STATISTICAL MODELS FOR DAILY RAINFALL DATA: A CASE STUDY IN SELANGOR, MALAYSIA

By

CHUAH HOCK LUNG

A dissertation submitted to the
Department of Mathematical and Actuarial Sciences,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the degree of
Master of Science
November 2016

ABSTRACT

STATISTICAL MODELS FOR DAILY RAINFALL DATA: A CASE STUDY IN SELANGOR, MALAYSIA

Chuah Hock Lung

Rainfall volume and occurrence analysis is one of the most commonly applied methods in rainfall data, while probability distributions such as Normal, Log-normal, Gamma, Gumbel and Weibull are among the important distributions that are commonly used in the rainfall analysis. In this study, the daily rainfall volume for a period of 10 years is investigated and fitted using various continuous distributions. The candidate distributions are selected from continuous distribution and beta related distributions. In addition, the analysis of distributions on rainfall occurrence is investigated and it is fitted to the daily rainfall data for the same 10 years period. The candidate distributions are Hurwitz-Lerch Zeta distribution, Eggenberger-Polya distribution, logarithmic distribution, truncated Poisson distribution and geometric distribution. Three new distributions are proposed in this research: the new generalized beta distribution, modified beta distribution and mixture of 2 modified lognormal distributions. The parameters of the distributions are estimated using the maximum likelihood estimation method; with the help of simulated annealing optimization method. The distributions are then plotted and compared to the histogram of daily rainfall volume and occurrence. Model selection techniques such as AIC and BIC are employed to examine the fittings of those distributions.

ACKNOWLEDGEMENT

First and foremost, I would like to express my greatest gratitude to my co-supervisor, Professor Dr. Kunio Shimizu, who has continually given me guidance on my research. His continuous support, patient, enthusiasm and motivation have undoubtedly helped me in overcoming many challenges. I am touched by Professor Shimizu's effort in sparing much time with me whenever he visits Malaysia, especially when he flew to Malaysia solely to attend my work completion seminar to give me support.

Secondly, my deepest appreciation goes to my former supervisor, Dr. Chua Kuan Chin. I would not have begun this study nor would I have completed it if without him. He has shown me his full support and guidance, particularly when I first commenced my research and was an amateurin the area of research. His guidance was not just confined to academic matters, but spiritual matters as well.

I am also immensely grateful to my current supervisor, Dr. Lee Min Cherng for his comments on an earlier version of the manuscript. He spared much of his time in understanding and improving my dissertation. Without his supervision and constant help this dissertation would not have been possible.

Last but not least, I would like to thank my parents, especially my late mother who showed me endless support, love and prayers.

APPROVAL SHEET

This dissertation entitled "**STATISTICAL MODELS FOR DAILY RAINFALL DATA: A CASE STUDY IN SELANGOR, MALAYSIA**" was prepared by CHUAH HOCK LUNG and submitted as partial fulfillment of the requirements for the degree of Master of Science at Universiti Tunku Abdul Rahman.

Approved by:

____

(Dr. Lee Min Cherng) Date:…01/11/2016…..
Assistant Professor/Supervisor
Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

____

(Professor Dr. Kunio Shimizu) Date:…01/11/2016…..
Professor/Co-supervisor
School of Statistical Thinking
The Institute of Statistical Mathematics, Tokyo

LEE KONG CHIAN FACULTY OF ENGINEERING AND SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 01 NOVEMBER 2016

SUBMISSION OF DISSERTATION

It is hereby certified that ***Chuah Hock Lung*** (ID No: ***12UEM01371***) has completed this dissertation entitled "STATISTICAL MODELS FOR DAILY RAINFALL DATA: A CASE STUDY IN SELANGOR, MALAYSIA" under the supervision of Dr. Lee Min Cherng (Supervisor) from the Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, and Professor Dr. Kunio Shimizu (Co-Supervisor) from the School of Statistical Thinking, The Institute of Statistical Mathematics, Tokyo.

I understand that University will upload softcopy of my dissertation in PDF format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

(*Chuah Hock Lung*)

DECLARATION

I, Chuah Hock Lung hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

(CHUAH HOCK LUNG)

Date 01 NOVEMBER 2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAM	Amman Airport Meteorological
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
ENGO	Environmental Non-Governmental Organization
HLZ	Hurwitz-Lerch Zeta
MLE	Maximum Likelihood Estimation
NEM	North East Monsoon
PDF	Probability Density Function
PMF	Probability Mass Function
SWM	South West Monsoon

CHAPTER 1

INTRODUCTION

Malaysia's climate is well known of its hot and humid tropical rainfall that is very much dependent on the monsoon seasons. A monsoon is defined as a seasonal wind that commonly brings changes to the precipitation. The monsoon seasons in Malaysia are divided into two periods: the north east monsoon (NEM) season and the south west monsoon (SWM) season. North east monsoon season occurs from November to March while south west monsoon season occurs from May to September. Between these seasons, Malaysia tends to experience a dry weather which is on average having approximate 100mm of rainfall per month. The annual rainfall volume for the Peninsular Malaysia and East Malaysia (Sabah and Sarawak state) are averaging at 2500 mm and 5080 mm respectively.

Figure 1.1 Map of Malaysia

Figure 1.1 shows the location of Peninsular Malaysia, East Malaysia and the Straits of Malacca.

Malaysia has more than hundred rivers system in Peninsular Malaysia and they contribute 97% of the total raw water supply to the Peninsular Malaysia. One particularly important river basin at Peninsular Malaysia is the Langat river basin. This is because Putrajaya, the federal administrative centre of Malaysia is located at this area, and the river basin supplies fresh water to about 67% of the state of Selangor which has 1.2 million people are living in the region. Langat river basin drains the westward water from the central region to the Straits of Malacca, occupying an area of $2200\,km^2$ at the southern region of Selangor State and there are 2 reservoirs and 8 water treating plants can be found at the river basin itself.

During the beginning of 2014, over 300,000 of households nearby Klang Valley, the central of Selangor state experienced water rationing for the whole month of March, as drought scorched the Peninsular Malaysia with 2 months dry spell have depleted the reservoirs. "The hot weather and lack of rain in catchment areas have caused all reservoirs in Selangor to recede," said the National water commission's chairperson Ismail Kasim. The dry weather may have affected the local businesses that are highly dependent on waters and also Malaysian economy, which its agricultural is the world's second largest producer of palm oil. During the same year, many areas of the Klang Valley were suffering flash flood due to the heavy rain during the monsoon season. According to the local newspaper, The Star, 22^{nd} of December, 2014 reported the downpour of rain lasted for at least 90 minutes with 77mm of rain fell within an hour. This had brought large amount of rain water since the average daily rainfall is only around 12.03mm. This had resulted many major roads and more than 100 vehicles submerged in the water causing serious damages.

Recent water crisis that occurred in Selangor of Malaysia has raised the importance of understanding the rainfall characteristics. Hence, it is important to model a rainfall processes, so that the runoff of water and natural disaster due to heavy rainfall can be supervised, and the appropriate solutions can be carried out. A frequency analysis of the rainfall data is the most commonly method used in understanding the rainfall characteristic. Generally it is assumed that a hydrological variable can be modelled by a certain type of distribution. Among the well known distributions that are used in hydrological analysis are normal, lognormal, gamma, Gumbel, and Weibull. The lognormal distribution is a transformed model from normal distribution which is known to be well fit in many distributions of hydrological variables (Aksoy, 2000). In a similar manner, the gamma distribution has the convenient properties of having only positive variable for application to hydrological data since hydrological variable, e.g. rainfall is always non-negative (Markovic, 1965). The Weibull and Gumbel distributions are usually being used to model data with exaggerated such as low flow values observed in river and frequency analysis of flood respectively (Gumbel, 1954).

The developments of rainfall occurrence models have expanded in demand. Considering Malaysia's economy is very much dependent on its agriculture, the rainfall occurrence models provide much useful information to the water resources management, hydrological and agricultural sectors. Besides, the climate might change over the years due to global warming that has recently caught world attention. Therefore, understanding the rainfall characteristics has become crucial issue for natural disaster management such as flood and draught.

In our research, different types of rainfall studies will be carried out. We divide the rainfall data into rainfall volume (in mm) per day and rainfall occurrence. Rainfall occurrences are divided into dry and wet spell. Dry spell is defined as the consecutively number of days remain as dry day after the last wet day, and vice versa for the wet spell. The objective of this study is to propose three new distributions to examine the rainfall characteristics and to do comparisons of their fittings with some existing distributions using the rainfall data from Langat river basin. We first proposed generalized beta distribution and modified beta distribution which can be reduced to generalized beta of the 1st kind, transformed Kumaraswamy, Gauss hypergeometric and arcsine distribution. However in this study we will be focusing on the proposed generalized beta distribution. Secondly, a mixture of 2 modified lognormal distributions is also proposed which it can be reduced to mixture of 2 lognormal distributions and modified lognormal distribution. We then fit the proposed distributions to an observed rainfall data taken from one of the station at Selangor, Malaysia. We also compared with its reduced model and various beta type and continuous distributions. Under the rainfall occurrence, Hutwitz-Lerch Zeta distribution and other discrete distributions that are well known in hydrology studies such as Eggenberger-Polya, logarithmic and geometric distributions were studied and fitted to dry and wet spell at six stations around Langat river basin. The probability density function (PDF) of the proposed distributions and distributions chosen will be plotted based on the parameter estimated using Maximum Likelihood Estimation (MLE) method. Comparisons between distributions are done using model selection criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) to help in selecting the best fit distribution. The result will be discussed in 2 parts, which are the results of continuous distributions based on rainfall volume, and the results of the discrete distributions based on rainfall day occurrence. This study provides factual information on the rainfall characteristics and some studies on the proposed new distributions. Results have shown the flexibility of the proposed distributions' probability density functions, and also the comparisons of the fitting on rainfall data at chosen station with the existing selected model.

The structure of the dissertation will be as follows: Chapter 2 reviews the literature about continuous distributions for rainfall volume and discrete distributions for rainfall occurrence. We will propose two new distributions based on beta-type distribution and one new mixture distribution on modified lognormal distribution in Chapter 3. The proposed distributions will then be compared with some existing distributions that are well known in hydrology analysis. This study will only examine the data taken from Langat river basin. In Chapter 4, the relative frequency of the rainfall will be plotted for the selected rainfall station (Station 3118102) from Langat river basin and its properties will be discussed. Followed by Chapter 5, the comparison among different distributions based on the AIC and BIC criteria will be discussed. The histograms and the probability density functions will also be plotted. We conclude our dissertation with conclusion and future work in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

In this chapter, we will be discussing the recent issues of climate change and rainfall modelling in Section 2.2. Many countries from different parts of the world are experiencing different level of effects resulting from climate change. The climate change is mainly due to global warming and has resulted in extreme weather like greater rainfall, longer draught and freezing winter. All of these phenomena have cost millions of losses towards the economy and even lives. Therefore researchers are inspired to study the climate particularly to identify the cycle of the rainfall trend. In Section 2.3, we will discuss the rainfall distribution in Malaysia and why rainfall study in Malaysia is important. The study on rainfall distribution can be separated into two types: the rainfall volume and the rainfall occurrence. Rainfall volume in this study is based on the daily rainfall volume in millimeter (mm), which assumed to follow continuous distributions. On the other hand, rainfall occurrence is assumed to follow discrete distribution. Both the continuous and discrete distributions will be discussed in Section 2.4 and 2.5.

2.2 Climate change and rainfall trend

In recent years, there has been an increasing interest in the study of climatic change as it is becoming a main challenge faced by the global leaders, Environmental Non-Governmental Organization (ENGO) leaders and others which they have sat down together in the effort of finding the best resolution to solve the problems caused by the change of climate. Such problems include extreme temperature, droughts, floods and many more. As mentioned in N.I. Obot *et al.* (2010), every aspect of the ecological system is being affected by the climate change, inclusive of flora and fauna, particularly us as human, from agricultural sectors to finance and insurance sectors. The climate plays a very important role in a country's economy. However, Onyenechere (2010) regrettably stated that countries that are most susceptible to the effect of climate change has the least in participation of fighting this global issue. The main reason is that they are vulnerable to climate change is because they do not have the related technology and sufficient resources to overcome it.

Based on the studies done by Adger *et al*. (2003), Novotny and Stefan (2007) and Frich *et al.* (2002), climate change can be assessed by rainfall trend or precipitation as it is one of the climate indicator. Hatzianastassiou *et al*. (2008) also stated that spatial and temporal distributions of rainfall are important in modelling and forecasting the weather and change in climate. Globally, many researches were done on rainfall analysis to understand the characteristics of precipitation so that the appropriate measure can be taken. For example, analysis on the trends of annual precipitation in Sri Lanka was carried out over 100 years from 15 meteorology stations, and the findings were mentioned by Jayawardene *et al.* (2008). It is found that there is a noticeable increasing rainfall trend at Colombo and decreasing trend at Nuwara Eliya and Kandy with the rate of 3.15mm per year, 4.87 mm per year and 2.88 mm per year respectively. Jayawardene *et al.* (2008) also concluded that the downward trends during the last 10 years are greater than the long term fluctuation. Another study done by Smadi and Zghoul (2006) focuses on inspecting the drift of the total precipitation and rainfall day occurrence in Jordan. The study was carried out for the period from 1922 to 2003 at Amman Airport Meteorological (AAM) station in Amman. The result showed that there was a sudden change of the total rainfall and today number of wet days during the year of 1957, which there was a decline in both rainfall volume and days for the last 46 years. At the same time, Partal and Kahya (2006) also have published a study regarding the trend of its precipitation in Turkey. They studied the rainfall trend from 1920 to 1993 over 96 stations around Turkey. The results showed that there are significant downwards trend in their rainfall volume, though there are few upwards trend.

2.3 Analysis of rainfall distribution in Malaysia

Malaysia is situated on a seismically stable plate, a geographical region that being protected from most major natural disasters e.g. earthquakes and volcanoes. Malaysia is less likely to be hit by tsunamis due to the surrounding landmasses and free from typhoons as it is not in the tropical cyclone basins. However, flash flood and drought are the two extreme contrary of natural disasters that are likely to occur in Malaysia within the same year in the recent years. These natural phenomenon can be very unhealthy to the Malaysia's economy when these calamities happen at the capital of Malaysia; which can cause million in losses, not to forget that the country's agricultural plantation is highly dependent on the rainfall. Furthermore, it is costing hundreds of life being taken away during the flood and the exposure to the risk of diseases. Besides flood, landslide is also one of the destructive disasters that caused by the heavy rain as there was a strong correlation between the rainfall and landslides (Ratnayake and Herath, 2005). Therefore, by understanding the characteristics of the rainfall, precaution steps to overcome or reduce problems can be planned and done earlier (Suhaila *et al*., 2011).

The precipitation in Malaysia cannot be detached from the influences of monsoon seasons; in fact, the rainfall of most countries in the area of tropical is greatly affected by the monsoon seasons. A study was carried out by Wong *et al*. (2009) to examine the rainfall's spatial pattern and time-variability in Peninsular Malaysia over 3 regions from 1971 to 2006. The regions are the east coast, west coast and inland of the Peninsular. The monsoon that hit on Malaysia can be shared by two different monsoon seasons, the north east monsoon (NEM) season from November to March, and the south west monsoon (SWM) season from May to September. The findings from the study showed that NEM season brought the most rainfall during the end of year as general. On average, there was 55% and 31% of rainfall received at the east coast region during NEM and SWM season respectively. On the other side of Peninsular Malaysia, the west coast regions had 37% and 41%

of the average rainfall during NEM and SWM seasons. Meanwhile, the inland region received 80% of its average yearly rainfall during the monsoon seasons.

2.4 Modelling of rainfall volume

Researchers have been finding the physical and statistical properties of precipitation based on past rainfall data. This can help in better understanding the rainfall characteristics, by fitting probability density function from many theoretical distributions to model the rainfall volume is one of the alternative (Meneghini and Jones, 1993). As mentioned by Kedem (1990), the PDF of a rainfall is indispensable from a meteorological and climatological viewpoints on the estimation of the total rainfall. However, Cho *et al*. (2004) stated that for a given set of parameters of a distribution, it may not fit at all areas though it may fit at certain location. Therefore there is a need of generalized distribution to model the precipitation.

One of the ways in examining the rainfall volume is to fit theoretical distribution with the assumption that daily rainfalls are independent. Under this study, continuous distribution is used to model the non-zero rainfall volume. Two types of continuous distributions will be discussed in the next section: the continuous distribution and beta type distribution.

2.4.1 Continuous distribution

The commonly used continuous distributions used in rainfall data are gamma distribution and lognormal distribution (Cho *et al*., 2004). Gamma distribution was used to describe daily rainfall at 2 sites in Ghana to model the forecasted rainfall, and its applicability on the model was assured by Adiku *et al*. (1997). Besides, other continuous distribution that are widely used are exponential distribution (Todorovic and Woolhiser, 1975), mixture of two exponential distribution (Whitesides *et al*., 1972) and two-parameter Gamma distribution (Buishand, 1978). In this study, we select 8 continuous distributions to model a rainfall station at Malaysia's Langat river basin and compare with each other. The chosen distributions are as follows:

1. Exponential distribution
2. Mixture of two exponential distributions
3. Lognormal distribution
4. Mixture of two lognormal distributions
5. Modified lognormal distribution
6. Pareto distribution
7. Marginal of linear and angular distribution
8. Gamma distribution

The PDF and the properties of the above distributions are presented in Equation 1 to Equation 8 and Table 2.1 to Table 2.8 respectively.

Table 2.1: Properties of exponential distribution

Parameters	$\beta > 0$	
Probability Density Function	$f(x) = \beta e^{-\beta x}$, $0 < x < \infty$	(1)
Cumulative Distribution Function	$F(x) = 1 - e^{-\beta x}$, $0 < x < \infty$	
Mean	$1/\beta$	
Variance	$1/\beta^2$	
General Moment	$n!/\beta^n$	

Table 2.2: Properties of mixture of 2 exponential distributions

Parameters	$\beta_1 > 0$, $\beta_2 > 0$, $0 \le p \le 1$	
Probability Density Function	$f(x) = p\beta_1 e^{-\beta_1 x} + (1-p)\beta_2 e^{-\beta_2 x}$, $0 < x < \infty$	(2)
Cumulative Distribution Function	$F(x) = 1 - pe^{-\beta_1 x} - (1-p)e^{-\beta_2 x}$, $0 < x < \infty$	
Mean	$p/\beta_1 + (1-p)/\beta_2$	
Variance	$2p\beta_1^{-2} - (p\beta_1^{-1} + (1-p)\beta_2^{-1})^2$ $+ (2-2p)\beta_2^{-2}$	
General Moment	$n!(p/\beta_1^{\,n} + (1-p)/\beta_2^{\,n})$	

The probability density function of the mixture of 2 exponential distributions (Equation (2)) will reduce to exponential distribution (Equation (1)) when $p = 0$.

Table 2.3: Properties of lognormal distribution

Parameters	$\sigma > 0$, $\mu \in \Re$	
Probability Density Function	$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$, $0 < x < \infty$	(3)
Cumulative Distribution Function	$F(x) = \frac{1}{2} + \frac{1}{2} Erf\left(\frac{\ln x - \mu}{\sqrt{2}\sigma}\right)$, $0 < x < \infty$	
Mean	$e^{\mu+\sigma^2/2}$	
Variance	$(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$	
General Moment	$e^{n\mu+(1/2)n^2\sigma^2}$	

Table 2.4: Properties of mixture of 2 lognormal distributions

Parameters	$\sigma_1, \sigma_2 > 0$, $\mu_1, \mu_2 \in \Re$, $0 \le p \le 1$	
Probability Density Function	$f(x) = \frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}}$, $0 < x < \infty$	(4)
Cumulative Distribution Function	$F(x) = \frac{1}{2}\left[pErfc\left(\frac{\mu_1 - \ln(x)}{\sqrt{2}\sigma_1}\right) + (1-p)Erfc\left(\frac{\mu_2 - \ln(x)}{\sqrt{2}\sigma_2}\right) \right]$,	

	$0 < x < \infty$
Mean	$pe^{\mu_1 + (\sigma_1{}^2)/2} + (1-p)e^{\mu_2 + (\sigma_2{}^2)/2}$
Variance	$(pe^{2(\mu_1 + \sigma_1{}^2)} + (1-p)e^{2(\mu_2 + \sigma_2{}^2)})$ $-(pe^{\mu_1 + (\sigma_1{}^2)/2} + (1-p)e^{\mu_2 + (\sigma_2{}^2)/2})^2$
General Moment	$pe^{n\mu_1 + (n^2\sigma_1{}^2)/2} + (1-p)e^{n\mu_2 + (n^2\sigma_2{}^2)/2}$

The probability density function of the mixture of 2 lognormal distribution (Equation (4)) will reduce to lognormal distribution (Equation (3)) when $p = 0$.

Table 2.5: Properties of modified lognormal distribution

Parameters	$\sigma > 0$, $\mu \in \Re$, $-1 \le c \le 1$	
Probability Density Function	$f(x) = 2\frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \times \int_0^{\left(\frac{x}{e^\mu}\right)^{\frac{c}{\sigma}}} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(\ln t)^2}{2}} dt$, $0 < x < \infty$	(5)

The probability density function of the modified lognormal distribution (Equation (5)) will reduce to lognormal distribution (Equation (3)) when $c = 0$.

