MODELLING THE EXTREME VALUE OF RIVER FLOW DATA IN WEST SABAH USING BAYESIAN APPROACH

CHEONG RI YING

FACULTY OF SCIENCE AND NATURAL RESOURCES
UNIVERSITY MALAYSIA SABAH
2018

MODELLING THE EXTREME VALUE OF RIVER FLOW DATA IN WEST SABAH USING BAYESIAN APPROACH

CHEONG RI YING

THESIS SUBMITTED IN FULLFILMENT FOR THE DEGREE OF MASTER OF SCIENCE

FACULTY OF SCIENCE AND NATURAL RESOURCES
UNIVERSITY MALAYSIA SABAH
2018

PUMS 99:1

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN TESIS

JUDUL : ____

____

____

IJAZAH : ____

____

SAYA : ________________________________ **SESI PENGAJIAN** : ____________________________

(HURUF BESAR)

Mengaku membenarkan tesis *(LPSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah 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 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 PUSTAKAWAN)

Alamat Tetap: ________________

____

____

______________________ ______________________

(NAMA PENYELIA)

TARIKH: ______________ TARIKH: ______________

Catatan:
*Potong yang tidak berkenaan.
*Jika tesis ini SULIT dan TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT dan TERHAD.
*Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana Secara Penyelidikan atau disertai bagi pengajian secara kerja kursus dan Laporan Projek Sarjana Muda (LPSM).

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excepts, equations, summaries and references, which have been duly acknowledged.

30th July 2018

....
CHEONG RI YING
MS1621073T

CERTIFICATION

NAME	: **CHEONG RI YING**
MATRIC NO.	: **MS1621073T**
TITLE	: **MODELLING THE EXTREME VALUE OF RIVER FLOW DATA IN WEST SABAH USING BAYESIAN APPROACH**
DEGREE	: **MASTER OF SCIENCE (MATHEMATICS WITH ECONOMICS)**
VIVA DATE	: **24th July 2018**

CERTIFIED BY;

1. SUPERVISOR Signature

Dr. Darmesah Gabda ....

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude and appreciation to my supervisor, Dr. Darmesah Gabda, for her expert guidance, support and understanding that helped and guided me to complete this thesis. I learnt a lot of technical knowledge and positive attitude towards given tasks from her.

In addition, special thanks to Mr. Mohd Rizal bin Abdul Ghani from Hydrology and Survey Division, Department of Irrigation and Drainage, Sabah for providing me information about the station inventory. This has improved the contents of my study.

Besides that, I would like to thank all lecturers for the invaluable motivations and knowledge they have taught me since my undergraduate studies. I would also like to show my gratitude to my fellow friends for helping me throughout my study. Their unconditional encouragement supported me during my research life. Thanks also the university mates in Bilik Pascasiswazah 4 who always tolerate and understand each other.

Last but not the least, I would like to express my deepest appreciation to my family for their financial and mental support. This made it possible for me to complete my study on time.

CHEONG RI YING
30th July 2018

ABSTRACT

This study aimed to model the annual maximum series data of river flow in several sites in Sabah with small sample size to Generalized Extreme Value (GEV) distribution. In flood frequency analysis, annual maximum river flow is always used as an indicator. Uncertainty in the model and prediction process leads to inaccurate estimation of extreme events. Therefore, Bayesian approach is suggested to cope with the uncertainty involved. Maximum likelihood estimation (MLE) was treated as the standard parameter estimation method due to wide application in extreme value analysis. A stationary model which holds all the parameters constant is compared to two non-stationary models which consists of linear time dependent to location parameter as well as both location and scale parameters. A simulation study among probability weighted moment (PWM), MLE and Bayesian Markov Chain Monte Carlo (MCMC) were conducted to determine the best parameter estimation of GEV distribution. The performances were compared using root mean square error as well as bias. The results showed that Bayesian MCMC was better than PWM and MLE in estimating GEV parameters especially with small sample size. Likelihood ratio test showed that the annual maximum river flow data over the homogeneous region followed the distribution with common shape parameter. Hence, the quantile estimation at 10-, 100-, 1000-year return period were obtained using new single model with Bayesian MCMC as the parameter estimation. This method was believed to consider the parameter uncertainty and provide a more reliable return level estimates. Most of the stations were found to exceed the maximum level once every 100 years.

