A-Rice: A Machine Learning-based Approach using Random Forest and Gaussian Process Regression for Forecasting Rice Production in Pampanga

Lance Angelo P. Arcega
Angeles University Foundation
College of Computer Studies
Angeles City
Pampanga, Philippines

Jeremy D. Samonte
Angeles University Foundation
College of Computer Studies
Angeles City
Pampanga, Philippines

Jomar Benedict B. Balagtas
Angeles University Foundation
College of Computer Studies
Angeles City
Pampanga, Philippines

Ian Miguel M. Bautista
Angeles University Foundation
College of Computer Studies
Angeles City
Pampanga, Philippines

Melissa M. Pantig
Angeles University Foundation
College of Computer Studies
Angeles City
Pampanga, Philippines

ABSTRACT

Rice production forecasting is critical for food security planning in the Philippines, yet existing estimation methods in provinces like Pampanga remain reliant on historical averages and manual assessments. This study presents A-Rice, a sequential hybrid machine learning model that integrates Random Forest (RF) and Gaussian Process Regression (GPR) for municipal-level rice production prediction across 20 municipalities over a 25-year period (2000–2025). The model was developed using historical agricultural data from the Philippine Statistics Authority and the Office of the Provincial Agriculturist, combined with six environmental variables from NASA POWER. Bayesian Small Area Estimation was applied to reconstruct missing municipal-level records for 2000–2016. In the hybrid framework, RF first generates baseline predictions and identifies key predictive features, after which GPR refines these predictions by modeling residual errors and quantifying uncertainty through calibrated confidence intervals. Evaluated against standalone RF and GPR models, the hybrid RF→GPR model achieved R^2 values exceeding 0.99 across all data splits, with an 80% reduction in RMSE on the test set (from 5,555 to 1,130 metric tons). Feature importance analysis revealed that municipality, cropping season, and relative humidity were the most influential predictors. The model was deployed as a web-based decision-support application providing production forecasts with confidence intervals, environmental diagnostics, and historical benchmarking. Results demonstrate the viability of sequential hybrid ML approaches for precision agriculture and data-driven agricultural planning in developing regions.

General Terms

Algorithms, Machine Learning, Prediction, Decision Support System

Keywords

Random Forest, Gaussian Process Regression, Hybrid Model, Rice Production, Crop Yield Prediction, Decision Support, Precision Agriculture, Pampanga

1. INTRODUCTION

Agriculture remains a cornerstone of the Philippine economy, employing over 24% of the national workforce and serving as a key livelihood source in rural areas [1]. Despite its importance, the sector contributed only 9.4% to national GDP in 2023 and 11% to the 2024 Regional GDP in Central Luzon. Rice, the country's most critical crop for food security, continues to face persistent productivity challenges. Central Luzon is the country's top rice-producing region, and Pampanga is among its key rice-producing provinces, with irrigated and rainfed palay cultivation spanning 20 municipalities across two cropping seasons annually. However, despite its agricultural significance, the province continues to experience declining yields, high input costs, aging irrigation infrastructure, and increasing environmental unpredictability.

Current methods of estimating rice yield rely on historical averages and generalized assumptions, which are insufficient for modern decision-making [2]. The Senate Economic Planning Office (2024) and the Central Luzon Regional Development Plan (CL-RDP) 2023–2028 have emphasized the urgent need for technological adoption in the agricultural sector. Without accurate, localized forecasting tools, agronomists and local government units (LGUs) are unable to make informed decisions about resource distribution, input allocation, and agricultural interventions—leading to suboptimal planning and continued production inefficiencies.

Machine learning (ML) models have demonstrated success globally in predicting crop yields with high accuracy [3][4]. Among these, Random Forest (RF) is valued for its robustness in handling complex datasets and its ability to rank feature importance, while Gaussian Process Regression (GPR) offers probabilistic forecasts that quantify prediction uncertainty through confidence intervals [5][6]. Both approaches have been validated individually in agricultural contexts, yet their sequential integration for crop production prediction remains largely unexplored.

Critically, no system currently exists for rice production prediction specifically tailored to Pampanga, and few studies have explored the combined use of RF and GPR in the

Philippine agricultural context [7][8]. While Lagrazon and Tan [7] applied GPR for rice yield prediction in Quezon Province, and Parreño and Anter [13] used RF at the national level, neither study combined these models sequentially or operated at the municipal level. This gap is significant because municipal-level predictions are necessary for LGU-level planning, which is where agricultural resource allocation decisions are actually made.

