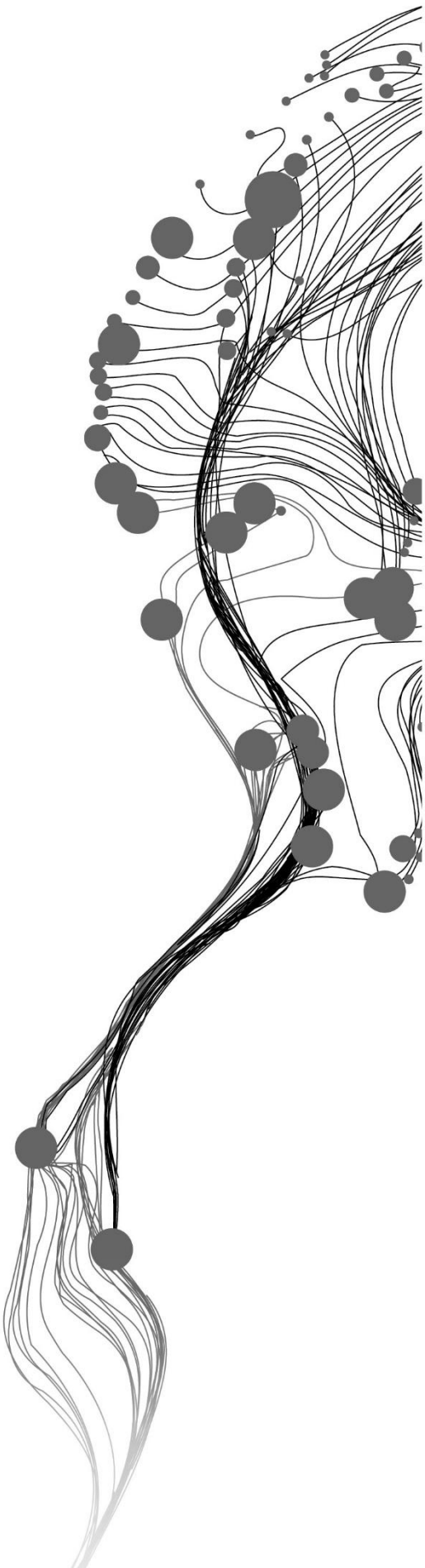
Estimation of Micronutrients in Maize yield in Malawi using Machine Learning and spatially explicit environmental data

XINRAN BIAN
June, 2025

SUPERVISORS:
Dr. M. Belgiu
Dr. M.T. Marshall



Estimation of Micronutrients in Maize yield in Malawi using Machine Learning and spatially explicit environmental data

XINRAN BIAN
Enschede, The Netherlands, June, 2025

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. M. Belgiu
Dr. M.T. Marshall

THESIS ASSESSMENT BOARD:
Prof. dr. ing. C. Persello (Chair)
Dr. R.M. Aguilar Bolivar (External Examiner, ITC-GIP)

ABSTRACT

Micronutrient deficiencies in staple crops contribute to "hidden hunger" across sub-Saharan Africa, yet field measurements of grain nutrient content remain spatially sparse and costly. This study develops a framework for predicting four key micronutrients—calcium (Ca), iron (Fe), zinc (Zn) and selenium (Se)—in Malawian maize grain by combining one-time ground observations with 61 terrain, climate, soil and multi-sensor remote-sensing variables. After quality filtering, georeferenced grain samples were paired with Sentinel-1/2 imagery, CHIRPS rainfall, MODIS land-surface temperature, MERIT DEM derivatives, SoilGrids variables and Global Agro-Ecological Zones (GAEZ). An XGBoost algorithm was trained for each micronutrient using spatial cross-validation approach to address the potential spatial autocorrelation bias. Model's performance varied by nutrient. Specifically, Zn and Ca obtained modest but useful test-set R^2 values of 0.14 and 0.15, whereas Fe reached 0.08 and Se showed low explanatory power. For RMSE, the model achieved test RMSE values of 18.92 for Ca, 25.05 for Fe, 3.67 for Zn, and 0.029 for Se, indicating variable predictive performance across nutrients. Feature-importance analysis highlighted elevation, seasonal precipitation, top-soil pH and organic carbon as among the most important variables for Ca, Fe and Zn estimations, whereas radar-derived canopy metrics proved to be very important for Se estimation. This emphasizes that Se uptake is tightly linked to short-term moisture and redox dynamics. The investigation of the spatial distribution of the maize nutrient revealed a consistent south-to-north decrease in micronutrient levels, aligning with the known gradients in soil pH, organic matter and climate. Despite several limitations, including small sample size, reliance on modelled soil layers and lack of variables related to management factors, the proposed framework shows the advantage of using freely available geodata for estimating maize nutrient content. These outputs can guide site-specific fertiliser recommendations, bio-fortified seed deployment and nutrition programmes, helping local authorities to address hidden hunger and advance Sustainable Development Goal 2. Future work should integrate farm-management surveys, or dynamic moisture–redox proxies to further improve accuracy, particularly for Se and Fe.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my two supervisors, Mariana and Michael. Both provided patient and thoughtful guidance during the most stressful periods of this research. Mariana's kindness and warmth made a profound difference during the most stressful periods of this research. Her gentle and reassuring support helped me realize that this project was not as overwhelming as I had feared. After experiencing a previous setback during my thesis proposal defence, I was uncertain about whether I could carry this work forward. However, both of my supervisors offered invaluable guidance with extraordinary patience and understanding. Their encouragement made me feel less like a student and more like someone genuinely cared for, almost like their own child.

I would also like to sincerely thank my study advisor, Marie-Chantel. During my master's studies, she showed genuine concern not only for my academic progress but also for my personal well-being. In moments when I needed support the most, she offered very much of help, including connecting me with psychologists and continuously following up on my mental health, daily life, and academic status afterward. At times when I tended to avoid or delay addressing issues, she would step in and check on me with care and initiative. Her support has had a significant and lasting impact on my journey, and I am truly grateful for her presence.

I sincerely appreciate my family for their financial and emotional support of all time. Furthermore, I am immensely grateful to my friends and colleagues, to Jorges, and to all the kind and crazy people I have met online. Their encouragement gave me strength and motivation during difficult times. This mutual support has been an unexpected and powerful source of inspiration.

I am grateful as well to the orchestras I have been part of — MSO and SHOT. They have given me invaluable emotional support and a vibrant social environment. I also want to thank for music, especially for orchestra works, which has been a deep and enduring part of my life since childhood. Music is far more than a hobby to me; it carries my memories, friendships, and some of the most meaningful connections in my life. Remembering those musical moments always brings me peace and joy. I feel truly lucky that music has stayed with me as both a passion and a source of strength, and through it, I've met many wonderful people and built lasting memories.

Lastly, I would like to thank myself — for maintaining order in my life and for never giving up. This journey has been as much about research as it has been about personal growth.

TABLE OF CONTENTS

1. Introduction 1
   1.1. Background 1
   1.2. Related work 1
   1.3. Research gap 4
   1.4. Objectives and research questions 5
2. Study area 6
3. Methodology 9
   3.1. Data collection 9
   3.2. Data preprocessing 11
   3.3. XGBoost modelling and optimisation 15
4. Results 18
   4.1. Nutrient spatial distribution maps 18
   4.2. Environmental variable spatial distribution 19
   4.3. Descriptive statistics and spatial distribution of maize grain nutrients 25
   4.4. XGBoost model performance 27
   4.5. Feature importance 31
   4.6. Partial dependence plots 33
5. Discussion 34
   5.1. Interpretation of findings 34
   5.2. Limitations of this study 36
6. Conclusion and future work 37
   6.1. Conclusion 37
   6.2. Future work 37
   6.3. Social impact and policy relevance 38
   6.4. Ethical Considerations 38

LIST OF FIGURES

Figure 1 Study area - Malawi 6
Figure 2 Workflow of integrating multi-source environmental variables for nutrients in maize yield estimation 9
Figure 3 Examples of 100m*100m buffers and maize mask 11
Figure 4 Spatial distribution of log-transformed maize grain micronutrient concentrations in Malawi (Ca, Zn, Fe, Se) 118
Figure 5 Average temperature of maize during growing season in MalawiFigure 4 Spatial distribution of log-transformed maize grain micronutrient concentrations in Malawi (Ca, Zn, Fe, Se) 11
Figure 6 Average precipitation of maize spatial distribution during growing season in Malawi 2011
Figure 7 Elevation (a) and slope (b) spatial distribution in Malawi 21
Figure 8 Soil pH distribution across soil depths in Malawi (0–100 cm, from left to right, up to down: 0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm) 22
Figure 9 SOC distribution (g/kg) across soil depths in Malawi (0–100 cm, from left to right, up to down: 0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm) 23
Figure 10 AEZ of Malawi Based on Climate, Elevation, and Land Limitations 24
Figure 11 Violin and box plots showing the spatial distributions of soil micronutrient concentrations for Ca, Fe, Zn (left), and Se_tripleq (right) across all sampling points 25
Figure 12 Log-transformed distributions of soil micronutrient concentrations (Ca, Fe, Zn, Se) across different AEZs in Malawi, visualized using combined violin and box plots 26
Figure 13 Spatial Clustering of Training Samples and Train/Test Split Distribution Across Malawi 27
Figure 14 Comparison of actual and predicted values of each micronutrient 29
Figure 15 Feature importance ranking of XGBoost model in estimating Ca, Fe, Zn and Se_tripleq concentration in maize (by Gain value from high to low) 32
Figure 16 PDP of the top three features of each element in the XGBoost model (Top 1 to Top 3 from left to right) 33

LIST OF TABLES

Table 1 Vegetation indices 12
Table 2 Model hyperparameters in models tuning with different nutrients as targets 28
Table 3 Overall model performance 30

1. INTRODUCTION

1.1. Background

Micronutrient deficiency, a form of "hidden hunger," undermines essential metabolic functions and tissue repair, currently affecting more than two billion people worldwide, and constitutes a significant barrier to achieving Sustainable Development Goal 2 (Zero Hunger) (WHO, 2023). Sub-Saharan Africa exhibits particularly high risks associated with hidden hunger, largely due to local diets heavily dependent on staple crops such as maize. These staple crops often lack essential minerals including Fe, Zn, and vitamin A (Wessells & Brown, 2012; Gödecke et al., 2018). In Malawi, maize dominates both agricultural production and dietary consumption, contributing more than 80% of caloric intake (Galani et al., 2021). However, changes in environmental factors such as soil, climate and topography have exacerbated micronutrient deficiencies in recent decades, further compromising maize nutritional quality (Voss, 1998; Botoman et al., 2022a). Traditional lab-based nutrient assessments remain costly and spatially limited, while most agricultural and public health policies continue to emphasize caloric yield over nutritional diversity (Yuan et al., 2023; Ichami et al., 2022; Pingali & Sunder, 2017). To align with SDG targets, there is an urgent need for innovative tools that use machine learning and spatially explicit environmental data (e.g., soil pH, precipitation trends, and land use patterns) to predict micronutrient levels in crops, enabling targeted interventions in resource-limited agricultural systems.

Addressing micronutrient deficiencies effectively demands a comprehensive understanding of how environmental variations influence maize nutritional quality across Malawi. Maize quality in Malawi, as a crucial staple crop, is influenced by a combination of genetic, environmental, and agronomic factors (Ennen et al., 2021). Environmental determinants, such as soil properties including pH and organic matter content, significantly influence the bioavailability of Fe and Zn (Bouis & Saltzman, 2017), while regional differences in temperature and rainfall alter nutrient uptake efficiency (Gashu et al., 2021). Critically, these environmental drivers interact non-linearly, creating complex spatial patterns that traditional statistical methods may fail to capture. In contrast, machine learning methods are uniquely equipped to address this limitation. By integrating multiple environmental datasets (e.g., soil data, climate data, and satellite-derived land use metrics), machine learning based predictive models can clarify which factors exert the most substantial influence on micronutrient variability, and how micronutrients are spatially distributed, thus bridging the scale mismatch between localized crop nutrition data and national-level policy interventions.

1.2. Related work

1.2.1. Micronutrients and maize productivity in Africa

Micronutrients such as Ca, Fe, Zn, and Se are essential for maize growth and grain quality. While macronutrient deficiencies (nitrogen (N), phosphorus (P), potassium (K)) are often addressed, shortages of micronutrients can also significantly constrain yields in Sub-Saharan Africa (Aliyu et al., 2021). For instance, apart from N, P, K, deficiencies of elements like Zn, Copper, Boron, and Sulphur have been reported to limit maize productivity in African soils (Aliyu et al., 2021). These nutrients play critical roles in plant physiology: Fe is required for chlorophyll formation and metabolic enzymes, Zn for enzyme activation and hormone regulation, and Ca for cell wall formation and stress resilience (Grabowski et al., 2024). In practice, hidden hunger for micronutrients is common in Africa since many African soils are inherently low in these elements, leading to both reduced crop yields and nutrient-poor harvests. In

Malawi, more than half of households are at risk of Ca, Zn, or Se deficiencies, reflecting low availability of these nutrients in diets and soils (Joy et al., 2015). Field studies demonstrate the impact of micronutrient management on maize. For example, adding Zn fertilizer in Malawi trials increased grain yield by ~11% and raised the Zn concentration in maize grain by 15% (Grabowski et al., 2024). Similarly, soil type affects the ability of grain to take up nutrients (Joy et al., 2015). These findings highlight that addressing micronutrient limitations (through soil amendments, breeding, or agronomic biofortification) is vital for improving maize productivity and nutritional quality in Africa (Aliyu et al., 2021; Grabowski et al., 2024). Ensuring adequate micronutrient availability can not only boost yields but also enhance the grain's contribution to dietary nutrition, which is critical given maize's role as a staple in the region (Mhlanga et al., 2021; Galani, 2022).

1.2.2. Use of spatially explicit environmental data in agriculture

Agricultural research increasingly uses spatially explicit environmental to analyse and predict crop performance (Chilimba et al., 2011; Botoman et al., 2022a). Traditional methods like laboratory spectroscopy or field sampling, while accurate, are impractical for national-scale assessments due to high costs and sparse spatial coverage (Bouis & Saltzman, 2017). Satellite remote sensing offers a promising alternative by enabling spatially continuous data collection on soil properties (e.g., organic matter, pH), climate variables (e.g., precipitation, temperature), and vegetation health (Bastiaanssen et al., 2000; Lobell et al., 2015). High-resolution satellite imagery, such as from the Sentinel constellation, provides detailed indicators of crop condition across space and time (Drusch et al., 2012; Thenkabail et al., 2017). In a Sahelian agroforestry landscape, for example, multi-temporal Sentinel-2 imagery explained 41–80% of the variation in field-level millet and sorghum yields, demonstrating the power of satellite time-series for yield estimation even in complex smallholder systems (Karlson et al., 2020). Optical data (e.g. Sentinel-2) supply visible and near-infrared bands that enable vegetation indices (NDVI, EVI, etc.) for assessing crop health, while radar data (e.g. Sentinel-1) offer complementary information on crop structure and moisture independent of clouds (Fathi et al., 2023). NDVI (Normalized Difference Vegetation Index) and thermal imagery have been linked to crop nutrient stress in maize systems (Ustin et al., 2020), while hyperspectral sensors can indirectly estimate soil micronutrients like Fe and Zn through spectral signatures (Goswami et al., 2020). Soil property databases like ISRIC's SoilGrids provide gridded layers of soil characteristics (texture, pH, organic carbon, etc.) at fine spatial resolution. Fathi et al. (2023) showed that integrating SoilGrids data with Sentinel-1/2 imagery in a deep-learning model improved maize yield prediction accuracy in the U.S. Corn Belt. The soil maps capture spatial variability in fertility and water-holding capacity that remote sensing alone might miss. Maize grown on calcareous, high-pH soils in Malawi contains significantly higher grain Ca, Zn, and Se levels than maize from the more prevalent acidic soils (Joy et al., 2015). Likewise, gridded climate data are indispensable in crop models. Datasets such as CHIRPS (Climate Hazards group InfraRed Precipitation with Station data) provide daily rainfall estimates at ~5 km resolution across Africa, and have been used to drive yield forecasting models (Lee et al., 2022). Lee et al. (2022) used only climate (rainfall, temperature) and vegetation index inputs to successfully forecast maize yields in multiple African countries, highlighting that spatial rainfall patterns and temperature extremes are key yield drivers. Finally, topographic data from digital elevation models (e.g. the MERIT DEM at 90 m resolution) are often incorporated to account for terrain effects on agriculture. Elevation and its derivatives (slope, wetness index) influence microclimates, soil drainage, and erosion. Accordingly, studies mapping nutrient distribution have included DEM-based features and found that topography significantly affects soil nutrient concentrations (Gohil, 2023). For instance, research in Malawi and Ethiopia showed that areas of higher elevation or particular landscape positions could be linked to higher grain Se or Zn, reflecting underlying soil differences (Gohil, 2023). In summary, the use of spatially explicit datasets – from satellites to soil and climate grids – has become a cornerstone of modern agricultural analyses, enabling more precise, location-specific insights into crop yields and nutrient status.

1.2.3. Machine learning for yield and nutrient estimation

Recent advances in machine learning and geospatial technologies offer transformative potential for crop nutrient estimation. Machine learning techniques have gained prominence in agricultural yield prediction and nutrient mapping due to their ability to handle complex, non-linear relationships in multi-source data. Ensemble tree-based models like Random Forests and Extreme Gradient Boosting (XGBoost) are especially popular. These methods can ingest diverse inputs (remote sensing indices, soil properties, weather variables, management data) and have demonstrated high predictive accuracy in crop studies (Mahesh & Soundrapandiyan, 2024). For example, Asamoah et al. (2024) trained a Random Forest model trained on soil, climate, environmental, and management factors (including fertilizer use) to predict maize yields in Ghana, achieving a model efficiency coefficient of 0.81 for yield prediction (Asamoah et al., 2024). Hybrid PROSAIL-PRO retrieval frameworks coupled with Gaussian Process Regression have been used to estimate aboveground nitrogen content, delivering $R^2 > 0.90$ and full uncertainty quantification (Berger et al., 2020). Back-propagation neural networks applied to over 40,000 paddy soil–rice samples predicted Zn bioaccumulation in rice grains with $R^2 = 0.93$ and a normalized RMSE of 0.21 (Wang et al., 2021). In hyperspectral nutrient mapping of Valencia-orange leaves, Random Forest outperformed Support Vector Machines and other regression methods, achieving R^2 above 0.85 for both macro- and micronutrient predictions (Osco et al., 2020). UAV-based models combining Random Forest and Partial Least Squares Regression with multispectral imagery and weather data estimated wheat shoot nitrogen concentration and the Nitrogen Nutrition Index with R^2 up to 0.82 (Tanaka & Gislum, 2025). Convolutional neural network and BiLSTM architectures integrating Sentinel-1/2 imagery with SoilGrids data have achieved an RMSE of 0.698 t/ha and an index of agreement of 84.7% in U.S. Corn Belt yield prediction, outperforming Random Forest baselines (Olisah et al., 2024). Ensemble methods such as XGBoost, LightGBM, and CatBoost have further surpassed single-tree algorithms in multi-crop yield and nutrient estimation tasks, often achieving $R^2 > 0.80$ in comparative benchmarks (Mahesh & Soundrapandiyan, 2024).

