Applied Water Science (2018) 8:145
https://doi.org/10.1007/s13201-018-0789-4

ORIGINAL ARTICLE

Multimodel response assessment for monthly rainfall distribution in some selected Indian cities using best-fit probability as a tool

Anumandla Sukrutha[1] · **Sristi Ram Dyuthi**[1] · **Shantanu Desai**[2]

Received: 12 March 2018 / Accepted: 8 August 2018 / Published online: 16 August 2018

Abstract

We carry out a study of the statistical distribution of rainfall precipitation data for 20 cites in India. We have determined the best-fit probability distribution for these cities from the monthly precipitation data spanning 100 years of observations from 1901 to 2002. To fit the observed data, we considered 10 different distributions. The efficacy of the fits for these distributions was evaluated using four empirical nonparametric goodness-of-fit tests, namely Kolmogorov–Smirnov, Anderson–Darling, Chi-square test, Akaike information criterion, and Bayesian information criterion. Finally, the best-fit distribution using each of these tests were reported, by combining the results from the model comparison tests. We then find that for most of the cities, generalized extreme value distribution or inverse Gaussian distribution most adequately fits the observed data.

Keywords Rainfall statistics · KS test · Anderson–Darling test · AIC · BIC

Introduction

Establishing a probability distribution that provides a good fit to the monthly average precipitation has long been a topic of interest in the fields of hydrology, meteorology, agriculture (Fisher 1925). The knowledge of precipitation at a given location is an important prerequisite for agricultural planning and management. Rainfall is the main source of precipitation. Studies of precipitation provide invaluable knowledge about rainfall statistics. For rain-fed agriculture, rainfall is the single most important agro-meteorological variable influencing crop production (Wallace 2000; Rockström et al. 2003). In the absence of reliable physically based seasonal forecasts, crop management decisions and planning have to rely on statistical assessment based on the analysis of historical precipitation records. It has been shown by Fisher (1925) that the statistical distribution of rainfall is more important than the total amount of rainfall for the yield of crops. Therefore, detailed statistical studies of rainfall data for a variety of countries have been carried out for more than 70 years along with fits to multiple probability distribution (Ghosh et al. 2016; Sharma and Singh 2010; Nguyen et al. 2002). We recap some of these studies for stations, both in India, as well as those outside India.

Mooley and Appa Rao (1970) first carried out a detailed statistical analysis of the rainfall distribution during southwest and northeast monsoon seasons at selected stations in India with deficient rainfall, and found that the gamma distribution provides the best fit. Stephenson et al. (1999) showed that the outliers in the rainfall distribution for the summers of 1986–1989 throughout India can be well fitted by the gamma and Weibull distributions. Deka et al. (2009) found that the logistic distribution is the optimum distribution for the annual rainfall distribution for seven districts in northeast India. Sharma and Singh (2010) found, based on daily rainfall data for Pantnagar spanning 37 years, that the lognormal and gamma distribution provide the best-fit probability distribution for the annual and monsoon months, whereas the generalized extreme value provides the best fit after considering only the weekly data. Most recently, Kumar et al. (2017) analyzed the statistical distribution of rainfall in Uttarakhand, India, and found that the Weibull distribution performed the best. However, one caveat with

✉ Shantanu Desai
shantanud@iith.ac.in

Anumandla Sukrutha
ee14btech11002@iith.ac.in

Sristi Ram Dyuthi
ee14btech11031@iith.ac.in

[1] Department of Electrical Engineering, IIT Hyderabad, Kandi, Telangana 502285, India

[2] Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India

some of the above studies is that only a handful of distributions were considered for fitting the rainfall data, and sometimes no detailed model comparison tests were done to find the most adequate distribution.

A large number of statistical studies have similarly been done for rainfall precipitation data for stations outside India. For brevity, we only mention a few selected studies to illustrate the diversity in the best-fit distribution found from these studies. In Costa Rica, normal distribution provided the best fit to the annual rainfall distribution (Waylen et al. 1996). A generalized extreme value distribution has been used for Louisiana (Naghavi and Yu 1995). Gamma distribution provided the best fit for rainfall data in Saudi Arabia (Abdullah and Al-Mazroui 1998), Sudan (Mohamed and Ibrahim 2015) and Libya (Şen and Eljadid 1999). Mahdavi et al. (2010) studied the rainfall statistics for 65 stations in the Mazandaran and Golestan provinces in Iran and found that the Pearson and log-Pearson distribution provide the best fits to the data. Nadarajah and Choi (2007) found that Gumbel distribution provides the most reasonable fit to the data in South Korea. Ghosh et al. (2016) found that the extreme value distribution provides the best fit to the Chittagong monthly rainfall data during the rainy season, whereas for Dhaka, the gamma distribution provides a better fit.

Therefore, we can see from these whole slew of studies that no single distribution can accurately describe the rainfall distribution. The selection depends on the characteristics of available rainfall data as well as the statistical tools used for model selection.

The main objective of the current study is to complement the above studies and to determine the best-fit probability distribution for the monthly average precipitation data of 20 selected stations throughout India, using multiple goodness-of-fit tests.

Datasets and methodology

The datasets employed here for our study span a 100-year period from 1901 to 2002, and is based on records collected by the Indian Meteorological Department. This data can be downloaded from http://www.indiawaterportal.org/met_data/. From these, we selected 20 stations, covering the breadth of the country for our study. The stations used for this study are Gandhinagar, Guntur, Hyderabad, Jaipur, Kohima, Kurnool, Patna, Aizawl, Bhopal, Ahmednagar, Cuttack, Chennai, Bangalore, Amritsar, Guntur, Lucknow, Kurnool, Jammu, Delhi, and Panipat. The location of these stations on a map of India is shown in Fig. 1. Detailed rainfall statistics for each of these stations can be found in Table 2.

