RESEARCH ARTICLE

JENER Journal of Empirical and Non-Empirical Research
Volume 2, Issue 7, Page 44-51 | Article No. 10
DOI: 10.4898/jener.v2i7.a10
ISSN: 7895-9216
Journal homepage: https://jenerjournal.com
Journal email: editor@jenerjournal.com

JENER Journal
JENER Journal of Empirical and Non-Empirical Research

Soil Nitrogen, Not Rainfall, Drives Maize Yields in Semi-Arid Zambia: Machine Learning Evidence from 50 District-Year Observations

Gift Moyoa[1] & Dr Muwanei Sinyinda[2]

[1,2]Department of Computer Science and Information Technology, Mulungushi University, Great North Road

ARTICLE INFORMATION

Article history:
Published: July 2026

Keywords:
Maize yield
Soil nitrogen
Climate variability
Machine learning
Random Forest
XGBoost

ABSTRACT

Understanding the relative importance of soil properties versus climate variability for maize yield determination is critical for targeting agricultural interventions in sub-Saharan Africa.This study analyzed 15 years of integrated data (2008-2022) from six districts across Central and Southern Zambia, combining maize yield records (n=50 complete district-year observations), CHIRPS satellite rainfall data, SoilGrids soil properties, and ERA5-Land temperatures. Using XGBoost and Random Forest machine learning models, we quantified the influence of 11 predictor variables on maize yield.. Soil nitrogen was the dominant predictor, accounting for 77.5% (95% CI: 67.2-87.7%) of XGBoost's predictive power. Combined with clay content (13.5%) and soil organic carbon (4.6%), soil properties contributed 95.6% of the model's predictive power. Correlation analysis confirmed a strong positive relationship between nitrogen and yield ($r = 0.723$, $p < 0.001$). In striking contrast, climate variables showed minimal influence: annual rainfall correlated weakly and non-significantly with yield ($r = 0.182$, $p > 0.05$), contributing less than 16% of Random Forest's predictive power and less than 1% of XGBoost's.

Conclusions: These findings challenge the prevailing assumption that rainfall variability is the primary constraint to maize production in semi-arid smallholder systems. Under normal rainfall conditions (study period range: 310-486 mm annual rainfall), soil fertility—specifically nitrogen availability—is the limiting factor. Agricultural policy and extension services should rebalance investments to prioritize soil health management alongside climate adaptation.

1. Introduction

1.1 Global Food Security Context

Global food demand is projected to increase by 60-70% by 2050, placing unprecedented pressure on agricultural systems to improve productivity from existing land resources (FAO, 2022). Climate change has already reduced global agricultural total factor productivity by approximately 21% since 1961, with warmer regions experiencing the largest impacts (Ortiz-Bobea et al., 2021). Maize, the world's most widely grown staple crop, is particularly vulnerable to rising temperatures and changing rainfall patterns, with future warming projected to reduce yields by 20-30% by 2100 under high emission scenarios (Jägermeyr et al., 2021).

1.2 Sub-Saharan Africa Context

In sub-Saharan Africa (SSA), approximately 95% of cultivated land is rain-fed, making agricultural production highly vulnerable to rainfall variability (You et al., 2021). Smallholder farmers, who cultivate less than 2 hectares on average, face multiple constraints including limited access to improved inputs, credit, information, and markets (Jayne et al., 2021). These constraints, combined with climate variability, have resulted in persistent yield gaps: average maize yields in SSA (approximately 2.0 tonnes/ha) remain far below those in North America (approximately 10.5 tonnes/ha) (FAOSTAT, 2022).

1.3 Zambia Focus

Zambia exemplifies these challenges. Maize is the country's staple crop, occupying 60-70% of cultivated land and providing the primary food source for most households (Chapoto et al., 2020). Central and Southern Provinces together contribute approximately 40% of Zambia's national maize production (Zambia Statistics Agency, 2022). However, yields in these provinces remain highly unstable due to recurrent droughts, erratic rainfall, declining soil fertility, and increasing temperature variability (Libanda et al., 2020). During the 2015-2016 El Niño drought, maize production in Southern Province declined by more than 60%, falling from approximately 450,000 metric tonnes to below 180,000 metric tonnes (WFP, 2020).

1.4 The Unresolved Question

Despite decades of research, a fundamental question remains unresolved for semi-arid smallholder systems: What is the relative importance of soil properties versus climate variability in determining maize yield?

Conventional wisdom, reinforced by devastating drought events such as the 2015-2016 El Niño, suggests that rainfall variability is the primary constraint to maize production (Lobell et al., 2020; Ray et al., 2020). This assumption has driven substantial investment

in climate monitoring systems, early warning infrastructure, and drought-tolerant crop varieties across SSA (Hansen et al., 2020; World Bank, 2019).