To address this gap, this study proposes A-Rice, a sequential hybrid RF–GPR model tailored to Pampanga's environmental and agricultural conditions. The model aims to serve as a decision-support tool for local farmers and policymakers by enabling precision agriculture and supporting evidence-based planning aligned with the CL-RDP 2023–2028. Specifically, this study aims to: (1) identify and prepare relevant agricultural and environmental datasets covering 20 municipalities from 2000 to 2025; (2) implement RF for initial prediction and feature importance analysis; (3) apply GPR for uncertainty modeling and prediction refinement; (4) evaluate the hybrid model against standalone approaches using RMSE, MAE, and R^2; and (5) deploy the model as a web-based decision-support application.

2. RELATED WORK

2.1 Machine Learning in Agriculture

ML has proven valuable in agriculture for forecasting crop yields, monitoring crop health, and optimizing farm operations. Unlike traditional methods that rely on historical averages, ML enables the analysis of complex datasets to reveal nonlinear relationships between environmental variables and crop yields [3]. Popular ML models include RF, valued for robustness and interpretability; Support Vector Machines (SVM), commonly used for classification; and GPR, known for modeling prediction uncertainty. More advanced models such as XGBoost and Artificial Neural Networks also offer high accuracy in large-scale applications.

Jabed and Murad [9] emphasized that ML models outperform traditional techniques, particularly when integrated with weather and soil data. Rashid et al. [10] highlighted ML's growing impact in yield prediction and the importance of selecting models based on dataset characteristics. Wolanin et al. [3] advocated for broader adoption of ML in agriculture. These studies affirm ML's capacity to advance precision agriculture in developing countries like the Philippines.

2.2 Small Area Estimation for Data Gap Filling

Agricultural studies often face challenges in obtaining reliable yield estimates due to limited sample sizes or missing observations. Small Area Estimation (SAE) techniques address these issues by integrating auxiliary and model-based information to produce reliable estimates for areas with scarce data [19]. Ghosh and Rao [19] demonstrated that SAE provides a statistically principled framework by borrowing strength from related areas through Bayesian hierarchical models. This method is crucial for agricultural applications where data collection is limited by cost, logistics, or environmental variability. In this study, Bayesian SAE was applied to reconstruct missing municipal-level data for 2000–2016, ensuring statistical validity across the complete dataset.

2.3 Random Forest in Crop Yield Prediction

RF is widely used in agriculture due to its ability to generalize across diverse datasets while reducing overfitting through aggregating multiple decision trees. Basha et al. [4] and Shawon et al. [11] reported low RMSE and high R^2 scores in yield prediction tasks. Zhou et al. [12] demonstrated strong performance ($R^2 \approx 0.85$) in mountainous rice-growing regions, confirming robustness across complex geographies. Choudhary et al. [14] used RF with Sentinel-2 imagery and environmental variables to predict rice yield, while Köse et al. [15] showed RF outperformed classical statistical and other ML models using climatic data.

In the Philippines, Parreño and Anter [13] found that RF outperformed other models in predicting rice and corn yields using national-level data. These findings collectively demonstrate RF's potential as a foundational component of the proposed hybrid model for rice yield forecasting.

2.4 Gaussian Process Regression in Crop Forecasting

GPR is a non-parametric, probabilistic model that provides both predictive estimates and confidence intervals, making it well-suited for risk-sensitive decision-making [5]. Ghosh et al. [5] and Liu et al. [16] highlighted GPR's effectiveness in modeling biophysical parameters across diverse conditions. Martínez-Ferrer et al. [6] demonstrated that GPR produced more accurate and interpretable forecasts compared to SVR and Random Forest Regression.

In the Philippines, Lagrazon and Tan [7] found GPR outperformed Decision Trees, Ensemble Models, Neural Networks, and SVM in predicting rice yields in Quezon Province, achieving the lowest RMSE (1785.04) and highest R^2 (0.968). Mabunga and Dela Cruz [23] applied GPR for soil moisture forecasting in the Philippines, achieving R^2 of 0.96 with an optimized model. Ekanayake and Wickramasinghe [17] confirmed GPR's superiority over ANN, SVMR, and MLR in paddy yield prediction in Sri Lanka. Hassan et al. [18] demonstrated GPR's effectiveness in mechanized rice farming contexts, outperforming ANN and linear regression. GPR's adaptability to nonlinear relationships and its ability to model uncertainty make it an excellent complement to RF in the proposed hybrid approach.

2.5 Hybrid and Comparative Models

Combining RF and GPR leverages both models' strengths for improved accuracy and flexibility. Hariyani et al. [8] demonstrated that ensemble methods outperform single-model approaches based on RMSE and MAE metrics. Rashid et al. [10] recommended RF-GPR hybrids for predicting crop yields under uncertain conditions. While individual RF and GPR models have been applied successfully in various contexts, few studies have explored their sequential integration, particularly in the Philippine setting. The hybrid RF-GPR approach offers a complementary blend—RF provides interpretability and computational efficiency, while GPR offers robustness in uncertainty estimation. This research gap motivates the present study's development of a sequential hybrid RF→GPR model specifically designed for rice production prediction in Pampanga.

