

Original Research Article

Forecasting Yield of Major Crops in Assam: A Time Series Approach

Abstract

This study aims to forecast the yield of five major crops in Assam—Rice, Wheat, Potato, Rapeseed & Mustard, and Arhar—using Autoregressive Integrated Moving Average (ARIMA) models. Time series data on crop yields were collected and analyzed to determine the most appropriate ARIMA models for each crop, based on model diagnostics and goodness-of-fit criteria. The selected models—ARIMA(0,1,1) for rice, ARIMA(2,1,0) for wheat, ARIMA(2,2,1) for potato, ARIMA(3,1,1) for rapeseed & mustard, and ARIMA(2,1,2) for arhar—were used to generate yield forecasts up to the year 2030. Results show a relatively stable yield trend for rice and wheat, while potato shows a sharp upward trend. Rapeseed & mustard and arhar exhibit moderate growth patterns. The widening confidence intervals suggest increasing uncertainty over time, emphasizing the need for continuous model updating. This study highlights the usefulness of ARIMA models in agricultural planning and policy formulation by providing reliable forecasts for strategic decision-making. The insights derived can support more efficient resource allocation, food security planning, and climate-resilient agricultural development in Assam.

Keywords: Box-Jenkins methodology, Stationary, AIC, BIC, Forecasting

1 Introduction

Agriculture is the backbone of Assam's economy, employing a significant portion of the population and contributing substantially to the state's Gross Domestic Product (GDP). According to the Agriculture Census of 2015-16, 70% of Assam's rural population depends on agriculture as their primary occupation, with the total number of farmer families recorded at 27,41,722 (State Level Bankers Committee, NE region). Accurate forecasting of crop yields is essential for efficient agricultural planning, food security, market regulation, and informed policymaking. In recent years, yield prediction has gained increasing importance in the face of growing uncertainties posed by climate variability, erratic rainfall, and changing agro-ecological conditions. Assam is known for its

diverse agricultural production, with rice, rapeseed and mustard, tea, various pulses, jute, sugarcane, potato, maize, oilseeds, and an array of horticultural crops contributing significantly to the state's agricultural output. Rice, the staple food crop, dominates in terms of both area and production. Other notable crops, oilseeds such as rapeseed and mustard, which are integral to edible oil production. Additionally, Assam produces a variety of pulses, vegetables, fruits, and spices, which cater to both domestic consumption and commercial markets. Timely and reliable forecasts of these crop yields can aid the government and agricultural stakeholders in making data-driven decisions regarding resource allocation, supply chain management, and climate-resilient interventions. This study employs the Autoregressive Integrated Moving Average (ARIMA) model—a widely used univariate time series forecasting technique—to model and forecast the yields of these five major crops namely rice, potato, wheat, rapeseed and mustard(R&M), and arhar(pigeon pea). Using historical yield data. ARIMA models are especially effective for capturing patterns in yield trends influenced by time-dependent factors and can provide short-term forecasts with a high degree of accuracy. By integrating statistical rigor with agricultural insights, this research aims to offer a forecasting framework that not only supports agronomic decision-making but also contributes to climate-resilient agricultural policy in Assam. The outcomes of this study will be valuable for researchers, planners, and policymakers in anticipating future production scenarios, mitigating risks, and enhancing sustainability in the state's agricultural sector.

