

MODELLING RAINFALL USING GENERALIZED  
LINEAR MODELS WHEN THE RESPONSE  
VARIABLE HAS A TWEEDIE DISTRIBUTION

By

Kanithi.K.Isaiah

I56/8084/2006

UNIVERSITY OF NAIROBI  
COLLEGE OF BIOLOGICAL AND PHYSICAL  
SCIENCES  
SCHOOL OF MATHEMATICS

---

A project submitted in  
partial fulfillment of  
the requirements for  
the Degree of  
Master of Science in Social Statistics  
at the  
University of Nairobi

August 2009


University of NAIROBI Library



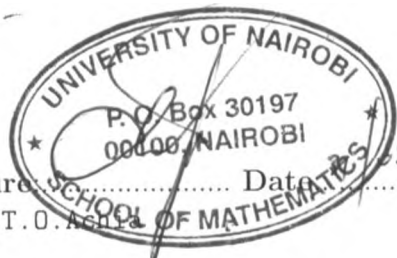
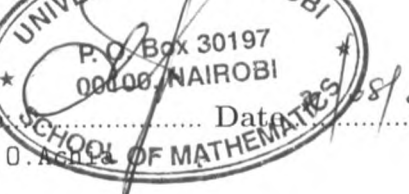
0378971 6

# Declaration

This project is my original work and has not been presented for a degree award in any other University.

Signature:  Date: <sup>TH</sup> 24 AUGUST 2009  
Kanithi.K.Isaiah

This Project has been submitted for examination with my approval as a University supervisor.

  
Signature:  Date: 29/8/09  
Dr. T. O. Asili

## Acknowledgements

The following people have assisted me in the successful completion of this project. I thank you all and will eternally be grateful.

- My supervisor Dr. T.O. Achia, University of Nairobi, who guided and inspired me throughout the course of this project. He also spent long hours in managing the huge dataset that we eventually worked with.
- Dr. Christopher Oludhe of The Department of Meteorology, who provided the data and assisted me with reference materials. He was also readily available for any explanations or information that I required.
- Mr. A. Njogu of Dagoretti Meteorological station who took time to explain to me the various climatological issues and introduced me to Instat+.
- Last but not least are Professor Biamah and Dr. Gichuki of The Department of Environmental Engineering, School of Engineering, University of Nairobi for their support.

## Dedication

This project is dedicated to my family for the continued support that they have given me over the years as I pursued my education. GOD bless YOU ALL.

# Contents

Acknowledgements	ii
Dedication	iii
List of Tables	vi
List of Figures	vii
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 BACKGROUND . . . . .	1
1.2 OBJECTIVES OF THE STUDY . . . . .	2
1.3 SIGNIFICANCE OF THE STUDY . . . . .	2
1.4 BRIEF LITERATURE REVIEW . . . . .	3
1.4.1 Modelling the Occurrence of Rainfall . . . . .	3
1.4.2 Modelling the Amount of Rainfall . . . . .	5
1.4.3 Modelling the Amount and Occurrence of Rainfall . . . . .	6
1.4.4 Generalized Linear Models . . . . .	6
1.4.5 Tweedie Distributions . . . . .	7
<b>2 DATA and METHODOLOGY</b>	<b>9</b>
2.1 Data and Climatology of the Study Area . . . . .	9
2.2 Generalized Linear Models . . . . .	10
2.2.1 Exponential Dispersion Models . . . . .	10
2.2.2 Generalized Linear Models . . . . .	12

2.3	Variance Functions . . . . .	12
2.4	Power-Variance(Tweedie)Distributions . . . . .	13
2.4.1	Series Evaluation . . . . .	15
2.4.2	Inversion of the Cumulant Generating Function . . . . .	18
2.4.3	Saddlepoint Approximation of Tweedie Densities . . . . .	19
2.4.4	Tweedie Distributions and the Quasi-Likelihood	19
2.5	Estimation of Parameters . . . . .	20
2.5.1	Estimation of $\beta$ . . . . .	20
2.5.2	Estimation of $\phi$ . . . . .	22
2.5.3	Estimation of p . . . . .	23
2.6	Diagnostic Testing . . . . .	23
2.6.1	Deviance . . . . .	23
2.6.2	Residuals . . . . .	25
<b>3</b>	<b>APPLICATION OF THE MODEL</b>	<b>27</b>
3.1	Tweedie Rainfall Model . . . . .	27
3.2	Software . . . . .	30
3.3	Missing Data . . . . .	30
3.4	Diagnostic Testing . . . . .	30
<b>4</b>	<b>RESULTS</b>	<b>31</b>
<b>5</b>	<b>CONCLUSION</b>	<b>45</b>

# List of Tables

2.1	Characteristics of exponential dispersion models .	13
2.2	Summary of Tweedie EDMs . . . . .	16
4.1	<b>A statistical summary of the monthly rainfall data for each month . . . . .</b>	<b>32</b>
4.2	Summary of rainfall data . . . . .	33
4.3	<b>January results . . . . .</b>	<b>36</b>
4.4	<b>February results . . . . .</b>	<b>39</b>
4.5	<b>September results . . . . .</b>	<b>40</b>
4.6	<b>S8635000 results . . . . .</b>	<b>42</b>
4.7	<b>S8641000 results . . . . .</b>	<b>42</b>
4.8	<b>S8840000 results . . . . .</b>	<b>43</b>
4.9	<b>S8935104 results . . . . .</b>	<b>43</b>
4.10	<b>S8937035 results . . . . .</b>	<b>44</b>

# List of Figures

2.1	Some Tweedie density functions . . . . .	16
2.2	A log-likelihood plot . . . . .	24
4.1	Histograms for some of the rainfall stations . . . . .	34
4.2	Plot of monthly rainfall amounts . . . . .	35
4.3	Normal probability plot S8635000 . . . . .	37
4.4	Normal probability plot S9135001 . . . . .	38
4.5	Normal probability plot S89338001 . . . . .	38



# Chapter 1

## INTRODUCTION

### 1.1 BACKGROUND

Rainfall in Kenya is characterized by a bimodal distribution where the country experiences long rains between March and April and short rains between October and December. This bimodal rainfall distribution has highly unpredictable occurrence and intensity. The uneven distribution of rainfall exposes agricultural enterprises to a range of mild to severe inter and intra-seasonal dry and wet spells.

Further it has been noted that considerable effort has been devoted to the collection of rainfall data, but little effort has been devoted to its analysis. This has contributed to the lack of knowledge regarding the trends of rainfall occurrence and hence constrained the possibility of solving perennial environmental and agricultural problems.

It is therefore necessary to understand as completely as possible the causes of precipitation variability as rainfall is of considerable importance to a wide range of human activities.

## **1.2 OBJECTIVES OF THE STUDY**

The major objective of this study is to demonstrate an alternative method of analysing rainfall records, which produces results that are directly applicable in the day to day activities and that are readily comprehensible to the various users of rainfall information.

The specific objectives include:

1. determination of an appropriate statistical distribution for the monthly occurrence and amounts of rainfall.
2. assessing the predictability of the characteristics of monthly rainfall.

## **1.3 SIGNIFICANCE OF THE STUDY**

Understanding of the occurrence processes governing rainfall is of considerable importance to a wide range of human activities.

This study is aimed at enhancing our understanding of the spatial and temporal characteristics of wet and dry rainfall spells in order to develop tools that can reduce vulnerability of the agricultural and other rain-dependent sectors to the negative impacts of extreme rainfall events.

The inter-annual and intra-annual variability of rainfall is the key determinant of success in agriculture. This is however complicated by the variability and unpredictability of the lengths and numbers of wet and dry spells in any month, season or year. This uneven distribution of rainfall exposes agricultural enterprises to a range of mild to severe inter and intra-seasonal dry and wet spells. The knowledge of dry -wet behavior will play as a guide to the type of crops that are viable in a particular location. It will also be indispensable in determining the start of the growing season and the harvesting period.

The knowledge of wet and dry spells is essential for defining appropriate domestic, industrial and urban water harvesting plans that maximize rainwater tapping and minimize rainwater losses while at the same time indicating potential for water storage.

Finally this will assist in improving the accessibility of weather information to users and thereby increase the probability of widening the scope of usage.

## 1.4 BRIEF LITERATURE REVIEW

One of the challenges researchers face while examining rainfall is its considerable variation from year to year. The other challenges encountered include the fact that rainfall is a somewhat skewed variable and that it is continuous at exact zero. 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. The second component focuses on the actual rainfall amount once a rainfall event has occurred.

### 1.4.1 Modelling the Occurrence of Rainfall

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

## 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. Markov Chains are also popular because of their largely non-parametric nature, ease of interpretability, and their well developed literature.

Whereas first order models have been studied extensively, research has also focused on higher orders. Gabriel and Neumann (1962), fitted a first order Markov Chain to Tel Aviv rainfall data. Katz (1981) studied zero, first, and second order models. A twelve order chain has also been fitted to hourly rainfall data. Hybrid Markov Chain models have also been developed, where wet spells are modelled as a first order but higher orders are used for dry spells. Other studies have also examined specific locations to test the use of Markov Chains of different orders at different locations.

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. Higher order Markov Chain models often have a lack of parsimony.

Markov Chain Models are however limited in their ability to efficiently model the amounts of rainfall. One method 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 (Lana and Burgueno, 1997). However this technique provides only limited information and is not efficient enough when dealing with extremely important management decisions such as crop growth.

### Alternating Renewal Process

This process considers a sequence of alternating wet and dry spells of varying length, with each spell having an assumed distribution. The type of distributions used to model the length of wet and dry spells as alternating renewal processes include: logarithmic series, truncated negative binomial and the truncated geometric distribution.

Other methods developed to model the occurrence of rainfall include the following; a mixture of geometric and negative binomial distributions and an autoregressive conditional Poisson model which deals with issues of discreteness, over-dispersion and correlation within the data. Time series models have been used to model rainfall. These models ensure that temporal dependence is included in the model. A type of time series model used to model occurrence of rainfall is the two-state discrete autoregressive moving average (DARMA).

#### 1.4.2 Modelling the Amount of Rainfall

Rainfall amount (daily intensity) is a continuous distribution. This type of data is usually modelled using a parsimonious member of the exponential family that fits the given data best. As rainfall is skewed to the right, distributions that follow this same pattern and are skewed to the right have proven to be the most useful, with the gamma distribution being the most commonly used.

Other distributions used to model the amount of rainfall include the exponential distribution which is a special form of the gamma distribution with  $\mu = 1$  and a mixed exponential distribution, which is a mixture of two different exponential distributions. Generalized Autoregressive Moving Average model (GARMA) have been used to model non-normal situations like rainfall. Chandler and Wheater (2002) suggest 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.

The methods described above only monitor either the amount of rainfall or the occurrence of rainfall and not both the occurrence and amount. A model incorporating both amount and occurrence of rainfall would be very desirable indeed.

