UNIVARIATE GENERALIZED EXTREME VALUE APPROACH FOR SPATIAL EXTREME EVENT WITH SMALL SAMPLE SIZE: AN APPLICATION TO EXTREME RAINFALL IN SABAH

SIOW CHEN SIAN

FACULTY OF SCIENCE AND NATURAL RESOURCES
UNIVERSITI MALAYSIA SABAH

2023

UNIVARIATE GENERALIZED EXTREME VALUE APPROACH FOR SPATIAL EXTREME EVENT WITH SMALL SAMPLE SIZE: AN APPLICATION TO EXTREME RAINFALL IN SABAH

SIOW CHEN SIAN

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

FACULTY OF SCIENCE AND NATURAL RESOURCES
UNIVERSITI MALAYSIA SABAH
2023

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL : **UNIVARIATE GENERALIZED EXTREME VALUE APPROACH FOR SPATIAL EXTREME EVENT WITH SMALL SAMPLE SIZE: AN APPLICATION TO EXTREME RAINFALL IN SABAH**

IJAZAH : **SARJANA SAINS**

BIDANG : **MATEMATIK DENGAN EKONOMI**

Saya **SIOW CHEN SIAN**, Sesi **2018-2023**, mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini 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 dalam AKTA RAHSIA 1972)
[ ]	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
[/]	TIDAK TERHAD	

Disahkan Oleh,

ANITA BINTI ARSAD
PUSTAKAWAN KANAN
UNIVERSITI MALAYSIA SABAH

SIOW CHEN SIAN
MS1721044T

(Tandatangan Pustakawan)

ASSOC. PROF. DR. DARMESAH GABDA
Deputy Dean (Academic & International)
Faculty of Science and Natural Resources
Universiti Malaysia Sabah

Tarikh : 20 Januari 2023

(Assoc. Prof. Dr. Darmesah Gabda)
Penyelia

DECLARATION

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

23 May 2022

Siow Chen Sian
MS1721044T

CERTIFICATION

NAME	:	**SIOW CHEN SIAN**
MATRIC NO.	:	**MS1721044T**
TITLE	:	**UNIVARIATE GENERALIZED EXTREME VALUE APPROACH FOR SPATIAL EXTREME EVENT WITH SMALL SAMPLE SIZE: AN APPLICATION TO EXTREME RAINFALL IN SABAH**
DEGREE	:	**MASTER OF SCIENCE**
FIELD	:	**MATHEMATICS WITH ECONOMICS**
VIVA DATE	:	**23 MAY 2022**

CERTIFIED BY:

Signature

SUPERVISOR
Assoc. Prof. Dr. Darmesah Gabda

ASSOC. PROF. DR. DARMESAH GABDA
Deputy Dean (Academic & International)
Faculty of Science and Natural Resources
Universiti Malaysia Sabah

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere appreciation to my supervisor, Associate Professor Dr. Darmesah Gabda, who has given me a lot of guidance, advice, suggestion and encouragement to complete this thesis. I learnt a lot of technical knowledge and gained a positive attitude from her.

In addition, I especially thank the Hydrology and Survey Division, Department of Sabah, Malaysia, for providing me with the rainfall data of each station from Sabah in detail. And thanks to the university for providing us with some financial support to conduct our research (SBP scheme and UMSGREAT Grant [GUG0209-1/2018]).

Besides, I would like to thank my seniors, Miss Farhana and Riying Cheong, who helped me a lot throughout my study. I would also like to show my gratitude to my fellow friends for helping me during my study. Their unconditional encouragement supported me during my research life. Thanks also to the faculty for providing us with a room for postgraduate students to let us focus on our studies.

Last but not least, I would like to express my deepest appreciation to my family for their financial and mental support. They were always there to support me when I felt like giving up.

