Hybrid Machine Learning Modelling for Rice Yield Forecasting under Climatic Variability in South Sudan

P. A. Kur[1]*, A. Waititu[2], T. Makoni[3]

[1]Pan African University for Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Jomo Kenyatta University of Agriculture and Technology, Kenya
[3]Great Zimbabwe University, Zimbabwe

ABSTRACT: Reliable agricultural yield forecasting is crucial for strengthening food security planning in South Sudan, where rice yields remain highly sensitive to climatic variability and structural production constraints. This study evaluates classical statistical and ensemble-based machine learning models for forecasting national rice yield using key agronomic and climatic determinants: cultivated land area, historical yield trends, rainfall, temperature, and soil pH. The empirical analysis draws on annual time-series data spanning nearly five decades (1961–2010), with the dataset partitioned chronologically into training (n=38) and testing (n=10) subsets to enable disciplined out-of-sample validation. Four modelling approaches were examined: Multiple Linear Regression (MLR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a Hybrid RF–XGBoost ensemble constructed via prediction aggregation. Model performance was assessed on the reserved test set using RMSE, MAE, R^2, and MAPE. Within the test sample (n=10), all models exhibited strong correlation ($R^2 > 0.91$). MLR achieved R^2=0.9128 but exhibited higher RMSE (295.60 kg/ha) and MAPE (31.75%) relative to the ensemble approaches, suggesting limited flexibility in capturing nonlinear yield responses. Random Forest improved performance (R^2=0.9647; RMSE=273.39; MAE=227.13; MAPE=28.26%), while XGBoost delivered R^2=0.9360 with RMSE=289.25, MAE=182.05, and MAPE=30.59%. The Hybrid model attained the lowest RMSE (259.14 kg/ha), lowest MAE (174.29), lowest MAPE (26.91%), and highest R^2 (0.9725). Although improvements are moderate in magnitude, the results suggest that hybrid ensemble aggregation can provide incremental gains in prediction accuracy under data-constrained conditions. Given the limited sample size, findings should be interpreted cautiously; nonetheless, this study offers an applied comparative assessment to support evidence-informed agricultural planning in fragile production environments.

KEYWORDS: Climate variability, Hybrid ensemble, Machine learning, Random Forest, Rice yield, XGBoost.

[Received Feb. 17, 2026; Revised Mar 5, 2026; Accepted Mar 5, 2026] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

Accurate crop yield forecasting plays a central role in agricultural planning, food security assessment, and climate adaptation strategy. Rising global food demand and pressures on land systems have intensified the need for more efficient and data-driven agricultural planning (Tilman et al., 2011). As production systems face increasing environmental and demographic stress, improving forecasting reliability becomes central to sustainable intensification strategies. Ensuring sustainable food supply for a growing global population remains a central development challenge, particularly in regions with climate-sensitive agricultural systems (Godfray et al., 2010). In such contexts, improved forecasting supports proactive rather than reactive food security policy. Reliable yield forecasts enable governments to anticipate potential shortfalls, guide import planning, allocate agricultural resources, and design risk-mitigation policies under climatic uncertainty. In developing economies, where agricultural systems are highly climate-sensitive and structurally constrained, forecasting reliability becomes particularly important for macro-level decision-making. Empirical analyses of global crop trends indicate that climatic variability has exerted measurable influence on yield patterns across multiple staple crops (Lobell, Schlenker and Costa-Roberts, 2011). These findings underscore the importance of incorporating climatic drivers into national yield forecasting frameworks.

*Corresponding author: *athian.peter@students.jkuat.ac.ke*

doi: http://doi.org/10.63746/njtd.v23i2.4454

Rice has become an increasingly important staple in Sub-Saharan Africa (SSA), driven by population growth, urbanization, and changing consumption patterns. In South Sudan, domestic rice yield remains vulnerable to rainfall variability, temperature fluctuations, and structural production constraints such as limited mechanization and weak statistical systems (FAO, 2012; World Bank, 2021; IPCC, 2021). Although substantial land resources exist, only a small proportion is cultivated, and national rice yield exhibits considerable inter-annual variability. These dynamics introduce uncertainty into agricultural planning and highlight the need for disciplined yield forecasting frameworks. Classical statistical approaches, including Multiple Linear Regression (MLR) and traditional time-series models, have long been used in agricultural yield forecasting due to their interpretability and transparency. However, such models typically rely on linearity and stationarity assumptions that may not fully capture nonlinear interactions among climatic and structural determinants. In climate-sensitive production systems, yield responses to environmental drivers often exhibit complex behaviour that challenges purely linear specifications.

Machine learning methods have gained prominence in yield forecasting because of their capacity to model nonlinear relationships and interaction effects with limited manual feature engineering (Jeong et al., 2016; Van Klompenburg, Kassahun and Catal, 2020). Recent evidence further suggests that machine learning approaches provide effective tools for assessing crop yield responses under climatic variability and climate change conditions (Crane-Droesch, 2018).

Such evidence supports the suitability of ensemble learning for agricultural modelling tasks involving nonlinear interactions. Recent comparative evaluations of machine learning algorithms under climatic variability further confirm the importance of rigorous out-of-sample validation when assessing predictive performance (Noorunnahar et al., 2023). Such studies emphasize that model comparison should be grounded in disciplined evaluation frameworks rather than reliance on in-sample goodness-of-fit statistics. Tree-based ensemble learners, such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost), have demonstrated promising prediction accuracy in agricultural applications across various contexts. Random Forest reduces prediction variance through bootstrap aggregation, while XGBoost reduces bias through sequential boosting and regularization. These approaches reflect complementary strategies within the bias–variance trade-off framework of statistical learning theory.

The suitability of highly flexible algorithms, however, depends critically on data structure. Many machine learning studies in yield forecasting rely on spatially disaggregated datasets, high-frequency remote sensing inputs, or relatively large sample sizes. Such data environments allow complex models to fully exploit their flexibility. In contrast, fragile agricultural systems characterized by short national time series and structural volatility present distinct modelling challenges. When only aggregated annual observations are available, overfitting risk increases and rigorous out-of-sample forecasting evaluation becomes essential. Under these data-constrained conditions, the incremental value of ensemble aggregation remains insufficiently examined. While hybrid models that combine multiple learners may theoretically balance bias and variance, their empirical contribution to out-of-sample yield forecasting performance in short national time-series datasets is not well established. Limited empirical evidence exists on the performance of hybrid tree-based ensemble aggregation using leakage-safe validation in data-scarce Sub-Saharan African contexts. Recent Sub-Saharan African yield forecasting studies have primarily focused on single-algorithm machine learning implementations rather than systematic hybrid ensemble evaluation under short national time-series conditions.