Table 2.6: Properties of Pareto distribution

Parameters	$\alpha > 0$, $\beta > 0$	
Probability Density Function	$f(x) = \frac{\alpha\beta}{(1+\beta x)^{\alpha+1}}$, $0 < x < \infty$	(6)
Cumulative Distribution Function	$F(x) = 1 - \left(\frac{1/\beta}{x + (1/\beta)}\right)^\alpha$,	

	$0 < x < \infty$	
Mean	$\dfrac{1}{\beta(\alpha-1)}$	, $\alpha > 1$
Variance	$\dfrac{\alpha}{(\alpha-2)(\alpha-1)^2\beta^2}$	, $\alpha > 2$
General Moment	$\dfrac{\beta^{-n}\Gamma(n+1)\Gamma(\alpha-n)}{\Gamma(\alpha)}$	, $\alpha > n$

Table 2.7: Properties of marginal of linear and angular distribution

Parameters	$\lambda > \kappa \geq 0$	
Probability Density Function	$f(x) = \sqrt{\lambda^2 - \kappa^2}\, I_0(\kappa x) e^{-\lambda x}$, $0 < x < \infty$ where I_0 is the modified Bessel function of the 1st kind and order 0	(7)
Mean	$\dfrac{\lambda}{\lambda^2 - \kappa^2}$	
Variance	$\dfrac{2\lambda^2 + \kappa^2}{(\lambda^2 - \kappa^2)^2}$	
General Moment	$\lambda^{-(n+1)}\sqrt{\lambda^2 - \kappa^2}\,\Gamma(n+1)$ $\times {}_2F_1\left((n+2)/2, (n+1)/2, 1, \kappa^2/\lambda^2\right)$	

The probability density function of the marginal of linear and angular distribution (Equation (7)) will reduce to exponential distribution (Equation (1)) when $\kappa = 0$. Modified Bessel function of the 1st kind and order n from equation (7) can be defined by the contour integral $I_n(z) = \dfrac{1}{2\pi i}\oint e^{(z/2)(t+1/t)} t^{-n-1} dt$, where the contour encloses the origin and is traversed in a counterclockwise direction.

Table 2.8: Properties of gamma distribution

Parameters	$k > 0$, $\theta > 0$
Probability Density Function	$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$, (8) $0 < x < \infty$ where $\Gamma(k)$ is the Gamma function
Cumulative Density Function	$F(x) = \frac{1}{\Gamma(k)} \gamma(k, x/\theta)$, (9) $0 < x < \infty$ where $\gamma(k, x/\theta)$ is lower incomplete gamma function
Mean	$k\theta$
Variance	$k\theta^2$
General Moment	$\frac{\theta^n \Gamma(k+n)}{\Gamma(k)}$

Gamma function from Equation (8) can be defined by $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ and lower incomplete gamma function from Equation (9) can be defined by $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$.

2.4.2 Continuous distribution: beta-type

In precipitation, many kinds of data are modelled by finite distribution including hydrological data. This section gives beta-type distribution with finite range for an advantage in modelling the data. An example of a beta-type distribution that is applied in hydrology is the Kumaraswamy distribution. It was introduced by Kumaraswamy (1980) for a double bounded random process and it has considerably caught attention in hydrology study. Kumaraswamy (1980) claimed that beta distribution does not fit hydrological random variable such as daily rainfall.

Beta densities are versatile and able to model many types of uncertainty since it can be unimodal, uniantimodal, increasing, decreasing or constant depending on the values of γ and q (Johnson *et al*., 1996). However, it is not so fabulous in some ways, as in the two-parameter distribution can only provide limited precision in fitting the data. It is preferable to have more parametrically flexible versions of beta to allow a richer empirical description of data, while offering more structure than a nonparametric estimator.

Therefore, we select a few beta-type distributions to describe the empirical data. The distributions are as follows:

1. Beta distribution
2. Gauss hypergeometric distribution
3. Generalized beta of the 1st kind (McDonald and Xu, 1995) distribution
4. Kumaraswamy distribution
5. Standard arcsine distribution

The PDF and properties of the above distributions are presented in Equation 10 to Equation 15 and Table 2.9 to Table 2.13 respectively.

Table 2.9: Properties of beta distribution

Parameters	$\gamma > 0$, $q > 0$
Probability Density Function	$f(x) = \dfrac{x^{\gamma-1}(1-x)^{q-1}}{B(\gamma,q)}$, (10) $0 < x < 1$
Cumulative Distribution Function	$F(x) = \dfrac{B(x;\gamma,q)}{B(\gamma,q)}$, (11) $0 < x < 1$, $B(x;\gamma,q)$ is the incomplete beta function
Mean	$\dfrac{\gamma}{\gamma+q}$
Variance	$\dfrac{\gamma q}{(\gamma+q)^2(\gamma+q+1)}$
General Moment	$\dfrac{\Gamma(\gamma+q)\Gamma(\gamma+n)}{\Gamma(\gamma)\Gamma(\gamma+q+n)}$

The incomplete beta function from Equation (11) can be defined by $B(z;a,b) = z^a \sum_{n=0}^{\infty} \frac{(1-b)_n}{n!(a+n)} z^n$ where $(x)_n = x(x+1)(x+2)\cdots(x+n-1)$ is Pochhammer's symbol.

Table 2.10: Properties of Gauss hypergeometric distribution

Parameters	$\theta > 0$, $\gamma > 0$, $-\infty < \sigma < \infty$
Probability Density Function	$f(x) = \dfrac{x^{\gamma-1}(1-x)^{\theta-1}(1+tx)^{-\sigma}}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)B(\theta,\gamma)}$, (12) $0 < x < 1$ where ${}_2F_1$ is the hypergeometric function.
Cumulative Distribution function	$F(x) = \dfrac{x^{\gamma}}{\gamma(\gamma+1)\,{}_2F_1(\gamma,\sigma;\gamma+\theta;-t)B(\theta,\gamma)} \times \begin{bmatrix} (\gamma+1)F_1(\gamma;-\theta,\sigma;\gamma+1;x,-xt) \\ +\gamma x F_1(\gamma+1;1-\theta,\sigma;\gamma+2;x,-xt) \end{bmatrix}$
Mean	$\dfrac{1}{{}_2F_1(\gamma,\sigma;\gamma+\theta;-t)B(\theta,\gamma)} \times \begin{bmatrix} B(\theta,\gamma+1)\,{}_2F_1(1+\gamma,\sigma;1+\gamma+\theta;-t) \\ -t^{-\sigma}(-1)^{\theta}B(\theta,\sigma-\gamma-\theta) \\ \times{}_2F_1(\sigma,\sigma-\gamma-\theta;\sigma-\gamma;-1/t) \end{bmatrix}$
General Moment	$\dfrac{1}{{}_2F_1(\gamma,\sigma;\gamma+\theta;-t)B(\theta,\gamma)} \times \begin{bmatrix} B(\theta,n+\gamma)\,{}_2F_1(n+\gamma,\sigma;n+\gamma+\theta;-t) \\ -t^{-\sigma}(-1)^{\theta}B(\theta,\sigma+1-n-\gamma-\theta) \\ \times{}_2F_1(\sigma,\sigma+1-n-\gamma-\theta;\sigma+1-n-\gamma;-1/t) \end{bmatrix}$

The hypergeometric function from Equation (12) can be defined by:

$${}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!}.$$

Table 2.11: Properties of generalized beta of the 1st kind distribution

Parameters	$a > 0$, $\gamma > 0$, $q > 0$	
Probability Density Function	$f(x) = \frac{ax^{a\gamma-1}(1-(x/b)^{a})^{q-1}}{b^{a\gamma}B(\gamma,q)}$, $0 < x < b$	(13)
Cumulative Distribution Function	$\frac{x^{a\gamma}\ {}_2F_1\left(1-q,\gamma;\gamma+1;(x/b)^{a}\right)}{\gamma\left(b^{a\gamma}\right)B(\gamma,q)}$	
Mean	$\frac{bB\left(q,\gamma+\frac{1}{a}\right)}{B(q,\gamma)}$	
General Moment	$\frac{b^{n}B\left(q,\gamma+\frac{n}{a}\right)}{B(q,\gamma)}$	

Table 2.12: Properties of Kumaraswamy distribution

Parameters	$\alpha > 0$, $\beta > 0$	
Probability Density Function	$f(x) = \alpha\beta x^{\alpha-1}(1-x^{\alpha})^{\beta-1}$, $0 < x < 1$	(14)
Cumulative Distribution Function	$F(x) = 1-(1-x^{\alpha})^{\beta}$ $0 < x < 1$	
Mean	$\beta B(\beta, 1/a+1)$	
General Moments	$\beta B(\beta, n/a+1)$	

Table 2.13: Properties of standard arcsine distribution

Probability Density Function	$f(x) = \dfrac{1}{\pi\sqrt{x(1-x)}}$, $0 < x < 1$	(15)
Cumulative Distribution Function	$F(x) = \dfrac{2\sin^{-1}(\sqrt{x})}{\pi}$, $0 < x < 1$	
Mean	1/2	
Variance	1/8	
General Moment	$\dfrac{\Gamma(n+1/2)}{\sqrt{\pi}\Gamma(n+1)}$	

2.5 Modelling of rainfall occurrence – discrete distribution

Besides modelling the rainfall amount, understanding the rainfall occurrence also important in studying the rainfall characteristic. The definition of rainfall occurrence or spell is based on the consecutive number of days of having dry (wet) day after every wet (dry) day. In this study, we considered a wet day when it has the rainfall volume greater or equal to 1mm for the day. One of the objectives of this study is to find the best fitting distribution in both dry spell and wet spell from a number of chosen distributions to a few selected important rainfall stations of Malaysia. Logarithmic distribution was suggested by Williams (1952), who was among the first to be involved in the study of the distribution of wet and dry sequences, using data from England and it is found that logarithmic distribution fitted very well to the distribution of dry spells. Other most frequently used model in dry and wet spell is the Markov chain. The Markov chain has been applied by many authors e.g. Gabriel and Neumann (1962), Richardson (1981) and Dubrovský (1997). However, the Markov chain model tends to exhibit geometric distribution for the probability of extreme long dry or wet spells. In fact, this model normally underestimates the long dry spells but overestimates the very short dry spells. Therefore, Berger and Goossens (1983) and Nobilis (1986) have reproduced the short and long spells by using higher order Markov chain and Eggenberger-Polya distribution. It is found that Eggenberger-Polya distribution is the best fit in the long spells. Later, a model constituting two different geometric distributions is proposed by Racsko *et al*. (1991). The mixed distributions were divided according to the length of dry sequences. Results suggested that mixed distributions could be favorable in reproduction the long dry spells while for wet spells, simple geometric distribution could be promising. At the same time, it was realized that Poisson distribution fits well to the excessive dry events at the Catalonia (69 stations) (Lana and Burgueño, 1998).

In this research, Hurwitz-Lerch Zeta (HLZ) distribution is studied and compared with the other well-known distributions since it is a relatively new distribution that has not been applied in precipitation studies. The candidate distributions are as follows:

1. Hurwitz-Lerch Zeta distribution
2. Eggenberger-Polya distribution

3. Logarithmic distribution
4. Truncated Poisson distribution
5. Geometric distribution

The probability mass function (PMF) and properties of above distributions are presented in Equation 16 to Equation 20 and Table 2.14 to Table 2.18 respectively.

Table 2.14: Properties of Hurwitz-Lerch Zeta distribution

Probability Mass Function	$P_k = \frac{1}{T(\theta,s,a)} \frac{\theta^k}{(k+a)^{s+1}}, \quad k = 1,2,\cdots \quad (16)$ where $T(\theta,s,a) = \sum_{k=1}^{\infty} \frac{\theta^k}{(k+a)^{s+1}}$, $0 < \theta \le 1$; $s \in C$; $a \ge -1$; $s > 0$ when $\theta = 1$
Mean	$\frac{T(\theta,s-1,a)}{T(\theta,s,a)} - a$

Table 2.15: Properties of Eggenberger-Polya distribution

Probability Mass Function	$P_k = \frac{h+(k-2)d}{(k-1)(1+d)} P_{k-1}, \; k = 2,3,\cdots \quad (17)$ $P_1 = (1+d)^{-h/d}$ $d = \frac{s^2}{h} - 1$ where $s,h \ge 0$

Table 2.16: Properties of logarithmic distribution

Probability Mass Function	$P_k = \frac{-\theta^k}{k\ln(1-\theta)}$, $k = 1,2,...$ where $0 < \theta < 1$	(18)
CDF	$F(k) = 1 + \frac{B(\theta; k+1, 0)}{\ln(1-\theta)}$ where B is the incomplete beta function	
Mean	$\frac{-1}{\ln(1-\theta)}\frac{\theta}{1-\theta}$	
Variance	$-\theta\frac{\theta+\ln(1-\theta)}{(1-\theta)^2\ln^2(1-\theta)}$	

Table 2.17: Properties of truncated Poisson distribution

Probability Mass Function	$P_k = \frac{\lambda^k}{(e^\lambda - 1)k!}$, $k = 1,2,...$ where $\lambda > 0$	(19)
Mean	$\frac{\lambda e^\lambda}{e^\lambda - 1}$	
Variance	$\frac{\lambda e^\lambda}{e^\lambda - 1}\left[1 - \frac{\lambda}{e^\lambda - 1}\right]$	

Table 2.18: Properties of geometric distribution

Probability Mass Function	$P_k = (1-p)^{k-1} \cdot p$, $k = 1,2,\ldots$ where $0 \le p \le 1$	(20)
CDF	$1-(1-p)^k$	
Mean	$\frac{1}{p}$	
Variance	$\frac{1-p}{p^2}$	

We will be discussing three proposed new distributions in the next chapter, where two new distributions will be based on beta-type distribution. A distribution chart between proposed distributions and the commonly used distributions in rainfall modelling discussed above is illustrated in the next chapter as well.

CHAPTER 3

PROPOSED NEW DISTRIBUTIONS

3.1 Background

In recent years, there has been as increasing interest in applying mixed distribution in modelling precipitation data. Fadhilah *et al*. (2007) had found that a mixed-exponential distribution is the best fit distribution among exponential, Weibull and gamma distribution. Suhaila and Jemain (2007) also claimed that a mixed lognormal distribution is a better fit in precipitation compared with lognormal and skewed normal distribution. Studies from both Suhaila *et al*. (2007) and Jamaludin and Jemain (2008) have concluded that a mixed-distribution will be better in modelling the rainfall of Peninsular Malaysia.

In this chapter, three new distributions will be proposed. The first and second proposed distribution will be beta-type distribution while another proposed model is a mixture distribution. The proposed beta-type distributions are generalized beta distribution that consists of 6 parameters and modified beta distribution with 5 parameters. On the other hand, the mixed proposed model is a mixture of 2 modified lognormal distributions with 7 parameters. Figure 3.3 illustrates the relationship between one distribution and another with the proposed distributions.

3.2 Proposed generalized beta distribution

The proposed generalized beta distribution with 6 parameters $\gamma, \rho, \alpha, \beta, \sigma$ and z is constructed from the generalized hypergeometric function. The proposed generalized beta distribution includes many beta-type distributions, e.g. Gauss hypergeometric by Armero and Bayarri (1994), generalized beta of the 1st kind by McDonald (1984), Kumaraswamy (1980) distribution and arcsine distribution. The probability density function of the proposed generalized beta distribution is given by:

$$f(x) = \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma} x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma} F(\alpha,\beta;\gamma;x)}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta; z/z-1)B(\gamma,\rho)} \quad ,$$

where $\sigma > 0$, $\gamma > 0$, $z < 0.5$, $(\gamma+\rho-\alpha-\beta) > 0$.

${}_3F_2$ is a generalized hypergeometric function and defined by:

$${}_3F_2(\alpha_1,\alpha_2,\alpha_3;\beta_1,\beta_2;z) = \sum_{k=0}^{\infty} \frac{(\alpha_1)_k(\alpha_2)_k(\alpha_3)_k}{(\beta_1)_k(\beta_2)_k} \frac{z^k}{k!}.$$

Given that

$$\int_0^1 x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma} F(\alpha,\beta;\gamma;x)dx = \frac{\Gamma(\gamma)\Gamma(\rho)\Gamma(\gamma+\rho-\alpha-\beta)}{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}(1-z)^{-\sigma} \times {}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta; z/z-1) ,$$

where $\mathrm{Re}\,\gamma > 0$, $\mathrm{Re}\,\rho > 0$, $\mathrm{Re}\,(\gamma+\rho-\alpha-\beta) > 0$, $|\arg(1-z)| < \pi$

(Gradshteyn and Ryzhik 2007, p. 813, 7.512.9).

Then it can be easily shown that

$$\int_0^1 \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma} x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma} F(\alpha,\beta;\gamma;x)}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;z/z-1)B(\gamma,\rho)} = 1.$$

The advantages of the proposed generalized beta distribution is it can be very versatile and flexible. The flexibility is able to provide good description to many different types of data, including unimodal, uniantimodal, increasing, decreasing, or bath-tub shape distribution depending on the value of its parameters. Table 3.1 and Figure 3.1 illustrate the shapes of the proposed generalized distribution's PDF given different values for the parameters.

Table 3.1: Values of parameters for proposed generalized beta distribution

	z	σ	α	β	ρ	γ
Green	0.3	2	1	1	1	2
Yellow	0.4	2	1	-1	2	1
Blue	0.4	2	-1	0.5	2	0.5
Black	0.4	0	1	0.5	1.5	0.5
Orange	0.4	2	-2	1	2	5
Purple	-1	2	-2	2	5	6
Red	0.4	0	1	0.5	5	1.5
Brown	0.4	0	-1	10	10	1

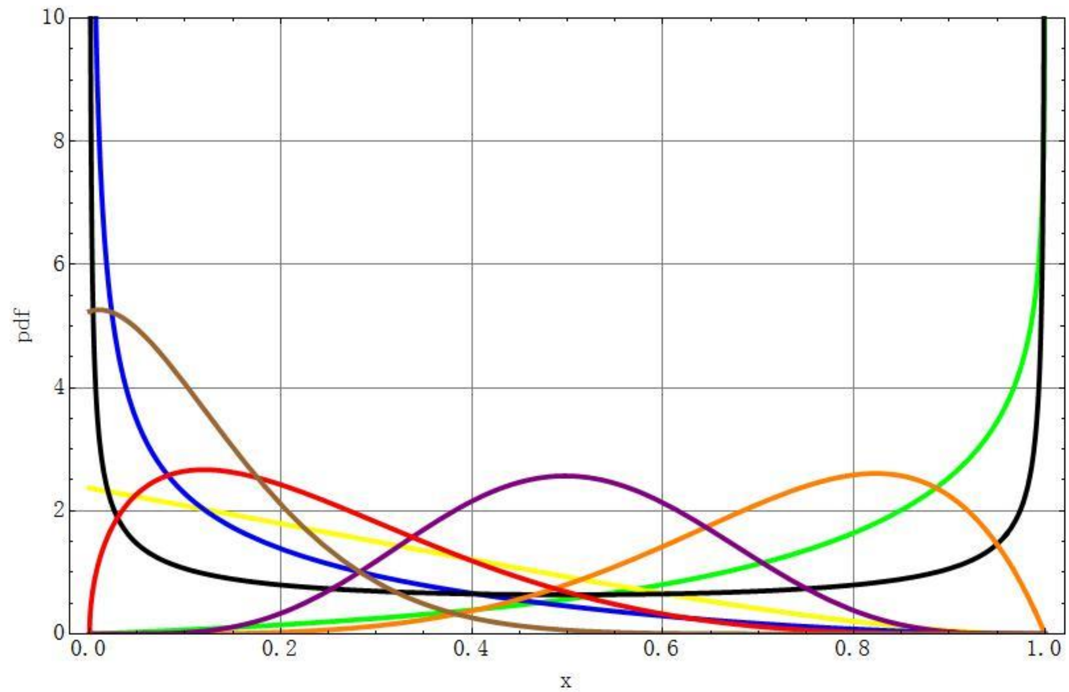

Figure 3.1: PDF of the proposed generalized beta density

Section 3.2.1 to 3.2.4 show how the proposed generalized beta distribution can be reduced to Gauss hypergeometric, generalized beta of the 1st kind, Kumaraswamy and arcsine distribution.

