Proceedings of the Pakistan Academy of Sciences:
B. Life and Environmental Sciences 53 (2): 121–128 (2016)

ISSN: 0377 - 2969 (print), 2306 - 1448 (online)

Pakistan Academy of Sciences

Research Article

Wakeby Distribution Modelling of Rainfall and Thunderstorm over Northern Areas of Pakistan

Naeem Sadiq*

Institute of Space and Planetary Astrophysics,
University of Karachi, Karachi 75270, Pakistan

Abstract: Five parametric Wakeby (L-moment) distribution has been compared with Gaussian, Gumbell and Generalized Extreme Value (GEV) for representing the rainfall and thunderstorm activities over four stations in the Northern areas of Pakistan (viz., Astore, Bunji, Garhi Dupatta and Muzaffarabad) for the period of 1961-2010. The test statistics of Kolmogorov-Smirnov test based on the empirical cumulative distribution function revealed that Wakeby modelling is quite suitable to model rainfall frequencies over all the considered stations while the same distribution also shows better model fittings for Astore, Garhi Dupatta and Muzaffarabad with exceptional case of Bunji where Gumbell distribution has a slightly better fit than the Wakeby. The intent of this paper is to present a contemporary statistical view of rainfall and thunderstorm investigation for the considered stations.

Keywords: Wakeby distribution, Gaussian distribution, Probabilistic modelling

1. INTRODUCTION

Thunderstorm (TS) and Rainfall (RF) events can cause severe damage to infrastructure in addition to tragic loss of lives [1]. However, predicting frequency of these parameters within a geographical area is an information problem. With the help of sufficiently long meteorological records, the distribution of frequency for a site may be estimated with a certain degree of accurately. In general, these distributions of the doubtful phenomena aren't apprehended with certainty. If they were known, even then their functional representation would likely have too many parameters to be of much applied usage [2]. The pragmatic issue is the selection of a simple, suitable and reasonable distribution to get a description of the phenomenon under consideration, and then estimate the parameters of that distribution which finally leads to risk estimates with reasonable and acceptable accuracy for the considered problem [3].

The modelling of hydro-meteorological related structures in weather modifications and climate changes monitoring is important and essential [4]. In this relation, rainfall and thunderstorm activities usually in monsoon season has been of great importance. In principle, climatologists have to fit different distributions to hydro-meteorological data to estimate a number return levels of extreme rainfalls [5].

Five parametric Wakeby (W5) distribution can copycat the outlines of a lot of more often than not used skewed distributions, as it comprises of five parameters. Because of its flexible nature it may prove to be sufficiently good fit to observed meteorological data parameter like rainfall etc. Many researchers used this distribution for other frequential analyses regarding different purposes [6-13] and they found this distribution to be more appropriate for the modelling of low flow discharges during flood frequency analyses.

From this place, W5 is used to a great degree in hydro-meteorological applications in a successful manner, especially for the modelling

Received, September 2015; Accepted, May 2016
Email: nsadiq@uok.edu.pk

purposes. Wilks and McKay [14] concluded that W5 furnished the best representations of extreme snowpack water equivalent values.

2. DATA AND METHODOLOGY

The approximate area covered by the Northern areas in Pakistan is 72,496 km^2. The extra tropical quadruple season type is the common type of climate observed over these areas. The range of the rainfall amount varying from 254 to 508 mm usually accompanied with thunderstorm activity [20].

For the approximation of W5 parameters, L moment [15] method was employed. An attempt was made to utilize W5 along with the said method to the TS and RF frequencies of four selected stations. viz.. Astore, Bunji, Garhi Dupatta and Muzaffarabad over Northern areas of Pakistan (Fig. 1). The frequency data of TS and RF for the period 1961-2010 were obtained from the Pakistan Meteorological Department, Government of Pakistan by in-situ observations recorded in accordance with the World Meteorological Organization (WMO) standards. The software MATLAB was used for the statistical purpose in this study.