Accordingly, the central question addressed in this study is whether hybrid ensemble aggregation provides measurable improvement over classical linear benchmarks in out-of-sample rice yield forecasting when applied to short, aggregated national time-series data characterized by structural volatility. To investigate this question, the study compares four forecasting approaches: (i) Multiple Linear Regression (MLR) as a classical benchmark, (ii) Random Forest (RF), (iii) Extreme Gradient Boosting (XGBoost), and (iv) a Hybrid RF–XGBoost ensemble constructed through prediction aggregation. The empirical analysis is based on annual national-level rice yield data for the period 1961–2010. Post-2010 observations are excluded due to documented structural discontinuities associated with institutional transition and statistical instability, ensuring temporal comparability within the forecasting framework. The primary contribution of this research lies in a disciplined and leakage-safe comparative assessment of classical and ensemble-based yield forecasting approaches under severe data constraints, rather than in proposing a novel algorithm. Model performance is evaluated strictly using out-of-sample forecasting metrics to emphasize predictive stability over in-sample fit. By focusing on methodological transparency and cautious interpretation, the study provides applied evidence on forecasting performance in fragile production environments where reliable data remain limited.

II. THEORETICAL AND CONCEPTUAL FRAMEWORK

This study is anchored in two interrelated theoretical traditions—agricultural production theory under climatic variability and statistical learning theory—together with empirical evidence from machine learning applications in crop yield forecasting. These perspectives provide a structured lens through which the relationship between structural determinants, climatic variability, and yield forecasting performance can be examined, particularly in data-constrained environments such as South Sudan. Agricultural production theory conceptualizes crop yield as the outcome of interactions among cultivated area, climatic conditions, soil characteristics, and temporal persistence. In predominantly rain-fed systems across Sub-Saharan Africa, rice yield responds dynamically to rainfall variability, temperature fluctuations, and agroecological constraints. Climatic variability introduces stochastic shocks that contribute to inter-annual yield instability and complicate agricultural planning. At the national level, yield dynamics reflect both

structural influences—such as land-use expansion and production capacity—and environmental variability. Lagged yield effects capture persistence arising from accumulated agronomic practices, infrastructure limitations, and structural constraints, which are particularly relevant in emerging or fragile agricultural systems.

Traditional statistical approaches, including Multiple Linear Regression (MLR), provide transparent baseline benchmarks for estimating relationships between explanatory variables and yield. However, linear models impose constant marginal effects and additive functional form assumptions that may not adequately represent nonlinear climate–yield interactions.

Empirical research in agricultural modelling has shown that yield responses to climatic drivers often exhibit nonlinear behavior, interaction effects, and threshold dynamics that challenge strictly linear specifications (Lobell and Burke, 2010; Liakos et al., 2018; Schlenker and Roberts, 2009)

These limitations motivate the consideration of flexible modelling frameworks capable of capturing more complex relationships.

Statistical learning theory prioritizes prediction accuracy and generalization performance. Within this framework, model performance is governed by the bias–variance trade-off. Linear models may exhibit relatively low variance but higher bias when nonlinear relationships are present. Conversely, highly flexible nonlinear models can reduce bias but risk overfitting, particularly in short time-series datasets with limited observations. In data-constrained settings, controlling model complexity and enforcing disciplined validation procedures are essential to ensure reliable out-of-sample forecasting performance. The bias–variance trade-off forms a central principle in statistical learning theory, highlighting the balance between model flexibility and generalization stability (Hastie, Tibshirani and Friedman, 2009). Models that are overly rigid may suffer from high bias, while excessively flexible models may overfit limited data.

Tree-based ensemble methods operationalize structured bias–variance management. Random Forest (Breiman, 2001; Biau, 2012) reduces prediction variance through bootstrap aggregation and randomized feature selection, stabilizing model outputs across multiple trees.

Extreme Gradient Boosting builds upon the gradient boosting framework originally proposed by Friedman (2001) and subsequently enhanced through regularization and computational optimization by Chen and Guestrin (2016).

These complementary mechanisms make RF and XGBoost suitable candidates for yield forecasting when nonlinear relationships are suspected but sample size remains limited. Ensemble learning theory formally demonstrates that combining multiple learners can reduce generalization error when base models exhibit partially uncorrelated error structures (Zhou, 2012). Early foundational work on ensemble methods emphasized variance reduction and improved predictive stability through model aggregation (Dietterich, 2000). These theoretical insights continue to inform contemporary ensemble applications in environmental forecasting. Aggregation mechanisms therefore provide a theoretical basis for hybrid modelling frameworks.

The hybrid modelling framework evaluated in this study extends these approaches by aggregating predictions from Random Forest and XGBoost through convex weighting. Conceptually, if the base learners exhibit partially distinct error structures, weighted aggregation may reduce overall out-of-sample forecasting error relative to either model individually. While hybrid ensemble methods are increasingly applied in environmental and agricultural modelling, empirical evidence regarding their incremental benefit in short, aggregated national time-series data remains limited. Under small-sample conditions, any improvement from aggregation is expected to be moderate and should be interpreted cautiously.

Conceptually, this study links structural determinants, climatic variability, and predictive modelling approaches to policy-relevant yield forecasting outcomes. Cultivated area, rainfall, temperature, soil characteristics, and yield persistence form the explanatory domain influencing national rice yield. These inputs are processed through alternative modelling pathways—linear benchmark, variance-reducing ensemble, bias-reducing ensemble, and hybrid aggregation—and evaluated strictly using out-of-sample forecasting metrics. The resulting forecasts are interpreted as decision-support tools rather than deterministic projections. Given the constraints of limited historical data and structural volatility, model outputs are assessed in terms of relative forecasting stability rather than structural dominance. In summary, structural and climatic determinants influence national rice yield; alternative modelling frameworks approximate this relationship under different functional assumptions; and disciplined out-of-sample evaluation determines the relative reliability of each approach for supporting agricultural planning in South Sudan.

III. MATERIALS AND METHOD

A. Data Used and Sources

This study makes use of annual national-level time series data on rice yield and selected agro-climatic variables covering the period from 1961 to 2010. In total, the dataset consists of 48 yearly observations and includes rice yield (kg/ha), cultivated area (ha), annual rainfall (mm), mean annual temperature (°C), and national average soil pH.

