Original Research Article

Stochastic Time Series Analysis, Modeling, and Forecasting of Weekly Rainfall Using SARIMA Model

Abstract

Rainfall holds critical significance for water resource applications, particularly in rainfed agricultural systems. This study employs the Autoregressive Integrated Moving Average (ARIMA) technique, a data mining approach commonly used for time series analysis and future forecasting. Given the increasing importance of climate change forecasting in averting unexpected natural hazards such as floods, frost, forest fires, and droughts, accurate weather data forecasting becomes imperative. The objective of this study was to develop a Seasonal Auto-Regressive Integrative Moving Average (SARIMA) model for forecasting weekly rainfall in Junagadh Station, Gujarat. Utilizing 53 years of historical data (1963 to 2016), the SARIMA model predicts weekly rainfall for the subsequent five years (2018 to 2022). Through comprehensive evaluation using ACF and PACF plots, AIC, SBC, MAPE, and MAE values,

the study identifies SARIMA (0,0,4)(0,1,1)52 as the optimal model, offering the most accurate prediction. The robust results affirm that the SARIMA model provides reliable and satisfactory weekly rainfall predictions. This research contributes valuable insights into the precision and efficacy of SARIMA models for rainfall forecasting, aiding in strategic water resource management in the Junagadh region.

Key Words : SARIMA, AIC, BIC, MAPE, SIC

1. INTRODUCTION

Efficient water resource management relies heavily on accurately forecasting rainfall for a given area or station (Kumar et al., 2021; Kumar et al., 2021a; Kumar et al., 2022). In the context of Indian agriculture, the southwest monsoon (June-September) plays a pivotal role in the agrarian economy, with adequate rainfall being essential for robust crop production (Kumar et al., 2021). Rainfall, among various hydrological parameters, is crucial for tasks such as irrigation planning, runoff modeling, and drought and flood management. The dynamic nature of rainfall patterns, influenced by changing climatic conditions, gives rise to challenges like flooding, landslides, and drought (Shivhare et al., 2017), significantly impacting agriculture and farming. In a country like India, where agriculture is a backbone, the success or failure of crops is a primary concern, and even slight variations in seasonal rainfall and temperature can have devastating effects on crops (Shivhare et al., 2018). The runoff characteristics, both in terms of quantity and quality, in the majority of watersheds, spanning from micro to macro scales, are significantly shaped and controlled by spatiotemporal variations in rainfall. (Ram, Bhavin et.al, 2023a)

Accurately predicting future climate data is a challenging task (Nikam and Meshram, 2013). The accuracy and adequacy of rainfall data serve as the essential cornerstone for determining the ultimate success of any progressive endeavors in natural resource management

(Ram, Bhavin et.al, 2023b). Despite the development of various algorithms, achieving precise forecasting remains a challenge. Time series models, integral in meteorology and hydrology, tackle the key problem of forecasting in statistics and Data Science. Data transforms into a time series when sampled based on a time-bound attribute like days, months, and years, inherently possessing an implicit order. Forecasting involves predicting future values using this ordered data. Stochastic models, evolving over time (Box and Jenkins, 1994), encompass autoregressive (AR) models, moving average (MA) models of different orders (Gupta and Kumar, 1994, and Verma, 2004), and auto-regressive moving average (ARMA) models of discrete orders (Katz and Skaggs, 1981; Chhajed, 2004; Katimon and Demon, 2004) for annual streamflow. Two widely used forecasting algorithms, ARIMA and SARIMA, address the challenge. ARIMA considers past values (autoregressive, moving average) to predict future values, while SARIMA incorporates seasonality patterns, making it more potent for forecasting complex data spaces containing cycles. The ARIMA model emerges as a valuable tool, handling various dimensions related to univariate time series model selection, parameter optimization, and prediction. In the current study, our focus was on developing a seasonal rainfall forecasting model to predict the weekly rainfall time series for Junagadh city in Gujarat, India, utilizing 58 years (1965-2022) of weekly rainfall data.

2. MATERIAL AND METHODS

2.1 Study location

Junagadh is geographically situated between latitude 21°31'23.29" N and longitudes 70°27'17.90" E, at an altitude of 86 meters above mean sea level in the South Saurashtra region of Gujarat state. The climate of the study area is subtropical and semi-arid, characterized by an average annual rainfall of 929.81 mm, which is concentrated between mid-June and mid-October. The average annual pan evaporation is 5.6 mm/day. The coldest month is January,

with a mean monthly temperature ranging from 7°C to 15°C. The maximum monthly temperature is recorded in May, varying between 29.50°C to 39.40°C. Relative humidity fluctuates between 45% and 89%, while wind speeds range from 2 to 9.70 km/h.

2.2 Data

In this study, weekly rainfall data spanning 58 years (1965-2022) were collected from the Agrometeorology Department of Junagadh Agricultural University, Junagadh. Forecasts were made for the five years (2018-2022) using a seasonal ARIMA model.

