Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling

Taryn Swan

Department of Mathematics and Computing
The University of Southern Queensland, Toowoomba, QLD

July 7, 2006

Abstract

Improvements in computer power, data gathering techniques and an increased understanding of atmospheric dynamics has lead to the availability of extensive sets of rainfall data. With these improvements has come a widespread influx of researchers attempting to adequately model rainfall. This has proved quite difficult due to the extreme variability and uncertainty of rainfall. However, as rainfall has a major influence on virtually all human activity, modelling this seemingly unpredictable variable is extremely important. The considerable variation of rainfall is one of the main difficulties when modelling rainfall. A second difficulty is that rainfall is a continuous variable with an exact zero. Finally rainfall is a highly skewed variable and there is a general consensus that it has a non-independent structure. This dissertation will focus on modelling rainfall processes by using generalized estimating equations, when the assumed distribution of the response variable is from the Tweedie family of distributions. The Tweedie distribution allows the different components of rainfall to be modelled simultaneously. When the Tweedie distributions are incorporated into generalized estimating equation estimation techniques, not only can the non-independent structure of data be incorporate but also multiple rainfall sites can be modelled concurrently. This dissertation will attempt to demonstrate the potential benefits of using generalized estimating equations for modelling and interpreting historical rainfall records by the examination of of monthly rainfall models at Emerald, Toowoomba and Gatton. The suitability and predictability of the models presented in this dissertation will then be discussed to determine if generalized estimating equations provide a unique method of modelling rainfall.

Acknowledgements

There are numerous people who have assisted in the successful completion of my Honours dissertation and to all of you, I am eternally grateful:

- My supervisor Peter Dunn, who has guided and inspired me throughout my studies. Without your encouragement and support I would not even have started an Honours project, yet alone finish it. I sincerely thank you.
- Other members of staff who have given their support and have given me the confidence to ensure the successful completion of my dissertation and studies: Christine McDonald, Dr Ashley Plank and Paul Fahey. I can not thank you all enough.
- My sister Megan, whose wisdom with words enabled me to perfect my Honours project and whose constant companionship, helped me to overcome some difficult times. A simple thank-you does not seem enough.
- My parents whose constant support and love is never-ending. You are always there for me and I hope you always will be. I love you both very much.
- Last, but most importantly, my family: Brad, my soon to be husband, you mean everything to me and words alone can not say how much I love you (thanks also for putting up with me in the final stages of my dissertation and agreeing to marry me!); and my two gorgeous boys, Bailey and Aiden, who when times are tough put life into perspective and make me realise that family is all that matters. I love you all tremendously.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Environmental data is often composed of two separate components: a discrete element at zero and a continuous element recorded on the positive real line. For example, there may be no recordable level of rainfall or pollutant at particular points in time which constitutes the discrete element and is often called the occurrence component. When an amount is record, a continuous quantity greater than zero is documented and is referred to as the amounts component. Insurance claims and non-recurrent expenditure also display this two component characteristic. The two components are generally modelled separately with observations being incorrectly assumed to be independent. The family of Tweedie distributions (Section 3.5) has been previously used to enable a single model for rainfall to be produced, however independence of observations is still assumed. This dissertation specifically examines rainfall data with the main focus of modelling rainfall using 'Generalized Estimating Equations'. These models enable a single model to be produced that incorporates the dependent structure of rainfall.

Rainfall strongly influences the design and operation of hydrological systems, irrigation systems, farm management systems, water resource systems and urban drainage systems (Dunn [27]; Srikanthan & McMahon [76]; Hughes et al. [51]). Rainfall models are also extremely useful for agricultural planning as they provide a better understanding of erosion problems and help in the development of crop growth models (Srikanthan & McMahon [76]; Hughes et al. [51]). As accurate rainfall data is needed for many different systems, and is required for important management decisions, an extremely efficient model is needed that takes into account the extreme variability of rainfall (Chandler & Wheater [12]).

This dissertation commences in Chapter 2 with the examination of previous researchers' attempts at modelling rainfall. The overview of past studies demonstrates the use of two separate models to represent rainfall: rainfall oc-

currence models which examine the number of 'wet' and 'dry' rainfall events; and the modelling of rainfall amounts greater than $0mm$. This chapter also covers the difficulties involved with modelling rainfall and how researchers have overcome these complexities in the past, as well as possible covariates to use in a rainfall model.

Chapter 3 introduces generalized linear models and discusses the framework, definitions and parameter estimations of these increasingly popular models. Diagnostic testing used with generalized linear models are also briefly addressed in this chapter. Of particular importance is the discussion on the Tweedie family of distributions and their application with modelling rainfall data. This forms a fundamental part of this dissertation.

Following on from Chapter 3 is the establishment of generalized estimating equations, which are an extension of generalized linear models. Specifications and estimation techniques for these estimating equations are explained in Chapter 4. This chapter also discusses the possibility of using generalized estimating equations to model rainfall data and describes the notations and motivations behind this dissertation. The chapter ends with an explanation of the diagnostics that are available to use with generalized estimating equations and how to fit a generalized estimating equation to a given data set.

Three rainfall data sets from Emerald, Toowoomba and Gatton, are examined in Chapter 5. Preliminary analyses are performed on these data sets to test their suitability for use in this dissertation and to determine if any unusual patterns or rainfall amounts are present in the data. Monthly rainfall amounts are considered for all three locations. The chapter also examines some of the possible predictors that may be used to model the rainfall at the three locations.

Chapter 6 demonstrates how to develop a code to model rainfall, using a generalized estimating equation, when the response variable (rainfall) is assumed to have a distribution that comes from the Tweedie family of distributions.

Chapters 7 and 8 give applications of modelling rainfall using a generalized estimating equation. A systematic approach to developing the models is discussed and a final rainfall model is found. Chapter 7 gives an example of the application of generalized estimating equations for modelling rainfall by showing the development of a model at single rainfall site, Emerald. Chapter 7 expands on the ideas presented in Chapter 8 to develop a rainfall model for multiple locations. Both chapters end with a discussion on diagnostics, model interpretation and model validation.

Finally, Chapter 9 summarizes the accomplishments of this dissertation and proposes some improvements which could be made. Recommendation

for further research into this area are also suggested.

Chapter 2

Literature Review

2.1 Modelling rainfall

Rainfall can be modelled using various timescales including hourly, daily, monthly and annual timescales (Dunn [27]). The most commonly studied timescale is daily because it can discriminate between the number of wet days and rainfall amounts when it does rain. This provides a more detailed understanding of the rainfall process (Chandler & Wheater [12]). Little research has been completed on annual or monthly rainfall since 1985 (Srikanthan & McMahon [76]). Annual rainfall models have little direct application, however they are used in disaggregation schemes to obtain monthly data, which is used in the estimation of water demand and the simulation of water supply systems (Srikanthan & McMahon [76]).

Even though modelling daily rainfall has proved to be the most valuable and have the most potential, this dissertation examines monthly models. This is because size of daily rainfall data sets is typically very large making the modelling process very time consuming and more difficult. The applications involved in this dissertation include the development of a new initiative to model rainfall and thus monthly rainfall data is used. Monthly models still have considerable applications and provide an excellent foundation for developing daily models.

One of the main difficulties researchers encounter when examining the rainfall variable is its considerable variation from year to year. Two other difficulties include the fact that rainfall is a somewhat skewed variable, and that it is continuous with an exact 0. This second difficulty is problematic because most models cannot cope with modelling a mixture of both discrete and continuous distributions concurrently. Therefore, to minimize this problem, rainfall is typically modelled using a two-component model. The first

component examines the occurrence of rainfall: this is the probability of a 'wet' or 'dry' event occurring, and is usually formulated using a Markov process. The second component focuses on the actual rainfall amount once a rainfall event has occurred. While several researchers have attempted to model these two processes together, only Dunn [27] has managed to create a model that simultaneously models both the occurrence and amount together using only one distribution.

2.2 Modelling the occurrence of rainfall

Rainfall occurrence can be viewed as a sequence of random variables $X(t)$, $t = t_1, t_2, \ldots, t_T$, where,

$$X(t) = \begin{cases} 1 & \text{if rainfall has occurred on a particular day or month,} \\ 0 & \text{if no rainfall has occurred on a particular day or month} \end{cases}$$

The occurrence of rainfall is a discrete process, therefore Markov Chains and renewal processes are the most common methods used to model the probability of a 'wet' rainfall event occurring. These two processes have been studied extensively and are discussed in further detail Sections 2.2.1 and 2.2.2. Generalized linear models (GLMs) are also becoming increasingly popular to model rainfall data. GLMs are also the backbone of generalized estimating equations (GEE), and GEEs are the main focus of this dissertation.

2.2.1 Markov Chains

Markov Chains are commonly used to model the proportion of 'wet' rainfall events. This is due to the flexibility and ease at which parameters can be estimated using Markov Chains, as well as the ability and ease the final fitted model gives for obtaining results that do not require the use of simulations (Stern & Coe [77]). Markov Chains are also popular because of their largely non-parametric nature, ease of interpretability, and the well-developed literature about their development.

A Markov Chain model for rainfall usually specifies two states for each day: either 'wet' or 'dry'. They are also used to develop a relationship between the state of the current rainfall event and the state of any proceeding rainfall events (Srikanthan & McMahon [76] and Chapman [15]). This is known as the order of the process, and is the number of proceeding rainfall events taken into account in the model. For example, a first order Markov Chain indicates that the probability of rain falling on any rainfall event depends only on the state of the previous rainfall event (Coe & Stern [17]).

Most Markov Chain models referred to in the literature are of first order, evident from Gabriel and Newman's research in 1962 (Coe & Stern [17]) to Stern and Coe's research in 1984 (Stern & Coe [77]).

While first order models have been studied extensively, more recent research has also focused on higher orders. Katz [54] studied zero, first and second orders. Lung and Grantham [59] has fitted up to a 12th order chain to rainfall data. Another area of research into the order of the Markov Chains involved a hybrid order, in which wet spells are modelled as a first order but higher orders are used for dry spells (Gyasi-Agyei & Willgoose [35] and Gyasi-Agyei [34]). Other studies have also examined specific locations to test the use of a Markov Chain at different locations. For example, Harrison and Waylen [40] examined the humid topics of America, and Robertson [70] used a hidden Markov model to describe rainfall occurrence at ten stations in northeast Brazil during the wet season.

While many studies have tested the different Markov Chain order models to discover which order is the most efficient in modelling the occurrence of rainfall, there is a general consensus among researchers that a first order Markov Chain is adequate for most locations. This is because it is able to adequately model the data while keeping the number of parameters at a minimum. The advantages of the Markov Chain first order model is also enhanced by the findings that higher order Markov Chain models have a lack of parsimony. Chin [16] did however, suggest that the order needs to be seasonally sensitive and this means second or higher orders are needed at some times during the year. Coe and Stern [17] have adopted this idea, allowing the Markov Chain to change throughout the year. In their research, Coe and Stern [17] used a first order Markov chain during the dry season and a second order Markov Chain during the wet season. Lall et al. [55] also suggests that seasonally varying transition probabilities may be chosen to represent the changes in data during different seasons, and uses fourier series methods to parameterize these seasonal variations.

A further extension of the Markov Chain concept has been investigated by Srikanthan and McMahon [75] who applied it to a multi-state model in which rainfall is grouped into up to seven classes. This application allowed for the dependence between the transition probabilities and rainfall amounts to be considered, and has therefore proved to be more successful in modelling seasonal variations than other first or second order Markov Chains.

Although Markov Chains have proved popular and have been studied extensively, they are limited by their ability to efficiently model the amount of rainfall. One method that has been used to overcome this limitation is the division of rainfall amounts into categories: no rain; less than 5mm of rain; between 5mm and 20mm of rain; and more than 20mm of rain. However,

this technique provides only limited information and is not efficient enough when dealing with extremely important management decisions, such as crop growth. Markov Chains are also limited by the need for special computer programs to evaluate the models, making this procedure quite complex (Coe & Stern [17]).

2.2.2 Alternating Renewal Process

Markov Chains are only one approach that can be used to model the occurrence of rainfall. Another approach is the alternating renewal process. This process considers a sequence of alternating wet and dry spells of varying length, with each spell having an assumed distribution. It is further assumed that all intervals are independent, and that the distributions may be different between wet and dry spells (Srikanthan & McMahon [76]). The type of distributions that have been investigated to model the length of wet or dry spells using an alternating renewal process include: logarithmic series (Williams [87]); modified logarithmic series (Green [32]); truncated negative binomial (Buishand [7]); and the truncated geometric distribution (Srikanthan & McMahon [76] and Chapman [15]. Of these distribution types, Buishand [7]) found from studies in the Netherlands, that an alternating renewal process with a truncated negative binomial distribution provided an excellent fit for rainfall data.

Although the alternating renewal process offers a different method to model the occurrence of rainfall, it also has its limitations. Similar to the Markov Chain procedure, rainfall amounts can not be modelled accurately. In addition, the assumption of independence between dry and wet spell lengths is difficult to justify at short timescales (Lall et al. [55]). Finally by considering spells rather than rainfall events there is a reduction in the sample size. This process also has the added disadvantage that it does not compute seasonality efficiently (Srikanthan & McMahon [76]). This was shown by Roldan and Woolhiser [71] who, using the Akaike Information Criterion (AIC), compared the alternative renewal process and a first order Markov Chain. They found that the Markov Chain was superior for each of the stations studied because the alternating renewal process could not deal with seasonality.

2.2.3 Generalized Linear Models

A third type of model used to model the occurrence of rainfall are generalized linear models (GLMs). Generalized linear models (Section 3) are becoming a popular method to compute data that has high levels of variability, such as rainfall. From their studies into GLMs Coe and Stern [17] concluded that in

comparison to the complicated process of analysing non-stationary Markov Chains, GLMs were more superior for modelling rainfall data. Chandler and Wheater [12] [10] extended this idea and modelled the binary series of wet and dry days using logistic regression, a form of a GLM. They used the logistic regression in response to their criticism that Markov Chains do not adequately represent temporal dependence. In the logistic regression model, the probability of rain for the ith case in the data set is denoted as being p_i, and $\mathbf{x}_i$ represents the predictor vector. The logistic regression model therfore becomes,

$$\ln\,(\frac{p_i}{1-p_i}) = \mathbf{x_i'}\beta.$$

GLMs have been found to be particularly useful in modelling rainfall and have several advantages. They provide a flexible and rigorous framework that is able to make distinctions between different climate change scenarios, and are useful for interpreting historical rainfall records. They are also able to model the rainfall at a daily timescale and are amenable to simulation. Finally they allow the uncertainty in prediction to be accounted for in the fact that the predictions are in the form of probabilities rather than point values.

2.2.4 Additional occurrence models

There have been several other methods developed to model the occurrence of rainfall. One method uses a mixture of geometric and negative binomial distributions (Feuerverger [29]). Another method uses a nonparametric technique by resampling from the historical records (Harold et al. [41]). A third method, developed recently by Henien [43], focuses on the autoregressive conditional Poisson model which deals with the issues of discreteness, overdispersion and correlation within the data. It claims to be more straightforward in its testing for autocorrelation than other approaches, such as GEEs, however it lacks the ability to model rainfall occurrence and intensity together as one.

A final class of models which have been used to model the occurrence of rainfall are time series models. These models ensure that temporal dependence is included in the model. A type of time series model is the two-state discrete autoregressive moving average (DARMA) model, which was first used by Buishand [7] and more recently by Chang et al. [13]. When comparing studies using the DARMA model with studies using the alternating renewal process (discussed in Section 2.2.2) it was found that the alternating renewal process models were more superior than the DARMA model when using data

from the Netherlands. However, when tropical and monsoon areas were examined the DARMA model proved more promising (Buishand [7]).

2.3 Modelling the amount of rainfall

Numerous different distributions have been used to model the amount of rainfall that occurs on a wet day. This rainfall amount is sometimes called "daily intensity" and it is a continuous distribution. This type of data is usually modelled using a parsimonious member of the exponential family that best the given data set. Dunn [27] states that as rainfall is highly skewed to the right, distributions that follow this same pattern and similarly are skewed to the right, have proven to be the most useful, with the gamma distribution being the most commonly used. There are also several other distributions following this pattern that are used by researchers: Srikanthan and McMahon [76] used the skewed Normal distribution to model rainfall amounts and Bardossy and Plate [4] examined a truncated power of the Normal distribution.

2.3.1 Gamma distribution

The gamma distribution is used in meteorology and climatology studies to represent the variations in rainfall amounts. A gamma distribution has the following density function, with the parameter, α, governing the shape of the distribution,

$$f(x) = \begin{cases} kx^{\alpha-1}e^{-x/\beta}, & x > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where $\alpha > 0$ (the shape parameter) and $\beta > 0$ (the scale parameter).

When using the gamma distribution to model rainfall amounts, it is common practice to keep the shape parameter, α constant. However Coe and Stern [17] state that the parameters of the gamma distribution should change throughout the year. This means that these distributions take into consideration the variances that occur in rainfall on any particular rainfall event. This variance is taken into consideration by varying μ, the mean amount of rain per wet day, as illustrated by Buishand [7] who fitted a Fourier series to the varying means of the gamma distribution.

Das [20] was one of the first researchers to model rainfall using a truncated gamma distribution on daily rainfall. This distribution measures rainfall for all days, not just those days on which rain occurred. Stern and Coe [77] also examined this idea by using a shifted gamma distribution. This model takes

into account the fact that some rain amounts under a certain value cannot be recorded (see Section 2.6.1 on trace values).

To classify different categories of rainfall amounts, Wilks [86] used three different gamma distributions. The gamma distributions had the common shape parameters, but used differing scale parameters. It was found that the mean rainfall generally increased from classes 1 to 3.

Another special form of the gamma distribution is the exponential distribution. This distribution occurs when μ is equal to 1. Todorovic and Woolhiser [81] used the exponential distribution to model rainfall amounts proposing it was easier to handle analytically. Wilks [86] further developed this model using a mixed exponential distribution, which is a mixture of two different exponential distributions.

Despite the popularity of the gamma distribution for modelling rainfall intensity, there are some disadvantages to using this method. One disadvantage is the maximum likelihood of the constant k is biased, and while this bias can be calculated when μ is constant, it cannot be calculated when μ varies. This limitation however, can be avoided when the number of observations are large (Coe & Stern [17]). A second disadvantage is that some gamma distribution methods rely on defining a threshold (See Section 2.6.1 on trace values), below which days are classified as dry. However, the question remains as to who defines this threshold and how reliable it is. An alternative method, the censored gamma distribution, uses the number of rainfall events which recorded 'small' amounts rather than the actual amounts. However using this model requires a modification to ordinary techniques for estimation parameters (Buishand [7]). Furthermore, another downfall is that fitting a gamma distribution to rainfall amounts and testing their goodness of fit usually requires specifically written computer programs which are complex and specific to each different model. Finally, the methods described in this section only monitor the amount of rainfall and not the occurrence; a model that describes both would be a more useful way to model rainfall.

2.3.2 Generalized Linear Models

As stated in Section 2.2.3 GLMs are useful for modelling rainfall data as they can model the high levels of variability effectively and can interpret different climate change situations. Coe and Stern [17] were the first to fit a gamma-based GLMs to rainfall data, and found that these types of models are easy to fit and interpret. They did not, however, incorporate temporal dependence into the model. Chandler and Wheater [12] [9] also fitted a gamma-based GLM to model rainfall amounts on wet days, however they did try to include temporal dependence.

The mean of the gamma distribution, $\boldsymbol{\mu}_i$ is determined by the values of various predictors, with $\mathbf{y}_i$ being the distribution of the rainfall amount for the ith wet day. The structure of the GLMwill then take the follwing form (Chandler and Wheater [9] [12]),

$$\ln \boldsymbol{\mu}_i = \ln E(\mathbf{y}_i) = \mathbf{x}_i \boldsymbol{\beta}, \tag{2.1}$$

where, $\mathbf{x}_i$ is a vector of predictors and $\boldsymbol{\beta}$ is a vector of corresponding unknown parameters. An advantage of using this method is the relatively simple diagnostics that are available to check the adequacy of a fitted model.

Chandler and Wheater [9] further demonstrate that the gamma distribution is the most appropriate to model rainfall intensity through the analysis of Anscombe residuals, which show a very satisfactory fit of the gamma model. They also go on to state that no one has given a rigorous justification for using the exponential or log-normal distributions and in fact, the exponential distribution provides an extremely poor fit to the data.

2.3.3 Resampling and non-parametric techniques

There are several resampling methods and non-parametric techniques that have been used to model the amount of rainfall. Lall et al. [55] used a technique similar to bootstrapping or sampling with replacement by resampling daily rainfall to produce a stochastic model. The probability distribution functions of alternating wet and dry spell lengths and of rainfall amounts were estimated using nonparametrically kernel density estimators. This method preserved the characteristics of wet and dry spells. The tests for selecting parametric distributions, such as the chi-square test, lack the power to discriminate between different candidate distributions, and the assumption of independence of wet and dry spells is debatable due to the heterogeneous nature of rainfall data (Lall et al. [55]). Therefore nonparametric techniques are becoming increasingly popular in stochastic hydrology and allow for no prior assumptions to be made about the overall functional form of the target function.

Disadvantages with this method are that larger sample sizes are needed in comparison to parametric estimation. In addition, spell definitions are questionable and are best at finer timescales such as daily. This is because sample sizes drop rapidly when longer time scales, such as monthly, are considered.

To improve on the kernel-based non-parametric simulation approach developed by Lall et al. [55], Rajagopalan and Lall [68] uses a multivariate, nonparametric time series simulation. The method used by Rajagopalan

and Lall [68] improves on the former approach by using a conditional bootstrap based on nearest neighbour probability estimation. It allows statistical properties of historical data to be honoured and again does not rely on prior assumptions to form the joint probability density function. However, this method does come with some drawbacks. Since it is a bootstrap, simulations can not produce values that have not been observed in the historical data, such as extreme values. Furthermore, the computational power needed is extremely large and there is a possibility that the bounds on some variables will be violated during the simulations.

2.3.4 Additional modelling of rainfall amounts

While researchers in the aforementioned studies examined only one type of distribution in each of their studies, Chapman's research [15] compared several distributions including the exponential, mixed exponential, skewed normal, gamma and the Kappa distributions. The skewed Normal distribution was found to be the most consistent, while the gamma and exponential distributions were the least consistent. It is unlikely that a single distribution will provide a good fit to rainfall data for all climatic regions (Walden & Guttorp [84]).

The effect of classifying rainfall amounts into one of three categories depending on the number of wet days recorded has also be investigated (Chapman [15]). These categories included: if it was a solitary day of rain; if rain fell on a day at the beginning or the end of a wet spell; and if the rain fell in the middle of a wet spell. Different distributions were used for each category and it was found that the models that take these categories into account generally perform better than those models that clump the data together (Chapman [15]). Buishand [7] also employed a similar approach and found that there was a small but significant correlation between rainfall amounts on successive wet days.

Another way of modelling the amount of rainfall was proposed using a class of generalized autoregressive moving average models (GARMA) to model non-Normal situations like rainfall (Benjamin et al. [60]). These models extend the GLM method by incorporating time dependence within observations. However, they can only use a set distribution and a set time dependent framework is needed.

2.4 Modelling the amount and occurrence of rainfall

While several researchers have tried to simultaneously model the occurrence of rainfall and the amount of rainfall, only Dunn [27] has been successful in producing a model that requires a single distribution. A Tweedie distribution was used to model both the discrete and continuous components of rainfall discussed in more detail in Section 3.5.

Several other researchers have attempted to model amounts and occurrence together. However they have not been successful in modelling the two components simultaneously and instead produce methods that use two separate models. Rajagopalan and Lall [69] developed a nonhomogeneous Markov model that used kernel methods to estimate a nonhomogeneous transition-probability matrix. This matrix models the rainfall occurrence, and estimates a corresponding nonstationary probability density function of daily rainfall amount. Another model proposed was a mulitstate Markov chain which treated rainfall as a mixed discrete and continuous variable and the transition probabilities are used to model the dependence structure (Haan et al. [36]). Yau et al. [88] used two generalized linear mixed models for analysing insurance claim data: one for the occurrence; and one for the intensity of claims. Although it does not model rainfall data, it models data which takes a similar form and uses random effects to model the correlation between individuals.

2.4.1 Generalized Additive Model

Grunwald and Jones [33] claim to combine the occurrence and intensity rainfall models into a single model for amount. They achieve this by using a first order Markov structure and a mixed transition density, with a discrete component at 0 and a continuous component for the positive sample space. Hyndman and Grunwald [48] used the same method, but they combined it with a generalized additive model (GAM) to relax the assumption that each year follows the same seasonal pattern. One advantage of the GAM methods is that it allows for the modelling of non-seasonal temporal variations, whereas GLM methods do not.

The amounts model takes the following form,

$$p_t(y|X_{t-1} = x_{t-1}) = [(1 - \pi_t(x_{t-1}))\delta_0(y) + \pi_t(x_{t-1})f_t(y|x_{t-1})],$$

with Y_t representing a random variable (rainfall) at time t; X_{t-1} is the vector of covariates; $\pi_t(x_{t-1})$ is the probability of getting rainfall on a given

event; and $f_t(y|X_{t-1})$ is the continuous density. In this example, a gamma distribution is used as f_t and $\delta_0(y)$ is a Dirac delta function.

Even though a single model is given in these studies, both admit that the functions and parameters are estimated separately. The occurrence distribution, $\pi_t(x_{t-1})$ is estimated first, followed by estimation for the the intensity distribution $f_t(y|x_{t-1})$. Thus, although these researchers do provide a single formula for rainfall occurrence and intensity, each component in rainfall still needs to be estimated separately.

2.4.2 Tweedie distributions

As mentioned in Section 2.4, the Tweedie distribution GLM is the only model that has successfully been developed to require only one distribution. The Tweedie family of distributions (named after Tweedie [83]) are a class of distribution that are able to model both discrete and continuous probabilities together as one model. Tweedie distributions are based upon generalized linear models and are classified by their variances, which takes the form $\text{var}[Y] = \phi\mu^p$ (See Section (3.5) for a discussion on these models). There are three main properties that makes the Tweedie distributions exceptional for modelling rainfall (Dunn [27]),

- The Tweedie distributions belong to the exponential family of distributions (See Section (3.1.1)), and form part of a larger group of models called the generalized linear model; This is advantageous because fitting techniques and diagnostics are readily available for generalized linear models, and thus for Tweedie distribution models;

- The motivation behind the setup of these models is simple and logical: total rainfall is considered the sum of rainfall on some smaller time scale;

- The Tweedie distributions provide a mechanism in which finer-scale structures can be understood through courser-scale data.

Two examples, the Charleville monthly rainfall and the Melbourne daily rainfall, were used to show how the Tweedie distribution worked for modelling rainfall (Dunn [27]). It was found using these examples, that the Tweedie family of distributions was useful in modelling rainfall on both a daily and monthly scale, and that modelling both the occurrence and amount of rainfall can be done simultaneously.

2.5 Modelling rainfall using multiple sites

Rainfall displays the largest variability among the meteorological variables in time and space, and thus dependence of rainfall at different sites should be accommodated for in models (Srikanthan [76]). Although this is an important issue, there has been limited work completed into creating a rainfall model using multiple locations. This is probably due to the difficulty involved when trying to simultaneously model more than one location.

Several different methods of modelling rainfall at multiple sites has been attempted. A family of multivariate models was used to represent the occurrence of rainfall at N sites (Zucchini & Guttorp [92]). This was done by introducing unobservable states to account for the different distributions of rainfall over the sites. This type of modelling is often referred to as a 'Hidden Markov Model'. This idea was extended to model the occurrence of rainfall using a nonhomogeneous hidden Markov model (Hughes et al. [51]). It was found that this model could provide scientists with a useful tool for generating realistic simulations of rainfall. Beersma and Buishand [6] used a nearest-neighbour resampling technique to generate multi-site sequences of daily rainfall in the Rhine basin. Finally, a Markov Chain for the occurrence and a mixed Exponential distribution for the intensity, was used to simultaneously generated rainfall at multiple locations (Wilks [86]).

2.5.1 Bayesian approach

Several researchers have used the Bayesian approach to model rainfall. An advantage of this method is its ability to overcome the ignorance of sampling errors that can lead to underestimating the range of the mean and variances (Srikanthan [76]). The Bayesian approach also allows for uncertainties in the parameter estimates to be taken into account and to be expressed through the posterior distribution. Sansø and Guenni [73] used a truncated and transformed multivariate normal distribution to model the rainfall at several different sites and made extensive use of an MCMC method.

2.6 Difficulties in rainfall modelling

Researchers face several difficulties when trying to model rainfall. The two factors that are often most challenging for researchers are the capricious nature of rainfall and its distributional form (it has a continuous form with an exact zero). Other factors that also create problems by adding to the complexity of rainfall modelling include: trace values; temporal and seasonal

variations; extreme events and; why and how many parameter to include in the model. The following discussion examines these other difficulties and how researchers have attempted to overcome them.

2.6.1 Trace values

When a very small amount of rainfall falls, it is extremely difficult to accurately record the exact amount, thus placing a limit on the smallest amount of rainfall that can be accurately recorded. The rainfall amounts that fall below this limit, are usually referred to as trace values. They are classified as any non-zero amount below some threshold, usually set at 0.1mm (Chandler and Wheater [12]). The problem with these values their classification: if a trace value is recorded, should the event be classified as wet or dry. Furthermore, once classified, the question arises as to what value should be assigned to these days: a zero, or a definite value. As trace values account for approximately 11% of wet events, it is quite important that these values are dealt with appropriately in any model calculating rainfall amounts.

When creating a rainfall model, a trace value may occur in two of the variables. It may occur in the predictor variable, or it could occur in the response variable if the predictors involve previous events' rainfall amounts. It is fairly straightforward to deal with trace values when they occur in the predictor variable. This is achieved by defining an extra predictor, which is 1 if an amount is recorded as a trace, or zero for no recorded amount. Trace values that occur in both the response and explanatory variables are not so easy to deal with. It has been suggested that one way to manage trace values in the response and explanatory variables is to treat the circumstance as a 'censored data' situation and therefore reformulate the likelihood function to take into account that some of the observations are not recorded accurately (Chandler & Wheater [12]). However, if a gamma distribution is used to model the rainfall amount, then a difficult integral is created than cannot be solved analytically. A better solution would be to replace each censored response value with its conditional expectation under the current model parameterization (Chandler & Wheater [12]). However this method could be quite costly if large data sets are involved. Another approach to deal with trace values when modelling rainfall occurrence would be to use a three-state Markov Chain (Stern & Coe [77]). Using the Markov Chain, the first two states would be the usual 'wet' or 'dry' situation. However a third state could be included to encompass those values classified as trace values.

2.6.2 Temporal dependence

Temporal dependence refers to the dependence of a variable upon past values of time, and implies that a variable is correlated. Correlated data occurs when data is collected on the same unit across successive points in time (Horton & Lipsitz [46]). Buishand [8] proposes that rainfall is neither independent nor identically distributed as many researchers assume and is, in fact, correlated. Therefore, temporal dependence should not be ignored when modelling rainfall. For rainfall data, temporal dependence implies that each rainfall amount or occurrence is dependent upon some number of previous rainfall events' amounts or occurrences. If temporal dependence is not taken into account when fitting a model, the standard errors of parameter estimates will not be valid and any inferences completed will not be replicable (Horton & Lipsitz [46]).

Chandler [11] suggests that the simplest way to deal with temporal dependence, within a GLM framework, is to include functions of previous rainfall values as extra covariates. Durban and Glasbey [28] use a multivariate latent Gaussian process to model rainfall, and a vector auto-regressive moving average (VARMA) to model the temporal dependence between rainfall variables. Although this model takes the dependence into consideration, it does not adequately fit the data for extreme values of rainfall. All Markov chain models incorporate the temporal dependence by choosing an order that takes into account previous occurrences of rainfall.

This dissertation examines a method developed to specifically deal with correlated data called Generalized Estimating Equations (GEEs). Although GEEs have been used in many different research areas such as clinical trials, health trials, and insurance claims, they have never been used to analyse rainfall data. As GEEs are specifically designed to deal with dependent data, such as rainfall, it is logical to use GEEs in this situation.

2.6.3 Spatial dependence

Spatial dependence is described by Barnsley [5] as a certain variable displaying similar (or different) values depending on the spatial location at which it is measured. For rainfall this is explained as the fact that two closely-spaced rainfall gauges should display closer recordings than a pair of gauges that are a greater distance apart, providing all other variables are kept constant. Spatial dependence is based on the assumption that gauges closer together receive rainfall from the same cloud, while at more distant locations the rainfall may vary due to variations in synoptic rainfall patterns.

The most common way to deal with spatial dependence is to use mul-

tivariate techniques. However Chandler [11] suggests that these techniques rely on the decomposition of an empirical covariance matrix, and as such, the implicit assumption that observations are independently and identically distributed through time. Multivariate techniques are also difficult to interpret as the results are made purely for mathematical convenience. Section 2.5 describes methods which have been developed to deal with the problem of multi-site modelling. Chandler [11] also suggests that inter-site dependence in the response may be able to be addressed using GEEs, however this is yet to be tried.

The lack of ordering of the sites in space means that spatial dependence cannot be dealt with naturally using a factorization technique. Instead, some researchers have fitted separate models to each site. This method however means that any systematic relationship between variables is lost. Although spatial dependence is important when modelling rainfall, little work has been completed in this area due to the difficulties involved. Most recently Chandler and Wheater [12] have developed some solutions to spatial dependence using the generalized linear models framework, however there is still a considerable amount of work needed in this area before a consistent model is created.

2.6.4 The number of parameters in the model

One of the main aims of modelling is to create a parsimonious model: a simple model that captures all of the important features of the data. The difficulty with modelling rainfall is that often a large number of parameters are needed. Rainfall amounts quite often follow a seasonal pattern, and this is the occurrence and amount of rainfall that occurs throughout the year changes every season. In order to take this seasonal effect into account, many researchers fit parameters that vary throughout the year. This creates an overall model that has a very large number of parameters and thus can become quite complex (Chapman [15]). In order to overcome this complexity and therefore minimise the number of parameters used, some researchers have fitted the parameter variation to a polynomial (Coe & Stern [17]) or Fourier Series (Roldan & Woolhiser [71]).

2.7 Covariates that may be used

The extreme variability of rainfall means numerous covariates could be included in a rainfall model and the choice of which covariates to consider is difficult. The distribution of rainfall depends on many different factors such as topography, elevation, proximity to forest covers and lakes and other

climatic conditions such as temperature (Sahur & Andres [72]). It is often difficult to obtain some predictors as they are have not been recorded accurately and thus often the researcher is limited to the choice of predictors that be can used when modelling rainfall.

Chandler and Wheater [12] suggest using previous rainfall amounts, time of year (for example month), variables representing topographic effects and regional differences in rainfall patterns as covariates in a rainfall model. They also consider the possibility of incorporating long-term climatic variability in the model by fitting a predictor that varies from year to year.

Through their research, Chandler and Wheater [10] found several predictors to be important in their final GLM for rainfall. These predictors include: gauge elevation; seasonal patterns in the form of sine and cosine curves; functions involving rainfall amounts occurring previously; and interactions of previous rainfall amounts. A GLM with a gamma distribution and a logarithm link function were used along with the listed covariates to model rainfall amounts.

The southern oscillation index (SOI) is another variable that is often associated with rainfall amounts and its association has been widely investigated (Stone & Auliciems [78] and Stone, Hammer & Marcussen [67] and Troup [82]). Five SOI phases relating the SOI to rainfall have also been investigated as possible predictors of rainfall in Eastern Australia (Stone & Auliciems [78]). However, Hyndman [47] shows that the SOI does not provide a strong predictor of rainfall, contrary to current meteorological practice. This study did not examine rainfall in Eastern Australia where the association between SOI and rainfall is the strongest.

Numerous covariates have been suggested by different researchers as possibilities in a rainfall model. The covariates that have been selected in this dissertation are those that are readily available and have been accurately recorded for the required time frame.

Chapter 3

Generalized Linear Models

For decades, simple linear regression models formed the basis of most analyses of continuous data. These models take the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where the elements of $\mathbf{e}$ are assumed to be independent and identically distributed with a Normal distribution $N(0, \sigma^2)$. $\mathbf{X}\boldsymbol{\beta}$ is a combination of covariates. The assumption that the error ($\mathbf{e}$) values follow a Normal distribution is often not correct, making these linear regression models inadequate. Recent advancements in statistical theory and computer software has led to improvements in these linear models by the creation generalized linear models (GLMs). GLMs are able to model situations in which the response variables have distributions other than, and including the Normal distribution, as well as model situations when the relationship between the response and explanatory variables is not of simple linear form (Dobson [21]). GLMs have been explored in the literature since 1970s when they were established by Nelder and Wedderburn [64].

