Best Fit Probability Distribution Analysis of Major Crop Paddy of Rice Bowl State of India-Telangana

L. Rajani Devi and V. V. HaraGopal

***Abstract* — Telangana state's population is mostly dependent on agriculture. The Telangana state's economy depends heavily on agriculture, as does the nation's and the state's ability to achieve food security. Combining art and science to fit a statistical distribution to data involves making trade-offs along the way. The secret to effective data analysis is striking a balance between improving distributional fit and preserving ease of estimation while keeping in mind that the analysis's ultimate goal is to help you make better decisions. A recurring issue in agricultural research was which distribution should be utilized to simulate the production data from an experiment.**

An analysis is then carried out utilizing the obtained distributions using the statistical method to fit probability distributions to data of variables. These distributions would be a representation of the properties of the variable data. The twenty distributions are: Cauchy, Error, Hypersequent, Gamma(3p), Laplace, Logistic, Log Pearson 3, Rayleigh (2p), Weibull (3p), Log Logistic (3p), Triangular, Gen Gamma, Gen.Gamma(4p), Gen.Extreme Value, Log Normal (3p), Pearson 5 (3p), Fatigue life (3p), Inv. Gaussian (3p), Nakagami, In order to suit a distribution research, rice distributions are used. Twenty probability distributions were computed, and the test statistic Kolmogorov-Smirnov test, Anderson-Darling test, Chi-Square test, and each data set were used to choose the distribution that fit the data the best. The probability distributions include Cauchy, Error, Hypersequent, Gamma, Laplace, Logistic, Log Pearson 3, Rayleigh (2p), Weibull (3p), Log Logistic (3p), Triangular, Gen. Gamma, Gen. Gamma (4p), Gen. Extreme Value, Log Normal (3p), Log Pearson 5, Fatigue life (3p), Inv. Gaussian (3p), and Nakagami.

***Keywords* — Anderson-darling test; Cauchy; Chi-square test; Distribution fitting; Fatigue life(3p); Gamma (3p); Generalized Extreme Value; Goodness-of-fit; Inverse Gaussian; Kolmogorov-Smirnov test; Laplace; Log Logistic; Log Pearson 3; Probability Distribution.**

I. Introduction

Agriculture provides a living for the bulk of Telangana's citizens. In order to provide food security for the state of Telangana and the nation as a whole, agriculture plays a crucial role in both the state's economy and national food security. A method of modelling and quantifying the results of production in the State of Telangana's paddy industry is probability distribution [1], [2]. Analysts can use the features of the distribution of a data set to forecast outcomes [3]. Because of its two clearly distinguishable parameters, mean and variance, simplicity, and the popular belief that most populations are distributed normally when sampled in large numbers, the normal distribution is one of the most often used probability distributions [4]. The Normal distribution, however, presupposes a few conditions, including.

It is important to note that the goal of the paper is not to promote the use of the Normal distribution or to demonstrate the applicability of a distribution to any set of data, but rather to promote the practice of finding the distribution that fits the data the best while refusing to assume that the data are normal [4]. In order to use the generated distribution for additional investigation, this work investigates strategies for identifying the probability distribution that fits a collection of data the best [5], [6]. Raw data are initially matched to probability distributions of interests that suggest a strong fit distribution for further Paddy Production data analysis [7]. The distribution in context is then picked with the best characteristics [8]. Following that, goodness of fit tests would be performed to choose the distribution that best suited the data [9], [10]. In the distribution.

Published on March 11, 2023.

L. R. Devi, Department of Mathematics& Statistics, Telangana Mahila Viswavidyalayam (UCW, Koti), Hyderabad-500026, India.
(corresponding e-mail: trdevi11@yahoo.com)
V. V. HoraGopal, Department of Statistics, Osmania University Hyderabad & BITS Pilani, Hyderabad Campus, India.
(e-mail: haragopalvajjha@gmail.com)

DOI: http://dx.doi.org/10.24018/ejmath.2023.4.2.179

II. Data Description and Source of Data

Here, the Secondary Data is collected from Agricultural Statistics Division, Directorate of Economics and Statistics, Telangana State. The Data is collected for the years 2008-2020 for the Seasons Kharif and Rabi from all the regions of the state Telangana before and after bifurcation.

