Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

District-Level Rice Yield Prediction in Karnataka: Evaluating Machine Learning Models Under Different Validation Strategies

Dhaksha Gowda
Buildemics
Vikram Ramavarapu
vikramr2@illinois.edu

University of Illinois Urbana-Champaign https://orcid.org/0009-0001-8875-7213

Research Article

Keywords: Rice yield, Karnataka, machine learning, Ridge Regression, Random Forest, XGBoost, rainfall, temperature, irrigation, PMFBY, walk-forward validation

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11712090/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

District-Level Rice Yield Prediction in Karnataka: Evaluating Machine Learning Models Under Different Validation Strategies

Dhaksha Gowda[1] and Vikram Ramavarapu[2]

[1]High School Student, Buildemics, Bengaluru, Karnataka, India.
[2]Siebel School of Computing and Data Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA.

Contributing authors: dhaksha.r.gowda@gmail.com; vikramr2@illinois.edu;

Abstract

Rice is an important food crop in Karnataka, and its yield is affected by different factors such as rainfall, temperature and irrigation. Machine learning models can be used to study these factors and predict rice yield, but the accuracy of these models may change depending on how the data is divided for training and testing. This study uses district-level rice yield and climate data from Karnataka for the period 2000–2023, with 636 observations from 27 harmonised districts. Rainfall data were collected from CHIRPS, temperature data from NASA POWER and rice area, production and yield data from the ICAR-IIRR Rice APY Dashboard. Three machine learning models, Ridge Regression, Random Forest and XGBoost, were evaluated using random train-test splitting and chronological walk-forward validation, and compared with two simple baselines. The study also looks at whether irrigation information and a PMFBY policy-era indicator improve prediction results. The models achieved reasonably good scores under random splitting (average $\boldsymbol{R^2}$ of 0.693–0.726), but their $\boldsymbol{R^2}$ was lower under chronological validation (0.643–0.664), which gave a more realistic picture of how the models perform when predicting later years. A simple previous-year yield baseline also had a lower mean absolute error than every model. The additional irrigation experiment showed mixed improvements, while interestingly, the PMFBY policy-era indicator improved prediction performance across all three models in the separate repeated random-split experiment, with $\boldsymbol{R^2}$ improvements of approximately 0.019 to 0.036. A variable importance analysis of XGBoost showed that district identity was the most important input in

every experiment that used the full dataset, followed by rainfall and then temperature, while the irrigation group was the most important input in the irrigation experiment. These results need to be interpreted carefully because of data limitationsf and do not establish a direct effect of the policy. The study more so highlights the importance of choosing suitable validation methods and conditions before concluding that a machine learning model is accurate.

Keywords: Rice yield, Karnataka, machine learning, Ridge Regression, Random Forest, XGBoost, rainfall, temperature, irrigation, PMFBY, walk-forward validation

1 Introduction

Rice is one of the most important food crops in India and is grown in different parts of Karnataka. Rice production depends on several factors, including rainfall, temperature, irrigation facilities and other agricultural conditions, and changes in these factors can affect the yield obtained by farmers. Machine learning is increasingly being used in agriculture to study patterns in data and make predictions. Models such as Ridge Regression, Random Forest and XGBoost can identify relationships between climate conditions and crop yield. However, the accuracy of a model does not always mean that it will perform well when predicting future values.

One important issue in machine learning is the way the data are divided into training and testing sets. A random split can place observations from different years in both sets, which may allow the model to learn patterns from years that come after the year being predicted. In comparison, chronological validation trains the model using earlier years and tests it on later years. This is closer to the actual situation in which future rice yield needs to be predicted using past information.

This study focuses on district-level rice yield prediction in Karnataka. It examines the performance of three machine learning models under different validation methods, and it also studies whether irrigation information and the introduction of the Pradhan Mantri Fasal Bima Yojana (PMFBY) policy period provide additional information for prediction. The main objectives of this study are:

- To examine rice yield prediction using climate conditions and district-level information in Karnataka.
- To compare the performance of Ridge Regression, Random Forest and XGBoost.
- To examine how model performance changes under random splitting and chronological walk-forward validation.
- To study whether irrigation categories provide additional information for rice yield prediction.
- To examine whether adding a PMFBY policy-era indicator changes prediction performance.
- To identify which variables were most important for accurate XGBoost predictions under each of these conditions.

2 Research Gap

De Clercq and Mahdi [1] studied rice yield prediction across 247 Indian districts. They used machine learning models with climate and remote-sensing data, but our study further examines whether the validation method affects the reported prediction accuracy or not. It also compares random splitting with chronological walk-forward validation and further examines whether random splitting gives a different picture of model performance compared with predicting later years.

For the Fofanah [2] Sierra Leone paper, the study uses national-level data to predict rice yield. It applied walk-forward validation and an anti-data-leakage approach, but it had only 25 annual data points available and did not have district-level specifications or details. Our study uses district-level Karnataka data, allowing differences in rice yield and climate conditions between these districts to be examined.

For the third study, the Czech cereal study [3], it compared random, spatial and temporal data-splitting methods. It showed that the choice of validation method is important when evaluating model performance. However, it focused on cereal crops in the Czech Republic, whereas our study examines the validation issue using rice yield data from Karnataka.