GLMs provide a flexible and rigorous framework that are able to deal with the high levels of variability such as in rainfall data (Chandler & Wheater [12]). The GLM approach has also proven to be a very powerful tool for interpreting historical rainfall records. The success of GLMs is due to the balance between simplicity and generality, that the design of these models has achieved both computationally as well as conceptually.

This chapter examines and defines GLMs, and provides diagnostic checks to determine the adequacy of a model. As GLMs form the basis of generalized estimating equations (GEEs), this chapter also contains important concepts needed in the formulation of GEE methodology.

3.1 The GLM framework

When modelling a data set using a GLM, three decisions need to be made before the model can be produced,

- What is the distribution of the response variable?
- What function of the mean will be modelled as linear in the predictors?
- What will the predictors be?

Deciding on the answers to these three questions defines the components needed to create a GLM (McCulloch & Searle [63]). The first is the existence of $n \times 1$ random variables $Y_1, \ldots, Y_N$ dependent on r predictors. These random variables form the response variables, which are assumed to share the same distribution and come from a specific family of distributions called the exponential dispersion model (EDM) family. The second component of a GLM is the link function, which relates the parameters of the distribution to various predictors. The last component uses a set of p unknown parameters, $\boldsymbol{\beta}$, and a set of $n \times r$ known explanatory variables $X_{n\times r} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]^T$, formed together so that $\mathbf{X}\boldsymbol{\beta}$ is a linear structure. This linear structure describes how the location of the response variable changes with the explanatory variables (Lindsey [57]).

As GLMs are traditionally formulated within the framework of a set of distributions which belong to the family of EDM, this family of distributions is described formally in Section 3.1.1 below. A formal definition of a GLM, that incorporates the information above, follows this (Section 3.2).

3.1.1 Exponential Dispersion Models

The theory of GLMs is based upon the exponential family of distributions. This formalisation recharacterises familiar functions into a formula that is theoretically more useful (Gill [31]). Exponential dispersion models are very versatile as they can be discrete, continuous, or can have a mixed distribution.

Definition

An exponential dispersion model (EDM) has a probability density function or a probability mass function, that can be written in the following form,

$$p(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{1}{\phi}[y\theta - \kappa(\theta)]\right\}, \tag{3.1}$$

where $\phi > 1$ is the dispersion parameter; μ is the position parameter and $\mu = \kappa'(\theta)$; Y is the variable of interest and θ is the *canonical parameter*. It is important to note that Y does not depend on the parameters θ and ϕ, and the function $a(y, \phi)$ cannot always be written in closed form. Also, it is necessary to ensure that the total summation of Y over the domain is one.

The notation $Y \sim ED(\mu, \phi)$ indicates that a random variable Y comes from the EDM family, with location parameter μ and dispersion parameter ϕ, as in Equation (3.1).

Distributions of the EDM Family

The Normal, binomial, Poisson, inverse Gaussian, exponential, gamma, and Tweedie distributions all have distributions that form part of the exponential dispersion model family. The binomial and Poisson distributions are both discrete distributions, with the Poisson distribution being used when the data involves counts. The binomial distribution is used when the data deal with proportions and the outcome is either a 'success' or 'failure'. The Normal, inverse Normal, exponential and gamma distributions are all continuous distributions. The gamma distribution is used when the response variable is skewed and the variance is not constant. The exponential distribution is a special case of the gamma distribution used when the shape parameter (α) is equal to one. Finally, the Tweedie distribution is a mixed distribution, which means that it can model data with both discrete and continuous components, such as the Poisson-gamma distribution. The Tweedie distribution is especially useful in modelling rainfall, as illustrated in Section 3.5. Table 3.1 provides information about several distributions that come from the EDM family, including their variance functions (See Section 3.2.2). These seven distributions demonstrate that EDMs can consist of discrete, continuous, or mixed distributions.

3.2 GLM definition

Formally, generalised linear models are stated to consist of two components (Dobson [21]; McCullagh & Nelder [62]; Dunn & Lennox [24]),

1. The response variable, y_i, follows an EDM with mean μ and dispersion parameter ϕ, such that,

$$y_i \sim \text{ED}(\mu_i, \phi/w_i),$$

where w_i are known prior weights (often one); and

Table 3.1: The characteristics of some of the distributions of the exponential dispersion model family (McCullagh & Nelder [62]).

Distribution	$\kappa(\theta)$	$\mu = E(Y)$	Variance Function
Normal	$\theta^2/2$	θ	1
Poisson	e^θ	e^θ	μ
Binomial	$\ln(1+e^\theta)$	$e^\theta/(1+e^\theta)$	$\mu(1-\mu)$
Gamma	$-\ln(-\theta)$	$-1/\theta$	μ^2
Inverse Gaussian	$-(-2\theta)^{\frac{1}{2}}$	-2θ	μ^3
Tweedie	$\theta(1-p)^{(2-p)/(1-p)}/(2-p)$ for $p \neq (1,2)$	$\kappa'(\theta)$	μ^p for $p \neq (0,1)$

2. The expected values of the y_i, say μ_i, are related to the covariates $\mathbf{x}_i$ through a monotonic differentiable *link function* $g(\cdot)$. This link function is described further in the following section.

3.2.1 Link Function

The link function, $g(\cdot)$, of a GLM relates the expected values of the y_i, commonly written as μ_i, to the covariates, $\mathbf{x}_i$, as follows:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

It is essential that the link function chosen is differentiable, so that $\boldsymbol{\beta}$ can be estimated and monotonic, to ensure that each value of $\mathbf{x}_i^T\boldsymbol{\beta}$ has only one corresponding μ_i value. The major advantage of the link function is that it can be chosen independently of the distribution.

Often the linear component, $\mathbf{x}_i^T\boldsymbol{\beta}$, is called the *linear predictor* and is given the symbol η_i, so that,

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}. \tag{3.2}$$

The most common link function to use, is the *canonical link function*, defined as $\eta = \theta = g(\mu)$. This function however, is not always the best function to use. One advantage of using the canonical link function for a given distribution is that all the unknown parameters $\boldsymbol{\beta}$ of the linear component η have sufficient statistics for distributions that are EDMs, which simplifies the fitting algorithm (Section 3.3) for GLMs. A list of commonly used link functions for the binomial, Poisson and gamma distributions and their canonical link functions can be seen in Table 4.1.

Table 3.2: The link functions commonly used for the binomial, Poisson and gamma generalized linear models; ϕ is the Normal cumulative distribution function, and μ and p are the mean value parameters (McCullagh & Nelder [62]).

Distribution	Canonical Link	Form	Other Links	Form
Binomial	logit	$\log[p/(1-p)]$	probit	$\Phi^{-1}(p)$
			c-log-log	$\log[-\log(1-p)]$
Poisson	log	$\log \mu$	identity	μ
			square root	$\log(\mu)$
Gamma	inverse	$1/\mu$	log	$\log(\mu)$
			identity	μ

3.2.2 Additional properties of GLMs

Mean and Variance

Members of the EDM, family written in the form of Equation (3.1), have a mean and variance defined as follows, where $\kappa(\theta)$ and ϕ are determined from Equation (3.1) (McCullagh & Nelder [62]),

- Mean of Y:
$$E[Y] = \mu = \kappa'(\theta). \tag{3.3}$$
- Variance of Y (var$[Y]$):
$$\text{var}[Y] = \phi\kappa''(\theta). \tag{3.4}$$

The variable θ is related to the mean μ through Equation (3.3). The relationship between μ and θ is often written as $\tau(\theta) = \kappa'(\theta) = \mu$ and $\theta = \tau^{-1}(\mu)$. The function $\tau(\theta)$ is referred to as the mean-value mapping and gives the functional relationship between μ and θ.

Variance Functions

Although not described in the original setup of a GLM, the variance function is important as it uniquely identifies a distribution within the class of EDMs. Equation 3.3 shows that $\kappa'(\theta)$ is a function of the mean and thus $\kappa''(\theta)$ is also dependent on the mean. For this reason $\kappa''(\theta)$ is often replaced by the variance function $V(\mu)$ so that,

$$V(\mu) = \kappa''(\theta).$$

The role of the variance function is to describe the mean-variance relationship of a distribution when the dispersion parameter is held constant. If Y follows an EDM with mean μ, variance function $V(\mu)$, and dispersion parameter ϕ, then the variance of Y can be written as,

$$\text{var}(Y) = \phi V(\mu).$$

The variance function that uniquely identifies the Normal, binomial, Poisson, inverse Gaussian, gamma, and Tweedie distributions is illustrated in Table 3.1. The Tweedie distributions, described in Section 3.5, are classified by a special form of the variance function ($\text{V}(\mu) = \mu^p$).

Deviance

One method to measure the appropriateness of a fitted model is to examine the difference between the fitted values $\hat{\boldsymbol{\mu}}$ and the observed values $\mathbf{y}$. In standard Normal distribution based regression, this measure is equivalent to the residual sum-of-squares (Hardin & Hilbe [38]). In the framework of GLM, this measure of difference is called the *deviance*, $\text{D}(y;\mu)$, and can be calculated as follows,

$$D(y;\mu) = \phi D^*(y;\mu) = 2\phi[\ell(y;y) - \ell(\hat{\mu};y)], \tag{3.5}$$

where D^* is called the scaled deviance and has only an approximate χ^2 distribution, and ℓ is the log-likelihood function. The deviance can be used to compare models. For further details on the deviance of GLM refer to Hardin & Hilbe [38] and Nelder & Wedderburn [64].

3.3 Estimation of parameters

To fit a model to a data set, estimates of the parameter values β_j are needed. The maximum likelihood method is used to estimate the parameters for GLMs, with the parameters being estimated numerically using a iterative procedure (Dobson [21]). To obtain the maximum likelihood estimators of the parameters β_j, the likelihood function is also needed. In general, the likelihood function is defined as,

$$L(\xi;y) = \prod_{i=1}^{n} f(y;\xi),$$

where n is the sample size of the data set, and ξ is the parameter of interest. Often, it is easier to work with the log-likelihood function, which is defined

as,

$$\begin{aligned}\ell(\xi; y) &= \log L(\xi; y) \\ &= \log \prod_{i=1}^{n} f(y; \xi) \\ &= \sum_{i=1}^{n} \log f(y; \xi).\end{aligned}$$

To use this theory for GLM methodology, the log-likelihood function needs to be applied to a EDM (Equation (3.1)). The log-likelihood of an EDM is,

$$\ell(\theta, \phi; y) = \sum_{i=1}^{n} a(\phi, y) + \frac{1}{\phi}[y\theta - \kappa(\theta)]. \tag{3.6}$$

The maximum likelihood estimates for β_j can now be found by taking the derivative of Equation (3.6) with respect to β_j. This is found through the following series of derivatives,

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta_i} \times \frac{d\theta_i}{d\mu_i} \times \frac{d\mu_i}{d\eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}. \tag{3.7}$$

Each of these four derivatives can be found individually using the following method,

- The first component $\partial \ell / \partial \theta_i$ is obtained directly by differentiating the log-likelihood function, seen in Equation (3.6):

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_i} &= \sum_{i=1}^{n} \frac{1}{\phi}[y_i - \kappa'(\theta_i)] \\ &= \sum_{i=1}^{n} \frac{1}{\phi}[y_i - \mu_i],\end{aligned}$$

  since $\mu_i = E[Y] = \kappa'(\theta_i)$.

- The second component uses the relationship $\mu_i = E[Y] = \kappa'(\theta_i)$ as well:

$$\begin{aligned}\mu_i &= \kappa'(\theta_i) \\ \frac{d\mu_i}{d\theta_i} &= \frac{d\kappa'(\theta_i)}{d\theta_i} \\ &= \kappa''(\theta_i) \\ &= V(\mu_i).\end{aligned}$$

  Inverting this final equation thus gives $d\theta_i / d\mu_i = 1/V(\mu_i)$.

- The third component differentiates the link function $g(\mu_i) = \eta_i$,

$$\begin{aligned} \eta_i &= g(\mu_i) \\ \frac{d\eta_i}{d\mu_i} &= \frac{g(\mu_i)}{d\mu_i} \\ &= g'(\mu_i). \end{aligned}$$

  Inverting this final equation gives $d\mu_i/d\eta_i = 1/g'(\mu_i)$.

- The final expression uses $\eta_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ir}$, where r is the rank of $\boldsymbol{\beta}$. Thus, the derivative of η_i with respect to β_j is x_{ij}.

Combining these four expressions shows that Equation (3.7) can be written as the 'score equation' for GLMs,

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}. \tag{3.8}$$

The maximum likelihood estimator is found by setting Equation (3.8) equal to 0 and solving for $j = 1, 2, \ldots, r$. When $\partial \ell / \partial \beta_j = 0$, the value of ϕ does not need to be known. This is an important concept of GLMs because an estimate of $\boldsymbol{\beta}$ can be found without knowing ϕ.

The set of Equations (3.8) can only be solved through numerical techniques involving iteration, such as the Newton-Raphson method or the method of scoring. See Dobson [21] and Hardin & Hilbe [38] for further details about these two methods.

3.4 Quasi-Likelihood methods

In many situations, some details of the distribution governing the data is known, however the distribution may not be able to be specified exactly. In addition, there are some cases for which the distribution is known, however it difficult to evaluate, such as the Tweedie distributions. This precludes the use of maximum likelihood, which requires exact specification of the distribution in order to construct the likelihood. The idea of quasi-likelihood addresses this concern (McCulloch & Searle [63]).

Quasi-likelihood methods were first proposed by Wedderburn [85], and are a methodology for regression that requires few assumptions about the distribution of the dependent variable. Hence they can be used with a variety of outcomes (Zeger & Liang [89]). In likelihood analysis, the actual form of

the distribution must be specified. However, in quasi-likelihood, only the relationship between the outcome mean and covariates, and the mean and variance, needs to be specified (Zeger & Liang [89]). The focus of quasi-likelihood is on methods for inference about β, and hence ϕ can be treated as a nuisance parameter.

A quasi-likelihood can be used if the researcher does not know the density function of the distribution, but knows its mean and variance. It is defined for one observation, Q, as,

$$Q(y;\mu) = \int \frac{(y-\mu)}{V(\mu)} d\mu. \tag{3.9}$$

This quasi-likelihood has the same properties as a true log-likelihood with regards to the derivatives of β, enabling GLMs and GEEs to be fitted for any distribution using a quasi-distribution. To define a quasi-likelihood function, only the relationship between the mean and variance needs to be specified through the variance function (Wedderburn [85]).

3.5 Power-variance (Tweedie) distributions

Of special interest within EDMs is a class of distributions with power mean-variance relationships $V(\mu) = \mu^p$. Any distribution whose variance function like this this belongs to the class of distributions known as the Tweedie family of distributions, named by Jørgensen [53] after Tweedie [83]. This section describes Tweedie distributions, and demonstrates how the these distributions can be used to model rainfall.

Most of the important distributions commonly associated with GLMs are contained within the Tweedie distribution framework, including the Normal ($p = 0$), Poisson ($p = 1$ and $\phi = 1$), gamma ($p = 2$), and inverse Gaussian distributions ($p = 3$). Tweedie models exist for all values of p outside the interval $(0, 1)$, however only the four distributions already mentioned have density functions which have explicit analytic forms (Dunn & Smyth [26]).

Tweedie distributions with $p > 1$ have strictly positive means, with $p > 2$ being continuous for positive Y, and a shape similar to the gamma, but more right skewed. Distributions with $p < 0$ are continuous on the entire real axis. Finally,, for $1 < p < 2$ the distributions are supported on non-negative real numbers, and the distributions are mixtures of the Poisson and gamma distributions, with a mass at zero (Dunn & Smyth [26]). These distributions have been called 'compound Poisson', 'compound gamma', and 'Poisson-gamma' distributions. Due to the characteristic of being able to model both discrete

and continuous combinations simultaneously, these distributions have a special use in being able to model both the occurrence and amount of rainfall.

The mean, μ, and canonical parameter, θ can be found for a Tweedie distribution by noting that $\kappa''(\theta) = d\mu/d\theta = \mu^p$ and the mean is given by $\mu = \kappa'(\theta)$. This allows the density function for a Tweedie distribution to be specified (See Dunn [**?**] for more information). Hence,

$$\begin{aligned}\mu^p &= \frac{\partial^2 \kappa}{\partial \theta^2} \\ &= \frac{\partial}{\partial \theta}\left(\frac{\partial \kappa}{\partial \theta}\right) \\ &= \frac{\partial \mu}{\partial \theta}.\end{aligned}$$

Taking the reciprocals of both sides and integrating with respect to μ gives,

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & p \neq 1, \\ \log \mu & p = 1. \end{cases}$$

By setting the arbitrary constant of integration to 0, and noting that $\mu = \kappa'(\theta)$ gives,

$$\kappa(\theta) = \begin{cases} \frac{\mu^{2-p}}{2-p} & p \neq 2, \\ \log \mu & p = 2. \end{cases}$$

The Tweedie densities can thus be written as,

$$f_p(y; \mu, \phi) = a_p(y, \phi) \exp\left\{\frac{1}{\phi}\left[y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right]\right\} \text{ for } p \neq (1, 2). \qquad (3.10)$$

Tweedie distribution and the Quasi-likelihood

Following Equation (3.9) the Tweedie distribution has the following quasi-likelihood distribution (when setting the arbitrary constant of integration to 0),

$$\begin{aligned}Q(\mu; y) &= \int \frac{(y-\mu)}{V(\mu)} d\mu \\ &= \int \frac{(y-\mu)}{\mu^p} d\mu \\ &= \int \frac{y}{\mu^p} - \mu^{1-p} d\mu \\ &= \int \left(y\mu^{-p} - \mu^{1-p}\right) d\mu \\ &= \frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}.\end{aligned}$$

This equation has the same likelihood function as Equation (3.10), except now there is no need to estimate $a(y, \phi)$. This is extremely helpful as often the $a(y, \phi)$ term can not be written in closed form, or is of a form which is extremely difficult to calculate.

3.5.1 Software

The program R is used extensively in this dissertation and is the program that was used to create the rainfall models in generated in Chapters 7 and 8. To fit a Tweedie generalized linear model the `tweedie` library is needed in the R program (Dunn [23]). There are two functions which were of particular use in this dissertation: the `tweedie.profile`; and the `tweedie` family model. The `tweedie.profile` function finds the most suitable Tweedie distribution for the given data set using maximum likelihood methods. This function only works for $p \geq 1$ and gives the maximum likelihood value of p and ϕ and a 95% confidence interval for p. The `tweedie.profile` function uses the default series expansion as the calculation method, however interpolation and inversion methods are also available. This dissertation uses the interpolation method to find the most appropriate Tweedie distribution.

To fit a GLM with a Tweedie distribution, the variance power (p found using `tweedie.profile`) and link function are needed to be specified. The default link function is the canonical link function, with the logarithm link function also available. The following command is used in R to fit a GLM,

```
family=tweedie(var.power=p,link.power=1-var.power).
```

3.5.2 Tweedie Distributions and Rainfall

To model rainfall using the Tweedie model, one vital assumption needs to be made: the amount of rainfall that occurs during any rain event follows a gamma distribution. Research shows that it is very common to use the gamma distribution to model the amount of rainfall (see Section 2.3.1) and thus, this assumption is valid. By following this assumption a Tweedie model can be set up to model rainfall (as completed by Dunn [27]).

Let i be a rainfall event, and R_i be the amount of rainfall that occurs during this event, it is assumed that each R_i follows a gamma distribution, with mean $-\alpha\gamma$ and variance $-\alpha\gamma^2$ ($\text{Gam}(-\alpha, \gamma)$). It is also assumed that the number of rainfall events during the time period (usually month or day), called N, follows a Poisson distribution with mean λ. Thus when no rainfall has occurred on that particular event, $N = 0$. Finally Y represents the total daily or monthly rainfall, and is represented as the Poisson sum of

gamma random variables, such that $Y = R_1 + R_2 + \ldots + R_N$, where N was defined earlier. This same setup can be applied to differing timescales. For example, if R_i represents the amount of rainfall per day, then Y is the total monthly rainfall. The resulting distribution of Y is called a Poisson-gamma distribution (Dunn [27]), and belongs to the class of Tweedie distributions when $1 < p < 2$.

A Poisson-gamma distribution has a complicated probability function, however Jørgensen [53] shows that it takes the following form,

$$\log\, f_p(y;\mu,\phi) = \begin{cases} -\lambda, & \text{for } \; y = 0 \\ -y/\gamma - \lambda - \log\, y + \log\, W(y,\phi,p), & \text{for } \; y > 0, \end{cases}$$

where $\gamma = \phi(p-1)\mu^{p-1}$, $\lambda = \mu^{2-p}/[\phi(2-p)]$, and W is an example of Wright's generalized Bessel function. It can be written as,

$$W(y,\phi,p) = \sum_{j=1}^{\infty} \frac{y^{-j\alpha}(p-1)^{\alpha j}}{\phi^{j(1-\alpha)}(2-p)^j j!\Gamma(-j\alpha)},$$

where $\alpha = (2-p)/(1-p)$. The mean of the Poisson-gamma distribution is μ, and its variance, as with all Tweedie distributions, is $\text{var}[Y] = \phi\mu^p$. The probability of obtaining no rainfall on any particular event is given by the following formula (Dunn [27]),

$$\Pr(Y = 0) = \exp(-\lambda) = \exp\left[-\frac{\mu^{2-p}}{\phi(2-p)}\right]. \tag{3.11}$$

3.6 Diagnostic Testing

The purpose of creating a model is to adequately summarize the important characteristics of the data by finding a parsimonious model that explains what is happening in the data without using meaningless, or too many parameters. In the creation of a model, often this model may show departures from the given data and thus not fit the data sufficiently. Diagnostic testing is used to determine whether the model adequately fits the data. There are a number of diagnostic tests that are available for GLM (some of which will be described below), and these include: a Q-Q plot; scatterplots of residuals and covariates; comparison of residual sizes; and residual deviances. These techniques allow the suitability of the link function and assumed distribution to be tested, as well as testing of the data for influential values, outliers, or pattern.

There are four main reasons why a fitted GLM may not adequately represent the data and these include,

- The model fits well for most observations, however a few isolated cases do not. These isolated cases are called outliers;
- The link function is incorrectly specified;
- The response variable, Y is incorrectly specified; and/or
- The linear predictor (η) may not be correctly specified, or is missing some terms.

Using a GLM to model rainfall data is not the main intention of this dissertation and thus the diagnostic testing available for these types of model will only be briefly discussed. The diagnostic testing that is available for GEE models differs from the testing available for GLMs and therefore for further discussions on diagnostic testing for GLMs see McCullagh & Nelder [62] and Dobson [21].

3.6.1 Residuals

A general tool used in diagnostic analysis is residuals. Residuals are a measure of how different expected values of the responses emerge from the observed responses. In simple regression models, the raw residuals $(y - \hat{y})$ are used, however these are generally inadequate when using a generalized linear model. The two most common residuals to use for GLMs are the Pearson residuals and deviance residuals. The Pearson residuals, which are also used in GEE models, have an approximate Normal distribution $N(0, \phi)$. Deviance residuals are related to the concepts of deviance $D(y; \mu)$, and also have an approximate Normal distribution. Quantile residuals have also been recently proposed by Dunn & Smyth [25] to be used with GLMs, and have an exact Normal distribution when μ and ϕ are known exactly.

Definition of Quantile Residuals

In continuous responses, the quantile residual is defined as,

$$r_{Q,i} = \Phi^{-1} F(y_i; \mu_i, \phi),$$

where $F(y_i; \mu_i, \phi)$ is continuous and is the distribution function of a random variable Y, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution.

In the discrete case, if $a_i = \lim_{y \uparrow y_i} F(y_i; \mu_i, \phi)$ and $b_i = F(y_i; \mu_i, \phi)$, then the quantile residuals are defined as,

$$r_{Q,i} = \Phi^{-1}(\mu_i),$$

where μ_i is a uniform random variable on the interval $(a_i, b_i]$.

3.6.2 Residual Plots

Any of the residuals discussed in section 3.6.1 can be plotted against a variety of statistics and other indices. Each provide different information about departures from the fitted model. Since residuals should ideally be random, any pattern observed in the plots indicate problems with the fitted model. These residual plots can therefore help the researcher determine if there are any isolated departures. Furthermore, by plotting the residuals against the fitted values, as well as against the covariates, systematic departures can also be determined (Chandler [11]).

Correct Distribution

One of the most important components of a GLM is that the correct distribution is chosen for the response variable. To check that the chosen distribution is adequate for the data, a normal probability or Q-Q (quantile) plot can be produced. If the model fits well, this plot should yield a straight line at 45 degrees. While quantile residuals are the ideal choice for GLMs, other residuals can be used.

Chapter 4

Generalized Estimating Equations

The class of generalized linear models (GLMs) introduced in Section 3.2 play a central role in regression problems which have discrete or continuous response variables. However they are based on the classical assumption that observations within a data set are independent. GLMs were extended by Liang and Zeger [56] so that longitudinal or correlated data (Section 4.1.1) could be analysed, and this approach is known as the Generalized Estimating Equation (GEE) method. This method has received wide use in medical and biological applications such as epidemiology, gerontology, and biology (Ballinger [3]), and is becoming increasingly popular in other disciplines such as organisational and psychological research. Much of the appeal of GEEs is due to their broad capabilities, including: modelling correlated responses; allowing for time-varying covariates; and facilitating regression analysis on dependent variables that are not normally distributed (Ballinger [3]).

4.1 Introduction

GEEs were introduced as a method of estimating the regression model parameters when the response variable is dependent. The GEE approach differs in a fundamental conceptual way from the techniques included under the rubric of 'random-effects', 'multilevel', and 'hierarchical' models which have previously been used to model correlated data. The techniques used in these models explicitly model and estimate the variations seen between observations, and incorporate these estimates and the residual variance into standard errors. The GEE method does not explicitly model the variation. Instead it focuses on, and estimates its counterpart: the similarity of the observations

(Hanley et al. [50]).

GEEs develop a population average or marginal model. In marginal models, the primary interest of the analysis is to model the marginal expectation of the response variable given the covariates. In other words, for every one-unit increase in a covariate across the population, the GEE tells the user how much the average response would change (Zorn [91]). The correlation, or more generally, the association between the response variables is modelled separately and is regarded as a nuisance parameter (Ziegler et, al. [2]). Thus, a basic premise of the GEE approach is that the researcher is primarily interested in the regression parameters $\boldsymbol{\beta}$ and is not interested in the variance-covariance matrix. GEEs are not meant to be used in situations in which scientific interest centres around the variance parameters.

This chapter focuses on the class of GEE models originally developed by Liang and Zeger [56]. This GEE approach is now commonly referred to as the GEE1 approach. Further developments are currently being made into different types of GEEs. While the focus of the chapter is on GEE1 models, the other types of GEEs are discussed briefly.

4.1.1 Longitudinal and correlated studies

GEEs are traditionally used to model correlated data from longitudinal or repeated measures units, as well as from clustered or multilevel studies. Longitudinal studies are defined by the characteristic that subjects are measured repeatedly throughout time. These studies require special statistical methods because the set of observations taken on one unit are usually intercorrelated (Diggle, Liang & Zeger [65]).

The issue of accounting for correlation also arises when analysing a single time series of measurements, such as rainfall. Although similar techniques can be applied to this type of data, inferences are usually less robust. The correlation must therefore be taken into account in order for valid scientific inferences to be made (Diggle, Liang & Zeger [65]).

The data examined in this dissertation is a single time series measurement, rainfall, which is measured over time. The site of the rainfall is considered as one unit, and the rainfall measured at each site would be the repeated measures over time. The prime advantage of studying rainfall in this manner is that multiple sites can be examined simultaneously and it is an effective way to study change. However, if more than one site is examined simultaneously in one model, it can be thought of as an example of a longitudinal study.

Correlated data has been examined through a variety of different approaches. The statistical methods for modelling longitudinal data are well

developed when the response variable is approximately Normal (Liang & Zeger [56]). Statistical models for non-Normal outcomes however, are not as developed. Where analysing longitudinal data there are two classical approaches which have been used in the past: the first is univariate mixed-model, split-plot, or repeated measures ANOVA; and the second is based on a multivariate ANOVA called MANOVA. Two other extensions to the classical approaches for modelling correlated data include multivariate modelling and mixed models. The former treats all measurements on the same unit as dependent variables, and models these simultaneously. The latter focuses on fixed and random effects within the model, with the correlation between the observations being a consequence of random effects (Dunlop [22]).

4.1.2 Notation

The following notation is used for the remainder of this dissertation: let $\mathbf{y}_{it}$ be a vector of responses with a set of corresponding r covariates or factors, $\mathbf{X}_{it}$, where i indexes the K units of analysis $i = 1, 2, \ldots, K$; and t indexes the time points $t = 1, 2, \ldots, n_i$ for each unit. Thus the number of clusters observed is K. Also, $N = \sum n_i$, and is the total number of observations across all units. The first element of $\mathbf{x}_{it}$ is set to 1 to allow the inclusion of an intercept.

Furthermore, let $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{in_i}]$ denote the corresponding column vector of observations on the response variable for unit i, and $\mathbf{X}_i = [X_{i1}, X_{i2}, \ldots, X_{in_i}]$ indicate the $n_i \times r$ matrix of covariates for unit i.

In the case of rainfall data, to correspond with the notation described above, the following notation is applied:

- Each site forms one unit or cluster. Therefore if only one site is examined, $K = 1$. If two sites are examined, then $K = 2$. Thus, $i = 1$ for one site and $i = 1, 2$ for two sites.

- The response variable, y_{it}, is the amount of rainfall recorded. Thus, if one site is examined, the response variable becomes $\mathbf{y}_{1t}$. If two sites are examined, there are two response vectors of $\mathbf{y}_{1t}$ and $\mathbf{y}_{2t}$.

- The observed time, t, corresponds with the time values at which the rainfall is measured. For example $t = 1, 2, 3$ would correspond with measurements taken at time point 1, time point 2, and time point 3.

- The number of time points is n_1 for site one and n_2 for site two.

4.1.3 Additional representations of GEEs

GEEs were originally defined by estimating the correlation parameters within the GEE using the method of moments (Sutradhar [79]). This means that only first order moments, that is, the mean structure, is estimated consistently. This approach, sometimes termed the GEE1 approach, allows the separation between the estimating equation for the regression parameters and the association parameters. Separating these parameters means that each are estimated individually. The GEE1 approach are the class of models that is most commonly found in statistical software implementation. Further work into GEEs has resulted in the development of the GEEJ_1 and GEEJ_2 approaches, which use similar principles as the GEE1 approach. Furthermore, another class of estimating equations, the GEE2 approach, has been developed. This approach however, does not separate the regression and association parameters (Hedeker [42]).

In the GEEJ_1 approach, as in the GEE1 approach, the regression parameters are estimated, but the 'working' correlation parameter α (See Section 4.3.1) is estimated by using a second set of estimating equations. The GEEJ_1 also requires formulas for the fourth-order moments, which are unknown and thus calculations can become quite complicated (Sutradhar [79]). Hall & Severini [37] avoided this problem, and estimated the 'working' correlation parameter, by using second-order moments only. This approach is called the GEEJ_2. From an efficiency point of view, the GEEJ_1 and the GEEJ_2 approached do not appear to produce better models than the GEE1 approach, and both are much more complicated than the GEE1 method (Sutradhar [79]).

Another form of the GEE is called the GEE2 approach. This uses a set of equations that allow the estimation of the first and second moments jointly and consistently, meaning that both the parameter and correlation estimates have to be approximated using a joint estimating equation approach (Ziegler et al. [2]). Another difference with this approach is that it constructs a true correlation structure and not a 'working' structure as the GEE1 approach does. However, it is extremely difficult, and as it has many convergence problems in the estimation of regression parameters the estimating equations become useless in the longitudinal setup (Sutradhar [79]). Extensions of the GEE2 include the GEGEE approach, the GGEE2 approach and the EGEE approach, all of which use similar principles to the GEE2 and have similar pitfalls.

Although other methods of estimation have been investigated, due to their difficulties and limited advancements, it has been decided that the simpler GEE1 method of estimation should be used for this dissertation. For the remainder of this dissertation, the GEE1 approach is simply referred to as

'GEE'.

4.1.4 Assumptions

Before explaining the concept of GEEs, there are four assumptions about the use of GEEs to model correlated data that need to be articulated. The most crucial assumption is that the following conditional expectation needs to be specified correctly,

$$\mu_{it} = E[y_{it}|x_{it}] = E[y_{it}|\mathbf{X}_i]. \tag{4.1}$$

Equation (4.1) implies the conditional mean μ_{it} of y_{it}, given the explanatory variable X_i, measured at all possible time points n_i, is equal to a set of the same point specific explanatory variables x_{it} (Dahmen & Ziegler [19]).

The second assumption is that the response variable y_{it} should have a mean and variance which are characterised by a GLM (Equation (4.2), Section 4.4). It is further assumed that a true conditional $n_i \times n_i$ covariance matrix exists (Dahmen & Ziegler [19]). Finally, it is imperative that any missing data is missing completely at random (MCAR), otherwise results become inconsistent (Dobson et al. [1]).

4.2 GEEs and rainfall

There is a general consensus that rainfall is correlated. For monthly rainfall this means that the rainfall observed during any particular month, depends on a number of previous months' conditions. Numerous studies have shown that this is the case, and thus the correlated structure of rainfall data should not be ignored when creating a model (Chandler & Wheater [10]; Beersma [6]; Buishand [7]).

Even though researchers have realised that rainfall data is correlated, introducing these dependencies into a model leads to difficulties. For example, parameter identification becomes difficult and models have an increased number of parameters. Thus, researchers typically assume that rainfall is independent. However, Lall et al. [55] state that if this independent assumption is violated, then the precision of any results obtained are over or underestimated and this leads to incorrect conclusions about the significance of parameters (Dahmen & Ziegler [19])

Past research thus shows that it is important to take the correlated structure of rainfall into account when creating a rainfall model. Generalized estimating equations are especially designed to handle correlated data and past reviews indicate that utilising this powerful estimating technique may be beneficial to rainfall modelling.

4.2.1 GEEs and the power-variance (Tweedie) GLM

Although several approaches have been used to model non-Normal continuous or discrete correlated data, no research has been completed on simultaneously modelling non-Normal data with continuous and discrete components. That is, GEE models have not been researched for the power-variance (Tweedie) GLM and thus this dissertation demonstrates an initiative approach to GEEs and rainfall modelling.

4.2.2 Multiple sites

Comparing rainfall across regions is traditionally computed using a direct standardisation approach that adjusts for confounding discrete factors (such as the SOI phase and month). Alternatively, The GEE approach can adjust for continuous, as well as discrete factors, and parameter estimation is more efficient when dealing with correlated longitudinal data (Ballinger [3]).

As stated earlier, limited research has been conducted into modelling rainfall generated from multiple sites, although the literature does state that multi-site modelling is important (Srikanthan & McMahon [76]). The techniques involved in modelling multiple sites has proved very difficult and tedious. It is a possibility that using GEEs to model rainfall may lead to new research into multi-site modelling. If each site is treated as a different cluster, then the GEE approach is viable.

4.3 Specification of GEEs

A basic feature of GEE models is that the joint distribution of a unit's response vector $\mathbf{y}_i$ does not need to be specified. Instead, only the marginal distribution of y_{it} at each time point needs specification. To clarify, assume there are two time points and the outcome variable is approximately Normal. GEEs only assume that the distribution of y_{i1} and y_{i2} are two univariate Normal distributions, rather than assuming that y_{i1} and y_{i2} form a (joint) bivariate Gaussian distribution. Thus, GEEs avoid the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time point (Hedeker [42]).

Since the GEE model can be thought of as an extension of GLMs for correlated data, the GEE specifications involve those of GLM, with one addition. Thus, GEE models require the user to specify the following,

- The linear predictor,

$$\eta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta},$$

where $\mathbf{x}_{it}$ is the covariate vector for unit i at time t.

- The link function, used to relate the response variable to the linear combination of the covariates,

$$g(\mu_{it}) = \eta_{it}.$$

- The variance as a function of the mean, and consequently the distribution of the response variable,

$$\text{Var}[Y_{it}] = \phi V(\mu_{it}).$$

- The correlation structure of the response variable.