2.1 Characterizing Statistical Parameters and Probability Distributions

The Gauss distribution is the most important and widely used distribution in many statistical applications. The Probability Density Function (PDF) of the distribution may be defined as

$$F(x) = \frac{\exp\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}{\sigma\sqrt{2\pi}}$$

where μ and σ are the location parameter and standard deviation, respectively. For $\mu = 0$ and $\sigma = 1$, this distribution may be referred as the *standard normal distribution*. The above said scale

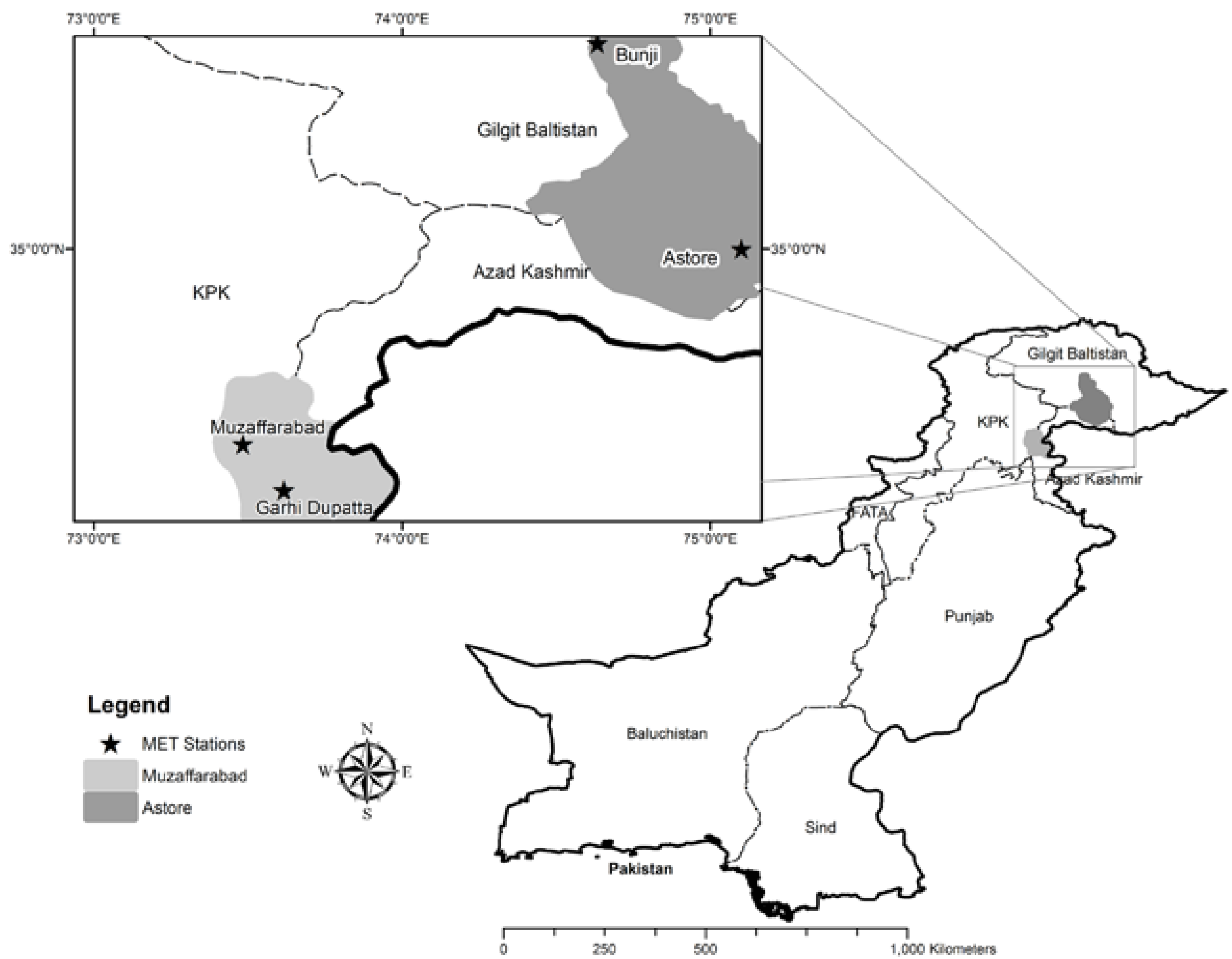

Fig. 1. Location map of selected cities in Northern Areas of Pakistan.

parameter (σ) should be greater than zero accompanied with the location parameter (μ), while the domain restriction is under -∞<x<+∞. In many connections it has been sufficient to use this simpler form since μ and σ simply may be regarded as a shift and scale parameter, respectively.

The Cumulative Distribution Function (CDF) may be defined as:

$$f(x) = \phi(\frac{x-\mu}{\sigma})$$

where ϕ is the Laplace Integral.

Gaussian distribution comprises of a function that assures the probability for any real observation to be fall between any two real limits, as the curve approaches zero on either side. This distribution is not uncommon in the science studies for real valued random variables whose distributions are not known.

