551.506.1 : 551.577.2

USE OF PROBABILITY DISTRIBUTIONS FOR THE ANALYSIS OF DAILY RAINFALL DATA OF NORTH EAST INDIA

1. Rainfall information forms the basis for designing water related structure in agriculture planning, in weather modification, in water management and also in monitoring climate changes. The most commonly measured and recorded information on rainfall is a daily value gauged. The equipment for observing daily values are also the simplest type of rain gauges which are fairly inexpensive, easy to maintain and read by local observer with little expertise.

In India, rainfall is generally confined to a brief period, the Indian Summer Monsoon season (June-September), which provide more than 80% of the annual rainfall in the large tract of the country. The main crop, *i.e.*, the kharif is also cultivated during this season. So far no rigorous work barring the work by Medhi (1976) pursued in the North East region of India, considerable effort has been made to graduate the rainfall of different time scales by fitting an appropriate frequency distributions. Barger and Thom (1949) showed that gamma distribution provides good fit to precipitation series in the United States. Gabriel and Neuman (1962) found that two state Markov Chain gave a good description of the occurrences of wet and dry days during rainy period at Tel Aviv.

The best-fit Gamma distributions were found by Simpson (1972) based on rather evaluated rainfall data. Mooley (1973) tested whether a suitable unified probability model exists or not for the distribution of monthly rainfall associated with the Asian Summer Monsoon. He found that gamma distribution is the most suitable probability model from among the Pearsonian models. Gamma distribution was also fitted by Stern and Coe (1984) for modeling rainfall amount. It was claimed that a comprehensive analysis of rainfall data should use daily records and not based on 7, 10 days or monthly total. Sharma (1996) claimed that the probability estimation for the Weibull pdf can be done by analytical integration which was not possible for normal, lognormal and gamma probability distributions. Aksoy (2000) investigated the amounts of daily rainfall and the ascension curve of the hydrograph by using 2-parameter gamma distribution. Muralidharan and Lathika (2005) have analyzed the rainfall occurrence based on modified version of Weibull distribution for two meteorological stations in India.

The two basic objectives of this paper are to judge the goodness of fit of the distributions fitted for daily rainfall observations sampled from seven stations of North East India and to detect the competing distributions.

2. In this study, seven distantly located stations in North East India, *viz*., Imphal, Mohanbari, Guwahati, Cherrapunji, Silcoorie, North Bank, Tocklai (Jorhat) have been selected. The locations of these seven stations of North East India are shown in Fig. 1. The study utilizes daily rainfall data in mm for five years (2001-2005). The series of daily rainfall are taken from Regional Meteorological Centre, Guwahati and Tocklai Experimental Station, Jorhat involving the aforesaid seven stations for the summer monsoon months of June, July, August and September in each year.

The distribution of rainfall on the rainy day should be described well by the model for rainfall amounts [Stern and Coe (1984)]. Daily rainfall data can be characterized by a probability distribution function known from the statistical literature. In this study, the two parameter Gamma distribution, the Left-truncated Normal distribution, 2-parameter Weibull distribution and 2-parameter Lognormal distribution is considered to find the best fitting probability distribution function of the daily rainfall data.

2.1. *Left-truncated normal distribution* **-** The probability density function of a normally distributed random variable x is given by

$$y = f(x/\mu_n, \sigma_n) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2\right\},$$

$$\sigma_n > 0, \; -\infty \le x \le \infty \qquad (1)$$

If the values of x below some value x_L cannot be observed due to censoring or truncation then, the resulting distribution is a left-truncated normal distribution with probability density function $f_{LTN}(x)$ given as :

$$f_{LTN}(x) = \begin{cases} 0, & -\infty \le x \le x_L \\ \dfrac{f(x)}{\int_{x_L}^{\infty} f(x)\mathrm{d}x}, & x_L \le x \le \infty \end{cases} \qquad (2)$$

where $f(x)$ is as defined in Equation (1) and μ_n and σ_n are the parameters of the distribution and are equal to mean $\overline{X}$ and standard deviation $\sqrt{V(X)}$ of the sampling distribution respectively.

