ICCAUA Proceedings Journal
Proceedings of the international conference of contemporary affairs in architecture and urbanism-ICCAUA
Volume 9 (December 2026), 2610149

Journal homepage: https://journal.iccaua.com/

DOI: https://doi.org/10.38027/ICCAUA2026EN0149

Enhancing Crop Yield Prediction Using Machine Learning and Geospatial Data

[1] Izukanji Muuka , [2] Kasongo Changwe , [3] Musoka Nyongolo , [4] Penjani Hopkins Nyimbili , [5] Masauso Sakala , [6] Erastus M. Mwanaumo , [7] Wellington D. Thwala
[1, 2, 3, 4 & 5] Department of Geomatic Engineering, School of Engineering, University of Zambia, Zambia
[6] Department of Civil and Environmental Engineering, School of Engineering, University of Zambia, Zambia
[3, 4 & 6] National Centre for Uncrewed Aircraft Systems (NACUAS), University of Zambia, Zambia
[4, 6 & 7] Built Environment and Information Technology, Faculty of Engineering, Walter Sisulu University, South Africa
[1] E-mail: izukanjimuuka19@gmail.com, [2] E-mail: kasongochangwe01@gmail.com, [3] E-mail: musokanyongolo@gmail.com,
[4] E-mail: penjani.nyimbili@unza.ac.zm, [5] E-mail: masauso.sakala@unza.ac.zm, [6] E-mail: erastus.mwanaumo@unza.ac.zm,
[7] E-mail: wdthwala@wsu.ac.za
[1] ORCID: https://orcid.org/0009-0009-1087-6763 , [2] ORCID: https://orcid.org/0009-0009-1087-6763 , [3] ORCID: https://orcid.org/0009-0008-6732-0654 , [4] ORCID: https://orcid.org/0000-0001-8271-5269 , [5] ORCID: https://orcid.org/0009-0008-1679-4569 , [6] ORCID: https://orcid.org/0000-0002-2911-3207 , [7] ORCID: https://orcid.org/0000-0002-8848-7823

Received: 24.04.2026
Revised: 22.06.2026
Accepted: 01.07.2026
Available online: 10.07.2026

This article has been selected and peer-reviewed for publication in this journal as part of the 9th International Conference of Contemporary Affairs in Architecture and Urbanism, held on 7–8 May 2026 in Istanbul, Türkiye.

Abstract

This study presents a geospatially informed machine learning approach to improve crop yield prediction in Zambia, where agriculture underpins rural livelihoods and national food security. The research integrates satellite-derived vegetation indices, principally the Normalised Difference Vegetation Index (NDVI), with meteorological indicators including seasonal rainfall distribution and temperature trends, across a 25-year wheat yield record (1999 to 2024) for a commercial farm in Chongwe District, Zambia. Datasets were harmonised through spatial standardisation, feature engineering, and temporal aggregation to produce a coherent input structure for a Random Forest regression model benchmarked against Extreme Gradient Boosting (XGBoost). Rainfall frequency, seasonal thermal accumulation, and vegetation vigour emerged as the most influential predictors of yield variability. Random Forest achieved a stable coefficient of determination (R2 = 0.389), while XGBoost exhibited apparent superiority (R2 = 0.950) attributable to overfitting on a small, spatially homogeneous dataset. The findings demonstrate that model selection is critical in small-sample agricultural prediction contexts and contribute an interactive, field-ready decision-support dashboard for precision agriculture in Zambia.

Keywords: Machine Learning; Crop Yield Prediction; NDVI; Remote Sensing; Precision Agriculture.

1. Introduction

Accurate crop yield prediction is a cornerstone of sustainable agricultural development, particularly in countries such as Zambia, where agriculture remains a primary driver of economic growth and food security. With a substantial share of the national population dependent on agriculture for livelihoods, the ability to forecast crop yields is essential for ensuring food availability, guiding investment, informing policy, and managing risks associated with climate variability (Mbow et al., 2021). Despite this importance, the integration of modern data-driven forecasting tools into Zambia's agricultural decision-making remains limited, constraining the sector's capacity to respond proactively to climatic shocks.

Historically, yield prediction approaches in Zambia and across the wider southern African region have relied on simple linear statistical models based on historical yields and basic weather summaries. These approaches frequently fail to capture the complex, non-linear interactions among environmental factors, including rainfall variability, temperature extremes, and soil characteristics, as well as farm management practices and localised climate anomalies (Siatwiinda et al., 2021; Ngoma et al., 2022). Consequently, predictions derived from such models are often unreliable, contributing to suboptimal planning, reduced productivity, and heightened vulnerability to climate shocks among both commercial and smallholder producers.

Recent advances in machine learning and Earth observation have opened new opportunities to address these limitations. Ensemble learning methods, particularly Random Forest and Extreme Gradient Boosting (XGBoost), have demonstrated strong capacity to model non-linear relationships between environmental predictors and crop yield outcomes across diverse agro-ecological settings (van Klompenburg et al., 2020; Zhou et al., 2022; Dhillon et al., 2023). When combined with satellite-derived vegetation indices such as the Normalised Difference Vegetation Index (NDVI), these methods enable continuous, spatially explicit monitoring of crop condition that complements conventional ground-based observation (Sishodia et al., 2020).

This research proposes a data-driven approach that integrates machine learning techniques with multi-source geospatial datasets to improve yield prediction for a representative Zambian farming context. Satellite-derived NDVI is combined with meteorological data, soil properties (pH and organic carbon), elevation, and slope to train a Random Forest regression model that captures non-linear patterns and delivers more reliable farm-scale forecasts. The study focuses on a wheat farm in Chongwe District, Lusaka Province. It spans the period 1999 to 2024, providing one of the longest continuous yield records analysed for this purpose in Zambia.
The study directly aligns with Sustainable Development Goal 2 (Zero Hunger) and Goal 13 (Climate Action), as well as Zambia's Eighth National Development Plan (8NDP), which emphasises enhancing agricultural productivity and building resilience through data-driven innovation. By demonstrating a replicable, field-ready predictive framework, the research contributes to the growing body of evidence on the applicability of ensemble machine learning for precision agriculture in data-scarce, climate-vulnerable agricultural systems.

