same missing value (say until convergence).

The previous two algorithms used an expectation-maximisation technique in that the row of best fit from the synthetic data was used directly without further modification. This algorithm differs from the former two as infilled values are permitted (and indeed expected) to come from different rows in the synthetic record. The advantage of this is inherently apparent. If the model is not fitting the spatial characteristics of the data (see Section 5.4.1), then the effects of this discrepancy can be minimised by sampling from multiple rows. Therefore, the iterative sampling algorithm (ISCDF) was developed as listed in Figure 5.32.

The permutation of the fitting order (Figure 5.32) is conducted so the fitting results are not biased by the order in which missing values are replaced. The sampling of the best fits is always completed out of the best 5% of the fits. This threshold is not necessarily optimal, however, an optimal threshold is likely to be both data and simulation dependent. Thresholds at 1% and 10% were tried but neither adjustment improved the results of the infilling but instead greatly worsened the fit. In any case, as the algorithm worked such adjustments were not high priority.

```
for all seasons, s
{
    let   H be the historical records during s,
          S be the 300-month simulation record for s,
          t be the historical time,
          i be the simulation time,
          j be a site in the region,
          n be the total number of sites in the region, and
          C be the cut off percentage to sample from (eg: 5%)
    for all rows in H containing missing data {
          if (the entire row is missing) {
              randomly select a row, i, from S
          } else {
              let   P be a random permutation of the sites of missing records in H_t
              for all sites, j, in P {
                    set   S = h(S);
                          where h is some heuristical function (Section 5.4)
                    minimise
                          SS_i = sum_{j=1}^{n} I_j * (CDF(H_{t,j}) - CDF(S_{i,j}))^2

                          where I_j is 1 if H_{t,j} is a valid point and 0 otherwise
                    sort SS
                    let   N be the number of potential fits
                    generate a random integer, U, distributed Uniform(1, C * N)
                    set   i = simulation time of SS_U
                    set   H_{t,j} = S_{i,j}; where j is the site index
              }
          }
    }
}
```

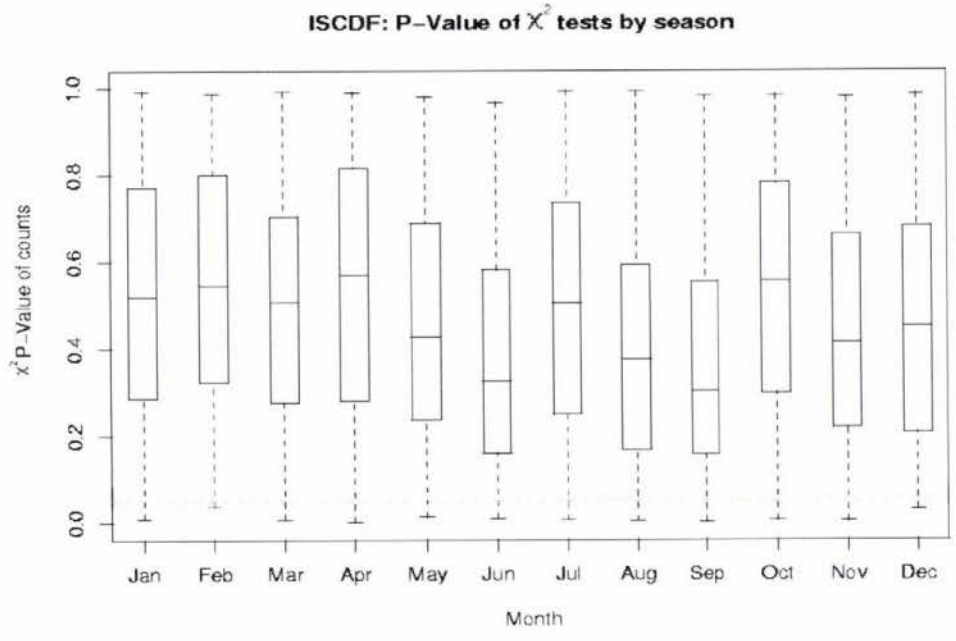
Figure 5.32: ISCDF least squares: algorithm

Analysis

Once again the analysis divides neatly into two sections dealing with the raw data and pooled statistics respectively. As expected, considerably more variation was seen in the intensity plots as, rather than just a few candidates for infilling, missing values could be infilled from multiple rows thus greatly widening the possible points used for infilling. Similarly, the quantile-quantile plots were expected to be closer to the historical regional distributions, and, if anything, to overestimate rather than underestimate the quantiles.

Infilling accuracy

The χ^2 test results for the wet/dry site counts are shown in Figure 5.33. It is evident that the sampling has fixed the problem with the incorrect site mixture at least at the levels examined. A deeper investigation could not be conducted using the current historical data, however, due to the sparseness of valid data within the historical record (Table 4.1). In any case, the infilling algorithm is a great improvement over the previous algorithms - at least in terms of the this test result.



5% significance level = dashed line

Figure 5.33: ISCDF χ^2 tests: infilled versus historical

The intensity plots for January and July (Figure 5.34 and Figures D.25 to D.30) show that there is more variation than seen in the earlier plots. It is evident that within the 100 samples, different points within the infilling algorithm are selected to infill the missing data. This implies that by sampling from the simulated record the sample size has effectively been increased while using the same amount of physical memory within the computer.

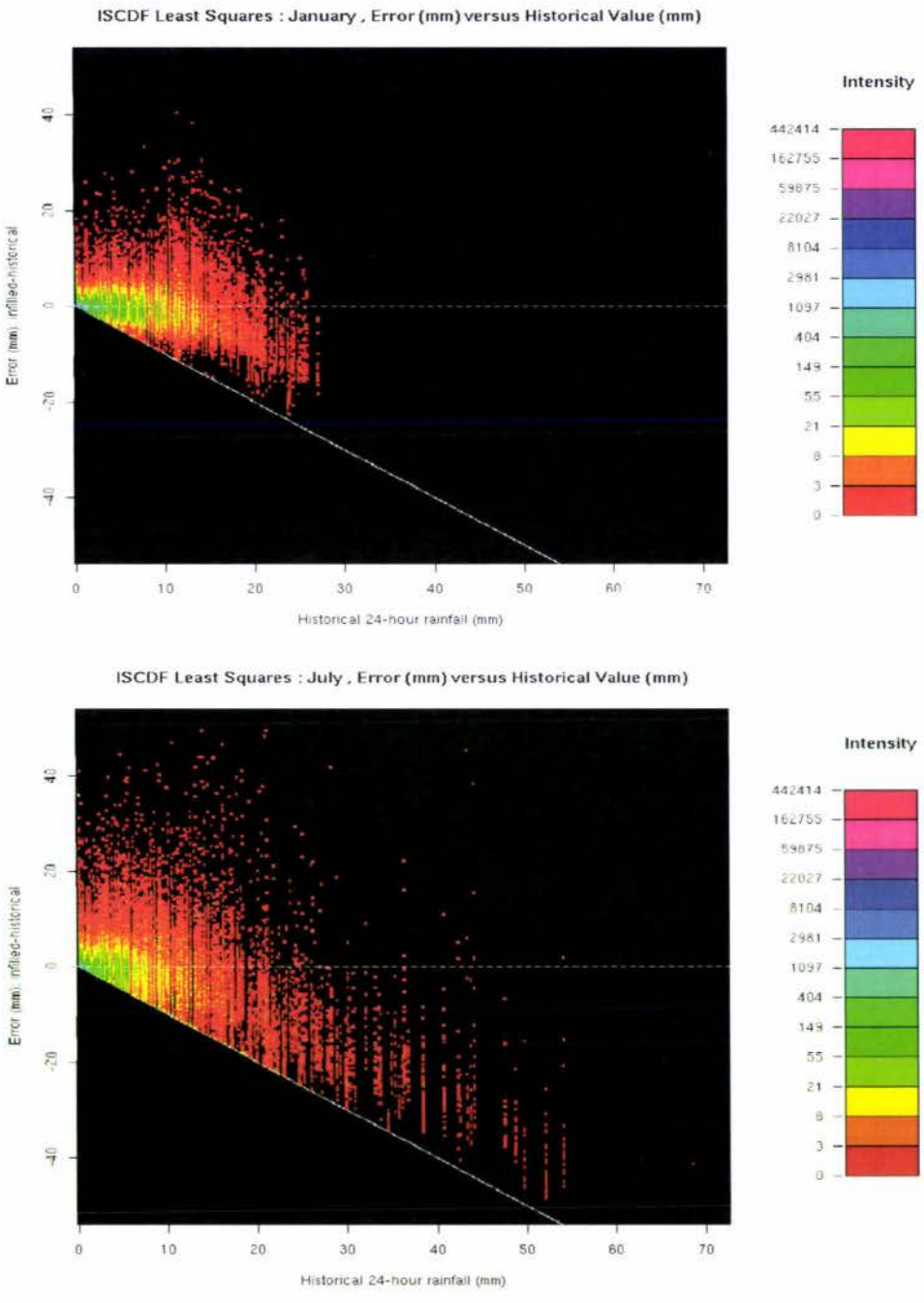
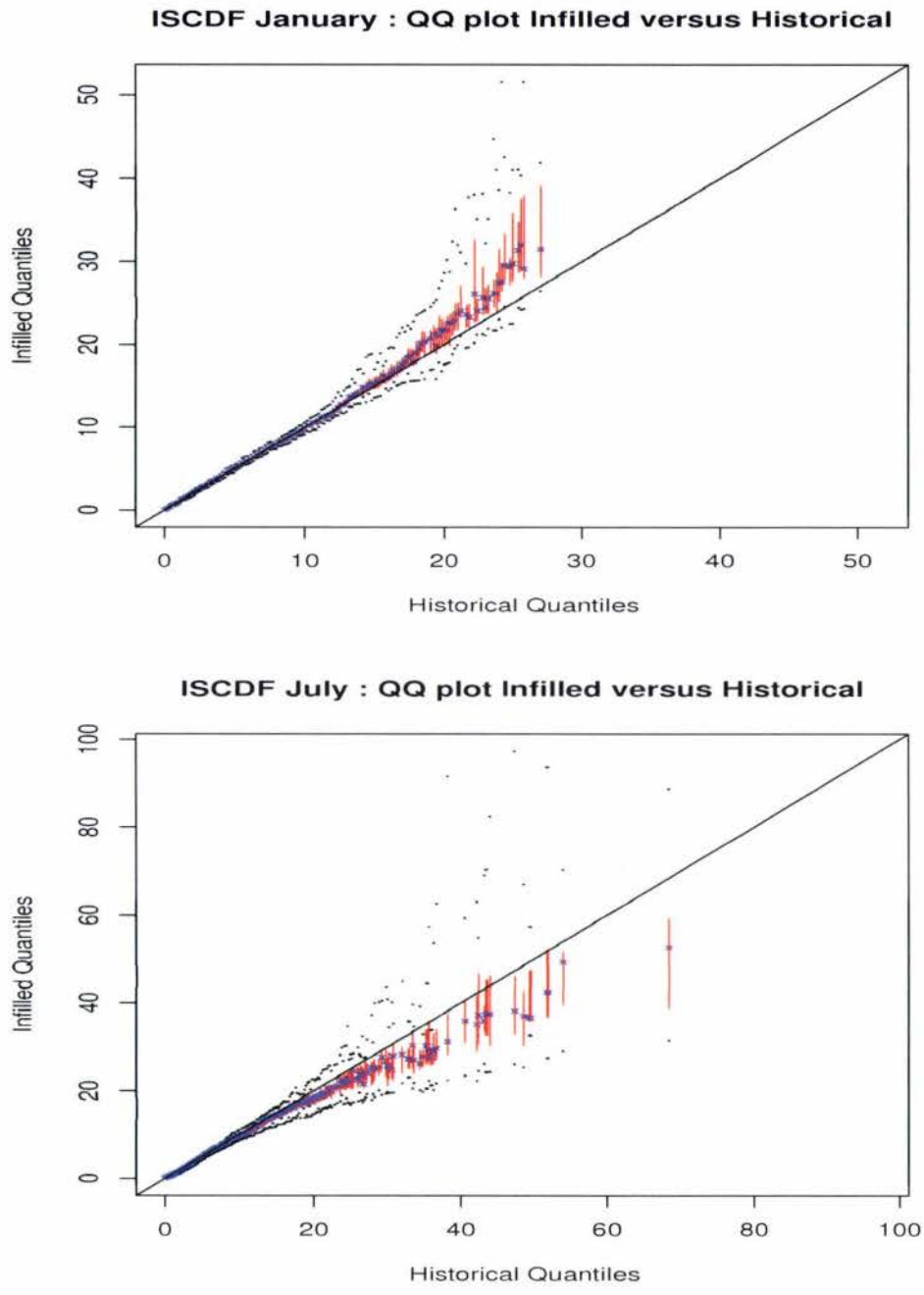


Figure 5.34: ISCDF intensity plots: January and July



cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure 5.35: ISCDF regional QQ plots: January and July

The quantile-quantile plots for January and July (Figure 5.35) show similar results to the BFLS algorithm. Note that January is slightly overestimated (as predicted) and, as such, lies more in accordance with the predicted smoothed estimates rather than the historical sample. There was no consistent pattern with over/underestimation of quantiles (Figures D.31 to D.36), however, underestimation was still more common.

Pooled statistics

The scatter plot of the pooled statistics (CV, skewness, and autocorrelation) in Figure 5.36, indicates that all the statistics have 'improved' over the previous distribution (see Section 5.5.5). Certainly there are more points above the expected line than previously for the CV and Skewness. As evident from the plot clustering, however, some of the months are still not being estimated well - this confirms some of the results from the quantile-quantile analysis of the raw results.

The quantile-quantile plots of the statistics (Figure 5.37) show that estimates of the skewness and CV tend to be overestimated compared to their historical counterparts. As before, the tendency is for either the whole month to be under estimated or overestimated. The main difference between this algorithm and the previous algorithms is that overestimation is now more common than underestimation.

The results for the cross-correlation are slightly worse than the other algorithms, for all months. In particular, the results for July do not lie on the expected quantile-quantile line at all (Figure 5.38). In general, the cross-correlation is under estimated (Figures D.37 to D.42). This is hardly a surprising result as any sampling method where results are taken from multiple source rows will tend to reduce the cross-correlation by default. Since the mixture of wet and dry days has been improved, it follows that the linear cross-correlation must have been worsened to some degree in order to get that improvement.

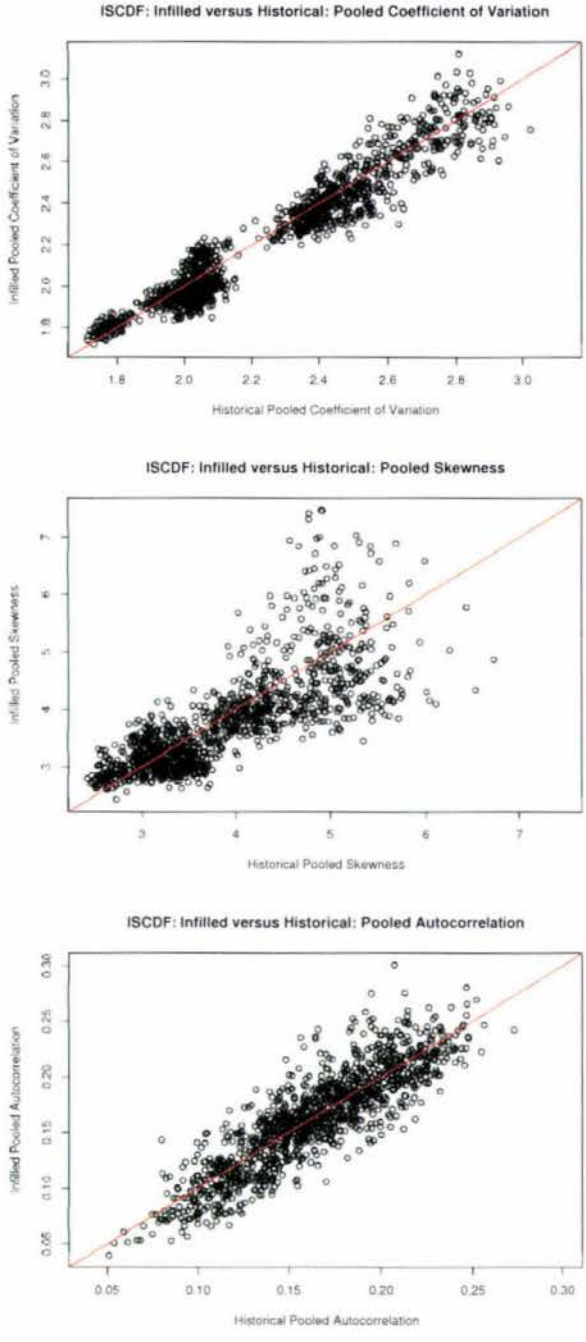


Figure 5.36: ISCDF pooled statistics: CV, skew, and acor

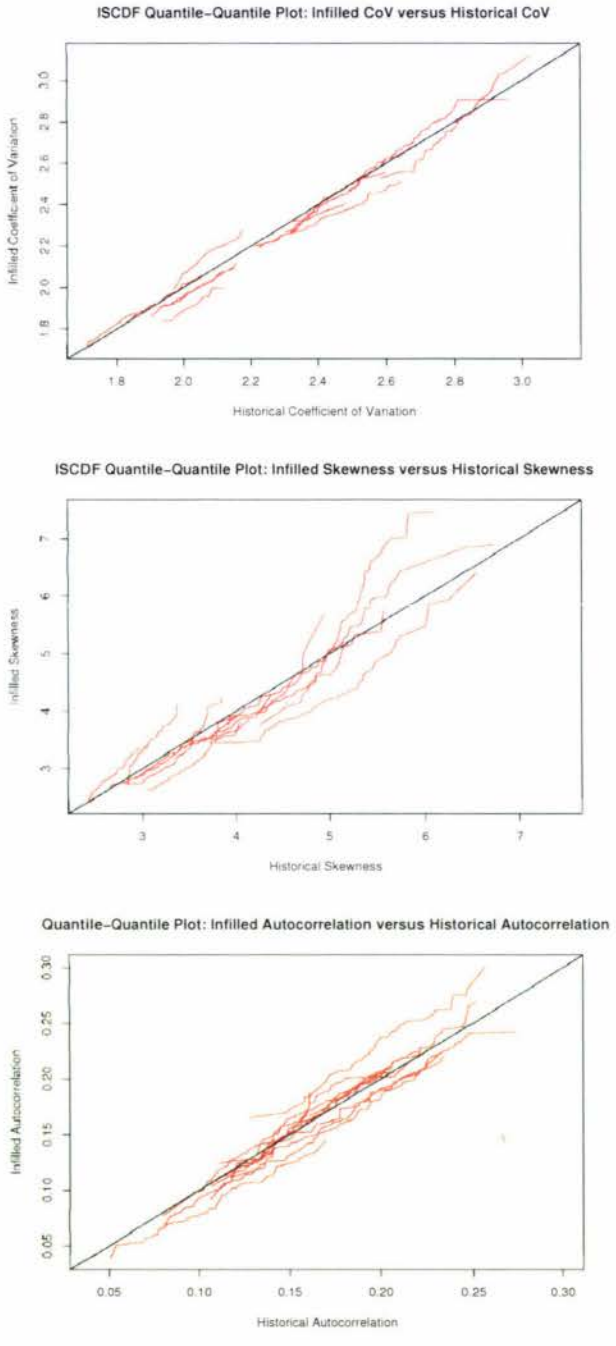


Figure 5.37: ISCDF pooled statistics QQ plots: CV, skew, and acor

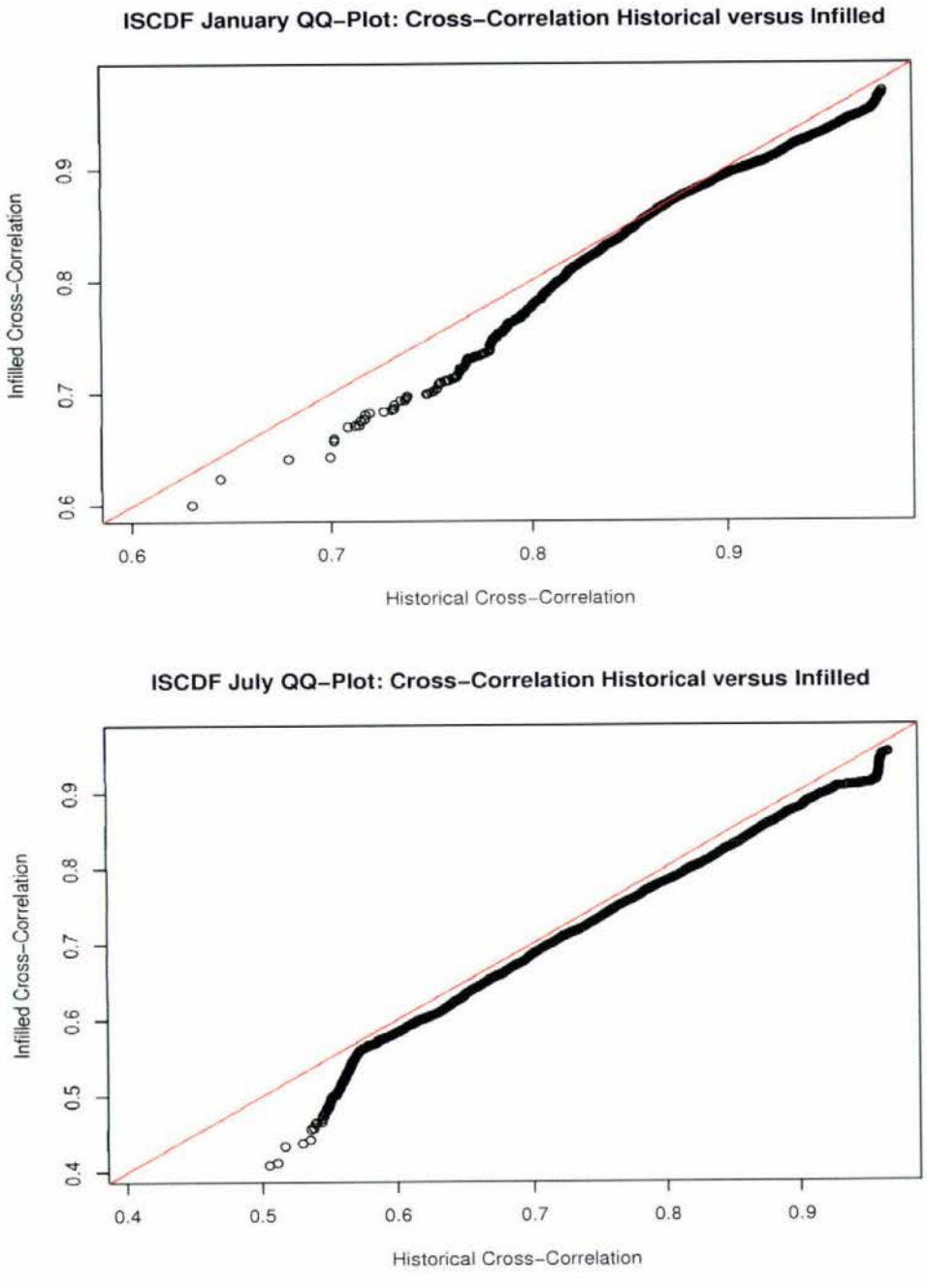


Figure 5.38: ISCDF cross-correlation QQ plots: January and July

5.5.5 *Comparison of algorithms*

Of the infilling algorithms considered, the iterative sampling algorithm (ISCDF) produces infilled records that are spatially more consistent with the historical record - in terms of the mixture of wet/dry sites (Section 5.5.4). However, it is also evident that the cross-correlation (see Section 5.5.4) has been reduced. Nevertheless, after taking into account the second observation, this infilling algorithm is definitely superior spatially.

Temporally, the results are not as clear. It was stated (Section 5.5.4) without numerical proof, that the ISCDF infilled statistics had a greater proportion of values above the historical estimates than either the BFLS algorithm or the BFCDFLS algorithm. A proportion test (z-test) was then conducted for each statistic by season (Table 5.13) comparing the ISCDF algorithm against the BFLS and the BFCDF algorithm. The proportion examined is the proportion of infilled statistics that are greater than their respective historical counterparts. A two-tailed test is conducted where the null hypothesis (H_0) is that the two proportions are equal.

Note that in this table (Table 5.13), the following notation is used.

Notataion in Table 5.13 and 5.14

H	= Historical
$p\{BF >= H\}$	= proportion of BFLS statistics >= H statistics
$p\{BFC >= H\}$	= proportion of BFCDF statistics >= H statistics
$p\{IS >= H\}$	= proportion of ISCDF statistics >= H statistics
P_A	= P-Value; $H_0 = p\{BF >= H\} == p\{IS >= H\}$
P_B	= P-Value; $H_0 = p\{BFC >= H\} == p\{IS >= H\}$

From Table 5.13, it is clear that the proportion of ISCDF statistics that are greater than the corresponding historical estimates is generally larger than the BFLS or BFCDF algorithms. The proportions BFLS versus ISCDF or BFCDF versus ISCDF are almost all significantly different. Note that all the infilling algorithms underestimate the autumn season (September to December).

From these results in Table 5.13, the tendency for the ISCDF algorithm in particular is for the historical estimates to be overestimated. Therefore, while

the algorithm gives a better spatial result, temporally one biased algorithm may just have been exchanged for another. Therefore, a Kolmogorov-Smirnov test is conducted to determine whether the bias is less for the ISCDF algorithm than the other two algorithms.

As the Kolmogorov-Smirnov test requires the two samples being compared to be independent, each distribution was separately sampled with repeats. Furthermore, the test was bootstrapped for each month 2000 times. The p-value for the test given in Table 5.14 is the median of p-values obtained from the bootstrap. Note that the first and third columns correspond to the median p-value for a one-sided test while the second and fourth columns correspond to the median p-value for a two-sided test.

The results in Table 5.14 show that, in general, the ISCDF algorithm is superior to the other two algorithms. Note, the blue results indicate that the ISCDF algorithm outperformed the alternative algorithm, a black result indicates neither algorithm is superior, whereas a red results implies the ISCDF algorithm was outperformed.

The only statistic where the ISCDF algorithm is less likely to be closer to the true historical statistics is for the coefficient of variation. For the skewness and autocorrelation, the ISCDF algorithm results are better (or indistinguishable) in all but one case (Skewness for June).

5.5.6 Other infilling derivations

Variations of the infilling algorithms discussed previously are presented briefly within this section. In all cases covered, the results did not improve on the base algorithm. The notation used within this section is as follows:

notation: H is historical records,
S is the 300-month simulation record,
t is the historical time,
i is the simulation time,
j is a site in the region,
n is the total number of sites in the region, and
I_j is 1 if the record $H_{t,j}$ is a valid record and 0 otherwise.

Season	$p\{BF>=H\}$	$p\{IS>=H\}$	P_A	$p\{BFC>=H\}$	$p\{IS>=H\}$	P_B
Coefficient of Variation						
Jan	0.08	0.89	0.000	0.20	0.89	0.000
Feb	0.05	0.21	0.002	0.07	0.21	0.008
Mar	0.06	0.81	0.000	0.25	0.81	0.000
Apr	0.15	0.99	0.000	0.25	0.99	0.000
May	0.03	0.46	0.000	0.10	0.46	0.000
Jun	0.07	0.55	0.000	0.80	0.55	0.000
Jul	0.22	0.55	0.000	0.46	0.55	0.258
Aug	0.05	0.52	0.000	0.08	0.52	0.000
Sept	0.06	0.10	0.434	0.09	0.10	1.000
Oct	0.01	0.19	0.000	0.03	0.19	0.001
Nov	0.02	0.10	0.037	0.00	0.10	0.004
Dec	0.02	0.00	0.477	0.06	0.00	0.038
Skewness						
Jan	0.23	0.82	0.000	0.44	0.82	0.000
Feb	0.55	0.24	0.000	0.44	0.24	0.005
Mar	0.26	0.40	0.051	0.40	0.40	1.000
Apr	0.10	0.97	0.000	0.35	0.97	0.000
May	0.26	0.27	1.000	0.38	0.27	0.131
Jun	0.13	0.35	0.001	0.82	0.35	0.000
Jul	0.33	0.41	0.305	0.66	0.41	0.001
Aug	0.21	0.57	0.000	0.20	0.57	0.000
Sept	0.10	0.07	0.612	0.19	0.07	0.021
Oct	0.09	0.36	0.000	0.27	0.36	0.223
Nov	0.15	0.30	0.018	0.09	0.30	0.000
Dec	0.10	0.04	0.166	0.29	0.04	0.000
Autocorrelation						
Jan	0.55	0.11	0.000	0.38	0.11	0.000
Feb	0.52	0.61	0.254	0.51	0.61	0.200
Mar	0.63	0.69	0.455	0.46	0.69	0.002
Apr	0.59	0.11	0.000	0.52	0.11	0.000
May	0.40	0.45	0.567	0.42	0.45	0.775
Jun	0.68	0.30	0.000	0.47	0.30	0.020
Jul	0.54	0.73	0.008	0.53	0.73	0.005
Aug	0.70	0.33	0.000	0.49	0.33	0.031
Sept	0.70	0.92	0.000	0.80	0.92	0.025
Oct	0.83	0.52	0.000	0.46	0.52	0.479
Nov	0.45	0.01	0.000	0.34	0.01	0.000
Dec	0.87	0.67	0.001	0.26	0.67	0.000

Table 5.13: ISCDF, BFLS, BFCDF: proportion test on the overestimation of historical, H, statistics

	BF vs IS		BFC vs IS	
$H_0 =$	$\{\lvert BF - H\rvert < \lvert IS - H\rvert\}$	$\{\lvert BF - H\rvert = \lvert IS - H\rvert\}$	$\{\lvert BFC - H\rvert < \lvert IS - H\rvert\}$	$\{\lvert BFC - H\rvert = \lvert IS - H\rvert\}$
Season	Coefficient of Variation			
Jan	0.002	0.004	0.000	0.000
Feb	0.018	0.037	0.298	0.281
Mar	0.852	0.037	0.001	0.001
Apr	0.000	0.000	0.000	0.000
May	0.914	0.000	0.779	0.016
Jun	0.990	0.000	0.852	0.024
Jul	0.018	0.037	0.008	0.016
Aug	0.961	0.000	0.990	0.000
Sep	0.613	0.111	0.613	0.054
Oct	0.852	0.000	0.779	0.010
Nov	0.237	0.037	0.613	0.010
Dec	0.000	0.000	0.000	0.000
	Skewness			
Jan	0.000	0.000	0.000	0.000
Feb	0.000	0.000	0.001	0.001
Mar	0.027	0.054	0.005	0.016
Apr	0.000	0.000	0.000	0.000
May	0.000	0.000	0.000	0.000
Jun	0.445	0.281	0.961	0.000
Jul	0.001	0.001	0.105	0.211
Aug	0.002	0.004	0.056	0.054
Sep	0.000	0.000	0.000	0.000
Oct	0.039	0.054	0.039	0.078
Nov	0.018	0.024	0.077	0.054
Dec	0.000	0.000	0.000	0.000
	Autocorrelation			
Jan	0.000	0.000	0.000	0.000
Feb	0.005	0.016	0.056	0.155
Mar	0.000	0.001	0.005	0.006
Apr	0.000	0.000	0.000	0.000
May	0.298	0.367	0.077	0.155
Jun	0.018	0.054	0.039	0.078
Jul	0.000	0.000	0.005	0.010
Aug	0.298	0.246	0.527	0.367
Sep	0.000	0.000	0.001	0.002
Oct	0.961	0.078	0.018	0.037
Nov	0.000	0.000	0.000	0.000
Dec	0.105	0.078	0.298	0.281

Blue numbers = ISCDF algorithm superior
Red numbers = BFLS/BFCDF algorithm superior

Table 5.14: ISCDF, BFLS, BFCDF: P-Values for KS test

The variations are discussed within three sections. The first section covers some alternatives to the use of least squares within the fitting summation to find the row(s) of best fit. The second section briefly outlines the results of including the surrounding time points within the fitting equations. The final section outlines the results of giving historical records different weights depending on their proximity to the to-be-infilled sites. The use of the linear cross-correlation is also suggested within this section as an alternative weighting measure.

Alternatives to the least squares summation

As an alternative fitting procedure to the least squares minimisation in Figures 5.19 to 5.32, the following functions were attempted. First, a scaled version of the least squares fitting was fitted (Equation 5.7) followed by the use of a minimum absolute difference calculation (Equation 5.8).

$$SS_i = \sum_{j=1}^{n} I_j * \left\{ \left(1 - \frac{S_{i,j}+1}{H_{t,j}+1} \right)^2 + \left(1 - \frac{H_{t,j}+1}{S_{i,j}+1} \right)^2 \right\} \tag{5.7}$$

$$SS_i = \sum_{j=1}^{n} I_j * |H_{t,j} - S_{i,j}| \tag{5.8}$$

Neither of these alternative fitting functions (Equations 5.7 or 5.8) significantly improved the algorithm's performance. In addition, a further attempt using a formula similar to the χ^2 formula (Equation 5.9) also failed to improve the fit.

$$SS_i = \sum_{j=1}^{n} I_j * \frac{(S_{i,j} - H_{t,j})^2}{(H_{i,j}+1)} \tag{5.9}$$

In equations 5.7 and 5.9 the addition of 1 in the denominator is to prevent a divide-by-zero occurring in the calculation of the fits.

Including surrounding data values

Also, for the above algorithms, the inclusion of the previous (and next) values were included in the fitting procedure. However, the inclusion of these points within the fitting procedure (when available) did not improve the effectiveness of the fitting significantly, but only markedly increased the infilling time for the algorithm. Therefore the subsequent algorithms did not make use of these surrounding values

but only concentrated on the current time points. The inclusion of surrounding values in the fit was expected to be of limited use anyway due to the low first lag autocorrelation (and even lower first lag cross-correlation).

Alternatives to equal weighting

The algorithms used give equal weighting to the valid historical data regardless of the distance between the observed data and the to-be-infilled point. As suggested in Haining (2003), the inclusion of spatial characteristics within a spatial process is desirable whenever possible. Therefore two attempted weighings of the fits were attempted as follows.

Firstly, a weighting of one over the Euclidean distance (km) between the sites was used. However, as for the inclusion of the surrounding values, all this accomplished was to slow the infilling algorithms down significantly without producing any observable benefit. Secondly, the cross-correlation was used as weights for the site values, however, this also did not procure any benefit.

A better weighting measure may be obtained from to use an exponential form of the inverse distance, to restrict the number of sites used in the selection of best points to the N closest sites, or to use all the sites but account for clustering (see Haining, 2003, pages 166-177).

6. CONCLUSIONS

I may not have gone where I intended to go, but I think I have ended up where I intended to be.

- Douglas Noel Adams, "Hitchhiker's Guide to the Galaxy"

6.1 Data analysis

The analysis of the historical records (Chapter 4) successfully identified many spurious periods within the historical record. While the erroneous data did not have an effect on the temporal statistics, the spatial statistics (especially the 24-hour cross-correlation) were significantly affected. All records identified as spurious were removed from the analysis.

6.2 Model

The spatial-temporal NSRP model is shown (Section 5.3.1) to fit the characteristics of the historical data for the pooled statistics (CV, skewness, first lag autocorrelation, and first order cross-correlation). The cross-correlation, although fitting reasonably, does not fit well at the higher distances - at the 24-hour level in particular, the higher distances are generally overestimated (Section 5.3.1). Overall, however, the fitted model reproduces the historical statistics used for fitting.

6.2.1 *Issues*

The proportion of dry sites, which is not used in the model fitting, does not match at the 24-hour level (Section 5.3.1). The 1-hour data, however, matches relatively closely over the region regardless of season (see Table 5.6).

The fitted model does not reproduce the correct spatial characteristics of the mixture of wet and dry sites at the 24-hour level (Section 5.4.1). Though it is not included herein, the 1-hour aggregation level also fails to produce the required behaviour.

The incorrect spatial characteristics described above are a result of the problems with the fitting as seen in Section 5.2.2. The overestimation of the cross-correlation at the 1-hour level at the greater distances is reflected in the overestimation of the proportion of all wet sites. The overestimation of the cross-correlation at the higher distances at the 24-hour level seems more of a reflection on the overestimation at the 1-hour aggregation level rather than a result of the model fitting at the 24-hour level. According to the fitted cross-correlation, as noted previously (Section 5.2.2), the 24-hour cross-correlation should tend to be underestimated not overestimated.

Underlying assumptions

The reason the model is not fitting the proportion dry adequately, will be due to one or more critical assumptions not being satisfied. Note that the criticality of the assumption will depend on how the model is to be used. Different model requirements may allow some assumptions to be 'invalid' as a fitted model, for that application, is able to produce satisfactory results.

The assumptions (Section 3.1.1) required are listed below for convenience:

(a) rainfall data are temporally stationary

(b) rainfall data are spatially stationary

(c) rain cells have zero-velocity

(d) cell intensity, duration, and radius are mutually independent

(e) cell x, y origins are independent within storms

(f) one cell type

(g) constant cell intensity over the cell area

In order to address the issues found with the model, the violated assumption(s) must be correctly identified and accounted for. It has been shown (Section 5.3.2) that, once standardised by dividing by the site-mean for each month, the data are spatially homogeneous. That is, the assumption of spatial stationarity can not be rejected.

The assumption of temporal stationarity has been shown to be satisfied (Section 4.4.2). The 34-year period shows no significant evidence of climate change.

The next four assumptions can not directly be confirmed as statistically violated. However, from the understanding of the physical process, none of these assumptions will be met. In addition, the model fitting is not mathematically tractable once these later assumptions (c)-(g) are relaxed. Therefore an alternative method of fitting like generalised maximum likelihood would have to be applied rather than the method of moments used here.

6.2.2 Simulation movie

In order to further clarify any problems with the spatial-temporal NSRP model, a movie of a sample 1-year simulation was constructed. As the model was used for infilling at the 24-hour aggregation level, the movie also used this aggregation level. Furthermore, the use of 24-hour data ensures the hard-drive usage was reasonable (a 1-year movie at a 1-hour aggregation level would use approximately $1.5Gb$ uncompressed) and the time for construction was feasible.

Description

$Model_B$ was used to generate synthetic, 24-hour data over an artificial $64 \times 64km$ region. Site locations were placed at $4km$ intervals and a scale parameter for each site by season was generated by using a random normal variates from a normal

distribution with mean, $\mu_{H,s}$ and standard deviation $\sigma_{H,S}$; where $\mu_{H,S}$ is the mean historical scale estimate for season S and $\sigma_{H,S}$ is the standard deviation of the historical scale estimates for season S.

Each 24-hour total was printed out on a log-scale to a 272×272 pixel bitmap file (16 pixels per site) for a full year (the bitmap production code was kindly provided by Bruce Mills). These images were then joined into an AVI file (CD:`Movie/RainMovie.avi`) at two frames per second so the transitional behaviour between 24-hour periods can be examined. The program, MakeAVI (see `http://makeavi.sourceforge.net/`), was used to join the bitmap files so they could be viewed as a movie.

The rainfall is such that *dark* pixels correspond to heavy rainfall and *light* pixels correspond to 0 or little rainfall. A log scale is used so it is easy to distinguish between no rainfall and light rainfall (see also CD:`Movie/README.txt`).

Discussion

From this movie of the synthetic record, some problems with the 24-hour record are evident. As expected, the majority of time points are either all wet or all dry (especially the latter). There is a more critical problem from the perspective of realistic rainfall simulation, however. A closer examination of the 24-hour time steps reveals that there are few (if any) storms that continue over a 24-hour period. Rather, each storm tends to begin and end within a 24-hour block. This is not characteristic of 'real' rainfall - especially over winter months. The lack of 'long' duration storms indicates a further point where the spatial-temporal NSRP model is failing to capture the behaviour of the historical rainfall data.

6.3 Infilling

The infilling algorithms use a 300-year simulated record to select suitable points for replacing missing values, as described in Section 3.3. The record is generated using the model fitted in Section 5.2. As discussed above, this model is deficient and does not reproduce all the characteristics of the historical data (for example,

the proportion dry). Nevertheless, the synthetic record is used as input into the infilling algorithms and an algorithm which avoids the model deficiencies while infilling sensible results is derived.

6.3.1 Best fit algorithms

The results from Sections 5.5.2 and 5.5.3 show that infilling using a best fit approach from synthetic data generated by the spatial-temporal NSRP model (Section 5.2) is not an appropriate method. Although the algorithm generally predicted historical values that were close to the true values, the pooled statistics are underestimated (eg: Section 5.5.2).

There are two possible sources considered for this underestimation. Firstly, it is observed that the method of least squares fit on the raw historical and synthetic data may be biasing the fit away from the extreme events. Secondly, as the model was not producing the correct 'mixture' of wet and dry sites any aggregation level, this is also a likely cause of the underestimation.

In order to address the first issue, the best fit least squares algorithm was modified to use CDFs to transform data prior to applying the summation. A CDF for each season and site was calculated from the synthetic data. In the least squares fit, the $CDF(x)$ and $CDF(y)$ were used as input (where x is the historical data some at time t, and y is a candidate record for infilling for the equivalent site). Therefore, points are fitted based on the distance between the probabilities of observing the points rather than based on their magnitude. This is advantageous as the records, particularly of the extreme values, are sparse. By transforming the data to a $[0, 1]$ scale, the effect of the magnitude of these extreme records on the fit has been significantly reduced.

The result of this algorithm adjustment was that the infilling algorithm is no longer as biased away from 'extreme' fitting. Indeed from some of the plots (eg: Figure 5.28), the algorithm has over compensated and the algorithm overestimated the quantiles. In general, however, there is still a tendency to under fit the true historical values. Furthermore, the pooled statistics are still underestimated (Section 5.5.3).

As the adjustment to use the CDFs does not fix the problem observed, it was concluded that if a fitted model does not produce the correct spatial-temporal characteristics (particularly relating to the proportion dry), then a method using a best fit algorithm will always be inadequate. Note also that other 'best fit' fitting functions and/or modifications (Section 5.5.6) also do not improve the algorithm.

6.3.2 Iterative sampling algorithms

The last algorithm considered, is derived from a sampling based method. Firstly, rather than infilling from an exact match, a 'best fitting row' is selected from the best 5% of the candidate rows. Secondly, each missing value is replaced iteratively, in a random site order, and the algorithm reapplied. Thus for this algorithm, infilled values within a record in the historical data, may be in different source rows in the simulated record. This is important as it enables the spatial deficiency of the fitted model to be mitigated.

Section 5.5.4 shows that the iterative sampling approach enables the infilled data to capture both the temporal and spatial characteristics reasonably well. The algorithm is not as 'accurate' as the best fit infilling algorithms discussed previously, however it is less biased (see Section 5.5.5). In general, considerably more variation in the predictions is evident in results obtained for the sampling algorithm (see for example, Figure 5.34).

This algorithm is superior to the best fit alternatives both spatially and temporally. However as the results for the coefficient of variation show (Table 5.14), the algorithm does not always infill results that are closer to the historical estimates than the best fit alternatives. For the months were this is the case (principally May through August), a hybrid approach could be taken that mixed the BFCDF algorithm and the ISCDF algorithm.

6.4 Conclusions

The derived infilling algorithm is likely to be adequate for the purpose for which it was designed even though the fitted model was shown to be spatially deficient. The iterative sampling infilling algorithm maintains the characteristics of the known historical data. The spatial mixture of wet and dry sites is not significantly different from the historical records for most cases. Nevertheless, the infilled values are close to the true historical values.

The developed algorithm is superior to the best fit least squares algorithm that is currently being used to select records for infilling. The sampling from more than one time point reduces the cross-correlation of the records, however the mixture of wet and dry sites is more likely to be maintained. Furthermore, the use of the CDF within the record selection procedure ensures that the algorithm is not biased away from infilling extreme points.

An extreme value frequency analysis also may need to be performed to compare the respective performance of the BFLS algorithm and the ISCDF algorithm against the historical values. This would be particularly useful if the record is to be used as a basis for flood frequency analysis where extreme event prediction is critical. In any case, it is expected that the ISCDF algorithm will be more effective than the BFLS algorithm purely due to the use of the CDFs in the record selection process. Similarly further tests, for example routing infilled data through a rainfall-runoff model, may be necessary before the algorithm is applied in practice.

Although some improvements could be made with the algorithm (see Section 6.5.2), it is likely that a better fitted model of the rainfall data is necessary to generate accurate infilled records. Nevertheless, as mentioned previously in Section 3.3.1, the infilling analysis described in this thesis is a harder task than infilling the missing data in the historical record. All the information that is removed for the testing of the algorithms, is available for the infilling of ordinary missing data. Therefore, it is likely that a better estimate of these missing records would be obtained relative to the 'missing values' predicted by the infilling algorithms in the testing results.

6.5 Future research

The results of the model fitting, heuristic analysis, and infilling algorithms have highlighted some areas where further development is necessary. This section is divided into two subsections. The first subsection examines further implementation specific directives for the applied algorithms. The second subsection is not related to this implementation, but rather discusses areas needing further investigation.

6.5.1 Internal algorithms

Simulation

The algorithm for generating the simulated records could be made more efficient. In order to generate a homogeneous point process, a sufficiently large region needs to be simulated over. In this case, a square region of length equal to the observed area plus four times the maximum cell width for each season was used. This is known to be excessive, but the simulation is designed to be conservative and as homogeneity is necessary requirement, the region size should be overestimated rather than underestimated. However, one side effect from generating rainfall over such a large region is that the number of rejected cells is over 99% for either fitted model.

Also, from Section 5.3.2 it is evident that even a 300-year simulation may not be completely adequate for infilling. Therefore, for the purpose of infilling, and as a by-product disaggregation, it may be better to simulate records that match with the observed monthly statistics rather than an independent set from the historical record. This approach would tend to oversample the observed monthly patterns, and thus is expected to be more beneficial for infilling and disaggregation algorithms.

Model fitting

The algorithm used to find the optimal model parameters used a Nelder-Mead simplex algorithm. This algorithm is known to be inefficient and may not converge. Although a parallel search seemed to increase the probability of finding the

optimal solution, this approach needs further confirmation. Furthermore, a recent modification of the Nelder-Mead algorithm which used bases and frames provides a provably convergent algorithm (see Price et al., 2002).

Further heuristics

Two further heuristics that may speed up the infilling algorithms are the reducing fit recalculation and the partitioning on total regional rainfall. The first heuristic does not require any additional resource allocation and, as a result, is the cheapest heuristic to implement. Both partitioning systems (wet/dry days and total rainfall) require additional memory to manage the segments. In particular, the proposed partitioning based on total rainfall, if used independently of the other heuristics, requires a sensible partitioning algorithm to balance the cost of management memory usage with the potential speed increase.

Reducing fit recalculation

This heuristic is based on the observed primary segmentation into all wet and all dry segments. However, unlike the first heuristic, it is heavily dependent on the cutoff point for wet/dry sites and may affect the fitting results.

Firstly it was observed, that the all wet days can still have considerable variation within what is considered wet (in this project an observation with a value of at least $0.1mm$ was considered *wet* as this was minimum recorded non-zero value in the 24-hour record), but the majority of the all dry days are $0.0mm$ for all sites. Therefore, the fit could be computed *once* as a difference from 0 at all valid sites, and this fit assigned to all records within the dry segment without recomputation.

For the historical record, this measurement will be exact, but, if an iterative fitting method is used, then it is possible for a missing datum to be infilled with a non-zero rainfall amount that is less than $0.1mm$ and therefore considered dry. Provided the number of sites is small, however, the total amount of error introduced into the fitting computation is not significantly high.

An equation for the maximum error can be obtained (Equation 6.1).

$$N \times (0.1)^2 + \sum_{i=1}^{N} (0.2 \times X_i) \tag{6.1}$$

where X_i is the maximum rainfall value at the ith site and N is number of valid sites in the record.

This is obvious from the least squares equation, $(O - E)^2$, where the error is introduced by assuming that E is 0. Remembering that the fit is proportional to X_i^2, it is clear that the relative error is high only as the historical estimate, O, tends to 0. This argument particularly applies if the raw data are used in the least squares fit. However, if the $CDF(x)$ is used in the fitting equation, as used in Section 5.5.3 and following, then no error is accumulated as $CDF(0)$ and $CDF(0.1 - \epsilon)$ (where 0.1 can be substituted by the granularity of the historical data) produces the same output.

A further observation derived from the behaviour of this heuristic is that the number of records of $0.0mm$ at each site in the simulated record could be reduced to one record of 0's (if, in fact, it was stored at all). Care has to be taken with this approach if data at the surrounding times are to be used in the fitting algorithm along with the current time. However, is expected that considerable storage space can be saved by making use of the sparseness of the rainfall data matrix even when accounting for surrounding records thus allowing for a longer simulation record to be generated if required.

Partitioning on total regional rainfall

A further heuristic, which could be considered, is to use the total amount of rainfall seen over the region from the currently valid data. The simulated record could be partitioned based on the total amount of rainfall seen according to a sensible partitioning system - so that the partitions are approximately equal in size. As with the wet/dry heuristic, memory overhead is added for each partition, and a careful analysis has to be done to find an optimal balance between memory usage and speed increase.

If this heuristic is implemented concurrently with the successful partitioning of the wet/dry sequences with this fitted model, then it would only make sense to partition the *all* wet partition. This is intrinsically obvious from the split of the partitions of the number of wet/dry days in the region (Table 5.11). From Table 5.11, it is evident that the only useful speed increase would be obtained from this all wet segment as the other records are not of a significant enough size and 'partitioning' an all zero sequence (the only other segment of a significant size) would be nonsensical.

If a more accurate model is fitted in which the mixture of wet and dry days are in the majority, then it may be beneficial to group the partitions given by the wet/dry heuristic and then repartition each group based on the total regional rainfall. In this case, it is expected that the speed increase originally generated by the wet/dry heuristic, could be maintained even when the matching of wet and dry sequences is distributed more evenly.

Infilling

It has so far been assumed that an entire synthetic record is generated prior to infilling. This assumption allows the model to be simulated once then stored as a complete simulation record in a single file. However, this limits the number of years that can be simulated due to the memory storage space necessary to hold the records. Therefore, in order to get around this issue, the simulation and infilling algorithms should be integrated.

When the simulation and infilling algorithms are integrated, then the simulated record can be obtained only for the season currently being infilled. This would effectively allow a simulation record of (say) 1000 years to be generated while using the same amount of memory that a full 300-year simulation record would use. This is obviously advantageous particularly when CDFs are to be calculated accurately.

6.5.2 *General improvements*

Data analysis

The algorithm used to select spatial plots could be further improved as outlined below.

Firstly, months where the problems are caused purely because of a short available record should generally not be plotted (though should be noted for subsequent checking especially if the mean and variance is 0). For example, in the spatial analysis, a series of plots may indicate a problem with a short record when it would have been more useful to have a table of results detailing these issues.

Secondly, the selection algorithm should also filter plots selected by site and year. For example, 10+ plots similar to Figure A.27, showing that for January 1990 site TW287141 had all values recorded as 0 and was in conflict with other sites, is superfluous. A random sample of these plots would be more useful especially when a large number sites are being analysed.

Thirdly, the regression analysis of the monthly maximums should be replaced with a more optimal method. The assumptions for the regression model are not satisfied and, as such, this is not the best way to determine whether months are spatially inconsistent.

There are several alternatives that avoid the problem with the regression analysis while producing the same results.

(a) An analysis of the distribution of the *difference* between monthly maximums. Any outliers within this distribution are points requiring further investigation.

(b) Further to (a), all site data could be first transformed to a CDF. This would prevent differences in scale from unduly influencing the selection of plots. Furthermore, this method is particularly applicable when the region analysed is non-homogeneous.

(c) While (a),(b), and the proportion dry selection methods examine overall monthly results, a method where a sequential integrity of the spatial data would also be useful. For example, a vector AR model could be constructed to model short-term and long-term deviance between two sites. From this model, an estimate of the probability of observing a discrepancy within a month could be obtained. As this model is for cross-validation, the month currently being examined would not be included in the model construction.

Note that if (c) were available, then this method would be the most useful as it gives a probability of observing the discrepancy between the sites. From this probability, a statistically justifiable decision can be made to either include or discard the records.

Model fitting procedure

The model fitting equations (Section 3.1.5) gave equal weight to all aggregation levels. This was done as the model was intended both for infilling and disaggregation so a close match is needed at both the 1-hour and 24-hour aggregation levels. However, a better model may be obtained by weighting each fit proportionally to the error associated with the statistic. For example, the 24-hour autocorrelation should not be given the same weight as the 1-hour autocorrelation.

Model improvements

As the proportion of dry sites is not matched well by the fitted model this could be included in the model fitting procedure. The overall proportion of dry sites could be used (that is, pooled across the region), however a more useful measure is likely to be the proportion of dry sites conditioned on the observed site values.

There are two assumptions that, once relaxed, may lead to a better fit to the historical record. Firstly, the inclusion of multiple rain cell types is likely to improve the fitting ability of the model. Secondly, a model where cell intensity, duration, and radius are not mutually independent is also expected to be beneficial. Note that it may be necessary to correct one of the assumptions listed above as the

inclusion of multiple cells, especially if duration and intensity were not assumed to be independent, may account for a relationship between duration, intensity, and cell radius. Alternatively, a dependence between cell intensity, duration, and radius, may negate the necessity for including multiple cell types.

Multiple rain cells

Multiple cell-types may help fix the cross-correlation problems seen in Section 5.2.2 - particularly with the apparent mismatch between the 1-hour and 24-hour cross-correlations. It is also possible that fitting multiple cells may allow one of the cell-types to be of longer duration, thus reducing the problem seen in the movie (Section 6.2.2). Furthermore, the discrepancy between the mixture of wet/dry days in the region would be expected to be at least partially alleviated by this model specification. If duration and intensity are permitted to be dependent and multiple cells types are modelled, then it is possible that this would account for any relationship between duration, intensity, and cell radius.

Dependent cell intensity, duration, and radius

The likelihood of cell radius, duration, and intensity being independent is unreasonable given the nature of the underlying physical process. As a temporal model assuming a relationship between duration and intensity was found to be beneficial (Kakou, 1998; Onof et al., 2000), it is expected that a model where cell radius, duration, and intensity are not assumed to be mutually independent would also improve the equivalent spatial-temporal model. Essentially, the model would change to modelling rainfall volume rather than just magnitude.

The advantage of including multiple cells is that the formulation of the model has already been derived (see Cowpertwait, 1995). Therefore, this would be the suggested next model to apply in the short term. However, a derivation of a model where cell intensity, cell duration, and cell radius are not mutually independent is highly likely to be an area worth investigating.

Infilling

The iterative sampling algorithm does not directly include a weighting for distance between historical sites and rainfall. Although a direct weighting via linear cross-correlation and weighting via distance (km) was not found to be useful (see Section 5.5.6), an algorithm that makes use of the regional information is likely to be better than an algorithm which does not.

The weighting was an attempt to obtain an evaluation over the whole region where the sites which were further away had less influence that sites that were closer. A better approach may be to cluster the results based on similarity and to filter out the candidate simulation records that do not maintain the same regional characteristics prior to the fitting. This would also increase the speed of the infilling algorithms if the computation cost of filtering is less than the cost of fitting. As the number of sites increases, this is likely to be true. A further advantage of this approach is that the effect of duplicated regional information is reduced (see Haining, 2003).

Another approach to infilling which was not considered, is to choose the infilled records based on how closely the pooled infilled monthly statistics matches the observed pooled historical monthly statistic. As the recalculation of statistics is expensive, this is probably best implemented via multiple imputation.

The algorithm would then be as follows. First, the month is infilled using iterative sampling (or some modified variant). Second, the pooled statistics are calculated both for the historical record and the infilled record. If the difference between any two equivalent statistics is too large (for example they are significantly different at the 5% threshold) then the infilled data are to be rejected and the algorithm is reapplied. Otherwise, if another month needs to be infilled then the algorithm operates on that month.

The direct benefit of this approach is that any seasonal fluctuation (for example dry / wet seasons) is correctly accounted for. If the original methods for infilling were applied, then this would not necessarily be the case as there is no split of the candidate records into wet/dry seasons.

BIBLIOGRAPHY

Balaji, R. (1995). *Nonparametric Stochastic Generation of Daily Precipitation and Other Weather Variables*. PhD thesis, Utah State University.

Bell, T. (1987). A space-time stochastic model of rainfall for satellite remote-sensing studies. *Journal of Geophysical Research*, 92:9631–9644.

Bennett, R., P., H. R., and Griffith, D. (1984). Review article: The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers*, 74:138–156.

Cameron, D., Beven, K., and Tawn, J. (2001). Modelling extreme rainfall using a modified random pulse Bartlett-Lewis model (with uncertainty). *Advances in Water Resources*, 24:203–211.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte-Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87.

Cowpertwait, P. (2001). A renewal cluster model for the inter-arrival times of rainfall events. *International Journal of Climatology*, 21:49–61.

Cowpertwait, P. S. P. (1994). A generalised point process model for rainfall. *Proceedings: Mathematical and Physical Sciences*, 447:23–37.

Cowpertwait, P. S. P. (1995). A generalized spatial-temporal model of rainfall based on a clustered point process. In *Proceedings Royal Society*, volume 450, pages 173–175. The Royal Society.

Cowpertwait, P. S. P. (1998). A poisson-cluster model of rainfall: high order moments and extreme values. In *Proceedings: Mathematical, Physical, and Engineering Sciences*, volume 454, pages 885–898. The Royal Society.

Cowpertwait, P. S. P., Kilsby, C. G., and O'Connell, P. E. (2002). A space-time Neyman-Scott model of rainfall: Empirical analysis of extremes. *Water Resources Research*, 8.

Cowpertwait, P. S. P., Lockie, T., and Davis, M. (2004). A stochastic spatial-temporal disaggregation model for rainfall. *Research Letters in the Information and Mathematical Sciences*, 6:109–122. Available online from `http://iims.massey.ac.nz/research/letters/`

Cowpertwait, P. S. P. and O'Connell, P. E. (1997). A regionalised Neyman-Scott model of rainfall with convective and stratiform cells. *Hydrology and Earth System Sciences*, 1:71–80.

Cowpertwait, P. S. P., O'Connell, P. E., Metcalfe, A., and Mawdsley, J. (1996). Stochastic point process modelling of rainfall. I. Single-site fitting and validation. *Journal of Hydrology*, 175:17–46.

Cox, D. R. and Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 317–328.

Cressie, N. (1993). *Statistics for spatial data.* John Wiley and Sons, revised edition.

De Lannoy, G., Verhoest, N., and De Troch, F. P. (2004). Characteristics of rainstorms over a temperate region derived from multiple time series of weather radar images. *Journal of Hydrology.* Available online from `www.sciencedirect.com`.

De Oliveira, V., Kedem, B., and Short, D. A. (1997). Bayesian prediction of transformed gaussian random fields. *American Statistical Association*, 92(440).

Dellaportas, P. and Roberts, G. O. (2002). An introduction to MCMC. In *Spatial Statistics and Computational Methods.* Springer-Verlag.

Dobi-Wantuch, I., Mika, J., and Szeidl (2000). Modelling wet and dry spells with mixture distributions. *Meteorology and Atmospheric Physics*, 73:245–256.

Favre, A.-C. and Overney, O. (1999). Investigations of the temporal and spatial properties of rainfall series: some insights into rainfall modelling. In *2nd Inter-Regional Conference on Environment-Water 99.*

Feyerherm, A. and Bark, L. D. (1964). Statistical methods for persistent precipitation patterns. *Journal of Applied Meteorology*, 4(3):320–328.

Gaither, C. and Cavazos-Gaither, A. (1996). *Statistically Speaking: A dictionary of quotations.* Institute of Physics Publishing.

Gerald, C. and Wheatley, P. (1984). *Applied numerical analysis.* Addison-Wesley Publishing Company, 3rd edition.

Glasbey, C., Cooper, G., and M.B., M. (1995). Disaggregation of daily rainfall by conditional simulation from a point-process model. *Journal of Hydrology*, 165:1–9.

Grunwald, G. and Jones, R. (2000). Markov models for time series with mixed distribution. *Environmetrics*, 11:327–339.

Güntner, A., Olsson, J., Calver, A., and Gannon, B. (2001). Cascade-based disaggregation of continuous rainfall time series: the influence of climate. *Hydrology and Earth System Sciences*, 5:145–164.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice.* Cambridge University Press.

Hox, J. (1999). A review of current software for handling missing data. *Kwantitaieve Methoden*, 62:123–138.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden markov model for precipitation occurrence. *Applied Statistics.*

Islam, S., Entekhabi, D., Bras, R., and Rodriguez-Iturbe, I. (1990). Parameter estimation and sensitivity analysis for the modified Bartlett-Lewis rectangular pulses model of rainfall. *Journal Geophysical Research*, 95:2093–2100.

Johns, C., Nychka, D., Kittle, T., and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, 98:796–806.

Kakou, A. (1998). A point process model for rainfall with dependent duration and intensity. Research Report, Department of Statistical Science, University College London.

Katz, R. W. and Parlange, M. B. (1998). Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11:591–601.

Kohn, R. and Ansley, C. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *Journal of the American Statistical Assocation*, 81:751–761.

Kong, A., Liu, J., and Wong, W. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Assocation*, 89:278–288.

Kottegoda, N., Natale, L., and Raiteri, E. (2003). A parsimonious approach to stochastic multisite modelling and disaggregation of daily rainfall. *Journal of Hydrology*, 274:47–61.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *Siam Journal Optimisation*, 9(1):112–147.

Le Cam, L. (1961). A stochastic description of precipitation. In Newman, J., editor, *Proceedings 4th Berkeley Symposium Mathematics, Statistics and Probability*, pages 165–186. University of California Press, Berkeley.

Maidment, D. R., editor (1993). *Handbook of Hydrology*. McGraw-Hill Inc.

Mark, O. and Hosner, M. (2002). *Urban Drainage modeling - a collection of experiences from the past decade*. DHI. Available online from `http://www.dhisoftware.com/book/index.htm`.

Matsumoto, M. (2004). Mersenne twister: A random number generator (since 1997/10). Available online. Last accessed February 2005 from `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html`

Mehrotra, R. and Singh, R. (1998). Spatial disaggregation of rainfall data. *Hydrological Sciences - Journal - des Sciences Hydrologiques*, 43.

Mellor, D. (1996). The Modified Turning Bands (MTB) model for space-time rainfall. i. model definition and properties. *Journal of Hydrology*, 175:113–127.

Mellor, D. and Metcalfe, A. (1996). The Modified Turning Bands (MTB) model for space-time rainfall. iii. estimation of the storm/rainband profile and a discussion of future model prospects. *Journal of Hydrology*, 175:161–180.

Mellor, D. and O'Connell, P. (1996). The Modified Turning Bands (MTB) model for space-time rainfall. ii. estimation of raincell parameters. *Journal of Hydrology*, 175:129–159.

Mohapl, J. (2002). A precipitation occurrence model. *Stochastic Environmental Research and Risk Assessment*, 16:143–154.

Northrop, P. (1998). A clustered spatial-temporal model of rainfall. *Proceedings of the Royal Statistical Society London A*, 454:1875–1888.

Olsson, J. and Burlando, P. (2002). Reproduction of temporal scaling by a rectangular pulses rainfall model. *Hydrological Processes*, 16:611–630.

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

Onof, C., Yameundjeu, B., Paoli, J., and Ramesh, N. (2002). A Markov modulated Poisson process model for rainfall increments. *Water Science and Technology*, 45.

Ormsbee, L. (1989). Rainfall disaggregation model for continuous hydrologic modeling. *Journal of Hydraulic Engineering*, 15.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2002). *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2nd edition.

Price, C., Coope, I. D., and Byatt, D. (2002). A convergent variant of the Nelder-Mead algorithm. *Journal of Optimization Theory and Applications*, 113(1):5–19.

Ramesh, N. (1998). Temporal modelling of short-term rainfall using Cox processes. *Environmetrics*, 9:629–643.

Rodriguez-Iturbe, I., Cox, D., and Eagleson, P. (1986). Spatial modelling of total storm rainfall. *Proceedings of the Royal Society London A*, 403:27–50.

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

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

Sansó, B. and Guenni, L. (1999). Venezuelan rainfall data analysed by using a Bayesian space-time model. *Applied Statistics*, 48:345–362.

Schafer, J. and Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7.

Skaugen, T., Astrup, M., Roald, L., and Førland, E. (2003). Scenarios of extreme daily precipitation for Norway under climate change. *Nordic Hydrology*, 35:1–13.

Smith, J. and DeVaux, R. (1994). A stochastic model relating rainfall intensity to raindrop processes. *Water Resources Research*, 30(3):651–665.

Smith, J. A. (1993). Marked point process models of raindrop-size distributions. *Journal of Applied Meteorology*, 32:284–296.

Smithers, J., Pegram, G., and Schulze, R. (2002). Design rainfall estimation in South Africa using Bartlett-Lewis rectangular pulse rainfall models. *Journal of Hydrology*, 258:83–99.

Srikanthan, R. and McMahon, T. (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth Sciences*, 5:653–670.

Stern, R. and Coe, R. (1984). A model fitting analysis of rainfall data. *Journal of the Royal Statistical Society. Series A*, 147:1–34.

Tilford, K., Sene, K., and Collier, C. (2003). Flood forecasting - rainfall measurement and forecasting. Research Report, Environment Agency. Available online from `http://www.environmental-agency.gov.uk/commondata/acrobat/w5c_013_4.tr.pdf`

Toth, V. T. (2004). Programmable calculators: Calculators and the gamma function. Available online. Retrieved October 10, 2004, from `http://www.rskey.org/gamma.htm`

Venugopal, V., Foufoula-Georgiou, E., and Sapozhnikov, V. (1999). A space-time downscaling model for rainfall. *Journal of Geophysical Research*, 104:19705–19721.

Verlarde, L. G. C., Migon, H. S., and Pereira, B. d. B. (2004). Space-time modeling of rainfall data. *Environmetrics*, 15:561–576.

Weisstein, E. W. (2003a). "Heaviside Step Function." from *MathWorld*–a Wolfram Web Resource. Wolfram Research, Inc. Retrieved December 23, 2004, from `http://mathworld.wolfram.com/HeavisideStepFunction.html`

Weisstein, E. W. (2003b). "Uniform Difference Distribution" from *MathWorld*–a Wolfram Web Resource. Wolfram Research, Inc. Retrieved December 23, 2004, from `http://mathworld.wolfram.com/UniformDifferenceDistribution.html`

Willems, P. (2001). A spatial rainfall generator for small spatial scales. *Journal of Hydrology*, 252:126–144.

Willems, P. and Luyckx, G. (1999). Stochastic generation of spatial rainfall for urban drainage areas. *Water Science and Technology*, 39:23–30.

Zhiquan, B. Shafiqul, I. and Eltahir, E. (1994). Aggregation-disaggregation properties of a stochastic rainfall model. *Water Resources Research*, 30(12):3423–3435.