3. METHODOLOGY

3.1 Research Design and Framework

This study employed a data science research design centered on the development, training, evaluation, and comparison of three machine learning models: standalone Random Forest Regression (RF), standalone Gaussian Process Regression (GPR), and a sequential hybrid of RF and GPR. The research aimed to generate a predictive model based on historical agricultural and environmental datasets to determine which model performed best in terms of accuracy and uncertainty estimation.

The research adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which provided a structured and adaptable methodology for conducting data-driven projects. CRISP-DM was selected because of its systematic approach to data mining and analytics and its suitability for machine learning applications in agriculture, with emphasis on iterative refinement and domain understanding. The framework consists of five major phases: (1) Data Understanding, involving the collection, exploration, and quality assessment of historical rice yield data and environmental variables from government sources; (2) Data Preparation, where raw data is cleaned, transformed, and organized into a structured format suitable for modeling, including handling missing values, standardizing numerical features, encoding categorical variables, and splitting the dataset; (3) Modeling, where the preprocessed data is used to train and test machine learning models using both standalone and sequential hybrid approaches; (4) Evaluation, where model performance is assessed using statistical metrics such as RMSE, MAE, MSE, and R^2, including uncertainty analysis from GPR; and (5) Deployment, where the best-performing model is integrated into a user-friendly web-based application.

3.2 Data Sources and Collection

Datasets were obtained from three complementary sources that collectively captured environmental, agronomic, and production-related variables influencing rice output in Pampanga.

The first source was the Philippine Statistics Authority's OpenSTAT repository, which provided provincial-level historical rice production data. This dataset included area harvested (in hectares) and volume of production (in metric tons) for two ecosystem categories: irrigated palay and rainfed palay. The records spanned from 2000 to the first quarter of 2025, available in quarterly, semiannual, and annual formats. The provincial dataset contained 532 quarterly observations across the coverage period. Yield per hectare was computed from the area harvested and production volume. The data was downloaded in CSV format as recommended by Mr. Al Jonnel L. Espiritu of the National Economic and Development Authority (NEDA).

The second source was the Office of the Provincial Agriculturist (OPA) in San Fernando, Pampanga, which supplied more recent and localized municipal-level data on rice farming for 2017–2024 across 20 municipalities. Variables captured in these records included area planted, area harvested, production (in metric tons), yield per hectare, and the number of farmers engaged in cultivation. Features related to palay crop growth were identified based on recommendations from domain experts at the Philippine Rice Research Institute (PhilRice) and OPA, obtained through face-to-face interviews and online consultations.

The third source was the NASA Prediction of Worldwide Energy Resources (POWER) database, which provided environmental variables across 20 municipalities in Pampanga. This dataset included six key parameters relevant to crop growth: UVB irradiance, maximum air temperature, minimum air temperature, relative humidity, precipitation, and root zone soil wetness. These variables capture essential aspects of plant stress, water availability, and climatic conditions affecting rice production.

3.3 Data Preprocessing

All datasets underwent a series of preprocessing and transformation procedures to ensure data quality, consistency, and compatibility for machine learning model development. All procedures were implemented using Python, primarily through the Pandas and NumPy libraries within a Google Colab environment.

In the cleaning phase, column names were standardized to lowercase with underscores via a snake_cols function to ensure consistent naming conventions. Temporal variables were converted to integer format using a parse_year_to_int function, and cropping seasons were normalized into two standardized categories—Dry Season (DS) and Wet Season (WS)—using a norm_season function. Non-numeric entries and typographical inconsistencies in production and area fields were converted into numerical form. Data entries that were invalid or missing were replaced by NaN for subsequent handling. Units of measurement were standardized across all datasets, with production expressed in metric tons and area in hectares.

In the transformation phase, numerical features were normalized and standardized using StandardScaler from scikit-learn, which is particularly important for GPR's sensitivity to the scale of input features. Categorical features, specifically cropping season and municipality, were encoded using OneHotEncoder. Missing values in continuous variables such as rainfall and temperature were imputed using SimpleImputer with mean substitution based on time-period averages. The transformation pipeline was organized using ColumnTransformer and Pipeline for cleaner and more maintainable code blocks.