1.1 Research Objectives

The primary objective of this study is to develop a machine learning model that integrates NDVI and meteorological data to improve the accuracy of crop yield prediction beyond traditional statistical methods. The secondary objectives are: (i) to identify the most influential environmental predictors of yield variation through correlation and regression analysis; (ii) to evaluate and compare the performance of Random Forest and XGBoost models under cross-validation, with particular attention to overfitting risk in small-sample agricultural datasets; and (iii) to develop a user-accessible decision-support dashboard for yield prediction by farmers, extension officers, and agronomists.

1.2 Climate-Smart Agriculture and Food Security

Agriculture remains central to Zambia's economy and rural livelihoods, contributing significantly to employment, household income, and national food security. The sector nonetheless faces increasing pressure from climate variability, erratic rainfall patterns, prolonged droughts, and rising temperatures. These climatic stressors directly affect crop productivity and increase vulnerability among both commercial and smallholder farmers, with maize and wheat among the staple crops most exposed to heat- and drought-related yield losses across the wider southern African region (Ngoma et al., 2022; Nxumalo et al., 2022).
Climate-smart agriculture has emerged as an important framework for enhancing resilience, improving productivity, and reducing environmental vulnerability within agricultural systems. The integration of remote sensing, geospatial analytics, and machine learning provides new opportunities for precision agriculture by enabling data-driven decision-making at both farm and regional scales (Kombat et al., 2021). Digital agriculture tools of this kind are increasingly recognised as a pathway for smallholder and commercial farmers alike to transition toward evidence-based input use and climate adaptation planning, although adoption across sub-Saharan Africa continues to be constrained by infrastructure, cost, and digital literacy barriers (Gumbi et al., 2023; Mhlanga & Ndhlovu, 2023).
Satellite-derived vegetation indices such as NDVI enable continuous monitoring of crop health and photosynthetic activity throughout the growing season. When combined with meteorological datasets and soil information, these data sources provide valuable predictors for yield-forecasting models, supporting irrigation planning, fertiliser optimisation, and climate adaptation strategies. The use of machine learning in agricultural prediction aligns strongly with Zambia's Eighth National Development Plan, which prioritises agricultural transformation, food security enhancement, and digital innovation. By improving predictive accuracy and enabling earlier intervention, geospatially informed machine learning systems can strengthen national resilience against climate-induced agricultural shocks.

2. Materials and Methods

2.1 Study Area

The study area is a commercial wheat farm located in Chongwe District, Lusaka Province, Zambia (Arc 1950 / UTM Zone 35S, EPSG:20935). Chongwe District lies in the transition zone between Zambia's semi-arid and sub-humid agro-ecological regions, making it broadly representative of conditions faced by both commercial and smallholder farmers across central Zambia. The farm's terrain is predominantly flat to gently sloping, as confirmed by slope analysis, which shows a maximum slope of approximately 2 degrees across the cultivated area, favouring mechanised agricultural operations.

Figure 1. Study area map of the wheat farm in Chongwe District, Lusaka Province, Zambia (Arc 1950 / UTM Zone 35S) (Developed by the Authors).

2.2 Data Acquisition

Multiple geospatial and meteorological datasets were integrated for this study, summarised in Table 1. Sentinel-2 imagery at 10 m resolution was used to extract NDVI values for the period 2019 to 2024, while Landsat 5, 7, and 8 imagery at 30 m resolution provided NDVI values from 1999 to 2018, ensuring continuity of the vegetation index time series across the full 25-year study period. A slope raster and elevation surface at 30 m resolution were derived from the Shuttle Radar Topography Mission (SRTM) digital elevation model. Soil pH and organic carbon rasters at 250 m resolution were obtained from the ISRIC SoilGrids database, which produces global soil property predictions using machine learning methods calibrated against approximately 240,000 soil profile observations worldwide (Poggio et al., 2021; Hengl et al., 2017). Daily precipitation estimates were obtained from the Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) dataset, processed via Google Earth Engine; CHIRPS has been independently validated as a reliable rainfall product for data-scarce regions of sub-Saharan Africa, performing favourably relative to other satellite-based rainfall estimates at daily, dekadal, and monthly time-scales (Geleta et al., 2021). Wheat yield data were obtained directly from farm records for the period 1999 to 2024.

Table 1: Summary of datasets used in the study.

Dataset	Source	Resolution	Period	Variables Extracted
Sentinel-2 NDVI	Copernicus/GEE	10 m	2019-2024	Mid-season, late-season NDVI
Landsat 5/7/8 NDVI	USGS/GEE	30 m	1999-2018	Seasonal NDVI time series
SRTM DEM	NASA/USGS	30 m	Static	Slope, elevation
ISRIC SoilGrids	ISRIC	250 m	Static	Soil pH, organic carbon
CHIRPS rainfall	CHIRPS/GEE	~5 km	1999-2024	Rainfall days count, total rainfall
Temperature data	ZMD / ERA5	~10 km	1999-2024	Average temperature, GDD, heat stress days
Wheat yield	Farm records	Field-scale	1999-2024	Annual yield (t/ha)

2.3 Data Pre-processing and Feature Engineering

All raster datasets were aligned and resampled to a common spatial reference and pixel resolution using QGIS. Bilinear interpolation was applied to continuous variables, including NDVI, rainfall, and soil pH, while nearest-neighbour resampling was used for categorical layers. A panchromatic raster at higher spatial resolution was used as the geometric reference to guide alignment across all layers, improving spatial consistency across the multi-source dataset.

Feature engineering was applied to transform the multi-temporal dataset into meaningful predictors. For NDVI, temporal aggregation techniques yielded early- and mid-season averages, peak values, and vegetation growth rate indicators. Temperature data were processed to create seasonal averages, maximum and minimum values, Growing Degree Days (GDD), and accumulated heat stress days. The initial feature matrix of 400 observations by 1,371 candidate features was reduced through systematic feature engineering and variable selection to a final structure of 400 observations by 80 predictors, minimising the risk of overfitting while preserving the most informative variables (Genuer et al., 2010; Belgiu & Drăguț, 2016).

