Statistical Modelling of Annual Maximum Rainfall for Sokoto Basin using Extreme Value Theory

Popoola, A.M., Adegbola, A.A., Olaniyan, O.S. & Ayinde, R.B.

Department of Civil Engineering, Ladoke Akintola University of Technology, Ogbomoso, Oyo State Nigeria.

Email: osolaniyan@lautech.edu.ng, malabipopoola@gmail.com.

Abstract - Water is one of the most important natural resources with diverse uses in agricultural, domestic and industrial sectors. Rainfall should be monitored overtime as significant changes in its intensity may lead to flood or drought. Extreme value theory gives us a statistical approach for the occurrences and the magnitude of these extreme cases that are beyond the scope of available data. In this study, statistical modelling of annual maximum rainfall in Sokoto Basin was done to predict extreme annual maximum rainfall trends and compute the return period.

Rainfall data was obtained from Sokoto Rima River Basin Development Authority for the period 1915-2018. The minimum, maximum, standard skewness and kurtosis of annual precipitation was determined. Parameter estimation methods (location, scale and shape) and Model selection criteria [negative log likelihood, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion)] were estimated. Extreme value distributions to mean annual rainfall were selected.

The minimum, maximum, standard skewness and kurtosis of the data were 342.2mm, 1477.5mm, 3.8217 and 3.7474, respectively. The location, scale and shape of the model ranged 140.94-636.89, 55.76-201.51 and -0.07-0.15, respectively. The negative log likelihood, AIC and BIC varied from 80.47-209.09, 166.95-1383.13 and 171.53-1391.06, respectively. The return levels were estimated for 5, 10, 20, 30, 50 and 100 years. The returned period estimated were 867.662, 974.6115, 1072.2101, 1126.2364, 1191.6743 and 1278.3797mm, respectively.

Generalized Pareto Distributions Model is adequate to predict extreme annual maximum rainfall trend in Sokoto Basin based. The return levels revealed that rainfall are increasing and may cause flooding in the nearest future.

Index Terms: Extreme Value Theory, Rainfall, R-package, Sokoto Basin.

1 INTRODUCTION

Water is a precious resources that exist naturally on the planet earth. Precipitation and temperature extremes are considered to be the most important climate events and have been extensively explored over the past several decades, according to [30], [32]. Water is one of the most important natural resources with diverse uses in agricultural, domestic and industrial sectors. Rainfall is the water falling in drops from vapor condensed in the atmosphere. Low or excessive rainfall can lead to water scarcity or floods, which may cause great damage to human health, the environment, and property, leading to massive social and economic losses. Rainfall has a substantial influence on agriculture, food security, infrastructure development, water quality and the economy [7]. Although knowledge of rainfall patterns over an area may be used for strategic economic planning, it is one of the most difficult meteorological parameters to study because of lack of reliable data and large variations of rainfall in space, and time [7].

Rainfall is a climate parameter that affects the way and manner man lives. It affects every facet of the ecological system, flora and fauna inclusive. Globally, lots of studies have been conducted on rainfall [17]. Rainfall and temperature are important climatic factors for crop production and affects humans' basic needs such as health, shelter, food and water in Sokoto Basin. However extreme rainfall causes flooding which can lead to loss of lives, properties among others. Extreme temperature also causes drought, hot and cold spells among others which have adverse implications on human beings and agriculture. Human losses from flooding are projected to increase by 70-80% if the global mean temperature increases above 1.5°C from the pre-industrial level [13]. It is therefore important to know about occurrences of such extreme events and their chances of occurring.

Most climate models show that, extreme rainfall events increase in response to increase in greenhouse gases. The greenhouse gases are not the sole agents responsible for extreme rainfall events. However, they

amplify the extreme rainfall induced by other causes. However, those extreme rainfall events will increase as global warming proceeds unabated. Hence, it is absolutely essential for us to be prepared to tackle more extreme rainfall events in this region in the future.

Extreme value theory gives us a statistical approach for the occurrences and the magnitude of these extreme cases or rare event that are beyond the scope of available data [23]. Therefore, developing methods that can suitably predict meteorological events is extremely valuable for both meteorologists and civil engineers in light of global climate change.

The main motivation for statistical modelling of weather extremes is reliability: the ability of a system or component to withstand the weather conditions for the whole of its working life. For instance, for civil engineers to determine how high the wall of a reservoir dam should be, they need to estimate what the heaviest (maximum) rainfall amount will be over some specified future time period. This can only be achieved by using historical data and fitting an appropriate statistical model for them. That is, a model for risk is developed by selecting a probability distribution that gives the best fit to the data at hand. Statistical models for extreme values are based on an asymptotic theory called extreme value theory [29].

2 Materials and Methodology

The Study Area

The study area is Sokoto, north-western Nigeria. The area is found between latitudes 10^0N and $13^0$58'N; and longitudes $4^0$8'E and $6^0$54'E. The area so defined covers a land area of approximately 62,000km^2. It lies to the north-west of Nigeria and shares its borders with Niger Republic to the north, Zamfara State to the South-East, Kebbi State to the South-West, and Benin Republic to the west. The southern boundary is arbitrarily defined by the Sudan savanna. Like the rest of West Africa, the climate of the region is controlled largely by the two dominant air masses affecting the sub-region. These are the dry, dusty, tropical-continental (cT) air mass (which originates from the Sahara desert), and the warm, tropical-maritime (mT) air mass (which originates from the Atlantic Ocean). The influence of both air masses on the region is determined largely by the movement of the Inter-Tropical Convergence Zone (ITCZ), a zone representing the surface demarcation between the two air masses. The interplay of these two air masses gives rise to two distinct seasons within the sub-region. The wet season is associated with the tropical maritime air mass, while the dry season is a product of the tropical continental air mass. The influence and intensity of the wet season decreases from the West African coast northwards. Therefore, precipitation in the whole sub-region of West Africa depends on thunderstorm activity which occurs along disturbance lines called "line squalls" and, about 80 percent of the total annual rainfall for most places is associated with line squall activities which are prevalent between June and September [1].

Source of Data

The data used for this study consist of annual maximum rainfall for Sokoto Basin extracted from the mean historical monthly rainfall for Sokoto during the time period 1915-2018. The data set was produced by the Sokoto Rima River Basin Development Authority. The choice of the time period is based on the quality and availability of data. The data produced by the monitoring team at Goronyo dam are of very high quality with no missing values. The data is required for model setup, model calibration and validation purposes.

(A) The Classical Extreme Value Distributions

The basic construction of the generalized extreme value distribution, according to [10], is as follows: Assume that the random variable X_i measures a daily, weekly, monthly or yearly quantity such as maximum temperature. Let $M_n = \max\{X_1,\ldots,X_n\}$ where $X_1,\ldots,X_n$ is a sequence of independent random variables and M_n corresponds to the maximum of n observations with a common distribution function F then, the distribution of M_n can be determined, theoretically, for all values of n. This is given by

$$\Pr\{Mn \leq z\} = \Pr\{X_1 \leq z, \ldots, X_n \leq z\} \quad (1)$$

$= \Pr\{X_1 \leq z\} \times \ldots \times \Pr\{X_n \leq z\}$

$= \{F(z)^n\}$

According to [10], this formulation is not helpful in practice because very small discrepancies in the estimates of F from observed data which are substituted into (1) can lead to substantial discrepancies for F^n. If the population distribution function F were known, the distribution function of M_n could be determined exactly. However, the population distribution function is often unknown. Therefore, the distribution of M_n is approximated by modeling F^n by means of asymptotic theory of M_n. This, however, has a disadvantage that as $n \rightarrow \infty$, the distribution of M_n degenerates to a point mass at the upper end point of F. In order to avert this problem, a linear renormalization of M_n is allowed which is analogous to the central limit theorem. The linear renormalization of M_n is given by the following formula:

$M_n = \left(\frac{M_n - b_n}{a_n} \leq x\right)$, for sequences $a_n > 0$ and b_n (2)

When suitable a_n and b_n are chosen the distribution of M_n is stabilized. This is formulated in theorem called the extremal types theorem by [14].

