

IDENTIFICATION OF A PROBABILITY DISTRIBUTION FOR  
EXTREME RAINFALL SERIES IN EAST MALAYSIA

ISMAIL BIN IERAHIM



KOLEJ UNIVERSITI TEKNOLOGI TUN HUSSEIN ONN

PERPUSTAKAAN KUI TTHO



3 0000 00117623 3



PTTA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH

# KOLEJ UNIVERSITI TEKNOLOGI TUN HUSSEIN ONN

## BORANG PENGESAHAN STATUS TESIS

JUDUL : IDENTIFICATION OF A PROBABILITY DISTRIBUTION FOR EXTREME RAINFALL SERIES IN EAST MALAYSIA

SESI PENGAJIAN : 2004 / 2005

Saya ISMAIL BIN IBRAHIM  
(HURUF BESAR)

mengaku membenarkan tesis (PSM / Sarjana / Doktor-Falsafah)\* ini disimpan di Perpustakaan dengan syarat-syarat kegunaan seperti berikut :

1. Tesis adalah hakmilik Kolej Universiti Teknologi Tun Hussein Onn.
2. Perpustakaan dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. \*\*Sila tandakan ( ✓ )

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

Disahkan oleh

  
(TANDATANGAN PENULIS)

  
(TANDATANGAN PENYELIA)

Alamat Tetap:

No. 11A JALAN 18/35C  
SEKSYEN 18,  
40200 SHAH ALAM  
SELANGOR DARUL EHSAN

Prof Ir Dr Amir Hashim Mohd Kassim  
Nama Penyelia

Tarikh : 02 - NOV - 2004

Tarikh : 03 / 11 / 2004

CATATAN :

- \* Potong yang tidak berkenaan
- \*\* Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT atau TERHAD.
- ♦ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda (PSM).

“~~Saya/Kami~~\* akui bahawa saya telah membaca karya ini dan pada pandangan  
saya/~~kami~~\* karya ini adalah memadai dari segi skop dan kualiti untuk tujuan  
penganugerahan Ijazah Sarjana Kejuruteraan Awam”



Tandatangan

.....

Nama Penyelia

: **Prof Ir. Dr Amir Hashim Mohd Kassim**

Tarikh

: **03 November 2004**

PTT AUTHM  
PERPUSTAKAAN TUN AMINAH

# **IDENTIFICATION OF A PROBABILITY DISTRIBUTION FOR EXTREME RAINFALL SERIES IN EAST MALAYSIA**

**ISMAIL BIN IBRAHIM**

**Laporan Projek ini dikemukakan  
sebagai memenuhi sebahagian daripada syarat  
penganugerahan Ijazah Sarjana Kejuruteraan Awam**



**Fakulti Kejuruteraan Awam & Alam Sekitar  
Kolej Universiti Teknologi Tun Hussein Onn**

**NOVEMBER, 2004**

**“Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya”.**



**PTTA UTHM**  
PERTUNJUKAN TINDAKAN AMINAH

**Tandatangan :** .....

**Nama Penulis**

**: ISMAIL BIN IBRAHIM**

**Tarikh**

**: 02 NOVEMBER 2004**

**For my beloved mother and my late father who never had a chance  
to see his son's success**

**and**

**For my dearest wife, Ayni, and two children, Mohd Izzuan and Nur Izni**

**May ALLAH protect and bless us all.**



PT TA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH

## **ACKNOWLEDGEMENTS**

I would like to send my regards and thank you to all parties that that had given me a helping in completing this report, either directly or indirectly. I would also like to express my deepest appreciation to my supervisor, Prof. Ir Dr. Amir Hashim bin Mohd Kassim, Dean, Faculty of Civil Engineering and Environment, KUiTTHO, and Dr Zalina bt Mohd Daud of Akademik Tentera Malaysia (ATMA), Sungai Besi, Kuala Lumpur, for their guidance and assistance during the course of this project. Not forgetting, I would also like to express my sincerest gratitude to my co-supervisor in Germany, Prof-Ing. Dr Wolfgang Geiger and his assistant Dipl.-Ing. Thorsten Mietzel, who had contributed in their own way in the preparations of this report.

A special appreciation is extended to the Department of Malaysia Meteorological Services in Petaling Jaya for being kind enough to supply me with all the data for this study. I would like also to express my thanks to all colleagues and friends who had given me invaluable assistances and encouragements throughout my study period. And a special thanks to KUiTTHO for granting me the study leave and financial support throughout the course of my study.

Lastly, I would like to express my deepest thanks and love to my family, especially to my wife Ayni Embong, my two children Mohd Izzuan and Nur Izni, and also to my mother Nafsiah bt Abdullah for their patience, moral supports, understanding and encouragements to make this project a success.

And above all, I thank Allah S.W.T for giving the strength and courage to undertake and complete this project.

## ABSTRACT

The goal of this study was to evaluate the goodness-of-fit of the alternate probability distributions to sequences of the annual maximum stream flows in the East Malaysian states of Sabah and Sarawak. We will never know with certainty, the actual amount of rainfall that will occur in the future. So a statistical analysis of this nature can provide guidance on which probability distributions can give reasonable approximation. Basically, this study is a statistical analysis on extreme annual rainfall series in East Malaysia. It will discuss the comparative assessment of eight candidate distributions in providing accurate and reliable maximum rainfall estimates for East Malaysia. The models considered were the Exponential (EXP), Gamma (GAM), Generalized Extreme Value (GEV), Generalized Logistic (GLO), Generalized Pareto (GPA), Gumbel (GUM), Pearson Type III (PE3) and Wakeby (WAK). Annual maximum rainfall series for one-hour resolution from a network of ten Principal Gauging Stations located five each in Sabah and Sarawak were selected for this study. On top of that, data for the fifteen-minutes were also taken for analysis to act as a check to the result. The length of rainfall records varies from seventeen to twenty-one years. Model parameters were estimated using the L-moment method. The quantitative assessment of the descriptive ability of each model was based on using the Probability Plot Correlation Coefficient (PPCC) test combined with Relative Root Mean Squared Error (RRMSE), Root Mean Squared Error (RMSE) and Maximum Absolute Error (MAE). Ranking of PPCC in descending order and the other three criteria on ascending orders were taken and the top three distributions from the ranking for each station were chosen. The GEV distribution came out on top that occurs frequently on most of the stations is selected as the best fitting distribution to describe the extreme rainfall series for East Malaysia.

## ABSTRAK

Tujuan utama kajian ini adalah untuk menilai ujian cocokan kuantitatif bagi setiap taburan kebarangkalian yang terjadi bagi taburan hujan maksimum di Sabah dan Sarawak. Kita tidak akan mengetahui dengan tepat berapa amaun hujan yang akan turun di masa akan datang jadi kajian statistik seperti ini perlu dijalankan untuk memberi sedikit sebanyak panduan tentang taburan kebarangkalian yang mana sesuai digunakan. Ini adalah merupakan kajian statistik untuk taburan hujan maksima di Malaysia Timur dan akan membincangkan mengenai penilaian ke atas lapan calon taburan frekuensi didalam memberikan anggaran yang tepat. Calon-calon untuk model taburan frekuensi tersebut adalah terdiri dari *Exponential (EXP)*, *Gamma (GAM)*, *Generalized Extreme Value (GEV)*, *Generalized Logistic (GLO)*, *Generalized Pareto (GPA)*, *Gumbel (GUM)*, *Pearson Type III (PE3)* and *Wakeby (WAK)*. Siri taburan maksima hujan tahunan untuk sela satu jam dari sepuluh tolok rakaman hujan automatik yang terletak lima di Sabah dan lima di Sarawak digunakan untuk kajian ini. Disamping itu data untuk sela lima belas minit juga digunakan sebagai semakan. Rekod untuk taburan hujan adalah selama antara tujuh belas hingga dua puluh satu tahun. Anggaran parameter model adalah berasaskan Kaedah momen-L manakala ujian cocokan kuantitatif untuk menilai keupayaan diskriptif setiap model adalah berasaskan ujian Koefisien Korelasi Plot Kebarangkalian (KKPK), dan tiga kriteria kejituan yang lain iaitu ralat Relatif Punca Min Kuasa Dua (RRPMKD), ralat Punca Min Kuasa Dua (RPMKD) dan Sisihan Mutlak Maksimum (SMM). KKPK diatur dalam susunan menurun manakala ketiga kriteria yang lain diatur dalam aturan meninggi, dan tiga taburan frekuensi yang teratas bagi setiap stesyen akan diambilkira sebagai calon terbaik. Dari analisa yang dijalankan bagi kajian ini, didapati bahawa taburan GEV adalah lebih sesuai dipilih sebagai taburan frekuensi untuk siri hujan ekstrim bagi Malaysia Timur.

