Ann. Agric. Res. New Series Vol. 46 (4) : 366-372 (2025)

Smart harvest: a multicrop, rainfall-integrated model comparison framework for yield prediction

Vidisha Rathi[1], Deeksha Sharma[2], Rayyan Amam Alam[3], Ranjit Kumar Paul[4], Md Yeasin[6], Anil Kumar[5] and Sanjeev Panwar[7]

Corresponding Author's Email: rathiv3777@gmail.com

Received: May 2025; Revised Accepted: August 2025

Abstract

This work introduces Smart Harvest, a predictive analytics model for predicting rice, wheat, jowar, and bajra crop yields. The model investigates the rainfall impact on yield through the use of historical production and climatic data. ARIMA, ARIMAX, SVR, ANN, and Random Forest models are considered and compared for performance and accuracy. The findings indicate the robustness of rainfall-integrated models in enhancing prediction accuracy. This method facilitates data-driven farm planning and risk management.

Key words: ARIMA, ARIMAX, SVR, ANN, machine learning, random forest, RMSE, MAPE

Introduction

Agriculture is the pillar of the Indian economy, providing jobs to over 50% of India's workforce and contributing significantly towards food security and rural prosperity. Crop yield, as the quantity of agricultural produce per unit of land area, is one of the most important key performance indicators in agriculture. Crop yield does not only determine farm-level income but also national food availability, inflation, trade, and policy choice. Yet crop yields are naturally variable, subject to a multifaceted interplay of biological, environmental, and socio-economic variables.

One of the most important environmental determinants of crop yield is rain. In India, where almost 60% of cultivable land is rain-fed, rainfall becomes a major determinant of crop prosperity or failure. The monsoon season, which provides most of the yearly rainfall, is a keystone of the crop cycle. The intensity, frequency, duration, and pattern of rainfall can all have a considerable impact on soil moisture, plant health, and nutrient uptake. Insufficient rainfall can result in drought stress, whereas unseasonable or too much rainfall can trigger flood or pest infestation leading to huge drops in crop yields.

Staple food grains in India, such as rice, wheat, jowar, and bajra, have differential sensitivities to rainfall. Rice, for example, needs standing water and is very much reliant on regular and sufficient rainfall. Wheat grows well in cooler and dryer conditions, where excess moisture is not beneficial. Millets such as jowar and bajra tend to be more drought-resistant but are still impacted by climatic extremes and erratic rainfall patterns. Consequently, knowledge of rainfall variability effects on the yields of these particular crops is essential for guaranteeing stable agricultural production in a changing climate.

Increased uncertainty in rainfall patterns as a result of climate change has further exacerbated the urgency for reliable crop yield forecasting systems. Historical average or manual estimate-based conventional methods of yield estimation are generally found to be inadequate in capturing the dynamic and non-linear dependencies between climatic factors and plant growth. To

improve this, contemporary techniques employ data analytics, machine learning (ML), and time series forecasting models for simulating the intricate dependencies and making more precise and timely estimations.

This research proposes to create a comparative model framework that uses past yield and rainfall data to predict rice, wheat, jowar, and bajra production. The research is about comparing a series of predictive models consisting of statistical models like ARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables) and machine learning algorithms like Support Vector Regression (SVR), Artificial Neural Networks (ANN), and Random Forest. All these models have different features of how they treat data, how they address nonlinearities, and how they deal with external variables like rainfall (Panwar, 2017).

The overall objective of this research is twofold:

1. For analyzing the influence of rainfall over crop yield across various crops.
2. For determining the best predictive model for each crop out of the most important performance indicators like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Through comparison of these models, Smart Harvest intends to develop a trustworthy and scalable yield forecasting instrument that can benefit farmers, researchers, and policymakers. Precise yield forecasting has the capability to facilitate efficient agricultural planning, risk management, and resource allocation, hence enhancing resilience in the context of climatic uncertainty. Overall, the project is a step toward the eventual goal of sustainable and data-informed agriculture in India.