Our study specifically focuses on district-level rice yield prediction in Karnataka. It combines rice yield, rainfall and temperature data for 27 harmonised districts covering 2000 to 2023, and it compares three machine learning models (Ridge Regression, Random Forest and XGBoost) using random train-test splitting and chronological walk-forward validation. It also compares the machine learning models with simple baseline methods like district historical mean and previous-year yield, to examine whether the models can predict later years reliably rather than depending only on random-split accuracy.

Irrigation was included because it is an important factor affecting rice cultivation and water availability. Rainfall alone may not explain all the differences in rice yield between the districts, since some districts depend more on irrigation while others depend more on rainfall. Our study examines whether adding irrigation categories (Irrigated, Mixed and Rain-fed) provides any additional information for prediction. Directly matched irrigation information was available for 79 observations, and this analysis is treated as a separate experiment.

PMFBY (Pradhan Mantri Fasal Bima Yojana) was introduced in 2016 [4]. It aims to provide crop insurance and financial protection to farmers against crop losses, and agricultural policies and support systems are also part of the changing agricultural conditions over time. Our study examines whether including a PMFBY policy-era indicator improves rice yield prediction. The indicator separates the years before PMFBY (2000 to 2015) and the years after its introduction (2016 to 2023); it does not represent actual district-level farmer participation, insurance enrolment or claims. The purpose is just to check whether the policy period provides additional predictive information, not to prove that PMFBY directly increases rice yield.

The main research gap is that previous studies have explored machine learning for crop yield prediction and the importance of validation methods, and our study brings these aspects together using district-level rice yield data from Karnataka. It examines whether reported model accuracy is reliable when predicting later years, and it further

investigates whether additional information from irrigation and the PMFBY policy period improves prediction performance. The main focus is to understand whether the machine learning models are genuinely useful for future prediction or whether their accuracy may be misleading because of the validation method used.

3 Methodology

3.1 Data

The study uses district-level rice yield and climate data for Karnataka covering the period from 2000 to 2023. The main data sources are:

- **Rice yield data:** ICAR-IIRR Rice APY Dashboard (Area, Production & Yield Analytics) [5]. This source provides information on rice area, production and yield.
- **Rainfall data:** CHIRPS rainfall data [6] were used to obtain annual and monthly rainfall information.
- **Temperature data:** NASA POWER data [7] were used to obtain temperature information.
- **Irrigation data:** District-wise agricultural reports from the Directorate of Economics and Statistics, Karnataka [8–10], were used for the irrigation analysis.
- **PMFBY information:** The policy introduction period [4] was used to create a separate policy-era indicator.

Fertilizer information was also considered as a possible additional variable. However, consistent district-level fertilizer data covering the full study period could not be obtained, so fertilizer data were not included in the final models.

The original district names and boundaries were checked before combining the datasets, because some districts had to be harmonised so that the data from different sources could be matched correctly. The district names were standardised, and the relevant district-level observations were combined: the records for Ramanagara were combined with Bengaluru Rural, Chikkaballapura with Kolar, Yadgir with Kalaburagi, and Vijayanagara with Ballari, and yield was recalculated as production divided by area. The rainfall and temperature data were then combined with the rice yield data using district and year as the common variables.

The final dataset contains 636 observations from 27 harmonised districts for the period 2000–2023. A complete dataset for 27 districts over 24 years would contain 648 observations, but 12 observations were not available in the source data: 11 records for Chikkamagaluru from 2000 to 2010 and one record for Bagalkot in 2017. These missing observations were not filled with estimated values. The final dataset was also checked for duplicate district-year records and for missing values in the main yield and climate variables, and none were found.

3.2 Variables and Preprocessing

The target variable is rice yield per hectare (kg/ha), which is the value the models are expected to predict. The climate variables used as predictors are annual, May, June and May–June rainfall, and annual, May, June and May–June temperature. District

information was also included to account for differences between districts, and the district variable was converted into numerical form using one-hot encoding.

Area and production were retained in the dataset for reference but were not used as input variables, because rice yield is calculated from them; including them could allow the models to indirectly access the information used to calculate the target variable, leading to target leakage and misleadingly high accuracy. The year variable was used to arrange the observations in chronological order for walk-forward validation, but it was not included as a predictor. Preprocessing steps such as imputation, one-hot encoding and scaling were fitted using the training data only, so that information from the test data was not used during training.

3.3 Machine Learning Models

Three machine learning models were used in this study. They were selected because they use different approaches, which allows their prediction performance to be compared. The same model settings were used in every experiment (Section 3.9).

Ridge Regression

Ridge Regression is a type of linear regression that reduces the effect of very large coefficients. It adds a penalty to the model to prevent it from depending too much on individual variables, which is useful when several input variables may be related to each other.

Random Forest

Random Forest uses multiple decision trees to make predictions. Each tree produces a prediction, and the model combines these predictions to obtain the final result. It can identify non-linear relationships between climate variables and rice yield.

XGBoost

XGBoost is a boosting-based machine learning model. It builds decision trees one after another, with each new tree trying to improve on the errors made by the earlier trees, so it can capture more complex patterns in the data.