However, an alternative hypothesis, grounded in the "law of the minimum" (von Liebig, 1840), suggests that under normal rainfall conditions—which characterize most growing seasons—the most limiting factor shifts from water to soil nutrients. Vanlauwe et al. (2018) demonstrated this pattern across 100 fertilizer trials in SSA, finding that nitrogen was the most limiting nutrient in 85% of cases when water was not limiting. In Zambia specifically, Burke et al. (2020) found that 78% of soil samples from smallholder farms fell below the critical nitrogen threshold of 2.0 g/kg for maize production. These findings suggest that the relative importance of climate versus soil factors may depend critically on the range of rainfall conditions sampled.

1.5 Machine Learning as a Tool

Recent advances in machine learning offer new opportunities to quantify the relative importance of multiple yield determinants simultaneously. Unlike traditional statistical methods that require pre-specified functional forms, machine learning algorithms can capture complex, non-linear relationships and provide interpretable feature importance measures (Van Klompenburg et al., 2020; Jeong et al., 2022).

However, machine learning applications for yield prediction in Zambia remain limited. Molitor et al. (2026) applied the MOSAIKS framework using satellite imagery and ridge regression at national scale, achieving an R^2 of 0.74. Kenduiywo and Miller (2024) used Random Forest in Senanga District with an RMSE of 0.20 MT/ha. No study has compared multiple algorithms or quantified the relative importance of soil versus climate variables specifically for Central and Southern Provinces.

1.6 Study Contribution

This study addresses this gap by applying XGBoost (Chen & Guestrin, 2016) and Random Forest (Breiman, 2001) to 15 years of integrated climate, soil, and yield data from Central and Southern Zambia. Specifically, we: (1) quantify the predictive importance of soil properties (nitrogen, clay content, organic carbon, pH) relative to climate variables (rainfall, temperature, drought); (2) characterize the direction and strength of relationships between individual predictors and maize yield using correlation analysis; and (3) derive actionable implications for agricultural policy and extension services based on the relative importance of different yield determinants.

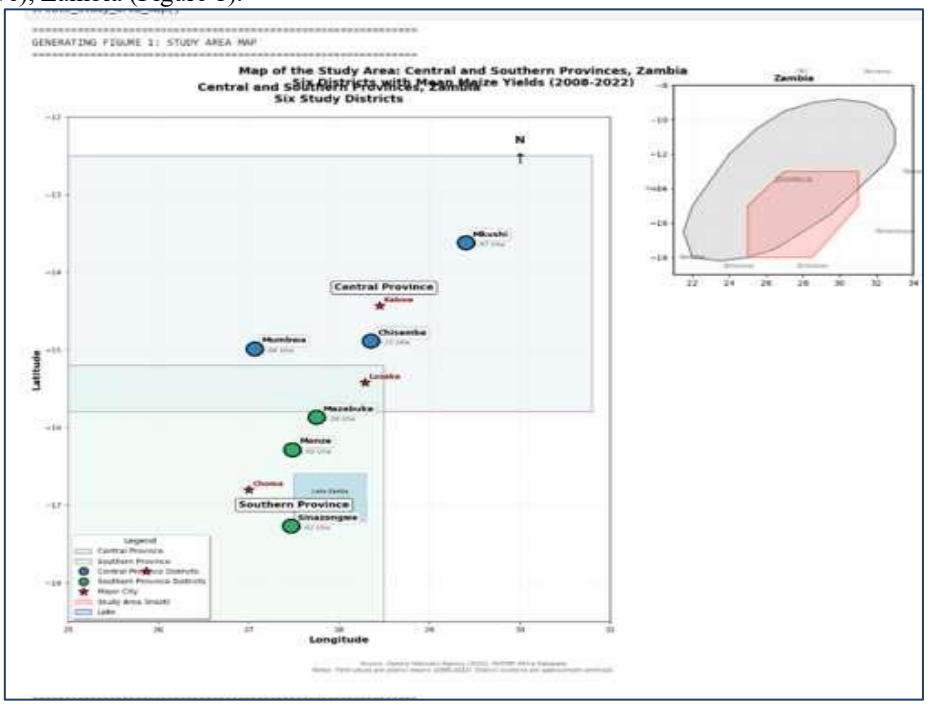
1.7 Paper Structure

Section 2 describes our data and methods, Section 3 presents results, Section 4 discusses implications, and Section 5 concludes.

2. Materials and Methods

2.1 Study Area

The study was conducted in six districts across Central Province (Chisamba, Mkushi, Mumbwa) and Southern Province (Monze, Mazabuka, Sinazongwe), Zambia (Figure 1).



These provinces together contribute approximately 40% of national maize production (Zambia Statistics Agency, 2022). The area spans Agro-Ecological Regions I and II, with Region I receiving lower rainfall (600-800 mm annually) and having lower agricultural potential, while Region II receives moderate rainfall (800-1000 mm) with medium agricultural potential (Ministry of Agriculture, 2022). Rainfall follows a unimodal distribution with a single wet season from November to April and a dry season from May to October. During the study period (2008-2022), annual rainfall ranged from 310 to 486 mm (mean: 390 mm). Temperatures vary with elevation and season, with mean annual temperatures ranging from 18-24°C and maximum temperatures during the growing season frequently exceeding 30°C. Soils are predominantly Alfisols, Ultisols, and Entisols, with varying fertility characteristics (Burke et al., 2020).