## LIST OF CONTENTS

CHAPTER	ITEM	PAGE
	Title	i
	Declaration	ii
	Dedication	iii
	Acknowledgements	iv
	Abstracts	v
	Table of Contents	vii
	List of Tables	x
	List of Figures	xi
	List of Abbreviations and Symbols	xii
	List of Probability Distributions	xv
<b>CHAPTER I -</b>	<b>INTRODUCTION</b>	<b>1</b>
	1.1 Background	1
	1.2 Statement of Problem	3
	1.3 Study Objectives	4
	1.4 Scope of Study	4
	1.5 Importance And Contribution of the Study	7
	1.6 Layout of Report	7
<b>CHAPTER II -</b>	<b>LITERATURE REVIEW</b>	<b>8</b>
	2.1 Introduction	8
	2.2 Data Series	9

2.3	Probability Distributions	11
2.4	Probability and Plotting Positions	13
2.5	Probability Plots	15
2.6	Chi-Square ( $\chi^2$ ) Test	16
2.7	Anderson-Darling Test	18
2.8	Kolmogorov-Smirnov Test	19
2.9	Probability Plot Correlation Coefficient (PPCC)	21
2.10	Theory of L-Moments	23
2.10.1	L-moment Ratio Diagram	24
2.11	Outliers	25
2.12	Parameter Estimation	26

### CHAPTER III -METHODOLOGY

3.1	Introduction	30
3.2	Selection of Probability Distributions	30
3.2.1	Statistical Probability Distributions	31
3.2.1.1	Exponential Distributions	32
3.2.1.2	Gumbel and Generalized Extreme Value Distributions	33
3.2.1.3	Gamma, Pearson Type III Distributions	34
3.2.1.3	Generalized Pareto Distributions	35
3.2.1.4	Generalized Logistic Distributions	36
3.2.1.5	Wakeby Distribution	37
3.2.2	Parameter Estimation	38
3.2.2.1	Method of L-moment	39
3.2.3	Goodness of Fit Tests	43
3.2.3.1	PPCC, RRMSE, RMSE & MAE	44
3.3	Computer Packaging	46



<b>CHAPTER IV</b>	<b>CASE STUDY</b>	<b>47</b>
4.1	Introduction	47
4.2	Selecting Data for Analysis	50
	4.2.1 Quantitative Tests	51
4.3	Summary	54
<b>CHAPTER V</b>	<b>RESULTS AND DISCUSSIONS</b>	<b>55</b>
5.1	Introduction	55
5.2	Results	56
5.3	Discussions	57
<b>CHAPTER VI</b>	<b>CONCLUSSIONS AND SUGGESTIONS</b>	<b>58</b>
6.1	Conclusions	58
6.2	Suggestions	59
	<b>REFERENCES</b>	<b>61</b>
	<b>APPENDICES</b>	<b>65</b>
	Appendices A-1 to A-11	65
	Appendices B-1 to B-10	76
	Appendix C	94



PTTA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH

**LIST OF TABLES**

<b>NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
1.1	Positions and Elevations of Each Stations	5
2.1	Commonly Used plotting Position Formulas	14
2.2	Theoretical and Empirical Formulas for Several Product Moments	27
4.1	Best Three Distributions Selected for 60-minute interval	52
4.2	Best Three Distributions Selected for 15-minute interval	53



**PTTA UTHM**  
PERPUSTAKAAN TUNKU TUN AMINAH

**LIST OF FIGURES**

<b>NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
1.1	Locations of Stations	6
2.1	Graph of Empirical Distribution Function Vs Normal Cumulative Distribution Function	20
4.1	Map of Malaysia	48



**PTTA UTHM**  
PERPUSTAKAAN TUNKU TUN AMINAH

## ABBREVIATIONS AND SYMBOLS

$A^2$	Anderson-Darling test statistic
AMS	Annual Maximum Series
$C_s$	Coefficient of Skewness
cdf	Cumulative distribution function
D	Distribution / Kolmogorov-Smirnov test statistic
DWD	Deutscher Wetterdienst (German Weather Bureau)
E	Estimation method
$E_i$	Expected frequency for bin $i$ (Chi-squared test)
EDF	Empirical distribution function
ENT	maximum Entropy method
$F$	Cumulative probability of non-exceedance
$F(x)$	Cumulative distribution function ( <i>cdf</i> )
$f(x)$	Probability distribution function ( <i>pdf</i> )
$G$	Skewness for theoretical sample
$i$	Rank of observation in ascending order
IEA	Institution of Engineers, Australia.
$k$	Kurtosis for theoretical sample
L-CV	L-coefficient of variation
LS	Least Square method
$m$	Rank of observation in descending order
MAE	Maximum Absolute Error
MLX	Mixed Moments method

<b>MLE</b>	<b>Maximum Likelihood Estimator</b>
<b>MMS</b>	<b>Department of Malaysia Meteorological Services</b>
<b>MOM</b>	<b>Method of Moments</b>
<b>n</b>	<b>Sample size</b>
<b>NERC</b>	<b>National Environmental Research Council</b>
$\nu$	<b>(Reserved frequency for bin <math>\nu</math> (Chi-squared test)</b>
$P$	<b>Probability of exceedance events</b>
<b>PDN</b>	<b>Probability Distributions</b>
<b>PDS</b>	<b>Partial duration series</b>
<b>PFT</b>	<b>Peak over threshold series</b>
<b>PPCC</b>	<b>Probability plot correlation coefficient</b>
<b>PWM</b>	<b>Probability weighted moments</b>
$Q$	<b>Flood magnitude</b>
$Q_T$	<b>Estimated flood magnitude in T-years</b>
$Q_r$	<b>(Reserved magnitude of <math>Q</math>) with probability <math>p_r</math></b>
$r$	<b>Correlation coefficient, probability plot correlation coefficient value</b>
<b>RMSE</b>	<b>Root mean squared error</b>
<b>RRMSE</b>	<b>Relative root mean squared error</b>
$s$	<b>Sample standard deviation</b>
$s^2$	<b>Variance for theoretical moment</b>
$T$	<b>Return period for AMS</b>
$T_r$	<b>Return period for PDS</b>
$u$	<b>th quantile</b>
$\bar{u}$	<b>Average value of fitted quantile</b>