If the probability density functions exhibit a characteristic heavy tail then it can be better modelled by W5 distribution, as our results (Fig. 2–5) revealed that this distribution provides markedly a good fit. The PDF of the distribution may be determined by the following method suggested by Johnson et al. [16].

$$f(x) = \frac{[1-F(x)]^{\delta+1}}{\gamma + \alpha[1-F(x)]^{\beta+\delta}}$$

where $F(x)$ is the CDF with α, β, γ, δ shape parameters. The inversed CDF of the W5 may be given by:

$$x(F) = \xi + \frac{\alpha}{\beta}[1-(1-F)]^{\beta} - \frac{\gamma}{\delta}[1-(1-F)]^{-\delta}$$

along with the following conditions or restrictions that must be apply among the various parameters:

$\gamma \geq 0$ and $\alpha + \gamma \geq 0$

If $\alpha = 0$ then $\beta = 0$

If $\gamma = 0$ then $\delta = 0$

either $\alpha \neq 0$ or $\gamma \neq 0$

while parametric domain comprises of:

$\xi \leq x < \infty$ if $\delta \geq 0$ and $\gamma > 0$

$\xi \leq x \leq \xi + \frac{\alpha}{\beta} - \frac{\gamma}{\delta}$ if $\delta < 0$ or $\gamma = 0$

The above parameterization has been explained by Hosking [17] which is unlike from that used by some other authors [8]. In fact, the parameterization [18] presents the W5 distribution as an extension of the Generalized Pareto Distribution (GPD) that provides guesstimates of the more stable parameters under small perturbed data [8]. In order that $x(F)$ in the equation [19] represents an inverse CDF, the conditions $\gamma \geq 0$ and $\gamma + \alpha \geq 0$ should be followed. As W5 is of supple nature, it can be utilized for the description of natural processes accompanied with multiple factors which should or else be modelled through the concoction of more than a few distributions.

External Type Theorem (ETT) is the base of Extreme Value Theory (EVT) which describes that the rescaled sample maxima converge in distribution to a variable having distribution, possibly within any one of the Gumbel, Frechet and Weibull (also called Type I, Type II and Type III) families, respectively. The amalgamation of these three types into a single family of models acquires distribution function in the form:

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

defined on the set $\{z: 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$. This is the GEV (generalized extreme value) family of distributions. The model has three different parameters viz. μ, σ, and ξ known as location, scale and shape parameters, respectively. Type I, II and III (classes of extreme value distribution) corresponds to $\xi = 0$ (i.e. Gumbell distribution model), $\xi > 0$ and $\xi < 0$ respectively. The Gumbel distribution model has also been consider in the study as in many cases it has proved to be more representative of the true values in monsoonal watersheds of the country.

3. RESULTS AND DISCUSSION

The CDF of the RF and TS frequencies were drawn for the Gaussian, Gumbell, GEV and W5 distribution. Careful observation of Fig. 1 and Fig. 2 show that the W5 distribution covers more area than the other distributions on the plotted histograms and hence appears to be the more appropriate fitted distribution for all the rainfall and most of the thunder frequencies over the Northern Areas of Pakistan.