3.4 Validation Strategies and Baselines

The first experiment used a random 80:20 train-test split, in which 80% of the observations (508) were used to train the models and the remaining 20% (128) were used to test them. The experiment was repeated using 30 different random seeds, and the average performance across these runs was calculated. The main purpose of this experiment was to understand how the models perform when the observations are randomly divided.

The second experiment used expanding-window walk-forward validation. In this method, the model is trained using earlier years and tested on a later year; after each test year, the training period is expanded to include more past observations before the next year is predicted. This method is closer to the actual use of a rice yield prediction model, because future yield would normally be predicted using information available

from previous years. Fourteen expanding-window folds were used, with test years covering 2010 to 2023, so the training data grew from 260 observations (2000–2009) to 609 observations (2000–2022). The pooled results contained 376 test predictions for each model.

The machine learning models were also compared with two simple baseline methods on the same 376 walk-forward test observations. The district historical mean predicts each district's yield as the average of its yield in all earlier years, using the overall training mean if a district has no earlier records. The previous-year yield method predicts each district's yield as its yield in the year before; if that record is missing, the district historical mean is used instead.

3.5 Model Evaluation

The models were evaluated using three measures, because looking at only one performance measure may not give a complete picture of a model's performance:

- **R-squared (R^2)** measures how much of the variation in rice yield is explained by the model. A higher value generally indicates better performance.
- **Mean Absolute Error (MAE)** measures the average absolute difference between actual and predicted yield. A lower value indicates better predictions.
- **Root Mean Squared Error (RMSE)** measures prediction errors while giving more weight to larger errors. A lower value indicates better performance.

For the walk-forward validation and the baselines, the measures were calculated on the 376 pooled test predictions, and R^2 was also calculated separately for each of the 14 test years.

3.6 Irrigation Analysis

Irrigation can affect rice yield because it influences the availability of water for crop growth, so irrigation information was included as a separate experiment. District-wise irrigation information was collected from the available agricultural reports [8–10]. The irrigated share was calculated as the total irrigated rice area divided by the total rice area across the reported seasons, and the observations were classified into three groups (Table 1).

Table 1 Irrigation group thresholds

Irrigated Share	Irrigation Group
Below 30%	Rain-fed
30% to below 70%	Mixed
70% or above	Irrigated

Directly matched irrigation information was available for 79 district-year observations from 2009, 2011 and 2017, and these observations were used for the irrigation

experiment. The remaining 557 observations did not have directly matched irrigation information, and they were not treated as if they had observed irrigation values. Two experiments were conducted. The first compared models with and without irrigation information using 30 paired random 80:20 train-test splits of the 79 observations (63 for training and 16 for testing), with the same split used for both conditions within each run. The second trained the models using the 53 observations from 2009 and 2011 and tested them on the 26 observations from 2017.

3.7 PMFBY Analysis

The Pradhan Mantri Fasal Bima Yojana was introduced in 2016 [4]. To examine whether the policy period was associated with changes in prediction performance, a separate indicator was created, with a value of 0 for the period 2000–2015 and 1 for the period 2016–2023. This variable represents the policy period only; it does not show whether a particular district or farmer participated in PMFBY, or the amount of insurance coverage received.

A separate experiment compared models without the PMFBY indicator with models that included it. The experiment used the same 30 random 80:20 train-test splits as the first experiment, with the same split used for both conditions within each run, so each run used 508 training observations and 128 test observations. The base predictors were district identity and the eight climate variables, and the PMFBY condition added only the policy-era indicator. Year, irrigation group, area and production were not included as predictors; irrigation was left out because directly matched irrigation information was available for only a small part of the complete dataset. This was a separate experiment from the main chronological walk-forward validation.

3.8 Variable Importance Analysis

To understand which variables the XGBoost model depended on for accurate predictions, a variable importance analysis was carried out under each of the conditions examined in this study: the 30 random splits, the 14 walk-forward test years, the 30 PMFBY splits with the indicator included, and the irrigation experiments with the irrigation group included. For every split, the model was trained on the training observations, and permutation importance was then calculated on the test observations. In this method, the values of one variable are randomly shuffled among the test observations while all other variables are kept the same, and the increase in MAE is recorded. A large increase means that the model relied on that variable to make accurate predictions, while an increase close to zero means that the variable made little difference. Each variable was shuffled 20 times and the increases were averaged. District identity was shuffled as one variable, so all of its one-hot encoded columns were treated together.

Several climate variables are strongly related to each other. For example, the correlation between annual and June rainfall is 0.93, and between June and May–June rainfall is 0.97, while the four temperature variables have correlations of 0.85 to 0.98 with each other. When one of these variables is shuffled, the model can still use the

related variables, so its importance can be understated. Therefore, permutation importance was also calculated for groups of variables, in which all four rainfall variables, or all four temperature variables, were shuffled together. As a second check, the mean absolute SHAP (SHapley Additive exPlanations) value of each variable was calculated using the TreeSHAP method built into XGBoost. This measures how much each variable changed the model's predictions, in kg/ha, rather than how much it affected accuracy.