2 Review of Literature

Box et al. (1970) pioneered the ARIMA methodology, which has since been widely adopted in agricultural forecasting. Prema Borkar(2014)Borkar (2014) in his study spanning 1949–2012 identified ARIMA(1,0,1) as the optimal model for potato yield forecasting, with diagnostic checking confirming model adequacy and forecasting precision. Nath et al. (2019) fitted an ARIMA(1,1,0) model to national wheat production data, projecting a steady upward trend through 2026–27 and validating forecasts via residual diagnostics. Mishra et al. (2021) further applied ARIMA for state-level pulse forecasting, including rice as a comparative crop, and generated reliable short-term forecasts for the period 2020–2029. Delvadiya et al. (2023) in their study modelled area and production using ARIMA, emphasizing the technique's significance for oilseed planning given India's status as the third-largest global producer. Several studies have applied ARIMA models specifically in the context of Assam. In a study conducted by Hazarika (2010), applied the ARIMA methodology to model and forecast tea production in Assam using annual data from 1961 to 2005. The ARIMA(1,1,1) model was identified as the most suitable model. Boruah et al. (2020) utilized ARIMA to forecast rice and wheat production at the national level, indicating high predictive accuracy for rice yields when fitting ARIMA(2,1,1) models to historical data (1961–2020). Hazarika and Phukon (2024) in their study applied ARIMA methodology to forecast sugarcane production in Assam, utilizing 60 years of time series data from 1962–63 to 2021–22. The ARIMA(1,2,1) model was identified as the best fit which forecasted an increasing trend in sugarcane production in the coming years. While several studies have focused on individual crops or regions, comprehensive forecasting studies involving multiple major crops in Assam remain limited. This gap underscores the need for a region-specific, multi-crop forecasting study using ARIMA that can inform both agricultural stakeholders and policymakers. This study addresses that gap by applying ARIMA models to forecast yields of five major crops of Assam viz. Rice, Wheat, Potato, Rapeseed and Mustard(R&M), and Arhar, with a view to support agricultural planning.

3 Objectives

The primary aim of this study is to develop a time series-based forecasting model for the yield of five major crops in Assam. In alignment with this aim, the specific objectives of the study are:

- Trend Detection in Yield of five major crops in Assam.
- Forecasting the Yield of five major crops in Assam using ARIMA model.

4 Methodology

4.1 Data Source

Annual yield data (kg/ha) for five crops from 1973–74 to 2022–23 were obtained from the Directorate of Economics and Statistics (DES), Govt. of Assam. Analysis was conducted in RStudio 2024.04.1.

4.2 Trend Analysis

Trend analysis in time series involves assessing both its magnitude and statistical significance. In this study, Sen’s slope estimator was used to measure strength of the trend, while the Mann–Kendall test was applied to detect the presence of statistically significant trends.

4.3 Mann-Kendall (MK) Test

The Objective of the test is to detect the presence of a monotonic (upward or downward) trend in each time series without assuming any particular distribution of the data. As a non-parametric test for trend detection, MK test was formulated by Mann (1945) and for testing the non-linear trend the test statistic distribution was given by Kendall (1975).

The Mann-Kendall test evaluates the presence of a monotonic trend in time series data using the hypotheses:

- H_0 : Observations are independent and identically distributed (no trend).
- H_a : A monotonic trend exists.

The test statistic S is calculated as:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \text{sgn}(X_j - X_k) \quad (1)$$

where:

$$\text{sgn}(X_j - X_k) = \begin{cases} +1 & \text{if } X_j > X_k, \\ 0 & \text{if } X_j = X_k, \\ -1 & \text{if } X_j < X_k. \end{cases} \quad (2)$$

Under H_0 , $E[S] = 0$. The variance $V[S]$ accounts for ties:

$$V[S] = \frac{n(n-1)(2n+5) - \sum_{j=1}^p t_j(t_j-1)(2t_j+5)}{18} \quad (4)$$

where p is the number of tied groups, and t_j is the size of the j -th group.

For $n > 10$, the standardized statistic Z is:

$$Z = \begin{cases} \frac{S-1}{\sqrt{V(S)}} & S > 0, \\ 0 & S = 0, \\ \frac{S+1}{\sqrt{V(S)}} & S < 0. \end{cases} \quad (5)$$

If $|Z| < Z_{\text{critical}}$ at the chosen significance level, H_0 is retained. The sign of S indicates trend direction.

4.4 Sen's Slope Estimator

Sen's slope was developed by Sen (1968) and it has been widely used to calculate the magnitude of trends in the long-term temporal data (Tabari and Talaei (2011)). The Objective of the test is to quantify the magnitude of any detected monotonic trend. Sen's Slope quantifies trend magnitude robustly. The slope d_k between pairs (i, j) is:

$$d_k = \frac{X_j - X_i}{j - i} \quad (1 \leq i < j \leq n). \quad (6)$$

Sen's slope b is the median of all d_k :

$$b = \text{Median}(d_k).$$

Intercepts a_t for time t are:

$$a_t = X_t - b \cdot t. \quad (7)$$

The final estimator L depends on N (total slopes):

$$L = \begin{cases} d_{(N+1)/2} & \text{odd } N, \\ \frac{1}{2} (d_{N/2} + d_{(N+2)/2}) & \text{even } N. \end{cases} \quad (8)$$

A positive L implies an upward trend; negative L indicates a downward trend.