Deep learning approaches are also increasingly applied in agricultural analytics. Convolutional Neural Networks (CNNs) can automatically extract spatial features from imagery (Kattenborn et al., 2021; Srivastava et al., 2021). Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, can capture temporal dependencies in time-series data (Ienco et al., 2017; Khaki et al., 2019). For instance, Fathi et al. (2023) proposed a hybrid 2D-CNN + BiLSTM that integrates Sentinel-1 radar, Sentinel-2 optical, and SoilGrids soil data to predict corn yields in Iowa. By fusing multi-temporal satellite data with static soil information, their deep learning model was able to capture short-term vegetation signals as well as longer-term site fertility effects, yielding more accurate predictions than baseline random forests (Fathi et al., 2023). Similarly, other studies have used CNNs on high-resolution imagery to estimate yields at the field scale, or LSTM-based frameworks to forecast yields mid-season using sequences of weather and NDVI data. These approaches have shown that deep models can match or exceed the accuracy of traditional methods, especially when large datasets are available. Nevertheless, simpler machine learning models remain competitive in many cases. For smaller datasets or more interpretable results, algorithms like Support Vector Machines (SVM) have also been employed for crop yield and quality prediction. For example, researchers in Senegal tested SVM, RF, and ANN models for regional yield prediction and found all machine learning methods yielded improvements over linear models (with the optimal choice varying by region) (Sarr & Sultan, 2023). In general, the literature suggests that no single algorithm is universally best. The performance of XGBoost, Random forests, SVMs, or neural networks depends on the context, but ensemble tree models are a strong starting point for tabular agri-environmental data (Mahesh & Soundrapandiyan, 2024), whereas deep learning is powerful when leveraging rich spatial or temporal remote sensing data (Sarr & Sultan, 2023). Overall, the application of

machine learning has enabled more accurate and scalable estimation of crop yields and even grain nutrient contents, which is a significant advancement for agricultural planning and food security analysis.

1.2.4. Micronutrient estimation in Malawi

Focusing on Malawi and the broader sub-Saharan context, several studies have directly tackled maize yield prediction and assessment of the spatial variation of micronutrient content.. Given maize's status as the national staple, there is high interest in understanding not only how much is produced but also its nutritional value. Ligowe et al. (2022) (as part of the GeoNutrition project) examined how agronomic practices affect maize yield and grain nutrients in Malawi. Their two-season field trials found that conservation agriculture practices (e.g. reduced tillage, residue retention) significantly increased maize grain yield (by 1.2–1.8×) and also enhanced grain Se content (by up to 70% higher) relative to conventional practices. However, the same trials noted varying effects on other micronutrients: grain Zn and Ca were not significantly changed by the treatment, while grain Fe and Mn concentrations actually decreased under conservation agriculture. This illustrates the complex interplay between soil management and micronutrient uptake. It also reinforces the need for spatially explicit studies. For instance, the authors observed that maize from calcareous soil sites had higher Fe, Zn, and Se than maize from acidic soil sites, indicating that underlying soil properties drive micronutrient availability (Galani, 2022). Recently, mapping of micronutrient concentrations in crops has been a research focus in Malawi and neighbouring countries. Botoman et al. (2022b) conducted a country-scale analysis of maize grain Zn in Malawi. Using a linear mixed-effects model, they related grain Zn measurements to an array of predictors including soil properties, climate, and topography. Their model identified soil pH and organic carbon as key predictors confirming the findings by Gashu et al. (2021) that soil pH and soil organic carbon correlate strongly with grain Ca, Fe, Zn, and Se levels in both Malawi and Ethiopia. Topographical factors (elevation, slope) also improved the Zn predictions, suggesting that landscape position influences how much Zn ends up in the grain (Galani, 2022). Importantly, Botoman et al. (2022b) noted that agronomic Zn interventions could have tangible benefits, namely by adding Zn fertilizer the maize Zn content and yields in on-farm trials increased. Complementary work by Gashu et al. (2020) mapped Se in Ethiopian grains, and together these studies form part of a growing effort to create "nutrient maps" for African staples. The insights from Malawi and Ethiopia emphasize that combining spatial data (soil, climate, remote sensing) with modern modelling can reveal nutrient deficiencies and guide interventions (like targeted fertilization or biofortification) to address human micronutrient needs.

1.3. Research gap

While existing literature establishes the link between soil health, environmental conditions, and crop nutrient levels, there remains a notable gap in studies that apply advanced machine learning models to estimate specific micronutrient concentrations in maize. Although prior research has explored the use of satellite imagery, environmental data, and machine learning techniques to assess nutrient status in Ethiopia, these studies have primarily utilized methods such as random forests or traditional statistical approaches (Ofori-Karikari, 2024), rather than developing a more advanced framework for estimating micronutrients in maize based on diverse environmental variables.

This study aims to address the research gap by developing an XGBoost-based model that uses environmental factors derived from satellite imagery to estimate the micronutrient concentration in maize crops across Malawi. While random forests have been commonly used in agricultural modelling due to their robustness with high-dimensional data and their ability to capture non-linear relationships (Breiman, 2001; Belgiu & Drăguţ, 2016), XGBoost's gradient boosting framework provides improved accuracy and computational efficiency, especially in handling structured data with complex interactions and missing values (Chen & Guestrin, 2016). These capabilities make it particularly well-suited for agricultural datasets

where data quality and completeness may vary. The selected environmental factors include soil properties such as pH and organic matter, climate data like temperature and precipitation, and topographical features including elevation and slope. Besides, Global Agro-Ecological Zones (GAEZ) was also introduced as a categorical input feature to capture broad-scale agroecological variability, covering climate, soil potential and land-suitability factors that influence micronutrient availability and uptake in maize. This research is significant in its potential to provide an efficient, scalable tool for nutrient management, enabling the identification of nutrient deficiencies across large agricultural areas. By precisely estimating individual nutrient concentrations (Ca, Fe, Zn, and Se), the model can inform more targeted interventions, such as optimized fertilizer application, to improve nutrient use efficiency, crop yield, and quality.

1.4. Objectives and research questions

The overall goal of this study is to develop and evaluate a framework, to predict maize micronutrient concentrations in Malawi using XGBoost-based machine learning model trained with spatially explicit environmental data.

To achieve this goal, the study will pursue the following specific objectives:

1) To characterize the spatial distribution of the environmental variables and maize micronutrient concentrations across Malawi.

2) To develop and validate an XGBoost machine learning model for predicting maize grain micronutrient levels using spatial environmental datasets.

3) To identify and rank the most influential environmental variables contributing to the variability in maize micronutrient concentrations.

Research questions:

1. How do environmental variables and maize micronutrients vary in Malawi?
2. How accurately does the XGBoost model estimate micronutrient levels in maize grain?
3. What are the most relevant environmental variables influencing micronutrients in maize yield?

2. STUDY AREA

Malawi is a located in southeastern Africa, nestled between Tanzania, Mozambique and Zambia, with its eastern boundary dominated by the long, narrow Lake Malawi and its western border dipping into the Zambezi Rift. The country's relief varies dramatically: steep escarpments and the Nyika Plateau in the north rise above 3,000 m, while the Rift Valley floor and lake shores lie near 37 m above sea level. Central and southern highlands form rolling plateaus at 800–1,200 m. This topographic diversity shapes a tropical to subtropical climate: a rainy season from November to April brings 500–1,800 mm of annual rainfall (increasing with elevation), followed by a cool, dry spell from May to August and a hot, dry interlude in September–October. Mean daily temperatures range from about 15 °C on the highest plateaus to upwards of 30 °C in lowland valleys.

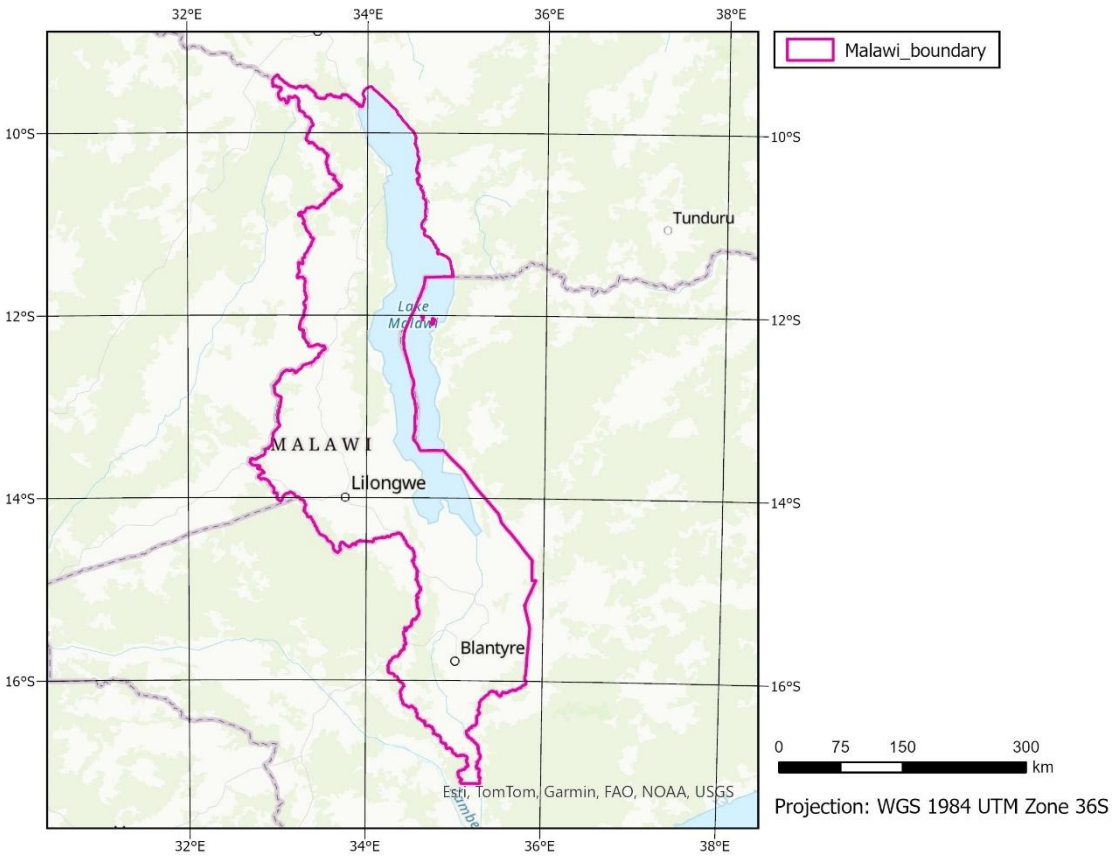


Figure 1 Study area - Malawi

Roughly two-thirds of Malawi's land is devoted to smallholder agriculture, with maize the flagship crop occupying over 60 % of cultivated area and supplying more than half of household calories. Farms average less than two hectares and typify mixed crop–livestock systems; tobacco, cassava and groundnuts are also common but play secondary roles. Soils are largely ferralsols and acrisols—deep and well-drained yet inherently infertile—with alluvial clays and loams confined to floodplains. Chronic acidification (pH < 5.5) and low organic matter foster widespread micronutrient shortages, especially of Zn, boron and manganese, limiting both maize productivity and grain quality.

Environmental pressures compound these natural constraints. Steep-slope cultivation and deforestation for fuelwood accelerate erosion and topsoil loss, while increasingly erratic rains and periodic droughts threaten yields. Socioeconomically, most farming households lack access to credit, quality inputs and

mechanization; landholdings average under one hectare, and labour shortages—intensified by health crises—delay critical planting and weeding operations. A growing body of agronomic research in Malawi has mapped soil fertility hotspots and tested Zn fertilization and organic amendments, yet these field-based studies remain spatially fragmented.

This complex mosaic of topography, climate, soil chemistry and land use make Malawi an ideal arena for machine learning and spatial analysis. By integrating remote sensing (e.g., multispectral imagery, digital elevation models) with ground-truth soil and yield data, predictive algorithms such as random forests or gradient boosting can generate high-resolution maps of soil micronutrient status and maize nutrient uptake. Such models promise to pinpoint zones where targeted interventions—whether micro-dosing of Zn or site-specific liming—will most effectively raise both yields and grain micronutrient concentrations, thereby strengthening food security and nutrition across Malawi's diverse landscapes.

3. METHODOLOGY

To systematically evaluate the soil, climate, topography and multi-source remote sensing factors that affect nutrients in maize yield, this study integrated soil data, climate data, digital elevation models (DEMs), and Sentinel-1/2 data. Feature extraction was performed based on sample points and their buffers, and a data set containing 61 input variables was constructed. The model was constructed and validated under the constraints of corn planting masks. Figure 2 shows the complete method flow chart.

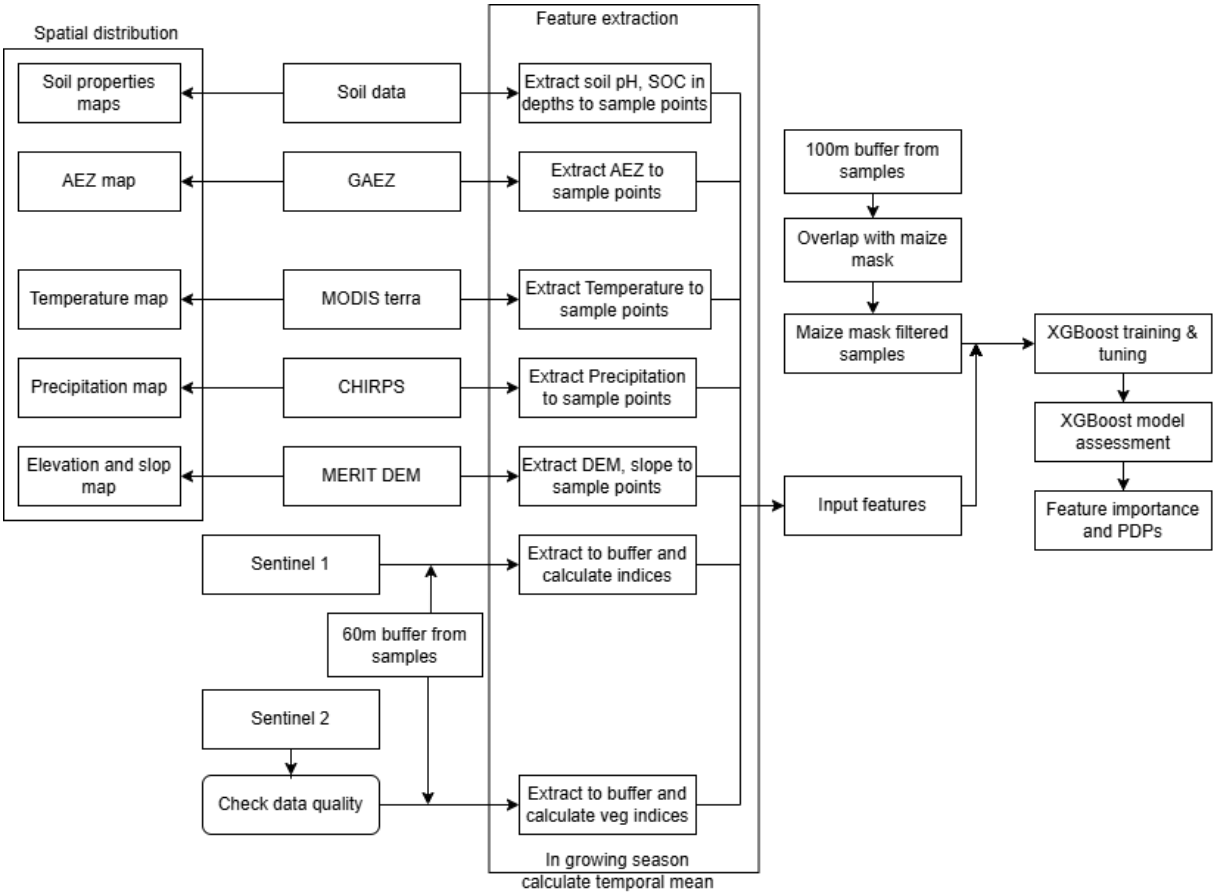


Figure 2 Workflow of integrating multi-source environmental variables for nutrients in maize yield estimation

3.1. Data collection

3.1.1. Micronutrient sample data

This research applied micronutrient data from the GeoNutrition project, which collected grain and soil samples from farmers' fields and grain stores across Malawi during April-June 2018, immediately following the 2017-2018 maize growing season. The sampling was conducted with informed consent from farmers and under ethical approval from the University of Nottingham's School of Sociology and Social Policy Research Ethics Committee (REC; BIO-1819-001 for Malawi). The study protocols were also formally recognized by the Director of Research at Lilongwe University of Agriculture and Natural Resources. The samples provide ground-truth data on concentrations of Ca, Fe, Zn, and Se in maize grain, which serve as the target variables for the machine learning models. In this study, the Se_tripleq values represent Se

concentrations measured using Triple Quadrupole ICP-MS. The samples are georeferenced with XY coordinates in WGS84 (EPSG:4326).

3.1.2. Environmental and spatial data

To capture the environmental factors that potentially influence micronutrient levels in maize, this study incorporates multiple datasets representing various aspects of environmental factors:

Satellite imagery

- Sentinel-2: launched in 2015 (Sentinel-2A) and 2017 (Sentinel-2B), these satellites provide global coverage every 5 days with a resolution of 10 m for RGB and key vegetation bands, and 20 m and 60 m for other bands. Sentinel-2 data is used to compute various vegetation indices that reflect crop health and environmental conditions.
- Sentinel-1: launched in 2014 (Sentinel-1A) and 2016 (Sentinel-1B), these satellites provide radar data every 6 days at 10 m resolution. Sentinel-1 data is particularly valuable for extracting soil moisture information and computing polarimetric indices, which are less affected by cloud cover compared to optical sensors.
- MODIS Terra: with a 1 km resolution and daily coverage, MODIS data is used to extract land surface temperature, an important factor affecting crop growth and nutrient uptake.

Climate Data

CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) dataset provides daily precipitation data at approximately 5.5 km spatial resolution, suitable for analysing rainfall patterns during the 2017-2018 growing season across Malawi.

Soil Data

SoilGrids global soil database (Hengl et al., 2017) provides predictions of soil properties at approximately 250 m resolution, including soil pH and organic carbon content, which are critical factors affecting nutrient availability to plants.

Topographic Data

MERIT DEM (Multi-Error-Removed Improved-Terrain Digital Elevation Model) has a resolution of 90 m resolution. This dataset provides elevation data that has been processed to reduce errors such as vegetation and building biases found in other DEMs. Slope data is derived from the MERIT DEM using Google Earth Engine (GEE), allowing for analysis of how terrain characteristics affect water flow and nutrient distribution.

Administrative Data

GADM (Global Administrative Areas Database) provides country boundaries for Malawi, used for spatial reference.

Malawi Maize Mask for 2017

This is a publicly licensed (Creative Commons Attribution 4.0) dataset produced by the World Bank (World Bank, 2017). Derived from Sentinel-2 satellite imagery and household survey plot labels, this binary GeoTIFF map identifies maize and non-maize areas across Malawi for 2017 at a 10 m spatial resolution.

3.2. Data preprocessing

3.2.1. Defining buffers around available grain nutrient samples and maize-mask filtering

In this study, 100 m × 100 m square buffers were generated around each sample point. The Malawi maize mask 2017 (10 m resolution GeoTIFF) was added into ArcGIS pro, and an intersection operation was performed between each buffer polygon and the maize-classified pixels. Buffers intersecting at least one maize pixel were retained, while those with no overlap were discarded. The resulting set of filtered buffers then provided the spatial framework for all subsequent remote-sensing feature extraction. Figure 3 shows a part of 100m*100m buffers and maize mask.