A. DATA INTEGRITY ANALYSIS: PLOTS AND TABLES

A.1 Temporal plots

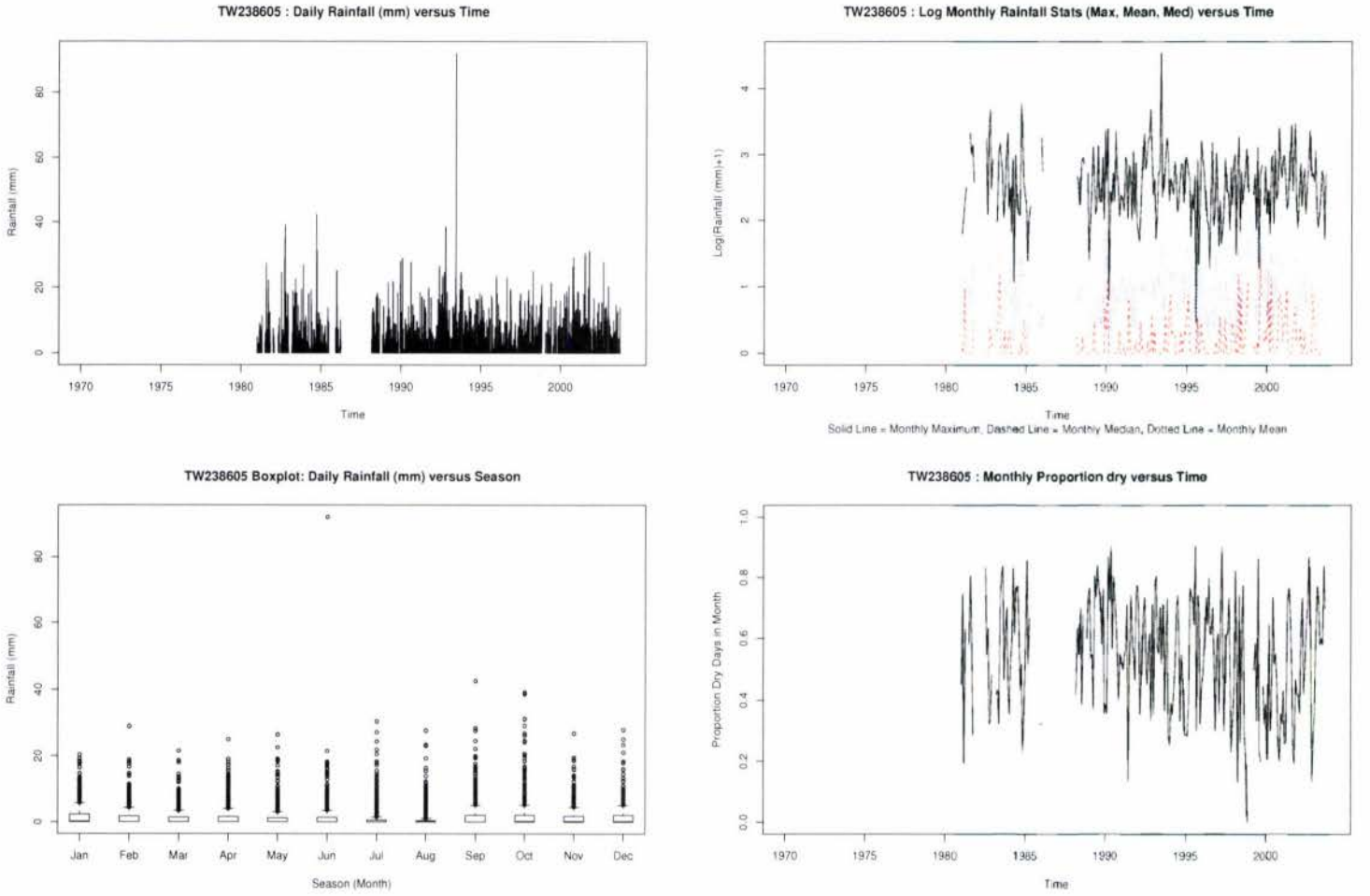


Figure A.1: Site TW238605 daily plots

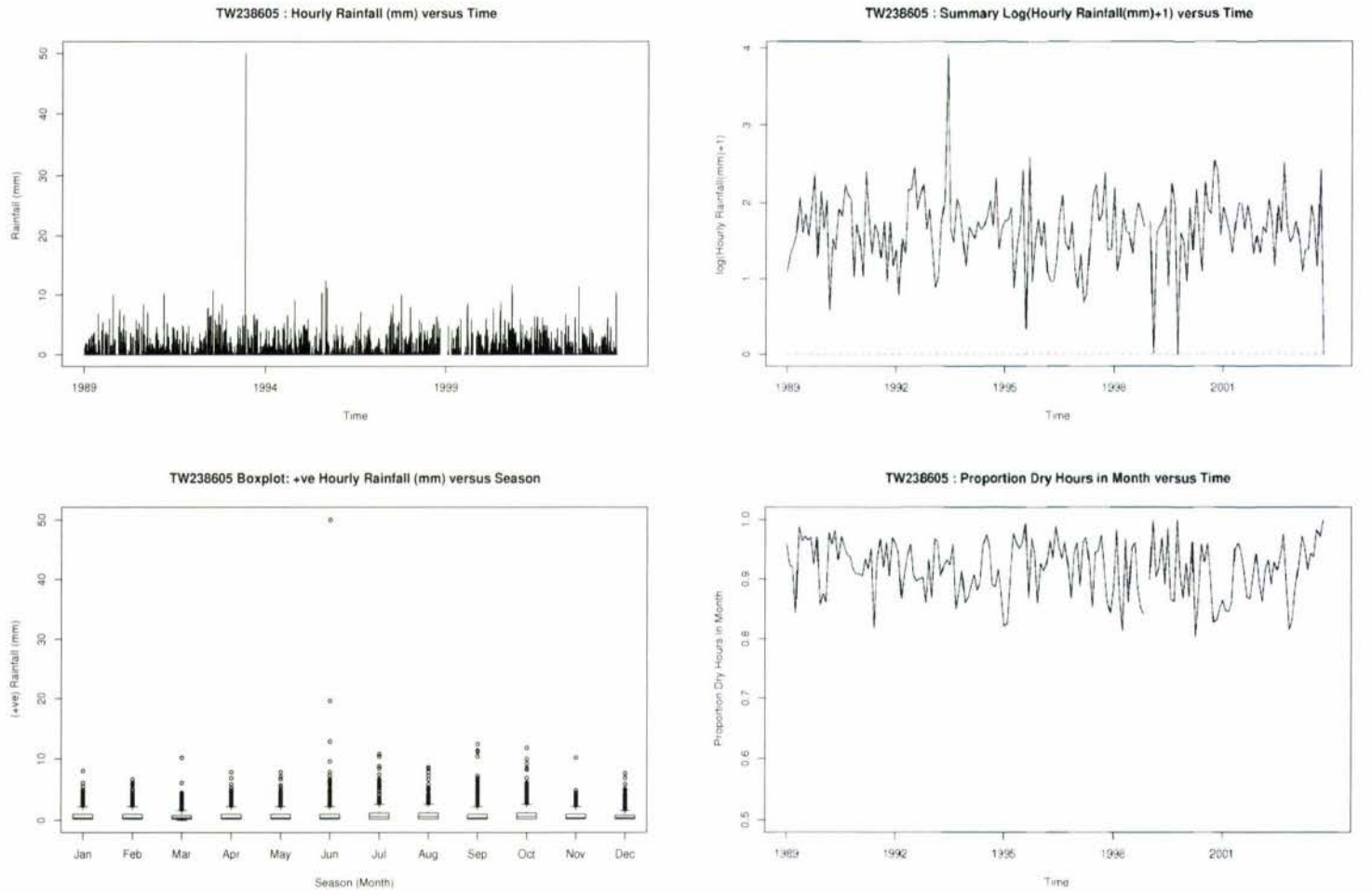


Figure A.2: Site TW238605 hourly plots

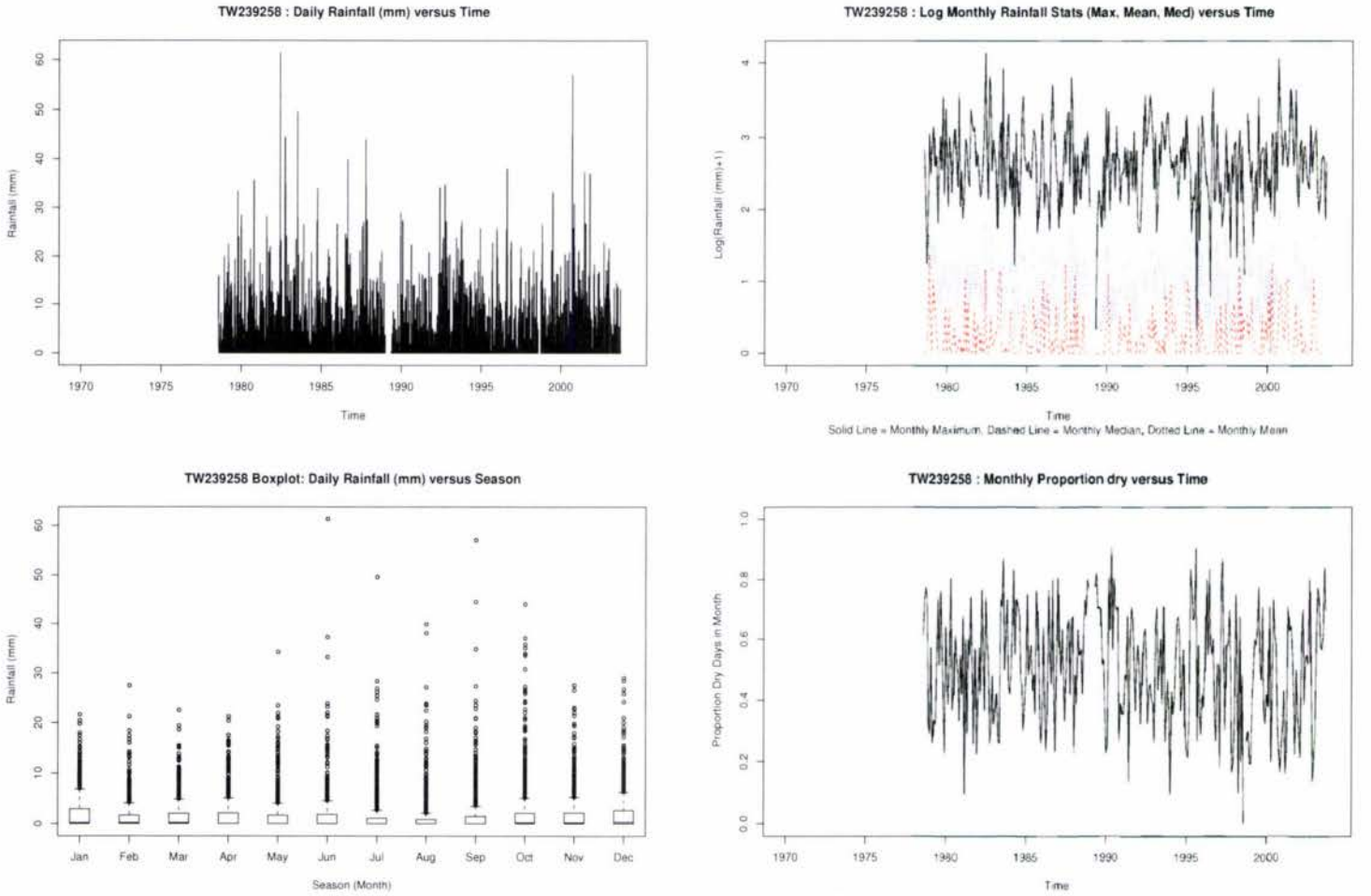


Figure A.3: Site TW239258 daily plots

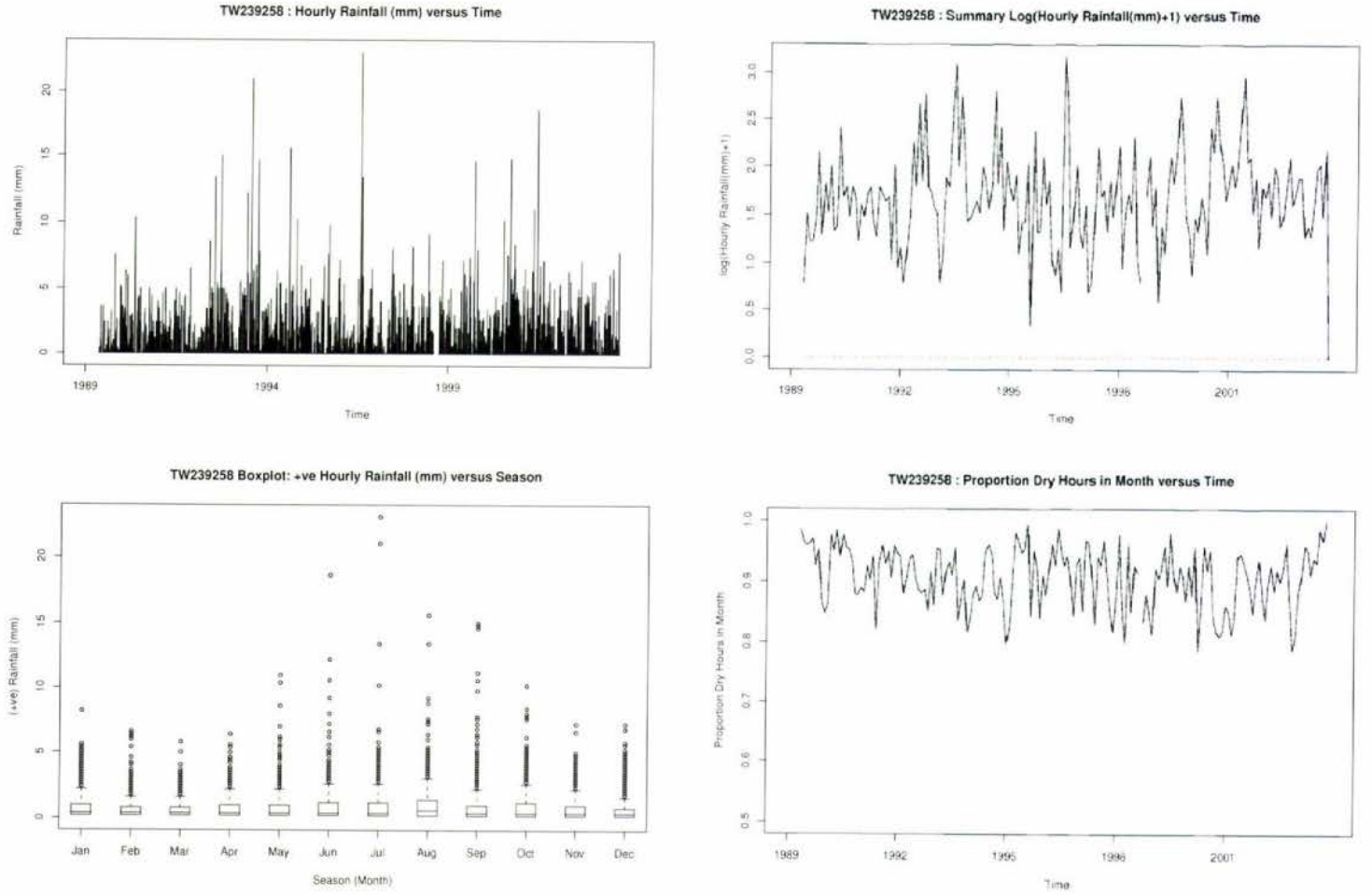


Figure A.4: Site TW239258 hourly plots

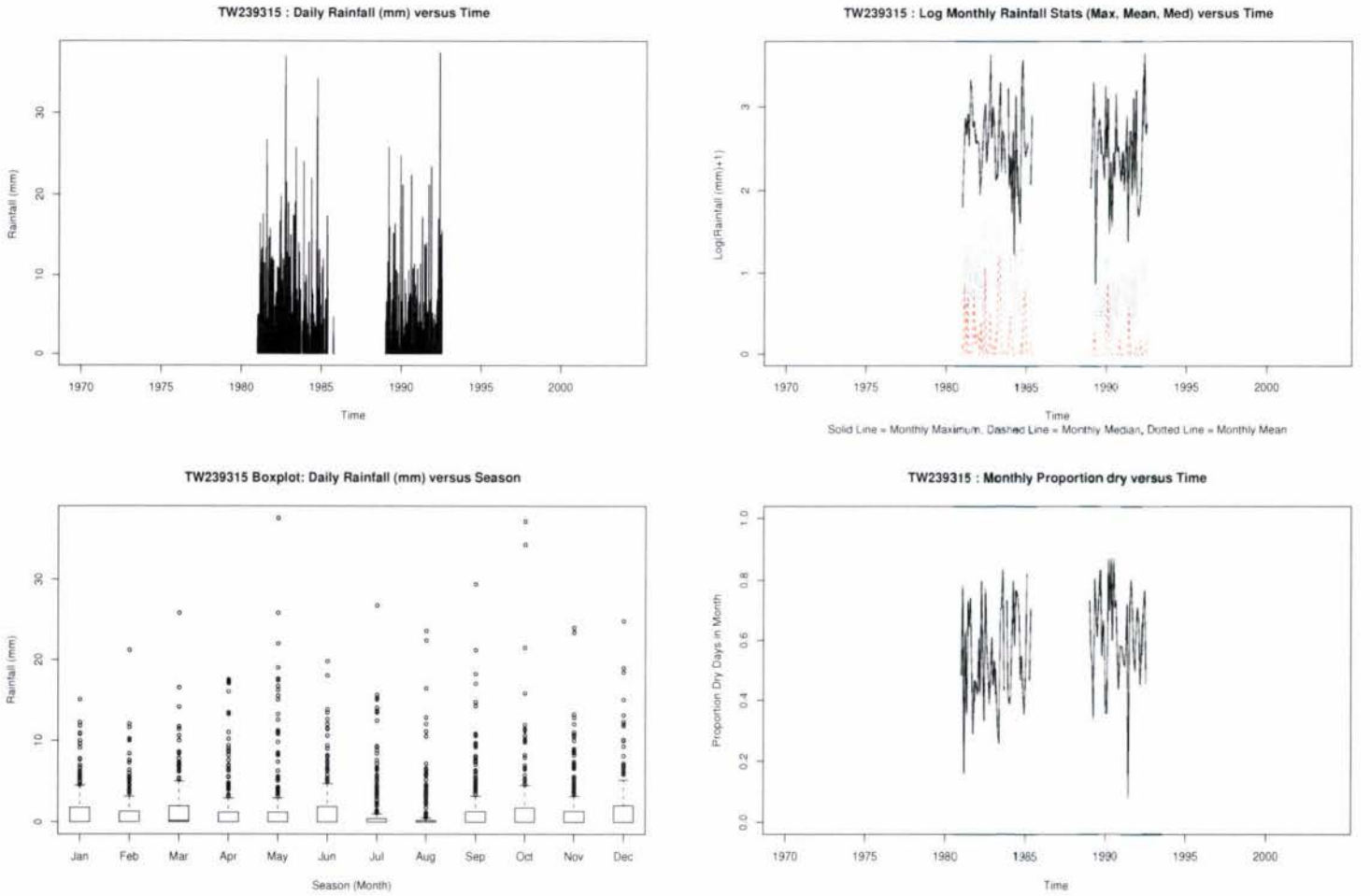


Figure A.5: Site TW239315 daily plots

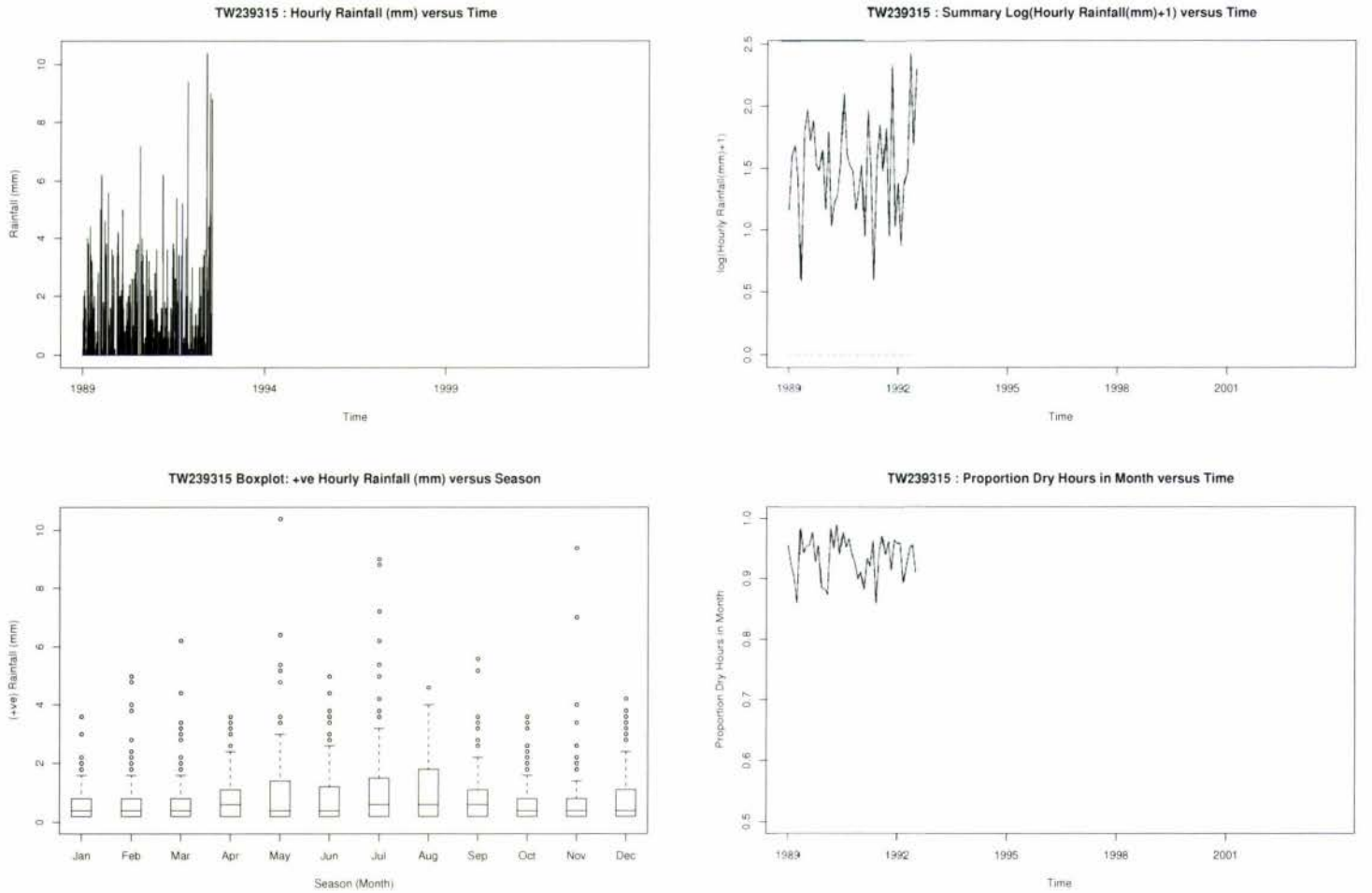


Figure A.6: Site TW239315 hourly plots

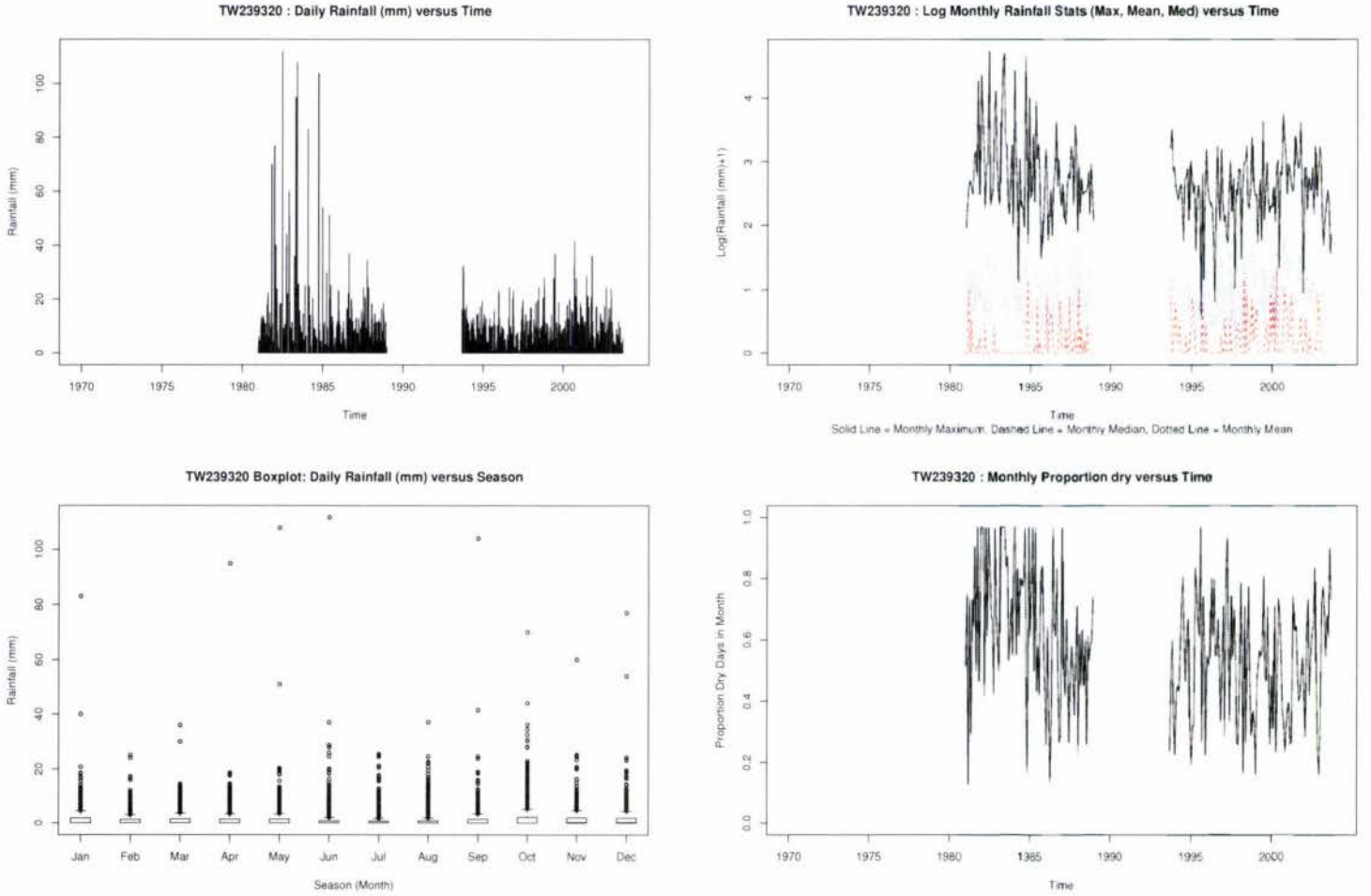


Figure A.7: Site TW239320 daily plots

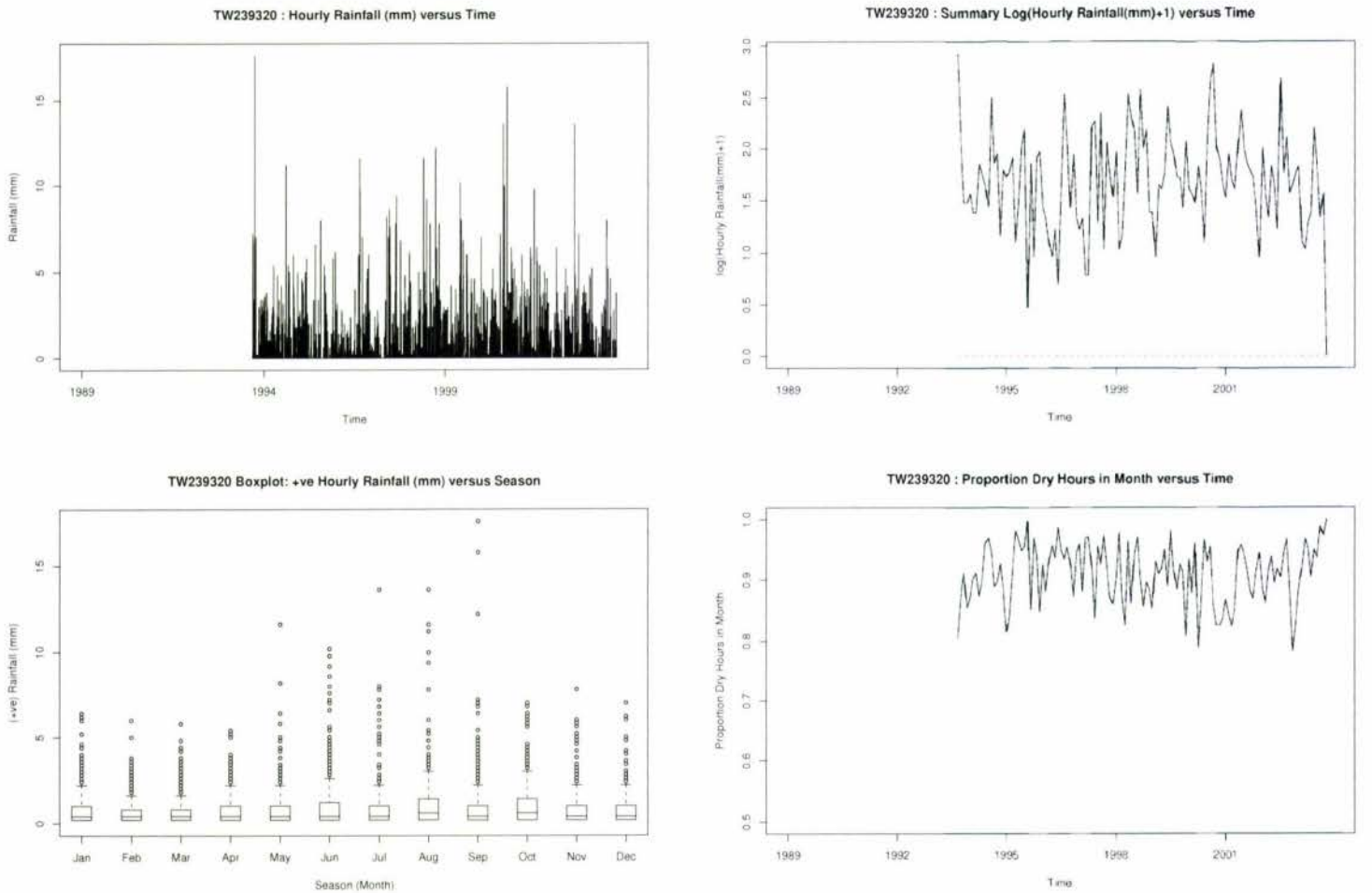


Figure A.8: Site TW239320 hourly plots

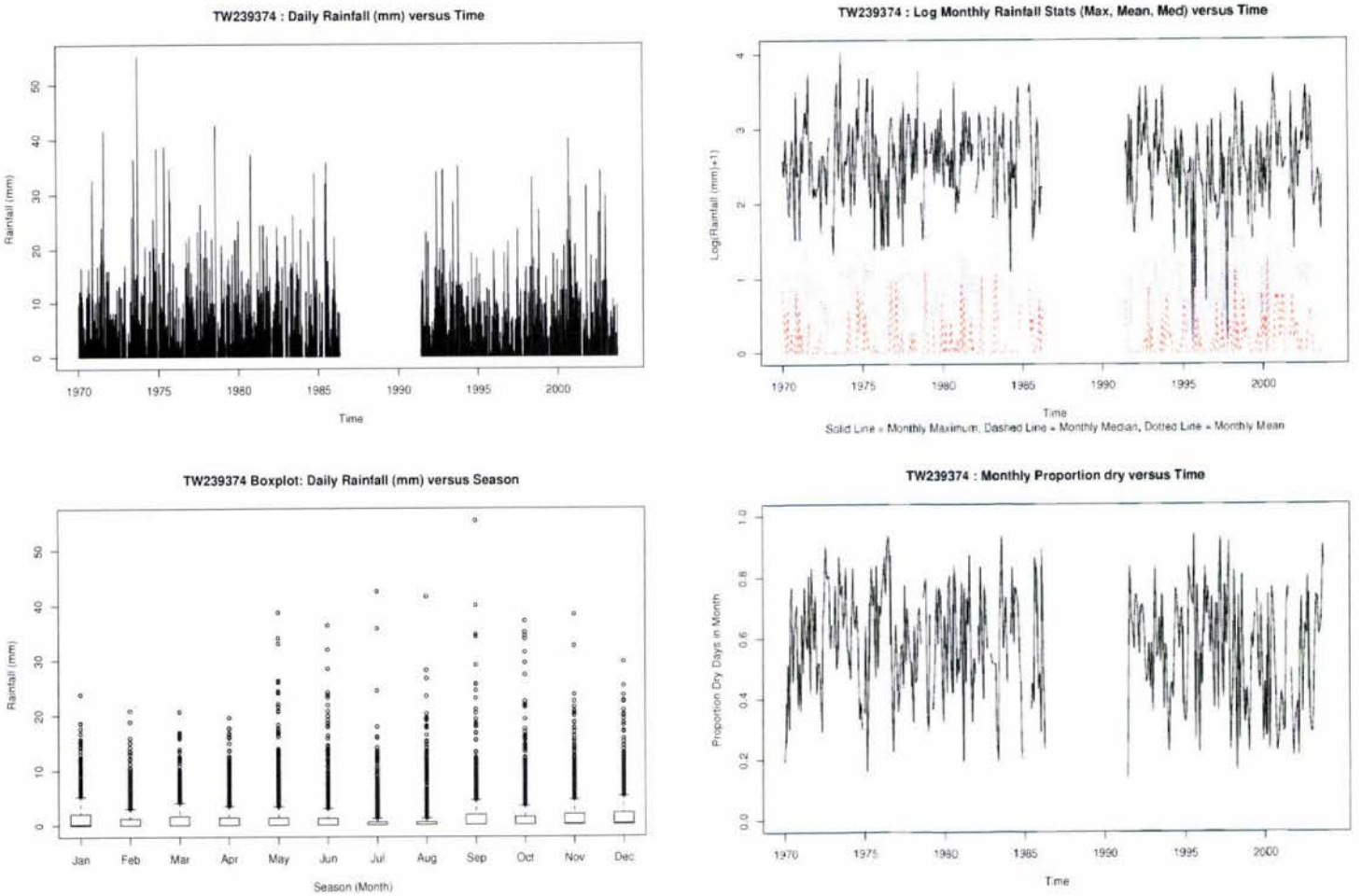


Figure A.9: Site TW239374 daily plots

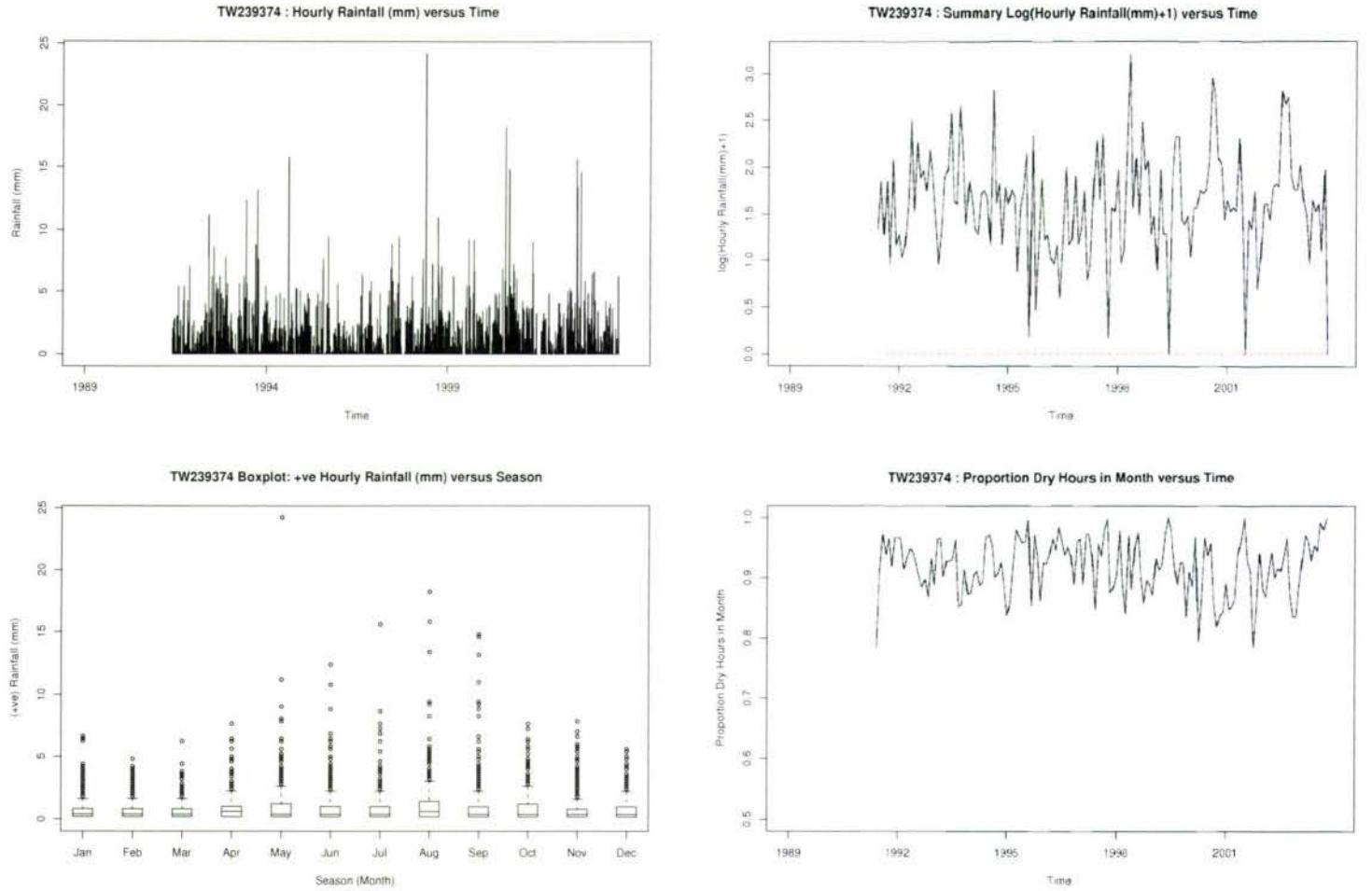


Figure A.10: Site TW239374 hourly plots

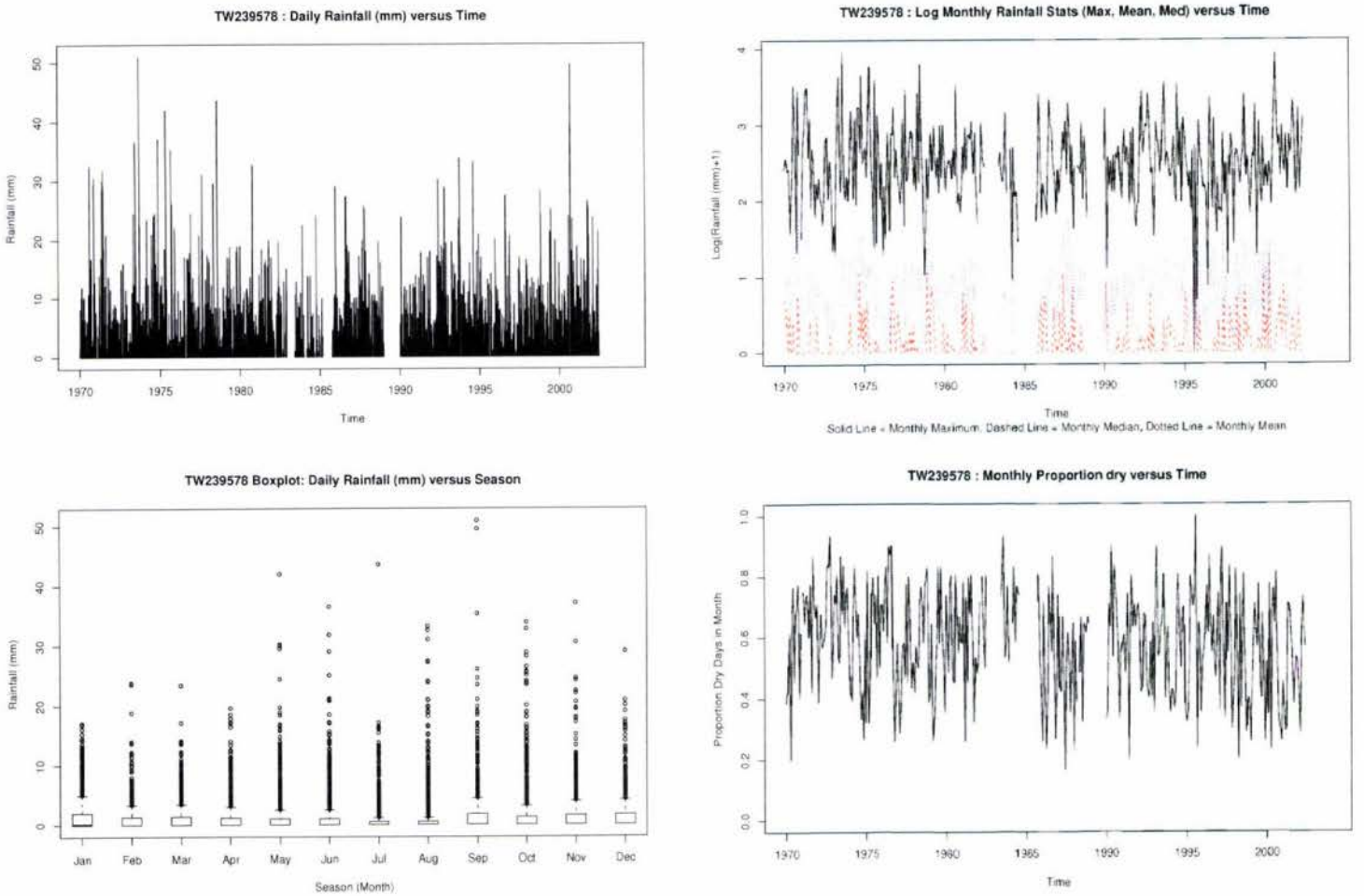


Figure A.11: Site TW239578 daily plots

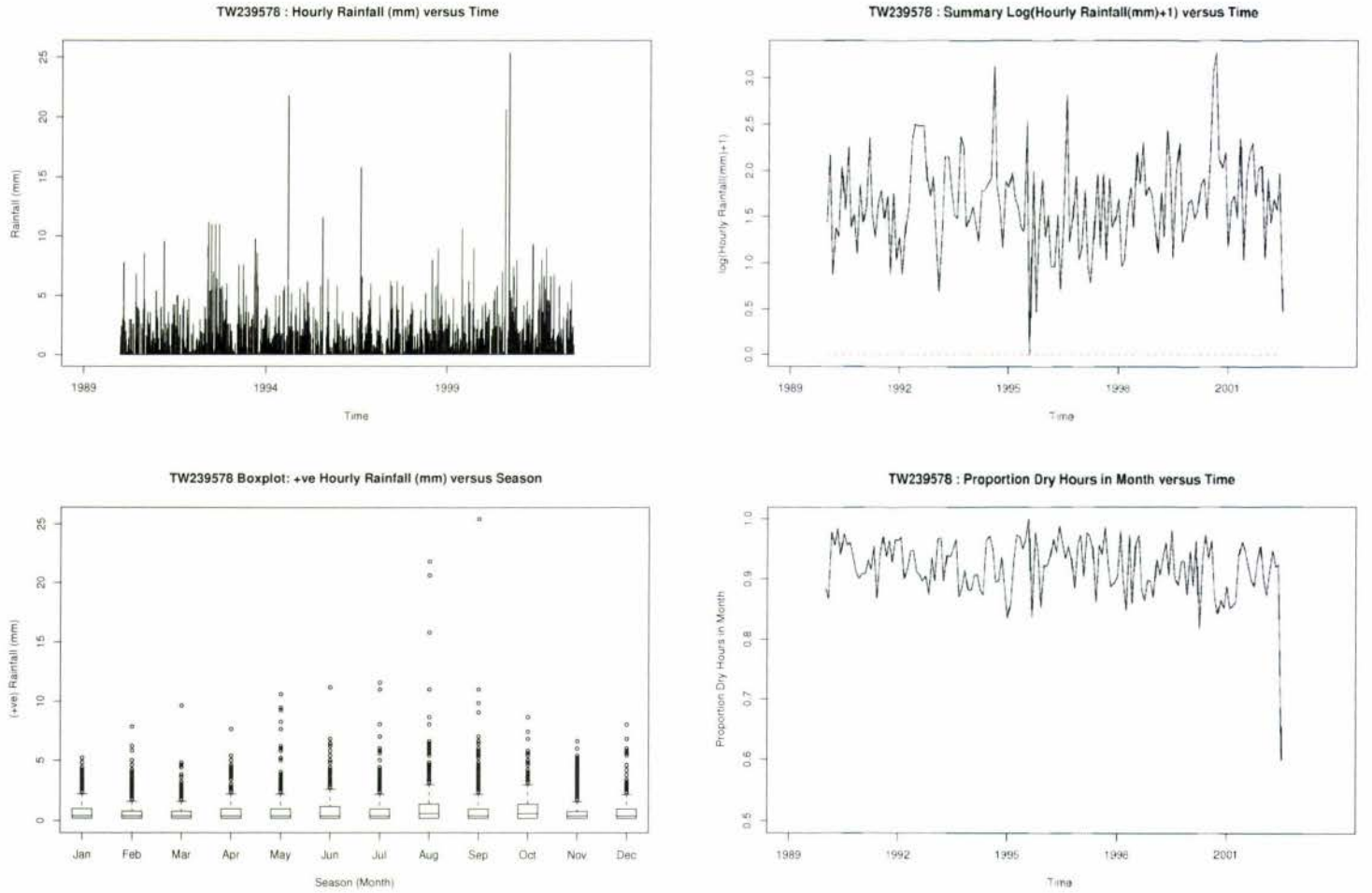


Figure A.12: Site TW239578 hourly plots

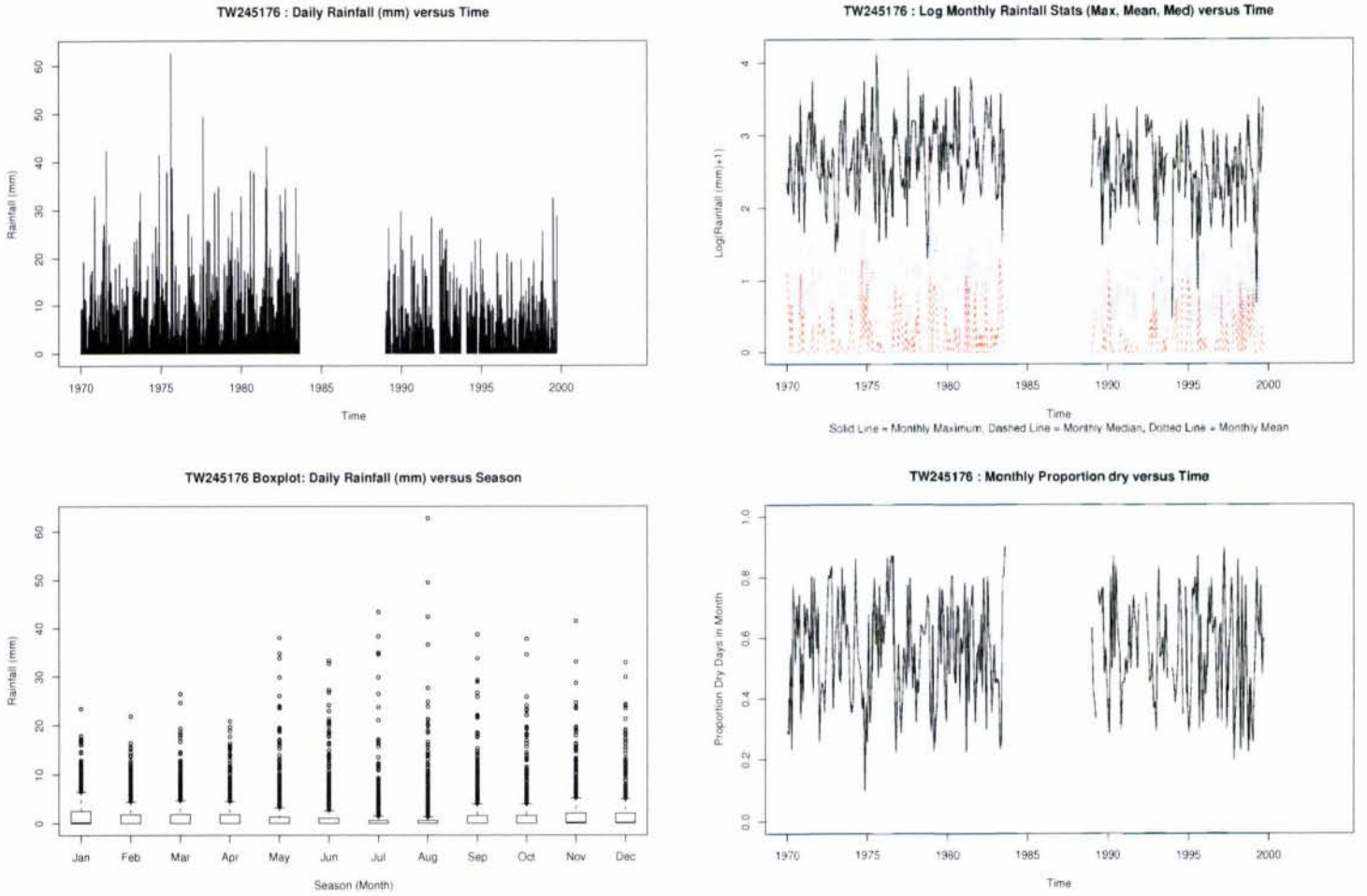


Figure A.13: Site TW245176 daily plots

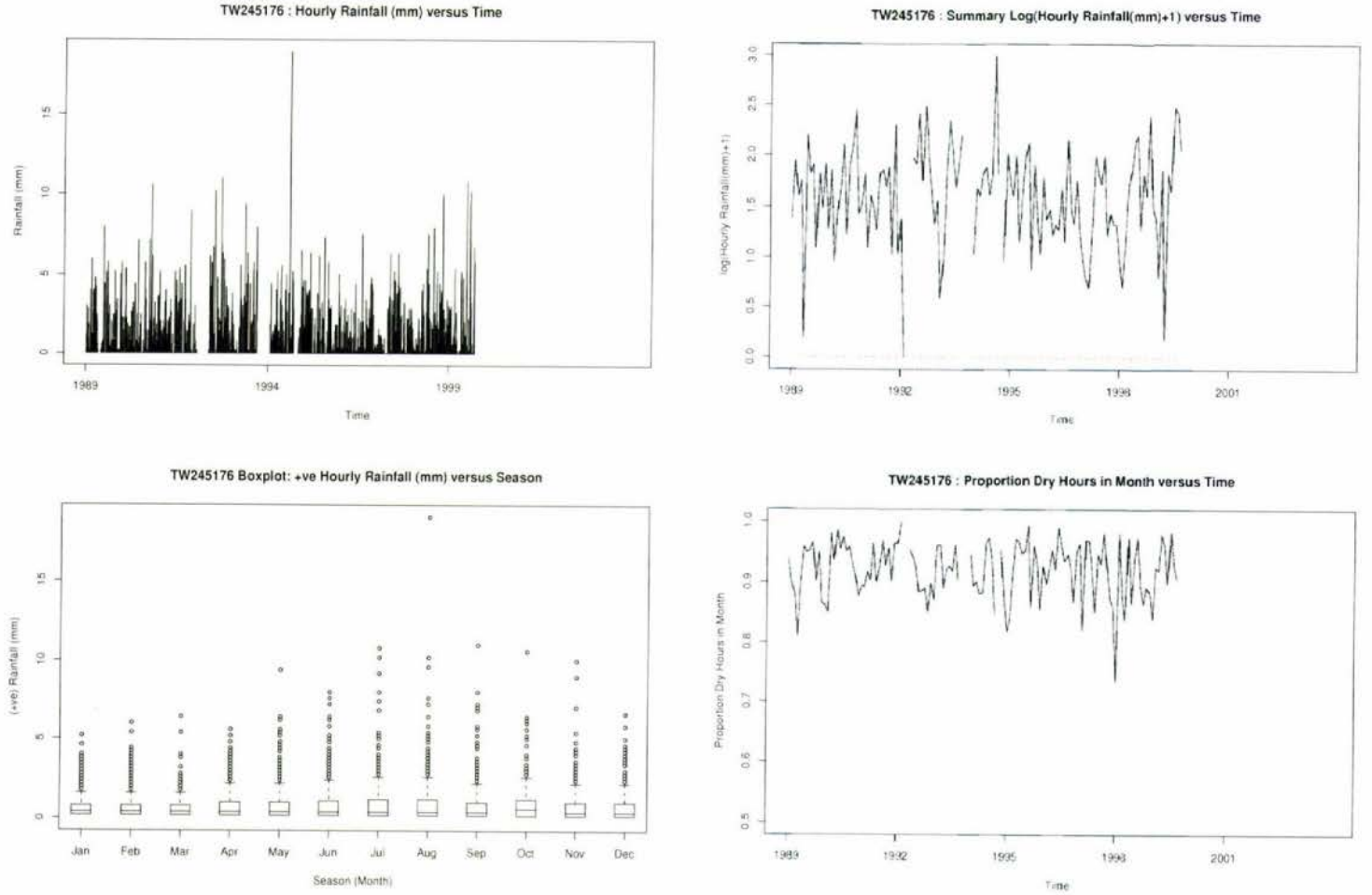


Figure A.14: Site TW245176 hourly plots

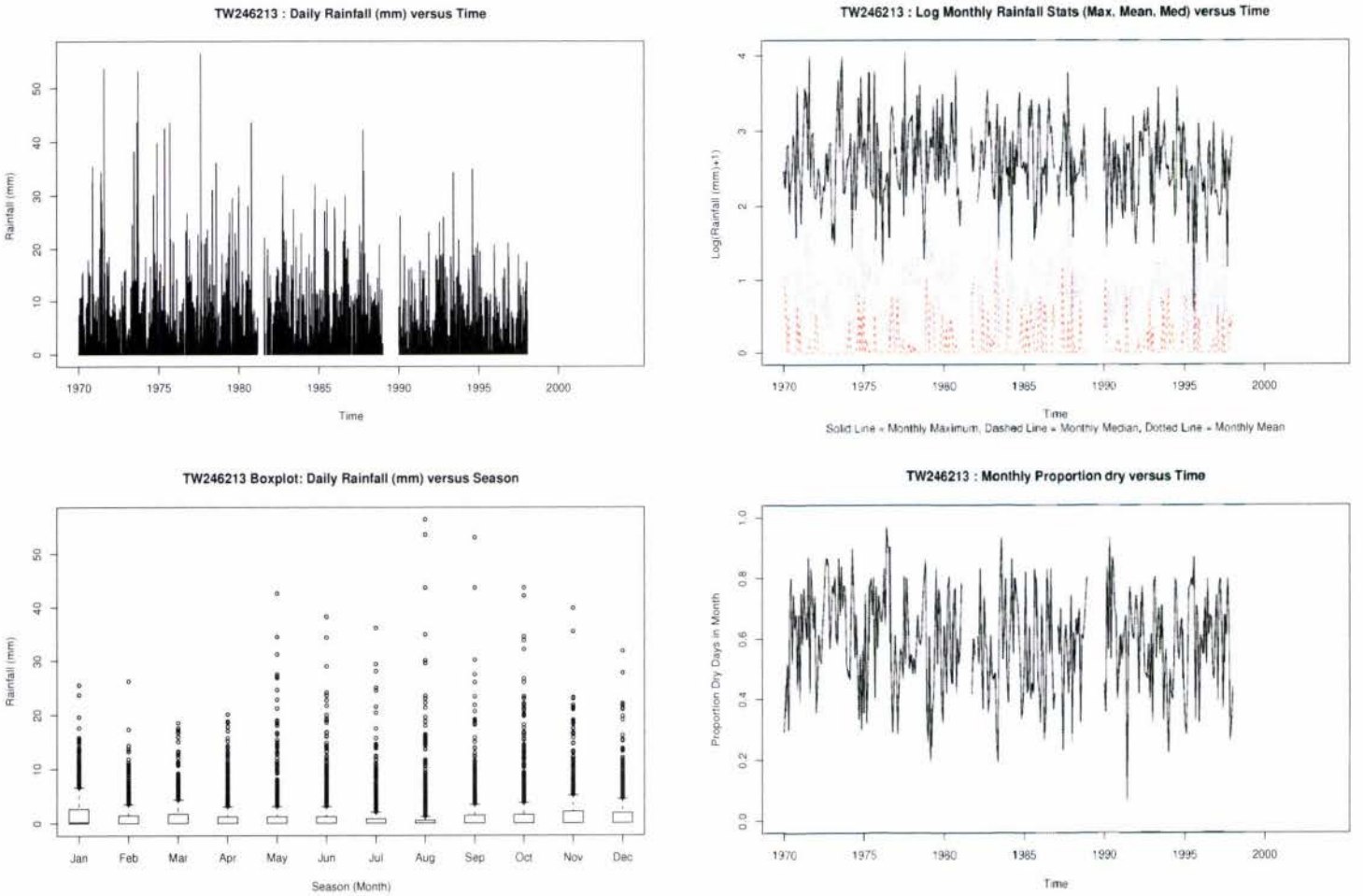


Figure A.15: Site TW246213 daily plots

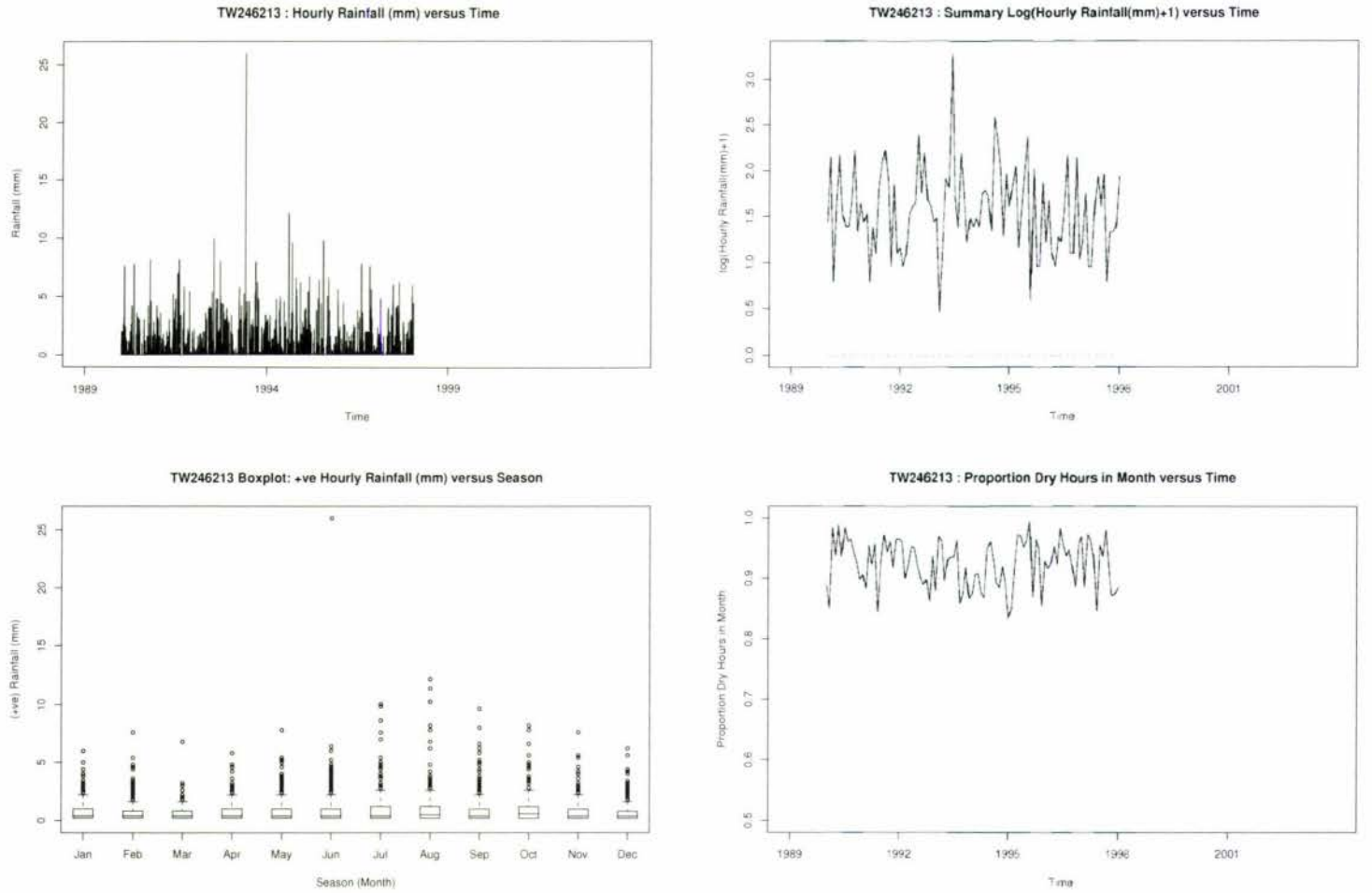


Figure A.16: Site TW246213 hourly plots

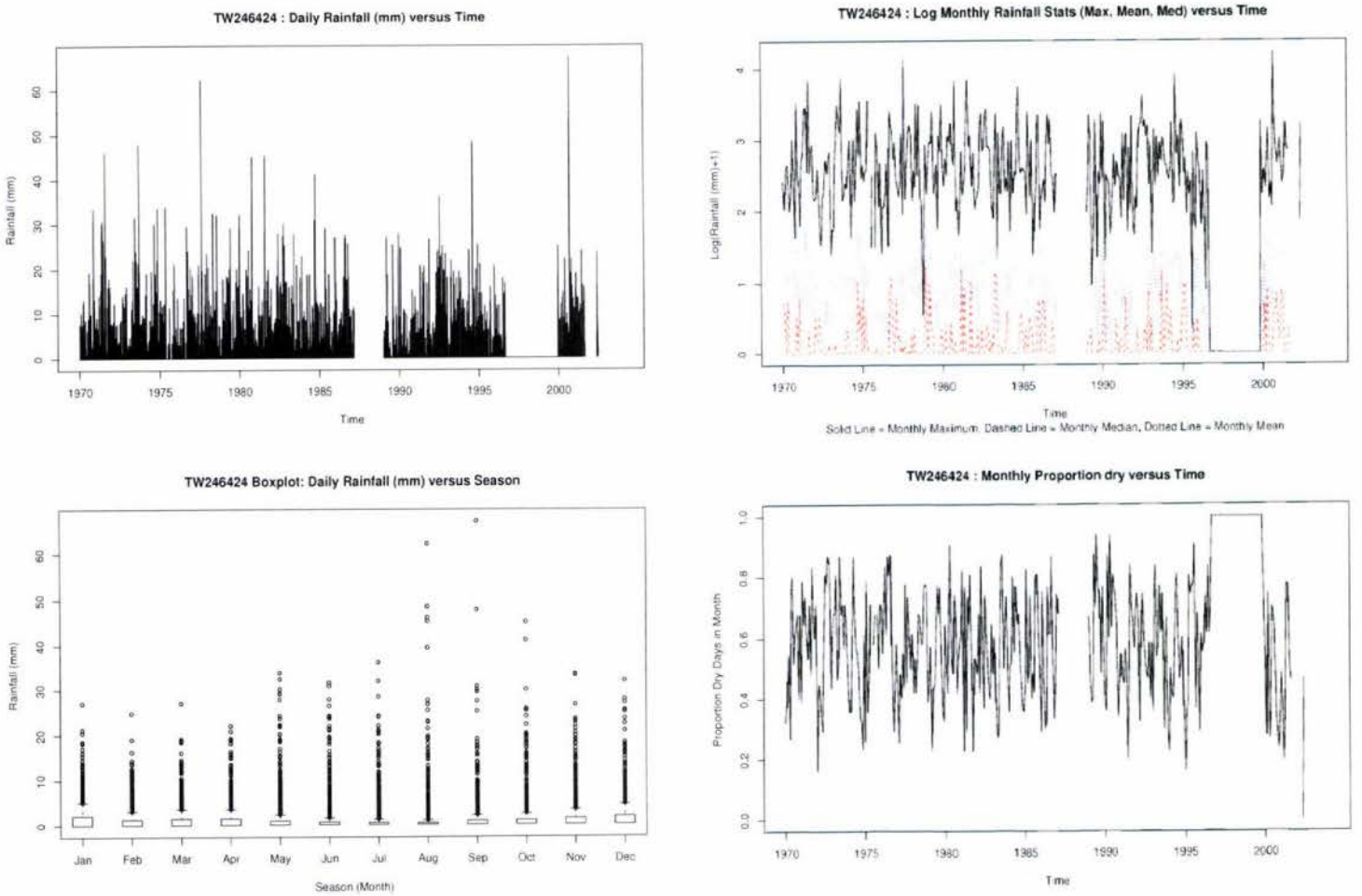


Figure A.17: Site TW246424 daily plots

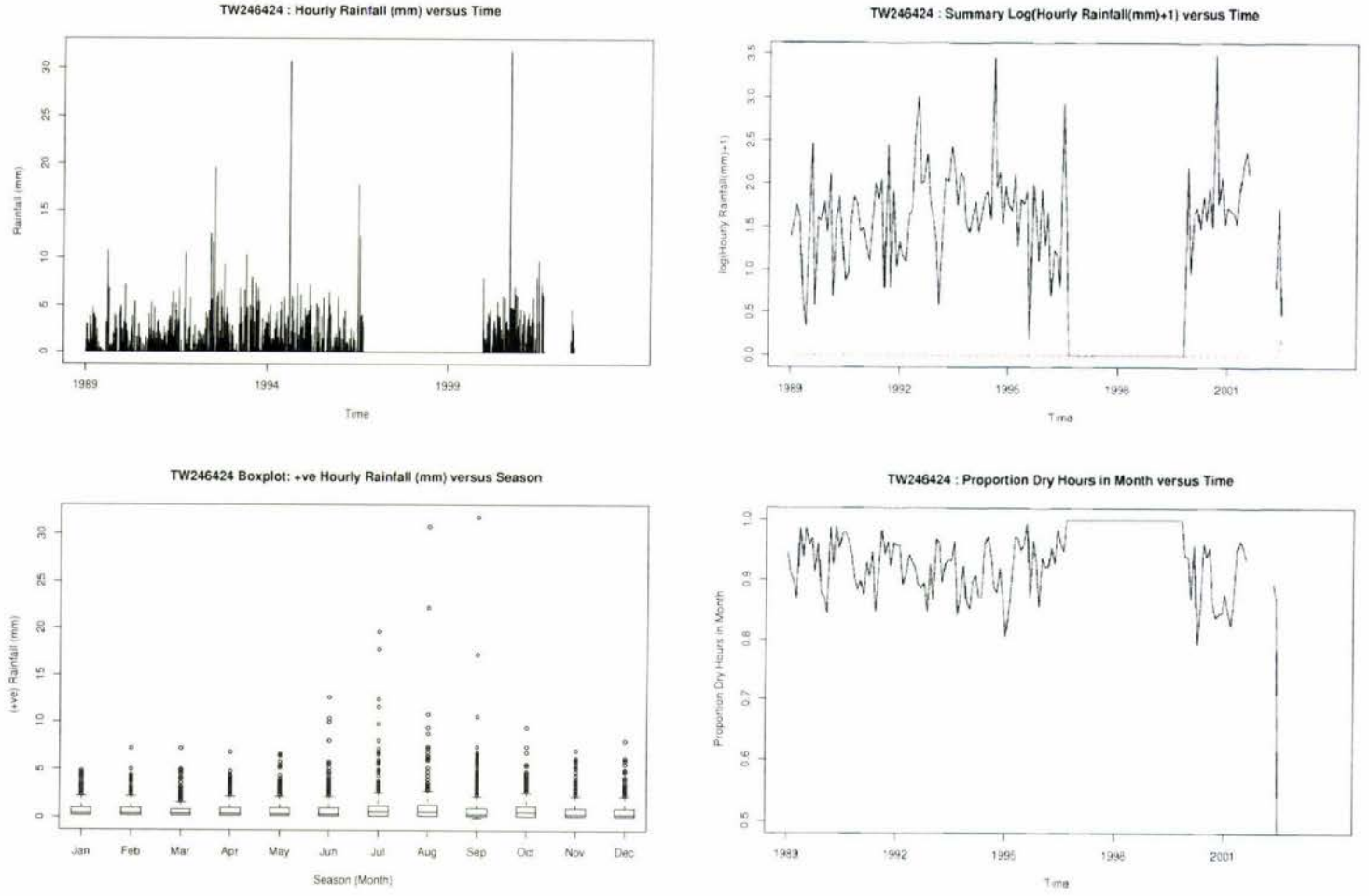


Figure A.18: Site TW246424 hourly plots

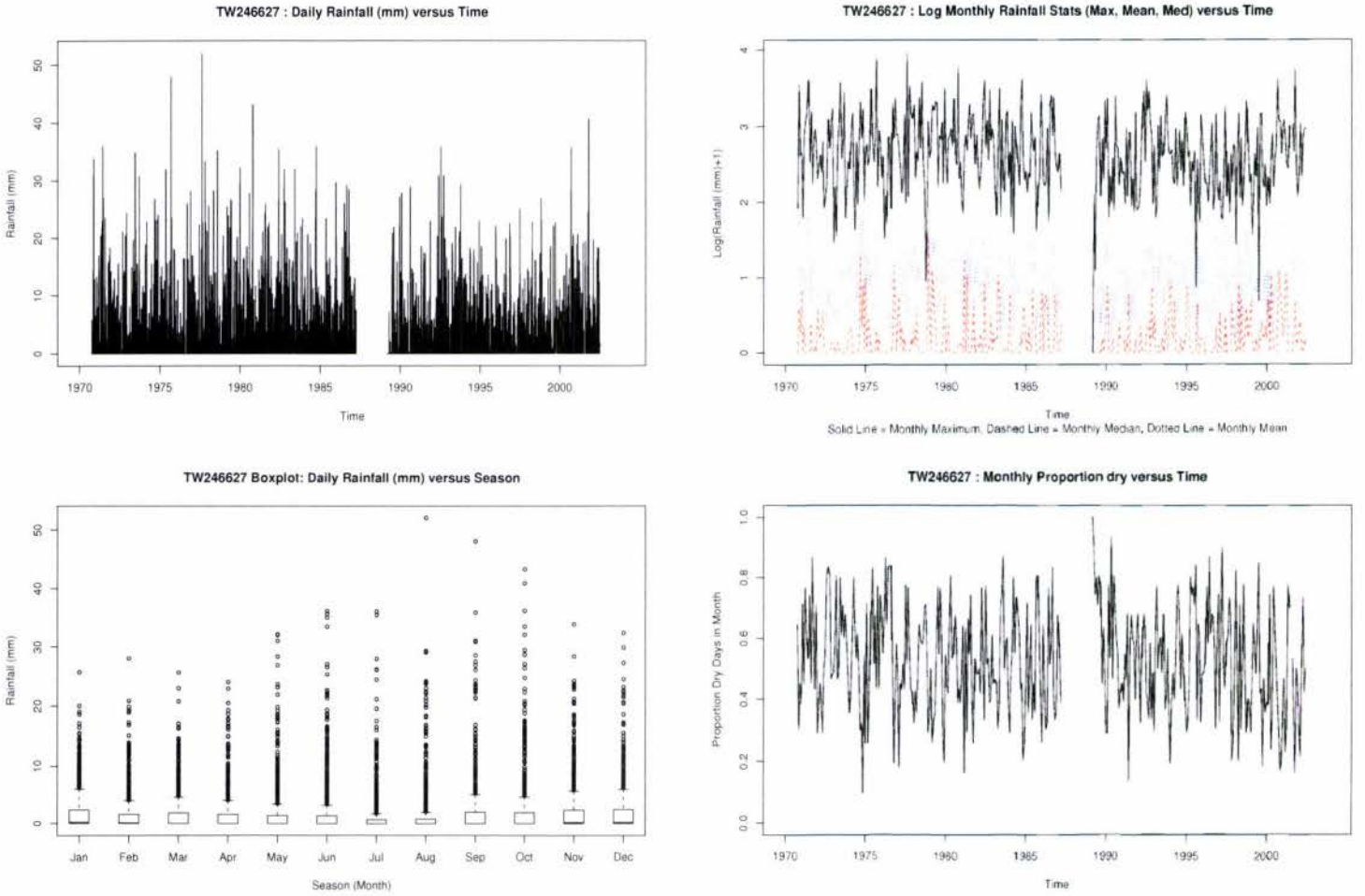


Figure A.19: Site TW246627 daily plots

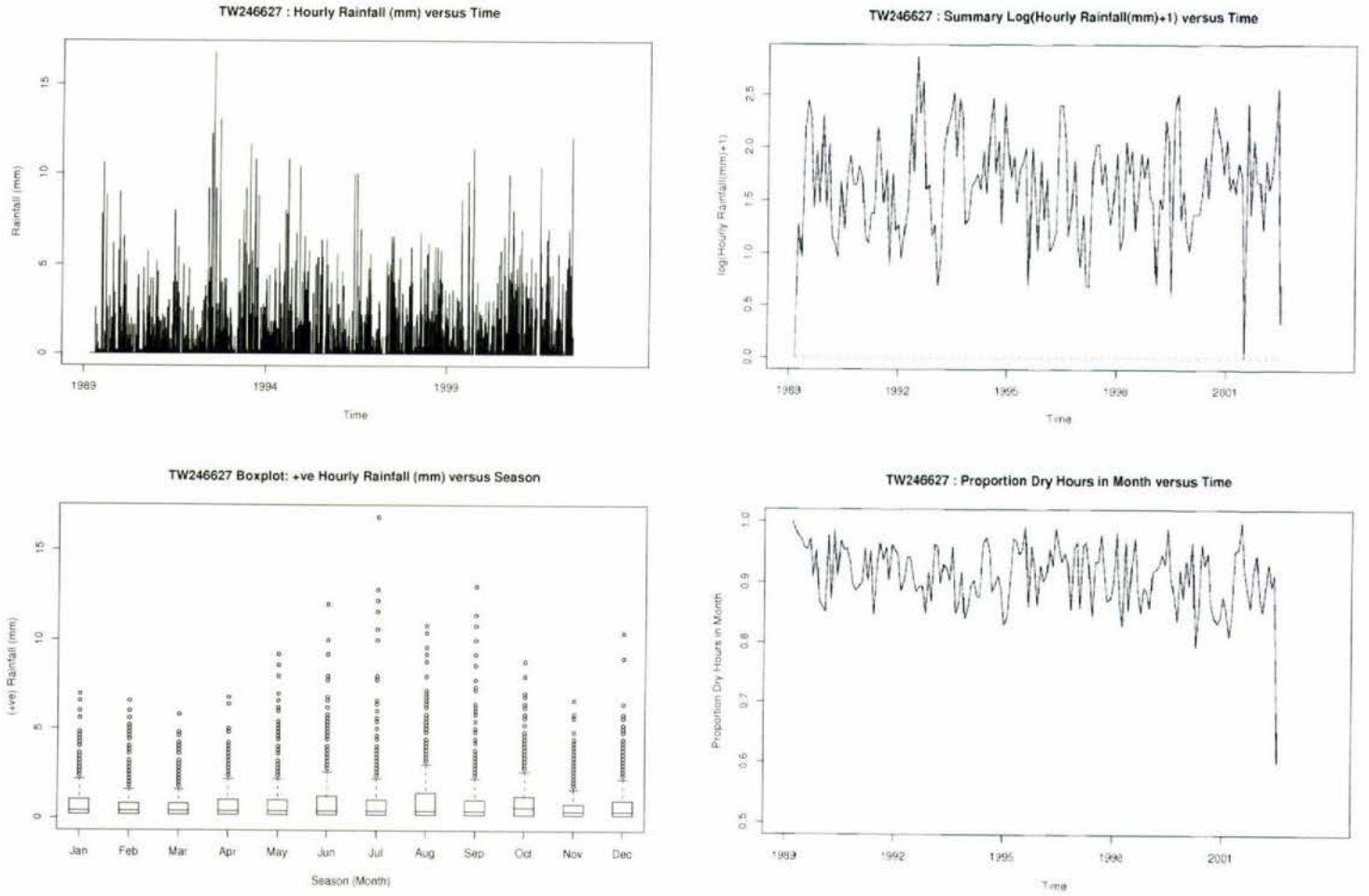


Figure A.20: Site TW246627 hourly plots

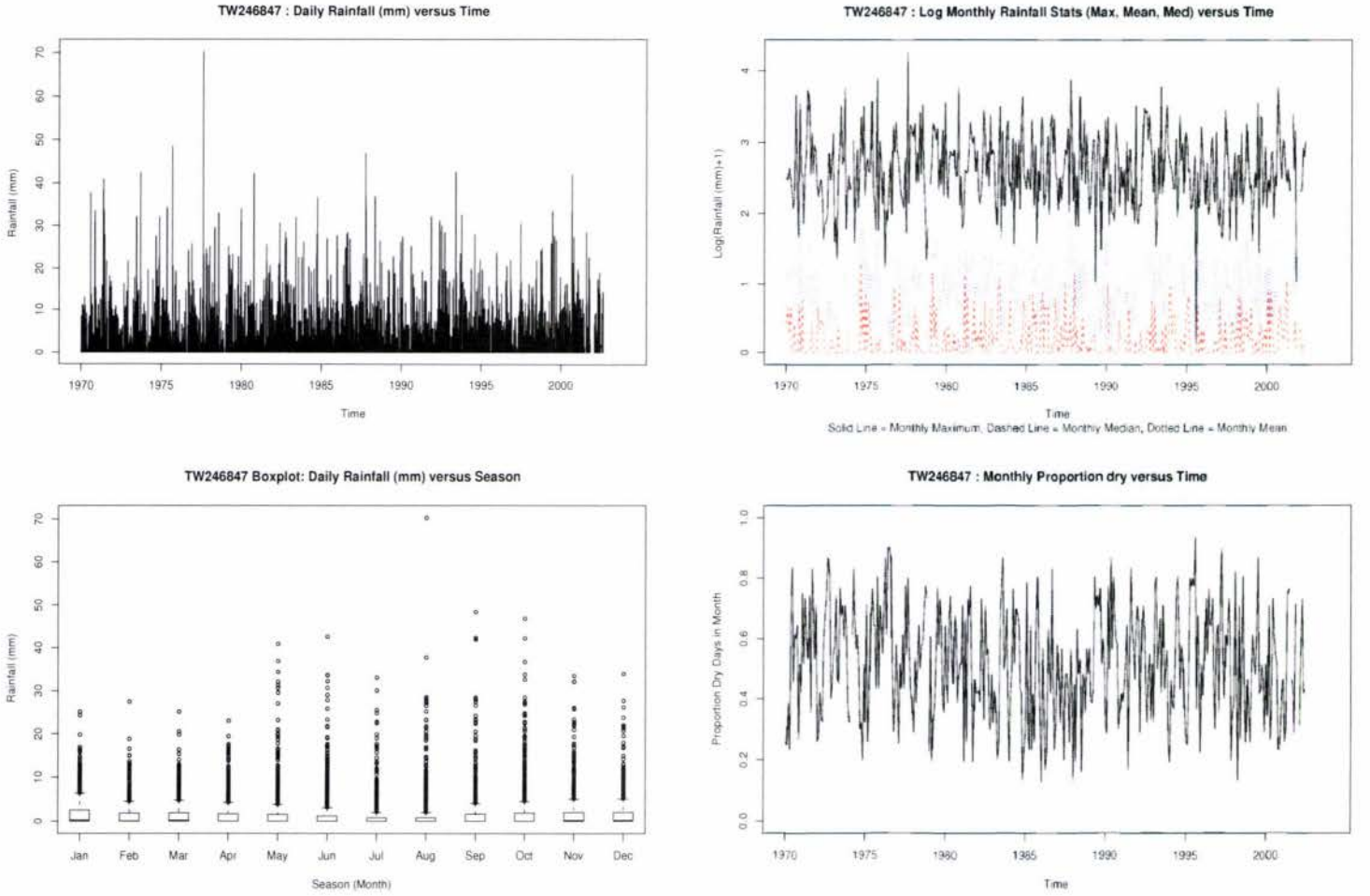


Figure A.21: Site TW246847 daily plots

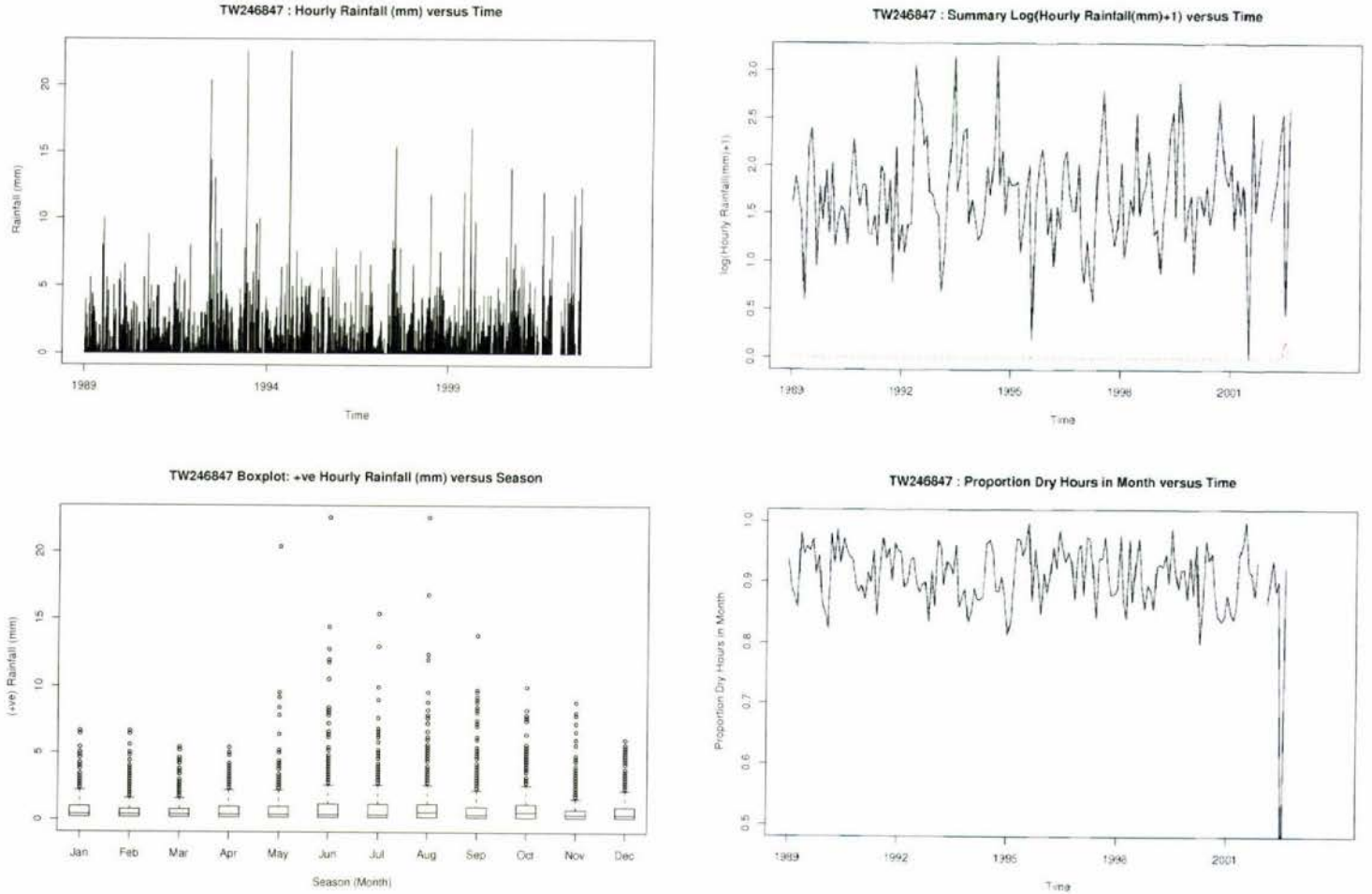


Figure A.22: Site TW246847 hourly plots

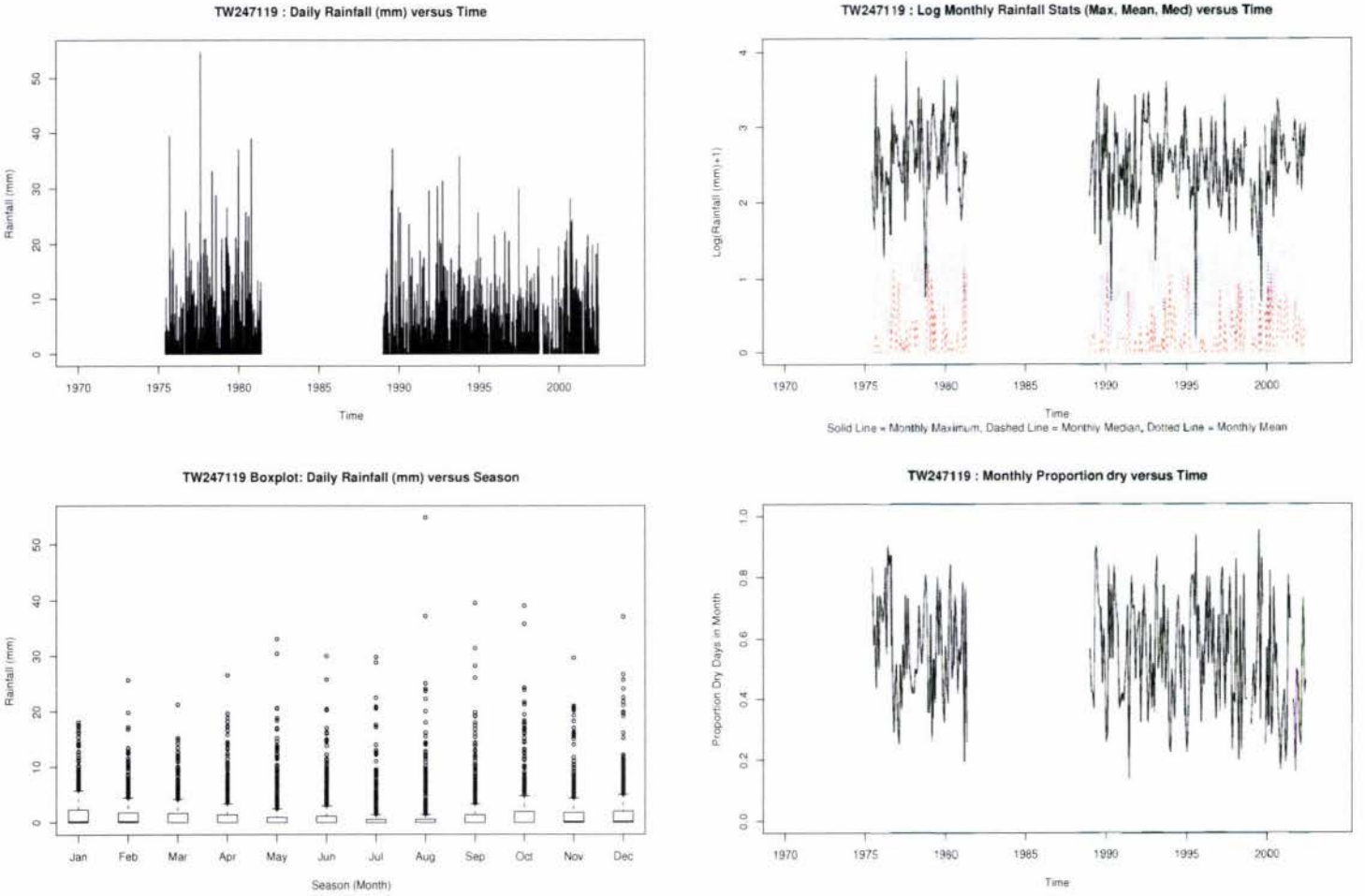


Figure A.23: Site TW247119 daily plots

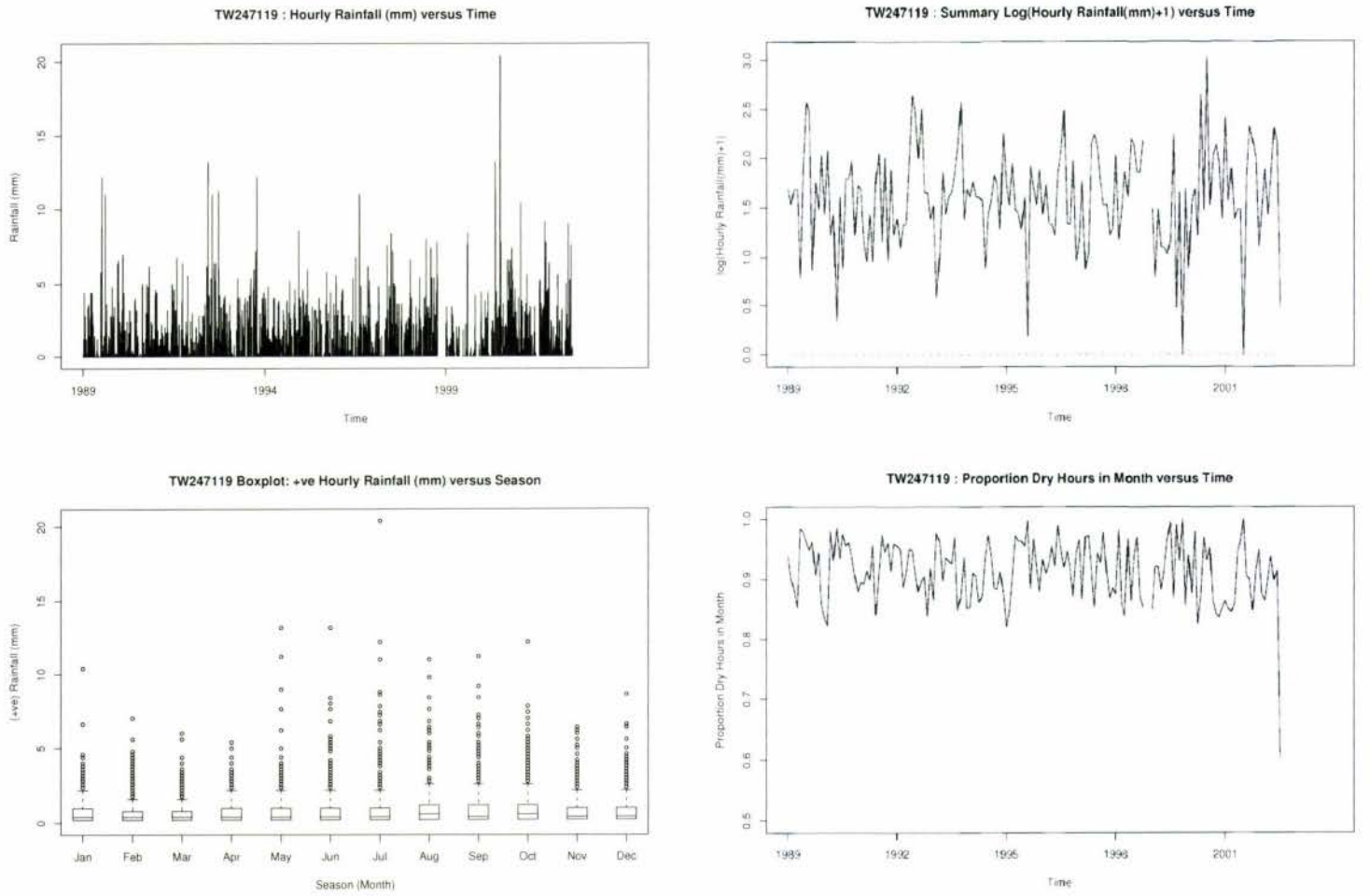


Figure A.24: Site TW247119 hourly plots

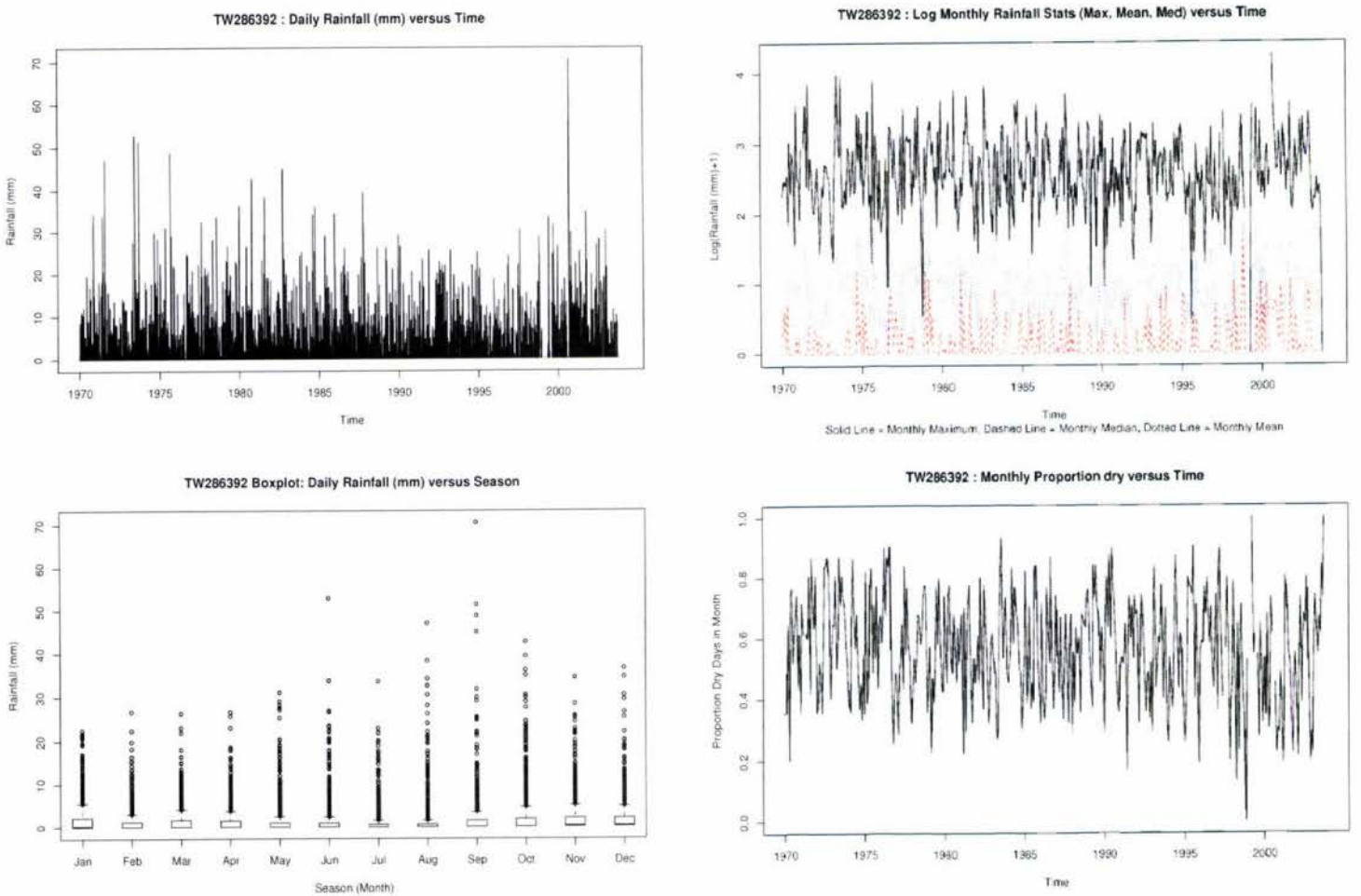


Figure A.25: Site TW286392 daily plots

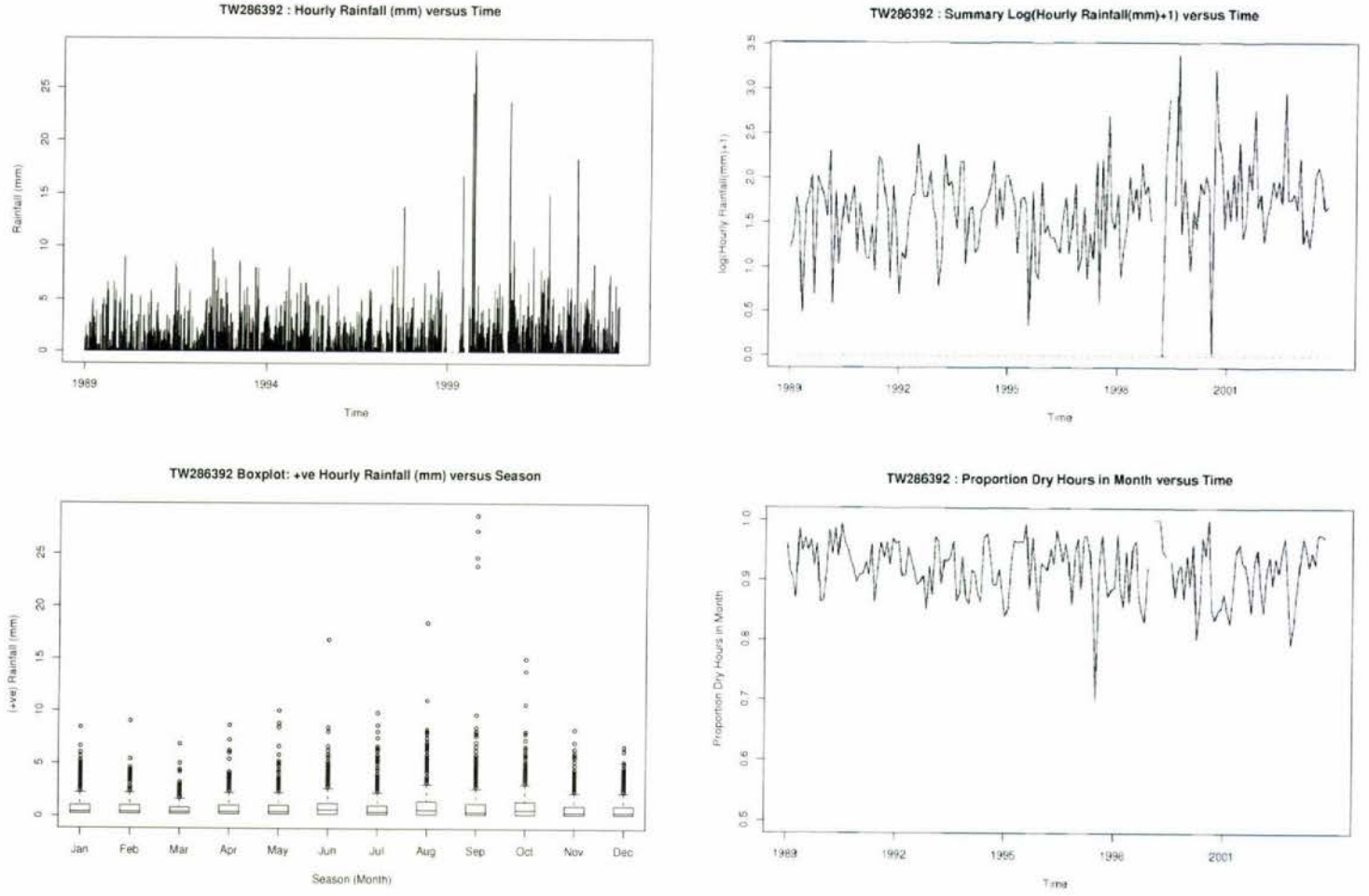