Siow Chen Sian

23 May 2022

ABSTRACT

This study aims to model the extreme event with small sample sizes using a univariate Generalized Extreme Value (GEV) distribution. The Maximum Likelihood Estimation (MLE) is the most recommended method for parameter estimation with GEV distribution due to the consistency of the results and wide application in extreme value analysis. However, the MLE performs poorly in small sample sizes, creating uncertainties that may lead to inaccurate estimation. Therefore, the Generalized Maximum Likelihood Estimation (GMLE) was suggested to improve the performance of MLE in modelling the small sample sizes of extreme events. A simulation study was conducted using several methods which are probability weighted moment (PWM), MLE, and GMLE to choose the most suitable parameter estimation of GEV distribution base on bias and root mean square error (RMSE). Other than that, the simulation results showed that GMLE performs better than PWM and MLE for GEV parameter estimations. A case study was conducted by fitting Sabah's annual maximum rainfall data with small sample sizes into GEV distribution with GMLE as the parameter estimation method. A stationary GEV model, which holds all parameters constant, is compared to a non-stationary model, consisting of a linear function of temperature as the covariate in the location parameter. From the results of the corrected Akaike's Information Criterion (AICc) and likelihood ratio test, there was insufficient evidence to prove the existence of a trend to the extreme rainfall. Besides, homogeneity testing was conducted for each district using the likelihood ratio test. It showed that all the rainfall stations from these five districts should be modelled independently without common shape parameters. Since the GEV was fitted independently at each site and the inter-dependency between sites was ignored, we applied the sandwich estimator to adjust the standard error. Hence, the quantile estimation at 10-, 100-, and 1000-years return period was carried out using a modified model. Most of the stations were found to be exceeded the maximum level once every 100-years.

ABSTRAK

PENDEKATAN NILAI EKSTRIM TERITLAK BAGI PERISTIWA EKSTRIM SPASIAL DENGAN SAMPEL SAIZ KECIL: APLIKASI KE ATAS HUJAN EKSTRIM DI SABAH

Kajian ini bertujuan untuk menyesuaikan peristiwa ekstrim dengan sampel saiz kecil menggunakan kaedah ekstrim univariat iaitu taburan nilai ekstrim teritlak (GEV). Kaedah kebolehjadian maksimum (MLE) merupakan pendekatan piawai bagi penganggaran GEV ini disebabkan ketekalan hasil dan aplikasi meluas dalam analisis nilai ekstrim. Walaubagaimanapun, MLE mempunyai prestasi yang lemah dalam saiz sampel yang kecil dengan kewujudan ketidakpastian dalam model. Ketidakpastian dalam model dan peramalan akan menyebabkan penganggaran peristiwa yang kurang tepat. Oleh itu, kaedah kebolehjadian maksimum teritlak (GMLE) dicadangkan untuk menambah baik prestasi MLE dalam saiz sampel yang kecil. Kajian simulasi dijalankan untuk membandingkan kaedah penganggaran iaitu kaedah momen berpemberat (PWM), kaedah MLE dan GMLE untuk memilih kaedah penganggaran GEV yang paling sesuai berdasarkan ketidakpincangan dan punca min ralat kuasa dua (RMSE). Keputusan menunjukkan GMLE merupakan kaedah yang paling sesuai berbanding dengan PWM dan MLE dalam saiz sampel yang kecil. Selanjutnya, kajian kes dijalankan dengan menyesuaikan hujan maksimum tahunan dengan saiz sampell kecil di Sabah menggunakan GEV dengan anggaran GMLE. Model pegun yang mengandaikan semua parameter sebagai malar telah dibandingkan dengan model tidak pegun yang bergantung kepada suhu cuaca pada parameter lokasi. Keputusan pengubahsuaian kriteria maklumat Akaike (AICc) dan ujian nisbah kebolehjadian (LR test) menunjukkan tidak terdapat bukti yang kukuh untuk membuktikan kewujudan trend dalam hujan ekstrim. Selain itu, keputusan ujian homogen bagi setiap daerah menunjukkan data hujan maksimum tahunan semua stesen dari lima daerah dimodelkan secara bebas tanpa parameter bentuk yang sama. Dalam kajian ini, GEV disesuaikan pada data hujan maksimum tahunan secara bebas antara stesen, maka kaedah penganggaran "sandwich" digunakan untuk pelarasan ralat piawai. Oleh itu, anggaran kuantitatif pada tahap pulangan 10-, 100-, dan 1000- tahun diperolehi dengan model terubahsuai. Kebanyakan nilai pulangan stesen dijangka melebihi tahap maksimum sekali setiap 100 tahun.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$+$	-	Addition
$-$	-	Subtraction
$\times$, $\cdot$	-	Multiplication
$-$, $/$	-	Division
$=$	-	Equal to
$\neq$	-	Not equal to
$>$	-	Greater than
$<$	-	Less than
$\geq$	-	Greater than or equal to
$\leq$	-	Less than or equal to
$\%$	-	Percentage
$\lvert\ \rvert$	-	Modulus
$\surd$	-	Square root notation
$\sum$	-	Summation notation
$\prod$	-	Product notation
∞	-	Infinity
$\rightarrow$	-	Tends to
∇	-	Vector differential operator
χ^2	-	Chi-square
θ	-	True value
μ	-	Location parameter
σ	-	Scale Parameter
ξ	-	Shape Parameter
α, β	-	Additional Parameters