In the reduction phase, dimensionality reduction was performed through feature selection based on correlation analysis and variance thresholds. Features exhibiting high multicollinearity were identified using a correlation matrix and removed to prevent redundancy. Low-variance features that provided little variability across the dataset were also excluded, as they did not contribute useful information to the learning process. This step improved model interpretability, reduced overfitting risk, and enhanced computational efficiency.

The dataset was split into training (data from years ≤2020), validation (2021), and testing (2022–2024) subsets to maintain temporal integrity and prevent data leakage, ensuring that the models were evaluated on genuinely unseen future data.

3.4 Addressing Data Gaps via Bayesian SAE

One of the key challenges encountered in this study was the limited availability of municipal-level rice production data. While OpenSTAT provided long-term provincial-level data, and OPA supplied municipal-level records, the OPA dataset covered only 2017–2024. This gap for 2000–2016 at the municipal level made it necessary to reconstruct the missing data to create a continuous series suitable for model development.

To address this, the study applied Bayesian Small Area Estimation (SAE) [19]. The method was chosen because it enables reliable estimation in small geographic domains where observations are scarce, by combining information from multiple sources to generate statistically consistent predictions. Provincial-level agricultural data from OpenSTAT and environmental covariates from NASA POWER—including precipitation, temperature, relative humidity, UVB irradiance, and soil moisture—served as auxiliary inputs to improve estimation accuracy.

The estimation procedure followed a hierarchical Bayesian framework. In the first stage, municipal rice production was modeled as a function of the environmental and climatic variables. In the second stage, prior distributions were assigned to model parameters so that information could be shared across municipalities and cropping seasons. Markov Chain Monte Carlo (MCMC) methods were used to approximate the posterior distributions and to obtain consistent estimates of municipal rice production for the years 2000 to 2016. Through this process, the missing municipal-level production data were reconstructed, producing a complete and continuous dataset spanning 2000 to 2025. This enhanced dataset served as the foundation for developing the RF, GPR, and hybrid models.

producing models.

3.5 Model Development

The hybrid model framework was developed using Python within the Google Colab environment, utilizing standard data science libraries from scikit-learn for modeling, evaluation, and visualization. The modeling proceeded in two sequential stages.

Stage 1: Random Forest for Initial Prediction and Feature Analysis. The Random Forest model was implemented as the first stage of the hybrid framework for generating baseline predictions and identifying the most influential agronomic and environmental variables. The model was configured with 600 decision trees (n_estimators = 600) to ensure model stability and reduce variance. Each tree was allowed to grow to its maximum depth (max_depth = None), while a minimum of two samples per leaf (min_samples_leaf = 2) was applied to prevent overly specific splits. The random seed (random_state = 42) ensured reproducibility, and all available processing cores (n_jobs = -1) were utilized to optimize computation time.

The model was trained on the full feature set including harvested area, precipitation, soil moisture, temperature (maximum and minimum), relative humidity, UV radiation, cropping season, municipality, and year. Predictions were generated for the training, validation, and testing subsets. To interpret the model's behavior and identify key predictive factors, two complementary feature importance methods were employed: (1) impurity-based importance, derived from the mean decrease in variance across all decision tree splits, where variables with higher scores contributed more to reducing prediction error; and (2) permutation importance, computed on the validation dataset by randomly shuffling the values of each feature and observing the resulting decrease in model performance.

Stage 2: Gaussian Process Regression for Refinement and Uncertainty. The GPR model was employed to refine the predictions from the Random Forest and to quantify predictive uncertainty. It was trained on the residuals—the differences between actual rice production values and the RF-predicted values. The covariance structure was defined by a composite kernel:

$$k(x,x') = C \times \mathrm{RBF}(\ell) + \mathrm{WhiteKernel}(\sigma^2 n),$$

where:

C is a scaling constant that controls the overall variance,

ℓ is the length-scale parameter of the Radial Basis Function (RBF) kernel that controls the smoothness of the function,

$\sigma^2 n$ represents the residual noise captured by the White Noise kernel.

Kernel hyperparameters were optimized automatically with three restarts of the optimizer (n_restarts_optimizer=3) to avoid local optima. The target variable was standardized internally (normalize_y=True) to ensure numerical stability during training. Predictive uncertainty was quantified by generating 90% and 95% confidence intervals from the predictive standard deviations.

Hybrid Integration. The final hybrid prediction for any input x was computed as:

$$\hat{y}\_\mathrm{hybrid}(x) = \hat{y}\_\mathrm{RF}(x) + \hat{r}\_\mathrm{GPR}(x)$$

where:

$\hat{y}\_\mathrm{RF}(x)$ represents the Random Forest baseline output

$\hat{r}\_\mathrm{GPR}(x)$ is the residual correction estimated by the Gaussian Process

This two-stage approach allows the RF to capture the dominant nonlinear patterns in the data, while the GPR models the remaining residual structure and provides calibrated uncertainty estimates. The sequential approach was chosen over joint modeling because it preserves RF's computational efficiency while allowing GPR to focus specifically on the error patterns that RF cannot capture.

