

Determination of the Best Fit Probability Distribution for Monthly Rainfall Data in Bangladesh

Sanjib Ghosh^{1,*}, Manindra Kumar Roy², Soma Chowdhury Biswas³

¹Department of Statistics, University of Chittagong, Bangladesh

²Ranada Prasad Shaha University, Narayangaj, Bangladesh

³University of Chittagong, Bangladesh

Abstract An attempt has been made to determine the best fitted distribution to describe the monthly rainfall data for the period of 1979 to 2013 of distantly located stations in Bangladesh such as Chittagong, Dhaka, Rajshahi and Sylhet. The Normal, Lognormal, Gamma, Weibull, Inverse Gaussain and Generalized Extreme Value distributions are fitted for these purposes using the method of L-moment. The performances of the distributions are evaluated using three goodness-of-tests namely Kolmogorov-Smirnov, Anderson-Darling and Chi-Square test. Finally, goodness of fit test result are compared and generalized extreme value distribution is empirically proved to be the most appropriate distribution of the monthly rainfall data for the three selected station only Dhaka station provides good fit gamma distribution.

Keywords Rainfall, Probability distribution and Goodness-of-fit tests

1. Introduction

Analysis of rainfall data strongly depends on its distribution pattern. It has long been a topic of interest in the fields of meteorology in establishing a probability distribution that provides a good fit to monthly rainfall. Several studies have been conducted in Bangladesh and abroad on rainfall analysis and best fit probability distribution functions.

In the advent of global warming, there are increased concerns regarding extreme weather events. As elsewhere across the globe, South Asian countries have been observing an increase in occurrence of extreme climate events in recent decades. Many researchers have found evidence of increasing extreme weather events such as heat waves, cold waves, floods, droughts and severe cyclones over the past few decades. Extreme rainfall events can have severe impacts on human life. An investigation of extreme rainfalls by the scientific community in any country serves several purposes such as: the estimation of extreme rainfalls for design purposes; the assessment of rarity of observed rainfall; and comparison of methods to estimate design rainfalls. This study also focuses the last purpose and a detailed regionalized study is practically useful for the planners and other users.

The selection of an appropriate model depends mainly on these evaluations that yield very different conclusions than

that of previous researches on this subject. Application of probability distributions to rainfall data have been investigated by several researchers from different regions of the world. Hirose (1994) have found that the weibull distribution is the best fit for the annual maximum of daily rainfall in Japan. Nadarajah and Withers (2001) and Nadarajah (2005) provided the application of extreme value distributions to rainfall data over sixteen locations spread throughout New Zealand and fourteen locations in West Florida, respectively. Further, Nadarajah and Choi (2007) have studied annual maxima of daily rainfall for the years 1961-2001 for five locations in South Korea, and the generalized extreme value distribution is fitted to data from each location to describe the extreme rainfall and to predict its future behavior. They suggested that the Gumbel distribution provides the most reasonable model for four of the five locations considered. Sharma and Singh (2010) analyzed the daily maximum rainfall data of Pantnagar, India for a period of 37 years annually, seasonally, monthly and weakly, and the best fitted probability distribution identified using the least square method among the 16 compared distributions. Deka and Borah (2009) have derived the best fitted probability distribution to describe the annual series of maximum rainfall data for the period 1966 to 2007 of nine distantly located stations in north east India, and they considered only five extreme distributions.

The main objective of the study is to determine the best fit probability distribution for monthly rainfall data of some selected stations in Bangladesh and best fit probability distribution was evaluated on the bases of several goodness of fit test. The resulting information is essential for several

* Corresponding author:

sanjib.stat@gmail.com (Sanjib Ghosh)

Published online at <http://journal.sapub.org/ajms>

Copyright © 2016 Scientific & Academic Publishing. All Rights Reserved

research programmers, agricultural and hydrological studies in Bangladesh.

2. Study Area

The daily rainfall data for the period 1979 to 2013 collected by the Department of Meteorology Government of People's Republic of Bangladesh have been employed in this study. We used long duration data to take appropriate statistical decision. In this study, the period between the months of May to October has been considered as the rainy season or monsoon period but in this period its rains irregularly. Because of this irregular variation we used monthly amount of rainfall data rather than daily or weekly rainfall data. The whole Bangladesh has divided into four regions according to the amount of monthly rainfall. Thus four important meteorological stations names Chittagong, Dhaka, Rajshahi and Sylhet have been selected from the four zones to analysis the rainfall data.