TABLE I: Production in Tones

Year	SEASON	Paddy Production
2008-2009	Kharif	4758712
	Rabi	3281633
2009-2010	Kharif	2453107
	Rabi	2450428
2010-2011	Kharif	5303722
	Rabi	4500202
2011-2012	Kharif	5042488
	Rabi	2678595
2012-2013	Kharif	4683792
	Rabi	2287562
2013-2014	Kharif	5651844
	Rabi	4219504
2014-2015	Kharif	4282235
	Rabi	2535038
2015-2016	Kharif	3295374
	Rabi	1275303
2016-2017	Kharif	4542720
	Rabi	5355523
2017-2018	Kharif	4418771
	Rabi	4975997
2018-2019	Kharif	6202378
	Rabi	3800748
2019-2020	Kharif	8977506
	Rabi	8850209

III. Methods and Materials

The user may characterize the behavior of the underlying Production data using the best-fitted distribution for a dataset. Usually, more than one distribution would be of interest in the matching process when fitting the data with a distribution. Raw data is gathered from the empirical data for several years in order to discover the distributions that would be fitted onto the data. For a visual depiction of the density, a histogram that represents the form of the distribution of the empirical data may be created. To fit the distributions, probability distributions from the three broad categories of parametric, non-parametric, and semi-parametric can be used [5]. Only the non-parametric and parametric distributions will be the subject of this study. Once the probability density function is determined to fit the data, Goodness of Fit tests are applied to find best fitting distribution.

A. *Probability Function*

The fitting of the distributions was done using the following twenty distributions: Cauchy, Error, Hypersecent, Gamma, Laplace, Logistic, Log Pearson 3, Rayleigh (2p), Weibull (3p), Log Logistic (3p), Triangular, Gen Gamma, Gen.Gamma (4p), Gen.Extream Value, Log Normal (3p), Pearson 5 (3p), Fatigue life (3p), I nv.Gaussian (3p), Naka The Goodness of Fit tests that would be important decision-making aids in selecting the best-fitting distribution are also listed in the section that follows [11], [12].
Some of the fitted distributions are:

1) *Log-Logistic Distribution*

α - indicates shape (α>0)
β- Indicates scale (β>0)
γ - indicates location
Three-Parameter Log-Logistic Distribution
Probability density function

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2} \; ; \gamma \le x < +\infty$$

2) *Laplace (Double exponential) Distribution*

λ – indicates inverse scale (λ >0)
μ- indicates location
Probability Density Function

$$f(x) = \frac{\lambda}{2}\exp(-\lambda|x-\mu|) \; ; -\infty < x < +\infty$$

3) *Pearson 5 Distribution*

α- indicates shape (α>0)
β- indicates scale (β>0)
γ –indicates
Three-Parameter Pearson 5 Distribution
Probability density function

$$f(x)=\frac{exp(-\beta(x-\gamma))}{\beta\Gamma(\alpha)((x-\gamma)/\beta)^{\alpha+1}}\ ;\quad \gamma<x<+\infty$$

4) *Generalized Extreme Value Distribution*

k- indicates shape
σ- indicates scale (σ>0)
μ- indicates location
Probability Density Function

$$f(x)=\begin{cases}\frac{1}{\sigma}\exp(-\ (1+kz)^{-1/k}\ )\ (1+kz)^{-1-1/k} & k\neq 0\\ \frac{1}{\sigma}\exp(-z-\exp(-z)) & k=0\end{cases}$$

5) *Lognormal Distribution*

σ- continuous parameter (σ>0)
μ- continuous parameter
γ - continuous location parameter(γ≡0 yields the two-parameter Lognormal distribution
Domain
γ<x<+∞
Three-Parameter Lognormal Distribution
Probability density function

$$f(x)=\frac{exp\left(-\frac{1}{2}\left(-\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^{2}\right)}{(x-\gamma)\sigma\sqrt{2\pi}}$$

6) *Fatigue Life(Birnbaum-Saunders) Distribution*

α - indicates shape (α>0)
β - indicates scale ($\beta > 0$)
γ - indicates location
Three-Parameter Fatigue Life Distribution
Probability Density function

$$f(x)=\frac{\sqrt{(x-\gamma)/\beta}+\sqrt{\beta/(x-\gamma)}}{2\alpha(x-\gamma)}.\Phi\left(\frac{1}{\alpha}\quad\left(\sqrt{\frac{x-\gamma}{\beta}}-\sqrt{\frac{\beta}{x-\gamma}}\right)\right);\ \gamma<x<+\infty$$