The fourth condition is what differentiates a GEE model from a GLM. Liang and Zeger [56] introduced a 'working' correlation structure to obtain consistent and efficient estimators for regression parameters when observations were correlated.

4.3.1 Working correlation matrix

It is assumed a true correlation between units exists, however it is very rare that this true correlation is actually known. Thus, a working correlation matrix, $\mathbf{R}$, is produced to obtain an estimate of the covariance matrix (Zorn [91]). This working correlation is of size $t \times t$ because one assumes that there are a fixed number of time points t at which units are measured. A given unit does not have to be measured at all t time points; each individual's correlation matrix $\mathbf{R}_i$ is of size $n_i \times n_i$, with the appropriate rows and columns removed if $n_i < t$.

It is further assumed that the correlation matrix $\mathbf{R}$, and thus $\mathbf{R}_i$, depend on a vector of association parameters, denoted by α. That is, the working correlation matrix, now fully defined as $R_i(\alpha)$, is completely specified by the vector of unknown parameters, α. This unknown vector of parameters has a structure which is determined by the investigator and is assumed to be the same for all units. It represents the average dependence among the observations.

Although $R_i(\alpha)$ is chosen at the researchers' own discretion, it is best to try to choose $R_i(\alpha)$ to be consistent with empirical correlations and on the basis of theoretical considerations (Dobson et al. [1]). This is because accurately representing the correlation matrix improves the efficiency of the GEE estimates. Despite this, there is little information available about how to choose the best correlation structure (Dahmen & Ziegler [19]), and often it

is difficult to determine. As long as μ_i is correctly specified however, and the covariance matrix converges to some fixed matrix, then consistent results can still be obtained, even if the incorrect $R_i(\alpha)$ structure is identified, (Dahmen & Ziegler [19]). Finally, any loss of efficiency is reduced as the number of units increases (Dobson et al. [1]).

The most common structures used to model the working correlation matrix are the independent, exchangeable, autoregressive, stationary, nonstationary, unstructured, and fixed correlation structures. The broad range of options available for specifying the correlation structure is another advantage for using the GEE approach. Some of these structures are examined in more detail below.

Independent Structure

The independent structure is the simplest form that the working correlation matrix can take, as it assumes that no correlation actually exists and observations within the series are independent. Because users assume that the responses within each unit are independent of each other, this approach sacrifices one of the benefits of GEE in that it does not account for within-subject correlation (Ballinger [3]). In general, this structure does not make logical sense for longitudinal data, since such data is usually highly correlated. Fitzmaurice [30] shows that using an independent structure for correlated data can lead to large efficiency loss of time-varying covariates. Thus, this structure would not be recommended for variables such as rainfall.

With this structure, the working correlation matrix becomes the identity matrix, $R_i(\alpha) = I$, and the resulting GEE is then called the Independent Estimating Equation (Dahmen & Ziegler [19]). No estimation of α is required, since no correlation is assumed to exist. This structure does not simply produce the algorithm used for a GLM, as it still involves the 'working' correlation matrix, which a GLM does not. For the independent structure, $R_i(\alpha)$ is defined as,

$$R_{u,v} = \begin{cases} 1 & u = v, \\ 0 & \text{otherwise.} \end{cases}$$

In matrix notation this becomes,

$$R_i = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Exchangeable Structure

The exchangeable structure assumes that there is a common correlation within observations. Thus, all of the correlations in $R_i(\alpha)$ are equal (Hedeker [42]). An exchangeable correlation may be used when each pair of observations within a time frame has approximately the same correlation. For the exchangeable structure, $R_i(\alpha)$ is defined as,

$$R_{u,v} = \begin{cases} 1 & u = v, \\ \alpha & \text{otherwise.} \end{cases}$$

In matrix notation this becomes,

$$R_i = \begin{bmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & & \alpha \\ \vdots & & \ddots & \vdots \\ \alpha & \alpha & \dots & 1 \end{bmatrix}.$$

Autoregressive Structure

For data that are correlated within cluster over time, an autoregressive correlation structure is specified to set the within-subject correlations as an exponential function of this lag period, which is determined by the user (Ballinger [3]). The autoregressive structure assumes time dependence for the association between observations and considers each time series to be an AR(m) process. The most difficult task for this structure is determining the correct order of the autoregressive process (Hardin & Hilbe [39]). It is common to choose an AR(1) structure, which is defined as,

$$R_{u,v} = \begin{cases} 1 & u = v, \\ \alpha^{|u-v|} & \text{otherwise.} \end{cases}$$

In matrix notation this becomes,

$$R_i = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ \alpha & 1 & \alpha & & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & & \vdots \\ \vdots & & & \ddots & \vdots \\ \alpha^{n-1} & \dots & & \alpha & 1 \end{bmatrix}.$$

Unstructured structure

The unstructured form of the working correlation matrix is the most general of all of the correlations discussed in this dissertation as no structure is imposed on the correlation matrix. This form requires all $n_i(n_i - 1)/K$ correlations of $R_i(\alpha)$ to be estimated, and thus when there are many time points this structure becomes very computationally burdensome.

An unstructured correlation matrix is used when there is no logical ordering for the observations in the cluster, and is recommended if the number of observations is small in a balanced and complete design (Horton & Lipsitz [45]). This correlation matrix is the most efficient structure, but is only useful when there are relatively few observations as its estimate is not guaranteed to be a positive number and there is often a problem with inverting $R_i(\alpha)$ (Hedeker [42]). For the unstructured structure, $R_i(\alpha)$ is defined as,

$$R_{u,v} = \begin{cases} 1 & u = v, \\ \alpha_{uv} & \text{otherwise.} \end{cases}$$

In matrix notation this becomes,

$$R_i = \begin{bmatrix} 1 & \alpha_{12} & \dots & \alpha_{1n_i} \\ \alpha_{21} & 1 & \dots & \alpha_{2n_i} \\ \vdots & & \ddots & \vdots \\ \alpha_{n_i1} & \alpha_{n_i2} & \dots & 1 \end{bmatrix}.$$

Fixed Correlation

A fixed correlation structure is fixed at some user-defined value and can be imposed if there is some knowledge of the structure of the correlation matrix from another source (Hardin & Hilbe [39]). With this structure, the working correlation is not estimated at each step, but instead takes the correlation as fixed throughout the entire process.

4.4 GEE Estimation

As GEEs can be thought of as a moderation in the GLM to incorporate correlated data, it makes sense that they involve a moderation to the estimating or score equation, U_j, used in GLMs (Section 3.3, Equation (3.8)). GEEs are modified by using the 'working' correlation matrix in the score equations to account for the correlations in the data (Hardin & Hilbe [39]).

To begin, the following terms need to be defined in order to setup the score equations for GEE models:

- The working correlation matrix, $\mathbf{R}(\alpha)$ was already defined in section 4.3.1, with α fully characterising $\mathbf{R}(\alpha)$. Note that $R_i(\alpha)$ is a $n_i \times n_i$ working correlation matrix for the i unit.

- A_i is defined as a $t \times t$ diagonal matrix, with the variance function $V(\mu_{it})$, as the tth diagonal element.

- Finally, a working variance-covariance matrix for $\mathbf{y}_i$, which incorporates the 'working' correlation matrix and thus the correlations of the data is defined as,
$$V_i(\alpha) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}. \tag{4.2}$$
This 'working' covariance matrix will be equal to $\text{cov}(Y_i)$ if $R_i(\alpha)$ is indeed the true correlation matrix for the response variable. It is a transformation of the variance $V(\boldsymbol{\mu}_i)$ term into a matrix form to account for the correlation between observations.

Generalized Estimating Equations Estimator

The generalized estimating equation estimator can now be defined as:

$$U_k(\beta) = \sum_{i=1}^{K} D_i^T [V_i]^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{4.3}$$

where D_i is a matrix of partial derivatives of $\boldsymbol{\mu}_i$ and $\boldsymbol{\beta}_i$ (where $D_{it} = \partial \mu_i / \partial \boldsymbol{\beta}_t$), and V_i is the working variance-covariance matrix of $\mathbf{y}_i$ (Equation (4.2)). This score equation for estimating $\boldsymbol{\beta}$ is the solution to a set of k 'quasi-score' differential equations (Zorn [91]), as Equation (4.3) only depends on the mean and variance of $\mathbf{y}_i$.

4.4.1 Estimation of $\boldsymbol{\beta}$

The ultimate aim of a GEE is to find the most adequate model to represent a given data set by finding values for the unknown $\boldsymbol{\beta}$ parameters. To estimate $\boldsymbol{\beta}$, the GEE estimator (Equation (4.3)) is rearranged to obtain the following (for the derivation of this formula see Appendix A.1),

$$\hat{\beta} = \sum_{i=1}^{K} (D_i^T \hat{V}_i^{-1} D_i^T)^{-1} \sum_{i=1}^{K} (D_i^T \hat{V}_i^{-1} \mathbf{y}_i). \tag{4.4}$$

As GEEs are not a likelihood-based method of estimation, computations based on likelihoods are not possible. Thus, in order to find a solution

for Equation (4.4), estimation may be accomplished either via generalized weighted least-squares, or through an iterative process (Zorn [91]). Essentially, solving the GEE involves the following steps:

1. Specifying,

   - the model parameters of interest and in particular the variable that indicates that the data is correlated;
   - the link function which will 'linearize' the regression equation;
   - the distribution of the dependent variable;
   - the structure of the 'working' correlation .

2. Computing an initial estimate of $\boldsymbol{\beta}$ using GLM methodology; thus assuming that observations are independent, with no correlation existing. This is done using GLM estimation techniques (Section (3.3)).

3. Given the initial estimates of $\boldsymbol{\beta}$, computing the Pearson's residuals,

$$e_{it} = \frac{y_{it} - \mu_{it}}{\sqrt{V(\mu_{it})}}. \tag{4.5}$$

4. An estimation of α, to be used in the working correlation matrix, is then computed using the Pearson's residuals and the assumed structure of R_i specified in step 2 (See Section **??** for the calculation of α for different structures). It should be noted that the number of nuisance parameters and the estimator of α vary depending on the correlation structure chosen. Liang and Zeger [56] introduced several formulas to calculated α. In addition, even though ϕ appears in all of the following formulas for α, it is not needed to obtain a consistent estimate of $\boldsymbol{\beta}$. Different texts use differing methods of calculating α, although most produce very similar values.

5. The working correlation matrix, R_i can now be specified using the α value calculated in step 4 and the assumed structure of R_i.

6. Using A_i, defined in Section 4.4 and $R_i(\alpha)$, defined in step 5, compute an estimate of the covariance V_i for the K units examined,

$$V_i = A_i^{\frac{1}{2}} \hat{R}_i(\alpha) A_i^{\frac{1}{2}}.$$

7. Finally, update $\hat{\boldsymbol{\beta}}$ using the following iteratively formula,

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left\{ \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right\}^{-1} \left\{ \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right\}. \quad (4.6)$$

8. Complete steps 3 to 6 until convergence.

4.4.2 Calculation of α

Independent Structure

When no correlation is assumed to exist, and an independence structure $\mathbf{R}(\alpha) = I$ is chosen, $\alpha = 1$. Thus no calculation of α is required.

Autocorrelation Structure

If an autocorrelation structure is chosen as the appropriate 'working' correlation matrix, then $\alpha = (\alpha_1, ..., \alpha_{n_i - 1})$. An estimator of α_t can then be given as,

$$\hat{\alpha}_t = \phi \sum_{i=1}^{K} \hat{e}_{it} \hat{e}_{i,t+1} / (N - r). \quad (4.7)$$

If the structure is specified specifically as an AR(1), then a common α is estimated as,

$$\hat{\alpha} = \sum_{t=1}^{n_i - 1} \hat{\alpha}_t / (n_i - 1). \quad (4.8)$$

For the AR(1) structure, all R_i will be identical as this is equivalent to a one-dependent model. Other m-dependent structures can be specified; see Hardin & Hilbe [39] for further examples.

Exchangeable

When an exchangeable correlation structure is chosen for $\mathbf{R}(\alpha)$, then α can be estimated as,

$$\hat{\alpha} = \phi \sum_{i=1}^{K} \sum_{t > t'} \hat{e}_{it} \hat{e}_{it'} / \left\{ \sum_{i=1}^{K} \frac{1}{2} n_i (n_i - 1) - r \right\}.$$

4.4.3 Properties of GEEs

Dispersion Parameter, ϕ

The dispersion parameter for a GEE can be estimated by,

$$\hat{\phi} = \frac{1}{N-r} \sum_{i=1}^{K} \sum_{t=1}^{n_i} e_{it}^2, \tag{4.9}$$

where $N = \sum n_i$ and is the total number of observations across all units, r is equal to the number of regression parameters, and e_{it} are the estimated Pearson's residuals (Hardin & Hilbe [39]). Although most software packages use Equation (4.9), some use,

$$\hat{\phi} = \frac{1}{N} \sum_{t=1}^{K} \sum_{t=1}^{n_i} e_{it}^2. \tag{4.10}$$

The advantage of Equation (4.9) over Equation (4.10) is model results for independent correlation exactly match GLM results. Liang & Zeger [56] state that any consistent estimate of ϕ is admissible.

Variance of $\boldsymbol{\beta}$

In order to perform hypothesis tests and construct confidence intervals, it is of interest to obtain standard errors associated with the estimated regression coefficients, $\boldsymbol{\beta}$. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\hat{\boldsymbol{\beta}})$. There are two different ways to calculate the variance of $\hat{\beta}$ within GEE methodology.

The first way is the naive or 'model-based' approach. This approach often underestimates the standard error of $\hat{\beta}$; however it is simple to calculate (Dobson et al. [1]). The second approach is called the robust or 'empirical' estimate, and yields more consistent results even when,

- $V(Y_{ij})$ is not equal to $\phi V(\mu_{ij})$; and
- $R_i(\alpha)$ is misspecified.

The naive approach gives the variance of $\hat{\beta}$ as,

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 \left[\sum_{i=1}^{K} D_i^T \hat{V}_i^{-1} D_i \right]^{-1}.$$

The empirical or robust approach gives the variance of $\hat{\beta}$ as,

$$\text{Var}(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1}, \text{ where}$$

- $M_0 = \sum_{i=1}^{K} D_i^T \hat{V}_i^{-1} D_i$, and
- $M_1 = \sum_{i=1}^{K} D_i^T (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{V}_i^{-1} D_i$.

It should be noted that if $\hat{\sigma}^2 \hat{V}_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T$, then the naive and empirical approaches are identical. This second estimator is often called the 'sandwich' estimator.

The consistency of the variance estimate of $\hat{\boldsymbol{\beta}}$ depends on proper specification of the working correlation structure, unlike the actual estimates of $\hat{\boldsymbol{\beta}}$ which do not. Misspecification of the working correlation structure yields estimates of $\text{Var}(\hat{\beta})$ which do not agree with the naive approach. Thus in practice, the robust estimator is nearly always used, since specification of the correct correlation matrix is difficult to achieve (Zorn [91]). However, if there are less than 20 units or clusters, the naive approach should be used as it gives better estimates for the variance of $\boldsymbol{\beta}$ (Horton & Lipsitz [45]).

4.5 Diagnostics

The main concern of researchers is finding a model that adequately describes the data as simply as possible. However, with GEEs the process of selecting model terms and the appropriate correlation structure is complicated by the correlation within observations. As observations are not independent of each other, the residuals are not independent either, and common likelihood-based methods of model fitting either cannot be used or need to be adjusted.

Although GEEs are increasing in popularity and improved research has refined the estimation of these equations, model selection techniques and diagnostics for GEEs has lagged (Ballinger [3]). There is still no universally accepted test for goodness of fit for GEE models. None of the diagnostic techniques discussed in the next section are available in any of the major statistical packages, meaning that checking the adequacy of a model is quite difficult.

The next section will outline some of the techniques that can be applied to evaluate GEE models. It should be noted that all of the criteria described below are meant only as a guide for when there is no scientific knowledge presented to the researcher. The main techniques discussed are the measures for evaluating the goodness of fit of the model, choosing the best correlation structure, and choosing the best subset of covariates for a given correlation structure. Section 3.6 should be read in conjunction with this Section 4.5 as it gives diagnostics for analysing a GLM model. Although diagnostics for GLMs should not be used with GEE models, they are the best approach available

for testing the link function and appropriateness of the assumed response variable's distribution.

4.5.1 The best correlation structure

In general, decisions about which correlation structure to use should be guided intially by theory. Despite this, choosing $\mathbf{R}(\alpha)$ on the basis of theoretical considerations is sometimes quite difficult to do (Hardin & Hilbe [39]). There is also very little information available about how to chose the best correlation structure. Hardin and Hilbe [39] suggest choosing a correlation structure by initially viewing the following guidelines:

- If the number of observations is small, and the design is balanced and complete, use an unstructured correlation structure.
- If the observations in a cluster are collected over time thereby making the clustered data longitudinal data, then the structure should be chosen to be time-dependent, that is, an autoregressive structure.
- If the observations are simply clustered and not collected over time, then an exchangeable structure is advisable.
- If the number of clusters is small, then the independent model may be the best to use.
- If one or more of the above points applies, then use the 'quasilikelihood under the independence model information criterion' (QIC) to determine the best structure. The QIC is explained below.

The QIC

Pan [66] recommends using a QIC to select the best correlation matrix for cases in which users may be undecided between two structures. The QIC is an extension of Akaike's information criterion (AIC) which uses the quasi-likelihood of a model rather than the log-likelihood. The QIC is called the 'quasi-likelihood under the independence model information criterion', and as its name infers, no matter what R_i is chosen, this criterion assumes independence: that is, $R = I$. It works by comparing the variance and magnitude of the squared deviances for an independence model to models that assume different sorts of correlation (for example, exchangeable, unstructured and autoregressive). It uses the model coefficient estimates and the correlation

in trying to calculate the most appropriate correlation structure. The QIC is defined as,

$$\text{QIC} = -2Q(y; g^{-1}(x\beta_R)) + 2\text{trace}(A_I^{-1}V),$$

where:

- $Q(y; g^{-1}(x\beta_R))$ is the value of the quasi-likelihood, computed using the coefficients from the model with the assumed correlation structure $\mathbf{R}$.
- $\mathbf{A_i}$ is the variance matrix of the independence model.
- $\mathbf{V_i}$ is the sandwich estimate of the variance using the assumed correlation matrix, $\mathbf{R}(\alpha)$.

The QIC can then be used to choose between several correlation structures, with the best structure being the one which has the lowest QIC value.

4.5.2 The best set of covariates to use

There are two methods sometimes employed to find the best subset of covariates to use in a model: the QIC_u, and the marginal R-squared.

The QICu

A similar technique to the QIC can be used to determine the best covariates to use in a given model. The new measure, called the QIC_u, is defined as:

$$\text{QIC}_u = -2Q(y; g^{-1}(x\boldsymbol{\beta}_R)) + 2r,$$

where $Q(y; g^{-1}(x\boldsymbol{\beta}_R))$ is the value of the quasi-likelihood, computed in similar fashion to the QIC and r is the number of coefficients in the model. The best subset of covariates is then the model that has the lowest QIC_u value.

Marginal R-squared

Another technique that can be used to determine which subset of covariates is appropriate is an extension of the R^2 statistic, referred to as 'marginal R-square' (R_m^2)(Ballinger [3]). Zheng [90] introduced this statistic to be used with GEE models that have continuous, binary and counted responses. The test measures improvement in fit between the estimated model and the intercept-only model. It does this by comparing two different quantities. Firstly it compares the predicted values produced from the model with the observed values, and secondly, it compares the squared deviations of the

observations from the mean values for the response variable. Marginal R-square is defined as follows,

$$R_m^2 = 1 - \frac{\sum_{t=1}^{n_i}\sum_{i=1}^{K}(y_{it} - \hat{y}_{it})^2}{\sum_{t=1}^{n_i}\sum_{i=1}^{K}(y_{it} - \bar{y}_{it})^2},$$

where, $\bar{y} = \frac{1}{Kn_i}\sum_{t=1}^{n_i}\sum_{i=1}^{K} y_{it}$ is the marginal mean across all time periods. The marginal R^2 is interpreted as the amount of variance in the response variable explained by the fitted model (Hardin & Hilbe [39]). It has similar properties as the statistic R^2, with the exception that it can take a negative value when the model gives a less accurate prediction than the intercept-only model (Ballinger [3]).

4.5.3 Analysis of residuals

Residuals are extremely important as a final check to see if the selected model adequately fits the data. However, there are limited techniques available to use with GEEs for checking the adequacy of a model using residuals. The raw residuals and Pearsons residuals are the only residuals that have currently been used to uncover any significant departures in the data. The raw residuals (rr) can be found via the simple formula of the observed values minus the predicted values,

$$rr_{it} = y_{it} - \hat{y}_{it}.$$

Visual inspection of the residuals, and a nonparametric test of the randomness of residuals are the two main methods of determining if the model produced adequately represents the given data. Model assessment is predominantly based on graphical visualisations for GEE models.

Raw Residuals

One method of checking the adequacy of the model is to use the raw residuals and a nonparametric test to check the randomness of residuals. Chang [14] suggests using the Wald-Wolfowitz run test to attempt to uncover possible patterns of nonrandomness within the raw residuals. The test begins by coding the raw residuals as '1' if the residual is positive, and a '-1' if the residual is negative. This test then assumes a null hypothesis that the signs of the residuals are distributed in a random sequence. It works by examining the sequence of codes produced and the count of the total number of runs of the two codes.

If n_p is the total number of positive residuals, n_n is the total number of negative residuals, and T indicates the number of observed runs in the

sequence, then the expected value and variance of T are,

$$\begin{aligned} E(T) &= \frac{2n_p n_n}{n_p + n_m} + 1 \\ V(T) &= \frac{2n_p n_n (2n_p n_n - n_p n_n)}{(n_p + n_n)^2 (n_p + n_n - 1)}. \end{aligned}$$

The test statistic for the hypothesis that the signs of the residuals are randomly distributed is,

$$W_Z = \frac{T - E(T)}{\sqrt{V(T)}}, \tag{4.11}$$

which has an approximately standard normal distribution, and thus the corresponding p-value can be determined using z-tables. Extreme values of W_Z indicate that the model does not adequately reflect the underlying structure of the data and may indicate one of many situations, such as,

- the underlying correlation structure has been misspecified;
- the covariates do not adequately represent the data;
- the incorrect distribution has been chosen to represent the response variable.

Graphical assessment

The first step in the graphical assessment of residuals is to include a graph of the raw residuals and then check for the presence of outlier values that may seriously affect the results (Diggle, Liang & Zeger [65]). The model can also be checked to ensure that the raw residuals follow a random pattern and do not form clusters around certain values; this can be further verified by using the Wald-Wolfowitz test described in S ection 4.5.3 (Hardin & Hilbe [39]). The Pearson residuals can be plotted against the linear predictor and the logarithm of the variance function to further assess model adequacy (Hardin & Hilbe [39]). Finally it should be ensured that the raw residuals do not show changes in patterns across the time periods as this could indicate that a different correlation structure is needed.

4.5.4 Summary of diagnostics

Overall there are limited diagnostics available to test the adequacy of GEE models. Most tests that can be performed have to be programmed by the analyst as most standard software do not perform the diagnostic tests described in this section. Also, no methods to assess whether the distribution

chosen to describe the response variable is adequate or which link function is appropriate, have been described in the literature.

The literature only provides a few model criterion measures to assess overall model goodness of fit. The QIC however, is particularly useful for choosing the best correlation structure for a GEE model. Similarly, the QIC_u measure is used for model selection. Standard model criterion measures, such as R^2, are available for GEE models, however it can be difficult to interpret for nonlinear models and experience may be the only method of correctly interpreting the magnitude of R^2 in particular situations. Finally, plots of the raw residuals and Pearson's residuals verse the fitted values, the linear predictor of the variance, can be used to assess a given models adequacy.

4.6 Fitting a GEE to a data set

When fitting a GEE model, a user should specify the requirements specified in Section (4.4.1). Details on how to make decisions required to accurately specify these conditions are discussed in turn below. Note that the first two steps are the same as for GLM; see Section (3.2) for further details.

Step 1 & 2 : Linear predictor and best link function

To model the expected value of the marginal response for the population $\mu_i = E(y_i)$ to a linear combination of the covariates, the user must specify a link transformation function that will allow the response variable to be expressed as a vector of parameter estimates (β) in the form of an additive model (McCullagh & Nelder [62]). The choices available for the link function depend primarily on the distribution specified, and a list of these available with GEE models can be seen in Table (4.1). This table gives the distributions and corresponding link functions currently available with GEE models in most statistical packages. Note that the Tweedie distribution does not appear here; it is not yet available with GEEs in any statistical packages.

Step 3 : Distribution of the response variable

The next step involves specifying the distribution of the outcome variable so that the variance might be calculated as a function of the mean response calculated in step 1 and 2 (Hardin & Hilbe [39]). GEEs, like GLMs, permit the specification of distributions from the exponential family of distributions (Section 3.1.1), including the Normal, inverse Normal, binomial, Poisson, negative binomial, and gamma distributions. This dissertation demonstrates

Table 4.1: The choice of link function will depend on the distribution of the underlying response variable. This table gives some brief directions on the different link functions currently available with GEE models currently available in most software packages (Ballinger [3]).

Distribution	Link Functions	Brief Description
Normal	Identity Link	This fits the same model as the GLM
	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response variable
Binomial	Logit Link	Fits logistic regression models
	Probit Link	Fits cumulative probability functions
	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response variable
Poisson	Log Link	
	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response variable
Negative Binomial	Power Link	Any power transformation
Gamma	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response
Multinomial	Cumulative Logit Link	

the use of the Tweedie distribution with GEEs. Misspecifications of the variance function, and thus the response distribution, can have important consequences and lead to incorrect statistical conclusions (Ballinger [3]).

In fitting a GEE (or any GLM), the user should make every reasonable effort to correctly specify the distribution of the response variable so that the variance can be efficiently calculated as a function of the mean and the regression coefficients can be properly interpreted (Ballinger [3]). It is usual for the user to have some prior knowledge of the distribution of the response variable.

Step 4 : Form of the correlation within the response variable

The final step involves the specification of the form of the correlation of responses within units or nested within a group in the sample. Even though GEE models are generally robust to misspecification of the correlation structure, it is still important that the user takes precautions in specifying this structure. This is because a structure that does not incorporate all of the information on the correlation of measurements within the cluster may result in inefficient estimators (Ballinger [3]).

The form of the correlation structure should be chosen from one of the structures described in Section 4.3.1.

Step 5 : Fitting the model and diagnostics

A GEE model can now be fitted to the data, however this usually takes considerable time and effort. Finally, and often most importantly, the model should be checked to see if it is adequate and justifiable using numerous diagnostic techniques (see Section 4.5).

4.6.1 Cautions regarding GEE

There are a few cautions that users should be aware of when fitting a GEE model. Firstly, users should be cautioned that using the robust approach to estimate the variance of $\boldsymbol{\beta}$ could be highly biased when the number of units or clusters examined is small. Horton and Lipsitz [45] suggest that the GEE robust variance estimate should only be used when there are more than 20 units or clusters, that is, K should be greater than 20. If a data set contains fewer than 20 units, the naive approach to estimating the variance should be used, as it gives better estimates for the variance of $\boldsymbol{\beta}$.

Secondly, although some researchers use the Wald chi-square statistic for model comparisons (Hedeker [42]) and many current statistical packages produce a deviance or chi-square statistic for a GEE model using this technique, such a statistic is only interpretable under certain unrealistic conditions. Thus, it is not recommended for use to test whether all of the variables in the estimate are different from one another and different from zero (Ballinger [3]). It is not interpretable when a user wants to model correlations using the autoregression correlation structure. Furthermore, this statistic is sensitive to large differences in the scale of different independent variables (Ballinger [3]). Thus this type of statistic is not suitable for this dissertation.

4.6.2 Advantages

The major, and most obvious advantage of GEEs is they can be used to model non-Normal, correlated longitudinal data. This makes GEEs an invaluable tool when analysing data that was previously modelled using uncorrelated models. This advantage is further strengthened by the broad range of options available that help specify the correlation between observations through the working correlation matrix. The incorporation of explicit knowledge about within-unit interdependence makes GEEs even more attractive (Zorn [91]). As well as the production of more efficient estimates of regression parameters due to the inclusion of the correlation, GEEs also produce reasonably accurate standard errors and hence, reasonably accurate confidence intervals with the correct coverage rates (Hanley et al. [50]).

Another advantage is that even if an incorrect working correlation matrix is specified, it is still possible to obtain consistent parameter estimates for $\hat{\boldsymbol{\beta}}$ that are asymptotically Normally distributed, provided the mean $\boldsymbol{\mu}_i$ has been correctly specified as a function of all possible explanatory variables $\mathbf{x}_i$ (Dahmen & Ziegler [19]). This is a clear advantage, as understanding the relationship of the correlation is often quite difficult (Zorn [91]). Also the GEE approach has some built-in robustness as it requires no specification of the full likelihood of the response variable's distribution.

As GEEs are an extension of GLMs, they allow the outcome variable to taken on numerous different forms, such as continuous, dichotomous, polychotomous, ordinal, or even count data. This makes their practicality even greater (Zorn [91]). Finally, as GEEs are becoming increasingly popular, more readily available packages have incorporated GEEs into their programs making the computations much easier.

4.6.3 Limitations

GEEs are gaining popularity, however there is some evidence that the use of an incorrect dependence structure within the GEE approach can produce worse results than if using an independent structure to model correlated data (Sutradhar & Das [80]; Crowder [18]). It has been further commented that solutions for $\hat{\alpha}$ may not exist for various reasons, leading to the complete breakdown of the estimation of the regression parameters.

Cologne et, al. [52] also found that when the true correlation structure was quite simple (for example exchangeable), then GEEs were quite efficient. However, when the structure is more difficult, the efficient results are often not obtained if the correlation structure is misspecified. In the case when the correlation structure is complicated, then every effort should be made to

approximate the true correlation structure correctly, as consistent results are not obtained when the correlation structure is misspecified.

Missing Data

One limitation with using GEEs to estimate parameters is that incomplete data sets can complicate the analysis. Often data sets have missing data, such as when rainfall is not recorded on a particular day. If data is missing completely at random (MCAR), consistent results can still be obtained; however the notation and calculations used become more complicated (Horton & Lipsitz [45]). In particular, the estimation of the working correlation matrix becomes quite tedious.

A series of approaches, when data is missing in the dependent variable, has been proposed recently. However, these methods are rarely used as they are extremely difficult and they are not available in accessible form with standard software (Dahmen & Ziegler [19]). Also, the analysis of a data set that contains missing observations produce differing results between differing packages (Horton & Lipsitz [45]). For a complete explanation on how to overcome missing data see Carlin et al. [49]. The three data sets that will be used in this dissertation do not have any missing data and thus this limitation is avoided.

4.6.4 GEE and software

The GEE algorithm has been incorporated into many major statistical software packages, including SAS, STATA, HLM, LIMDEP, GAUSS, SUDANN, R, and S-Plus. However most of the packages are restricted to only modelling a limited number of response outcome distributions (See Table (4.1 for a list of these distributions). Further advancements in the area of GEE software is continuously occurring, and existing software is being constantly revised and updated to include new research. For an overview of software packages offering GEE methodology see Zorn [91] and Horton and Lipstiz [45].

4.6.5 Summary

This section has described the GEE approach for modelling longitudinal and correlated data. This approach has several features which makes it particularly useful and popular. Because it is a generalisation of GLM, many types of dependent variables can be accommodated within the GEE family of models. Also, the selection of the variance-covariance matrix is not as critical as with other models because GEEs provide standard errors that are robust

to misspecification of the variance-covariance matrix. This is an attractive feature, especially for situations where the scientific interest is in estimation and inference of the regression parameters and not of the variance-covariance structure. The converse of this is that if there is scientific interest in the variance-covariance structure of the longitudinal data; then GEEs are not appropriate (at least in its GEE1 implementation).

Liang & Zeger [56] applied the name 'GEE' to emphasise the nature of the generalisation of the original estimating equation due to the focus on the marginal distribution. These models do not start with a probability-based model, nor a likelihood. There is an implied quasilikelihood form to the GEE model which may or may not coincide to a probability-based model.

The GEE model was extended assuming a correlation structure that was estimated by combining information across panels. The ancillary parameter (α) was estimated to get a working correlation matrix. By applying the correlation matrix to each unit, the $\boldsymbol{\beta}$ regression coefficients can be estimated. Thus, the focus is on the marginal distribution, where the units are summed together after taking into account the correlation.

The remainder of this dissertation will focus on creating a program in R that will model data which has an assumed Tweedie distribution, using a GEE approach. This is a new approach to modelling data using a GEE model, and the results obtained will be of practical use in the future.

Chapter 5

Data and Preliminaries

To demonstrate the practicality of using GEEs with the Tweedie distribution, three Australian rainfall data sets are investigated. A rainfall data set from Emerald is used to develop appropriate codes for constructing a GEE model for rainfall. This code is further developed to allow the simultaneous modelling of more than one rainfall location, and Toowoomba and Gatton rainfall data sets are used to demonstrate this. Due to the complexity of some of the techniques involved in GEE models, only monthly rainfall is examined in all three data sets. While modelling daily data offers clear benefits, such as being able to analyse the number of wet days in the month and the rainfall amounts when wet, as well as the provision of a more detailed understanding of many different aspects of rainfall processes, only monthly data is used. This is due to the high level of noise present in daily data and the size of the data set when such data is used. Working with the monthly timescale therefore filters out some of the noise and allows for smaller data sets to be used. Although using daily data presents several advantages, interesting outcomes and different properties of rainfall can still be examined when monthly data is used. Future research can build on this study of monthly data by examining daily data.

This chapter examines the three nominated rainfall data sets, and a preliminary analysis is conducted to highlight any problems that may be faced when creating a rainfall model. Relevant covariates which may be used to adequately represent the variability of rainfall are also discussed.

5.1 Covariates and Factors

Various climatic and time covariates were considered when modelling the rainfall data, and these covered numerous different sources of the variability

of rainfall. The variables that were collected and used to model monthly rainfall for both the single and multiple site models included: month; year; monthly southern oscillation index; monthly southern oscillation index phases; sine and cosine terms; and location predictors.

5.1.1 Southern Oscillation Index

The Southern Oscillation Index (SOI) is the computed standardized difference between Darwin (Australia) and Tahiti's air pressure, multiplied by a factor of 10 (Troup [82]). Records of the monthly average SOI have been collected since January 1879, with any missing values being computed by interpolation. Relationships between the SOI and rainfall have been extensively explored since the early century and numerous authors have shown its relationship to Australian rainfall (Stone, Hammer & Marcussen [67]). Despite the depth of research in this relationship, other authors claim that the SOI does not provide a strong predictor of precipitation occurrence (Hyndman [47]). Furthermore it is proposed that the SOI values prior to 1935 should be used with caution, as there are questions regarding the consistency and quality of the Tahiti pressure values prior to this year. However, as the SOI is used as a predictor of rainfall in current meteorological practices, it is considered as a covariate in this dissertation. Its use, though, is approached with caution.

Further research into the SOI has also found that an index which classifies seasons into 5 phases depending on the value and rate of change in the SOI would be useful when modelling rainfall. Stone and Auliciems [78] used a principal components analysis and cluster analysis to group all sequential two-month pairs of the SOI into five groups called the SOI phases. The SOI phases are recorded monthly, indicating which phase each month appears to be in. Generally, the use of SOI phases to calculate future seasonal rainfall probabilities gives a more accurate result than using SOI averages. The five phases can be stated generally in the following terms (Dunn & Lennox [24]),

- Phase 1 - termed 'consistently negative', indicates that the SOI values for the two previous months are both negative;
- Phase 2 - termed 'consistently positive', indicates that the SOI values for the two previous months are both positive;
- Phase 3 - termed 'rapidly falling', indicates a marked decrease in the SOI from the previous month to the current month;
- Phase 4 - termed 'rapidly rising', indicates a marked increase in the SOI from the previous month to the current month;

- Phase 5 - termed 'consistently near zero', indicates that both the SOI values for the previous two months are close to zero.

5.1.2 Time and seasonal predictors

For each data set examined in this dissertation, two time predictors were given: year and month. The year predictor ranges from 1889 until 2001 for the Emerald data set and from 1980 until 2001 for both Toowooomba and Gatton. The month predictor has 12 factors: one for each month. Two further time predictors were also created to use in the modelling process: season; and a 'wet-season' factor.