### **1.4.3 Modelling the Amount and Occurrence of Rainfall**

Several researchers have tried to model the occurrence and amount of rainfall simultaneously. Rajagopalan and Lall (1996) used a multi-state Markov Chain which treated rainfall as a mixed discrete and continuous variable and the probabilities are used to model the dependence structure.

Grunwald and Jones (2000) used 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 (2000) 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.

However, in this model, the functions and parameters are estimated separately. The occurrence distribution is estimated first followed by the estimation of the intensity distribution.

### **1.4.4 Generalized Linear Models**

Generalized Linear Models have been used to compute data that has high levels of variability such as rainfall. Coe and Stern (1982) used GLMs to model rainfall data and found it superior to non-stationary Markov Chains.

Wheater (2002) used logistic regression model, a form of a GLM

to model the binary series of wet and dry days as they were easy to fit and interpret. Chandler and Wheater(1998) fitted a gamma-based GLM to model rainfall on wet days.

GLMs also provide a flexible and rigorous framework that are able to deal with the high levels of variability such as in rainfall data. The GLM approach has also proven to be a very powerful tool for interpreting historical rainfall records.

### 1.4.5 Tweedie Distributions

As rainfall process involves both discrete(rainfall=0mm) and continuous(rainfall>0mm) parts,two separate models have previously been fitted and the information from the two models combined in order to provide a summary of the rainfall process. The Tweedie distribution however is able to combine both aspects to provide one complete rainfall process. They thus have the potential to allow improved rainfall models to be developed. This will result in a more accurate, reliable and practical model that can be incorporated into other areas such as crop growth systems.

Tweedie distributions are based upon Generalized Linear Models and are classified according to their variance. The properties that make the tweedie distributions suitable for modelling rainfall include:

1. The tweedie distributions belong to the exponential family of distributions and form a part of the larger group of Generalized Linear Models
2. Are simple and logical
3. The tweedie distributions provide a mechanism in which finer-scale structures can be understood through courser-scale data

Apart from modelling rainfall, Tweedie distributions have been used in Actuarial studies (B Jorgensen and Paes de Souza MC, 1994 and G K. Smyth and B Jorgensen, 1999), assay analysis, modelling time spent splicing telephone cables (J A. Nelder, 1994), modelling money spent on hiring outside labour (B Jorgensen, 1987), ecological studies, analysis of medical data, analysis of alcohol consumption by british teenagers (Gilchrist, Robert and Drinkwater, Denise, 1999), among others.



## Chapter 2

# DATA and METHODOLOGY

The methods that are adapted to investigate the various characteristics of rainfall to achieve the objectives stated include;

### 2.1 Data and Climatology of the Study Area

Kenya lies approximately within the latitude  $1^{\circ}00'N$  and longitude  $38^{\circ}00'E$  covering  $583,670 \text{ km}^2$ .

It is hot and humid at the coast, temperate in the inland and very hot and dry in the north and north eastern parts of the country.

Rainfall in Kenya is characterized by a bimodal distribution with the country receiving long rains from March to April and short rains from October to December.

The Kenya rainfall data is used to fit a GLM with Tweedie Distribution. Rainfall data has been investigated using hourly, daily, monthly or even yearly time scales. Monthly rainfall data from 1961 to 2001 is used in this study. The rainfall data was collected from a network of twenty five rain stations situated in different parts of the country. The rainfall records have at least forty (40) years duration which is adequate to assess the temporal and spatial characteristics of rainfall.

The software R has been used to demonstrate the application

of modelling rainfall using Generalized Linear Models (GLMs) as it has the necessary requirement of being able to produce a GLM with a Tweedie distribution.

## 2.2 Generalized Linear Models

To model a dataset using GLMs three decisions need to be made:

1. What is the distribution of the response variable?
2. What function of mean will be modelled as linear in the predictors?
3. What will the predictors be?

The answers to these three questions defines the components needed to create a GLM. First, is the existence of  $n \times 1$  random variables

$$Y_1, \dots, Y_n$$

dependent on  $t$  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 the various predictors. The link function uses a set of  $p$  unknown parameters,  $\beta$ , and a set of  $n \times t$  known explanatory variables

$$X_{n \times t} = [X_1^T, \dots, X_n^T]$$

formed together so that  $X\beta$  is a linear structure.

### 2.2.1 Exponential Dispersion Models

GLMs are formulated within the framework of the set of distributions which belong to the family of Exponential Dispersion Models (EDMs). An Exponential Dispersion Model (EDM)

is a two parameter family of distributions consisting of a linear exponential family with an additional dispersion parameter. An Exponential Dispersion Model (EDM) has a probability density function that can be written in either of the following forms

$$p(y; \theta, \phi) = a(y, \phi) \exp\left\{\frac{1}{\phi}[y\theta - k(\theta)]\right\} \quad (2.1)$$

or

$$p(y; \theta, \phi) = b(y, \phi) \exp\left\{\frac{1}{2\phi}[-d(y, \mu)]\right\} \quad (2.2)$$

where

$\phi > 1$  is the dispersion parameter,  $\mu = k'(\theta)$  is the position parameter,  $d(y, \mu)$  is the unit deviance,  $k(\theta)$  is the cumulant function,  $\theta$  is the canonical parameter and  $y$  is the variable of interest.

The family includes discrete and continuous densities as well as mixed densities.

The normal, binomial, poisson, inverse gaussian, exponential, gamma and the tweedie distributions all have distributions that form part of exponential dispersion model. The mean and variance of EDMs can be defined as follows:

Mean of  $y$ :

$$E[y] = \mu = k'(\theta) \quad (2.3)$$

Variance of  $y$ :

$$Var[y] = \phi k''(\theta) \quad (2.4)$$

$\theta$  is related to the mean  $\mu$  through equation (2.3). The relationship between  $\mu$  and  $\theta$  is often written as  $\tau(\theta) = k'(\theta)$  and  $\theta = \tau^{-1}(\mu)$ . The function  $\tau(\theta)$  is often written as the mean-value mapping and gives the functional relationship between  $\mu$  and  $\theta$ .

## 2.2.2 Generalized Linear Models

Generalized Linear Models consist of two components

1. The response variable,  $y_i$ , which follows an EDM with mean  $\mu$  and dispersion parameter  $\phi$ , such that

$$y \sim ED(\mu_i, \phi/w_i) \quad (2.5)$$

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

2. The link function

$$g(\mu_i) = x_i^T \beta \quad (2.6)$$

which relates the expected values of the  $y_i$  that is  $\mu_i$ , to the covariates  $x_i$ .

$x_i^T \beta$ , the linear component of the link function is called the linear predictor and is given by the symbol  $\eta$ , so that,

$$g(\mu_i) = \eta = x_i^T \beta \quad (2.7)$$

## 2.3 Variance Functions

The variance function uniquely identifies a distribution within the class of EDMs (that is, an exponential dispersion model is characterized within the class of all exponential dispersion models by its variance function). Equation (2.3) shows that  $k'(\theta)$  is a function of the mean and thus  $k''(\theta)$  is also dependent on the mean. Thus  $k''(\theta)$  is often replaced by the variance function  $V(\mu)$  so that

$$V(\mu) = k''(\theta) \quad (2.8)$$

The role of the variance function is to describe the mean-variance relationship of a distribution when the dispersion parameter is

Table 2.1: The characteristics of some of the distributions of the exponential dispersion models (McCullagh and Nelder, 1989)

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

held constant. If  $Y$  follows an EDM with mean  $\mu$ , variance function  $V(\mu)$ , and dispersion parameter  $\phi$ , then the variance of  $Y$  can be written as

$$\text{Var}(Y) = \phi V(\mu) \quad (2.9)$$

Table(2.1) provides information about several distributions that come from the EDM family, including their variance functions.

## 2.4 Power-Variance(Tweedie)Distributions

Within EDMs, there exists a class of distributions with power-mean variance relationships known as the Tweedie Family of Distributions.

The Tweedie family is a three parameter family of distributions in  $\mu$  (the mean),  $\phi > 0$  (the dispersion) and  $p$ . Tweedie distributions with fixed  $k(\cdot)$  and  $a(\cdot, \cdot)$  and variable  $\theta$  and  $\phi$

exists such that the variance is of the form

$$V(\mu) = \mu^p \tag{2.10}$$

for some exponent  $p \in (-\infty, 0] \cup [1, \infty)$ . The parameter  $p$  is called the index parameter and determines the shape of the Tweedie Distributions.

Most of the important distributions commonly associated with GLMs are contained within the Tweedie Distributions framework including the Normal( $p=0$ ), Poisson( $p=1$ ), Gamma( $p=2$ ), and Inverse Gaussian( $p=3$ ).

Tweedie models exist for all values of  $p$  outside the interval  $(0,1)$ . Apart from the four distributions stated above, none of the Tweedie models have density functions which have explicit analytic forms or which can be written in closed form. Instead, the densities can be represented as infinite summations derived from series expansions, evaluating infinite oscillating integrals, the method of interpolation, inversion of the cumulant generating function, by the saddlepoint approximation method or by evaluating the corresponding quasi-likelihood of the distribution.

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. Table (2.2) shows a summary of Tweedie EDMs and Figure (2.1) shows a plot of some of the Tweedie densities.

Due to their ability to model both discrete and continuous data simultaneously they are very useful in rainfall modelling.

The mean,  $\mu$ , and the canonical parameter,  $\theta$  can be found for a Tweedie Distributions by noting that  $k''(\theta) = \frac{d\mu}{d\theta} = \mu^p$  and the

mean is given by  $\mu = k'(\theta)$ . Hence

$$\mu^p = \frac{\delta^2 k}{\delta(\theta^2)} = \frac{\delta}{\delta\theta} \left( \frac{\delta k}{\delta\theta} \right) = \frac{\delta\mu}{\delta\theta} \quad (2.11)$$

Taking the reciprocals on both sides and integrating with respect to  $\mu$  gives

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p}, & p \neq 1 \\ \log\mu, & p=1 \end{cases} \quad (2.12)$$

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

$$k(\theta) = \begin{cases} \frac{\mu^{2-p}}{2-p}, & p \neq 2 \\ \log\mu, & p=2 \end{cases} \quad (2.13)$$

Thus the Tweedie densities can be written as

$$f_p(y; \mu, \phi) = a_p(y, \phi) \exp\left\{ \frac{1}{\phi} \left[ y \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p} \right] \right\} \quad (2.14)$$

for  $p \neq (1, 2)$ .

Tweedie distributions are the only EDMs which are closed under re-scaling of the response variable, thus if  $y \sim ED_p(\mu, \phi)$ , then  $cy \sim ED_p(c\mu, c^{2-p}\phi)$ . This makes Tweedie EDMs an obvious choice for modelling data when the unit of measurement is arbitrary.

The function  $a_p(y, \phi)$  cannot generally be written in closed form apart from some of the special cases given above. The following methods are used to evaluate that part of the distribution.

