ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Extreme Rainfall: Candidature Probability Distribution for Mean Annual Rainfall Data.

Retius Chifurira[1]*, and Delson Chikobvu[2].

[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban 4000, South Africa.
[2]Department of Mathematical statistics and Actuarial Science, University of Free State, Box 339, Bloemfontein 9300, South Africa.

ABSTRACT

Although a lot of literature exists on the selection of an appropriate probability distribution for daily or monthly rainfall, few studies have examined which probability distributions are most suitable to fit mean annual rainfall. The aims of this study are to select the best probability distribution to estimate mean annual rainfall and assess the effects of data length on the selection of a suitable probability distribution for Zimbabwe. The theoretical parent frequency distribution; log-logistic, lognormal and gamma distributions are fitted to mean annual rainfall. The performance of the fitted distributions are assessed using the goodness-of-fit tests, namely: relative root mean square error (RRMSE), relative mean absolute error (RMAE) and probability plot correlation coefficient (PPCC). Results show that the gamma distribution had more fitness with data series. However, the parent distribution sometimes diverges in predicting extreme maxima annual rainfall. Therefore, we compared the relative performance of the gamma distribution against the two-parameter exponential distribution in modeling extreme maxima mean annual rainfall. Results show that the two-parameter exponential distribution provide the best fit against all fitted distributions in all statistical periods, thereby providing a good alternative candidate for modelling mean annual rainfall extremes. The mean return period of mean annual rainfall amounts are calculated and return level of 1193mm (recorded high mean annual rainfall amount) is associated with a mean return period of approximately 579 years. This paper provides the first application of parent distributions and two-parameter exponential distribution derived from extreme value theory to mean annual rainfall from a drought prone country such as Zimbabwe.

Keywords: Mean annual rainfall, Goodness-of-fit, gamma distribution, lognormal distribution, log-logistic distribution, two-parameter exponential distribution.

Corresponding author

INTRODUCTION

Dry spells of varying severities are regular occurrences in Zimbabwe usually resulting in drought. Drought is a calamity with severe impacts on society. It contributes to loss of crops, animals and valuable property. Although knowledge of rainfall patterns over an area may be used for such disaster prevention purposes, it is one of the most difficult meteorological parameters to study because of a lack of reliable data and the large variations of rainfall in space, and time. Developing methods that can give a suitable prediction of hydrologic events is always interesting for both hydrologists and statisticians, mainly because of its importance in infrastructure development, water resource management and agriculture. Knowledge of rainfall characteristics, its temporal and spatial distribution plays a major role in drought-prone southern Africa, where economies are mainly driven by rain-fed agriculture (Jury, 1996). In modelling rainfall data, hydrologists and statisticians face difficulties, in most cases, the available amount of data is limited (Aksoy, 2000). To improve on modelling rainfall processes, many researchers have been searching for physical and statistical properties of rainfall using observational data. One area of interest is the parent probability distribution of rainfall amount (Cho *et al*., 2004; Deka *et al*., 2009). Modelling rainfall data using various mathematical models has been an important research area in meteorology and hydrology for the past three decades. Suhaila and Jemain (2007), Dan'azumi *et al*. (2010), Husak *et al*. (2007), Martins and Stedinger (2000) provide most recent research on mathematical modelling of rainfall patterns. It is generally assumed that a hydrological variable follows a certain probability distribution. Many probability distributions have been considered, in many different situations. There are many types of parent probability distributions used to fit rainfall data. These distributions are the gamma distribution (Aksoy, 2000; McKee *et al*, 1993; Cho *et al*, 2004; Adiku *et al*, 1997; Husak *et al*, 2007; Stagge *et al*, 2015), lognormal distribution (Deka *et al*, 2009; Cho *et al*, 2004; Suhaila *et al*, 2011), the generalized extreme value distribution (Roth *et al*, 2014; Madsen *et al*, 1997; Bulu and Askoy, 1998; Aksoy, 2000; Coles, 2001; Martins and Stedinger, 2000) and the log-logistic distribution (Fitzgerald, 2005; Almad *et al*, 1988).