7) *Cauchy Distribution*

σ- indicates scale ("σ">0)
μ- indicates location
Probability Density function

$$f(x)=\left(\pi\sigma\left(1+\left(\frac{x-\mu}{\sigma}\right)^{2}\right)\right)^{-1}\ ;\ -\infty<x<+\infty$$

8) *Generalized Gamma Distribution*

k- indicates shape (k>0)
α - indicates shape (α>0)
β – indicates scale (β>0)
γ - indicates location
Four-Parameter Generalized Gamma Distribution
Probability Density function

$$f(x)=\ \frac{k(x-\gamma)^{k\alpha-1}}{\beta^{k\alpha}\Gamma(\alpha)}\exp\ (-((x-\gamma)/\beta)^{\ k})\ ;\ \gamma\leq x<+\infty$$

9) *Gamma Distribution*

α - indicates shape (α>0)
β - scale ($\beta > 0$)
γ - location
Three-Parameter Gamma Distribution
Probability Density function

$f(x)=\frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}$ exp(-(x-γ)/β); $\gamma \le x < +\infty$

10) Inverse Gaussian Distribution

λ - continuous parameter (λ>0)
μ- continuous parameter (μ>0)
γ - location
Three-Parameter Inverse Gaussian Distribution
Probability Density function

$f(x)=\sqrt{\frac{\lambda}{2\pi(x-\gamma)^3}}\exp\left(-\frac{\lambda(x-\gamma-\mu)^2}{2\mu^2(x-\gamma)}\right)$; $\gamma < x < +\infty$

B. Goodness of Fit Tests

1) Kolmogorav-smirnov test

This test determines if a sample originates from a presumed continuous distribution. The empirical cumulative distribution function serves as its foundation (ECDF). Assume that we have a CDF F distribution and a random sample of size n, As shown by, the empirical CDF is

$F_n(x)=\frac{1}{n}$ (no. of observations ≤ x]

Definition

Kolmogorav-Smirnov Statistic(D) is based on the largest vertical difference between the theoretical and empirical cumulative distribution function.

D= Max. $1 \le i \le n$ [$F(X_i)$ – ($\frac{i-1}{n}, \frac{i}{n}$ -$F(x_i)$]

2) Anderson darling test

A general test to determine if an observed cumulative distribution function fits an expected cumulative distribution function is the Anderson Darling technique. Compared to the Kolmogorav-Smirnov Test, this test gives the stories greater weight.

Definition: The Anderson Darling Statistic (A2) is defined as

$A^2= -n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)$ [ln $F(x_i)$+ ln(1-$F(x_{n-i+1})$)]

3) Chi square test

To ascertain if a sample is drawn from a population with a certain distribution, the Chi Squared test is performed. The value of the test statistic relies on how the data was binned because this test is conducted to binned data. Only continuous sample data are available for use with this test.

There are numerous formulas that may be used to determine the number of bins(k) depending on the sample size, even though there is no ideal solution for this number (n). Here the following empirical formula for k is k =1+log_2n

The information can be categorized into intervals with equal probabilities or widths. It is necessary to join some adjacent bins together in order to satisfy the requirement that each bin contains at least 5 data points.

Definition: The Chi Squared statistic is defined as $\chi^2= \sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$

Where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i calculated by $E_i= F(x_2) - F(x_1)$. Where F is the CDF of the probability distribution being tested, and x_1, x_2 are the limits for bin i.

IV. Results and Discussion

A. Descriptive Statistics

The Production Data for Paddy of Telangana State for the Period of 2008-2020 for both the Seasons Kharif and Rabi were analyzed to understand basic Statistical Characteristics in Table II.

TABLE II: Descriptive Statistics

Percentile	Value
Min	1275303
5%	1528367.75
10%	2368995.0
25% (Q1)	2829354.5
50% (Median)	4459486.5
75% (Q3)	5238413.5
90%	7526293.5
95%	8945681.75
Max	8977506

Coefficient of Variation is observed as 0.4214375 which is less than 1, considered as low variance in the Distribution and also Skewness is observed as 0.9055390 shows that the Data are moderately skewed. Also, excess Kurtosis observed as 1.3458776 indicates the data follows Leptokurtic and the distribution shows heavy tails on either side indicating large outliers.

B. Best Fit Distribution

In this Section results documented for the best fit Probability Distribution for all the Productions for different years are focused. The Production data were assessed with twenty different Probability distributions are described with Probability density function Plots are given in Figures.

