Original Research Article

AnalysisandModellingof ExtremeRainfall: Acasestudyfor Dodoma, Tanzania

Abstract : *The analysis of climate change, climate variability and their extremes has become more important as they clearly affect the human society and ecology. The impact of climate change is reflected by the change of frequency, duration and intensity of climate extreme events in the environment and on the economic activities. Climate extreme events, such as extreme rainfall threaten to environment, agricultural production and loss of people's lives. Dodoma daily rainfall data exported from R-Instat software were used after being provided by Tanzania Meteorological Agency. The data were recorded from 1935 to 2011. In this essay, we used climate indices of rainfall to analyse changes in extreme rainfall. We only used 6 rainfall indices related to extremes to describe the change in rainfall extremes. Extreme rainfall indices did not show statistical evidence of a linear trend in Dodoma rainfall extremes for 77 years. Apart from the extreme rainfall indices, this essay utilized two techniques in extreme value theory namely the block maxima approach and peak over threshold approach. The two extreme value approaches were used for univariate sequences of independent identically distributed (iid) random variables. Using Dodoma daily rainfall data, this essay illustrated the power of the extreme value distributions in modelling of extreme rainfall. Annual maxima of Dodoma daily rainfall from 1935 to 2011 were fitted to the Generalized Extreme Value (GEV) model. Gumbel was found to be the best fit of the data after likelihood ratio test of GEV and Gumbel models. The Gumbel model parameters were considered to be stationary and non-stationary in two different models. The stationary Gumbel model was found to be good fit of Dodoma maximum rainfall. Later, the levels at which maximum Dodoma rainfall is expected to exceed once, on average, in a given period of time T = 2, 5, 10, 20, 30, 50 and 100 years, were obtained using stationary Gumbel model. Lastly, the data of exceedances were fitted to the Generalized Pareto (GP) model under stationary climate assumption.*

Keywords - *Climate extreme indices; Extreme value theory; Generalized Extreme Value Distribution; Generalized Pareto Distribution; Block Maxima; Peak Over Threshold and Tail Distribution; Return level.*

INTRODUCTION

Extreme weather causes substantial damage to our lives through events such as extreme rainfall, floods and ecological disturbances as they affect human activities and the economy (Hasan et al., 2013). In Tanzania, flooding has been reported in 5 regions since mid January, 2016. At least 400 people have been displaced in Dodoma municipality after 70 houses were destroyed or damaged after heavy rains between 17 and 18 January 2016. Since then, flooding has been reported in Morogoro, Katavi, Mtwara and Dar es Salaam (Floods in Tanzania, January 2015). Some examples of the loss caused by floods in the region are the damage both to life and property experienced throughout the country during the 1997-1998 El Nino associated with floods, and the 2011 floods that wrecked the coastal city of Dar es Salaam. In recent years (2009-2011), heavy rains accompanied with strong winds have left thousands of people displaced and without food in Muleba, Kilosa, Same and Dar es Salaam. The flooding of 2009-2010 in Kilosa proved as serious, that over three quarters of the farmers reported their households were affected (Mboera,2011). Furthermore, in 2010, floods occurred in Kilosa (Morogoro), Mpwapwa and Kondoa (Dodoma) where more than 50000 people were affected, 5100 hectares of crops were destroyed and agricultural land was covered with mud and sand; public facilities were also destroyed (Source: arcjournals,2016).

2.0 METHODOLOGY

Various methods were applied to achieve the objectives of the study. Some of the methods were

2.1.1 Climate Extreme Indices

Climate indices allow a statistical study of variations of the dependent climatological aspects, such as analysis and comparison of time series, means, extremes and trends (Santikayasa,2015).

The World Meteorology Organization (WMO) developed the 27 indices which describe the changes in extremes. Indices are driven from the daily maximum and minimum temperatures and daily rainfall. In this paper, we only defined some extreme rainfall indices which are related to the objectives of the study.

2.1.2.1 *Extreme Rainfall Indices*

Six indices of rainfall extremes were considered. Some of them are percentile based; very wet days (R95p) and extremely wet days (R99p). Indices which represent maximum value within a year; highest daily precipitation (RX1day) and highest 5 consecutive days precipitation amount (RX5day) were analysed. Indices which represent the number of days on which the rainfall value falls above a fixed threshold; heavy rainy days (R20) and very heavy rainy days (R50) were also analyzed. In Table 2.1 below, each index was shortly defined.

Table 2.1: Definition of extreme rainfall indices

Extreme rainfall indices		
Index	Indicator name	Definition
R20	Heavy rainy days	Annual count of days when PRCP $\geq 20mm$
R50	Very heavy rainy days	Annual count of days when PRCP $\geq 50mm$ (threshold)
R95p	Very wet days	Annual total PRCP when RR > 95^{th} percentile
R99p	Extremely wet days	Annual total PRCP when RR > 99^{th} percentile
RX1day	Maximum 1-day rainfall amount	Annual maximum 1-day rainfall
RX5day	Maximum 5-day rainfall amount	Annual maximum 5-day rainfall

2.1.3 Observed change/trend in extreme Rainfall.

Changes in extreme rainfall in Dodoma were analysed through the annual and daily occurrence of rainfall. Changes in extreme rainfall can be studied by looking at the change in the frequency of days with precipitation exceeding some threshold; R10mm, R20mm and Rnnmm where nn represents any fixed threshold (Stephenson et al.,2014). Extreme rainfall is defined also as the highest daily precipitation (RX1day) or the highest 5 consecutive days precipitation amount (RX5day) per year or again extreme rainfall is a heavy rainfall event (R95p and R99p). The indices were chosen primarily for the assessment of many aspects of a changing global climate which include changes in intensity, frequency and duration of precipitation events. They represent events that occur several times per season or year giving them more strong statistical properties than measures of extremes which are far enough into the tails of the distribution so as not to be observed during some years (Stephenson et al.,2014).

This paper used the linear regression model to describe change of extreme rainfall over the time. Let *Y* be response variable and *T* be independent variable (Time). So, we fitted the following simple model:

$$Y_i = \alpha_0 + \alpha_1 T_i + Z_i \qquad i = 1 \ldots n$$

where α_0 is an intercept and α_1 is the slope which describes the change of extreme rainfall over time. After fitting this model to the data, we made the following inference,

$$H_0 : \alpha_1 = 0 \qquad against \qquad H_1 : \alpha_1 \neq 0$$

to check if there is a relationship between the extreme rainfall and time. To test the statistical significance of relationship between time T and the extreme rainfall Y, the significance level of 0.01 was used.

All climate extremes indices for rainfall presented in Table 2.1 are calculated using data from Dodoma and the analysis and results are presented in chapter 3. Climate extremes indices can be used to define extremes and analyse changes in extremes. However, those indices do not give the answer to the question of return levels of extreme rainfall. Thus, extreme value distributions are introduced in the next section.

2.1.4 Extreme Value Distributions

In this section we reviewed the model which focuses on the statistical behaviour of

$$M_n = \max\{X_1, X_2, ..., X_n\},$$

where $X_1, X_2, \ldots, X_n$, is a sequence of independent random variables having a common distribution function F (Coles et al., 2001). In applications, X_i usually represent values of a process measured on a regular time-scale, then we take the maxima over particular blocks of time to extract the upper extreme values from a set of data. For example, in this essay $X_1, X_2, \ldots, X_n$ represent Dodoma daily rainfall since 1935 to 2011. If n is the number of observations in a year, then M_n corresponds to the annual maximum of the daily rainfall over $1935-2011$ period.

Now, could we derive the distribution for M_n for all n? to answer this question, we used the probability theory to find the possible limit distributions of the maxima M_n. From probability theory, $F(x)$ the cumulative distribution function of X is defined as

$$P(X \leq x) = F(x).$$

If F is known, the distribution of M_n is derived exactly for all values of n as follow:

$$P[M_n \leq x] = P[X_i \leq x]; \qquad i = 1, 2, \ldots, n. \tag{2.2.1}$$

By using the fact of independence Equation 2.2.1 becomes

$$\begin{aligned}\mathbb{P}[X_i \leq x] &= \mathbb{P}[X_1 \leq x]\,\mathbb{P}[X_2 \leq x] \ldots \mathbb{P}[X_n \leq x], \\ &= (\mathbb{P}[X \leq x])^n.\end{aligned}$$

As the X_i are independent identically distributed with a common distribution F

$$P[M_n \leq x] = F^n(x). \tag{2.2.2}$$

For unknown distribution F, we use the limit laws of convergence in distribution to approximate F^n for large n.

Theorem (Fisher-Tippett 1928; Gnedenko, 1943). *If the sequence $\{X_i\}$ are iid random variables with the distribution function F and $\{a_n > 0\}$, $\{b_n\}$ are sequences of normalizing constants. Then, if there*

~~*exist constants* $\{a_n\}$, $\{b_n\}$ *and a non-degenerate distribution function* G *so that,*~~

as $n \to \infty$,

$$P\left[\frac{M_n - b_n}{a_n} \leq x\right] = F^n(a_n x + b_n) \xrightarrow{d} G(x)$$

then it must be of the same type as one of the following three types of distributions:

Gumbel distribution

$$G(x) = exp\left\{-exp\left(-\frac{x - b}{\alpha}\right)\right\}, -\infty < x < +\infty$$

Weibull distribution

$$W(x) = \begin{cases} exp\left\{-\left[-\left(\frac{x-b}{\alpha}\right)\right]^{\alpha}\right\}, & if\ x < b \\ 1, & if \geq b; \end{cases}$$

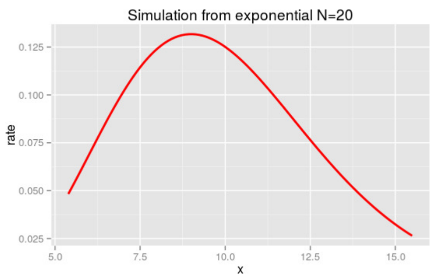
Frechet distribution

$$F(x) = \begin{cases} exp\left\{-\left[-\left(\frac{x-b}{\alpha}\right)\right]^{-\alpha}\right\}, & if\ x < b \\ 0, & if \leq b \end{cases}$$

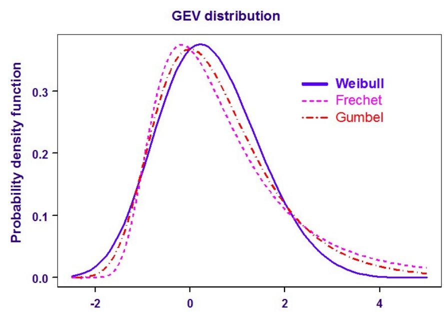
with parameters α, b *and* $\alpha > 0$ *namely scale, location and shape parameters respectively.*

2.1.1 Remark. The Theorem 2.2.1 is also known as **Extremal type's theorem** while the three max-stable distributions are **Gumbel, Weibull** and **Fréchet**.

Convergence of sample maxima



Sample maxima $N = 1000$



$As\ n \to \infty$

Figure 2.1: The extremal types theorem: the power of this theorem is to approximate the distribution of sample maxima as n increases to be max-stable distribution regardless of the parent population X_i.

