Brief Report

Not peer-reviewed version

Wheat Yield Prediction Based on Random Forest Method

Yared Semahegn *

Posted Date: 25 March 2026

doi: 10.20944/preprints202502.1837.v2

Keywords: crop yield prediction; sustainability; machine learning; random forest; wheat

Brief Report

Wheat Yield Prediction Based on Random Forest Method

Yared Semahegn

Holetta Agricultural Research Center, Holetta, Ethiopia; yaredsemahegnb@gmail.com

Abstract

Crop yield prediction is essential for enhancing agricultural productivity, managing risks, ensuring food security, and improving the sustainability of farming systems. This study aimed to evaluate the effectiveness of Random Forest (RF) regression for predicting wheat yield using a time-series dataset comprising climate- and soil-related variables, and to identify the key factors influencing wheat yield. The performance of RF was compared with Multiple Linear Regression (MLR) as a benchmark model. The results showed that RF outperformed MLR in predicting wheat yield. Model performance was evaluated using mean absolute error (MAE) and root mean squared error (RMSE). The RF model achieved an MAE of 135.88 and an RMSE of 163.90, whereas the MLR model produced substantially higher errors, with an MAE of 435.74 and an RMSE of 653.39. Variable importance analysis from the RF model indicated that year and CO_2 emission were the most influential predictors. Nitrogen fertilizer use, wheat cultivated area, and phosphate fertilizer application were also associated with improved wheat yield. The partial dependence plot for year revealed an increasing trend in wheat yield from 2000 to 2010, followed by a yield plateau after 2010. Overall, these findings demonstrate that RF provides superior predictive performance compared with MLR and represents a robust tool for wheat yield forecasting and identifying key drivers of yield variation.

Keywords: crop yield prediction; sustainability; machine learning; random forest; wheat

---

1. Introduction

Crop yield prediction is essential for enhancing productivity, managing risks, ensuring food security, and improving the overall sustainability of agriculture [2]. Accurate yield predictions help farmers allocate resources such as water, fertilizers, and labor more efficiently. Predicting crop yields in advance also aids in the planning of marketing and distribution, allowing stakeholders from farmers to food processors and retailers, to prepare for demand and supply fluctuations, reducing waste and ensuring food availability [1,9]. Additionally, weather patterns, pests, and diseases can greatly affect crop yields. By predicting yields, farmers can take proactive measures, such as adjusting planting schedules or implementing pest control strategies. Governments and organizations also rely on crop yield predictions to make informed decisions regarding agricultural policy, subsidies, and crop insurance, ensuring that farmers receive adequate support during poor harvests or unforeseen challenges [2].

Crop yield prediction is crucial for food security. As climate change alters growing conditions, crop yield predictions can help farmers adapt by identifying potential changes in optimal planting times or locations [10]. This information allows for the modification of agricultural practices to maintain high productivity in a changing climate. Yield predictions also support farmers and agribusinesses in making investment decisions, such as whether to plant a particular crop or invest in new technology [10,11]. Furthermore, it aids in assessing the potential profitability of different crops.

Wheat production plays a significant role in Ethiopian agriculture, contributing to both food security and the economy. It is a key staple, particularly in urban areas where it is consumed in the form of bread, pasta, and other processed foods. With urbanization and changing dietary patterns,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202502.1837.v2

the demand for wheat has been rising [8]. Wheat production in Ethiopia is concentrated in the highland areas, which offer favorable climates with cooler temperatures and fertile soil, essential for wheat farming. However, Ethiopia's wheat production is vulnerable to the impacts of climate change, such as unpredictable rainfall patterns and rising temperatures. Additionally, pests like wheat rusts and various diseases can severely affect yields if not properly managed [8]. Unsustainable farming practices, such as soil erosion and degradation, further threaten wheat yields.