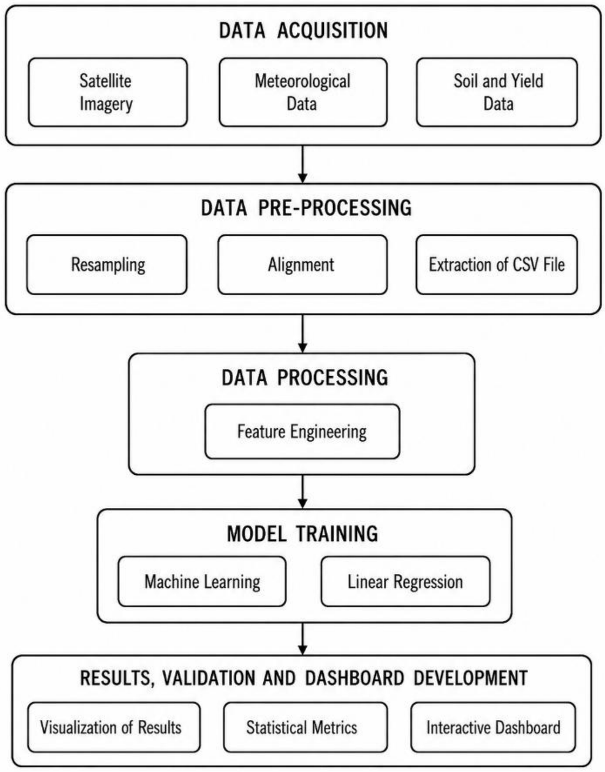

Figure 2. Methodological flowchart for the crop yield prediction pipeline, from data acquisition through pre-processing, model training, and dashboard development (Developed by the Authors).

2.4 Regression Analysis

Pearson correlation analysis and simple linear regression were performed to examine the relationships between annual yield variation and key environmental variables and to establish baseline relationships before machine learning model development. Scatter plots with regression trend lines, classified by year-performance category (good, average, or poor), were generated for each predictor variable to assess the strength and direction of association visually.

2.5 Machine Learning Model Training

Two supervised machine learning algorithms were evaluated: Random Forest and Extreme Gradient Boosting (XGBoost). Both were implemented in Python using the scikit-learn and XGBoost libraries. Five-fold cross-validation was applied to all models to minimise bias and provide robust performance estimates. Model performance was evaluated using the coefficient of determination (R2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The target variable was the detrended annual wheat yield (tonnes per hectare), obtained by removing the long-term positive trend of approximately 0.129 t/ha per year to isolate the interannual variability attributable to environmental predictors rather than to gradual improvements in farm management over time.

2.6 Dashboard Development

An interactive prediction dashboard was developed to make the trained model accessible to non-technical users, including farmers, extension officers, and policymakers. The dashboard provides input fields for NDVI (early and mid-season), elevation, slope, soil organic carbon, growing-season rainfall, temperature, soil pH, planting date, and irrigation status. The dashboard generates a yield prediction and is structured to be adapted to other crop types by adjusting the underlying set of input parameters.

2.7 Random Forest and Ensemble Learning Framework

Random Forest is an ensemble machine learning algorithm that constructs multiple independent decision trees using bootstrap sampling and random feature selection at each split. The final prediction is generated by aggregating the predictions of all trees within the ensemble. This architecture reduces model variance and minimises overfitting, particularly in datasets with limited sample sizes or noisy environmental variables (Belgiu & Drăguț, 2016; Dhillon et al., 2023). The Random Forest prediction function is expressed in Equation 1.

$$fRF(x) = (1/B)\ \Sigma\ Tb(x), \text{ for } b = 1 \text{ to } B$$ **Equation 1.** Random Forest prediction function.

where B represents the number of trees in the ensemble, and Tb(x) represents the prediction generated by each tree. Random Forest models are particularly suitable for agricultural applications because they effectively capture complex, non-linear interactions among environmental variables such as rainfall, temperature, soil characteristics, and vegetation indices, consistent with findings reported across multiple recent applications of ensemble learning to crop yield prediction (Shahhosseini et al., 2021; Zhou et al., 2022).

Model performance was evaluated using the Root Mean Square Error, the Mean Absolute Error, and the coefficient of determination. RMSE is calculated as shown in Equation 2.

$$RMSE = sqrt[\, (1/n)\, \Sigma\, (yi - \hat{y}i)^2\,]$$ **Equation 2**. Root Mean Square Error.

The coefficient of determination is expressed in Equation 3.

$$R^2 = 1 - [\, \Sigma\, (yi - \hat{y}i)^2 / \Sigma\, (yi - \bar{y})^2\,]$$ **Equation 3**. Coefficient of determination.

These metrics collectively evaluate predictive accuracy, error magnitude, and the proportion of yield variability explained by the model, providing a comprehensive basis for comparing Random Forest against XGBoost under identical cross-validation conditions.

3. Results

3.1 Preliminary Spatial Analysis

Spatial analysis of the study area revealed considerable variation in soil properties, vegetation health, and terrain characteristics. The soil pH raster showed predominantly acidic conditions (pH 4 to 5) across most of the wider district, with localised near-neutral patches (pH 6 to 7) representing the most suitable zones for wheat production within the farm boundary itself. The slope analysis confirmed predominantly flat to gently sloping terrain of 2 degrees or less, favourable for mechanised agriculture. The NDVI raster showed variable vegetation performance across the district, with areas of high photosynthetic activity in well-managed zones contrasting with sparse vegetation or bare soil.