The four seasons (Summer, Autumn, Winter and Spring), together with the 12 months, can be used to take into account the seasonal variation of rainfall. The four seasons used in this dissertation are the Southern hemisphere seasons. Although the variable 'season' seems to have a relationship with the amount of rainfall per month, when examining the monthly rainfall amounts for Emerald, it was noticed that some months had similar rainfall distributions and these were not confined to the four seasons. Dunn & Lennox [24] suggest the fitting of a seasonality term that represents the three distinct rainfall periods: a 'Dry' period for the months April to September; a 'Wet' period for the months December and January; and a 'Transitional' period for February, March, October and November. These three distinct rainfall periods are called the 'wet-season' factor. Figure 5.4 presents a visual explanation of this factor for Emerald.

Toowoomba and Gatton did not have such a distinct classification of this 'wet-season' factor, with the four seasons giving a better representation of the seasonal variation of rainfall for these locations. Thus the 'wet-season' factor is not used as a covariate for Toowoomba and Gatton.

Sine and Cosine Terms

Chandler and Wheater [12] found that both a sine and cosine wave could be included as possible covariates to represent season variation of rainfall. Various forms of these waves were introduced in this dissertation to cover a variety of possible cyclical patterns of various lengths in the model, and attempt to include information about the cyclical nature of rainfall from one period to another.

The sine and cosine waves used by Chandler and Wheater [12] to represent the season variation of rainfall are as follows (these are known as annual frequency sine and cosine terms (Dunn & Lennox [24]),

- Sine term: $\sin\left(\frac{2\pi}{12} \times \text{ month}\right)$;

- Cosine term: cos ($\frac{2\pi}{12}\times$ month).

Dunn & Lennox [24] also successfully used the following alterations to the sine and cosine terms to model the seasonal variation of rainfall (these are known as six-monthly frequency sine and cosine terms),

- Sine term: sin ($\frac{4\pi}{12}\times$ month);
- Cosine term: cos ($\frac{4\pi}{12}\times$ month).

The use of either representation of the sine and cosine terms was a possibility for this dissertation, and both were employed. Other slight variations may also prove to be worthwhile in future research.

5.1.3 Temporal dependence

It has been suggested that several previous month's rainfall amounts should be included in a model for rainfall to account for any temporal dependencies that may occur from one month to another (Dunn & Lennox [24]). However this was unnecessary in this dissertation as one of the main purposes of GEE models is to account for this correlation without using extra covariates. Although despite this, based on recommendations by Chandler and Wheater [12], indicators for the previous 12 months and persistent indicators for the previous two or three months are examined. An 'indicator' specifies that any recorded rainfall is coded as one, and no rainfall is coded as zero for a given period. A 'persistent indicator' is one which specifies that rain has occurred on both of the last two, or all of the last three months respectively.

5.1.4 Location covariates

Three location parameters: longitude, latitude, and altitude, are all examined when implementing the multiple site model for Toowoomba and Gatton. Table 5.1 gives the values of these three variables for the two different locations.

5.1.5 Interactions

It is common for climatic variables to interact with one another, meaning that the effect of one predictor may depend on the values of others (Chandler & Wheater [12]). For example, the effect of the SOI on rainfall may be of a differing intensity in different seasons or months. Interaction terms incorporate these relationships into a model by adding an extra predictor to

Table 5.1: The latitude, longitude and altitude of Toowoomba and Gatton. All values are in decimal form and the altitude measurements were taken from the Queensland Department of Primary Industries Research Station.

Location	Latitude (Degrees)	Longitude (Degrees)	Altitude (Metres)
Toowoomba	−27.55	151.95	674.9
Gatton	−27.58	152.28	93

the model. The value of this extra predictor is the product of the interacting predictors. Only those terms that have practical significance and are of second-order (meaning that only two predictors are used in the interaction) are examined in this dissertation.

5.2 Emerald Rainfall Data

The first data set that is used to determine a possible rainfall model for Emerald was collected at the Emerald Post Office. Figure 5.1 shows the location of Emerald within Australia. This data was obtained by the Queensland Department of Primary Industries. At the time of analysis, data that was available was from January 1889 to December 2001.

5.2.1 Model validation

Model validation is important in the model building process and is defined as the confirmation that a given model acquires a satisfactory level of accuracy, consistent with the intended application of the model (Hicks & Earl [44]). There are various validation techniques and tests used in model validation. Diagnostic testing is included as a validation technique and numerous tests will be performed in this dissertation. Another validation technique is historical data validation: when historical data exists, part of the data is used to build the model and the remaining data is kept to test the produced model (Sargent [74]). Diagnostic testing and historical model validation are used in this dissertation.

To enable model validation for the Emerald rainfall data, a portion of the data is used to estimate the rainfall model for Emerald and the rest of the data is kept for validation of the final model. Data from January 1889 to December 1992 was used for model estimation, which constitutes 91% of

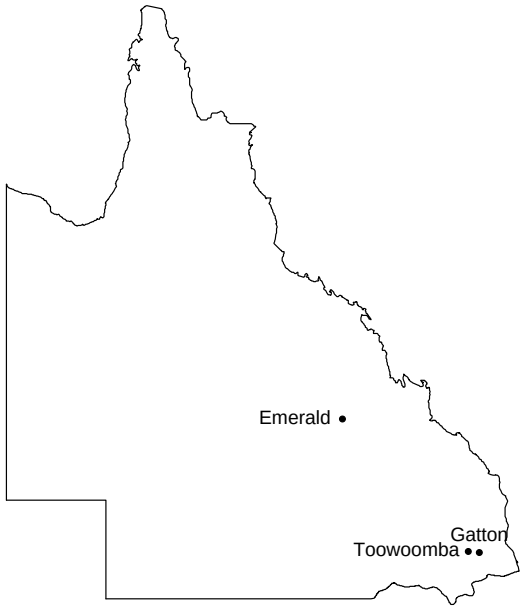

Figure 5.1: The location of Emerald, Toowoomba and Gatton in Queensland (Australia).

the data, the remainder was kept for validation purposes. The proportion of data kept for validation purposes is debatable (Dunn & Lennox [24]) and as the principal intention of the model validation in this dissertation is to demonstrate understanding, the proportions used for model estimation and for model validation is incidental. The remainder of this section examines the entire Emerald rainfall data from 1889 to 2001, to determine the suitability of using this data set for the development of a GEE model.

5.2.2 Emerald data

The Emerald data consists of 1356 monthly recordings in which 102 (7.5%) months were recorded as 'dry' (no rainfall recorded in that month). As there were some months where the rainfall was recorded as zero, a model that combines the monthly rainfall occurrence and rainfall amounts is necessary. A histogram of Emerald monthly rainfall amounts (Figure 5.2) shows that the rainfall amounts are right skewed, with the majority of months experiencing less than $100mm$ of rain. No obvious outliers can be seen. The mean amount of rain recorded in a month was $53.18mm$, whereas the median amount was $33.7mm$, which supports the idea that the Emerald rainfall data is skewed. The rainfall amounts are quite spread out, ranging from $0mm$ of rainfall to $556.3mm$ of rainfall in a month. A summary of the Emerald rainfall data can be seen in Table 5.2.

A plot of the individual rainfall amounts for each month from 1889 to 2001 can be seen in Figure 5.3. This graph shows that three months recorded a high rainfall amount: January 1918; February 1954; and January 1974. With the removal of these high values, the mean and standard deviation do not change significantly. Thus, these three months were not excluded from the analysis. No trace values were recorded for the Emerald rainfall data.

Figure 5.4 and Table 5.3 provide a statistical summary of the rainfall amounts for each month. From these illustrations it can be seen that January is the wettest month, recording an average of $105.24mm$ of rainfall, and August is the driest month, recording an average of $21.59mm$. Figure 5.5 shows the rainfall amounts per season, with the Summer months recording the highest average rainfall of $96.81mm$ per month and the Winter months recording the lowest average of $27.19mm$ per month.

There were 1254 months (92.5%) which can be classified as 'wet', which means that these months experienced a rainfall amount of greater than $0mm$. The distribution of these 'wet' months is also right skewed (Figure 5.6), with a mean of $57.51mm$ recorded per month and a median of $38.15mm$. No outliers are evident. Table 5.3 summarizes the statistics for the 'wet' months.

Analysis of the Emerald data indicates that the use of the Tweedie dis-

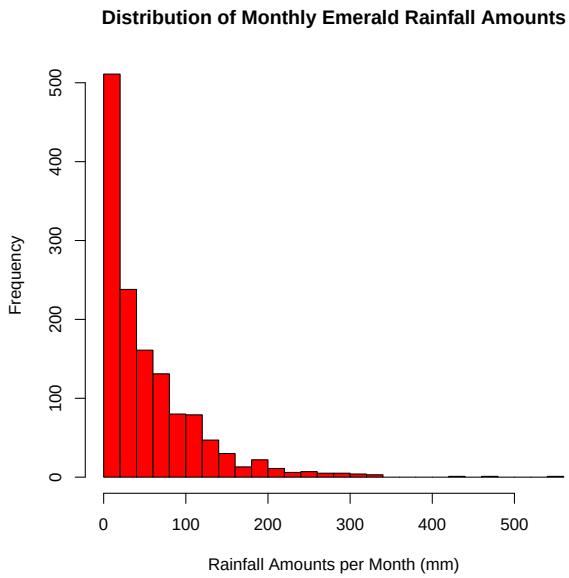

Figure 5.2: Emerald's monthly rainfall amounts for all months.

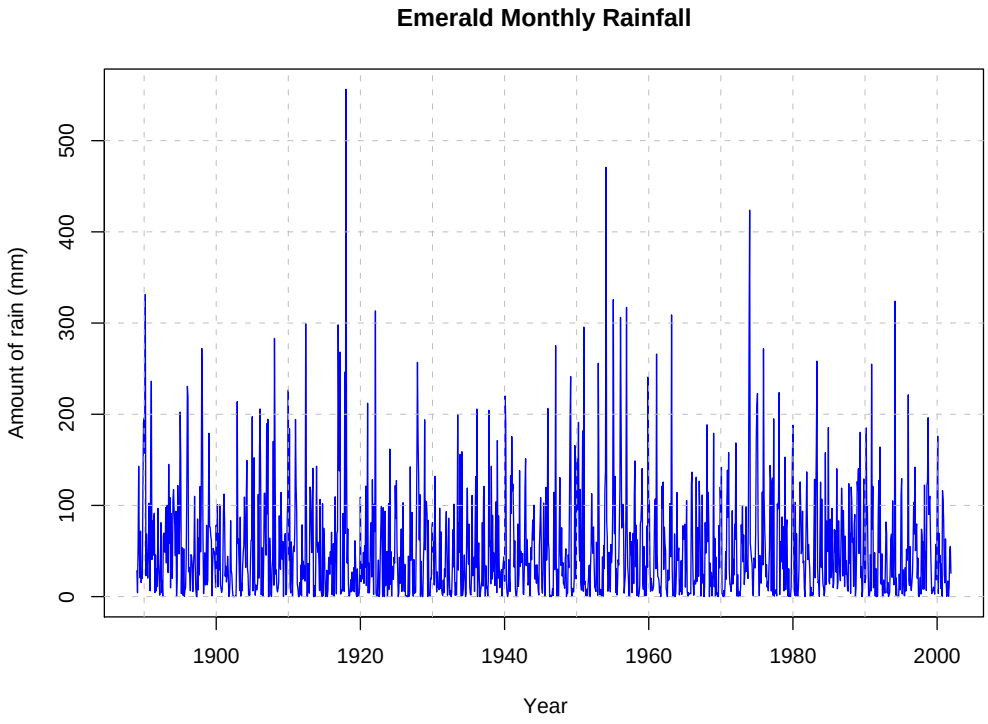

Figure 5.3: Each individual month's rainfall amounts for Emerald.

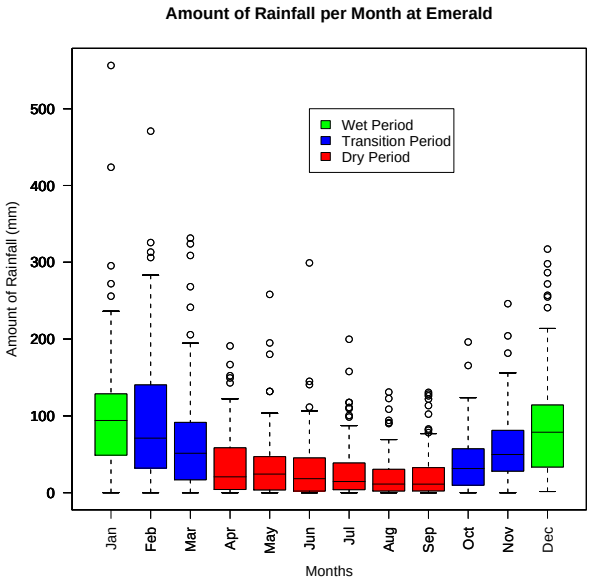

Figure 5.4: Rainfall amount per month for Emerald. The plot also shows the months included in each of the 'wet-season' factors.

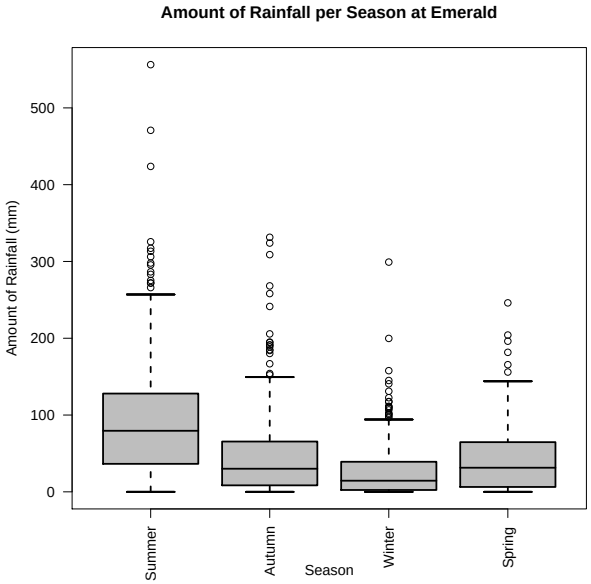

Figure 5.5: Rainfall amount per season for Emerald.

Table 5.2: A statistical summary of the monthly rainfall data for Emerald from 1889 to 2001. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$ (all months) and monthly rainfall amounts $> 0mm$ (rain only months). All measurements are recorded in millimeters.

Statistic	All Months	Rain months only
Minimum	0	0.2
Maximum	556.3	556.3
Mean	53.18	57.51
Median	33.7	38.15
Standard Deviation	61.42	61.89
IQR	68.33	67.10
Shape	Right skewed	right skewed

tribution to model rainfall is reasonable. This can be concluded because the data has a discrete component when rain is $0mm$ for a month, and a continuous component when the rainfall amount recorded is greater than $0mm$. This preliminary analysis has also highlighted that some values may affect the modelling fitting process, and also outlines some of the patterns occurring with this data set.

5.2.3 Comparison of validation and estimation sets

As mentioned in Section 5.2.1, the Emerald data set was divided into two different sections: a validation set, comprising of 9% of the data; and a estimation set, comprising of 91% of the data. This section briefly examines the two sets to determine their suitability for use in model validation. The validation set ($n = 108$) has a slightly lower mean monthly rainfall amount than the estimation set ($n = 1236$): $43.43mm$ compared with $53.79mm$. The percentage of months that do not experience any rainfall is similar for both the validation and estimation data sets: the validation set has 6 observations which were recorded as 'dry' (5.6%); and the estimation set has 96 observations (7.8%). Properties of the two sets of data can be seen in Table 5.4.

Table 5.3: A statistical summary of the monthly rainfall data for Emerald for each month. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$. All measurements are recorded in millimeters.

Month	Mean	Median	Standard Deviation	IQR
January	105.24	94.0	85.43	79.8
February	96.30	71.0	84.75	108.3
March	68.60	51.3	69.12	74.7
April	36.50	20.6	40.84	54.3
May	34.81	24.2	42.84	43.4
June	32.23	18.2	41.66	43.1
July	27.75	14.5	35.74	34.7
August	21.59	11.2	27.50	28.20
September	24.46	11.1	31.70	30.3
October	41.36	31.4	39.19	47.4
November	60.34	49.7	45.91	53.1
December	88.90	78.7	67.67	80.7

5.3 Multiple Site Data

To demonstrate the functionality of GEE models and show that they can not only simultaneously model rainfall occurrence and rainfall amount, but can also simultaneously model more than one rainfall data set, two sites are modelled together: Toowoomba; and Gatton. These two sites are located close together (See Figure 5.1) and thus simultaneous modelling of these two sites is practical and efficient. Due to the complexity and time constraints involved when creating a multiple site model, data is only be examined from January 1980 to December 2001 for both Toowoomba and Gatton. Also for the same reasons the data is not portioned into two set for estimation and validation. Thus the only validation that is done on the multi-site models is diagnostic tests. Validation on the multiple site model is important and the development of this technique is a possibility for future research.

5.3.1 Toowoomba Rainfall Data

The first data set that is analysed in the multiple site model of rainfall is Toowooomba. The location of this site in Queensland, Australia can be seen

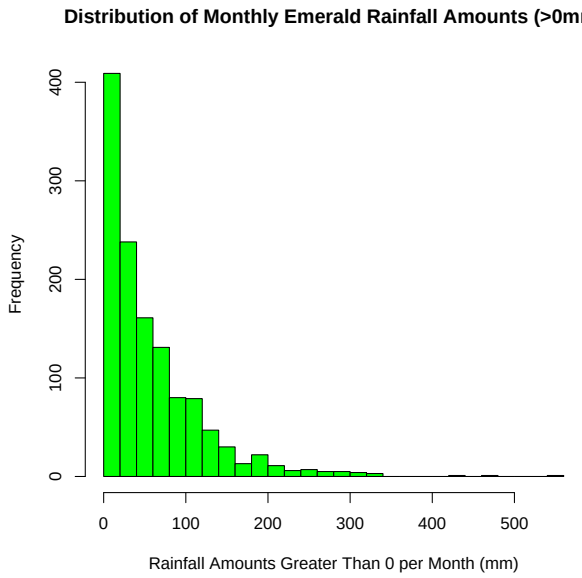

Figure 5.6: Emerald's monthly rainfall amounts for the months which recorded a rainfall amount greater than $0mm$.

in Figure 5.1. This data was obtained from the Queensland Department of Primary Industries.

The Toowoomba data consists of 264 monthly recordings in which 1 (0.38%) month was recorded as 'dry' (no rainfall recorded in that month). As there was a month where the rainfall was recorded as zero, a model that combines the monthly rainfall occurrence and rainfall amounts is necessary. A histogram of Toowoomba's monthly rainfall amounts (Figure 5.7) indicates that the distribution of rainfall is right skewed, with the majority of months experiencing less than $100mm$ of rain. It could be argued that there appears to be three outliers, as seen in Figure 5.7. However, it could also be counter-argued that these values are a continuation of the tail of the distribution of rainfall, and therefore contribute to the overall characteristics of rainfall. The outliers, highlighted in Table 5.5 were examined carefully to investigate their contribution and effect on the overall model produced.

The mean amount of rain recorded in Toowoomba was $76.62mm$ per month, whereas the median amount was $55mm$. These results support the idea that the Toowoomba rainfall data is skewed. The rainfall amounts range from $0mm$ to $519.6mm$ of rain in a month. A summary of the Toowoomba rainfall data is produced in Table 5.6.

Figure 5.8 is a plot of the individual rainfall amounts for each month from

Table 5.4: Summary statistics of monthly rainfall data for the estimation and validation data sets. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$ (all months) and monthly rainfall amounts $> 0mm$ (rain only months). All measurements are recorded in millimeters.

	All data		Estimation		Validation	
Statistic	All months	Rain months	All months	Rain months	All months	Rain months
Minimum	0	0.2	0	0.2	0	0.2
Maximum	556.3	556.3	556.3	556.3	324.0	324.0
Mean	53.18	57.51	53.79	58.31	43.32	47.99
Median	33.70	38.15	34.50	38.85	26.40	30.45
Standard Deviation	61.42	61.89	62.11	62.60	53.03	53.38
IQR	68.33	67.10	69.40	67.63	56.98	58.70
n	1356	1254	1236	1140	108	102

Table 5.5: The three most extreme rainfall amount values in the Toowoomba rainfall data, together with their corresponding dates.

Month	Year	Rainfall Amount (mm)
February	1981	421.4
April	1988	421.2
May	1996	519.6

1980 to 2001. This graph also highlights the three months which recorded a high rainfall amount: February 1981; April 1988; and May 1996. With the removal of these high values the mean and standard deviation do not change significantly. No trace values were recorded for the Toowoomba rainfall data.

A statistical summary of the rainfall amounts for each month is produced in Figure 5.9 and Table 5.7. These illustrations show that December is the wettest month, recording an average of $130.22mm$ of rainfall, and August is the driest month, recording an average of $30.49mm$. Figure 5.10 shows the rainfall amounts per season. It shows that the Summer months recorded the highest average rainfall of $120.25mm$ per month and the Winter months recorded the lowest average of $41.10mm$ per month.

Analysis of the Toowoomba data indicates that the use of the Tweedie

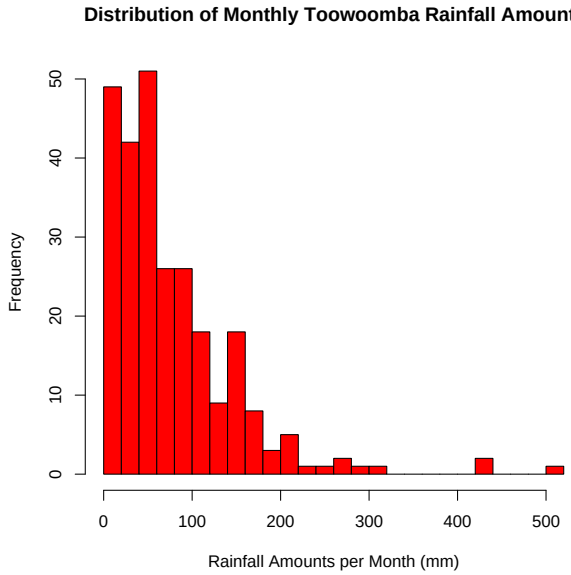

Figure 5.7: Toowoomba's monthly rainfall amounts for all months.

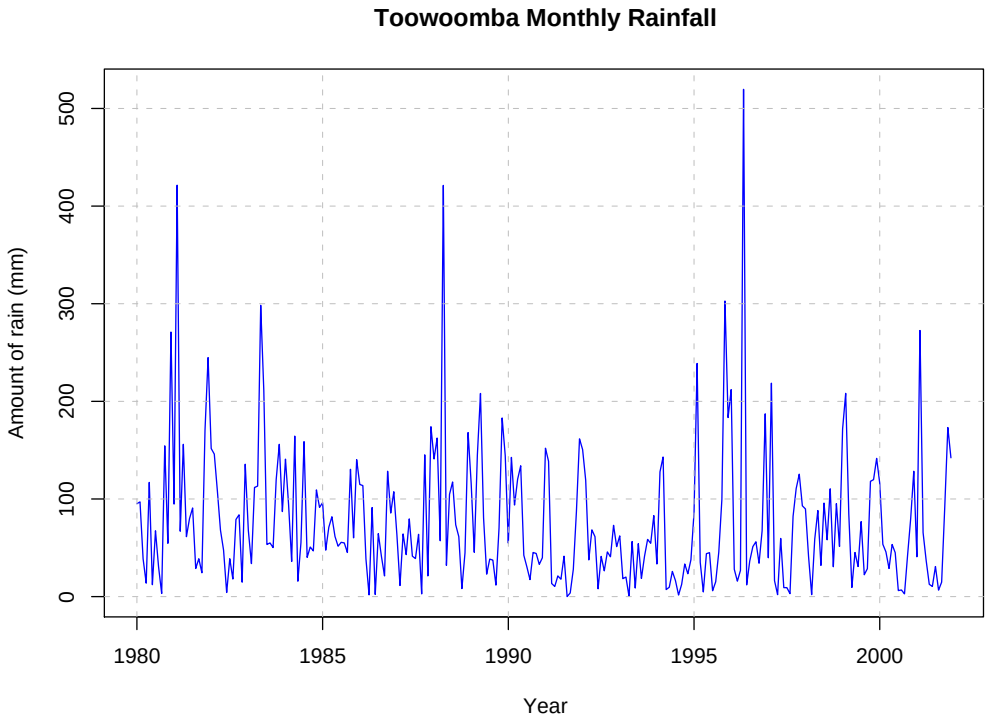

Figure 5.8: Each individual month's rainfall amount for Toowoomba.

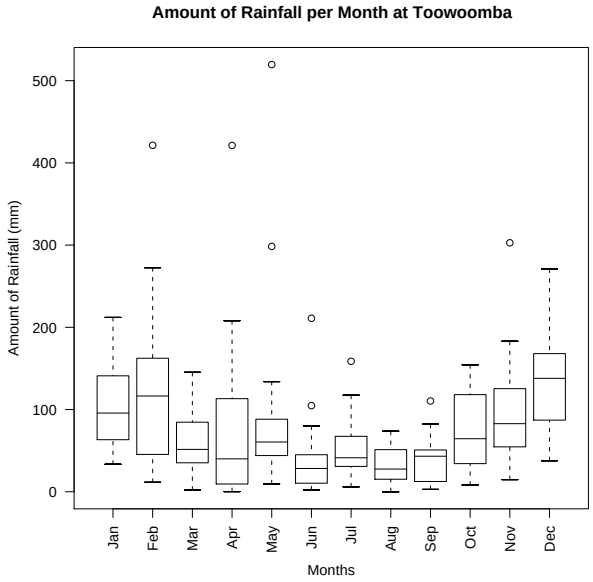

Figure 5.9: Rainfall amounts per month for Toowoomba.

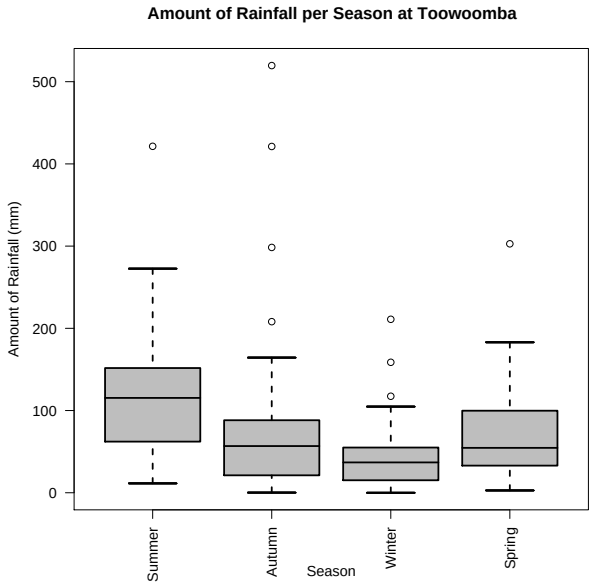

Figure 5.10: Rainfall amounts per season for Toowoomba.

Table 5.6: A statistical summary of the monthly rainfall data for Toowoomba from 1980 to 2001. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$ (all months) and monthly rainfall amounts $> 0mm$ (rain only months). All measurements are recorded in millimeters.

Statistic	All Months	Rain months only
Minimum	0	0
Maximum	519.6	519.6
Mean	76.62	76.914
Median	55	55
Standard Deviation	71.83	71.81
IQR	79.03	79.15
Shape	Right skewed	right skewed

distribution to model rainfall is reasonable. This can be concluded because the data has a discrete component when rain is $0mm$ for a month, and a continuous component when the rainfall amount recorded is greater than $0mm$. This preliminary analysis has also highlighted that some values may affect the modelling fitting process, and also outlines some of the patterns occurring with this data set.

5.3.2 Gatton Rainfall Data

The second data set in the multiple site model of rainfall is from Gatton. The location of this site in Queensland, Australia can be seen in Figure 5.1. This data was obtained from the Queensland Department of Primary Industries. As stated earlier, the data from Gatton is only examined from January 1980 to December 2001.

The Gatton data consists of 264 monthly recordings in which 9 months (3.51%) were recorded as 'dry' (no rainfall recorded in that month). As there were months where the rainfall was recorded as zero, a model that combines the monthly rainfall occurrence and rainfall amounts is necessary. A histogram of Gatton's monthly rainfall amounts (Figure 5.11) indicates that the distribution of rainfall is right skewed, with the majority of months experiencing less than $100mm$ of rain. It could be argued that there appears to be two outliers, as seen on Figure 5.11). However it could also be counter-argued that these values are a continuation of the tail of the distribution of rainfall, and therefore contribute to the overall characteristics of rainfall. The outliers, highlighted in Table 5.8, correspond with two of the three outliers

Table 5.7: A statistical summary of the monthly rainfall data for Toowoomba for each month. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$. All measurements are recorded in millimeters.

Month	Mean	Median	Standard Deviation	IQR
January	104.13	95.70	47.37	76.53
February	126.39	116.45	99.74	112.35
March	59.60	51.50	40.78	45.80
April	74.79	40.00	97.83	95.70
May	90.48	60.50	113.10	42.05
June	39.67	28.30	46.13	34.20
July	53.13	41.30	37.36	34.48
August	30.49	27.60	21.89	32.95
September	38.80	43.20	28.78	37.58
October	74.74	64.50	44.78	80.40
November	97.04	82.80	68.16	69.40
December	130.22	137.95	61.97	78.23

observed in the Toowoomba data. These values are examined carefully to investigate their contribution and effect on the overall model produced.

The mean amount of rain recorded in Gatton was $62.42mm$ per month, whereas the median amount was $43.6mm$. These results support the idea that the Gatton rainfall data is skewed. These results are also slightly lower than the averages recorded in Toowoomba, with the rainfall amounts ranging from $0mm$ to $449.3mm$ of rain in a month. A summary of the Gatton rainfall data is provided in Table 5.9.

Table 5.8: The two most extreme rainfall amount values in the Gatton rainfall data, together with their corresponding dates.

Month	Year	Rainfall (mm)
April	1988	393.4
May	1996	449.3

Figure 5.12 is a plot of the individual rainfall amounts for each month from 1980 to 2001. This graph also highlights the two months which recorded

a high rainfall amount: April 1988; and May 1996. With the removal of these high values, the mean and standard deviation do not change significantly. No trace values were recorded for the Gatton rainfall data.

Table 5.9: A summary of the monthly rainfall data for Gatton from 1980 to 2001. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$ (all months) and monthly rainfall amounts $> 0mm$ (rain only months). All measurements are recorded in millimeters.

Statistic	All Months	Rain months only
Minimum	0	0
Maximum	449.3	449.3
Mean	62.42	64.62
Median	43.6	47.0
Standard Deviation	61.56	61.49
IQR	67.95	65.30
Shape	Right skewed	right skewed

A statistical summary of the rainfall amounts for each month is provided in Figure 5.13 and Table 5.10. These illustrations show that December is the wettest month, recording an average of $106.24mm$ of rainfall, and August is the driest month, recording an average of $22.59mm$. Figure 5.14 shows the rainfall amounts per season. It shows that the Summer months recorded the highest average rainfall of $99.06mm$ per month and the Winter months recorded the lowest average of $30.50mm$ per month.

Analysis of the Gatton data indicates that the use of the Tweedie distribution to model rainfall is reasonable. This can be concluded because the data has a discrete component when rain is $0mm$ for a month, and a continuous component when the rainfall amount recorded is greater than $0mm$. This preliminary analysis has also highlighted that some values may affect the modelling fitting process, and also outlines some of the patterns occurring with this data set.

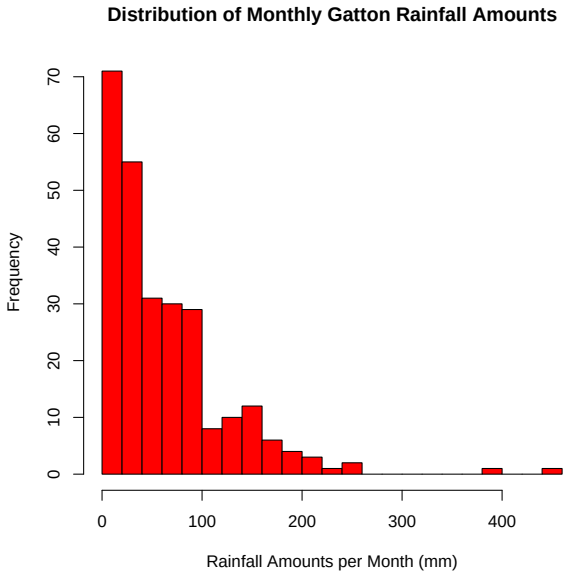

Figure 5.11: Gatton's monthly rainfall amounts for all months.

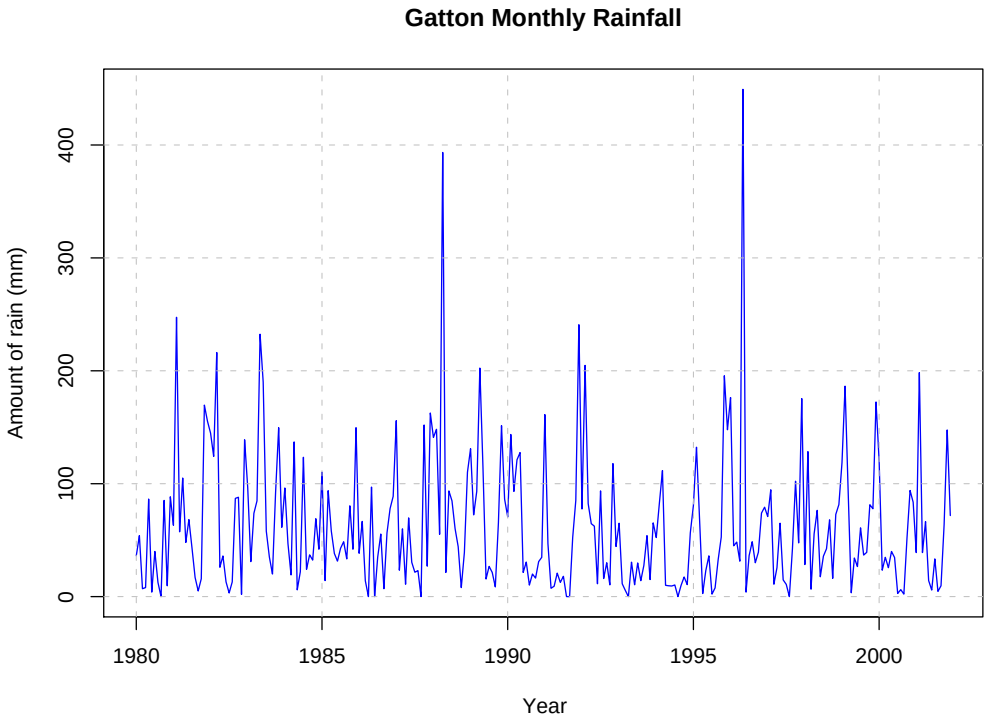

Figure 5.12: Each individual month's rainfall amount for Gatton.

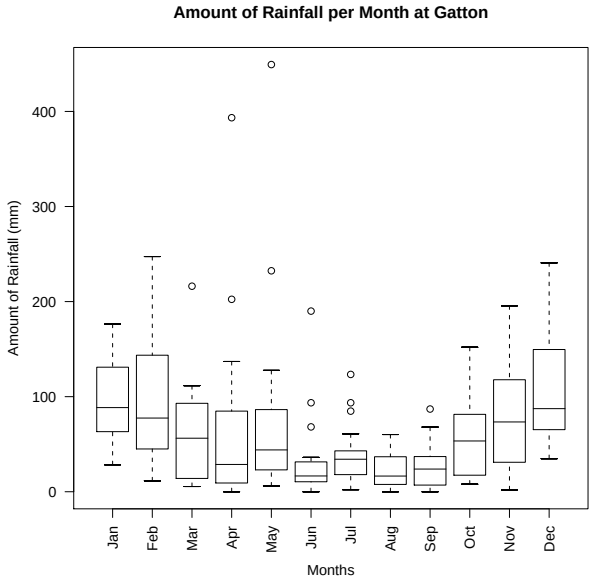

Figure 5.13: Rainfall amount per month for Gatton.