2.3 Methodological Description

Seasonal ARIMA (SARIMA) modelling

An autoregressive model of order p is conventionally classified as AR(p), and a moving average model with q terms is known as MA(q). A combined model that includes p AR-terms and q MA-terms is referred to as an ARMA(p, q) model. To address non-stationarity, a generally non-stationary time series is transformed into a stationary one by computing differences shifted by d lags, where in most cases, d=1. Such a model is then categorized as ARIMA(p, d, q), where the symbol "I" signifies "integrated." The general form of the above model, describing the current value y(t) of a time series by its own past, is expressed as:

$$\boldsymbol{\phi_p(B)\emptyset_P(B^s)\nabla^d\nabla_s^{\,D}y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t} \qquad (1)$$

Where $\phi_p(B) = 1 - \phi_1 B - \cdots \phi_p B^p$= Non Seasonal autoregressive (AR) operator; $\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$=Non Seasonal moving average operator (MA) operator; $\emptyset_P(B^s) = 1 - \emptyset_1 B^s - \cdots - \emptyset_p B^{sP}$= Seasonal auto regressive (SAR) operator; $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \cdots - \Theta_Q B^{sQ}$= Seasonal moving average operator (SMA). Here, B = backshift operator (i.e. $B^1Yt = Y_{t-1}$, $B^2Y_t = Y_{t-2}$ and so on); s = the seasonal lag; εt= sequence of independent normal error variables with mean zero and variance σ2; p and q are orders of non-

seasonal auto-regression and moving average parameters respectively and P and Q are that of the seasonal auto regression and moving average parameter respectively; d and D denote the non-seasonal and seasonal differences respectively.

The main stages in setting up an ARIMA forecasting model include model identification, model parameter estimation, and diagnostic checking for the identified model's appropriateness for modeling and forecasting. The classical Box-Jenkins model describes stationary time series. Thus, tentatively identifying a Box-Jenkins model requires verifying the time series for stationarity. Stationary models assume that the process remains in equilibrium around a constant mean level, indicated when the plotting shows that the data fluctuates around its constant mean. A cursory examination of the graph of the data and the structure of autocorrelation and partial correlation coefficients at various lags may provide clues to the presence of stationarity. If the model is found to be non-stationary, stationarity could mostly be achieved by differencing the series. The next step in the identification process is to find the initial values for the orders of seasonal and non-seasonal parameters, p, q, and P, Q. These values could be obtained by looking for significant autocorrelation and partial autocorrelation coefficients.

After choosing the most appropriate model (step 1 above), the model parameters are estimated (step 2) using the least square method. In this step, values of the parameters are chosen to minimize the Sum of the Squared Residuals (SSR) between the real data and the estimated values. Generally, a nonlinear estimation method is used to estimate the identified parameters to maximize the likelihood (probability) of the observed series given the parameter values. The methodology uses the following criteria in parameter estimation:

a) The estimation procedure stops when the change in all parameter estimates between iterations reaches a minimal change of 0.001.

b) The parameters estimation procedure stops when the SSR between iterations reaches a minimal change of 0.0001.

In the diagnostic checking step (step three), the residuals from the fitted model are examined for adequacy. This is typically done through correlation analysis using residual ACF plots and goodness-of-fit tests via Chi-square statistics. If most of the sample autocorrelation coefficients of the residuals are within the limits ±1.96/√N, where N is the number of observations upon which the model is based, then the residuals are white noise, indicating that the model is a good fit. Otherwise, if the autocorrelations are not white noise, the model may not adequately represent our time series. In the last phase, i.e., forecasting, we calculate the point extrapolated prognosis of the time series and eventually the confidence lag of the prognosis.

Evaluation Criteria

The other statistical criteria adopted in the study are:

1) Akaike Information Criterion (AIC)

The AIC is given by

$$AIC = n \ln \sigma^2 + n + \frac{2(k+1)}{n-k-2} \qquad (2)$$

Where n is the size of the sample used for fitting, k is the number of parameters excluding constant terms, and $\sigma^2(\varepsilon)$ is the maximum likelihood estimate of the residual variance.

2) **Schwarz information criterion** (SIC)

The SIC is given by

$$SIC = n \ln \sigma^2(\varepsilon) + n + k \ln n \qquad (3)$$

Where n, k and $\sigma^2(\varepsilon)$ are defined in the same way as for the AIC statistic.

3) Mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{X_t - O_t}{O_t} \right| \times 100 \qquad (4)$$

Where X_t= forecast value at time t; O_t= actual value at time t; N= number of weeks considered for forecasting.

4) Mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_t - O_t| \qquad (5)$$

Where X_t= forecast value at time t; O_t= actual value at time t; N= number of weeks considered for forecasting.