The list of probability distributions considered for fitting the rainfall data includes: gamma, Fisher, inverse Gaussian,

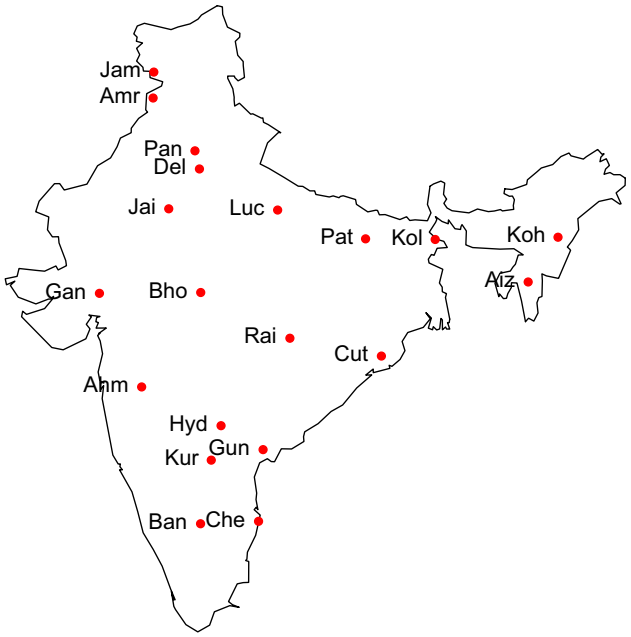

Fig. 1 Map showing location of various stations throughout India for which rainfall statistics and best-fit distributions were obtained. Each red point represents a station and next to it we show its first three letters. The full names of the cities can be found in Table 2. This plot has been made with the `ggplot` (Kahle and Wickham 2013) data visualization package in the `R` programming language, where "gg" in `ggplot` is an abbreviation for "Grammar of graphics"

normal, Student's *t*-, lognormal, generalized extreme value, Weibull and beta distributions. The mathematical expressions for the probability density functions of these distributions can be found in Table 1, and have been adapted from VanderPlas et al. (2012), Ghosh et al. (2016). All of these distributions have been previously used for similar studies (eg. Ghosh et al. 2016; Sharma and Singh 2010) and the other references listed in the introductory section). For each station, we find the best-fit parameters for each of these probability distribution using maximum-likelihood analysis. To select the best-fit distribution for a given station, we then use multiple model comparison techniques to rank each distribution for every city. We now describe the model comparison techniques used.

Model comparison tests

We use multiple model comparison methods to carry out hypothesis testing and select the best distribution for the precipitation data.

For this purpose, the goodness-of-fit tests used include nonparametric distribution-free tests such as Kolmogorov–Smirnov test, Anderson–Darling test, Chi-square test, and information-criterion tests such as Akaike and Bayesian information criterion. For each of the probability

Table 1 Probability density functions of different distributions used to fit the rainfall data. VanderPlas et al. (2012), Ghosh et al. (2016)

Distribution	Probability density function
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$
Lognormal	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp - \frac{(\ln x-\mu)^2}{2\sigma^2}$
Gamma	$f(x) = \frac{1}{\theta_k} \frac{x^{k-1}\exp(-x/\theta)}{\Gamma(k)}$
Inverse Gaussian	$f(x) = \frac{\lambda}{2\pi x^3}^{0.5} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x}$
GEV	$f(x) = \frac{1}{\sigma}\left[1 - k\frac{x-\mu}{\sigma}\right]^{1/k-1} \exp\left[-(1-k\frac{x-\mu}{\sigma})\right]^{1/k}$
Gumbel	$f(x) = 1/\beta \exp(-z + \exp(-z))), z = \frac{x-\mu}{\beta}$
Student's *t*	$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma(v/2)}\left(1 + \frac{x^2}{v}\right)^{\frac{-v-1}{2}}$
Beta	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}$
Weibull	$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$
Fisher	$\frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{xB\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$

distributions, we find the best-fit parameters for each of the stations using least-squares fitting and then carry out each of these tests. We now describe these tests.

Kolmogorov–Smirnov test

The Kolmogorov–Smirnov (K–S) test (VanderPlas et al. 2012) is a nonparametric test used to decide if a sample is selected from a population with a specific distribution. The K–S test compares the empirical distribution function (ECDF) of two samples. Given N ordered data points $y_1, y_2, ..., y_N$, the ECDF is defined as

$$E_N = n(i)/N, \tag{1}$$

where $n(i)$ indicates the total number of points less than y_i, after sorting the y_i in increasing order. This is a step function, whose value increases by $1/N$ for each sorted data point.

The K–S test is based on the maximum distance (or supremum) between the empirical distribution function and the normal cumulative distributive function. An attractive feature of this test is that the distribution of the K–S test statistic itself does not depend on the statistics of the parent distribution from which the samples are drawn. Some limitations are that it applies only to continuous distributions and tends to be more sensitive near the center of the distribution than at the tails.