Extremal types theorem: If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$P_r\{\left(\frac{M_n - b_n}{a_n} \leq z\right) \rightarrow G(z)$ *where Ga non-degenerate distribution function, then G belongs to one of the following families:*

I: $G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}$, if $-\infty < x < \infty$ (3)

II: $G(x) = \begin{cases} 0 \\ \exp\left\{-\left(\frac{x-b}{a}\right) - \alpha\right\}, \end{cases}$ (4)

III: $G(x) = \begin{cases} exp \\ 1 \end{cases}\left\{-\left[-\left(\frac{x-b}{a}\right)\alpha\right]\right\}$ (5)

For parameters $\sigma > 0$, $\mu \in R$, and $\varepsilon > 0$. Here, ε is the shape parameter, σ is the scale parameter and μ is the location parameter. The three distributions are referred to as the extreme value distributions. Equations (3), (4) and (5) are known respectively as the Gumbel, Frechet and Weibull families of distributions.

The three types of distributions stated above have distinct forms of behavior corresponding to different forms of tail behavior for the distribution function F of the X_i. The question, therefore, becomes which of the three distributions should be used when analyzing a set of data. To solve this problem the three models were reformulated into a single model by [10], [18]. The reformulation resulted in a model called the Generalized Extreme Value (GEV) family of distributions. This is given by the following equation:

$$G(z) = exp\left\{-\left[1 + \varepsilon\left(\frac{z-\mu}{\sigma}^{-1/\varepsilon}\right)\right]\right\} \quad (6)$$

This is defined on the set $\left[z: 1 + \varepsilon\left(\frac{z-\mu}{\sigma}\right)\sigma > 0\right]$ with ε, σ and μ representing respectively the shape, scale and location parameters. The parameters also satisfy $-\infty < \mu < \infty$, $-\infty < \varepsilon < \infty$ and $\sigma > 0$. When $\varepsilon > 0$ the distribution is known as Frechet distribution and it has a fat tail. The larger the shape parameter is, the more fat-tailed the distribution. Also if $\varepsilon < 0$, the distribution is the Weibull distribution and finally a Gumbel distribution when $\varepsilon = 0$.

According to [10], the unification of the three families of extreme value distributions into a single family distribution simplifies statistical implementation. The type II and type III classes of extreme value distributions given by equations (4) and (5) correspond respectively to the cases where $\varepsilon > 0$ and $\varepsilon < 0$ in this parameterization. The type I class indicated by equation (3) corresponds to the Gumbel distribution family for $\varepsilon = 0$.

Fundamental to the study of extremes is the concept of max-stability. According to [10], a distribution is max-stable if and only if it is of the same type as an extreme value distribution. The class of max-stable distributions characterized the family of extreme value distributions.

(a) The Peaks-Over-Threshold Model

This approach contrasts with the block maxima model through the characterization of an observation as extreme if it exceeds a given high threshold. It has the advantage over the block maxima model in that data are more efficiently used. In the block maxima model data are wasted if one block happens to contain more extreme events than other blocks [3], [4], [10]. The events exceeding some chosen high threshold have an approximate Generalized Pareto distribution (GPD) that governs the intensity of the events.

By the Peaks-Over-Threshold (POT) method, if $X_1, X_2,\ldots X_n$ is a sequence of independent and identically distributed random variables with a marginal distribution function F, then extreme events are those of the X_i that exceed some high threshold denoted as u. If an arbitrary term in the X_i sequence is X, then a description of the stochastic behavior of extreme events is given by the conditional probability

$$\Pr\{X > u + y/x > u\} = \frac{1-F(u+y)}{1-F(u)},\ y > 0 \qquad (7)$$

Again, just as in the block maxima approach, since the parent distribution of F is unknown, the distribution of threshold excesses as given in equation (7) are unknown. In this in practical applications, approximations that are broadly applicable for high values of threshold are sought similar to the distribution of maxima of long sequences in the GEV case described earlier. This is provided by a General Pareto Distribution (GPD). The GPD asymptotic model description is given by the following theorem:

Theorem 1: Let $X_1, X_2,\ldots$ be a sequence of independent random variables with a common distribution function F, and let M_n = max $\{X_1,\ldots X_n\}$. Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies the extremal types theorem so that for large n, Pr $\{M_n \leq z\} \approx G(z)$, where G(z) is the families of GEV distributions, then for large u the distribution function of (x-u), conditional on $x > u$ is approximately

$$H(y) = 1 - (1 + \varepsilon y/\sigma_u) \qquad (8)$$

defined on $\{y: y > 0 \text{ and } \left(1 + \frac{\varepsilon y}{\sigma u}\right) > 0\}$ where $\sigma_u = \sigma + \varepsilon(u + \mu)$, σ_u is the GPD scale parameter which is dependent on the threshold σ and ε are the corresponding scale and shape parameters of the GEV distribution and is the threshold excesses [10].

In modelling extreme events by the POT method, inference consists of fitting the generalized Pareto family distributions to the observed threshold of excesses, followed by verification of the threshold and extrapolation.

(b) Selection of Threshold in the POT Model

The issue of threshold selection is similar to that of selection of block size in the block maxima approach. The choice of the threshold is not straightforward and usually a compromise has to be found. According to [5] and [10], a high threshold value reduces the bias as this satisfies the convergence towards the extreme value theory but however increases the variance for the estimators of the parameters of the GPD, as there will be fewer data from which to estimate the parameters. A low threshold value on the other hand, results in the opposite i.e. a high bias but a low variance of the estimators, but there is more data with which to estimate the parameters. Consequently, various graphical techniques used for in selecting an appropriate threshold include mean excess plot, parameter stability plot and selection based on empirical quantiles.

(i) Mean Excess Plot

Davison and Smith [12] suggest this graphical method for the selection of the threshold.

The method is based on the mean of the GPD given by $E(Y) = \frac{\sigma}{1-\varepsilon}$. Suppose the GPD is valid as a model for the excesses of a threshold u_0 generated by a series $X_1,\ldots X_n$. By the mean formula, then, $E(X - u_0/x > u_0) = \frac{\sigma u_0}{1-\varepsilon}$ provided that $\varepsilon < 1$ and where σ_{u_0} is the GPD scale parameter for exceedances over threshold u_0. The threshold stability property of the GPD means that if the GPD is a valid model for excesses over some threshold u_0, then it is valid for excesses over all thresholds $u > u_0$. Thus, for all $u > u_0$, $E(X - u/X > u) = \frac{\sigma u}{1-\varepsilon} = \frac{\sigma u_0 + \varepsilon u}{1-\varepsilon}$ where $\sigma_u = \sigma + \varepsilon(u - \mu)$. Thus for all, $u > u_0$, $E(X - u/X > u)$ is a linear function of u. Also, $E(X - u/X > u)$ is the mean of excesses of the threshold u, and can be estimated by the sample mean of the threshold excesses. This leads to the mean residual life plot defined by the locus of points $\{(u, 1/n_u \sum_{i=1}^{n_u}(x_{(i)} - u));\ u < x_{max}\}$, where $x_{(1)},\ldots,x_{(nu)}$ consists of the n_u observations

that exceed u, and x_{max} is the largest of the X_i. If the GPD assumption is correct, then the plot should be linear with the intercept to be $\frac{\sigma u}{1-\varepsilon}$ and slope is $\frac{\varepsilon}{1-\varepsilon}$. The mean excess plot of the data can be used to distinguish between light-and heavy-tailed models. The plot of a heavy-tailed distribution shows an upward trend, a medium tail shows a horizontal line, and the plot is downward-sloped for light-tailed data.