4.5 ARIMA(p, d, q) Model Overview

The Autoregressive Integrated Moving Average (ARIMA) model, introduced by Box et al. (1970), is a widely used method for time series forecasting. It combines three components:

- **Autoregressive (AR)**: Captures dependency on past values (order p).
- **Integrated (I)**: Stabilizes non-stationary data through differencing (order d).
- **Moving Average (MA)**: Accounts for error term dependencies (order q).

ARIMA is particularly effective for short-term predictions in datasets without seasonal patterns.

4.5.1 Mathematical Formulation

The general ARIMA(p, d, q) model using the backshift operator B is:

$$\phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t \quad (9)$$

where:

- Y_t : Crop yield at time t (target variable)
- B : Backshift operator ($B^k Y_t = Y_{t-k}$)
- $\phi(B)$: AR polynomial of order p : $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$
- $(1 - B)^d$: Differencing operator of order d
- $\theta(B)$: MA polynomial of order q : $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$
- ε_t : White noise error term

4.5.2 Model Building Steps

Step 1: Model Identification

- **Stationarity Check:** Use visual inspection, ACF/PACF plots, and Augmented Dickey-Fuller (ADF) test ($p < 0.05$ indicates stationarity).
- **Differencing:** Apply d -th order differencing if non-stationary.
- **Order Selection:** Determine p and q using ACF (identifies MA order) and PACF (identifies AR order) plots.

Step 2: Parameter Estimation

- Estimate AR(p) and MA(q) coefficients via maximum likelihood estimation.
- Select model with lowest Akaike (AIC) and Bayesian (BIC) information criteria.

Step 3: Diagnostic Checking

- **Residual Analysis:** Ensure residuals resemble white noise (no autocorrelation).
- **Ljung-Box Q Test:** Validate residual independence ($p > 0.05$ preferred).
- Refine model if assumptions are violated.

Step 4: Forecasting

- Generate predictions with 95% confidence intervals.
- Validate forecast accuracy using metrics like RMSE or MAE.

Table 1: Summary statistics of yield of five major crops of Assam

Crop	Min	Max	Mean	SD	MK Stat	Sen's Slope
Rice	949.8	3315.9	1538.7	511.9	8.749***	25.67
Wheat	936.5	1378.7	1185.7	104.6	2.694**	2.86
Potato	3094.0	9331.0	6362.1	1140.7	4.885***	50.03
R&M	365.3	760.2	516.1	72.3	7.009***	3.82
Arhar	671.9	936.9	737.5	64.9	3.178**	1.22

* 5% level of significance, **1% level of significance, ***0.1% level of significance

5 Results and Discussion

5.1 Descriptive Statistics:

Some basic descriptive statistics describing the yield of five major crops of Assam for the last 50 years is presented in the Table 1.

The Mann-Kendall test results indicated statistically significant positive trends for all five crops, as $p < 0.05$. Corresponding Sen's slope estimates quantify the magnitude of these trends, suggesting a consistent and meaningful upward progression over time. The time series plot of yield of the five major crops from 1973-74 to 2022-23 are depicted in fig. 1, which indicate non-seasonality and non-stationarity.

The ADF test was employed to check the stationarity which was found to be non-significant indicating non-stationarity of the time series data. Except for potato, the first differenced time series was found to be stationary with significant Dickey-Fuller value ($p < 0.05$). For Potato second order differencing was done to get significant Dickey-Fuller value. Therefore, degree of differencing $d = 1$ for rice, wheat, R&M, Arhar and $d = 2$ for potato. It is visualised in the ACF and PACF plot as given in fig. 2.