3.9 Implementation and Model Settings

All experiments were run in Python 3.12. The models, preprocessing steps, data splits and evaluation measures used scikit-learn 1.9.1 [11], except for the XGBoost model, which used the XGBoost 3.4.1 library [12]. Data handling used pandas 3.1.1 [13] and NumPy 2.5.4 [14], and the figures were drawn with Matplotlib 3.11.2 [15]. The settings used in every experiment are listed in Table 2; any parameter not listed was left at the library default. For each training set (each random split, each walk-forward fold and each irrigation split), the preprocessing steps and the model were fitted together in a single scikit-learn pipeline, so the median values, scaling and district categories were learned from the training observations only. The data and code used for all experiments are available at https://github.com/dhakshargowda7-web/karnataka-rice-yield-research.

Table 2 Software components and settings used in all experiments

Step	Implementation	Settings
Climate predictors	`SimpleImputer`, `StandardScaler`	Median imputation; standardised to zero mean and unit variance
District (and irrigation group)	`SimpleImputer`, `OneHotEncoder`	Most-frequent imputation; one-hot encoding, categories not seen in training ignored
Ridge Regression	`Ridge`	$\alpha = 1.0$
Random Forest	`RandomForestRegressor`	300 trees; `random_state` = 42
XGBoost	`XGBRegressor`	300 trees; maximum depth 3; learning rate 0.05; squared-error objective; `random_state` = 42
Random splits	`train_test_split`	Test size 0.20; random seeds 0 to 29
Evaluation	`r2_score`, `mean_absolute_error`, `mean_squared_error`	RMSE is the square root of the mean squared error
Variable importance	`permutation_importance`; XGBoost `pred_contribs`	Increase in MAE; 20 shuffles per variable; TreeSHAP values for the SHAP check

4 Results

4.1 Results of Random Train-Test Split

The random 80:20 train-test experiment was repeated using 30 different random seeds, and the distribution of results across the runs is shown in Fig. 1. Ridge Regression achieved the highest average R^2 (0.726) and the lowest average MAE (296.571 kg/ha) and RMSE (400.171 kg/ha) among the three models. Random Forest (R^2 0.699, MAE 311.699 kg/ha, RMSE 419.361 kg/ha) and XGBoost (R^2 0.693, MAE 315.570 kg/ha, RMSE 423.545 kg/ha) also showed reasonably good performance, with slightly higher prediction errors than Ridge Regression in this experiment. These results show that all three models were able to identify patterns in the dataset under random splitting. The main takeaway is that even though the random split identifies patterns better than the walk-forward, it does not show how well the models would predict rice yield in future years as accuracy is inflated from being able to see into the future.

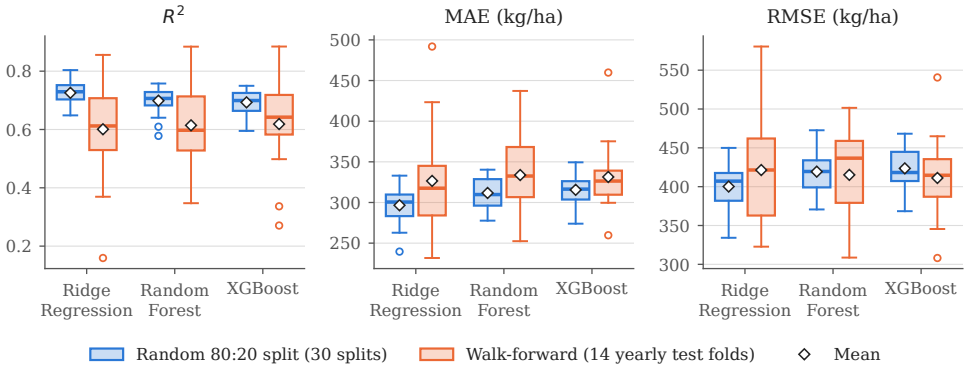


Fig. 1 Distribution of R^2, MAE and RMSE for each model under the two validation strategies: 30 random 80:20 train-test splits (blue) and the 14 yearly test folds of the walk-forward validation, 2010–2023 (orange). Boxes show the interquartile range, the line inside each box is the median and diamonds mark the mean; whiskers extend to 1.5 times the interquartile range and circles are values beyond it.

4.2 Results of Chronological Walk-Forward Validation

The models were also tested using expanding-window walk-forward validation, in which earlier years were used for training and later years for testing, and the pooled results from the 376 test predictions are shown in Table 3. XGBoost achieved the highest pooled R^2 value of 0.664, Random Forest had a similar value of 0.655, and Ridge Regression had a value of 0.643. Ridge Regression had the lowest MAE, while XGBoost had the lowest RMSE. The average R^2 across the 14 individual yearly test folds was 0.601 for Ridge Regression, 0.614 for Random Forest and 0.619 for XGBoost.

Table 3 Pooled model performance under chronological walk-forward validation, 376 test predictions (MAE and RMSE in kg/ha)

Model	R^2	MAE	RMSE
Ridge Regression	0.643	326.379	426.798
Random Forest	0.655	333.541	419.664
XGBoost	0.664	331.250	414.614

All three models had a lower R^2 and a higher MAE under walk-forward validation than their averages under random splitting. For example, the R^2 of Ridge Regression fell from 0.726 to 0.643, Random Forest from 0.699 to 0.655 and XGBoost from 0.693 to 0.664. The change in RMSE was not the same for every model: it increased for Ridge Regression (400.171 to 426.798 kg/ha), stayed almost the same for Random Forest (419.361 to 419.664 kg/ha) and decreased for XGBoost (423.545 to 414.614 kg/ha). The results show that the models were able to predict later-year rice yield to some extent, but their performance was not the same across all test years. The yearly R^2 ranged from 0.159 to 0.856 for Ridge Regression and from about 0.27 to 0.88 for the tree-based models, with 2013 being the most difficult year for Ridge Regression and XGBoost, as the wider spread of the walk-forward folds in Fig. 1 shows.