If the Theorem 2.2.1 holds for suitable choices of α_n and b_n then we say that G is an extreme value cumulative distribution and F is in the domain of attraction of G, written as $F \in D(G)$. However, G can take the form of the generalized extreme value distribution which unifies three extreme value distributions known as **Gumbel**,**Weibull** and **Frechet** (Coles et al.,2001). The unified extreme value distributions G is defined by

$$P(X \leq x) = G(x) = exp\left[-\left(1+\xi\left(\frac{x-b}{\alpha}\right)\right)^{\frac{-1}{\xi}}_{+}\right], -\infty < \mu, \qquad \xi < +\xi \; and \; \sigma > 0$$

(2.2.3)

with $z+ = max\{z, 0\}$. From Equation 2.2.3, we derive the GEV density function by using the probability theory of cumulative and density function by applying derivative of cumulative distribution as follows

$$g(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-b}{\alpha}\right)\right]^{\frac{-1}{\xi}-1} + exp\left[-\left(1+\xi\left(\frac{x-b}{\alpha}\right)\right)^{\frac{-1}{\xi}}_{+}\right], -\infty < \mu, \qquad \xi < +\xi \; and \; \sigma > 0$$

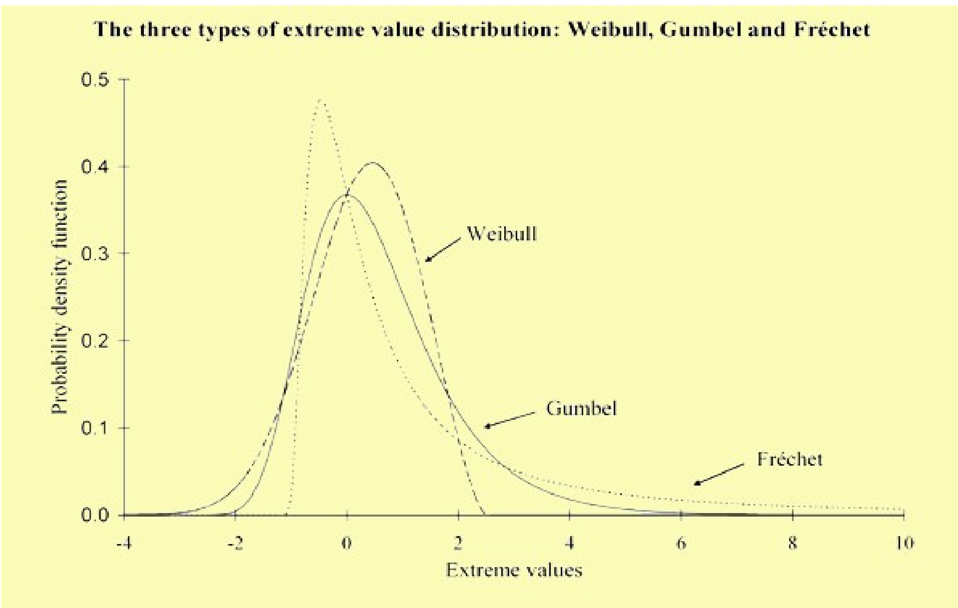
with $z+ = max\{z, 0\}$. (2.2.4)

(2.2.4)

The GEV distribution and its density function have three parameters namely (μ, ξ, σ), location, shape and scale parameters respectively. $G(x)$ and $g(x)$ can be denoted by $G(\mu, \xi, \sigma)$ and $g(\mu, \xi, \sigma)$ respectively. The x are the extreme values from the block maxima.

Remark. The shape parameter ξ governs the tail behaviour of the distribution. When fitting the GEV model to sample data, the sign of the shape parameter ξ will usually indicate which one of the three models best describes the random process we are dealing with (Coles et al.,2001).

- For $\xi \longrightarrow 0$, light tail (Gumbel type)
- For $\xi < 0$, bounded upper tail (Weibull type)
- For $\xi > 0$, heavy tail (Fréchet type)



Modelling by Generalized Extreme Value Distributions

The Block Maxima approach description

In ordinary statistics, we describe the main part of the distribution; many ignore outliers. However, in the statistics of extremes we characterise the tail of the distribution by keeping only the extreme observations. We do not care about mean and variance, we care only about tails. If we fit the one distribution to entire data sets, we shall often miss the tail. Therefore, we take data and we extract some data which are said to be extreme. One of the methods of extracting extreme data is the *block maxima method*. In this method, the idea is to break the data into the monthly/annual blocks of equal length then extract the maximums from each month/year and fit the model to that data (monthly/annual maxima) (Coles et al.,2001). The right distribution to fit block maxima is the *generalised extreme value (GEV) distribution* as shown in Equation 2.2.3. In practice, the implementation of this model for any particular data, to choose the block size is critical because of the following reasons:

i. By the limit model in Theorem 2.2.1, blocks that are too small are likely to have poor approximation, which leads to bias in estimation and extrapolation.

ii. Large blocks generate few block maxima, leading to large estimation variance.

a) Maximum likelihood estimation

Let us denote the maximum of a sample $X_1, X_2, \ldots, X_n$ by Y. So, a sample $Y_1, Y_2, \ldots, Y_n$ of independent sample maxima has a common GEV distribution. The parameters σ, μ and ξ of GEV distribution can be estimated by using different methods. Various methods of estimation for fitting GEV model have been proposed: least squares estimation, maximum likelihood estimation, probability weighted moments and others. In this essay, we focus on the maximum likelihood (ML) method because of its flexibility to any model.

Consider $Y_1, Y_2, \ldots, Y_m$ independent random variables such that

$$Y_i \sim G(y;\, \sigma, \xi, \mu),\, i = 1, 2, \ldots, m.$$

The GEV log-likelihood function is:

$$\log\left(L(\sigma,\xi,\mu)\right) = \begin{cases} -m\log\sigma - (\xi^{-1}+1)\sum_{i=1}^{m}\log\left(1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right) - \sum_{i=1}^{m}\left(1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ -m\log\sigma - \sum_{i=1}^{m}\exp\left\{-\left(\frac{y_i-\mu}{\sigma}\right)\right\} - \sum_{i=1}^{m}\left(\frac{y_i-\mu}{\sigma}\right), & \text{if } \xi = 0; \end{cases}$$

defined when $\left\{1+\xi\left(\frac{y_i-\mu}{\sigma}\right) > 0,\ i = 1, 2, \ldots \ldots m.\right\}$ (2.2.5)

The ML estimates with respect to the entire GEV family are obtained by maximising the Equation 2.2.5 with respect to the parameter vector (σ, ξ, μ). It is possible to obtain the maximum likelihood estimator explicitly, usually by differentiating the log-likelihood and equating to zero.

b) Inference for return levels

Definition. A return period, also known as a recurrence interval is defined as an estimate of the likelihood of an event, such as extreme rainfall, flood or a river discharge flow to occur.

In simple terms, the **return level** is associated with the corresponding return period and indicates the maxima can reach within such a return period. We used the annual block maxima approach which consists of fitting the GEV model to a series of annual maximum data with n taken to be the number of iid events in a year. The T-year return value is formally defined by setting Equation 2.2.3 to

$1-\frac{1}{T}$

x_p is then the solution to the resulting equation. We need to choose an optimal threshold x_p such that the probability that an observed value exceeds x_p is equal to p, where $p=\frac{1}{T}=P[M_n > x_p]$ is the upper tail probability.

$$exp\left[-\left\{\xi\left(\frac{x_p-\mu}{\sigma}\right)\right\}^{\frac{-1}{\xi}}\right]=1-\frac{1}{T} \qquad (2.2.6)$$

Solving Equation 2.2.6 for x_p, we obtain

$$x_p=\begin{cases}\mu-\frac{\sigma}{\xi}\left(1-[-\log(1-p)]^{-\xi}\right), & for \quad \xi\neq 0;\\ \mu-\sigma\log[-\log(1-p)], & for \quad \xi=0.\end{cases} \qquad (2.2.7)$$

In terms of extreme value terminology, x_p is the return level associated with the return period p and it is common to extrapolate the relationship (2.2.7) to obtain estimates of return levels considerably beyond the end of the data to which the model is fitted. After estimating the GEV parameters by maximum likelihood method, we obtain the maximum likelihood estimates of, x_p by substituting estimated GEV parameters into Equation 2.2.7

$$\hat{x}_p=\begin{cases}\hat{\mu}-\frac{\hat{\sigma}}{\hat{\xi}}\left(1-y_p^{-\hat{\xi}}\right), & for \quad \hat{\xi}\neq 0;\\ \hat{\mu}-\hat{\sigma}\log y_p, & for \quad \hat{\xi}=0;\end{cases} \qquad (2.2.8)$$

The p–year return level, x_p , is the level an extreme is expected to exceed once every n time-units.

Modelling by Generalised Pareto Distributions.

The Peak Over Threshold approach description

Modelling by generalized extreme value distribution is based on the block maxima approach. However, the block maxima approach does not consider all maximums. It considers only the highest value in all maximum values. Therefore, sometimes using only the block maxima can be wasteful if it ignores much of the data. It is often more useful to look at exceedances over a fixed high threshold instead of simply the maximum or minimum of the data. Consider values of X_i to be extreme if they are above (below) a high (low) threshold u. In peak over threshold method, we fix the threshold and we extract the data exceeding the threshold. Let $\{X_i\}$ be the sequence of independent random variables with common distribution function F and M_n be the sample maxima of the sequence $\{X_i\}$ (Coles et al., 2001).

Theorem. *Denote an arbitrary term in the* X_i *sequence by* X, *and suppose that* F *satisfies* Theorem 2.2.1.

By Theorem 2.2.1, for a large n,

$$p[M_n \leq x] \approx G(x),$$

where

$$G(x)=exp\left\{-\left(1+\xi\left(\frac{x-b}{\alpha}\right)\right)_+^{\frac{-1}{\xi}}\right\}, \text{ for some } \mu,\sigma>0 \text{ and } \xi$$

Then, for large enough μ, the distribution of $Y=X-\mu$, conditional on $(X>n)$, is approximately

$$p(Y\leq y)=p(y)=1-\left(1+\frac{\xi y}{\sigma+\xi(u-\mu)}\right)^{\frac{-1}{\xi}}, if\ \xi\neq 0 \qquad (2.2.9)$$

Defined on $\{y: y > 0$ and $(\sigma + \xi(u - \mu)) > 0.\}$

For $\xi = 0$, which is interpreted as limit $\xi \to 0$ in (2.2.9), leading to

$$p(y) = 1 - exp\left(-\frac{y}{\sigma_u}\right), \tag{2.2.10}$$

Where $\sigma_u = \sigma + \xi(u - \mu)$.

As $y = x - u$, the two Equation 2.2.10 and (2.2.9) can be written as

$$p(x) = \begin{cases} 1 - \left(1 + \xi\left(\frac{x-b}{\sigma_u}\right)\right)^{\frac{-1}{\xi}}, & if\ \xi \neq 0 \\ 1 - exp\left(-\frac{x-u}{\sigma_u}\right), & if\ \xi = 0, \end{cases} \tag{2.2.11}$$

The family of distributions defined by Equation 2.2.9 is known as **generalised Pareto family.** Therefore, if block maxima have approximating distribution G, then threshold excesses have a corresponding approximate distribution within the generalised Pareto family (Coles et al.,2001).nFrom Equation 2.2.11, we derive the density function of the generalised Pareto distribution

$$p(x) = \begin{cases} \frac{1}{\sigma_u}\left(1 + \xi\left(\frac{x-b}{\sigma_u}\right)\right)^{\frac{-1}{\xi}-1}, & if\ \xi \neq 0 \\ \frac{1}{\sigma_u} exp\left(-\frac{x-u}{\sigma_u}\right), & if\ \xi = 0, \end{cases} \tag{2.2.12}$$

Remark. There are three types of generalised Pareto distribution which are:

Exponential $(\xi = 0)$, Pareto $(\xi > 0)$ and Beta $(\xi < 0)$.

a) Threshold selection

One consideration for POT modelling is the right choice of threshold. In practice, the implementation of this model for any particular data set to choose the right threshold is critical because of the following reasons: the threshold that is too low is likely to violate the asymptotic basis of the threshold model, which leads to bias in estimation and extrapolation (Coles et al.,2001). Too high threshold generates few excesses, leading to high estimation variance. To handle this challenge, two methods are available: the first method is an exploratory technique carried out prior to model estimation. The second is to assess the stability of parameter estimates, based on the fitting of models across a range of different thresholds. There are two common graphical tools that can help in choice of the threshold. The first is the mean excess plot.

Remark. Above a threshold u_0 at which the generalised Pareto distribution provides a valid approximation to excess distribution, the mean residual life plot should be approximately linear in u.

In the second method, we plot the parameter estimates and confidence intervals at different thresholds. The estimated parameters remain constant above the threshold at which the asymptotic approximation is valid.

Above a level u_0 at which the asymptotic motivation for the generalised Pareto distribution is valid, estimates of the shape parameter ξ should be approximately constant, while estimates of σ_u should be linear in u.

b) Maximum likelihood estimation

After determining the threshold, the generalised Pareto distribution parameters can be estimated by using the maximum likelihood method. Let $y_1, y_2, \dots, y_k$ be k excesses of a threshold u. The log-likelihood is derived from (2.2.12) as

$$l(\sigma,\xi,\mu) = \begin{cases} -k\log\sigma_u - \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{k}\log\left(1+\xi\left(\frac{x_i-\mu}{\sigma_u}\right)\right) & for\ \xi \neq 0 \\ k\log\sigma_u - \frac{1}{\sigma_u}\sum_{i=1}^{m}((x_i-\mu)), & for\ \xi = 0, \end{cases} \qquad (2.2.13)$$

defined $1+\xi\left(\frac{x_i-\mu}{\sigma_u}\right)$, $i = 1, 2, \dots, k$. We obtain the ML estimates $(\hat{\sigma}_u, \hat{\xi})$ for (σ_u, ξ) by maximizing numerically Equation 2.2.13.

c) Inference on the return levels

The more convenient way of interpreting extreme value models is using the quantiles or return levels, rather than individual parameter values. So, we suppose that a generalized Pareto distribution with parameters σ and ξ is a suitable model for exceedances of a threshold u by a variable X. For $x > u$,

$$P[X > x|X > u] = \left(1+\xi\left(\frac{x-u}{\sigma}\right)\right)^{\frac{-1}{\xi}}$$

It means that

$$P[X > x|X > u] = \tau_u\left(1+\xi\left(\frac{x-u}{\sigma}\right)\right)^{\frac{-1}{\xi}}$$

where $\tau_u = p[X > u]$. Thus, for $\xi \neq 0$ the level x_m that is exceeded on average once every m observations is the solution of

$$\tau_u\left(1+\xi\left(\frac{x-u}{\sigma}\right)\right)^{\frac{-1}{\xi}} = \frac{1}{m}.$$

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left[(m\tau_u)^{\xi} - 1\right], & for \quad \xi \neq 0 \\ u + \sigma\log(m\tau_u), & for \quad \xi = 0, \end{cases} \qquad (2.2.14)$$

provided m is sufficiently large to ensure that $x_m > u$.