#### 2.4.1 Series Evaluation

If  $Y \in ED_p(\mu, \phi)$  with  $1 < p < 2$ , then  $Y$  can be represented as

$$Y = X_1 + X_2 + X_3 + \dots + X_N \quad (2.15)$$

Table 2.2: Summary of Tweedie EDMs.  $p$  is the index parameter of the Tweedie distribution and  $y$  is the variable of interest

Distribution	$p$	Range of $y$
Extreme stable	$p < 0$	$(0, \infty)$
Normal	$p = 0$	$\mathbb{R}$
Do not exist	$0 < p < 1$	
Poisson	$p = 1$	$(0, \infty)$
Compound Poisson	$1 < p < 2$	$(0, \infty)$
Gamma	$p = 2$	$(0, \infty)$
Positive Stable	$p > 2$	$(0, \infty)$
Extreme Stable	$p \rightarrow \infty$	$\mathbb{R}$

### Tweedie density functions

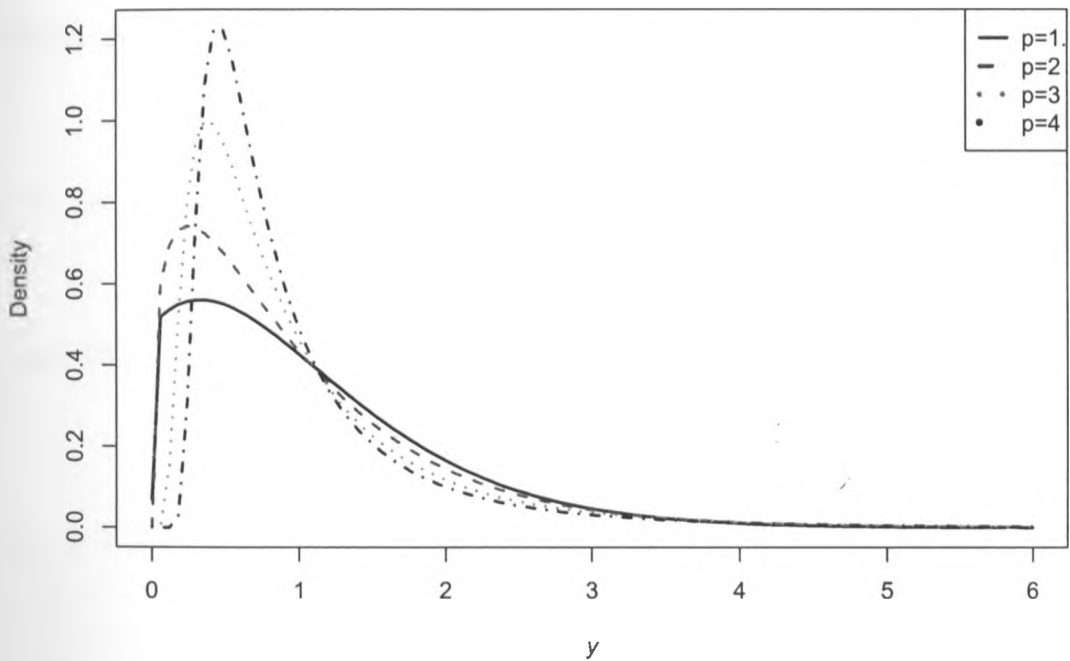


Figure 2.1: Some Tweedie density functions.  $p=1.5$  corresponds to Compound Poisson distribution,  $p=2$  corresponds to a Gamma distribution and  $p=3$  corresponds to the Inverse Gaussian distribution. In each case  $\phi$  is fixed at 0.75, the mean and variance are fixed at unity.



where  $N$  has a Poisson distribution

$X_i$  are independent gamma random variables

Let  $\lambda$  be the mean of  $N$  and let  $\alpha$  and  $\gamma$  be the shape and scale parameters of the  $X_i$ , with  $\alpha\gamma$  and  $\alpha\gamma^2$  the mean and variance of  $X_i$  respectively. Then the parameters are related by

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}$$

$$\alpha = \frac{(2-p)}{(1-p)}$$

$$\gamma = \phi(p-1)\mu^{p-1}$$

From this or otherwise it follows that

$$P(Y = 0) = \exp(-\lambda) = \exp\left\{\frac{-\mu^{2-p}}{(2-p)}\right\} \quad (2.16)$$

and for  $Y > 0$  that

$$a(y, \phi) = \frac{1}{y} W(y, \phi, p) \quad (2.17)$$

with  $W(y, \phi, p) = \sum_{j=1}^{\infty} W_j$  and

$$W_j = \frac{y^{-j\alpha}(p-1)^{\alpha j}}{\phi^{j(1-\alpha)}(2-p)^{j\alpha} \Gamma(-j\alpha)}$$

For  $p > 2$

$$a(y; \phi) = \frac{1}{\pi y} V(y, \phi, p) \quad (2.18)$$

with

$$V = \sum_{k=1}^{\infty} V_k$$

and

$$V_k = \frac{\Gamma(1 + \alpha k) \phi^{k(\alpha-1)} (p-1)^{\alpha k}}{\Gamma(1+k) (p-2)^k y^{\alpha k}} (-1)^k \sin(-k\pi\alpha)$$

Note that  $W_j$  are all positive while  $V_k$  are both positive and negative. The series method has been used in this project to evaluate the Tweedie distribution.

### 2.4.2 Inversion of the Cumulant Generating Function

Tweedie densities have cumulant generating function of the form

$$k(t) = [k(\theta + \phi t) - k(\theta)]/\phi \quad (2.19)$$

Given the cumulant generating function, one method of evaluating the distributions is by using the Fourier inversion to invert the cumulant generating function using

$$p(y; \mu, \phi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\{k(it) - ity\} dt \quad (2.20)$$

where  $i = \sqrt{-1}$ . In the case  $1 < p < 2$ , the continuous conditional density  $Y/Y > 0$  is used to obtain

$$p_{Y/Y>0}(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left\{ \frac{M_{Y(it)} - \exp(-\lambda)}{1 - \exp(-\lambda)} \right\} \exp(-ity) dt$$

In both cases, the infinite oscillating integral is evaluated by converting it into a series by determining the zeroes of the integrand and integrating between them. Analytical analysis of the integrand assists in locating the required zeroes and ensuring that the algorithms are known to converge. The convergence is made faster and more reliable by using an acceleration technique called the W-transformation implemented using the W-algorithm

The series and the inversion methods work best in different parts of the parameter space as shown in the table below.

	small y	large y
$1 < p < 2$	both work well	both are OK
$p > 2$	inversion best	series best

### 2.4.3 Saddlepoint Approximation of Tweedie Densities

In this method, the part of the density that cannot be written in closed form is replaced by a simple analytic expression as follows

$$f(y; \mu, \phi, ) = [2\pi\phi y^p]^{-1/2} \exp\{-d(y, \mu)/(2\phi)\} \{1 + O(\phi)\} \quad (2.21)$$

as  $\phi \rightarrow 0$  for Tweedie densities. The ratio of this approximation to the form of the density in (1) is  $\rho = b_p(y, \phi)\sqrt{2\pi\phi y^p}$ . Thus

$$f_p(y; \mu, \phi, ) = \frac{1}{y} b_p(1, \xi) \exp\{-d(y, \mu)/(2\phi)\} \quad (2.22)$$

where  $\xi = \phi y^{p-2}$ , and so the ratio of the density to the saddlepoint approximation can be expressed as  $\rho = b_p(1, \xi)\sqrt{2\pi\xi}$ .  $\rho$  is a function of  $p$  not  $\mu$ , and is a function  $y$  and  $\phi$  only through  $\xi$ . The ratio  $\rho$  is for each  $p$  and is an increasing monotonic function of  $\rho = \phi y^{p-2}$  for  $p > 3$  and a decreasing monotonic function of  $\rho$  for  $1 < p < 3$  provided  $p$  is not close to 1.

Saddlepoint approximation evaluates Tweedie densities as follows: the density is evaluated on a grid of values given the roots of a Chebyshev polynomial and then form the ratio  $\rho$ . For any necessary evaluation, a two-dimensional Chebyshev interpolation scheme is used to interpolate any values of the parameters, and hence find  $\rho$ . From  $\rho$  the density can be reconstructed.

### 2.4.4 Tweedie Distributions and the Quasi-Likelihood

The Tweedie Distribution has the following Quasi-Likelihood distribution (when setting the arbitrary constant of integration to zero)

$$\begin{aligned} Q(\mu; y) &= \int \frac{(y - \mu)}{V(\mu)} d\mu \\ &= \int \frac{(y - \mu)}{\mu^p} d\mu \end{aligned} \quad (2.23)$$

$$\begin{aligned}
&= \int \left( \frac{y}{\mu^p} - \mu^{1-p} \right) d\mu \\
&= \int (y\mu^{-p} - \mu^{1-p}) d\mu \\
&= \frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}
\end{aligned} \tag{2.24}$$

This equation has the same likelihood function as equation (2.14) except now there is no need to estimate  $a(y, \phi)$ . This is useful since the term  $a(y, \phi)$  cannot always be written in a closed form or is of a form which is extremely difficult to evaluate.

## 2.5 Estimation of Parameters

### 2.5.1 Estimation of $\beta$

Estimates of the parameter values,  $\beta$ , are needed to fit a model to a dataset. The maximum likelihood method is used to estimate the parameters for GLMs, with the parameters being estimated numerically using an iterative procedure. The likelihood function is defined as,

$$l(\xi; y) = \prod_{i=1}^n f(y; \xi) \tag{2.25}$$

where

- $n$  is the sample size of the data set, and
- $\xi$  is the parameter of interest

The corresponding log-likelihood is given as

$$l(\xi; y) = \log l(\xi; y) = \log \prod_{i=1}^n f(y; \xi) = \sum_{i=1}^n \log f(y; \xi) \tag{2.26}$$

The log-likelihood of an EDM is given as

$$l(y; \theta, \phi) = \sum_{i=1}^n a(\phi, y) + \frac{1}{\phi} [y\theta - k(\theta)] \tag{2.27}$$

Taking the derivative of  $l(y; \theta, \phi)$  with respect to  $\beta_j$  in order to find the maximum likelihood estimates of  $\beta_j$ .

$$\frac{dl}{d\beta_j} = \frac{dl}{d\theta_i} \times \frac{d\theta_i}{d\mu_i} \times \frac{d\mu_i}{d\eta_i} \times \frac{d\eta_i}{d\beta_j} \quad (2.28)$$

Each of these derivatives can be found individually.