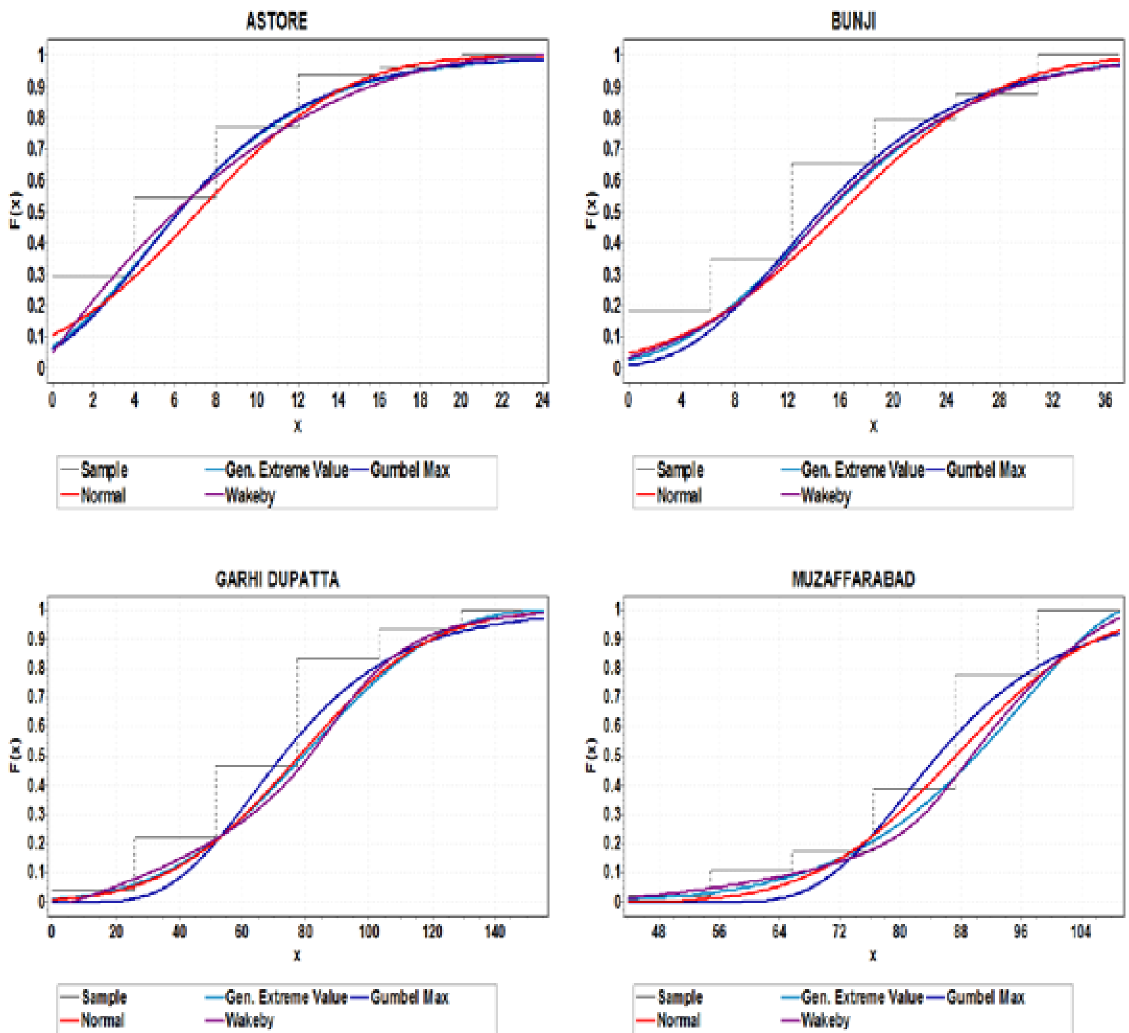

Fig. 2. Cumulative distribution function of selected cities for thunderstorm distribution.

Table 1. Summary of hypothesis testing for thunderstorm.

S. No.		Astore	Bunji	Garhi Dupatta	Muzaffarabad
1	Wakeby	0.0845	0.0984	0.0844	0.0618
	P_{Wa}-Value	0.8543	0.6924	0.8474	0.9902
2	Gumbell	0.1029	0.0880	0.1562	0.1995
	P_{Gu}-Value	0.6520	0.8107	0.1645	0.0442
3	GEV	0.0926	0.1061	0.0955	0.0797
	P_{Ge}-Value	0.7699	0.6013	0.7269	0.9097
4	Normal	0.1264	0.1365	0.0955	0.1289
	P_N-Value	0.3943	0.2936	0.7264	0.3960

3.1 Probability-Probability Plot (P-P Plot)

It depict the plotted values of theoretical CDF verses empirical CDF to observe the fitted accuracy of different distributions to the data. In case of appropriate selection of distribution, p-p plot should be close to linear model. These plots are drawn to illustrate the above goodness-of-fit test results, for the annual rainfall of cities under consideration (Fig. 3-4). It is evident from the