4.3 Comparison with Baseline Methods

The machine learning models were compared with the district historical mean and previous-year yield baselines using the same 376 test observations (Table 4). All three machine learning models performed better than the district historical mean on all three measures, and they had a higher R^2 and a lower RMSE than the previous-year yield method. However, none of them achieved a lower MAE than the previous-year yield method (312.4 kg/ha). This shows that simple methods can also provide useful predictions.

Table 4 Performance of baseline methods on the 376 walk-forward test observations (MAE and RMSE in kg/ha)

Baseline Method	R^2	MAE	RMSE
District Historical Mean	0.622	342.4	439.2
Previous-Year Yield	0.614	312.4	443.9

4.4 Irrigation Analysis Results

Directly matched irrigation information was available for 79 observations, distributed across the irrigation groups as shown in Table 5. The first experiment compared model

performance with and without the irrigation group variable, using the same random train-test split for both conditions in each of the 30 runs. Fig. 2 shows the distribution of the change in performance, calculated as the result with irrigation minus the result without irrigation; for MAE and RMSE, a negative change means that the prediction error decreased.

Table 5 Number of observations with directly matched irrigation information, by irrigation group

Irrigation Group	Number of Observations
Irrigated	51
Mixed	12
Rain-fed	16
Total	79

On average, adding irrigation information did not improve every model in the same way. Ridge Regression had a small decrease in R^2 (-0.0244), but its MAE and RMSE improved (-2.00 and -4.59 kg/ha). Random Forest showed improvement across all three measures (R^2 $+0.0236$, MAE -7.16 kg/ha, RMSE -19.65 kg/ha). XGBoost showed improvement in R^2 ($+0.0242$) and RMSE (-10.64 kg/ha), but its MAE increased ($+2.41$ kg/ha). When the results were examined across the 30 runs, Random Forest showed an improvement in R^2 in 21 out of 30 runs, in MAE in 20 runs and in RMSE in 21 runs, with all three measures improving together in 20 runs. This suggests that irrigation information may provide additional predictive information, particularly for Random Forest. However, the results are mixed, the changes vary widely from run to run (Fig. 2), and the number of observations is small.

A second experiment used the 2009 and 2011 observations for training and the 2017 observations for testing (Table 6). Ridge Regression and Random Forest showed improvements in both R^2 and MAE after irrigation information was added, while XGBoost showed an improvement in R^2 but a slight increase in MAE. All of the R^2 values in this experiment were low (0.09 to 0.35). Overall, the irrigation experiments suggest that irrigation information may help improve rice yield prediction. However, these findings should be treated carefully because the available data cover only three reporting years and 79 district-year observations.

4.5 PMFBY Analysis Results

A separate experiment was conducted to examine whether adding the PMFBY policy-era indicator changed model performance, and the results across the 30 paired runs are shown in Fig. 3. All three models had higher R^2 values and lower MAE and RMSE values after the PMFBY indicator was added. For Ridge Regression, R^2 increased from 0.726 to 0.744, MAE decreased from 296.571 to 287.079 kg/ha and RMSE decreased from 400.171 to 386.470 kg/ha. For Random Forest, R^2 increased from 0.699 to 0.727,

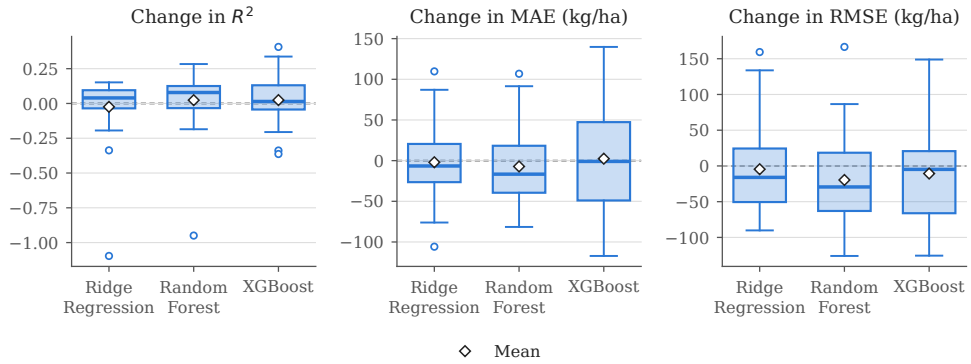


Fig. 2 Change in R^2, MAE and RMSE after adding the irrigation group variable (with irrigation minus without irrigation) across 30 paired random 80:20 splits of the 79 observations with directly matched irrigation information. The dashed line marks no change; for MAE and RMSE, values below it mean lower prediction error. Box elements are as in Fig. 1.

