Best Fit Probability Distributions for Monthly Radiosonde Weather Data

Athulya P. S[1] and K. C James[2]
[1]M.Tech III Semester, [2]Professor Department of statistics
Cochin University of Science and Technology
athulyaps430@gmail.com, jamesmech@cusat.ac.in

Abstract

Study of weather in the tropical regions like India is a major challenge due to the sophisticated and dynamic nature weather system. Applications of probability distributions to weather data have been investigated by several researchers from different regions of the world. This paper explains a methodology for fitting the probability distribution to weather parameters with the help of testing goodness of fit tests. Data recorded from the location thiruvanathapuram on a period of 10 years (1892-1991) is used to establish best fit probability distributions from three different statistical tests and best one is selected by a ranking method. It is observed that Johnson SB distribution and Gamma distribution gives good fit in most of the instances.

Keywords: **Probability Distribution, Weather Parameter, goodness of fit, Johnson SB distribution, Gamma distribution**

1. INTRODUCTION

The prognosis of weather in the tropical regions like India is a major challenge due to the sophisticated and dynamic nature weather system. The day to day changes of weather such as pressure, temperature, wind speed and humidity are the meteorological parameters to be monitored on a continuous basis. Probability distribution fitting is finding an appropriate probability distribution to a data set of the repeated measurement of a variable phenomenon. The aim of fitting the distribution is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval. There are lot of probability distributions of which some can be fitted more closely to the observed frequency of the data, depending on the nature of the phenomenon and of the distribution. The best fit distribution lead to a good prediction. In distribution fitting, therefore, one needs to select a distribution that suits the data well. Weather conditions are necessary to be predicted not only for future plans in agriculture and industries but also in many other fields like defence, mountaineering, shipping and aerospace navigation etc. It is often used to warn about natural disasters are caused by abrupt change in climatic conditions. At macro level, weather forecasting is usually done using the data gathered by remote sensing satellites. Weather parameters like maximum temperature, minimum temperature, pressure, wind streams and their directions, are projected using images and data taken by these meteorological satellites to access future trends.

Applications of probability distributions to weather data have been investigated by several researchers from different regions of the world. Biswas and Khambete [4] computed the lowest amount of rainfall at different probability level by fitting gamma distribution probability model to week by week total rainfall of 82 stations in dry farming tract of Maharashtra. Duan et al., [5] suggested that for modeling daily rainfall amounts, the weibull and to a lesser extent the exponential distribution is suitable. Upadhaya and Singh [6] stated that it is possible to predict rainfall fairly accurate using various probability distributions for certain returns periods although the rainfall varies with space, time and have erratic nature. Sen and Eljadid [7] reported that for monthly rainfall in arid regions, gamma probability distribution is best fit.

Rai and Jay [8] studied humidity and upper winds temperature over Madras in relation to precipitation occurrence and found the vertical distribution of temperature and humidity associated with dry or wet days over the same area. Benson [9] adopted a large scale planning for improved flood plain management and expending water resources development and he suggested adopting a procedure where records are available for all government agencies. Along with Pearson type I, Gumble’s and log normal distribution, the log Pearson type III distribution has been selected as the based method with provision for departure from the base method were justified continuing study leading towards improvements or revision of method is recommended. Kulkarni and Pant [10] studied the cumulative frequency distribution of rainfall of different intensities during south-west monsoon for 20 stations in India. The distribution was found to be exponential and curves were fitted to observed date by the method of least square.

2. METHODOLOGY

This study was done on collected radiosonde data. A radiosonde is a small weather station linked with a radio transmitter. The radiosonde is connected to a helium or hydrogen-filled balloon, usually called a weather balloon, and this balloon lifts the radiosonde to heights exceeding 115,000 feet. During the radiosonde’s rise, it transmits data on various parameters like temperature, pressure, and humidity to ground-based receiving station. These data were recorded on a period of 10 years (1892-1991). Building a best fit probability distribution for different weather parameter has long been a topic of interest in the field of meteorology. This study is planned to identify the best fit probability distribution based on distribution pattern for different radiosonde weather data set.