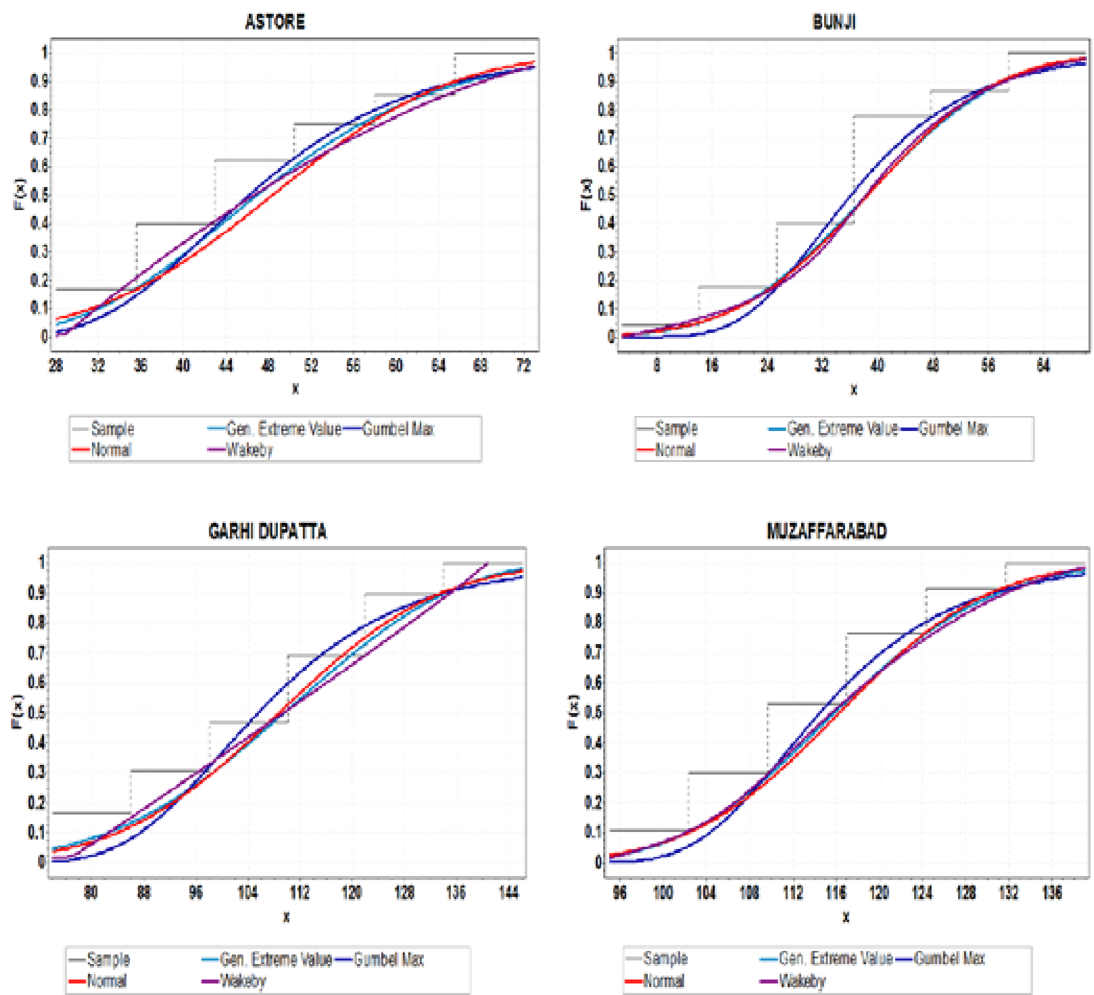

Fig. 3. Cumulative distribution function of selected cities for rainfall distribution.

Table 2. Summary of hypothesis testing for rainfall.

S. No.		Astore	Bunji	Garhi Dupatta	Muzaffarabad
1	Wakeby	0.0867	0.0953	0.0615	0.0617
	P_{Wa}-Value	0.8327	0.7735	0.9871	0.9890
2	Gumbell	0.1219	0.1423	0.1281	0.1170
	PGu-Value	0.4391	0.2929	0.3662	0.5036
3	GEV	0.1108	0.1148	0.0868	0.0766
	PGe-Value	0.5602	0.5546	0.8230	0.9259
4	Normal	0.1229	0.1132	0.0974	0.0803
	P_N-Value	0.4287	0.5727	0.7042	0.8981

figure that the deviation of observed data points from theoretical CDF values is comparatively more in other distributions than in the W5 distribution. As per criteria, the lesser the deviation - the better fitted will be the distribution, therefore, the W5 distribution again appears to be the good fit distribution for the data.

3.2 Testing Hypothesis

3.2.1 Goodness of Fit Test

To compare the 'distance' to threshold value and to measure the distance between the data and the fitted distribution, Kolmogorov-Smirnov