Table 6 Performance with and without irrigation information when training on 2009 and 2011 and testing on 2017 (MAE in kg/ha)

	R^2		MAE	
Model	Without	With	Without	With
Ridge Regression	0.2524	0.3493	513.77	445.51
Random Forest	0.1420	0.2994	537.03	504.63
XGBoost	0.0881	0.2666	517.58	523.69

MAE decreased from 311.699 to 293.954 kg/ha and RMSE decreased from 419.361 to 399.533 kg/ha. For XGBoost, R^2 increased from 0.693 to 0.729, MAE decreased from 315.570 to 299.180 kg/ha and RMSE decreased from 423.545 to 397.673 kg/ha. The improvement was also consistent across runs: Ridge Regression improved in all 30 paired runs for R^2 and RMSE and in 29 of the 30 runs for MAE, while Random Forest and XGBoost improved in all 30 paired runs across all three evaluation measures.

This suggests that the policy-era indicator contains information that is associated with rice yield patterns in the dataset. However, it does not prove that PMFBY directly caused an improvement in rice yield or model performance. The indicator only separates the period before and after 2016, and other changes in agriculture, climate, government schemes and farming practices may also have occurred during this period. This is an interesting finding, however, as the era indicator alone is associated with improved model performance. While it is possible that PMFBY may have contributed to this improvement due to factors such as improved access to agricultural insurance

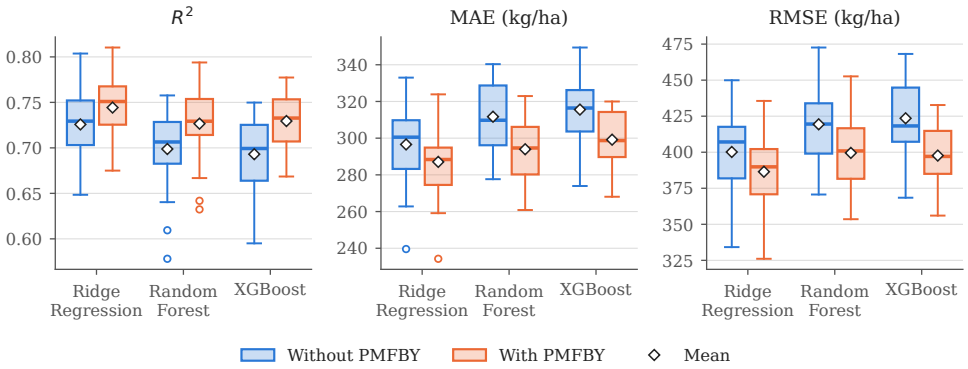


Fig. 3 Distribution of R^2, MAE and RMSE without (blue) and with (orange) the PMFBY policy-era indicator across 30 paired random 80:20 splits (508 training and 128 test observations per split). Box elements are as in Fig. 1.

and intervention, there isn't enough evidence for the observed association to be interpreted as evidence of a direct policy effect. Therefore, the PMFBY result should be understood as a predictive association rather than evidence of a direct policy impact.

4.6 Variable Importance

The variable importance of the XGBoost model under each condition is shown in Fig. 4, and the results for groups of variables are summarised in Table 7. In every experiment that used the full dataset, district identity was by far the most important variable. Shuffling the district increased the MAE by 331.7 kg/ha under random splitting, by 300.8 kg/ha under walk-forward validation and by 343.9 kg/ha in the PMFBY experiment. This shows that much of the model's accuracy came from learning the typical yield level of each district, which is similar to the information used by the district historical mean baseline.

Rainfall was the second most important group of variables in all of these experiments, and temperature was third. Among the individual variables, June rainfall (63.0 kg/ha), annual rainfall (54.5 kg/ha) and May rainfall (42.7 kg/ha) were the most important climate variables under random splitting. Under walk-forward validation, annual rainfall (60.7 kg/ha) and June rainfall (52.4 kg/ha) remained important, but the importance of May rainfall fell to 7.4 kg/ha, which suggests that its relationship with yield did not carry over as well to later years. The importance of the rainfall group was also lower under walk-forward validation (63.7 kg/ha) than under random splitting (84.2 kg/ha), while temperature stayed at a similar, lower level (29.0 to 30.3 kg/ha). Each individual temperature variable increased the MAE by less than 15 kg/ha in all three experiments, and May–June rainfall had almost no importance on its own, because it is very closely related to June rainfall. In the PMFBY experiment, the policy-era indicator increased the MAE by 21.5 kg/ha when shuffled, which placed it after district, rainfall and temperature as a group. This is consistent with the small but consistent improvement found when the indicator was added.

Table 7 Grouped permutation importance of XGBoost under each condition: average increase in MAE (kg/ha) when all variables in a group are shuffled together. A dash means that the variable was not included in that experiment.

	Random	Walk-	With	With irrigation group	
Variable group	split	forward	PMFBY	Random splits	Test 2017
District	331.7	300.8	343.9	50.1	21.4
Rainfall (4 variables)	84.2	63.7	75.2	54.3	87.9
Temperature (4 variables)	29.0	30.3	26.1	6.7	−7.5
PMFBY indicator	–	–	21.5	–	–
Irrigation group	–	–	–	147.2	217.1

The results were different in the irrigation experiments, which used only the 79 observations with directly matched irrigation information. Here, the irrigation group was the most important variable, increasing the MAE by 147.2 kg/ha in the random splits and by 217.1 kg/ha when the 2017 observations were predicted, followed by rainfall. District identity was much less important than in the full dataset (50.1 and 21.4 kg/ha), because each district has at most three observations in this subset, so the irrigation group carries much of the information that would otherwise come from the district. However, the importance values in the irrigation random splits varied widely from split to split (Fig. 4), because each test set contained only 16 observations.