3. RESULTS AND DISCUSSION

In the present study the time series of weekly rainfall data from 1965 to 2017 were used to develop the Seasonal ARIMA (SARIMA) model and the prediction was made for next five years (2018-2022) using the developed model. The forecasted values than used for validation of developed SARIMA model.

3.1 Analysis of Weekly Rainfall Time Series used for Model Development

Data of weekly rainfalls were analysed using Statistical Analysis System (SAS) software. Auto correlation function (ACF) and Partial Auto correlation function (PACF) of the original time series of weekly rainfall are shown in figure 1.

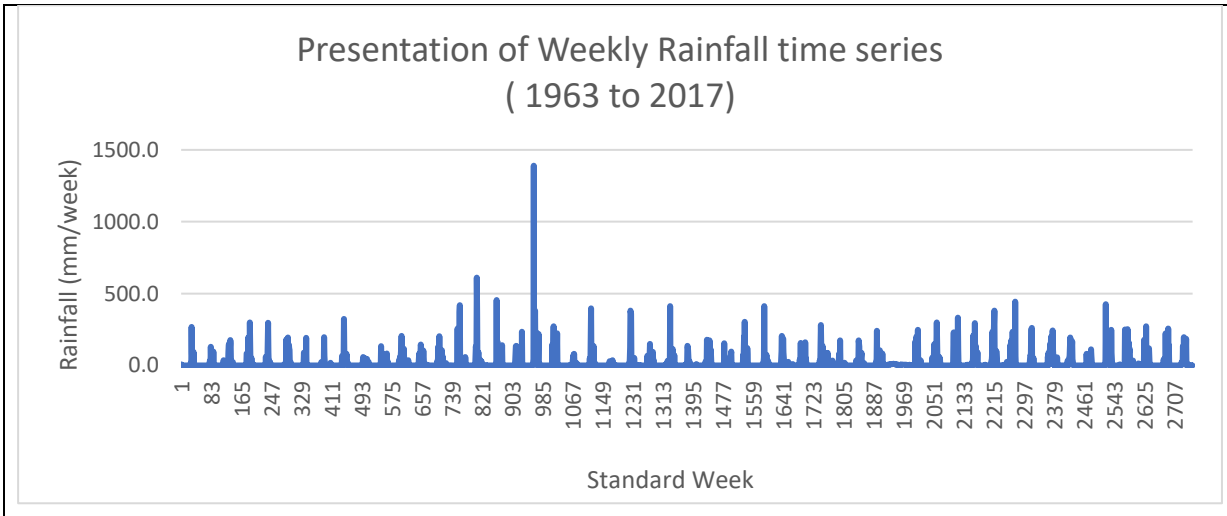

Figure:1 Graph of weekly Rainfall data series from 1965 to 2017

Key statistics summarizing the weekly rainfall time series data used for prediction spanning the period from 1963 to 2017 were computed. The mean weekly rainfall is calculated at 17.43, with a standard deviation of 56.75, indicating a notable variability in the data. The dataset consists of 2743 observations (N). The Augmented Dickey-Fuller (ADF) test results are presented, revealing significant negative values for the Zero Mean ADF (-36.45), Single Mean ADF (-39.03), and Trend ADF (-39.06). These ADF test statistics suggest a high likelihood of stationarity in the time series data, particularly with the consistently low p-values associated with the ADF tests, indicating a rejection of the null hypothesis of non-stationarity. The negative values further reinforce the presence of a stable trend in the data, laying a foundation for the application of time series forecasting models.

The table 2 provides diagnostic measures for a time series, showcasing autocorrelation (AutoCorr) and partial correlation (Partial) coefficients at different lags. The Ljung-Box Q statistic with associated p-values is used to test the null hypothesis of no autocorrelation in the residuals. Notably, all autocorrelation coefficients at various lags are significant, as indicated by the low p-values (<0.0001). The decreasing pattern in autocorrelation coefficients with increasing lags suggests a declining influence of past observations on the current one. The negative partial correlation coefficients imply that the effect of past observations is adequately captured by the model. These results support the suitability of the model for forecasting as they align with the assumption of white noise residuals, essential for robust time series modeling.

Table:1 Time series Basic diagnostics

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	-	-	0	1.0000
1	0.2847	222.561	<.0001*	1	0.2847
2	0.1443	279.741	<.0001*	2	0.0688
3	0.1365	330.972	<.0001*	3	0.0861
4	0.1840	424.070	<.0001*	4	0.1294
5	0.1233	465.845	<.0001*	5	0.0310
6	0.1141	501.685	<.0001*	6	0.0489