Modelling of rainfall data have been investigated by several authors from different regions of the world. Rakhecha *et al*. (1994) analyzed the annual extreme rainfall series from India, covering over 80-years of rainfall data. Koutsoyiannis and Baloutsos (2000) analyzed rainfall data from Greece. Nadarajah (2005) provided the application of extreme value distributions to rainfall data from West Central Florida. Sakulski *et al*. (2014) fitted the log-logistic, Singh-Maddala, lognormal, generalized extreme value, Frechet and Rayleigh distributions to spring, summer, autum and winter rainfall data from the Eastern Cape province, South Africa and found that the Singh-Maddala to be the best fitting distribution to all four seasons rainfall data. Stagge *et al*. (2015) fitted seven candidate distributions to standardized precipitation index (SPI) and standardized evapo-transpiration index (SPEI) for Europe and recommended the two-parameter gamma distribution for modelling SPI and generalized extreme value distribution for modelling SPEI. Suhaila and Jemain (2007) found the mixed Weibull distribution as the best fitting distribution than single distributions in modelling rainfall amounts in Peninsular Malaysia. Zin *et al.* (2009) found the generalized lambda distribution as the best fitting distribution for rainfall amounts in Peninsular Malaysia as well. Those results differ from the results obtained by Suhaila and Jemain (2007). Therefore, each kind of probability distribution has its own applicability and limitations. A regionalized study on the statistical modelling of annual rainfall is very much essential as the statistical models may vary according to the geographical locations of the area considered and the length of the rainfall data series. In this study, an attempt has been made to fit the gamma, lognormal, log-logistic distributions to mean annual rainfall data series for Zimbabwe. These distributions are referred to as parent distributions since they fit to the whole body of the data. However, the tail of the theoretical parent distributions sometimes diverges in the extreme minima or maxima rainfall region. Extreme value theory is an alternative to fit minima and maxima mean annual rainfall (Chikobvu and Chifurira, 2015). Berger *et al*., (1982) and Surman *et al*., (1987) showed that the two-parameter exponential distribution fits well to extreme weather and atmospheric data. Therefore, the purposes of this study are: (1) Compare the relative performance of the gamma, lognormal, log-logistic distributions and the two-parameter exponential distributions in fitting the mean annual rainfall for Zimbabwe, (2) investigate the performance of the candidate distributions at 25, 50 and 75-year periods. (3) Select the most robust model using goodness-of-fit tests namely relative root mean square error (RRMSE), relative mean absolute error (RMAE) and probability plot correlation coefficient (PPCC) and (4) estimate the mean return period for specific return levels. The rest of the paper is organized as follows. In Section 2, we provide some background theory on the gamma, lognormal, log-logistic and the two-parameter exponential distributions. The data used in the study are described in Section 3. Section 4 presents data analysis on models fitted. Finally, Section 5 concludes this work.

METHODOLOGY

In order to describe the rainfall pattern at a particular area, it is necessary to identify the distribution(s), which best fit the data. In this section, we present some background theory on four theoretical distributions, namely two-parameter gamma, lognormal, log-logistic and two-parameter exponential distributions fitted to the mean annual rainfall. The parameters are estimated by the method of maximum likelihood. This procedure always gives minimum variance estimate of parameters. Here, we present the distribution functions for the proposed candidate models.

Two-parameter gamma distribution

The two-parameter gamma distribution is recommended for hydrological/meteorological frequency analysis (Mckee *et al.* 1993; Hosking, 1990). The cumulative probability function of the two-parameter gamma distribution used in the calculation of the standardized precipitation index, a widely used drought monitoring tool (Tigkas *et al*. 2015). The two-parameter gamma distribution is defined by the density function

$$f_{ga}(x) = \frac{x^{\beta_g - 1}}{\alpha_g{}^{\beta_g}\Gamma(\beta_g)} \exp\left(-\frac{x}{\alpha_g}\right), \quad x > 0 \qquad (1)$$

where α_g, $\beta_g > 0$. $\alpha_g$is the scale parameter and β_g is the shape parameter of the gamma distribution. The quantile function of the gamma distribution has no explicit form. In this paper, we estimate the quantiles for the gamma distribution using the R package “fitdistrplus”.

Maximum likelihood estimation for gamma distribution

Under the assumption that the observed n independent data points $X_1, \dots, X_n$ have a gamma distribution, the log-likelihood for the two-parameter gamma distribution is

$$l(\alpha_g, \beta_g) = (\beta_g - 1)\sum_{i=1}^{n} \log(X_i) - n\log\Gamma(\beta_g) - n\beta_g \,\text{alog}(\alpha_g) - \frac{1}{\alpha_g}\sum_{i=1}^{n} X_i \qquad (2)$$

The estimate of α_g is found to be

$$\hat{\alpha}_g = \frac{\bar{X}}{\beta_g}$$

and substituting this into the log-likelihood gives

$$l(\alpha_g, \beta_g) = n(\beta_g - 1)\sum_{i=1}^{n} \overline{\log(X_i)} - n\log\Gamma(\beta_g) - n\beta_g \log(\alpha_g) - n\beta_g \log(\bar{X}) + n\beta_g \log(\beta_g) - n\beta_g \qquad (3)$$

The estimate of β_g is obtained by maximizing (3) via the generalized Newton algorithm. The estimate of $\beta_g$is given by

$$\hat{\beta}_g \approx \frac{0.5}{\log\bar{X} - \overline{\log X}} \qquad (4)$$

where $\log\bar{X} > \overline{\log X}$.