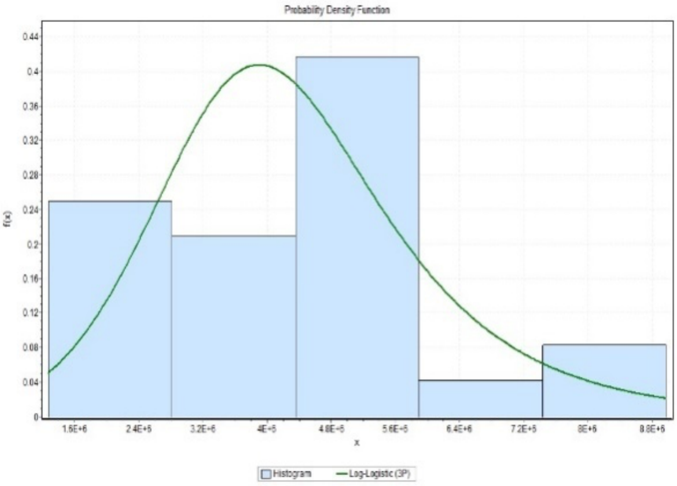

Fig. 1. PDF of Log Logistic Distribution.

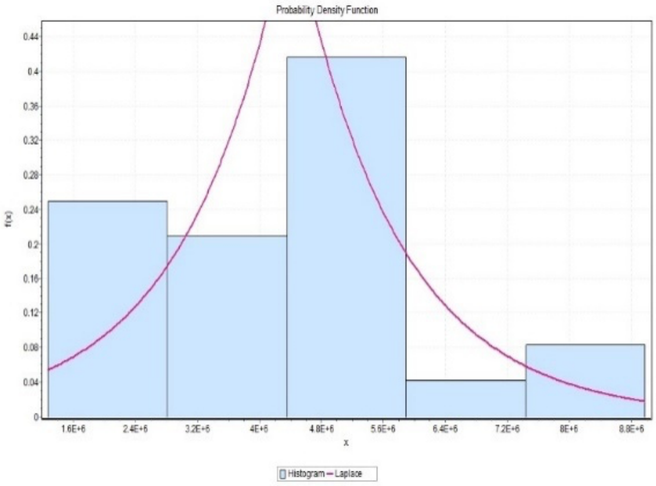

Fig. 2. PDF of Laplace Distribution.

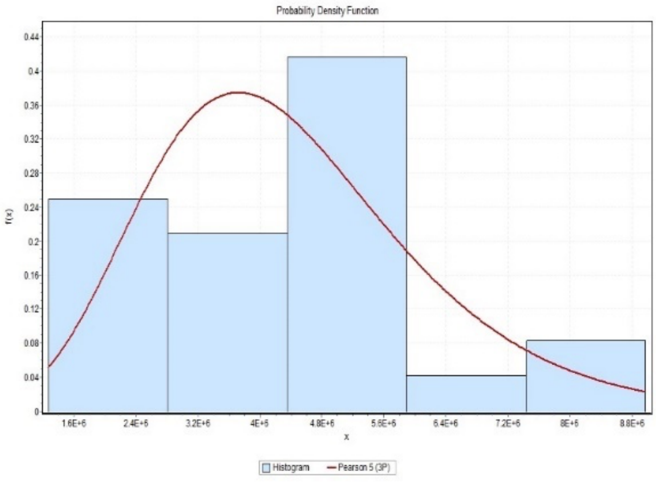

Fig. 3. PDF of Pearson 5 Distribution.

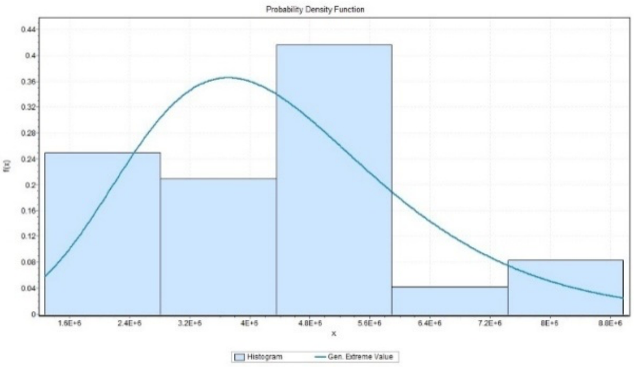

Fig. 4. PDF of General Extreme Value Distribution.