Table:1 Time series Basic diagnostics					
Lag	**AutoCorr**	**Ljung-Box Q**	**p-Value**	**Lag**	**Partial**
7	0.1304	548.496	<.0001*	7	0.0638
8	0.0290	550.819	<.0001*	8	-0.0650
9	0.0099	551.092	<.0001*	9	-0.0261
10	-0.0099	551.362	<.0001*	10	-0.0429
11	-0.0133	551.851	<.0001*	11	-0.0339
12	-0.0289	554.157	<.0001*	12	-0.0245
13	-0.0032	554.185	<.0001*	13	0.0130
14	-0.0651	565.877	<.0001*	14	-0.0637
15	-0.0477	572.147	<.0001*	15	0.0020
16	-0.0668	584.453	<.0001*	16	-0.0352
17	-0.0785	601.493	<.0001*	17	-0.0396
18	-0.0778	618.215	<.0001*	18	-0.0214
19	-0.0813	636.475	<.0001*	19	-0.0338
20	-0.0868	657.292	<.0001*	20	-0.0355
21	-0.0848	677.179	<.0001*	21	-0.0150
22	-0.0872	698.220	<.0001*	22	-0.0334
23	-0.0869	719.121	<.0001*	23	-0.0216
24	-0.0885	740.823	<.0001*	24	-0.0265
25	-0.0880	762.299	<.0001*	25	-0.0295

