EVALUATING SUSTAINABLE INTENSIFICATION IN SUB-SAHARAN AFRICA: APPROACHES LINKING ECOSYSTEM SERVICES TO LIVELIHOOD CAPITALS AND AGRICULTURAL DECISION-MAKING

By

TAISHA VENORT

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA
2023

To my parents, my greatest supporters

ACKNOWLEDGMENTS

I thank my advisors Dr. Cheryl Palm and Dr. Rafael Muñoz-Carpena for their mentorship, academic advising and unwavering support throughout my PhD journey. I would also like to thank my advisory committee members, Dr. Gregory Kiker, Dr. Jiangxiao Qiu and Dr. David Kaplan for their valuable insights on my work. I would also like to thank my colleague and mentor Dr. Geraldine Klarenberg for her diligence with providing parts of the data needed to accomplish this work and her insights, and my colleague Alvaro Carmona for his help throughout. I am also very thankful towards the leadership and administrative community in the Department of Agricultural and Biological Engineering at the University of Florida, particularly Dr. Kati Migliaccio, Shannon Noble and many others for their unwavering support in the pursuits of fellowships and scholarships. I am also grateful for all colleagues and friends of ABE - GSO and the ABE Biocomplexity group for making this journey a pleasant one. Finally, I thank my parents, my family, and my friends for all their support and encouragement throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

Table page

LIST OF FIGURES

LIST OF ABBREVIATIONS

CP	Crop Production
CS	Carbon Storage
CII	Combined Input Intensity
ES	Ecosystem Services
ESC	Ecosystem Services Cluster
FC	Financial Capital
GSA	Global Sensitivity Analysis
GSUA	Global Sensitivity and Uncertainty Analysis
HC	Human Capital
HDC	Human-derived Capital
K_PNB	Potassium Partial Nutrient Budget
LC	Livelihood Capitals
LR	Logistic Regression
LSC	Livelihood Strategy Cluster
MCF	Monte Carlo Filtering
MI	Management Intensity
ML	Machine Learning
NC	Natural Capital
N_PNB	Nitrogen Partial Nutrient Budget
PC	Physical Capital
P_PNB	Phosphorous Partial Nutrient Budget
RF	Random Forest

SAGCOT	The Southern Agricultural Growth Corridor of Tanzania (SAGCOT)
SAI	Sustainable Agricultural Intensification
SC	Social Capital
SIIL	Sustainable Intensification Innovation Lab
SMI	Soil Management Intensity
SOM	Self-Organizing Map
SSA	Sub-Saharan Africa
VIF	Variance Inflation factor
WS	Water Storage

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

EVALUATING SUSTAINABLE INTENSIFICATION IN SUB-SAHARAN AFRICA: APPROACHES LINKING ECOSYSTEM SERVICES TO LIVELIHOOD CAPITALS AND AGRICULTURAL DECISION-MAKING

By

Taisha Venort

December 2023

Chair: Rafael Muñoz-Carpena
Co-chair: Cheryl Palm
Major: Agricultural & Biological Engineering

Agricultural systems are multi-dimensional and complex. In the context of advancing sustainable intensification (*increase in production per unit of land while mitigating environmental and social impacts*) in Sub-Saharan Africa (SSA),novel frameworks for evaluating the performance of agricultural systems must be more attentive to the heterogeneity of micro-scale decision-making and socio-economic factors affecting productivity, and to farm multifunctionality metrics by which performance can be further evaluated beyond yield.

First, this work proposes a farm/household scale modeling framework that links social-ecological explanatory indicators (i.e., livelihood capitals, intensification decision-making) to ecosystem services performance metrics (i.e., crop yield and soil health indicators). The framework was applied in two smallholder landscapes in SSA, for 210 farming households in the Southern Agricultural Corridor of Tanzania, and 500 farming households in the Upper Ewaso Ng'iro basin of Kenya. Second, statistical and clustering analyses were conducted to identify patterns of relationships amongst livelihood capital, management intensity and ecosystem services indicators in both case

studies. Results from the analyses provided evidence of synergistic relationships between human-derived capital, farm input intensity and crop production across case studies. Third, a suite of methods including machine learning predictive modeling approaches, high dimensional factor importance analysis, and monte carlo filtering, were applied to one case study to inform the design of effective intervention strategies for sustainable intensification. The findings revealed highly interactive soil health indicators (i.e.,soil nutrient balances), household-level and landscape level human and physical capital indicators (i.e., labor, distance to river, dependency ratio, input use) as the most influential factors driving productivity. Effective intervention strategies able to successfully move all low productivity farms (86% with maize yield below 1.5 t/ha) to mid-productivity (89% with maize yield between 1.5 and 2.5 t/ha) and high productivity (10% with maize yield above 2.5 t/ha) encompassed nutrient imbalance remediation, soil health, farm/household physical and human capital development interventions.

This work is new in linking livelihood strategies to ecosystem services relationships, and in successfully modeling productivity using such approaches. It makes a compelling case to agricultural scientists and policymakers for further integrating social-ecological approaches in agricultural systems modeling.