Rice yield and cultivated area data were obtained from the Food and Agriculture Organization (FAOSTAT) database. These records were carefully cross-checked with data from the National Bureau of Statistics of South Sudan and the Ministry of Agriculture to ensure consistency and reliability. Annual rainfall and temperature data were sourced from the South Sudan Meteorological Department, while soil pH values were drawn from available agricultural and environmental reports reflecting national averages.

Although South Sudan became independent in 2011, agricultural statistics recorded after 2010 show clear inconsistencies linked to institutional restructuring, data system adjustments, and conflict-related disruptions. Including these years could introduce structural shifts that may distort estimation and weaken model stability. For this reason, the analysis was confined to the 1961–2010 period to maintain temporal continuity and ensure methodological consistency.

Structural and institutional challenges affecting agricultural data systems in South Sudan have been documented in sector reviews (FAO, 2012). Such discontinuities justify restricting the analysis to the pre-2011 period to preserve temporal consistency.

All variables used in this study are aggregated at the national level. While this level of aggregation supports long-term macro-level forecasting and aligns with national policy planning, it does not reflect subnational variations or farm-level heterogeneity. As a result, the findings should be interpreted as representing national production dynamics rather than localized agronomic responses. Given the limited size of the dataset (n = 48), special attention was given to maintaining a balance between model complexity and available observations. The training dataset contains 38 observations, while 10 observations were reserved for out-of-sample testing. The relatively small sample size limits the degrees of freedom available for estimation and requires cautious interpretation of performance metrics. Therefore, the modelling strategy prioritizes stability and disciplined validation over maximizing in-sample fit. Before modelling, all datasets were aligned using the year variable as a common temporal reference. The cleaning process involved verifying measurement units, checking for inconsistencies across sources, and correcting evident reporting errors. Where missing values were identified, they were addressed using conservative interpolation based on neighbouring years to preserve continuity without artificially altering structural patterns.

To reflect persistence in agricultural output, one-year and two-year lagged variables were generated for yield and selected climatic predictors. However, to prevent overfitting and excessive parameterization relative to sample size, only theoretically justified lag terms were retained. Predictor selection was intentionally constrained to maintain a reasonable ratio between variables and observations. The final dataset consisted of harmonized, validated, and temporally aligned annual observations prepared for chronological division into training and testing subsets. Importantly, no information from the test period was used during preprocessing, feature construction, or parameter tuning.

Table 1 Data used, sources, and their properties

Data	Source	Description
Rice Yield	FAOSTAT; National Bureau of Statistics; Ministry of Agriculture	Annual national rice yield (kg/ha)
Cultivated Area	FAOSTAT; Ministry of Agriculture	Annual harvested rice area (ha)
Rainfall	South Sudan Meteorological Department	Total annual rainfall (mm)
Temperature	South Sudan Meteorological Department	Mean annual temperature (°C)
Soil pH	Agricultural environmental reports	National average soil acidity level
Lagged Terms	Constructed from above variables	One- and two-year lagged values of key predictors

B. Data collection and processing.

Following compilation and validation, the dataset was organized using the year variable as the temporal index. All variables were aligned chronologically to ensure consistency across sources before modelling. Data preprocessing involved unit verification, consistency checks, and correction of identifiable reporting discrepancies. Where missing observations were detected, conservative interpolation based on adjacent-year values was applied to maintain continuity while preserving underlying structural variation. To reflect inter-annual persistence in agricultural yield, one-year and two-year lagged terms were constructed for yield and selected climatic predictors. The inclusion of lagged variables allows the models to capture delayed agronomic responses and temporal dependence. However, given the limited size of the training dataset (n = 38), only theoretically justified lag structures were retained to prevent overparameterization and unstable estimation.

The dataset was divided chronologically into a training set (1961–1998) and a test set (1999–2010). This time-ordered partition preserves temporal integrity and ensures that forecasting performance is evaluated under genuine out-of-sample conditions. No information from the test period was used during preprocessing, feature engineering, or model tuning. All predictor variables were standardized using parameters derived exclusively from the training data. This transformation was performed after the chronological split to prevent information leakage and to ensure that evaluation metrics reflect true predictive generalization rather than in-sample optimization.

The overall processing workflow therefore consisted of: (i) chronological alignment; (ii) data cleaning and validation; (iii) lag construction; (iv) training–testing partitioning; and (v) training-based standardization. These steps were implemented to ensure methodological transparency, reproducibility, and disciplined model evaluation. The overall methodological framework adopted in this study is presented in Figure 1. The framework summarizes the key stages, including data preparation, model development, and out-of-sample evaluation, ensuring a structured and leakage-free forecasting process.

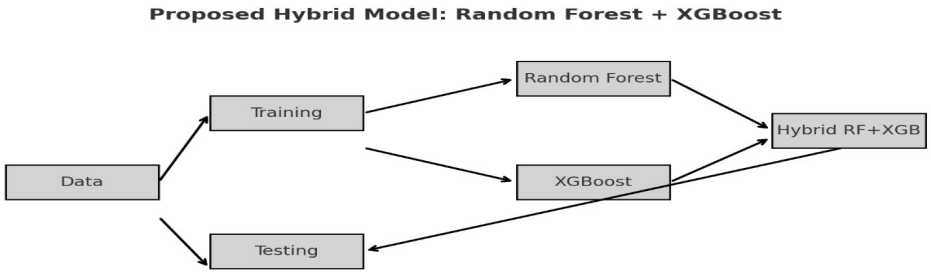


Figure 1 Overall methodological framework of the study

1) Baseline Model: Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) was implemented as the baseline statistical model to examine the linear relationship between rice yield and selected agro-climatic predictors. MLR estimates the conditional mean of the dependent variable as a linear combination of explanatory variables and provides an interpretable benchmark for evaluating nonlinear ensemble methods.

In this study, rice yield is modelled as a function of cultivated area, rainfall, temperature, soil pH, and selected lagged terms. The general MLR specification is expressed in Equation

$$Y_t = \beta_0 + \beta_1 A_t + \beta_2 R_t + \beta_3 T_t + \beta_4 S_t + \beta_5 Y_{\{t-1\}} + \beta_6 A_{\{t-1\}} + \varepsilon_t \quad (1)$$

where:

- $Y_t$represents rice yield at time t
- $A_t$denotes cultivated area
- $R_t$denotes annual rainfall
- $T_t$represents mean annual temperature
- $S_t$denotes soil pH
- $Y_{t-1}$and A_{t-1}represent lagged variables
- $\beta_0$is the intercept term
- $\beta_1, \ldots, \beta_6$are regression coefficients
- $\varepsilon_t$is the random error term

The regression coefficients were estimated using the Ordinary Least Squares (OLS) method, which minimises the sum of squared residuals between observed and predicted values. Under classical linear regression assumptions, the OLS estimator is unbiased and efficient, provided that error terms are homoscedastic and uncorrelated (Gujarati and Porter, 2009). These assumptions were assessed within the training dataset to ensure baseline model validity.