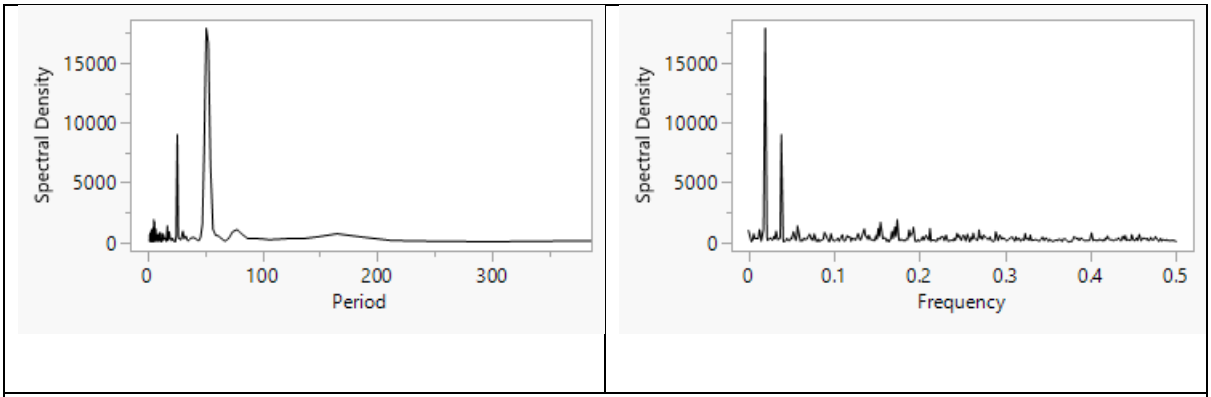

Figure:2 Spectral density plots of weekly rainfall time series

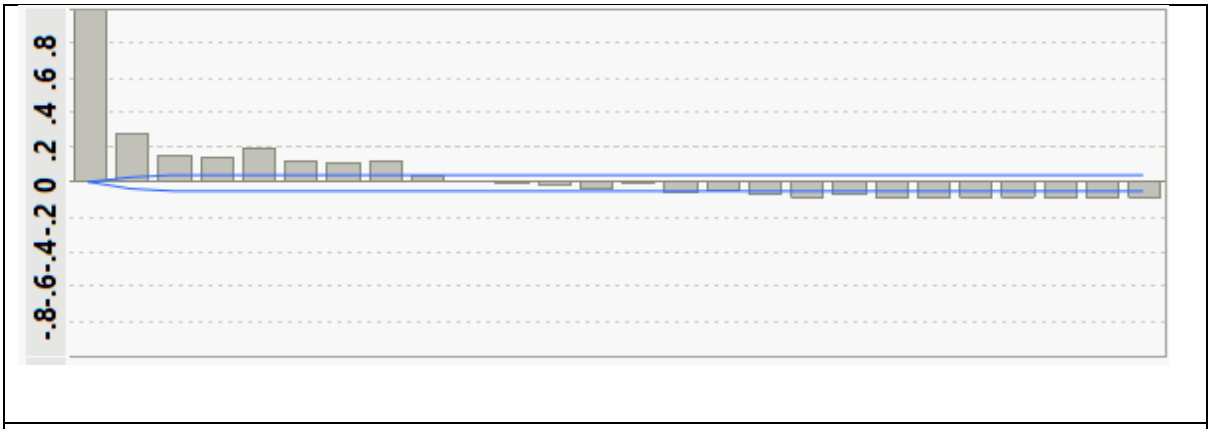

Figure: 3 ACF plot of weekly rainfall time series

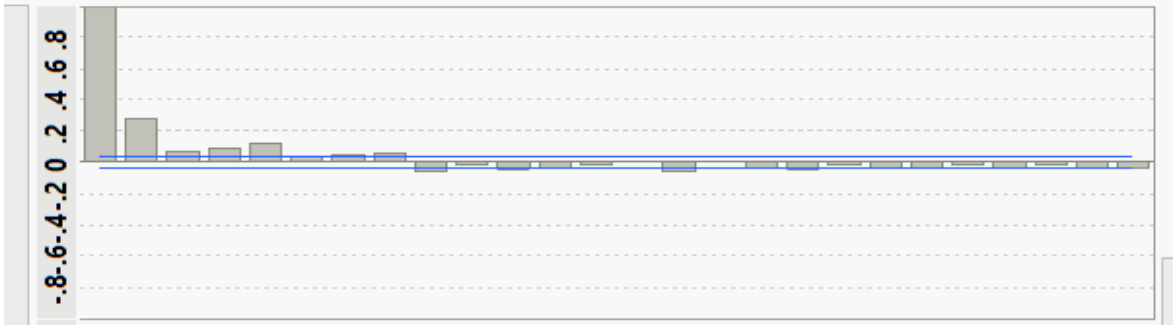

Figure : 4 PACF plot of weekly rainfall time series

3.2 Model Development and Parameter Estimation

Figures 3 and 4 provide a detailed depiction of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), offering profound insights into the periodic nature of the variables associated with weekly rainfall. These graphical representations consistently reveal patterns indicative of seasonal variations within the time series. Building upon these findings, we assume a yearly period of 52 weeks for the given rainfall time series.

Figures 5 and 6 provide a concise overview of the SARIMA (0,0,4)(0,1,1) model's performance in predicting weekly rainfall. Figure 5 illustrates the model's predictions, showcasing its ability to capture both non-seasonal and seasonal components. The parameters (0,0,4) indicate the absence of non-seasonal autoregressive and moving average effects, while (0,1,1) signifies first-order differencing in the seasonal part for stationarity. This visualization offers a clear representation of how well the SARIMA model aligns with observed weekly

rainfall trends. In Figure 6, the Residual Plot for SARIMA (0,0,4)(0,1,1) allows for a quick assessment of model residuals. A well-behaved residual plot indicates a well-fitted model, and analyzing it provides insights into the accuracy and reliability of the SARIMA model in predicting weekly rainfall.

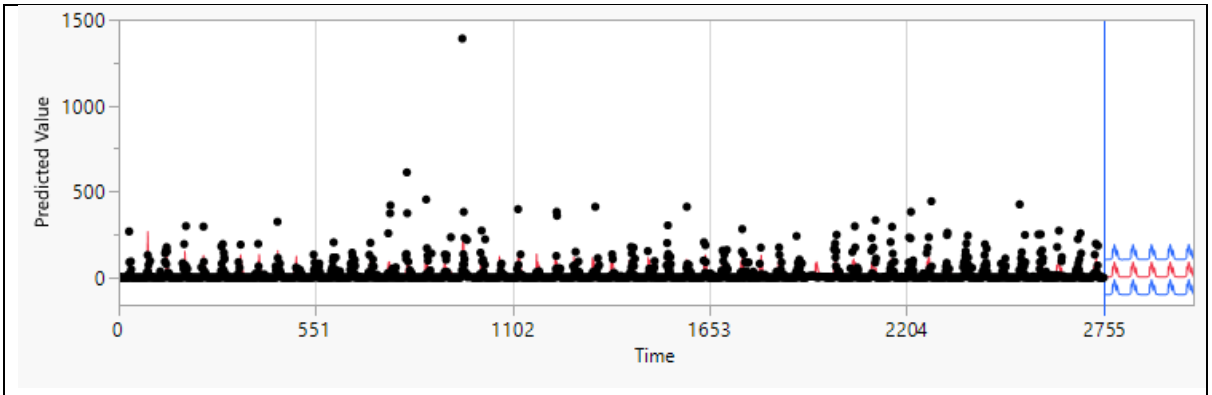

Fig:5 Prediction of weekly rainfall using SARIMA (0,0,4) (0,1,1)

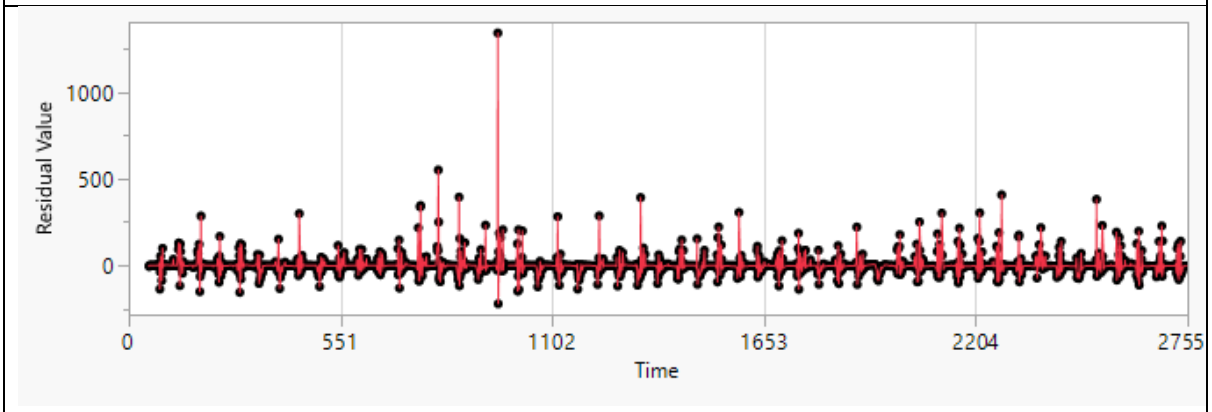

Fig:6 Residual Plot for SARIMA (0,0,4) (0,1,1)