(ii) Parameter Stability Plot

Another graphical method which is widely used to determine the threshold u is the parameter stability plot. The idea of this plot is that if the excesses of a high threshold u_0 follows a GPD with parameters ε and σ_{u0}, then for any threshold u such that $u > u_0$, the excesses still follow a GPD with shape parameter $\varepsilon_u = \varepsilon$ and scale parameter $\sigma_u = \sigma_{u0} + \varepsilon(u-u_0)$. Letting $\sigma^* = \sigma_{u0} - \varepsilon_u u$, this new parameterization does not depend on u any longer, given that u_0 is a reasonable threshold. The plot defined by the locus of points $\{(u, \sigma^*); u < x_{max}\}$ and $\{(u, \varepsilon_u); u < x_{max}\}$, where x_{max} is the maximum of the observations. Estimates of σ^* *and* ε_u are constant for all asymptotic approximations. The threshold is chosen at the value where the shape and scale parameters remain constant.

(iii) Empirical Quantiles

According to [10] and [26] the simplest way to select a threshold is to choose from the raw data at a specified empirical quantiles in the range of 90% to 97%. This procedure consists in choosing one of the sample points as a threshold. The choice is practically equivalent to estimation of the kth upper order statistic X_{n-k+1} from the ordered sequence $X_1,...X_n$. Frequently used is the 90% quantile, but this is inappropriate from a theoretical point of view [28].

(B) Tests for Stationarity

Time series data such as temperature and rainfall may be stationary or non-stationary. Stationary series are characterized by a statistical equilibrium around a constant mean level as well as a constant dispersion around that mean level [6]. Non-stationary series that lack mean stationarity have no mean attractor toward which the levels tend over time. According to [16], unstable and indefinitely growing variances inherent in non-stationary series not only complicate significance tests but also render forecasting problematic as well. According to [5], the Peaks-Over-Threshold method is only valid for extreme if we can assume stationarity of the data. Rainfall data are likely to be non-stationary as they have a tendency to show strong seasonal patterns. Rainfall may increase during the rainy season but stop during the dry season. Therefore, to prepare data that are likely to be non-stationary for any statistical modelling, it is important to subject the data to tests of stationarity to transform them to stationarity before analyzing them.

Various methods exists to transform the data before any analysis, if they are non-stationary, such as by taking the natural logarithm, by taking a difference or by taking residuals from a regression. There are objective tests which are conducted to determine the stationarity or otherwise of a time series. The objective test used include: graphical techniques such as Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) and unit root tests such as Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. In this study, ADF and KPSS tests were used to test for stationarity.

(i) The ADF Test

The Dickey-Fuller (DF) and Augmented Dickey-Fuller (ADF) tests have been used to check the stationarity and presence of unit root of a process. The DF test is considered valid for only AR (1) and when the residuals are not auto correlated. However, when there is higher order correlation or where there is autocorrelation in the residuals, the ADF test is employed to test for unit root. The test presumes that the errors are independent of one another. In other words they are distributed as white noise and homogenous.

There are three versions of ADF which can be used to test for the presence of unit roots. These are test for a unit root, test for a unit root with drift and test for unit root with drift and a deterministic time trend. These situations have their respective equations given by Equations

$$\Delta y_t = \gamma y_{t-1} + \sum^{p-1}_{i=1} \varphi_i y_{t-1} + \varepsilon_t \quad (9)$$

$$\Delta y_t = \beta_0 + \gamma y_{t-1} + \sum^{p-1}_{i=1} \varphi_i y_{t-1} + \varepsilon_t \quad (10)$$

$$\Delta y_t = \beta_0 + \gamma y_{t-1} + \sum^{p-1}_{i=1} \varphi_i y_{t-1} + \beta_1 t + \varepsilon_t \quad (11)$$

Where y_t is the time series being tested, β_0 is the drift term, t is a linear trend term, β_1 is the coefficient of the linear trend term, φ_i are parameters of the model, p - 1 is the number of lags which are added to ensure the model residuals are white noise. Also $\Delta y_t = y_t - y_{t-1}$.
The null and alternative hypotheses are respectively;
$H_0: \gamma = 0$ *(non-stationary)*
$H_1: \gamma < 0$ *(stationary)*
The test statistic is $\tau = \gamma / \sqrt{\text{var}(\gamma)}$
This value is compared to the corresponding critical value at different significant levels.

(ii) The KPSS Test

Kwiatkowski *et al.,* [20] proposed an alternative test in which y_t is assumed to be stationary under the null hypothesis. The test can be computed by firstly regressing the dependent variable y_t on a constant and a time trend variable t. KPSS assesses the null hypothesis that a time series is trend stationary against the alternative that it is a non-stationary unit root process. The test uses the structural model

$$y_t = c_t + \delta t + u_{1t} \quad (12)$$

Where $c_t = c_{t-1} + u_{2t}$, and δ is the trend coefficient, u_{1t} is a stationary process and u_{2t} is an independent and identically distributed process with mean zero and variance σ^2. The null hypotheses are respectively;

$$H_0: \sigma^2 = 0$$
$$H_1: \sigma^2 > 0$$

The null hypothesis implies that the random walk term is constant and acts as the model intercept and the alternative hypothesis introduces the unit root in the random walk. The test statistic is given as;

$$\text{KPSS} = \frac{\sum_{t=1}^{T} St2}{s2T2} \quad (13)$$

T is the sample size, s^2 is the estimate of the long-run variance and s_t sum of the errors.

(C) Model Selection Criteria

When there are two competing candidate models for a set of data, it important to subject them to test to see which of them better fits the data well. There are many measures that can be used for estimating how well the model fits the data. Two of these models employed in this study are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The AIC is a measure which uses the log-likelihood but adds a penalizing term associated with the number of variables. The fit of a model can be improved by adding more variables. As a result, the AIC tries to balance the goodness-of-fit versus the inclusion of variables in the model. The AIC is given as;

$$AIC = 2k - 2ln(L) \quad (14)$$

Where, k is the number of unknown parameters included in the model, lnL is the log-likelihood function of the model. If the model errors are normally and independently distributed and n the number of observations with the residual sum of squares (RSS) defined by $RSS = \sum^{n}_{i=1} \check{\varepsilon}_i^2$, then the AIC is redefined as

$$AIC = 2k + n[ln(2\pi RSS/n) + 1] \quad (15)$$

A variation of the AIC which is used to test model fit is the Consistent AIC which includes an added penalty for models that have a greater number of parameters, k, for a given sample size n. It is defined by;

$$CAIC = -2lnL + [ln(1) + 1] \quad (16)$$

Another comparative fit measure which is the Bayesian information criterion (BIC). It is given by the following formula;

$BIC= 2ln(L) + kln(n)$ (17)

Where, n is the sample size, k is the number of free parameters to be estimated, L is the maximized value of the log-likelihood function for the estimated model.

Under the assumption that the model errors or disturbances are normally distributed, the BIC becomes;

$BIC= nln\left(\frac{RSS}{n}\right)+kln(n)$ (18)

The model with the lowest BIC value is preferable.

(D) Model Diagnostics

The reason for fitting any statistical model to data is to draw inferences about some aspects of the population from which the data were drawn. Since conclusions drawn can be sensitive to the accuracy of the fitted model, it is important to check that the model fits the data well. The reliability of fitting a result of a statistical model is assessed by goodness-of fit methods. In extreme value modelling there are various methods which are used for model diagnostics. These include probability-probability (PP) plots, quantile-quantile (QQ) plots, return level plots and empirical distribution function plots. A brief description of the plots used in this study is as follow.