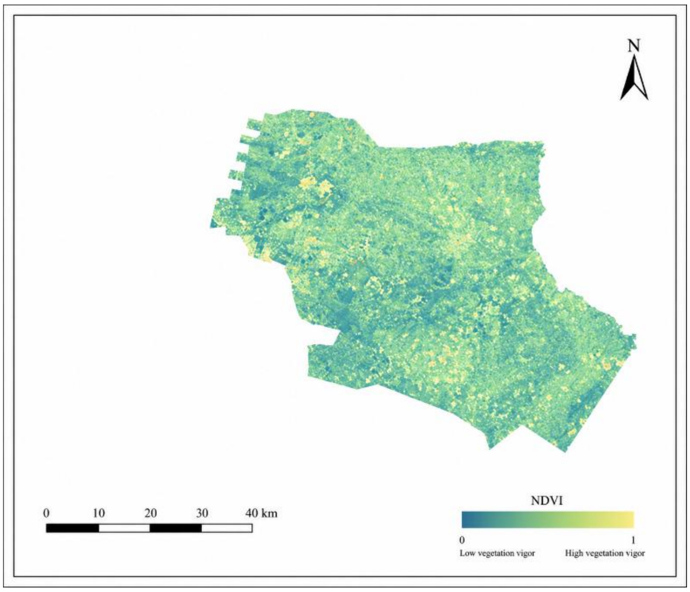

Figure 3. NDVI raster map of the study area in Chongwe District, illustrating spatial variation in vegetation health and crop vigour (Developed by the Authors).

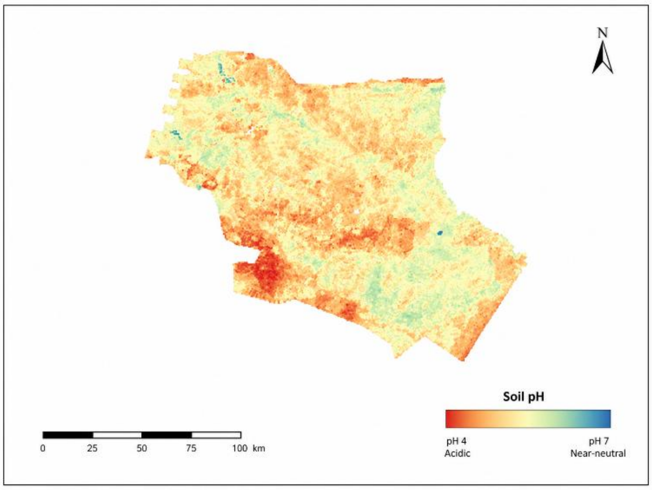

Figure 4. Soil pH raster map of the study area, showing predominantly acidic conditions with localised near-neutral areas suitable for wheat (Developed by the Authors).

3.2 Regression Analysis Results

Soil pH showed essentially no relationship with wheat yield variation within the cultivated farm boundary (r = 0.020, p = 0.924), consistent with the narrow pH range observed across the immediate study area (6.16 to 6.20), where variation was negligible. This narrow range falls within the generally accepted optimal pH band for wheat production, consistent with evidence that soil acidification meaningfully constrains yield only below approximately 5.5 (Warner et al., 2023; Zhang et al., 2023). In contrast, rainfall frequency, measured as the count of rainfall days during the growing season, exhibited a statistically significant moderate positive correlation with yield variation (r = 0.444, p = 0.026), with each additional rainfall day associated with an approximate yield increase of 1.14 t/ha.

Growing Degree Days (GDD) showed a statistically significant negative correlation with yield variation (r = -0.406, p = 0.044), indicating that higher heat accumulation during the growing season is associated with lower yields. This finding is consistent with the physiological literature on wheat heat stress, in which elevated temperatures during grain filling shorten the effective growing season, accelerate senescence, and reduce both grain weight and total yield (Farhad et al., 2023). Average seasonal temperature also showed a significant negative correlation with yield variation (r = -0.403, p = 0.046), with each 1 degree Celsius increase in seasonal temperature associated with a yield decrease of approximately 0.86 t/ha.

A 25-year yield trend analysis covering the period 2000 to 2024 revealed an overall positive yield trajectory of approximately 0.129 t/ha per year, attributable to gradual improvements in farm management practices over the study period. The best-performing year was 2010/2011, with a yield of 9.9 t/ha, while the lowest yield occurred in 2000/2001, at 5.0 t/ha. Above-average yield years were consistently associated with cooler growing conditions and higher rainfall frequency, reinforcing the regression findings reported above.

Table 2: Pearson correlation coefficients between wheat yield variation and key environmental predictors.

Predictor Variable	r (Pearson)	p-value	Significance	Direction
Rainfall Days Count	0.444	0.026	Significant	Positive
Average Seasonal Temperature	-0.403	0.046	Significant	Negative
GDD Season Total	-0.406	0.044	Significant	Negative
Minimum Average Temperature	-0.460	< 0.05	Significant	Negative
Heat Stress Days Total	-0.350	< 0.05	Significant	Negative
Soil pH	0.020	0.924	Not significant	Near zero
NDVI Maximum	-0.280	> 0.05	Not significant	Weak negative

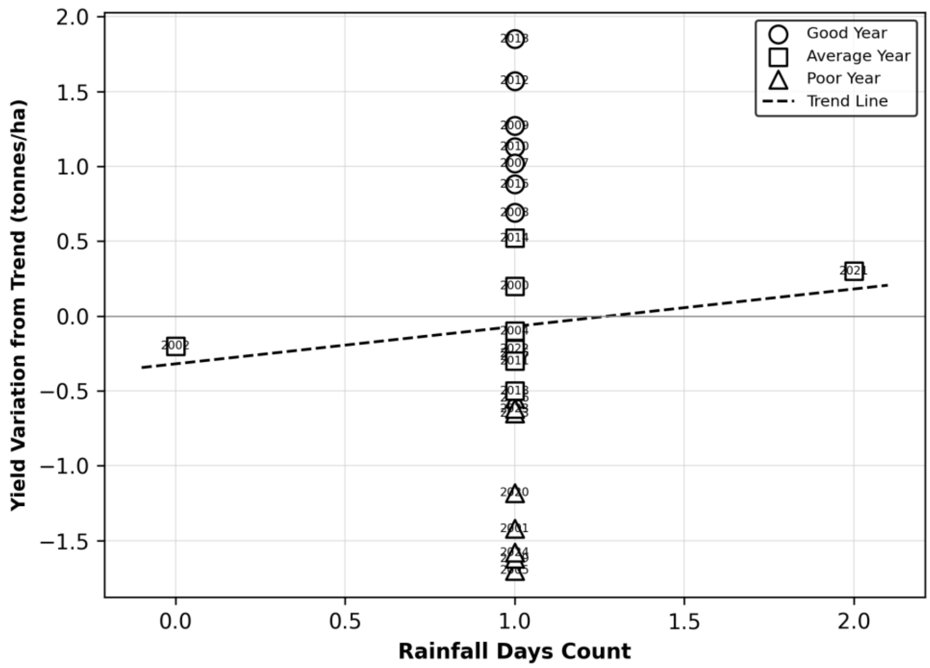

Figure 5. Scatter plot of rainfall days count against wheat yield variation, showing a significant positive correlation (Developed by the Authors).