Following the essential stationarities steps, we systematically explored various orders of Seasonal ARIMA models tailored to the weekly rainfall series. The model selection process involved a thorough assessment, considering not only the ACF and PACF charts but also key metrics such as the Akaike Information Criterion (AIC), Mean Absolute Percentage Error

(MAPE), and Mean Absolute Error (MAE). Following a meticulous evaluation, the Seasonal ARIMA model (0,0,4) (0,1,1) 52 emerged as the optimal choice for accurately forecasting weekly rainfall in the Junagadh region. Subsequently, a comprehensive five-year forecast spanning 2018 to 2022 was executed. Visual representations of the selected model dynamics are thoughtfully presented in Figures 7 and 8, while a detailed breakdown of parameters and statistical insights is thoroughly documented in Table 2 and Table 3.

Table:2 SARIMA (0,0,4) (0,1,1) Model Summary	
DF	2672
Sum of Squared Innovations	7060778.92
Sum of Squared Residuals	7258474.63
Variance Estimate	2642.50708
Standard Deviation	51.4053215
Akaike's 'A' Information Criterion	28827.5382
Schwarz's Bayesian Criterion	28862.8951
RSquare	0.16994209
RSquare Adj	0.16839635
MAPE	.
MAE	19.5382271
-2LogLikelihood	28815.5382

Table-3 SARIMA (0,0,4) (0,1,1) Parameter Estimates						
Term	**Factor**	**Lag**	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
MA1,1	1	1	-0.1159690	0.0193688	-5.99	<.0001*
MA1,2	1	2	0.0346565	0.0188537	1.84	0.0661
MA1,3	1	3	0.0436783	0.0193018	2.26	0.0237*
MA1,4	1	4	-0.0665500	0.0196808	-3.38	0.0007*
MA2,52	2	52	0.9500147	0.0106861	88.90	<.0001*
Intercept	1	0	0.1583824	0.0904021	1.75	0.0799

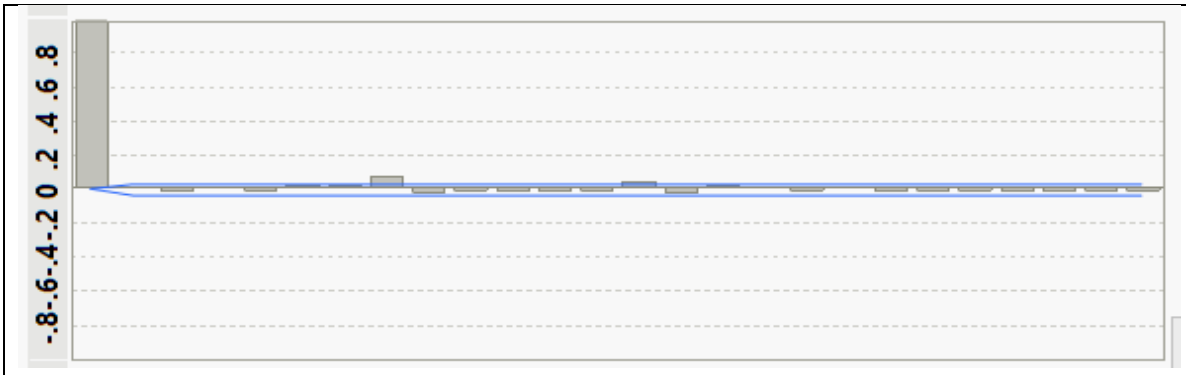

Fig:7 ACF Plot for SARIMA (0,0,4) (0,1,1)