Materials and Methods

Data Description

The research draws on two main datasets

Crop Yield Data: Crop yields of rice, wheat, jowar, and bajra from past years were obtained from the Ministry of Agriculture & Farmers Welfare, Government of India. (2020) and Agricultural Statistics at a Glance. The data comprises yearly production (in tonnes) between 1950 to 2020, which provides a wide time horizon to analyse trends accurately.

(Source: https://agriwelfare.gov.in/en/Agricultural_Statistics_at_a_Glance)

Rainfall Data: Annual rainfall data for the same time frame was collected from the **Indian Meteorological Department (IMD) [2020]**. The figures are provided in millimetres and represent the total annual rainfall in the entire nation.

(Source: https://www.tropmet.res.in)

Datasets were preprocessed to make them consistent in time index (e.g., transforming "1950-51" into a single representative year such as "1950") and merged with respect to the year. Handling missing values was done via interpolation techniques, and data was normalized where necessary for machine learning models.

Stochastic Models

Stochastic models make use of time series properties and are applied extensively in forecasting because they are capable of describing temporal dependencies.

Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA is a model for forecasting time series that captures autocorrelations in past crop yields data. It takes three parameters to capture trends and patterns:

- p (autoregression)
- d (differencing)
- q (moving average)

The model has the assumption of future values as a linear combination of previous values and residuals. It is applicable for univariate data and does not take external factors such as rainfall into account. Parameter tuning was performed by grid search with Akaike Information Criterion (AIC) as the criterion for choosing parameters. The model was separately fit on rice, wheat, jowar, and bajra yield data.

$$Y_t = c + \Sigma (\varphi_i Y_{t-i}) + \Sigma (\theta_j - \varepsilon_{t-i}) + \varepsilon_t$$

where:

Y_t = Actual value at time t

φ_i = Coefficients of the autoregressive series

θ_j = Coefficients of the moving average series
ε_t = Random error (or white noise)
p, d, q = The orders of autoregression, differencing, and moving average respectively
c = Constant term

ARIMAX (ARIMA with Exogenous Variables)

ARIMAX extends ARIMA to incorporate exogenous variables here being annual rainfall as predictors. This model captures the effect of climatic variables on crop production. Similar to ARIMA, it involves p, d, q parameters in addition to the coefficients for the exogenous input. It is especially suitable when crop yield is heavily influenced by external trend like variability in rainfall. The model was tested for every crop using rainfall as input and measured based on AIC and forecasting performance measures.

$$Y_t = c + \Sigma (\varphi_i Y_{t-i}) + \Sigma (\theta_j - \varepsilon_{t-i}) + \beta X_t + \varepsilon_t$$

where:
X_t = Exogenous variable (for example, rainfall during time t)
β = Coefficient for the exogenous input
Keeping the remaining terms as defined in ARIMA model

Machine Learning Models

Some machine learning models were used to identify nonlinear relationships between rainfall and crop yield.

Support Vector Regression (SVR)

SVR is a supervised learning algorithm that seeks out a hyperplane to fit the data best within a tolerance margin (epsilon). SVR performs well in identifying non-linear relationships between features such as rainfall and crop yield. RBF kernel functions were utilized in coping with complicated patterns. The algorithm was trained and optimized via grid search and tested on each crop alone.