1. 
$$\frac{dl}{d\theta_i} = \sum_{i=1}^n \frac{1}{\phi} [y_i - k'(\theta)] = \sum_{i=1}^n \frac{1}{\phi} [y_i - \mu_i] \quad (2.29)$$

since  $\mu_i = E[y] = k'(\theta_i)$

2. The second component uses the relationship  $\mu_i = E[y] = k'(\theta_i)$

$$\frac{d\mu_i}{d\theta_i} = \frac{dk'(\theta_i)}{d\theta_i} = k''(\theta_i) = V(\mu_i) \quad (2.30)$$

Taking the inverse

$$\frac{d\theta_i}{d\mu_i} = \frac{1}{V(\mu_i)} \quad (2.31)$$

3. The third component differentiates the link function  $g(\mu_i) = \eta_i$

$$\frac{d\eta_i}{d\mu_i} = \frac{g(\mu_i)}{d\mu_i} = g'(\mu_i) \quad (2.32)$$

Taking the inverse

$$\frac{d\mu_i}{d\eta_i} = \frac{1}{g'(\mu_i)} \quad (2.33)$$

4. The fourth component uses

$$\eta_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_r x_{ir} \quad (2.34)$$

Thus the derivative of  $\eta_i$  with respect to  $\beta_j$  is  $x_{ij}$

Therefore

$$\frac{dl}{d\beta_j} = \frac{1}{\phi} \sum_{i=1}^n \frac{(y_i - \mu_i)}{(V\mu_i)} \frac{x_{ij}}{g'(\mu_i)} \quad (2.35)$$

The maximum likelihood is found by setting the above equation to zero and solving for  $j=1,2,\dots, r$ .

The set of equations (2.35) can only be solved through numerical techniques involving iteration, such as the Newton-Raphson method or the Fisher Scoring iteration.

The maximum likelihood estimate of  $\beta$  does not depend on the dispersion parameter  $\phi$  but depends on  $\theta$ .

The scoring iteration for  $\beta$  can be written as

$$\beta^{k+1} = (X^T W X)^{-1} X^T W z \quad (2.36)$$

where  $W = \text{diag}\left\{\left[\frac{d(\mu_i)}{d\mu}\right]^{-2} \frac{w_i}{\phi_i V(\mu_i)}\right\}$  with variance function  $V(\mu) = \mu^p$ ,  $Z$  is the working vector with components  $z_i = \frac{d(\mu_i)}{d\mu}(y - \mu_i) + g(\mu_i)$  and all the terms are evaluated at the previous iterate  $\beta^k$ . The iteration may be started at  $\mu_i = y_i$  and converges reliably to the maximum likelihood estimate  $\hat{\beta}$  for most link functions.

### 2.5.2 Estimation of $\phi$

The maximum likelihood estimate for parameter  $\phi$  in the normal and inverse gaussian cases, is the mean-deviance estimator

$$\hat{\phi} = \frac{1}{n} \sum_{i=1}^n w_i d(y_i, \hat{\mu}_i) \quad (2.37)$$

where  $d(y_i, \hat{\mu}_i)$  is the unit deviance.

In other cases, the unit deviances are not sufficient for  $\phi$  and the maximum likelihood estimate of  $\phi$  must be computed iteratively from the full data.

However, given the estimated values for  $\beta$  and  $\theta$ , an unbiased estimate of  $\phi$  can be obtained from

$$\tilde{\phi} = \sum_{i=1}^n \frac{[y_i - \mu_i(\hat{\beta})]^2}{(\mu_i(\hat{\beta}))^\theta} \quad (2.38)$$

### 2.5.3 Estimation of $p$

To fit a Tweedie GLM, 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 Tweedie profile function finds the most suitable Tweedie distribution for the given dataset using maximum likelihood methods.

The choice of  $p$  determines 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 residual plots. A log-likelihood plot is shown in figure 2.2

## 2.6 Diagnostic Testing

Diagnostic testing is used to determine whether the model adequately fits the data. There are various diagnostic tests available for GLMs including Q-Q plots; scatterplots of residuals and covariates; comparison of residuals sizes; and residual deviance. 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 patterns.

### 2.6.1 Deviance

To measure the appropriateness of a fitted model the difference between the fitted values  $\hat{\mu}$  and the observed values  $y$  is observed. This measure of difference is called the deviance,  $D(y; \mu)$ , and can be calculated as follows.

$$D(y; \mu) = \phi D^*(y; \mu) = 2\phi[l(y; y) - l(\hat{\mu}; y)] \quad (2.39)$$

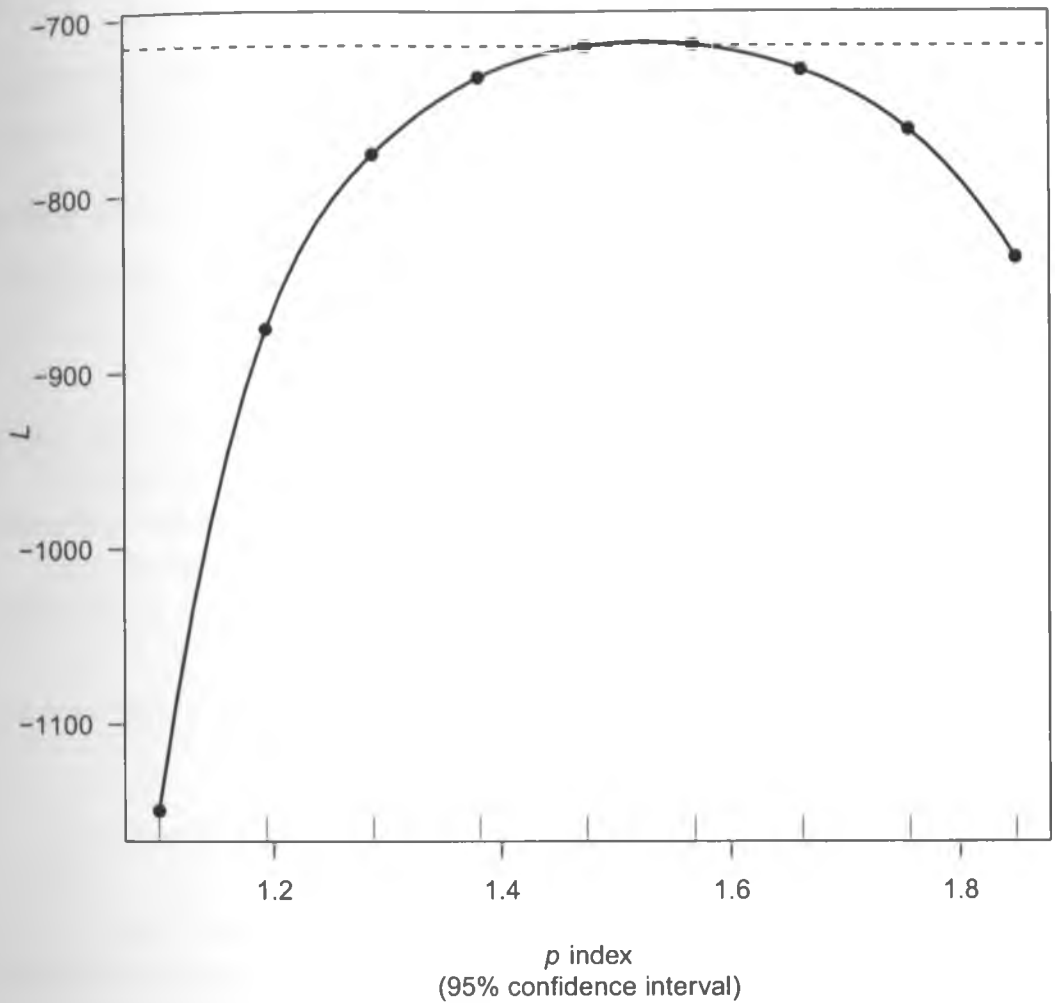


Figure 2.2: A typical log-likelihood plot. This plot estimates the maximum likelihood value of  $p$ . The points represent the computed likelihood values for differing  $p$  estimates, the solid line is a cubic-spline smooth interpolation through these points and the dotted line represents a 95% confidence interval for  $p$ . The estimate for this graph is 1.53



where  $D^*$  is called the scaled deviance and has an approximate  $\chi$  distribution and  $l$  is the log-likelihood function. The deviance is used to compare models.

## 2.6.2 Residuals

Residuals are a measure of how different expected values of the responses emerge from the observed responses.

### Pearson and Deviance Residuals

The Pearson residual is defined by

$$r_{p,i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)^{1/2}} \quad (2.40)$$

where  $\hat{\mu}$  is the fitted value for  $\mu$ .

The Pearson residuals have mean zero and variance  $\phi$ , if the sampling variability in  $\hat{\mu}_i$  is small.

The deviance residuals are defined in terms of the unit deviances. Let

$$t(y, \mu) = y\theta - k(\theta) \quad (2.41)$$

The deviance residual is

$$r_{d,i} = d(y_i, \hat{\mu})^{1/2} \text{sign}(y_i - \hat{\mu}_i) \quad (2.42)$$

Pearson and Residual deviances converge to normality as  $\phi \rightarrow 0$  relative to the  $\mu_i$ , the Pearson residuals at rate  $O(\phi^{1/2})$  by the central limit theorem and the deviance residuals at  $O(\phi)$  by the saddlepoint approximation to  $f(y; \theta, \phi)$ .

### Randomized Quantile Residuals

Let  $F(y; \mu, \phi)$  be the cumulative distribution function of  $f(\mu, \phi)$ . If  $F$  is continuous, then  $F(y; \mu, \phi)$  are uniformly distributed on the unit interval. In this case the quantile residuals are defined by

$$r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}, \hat{\phi})\} \quad (2.43)$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. In the discrete case, if  $a_i = \lim_{y \uparrow y_i} F(y; \hat{\mu}_i, \hat{\phi})$  and  $b_i = F(y_i; \hat{\mu}_i, \hat{\phi})$  the randomized quantile residual for  $y_i$  is defined by

$$r_{q,i} = \Phi^{-1}(\mu) \quad (2.44)$$

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

In both cases the  $r_{q,i}$  are exactly standard normal, apart from sampling variability in  $\hat{\mu}$  and  $\hat{\phi}$ . This implies that the distribution of  $r_{q,i}$  converges to standard normal if  $\beta$  and  $\phi$  are consistently estimated. Quantile residuals have an exact standard normal distribution (apart from sampling error) provided that the correct distribution is used.

## Chapter 3

# APPLICATION OF THE MODEL

### 3.1 Tweedie Rainfall Model

To model rainfall using the Tweedie distribution two assumptions need to be made.

- The amount of rainfall that occurs during any rain event follows a gamma distribution, with mean  $\alpha\gamma$  and variance  $\alpha\gamma^2$  ( $Gam(\alpha, \gamma)$ ).
- The number of rainfall events during the time period (usually day or month) called  $N$  follows a Poisson distribution with mean  $\lambda$

Let  $i$  be a rainfall event and  $R_i$  be the amount of rainfall that occurs during this event.  $Y$  represents the total daily or monthly rainfall, and is represented as the Poisson sum of gamma random variables, such that

$$Y = R_1 + R_2 + \cdots + R_N \quad (3.1)$$

This same setup can be applied to different timescales. For example if  $R_i$  represents the amount of rainfall per month then  $Y$

is the total annual rainfall. Further it is assumed that  $Y$  follows a Tweedie distribution such that

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

with mean  $k'(\theta)$  and variance  $\phi V(\mu)$  where  $V(\mu) = \mu^p$ .