PITA UTHM

PERPUSTAKAAN TUNKU TUN AMINAH

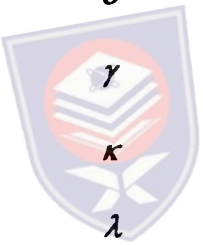


<b>MLE</b>	<b>Maximum Likelihood Estimator</b>
<b>MMS</b>	<b>Department of Malaysia Meteorological Services</b>
<b>MOM</b>	<b>Method of Moments</b>
<b><math>N</math></b>	<b>Sample size</b>
<b>NERC</b>	<b>National Environmental Research Council</b>
<b><math>O_i</math></b>	<b>Observed frequency for bin <math>i</math> (Chi-squared test)</b>
<b><math>P</math></b>	<b>Probability of exceedance events</b>
<b>PDs</b>	<b>Probability Distributions</b>
<b>PDS</b>	<b>Partial duration series</b>
<b>POT</b>	<b>Peak over threshold series</b>
<b>PPCC</b>	<b>Probability plot correlation coefficient</b>
<b>PWM</b>	<b>Probability weighted moments</b>
<b><math>Q, q</math></b>	<b>Flood magnitude</b>
<b><math>Q_T</math></b>	<b>Estimated flood magnitude in T-years</b>
<b><math>\hat{Q}</math></b>	<b>Observed magnitude of <math>Q</math></b>
<b><math>q_i</math></b>	<b><math>i</math>th plotting position</b>
<b><math>r</math></b>	<b>Correlation coefficient, probability plot correlation coefficient value</b>
<b>RMSE</b>	<b>Root mean squared error</b>
<b>RRMSE</b>	<b>Relative root mean squared error</b>
<b><math>s</math></b>	<b>Sample standard deviation</b>
<b><math>s^2</math></b>	<b>Variance for theoretical moment</b>
<b><math>T</math></b>	<b>Return period for AMS</b>
<b><math>T_E</math></b>	<b>Return period for PDS</b>
<b><math>w_i</math></b>	<b><math>i</math>th quantile</b>
<b><math>\bar{w}</math></b>	<b>Average value of fitted quantile</b>



$x_i$	$i$ th ordered observation
$x_p$	$p$ th quantile or 100 $p$ percentile
$\bar{x}$	Average value of the observation / mean for theoretical moment
$x(F)$	Quantile function
$Y_l$	Lower limit for class $i$ (Chi-squared test)
$Y_u$	Upper limit for class $i$ (Chi-squared test)

$\alpha$	Scale parameter
$\beta, \xi$	Location parameter
$\beta_r$	$r$ th probability weighted moment
$\chi^2$	Chi-squared test
$\delta$	Parameter for Wakeby distribution
$\gamma$	Parameter for Wakeby distribution / Skewness for sample moment
$\kappa$	Shape parameter / kurtosis for sample moment
$\lambda$	Mean number of peaks per year
$\lambda_r$	$r$ th L-moment
$\mu$	Mean for sample moment
$\sigma^2$	Variance for sample moment
$\tau_2$	L-coefficient of variation
$\tau_3$	L-skewness
$\tau_4$	L-kurtosis
$\Gamma$	Gamma function
$\Sigma$	Summation



PT TAA UTHM  
PERPUSTAKAAN TUNJUNGIN AMINAH

**PROBABILITY DISTRIBUTIONS**

<b>EV1</b>	<b>Extreme Value Type 1 distribution</b>
<b>EV2</b>	<b>Extreme Value Type 2 distribution</b>
<b>EV3</b>	<b>Extreme Value Type 3 distribution</b>
<b>GAM</b>	<b>Gamma distribution</b>
<b>GEV</b>	<b>Generalized Extreme Value distribution</b>
<b>GUM</b>	<b>Gumbel distribution</b>
<b>GLO</b>	<b>Generalized Logistic distribution</b>
<b>GPA</b>	<b>Generalized Pareto distribution</b>
<b>LP3</b>	<b>Log Pearson Type 3 distribution</b>
<b>PE3</b>	<b>Pearson Type 3 distribution</b>
<b>WAK</b>	<b>Wakeby distribution</b>



**PTTHM**  
PERPUSTAKAAN TUNKU TUN AMINAH

# CHAPTER

# I

PTTAUTHM

PERPUSTAKAAN TUNKU TUN AMINAH



# CHAPTER I

## INTRODUCTION

### 1.1 Background

After five decades, the field of statistical hydrology continues to evolve and remains a very active area of investigation. Researchers continue to examine various distributions, methods of estimation of parameters, and problems related to regionalization. However, much of this material appears in journals and reports and usually in a form which is not easily accessible to practitioners and students and hence producing a bigger gap between research and practice.

The eighties proved to be important years with many significant contributions. Due to its large economical and environmental impact, flood frequency analysis remains a subject of great important and interest, and research on improved methods for obtaining reliable flood estimates has continued into the nineties, although with different emphasis. In the seventies and eighties much effort was spent on developing efficient at-site flood frequency procedures. New distributions and estimation methods were introduced in the hydrologic journals, some of them developed specifically for flood frequency analysis. It seems that this tendency has decelerated somewhat at the beginning of the nineties. Researchers are gradually realizing that the lack of sufficiently long data series imposes an upper limit on the degree of sophistication that can reasonably be justified in at-site flood frequency analysis. It has been emphasized by many that instead of developing new methodologies for flood frequency analysis, effort should be spent on comparing existing ones and on looking for other sources of information (Potter, 1987; Bobée

et al., 1993). Regionalization is probably the most viable avenue for improving flood estimates, and fortunately this seems to be the direction that the researches have taken in the nineties.

Before designing a variety of engineering works in water resources planning and other water related projects, engineers often require flood estimates at a particular proposed site or project location. Flood volume estimate is very important in predicting or estimating the return period of rare events such as extreme rainfalls or precipitation for a site or a group of sites.

The general purpose of frequency analysis is to relate the magnitude of extreme events to their frequency of occurrence through the use of probability distributions (Chow et al., 1988). The data observed over an extended period of time in a hydrologic system are analyzed in frequency analysis and are assumed to be independent and identically distributed. Further, it is assumed that the floods have not been affected by natural or manmade changes in the hydrological regime in the system.

In practice, the actual probability distribution for both at-site and a regional data is quite unknown. For this study of the East Malaysian states of Sabah and Sarawak, the data collected is quite short in relation to other countries and began in the year 1951 for monthly rainfall and only from 1979/1981 for the detailed hourly duration rainfalls. So due to the very limited short data available, it is quite possible that some output from the analysis might not be that accurate. Hence it is quite necessary that a more detailed study should be done for better estimates of design storms based on a more reliable and longer period of data.

With the advancement of computer software, the efficiently quantitative method for goodness-of-fit such as the probability plot correlation coefficient test will provide a more reliable output of the existing techniques. And also with the introduction of the L-moments (Hosking, 1990) numerous researchers have recommended them to assess the goodness-of-fit of various probability distributions to data samples of stream flow and precipitation (Chowdhury et al., 1991, Hosking and Wallis, 1993, Stedinger et al., 1993).

## 1.2 Statement of Problem

The focus of this study is to determine the most appropriate probability distribution of extreme rainfall that are best suited to East Malaysia, which comprises the states of Sabah and Sarawak. For that purpose, the data from the annual maximum series were chosen and more preferred to be used instead of the partial duration series by virtue of its simplicity and easier to extract and analyze.

The process of probability fitting involved several steps. The problems of finding the optimal combination of the best estimation techniques with the most suitable distribution have actively been discussed over the past couple of decades. Hosking and Wallis (1993) organized regional frequency analysis into four stages: (1) Screening of the data; (2) identification of homogeneous region; (3) choice of a regional probability distribution; and (4) estimation of the regional probability distribution. Normally, if the final goal is to estimate a regional flood frequency, the process should proceed from one stage to the next without skipping intermediate stages. The identification of homogeneous region is normally important for identifying a regional shape or skew parameter to be used with the regional estimation procedure. So a comprehensive and global study on this subject matter will be an enormous task to undertake, and let alone the stochastic nature of the empirical rainfall data is a big hurdle to overcome along with the spatial and temporal differences of the region. Because of that, most studies are done locally, using at-site data and also restricted to a few commonly used distributions.

It should be noted also that despite the availability of rainfall data from either the Malaysia Meteorological Services Department or the Department of Drainage and Irrigations Malaysian, a formal study on the probability distribution of annual extreme rainfall series in East Malaysia has never been conducted. Thus there is an urgent need for this research to done immediately so that the existing rainfall atlas can be revised to the latest development.

Regionalization study is also important in determining the spatial and temporal patterns of rainfall in the region which are directly affected by the monsoon seasons. The mechanisms, which bring heavy rains to different parts of

the region, should be incorporated into the homogeneity study, which, at the present moment are very lacking. Unfortunately, this study is not included in the scope of this report.