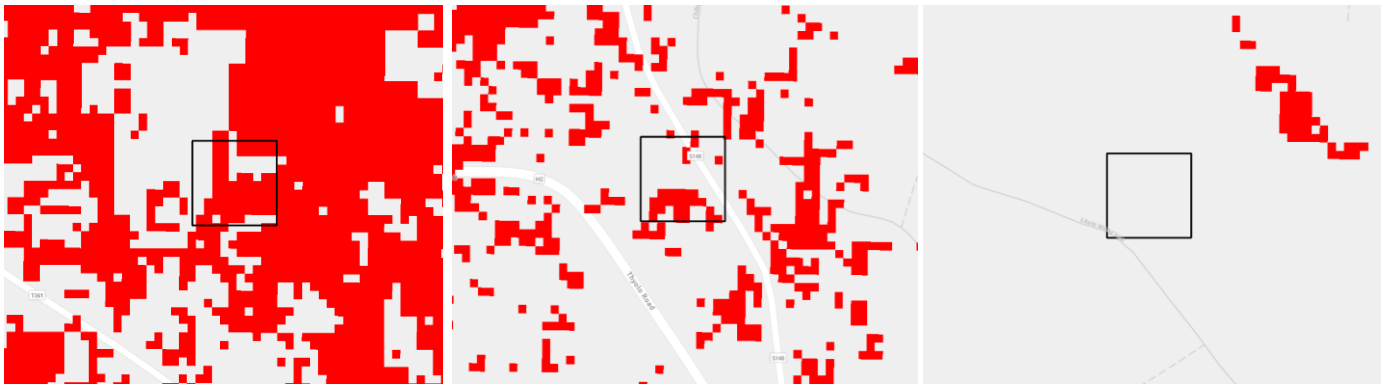


Figure 3 Examples of 100m*100m buffers and maize mask
(From left to right, the buffers in the first and second images were retained, while the buffer in the third image was removed.)

3.2.2. Sentinel-2 data quality assessment and buffer sampling

A spatial buffer of 60m × 60m was established around each field-validated sampling site to match the spatial resolution of Sentinel-2/-1 and reduce errors due to geolocation bias. The entire processing pipeline is based on the S2_SR_HARMONIZED dataset in Google Earth Engine (GEE), and the core can be divided into several steps: first, cloud cover is filtered for each pixel to keep only completely cloud-free observations to ensure the original quality of the spectral data; second, the pixels with less than 50% valid observations during the entire growing season (November 1, 2017 to April 30, 2018) are eliminated through the validity screening mechanism to ensure temporal representativeness; next, the seasonal average value during the period is calculated based on the retained pixels to generate a robust composite image; thereafter, the average reflectance of all pixels in each buffer is statistically calculated, and the NDVI variance is attempted to assess the spatial consistency of the vegetation index within the buffer (although the variance is not ultimately included in the model construction and is only used for preliminary quality assessment and noise detection); finally, the buffer-level point features, including the Sentinel-2 full-band mean (B1–B12) and the average values of several commonly used vegetation indices (such as NDVI, EVI, MSAVI2, etc.; seen Table 1) are exported as structured CSV The file is used for subsequent modelling and analysis. This process combines cloud detection, time validity control and neighbourhood statistics, which not only strengthens the temporal stability of spectral features, but also considers spatial representativeness, providing the model with a set of spectral prediction variables with high generalization ability.

Table 1 Vegetation indices used as input variables for the developed machine learning based maize nutrient content estimation

Abbreviation	Index name	Expression	Application / characteristics	Citation
ARI	Anthocyanin Reflectance Index	(1 / B3) - (1 / B5)	Anthocyanin-physiological status indicator for different plant stress types	Gitelson et al., 2003
ARVI	Atmospherically Resistant Vegetation Index	(B8 - B6 - (B2 - B6)) / (B8 + B6 - (B2 - B6))	Corrects atmospheric scattering using blue-light reflectance	Kaufman & Tanré, 1992
CI_RE	Chlorophyll Index - red edge	(B8 / B5) - 1	Chlorophyll content	Gitelson et al., 2003
DSWI	Disease Water Stress Index	(B8 - B3) / (B11 + B4)	Sensitive to water shortage and plant damage	Bochenek et al., 2018
EVI	Enhanced Vegetation Index	2.5 * ((B8 - B4) / (B8 + 6 * B4 - 7.5 * B2 + 1))	Improved NDVI that reduces atmospheric influences	Matsushita et al., 2007
EVIredEdge	Red-Edge Enhanced Vegetation Index	2.5 * ((B8 - B6) / (B8 + 6 * B6 - 7.5 * B2 + 1))	Estimates LAI, chlorophyll and canopy water content	
GCI	Green Chlorophyll Index	(B5 / B3) - 1	Estimates chlorophyll content	Gitelson et al., 2003
GNDVI	Green Normalized Difference Vegetation Index	(B8 - B3) / (B8 + B3)	More sensitive than NDVI to chlorophyll variations (linked to N)	Gitelson & Merzlyak, 1998
HMSSI	Heavy-Metal Stress Sensitive Index	(B8 - B5 - 1) / ((B5 - B2) / B3)	Heavy-metal stress detection	Z. Zhang et al., 2018
IRECI	Inverted Red-Edge Chlorophyll Index	(B7 - B4) / (B5 / B6)	Canopy chlorophyll content	Jiang et al., 2023
MCARI	Modified Chlorophyll Absorption Ratio Index	((B5 - B4) - 0.2 * (B5 - B3)) * (B5 / B4)	Responsive to leaf chlorophyll & ground reflectance	Wu et al., 2008
MSR_RE	Modified Simple Ratio - red edge	((B8 / B4) - 1) / ((B8 / B4) + 1) ** 0.5	Chlorosis; high sensitivity to vegetation biophysical parameters	Wu et al., 2008
MTCI	MERIS Terrestrial Chlorophyll Index	(B6 - B5) / (B5 - B4)	Chlorophyll content of canopies	Dash & Curran, 2004
NDTI	Normalized Difference Turbidity Index	(B11 - B3) / (B11 + B3)	Water turbidity / suspended particles	Bid & Siddique, 2019
NDVI	Normalized Difference Vegetation Index	(B8 - B4) / (B8 + B4)	Green biomass, LAI	Sims & Gamon, 2002
NDVI_RE	Normalized Difference Vegetation Index - red edge	(B8 - B5)/(B8 + B5); (B8 - B6)/(B8 + B6); (B8 - B7)/(B8 + B7)	Chlorophyll content	Gitelson & Merzlyak, 1994
NDWI	Normalized Difference Water Index	(B8 - B12) / (B8 + B12)	Presence & abundance of water	Gao, 1996
NPCI	Normalized Pigment Chlorophyll Index	(B4 - B2) / (B4 + B2)	Chlorophyll content	Huang et al., 2014
NRI	Nitrogen Reflectance Index	(B3 - B4) / (B3 + B4)	Nitrogen concentration	Huang et al., 2014
PhRI	Physiological Reflectance Index	(B3 - B2) / (B3 + B2)	Solar utilisation efficiency; disease & abiotic stress	Huang et al., 2014
PSRI	Plant Senescence / Reflectance Index	(B5 - B2) / B3	Plant senescence	Yu et al., 2018; Z. Zhang et al., 2018
PSSRa	Pigment-Specific Simple Ratio (Chl-a)	B7 / B4	Chlorophyll-a index	Psomiadis et al., 2017
RERVI	Red-Edge Ratio Vegetation Index	B8 / B6	Biomass & chlorophyll estimation	
RVI	Ratio Vegetation Index	B8 / B4	Mitigates irradiance & transmittance effects	Y. Tan et al., 2019

RVSI	Red-Edge Vegetation Stress Index	((B5 + B6) / 2) - B6	Early stress detection & vegetative health assessment	
S2REP	Sentinel-2 Red-Edge Position Index	705 + 35 * (((B4 + B7) / 2 - B5) / (B6 - B5))	Chlorophyll, N status & growth monitoring	Eleveld et al., 2018
SAVI	Soil-Adjusted Vegetation Index	((B8 - B4) / (B8 + B4 + 0.5)) * (1 + 0.5)	Reduces soil-brightness effects	Huete, 1988
SIPI	Structure-Insensitive Pigment Index	(B8 - B2) / (B8 + B2)	Carotenoid / chlorophyll-a ratio; canopy stress; LAI	Yu et al., 2018
TCARI	Transformed Chlorophyll Absorption & Reflectance Index	3 * ((B5 - B4) - 0.2 * (B5 - B3)) * (B5 / B4)	Chlorophyll content, LAI	Wu et al., 2008
TVI	Triangular Vegetation Index	0.5 * (120 * (B6 - B3) - 200 * (B4 - B3))	Green LAI; sensitive to chlorophyll rise with canopy density	Qian et al., 2022
WDRVI	Wide Dynamic Range Vegetation Index	(0.2 * B8 - B4) / (0.2 * B8 + B4)	Vegetation fraction; LAI sensitivity	Gitelson, 2004

3.2.3. Sentinel-1 and polarimetric indices extraction

The Sentinel-1 GRD archive was filtered to the same temporal window and spatial buffers as Sentinel-2 data (November 1, 2017–April 30, 2018), selecting only scenes containing both VV and VH polarizations. A custom function computed four polarimetric indices on each image:

$DPSVI\_S1 = VV^2 / (VV \times VH)$	(dos Santos et al., 2021)
$RVI\_S1 = 4 \times VH / (VH + VV)$	(Nasirzadehdizaji et al., 2019)
$Pol\_S1 = (VV - VH) / (VV + VH)$	(Hird et al., 2017)
$CR\_S1 = VV / VH$	(Frison et al., 2018)

These indices, together with the original VV and VH bands, formed an indexed image collection. The collection was then averaged to produce mean composites for each band and index. Finally, mean values within each 60 m buffer were extracted and exported as a per-buffer feature table for downstream modelling.

3.2.4. Temperature data processing

The MODIS Terra Land Surface Temperature (LST) Day product (MOD11A1) was used to extract average daytime temperatures for the November 1, 2017–April 30, 2018 period, the growing season of maize (Mloza Banda et al., 2024). The image collection was first filtered spatially to the Malawi boundary and temporally to the growing-season window. Each scene's LST_Day_1km band was converted from scaled Kelvin to degrees Celsius ($\times$ 0.02 – 273.15) to produce physically meaningful temperature values. A per-pixel mean was then calculated across all valid observations to generate a single seasonal composite, which was clipped to the national border. To preserve the native sensor detail, the nominal 1 km resolution was used when extracting the mean temperature at each sampling point. The resulting point-level temperature metrics were exported as a CSV table for integration with other predictors. A TIFF file of mean temperature in Malawi was also exported and was clipped to Malawi boundary for temperature distribution mapping.

3.2.5. Precipitation data processing

Daily precipitation estimates were obtained from the CHIRPS dataset for the same November 2017–April 2018 interval. The collection was filtered by study-area boundary and date range, then averaged on a per-pixel basis to create a seasonal rainfall composite. After clipping to the Malawi outline, the dataset's native spatial resolution (250m) was employed to extract mean precipitation at each sample location, ensuring consistency with the inherent grid cell size. These point-level precipitation values were likewise exported as a CSV file for use alongside the temperature, and other features in subsequent modelling. A TIFF file of mean precipitation in Malawi was also exported and was clipped to Malawi boundary for precipitation distribution mapping.

3.2.6. Topology data processing

The MERIT DEM v1.0.3 was used to capture elevation across Malawi. The DEM was first clipped to the country boundary to limit the dataset to the study area. From this clipped surface, slope was calculated in degree using a standard terrain-analysis algorithm, producing a second layer that describes the steepness of the landscape. Elevation and slope together are key control factors on soil moisture, erosion potential, and nutrient redistribution factors known to affect crop nutrient status.

Both layers were merged into a single multi-band image, maintaining the native resolution of roughly 90 m (resampled to a 100 m scale during extraction). Using the same sampling points as for the spectral variables, the elevation and slope values were extracted per location. These point-level topographic

metrics were exported as a CSV table, completing the suite of environmental predictors for subsequent machine-learning analysis.

3.2.7. Soil data processing

Soil property layers (soil organic carbon and pH in water) were obtained from the SoilGrids250m dataset (Hengl et al., 2017). Because SoilGrids imposes a maximum download window of 2° × 2°, the Malawi extent was partitioned into adjacent tiles of this size. Each tile, covering the target depth interval was downloaded separately and then mosaicked into a one mosaic raster at the national level. The mosaic was clipped to the official Malawi boundary to remove data outside the study area and for soil properties distribution mapping.

All soil layers share a 250 m native resolution. Using the same georeferenced sampling points employed for climate and topographic variables, the soil-property value was extracted at each point using 'rasterio' package in python. These point-level soil metrics were exported as CSV tables. This workflow ensures consistent spatial alignment and uses the globally standardized SoilGrids predictions for integration with other predictors in the subsequent machine-learning models.

3.2.8. GAEZ extraction

GAEZ version 4 dataset for Malawi was used in this study, retrieved from the FAO–IIASA portal (Food and Agriculture Organization & International Institute for Applied Systems Analysis, 2021). GAEZ maps were obtained as a categorical raster covering world croplands. The GAEZ layer was first clipped to the Malawi boundary to limit it to the study area. Using the same georeferenced sampling points, the GAEZ class value at each location was extracted directly from the clipped raster, generating a categorical feature that describes the local Agro-Ecological Zone (AEZ; e.g. rainfall regime, temperature class, soil–water balance). This categorical variable captures broad, integrated controls on crop growth, combining climate, soil, and terrain into a single descriptor and was added to the point-level dataset alongside the continuous spectral, polarimetric, climatic, terrain, and soil predictors.

3.2.9. Data integration and preparation for modelling

All point-level feature tables (Sentinel-2 reflectance and indices, Sentinel-1 backscatter features and polarimetric indices, temperature, precipitation, elevation, slope, soil properties, and GAEZ) were joined on the unique sample-point identifier to form a single comprehensive dataset. Irrelevant columns, such as ID, raw geometry fields, variance, and export metadata, were removed to streamline the table.

3.3. XGBoost modelling and optimisation

3.3.1. Spatially informed data partitioning

To address the influence of spatial autocorrelation, a geographically structured partitioning strategy was employed. Sample locations, defined by geographic coordinates (longitude and latitude), were grouped into spatial clusters using K-Means clustering (k = 10). Data were then partitioned into training and testing sets using GroupShuffleSplit, treating each spatial cluster as an indivisible unit. This ensures that no samples from the same geographic region are present in both subsets, thus reducing the risk of spatial information leakage.

3.3.2. Hyperparameter tuning and XGBoost model

To model the relationship between environment features and multiple nutrient targets, the XGBoost algorithm (Chen & Guestrin, 2016), was employed, chosen for its scalability, ability to handle mixed data types (including categorical variables), and robustness to multicollinearity and missing data. XGBoost builds an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the

residuals of previous trees using gradient descent optimization. Its regularization features also help prevent overfitting, making it a suitable choice for our high-dimensional feature set.

Given the complexity of our dataset and the potential for overfitting, particularly due to spatial autocorrelation and collinearity among predictors, an extensive hyperparameter tuning process was performed. We used a randomized search strategy across a defined hyperparameter space to efficiently identify optimal model configurations. The hyperparameters included the number of boosting rounds (n_estimators), learning rate, maximum tree depth, minimum child weight, subsample ratio, column sample ratio, and regularization terms (gamma, reg_alpha, reg_lambda). These were sampled from a mix of uniform, integer, and log-uniform distributions to ensure adequate coverage of the search space while maintaining computational feasibility.

The search procedure used 5-fold cross-validation with shuffling to robustly estimate model generalization. A negative root mean squared error (RMSE) metric was used as the objective for tuning, as it directly reflects prediction error in the original units of the log-transformed target variable. Each nutrient target was modelled independently, enabling separate optimization and evaluation while allowing comparisons across nutrients.

Once the best hyperparameter set was identified, a final model was retrained on the entire training subset (still using log-transformed targets). This model represents the best compromise between fitting complexity and generalization ability, given the available data and the specified search ranges. To evaluate how well the model predicted actual concentration values, predictions on both the training and test subsets were first produced in log space, as the target distribution is skewed, and then transformed back to the original concentration scale by applying the inverse log operation.

3.3.3. Model performance assessment

After hyperparameter tuning, the best-performing models for each nutrient target were selected and further evaluated on both training and test subsets. To handle non-normality and skew in the target variables, log-transformation was applied during training, and predictions were exponentiated back to the original scale for interpretation and error calculation.

Model performance was assessed using several standard regression metrics, including Root Mean Squared Error (RMSE), normalized RMSE (nRMSE), and the coefficient of determination (R^2). These were calculated as follows:
RMSE quantifies the standard deviation of prediction errors in the same unit as the target variable.

$$RMSE = \sqrt{[\Sigma\ (y_i - \hat{y}_i)^{\,2} / n]}$$

Where:

- y_i is the observed values,
- $\hat{y}_i$ is the predicted values,
- n is the number of observations.

nRMSE scales RMSE relative to the mean of the actual values, allowing comparison across nutrients.

$$nRMSE = RMSE / \bar{y}$$

Where:

- $\bar{y}$ is the mean of the observed values.

R^2 indicates the proportion of variance explained by the model, with values closer to 1 suggesting better predictive power.

$$R^2 = 1 - [\Sigma\ (y_i - \hat{y}_i)^{\ 2} / \Sigma\ (y_i - \bar{y})^{\ 2}]$$

Where:

- y_i is the observed values,
- $\hat{y}_i$ is the predicted values,
- $\bar{y}$ is the mean of the observed values.

MAE is less sensitive to outliers compared to RMSE and provides an additional perspective on prediction accuracy.

$$\text{MAE} = \Sigma\ |y_i - \hat{y}_i|$$

Where:

- y_i is the observed values,
- $\hat{y}_i$ is the predicted values.

To assess the accuracy of the models, actual-versus-predicted value plots were generated, which help visualize systematic biases or heteroscedasticity in the predictions. Moreover, feature importance was quantified using the gain metric, which reflects the relative contribution of each feature to improving the model's predictive performance. In the context of regression, gain measures the average reduction in the loss function (typically mean squared error) achieved by splits on a given feature across all trees in the ensemble. Sorting these scores revealed which environmental variables (for example, soil pH, elevation, precipitation, or spectral indices) contributed most to reducing prediction error. Horizontal bar charts were produced to visualize the top features for each nutrient, offering insight into the primary drivers of nutrient variability. These insights support model interpretability and help identify the most influential predictors.

Finally, Partial Dependence Plots (PDPs) were created for selected features to illustrate their marginal effect on the predicted outcomes. PDPs offer an intuitive view of how the model's predictions change with a given feature while averaging out the effects of others, thus aiding in the interpretation of potentially non-linear and interactive relationships (Friedman, 2001).

4. RESULTS

4.1. Nutrient spatial distribution maps

The spatial distribution maps of different nutrients were generated using all available sampling points, prior to data cleaning or model-specific filtering. These maps illustrate the general spatial patterns of each micronutrient across Malawi (Figure 4). All nutrient concentrations were log-transformed (natural log) prior to visualization and modelling. This transformation was applied to reduce right-skewness, stabilize variance, and improve the statistical properties of the data for predictive modelling (Osborne, 2010). Moreover, since predictive models were trained using log-transformed nutrient concentrations, visualizing the spatial distribution in the same scale allows for consistent interpretation.

The maps show that maize grain nutrient levels are generally highest in southern Malawi and lowest in the north (Figure 4). On the map showing the Ca values, the darkest red areas appear around the Shire Valley and near Blantyre in the south, while central Malawi is mostly orange and the far north is pale yellow. The Zn map follows a similar pattern, with the deepest greens clustered in the south, medium greens in the central region, and the lightest greens up north. Fe concentrations also peak in the south, where the darkest blue dots are most frequent, tapering to medium blues in central areas and light blues in the north. Se is especially low throughout the north and centre (pale purple) but shows higher values (darker purple) in the southern Shire Valley. In each case, nutrient-rich maize grain is found primarily in the south, with steadily lower levels as one moves northward.