Using goodness-of-fit measures AIC and BIC values, various models were compared in order to find the best model for the data. The model with lowest AIC and BIC values was considered to be the best fitted model as detailed in table 2

Table 2: Best Fitted Model with AIC and BIC values for all the five crops

Crop	Model	AIC	BIC
Rice	ARIMA(0,1,1)	661.2882	665.0718
Wheat	ARIMA(2,1,0)	591.5961	597.2716
Potato	ARIMA(2,2,1)	789.7411	797.2259
R&M	ARIMA(3,1,1)	511.9777	521.4368
Arhar	ARIMA(2,1,2)	445.8138	455.2729

The parameter of the selected model was estimated using maximum likelihood method and detailed in table 3

Table 3: Parameter estimates of selected ARIMA models

Rice			Wheat			Potato		
Variable	Estimate	Std Error	Variable	Estimate	Std Error	Variable	Estimate	Std Error
MA(1)	-0.3453	0.137	AR(1)	-0.779	0.1222	AR(1)	-0.6939	0.1488
			AR(2)	-0.4936	0.1209	AR(2)	-0.3942	0.1514
						MA(1)	-1	0.1008

R&M			Arhar		
Variable	Estimate	Std Error	Variable	Estimate	Std Error
AR(1)	-1.1419	0.1832	AR(1)	0.1106	0.1746
AR(2)	-0.7121	0.2089	AR(2)	-0.5516	0.1339
AR(3)	-0.4877	0.1482	MA(1)	0.0238	0.0888
MA(1)	0.766	0.1598	MA(2)	1	0.1305

From the table 3, the estimated models for all the crops can be written as follows

Rice: $Y_t = Y_{t-1} + \varepsilon_t - 0.3453\varepsilon_{t-1}$

Wheat: $Y_t = 0.2210 Y_{t-1} - 0.4936 Y_{t-2} + \varepsilon_t$

Potato: $Y_t = 1.3061 Y_{t-1} - 0.0064 Y_{t-2} + 0.0945 Y_{t-3} - 0.3942 Y_{t-4} + \varepsilon_t - \varepsilon_{t-1}$

R&M: $Y_t = -0.1419 Y_{t-1} + 0.4298 Y_{t-2} + 0.2244 Y_{t-3} + 0.4877 Y_{t-4} + \varepsilon_t + 0.7660 \varepsilon_{t-1}$

Arhar: $Y_t = 1.1106 Y_{t-1} - 0.6622 Y_{t-2} + 0.5516 Y_{t-3} + \varepsilon_t + 0.0238 \varepsilon_{t-1} + \varepsilon_{t-2}$

The adequacy of the fitted ARIMA model was evaluated by analyzing the residuals' ACF and PACF plots. In this case, the residual plots did not display any discernible trends, suggesting that the model's assumptions were satisfied and it provided a good fit to the observed data. Additionally, there were no significant spikes observed in the ACF and PACF plots of the residual, which indicated no autocorrelation between the residuals. The ACF and PACF plots further showed that the residuals from these models were random and resembled white noise. Histogram plot also suggested residuals were normally distributed. For diagnostic checking, the presence of autocorrelation between the residuals in the fitted ARIMA models was checked by the Ljung-Box Q statistic under null hypothesis of no autocorrelation. The non-significant Ljung-Box Q statistic (p -value for *rice* = 0.8824, for *wheat* = 0.4928, for *potato* = 0.7068, for *R&M* = 0.9416 and for *arhar* = 0.6115) for all the crops, indicated absence of autocorrelation and it followed white noise. This indicated that the models could be effectively utilized for predicting the future yield values. Utilizing the fitted ARIMA models, yield forecast for the five major crops in Assam was done spanning from 2023 to 2030. These forecasts, along with their respective 95% Upper Control Limits (UCL) and Lower Control Limits (LCL) have been detailed in table 4 and visualized in Fig. 3. It depicts the forecasted yield of all the five major crops in Assam from 2023 to 2030, with the blue line representing the point forecast and the shaded region representing 95% confidence interval.

6 Conclusion

This study aimed to forecast the yield of five major crops in Assam—rice, wheat, potato, rapeseed & mustard, and arhar—using ARIMA models based on historical time series data. The results

Table 4: Forecast Results with 95% Confidence Intervals from 2023 to 2030.