Mathematically, the distributions are best analyzed using the  $(\mu, \phi, p)$  parameterization; climatologically parameterization in terms of  $(\lambda, \alpha, \gamma)$  is more appropriate where:

- $\lambda$  is the mean number of rainfall events per month
- $\gamma$  is the shape of the rainfall distribution when rain occurs during the month
- $\alpha\gamma$  is the mean amount of rainfall per rainfall event

The moment generating function of  $Y$  is

$$\begin{aligned} M(t) &= \int \exp\left\{\frac{1}{\phi}[y(\theta + t\phi) - k(\theta)] + a(y, \phi)\right\} dy \\ &= \exp\left\{\frac{1}{\phi}[k(\theta + t\phi) - k(\theta)]\right\} \end{aligned} \quad (3.2)$$

so the cumulant generating function is

$$\begin{aligned} K(t) &= \log M(t) = \frac{1}{\phi}[k(\theta + t\phi) - k(\theta)] \\ &= \frac{1}{\phi} \frac{\mu^{2-p}}{2-p} [(1 + t\phi(1-p)\mu^{p-1})^{(2-p)/(1-p)} - 1] \end{aligned} \quad (3.3)$$

which is comparable to the cumulant generating function of

$$Z = R_1 + R_2 + \dots + R_N \quad (3.4)$$

, where  $N$  is  $Poisson(\lambda)$  and, conditional on  $N$ , the  $R_i$  are independent  $gamma(\alpha, \gamma)$ , which is

$$\log M(t) = \lambda[(1 - \gamma t)^{-\alpha} - 1] \quad (3.5)$$

Note that  $Z$  is a Poisson mixture of gamma distributions since  $Z$  given  $N$  is  $gamma(N\alpha, \gamma)$ . It can be seen that by identifying terms in the cumulant generating functions  $Y$  has the same distribution as  $Z$  with

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}$$

$$\alpha = \frac{(2-p)}{(p-1)}$$

$$\gamma = \phi(p-1)\mu^{p-1}$$

$\lambda > 0$  and  $\gamma > 0$  imply that  $\mu > 0$  and  $\phi > 0$  also. The density function can now be written as

$$f(y; \mu, \phi) = P(N=0)d_0(y) + \sum_{j=1}^{\infty} P(N=j)f_{Z|N=j}(y)$$

$$= e^{-\lambda}d_0(y) + \sum_{j=1}^{\infty} \frac{\lambda^j e^{-\lambda} y^{j\alpha-1} e^{-y/\gamma}}{j! \gamma^{j\alpha} \Gamma(j\alpha)} \quad (3.6)$$

where  $d_0$  is the Dirac delta function at zero and  $f_{Z|N}$  is the conditional density of  $Z$  given  $N$ . Therefore

$$\log f_p(y; \mu, \phi) = \begin{cases} -\lambda, & \text{for } y=0 \\ \frac{-y}{\gamma} - \lambda - \log y + \log W(y, \lambda, \alpha, \gamma), & \text{for } y > 0. \end{cases} \quad (3.7)$$

where

$$W(y, \lambda, \alpha, \gamma) = \sum_{j=1}^{\infty} \frac{\lambda(y/\gamma)^{j\alpha}}{j! \Gamma(j\alpha)}$$

As  $p \rightarrow 2$  the distribution approaches a  $gamma(\alpha, \gamma)$  distribution. As  $p \rightarrow 1$  the distribution approaches  $Poisson(\lambda)$

The mean of the Poisson-gamma distribution is  $\mu$ , and its variance is  $var[Y] = \phi\mu^p$ . The probability of obtaining no rainfall on any particular event is given by

$$Pr[Y=0] = \exp(-\lambda) = \exp\left[\frac{-\mu^{2-p}}{\phi(2-p)}\right] \quad (3.8)$$

## 3.2 Software

The software R has been used in analysis of rainfall data. In R there are only two(2) link functions that are available for use with the Tweedie distributions: the logarithm and the canonical link functions. The link function used in this study is the logarithm link function. Upon comparison of residual deviances produced by the two(2) link functions it was found that the logarithm link function produced the lowest residual deviances and thus is the most suitable link function to use.

## 3.3 Missing Data

One limitation of using GLMs to estimate parameters is that incomplete datasets can complicate the analysis. If the data is missing completely at random (MCAR), consistent results can still be obtained. Approaches of dealing with missing climatic data include simple long-term averages, cross-correlation between nearest rainfall stations, interpolation with surrounding stations, isopleths or Thiessen polygon methods.

The dataset contained a lot of missing values. However, the methods indicated were not used as some of these are not available in standard software.

## 3.4 Diagnostic Testing

To assess the quality of the fitted distributions, quantile residuals have been used to infer the level of randomness of the deviances. Quantile residuals have an exact standard normal distribution (apart from sampling error) provided that the correct distribution is used.

## Chapter 4

# RESULTS

Rainfall records for twenty five stations spanning forty one(41) years from 1961 to 2001 were analyzed and the following results were obtained.

There were 12300 months in total. A total of 961(7.8%) months either did not have any data or negligible amounts of rainfall were recorded. All the stations had some months with missing data. The station with the highest number of months with missing data was S8834098 which did not have data for 219(44.5%) months while the station with the least number of months with missing data was S9034025 which lacked data for only 3(0.6%) months. A total of 1,073(8.7%) months had recorded zero(0)mm rainfall. Only two stations did not have months with zero rainfall recorded, that is S9034025 and S9034088 but they also had three(3) and thirty eight(38) months of missing data respectively. The station with the highest number of months with zero rainfall was S8641000 which had 159(32.3%) months. There were 10,266(83.5%) wet months, that is months that recorded rainfall amounts greater than 0mm. The maximum rainfall recorded in any month was 6.344.20mm in March 1978 in Station no. S9137089.

Table (4.1) provides a statistical summary of the rainfall amounts for each month. From this analysis it can be seen that April is the wettest month, recording an average of 155.57mm of

Table 4.1: A statistical summary of the monthly rainfall data for each month

MONTH	Max	Mean	Median	STD
January	480	41	14.10	68.85
February	486.8	37.02	12.10	54.92
March	6344.20	92	53.15	257.35
April	1299.10	155.57	136.90	106.18
May	1524.80	142.377	106.60	145.58
June	740.10	68.36	36.65	80.63
July	587.80	57.21	24.70	69.41
August	700.10	58.82	25.55	76.34
September	487	48.17	22.15	62.60
October	4656.60	78.57	49.40	167.62
November	622.60	133.35	92.60	96.71
December	3236.20	74.23	49.20	162.72

rainfall and February is the driest month, recording an average of 37.02mm of rainfall.

Table(4.2) shows the minimum rainfall, the median rainfall, mean rainfall, maximum rainfall, interquartile range, standard deviation and number of months with missing values per station. As the median is consistently lower than the mean, this suggests that the rainfall data is skewed. A plot of the individual rainfall amounts for each month from 1961-2001 is shown in Fig 4.2. The graph shows that the following 7 months recorded abnormally high rainfall amounts; October 1962, March 1963, March 1967, May 1967, March 1978, March 1981 and December 1985. With the removal of these high values, the mean and standard deviation do not change significantly.

This analysis indicates that the use of the Tweedie distribution to model rainfall is reasonable because the data has a discrete component when rainfall amount recorded is zero(0)mm in a month and a continuous component when the rainfall amount recorded is greater than zero(0)mm.

The maximum likelihood values of  $p$  and  $\phi$  for each month



Table 4.2: A statistical summary of the monthly rainfall data for each station

STATION	MIN	MEDIAN	MEAN	MAX	IQR	STD	N/A
S8635000	0	2.60	22.86	3104.80	19.30	144.18	12
S8639000	0	27	63.21	1425.70	69.30	100.56	11
S8641000	0	1.70	23.82	621.70	24.60	51.50	14
S8834001	0	103.80	112.50	1524.80	116.50	103.64	22
S8834098	0	100.20	104.60	341	110.35	71.81	219
S8840000	0	5.60	27.55	621.60	32.70	55.50	14
S8934096	0	158.20	164.60	551.40	120.15	90.57	9
S8935104	0	71	90	442.90	97.65	79.96	48
S8937022	0	46.40	58.70	301.40	60.70	48.97	50
S8937035	0	7.95	31.14	418.20	43.70	51.06	49
S9034025	0.30	93.95	114.63	517.90	92.98	78.62	3
S9034088	5.9	159.50	179.60	1299.10	111.80	116.17	38
S9036025	0	85.05	97.04	735.60	83.20	71.10	11
S9039000	0	7.15	32.82	700.10	35.50	62	21
S9134009	0	57.80	80.03	570.90	83	80.11	86
S9135001	0	41.70	74.20	4656.60	73	221.37	18
S9136130	0	40.60	82.74	2836.50	96.55	158.46	28
S9136164	0	47.3	87.80	622.60	105.60	100.48	22
S9137089	0	25	82.95	6344.20	90.20	341.12	25
S9237000	0	12.45	56.86	2191.50	79.38	128.47	21
S9240001	0	39.60	85.07	917.10	106.55	121.80	27
S9338001	0	18.20	49.50	384.80	65.10	67.86	27
S9339036	0	70.60	109.50	742.10	123.20	117.60	38
S9340007	0	57.70	97.87	656.10	124	109.50	38
S9340009	0	48.60	89.28	600.60	116.80	105.20	72