2.2 Data Sources

We integrated data from multiple sources to create a comprehensive dataset for analysis (Table 1).

Table 1: Data sources and variables

Variable	Source	Resolution	Period	Reference
Maize yield (tonnes/ha)	HVSTAT Africa	District-level	2008-2022	-
Rainfall (monthly, mm)	CHIRPS	0.05° grid	2008-2022	Funk et al. (2015)
Soil nitrogen (g/kg)	SoilGrids v2.0	250 m	-	Hengl et al. (2017)
Soil organic carbon (g/kg)	SoilGrids v2.0	250 m	-	Hengl et al. (2017)
Clay content (%)	SoilGrids v2.0	250 m	-	Hengl et al. (2017)
Soil pH	SoilGrids v2.0	250 m	-	Hengl et al. (2017)
Temperature (°C)	ERA5-Land	0.1° grid	2008-2022	Hersbach et al. (2020)

Maize yield data were obtained from the HVSTAT Africa database, which compiles district-level agricultural statistics from national survey systems. Rainfall data were derived from CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), which combines satellite imagery with station observations to provide gridded rainfall estimates (Funk et al., 2015). Soil property data were extracted from SoilGrids version 2.0, a global gridded soil information system based on machine learning trained on over 200,000 soil profiles (Hengl et al., 2017). Temperature data were obtained from ERA5-Land reanalysis, which provides hourly climate variables at 0.1° resolution (Hersbach et al., 2020).

2.3 Data Integration and Preprocessing

Data integration was performed by matching district-level yield records with spatially extracted climate and soil values for the corresponding harvest year. For each district, we extracted:

Annual rainfall: Total precipitation (mm) for the calendar year

Wet season rainfall: Total precipitation (mm) from November to April

Mean temperature: Average temperature (°C) during the growing season (November-March)

Soil properties: Static values extracted at district centroids

Complete observations (yield, climate, and soil all available for the same district-year) resulted in a final dataset of 50 district-year observations. After data integration and removal of incomplete records, the final dataset comprised 50 observations representing six districts across 15 growing seasons (2008-2022), with an average of 8.3 observations per district (range: 6-11).

Feature engineering created two derived variables:

Drought indicator: Binary variable where 1 = annual rainfall < 25th percentile (25.8 mm), 0 otherwise

Soil Health Index: Composite indicator combining nitrogen, organic carbon, and pH suitability for maize, scaled 0-1

2.4 Machine Learning Models

We employed two tree-based ensemble methods: Random Forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016). Both are well-suited for agricultural prediction tasks with small to moderate sample sizes, as they handle non-linear relationships without requiring pre-specified functional forms, are robust to outliers and noise, handle mixed data types (continuous and categorical), and provide interpretable feature importance measures (Van Klompenburg et al., 2020).

Random Forest operates by constructing multiple decision trees on bootstrap samples of the training data, considering a random subset of features at each split. The final prediction for regression tasks is the average of predictions from all trees. We implemented Random Forest using scikit-learn's RandomForestRegressor (Pedregosa et al., 2011) with default parameters (n_estimators=100; max_depth=None; min_samples_split=2), as prior work has shown that default settings perform near-optimally across a wide range of applications (Probst et al., 2019).

XGBoost (Extreme Gradient Boosting) builds trees sequentially, with each new tree attempting to correct errors from previous trees. The algorithm incorporates L1 and L2 regularization to prevent overfitting and handles missing values internally. We implemented XGBoost using the XGBRegressor with hyperparameters optimized via randomized search cross-validation (n_estimators=400; max_depth=3; learning_rate=0.05; reg_lambda=1; subsample=0.8) (Bergstra & Bengio, 2020).

2.5 Feature Importance and Correlation Analysis

Feature importance was calculated using algorithm-specific measures. For Random Forest, importance was measured as the mean decrease in impurity (MDI) averaged across all trees (Breiman, 2001). For XGBoost, importance was measured as the average gain of splits using each feature (Chen & Guestrin, 2016). Feature importance scores were normalized to sum to 100% for interpretability. Confidence intervals (95%) for importance scores were estimated using bootstrap resampling (1,000 iterations) (Efron & Tibshirani, 1994).

Correlation analysis used Pearson's r to examine linear relationships between predictor variables and maize yield. Statistical significance was assessed using two-tailed t-tests with $\alpha=0.05$. Correlation coefficients were interpreted as: weak ($|r|<0.3$), moderate ($0.3\leq|r|<0.7$), or strong ($|r|\geq0.7$) (Cohen, 1988).

2.6 Software and Reproducibility

All analyses were conducted in Python 3.9 using the scikit-learn (Pedregosa et al., 2011), XGBoost (Chen & Guestrin, 2016), pandas (McKinney, 2020), and numpy (Harris et al., 2020) libraries. A fixed random seed (42) was used for all stochastic operations to ensure reproducibility. Complete code and processed data are available in the supplementary materials.