Various process-based and statistical models have been used to predict crop yields. Although process-based models are valuable, their calibration and validation can be challenging under certain conditions [5,6,12]. Statistical models, while easier to interpret and able to work with smaller datasets, often struggle to capture complex relationships, especially those involving non-linearities, interactions, or high-dimensional data. They also require assumptions about data distribution and the relationship between variables, which may not always hold true in real-world scenarios [7]. To address these limitations, machine learning approaches offer a promising alternative. Machine learning, a branch of artificial intelligence, has gained importance in crop yield prediction. Machine learning algorithms, such as decision trees, support vector machines, ensemble methods, and neural networks, can handle non-linear relationships and complex interactions between features that traditional statistical models might miss [6,9,10]. When provided with large datasets, machine learning models often outperform traditional statistical models in terms of prediction accuracy. These models can continuously improve as more data is introduced, allowing them to adapt to changing conditions, such as climate change and evolving farming techniques [9]. The present study aimed to evaluate the effectiveness of Random Forest (RF) regression for predicting wheat yield using a time-series dataset comprising climate- and soil-related variables, and to identify the key factors influencing wheat yield.

2. Materials and Methods

2.1. Dataset

The data used in this analysis were obtained from various sources, covering the period from 1980 to 2018, for model training and testing. Data on wheat yield, area coverage, and fertilizer use (nitrogen and phosphate) were sourced from the FAO database (https://www.fao.org), freely accessed on December 20, 2024. Climate data were obtained from the National Meteorological Agency of Ethiopia, while CO2 emissions (in kilotons) for the specified period were retrieved from the World Bank database (https://data.worldbank.org), also freely accessed in December 2024. These sources were combined to create a dataset containing 1,092 data points for the present analysis. Wheat yield was treated as the target variable, with the other variables used as predictors (Table 1). The analysis incorporated five climatic variables, monthly maximum temperature, monthly minimum temperature, monthly rainfall, monthly relative humidity, and CO_2, as well as three soil-related variables: nitrogen and phosphate fertilization uses, and crop harvested area. For the climate variables, data from the growing season (June to November) were considered except for CO_2 which is on annual bases. Fertilizer use for each nutrient was adjusted for specific wheat crop land. Additionally, year was included as a predictor to account for temporal variability due to technological advances, such as improvements in genetics and management practices [9].

Table 1. Target variable and features considered in the present dataset.

Name	Description
Year	The period from 1980-2018
Wheat yield (kg/ha)	Wheat production divided by wheat harvested area
Wheat area (ha)	Amount of crop land allocated to wheat production

 doi:10.20944/preprints202502.1837.v2

Name	Description
Monthly maximum temperature (°c)	Average of daily maximum temperatures registered in the crop growing season (June to November)
Monthly minimum temperature (°c)	Average of daily minimum temperatures registered in the crop growing season (June to November)
Monthly rainfall (mm)	Total amount of daily rainfall registered in the crop growing season (June to November)
Monthly relative humidity (%)	Average of daily relative humidity registered in the crop growing season (June to November)
Nitrogen use (kg)	The total nitrogen use of wheat crop land
Phosphate (P_2O_5) (kg)	The total phosphate ((P_2O_5) use of wheat crop land
CO_2 emission (in kilotons)	Annual CO2 emission

2.2. Data Preprocessing

Data preprocessing, along with the subsequent tasks of training and testing the models, was performed using relevant machine learning libraries in Python version 3.12. Missing values in the dataset were imputed using the mean. The data were then split into training and testing sets, with 80% of the data used for training and the remaining 20% allocated for testing.

2.3. Machine Learning Techniques

Random Forest (RF)

A random forest (RF) regression model was used to predict wheat yield. RF is an ensemble machine learning method that combines multiple decision trees to enhance prediction performance and accuracy. Each tree is trained on a random subset of the data, and predictions are made by aggregating the results from all the trees through averaging. This approach is robust against overfitting [3]. Feature importance and partial dependence plots were created to assess how much the model relies on a particular feature when making predictions. The partial dependence plots illustrate how the model's predictions change with respect to the most important feature, while keeping all other features constant [4].