Chittagong: The port city Chittagong is the second largest city in Bangladesh. It is situated in the south-eastern portion of the country and was built on the bank of the Kornofulli River. Its graphical coordinates are $22^{\circ}21'49''$ North, $91^{\circ}48'12''$ East. The total area of the district 5282.98 square kilometer. The average monthly rainfall is 441.58 mm of the rainy season considering the data set 1979 to 2013.

Dhaka: Dhaka is the capital of Bangladesh. It's located at $23^{\circ}42'0''$ North, $90^{\circ}22'30''$ East on the eastern banks of the Buriganga River. The geographical area of Dhaka is 815.85

square kilometers. The average monthly rainfall is 310.24 mm during the period 1979 to 2013.

Rajshahi: Rajshahi is famous for its silk industries. Rajshahi is also well known for its fruits, especially for Mango and Lychee. It's located at $24^{\circ}22'26.54''$ North, $88^{\circ}36'34.23''$ East in the north eastern region of Bangladesh. The average monthly rainfall of Rajshahi during the period 1979 to 2013 is 231.38 mm.

Sylhet: Sylhet is a major city in north-eastern Bangladesh. It's located on the banks of the surma River and is surrounded by the Jaintia, Khasi and Tripura hills. The Sylhet region is well known for its tea gardens and tropical forests. It's located at $24^{\circ}53'30''$ North, $91^{\circ}53'00''$ East in the north eastern region of Bangladesh. The average monthly rainfall of Sylhet during the period 1979 to 2013 is 585.58mm.

3. Methodology

In order to describe the amount of monthly rainfall data, it is necessary to identify the distributions, which best fit to the data. In this study, six continuous probability distribution viz. Normal, Lognormal, Gamma, Weibull, Inverse Gaussain and Generalized Extreme value distribution are considered to test the goodness of fit. The analysis of rainfall data is prepared with the help of EASY FIT and MICROSOFT EXCEL software's.

The probability density function of above distribution are exhibited in the following table 1.

Table 1.

Distribution	Probability density function	Range
Noomal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2 / 2\sigma^2}$	$-\infty \leq x \leq \infty$
Lognormal	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}$	$-0 \leq x \leq \infty$ $-\infty < \mu < \infty, \sigma > 0$
Gamma	$f(x) = \frac{1}{\beta^\alpha \Gamma \alpha} x^{\alpha-1} e^{-x/\beta}$	$0 \leq x \leq \infty$ $\alpha \geq 0, \beta \geq 0$
Weibull	$f(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{x}{\beta}\right)^\alpha\right\}$	$x \geq 0$
Inverse Gaussain	$f(x) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}$	$x > 0, \mu > 0 \text{ and } \lambda > 0$
Generalized Extreme value	$f(x) = \frac{1}{\sigma} \left[1 - k\left(\frac{x-\mu}{\sigma}\right)\right]^{1/k-1} \exp\left[-\left\{1 - k\left(\frac{x-\mu}{\sigma}\right)\right\}^{1/k}\right]$	$-\infty \leq x \leq \infty$ $k \neq 0$

3.1. Method of L-moment

The L-moments (LMOM) were introduced by Sillitto (1951) and comprehensively reviewed by Hosking (1986, 1990) for estimating the parameters of certain statistical distributions. The L-moments are linear functions of the expectations of order statistics and they can be viewed as an alternative system of describing shapes of probability distributions. The main advantage of using the method of LMOM are that the parameter estimates are more reliable (i.e. smaller mean-squared error estimation) and are most robust, and usually computationally more tractable than the conventional moments and maximum likelihood.

Let $X_{1:n}, X_{2:n}, X_{3:n}, \dots, X_{n:n}$ be a sample from a continuous distribution function $F(\cdot)$ with quantile function $Q(F)$ and let $X_{1:n} \leq X_{2:n} \leq X_{3:n} \leq \dots \leq X_{n:n}$ denotes the order statistics. Then r^{th} L-moment λ_r is given by

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(X_{r-k:n}), \quad r=1, 2, \dots$$

The details on the estimation of parameters for each of the aforesaid distributions can be found in Hosking and Wallis (1997).

3.2. Goodness-of-Fit Tests (GOF)

GOF is used to determine the best model among the distributions tested in rain characteristic. The goodness-of-fit test is performed in order to test the following hypothesis:

H_0 : The amount of monthly rainfall data follow the specified distribution

H_1 : The amount of monthly rainfall data do not follow the specified distribution

A couple of goodness-of-fit test have been conducted such as Kolmogorov-Smirnov test, Anderson-Darling test along with the chi-square test at significance level ($\alpha=0.05$) for choosing the best probability distribution (Sharma and Singh, 2010).