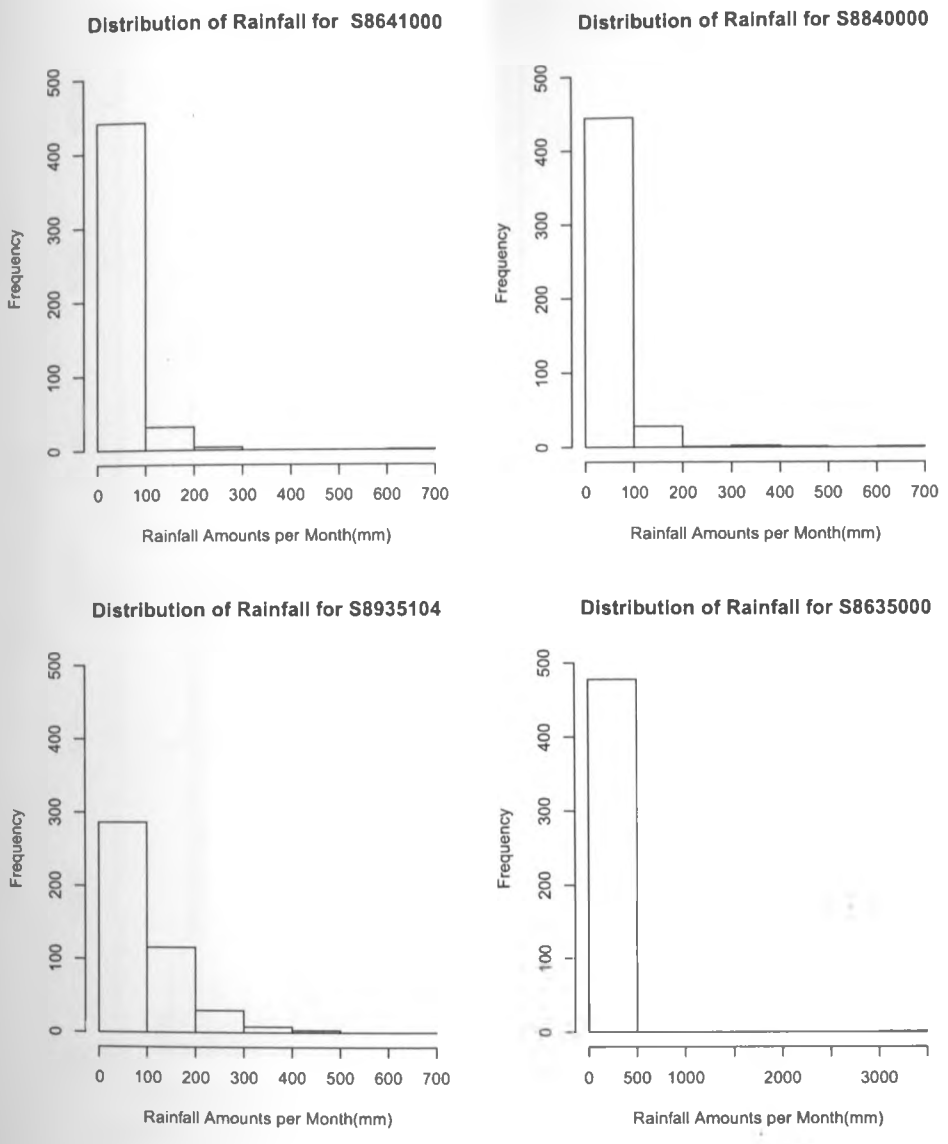


Figure 4.1: Typical histograms for some of the rainfall stations. The histograms show that rainfall is skewed to the right.

### Kenya Monthly Rainfall

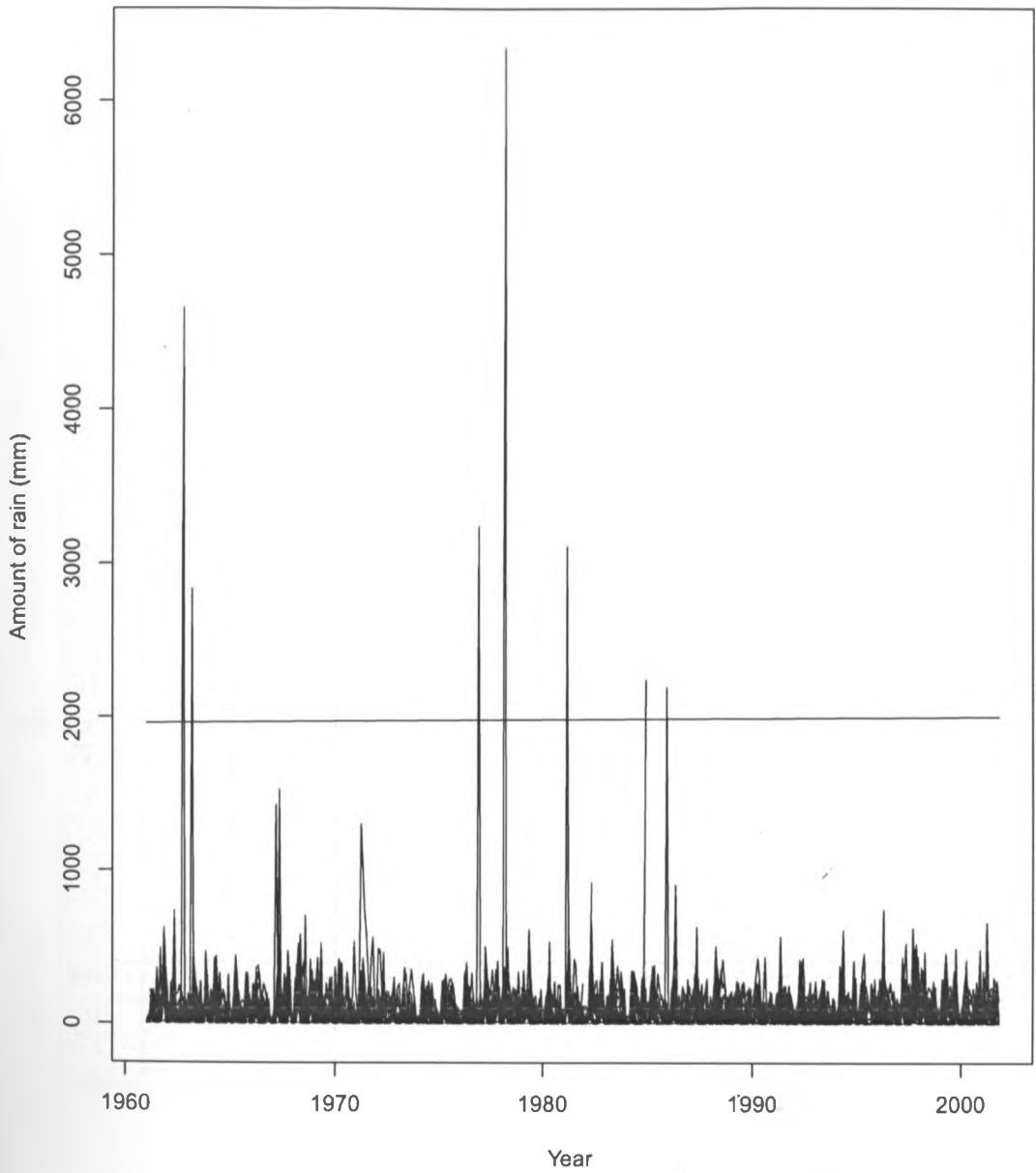


Figure 4.2: Each individual month's rainfall amounts for the Kenya rainfall data.

and the parameters expressed in terms of  $\lambda$  (the mean number of rainfall events per month),  $\gamma$  (the shape of the rainfall distribution when rain occurs during the month) and  $\alpha\gamma$  (the amount of rainfall per rainfall event) are as shown below

Table 4.3: January results

STATION	p	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$
S8635000	1.679	6.363	8.224	0.962	0.474	18.038	8.545
S8639000	1.671	5.339	20.739	1.544	0.49	27.452	13.434
S8641000	1.536	7.545	1.958	0.39	0.867	5.793	5.02
S8834001	1.536	7.833	27.688	1.285	0.867	24.86	21.545
S8834098	2.199	0.809	29.108	-3.191	-0.166	55.081	-9.122
S8840000	1.603	8.789	11.508	0.756	0.658	23.127	15.223
S8934096	1.536	4.787	80.964	3.460	0.867	26.997	23.397
S8935104	1.637	6.804	33.415	1.447	0.571	40.467	23.086
S8937022	1.761	2.54	27.086	3.624	0.314	23.826	7.473
S8937035	1.614	6.665	16.861	1.156	0.628	23.219	14.58
S9034025	1.839	1.348	79.6	9.322	0.192	44.424	8.539
S9034088	1.476	4.126	112.96	5.51	1.102	18.606	20.502
S9036025	1.491	7.923	46.433	1.75	1.037	25.582	26.539
S9039000	1.723	6.101	17.725	1.312	0.384	35.179	13.515
S9134009	1.838	1.407	59.86	8.51	0.194	36.322	7.034
S9135001	2.363	0.208	89.748	-2.581	-0.267	130.49	-34.771
S9136130	1.743	5.449	69.408	2.123	0.346	94.428	32.687
S9136164	1.771	4.058	72.385	2.869	0.296	85.148	25.229
S9137089	1.715	4.008	61.683	2.834	0.398	54.686	21.765
S9237000	1.783	4.229	44.567	2.483	0.278	64.626	17.947
S9240001	1.754	10.241	15.014	0.773	0.326	59.61	19.418
S9338001	2.316	0.754	46.572	-1.244	-0.240	155.794	-37.439
S9339036	1.747	3.719	25.972	2.422	0.339	31.645	10.721
S9340007	1.715	5.451	16.577	1.433	0.398	29.059	11.566
S9340009	1.691	4.176	9.134	1.535	0.448	13.298	5.952

The results for January indicate that 3 months had p values greater than 2 thus yielding negative values of  $\alpha\gamma$ , the mean amount of rainfall. S8834098 had 18(44%) missing observations

for the month of January. Therefore data was not adequate to find the correct estimates of  $p$ . Figure(4.4) and figure(4.5) show that most of the residuals for S9135001 and S9338001 lie away from the line indicating normality, showing that the distribution is not appropriate for modelling rainfall in these 2 stations for the month of January.

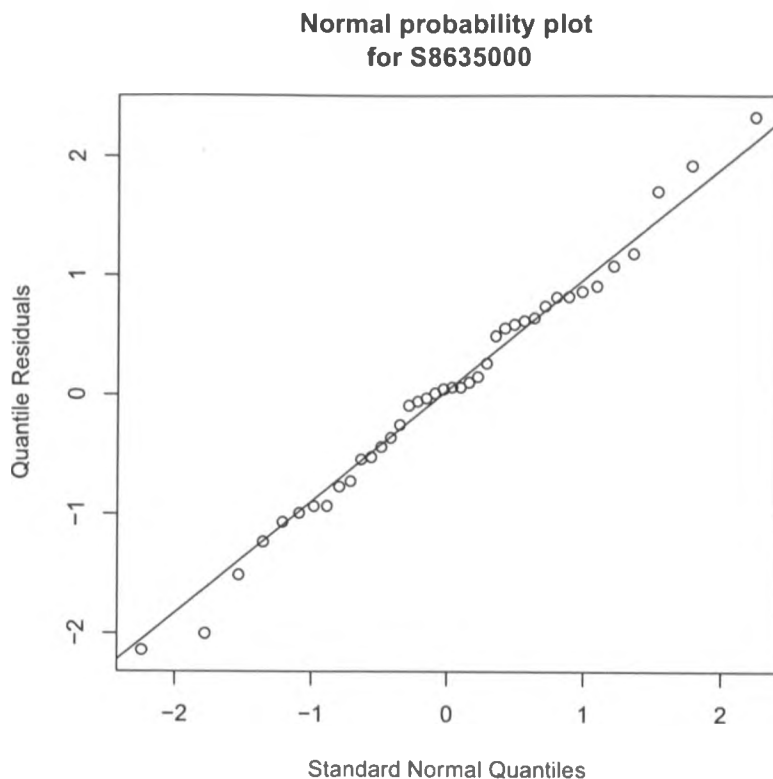


Figure 4.3: The Normal probability plot of the quantile residuals for S8635000 which suggests that the fit is appropriate as the residuals lie close to the line of Normality