3.2.1 Gauss hypergeometric

When $\alpha = \gamma;\ \ \rho = \beta + \theta;\ \ z = -t,$

$F(\gamma, \beta; \gamma; x) = {}_1F_0(\beta; x) = (1-x)^{-\beta},$

$F(\alpha, \beta; \gamma; z) = (1-z)^{-\alpha} F(\alpha, \gamma - \beta; \gamma; z/z-1),$

$$\frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma} x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma} F(\alpha,\beta;\gamma;x)}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;z/z-1)B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(\beta+\theta)\Gamma(\gamma+\theta)}{\Gamma(\gamma+\beta+\theta)\Gamma(\theta)}(1+t)^{\sigma} x^{\gamma-1}(1-x)^{\beta+\theta-1}(1+tx)^{-\sigma}(1-x)^{-\beta}}{{}_3F_2\left(\beta+\theta,\sigma,\theta;\beta+\theta,\gamma+\theta;\frac{z}{z-1}\right)B(\gamma,\beta+\theta)}$$

$$= \frac{\frac{\Gamma(\beta+\theta)\Gamma(\gamma+\theta)}{\Gamma(\gamma+\beta+\theta)\Gamma(\theta)}(1+t)^{\sigma}x^{\gamma-1}(1-x)^{\theta-1}(1+tx)^{-\sigma}}{{}_2F_1\left(\sigma,\theta;\gamma+\theta;\frac{z}{z-1}\right)\frac{\Gamma(\gamma)\Gamma(\beta+\theta)}{\Gamma(\gamma+\beta+\theta)}}$$

$$= \frac{\frac{\Gamma(\gamma+\theta)}{\Gamma(\gamma)\Gamma(\theta)}(1+t)^{\sigma}x^{\gamma-1}(1-x)^{\theta-1}(1+tx)^{-\sigma}}{{}_2F_1\left(\sigma,\theta;\gamma+\theta;\frac{z}{z-1}\right)}$$

$$Let\, u = \gamma+\theta \implies \theta = u-\gamma$$

$$\begin{aligned} &{}_2F_1\left(\sigma,\theta;\gamma+\theta;\frac{z}{z-1}\right) \\ &= {}_2F_1\left(\sigma,u-\gamma;u;\frac{z}{z-1}\right) \\ &= {}_2F_1(\sigma,\gamma;\gamma+\theta;z)(1-z)^{\sigma} \\ &= {}_2F_1(\sigma,\gamma;\gamma+\theta;-t)(1+t)^{\sigma} \end{aligned}$$

$$= \frac{\frac{\Gamma(\gamma+\theta)}{\Gamma(\gamma)\Gamma(\theta)}(1+t)^{\sigma}x^{\gamma-1}(1-x)^{\theta-1}(1+tx)^{-\sigma}}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)(1+t)^{\sigma}}$$

$$= \frac{x^{\gamma-1}(1-x)^{\theta-1}(1+tx)^{-\sigma}}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)B(\theta,\gamma)}, 0 < x < 1.$$

This is the pdf of Gauss hypergeometric distribution.

3.2.2 Generalized beta of the 1st kind

Let $z = 0; \quad \alpha = \gamma; \quad \rho - \beta = q,$

$F(\gamma,\beta;\gamma;x) = {}_1F_0(\beta;x) = (1-x)^{-\beta},$

$$\frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)}{{}_3F_2\left(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1}\right)B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}x^{\gamma-1}(1-x)^{\rho-1}F(\alpha,\beta;\gamma;x)}{B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(\gamma+\rho-\gamma)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\gamma-\beta)}x^{\gamma-1}(1-x)^{\rho-1}(1-x)^{-\beta}}{\frac{\Gamma(\gamma)\Gamma(\rho)}{\Gamma(\gamma+\rho)}}$$

$$= \frac{\frac{\Gamma(\rho)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\rho-\beta)} x^{\gamma-1}(1-x)^{\rho-\beta-1}}{\frac{\Gamma(\gamma)\Gamma(\rho)}{\Gamma(\gamma+\rho)}}$$

$$= \frac{\Gamma(\gamma+\rho-\beta)x^{\gamma-1}(1-x)^{\rho-\beta-1}}{\Gamma(\gamma)\Gamma(\rho-\beta)}$$

$$= \frac{\Gamma(\gamma+q)x^{\gamma-1}(1-x)^{q-1}}{\Gamma(\gamma)\Gamma(q)}$$

$$= \frac{x^{\gamma-1}(1-x)^{q-1}}{\mathrm{B}(\gamma, \mathrm{q})}$$

By transformation

Let $y = bx^{1/a} \Longrightarrow x = \left(\frac{y}{b}\right)^a$

$$\frac{dx}{dy} = \frac{ay^{a-1}}{b^a}$$

$$f_y(y) = f_x(x)\frac{dx}{dy}$$

$$= \frac{\left[\left(\frac{y}{b}\right)^a\right]^{\gamma-1}\left[1-\left(\frac{y}{b}\right)^a\right]^{q-1}}{B(\gamma,q)} \cdot \frac{ay^{a-1}}{b^a}$$

$$= \frac{ay^{a\gamma-1}\left[1-\left(\frac{y}{b}\right)^a\right]^{q-1}}{b^{a\gamma}B(\gamma,q)} \quad , \quad 0 < y < b.$$

This is the pdf of generalized beta of the 1st kind.

3.2.3 Kumaraswamy distribution

Let $z = 0; \quad \alpha = \gamma; \quad \rho - \beta = q; \quad b = 1; \quad \gamma = 1$

$$F(\gamma, \beta; \gamma; x) = {}_1F_0(\beta; x) = (1-x)^{-\beta}$$

$$\frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)}{{}_3F_2\left(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1}\right)B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)} x^{\gamma-1}(1-x)^{\rho-1}F(\alpha,\beta;\gamma;x)}{B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(\gamma+\rho-\gamma)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\gamma-\beta)} x^{\gamma-1}(1-x)^{\rho-1}(1-x)^{-\beta}}{\frac{\Gamma(\gamma)\Gamma(\rho)}{\Gamma(\gamma+\rho)}}$$

$$= \frac{\frac{\Gamma(\rho)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\rho-\beta)} x^{\gamma-1}(1-x)^{\rho-\beta-1}}{\frac{\Gamma(\gamma)\Gamma(\rho)}{\Gamma(\gamma+\rho)}}$$

$$= \frac{\Gamma(\gamma+\rho-\beta)x^{\gamma-1}(1-x)^{\rho-\beta-1}}{\Gamma(\gamma)\Gamma(\rho-\beta)}$$

$$= \frac{\Gamma(\gamma+q)x^{\gamma-1}(1-x)^{q-1}}{\Gamma(\gamma)\Gamma(q)}$$

$$= \frac{x^{\gamma-1}(1-x)^{q-1}}{\mathrm{B}(\gamma,\mathrm{q})}$$

By transformation

Let $y = bx^{\frac{1}{a}} \Longrightarrow x = (\frac{y}{b})^{a}$

$$\frac{dx}{dy} = \frac{ay^{a-1}}{b^{a}}$$

$$f_y(y) = f_x(x)\,\frac{dx}{dy}$$

$$= \frac{\left[\left(\frac{y}{b}\right)^{a}\right]^{\gamma-1}\left[1-\left(\frac{y}{b}\right)^{a}\right]^{q-1}}{B(\gamma,q)} \cdot \frac{ay^{a-1}}{b^{a}}$$

$$= \frac{ay^{a\gamma-1}\left[1-\left(\frac{y}{b}\right)^{a}\right]^{q-1}}{b^{a\gamma}B(\gamma,q)} \qquad 0 < y < b$$

$$= \alpha q y^{\alpha-1}(1-y^{\alpha})^{q-1}, \qquad 0 < y < 1.$$

This is the pdf of Kumaraswamy distribution.

3.2.4 Standard arcsine distribution

Let $\quad z = 0; \quad \alpha = 0; \quad \beta = 0; \quad \gamma = 0.5; \quad \rho = 0.5$

$$\frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)}{{}_3F_2\left(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1}\right)B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)} x^{\gamma-1}(1-x)^{\rho-1}F(\alpha,\beta;\gamma;x)}{B(\gamma,\rho)}$$

$$= \frac{\frac{\Gamma(1-\alpha)\Gamma(1-\beta)}{\Gamma(1-\alpha-\beta)} \ {}_2F_1(\alpha, \beta; 0.5; x)}{\pi\sqrt{x(1-x)}}$$

$$= \frac{1}{\pi\sqrt{x(1-x)}}.$$

This is the pdf of standard arcsine distribution.

3.3 Proposed modified beta distribution

The proposed modified beta distribution is constructed based on the generalized hypergeometric function as well. It has 5 parameters $\gamma, \rho, \alpha, \beta, \sigma$. Table 3.2 shows the basic properties of the modified beta distribution:

Table 3.2: Properties of modified beta distribution

Probability Density Function	$f(x) = \frac{1}{B(\rho,\sigma)_3F_2(\alpha,\beta,\rho;\gamma,\rho+\sigma;1)}$, $\times x^{\rho-1}(1-x)^{\sigma-1} {}_2F_1(\alpha,\beta;\gamma;x)$ $0 < x < 1$
Mean	$\frac{\rho({}_3F_2(\alpha,\beta,\rho+1;\gamma,\rho+\sigma+1;1))}{(\rho+\sigma)_3F_2(\alpha,\beta,\rho;\gamma,\rho+\sigma;1)}$
Variance	$E(X^2)\text{-}[E(X)]^2$
General Moment	$\frac{B(\rho+n,\sigma)({}_3F_2(\alpha,\beta,\rho+n;\gamma,\rho+\sigma+n;1))}{B(\rho,\sigma)_3F_2(\alpha,\beta,\rho;\gamma,\rho+\sigma;1)}$

Given that

$$\int_0^1 x^{\rho-1}(1-x)^{\sigma-1}F(\alpha,\beta;\gamma;x)dx = \frac{\Gamma(\rho)\Gamma(\sigma)}{\Gamma(\rho+\sigma)} \times {}_3F_2(\alpha,\beta,\rho;\gamma,\rho+\sigma;1)$$

where $\operatorname{Re}\sigma > 0$, $\operatorname{Re}\rho > 0$, $\operatorname{Re}(\gamma+\rho-\alpha-\beta) > 0$ (Gradshteyn and Ryzhik 2007, p. 813, 7.512.5).

Then it can be easily shown that

$$\int_0^1 \frac{x^{\rho-1}(1-x)^{\sigma-1}{}_2F_1(\alpha,\beta;\gamma;x)}{B(\rho,\sigma){}_3F_2(\alpha,\beta,\rho;\gamma,\rho+\sigma;1)} = 1$$

The advantages of modified distribution are it is very flexible and relates with many existing distributions such as beta, Gauss hypergeometric, generalized beta and arcsine distributions. However, the disadvantage is the difficulty to get mathematical properties of the distribution because of its complex form of the pdf.

3.4 Proposed mixture of 2 modified lognormal distributions

The proposed mixture of 2 modified lognormal distributions consists of 7 parameters $\mu_1, \sigma_1, c_1, \mu_2, \sigma_2, c_2, p$. The mixture of 2 modified lognormal distributions can be reduced to mixture of 2 lognormal distributions when *c=0*, modified lognormal when *p=1*, and lognormal distribution when *c=0* and *p=1*. The probability density function of the mixture of 2 modified lognormal distributions is given by:

$$f(x) = \frac{2p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} \int_0^{\left(\frac{x}{e^{\mu_1}}\right)^{c1/\sigma_1}} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(\ln t)^2}{2}} dt$$
$$+ \frac{2(1-p)}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}} \int_0^{\left(\frac{x}{e^{\mu_2}}\right)^{c2/\sigma_2}} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(\ln t)^2}{2}} dt ,$$

where $\sigma_1, \sigma_2 > 0$, $\mu_1, \mu_2 \in \Re$, $-1 \le c_1, c_2 \le 1$ and $0 \le p \le 1$.

It can be shown that the integration of the probability density function of the mixture of 2 modified lognormal distributions equals to 1 since the mixture of 2 modified lognormal distribution is a transformation distribution from skewed normal distribution. A skew-normal distribution has been introduced by Azzalini (1985). Its probability density function is of the form with parameter c ($-\infty < c < \infty$).

$$f(u) = 2\phi(u)\Phi(cu)$$

where ϕ denotes a standard normal PDF and Φ a standard normal distribution function. Transforming $u = \log y$, we have the skewed lognormal pdf:

$$f(y)dy = 2 \times \frac{1}{\sqrt{2\pi} \cdot y} e^{-\frac{(\ln y)^2}{2}} \left[\int_0^{y^c} \frac{1}{\sqrt{2\pi} \cdot t} e^{-\frac{(\ln t)^2}{2}} dt \right] dy$$

Let $\ln y = \frac{\ln x - \mu}{\sigma} \rightarrow y = e^{\frac{1}{\sigma}\ln\left(\frac{x}{e^{\mu}}\right)}$; $\frac{1}{y} dy = \frac{1}{\sigma x} dx$

$$= \frac{2}{\sqrt{2\pi} \cdot \sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \left[\int_0^{\left(\frac{x}{e^{\mu}}\right)^{\frac{c}{\sigma}}} \frac{1}{\sqrt{2\pi} \cdot t} e^{-\frac{(\ln t)^2}{2}} dt \right] dx$$

This is the PDF of the modified lognormal distribution. From there, the mixture of 2 modified lognormal distribution is derived. The advantage of the mixture of 2 modified lognormal is the PDF can be flexible e.g. unimodal and bimodal. The disadvantage is it is very difficult to get mathematical properties of the distribution because of its number of parameters and the complex form of the pdf. Figure 3.2 shows the graphs of its PDF with the parameter given in Table 3.3.

Table 3.3: Parameter values for mixture of 2 modified lognormal distributions

	μ_1	σ_1	c_1	μ_2	σ_2	c_2	p
Green	1	1	-1	1	1	-1	0.8
Blue	1	1	1	1	1	1	0.8
Red	1	5	0.5	1	2	1	0.8
Black	0	1	0.2	1.5	0.1	-0.1	0.85

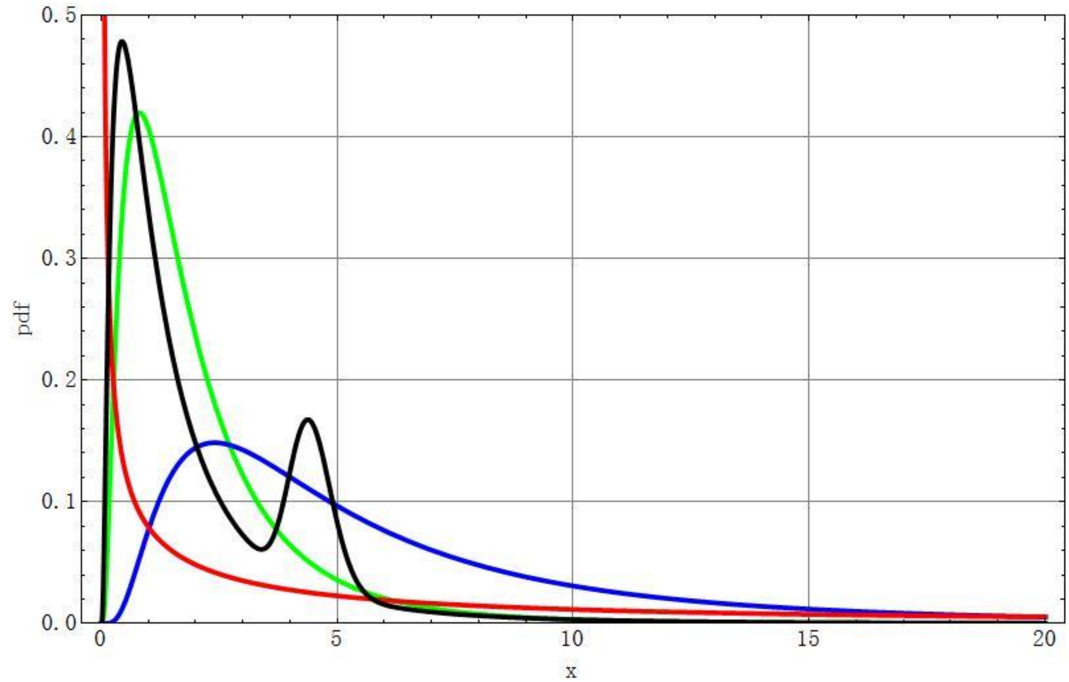

Figure 3.2: PDF of mixture of 2 modified lognormal distributions with difference parameter values

Figure 3.3 shows the relationship between the proposed distributions and some other distributions in chart with certain respective parameters.

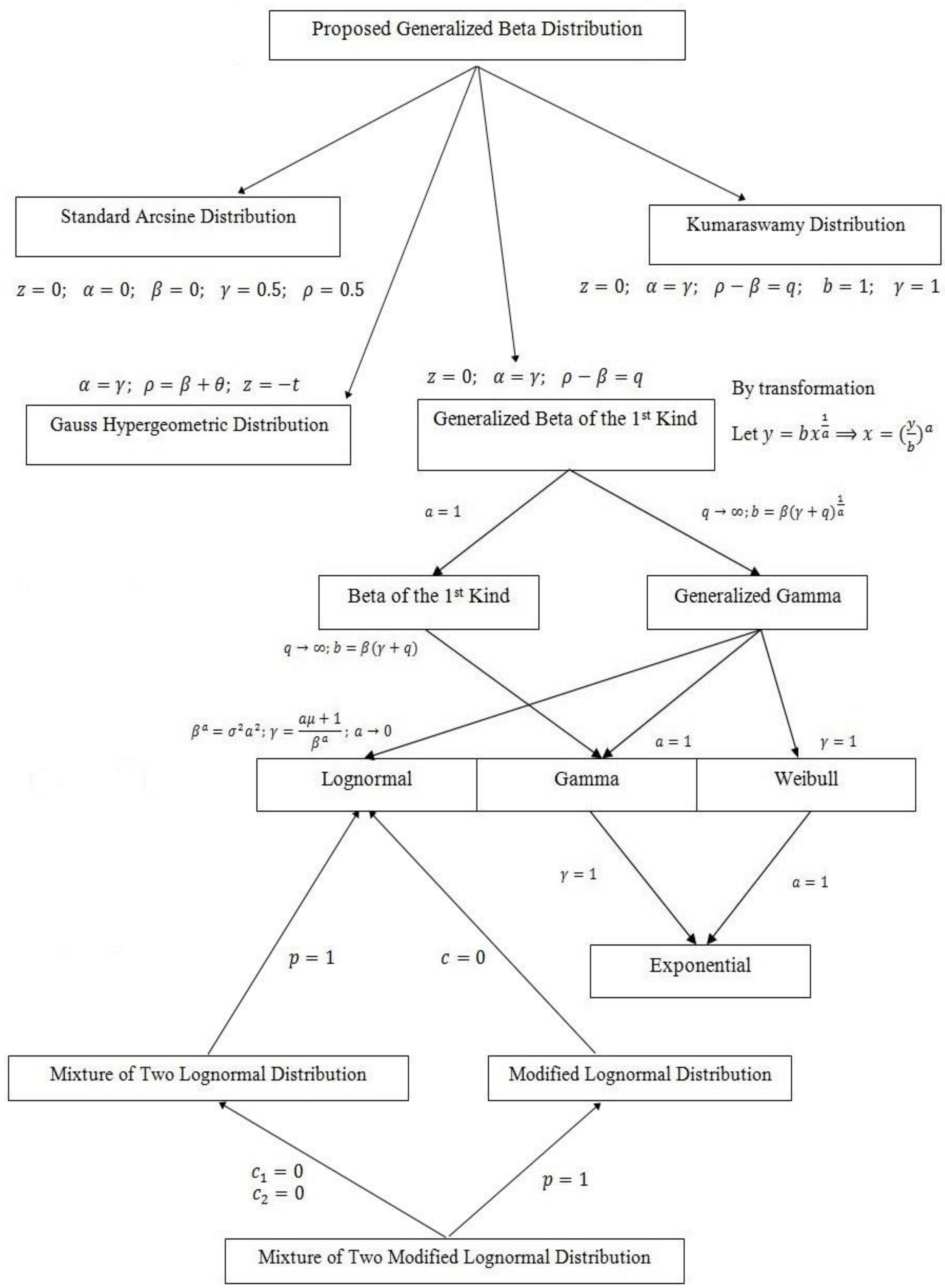

Figure 3.3: Relationship chart of various distributions

CHAPTER 4

DATA AND METHODOLOGY

4.1 Data collection and statistics (with maps)

The rainfall data were collected from Ministry of Natural Resources and Environment Malaysia and were recorded at 15 minutes interval from 1st of January 1995 to 31st of December 2004. However, the data were tabulated into daily interval as it illustrates simple method of analysis (Stern *et al*., 1982). In this study, we selected the Langat river basin as it is an important water catchment area that provides water supply to the 1.2 million people living in Kuala Lumpur and Selangor, the capital city of Malaysia.