Figure A.26: Site TW286392 hourly plots

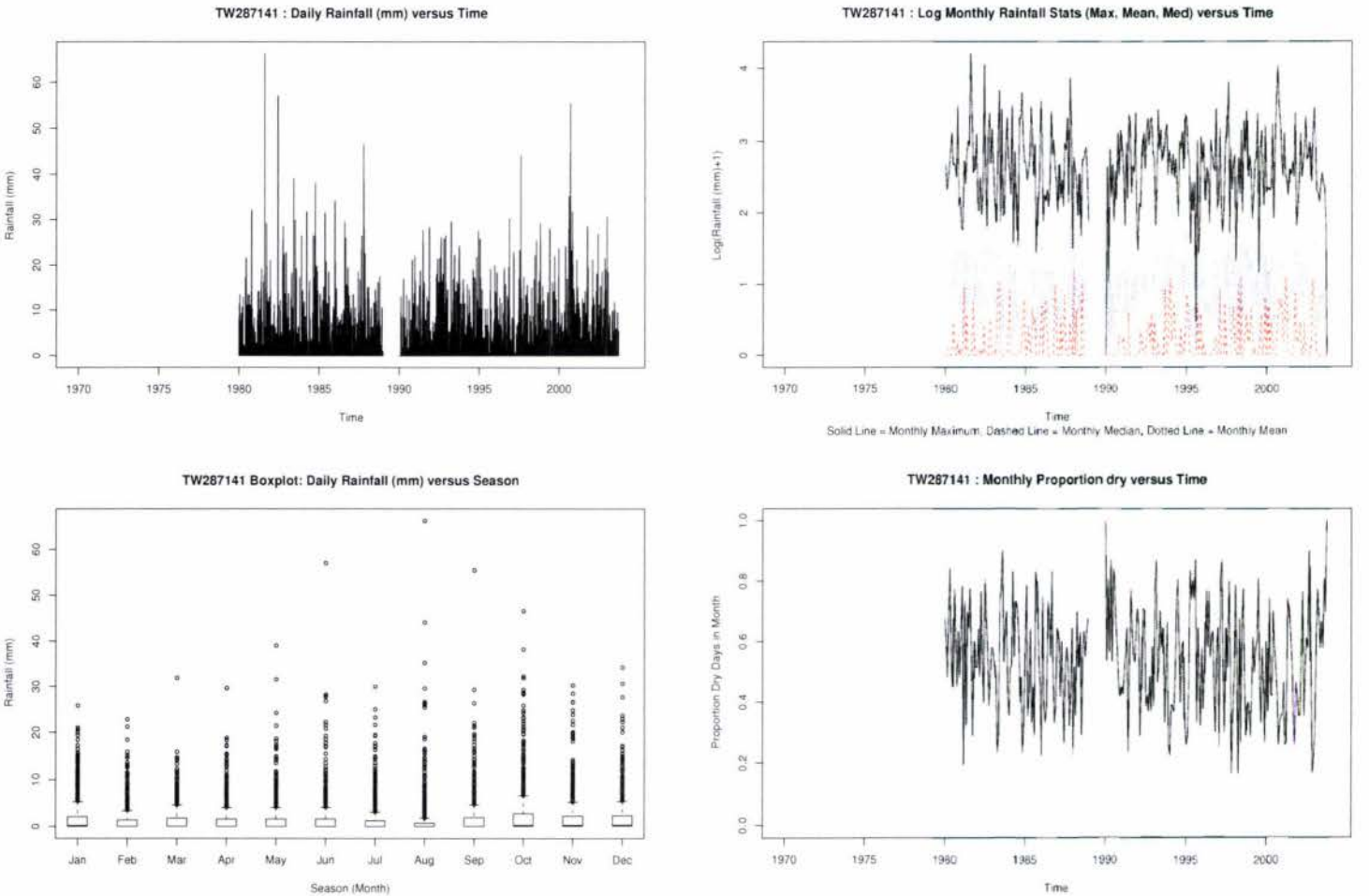


Figure A.27: Site TW287141 daily plots

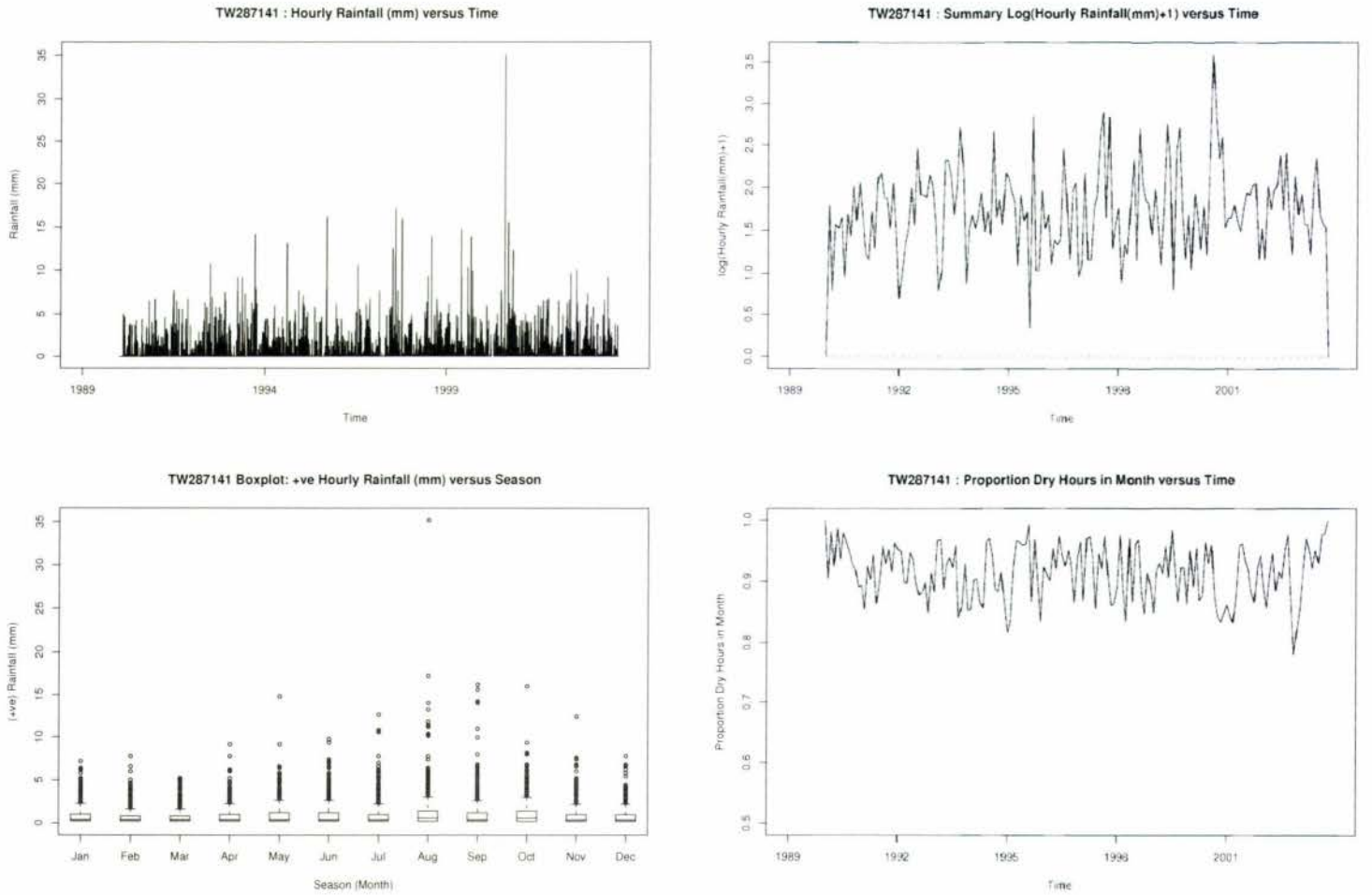


Figure A.28: Site TW287141 hourly plots

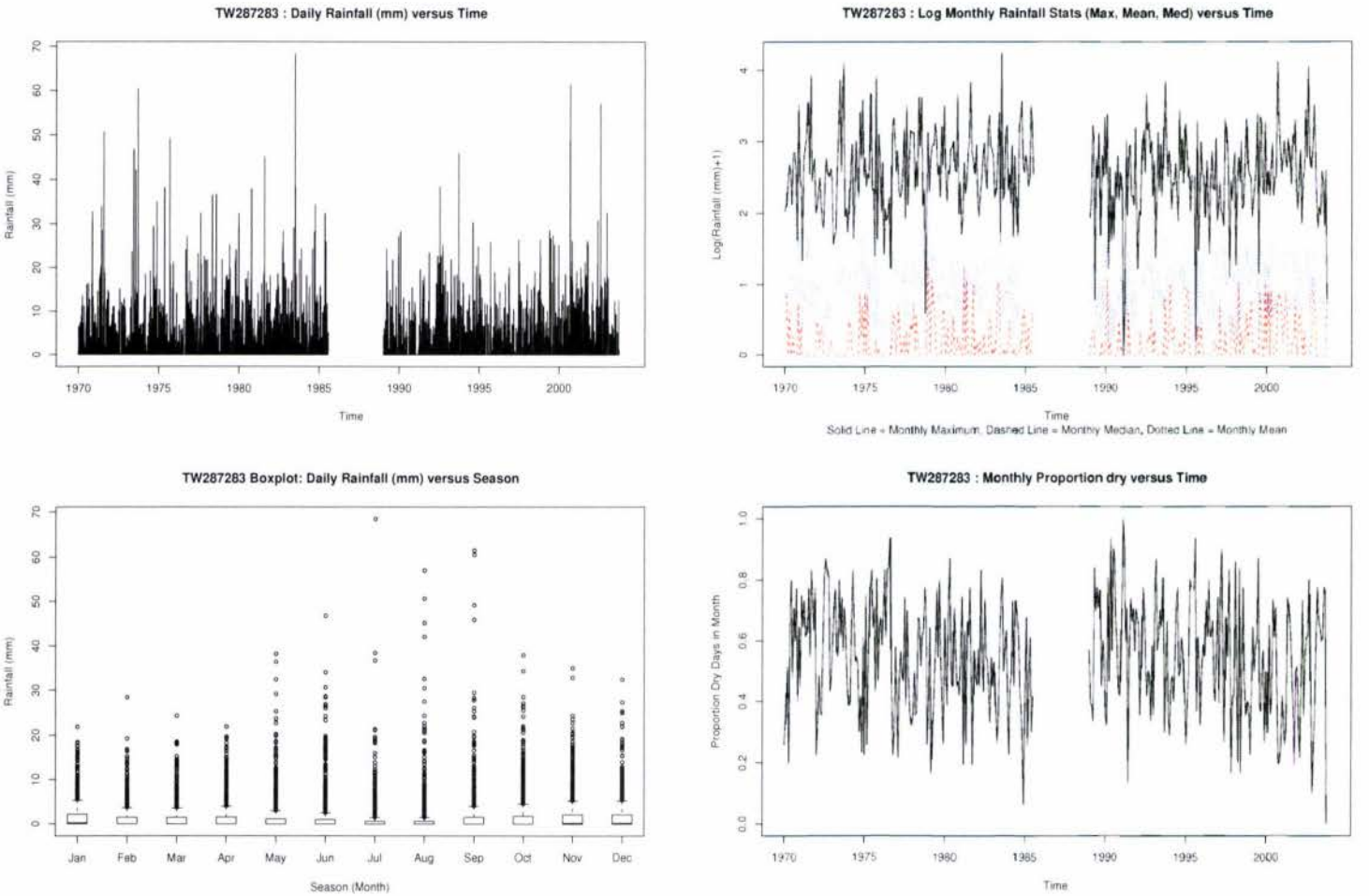


Figure A.29: Site TW287283 daily plots

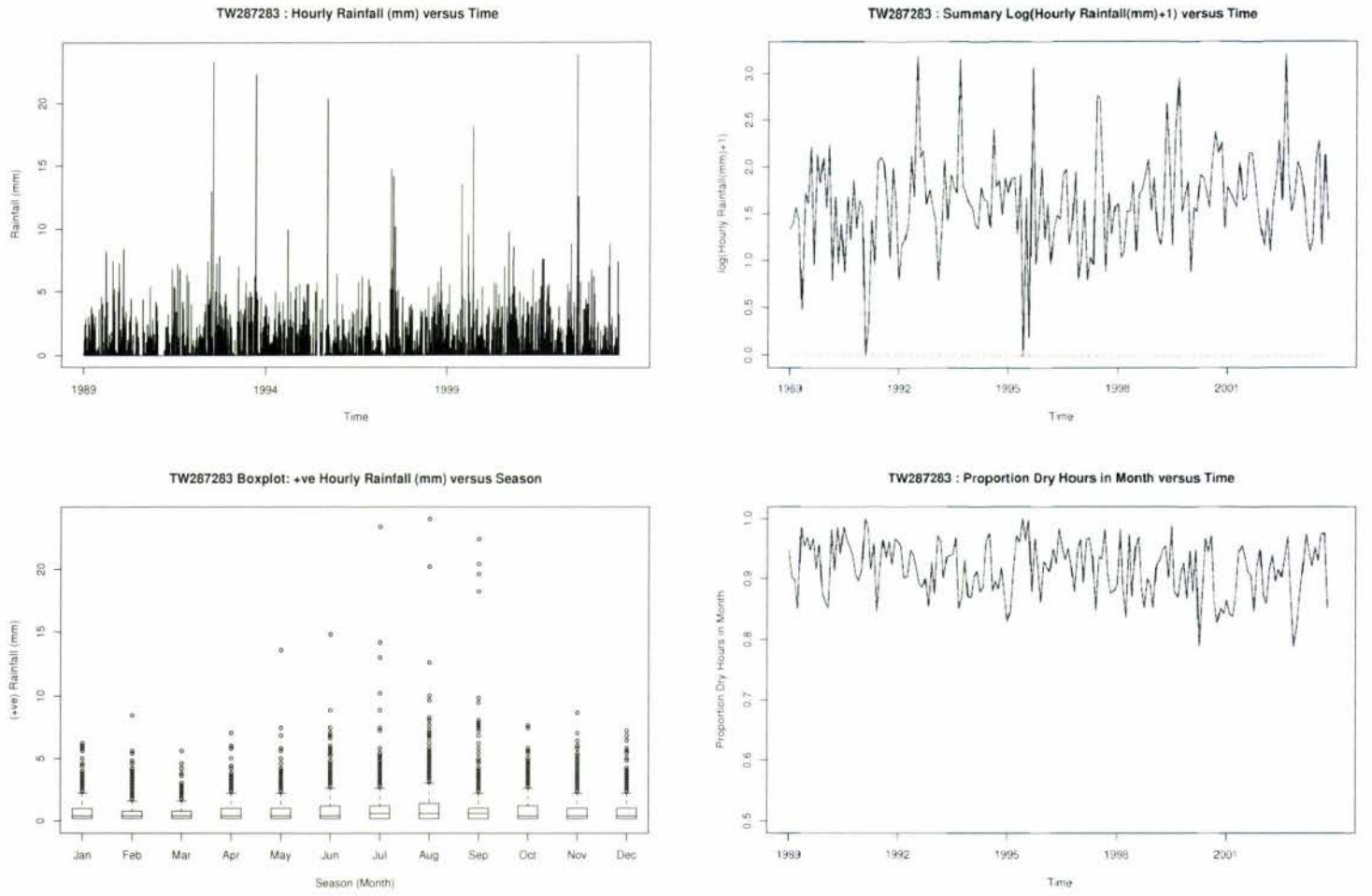


Figure A.30: Site TW287283 hourly plots

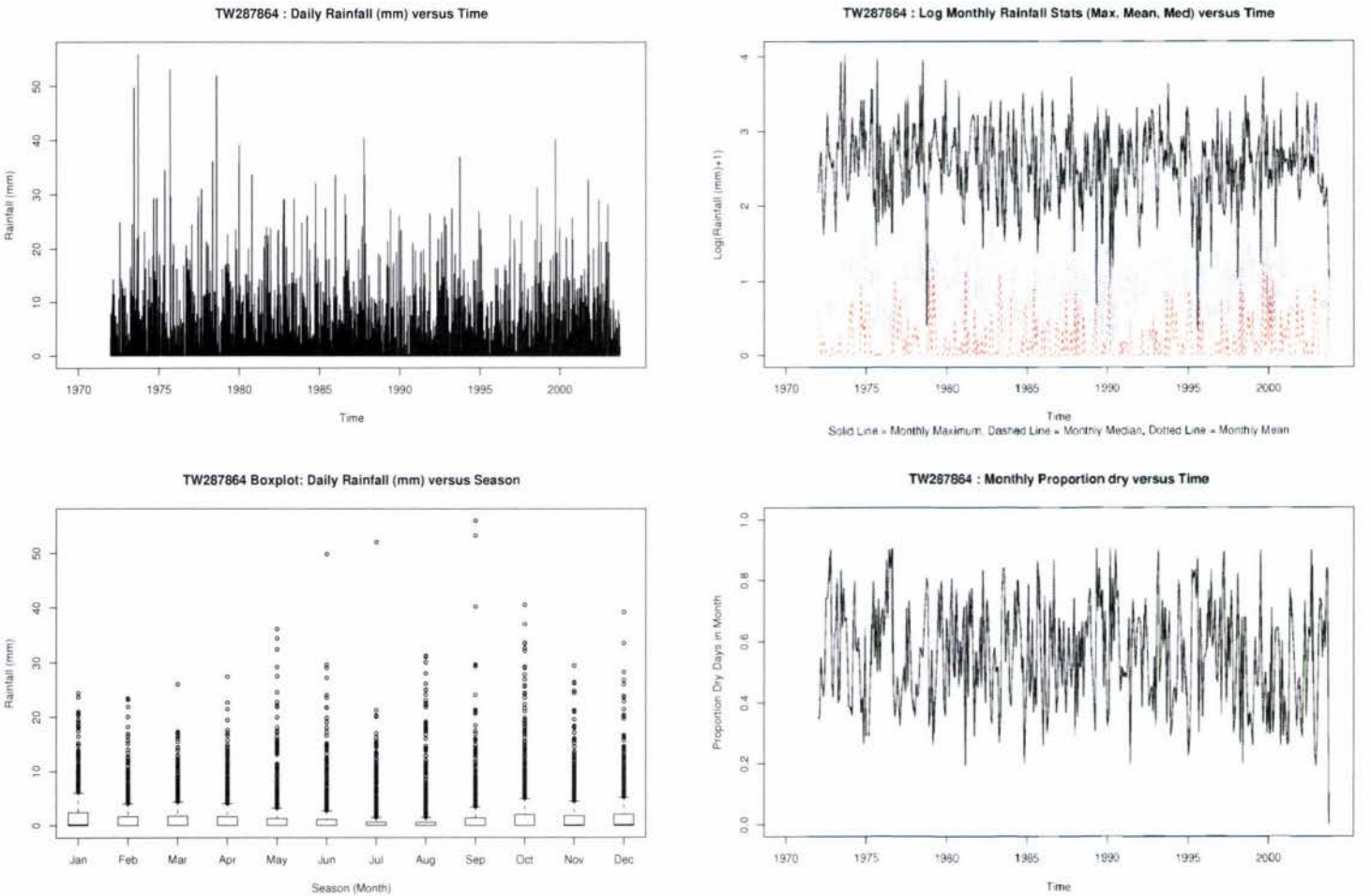


Figure A.31: Site TW287864 daily plots

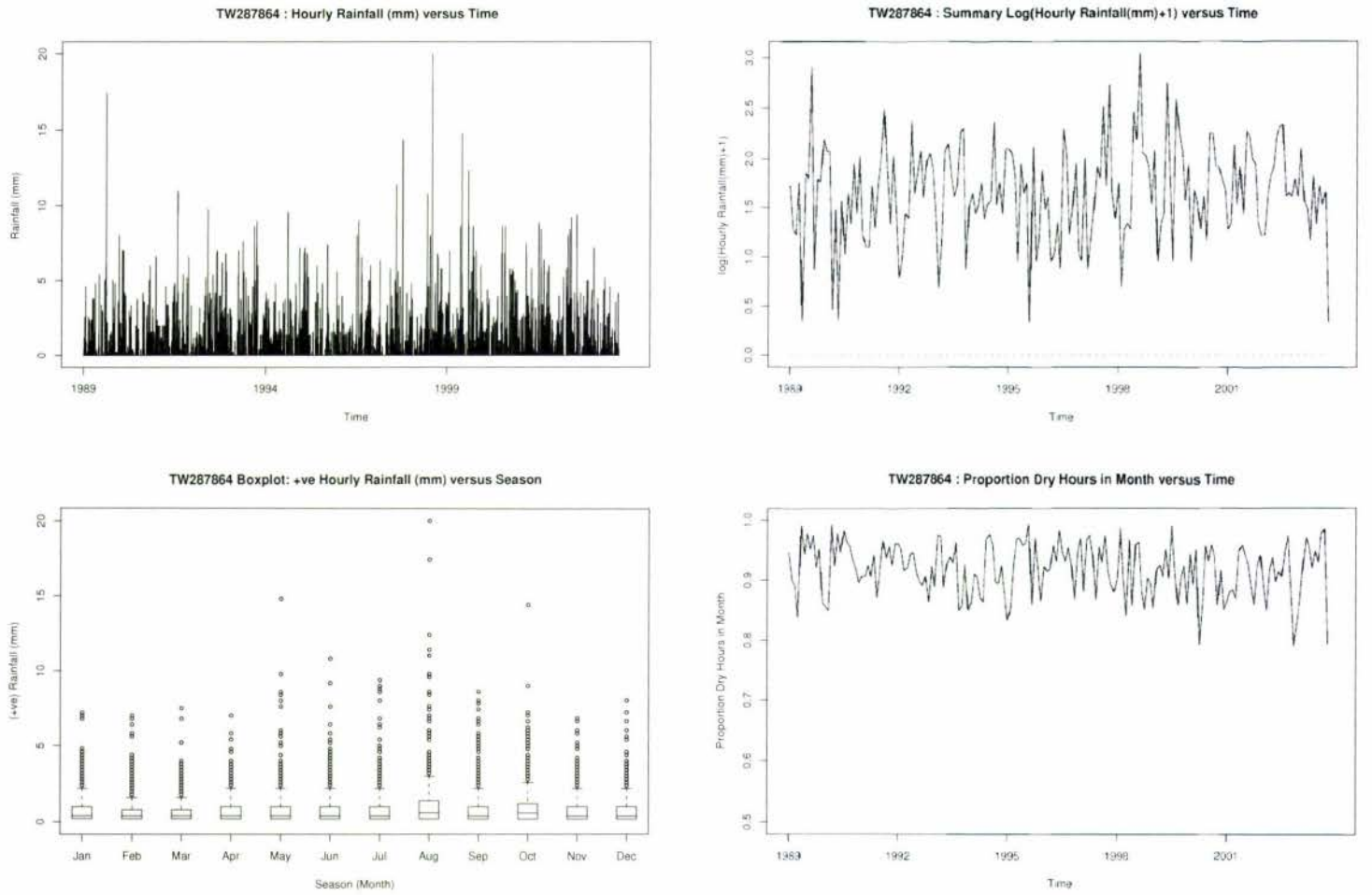


Figure A.32: Site TW287864 hourly plots

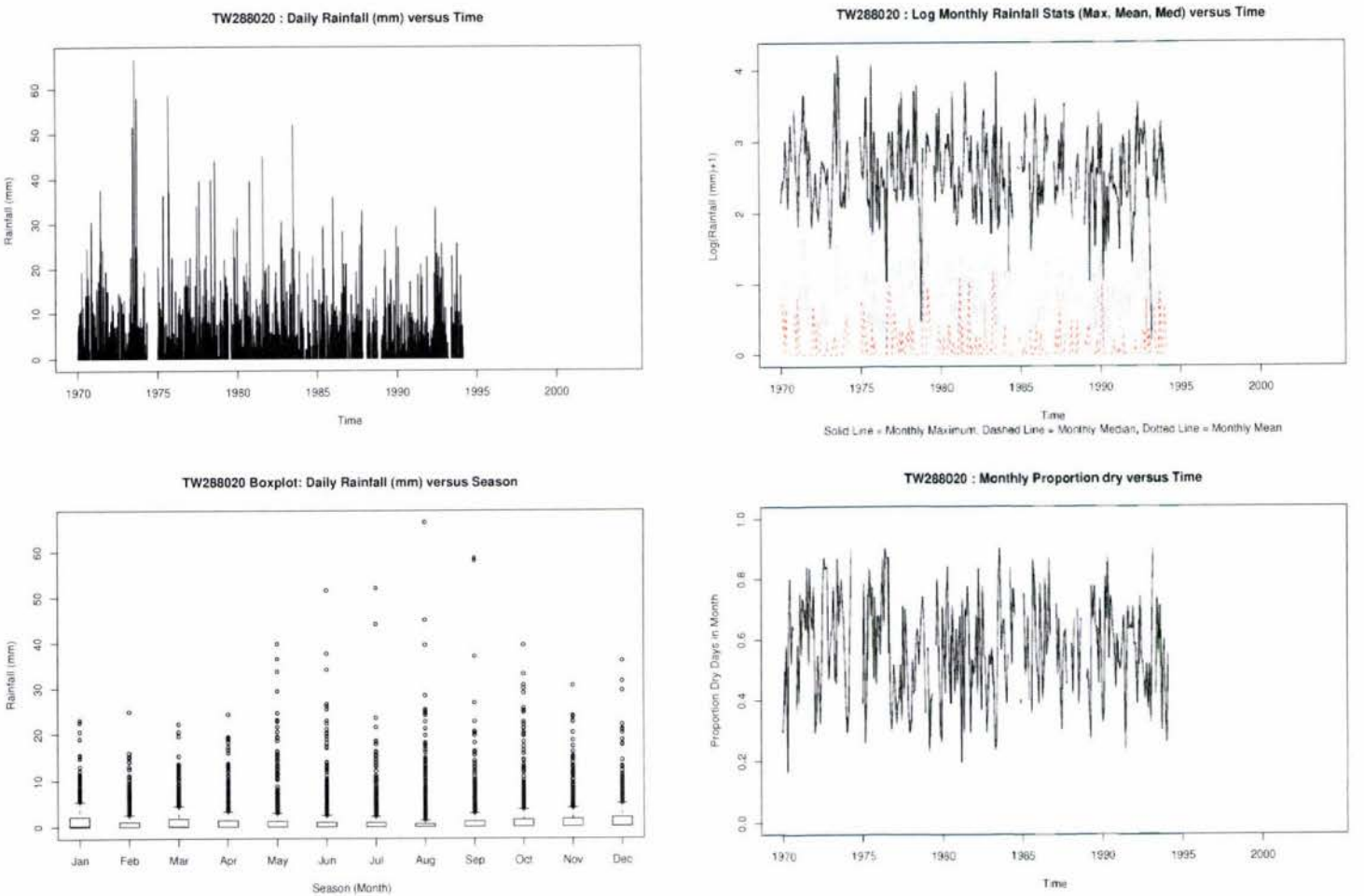


Figure A.33: Site TW288020 daily plots

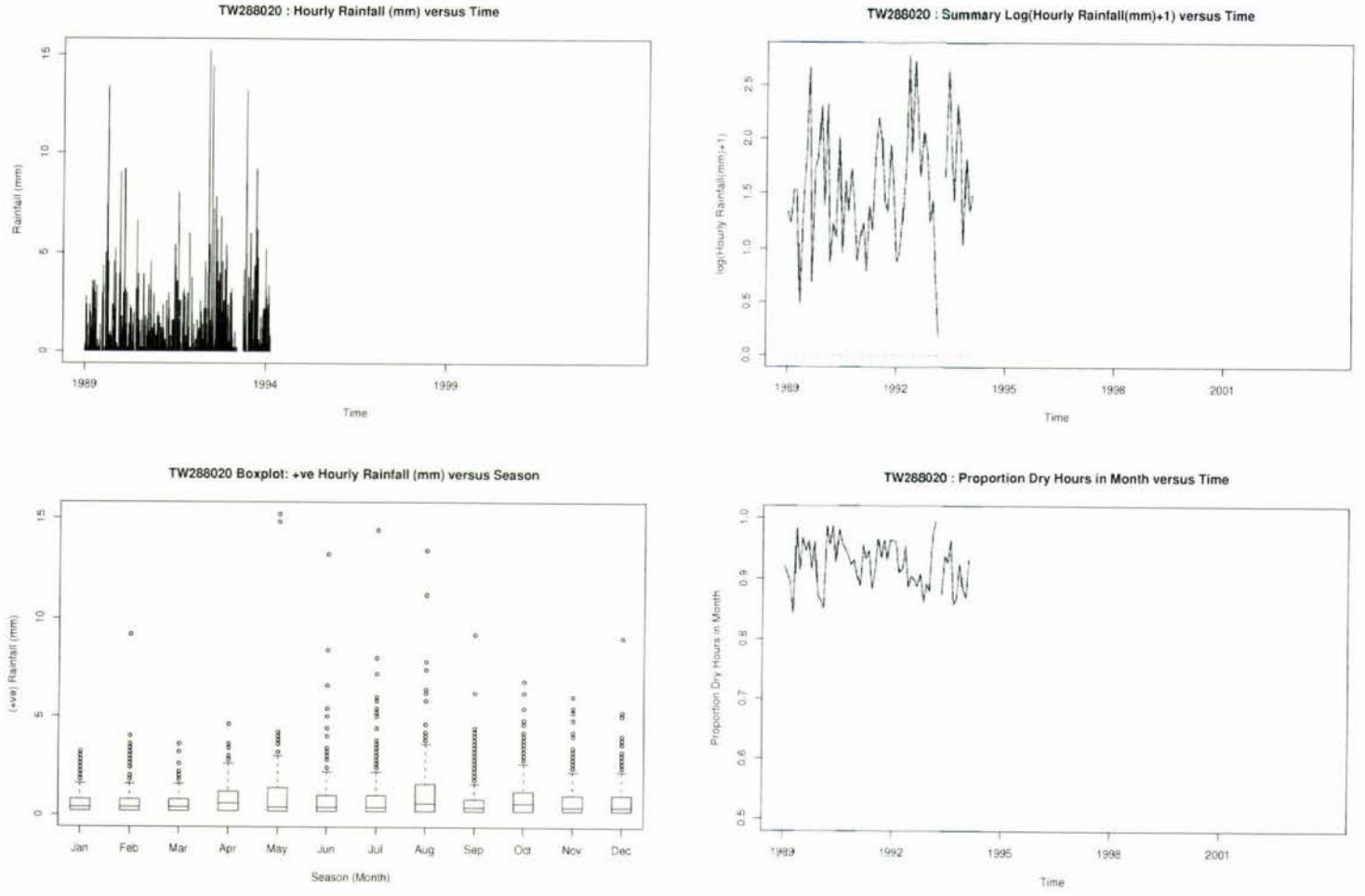


Figure A.34: Site TW288020 hourly plots

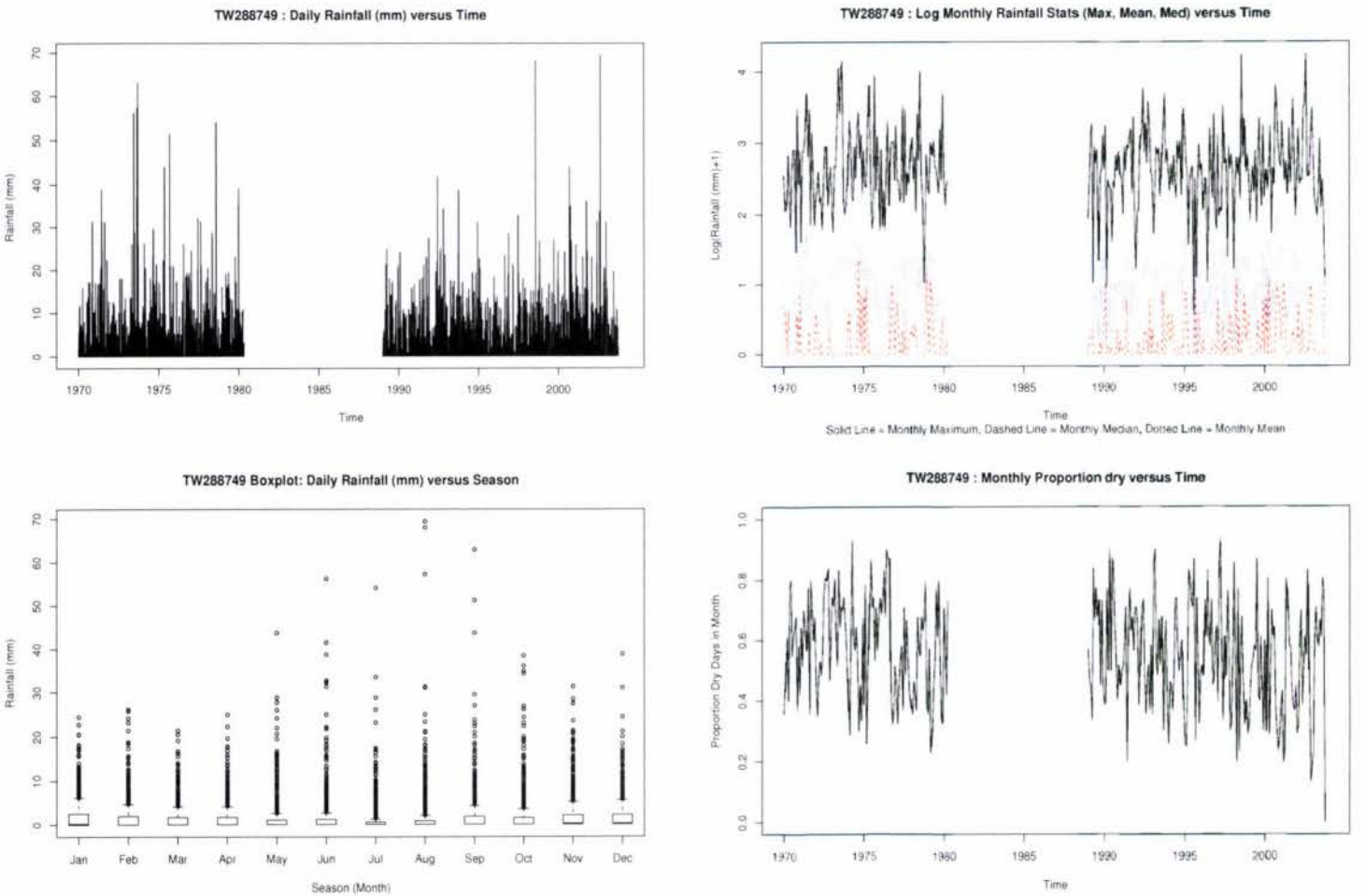


Figure A.35: Site TW288749 daily plots

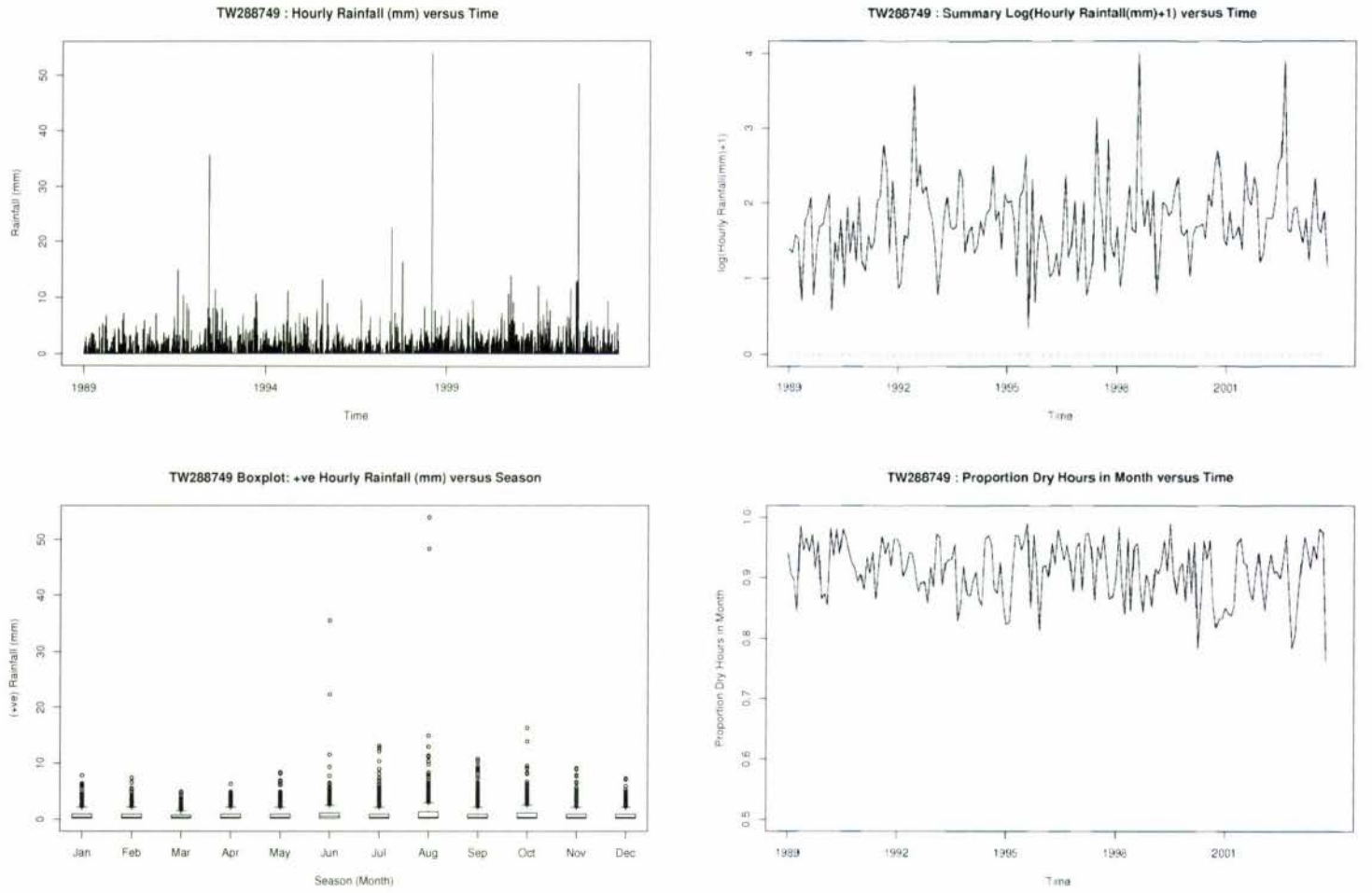


Figure A.36: Site TW288749 hourly plots

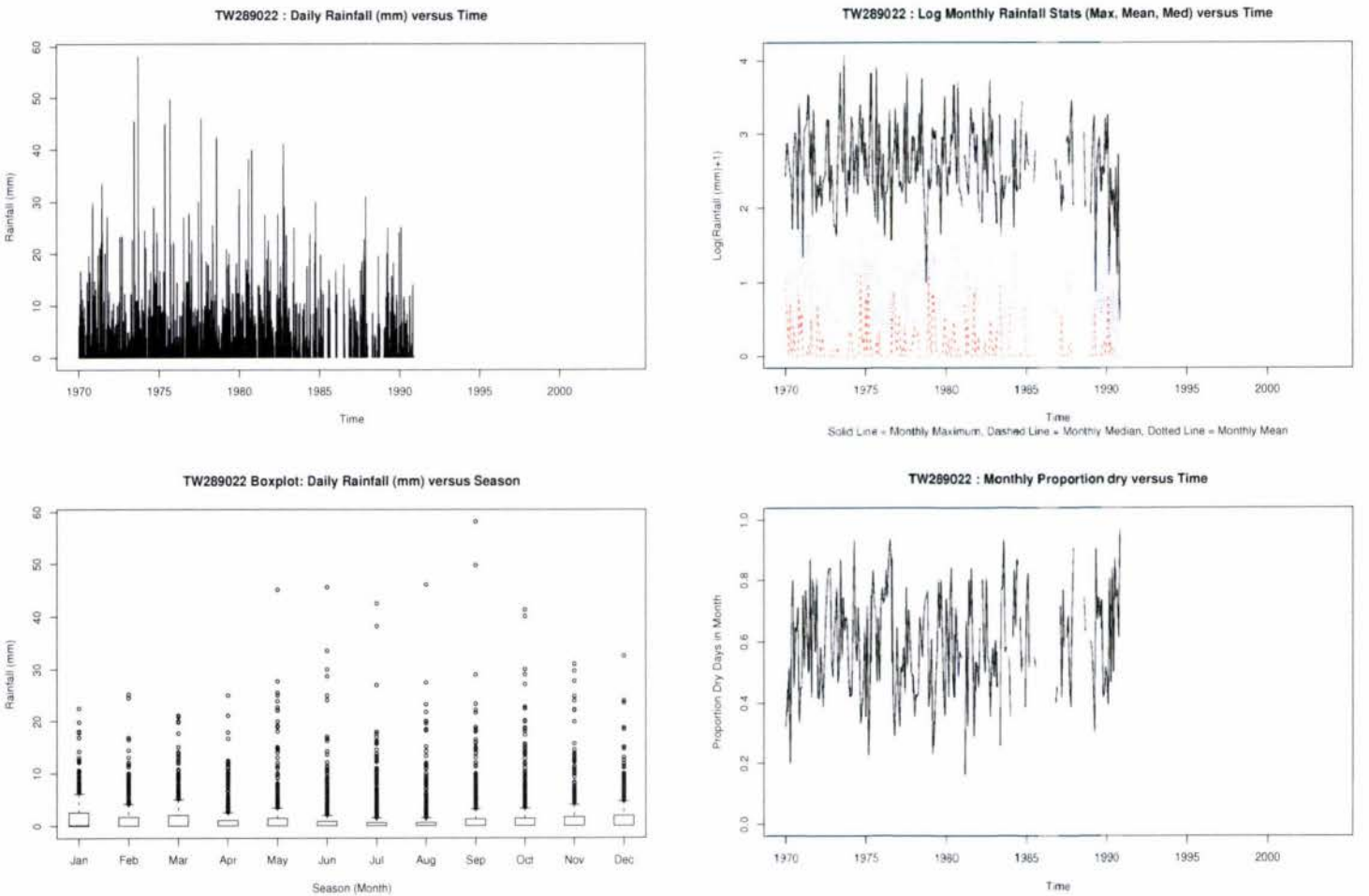


Figure A.37: Site TW289022 daily plots

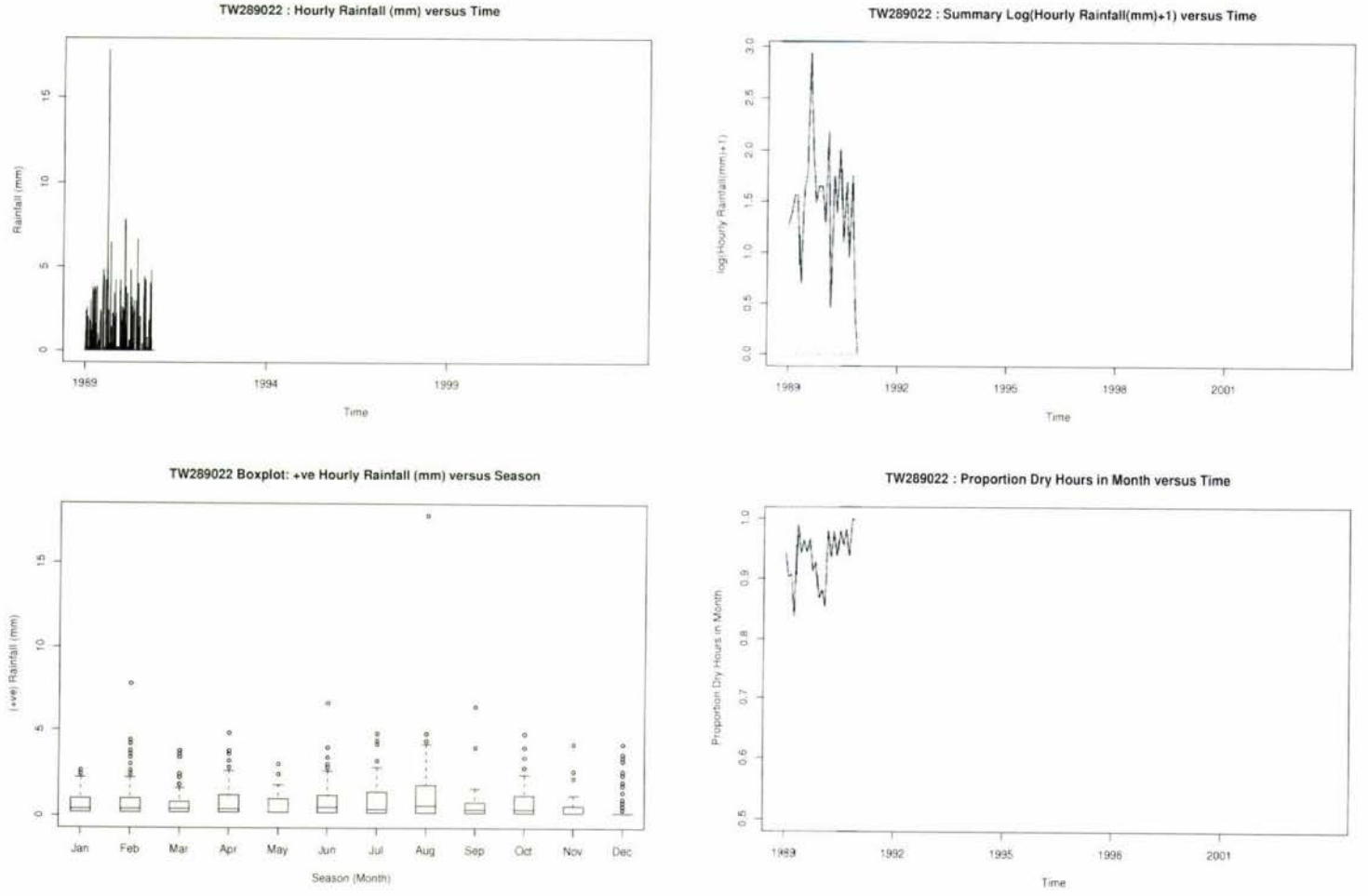


Figure A.38: Site TW289022 hourly plots

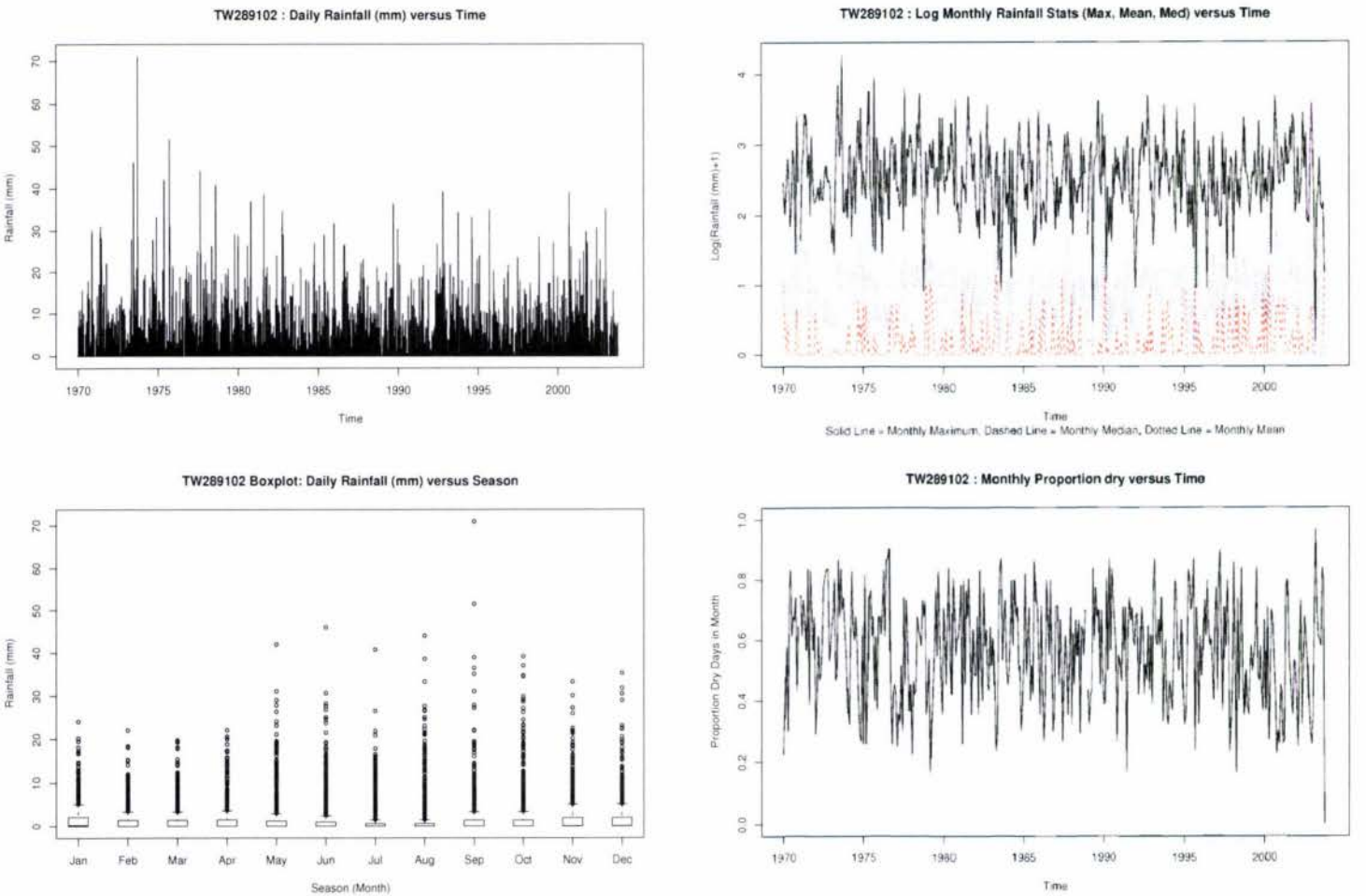


Figure A.39: Site TW289102 daily plots

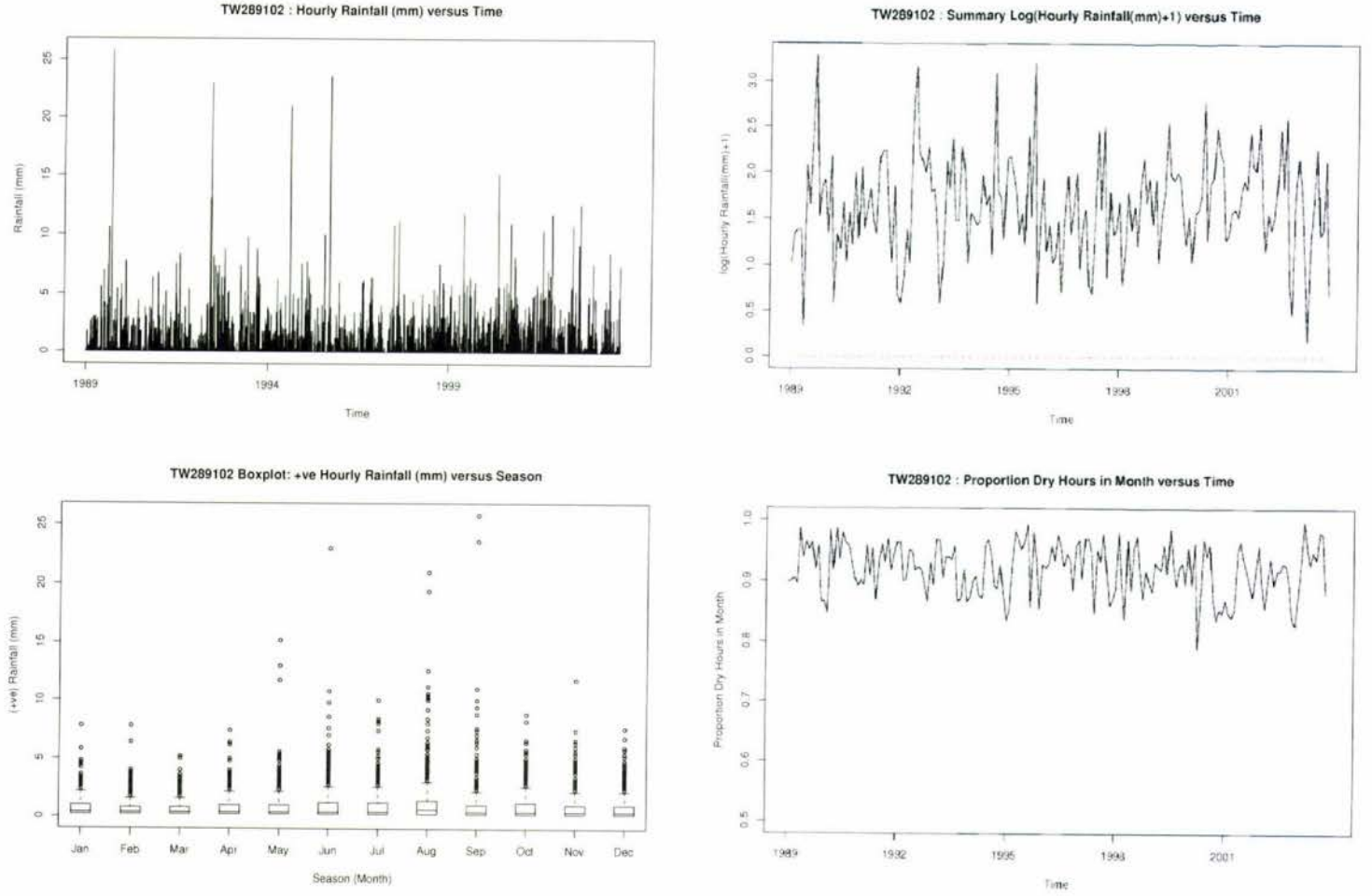


Figure A.40: Site TW289102 hourly plots

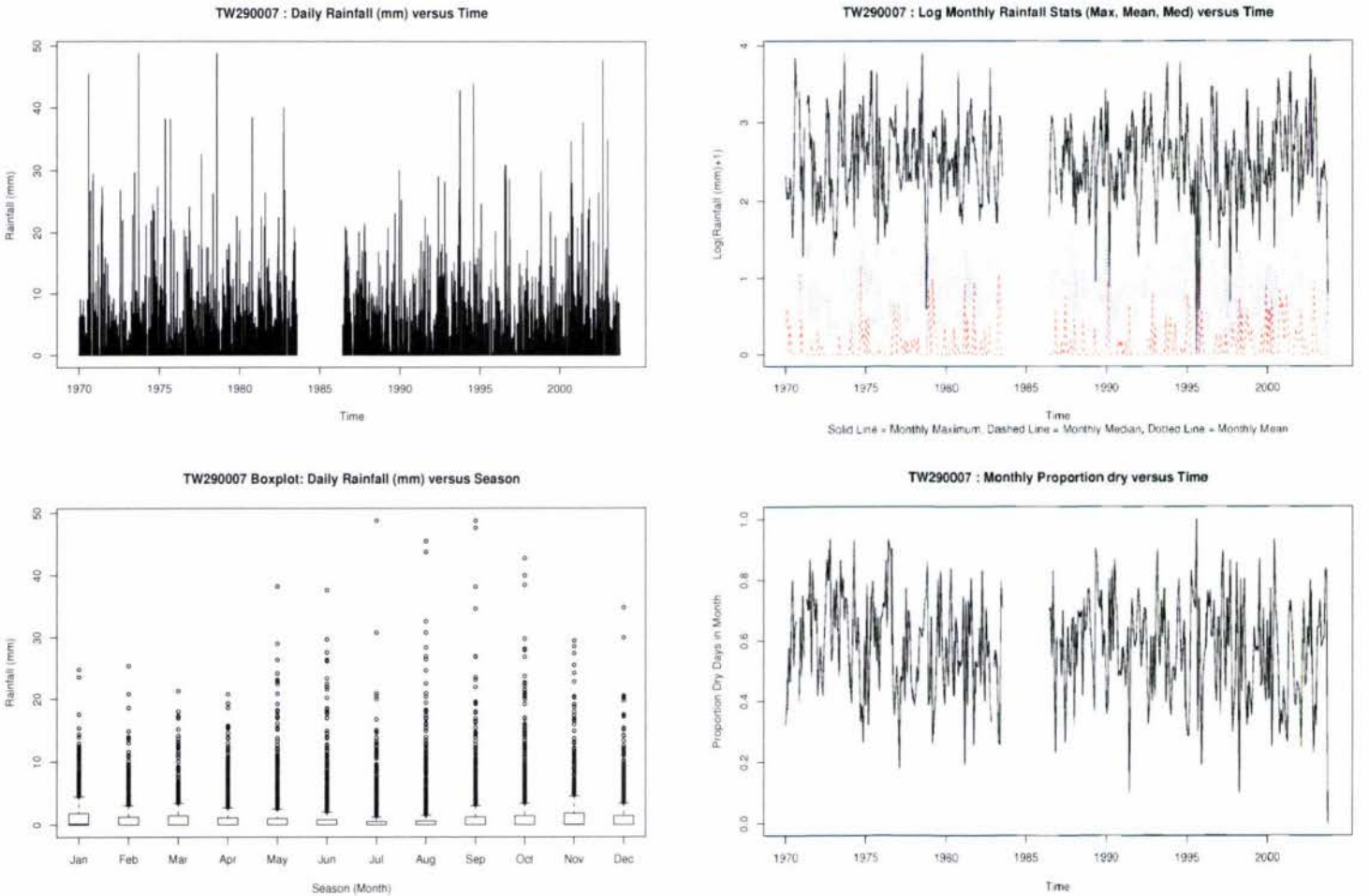


Figure A.41: Site TW290007 daily plots

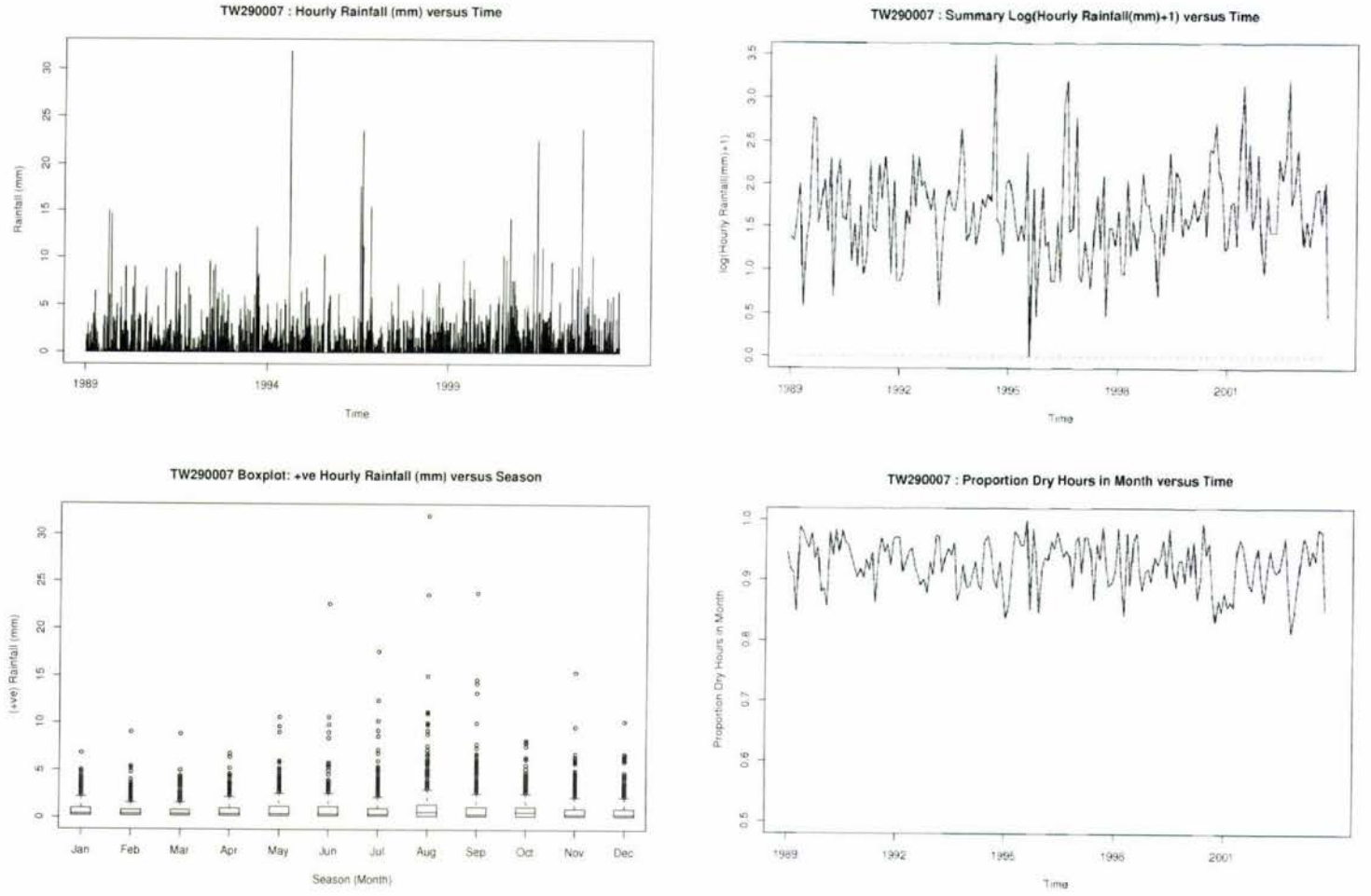


Figure A.42: Site TW290007 hourly plots

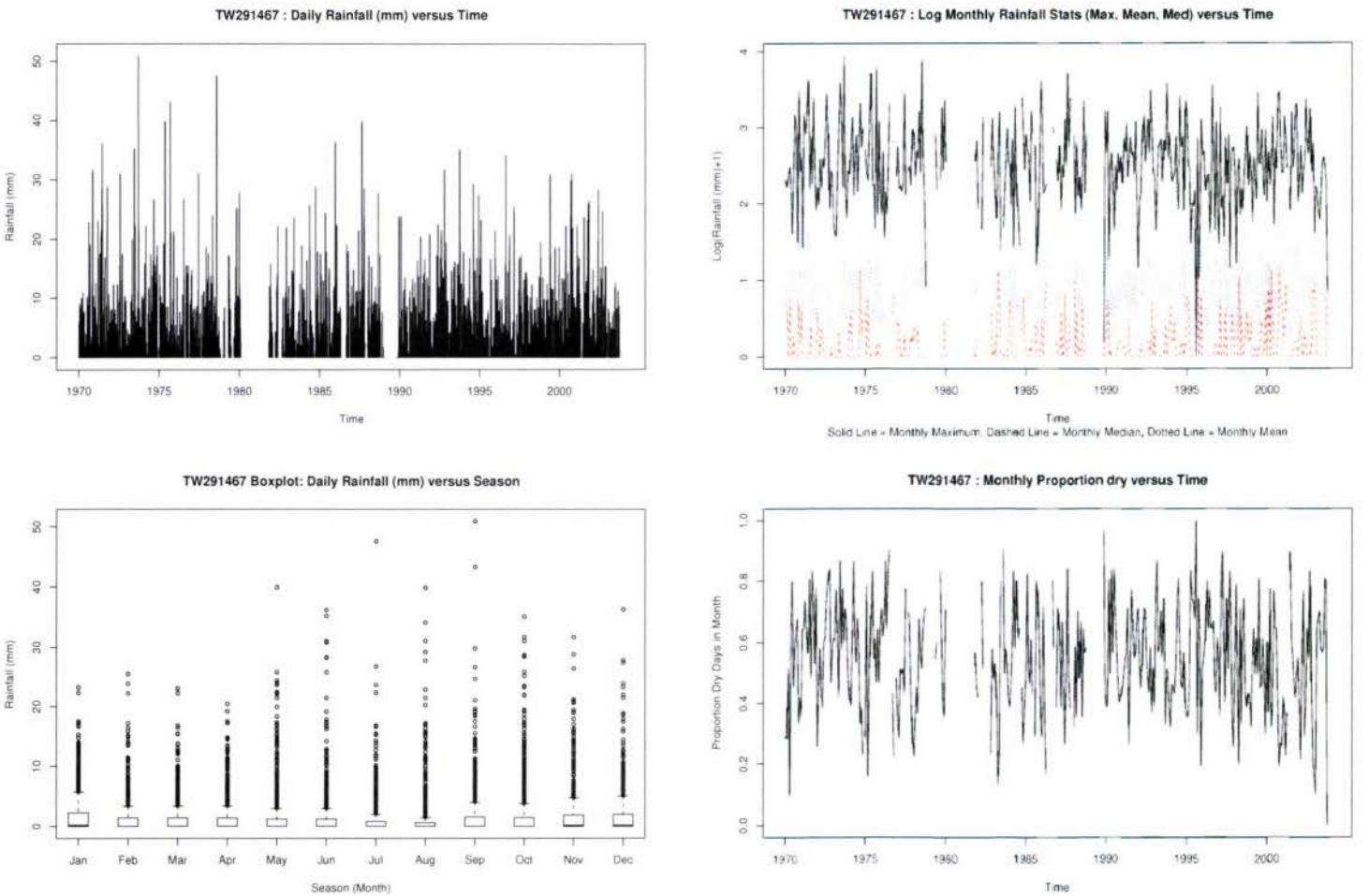


Figure A.43: Site TW291467 daily plots

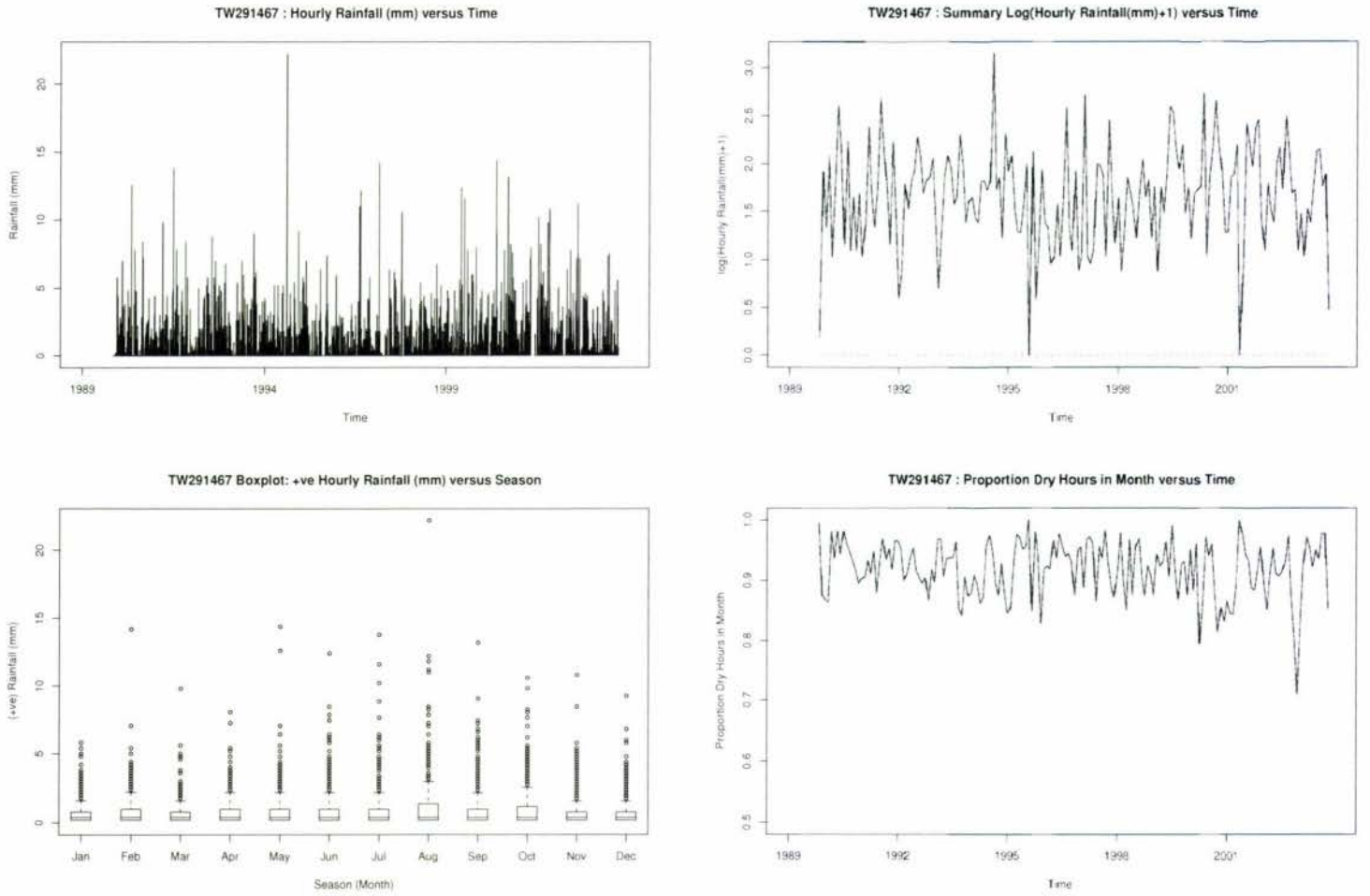
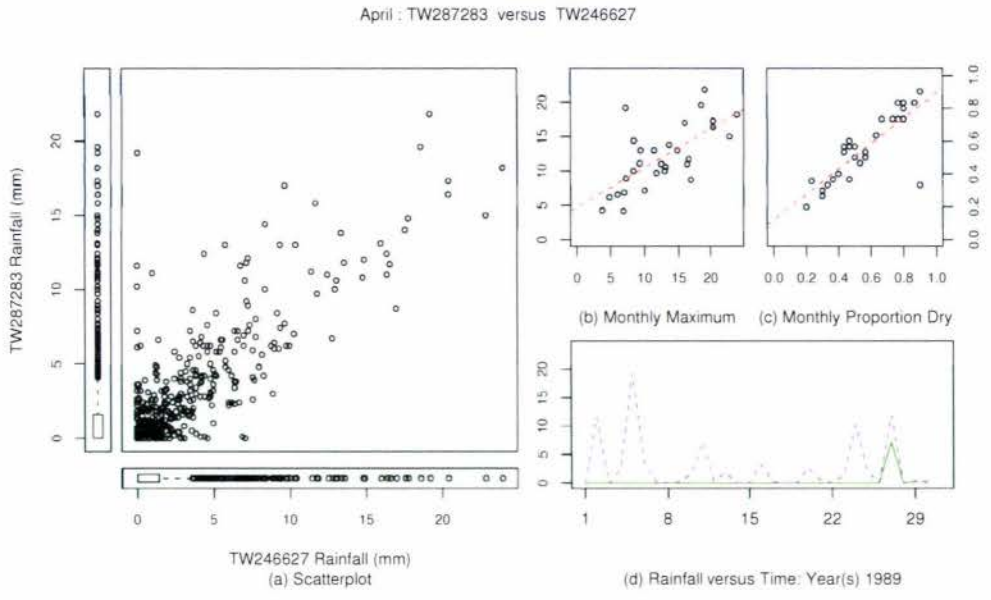


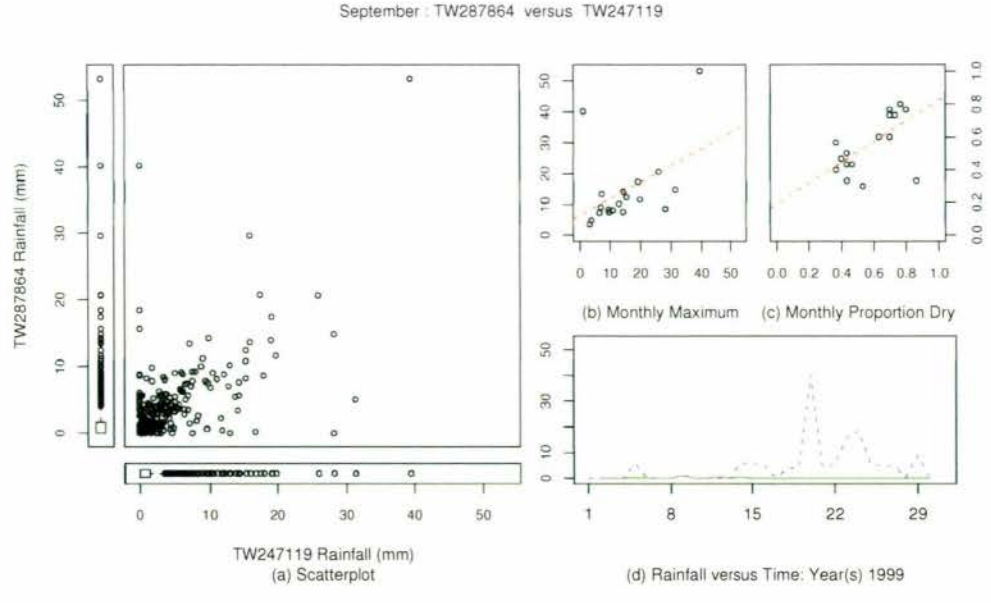
Figure A.44: Site TW291467 hourly plots

A.2 Spatial plots



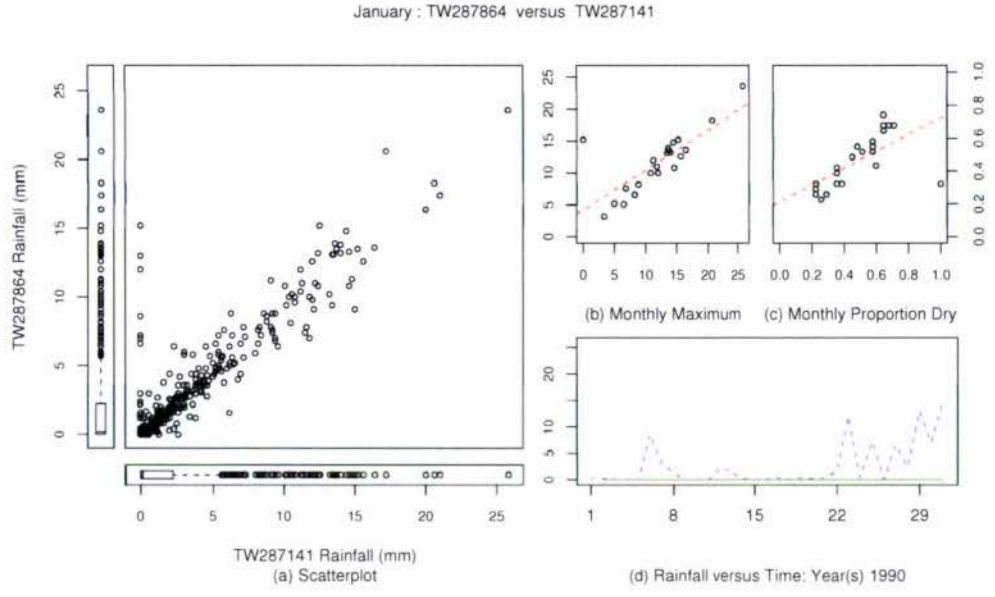
Note: plot (d), Solid line is site TW246627, dashed line is site TW287283

Figure A.45: Daily data, April, site TW287283 versus site TW246627



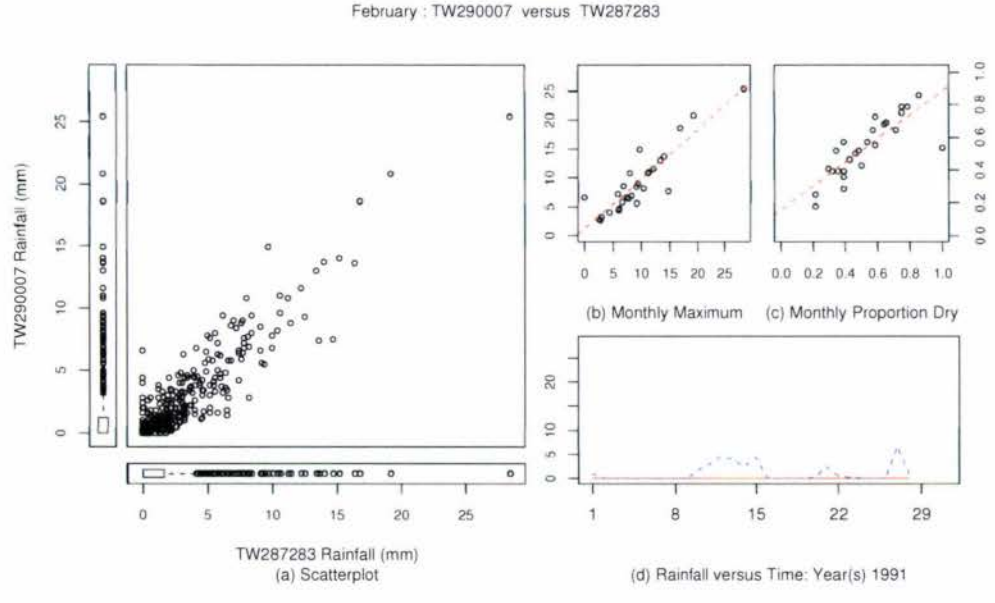
Note: plot (d), Solid line is site TW247119, dashed line is site TW287874

Figure A.46: Daily data, September, site TW287874 versus site TW247119



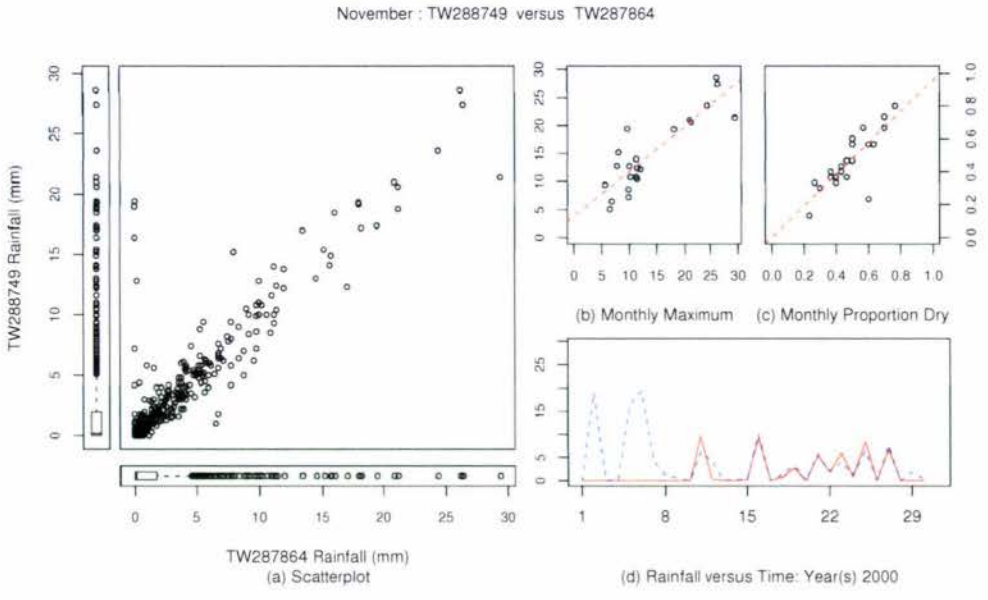
Note: plot (d), Solid line is site TW287141, dashed line is site TW287874

Figure A.47: Daily data, January, site TW287874 versus site TW287141