$f(x) = wW^{\circ}x + b$
Under constraint:
$|y_i^{\circ} - f(x_i)| \leq d$
Kernel (in case of nonlinear SVR):
$K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2)$
where:
w = Weight vector
b = Bias
ε = Tolerance margin
γ = Kernel coefficient in RBF kernel

Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are machine learning models that are heavily influenced by the neural structure of the brain and are able to learn complex and nonlinear relationships within data. An ANN has an input layer, hidden layers, and an output layer and has its neurons in each of these layers connected by weighted links. Each neuron applies inputs to a weighted sum and feeds the output through an activation function such as ReLU or sigmoid in order to add nonlinearity. Training of the network is performed using a method known as backpropagation, which updates weights depending on the difference between predicted and real outputs. ANNs in crop yield prediction are extremely efficient at capturing the complex relationships between variables such as past yield and rainfall. Because they are flexible, ANNs are able to capture variations in yield patterns between crops and seasons more adequately than conventional statistical models. They are, however, sensitive to large amounts of data, careful hyperparameter tuning, and the use of regularization to prevent overfitting. In this research, ANN was shown to be highly accurate, especially for rain-sensitive crops such as rice and wheat. That it can learn from complicated patterns makes it a very useful tool for data-based agricultural planning.

$$y = f(\Sigma W_j \cdot \sigma(\Sigma V_{ij}, X_i + b_j) + b)$$

where:
x_i = Input variables (rainfall, previous yield, etc.)
V_{ij} = Weights between input and hidden layer
$W_{i'}$ = Weights between hidden and output layer
σ = Activation function (e.g., sigmoid or ReLU)
f = Output function (e.g., linear or SoftMax)
b_j, b = Biases

Random Forest

Random Forest is an ensemble learning algorithm that constructs a number of decision trees and combines their outputs to make more accurate predictions. It is good at dealing with nonlinearities and eliminating overfitting. The model

was trained on rainfall and yield data, with hyper parameters optimized by cross-validation. It made stable forecasts for all crops because of its stability to data noise and variability.

$\hat{y} = (1/T) \Sigma h_t$ ‰)

where:

$\hat{y}$ = Final predicted value

T = Decision tree count

h(x) = Prediction from the t-th tree

Each tree is trained on a random sample (bootstrap) from the data

Model Comparison

To select the most appropriate model for crop yield prediction, all models that were put into practice—ARIMA, ARIMAX, SVR, ANN, and Random Forest—were tested and compared on the basis of both quantitative measures and qualitative factors. All the models were trained on 80% of data and tested on the other 20%, while assessment was done using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The accuracy measures are defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

All the analyses were conducted in Python (v3.x) with the help of libraries such as statsmodels, scikit-learn, Tensor Flow, and Keras for machine learning and time series modelling, and pandas and matplotlib for data handling and plotting.

Results and Discussion

The relative performance of crop yield forecast models—tested on both the training and testing datasets using RMSE, MAE, and MAPE—demonstrate the obvious dominance of machine learning over conventional time-series models in this research. Of all the models used for the four leading crops (Rice, Wheat, Jowar, and Bajra) to forecast yield, the Artificial Neural Network (ANN) provided the overall optimal performance. It recorded the lowest MAPE values on the test set, for example, 3.21% for rice and 4.12% for wheat, indicating significantly its capacity to identify and learn from intricate nonlinear relationships present in agricultural datasets. Its performance being relatively stable across all the crops and metrics is evidence of ANN's capacity to generalize impressively from training to unseen data, hence highly dependent for yield forecasting tasks. The second-best model in performance was the Random Forest model. The model also had excellent predictive abilities with test MAPE values of 4.93% for rice and 9.39% for wheat, and moderately high performance for jowar and bajra. Random Forest, being an ensemble learning technique, takes advantage of the combination of many decision trees, preventing overfitting and enhancing the robustness of the model. It was especially useful for dealing with high-dimensional data and extracting variable importance, which is essential in crop prediction where several climatic and agronomic variables interact. Support Vector Regression (SVR) was third, showing fairly good prediction accuracy. It showed low values of MAPE for rice (8.71%) and wheat (4.24%) but lagged behind for bajra (14.29%) and jowar (8.71%). This indicates that even though SVR is effective for some crop yield patterns, it can fail when dealing with datasets with high variance or nonlinearities outside of its kernel's capacity. However, it was more consistent than the statistical models and can be a good model for certain

