MAUSAM, **74**, 4 (October 2023), 1081-1092

MAUSAM

DOI : https://doi.org/10.54302/mausam.v74i4.805
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No. 551.577 : 551.509.3 : 519.676

Probability analysis and rainfall forecasting using ARIMA model

CHANDRAN S., SELVAN P., NAMITHA M. R.*, PRADEEP MISHRA** and KUMAR V***

Thiagarajar College of Engineering, Madurai – 625 015, India

**KCAET, Tavanur-Malapuram (Kerala) – 679 573, India*

***College of Agriculture, Rewa, J. N. K. V. V. – 486 001 (M.P.), India*

****ACRI-TNAU-Madurai Campus – 625 104, Tamil Nadu, India*

(*Received 10 July 2021, Accepted 16 February 2023*)

e mail : waterchandran@gmail.com

सार — **तमिलनाडु में वैगई नदी की दस उप-द्रोणियों के** 1976 **से** 2009 **तक** 34 **वर्षों के वर्षा के आँकड़े एकत्र किए गए और विभिन्न संभाव्यता वितरण फलनों का उपयोग करके सांख्यिकीय रूप से इनका विश्लेषण किया गया। अध्ययन क्षेत्र के लिए दो उत्कृष्ट उपयुक्त परीक्षणों का उपयोग करके वार्षिक, मासिक और ऋतुनिष्ठ वर्षा के लिए सबसे उपयुक्त संभाव्यता वितरण पाए गए। मॉडल की पहचान, नैदानिक जांच और अध्ययन क्षेत्र की वार्षिक वर्षा पूर्वानुमान के लिए बॉक्स-जेनकिंस ऑटोरिग्रेसिव इंटीग्रेटेड मूविंग एवरेज** (ARIMA) **पद्धति को अपनाया गया। प्रत्येक उप-द्रोणी के लिए सर्वश्रेष्ठ** ARIMA **मॉडल का चयन किया गया, और** 2010, 2015, 2020 **और** 2025 **के लिए औसत वार्षिक वर्षा का पूर्वानुमान दिया गया। पूर्वानुमानित परिणाम की तुलना** 2020 **तक प्रेक्षित किए गए डेटा से अच्छी तरह से की गई, जो मॉडल की उपयुक्तता को दर्शाता है।**

ABSTRACT. A 34-year rainfall data from 1976 to 2009 of ten sub-basins of the Vaigai River in Tamil Nadu were collected and analysed statistically using various probability distribution functions. The best-fit probability distributions for the annual, monthly and seasonal rainfall for the study area were found using two goodness-of-fit tests. The Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) methodology has been adopted for model identification, diagnostic checking and forecasting the study area's annual rainfall. The best ARIMA models were selected for each sub-basin and the average annual precipitation for 2010, 2015, 2020 and 2025 has been forecasted. The forecasted result compared well with observed dataup to 2020, which indicates the appropriateness of the model.

Key words – ARIMA, Rainfall, Probability analysis, Forecasting.

1. Introduction

The rainfall distribution pattern, duration, temporal and spatial variation significantly influence the agricultural systems. India's economy is mainly dependent on agriculture, which is affected directly by the variation in rainfall. The varying seasonal, annual and monthly rainfall trends are useful in managing the cropping system and applying irrigation properly. Several studies have been conducted in India and abroad on rainfall analysis and best fit probability distribution functions were found to analyse the trend of rainfall (Sharma and Singh, 2010). Ray *et al*. (1980) stated that the weekly, monthly and seasonal pattern of rainfall and their probabilities help crop planning by identifying the periods of drought, normal and excess rain. The influence of rainfall on wheat yield in Rothamsted was studied by Fisher (1925).

The early warning of rainfall helps manage water resources and it also helps in taking preventive measures against natural calamities like floods and drought. In India, many researchers have done forecasting rainfall all over the country with various spatial and temporal resolutions (Kaushik and Singh, 2008; Chattopadhyay and Chattopadhyay, 2010; Narayanan *et al*., 2013). Eni and Adeyeye (2015) did a seasonal ARIMA modeling and forecasting rainfall in Warri town, Nigeria. Several empirical approaches, *viz*., regression, Autoregressive Integrated Moving Average (ARIMA), fuzzy logic, artificial neural network (ANN), etc., are widely used for rainfall forecasting. The empirical approaches for rainfall forecasting deal with evaluating past rainfall data and setting a link with self or other meteorological variables (Narayanan *et al*., 2016). Valipour (2015) investigated the ability of the seasonal autoregressive integrated moving