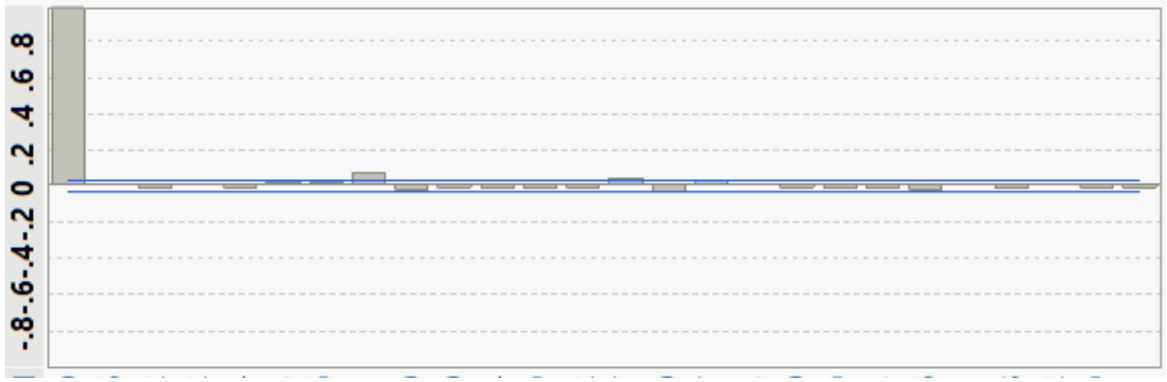

Fig:8 PACF Plot for SARIMA (0,0,4) (0,1,1)

3.3 Comparison of Actual and Predicted weekly rainfall value

Figure 9 serves as a visual guide for comparing the actual and predicted values of weekly rainfall over the five-year span from 2018 to 2022. The graph offers a detailed examination of how well the SARIMA model performs in forecasting weekly rainfall. A closer inspection reveals a remarkable proximity between the predicted time series and the actual data series. This visual coherence signifies the SARIMA model's exceptional capability to provide accurate and reliable forecasts of rainfall patterns. The model adeptly captures the nuances and fluctuations present in the observed data, emphasizing its effectiveness as a valuable forecasting tool. The visual representation in Figure 9 serves as a compelling endorsement of the SARIMA model's robust performance in predicting weekly rainfall values.

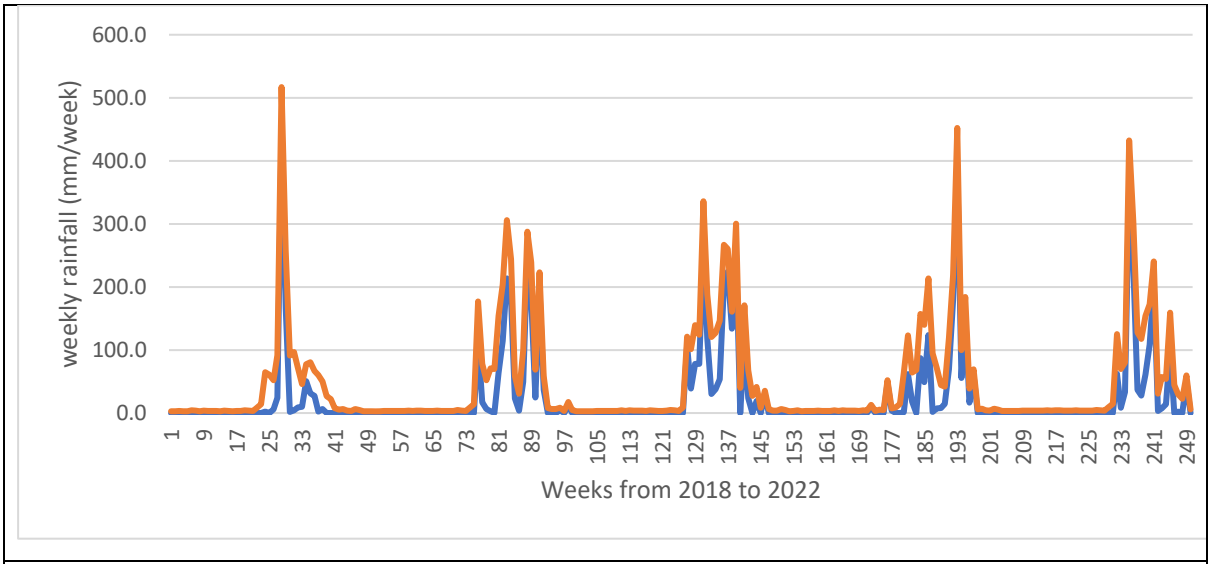

Fig:9 Comparison of Actual and Predicted weekly rainfall value of five years (2018-2022)

4. Conclusion

The study conclusively asserts the efficacy of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model as an indispensable tool for forecasting weekly rainfall in the Junagadh region. Boasting commendable accuracy, as evidenced by robust statistical measures, the SARIMA model emerges as a reliable asset for predicting the intricate patterns of weekly rainfall. This finding underscores the pivotal role of the Box-Jenkins methodology, which, through SARIMA, equips decision-makers with valuable insights. Decision-makers are empowered to forge better strategies and prioritize actions to fortify themselves against impending weather changes. Such strategic planning is particularly crucial given the potential enduring impacts of weather fluctuations on the water resources in Junagadh.