3.2.1. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (Chakravart, Laha, and Roy, 1967) is used to decide if a sample comes from a population with a specific distribution.

The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given N ordered data points Y_1, Y_2, \dots, Y_N , the ECDF is defined as

$$E_N = \frac{n(i)}{N}$$

Where $n(i)$ is the number of points less than Y_i and the Y_i are ordered from smallest to largest value. This is a step function that increases by $1/N$ at the value of each ordered data point.

Test Statistic: The Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1 \leq i \leq N} \left[F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right]$$

Where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).

The hypothesis regarding the distributional form is rejected if the test statistic, D , is greater than the critical value obtained from a table.

3.2.2. Anderson –Darling Test

The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data comes from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I, and logistic distributions.

The Anderson-Darling test statistic is defined as

$$A^2 = -N - \frac{1}{N} \sum_{i=1}^N (2i-1) [\ln F(X_i) + \ln(1-F(X_{N-i+1}))]$$

F is the cumulative distribution function of the specified distribution. Note that the Y_i are the ordered data.

The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas have been published (Stephens, 1974) for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A , is greater than the critical value.

3.2.3. Chi-square Test

The Chi-square test assumes that the number of observations is large enough so that the chi-square distribution provides a good approximation as the distribution of test statistic. The Chi-squared statistic is defined as

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Where

O_i = observed frequency

E_i = expected frequency

'i' = number observations (1, 2, ..., k)

Calculated by $E_i = F(X_2) - F(X_1)$

F = the CDF of the probability distribution being tested

The observed number of observation (k) in interval 'i' is computed from equation given below

$$K = 1 + \log_2 n$$

$$n = \text{sample size}$$

This equation is for continuous sample data only and is used to determine if a sample comes from a population with a specific distribution (Sharma and Sing, 2010).

4. Result and Discussion

The statistical parameters for the amount of monthly rainfall data of four selected stations are summarized in Table 2. Where the minimum, maximum, mean, standard deviation (SD), coefficient of variation (CV), skewness and kurtosis are given.

The distribution of monthly rainfall data was found to be positively skewed at the selected station. The monthly rainfall data set indicates that the monthly rainfall was strongly positively skewed for Rajshahi station. It is also observed that the monthly rainfall data set of Rajshahi station was highly leptokurtic and Sylhet station was platykurtic. Coefficient of variation of rainfall around 50-67% in Bangladesh.

The test statistics for Kolmogorov-Smirnov test (D), Anderson-Darling Test (A^2) and Chi-square test (χ^2) for monthly rainfall data were computed for six probability distributions. The probability distribution having the first rank along with their test statistic is presented in Table 3.

Using Kolmogorov-Smirnov test (D), it is observed that generalized **extreme value** distribution provides good fit to the monthly rainfall data during rainy season at Chittagong, Rajshahi and Sylhet station. Only for Dhaka station, the

gamma distribution provides good fit to the monthly rainfall data during rainy season.

Using Anderson-Darling Test (A^2), it is also observed that the Generalized Extreme Value distribution provides good fit Chittagong, Dhaka and Rajshahi station to the monthly rainfall data during rainy season in Bangladesh. Only for Sylhet station provides good fit Gamma distribution to the monthly rainfall data during rainy season.

Finally, Chi-square test has been applied to test the goodness-of-fit. It has been shown that the Generalized Extreme Value distribution provides good fit at Chittagong, Dhaka and Sylhet station to the monthly rainfall data during rainy season in Bangladesh. Only for Rajshahi station provides good fit Normal distribution to the monthly rainfall data during rainy season.

The identified distributions are listed in Table 4 with the estimated parameters for monthly rainfall data. It is also observed that some of the probability distributions have the first rank in both Kolmogorov-Smirnov and Anderson-Darling tests.

The estimated parameters were used to generate random numbers for monthly rainfall data set and the least square method was used for the rainfall analysis.

The best selected probability distribution for monthly rainfall data set are presented in Table 5.

Comparing three goodness-of-fit tests it has been observed that the Generalized Extreme Value distribution provides a good fit of the three selected station only Dhaka station provides good fit Gamma distribution to the monthly rainfall data of during rainy season in Bangladesh.