### 1.3 Study Objectives

The objectives of this study to perform a detailed assessment of various probability distributions to determine the best distribution model for describing the extreme rainfall series for East Malaysia. This is important because there is no formal study on the subject has ever been done and hence the need for it in order to identify the pattern of distribution for these areas. Apart from that, it is to provide a bit on theoretical knowledge of spatial and temporal pattern of extreme and monthly rainfalls in East Malaysia.

### 1.4 Scope of Study

The study will be basically a statistical analysis to determine the best probability distribution that can be applied to East Malaysia based on annual extreme rainfall series and not from the partial duration series. All the data was collected from the Malaysia Meteorological Services Department located in Petaling Jaya, Selangor . The data for the annual maximum series collected was for five principal stations in Sabah and another five stations for Sarawak. These stations are located in Labuan, Kota Kinabalu, Kudat, Sandakan and Tawau for Sabah, and Kuching, Sri Aman, Sibul, Bintulu and Miri for Sarawak. The locations and altitudes of each station is as shown in Table 1.1 and Figure 1.1 below.

**Table 1.1 : Position and Elevations of Each Principal Stations**

<u>Station</u>	<u>Latitude</u>	<u>Longitude</u>	<u>Ht above M.S.L (m)</u>
<b>Kuching</b>	1° 29' N	110° 20'E	21.7
<b>Sri Aman</b>	1° 13' N	111° 27'E	9.6
<b>Sibu</b>	2° 15' N	111° 58'E	30.9
<b>Bintulu</b>	3° 12' N	113° 02'E	3.1
<b>Miri</b>	4° 20' N	113° 59'E	17.0
<b>Labuan</b>	5° 18' N	115° 15'E	29.3
<b>Kota Kinabalu</b>	5° 56' N	116° 03'E	2.3
<b>Kudat</b>	6° 55' N	116° 50'E	3.5
<b>Tawau</b>	4° 16' N	117° 53'E	19.8
<b>Sandakan</b>	5° 54' N	118° 04'E	10.3

The study will also be focusing on using the L-moments method for parameter estimations which is highly recommended by researchers. For the goodness-of-fit test, the probability plot correlation coefficient (PPCC) will be the main test considered as well as the other three tests for comparison such as the relative root mean square error (RRMSE), root mean square error (RMSE) and the maximum absolute error (MAE).

For the probability distribution, a set of eight probability distributions commonly used in distribution fitting studies are chosen and they are the two-parameter Exponential (EXP), Gamma (GAM) and Gumbel (GUM) distributions, the three-parameter Generalized Extreme Value (GEV), Generalized Logistic (GLO), Generalized Pareto (GPA) and Pearson Type III (PE3) distributions, and the five-parameter Wakeby (WAK) distribution.

# Malaysian Meteorological Service Station Network

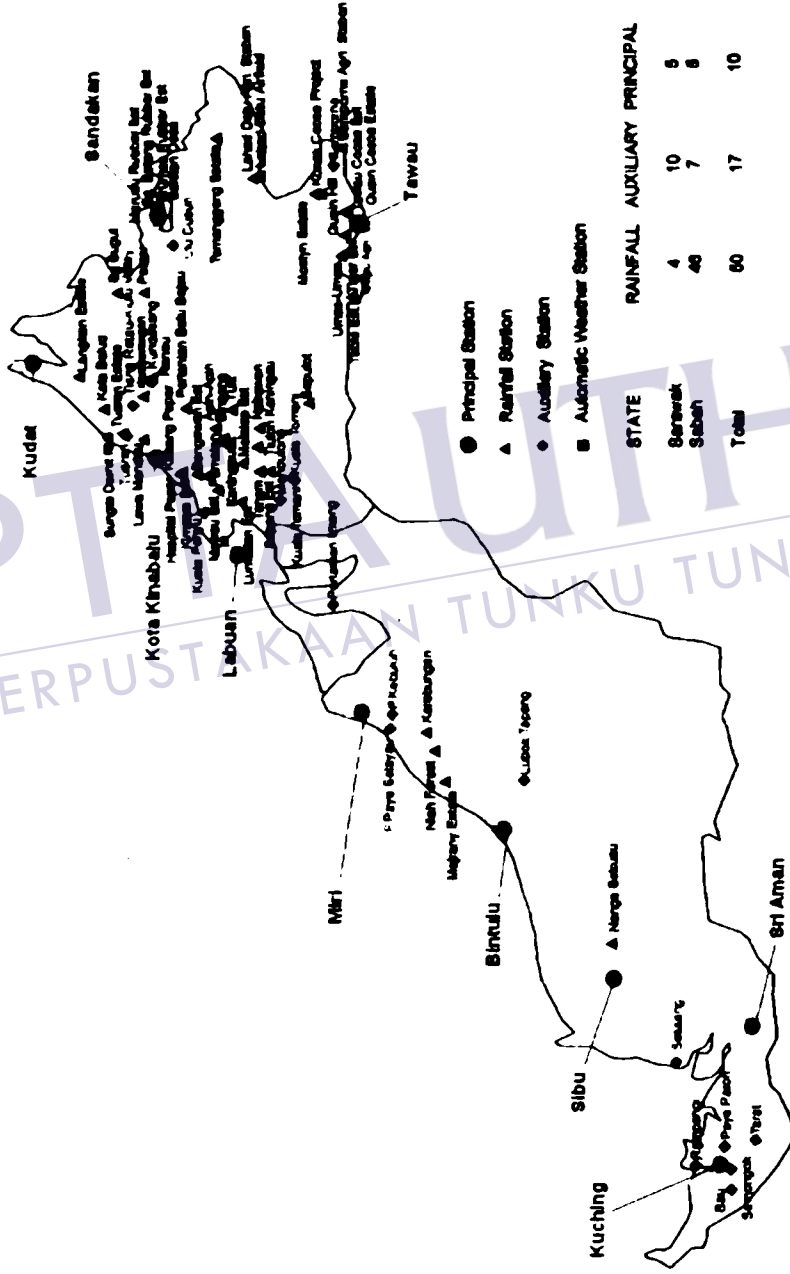


Figure 1.1 : Locations of Stations

## **1.5 Importance and Contribution of the Study**

The study is meant to identify the best fitting probability distribution that best describe the extreme rainfall in East Malaysia and hence can be of great importance for future practitioners to design for water related structures in both the states. It is made more important to various authorities in those two states or federal government because so far there is no formal study on probability distribution of extreme rainfall series has ever been conducted to cover the whole of East Malaysia.

At the moment, Gumbel has been accepted to be used as a probability distribution for the whole of Malaysia. This is done so without any research done on this matter. So with this perception, it is hopefully beneficial to all that the outcome of this study will contribute to a better knowledge of the rainfall behavior in East Malaysia

## **1.6 Layout of Report**

This report is presented in six chapters and three appendices. The objectives and scope of the study are covered in Chapter I, that also including the problem statement and significant of study. Chapter II will basically be dealing with literature review of all the procedures and methods that are in existence. The methodology, explaining all the methods and procedures that are to be used in this study are described in details in Chapter III. The case study and analysis of the probability distribution for this study is discussed in Chapter IV. The results and discussions based on this analysis are presented in Chapter V. The overall conclusion of the study and further suggestions for future research are discussed in Chapter VI. Appendices A will be the summary of all the results from the analysis while Appendices B will tabulate the data collected from MMS. And lastly Appendix C is the sample of Fortran routine for computing L-moment parameters for GEV distribution.

# CHAPTER

## II

PTTAUTHM  
PERPUSTAKAAN TUNKU TUN AMINAH



## **CHAPTER II**

### **LITERATURE REVIEW**

#### **2.1 INTRODUCTION**

Fitting a probability distribution for extreme rainfall series is essential in the design of water-related structures, in agriculture, in weather modification and monitoring of climatic changes. It is also fundamental to the building of design storms and hence become the main issue to be discussed in this study.