ABSTRAK

PEMODELAN NILAI EKSTRIM DENGAN MENGGUNAKAN KAEDAH BAYESIAN PADA DATA ALIRAN SUNGAI DI BARAT SABAH

Kajian ini bertujuan untuk menyuaikan data siri aliran maksimum tahunan sungai dari beberapa sungai di Sabah yang mempunyai saiz sampel kecil dengan menggunakan taburan nilai ekstrim teritlak (GEV). Aliran maksimum tahunan sungai sentiasa dijadikan sebagai penunjuk dalam analisis kekerapan banjir. Ketidakpastian dalam model dan peramalan akan menyebabkan penganggaran peristiwa yang kurang tepat. Pendekatan Bayesian adalah dicadangkan untuk mengambilkira ketidakpastian yang terlibat. Kaedah kebolehjadian maksimum (MLE) dijadikan sebagai pendekatan piawai disebabkan kebiasaan aplikasi ini dalam analisis nilai ekstrim. Model pegun yang mengandaikan semua parameter sebagai malar telah dibandingkan dengan dua model tidak pegun yang bergantung kepada masa pada parameter lokasi serta parameter skala. Kajian simulasi antara kaedah penganggaran telah dijalankan untuk menentukan penganggaran parameter yang terbaik iaitu antara kaedah momen berpemberat (PWM), MLE dan Bayesian Rantaian Markov Monte Carlo (Bayesian MCMC). Prestasi kaedah tersebut dibandingkan dengan menggunakan punca min ralat kuasa dua (RMSE) dan bias. Keputusan menunjukkan bahawa Bayesian MCMC adalah kaedah yang lebih baik berbanding dengan PWM dan MLE terutama sekali melibatkan saiz sampel yang kecil. Ujian nisbah kebolehjadian (LR-Test) menunjukkan bahawa aliran maksimum tahunan sungai di kawasan homogen mematuhi taburan yang mempunyai parameter bentuk yang sama. Oleh itu, anggaran kuantitatif pada tahap pemulangan 10-, 100- dan 1000-tahun telah diperolehi dengan menggunakan taburan baru ini berdasarkan kaedah Bayesian MCMC sebagai penganggaran parameter. Kaedah ini dipercayai dapat mengurangkan ketidakpastian pada parameter dan memberi nilai pulangan yang lebih kukuh. Kebanyakan nilai pulangan stesen dijangka akan melebihi tahap maksimum sekali pada setiap 100 tahun.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	-	Anderson-Darling
AIC	-	Akaike information criterion
AICc	-	Corrected Akaike information criterion
AMS	-	Annual maximum series
BIC	-	Bayesian information criterion
BM	-	Block maxima
cdf	-	Cumulative distribution function
CLT	-	Central limit theorem
EVT	-	Extreme Value Theory
GEV	-	Generalized extreme value
GPD	-	Generalized Pareto Distribution
i.i.d	-	independent and identical distributed
IPO-PDO	-	Interdecadal Pacific Oscillation-Pacific Decadal Oscillation
KS	-	Kolmogrov-Smirnov
MAX	-	Maximum
MCMC	-	Markov Chain Monte Carlo
MIN	-	Minimum
MLE	-	Maximum likelihood estimation
pdf	-	Probability distribution function
PDS	-	Partial duration series

POT	-	Peak-over-threshold
P-P	-	Probability-probability
PWM	-	Probability weighted moment
Q-Q	-	Quantile-quantile
RMSE	-	Root mean square error
Stan. dev	-	Standard deviation

LIST OF SYMBOLS

$+$	-	Addition
$-$	-	Subtraction
$\times, \cdot$	-	Multiplication
$-, /$	-	Division
$=$	-	Equal to
$\neq$	-	Not equal to
$>$	-	Larger than
$<$	-	Less than
$\leq$	-	Less than or equal to
$\propto$	-	Proportional
$\%$	-	Percentage
$\lvert \ \rvert$	-	Modulus
$\sqrt{\ }$	-	Square root notation
$\sum$	-	Summation notation
$\prod$	-	Product notation
∞	-	Infinity
$\rightarrow$	-	Tends to
$\therefore$	-	Hence
χ^2	-	Chi-square

γ	-	Likelihood ratio test
α	-	Significance level
exp	-	Exponential
ln	-	Natural logarithm
log	-	Logarithm
H_0	-	Null hypothesis
H_a	-	Alternative hypothesis
n	-	Number of observations
$\bar{x}$	-	Mean value of sample
p	-	Probability
t	-	Time covariate
θ	-	True parameter
μ	-	Location parameter
σ	-	Scale parameter
ξ	-	Shape parameter
$\beta_0, \beta_1, \alpha_0, \alpha_1$	-	Additional parameters
$\hat{\theta}$	-	Estimated parameter
$\hat{\mu}$	-	Estimated location parameter
$\hat{\sigma}$	-	Estimated scale parameter
$\hat{\xi}$	-	Estimated shape parameter
$\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\alpha_0}, \widehat{\alpha_1}$	-	Additional estimated parameters