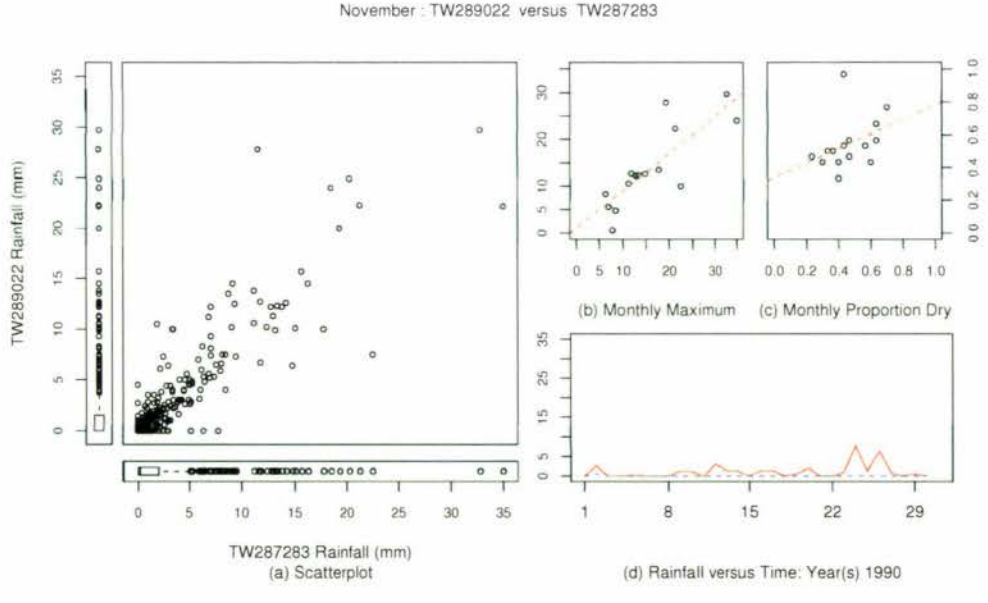
Note: plot (d), Solid line is site TW287283, dashed line is site TW290007

Figure A.48: Daily data, February, site TW290007 versus site TW287283



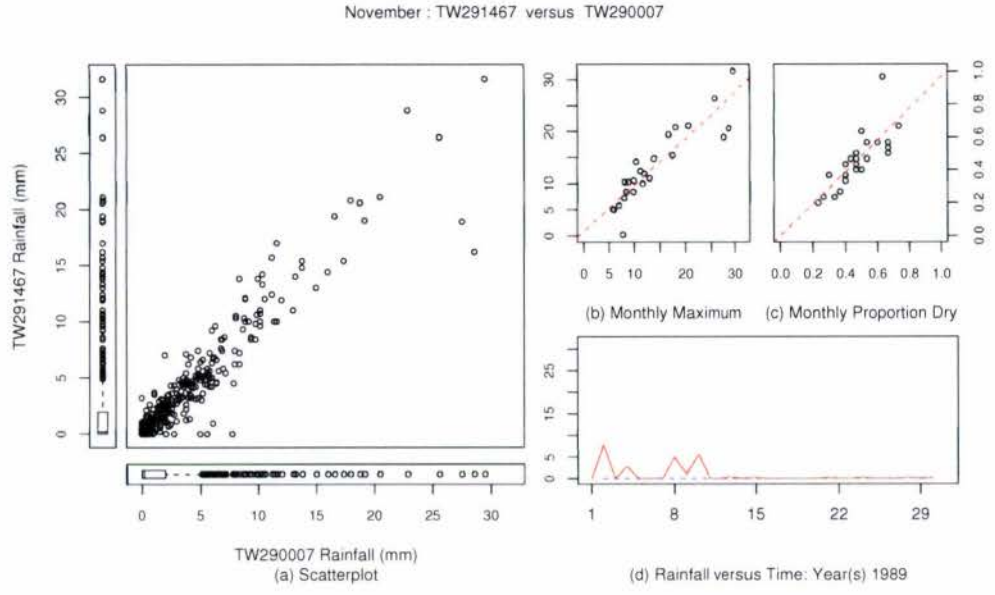
Note: plot (d), Solid line is site TW287864, dashed line is site TW288749

Figure A.49: Daily data, February, site TW288749 versus site TW287864



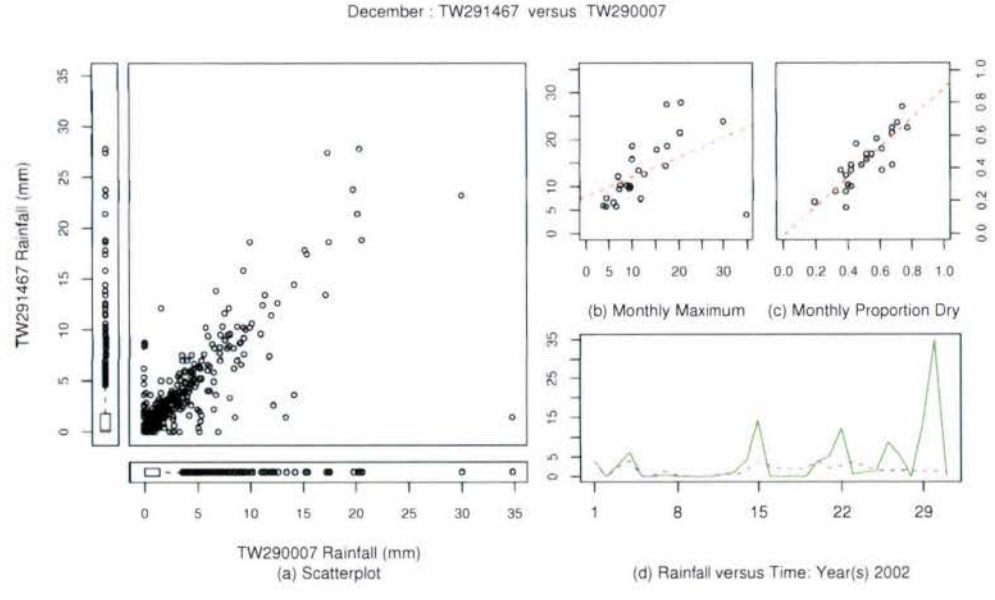
Note: plot (d), Solid line is site TW287283, dashed line is site TW289022

Figure A.50: Daily data, November, site TW289022 versus site TW287283



Note: plot (d), Solid line is site TW290007, dashed line is site TW291467

Figure A.51: Daily data, November, site TW291467 versus site TW290007



Note: plot (d), Solid line is site TW290007, dashed line is site TW291467

Figure A.52: Daily data, December, site TW291467 versus site TW290007

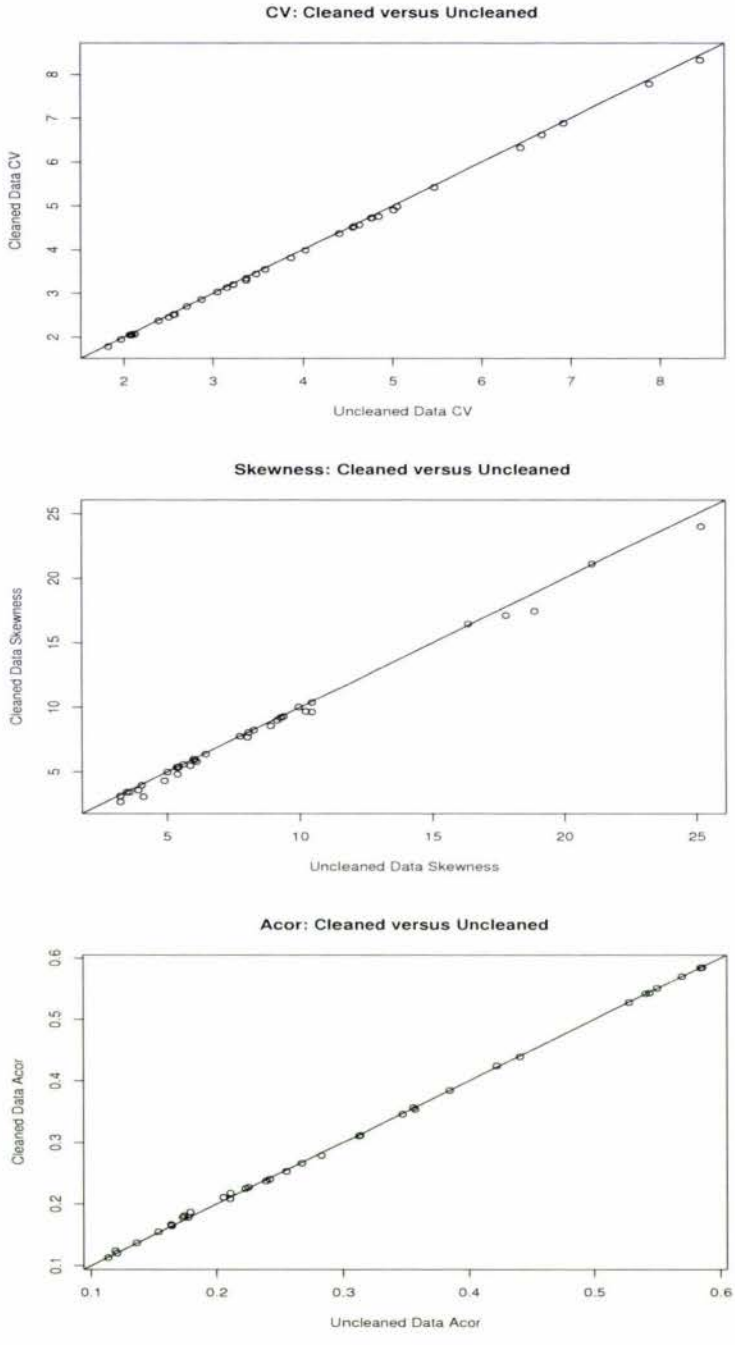


Figure A.53: Pooled statistics: cleaned data versus uncleaned data (CV, skewness, autocorrelation)

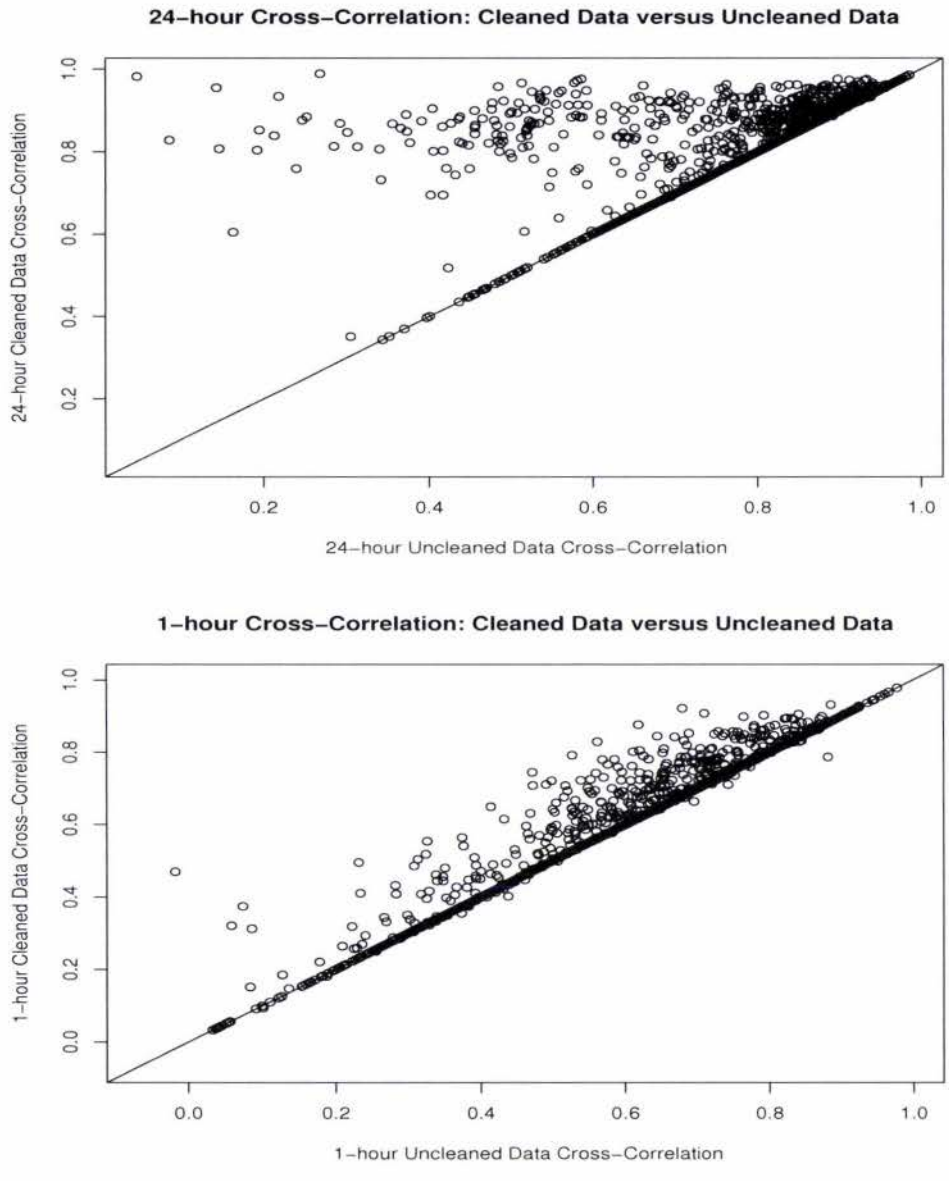
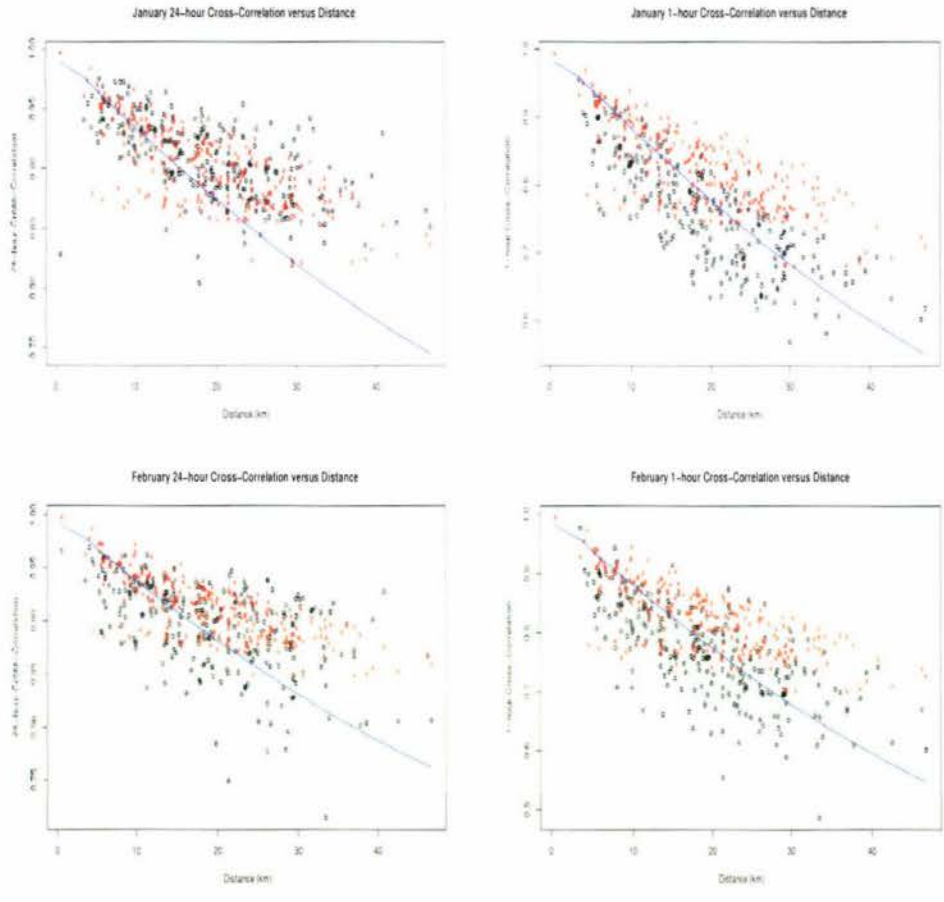


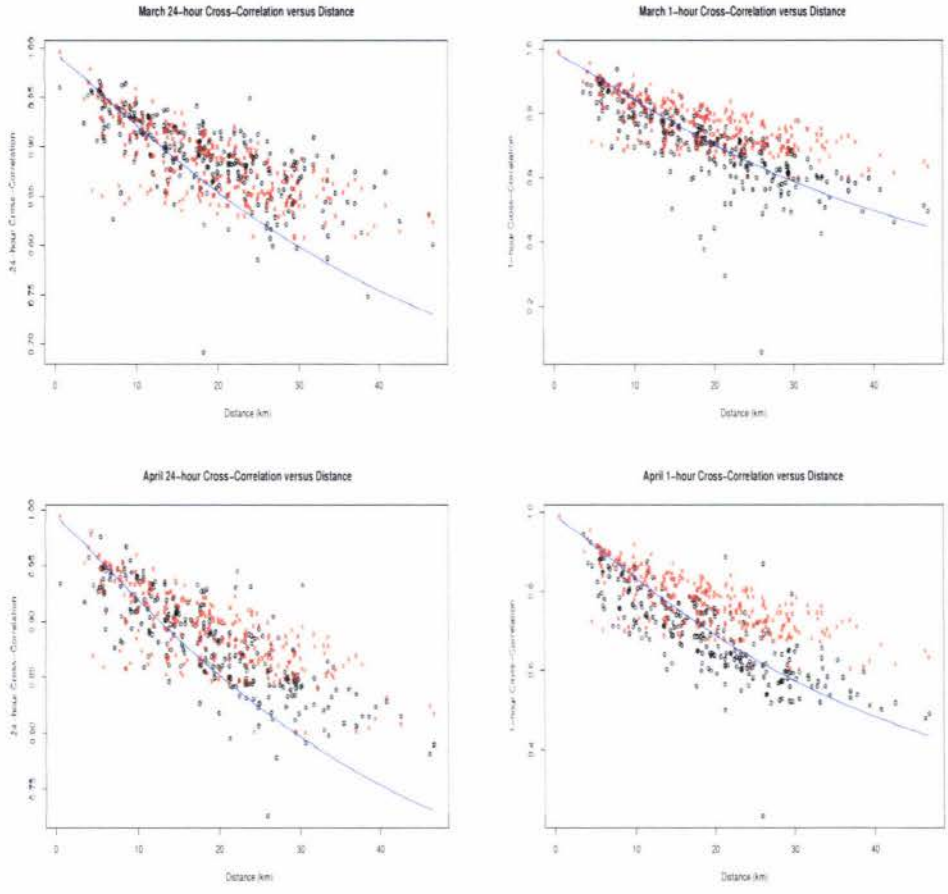
Figure A.54: Pooled statistics: cleaned data versus uncleaned data (Cross-Correlation)

B. MODEL FITTING: PLOTS



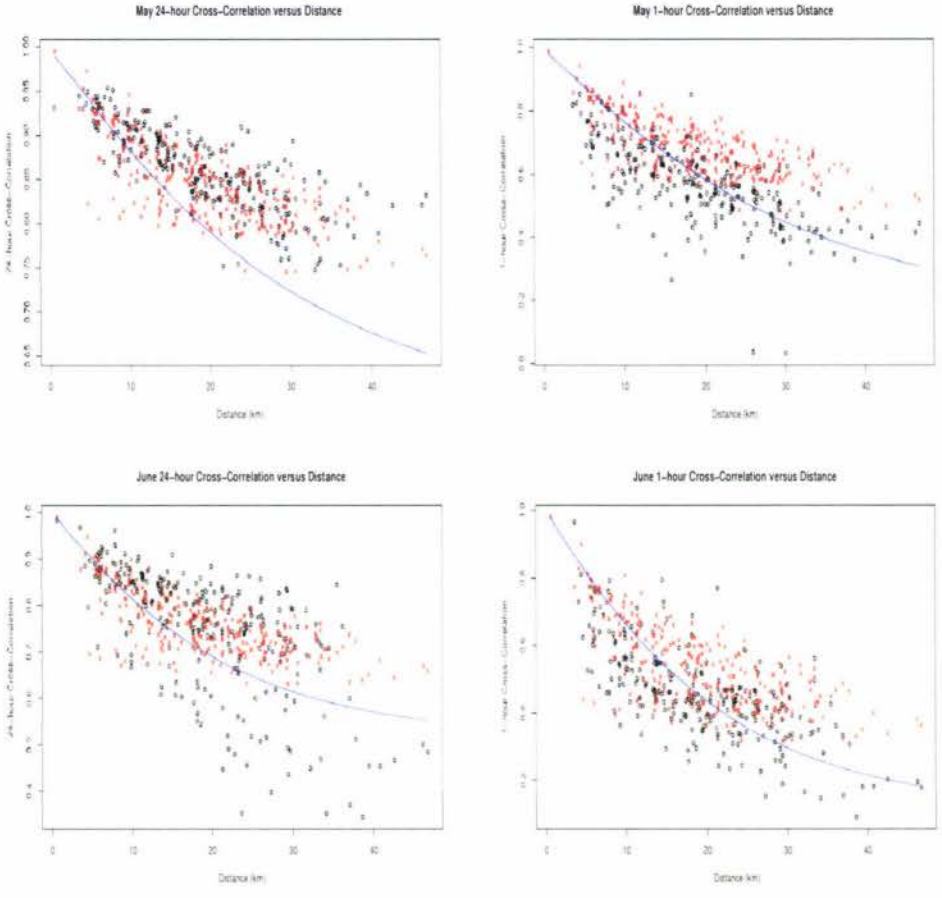
The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure B.1: $Model_B$ cross-correlation versus distance - January, February



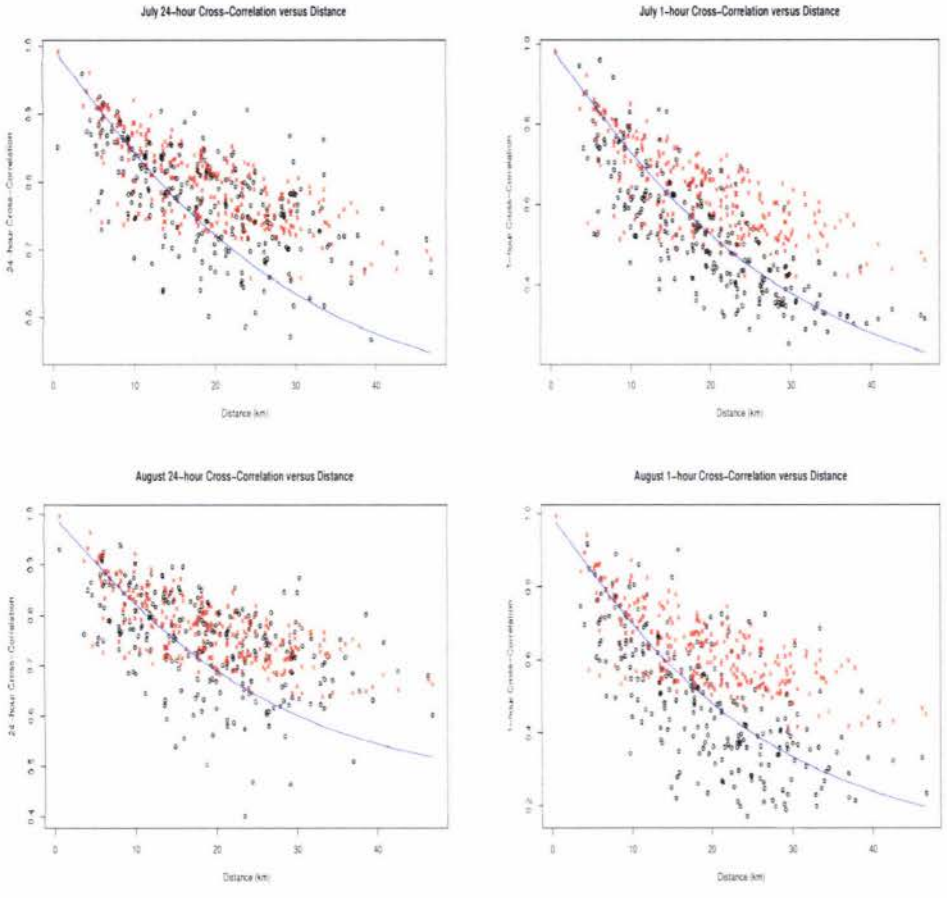
The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure B.2: $Model_B$ cross-correlation versus distance - March, April



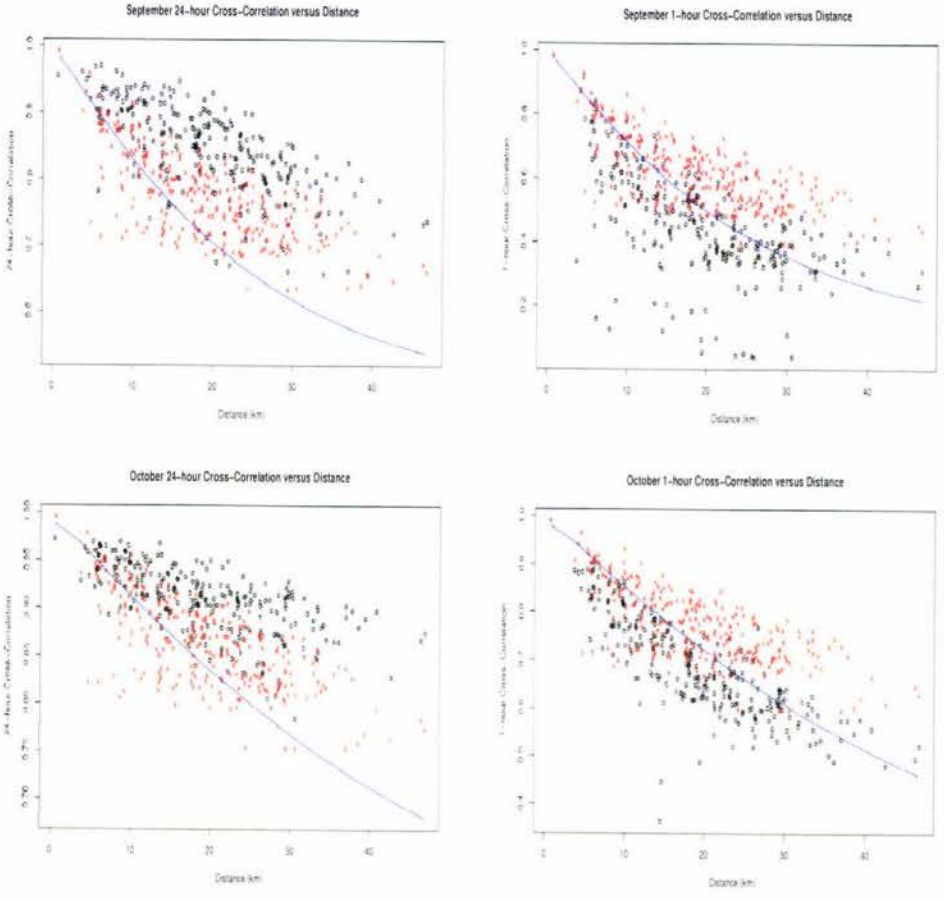
The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure B.3: $Model_B$ cross-correlation versus distance - May, June



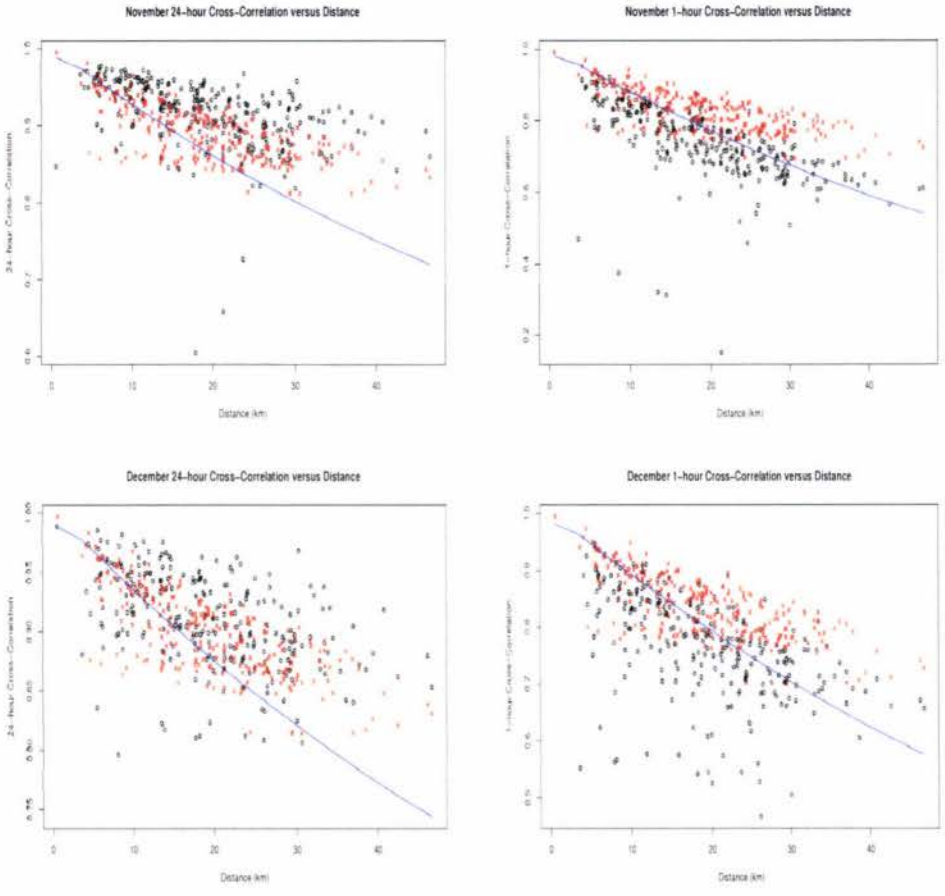
The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure B.4: $Model_B$ cross-correlation versus distance - July, August



The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure B.5: $Model_B$ cross-correlation versus distance - September, October



The curved line is the fitted value under the model,
the x's are the 300-year sample simulation cross-correlation,
the black o's are the historical cross-correlation.