The SARIMA model's predictive prowess not only enhances forecasting precision but also facilitates proactive decision-making to navigate the dynamic nature of climatic conditions. By embracing SARIMA within the Box-Jenkins framework, decision-makers can not only anticipate and plan for upcoming weather variations but also establish resilient strategies for long-term water resource management. This holistic approach aids in setting priorities and allocating resources efficiently. In essence, the SARIMA model, bolstered by the

Box-Jenkins methodology, emerges as a key ally for decision-makers, offering a strategic advantage in mitigating the effects of weather changes and fortifying the water resources of the Junagadh region against the uncertainties of the future.

REFERENCES

Bender, M. and Stohodan S., (1994) Time-series modelling for long-range stream flow forecasting. Journal of Water Resources Planning and Management, ASCE, 120(6): 857-870.

Box, G.E.P. and Jenkins, G.M. (1994). Time series analysis, forecasting and control, Revised Edition, Holden-Day, San Francisco, California, United States.

Gupta, R.K. and Kumar, R. 1994. Stochastic analysis of weekly evaporation values, Indian J. Agric. Eng., 4(3-4):140-142.

Inderjeet K., and Singh. S. M., (2008) seasonal arima model for forecasting of monthly rainfall and temperature Journal of Environmental Research and Development Vol. 3 No. 2.

Katz, R.W. and Skaggs, R.H. 1981. On the use of autoregressive moving average processes to model meteorological time series, Monthly Weather Rev., 109: 479-484.

Kumar, U., Meena, V.S., Singh, S., Bisht, J.K. and Pattanayak, A. (2021a).Evaluation of digital elevation model in hilly region of Uttarakhand: FA case study of experimental farm Hawalbagh. Indian J. Soil Conserv., 49: 77-81.

Kumar, U., Panday, S.C., Kumar, J., Parihar, M., Meena, V.S., Bisht, J.K.and Kant, L. (2022). Use of a decision support system to establish the best model for estimating reference evapotranspiration in subtemperate climate: Almora, Uttarakhand. Agric. Eng. Int. CIGR J., 24(1): 41-50.

Kumar, U., Singh D.K., Panday, S.C., Bisht, J.K and Kant, L. (2022). Development and evaluation of seasonal rainfall forecasting (SARIMA) model for Kumaon region of Uttarakhand. Indian Journal of Soil Conservation, Vol. 50, No. 3, pp 190-198, 2022.

276 Kumar, U., Srivastava, A., Kumari, N., Rashmi, Sahoo, B., Chatterjee, C and Raghuwanshi,
277 N.S. (2021b). Evaluation of spatio-temporal evapotranspiration using satellite-based
278 approach and lysimeter in the agriculture dominated catchment. J. Indian Soc. Remote
279 Sens., 49: 1939-1950.
280 Mohan, S. and Arumugam N., (1995). Forecasting weekly reference evapotranspiration series.
281 Hydrological Science. Vol. 40(6), 689-702.
282 Nikam, Valmik B. and Meshram, B. B., (2013) "Modeling rainfall Prediction using data mining
283 method: A Bayesian approach",Computational Intelligence, Modelling and Simulation
284 (cimsim), 2013 Fifth International Conference on, 132-136,IEEE.
285 Nury A .H., Koch M., and M.J.B. Alam (2013) Time Series Analysis and Forecasting of
286 Temperatures in the Sylhet Division of Bangladesh, Environmental Science. 65-68
287 Popale P. G., and S. D. Gorantiwar (2014) Stochastic Generation and Forecasting Of Weekly
288 Rainfall for Rahuri Region International Journal of Innovative Research in Science,
289 Engineering and Technology An ISO 3297: 2007 Certified Organization Volume 3,
290 Special Issue 4, April 2014 Two days National Conference – VISHWATECH 2014.
291 Ram , B., Gaur , M. L., Patel , G. R., Kunapara , A. N., & Damor , P. A. (2023a). Stochastic
292 Disaggregation of Daily Rainfall Using Barlett Lewis Rectangular Pulse Model
293 (BLRPM): A Case Study of Middle Gujarat. International Journal of Environment and
294 Climate Change, 13(4), 37–47. https://doi.org/10.9734/ijecc/2023/v13i41710
295 Ram, Bhavin & Gaur, Murari & Patel, Gautam & Kunapara, A. & Pampaniya, Nirav & Damor,
296 P. & Balas, Duda. (2023b). Assessment of Diurnal Variability and Region-Specific
297 Connection across Intensity, Depth & Duration of Rainfall. International Journal of
298 Environment and Climate Change. 13. 595-606. 10.9734/ijecc/2023/v13i92275.

299 Shivhare, N., Kumar, A.L., Dwivedi, S.B., and Dikshit, P.K.S. (2019). ARIMA based daily
300 weather forecasting tool: A case study for Varanasi. MAUSAM, 70, 1 (January 2019),
301 133-140.
302 Shoba, G. and Shobha, G., (2014). “Rainfall prediction using Data Mining techniques: A
303 Survey”, Int. J. of Eng. and Computer Science, 3, 5, 6206-6211.