TABLE 1

Description of various probability distribution functions

S. No.	Name of the probability distribution	Probability density function	Range	Parameters
1.	Johnson SB	$z = \gamma + \delta \log\left(\frac{x-\xi}{\xi+\lambda-x}\right)$	$\gamma, \xi\ \delta > 0,\ \lambda > 0$	γ, δ, λ and ξ
2.	Dagum	$z = \frac{\alpha k\left(\frac{x}{b}\right)^{\alpha k-1}}{\beta\left[1+\left(\frac{x}{\beta}\right)^{\alpha}\right]^{k+1}}$	k, α, β > 0	k, α and β
3.	General Pareto	$\frac{1}{\sigma}\left[1+k\frac{(x-\mu)}{\sigma}\right]^{-1-1/k}$	k, μ, σ > 0	k, μ and σ
4.	Beta	$\frac{1}{B(\alpha_1,\alpha_2)}\frac{(x-\alpha_1)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-\alpha)^{\alpha_1+\alpha_2-1}}$	$\alpha_1, \alpha_2 > 0$ and $\alpha > b$	$\alpha_1, \alpha_2, \alpha$ and b
5.	Generalized Extreme Value	$\frac{1}{\sigma}\exp\left[-z-\exp(-z)\right]$	σ > 0	σ
6.	Log-Logistic (3P)	$\frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{-2}$	α, β, γ > 0	α, β and γ
7.	Pert	$\frac{1}{B(\alpha_1,\alpha_2)}\frac{(x-\alpha)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-\alpha)^{\alpha_1+\alpha_2-1}}$	$\alpha_1, \alpha_2 > 0$ and $\alpha > b$	$\alpha_1, \alpha_2, \alpha$ and b
8.	Weibull	$\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$	α, β > 0	α and β
9.	Rayleigh	$\frac{x}{\sigma^2}\exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right]$	σ > 0	σ
10.	Power function	$\frac{\alpha(x-\alpha)^{\alpha-1}}{(b-\alpha)^{\alpha}}$	α, $b > 0$	α and b
11.	Burr	$z = \frac{\alpha k\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta\left[1+\left(\frac{x}{\beta}\right)^{\alpha}\right]^{k+1}}$	k, α, β > 0	k, α and β
12.	Gamma	$\frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp(-x/\beta)$	α, β > 0	α and β
13.	Reciprocal	$\frac{1}{x[ln(b)-ln(a)]}$	$a,\ b > 0$	a and b
14.	Uniform	$\frac{1}{b-a}$	$b < a$	a and b

TABLE 1 (*Contd.*)

S. No.	Name of the probability distribution	Probability density function	Range	Parameters
15.	Error	$c_{1\sigma^{-1}}\exp\left(-\left\|c_0\frac{x-\mu}{\sigma}\right\|\right)$	$\mu, \sigma > 0$	μ and σ
16.	Inv. Gaussian	$\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right]$	$\mu,\ \lambda > 0$	μ and λ
17.	Log-Pearson 3	$\frac{1}{x\lvert\beta\rvert\Gamma(\alpha)}\left[\frac{ln(x)-\gamma}{\beta}\right]^{\alpha-1}\exp\left[\frac{ln(x)-\gamma}{\beta}\right]$	$\alpha, \beta, \gamma > 0$	α, β and γ
18.	Lognormal (3P)	$\frac{\exp\left\{-\frac{1}{2}\left[\frac{ln(x-\alpha)-\mu}{\sigma}\right]^2\right\}}{(x-\alpha)\sigma\sqrt{2\pi}}$	$\mu, \sigma, \alpha > 0$	μ, α and σ
19.	Pearson 5 (3P)	$\frac{\exp[-\beta/(x-\gamma)]}{\beta\Gamma(\alpha)[(x-\gamma)/\beta]^{\alpha+1}}$	$\alpha, \beta, \gamma > 0$	α, β and γ
20.	Gen. Gamma	$\frac{kx^{k\alpha-1}}{\beta^{k\alpha}\Gamma(\alpha)}\exp\left[-\left(\frac{x}{\beta}\right)\right]^k$	$\alpha, \beta, k > 0$	α, β and k

average (SARIMA) and autoregressive integrated moving average (ARIMA) models for long-term runoff forecasting in the United States.

In this study, the ARIMA model was used to forecast the average annual rainfall for ten different sub-basins of the Vaigai River in Tamil Nadu.

2. Materials and method

Daily rainfall data for 34 years (1976-2009) of 10 sub-basins in Vaigai basin, Tamil Nadu were collected from the Tamil Nadu state Public Works Department (PWD) observatory.These sub-basins lie in the southernmost part of the Indian subcontinent and are located between 9°30′ and 10°10′ North latitudes and 77°10′ and 77°40′ East longitudes. The ten sub-basins monthly, seasonal and annual rainfall patterns were statistically analysed using various probability distribution functions. Johnson SB, General Pareto, Dagum and others. (Table 1).

Seasonal rainfall was classified into four categories, *viz*., southwest monsoon (June-September), northeast monsoon (October-November), summer (March-May) and winter (December-February). The probability distributions were fitted to the data using the data analyser and simulation software, Easy fit.

The Kolmogorov Smirnov and Anderson Darling statistical goodness-of-fit tests were carried out to select the best fit probability distribution based on the highest rank with minimum test statistic value. The annual rainfall pattern for the next 16 years (until 2025) was forecasted using Box-Jenkin's ARIMA method based on the best fit probability distribution. The statistical software SPSS was used for predicting the annual rainfall.

2.1. *Box-Jenkins model*

The Box-Jenkins model (Box and Jenkins,1976) is the best used computer-calculated forecasting model based on time-series data regression studies. The basic principle behind this methodology is that the present value of the series is in any way related to its past values. Given a time series of data X_t, the ARMA model is a tool for understanding and perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) component and a moving average (MA) part. The model is usually referred to as the ARMA

(*p*,*q*) model, where *p* is the order of the autoregressive part and *q* is the order of the moving average part (as defined below).

2.1.1. *Autoregressive model*

The notation AR (*p*) refers to the autoregressive model of order *p*. The AR (*p*) model is written as:

$$X_t = c + \sum_{i=1}^{P} \rho_i X_t + \varepsilon_t$$

where, $\rho_1, \rho_2 \ldots\ldots \rho_p$ are the parameters of the model, *c* is a constant and ε_t is white noise. Sometimes the constant term is neglected.

2.1.2. *Moving Average model*

The notation MA (*q*) refers to the moving average series of order *q*:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where, the $\theta_1, \theta_2 \ldots \theta_q$ are the parameters of the model, μ is the expectation of X_t (often assumed as zero) and $\varepsilon_t, \varepsilon_{t-1}$ are white noise at respective time intervals.