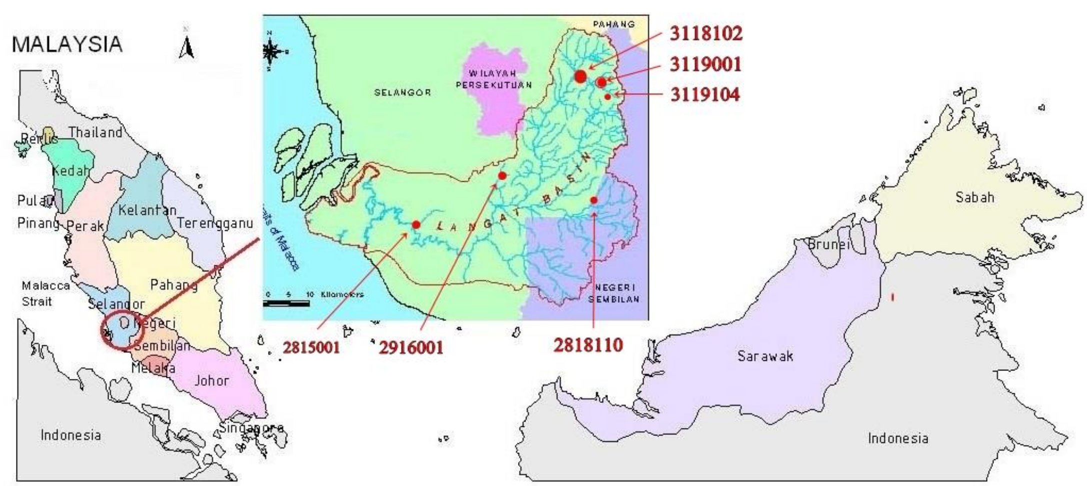

Figure 4.1: The map location for chosen rainfall stations in Selangor, Malaysia

Figure 4.1 shows the location of where all six rainfall stations 2815001, 2916001, 2818110, 3118102, 3119001 and 3119104 (in red dot) on the map of Malaysia. The Langat river basin is located at the region of Selangor and Negeri Sembilan states, where it is near to the straits of Malacca. This rainfall over this region is likely to be influenced by the south west monsoon over the north east monsoon seasons where the wind on average is below 15 knots.

For simplicity, the stations will be renamed using alphabets instead of station number as follows:

Table 4.1: Renamed rainfall station

Original Rainfall Station Number	Renamed Rainfall Station
2815001	A
2916001	B
2818110	C
3118102	D
3119001	E
3119104	F

We use one of the rainfall stations (Station D) that is located at Sungai Lui of the Langat river basin. The rainfall data set has 2057 days (56.31%) of zero rainfall days out of 3653 days (10 years) recorded. Non-zero rainfall is studied and the rainfall volume is measured in millimeters (mm). During the 10 years period, the highest recorded volume of rainfall per day is 121.5mm. It has mean of 12.03mm with variance 264.57mm, while median, mode, kurtosis and skewness are

5.50mm, 0.50mm, 6.80mm and 2.37mm respectively. The modelling of rainfall volume will only to be fitted on rainfall data from Station D. Figure 4.2 shows the histogram of the relative frequency versus rainfall volume.

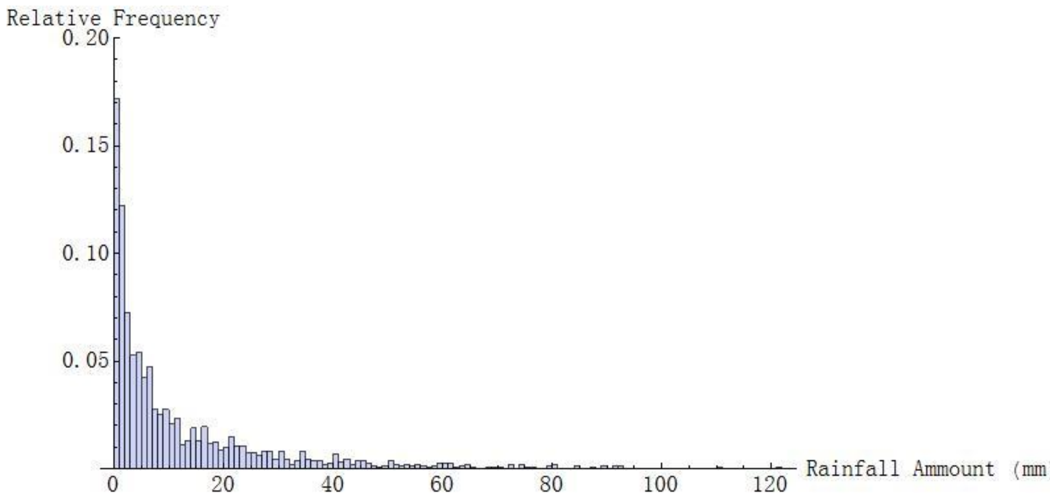

Figure 4.2: Relative frequency versus rainfall volume (mm)

With the same source of data from year 1995-2004, we tabulated the data into number rainfall occurrence or dry/wet spell. In this instance, we consider the day as dry day when the rainfall volume is less than 1mm per day. A dry spell is the number of days consecutively having dry day after each wet day. While for the wet spell, it is defined by the number of having wet days consecutively after each dry day. Therefore dry/wet spell data follow a discrete distribution. The modelling of rainfall occurrence will be fitted on data from 6 rainfall stations which are the Station A, Station B, Station C, Station D, Station E and Station F. Figure 4.3 shows the histogram of the relative frequency versus number of dry spell for Station D while Figure 4.4 shows the histogram of the relative frequency versus number of wet spell for Station D.

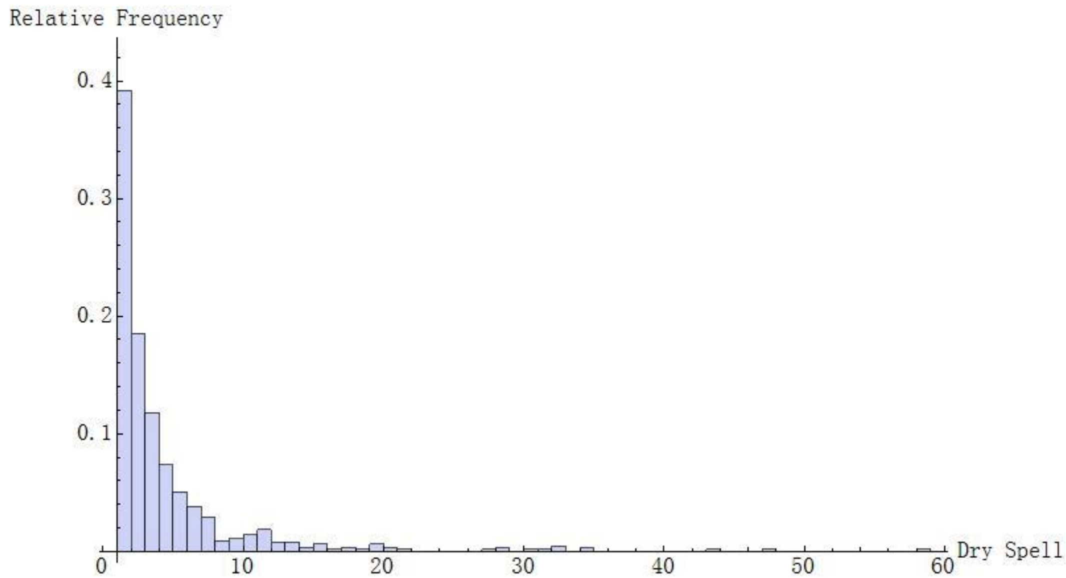

Figure 4.3: Relative frequency versus dry spell days

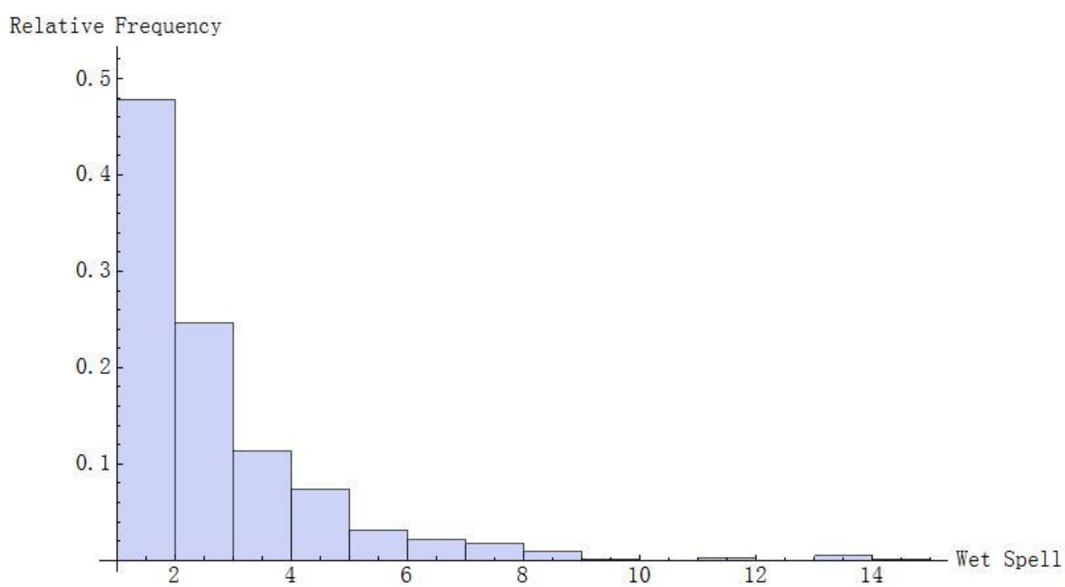

Figure 4.4: Relative frequency versus wet spell days

4.2 Parameter estimation

The Maximum Likelihood Estimation (MLE) is the backbone of statistical estimation and the basis for deriving estimation of unknown parameters for a given model or a set of data. MLE method is used since MLE is asymptotically normally distributed and able to provide consistent approach to the problem of parameter estimation. Therefore it is optimal for large samples and has the minimum variance unbiased estimators when the sample size is large.

Suppose $X_1, X_2, \cdots, X_n$ is the random variables with its probability density function

$$f_{\boldsymbol{\theta}}(X_1, X_2, \cdots, X_n) = f(X_1, X_2, \cdots, X_n|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is the vector $(\theta_1, \theta_2, \cdots, \theta_k)$ for the function.

Given observed values $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, the likelihood of $\boldsymbol{\theta}$ which is the probability of observing the given data as a function of $\boldsymbol{\theta}$ is given by:

$$L(\boldsymbol{\theta}; x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n|\boldsymbol{\theta}).$$

If $X_1, X_2, \cdots, X_n$ are independent and identically distributed with $f_{\boldsymbol{\theta}}(x_i)$, then the likelihood function will be $L(\boldsymbol{\theta}; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$. The maximum likelihood estimator of $\boldsymbol{\theta}$ is the values of parameters where it maximized the likelihood function $L(\boldsymbol{\theta}; x_1, x_2, \cdots, x_n)$. To some extent maximizing the product function can be quite tedious, we often maximize the log-likelihood since logarithm is also an increasing function.

$$l(\theta) = \sum_{i=1}^{n} \log\left(f(x_i \mid \theta)\right).$$

In regular cases, we equivalently maximize $L(\boldsymbol{\theta}; x_1, x_2, \cdots, x_n)$ by solving

$$\frac{\partial}{\partial \theta} \log L(\theta; x_1, x_2, \cdots, x_n) = 0,$$

which called the likelihood equation. However, the complexity of the proposed generalized distributions create serious complications for parameter estimation. This resulted in the likelihood function formed could be numerically intractable as the likelihood function cannot be solved analytically. Therefore, a global optimization method, simulated annealing algorithm is used in the assist to find the maximize likelihood estimator (Goffe, 1996). Simulated annealing algorithm is a form of stochastic optimization has been chosen since it is explicitly designed for multi-optimal points function. It optimizes the function by exploring the entire surface of the function and it makes fewer assumptions than classical optimization methods such as Newton-Raphson, the Davidon-Fletcher-Powel, and the simplex method. The simulated annealing algorithm method will evaluate the cost function by random so that it is possible for the transitions to be out of local maximum and it is able to differentiate the “gross behavior” and the finer “wrinkles” of the function. The algorithm could first find the global maximum of the function domain, and then it will continue with its finer details in finding the optimal maximum, or a point that is very near to it. Corana *et al*. (1989) has proven that it is much more robust and reliable than those classical algorithms as it always able to discover the optimum point or a point that is close to it. However, the drawback will be it takes longer time to compute. Since the proposed generalized distributions have more than 3 parameters, the simulated annealing algorithm is the alternate solution to compute the MLE.

4.3 Model selection criteria

We considered Akaike's information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978) as measures of relative quality among all the distributions for each station. The AIC and BIC are used to determine the best fit distributions. The AIC and BIC consider the statistical goodness of fit with the number of parameters of the model taken into account as a penalty for increasing in the number of parameters, with BIC penalizes more on the number of parameters compared to AIC.

$$AIC = 2k - 2\ln L,$$

$$BIC = k\ln n - 2\ln L,$$

where

k = number of parameter in the model,

L= maximum likelihood value,

n = number of data .

Therefore, with the same value of maximum likelihood, the less parameter will return a smaller value for AIC and BIC. In other words, the model that with the minimum index of AIC and BIC among others will be preferred as the adequate fit to the data. Noted that the values of AIC and BIC themselves only provide the comparison of the quality of each model relative to other distributions, whose the values by themselves do not provide any useful information.

4.4 Research procedures

The modelling of rainfall data in Langat river basin will be separated into modelling the daily rainfall volume and rainfall occurrence. Both will be carried out as follows:

(i) Estimate the parameters of each distribution discussed earlier by using maximum likelihood estimation method.

(ii) An optimal model will be chosen based on the criteria of AIC and BIC

(iii) Based on the parameters estimated, the graph of the PDF of the distributions will be plotted and compared with the histogram of data.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Overview

The first part of this chapter will discuss the modelling of rainfall volume using continuous distributions on station D. The results of the AIC and BIC of all distributions will be organized in a table and graphs for comparisons. The graphs of every candidate model's PDF will be plotted together with the histogram of data for comparison. Next, the best model will be used to examine whether there is changes on rainfall pattern by using hypothesis testing.

The second part of this chapter will focus on modelling of the rainfall occurrence for both dry and wet spells using discrete distributions on the 6 chosen rainfall stations (A, B, C, D, E and F). The results will be organized in the tables. However, only the graphs of Hurwitz-Lerch Zeta (HLZ) distribution will be plotted for the comparison with the histogram of the rainfall data since HLZ distribution have the best fitting among all in general. In addition, it is relatively new model in exploring the hydrology analysis compared with other existing distributions that have already been studied by many researchers. Therefore we would like to emphasis on the performance of the HLZ distribution in model fitting.

5.2 Results on daily rainfall volume modelling

Based on the maximum likelihood estimation that computed from the simulated annealing algorithm, we obtained the maximum log-likelihood function's values. We compiled and listed out the results of AIC and BIC computed from the maximum log-likelihood in Table 5.1.

Table 5.1: Comparison of AIC and BIC for difference distributions

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Transformed Beta distribution	-5492.99	10989.99	11000.74
2	Transformed Gauss hypergeometric distribution	-5458.32	10924.66	10942.59
3	Generalized beta of the 1st kind distribution	-5424.28	10854.56	10870.69
4	Lognormal distribution	-5404.83	10813.66	10824.41
5	**Mixture of two lognormal distribution**	**-4442.28**	**8894.56**	**8921.44**
6	Modified lognormal distribution	-5404.52	10815.06	10831.18
7	Exponential distribution	-5566.35	11134.70	11140.08

8	Mixture of two exponential distribution	-5391.38	10788.75	10804.88
9	Transformed Kumaraswamy distribution	-5470.50	10945.00	10955.75
10	Pareto distribution	-5446.70	10897.40	10908.15
11	Marginal of linear and angular distribution	-5395.70	10795.40	10806.15
12	Gamma distribution	-5466.71	10937.42	10948.17
13	Proposed generalized beta distribution	-5418.35	10848.71	10880.96
14	Modified beta distribution	-5467.35	10944.69	10971.57

From the results based on the Table 5.1, the best fit model has been bolded and it is obvious that the mixture of two lognormal distributions has the lowest index in both AIC and BIC. The comparisons in results of the AIC and BIC among all the distributions are also shown in Figure 5.1 and Figure 5.2.

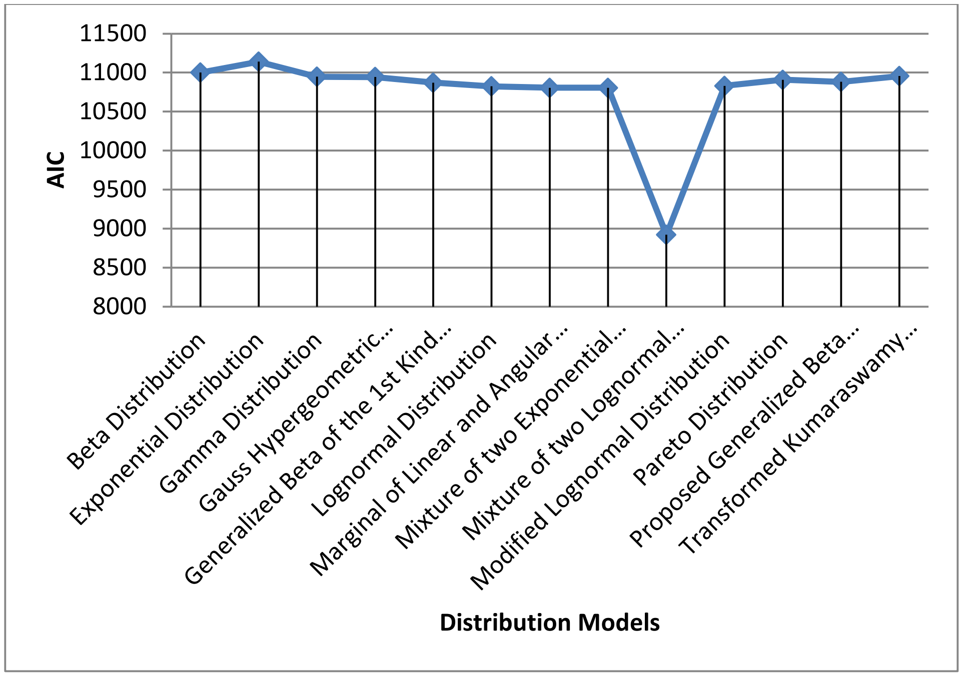

Figure 5.1: Graph of AIC for different distributions

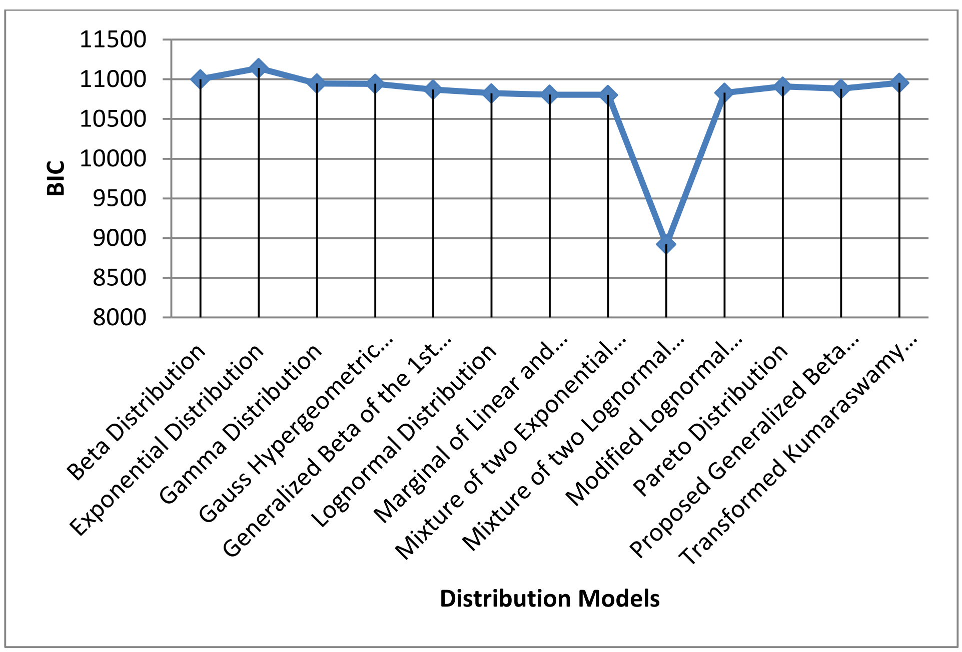

Figure 5.2: Graph of BIC for different distributions

From both the graphs on Figure 5.1 and Figure 5.2, it is very obvious that the mixture of two lognormal has a significant different in both the index of AIC and BIC compared with other distributions. Different distribution might be selected by different criteria. However for this data set, the same distribution which is mixture of two lognormal was selected using different AIC and BIC criteria.