Figure B.6: $Model_B$ cross-correlation versus distance - November, December

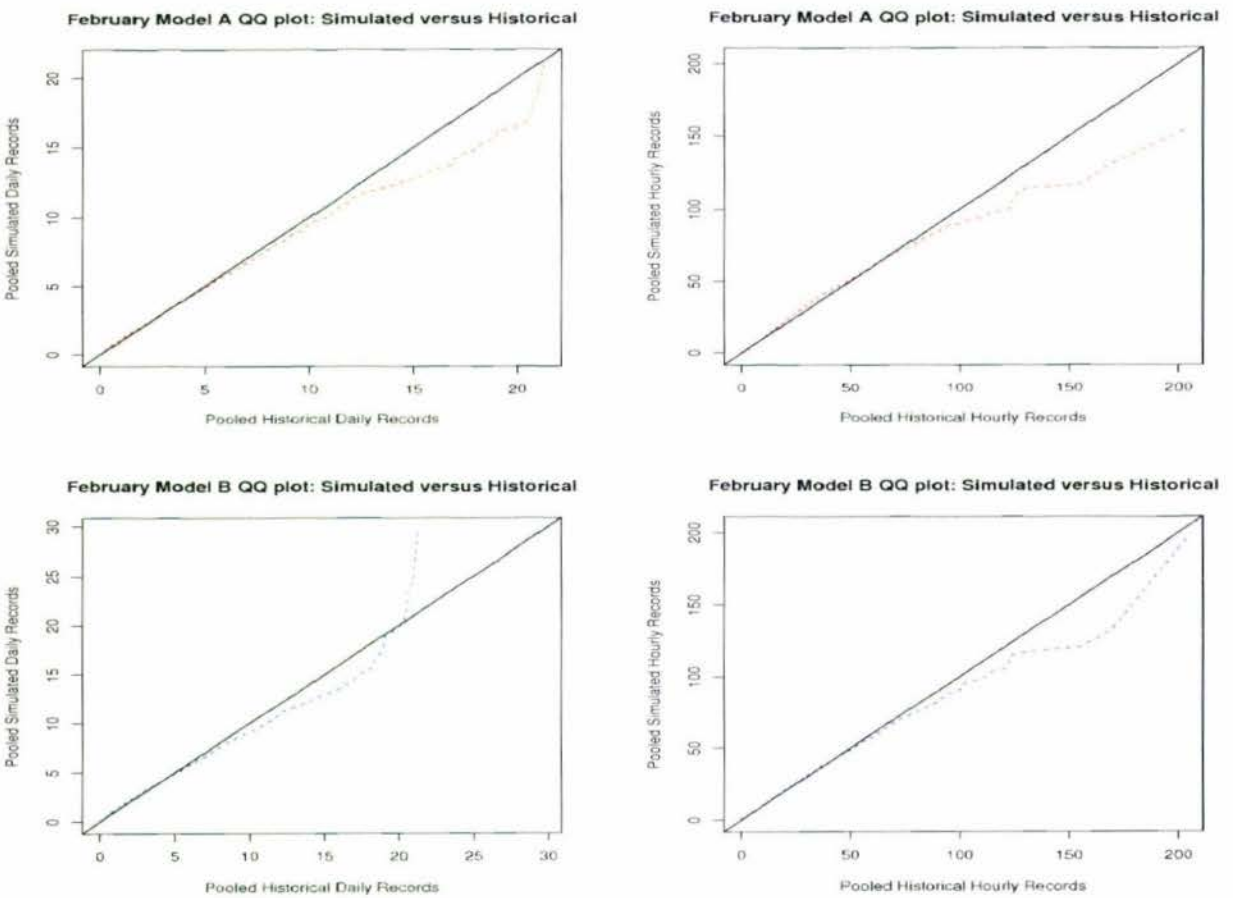


Figure B.7: Quantile-Quantile plots: February $Model_A$, $Model_B$

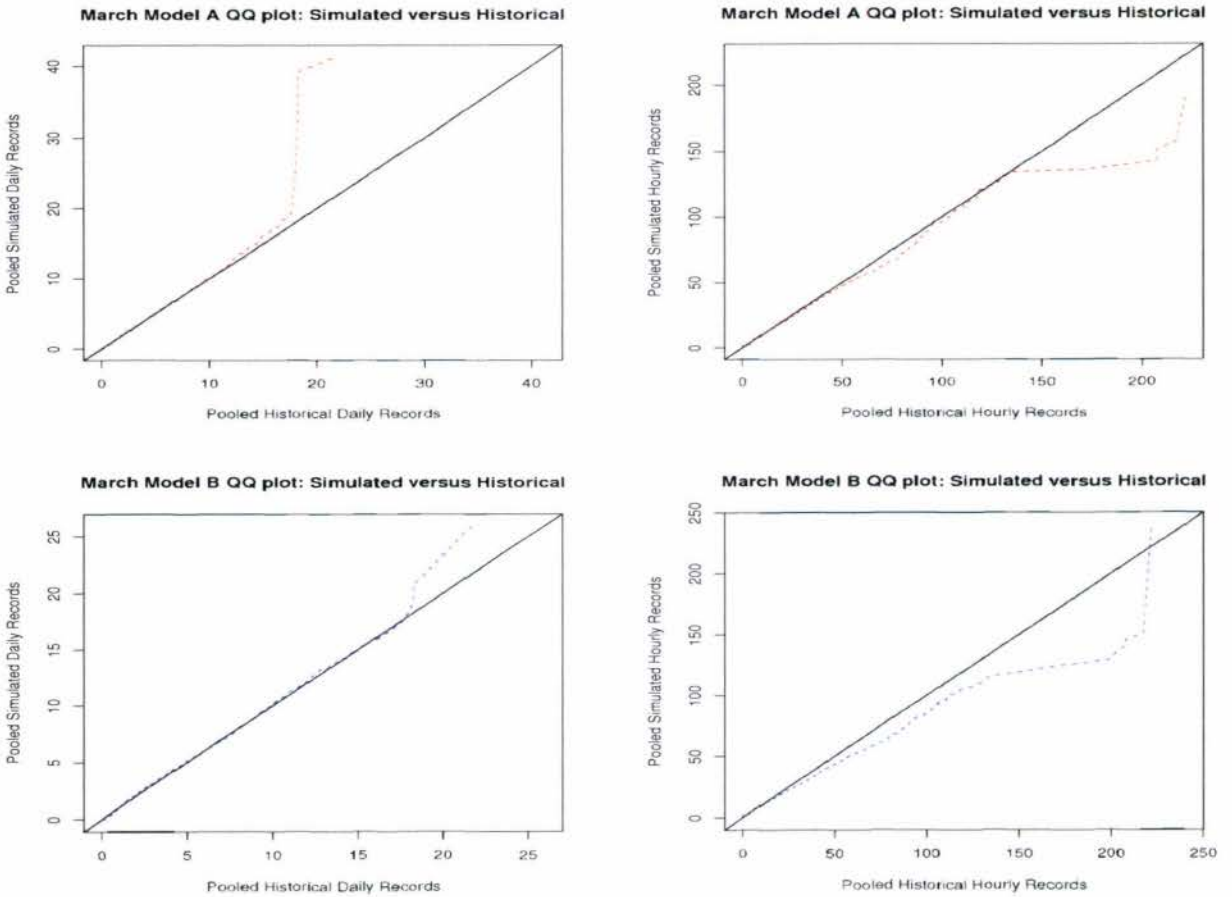


Figure B.8: Quantile-Quantile plots: March $Model_A, Model_B$

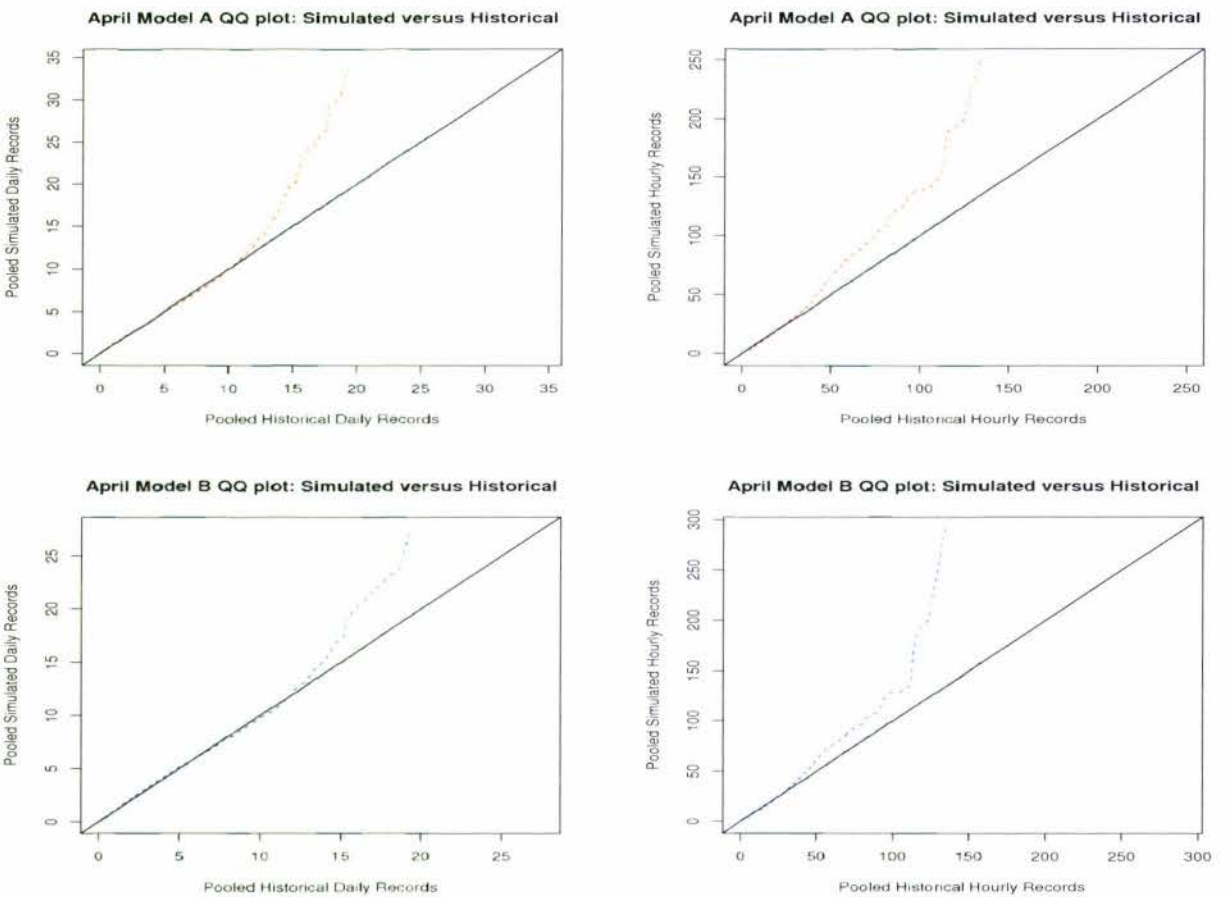


Figure B.9: Quantile-Quantile plots: April $Model_A, Model_B$

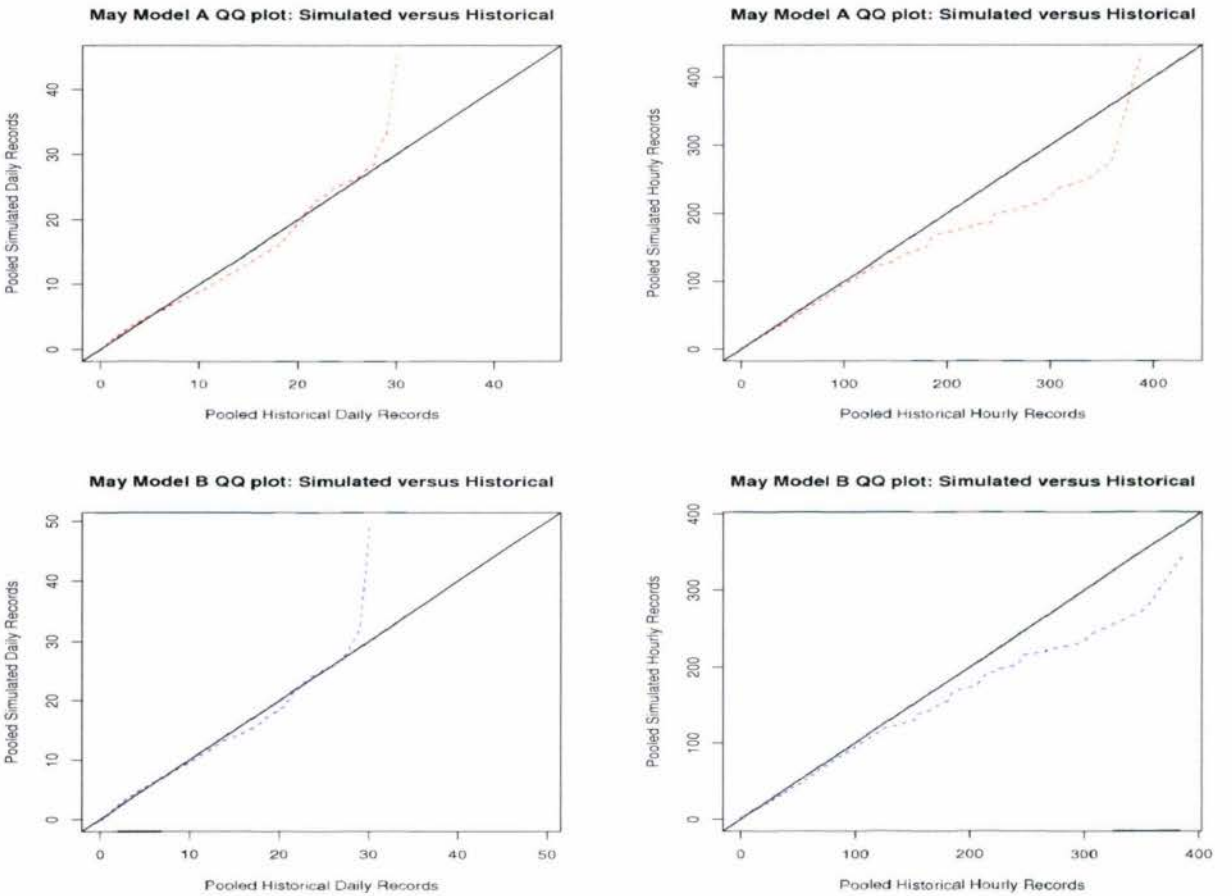


Figure B.10: Quantile-Quantile plots: May $Model_A, Model_B$

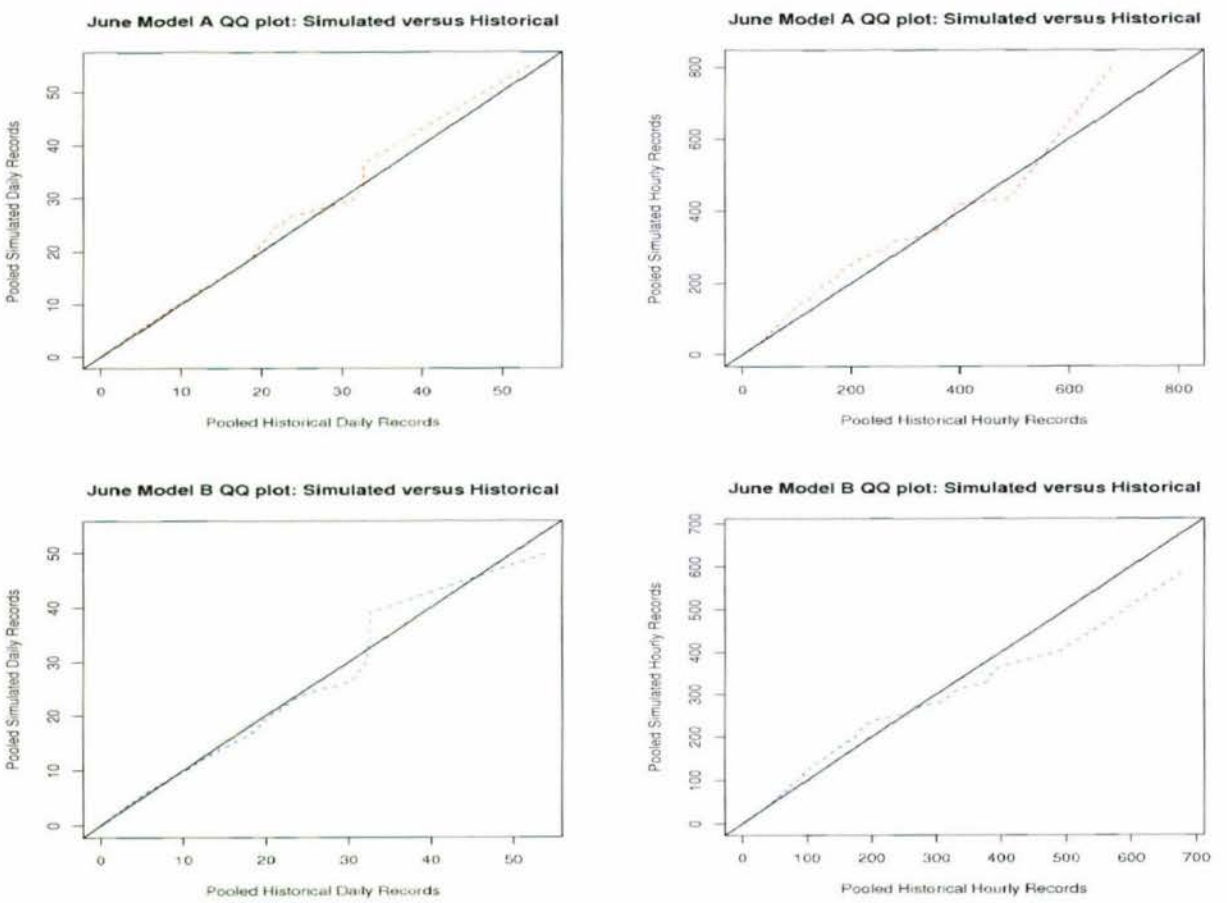


Figure B.11: Quantile-Quantile plots: June $Model_A$, $Model_B$

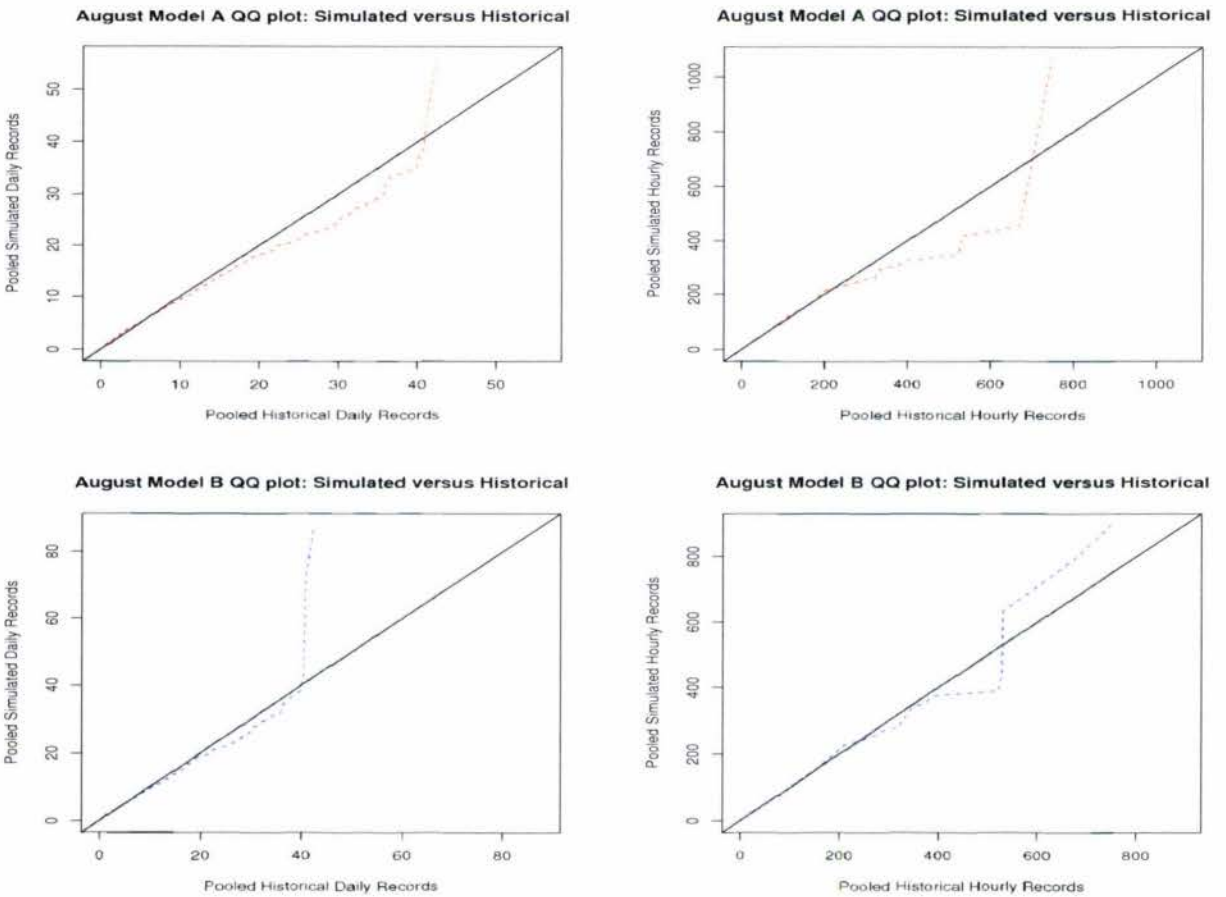


Figure B.12: Quantile-Quantile plots: August $Model_A, Model_B$

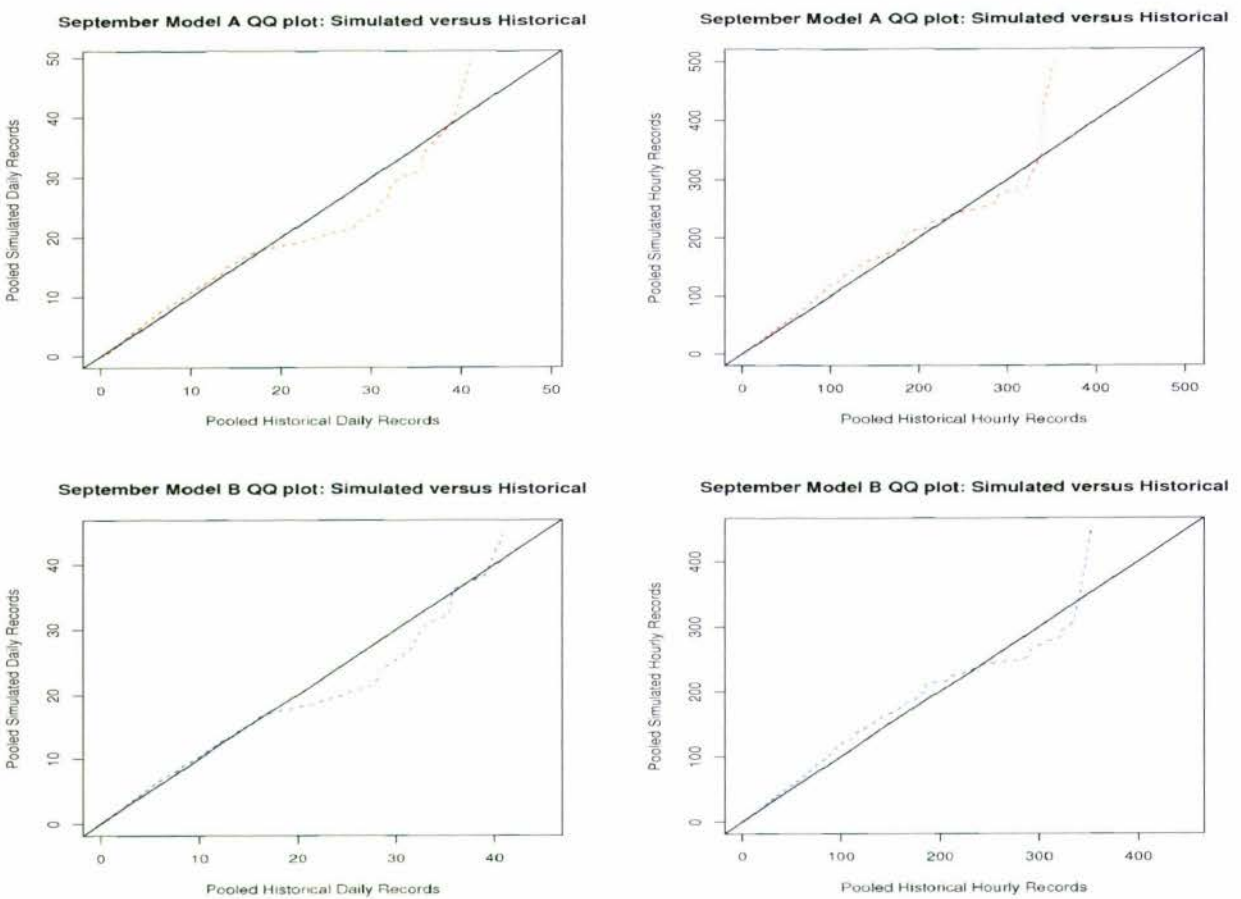


Figure B.13: Quantile-Quantile plots: September $Model_A, Model_B$

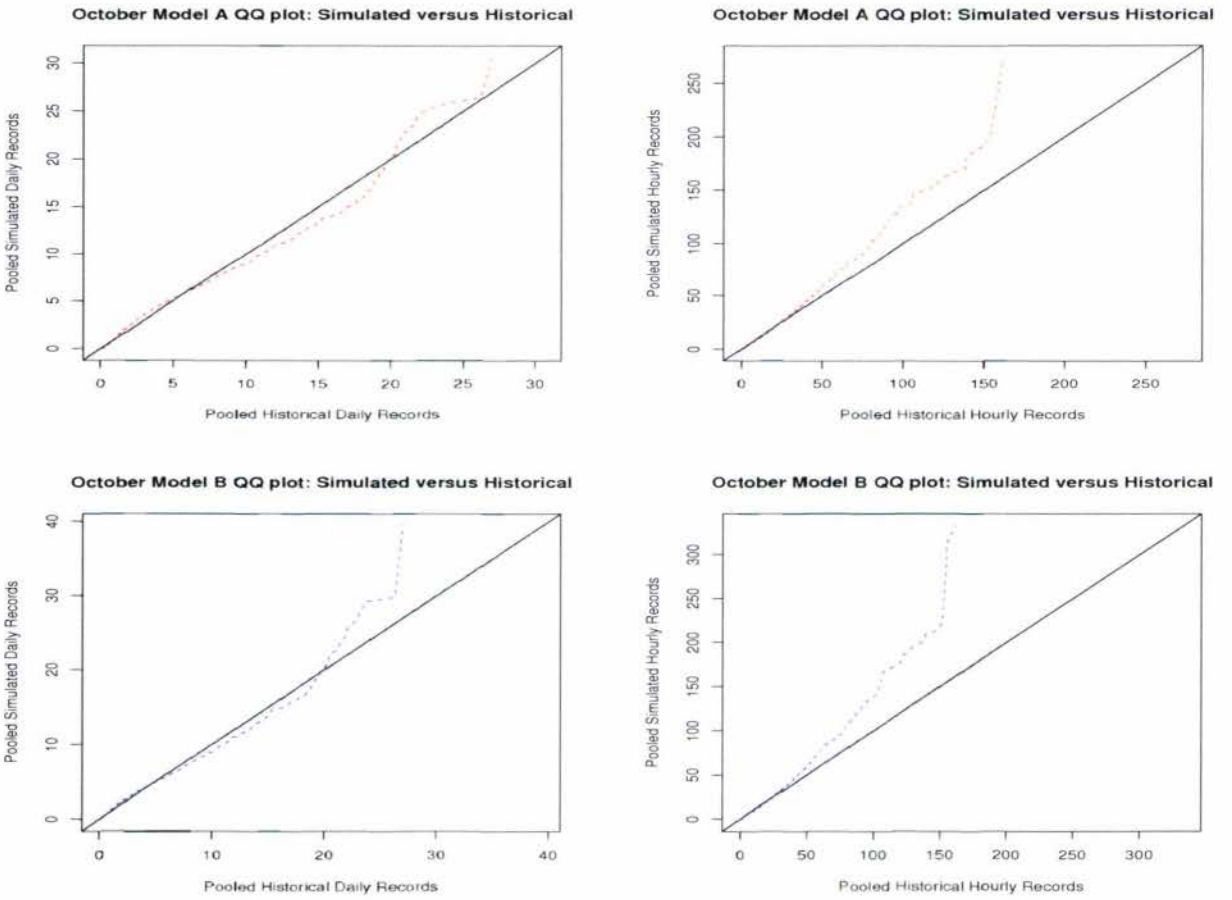


Figure B.14: Quantile-Quantile plots: October $Model_A, Model_B$

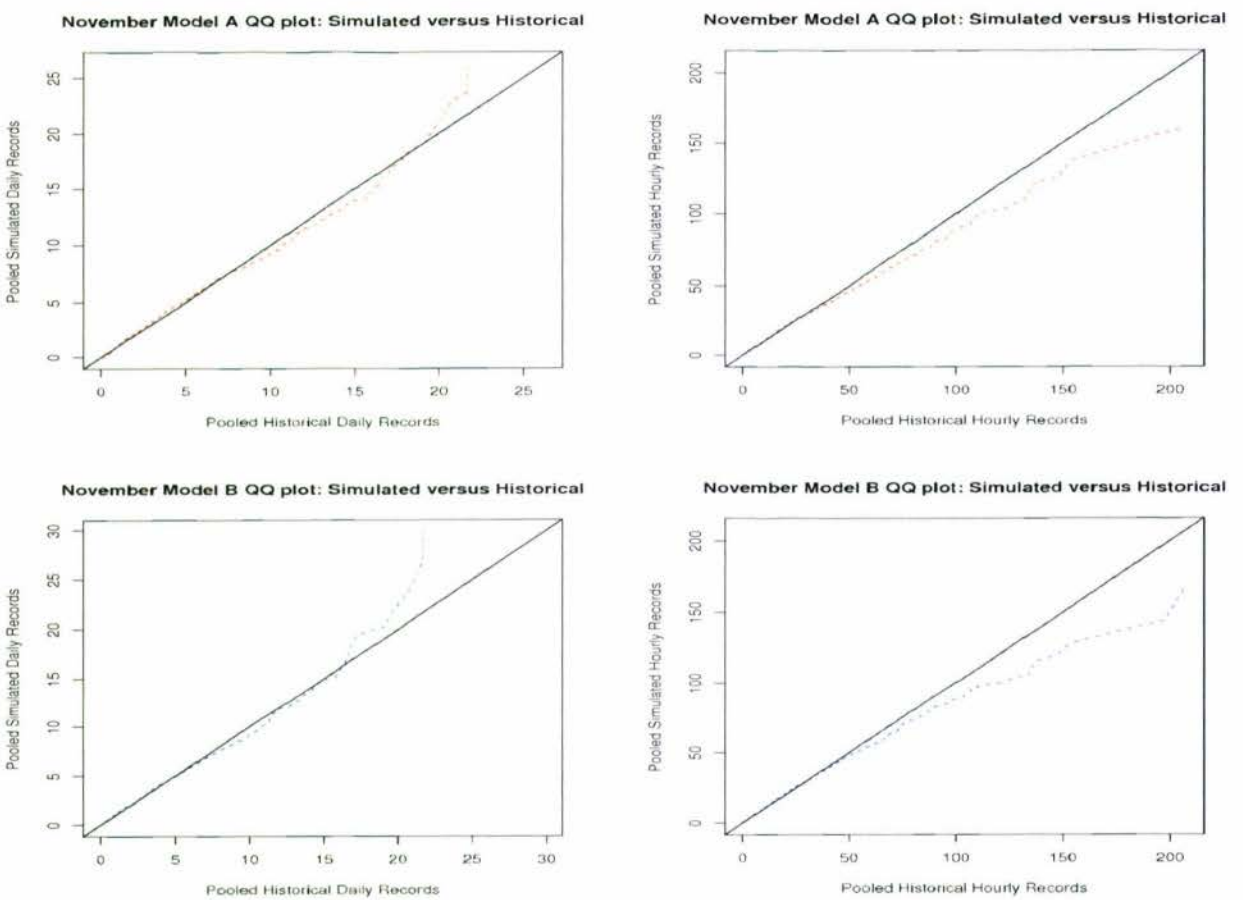


Figure B.15: Quantile-Quantile plots: November $Model_A, Model_B$

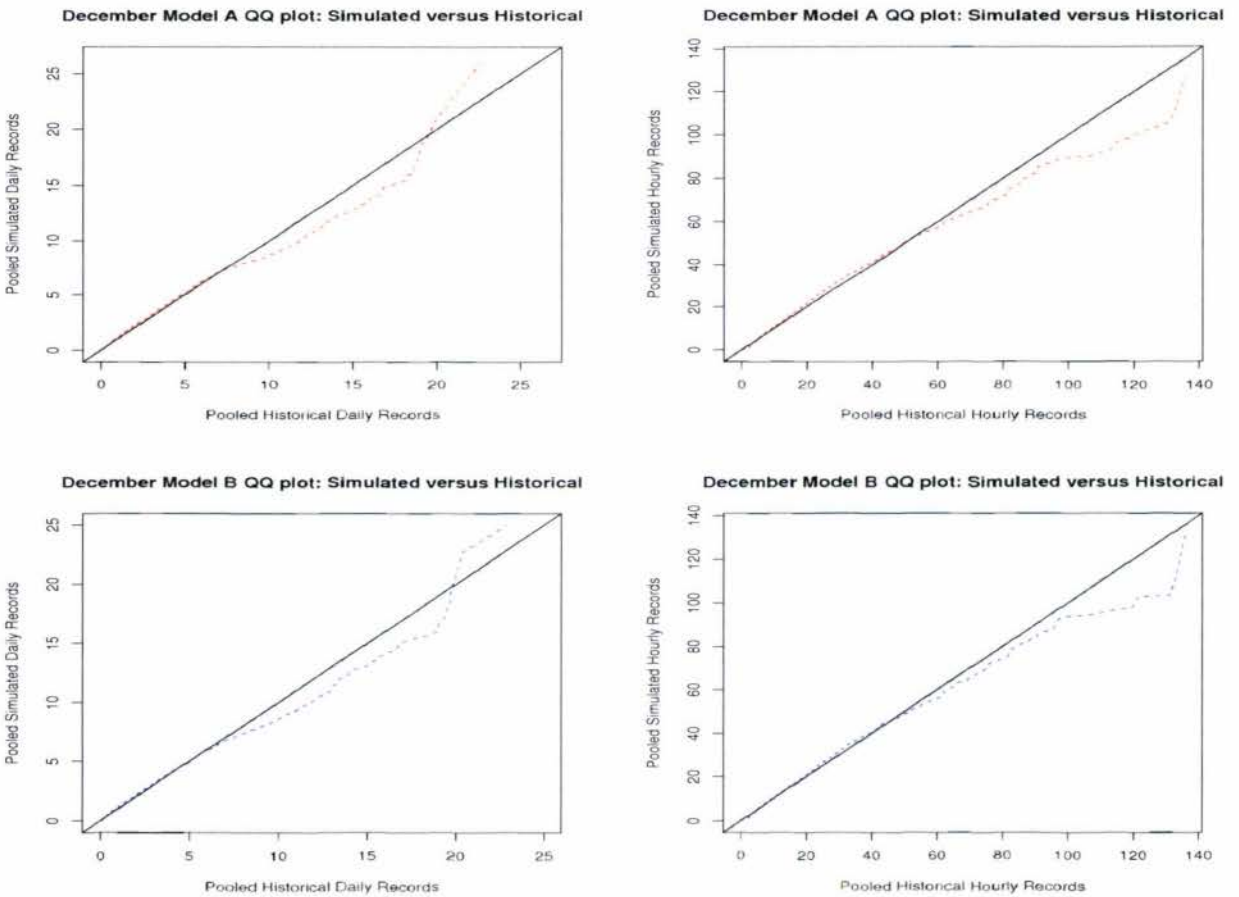
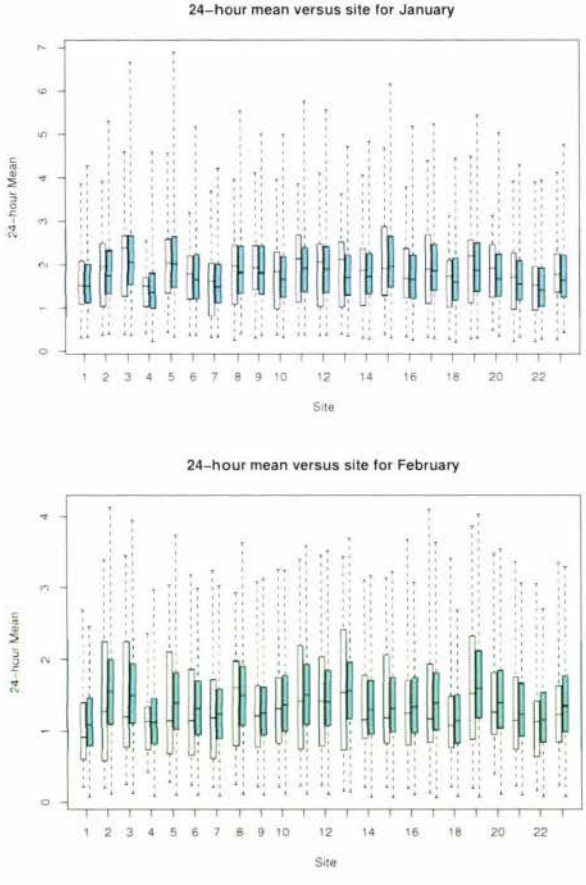


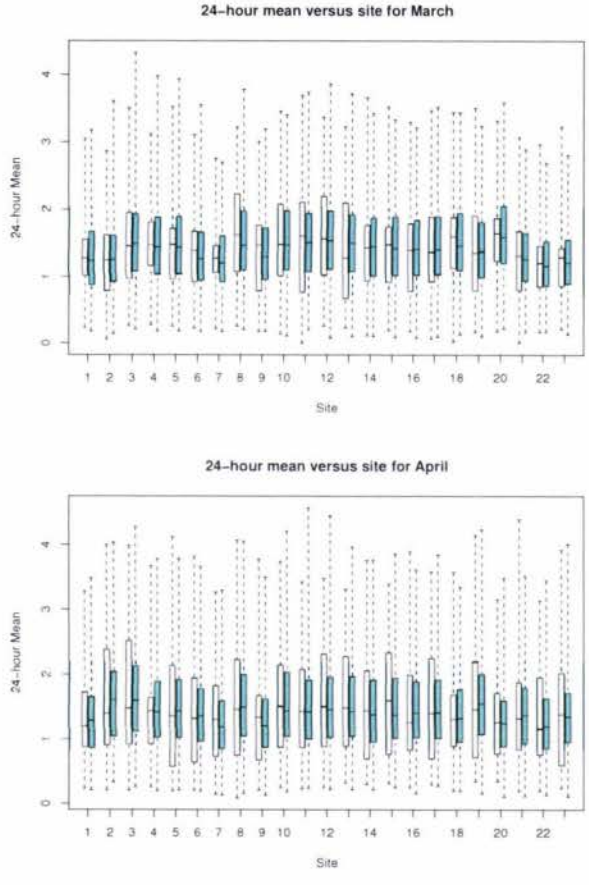
Figure B.16: Quantile-Quantile plots: December $Model_A, Model_B$