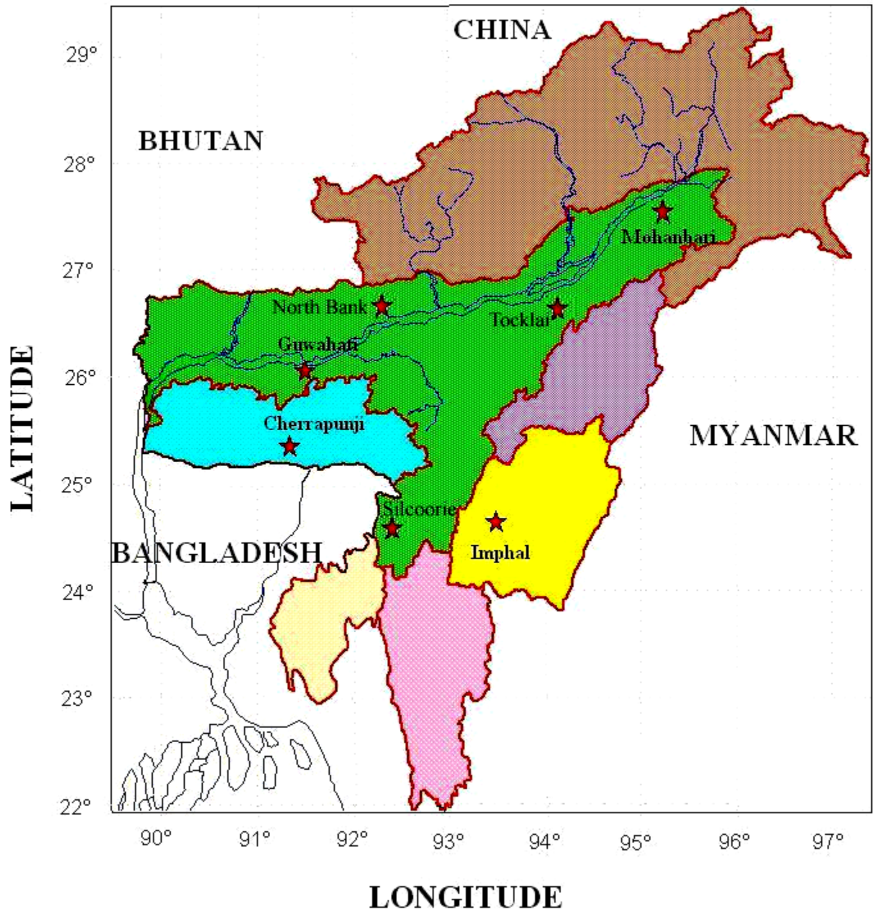

Fig. 1. Locations of seven stations of NE India

2.2. *Lognormal distribution* - The probability density function of the lognormal distribution is given by

$$y = f(x/\mu_l, \sigma_l) = \frac{1}{x\sigma_l\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{z-\mu_l}{\sigma_l}\right)^2\right\}, \quad (3)$$

where $z = \log x$ and μ_l and σ_l are the parameters of the distribution and can be evaluated using the following relationship

$$\hat{\mu}_l = \bar{z} \quad (4)$$

$$\hat{\sigma}_l = \left[n^{-1}\sum_{j=1}^{n}(z_j - \bar{z})^2\right]^{1/2} \quad (5)$$

where $\bar{z} = n^{-1}\sum_{j=1}^{n} z_j$ (assuming that $x_1, x_2, x_3, \ldots. x_n$ are independent random variables each having the same lognormal distribution).