5.2.1 Graphs of histogram and probability density function

Graphs of probability density function versus rainfall volume are plotted for every distribution and compare with the histogram of rainfall data from station D. Note that since the range of x for the beta-type distribution is from 0 to 1, therefore transformation for beta type distribution is needed to extend the range of x to fit into the data set. The maximum daily rainfall volume at station D throughout the 10 years periods is 121.4mm. It is prudent to assume the maximum rainfall volume per day to be 150mm. To transform, we let $x = \frac{y}{150}$, then $dx = \frac{1}{150} dy$,

$$\int_0^x f(x)dx = \int_0^{y/150} \frac{1}{150} f(y/150)dy, \text{ where } 0 \le y \le 150 .$$

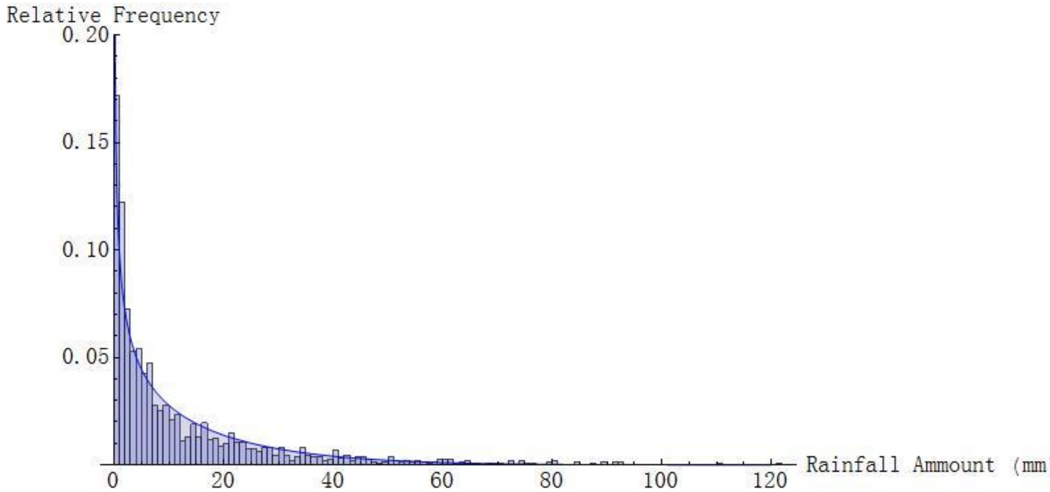

Figure 5.3: PDF of transformed beta distribution and histogram for rainfall amount

Figure 5.3 shows the histogram of the rainfall amount and the transformed beta distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(y) = \frac{(y/150)^{\gamma-1}(1-y/150)^{q-1}}{150B(\gamma,q)}$$

$$\gamma = 0.6115$$

$$q = 6.7800.$$

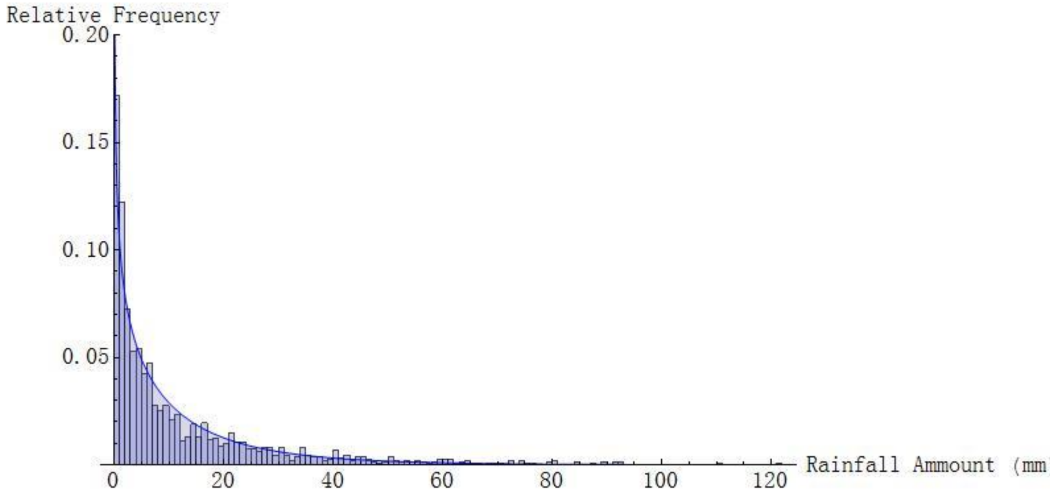

Figure 5.4: PDF of transformed Gauss hypergeometric distribution and histogram for rainfall amount

Figure 5.4 shows the histogram of the rainfall amount and the transformed Gauss hypergeometric distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(y) = \frac{(y/150)^{\gamma-1}(1-y/150)^{\theta-1}(1+ty/150)^{-\sigma}}{150\,{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)B(\theta,\gamma)}$$

$$\gamma = 0.7011$$

$$\theta = 0.7044$$

$$\sigma = 15.3543$$

$$t = 0.6741.$$

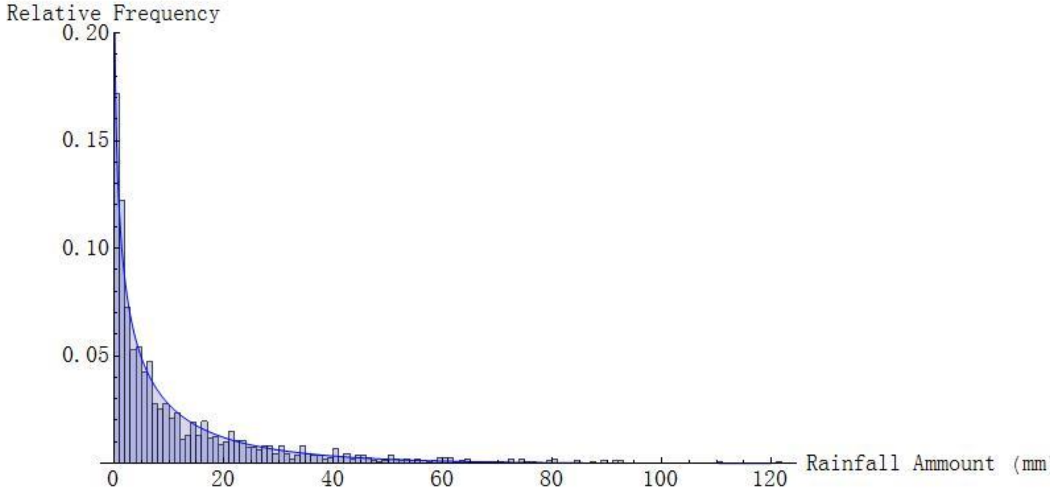

Figure 5.5: PDF of generalized beta of the 1st kind distribution and histogram for rainfall amount

Figure 5.5 shows the histogram of the rainfall amount and the generalized beta of the 1st kind distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \frac{a x^{a\gamma-1}(1-(x/b)^{a})^{q-1}}{b^{a\gamma}B(\gamma,q)}$$

$$\alpha = 0.0389$$

$$\gamma = 32.2878$$

$$q = 4.5343$$

$$b = 150.$$

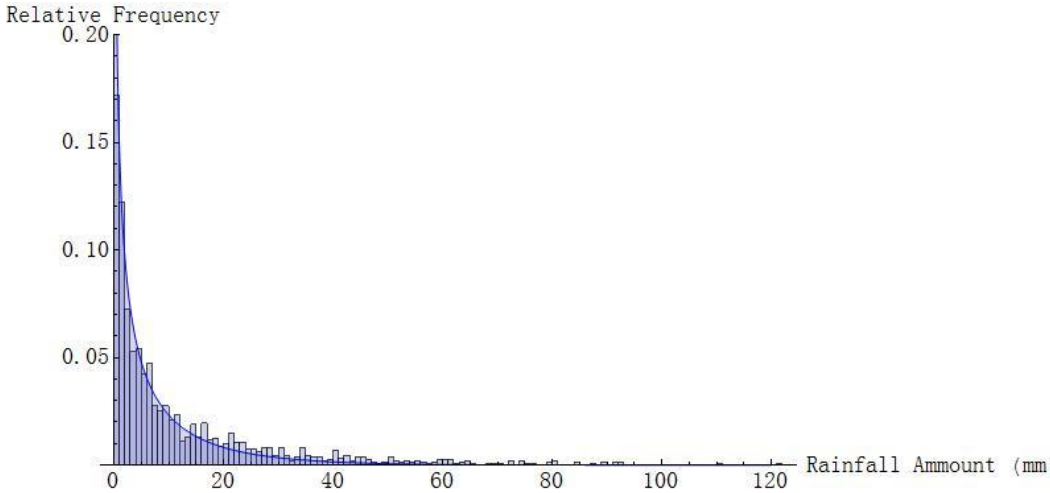

Figure 5.6: PDF of transformed Kumaraswamy distribution and histogram for rainfall amount

Figure 5.6 shows the histogram of the rainfall amount and the transformed Kumaraswamy's distribution probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(y) = [\alpha\beta(y/150)^{\alpha-1}(1-y/150)^{\alpha})^{\beta-1}]/150$$

$$\alpha = 0.0625$$

$$\beta = 0.6630.$$

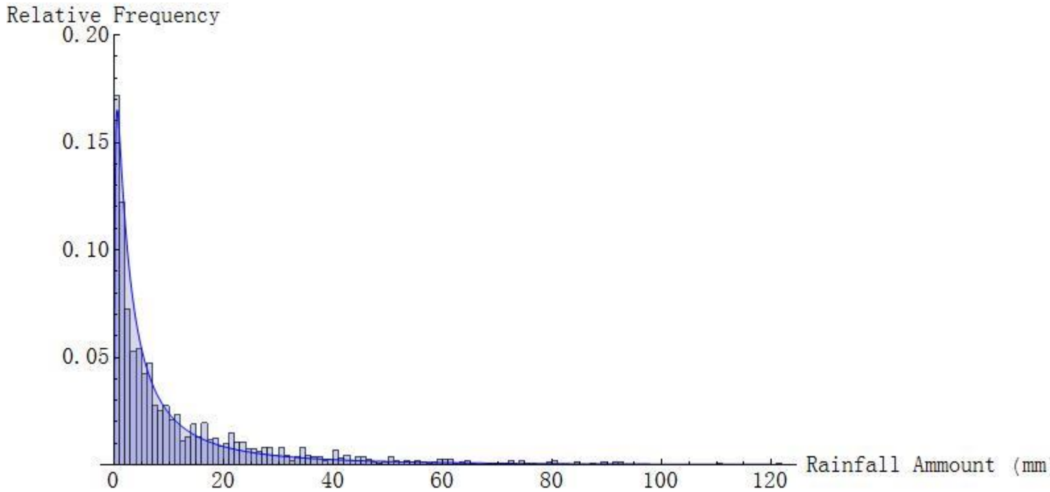

Figure 5.7: PDF of lognormal distribution and histogram for rainfall amount

Figure 5.7 shows the histogram of the rainfall amount and the lognormal distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

$$\mu = 1.5801$$

$$\sigma = 1.4733.$$

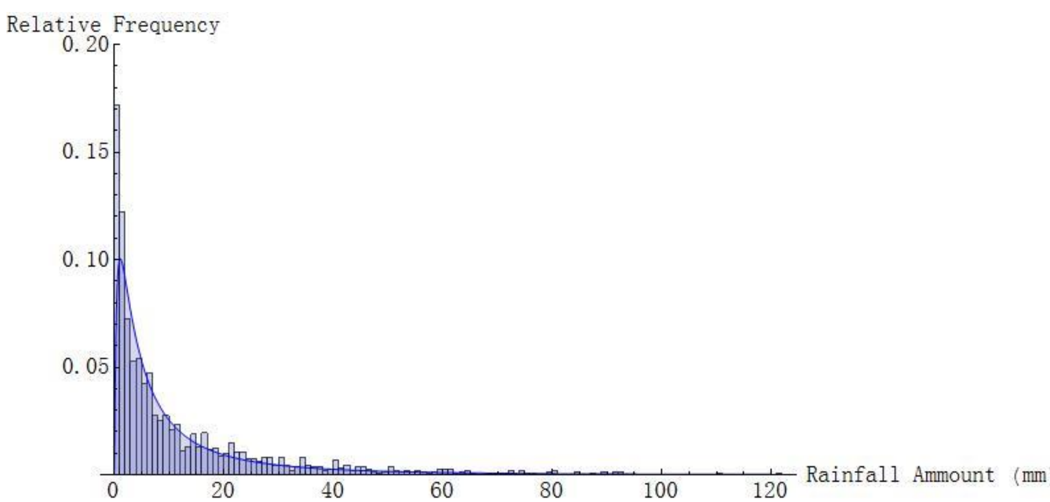

Figure 5.8: PDF of mixture of 2 lognormal distributions and histogram for rainfall amount

Figure 5.8 shows the histogram of the rainfall amount and the mixture of 2 lognormal distributions' probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}}$$

μ_1 = 1.8561

σ_1 = 1.3155

μ_2 = -0.6931

σ_2 = 0.0010

p = 0.8917.

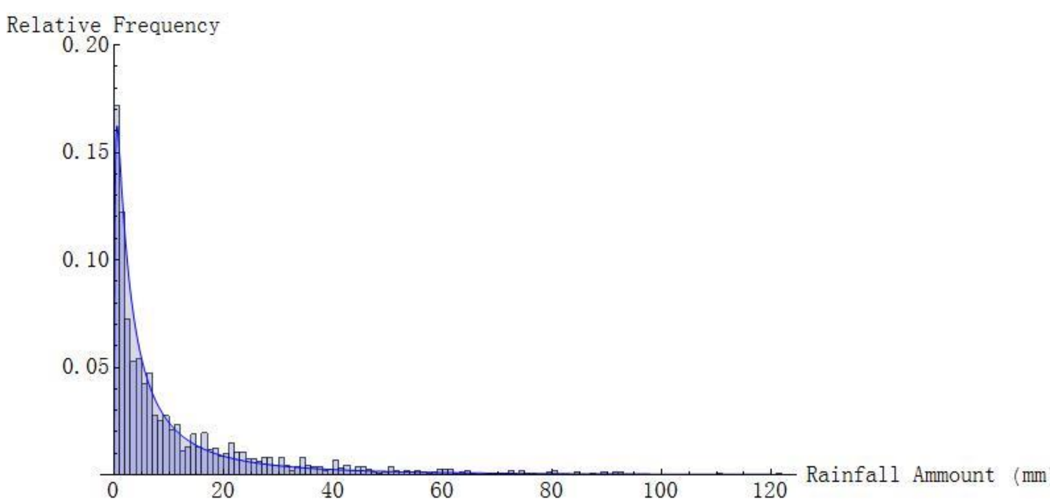

Figure 5.9: PDF of modified lognormal distribution and histogram for rainfall amount

Figure 5.9 shows the histogram of the rainfall amount and modified lognormal distributions' probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = 2\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \times \int_0^{\left(\frac{x}{e^\mu}\right)^{\frac{c}{\sigma}}} \frac{1}{\sqrt{2\pi t}}e^{-\frac{(\ln t)^2}{2}}dt$$

$$\mu = 2.3638$$

$$\sigma = 1.6687$$

$$c = -0.7272.$$

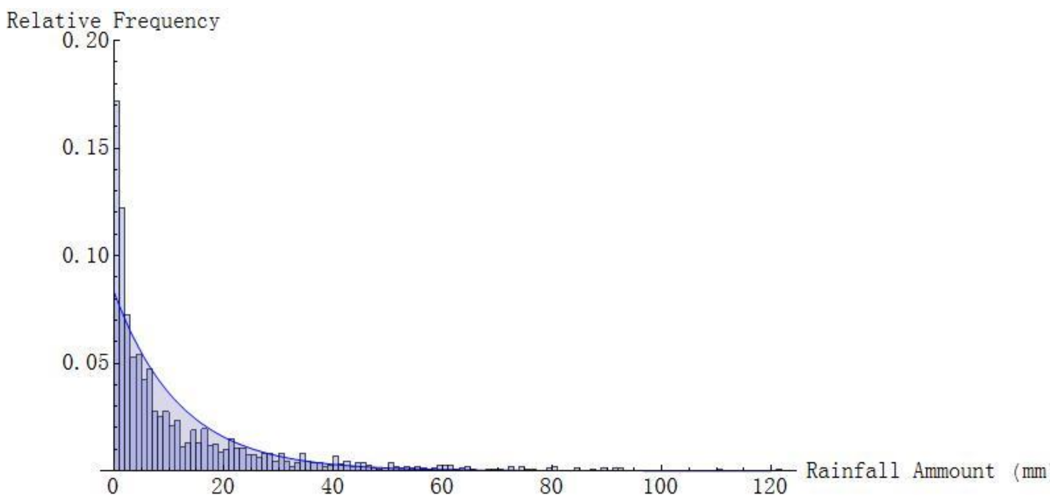

Figure 5.10: PDF of exponential distribution and histogram for rainfall amount

Figure 5.10 shows the histogram of the rainfall amount and exponential distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \beta e^{-\beta x}$$

$$\beta = 0.0831$$

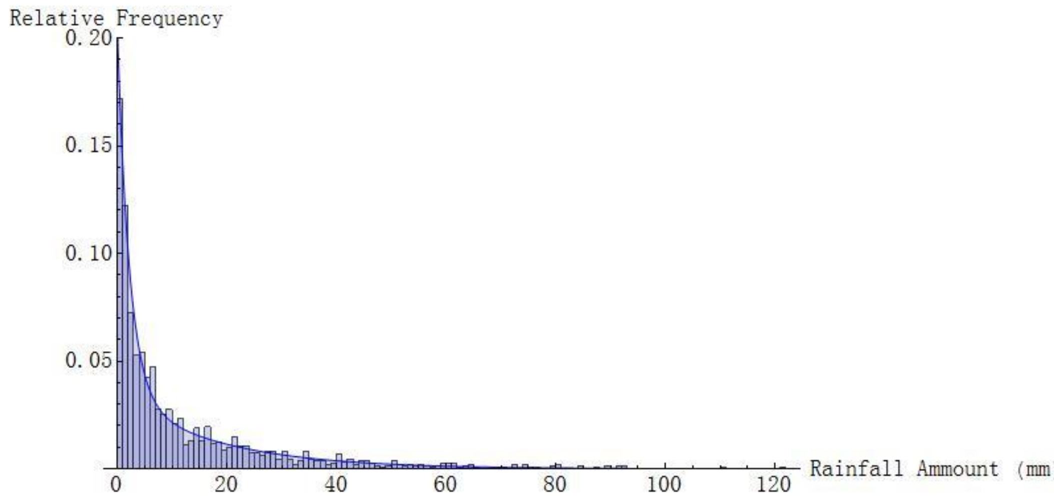

Figure 5.11: PDF of mixture of 2 exponential distributions and histogram for rainfall amount

Figure 5.11 shows the histogram of the rainfall amount and the mixture of 2 exponential distributions' probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = p\beta_1 e^{-\beta_1 x} + (1-p)\beta_2 e^{-\beta_2 x}$$

$$\beta_1 = 0.0556$$

$$\beta_2 = 0.4546$$

$$p = 0.6231.$$

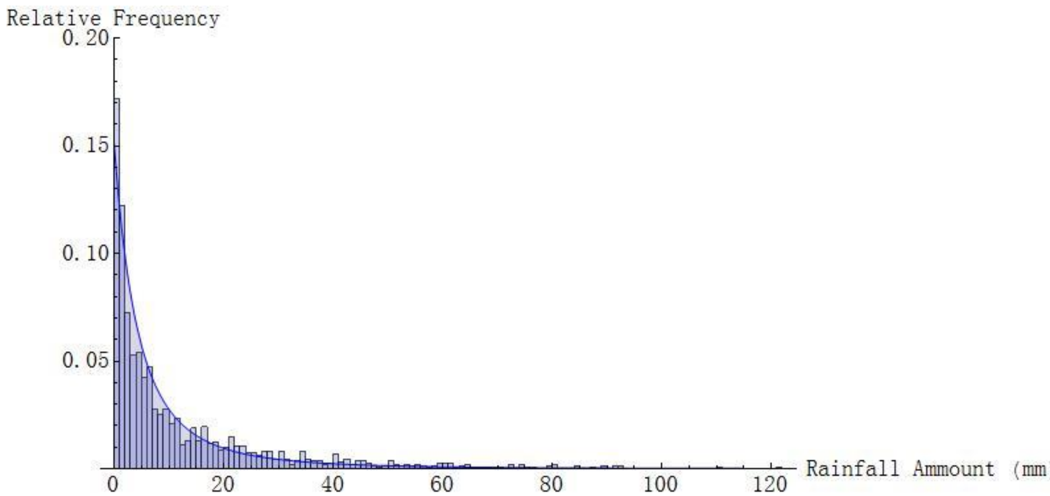

Figure 5.12: PDF of Pareto distribution and histogram for rainfall amount

Figure 5.12 shows the histogram of the rainfall amount and the Pareto distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \frac{\alpha\beta}{(1+\beta x)^{\alpha+1}}$$

$$\alpha = 1.9059$$
$$\beta = 0.0794.$$

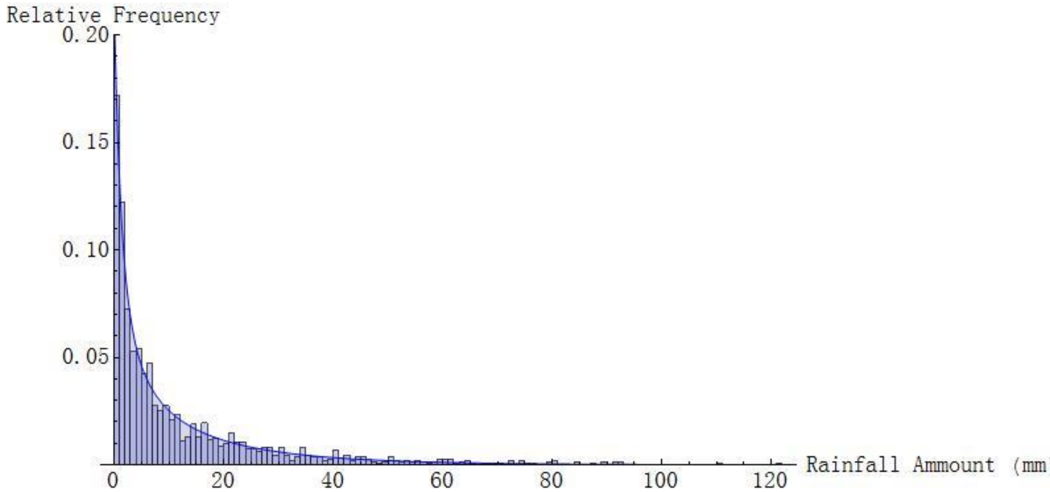

Figure 5.13: PDF of marginal of linear and angular distribution and histogram for rainfall amount