		TRAIN			TEST			
MODEL	CROP	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RANK
ARIMA	RICE	132.329	88.718	7.97	182.511	156.127	6.113	5
	WHEAT	131.904	88.718	7.719	1224.61	679.769	12.599	
	JOWAR	89.618	67.291	11.38	135.149	109.005	11.266	
	BAJRA	108.772	80.647	16.374	363.907	336.152	24.487	
ARIMAX	RICE	137.93	97.21	8.94	428.19	392.62	16.95	4
	WHEAT	131.57	89.51	7.73	136.61	113.05	3.91	
	JOWAR	89.51	66.2	11.39	149.78	132.32	14.57	
	BAJRA	110.24	85.74	18.47	378.15	337.45	29.89	
SVR	RICE	97.24	73.18	5.84	82.33	65.34	2.85	3
	WHEAT	101.38	85.25	6.19	176.12	119.24	4.24	
	JOWAR	77.04	57.63	9.11	97.82	76.77	8.71	
	BAJRA	119.04	87.38	17.49	159.07	146.83	14.29	
ANN	RICE	82.12	63.29	4.98	92.1	73.66	3.21	1
	WHEAT	85.49	71.85	5.39	159.61	118.6	4.12	
	JOWAR	59.84	45.96	7.44	109.33	85.79	9.75	
	BAJRA	86.28	54.61	10.47	198.27	155.79	15.07	
RANDOM FOREST	RICE	79.82	62.17	4.93	386.98	358.93	15.44	2
	WHEAT	102.42	82.01	5.88	344.29	285.87	9.39	
	JOWAR	66.35	48.62	7.42	108.22	93.15	10.34	
	BAJRA	103.18	74.19	15.41	339.07	302.29	27.03	

well-behaved data cases. Conversely, ARIMAX, which is a variation of ARIMA that accounts for exogenous variables (rainfall in the present scenario), did better than the simple ARIMA because it could consider external factors. For instance, ARIMAX generated MAPE values of 3.91% for wheat and 16.95% for rice, which are less compared to ARIMA's respective values. But both the models had high errors in crops such as bajra and wheat because of their own linearity and stationarity assumptions, which do not capture the dynamic nature of agricultural data entirely. The ARIMA model ranked the lowest of them. It had very high errors in test data—e.g., wheat MAPE of 12.60% and bajra MAPE of 24.49%—pointing towards its incapability to deal with seasonal effects and external variables. Overall, the assessment firmly deems that machine learning models, particularly ANN and Random Forest, are much more effective for crop yield prediction tasks because they are flexible, capable of nonlinear modelling, and possess higher accuracy. They can efficiently handle the variability and multivariate nature of real-world agricultural data, performing better than conventional statistical models such as ARIMA and ARIMAX. Hence, for the future decision support and predictive models in precision agriculture, the incorporation of superior machine learning models should be