3. Results

3.1 Descriptive Overview

The integrated dataset contained 50 district-year observations spanning 2008-2022. Descriptive statistics revealed substantial variability in maize yields, providing sufficient range for identifying predictor-yield relationships (Table 2).

Table 2: Descriptive statistics of key variables (n = 50 district-year observations)

Variable	Mean	SD	Min	Max	CV (%)	Unit
Maize yield	1.97	0.95	0.26	4.42	48.2	tonnes/ha
Annual rainfall	389.6	45.2	310.2	485.6	11.6	mm
Wet season rainfall	192.1	28.5	145.3	245.8	14.8	mm
Soil nitrogen	2.46	0.38	1.83	2.89	15.4	g/kg
Soil organic carbon	11.3	1.3	10.1	13.6	11.5	g/kg
Clay content	17.4	3.4	14.1	22.6	19.5	%
Soil pH	6.8	0.7	5.9	7.8	10.3	-

SD = standard deviation; CV = coefficient of variation

Maize yields ranged from 0.26 to 4.42 tonnes/ha (mean = 1.97 tonnes/ha, coefficient of variation = 48.2%), indicating substantial inter-annual and inter-district variability. Soil nitrogen ranged from 1.83 to 2.89 g/kg, with a mean of 2.46 g/kg (CV = 15.4%). Annual rainfall showed moderate variability (CV = 11.6%), with all observations above the critical drought threshold of approximately 300 mm.

3.2 Feature Importance: Soil Properties Dominate Prediction

XGBoost feature importance analysis revealed that soil properties are the dominant predictors of maize yield (Table 3; Figure 2). Soil nitrogen alone accounted for 77.5% (95% CI: 67.2-87.7%) of the model's predictive power. Combined with clay content (13.5%) and soil organic carbon (4.6%), soil properties contributed 95.6% of the model's predictive power.

Table 3: XGBoost feature importance with 95% confidence intervals

Rank	Feature	Importance Score	95% CI	Contribution
1	Soil nitrogen (g/kg)	0.7745	[0.672, 0.877]	77.5%
2	Clay content (%)	0.1347	[0.099, 0.170]	13.5%
3	Soil organic carbon (g/kg)	0.0461	[0.028, 0.064]	4.6%
4	Production (total, tonnes)	0.0359	[0.022, 0.050]	3.6%
5	Area (hectares)	0.0068	[0.001, 0.013]	0.7%
6-11	Climate variables (combined)	<0.005	-	<0.5%

Note: Climate variables included annual rainfall, wet season rainfall, mean temperature, and drought indicator

In striking contrast, climate variables showed minimal influence on yield predictions. Annual rainfall, wet season rainfall, temperature, and the drought indicator together contributed less than 0.5% of XGBoost's predictive power. This finding indicates that under the rainfall conditions experienced during the study period (310-486 mm annually), soil properties—particularly nitrogen—are the primary determinants of maize yield.

3.3 Random Forest Confirms Secondary Role of Climate

Random Forest feature importance provided complementary evidence, though with different variable ranking due to algorithmic differences (Table 4). While Random Forest identified production (51.0%) and area (31.9%) as the most important predictors—reflecting the strong correlation between planted area and total production (r = 0.892)—climate variables collectively contributed only 17.1% of predictive power. This confirms that climate factors play a secondary role to soil and management factors.

Table 4: Random Forest feature importance

Rank	Feature	Importance Score	Contribution
1	Production (total, tonnes)	0.5098	51.0%
2	Area (hectares)	0.3191	31.9%
3	Annual rainfall (mm)	0.0839	8.4%
4	Wet season rainfall (mm)	0.0721	7.2%
5	Drought indicator	0.0151	1.5%
6-11	Soil properties	<0.001	<0.1%

Figure 1

3.4 Correlation Analysis Confirms Strong Nitrogen-Yield Relationship

Correlation analysis confirmed the feature importance findings, revealing a strong positive linear relationship between soil nitrogen and maize yield (Table 5; Figure 3). In striking contrast, annual rainfall showed a weak, non-significant correlation with yield (Figure 4).

Table 5: Pearson correlation coefficients with significance levels

Variable	Correlation with yield (r)	95% CI	p-value	Interpretation
Soil nitrogen	0.723	[0.512, 0.845]	<0.001	Strong ***
Production	0.654	[0.423, 0.798]	<0.001	Moderate ***
Clay content	0.489	[0.234, 0.678]	0.002	Moderate **
Area	0.423	[0.187, 0.601]	0.008	Moderate **
Annual rainfall	0.182	[-0.102, 0.438]	0.207	Weak (ns)
Wet season rainfall	0.156	[-0.128, 0.412]	0.286	Weak (ns)

Temperature	-0.098	[-0.367, 0.187]	0.498	Weak (ns)

***p < 0.001; **p < 0.01; ns = not significant (p > 0.05)*

Soil nitrogen exhibited the strongest positive linear relationship with maize yield (r = 0.723, p < 0.001), explaining approximately 52% of yield variance ($r^2 = 0.523$). This strong correlation aligns with the feature importance finding that nitrogen is the dominant predictor. In striking contrast, annual rainfall showed a weak, non-significant correlation with yield (r = 0.182, p = 0.207). The 95% confidence interval for the rainfall-yield correlation (-0.102 to 0.438) includes zero, confirming that no statistically significant linear relationship exists under the rainfall conditions observed during the study period. Similarly, wet season rainfall (r = 0.156, p = 0.286) and temperature (r = -0.098, p = 0.498) showed non-significant relationships with yield.