$\hat{\theta}$	-	Estimated true value
$\hat{\mu}$	-	Estimated location parameter
$\hat{\sigma}$	-	Estimated scale parameter
$\hat{\xi}$	-	Estimated shape parameter
$\hat{\alpha}, \hat{\beta}$	-	Estimated additional parameters
Γ	-	gamma
exp	-	Exponentional
ln	-	Natural logarithm
log	-	Logarithm
H_0	-	Null hypothesis
H_1	-	Alternative hypothesis
n	-	Number of observations
t	-	Temperature covariate
k	-	Number of parameter
p	-	Probability
γ	-	Likelihood function
Z_p	-	Return level
$\pi(x)$	-	Beta prior
L	-	Likelihood function
$\ell(\hat{\theta})$	-	Log likelihood function
E	-	Expected value
$J(\hat{\theta})$	-	Partial derivative of log likelihood function
$I(\hat{\theta})$	-	Fisher information matrix

LIST OF ABBREVIATIONS

AIC	-	Akaike's Information Criterion
AICc	-	Corrected Akaike's Information Criterion
AMS	-	Annual maximum series
BIC	-	Bayesian Information Criterion
BM	-	Block maxima
cdf	-	Cummulative frequency function
CLT	-	Central limit theorem
EVT	-	Extreme Value Theory
GEV	-	Generalized extreme value
GLO	-	Generalized logistic
GMLE	-	Generalized maximum likelihood estimation
GOF	-	Goodness-of-fit
GPD	-	Generalized pareto distribution
GPWM	-	Generalized probability weighted moment
i.i.d	-	Independent and identical distributed
Max	-	Maximum
Min	-	Minimum
MCMC	-	Markov Chain Monte Carlo
MLE	-	Maximum likelihood estimation
pdf	-	Probability density function
PDS	-	Partial duration series
POT	-	Peak of threshold
PPCC	-	Probability plot correlation coefficient
PWM	-	Probability weighted moment

P-P	-	Probability-probability
Q-Q	-	Quantile-quantile
RMSE	-	Root mean square error
VaR	-	Value at risk

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Extreme Value Theory (EVT) is a branch of statistics that deals with statistical techniques for modelling and estimating rare events. EVT differs from most traditional statistical analyses that deal with the centre of the underlying distribution. Other than that, EVT focuses on the behaviour of the tails of the distribution function. The outcome of EVT is the examination of extreme observations (Minkah, 2016; Ramadhani *et al.*, 2016). Historically, Nicholas Bernoulli started to work on EVT in 1709. EVT has started by utilizing and rejecting outlying observations from the needs of astronomers. Then, the general theory began to develop with the publication of Bortkiewicz in 1992, which worked on the range distribution in random samples from a normal distribution (Kotz & Nadarajah, 2000). In recent decades, other than geology and hydrology events, EVT has been widely applied in applied science and other disciplines, such as financial risk, the insurance industry and traffic prediction.

There are two different models in EVT, which are block maxima and peak over threshold (POT). These two models are differentiated by the classification method of observation of extreme events that are used in the data analysis process. By the block maxima method, the samples are taken from a particular period, such as hourly, weekly, or yearly. Each period's maximum or minimum observations will be used as extreme observations. Using the POT method, a predetermined threshold is needed. The observations that exceed the given threshold will be considered extreme (Ramadhani *et al.*, 2016).