C. MODEL VALIDATION: PLOTS



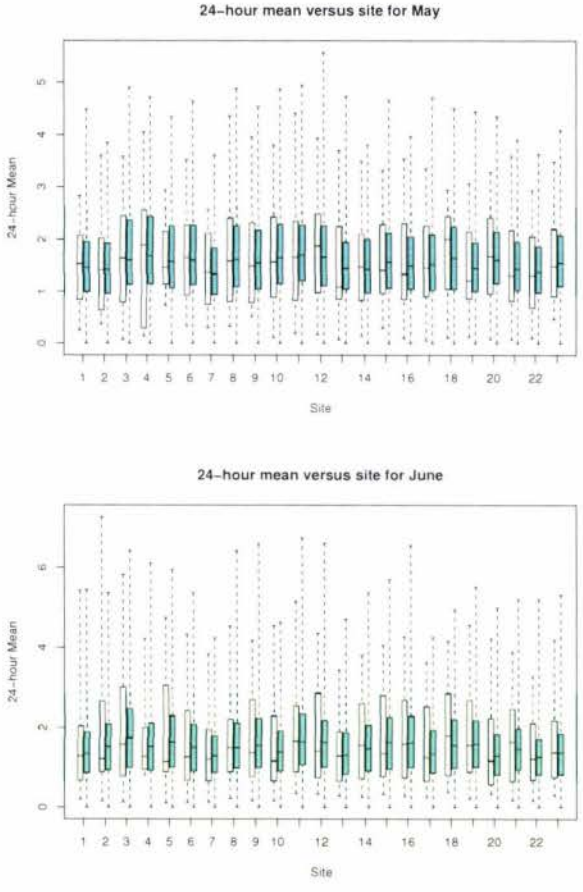
White = Historical; Shaded = $Model_B$

Figure C.1: Monthly Means by site - January and February



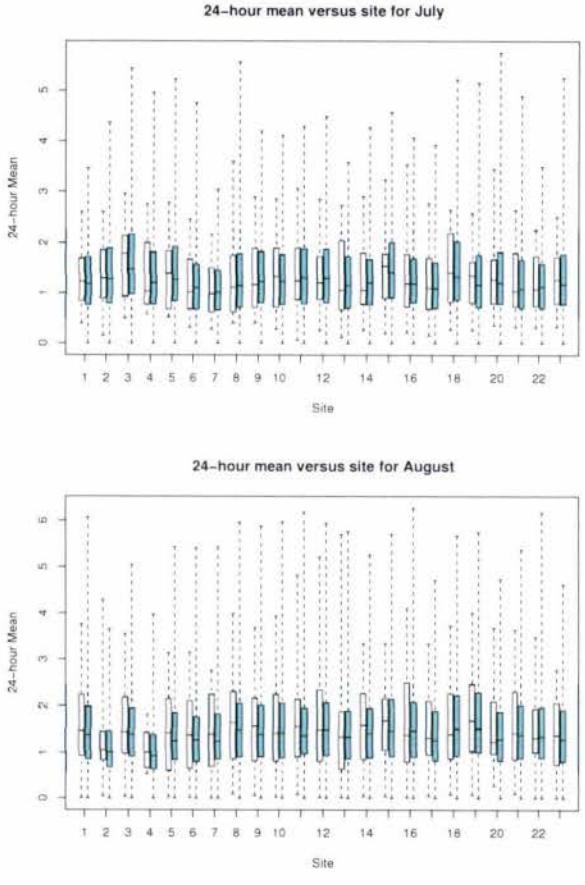
White = Historical; Shaded = $Model_B$

Figure C.2: Monthly Means by site - March and April



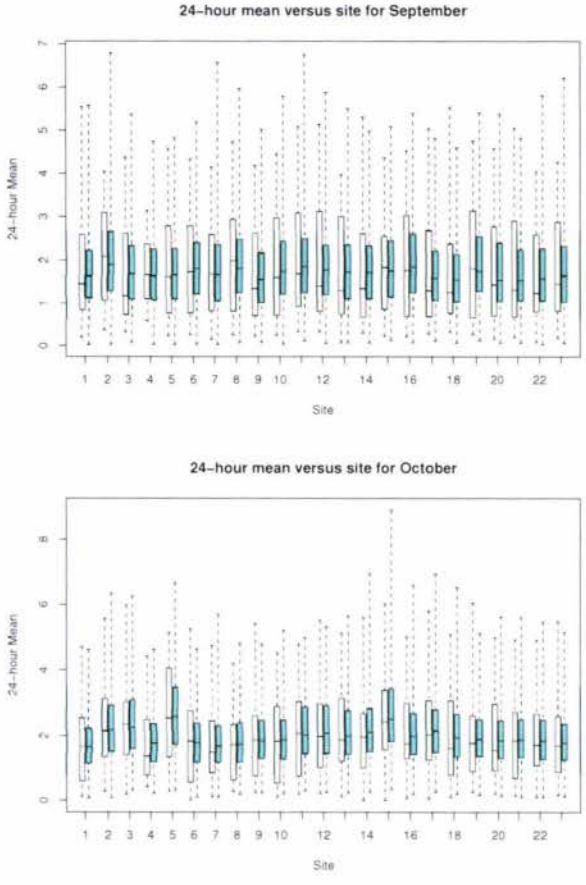
White = Historical; Shaded = $Model_B$

Figure C.3: Monthly Means by site - May and June



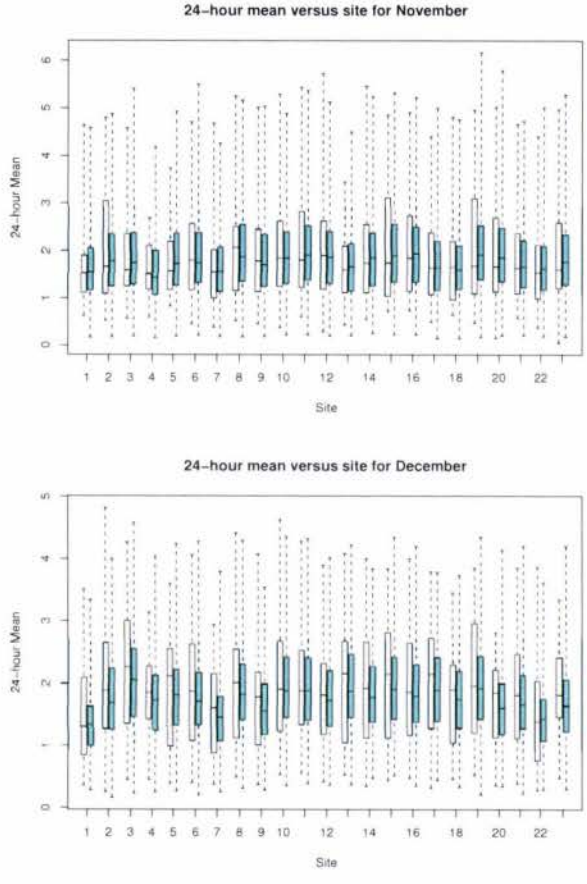
White = Historical; Shaded = $Model_B$

Figure C.4: Monthly Means by site - July and August



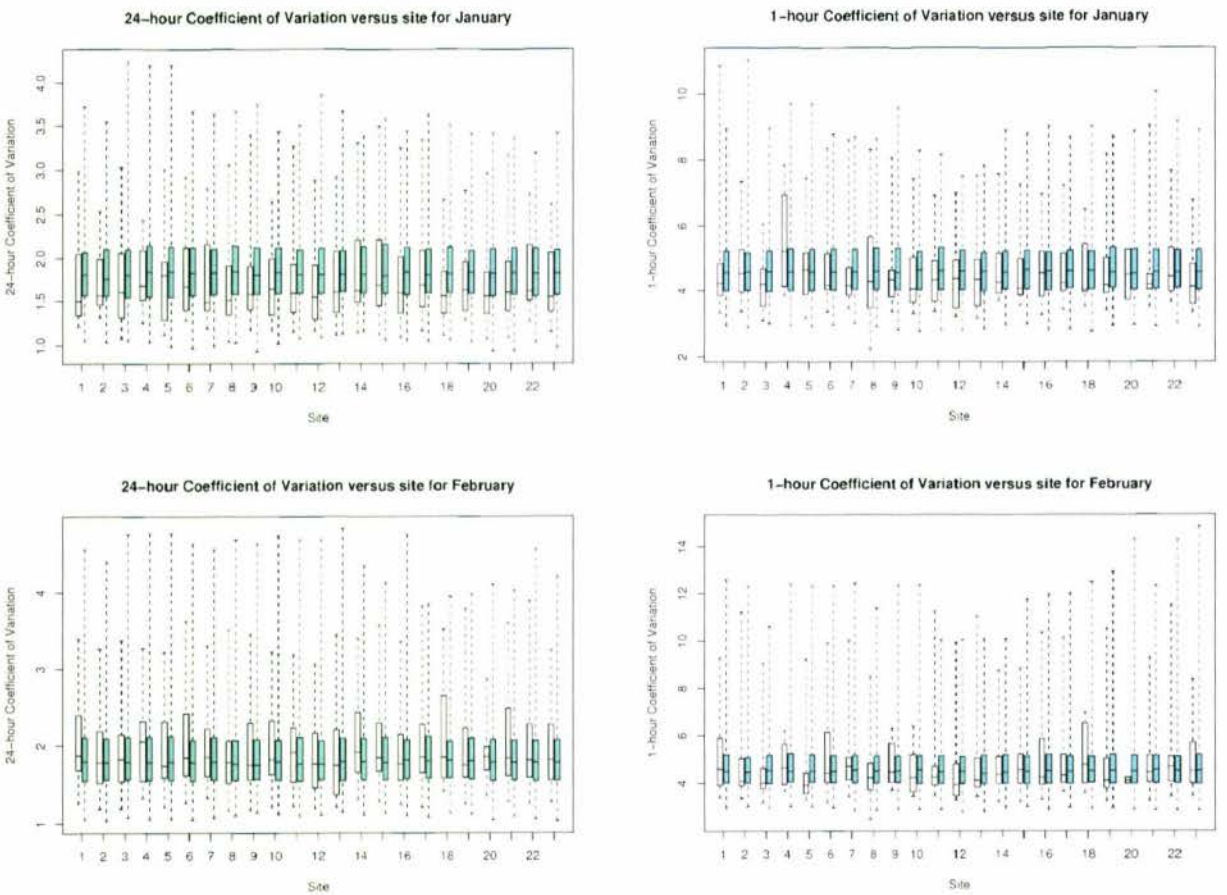
White = Historical; Shaded = $Model_B$

Figure C.5: Monthly Means by site - September and October



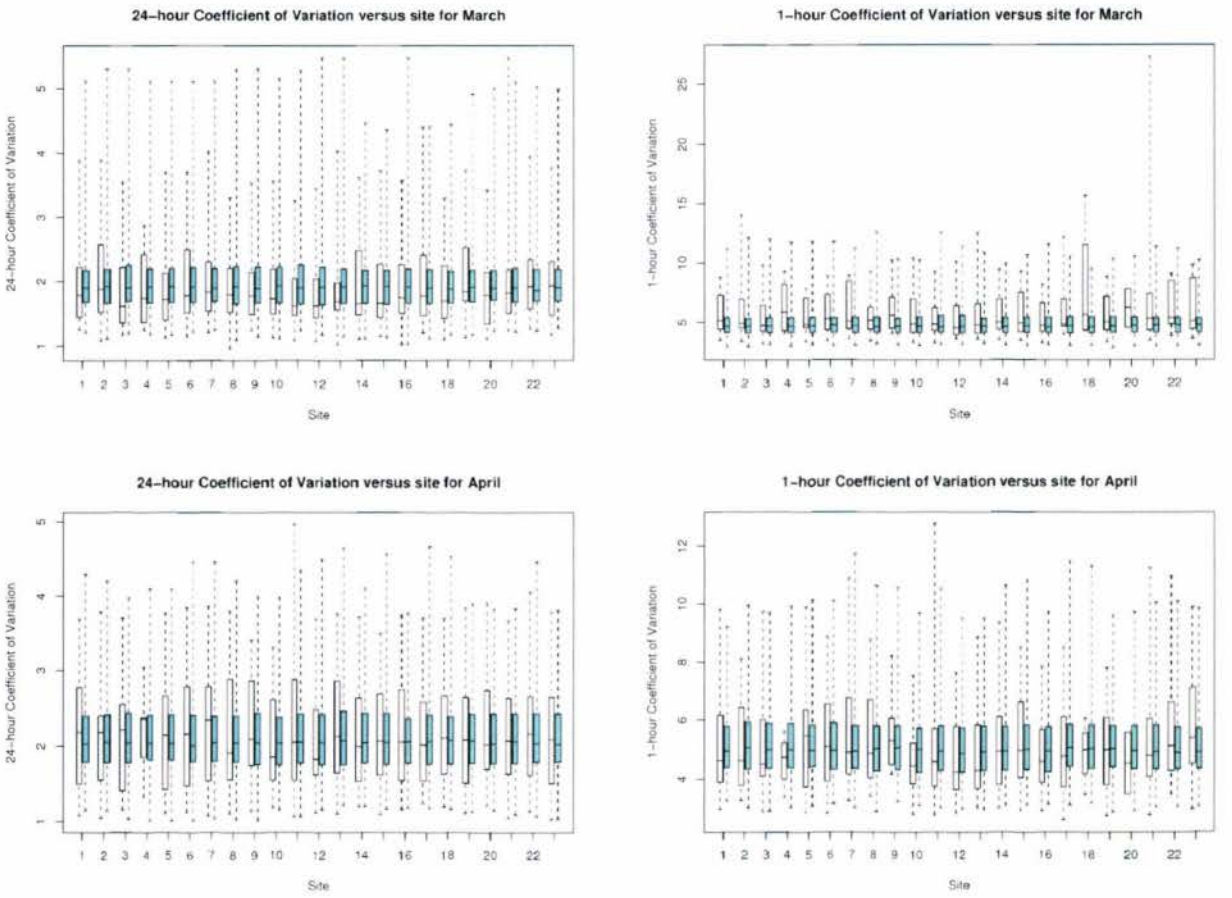
White = Historical; Shaded = $Model_B$

Figure C.6: Monthly Means by site - November and December



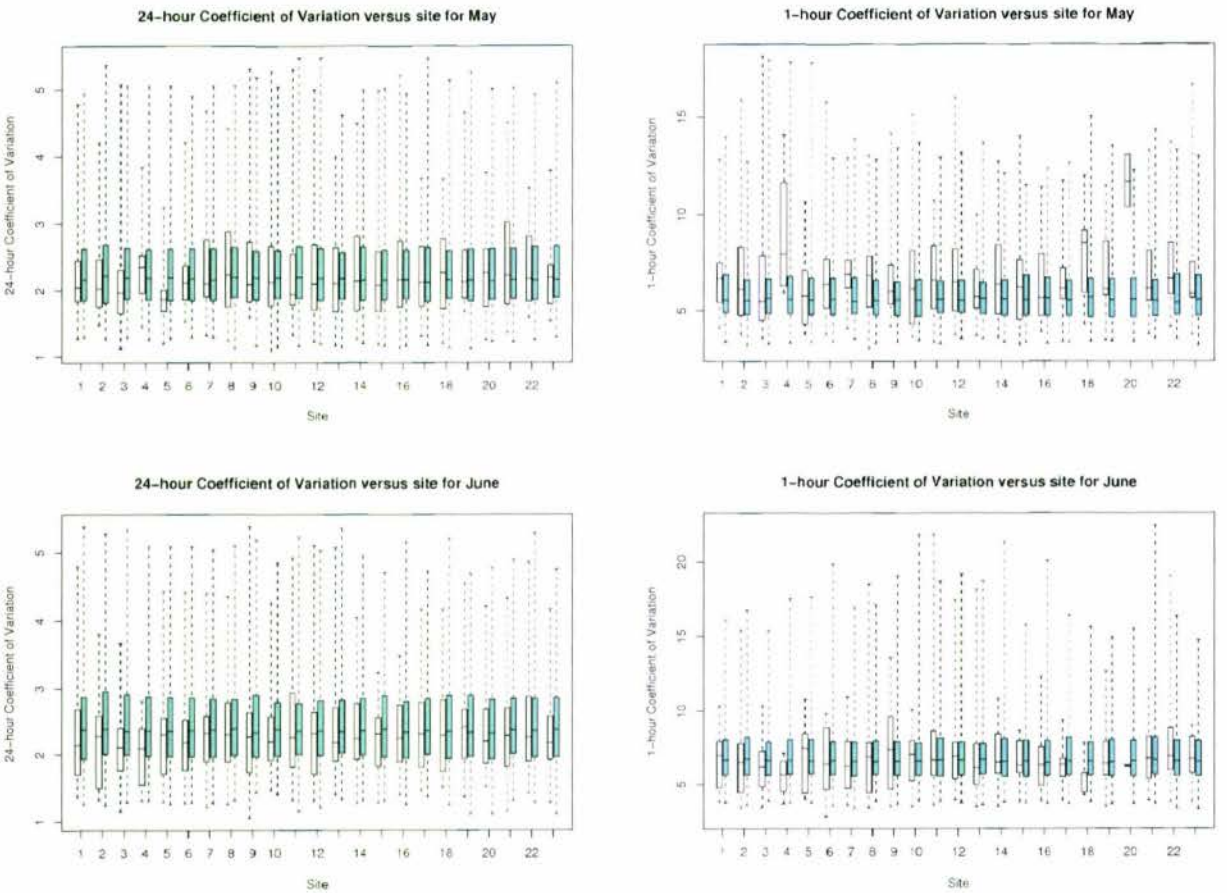
White = Historical; Shaded = $Model_B$

Figure C.7: Monthly Coefficient of Variation by site - January and February



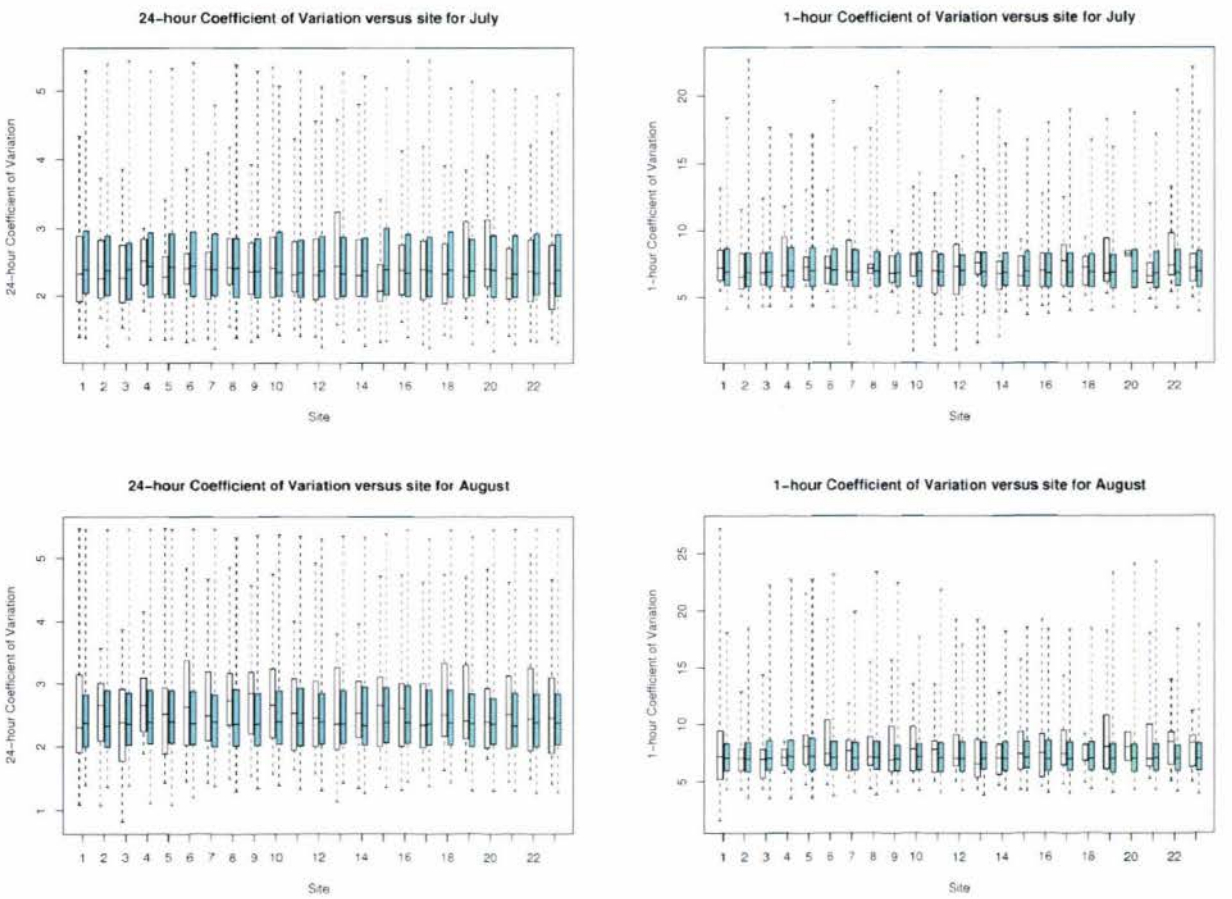
White = Historical; Shaded = $Model_B$

Figure C.8: Monthly Coefficient of Variation by site - March and April



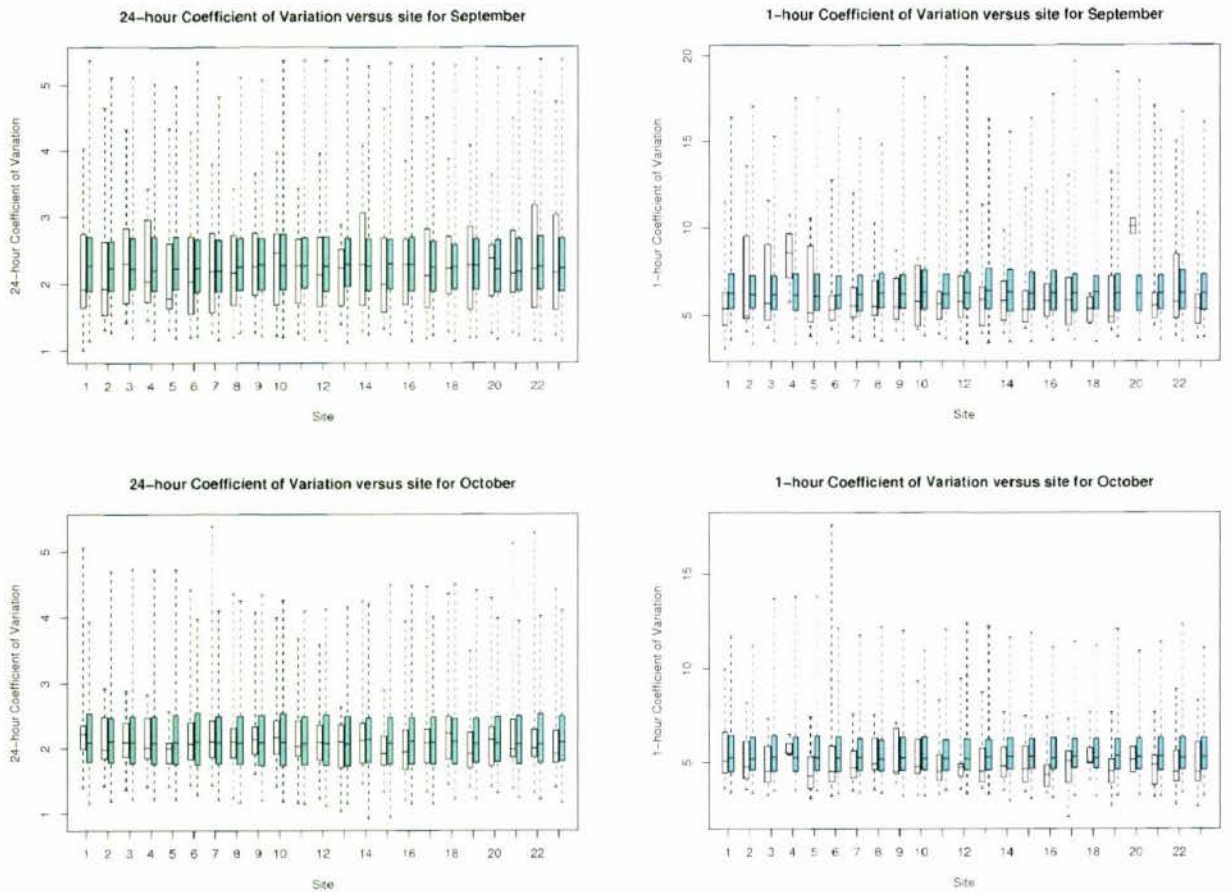
White = Historical; Shaded = $Model_B$

Figure C.9: Monthly Coefficient of Variation by site - May and June



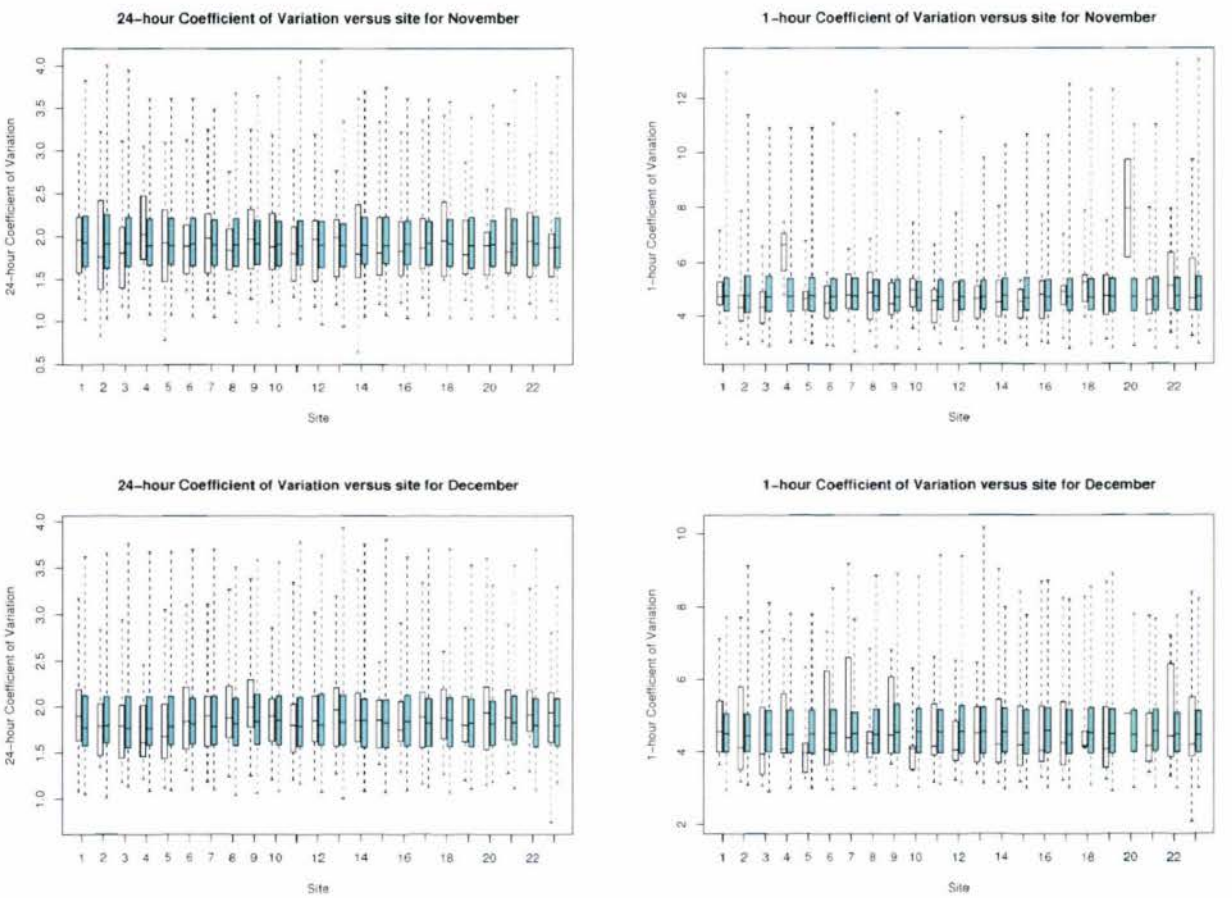
White = Historical; Shaded = $Model_B$

Figure C.10: Monthly Coefficient of Variation by site - July and August



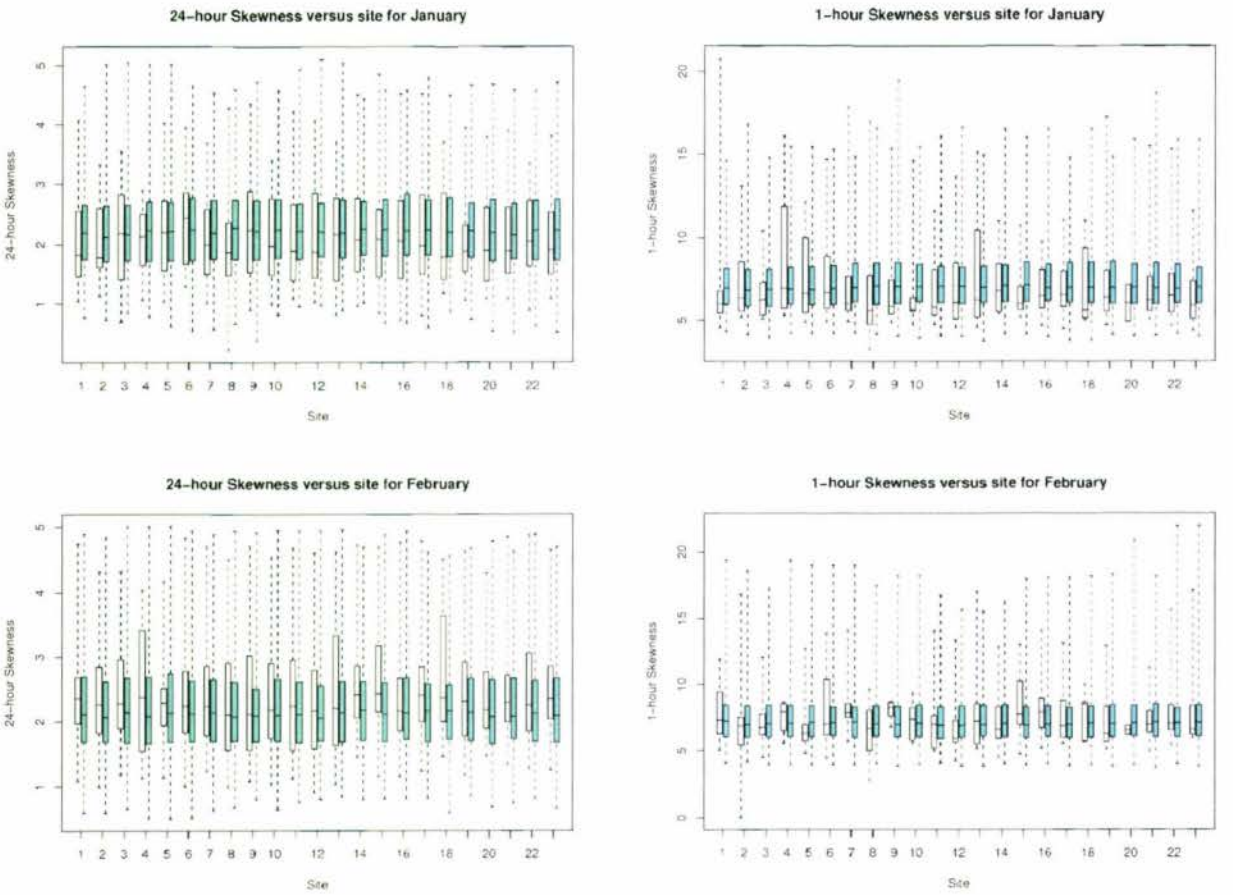
White = Historical; Shaded = $Model_B$

Figure C.11: Monthly Coefficient of Variation by site - September and October



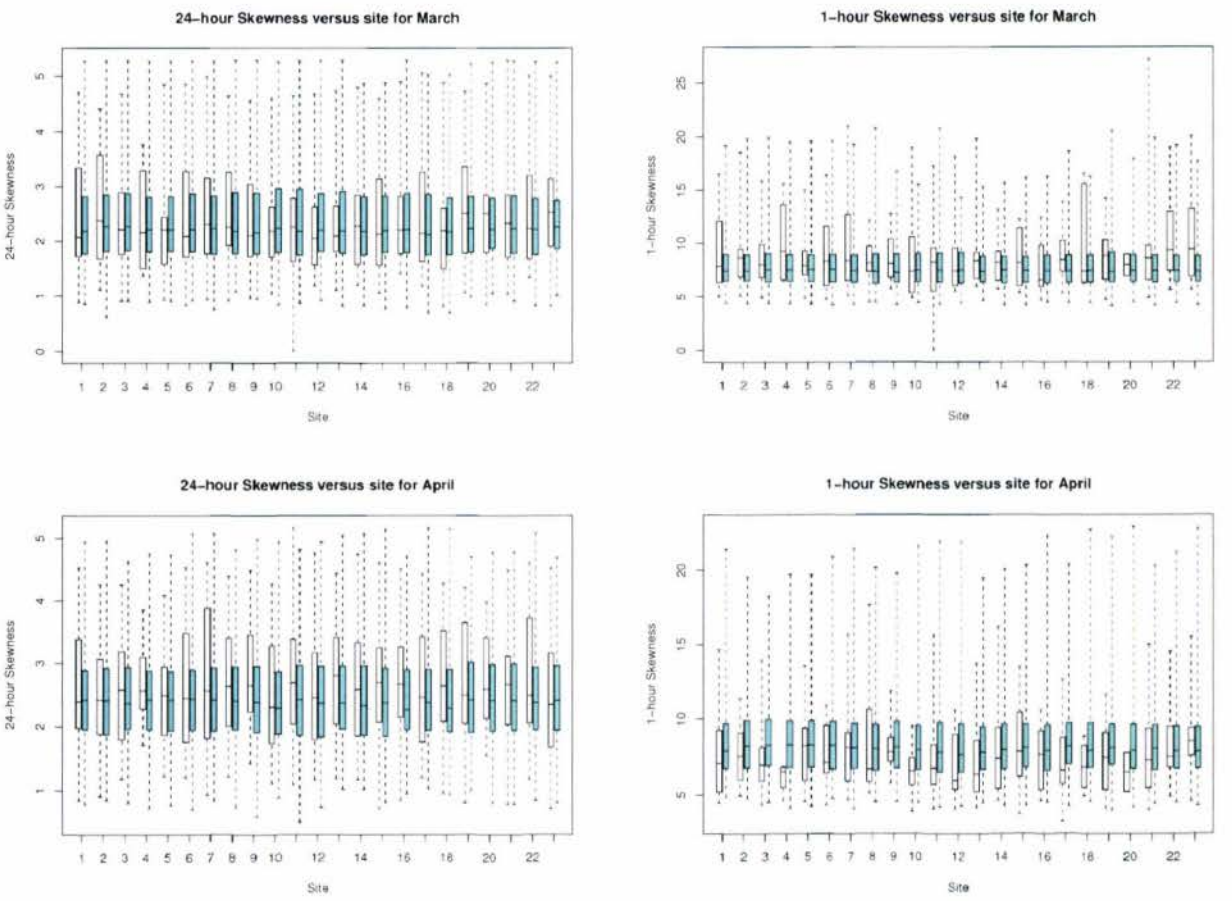
White = Historical; Shaded = $Model_B$

Figure C.12: Monthly Coefficient of Variation by site - November and December



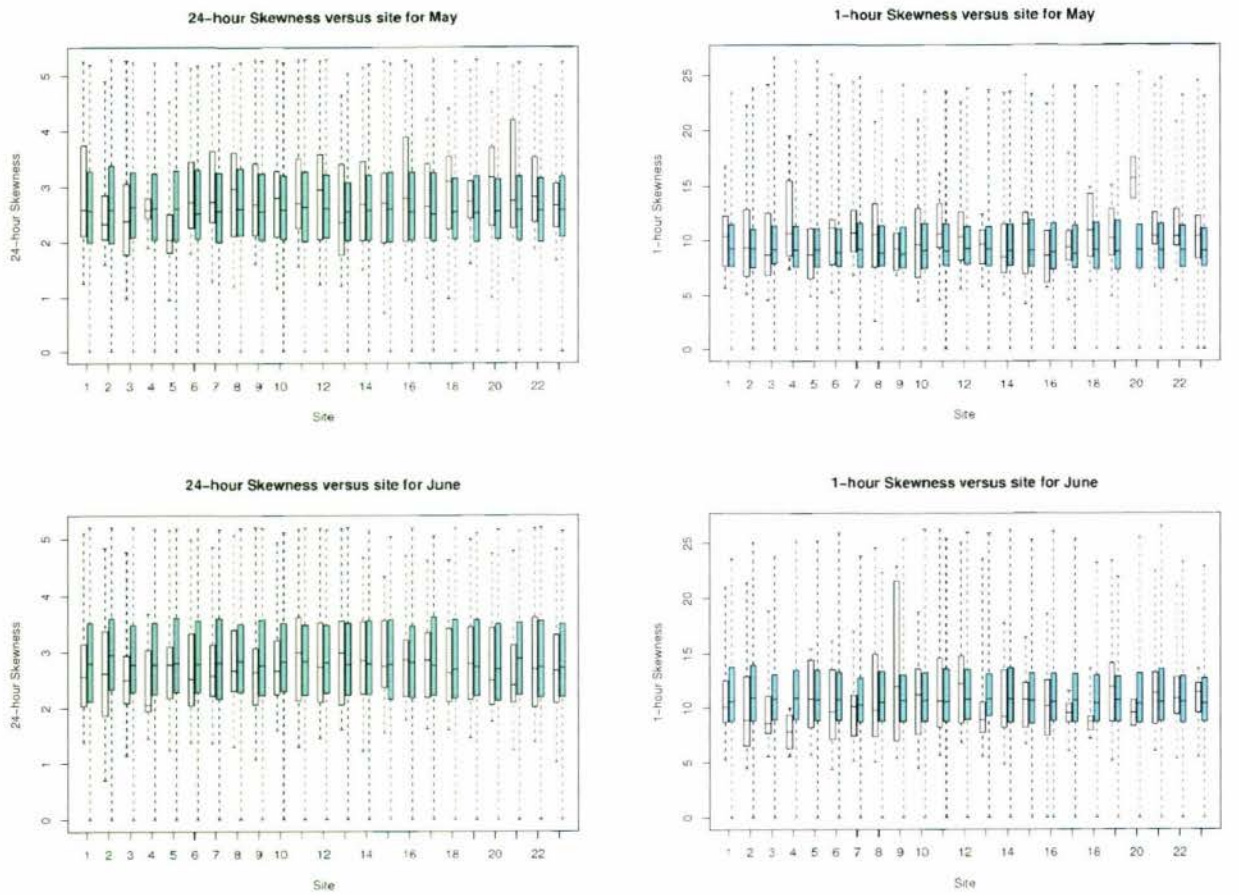
White = Historical; Shaded = $Model_B$

Figure C.13: Monthly Skewness by site - January and February



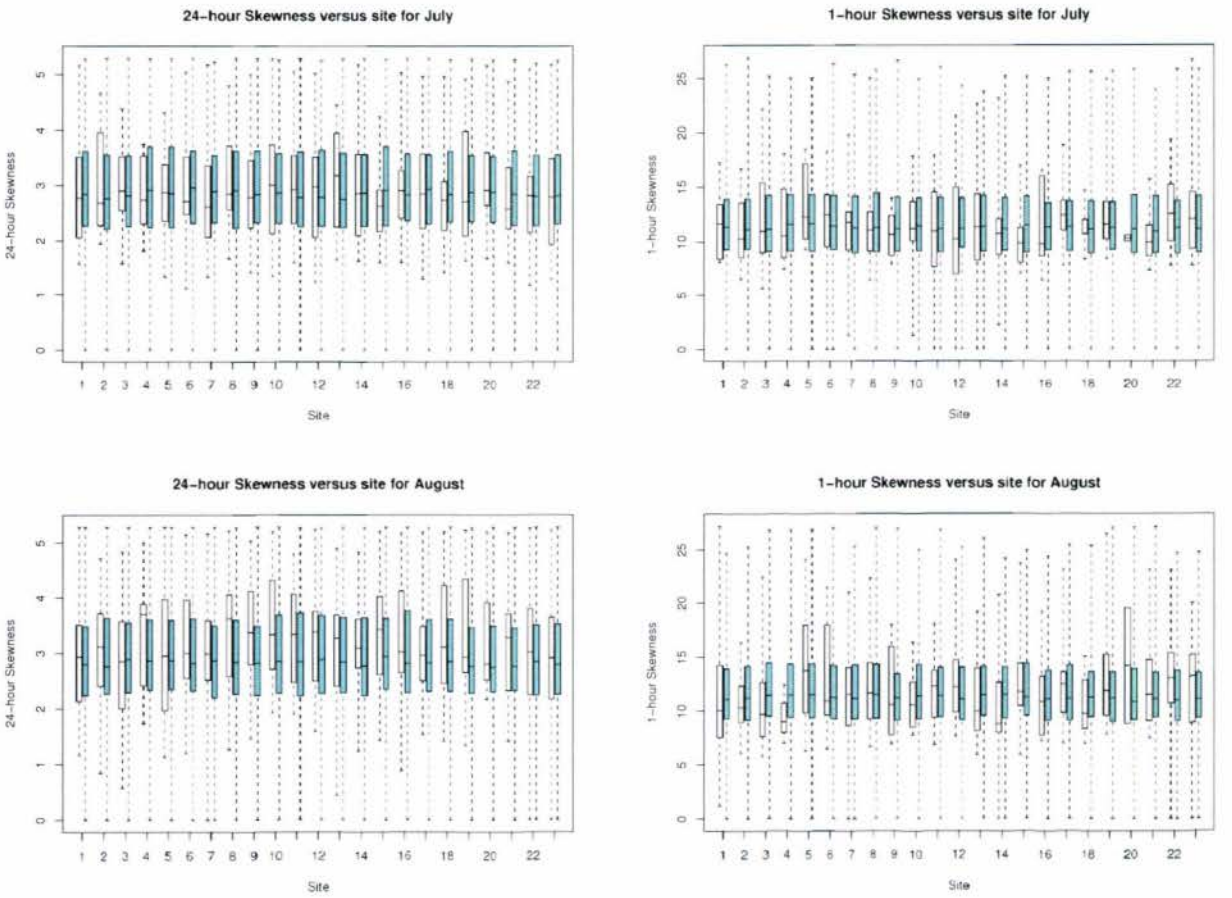
White = Historical; Shaded = $Model_B$

Figure C.14: Monthly Skewness by site - March and April



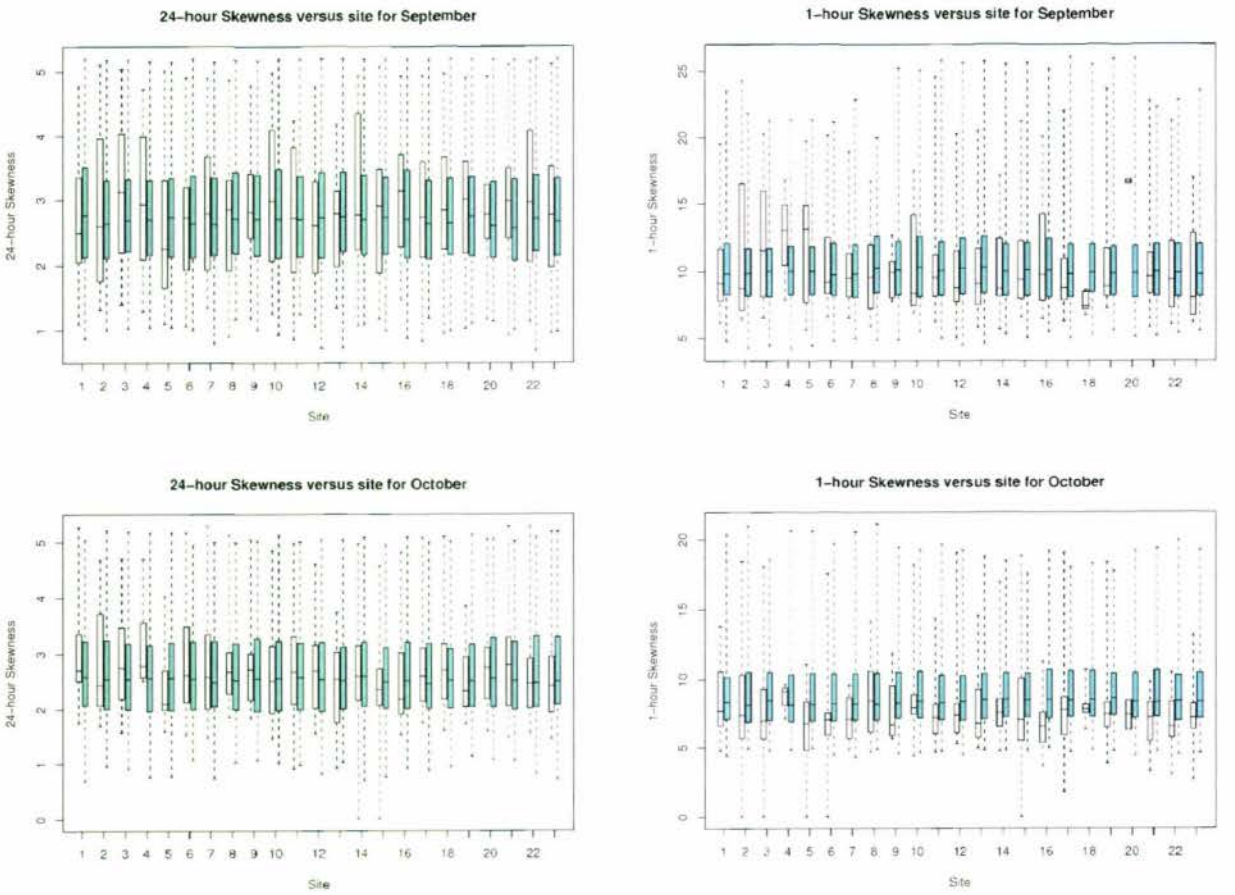
White = Historical; Shaded = $Model_B$

Figure C.15: Monthly Skewness by site - May and June



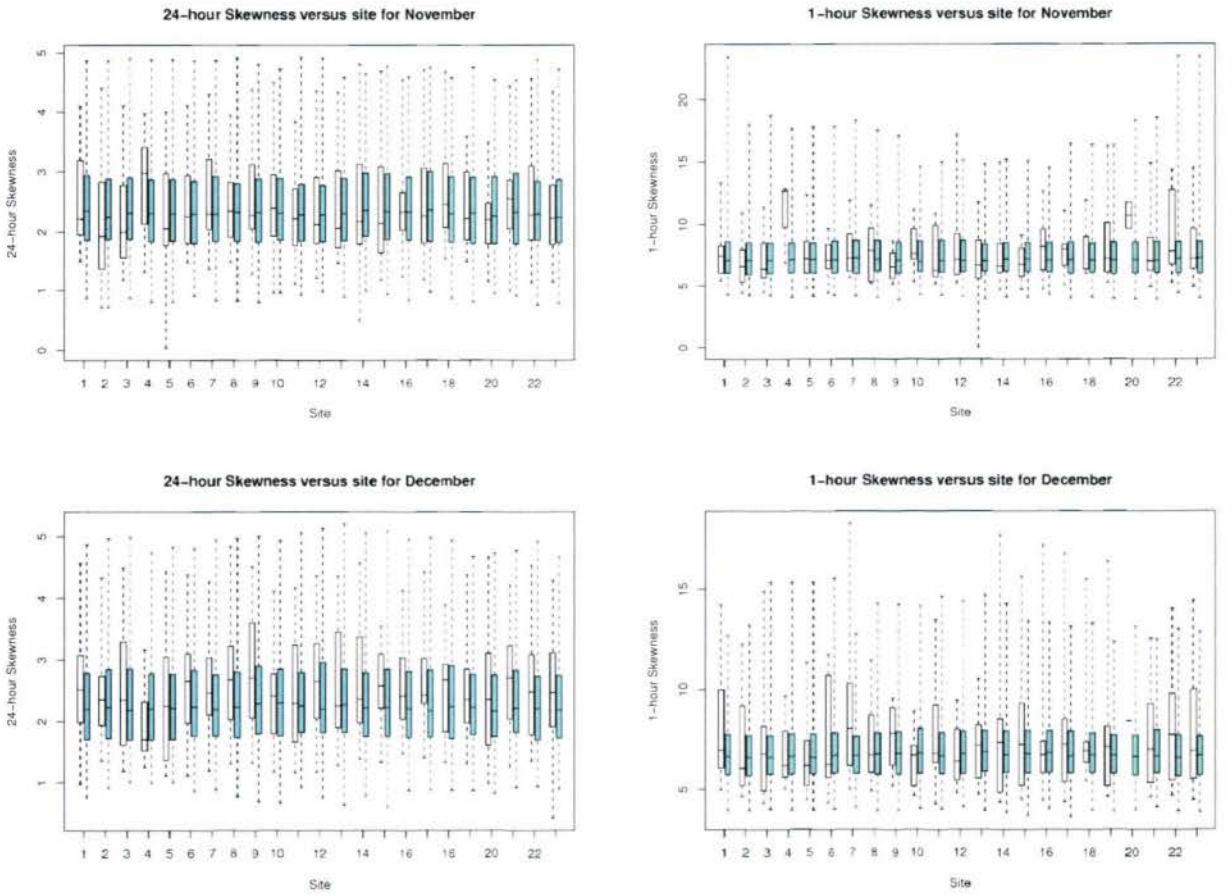
White = Historical; Shaded = $Model_B$

Figure C.16: Monthly Skewness by site - July and August



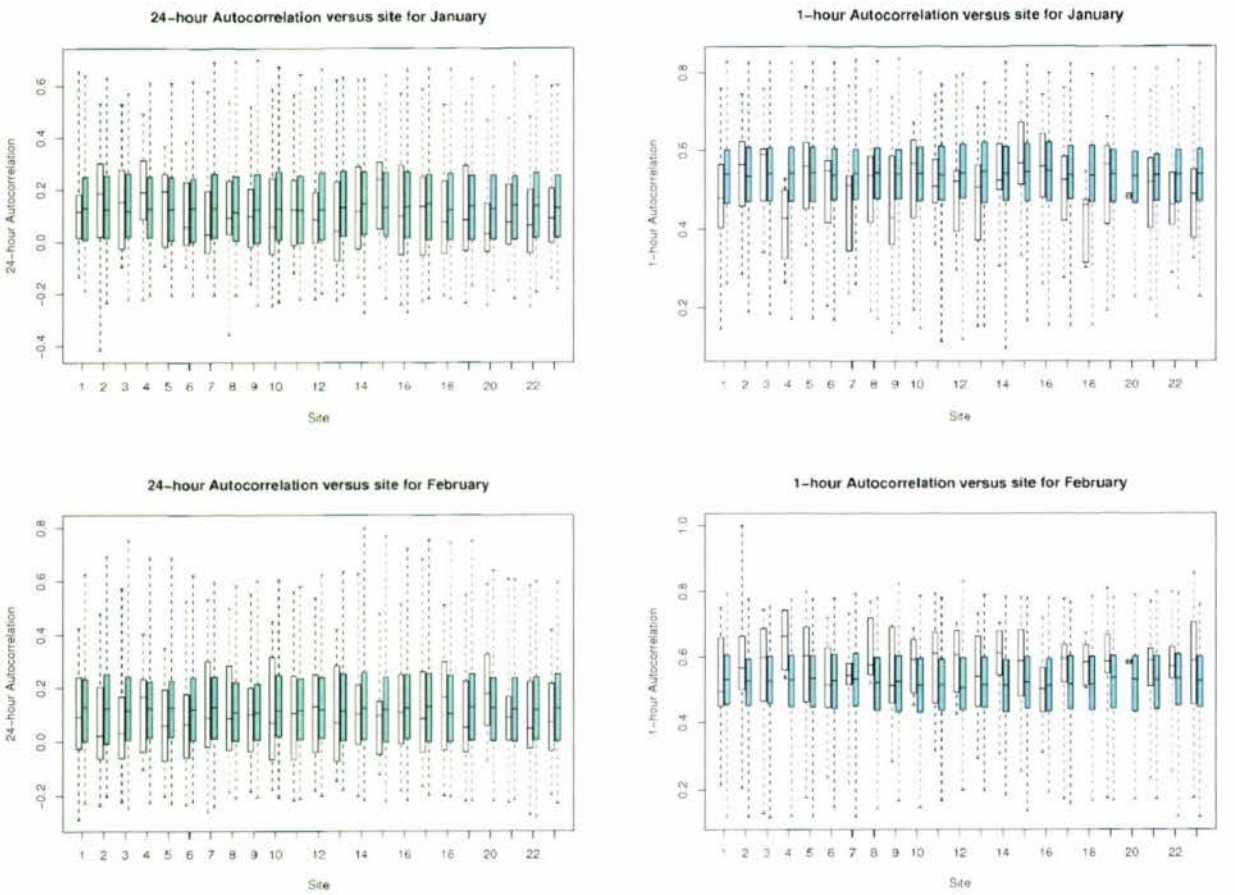
White = Historical; Shaded = $Model_B$

Figure C.17: Monthly Skewness by site - September and October



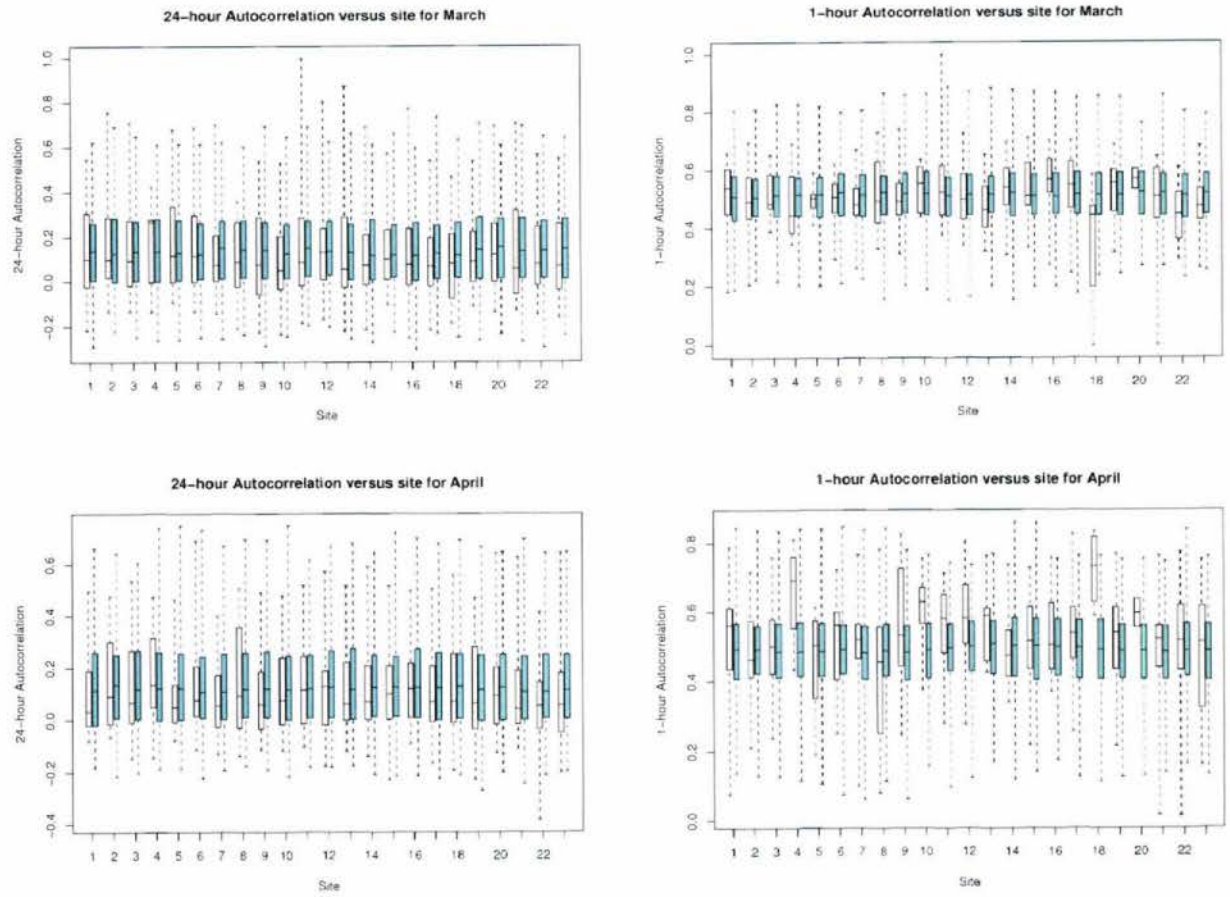
White = Historical; Shaded = $Model_B$

Figure C.18: Monthly Skewness by site - November and December



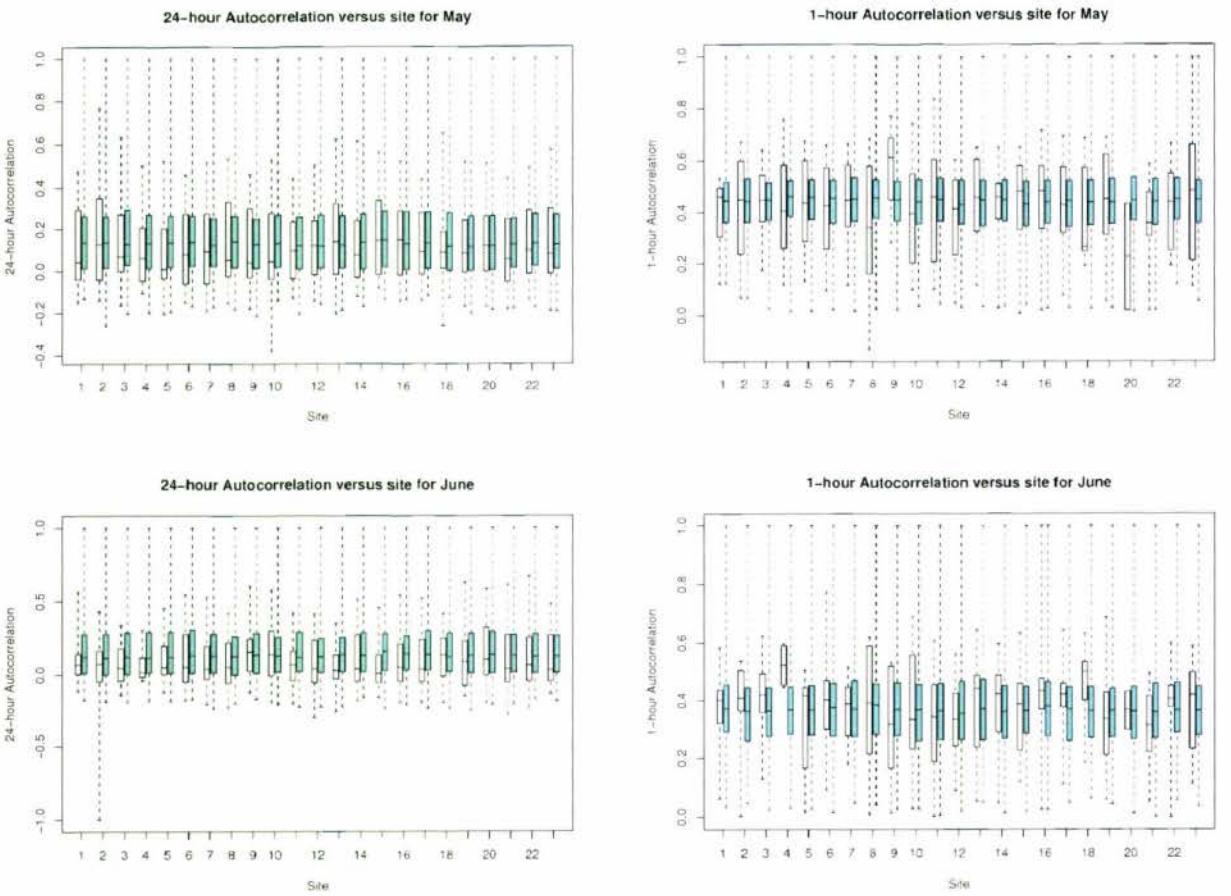
White = Historical; Shaded = $Model_B$

Figure C.19: Monthly Autocorrelation by site - January and February



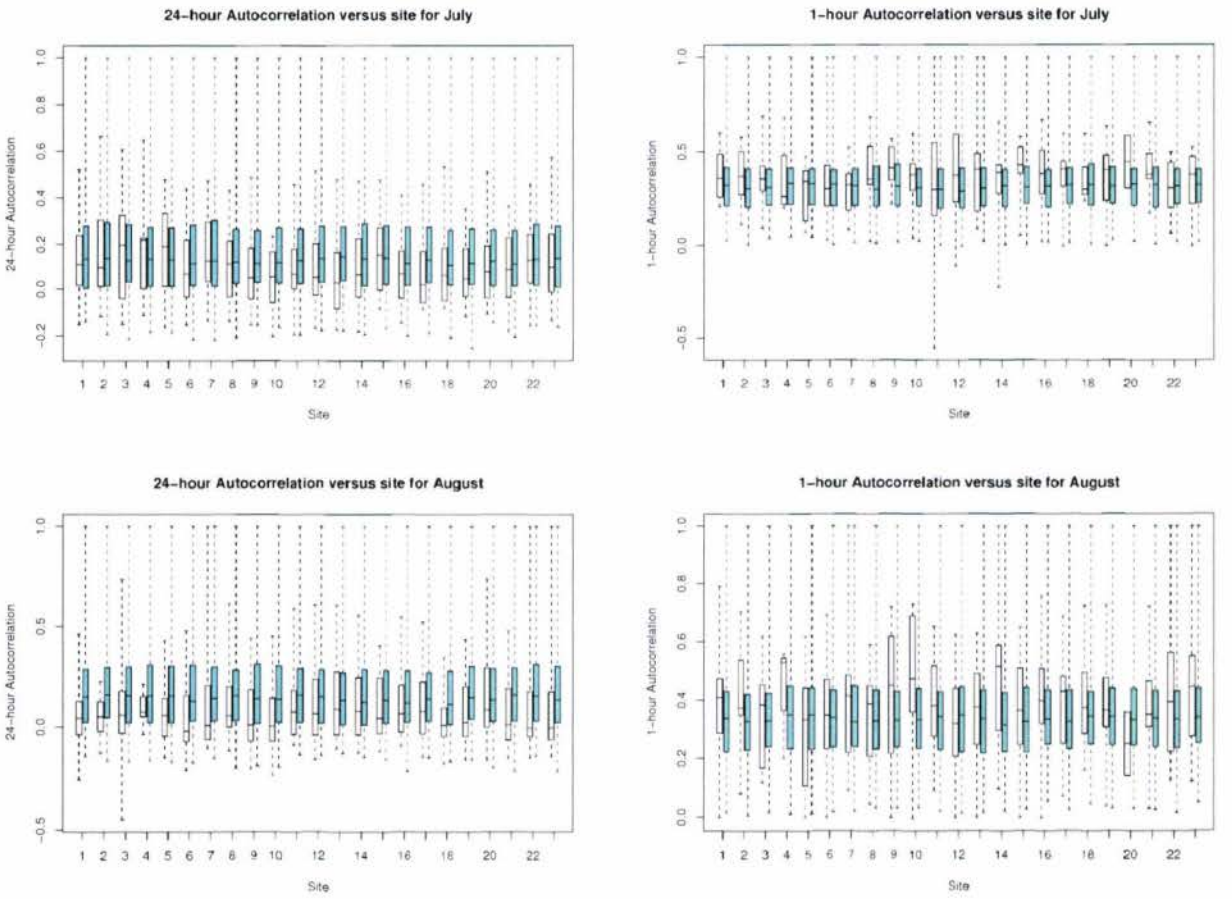
White = Historical; Shaded = $Model_B$

Figure C.20: Monthly Autocorrelation by site - March and April



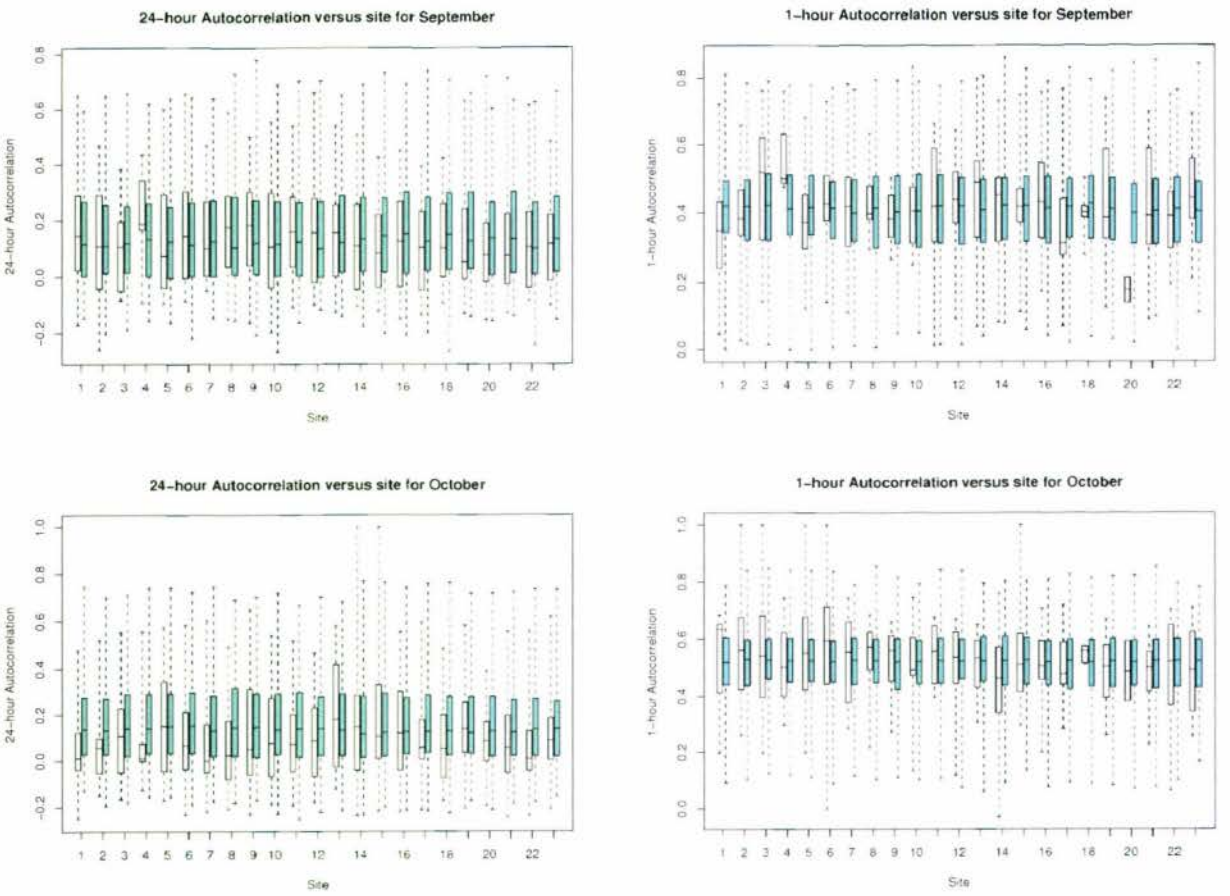
White = Historical; Shaded = $Model_B$

Figure C.21: Monthly Autocorrelation by site - May and June



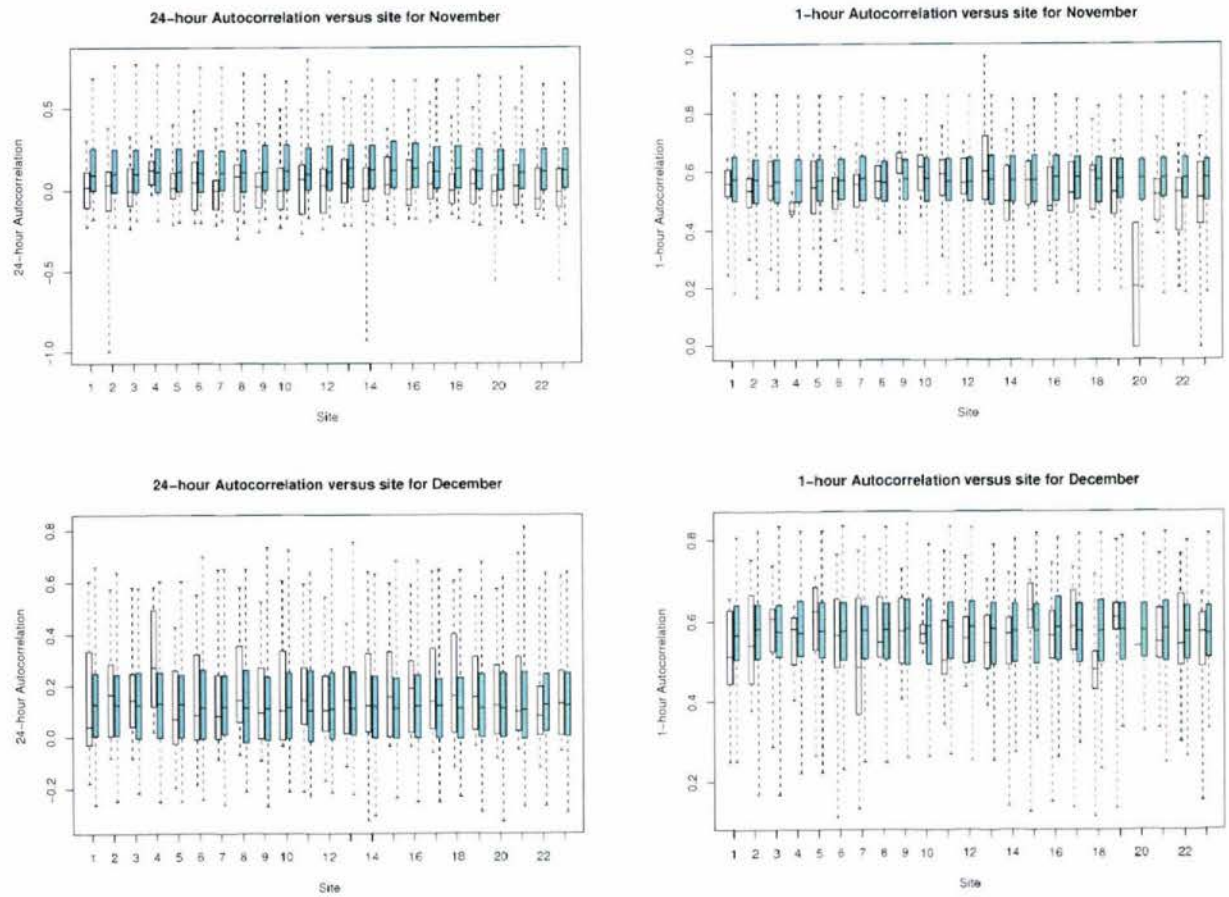
White = Historical; Shaded = $Model_B$

Figure C.22: Monthly Autocorrelation by site - July and August



White = Historical; Shaded = $Model_B$

Figure C.23: Monthly Autocorrelation by site - September and October



White = Historical; Shaded = $Model_B$

Figure C.24: Monthly Autocorrelation by site - November and December

D. INFILLING PLOTS

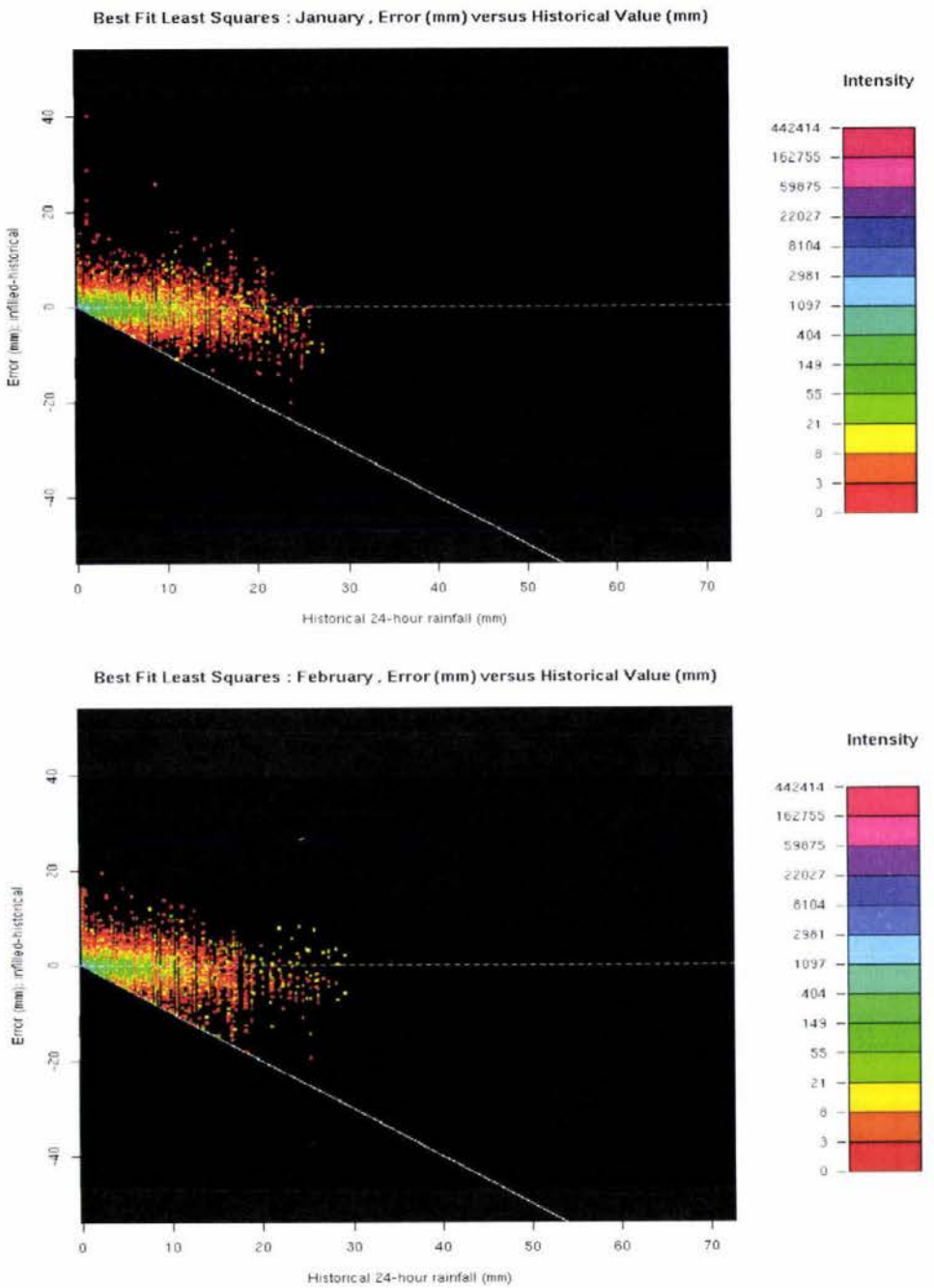


Figure D.1: BFLS Intensity plots: January and February

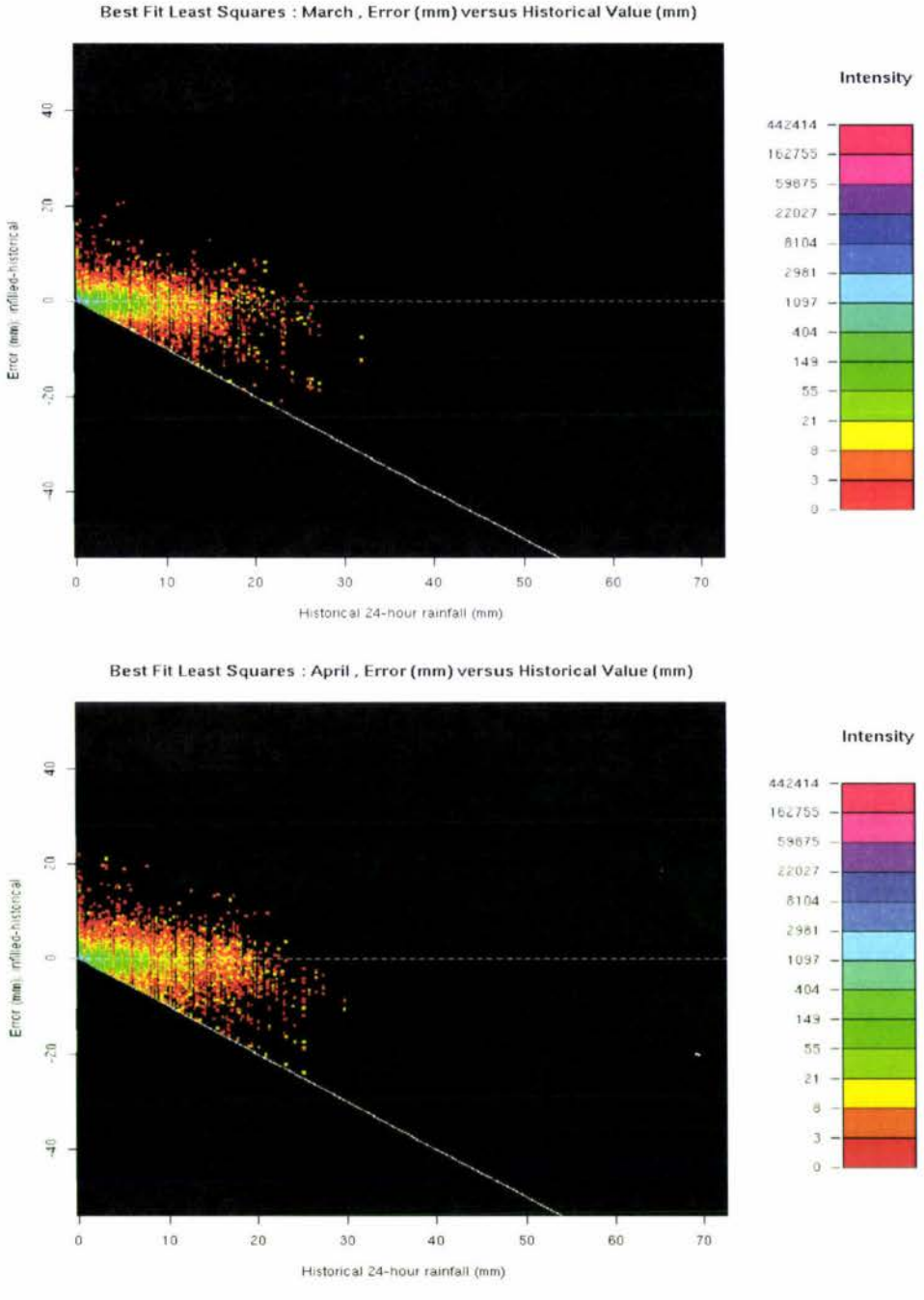


Figure D.2: BFLS Intensity plots: March and April

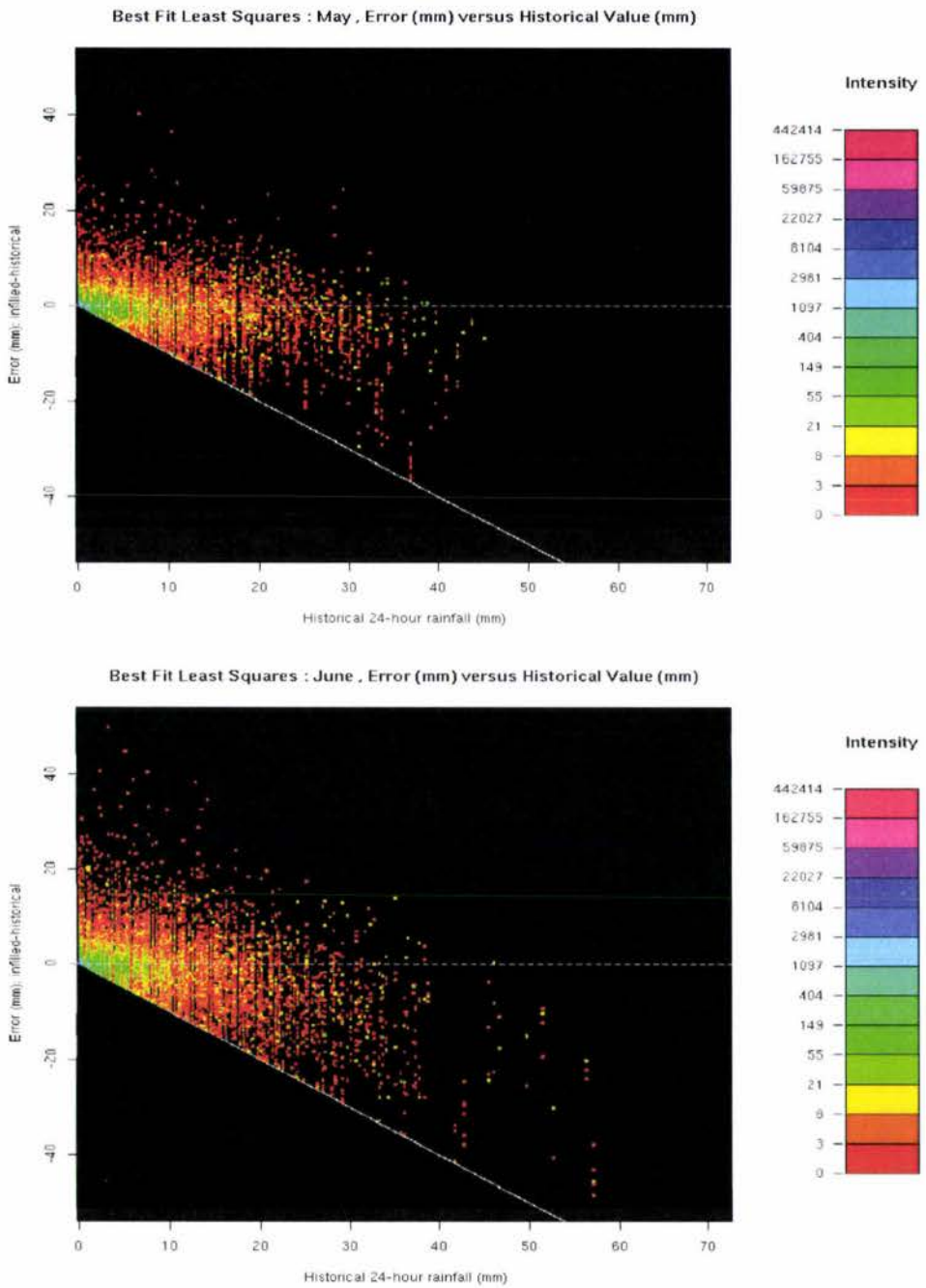


Figure D.3: BFLS Intensity plots: May and June

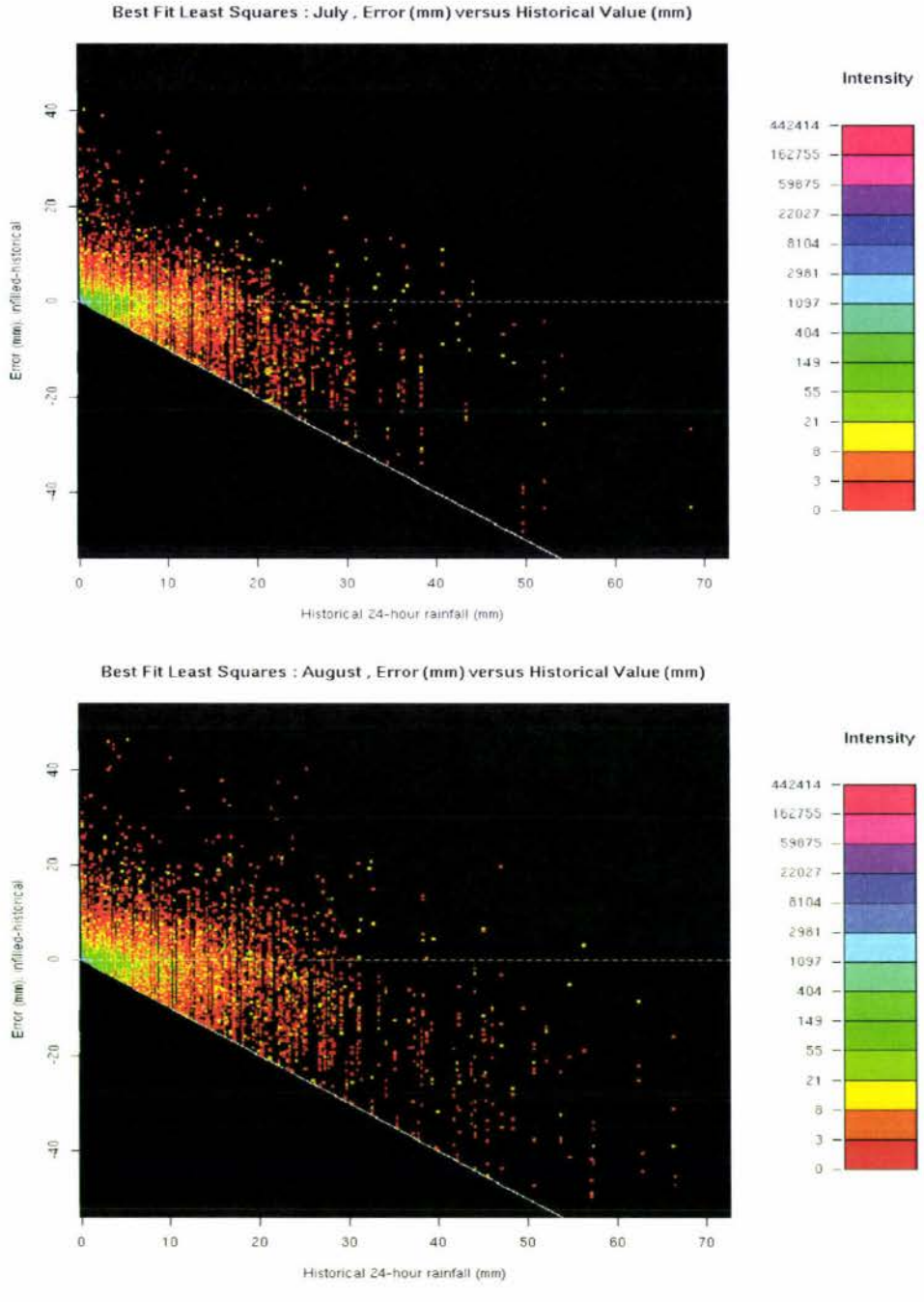


Figure D.4: BFLS Intensity plots: July and August

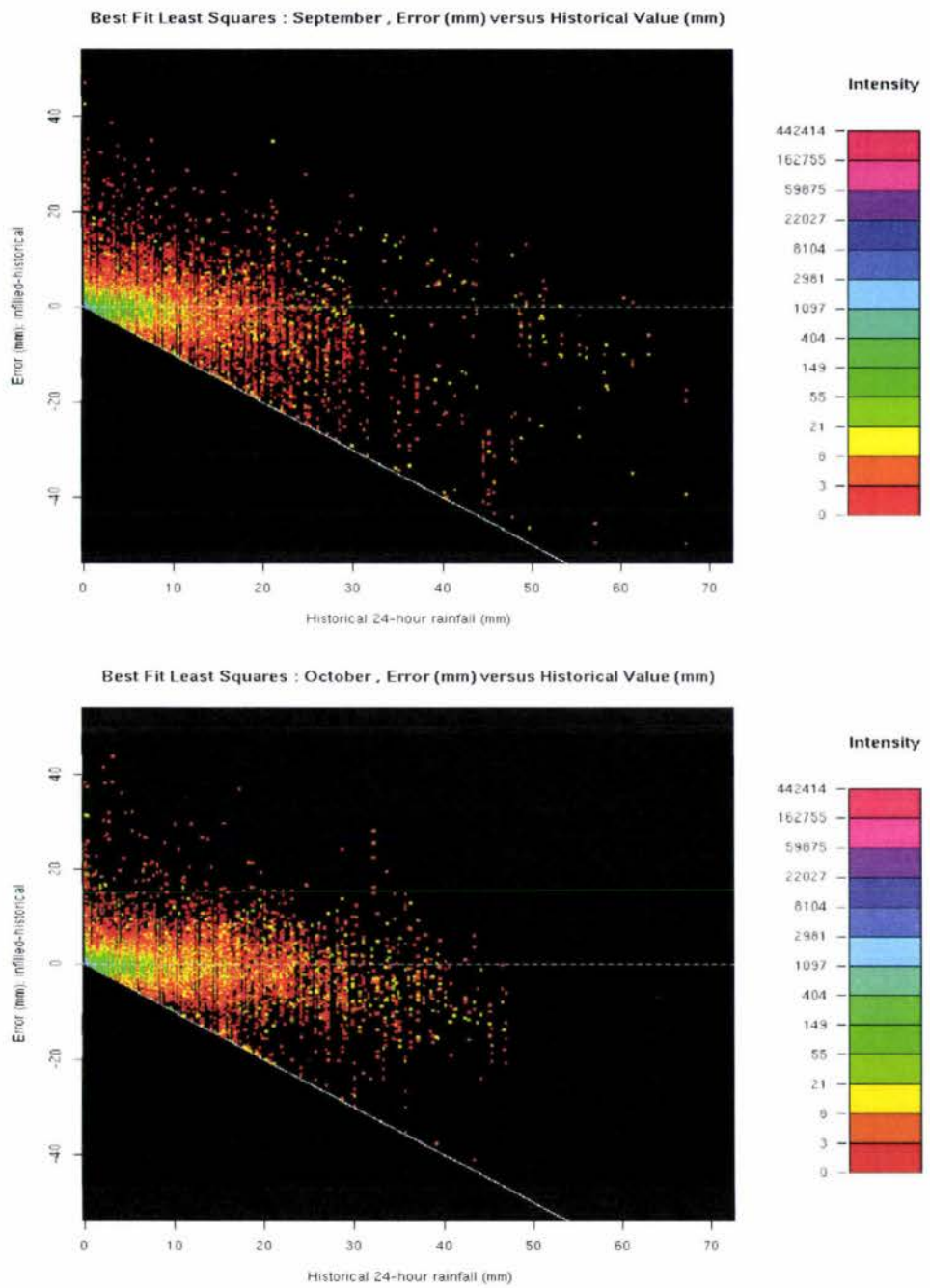


Figure D.5: BFLS Intensity plots: September and October

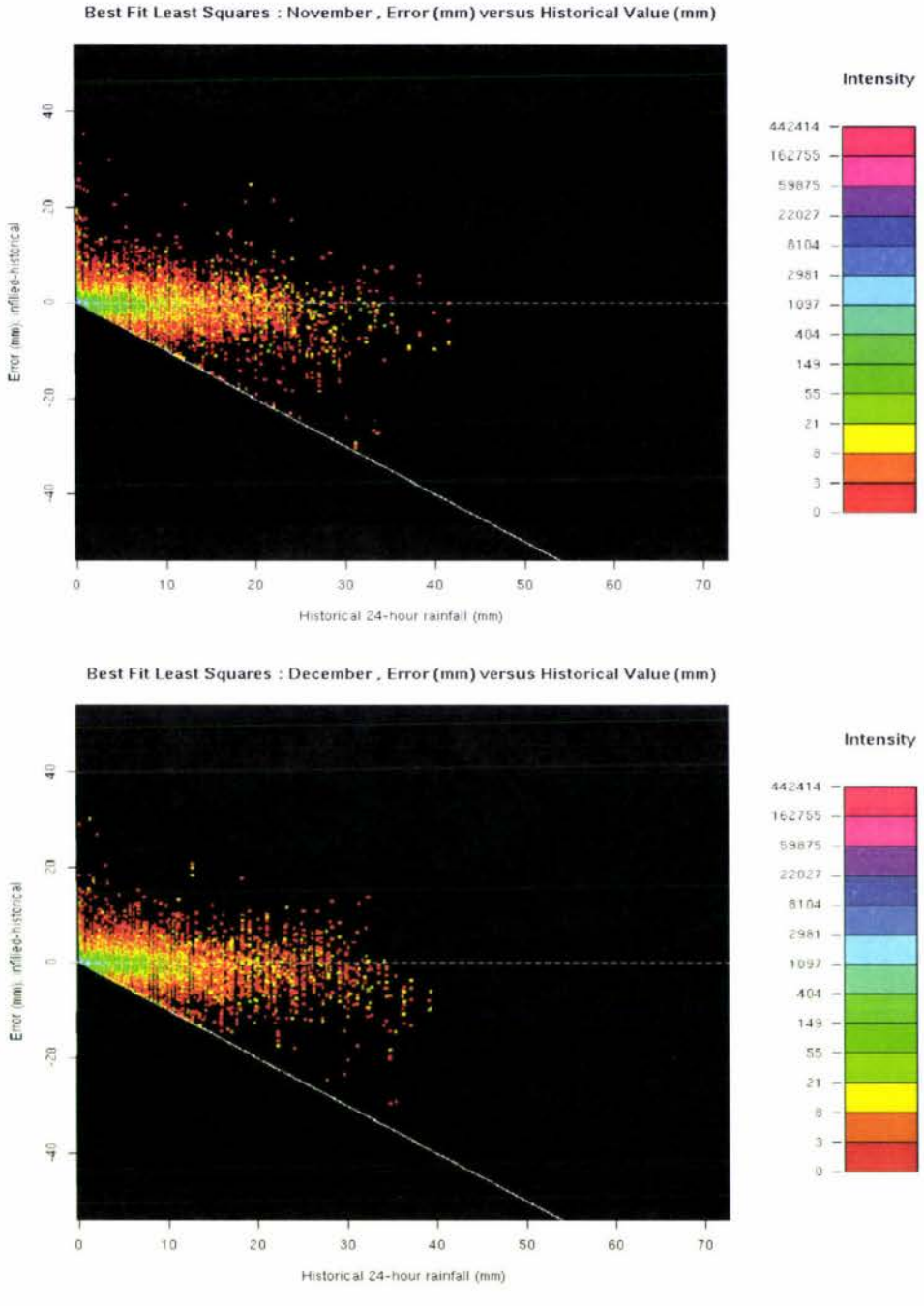
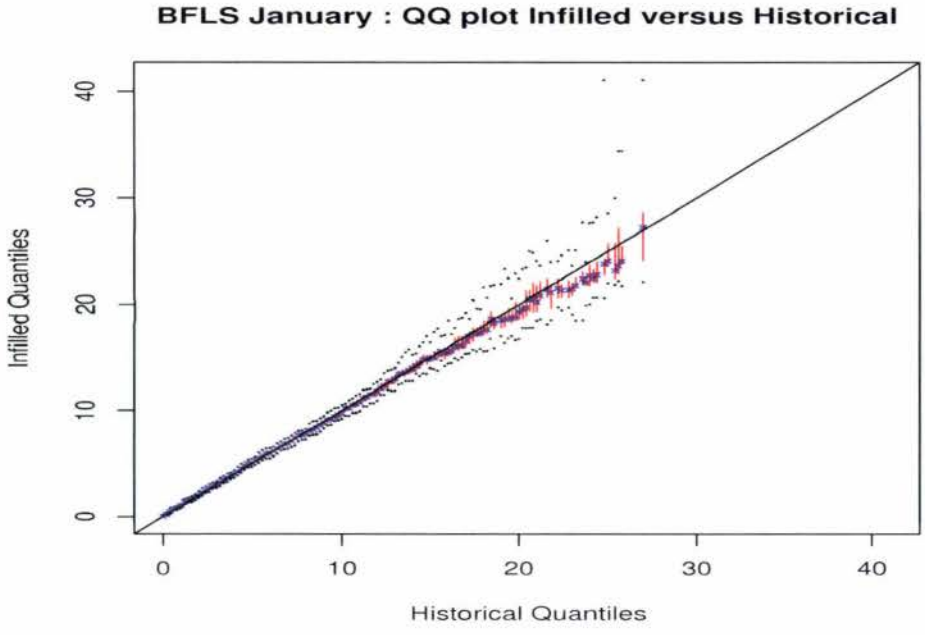
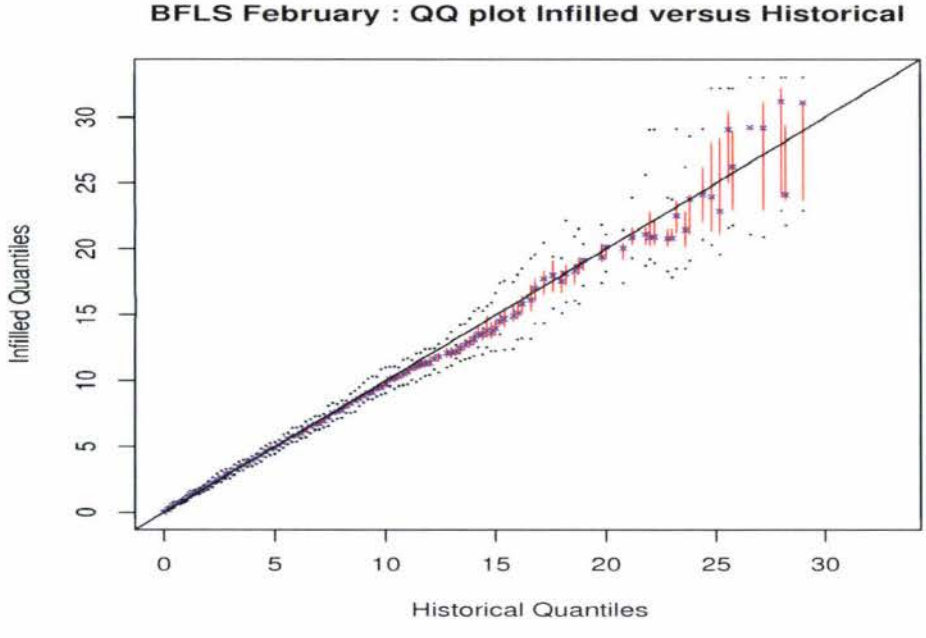