$\boldsymbol{z_p}$	-	Return level
$\boldsymbol{\ell}$	-	Log-likelihood function
$\boldsymbol{L}$	-	Likelihood function
$\boldsymbol{\alpha_i}$	-	Acceptance probability
$\boldsymbol{\theta}^*$	-	Proposed value
$\boldsymbol{\pi}(\cdot)$	-	Target distribution
$\boldsymbol{\pi}(\boldsymbol{\theta})$	-	Prior distribution density
$\boldsymbol{f}(\boldsymbol{\theta}\|\boldsymbol{x})$	-	Posterior probability density
$\boldsymbol{f}(\boldsymbol{x}\|\boldsymbol{\theta})$	-	Likelihood of observations

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Extreme value theory (EVT) is a branch of statistic that deals with statistical techniques for modelling and estimation of rare events (Minkah, 2016). Most of the traditional statistical analyses study the body of the underlying distribution. However, EVT focuses on the tail estimation, where the extreme observations are located, over a certain time period. In general, the outcome of EVT is the estimation of the probability of the event which is more extreme than the observations itself. Historically, Nicolas Bernoulli started to work on extreme value problems in 1709. The theory began to develop after the publication of von Bortkiewicz in 1922 on the distribution of range in random samples from a normal distribution (Kotz & Nadarajah, 2000). In recent decades, besides geology and hydrology event analysis, EVT has been widely used in applied science and many other disciplines. For instance, financial markets, risk management and survival analysis.

Extreme risks are the risks of bad outcomes with low probability. Extreme events can be defined according to the cases and situations. The events might be caused by human or nature which includes terrorist attack, biosecurity and extreme natural hazards such as floods, earthquakes and landslides. The impacts can be measured in terms of death as well as cost. Extreme event risk management aims to model the rare but damaging events and hence to provide adequate decisions.

Block maxima or the largest value within a block is the beginning of the development as a method of modelling in extreme value theory. Maximum precipitation has always been treated as extreme values in rainfall analysis. Since rainfall is the primary cause of flooding hazard, it follows that the lasting period of water management and water reservoir seems important. Extreme value theory provides such a framework for these preventions. In the study of Coles and Tawn (1996), local data of 54-year series of daily rainfall was used to extrapolate the 100-year return level. Also, in wind speed analysis (Soukissian & Tsalis, 2015; Rajabi & Modarres, 2008), the results obtained are vital for the design of offshore platforms and coastal marine structures, coastal management, wind climate and structural safety. EVT is employed due to higher possibility in obtaining better estimate in strong wind events (D'Amico *et al.*, 2015). In modelling and predicting earthquake magnitudes, the maximum possible earthquake size is favourable for engineers to take necessary decisions in the construction of buildings and for insurers to make considerations on insurance policy (Pisarenko *et al.*, 2014).

EVT also performs well in extreme small value. It is believed that the safety of materials is determined by the stress level. If the strength of the material fails to overcome the stress level, the system of the material will break down. On the other hand, the fibre strength was collected and studied by Smith and Naylor (1987). The concerned value in the data set is the outlier which is believed that this influences the fibre strength. If the smallest fibre breaks, then the entire fibre breaks too. Therefore, a study of the smallest value in the data is important and can be carried out by the employment of extreme value theory.

In addition, the probability distribution can also be measured in terms of threshold, the analogue of annual maxima. In financial markets, Value-at-Risk (VaR) is the capital sufficient to cover losses in a given period and thereby estimates the potential size of the loss. The obtained return level can then be used as a measure of the maximum loss (Gilli & Këllezi, 2006). Ability in computing the tail risk measure and liquidity risk has encouraged the application of EVT in financial industry (Meula *et al.,* 2017). Also, VaR in insurance market is treated as a benchmark for risk estimation (Adesina *et al.*, 2016) in which the idea concerns more data over the series.

In road-safety analysis, EVT is employed to estimate the head-on-collision probability in passing manoeuvres (Farah & Azevedo, 2017). EVT is able to link crash frequency estimation and traffic conflict analysis using a single probabilistic framework. The results evaluated using EVT has a high probability to be less pessimistic as compared to the stochastic model (Mouradian, 2016). Moreover, Zheng *et al.,* (2014) applied EVT in transportation engineering due to the advantages over regression models in which EVT is able to estimate the return level from short data. The advantages had made EVT a powerful tool in area of risk management.