3.6 Evaluation Metrics

Model performance was quantified using the following standard regression metrics: Root Mean Square Error (RMSE), which measures the average magnitude of prediction error with higher penalties for larger deviations; Mean Absolute Error (MAE), which represents the average absolute difference between predicted and actual values; and the Coefficient of Determination (R^2), which expresses the proportion of variance in the observed data explained by the model. For models with probabilistic components (GPR and hybrid), Prediction Interval (PI) coverage was additionally computed to evaluate the reliability of uncertainty estimates, measuring how often the true value falls within the predicted confidence interval.

4. RESULTS AND DISCUSSION

4.1 Data Identification and Preparation

The data preparation phase resulted in a unified dataset combining agricultural production records from PSA and OPA with environmental covariates from NASA POWER. The Bayesian SAE procedure successfully reconstructed municipal-level production data for 2000–2016, producing a complete and continuous panel spanning 25 years across 20 municipalities and two cropping seasons. Standard Python libraries (Pandas, NumPy, scikit-learn) were used for all data manipulation, cleaning, and feature engineering. The final preprocessed dataset contained records organized by municipality, cropping season, and year, with features including harvested area, precipitation, soil moisture, temperature (max/min), relative humidity, UVB irradiance, and production volume as the target variable.

4.2 Random Forest Performance

The standalone RF model achieved high accuracy across all data splits. On the training set, it recorded an RMSE of 946 metric tons, MAE of 443 metric tons, and R^2 of 0.995. On the validation set, performance remained strong with RMSE of 3,571t, MAE of 1,694t, and R^2 of 0.950. On the test set (2022–2024), the model achieved RMSE of 5,555t, MAE of 2,737t, and R^2 of 0.898, indicating that the model effectively captured the dominant patterns in the data while maintaining acceptable generalization to unseen test samples.

Feature importance analysis using both impurity-based and permutation methods produced consistent patterns. The municipality of Candaba emerged as the top predictor, indicating a strong spatial influence on production variability. This is expected given Candaba's status as one of the largest rice-producing municipalities in Pampanga. The Wet Season (WS) and Dry Season (DS) categories followed as major contributors, reflecting the significant seasonal variation in rice production. Among continuous meteorological variables, relative humidity and soil moisture were identified as the most significant predictors, suggesting that climatic moisture conditions substantially affect rice productivity in Pampanga. Temperature and precipitation variables, while still relevant, contributed less to overall model performance. These findings align with Alebele et al. [24], who emphasized rainfall and temperature as influential factors for rice production, and support the feature selection decisions made during preprocessing.

4.3 GPR Performance

The standalone GPR model achieved RMSE of 4,721t, MAE of 2,679t, and R^2 of 0.927 on the test set—outperforming standalone RF on point-prediction accuracy. This improvement is consistent with findings by Martínez-Ferrer et al. [6], who demonstrated GPR's effectiveness in modeling nonlinear agricultural relationships. Critically, GPR additionally provides well-calibrated confidence intervals essential for risk-aware agricultural planning, a capability that standalone RF cannot offer. The 90% confidence interval achieved 96.8% coverage on training data and 91.2% on test data, while the 95% CI achieved 97.7% and 95.6% coverage respectively. These values indicate that the model's uncertainty quantification was well-calibrated—the actual coverage rates closely matched or exceeded the nominal confidence levels.

The average confidence interval widths increased across data splits: from 5302 metric tons on training to 7868t on validation and 12372t on test data at the 90% level. This pattern of increasing uncertainty for more recent years reflects greater variability in agricultural conditions during the 2022–2024 period, which is expected given evolving climate patterns and changing farming practices. The ability to quantify this increasing uncertainty is a key advantage of GPR over deterministic models like RF, as it allows decision-makers to understand not just the predicted production level but also the range of plausible outcomes. These results are consistent with findings by Martínez-Ferrer et al. [6], who demonstrated GPR's value in providing interpretable uncertainty estimates for crop yield forecasting.

4.4 Hybrid Model Performance

The hybrid RF→GPR model substantially outperformed both standalone models across all evaluation metrics. On the training set, it achieved RMSE of 581t, MAE of 228t, and R^2 of 0.998. On the validation set, RMSE was 627t, MAE was 315t, and R^2 was 0.998. Most importantly, on the test set (2022–2024), the hybrid model achieved RMSE of 1130t, MAE of 556t, and R^2 of 0.996. The 90% prediction interval coverage on the test set was 88.7%, confirming that the model's uncertainty intervals were well-calibrated and reliable.

