Estimation of Rice Productivity in South Sulawesi, Indonesia Using Meteorological Data

Masayuki Matsuoka[1]*, Nur Azizah[1]

[1]*Mie University,1577 Kurima-machiya, Tsu, Mie, Japan. matsuoka@info.mie-u.ac.jp*

Abstract: *Rice is the staple food in Indonesia and accurate prediction of its productivity is crucial for ensuring food security, especially in South Sulawesi where demand is rising and yield fluctuations often increase reliance on imports. This study investigates the potential of meteorological data (temperature, precipitation, and solar radiation) for predicting rice productivity at the district level using machine learning approaches. Data were obtained from ERA5-Land reanalysis and CHIRPS precipitation datasets for the period 2018–2024, while district-level rice productivity statistics were collected from official reports. Preprocessing involved cleaning, normalization, and feature extraction from monthly climatic variables. Two modeling strategies were explored: (1) per-month regression to evaluate the contribution of individual months, and (2) multi-month modeling using Random Forest and Gradient Boosting. Monthly RMSE analysis revealed variability in model accuracy across the calendar year, with April and November achieving the lowest normalized RMSE (~5%), while May and December exhibited the highest errors (~9%). These findings highlight the seasonal sensitivity of rice yields to climatic conditions. When advanced models were applied, Random Forest consistently outperformed both Linear Regression and Gradient Boosting. Using all months as predictors, Random Forest achieved an RMSE of ~0.6 and R^2 of ~0.99, substantially better than Gradient Boosting (RMSE ~2.4, R^2 ~0.80) and Linear Regression (RMSE ~5.3, R^2 ~0.0).*

This research demonstrates that meteorological variables derived from reanalysis datasets can reliably predict rice productivity. The results emphasize the potential of Random Forest for operational yield forecasting in South Sulawesi, supporting early warning systems and agricultural planning. Future work should test region-based cross-validation to mitigate possible overfitting and integrate additional agronomic variables such as soil type and land use.

Keywords: *Agriculture, Machine Learning, Meteorology, Rice, Yield Estimation*

1. Introduction

Rice is the main staple food for Indonesia, where demand continues to rise with population growth. South Sulawesi is one of the country's key rice-producing provinces especially in Sulawesi Island, yet fluctuations in climate and inconsistent yields have increased reliance on imports. Accurately forecasting rice productivity is crucial for ensuring food security and guiding policy decisions. Advances in remote sensing and machine learning provide opportunities to integrate meteorological data into yield estimation models. This paper investigates how meteorological variables such as temperature, precipitation, and solar radiation can be used to predict rice productivity using different machine learning approaches.

2. Methodology

Meteorological data were collected from ERA5-Land (temperature and solar radiation) and CHIRPS (precipitation) for the period 2018–2024. District-level rice productivity statistics were obtained from official agricultural reports. Data preprocessing included cleaning string-formatted solar values, converting them into floats, and aligning meteorological variables with productivity records.

Two approaches were implemented:
Per-month RMSE analysis: The dataset consists of 42 records. Each record includes 36 predictors (12 months × 3 variables) representing temperature, precipitation, and solar radiation, along with the observed rice productivity value. Linear Regression was applied separately to each month’s meteorological variables to evaluate the monthly contribution to productivity. RMSE and R^2 was calculated to compare model performance across months.

Multi-month modeling: All 36 predictors (12 months × 3 variables) were used as features in Random Forest, Gradient Boosting, and Linear Regression models. Cross-validation was performed to estimate RMSE and R^2 values. Feature importance analysis was used to identify key climatic drivers.

3. Results/Findings

The per-month RMSE analysis showed clear seasonal variability. April and November had the lowest nRMSE (~5%), while May and December showed the highest errors (~9%) as shown in Figure 1. This suggests certain months provide more predictive information about rice yields than others. April and November showed the lowest RMSE, partly due to more consistent relationships between meteorological variables and productivity.In contrast, May and December were less predictable, possibly due to transitional climatic condition.

For the multi-month modeling, Random Forest significantly outperformed other methods, achieving an RMSE of ~0.6 and R^2 of ~0.99 . Gradient Boosting performed moderately well (RMSE ~2.4, R^2 ~0.80), while Linear Regression performed poorly (RMSE ~5.3, R^2 ~0). These results confirm the advantage of non-linear ensemble models for capturing complex climate-yield relationships. Feature importance analysis indicated that solar radiation (April, October) and precipitation variables were particularly influential, aligning with the crop’s critical growth phases.

Although Random Forest showed very high accuracy, there is potential risk of overfitting given the limited dataset. Future validation using region- or year-based splits is recommended. Nevertheless, the results highlight the feasibility of using meteorological predictors and machine learning for yield forecasting in South Sulawesi.