To estimate the return levels, we substitute the parameters by their corresponding maximum likelihood estimates. However, the probability of an individual observation exceeding the threshold u has a natural estimator of $\tau_u = \frac{k}{n}$,

the sample proportion of points exceeding u (Coles et al., 2001).

Stationary and non-stationary model.

Most of the time when one deals with real life data some assumptions are violated. Therefore, in this essay we considered both assumptions, stationarity and non-stationarity of climate extremes data. Climate is change over period and the reliable future projections of extreme rainfall cannot rely only on stationary assumption. Under the assumption of non-stationarity, we have non-stationary model with a linear trend in location parameter. Using the notation $GEV\ (\mu, \sigma, \xi)$ to denote the GEV distribution with parameters $GEV\ (\mu, \sigma, \xi)$ it follows that a suitable model for X_t, the annual maximum Dodoma rainfall in year t, might be

$$X_t \sim GEV\ (\mu_t, \sigma, \xi),$$

where

$$\mu_t = \mu_0 + \mu_{1t}$$

with parameters μ_0 and μ_1. In this way, variations through time in the observed process are modelled as a linear trend in the location parameter of the appropriate extreme value model, which in this case is the GEV model. The parameter μ_1 corresponds to the annual rate of change in annual maximum rainfall. Non-stationarity can be expressed in terms of the location parameter as follow:

$$\boxed{GEV\,(\mu_t, \sigma_t, \xi_t) = \exp\left[-\left(1 + \xi_t\left(\frac{x_t - \mu_t}{\sigma_t}\right)\right)_+^{-\frac{1}{\xi_t}}\right],} \tag{2.2.15}$$

where

$$\mu_t = \mu_0 + \mu_{1t},$$

$$\sigma_t = \sigma,$$

$$\xi_t = \xi.$$

For $\xi = 0$,

$$G(\mu, \sigma) = exp\left\{-exp\left\{-\frac{x_t - \mu_t}{\sigma_t}\right\}\right\}. \tag{2.2.16}$$

The advantage of maximum likelihood over other techniques of parameter estimation is its adaptability to the changes in model structure (Coles et al., 2001). That is why for this non-stationary model, we did not change our previous model. We maximised the Equation 2.2.5 by considering a linear trend in location parameter. Note that, for stationarity, the GEV and GP models assume that the parameter location, scale and shape are time-independent (parameters are constants).

Likelihood Ratio (LR) Test and Model Diagnostics or goodness-of-fit checks.

As in any statistical model, after fitting model, we check the good of fit of the model. The Likelihood Ratio (LR) test is used to compare the fit of two models where the null model, H_0 is a special case of the other (alternative model, H_1) (Hasan et al., 2013). The best model is determined by deriving the probability or p-value of the difference in γ, the LR test statistic, defined as $\gamma = -2\ln\left(\frac{H_0}{H_1}\right)$, where γ has a chi-square distribution. However, LR test is applied to nested models, which means that comparison can only be made between one complex model and one simpler model (Hasan et al.,2013). In the model checking we are comparing the observed data to GEV or GP estimates. We use probability Plots, quantile plots, return level plots and density plots to assess the quality of a fitted GEV or GP Model. The probability plot compares the empirical and fitted distribution functions.

i. The probability plot should lie close to the unit diagonal. In probability plot, we Look for linearity and deviations in tails.

ii. Quantile plot compares observed quantiles in data to quantiles estimated by the GEV. In quantile plot, we also Look for linearity and deviations in tails.

iii. The return level plot.

iv. The density function of fitted GEV or GP model is compared to histogram of block maxima (histogram of exceedances for GP model).

Data preparation.

The daily rainfall data obtained for Dodoma starts in January 1935 and lasts in December 2013. The data had no missing values apart from the last two years. The data were supplied by Tanzania Meteorological Agency in 2013.Table 3.1 below details the information of the missing values.

Table 3.1: Missing values in Dodoma daily rainfall data

Variable	Period	Month	Number of missing values
Rain	2013	November	30
		December	31
	2012	November	30

As shown in Table 3.1, all years had values except the last two years. Hence, we chose to use the data to 2011. We shifted years so that we obtain all extreme rainfall in the same season for Dodoma. Then, the daily rainfall data starts from August 1935 and ends in July 2011 (see Figure 3.4). The number of observations did not change because we brought half of the data for 1935 to 1934 and the last year ends in July 2011.

Data description.

We put our data into two main groups; rainy days and dry days to get rainy season data for extreme rainfall. As we were interested in studying the behaviour of maximum rainfall in Dodoma, we considered rainy days (Rain > 10.0*mm* in our data). The Dodoma data has 4 variables; Year, Month, Date and Rain. The statistical summaries for rainy days between 1935 and 2011 are presented in a table below.

Table 3.2: Statistical summaries of Dodoma rainy days.

1st Qu.	Mean(mm)	3rd Qu.	Min	Max	Std(mm)	Median	observations(days)
14.3	27.1	33.5	10.2	119.8	17.6	21.1	1337

Table 3.2 shows that in our Dodoma daily rainfall data approximately 5% of the daily rainfall exceeds 10*mm*, and this was 1337 out of 28124 days. In total, we have 1337 rainy days for Dodoma from 1935 to 2011 and the data for rainfall were recorded in millimetre (*mm*). The maximum rainfall in our data was 119.8*mm* which occurred on 02 Feb 1964. The average daily rainfall was 27.1*mm*. The table below describes the Dodoma daily rainfall on the monthly basis.

Table 3.3: Statistical monthly summaries of rainfall from 1935 to 2011

R10(days)	**R10 per year(days)**	**chance to rain each year** (%)	**Mean** (*mm*)	**Max**(mm)	**Std**(mm)	**Month**
315	4	13	27.2	113.00	18.3	**Jan**
274	4	13	25.9	119.8	16.3	**Feb**
259	3	11	28.2	102.5	18.9	**Mar**
117	2	5	26.7	92.5	16.9	**Apr**
7	0	0	19.1	41.7	11.8	**May**
2	0	0	10.8	11.4	0.8	**Jun**
0	0	0	-	0.6	-	**Jul**
0	0	0	-	0.8	-	**Aug**
0	0	0	-	2.2	-	**Sep**
9	0	0	30.7	54.4	12.9	**Oct**
60	1	2	24.5	90.4	16.7	**Nov**
294	4	12	27.8	107.0	17.4	**Dec**

In Table 3.3, the column with R10 represents the number of days in each month with daily rainfall greater than 10*mm* for 77 years. Mean column is the average monthly rainfall for 77 years. The chance to rain was calculated taking R10 per year divided by days of the month. The Table 3.3above shows, for example, July, August and September were found to be months with almost no rain, while January, February, March and December were the wettest months with some possibility of the daily rainfall being greater than 10*mm*. To better understand the behaviour of the daily maximum rainfall for Dodoma, some analyses were investigated using graphical methods.

d) Different graphical Description of Dodoma daily rainfall data

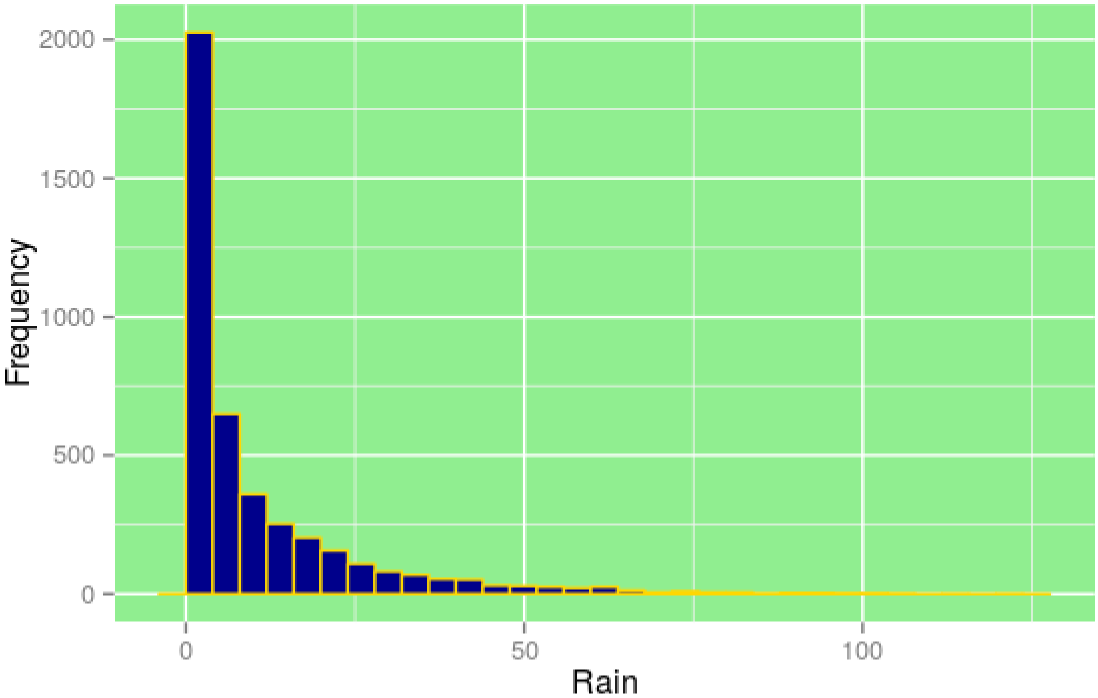


Figure 3.2: Histogram of Dodoma daily rainfall

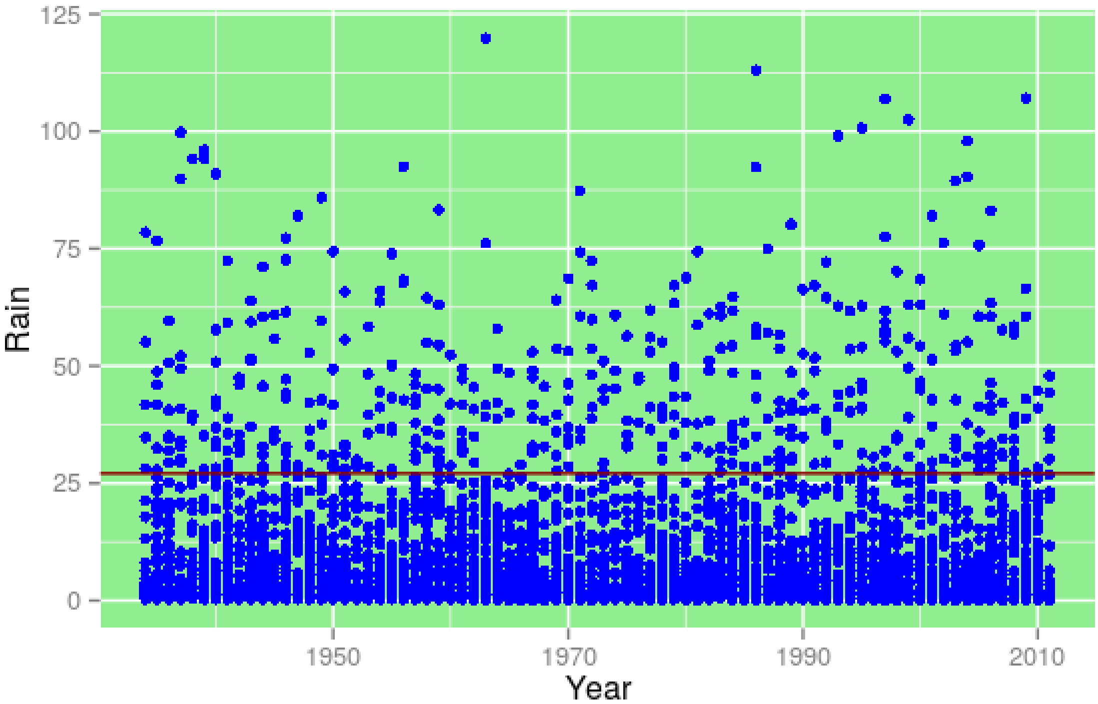


Figure 3.3: **Daily rainfall scatter plot**

Figure 3.3 above shows the daily rainfall. We observe that all years had rainfall above 27.1*mm*, which is the average rainfall of rainy days for 77 years (seeTable 3.2) . The red line represents the average rainfall in Dodoma for 77 years. This scatter plot can give us an idea about the extreme rainfall by studying the behaviour (distribution) of rainfall exceeding the average rainfall. We defined the year to start in August and end in July as shown in a monthly boxplot in Figure 3.4 below.