Compared to the standalone models, the hybrid achieved an 80% reduction in RMSE on the test set (from 5,555t to 1,130t) and a 79% reduction in MAE (from 2,737t to 556t). The R^2 improved from 0.898 to 0.996, meaning the hybrid model explained 99.6% of the variance in rice production compared to 89.8% for standalone approaches. These improvements demonstrate the substantial value of the sequential two-stage architecture, where GPR's residual modeling captures patterns that RF alone cannot detect.

4.5 Comparative Analysis

Table 1 summarizes the comparative performance across all three models on the test set (2022–2024).

Table 1. Comparative Model Performance on Test Set

Model	RMSE (t)	MAE (t)	R^2
Random Forest	5,555	2,737	0.898
GPR	4,721	2,679	0.927
Hybrid RF-GPR	1130	556	0.996

The hybrid model consistently outperformed both standalone models across all data splits. The superior performance is attributed to the complementary nature of the two-stage architecture: RF captures the dominant nonlinear relationships between environmental and agronomic variables and rice production, effectively handling the high-dimensional feature space with categorical and continuous variables. GPR then models the residual errors left by RF, capturing fine-grained temporal and spatial patterns that the ensemble of decision trees cannot fully represent. Additionally, GPR provides probabilistic uncertainty quantification that standalone RF cannot offer, enabling confidence-aware forecasting.

These results are consistent with findings by Hariyani et al. [8], who demonstrated that hybrid ensemble approaches outperform standalone models in crop yield prediction, and extend the work of Lagrazon and Tan [7] by showing that GPR's refinement of RF predictions yields significantly greater accuracy than either model alone. The 88.7% coverage of the 90% prediction interval on the test set also confirms that the hybrid model maintains reliable uncertainty estimates even for out-of-sample predictions, which is essential for practical agricultural decision-making where stakeholders need to understand both the expected outcome and the associated risk. The performance difference between standalone RF (R^2 = 0.898) and GPR (R^2 = 0.927) reflects GPR's superior capacity to model nonlinear residual structure in the data. The hybrid architecture amplifies this advantage: by training GPR specifically on RF's residual errors, the model directs GPR's probabilistic strengths toward the patterns RF cannot capture, yielding an 80% RMSE reduction over either standalone approach. To further demonstrate the model's practical utility, three agricultural scenarios were simulated for Candaba municipality during the 2024 wet season. As shown in Table 2, severe drought conditions—combining a 50% precipitation reduction, decreased soil moisture, elevated temperatures, and a 20% contraction in harvestable area—produced a 22.4% decline in predicted production alongside a substantially wider confidence interval (±5,717t), reflecting heightened forecasting uncertainty under climate stress. Optimal conditions yielded a 5.1% production gain with a narrower interval, indicating greater predictive stability when environmental inputs are favorable. These results demonstrate the model's sensitivity to compound agro-climatic changes and its capacity to support scenario-based LGU planning.

Table 2. Scenario Analysis — Hybrid RF→GPR Predictions for Candaba, WS 2024

Scenario	Predicted (MT)	90% CI Low	90% CI High	Change
Baseline (normal)	17,467	16,411	18,523	—
Severe drought (−50% precip, −20% area)	13,559	7,842	19,275	−22.4%
Optimal conditions (+10% area)	18,362	16,874	19,849	+5.1%

4.6 Web Application Deployment

The hybrid model was integrated into a web-based decision-support application built using Streamlit, designed for local agricultural officers and planners. The application's primary goal is to make the complex prediction tool simple and useful for non-technical users. Instead of requiring specialized software or technical expertise, users can input key farming conditions—including area harvested, temperature ranges, total rainfall, and soil moisture—through interactive sidebar controls. Once the prediction runs, the application immediately generates a forecast for total rice production and yield per hectare.

The application provides five core decision-support functionalities: (1) Model Confidence, displaying 95% prediction intervals and standard deviation for production risk assessment; (2) Environmental Diagnosis, evaluating soil moisture, humidity, precipitation, and UVB radiation against optimal thresholds; (3) Feature Contribution, showing each input variable's positive or negative effect on the prediction; (4) Historical Comparison, benchmarking forecasts against municipal historical averages; and (5) Historical Trend, displaying yield trends by Dry and Wet seasons for contextual planning.