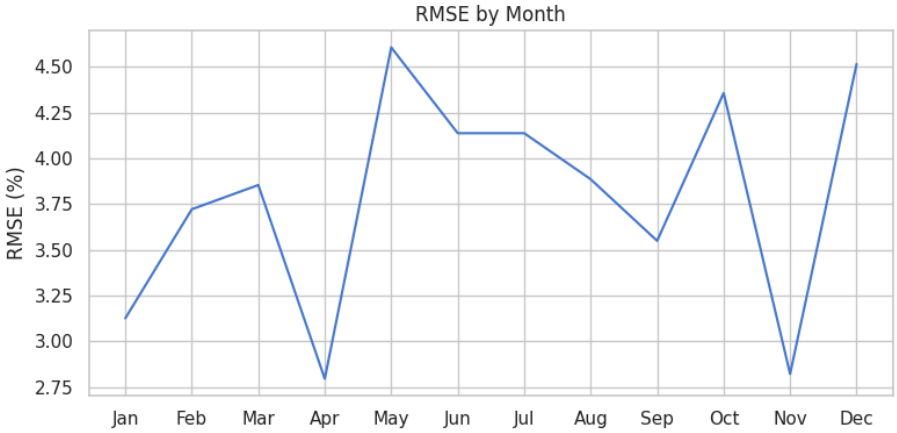

Figure 1: RMSE across months

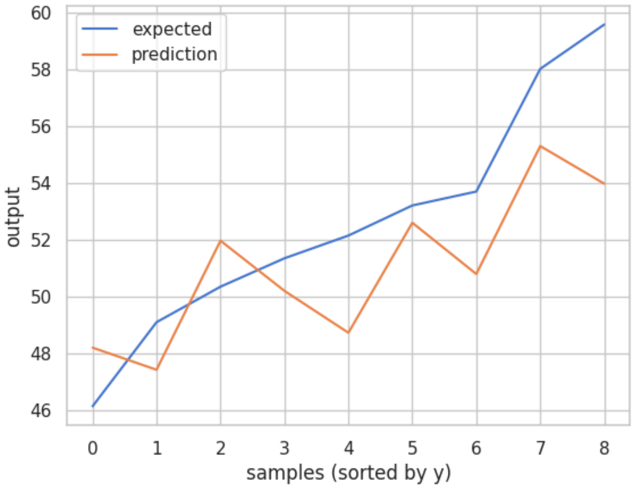

Figure 2: April Prediction on Productivity

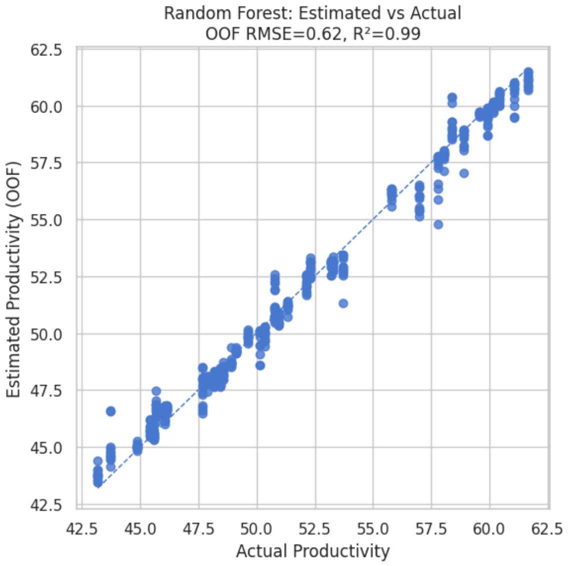

Figure 3: Random Forest Prediction on Productivity

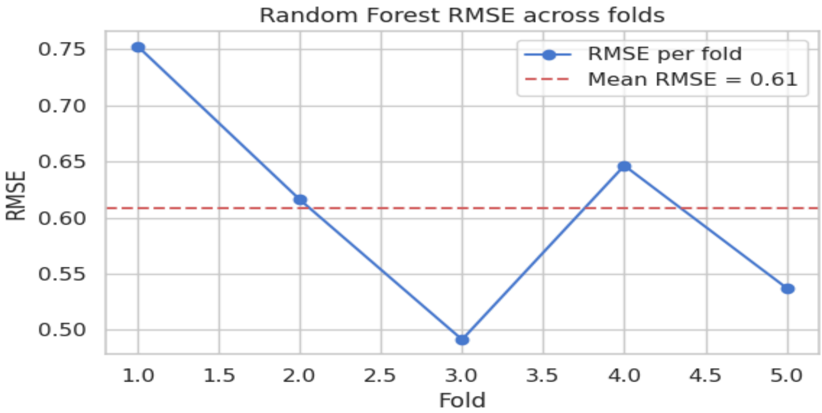

Figure 4: Random Forest RMSE across all variables

Table 1: Comparison of machine learning model based on RMSE and R²

Model	RMSE (Mean ± std)	R² (Mean ± std)
Multiple Linear Regression	5.309 ± 0.291	-0.006 ± 0.036
Random Forest	0.609 ± 0.091	0.986 ± 0.005
Gradient Boost	2.369 ± 0.055	0.799 ± 0.014

4. Conclusion

This study demonstrates that rice productivity in South Sulawesi can be effectively estimated using meteorological data and machine learning models. Monthly RMSE analysis provided insight into seasonal influences, with April and November emerging as the most predictive months. Random Forest achieved the highest accuracy, suggesting its suitability for operational yield forecasting. Future research should expand the dataset, incorporate additional agronomic variables such as soil type, fertilizer, paddy area and cropping calendars, and apply robust validation techniques to strengthen generalizability. These improvements would support the development of early warning systems and more resilient agricultural planning in Indonesia.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP21K05669, JP22H05004, and JP20K20487

References

ECMWF. (2019, June 23). ERA5-Land monthly averaged data from 1950 to present. https://cds.climate.copernicus.eu/doi/10.24381/cds.68d2bb30

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., & Hoell, A. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data, 2(1), 1–21. https://doi.org/doi:10.1038/sdata.2015.66