In application, the block maxima model is mostly used in climatological and hydrological data to determine the distribution of the maxima values. Maximum precipitation has always been treated as the extreme values in extreme rainfall analysis. Since rainfall is the primary cause of flooding, water management and water reservoir seem important. In wind speed analysis, the results obtained were essential for the design of offshore platforms, coastal marine structure, coastal management, wind climate and wind structural safety. Note that EVT has the ability to predict a better estimate of strong wind events (Rajabi & Modarres, 2008; Soukissian & Tsalis, 2015). Not only in wind speed analysis, but EVT also performed well in modelling and predicting earthquake magnitudes. The maximum possible size of an earthquake is beneficial for construction engineers and insurers to make considerations (Pisarenko *et al.*, 2014). The stress level was used to measure the safety of materials as the materials will easily break down when fails to overcome the stress level. EVT can also apply to the extremely small value. For example, the fibre strength collected by Smith & Naylor (1987). The outliers in a data set may influence the fibre strengths; if broken of smallest fibre may cause the entire fibre breaks too. Thus, EVT can be applied to study the smallest value in the data.

On the other hand, the POT model is mostly used in financial and insurance data which tends to be time independent. In the financial market, Value-at-Risk (VaR) is the capital sufficient to cover losses in a given period and estimate potential losses. EVT is able to compute the tail risk measure and the liquidity risk in the financial market. The return level of the distribution can be used to measure the maximum loss (Gilli, 2006). Besides, VaR in the insurance market is treated as the benchmark for risk estimation. VaR help in the estimate of the minimum amount of claims insurance in a given period (Adesina *et al.*, 2016). In the road safety analysis, EVT was applied to estimate the head-on-collision probability in passing manoeuvres (Farah & Avezedo, 2017). Apat from that, EVT is able to link frequency estimation and traffic conflict analysis using a single probabilistic framework. The result estimated using EVT has a high probability of fewer mistakes than the stochastic model (Mouradian, 2016). Furthermore, application in transportation engineering is due to the benefits of EVT over regression models, which are able to estimate the return level from short data (Zheng *et al.*, 2014).

Nowadays, it can be seen that there is an increase in the occurrence of natural disasters due to the global warming all over the world. Human activities in the environment are the main cause of global warming. Due to global warming, the surface temperature has been increasing rapidly. Extreme high and low temperature on the global surface lead to heat and cold waves (Hasan *et al.*, 2012). Additionally, human activities and natural disasters may be affected by this phenomenon. Natural disasters bring a lot of negative impact on a country and even the world economy. According to the annual climate catastrophe report (Benfield, 2017), the top three risks in the year 2012-2016 were floods, earthquakes and severe weather. The costliest weather events among these are the flooding event in the Yangtze River in China. A flood is defined as an unusually high stage of river flow. This happens when the stream channel is filled, and the water covers the land outside the normal confines (Zakaullah *et al.*, 2012). According to United Nations Intergovernmental Panel on Climate Change (IPCC) report in 2021, climate change will increase in all regions in the coming decades. It is likely that hot extremes, heat waves, and heavy precipitation events will continue to become more frequent.

Statistical modelling of extreme events is important for civil engineering and planners since the results could be used to estimate the ability to build the structure to survive under the utmost extreme conditions (Eli, 2012). Besides, the determination of hydrology extreme events is important for water resources management and the designs of the hydraulic structure such as pumping stations, tunnels, dams, and spillways (Chung & Kim, 2013). Moreover, extreme events may cause huge economic losses and negative impacts on agriculture and people. Consequently, it is very important to study the distribution as the return period of extreme events contributes to flooding risk, reservoir management, and construction development. Selection of a suitable probability distribution is always the first step in modelling extreme events, analysing data in the form of cumulative distribution, and determining the best fitting distribution function. In geophysical processes, EVT is applied to improve the prevention, preparedness, and mitigation of natural disasters.