Since the early studies by the U.S Water Resources Council (“Guidelines” 1967), which recommended the wide-scale adaptation of the Log Pearson Type III Distribution (LP3) as the base distribution for flood frequency analyses in the USA, numerous literatures pertaining to the subject matter were discussed and presented in engineering journals and publications. However, in most of the literatures, if not all, the involvement of the statistical analysis in the impending investigation cannot be avoided. The selection of the most suitable distribution that could provide accurate estimation of the extreme rainfall has to systematically be performed and analyzed on the various probability models.

Design of engineering structures often requires estimation of flood of large return periods. Some researchers therefore resort to analyzing the larger sample events, due to the fact that observed data often exhibit two or more segment that could not fit into one smooth curve. Hence fitting a single function to the whole data set might lead to errors in choosing distribution, which consequently lead to errors in the estimation of large return period events.

The small observations are less relevant in flood analysis because they represent low return periods but can have a considerable effect on the estimation of the cumulative distribution functions

## 2.2 Data Series

In frequency analysis, a unique relationship between a flood magnitude  $q$  and the corresponding recurrence interval  $T$  is sought. The task is to extract information from a flow record to estimate the relationship between  $q$  and  $T$ . There are mainly two types of data series involved in the frequency modeling. The first and most frequently used is the annual maximum series (AMS), which, only the maximum or the peak flow value observed for each year is extracted to form the empirical data set with a specific and suitable time interval. However, the use of an AMS series may involve some loss of information. For example, the second or third peak within a year may be greater than the maximum flow in other years and yet they are ignored. Assuming independence and stationary, a statistical distribution is fitted to the data using a given estimation method, and a flood with a specific exceedance probability can then be inferred from this distribution. Much of the recent research on at-site flood frequency procedures has focused on the adequate choice of distributions (D) and estimation methods (E), in terminology of Distribution-Estimation (D/E) procedure.

Two components have to be considered when evaluating a particular D/E procedure, mainly the descriptive and predictive ability (Cunnane, 1987). The former relates to the ability of a distribution to reproduce selected statistical characteristics of the data, in particular the density shape typically expressed by skewness and kurtosis measures. The predictive ability, on the other hand, refers to the accuracy, usually measured in terms of bias and mean square error, with which flood quantiles can be estimated. It is also a relative measure, because it necessitates an assumption on the true distribution of floods, which in practice is always unknown. Commonly, it is assumed that the fitted and the parent distributions belong to the same family, and under this assumption, the bias, the root mean square

error, or some other figure of performance can be computed either analytically or by simulation.

With the usage of modern technologies, computer software can easily extract data to give a more accurate scanning over the entire data set to locate the maximum for a specific time period. However, there is also an argument that the annual maximum over a certain year is being exceeded by a non-maximum of another year. This has given rise to another second method of extracting data which produces a peak over threshold series (POT) or also known as partial duration series (PDS).

POT or PDS is a data series comprising of extreme values that exceed a certain identified threshold value. Any values that exceed the certain base value one will be included in the data set. The base is usually selected low enough to include at least one event in each year. So it is not confined to a single observation for any year but may be taken as more than one on any or all of the years. This PDS model, however, is limited by the fact that observations may not be independent (Chow et al., 1988). According Cunnane (1989), the AMS model is statistically more efficient than the PDS model when  $\lambda$  is small ( $\lambda < 1.65$ ) where  $\lambda$  is the mean number of peaks per year included in the PDS. The return period  $T_E$  for a PDS model is related to the return period  $T$  of an AMS model by equation 2.1 (Chow, 1964).

$$T_E = \left[ \log \left( \frac{T}{T-1} \right) \right]^{-1} \quad (2.1)$$

The relative difference between  $T_E$  and  $T$  is greatest for small values of  $T$  and converges to 0.5 as  $T$  increases.

Research of flood frequency analysis based on partial duration series (PDS) has in recent years mainly focused on the use of the generalized Pareto (GPA) distribution for modeling exceedances. Davison and Smith (1990) mention several advantages of this distribution for use with PDS flood data including the observed values really display the maximum events that could occur and a certain year might

have more than once the occurrence of the extreme events and this is something to look into by the researchers.

However, this method is rarely been used due to the lack of general guidelines for its application and remains unpopular for design practice. The problem of deciding on the base value and detainment of the peaks has to be resolved first. The peak has to be taken from different storm events in order to preserve independence and it is still unclear for how long an interval should be inserted in between peaks to determine it. Little has been done to include PDS models in a regional estimation scheme, and this is perhaps the main reason that PDS analysis remains less used in practice than the annual flood method. Future research should focus on developing regional estimation procedures for use with PDS data. AMS does not have this drawback as the selection of one largest value per year generally leads to identical and independently distributed events.

The usage of AMS is very significant for a design rainfall estimate of longer return periods of more than 5 years, but the PDS will give a more accurate and reliable estimate for the shorter return periods of two years or less.

### **2.3 Probability Distributions**

A probability distribution (PD) is a function representing the frequency of occurrence of the value of a random variable. By fitting a distribution to a set of data, a great deal of the probabilistic information in the sample can be compactly summarized in the function and its associated parameters.

There are probably fifteen or more of the probability distributions (PDs) that are generally in used at present moment that falls under several family distributions. The Normal Family of PDs includes the two-parameter Normal, three-parameter Log-Normal and Generalized Normal distributions. Under the Gamma family there are the two-parameter Gamma, three-parameter Gamma or also known as Pearson Type III. The Extreme-value family comprises the Extreme Value Type I (EV1),

Extreme Value Type II (EV2), Extreme Value Type III (EV3) and also the two-parameter Gumbel. Gumbel was the most popular and widely used distribution for describing extreme processes in the 1970s. In 1995, the three forms of the extreme value family, EV1, EV2 and EV3, were combined into a single distribution called the generalized extreme value (GEV) distribution which is a three-parameter distribution and hence it is more flexible. When the shape parameter of the GEV is reduced to zero, then GEV becomes Gumbel. The Weibull also belongs to this family although it is more suitable for describing minimum processes and it is bounded below by zero. Other distributions often used in hydrological study include the various forms the Exponential distribution, the three-parameter Generalized Logistic and the Generalized Pareto, the four-parameter Kappa, and the five-parameter Wakeby distribution

As mentioned earlier, the LP3 was recommended by the U.S Water Resources Council ("Guidelines" 1967), and to be adopted in the United States. However, re-evaluation study done by Wallis and Wood (1995) has shown that Generalized Extreme Value or the Wakeby is more suitable to be adopted in the U.S. In the United Kingdom, the Natural Environmental Research Council (NERC) has decided on the use of Generalized Extreme Value (GEV) distribution. Other countries have their own standard distribution to be used for describing flood flow. The Institute of Engineers in Australia (IEA) has also recommended the LP3. But recent research has proposed a separate division of Australia into winter and non-winter rainfall regions. The research concluded that GEV and GPA performed better than the adopted LP3 at representing flood flow data outside the winter dominated region. In India, the GPA has been concluded by researchers to be the best fitting model for central India. The German Weather Bureau (Deutscher Wetterdienst DWD) applied another way of evaluating extreme rainfall data. Its target is to provide rainfall statistics for any point within the Federal Republic of Germany. Hence, rather simple methods were invoked to allow for regionalization of point rainfall. The Extreme Value Type I (EV1) distribution has been chosen as the only probability distribution to be used at all stations throughout the country (Bartels, 1997).

The study of probability distributions, either fitting hypothetical distribution or empirical data sets, or investigating of estimation methods, goodness of fit procedures or merely building tables of critical values which requires simulation of hypothetical distributions has been and is still being investigated and improved all over the world, meaning that the process is on going and will never be stopped.