It can be observed that in the quantile residual plots, some large values deviate from the normality line, indicating that a Tweedie distribution does not fit extreme values very well. Most of these deviations are observed in the rainy months of April, May,

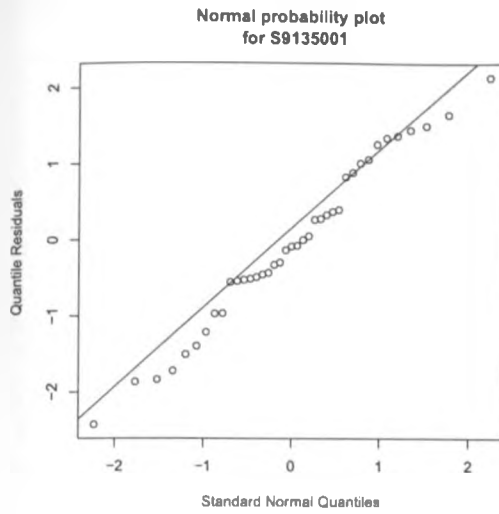


Figure 4.4: The Normal probability plot of the quantile residuals for S9135001 which suggests that the fit is inappropriate as some of the residuals lie away from the line of Normality

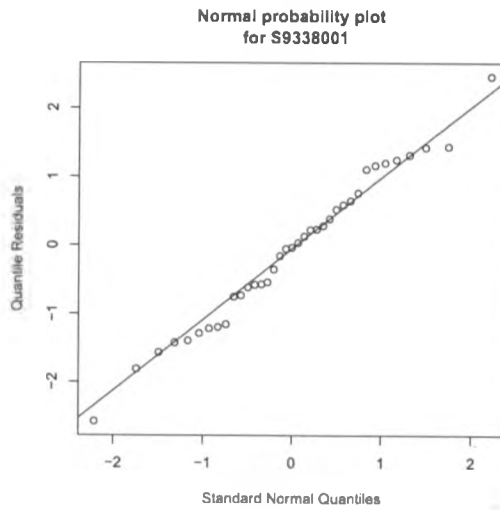


Figure 4.5: The Normal probability plot of the quantile residuals for S89338001 which suggests that the fit is inappropriate as some of the residuals lie away from the line of Normality

November and december when high amounts of rainfall are expected.

Table 4.4: February results

STATION	$p$	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$
S8635000	1.614	5.177	8.611	1.15	0.628	11.934	7.5
S8639000	1.626	4.589	19.356	1.765	0.6	18.32	10.968
S8641000	1.757	12.058	4.581	0.494	0.321	28.91	9.374
S8834001	1.626	5.25	48.62	2.152	0.573	39.43	22.593
S8834098	1.62	4.624	44.168	2.4	0.612	30.082	18.466
S8840000	1.605	5.719	7.745	0.994	0.653	11.942	7.794
S8934096	1.452	6.637	104.481	3.512	1.212	24.539	29.75
S8935104	1.571	6.089	35.84	1.795	0.75	26.632	19.974
S8937022	1.598	5.235	22.2	1.652	0.672	19.981	13.434
S8937035	1.625	13.552	13.791	0.526	0.601	43.57	26.2
S9034025	1.805	1.716	83.013	7.0732	1.242	48.481	60.318
S9034088	1.870	0.838	116.164	17.044	1.149	45.778	52.6
S9036025	1.376	8.438	51.162	2.216	1.663	13.886	23.1
S9039000	1.648	5.459	4.641	0.89	0.543	9.563	5.196
S9134009	1.614	4.768	64.457	2.712	0.628	37.85	23.762
S9135001	1.558	4.408	72.603	3.41	0.792	26.9	21.29
S9136130	1.678	5.741	41.846	1.8	0.476	48.827	23.257
S9136164	1.611	5.63	48.667	2.069	0.636	37	23.522
S9137089	1.575	7.972	37.114	1.3722	0.741	36.52	27.048
S9237000	1.666	6.742	29.79	1.379	0.5	43.114	21.589
S9240001	1.611	7.507	4.5	0.62	0.636	11.5	7.314
S9338001	1.729	4.451	24.17	1.965	0.373	33.015	12.3
S9339036	1.741	6.897	17.236	1.17	0.35	42.103	14.73
S9340007	1.738	6.837	10.237	1.026	0.356	28.056	9.973
S9340009	1.79	6.767	10.45	1.15	0.267	34.114	9.08

Table 4.4 shows that the mean amount of rainfall expected in the month of February ranges between 4.5mm and 116.164mm with most observations in the range 20mm-75mm. Mean number of rainfall events falls between 0.62-17.44 events. The corresponding expected amount of rainfall in February is between 5.2mm

and 60mm of rainfall. Results of the other months follow along similar lines

Table 4.5: September results

STATION	$p$	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$
S8635000	1.703	12.197	5.213	0.451	0.422	27.378	11.563
S8639000	1.691	2.78	20.8	2.954	0.448	15.733	7.042
S8641000	1.654	6.465	1.561	0.522	0.529	5.658	2.993
S8834001	1.261	10.688	103.028	3.894	2.836	9.329	26.456
S8834098	1.456	3.471	98.692	6.103	1.102	14.677	16.172
S8840000	1.679	5.674	3.82	0.8434	0.474	9.559	4.528
S8934096	1.166	10.1	171.4	8.677	5.04	3.919	19.752
S8935104	1.166	11.188	45.331	2.582	5.040	3.483	17.555
S8937022	1.847	0.88	49.452	13.5	0.181	20.274	3.664
S8937035	1.51	5.82	2.646	0.565	0.962	4.871	4.686
S9034025	1.418	3.533	90.19735	6.684	1.393	9.686	13.495
S9034088	1.501	2.205	161.58	11.5	0.996	14.122	14.062
S9036025	1.572	1.356	104.794	12.601	0.747	11.136	8.316
S9039000	1.679	2.703	7.426	2.192	0.474	7.151	3.388
S9134009	1.378	5.033	32.852	2.806	1.649	7.102	11.709
S9135001	1.525	3.324	25.053	2.926	0.907	9.444	8.562
S9136130	1.679	2.749	18.597	2.895	0.474	13.559	6.423
S9136164	1.679	2.467	26.338	3.61	0.474	15.404	7.297
S9137089	1.593	5.073	5.714	0.985	0.687	8.452	5.804
S9237000	1.63	3.904	2.773	1.01	0.588	4.671	2.748
S9240001	2.426	0.242	53.038	-1.794	-0.3	99.026	-29.56
S9338001	1.74	3.457	13.585	2.192	0.352	17.621	6.198
S9339036	2.4	0.108	66.287	-6.342	-223	46.777	-10.453
S9340007	3	0.0242	53.678	-0.767	-0.5	139.865	-69.897
S9340009	1.893	1.122	40.246	12.365	0.12	27.124	3.255

Note that as predictors have not been included,  $\mu$ , has been estimated using sample mean. Maximum likelihood estimates for  $\mu$  can be found based on the first 2 moments, even in the full generalized linear model case. This implies that only the first 2 moments of the distribution are necessary for maximum likelihood



estimation of the linear predictor based on the distributional assumptions.

Table 4.6 to Table 4.10 show the results of some of the stations. In addition to the mean number of rainfall events and mean amount of rainfall per month, the tables also show the probability of obtaining no rainfall in any particular month ( $\Pr(Y=0)$ )

Table 4.6 shows the results for station S8635000. The maximum rainfall expected is 96.4mm in March with 49mm expected in April. Three(3) rainfall events of 43.6mm of rainfall each are expected in March with the probability of no rainfall being 11%. Two(2) rainfall events each with about 31.5mm are expected in April with the probability of obtaining no rainfall in April being 21.1%. The driest month is September when expected rainfall is 5.2mm with about 1 rainfall event with 11.6mm of rainfall. The probability of obtaining no rainfall in September is 63.7%. Table 4.6 shows that expected rainfall is low in January and February(8.5mm). It increases to 96.4mm in April and decreases gradually to about 5.2mm in September and increases during the shorts rains to 27mm in November. On average about 1 or 2 rainfall events can be expected in each month. However high amounts of rainfall are received in each rain event. The probabilities of not receiving any rainfall are highest in June and September at 50.2% and 63.7% respectively.

Table 4.7 shows that the highest rainfall expected in station S8641000 is 91.6mm in April. High rainfall is also expected in October and November. The maximum number of rain events expected in this station are 4 in April and November. Two(2) rain events are expected in March. In all other months just one rainfall event is expected. January and February are most unlikely to receive any rainfall with the probability of receiving no rainfall being 67.7% and 61.1% respectively.

Table 4.8 shows that expected rainfall in station S8840000

Table 4.6: S8635000 results

S8635000	p	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$	P(Y=0)
January	1.679	6.363	8.224	0.962	0.474	18.038	8.545	38.2%
February	1.614	5.177	8.611	1.15	0.628	11.934	7.5	31.7%
March	1.877	6.437	96.351	2.2115	0.141	309.3	43.6113	11%
April	1.941	1.275	48.96	1.555	0.805	39.125	31.488	21.1%
May	1.648	4.895	21.56	1.17	0.543	23.206	12.6	31%
June	1.554	8.177	7.917	0.69	0.805	14.257	11.474	50.2%
July	1.7	6.487	19.144	1.246	0.429	35.36	15.384	28.3%
August	1.654	8.024	8.662	0.76	0.53	21.543	11.418	46.8%
September	1.703	12.197	5.213	0.451	0.422	27.378	11.563	63.7%
October	1.691	5.786	9.539	2.288	0.448	93.167	41.7	10.2%
November	1.752	5.544	27.059	1.648	1.33	49.8	66.214	19.2%
December	1.74	8.386	13.092	0.895	0.352	41.592	14.629	40.9%

Table 4.7: S8641000 results

S8641000	p	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$	Pr(Y=0)
January	1.536	7.545	1.958	0.39	0.867	5.793	5.02	67.7%
February	1.757	12.058	4.581	0.494	0.321	28.91	9.374	61.1%
March	1.66	5.043	18.302	1.567	0.517	22.587	11.678	20.9%
April	1.421	6.93	91.561	3.403	1.373	19.595	26.902	3.3%
May	2.214	0.842	43.21	-2.474	-0.176	98.97	-17.419	-
June	1.539	3.136	0.593	0.566	0.805	1.3	1.046	56.8%
July	1.617	4.246	1.863	0.781	0.62	3.85	2.386	45.8%
August	1.421	1.812	0.468	0.615	1.373	0.554	0.761	54.1%
September	1.654	6.465	1.561	0.522	0.529	5.658	2.993	59.3%
October	2.444	0.28	57.455	-1.330	-0.307	140.485	-43.188	-
November	1.721	2.925	49.503	3.639	0.473	35.228	16.672	2.6%
December	1.679	7	11.034	0.962	0.474	24.218	11.472	38.2%

ranges between 86.5mm in November to about 3.8mm in September. The highest number of rain events is 4 in November with all other months expecting 1 or 2 rain events. August is the driest month with the probability of receiving no rainfall being