A time series $\{X_t\}$ is stationary and if for every *t*,

$$X_t - \phi_1 X_{t-1} - \ldots\ldots\ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots\ldots\ldots + \theta_1 Z_{t-q}$$

where, $\{Z_t\} \sim WN(0,\sigma^2)$ and the polynomials

$$\left(1 - \phi_1 Z - \ldots\ldots\ldots - \phi_p Z^p\right) \text{and} \left(1 + \theta_1 Z + \ldots\ldots\ldots + \theta_q Z^q\right)$$

have no common factors.

where, *p* and *q* are respectively the AR and MA terms.

Stationarity

Box and Jenkins (1976), Anderson (1976), Judge *et al.* (1982), Chatfield (1984) and Pankratz (1983) pointed out that for the process to be strictly stationary, the joint distribution function describing the process must be invariant concerning time, where $F(zt,\ldots,zt+k) = F(zt+s,\ldots,zt+s+k)$ for *s* and all *k*. This strong stationarity condition implies that the mean, variance and covariance are constant.

2.2. *Autoregressive Moving Average model (ARMA)*

The acronym ARIMA stands for "Auto-Regressive Integrated Moving Average". Lags of the differenced series appearing in the forecasting equation are called "auto-regressive" terms, lags of the forecast errors are called "moving average" terms and a time series that needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series. A non-seasonal ARIMA model is classified as an "ARIMA (p,d,q)" model, where:

(*i*) p is the number of autoregressive terms,

(*ii*) d is the number of non-seasonal differences and

(*iii*) q is the number of lagged forecast errors in the prediction equation.

This method consists of four steps, identification, estimation, diagnostic checking and forecasting.

2.2.1. *Identification*

The problem is to find out the appropriate values of p, d and q. One of the essential tools for identifying is the autocorrelation function (ACF), the partial autocorrelation function (PACF) and the resulting correlograms, which are simply the ACF plots and PACFs against the lag length. One way of accomplishing this is to consider the ACF and PACF and the associated correlograms of a selected number of ARMA processes, such as AR(1), AR(2), MA(1), MA(2), ARMA (1,1), ARMA (2) and so on. Since each of these stochastic processes exhibits a typical ACF pattern and PACF, if the time series under study fits one of these patterns, we can identify the time series with that process. Of course, one will have to apply diagnostic tests to determine if the chosen ARIMA model is reasonably accurate.

2.2.2. *Estimation*

After identifying the appropriate values of p, d and q, the next step is to estimate the parameters of the autoregressive and moving average terms included in the model. Sometimes this calculation can be done by simple least-squares,' but sometimes, one will have to resort to the nonlinear (in parameter) estimation method.

2.2.3. *Checking the model accuracy's*

Among the competitive Box- Jenkins model, the best model is selected based on maximum R^2, minimum root mean square error (RMSE), minimum mean absolute percentage error (MAPE), minimum of maximum average

TABLE 2

Monthly summary statistics for the Lower Vaigai basin

S. No.	Month	Mean ($\bar{x}$)	SD (σ)	CV	Maximum	Minimum	Kurtosis	Skewness (γ)
1.	January	30.75	45.95	1.01	157.60	0.00	4.78	1.71
2.	February	32.02	51.86	0.92	235.73	0.00	9.35	2.50
3.	March	27.66	53.72	0.71	290.80	0.00	18.33	3.76
4.	April	54.40	70.56	0.74	379.48	0.74	14.60	3.15
5.	May	39.79	36.13	0.65	153.54	0.44	4.76	1.42
6.	June	19.15	19.28	0.60	73.39	0.00	3.78	1.27
7.	July	32.33	29.63	0.80	145.84	0.00	7.34	1.79
8.	August	36.50	25.94	1.49	107.73	0.00	2.97	0.67
9.	September	62.50	46.35	1.62	182.85	0.00	3.44	1.14
10.	October	218.00	142.48	1.94	557.98	0.00	2.99	0.80
11.	November	225.63	136.40	1.30	459.42	0.00	1.82	0.14
12.	December	122.90	98.78	0.91	419.20	0.00	3.96	1.18

percentage error (MaxAPE), minimum of maximum absolute error (MaxAE) and minimum of Normalized BIC. Any model which has fulfilled most of the above criteria is selected. This section provides definitions of the goodness-of-fit measures used in time series modeling.

2.3. *R-squared*

An estimate of the proportion of the total variation in the series explained by the model. This measure is most useful when the series is stationary. Positive values mean that the model under consideration is better than the baseline model (Mishra, 2021).

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{X}_i - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

2.4. *Root Mean Square Error (RMSE)*

A measure of how much a dependent series varies from its model-predicted level, expressed in the same units as the dependent series.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \hat{X}_i\right)^2}{n}}$$

2.5. *Mean Absolute Percentage Error (MAPE)*

A measure of how much a dependent series varies from its model-predicted level. It is independent of the units used and can therefore, be used to compare series with different units.

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{X_i - \hat{X}_i}{X_i} \right|}{n} \times 100$$

2.6. *Mean absolute error (MAE)*

Measures how much the series varies from its model-predicted level. MAE is reported in the original series units.