## 2.4 Probability and Plotting Positions

Maximum floods do not occur with any fixed pattern of time or magnitude, and the time intervals between floods also vary. The definition of return period is the average of these inter-event times between flood events (Cunnane, 1989) and may not involve any reference to probability. Large floods naturally have large return periods or in another word the tendency of reoccurrence is smaller, and vice-versa. However, a relationship between the probability of occurrence of a flood and its return period can be justified. A given flood  $q$  with a return period  $T$  may be exceeded once in  $T$  years. Hence, the probability of exceedance is  $P(Q_T > q) = 1/T$ . The cumulative probability of non-exceedance,  $F(Q_T)$  is given by Equation 2.2 :

$$F(Q_T) = P(Q_T \leq q) = 1 - P(Q_T > q) = 1 - \frac{1}{T} \quad (2.2)$$

Equation 2.2 is the basis for estimating the magnitude of a flood,  $Q_T$ , given its return period  $T$ . Substituting  $F(Q_T) = 1 - (1/T)$  in a known statistical distribution function, the magnitude of  $Q_T$  can easily be solved. Normally the data are plotted on a probability paper to check whether they follow a particular pattern and to detect if any errors occur. Probability plots require an initial estimate of the probability of non-exceedance  $F = \{F(Q_T)\}$ , which is known as a 'plotting position.' Equation 2.3 is a formula for plotting position developed by Hosking (1990) :

$$F = \frac{i - 0.35}{N}, i = 1, 2, \dots, N \quad (2.3)$$

where,

$N$  = sample size

$i$  = the rank of the observations in ascending order

Equation 2.3 above is believed to give an acceptable results for some commonly used three-parameter distributions. Cunnane (1989), have also given some other formulas that are commonly used for plotting positions as shown in Table 2.1 below :

**Table 2.1 : Commonly Used Plotting Position Formulas**

Plotting Position	Formula	
	$T$	$F = 1 - (1/T)$
Hazen	$\frac{N}{m - 0.5}$	$\frac{i - 0.5}{N}$
California	$\frac{N}{m}$	$\frac{i - 1}{N}$
Weibull	$\frac{N + 1}{m}$	$\frac{i}{N + 1}$
Chegodayev	$\frac{N + 0.4}{m - 0.3}$	$\frac{i - 0.3}{N + 0.4}$
Blom	$\frac{N + 0.25}{m - 0.375}$	$\frac{i - 0.375}{N + 0.25}$
Gringorton	$\frac{N + 0.12}{m - 0.44}$	$\frac{i - 0.44}{N + 0.12}$
Cunnane	$\frac{N + 0.2}{m - 0.4}$	$\frac{i - 0.4}{N + 0.2}$
Adamowski	$\frac{N + 1}{m}$	$\frac{i}{N + 1}$

Where

$i$  = rank in ascending order =  $N - m + 1$

$m$  = rank in descending order =  $N - i + 1$

$N$  = number of observations

## 2.5 Probability Plots

One of the most widely used method of determining best fit distribution is the probability plots method. The probability plot is a graphical technique for assessing whether or not a data set follows a given distribution. They are used to visually evaluate the agreement between distributions and observed data. Observed data are plotted against the values estimated from the fitted distribution. A straight line of a  $45^\circ$  slope through the origin should appear if the fitted distribution is the exact parent distribution. Estimates of the location and scale parameters of the distribution are given by the intercept and slope. Probability plots can be generated for several competing distributions to see which provides the best fit, and the probability plot generating the highest correlation coefficient is the best choice since it generates the straightest probability plot.

Another method that is being used widely to determine the best fit distribution is the correlation coefficient. Associated with the linear fit to the data in the probability plot is a measure of the goodness of fit. For distributions with shape parameters (not counting location and scale parameters), the shape parameters must be known in order to generate the probability plot. For distributions with a single shape parameter, the probability plot correlation coefficient (PPCC) plot provides an excellent method for estimating the shape parameter.

Stedinger et al. (1993) has shown that in the case of two-parameter distributions, a similar procedure may be applied before the parameters are estimated. For a given two-parameter distribution, the function can be written in the form of Equation 2.4,

$$F = G\left(\frac{Q - \alpha}{\beta}\right) \quad (2.4)$$

Where  $\alpha$  and  $\beta$  are the the distribution location and scale parameters, respectively, and thus the estimated quantile may be obtained from Equation 2.5,

$$\hat{Q} = \alpha + \beta G^{-1}(F) \quad (2.5)$$

Equation 2.5 represents a linear relationship between  $\hat{Q}$  and  $G^{-1}(F)$ . If the estimate of the observed  $Q$  is  $\hat{Q}$ , then the relationship between the observed  $Q$  and  $G^{-1}(F)$  should be linear if the fitted distribution is the exact parent distribution.

For two-parameter distributions,  $G^{-1}(F)$  depends only on  $F$ , which can be estimated by using a suitable plotting position formula (Section 2.4). By plotting the observed data  $q_i$  against  $G^{-1}(F)$ , those distributions which give a straight line relationship on the probability plot can be selected. A special plotting paper is prepared in advance so as the relationship between  $F$  and  $Q$  will appear as a straight line for each distribution. The values of  $Q_i$  can be directly plotted on the paper against the plotting position of  $F_i$ .

For three-parameter distributions, the function of  $G^{-1}(F)$  depends on the skewness coefficient,  $C_s$ , and probability plots can be used only for a specific value of predetermined  $C_s$ . Probability plots, although helpful in choosing between alternative distributions, suffer from a distinct possibility of error due to large variations in the plotting behavior of samples drawn from a given distribution.

## 2.6 Chi-Square ( $\chi^2$ ) Test

The Chi-square test (Snedecor and Cochran, 1989) is used to test if a sample of data came from a population with a specific distribution. An attractive feature of the Chi-square goodness-of-fit test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The Chi-square goodness-of-fit test is applied to binned data (i.e. data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the Chi-square test. However, the value of the Chi-square test statistic is dependent on how the data is binned.

Another disadvantage of the Chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

The Chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. The Chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

For the Chi-square goodness-of-fit computation, the data are divided into  $k$  bins and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^k \frac{[O_i - E_i]^2}{E_i} \quad (2.6)$$

where

$O_i$  is the observed frequency for bin  $i$

$E_i$  is the expected frequency for bin  $i$ .

The expected frequency is calculated by

$$E_i = N[F(Y_u) - F(Y_l)] \quad (2.7)$$

where

$F$  is the cumulative Distribution function for the distribution being tested,

$Y_u$  is the upper limit for class  $i$ ,

$Y_l$  is the lower limit for class  $i$ , and

$N$  is the sample size.

This test is sensitive to the choice of bins. There is no optimal choice for the bin width (since the optimal bin width depends on the distribution). Most reasonable choices should produce similar, but not identical, results. Data plot uses  $0.3*s$ , where ' $s$ ' is the sample standard deviation, for the class width. The lower and

upper bins are at the sample mean plus and minus  $6.0*s$ , respectively. For the Chi-square approximation to be valid, the expected frequency should be at least 5. This test is not valid for small samples, and if some of the counts are less than five, you may need to combine some bins in the tails.

## 2.7 Anderson-Darling Test

The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution.

The Anderson-Darling test is an alternative to the chi-square and Kolmogorov-Smirnov goodness-of-fit tests. The Anderson-Darling test statistic is defined as :

$$A^2 = -N - S \quad (2.8)$$

where

$$S = \sum_{i=1}^N \frac{(2i-1)}{N} [\ln F(Y_i) + \ln \{1 - F(Y_{N+1-i})\}] \quad (2.9)$$

$F$  is the cumulative distribution function of the specified distribution. Note that the  $Y_i$  are the ordered data. The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic,  $A$ , is greater than the critical value.

Note that for a given distribution, the Anderson-Darling statistic may be multiplied by a constant (which usually depends on the sample size,  $N$ ). These constants are given in the various papers by Stephens (1974). This is what should be compared against the critical values. Take a very particular attention and be aware that different constants and therefore different critical values have also been published. You just need to be aware of what constant was used for a given set of critical values (the needed constant is typically given with the critical values).