Two-parameter lognormal distribution

Another distribution that is commonly used to model rainfall amounts is the lognormal distribution (Cho *et al*., 2004). The lognormal distribution is similar in appearance to the gamma distribution. An assumption of the lognormal distribution is that the logarithms of the data are normally distributed. The two-parameter lognormal distribution is defined by the density function

$$f_l(x) = \frac{1}{x\beta_l\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \alpha_l)^2}{2\beta_l^2}\right], x > 0 \qquad (5)$$

where $-\infty < \alpha_l < \infty$ and $\beta_l > 0$. α_l and β_l are the scale and shape parameters of the lognormal density function, respectively. Due to a close relationship with the normal distribution, the scale parameter α_l, may be interpreted as the mean of the logarithm of the random variable, while the shape parameter β_l, maybe interpreted as the standard deviation of the logarithmically transformed variables. Modelling with the log-normal distribution allows the use of normal-theory statistics on a logarithmic scale, and parameter estimation is then straightforward (Manning and Mullahy, 2001). The quantile function of the lognormal distribution is given by

$$Q(F_l) = \exp(\alpha_l + \beta_l \Phi^{-1}(F_i)) \quad (6)$$

where $\Phi^{-1}(.)$ has a standard normal distribution with mean zero and unit variance and F_i, is the i^{th} Gringorten plotting position given by $F_i = \frac{i-0.44}{n+0.12}$. i is the order statistics of mean annual rainfall.

Maximum likelihood estimation for lognormal distribution

Under the assumption that the observed n independent data points $X_1, \ldots, X_n$ have a lognormal distribution, the log-likelihood for the two-parameter lognormal distribution is derived by taking the product of the probability densities of the individual $X_i s$:

$$l(\alpha_l, \beta_l) = \ln\left((2\pi\beta_l^2)^{-\frac{n}{2}} \prod_{i=1}^{n} X_i^{-1} \exp\left[\frac{(\ln(X_i)-\alpha_l)^2}{\beta_l}\right]\right) \quad (7)$$

The estimates of α_l and β_l are obtained by maximising equation (7). The maximum likelihood parameter estimates are

$\hat{\alpha}_l = \frac{\sum_{i=1}^{n} \ln(X_i)}{n}$ and

$$\hat{\beta}_l = \frac{\sum_{i=1}^{n}\left(\ln(X_i) - \frac{\sum_{i=1}^{n}\ln(X_i)}{n}\right)^2}{n}.$$

Two-parameter log-logistic distribution

The log-logistic distribution is related to the logistic distribution in the same manner as the lognormal distribution is related to the normal distribution. A logarithmic transformation of the logistic distribution generates the log-logistic distribution. The log-logistic distribution is defined by the density function

$$f_{ll}(x) = \frac{(\beta_{ll}/\alpha_{ll})(x/\alpha_{ll})^{\beta_{ll}-1}}{[1+(x/\alpha_{ll})^{\beta_{ll}}]^2}, x > 0 \quad (8)$$

where $\alpha_{ll} > 0$ is the scale parameter of the two-parameter exponential distribution, and $\beta_{ll} > 0$ is the shape parameter of the distribution. The log-logistic distribution has different shapes: It can be strictly decreasing, right-skewed, or unimodal. This flexibility property enables the log-logistic distribution to fit data from many different fields, including engineering, economics, hydrology, and survival analysis. The quantile function of the log-logistic distribution is given by

$$Q(F_{ll}) = \alpha_{ll}\left(\frac{F_i}{1-F_i}\right)^{\frac{1}{\beta_{ll}}} \quad (9)$$

Maximum likelihood estimation for log-logistic distribution

Under the assumption that the n observations, denoted by $X_1, \ldots, X_n$ are from a log-logistic distribution, the log-likelihood function is

$$l(\alpha_{ll}, \beta_{ll}) = n\log(\beta_{ll}) - n\beta_{ll}\log(\alpha_{ll}) + (\beta_{ll} - 1)\sum_{i=1}^{n}\log(X_i) - 2\sum_{i=1}^{n}\log\left[1 + \left(\frac{X_i}{\alpha_{ll}}\right)^{\beta_{ll}}\right] \quad (10)$$

The parameter estimates are obtained by differentiating the log-likelihood function with respect to α_{ll} and β_{ll}, and equating to zero. We use the R package 'fitdistrplus' to obtain the maximum likelihood parameter estimates of the gamma, lognormal and log-logistic distributions.

Two-parameter exponential distribution derived from extreme value theory

Extreme value theory has emerged as one of the most important statistical disciplines for meteorological sciences over the last 50 years (Li *et al*., 2015). The two-parameter exponential distribution is recommended to model extreme events (Lu, 2004). A random variable X is said to have a two-parameter exponential distribution (Exp(α_E, β_E)) if its probability density function is of the form

$$f_{\mathrm{e}}(x) = \frac{1}{\beta_{\mathrm{e}}} \exp - \left(\frac{x-\alpha_{\mathrm{e}}}{\beta_{\mathrm{e}}}\right), -\infty < x < \infty,\ \beta_{\mathrm{e}} > 0 \quad (11)$$

where α_{e} is the location parameter and β_{e} is the scale parameter. Berger *et al*. (1982) derived the two-parameter exponential distribution, F_{e}, from extreme value theory to represent the cumulative frequency distribution of maxima random variable over a specific percentile,

$$F_{\mathrm{e}} = 1 - \exp[-y] \quad (12)$$

where $y = \frac{x-\alpha_E}{\beta_{\mathrm{e}}}$. The mean annual rainfall below a specific drought threshold was chosen from the complete set of mean annual rainfall to fit the two-parameter exponential distribution. The estimated cumulative probability $\bar{F}_{\mathrm{e}}$ can be calculate from the maxima mean annual rainfall x from

$$\hat{F}_{\mathrm{e}}(x_i) = \frac{N-r+1}{N+1} = P_r \quad (13)$$

where N is the size of the chosen minima mean annual rainfall. P_i is the probability of a value that is ranked r out of N values. Therefore, the relationship between the variate y and P_r is given by

$$y(i) = \ln(1 - P_r) \quad (14)$$

The estimates of α_{e} and β_{e} can be estimated by least-squares method since y and x are linearly related. The quantile function of the two-parameter exponential distribution is given by

$$Q(F_{\mathrm{e}}) = \alpha_{\mathrm{e}} - \beta_{\mathrm{e}} \ln(1 - F_i) \quad (15)$$