(i) Q-Q Plot

The Q-Q plot graphs the sample quantiles against the theoretical quantiles of the distribution F and then a visual check is made to see whether or not the points are closed to a straight line. This plot if it fits the data well should converge to a straight line as the sample size increases. It is a graphical device used to test of goodness of fit of a sample $X_1,...,X_n$ to some distribution F in an exploratory way. It measures how close the sample quantiles are to the theoretical quantile. Rather than considering individual quantiles the Q-Q plot considers the sample quantiles against the theoretical quantiles of a specified target distribution F.

For a quantile plot, a given ordered sample of independent observations $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ from a population with estimated distribution function $\hat{F}$, has a quantile plot consisting of the points $\{(F^{-1}\frac{i}{n+1}) (x_{(i)}): i =1,...,n\}$

(ii) Probability-Probability Plot

P-P plot (probability-probability plot), also known as percent-percent plot, is a graphical technique to assess if a fitting result of probability distribution is a reasonable model by comparing theoretical and empirical probability. Given an ordered sample of independent observations $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ from a population with estimated distribution function F, then a probability plot consists of the points $\{(F(x_{(i)}), \frac{i}{n+1}): i =1,...,n\}$. If $\hat{F}$ is a reasonable model for the population distribution function, the points of the probability plot should lie in a straight line.

(iii) Return Level Plot of GEV Distribution

Return level plots are considered convenient for both presentation and validation of extreme value models. The return level plot for the GEV distribution is stated in equation (19) which is obtained by inverting the GEV distribution function.

$$z_p = \begin{cases} \mu - \sigma/\varepsilon[1 - \{-log(1-p)\} - \varepsilon], for \neq 0 \\ \mu - \sigma log[-log(1-p)], for\ \varepsilon = 0 \end{cases} \quad (19)$$

If $-log\ (1-p) = y_p$ so that the equation becomes after substitution,

$$z_p = \begin{cases} \mu - \sigma/\varepsilon[1 - \{-log(yp)\} - \varepsilon], for \neq 0 \\ \mu - \sigma log[-log(yp)], for\ \varepsilon = 0 \end{cases} \quad (20)$$

a plot of z_p against y_p on a logarithmic scale will produce a straight line when $\varepsilon = 0$ but has no finite limit when ε is less than zero.

(iv) Return Levels of GPD

It often of interest to evaluate return levels, for example the *N-year* return level denoted by x_N that is exceeded once in *N* years [4]. Given that the GPD model with parameters ε and σ is most suitable for exceedances, then the probability of exceedances of a variable *X* over a high threshold *u* is written as

$$P\{X > x/X > u\} = \left[1 + \varepsilon\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}} \quad (21)$$

Provided $x > u$ and $\varepsilon \neq 0$.

Also letting $\pi_u = p\{X > u\}$ where π_u represents the probability of occurrence of an excess of a high threshold *u* where the subscript *u* emphasizes that this value depends on the choice of threshold *u* then

$$P\{X > x\} = \pi_u\left[1 + \varepsilon\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}} \quad (22)$$

Therefore, the level x_m that is exceeded on average once every *m* observations will be obtained by solving the following equation

$$\pi_u\left[1 + \varepsilon\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}} = \frac{1}{m} \quad (23)$$

Rearrangement of equation 23 results in

$$x_m = u + \frac{\sigma}{\varepsilon}[(m\pi_u)^{\varepsilon} - 1] \quad (24)$$

Equation 23 according to [4] is only valid when *m* is large enough to ensure that $x > u$.

In the case where $\varepsilon = 0$, applying the same preceding procedures to the GPD leads to

$$x_m = u + \sigma log(m\pi_u) \quad (25)$$

Where x_m is called the m-observation return level. In the case where the N-year return level is of interest and given that the number of observations per year is n_y then

$$m = Nn_y \quad (26)$$

Hence, the N-year return level is given by

$$x_N = \begin{cases} u + \frac{\sigma}{\varepsilon}[(Nny\pi u)\varepsilon - 1], \varepsilon \neq 0 \\ u + \sigma log(Nny\pi u), \varepsilon = 0 \end{cases} \quad (27)$$

An estimate of π_u is also given the

$$\pi_u = \frac{k}{n} \quad (28)$$

Where *k* is the number of exceedances and *n* is the sample size. Coles [10], stated that the number of exceedances of *u* follows a Binomial distribution, Bin (n, π_u).

3 Results and Discussion

Summary of Statistics Result of Annual Precipitation

The graph of Maximum rainfall in Sokoto Basin from 1915-2018 is presented in Figure 1. Furthermore, the highest mean annual rainfall amount recorded for the area is 1477.5mm. The summary statistics for annual rainfall of is shown in Table 1. It includes measures of central tendency, variability, and shape. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the sample comes from a normal distribution. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate any statistical test regarding the standard deviation. In this case, the standardized skewness and kurtosis value is not within the range expected for data from a normal distribution.

Parameters Estimation Method and Model Selection Criteria

Table 2 shows result of the estimated parameters to determine which of the model is best fit. There is no much difference between the values for the test of location parameter (μ) of GEV and Gumbel compared to Poisson Process, which has the lowest value. The GP model does not have location parameter. The same goes for scale parameter (σ) of GEV whose value has a slight difference to Gumbel but has larger difference when compared with Poisson Process value. The GP model has the largest value for the scale parameter with value (201.5119). The shape parameter (ε) for both the GEV and GP model has negative value of (-0.06776) and (-0.148203), respectively. The Poisson Process value was (0.1463287) and Gumbel did not have shape parameter.

For model selection criteria, the Negative Log-Likelihood (NNL), Akaike Information Criterion and Bayesian Information Criterion were used for each model. NLL value for GEV and Gumbel were closed to each other compared to others. AIC value for GEV and Gumbel were also close to each other's as compared to other models. The same can also be said of BIC value for all the models, which follow the same trend in with GEV, Gumbel, Poisson Process and GP models values of 1391.063, 1387.6340, 171.5326 and 425.2249, respectively.
From Table 2, Poisson Process of Generalized Pareto appear to be the best model to fit the data since it has lowest AIC (166.9535), and BIC (171.5326) value.

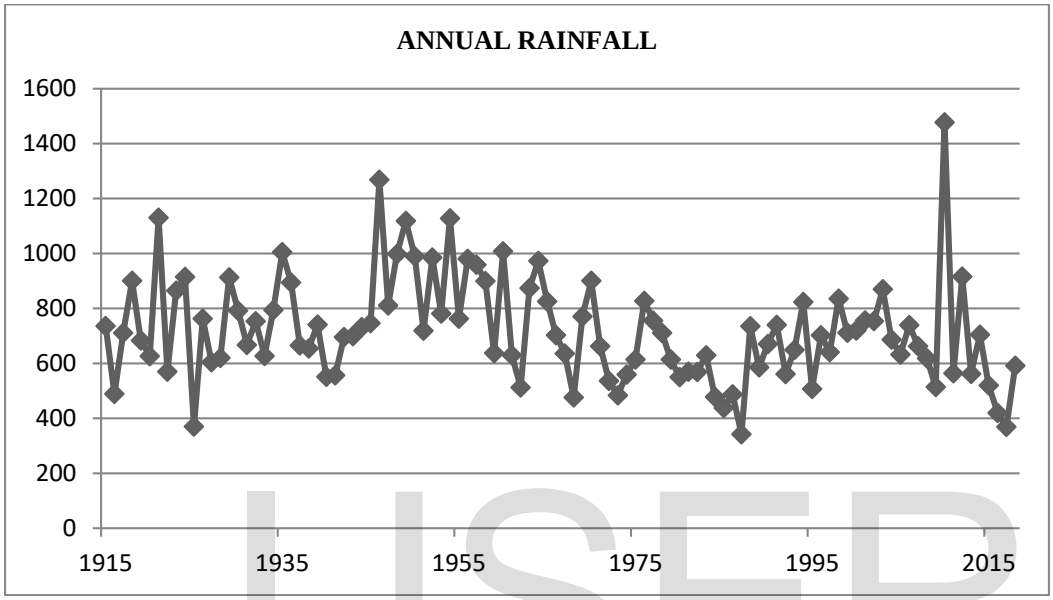

Figure 1: Graph of Annual Maximum Rainfall for Sokoto Basin (1915-2018)