Data containing weather parameter was analyzed to identify the best fit probability distribution for each month wise data set. Here basically three statistical goodness of fit test (Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test)were carried out in order to select the best fit probability distribution on the basis of highest rank with minimum value of test statistic. The correct probability distributions are found for the different dataset using the results obtained from three selected goodness of fit tests.

The probability distributions viz., gamma, weibull, Pearson, generalized extreme value were fitted to the data for evaluating the best fit probability distribution for weather parameters. In addition, the different forms of these distributions were also tried and thus total 13 probability distributions viz. gamma (3P), generalized gamma (4P),weibull (3P), pearson 6 (4P), Johnson SB ,Beta, Log-Logistic (3P), Triangular, Burr (4P), Cauchy, Pearson 5 (3P), Inv. Gaussian (3P), Gen. Pareto were applied to find out the best fit probability distribution.

3. FITTING THE PROBABILITY DISTRIBUTION

3.1 Testing the goodness of fit

The goodness of fit test actually tests the compatibility of random sample with the theoretical probability distribution. The goodness of fit tests is applied for testing the following null hypothesis:

Ho: the weather parameter data follow the specified distribution

H1: the weather parameter data does not follow the specified distribution

3.1.1 Kolmogorov-Smirnov Test : The Kolmogorov-Smirnov test [3] is used to decide if a sample comes from a population with a specific distribution.

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

$$D = \max_{1\le i\le n}\left[F\left(X_i\right)-\frac{i-1}{n},\frac{i}{n}-F\left(X_i\right)\right] \qquad (1)$$

Where,

Xi = random sample, i =1, 2,….., n.

$$\text{CDF} = F_n(X) = \frac{1}{n}[\text{Number of Observations} \le X] \qquad (2)$$

This test is used to decide if a sample comes from a hypothesized continuous distribution

3.1.2 Anderson-Darling Test: The Anderson-Darling test [4] 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 it gives more weight to the tails than does the K-S test. The K-S test is distribution free in the meaning 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. The Anderson-Darling test statistic is defined as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1).[\ln F(X_i) + \ln(1 - F(X_{n-i+1}))] \qquad (3).$$

3.1.3 Chi-Squared Test : The Chi-Squared statistic is defined as

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

Where,
Oi = observed frequency,
Ei = expected frequency,
'i'= number of observations (1, 2, …….k)
Ei is calculated by the following computation

$$E_i = F\left(X_2\right) - F\left(X_1\right) \qquad (5)$$

F is the CDF of the probability distribution being tested.
The observed number of observation (k) in interval 'i' is computed from equation given below

$$k = 1 + \log_2 n \qquad (6)$$

Where, n is the sample size.
This test is for continuous sample data only and is used to determine if a sample comes from a population with a specific distribution [2].

3.2 Probability Distribution of Weather Data.

Descriptive statistics is basically interested in exploring and describing a sample of data. Here descriptive statistics of the weather data set is given in Table 1.

Table 1. Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Pressure	5892	7	956.08	258.4726	267.45048
Temperature	5892	-87.26	27.18	-36.7721	29.43914
N	5892				

The methodology depicted in section 3.1 is applied to the 10 years weather data classified into monthly basis. These data sets used to study the distribution pattern at different levels. The test statistic D, A^2 and $\chi^2$for each data set are computed and shown in the Table 2.

Table 2. Distributions fitted for pressure data sets

Study period	Test ranking first position					
	Kolmogorov Smirnov		Anderson Darling		Chi-square	
	Distribution	Statistic	Distribution	Statistic	Distribution	Statistic
January	Johnson SB	0.04297	Johnson SB	0.9187	Johnson SB	2.8018
February	Gen. Gamma (4P)	0.05604	Gamma (3P)	3.779	Gamma (3P)	27.054
March	Johnson SB	0.02749	Johnson SB	0.38205	Johnson SB	1.5856
April	Pearson 6 (4P)	0.04701	Johnson SB	1.9771	Pearson 6 (4P)	2.4498
May	Weibull (3P)	0.05209	Johnson SB	2.3878	Johnson SB	10.86
June	Beta	0.04162	Beta	1.7418	Beta	6.7056
July	Gen. Gamma (4P)	0.05688	Johnson SB	1.6672	Johnson SB	16.735
August	Beta	0.04485	Beta	2.2464	Johnson SB	3.9355
September	Gen. Gamma (4P)	0.05588	Weibull (3P)	7.59	Johnson SB	16.03
October	Gen. Gamma (4P)	0.03801	Weibull (3P)	3.9634	Weibull (3P)	21.697
November	Gamma (3P)	0.04692	Gamma (3P)	2.3234	Gamma (3P)	13.043
December	Weibull (3P)	0.05889	Johnson SB	1.9128	Johnson SB	5.0578