The system aligns with established decision support system (DSS) principles: it is designed to support rather than replace human decision-making [20], provides an interactive and user-friendly interface accessible to users without technical expertise [21], integrates historical data and analytical models to generate forecasts [22], and delivers timely, contextual information at the municipal level for planning decisions. The model's combined ability to identify production trends, assess climatic impacts, and forecast yields with quantified uncertainty provides actionable insights for strengthening Pampanga's agricultural resilience, enhancing resource efficiency, and supporting long-term food security initiatives.

4.7 Limitations

Several limitations of this study should be acknowledged. First, the Bayesian SAE-reconstructed data for 2000–2016, while statistically principled, introduces estimation uncertainty that may propagate into model training; the reconstructed values are modeled estimates rather than direct observations, and their accuracy depends on the quality of the auxiliary environmental covariates used. Second, the model does not account for unpredictable factors such as pest infestations, typhoons, or socio-political disruptions that can cause sudden and severe production losses outside of normal environmental variability. Third, economic factors and farm-level practices—including labor availability, fertilizer application rates, and market prices—are not modeled due to data unavailability, despite their known influence on production decisions. Fourth, additional variables such as vegetation indices (NDVI), soil nutrient composition, and fertilizer-related data were excluded from the study due to limited availability in the respective repositories; incorporating these variables could further improve prediction accuracy. Fifth, the model is tailored specifically for Pampanga and may require retraining and recalibration for use in other provinces with different agro-climatic conditions. Sixth, GPR's computational complexity scales cubically with dataset size ($O(n^3)$), which may present challenges for substantially larger datasets or real-time applications without approximation methods such as sparse GPR or variational inference. These limitations suggest important directions for future improvement while not diminishing the demonstrated effectiveness of the hybrid approach within its intended scope and application context.

5. CONCLUSION AND FUTURE WORK

This study developed A-Rice, a sequential hybrid RF→GPR model for municipal-level rice production prediction in Pampanga, Philippines. The model was trained on a comprehensive dataset spanning 20 municipalities over 25 years (2000–2025), integrating agricultural production data from PSA and OPA with six environmental variables from NASA POWER.

The hybrid model achieved R^2 of 0.996 on the test set and reduced RMSE by 80% compared to standalone models (from 5,555 to 1,130 metric tons), confirming that sequential integration of ensemble and probabilistic approaches yields a substantially more accurate and reliable forecasting system. Based on the findings, the following conclusions were drawn:

First, comprehensive and integrated datasets combining agricultural production data from PSA and OPA with meteorological data from NASA POWER can effectively support predictive modeling of rice production. The Bayesian SAE methods successfully addressed missing municipal-level data for 2000–2016, resulting in a complete and coherent data panel spanning 25 years.

Second, the Random Forest model provided robust baseline predictions and valuable insights into variable importance, revealing that relative humidity, soil moisture, municipality, and cropping season are the most influential factors affecting rice production in Pampanga.

Third, the Gaussian Process Regression model enhanced predictive precision by modeling residual errors and quantifying uncertainty through well-calibrated confidence intervals, which is crucial for risk-aware agricultural planning.

Fourth, the hybrid RF→GPR model achieved the best overall performance with R^2 of 0.996 on the test set and substantially reduced RMSE and MAE, confirming that the sequential integration of ensemble and probabilistic approaches yields a more accurate and stable forecasting system than either approach alone.

Fifth, the web-based application successfully made complex analytical results accessible to agricultural stakeholders through an interactive interface with model confidence visualization, environmental diagnostics, feature contributions, and historical benchmarking capabilities.

From a practical standpoint, the A-Rice system provides LGUs and provincial agricultural offices with a concrete tool for evidence-based resource allocation, production advisories, and seasonal planning. The model's uncertainty estimates enable stakeholders to prepare for best-case and worst-case production scenarios, rather than relying on single-point estimates that mask the inherent variability in agricultural production. This capability directly addresses the CL-RDP 2023–2028's call for modernized agricultural planning tools and contributes to strengthening food security in the region.

Future work should expand the model by incorporating additional predictors such as soil nutrient content, vegetation indices (NDVI), and fertilizer data to further enhance accuracy. Real-time data integration from weather stations and satellite feeds would enable continuous updating and near real-time forecasting. Deep learning approaches such as Long Short-Term Memory (LSTM) networks or Convolutional Neural Networks (CNNs) could be explored to capture temporal and spatial dependencies more effectively. Hyperparameter tuning methods such as grid search or Bayesian optimization should be applied to further optimize model performance. The framework should be validated in other provinces and crop types to assess scalability and generalizability. Finally, collaboration among agencies such as PSA, OPA, and PAGASA should be strengthened to establish a centralized database system for agricultural analytics and forecasting.