The Kolmogorov–Smirnov test statistic is defined as:

$$\max_{1 \leqslant i \leqslant N}\left(F(y_i) - \frac{i-1}{N}, \frac{i}{N} - F(y_i)\right), \tag{2}$$

where F is the cumulative distribution function of the samples being tested. If the probability that a given value of D

Table 2 Summary statistics of monthly precipitate data for the selected stations during the years (1901–2002). We note that all units of dimensional quantities are in mm

	Min.	Max.	Mean	SD	Coeff. of variation	Coeff. of skewness	Kurtosis
Kohima	0	802.43	196.33	177.67	0.91	0.77	− 0.24
Jaipur	0	517.61	48.6	83.53	1.72	2.28	5.26
Kolkata	0	892.15	132.15	148.63	1.13	1.31	1.474
Raipur	0	635.98	105.38	140.33	1.33	1.33	0.72
Gandhinagar	0	694.2	56.42	105.18	1.86	2.33	5.36
Hyderabad	0	544.26	70.06	89.41	1.28	1.53	2.19
Aizawl	0	1065.92	227.2	221.48	0.98	0.8	− 0.311
Bhopal	0	725.72	89.53	140.91	1.57	1.73	2.18
Ahmednagar	0	611.13	70.73	96.63	1.37	1.58	2.33
Cuttack	0	506.19	106.32	115.32	1.09	0.91	− 0.34
Chennai	0	768.91	96.89	118.27	1.22	1.99	4.82
Bangalore	0	360.95	69.89	68.66	0.98	1.08	0.78
Patna	0	534.69	90.96	121.9	1.34	1.39	0.9
Amritsar	0	416.06	39.16	59.15	1.51	2.61	8.02
Guntur	0	438.45	65.66	74.58	1.14	1.44	2.24
Lucknow	0	619.08	74.85	113.6	1.52	1.76	2.43
Kurnool	0	374.53	45.19	53.93	1.19	1.85	4.69
Jammu	0	704.43	60.88	83.41	1.37	2.59	8.35
Delhi	0	511.54	47.45	80.67	1.7	2.47	6.58
Panipat	0	463.83	43.58	69.103	1.59	2.33	5.87

is very small (less than a certain critical value, which can be obtained from tables), we can reject the null hypothesis that the two samples are drawn from the same underlying distributions at a given confidence level.

Anderson–Darling test

The Anderson–Darling test (VanderPlas et al. 2012) is another test (similar to K–S test), which can evaluate whether a sample of data came from a population with a specific distribution. It is a modification of the K–S test, and gives more weight to the tails compared to the K–S test. Unlike the K–S test, the Anderson–Darling test makes use of the specific distribution in calculating the critical values. This has the advantage of allowing a more sensitive test. However, one disadvantage is that the critical values must be calculated separately for each distribution. The Anderson–Darling test statistic is defined as follows (VanderPlas et al. 2012):

$$A^2 = -N - \sum_{i=1}^{N} \frac{(2i-1)}{N} \left[\log F(y_i) + \log(1 - F(y_{N+1-i}))\right] \tag{3}$$

where F is the cumulative distribution function of the specified distribution and y_i denotes the sorted data. The test is a one-sided test and the hypothesis that the data are sampled from a specific distribution is rejected if the test statistic, A, is greater than the critical value. For a given distribution, the Anderson–Darling statistic may be multiplied by a constant (depending on the sample size, n). These constants have been tabulated by Stephens (1974).

Chi-square test

The Chi-square test (Cochran 1952) is used to test if a sample of data is obtained from a population with a specific distribution. An attractive feature of the Chi-square goodness-of-fit test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The Chi-square goodness-of-fit test is usually applied to binned data. The Chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson distributions. The Kolmogorov–Smirnov and Anderson–Darling tests can only be applied to continuous distributions. For the Chi-square goodness-of-fit computation, the data are subdivided into k bins and the test statistic is defined as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{4}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l)), \tag{5}$$

where F is the cumulative distribution function for the distribution being tested, Y_u is the upper limit for class i, Y_l is the lower limit for class i, and N is the sample size.

This test is sensitive to the choice of bins. There is no optimal choice for the bin width (since the optimal bin width depends on the distribution). For our analysis, since there were a total of 1224 data points, we have chosen 100 bins, so that there were sufficient data points in each bin. For the Chi-square approximation to be valid, the expected frequency of events in each bin should be at least five. The test statistic follows, approximately, a Chi-square distribution with $(k - c)$ degrees of freedom, where k is the number of non-empty cells, and c is the number of estimated parameters (including location, scale and shape parameters) for the distribution + 1. Therefore, the hypothesis that the data are from a population with the specified distribution is rejected if:

$$\chi^2 \geqslant {\chi_{1-\alpha,k-c}}^2, \tag{6}$$

where ${\chi_{1-\alpha,k-c}}^2$ is the Chi-square critical value with $k - c$ degrees of freedom and significance level α.

AIC and BIC

The Akaike information criterion (AIC) (Liddle 2004; Kulkarni and Desai 2017) is a way of selecting a model from an input set of models. It can be derived by an approximate minimization of the Kullback–Leibler distance between the model and the truth. It is based on information theory, but a heuristic way to think about it is as a criterion that seeks a model, which has a good fit to the truth with very few parameters.

It is defined as (Liddle 2004):

$$\mathrm{AIC} = -2\log(\mathcal{L}) + 2K \tag{7}$$

where $\mathcal{L}$ is the likelihood which denotes the probability of the data given a model, and K is the number of free parameters in the model. AIC scores are often shown as ΔAIC scores, or difference between the best model (smallest AIC) and each model (so the best model has a ΔAIC of zero).

The bias-corrected information criterion, often called AICc, takes into account the finite sample size, by essentially increasing the relative penalty for model complexity with small datasets. It is defined as (Kulkarni and Desai 2017):

$$\mathrm{AICc} = -2\log(\mathcal{L}) + 2\frac{K(K+1)}{N-K-1} \tag{8}$$

where $\mathcal{L}$ is the likelihood and N is the sample size. For this study, we have used AICc for evaluating model efficacy.