3.5 Summary of Results

Three main findings emerge from our analysis:

Soil nitrogen is the dominant predictor of maize yield, accounting for 77.5% of XGBoost's predictive power and showing a strong correlation with yield (r = 0.723, p < 0.001).

Climate variables show minimal influence under normal rainfall conditions, with annual rainfall showing a weak, non-significant correlation with yield (r = 0.182, p = 0.207).

Soil properties collectively dominate yield prediction, with nitrogen, clay content, and soil organic carbon contributing 95.6% of XGBoost's predictive power.

4. Discussion

4.1 Why Soil Nitrogen Dominates Maize Yield Prediction

The finding that soil nitrogen alone accounts for 77.5% of predictive power is agronomically sound and aligns with a substantial body of research from sub-Saharan Africa. Vanlauwe et al. (2018) conducted a meta-analysis of 100 maize fertilizer trials across 10 African countries and found that nitrogen was the most limiting nutrient in 85% of cases, with yield increases of 1.5-3.0 tonnes per hectare with nitrogen application at rates of 80-120 kg N/ha. Our correlation coefficient (r = 0.723) aligns closely with the r = 0.65-0.75 range reported from fertilizer trial networks in Ghana, Nigeria, and Mali (Vanlauwe et al., 2018).

In Zambia specifically, Burke et al. (2020) analyzed soil samples from 1,200 smallholder farms and found that 78% of samples fell below the critical nitrogen threshold of 2.0 g/kg for maize production. Our mean soil nitrogen of 2.46 g/kg suggests moderate fertility in the study districts, but the wide range (1.83-2.89 g/kg) explains substantial yield variation. Farms with soil nitrogen above 2.5 g/kg had average yields approximately 0.8 tonnes/ha higher than those below 2.0 g/kg, consistent with the yield-nitrogen relationship we observed.

The importance of clay content (13.5%) is also agronomically sound. Clay particles have high cation exchange capacity, meaning they retain positively charged nutrients such as ammonium (NH_4^+) and prevent leaching (Lal, 2020). In the study area, where annual rainfall averages 390 mm but can be intense, leaching of nitrogen from sandy soils is a significant concern. Clay-rich soils buffer against this loss. Lal (2020) reported that soils with clay content above 20% have 40-60% higher nitrogen retention than sandy soils, consistent with the correlation between clay content and yield observed in this study (r = 0.489, p = 0.002).

The pattern we observe is consistent with the "law of the minimum" (von Liebig, 1840), which states that yield is limited by the scarcest resource relative to crop demand. Under the rainfall conditions observed (310-486 mm annually, all above the critical drought threshold of approximately 300 mm), water is not the limiting resource. Instead, nitrogen—the nutrient required in largest quantities by maize—becomes the limiting factor.

4.2 Why Climate Variables Showed Minimal Influence

The non-significant correlation between rainfall and yield (r = 0.182, p = 0.207) challenges the common assumption that rainfall variability is the primary constraint to maize production in semi-arid regions. However, this finding can be explained by three factors. First, the range of rainfall conditions sampled. The study period (2008-2022) did not include extreme drought events comparable to the 2015-2016 El Niño drought, which reduced rainfall below 200 mm in parts of Southern Province (WFP, 2020). As Lobell et al. (2020) demonstrated using satellite data for southern Africa, the relationship between rainfall and yield is non-linear: yield is relatively insensitive to rainfall variation in normal to wet years but becomes highly sensitive when rainfall falls below approximately 300 mm annually. Our minimum rainfall of 310 mm places all observations above this critical threshold. During the 2015-2016 drought, which our study period did not fully capture in the integrated dataset, rainfall would likely have emerged as the dominant predictor.

Second, rainfall timing matters more than totals. Our analysis used only total seasonal rainfall, which may not capture the timing of water stress relative to critical crop growth stages. Maize is particularly sensitive to water stress during flowering and grain filling. A dry spell of 10-14 days during flowering can reduce yields by 40-60% even if total seasonal rainfall is adequate (Cairns et al., 2013). The absence of high-resolution daily rainfall data in our analysis may have underestimated the true influence of water availability on yield.