Changes in extremes have large impacts on human adaptation issues since they occur out of the coping range. In geophysical processes, EVT is applied to obtain better natural disasters' preparedness, prevention and mitigation. Natural disasters bring serious impact to a country and even to the world economy. Natural disasters damage tangible assets such as facilities, constructions and human capitals, hence, reduce the productivity and bring economic losses. In recent decades, human activities to the environment cause huge effects on the global climate. Due to global warming, the surface temperature on global surface has been increasing rapidly. Extremely high or low temperature leads to heat waves and cold waves, respectively. These phenomena affect human socio-economic activities and cause other natural catastrophe (Hasan *et al.*, 2012a; Lee, 2017). The mountain snow cover is known as one of the water resources, however, the extreme high temperature or extreme snowfall leads to flood (Blanchet *et al.*, 2009). According to 2016 Aon Benfield Catastrophe Report (Aon Benfield, 2017), the top three hazards are flooding, earthquake and severe weather. Among them, flooding is the costliest for four consecutive years which reached overall peril at $62 billion in year 2016. This amount was 30% of the total losses. The most remarkable floods were along the Yangtze River in China and in the US state of Louisiana. Both events cause $28 billion in damage and $10 billion to $15 billion in losses, respectively.

Flood is defined as an unusual high stage of river flow. This happens when the stream channel is filled and land is covered by water outside the normal confine (Zakaullah *et al.*, 2012). In other words, flood occurs when the water flow exceed the river channel capacity. Flood frequency analysis is a proposed method in determining design flood. A fairly accurate estimation of extreme flows with given return period is important in determining flood risk, reservoir management and construction of hydraulic structures such as tunnels, pumping stations, dams and spillways. An overestimation of flood magnitude may lead to waste of investment;

however, an underestimation of flood potential will lead to severe damages and casualty (Saghafian *et al.*, 2014; Ellouze & Abida, 2008). In previous studies, the annual maximum river flow was used as an indicator for the analysis of flood trends. It is believed that a good understanding of the probability distributions of river flows may be useful for water resource planning and management, including flood control and water supply during drought. Hence, in the effort of flood risk management in Malaysia, the study on maximum river flow has been conducted by Nur and Ani (2017) in Johor as well as Noor and Zulkifli (2017) in Segamat. The former suggested three-parameter log-normal distribution while the latter suggested generalized Pareto distribution (GPD) to describe the maximum river flows. Generalized extreme value (GEV) distribution and generalized logistic (GLO) distribution were also applied for regional flood frequency analysis in Sarawak (Lim & Lye, 2003). This is always the first step in modelling extreme value, analysing the data in the form of cumulative distribution and then determining the best fitting distribution function.

Most of the flood frequency analysis in technical literature focuses on stationarity of the distribution. Recent decade, statisticians and scientists have put non-stationarity in environmental extreme time series under spotlight. This mostly may be caused by climate variability as well as human behaviour (Toonen, 2015). River management, global warming, land use and nature protections are the leading causes of human-induced non-stationarity in flood frequency analysis. When global warming occurs, more water will evaporate from ocean and more melting glaciers will happen causing the atmosphere to carry much more water than usual. Changes in rainfall, snow or fog precipitation intensity are also due to the increasing in temperature and carbon dioxide concentration leading to droughts and flash floods occurring more frequently. River discharge will then gets out of control and burdens the hydraulic reservoir construction (Woodford, 2012). The impacts of climate change on river discharge have been studied by Xu and Luo (2015) in China and Parajuli *et al.* (2015) in Nepal. Infrastructure of impermeable rock or land will increase the river's capacity because the water is unable to soak into the ground.

A suitable covariate will improve the model fitting and reduce the modelling uncertainty (Jonathan & Ewans, 2013). There are many expressions of parameters which can account the non-stationarity in the selected model. Generally, location parameter and scale parameter are assumed to be time dependent or other covariate dependent. Time is usually observed in both parameters simultaneously (El Adlouni *et al.*, 2007) whereas the location parameter could be expressed in linear or nonlinear fully parametric, quadratic and exponential function. Also, the scale parameter is always concerned as a positive value. On the other hand, shape parameter is fixed and constant because the value is hard to be reliably estimated (Coles, 2001). The presence of trends in the data may influence the design values estimation (Cunderlik & Burn, 2003). However, argument about the data used in non-stationary model may bias the results obtained since the past condition and present situation might not be similar (Toonen, 2015).