2.3. *Gamma distribution* - Gamma distribution is next to the normal distribution in simplicity and the same time it covers a wide range of skewness. We therefore decided to test the fit of daily rainfall to gamma distribution for which the probability density function is given by

$$f(x/\lambda, \eta) = \frac{1}{\eta^{\lambda}\Gamma(\lambda)} x^{\lambda-1} e^{-x/\eta}, \quad \text{for } x > 0;\ \lambda, \eta > 0 \quad (6)$$

$$= 0, \quad \text{otherwise} \quad (7)$$

TABLE 1

Fitting of probability distributions for daily rainfall data (2001-2005) of Mohanbari during Indian summer monsoon season

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=16.6879 σ_n=21.2957	Lognormal μ_l =1.9568 σ_l =1.5023	Gamma λ =.7039 η=23.7079	Weibull α =.1224 β =.7859
0-14	276	132	300	269	277
14-28	87	143	64	90	85
28-42	34	102	28	42	39
42-56	21	48	15	21	20
56-70	11	15	9	11	10
70-84	4	3	6	6	6
84-98	8	0	4	3	3
98-112	2	0	3	2	2
112-126	0	0	2	1	1
126-140	2	2	14	0	2
Kolmogorov –Smirnov D Statistics		.3226	.0549	.0164	.0086
	χ^2	264.8085	25.3793	8.4726	3.5084
	d.f	3	5	4	4
	p-value	4.09828e-57	0.00012	0.07572	0.47660

TABLE 2

Fitting of probability distributions for daily rainfall data (2001-2005) of Guwahati during Indian summer monsoon

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=13.0294 σ_n=19.6634	Lognormal μ_l =1.5200 σ_l =1.6097	Gamma λ=.5913 η=22.0360	Weibull α =.2012 β =.6952
0-14	282	141	300	273	284
14-28	59	136	45	68	61
28-42	25	81	18	29	25
42-56	12	30	10	13	12
56-70	9	7	6	6	6
70-84	5	1	4	3	3
84-98	0	0	3	2	2
98-112	1	0	2	1	1
112-126	1	0	1	0	1
126-140	2	0	7	1	1
Kolmogorov –Smirnov D Statistics		.3558	.0444	.0235	.0089
	χ^2	246.6116	14.5292	4.1876	1.7047
	d.f	2	4	3	3
	p-value	2.81172e-54	0.00578	0.24191	0.63589

TABLE 3

Fitting of probability distributions for daily rainfall data (2001-2005) of Imphal during Indian summer monsoon

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=10.5100 σ_n=16.3932	Lognormal μ_l =1.3093 σ_l =1.6168	Gamma λ=.5934 η=17.7122	Weibull α =.2321 β =.6976
0-14	329	189	342	321	331
14-28	61	159	43	68	60
28-42	21	67	17	25	22
42-56	9	14	9	10	9
56-70	4	2	5	4	4
70-84	3	0	3	2	2
84-98	1	0	2	1	1
98-112	2	0	2	0	1
112-126	0	0	1	0	0
126-140	1	0	7	0	1
Kolmogorov –Smirnov D Statistics		.3257	.0312	.0177	.0051
	χ^2	196.6883	14.8131	3.9457	.5187
	d.f	1	4	2	2
	p-value	1.10293e-44	0.00510	0.13906	0.77155

TABLE 4

Fitting of probability distributions for daily rainfall data (2001-2005) of Cherrapunji during Indian summer

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=77.7161 σ_n=106.9166	Lognormal μ_l =3.3920 σ_l =1.6223	Gamma λ=.6373 η=121.9516	Weibull α =.0469 β =.7360
0-80	373	190	386	355	367
80-160	74	187	64	96	88
160-240	39	108	27	41	37
240-320	19	36	15	19	17
320-400	13	7	9	9	9
400-480	5	1	6	4	5
480-560	3	0	4	2	3
560-640	1	0	3	1	2
640-720	1	0	2	1	1
720-800	1	0	13	1	0
Kolmogorov –Smirnov D Statistics		.3464	.0337	.0334	.0151
	χ^2	328.652	23.1995	8.2741	4.4465
	d.f	2	5	3	4
	p-value	4.30651e-72	0.00031	0.04067	0.34893