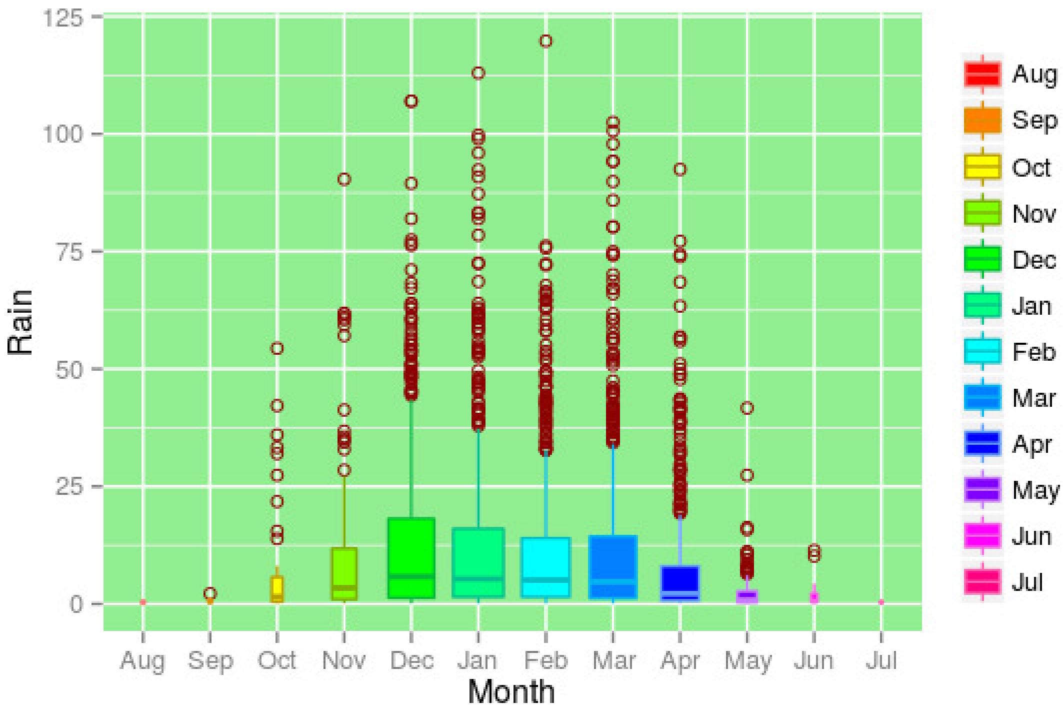


Figure 3.4: Monthly boxplot of Dodoma daily rainfall.

This box plot is showing the variability in the daily rainfall on monthly basis across the years for the Dodoma station. The daily rainfall rises during the wet season (from November to December and again from January to April) and declines during the dry period (from May to September). Several periods of heavy rainfall in Tanzania since 14 January 2016 have caused flooding in the regions of Mwanza and Dodoma. The Tanzania Meteorological Agency issued a warning of severe weather in most parts of the country, with possible rainfall of 50*mm* in 24 hours expected in many areas until 16 April (Rainfall and forecasts, 14 January 2016). We used records of extreme rainfall causing floods in some regions, to fix the threshold to describe extreme rainfall in our Dodoma data. The plots below show the daily rainfall of Dodoma from 1935 to 2011 exceeding some heavy rainfall causing floods in Dodoma.

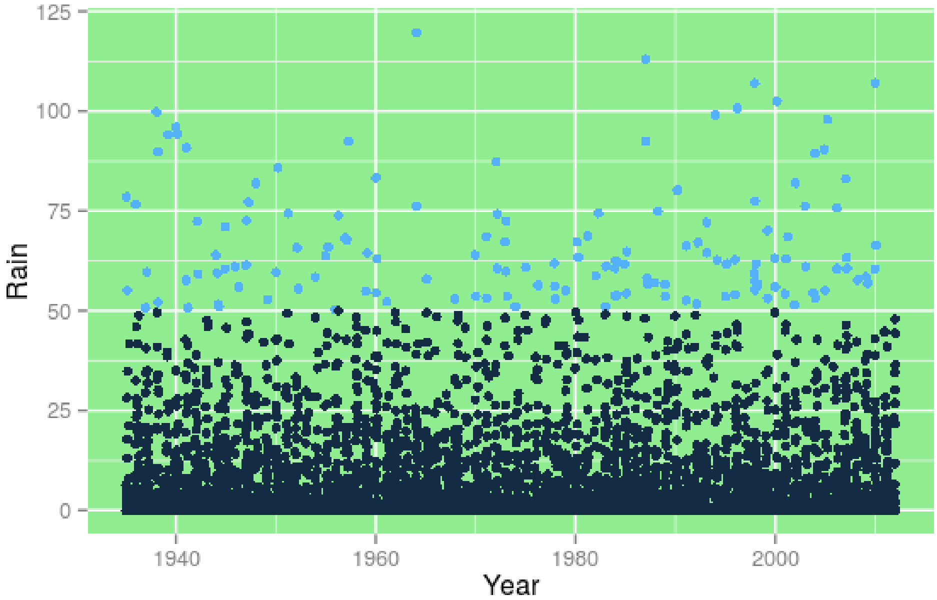


Figure 3.5: Daily rainfall for Dodoma with rain exceeding 50*mm*

The blue points represent the daily rainfall in Dodoma exceeding 50*mm* each year for 77 years. Data with rainfall greater than 50*mm* is much scattered compared to others.

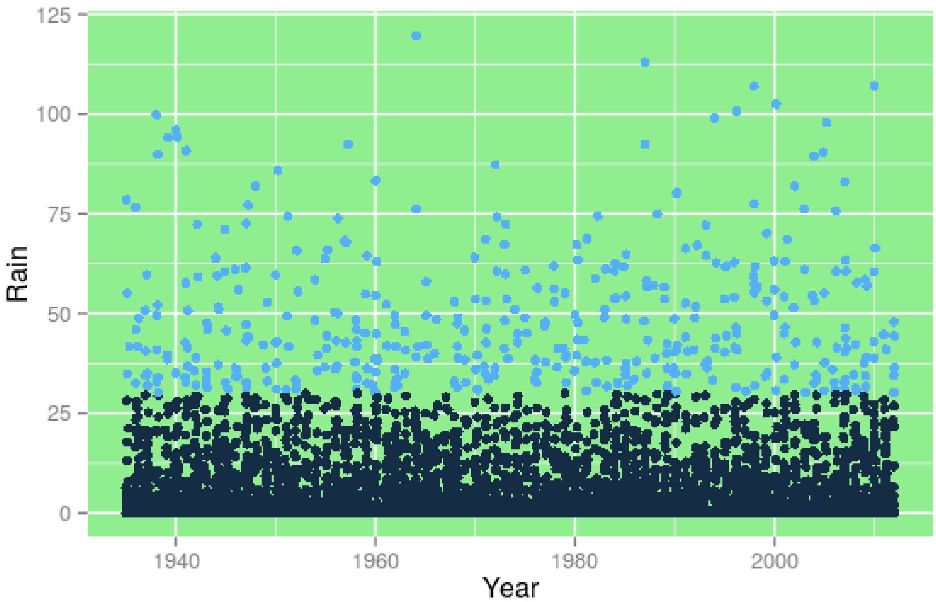


Figure 3.6: Daily rainfall for Dodoma with rain exceeding 30*mm*

The blue points represent daily rainfall in Dodoma exceeding 30*mm* each year since 1935. Data with rainfall greater than 30*mm* is scattered compared to others. This scatter plot does not show obvious trend in daily rainfall greater than 30*mm*. Even if the above scatter plots did not show an obvious trend in rainfall exceeding some maximum rainfall, we need some statistical evidences to confirm this. In the next section we used precipitation indices to see whether there was a linear trend in
Extreme rainfall or not.

Analysis of a linear trend in extreme rainfall over time using rainfall indices

Changes in extreme rainfall in Dodoma were analysed through the annual and daily occurrence of rainfall. In Table 2.1, we described some useful indices to analyse extreme rainfall. Changes in extreme rainfall can be studied by looking at the change in the frequency of days with precipitation exceeding some threshold; $R10mm$, $R20mm$ and $Rnnmm$ where nn represents any fixed threshold (Stephenson et al., 2014). Extreme rainfall is defined also as the highest daily precipitation ($RX1day$) or the highest 5 consecutive days precipitation amount ($RX5day$) per year or again extreme rainfall is a heavy rainfall event (R95p and R99p) (Alexander et al.,2006). As extremes are defined based on the occurrence, frequency and intensity, the plots below are based on some threshold of exceedance, intensity and frequency of rainfall in Dodoma. To study the trend in rainfall extremes over period of 77 years, linear regression model was used and the fitted line (in red) indicates linear trend in occurrence, frequency and intensity of extreme rainfall. How often does extreme rainfall in Dodoma occur? Is there any statistical evidence of change in extreme rainfall in Dodoma over a period of 77 years?

Using the indices described inTable 2.1, we can answer those questions. In this essay, 7 precipitation indices related to exceedances, frequency and duration of rainfall were analysed. To determine whether there exist a linear trend, a linear regression of rainfall indices against year was fitted. The slopes of the annual trends of extreme rainfall indices were calculated based on a least square linear fitting. Trends were obtained for each index and the statistical significance of the trends were assessed using a p-value. The trends were considered to be statistically significant at 99% confidence level.

The observed linear trend of extreme rainfall indices are presented below

Table 3.4: Summary of linear regression model

index	p-value(slope)	R-squared	Trend line
R20	0.79	0.00	$Y = -0.01 \times T + 18.59$
R50	0.18	0.01	$Y = 0.01 \times T - 15.71$
R95p	0.20	0.01	$Y = 0.59 \times T - 1038.52$
R99p	0.74	0.00	$Y = 0.11 \times T - 175.36$
RX1day	0.16	0.01	$Y = 0.13 \times T - 195.39$
RX5day	0.14	0.02	$Y = 0.28 \times T - 442.84$

The line in red is a trend-line computed by least square fit and the corresponding regression equation is presented for each index in Table 3.4.

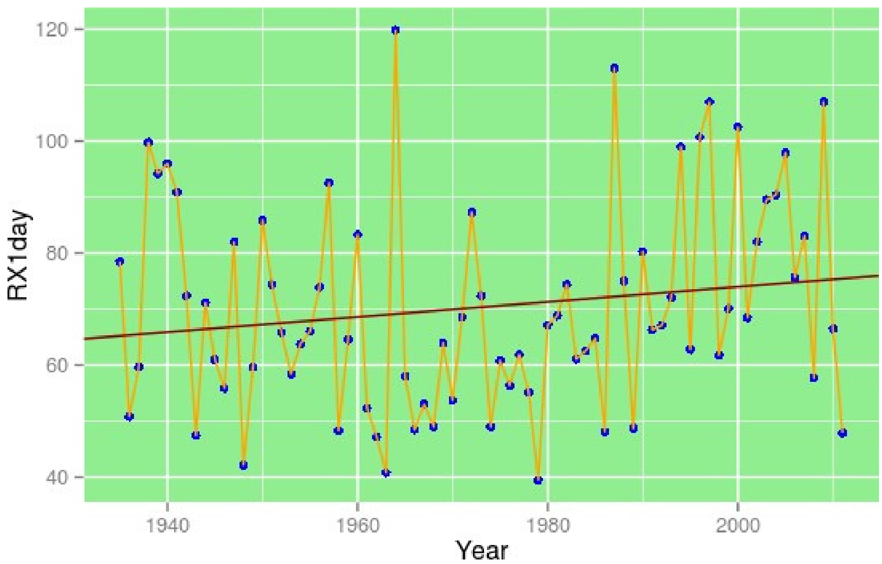


Figure 3.7: Annual daily maximum rainfall with a regression line $y = 0.13t - 195.39$.