Table 1: Summary of Statistics Result for Annual Precipitation

Count	104
Average	720.228
Standard deviation	192.326
Coefficient of variation	26.7034%
Minimum	342.2
Maximum	1477.5
Range	1135.3
Standardized skewness	3.82179
Standardized kurtosis	3.74741

Table 2: Parameter Estimation based on GEV, Generalized Pareto, Gumbel and Poisson Process

Estimation Methods				
	GEV Model	Gumbel	Poisson Process of GP Model	GP Model
Location	636.8953	631.9885	140.94048	-
Scale	161.8049	160.0016	55.7572989	201.5119
Shape	-0.06776	-	0.1463287	-0.148203

Model Selection Criteria				
Negative Log-Likelihood	688.5649	689.1727	80.476	209.0861
AIC	1383.13	1382.3450	166.9535	422.1722
BIC	1391.063	1387.6340	171.5326	425.2249

Stationarity Test

Computes the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for the null hypothesis that x is level or trend stationary. KPSS Level = 0.58349, Truncation lag parameter = 2, p-value = 0.02414. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time.

With alpha being 0.05, and p-value of the test statistics =0.02414, we reject the null hypothesis, and say that is rainfall is not level or trend stationary. The diagnostic plots of GEV distribution are presented in Figure 2.

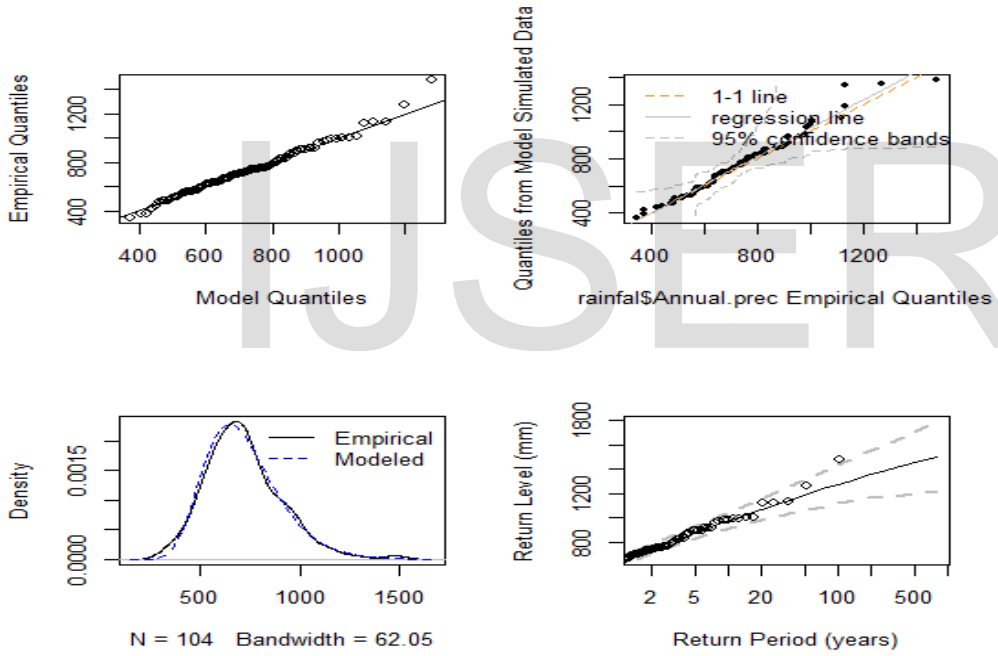

Figure 2: Diagnostic Plots of the GEV Distribution

Comparison of GEV and GPD Models

The stationary maximum rainfall return series was modelled using both the generalized extreme value and generalized pareto distribution. The negative value of the shape parameter suggests that the Weibull distribution fits the data well for both the GEV and GPD, tentatively.

Some graphical techniques were then employed to ascertain the fitness of the GEV and GPD to the data. The plots used were the model quantile plot, empirical quantile plot, density plot, and return level plot based on randomly generated data for the fitted GEV and GPD function as shown in Figure 2 for GEV and Figure 3 for GPD. Both Figures shows that the model quantiles and the empirical quantiles plot against those derived from the GEV and GPD function produced a straight line along the unit diagonal. This confirmed the asymptotic assumption for using both the GEV and GPD function. To further buttress the fit of the GEV and GPD, the density

plot shown good agreement between the empirical density shown by the solid black line and that of the fitted GEV and GPD function shown by the blue line.

Finally, the return level plot also shown that the plot on the log scale could be that heavy tail, bounded-tail and light-tail cases are concave, convex and linear, respectively. A visual inspection shows that it is convex which suggests that the shape parameter is negative. This confirms the negative value in the shape parameter as given in Table 2. The convexity of the return level plot also shows that the distribution that best fits the data from both the GEV and GPD family is of the Weibull type distribution, which has an upper bound. Furthermore, the return level plot shows the return periods in years for 2, 5, 20, 100, 500 and 1000 return periods.

The linearity of the model quantiles and the empirical quantiles from Figure 2 and 3 indicate that, the model is valid. From the histogram, the density also appears to be consistent with the data points. The diagnostic plots indicate a good fit for both the GEV and GPD model.