Figure 5.13 shows the histogram of the rainfall amount and the marginal of linear and angular distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \sqrt{\lambda^2 - \kappa^2}\, I_0(\kappa x) e^{-\lambda x}$$

$$\lambda = 0.5483$$

$$\kappa = 0.5051$$

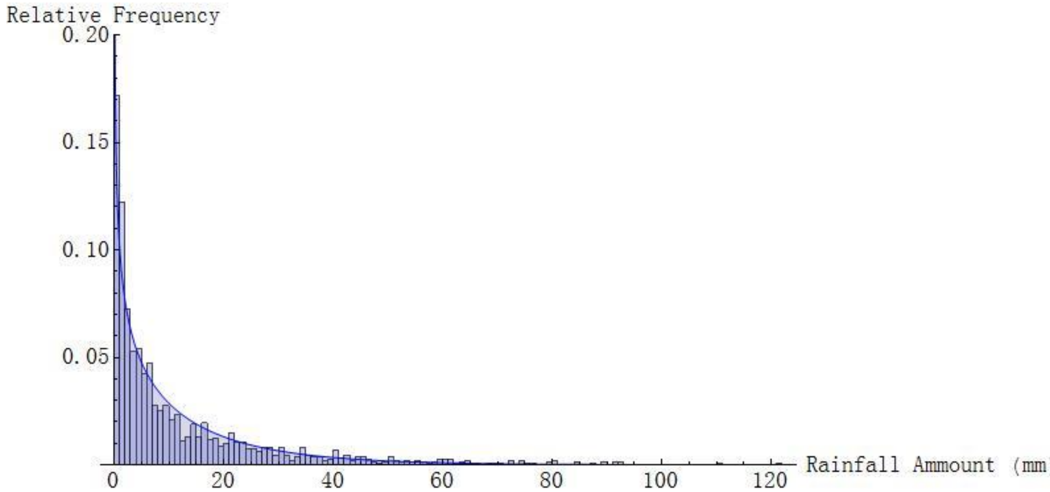

Figure 5.14: PDF of gamma distribution and histogram for rainfall amount

Figure 5.14 shows the histogram of the rainfall amount and the gamma distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$$

$$k = 0.66998$$

$$\theta = 17.9609.$$

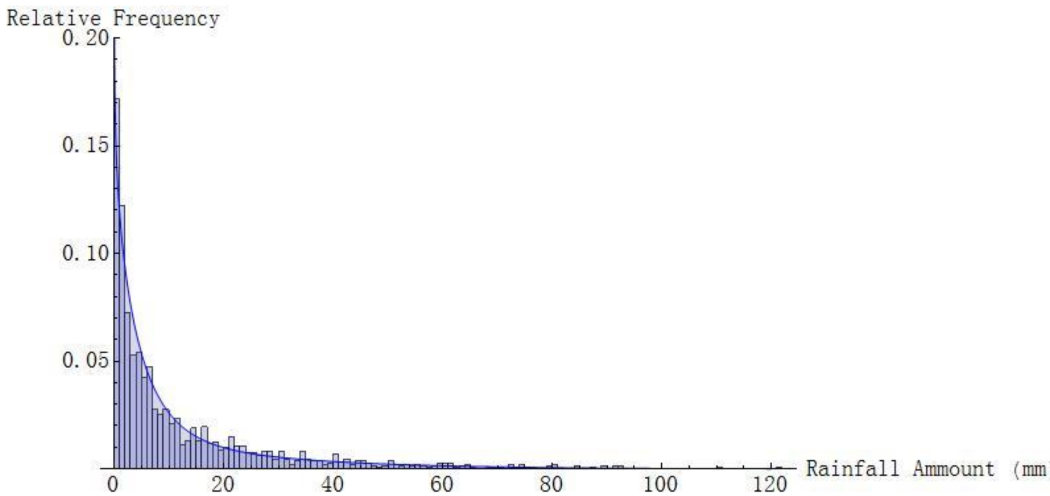

Figure 5.15: PDF of transformed proposed generalized beta distribution and histogram for rainfall amount

Figure 5.15 shows the histogram of the rainfall amount and the transformed proposed generalized beta distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(y) = \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}(y/150)^{\gamma-1}(1-y/150)^{\rho-1}(1-zy/150)^{-\sigma}F(\alpha,\beta;\gamma;y/150)}{150\,{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;z/z-1)B(\gamma,\rho)}$$

γ = 0.8743

ρ = 5.125

α = -29.9999

β = 0.6393

σ = 6.8720

z = 0.29588.

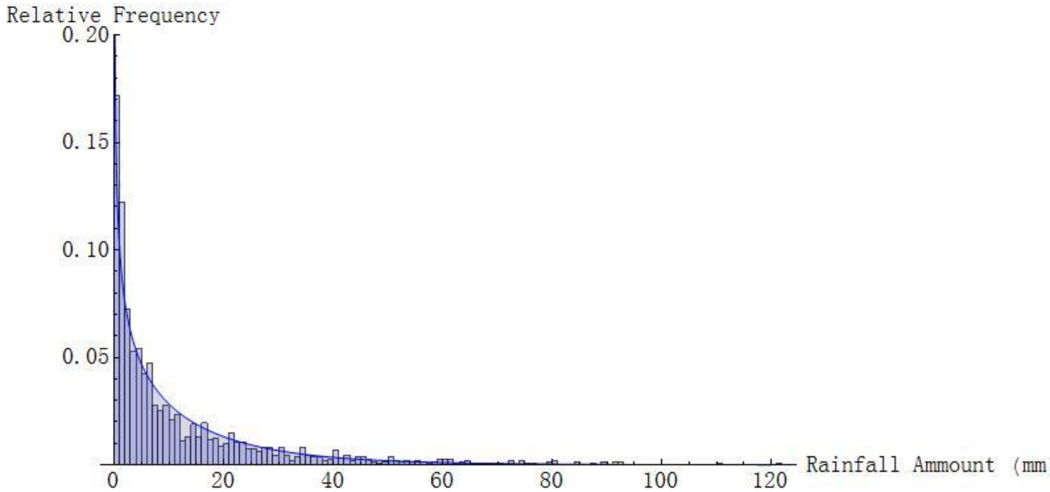

Figure 5.16: PDF of transformed modified beta distribution and histogram for rainfall amount

Figure 5.16 shows the histogram of the rainfall amount and the modified beta distribution's probability density function using the estimated parameters obtained by using the MLE approach. The PDF and estimated parameters are as follows:

$$f(y) = \frac{(y/150)^{\rho-1}(1-y/150)^{\sigma-1}\,{}_2F_1(\alpha,\beta;\gamma;y/150)}{B(\rho,\sigma)\,{}_3F_2(\alpha,\beta,\rho;\gamma,\rho+\sigma;1)}$$

$$\gamma = 0.0009$$
$$\rho = 0.6543$$
$$\alpha = 5.3545$$
$$\beta = 0.0001$$
$$\sigma = 9.4622$$

Since both the proposed generalized beta and modified beta distribution are the beta-type distribution, therefore transformation is needed to transform the maximum of the variable from 1 to 150 as well.

5.2.2 Discussion on daily rainfall volume modelling

Based on the selection of minimum AIC as the model selection criterion, Table 5.1 shows that mixture of 2 lognormal distributions has the lowest AIC and BIC values which are 8894.56 and 8921.44 among all the distributions compared. This result indicates that mixture of 2 lognormal distributions is an optimal fit to the rainfall data among the distributions considered. Although from Figure 5.8, the mixture of 2 lognormal distributions' PDF may not fit perfectly well during the low rainfall amount, that should not affect much on the heavy rainfall modelling. This is because in the view point of water resources management, it is more important to understand the rainfall behavior that brings high amount of rainfall. Note that the mixture of 2 modified lognormal distributions has not been estimated, due to its complexity of 7 parameters that make it very difficult in computing of its maximum likelihood. However, since mixture of 2 modified lognormal is a general model to mixture of 2 lognormal distributions, it is presumed that the mixture of 2 modified lognormal will perform close to the mixture of 2 lognormal distributions.

Among the beta-type distributions, the proposed generalized beta distribution was fitted better than the other beta-type distributions, in this rainfall data set. Based on the plotted graphs, the proposed generalized beta distribution, generalized beta of the 1st kind and Kumaraswamy distribution are all generally fitted well in the rainfall data. Beta-type distributions generally have longer tail as it can be seen in the plotted graphs. The longer tails show that they have the higher tendency to have high rainfall amount per day, which usually resulted in flash flood due to heavy rainfall. Therefore these models are better fit for natural disaster prevention.

Generally a mixture distribution will create a longer tail as compared to its sole distribution (e.g. mixture of exponential distribution versus exponential distribution) and performs better than its sole distribution. It has a longer tail because it involved more parameters that able to give its probability density function's curve more versatile to describe a data set. The reason a mixture distribution will have a better fit is because of the seasonal effect of rainfall in Malaysia that can be divided to dry and wet season, or monsoon season. Pareto distribution is also a mixture of exponential and gamma distribution. Therefore Pareto distribution is better fit than exponential and gamma distributions.

5.2.3 Change of rainfall pattern within study period

Since the study period is from 1995 to 2004, we do concern if there is a change of rainfall pattern throughout this 10 years. Therefore, a hypothesis testing is carried out to examine on the change of rainfall pattern. The mixture of 2 lognormal distributions has the best fit among all the other continuous distributions that has been compared. Therefore it is chosen to test on the change of climate by using hypothesis test. The hypothesis test will be based on likelihood ratio test under the assumption that the rainfall distributions of the two seasons are given by

$$f_j(x) = \frac{p_j}{\sqrt{2\pi}\sigma_{1j}x} e^{-\frac{(\ln x - \mu_{1j})^2}{2\sigma_{1j}^2}} + \frac{1-p_j}{\sqrt{2\pi}\sigma_{2j}x} e^{-\frac{(\ln x - \mu_{2j})^2}{2\sigma_{2j}^2}}, \ j = 1,2.$$

Let $\theta_j = (\mu_{1j}, \sigma_{1j}, \mu_{2j}, \sigma_{2j}, p_j)'$, *j=1,2,* where θ_1 is the parameter vector estimated for the first 5 years, and θ_2 is the parameter vector estimated for the second 5 years, out of 10 years. Let $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, p\ )'$ be the parameter vector estimated for the whole 10 years. Then, null vs. alternative hypotheses are

$$H_0 : \theta_1 = \theta_2 \quad \text{vs.} \qquad H_1 : \theta_1 \neq \theta_2 .$$

The likelihood of θ_j is

$$L(\theta_j) = \prod_{i=1}^{n_j} \left(\frac{p_j}{\sqrt{2\pi}\sigma_{1j} x_i} e^{-\frac{(\ln x_i - \mu_{1j})^2}{2\sigma_{1j}^2}} + \frac{1-p_j}{\sqrt{2\pi}\sigma_{2j} x_i} e^{-\frac{(\ln x_i - \mu_{1j})^2}{2\sigma_{1j}^2}} \right),$$

with joint likelihood of θ_1 and θ_2 is given by

$$L(\theta_1, \theta_2) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} L(\theta_j),$$

and that of θ is

$$L(\theta) = \prod_{i=1}^{n} \left(\frac{p}{\sqrt{2\pi}\sigma_1 x_i} e^{-\frac{(\ln x_i - \mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2 x_i} e^{-\frac{(\ln x_i - \mu_2)^2}{2\sigma_2^2}} \right),$$

where n_j is the number of days for positive rainfall amount during first and second 5 years, and $n = n_1 + n_2$ for total 10 years. The maximized log-likelihood values of θ_1, θ_2 and θ are:

$$\ln L(\theta_1) = -1869.52,$$

$$\ln L(\theta_2) = -2540.28 ,$$

$$\ln L(\theta) = -4442.28 .$$

Therefore, $\ln L(\theta_1, \theta_2) = \ln L(\theta_1) + \ln L(\theta_2) = -4409.80,$ and we have

Test statistic $D = -2\ln\left(\frac{L(\theta)}{L(\theta_1,\theta_2)}\right) = -2[\ln L(\theta) - \ln L(\theta_1,\theta_2)] = 64.96$

$D \sim \chi^2_{\alpha=0.05;\nu=10-5}$; P-value < 0.00001.

Since P-value < α = 0.05, therefore null hypothesis is rejected. This may suggest that there is change in climate throughout the 10 years from statistical point of view.

The same hypothesis test is carried out on lognormal distribution

$$f_j(x) = \frac{1}{\sqrt{2\pi}\sigma_j x} e^{-\frac{(\ln x - \mu_j)^2}{2\sigma_j^2}}, j = 1, 2.$$

Let $\theta_j = (\mu_j, \sigma_j)'$, *j=1,2*, where θ_1 is the parameter vector estimated for the first 5 years, and θ_2 is the parameter vector estimated for the second 5 years, out of 10 years. Let $\theta = (\mu, \sigma)'$ be the parameter vector estimated for the whole 10 years. Then, null vs. alternative hypotheses are

$$H_0 : \theta_1 = \theta_2 \quad \text{vs} \quad H_1 : \theta_1 \neq \theta_2$$

The joint likelihood of θ_1 and θ_2 is written by:

$$L(\theta_1,\theta_2) = L(\theta_1)L(\theta_2),$$

where the likelihood of θ_j is

$$L(\theta_j) = \prod_{i=1}^{n_j}\left(\frac{1}{\sqrt{2\pi}\sigma_j x_i} e^{-\frac{(\ln x_i - \mu_j)^2}{2\sigma_j^2}}\right)$$

where n_j is the number of days for positive rainfall amount during first and second 5 years, and $n = n_1 + n_2$ for total 10 years. The likelihood function is written by $L(\theta)$ for the total lognormal rainfall. The maximized log-likelihood values of θ_1, θ_2 and θ are:

$$\ln L(\theta_1) = -2664.53 \ ,$$

$$\ln L(\theta_2) = -2736.98 \ ,$$

$$\ln L(\theta) = -5404.83 \, .$$

Therefore, $\ln L(\theta_1, \theta_2) = \ln L(\theta_1) + \ln L(\theta_2) = -5401.51$, and we have

$$\text{Test statistic } D = -2\ln\left(\frac{L(\theta)}{L(\theta_1, \theta_2)}\right) = -2[\ln L(\theta) - \ln L(\theta_1, \theta_2)] = 6.64$$

$$D \sim \chi^2_{\alpha=0.05;\nu=10-5} \ ; \text{P-value} = 0.036153.$$

Since P-value < α = 0.05, therefore null hypothesis is rejected. This shows the same result with mixture of two lognormal distributions as it may suggest that there is change in rainfall pattern throughout the 10 years from statistical point of view.

5.3 Results on rainfall occurrence modelling

Based on the maximum likelihood estimation that computed from the simulated annealing algorithm, we obtained the maximum log-likelihood function's values. We compiled and listed out the results of AIC and BIC computed from the maximum log-likelihood in Table 5.2.1 to Table 5.7.2.

Table 5.2.1: Result of the model selection at Station A (Dry Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	**Hurwitz-Lerch Zeta distribution**	**-1372.95**	**2751.903**	**2765.402**
2	Eggenberger-Polya distribution	-1394.18	2792.366	2801.366
3	Logarithmic distribution	-1381.90	2765.809	2770.308
4	Truncated Poisson distribution	-2351.63	4705.251	4709.75
5	Geometric distribution	-1440.01	2882.019	2886.518

From Table 5.2.1, it is shown that the HLZ distribution has the best fit among all from the criteria of both AIC and BIC at Station A under dry spell modelling.

Table 5.2.2: Result of the model selection at Station A (Wet Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Hurwitz-Lerch Zeta distribution	-803.18	1612.352	1625.851
2	**Eggenberger-Polya distribution**	**-803.06**	**1610.12**	1619.119
3	**Logarithmic distribution**	**-804.23**	1610.453	**1614.904**
4	Truncated Poisson distribution	-860.39	1722.777	1727.277
5	Geometric distribution	-807.52	1617.049	1621.548

From Table 5.2.2, it is shown that the Eggenberger-Polya distribution has the best fit among all from the criteria of AIC while logarithmic distribution has the best fit in term of BIC at Station A under wet spell modelling.

Table 5.3.1: Result of the model selection at Station B (Dry Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	**Hurwitz-Lerch Zeta distribution**	**-1277.08**	**2560.161**	**2573.381**
2	Eggenberger-Polya distribution	-1307.16	2618.32	2627.134

3	Logarithmic distribution	-1293.68	2589.37	2593.777
4	Truncated Poisson distribution	-2990.33	5982.663	5987.069
5	Geometric distribution	-1390.65	2783.306	2787.713

From Table 5.3.1, it is shown that the HLZ distribution has the best fit among all from the criteria of both AIC and BIC at Station B under dry spell modelling.

Table 5.3.2 Result of the model selection at Station B (Wet Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Hurwitz-Lerch Zeta distribution	-774.86	1555.725	1568.94
2	Eggenberger-Polya distribution	-776.27	1556.538	1565.349
3	**Logarithmic distribution**	**-776.48**	**1554.836**	**1559.241**
4	Truncated Poisson distribution	-852.88	1707.767	1712.172
5	Geometric distribution	-781.19	1564.376	1568.782

From Table 5.3.2, it is shown that the logarithmic distribution has the lowest value of AIC and BIC at Station B under wet spell modelling. Therefore it suggests that logarithmic distribution is the best fit at Station B for wet spell.

Table 5.4.1: Result of the model selection at Station C (Dry Spells)

No.	**Distribution**	**Maximum Log-Likelihood**	**AIC**	**BIC**
1	**Hurwitz-Lerch Zeta distribution**	**-1319.41**	**2644.813**	**2658.17**
2	Eggenberger-Polya distribution	-1333.39	2670.776	2679.68
3	Logarithmic distribution	-1326.17	2654.345	2658.797
4	Truncated Poisson distribution	-2456.02	4914.046	4918.498
5	Geometric distribution	-1404.13	2810.26	2814.712

From Table 5.4.1, it is shown that the HLZ distribution has the best fit among all from the criteria of both AIC and BIC at Station C under dry spell modelling.

Table 5.4.2: Result of the model selection at station C (Wet Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Hurwitz-Lerch Zeta distribution	-823.67	1653.35	1666.701
2	**Eggenberger-Polya distribution**	**-823.10**	**1650.207**	**1659.108**
3	Logarithmic distribution	-828.89	1659.776	1664.226
4	Truncated Poisson distribution	-868.53	1739.06	1743.511
5	Geometric distribution	-1403.07	2808.14	2812.59

From Table 5.4.2, it is shown that the Eggenberger-Polya distribution has the best fit among all from the criteria of both AIC and BIC at station C under wet spell modelling.

Table 5.5.1: Result of the model selection at Station D (Dry Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	**Hurwitz-Lerch Zeta distribution**	**-1378.32**	**2762.644**	2776.089
2	Eggenberger-Polya distribution	-1389.40	2782.797	2791.76

3	**Logarithmic distribution**	**-1382.67**	2767.346	**2771.827**
4	Truncated Poisson distribution	-2395.27	4792.548	4797.029
5	Geometric distribution	-1458.38	2918.76	2923.242

From Table 5.5.1, it is shown that the both HLZ and logarithmic distribution shared the best fitting at Station D for dry spell. HLZ distribution has the best fit among all from the criteria of AIC while logarithmic distribution has the best fit in term of BIC at Station D under dry spell modelling.

Table 5.5.2: Result of the model selection at Station D (Wet Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Hurwitz-Lerch Zeta distribution	-990.03	1986.059	1999.509
2	**Eggenberger-Polya distribution**	**-990.52**	**1985.05**	1994.016
3	Logarithmic distribution	-996.24	1994.48	1998.963
4	Truncated Poisson distribution	-1093.55	2189.105	2193.588

5	**Geometric distribution**	**-993.28**	1988.566	**1993.049**

From Table 5.5.2, it is shown that the Eggenberger-Polya distribution has the best fit among all from the criteria of AIC while geometric distribution has the best fit in term of BIC at Station D under wet spell modelling.

Table 5.6.1: Result of the model selection at Station E (Dry Spells)

No.	**Distribution**	**Maximum Log-Likelihood**	**AIC**	**BIC**
1	**Hurwitz-Lerch Zeta distribution**	**-941.53**	**1889.056**	**1901.404**
2	Eggenberger-Polya distribution	-1005.16	2014.328	2022.56
3	Logarithmic distribution	-1004.93	2011.864	2015.98
4	Truncated Poisson distribution	-4486.36	8974.729	8978.845
5	Geometric distribution	-1204.87	2411.74	2415.856

From Table 5.6.1, it is shown that the HLZ distribution has the best fit among all from the criteria of both AIC and BIC at Station E under dry spell modelling.