The OLS estimator is given in Equation (2).

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2)$$

where X is the matrix of explanatory variables and Y is the vector of observed yield values.

The MLR model assumes linearity between predictors and yield, independence of errors, homoscedasticity, and absence of perfect multicollinearity among explanatory provides a low-variance reference model with restricted complexity. Model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), variables. Diagnostic checks were conducted to assess these assumptions within the training dataset. MLR serves as a transparent and interpretable benchmark against which the predictive performance of Random Forest, XGBoost, and the hybrid ensemble model can be evaluated. Given the limited training sample (n = 38), MLR also

Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) computed on the out-of-sample test dataset

2). Random Forest Model

Random Forest (RF) was implemented as a nonlinear ensemble learning method to capture potential interaction effects and nonlinear relationships between rice yield and agro-climatic predictors. RF is a bagging-based algorithm that constructs multiple regression trees using bootstrap samples of the training data and aggregates their predictions to improve generalization performance. For regression tasks, the Random Forest prediction is obtained by averaging the outputs of individual decision trees. The RF estimator is expressed in Equation (3):

$$\hat{Y}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (3)$$

where:

B is the total number of trees in the forest

$T_b(x)$ denotes the prediction from the b-th regression tree

x represents the vector of predictor variables

Each tree is constructed using a bootstrap sample drawn from the training dataset. At each node split, a random subset of predictors (mtry) is considered, which reduces correlation among trees and enhances ensemble stability.

The RF algorithm minimises the mean squared error (MSE) within each tree through recursive binary partitioning. The split criterion selects the predictor and threshold that produce the largest reduction in impurity, measured for regression as:

$$\Delta I = I_{parent} - (I_{left} + I_{right}) \quad (4)$$

where impurity is defined as the within-node variance.

3. Extreme Gradient Boosting (XGBoost) Model

Extreme Gradient Boosting (XGBoost) was employed to model nonlinear relationships and interaction effects among agro-climatic predictors. Unlike Random Forest, which constructs trees independently, XGBoost builds trees sequentially, where each subsequent tree is fitted to the residual errors of the previous ensemble.

The final prediction is expressed as:

$$\hat{Y}_{XGB}(x) = \sum_{k=1}^{K} f_k(x) \qquad (5)$$

where $f_k(x)$denotes the k-th regression tree and $K$is the number of boosting iterations.

Model estimation minimises a regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (6)$$

where $l(\cdot)$represents the squared error loss and $\Omega(f_k)$is a regularisation term that penalises model complexity.

Given the limited training sample (n = 38), conservative hyperparameter settings were adopted. Tree depth and boosting rounds were restricted, and regularization parameters were tuned to mitigate overfitting risk. Subsampling procedures were also applied to enhance stability.

Model performance was evaluated using RMSE, MAE, MAPE, and R^2 computed strictly on the independent test dataset to ensure genuine out-of-sample assessment.

4. Hybrid RF–XGBoost Ensemble

To leverage the complementary strengths of bagging and boosting approaches, a hybrid ensemble combining Random Forest (RF) and XGBoost (XGB) was constructed. The hybrid model aggregates predictions from both base learners using convex weighting.

The hybrid prediction is defined as:

$$\hat{Y}_{Hybrid} = \alpha \hat{Y}_{RF} + (1-\alpha)\hat{Y}_{XGB} \qquad (7)$$

where $0 \leq \alpha \leq 1$represents the optimal weight assigned to the RF model, and $(1-\alpha)$corresponds to the weight for XGBoost.

To prevent information leakage, hybrid weights were estimated exclusively within the training dataset using five-fold cross-validation. In each fold, RF and XGBoost were trained on four subsets and used to generate out-of-fold (OOF) predictions for the held-out subset.

Out-of-fold (OOF) predictions used for weight optimization were generated strictly from models trained on disjoint training folds, ensuring that no observation contributed simultaneously to model fitting and weight estimation within the same fold.

The optimal weight was obtained by minimizing the squared error between observed yield and OOF predictions.

After determining α*, both RF and XGBoost were re-trained on the full training dataset, and final hybrid predictions were generated for the independent test period.

$$\alpha^* = \arg\min_{\alpha} \sum_{i=1}^{n} \left(y_i - \alpha \hat{y}_{RF,i}^{OOF} - (1-\alpha)\hat{y}_{XGB,i}^{OOF}\right)^2 \qquad (8)$$

No test observations were used during weight optimisation. This design ensures that reported hybrid performance reflects genuine out-of-sample evaluation rather than in-sample optimization. Given the limited training sample (n = 38), improvements from hybrid aggregation are interpreted cautiously.

The objective of the hybrid approach is not to introduce additional model complexity, but to reduce prediction variance when base learners exhibit partially distinct error structures.

Hybrid aggregation was limited to RF and XGBoost because both are nonlinear tree-based ensemble learners with complementary bias–variance properties.

Combining these models allows variance reduction (RF) and bias reduction (XGBoost) mechanisms to interact. Hybridization with MLR was not pursued, as combining a linear parametric estimator with nonlinear ensemble learners may introduce structural bias rather than complementary variance structures, particularly under small-sample conditions. Model performance was assessed using RMSE, MAE, MAPE, and R^2 computed on the reserved test data.

5). Model Evaluation

Model performance was assessed using standard regression accuracy metrics computed strictly on the independent test dataset (1999–2010, n = 10). No training observations were used in performance reporting. This ensures that the evaluation reflects genuine out-of-sample forecasting behavior. The selection of RMSE, MAE, and MAPE follows established guidance in forecast accuracy evaluation, where error-based measures are recommended over sole reliance on goodness-of-fit statistics (Hyndman and Koehler, 2006). These metrics provide complementary insights into scale-dependent and proportional forecasting error.

Given the limited size of the test sample, performance statistics are interpreted cautiously and represent model behavior within the evaluated period rather than evidence of structural permanence or large-sample generalization.