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| X_i - \hat{X}_i \right|}{n}$$

2.7. *Maximum absolute percentage error (MaxAPE)*

The largest forecasted error, expressed as a percentage. This measure is useful for imagining a worst-case scenario for your forecasts.

$$\text{MaxAPE} = 100 \max\left[\left| \frac{X(t) - \hat{X}(t)}{X(t)} \right| \right]$$

TABLE 3

Summary statistics for various seasons for the study area

S. No.	Sub basin	Season	Mean ($\bar{x}$)	SD (σ)	CV	Maximum	Minimum	Kurtosis	Skewness (γ)
1.	Lower Vaigai	S.W.	150.49	70.69	0.47	331.70	23.70	3.22	0.73
		N.E.	566.52	253.07	0.45	1015.20	26.50	2.53	0.10
		Summer	62.77	72.77	1.16	286.90	0.00	4.07	1.30
		Winter	121.84	91.67	0.75	476.30	24.50	8.89	2.29
2.	Theniyaru	S.W.	116.41	61.80	0.53	299.8	37.90	4.52	1.27
		N.E.	337.41	138.10	0.41	616.5	68.70	2.66	0.37
		Summer	31.98	40.85	1.28	157.6	0.00	4.38	1.50
		Winter	149.42	62.92	0.42	283.00	44.40	2.25	0.42
3.	Manjalaru	S.W.	194.38	113.75	0.59	418.40	31.30	1.80	0.21
		N.E.	315.91	189.23	0.60	835.20	77.50	3.01	0.74
		Summer	20.47	27.80	1.36	131.00	0.00	9.37	2.46
		Winter	143.34	87.54	0.61	328.10	7.00	2.57	0.53
4.	Suruliyaru	S.W.	249.18	19.30	0.46	587.70	75.30	3.62	0.74
		N.E.	359.57	25.29	0.42	741.80	143.60	2.73	0.61
		Summer	28.40	6.45	1.34	184.70	0.00	9.66	2.42
		Winter	166.49	12.89	0.46	305.60	13.40	2.21	0.15
5.	Sathaiyaru	S.W.	269.93	83.20	0.31	440.60	131.20	1.99	0.18
		N.E.	379.95	166.38	0.44	740.80	87.40	2.66	0.45
		Summer	23.44	38.25	1.63	159.70	0.00	7.15	2.17
		Winter	135.20	65.84	0.49	329.60	45.20	4.30	1.23
6.	Sirumalaiyaru	S.W.	250.71	90.30	0.36	463.20	86.00	2.50	0.10
		N.E.	401.79	187.28	0.47	882.60	43.80	3.13	0.41
		Summer	23.13	36.74	1.59	142.60	0.00	6.61	2.10
		Winter	162.20	68.70	0.42	357.70	58.50	3.29	0.71
7.	Upparu	S.W.	292.11	79.23	0.27	476.80	110.80	3.16	0.10
		N.E.	367.47	143.77	0.39	690.20	126.80	2.76	0.59
		Summer	24.48	39.44	1.61	133.60	0.00	4.69	1.78
		Winter	116.61	63.60	0.55	306.00	28.60	5.03	1.31
8.	Varaganadhi	S.W.	186.90	96.74	0.52	485.30	52.90	5.30	1.44
		N.E.	438.01	184.18	0.42	925.00	102.70	3.07	0.53
		Summer	43.17	54.98	1.27	207.60	0.00	5.27	1.75
		Winter	200.95	91.75	0.46	376.20	58.60	1.80	0.17
9.	Varattar-Nagalaru	S.W.	163.00	190.48	0.48	347.10	69.60	2.46	0.82
		N.E.	361.93	423.88	0.49	916.20	83.70	4.02	0.90
		Summer	31.24	46.57	1.41	205.70	0.00	9.63	2.49
		Winter	141.10	161.66	0.42	300.20	37.40	3.31	0.63
10.	Upper Vaigai	S.W.	152.65	75.33	0.49	347.20	14.10	3.01	0.16
		N.E.	345.00	141.45	0.41	628.40	128.00	2.12	0.26
		Summer	28.94	38.79	1.34	195.00	0.00	11.01	2.56
		Winter	150.11	76.46	0.51	315.70	9.90	2.79	0.39

TABLE 4

First ranked probability distribution for monthly rainfall data of LowerVaigai using different goodness-of-fit tests

S. No.	Month	Best-Fit Test Statistic Results		First Ranked Distribution
		Kolmogorov Smirnov	AndersonDarling	
1.	January	JSB (0.100)	Rec. (0.625)	JSB
2.	February	Uniform (0.133)	CS-2P (0.484)	Uniform
3.	March	Uniform (0.206)	Rec. (0.625)	Uniform
4.	April	Uniform (0.206)	Rec. (0.625)	Uniform
5.	May	Pert (0.104)	Rec. (0.625)	Pert
6.	June	JSB (0.093)	GP (0.294)	JSB
7.	July	JSB (0.065)	GEV (0.239)	JSB
8.	August	Error (0.068)	JSB (0.466)	Error
9.	September	JSB (0.057)	JSB (0.454)	JSB
10.	October	JSB (0.051)	Frechet (0.612)	JSB
11.	November	JSB (0.055)	CS-2P (0.396)	JSB
12.	December	JSB (0.050)	Error (0.127)	JSB

*JSB-Johnson SB, CS-Chi-Squared, GEV-Generalised Extreme Value, GP-General Pareto, Rec.-Reciprocal

TABLE 5

First ranked probability distribution for annual rainfall of different sub-basins using different goodness-of-fit tests

S. No.	Sub basin	Best-Fit Test Statistic Results		First Ranked Distribution
		Kolmogorov Smirnov	AndersonDarling	
1.	Lower Vaigai	JSB (0.037)	CS-2P (0.428)	JSB
2.	Theniyaru	Error (0.069)	JSB (0.137)	Error
3.	Manjalaru	GEV (0.059)	GEV (0.160)	GEV
4.	Suruliyaru	LL-3P (0.086)	LL-3P (0.259)	LL-3P
5.	Sathaiyaru	Logistic (0.084)	LL-3P (0.309)	Logistic
6.	Sirumalaiyaru	LP3 (0.071)	GEV (0.171)	LP3
7.	Upparu	Frechet-3P (0.065)	GEV (0.189)	Frechet-3P
8.	Varaganadhi	Kumaraswamy (0.080)	Error (0.212)	Kumaraswamy
9.	Varattar-Nagalaru	Weibull (0.073)	JSB (0.203)	Weibull
10.	Upper Vaigai	GEV (0.052)	Gamma (0.313)	GEV