Bayesian information criterion (BIC) is also an alternative way of selecting a model from a set of models. It is an approximation to Bayes factor between two models. It is defined as (Liddle 2004):

$$\mathrm{BIC} = -2\log(\mathcal{L}) + K\log(N) \tag{9}$$

When comparing the BIC values for two models, the model with the smaller BIC value is considered better. In general, BIC penalizes models with more parameters more than AICc does.

Results and discussion

The summary statistics for the amount of monthly precipitation data for the above-mentioned stations are summarized in Table 2, where the minimum, maximum, mean, standard deviation (SD), coefficient of variation (CV), skewness and kurtosis are shown. The monthly rainfall dataset indicates that the monthly rainfall was strongly positively skewed for Gandhinagar, Jaipur, Amritsar, Delhi and Panipat stations. Aizawl, Kohima and Cuttack show negative values of kurtosis. The distributions listed above are fitted for each of the selected locations. For brevity in this manuscript, we show the plots for only four cities. These can be found in Figs. 2, 3, 4 and 5, which illustrate the fitted distribution for Kurnool, Hyderabad, Jammu and Patna. These plots are mainly for illustrative purposes. More detailed information about the rainfall distribution can be gleaned from the statistical fits to different distributions. Similar plots for the remaining stations have been uploaded on a Google Drive, whose link is provided at the end of this manuscript.

The test statistics for K–S test (D), Anderson–Darling test (A^2), Chi-square test (χ^2), AICc, and BIC were computed for the 10 probability distributions. The AICc and BIC values for each of these 10 distributions and 20 cities can be found on the Google Drive, which documents this analysis. The probability distribution that fits a given data the best (using the largest p value) according to each of the above criterion is shown in Table 3.

For each station, we ranked all the probability distribution functions, using each of the four model comparison techniques in decreasing order of its p value. The best-fit distribution among these for each city was found after summing these ranks and choosing the function with the smallest cumulative rank. A similar technique was also used in Sharma and Singh (2010) to find the best distribution, which fits the rainfall data using multiple model comparison techniques. The best-fit distribution for each station using this ranking technique is shown in Table 4. After obtaining the best fit, similar to Ghosh et al. (2016), we then calculate the first four sample L-moments for each station. L-moments

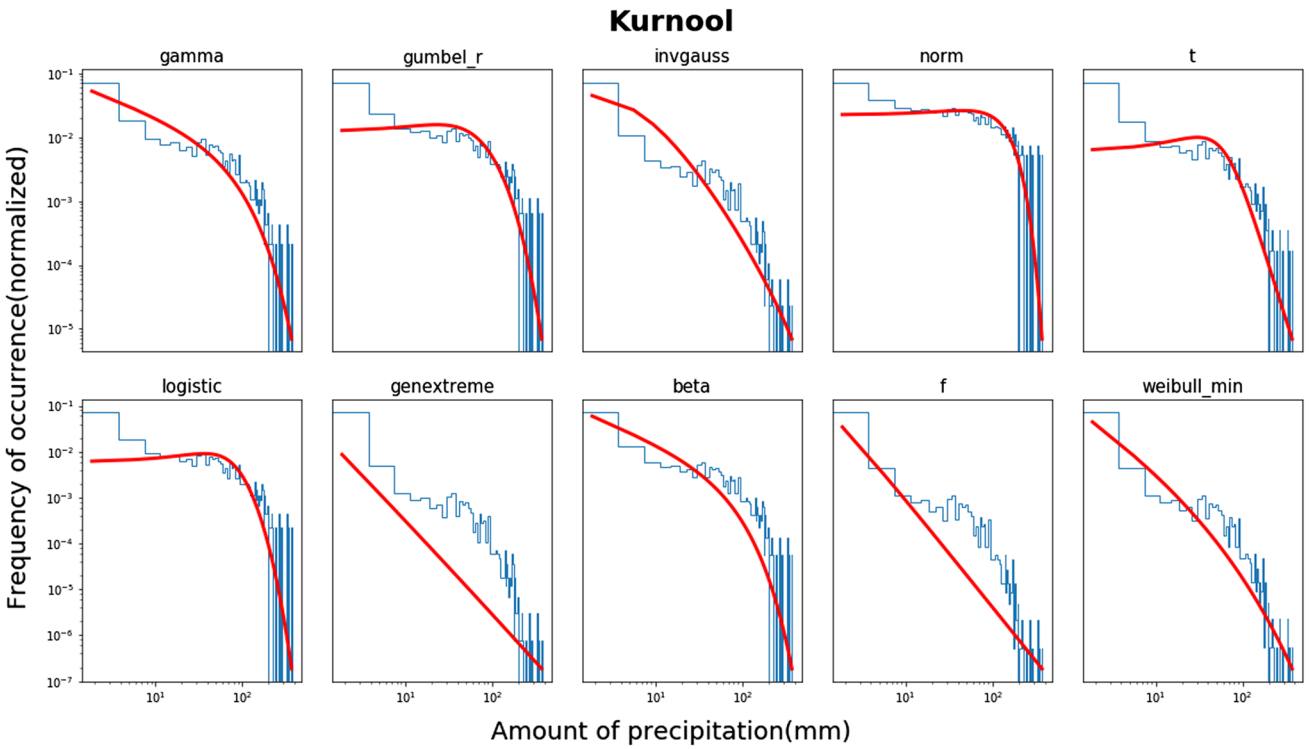

Fig. 2 Histogram of the monthly precipitate data at Kurnool (blue lines) along with best fit for each of the 10 probability distributions functions considered