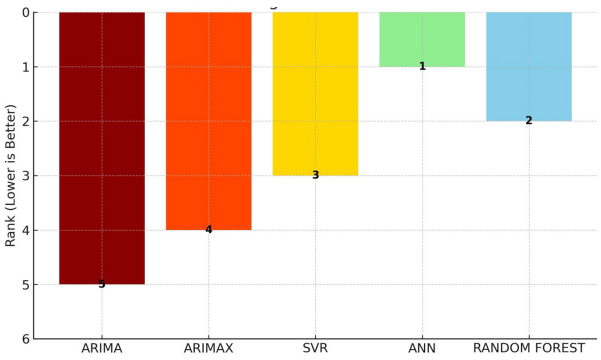

Fig. 1. Model ranking based on performance

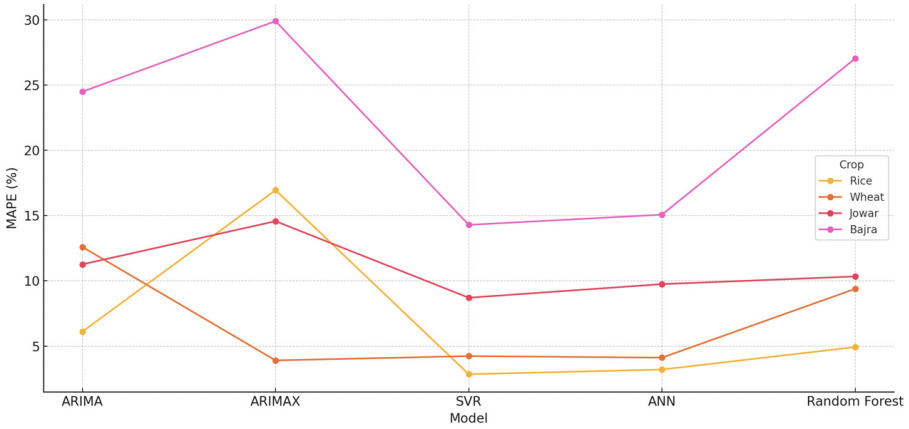

Fig. 2. Model-wise MAPE comparison across crops

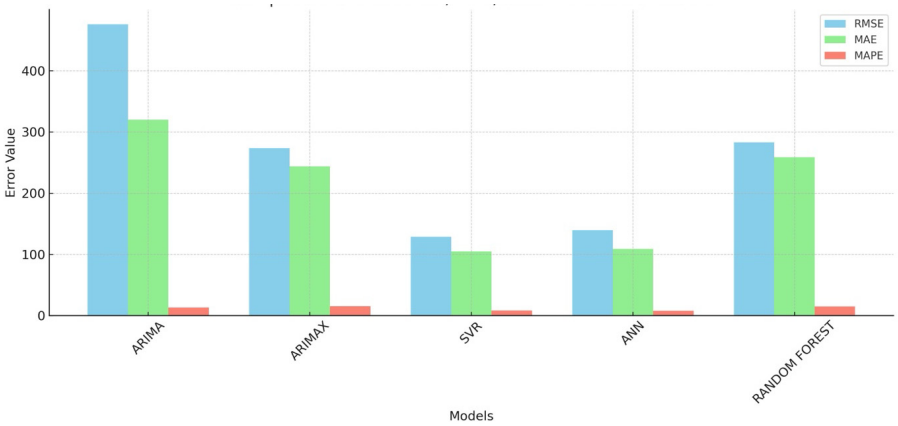

Fig. 3. Comparison of Test RMSE, MAE, and MAPE Across models

given precedence to enable greater reliability and actionable information (Paul, 2024).

Conclusion

The graphical comparison of MAPE values for various models and crops evidently shows that machine learning models, specifically ANN and SVR, are more accurate than conventional time-series models such as ARIMA and ARIMAX. ANN performed the lowest MAPE in three out of four crops, which suggests superior generalization and flexibility to the hidden data patterns. SVR also provided very competitive results, particularly for Rice and Wheat, validating its capability for handling non-linear relationships. Random Forest was the second-best overall and performed well across all crops but at slightly lower accuracy than ANN. ARIMA and ARIMAX, on the other hand, provided much larger MAPE values, particularly for Bajra and Rice, indicating the inability to handle complexity in modelling as well as external factors. Overall, the visual trends confirm that ANN is the most dependable model, followed by SVR and Random Forest, which makes them better fit for precise crop yield prediction in data-intensive and non-linear agricultural systems.

The following bar chart compares the error value of all five models for three different measures i.e. RMSE, MAE, and MAPE. It clearly illustrates that SVR and ANN have the lowest error values, proving greater accuracy. ARIMA performs worst in all measures of error.

The following graph shows the overall ranking of every model with lower values reflecting greater performance. ANN ranked first (Rank 1) and then Random Forest and SVR. The ranking was combined evaluation using RMSE, MAE, and MAPE.