Although classical ARIMA models are widely used in agricultural time-series forecasting, their application in this study was constrained by the short sample length (48 annual observations) and potential structural instability. Preliminary stationarity diagnostics indicated limited robustness for ARIMA parameter estimation under these conditions. Therefore, the empirical comparison focused on cross-sectional and ensemble learning frameworks that rely less heavily on strict stationarity assumptions. The following metrics were employed:

5.1 Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared prediction error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (9)$$

RMSE is expressed in kilograms per hectare (kg/ha), allowing direct interpretation of forecast error magnitude in agricultural terms. Lower RMSE indicates improved predictive precision.

5.2. Mean Absolute Error (MAE)

MAE captures the average absolute deviation between observed and predicted yield:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \qquad (10)$$

Unlike RMSE, MAE does not disproportionately weight large errors and therefore provides a complementary measure of average forecast deviation.

5.3 Mean Absolute Percentage Error (MAPE)

MAPE expresses prediction error relative to observed yield values:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} | \frac{y_i - \hat{y}_i}{y_i} | \qquad (11)$$

This metric enables scale-independent comparison across models and facilitates interpretation of proportional error.

5.4 Coefficient of Determination (R^2)

The coefficient of determination, R^2., (equation 12) is indicates the proportion of variation in test-period yield explained by model predictions. However, with a small test sample (n = 10), R^2 may be sensitive to individual observations and should not be interpreted as a definitive measure of structural explanatory power. For this reason, model comparison emphasizes error-based metrics (RMSE and MAE) rather than relying solely on R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (12)$$

All metrics were calculated on the same reserved test dataset to ensure comparability across MLR, Random Forest, XGBoost, and the hybrid ensemble. Differences in error values are interpreted in practical terms (kg/ha) rather than solely as statistical improvements.

IV. RESULTS AND DISCUSSION

The descriptive statistics in Table 2 provide an overview of the distributional properties of rice yield and associated environmental variables over the study period. The mean rice yield was 1,434.08 kg/ha, with a wide observed range of 4,261.90 kg/ha, indicating substantial inter-annual variability. The median yield (1,027.80 kg/ha) is notably lower than the mean, suggesting the presence of higher-yield years that influence the distribution.

Cultivated area exhibits considerable dispersion, with a range of 35,606 ha, reflecting structural fluctuations in production scale over time. Rainfall shows moderate variability, while temperature and soil pH display relatively narrower ranges, indicating more stable climatic and soil conditions compared to yield and area.

The lagged yield variables exhibit similar distributional characteristics to current yield, reflecting temporal persistence

in production patterns. These descriptive patterns justify the inclusion of lagged terms and nonlinear modelling approaches in subsequent analysis.

Table 2 Summary Statistics of Rice Yield and Related Environmental Variables (1961–2010)

Statistic	Area (ha)	Yield (kg/ha)	Soil pH	Temperature (°C)	Rainfall (mm)	$Yield_{t-1}$ (kg/ha)	$Yield_{t-2}$ (kg/ha)
Mean	7,726.13	1,434.08	6.92	26.85	234.79	1,395.03	1,350.03
Median	4,500.00	1,027.80	7.22	26.84	233.00	1,027.80	1,027.80
Range	35,606.00	4,261.90	1.47	2.29	180.00	4,261.90	4,261.90
Q1 (25%)	1,425.00	858.70	6.41	26.44	218.25	858.70	858.70

Figure 2 presents the Pearson correlation matrix among the principal study variables. Rice yield exhibits a strong positive correlation with cultivated area (r = 0.91) and a moderate association with time (year), suggesting structural and expansion-related influences within the aggregated national dataset. Climatic variables show comparatively weaker linear correlations with yield, while soil pH demonstrates a moderate negative association. These correlations reflect linear relationships and do not imply causation. Moreover, tree-based models are capable of capturing nonlinear interactions beyond pairwise correlation patterns.

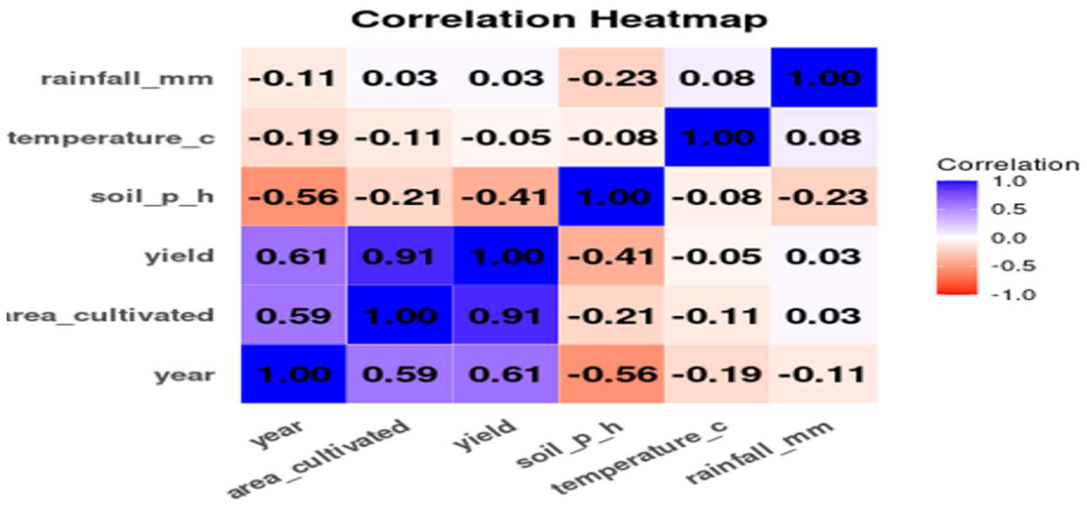
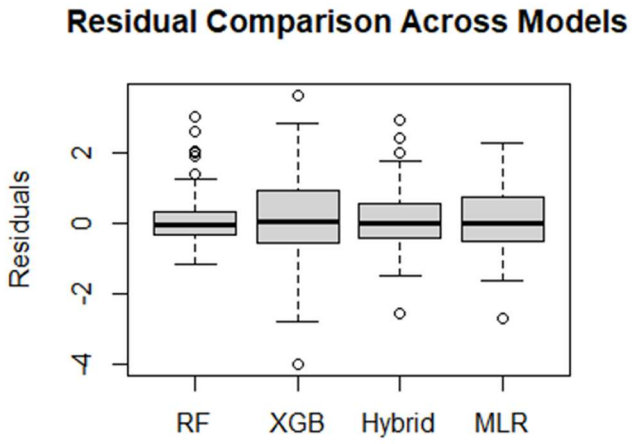