Year	Rice			Wheat			Potato		
	95% Confidence Interval			95% Confidence Interval			95% Confidence Interval		
	Forecast	Lower	Upper	Forecast	Lower	Upper	Forecast	Lower	Upper
2023	2687.84	2300.34	3075.34	1326.34	1141.31	1511.37	8011.53	6462.37	9560.70
2024	2687.84	2224.67	3151.01	1283.07	1093.58	1472.57	8226.59	6597.15	9856.03
2025	2687.84	2159.74	3215.95	1309.38	1110.05	1508.72	8788.24	7034.03	10542.45
2026	2687.84	2101.95	3273.73	1310.25	1079.28	1541.21	8504.34	6489.04	10519.65
2027	2687.84	2049.38	3326.31	1296.59	1057.02	1536.16	8670.51	6528.09	10812.93
2028	2687.84	2000.81	3374.87	1306.80	1054.89	1558.71	8857.70	6568.94	11146.46
2029	2687.84	1955.46	3420.22	1305.59	1037.07	1574.10	8822.89	6399.13	11306.66
2030	2687.84	1912.76	3462.92	1301.49	1022.92	1580.06	8973.02	6383.69	11562.35

Year	R&M			Arhar		
	Forecast	Lower limit	Upper limit	Forecast	Lower limit	Upper limit
2023	664.18	585.39	742.98	953.62	914.18	993.06
2024	667.56	574.67	760.44	966.77	907.15	1026.39
2025	684.46	587.76	781.15	958.99	873.27	1044.72
2026	709.60	606.77	812.44	950.88	845.93	1055.83
2027	667.21	550.09	784.33	954.27	837.53	1071.02
2028	689.47	566.61	812.32	959.12	831.86	1086.38
2029	681.98	552.84	811.12	957.79	818.83	1096.75
2030	695.36	560.18	830.53	954.97	804.91	1105.02

revealed that distinct ARIMA specifications were optimal for each crop: ARIMA(0,1,1) for rice, ARIMA(2,1,0) for wheat, ARIMA(2,2,1) for potato, ARIMA(3,1,1) for rapeseed & mustard, and ARIMA(2,1,2) for arhar. The selection of these models was based on diagnostic checks, including AIC and BIC values, residual analysis, and the stability of the models. The forecasts suggest that rice and wheat yields are expected to remain relatively stable through 2030, with point estimates hovering around 2687.84 kg/ha and 1300 kg/ha respectively. Potato, in contrast, is projected to show a strong upward trend, with yields rising from 8011.53 kg/ha in 2023 to over 8973 kg/ha by 2030. Rapeseed & mustard and arhar also exhibit gradual increases in yield, although at a slower pace, indicating moderate but consistent growth. Despite these positive trends, the confidence intervals widen over time for all crops, pointing to increasing uncertainty in long-term projections. Overall, the study demonstrates the effectiveness of ARIMA models in capturing yield dynamics and providing actionable forecasts. The findings are particularly valuable for policymakers, researchers, and agricultural planners in Assam, enabling them to anticipate supply trends, allocate resources efficiently, and develop strategies for enhancing crop productivity. Future research can incorporate exogenous variables such as rainfall, temperature, and input use to further refine forecasts and support climate-resilient agricultural planning

References

- Borkar, P. (2014). Forecasting of potato production in india using arima model. *International Journal of Humanities and Social Sciences*, 4(2):213–220.
- Boruah, B. B., Roy, P., Mahanta, K. K., and Barhoi, K. (2020). Forecasting rice and wheat production of india using arima models. *International Journal of Management*, 11(11):3117–3128.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Prentice-Hall, 3rd edition.
- Delvadiya, J. B., Patel, U. B., Padaliya, M., and Gohil, V. M. (2023). An application of arima for forecasting rapeseed and mustard area in gujarat. *International Journal of Statistics and Applied Mathematics*, 8(5S):524–527.
- Hazarika, J. (2010). Development of arima model for tea production of assam, india: A case study. *International Journal of Agricultural and Statistical Sciences*, 6(1):11–18.
- Hazarika, M. and Phukon, K. K. (2024). Development of arima model for forecasting sugarcane production in assam. *Indian Journal of Agricultural Research*.
- Kendall, M. G. (1975). *Rank Correlation Methods*. Charles Griffin, London, U.K., 4th edition.
- Mann, H. B. (1945). Non-parametric test against trend. *Econometrica*, 13:245–259.
- Mishra, P., Yonar, A., Yonar, H., Kumari, B., Abotaleb, M., Das, S. S., and Patil, S. G. (2021). State of the art in total pulse production in major states of india using arima techniques. *Current Research in Food Science*, 4:800–806.
- Nath, B., Dhakre, D. S., and Bhattacharya, D. (2019). Forecasting wheat production in india: An arima modelling approach. *Journal of Pharmacognosy and Phytochemistry*, 8(1):2158–2165.
- Sen, P. K. (1968). Estimates of the regression coefficient based on kendall’s tau. *Journal of American Statistical Association*, 39:1379–1389.
- Tabari, H. and Talaei, P. H. (2011). Analysis of trends in temperature data in arid and semi-arid regions of iran. *Global and Planetary Change*, 79:1–10.