*JSB-Johnson SB, CS-Chi-Squared, LL-Log Logistic, GEV-Generalised Extreme Value, LP3-Log Pearson 3

2.8. *Maximum absolute error (MaxAE)*

The largest forecasted error, expressed in the same units as the dependent series. Like MaxAPE, it is useful for imagining the worst-case scenario for the forecasts.

$$\text{MaxAPE} = \max\left[\left|X(t) - \hat{X}(t)\right|\right]$$

2.9. *Normalized Bayesianinformation criterion (Normalized BIC)*

A general measure of the overall fit of a model that attempts to account for model complexity. It is a score based upon the mean square error and includes a penalty for the number of parameters in the model and the series' length. The penalty removes the advantage of models with

TABLE 6

Season-wise first ranked probability distribution using different goodness-of-fit tests

S. No.	Sub basin	Season	Best-Fit Test Statistic Results		First Ranked distribution
			Kolmogorov Smirnov	Anderson Darling	
1.	Lower Vaigai	S.W.	JSB (0.038)	JSB (0.197)	JSB
		N.E.	JSB (0.042)	CS-2P (0.489)	JSB
		Winter	Beta (0.107)	Rec. (0.625)	Beta
		Summer	JSB (0.082)	Rec. (0.625)	JSB
2.	Theniyaru	S.W.	P6 (0.077)	Burr (0.226)	P6
		N.E.	Laplace (0.087)	LL-3P (0.357)	Laplace
		Winter	Gamma (0.124)	JSB (0.411)	Gamma
		Summer	GEV (0.072)	Weibull-3P (0.235)	GEV
3.	Manjalaru	S.W.	GP (0.078)	GP (0.244)	GP
		N.E.	GP (0.056)	GP (0.185)	GP
		Winter	GEV (0.105)	GP (0.544)	GEV
		Summer	GEV (0.059)	GEV (0.160)	GEV
4.	Suruliyaru	S.W.	Pert (0.103)	GEV (0.329)	Pert
		N.E.	GEV (0.074)	JSB (0.165)	GEV
		Winter	GEV (0.122)	GP (0.513)	GEV
		Summer	Rayleigh (0.095)	GEV (0.370)	Rayleigh
5.	Sathaiyaru	S.W.	Beta (0.100)	JSB (0.315)	Beta
		N.E.	LL-3P (0.058)	GEV (0.130)	LL-3P
		Winter	Gamma (0.117)	JSB (0.779)	Gamma
		Summer	Burr (0.053)	Burr (0.117)	Burr
6.	Sirumalaiyaru	S.W.	P5-3P (0.067)	LL-3P (0.163)	P5-3P
		N.E.	JSB (0.096)	GP (1.049)	JSB
		Winter	Pert (0.088)	GEV (0.218)	Pert
		Summer	LP3 (0.071)	GEV (0.172)	LP3
7.	Upparu	S.W.	Logistic(0.069)	LL-3P (0.164)	Logistic
		N.E.	IG (0.072)	LL-3P (0.213)	IG
		Winter	Gamma (0.176)	GP (1.405)	Gamma
		Summer	Dagum (0.084)	Dagum (0.237)	Dagum
8.	Varaganadhi	S.W.	Burr (0.064)	Burr (0.186)	Burr
		N.E.	Weibull-3P (0.084)	GEV (0.160)	GEV
		Winter	GEV (0.115)	GP (0.394)	GEV
		Summer	JSB (0.061)	GP (0.168)	JSB
9.	Varattar-Nagalaru	S.W.	LL-3P (0.095)	GP (0.397)	LL-3P
		N.E.	Weibull-3P (0.066)	GEV (0.138)	Weibull-3P
		Winter	Gamma (0.117)	GP (0.567)	Gamma
		Summer	GumbelMax (0.085)	LL-3P (0.260)	Gumbel Max
10.	Upper Vaigai	S.W.	Burr-4P (0.093)	Burr-4P (0.234)	Burr-4P
		N.E.	JSB (0.014)	P6-4P (0.217)	JSB
		Winter	GP (0.227)	GP (0.322)	GP
		Summer	P6-4P (0.679)	GEV (0.389)	GEV

*JSB-Johnson SB, CS-Chi-Squared, LL-Log Logistic, GEV-Generalised Extreme Value, GP-General Pareto, P6-Pearson 6, IG-Inverse Gaussian, LP3-Log Pearson 3, P5-Pearson 5, Rec.-Reciprocal.

more parameters, making it easy to compare different models for the same series.

$$\mathrm{NBIC} = \ln(\mathrm{MSE}) + k\frac{\ln(n)}{n}$$

2.9.1. *Diagnostic checking*

In this step, one can see whether the chosen model fits the data reasonably well. One simple test of the selected model is to see if the residuals estimated from this model white noise; if they are, one can accept the precise fit; if not, one starts the process afresh; thus, the Box-Jenkins Methodology is an iterative process.

2.9.2. *Forecasting*

One of the reasons for the popularity of ARIMA modeling is its success in forecasting.To forecast the values of a time series, the basic Box-Jenkins strategy is as follows.

(*i*) First, examine the stationarity. This step can be done by computing the autocorrelation function (ACF) and partial autocorrelation (PACF) or a standard root analysis.

(*ii*) If the time series is not stationary, the difference of the time series, one or more times to achieve stationarity.

(*iii*) The stationary time series' ACF and PACF are then computed to determine if the series is purely autoregressive or purely of the moving average type or a mixture of the two.

(*iv*) The tentative model is then estimated.