Table 2. Summary Statistics of the monthly rainfall data for the selected stations during the year (1979-2013)

Descriptive measure \ Station	Chittagong	Dhaka	Rajshahi	Sylhet
Minimum(mm)	5	30	4	29
Maximum(mm)	1457	839	763	1394
Mean	441.58	310.24	231.38	585.58
Standard deviation	292.42	157.53	135.33	303.27
Co-efficient of variation	66.22	50.77	58.49	51.79
Co-efficient of Skewness	0.89	0.75	0.96	0.39
Kurtosis	3.52	3.31	3.46	2.59

Table 3. Station wise first ranked probability distribution using different goodness of fit test

Study Location	First position in ranking of different tests					
	Kolmogorov-Smirnov test		Anderson-Darling Test		Chi-square test	
	Distribution	Statistic	Distribution	Statistic	Distribution	Statistic
Chittagong	GEV	0.0443	GEV	0.467	GEV	7.353
Dhaka	Gamma	0.0354	GEV	0.207	GEV	3.408
Rajshahi	GEV	0.0459	GEV	0.414	Normal	3.682
Sylhet	GEV	0.0357	Gamma	0.056	GEV	3.650

*GEV=Generalized extreme value distribution

Table 4. Parameters estimates of the best fitted distributions

Study Location	Distribution	Parameters (L-moment)
Chittagong	Generalized Extreme value	$\mu=304.9, \sigma=229.2, k=0.019$
Dhaka	Generalized Extreme value Gamma	$\mu=238.9, \sigma=129.5, k=-0.023, \alpha=3.88, \beta=79.99$
Rajshahi	Generalized Extreme value Normal	$\mu=171.9, \sigma=111.6, k=-0.05, \mu=231.4, \sigma=153.3$
Sylhet	Generalized Extreme value Gamma	$\mu=458.6, \sigma=278.4, k=-0.14, \alpha=3.73, \beta=157.1$

Table 5. Best fitted probability distribution

Study Location	Best fit distribution
Chittagong	Generalized Extreme value
Dhaka	Gamma
Rajshahi	Generalized Extreme value
Sylhet	Generalized Extreme value

5. Conclusions

A systematic assessment procedure was applied to evaluate the performance of different probability distribution with view to identifying the best fit probability distribution for monthly rainfall data at four selected stations of Bangladesh. The data showed that the monthly minimum and maximum rainfall at any time at any station ranged from 4 mm (Rajshahi) to 1457 mm (Chittagong) which is obviously indicating a huge range of fluctuation during the period of the study. It was observed that the Generalized Extreme Value distribution provides a good fit of the three selected station only Dhaka station provides good fit Gamma distribution to the monthly rainfall data of during rainy season in Bangladesh. Identifying the distribution amount of monthly rainfall data could have a wide range of applications in agriculture, hydrology, engineering design and climate research.

REFERENCES

- [1] Deka S., and Borah M., (2009). Distribution of Annual Maximum Rainfall Series of North east India, European Water Publications 27/28: 3-14.
- [2] Hirose, H. (1994). Parameter Estimation in the Extreme-Value distributions using the Continuation Method, Information processing society of Japan vol. 35, 9.
- [3] Nadarajah. S., and Choi. D. (2007). Maximum daily rainfall in South Korea. J. Earth Syst. Sci 116, No. 4, pp. 311-320, DOI: 10.1007/s 12040-007-0028-0.
- [4] Nadarajah. S., and Withers, C. S. (2001). Evidence of trend in returns levels for daily rainfall in New Zealand, Journal of Hydrology New Zealand, Volume 39, p.155-166.
- [5] Nadarajah, S. (2005). The exponential Gumbel distribution with climate application, Environmetrics, 17(1), 13-23, DOI: 10.1002/env.739
- [6] Sharma, M.A. and Singh, J.B. (2010). Use of Probability Distribution in rainfall Analysis, New York science Journal, pp.40-49.
- [7] Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, Vol. 69, pp. 730-737.
- [8] Stephens, M. A. (1977). "Goodness of Fit with Special Reference to Tests for Exponentially". Tech. Rep. No. 262, Department of Statistics, Stanford, CA.
- [9] Sillitto, G. P. (1951), Interrelations between certain linear systematic statistics of samples from any continuous population, Biometrika, 38, 377-382.
- [10] Hosking, J.R.M., (1986). The theory of probability weighed moments, Research report RC 12210, IBM Research Division, and Yorktown Heights.
- [11] Hosking, J.R.M., (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc, Ser B, 52(1), 105-124.
- [12] Hosking, J.R.M., Wallis, J.R. (1990). Regional frequency analysis: an approach based on L-moments. University press, Cambridge.
- [13] Shabri, A. and Jemain, A.A. (2007). LQ-Moment: Applications to Generalized extreme value, Journal of Applied Sciences 7(1), 115-120.
- [14] Chakravarti, Laha, and Roy, 1967. Handbook Methods of Applied Statistics. Volume I, John Wiley and Sons, pp. 11-27.