References

Sanjeev Panwar, K.N., Singh, Anil Kumar, Susheel Kumar Sarkar, Ranjeet Paul, Abhishek Rathore and Sivaramane, N. 2014. Forecasting of growth rates of wheat yield of Uttar Pradesh through non-linear growth models. *Indian Journal of Agricultural Sciences*. **84**(7).

Singh, K.N., Singh, K.K., Sudheer Kumar, Sanjeev Panwar and Bishal Gurung, 2019. Forecasting crop yield through weather indices through LASSO. *Indian Journal of Agricultural Sciences*, **89** (3).

Garai, S., Paul, R.K., Yeasin, M. and Paul, A.K. 2024. CEEMDAN-Based Hybrid Machine Learning Models for Time Series Forecasting Using MARS Algorithm and PSO-Optimization. Neural Processing Letters, https://doi.org/10.1007/s11063-024-11552-w

Paul, R.K., Vennila, S., Bhat, M.N., Yadav, S.K., Sharma, V.K., Nisar, S. and Panwar, S. 2019. Prediction of early blight severity in tomato (*Solanum lycopersicum*) by machine learning technique. *Indian Journal of Agricultural Sciences,* **89** (11): 169-175.

Paul, R.K. and Garai, S. 2021. Performance comparison of wavelets-based machine learning technique for forecasting agricultural commodity prices. *Soft Computing*, **25**(20), 12857-12873.

Sanjeev Panwar*, K.N. Singh, Anil Kumar, Bishal Gurung, Susheel Kumar Sarkar, Sivaramane N and Abhishek Rathore, 2017. Pre harvest forecasting of crop yield using non linear regression modelling: A concept. *Indian Journal of Agricultural Sciences*, **87** (5).

Sanjeev Panwar, Anil Kumar, K.N. Singh, Ranjit Kumar Paul, Bishal Gurung, Rajeev Ranjan, N M Alam and Abhishek Rathore, 2018. Forecasting of crop yield using weather parameters– two step nonlinear regression model approach. *Indian Journal of Agricultural Sciences*, **88** (10).

Paul, R.K., Das, T. and Yeasin, M. 2023. Ensemble of Time Series and Machine Learning Model for Forecasting Volatility in Agricultural Prices. National Science Academy Science Letter. https://doi.org/10.1007/s40009-023-01218-x

Garai, S., Paul, R.K., Rakshit, D., Yeasin, M., Emam, W., Tashkandy, Y. and Chesneau, C. 2023. Wavelets in Combination with Stochastic and Machine Learning Models to Predict Agricultural Prices. Mathematics, **11**, 2896.

Garai, S., Paul, R.K., Kumar, M. and Choudhury, A. 2023. Intra-annual National Statistical Accounts Based on Machine Learning Algorithm. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS3202870

Singh, A. K., Sarkar, A., Paul, R.K., Yeasin, M., Roy, H.S., Kumar, P. and Paul, A. K. 2024. Ensemble Machine Learning Techniques for Prediction of Rainfall in India. *Journal of Agricultural Physics*, **24**(1): 1-8

Paul, R. K., Shankar, S. V. and Yeasin, M. 2024. Forecasting area and yield of cereal crops in India: intelligent choices among stochastic, machine learning and deep learning techniques. *Proceedings of the Indian National Science Academy*, 1-7.

India Meteorological Department. 2020. All-India Rainfall Data. Retrieved from https://www.tropmet.res.in

Ministry of Agriculture & Farmers Welfare, Government of India. 2020. https://agriwelfare.gov.in/en/Agricultural_ Statistics_at_a_Glance Indian

Alam, N.M., Mazumdar, N.M., Mitra, S., Ritesh Saha, Pandey, S.K., Sanjeev Panwar and Gouranga Kar 2021. Predictive model for fibre yield estimation of tossa jute (*Corchorus olitorius*) in India. *Indian Journal of Agricultural Sciences*, **91**(6).