TABLE 5

Fitting of probability distributions for daily rainfall data (2001-2005) of Silcoorie during summer monsoon

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=15.5706 σ_n=18.0755	Lognormal μ_l =2.1468 σ_l =1.1634	Gamma λ=.9681 η=16.0844	Weibull α =.0771 β =.9434
0-14	291	153	303	272	276
14-28	97	163	83	108	104
28-42	27	98	31	45	43
42-56	21	34	15	18	18
56-70	9	6	8	8	8
70-84	7	1	5	3	4
84-98	2	0	3	1	2
98-112	1	0	2	1	1
112-126	0	0	1	0	0
126-140	1	1	5	0	0
Kolmogorov –Smirnov D Statistics		.3019	.0257	.0423	.0328
	χ^2	225.6039	11.3778	17.4726	10.1506
	d.f	2	5	3	3
	p-value	1.02503e-49	0.0444	0.0002	0.0173

TABLE 6

Fitting of probability distributions for daily rainfall data (2001-2005) of North Bank during summer monsoon

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=17.8168 σ_n=21.6805	Lognormal μ_l =2.0529 σ_l =1.4704	Gamma λ=.7268 η=24.5152	Weibull α =.1088 β =.8044
0-20	269	169	297	277	282
20-40	81	155	51	76	71
40-60	30	64	20	29	27
60-80	16	12	10	12	11
80-100	3	1	6	5	5
100-120	2	0	4	2	2
120-140	0	0	3	1	1
140-160	0	0	2	0	1
160-180	0	0	1	0	0
180-200	1	1	8	0	2
Kolmogorov –Smirnov D Statistics		.2481	.0702	.0199	.0331
	χ^2	117.1346	42.9118	2.4278	6.9138
	d.f	1	4	2	3
	p-value	12.6824e-27	1.0793e-08	0.2970	0.0747

TABLE 7

Fitting of probability distributions for daily rainfall data (2001-2005) of Tocklai during summer monsoon

Rainfall (mm)	Observed frequencies	Theoretical frequencies			
		Truncated normal μ_n=13.0146 σ_n=17.1076	Lognormal μ_l =1.6966 σ_l =1.4449	Gamma λ=.6957 η=18.7077	Weibull α =.1556 β =.7732
0-12	271	135	292	261	270
12-24	69	140	58	81	76
24-36	33	86	22	34	31
36-48	13	42	14	19	18
48-60	12	9	7	8	8
60-72	12	1	5	4	4
72-84	0	0	3	2	2
84-96	1	0	2	1	1
96-108	1	0	2	1	1
108-120	1	0	8	2	2
Kolmogorov –Smirnov D Statistics		.3305	.0511	.0249	.0206
	χ^2	254.6012	32.2358	8.5851	6.6664
	d.f	2	5	3	3
	p-value	5.17681e-56	5.3355e-06	0.0353	0.0833

where η and λ are scale and shape parameters, respectively. The exponential distribution is a particular case when λ = 1. The maximum likelihood estimates $\hat{\lambda}$ and $\hat{\eta}$ of the parameters can be obtained by solving the equations

$$n^{-1}\sum_{j=1}^{n}\log X_j = \log\hat{\eta} + \psi(\hat{\lambda}), \tag{8}$$

$$\overline{X} = \hat{\lambda}\hat{\eta} \tag{9}$$

where $\overline{X}$ is the arithmetic mean of the rainfall amounts $x_1, x_2, x_3, \ldots x_n$ and

$$\psi(\hat{\lambda}) = \frac{\partial \log\Gamma(\hat{\lambda})}{\partial\hat{\lambda}} \text{ = di-gamma function} \tag{10}$$