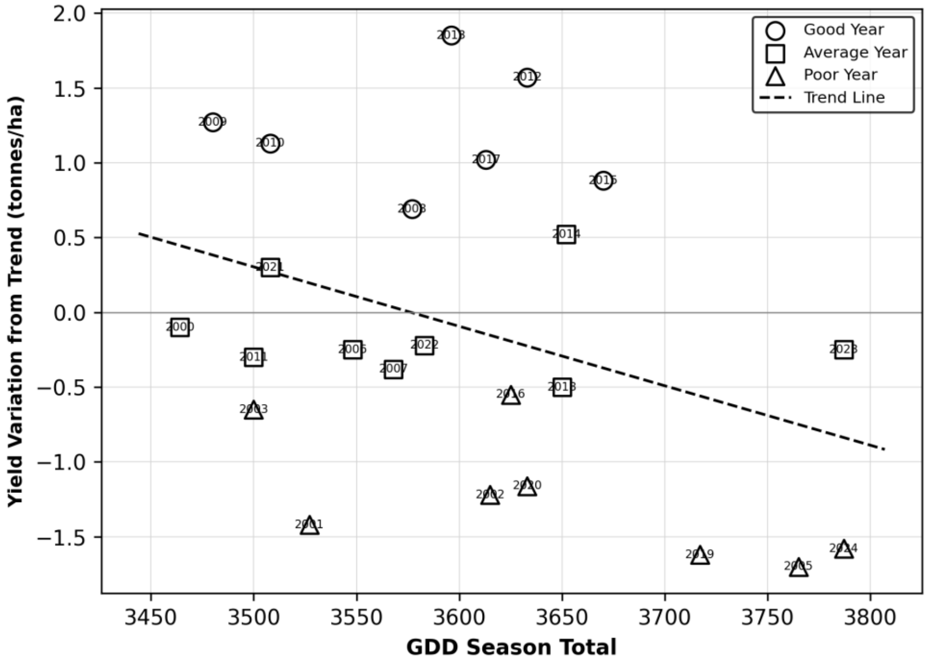

Figure 6. Scatter plot of Growing Degree Days (GDD) season total against wheat yield variation, showing a significant negative correlation (Developed by the Authors).

3.3 Machine Learning Model Performance

Two models were evaluated under five-fold cross-validation. XGBoost achieved a coefficient of determination of 0.950 (standard deviation 0.091), while Random Forest achieved 0.389 (standard deviation 0.103). Although XGBoost's apparent score is markedly higher, this result is attributable to overfitting: with a small, spatially homogeneous dataset characterised by minimal soil pH variation and consistent management practices, XGBoost's sequential boosting architecture memorised patterns in the training data rather than learning generalisable relationships. The relatively high variance across folds (standard deviation of 0.091) further indicates instability in the XGBoost cross-validation results.

Random Forest's more modest yet stable coefficient of determination of 0.389 better reflects the dataset's genuine predictive signal. Its ensemble architecture, which builds independent decision trees on random subsets of features, is specifically designed to resist overfitting (Genuer et al., 2010). The model successfully captured rainfall frequency patterns, temperature fluctuations, and NDVI seasonal variation as the primary drivers of inter-annual yield variability, consistent with comparable applications of Random Forest to crop modelling under similarly constrained sample sizes (Shahhosseini et al., 2021; Dhillon et al., 2023).

Table 3: Model performance comparison between XGBoost and Random Forest under five-fold cross-validation.

Model	R2 (mean +/- SD)	Interpretation	Overfitting Risk
XGBoost	0.950 +/- 0.091	Excellent apparent fit	High: memorisation on a small dataset
Random Forest	0.389 +/- 0.103	Moderate, genuine generalisation	Low: appropriate for dataset size

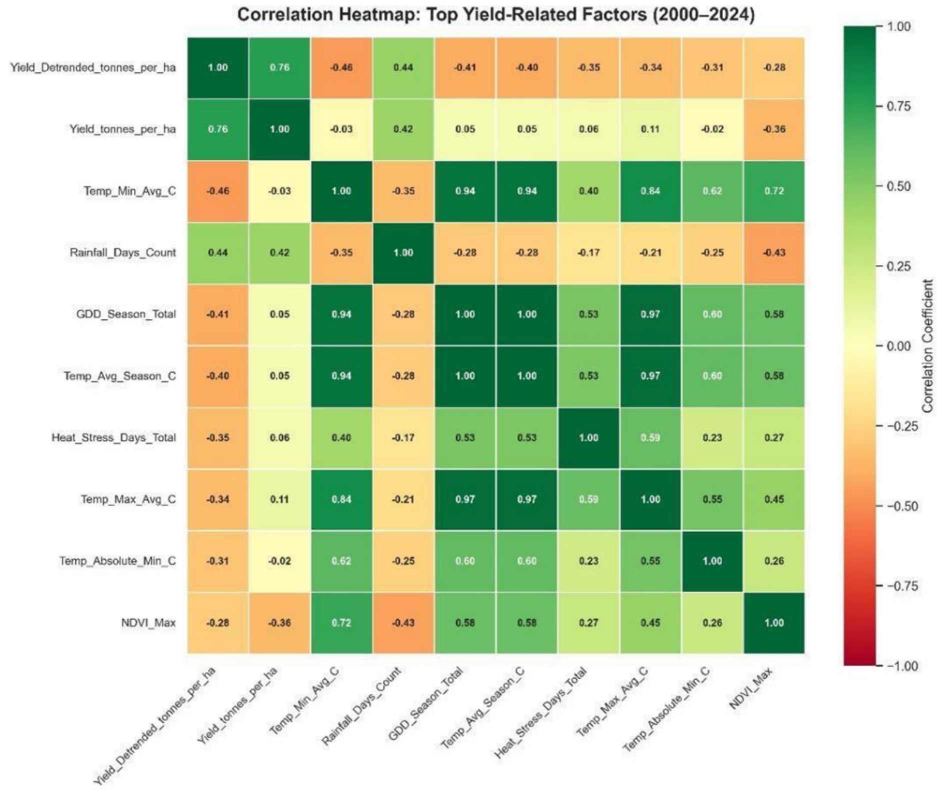

Figure 7. Correlation heatmap of yield and key environmental and meteorological variables for the Chongwe District wheat farm, 2000 to 2024 (Developed by the Authors).