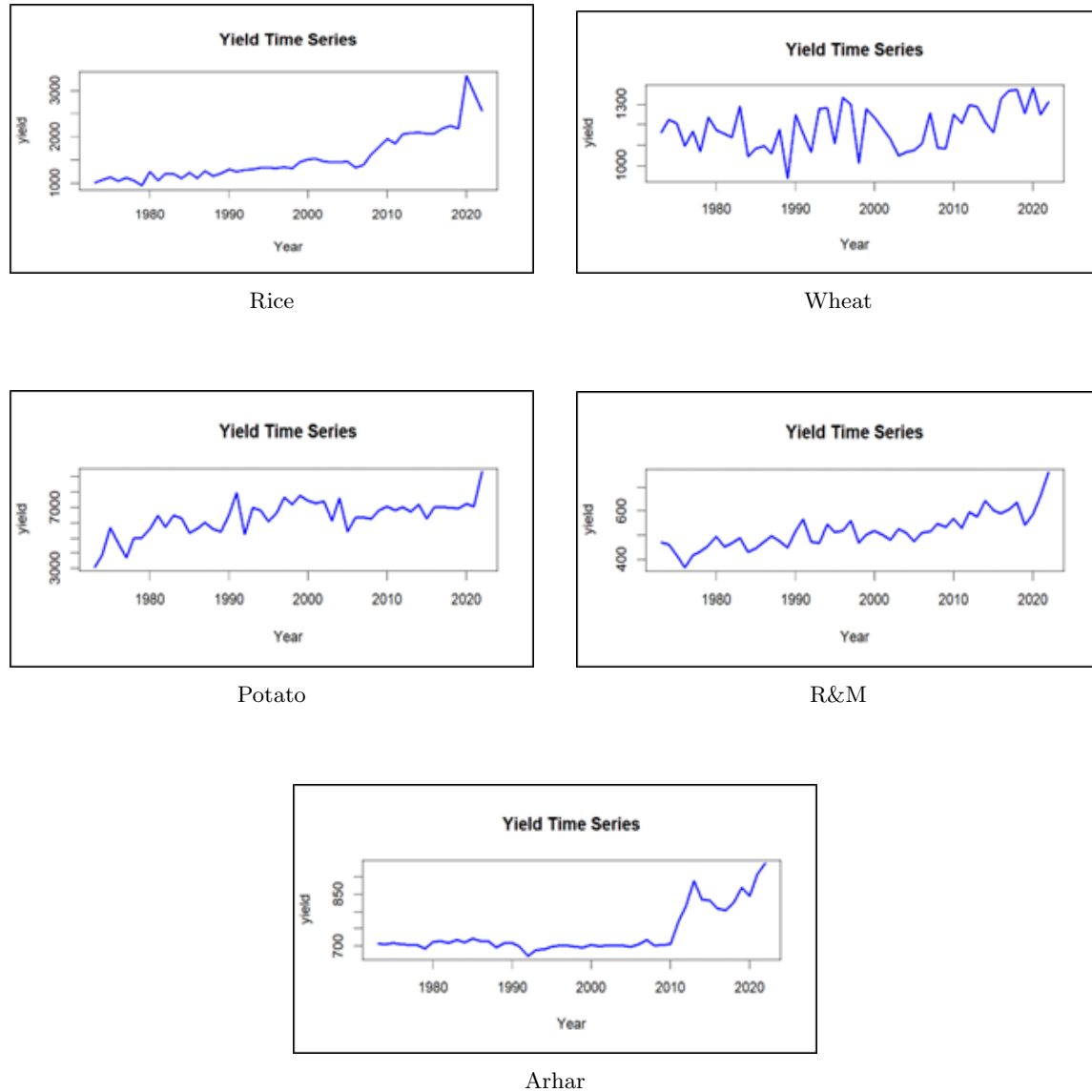


Figure 1: Plot of Yield of five major crops of Assam from 1973–74 to 2022–23