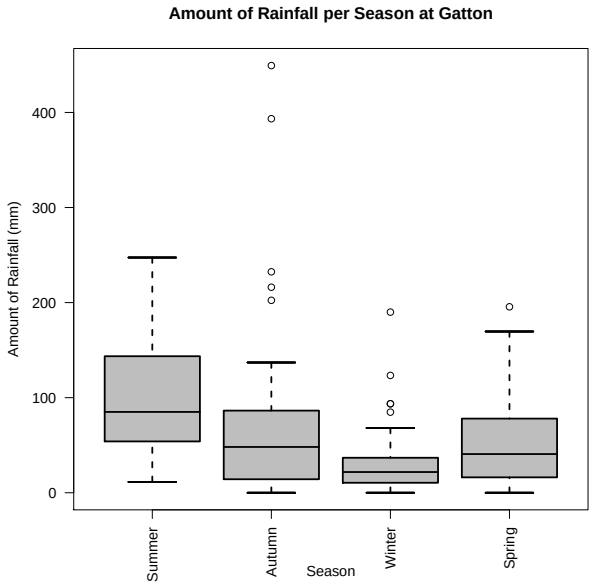

Figure 5.14: Rainfall amount per season for Gatton.

Table 5.10: A statistical summary of the monthly rainfall data for Gatton for each month. Statistics have been recorded for monthly rainfall amounts $\geq 0mm$. All measurements are recorded in millimeters.

Month	Mean	Median	Standard Deviation	IQR
January	94.29	88.50	44.53	64.70
February	96.65	77.50	69.33	95.23
March	58.65	56.30	49.30	75.05
April	65.52	28.70	90.34	70.80
May	77.43	44.00	97.53	59.03
June	30.59	16.60	41.68	20.43
July	38.32	34.20	30.66	24.05
August	22.59	16.50	18.81	28.03
September	25.18	23.85	22.68	28.70
October	56.06	53.40	36.14	60.03
November	77.56	73.40	56.52	78.85
December	106.24	87.40	53.82	82.83

Chapter 6

GEEs and the Tweedie distribution

As demonstrated in Chapter 5, monthly rainfall at Emerald, Toowoomba and Gatton display both discrete and continuous amounts of rainfall in any given month. Thus, a rainfall model that not only simultaneously models rainfall occurrence and rainfall amount, but also takes into account the dependency of rainfall and is able to model several sites concurrently, will be useful, practical and completely novel. Using GEEs to create a model when the response variable is from the power-variance (Tweedie) family of distributions will be an innovative approach to modelling rainfall. The difficulty in using GEEs in this situation is that there is no software available to compute the appropriate parameters in the model.

6.1 The Tweedie distribution, GEEs and rainfall

It has already been shown by Dunn and Lennox [24] that using Tweedie generalized linear models to model rainfall is an appropriate and practical technique for this type of data. However, by expanding the work completed by Dunn and Lennox [24] to incorporate the important fact that rainfall data is dependent, other difficulties in modelling rainfall can be addressed.

The other difficulties in modelling rainfall, as discussed in Chapter 2, are trace values, temporal dependence, and spatial dependence. No trace values were recorded for the three rainfall sites analysed in this dissertation, and GEEs allow the other two difficulties to be dealt with as they are specifically designed to model correlated data and to allow for more than one unit to be examined simultaneously.

Due to the difficulties that can be resolved by using GEEs to model rainfall, GEEs are not only an innovative way to model rainfall, but are also a practical and useful approach which may have an important use in the future of rainfall modelling.

6.2 Implementing GEEs and rainfall

Many different software packages are available to model a given data set using GEEs (see Section 4.6.4), however no packages have yet been produced that use GEEs for a Tweedie distribution. In order to create a model for rainfall data, using GEEs with a Tweedie distribution, the program R has been used to find an estimation of the $\boldsymbol{\beta}$ values. This software package, unlike other packages, enables the fitting of GLMs using the Tweedie family of distributions and this is needed to fit a GEE with a Tweedie distribution. The following section outlines the steps involved in creating such a program. In order to create this program, the steps used to fit a GEE to data (Section 4.6) are used.

Step 1 : Specification of parameters

Before a model can be created, several parameters need to be specified. Firstly, the variable that is correlated needs to be identified and the other variables of interest need to be specified. Rainfall data has been documented as dependent (Chandler & Wheater [12]; Buishand [8]; Horton & Lipsitz [46]), and thus this variable will be the response variable. The list of possible covariates are examined in Chapter 5. Secondly, the link function that will 'linearise' the regression equation needs to be specified. When using a Tweedie distribution, the most common link function to use is the log link function, and this is what is used in this dissertation. Thirdly, the distribution of the dependent variable needs to be ascertained. Previous investigations (Dunn [27] and Dunn & Lennox [24]) shows that a Tweedie family of distributions is a novel way of modelling rainfall data. As demonstrated in Section 3.5, the class of Tweedie distributions when $1 < p < 2$, is called a Poisson-gamma distributions. This Poisson-gamma distribution can be used to model rainfall (Section 3.5.2). Finally, an initial structure of the 'working' correlation matrix needs to be decided. For data that are correlated over time, an autoregressive correlation structure can be specified. Due to the dependent nature of rainfall over time, this type of structure seems appropriate. It is most common to chose an AR(1) structure and thus, this structure will be used as the initial 'working' correlation matrix. Dunn and

Lennox [24] showed that an AR(1) appears to be an appropriate choice to model rainfall.

Within a class of GLMs for correlated data, the initial choice of the variance function is driven by the range and nature of the response variable. Thus, a Tweedie distribution which can accommodate data with both discrete and continuous components is an obvious choice for modelling rainfall. Likewise, the initial choice of the link function for a particular model is usually chosen based upon the range of the response variable. In most cases the canonical link is used. The choice of the link function does not affect the outcome of the analysis in GEE models, but can affect the calculation of the sandwich estimate of variance. A logarithm link function is the most sensible choice to use when modelling rainfall. It is the most common to use with the Tweedie distribution and maps $-\infty < \eta < \infty$ to $\mu > 0$, which is sensible for modelling rainfall data. Although other choices of link function are available and may be appropriate, the logarithm link function is used in the following applications due to its feature of mapping into a non-negative continuous outcome.

Step 2 : Fitting a GLM to the data

After the parameters have been specified, initial estimates of $\boldsymbol{\beta}$ are to be estimated using GLM methodology. In order to do this, a usually-robust iterative procedure called iteratively reweighted least-squares (McCullagh & Nelder [62]) is used. To specify the Tweedie distribution, the mean (μ), dispersion parameter (ϕ), and the variance power (p) are needed. Standard algorithms are used to estimate μ, and maximum likelihood estimation is used to find ϕ. Finally, to estimate p, maximum likelihood estimation is used by employing a profile log-likelihood plot. Although this is computationally difficult, the R package has a function to compute p relatively easily using the tweedie packages: `tweedie.profile`. For an explanation of this function, see Section 3.5.1.

Step 3 : Compute Pearson's residuals

The Pearson's residuals are computed using Equation 4.5. Some of the terms in this equation can be simplified as the distribution and the variance function are known. The following is known:

- $\mathbf{y}_{it}$ represents the vector of responses (Section 4.1.2), which in the case of rainfall, is the rainfall amounts observed every month. This will not change as the algorithm to fit $\boldsymbol{\beta}$ is implemented.

- $\boldsymbol{\mu}_{it}$ is the mean of responses and is the 'predicted values'. Predicted or fitted values are those which have been predicted by the model created using the predictors set by the researcher.

- $V(\boldsymbol{\mu}_{it})$ represents the variance function, and in the case of a Tweedie family this is simply μ^p, where p will be a value between 1 and 2.

Step 4 : Calculating α

As stated in step 1, the working correlation structure that is used is the AR(1) structure, and thus the calculation of common α can be calculated as per Equation (4.7) and Equation (4.8) using the Pearson's residuals found in step 3.

It is generally advisable to choose a working correlation structure that is similar to the structure of observed correlations. This is because, although the GEE is robust to misspecification of the correlation structure, efficiency is increased to the extent that the specified structure is correct (Hedeker [42]).

Step 5 : Calculating $\mathbf{R}(\alpha)$

The working correlation matrix $\mathbf{R}(\alpha)$ is specified using the common α found in step 4. As the working correlation matrix is specified as a AR(1) structure, it will have the following form,

$$R_i = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ \alpha & 1 & \alpha & & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & & \vdots \\ \vdots & & & \ddots & \vdots \\ \alpha^{n-1} & \dots & & \alpha & 1 \end{bmatrix}.$$

Step 6 : Estimate the covariance, V_i

$\mathbf{A}_i$ is a $t \times t$ diagonal matrix with the variance function as the tth diagonal. Thus, the diagonal values are $\boldsymbol{\mu}^p$, as this is the variance function of the Tweedie family distribution. The covariance is then found using Equation (4.2).

Step 7 : Find an updated value of $\hat{\boldsymbol{\beta}}$

Using Equation (4.6), an updated value of $\hat{\boldsymbol{\beta}}$ can be found. The following points show the simplification of some of the values used this formula when a Tweedie distribution is used as the distribution of the response variable,

- $D_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is equal to $\mathbf{X}_i \times \boldsymbol{\mu}$;
- Use the V_i that was calculated in step 6;
- $\mathbf{y}_i$ is the value of the original response variable values (rainfall amounts per month);
- $\boldsymbol{\mu}_i$ is the value of the fitted rainfall amounts.

To find the final $\boldsymbol{\beta}$ values, iterations of steps 3 to 7 continue until a predetermined criterion of convergence is reached. The convergence criteria that was used in this dissertation was,

```
abs(dev - devold)/(0.1 + abs(dev)) > epsilon.
```

6.3 Choosing the correct number of parameters

In order to find the most suitable, adequate, and parsimonious model that represents the data satisfactorily, it is necessary to run the program numerous times with a variety of covariates, factors, and interaction terms. Two diagostic approaches, the QIC_u and R^2 are used to determine the most appropriate set of covariates. As it is impractical to fit all possible models due the length of the running time of the program, it is necessary to use a systematic approach to find the most appropriate model. Firstly, predictors suspected to be more likely to contribute to the rainfall variability are fitted first. Those with the lowest QIC_u were considered to represent the data more accurately. When a final model was found, each predictor was systemically removed to determine if they were necessary.

Chapter 7

Application of single site modelling

The next two chapters provide an illustration of the ideas presented in previous chapters concerning GEE models, using an application in rainfall modelling. The data consists of monthly rainfall amounts measured in millimeters, as described in Chapter 5. This chapter develops the program needed to produce a GEE model with a Tweedie distribution using a single location, Emerald.

7.1 Fitting procedures

A good model-building strategy is essential when dealing with large and complex data sets, such as the rainfall data sets which are generally quite large. When dealing with such complex data sets it is possible that more than one model may represent the given data adequately (Dunn & Lennox [24]). There is thus a high possibility that more than one rainfall model will be found to be appropriate for this data and thus several models may be examined in this dissertation. The final model specified in this dissertation may be just one of many possibilities, but it will be one of the most appropriate of the models produced.

There is no software available that fits a GEE with a Tweedie distribution therefore it was necessary to write a code so this type of modelling could be implemented (See Appendix B.1 for the described code). The program R was used to demonstrate the application of modelling rainfall using GEEs as this program had the necessary requirement of being able to produce a GLM with a Tweedie distribution.

In order to develop the required code necessary to implement a GEE with

a Tweedie distribution, a single site (and thus a single unit, $K = 1$) was used to check the running and performance of the program. Emerald (located in Queensland, Australia) was chosen to show the application of the GEE program developed. As previously noted, the model is fitted to the data from 1889 to 1992, the remainder is kept for validation purposes (Section 5.2.1). Once the program was checked using the single site, further developments were made to incorporate multi-sites into the modelling process.

The Emerald data set is very large, with rainfall data being recorded from January 1889 to December 2001 and thus it would be very time consuming to examine all possible rainfall models. Therefore a systematic approach had to be developed in order to find the most suitable model for rainfall and below gives an outline of this approach:

1. Produce a basic rainfall model which contains only a single covariate. A model with each proposed single predictor will be developed and the corresponding QIC_u will be examined for each model. The models with the lowest QIC_u will be used and examined in further detail.

2. Further predictors will be added to the model, in a systematic approach: blocks of covariates will be added successively with predictors being added in perceived order of importance.

3. Once an approach model has been formed with single covariates only, each covariate will be systematically excluded from the model, to determine if an improvement in the model can be made. These deletions ensure that the overall size of the model remains manageable throughout.

4. Interaction terms will be added to the model to determine if any of there terms are relevant in a rainfall model. Only those terms which have a practical significance or meaning will be included in the rainfall model.

5. Once an adequate model has been determined, the covariates will be removed one by one to see if any improvements in the rainfall model can be made.

6. Diagnostic checks will be completed on the final models to determine if they adequately model the data given.

To fit a Tweedie GLM to the rainfall data (the initial requirement of finding a GEE model), an appropriate value of the variance power, p needs to be found. This is determined by using the profile log-likelihood function, with

the maximum likelihood value from this function corresponding to the most appropriate value for p. The choice of p determines the which member of the Tweedie family of distributions will be used in the analysis. Confidence intervals for p are also produced (95%) and any p value within the 95% confidence interval produce very similar estimates, models and residuals plots (Dunn & Lennox [24]).

The procedure QIC$_u$ will be used as a preliminary diagnostic tool to determine the best set of covariates to use. The best subset of covariates is then the model that has the lowest QIC$_u$ value. This diagnostic technique does not determine the most appropriate and most parsimonious model, but rather the set of covariates that most efficiently represents the given data set. For example, adding a certain predictor may lower the QIC$_u$ but only slightly and thus it could be argued that this predictor may not be necessary in the overall rainfall model. Therefore careful examination and further diagnostics of the overall model produced is needed before any model interpretations can be made.

The order in which the predictors are fitted is important in any testing of the predictors for a GLM, however for a GEE model, the order does not matter. Despite this, it is still important that the predictors are systematically fitted in order of suspected importance, as the fitting procedure is quite tedious for a GEE model created for this dissertation.

7.2 Preliminary modelling of rainfall

The starting point for developing a rainfall model using a GEE was to produce several simple models incorporating only one predictor. These simple models served as a base against which to judge more complex representations. So that any rainfall model produced could be easily written, each predictor was represented by the following definitions (see Table 7.1 for further clarification),

- Month (M), where 1 = January, 2 = February, 3 = March, etc.;
- Wet-season factor (WS), a seasonality term representing three distinct rainfall periods;
- Season (S), where 1 = Summer, 2 = Autumn, 3 = Winter, 4 = Spring;
- An annual frequency sine (SIN) and cosine term (COS): sin $(2\pi\times$ Month$)/12)$ and cos $(2\pi\times$ Month$)/12)$;

- A six-monthly frequency sine (S1)and cosine term (C1); sin $(4\pi\times$ Month$)/12)$ and cos $(4\pi\times$ Month$)/12)$;
- Southern Oscillation Index (SOI);
- Southern Oscillation Index Phases (P);
- Indicators for the previous 12 months are represented as IND1 for an indicator for the previous month, IND2 for an indicator of rainfall two months ago, up to IND12 for an indicator of rainfall 12 months ago;
- A persistent indicator for the previous two months will be represented by IND1.2 and a persistent indicator for the previous three months will be represented by IND1.2.3;
- Year (Y).

Table 7.1 gives a summary of the QIC_u and R^2 value for each of the predictors when they are used as a single predictor in the model for rainfall at Emerald. Only the first three indicators and a persistent indicator were included in Table 7.1. All of the other indicators produced similar results to the indicators included in the table.

Table 7.1: A summary of the QIC_u when each of the covariates was used singularly in the model for rainfall at Emerald. The lower the QIC_u and higher the R^2 value, the better the covariate is at representing the data.

Predictor	Definition	QIC_u	R^2
Month	M	51876	0.219
Season	S	48250	0.193
Wet-season	WS	47176	0.179
Cosine & Sine	COS & SIN	49092	0.022
Cosine(1) & Sine(1)	C1 & S1	42638	0.029
SOI	SOI	43620	0.036
SOI phases	P	43682	0.024
Year	Y	43021	0.00002
Indicator 1	IND1	43662	0.019
Indicator 2	IND2	43798	0.005
Indicator 3	IND3	43635	0.024
Indicator 1.2	IND1.2	43012	0.001

An example of one of the models produced for Table 7.1 would be if season was used as the single predictor for rainfall at Emerald. The model would be of the following form,

$$\log \mu = 1 + S,$$

where μ is the expected rainfall amount per month, S is the factor season, the assumed distribution is the Tweedie distribution and the link function used is the logarithm link function. Note that if the predictor 'cosine' is used, then the corresponding 'sine' predictor must also appear in the model and visa versa (Dunn & Lennox [24]). This is because both the sine and cosine waves represent a variety of possible cyclical patterns of various lengths when appearing in a model together.

Table 7.1 shows that the predictors season, month and the wet-season factor have the highest R^2 values. These values when used as a single predictor for rainfall explain between approximately 18% to 22% of the variance seen in rainfall. Thus it was initially decided that these three predictors would provide the starting point for the modelling fitting process. Many of the other predictors have quite low QIC_u values, however their corresponding R^2 value is also very low. Therefore the following covariates were further developed: month; season; and the wet-season factor.

7.3 Further model developing

In order to build upon the models described in Section 7.2 to see if any improvements could be made, covariates were added one at a time to the predictors month, season, and the wet-season factor. Covariates were added in a logical order depending on their expected contribution to the variability of rainfall.

It was expected that seasonal factors would be the main contributors to the amount of rainfall that falls during the year, as rainfall is very much dependent on season variations. Preliminary analyses of the data for Emerald confirmed this as the three seasonal factors (wet-season factor, month and season) contributed the most to the variability of rainfall (Section 7.2). These three seasonal predictors produced the highest R^2 value when a rainfall model was developed with only one covariate. It should be noted that the wet-season factor and the predictors season, sine and cosine all involve the predictor month in their calculations and thus month can not be combined with any of these covariates.

Chandler & Wheater [12] suggested incorporating the seasonal variation into a rainfall model in the form of an annual frequency sine and cosine term (SIN and COS). The inclusion of the sine and cosine term attempts to

incorporate the cyclical nature of rainfall from one season to another. This term was the next predictor added to the rainfall model. Dunn & Lennox suggested a variation to the sine and cosine term suggested by Chandler & Wheater [12] called a six-monthly frequency sine and cosine term (S1 and C1).

The southern oscillation index (SOI) as well as the SOI phases (Stone & Auliciems [78]) were the next predictors included into the rainfall model. Stone & Auliciems [78], McBride & Nicholls [61] and Lough [58] state that the these predictors have an important influence in the modelling of rainfall, especially within the Queensland and northern New South Wales region.

Indicators for previous month's rainfall amounts and two persistent indicators were first described by Chandler & Wheater [12] and more recently used by Dunn & Lennox [24] as predictors for rainfall. The indicator variables were also added in a systematic way to the rainfall model so that all indicators could be investigated: an indicator for the previous month was added first; then an indicator for two months ago; and so on. The persistent indicators were incorporated into the rainfall model, after the indicator predictors had been determined.

The predictor, Year, was the last covariate to be added into the rainfall model. Chandler & Wheater [12] stated that this predictors could be used to incorporate long-term climatic variability or linear trend in a rainfall model. Finally, the interaction terms which have a practical or meaningful interpretation were investigated.

7.3.1 Model with no interactions

After examining Table 7.1 and conducting preliminary analyses on the data, it was decided that three different models would be developed: One involveing the predictor month; another the wet-season factor; and the last would involve the season predictor. These three predictors are all seasonal predictors and consequently, as expected, seasonal predictors are the fundamental contributes to rainfall.

Each of the three models were developed by systematically adding predictors in the perceived order of importance, until an adequate model was found. In the preliminary stages, an adequate model was determined as one which produces a low QIC_u value and a corresponding high R^2 value. This can be difficult as often the addition of a predictor will produce a slight increase in the QIC_u value but a large increase in the R^2 value with. In this case, the predictor would be included in the rainfall model as a significant predictor. Therefore careful judgement is needed with the inclusion of predictors and finding the final model is a very time consuming and complex procedure. No

interaction terms were included at this stage of the model formulation.

7.3.2 Month factor

Each covariate listed in Section 7.2 were combined with the month factor and cautiously included into the rainfall model, to determine if any improvement could be make. It was found that adding the SOI phase predictor and one of the persistent indicators (IND1.2), reduced the QIC_u to 49777 and increased the R^2 value to 25.49%. This was a significant improvement on the 'month only' model and thus SOI phases and IND1.2 appear to enhance the rainfall model. This new model can be written as,

$$\log(\mu) = 1 + M + P + IND1.2. \tag{7.1}$$

A variance power (p) value of 1.59 was produced when using the log-likelihood profile function with predictors month, SOI phase, and a persistent indicator. One difficultly with using month as the seasonal predictor is for each month, an extra coefficient or β value is needed. For example, Equation (7.1) would be written as the following when the corresponding β values are added to the model,

$$\begin{aligned}\log(\mu) = {} & \beta_0 + \beta_1 M_2 + \beta_2 M_3 + \beta_3 M_4 + \beta_4 M_5 + \beta_5 M_6 \\ + {} & \beta_6 M_7 + \beta_7 M_8 + \beta_8 M_9 + \beta_9 M_{10} + \beta_{10} M_{11} + \beta_{11} M_{12} \\ + {} & \beta_{12} P_2 + \beta_{13} P_3 + \beta_{14} P_4 + \beta_{15} P_1 5 + \beta_{16} IND1.2,\end{aligned} \tag{7.2}$$

where M_2 to M_{12} represent the months February to December (January is used as the reference month) and take on a value of '1' when the month examined needs representation and '0' otherwise. Furthermore S_2 to S_5 represent phases 2 to 5 of the SOI phases (taking on a '1' or '0' in similar fashion to the month factor). Equation (7.2) shows that 17 separate terms need to be included in the rainfall model when month is the seasonal predictor. The use of season or the wet-season factor will help to reduce the large number of coefficients required when using month combined with the SOI phases and yet maintain a degree of accuracy.

7.3.3 Wet-Season factor

Systematically fitting each predictor listed in Section 7.2 to the rainfall model with wet-season factor already fitted as a predictor, indicated that adding more predictors did not lower the QIC_u value. However a significant improvement was seen in the R^2 value (with a corresponding slight increase in the QIC_u value) with the inclusion of the following terms,

- Six-monthly frequency sine and cosine term (SIN1 and COS1);
- Persistent indicator - An indicator for the previous three months rainfall (IND1.2.3).

This combination of predictors produced a QIC_u value of 48056 and a R^2 value of 20.50%. Due to the uncertainty as to which model 'best' describes the variability of rainfall when the wet-season factor is used as the seasonal predictor, two models were further examined. The first, involving only the wet-season factor, can be written as,

$$\log(\mu) = 1 + WS. \tag{7.3}$$

The second, involves the wet-season factor, a six-monthly frequency sine and cosine term and a persistent indicator,

$$\log(\mu) = 1 + WS + SIN1 + COS1 + IND1.2.3. \tag{7.4}$$

7.3.4 Season factor

After exploring different models using season as the leading predictor, it was found that adding a six-monthly frequency sine and cosine term and two persistent indicators (IND1.2 and IND1.2.3) produced a rainfall model with a lower QIC_u and a higher R^2 value than a model using season as the only predictor. The wet-season factor was also tried as a predictor, however it did not make any improvement to the model. This new model, written as

$$\log(\mu) = 1 + S + SIN1 + COS1 + IND1.2 + IND1.2.3, \tag{7.5}$$

produced a QIC_u value of 48056 and a R^2 value of 20.50%.

7.3.5 Other leading factors

When examining Table 7.1 it was noticed that the combination of the six-monthly sine and cosine term produced the lowest QIC_u value of all of the predictors. As this term is also a seasonal predictor, it was further investigated. However, using the sine and cosine term was inappropriate without the inclusion of the season or wet-season factor.

To complete the investigation of rainfall models, other combinations of factors were examined which did not involve month, the wet-season factor or season. No other models, however, proved to provide a better fit for the rainfall data at Emerald. Consequently, four separate models (Model (7.1), Model (7.3), Model (7.4) and Model (7.5)) were analysed further.

Interactions

The inclusion of interaction terms did not improve any of the four models identified as potential adequate models for rainfall at Emerald. Thus it was concluded that each model would only involve single predictors and no interaction terms.

7.3.6 Fitted model

The final four models, (Model (7.1), Model (7.3), Model (7.4) and Model (7.5)) found after initial investigations to model the rainfall at Emerald using a GEE model and a Tweedie distribution were examined in more detail, to determine if any of these models were adequate and are able to be used to represent the rainfall at Emerald. The information for these four models can be seen in Table 7.2.

Table 7.2: A summary of the four models that were found to be representative of the rainfall data at Emerald after initial diagnostics only. Displayed is each model with their corresponding QIC$_u$, R^2 and variance power (p) values. The lower the QIC$_u$ and the higher the R^2 value, the better the model is at representing the rainfall data.

Model	Model No.	p	QIC_u	$R^2(\%)$
1+M+P+IND1.2	7.1	1.59	49777	25.5
1+WS	7.3	1.61	47176	17.9
1+WS+COS1+SIN1+IND1.2.3	7.4	1.61	48056	20.5
1+S+COS1+SIN1+IND1.2+IND1.2.3	7.5	1.61	48156	20.1

Table 7.2 shows that Model (7.3) has the lowest QIC$_u$ value and Model (7.1) has the highest R^2 value. Therefore, after preliminary diagnostics and analyses, none of the four models stand out as being better than the others. Further diagnostic testing and examination of the four models was needed to determine if any are adequate for modelling the rainfall at Emerald. The next section determines the suitability of the models listed in Table 8.2.

7.4 Diagnostics

Before attempting to interpret any of the results obtained, it is necessary to carry out thorough checks. For a statistical model, such checks fall broadly

into three categories: assessment of predictive ability; checks on probability structure; and checks for systematic structure. To assess the probability structure, testing the corresponding GLM is the best method of determining the appropriateness of the assumed distribution of the response variable and the chosen link function. Although this is not suggested in literature, no other methods of testing have been recommended. Checks for systematic structure were performed via the QIC_u, R^2 (which have already been performed) and the examination of residuals. Lastly assessment of the predictive ability of the final model/s are completed in Section 7.6.

7.4.1 Residuals

Initially the Wald-Wolfowitz randomness test (Section 4.5.3) was performed on each model listed in Table 8.2 to test that the raw residuals are distributed in a random sequence. This test produces a W_Z value, which is similar to a z-score, and a corresponding p-value. An extreme value of W_Z indicates that the model does not adequately reflect the underlying structure of the data and the residuals are not randomly distributed. Table 7.3 indicates that Models (7.4) and (7.5) have p-values below 5% and thus at the 5% significance level, there is enough evidence state that the residuals are non-randomly distributed, and thus these models should not be used as rainfall models. At the 10% level of significance Model (7.3) also has non-random residuals. Model (7.3) has a p-value between 5 and 10% and is not the borderline of having non-random or random residuals. It was therefore decided that Model (7.3) and Model (7.1) should be analysed further.

Table 7.3: The final models with their W_Z value from the Wald-Wolfowitz randomness test and the corresponding p-value. This tests the null hypothesis that the signs of the raw residuals are distributed in a random sequence.

Model	Model No.	W_Z	p-value
1+M+P+IND1.2	(7.1)	−1.110	0.1335
1+WS	(7.3)	−1.608	0.0537
1+WS+COS1+SIN1+IND1.2.3	(7.4)	−1.742	0.0409
1+S+SIN1+COS1+IND1.2+IND1.2.3	(7.5)	−2.588	0.0048

Residual Plots

There are three different residuals plots that are appropriate to use with GEE models,

- Pearson residuals versus predicted values;
- Raw residuals versus observation number;
- Pearson residuals versus linear predictor.

The Pearson residuals versus predicted values plot is used if there is any indication that the residuals depend on the unit identifier. As there is only one unit in this application ($K = 1$), this plot is not relevant. Therefore only the last two plots in the above list were produced for Model (7.1) and Model (7.3) to determine the adequacy of the models.

Model (7.3) was examined first to determine if the residuals plots indicate any inadequacies in the model. Figure 7.1 shows two separate problems. Firstly, the magnitude of the positive residuals is much larger than the magnitude of the negative residuals. This is not a major concern as there are reasons for why this could occur. It indicates that this model for rainfall does not predict extreme rainfall values very accurately. Also rainfall has lower bound at 0mm of rain, meaning a negative rainfall amount is impossible and thus it is expected that the negative residuals will have a lower magnitude. The second problem with Figure 7.1 is concerning. There is a reasonable congregation of points when the raw residual is equal to approximately -20. This indicates that the points may not be randomly distributed. The Wald-Wolfowitz test indicated the residuals were non-random at a 6% level of significant (p-value = 5.37%). It seems likely that this model has violated the randomness of residuals requirement and should not be used to model rainfall. However, further testing was performed to clarify this.

Figure 7.2 represents a plot of the Pearson residuals versus the linear predictor ($\eta = \log(\mu)$). If Model (7.3) accurately represents the rainfall data, Figure 7.2 should have a uniform spread of point. As the plot does not show uniformity, Model (7.3) provides a poor fit for the rainfall data at Emerald. As Model (7.3) has shown more than one violation, and has shown to provide a poor fit to the Emerald rainfall data, it is not considered an adequate model.

The residual plots for Model (7.1) produced better results than those for Model (7.3). Figure 7.3 demonstrates a plot of the raw residuals. This figure shows also that the magnitude of the positive residuals is much larger than the magnitude of the negative residuals. However, similar to Model (7.3), this could be due to the model not predicting the extreme values very accurately or because rainfall has a lower bound at 0mm meaning residuals

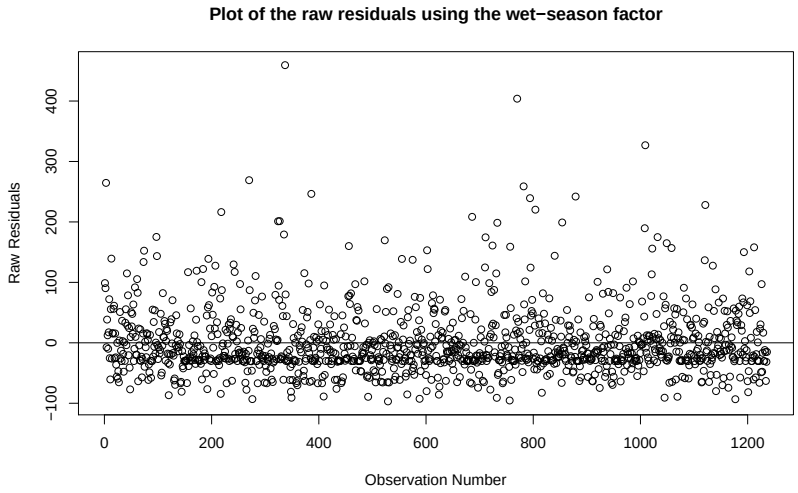

Figure 7.1: A plot of the raw residuals using the wet-season factor as the only predictor for rainfall at Emerald. The plot shows that the magnitude of the positive residuals is larger than that of the negative residuals.

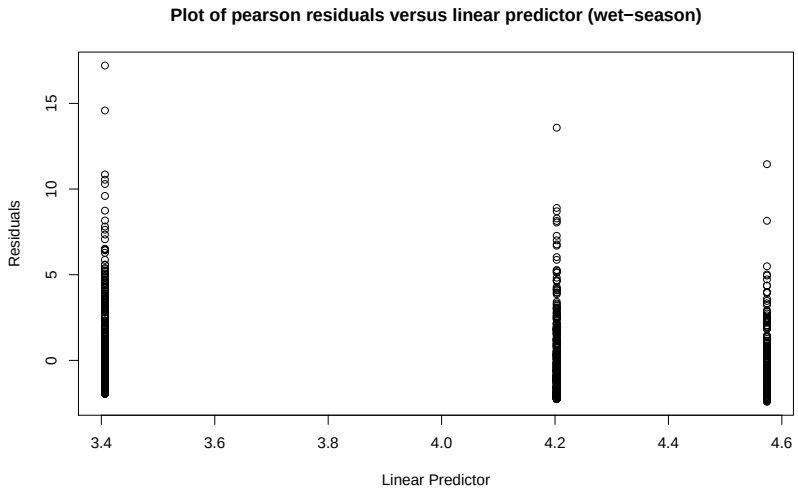

Figure 7.2: The Pearson residuals plotted against the linear predictor ($\eta = \log(\mu)$) using the wet-season factor as the only predictor.

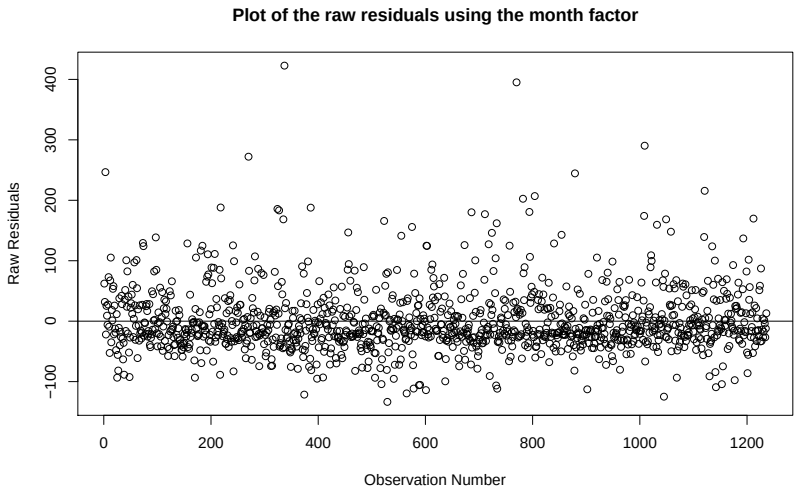

Figure 7.3: A plot of the raw residuals using month, SOI phases and a persistent indicator as the predictors for rainfall at Emerald. The plot shows that the magnitude of the positive residuals is larger than that of the negative residuals.

have a limit of how low they can be. Figure 7.4 gives an indication that the model is appropriate as the points on the plot of the linear predictor versus the Pearson residuals are uniformly spread.

The examination of the QIC_u, R^2, the Wald-Wolfowitz randomness test and several plots of the residuals indicates that Model (7.1) is the most appropriate of those models examined for the Emerald rainfall data. Note that not all combination of models were examined, as this would be an extremely time consuming and difficult task. This is why a systematic approach was developed. The final diagnostic test involved checking the suitability of the Tweedie distribution and the logarithm link function.

Checking the properties of the GLM

It is important to do a diagnostic check to determine if the underlying assumed response variable's distribution is correct and if the correct link function has been chosen to represent the data. However there is no literature stating how to do this for a GEE model and thus this section will verify that the associated GLM has a suitable link function and distribution, using techniques described in Section 3.6.

A Normal probability plot of the quantile residuals is an efficient method of determining if the distribution chosen to represent the data is appropriate

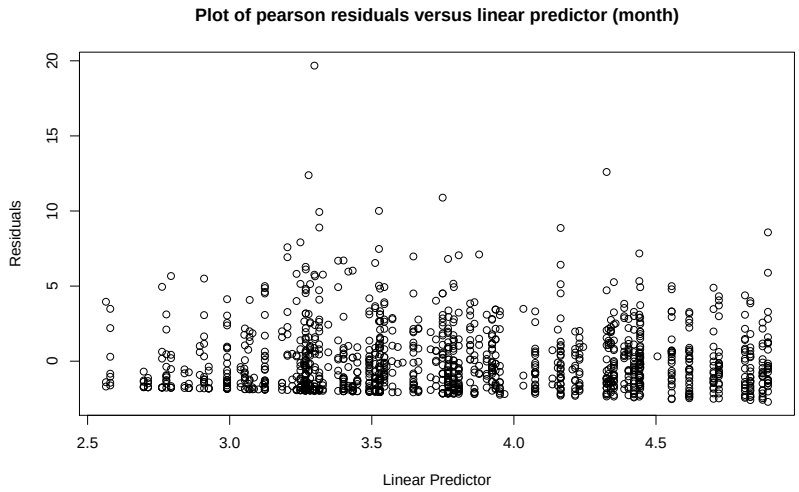

Figure 7.4: A plot of the Pearson residuals plotted against the linear predictor ($\eta = \log(\mu)$) using month, SOI phases and a persistent indicator as the predictors.

for the fitted model (Model (7.1)). Figure 7.5 shows the Normal probability plot of the quantile residuals for Model (7.1) using a GLM. It suggests that the model is appropriate for the given data as nearly all of the residuals lie close to the line indicating Normality. Some of the larger values do deviate from the Normality line, however these points only represent 0.7% of the data and with such a large data set ($n = 1236$) it is expected that there will be some minor deviations. Thus a Tweedie distribution, with a power index parameter of $\hat{p} = 1.59$ and $\hat{\phi} = 4.38$, suitability fits the monthly rainfall data at Emerald.