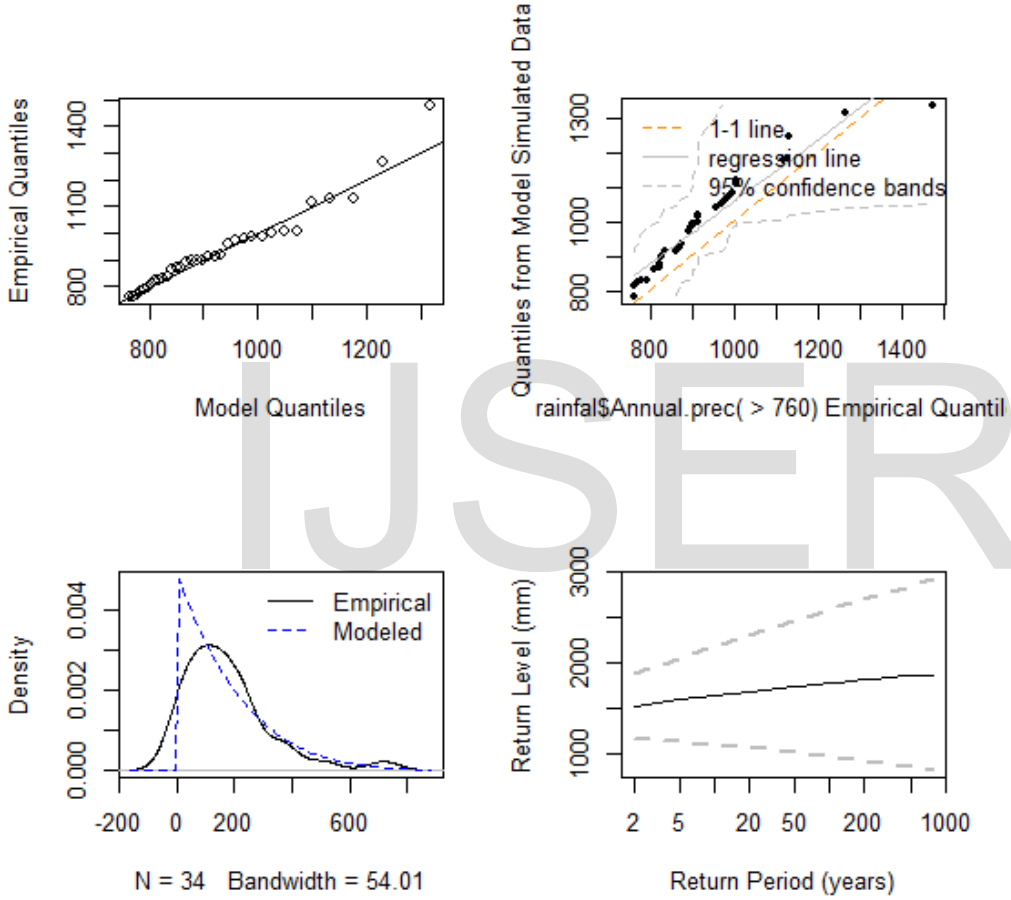

Figure 3: Diagnostic Plots of GPD Model

Fitting of Poisson Process to Generalized Pareto Distribution

The diagnostic plots suggest that both the GEV and GPD fit the data of maximum rainfall returns. However, their AIC and BIC values shown in Table 2 revealed that the Poisson Process of GP model is superior in fitting the data because it has the lowest AIC and BIC values. In order to validate this conclusion, further goodness-of-fit tests were conducted using three non-parametric tests, namely, Kolmogorov-Smirnov (KS), Cramer-von Mises (CM) and the Anderson-Darling (AD) tests. Table 3 shows the result of GP distributions based on the three tests. The data is not normally distributed, hence the use of Extreme Value model to capture to tail area of the distribution.

Table 3: Results of Goodness-of-Fit Tests

Normality Test: H_0: Normal

Type	Test Statistics	p-value
Anderson-Darling	1.0044	0.01152
Cramer-von Mises	0.17368	0.0114
Kolmogorov-Smirnov	0.10431	0.007242

The Threshrange Plot

The first step in fitting the GPD is the selection of an appropriate threshrange level for the tails of the distribution. Two methods for selecting an adequate threshold, namely parameter stability plot and mean residual life plot were used in the selection of the threshold. Firstly, in the parameter stability plot, a range of thresholds was arbitrary selected with the nature of the data in mind for a total number of 20 thresholds. Based on variation of the minimum and maximum thresholds using the *in2extreme* package in R, it was revealed that a minimum threshold of 200 and a maximum threshold of 800 yielded the best stability of the parameters (see Appendix A.1). The choice of minimum threshold and maximum threshold is based on the output generated by the software for arbitrary choices made based on the data and stability plot. The stability plots in Figure 4 show that the parameter estimates do not appear to vary considerably for even small values. Vertical line segments display the 95% normal approximation confidence intervals for each parameter estimate. Therefore, any threshold that lies between 200 and 800 could be used to fit the GPD. Based on this criterion, a threshold of 760 was chosen bearing in mind the fact that too high a threshold will result in few excesses that will result in a large variance.

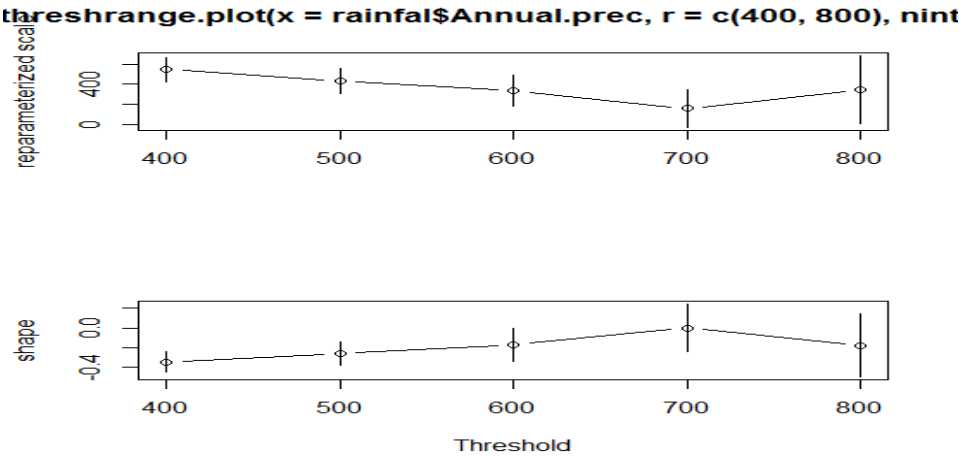

Figure 4: Threshrange plot

Mean Excess Plot

The second method, which was used to choose the appropriate threshold, is the mean excess plot. The idea is to find the lowest threshold whereby a straight line could be drawn from that point to higher values and still be within the uncertainty bounds indicated by blue dashed lines as shown in Figure 5. From Figure 5, a threshold of 760 was chosen to fit the model. The downward behaviour of the plot also suggests a light-tail

distribution. Mean Residual Life Plot play similar role with Mean excess plot. Figure 6 show that the plot was smooth up to a point except somewhere around 800, which conforms to that of mean excess plot.

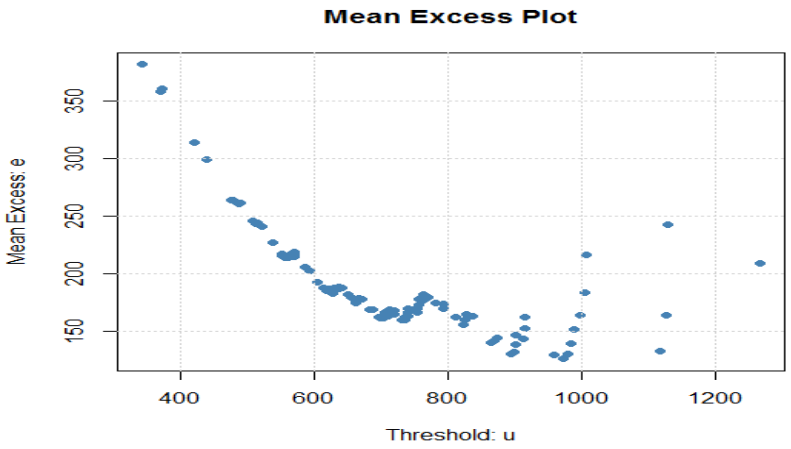

Figure 5: Mean Excess Plot

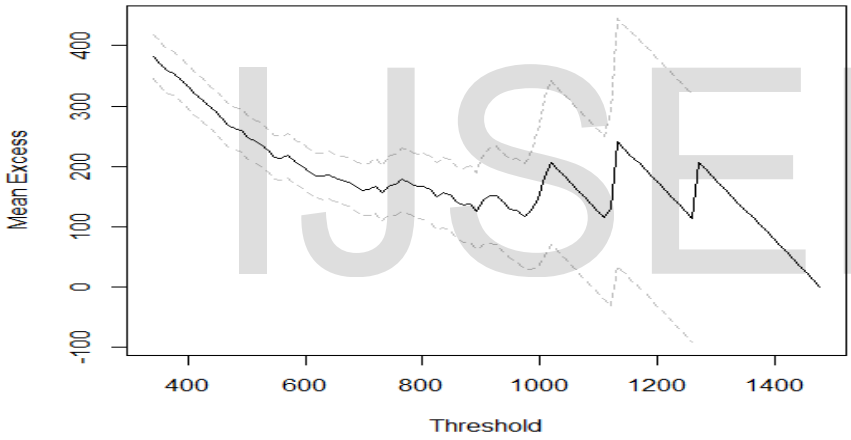

Figure 6: Mean Residual Life Plot

Fitting of Threshold Value

From Table 2, Poisson Process of Generalized Pareto appear to be the best model to fit the data since it has lowest AIC (166.9535), and BIC (171.5326) value.