It is observed that Johnson SB distribution is fitted in more than 50% of months. The parameters of the distributions identified are listed in Table 3. Random numbers are generated using the estimated parameters for monthly pressure data set and the least square method was used for finding best fit distribution [11]. Best selected probability distributions of pressure data are presented in Table 4.

Table 3. Parameters of the distributions fitted for pressure data sets.

Study period	Distribution	Parameter
January	Johnson SB	$\gamma=0.77689, \delta=0.77689, \lambda=0.77689, \xi=0.77689$
February	Gamma (3P)	$\alpha=0.58082 \ \beta=421.93 \ g=7.273$
	Gen. Gamma (4P)	$k=1.0314 \ \alpha=0.55953, \beta=423.7 \ \xi=7.273$
March	Johnson SB	$\gamma=0.71702 \ \delta=0.54444, \lambda=1009.6 \ \xi=-0.28684$
April	Pearson 6 (4P)	$\alpha1=0.61327 , \alpha2=22.601, \beta=8997.2 \ \xi=7.049$
	Johnson SB	$\gamma=0.85923 \ \delta=0.47498, \lambda=987.45 \ \xi=0.02358$
May	Johnson SB	$\gamma=0.88903 \ \delta=0.46306, \lambda=981.92 \ \xi=2.1275$
	Weibull (3P)	$\alpha=0.70901 \ \beta=196.99 \ \gamma=7.048$
June	Beta	$\alpha1=0.50476 \ \alpha2=1.0825, a=7.353 \ b=951.91$
July	Johnson SB	$\gamma=0.87108 \ \delta=0.47162 \ \lambda=977.59 \ \xi=4.7911$
	Gen. Gamma (4P)	$k=0.98318 \ \alpha=0.6987 \ \beta=346.05 \ \gamma=7.288$
August	Beta	$\alpha1=0.43974 \ \alpha2=1.0144 \ a=7.232 \ b=953.84$
	Johnson SB	$\gamma=0.74508 \ \delta=0.48069 \ \lambda=985.82 \ \xi=-2.7184$
September	Weibull (3P)	$\alpha=0.6823 \ \beta=179.29 \ g=7.0$
	Johnson SB	$\gamma=0.91857 \ \delta=0.45139 \ \lambda=979.71 \ \xi=1.6486$
	Gen. Gamma (4P)	$k=0.59438 \ \alpha=1.2884 \ \beta=110.32 \ \gamma=7.0$
October	Weibull (3P)	$\alpha=0.77829 \ \beta=220.89 \ \gamma=7.059$
	Gen. Gamma (4P)	$k=1.6748 \ \alpha=0.31601 \ \beta=708.61 \ \gamma=7.059$
November	Gamma (3P)	$\alpha=0.5752 \ \beta=439.02 \ \gamma=7.03$
December	Weibull (3P)	$\alpha=0.73956 \ \beta=210.25 \ \gamma=7.0$
	Johnson SB	$\gamma=0.84835 \ \delta=0.48087 \ \lambda=991.64 \ \xi=0.00225$

Table 4. Best fit probability distribution for Pressure

Study Period	Best-Fit
January	Johnson SB
February	Gamma (3P)
March	Johnson SB
April	Johnson SB
May	Pearson 6 (4P)
June	Beta
July	Johnson SB
August	Beta
September	Johnson SB
October	Weibull (3P)
November	Gamma (3P)
December	Johnson SB

Johnson SB Distribution is observed six times in the monthly data sets, means January, March, April, July, September and December indicating the highest contribution of the distribution. Further, we observe that Gamma (3P), Weibull (3P), Beta, Pearson 6 (4P) are found as the best fitted probability distributions for the monthly pressure data sets.