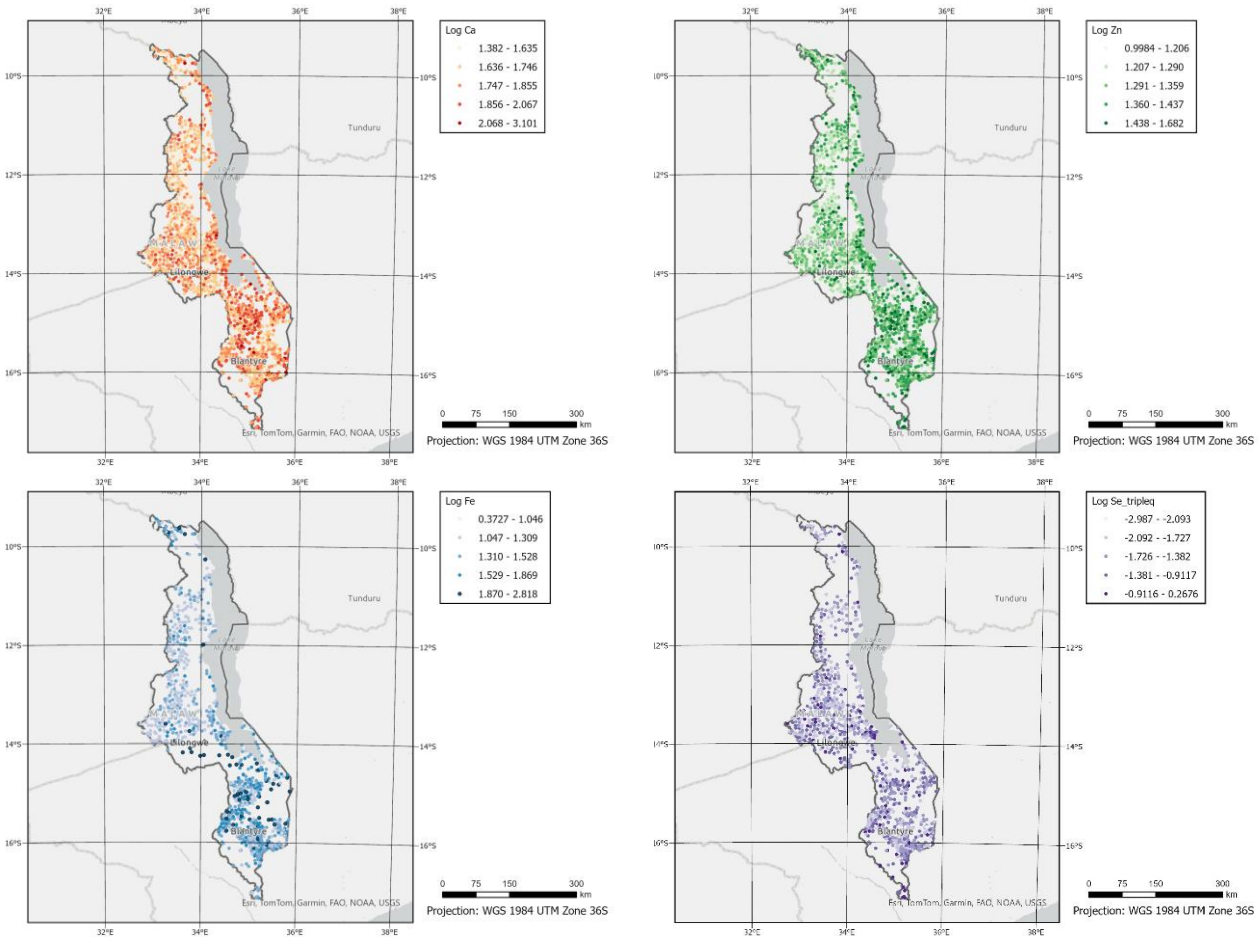


Figure 4 Spatial distribution of log-transformed maize grain micronutrient concentrations in Malawi (Ca, Zn, Fe, Se)

4.2. Environmental variable spatial distribution

4.2.1. Temperature spatial distribution

Malawi's average temperatures during the growing season from November 2017 to April 2018 showed a clear north-south difference (Figure 5). Temperatures ranged from about 19.4°C to 43.4°C, with most areas between 22–28°C. Cooler temperatures were found in the northern highlands and along the shores of Lake Malawi, especially near Mzuzu, shown as blue areas with temperatures below 22°C. In the south, especially near Blantyre, Nsanje, and Mangochi, temperatures were significantly higher, reaching over 30°C, shown in orange-red. This spatial variation is closely related to changes in topography height and latitude, with low altitudes and high temperatures in the south and cooler highlands in the north. Temperature distribution is particularly important for agriculture, as it can affect crop growth cycles and yields.

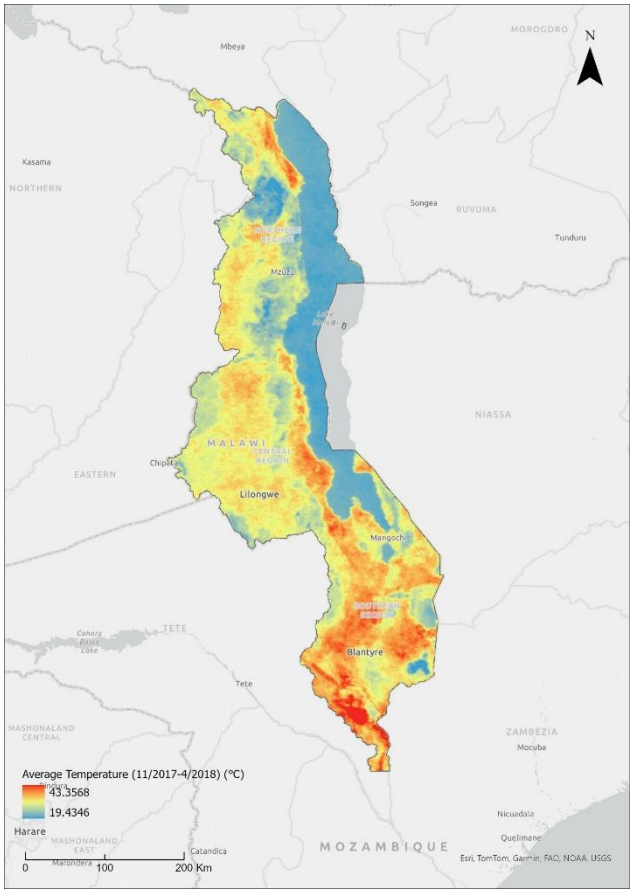


Figure 5 Average temperature of maize during growing season in Malawi

4.2.2. Precipitation spatial distribution

Figure 6 shows the average precipitation distribution in Malawi during the crop growing season from November 2017 to April 2018 (unit: mm/day). Overall, the spatial distribution of precipitation in Malawi shows more precipitation in the north and less in the south, more in the east and less in the west.

The northern region, especially Mzuzu and its surrounding areas, shows a higher average precipitation, with an average daily precipitation of more than 8 mm/day, and local areas reaching more than 10 mm/day, which is the darker area in the figure. This may be related to the large terrain variations and high altitude in the region, which promotes more precipitation to gather. The central region (such as Lilongwe) has relatively moderate precipitation, with average values mostly in the range of 5–7 mm/day, which is the

light blue area. The precipitation distribution is relatively uniform, which is suitable for the development of general dryland agriculture. In contrast, the average precipitation in the southern region (such as Blantyre, Nsanje, and Mangochi) is relatively low, with most areas below 4 mm/day, and some areas even below 3 mm/day. These areas are shown in light blue on the map, indicating that they have relatively scarce precipitation resources and may be more dependent on irrigation or face drought stress.

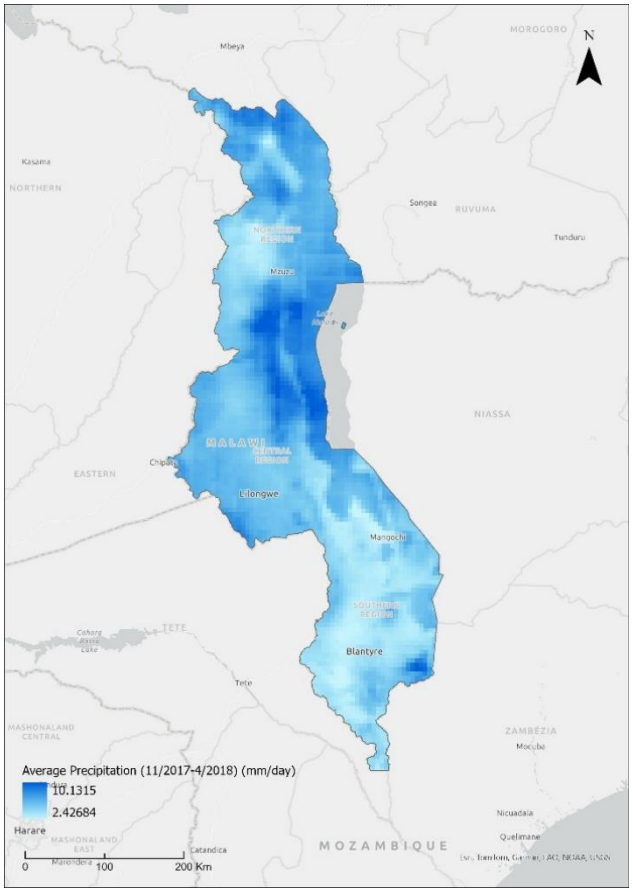


Figure 6 Average precipitation of maize spatial distribution during growing season in Malawi

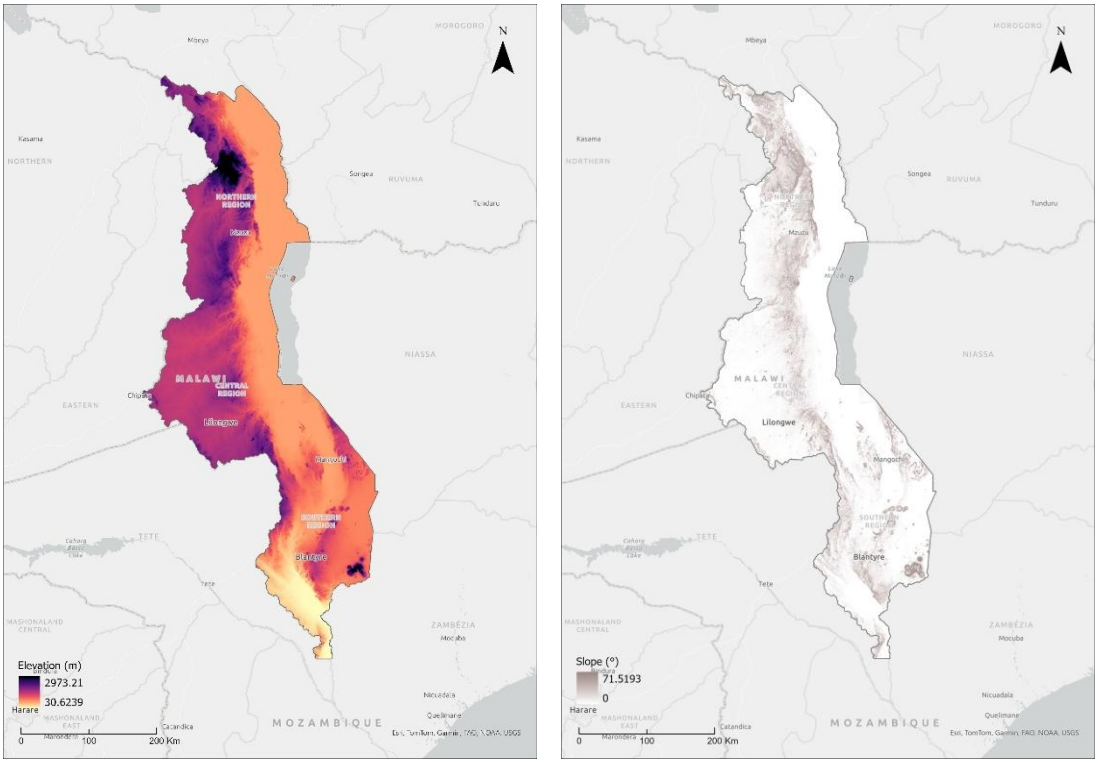
4.2.3. Elevation and slope spatial distribution

Malawi has a very diverse terrain, ranging from a minimum of about 30 meters to a maximum of more than 2,900 meters. As can be seen from the map (Figure 7a), the overall elevation shows a trend of gradually decreasing from the western and southern edges to the central and eastern parts.

The northern region (especially north of Mzuzu) has many high mountain areas and is the highest area in Malawi, highlighted in dark purple. The central region (Lilongwe and its surroundings) is relatively flat, mostly between 800–1,200 meters above sea level, and is a typical plateau. The eastern area along Lake Malawi, especially near the lakeshore, has the lowest elevation, shown in light yellow. These low-altitude areas are usually fertile and humid and are one of the important agricultural belts in Malawi. The spatial distribution of elevation affects temperature, precipitation and land suitability, and is also the basic condition for slope formation.

The slope distribution (Figure 7b) in Malawi is closely related to its topography. Darker colours represent steeper slopes. As shown in the figure, the areas with large slopes are mainly concentrated around high-altitude areas, especially in the northern and southern marginal areas, where the slope can exceed 30° and even reach 70° in some areas. The central plateau area (near Lilongwe) has a smaller slope, and a large area has a slope between 0–5°, which is shown in light grey in the figure. These areas are flat and suitable for

mechanized agricultural development. In contrast, the slopes in the south (such as Blantyre) and around the northern mountains are steep, which are not suitable for construction and large-scale farming, and there is a potential risk of soil erosion and landslides.



a b

Figure 7 Spatial distribution maps of elevation (a) and slope (b) in Malawi.

4.2.4. Soil pH spatial distribution

The soil pH distribution in Figure 8 is expressed in pH × 10. The colour classification reflects the range of acidity and alkalinity. The soil in Malawi is generally acidic to slightly acidic, mainly distributed between pH 5.0–6.5. Among them, the soil pH in the central region (especially Lilongwe and its surrounding areas) is relatively moderate, mostly between 5.5–6.5, which is suitable for the growth of common dryland crops. In the southern and northern highlands, the shallow soil still shows high acidity (pH < 5.5), but this is alleviated in the deep layer, showing a trend of slightly increasing pH with depth, which may be related to leaching and accumulation of surface organic matter.

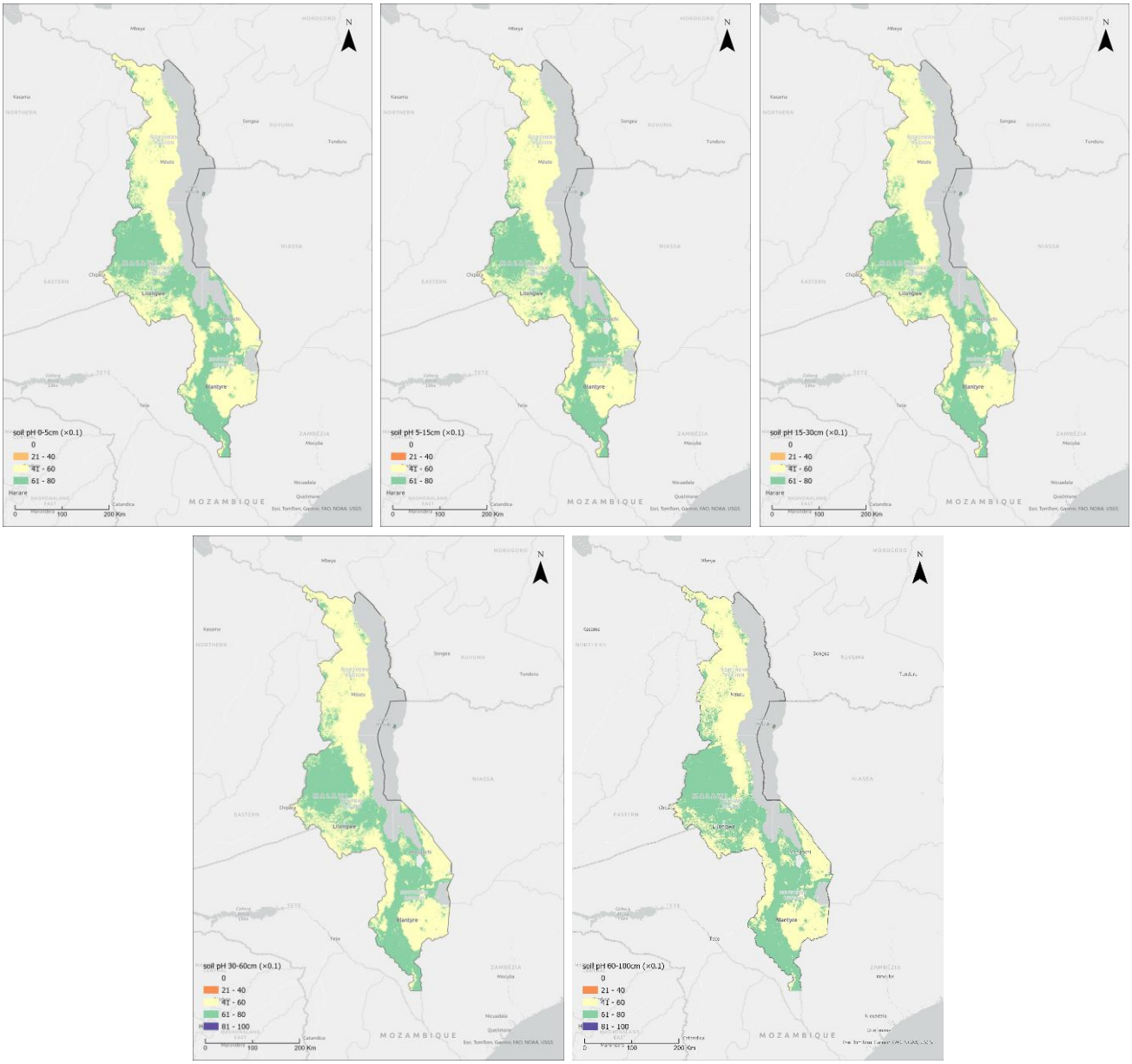


Figure 8 Soil pH distribution across soil depths in Malawi (0–100 cm, from left to right, up to down: 0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm)

Overall, the soil pH in most areas of Malawi is acidic, which may limit the absorption of nutrients by maize.

4.2.5. Soil of Carbon (SOC) spatial distribution

The set of figures (Figure 9) shows the spatial distribution of SOC content in different soil layers (from 0–100 cm) in Malawi, in kg/ha, and the data comes from SoilGrids. In general, the SOC content is unevenly distributed in the study area and decreases significantly with the depth of the soil layer. The surface layer (0–5 cm) has the highest organic carbon content, with some areas exceeding 3000 kg/ha, mainly concentrated in the mountainous areas and lakeshore areas in northern Malawi. These areas have high terrain and good vegetation coverage, which promotes the accumulation of surface organic matter. As the soil layer deepens, the SOC content gradually decreases. The 5–15 cm and 15–30 cm layers still retain certain carbon-rich patches, but the distribution area and concentration have been significantly reduced.

The organic carbon content of the 30–60 cm and 60–100 cm layers is generally low, with most areas less than 1000 kg/ha, and only a few high-carbon areas have relatively high carbon reserves in the deep layer. Vertically, SOC shows an obvious "rich on the surface and poor in the deep" distribution pattern, indicating that the surface soil is key in the carbon cycle. It also suggests that the deep carbon pool is relatively weak in response to natural and human disturbances.