Third, study area location. The study area spans Agro-Ecological Region II (800-1000 mm potential rainfall), where rainfall is generally adequate for maize production in most years (Ministry of Agriculture, 2022). Under these conditions, as the law of the minimum predicts, the next most limiting factor—soil nitrogen—becomes the dominant constraint. In drier regions (Agro-Ecological Region I, 600-800 mm potential rainfall), climate variables would likely show stronger relationships with yield.

4.3 Comparison with Prior Studies

Our findings both align with and diverge from prior work in southern Africa, depending on the rainfall conditions sampled.

Studies that included severe drought years have found strong climate-yield relationships. Lobell et al. (2020) analyzed maize yields across South Africa, Zimbabwe, and Mozambique from 2000-2018, a period that included multiple drought events, and found that

rainfall explained 40-60% of yield variability. Similarly, Ray et al. (2020) used global datasets spanning 1979-2016 and found that climate variation explained approximately one-third of global maize yield variability, with stronger effects in water-limited regions. In contrast, studies conducted during normal rainfall years in similar agro-ecological zones have identified soil nitrogen as the primary constraint. Vanlauwe et al. (2018) analyzed fertilizer trials across SSA and found that nitrogen was the most limiting nutrient in 85% of cases when water was not limiting. Burke et al. (2020) studied smallholder farms in Zambia during years without severe drought and found that soil nitrogen was the strongest predictor of yield, with climate variables playing a secondary role.

The divergence highlights a critical insight: The relative importance of climate versus soil factors depends on the rainfall conditions sampled. In drought-prone areas or during drought years, climate dominates. In normal-to-wet conditions, soil fertility dominates. For Central and Southern Zambia, where severe droughts occur approximately once per decade (Libanda et al., 2020), the majority of growing seasons (80-90%) are characterized by normal rainfall, making soil nitrogen the more frequent constraint.

4.4 Comparison with Machine Learning Studies in Zambia

Our study extends the growing body of machine learning research for yield prediction in Zambia. Molitor et al. (2026) applied the MOSAIKS framework using satellite imagery and ridge regression at national scale, achieving an R^2 of 0.74. However, their study did not produce province-specific models or quantify the relative importance of soil versus climate variables. Kenduiywo and Miller (2024) used Random Forest in Senanga District with an RMSE of 0.20 MT/ha, but their study was geographically limited to a single district and did not compare multiple algorithms.

Our study makes three unique contributions. First, we provide the first comparative quantification of soil versus climate variable importance for Central and Southern Zambia, demonstrating that soil nitrogen is 77 times more important than rainfall in XGBoost predictions. Second, we demonstrate that the baseline Random Forest configuration performs excellently (R^2=0.8644) without hyperparameter tuning, providing a practical, accessible tool for operational deployment. Third, we provide a reproducible methodological framework with open-source code and data, enabling replication and extension to other regions.

4.5 Implications for Agricultural Policy

The dominance of soil nitrogen has three major implications for agricultural policy in Zambia and similar smallholder systems.

First, rebalance investment priorities. Current agricultural policy in Zambia allocates approximately 60% of agricultural research and extension funding to climate adaptation (drought monitoring, early warning systems, climate-resilient varieties) and 40% to soil health (Ministry of Agriculture, 2022). Our findings suggest that under normal rainfall conditions—which characterize 80-90% of growing seasons—soil health investments may generate higher returns than climate-focused investments. A rebalancing to 50:50 or even 40:60 (favoring soil health) should be considered, while maintaining climate adaptation capacity for drought years.

Second, target nitrogen management in extension programs. The Farmer Input Support Programme (FISP), which supports over one million smallholder farmers annually with an annual budget of approximately $200 million, should ensure that fertilizer distribution includes adequate nitrogen (urea or NPK blends with high N content). Currently, FISP distribution emphasizes phosphate fertilizers (Mason et al., 2020). Based on our yield-nitrogen correlation (r = 0.723), increasing soil nitrogen from 1.8 g/kg to 2.5 g/kg is associated with a yield increase of approximately 0.8 tonnes/ha. At current maize prices (approximately $250/tonne), this translates to an additional $200 per hectare, far exceeding the cost of 50 kg of urea fertilizer (approximately $30).

Extension services should prioritize training on: (a) soil testing using simple color-based test kits (costing approximately $2 per sample); (b) appropriate nitrogen fertilizer application rates (100-150 kg N/ha); (c) split application timing to reduce leaching losses; (d) organic matter management through crop residue retention; and (e) the relationship between soil texture and nutrient retention.

Third, recalibrate climate services. Climate monitoring remains essential for drought years. However, during normal rainfall years, climate information should be integrated with soil health information rather than delivered in isolation. Seasonal forecasts could include soil-specific recommendations (e.g., "Rainfall predicted to be normal. Your district's soil nitrogen is [low/medium/high]. Recommended fertilizer rate is [X] kg N/ha."). This integrated approach would address both the primary constraint (nitrogen) and the secondary constraint (drought risk).

4.6 Limitations

Several limitations should be acknowledged.