Multiple Linear Regression (MLR)

Multiple linear regression (MLR) was used as a benchmark model for predicting wheat yield. MLR is a statistical technique that models the relationship between input features and a continuous target variable by fitting a linear equation to the observed data [6]. MLR is one of the most commonly used statistical models for predicting crop yield [6,9,11]. The same dataset used for the RF models was employed to construct the MLR model. The details of the dataset and definitions of the variables used in the study are presented in Table 1.

2.4. Evaluation Metrics

The models made predictions on the test set, and their performance was evaluated using mean absolute error (MAE) and root mean squared error (RMSE). These are common evaluation metrics for crop prediction, with lower values indicating better model performance [7]. The metrics were calculated as follows:

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

 doi:10.20944/preprints202502.1837.v2

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

Where, n is the number of observations; y_i is the observed value of the target variable; $\hat{y}_i$ is the predicted value of the target variable, and $y_i - \hat{y}_i$ is an error for each observation

3. Results and Discussion

The results demonstrate that the Random Forest (RF) model is an effective tool for predicting wheat yield and outperforms the Multiple Linear Regression (MLR) model. The RF model achieved a Mean Absolute Error (MAE) of 135.88 and a Root Mean Square Error (RMSE) of 163.90 when predicting wheat yield over a 39-year period using the test dataset. In contrast, the MLR model showed substantially poorer performance, with an MAE of 435.74 and an RMSE of 653.39 (Table 2). The comparison between observed and predicted values further indicates that the RF model closely tracked the actual wheat yield values (Table 3 and Figure 1), whereas the MLR model exhibited larger deviations and less consistent predictions (Table 4 and Figure 2). Since both models were trained and tested using the same dataset, the improved performance of RF can be attributed to its ability to capture nonlinear relationships and complex interactions among predictors. Similar superior performance of RF over traditional regression approaches has also been reported in previous studies [9,10,13].

Table 2. Random forest and Multiple regression models performance. .

Model	Metrics	
	MAE	RMSE
Random forest	135.879	163.898
Multiple linear regression	435.741	653.394

Table 3. Observed and predicted wheat yield for RF model. .

Year	Observed (kg/ha)	Predicted (kg/ha)
2013	2444.6	2268.410
2016	2675.5	2594.297
1984	1064.6	1115.205
1993	1548.7	1228.717
2010	1838.5	2018.509
2006	1520.4	1577.183
1986	996.0	1184.546
2007	1671.1	1637.389

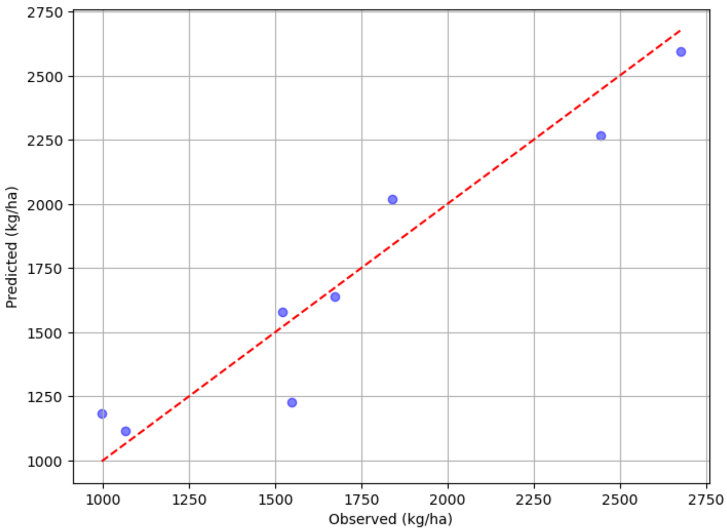

Figure 1. Graphical view of observed and predicted yield of wheat for RF model.

Table 4. Observed and predicted wheat yield for MLR model.