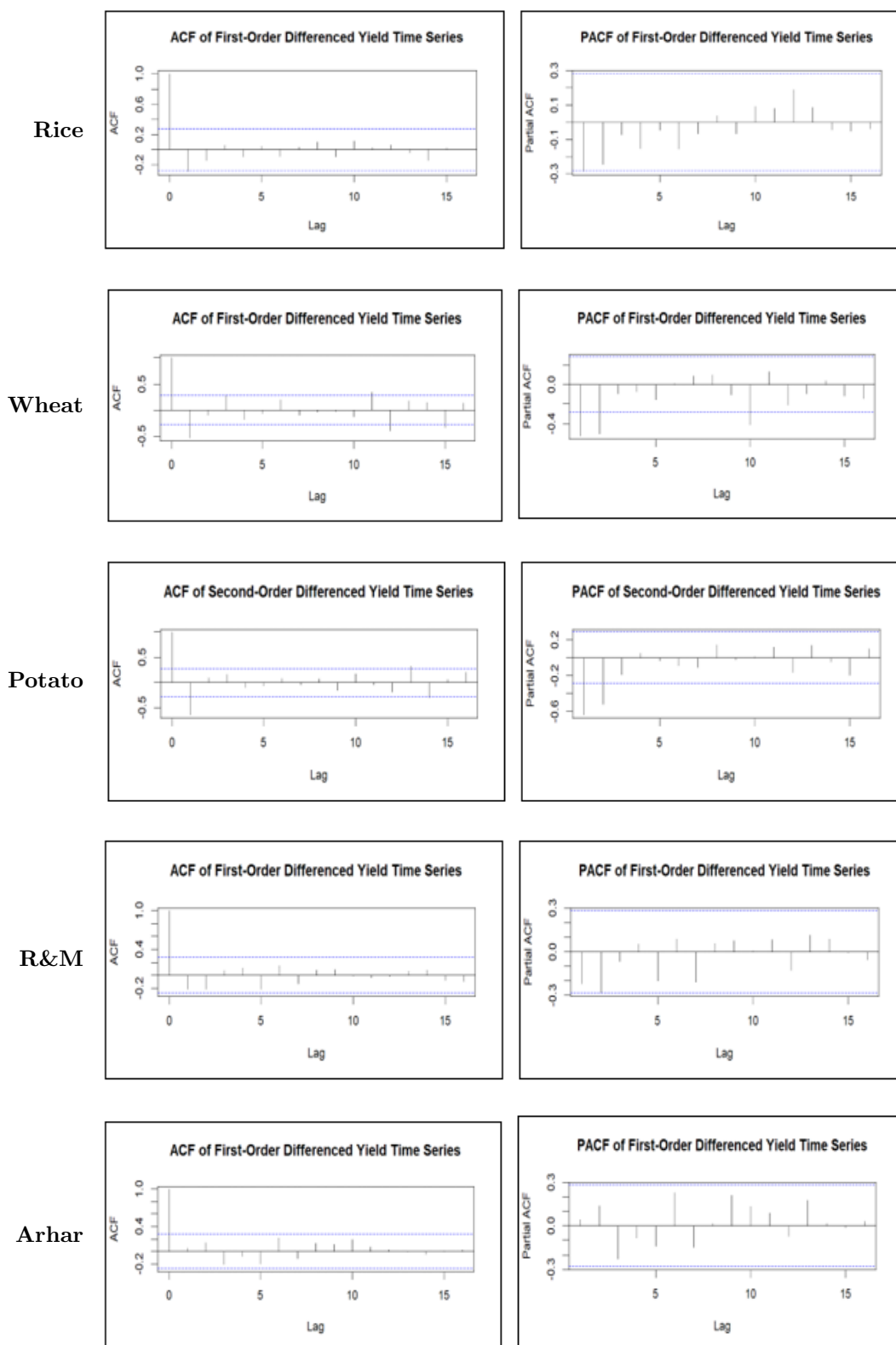


Figure 2: ACF and PACF of the Differenced Data of Yield of five crops