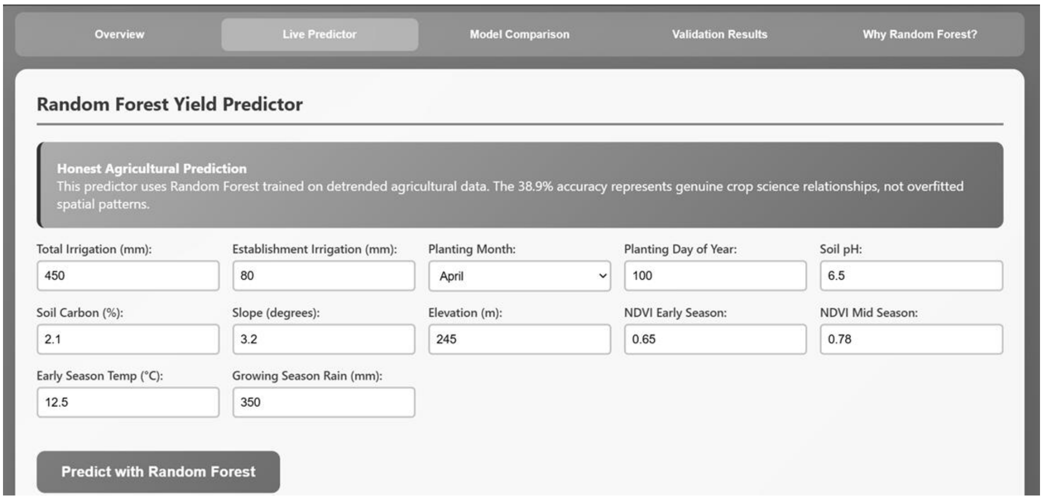

Figure 8. Yield prediction dashboard interface, showing input fields for NDVI, rainfall, temperature, and soil parameters, with prediction output (Developed by the Authors).

4. Discussion

The regression and machine learning results align with and extend the existing literature on crop yield determinants in semi-arid southern African environments. Rainfall frequency emerged as the dominant positive predictor, consistent with studies showing that the timing and frequency of rainfall events, rather than total precipitation alone, are the critical determinants of crop water availability during sensitive phenological stages (Siatwiinda et al., 2021). This pattern has also been documented at the regional scale, where climate change projections for Zambia indicate that shifts in rainfall distribution, rather than changes in total seasonal rainfall, are likely to be the primary driver of future maize and wheat yield variability (Ngoma et al., 2022).

The negative temperature-yield relationship observed in this study ($r = -0.403$ for average seasonal temperature and $r = -0.406$ for Growing Degree Days) is consistent with the physiological literature on wheat heat stress. Elevated temperatures during grain filling shorten the effective growing season and accelerate senescence, reducing grain weight and total yield (Farhad et al., 2023). This relationship has been observed across multiple wheat-producing regions, including South Africa, where meteorological drought variability and elevated thermal stress have consistently reduced wheat productivity over multi-decadal periods (Nxumalo et al., 2022). The consistency of this finding across different agro-ecological contexts strengthens confidence in the present study's conclusions regarding the centrality of thermal stress as a constraint on wheat yield in Zambia.

The negligible influence of soil pH ($r = 0.020$) observed within the cultivated farm boundary is attributable to the narrow pH range present across the study area (6.16 to 6.20), which falls within the optimal range for wheat production. This finding should not be interpreted as evidence that soil pH is unimportant for wheat productivity in general; rather, it reflects the specific edaphic conditions of the farm under study. Soil acidity is well documented as a major constraint on cereal productivity across sub-Saharan Africa, with acidic soils estimated to cover approximately 30 per cent of the region and substantially reduce yields when pH falls below 5.5 (Warner et al., 2023; Zhang et al., 2023). The absence of a detectable pH effect in the present study therefore validates the analytical focus on climatic rather than edaphic predictors for this particular farm, while underscoring that soil pH should not be dismissed as a predictor in studies covering more heterogeneous soil conditions.

The comparison between XGBoost and Random Forest provides an important methodological lesson for small-dataset agricultural prediction studies. XGBoost's apparently high coefficient of determination (0.950) is a classic overfitting artefact when applied to small, spatially homogeneous datasets with limited feature variation. This pattern echoes broader findings in the crop yield prediction literature, which caution that boosting-based ensemble methods, while powerful, are particularly susceptible to memorisation when sample sizes are small relative to feature dimensionality (van Klompenburg et al., 2020). Random Forest's ensemble construction, with each tree built on a random subset of features and a bootstrap sample of observations, provides genuine protection against this risk, delivering a stable, interpretable model that better reflects the actual predictive signal present in the data (Belgiu & Drăguţ, 2016; Genuer et al., 2010). This finding reinforces recent calls in the literature for greater attention to model selection criteria beyond raw accuracy metrics when working with small, single-site agricultural datasets common in data-scarce regions such as Zambia (Dhillon et al., 2023; Shahhosseini et al., 2021).

The interactive dashboard developed in this study represents a practical contribution to agricultural planning in Zambia. By providing a parameter-driven interface that agronomists and extension officers can use to generate yield forecasts from routinely collected environmental data, the study operationalises its research findings beyond the academic domain. This contribution is particularly relevant given that digital agriculture tools remain underutilised across much of sub-Saharan Africa, with adoption constrained by infrastructure, cost, and digital literacy barriers (Mhlanga & Ndhlovu, 2023; Gumbi et al., 2023). A lightweight, locally calibrated decision-support tool of the kind developed here addresses a documented gap between the technical sophistication of academic yield prediction research and the practical accessibility required for uptake by farmers and extension services on the ground.