Return period

In order to develop an effective early warning drought monitoring strategy, we estimate how often the extreme quantiles occur with a certain return level. The cumulative two-parameter exponential distribution is used to calculate return period as suggested by Berger et al. (1982). The return period is given by

$$R(x) = \frac{1}{[(1-f)(1-F_{\mathrm{e}}(x)} \quad (17)$$

where $R(x_c)$ is the return period in years of the mean annual rainfall, x and f is the chosen specific percentile.

Model adequacy and Goodness-of-fit

Model adequacy was performed using the Anderson-Darling (AD) tests and we judge goodness-of-fit for the fitted distributions using the relative root mean squared error (RRMSE), relative mean absolute error (RMAE) and probability plot correlation coefficient (PPCC).

Anderson-Darling test

The AD test is an improvement of the Kolmogorov-Smirnov test. It gives more weight to the tails of the distribution (Farrel and Stewart, 2006). The AD statistic is defined as

$$W_n^2 = n\int_{-\infty}^{\infty}[F_n(x) - F^*(x)]^2 \psi(F^*(x))dF^*(x) \quad (16)$$

where ψ is a nonnegative weight function which can be computed from

$$\psi = [F^*(x)(1 - F^*(x))]^{-1}$$

In order to make the computation of the statistic easier, Arshad *et al*. (2003) redefined the KS statistic as

$$W_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log F^*(x_i) + \log(1 - F^*(x_{n+1-1}))] \quad (17)$$

where $F^*(x_i)$ is the cumulative distribution function of the specified distribution, x_i is ordered data and n is the sample size. According to Arshad *et al*. (2003), the AD test is the most powerful empirical distribution function test. Since the AD statistics is measure of the distance between the empirical and hypothesized distribution functions, the fitted distribution with the smallest AD statistic value will be selected as the best fitting model.

Goodness-of-fit tests

RRMSE and RMAE assesses the difference between the observed values and the expected values of the assumed distributions. The PPCC measures the correlation between the ordered values and the corresponding expected values of the assumed distribution. The formulae for the tests are

$$RRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{i:n} - \hat{Q}(F_i)}{x_{i:n}}\right)^2} \quad (18)$$

$$RMAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_{i:n} - \hat{Q}F_i}{x_{i:n}}\right| \quad (19)$$

$$PPCC = \frac{\sum_{i=1}^{n}(x_{i:n} - \bar{x})(\hat{Q}(F_i) - \bar{Q}(F))}{\sqrt{\sum_{i=1}^{n}(x_{i:n} - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(\hat{Q}(F_i) - \bar{Q}(F))^2}} \quad (20)$$

where $x_{i:n}$ is the observed values of the i^{th} order statistics of a random sample of size n. $\hat{Q}(F_i)$ is the estimated quantile value of the assumed distribution associated with the i^{th} Gringorten plotting position, $F_i = \frac{i-0.44}{n+0.12}$. $\bar{Q}(F_i)$ and $\bar{x}$ are the averages of $\hat{Q}(F_i)$ and $x_{i:n}$ respectively. The fitted distribution with the smallest values of the $RRMSE$ and $RMAE$ is selected as the best fitting distribution while the distribution with the computed $PPCC$ closest to 1 indicates the best fitting distribution.

DATA

We use the mean annual rainfall data series for the period 1901 to 2009. The time series plot in Figure 1 shows that the mean annual rainfall seems to be stationary. The Augmented-Dickey-Fuller test is used to formally test for stationarity in mean and variance. The null hypothesis is that the mean annual rainfall series is non-stationary and the alternative hypothesis is the mean annual rainfall series is stationary. The Augmented-Dickey Fuller statistic is -4.378 with *p*-value = 0.01< 0.05 rejects the null hypothesis at 5% significance level implying that the mean annual rainfall data are stationary.