(*v*) The residuals from this model are examined to find out if they are white noise. If they are, the tentative model is probably a good approximation to the underlying stochastic process. If they are not, the process is started all over again. Therefore, the Box-Jenkins method is an iterative one. The model finally selected can be used for forecasting.

3. Results and discussion

The 34 years (1976-2009) data from 10 different sub-basins were statistically analysed for getting the monthly and seasonal variation of rainfall. The test statistics for these data were generated and compared to examine the nature of each series. Table 2 shows the monthly summary statistics for the Lower Vaigai basin. The maximum monthly average rainfall was received during October, November and December, which the north-east monsoon contributed. The average monthly rainfall for the other nine sub-basins also followed a similar trend.

Table 3 shows the summary statistics of seasonal rainfall for the study area. It is evident from the table that the North-East monsoon period received maximum average rainfall. The Lower Vaigai basin received the highest and the Manjalaru basin received the lowest average seasonal rainfall of 566.52 mm and 315.91 mm, respectively, during the North-East monsoon season. The monthly and seasonal rainfall data were fitted with multiple probability distributions. The first ranked probability distribution was selected for seasonal, annual and monthly rainfall for all ten sub-basins. The first rank probability distribution was calculated based on the two goodness-of-fit tests, *viz*., Kolmogorov Smirnov and Anderson Darling tests. The probability distribution having the highest rank with lower test statistics was chosen as the best-fit probability distribution for the particular data series.

The first ranked probability distribution for the Lower Vaigai basin's monthly rainfall data using different goodness-of-fit tests was given in Table 4.

The monthly rainfall data of the Lower Vaigai basin for January, June, July, September, October, November and December fits best with the Johnson SB distribution. In contrast, the data for February, March and April does best with the uniform distribution. The rainfall data of May month for the basin was fitted well with the Pert distribution function.

The first ranked probability distribution for different sub-basins' annual rainfall using other goodness-of-fit tests was given in Table 5. Johnson SB distribution was the best-fit probability distribution for the annual rainfall data of the Lower Vaigai basin. In contrast, GEV was the best fit distribution for the Manjalaru basin for the same time series (Table 5). Kumaraswamy's double bounded distribution fitted well with the annual rainfall data of the Varaganadhi basin with a test statistic of 0.080.

Table 6 shows the first ranked probability distribution for seasonal rainfall data. The annual rainfall of the study area was distributed into four seasons, *viz*., southwest monsoon, northeast monsoon, winter and summer. The northeast monsoon contributed approximately 60% of the rainfall in the study area.

The Johnson SB distribution fitted best with the northeast monsoon data of the Lower Vaigai, Sirumalaiyaru and Upper Vaigai basins. GEV was the best fit distribution for N-E monsoon data of Suruliyaru and Varaganadhi basin.

TABLE 7

Model validation

S. No.	Sub basin	ARIMA Model Type (p,d,q)	Model Fit statistics						
			R^2	RMSE	MAPE	MAE	MaxAPE	MaxAE	Norm. BIC
1.	Lower Vaigai	(0,1,0)	-0.87	398.76	47.05	310.21	537.72	1010.71	12.08
2.	Theniyaru	(1,1,2)	0.92	185.09	25.42	94.95	356.42	10.97	0.92
3.	Manjalaru	(0,0,0)	0.19	283.61	46.46	218.70	263.16	850.49	11.50
4.	Suruliyaru	(0,0,0)	0.73	232.55	24.34	164.47	561.43	11.00	0.74
5.	Sathaiyaru	(0,0,0)	0.13	210.49	22.59	165.33	67.93	429.56	10.91
6.	Sirumalaiyaru	(1,0,1)	0.20	186.05	18.47	146.23	65.99	419.72	10.76
7.	Upparu	(1,1,2)	0.68	189.17	18.55	137.87	92.64	425.21	10.91
8.	Varaganadhi	(1,1,1)	0.05	238.79	22.67	183.51	50.04	656.45	11.27
9.	Varattar-Nagalaru	(1,1,0)	-0.22	250.34	32.23	206.70	93.17	497.18	11.26
10.	Upper Vaigai	(1,1,5)	0.77	223.44	30.24	167.66	227.70	459.05	11.56

TABLE 8

Forecasting of annual rainfall of different sub-basins using the ARIMA model

S. No.	Sub basin	Annual Rainfall (mm)										
		2008		2009		2010		2015		2020		2025
		Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Exp.
1.	Lower Vaigai	1281.9	1297.2	953.3	968.6	821.0	983.9	1124.0	1060.2	755.1	1136.6	1213.0
2.	Theniyaru	608.0	594.4	617.8	821.3	665.8	642.9	581.4	665.2	838.9	668.9	669.2
3.	Manjalaru	877.4	896.3	504.8	509.0	936.6	923.3	880.0	990.6	1193.0	1057.9	1130.0
4.	Suruliyaru	862.2	900.3	748.3	864.8	956.1	890.0	807.3	901.0	880.3	910.0	920.0
5.	Sathaiyaru	885.4	933.9	759.9	742.0	805.8	950.1	946.2	990.6	897.5	1031.0	1071.5
6.	Sirumalaiyaru	920.0	903.0	713.0	730.0	837.8	898.0	972.5	906.0	842.9	914.0	922.0
7.	Upparu	816.0	851.9	742.1	776.0	800.7	831.9	836.4	785.5	947.8	821.7	796.2
8.	Varaganadhi	832.5	802.0	670.9	658.0	896.0	874.9	878.4	916.9	1129.4	938.2	959.5
9.	Varattar-Nagalaru	764.4	728.0	588.1	577.0	697.3	630.4	765.4	642.8	849.6	638.4	634.6
10.	Upper Vaigai	701.5	711.0	700.4	699.0	821.0	726.7	849.0	817.9	836.6	845.5	875.3

*Obs. - Observed, Exp. – Expected.