Selection of a probability distribution is always the first step in modelling extreme events, analysing the data set in the form of cumulative distribution, and then determining the best fitting distribution function. Most previous studies recommended the Generalized Extreme Value (GEV) distribution as a probability distribution to model extreme events. In the effort of flood risk management, Lim & Lye (2003) conducted a study on maximum river flow for 23 gauged river basins in Sarawak using GEV distribution. Besides, Zalina *et al.* (2002) proved that the GEV distribution is the most appropriate probability distribution among eight candidates of probability distribution to model extreme rainfall in Peninsular Malaysia, as GEV distribution has good descriptive and predictive abilities. The annual maximum rainfall data in Alor Setar, Kedah, also modelled using GEV distribution, was conducted by Eli & Shaffie (2012). Other than river flow and extreme rainfall, the annual maximum temperature data modelled by GEV distribution was carried out by previous studies where the data were collected from numerous locations in Malaysia, India, and Europe with the period of 32 years, 117 years, and 68 years, respectively (Hasan *et al.*, 2012; Gurung *et al.*, 2021; Auld *et al.*, 2021).

When sample data are collected from numerous locations, it is considered spatial data. Normally, spatial and multivariate extreme analyses were employed to model spatial data to capture the dependency between sites. Since several locations refer to the multivariate variables, multivariate extreme value distribution is suitable to model spatial data collected from several locations (Dixon & Tawn, 1999). From the study of Coles & Tawn (1991), they applied a trivariate distribution to oceanographic data to capture the dependencies. Apart from that, a joint estimation proposed by Buishand (1991) is one of the methods for spatial extreme modelling to capture inter-site dependencies. Coles & Tawn (1996) model spatial extreme analysis to extreme rainfall in England. From their study, geographical bias was corrected, and spatial dependence was captured. However, these two methods may lead to high dimensional difficulties (Gabda & Tawn, 2017). There were previous studies that proposed a method by modelled spatial data using a univariate extreme approach independently using GEV distribution. However, the results that were modelled by the univariate extreme approach were based on the wrong assumption since the dependency was ignored. Therefore, a standard error modification is needed to capture the data dependency. The standard error modification proposed by Smith (1990) was the sandwich estimator. This method modified the standard error to capture the dependencies without affecting the parameters estimated independently. On the other hand, Zheng *et al*. (2015) compare the performance of the independent method with a sandwich estimator and three spatial extreme models. They presented that the independent method with a sandwich estimator performs better in terms of bias and root mean square error. Additionally, Northdrop & Jonathan (2011) have applied the sandwich estimator to model the hurricane-induced wave heights independently. Both studies showed that modelling independently with a sandwich estimator has a more efficient computation and avoids model misspecification. Among these aforementioned methods, this study will apply a sandwich estimator as the standard error modification with marginal estimation to avoid model misspecification and for efficient computation.

Parameter estimation is required in a probability distribution to determine the experimental values of the parameters. Several estimators can be employed with GEV distribution to obtain parameters and quantiles of the probability distribution, such as Maximum Likelihood Estimation (MLE), probability weighted moment (PWM) as well as L-moment. Note that MLE was the most recommended model because this method can be easily extended to non-stationary cases (Coles, 2001). MLEs are popular due to their inferences, such as standard error and statistical tests. However, scarcity of data is always a weakness in MLE. The sample sizes that are less than 50 are considered small sample sizes since MLE is only preferable when the sample sizes are modes$(n \geq 50)$ (Martin & Stedinger, 2000). Therefore, the employment of an alternative method of MLE is important to improve the performance in small sample sizes to reduce uncertainties. MLE was a physically infeasible shape parameter for small samples (Hosking *et al.*, 1985). Meanwhile, Generalized Maximum Likelihood Estimation (GMLE) is a modified MLE proposed by Martin & Stedinger (2000). This method is able to improve analysis and reduce the uncertainties in small sample sizes by adding a beta prior function to the shape parameter. In previous studies, GMLE was able to improve the performance of MLE without affecting the advantage of MLE, which retained the modelling flexibility and large sample optimality. Apart from that, Song *et al.* (2018) showed that GMLE outperformed in most of the cases in their study to estimate the