Figure 2 Correlation matrix of key variables

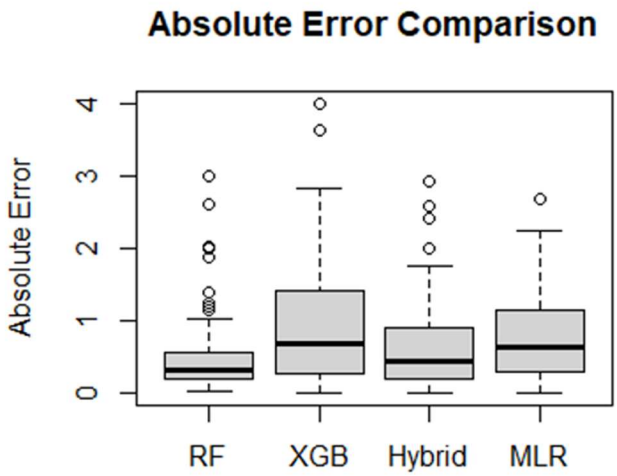
The residual plots show that all models produce errors centred around zero, which indicates the absence of systematic overestimation or underestimation. However, the variability of residuals differs across models. Random Forest and the Hybrid model display a narrower spread of residuals, suggesting more consistent predictions. XGBoost shows a wider dispersion and more extreme values, indicating less stability across observations. Multiple Linear Regression also exhibits noticeable spread, though not as large as XGBoost, reflecting moderate inconsistency. The absolute error distribution supports this pattern. Random Forest has the smallest median absolute error and a compact range, indicating relatively precise predictions. The Hybrid model follows with slightly higher variability but still maintains controlled error levels. XGBoost shows larger and more dispersed errors, while Multiple Linear Regression records moderate error magnitudes, higher than Random Forest and the Hybrid model.

The radar chart summarises model performance using RMSE, MAE, R^2, and MAPE. The Hybrid model shows balanced performance across all metrics, combining low error values with high explanatory power. Random Forest performs well in reducing errors, while XGBoost achieves high R^2 but with larger error measures. Multiple Linear Regression shows acceptable explanatory ability but higher prediction errors compared to the ensemble models.

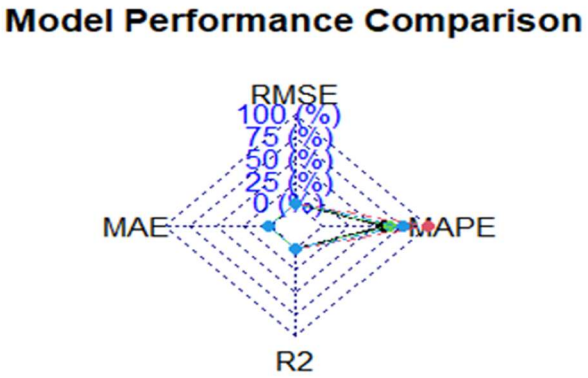
These results indicate that ensemble-based approaches provide more stable and accurate predictions in this application, although the small test sample limits the strength of inference.



a) Residual comparison across RF, XGB, Hybrid Random Forest-XGB, and MLR models



b). Absolute error comparison across (RF), (XGB), Hybrid (Hybrid), and (MLR) models



c) Radar chart comparing the performance of Random Forest (RF), Extreme Gradient Boosting (XGB), Hybrid RF-XGB Model.

Figure 3 Comparative performance evaluation of RF, XGB, Hybrid Random Forest-XGB, and MLR models using (a) residuals, (b) absolute errors, and (c) performance metrics (RMSE, MAE, MAPE, and R^2).

As illustrated in Table 3, the XGBoost model was trained on 38 observations and evaluated on an independent test set of 10 observations. Within the evaluated test sample, the model achieved an RMSE of 289.25 kg/ha and an MAE of 182.05 kg/ha, with an R^2 of 0.9360. While the R^2 value indicates high test-sample correlation between observed and predicted yield, interpretation is undertaken cautiously due to the limited sample size. Error-based measures suggest reasonable predictive performance under data-constrained conditions, though results are assessed comparatively against alternative modelling approaches rather than in isolation.

Table 3 Out-of-Sample Performance of the XGBoost Model

Model	RMSE (kg/ha)	MAE (kg/ha)	R^2	MAPE (%)	Train	Test
XGBoost	289.25	182.05	0.9360	30.59	38	10

Table 4 summarizes the predictive performance of the Random Forest model evaluated on the independent test dataset. The model achieved an RMSE of 273.39 kg/ha and an MAE of 227.13 kg/ha, with an R^2 of 0.9647 within the evaluated sample. While the R^2 suggests high alignment between predicted and observed values within the test sample, error-based metrics provide a more reliable indication of forecasting accuracy. The relatively higher MAE compared to some competing models indicates sensitivity to certain annual deviations, despite favourable RMSE performance. Lastly, the Random Forest model demonstrates robust predictive capability under the data constraints of this study.

Table 4 Out-of-Sample Performance of the Random Forest Model (Test Period: 1999–2010)

Metric	**Value**
RMSE (kg/ha)	273.39
MAE (kg/ha)	227.13
R^2	0.9647
MAPE (%)	28.26

Table 5 below presents the out-of-sample forecasting performance of the models over the independent test period (1999–2010). All models show high test-sample correlation ($R^2 > 0.91$). However, these values are interpreted with caution due to the small test sample (n = 10) and are not relied upon as the main criterion for comparison. Instead, error-based measures provide a more robust basis for evaluation. The hybrid ensemble model records the lowest RMSE (259.14 kg/ha) and MAE (174.29 kg/ha), indicating improved predictive accuracy compared to the baseline Multiple Linear Regression and the individual machine learning models. Random Forest shows relatively low RMSE but higher MAE, suggesting variation in error distribution. XGBoost demonstrates comparable performance, although it does not consistently exceed the hybrid model across all metrics.