After finding the best-fit probability distribution for each data series, the annual rainfall for the upcoming years was forecasted using Box-Jenkins's methodology. The data from the year 2007to 2009 was taken for model validation. The forecasting of the rainfall data was done using the corresponding best fit probability distributions. Table 7 shows the results of model validation. The data from different sub-basins were fitted with other ARIMA models. The model with the maximum R^2, minimum RMSE, minimum MAPE, minimum MaxAPE, minimum MaxAE and minimum Normalized BIC was chosen as the best model.

The validation results show that the ARIMA (1,1,2) model fitted with the Theniyaru basin data is the best compared with the other models. The ARIMA (1,1,5), ARIMA (0,0,0) and ARIMA (1,1,2) models of the Upper Vaigai, Suruliyaru and Upparu basin could also be considered as the best models for the respective data sets. R^2 value represents the proportion of variance explained by the fit. A negative R^2 value for the Lower Vaigai and Varattar-Nagalaru basin indicates that the fit is worse than just fitting a horizontal line. The R^2 is negative only when the chosen model does not follow the trend of the observed data.

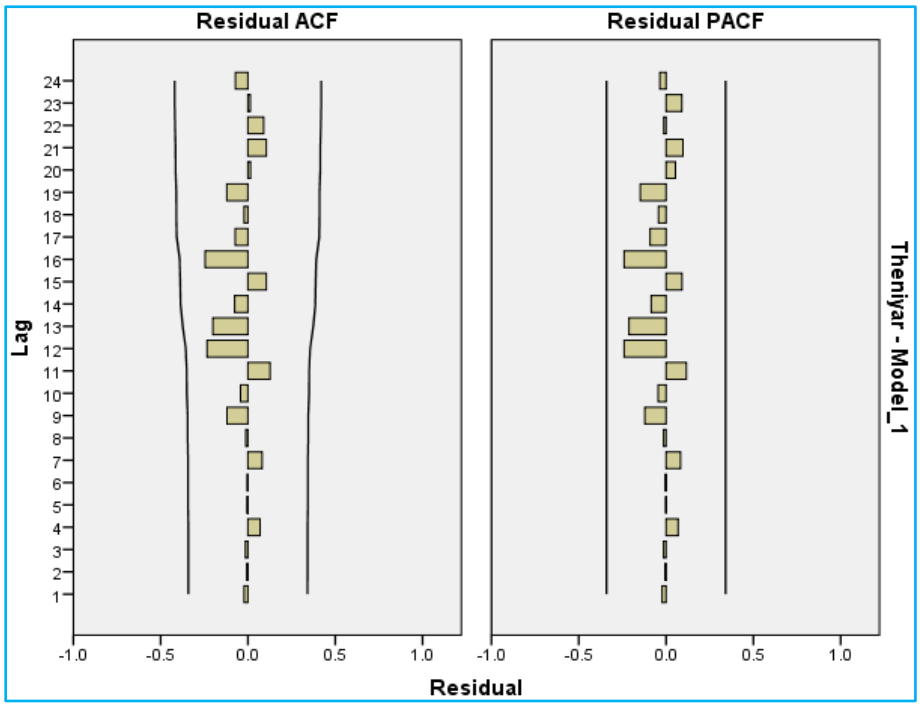

Fig. 1. ACF and PACF graphs of residuals for the best fit model of the annual rainfall of the Theniyarubasin

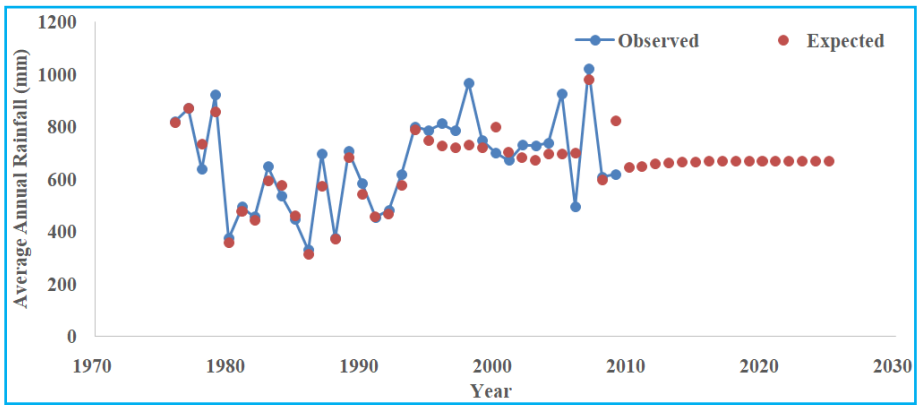

Fig. 2. Observed and expected annual rainfall of the Theniyaru basin

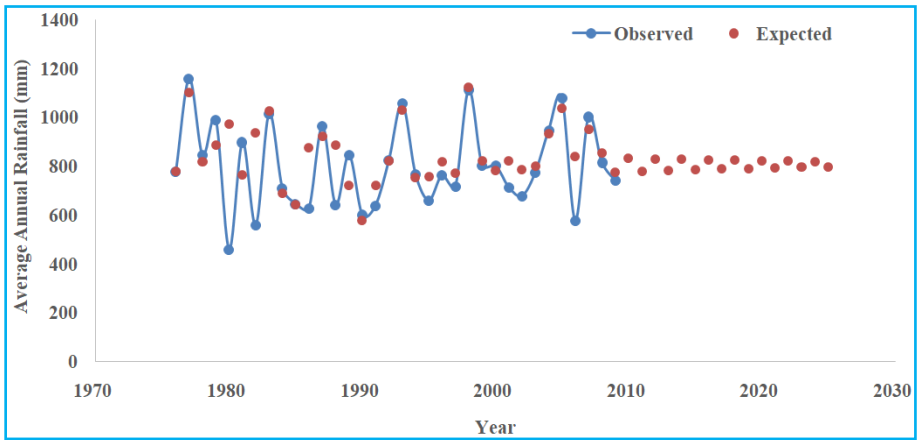

Fig. 3. Observed and expected annual rainfall of the Upparu basin

The best fit models were put under the diagnostic check of the residuals with the ACF and PACF graph's help. Fig. 1 shows the ACF and PACF graphs of residuals for the best fit model of the annual rainfall of the Theniyaru basin.