Using the threshold value of 760 from mean excess plot, the estimated parameters of the GPD is shown in Table 2. The shape parameter, which is dominant in determining the qualitative behavior of the Poisson Process of GP distribution, is positive. The value of the shape parameter of the GPD (-0.148203) is not close to the estimated shape parameter value (-0.06776) in the GEV estimate. This shows that the distribution of excesses has an upper bound or upper end point and also is short-tailed. Just as in the GEV case, this is a Frechet distribution in the family of the generalized Pareto distributions.

Furthermore, to confirm that the threshold selected is good to use in fitting the GPD, diagnostic plots were plotted based on the selected threshold of 760. Figure 7 indicate that the assumptions for fitting the GPD to excesses over threshold were met. The diagnostic plots agreed with those of the GEV distribution function. The QQ-plot in Figure 7 shows that all the points are approximately linearly distributed along the unit diagonal showing a good fit of the GPD for maximum rainfall returns. This agrees with QQ-plot generated by randomly

selected data from the GPD against the empirical quantiles. The empirical density plot also affirms how adequate the GPD is in terms of modelling the data. It was observed that the number of excesses is 70 for the chosen threshold. The return level plot is also convex as in the GEV distribution case. Apart from a few points at the upper portion which show departure, the rest of the points lie on the line.

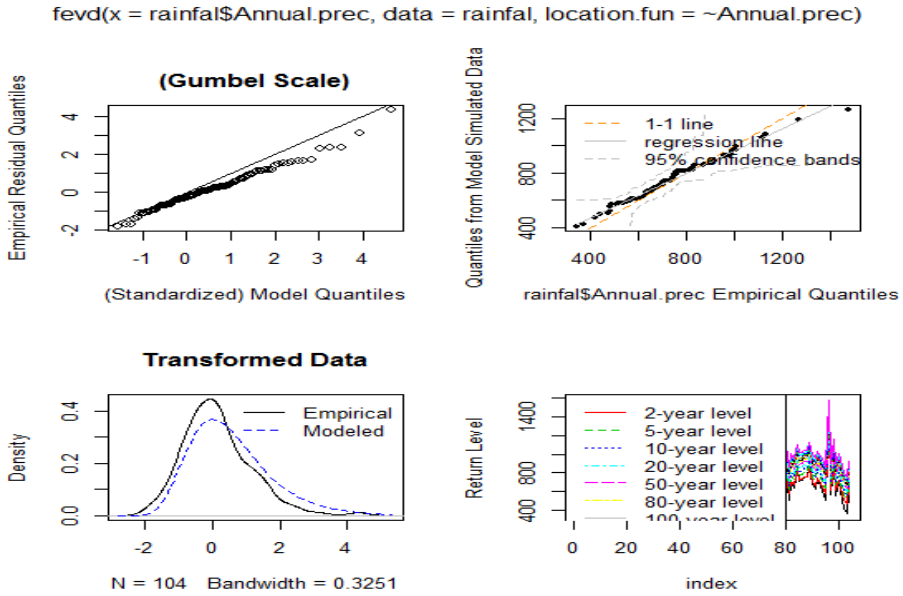

Figure 7: Diagnostic Plots for Chosen Threshold

Return Levels Estimate

In this theory, the return period (or the average recurrence interval) corresponds to the probability p of a return level that has a 100% chance of being exceeded in given year. The concepts of return level and return period are commonly used to convey information about the likelihood of rare events such as floods. A return level with a return period (years) of $T = 1/p$ is a high threshold x_p (e.g., maximum annual rainfall) whose probability of exceedance is p, [27].

The return level for 5-year, 10-year, 20-year, 30-year, 50-year, 100-year was carried out in this study and shown in Table 4. The 5-year return level shows that 867.662 mm of annual rainfall is expected at least once between the year 2018 and 2023. The 10-year return level shows that 974.6115 mm of annual rainfall is expected at least once between the year 2018 and 2028. The 20-year return level shows that 1072.2101 mm of annual rainfall is expected at least once between the year 2018 and 2038. Return level for 30, 50 and 100-year are 1126.2364, 1191.6743), and 1278.3797, respectively.

Table 4: Return Level Estimate

Return Level	95% Lower CI	Estimate	95% Upper CI
5-year	817.4591	867.6662	917.8733
10-year	910.0324	974.6115	1039.1907
20-year	987.2178	1072.2101	1157.2025
30-year	1026.3426	1126.2364	1226.1302

50-year	1070.0334	1191.6743	1313.3152
100-year	1120.3415	1276.3797	1432.4180

4 Conclusion and Recommendations

Conclusion

In this study, the annual maximum rainfall from December, 1915 up to December, 2018 were studied using two extreme value distribution models. Before fitting the models to the data, the maximum rainfall data were transformed logarithmically in order to ensure that they are stable. The following conclusions were made
(i) Maximum rainfall values followed a log-quadratic trend.
(ii) Generalized Pareto Distributions was adequate model for the maximum rainfall data in Sokoto Basin based on evidence from diagnostic plots, model stability checks and model comparison techniques.
(iii) The return levels revealed that rainfall are increasing and can reach unbearable levels in the far future.

Recommendations

The following recommendations were made based on the outcome of the study;
i. Farmers should be encouraged on climate change adaptation measures through agro-forestry, soil and water management. Alternative sources of water for crop irrigation should be encouraged such as water harvesting culture during the rainy season for later use.
ii. The study can be replicated in other basins within Nigeria to study the effect of global warming in the country.

REFERENCES

[1] Adefolalu, D.O. Further aspects of Sahelian droughts as evident from rainfall regime in Nigeria. Arch. Met. Geoph. Biolol. Ser. B., 36: 277-295, 1986.
[2] Allen, R.G., Pereira, L.S., Raes, D., Smith, M. Crop evapotranspiration - guidelines for computing crop water requirements - FAO irrigation and drainage paper 56. Rome: FAO; 1998.
[3] Babanazarov, B. Distribution of New temperature Extremes. Texas Technical University of Health Sciences. Unpublished M.Sc. Thesis. 2006.
[4] Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. Statistics of Extremes: Theory and Applications. Wiley, England. 2004.
[5] Bommier, E. Peaks-Over-Threshold Modelling of Environmental data. University of Uppsala, Sweden. An Unpublished M.Sc. Thesis. 2014.
[6] Box, G.E.P., and Jenkins, G.M. Time Series Analysis, Forecasting and control. Oakland, CA. Holden Day. 1976.
[7] Chifurira, R. Modelling Mean Annual Rainfall for Zimbabwe. University of the Free State, South Africa. An Unpublished PhD. Thesis. 2018.
[8] Chikobvu, D. and Chifurira, R. Modelling of extreme minimum rainfall using generalized extreme value distribution for Zimbabwe. South African Journal of Science, 111(9-10), pp.01-08, 2015.
[9] Coates, L.; Haynes, K.; Gissing, A.; Radford, D. The Australian experience and the Queensland Floods of 2010–2011. In The Handbook of Drowning: Prevention, Rescue, Treatment, 2nd ed.; Bierens, J.J.L.M., Ed.; Springer-Verlag Publishing: Berlin, Germany, Chapter 10–17, 2012, in press.
[10] Coles, S. An introduction to Statistical Modelling of Extreme Values. Springer-Verlag, London, 2001.
[11] Coles, S. and Tawn, J. Bayesian modelling of extreme surges on the UKeast coast. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 363(1831), pp.1387-1406, 2005.
[12] Davison, A.C. and Smith, R.L. Models or Exceedances over High Thresholds. Journal of the Royal Society, B 52, 393-442, 1990.
[13] Dottori, F., Szewczyk, W., Ciscar, J.-C., Zhao, F., Alfieri, L., Hirabayashi, Y., Bianchi, A., Mongelli, I., Frieler, K., Betts, R. A. and Feyen, L.: Increased human and economic losses from river flooding with anthropogenic warming, Nat. Clim. Chang, 2018; doi:10.1038/s41558-018-0257-z.
[14] Fisher, R.A. and Tippett, L.H.C. On the estimation of the frequency distributions of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 24, 180-190, 1928.
[15] Gilleland, E., and Katz, R.W. extRemes 2.0: An Extreme Value Analysis Package in R. Journal of Statistical Software, 2016; 72(8), 1-39. doi:10.18637/jss.v072.i08.