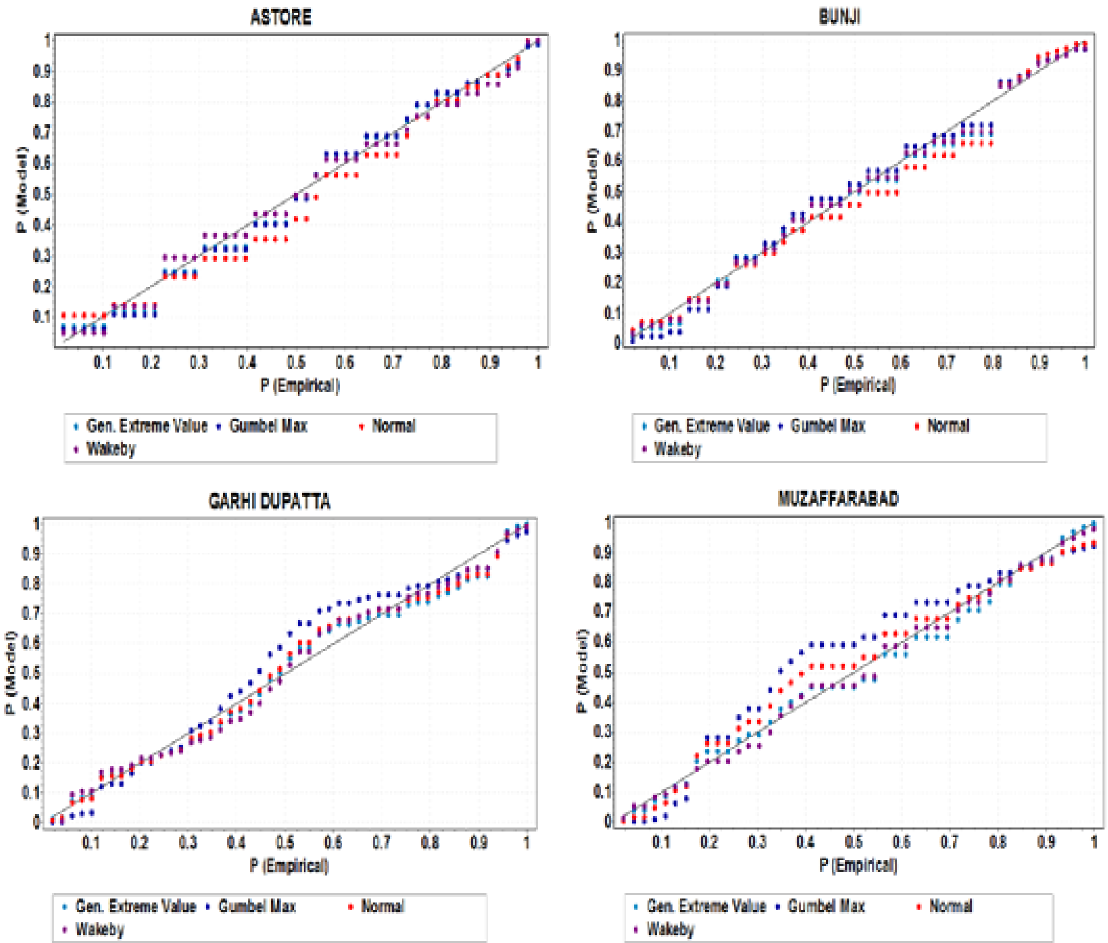

Fig. 4. Probability-probability (P-P) of selected cities for thunderstorm distribution.

goodness-of-fit test is employed. It helps to examine and note the similarities or differences of an observed CDF to the function of an expected cumulative distribution. The distribution is based on the largest vertical difference between the empirical and theoretical CDF via:

$$D = \max_{1 \leq i \leq n} \left\{ F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i) \right\}$$

This test is based on the Empirical Cumulative Distribution Function (ECDF) and helpful to make a decision that if a sample appears from a hypothesized continuous distribution. If a random sample $x_1, \ldots, x_n$ from some distribution with CDF $F(x)$. The ECDF is denoted by

$$F_n(x) = \frac{1}{n}[number\ of\ observations \leq x]$$

The hypothesis concerning the distributional form is not acceptable at the specified chosen significance level (α) if the test statistic, value obtained from, is greater than the critical value (0.19221 in our case). The fixed value of α (0.05 in our case) is used to evaluate the null hypothesis (H_o). Table 1 shows the summary of the goodness-of-fit test for the different distributions. For instance, the estimated *D* for the TS for Astore comes out as 0.08451 which is smaller than the 95th percentile value of 0.19221. Hence, H_o suggests that annual extreme rainfall data of Astore cannot be rejected even at the 5 % level. Likewise, for the other stations estimated *D* values are obtained with a 5 % significance level (Table 1 & 2). Thus, more closely, data of rainfall and most of the thunderstorms for the four stations have been well drawn from the W5 distribution.