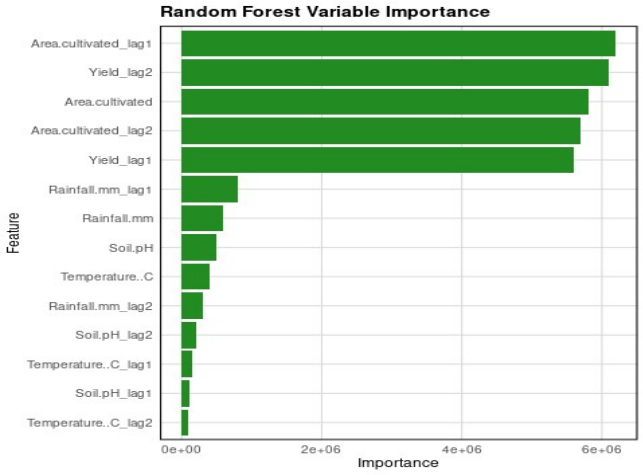
The results indicate that combining models can improve predictive performance, although the magnitude of improvement remains moderate within the test sample. Similar findings have been reported in systematic reviews of agricultural forecasting studies where ensemble and hybrid machine learning methods often outperform individual algorithms by exploiting complementary prediction strengths (Van Klompenburg, Kassahun and Catal, 2020)

The relative importance of predictor variables across the ensemble models is shown in Figure 3. The results indicate that both climatic factors and lagged variables contribute to explaining variations in rice yield within the modelling framework

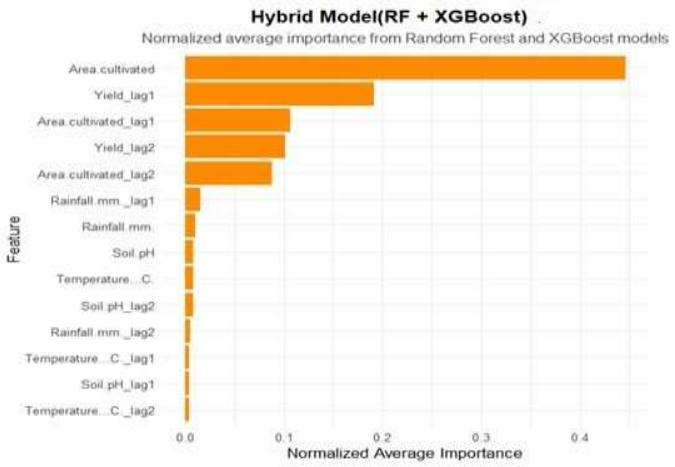
Table 5 Out-of-Sample Performance Comparison of Predictive Models (Test Period: 1999–2010)

Model	RMSE (kg/ha)	MAE (kg/ha)	R^2	MAPE (%)
MLR	295.60	190.42	0.9128	31.75
Random Forest	273.39	227.13	0.9647	28.26
XGBoost	289.25	182.05	0.9360	30.59
Hybrid	**259.14**	**174.29**	0.9725	**26.91**

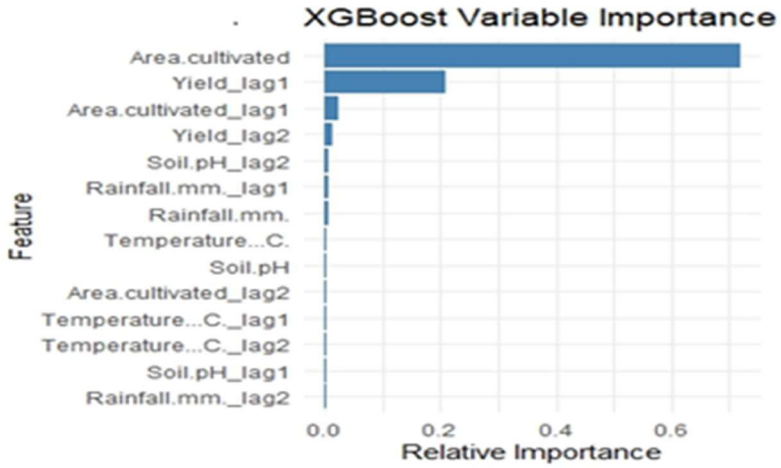
The hybrid ensemble exhibits a comparable importance distribution, with smoother weighting due to averaging across model structures. This consistency reinforces the prominence of structural and lagged variables within the evaluated dataset. Climatic variables—including rainfall, temperature, and soil pH (both current and lagged)—display comparatively lower importance scores across models. However, it is important to emphasize that tree-based importance metrics are sensitive to variable variance, scale, and temporal persistence. Lower relative importance does not imply negligible agronomic relevance; rather, it reflects their contribution within the specific modelling framework and aggregated national data used in this study. Overall, the variable importance results suggest that yield persistence and cultivated area dynamics provide substantial explanatory structure for national-level rice yield forecasting over the study period, while climatic variables contribute complementary information. These findings are interpreted cautiously in light of data aggregation and limited sample size.



a) Random forest Variable Importance



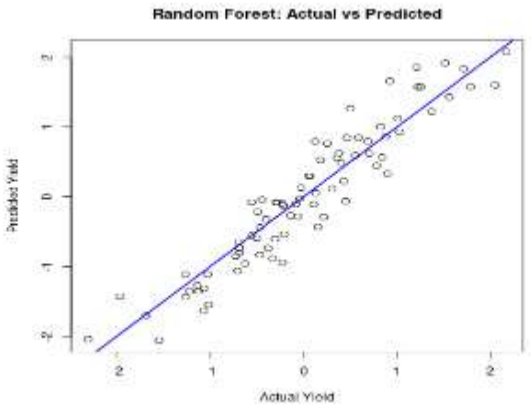
b) Hybrid RF-XGB Variable Importance



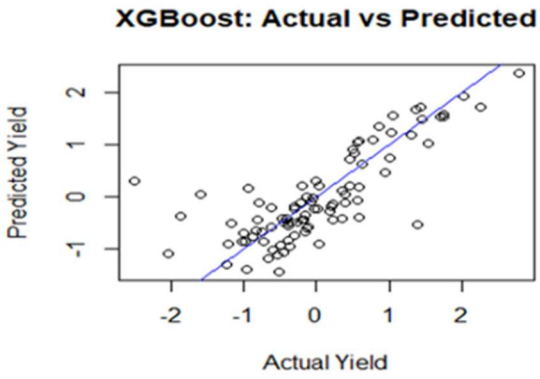
c) XGBoost Variable Importance

Figure 4 Variable importance across models: (a) Random Forest, (b) Hybrid RF–XGBoost, (c) XGBoost