2.4. *Weibull distribution* - The probability density function of the two parameter Weibull distribution is given by

$$y = f(x/\alpha, \beta) = \alpha\,\beta\, x^{\beta-1}\exp(-\alpha x^{\beta}), x > 0 \tag{11}$$

where α be the scale parameter and β be the shape parameter of the distribution. The maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β respectively satisfy the equations

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i^{\hat{\beta}}\log x_i}{\sum_{i=1}^{n} x_i^{\hat{\beta}}} - \frac{\sum_{i=1}^{n}\log x_i}{n} \tag{12}$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} x_i^{\hat{\beta}}} \tag{13}$$

The value of $\hat{\beta}$ has to be obtained from Eqn. (11) and then used in Eqn. (12) to obtain $\hat{\alpha}$.

2.5. *Test for goodness of fit* - The test applied for judging the goodness of fit of the distributions for rainfall series are namely Chi-square test and Kolmogorov-Smirnov test. However, Chi-square test has been carried out with caution considering its limitations in application

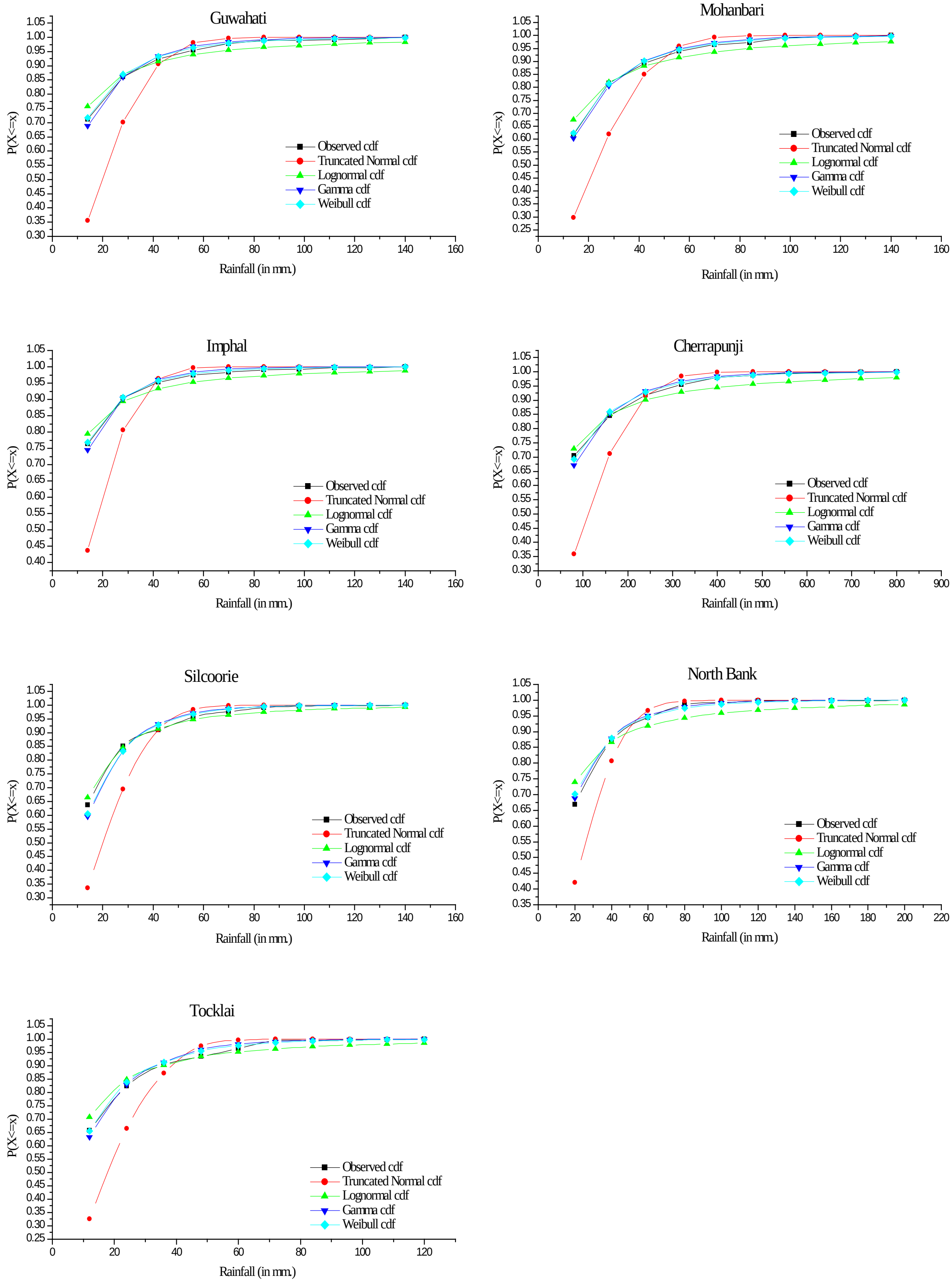

Fig. 2. Curve for probability distribution functions

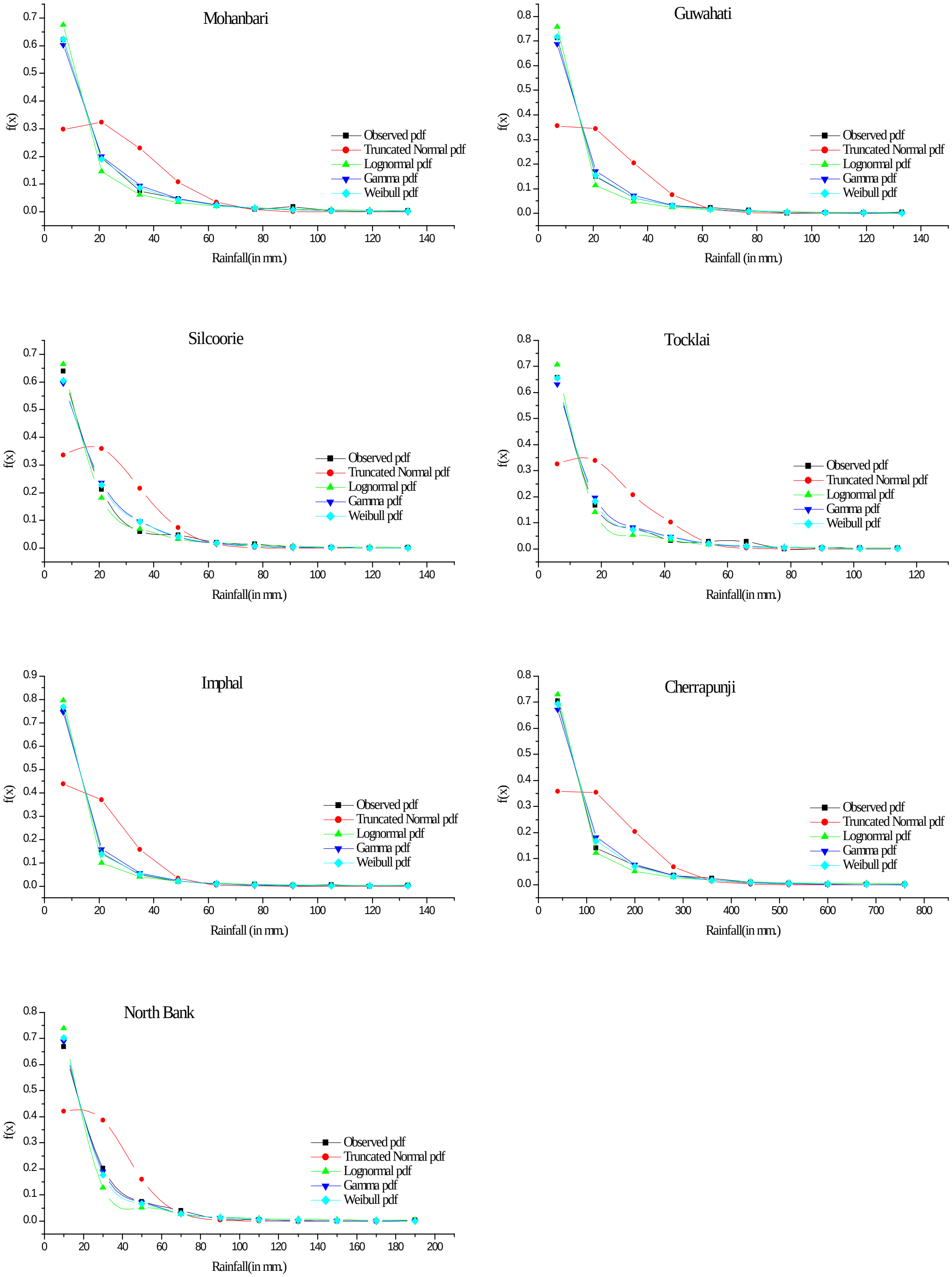

Fig. 3. Curve for probability density functions