Year	Observed (kg/ha)	Predicted (kg/ha)
2013	2444.6	2599.199
2016	2675.5	2156.834
1984	1064.6	1133.232
1993	1548.7	1510.493
2010	1838.5	2193.276
2006	1520.4	3161.813
1986	996.0	492.161
2007	1671.1	1465.302

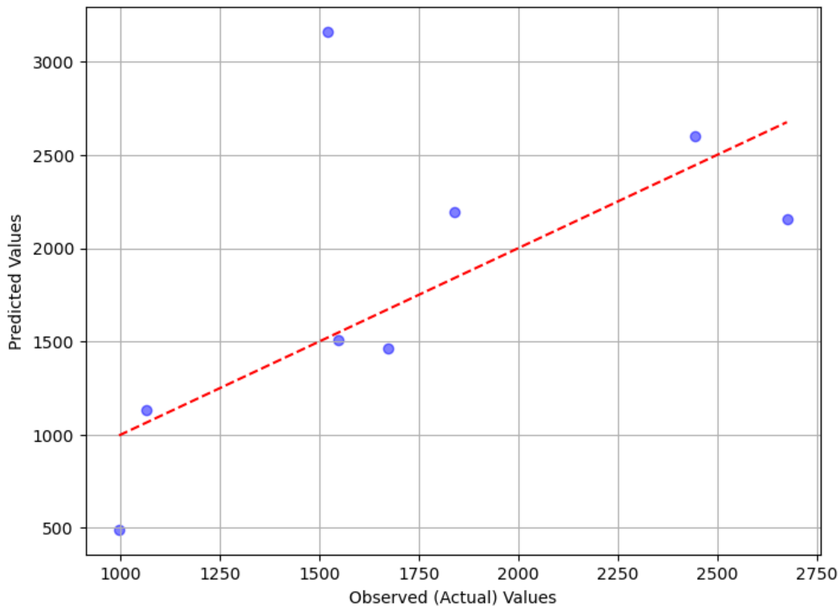

Figure 2. Graphical view of observed and predicted yield of wheat for MLR model.

 doi:10.20944/preprints202502.1837.v2

Variable importance analysis from the RF model identified year and CO_2 emission as the most influential predictors (Figure 3). Furthermore, nitrogen fertilizer use, wheat cultivated area, and phosphate fertilizer application were also associated with improved wheat yield. The dominance of the year variable likely reflects the cumulative effects of technological progress, improved cultivars, agronomic management, and policy changes over time. These findings suggest that both input intensification and broader environmental factors contribute to yield improvement. However, because cultivated land is a limited resource, future yield gains will likely depend more on genetic improvement, optimized nutrient management, and improved agronomic practices rather than area expansion. Therefore, integrating improved crop varieties with efficient fertilizer use and modern production technologies could enhance wheat productivity while supporting sustainable agricultural development [9,13]. The partial dependence plot for year, which isolates its marginal effect, showed an increasing trend in wheat yield from 2000 to 2010, followed by a yield plateau after 2010 (Figure 4). This pattern may indicate that gains from technological improvements and management practices slowed during the later period, highlighting the need for renewed innovations to sustain yield growth. Similar temporal trends influenced by modernization and technological advancement have also been reported in earlier studies [10,13].

JunTmax-NovTmax: Maximum temperature (°c) for the months of June to November; JunTmin-NovTmin: Minimum temperature for the months of June to November (°c); JunRF-NovRF: Total amount of rainfall (mm) for the months of June to November ; JunRH-NovRH: Relative humidity (%) for the months of June to November; Nitro: Nitrogen fertilizer use (kg); Phosp: Phospahte fertilizer use (kg); CO_2 : CO2 emission (kt)