Figure 1: Time series plot of mean annual rainfall

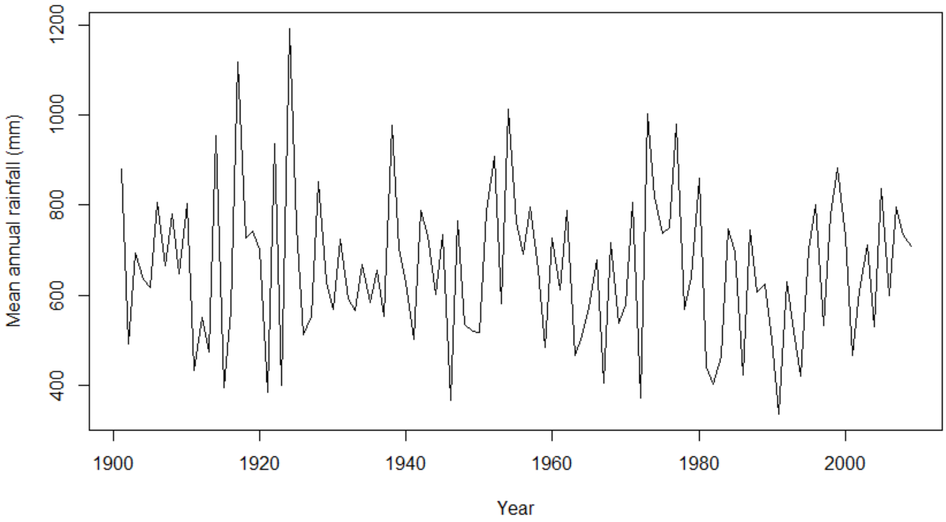

Table 1 presents the descriptive statistics for the mean annual rainfall data. The positive skewness and negative excess kurtosis clearly illustrates the non-normality of the distribution.

Table 1: Descriptive statistics of mean annual rainfall

No. of obs	Mean	Std. dev.	Min	Max	Skewness	Excess Kurtosis
109	659.9312	169.2457	335.3000	1192.6000	0.4455	0.1222

The table reports summary statistics for the mean annual rainfall for Zimbabwe.

Fitting a statistical distribution usually assumes that the data are independent and identically distributed (i.e., randomness), with no serial correlation, and no heteroscedasticity. We tested for randomness using the Brock-Dechert-Scheinkman (BDS), Box and Pierce (1970) and Bartels (1982) tests. The null hypothesis for the tests is that the annual rainfall is independent and identically distributed (i.i.d). The corresponding *p*-values based on the mean annual rainfall are given in Table 2.

Table 2: *p*-values of the tests for randomness

Test	*p*-values
BDS	0.3970
Box and Pierce (1970)	0.5870
Bartels (1982)	0.7570
Rank	0.8500
Cox and Stuart (1955)	0.5040

We tested for no serial correlation using the Ljung-Box, the Durbin-Watson and the Godfrey-Breusch tests. The null hypothesis for the tests is the annual rainfall is not serially correlated. The corresponding *p*-values based on the mean annual rainfall are given in Table 3.

Table 3: *p*-values of the tests for no serial correlation

Test	*p*-values
Ljung-Box	0.1270
Durbin-Watson	0.3795
Breusch- Godfrey	0.7838

We tested for no heteroscedasticity using the ARCH LM and Breush-Pagan tests. The null hypothesis for the tests is the annual rainfall data has no presence of heteroscedasticity. The corresponding *p*-values based on the annual rainfall are given in Table 4.

Table 4: *p*-values of the tests for no heteroscedasticity

Test	*p*-values
ARCH LM	0.2790
Breush-Pagan	0.1495

All the tests reported in Tables 2, 3 and 4 are non-parametric in nature, i.e., no distributional assumptions are made about the data. The tests confirm that the mean annual rainfall data are independent and identically distributed, with no serial correlation and have no heteroscedasticity.

RESULTS AND DISCUSSION

The three parent distributions were fitted to the data described in Section 2. Figure 2 show the c.d.f of three theoretical parent distributions and mean annual rainfall for Zimbabwe.

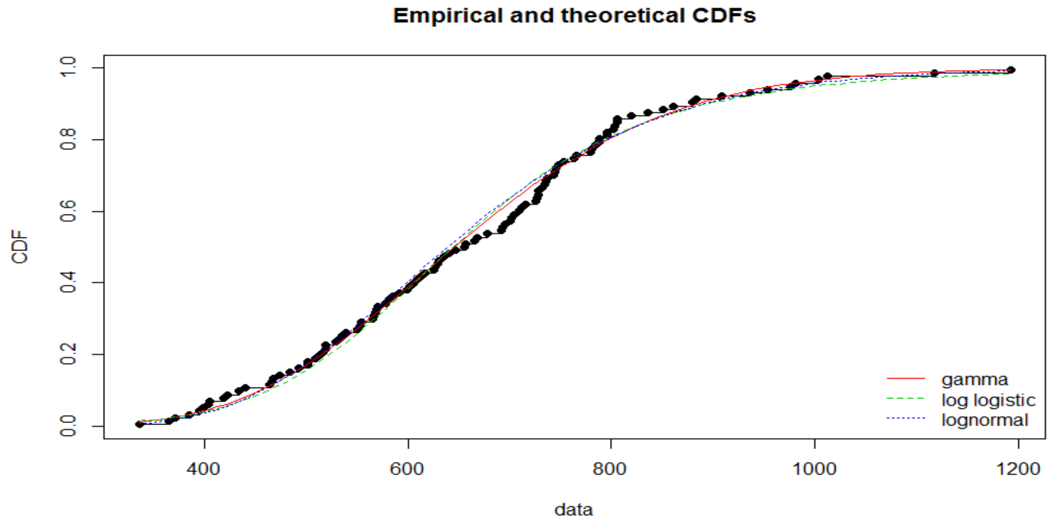

Figure 2: The c.d.f. of three theoretical parent distributions and mean annual rainfall for Zimbabwe

From Figure 2, the c.d.f. of the three fitted theoretical distributions seems similar to the frequency distribution of the data. The distribution parameters i.e. scale and shape parameters can be estimated by the maximum likelihood estimation procedure. The parameter estimates with their standard errors, AIC values and *p*-values of AD for the fitted distributions are shown in Table 5.