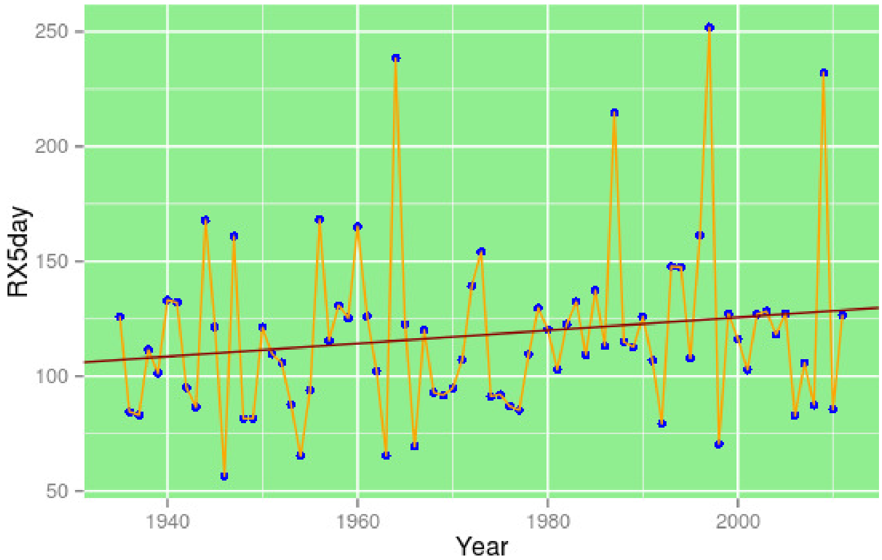


Figure 3.8: Annual maximum of 5-day consecutive rainfall with a regression line $y = 0.28t - 442.83$.

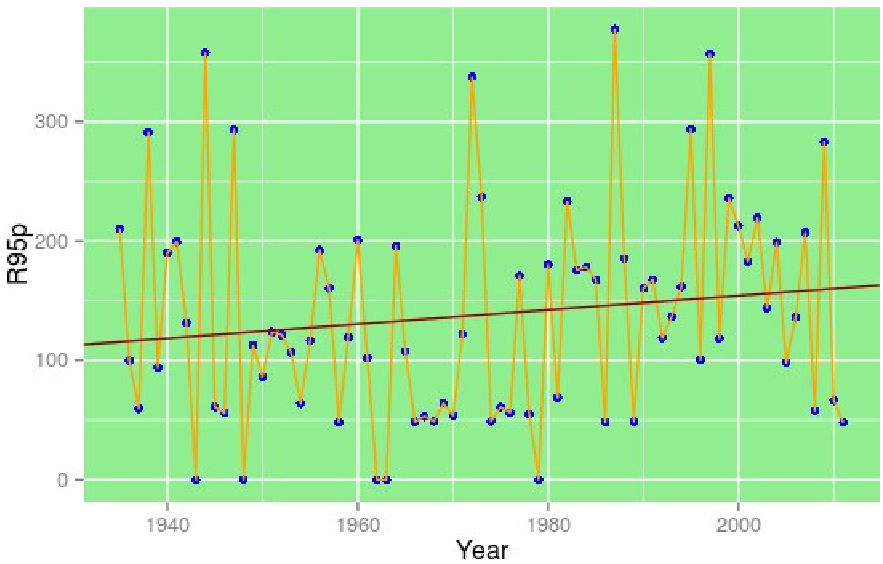


Figure 3.9: The exceedance of 95 percentile threshold with a regression line $y = 0.59t - 1038.51$.

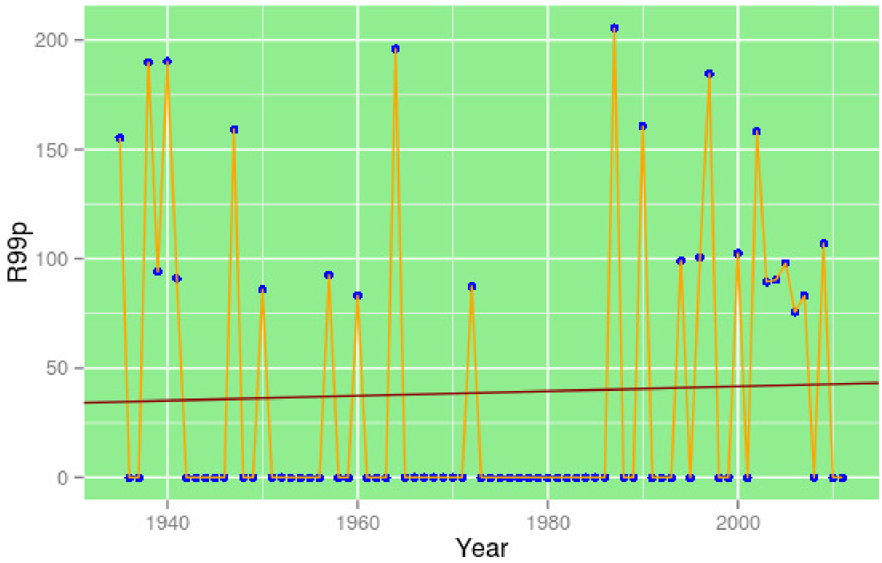


Figure 3.10: The exceedance of 99 percentile threshold with a regression line $y = 0.11t - 175.36$.

Figure 3.10 and Figure 3.9above represent very wet days and extremely wet days: The 95^{th} and 99^{th} percentiles describe the annual precipitation amount accumulated on days when daily precipitation is greater than the (95^{th}) and (99^{th}) percentiles threshold of the wet-day precipitation (Rain> $1mm$).

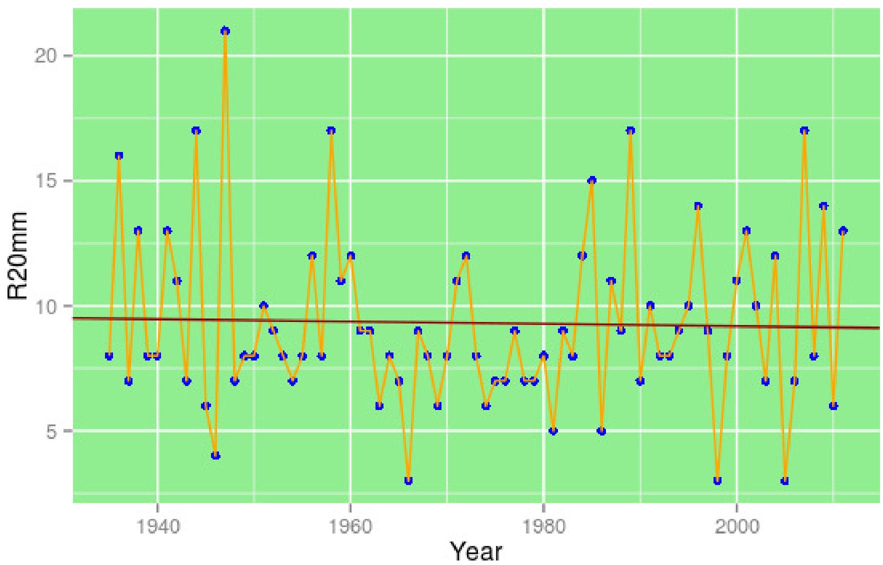


Figure 3.11: Annual counts of days with daily rainfall exceeding $20mm$ with a regression line $y = -0.01t + 18.59$.

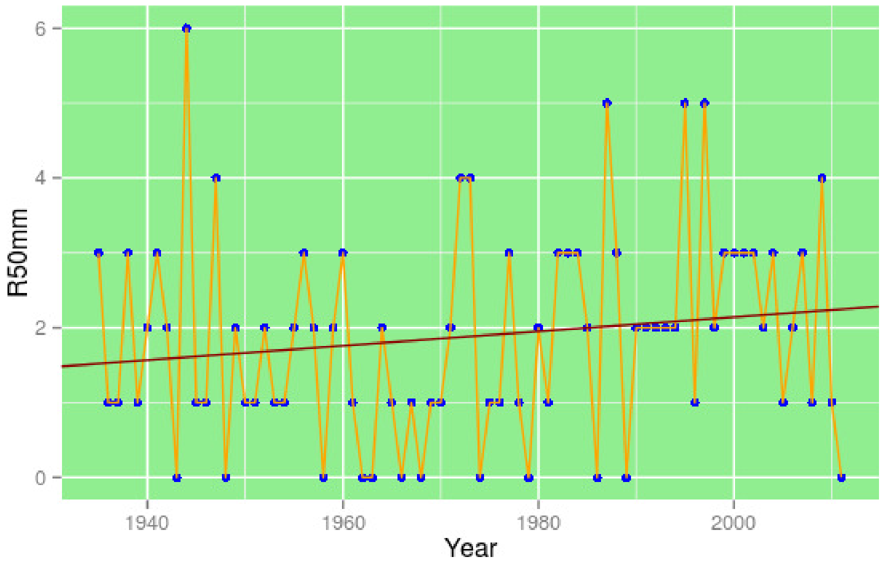


Figure 3.12: Annual counts of days with daily rainfall exceeding 50*mm* with a regression line $y = 0.01t - 15.71$.

Figure 3.11 and Figure 3.12 above represent the heavy rainfall days (Rain $> 20, 50mm$). An increase shown in annual counts of days with rainfall exceeding 50*mm* could indicate an increase of extreme rainfall in Dodoma for 77 years. We observed a positive slope of the trend line to all indices except R20, which means that the rainfall extremes increased over time. However, we need statistical test to confirm this change in extreme rainfall. From Table 3.4, all the p-values were greater than 0.01 level of significance. We therefore did not have enough evidence to reject the null hypothesis, since our test was not significant. In addition, the value of the R-squared is very small. This implies there is a very small (for example 1 % for R50) variability in extreme rainfall that can be explained by the change in time. Then, we concluded that the rainfall indices showed that there is no statistical evidence of the change in rainfall extremes in Dodoma between 1935 and 2011.

One of the unanswered questions clearly by the above indices is the distribution of observed extreme rainfall in Dodoma since 1935. As extreme events are also defined as those in the tail of the distribution, to accurately assess potential changes in the shape of the distribution of rainfall observations requires additional rigorous analysis rather than using the rainfall indices. To study the distribution of extreme rainfall over period in Dodoma, extreme value distributions were used. In the next section, extreme value distributions with stationary and non-stationary parameters were fitted to observations of Dodoma daily rainfall to model trends in extreme rainfall and to determine return levels.

Modelling of Extreme Rainfall using Extreme Value Distributions

This modelling is based on the time series of daily rainfall for Dodoma recorded from 1935 up to 2011 as described in section 3.1. In this section, we applied the theory of extreme value distributions presented in section 2.2. At the first, we used block maxima (BM) approach. Secondly, we considered model with stationary and non-stationary extreme value distribution parameters. We finally used the peak over a threshold (POT) approach to model the data of exceedances.

Fitting the model to the data by BM Approach.

One of the most important things to do before applying GEV model is to obtain $Y_t, n, t = 1, \ldots, m$, the maximum observations in m blocks of length n related to the period $[(t-1)n+1, t_n]$. For this, we need to choose a block of equal length n, and discard all values, remaining with only the maximum value in each block. First, we have extracted the block maxima of annual maximum from Dodoma daily rainfall. Therefore, for Dodoma rainfall $m = 77, n = 365, 366$ for annual blocks of maxima.