To verify the suitability of a link function the residual deviance can be calculated for numerous link functions. The link function with the lowest associated residual deviance is the most suitable to use. In R there are only two link function that are available to use with the Tweedie distribution: the logarithm and canonical link functions. Table 7.4 gives the residual deviances and residual degrees of freedom for a Tweedie GLM using first a logarithm and then a canonical link function for Model (7.1). The logarithm link function produces the lowest residual deviance indicating that it was the most suitable link function to use when modelling the rainfall at Emerald.

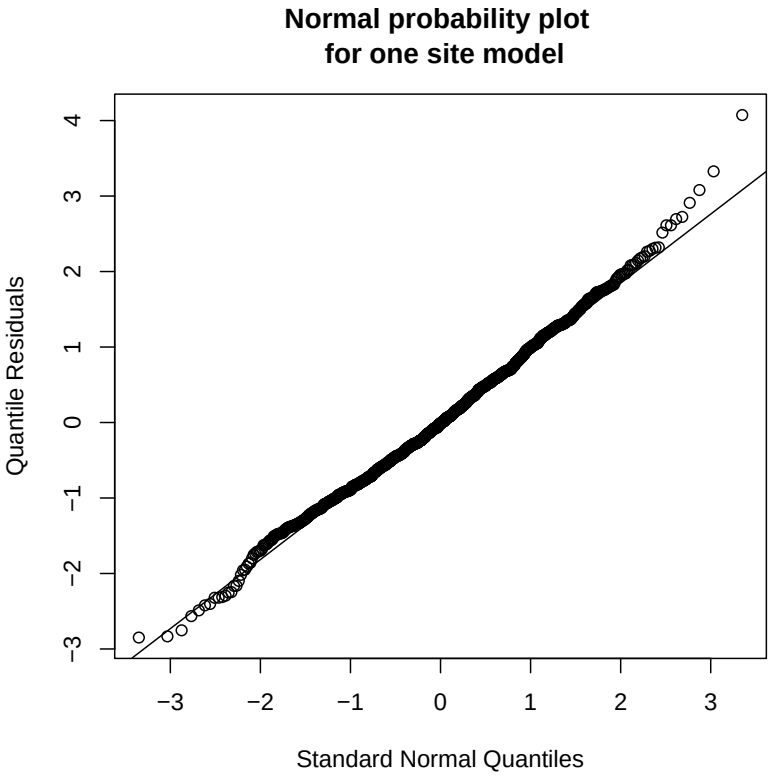

Figure 7.5: A Normal probability plot of the quantile residuals for the GLM of Model (7.1), suggests that the model is appropriate

Table 7.4: The residual deviances with the residual degrees of freedom for a Tweedie GLM with differing link functions for Model (7.1), which has month, SOI phases and a persistent indicator as covariates.

Link Function	Residual Deviance	Residual df
Logarithm	6754.85	1219
Canonical	6828.20	1219

Final model

After conducting numerous modelling fitting techniques and diagnostic testing, it was found that one model was superior to all others investigated for this dissertation when modelling the rainfall at Emerald using a GEE model. This model identified the following covariates as being significant for modelling the rainfall at Emerald,

- Month (M);
- SOI phases (P);
- A persistent indicator variable for the previous two months rainfall (IND1.2).

To fit this model to the Emerald rainfall data, an appropriate estimate for the variance power p was found using the profile log-likelihood plot. Figure 7.6 shows the profile log-likelihood plot suggesting a maximum likelihood estimate of $\hat{p} = 1.59$ with a 95% confidence interval of between 1.56 and 1.62.

After running the code in Appendix B.1, the coefficients and covariates of Model (7.1) are,

$$\begin{aligned}\log(\mu) \;=\; & 4.206 - 0.062M_2 - 0.453M_3 - 1.088M_4 - 1.126M_5 \\ & -1.105M_6 - 1.350M_7 - 1.626M_8 - 1.493M_9 - 0.971M_{10} \\ & -0.542M_{11} - 0.173M_{12} + 0.491S_2 - 0.015S_3 + 0.472S_4 \\ & +0.214S_5 + 0.197IND1.2,\end{aligned}$$

where,

- μ is the expected monthly rainfall amount at Emerald;

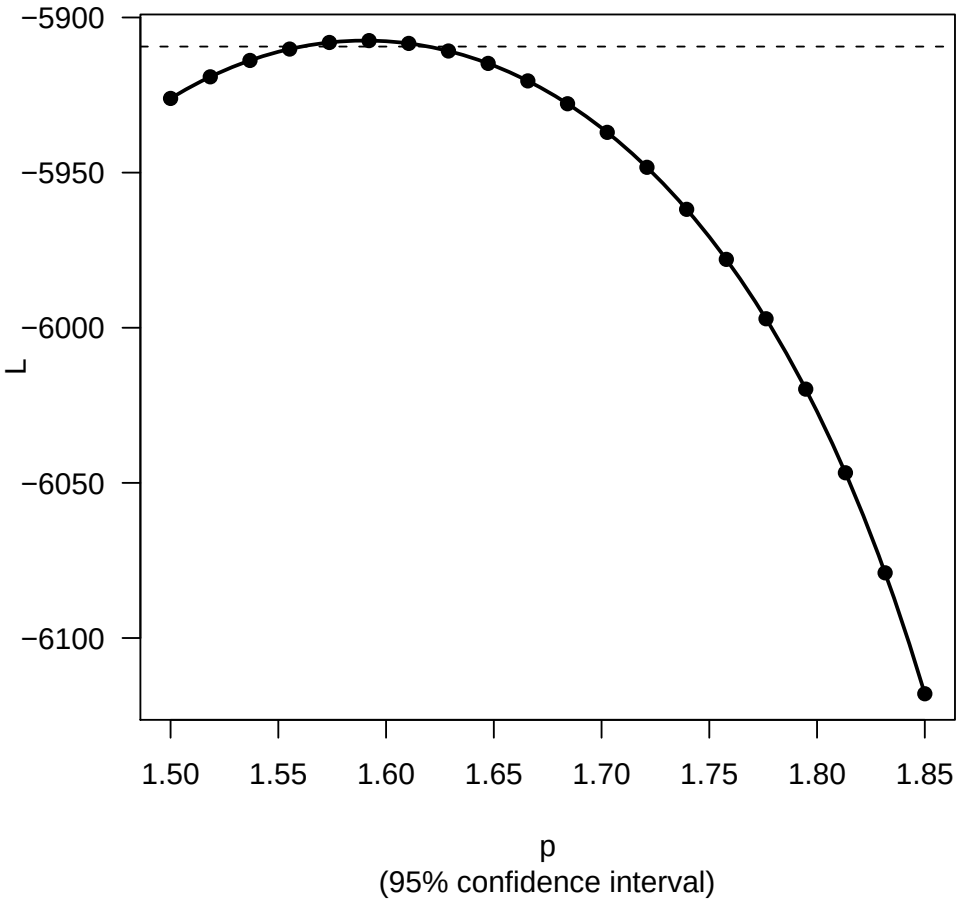

Figure 7.6: The profile log-likelihood plot for the monthly model at Emerald using month, SOI phase and a persistent indicator as covariates. This plot was used to estimate the maximum likelihood value of p. The points represent the computed likelihood values for differing p estimates, the solid line is a cubic-spline smooth interpolation through these points and the dotted line represents a 95% confidence interval for p. The estimate for p from this graph is $\hat{p} = 1.59$.

- Each month (M_2 = February, M_3 = March, up to M_12 = December) takes on a value of 1 or 0 depending on which month is being represented. The first month (M_1 = January) is the reference month and thus does not appear in the equation. This is an application of the use of treatment contrasts;

- Each SOI phase (S_2 is the SOI phase 2 up to S_5 which is the SOI phase 5) are represented in the same way as the months, taking on a value of 1 or 0 and the first SOI phase (S_1) being the reference phase;

- The persistent indicator variable for the previous two months rainfall; written $IND1.2$, where $IND1.2 = 1$ when rainfall was received during the two previous months, and is zero otherwise.

This model has a QIC_u value of 49777 and a R^2 value of 25.5%. Figure 7.7 shows the predicted rainfall values from Model (7.1) together with the observed rainfall values. Extreme values have not been modelled with much accuracy (the residuals plots have the same indication). Overall however, according to the diagnostics checks performed, Model (7.1) provides a good representation of the structure in the data, and the distributional assumptions are satisfied.

7.5 Model interpretation

It is not surprising that the month factor was found to be a significant predictor of rainfall at Emerald as Figure 5.4 indicates that there is a variation between the rainfall amounts for different months of the year. Figure 7.8 confirms the inclusion of the SOI phases in the final rainfall model, as the boxplot shows increased rainfall amounts during phase 2 and phase 4.

R^2 indicates how well the model does at predicting the variability of rainfall and initially it may seem that with an R^2 of only 25.5% Model (7.1) is not efficient at explaining the variance. However rainfall is extremely hard to predict thus a R^2 value of 25.5% is reasonable. The difficulty in predicting extreme rainfall events and the limited number of covariates examined, may be two of the many reasons why the percentage is low.

To interpret the coefficients of the final fitted model, recall that the model is written in terms of logarithms (that is $\log(\mu)$) and thus the coefficients imply a multiplicative factor in the model. To understand this effect, consider the equation $\log(\mu) = 4.206 - 0.062M_2 - 0.453M_3$ (which is the first three

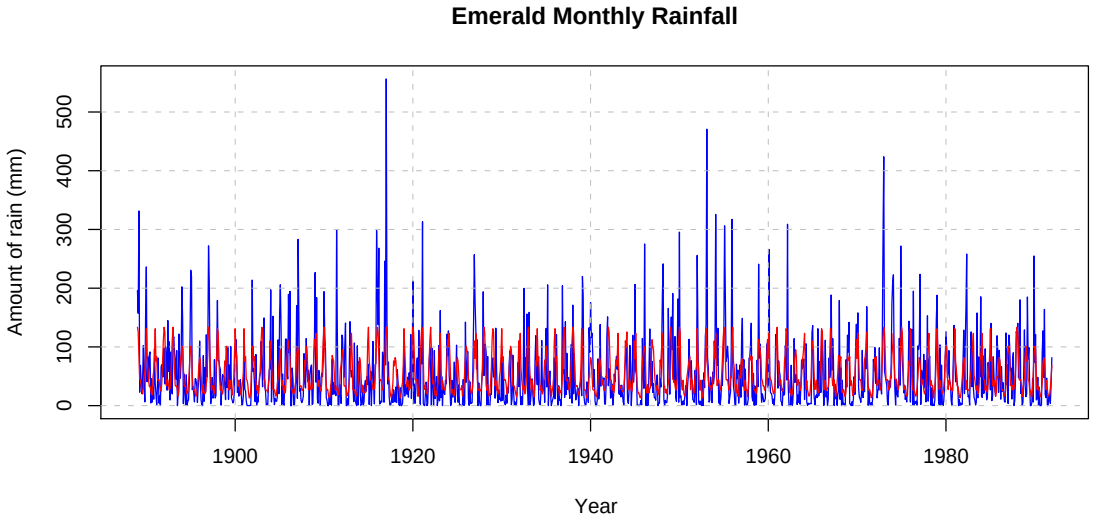

Figure 7.7: Time series plot of the observed (blue) and predicted (red) monthly rainfall amounts ($\geq 0mm$) for Emerald, obtained through the use of a Tweedie GEE ($p = 1.59$) with a logarithm link function. The fitted covariates include month, SOI phase and a persistent indicator.

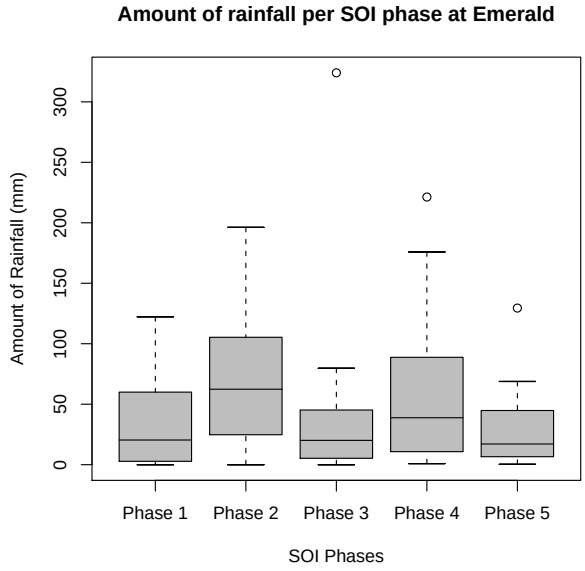

Figure 7.8: The amount of rainfall per SOI phase for Emerald.

months of the final equation) which can be written as,

$$\begin{aligned}\mu &= \exp(4.206 - 0.062M_2 - 0.453M_3) \\ &= \exp(4.206) \times \exp(-0.062M_2) \times \exp(-0.453M_3) \\ &= 67.09 \times 0.94^{M_2} \times 0.64^{M_3}.\end{aligned}$$

Thus for the month March (that is $M_3 = 1$) a coefficient of -0.453 implies an average decrease in the rainfall (when compared with January ($M_1 = 1$) by a factor of 0.64. All predictors involved in the final equation are indicator variables meaning they can only take on a value of 0 or 1. Therefore, the predictors' coefficients in the final model effect the overall predicted rainfall by a multiplicative factor only when the predictor takes on a value of 1.

Information about rainfall events

Another functional characteristic of the Tweedie distribution when used to model rainfall is that it can provide some useful information about different rainfall events. When $1 < p < 2$ the Tweedie parameters (μ, p, ϕ) can be reparameterized to the Poisson and gamma parameters $(\lambda, \gamma, \alpha)$ which can be used to provide different information about rainfall events. These transformations are,

- $\lambda = \mu^{2-p}/(\phi(2-p))$;
- $\gamma = \phi(p-1)\mu^{p-1}$;
- $\alpha = (p-2)/(1-p)$,

where λ is the mean number of rainfall events per month, γ is the shape of the rainfall distribution when rain occurs during the month and $\alpha\gamma$ is the amount of rain per rainfall event (Dunn [27]).

To best understand these transformations, an example is given. Consider the last observation in the estimation data set (December 1992) where $\hat{p} = 1.59$, $\hat{\phi} = 5.00$ and $\mu = 68.77$. Reparameterizing to Poison and gamma parameters gives: the predicted mean number of rainfall events for this month as $\lambda = 2.76$; the shape of the rainfall gamma distribution as $\gamma = 35.80$; and the mean amount of rain for December 1992 as $\alpha\gamma = 24.88$.

The probability of obtaining no precipitation on any particular month can also be calculated (Dunn [27]),

$$\Pr(Y = 0) = \exp(-\lambda) = \exp\left[-\frac{\mu^{2-p}}{\phi(2-p)}\right].$$

For the example just given (December 1992), where $\lambda = 2.76$, the probability of obtaining no rain in that month is 6.3%.

7.6 Model validation

Model validation is an important in the model building process as can be used to determine how accurately a model is at predicting data (historical data validation). In the previous section it was demonstrated that Model (7.1) performed satisfactorily and accurately fitted the estimation data set well. This section demonstrates that the rainfall data simulated from the final model (Model (7.1)) has similar properties to the actual rainfall data. Data from January 1993 to December 2001 was kept for validation purposes.

Model (7.1) was used estimate the mean of the rainfall distribution for each month using a series of one-step ahead forecasts. A random number was then drawn from corresponding Tweedie distribution based on the the forecasted μ, the previously fitted value of $p = 1.59$ and the previously found $\phi = 5.0$. This allows a predicted rainfall value to be calculated that occurs with a Tweedie distribution of mean value μ. The random number function used in R to find the predicted value is,

```
round(rtweedie(mu=fits,phi=phi,p=p,length(rain))).
```

The predicted values were then compared with the actual data in the validation set.

The results were examined for a variety of random number seeds that the results presented in this section are an actual representation of all of the results obtained.

The distribution of each month of rainfall was examined for the actual and validated data. Figure 7.9 shows the rainfall distribution for March and Figure 7.10 shows the rainfall distribution for December. From these two months it can been seen that the comparisons are quite good. The distributions for March are extremely similar and indicate that Model (7.1) provides an excellent prediction for this month. December does not provide such an exact fit as March does, however Model (7.1) still provides a reasonable prediction for this month. The monthly model performs well for most months, however for March and October it does not provide a good prediction of rainfall. This may have been improved if more data points were kept for the validation set.

When looking at the overall statistics of both the actual data from January 1993 to December 2001, and the validation set for this period, it can be seen that the validation set is satisfactory of predicting the overall mean and median of rainfall at Emerald (Table 7.5). The validation set displays a greater variance than the actual data. The validation set had 17 months of

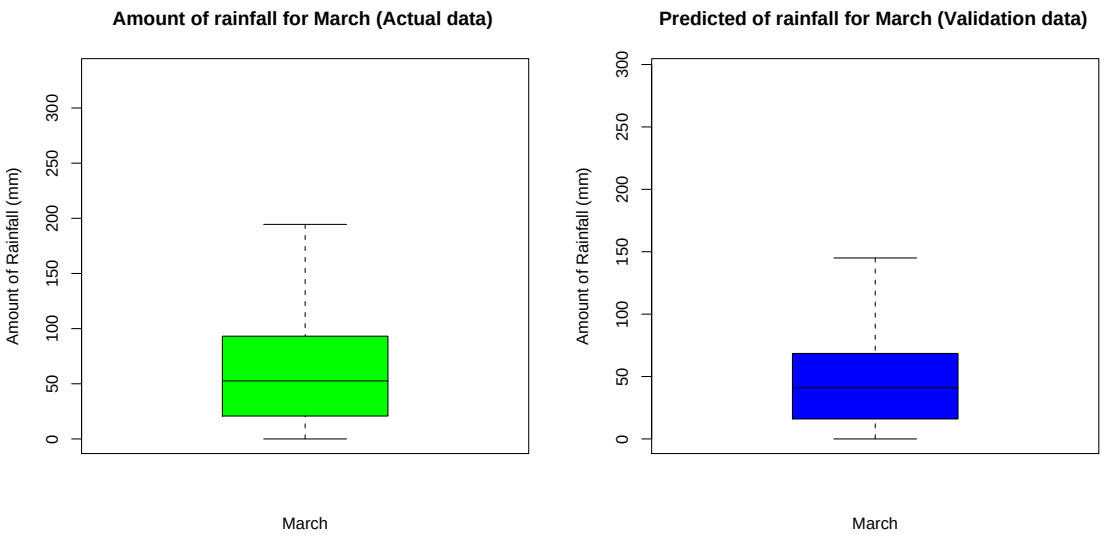

Figure 7.9: The distribution of the actual monthly rainfall (green) and the simulated data using the validation data set (blue) for March. The horizontal line in each box indicates the median of the distribution.

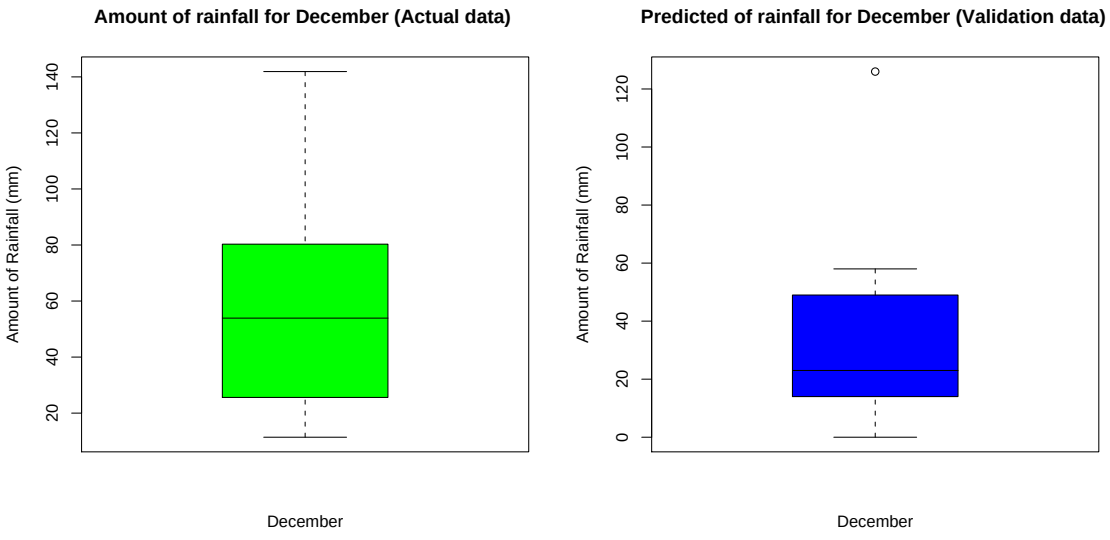

Figure 7.10: The distribution of the actual monthly rainfall (green) and the simulated data using the validation data set (blue) for December.

no rain whereas the actual data set had only 6 months: the monthly model often over-estimated the number of dry months.

Table 7.5: Summary statistics of monthly rainfall data for the actual and validation data sets from January 1993 to December 2001 . All measurements are recorded in millimeters.

Statistics	Actual data	Validation data
Mean	45.32	43.39
Median	26.40	21.50
Standard Deviation	53.03	61.85
IQR	56.98	45.00
n	108	108

Chapter 8

Application of multi-site modelling

Literature cites (Srikanthan [76]; Chandler & Wheater [12]); Chandler [11]) that modelling the spatial dependence of rainfall at different sites should be accommodated in rainfall models (Chapter 2, Section 2.5), however due to the cumbersome nature of this type of modelling, very few researchers have attempted to do so. Researchers that have used multiple sites, have done so only for either rainfall occurrence or rainfall amounts, and have not combined these categories. This section demonstrates that GEEs can be used to model rainfall at multiple sites, with an application demonstrating the simultaneous modelling of rainfall at two differing locations, Toowoomba and Gatton. Toowoomba and Gatton were chosen to demonstrate the usefulness of GEEs due the proximity of their locations.

8.1 Fitting procedures

The same systematic approach was used to find an adequate multi-site model for rainfall as for the single site rainfall model (Chapter 7). There was only two differences. The first was that three extra covariates were also examined: latitude; longitude; and altitude. These predictors were incorporated last in the modelling process, after the predictor 'year' has been investigated, and were assigned the following variable names,

- Longitude = 'LONG';
- Latitude = 'LAT';
- Altitude = 'ALT'.

The second difference was the wet-season factor was no longer considered as a predictor of rainfall. Toowoomba and Gatton do not display the same characteristics as Emerald for this factor and the classification of the wet-season factor is therefore irrelevant.

Again no software is available that fits models using a GEE with a Tweedie distribution, and thus an extension to the code formulated for the single site rainfall model has to be written so this multi-site model can be implemented. The code written is similar to the single site code, with a few minor adjustments having to be made. The final code for the multi-site GEE rainfall model can be seen in Appendix B.2.

The profile log-likelihood plot has to be computed for each model investigated to find the maximum likelihood estimate for the index parameter p. As the QIC_u depends upon the chosen $\hat{p}$ value, it is important that the correct p value is chosen.

8.1.1 Single terms

To find the most appropriate model for rainfall when multiple sites are examined, each covariate was added separately to the model to determine which predictors contribute the most information to the overall variability of rainfall. To re-clarify, each predictor is classified by the following notation (written in expected importance to the variability of rainfall),

- Month (M), where 1 = January, 2 = February, 3 = March, etc.;
- Season (S), where 1 = Summer, 2 = Autumn, 3 = Winter, 4 = Spring;
- An annual frequency sine (SIN) and cosine term (COS): sin ($2\pi\times$ Month)/12) and cos ($2\pi\times$ Month)/12);
- A six-monthly frequency sine (S1)and cosine term (C1); sin ($4\pi\times$ Month)/12) and cos ($4\pi\times$ Month)/12);
- Southern Oscillation Index (SOI);
- Southern Oscillation Index Phases (P);
- Indicators for the previous 12 months are represented as IND1 for an indicator for the previous month, IND2 for an indicator of rainfall two months ago, up to IND12 for an indicator of rainfall 12 months ago;
- A persistent indicator for the previous two months are represented by IND1.2 and a persistent indicator for the previous three months aree represented by IND1.2.3;

- Year (Y);
- Longitude (LONG);
- Latitude (LAT);
- Altitude (ALT).

As for a single site, this section was developed by starting with a simple single predictor model. All predictors were examined to determine which seem to be important in explaining the variance of rainfall. Two diagnostic techniques were initially applied to each model, the QIC_u and R^2. Those models with both a lower QIC_u and a higher R^2 will be considered to be the most adequate models at this stage.

Table 8.1 gives a summary of the QIC_u and R^2 values for each of the covariates when they are used as a single predictor for rainfall. Only the indicators that performed the 'best' with the QIC_u and R^2 were included in the table. If the Southern Oscillation Index is used as the single predictor of rainfall then the Tweedie generalized estimating equation model can be written as,

$$\log \mu = 1 + \text{SOI},$$

where μ is the expected amount of rainfall each month and the link function used is a logarithm link function. Note that if the 'cosine' predictor is used, it must appear with the corresponding 'sine' predictor (Dunn & Lennox [24]) and visa versa.

From Table 8.1 it can be seen that the seasonal predictors, month and season, and the combination of the sine and cosine terms explain approximately 15% of the variance seen in rainfall when each is used as a single predictor of rainfall. It was thus initially decided that these three predictors would provide the starting point for the modelling fitting process.

8.1.2 Model with no interactions

Due to preliminary analyses and previous information regarding rainfall data, it was decided that two different models would be developed: one to involve the predictor month; and the other to involve season. These two models were developed with predictors being added successively in perceived order of importance (Section 8.1.1), until an adequate model was found. No interaction terms were included at this stage of the model formulation.

Table 8.1: A summary of each of the covariates with their corresponding QIC_u value and R^2 value when each is used singularly in a model for rainfall. The lower the QIC_u and the higher the R^2 value, the better the predictor is at representing the rainfall data. Only those indicators which had the highest R^2 value were included in this table.

Predictor	Definition	Variance power (p)	QIC_u	R^2
Month	M	1.61	21950	0.198
Season	S	1.63	21985	0.157
Cosine & Sine	COS & SIN	1.64	19774	0.142
Cosine(1) & Sine(1)	C1 & S1	1.67	20054	0.021
SOI	SOI	1.67	18876	0.022
SOI phases	P	1.67	18883	0.024
Year	Y	1.68	18900	0.0134
Altitude	ALT	1.68	18551	0.00032
Longitude	LONG	1.68	18551	0.00092
Latitude	LAT	1.68	18551	0.00092
Indicator 2	IND2	1.68	18560	0.0031
Indicator 4	IND4	1.68	18561	0.0043
Indicator 12	IND12	1.68	18538	0.0097
Indicator 1.2	IND1.2	1.68	18558	0.00451

Month

To find if any other predictors improve the rainfall model, each predictor listed in Section 8.1.1 was systematically added to the model. As with the single site model, season and the two sine and cosine terms all involve the predictor month in their calculations and thus month can not be combined with any of these covariates. It was found that adding the predictors year and one of the persistent indicators appeared to improve the rainfall model by not only lowering the QIC_u value but also increasing the R^2 value. Although the R^2 value increased for many of the models fitted, the QIC_u value did not improve for any other model. A combination of an increase in the R^2 value and a decrease in the QIC_u value is needed for the model to be more appropriate. Thus the most adequate rainfall model when using month as the seasonal indicator, can be written as,

$$\log \mu = 1 + M + Y + IND1.2. \tag{8.1}$$

A Tweedie log-likelihood profile with month, year and a persistent indi-

cator as predictors was used to estimate the maximum likelihood value of p. The resulting maximum likelihood value from this graph for the variance power p is 1.61. This model had a QIC$_u$ value of 21905.76 and a R^2 value of 20.7%.

The shortfall of this model is that for each month, an extra coefficient or β value is needed. Thus the predictor month has eleven corresponding coefficients, one for each month minus one which is used as the reference month. The use of the season factor might help reduce the reasonably large number of coefficients required when using month as a predictor and yet keep a certain amount of accuracy.

Season

Systematically fitting each predictor listed in Section 8.1.1 to the rainfall model with season already fitted as a predictor, found that the following terms were significant predictors contributing to the modelling of rainfall,

- An annual frequency sine and cosine term (SIN and COS);
- Indicator 1 : An indicator for rainfall one month ago (IND1);
- Indicator 2 : An indicator for rainfall two months ago (IND2);
- Indicator 12 : An indicator for rainfall twelve months (or one year) ago (IND12).

This combination of predictors produced not only the lowest QIC$_u$ value but also produced the highest R^2 value (out of the models fitted with season already included in the model). Adding the extra five covariates lowered the QIC$_u$ to 21066 and increased the R^2 value to 18.4%. The best estimate of the variance power, using the log-likelihood profile, is now estimated to be 1.63 with a 95% confidence interval ranging from 1.56 to 1.69. The rainfall model can now be written as,

$$\log \mu = 1 + S + SIN + COS + IND1 + IND2 + IND12. \tag{8.2}$$

Sine and Cosine terms (Annual frequency)

After further exploration into different models it was found that when the sine and cosine terms were combined with other predictors (except month and season), this combination of predictors made substantial improvements to the rainfall model. The combination of the sine and cosine terms, together with the SOI and an indicator term, produced a QIC$_u$ value of 21025.38 which

is less than Model (8.1) and Model (8.2), however the R^2 value of 16.10% is also lower. The following model, which has a variance power value of $p = 1.6286[1.56, 1.69]$, will also be investigated more thoroughly in the next sections:

$$\log \mu = 1 + SIN + COS + SOI + IND2. \tag{8.3}$$

Thus there are now three models that need further analysis.

8.1.3 Interaction terms

Model (8.1), Model (8.2) and Model (8.3) were further examined to see if the inclusion of any interaction term would make an improvement in the overall rainfall model. As there are many different interaction terms available, only those interaction terms which have practical significance and are of second-order only, were examined in this dissertation. Again the variance power has to be estimated for each fitted model.

Month

No improvements could be made to Model (8.1) with the inclusion of interaction terms.

Season

An interaction between season and the SOI has a valid practical significance to rainfall. It may be the case that the effect of the SOI on rainfall may be different in different seasons, and when this term (S:SOI) was added into model 8.2 it slightly improved the rainfall model. The QIC_u remained very similar to the model without the interaction term added, however the R^2 was improved to 18.4%.

However, as stated in Section 7.1, a systematic approach is needed in order to find the most suitable model for rainfall and step 5 states that 'once an adequate model has been determined, the covariates will be removed one by one to see if any improvements in the rainfall model can be made'. Thus each of the covariates were removed to determine if any improvements could be made. It was found that the indicator terms (IND1, IND2 and IND12) were not needed, as a model with season, sine and cosine terms and the interaction term of season and SOI, produced a slightly improved QIC_u and R^2 values of 21063 and 18.5% respectively.

The best estimate of the variance power, using the log-likelihood profile, is now estimated to be 1.62 with a 95% confidence interval ranging from 1.55

to 1.69. The rainfall model can now be written as,

$$\log \mu = 1 + S + SIN + COS + S : SOI. \tag{8.4}$$

Sine and Cosine Terms (Annual frequency)

An improvement to the sine and cosine model 8.3 could be made with the inclusion of an interaction term between SOI and sine (SOI:SIN). It could be argued that this interaction term is not needed as only a very slight improvement was made with its inclusion; the QIC_u became 21023.7 and the R^2 was 16.12%. The best estimate of the variance power is estimated to be 1.63 with a 95% confidence interval ranging from 1.56 to 1.70. The sine and cosine rainfall model can now be written as,

$$\log \mu = 1 + SIN + COS + SOI + IND2 + SOI : SIN. \tag{8.5}$$

It was decided that the sine and cosine model with (Model (8.3)) and without (Model (8.5)) the interaction term would be examined in further detail to determine if either of these models were an adequate rainfall model for Toowoomba and Gatton. This is because the QIC_u and R^2 values were very similar.

8.1.4 Fitted model

Four final models were fitted to the rainfall data at Toowoomba and Gatton and the information pertaining to these models can be seen in Table 8.2. All four will be examined in further detail, to determine if any of the models are adequate and can be used to represent the rainfall at these two locations.

From the preliminary analysis and diagnostic tests, Model (8.5) has the lowest QIC_u value and Model (8.1) has the highest R^2 value, however none of the models stand out as being better than the others at this stage. Further diagnostic testing was required to determine if any of the models given in Table 8.2 were an adequate rainfall model. It is interesting to note that 'indicator 2' (an indicator for rainfall two months ago) appears in two of the four models. It is uncertain why this indicator and not an indicator for rainfall one month ago, improved the rainfall model. The next section performs further diagnostics on the final four models to determine their suitability for modelling rainfall.

Table 8.2: A summary of each of the four models that were found to be representative of the rainfall data at Toowoomba and Gatton after initial diagnostics only. The table displays the four models, with their corresponding QIC$_u$, R^2 and variance power (p) values. The lower the QIC$_u$ and the higher the R^2 value, the better the model is at representing the rainfall data.

Model	Model No.	p	QIC$_u$	R^2(%)
1+M+Y+IND1.2	8.1	1.61	21906	20.5
1+S+SIN+COS+S:SOI	8.4	1.62	21063	18.5
1+SIN+COS+SOI+IND2	8.3	1.63	21025	16.10
1+SIN+COS+SOI+IND2+SOI:SIN	8.5	1.63	21024	16.12

8.2 Diagnostics

The three main diagnostics available in GEE models are the QIC$_u$ and R^2 (which have both been tested throughout the modelling building process) and the examination of the residuals, including the Wald-Wolfowitz randomness test. Each of the four models described in Section 8.1.4 were thoroughly checked using each of these diagnostic techniques. Furthermore although literature does not suggest any methods of checking the assumed distribution of the response variable and the appropriate link function, tests were performed on the corresponding GLM, produced in the initial stages of the GEE methodology. Although this is not suggested anywhere in literature, it will hopefully show that the Tweedie distribution and the logarithm link function are appropriate to use.

8.2.1 Residuals

The Wald-Wolfowitz randomness test (Section 4.5.3) was performed on each of the models to test that the signs of the raw residuals were distributed in a random sequence. If the residuals are not distributed in a random sequence, an extreme W_Z value and corresponding lower p-value is found. Extreme values of W_Z (and thus a low p-value) indicate that the model does not adequately reflect the underlying structure of the data. Table 8.3 shows that Models (8.1), (8.3) and (8.5) all have p-values below 5% and thus at the 5% level there is enough evidence to reject the null hypothesis that the residuals are random. This could be due to numerous reasons and indicates that these three models should not be used as rainfall models, as the residuals are not

distributed randomly. Only Model (8.4) indicates that the residuals from the model are randomly distributed. As models (8.1), (8.3) and (8.5) have all violated one of the requirements of a GEE model, Model (8.4) will be the only model to be examined further.

Table 8.3: Each of the four final models are shown with the W_Z value calculated from the Wald-Wolfowitz randomness test and the corresponding p-value. This randomness investigation tests the null hypothesis that the signs of the raw residuals are distributed in a random sequence.

Model	Model No.	W_Z	p-value
1+M+Y+IND1.2	8.1	-3.355	0.0004
1+S+SIN+COS+S:SOI	8.4	-0.964	0.1675
1+SIN+COS+SOI+IND2	8.3	-2.000	0.0228
1+SIN+COS+SOI+IND2+SOI:SIN	8.5	-2.403	0.0081

Numerous residual plots can now be drawn to check the adequacy of Model (8.4). Firstly Figure 8.1 gives no indication that the residuals depend on either the unit identifier (K, which determines which site is being modelled) or on the repeated measures identifier (t, the time points for the data). The two plots in Figure 8.1 show that all the plots are similar and thus Model (8.4) seems adequate. Secondly a plot of the observation number versus the raw residuals can be plotted (Figure 8.2). This figure shows that the magnitude of the positive residuals is much larger than the magnitude of the negative residuals. However this is caused by the fact that a model for rainfall does not predict extreme rainfall values very accurately, thus causing large positive residuals for these extreme values. Although a plot of this form does indicate a misspecification of indeterminate nature, it is not a cause for concern and does not indicate that the model is inadequate, just that it does not model extreme values very accurately. It can also be explained (as stated in Chapter 7) by the fact that rainfall has a lower bound at $0mm$ (meaning you can not attain a rainfall value less than $0mm$) and therefore the magnitude of the negative residuals is also bounded. Finally a plot of the Pearson residuals versus the linear predictor ($\eta = \log(\mu)$) (Figure 8.3) shows that the the points are fairly uniformly spread out, indicating that the model is adequate for the data.