## 2.8 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. (Chakravarti et al., 1967). It is based on the empirical distribution function (EDF). Given  $N$  ordered data points  $Y_1, Y_2, \dots, Y_N$ , the EDF is defined as

$$E_N = \frac{n(i)}{N} \quad (2.10)$$

where  $n(i)$  is the number of points less than  $Y_i$  and the  $Y_i$  are ordered from smallest to largest value. This is a step function that increases by  $1/N$  at the value of each ordered data point.

Figure 2.1 below is a plot of the empirical distribution function with a normal cumulative distribution function for 100 normal random numbers. The K-S test is based on the maximum distance between these two curves.

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the Chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid).

Despite these advantages, the K-S test has several important limitations:

- i. It only applies to continuous distributions.
- ii. It tends to be more sensitive near the center of the distribution than at the tails.
- iii. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

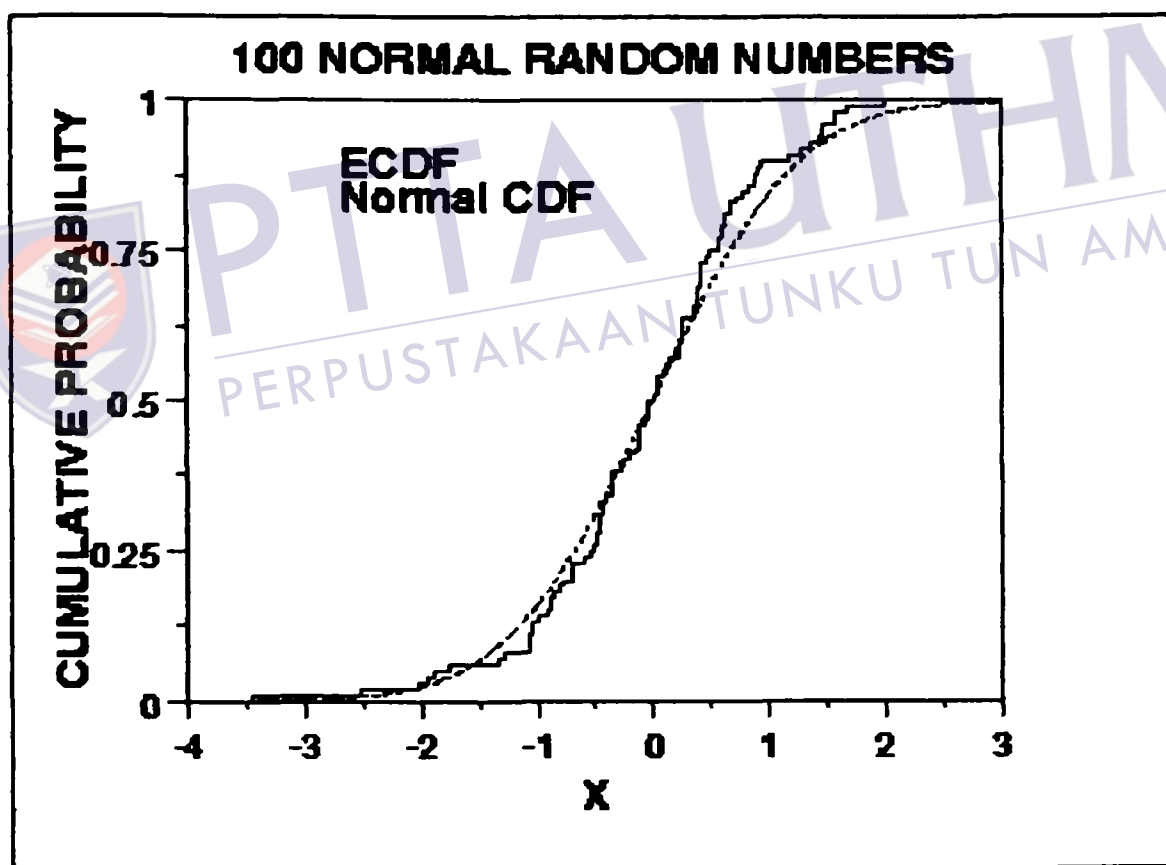


Figure 2.1 : Graph of Empirical Distribution Function Vs Normal Cumulative Distribution Function

Due to limitations 2 and 3 above, many analysts prefer to use the Anderson-Darling goodness-of-fit test. However, the Anderson-Darling test is only available for a few specific distributions.

The Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1 \leq i \leq n} \left| F(Y_i) - \frac{i}{N} \right| \quad (2.11)$$

where,

$F$  is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).

The hypothesis regarding the distributional form is rejected if the test statistic,  $D$ , is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scaling for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated.

## 2.9 The Probability Plot Correlation Coefficient (PPCC)

The probability plot correlation coefficient (PPCC) plot is a simple but powerful goodness-of-fit test developed by Filliben (1975). The test uses the correlation  $r$  between the ordered observations  $x_i$  and the corresponding fitted quantiles  $w_i = G^{-1}(1 - q_i)$ , determined by plotting positions  $q_i$  for each  $x_i$ . Values of  $r$  near 1.0 suggest that the observations could have been drawn from the fitted distribution. Essentially,  $r$  measures the linearity of the probability plot, providing a

quantitative assessment of fit. If  $\bar{x}$  denotes the average value of the observations and  $\bar{w}$  denotes the average value of the fitted quantiles, then

$$r = \frac{[\sum (x_i - \bar{x})(w_i - \bar{w})]}{\sqrt{[\sum (x_i - \bar{x})^2 \sum (w_i - \bar{w})^2]}} \quad (2.12)$$

Generally the PPCC plot is generated using the following procedures. For a series of values for the shape parameter, the correlation coefficient is computed for the probability plot associated with a given value of the shape parameter. These correlation coefficients are plotted against their corresponding shape parameters and the maximum correlation coefficient corresponds to the optimal value of the shape parameter. For better precision, two iterations of the PPCC plot can be generated; the first is for finding the right neighborhood and the second is for fine tuning the estimate.

The PPCC plot is used first to find a good value of the shape parameter. The probability plot is then generated to find estimates of the location and scale parameters and in addition to provide a graphical assessment of the adequacy of the distributional fit.

Apart from finding a good choice for estimating the shape parameter of a given distribution, the PPCC plot can be useful in deciding which distributional family is most appropriate. For example, given a set of reliability data, you might generate PPCC plots for a Weibull, lognormal, gamma, and possibly others, on a single page. This one page would show the best value for the shape parameter for several distributions and would additionally indicate which of these distributional families provides the best fit (as measured by the maximum probability plot correlation coefficient). That is, if the maximum PPCC value for the Weibull is 0.99 and only 0.94 for the lognormal, then we could reasonably conclude that the Weibull family is the better choice.

When comparing distributional models, the maximum PPCC value should not be simply chosen. In many cases, several distributional fits provide comparable

PPCC values. For example, a lognormal and Weibull may both fit a given set of reliability data quite comfortably. Typically, we would first consider the complexity of the distribution, that is, a simpler distribution with a marginally smaller PPCC value may be preferred over a more complex distribution. Likewise, there may be theoretical justification in terms of the underlying scientific model for preferring a distribution with a marginally smaller PPCC value in some cases. In other cases, we may not need to know if the distributional model is optimal, only that it is adequate for our purposes. That is, we may be able to use techniques designed for normally distributed data even if other distributions fit the data somewhat better.

Many statistical analyses are based on distributional assumptions about the population from which the data have been obtained. However, distributional families can have radically different shapes depending on the value of the shape parameter. Therefore, finding a reasonable choice for the shape parameter is a necessary step in the analysis. In many analyses, finding a good distributional model for the data is the primary focus of the analysis. In both of these cases, the PPCC plot is a valuable tool.

## 2.10 Theory of L-Moments

L-Moments are analogous to the conventional moments but are estimated by linear combinations of an ordered data set called L-statistic. L-Moments have a theoretical advantage over conventional moments of being able to characterize a wider range of distributions and, when estimated from a sample, of being more robust to the presence of outliers in the data. L-Moments and probability weighted moments (PWM) are analogous to ordinary moments in that their purpose is to summarize theoretical probability distributions and observed samples. Similar to ordinary product moments, L-Moments can also be used for parameter estimation, interval estimation and hypothesis testing.