There is a need to know the risks of extreme events in agriculture, especially those that are damaging, such as heavy rainfall. Too much rainfall can affect the quality and productivity of crops. Figure 3.13 shows the extreme daily rainfall in a year. In 1964 we had the highest rainfall amount of 119.8*mm* whereas, the lowest extreme amount of rainfall was 29.0*mm*.

e) The figure below represents the annual block maxima for Dodoma daily rainfall

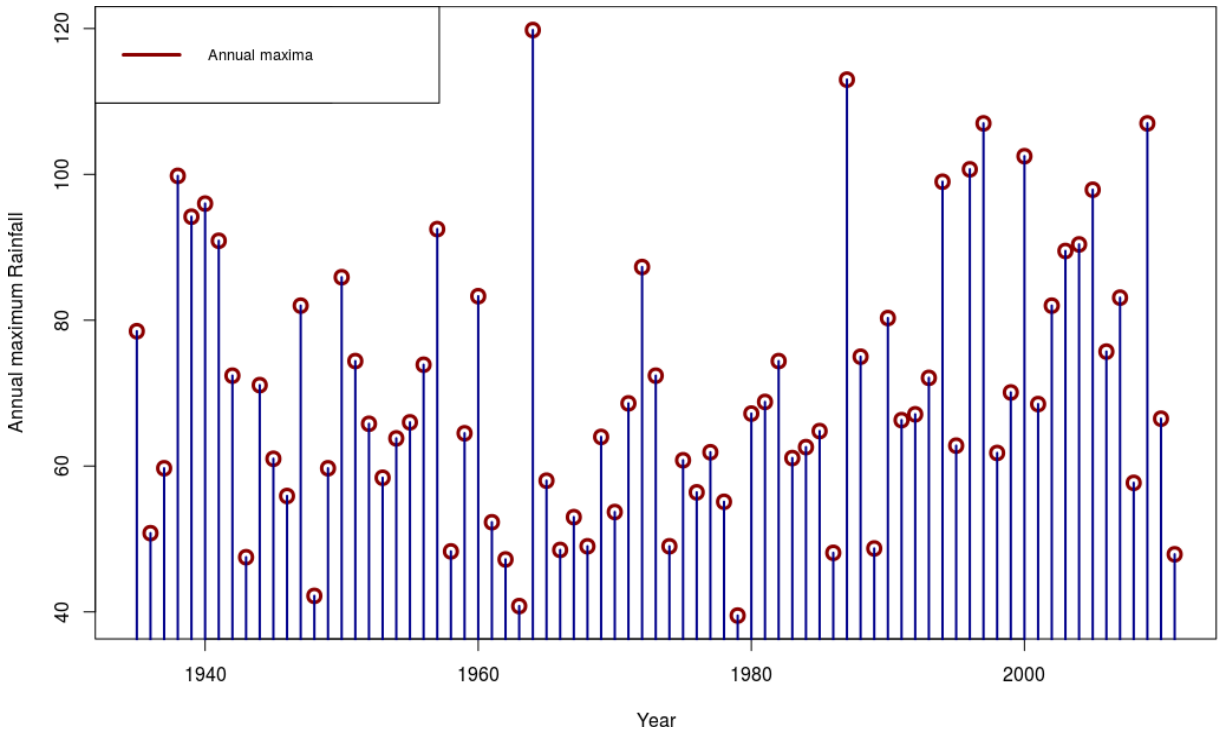


Figure 3.13: The extreme rainfall in Dodoma since 1935 to 2011

Fitting the data to a GEV model with stationary parameters

We assumed that the pattern of variation of extreme rainfall has stayed constant over the period 1935- 2011, so we modelled the daily Dodoma rainfall as independent observations from the GEV distribution. After filtering 77 blocks of maximum, we fitted the annual block maxima to GEV model and we estimated the parameters μ, σ and ξ by maximum likelihood method. The GEV log-likelihood of annual maxima $L(\varphi)$ with $\varphi = (\mu, \sigma, \xi)$ is given by

$$L(\varphi) = -77\log\sigma - (\xi^{-1}+1)\sum_{i=1}^{77}\log\left(1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right) - \sum_{i=1}^{77}\left(1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}.$$

(3.4.1)

If $\xi = 0$, we have Gumbel model. The log-likelihood of annual maxima L(φ) with $\varphi = (\mu, \sigma)$ is given by

$$L(\varphi) = -77\log\sigma - \sum_{i=1}^{77}\left(\frac{y_i-\mu}{\sigma}\right) - \sum_{i=1}^{77} exp\left(-\left(\frac{y_i-\mu}{\sigma}\right)\right)$$

(3.4.2)

Maximisation of the log-likelihood (Equation 3.4.1 and Equation 3.4.2) numerically using R, leads to the estimates presented in Table 3.5.

Table 3.5: GEV and Gumbel parameter estimates with their 99% confidence intervals

	GEV parameters		
Parameter	Estimate	Standard Error	CI (99%)
Shape($\hat{\xi}$)	-0.19	0.08	(−0.39, 0.01)
Location($\hat{\mu}$)	61.91	2.29	(56.00, 67.83)
Scale($\hat{\sigma}$)	18.23	1.61	(14.08, 22.37)
	Gumbel parameters		
Parameter	Estimate	Standard Error	CI (99%)
Location($\hat{\mu}$)	60.08	2.12	(54.62, 65.54)
Scale($\hat{\sigma}$)	17.69	1.48	(13.87, 21.51)

Maximization of Equation 3.4.1for Dodoma annual maxima rainfall data leads to the estimate $(\hat{\mu}, \hat{\sigma}, \hat{\xi}) =$ (61.91, 18.23, −0.19), for which the only one parameter ($\hat{\xi}$) was statistically insignificant. But all parameters were statistically significant $(\hat{\mu}, \hat{\sigma}) =$ (60.08, 17.69) when we maximised the Equation 3.4.2.

3 parameter (GEV) versus 2 parameter (Gumbel) model

The shape parameter ξ is the only parameter which governs the tail behaviour of the distribution. After fitting the GEV model to annual maxima data, ξ indicates which one of the three models best describe the Dodoma annual maxima rainfall. We used likelihood ratio-test to test GEV and Gumbel models with the following hypothesis

H_0: *Gumbel model* ($\xi =$ 0), H_1:*GEV model* (ξ $\neq$0).

The smaller the p-value, the stronger the evidence against H_0 provided by the data. Using function **lr.test()** in extRemes package for likelihood ratio-test, we got the p-value 0.024 at alpha ($\alpha = 0.01$), therefore, we failed to reject the null hypothesis. The zero belongs to the shape parameter's confidence interval ($0 \in 99\% CI(\xi)$) (see Table 3.5). Thus, we did not reject the null hypothesis H_0, which means that the suitable model for our Dodoma extreme rainfall belongs to Gumbelmodel.

Diagnostic plots of GEV and Gumbel model

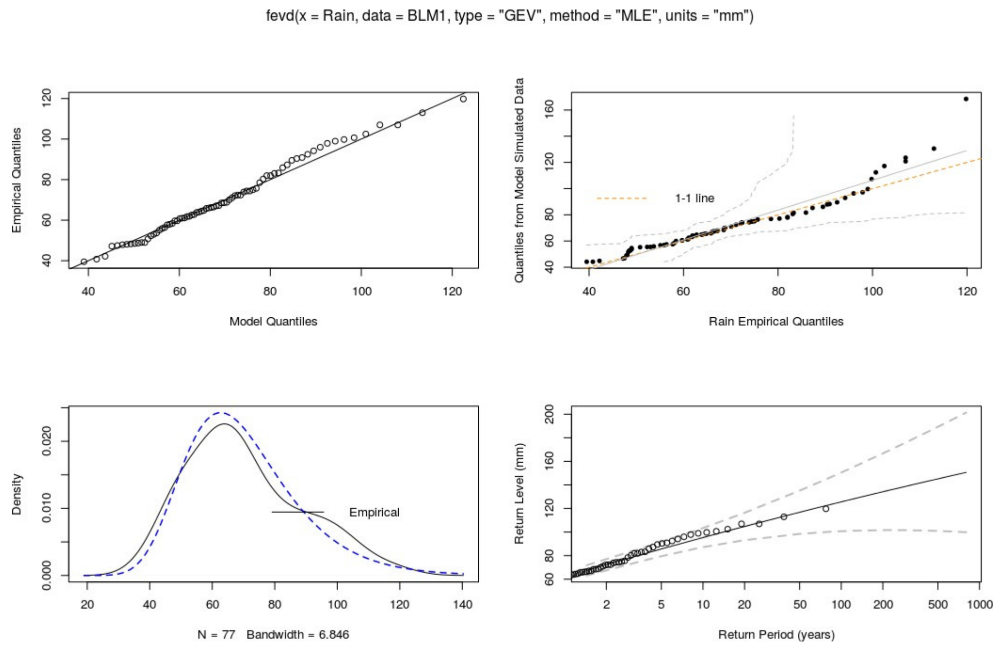


Figure 3.14: GEV model

On top, there are two plots; probability and quantile plots and at the bottom we have plot of the fitted GEV density superimposed onto the empirical density of the actual data (bottom left) and retun level plot.

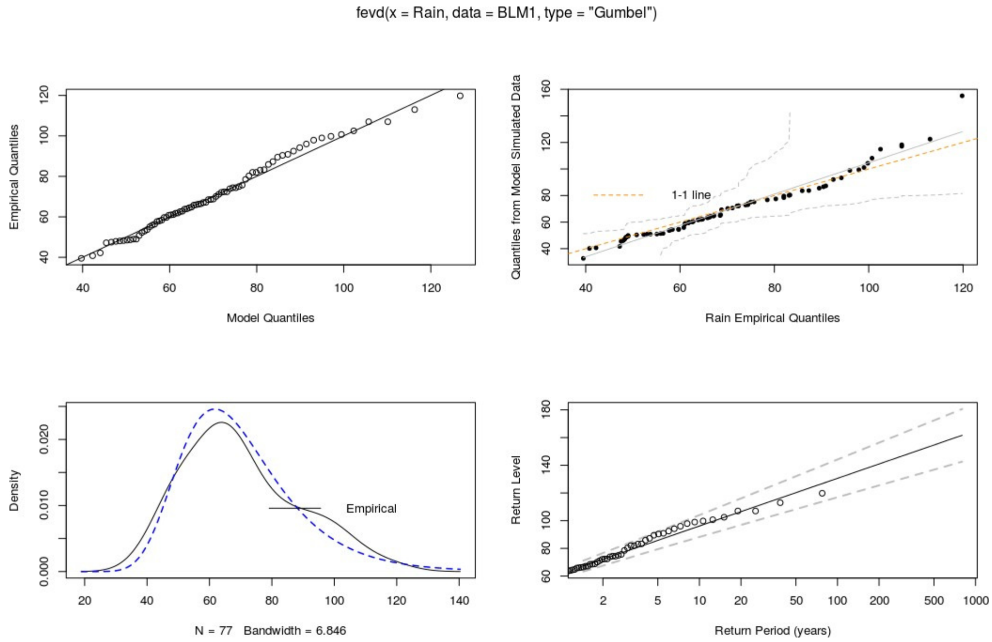


Figure 3.15: Gumbel model

The above plots are four diagnostic plots; probability plot, quantile plot, return level plot and density plot of Gumbel model.

Modelling a linear trend in extreme rainfall using Gumbel with non-stationary parameters

After finding that Gumbel model is the best fit of Dodoma maximum rainfall, we modelled linear trend in extreme rainfall using Gumbel model. In the context of environmental processes, non-stationarity is often apparent because of seasonal effects, perhaps due to different climate patterns in different months, or in the form of trends, possibly due to long-term climate changes. Due to climate change, the trend in frequency and intensity of extreme weather events occurs through time. To model change in extreme rainfall, extreme value distributions with non-stationary parameters could be used. Non-stationarity can be expressed in terms of location parameter with trend. Thus, we used Gumbel with two parameters as follow:

$$\mu_t = \mu_0 + \mu_{1t},\ t = 1, 2, ..\ \sigma_t = \sigma.$$

The classical Gumbel; $Gu(x, \mu, \sigma)$ model assumes that the two parameters of location and scale are time independent (stationary parameters). However, if trends are detected in the data sample, the non-stationarity case where parameters are no longer constants but expressed as covariates (e.g.time), should be considered. To study linear trend checking whether there exist a trend in change of extreme rainfall, two models were considered; stationary model1 (classical Gumbel), and non stationary model2 (Gumbel with time as covariates).

Model 1 *without trend:* μ, σ *are constants*

Model 2 *with trend:* $\mu_t = \mu_0 + \mu_{1t},$ σ *is constant,* where t refers to units of the selection period (for example, $t = 1$ corresponds to 1935 for Dodoma daily rainfall data).

We fitted Gumbel without trend in model 1 to compare with model 2 including trend. The second model is Gumbel fitted with linear trend in location parameter. The detail result of each model is shown below:

Table 3.6: Gumbel parameter estimates for Model1 and Model2 with their 99% confidence intervals

	Model1		
Parameter	Estimate	Standard Error	CI (99%)
Location($\hat{\mu}$)	60.1	2.1	(54.6, 65.5)
Scale($\hat{\sigma}$)	17.7	1.5	(13.9, 21.5)
	Model2		
Parameter	Estimate	Standard Error	CI (99%)
Location($\hat{\mu}_0$)	61.4	201.8	(−458.3, 581.2)
Location($\hat{\mu}_1$)	-0.001	0.1	(−0.3, 0.3)
Scale($\hat{\sigma}$)	17.7	1.5	(13.7, 21.6)

Fitting the model without a trend in location parameter, we got that all two parameters are statistically significant at 99 %. The second model is Gumbel fitted with linear trend in only location parameter and we got only one parameter statistically significant ($\hat{\sigma}$) but location parameters were not.