Table 5.6.2: Result of the model selection at Station E (Wet Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Hurwitz-Lerch Zeta distribution	-717.07	1440.131	1452.472
2	Eggenberger-Polya distribution	-712.37	1428.74	1436.967
3	Logarithmic distribution	-727.67	1457.343	1461.457
4	Truncated Poisson distribution	-754.50	1510.994	1515.108
5	**Geometric distribution**	**-712.53**	**1427.05**	**1431.164**

From Table 5.6.2, it is shown that the Geometric distribution has the best fit among all from the criteria of both AIC and BIC at Station E under wet spell modelling.

Table 5.7.1: Result of the model selection at Station F (Dry Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	**Hurwitz-Lerch Zeta distribution**	**-1125.38**	**2256.75**	**2269.647**
2	Eggenberger-Polya distribution	-1165.99	2335.978	2344.576

3	Logarithmic distribution	-1160.49	2322.975	2327.274
4	Truncated Poisson distribution	-3283.82	6569.636	6573.935
5	Geometric distribution	-1306.43	2614.86	2619.159

From Table 5.7.1, it is shown that the HLZ distribution has the best fit among all from the criteria of both AIC and BIC at Station F under dry spell modelling.

Table 5.7.2: Result of the model selection at Station F (Wet Spells)

No.	Distribution	Maximum Log-Likelihood	AIC	BIC
1	Hurwitz-Lerch Zeta distribution	-802.52	1611.043	1623.934
2	**Eggenberger-Polya distribution**	**-802.93**	**1609.86**	1618.454
3	Logarithmic distribution	-806.89	1615.773	1620.071
4	Truncated Poisson distribution	-882.20	1766.398	1770.695

5	**Geometric distribution**	-805.46	1612.926	**1617.223**

From Table 5.7.2, it is shown that the Eggenberger-Polya distribution has the best fit among all from the criteria of AIC while geometric distribution has the best fit in term of BIC at Station F under wet spell modelling.

As an overall comparison, the best fit model of all stations for dry and wet spell will be summarized in a table below:

Table 5.8: Summary of best fit distributions

Station	**Spell**	**AIC**	**BIC**
A	Dry	HLZ	HLZ
	Wet	EP	L
B	Dry	HLZ	HLZ
	Wet	L	L
C	Dry	HLZ	HLZ
	Wet	EP	EP
D	Dry	HLZ	L
	Wet	EP	G
E	Dry	HLZ	HLZ
	Wet	G	G
F	Dry	HLZ	HLZ
	Wet	EP	G

HLZ = Hurwitz-Lerch Zeta; EP = Eggenberger-Polya;

L = Logarithmic; G = Geometric distribution

5.3.1 Graphs of histogram and probability mass function

Graphs of relative frequency versus rainfall volume are plotted for HLZ distribution and compared with the rainfall data's histogram. The probability mass function of HLZ is given by:

$$P_k = \frac{1}{T(\theta, s, a)} \frac{\theta^k}{(k+a)^{s+1}} \qquad k = 1, 2, \cdots$$

where $T(\theta, s, a) = \sum_{k=1}^{\infty} \frac{\theta^k}{(k+a)^{s+1}}$.

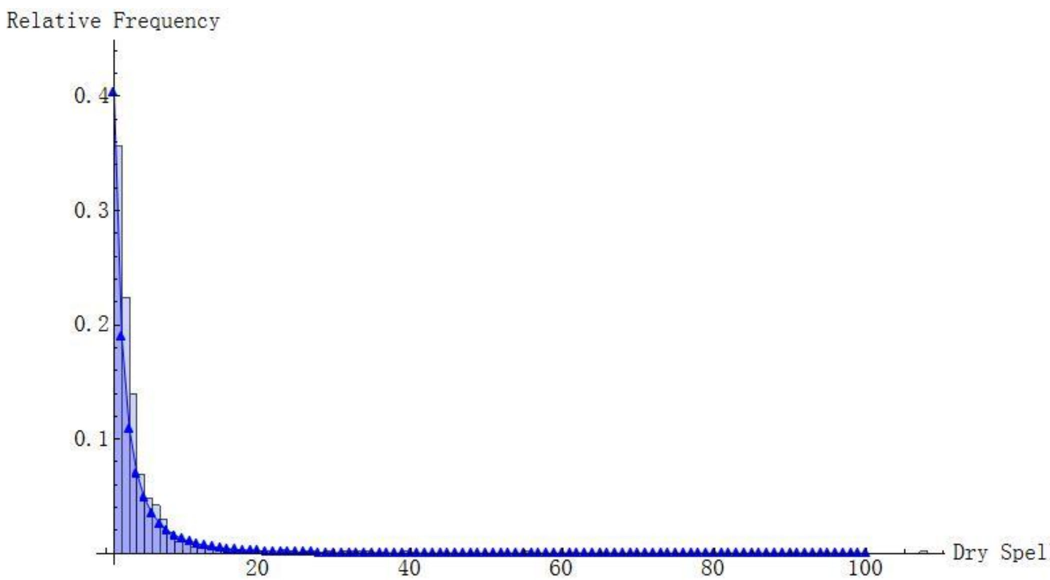

Figure 5.17: PMF of HLZ distribution and dry spell (days) histogram at Station A

Figure 5.17 shows the histogram of the rainfall dry spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station A. The estimated parameters are as follows:

$$\theta = 0.9408$$

s = 0.7088
a = 0.9999

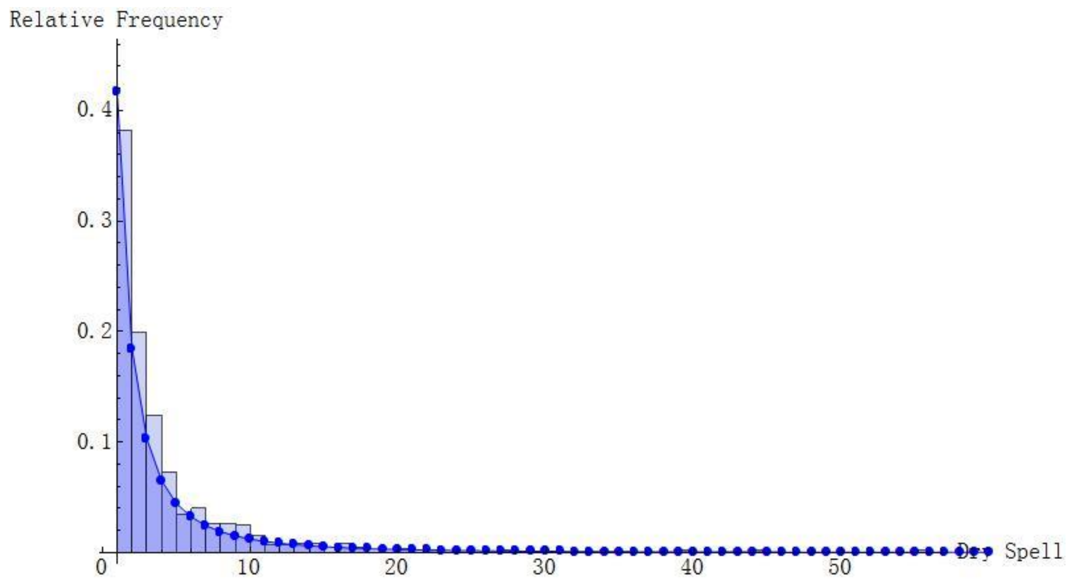

Figure 5.18: PMF of HLZ distribution and dry spell (days) histogram at Station B

Figure 5.18 shows the histogram of the rainfall dry spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station B. The estimated parameters are as follows:

θ = 0.9754
s = 0.9450
a = 0.9999

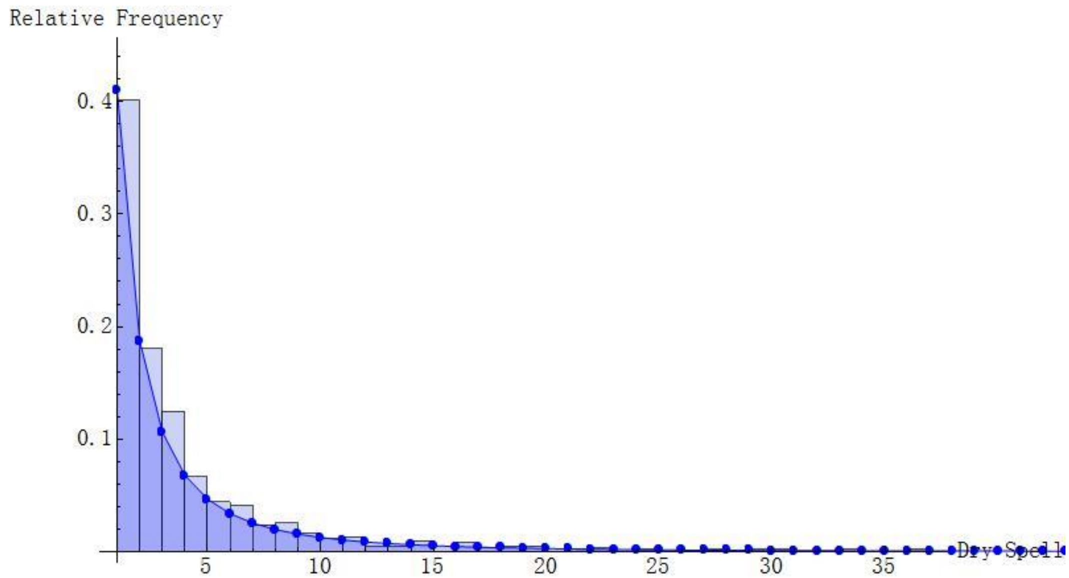

Figure5.19: PMF of HLZ distribution and dry spell (days) histogram at Station C

Figure 5.19 shows the histogram of the rainfall dry spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station C. The estimated parameters are as follows:

$$\theta = 0.9583$$
$$s = 0.8265$$
$$a = 0.9999$$

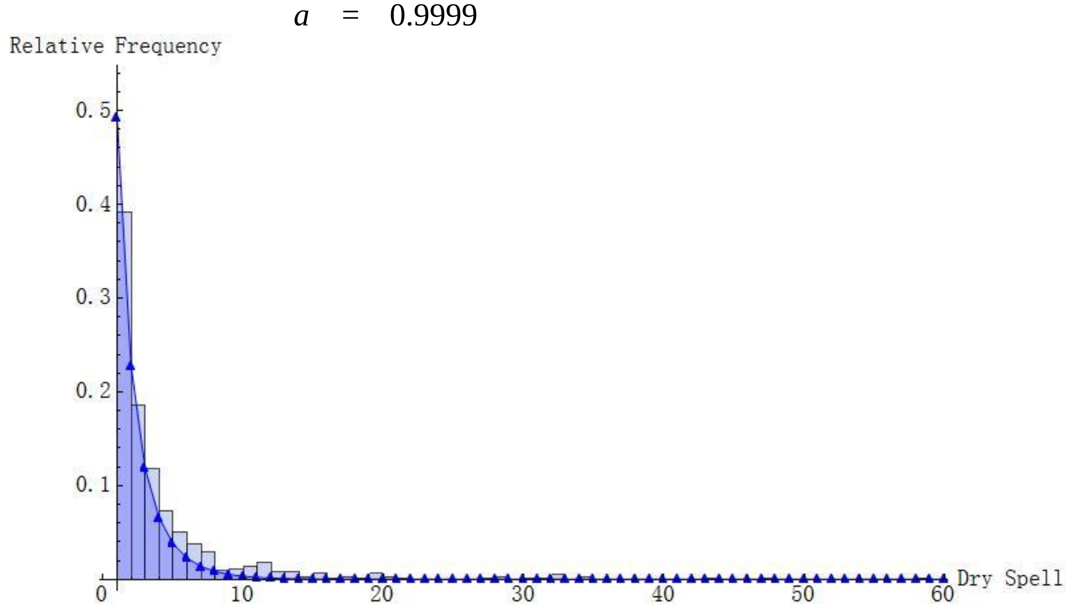

Figure 5.20: PMF of HLZ distribution and dry spell (days) histogram at Station D

Figure 5.20 shows the histogram of the rainfall dry spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station D. The estimated parameters are as follows:

$$\theta = 0.9531$$
$$s = 0.7556$$
$$a = 0.9999$$

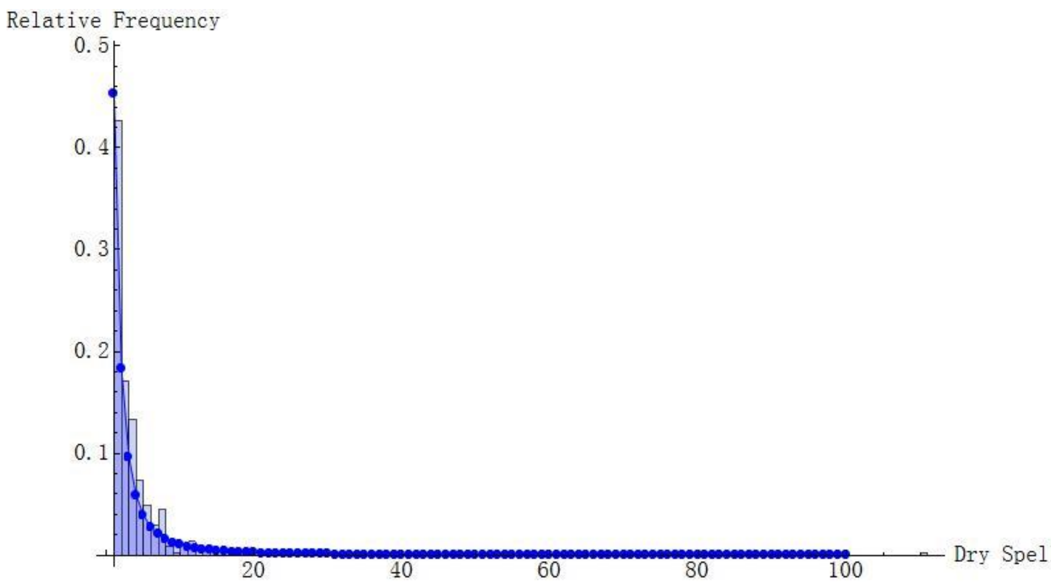

Figure 5.21: PMF of HLZ distribution and dry spell (days) histogram at Station E

Figure 5.21 shows the histogram of the rainfall dry spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station E. The estimated parameters are as follows:

$$\theta = 0.9998$$
$$s = 1.2328$$
$$a = 0.9999$$

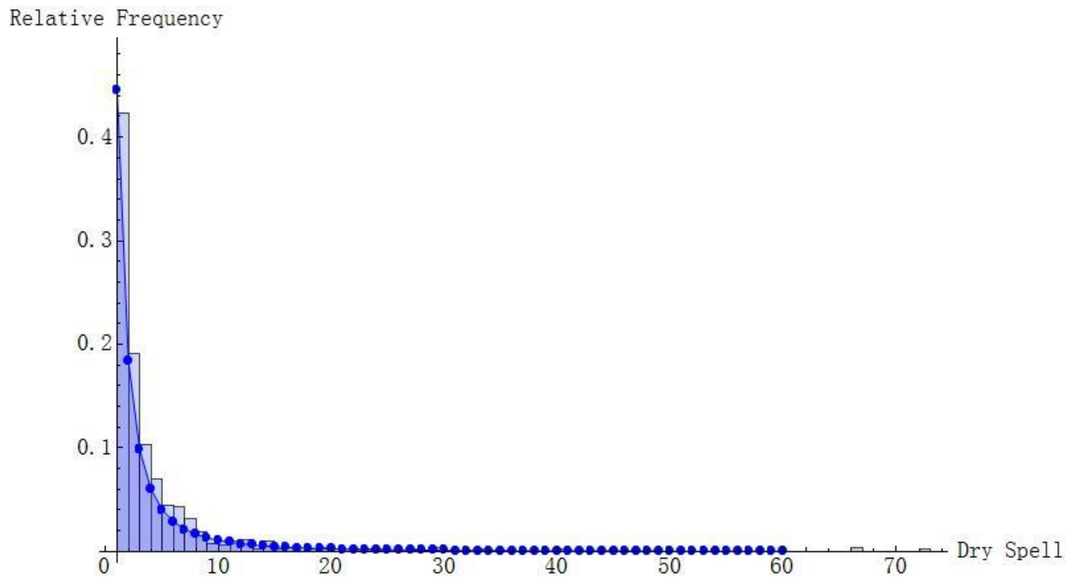

Figure 5.22: PMF of HLZ distribution and dry spell (days) histogram at Station F

Figure 5.22 shows the histogram of the rainfall dry spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station F. The estimated parameters are as follows:

$$\theta = 0.9947$$
$$s = 1.1689$$
$$a = 1.0000$$

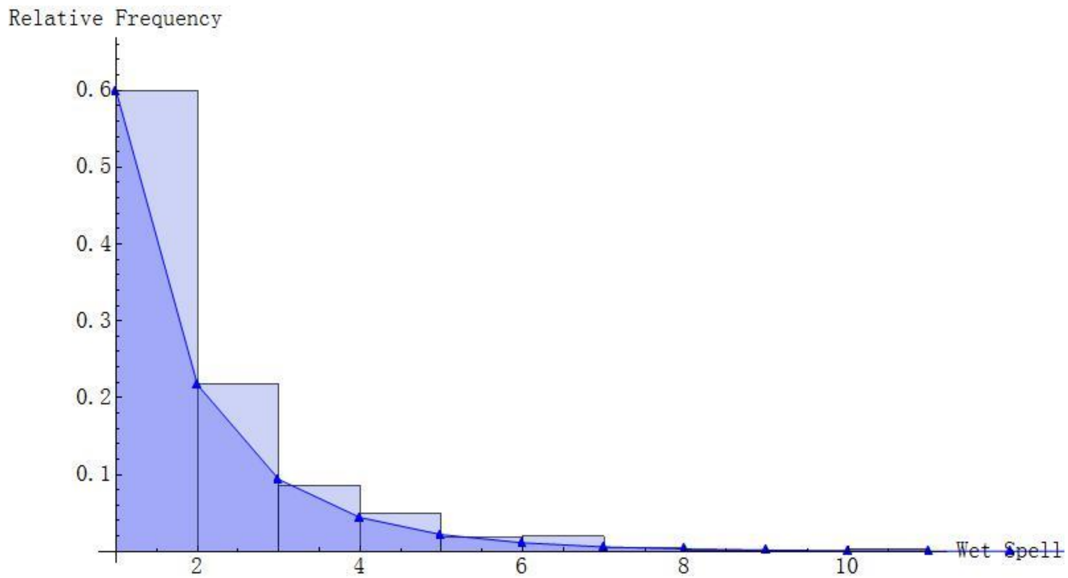

Figure 5.23: PMF of HLZ distribution and wet spell (days) histogram at Station A

Figure 5.23 shows the histogram of the rainfall wet spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station A. The estimated parameters are as follows:

$$\theta = 0.6060$$
$$s = 0.0003$$
$$a = 0.4849$$

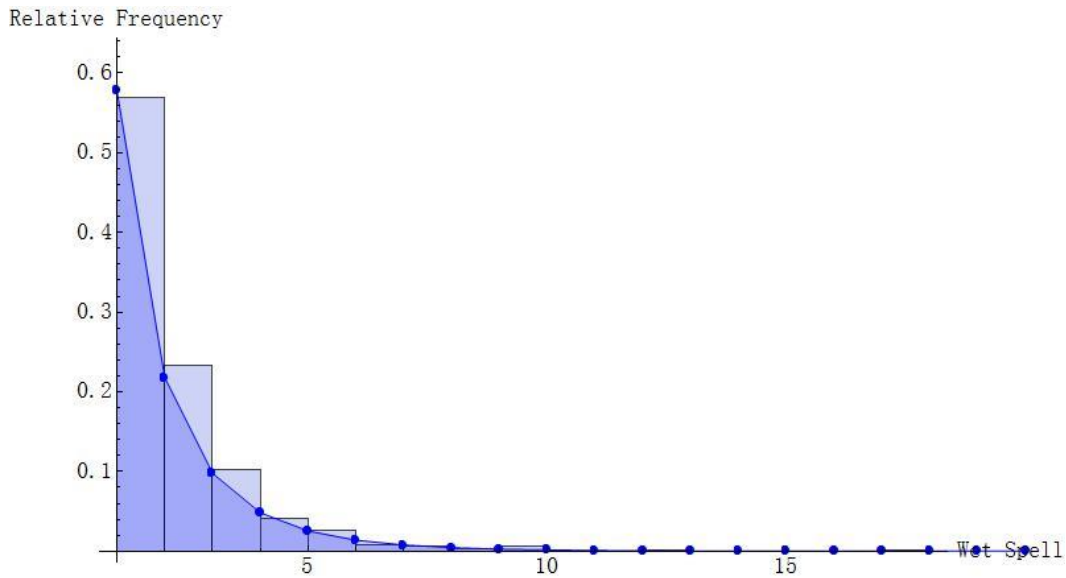

Figure 5.24: PMF of HLZ distribution and wet spell (days) histogram at Station B

Figure 5.24 shows the histogram of the rainfall wet spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station B. The estimated parameters are as follows:

$$\theta = 0.6859$$
$$s = 0.4762$$
$$a = 0.9998$$

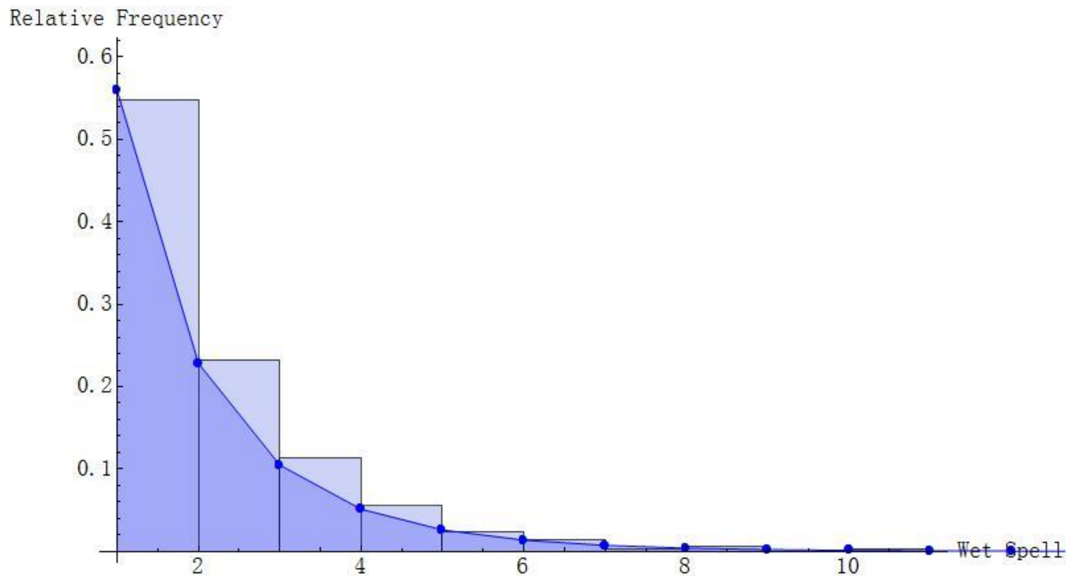

Figure 5.25: PMF of HLZ distribution and wet spell (days) histogram at Station C

Figure 5.25 shows the histogram of the rainfall wet spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station C. The estimated parameters are as follows:

$$\theta = 0.6118$$
$$s = 0.00004$$
$$a = 0.9999$$

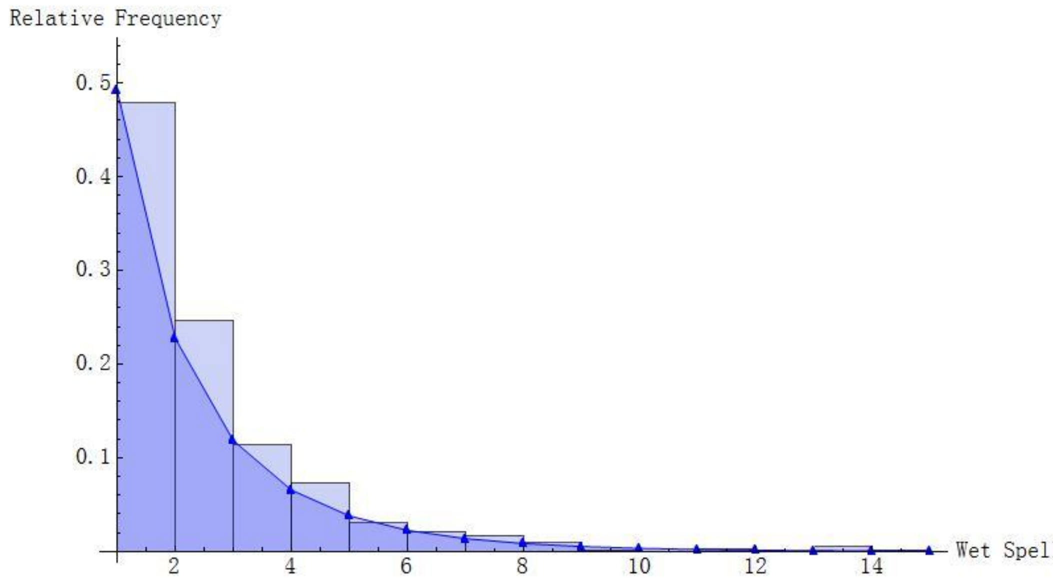

Figure 5.26: PMF of HLZ distribution and wet spell (days) histogram at Station D

Figure 5.26 shows the histogram of the rainfall wet spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station D. The estimated parameters are as follows:

$$\theta = 0.6930$$
$$s = 0.00002$$
$$a = 0.9999$$

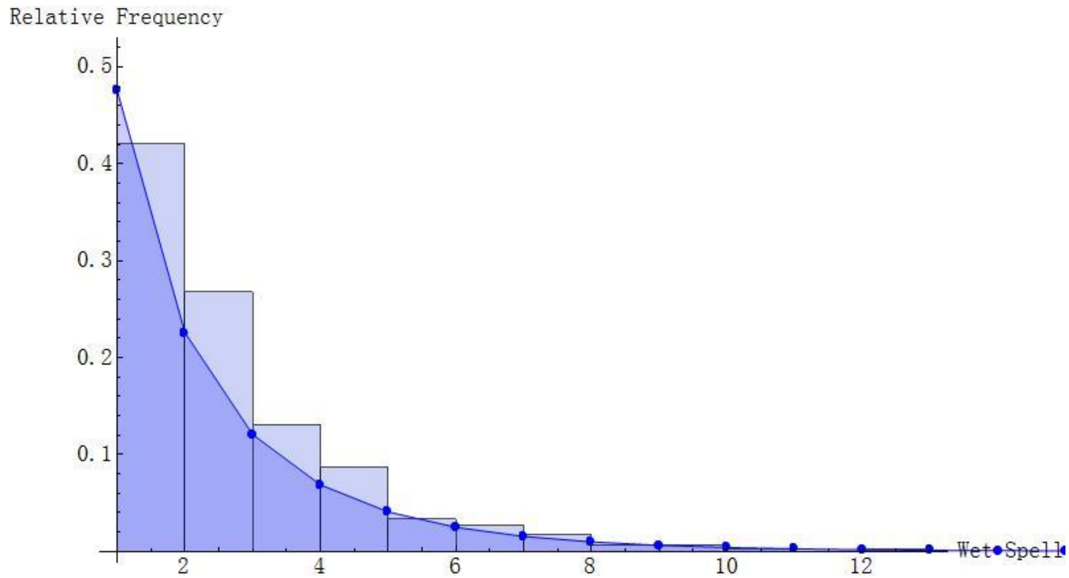

Figure 5.27: PMF of HLZ distribution and wet spell (days) histogram at Station E

Figure 5.27 shows the histogram of the rainfall wet spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station E. The estimated parameters are as follows:

$$\theta = 0.7119$$
$$s = 0.00001$$
$$a = 0.9999$$

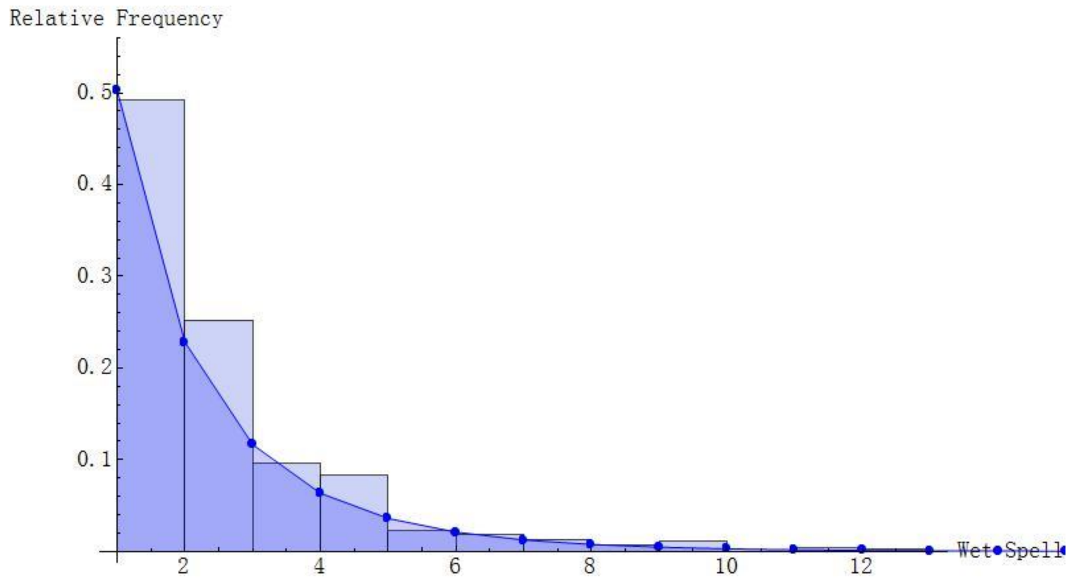

Figure 5.28: PMF of HLZ distribution and wet spell (days) histogram at Station F

Figure 5.28 shows the histogram of the rainfall wet spell and the HLZ distribution's probability density function using the estimated parameters obtained by using the MLE approach for Station F. The estimated parameters are as follows:

$$\theta = 0.6806$$
$$s = 0.00006$$
$$a = 0.9998$$

5.3.2 Discussion on rainfall occurrence modelling

For dry spell, Hurwitz-Lerch Zeta distribution is a better fit among all other distributions under the criterion of AIC. Under the BIC criterion, HLZ distribution has the best fit among all stations except at Station D (Table 5.5.1) as logarithmic distribution has the lowest index for BIC at Station D.

In the case of wet spell, Eggenberger-Polya distribution has the best fit under the AIC criterion in wet spell for 4 stations, which are Station A (Table 5.2.2), Station C (Table 5.4.2), Station D (Table 5.5.2) and Station F (Table 5.7.2). Logarithmic and geometric distributions both have the best fit in wet spell on Station B (Table 5.3.2) and Station E (Table 5.6.2) respectively under the criterion of AIC. However, using BIC, it suggests that logarithmic distribution is the best fit model at Station A and Station B, while geometric distribution has the best fit at Station D, E and F, with Eggenberger-Polya distribution has the best fit at Station C for wet spell.

In general, HLZ distribution has the best fit for the dry spell among other distributions, while for wet spell, Eggenberger-Polya distribution generally fits slightly better than all other distributions as it has 4 best fit across all stations in terms of AIC or BIC. Although Eggenberger-Polya distribution fits better than HLZ distribution in the wet spell, the indexes of AIC and BIC of HLZ distribution are not too much different from the Eggenberger-Polya distribution and other distributions. The graphs of HLZ's PMF on the histogram of wet spell also suggest that HLZ fits well in the wet spell data graphically.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Rainfall modelling on the rainfall volume and rainfall occurrence is essential in assisting us in understanding the characteristic of the precipitation, especially in a country like Malaysia that is located at the tropical region. By understanding the precipitation, many precaution procedures can be done before devastating disasters strike. When it happens, the people are more ready in overcoming it. For example, a water reservoir can be built to collect the rain water when raining season is happening, and the water in the reservoir can be supplied to the people or plantation during the dry season. Besides, safety precaution like expanding the drainage system can be done appropriately especially at the city area to cope with the large amount of rainfall during the wet season. Therefore reducing the damages caused by the flood and drought. Hence, a study on the rainfall is important to evaluate the worthiness of building a water reservoir or any other precaution solution. Besides, the insurance company is able to quantify the risk that associate with the rainfall, e.g. flash flood in the city area. Therefore many damages that may cause by the flash flood can be insured.

In rainfall modelling, it is important to discover what types of distribution describe the rainfall pattern well. We studied the rainfall data around the Langat river basin. The types of distribution used for rainfall volume are the continuous distributions. The past thirty years have seen increasingly rapid advances in the field of applying generalized or mixed distribution in modelling the rainfall data. Therefore we proposed a generalized beta distribution and modified beta distribution for beta-type distributions, and mixture of two modified lognormal distributions.

From our research, we found that in general the mixture of two lognormal distributions has the best fit among all the distributions. Regrettably, it was unable to compute the maximum likelihood of the mixture of two modified lognormal distributions due to its complexity in parameters. However, since the mixture of 2 modified lognormal fits well in the rainfall data, thus it is presumed that the mixture of 2 modified lognormal will fit well too as the mixture of 2 modified lognormal distributions is a general distribution to mixture of 2 lognormal distributions. It can be seen via the performance of the proposed beta generalized distribution fits better than its sub-distributions. Therefore, the proposed generalized beta distribution has the best fit to the rainfall data among the beta-type distributions. As a conclusion for rainfall volume modelling, the mixture of two lognormal is the best model to describe the rainfall data at station D.

Meanwhile, for the study of rainfall occurrence modelling, HLZ distribution fits the best among other distributions for all stations, for dry spell. While Eggenberger-Polya distribution describes the rainfall wet spell slightly better than the other distributions discussed.

Future work includes studying the general properties of the proposed generalized beta distribution, modified beta distribution with graphs and mixture of two modified lognormal distributions including the mean, variance and the general moments. Thus help to understand the new distributions better. Besides, more work can be done on solving the computational problem on computing the maximum likelihood estimator of the mixture of two modified lognormal distributions due to its complexity of 7 parameters. Other potential work includes studying the performances of the proposed distributions on the most recent rainfall data. This can be done by dividing the study period into different smaller intervals to examine the fitting of the distributions. Last but not least, forecasting future rainfall pattern using the proposed distributions can be considered. Multivariate analyses such as regression models and Bayesian statistics can be applied to our proposed distributions.

LIST OF REFERENCES

Adger, W.N., Huq, S. & Brown, K., 2003. Adaptation to climate change in the developing world. *Progress in Development Studies* 3, 3, pp.179–195.

Adiku, S.G.K., Dayananda, P.W.A, Rose, C.W. & Dowuona, G.N.N., 1997. An analysis of the within-season rainfall characteristics and simulation of the daily rainfall in two savanna zones in Ghana. *Agricultural and Forest Meteorology*, 86(1-2), pp.51–62.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, 19(6), pp.716–723.

Aksoy, H., 2000. Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, 24(6), pp.419–428.

Armero, C. & Bayarri, M.J., 1994. Prior assessments for prediction in queues. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(1), pp.139–153.

Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12:171-178.

Berger, A. & Goossens, C., 1983. Persistence of wet and dry spells at uccle (Belgium). *Journal of Climatology*, 3(1), pp.21–34.

Buishand, T.A., 1978. Some remarks on the use of daily rainfall models. *Journal of Hydrology*, 36(3-4), pp.295–308.

Cho, H.-K., Bowman, K.P. & North, G.R., 2004. A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43(11), pp.1586–1597.

Corana, A., Marchesi M., Martini, C. & Ridella, S., 1989. Corrigenda: "Minimizing multimodal functions of continuous variables with the 'Simulated Annealing' Algorithm." *ACM Transactions on Mathematical Software*, 15(3), p.287.

Dubrovský, M., 1997. Creating daily weather series with use of the weather generator. *Environmetrics*, 8(5), pp.409–424.

Fadhilah, Y., Zalina, Md., Nguyen, V-T-V., Suhaila, S. & Zulkifli, Y., 2007. Fitting the best–fit distribution for the hourly rainfall amount in the Wilayah Persekutuan. *Jurnal Teknologi*, 46(C), pp.49–58.

Frich, P., Alexander, L. V., Della-Marta, P., Gleason, B., Haylock, M. Tank Klein, A.M.G. & Peterson, T., 2002. Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, 19(3), pp.193–212.

Gabriel, K.R. & Neumann, J., 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375), pp.90–95.

Goffe, W.L., 1996. SIMANN: A global optimization algorithm using simulated annealing. *Studies in Nonlinear Dynamics & Econometrics*, 1, pp.169–176.

Gradshteyn, I.S. & Ryzhik, I.M., 2007. 6–7 - Definite integrals of special functions. in *table of integrals, Series, and Products (Seventh Edition)*. Boston: Academic Press, pp. 631–857.

Gumbel, E.J., 1954. *Statistics of Extremes*, Courier Corporation.

Hatzianastassiou, N., Katsoulis, B., Pnevmatikos, J. & Antakis, V., 2008. Spatial and temporal variation of precipitation in Greece and surrounding regions based on global precipitation climatology project data. *Journal of Climate*, 21, pp.1349–1370.

Jamaludin, S. & Jemain, A.A., 2008. Fitting the statistical distribution to the daily rainfall amount in Peninsular Malaysia based on AIC criterion. *Journal of Applied Sciences Research*, 4(12), pp.1846–1857.

Jayawardene, H., Sonnadara, D. & Jayewardene, D., 2008. Trends of rainfall in Sri Lanka over the last century. *Sri Lankan Journal of Physics*, 6(0), pp.7–17.

Johnson, N.L., Kotz, S. & Balakrishnan, N., 1996. *Continuous univariate distributions*, *Computational Statistics & Data Analysis*, 2(1), pp.119.

Kedem, B., 1990. Estimation of mean rain rate: Application to satellite observations. *Journal of Geophysical Research*, 95(89), pp.1965–1972.

Kumaraswamy, P., 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46, pp.79–88.

Lana, X. & Burgueño, A., 1998. Probabilities of repeated long dry episodes based on the poisson distribution. an example for Catalonia (NE Spain). *Theoretical and Applied Climatology*, 60(1-4), pp.111–120.

Markovic, R.D., 1965. Probability functions of best fit to distributions of annual precipitation and runoff. *Hydrology Papers 8*, (8), p.39.

McDonald, J.B., 1984. Some generalized functions for the size distribution of income. *Econometrica*, 52, pp.647–663.

McDonald, J.B. & Xu, Y.J., 1995. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1-2), pp.133–152.

Meneghini, R. & Jones, J.A., 1993. An approach to estimate the areal rain-rate distribution from spaceborne radar by the use of multiple thresholds. *Journal of Applied Meteorology*, 32(2), pp.386–398.

N.I. Obot, M.A.C Chendo, S.O.U. and I.O.E., 2010. Evaluation of rainfall trends in Nigeria for 30 years (1978-2007). *International Journal of the Physical Sciences*, 5(14), pp.2217–2222.

Nobilis, F., 1986. Dry spells in the Alpine country Austria. *Journal of Hydrology*, 88(3-4), pp.235–251.

Novotny, E. V. & Stefan, H.G., 2007. Stream flow in Minnesota: Indicator of climate change. *Journal of Hydrology*, 334(3-4), pp.319–333.

Onyenechere, E.C., 2010. Climate change and spatial planning concerns in Nigeria: Remedial measures for more effective response. *Journal of Human Ecology*, 32(3), pp.137–148.

Partal, T. & Kahya, E., 2006. Trend analysis in Turkish precipitation data. *Hydrological Processes*, 20(9), pp.2011–2026.

Racsko, P., Szeidl, L. & Semenov, M., 1991. A serial approach to local stochastic weather models. *Ecological Modelling*, 57(1-2), pp.27–41.

Ratnayake, U. & Herath, S., 2005. Changing rainfall and its impact on landslides in Sri Lanka. *Journal of Mountain Science*, 2(3), pp.218–224.

Richardson, C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1), pp.182–190.

Schwarz, G., 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), pp.461–464.

Smadi, M.M. & Zghoul, A., 2006. A sudden change in rainfall characteristics in Amman, Jordan during The mid 1950s. *American Journal of Environmental Sciences*, 2(3), pp.84–91.

Stern, R.D., Dennett, M.D. & Dale, I.C., 1982. Analysing daily rainfall measurements to give agronomically useful results. I. Direct Methods. *Experimental Agriculture*, 18(03), pp.223–236.

Suhaila, J. & Jemain, A.A., 2007. Fitting daily rainfall amount in Peninsular Malaysia using several types of exponential distributions. *Journal of Applied Sciences Research*, 3(10), pp.1027–1036.

Suhaila, J., Jemain, A.A., , 2011. Introducing the mixed distribution in fitting rainfall data. *Open Journal of Modern Hydrology*, 01(02), pp.11–22.

Suhaila, J. & Jemain, A.A., 2007. Fitting daily rainfall amount in Malaysia using the normal transform distribution. *Journal of Applied Sciences*, 7(14), pp.1880–1886.

Todorovic, P. & Woolhiser, D.A., 1975. A stochastic model of *n* -day precipitation. *Journal of Applied Meteorology*, 14(1), pp.17–24.

Whitesides, G.M., Panek, E.J. & Stedronsky, E.R., 1972. Stochastic daily precipitation models 2. A comparison of distributions of amounts. *Journal of the American Chemical Society*, 18(5), pp.232–239.

Williams, C.B., 1952. Sequences of wet and of dry days considered in relation to the logarithmic series. *Quarterly Journal of the Royal Meteorological Society*, 78(335), pp.91–96.

Wong, C.L., Venneker, R., Uhlenbrook, S., Jamil, A. B. M. & Zhou, Y., 2009. Variability of rainfall in Peninsular Malaysia. *Hydrology and Earth System Sciences Discussions*, 6(4), pp.5471–5503.