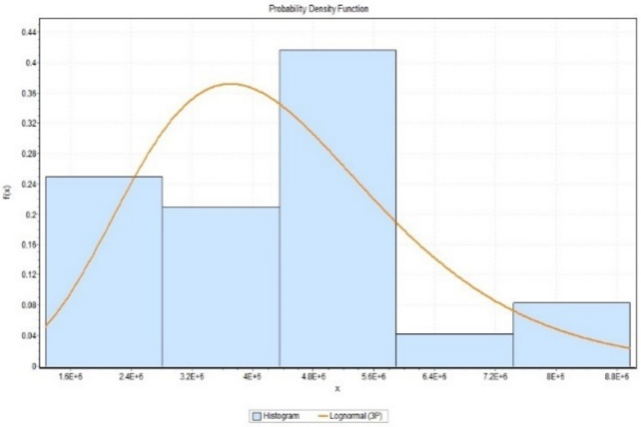

Fig. 5. PDF of Log Normal Distribution.

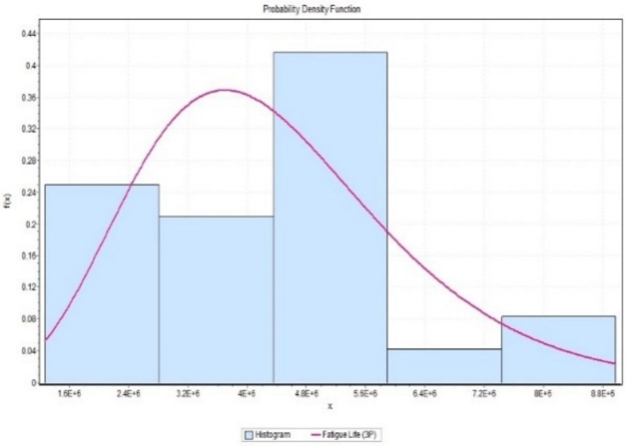

Fig. 6. PDF of Fatigue Life Distribution

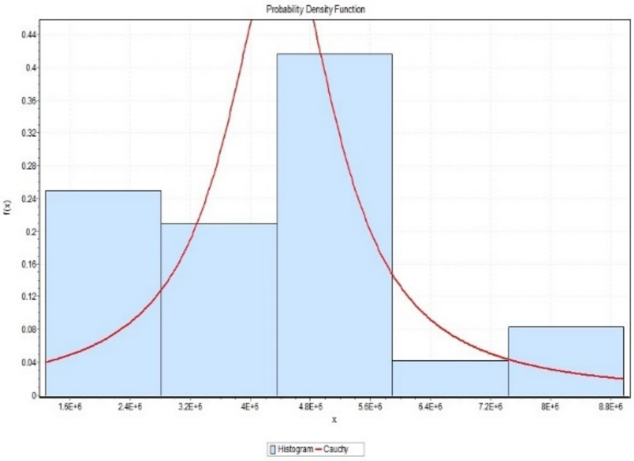

Fig. 7. PDF of Cauchy Distribution.

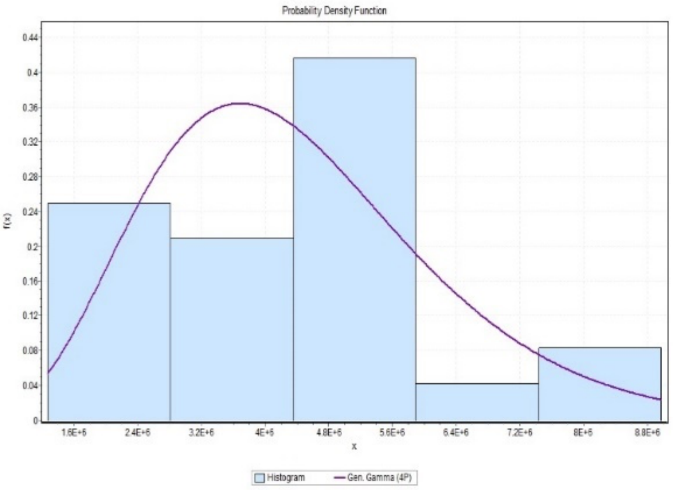

Fig. 8. PDF of Generalized Gamma Distribution.