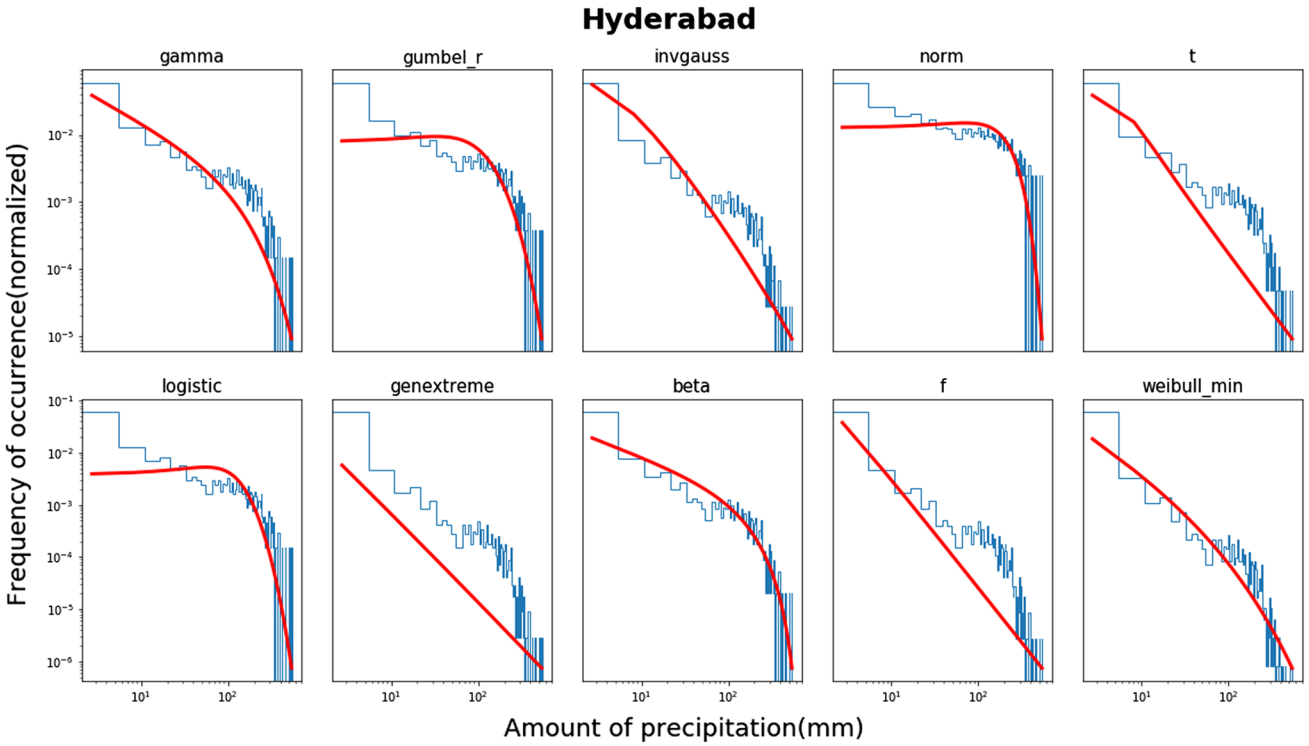

Fig. 3 Histogram of the monthly precipitate data at Hyderabad (blue lines) along with best fit for each of the 10 probability distributions functions considered