Table 5: Fitted distributions, parameter estimates with standard errors in brackets

Distribution	$\hat{\alpha}$	$\hat{\beta}$	AIC	*p*-value for AD
Gamma	0.0232 (0.0030)	15.2801 (1.9752)	1426.5270	0.2835
Lognormal	6.4591 (0.0249)	0.2598 (0.0176)	1427.5860	0.4637
Log-logistic	644.0337 (16.1974)	6.6773 (0.5288)	1420.9270	0.5125

The *p*-values of the AD which are all greater than 0.05 as reported in Table 5. Overall, the log-logistic distribution gives the best fit by having the least AIC and the largest *p*-value for the AD test. Subsequent analysis involves selection of the best fitting distribution out of the three candidate distributions using the goodness-of-fit tests. Results of the goodness-of-fit tests are presented in Table 6.

Table 6: Outcomes of the GOF tests

Distribution	RRMSE	RMAE	PPCC
Gamma	**0.0218**	**0.0156**	**0.9970**
Log normal	0.0260	0.0200	0.9950
Log logistic	0.0318	0.0266	0.9879

The distribution that is found best at least twice out of the three goodness-of-fit tests is selected as the best fitting distribution. Results indicate that the gamma distribution is the best fitting parent distribution since it has the least RRMSE and RMAE and the highest PPCC values. We compare the relative performance of the fitted distributions at 25, 50 and 75 year periods. Table 7 shows the goodness-of-fit test results at different statistical periods.

Table 7: Outcomes of the GOF tests at different statistical periods

period	Distribution								
	Gamma			Lognormal			Log-logistic		
	RRMSE	RMAE	PPCC	RRMSE	RMAE	PPCC	RRMSE	RMAE	PPCC
25	0.5821	0.5147	0.2359	0.5864	0.5135	0.2317	0.6141	0.5311	0.2260
50	0.3061	0.2718	0.9559	0.3089	0.2713	0.9389	0.3168	0.2739	0.9191
75	0.1774	0.1545	0.9657	0.1825	0.1537	0.9500	0.1998	0.1627	0.9257

From Table 7, the gamma distribution is found to be the best fitting distribution at each period, having the lowest RRMSE and RMAE values and the highest PPCC value which is closer to one. The results also show that by increasing the period the performance of the fitted distribution improves. The PPCC value is close to one when the period is 50 years or more. This suggests that analyzing mean annual rainfall data using parent distributions requires data of length at least 50 or better.

Parent distributions are known to diverge in predicting high or low rainfall amounts. We fit a two-parameter exponential and compare their relative performance against the best fitting gamma distribution. To fit a two-parameter exponential distribution to extreme maxima rainfall, we select a drought threshold. The Department of Meteorological Services of Zimbabwe determines the drought threshold value of 75% of the average annual rainfall of a 30-year rainfall time series obtained from aerially averaging ten rainfall stations with long enough rainfall data sets. We use a drought threshold value of 473mm. Rainfall amounts above 473mm are selected. We fit the two-parameter exponential distributions to the selected data using least-squares method since $y(i)$ and x are linearly related. Figure 3 show the theoretical line of the variate, $y(i)$, and mean annual rainfall over the drought threshold value of 473 mm.