6. ACKNOWLEDGMENTS

The authors express their gratitude to their thesis adviser, Melissa M. Pantig, MCS, for her guidance throughout this research. Special thanks to the Philippine Statistics Authority, the Office of the Provincial Agriculturist of Pampanga, the Philippine Rice Research Institute (PhilRice), and the National Economic and Development Authority (NEDA) for providing access to essential datasets and domain expertise. The authors also acknowledge the College of Computer Studies at Angeles University Foundation for institutional support.

7. REFERENCES

[1] Philippine Statistics Authority. 2023. OpenSTAT: Philippine Statistical Database. https://openstat.psa.gov.ph/.

[2] National Economic and Development Authority Regional Office III. 2023. Central Luzon Regional Development Plan 2023–2028. City of San Fernando, Pampanga.

[3] Wolanin, A., Mateo-García, G., Camps-Valls, G., et al. 2020. Estimating and Understanding Crop Yields with Explainable Deep Learning in the Indian Wheat Belt. Environmental Research Letters, 15(024019), 1-12.

[4] Basha, S. M., Rajput, D., et al. 2020. Crop Yield Prediction Using Random Forest. Scalable Computing: Practice and Experience, 21(4), 593–602.

[5] Ghosh, S. S., Dey, S., et al. 2022. Gaussian Process Regression Model for Crop Biophysical Parameter Retrieval from Multi-Polarized C-Band SAR Data. Remote Sensing, 14(4), 934-960.

[6] Martínez-Ferrer, L., Piles, M., and Camps-Valls, G. 2021. Crop Yield Estimation and Interpretability with Gaussian Processes. IEEE Geoscience and Remote Sensing Letters, 18(12), 2043-2047.

[7] Lagrazon, P. G. and Tan, J. 2024. Predicting Crop Yield in Quezon Province, Philippines Using Gaussian Process Regression. 2023 International Conference on Modeling & E-Information Research, 6-12.

[8] Hariyani, G., Karad, V., et al. 2024. Analysis on Crop Yield Prediction using Various Ensemble Methods. 8th International Conference on Computing, Communication, Control and Automation, 1-6.

[9] Jabed, Md. A. and Azmi Murad, M. A. 2024. Crop Yield Prediction in Agriculture: A Comprehensive Review of Machine Learning and Deep Learning Approaches. Heliyon, 10(24), e40836.

[10] Rashid, M., Bari, B. S., et al. 2021. A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches. IEEE Access, 9, 63406–63439.

[11] Shawon, S., Ema, F., et al. 2023. Crop Yield Prediction: Robust Machine Learning Approaches for Precision Agriculture. 26th International Conference on Computer and Information Technology, 1-6.

[12] Zhou, W., et al. 2022. Random Forest-based Rice Yield Prediction in Mountainous Regions. Agricultural Systems.

[13] Parreño, S. J. E. and Anter, M. C. J. 2024. New Approach for Forecasting Rice and Corn Production in the Philippines Through Machine Learning Models. Multidisciplinary Science Journal, 6(9), 2024168.

[14] Choudhary, K., et al. 2022. Random Forest for Rice Yield Prediction Using Sentinel-2 Imagery and Environmental Variables.

[15] Köse, U., et al. 2023. Random Forest Outperformance in Rice Yield Prediction Using Climatic and Historical Data.

[16] Liu, F., Su, L., et al. 2024. Prediction Models of Growth Characteristics and Yield for Chinese Winter Wheat Based on Machine Learning. Agronomy, 14, 839–856.

[17] Ekanayake, P. and Wickramasinghe, L. 2022. GPR-based Paddy Yield Prediction Using Climate Variables in Sri Lanka.

[18] Hassan, M. A., et al. 2025. GPR Effectiveness in Mechanized Rice Farming Yield Forecasting.

[19] Ghosh, M. and Rao, J. N. K. 1994. Small Area Estimation: An Appraisal. Statistical Science, 9(1), 55–93.

[20] Keen, P. G. W. and Scott Morton, M. S. 1978. Decision Support Systems: An Organizational Perspective. Addison-Wesley.

[21] Power, D. J. 2002. Decision Support Systems: Concepts and Resources for Managers. Quorum Books.

[22] Sprague, R. H. and Carlson, E. D. 1982. Building Effective Decision Support Systems. Prentice Hall.

[23] Mabunga, Z. and Dela Cruz, J. 2022. An Optimized Soil Moisture Prediction Model for Smart Agriculture Using Gaussian Process Regression. IEEE 18th International Colloquium on Signal Processing & Applications, 243-247.

[24] Alebele, Y., Wang, W., et al. 2021. Estimation of Crop Yield From Combined Optical and SAR Imagery Using Gaussian Kernel Regression. IEEE JSTARS, 14, 10520–10534.