and the suggestion made by Massey (1951). Massey showed that Kolmogorov-Smirnov test treats individual observation separately leading to no loss of information in grouping while loss of information in Chi-square procedure is large. Pal (1998) mentioned that the Chi square test's sensitivity to very small cell frequencies make itself unsuitable when expected frequencies work out at less than 5 in 20 per cent of the cells. In the present case it is found that more than 50% of the cell frequencies are less than 5. Also according to Keeping (1962), Kolmogorov Smirnov test can be applied in situations where the theoretical distribution function is continuous. Here also the theoretical distribution functions considered are continuous since the parameters are positive and *x* can assume values greater than zero. The test statistics used is

$$D_n = \max \left| S_n(x) - F(x) \right|$$

where $S_n(x)$ and $F(x)$ are empirical and theoretical distribution functions, respectively. The distribution of D_n is independent of $F(x)$. The theoretical distribution function however, has to be completely specified. In this study the theoretical distribution function have been calculated by using the estimated parameters of the distribution in each case. The significance of a critical value of D_n depends on *n*, the no. of observations. If *n* is over 35, the critical values of *D* at .05 level of significance can be determined by the formula $1.36/\sqrt{n}$. Any D_n equal to or greater than $1.36/\sqrt{n}$ will be significant at .05 level (two tailed test).

3. A day with rainfall of more than 0 mm. or a trace be designated as a rainy day and with no rainfall as dry. After defining a rainy day, it is necessary to determine the amount of rainfall on such a day. In the present study, different distributions are considered as the probability distribution function of the daily rainfall data. The parameters for each distribution are estimated by maximum likelihood method from the daily rainfall data for each station separately and are provided in Tables 1-7. The tables also include the observed frequencies, expected frequencies obtained from the different fitted distributions. The values of Kolmogorov- Smirnov *D*-statistics, values of χ^2 along with degrees of freedom and the corresponding *p*-value for χ^2 are also provided as an evidence in support of goodness of fit.

The Chi-square test of goodness of fit is applied to daily rainfall. The no. of class intervals was found to be 10 over which the computations were done. Further more than 50% of the cell frequencies were found to be less than 5 for almost every cases. Accordingly, Kolmorgorov-Smirnov test was applied in all cases following by the suggestion made by Pal (1998). Barring Truncated Normal distribution, other distributions, *viz*., Log-normal, Gamma, and Weibull have been found satisfactory to model the rainfall series as evidenced by Kolmogorov-Smirnov test.