While looking into the temperature data set, the same methods are followed as explained above for the pressure data. Table 5 listed the distributions with rank 1 for the three goodness of fit tests. Parameter of these identified distributions for each data set is mentioned in the Table 6.The best selected probability distributions for monthly temperature data set are presented in Table 7.

Table 5. Distributions fitted for temperature data sets.

Study period	Test ranking first position					
	Kolmogorov Smirnov		Anderson Darling		Chi-square	
	Distribution	Statistic	Distribution	Statistic	Distribution	Statistic
January	Johnson SB	0.04417	Beta	2.2877	Gamma (3P)	21.2
February	Triangular	0.06008	Gamma (3P)	3.1373	Log-Logistic (3P)	32.562
March	Johnson SB	0.03315	Beta	1.2873	Beta	16.336
April	Pearson 6 (4P)	0.0537	Pearson 6 (4P)	2.4933	Log-Logistic (3P)	34.428
May	Pearson 6 (4P)	0.05273	Burr (4P)	3.3964	Cauchy	46.905
June	Johnson SB	0.03616	Beta	2.4033	Beta	17.139
July	Gen. Gamma (4P)	0.06384	Gamma (3P)	3.5981	Log-Logistic (3P)	34.032
August	Johnson SB	0.05133	Gamma (3P)	3.9581	Gamma (3P)	37.785
September	Burr (4P)	0.06123	Log-Logistic (3P)	3.8865	Log-Logistic (3P)	52.929
October	Johnson SB	0.05424	Gamma (3P)	2.9896	Gamma (3P)	24.451
November	Gen. Pareto	0.06436	Gamma (3P)	3.2132	Pearson 5 (3P)	48.748
December	Weibull (3P)	0.05994	Gamma (3P)	2.8038	Inv. Gaussian (3P)	33.689

Table 6. Parameters of the distributions fitted for temperature data sets

Study period	Distribution	Parameter
January	Johnson SB	$\gamma=0.34791$ $\delta=0.62009$ $\lambda=109.17$ $\xi=-79.176$
	Beta	$\alpha1=1.0992$ $\alpha2=1.3526$ $a=-83.32$ $b=26.72$
	Gamma (3P)	$\alpha=2.5064$ $\beta=20.749$ $\gamma=-86.936$
February	Triangular	$m=-73.282$ $a=-87.558$ $b=38.615$
	Gamma (3P)	$\alpha=2.3989$ $\beta=20.373$ $\gamma=-88.392$
	Log-Logistic (3P)	$\alpha=2.9493$ $\beta=48.369$ $\gamma=-95.141$
March	Beta	$\alpha1=1.0397$ $\alpha2=1.2014$ $a=-84.515$ $b=26.119$
	Johnson SB	$\gamma=0.26152$ $\delta=0.61924$ $\lambda=111.71$ $\xi=-82.078$
April	Pearson 6 (4P)	$\alpha1=3.2005$ $\alpha2=5.1458E+6$ $\beta=8.9882E+7$ $\gamma=-92.757$
	Log-Logistic (3P)	$\alpha=4.0417$ $\beta=67.405$ $\gamma=-109.65$
May	Pearson 6 (4P)	$\alpha1=2.1796$ $\alpha2=92.904$ $\beta=1986.7$ $\gamma=-84.502$
	Burr (4P)	$k=37.854$ $\alpha=1.5469$ $\beta=517.84$ $\gamma=-83.017$
	Cauchy	$\sigma=16.581$ $\mu=-45.259$
June	Johnson SB	$\gamma=0.29354$ $\delta=0.61902$ $\lambda=110.14$ $\xi=-79.176$
	Beta	$\alpha1=1.1853$ $\alpha2=1.3162$ $a=-85.041$ $b=27.495$
July	Gen. Gamma (4P)	$k=1.5634$ $\alpha=1.0638$ $\beta=48.51$ $\gamma=-83.263$
	Gamma (3P)	$\alpha=2.3454$ $\beta=19.752$ $\gamma=-84.034$
	Log-Logistic (3P)	$\alpha=2.7192$ $\beta=42.667$ $\gamma=-87.945$
August	Johnson SB	$\gamma=0.41838$ $\delta=0.56421$ $\lambda=102.31$ $\xi=-73.605$
	Gamma (3P)	$\alpha=3.1773$ $\beta=17.095$ $\gamma=-88.884$
September	Log-Logistic (3P)	$\alpha=2.7583$ $\beta=39.505$ $\gamma=-85.09$
	Burr (4P)	$k=0.13666$ $\alpha=4.4190E+8$ $\beta=1.8445E+9$ $\gamma=-1.8445E+9$
October	Johnson SB	$\gamma=0.47585$ $\delta=0.60157$ $\lambda=107.43$ $\xi=-76.34$
	Gamma (3P)	$\alpha=2.4609$ $\beta=19.995$ $\gamma=-85.827$
November	Weibull (3P)	$\alpha=1.5192$ $\beta=48.979$ $\gamma=-82.246$
	Gamma (3P)	$\alpha=2.0774$ $\beta=21.761$ $\gamma=-83.276$
December	Weibull (3P)	$\alpha=1.746$ $\beta=56.743$ $\gamma=-87.819$
	Gamma (3P)	$\alpha=2.8494$ $\beta=18.667$ $\gamma=-90.491$