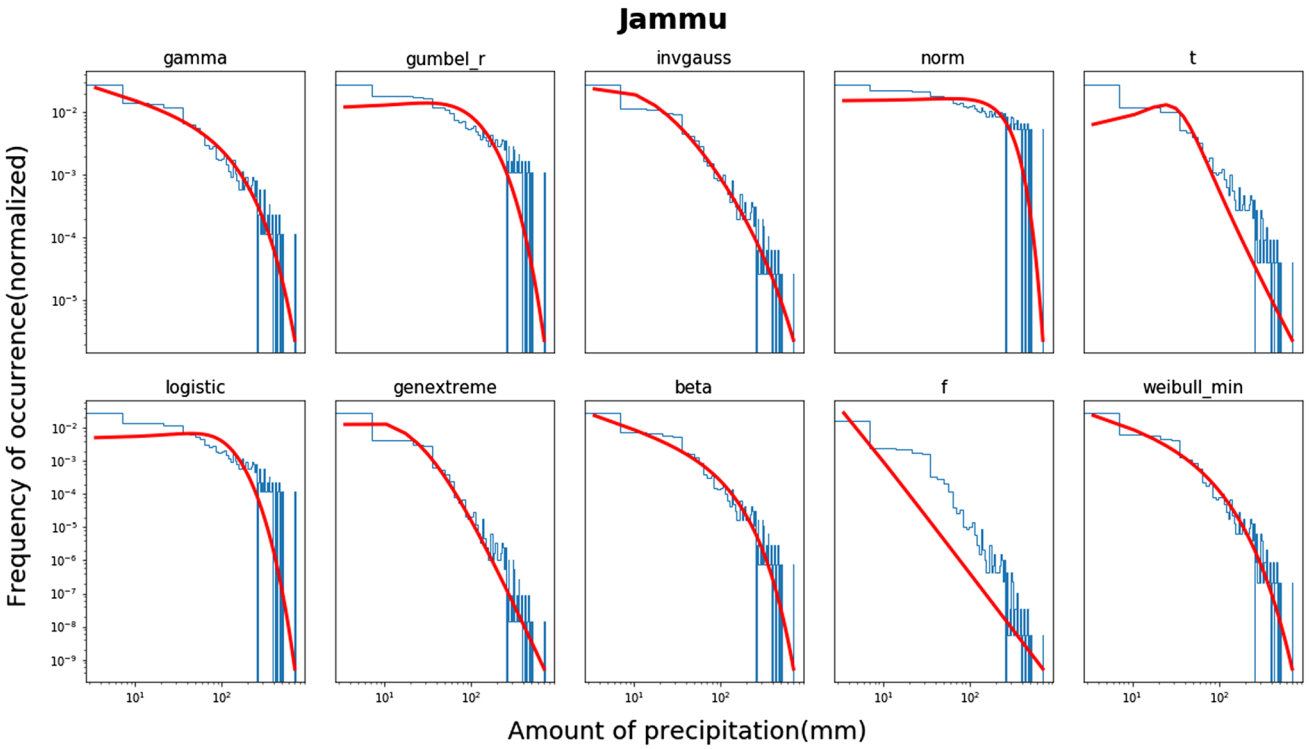

Fig. 4 Histogram of the monthly precipitate data at Jammu (blue lines) along with best fit for each of the 10 probability distributions functions considered

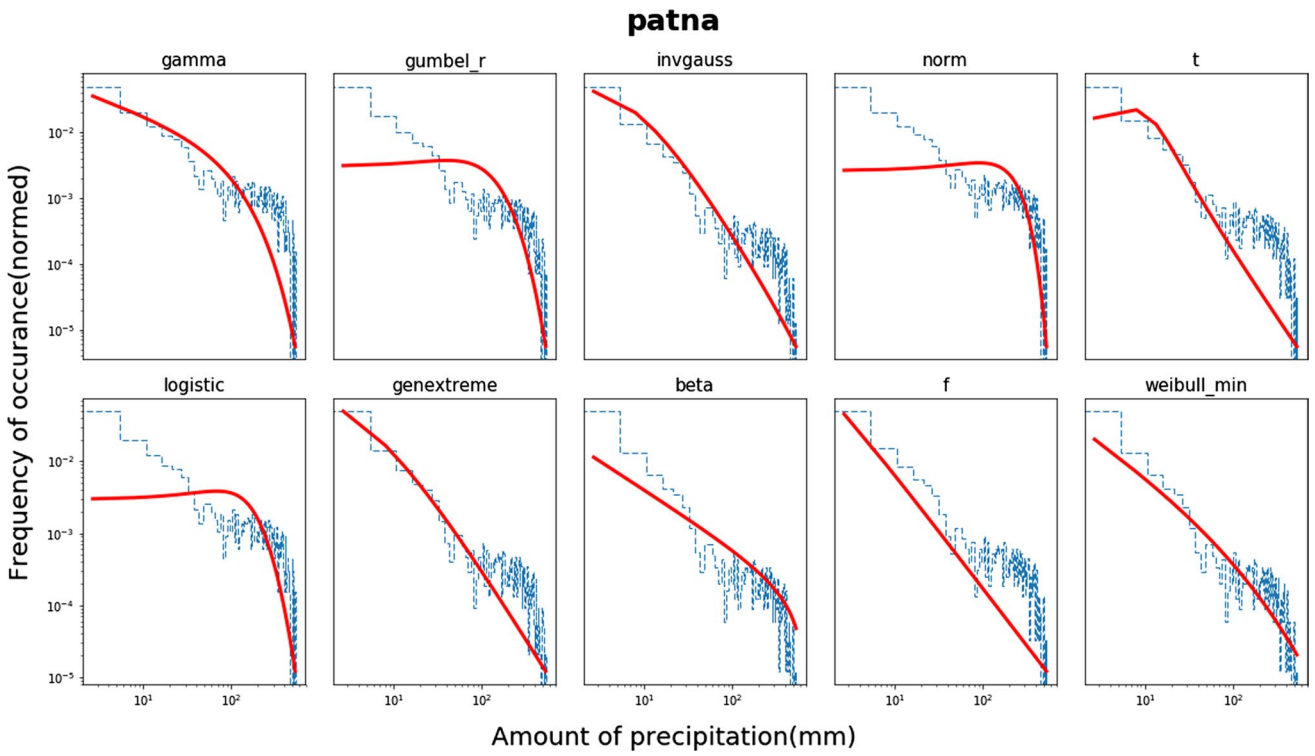

Fig. 5 Histogram of the monthly precipitate data at Patna (blue lines) along with best fit for each of the 10 probability distributions functions considered

Table 3 Station-wise best ranked probability distribution using different goodness-of-fit tests

Study location	KS	AD	Chi square	AIC(c)	BIC
Patna	F	F	GEV	F	Beta
Kurnool	F	F	Weibull	F	Beta
Jaipur	F	F	Inv. Gauss	F	Beta
Chennai	F	F	Gamma	F	Beta
Hyderabad	F	F	Inv Gauss	F	Beta
Lucknow	F	F	Inv. Gauss	F	Beta
Bangalore	F	F	Weibull	F	Beta
Kohima	Weibull	Beta	Beta	Weibull	Beta
Aizawl	Weibull	Beta	Gamma	Weibull	Beta
Guntur	F	F	F	F	Beta
Panipat	F	F	GEV	F	Beta
Amritsar	F	F	Inv. Gauss	F	Beta
Cuttack	F	F	GEV	F	Beta
Gandhinagar	F	F	Beta	GEV	t
Ahmednagar	F	F	Inv. Gauss	t	Beta
Raipur	F	F	GEV	F	Beta
Jammu	F	F	Weibull	F	Beta
Kolkata	F	F	F	F	Beta
Bhopal	F	F	Inv. Gauss	F	Beta
Delhi	F	F	Inv. Gauss	F	Beta