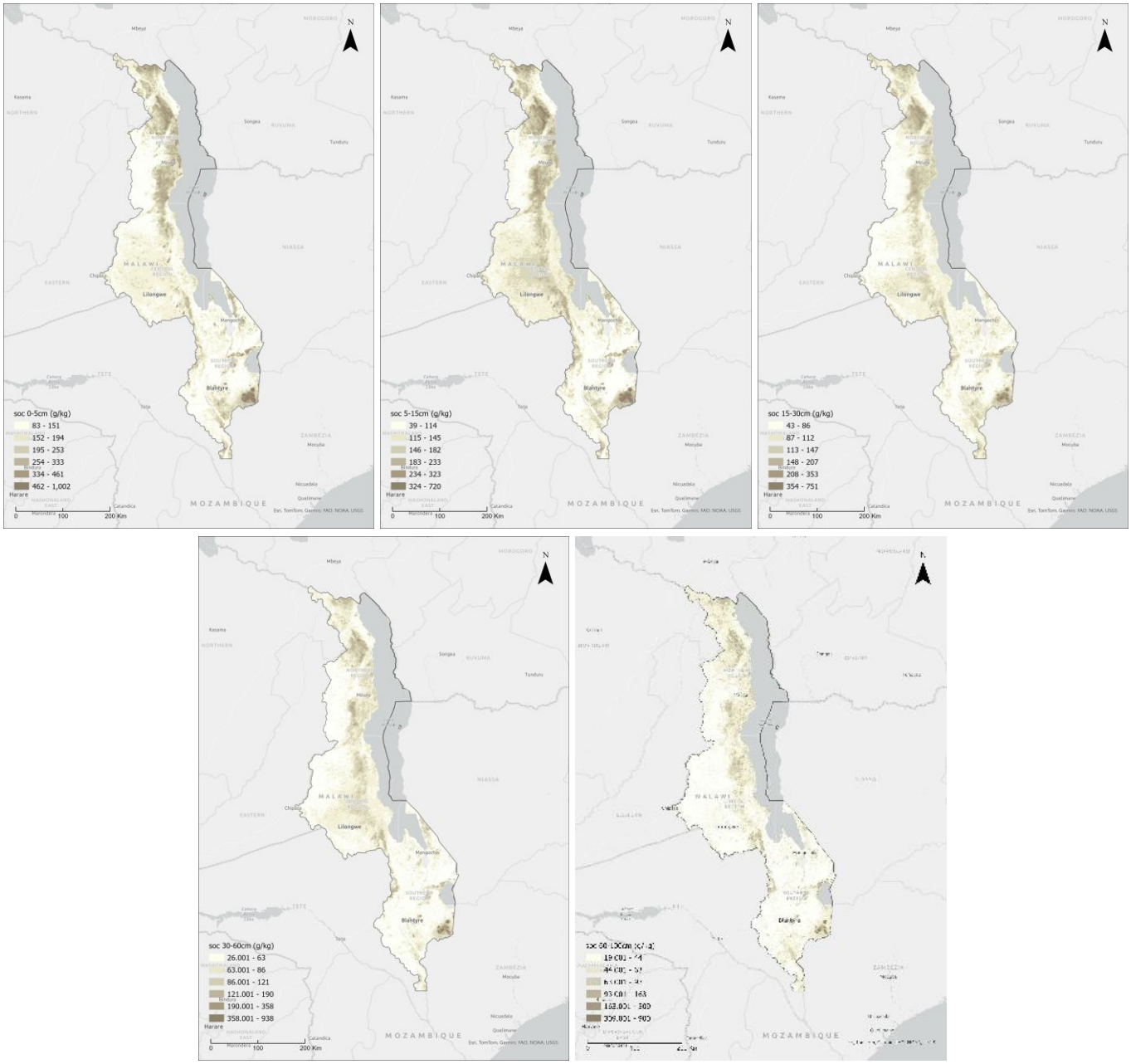


Figure 9 SOC distribution (g/kg) across soil depths in Malawi (0–100 cm, from left to right, up to down: 0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm)

4.2.6. Agroecological zones in Malawi

Malawi's agricultural ecological regionalization shows significant latitudinal gradients and topographic control effects. As can be seen in Figure 10, most areas are classified as tropical lowland semi-arid (Tropics, lowland; semi-arid) and tropical lowland sub-humid (Tropics, lowland; sub-humid), represented by light pink and pink respectively, which are widely distributed in the central and southern plains and are the country's main food production areas. Some areas, especially the central and northern highlands, are classified as tropical highland semi-arid or sub-humid areas (such as magenta and purple areas). These areas have complex terrain, high precipitation but limited land resources.

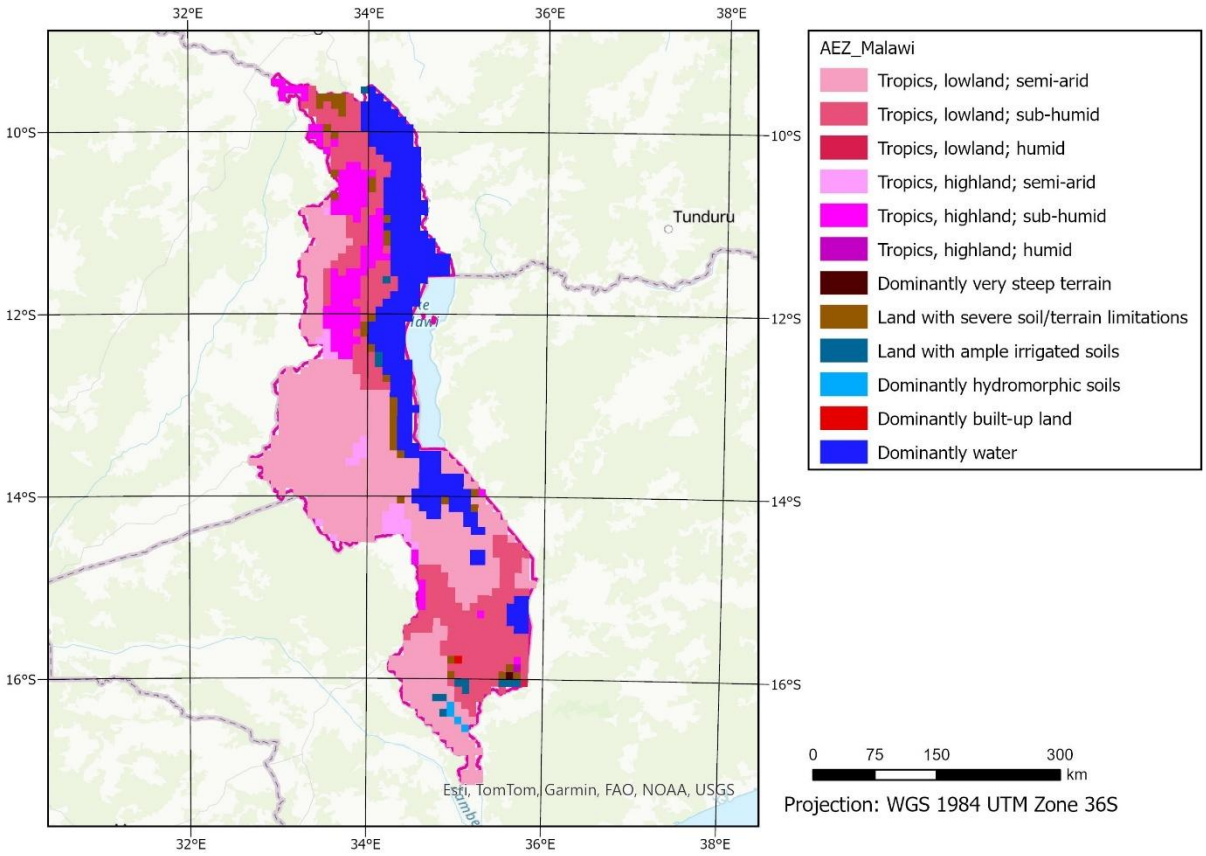


Figure 10 AEZ of Malawi based on climate, elevation, and land limitations

There are significant areas of terrain or soil restrictions (dark brown) in central and northern Malawi and along the eastern shore of the lake, which may include steep slopes, shallow soil, rocky surfaces, etc., which restrict agricultural production. At the same time, a small amount of well-irrigated land (light blue) and wetland areas (blue green) are also marked in the figure, indicating that the local water resource distribution is significantly uneven.

In addition, the blue area along Lake Malawi represents a large area of water, showing the important position of the lake in the geographical pattern. The red block area represents the main towns or built-up areas, which are small but concentrated in the south and along the lake coast, reflecting the concentrating of population and economic activities.

Overall, Malawi's agricultural ecological zones have obvious north-south and high-low zone differences, which have important guiding value for crop and land management. The tropical lowland semi-arid and sub-humid areas are the current key agricultural development areas, whereas the highlands and terrain-restricted areas need to adopt corresponding sustainable management strategies.

4.3. Descriptive statistics and spatial distribution of maize grain nutrients

All subsequent analyses, including descriptive statistics and modelling, were based on the pre-processed and filtered sample set.

4.3.1. Descriptive statistics

The descriptive statistics of nutrient composition in various crops, measured in mg·kg−1, reveal significant differences in nutrient densities based on median values. For Ca (n = 319), the mean concentration is 61.00 mg·kg^{-1} and the median is 57.05 mg·kg^{-1}, with a standard deviation of 18.31 and an interquartile range of 50.38 – 68.04 mg·kg^{-1}. Fe (n = 270) has a mean of 26.94 mg·kg^{-1}, a median of 20.15 mg·kg^{-1}, a standard deviation of 28.18, and an interquartile range of 15.41 – 28.24 mg·kg^{-1}. Zn (n = 319) shows a mean of 22.03 mg·kg^{-1} and a median of 21.85 mg·kg^{-1}, with a standard deviation of 3.97 and a 25th–75th percentile range of 19.31 – 24.35 mg·kg^{-1}. Se_tripleq (n = 287) has much lower values: a mean of 0.0401 µg·kg^{-1}, a median of 0.0227 µg·kg^{-1}, a standard deviation of 0.0484, and an interquartile range of 0.0118 – 0.0475 µg·kg^{-1}.

Figure 11 presents violin plots with overlaid boxplots for these same targets. The left panel shows that Ca, Fe, and Zn distributions are right-skewed: Fe spans several orders of magnitude (approximately 5 – 100 mg·kg^{-1}), whereas Ca and Zn are more tightly clustered around their medians (≈ 57 mg·kg^{-1} for Ca; ≈ 21.85 mg·kg^{-1} for Zn). The right panel isolates Se_tripleq, which is highly skewed toward small values (0.001 – 0.10 µg·kg^{-1}); its mean lies well above its median. These results support the use of a log transformation in subsequent modelling to stabilize variance and reduce skewness.

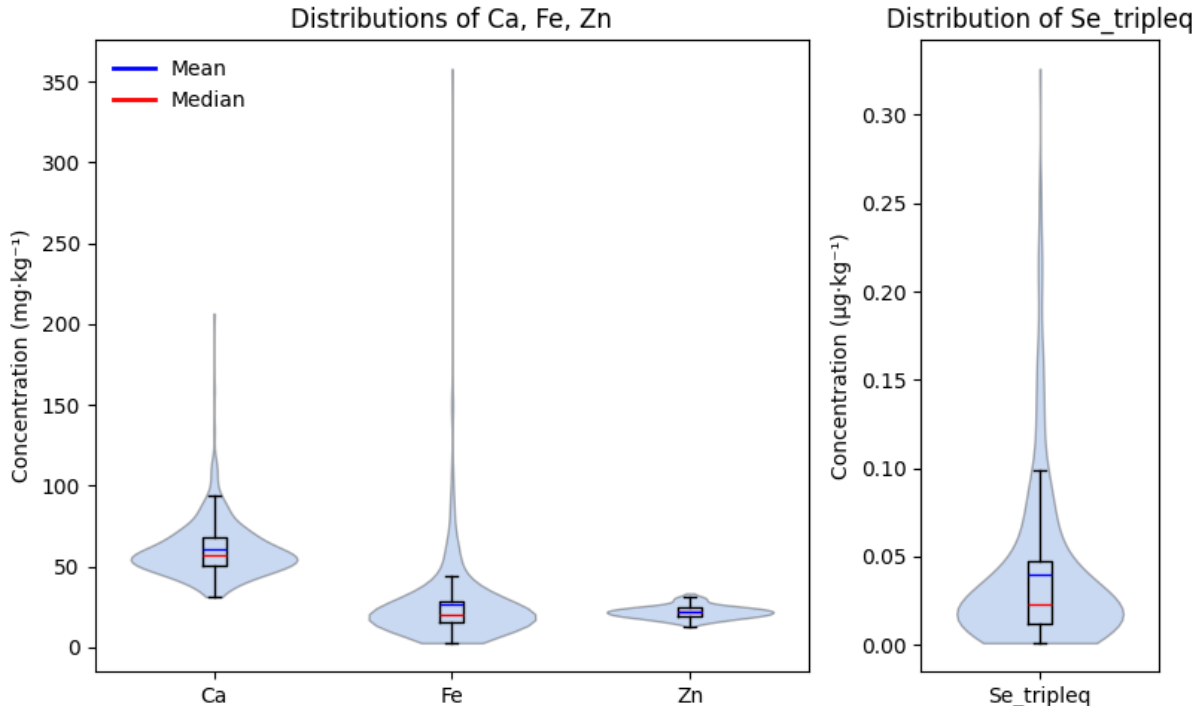


Figure 11 Violin and box plots showing the spatial distributions of soil micronutrient concentrations for Ca, Fe, Zn (left), and Se_tripleq (right) across all sampling points

Figure 12 presents the distributions of Ca, Fe, Zn, and Se concentrations across three AEZs in Malawi, using combined violin and boxplots. Overall, notable differences in micronutrient concentrations are observed between AEZs, particularly between lowland and highland regions.

Ca (Figure 12a) exhibits substantial variation across AEZs, with the Lowland Semi-arid and Lowland Sub-humid zones displaying higher median values and broader distributions compared to the Highland Semi-

arid zone. The violin plots in lowland areas show positive skewness and long upper tails, suggesting the presence of high-concentration outliers and greater variability. In contrast, the highland zone presents a narrower distribution with lower overall values, indicating more uniform but limited Ca availability.

A similar spatial pattern is observed for Fe (Figure 12b), though with even more pronounced skewness. The Lowland Semi-arid zone shows a wide distribution and exceptionally high Fe concentrations in some samples, as indicated by the elongated upper tail. The Highland Semi-arid zone again exhibits the lowest and least variable Fe concentrations, reinforcing the trend of reduced micronutrient availability at higher elevations.

Zn (Figure 12c) presents a more balanced distribution across AEZs. While concentrations are still higher in the two lowland zones, particularly in the Lowland Sub-humid region, the differences among zones are less pronounced than those observed for Ca and Fe. The distributions are more symmetrical and moderately spread, indicating relatively consistent Zn levels across the zones.

Se (Figure 12d), measured in micrograms per kilogram, shows the lowest absolute concentrations and the strongest right skew among the four nutrients. The Lowland Semi-arid zone displays the highest Se values, although the overall range remains small. The Highland Semi-arid zone shows both the lowest median and the narrowest distribution, suggesting that Se deficiency may be most severe in upland areas. This spatial pattern likely reflects the influence of parent material, drainage, and topographic conditions on Se availability.

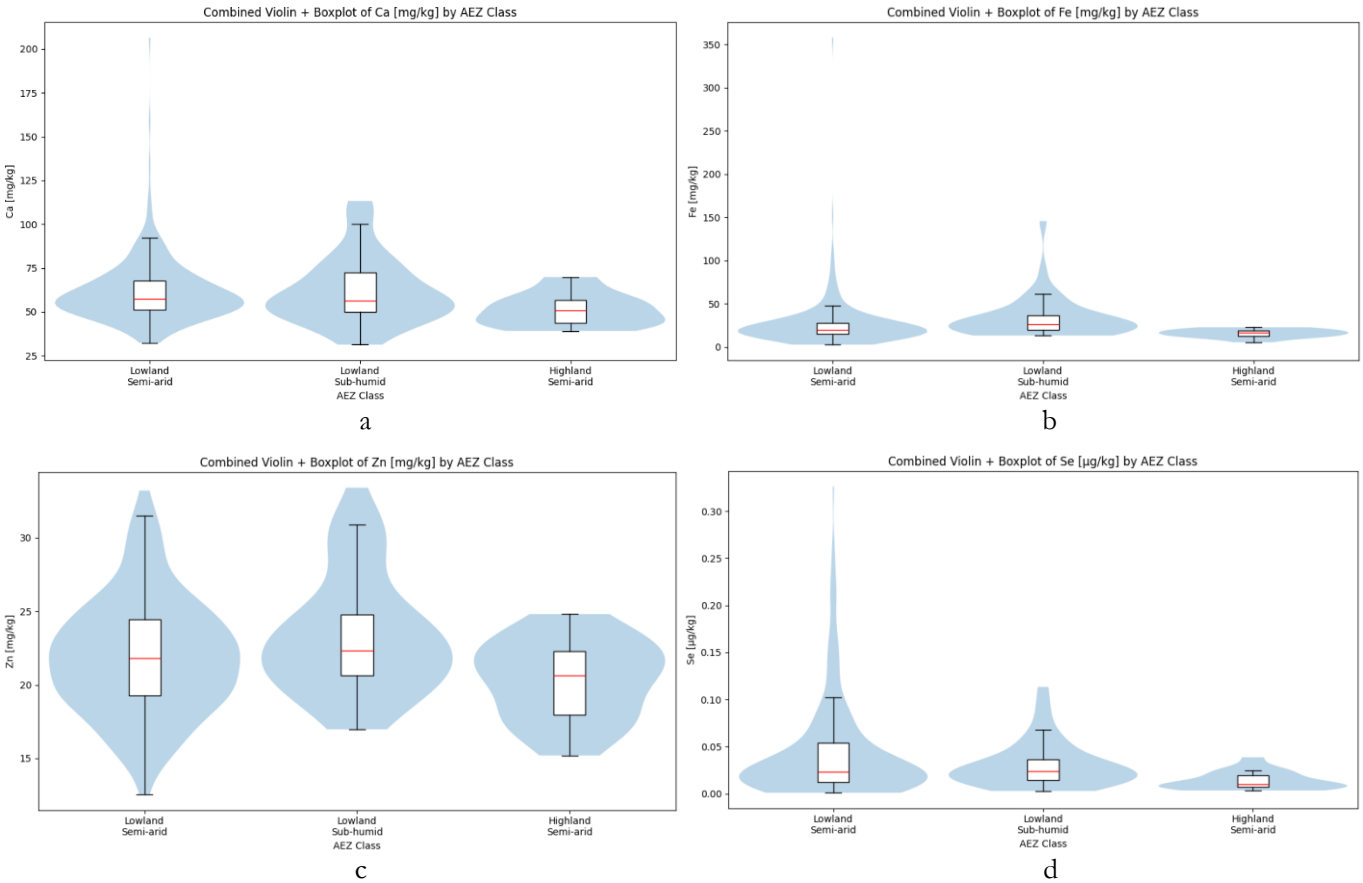


Figure 12 Combined violin and boxplots of nutrient concentrations in maize yield by AEZ class: (a) Ca, (b) Fe, (c) Zn, (d) Se.

In summary, these distributions reveal clear spatial heterogeneity in micronutrient concentrations across Malawi. The consistently lower concentrations in the Highland Semi-arid zone highlight the need for targeted nutrient management or biofortification efforts in highland regions.

4.3.2. Spatial distribution of filtered samples

The resulting spatial clusters and the train/test partition are illustrated in Figure 13. Each cluster is represented by a different colour. Training samples are marked with black circles, while testing samples are shown without border. The map is overlaid with the national boundary of Malawi for geographic context.

It can be observed that the spatial split preserves geographic separation between training and testing data, thereby enabling spatially informed partitioning strategy.