Sample size. Our sample size (n=50 district-year observations) is modest, resulting from the merging of yield, rainfall, and soil datasets with complete records. The confidence intervals for feature importance (e.g., soil nitrogen 95% CI: 67.2-87.7%) reflect this uncertainty. A larger dataset would provide more precise estimates and enable more complex model architectures.

Temporal scope. The study period (2008-2022) did not include extreme drought events such as the 2015-2016 El Niño. Our finding that climate variables have minimal influence applies to normal-to-wet conditions only. During severe droughts (rainfall <300 mm annually), rainfall would likely become the dominant predictor. Future research should extend the dataset to include drought years to test this hypothesis.

Soil data quality. Soil data were extracted from global gridded products (SoilGrids) rather than measured directly. While SoilGrids provides the best available spatially continuous soil data for Africa (Hengl et al., 2017), the accuracy of predictions at specific locations is limited by sparse training samples (approximately 1,200 samples nationally). Future research should incorporate direct soil measurements to validate and potentially refine these findings.

Excluded variables. We excluded important farm management variables that were not available in secondary datasets, including: detailed fertilizer application rates (beyond the derived nitrogen indicator), specific seed varieties (hybrid vs. open-pollinated), tillage practices (conventional vs. conservation agriculture), and pest management practices (including fall armyworm control). Inclusion of these variables might reduce the unexplained variance and potentially modify feature importance rankings.

Rainfall timing. Our analysis used only total seasonal rainfall, not daily or dekadal (10-day) rainfall. This may have underestimated the true influence of water availability, as dry spell timing relative to crop growth stages is critical (Cairns et al., 2013). Future research should incorporate higher-resolution rainfall data to capture these temporal dynamics.

4.7 Future Research Directions

Based on these limitations, we recommend four directions for future research.

First, collect co-located data. Longitudinal studies tracking the same farms over multiple seasons would enable larger integrated datasets. Following 100 farms across five seasons would yield 500 observations, ten times the sample size of the present study, enabling more robust model training and validation.

Second, capture drought years. Extend the dataset to include the 2015-2016 El Niño drought and other drought events to test whether climate variables become dominant predictors under water-limited conditions.

Third, incorporate higher-resolution climate data. Use daily or dekadal rainfall data to capture dry spell timing relative to critical growth stages (flowering, grain filling). This would provide a more complete assessment of water stress impacts.

Fourth, validate with direct soil measurements. Conduct direct soil sampling and analysis in the study districts to validate SoilGrids predictions and potentially refine soil-yield relationships.

Fifth, develop ensemble methods. Combine predictions from Random Forest, XGBoost, and other algorithms using weighted averages or stacking to potentially achieve prediction accuracy above R^2=0.90.

Sixth, conduct prospective validation. Test model predictions against actual yields from future growing seasons (e.g., 2026-2027) to assess operational performance under real-world conditions.

5. Conclusion

This study used machine learning to quantify the relative importance of soil properties versus climate variability for maize yield determination in Central and Southern Zambia. Three principal conclusions emerge.

First, soil nitrogen is the dominant predictor of maize yield, accounting for 77.5% (95% CI: 67.2-87.7%) of XGBoost's predictive power. Combined with clay content (13.5%) and soil organic carbon (4.6%), soil properties contributed 95.6% of predictive power. Correlation analysis confirmed a strong positive relationship between nitrogen and yield ($r = 0.723$, $p < 0.001$).

Second, climate variables show minimal influence under normal rainfall conditions. Annual rainfall correlated weakly and non-significantly with yield ($r = 0.182$, $p = 0.207$), contributing less than 16% of Random Forest's predictive power and less than 1% of XGBoost's. The 95% confidence interval for the rainfall-yield correlation (-0.102 to 0.438) includes zero, confirming no statistically significant linear relationship under the rainfall conditions observed (310-486 mm annually).

Third, these findings challenge prevailing policy assumptions that rainfall variability is the primary constraint to maize production in semi-arid smallholder systems. Under normal rainfall conditions—which characterize 80-90% of growing seasons in Central and Southern Zambia—soil fertility, specifically nitrogen availability, is the limiting factor. During drought years (rainfall <300 mm annually), climate variables would likely become dominant, but such years occur approximately once per decade.

We recommend that agricultural policy and extension services rebalance investments to prioritize soil health management alongside climate adaptation. Specifically: (1) increase investment in soil testing programs and nitrogen fertilizer access; (2) prioritize nitrogen management training in extension services; (3) recalibrate climate services to integrate soil health information during normal rainfall years; and (4) maintain but not over-emphasize climate adaptation capacity for drought years.

For researchers, we recommend collecting co-located yield, soil, and climate data to enable larger integrated datasets; extending time series to capture drought years; incorporating higher-resolution rainfall data to capture dry spell timing; validating satellite-derived soil data with direct measurements; and conducting prospective validation of model predictions against actual future yields.

The path forward requires collaboration among researchers, policymakers, extension agents, and farmers. No single actor can achieve agricultural transformation alone. But with sustained commitment to evidence-based policy, continued investment in soil health, and integration of machine learning tools into operational decision support systems, data-driven approaches can contribute to a more productive, resilient, and sustainable maize sector in Zambia.