Writing a location parameter with linear trend $\mu_t = 61.4 - 0.001t$, where t is an index for year, with $t = 1$ corresponding to 1935. The parameter $\hat{\mu}_1 = -0.001$ corresponds to the annual rate of change in yearly maximum rainfall in Dodoma. However, all estimates of location parameter in model2 were not statistically significant at 99%. We used the likelihood-ratio test to test two models. We got a big p-value (≈ 1) indicating that the stationary model (Model1) should be accepted. This implies there is no evidence of a linear trend in location parameter.

f) Annual maximum rainfall with a fitted line of a linear trend in location parameter

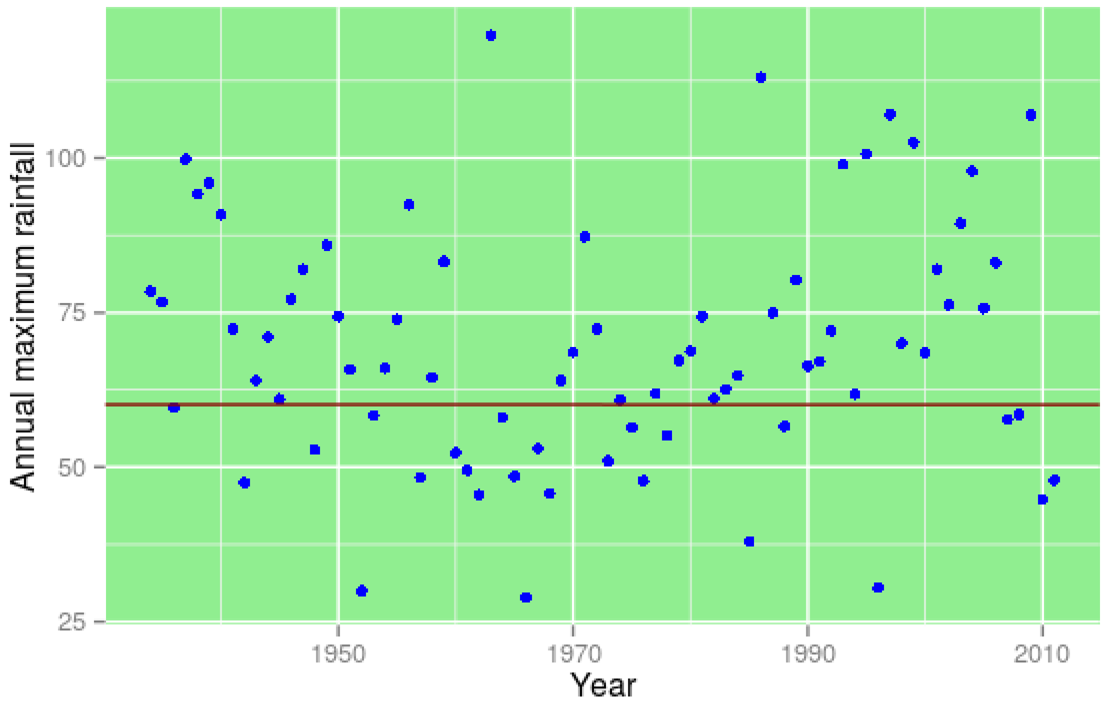


Figure 3.16: Fitted estimates for μ in linear trend Gumbel model of Dodoma annual maximum rainfall. The red line represents location parameter with a linear trend $\mu = 61.4 - 0.001t$.

1.1.1 Remark. The location parameter is analogous to the mean of a normal distribution, so increase in μ uniformly shifts the distribution to higher values, increasing all extremes equally. Whereas σ and ξ

determine the rate at which the magnitude of extremes alters with rarity (Brown et al., 2008).

g) Diagnostic plots for Gumbel with and without trend model

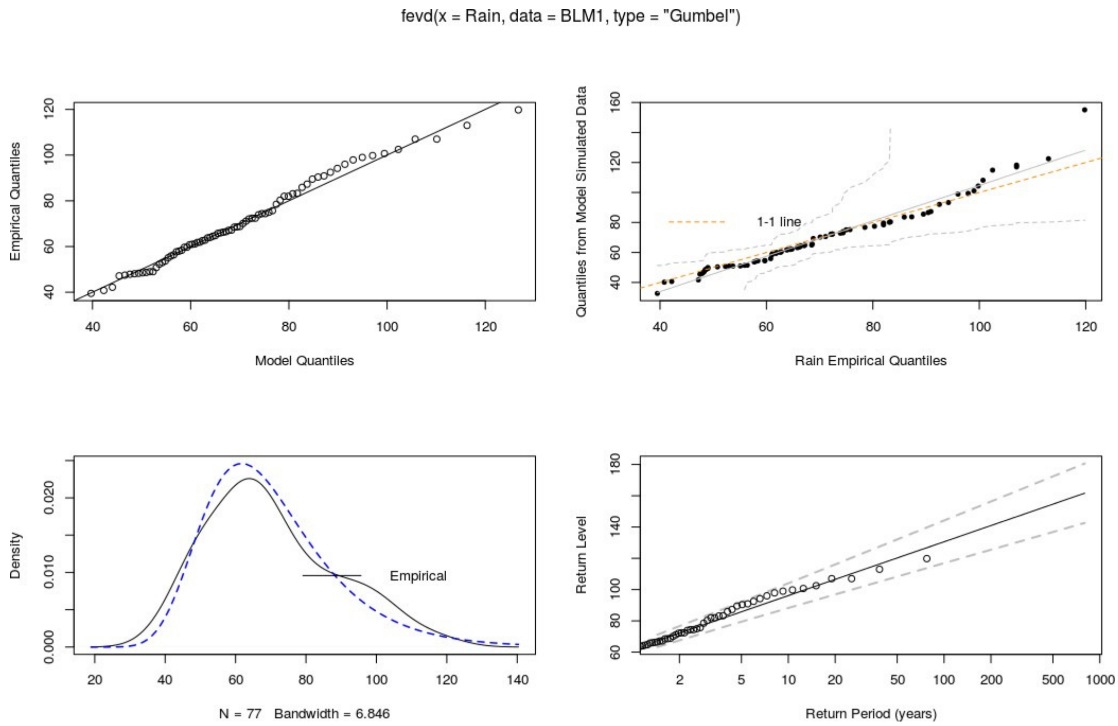


Figure 3.17: Gumbel without any trend.

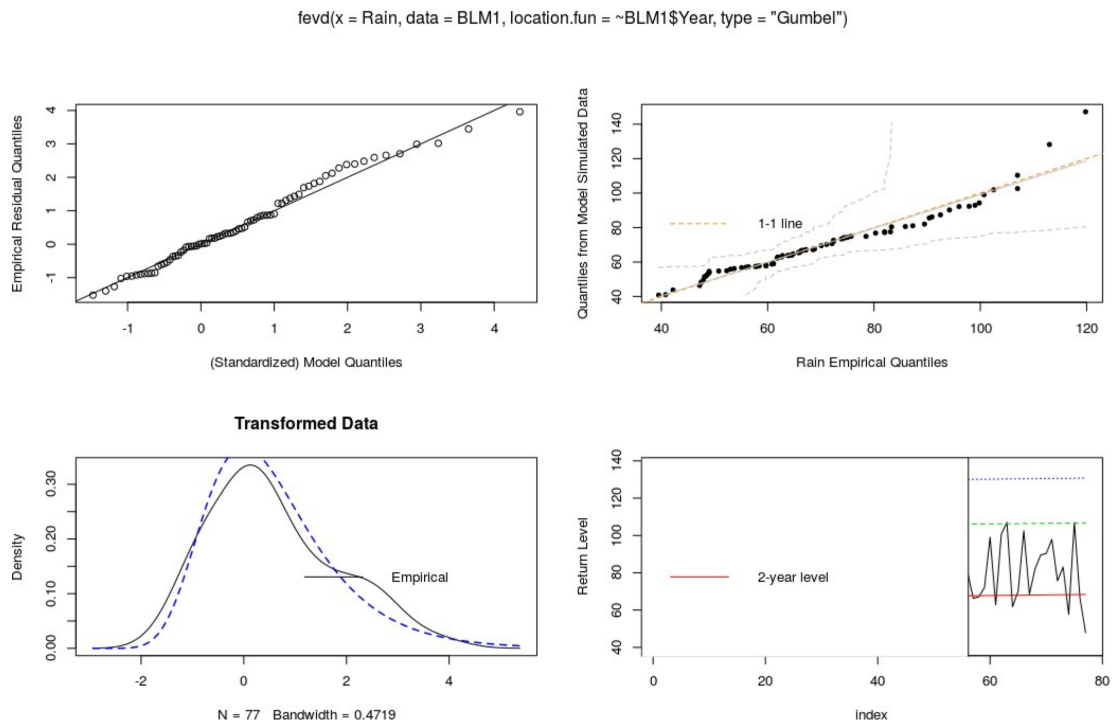


Figure 3.18: Gumbel with linear trend in location parameter.

Return levels and their $(1-\alpha)\%$ confidence limits

After estimating the Gumbel parameters, we estimated the return levels of extreme rainfall in Dodoma and we extrapolated to obtain estimates of return levels beyond the end of the data we have. Under the ideal of stationarity, the return level calculated from one period of the data should be approximately the same value if it was calculated from any other period of the same data. However, this is not the case if climate is changing. We were able to predict the estimate of the daily rainfall we would expect to see in Dodoma,

- ◊ Once in $T = 2$ years,
- ◊ Once in $T = 5$ years,
- ◊ Once in $T = 10$ years, so on.

We used our fitted stationary Gumbel model to extrapolate beyond the range of our data to estimate such return levels. The results are presented in Table 3.7 below.

Table 3.7: Gumbel return level estimates with their 99% confidence intervals

Return period T	T-year return level in *mm*	
	Estimated return level ($\hat{x}_T$) in *mm*	CI (99%)
2-year return level	66.6	(60.5,72.6)
5-year return level	86.6	(77.5, 95.7)
10-year return level	99.9	(88.3, 111.5)
20-year return level	112.6	(98.5, 126.7)
50-year return level	129.1	(111.6, 146.6)
100-year return level	141.5	(121.4,161.5)

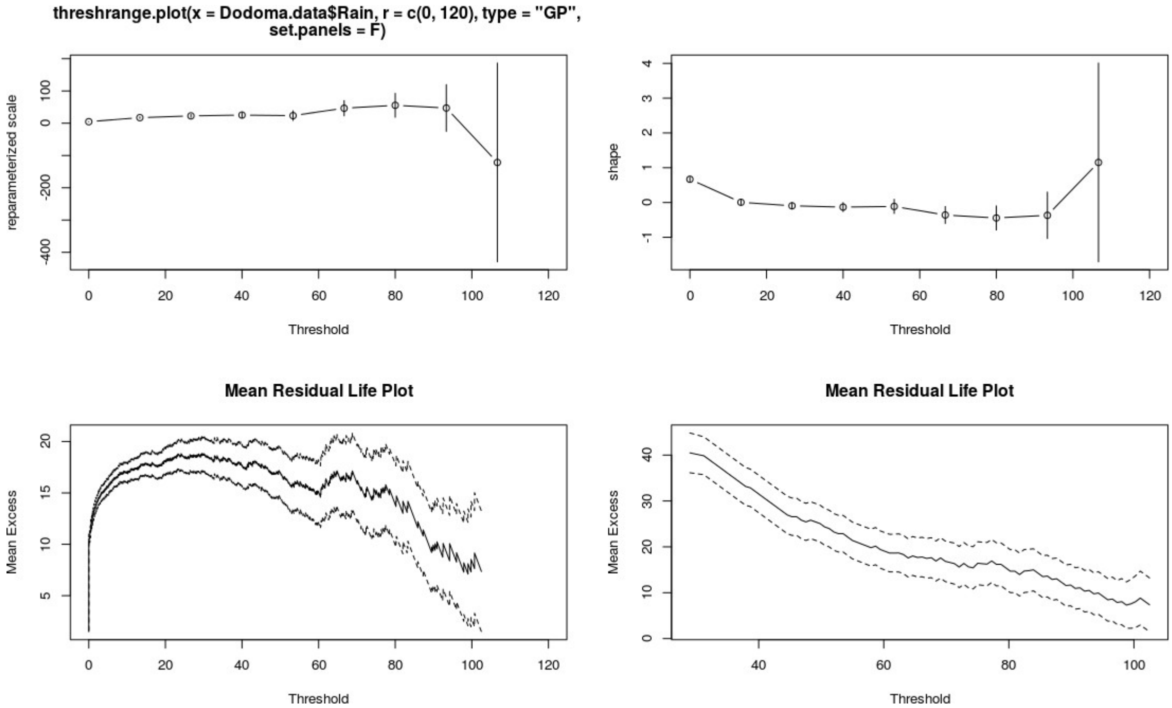
In this table above, the 5−year return level, 86.6, is the level extreme rainfall is expected to occur once in a period of 5 years. We would say that extreme rainfall of 86.6*mm* in Dodoma has 20% chance of being exceeded in any one year. According to estimated return levels in Table 3.7, there is a probability of 1% in Dodoma extreme rainfall to exceed 141.5*mm* in any one year. The results presented in the above Table 3.7 were obtained under a stationary Gumbel model.