53.6%. November is the wettest month with the probability of not receiving any rainfall being 1.6%

Table 4.8: S8840000 results

S8840000	p	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$	Pr(Y=0)
January	1.603	8.789	11.508	0.756	0.658	23.127	15.223	47%
February	1.605	5.719	7.745	0.994	0.653	11.942	7.794	37%
March	1.66	4.522	35.8	2.2	0.52	31.52	16.390	11.1%
April	2.336	0.122	86.401	-5.465	-0.251	62.904	-15.81	-
May	1.611	4.91	33.988	2.064	0.636	25.892	16.467	12.7%
June	1.613	4.443	3.034	0.894	0.631	5.382	3.394	40.9%
July	1.568	3.082	2.308	1.078	0.759	2.818	2.14	34%
August	1.482	4.244	1.837	0.623	1.074	2.743	2.946	53.6%
September	1.679	5.674	3.82	0.843	0.474	9.559	4.528	43%
October	1.666	5.338	34.501	1.83	0.501	37.653	18.855	16%
November	1.797	2.956	86.488	4.12	0.255	82.381	20.991	1.6%
December	1.581	6.07	26.311	1.548	0.722	23.528	16.995	21.3%

Table 4.9: S8935104 results

S8935104	p	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$	Pr(Y=0)
January	1.637	6.804	33.415	1.447	0.571	40.467	23.086	23.52%
February	1.571	6.089	35.84	1.795	0.75	26.632	19.974	16.6%
March	1.62	3.802	78.41	3.629	0.612	35.32	21.616	2.65%
April	1.203	20.5	166.1681	3.602	3.923	11.761	46.132	2.73%
May	1.417	5.994	139.35	5.097	1.4	19.52	27.328	0.6%
June	2.188	0.423	78.335	-222.9	-0.011	31.233	-0.351	-
July	1.356	7.956	106.272	3.942	1.81	26.951	48.781	1.94%
August	1.1	24	116.87	3.361	9	3.863	34.767	3.47%
September	1.166	11.188	45.331	2.582	5.040	3.483	17.555	7.58%
October	1.428061	6.994	95.392	3.39	1.336	21.067	28.148	3.38%
November	2.524	0.053	135.233	-2.769	-0.344	142.155	-48.845	-
December	1.159	11.795	41.554	2.314	5.278	3.402	17.955	9.9%

Interpretation of results for other stations follow along similar lines.

Table 4.10: S8937035 results

S8937035	$\rho$	$\phi$	$\mu$	$\lambda$	$\alpha$	$\gamma$	$\alpha\gamma$	Pr(Y=0)
January	1.614	6.665	16.861	1.156	0.628	23.219	14.58	31.5%
February	1.625	13.552	13.791	0.526	0.601	43.57	26.2	51.9%
March	1.498	7.068	30.33	1.563	1	19.25	19.25	21%
April	1.648	1.568	82.930	8.576	0.542	17.83	9.67	1%
May	1.575	3.694	34.85	2.883	0.74	16.32	12.077	5.6%
June	1.62	10.547	4.581	0.445	0.612	16.822	10.293	64%
July	1.328	10.078	3.85	0.365	2.048	5.145	10.538	69%
August	1.65	9.468	3.428	0.464	0.54	13.71	7.403	62.9%
September	1.51	5.82	2.646	0.565	0.962	4.871	4.686	56.8%
October	1.642	4.305	39.686	2.424	0.558	29.343	16.374	8.9%
November	2.0745	0.681	99.286	-14	-0.07	102.327	-7.094	-
December	1.591	5.163	46.293	2.274	0.693	29.399	20.361	10.3%

The tables show that wet(dry) months are likely to be followed by wet(dry) months which is indicative of the high degree of persistence in the prevailing weather conditions during the wet or dry months.

## Chapter 5

### CONCLUSION

A rainfall model in which a GLM which follows a Tweedie distribution has been proposed. It has been demonstrated that it is possible to model occurrence and amount of rainfall simultaneously. The results show that it is possible to obtain a description of the behaviour of rainfall for a large number of stations some of which present data that are not of the highest quality. The results also provide valuable information regarding the length of rain events, the number of wet events and the amount of rainfall when rain occurs. They also establish clear differences among the wet and dry months. Further several extensions and refinements of the model are possible.

The use of a GLM 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. Further there is a theoretical justification of using these distributions, since for  $1 < p < 2$ , they can be seen as poisson sum of gamma distributions. The number of precipitation events has been modelled as a poisson distribution and the rainfall amounts with a gamma distribution.

An intercept only model was used enabling the predicted amount of rainfall for a particular month to be determined from

the sample mean. It is however possible to include covariates in the model unlike in Markov chains or in the Exponential and Gamma distributions.

The Tweedie GLM model has a few shortcomings though. First the distribution is not able to model extreme events accurately as shown in the normal probability plots. Then GLMs assume that the responses are independent, but there is a consensus amongst researchers that rainfall is correlated and any rainfall model must take into consideration previous days' or months' rainfall into account, a phenomenon known as temporal dependence. Further research into this area of modelling may include:

- Examining of different timescales for example daily or yearly. The Tweedie family of distributions is not only applicable to the monthly but can easily be applied to yearly or daily timescales.
- Modelling rainfall taking into consideration that rainfall occurrences are not independent but rather are correlated. Generalized Additive Models (GAM) and Generalized Estimating Equations (GEEs) are some of the techniques that may be appropriate.
- Each site has been modelled separately. Further research should be carried out on the likelihood of modelling various sites simultaneously.
- Rainfall models that include predictors of rainfall should also be investigated.

## Bibliography

- [1] A F.Siegel, *Modelling Data Containing Exact Zeroes using Zero Degrees of Freedom*, Journal of Royal Statistical Society, Series B:Methodological **47**:267-271,1985
- [2] A J.Nelder, *An alternative view of the splicing data*, Journal of Royal Statistical Society **C 43**:469-476,1994
- [3] A J.Nelder and Y Lee, *Likelihood,quasi-likelihood and pseudo-likelihood:Some comparisons* Journal of Royal Statistical Society:Series B Statistical Methodology,**54**:273-284
- [4] A J.Nelder and R W.M.Wedderburn, *Generalized Linear Models*, Journal of Royal Statistical Society, Series A(General),**135(3)**:370-384,1972
- [5] A J.Dobson, *An Introduction to Generalized Linear Models*, Chapman and Hall,London,2<sup>nd</sup> Edition,2002.
- [6] B Jorgensen, *Exponential Dispersion Models(with discussion)*, Journal of The Royal Statistical Society, **49**:127-162,1987
- [7] B Jorgensen, *Fitting Tweedie's Compound Poisson Model to Insurance Claims Data*, Scandinavian Actuarial Journal,**1**:69-93,1994
- [8] B Rajagopalan and U Lall, *Nonhomogenous markov model for daily precipitation*, Journal of Hydrolic Engineering,**1(1)**:33-40,1996.

- [9] C C.Kokonendji,S Dossou-Gbete,C G.D.Demetrio, *Some Discrete Exponential Dispersion Models:Poisson-Tweedie and Hinde-Demetrio Classes*, SORT 28(2)July-December 2004:201-214
- [10] G K.Smyth, *Regression analysis of Quantity Data with Exact Zeroes*, Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering,Technology and Management,Technology Management Centre,University of Queensland:572-580
- [11] G K.Smyth and B Jorgensen 1999, *Fitting Tweedie's compound Poisson model to insurance claims data:dispersion modelling*, In Proceedings of the 52nd session of the International Statistical Institute,Helsinki,Finland,10-18 August,Contributed Paper Meeting **68**:Statistics and Insurance
- [12] G K.Grunwald and R H.Jones, *Markov models for time series with mixed distribution*, Environmetrics,11:327-339,2000
- [13] Gilchrist,Robert and Drinkwater,Denise *Fitting Tweedie models to data with Probability of Zero Responses*, in Proc.of the 14<sup>th</sup> International Workshop on Statistical Modelling,Graz,Austria,July 19-23,1999
- [14] J R.Green, *A Model for Rainfall Occurrence*, Journal of Royal Statistical Society,**B64**:345-353,1964
- [15] K R.Gabriel and J Neumann *A Markov Chain model for Daily Rainfall Occurrence at Tel Aviv*. Quart.J.Royal Met.Soc.,**88**:90-95,1962
- [16] M.C.K.Tweedie, *An Index which Distinguishes between some Important Exponential Families*, Proceedings of the



Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions Calcutta, 16<sup>th</sup> Dec – 19<sup>th</sup> Dec, 1981, pp:579-604

- [17] P.K.Dunn and G.K.Smyth, *Randomized Quantile Residuals*, Journal of Computational and Graphical Statistics, **5**:236-244, 1996
- [18] P.K.Dunn and G.K.Smyth, *Tweedie Family of Densities: Methods of Evaluation*, Proceedings of the 16<sup>th</sup> International Workshop on Statistical Modelling, Odense, Denmark, 2-6 July, 2001
- [19] Peter K. Dunn, *Occurrence and Quantity of Precipitation can be Modelled Simultaneously*, International Journal of Climatology, **24**:1231-1239, 2004
- [20] P.K.Dunn and G.K.Smyth, *Series Evaluation of Tweedie Exponential Dispersion Model Densities*, Statistics and Computing **15**:267-280, 2005
- [21] P McCullagh and J A. Nelder *Generalized Linear Models*, 2<sup>nd</sup> Edition, Chapman and Hall: London, 1989
- [22] R Chandler and H Wheeler, *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*, Report Research 195, Department of Statistical Science, University College of London, June 1998
- [23] R Chandler and H Wheeler, *Analysis of rainfall using generalized linear models: A case study from the west of Ireland*. Water Resources Research, **38**(10), October 2002.
- [24] R.Coe and R.D.Stern, *Fitting Models to Daily Rainfall Data*, Journal of Applied Meteorology, **21**:1024-1031, February 1982

- [25] R J.Hyndman and G K.Grunwald, *Generalized Additive Modeling of mixed distributions markov models with application to Melbourne's rainfall*, Australian and New Zealand Journal of Statistics,42(2):145-158,2000.
- [26] R W.Katz, *Precipitation as a Chain Dependent Process*, Journal of Applied Meteorology,16:671-676,1977
- [27] X Lana and A Burgueno, *Daily-Dry-Wet Behaviour in Catalonia(NE Spain) from the viewpoint of Markov Chains*, International Journal of Climatology,18:793-815,1998
- [28] W Feller, *An Introduction to Probability and its Applications*, Volume I,John Wiley,New York,1968
- [29] W O.Ochola and P Kerkides, *A Markov Chain Simulation Model for Predicting Critical Wet and Dry Spells in Kenya:Analysing Rainfall Events in the Kano Plains*, Irrigation and Drainage,52:327-342,2003