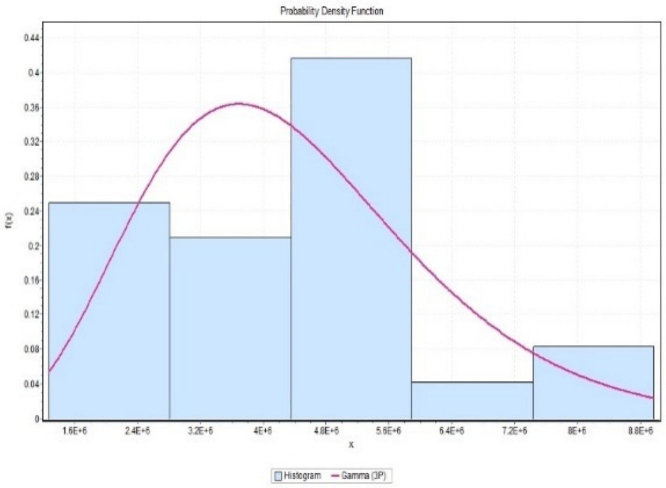

Fig. 9. PDF of Gamma Distribution.

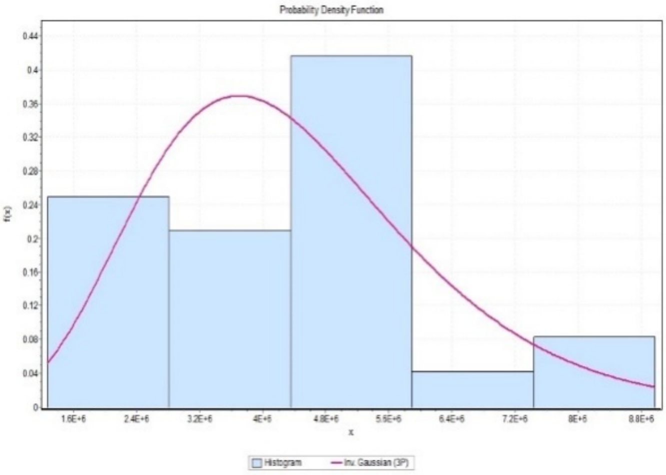

Fig. 10. PDF of Inverse Gaussian Distribution.

To identify which Probability distributions would be most appropriate for characterising the data variance for the Production data reported in Table I, a summary of the resulting Parameters is shown in Table. III. These parameters and the distribution may be applied to any production-related modelling and policy-making from the perspective of agricultural production requirements. The best fit distributions, including Cauchy, Error, Hypersecent, Gamma, Laplace, Logistic, Log Pearson 3, Rayleigh (2p), Weibull (3p), Log Logistic (3p), Triangular, Gen Gamma, Gen.Gamma (4p), Gen.Extream Value, Log Normal (3p), Pearson 5 (3p),Fatigue life (3p), I nv.Gaussian (3p), Nakagami, and Rice distributions, The probability distribution that had the lowest rating was deemed to have the greatest match.