The SHAP values gave the same order of importance for the variable groups in all three full-dataset experiments: district first, rainfall second and temperature third. The SHAP values of the individual temperature variables were larger than their permutation importance, which means that the model did use temperature in its predictions, but shuffling it did not reduce accuracy very much. This is likely partly because 82–89% of the variation in each temperature variable is between districts rather than between years, so much of its information is already provided by district identity.

5 Discussion

The main purpose of this study was to examine whether machine learning models can predict district-level rice yield in Karnataka and whether their reported accuracy is reliable. The random split experiment showed reasonably good model performance, with Ridge Regression achieving the best results. However, random splitting does not fully represent the situation in which a model is trained on past data and used to predict future years, because observations from different years can appear in both the training and testing sets. Chronological walk-forward validation provides a more realistic test because the model only uses earlier observations to predict later years. Under this validation, every model had a lower R^2 and a higher MAE than under random splitting, and the best model also changed: XGBoost had the highest pooled R^2 and lowest RMSE, while Ridge Regression had the lowest MAE. This shows why the validation method is important when evaluating agricultural prediction models.

The comparison with baseline methods also provides an important finding. Although the machine learning models performed better than the district historical

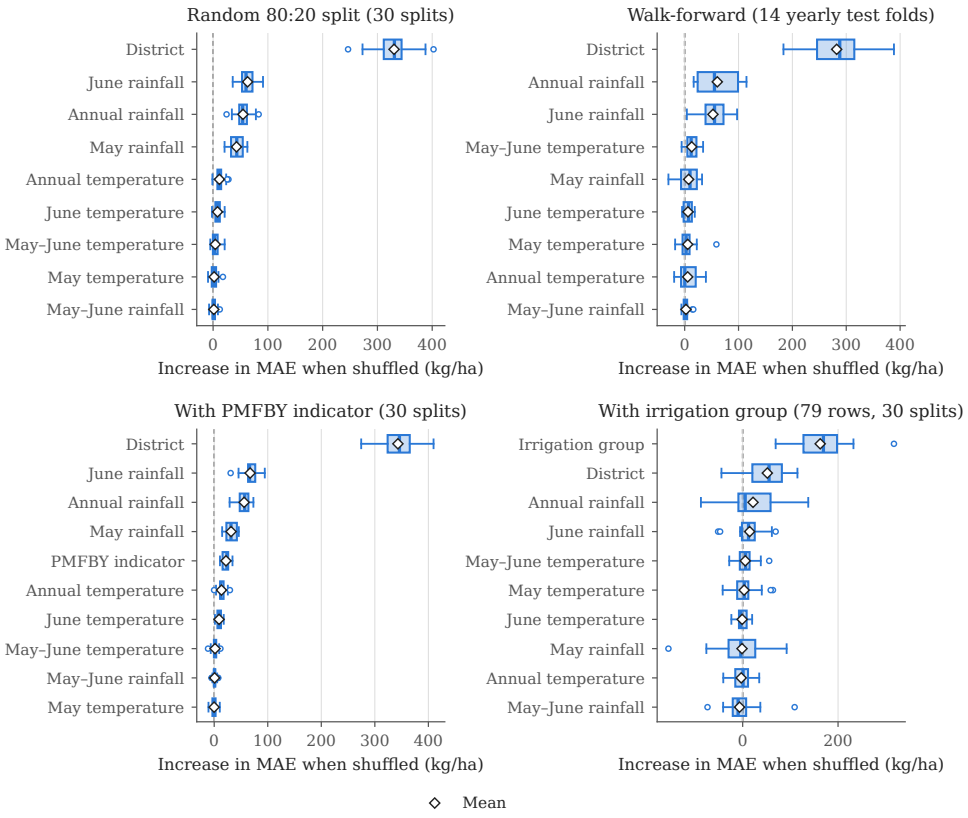


Fig. 4 Permutation importance of each variable for XGBoost under four conditions, measured as the increase in MAE when the variable is shuffled among the test observations. Each box summarises the 30 random splits or the 14 walk-forward test years; a value close to zero (dashed line) means the variable made little difference to accuracy. Box elements are as in Fig. 1.

mean on all three measures, the previous-year yield method had a lower MAE than every model. This means that more complex models do not automatically provide better predictions than simpler methods. Overall, the findings show that model accuracy should be examined using different validation methods and compared with simple baseline predictions, and a high score under random splitting should not be treated as sufficient evidence that a model can accurately predict future rice yield.

The irrigation analysis suggests that irrigation information can provide useful additional information, but the improvement was not consistent across all models and evaluation measures, and the small number of observations and limited reporting years make it difficult to draw strong conclusions. The PMFBY experiment showed that the policy-era indicator improved the reported performance of all three models. However, the indicator only represents whether an observation belongs to the period before or

after 2016 and does not measure actual PMFBY participation or insurance coverage, so the improvement may also reflect other changes that occurred after 2016.