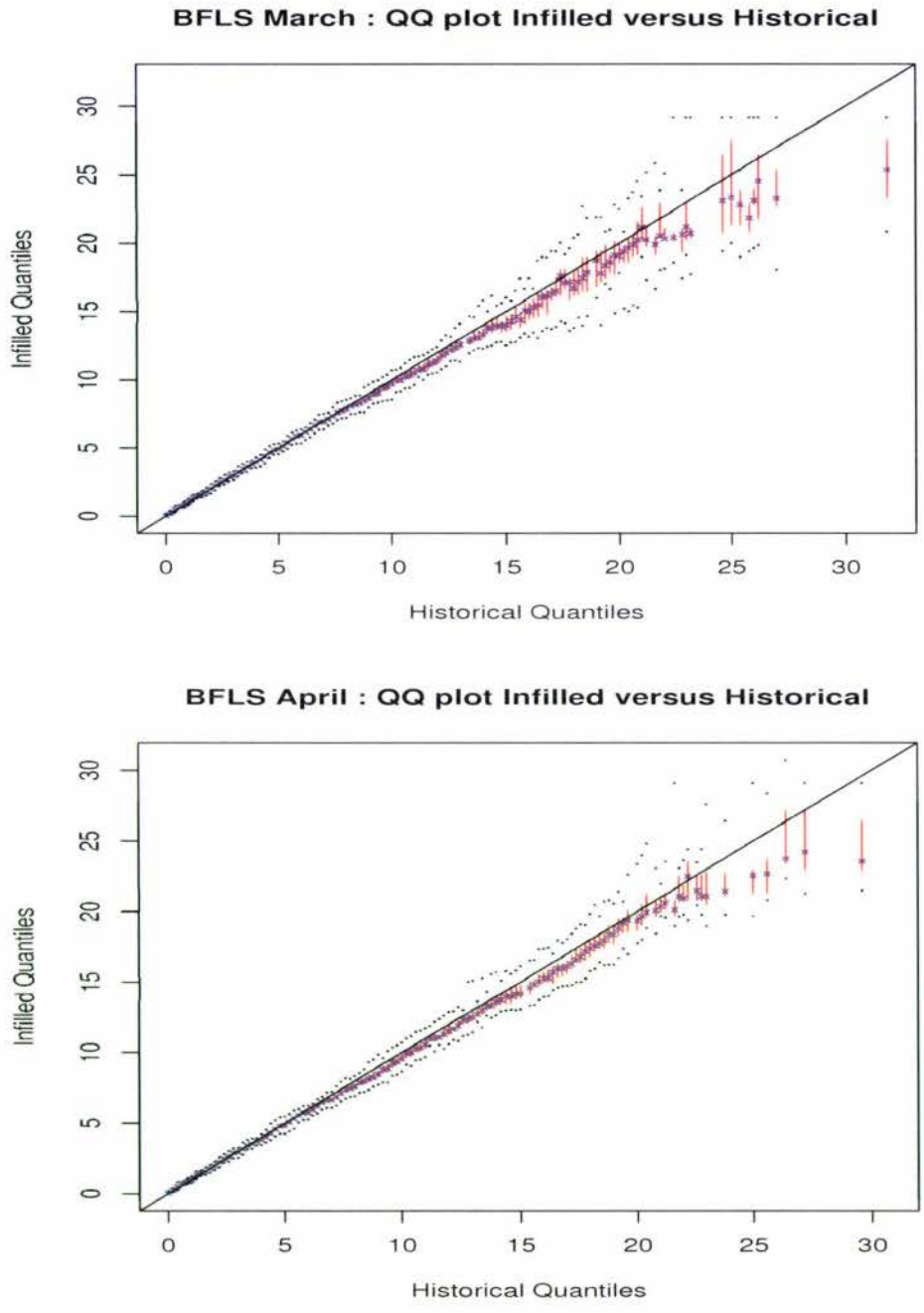
Figure D.6: BFLS Intensity plots: November and December





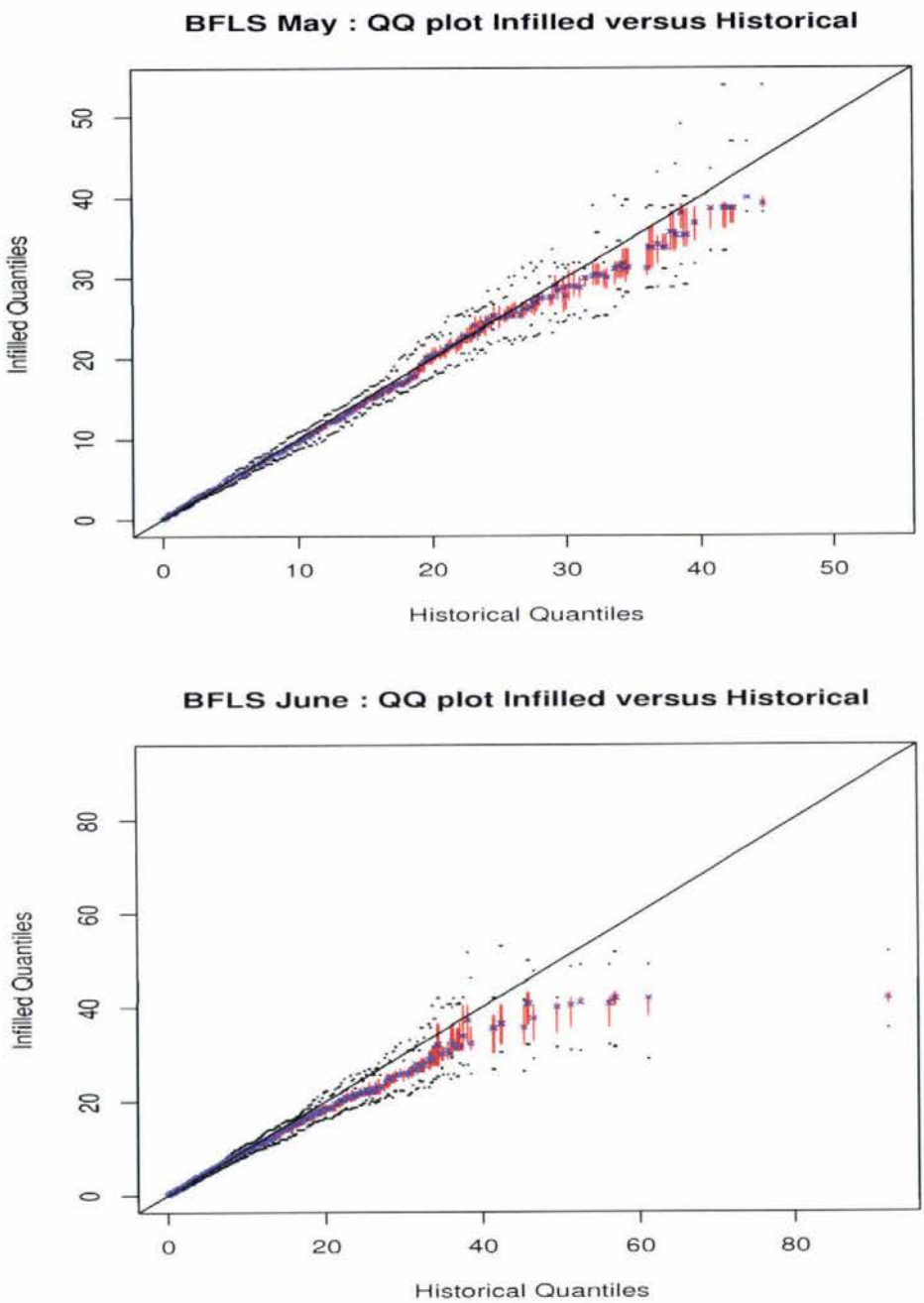
cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.7: BFLS Regional QQ plots: January and February



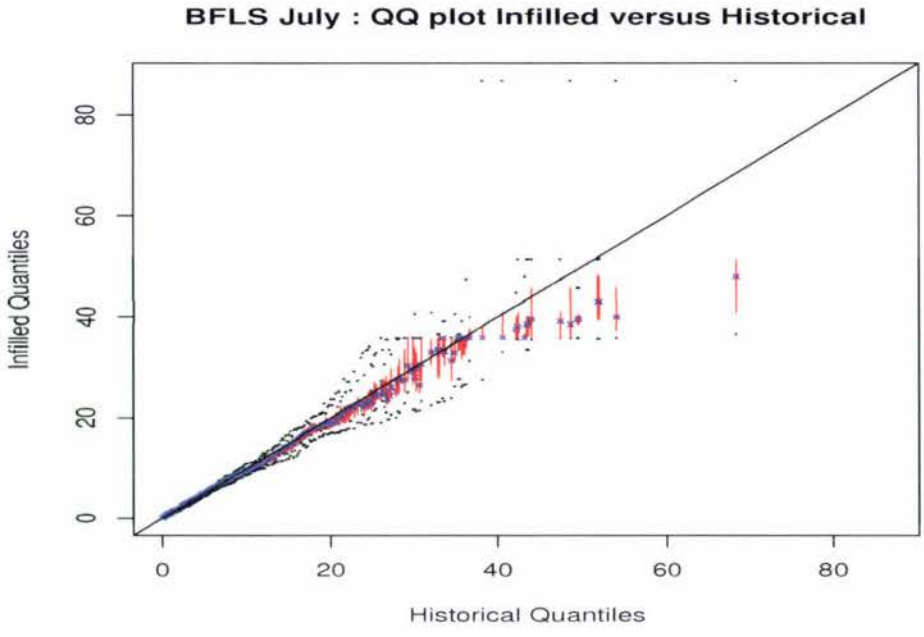
cross=median; vertical line = IQR error bar; dots = minima or maxima

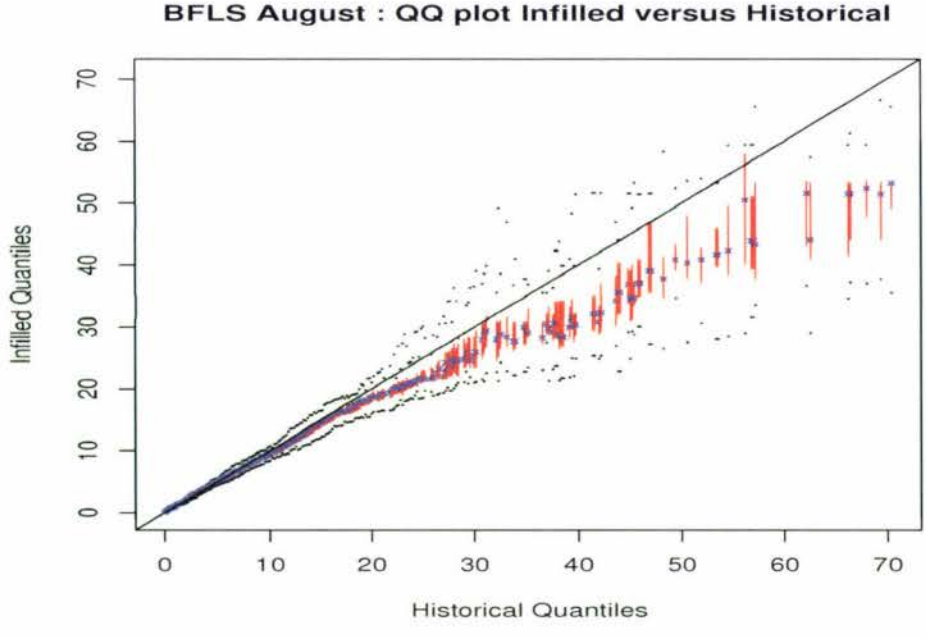
Figure D.8: BFLS Regional QQ plots: March and April



cross=median; vertical line = IQR error bar; dots = minima or maxima

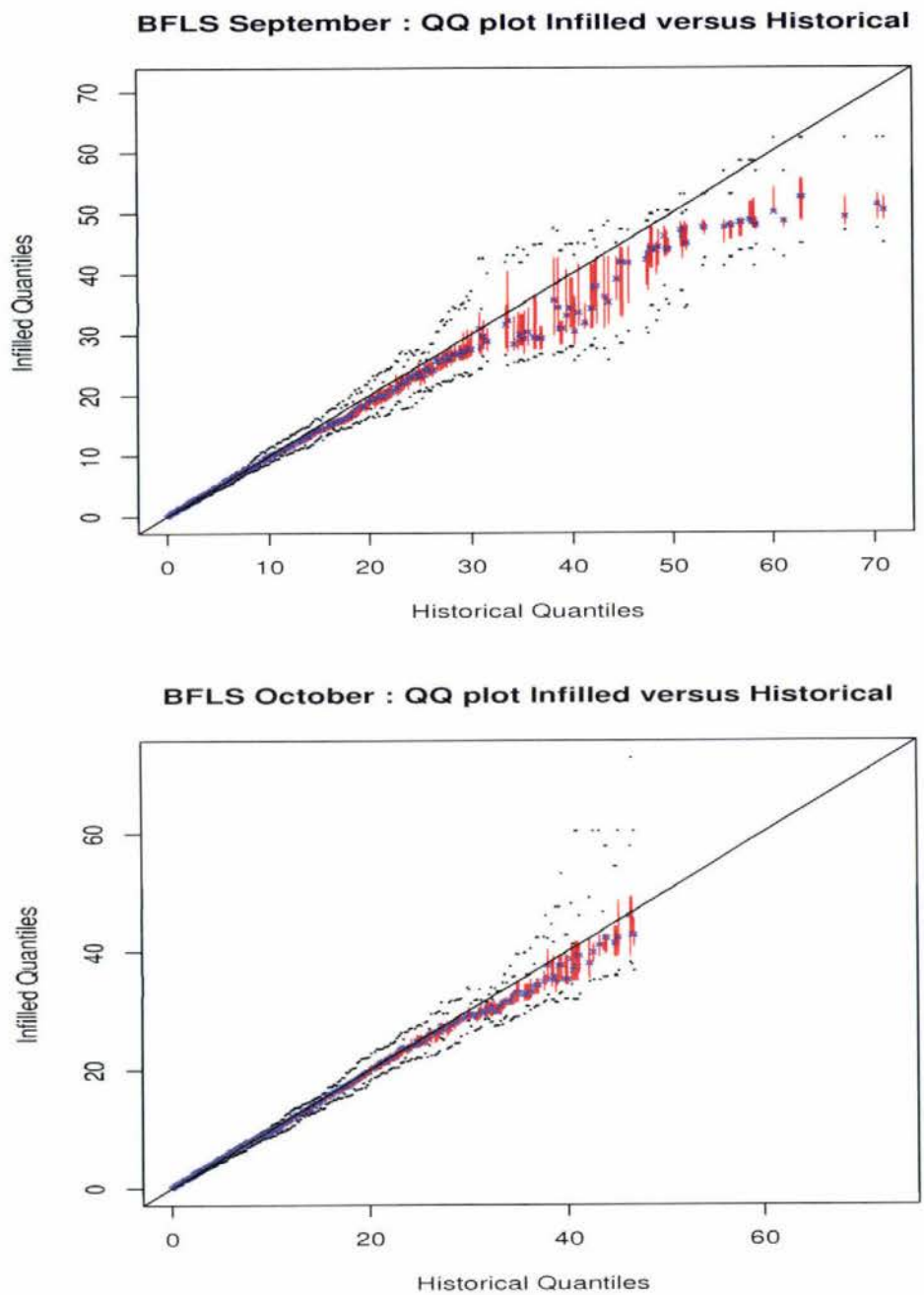
Figure D.9: BFLS Regional QQ plots: May and June





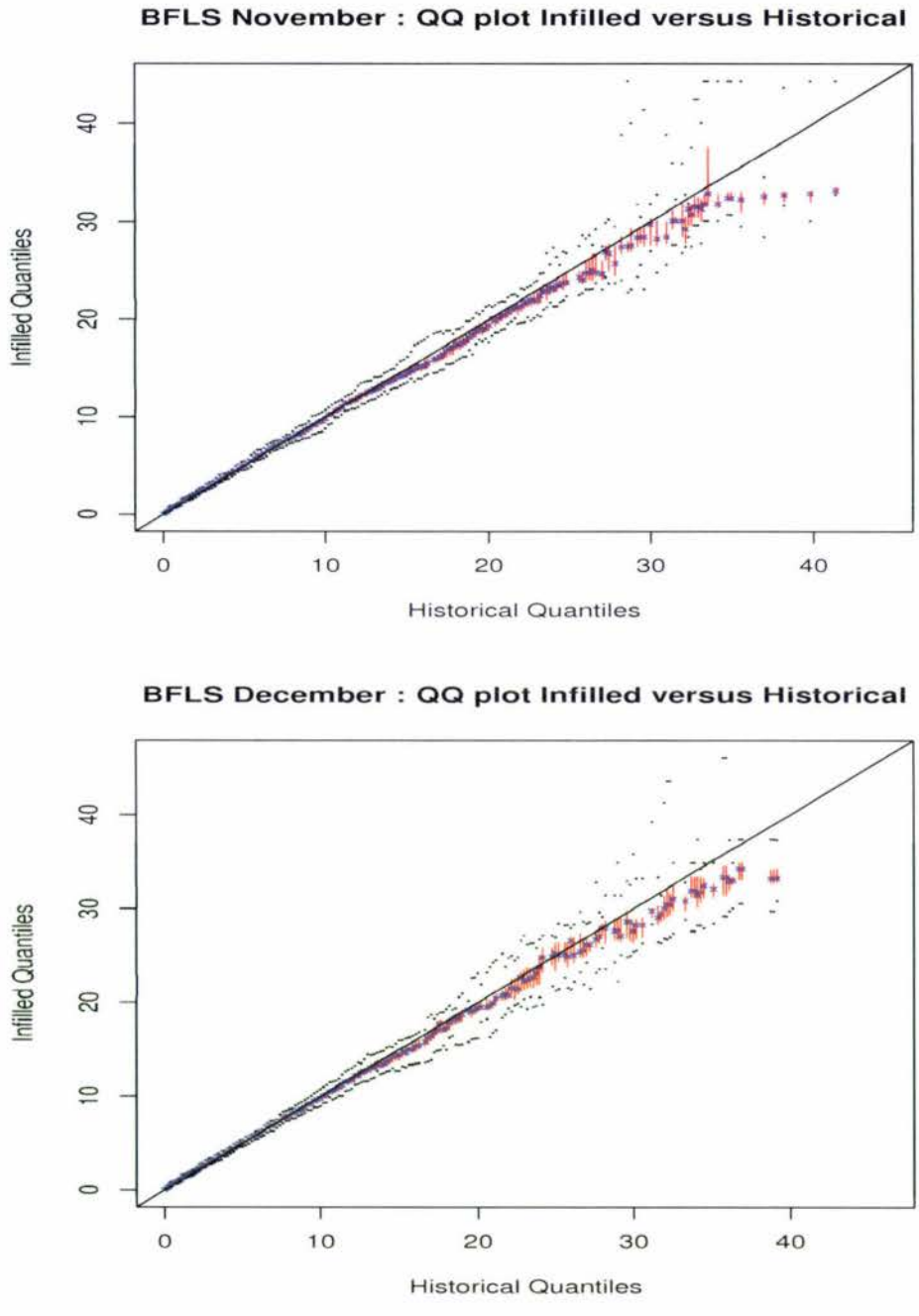
cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.10: BFLS Regional QQ plots: July and August



cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.11: BFLS Regional QQ plots: September and October



cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.12: BFLS Regional QQ plots: November and December