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

Sustainably meeting food demands of a growing population is one of humanity's greatest future challenge. Sub-Saharan Africa (SSA) comprises of fastest growing populations globally, and is projected to face serious food insecurity and livelihood risks by mid-century, placing significant pressure on the region to increase food production while preserving natural capital under rising climate and soil degradation risks (Bradshaw & Di Minin, 2019; Tully et al., 2015; Van Ittersum et al., 2016). Over the last two decades, agricultural extensification (expansion *of cropland area*) has played the larger role in the region's increase in food production over agricultural intensification *(increase in production per unit of input on existing cropland area*) (Nin Pratt, 2015). Despite these efforts, average yields for staple grains grown across the region such as maize, wheat and rice are still low compared to global averages (Erenstein et al., 2022; Jayne & Sanchez, 2021; USDA, 2022; Yuan et al., 2021). If yield gaps for staple crops are to be narrowed while adverse effects of cropland expansion also mitigated, agricultural intensification will need to play the leading role in meeting the region's 60% increase in food demand by 2050 (Jayne et al., 2018; Van Ittersum et al., 2016). More importantly, if natural capital *(stocks of natural resources)* and ecosystem services (*the benefits people obtain from ecosystems*' *natural resources*) are to be sustained, and food insecurity and livelihood risks reduced in the region, sustainable intensification (SAI) (*increase in production per unit of land while mitigating environmental and social impacts)* will need to be at the forefront of advisable solutions (Godfray, 2015; Pretty & Bharucha, 2014).

Sustainable intensification practices are being widely promoted in SSA. Practices being promoted focus on soil and water conservation, improved nutrient management, and cropping systems diversity, among many others (Holden, 2018; Kuyah et al., 2021; Pretty et al., 2011). Conventional approaches to evaluating the performance of sustainable intensification strategies focus on comparing agronomic impacts of practices (Fortmann, 2011; Kuyah et al., 2021). More integrated approaches are recently proposed towards more a comprehensive suite of metrics able to capture the agronomic, environmental, socio-economic and sustainability dimensions of different agricultural systems (Mouratiadou et al., 2021; Musumba et al., 2021; Smith et al., 2017). The Sustainable intensification framework has been mostly used for evaluating sustainable intensification practices in SSA (Eissler et al., 2021; Hammond et al., 2021; Musumba et al., 2021; Rahman et al., 2020). It considers a diverse set of indicators under their respective domains (productivity, economic, environment, human condition, and social) (Musumba et al., 2017, 2021).

Although these approaches allow for a broader understanding of the relative performance of sustainable intensification practices across agricultural systems, they fall short of capturing relationships between farm characteristics, productivity, and sustainability performance, all crucial for a more comprehensive evaluation of farm performance. Farmers combine several management decisions about agricultural practices to achieve agronomic goals (Cassman & Grassini, 2020). These decisions about livelihood strategies are mainly based on their level of different capitals/resources (Pretty, 2008). Useful metrics are needed to capture the wide range of agricultural intensification strategies of smallholders. Furthermore, each strategy affects soil ecosystem services functions necessary to achieve crop production differently, considering the heterogeneity of the soil conditions of farms. Soil health metrics that pertains to soil sustainability, particularity in SSA where soil degradation challenges are notable, are important sustainability metrics to consider in farm performance (Tully et al., 2015; Vanlauwe et al., 2015).

We have identified the following research gaps in our current knowledge on sustainable intensification evaluation in SSA:

- The need to account for the heterogeneity of micro-scale decision-making and socio-economic factors that can affect farm performance.
- The need to account for farm multifunctionality (i.e., relationships between many ecosystem service benefits of farming).

These research gaps have informed a framework (Figure 1) for defining research questions around the evaluation of sustainable intensification in SSA.

1.2 Conceptual Framework

We propose a social-ecological system (*interdependent human-natural systems*) (Biggs et al., 2021)) framework (Figure 1-3) at the farm scale. This framework is a merging of the livelihood capital concept adapted from the Sustainable Livelihoods Framework (Jones et al., 2016; Scoones, 1998; Solesbury, 2005) (Figure 1-1) with the system approach to ecosystem service social-ecological concept (Figure 1-2) (B. Fisher et al., 2008; Jones et al., 2016), where natural capital and human-derived capitals (i.e., human, social, financial) are established as co-producers of ecosystem services via interactions with users/beneficiaries. The framework integrates the following components:

(1) Social-ecological factors (Figure 1-3)

Livelihood capitals (LC) - The livelihood base of the farm unit (i.e., farm household) gathers natural, human, social, financial, physical capitals (Figure 1-1) deployed in pursuing livelihood strategies such as agricultural intensification. Natural capital involves stocks of natural resources. These resources include water, soil, trees, and many others. Human capital involves educational and training skills, and availability of labor. Social capital includes integration and networking through social or political institutions. Financial capital includes economic assets such as cash, credit, debts. Physical capital includes transportation and communication infrastructures, machinery assets (Carney, 2003; Scoones, 1998; Solesbury, 2005). Data on human-derived capitals must be collected at the farm-household level, some data on infrastructure can be collected at the landscape scale. Capital levels are assessed using evaluation scores based on a set of indicators embracing the different types of capitals.