[16] Greene, W. H. Econometric analysis 3rd edition. Englewood Cliffs, NJ: Prentice Hall. 1997.
[17] Ismail, A. and Oke, I.A. Statistical and Trend Analyses of Rainfall in Sokoto. International Research Journal of Engineering Science, Technology and Innovation (IRJESTI) vol. 1(6) pp.161-174, September 2012.
[18] Jenkinson, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quarterly Journal of the Royal Meteorological Society, 81(348), pp. 158-171, 1955.
[19] Khapalova, E.A., Jandhyala, V.K. and Fotopoulos, S.B. Change-Point Analysis of Annual Mean Precipitation for Northern, Tropical and Southern Latitude of the Globe in the Past Century. Journal of Environmental statistics. Vol.4, Issue 3: 1-21, 2013.
[20] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. "Testing the null hypothesis of stationarity against the alternative of unit root". *Journal of Econometrics* 54, 159-178, 1992.
[21] Leadbetter, M.R., Lindgren, G. and Rootzn, H. Extremes and related properties of random sequences and processes. Springer Science & Business Media, 2012.
[22] Nadarajah, S. and Choi, D. Maximum daily rainfall in South Korea. Journal of Earth System Science, 116(4), pp.311-320, 2007.
[23] Nkrumah, S. Extreme Value Analysis of Temperature and Rainfall. University of Ghana, Ghana. An Unpublished M-PHIL Thesis. 2017.
[24] Resnick, S.I. Extreme values, regular variation and point processes. Springer, 2013.
[25] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019; URL https://www.R-project.org/.
[26] Sanchez, M.C. Statistical Modelling and Analysis of Summer Very Hot Events in Mainland Spain. An unpublished PhD. Thesis. Dept. of Geog., Centre for Climate Change, Virgili University. 2014.
[27] Santos, E.B., Lucio, E.S. and Santos E. Silva., C.M. Seasonal Analysis of Return Periods for Maximum Daily Precipitation in the Brazilian Amazon, Journal of Hydrohydrometeology, Vol 16, 2015.
[28] Scarrot, C., and McDonald, A. A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. REVSTAT-Statistical Journal, 10(1): 33-60, 2012.
[29] Thupeng, W.M. Statistical Modelling of Annual Maximum Rainfall for Botswana using Extreme Value Theory. International Academy and Science, Engineering and Technology (IASET). 2019.
[30] Thuto Mothupi, Wilson Moseki Thupeng, Baitshephi Mashabe and Botho Mokoto. Estimating Extreme Quantiles of the Maximum Surface Air Temperatures for the Sir Seretse Khama International Airport Using the Generalized Extreme Value Distribution. American Journal of Theoretical and Applied Statistics. Vol. 5, No. 6, pp. 365-375, 2016.
[31] Von MISES, R. La distribution de la plus grande de n valeurs. American Mathematical Society, Providence, RI, Selected Papers, 2, pp. 271-294, 1954.
[32] Wen, X., Fang, G., Qi, H., Zhou, L. and Gao, Y. Changes of temperature and precipitation extremes in China: past and future. Theoretical and Applied Climatology. 2015; Vol.10.1007/s00704-015-1584-x, 369-383.
[33] Wuertz, D., Setz, T. and Chalabi, Y. fExtremes: Rmetrics - Modelling Extreme Events in Finance. R package version, 2017; 3042.82.https://CRAN.R-project.org/package=fExtremes

APPENDIX A.1

R Code for precipitation

```
rainfal<-read.csv("C:/Users/TaiwoAdesina/Documents/Rasaq/rainfall.csv")

attach(rainfal)

#levelor trend stationary test
require(tseries)

kpss.test(rainfal[,2])
#Normality test
require(nortest)
#Anderson-Darling test
ad.test(rainfal[,2])
#Cramer-Von test
cvm.test(rainfal[,2])
#Lilli test
lillie.test(rainfal[,2])
#qqnorm plot
require(fBasics)
```

```
Qqnormplot(rainfal[,2])
require(fExtremes)

require(extRemes)

threshrange.plot(rainfal$Annual.prec, r = c(400, 800), nint=5)

mePlot(rainfal$Annual.prec)

mrlplot(rainfal$Annual.prec)

fit<- fevd(rainfal$Annual.prec,rainfal, units="mm")

fit

plot(fit)

par(mfrow=c(1,1))

plot(fit, "probprob")

plot(fit, "hist", col="gold",ylim=c(0,0.01))

plot(fit, "density", ylim=c(0,0.01))

plot(fit, "trace")
```

 # GEV

```
fit1 <- fevd(rainfal$Annual.prec,rainfal, units="mm")

fit1

plot(fit1)

plot(fit1, "trace")

return.level(fit1)

return.level(fit1, do.ci=TRUE)

ci(fit1, return.period=c(5,10,20,30,50,100)) # Same as above.
```

GP df

```
fit<- fevd(rainfal$Annual.prec,rainfal, threshold=760, type="GP", units="mm", verbose=TRUE)

fit

plot(fit)

plot(fit, "trace")

ci(fit, type="parameter")

par(mfrow=c(1,1))

ci(fit, type="return.level", method="proflik", xrange=c(4,7.5), verbose=TRUE)

# Can check using locator(2).
```

#PP

```
fit<- fevd(Fort$Prec, threshold=0.395, type="PP",
optim.args=list(method="Nelder-Mead"), units="inches", verbose=TRUE)
fit
plot(fit)
plot(fit, "trace")
ci(fit, type="parameter")
distill(fit)
distill(fit, cov=FALSE)
fit2 <- fevd(rainfal$Annual.prec,rainfal, location.fun=~Annual.prec)
fit2
plot(fit2)
##
# plot(fit2, "trace") # Gives warnings because of some NaNs produced
# (nothing to worry about).
lr.test(fit, fit2)
ci(fit)
ci(fit, type="parameter")
fit0 <- fevd(rainfal$Annual.prec,rainfal, type="Gumbel")
fit0
plot(fit0)
lr.test(fit0, fit)
plot(fit0, "trace")
ci(fit, return.period=c(5,10,20,30,50,100))
ci(fit, type="return.level", method="proflik", return.period=20, verbose=TRUE)
ci(fit, type="parameter", method="proflik", which.par=3, xrange=c(-0.1,0.5), verbose=TRUE)
# L-moments
fitLM<- fevd(rainfal$Annual.prec,rainfal, method="Lmoments", units="inches")
fitLM # less info.
plot(fitLM)
# above is slightly slower because of the parametric bootstrap
```

```
# for finding CIs in return levels.

par(mfrow=c(1,1))

plot(fitLM, "density", ylim=c(0,0.01))
```