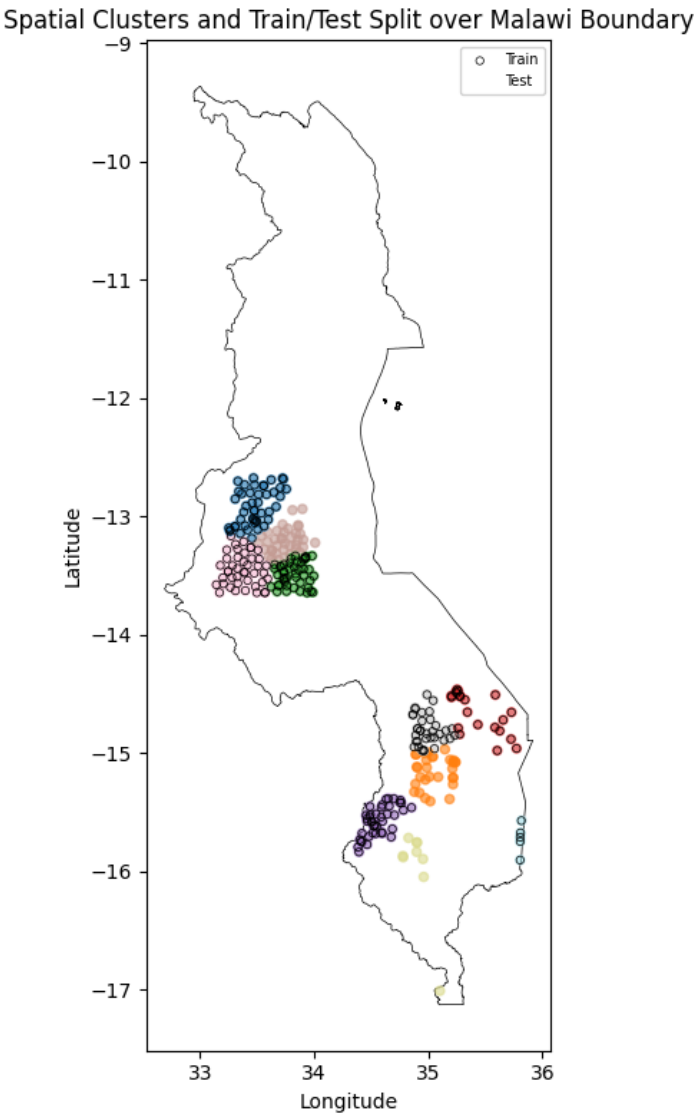


Figure 13 Spatial Clustering of Training Samples and Train/Test Split Distribution Across Malawi

4.4. XGBoost model performance

4.4.1. Hyperparameters

After adjusting parameters with different nutrients as targets, the hyperparameters of the four models were obtained, as shown in Table 2 below. The optimal hyperparameters varied considerably across the four nutrient targets, reflecting differences in data structure and model complexity requirements. For example, Ca and Se_tripleq models adopted deeper trees and lower learning rates. In contrast, Fe achieved optimal performance with a higher learning rate and more conservative tree depth. The Zn model showed a strong L1 regularization (reg_alpha) and nearly full sampling (subsample ≈ 0.99), implying sensitivity to input

features and possible overfitting risk. Meanwhile, Se_tripleq required stronger L2 regularization (reg_lambda), possibly due to weaker or noisier feature-target relationships.

Table 2 Model hyperparameters in models tuning with different nutrients as targets

Parameter	Ca	Fe	Zn	Se_tripleq
colsample_bytree	0.83446203	0.687270059	0.635674516	0.886122385
gamma	0.329130195	1.863096003	0.161868646	5.35E-07
learning_rate	0.003717036	0.06504857	0.007651054	0.001031998
max_depth	9	6	5	4
min_child_weight	3	5	9	3
n_estimators	565	664	206	642
reg_alpha	1.57E-08	3.69E-05	0.026156272	0.004906091
reg_lambda	0.00145807	6.31E-08	3.95E-08	0.02096063
subsample	0.58855534	0.729624446	0.993443468	0.802979987

4.4.2. Comparison between actual and predicted nutrient concentrations

Figure 14 shows the comparison between the actual concentrations of four micronutrient elements in corn grains (Ca, Fe, Zn and Se) and the model prediction values, presented in the form of a scatter plot. The horizontal axis is the actual observed value, the vertical axis is the model prediction value, and the red dotted line represents the ideal prediction line (i.e., predicted value = actual value). By analysing the distribution trend of the scatter points in the figure, the accuracy and bias of the model prediction for different elements can be intuitively evaluated.

In the prediction graph of Ca (Figure 14a), most sample points are concentrated in the low to medium concentration range of 45–80 mg/kg, and the scattered points are relatively closely distributed near the ideal line, showing a certain fitting trend, indicating that the overall prediction ability of the model for Ca is relatively stable. However, in some high-concentration samples (such as >100 mg/kg), the predicted values are significantly lower than the actual values, showing a certain underfitting or "regression to the mean" phenomenon, which means the model is difficult to accurately capture extremely high values, resulting in a low prediction of high-concentration Ca. Overall, the prediction R^2 of Ca is 0.15 and the RMSE is 18.92..

The prediction graph of Fe (Figure 14b) reflects large deviations and instability. Although some points in the low concentration area (<50 mg/kg) are close to the diagonal line, the overall predicted value is generally lower than the actual value, especially in the high concentration area (such as >100 mg/kg), where multiple points deviate significantly from the ideal line, forming a significant underestimation trend. This underestimation phenomenon indicates that the model has great difficulty in learning the distribution law of Fe, which may be related to the extremely skewed distribution of the original Fe data. In addition, the large span of Fe concentration may also make it difficult for the model to adapt to its distribution changes. The R^2 of this group of predictions is only 0.08, and the RMSE is 25.05, which shows a poor fitting effect.

The prediction results of Zn (Figure 14c) are relatively good, and most of the sample points are concentrated in the range of 15–25 mg/kg and are close to the ideal line. Although there are a few underestimations, the overall linear relationship and consistency are strong, showing good generalization ability in Zn concentration prediction, and can accurately identify the relative changes between different samples. The R^2 of the model is 0.14 and the RMSE is 3.67, indicating that its prediction stability and reliability are high.

The prediction of Se (Figure 14d) faces greater challenges. First, the concentration level of Se is extremely low (usually in the μg/kg range). As shown in the figure, most of the scattered points are concentrated in the area with low actual values (<0.05 μg/kg), and the model prediction values are correspondingly low. However, in the medium and high concentration samples (such as >0.10 μg/kg), the model prediction values are significantly lower and underestimated, indicating that the model has a weak ability to identify high Se values. In addition, the unit of Se is different from that of the other three elements (μg/kg instead of mg/kg), which may also increases the modeling difficulties caused by the difference in numerical scales. The R^2 of this group of predictions is -0.51 and the RMSE is 0.03, indicating that the model not only fails to effectively explain the changes in Se concentration, but also produces large errors.

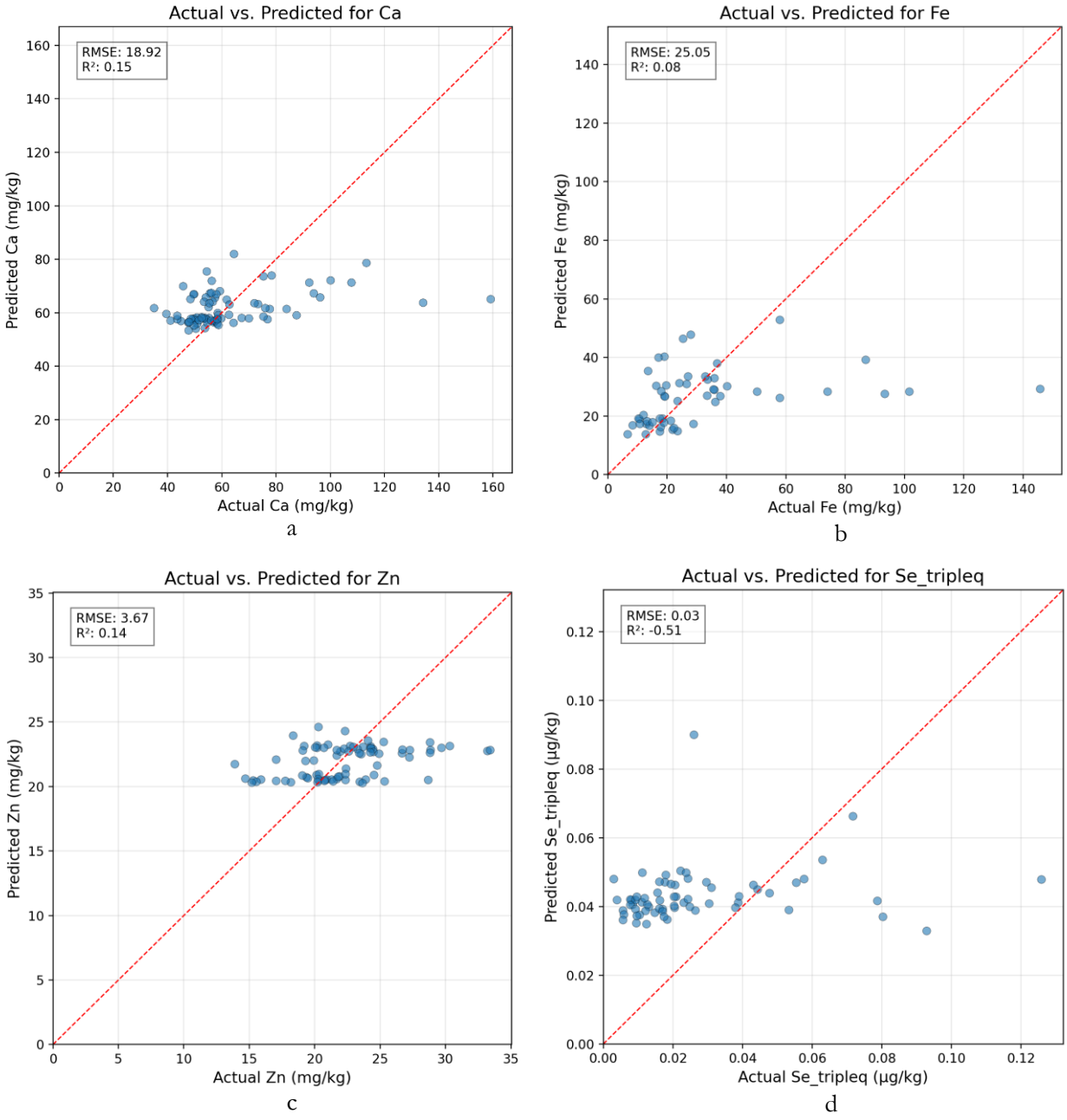


Figure 14 Comparison of actual and predicted values of each micronutrient: (a) Ca, (b) Fe, (c) Zn, (d) Se.

Overall, the model has a good fitting effect on Zn and Ca, and the predicted values are basically consistent with the actual values. However, Fe and Se have significant underestimation problems, especially in areas with high concentrations or extreme values.

4.4.3. Overall model performance

In the prediction of all elements, the model's fitting ability for extreme values is weak, and the prediction results tend to regress to the central value, reflecting its insufficient generalization ability when it exceeds the mean behaviour range, which is also consistent with the generally low R^2 phenomenon of the test set. For each nutrient element, Train_RMSE and Train_nRMSE reflect the model's fitting ability on the training data (expressed in original units and standardized units), while Train_R^2 indicates the proportion of variance that can be explained during the training process. Similarly, Test_RMSE, Test_nRMSE, Test_R^2, and Test_MAE measure the generalization error of the model on unseen data (Table 3).

Table 3 Overall model performance

Nutrient	Train_RMSE	Train_nRMSE	Train_R2	Test_RMSE	Test_nRMSE	Test_R2	Test_MAE
Ca	13.38438	0.22135	0.41148	18.92374	0.30217	0.15047	12.43531
Fe	24.16926	0.93355	0.28121	25.05130	0.80215	0.07867	14.16494
Zn	3.28879	0.14993	0.31313	3.66949	0.16447	0.13722	2.79445
Se_tripleq	0.04199	0.95158	0.36592	0.02879	1.08884	-0.51345	0.02515

For Ca, the model showed a certain fitting ability on the training set, with Train_R^2 of 0.41148, indicating that the model can explain about 41% of the variance in Ca concentration. However, on the test set, Test_R^2 dropped to 0.15, and the prediction performance decreased significantly, indicating that the model tends to overfit on new data. In addition, Test_RMSE is 18.92, which is significantly higher than 13.38 in the training set, which also confirms this hypothesis. Although the normalized error (Test_nRMSE = 0.30217) is not extreme, the overall predictive ability is weak. This result is consistent with the trend of significant underestimation of high concentrations of Ca in the figure, and the model's performance at the tail of the distribution needs to be strengthened.

The prediction performance of Fe is the weakest, with Train_R^2 of 0.28121 and Test_R^2 even lower, at only 0.07867, which has almost no explanatory power. Although the training error is slightly lower (Train_RMSE = 24.17), the test error has almost no improvement (Test_RMSE = 25.05), and the normalized error is still high (Test_nRMSE = 0.80215), indicating that the model has serious difficulties in fitting the Fe concentration and has very poor generalization ability. The image also shows that Fe is significantly underestimated in the high value area, which may be greatly affected by the skewed distribution of its data, making it difficult for the model to learn its rules.

For Zn, the model prediction performance is relatively good, with R^2 of 0.31 and 0.13 for training and testing respectively. Although the performance of the test set has declined, the Test_RMSE is 3.67, the normalized error is only 0.16447, and the Test_MAE is relatively low (2.79), indicating that the model has a certain generalization ability. Combined with the image, the overall distribution of the prediction results of Zn is reasonable, and a small amount of underestimation does not significantly affect the overall trend. It is the most balanced one among the four elements.

Se performed well in training (Train_R^2 = 0.36592), but performed extremely poorly on the test set, with a negative value of -0.51345, indicating that the model's prediction effect is even worse than simply replacing it with the mean. Although the test error is small in absolute value (Test_RMSE = 0.02879), due to the extremely low concentration of Se (μg/kg), its normalized error is as high as 1.08884, reflecting the serious underfitting problem of the model in Se prediction.

Overall, the model has good prediction stability for Zn and Ca, but poor prediction for Fe and Se, especially the weak generalization ability on the test set. All elements have a certain degree of performance degradation on the test set, indicating that the current model has not been able to cope with the complexity of the data well, especially the lack of modelling ability for outliers or the tail of the distribution.

4.5. Feature importance

Figure 15 shows the feature importance of each model in estimating different nutrients in maize. Elevation and precipitation are among the most relevant predictors for Ca, Fe and Zn, proving that topography-derived covariates and rainfall consistently explain most spatial variance in grain or soil micronutrients (Gohil, 2023). Their dominance in our model, therefore, aligns with the broader literature that identifies terrain-controlled hydrology and weathering as first-order nutrient drivers (Páez-Bimos, 2023).

The Ca model identifies the elevation as the most relevant variables, followed by seasonal rainfall and the acidity level of surface soil (pH 5–15 cm). Higher regions in Malawi are cooler and better drained, and, together with moderate rain and near-neutral pH, makes Ca easier for maize roots to absorb (Chilimba et al., 2011). Vegetation indices from Sentinel-2 are among the next most relevant input variables. Greener and thicker canopies is an indication that the crop has a high Ca concentration (Gatti et al., 2023).

In the case of Fe, the importance of elevation is even more pronounced. Alongside soil properties such as SOC_5_15 and pH, multiple spectral bands (e.g., B2_mean, B7_mean) play a strong role. This implies Fe availability is shaped by both soil characteristics and surface reflectance properties.

For Zn model, elevation, rainfall and soil carbon are very important for Ca estimation. In addition, polarimetric index proves also to be very relevant. Microwave back-scatter rises with canopy biomass and moisture, both linked to the period when maize loads Zn into the grain (Khabbazan et al., 2022).

Se stands out: its top-ranked features are almost exclusively spectral indices—CR_S1, PSRI_mean, DWSI_S1, and ARVI_mean—with elevation and soil variables appearing further down the ranking. This pattern suggests that Se distribution is more tightly linked to remote sensing signals, possibly reflecting influences like moisture or vegetation health in Se-limited regions.

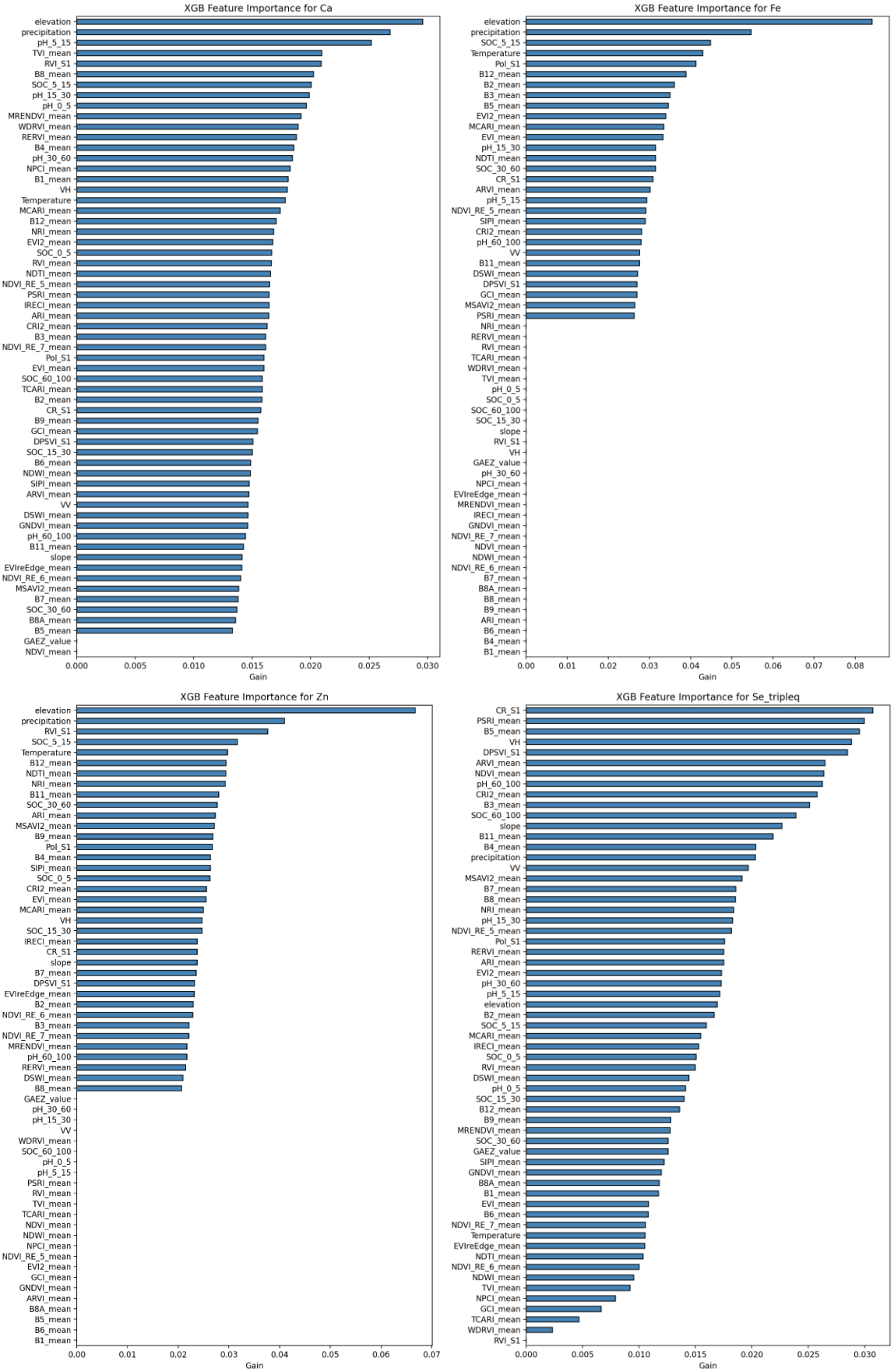


Figure 15 Feature importance ranking of XGBoost model in estimating Ca, Fe, Zn and Se_tripleq concentration in maize (by Gain value from high to low)

4.6. Partial dependence plots

The PDPs in Figure 16 describe how each model's three most influential covariates—plotted left-to-right in order of ranked importance—alter the predicted micronutrient content when all other features are held at their empirical distribution.