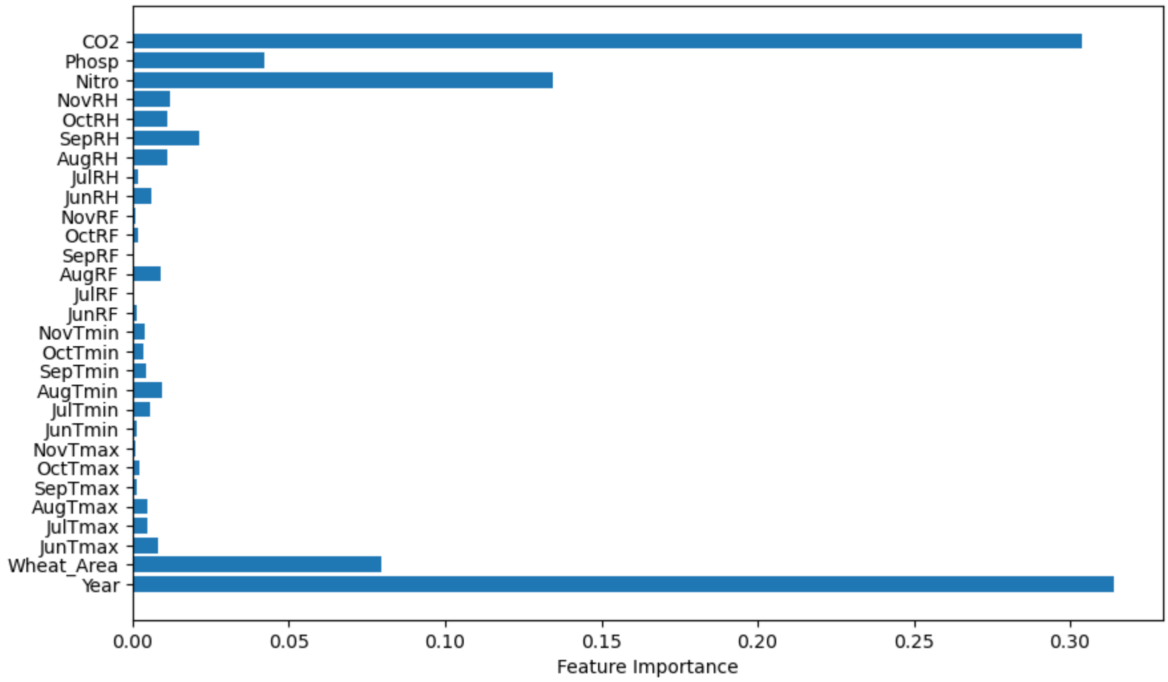

Figure 3. Feature importance measures of random forest model. .

doi:10.20944/preprints202502.1837.v2

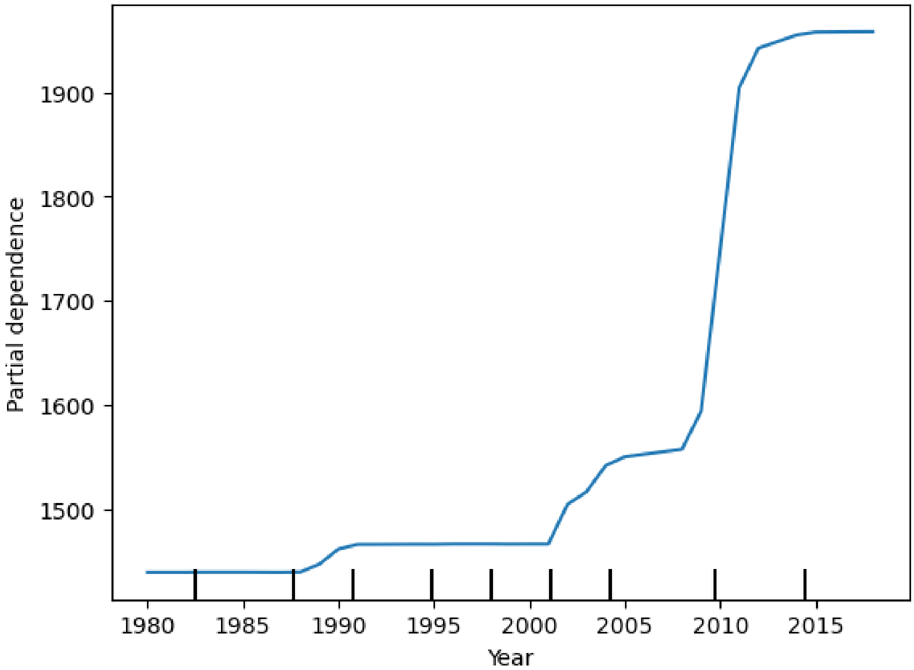

Figure 4. Partial dependence plot for the top ranked predictor feature or variable. .