Table 7. Best fit probability distribution for temperature

Study Period	Best -Fit
January	Gamma (3P)
February	Gamma (3P)
March	Beta
April	Pearson 6 (4P)
May	Pearson 6 (4P)
June	Beta
July	Gamma (3P)
August	Gamma (3P)
September	Log-Logistic (3P)
October	Gamma (3P)
November	Gamma (3P)
December	Gamma (3P)

4. CONCLUSION

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 radiosonde weather data. It is observed that Johnson SB distribution is fitted in more than 50 percent of months. The best fit probability distribution of monthly data was found to be different for each month. Gamma (3P) was observed in most of the monthly temperature data. Beta distribution was best fit for March and June. Log-Logistic (3P) and Pearson 6 (4P) was the best fit distribution for rest of months. Identifying the distribution of weather data has a wide range of applications in agriculture field and climate research.

REFERENCES

[1] Sharma, M.A. and Singh, J.B. (2010).Use of Probability Distribution in rainfall Analysis, New York science Journal, pp.40-49.

[2] Chakravarti, Laha, and Roy, 1967. Handbook Methods of Applied Statistics. Volume I, John Wiley and Sons, pp. 11-27.

[3] Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, Vol. 69, pp. 730-737.

[4] Biswas, B.C. and Khambeta, N.K. Distribution of short period rainfall over dry farming tract of Maharashtra. Journal of Maharashtra Agricultural University. 1989

[5] Duan, j., Sikka, A.K., and Grant, G.E. A comparison of stochastic models for generating daily precipation at the H.J. Andrews Experiment Forest. Northwest Science. 1995; 69(4):pp .318-329.

[6] Upadhaya, A., and Singh, S.R. Estimation of consecutive day's maximum rainfall by various methods and their comparison. Indian Journal of S. Cons. 1998;26 (2):pp. 193-2001.

[7] Sen, Z., and Eljadid, A.G. Rainfall distribution functions for Libya and Rainfall Prediction. Hydrol. Sci. J. 1999;4(5):pp. 665-680.

[8] Rai Sircar, N.C. and Jay Raman, N.C. (1966): Study upper winds, temperature and humidity over Madras in relation to precipitation occurrences there during the monsoon season. Indian Journal of Meteorological Geophysics, Vol. 17, No. 4, pp. 649-651.

[9] Benson, Manuel, A. (1968): Uniform flood frequency estimating methods for federal agencies. Water resources research, Vol. 4, No. 5, pp. 891-895.

[10] Kulkarni, N.S. and Pant, M.B. (1969): Cumulative frequency distribution of rainfall of different intensities. Indian Journal of Meteorology Geophysics, Vol.20, No. 2, pp. 109-114.

[11] S Ghosh,,M.K Roy,S.C Biswas(2016): Determination of the Best Fit Probability Distribution for Monthly Rainfall Data in Bangladesh.American Journal of Mathematics and Statistics , 6(4): 170-174