Checking the QIC_u value, R^2 value and several plots of the residuals indicates that Model (8.4) seems to be an adequate model for rainfall. The

final diagnostic testing involves checking that the Tweedie distribution is appropriate and the logarithm link function is adequate.

8.2.2 Checking the properties of the GLM

As there is no literature stating what procedures are available to check the suitability of the chosen distribution of the response variable and to check that the link function chosen is adequate for a GEE model. Therefore this section checks that the associated GLM has a suitable link function and distribution.

A Normal probability plot of the quantile residuals is used to check that the distribution chosen to represent the response variable is appropriate for the fitted model. Figure 8.4 shows the Normal probability plot of the quantile residuals for the Model (8.4) using a GLM. This plot shows that nearly all the residuals lie close to the line indicating Normality. There are some larger values which do deviate from the Normality line and this is an indication that the model does not fit extreme values very well. However for a large data set, it is expected that there will be some minor deviations from the Normal line and these points only represent 2% of the data. Thus a Tweedie distribution with a power index parameter of $\hat{p} = 1.62$ and $\hat{\phi} = 3.83$ fits the monthly rainfall data appropriately when using a GLM to model the data.

The logarithm link function can be checked to determine if it is the most suitable link function to use. This is done by refitting the GLM with the most suitable covariates (Model (8.4)) using the canonical link function and calculating the residual deviances for both link functions. The link function with the lowest associated residual deviance is the most suitable link function to use. Table 8.4 shows the residual deviances and residual degrees of freedom for a Tweedie GLM with either a logarithm or canonical link function and with the predictors season, cosine and sine terms and an interaction term between season and SOI. As can be seen in this table the logarithm link function has the lowest residual deviance indicating that it was appropriate to use it as the link function in the modelling of the rainfall data.

8.2.3 Final model

Systematic testing of covariates, using the QIC_u and R^2, was performed to determine the best set of predictors to use for a multi-site model for rainfall, using a GEE model. This produced four different models for rainfall. After performing numerous diagnostic testing it was found that one model was superior to the others when modelling the rainfall at Toowoomba and

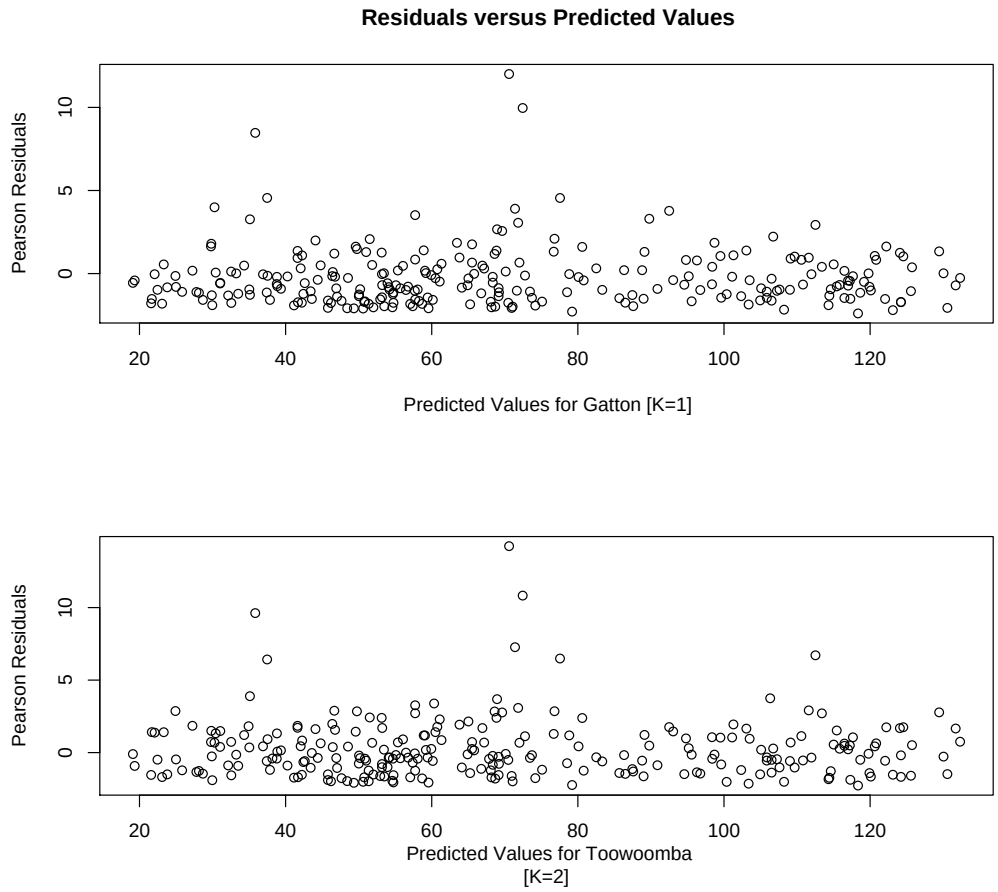

Figure 8.1: Plots of the predicted values for Gatton (top plot) and Toowoomba (bottom plot) versus the corresponding Pearson residuals. As both plots are similar there is no indication that the residuals depend on the location or on the time points.

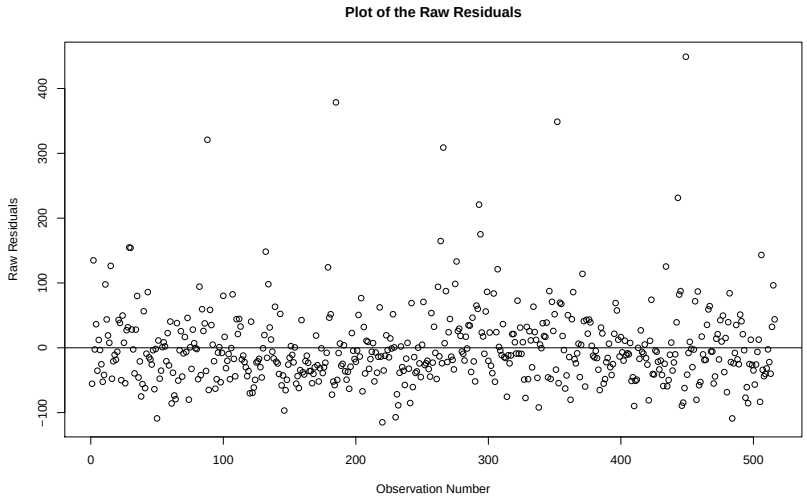

Figure 8.2: A plot of the raw residuals. The plot shows that the magnitude of the positive residuals is larger than that of the negative residuals.

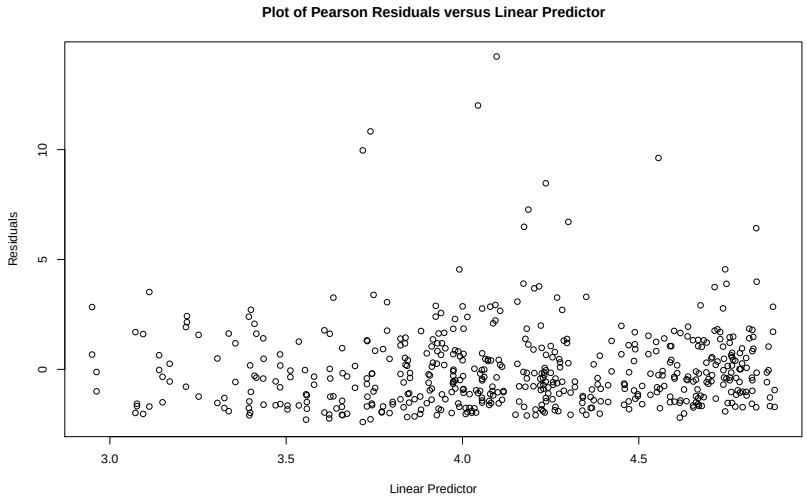

Figure 8.3: A plot of the Pearson residuals plotted against the linear predictor ($\eta = \log(\mu)$). The plot shows that Model (8.4) is appropriate as the points are uniformly spread.

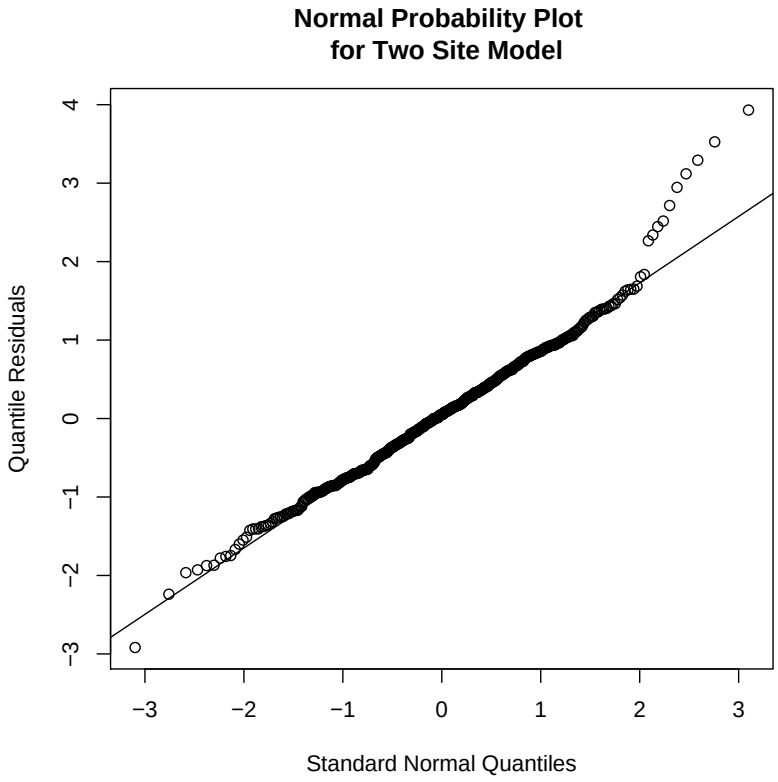

Figure 8.4: This Normal probability plot of the quantile residuals for the GLM of Model (8.4), suggesting that the model is appropriate as the residuals lie close to the Normality line.

Table 8.4: The residual deviances with the residual degrees of freedom for a Tweedie GLM with differing link functions for Model (8.4), which had season, cosine and sine terms (annual) and an interaction term between season and SOI as covariates.

Link Function	Residual Deviance	Residual df
Logarithm	1771.88	506
Canonical	1787.33	506

Gatton using a GEE model. The following variables were identified as being significant for modelling rainfall at Toowoomba and Gatton,

- Season (S);
- An annual frequency sine (SIN) and cosine term (COS): sin ($2\pi\times$ Month)/12) and cos ($2\pi\times$ Month)/12);
- An interaction term between season and the SOI (S:SOI).

To fit this model to the monthly rainfall data, first the appropriate estimate for the variance power p was found using the profile log-likelihood plot. This plot (Figure 8.5) suggests using a maximum likelihood estimate of $\hat{p} = 1.62$ with a 95% confidence interval of [1.55, 1.69].

After running the code in Appendix B.2, the coefficients and covariates of this model can be written as:

$$\begin{aligned}\log(\mu) &= 4.493 - 0.210S_2 - 0.627S_3 - 0.310S_4 \\ &+ 0.197SIN + 0.193COS + 0.009S_1 : SOI + 0.025S_2 : SOI \\ &+ 0.029S_3 : SOI + 0.010S_4 : SOI,\end{aligned}$$

however as SIN = sin($2\pi\times$ month / 12) and COS = cos($2\pi\times$ month / 12), this can be written as,

$$\log(\mu) = 4.493 - 0.210S_2 - 0.627S_3 - 0.310S_4 + 0.197 \times \sin(2\pi \times M/12) \quad (8.6)$$

$$\begin{aligned}&+ 0.193 \times \cos(2\pi \times M/12) + 0.009S_1 : SOI \\ &+ 0.025S_2 : SOI + 0.029S_3 : SOI + 0.010S_4 : SOI,\end{aligned}$$

where μ is the expected monthly rainfall amount and M takes on a value of 1 for January, 2 for February, 3 for March , up to 12 for December. Each season

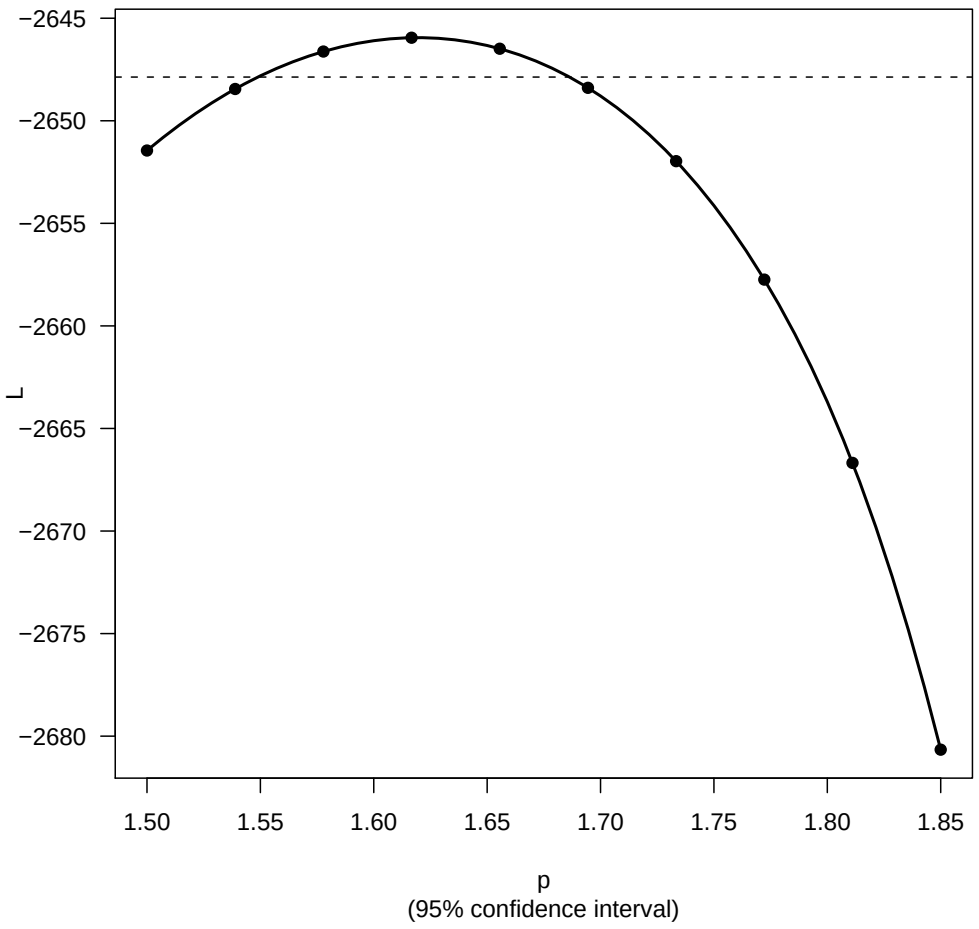

Figure 8.5: The profile log-likelihood plot for the monthly model using season, sine and cosine terms, and an interaction term between season and SOI, as covariates, which was used to estimate the maximum likelihood value of p. In this Figure, the points represent the computed likelihood values for differing p estimates, the solid line is a cubic-spline smooth interpolation through these points and the dotted line represents a 95% confidence interval for p. The estimate for p from this graph is $\hat{p} = 1.62$.

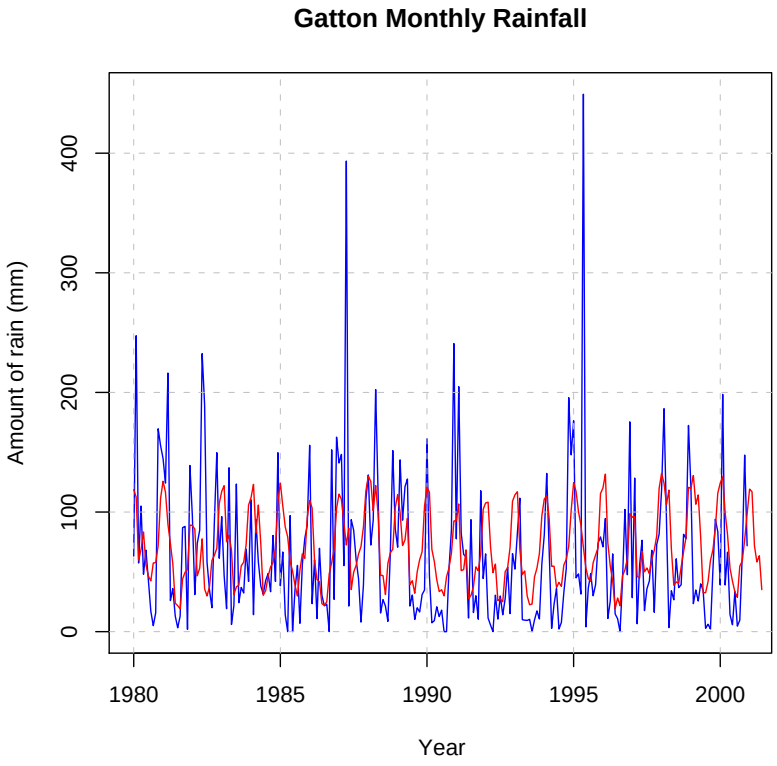

Figure 8.6: Time series plot of the observed and predicted monthly rainfall amounts (≥ 0) for Gatton, obtained through the use of a Tweedie GEE (p=1.62) with a logarithm link function. The fitted covariates include a season factor, an annual sine and cosine term and an interaction term between season and SOI.

(S_2 = Autumn, S_3 = Winter, S_4 = Spring) takes on a value of either 1 or 0 depending on which season is being represented, that is treatment contrasts are being used. The first season, Summer (S_1) is the reference season and thus it does not appear in the Equation (8.6).

This model has a QIC_u value of 21062.5 and a R^2 value of 18.7%. The plot of the predicted rainfall values from this model can be seen in Figures 8.6 and 8.7, in which the observed rainfall values have also been plotted. Figure 8.6 shows the predicted and observed values for Gatton and Figure 8.7 shows the values for Toowoomba. As can be seen in these plots, extreme rainfall values have not been modelled very accurately, with the residual plots and Normal probability plot confirming this notion earlier. Extreme rainfall events are extremely difficult to model and thus this shortfall is expected.

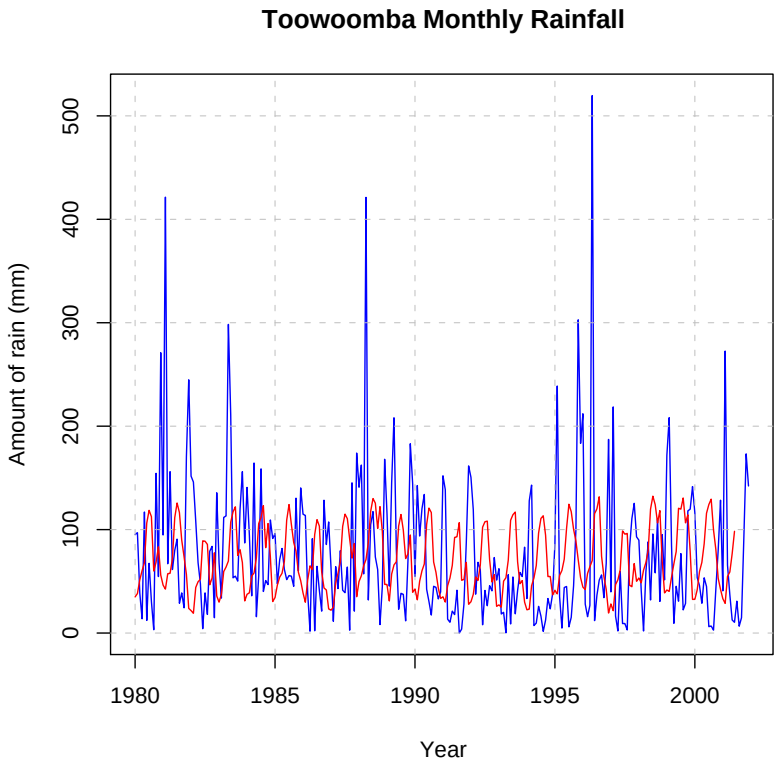

Figure 8.7: Time series plot of the observed and predicted monthly rainfall amounts (≥ 0) for Toowoomba, obtained through the use of a Tweedie GEE (p=1.62) with a logarithm link function. The fitted covariates include a season factor, an annual frequency sine and cosine term and an interaction term between season and SOI.

8.3 Model interpretation

From the fitted model (Model 8.6) the inclusion of the season factor in the final model indicates there is a variation in rainfall amounts for the different seasons. However it is surprising that both the SOI and SOI phases were not required in the final model for rainfall.

The model confirms the belief that the Summer months are the wettest months. From Model (8.6) it can be seen that all seasons (except Summer which is the reference month) have a negative coefficient indicating that rainfall is lower in Autumn, Spring and Winter than in Summer.

R^2 is an indicator of how well a model does at predicting the variability of rainfall. At first glance, with an R^2 of only 18.5%, it would appear that this model does not do a very good job of explaining the variability of rainfall. However even though the R^2 value indicates a variance explanation of approximately 20%, Hardin and Hilbe [39] suggest that interpretation is difficult and is not a true indication of the adequacy of the model. Sometimes only the researcher knows the full interpretation of this value. Rainfall is an extremely fluctuating occurrence and is very difficult to predict. Thus, a percentage this low is expected. Finally, it has been shown that the extreme values of rainfall are hard to model and predict and this would cause the percentage to be lowered.

Dunn and Lennox [24] suggest that the power-variance parameters (μ, p, ϕ) can be reparameterized to the Poisson and Gamma parameters $(\lambda, \gamma, \alpha)$ which aids in providing information about certain rainfall events. The transformation when $1 < p < 2$,

- $\lambda = \mu^{2-p}/(\phi(2-p))$;
- $\gamma = \phi(p-1)\mu^{p-1}$;
- $\alpha = (p-2)/(1-p)$,

where λ is the mean number of rainfall events per month, γ is the shape of the rainfall distribution when rain occurs during the month and $\alpha\gamma$ is the amount of rain per rainfall event (as stated in Chapter 7, Section 7.5).

For the multiple site application consider the observation from January 2000, where $\hat{p} = 1.62, \hat{\phi} = 3.83$, and the prediction is $\mu = 119.79$ for Gatton and $\mu = 41.63$ for Toowoomba. Reparameterizing to Poison and Gamma parameters gives the mean number of rainfall events for this month predicted to be $\lambda = 4.22$ for Gatton and $\lambda = 2.83$ for Toowooomba; the shape of the rainfall gamma distribution is $\gamma = 46.62$ for Gatton and $\gamma = 24.17$ for Toowoomba; and the mean amount of rain per this month is $\alpha\gamma = 28.40$ for Gatton and $\alpha\gamma = 14.73$ for Toowoomba.

The probability of obtaining no precipitation on any particular month is given by (Dunn [27]),

$$\Pr(Y = 0) = \exp(-\lambda).$$

For the example just given from January 2000, where $\lambda = 4.22$ for Gatton and $\lambda = 2.83$ for Toowoooomba, the probability of obtaining no rain in that month was 1.47% for Gatton and 5.9% for Toowoomba.

This section developed a model for monthly rainfall data for multiple sites at Toowoomba and Gatton using GEE and the Tweedie distribution. The use of the GEE model with the Tweedie distribution allows not only the simultaneous modelling of both the occurrence and amount of rainfall per month, but also allows for multi-site modelling of rainfall, whilst taking into account the dependent nature of rainfall data.

8.4 Model validation

Even though model validation is important when creating a model, to test the predictability of the model, due to time constraints this was not done for the multi-site model. It was demonstrated that Model (8.4) performed satisfactorily in all diagnostic tests and accurately fitted the data set well. However, validation should be completed and this is a possibility for future research.

Chapter 9

Conclusion

This dissertation has attempted to demonstrate the potential benefits of using GEEs for modelling and interpreting historical rainfall records. GEEs are especially designed to handle data that is dependent and as there is a general consensus that rainfall is correlated, GEEs offer a clear advantage for modelling rainfall. Past reviews have indicated that utilising GEEs to model rainfall may be beneficial to rainfall (Chandler & Wheater [10]; Beersma [6]; Buishand [7]). Chandler [11] also suggested previously that inter-site dependence within rainfall may be able to be tackled using GEEs and limited work has been completed in this area (Srikanthan & McMahon [76]). Thus the motivation and applicability for the development of GEEs to model rainfall, as developed in this dissertation, is apparent.

Simultaneously modelling the occurrence and amount of rainfall, combined with incorporating the correlated structure of rainfall and being able to model multi-sites concurrently has not previously be done. Thus this dissertation is not only applicable but is an initiative approach to modelling rainfall.

A model for monthly rainfall for Emerald, Toowoomba and Gatton using GEEs and Tweedie distributions was developed in this dissertation. The use of the GEEs to model rainfall, combined with the use of a Tweedie distribution, simplifies rainfall modelling by using only one model for the occurrence and amount of rainfall and provides an extension and advancement in modelling rainfall. Tweedie distributions to model rainfall has proven to be a very suitable and practical alternative to existing models and thus this distribution combined with a GEE model to incorporate correlation, is an extremely useful and powerful research.

A GEE model was developed at a single site, Emerald, to show the potential of using this technique for modelling rainfall. An extension of the single site model that incorporates multiple sites at Toowoomba and Gatton, is an

exciting development in rainfall modelling. It was found that the both model performed satisfactorily, however they were unable to model extreme events accurately. However, Hardin and Hilbe [39] stated that there is usually only a small relative gain in using a GEE model over the independence model (which assumes that no correlation exists) is relatively small if the number of units in the dataset is small. It is recommend that the independence model be used when there are less than 30 units in a data set. The models produced for this dissertation only had one unit for the single site model and two units for the multi-site model. Thus, it is expected that the models may not be as accurate as possible. The main purpose of this dissertation however, was to demonstrate an understanding and develop an initial baseline for further advancements and thus the GEE and not the independence model, was used. Simultaneously modelling rainfall at more sites is a definite possibility for further research and will hopefully prove that the GEE method for modelling rainfall is most appropriate.

It was found that for the single site model the significant covariates were month, SOI phase and a persistent indicator. For the multi-site model the most appropriate predictors were found to be season, an annual frequency sin and cosine term and an interaction term between season and SOI. The single site was validated using historical model validation and it was found that the final model produced replicated the actual data well. The shortfall with both models was the difficulty in modelling extreme rainfall events and the time it takes to run each model. Due to the matrices involved in the GEEs the running time for each model, when there were a large number of time points (as in the Emerald data set), is quite large and thus producing the final fitted model is very time consuming.

The Tweedie distribution used with GEEs to model rainfall appears to be a suitable alternative to modelling rainfall amounts and further research into this area of modelling may be very viable. Possibilities for further research include,

- Examining different timescales, for example daily. The Tweedie family of distributions is not only applicable to the monthly timescale but can just as easily be applied to yearly or daily timescales (Dunn & Lennox [24]);
- Examining other correlation structures and comparing them with the AR(1) structure used in this dissertation;
- Looking at including more than two sites in the rainfall model;
- Modelling other locations (world-wide);

- Investigating the possibility of using alternative covariates;
- Producing further model validation techniques and validating the multi-site model;
- Fitting more than one model for each month, season or other periods of time to incorporate the extreme variability of rainfall.

Appendix A

Appendix

A.1 Deriving a formula for β

To find an estimate for β in the GEE case, the GEE estimator equation used is:

$$\sum_{i=1}^{N} D_i^T V_i^{-1}(y_i - \mu_i) = 0$$

with

- $V_i = A_i^{\frac{1}{2}} \times R_i(\alpha) \times A_i^{\frac{1}{2}}$
- D_i = matrix of partial derivatives of μ and β

It is known that $D_i^T = \frac{\partial \mu_i}{\partial \beta_j}$ and thus it can be said that $\mu_i = D_i^T \beta$ after performing integration and setting the integration constant to 0. Therefore the GEE estimator equation now becomes,

$$\sum_{i=1}^{N} D_i^T V_i^{-1}(y_i - D_i^T \hat{\beta}) = 0$$

An estimate for β can now be found,

$$\sum_{i=1}^{N} D_i^T V_i^{-1} y_i = \sum_{i=1}^{N} D_i^T V_i^{-1} D_i^T \beta$$

Finally,

$$\hat{\beta} = \sum_{i=1}^{N} (D_i^T V_i^{-1} D_i^T)^{-1} \sum_{i=1}^{N} (D_i^T V_i^{-1} y_i)$$

Appendix B

Code for producing a GEE

B.1 Single-site code

The following code was written to produce a GEE in which the response variable's distribution was a Tweedie distribution. The initial set up of the data is seen first (data from Emerald), followed by the implementation of the GEE and finally a series of tests for GEEs can be seen last.

```
#################################################################
# Initial setting of data
# Load libraries that are needed to perform calculations
library(stats)
library(statmod)
library(tweedie)

# Set the directory
setwd("//usq/sciences/home/swan/My Documents/Thesis/data")
rm(list=ls())   # remove any previous lists

# Load the data
emerald<-read.table("emeraldall.txt", header=TRUE)

# Define the cos and sin terms - annual terms
emerald$cos=NULL
emerald$cos=cos(2*pi*emerald$month/12)

emerald$sin=NULL
emerald$sin=sin(2*pi*emerald$month/12)
```

```
# Define the cos1 and sin1 terms - 6 monthly terms
emerald$cos1=NULL
emerald$cos1=cos(4*pi*emerald$month/12)

emerald$sin1=NULL
emerald$sin1=sin(4*pi*emerald$month/12)

# Define the factors soiphase and month
emerald$soiphase=factor(emerald$soiphase)
emerald$month=factor(emerald$month)

# Define seasons (Summer = 1, Autumn = 2, Winter = 3,
# Spring = 4)
emerald$season=NULL
emerald$season[emerald$month==1|emerald$month==2|emerald$month==12]=1
emerald$season[emerald$month==3|emerald$month==4|emerald$month==5]=2
emerald$season[emerald$month==6|emerald$month==7|emerald$month==8]=3
emerald$season[emerald$month==9|emerald$month==10|emerald$month==11]=4
emerald$season=factor(emerald$season)

# Define oddseasons
emerald$oddseasons <- array( dim=length(emerald$season))
emerald$oddseasons[emerald$month %in% (4:9)] <- "Dry"
emerald$oddseasons[emerald$month %in% c(1,12)] <- "Wet"
emerald$oddseasons[emerald$month %in% c(2,3,10,11)] <-"Transition"
emerald$oddseasons=factor(emerald$oddseasons)

# Define Indicators as to where or not rain occurred in
# previous months
emerald1 <- emerald[13:length(emerald$rain),]
emerald1$rain1 <- emerald$rain[12: (length(emerald$rain)-1)]
emerald1$rain2 <- emerald$rain[11: (length(emerald$rain)-2)]
emerald1$rain3 <- emerald$rain[10: (length(emerald$rain)-3)]
emerald1$rain4 <- emerald$rain[9: (length(emerald$rain)-4)]
emerald1$rain5 <- emerald$rain[8: (length(emerald$rain)-5)]
emerald1$rain6 <- emerald$rain[7: (length(emerald$rain)-6)]
emerald1$rain7 <- emerald$rain[6: (length(emerald$rain)-7)]
emerald1$rain8 <- emerald$rain[5: (length(emerald$rain)-8)]
emerald1$rain9 <- emerald$rain[4: (length(emerald$rain)-9)]
emerald1$rain10 <- emerald$rain[3: (length(emerald$rain)-10)]
emerald1$rain11 <- emerald$rain[2: (length(emerald$rain)-11)]
```

```
emerald1$rain12 <- emerald$rain[1: (length(emerald$rain)-12)]

# Now indicators only
emerald1$ind1 <- emerald1$rain1>0
emerald1$ind2 <- emerald1$rain2>0
emerald1$ind3 <- emerald1$rain3>0
emerald1$ind4 <- emerald1$rain4>0
emerald1$ind5 <- emerald1$rain5>0
emerald1$ind6 <- emerald1$rain6>0
emerald1$ind7 <- emerald1$rain7>0
emerald1$ind8 <- emerald1$rain8>0
emerald1$ind9 <- emerald1$rain9>0
emerald1$ind10 <- emerald1$rain10>0
emerald1$ind11 <- emerald1$rain11>0
emerald1$ind12<- emerald1$rain12>0
emerald1$ind1.2 <- emerald1$ind1 & emerald1$ind2
emerald1$ind1.2.3 <- emerald1$ind1 & emerald1$ind2 & emerald1$ind3


# Set the up to 1992 as the estimation data and 1993 onwards as the
# validation set
estimation<-emerald1[emerald1$year<1993,]
validation<-emerald1[emerald1$year>=1993,]

#Attach all the data and perform calculations on the estimation set
attach(estimation)

#################################################################
#FITTING THE GENERALISED ESTIMATING EQUATION

# STEP 1 - Compute an inital estimate of beta using glm
# metholodogy.  Calculate "p", to be used in the variance function
# of the Tweedie distribution using profile likelihood function.
power=tweedie.profile(rain~month+soiphase+ind1.2,
        p.vec=seq(1.5,1.85,length=10),
        do.plot=TRUE, smooth=TRUE,do.ci=TRUE, method="interpolation")
p=power$p.max

# Fitting a Tweedie model to this data, with "p" value found
# using the profile likelihood function and a log link function
glmmodel<-glm(rain~month+soiphase+ind1.2,
```

```
    family=tweedie(var.power=p, link.power=0),x=TRUE)

# Initialize values - variables used in the first repetition
fits<-glmmodel$fitted.values
beta=glmmodel$coefficients

phi = power$phi.max
n=length(rain)

# Let "r" be the number of beta values (number of covariates +1)
r=glmmodel$rank

# Set the variables to be used in the convergence criteria
dev=sum(tweedie.dev(rain,fits,p))
devold=100*dev
epsilon = 1e-8


#################################################################
# Create the recusive (repeating steps 2 to 5), using a
# convergence criteria.
# Create a new set of fitted values for the new beta values found

# Convergence criteria
while (abs(dev - devold)/(0.1 + abs(dev)) > epsilon) {

#_____________________________________________________________
# Step 2 - Compute the Pearson's residuals for the model
p.residuals=(rain-fits)/sqrt(fits^p)

#_____________________________________________________________
# Step 3a - Calculate alpha

# Calculate the new phi value and alpha
phi<-sum(p.residuals^2)/(n-r)

# Initialize alpha
alpha=NULL

# Obtain alpha value
alpha=sum(p.residuals[1:(n-1)]*p.residuals[2:n])
```

```
alpha=((phi)*alpha)/(n*(n-r))

# Step 3b - Calculate R using the alpha values found in
# step 3a (using AR(1))
index<-seq(0,n-1,by=1)
longindex<-c(seq(n-1,1,by=-1),index)

# Calculate R
i=0
R=matrix(nrow=n,ncol=n)
while(i<n){
  R[i+1,]=alpha^longindex[(n-i):(2*n-i-1)]
  i=i+1
  }

#-----------------------------------------------------------------
# Step 4 - Calculate an estimate of the covariance matrix V
# using R found in step 3b.

# Calculate A:
A=diag(fits)^(p/2)

# Calculate V:
V = (A %*% R %*% A)

#-----------------------------------------------------------------
# Step 5 - Find an updated version of beta
# Firstly find (partial mu / partial beta), let
# (partial mu/partial beta) be matrix "D"

xmat=glmmodel$x

D = matrix(nrow=n,ncol=r)

#Add the values to matrix "D"
for(i in (1:r)){
  D[,i]=fits*xmat[,i]
  }

# To find matrix beta(r+1) use the following notations
# beta=beta+inverse(C)*B,
```

```
# where C = transpose(D)*inverse(V)*D and
# B = transponse(D)*inverse(V)*(actual-fitted)

# Firstly find C
C=t(D) %*% solve(V) %*% D

# Find B
B=t(D) %*% solve(V) %*% (rain-fits)

beta=beta + (solve(C) %*% B)

# Fit the new values of dev, devold and fits for use in the
# covergence criteria
fits<-exp(t(beta) %*% t(xmat))
fits<-as.vector(fits)
devold<-dev
dev<-sum(tweedie.dev(rain,fits,p))

}

#####################################################################
# DIAGNOSTICS

# Calculate QICu
# Calculate the quasi-likelihood first
quasi<-sum((rain*fits^(1-p)/(1-p))-((fits^(2-p))/(2-p)))

# Next calculate the QICu which is to find the best covariates to use
qicu<-(-2*quasi)+(2*r)

#____________________________________________________________
# Calculate the Marginal R squared
marginal=(1/n)*sum(rain)            # marginal component of R^2
top=sum((rain-fits)^2)              # numerator of R^2
bottom=sum((rain-marginal)^2)       # denominator of R^2

R2 = 1 - (top / bottom)

#____________________________________________________________
# Calculate the Wald-Wolfowitz run test to detect if the model
# is adequate and residuals are random.
```

```
# Calculate the raw residuals
residuals=rain-fits

# Initialise the values to use
run=NULL
nn=0
np=0
j=1

# Start the test
while(j<=n){
  if(residuals[j]<=0){
     run[j]=-1
     nn=nn+1}
  if(residuals[j]>0){
     run[j]=1
     np=np+1}
  j=j+1
  }

# Find the components E(T) and V(T) needed in the randomness test
ET=(2*np*nn)/(np+nn)+1
VT=(2*np*nn*(2*np*nn-np-nn))/((np+nn)^2*(np+nn-1))

# Find the total number of observed runs in the sequence
T=0
j=1

while(j<=(n-1)){
  if(run[j]!=run[j+1]){
     T=T+1}

j=j+1
}

T=T+1

# Find the test statistic W
W=(T-ET)/sqrt(VT)
```

```
#Print out all the relevant information
output1<-data.frame(Diagnostic=c("alpha","QICu","R2","W"),
    Data=c(alpha,qicu,R2,W))
output2<-data.frame(BetaValues=c(beta),
    Names=c(names(glmmodel$coefficient)))

print(output1)
print(output2)

#################################################################
# RESIDUALS PLOT FOR MODEL (1 + WS)

# Plot of the Raw residuals
win.graph(width=11,height=7)        # graphic size
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
main="Plot of the raw residuals using the wet-season factor")
abline(0,0)             # add a horizontal line at 0
dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/rawresidualsone.pdf")


# Plot of Pearson Residuals versus linear predictor (eta=log(mu))
win.graph(width=11,height=7)
plot(log(fits),p.residuals,xlab="Linear Predictor", ylab="Pearson
Residuals",main="Plot of pearson residuals versus linear predictor
    (wet-season)")
dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/linearresidualsone.pdf")


#################################################################
# RESIDUALS PLOT FOR MODEL (1 + M + P + IND1.2)

# Plot of the Raw residuals
win.graph(width=11,height=7)        # graphic size
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
main="Plot of the raw residuals using the month factor")
abline(0,0)                 # add a horizontal line at 0
dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/rawresidualsmonth.pdf")
```

```
# Plot of Pearson Residuals versus linear predictor (eta=log(mu))
win.graph(width=11,height=7)
plot(log(fits),p.residuals,xlab="Linear Predictor", ylab="Pearson
Residuals",main="Plot of pearson residuals versus linear predictor
    (month)")
dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/linearresidualsmonth.pdf")

################################################################
# PREDICTED VALUES for (1 + M + P + IND1.2)
# Plot of predicted values for Emerald
win.graph(width=12,height=6)        # graphic size

# A time series plot of the amount of rain recorded during each
# dry and wet month and a plot of the predicted values for the
# amount of rain per month

# Observed rainfall:
plot(ts(rain,start=c(1889,1),frequency=12),
    plot.type="single",col="blue", xlab="Year",ylab="Amount of rain (mm)",
    main="Emerald Monthly Rainfall")
abline(h=c(0,100,200,300,400,500),v=c(1900,1920,1940,1960,1980,2000),
    lty=2,lwd=.1,col="gray",las=2)

#Predicted Rainfall:
emerald.fitted<-ts(fits,start=c(1889,1),frequency=12)
points(emerald.fitted,type="l",col="red")           #Add to the plot

dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/emeraldobsandpredict.pdf")


################################################################
# Finding a suitable link function
glmmodel<-glm(rain~month+soiphase+ind1.2,
    family=tweedie(var.power=p, link.power=0),x=TRUE)   # Logarithm

glmmodel.other<-glm(rain~month+soiphase+ind1.2,
    family=tweedie(var.power=p),x=TRUE)                 # Canonical
```

```
#Deviances
glmmodel$deviance
glmmodel.other$deviance

#Df Residuals
glmmodel$df.residual
glmmodel.other$df.residual

#################################################################
# Normal probability plot for the model (1 + M + P + IND1.2)
# First print the profile log-likelihood plot
win.graph(width=6,height=6)
power=tweedie.profile(rain~month+soiphase+ind1.2,
    p.vec=seq(1.5,1.85,length=20),
    do.plot=TRUE, smooth=TRUE,do.ci=TRUE, method="interpolation")
dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/logplotone.pdf")

p=power$p.max
glmmodel<-glm(rain~month+soiphase+ind1.2,family=tweedie(var.power=p,
    link.power=0),x=TRUE)

win.graph(width=6,height=6)                 # graphic size
quantile=qres.tweedie(glmmodel)             # Quantile residuals
qqnorm(quantile, main = "Normal probability plot \n for one site model",
    xlab="Standard Normal Quantiles", ylab="Quantile Residuals")
qqline(quantile)                            # Normality line
dev.print(pdf,"C:/Documents and Settings/owner/My Documents/Taryn
    = uni/Thesis/22 June 06/Pictures/quantileonemonth.pdf")
```