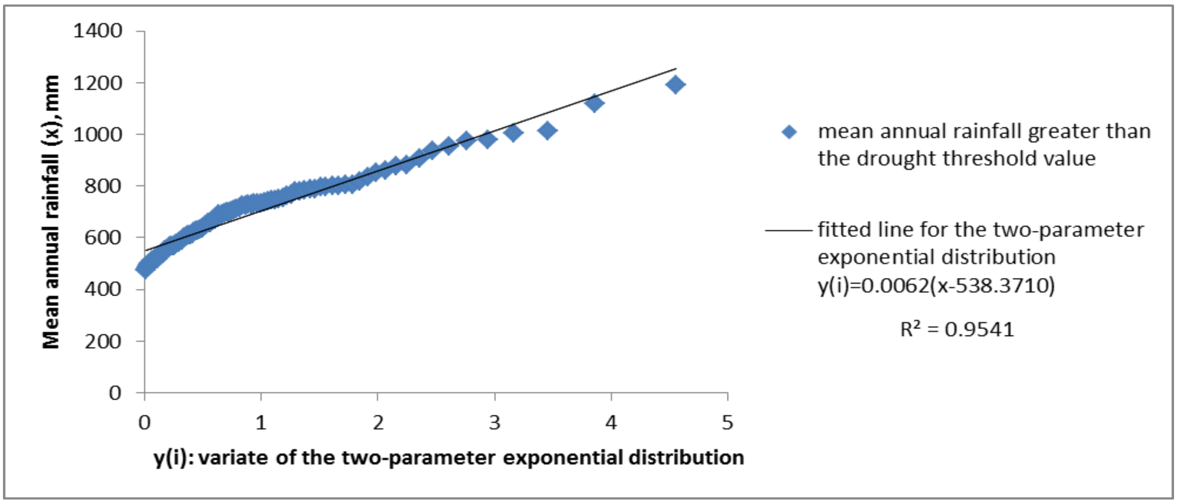

Figure 3: The fitted theoretical line of variate and mean annual rainfall above the drought threshold value of 473 mm by the two-parameter exponential distribution

The regression model result shows that the coefficient of determination is greater than 0.95 which indicates that the two-parameter exponential distribution can fit maxima mean annual rainfall well. The F_e can be calculated form Equation (12). The fitted two-parameter exponential distribution is $F_e(x) = 1 - \exp[-0.0062\ (x - 538.3710)]$. The goodness-of-fit tests are used to compare the relative performance of the two-parameter distribution against the best fitting parent distribution. Table 8 shows the goodness-of-fit tests for the gamma and two-parameter exponential distributions.

Table 8: Outcomes of the GOF tests for gamma and two-parameter exponential distributions

Distribution	RRMSE	RMAE	PPCC
Gamma	0.0323	0.0220	0.9887
Two-parameter exponential	**0.0136**	**0.0012**	**0.9998**

From Table 8, the goodness-of-fit results shows that the two-parameter exponential distribution fits the data better than the best fitting parent distribution, the gamma distribution. We also compare the two distributions at different periods. Table 9 shows the goodness-of-fit tests results at different periods.

Table 9: Outcomes of the GOF tests for gamma and two-parameter exponential distributions at different periods

period	Distribution					
	Gamma			Two-parameter exponential		
	RRMSE	RMAE	PPCC	RRMSE	RMAE	PPCC
25	0.5821	0.5147	0.2359	0.0230	0.00136	0.9892
50	0.3061	0.2718	0.9559	0.0254	0.0124	0.9887
75	0.1774	0.1545	0.9657	0.0262	0.0117	0.9900

From Table 9, the two-parameter exponential distribution is found to be the best performing distribution in fitting mean annual rainfall data series at all given periods. It can also be seen that the PPCC value for the two-parameter exponential distribution is closer to one at all periods. This indicates that the two-parameter exponential distribution fits the data well with even small data length. Thus, the distribution is a good candidate distribution for fitting and modelling extreme mean annual rainfall data regardless of the sample size of the data.

The F_E and return period of mean annual rainfall is calculated from Equations (12) and (17). For example, $F_E(473\ \text{mm}) = 1 - \exp[-0.0062(473 - 538.3701)] = -0.4997$. The chosen data above the drought threshold value of 473mm corresponds to 90th percentile, i.e. $f = 0.90$. Then the return period of 473mm (drought threshold value) is $R(x_c) = \frac{1}{[(1-0.9)(1+0.4997)]} \approx 7$ years, so a mean annual rainfall amount of 473 mm is expected to return in 7 years' time. The maximum mean annual rainfall for Zimbabwe is 1192.6 mm recorded the in 1923/24 rainfall season. The mean return period associated with a return level estimate of 1193mm is approximately 579 years. This suggests that extreme flood of this magnitude is likely to return in year 579 period on average.

CONCLUSION

This study mainly investigates the relative performance of three commonly used probability distributions for mean annual rainfall, with the purpose of both providing recommendation in the selection of suitable distribution for frequency analysis of mean annual rainfall. Results show that the gamma distribution is the most suitable parent distribution. The worst performing is the log-logistic distribution. The performance of the best fitting parent distribution is compared to the performance of the two-parameter exponential distribution in fitting high rainfall extremes. It is found that the fitted two-parameter exponential agrees better with the actual data than the best fitting parent distribution at all different lengths of data outperforming the gamma distribution. This leads to the recommendation that the two-parameter exponential distribution is preferable to model mean annual rainfall. The return level estimates, which is the return level expected to be exceeded in a certain period of time *T* in years are calculated for Zimbabwe rainfall using the two-parameter exponential distributions. The highest mean annual rainfall amount recorded for the country is 1192.6mm. A return level of 1193mm is associated with a mean return period of 579 years, on average. Although, national data are analysed, the results of this study can be extended to station data in Zimbabwe. The findings of this study provide useful information for early drought monitoring management and provides a good alternative candidate for modelling mean annual rainfall extremes.

REFERENCES

[1] Adiku, S.G.K., Dayananda, P.W.A., Rose, C.W. and Dowuona, G.N.N., (1997). An analysis of the within-season rainfall characteristics and simulation of the daily rainfall in two savannah zones in Ghana. *Agricultural and forest meteorology*, *86*(1), 51-62.

[2] Aksoy, H. (2000). Use of Gamma distribution in hydrological analysis, *Turk Journal of Engineering Science*, 24: 419-428.