The variable importance analysis helps to explain these results. In all of the full-dataset experiments, XGBoost relied mostly on district identity, followed by rainfall, with temperature contributing much less. This means that the models mainly learned the typical yield level of each district and then adjusted it using rainfall, which is also why the district historical mean and previous-year yield baselines performed close to the machine learning models. The rainfall variables were less important under walk-forward validation than under random splitting, especially May rainfall, which suggests that some of the climate–yield relationships learned from random splits did not hold as well for later years.

6 Limitations of the Study

There are some limitations in this study:

- The dataset contains 636 observations instead of the 648 observations expected in a complete panel, because twelve district-year records were not available in the source data.
- Irrigation information was available for only 79 district-year observations from three reporting years (2009, 2011 and 2017), which limits the amount of information available for the irrigation analysis.
- The PMFBY variable is only a policy-era indicator. It does not contain district-level information about farmer participation, insurance enrolment, claims or settlement amounts.
- Fertilizer data were not included because consistent district-level information covering the full study period could not be obtained.
- The study uses rainfall and temperature as the main climate predictors. Other factors such as soil quality, crop variety, farming practices, pest attacks and access to agricultural inputs may also affect rice yield.
- The variable importance results show which variables the XGBoost model relied on, not whether those variables cause changes in rice yield. Because several climate variables are strongly related to each other, the importance of the individual variables should be interpreted together with the grouped results.
- The same model settings were used in every experiment, and they were not tuned. Other settings could change the results of individual models.
- The results are based on the available historical data and may not represent all future agricultural conditions.
- The irrigation and PMFBY experiments are separate additional analyses, and their results should not be directly compared with the main chronological model results as if they were the same experiment.

7 Conclusion

This study examined district-level rice yield prediction in Karnataka using rainfall, temperature and district information, testing Ridge Regression, Random Forest and

XGBoost with random train-test splitting and chronological walk-forward validation. The random split experiment produced reasonably good results, but chronological validation provided a more realistic way of testing how the models perform when predicting later years, and it gave lower R^2 values for every model. The comparison with baseline methods also showed that machine learning models do not always perform better than simpler prediction methods.

The additional irrigation analysis showed that irrigation information may improve prediction performance, particularly for Random Forest, although the results were not consistent across all models. The PMFBY experiment also showed improvements after adding a policy-era indicator, but this should not be considered proof that PMFBY directly caused changes in rice yield. The variable importance analysis showed that XGBoost relied mostly on district identity and rainfall, with temperature contributing much less. The main finding of this study is that model accuracy should not be judged using only one validation method or one performance measure. It is important to test models using chronological data and compare them with simple baselines before considering them reliable for future prediction.

Acknowledgements

The authors would like to thank Rainy Jain and Siddhant Gahlot at Buildemics for providing the collaborative network to make this work possible.

AI Use Statement

Artificial intelligence tools (Claude Opus 5.5) were used to assist with data analysis, code development, and figure generation in this study, as well as grammar and wording review. All results were independently verified by the authors.

References

[1] De Clercq, D., Mahdi, A.: Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data. Agricultural Systems **220**, 104099 (2024) https://doi.org/10.1016/j.agsy.2024.104099

[2] Fofanah, I.D.: Can machine learning forecast rice yields in data-constrained settings? Satellite climate data, national crop statistics, and lessons from Sierra Leone. Preprint at https://arxiv.org/abs/2606.13959 (2026)

[3] Petrovic, B., Paluba, D., Nofrizal, A.Y., Kucera, A.: Analyzing meteorological effects on crop yield and yield prediction through machine learning with different data partitioning approaches: A case study from Czech Republic. Journal of Agriculture and Food Research **26**, 102662 (2026) https://doi.org/10.1016/j.jafr.2026.102662

[4] Press Information Bureau, Government of India: Implementation of Pradhan Mantri Fasal Bima Yojana. Press release, 5 August 2016. https://www.pib.gov.in/newsite/PrintRelease.aspx?lang=2&reg=3&relid=148441 (2016)

[5] ICAR–Indian Institute of Rice Research: Rice APY Dashboard: Area, Production & Yield Analytics. https://icar-iirr.org/Rice_APY_Dashboard/. Accessed 20 September 2026

[6] Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data **2**, 150066 (2015) https://doi.org/10.1038/sdata.2015.66

[7] National Aeronautics and Space Administration, Langley Research Center: Prediction of Worldwide Energy Resources (POWER) Project. https://power.larc.nasa.gov/

[8] Government of Karnataka, Directorate of Economics and Statistics: District-wise report, 2009–10 (2010)

[9] Government of Karnataka, Directorate of Economics and Statistics: District-wise report, 2011–12 (2012)

[10] Government of Karnataka, Directorate of Economics and Statistics: District-wise report, 2017–18 (2018)

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[12] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, ??? (2016). https://doi.org/10.1145/2939672.2939785

[13] McKinney, W.: Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010). https://doi.org/10.25080/Majora-92bf1922-00a

[14] Harris, C.R., Millman, K.J., Walt, S.J., *et al.*: Array programming with NumPy. Nature **585**, 357–362 (2020) https://doi.org/10.1038/s41586-020-2649-2

[15] Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007) https://doi.org/10.1109/MCSE.2007.55