Before the introduction of Bayesian paradigm in extreme value analysis, the studies on stationary and non-stationary frequency analysis adopted frequentist statistical tool. For instance, maximum likelihood estimation (MLE) is raised naturally in GEV distribution (Serinaldi & Kilsby, 2015). Also, inferences provided in MLE, such as standard error and statistical tests, are one of the factors that popularizes MLE as a frequentist method. Besides that, there are comparisons of the suitability among frequentist paradigm, such as L-moment as well as probability weighted moment (Engeland *et al.*, 2004). In the early of 1990, employment of Bayesian approach was extremely limited due to poor understanding. Contribution of Coles (2001) by applying Bayesian method in GEV distribution and generalized Pareto distribution widen the choices of parameter estimation in extreme value analysis. Advantages of prior belief in Bayesian paradigm are able to improve analysis when scarcity of data occurs. Also, performance in uncertainty analysis and accuracy of prediction are the benefits of Bayesian approach. Application of MLE and Bayesian paradigm in frequency analysis has been discussed and analysed in technical literature see for example Houkpè *et al.* (2015) and Silva *et al.* (2017).

1.2 Problem Statement

Every probability distribution of the random variable can be specified by the probability distribution function given by:

$$F(x) = \Pr\{X \leq x\} \quad (1.1)$$

where $F(x)$ is a non-decreasing function of the random variable x such that the lower and upper limit of sample space is 0 and 1, respectively. An aggregation process in statistical modelling will result a distribution with central part (mean) and outmost part (extremes). A clear explanation can be observed from Figure 1.1.

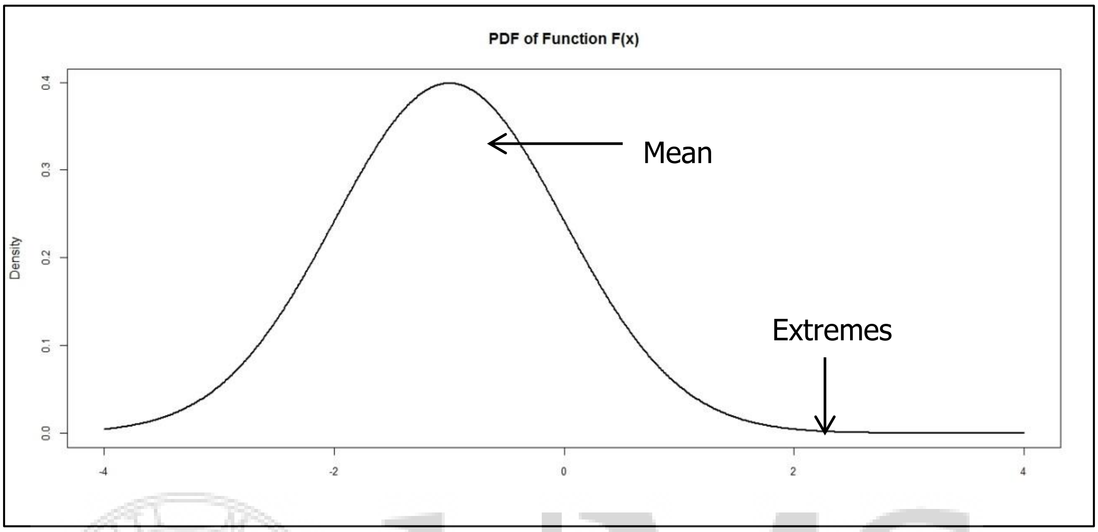

Figure 1.1: Probability density function of $F(x)$.

Central limit theorem (CLT) focuses on the changes around the mean; while EVT provides information on the extreme behaviour. CLT can be explained as follows (Devore, 2012),

Suppose that $x_1, x_2, \cdots, x_n$ be the random sample from a common distribution $F(x)$ with mean μ and finite variance σ^2, the function will converge to Gaussian distribution when n tends to infinity. A better approximate with the increasing of n.

This suggests that the region surrounded by mean can be explained by normal distribution but the lower and upper tails might follow other types of distribution. EVT hence provides the information regarding the tail distribution through continuous distribution according to their extreme realizations. In extreme events analysis, the distributional of the maximum and minimum values are determined by the upper tail and lower tail of the distribution, respectively.

Extreme value analysis might involve spatial analysis, univariate or multivariate model. Spatial analysis is a tool applied to analyse the events located over geographical region (Fischer, 2001). Merz and Blöschl (2008) studied the flood