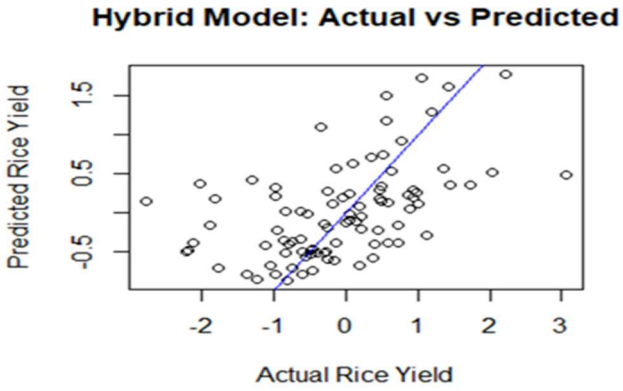
The three scatter plots present a visual assessment of the forecasting performance of the Hybrid, XGBoost, and Random Forest models by comparing observed (actual) rice yield with predicted values on the test sample. The 45-degree reference line represents perfect forecasts, where predicted values exactly match observed outcomes. In a forecasting context, the closeness of observations to this line indicates the model's ability to produce accurate out-of-sample predictions. The Random Forest model exhibits the strongest forecasting performance, with predictions closely clustered along the line of perfect prediction, suggesting lower forecast errors and higher reliability. The XGBoost model also demonstrates good forecasting capability, although with slightly greater dispersion, indicating moderate deviations from actual values. In contrast, the Hybrid model shows a wider spread of points around the reference line, reflecting higher forecast uncertainty and reduced precision in out-of-sample predictions. Overall, while all models capture the general pattern of rice yield variability, the Random Forest model provides the most accurate and stable forecasts, followed by XGBoost, whereas the Hybrid model appears less effective under the given data conditions. These visual findings are consistent with the quantitative evaluation metrics, reinforcing the robustness of the Random Forest model for forecasting rice yield.



a) random forest model (actual vs predicted).



b) XGBoost model (Actual vs predicted)



c) Hybrid Model (Actual vs predicted)

Figure 5 Comparison of Predictive Performance across the models: (a) random forest (actual vs predicted), (b) XGBoost (Actual vs predicted Hybrid), (c) Hybrid Model RF-XGB (Actual vs predicted)

V. CONCLUSION

This study assessed the out-of-sample forecasting performance of classical and machine learning approaches for national rice yield prediction in South Sudan under constrained data conditions. Using annual observations from 1961 to 2010, Multiple Linear Regression, Random Forest, XGBoost, and a hybrid RF–XGBoost ensemble was evaluated within a strictly time-ordered validation framework to preserve forecasting integrity. The empirical results show that all models achieved high test-sample correlation; however, given the limited size of the test set (n = 10), these values are treated cautiously. Greater emphasis was placed on error-based measures. Within this context, the hybrid ensemble produced the lowest RMSE and MAE, indicating improved predictive precision relative to both the linear benchmark and the individual tree-based models. Random Forest demonstrated stable performance, particularly in terms of RMSE, although its higher MAE suggests sensitivity to certain yearly deviations. XGBoost delivered competitive results, but its performance was less consistent when compared across all evaluation metrics.

The gains from hybrid aggregation are present but remain moderate, reflecting the constraints imposed by short, aggregated longitudinal data. Recent studies have shown that machine learning performance estimates derived from small datasets can be unstable and may not generalize reliably beyond the evaluation sample (Hawinkel et al., 2023).

These results suggest that combining learners with complementary bias–variance properties can enhance prediction stability, even when improvements in absolute error are incremental. At the same time, the magnitude of these gains highlights the importance of cautious interpretation, as small samples limit the generalisability of model performance. From an applied perspective, the findings indicate that ensemble-based approaches can serve as practical forecasting tools in data-limited environments. Rather than replacing classical models, they provide an additional layer of robustness by capturing nonlinear patterns and reducing variability in predictions. This is particularly relevant for agricultural planning in fragile systems, where decisions must often be made under uncertainty and with imperfect data.

Future research should focus on expanding both the depth and structure of available data. Recent advances in crop forecasting suggest that machine learning and deep learning frameworks can achieve substantial improvements in predictive accuracy when larger and more diverse datasets become available (Khaki and Wang, 2019; Van Klompenburg, Kassahun and Catal, 2020). Further work is also needed to examine the robustness of hybrid models under alternative validation schemes, including rolling-origin and cross-validation approaches tailored to time series data. In addition, incorporating benchmark time-series models such as ARIMA, where data conditions permit, would provide a more comprehensive comparative framework. Finally, exploring uncertainty quantification and probabilistic forecasting methods would enhance the practical relevance of yield predictions for policy and decision-making under risk.

AUTHOR CONTRIBUTIONS

P. A. Kur: Conceived the study, designed the methodology, conducted analysis, and drafted the manuscript. **A. Waititu:** Supervised the research and contributed to methodological refinement. **T. Makoni**: Provided technical review and contributed to manuscript editing.

ACKNOWLDGEMENT

The authors gratefully acknowledge the Pan-African University Institute for Basic Sciences, Technology and Innovation (PAUSTI) for providing academic support and a conducive research environment. Appreciation is also extended to the National Bureau of Statistics of South Sudan and the Ministry of Agriculture for facilitating access to relevant agricultural and climatic data. The authors further thank colleagues and reviewers for their constructive feedback, which significantly enhanced the clarity, coherence, and academic rigor of this study.

REFERENCES

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). Time Series Analysis: Forecasting and Control. 5th ed. Hoboken: John Wiley & Sons.

Biau, G. (2012). Analysis of a random forests model. Journal of Machine Learning Research, 13, pp. 1063–10

Breiman, L. (2001). Random forests. Machine Learning, 45(1), pp. 5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In: Multiple Classifier Systems. Lecture Notes in Computer Science, vol. 1857, pp. 1–15.

FAO (2012). South Sudan: Agricultural Sector Review. Rome: Food and Agriculture Organization of the United Nations.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M. and Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. Science, 327(5967), pp. 812–818.

Gujarati, D.N. and Porter, D.C. (2009). Basic Econometrics. 5th ed. New York: McGraw-Hill.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

Hawinkel, S., Fiers, M., Verbist, B., Foré, S. and others (2023). A broken promise: Machine learning for prediction is difficult in small datasets. Patterns, 4(9), 100804.

Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), pp. 679–688.

Intergovernmental Panel on Climate Change (IPCC). (2021). Climate Change 2021: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report. Cambridge: Cambridge University Press.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M. and Reddy, V.R. (2016). Random forests for global and regional crop yield predictions. PLoS ONE, 11(6), e0156571.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), pp. 1189–1232.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

Lobell, D.B. and Burke, M.B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), pp. 1443–1452.

Molnar, C. (2022). Interpretable Machine Learning. 2nd ed. Munich: Leanpub.

Schlenker, W. and Roberts, M.J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), pp. 15594–15598.

Van Klompenburg, T., Kassahun, A. and Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.