Although the theory and application of L-Moments are parallels to those of the conventional moments, but L-Moments have several important advantages. The

**L-Moments ratio diagram**, being a graphical analyses of distribution fitting, has an advantage of being able to compare fitting of several distributions in a single graphical diagram (Vogel et al., 1993). It compares sample estimates of the coefficient of variance  $\tau_2$ , skewness  $\tau_3$ , and kurtosis  $\tau_4$  with their theoretical counterparts for a range of distributions.

Since sample estimators of L-Moments are always linear combinations of the ranked observations, they are subjected to less bias than ordinary product moments. The reason is that ordinary product moments estimators such as skewness and kurtosis requires squaring and cubing observations, which causes them to give greater weights to the observations far from the mean, resulting in substantial bias and variance.

### 2.10.1 L-moment Ratio Diagram

The theoretical L-Moments ratio diagram is constructed by first calculating the L-moments of the theoretical distributions, and then the values of  $\tau_3$  versus  $\tau_4$  for each distribution are plotted. Next step is to identify the L-Moments ratio of the samples and the values of  $\tau_3$  versus  $\tau_4$  for every station is plotted onto the theoretical L-Moments ratio diagram. The best approximation to the distribution of observed data is chosen from the plot.

The following equations will give the values of several distributions:

$$\tau_4^{pe3} = 0.1224 + 0.30115\tau_3^2 + 0.95812\tau_3^4 - 0.57488\tau_3^6 + 0.19383\tau_3^8 \quad (2.13)$$

$$\tau_4^{gev} = \frac{[1 - 6(2^{-\kappa}) + 10(3^{-\kappa}) - 5(4^{-\kappa})]}{(1 - 2^{-\kappa})} \quad (2.14)$$

# REFERENCES



PTTA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH

## REFERENCES

- Ashkar, F., Bobée, B., Bernier, J. (1992). "Separation of Skewness: Reality or Regional Artifact?" *Journal of Hydraulic Engineering*.ASCE. 460 – 475.
- Bartels, H. (1997). "Precipitation Depths for Germany – KOSTRA (Coordinated Storm Precipitation Regionalisation Analysis)." *Deutscher Wetterdienst (DWD)*. Offenbach/main, Germany.
- Bobée, B., and Rasmussen, P.F.(1995). "Recent Advances in Flood Frequency Analysis." *Reviews of Geophysics*. Vol. 33.
- Bobée, B., Cavadias, G., Ashkar, F., and Rasmussen, P.F. (1993). "Towards a Systematic Approach to Comparing Distributions used in Flood Frequency Analysis." *Journal of Hydrology*. 142. 121-136.
- Chakravarti, Laha, and Roy, (1967). "Handbook of Methods of Applied Statistics, Volume 1." John Wiley & Sons, Inc, USA. 392-394.
- Chow, V.T. (1964).Editor-in-Chief. "Handbook of Applied Hydrology." McGraw-Hill New York, NY.
- Chow, V.T., Maidment, D.R., and Mays, L.W. (1988). "Applied Hydrology." McGraw-Hill International, USA. 380 - 410.
- Chowdhury, J.U., Stedinger, J.R., and Lu, L.H. (1991). "Goodness-of Fit Tests for regional Generalized Extreme Value Flood Distributions." *Water Resources Research*. 27(7). 1765 - 1776.

- Cunnane, C. (1987). "Review of statistical methods for flood frequency estimation." in Singh, V.P. ed. *Hydrologic Frequency Modeling*. D. Reidel, Dordrecht. 49 - 95.
- Cunnane, C. (1989). "Statistical Distributions for Flood Frequency Analysis." *World Meteorological Organization Operational Hydrology*. Report No. 33, WMO – No. 718. Secretariat of the WMO, Geneva, Switzerland.
- Davison, A.C., and Smith, R.L. (1990). "Models for exceedances over high thresholds." *Journal of Royal Statistical Society. Series B.* 52(3), 393-442.
- Filliben, J. J. (1975). "The Probability Plot Correlation Coefficient Test for Normality." *Technometrics*. 17(1). 111-117.
- Greenwood, J.A., Landwehr, J.M., Matalas, N.C., and Wallis, J.R. (1979). "Probability Weighted Moments : Definition and Relation to Parameters of Several Distributions Expressible in Inverse Form." *Water Resources Research*. 15(5). 1049 - 1054.
- Helsel, D.R., and Hirsch, R.M., (1992). "Statistical Methods in Water Resources." Elsevier Science Publications Co., Inc. Netherlands.
- Hosking, J.R.M., (1991). "The Fortran Routines for Use With the Method of L-moments, Version 2." *Research Report RC17097*. IBM Research Division, Yorktown Heights, New York 10598.
- Hosking, J.R.M., (1990). "L-moments : Analysis and Estimation of Distributions using Linear Combinations of Order Statistic." *Journal of Royal Statistical Society. Series B.* 52(1). 105-124.
- Hosking, J.R.M., (1986). "The Theory of Probability Weighted Moments." *Research Report RC12210*. IBM Research Division, Yorktown Heights, New York 10598.

- Hosking, J.R.M, and Wallis, J.R. (1993). "Some statistics useful in regional frequency analysis." *Water Resources Research*. 29(2). 271-281.
- Houghton, J.C. (1978). "The Incomplete Means Estimation Procedure Applied to Flood Frequency Analysis." *Water Resources Research*. 14(6). 1111 - 1115.
- Mays, L. W. (2001). "Water Resources Engineering." John Wiley & Sons, Inc. USA. 309 - 342.
- Naghavi, B, and Fang, X. Y. (1995). "Regional Frequency Analysis of Extreme precipitation in Louisiana." *Journal of Hydraulic Engineering*. November 1995. 819 - 827.
- Potter, K. W.(1987). "Research on Flood Frequency Analysis: 1983-1986." *Reviews of Geophysics*. 25(2), 113-118.
- Ramachandra, A.R., and Khaled, H.H.(2000). "Flood Frequency Analysis." CRC Press LLC. Boca Raton, Florida, USA.
- Schaeffer, M.G.(1990). "Regional Analyses of Precipitation Annual Maxima in Washington State." *Water Resources Research*. 26(1). 119 - 131.
- Sevruk, B, and Geiger. H. (1981). "Selection of Distribution Types for Extremes of Precipitation." *World Meteorological Organization Operational Hydrology*. Report No. 15, WMO – No. 560. Secretariat of the WMO, Geneva, Switzerland.
- Snedecor, G.W., and Cochran, W.G. (1989). "Statistical Methods, Eighth Edition." Iowa State University Press.
- Stedinger, J.R, Vogel, R.M., and Foufoula-georgiou, E. (1993). "Frequency Analysis of Extreme Events." in Maidment, D.R. ed. "*Handbook of Hydrology*." McGraw-Hill Book Co.Inc., New York.

Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association*. **69**. 730 - 737.

U.S. Water Resources Council (USWRC), (1967). "Guidelines for Determining Flood Flow Frequency." *Bulletin No. 15*. Hydrology Subcommittee, Washington, D.C.

Vogel, R.M., McMahon, T.A., and Chiew, F.H.S. (1993). "Flood Flow Frequency Model Selection in Australia." *Journal of Hydrology*. **146**. 421 - 449.

Vogel, R.M., and Wilson, I. (1996). "Probability Distribution of Annual Maximum, Mean and Minimum Streamflows in The United States." *Journal of Hydrologic Engineering*. **1**(2), April 1996. 69 - 76.

Zalina Mohd Daud (2001). "Statistical Modelling of Extreme Rainfall Processes in Malaysia." Universiti Teknologi Malaysia. Ph. D thesis.



PTTA UTHM  
PERPUSTAKAAN TUNKU TUN AMINAH