The average annual rainfall for 2010, 2015, 2020 and 2025 for the ten sub-basins was predicted and the validated results for the observed annual rainfall in 2010, 2015 and 2020 are given in Table 8.

For example, the observed and expected annual rainfall for the Theniyaru basin (Fig. 2) shows that the observed data agrees with the expected annual rainfall. It indicates the model accuracy's in forecasting the annual rainfall.

For the Theniyaru basin, the expected rainfall for 2010, 2015, 2020 and 2025 was 642.9 mm, 665.2 mm, 668.9 mm and 669.2 mm, respectively. For the year 2025, an average annual rainfall of 875.3 mm, 920 mm and 796.2 mm is expected in the Upper Vaigai, Suruliyaru and Upparu basins. Out of these four sub-basins, the Upparu basin shows a decreasing trend in average annual rainfall from 2009 to 2025 (Fig. 3).

The average annual rainfall of the Upparu basin is expected to decrease from 816 mm in 2009 to 769.2 mm in 2025.

The forecasted annual rainfall will be helpful in crop planning and irrigation planning in the study area. Moreover, it will also be beneficial in taking precautions against the probable extreme natural events like floods and drought in the study area.

4. Conclusion

Our study statistically analysed the ten sub-basins (Vaigai river) rainfall data from 1976 to 2009 with various probability distributions and found the best fit probability distribution for monthly, seasonal and annual data series. The Kolmogorov Smirnov and Anderson Darling goodness-of-fit tests were used for ranking the distribution of various time series of data. Later, the ARIMA model using Box-Jenkin's methodology was used for forecasting the average annual rainfall for future years. The model validation was done from 2008 to 2010, 2015 and 2020.

Our study found ARIMA (1,1,2), ARIMA (1,1,5), ARIMA (0,0,0) and ARIMA (1,1,2) to be the best models for forecasting the annual rainfall of Theniyaru, Upper Vaigai, Suruliyaru and Upparu basin respectively. The study predicted the average annual rainfall of 10 sub-basins for 16 years from 2009 to 2025. The observed and predicted rainfall showed a good agreement, evident in the model's accuracy in predicted data. The forecasting results would help plan and manage irrigation of the crops in Vaigai river sub-basins.

Disclaimer : The contents and views expressed in this study are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Anderson, O., 1976, "Time Series Analysis and Forecasting: The Box-Jenkins Approach", London: Butterworths.

Box, G. E. P. and Jenkins, G. M., 1976, "Time series analysis: Forecasting and control, Revised Edition", San Francisco: Holden-Day.

Chatfield, C., 1984, "The Analysis of Time Series. An Introduction, 3rd edition", London : Chapman and Hall.

Chattopadhyay, S. and Chattopadhyay, G., 2010, "Univariatemodelling of summer monsoon rainfall time series : comparison between ARIMA and ARNN", *C. R. Geosci*., **342**, 100-107.

Eni, D. and Adeyeye, F. J., 2015, "Seasonal ARIMA Modeling and Forecasting of Rainfall in Warri Town, Nigeria", *Journal of Geoscience and Environment Protection*, **03**, 91-98.

Fisher, R.A., 1925, "The influence of the rainfall on the yield of wheat at Rothamsted", *Philosophical Transactions of the Royal Society of London*, Series B, **213**, 89-142.

Judge, G., Hill, R. C., Griffiths, W. E., Lutkepohl, H. and Lee, T. C., 1982, "Introduction to the Theory and Practice of Econometrics", Wiley, New York.

Kaushik, I. and Singh, S. M., 2008, "Seasonal ARIMA model for forecasting of monthly rainfall and temperature", *Journal of Environmental Research and Development*, **3**, 2.

Mishra, P., Fatih, C., Vani, G., Lavrod, J. M., Jain, V., Mishra, P. C., Choudhary, A. K. and Dubey, A., 2021, "Modeling and forecasting of metrological factors using ARCH process under different errors distribution specification", *MAUSAM*, **72**, 2, 301-312.

Narayanan, P., Sarkar, S., Basistha, A. and Sachdeva, K., 2013, "Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India", *C. R. Geosci*., **345**, 22-27.

Narayanan, V. L. S., Gurubaran, K. S. and Emperumal, K., 2016, "Shrinking equatorial plasma bubbles", *J. Geophys.Res. Space Physics*, 121. doi : 10.1002/2016JA022633.

Pankratz, A., 1983, "Forecasting with Univariate Box-Jenkins Model", New York : John Wiley and Sons.

Ray, C. R., Senapati, P. C. and Lal, R., 1980, "Rainfall analysis for crop planning at Gopalpur, Orissa", *J. Agril. Eng*., I.S.A.E., **17**, 384.

Sharma, M. and Singh, J. B., 2010, "Use of probability distribution in rainfall analysis", *New York Science Journal*, **3**, 40-49.

Valipour, M., 2015, "Long-term runoff study using SARIMA and ARIMA models in the United States",*Meteorol.Appl*., **22**, 592-598.