F stands for the Fisher distribution, t stands for Student's *t*-distribution and GEV stands for generalized extreme value distribution

Table 4 Station-wise best-fit distribution obtained by summing the ranks of each of the distributions from all the model comparison tests considered in Table 3

Study location	Best fit
Kohima	Genextreme
Jaipur	Invgauss
Kolkata	Genextreme
Raipur	Genextreme
Gandhinagar	GHenextreme
Hyderabad	Invgauss
Aizawl	Gamma
Bhopal	Invgauss
Ahmednagar	Invgauss
Cuttack	Genextreme
Chennai	Invgauss
Banglore	Genextreme
Patna	Genextreme
Amritsar	Invgauss
Guntur	Gumbel
Lucknow	Invgauss
Kurnool	Gumbel
Jammu	Invgauss
Delhi	Invgauss
Panipat	Genextreme

are linear combinations of expectations of order statistics and are reviewed extensively in Hosking (1990). They are more robust estimates of the central moments than the conventional moments. The first four L-moments are analogous to mean, standard deviation, skewness and kurtosis. These L-moments are shown in Table 5.

Our results from each of the model comparison tests are summarized as follows:

- Using K–S test (D), we find that the Fisher distribution provides a good fit to the monthly precipitation data for all cities except Kohima and Aizawl. For these cities, Weibull distribution provides the best fit.
- Using Anderson–Darling test (A^2), it is observed that the Fisher distribution is the best fit for all the cities except (again) for Kohima and Aizawl, for which the beta distribution gives the best fit for both the cities.
- Using Chi-square test (χ^2), there is no one distribution which consistently provides the best fit for most of the cities. Inverse Gaussian is the optimum fit for seven cities, whereas Weibull and generalized extreme for three cities, beta and Fisher for two cities each. The locations of the corresponding cities can be found in Table 3.
- Using AICc, it is observed that the Fisher distribution provides best distribution for about 16 cities. The exceptions are again Kohima and Aizawl, for which Weibull is the most appropriate distribution. Generalized extreme value distribution provides the best fit for Gandhinagar, whereas Student's *t*-distribution provides the best fit for Ahmednagar.
- For BIC, we find that the beta distribution provides best distribution for all districts except Gandhinagar. Student's *t*-distribution provides best fit for Gandhinagar.

If we then determine the best distribution from a combination of the above model comparison techniques using the ranking technique, we find (cf. Table 4) that the generalized extreme value distribution is the most appropriate for eight cities, inverse Gaussian for nine cities, Gumbel for two cities, and gamma for one city. Therefore, although no one distribution provides the best fit for all stations, most of them can be best fitted using either the generalized extreme value or inverse Gaussian distribution.

Implementation

We have used the python v2.7 environment. In addition, Numpy, pandas, matplotlib, scipy packages are used. Our codes to reproduce all these results can be found in http://goo.gl/hjYn1S. These can be easily applied to statistical studies of rainfall distribution for any other station.

Table 5 Parameters estimates using sample L-moments [mean (L1), variance (L2), skewness (L3) , kurtosis (L4)] of the best-fit distributions

Study location	Best-fit	Mean (L1)	Variance (L2)	Skewness (L3)	Kurtosis (L4)
Kohima	GEV	196.33	98.56	0.21	0.02
Jaipur	Inv Gauss	48.6	36.27	0.57	0.27
Kolkata	GEV	132.15	77.77	0.33	0.07
Raipur	GEV	105.38	70.01	0.43	0.09
Gandhinagar	GEV	56.42	44.44	0.61	0.29
Hyderabad	Inv Gauss	70.06	45.1	0.4	0.1
Aizawl	Gamma	227.2	121.87	0.24	0.01
Bhopal	Inv Gauss	89.53	65.42	0.53	0.18
Ahmednagar	Inv Gauss	70.73	47.78	0.43	0.11
Cuttack	GEV	106.32	62.07	0.3	0.01
Chennai	Inv Gauss	96.89	58.01	0.39	0.17
Bangalore	GEV	69.89	37.14	0.26	0.07
Patna	GEV	90.96	60.58	0.44	0.11
Amritsar	Inv Gauss	39.16	26.01	0.51	0.27
Guntur	Gumbel	65.66	38.73	0.33	0.08
Lucknow	Inv Gauss	74.85	53.11	0.51	0.18
Kurnool	Gumbel	45.19	27.06	0.36	0.13
Jammu	Inv Gauss	60.88	37.41	0.48	0.26
Delhi	Inv Gauss	47.45	34.68	0.57	0.28
Panipat	GEV	43.58	30.62	0.54	0.25

Comparison to previous results

A summary of some of the previous studies of rainfall distribution for various stations in India is outlined in the introductory section. An apples-to-apples comparison to these results is not straightforward, since they have not used the same model comparison techniques or considered all the 10 distributions which we have used. Moreover, the dataset and duration they have used is also different. Nevertheless, we compare and contrast the salient features of our conclusions with the previous results.

Among the previous studies, Sharma and Singh (2010) have also found that generalized extreme value distribution fits the weekly rainfall data for Pantnagar. We also find that this distribution provides the best fit for eight cities. The best-fit distribution which we found for Aizawl agrees with the results from Mooley and Appa Rao (1970), Kulandaivelu (1984), Bhakar et al. (2006). None of the previous studies have found the inverse Gaussian or the Gumbel distribution to be an adequate fit to the rainfall data. However, this could be because these two distributions were not fitted to the observed data in any of the previous studies. Inverse Gaussian and the Gumbel distribution have only recently been considered by Ghosh et al. (2016) and Nadarajah and Choi (2007) for fitting the rainfall data in Bangladesh and Korea, respectively. We hope our results spur future studies to consider these distributions for fitting rainfall data in India.