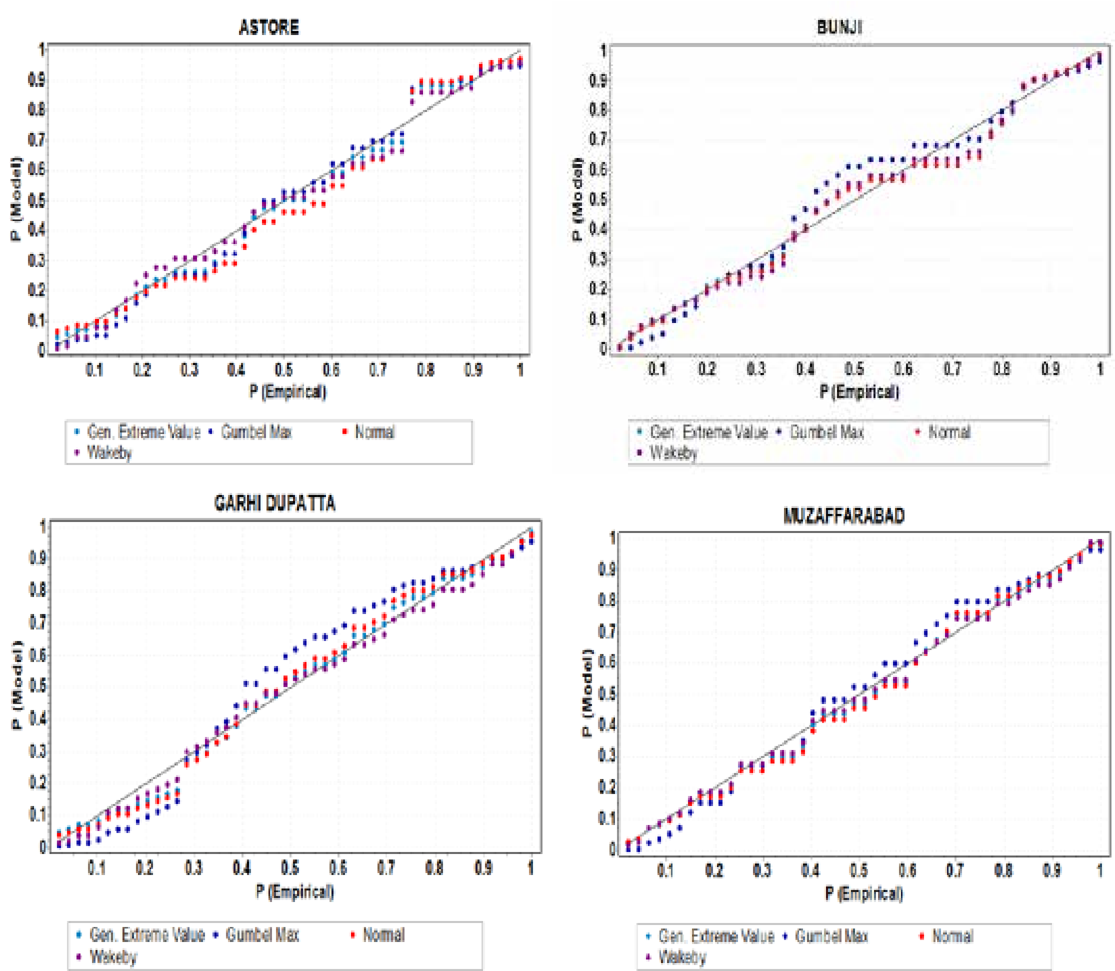

Fig. 5. Probability-probability (P-P) of selected cities for rainfall distribution.

In a similar fashion, the goodness-of-fit test is also applied to the other distributions. The values are also found which are smaller than the 95th percentile mentioned in their respective columns (Table 1 & 2), indicating that data of the four stations may be drawn for this distribution. Since calculated values of *D* for the W5 distribution are markedly smaller compared to the estimated values for the other distribution, hence for our data, the W5 distribution is found to be the more appropriate fitted distribution model with the exceptional case of Bunji's Thunderstorm case for which Gumbell distribution model fit a bit better than W5.

3.2.2 P-Value

Instead of immovable α value, its value is estimated and based on the test statistic. It indicates the limit estimation of the significance level such that the null hypothesis (H_o) will be accepted for all P-values greater than the values of α. As for instance, if P = 0.025, the null hypothesis can be putative at all significance levels which are not greater than P (i.e. 0.01 and 0.02), and rebuffed at advanced levels, be made up out of 0.05 and 0.1. P-values for Wakeby (P_{Wa}), Gumbell (P_{Gu}), GEV (P_{Ge}) and Gauss (P_N) distributions in Table 1 and 2 depict the validation of the goodness of fit test in this regard also.