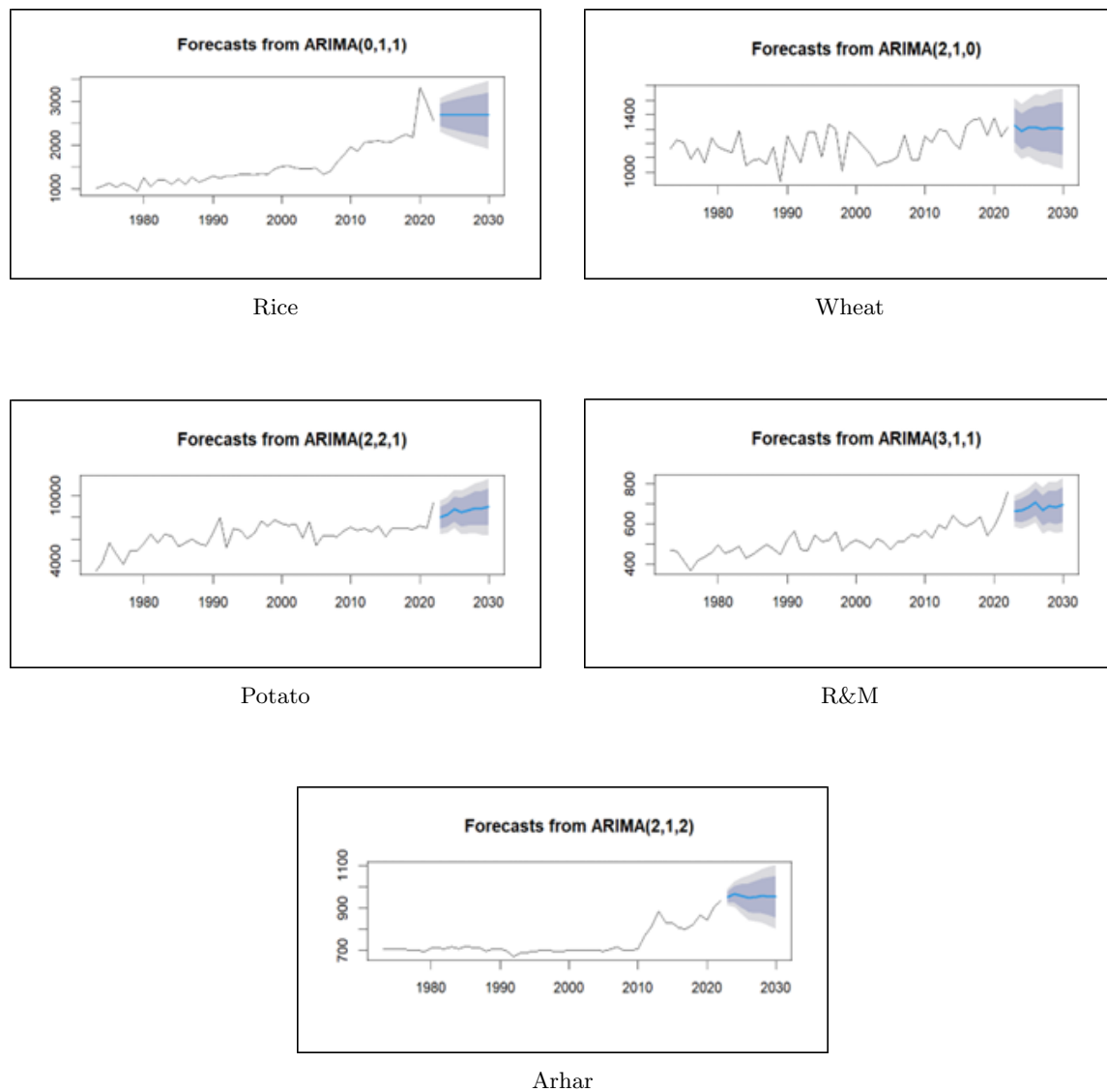


Figure 3: Forecasted Yield of five major crops of Assam for the period 2023–2030