The findings also carry implications for the broader integration of satellite Earth observation into Zambia's agricultural early-warning and planning systems. The reliable performance of CHIRPS rainfall estimates, as demonstrated in independent validation studies across sub-Saharan Africa (Geleta et al., 2021; Gebrehiwot et al., 2023), supports the continued use of satellite-derived precipitation data as a cost-effective substitute for sparse ground-based weather station networks, particularly in remote farming districts where conventional meteorological infrastructure is limited. Combined with the demonstrated predictive value of NDVI and thermal accumulation indices, this suggests a viable pathway toward a low-cost, satellite-driven agricultural monitoring framework appropriate for national-scale deployment.

4.1 Limitations

Several limitations should be acknowledged when interpreting these findings. First, the short temporal overlap of high-resolution Sentinel-2 data with the full yield record (2019 to 2024 only) limited the integration of fine-resolution spatial data with the 25-year yield dataset, requiring reliance on coarser-resolution Landsat imagery for the earlier portion of the time series. Second, the absence of predicted meteorological data prevented the development of forward-looking, pre-season yield forecasts, restricting the model's current utility to retrospective and in-season analysis rather than seasonal forecasting. Third, the spatial resolution of soil and meteorological variables, at 250 m and 5 to 10 km, respectively, does not capture field-level variation that may be agronomically significant at finer scales. Fourth, the study's single-farm, single-crop focus limits the immediate transferability to other crops and agro-ecological zones across Zambia without further validation. Finally, the modest sample size (400 observations before feature reduction, drawn from a single 25-year time series) constrains the statistical power available for more complex model architectures, reinforcing the methodological conclusion that simpler, variance-resistant models such as Random Forest are more appropriate than highly flexible boosting algorithms in this data context.

5. Conclusions

This study demonstrates the feasibility and value of integrating satellite-derived geospatial data with machine learning to predict crop yields in Zambia. The Random Forest model, trained on a 25-year yield record with multi-source environmental predictors, successfully identified rainfall frequency and seasonal temperature as the primary drivers of inter-annual yield variability, with clear implications for precision agriculture, irrigation scheduling, and climate adaptation planning in Zambia's wheat-producing districts.

The central methodological contribution of this research is the demonstration that model selection in small-dataset agricultural prediction contexts is critical to obtaining genuinely useful predictive tools. XGBoost's apparent superiority, with a coefficient of determination of 0.950, masked systematic overfitting on a small, spatially homogeneous dataset, while Random Forest's more modest but stable performance, with a coefficient of determination of 0.389, better represents the genuine predictive signal available from the data. This finding directly addresses a gap in the existing literature, in which the comparative overfitting risk of boosting versus bagging ensemble methods under small-sample agricultural conditions has received limited explicit attention, and offers a transferable methodological caution for researchers undertaking similar single-site or data-scarce yield prediction studies elsewhere in the region.

In relation to the study's original research objectives, the findings confirm that a machine learning model integrating NDVI and meteorological data can improve crop yield prediction accuracy beyond what conventional statistical approaches typically achieve in similarly constrained data settings, while also revealing that the most influential predictors

are rainfall frequency and seasonal thermal accumulation rather than soil pH, within the specific edaphic context of the study farm. The comparative evaluation of Random Forest and XGBoost further demonstrates that predictive accuracy alone is an insufficient basis for model selection in small-sample contexts. That genuine generalisation performance under cross-validation must be weighed against apparent fit. The development of the interactive yield prediction dashboard operationalises these findings into a tool with immediate practical utility for non-specialist users.
This study contributes to the growing body of literature on geospatial machine learning for agricultural decision support in sub-Saharan Africa by providing one of the longest continuous farm-scale yield records analysed using this combined NDVI-meteorological-machine learning framework in Zambia specifically, and by offering empirical evidence on overfitting risk that is directly relevant to the wider community of researchers working with similarly small, single-site agricultural datasets across the region. Future research should incorporate longer high-resolution satellite time series as Sentinel-2 archives continue to grow, integrate seasonal climate forecasts to enable forward-looking, pre-season yield predictions, and extend the present framework to multiple crop types and agro-ecological zones across Zambia, including maize and soybean systems that face comparable climate-related yield risks (Ngoma et al., 2022; Siatwiinda et al., 2021). The interactive prediction dashboard developed in this study provides an immediately applicable decision-support tool for farmers, extension officers, and agricultural policymakers in Zambia, operationalising the research findings in a format accessible to non-specialist users and contributing a practical, field-ready output that directly supports the precision agriculture and climate resilience priorities articulated in Zambia's Eighth National Development Plan.

Acknowledgements
The authors acknowledge the supervision and technical guidance of Dr F. A. S Banda, Dr P. H. Nyimbili, and Mr M. Sakala M and thank the University of Zambia, Department of Geomatic Engineering, for academic and resource support throughout this research.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest
The authors report no conflicts of interest.

Data Availability Statement
The farm-level wheat yield data used in this study are held under a confidentiality agreement with the participating commercial farm and are not publicly available. Satellite imagery (Sentinel-2, Landsat 5/7/8), CHIRPS rainfall data, SRTM elevation data, and ISRIC SoilGrids data are openly accessible through Google Earth Engine and the respective provider repositories. Processed datasets and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request.

Institutional Review Board Statement
This study did not involve human participants, human data, or animals, and therefore did not require ethics committee approval. All environmental, satellite, and farm yield data used were either obtained from open-access repositories or shared under a data-use agreement with the participating farm.

CRediT Author Statement
Muuka I.: Conceptualisation, Data Processing, Model Training, Writing - original draft. Changwe K.: Conceptualisation, Data Collection, Writing - original draft. Nyongolo M.: Writing - review and editing. Nyimbili P.H.: Supervision, Writing - review and editing. Sakala M.: Supervision, Writing - review and editing. Banda F.A.S.: Supervision. Mwanaumo E.M.: Supervision, Resources. Thwala W.D.: Validation, Project Administration. All authors have read and approved the final manuscript.