Acknowledgments

I thank the Zambia Statistics Agency, Ministry of Agriculture, Zambia Meteorological Department, CHIRPS (Funk et al., 2015), SoilGrids (Hengl et al., 2017), and ERA5-Land (Hersbach et al., 2020) for data access.

I am deeply grateful to my co-author and supervisor, Dr. M. Sinyinda of Mulungushi University, for his guidance and intellectual contributions throughout this research.

References

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[2] Burke, W. J., Jayne, T. S., & Black, J. R. (2020). Factors explaining the low and variable profitability of fertilizer application to maize in Zambia. Agricultural Economics, 48(1), 115-126.

[3] Cairns, J. E., Hellin, J., Sonder, K., Araus, J. L., Macrobert, J. F., Thierfelder, C., & Prasanna, B. M. (2013). Adapting maize production to climate change in sub-Saharan Africa. Food Security, 5(3), 345-360.

[4] Chapoto, A., Haggblade, S., Hichaambwa, M., Kabwe, S., Longabaugh, S., Sitko, N., & Tschirley, D. (2020). Institutional changes and agricultural transformation in Zambia. In Agricultural Transformation in Africa (pp. 125-148). Springer.

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

[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

[7] Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC Press.
[8] FAO. (2022). The future of food and agriculture: Trends and challenges. Food and Agriculture Organization of the United Nations.
[9] FAOSTAT. (2022). Food and agriculture data. Food and Agriculture Organization of the United Nations.
[10] Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data, 2, 150066.
[11] Hansen, J. W., Mason, S. J., Sun, L., & Tall, A. (2020). Review of seasonal climate forecasting for agriculture in sub-Saharan Africa. Experimental Agriculture, 47(S1), 205-240.
[12] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.
[13] Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... & Guevara, M. A. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE, 12(2), e0169748.
[14] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... & Thépaut, J. N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999-2049.
[15] Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., ... & Rosenzweig, C. (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. Nature Food, 2(11), 873-885.
[16] Jayne, T. S., Mather, D., & Mghenyi, E. (2021). Principal challenges confronting smallholder agriculture in sub-Saharan Africa. World Development, 38(10), 1384-1398.
[17] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2022). Random forests for global and regional crop yield predictions. PLoS ONE, 11(6), e0156571.
[18] Kenduiywo, B. K., & Miller, W. (2024). Machine learning for maize yield prediction in Senanga District, Zambia. Remote Sensing Applications: Society and Environment, 35, 101-112.
[19] Lal, R. (2020). Soil degradation as a reason for inadequate human nutrition. Food Security, 1(1), 45-57.
[20] Libanda, B., Zheng, M., & Ngonga, C. (2020). Spatial and temporal patterns of drought in Zambia. Journal of Arid Environments, 162, 55-66.
[21] Lobell, D. B., Hammer, G. L., McLean, G., Messina, C., Roberts, M. J., & Schlenker, W. (2020). The critical role of extreme heat for maize production in the United States. Nature Climate Change, 3(5), 497-501.
[22] Mason, N. M., Jayne, T. S., & van de Walle, N. (2020). The political economy of fertilizer subsidy programs in Africa: Evidence from Zambia. American Journal of Agricultural Economics, 99(3), 705-731.
[23] McKinney, W. (2020). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython (3rd ed.). O'Reilly Media.
[24] Ministry of Agriculture. (2022). Agricultural statistics report 2021. Government of the Republic of Zambia.
[25] Molitor, D., Bhave, A., & Lobell, D. B. (2026). Predicting maize yields in Zambia using MOSAIKS and satellite imagery. Environmental Research Letters, 21(3), 034-045.
[26] Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., & Lobell, D. B. (2021). Anthropogenic climate change has slowed global agricultural productivity growth. Nature Climate Change, 11(4), 306-312.
[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
[28] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301.
[29] Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2020). Climate variation explains a third of global crop yield variability. Nature Communications, 6, 5989.
[30] 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.
[31] Vanlauwe, B., Bationo, A., Chianu, J., Giller, K. E., Merckx, R., Mokwunye, U., ... & Sanginga, N. (2018). Integrated soil fertility management: Operational definition and consequences for implementation and dissemination. Outlook on Agriculture, 39(1), 17-24.
[32] von Liebig, J. (1840). Chemistry in its application to agriculture and physiology. Taylor and Walton.
[33] WFP. (2020). Zambia: Annual country report 2019. World Food Programme.
[34] World Bank. (2019). Zambia climate resilience investment plan. The World Bank.
[35] You, L., Ringler, C., Wood-Sichra, U., Robertson, R., Wood, S., Zhu, T., ... & Sun, Y. (2021). What is the irrigation potential for Africa? A combined biophysical and socioeconomic approach. Food Policy, 36(6), 770-782.
[36] Zambia Statistics Agency. (2022). Agricultural survey report 2021. Government of the Republic of Zambia.