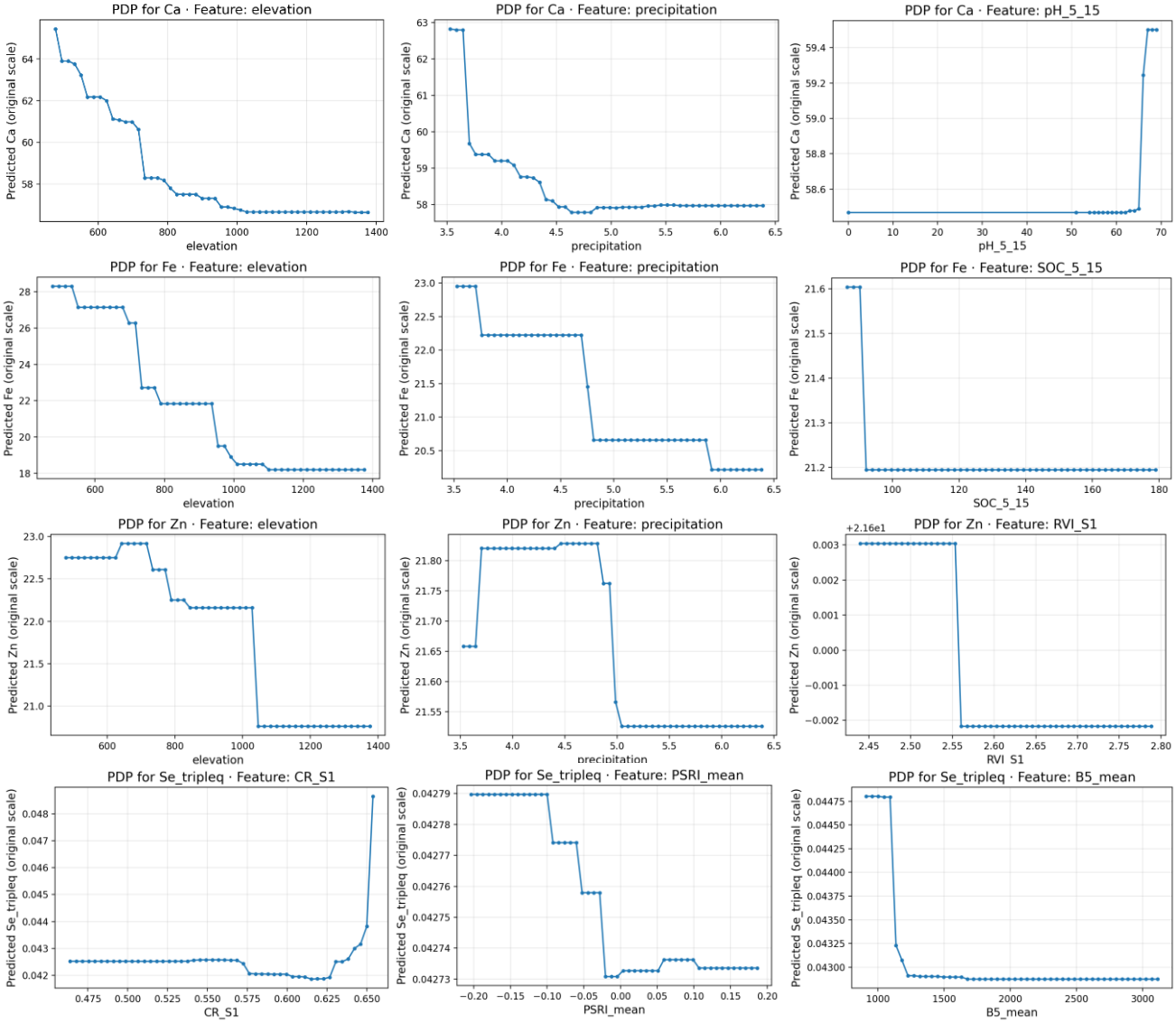


Figure 16 PDP of the top three features of each element in the XGBoost model (Top 1 to Top 3 from left to right)

In general, Ca, Fe, and Zn all show significant negative effects on the dual gradient of "altitude-precipitation" - when the altitude increases by about 700 m and the daily average precipitation exceeds ≈4.5 mm d^{-1}, the model prediction value drops synchronously, echoing the law of nutrient loss caused by intensified leaching and limited weathering of parent materials in mountainous areas and heavy rainfall scenarios (Kamal et al., 2023; Oishy et al., 2025).

For Ca, Fe and Zn the plots show a clear downward trend with rising elevation and heavier daily rainfall. Similar patterns have been reported in mountain soils where cooler and wetter settings slow weathering and enhance leaching, lowering micronutrient supply to maize grain (He et al., 2016; Li et al., 2019). Ca displays an additional jump when top-soil pH rises above about 6.8, matching the well-known increase in Ca^{2+} availability under mildly alkaline conditions (Obreza & Morgan, 2008).

Fe responds differently to chemistry rather than pH: once the 5–15 cm layer exceeds roughly 50 g kg^{-1} organic carbon, predicted Fe drops sharply and then levels out. High organic matter can bind or co-precipitate ferric oxides, making less Fe available for plant uptake, a mechanism described in agronomic extension literature (Schulte & Kelling, 1999).

Zn again declines with altitude and rainfall, but its third most important variable—the red-edge ratio RVI_S1—shows an abrupt cliff beyond about 2.55. Extremely nitrogen-rich canopies can dilute Zn in grain, an effect confirmed by large-scale fertiliser trials and meta-analyses (Liu et al., 2022). The model captures that threshold with an almost step-like decrease.

Se behaves quite differently. Its predicted concentration is highly sensitive to three features: CR_S1, PSRI_mean, and B5_mean. For CR_S1, Se levels remain relatively stable until approximately 0.64, beyond which there is a sharp increase, suggesting a threshold effect. PSRI_mean displays a generally negative relationship with Se, where higher PSRI_mean values lead to a noticeable decline in predicted Se. Lastly, B5_mean demonstrates a distinct breakpoint at around 1000; Se concentrations are higher below this value and drop sharply above it, indicating a strong inverse relationship beyond that point. These patterns suggest that Se distribution is influenced by nonlinear and threshold-dependent interactions with spectral and remote sensing variables (Molnar, 2020).

5. DISCUSSION

5.1. Interpretation of findings

The testing accuracy of our maize-nutrient models ($R^2 \approx 0.26$ for grain Zn) is largely a biological rather than a statistical limitation. Extreme genotype × management × environment (G × M × E) heterogeneity across Malawi dilutes the predictive signal and inflates the residuals. Nevertheless, the models and PDPs point to a coherent set of drivers led by soil pH, SOC, temperature and rainfall. A comparison with Ethiopia (Ofori-Karikari, 2024) shows that steeper altitudinal gradients there sharpen nutrient contrasts and make spatial prediction easier than in Malawi's relatively flat terrain.

5.1.1. Nutrients spatial distribution

Several factors contribute to higher micronutrient levels in maize grain from southern Malawi compared with the central and northern regions. First, the soils in the Shire Valley (southern Malawi) are often Vertisols formed on geologically distinct parent materials that naturally contain more Ca, Zn, Fe, and Se. These soil types also tend to have higher pH, which improves the availability of Zn and Se for plant uptake (Kumssa et al., 2022). Second, soil organic carbon is generally higher in the south and helps retain Zn in forms that maize can absorb. Third, the southern region has a slightly warmer climate and different moisture regimes (as indicated by higher mean annual temperature), both of which promote greater Se interactions and Zn solubility in the soil, respectively. These soil and climate conditions mean that corn grown in the south can accumulate more trace elements, whereas cooler temperatures, lower pH, and less organic matter in central and northern soils limit micronutrient availability and crop uptake.

5.1.2. Low testing accuracy

The findings of this study revealed a clear variation in model performance across different micronutrients, highlighting the complexity of nutrient–environment interactions and the limitations of current data

coverage. Among the four target elements (Ca, Fe, Zn, and Se), Zn and Ca demonstrated relatively better predictive results, with R^2 values of 0.14 and 0.15 respectively. These modest but meaningful predictions suggest that the spatial patterns of Zn and Ca in maize grain are largely governed by environmental variables included in the model, particularly soil pH, SOC, and topographic features. This aligns with established understanding that soil chemical conditions significantly influence Zn and Ca availability (Chaudhry & Loneragan, 1972; Thomas, 2016). In contrast, predictions for Fe and Se were notably weaker, with Se predictions even yielding negative R^2 values, indicating performance worse than a simple mean. Several factors likely contributed. First, Fe and Se distributions were highly skewed, with many low-range samples and a few extreme high values that the model failed to capture, resulting in consistent underestimation of outliers.

Malawian smallholders plant a mixture of hybrids and local open-pollinated varieties, each with distinct nutrient loading efficiencies (Nyirenda et al., 2021). Those genetics interact with highly variable on-farm practices that were not recorded but might influence grain composition. Se and Fe have fewer than 300 points, heavily clustered in the south-central plateau. A spatially stratified split therefore reserves some entire eco-regions solely for testing, forcing the model to inference beyond the training range —excellent for unbiased assessment, but leading to lower R^2.

Several key processes were missing from the input variables. No variable directly captures in-field fertiliser rates, liming history or short-term redox oscillations that govern Se speciation (Hu et al., 2023; Huang et al., 2017; Couture et al., 2015). The resulting omitted-variable bias is clearest in Se predictions, which regress toward the mean and yield negative R^2 despite visually coherent spatial patterns. Critical environmental drivers for these nutrients, such as soil redox potential and speciation, were missing from the input. Se bioavailability in soil is also strongly affected by redox state, and microbial activity, of which were not directly captured in our dataset (Saha et al., 2017; Zhang et al., 2024).

5.1.3. Key predictors to estimate micronutrient content in maize in Malawi

Feature importance analysis highlighted soil pH and SOC as consistent top predictors, verifying their central roles in micronutrient uptake (Thomas, 2016; Jalal et al., 2023). Sentinel-2 vegetation indices (e.g., NDVI, MSAVI2) also showed importance for some elements such as Ca and Se, indicating that canopy health reflects underlying soil nutrient status. Topographic factors such as elevation and slope were especially influential for Se, suggesting that water movement and microclimate contribute to Se mobilization.

Figure 15 ranks features by nutrient importance: elevation, seasonal precipitation, and topsoil pH determine Ca, Fe, and Zn, while cross-polarization radar ratio and red-edge vegetation index determine Se. Topsoil pH affects Ca and Zn because they are more soluble and less leached in near-neutral soils (Abedi et al., 2022); This explains why nutrient "hotspots" cluster on neutral Vertisols in the southern Shire Valley. Organic carbon comes in second; it complexes with Zn and buffers soil moisture, so areas with SOC > 2% have a slow but steady rise in predicted Zn across the PDP. Altitude and rainfall have a dual effect: with increasing elevation and daily rainfall exceeding ≈4.5 mm d^{-1}, cooler temperatures slow weathering, while leaching accelerates, dragging all three cationic nutrients downward simultaneously. Se is an outlier. Its uptake depends on instantaneous redox pulses rather than long-term averages; thus, radar-measured moisture proxies and stress indices mask static soil variables in the gain plot.

5.1.4. Interpretation of the partial-dependence plots

The soil-pH PDP derived for grain Zn increases almost linearly from pH ≈ 4.5 to 6.3 and then flattens, while the Ca PDP follows the same trajectory but bends more sharply near pH 6.8, indicating that both nutrients become progressively more available as strong acidity is neutralised before reaching an upper

limit in mildly alkaline soils (Botoman et al., 2022a). These results align with well-established chemistry: very acidic conditions keep Zn^{2+} strongly attached to oxide and clay surfaces, suppressing desorption, whereas the displacement of H^+ from the cation-exchange complex during liming releases both Zn and base cations such as Ca^{2+} into the soil solution (Hamzah Saleem et al., 2022). The subsequent plateau is consistent with the onset of hydroxy- and carbonate-Zn precipitation (e.g., $Zn(OH)_2$, $ZnCO_3$) that begins around pH 6.5–7 and caps further in soluble or exchangeable Zn (Saeed & Fox, 1977). Field data corroborate the model inflection: the highest grain-Zn concentrations occur on the neutral Vertisols of the Shire Valley in southern Malawi, where the has a pH typically above 6 and where the maize grain Zn is about 30 % higher than compared with the maize grown on the more acidic upland Ultisols (Gashu et al., 2021; Botoman et al., 2022a). Other PDPs reinforce element-specific controls. Predicted grain Zn rises with SOC content to roughly 2.5 %, then levels off, alignign with the previous studies that showed that moderate carbon enrichment enhances Zn mobilisation but that additional humification offers diminishing returns (Botoman et al., 2022a). Se behaves differently: it shows minimal direct response to pH or SOC in our model but increases sharply once the radar-based canopy-moisture index crosses a narrow threshold, emphasizing the sensitivity of the Se to episodic wetting/drying cycles and the redox oscillations in the root zone (Liao et al., 2014).

5.1.5. Malawi contrasts with Ethiopia

Ethiopia's cereal micronutrient maps, generated with Random-Forest models that fuse Sentinel-2 imagery and detailed soil layers, display far sharper gradients than Malawi's (Ofori-Karikari, 2024). Grain Zn and Fe concentrations often double between Rift-valley lowlands (<800 m) and highland plateaux (>2 500 m) within 250 km. Ofori-Karikari's (2024) thesis shows that elevation alone explains ≈35 % of the variance in Ethiopian grain Zn across 19 AEZs, versus <10 % in this Malawian study. From a SELPR perspective, where production-landscape resilience is assessed by the balance between ecological and socio-economic services, Ethiopia's topography create highly heterogeneous nutrient content in the cultivated crops. High-SELPR zones in the central highlands maintain dense organic soils and reliable rainfall, sustaining nutrient-rich cereals despite limited fertiliser access (Zhang et al., 2020). Malawi's relief produces broader, less contrasting zonation: its south-to-north decline in Zn and Ca is gradual.

5.1.6. Ecological explanation of the spatial nutrient patterns

There is an ecological explanation of the spatial nutrient patterns Higher Ca and Zn concentrations were observed in southern Malawi, corresponding with areas having neutral to slightly alkaline soils, warmer climates, and Vertisol parent geology, conditions known to retain these micronutrients more effectively (Thomas, 2016). Northern highland zones, typically cooler, more acidic, and erosion-prone showed lower concentrations, particularly for Se and Fe.

5.2. Limitations of this study

Despite its strengths, this study faces several important limitations that may affect the accuracy, transferability, and interpretability of its findings:

- Multi-resolution dataset inconsistencies: the integration of environmental predictors spanning several spatial resolutions introduces potential aggregation and scale mismatches. These inconsistencies can obscure fine-scale nutrient variability in maize grain.
- SoilGrids predictive uncertainty: soil input variables (e.g., pH, SOC) were derived from the global SoilGrids model, which estimates soil properties using machine learning and presents explicit uncertainty measures. Independent evaluations have shown substantial inaccuracy in regions with sparse ground observations (e.g., R^2 as low as 0.04–0.27 for texture in Croatia), raising concerns

about reliability in poorly sampled areas like parts of Malawi (Radocaj et al., 2023; Poggio et al., 2021).

- Limited ground-truth sample size and uneven distribution: although over 300 maize samples were collected, samples for Se and Fe remained under 300 and were spatially clustered. Such imbalance limits representativeness and increases variance, reducing model performance in under-sampled regions (Sharma et al., 2019; Moran et al., 2000).
- Incomplete environmental and agronomic covariates: important agronomic factors such as fertilization rates, soil amendments (e.g., $CaCO_3$, micronutrient fertilizer), irrigation, and soil microbial activity were absent. These are known to influence micronutrient uptake but were unrepresented in the model (Rahman et al., 2021).
- Mask uncertainty from maize delineation: the maize mask used to filter remote sensing-derived predictors is subject to classification error. As evidenced by decadal studies, even small misclassifications can bias nutrient predictions by inaccurately assigning environmental covariates (Liu et al., 2024).
- One time ground sampling without temporal assessment: around measurements were collected only during one growing season, limiting the model's ability to capture seasonal or inter-annual variability. Time-series ground measurements or repeated sampling are essential to understand temporal dynamics (Smith & Myers, 2018).
- Black-box model structure lacking processual controls: XGBoost's fully data-driven nature may capture spurious correlations when key biophysical processes (e.g., soil redox dynamics affecting Se) are missing. Hybrid process-driven and data-driven frameworks could reduce this risk and improve model robustness (Lu et al., 2023).

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion

This study shows that machine-learning models trained using satellite, climate, terrain and SoilGrids data can give a first, country-wide assessment of Ca, Fe, Zn and Se levels in the maize grain. Elevation, rainfall and top-soil pH emerged as the main controls on Ca, Fe and Zn, matching agronomic knowledge that these nutrients are most available on well-drained, near-neutral soils under moderate moisture. Se behaved differently: radar-based vegetation proved to be very relevant predictors, emphasizing the key role of short-term soil-moisture and redox changes that microwave sensors can detect. Although the model faced challenges with extreme nutrient values, especially for Se and Fe, the resulting maps already highlight "cold-spots" where hidden hunger is likely greatest, aligning with earlier field surveys in the country.

This work provides a spatial tool for targeting fertiliser, bio-fortified seed and nutrition programmes. Future iterations should incorporate detailed farm-management data, time-resolved moisture and redox proxies, and a loss-weighted training scheme so that the model gives appropriate emphasis to the nutritionally critical—but infrequently observed—high-concentration samples.

6.2. Future work

While this study establishes a valuable spatial mapping framework for maize micronutrients in Malawi, several extensions could substantially enhance its predictive power and applicability.

First, adopting weighted loss functions tailored to high-concentration or rare samples such as those found in Se and Fe can mitigate bias toward common values. Weighted MSE or output-weighted losses have proven effective in improving rare-event estimation in imbalanced regression contexts (Ren et al., 2022; Zhou et al., 2023). Additionally, data augmentation for regression, which combines resampling and synthetic sample generation, can help expand the training distribution and improve model robustness.

Enriching the feature set by integrating additional environmental and soil parameters including cation exchange capacity, $CaCO_3$ content, irrigation indices, and soil microbial proxies would help account for variables currently missing from the model. These factors are known to significantly influence micronutrient bioavailability but are not captured by standard remote sensing or existing soil datasets at the global level.

Feature reduction methods such as recursive feature elimination or regularization can also be investigated to identify and remove redundant or noisy predictors, thereby reducing overfitting and improving interpretability.

By integrating these enhancements, namely advanced learning strategies (weighted loss, augmentation), enriched predictors, expressive ensemble models, and streamlined feature sets, the proposed framework can yield more accurate, robust, and interpretable predictions. These improvements are essential for deploying reliable nutrient-mapping tools to support precision agriculture and address micronutrient deficiencies across smallholder farms.

6.3. Social impact and policy relevance

This study offers a powerful tool for identifying regions with low micronutrient levels in maize, especially Zn, Fe, and Se, helping to identify where hidden hunger is most pressing. By translating environmental and crop data into actionable maps, it supports targeted interventions like biofortification and micronutrient-enriched fertilization rather than broad, untailored efforts.

Stakeholders—including farmers, extension services, NGOs, and government agencies—can use this framework to pinpoint the most nutrient-deficient areas and channel scarce inputs — such as fertiliser subsidies, bio-fortified seed, advisory visits, and credit — toward communities where they will yield the greatest nutritional and economic return. Better-aligned agricultural practices not only boost crop yield and grain quality but also reinforce livelihoods and build more resilient food systems.

While the current study is grounded in agricultural science, it establishes a strong platform for future public-health collaborations that could drive targeted interventions to reduce micronutrient deficiencies, strengthen food security, and ultimately enhance community well-being at scale.