References
Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011
Briffa, J., Sinagra, E., & Blundell, R. (2020). Heavy metal pollution in the environment and its toxicological effects on humans. Heliyon, 6(9), e04691. https://doi.org/10.1016/j.heliyon.2020.e04691
Dhillon, M. S., Dahms, T., Kuebert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I., & Ullmann, T. (2023). Integrating random forests and crop modelling improves crop yield prediction for winter wheat and oilseed rape. Frontiers in Remote Sensing, 3, 1010978. https://doi.org/10.3389/frsen.2022.1010978
Farhad, M., Kumar, U., Tomar, V., Bhati, P. K., Krishnan, J. N., Barek, V., Brestic, M., & Hossain, A. (2023). Heat stress in wheat: A global challenge to feed billions in the current era of the changing climate. Frontiers in Sustainable Food Systems, 7, 1203721. https://doi.org/10.3389/fsufs.2023.1203721
Gao, X., Yang, T., Zhou, Z., Yang, B., & Yan, F. (2021). Effects of soil acidification on nitrogen and phosphorus uptake of crops: A meta-analysis. Journal of Soils and Sediments, 21(7), 2552-2563. https://doi.org/10.1007/s11368-021-02938-0
Gebrehiwot, S. G., Bewket, W., & Mengistu, T. (2023). Accuracy of satellite and reanalysis rainfall estimates over Africa: A multi-scale assessment of eight products for continental applications. Journal of Hydrology: Regional Studies, 49, 101512. https://doi.org/10.1016/j.ejrh.2023.101512

Geleta, T. A., Gemeda, D. O., Tareke, K. A., & Sileshi, M. A. (2021). Evaluation of Climate Hazards Group InfraRed Precipitation Station (CHIRPS) satellite-based rainfall estimates over Finchaa and Neshe watersheds, Ethiopia. Engineering Reports, 3(12), e12338. https://doi.org/10.1002/eng2.12338

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using Random Forests. Pattern Recognition Letters, 31(14), 2225-2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gumbi, N., Gumbi, L., & Twinomurinzi, H. (2023). Towards sustainable digital agriculture for smallholder farmers: A systematic literature review. Sustainability, 15(16), 12530. https://doi.org/10.3390/su151612530

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE, 12(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hunter, M. C., Smith, R. G., Schipanski, M. E., Atwood, L. W., & Mortensen, D. A. (2020). Agriculture in 2050: Recalibrating targets for sustainable intensification. BioScience, 67(4), 386-391. https://doi.org/10.1093/biosci/bix010

Kombat, R., Sarfatti, P., & Fatunbi, O. A. (2021). A review of climate-smart agriculture technology adoption by farming households in sub-Saharan Africa. Sustainability, 13(21), 12130. https://doi.org/10.3390/su132112130

Mbow, C., Rosenzweig, C., Barioni, L. G., Benton, T. G., Herrero, M., Krishnapillai, M., Liwenga, E., Pradhan, P., Rivera-Ferre, M. G., Sapkota, T., Tubiello, F. N., & Xu, Y. (2021). Food security. In P. R. Shukla et al. (Eds.), Climate Change and Land: An IPCC Special Report (pp. 437-550). Cambridge University Press. https://doi.org/10.1017/9781009157988.007

Mhlanga, D., & Ndhlovu, E. (2023). Digital technology adoption in the agriculture sector: Challenges and complexities in Africa. Human Behavior and Emerging Technologies, 2023, 6951879. https://doi.org/10.1155/2023/6951879

Mulungu, K., Tembo, G., Bett, H., & Ngoma, H. (2021). Climate change and crop yields in Zambia: Historical effects and future projections. Environment, Development and Sustainability, 23(8), 11859-11880. https://doi.org/10.1007/s10668-020-01146-6

Ngoma, H., Lupiya, P., Kabisa, M., & Hartley, F. (2022). Climate change impacts on maize and soybean yields in Zambia. Agronomy Journal, 114(5), 2430-2444. https://doi.org/10.1002/agj2.21092

Nxumalo, G., Bashir, B., Alsafadi, K., Bachir, H., Harsányi, E., Arshad, S., & Mohammed, S. (2022). Meteorological drought variability and its impact on wheat yields across South Africa. International Journal of Environmental Research and Public Health, 19(24), 16469. https://doi.org/10.3390/ijerph192416469

Phakela, K., Lannucci, A., Akin, D. S., Mason, R. E., Subramanian, N. K., Tan, K., & Hays, D. B. (2022). Wheat grain quality response to heat and drought stress and the relationship to associated quantitative trait loci. Cereal Chemistry, 99(1), 80-93. https://doi.org/10.1002/cche.10489

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. SOIL, 7(1), 217-240. https://doi.org/10.5194/soil-7-217-2021

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports, 11(1), 1606. https://doi.org/10.1038/s41598-020-80820-1

Siatwiinda, S. M., Supit, I., Boogaard, H., van Ittersum, M. K., Ngoma, H., & Yerokun, O. (2021). Climate change impacts on rainfed maize yields in Zambia under conventional and optimized crop management. Climatic Change, 167(3-4), 39. https://doi.org/10.1007/s10584-021-03191-0

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. Remote Sensing, 12(19), 3136. https://doi.org/10.3390/rs12193136

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

Warner, J. M., Mann, M. L., Chamberlin, J., & Tizale, C. Y. (2023). Estimating acid soil effects on selected cereal crop productivities in Ethiopia: Comparing economic cost-effectiveness of lime and fertilizer applications. PLOS ONE, 18(1), e0280230. https://doi.org/10.1371/journal.pone.0280230

Zhang, T., Li, J., Fan, J., Zhang, F., Li, F., & Xiang, Y. (2023). Optimizing soil pH for higher crop productivity in acidic croplands of China: A meta-analysis. Field Crops Research, 295, 108888. https://doi.org/10.1016/j.fcr.2023.108888

Zhou, G., Wang, Z., Xu, X., Jiang, W., Wang, Y., & Bai, X. (2022). Application of random forest and gradient boosting in predicting crop yields using remote sensing data. Remote Sensing, 14(9), 2133. https://doi.org/10.3390/rs14092133