[3] Arshad, M., Rasool, M.T. and Ahmad, M.I. (2003). Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution. *Pakistan Journal of Applied Sciences* 3(2), 85-88.

[4] Berger, A., Melice, J. L., and Demuth, C. (1982). Statistical distributions of daily and high SO_2-concentrations. *Atmospheric Environment*, 16(12), 2863-2877.

[5] Bulu, A. and Aksoy, H. (1998). Low flow and drought studies in turkey, proceedings low flows expert meeting, 10-12 June 1998, University of Belgrade.

[6] Chikobvu, D., and Chifurira, R. (2015). Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, *111*(9-10), 01-08.

[7] Cho, H.K., Bowman, K.P. and North, G.R., (2004). A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, *43*(11), 1586-1597.

[8] Coles, S (2001). *An introduction to statistical modelling of extreme values*. Great Britain: Springer.

[9] Deka, S., Borah, M., and Kakaty, S. C. (2009). Distribution of Annual Maximum Rainfall of North-East India. *European Water* 27/28:3-14.

[10] Dan'azumi, S., Shamsudin, S. and Aris, A., (2010). Modeling the distribution of rainfall intensity using hourly data. *American Journal of Environmental Sciences*, *6*(3), 238-243.

[11] Farrell, P.J. and Rogers-Stewart, K., (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, *76*(9), 803-816.

[12] Fitzgerald, D. L. (2005). Analysis of extreme rainfall using the log logistic distribution. Stochastic Environmental Research and risk Assessment, 19, 249-257.

[13] Hosking, J.R., (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 105-124.

[14] Husak, G.J., Michaelsen, J. AND Funk, C., (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, *27*(7), 935-944.

[15] Jury, M.R., (1996). Regional tele-connection patterns associated with summer rainfall over South Africa, Namibia and Zimbabwe. *International Journal of Climatology*, *16*(2), 135-153.

[16] Koutsoyiannis, D., and Baloutsos, G. (2000). Analysis of a long record of annual maximum rainfall in Athens, Greece, and design rainfall inferences. *Natural Hazards*, *22*(1), 29-48.

[17] Li, Z., Brissette, F., and Chen, J. (2013). Finding the most appropriate precipitation probability distribution for stochastic weather generation and hydrological modelling in Nordic watersheds. *Hydrological Processes*, *27*(25), 3718-3729.

[18] Lu, H. C. (2004). Estimating the emission source reduction of PM 10 in central Taiwan. *Chemosphere*, *54*(7), 805-814.

[19] Madsen, H., Pearson, C.P., and Rosbjerg, D., (1997) Comparison of annual maximum series and partial duration series methods for modelling extreme hydrological events II: regional modelling. *Water Resource Research*. 17, 1421-1432.

[20] Manning, W. G., and Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of health economics*, *20*(4), 461-494.

[21] Martins, E.S. and Stedinger, J.R., (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*(3), 737-744.

[22] McKee, T.B., Doesken, N.J., and Kleist, J., (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, 17(22), 179-183.

[23] Nadarajah, S. (2005). Extremes of daily rainfall in West Central Florida. *Climatic change*, *69*(2), 325-342.

[24] Rakhecha, P. R., and Soman, M. K. (1994). Trends in the annual extreme rainfall events of 1 to 3 days duration over India. *Theoretical and Applied Climatology*, *48*(4), 227-237.

[25] Roth, M., Buishand, T. A., Jongbloed, G., Tank, A. K., and Van Zanten, J. H. (2014). Projections of precipitation extremes based on a regional, non-stationary peaks-over-threshold approach: A case study for the Netherlands and north-western Germany. *Weather and Climate Extremes*, *4*, 1-10.

[26] Sakulski, D., Jordaan, A., Tin, L., and Greyling, C. (2014). Fitting Theoretical Distributions to Rainy Days for Eastern Cape Drought Risk Assessment. In *Proceedings of DailyMeteo. org/2014 Conference* (p. 48).

[27] Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F., and Stahl, K. (2015). Candidate distributions for climatological drought indices (SPI and SPEI). *International Journal of Climatology*, *35*(13), 4027-4040.

[28] Suhaila, J. and Jemain, A.A., (2007). Fitting daily rainfall amount in Peninsular Malaysia using several types of exponential distributions. *Journal of Applied Sciences Research*, *3*(10), 1027-1036.

[29] Surman, P. G., Bodero, J., and Simpson, R. W. (1987). The prediction of the numbers of violations of standards and the frequency of air pollution episodes using extreme value theory. *Atmospheric Environment (1967)*, *21*(8), 1843-1848.

[30] Tigkas, D., Vangelis, H., and Tsakiris, G. (2015). DrinC: a software for drought analysis based on drought indices. *Earth Science Informatics*, *8*(3), 697-709.

[30] Zin, W. Z. W., Jemain, A. A., and Ibrahim, K. (2009). The best fitting distribution of annual maximum rainfall in Peninsular Malaysia based on methods of L-moment and LQ-moment. *Theoretical and applied climatology*, *96*(3-4), 337-344.