6.4. Ethical Considerations

In this study, the crop nutrient data was provided by the GeoNutrition project. Informed consent was obtained from every farmer before grain and soil sampling took place in both field plots and grain stores (Kumssa et al., 2022). The project received formal clearance from the University of Nottingham's School of Sociology and Social Policy Research Ethics Committee (REC) under reference BIO-1718-0004 for the activities carried out in Malawi. This approval was also endorsed by the Directors of Research at Lilongwe University of Agriculture and Natural Resources (Malawi).

AI-assisted tools have been employed to smoothen language and check for grammar errors.

LIST OF REFERENCES

Abedi, T., Gavanji, S., & Mojiri, A. (2022). Lead and zinc uptake and toxicity in maize and their management. Plants, 11(15), 1922. https://doi.org/10.3390/plants11151922

Aliyu, K. T., Huising, J., Kamara, A. Y., Jibrin, J. M., Mohammed, I. B., Nziguheba, G., … Vanlauwe, B. (2021). Understanding nutrient imbalances in maize (Zea mays L.) using the diagnosis and recommendation integrated system (DRIS) approach in the Maize belt of Nigeria. Scientific Reports, 11(1), 16018. https://doi.org/10.1038/s41598-021-96018-0

Andrés-Caballero, R., García-Ferrer, A., Benito-Andrés, P., & Moral, F. J. (2023). Spatial prediction of soil micronutrients using supervised self-organising maps. Nutrient Cycling in Agroecosystems, 125(3), 345–360. https://doi.org/10.1007/s10705-023-10303-y

Asamoah, E., Heuvelink, G. B. M., Chairi, I., Bindraban, P. S., & Logah, V. (2024). Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana. Heliyon, 10, e37065. https://doi.org/10.1016/j.heliyon.2024.e37065

Bastiaanssen, W. G., Molden, D. J., & Makin, I. W. (2000). Remote sensing for irrigated agriculture: examples from research and possible applications. Agricultural Water Management, 46(2), 137–155. https://doi.org/10.1016/S0378-3774(00)00080-9

Belgiu, M., & Drăguţ, 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

Berger, K., Verrelst, J., Féret, J.-B., Hank, T., Wocher, M., Mauser, W., & Camps-Valls, G. (2020). Retrieval of aboveground crop nitrogen content with a hybrid machine learning method. International Journal of Applied Earth Observation and Geoinformation, 92, 102174. https://doi.org/10.1016/j.jag.2020.102174

Bouis, H. E., & Saltzman, A. (2017). Improving nutrition through biofortification: A review of evidence from HarvestPlus, 2003 through 2016. Global Food Security, 12, 49–58. https://doi.org/10.1016/j.gfs.2017.01.009

Botoman, L., Chagumaira, C., Mossa, A. W., Amede, T., Ander, E. L., Bailey, E. H., Chimungu, J. G., Gameda, S., Gashu, D., Haefele, S. M., Joy, E. J. M., Kumssa, D. B., Ligowe, I. S., McGrath, S. P., Milne, A. E., Munthali, M., Towett, E. K., Walsh, M. G., Wilson, L., Young, S. D., Broadley, M. R., Lark, R. M., & Nalivata, P. C. (2022a). Soil and landscape factors influence geospatial variation in maize grain zinc concentration in Malawi. Scientific Reports, 12(1), 7986. https://doi.org/10.1038/s41598-022-12014-w

Botoman, L., Chimungu, J. G., Bailey, E. H., Munthali, M. W., Ander, E. L., Mossa, A. W., … & Nalivata, P. C. (2022b). Agronomic biofortification increases grain zinc concentration of maize grown under contrasting soil types in Malawi. Plant Direct, 6(11), e458. https://doi.org/10.1002/pld3.458

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cereal grain mineral micronutrient and soil chemistry data from GeoNutrition surveys in Ethiopia and Malawi. Scientific Data 2022 9:1, 9(1), 1–12. https://doi.org/10.1038/s41597-022-01500-5

Chaudhry, M. A., & Loneragan, J. F. (1972). Effect of pH on the uptake of zinc and cobalt from soil by subterranean clover. Plant and Soil, 36(1), 171-182. https://doi.org/10.1007/BF01373468

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

Chen, T., Li, J., Wang, L., & Zhang, X. (2021). Interactive effects of temperature and redox on soil carbon and iron cycling. Soil Biology & Biochemistry, 160, 108337. https://doi.org/10.1016/j.soilbio.2021.108337

Chilimba, A. D. C., Young, S. D., Black, C. R., Rogerson, K. B., Ander, E. L., Watts, M. J., Lammel, J., & Broadley, M. R. (2011). Maize grain and soil surveys reveal suboptimal dietary selenium intake is widespread in Malawi. Scientific Reports, 1, 72. https://doi.org/10.1038/srep00072

Couture, R. M., Charlet, L., Markelova, E., Madé, B., & Parsons, C. T. (2015). On–off mobilization of contaminants in soils during redox oscillations. Environmental science & technology, 49(5), 3015-3023. https://doi.org/10.1021/es5056669

Gatti, V. C. M., Barata, H. S., Silva, V. F. A., Cunha, F. F., Oliveira, R. A., Oliveira, J. T., & Silva, P. A. (2023). Influence of calcium on the development of corn plants grown in hydroponics. AgriEngineering, 5(1), 623–630. https://doi.org/10.3390/agriengineering5010039

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernández, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Ennen, R., Thorson, E., Jeschke, M., Hoss, N., & Johnson, J. (2021, March 26). Building better quality seed with rigorous vigor testing. Pioneer Seeds. Retrieved May 6, 2025, from https://www.pioneer.com/us/agronomy/building-better-quality-seed-with-rigorous-vigor-testing.html

EOS Data Analytics. (2022). Chlorophyll index in agriculture: CIred-edge explained. https://eos.com/make-an-analysis/chlorophyll-index/

Fathi, M., Shah-Hosseini, R., & Moghimi, A. (2023). Enhancing corn yield prediction in Iowa: A concatenate-based 2D-CNN-BiLSTM model with integration of Sentinel-1/2 and SoilGRIDs data. Environmental Sciences Proceedings, 29(1), 2. https://doi.org/10.3390/ECRS2023-15852

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29, 1189–1232. ttps://doi.org/10.1214/aos/1013203451

Food and Agriculture Organization of the United Nations & International Institute for Applied Systems Analysis. (2021). Global Agro-Ecological Zones (GAEZ v4) – Data Portal user's guide. Rome: FAO. https://doi.org/10.4060/cb5167en

Galani, Y. J. H., Ligowe, I. S., Kieffer, M., Kamalongo, D., Kambwiri, A. M., Kuwali, P., … & Orfila, C. (2022). Conservation agriculture affects grain and nutrient yields of maize (Zea mays L.) and can impact food and nutrition security in Sub-Saharan Africa. Frontiers in Nutrition, 8, 804663. https://doi.org/10.3389/fnut.2021.804663

Gashu, D., Nalivata, P. C., Amede, T., Ander, E. L., Bailey, E. H., Botoman, L., … & Broadley, M. R. (2021). The nutritional quality of cereals varies geospatially in Ethiopia and Malawi. Nature, 594(7861), 71–76. https://doi.org/10.1038/s41586-021-03559-3

Gödecke, T., Stein, A. J., & Qaim, M. (2018). The global burden of chronic and hidden hunger: trends and determinants. Global Food Security, 17, 21–29. https://doi.org/10.1016/j.gfs.2018.05.001

Gohil, J. H. (2023). Estimating micronutrient concentrations in maize grains with Sentinel-1 and -2 images (Master's thesis). University of Twente.

Grabowski, P., Slater, D., Gichohi-Wainaina, W., Kihara, J., Chikowo, R., Mwangwela, A., … & Bekunda, M. (2024). Research agenda for holistically assessing agricultural strategies for human micronutrient deficiencies in east and southern Africa. Agricultural Systems, 220, 104094. https://doi.org/10.1016/j.agsy.2024.104094

Hall, S. J., Curtinrich, H. J., & Sebestyen, S. D. (2022). Simulated hydrological dynamics and coupled iron redox cycling in peatlands under climate change. Journal of Geophysical Research: Biogeosciences, 127, e2021JG006662. https://doi.org/10.1029/2021JG006662

Hamzah Saleem, M., Usman, K., Rizwan, M., Al Jabri, H., & Alsafran, M. (2022). Functions and strategies for enhancing zinc availability in plants for sustainable agriculture. Frontiers in Plant Science, 13, 1033092. https://doi.org/10.3389/fpls.2022.1033092

He, X., Hou, E., Liu, Y., & Wen, D. (2016). Altitudinal patterns and controls of plant and soil nutrient concentrations and stoichiometry in subtropical China. Scientific Reports, 6, 24235. https://doi.org/10.1038/srep24235

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez González, M., Kilibarda, M., Blagotić, A., … 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

Holden, S. T., & Fisher, M. (2015). Subsidies promote use of drought tolerant maize varieties despite variable yield performance under smallholder environments in Malawi. Food Security, 7, 1225–1238. https://doi.org/10.1007/s12571-015-0494-y

Hu, C., Nie, Z., Shi, H., Peng, H., Li, G., Liu, H., ... & Liu, H. (2023). Selenium uptake, translocation, subcellular distribution and speciation in winter wheat in response to phosphorus application combined with three types of selenium fertilizer. BMC Plant Biology, 23(1), 224. https://doi.org/10.1186/s12870-023-04227-6

Huang, G., Ding, C., Guo, F., Li, X., Zhang, T., & Wang, X. (2017). Underlying mechanisms and effects of hydrated lime and selenium application on cadmium uptake by rice (Oryza sativa L.) seedlings. Environmental Science and Pollution Research, 24, 18926-18935. https://doi.org/10.1007/s11356-017-9510-7

Ichami, S. M., Karuku, G. N., Sila, A. M., Ayuke, F. O., & Shepherd, K. D. (2022). Spatial approach for diagnosis of yield-limiting nutrients in smallholder agroecosystem landscape using population-based farm survey data. PLOS ONE, 17(2), e0262754. https://doi.org/10.1371/journal.pone.0262754

Joy, E. J. M., Broadley, M. R., Young, S. D., Black, C. R., Chilimba, A. D. C., Ander, E. L., Barlow, T. S., & Watts, M. J. (2015). Soil type influences crop mineral composition in Malawi. Science of the Total Environment, 505, 587–595. https://doi.org/10.1016/j.scitotenv.2014.10.038

Kamal, A., Mian, I. A., Akbar, W. A., Rahim, H. U., Irfan, M., Ali, S., ... & Zaman, W. (2023). Effects of soil depth and altitude on soil texture and soil quality index. Applied Ecology & Environmental Research, 21(5). https://doi.org/10.15666/aeer/2105_

Karlson, M., Ostwald, M., Bayala, J., Bazié, H. R., Ouedraogo, A. S., Soro, B., … & Reese, H. (2020). The potential of Sentinel-2 for crop production estimation in a smallholder agroforestry landscape, Burkina Faso. Frontiers in Environmental Science, 8, 85. https://doi.org/10.3389/fenvs.2020.00085

Khabbazan, S., Steele-Dunne, S. C., Vermunt, P., Judge, J., Vreugdenhil, M., & Gao, G. (2022). The influence of surface canopy water on the relationship between L-band backscatter and biophysical variables in agricultural monitoring. Remote Sensing of Environment, 268, 112789. https://doi.org/10.1016/j.rse.2021.112789

Kumssa, D. B., Mossa, A. W., Amede, T., Ander, E. L., Bailey, E. H., Botoman, L., ... & Nalivata, P. C. (2022). Cereal grain mineral micronutrient and soil chemistry data from GeoNutrition surveys in Ethiopia and Malawi. Scientific Data, 9(1), 443. https://doi.org/10.1038/s41597-022-01500-5

Lee, D., Davenport, F., Shukla, S., Husak, G., Funk, C., Harrison, L., … & Verdin, J. (2022). Maize yield forecasts for Sub-Saharan Africa using Earth Observation data and machine learning. Global Food Security, 33, 100643. https://doi.org/10.1016/j.gfs.2022.100643

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., & Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. Global Change Biology, 25(7), 2325–2337. https://doi.org/10.1111/gcb.14628

Liao, Q., Xing, Y., Li, A. M., Liang, P. X., Jiang, Z. P., Liu, Y. X., & Huang, D. L. (2024). Enhancing selenium biofortification: strategies for improving soil-to-plant transfer. Chemical and Biological Technologies in Agriculture, 11(1), 148. https://doi.org/10.1186/s40538-024-00581-1

Liu, Z., Cakmak, I., & White, P. J. (2022). Global analysis of nitrogen fertilization effects on grain zinc and iron concentrations in cereals. Environmental and Experimental Botany, 201, 104610. https://doi.org/10.1016/j.envexpbot.2022.104610

Mahesh, P., & Soundrapandiyan, R. (2024). Yield prediction for crops by gradient-based algorithms. PLoS ONE, 19(8), e0291928. https://doi.org/10.1371/journal.pone.0291928

Jalal, A., Júnior, E. F., & Teixeira Filho, M. C. M. (2024). Interaction of zinc mineral nutrition and plant growth-promoting bacteria in tropical agricultural systems: a review. Plants, 13(5), 571. https://doi.org/10.3390/plants13050571

Mhlanga, B., Mwila, M., & Thierfelder, C. (2021). Improved nutrition and resilience will make conservation agriculture more attractive for Zambian smallholder farmers. Renewable Agriculture and Food Systems, 36(5), 443–456. https://doi.org/10.1017/S1742170521000028

Mloza Banda, M. L., Cornelis, W., & Mloza Banda, H. R. (2024). Seasonal, decadal, and El Niño–Southern Oscillation-related trends and anomalies in rainfall and dry spells during the agricultural season in central Malawi. Geographies, 4(3), 563–582.

Molnar, C. (2020). Interpretable machine learning. Lulu. com. https://christophm.github.io/interpretable-ml-book/

Nyirenda, H., Mwangomba, W., & Nyirenda, E. M. (2021). Delving into possible missing links for attainment of food security in Central Malawi: Farmers' perceptions and long term dynamics in maize (Zea mays L.) production. Heliyon, 7(5). https://doi.org/10.1016/j.heliyon.2021.e07010

Obreza, T. A., & Morgan, K. T. (2008). Soil pH and nutrient availability (SL256). University of Florida IFAS Extension. https://edis.ifas.ufl.edu/publication/SS480

Ofori-Karikari, K. A. G. (2024). Integrating environmental data with satellite imagery for large-scale crop grain nutrient mapping (Master's thesis, University of Twente).

Oishy, M. N., Shemonty, N. A., Fatema, S. I., Mahbub, S., Mim, E. L., Raisa, M. B. H., & Anik, A. H. (2025). Unravelling the effects of climate change on the soil-plant-atmosphere interactions: A critical review. Soil & Environmental Health, 100130. https://doi.org/10.1016/j.seh.2025.100130

Olisah, C. C., Smith, L., Smith, M., Lawrence, M. O., & Ojukwu, O. (2024). Corn yield prediction model with deep neural networks for smallholder farmer decision support system. arXiv Preprint arXiv:2401.03768.

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation, 15(1). https://doi.org/10.7275/qbpc-gk45

Osco, L. P., Ramos, A. P. M., Faita Pinheiro, M. M., Moriya, É. A. S., Imai, N. N., Estrabis, N., Ianczyk, F., de Araújo, F. F., Liesenberg, V., & de Castro Jorge, L. A. (2020). A machine learning framework to predict nutrient content in Valencia-orange leaf hyperspectral measurements. Remote Sensing, 12(6), 906. https://doi.org/10.3390/rs12060906

Páez-Bimos, S., Molina, A., Calispa, M., Delmelle, P., Lahuatte, B., Villacís, M., ... & Vanacker, V. (2023). Soil–vegetation–water interactions controlling solute flow and chemical weathering in volcanic ash soils of the high Andes. Hydrology and Earth System Sciences, 27(7), 1507-1529. https://doi.org/10.5194/hess-27-1507-2023

Pingali, P., & Sunder, N. (2017). Transitioning toward nutrition-sensitive food systems in developing countries. Annual Review of Resource Economics, 9(1), 439–459. https://doi.org/10.1146/annurev-resource-100516-053552

Ren, J., Zhang, M., Yu, C., & Liu, Z. (2022). Balanced mse for imbalanced visual regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7926-7935). https://doi.org/10.1109/CVPR52688.2022.00778

Saeed, M., & Fox, R. L. (1977). Relations between suspension ph and zinc solubility in acid and calcareous solls. Soil Science, 124(4), 199-204. https://doi.org/10.1097/00010694-197710000-00003

Saha, U., Fayiga, A., & Sonon, L. (2017). Selenium in the soil–plant environment: A review. International Journal of Applied Agricultural Sciences, 7(1), 1–18. https://doi.org/10.11648/j.ijaas.20170701.11

Sarr, A. B., & Sultan, B. (2023). Predicting crop yields in Senegal using machine learning methods. International Journal of Climatology, 43(4), 1817–1838. https://doi.org/10.1002/joc.7814

Schulte, E. E., & Kelling, K. A. (1999). Soil and Applied Iron (A3554). University of Wisconsin Extension. https://learningstore.extension.wisc.edu/products/soil-and-applied-iron-a3554

Tanaka, T., & Gislum, R. (2025). Prediction of winter wheat nitrogen status using UAV imagery, weather data, and machine learning. European Journal of Agronomy, 164, 127534. https://doi.org/10.1016/j.eja.2025.127534

Thenkabail, P. S., Ceccato, P., Patel, N., Wotton, B., & Mouat, J. (2017). Using high-resolution Sentinel-2 imagery to monitor crop condition and forecast drought in agricultural landscapes. International Journal of Applied Earth Observation and Geoinformation, 55, 123–133. https://doi.org/10.1016/j.jag.2016.10.011

Thomas, G. W. (2016). Soil pH [Wikipedia entry]. Soil acidity controls the availability of micronutrients including Zn, Ca, and Fe.

The World Bank. (2017). Malawi maize mask for 2017 [Data set]. World Bank Data Catalog. https://datacatalog.worldbank.org/dataset/0037935

Voss, R. (1998). Micronutrients. Iowa State University Soil Fertility. Retrieved June 14, 2025, from https://www.agronext.iastate.edu/soilfertility/info/Micronutrients_VossArticle.pdf

Wang, Y., Yu, T., Yang, Z., Bo, H., Lin, Y., Yang, Q., Liu, X., Zhang, Q., Zhuo, X., & Wu, T. (2021). Zinc concentration prediction in rice grain using back-propagation neural network based on soil properties and safe utilization of paddy soil: A large-scale field study in Guangxi, China. Science of the Total Environment, 798, 149270. https://doi.org/10.1016/j.scitotenv.2021.149270

Wessells, K. R., & Brown, K. H. (2012). Estimating the global prevalence of zinc deficiency: Results based on zinc availability in national food supplies and the prevalence of stunting. PLoS ONE, 7(11), e50568. https://doi.org/10.1371/journal.pone.0050568

World Health Organization. (2023). Micronutrient deficiencies. Retrieved from https://www.who.int/data/nutrition

Zhang, H., Chen, H., Geng, T., Liu, D., & Shi, Q. (2020). Evolutionary characteristics and trade-offs' development of social–ecological production landscapes in the loess plateau region from a resilience point of view: A case study in Mizhi County, China. International Journal of Environmental Research and Public Health, 17(4), 1308. https://doi.org/10.3390/ijerph17041308

Zhang, X., et al. (2024). Selenium dynamics in plants: Uptake, transport, toxicity, and implications. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2023.169246

Zhou, Z., Zheng, C., Liu, X., Tian, Y., Chen, X., Chen, X., & Dong, Z. (2023). A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data. Remote Sensing, 15(7), 1768. https://doi.org/10.3390/rs15071768