In order to confirm the goodness of fit for the above three distributions we additionally applied graphical plots of theoretical and observed cumulative distribution functions. The estimation of the cumulative distribution functions, $S_n(x) = P\,(X \le x)$ for various pre assigned values of *x* for each distribution, *viz*., normal, log-normal, Gamma, and Weibull distribution were calculated and graphs were drawn taking probabilities as ordinate and rainfall amount as abscissa (c.f. Fig. 2.). Graphic plots for pdf have been also done in Fig. 3. The Computations have been carried out in the workstation Matlab 7.0.

The following salient features have been revealed from the goodness of fit tests and graphs:

(*i*) In general, Truncated Normal distribution appears to poorly represent the distribution of daily rainfall as evidenced by the tests and evinced by the graphs for pdf and cdf.

(*ii*) The Gamma and Weibull pdf can be regarded to compete with each other as both of them preserve the 'sigmoid' shape of the observed cumulative distribution of daily rainfall series. Also it is seen that Gamma and Weibull pdf are quite close to observed pdf plot.

(*iii*) Log-normal distribution, although accepted to be well fitted on the basis of Chi-square and Kolmogorov-Smirnov test, does not seem to compete with Gamma and Weibull distribution and also is observed to be quite a distance from the observed plot.

4. The following conclusions are drawn on the basis of the results and analysis made in this study.

The Gamma and Weibull distributions are observed to be competing each other and both are very close to the observed distributions. It is well evidenced by the graphic plots animated on the basis of cdf and pdf. So far as goodness of fit of these two distributions are concerned, they are judged to be well fitted as evidenced by chi-square and Kolmogorov-Smirnov test.

5. The authors wish to thank the anonymous referees for carefully reading the manuscript and their valuable comments and suggestions that resulted in improving the presentation.

References

Aksoy, H., 2000, "Use of Gamma distribution in hydrological analysis", *Turk J. Engi. Environ Sci*., **24**, 419-428.

Barger, Gerald L. and Thom, Herbert, C. S., 1949, "Evaluation of Drought Hazard," *Agronomy Journal*, **41**, 11, Geneva, N. Y., 519-526.

Gabriel, K. R. and Neumann, J., 1962, "A Markov Chain model for daily occurrence at Tel Aviv", *Quart. J. R. Met. Soc.*, **88**, 90-95

Keeping, E. S., 1962, "Introduction to Statistical Inference", D. Van Nostrand Co., Inc., Princeton, N.J., 451.

Massey J. Frank, Jr., 1951, "The Kolmogorov Test for Goodness of Fit". *JASA*, **46**, 253, Washington, D.C., 68-78.

Mooley, D. A., 1973, "Gamma Distribution Probability Model for Asian summer monsoon monthly rainfall". *Mon. Wea. Rev.*, **101**, 2, 160-176.

Muralidharan, K. and Lathika, P., 2005, "Statistical modeling of rainfall data using modified Weibull distribution", *Mausam*, **56**, 4, 765-770.

Medhi, J., 1976, "A Markov Chain for the occurrence of wet and dry days", *Indian J. Met. Hydro. & Geophys*., **27**, 431-435.

Pal, S. K., 1998, "Statistics for Geoscientists: Techniques and Applications" Concept Publishing Company, New Delhi, 383-400.

Sharma, T. C., 1996, "A Markov-Weibull Rain-sum Model for designing rain water catchments systems", *Water Resources Management*, **10**, 147-162.

Simpson, J., 1972, "Use of the Gamma Distribution in single cloud rainfall analysis", *Mon. Wea. Rev.*, **100**, 4, 309-312.

Stern, R. D. and Coe, R., 1984, "A model fitting analysis of daily rainfall data", *Journal of the Royal Statistical Society*. Series A (General), **147**, 1, 1-34.

S. DEKA
S. C. KAKATY*
M. BORAH

Department of Mathematical Sciences, Napaam, Tezpur, Assam, India
**Department of Statistics, Dibrugarh University, Dibrugarh, Assam, India*
(*Received 20 March 2007, Modified 7 April 2008*)
e mails : surobhi@tezu.ernet.in;
mborah@tezu.ernet.in