1.1.2 *Fitting the model to the data by POT Approach .*

After BM approach, we turned to another alternative approach to the extreme value statistics based on exceedances over a threshold. The basic idea is to pick a high threshold u and to study all the exceedances of u. Those selected exceedances are said to follow the generalised Pareto distribution (*Check* : *Equation* 2.2.9). However, the main challenge of this approach is the selection of proper threshold. In subsection 2.2.6, we discussed POT approach and in this subsection we used the Dodoma daily rainfall data for the application.

The two plots were used for this selection of threshold: the mean residual life plot and the threshold range plot for parameters. We finally chose the best indicated threshold by two plots to estimate the GPD parameters.

Figure 3.19: **The threshold selection plots**



The above plots were used before making a final decision. We selected a threshold such that the mean residue life plot is approximately linear above the selected proper threshold u_0. The 30*mm* was found to be a reasonable threshold.

h) Fitting the Dodoma daily rainfall data to a GP model with stationary parameters

After selecting the threshold, the Dodoma daily rainfall data were fitted to GP with a threshold of 30*mm* and we estimated GP parameters using MLE. The results are represented in the table below.

Table 3.8: Parameter estimates with their 99.5% confidence intervals of the GP fitted to the daily Dodoma rainfall exceeding the threshold $u_0 = 30mm$.

	Threshold $u_o = 30mm$		
Parameter	Estimate	Standard Error	CI (99.5%)
Shape($\hat{\xi}$)	-0.13	0.05	(−0.26, 0.01)
Scale($\hat{\sigma}$)	21.24	1.47	(17.11, 25.37)

After estimating the generalised Pareto model parameters, we tested the shape parameter to know the best model of exceedances.

$$H_0 : \textit{exponential model } (\xi = 0),$$

$$H_1 : \textit{Beta or Paretomodel } (\xi \neq 0).$$

A likelihood ratio test gave us a big p-value ≈ 0.02 ($\alpha = 0.005$) for two models (H_0 and H_1). Consequently, this implies there is no evidence of rejecting the null hypothesis. Then, the exponential model is the appropriate model for the data of exceedances.

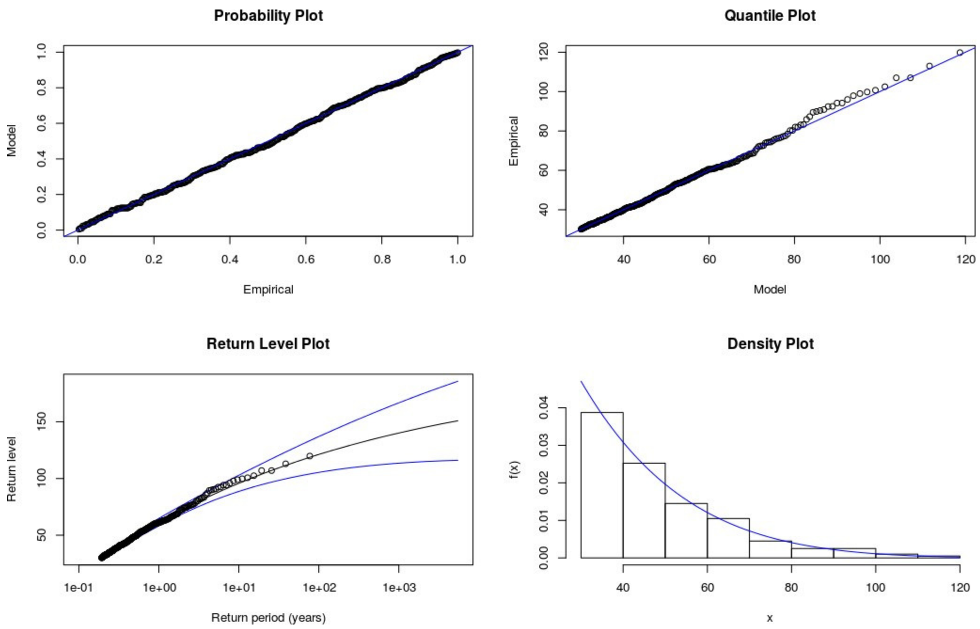


Figure 3.20: The Diagnostic plots from the GP fitted to Dodoma daily rainfall. Quantile-quantile plot (top right), quantiles from a sample drawn from the fitted GP against the empirical data quantiles (top left), density plots of empirical data and fitted GP (bottom right), and return level plot with pointwise normal approximation confidence intervals (bottom left).

We have already estimated the GP parameters, therefore we can estimate the return levels by using Equation 2.2.13. Table 3.9describes the return level estimates with their 99.5% CI at different return periods of daily exceedances over 30mm.

Table 3.9: GP return level estimates with their 99.5% confidence intervals

	T-year return level in *mm*	
Return Period T	Estimated Return level	CI (99.5%)
2	72.9	(66.4,79.5)
5	86.5	(77.87, 95.3)
10	95.8	(84.6, 106.9)
20	104.3	(89.9, 118.6)
50	114.4	(95.0, 133.676)
100	121.2	(97.7,144.8)

2. *Conclusion and Recommendations*

2.1 *Conclusion*

In this essay we used three approaches to analyse and model rainfall extremes in Dodoma. Using 6 rainfall extremes indices, we analysed trends in change in extreme rainfall. We used the least square method to estimate the parameters(slope and intercept) of a linear regression line. All estimated parameters were statistically insignificant at 0.01 level of significance. Then, there was no statistical evidence of the linear change in rainfall extremes in Dodoma.

Apart from the rainfall extremes indices to analyse change in extreme rainfall over the time, this essay used the annual block maxima approach, a method which fitted GEV model to the maximum rainfall. Using this approach we extracted the sample data and we fitted it to GEV model. After the likelihood ratio test of GEV model and Gumbel model, the Gumbel model was found to be appropriate model to describe the annual maximum Dodoma rainfall. The Gumbel model with a linear trend was not showing any statistical evidence of a linear trend in Dodoma rainfall extremes. However, more research is needed especially about cyclic variations because of seasonality.

We estimated the return levels of extreme rainfall under Gumbel model with stationary parameters. But we did not estimate the return levels for extreme rainfall under a changing climate. We were able to predict the estimate of the daily rainfall we would expect to see in Dodoma once in $T = 2, 5, 10, 20, 50$ and 100 years. However, this essay did not extend the concept of return level to non-stationary climate.

As block maxima approach ignores other important extreme rainfall data, especially those greater than the annual maximum. The peak over threshold (POT) approach were also used, where we focused on the distribution of values that have exceeded a threshold of $30mm$. Using the likelihood ratio test, we tested the shape parameter and exponential model was found to be the extreme value model which can describe Dodoma rainfall exceeding $30mm$. We have modelled exceedances data under the stationary climate only. For the GPD, it is not always clear how to interpret some parameters, such as return levels because the rate of exceeding the threshold may vary seasonally. The choice of an appropriated threshold in this approach remains a challenge.

As shown with Dodoma daily rainfall data analysis, BM and POT approaches can be used for stationary and non-stationary extreme data. But still there is work to be done on the general theory to be extended for non-stationary series especially in describing the trend variation in the data of exceedances using the fitted GP parameters.

2.2 *Recommendations*

In this essay, we used **extRemes** package in R software. If this package can be incorparated in R-Instat software,
this could be easy for some people especially those who want to analyse extremes in R-Instat software. Analysis and modelling of climate extremes need reliable and long period data. However, the data may have
some missing values and this can provide the wrong predictions. It is always difficult to find a well detailed historic climate data. There is often a percentage of data missing, which if not well handled, can
give wrong analysis.Thus, better ways of handling missing information should be considered. In this essay,
we used only Dodoma station data. However, other stations in Tanzania or elsewhere can be used to study the distribution and change in extreme rainfall using the same theory applied in this essay.

2.3 *Further work*

More research is needed to learn which model would be preferable to convey uncertainty about extreme events under climate change. To have more experience in this field of climate extremes, this essay will be extended as my future work to *modelling non-stationary extreme rainfall and temperatures in Rwanda using extreme value distributions*. Apart from considering a linear trend, in this work the cyclic variations of extremes will be taken into account.

REFERENCES

[1] L. V. Alexander, X. Zhang, T. C. Peterson, J. Caesar, B. Gleason, A. M. G. Klein Tank, M. Haylock,D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K. Rupa Kumar, J. Revadekar, G. Griffiths,
L. Vincent, D. B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rus ticucci, and J. L. Vazquez-Aguirre. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, 111(D5): n/a–n/a, 2006.

[2] S. J. Brown, J. Caesar, and C. A. T. Ferro. Global changes in extreme daily temperature since 1950. *Journal of Geophysical Research: Atmospheres*, 113(D5):n/a–n/a, 2008. ISSN 2156-2202.

[3] D. Chikobvu and R.Chifurira. Modelling of extreme minimum rainfall using generalised extreme value distribution for zimbabwe. September, 2015.

[4] S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[5] Earthquake in Tanzania. Earth changes. Website, https://www.sott.net/article/ 327978-Update-At-least-11-dead-192-injured-following-earthquake-in-Tanzania, Accessed 3 May 2017.

[6] Floodlist (22 March 2015). Tanzania floods – 5 killed in dar es salaam after 91mm of rain in 24 hours. Website,http://floodlist.com/africa/tanzania-floods-5-killed-dar-es-salaam, Accessed 16 May 2017.

[7] Floods in Tanzania (January 2015). Poor distribution of rainfall. Website,http://floodlist.com/africa/ poor-distribution-rainfall-leads-floods-droughts-southern-africa, Accessed 3 May 2017.

[8] E. Gilleland and R. W. Katz. extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, 72(8):1–39, 2016.

[9] H. Hasan, N. Salam, and M. B. Adam. Modeling extreme temperature in malaysia using generalized extreme value distribution. *World Academy of Science and Technology*, 78:435–441, 2013.

[10] Heavy rainfall in Tanzania(March 2015). Tanzanian die in heavy flood. Website,http://ejuratalks. blogspot.com/2015/03/tanzanian-die-in-heavy-flood.html, Accessed 3 May 2017.

[11] International Journal of Science and Research (IJSR). Non homogeneous poisson process mod- elling of seasonal extreme rainfall events in tanzania. Website, https://www.ijsr.net/archive/v5i10/ ART20162322.pdf, Accessed 4 May 2017.

[12] S. Jeon. Data analysis in extreme value theory: Non-stationary case. 2009.

[13] B. K. K. E. J. M. H. D. Mboera, L. E. G.; Mayala. Impact of climate change on human health and health systems in tanzania: a review. 2011.

[14] P. L. Myriam Charras-Garrido. Extreme value analysis : an introduction. *Journal de la Societe Francaise de Statistique, Societe Francaise de Statistique et Societe Mathematique de France*, 2013.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URLhttp://www.R-project.org/.

[16] Rainfall and forecasts(14 January 2016). Tanzania – 400 displaced and 1 dead after floods in mwanza and dodoma – further flood warnings issued. Website,http://floodlist.com/africa/ tanzania-floods-mwanza-dodoma-further-flood-warnings-issued, Accessed 17 May 2017.

[17] R.-D. Reiss and M. Thomas. Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields. 2007.

[18] I. P. Santikayasa. Climate change indices. Technical report, Water Engineering and Management (WEM) School of Engineering and Technology (SET), Asian institute of Technology (AIT), 2015.

[19] Source: arcjournals(2016). Modelling of extreme maximum rainfall using extreme value theory for tanzania. Website,https://www.arcjournals.org/pdfs/ijsimr/v4-i3/7.pdf, Accessed 5 May 2017.

[20] T. S. Stephenson, L. A. Vincent, T. Allen, C. J. Van Meerbeeck, N. McLean, T. C. Peterson, M. A. Taylor, A. P. Aaron-Morrison, T. Auguste, D. Bernard, J. R. I. Boekhoudt, R. C. Blenman, G. C. Braithwaite, G. Brown, M. Butler, C. J. M. Cumberbatch, S. Etienne-Leblanc, D. E. Lake, D. E. Martin, J. L. McDonald, M. Ozoria Zaruela, A. O. Porter, M. Santana Ramirez, G. A. Tamar, B. A. Roberts, S. Sallons Mitro, A. Shaw, J. M. Spence, A. Winter, and A. R. Trotman. Changes in extreme temperature and precipitation in the caribbean region, 1961–2010.