Conclusions

We carried out a systematic study to identify the best-fit probability distribution for the monthly precipitation data at twenty selected stations distributed uniformly throughout India. The data showed that the monthly minimum and maximum precipitation at any time at any station ranged from 0 to 802 mm, which obviously indicates a large dynamic range. So identifying the best parametric distribution for the monthly precipitation data could have a wide range of applications in agriculture, hydrology, engineering design and climate research.

For each station, we fit the precipitation data to 10 distributions as described in Table 1. To determine the best fit among these distributions, we used five model comparison tests, such as K–S test, Anderson–Darling test, Chi-square test, Akaike and Bayesian information criterion. The results from these tests are summarized in Table 3. For each model comparison test, we ranked each distribution according to its p value and then added the ranks from all the four tests. The best-fit distribution for each city is the one with the minimum total rank and is tabulated in Table 4. We find that no one distribution can adequately describe the rainfall data for all the stations. For about nine cities, the inverse Gaussian distribution provides the best fit, whereas generalized extreme value can adequately fit the rainfall distribution for about eight cities. Our study is the first one, which finds the inverse Gaussian distribution to be the optimum fit for any station. Among the remaining cities, Gumbel and gamma distributions are the best fit for two and one city, respectively.

In the hope that this work would be of interest to researchers wanting to do similar analysis and to promote transparency in data analysis, we have made our analysis codes as well as data publicly available for anyone to reproduce this results as well as to do similar analysis on other rainfall datasets. This can be found at http://goo.gl/hjYn1S

References

Abdullah M, Al-Mazroui M (1998) Climatological study of the southwestern region of Saudi Arabia. i. Rainfall analysis. Clim Res 9:213–223

Bhakar S, Bansal AK, Chhajed N, Purohit R (2006) Frequency analysis of consecutive days maximum rainfall at Banswara, Rajasthan, India. ARPN J Eng Appl Sci 1(3):64–67

Cochran WG (1952) The $\chi 2$ test of goodness of fit. Ann Math Stat 23:315–345

Deka S, Borah M, Kakaty S (2009) Distributions of annual maximum rainfall series of North-East India. Eur Water 27(28):3–14

Fisher R (1925) The influence of rainfall on the yield of wheat at rothamsted. Philos Trans R Soc Lond B Biol Scie 213(402–410):89–142

Ghosh S, Roy MK, Biswas SC (2016) Determination of the best fit probability distribution for monthly rainfall data in Bangladesh. Am J Math Stat 6(4):170–174

Hosking JR (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc Ser B (Methodological) 52:105–124

Kahle D, Wickham H (2013) ggmap: spatial visualization with ggplot2. R J 5(1):144–161

Kulandaivelu R (1984) Probability analysis of rainfall and evolving cropping system for coimbatore. Mausam 35(3):257

Kulkarni S, Desai S (2017) Classification of gamma-ray burst durations using robust model-comparison techniques. Astrophys Space Sci 362(4):70

Kumar V (2017) Statistical distribution of rainfall in Uttarakhand, India. Appl Water Sci 7:1–12

Liddle AR (2004) How many cosmological parameters. Mon Not R Astron Soc 351(3):L49–L53

Mahdavi M, Osati K, Sadeghi SAN, Karimi B, Mobaraki J (2010) Determining suitable probability distribution models for annual

precipitation data (a case study of mazandaran and golestan provinces). J Sustain Dev 3(1):159
Mohamed TM, Ibrahim AAA (2015) Fitting probability distributions of annual rainfall in Sudan. J Sci Technol 17(2)
Mooley D, Appa Rao G (1970) Statistical distribution of pentad rainfall over india during monsoon season. Indian J Meteorol Geophys 21:219–230
Nadarajah S, Choi D (2007) Maximum daily rainfall in South Korea. J Earth Syst Sci 116(4):311–320
Naghavi B, Yu FX (1995) Regional frequency analysis of extreme precipitation in louisiana. J Hydraul Eng 121(11):819–827
Nguyen VTV, Tao D, Bourque A (2002) On selection of probability distributions for representing annual extreme rainfall series. In: Global solutions for urban drainage, pp 1–10
Rockström J, Barron J, Fox P (2003) Water productivity in rain-fed agriculture: challenges and opportunities for smallholder farmers in drought-prone tropical agroecosystems. Water Prod Agric Limits Oppor Improv 85199(669):8
Şen Z, Eljadid AG (1999) Rainfall distribution function for libya and rainfall prediction. Hydrol Sci J 44(5):665–680
Sharma MA, Singh JB (2010) Use of probability distribution in rainfall analysis. NY Sci J 3(9):40–49
Stephens MA (1974) Edf statistics for goodness of fit and some comparisons. J Am Stat Assoc 69(347):730–737
Stephenson D, Kumar KR, Doblas-Reyes F, Royer J, Chauvin F, Pezzulli S (1999) Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon. Mon Weather Rev 127(9):1954–1966
VanderPlas J, Connolly AJ, Ivezić Ž, Gray A (2012) Introduction to astroml: machine learning for astrophysics. In: 2012 Conference on intelligent data understanding (CIDU). IEEE, pp 47–54
Wallace J (2000) Increasing agricultural water use efficiency to meet future food production. Agric Ecosyst Environ 82(1–3):105–119
Waylen PR, Quesada ME, Caviedes CN (1996) Temporal and spatial variability of annual precipitation in Costa Rica and the Southern oscillation. Int J Climatol 16(2):173–193