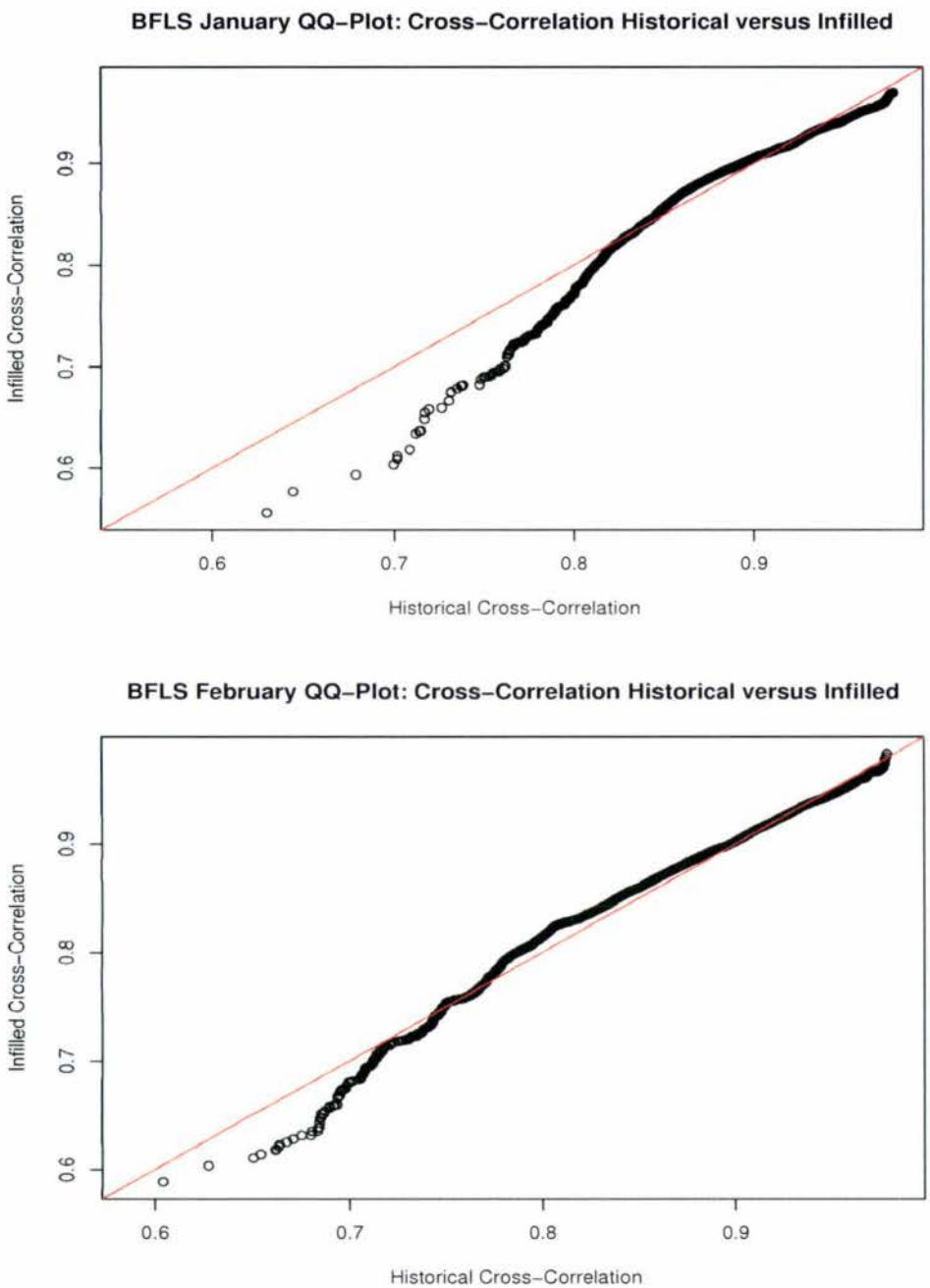


Figure D.13: BFLS cross-correlation QQ plots: January and February

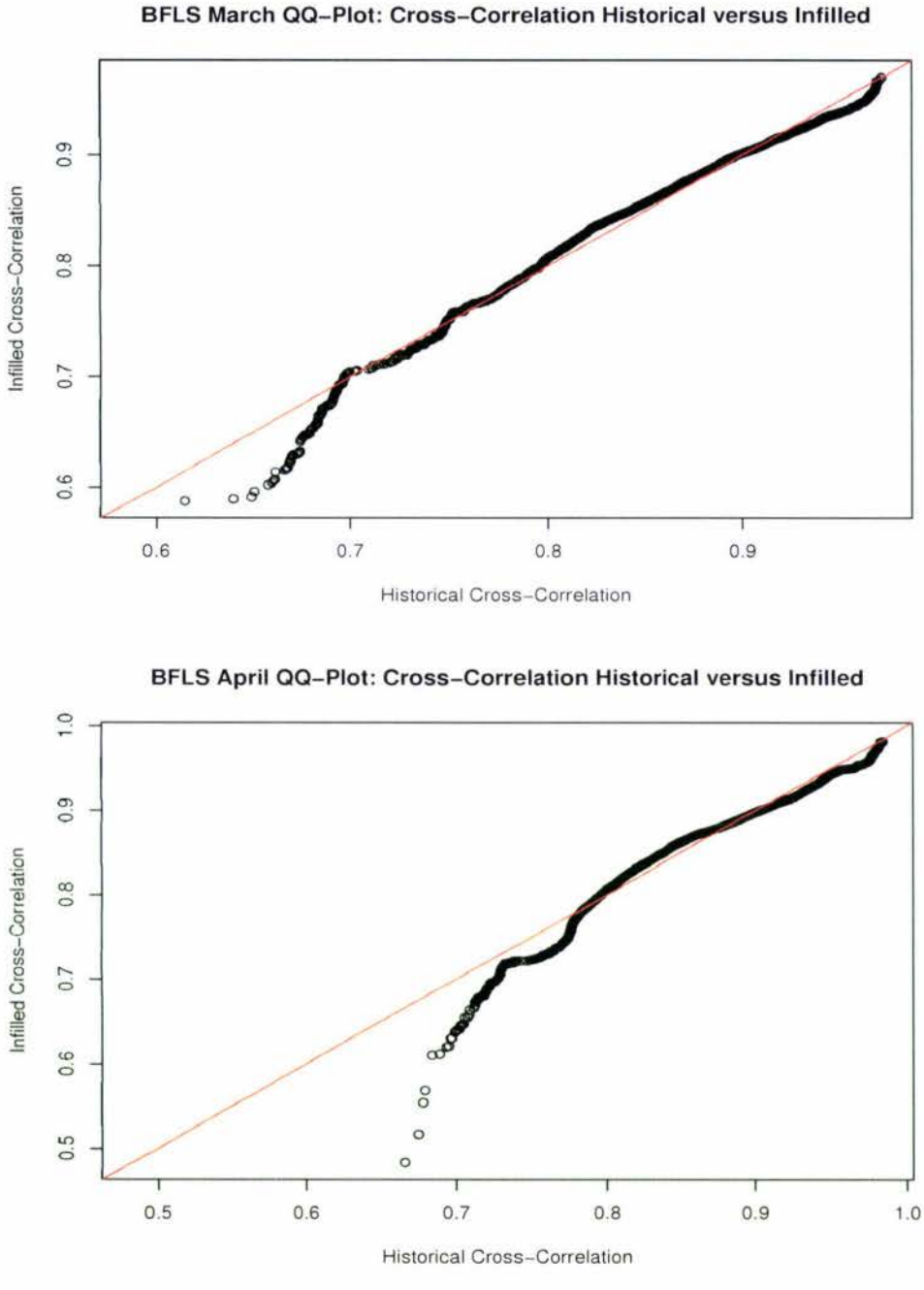


Figure D.14: BFLS cross-correlation QQ plots: March and April

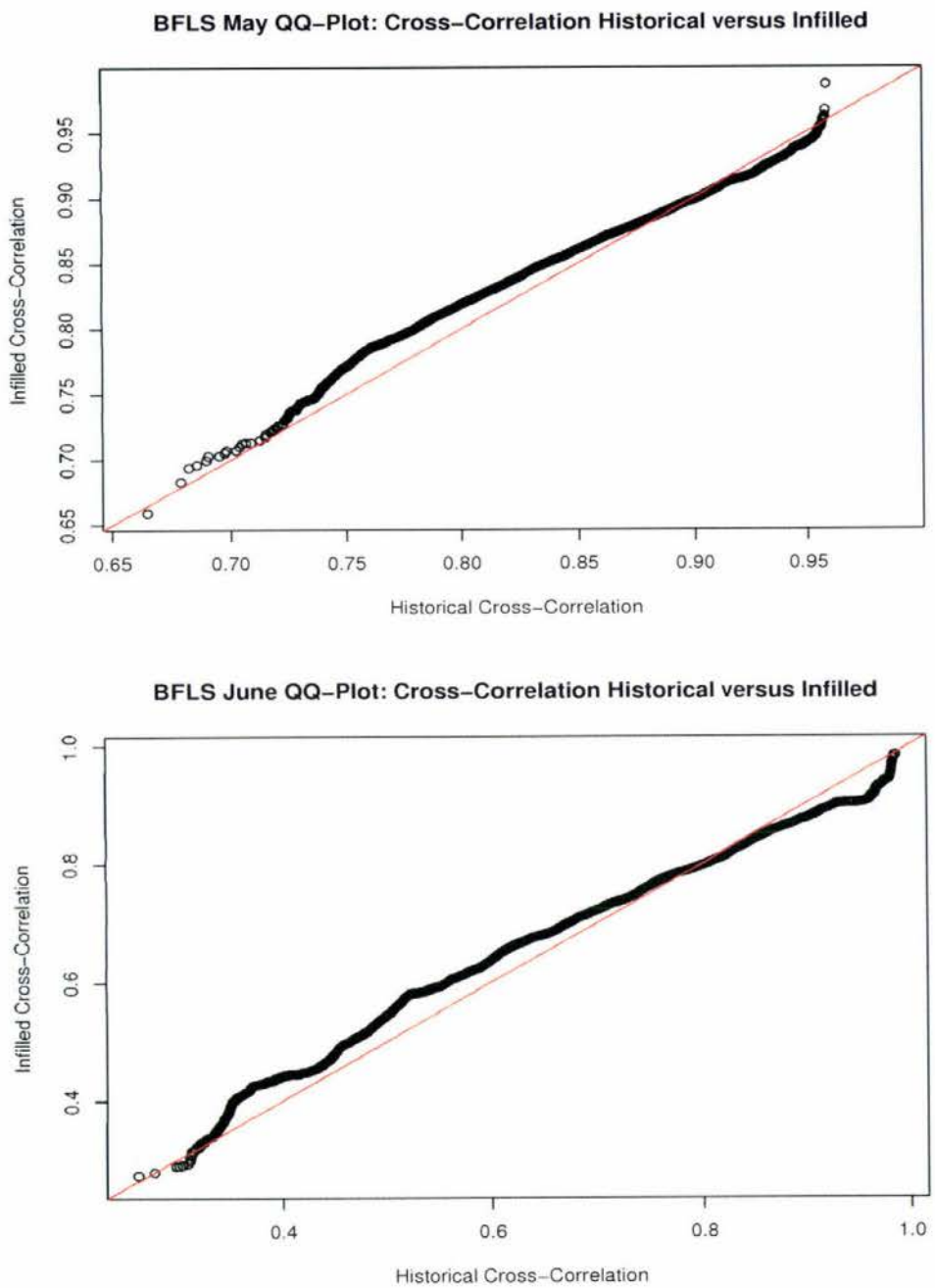


Figure D.15: BFLS cross-correlation QQ plots: May and June

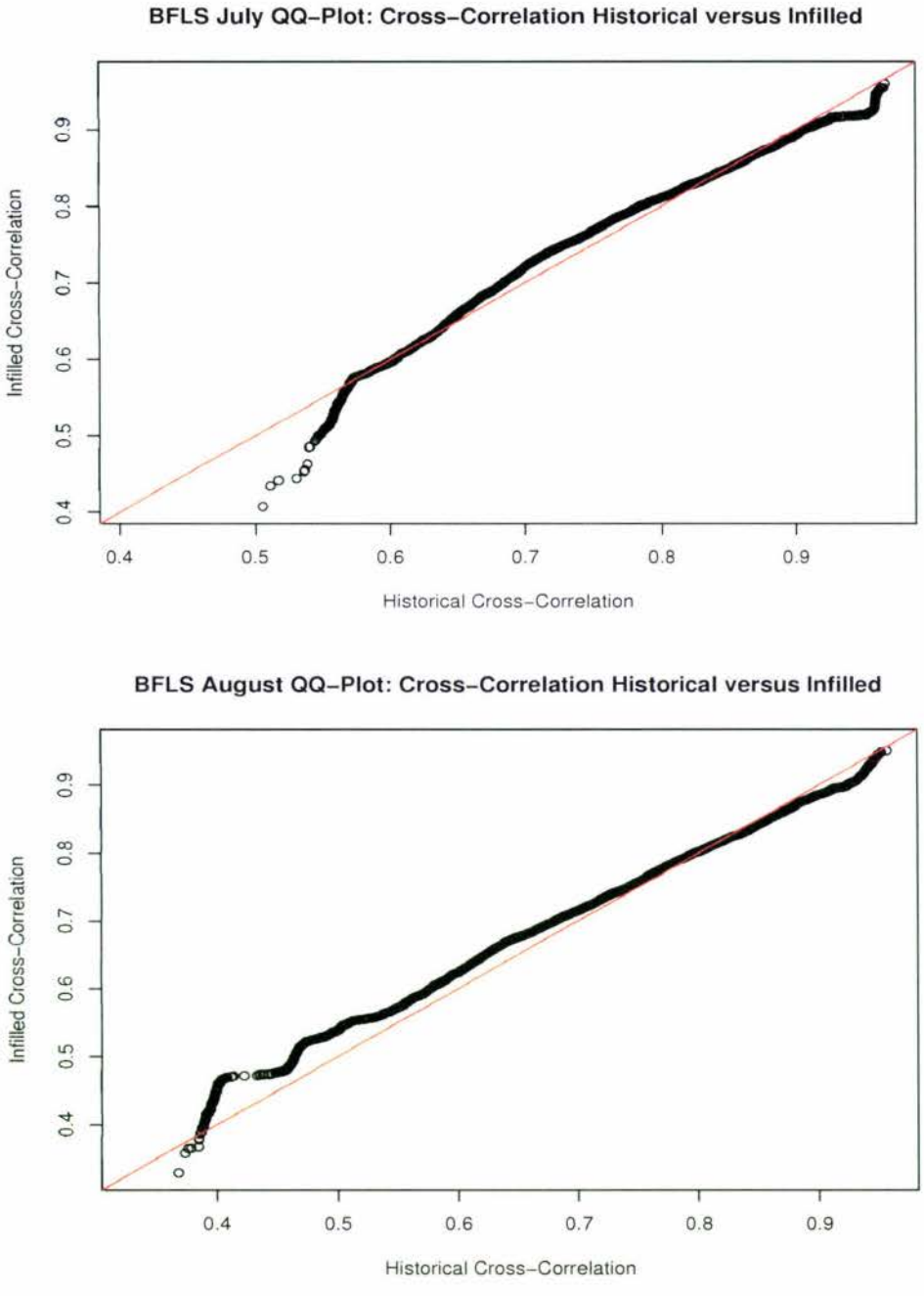


Figure D.16: BFLS cross-correlation QQ plots: July and August

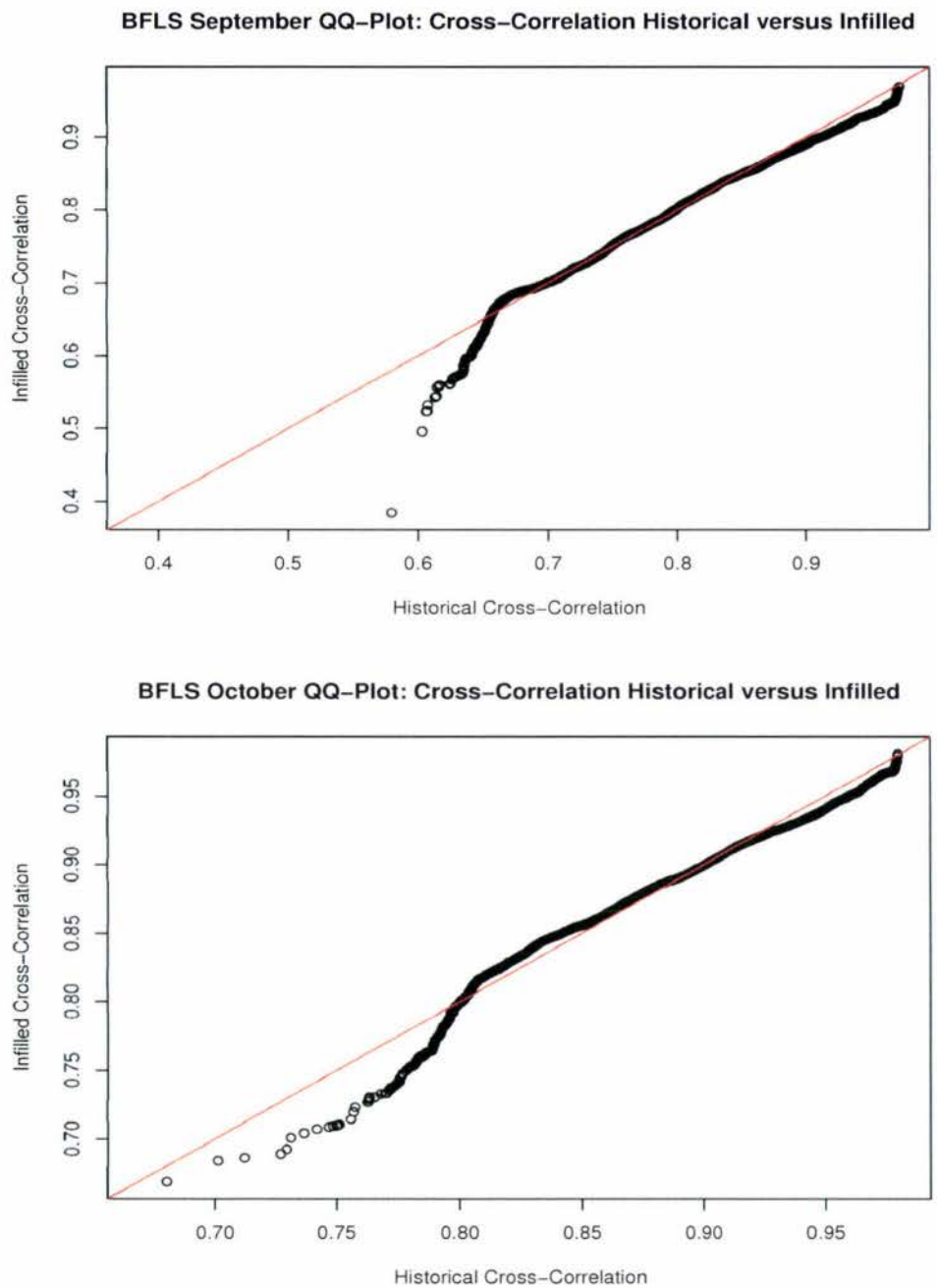


Figure D.17: BFLS cross-correlation QQ plots: September and October

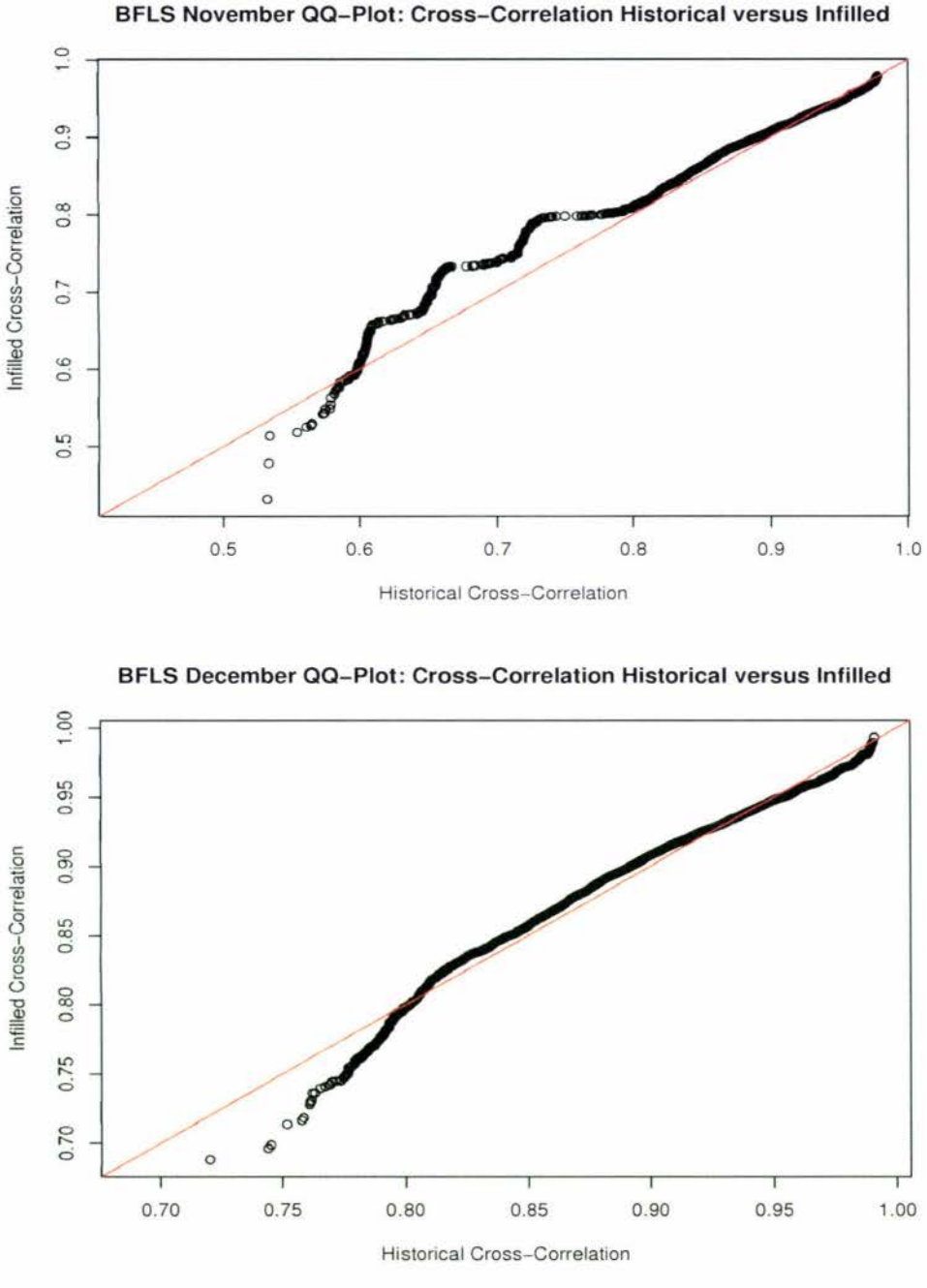


Figure D.18: BFLS cross-correlation QQ plots: November and December

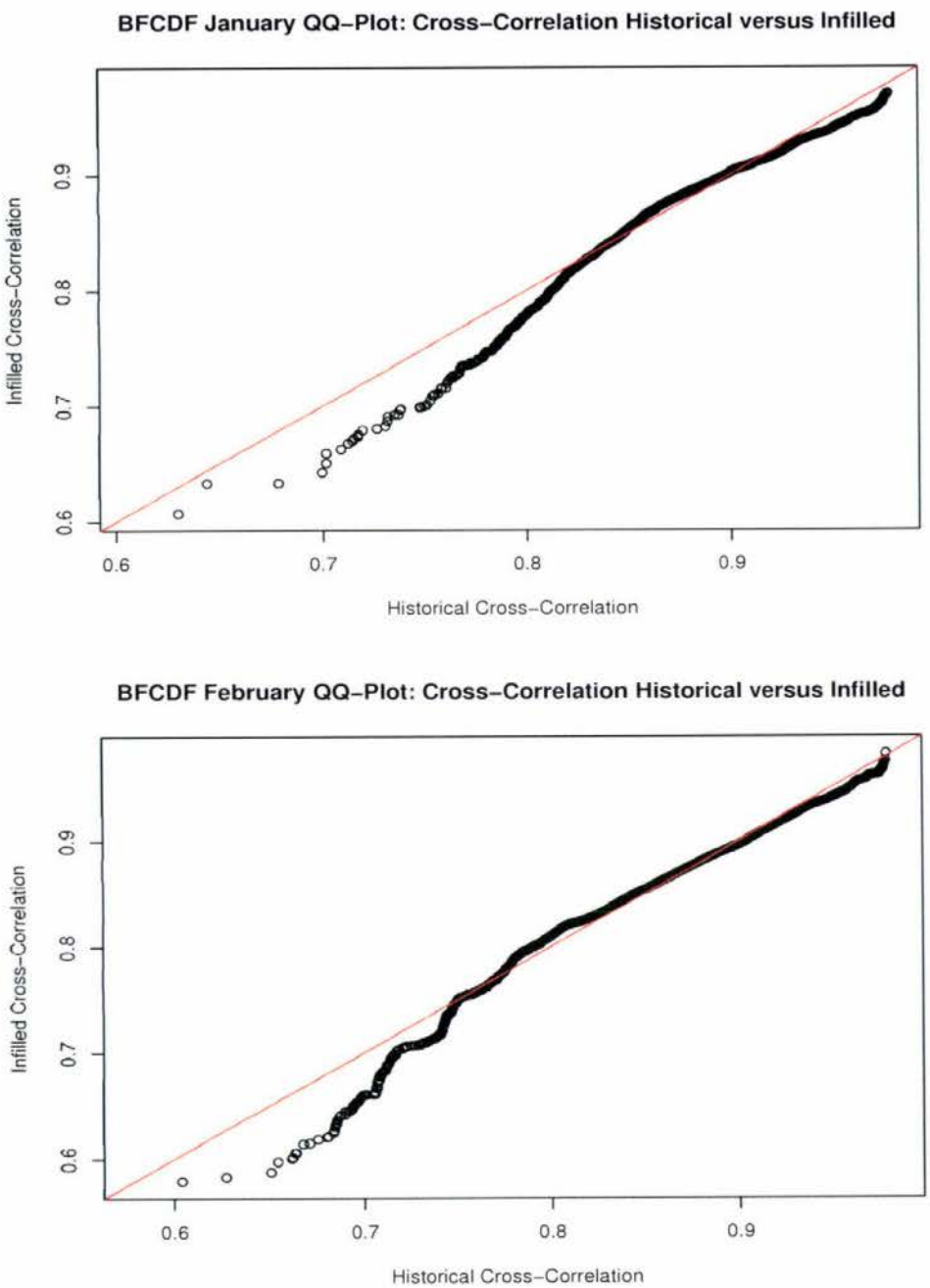


Figure D.19: BFCDF cross-correlation QQ plots: January and February

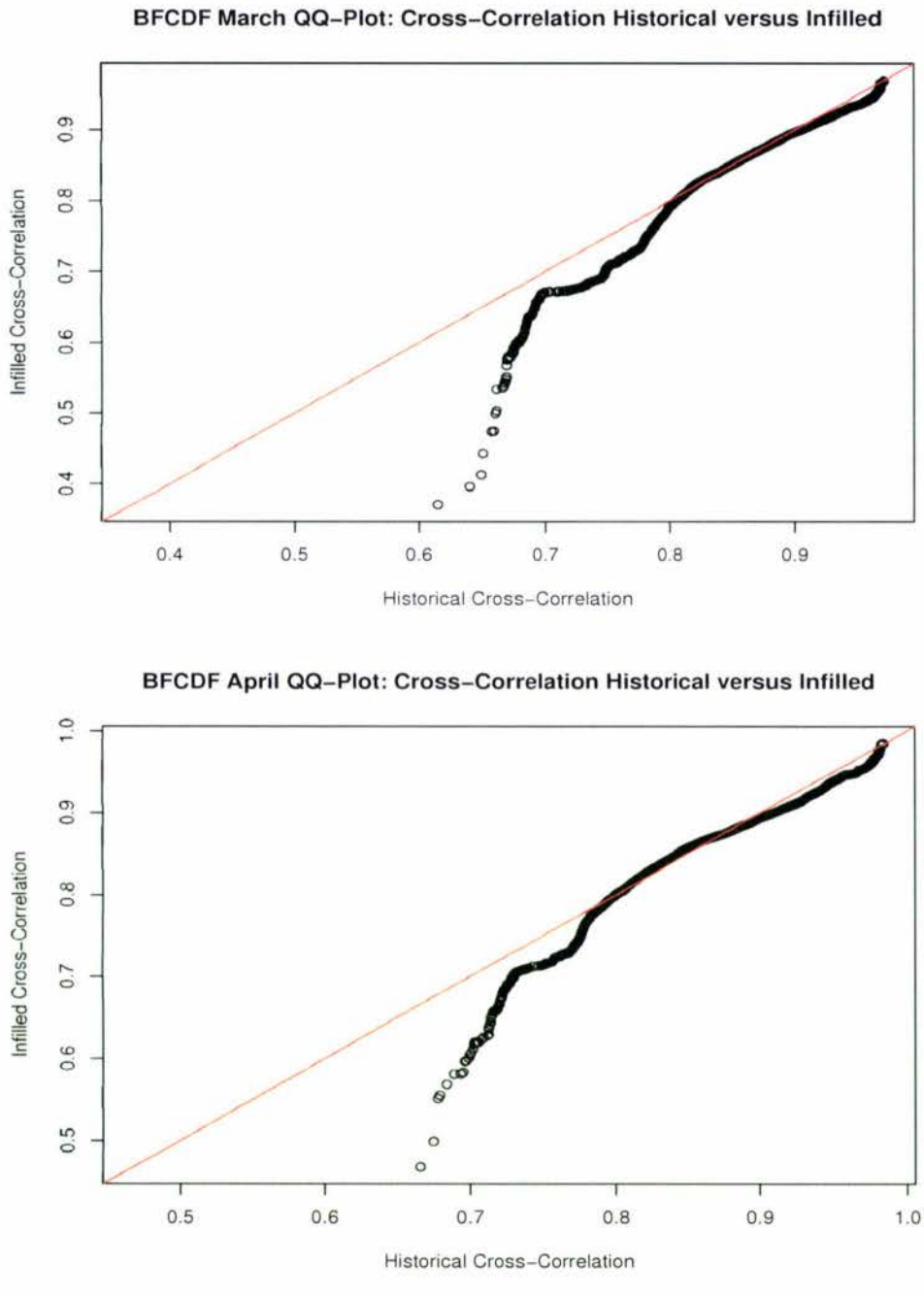


Figure D.20: BFCDF cross-correlation QQ plots: March and April

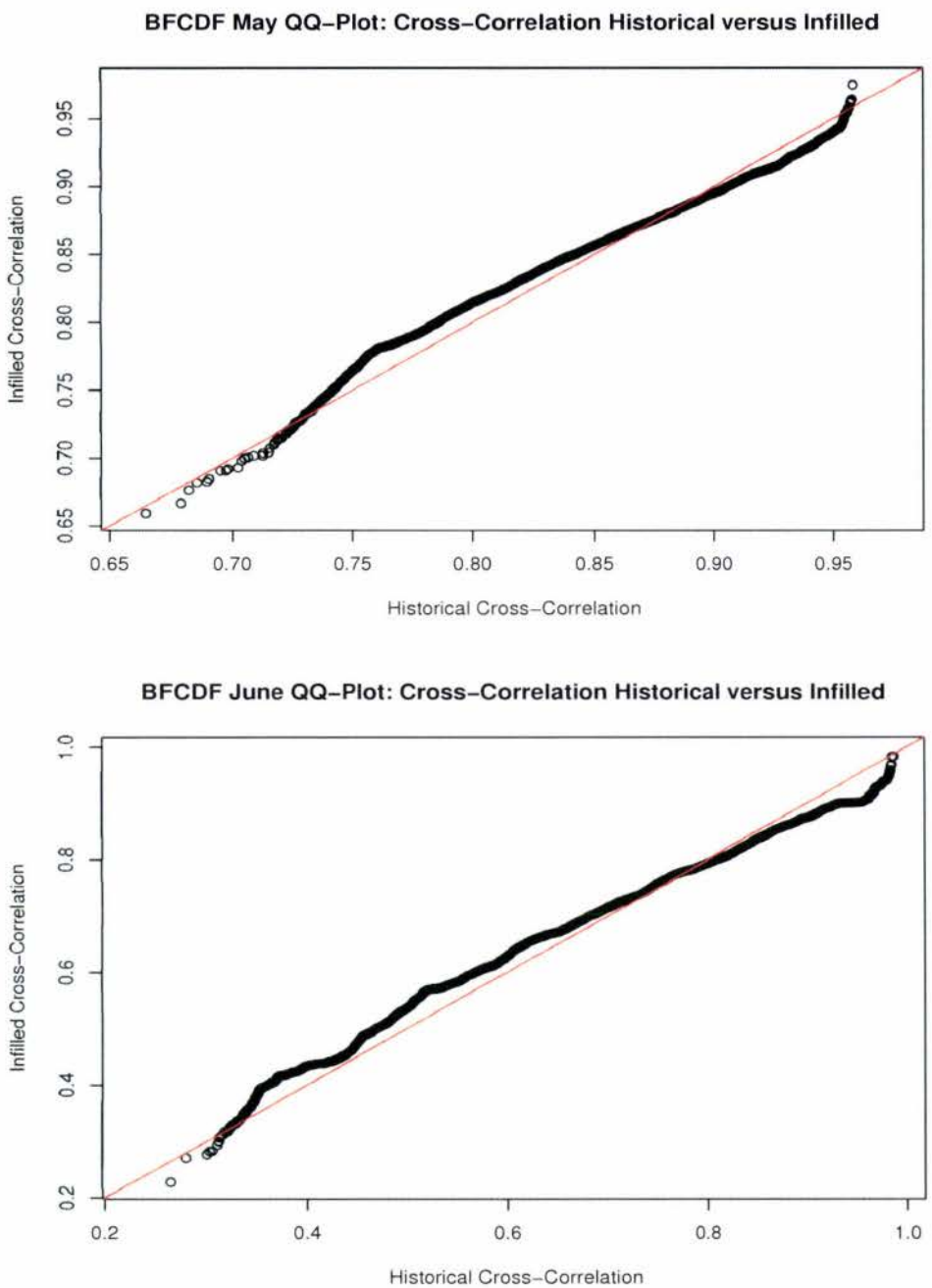


Figure D.21: BFCDF cross-correlation QQ plots: May and June

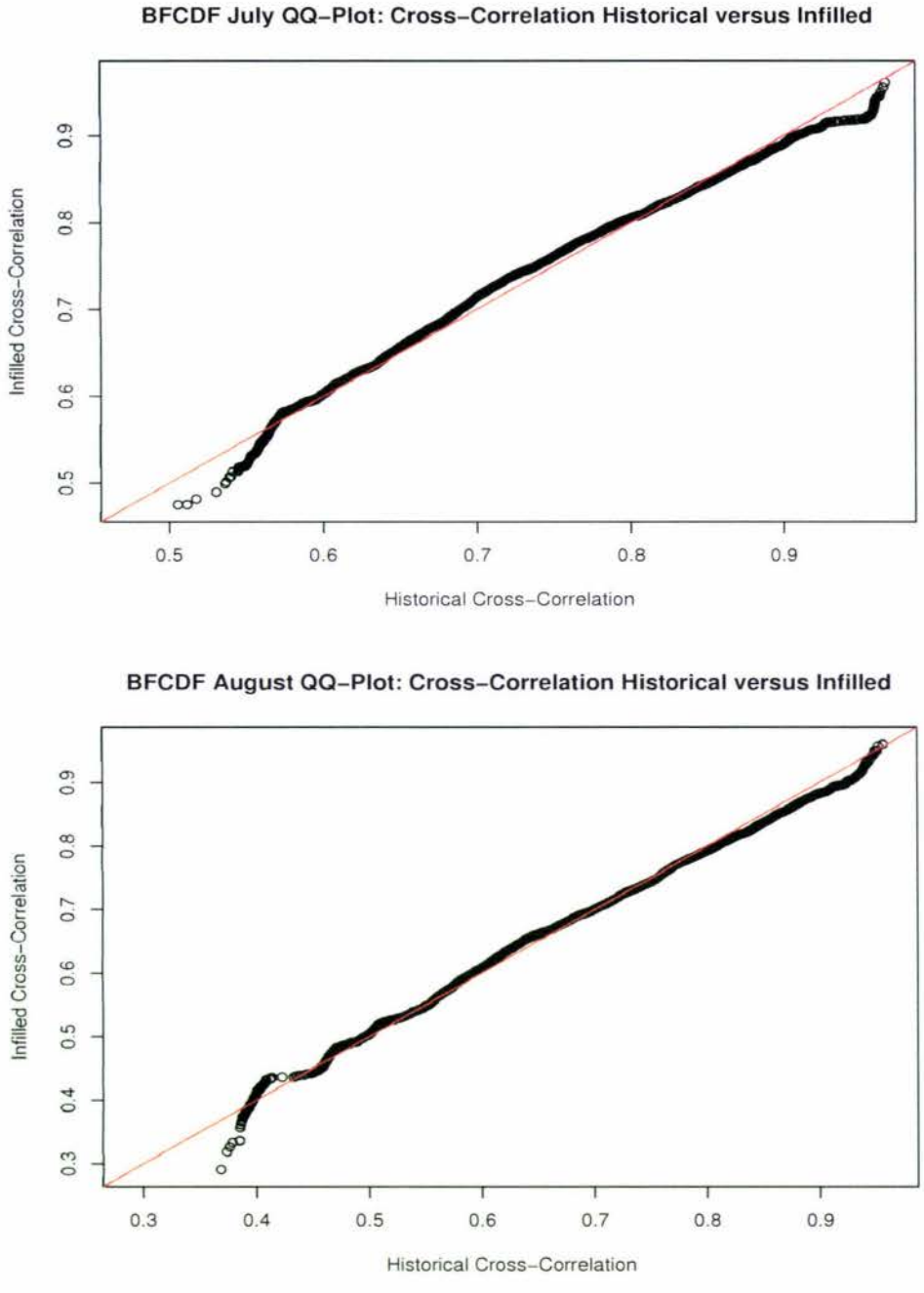


Figure D.22: BFCDF cross-correlation QQ plots: July and August

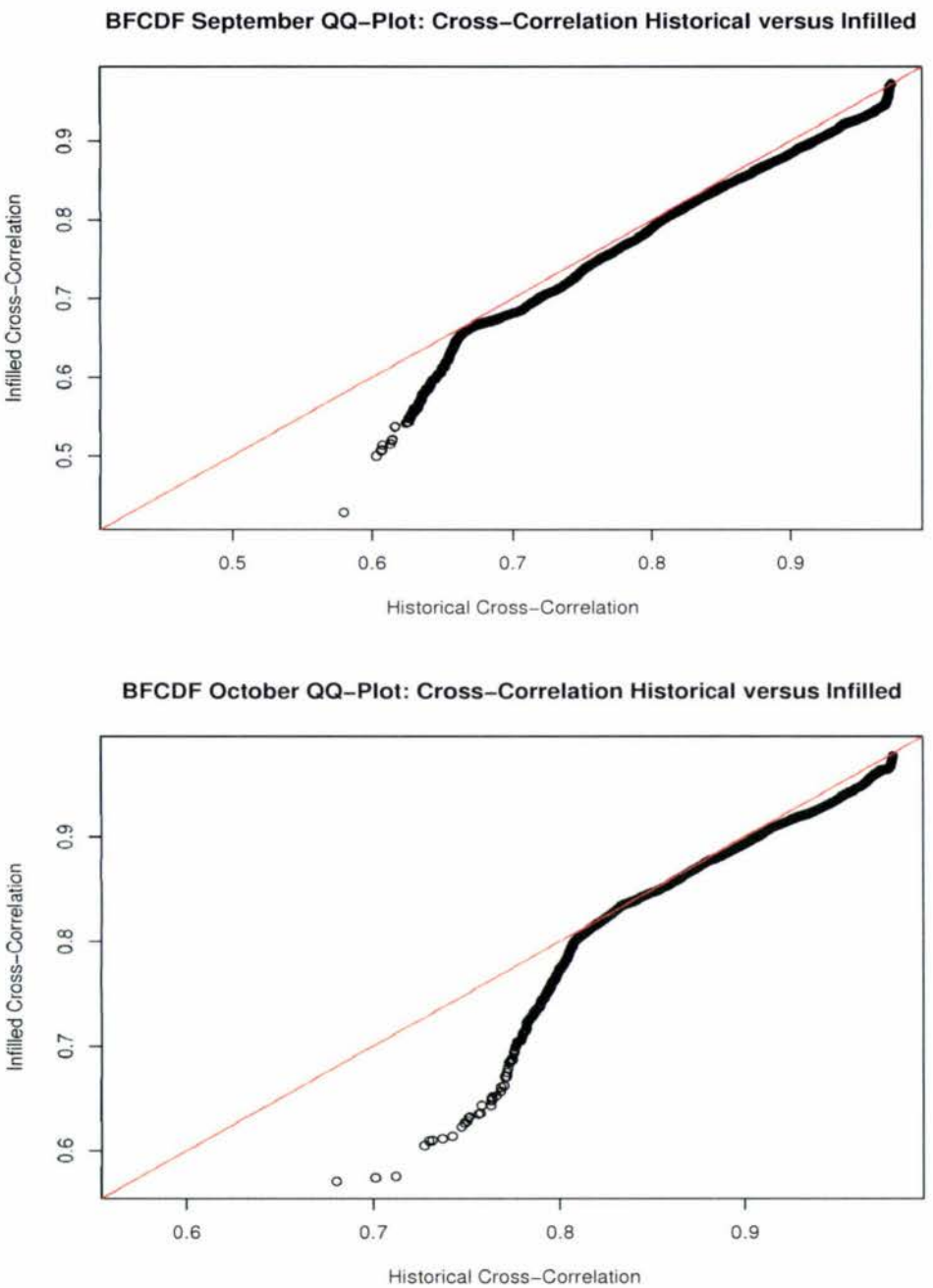


Figure D.23: BFCDF cross-correlation QQ plots: September and October

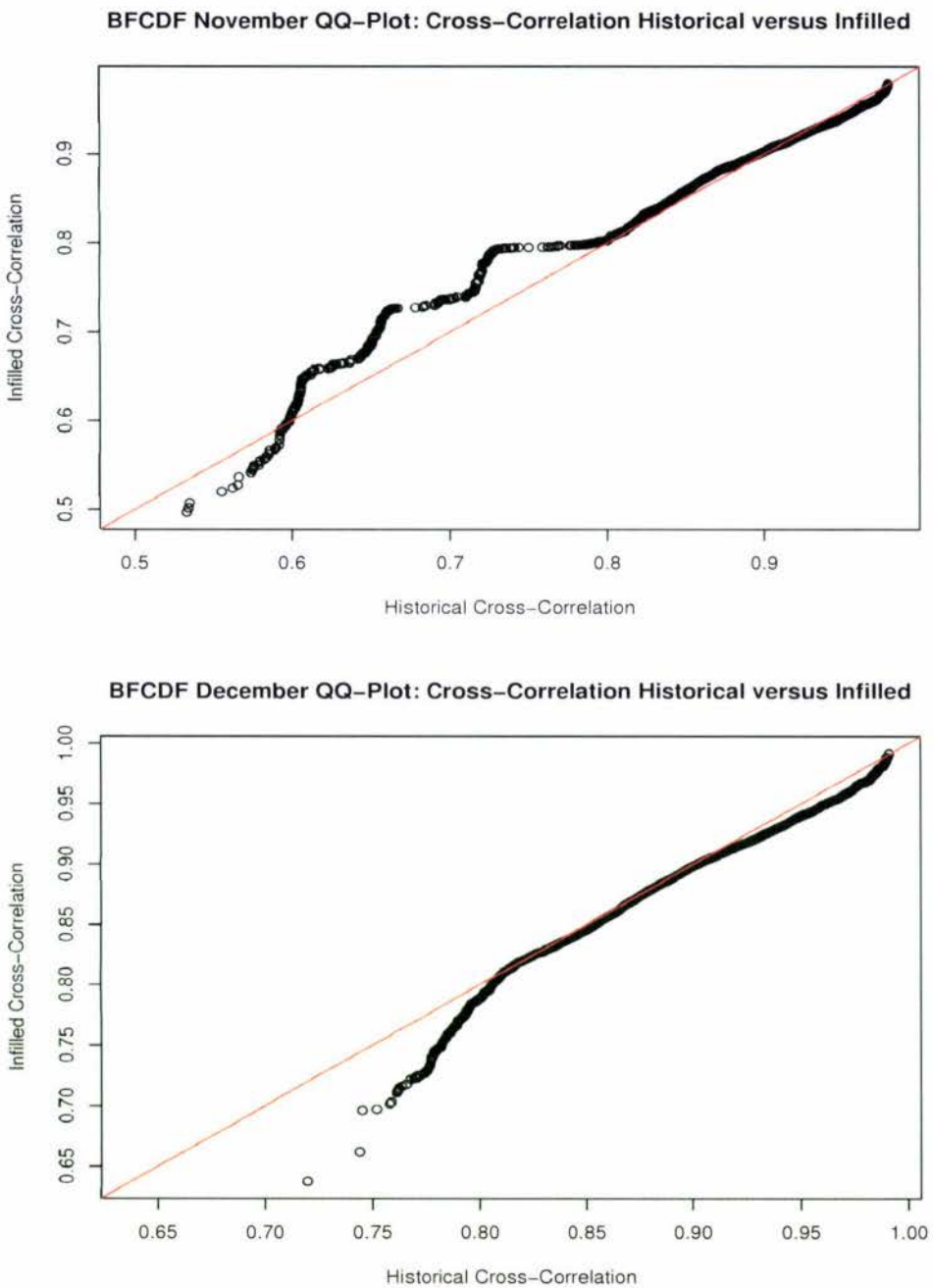


Figure D.24: BFCDF cross-correlation QQ plots: November and December

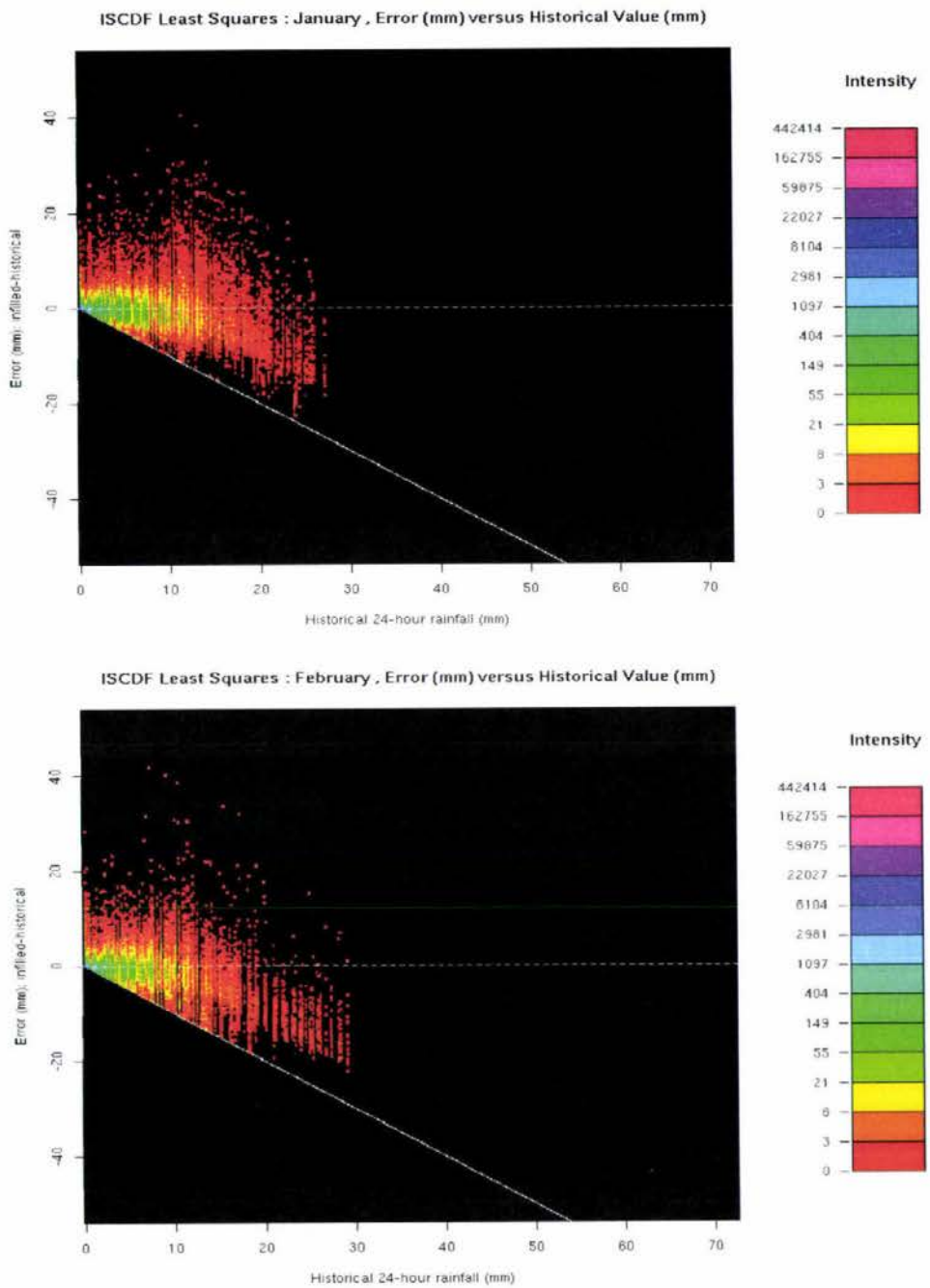


Figure D.25: ISCDF Intensity plots: January and February

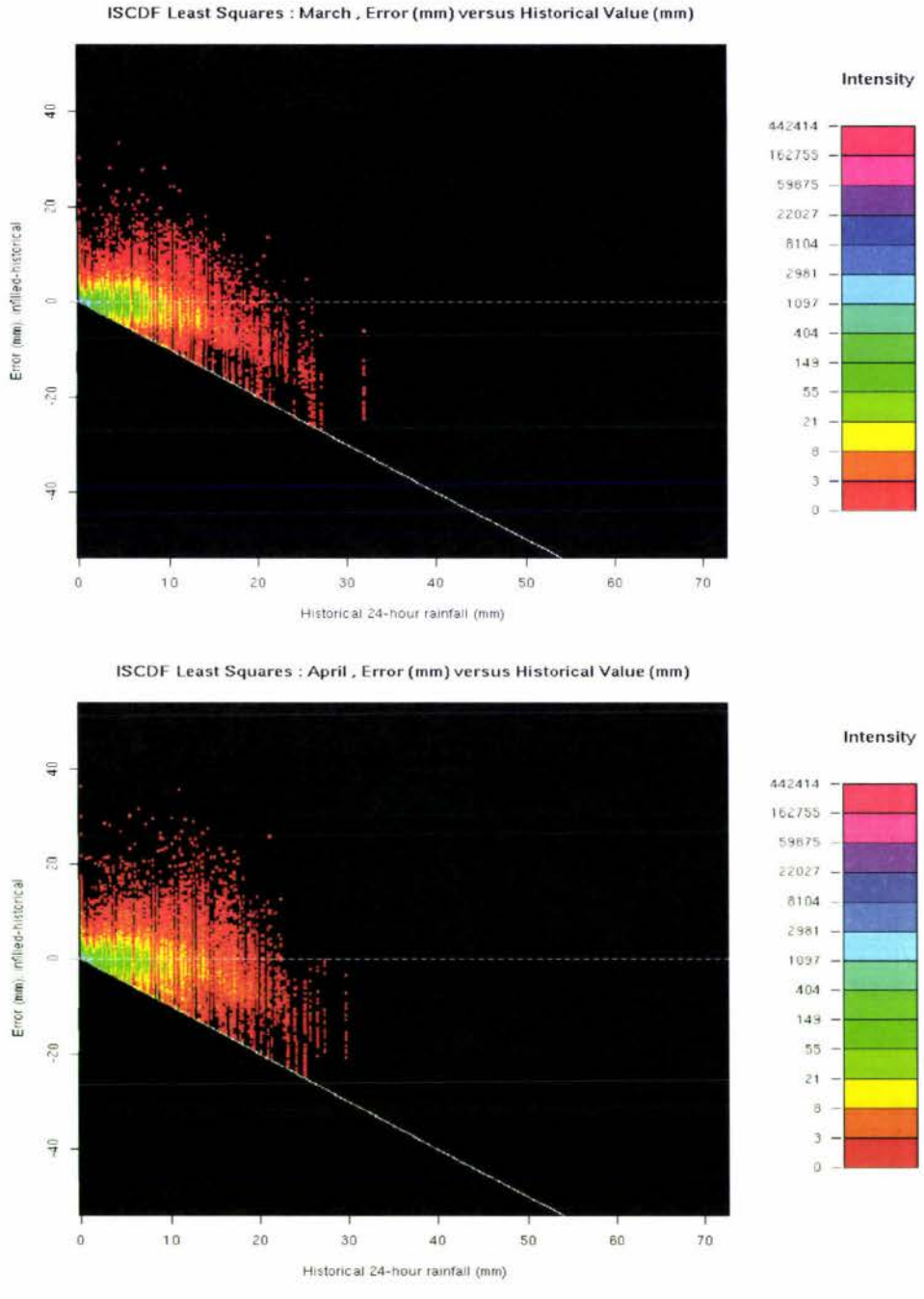


Figure D.26: ISCDF Intensity plots: March and April

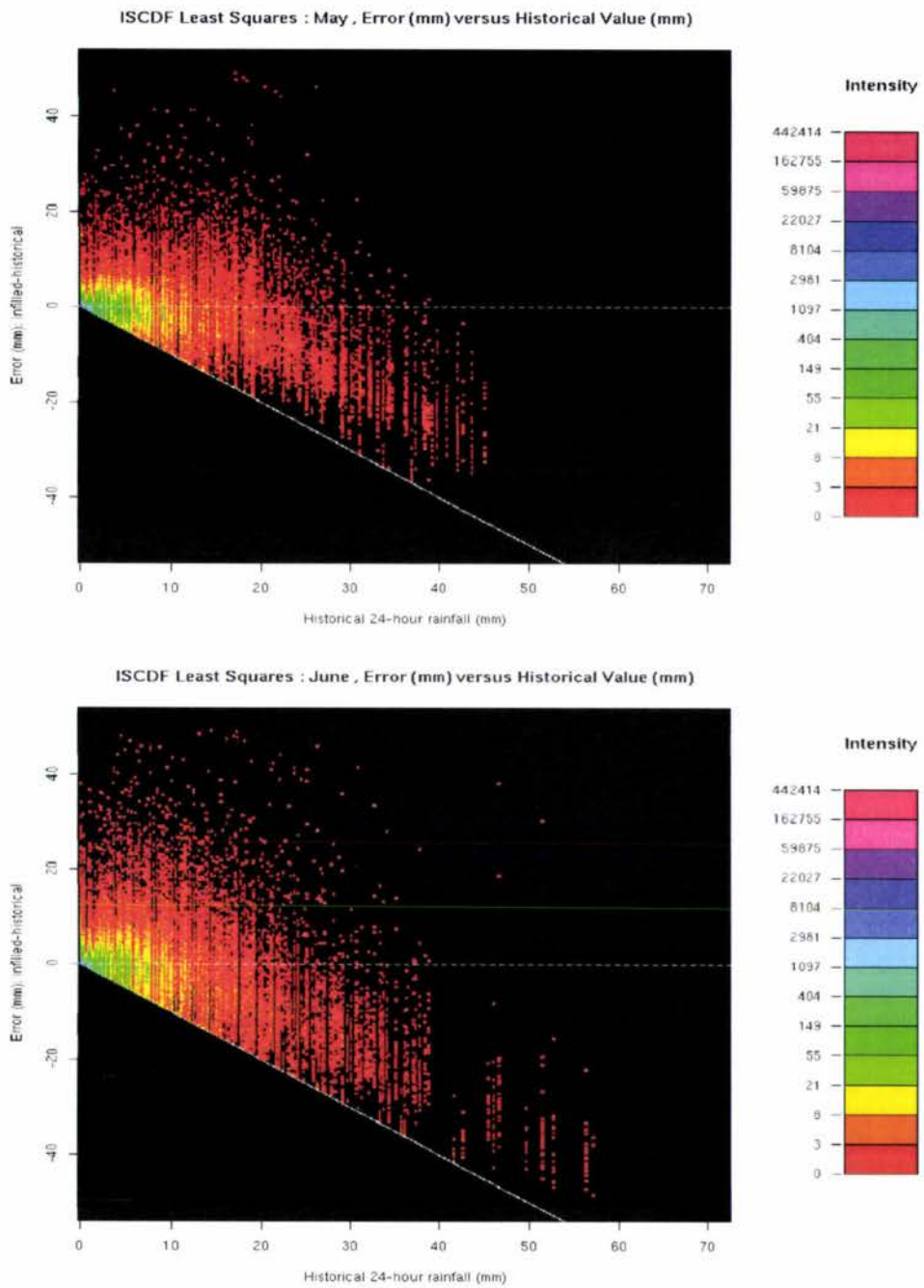


Figure D.27: ISCDF Intensity plots: May and June

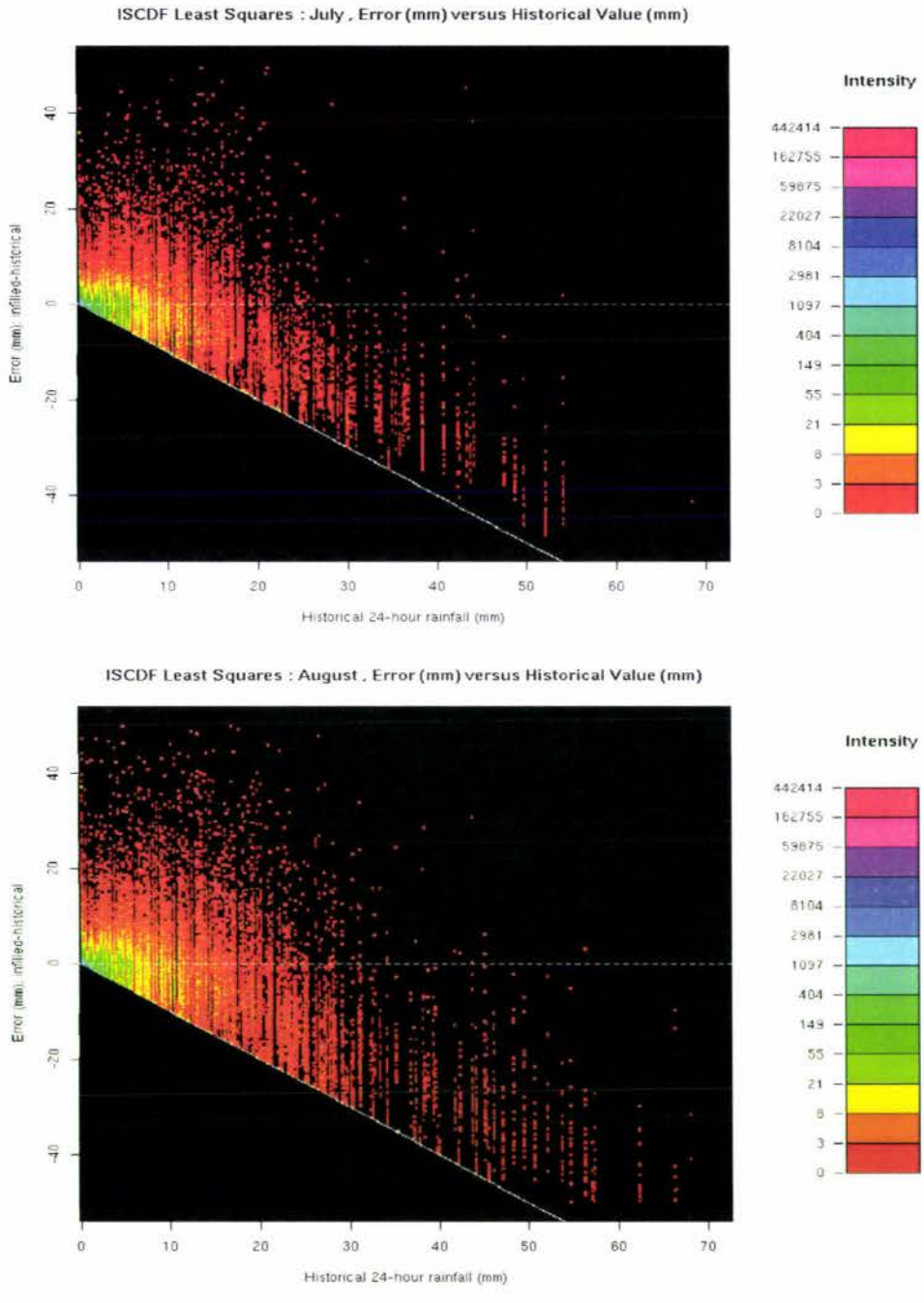


Figure D.28: ISCDF Intensity plots: July and August

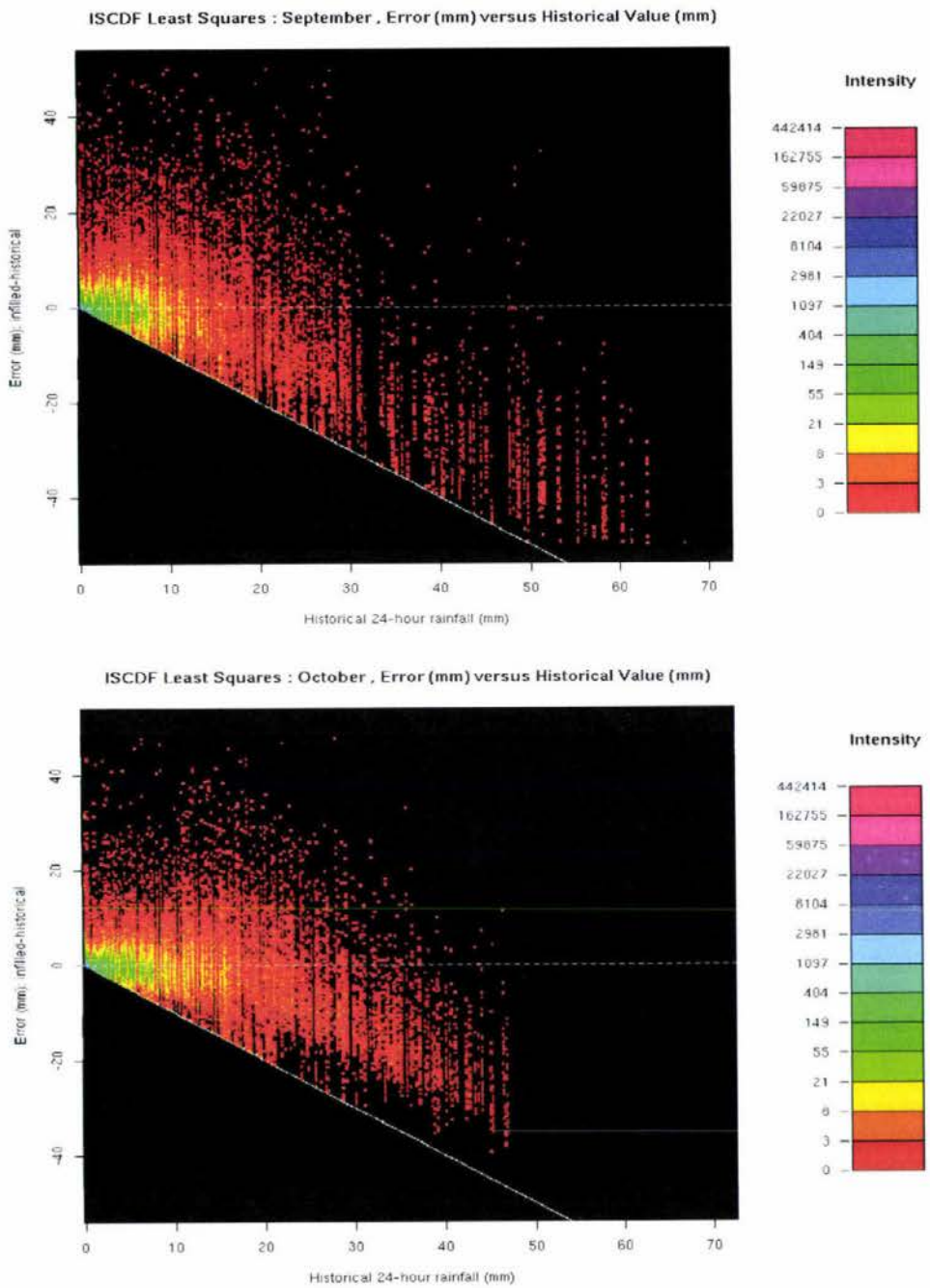


Figure D.29: ISCDF Intensity plots: September and October

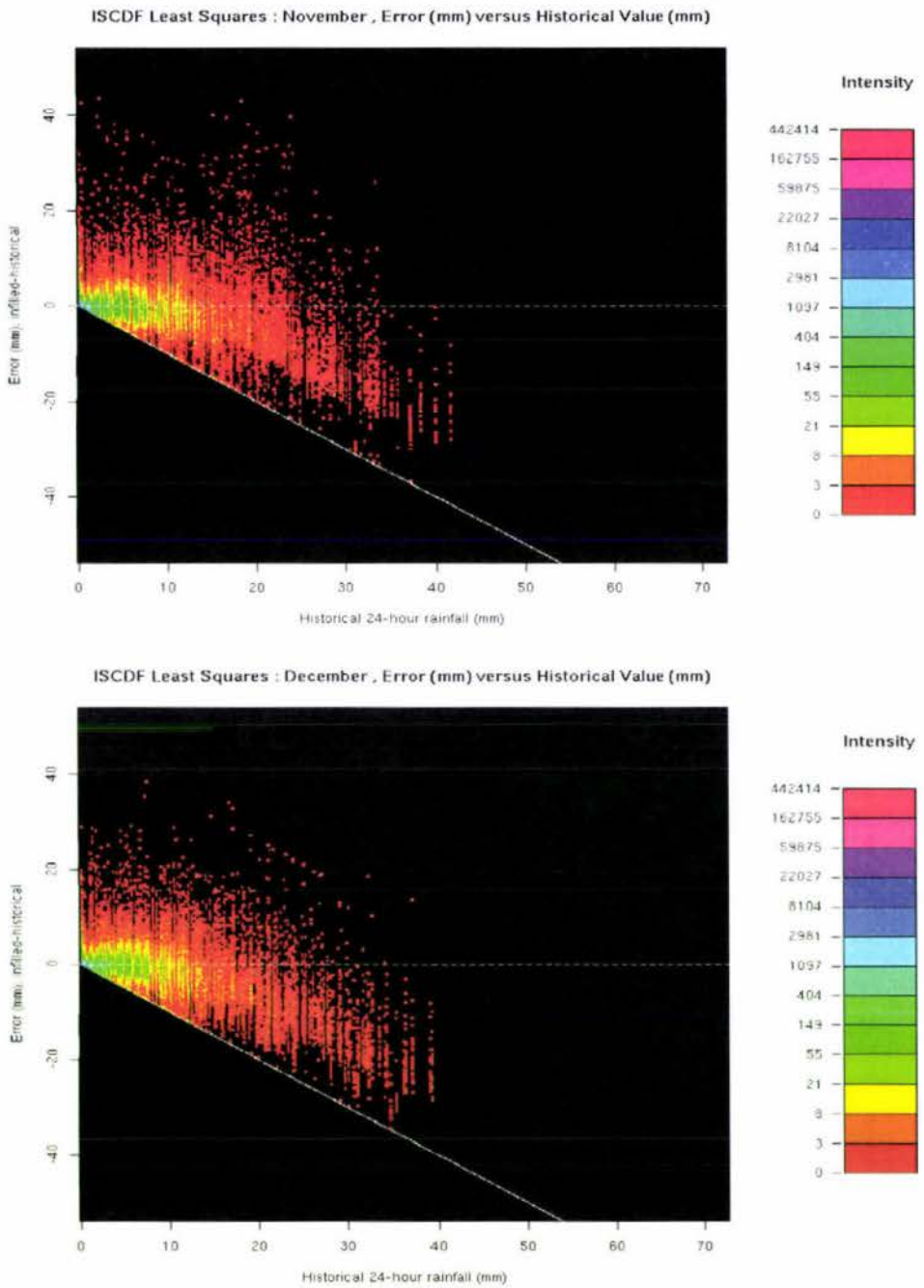
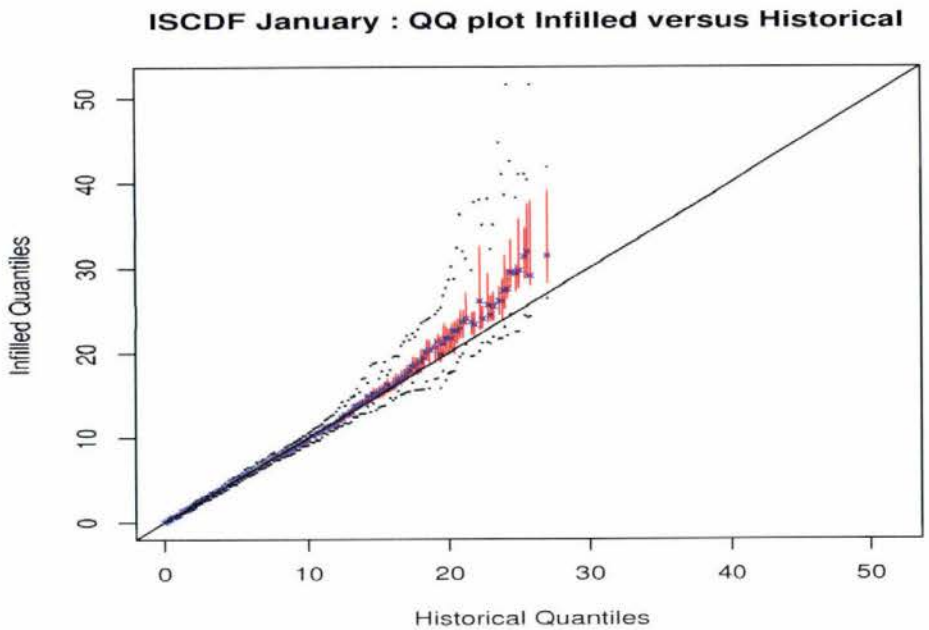
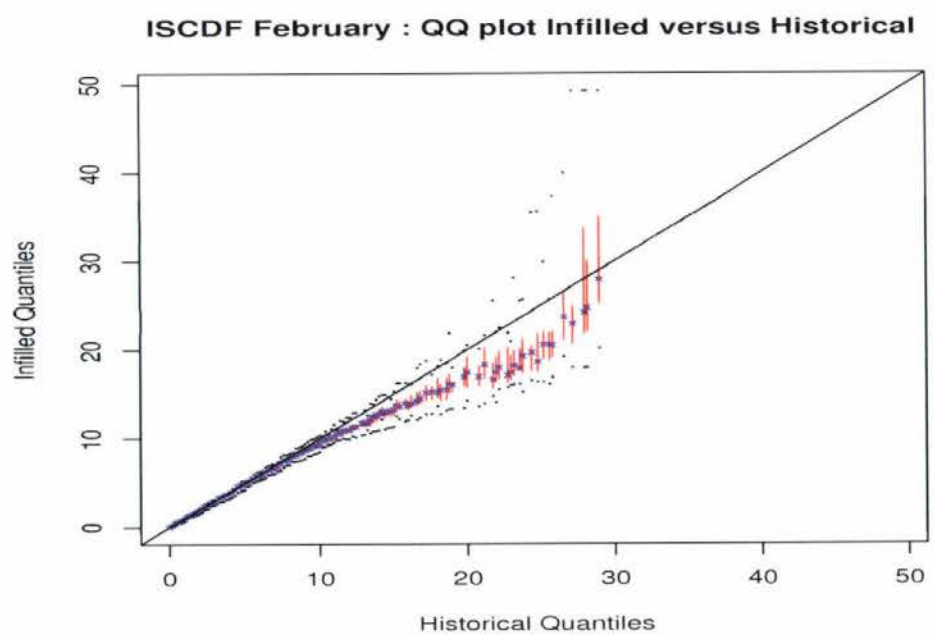


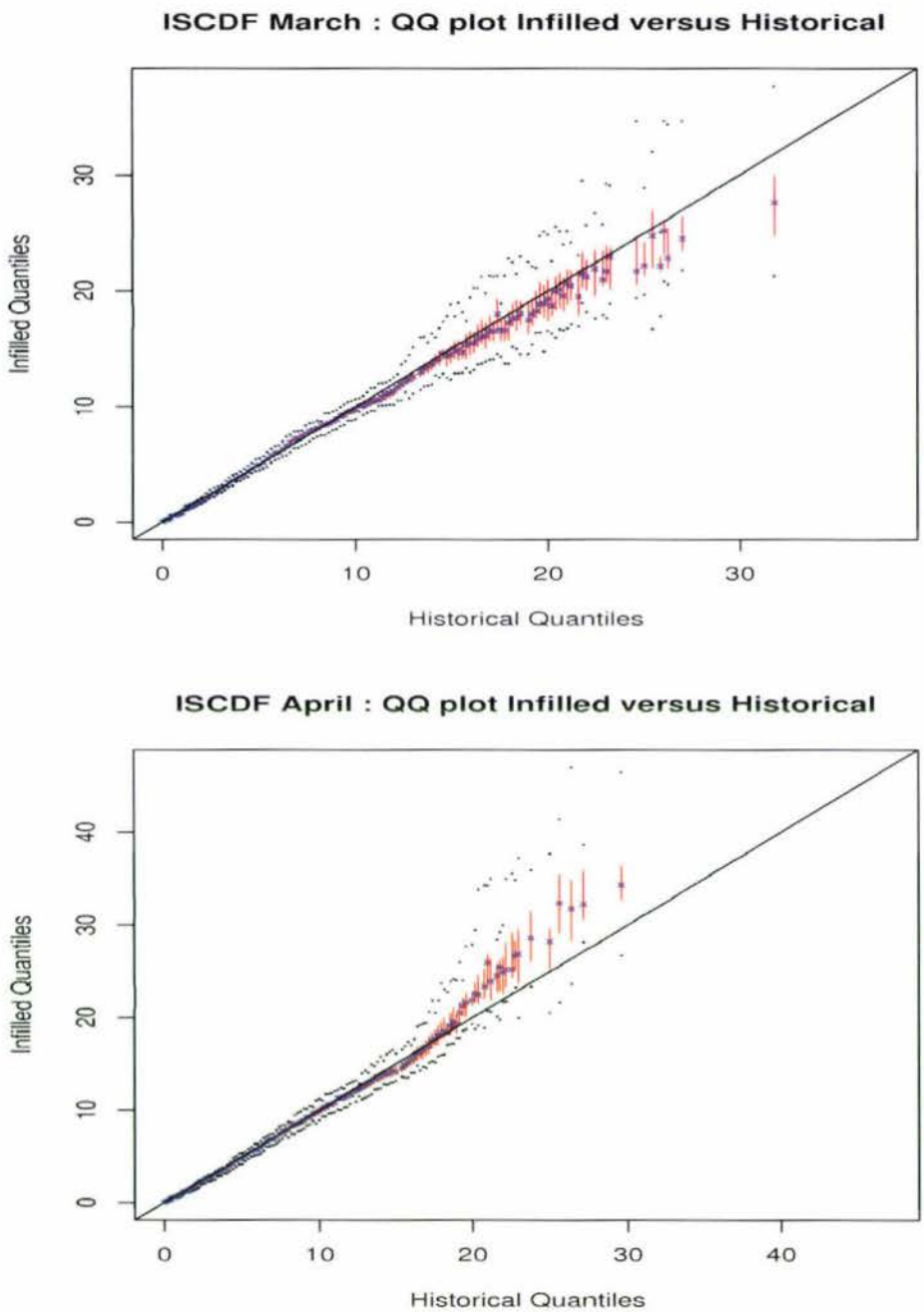
Figure D.30: ISCDF Intensity plots: November and December





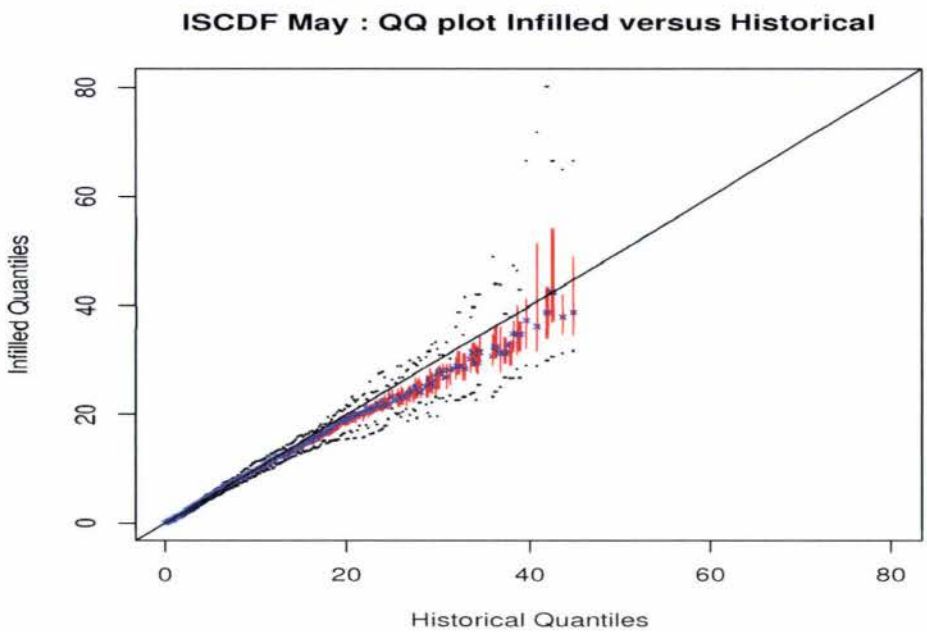
cross=median; vertical line = IQR error bar; dots = minima or maxima

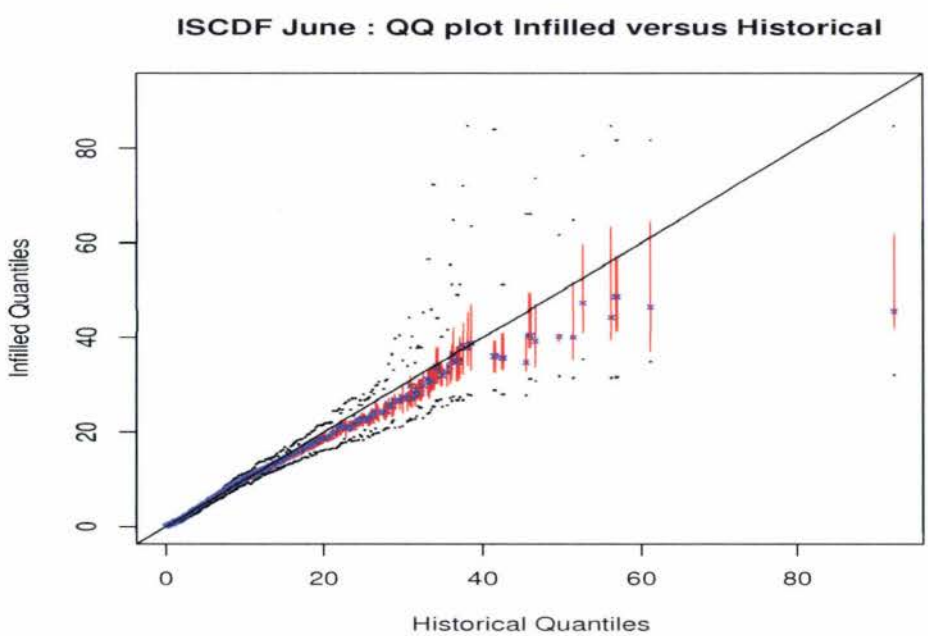
Figure D.31: ISCDF Regional QQ plots: January and February



cross=median; vertical line = IQR error bar; dots = minima or maxima

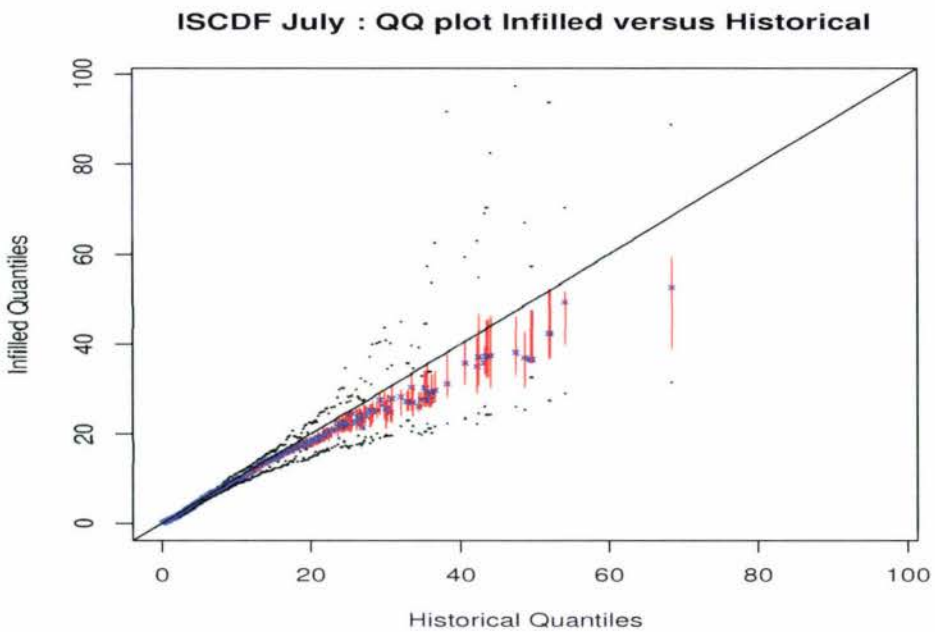
Figure D.32: ISCDF Regional QQ plots: March and April

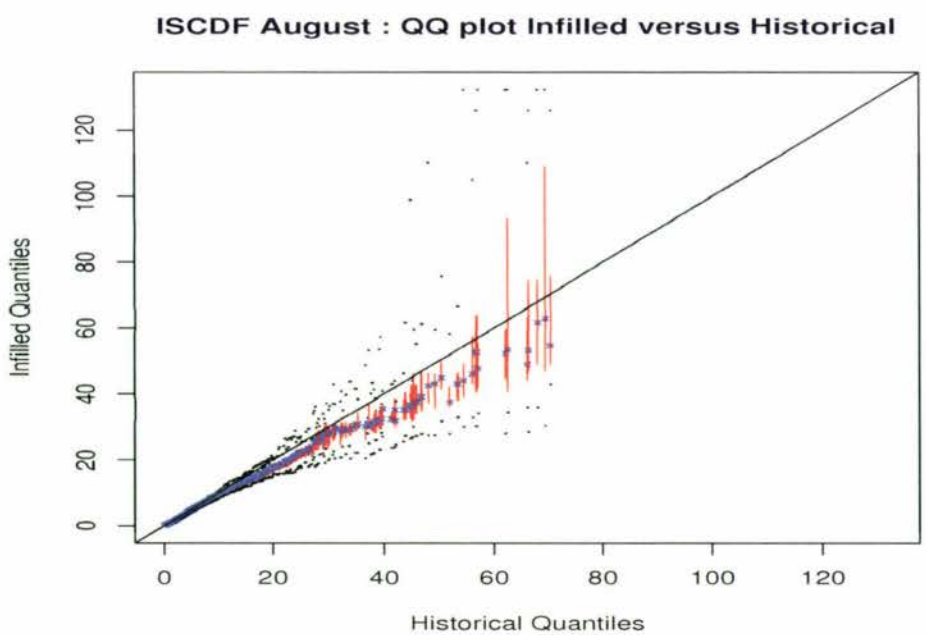




cross=median; vertical line = IQR error bar; dots = minima or maxima

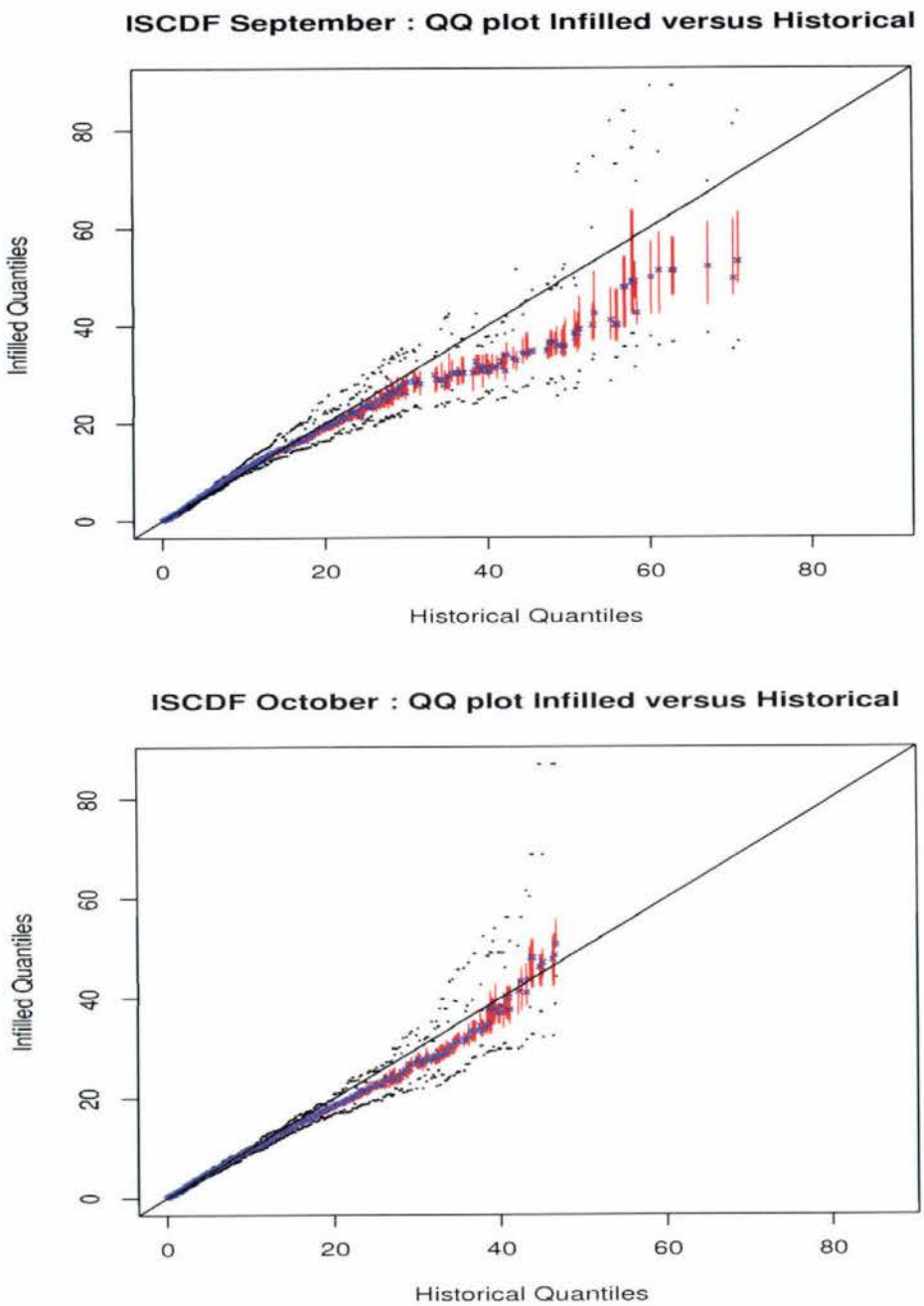
Figure D.33: ISCDF Regional QQ plots: May and June





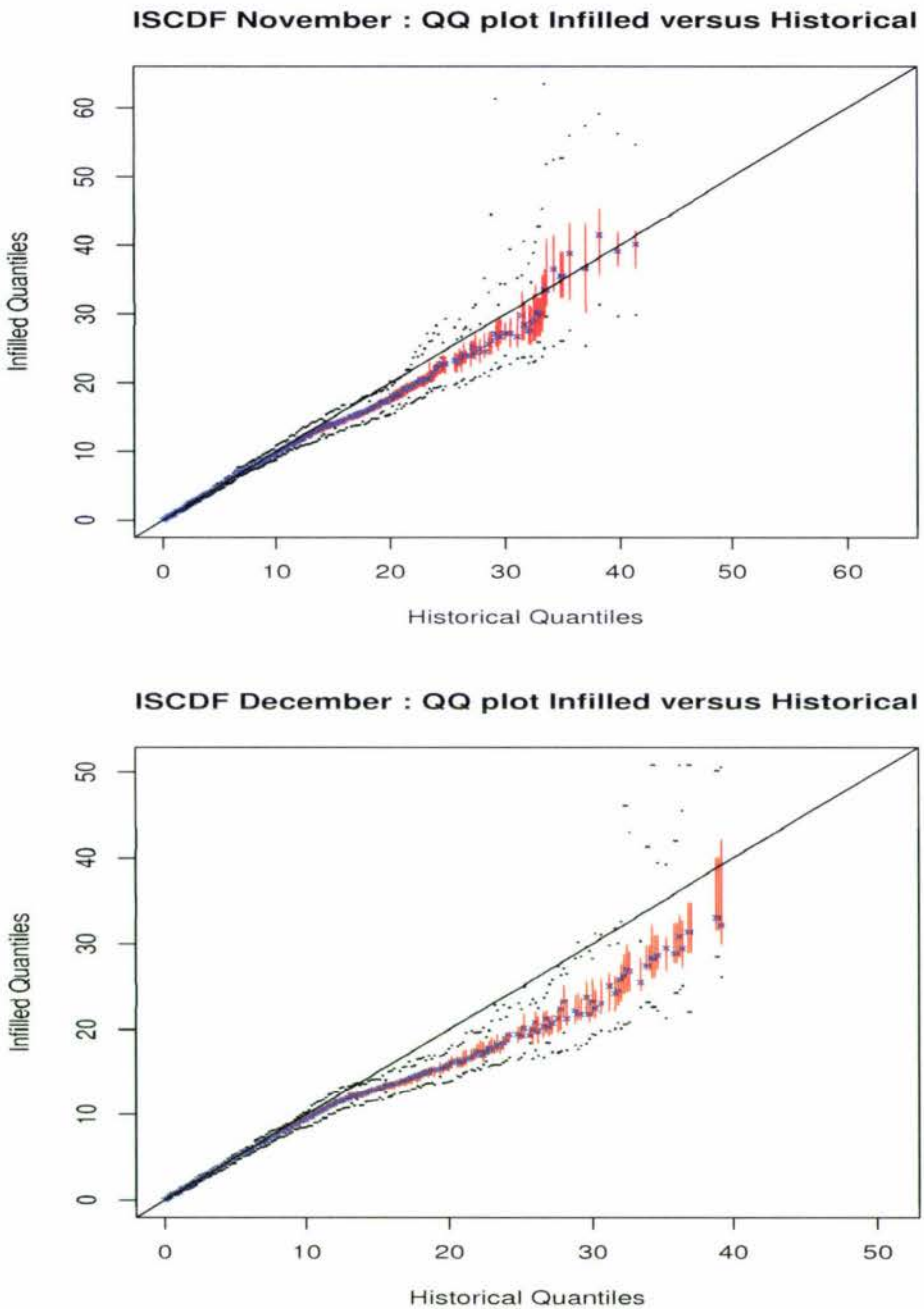
cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.34: ISCDF Regional QQ plots: July and August



cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.35: ISCDF Regional QQ plots: September and October



cross=median; vertical line = IQR error bar; dots = minima or maxima

Figure D.36: ISCDF Regional QQ plots: November and December

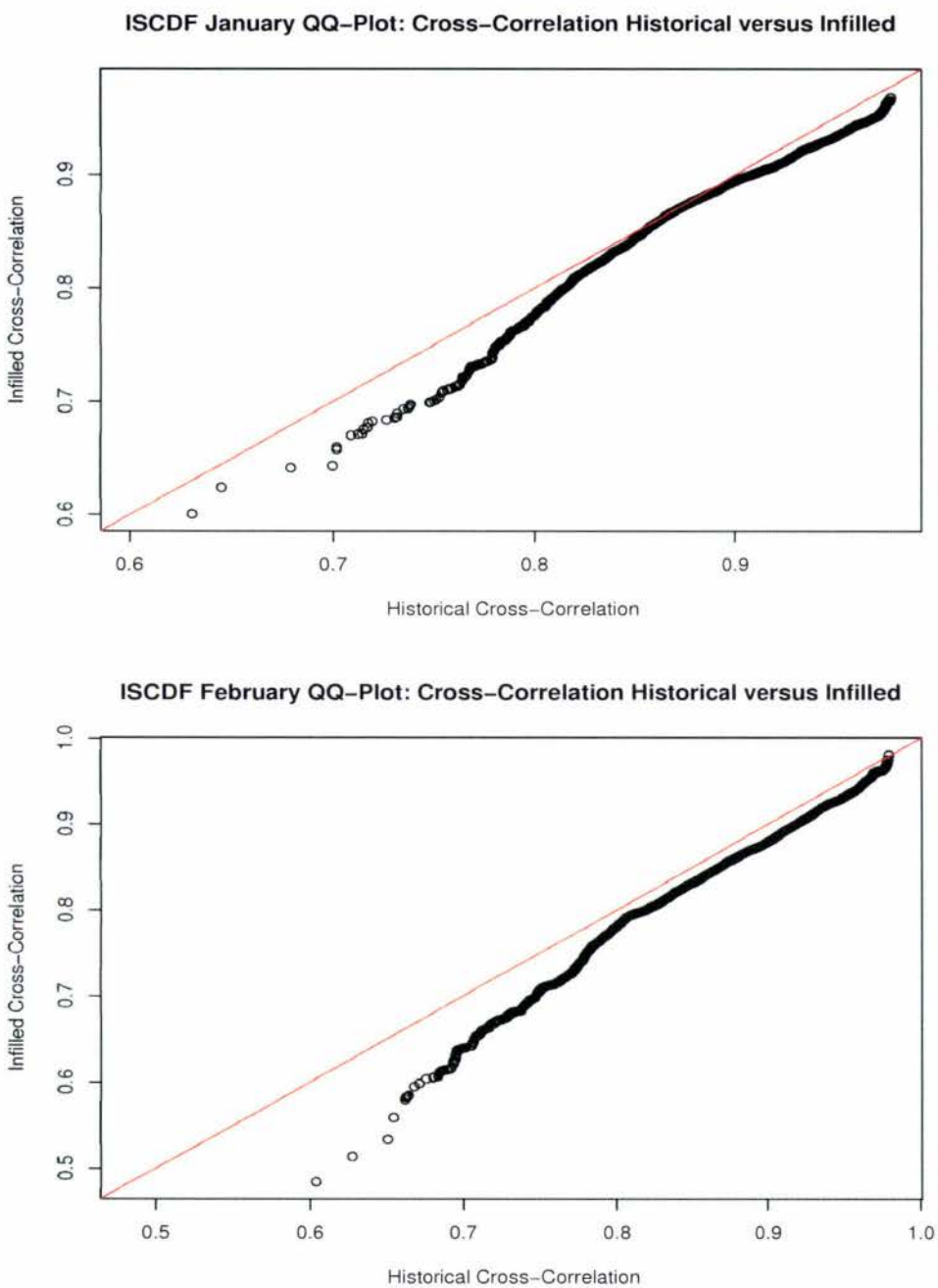


Figure D.37: ISCDF cross-correlation QQ plots: January and February

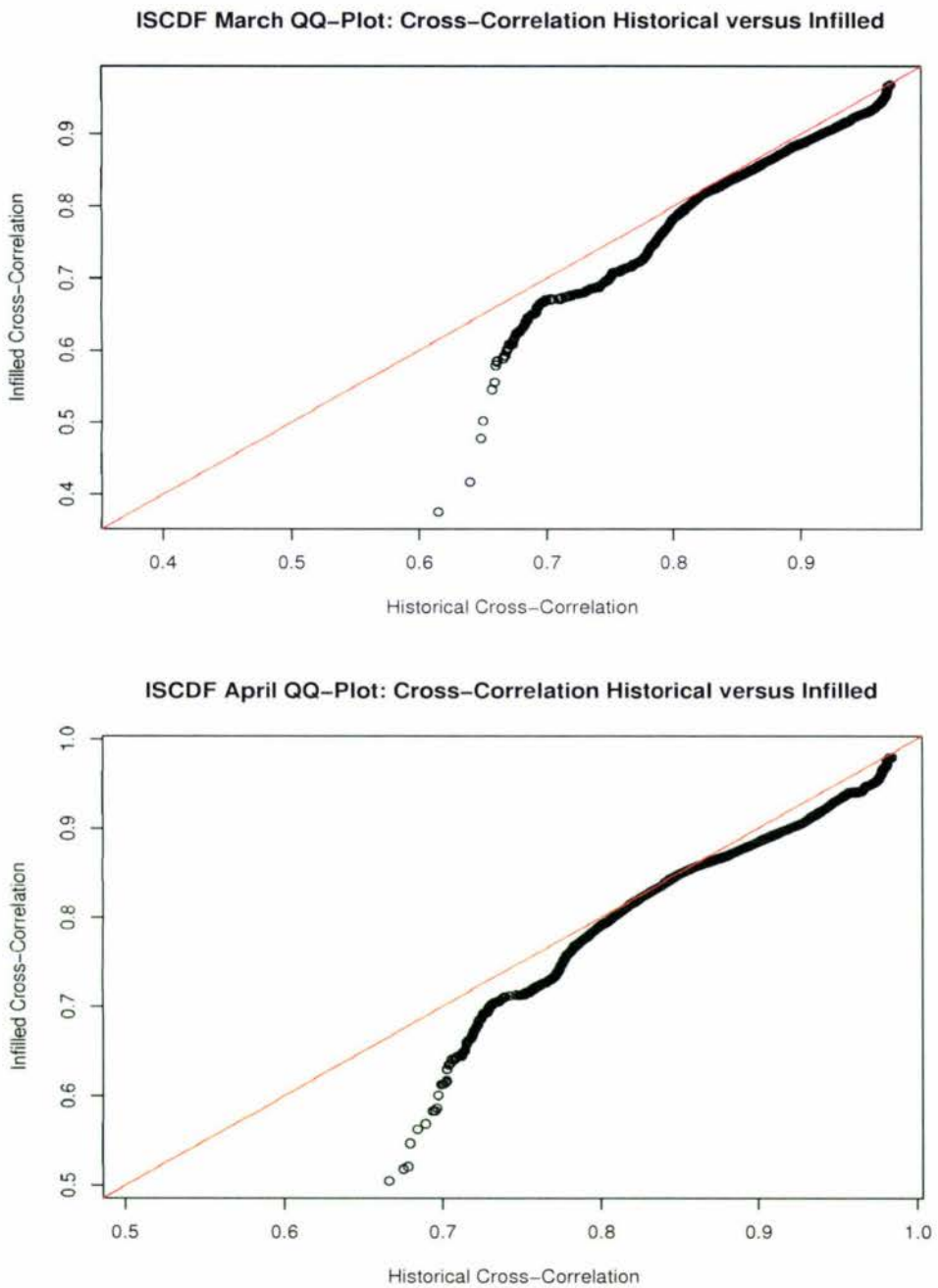


Figure D.38: ISCDF cross-correlation QQ plots: March and April

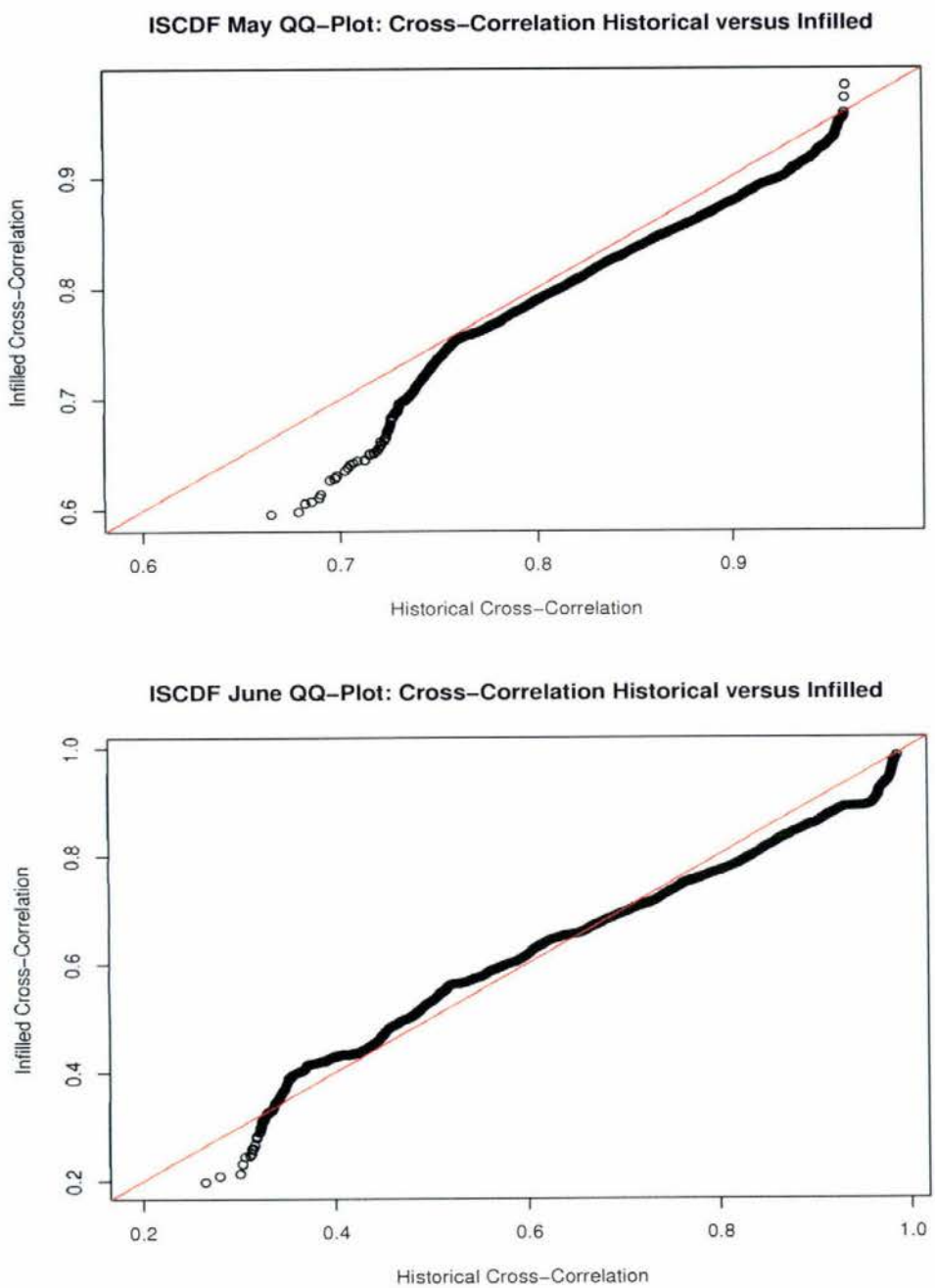
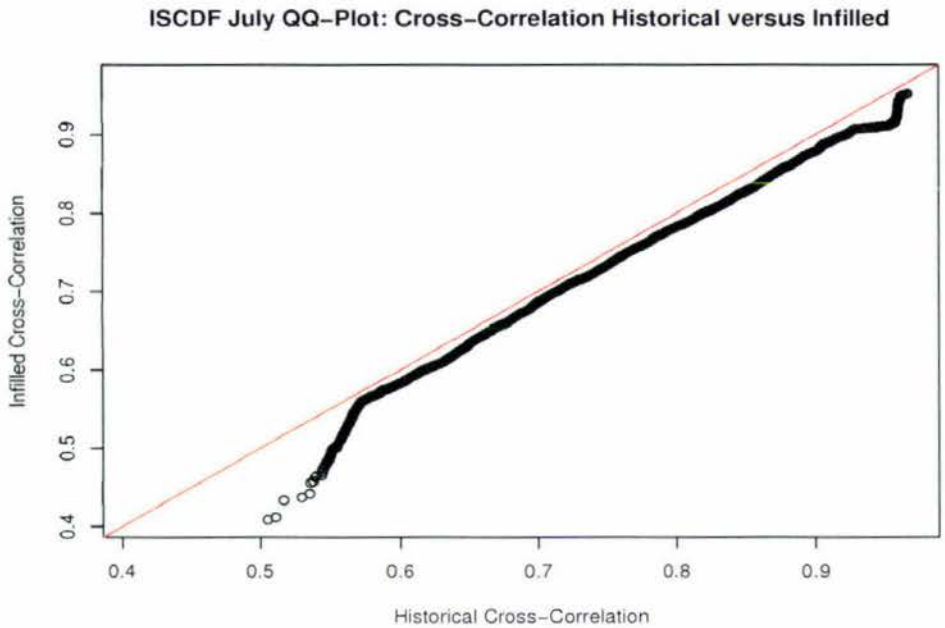


Figure D.39: ISCDF cross-correlation QQ plots: May and June



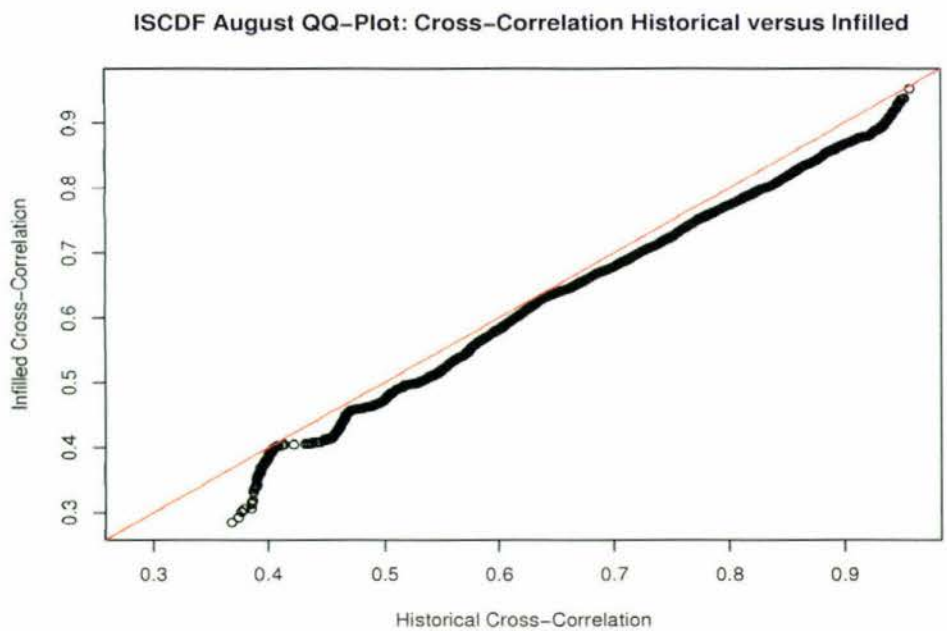


Figure D.40: ISCDF cross-correlation QQ plots: July and August

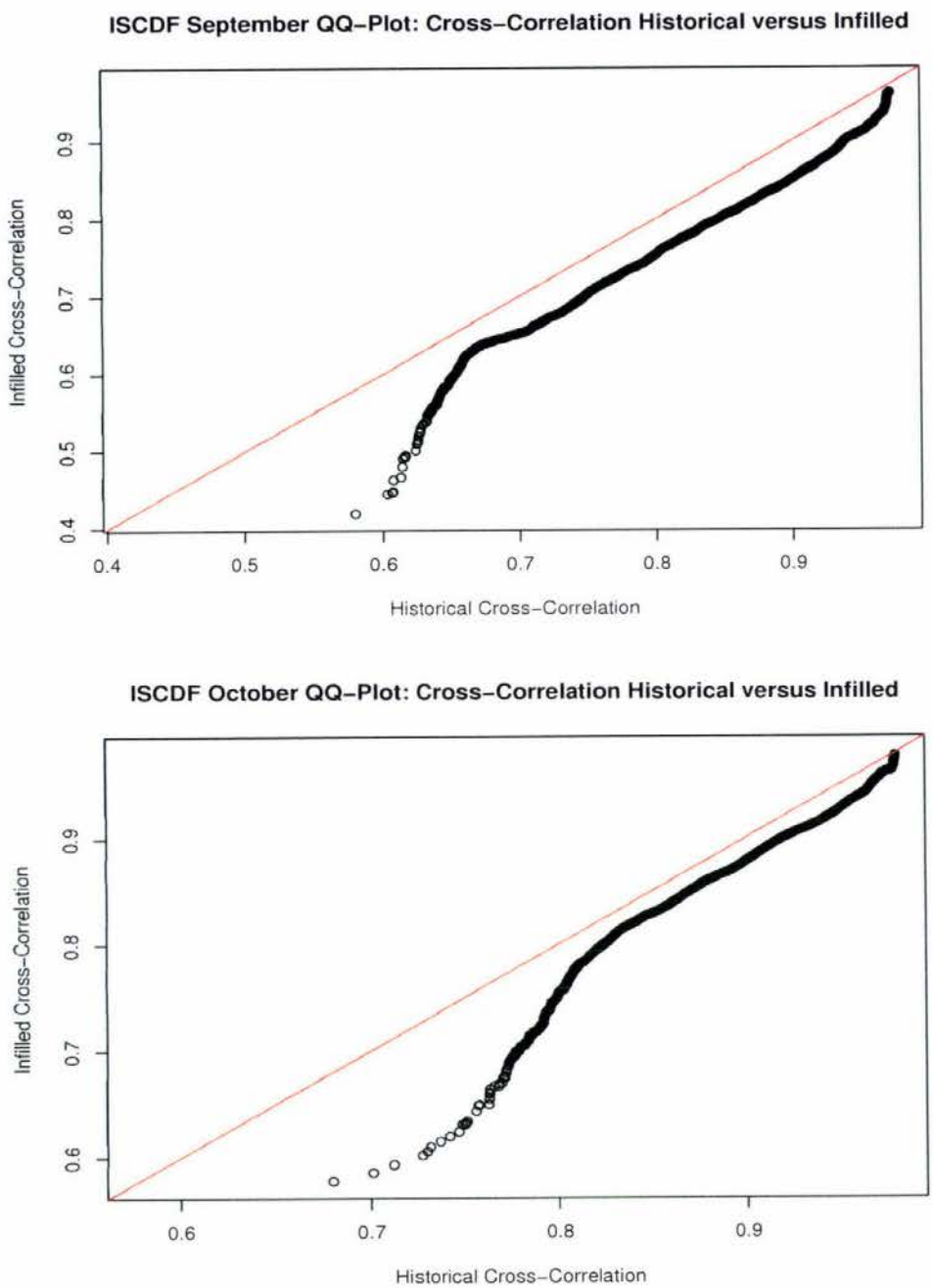


Figure D.41: ISCDF cross-correlation QQ plots: September and October

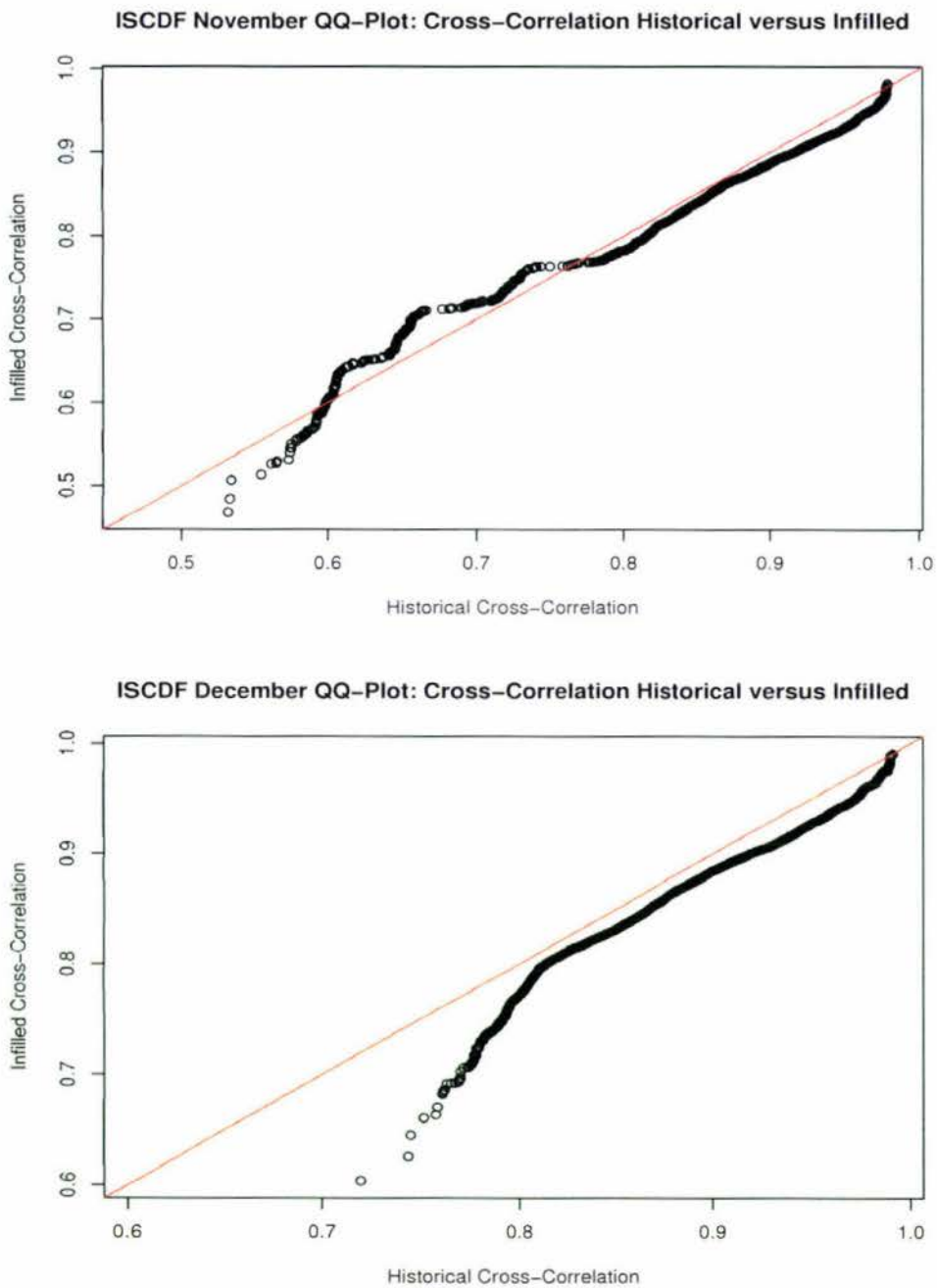


Figure D.42: ISCDF cross-correlation QQ plots: November and December