4. Conclusion

This study evaluated the effectiveness of the Random Forest (RF) regression model, using Multiple Linear Regression (MLR) as a benchmark, for modeling the complex responses of wheat yield. The results demonstrated that RF outperformed MLR in predictive accuracy, indicating its superior ability to capture nonlinear relationships and interactions among predictors. These findings suggest that RF is a robust tool for wheat yield prediction, particularly when a wide range of explanatory variables is included. The variable importance analysis identified year and CO_2 emission as the most influential predictors. Nitrogen fertilizer application, expansion of wheat cropped area, and phosphate fertilizer use were also associated with higher wheat yield levels. The partial dependence plot for year, while accounting for the effects of other variables, revealed an increasing trend in wheat yield from 2000 to 2010, followed by a plateau thereafter. This pattern may reflect technological improvements during the early period and a subsequent stabilization in productivity gains. Overall, this study provides preliminary evidence supporting the use of RF for modeling wheat yield. Future research should incorporate larger datasets, additional environmental and management variables, and multi-region data to further improve model robustness and predictive performance.

Authors' contribution: Sole author.

Funding: This study received no external funding.

Disclosure statement: No conflict of interest.

References

1. M.F Aslan, K. Sabanci, B. Aslan, Artificial intelligence techniques in crop yield estimation based on Sentinel-2 data: A comprehensive survey, Sustainability, 16 (2024), 8277.
2. B. Basso, D. Cammarano, E. Carfagna, Review of crop yield forecasting methods and early warning systems. In Proceedings of the first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics, FAO Headquarters, Rome, Italy, 2013.
3. L. Breiman, Random forests. Machine learning, 45 (2001), 5-32.

 doi:10.20944/preprints202502.1837.v2

4. L. Breiman, Manual on setting up, using, and understanding random forests. Berkeley: Statistics Department University of California Berkeley, 2002.
5. M.S Dhillon, T. Dahms, C. Kuebert-Flock, T. Rummler, J. Arnault, I. Steffan-Dewenter, T. Ullmann, Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape, Frontiers in Remote Sensing, 3 (2023), 1010978.
6. A. González Sánchez, J. Frausto Solís, W. Ojeda Bustamante, Predictive ability of machine learning methods for massive crop yield prediction, 2014.
7. S.A. Haider, S.R. Naqvi, T. Akram, G.A. Umar, A. Shahzad, M.R, Sial…, M. Kamran, LSTM neural network-based forecasting model for wheat production in Pakistan, Agronomy, 9 (2) (2019), 72.
8. D.P. Hodson, M. Jaleta, K. Tesfaye, C. Yirga, H. Beyene, A. Kilian..., O. Erenstein, Ethiopia's transforming wheat landscape: tracking variety use through DNA fingerprinting, Scientific reports, 10 (1) (2020), 18532.
9. J.H. Jeong, J.P. Resop, N.D. Mueller, D.H. Fleisher, K. Yun, E.E Butler..., S.H. Kim, Random forests for global and regional crop yield predictions, PloS one, 11(6) (2016), e0156571.
10. K. Jhajharia, P. Mathur, S. Jain, S. Nijhawan, Crop yield prediction using machine learning and deep learning techniques, Procedia Computer Science, 218 (2023), 406-417.
11. T. Van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: A systematic literature review, Computers and electronics in agriculture, 177 (2020), 105709.
12. K. Matsumura, C.F. Gaitan, K. Sugimoto, A.J. Cannon, W.W. Hsieh, Maize yield forecasting by linear regression and artificial neural networks in Jilin, China, The Journal of Agricultural Science, 153 (3) (2015), 399-410.
13. Y.E. Roell, A. Beucher, P.G. Møller, M.B., Greve, M.H. Greve, Comparing a random forest-based prediction of winter wheat yield to historical yield potential, Agronomy, 10 (3) (2020), 395.