TABLE III: PARAMETERS OF THE DISTRIBUTION

S.No	Distribution	Parameters

1	Cauchy	σ=8.8E+5
		μ=4.4E+6

2	Fatigue Life (3P)	α=0.28
		β=6.3E+6
		γ=-2.1E+6

3	Gamma (3P)	α=6.4
		β=7.2E+5
		γ=-1.9E+5

4	Gen. Extreme Value	k=-0.07
		σ=1.6E+6
		μ=3.6E+6

5	Gen. Gamma (4P)	k=0.98
		α=6.7
		β=6.7E+5
		γ=-2.0E+5

6	Inv. Gaussian (3P)	λ=8.7E+7
		μ=6.6E+6
		γ=-2.2E+6

7	Laplace	λ=7.6E-7
		μ=4.4E+6

8	Log-Logistic (3P)	α=6.6
		β=6.4E+6
		γ=-2.2E+6

9	Lognormal (3P)	σ=0.27
		μ=16.0
		γ=-2.1E+6

10	Pearson 5 (3P)	α=24.0
		β=2.0E+8
		γ=-4.1E+6

TABLE IV: GOODNESS OF FIT-SUMMARY

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Cauchy	0.12	5	0.43	7	0.76	19
2	Error	0.1	1	0.44	12	0.32	4
3	Fatigue Life (3P)	0.14	14	0.43	6	0.42	11
4	Gamma (3P)	0.14	11	0.44	9	0.43	12
5	Gen.Extreme Value	0.14	12	0.42	4	0.38	7
6	Gen.Gamma	0.14	19	0.44	11	0.44	14
7	Gen.Gamma (4P)	0.14	13	0.44	8	0.43	13
8	Hypersecant	0.11	2	0.44	13	0.33	5
9	Inv.Gaussian (3P)	0.14	15	0.44	10	0.46	15
10	Laplace	0.12	4	0.41	2	1.5	20
11	Log-Logistic (3P)	0.13	9	0.39	1	0.35	6
12	Log-Pearson 3	0.13	6	0.46	14	0.42	9
13	Logistic	0.12	3	0.49	17	0.32	3
14	Lognormal (3P)	0.14	16	0.43	5	0.42	10
15	Nakagami	0.14	18	0.57	18	0.09	2
16	Pearson 5 (3P)	0.14	17	0.42	3	0.41	8
17	Rayleigh (2P)	0.13	7	0.48	15	0.47	16
18	Rice	0.14	20	0.6	19	0.06	1
19	Triangular	0.13	10	2.7	20	0.58	18
20	Weibull (3P)	0.13	8	0.49	16	0.47	17

V. CONCLUSIONS

Though the Variation in the Data is low, and is because of Climatic factor, Geographical Settings and other Environmental Challenges which are difficult to understand. Climatic change issues are most significant Concerns for the Scientific Community involved in Agriculture Planning. In Order to minimize the Uncertainty in using past records and analyzing an effort has been made with well-organized Procedure to understand the best fit distribution of Paddy Production in the State Telangana where agriculture is main Source for many Rural People.

The twenty different Probability distributions, including Cauchy, Error, Hypersequent, Gamma, Laplace, Logistic, Log Pearson 3, Rayleigh (2p), Weibull (3p), Log Logistic (3p), Triangular, Gen Gamma, Gen.Gamma(4p), Gen.Extream Value, Log Normal (3p), Pearson 5 (3p), Fatigue life (3p), Inv.Guassian (3p), Nakagami, and Rice distributions, were applied The optimal probability distribution was determined to be the Log-Logistic (3P) distribution. The distribution giving a close fit is supposed to lead good predictions and calculations and apply the results to make well grounded agricultural policies. The main issue in every other subject is the selection of an appropriate probability distribution.

Acknowledgment

L Rajani Devi expresses sincere gratitude and appreciation to Prof V V Haragopal for his continuous support and encouragement in carrying out the study. We authors thank Agricultural Statistics Division, Directorate of Economics and Statistics, Telangana State for providing the data.

Conflict of Interest

Authors declare that they do not have any conflict of interest.

References

[1] Dikko HG, David IJ, Bakari HR. Modeling the distribution of rainfall intensity using quarterly data. *IOSR Journal of Mathematics*. 2013; 9(1): 11-16.

[2] Jamaludin S, Jemain AA. Fitting the statistical distributions to the daily rainfall amount in peninsular Malaysia. *Journal Technology*. 2007: 33â-48.

[3] Ghosh S, Roy MK, Biswas SC. Determination of the best fit probability distribution for monthly rainfall data in Bangladesh. *American Journal of Mathematics and Statistics*. 2016; 6(4): 170-174.

[4] Gupta SC, Kapoor VK. *Fundamentals of mathematical statistics*. 11th ed. Sultan Chand & Sons; 2002.

[5] Kumar V, Bala S. Best fit probability distribution analysis of precipitation and potential evapotranspiration of India’s highly dense population state-Bihar. *MAUSAM*. 2022; 73(1): 139-150.

[6] Chaudhari RH, Khokhar AN, Paramr DJ, Patel HV, Kumar P, Kumar R. Fitting of the distribution for CV value of the cotton and tobacco experiment. *Journal of Pharmacognosy and Phytochemistry*. 2020; (5S): 884-890.

[7] Amin MT, Rizwan M, Alazba AA. A best-fit probability distribution for the estimation of rainfall in northern regions of Pakistan. *Open Life Sciences*.2016; 11(1): 432-440.

[8] Beckman RJ, Tiet jen GL. Maximum likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation*. 1978; 7(3-4): 253-258.

[9] Smirnov N. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics.* 1948; 19(2): 279-281.

[10] Alam MA, Emura K, Farnham C, Yuan J. Best-fit probability distributions and return periods for maximum monthly rainfall in Bangladesh. *Climate*. 2018; 6(1): 9.

[11] Darling DA. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*. 1957; 28(4): 823-838.

[12] Anderson TW, Darling DA. A test of goodness of fit. *Journal of the American statistical Association.* 1954; 49(268): 765-769.