B.2 Multi-site code

The following code was written to produce a GEE in which the response variable's distribution was a Tweedie distribution and multiple sites are modelled concurrently. The initial set up of the data is seen first (data from Toowoomba and Gatton), followed by the implementation of the GEE and finally a series of tests for GEEs can be seen last.

```
#################################################################
```

```
# Initial setting of data for Toowoomba and Gatton
#Load libraries that are needed to perform calculations
library(stats)
library(statmod)
library(tweedie)

# Remove all previous data
rm(list=ls(all=TRUE))

# Read in the data, Jan 1980 until Dec 2001, totalling 22 years.
gatton<-read.table("gatton.dat", header=TRUE)
toow<-read.table("toowoomba.dat",header=TRUE)
double<-read.table("double.txt",header=TRUE)

# Assume a '1' to Gatton data and '2' to Toowoomba data
gatton$id<-seq(1,1,264)
toow$id<-seq(2,2,264)
double$id[1:264]<-gatton$id
double$id[265:528]<-toow$id

# Convert the Daily Rainfall amounts to Monthly Rainfall amounts
# GATTON
double$rain=NULL
count=1
i=1980
while(i<=2001){
double$rain[count]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Jan"])
double$rain[count+1]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Feb"])
double$rain[count+2]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Mar"])
double$rain[count+3]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Apr"])
double$rain[count+4]=sum(gatton$rain[gatton$Year==i&gatton$Month=="May"])
double$rain[count+5]=sum(gatton$rain[gatton$Year==i&gatton$Month=="June"])
double$rain[count+6]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Jul"])
double$rain[count+7]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Aug"])
double$rain[count+8]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Sep"])
double$rain[count+9]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Oct"])
double$rain[count+10]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Nov"])
double$rain[count+11]=sum(gatton$rain[gatton$Year==i&gatton$Month=="Dec"])
i=i+1
count=count+12
}
```

```
# TOOWOOMBA
count=265
i=1980
while(i<=2001){
double$rain[count]=sum(toow$rain[toow$Year==i&toow$Month=="Jan"])
double$rain[count+1]=sum(toow$rain[toow$Year==i&toow$Month=="Feb"])
double$rain[count+2]=sum(toow$rain[toow$Year==i&toow$Month=="Mar"])
double$rain[count+3]=sum(toow$rain[toow$Year==i&toow$Month=="Apr"])
double$rain[count+4]=sum(toow$rain[toow$Year==i&toow$Month=="May"])
double$rain[count+5]=sum(toow$rain[toow$Year==i&toow$Month=="June"])
double$rain[count+6]=sum(toow$rain[toow$Year==i&toow$Month=="Jul"])
double$rain[count+7]=sum(toow$rain[toow$Year==i&toow$Month=="Aug"])
double$rain[count+8]=sum(toow$rain[toow$Year==i&toow$Month=="Sep"])
double$rain[count+9]=sum(toow$rain[toow$Year==i&toow$Month=="Oct"])
double$rain[count+10]=sum(toow$rain[toow$Year==i&toow$Month=="Nov"])
double$rain[count+11]=sum(toow$rain[toow$Year==i&toow$Month=="Dec"])
i=i+1
count=count+12
}

#Define location predictors - longitude, latitude and altitude
#LONGITUDE
long1<-seq(152.28,152.28,length=264)    #Gatton
long2<-seq(151.95,151.95,length=264)    #Toowoomba
double$long[1:264]<-long1
double$long[265:528]<-long2

#LATITUDE
lat1<-seq(-27.58,-27.58,length=264) #Gatton
lat2<-seq(-27.55,-27.55,length=264) #Toowoomba
double$lat[1:264]<-lat1
double$lat[265:528]<-lat2


#ALTITUDE
alt1<-seq(674.9,674.9,length=264)   #Gatton
alt2<-seq(93,93,length=264)     #Toowoomba
double$alt[1:264]<-alt1
double$alt[265:528]<-alt2
```

```
#Define the cos and sin terms to be used to model rainfall
double$cos=NULL
double$cos=cos(2*pi*double$month/12)

double$sin=NULL
double$sin=sin(2*pi*double$month/12)

# Define the second lot of sin and cos terms defined by Dunn
# and Lennox

double$cos1=NULL
double$cos1=cos(4*pi*double$month/12)

double$sin1=NULL
double$sin1=sin(4*pi*double$month/12)

#Define the factors - soiphase and month to be used for
#modelling rainfall
double$soiphase=factor(double$soiphase)
double$month=factor(double$month)

#Define seasons where:
#Summer = 1,
#Autumn = 2,
#Winter = 3 and
#Spring = 4

double$season=NULL
double$season[double$month==1|double$month==2|double$month==12]=1
double$season[double$month==3|double$month==4|double$month==5]=2
double$season[double$month==6|double$month==7|double$month==8]=3
double$season[double$month==9|double$month==10|double$month==11]=4
double$season=factor(double$season)

#Define oddseasons, defined as:
#Jan and Dec - 'Wet'
#Feb, March, Oct and Nov - 'Transition'
#April to Sept - 'Dry'
double$oddseasons <- array( dim=length(double$season) )
double$oddseasons[ double$month %in% (4:9) ] <- "Dry"
```

```
double$oddseasons[ double$month %in% c(1,12) ] <- "Wet"
double$oddseasons[ double$month %in% c(2,3,10,11) ] <- "Transition"
double$oddseasons=factor(double$oddseasons)

#Define Indicators as to where or not rain occurred in previous months
double1 <- double[13:length(double$rain),]
double1$rain1 <- double$rain[12: (length(double$rain)-1)]
double1$rain2 <- double$rain[11: (length(double$rain)-2)]
double1$rain3 <- double$rain[10: (length(double$rain)-3)]
double1$rain4 <- double$rain[9: (length(double$rain)-4)]
double1$rain5 <- double$rain[8: (length(double$rain)-5)]
double1$rain6 <- double$rain[7: (length(double$rain)-6)]
double1$rain7 <- double$rain[6: (length(double$rain)-7)]
double1$rain8 <- double$rain[5: (length(double$rain)-8)]
double1$rain9 <- double$rain[4: (length(double$rain)-9)]
double1$rain10 <- double$rain[3: (length(double$rain)-10)]
double1$rain11 <- double$rain[2: (length(double$rain)-11)]
double1$rain12 <- double$rain[1: (length(double$rain)-12)]

# Now indicators only
double1$ind1 <- double1$rain1>0
double1$ind2 <- double1$rain2>0
double1$ind3 <- double1$rain3>0
double1$ind4 <- double1$rain4>0
double1$ind5 <- double1$rain5>0
double1$ind6 <- double1$rain6>0
double1$ind7 <- double1$rain7>0
double1$ind8 <- double1$rain8>0
double1$ind9 <- double1$rain9>0
double1$ind10 <- double1$rain10>0
double1$ind11 <- double1$rain11>0
double1$ind12<- double1$rain12>0
double1$ind1.2 <- double1$ind1 & double1$ind2
double1$ind1.2.3 <- double1$ind1 & double1$ind2 & double1$ind3

double<-double1

# Attach all of the variables to the name "double" so that calculations
# can be performed
attach(double)
```

```
#################################################################
# FITTING THE GENERALISED ESTIMATING EQUATION FOR BOTH GATTON
# AND TOOWOOMBA. Set the initial variables of K (K is 2 as there
# are two places and n_i (write as n rather than n_i where it is
#the number of time points per place
K = 2
n = length(double$rain)/K
N = K*n

# STEP 1 - Compute an inital estimate of beta using glm
# metholodogy.  Calculate "p", to be used in the variance function
# of the Tweedie distribution using profile likelihood function.
power=tweedie.profile(rain~month+soiphase+ind1.2,
        p.vec=seq(1.5,1.85,length=10),
        do.plot=TRUE, smooth=TRUE,do.ci=TRUE, method="interpolation")
p=power$p.max

# Fitting a Tweedie model to this data, with "p" value found
# using the profile likelihood function and a log link function
glmmodel<-glm(rain~month+soiphase+ind1.2,
    family=tweedie(var.power=p, link.power=0),x=TRUE)


# Initialize values - variables used in the first repetition,
# where fits is the fitted values obtained by running the glm.
# The variable beta contains the values of the regression parameters
#obtained by running the glm on the data.
fits=matrix(nrow=K,ncol=n)
fits[1,]<-glmmodel$fitted.values[1:n]
fits[2,]<-glmmodel$fitted.values[(n+1):N]

beta=glmmodel$coefficients

# Convert the "rain1" values into a matrix form to indicate that
# the values are from different locations
rain1<-matrix(nrow=K,ncol=n)
rain1[1,]<-rain[1:n]
rain1[2,]<-rain[(n+1):N]

# Use an initial estimate of phi by using the phi obtained when fitting
# the glm model and the profile likelihood function
```

```
phi = power$phi.max

#Let "r" be the number of beta values (number of covariates +1)
r = glmmodel$rank

# Find the deviance
dev = sum(tweedie.dev(rain1,fits,p))
devold = 100*dev
epsilon = 1e-8


#####################################################################
# Create the recusive (repeating steps 2 to 5), using a
# convergence criteria
# Create a new set of fitted values for the new beta values found

while (abs(dev - devold)/(0.1 + abs(dev)) > epsilon) {

#____________________________________________________________
# Step 2 - Compute the Pearson's residuals for the model
p.residuals=(rain1-fits)/sqrt(fits^p)

#____________________________________________________________
# Step 3a - Calculate alpha

# Calculate the new phi value and alpha by firstly finding the new
# phi value
phi<-sum(p.residuals^2)/(N-r)

# Obtain alpha value
alpha=NULL
k=1
while(k<n){
    alpha[k]=p.residuals[1,k]*p.residuals[1,k+1]+
        p.residuals[2,k]*p.residuals[2,k+1]
    k=k+1
    }

alpha=sum(alpha)
alpha=(phi*alpha)/(n*(N-r))
```

```
# Step 3b - Calculate R using the alpha values found in step 3a
# using an AR(1)
index<-seq(0,n-1,by=1)
longindex<-c(seq(n-1,1,by=-1),index)

i=0
R=matrix(nrow=n,ncol=n)
while(i<n){
  R[i+1,]=alpha^longindex[(n-i):(2*n-i-1)]
  i=i+1
  }

#_____________________________________________________________
# Step 4 - Calculate an estimate of the covariance matrix V using
# R found in step 3b.

#Calculate A:
A1=diag(fits[1,])^(p/2)
A2=diag(fits[2,])^(p/2)

#Calculate V:
V1 = (A1 %*% R %*% A1)
V2 = (A2 %*% R %*% A2)

#_____________________________________________________________
# Step 5 - Find an updated version of beta
# Firstly find (partial mu / partial beta), let
# (partial mu/partial beta) be matrix "D"

xmat=glmmodel$x

D1 = matrix(nrow=n,ncol=r)
D2 = matrix(nrow=n,ncol=r)

#Add the values to matrix "D"
for(i in (1:r)){
  D1[,i]=t(fits[1,])*xmat[,i][1:n]
  }

for(i in (1:r)){
  D2[,i]=t(fits[2,])*xmat[,i][(n+1):N]
```

```
  }

# To find matrix beta(r+1) use the following notations
# beta=beta+inverse(C)*B, where C = transpose(D)*inverse(V)*D and
# B = transponse(D)*inverse(V)*(actual-fitted)

# Firstly find C
C1=t(D1) %*% solve(V1) %*% D1
C2=t(D2) %*% solve(V2) %*% D2

# Find B
B1=t(D1) %*% solve(V1) %*% (rain[1:n]-fits[1,])
B2=t(D2) %*% solve(V2) %*% (rain[(n+1):N]-fits[2,])

# As per the formula, add together B1 and B2 and C1
# and C2 to get a final answer for beta
C=C1+C2
B=B1+B2

beta=beta + (solve(C) %*% B)

#Fit the new values of dev, devold and fits
newfits<-exp(xmat %*% beta)
fits[1,]<-newfits[1:n]
fits[2,]<-newfits[(n+1):N]
devold<-dev
dev<-sum(tweedie.dev(rain1,fits,p))

j=j+1

}

#####################################################################
# DIAGNOSTICS
# Calculate QICu to find the best set of covariates to use.
# Firstly calculate the quasi-likelihood
quasi<-sum((rain1*fits^(1-p)/(1-p))-((fits^(2-p))/(2-p)))

#Next calculate the QICu
qicu<-(-2*quasi)+(2*r)
```

```
#####################################################################
# Calculate the Marginal R squared
fits1=rain
fits1[1:n]=fits[1,]
fits1[(n+1):N]=fits[2,]

marginal=(1/N)*sum(rain)

top=sum((rain-fits1)^2)                          # numerator
bottom=sum((rain-marginal)^2)                    # denominator

R2 = 1 - (top / bottom)

#####################################################################
# Calculate the Wald-Wolfowitz run test to detect if the model
# is adequate and residuals are random.

residuals=rain-fits1
run=NULL
nn=0
np=0
j=1
while(j<=N){
  if(residuals[j]<=0){
     run[j]=-1
     nn=nn+1}

  if(residuals[j]>0){
     run[j]=1
     np=np+1}

  j=j+1
  }

ET=(2*np*nn)/(np+nn)+1

VT=(2*np*nn*(2*np*nn-np-nn))/((np+nn)^2*(np+nn-1))

#Find the total number of observed runs in the sequence
T=0
j=1
```

```
while(j<=(N-1)){
  if(run[j]!=run[j+1]){
     T=T+1}

j=j+1
}
T=T+1

# Find the test statistic W
W=(T-ET)/sqrt(VT)

#####################################################################
# Print out all the relevant information
output1<-data.frame(Diagnostic=c("alpha","QICu","R2","p","CI.l",
    "CI.u"),Data=c(alpha,qicu,R2,p,power$ci[1],power$ci[2]))
output2<-data.frame(BetaValues=c(beta),
    Names=c(names(glmmodel$coefficient)))

print(output1)
print(output2)

#####################################################################
# RESIDUALS PLOT FOR MODEL (1 + S + SIN + COS + S:SOI)

# Predicted Values Versus Raw residuals
win.graph(width=11,height=10)       # graphic size
par(mfrow=c(2,1))                   # plots 2 graphs in 1 frame
plot(fits[1,],p.residuals[1,],xlab="Predicted Values for Gatton
    [K=1]", ylab="Pearson Residuals", main="Residuals versus
    Predicted Values")
                                    # residuals from Gatton

plot(fits[2,],p.residuals[2,],xlab="Predicted Values for Toowoomba
[K=2]",ylab="Pearson Residuals")    # residuals from Toowoomba
dev.print(pdf,"H:/My Documents/Thesis/data/Pictures/fittedversusresiduals.pdf

# Plot of the Raw residuals
win.graph(width=11,height=7)        # graphic size
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",
main="Plot of the Raw Residuals")
```

```
abline(0,0)                          # add a horizontal line at 0
dev.print(pdf,"H:/My Documents/Thesis/data/Pictures/rawresiduals.pdf")

# Plot of Pearson Residuals versus linear predictor (eta=log(mu))
win.graph(width=11,height=7)
plot(log(fits1),p.residuals,xlab="Linear Predictor", ylab="Pearson
Residuals",main="Plot of Pearson Residuals versus Linear Predictor")
dev.print(pdf,"H:/My Documents/Thesis/data/Pictures/linearresiduals.pdf")


#################################################################
# PREDICTED VALUES for (1 + S + SIN + COS + S:SOI)
# Plot of predicted values for Toowooomba and Gatton
win.graph(width=6,height=6)     # graphic size

# GATTON
# A time series plot of the amount of rain recorded during each dry
# and wet month and a plot of the predicted values for the amount
# of rain per month

# Observed rainfall:
plot(ts(double$rain[id==1],start=c(1980,1),frequency=12),
    plot.type="single",col="blue", xlab="Year",ylab="Amount of rain (mm)",
    main="Gatton Monthly Rainfall")
abline(h=c(0,100,200,300,400,500),v=c(1980,1985,1990,1995,2000),
    lty=2,lwd=.1,col="gray",las=2)

#Predicted Rainfall:
gatton.fitted<-ts(fits[1,],start=c(1980,1),frequency=12)
points(gatton.fitted,type="l",col="red")            #Add to the plot

dev.print(pdf,"H:/My Documents/Thesis/data/Pictures/gattonobsandpredict.pdf")

# TOOWOOMBA
# A time series plot of the amount of rain recorded during each
# dry and wet month and a plot of the predicted values for the amount
# of rain per month

#Observed Rainfall:
plot(ts(double$rain[id==2],start=c(1980,1),frequency=12),
    plot.type="single",col="blue",xlab="Year",ylab="Amount of rain (mm)",
```

```
    main="Toowoomba Monthly Rainfall")
abline(h=c(0,100,200,300,400,500),v=c(1980,1985,1990,1995,2000),
    lty=2,lwd=.1,col="gray",las=2)

#Predicted Rainfall:
toow.fitted<-ts(fits[2,],start=c(1980,1),frequency=12)
points(toow.fitted,type="l",col="red")          # Add to the plot

dev.print(pdf,"H:/My Documents/Thesis/data/Pictures/toowobsandpredict.pdf")

#####################################################################
# Finding a suitable link function
glmmodel<-glm(double$rain~season+sin+cos+season:soi,
    family=tweedie(var.power=p, link.power=0),x=TRUE)   # Logarithm

glmmodel.other<-glm(double$rain~season+sin+cos+season:soi,
    family=tweedie(var.power=p),x=TRUE)                 # Canonical

#Deviances
glmmodel$deviance
glmmodel.other$deviance

#Df Residuals
glmmodel$df.residual
glmmodel.other$df.residual

#####################################################################
# Normal probability plot for the model (1 + S + SIN + COS + S:SOI)
win.graph(width=6,height=6)             # graphic size

quantile=qres.tweedie(glmmodel)             # Quantile residuals
qqnorm(quantile, main = "Normal Probability Plot \n for Two Site Model",
    xlab="Standard Normal Quantiles", ylab="Quantile Residuals")
qqline(quantile)                    # Normality line
dev.print(pdf,"H:/My Documents/Thesis/data/Pictures/quantiletwo.pdf")
```

Bibliography

[1] D. Puride A. Dobson and G. Williams. Application of generalized estimating equations to longitudinal data. Longitudinal Studies, The University of Queensland, 2003.

[2] U. Gromping A. Ziegler, C. Kastner and M. Blettner. The generalized estimation equations in the past ten year: An overview and a biomedical application. April 1996.

[3] G. Ballinger. Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, 7(2):127–150, April 2004.

[4] A. Bardonssy and E. Plate. Space time model for daily rainfall using atmospheric circulation patterns. *Water Resource Reserve*, 28:1247–1259, 1992.

[5] M. Barnsley. Spatial dependence and the semi-variogram. available at http://stress.swan.ac.uk/ mbarnsle/teaching/ envmod04/lectures/variogram.pdf, on 25th February 2005 2004.

[6] J. Beersma and A. Buishand. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Climate Research*, 25(2):121–133, 2003.

[7] T. A. Buishand. Some remarks on the use fo daily rainfall models. *Journal of Hydrology*, 36:295–308, 1978.

[8] T. A. Buishand. The effect of seasonal variation and serial correlation on the extreme value distribution of rainfall data. *Journal of climate and Applied Meteorology*, 24:154–173, 1985.

[9] R. Chandler and H. Wheater. Climate change detection using generalized linear models for rainfall — A case study from the west of Ireland II. Modelling of rainfall amounts on wet days. Research Report 195,

Department of Statistical Science, University College of London, June 1998.

[10] R. Chandler and H. Wheater. Climate change detection using generalized linear models for rainfall — A case study from the west of Ireland I. Preliminary analysis and modelling of rainfall occurrence. Research Report 194, Department of Statistical Science, University College of London, June 1998.

[11] R. E. Chandler. On the use of generalized linear models for interpreting climate variability. Research Report 232, Department of Statistical Science, University College London, February 2003.

[12] R. E. Chandler and H. S. Wheater. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, 38(10), October 2002.

[13] T. J. Chang, M. L. Kavvas, and J. W. Delleur. Daily precipitation modeling by discrete autoregressive moving average processes. *Water Resource Reserve*, 20:565–580, 1984.

[14] Y. Chang. Residual analysis of the generalized linear models for longitudinal data. *Statistics in Medicine*, 19:1277–1293, 2000.

[15] T. Chapman. Stochastic modelling of daily rainfall: The impact of adjoining wet dats on the distribution of rainfall amounts. *Environmental Modelling and Software*, 13:317–324, 1998.

[16] E. H. Chin. Modeling daily precipitation occurrence process with Markov chain. *Water Resources Research*, 13:949–956, 1977.

[17] R. Coe and R. D. Stern. Fitting models to daily rainfall data. *Journal of Applied Meteorology*, 2:1024–1031, February 1982.

[18] M. Crowder. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82(2):407–410, June 1995.

[19] G. Dahmen and A. Ziegler. Generalized estimating equations in controlled clinical trials: Hypothesis testing. *Biometrical Journal*, 46:214–232, 2003.

[20] S. C. Das. The fitting of truncated type III curves to daily rainfall data. *Australian Journal of Physics*, pages 198–304, 1955.

[21] A. J. Dobson. *An introduction to generalized linear models*. Chapman and Hall, London, 2nd edition, 2002.

[22] D. D. Dunlop. Regression for longitudinal data: A bridge from least squares regression. *The American Statistican*, 48:299–303, 1994.

[23] P. Dunn. Tweedie: Tweedie exponential family models. *R package version 1.02*, 2004. http://www.sci.usq.edu.au/staff/dunn/twhtml/home.html.

[24] P. Dunn and S. Lennox. Simultaneous analysis of rainfall occurrence and amounts using power-variance generalized linear models. *Water Resources Research*, Submitted 2006.

[25] P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:1–10, 1996.

[26] P. K. Dunn and G. K. Smyth. Series evaluations of Tweedie exponential dispersion models. *Statistics and Computing*, 2005. Accepted for publication.

[27] Peter Dunn. Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24:1231–1239, May 2004.

[28] M. Durban and C. A. Glasbey. Weather modeling using a multivariate latent gaussian model. *Agricultural and Forest Meteorology*, 109:187–201, June 2001.

[29] A. Feuerverger. On some methods of analysis for weather experiments. *Biometrika*, 66:655–658, 1979.

[30] G. Fitzmaurice. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 51:309–317, 1995.

[31] J. Gill. *Generalized Linear Models - A Unified Approach*. SAGE Publications, Inc., United States of America, 2001.

[32] J. R. Green. A model for rainfall occurrence. *Journal of Royal Statistic Society*, B64:345–353, 1964.

[33] G. K. Grunwald and R. H. Jones. Markov models for time series with mixed distribution. *Environmetrics*, 11:327–339, 2000.

[34] Y. Gyasi-Agyei. Modelling diurnal cycles in point rainfall properties. *Hydrological Processes*, 15:595–608, 2001.

[35] Y. Gyasi-Agyei and G. Willgoose. A hybrid model for point rainfall modelling. Technical report, The University of Newcastle, Australia, Department of Civil, Surveying and Environmental Engineering, 1997.

[36] C. Haan, D. Allen, and J. Street. A markov chain of daily rainfall. *Water Resources Research*, 12(3):443–449, 1976.

[37] D. Hall and T. Severini. Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, 93:1365–1375, 1998.

[38] J. Hardin and J. Hilbe. *Generalized Linear Models and Extensions*. Stata Corporation, United States of America, 2001.

[39] J. W. Hardin and J. M. Hilbe. *Generalized estimating equations*. Chapman and Hall, United States of America, 2003.

[40] M. Harrison and P. Waylen. A note concerning the proper choice for Markov model order for daily precipitation in the humid tropics: A case study in Costa Rica. *International Journal of Climatology*, 20:1861–1872, 2000.

[41] T. Harrold, A. Sharma, and S. Sheather. A nonparametric model for stochastic generation of daily rainfall occurrence. *Water Resources Research*, 39(10):1029–1039, October 2003.

[42] D. Hedeker. *Longitudinal Data Analysis*, chapter 8 : Generalized Estimating Equations, pages 193–228. University of Illinois, Chicago, United States of America, 2005.

[43] A. Heinen. Modelling time series count data: An autoregressive conditonal Poisson model. July 2003.

[44] C. Hicks and C. F. Earl. The validation of simulation models. *Assembly Automation*, 21(3), September 2001.

[45] N. J. Horton and S. R. Lipstiz. Review of software to fit generaized estimating equation regression models. *The American Statistican*, 53:??, 1999.

[46] N.J. Horton and S. R. Lipsitz. Review of software to fit generalized estimating equation regression models. *The American Statistican*, 53, 1999.

[47] R. J. Hyndman. Nonparametric additive regression models for binary time series. Clayton, Victoria, March 1999. Department of Econometrics and Business Statistics, Monash University.

[48] R. J. Hyndman and G. K. Grunwald. Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall. *Australian and New Zealand Journal of Statistics*, 42(2):145–158, 2000.

[49] C. Coffey J. B. Carlin, R. Wolfe and G. C. Patton. Analysis of discrete binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: Prevalence and incidence of smoking in adolescent cohort. *Statistics in Medicine*, 18:2655–2679, 1999.

[50] M. Edwardes J. Hanley, A. Negassa and J. Forrester. Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology*, 157(4):364–375, 2003.

[51] P. Guttory J. Hughes and S. Charles. A non-homogeneous hidden Markov model for precipitation. *Journal of the Royal Statistical Society: Series C (Applied Statistics*, 48(1):15–30, 1999.

[52] S. Fujita J.B. Cologne, R. L. Carter and S. Ban. Application of generalized estimating equations to a study of in vitro radiation sensitivity. *Biometrics*, 49(3):927–934, September 1993.

[53] B. Jørgensen. Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 49:127–162, 1987.

[54] R. W. Katz. Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16:671–676, 1977.

[55] U. Lall, B. Rajagopalan, and K. Tarboton. A nonparametric wet/dry spell model for resampling daily precipitation. *Water Resources Research*, 32(9):2803–2823, September 1996.

[56] K. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, April 1986.

[57] James K. Lindsey. *Applying Generalized Linear Models*. Springer-Verlag New York, Inc., New York, 1997.

[58] J. M Lough. Variations of sea-surface termperatures of North-Eastern Australia an associations with rainfall in Queensland. *International Journal of Climatology*, 12:765–782, 1992.

[59] I. A. Lung and D. D. Grantham. Persistence, runs and recurrence of precipitation. *Journal of Applied Meteorology*, 16:346–358, 1977.

[60] R. Rigby M. Benjamin and M. Stasinopoulos. Modelling exponential family time series data. Statistical Modelling: Proceedings of the 13th International Workshop on Stastical Modelling, 1998.

[61] J. L. McBride. Seasonal relationships between Australian rainfall and the Southern Oscillation. *Monhtly Weather Review*, 111:1998–2004, 1983.

[62] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

[63] C. E McCulloch and S. R. Searle. *Generalized, Linear and Mixed Models*. John Wiley and Sons Inc., Canada, 2001.

[64] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of Royal Stastistical Society. Series A (General)*, 135(3):370–384, 1972.

[65] K. Liang P. Diggle and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press Inc., New York, 1994.

[66] W. Pan. Akaike's information criteria in generalized estimating equations. *Biometrics*, 57:120–125, 2001.

[67] G. Hammer R. Stone and T. Marcussen. Prediction of global rainfall probabilities using phases of the southern oscillation index. *Nature*, 384:252–255, 1996.

[68] B. Rajagopalan and U. Lall. A k-nearest-neighbour simulator for daily precipitation and other weather variables. *Water Resources Research*, 35(10):3089–3101, 1996.

[69] B. Rajagopalan and U. Lall. Nonhomogeneous Markov model for daily precipitation. *Journal of Hydrologic Engineering*, 1(1):33–40, January 1996.

[70] A. W. Robertson. Hidden Markov models for modeling daily rainfall occurrence over Brazil. Technical Report UCI-ICS 03-27, University of California, Irvine, November 2003.

[71] J. Roland and D. A. Woolhiser. Stochastic daily precipitation models 1. a comparison of occurrence models. *Water Resources Research*, 18(5):1451–1459, 1982.

[72] M. A. Sabur and D. D. Andres. *Stochatic and Statistical Methods in Hydrology and Environmental Engineering: Extreme values: Floods and Droughts*, volume 1, chapter - The regionalization of runoff coefficients for prairies of Alberta, pages 341–354. Kluwer Academic Publishers, Dordrecht, 1994.

[73] B. Sanso and L. Guenni. Venezuelan rainfall data analysed by using a bayesian space-time model. *Applied Statistics*, 48(3):345–362, 1999.

[74] R. G. Sargent. Verification and validation of simulation models. In *Proceedings of the 1998 Winter Simulation Conference*, pages 121–130, Syracuse, New York, 1998. Simulation Research Group, Syracuse University.

[75] R. Srikanthan and T. McMahon. Stochastic generation of rainfall and evaporation data. Technical Paper 84, Australian Water Resources Council, Canberra, 1985.

[76] R. Srikanthan and T. A. McMahon. Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, 5(4):653–670, 2001.

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

[78] R. Stone and A. Auliciems. SOI Phase relationships with rainfall in eastern australia. *International Journal of Climatology*, 12:625–636, 1992.

[79] B. Sutradhar. An overview on regression models for discrete longitudinal responses. *Statistical science*, 18(3):377–393, 2003.

[80] B. C Sutradhar and K. Das. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Birometrika*, 86(2):459–465, June 1999.

[81] P. Todorovic and D. Woolhiser. A stochastic model of n-day precipitation. *Agricultural Meteorology*, 26:35–50, 1975.

[82] A. Troup. The southern oscillation. *The Quarterly Journal of the Royal Meteorological Society*, 91:490–506, 1965.

[83] M. C. K. Tweedie. An index which distinguishes between some important exponential families. In *Statistics - Application and New Directions*, pages 579–604, Indian Statistical Institute, Calcutta, 1984. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.

[84] A. T. Walden and P. Guttorp. *Statistics in the Environmental and Earth Sciences*. John Wiley and Sons Inc., New York, 1992.

[85] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, December 1974.

[86] D. S. Wilks. Interannual variability and extreme-value characteristics of several stochastic daily precipition models. *Agricultural and Forest Meteorology*, 93:153–169, 1999.

[87] C. B. Williams. The log series and its application to biological problems. *Journal of Ecology*, 34:253–272, 1947.

[88] K. W. Yau, A. H. Lee, and A. Ng. A zero-augmented gamma mixed model for longitudinal data with many zeros. *Aust. N. Z J. Stat*, 22(2):177–183, 2002.

[89] S. L. Zeger and K. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, March 1986.

[90] B. Zheng. Summarizing the goodness of fit on generalized linear models for longitudinal data. *Statistics in Medicine*, 19:1265–1275, 1988.

[91] C. J. W. Zorn. Generalized estimating equation models for correlated data: a review with applications. *American Journal of Political Science*, 45(2):470–490, April 2001.

[92] W. Zucchini and P. Guttorp. A hidden Markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, August 1991.