4. CONCLUSIONS

Gaussian, Wakeby, Gumbell and GEV probabilistic approaches were utilized to model the RF and TS frequency data. Though Gaussian distribution is considered as the good distribution to represent many hydro-meteorological applications, in this study, for given eight data sets (i.e., four for Thunderstorm + four for rainfall), the Wakeby distribution produced markedly better results for seven cases (i.e., three thunderstorm + four rainfall cases). When the outcome of modified nonparametric Kalmogorov Simrnov test were considered, the results yielded by Wakeby distribution, with its parameters estimated by the L-moment, mostly produced ameliorating results as compared to other distributions with their parameters. When the results of the KG statistic for the four highest observed and distributions-predicted values were considered, then again the results produced by Wakeby distribution were chiefly improved than those by others. Also, the dominance of the five-parameter Wakeby distribution over Gaussian was observed. A good understanding of the statistical characteristics of rainfall and thunderstorm activity in the Northern areas of Pakistan may be helpful for the water resources planning and management, together with predictions for flood control.

5. ACKNOWLEDGEMENTS

Financial support provided by the Dean Faculty of Science, University of Karachi for this work is hereby acknowledged.

6. REFERENCES

1. Sadiq, N. & H. Naz. Spatial interpolation of the frequential distribution of thunderstorm and rainfall over Balochistan, Pakistan. *The Nucleus* 51 (3): 329-334 (2014).
2. Sadiq, N. Stochastic modelling of the daily rainfall frequency and amount. *Arabian Journal for Science and Engineering* 39 (7): 5691-5702 (2014).
3. Forbes, C., M. Evans, N. Hastings & B. Peacock. *Statistical Distributions,* 4th ed. Jhon Wiley and Sons, USA, p. 5–9 (2011).
4. Huff, F.A. & J.R. Angel. *Rainfall Distributions and Hydroclimatic Characteristics of Heavy Rainstorms in Illinois*. *Illinois State Water Survey.* Bulletin 70: Champaign, IL, USA (1989).
5. Sadiq, N. & M.S. Qureshi. Estimating Recurrence Intervals of Extreme Rainfall Through a Probabilistic Modeling Approach for Different Urban Cities of Pakistan. *Arabian Journal for Science and Engineering* 39 (1): 191-198 (2014).
6. Houghton, J.C. Birth of a parent: The Wakeby distribution for modeling flood flows. *Water Resources Research* 14 (6): 1105–1109 (1978).
7. Landwehr, J.M., N.C. Matalas & J.R. Wallis. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research* 15 (5): 1055–1064 (1979).
8. Landwehr, J.M., N.C. Matalas & J.R. Wallis. Estimation of parameters and quantiles of Wakeby distributions. *Water Resources Research* 15 (6): 1361– 1372 (1979).
9. Kumar, A. & S. Chander. Statistical flood frequency analysis: an overview. In: *Proceedings of the International Symposium on Flood Frequency and Risk Analyses. Hydrologic Frequency Modelling*. D. Reidel Publishing Company: Dordrecht, The Nederlands, p. 19–35 (1987).
10. Rao, A.R. & P.S. Arora. An empirical study of probability distributions of annual maximum floods. In: *Proceedings of the International Symposium on Flood Frequency and Risk Analyses, Hydrologic Frequency Modelling*. D. Reidel Publishing Company: Dordrecht, The Nederlands, p. 449–465 (1987).
11. Haktanir, T. Comparison of various flood frequency distributions using annual flood peaks data of rivers in Anatolia. *Journal of Hydrology* 136: 1–31 (1992).
12. Haktanir, T. & H.B. Horlacher. Evaluation of various distributions for flood frequency analysis. *Hydrological Sciences* 38 (1): 15–32 (1993).
13. Vogel, R.M., T.A. McMahon & F.H.S. Chiew. Flood flow frequency model selection in Australia. *Journal of Hydrology* 146: 421–449 (1993).
14. Wilks, D S. & M. McKay. Extreme-value statistics for snowpack water equivalent in the north-eastern United States using the cooperative observer network. *Journal of Applied Meteorology* 35: 706–713 (1996).
15. Hosking, J.R.M. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society*. *Series B* 52: 105–124 (1990).
16. Johnson, N.L., S. Kotz & N. Balakrishnan. *Continuous Univariate Distributions*. Vol. 1, Wiley, New York (1994).
17. Hosking, J.R.M. *The Wakeby Distribution. Research Report RC12302*. IBM Research Division, Yorktown Heights, New York (1986).
18. Sen, A. K. Complexity analysis of river-flow time series. *Stochastic Environmental Research and Risk Assessment* 23: 361 (2009).
19. Marsh, T. J. & M. Dale. The UK floods of 2000/01 – a hydrometeorological Appraisal. *CIWEM Journal* 16: 180 (2002).
20. Shamshad, K.M. The *Meteorology of Pakistan*. (First Edition), Royal Book Company, Karachi. p. 207–208 (1988).