Management Intensity Decision-making (MI) – To account for the heterogeneity of management decision-making at the farm/household scale, intensification strategies across diverse farms are quantified and characterized. Here, we propose two metrics of management intensity : a combined farm input intensity index and a soil management intensity index. The combined farm input intensity index is used to capture the intensity of farming practices. It is obtained by ranking the grouping structure of the use or no use of inputs based on their relative potential impact on productivity in the following order: fertilizer use > improved seeds use > mechanization > irrigation> pesticide use. The soil management intensity index is obtained by ranking the grouping structure of practices based on their relative impact on soil disturbance in the following order: soil tillage > residue retention practice. The indexes value increases with more combinations of practices.

(2) Farm multifunctionality performance metrics (Figure 1-3)

Ecosystem services (ES) – Farm performance is measured through ecosystem services to capture the multi-functional benefits of farming through yield and relevant on-farm ecosystem services, as productivity and sustainability performance metrics, respectively (Garbach et al., 2017; Robertson et al., 2014). We focus on food provisioning services (i.e., primary crops production), and two soil health related regulating services (i.e., soil carbon storage, soil water storage), and one soil health related supporting services (i.e., soil nutrient supply). Indicators were established for quantifying these services. Provisioning services were measured using crop yield. Soil health functions were established for regulating and supporting services. These soil health indicators were selected to best reflect the impact of farm management practices on soils. Additional information on the soil health functions proposed are provided in Appendix A.

The conceptual framework proposed is explored in two case studies of smallholder farming systems (*land size less than 2 hectares*) (Lowder et al., 2016) in Sub-Saharan Africa: The Southern Agricultural Growth Corridor of Tanzania (SAGCOT) and the Upper Ewaso Ngiro Basin in Kenya. These regions are representative examples of complex social-ecological systems, characterized by heterogenous farming households operating under diverse agro-ecological conditions, both facing pressure to increase food production while preserving natural resources.

1.3 Research Questions, Objectives and Dissertation Outline

Based on the research gaps identified, this research aims to answer the following overarching research questions (RQ1, RQ2) and their research objectives (OB1, OB2, OB3, OB4) :

RQ1: What are the relationships patterns between social-ecological factors (i.e., livelihood capitals, management intensity) and ecosystem services (i.e., crop yield, soil carbon and water storage, soil nutrient supply) in smallholder systems of SSA?

> OB1 : Test for significant trade-offs and synergies between ecosystem services, livelihood capitals, and management intensity.
>
> OB2 : Identify dissimilar group of farms based on ecosystem services relationships and livelihood capitals/management intensity relationships.

RQ2: Can social-ecological factors (i.e., livelihood capitals, management intensity), and soil health indicators (i.e., soil carbon and water storage, soil nutrient supply) be good predictors of productivity (i.e., crop production) ? If so, how can these relationships be modeled to inform the design of intervention strategies able to boost farm productivity in smallholder systems of SSA?

> OB3 : Develop a predictive model able to establish relationships between productivity (i.e., crop yield), social-ecological factors (i.e., livelihood capitals, management intensity) and soil health indicators (i.e., soil carbon and water storage, soil nutrient supply) as explanatory variables.

OB4: Couple predictive model with Global Sensitivity and Uncertainty Analysis (GSUA) to test model reliability.

OB5: Use predictive model to guide the design of effective intervention strategies for boosting farm productivity.

This dissertation is composed of five chapters. The first chapter addresses the development and application of the conceptual framework. Chapter 2 addresses RQ1 and meets OB1 and OB2 of this research. In chapter 2, patterns of relationships between livelihood capitals/management intensity, and ecosystem services are assessed in the two case studies using correlation analysis towards identifying synergies and trade-offs between indicators, and using self-organizing map clustering algorithms, towards capturing intricate non-linear relationships that characterize farms as human-natural systems. It focuses on uncovering the linkages between social-ecological factors, productivity and soil health metrics. Chapter 3 addresses RQ2 and meets OB3 and OB4 of this research. In Chapter 3, predictive machine learning models are developed to test their ability to predict the productivity level of farms (using crop production as a key performance metric), based on social-ecological factors (i.e., livelihood capitals, management intensity and soil health indicators (i.e., soil carbon and water storage, soil nutrient supply) as explanatory factors. Global Sensitivity Analysis (GSA) is used to identify influential factors of the best performing model. Chapter 4 addresses RQ2 and meets OB5 of this research. In Chapter 4, the best performing model is used to inform the design of effective strategies for increasing productivity in the studied SSA smallholder systems, using decision-support tools such as Global Sensitivity and Uncertainty Analysis (GSUA) and Monte Carlo Filtering (MCF). Chapter

5 summarizes the findings of this research and concludes with limitations and future works.

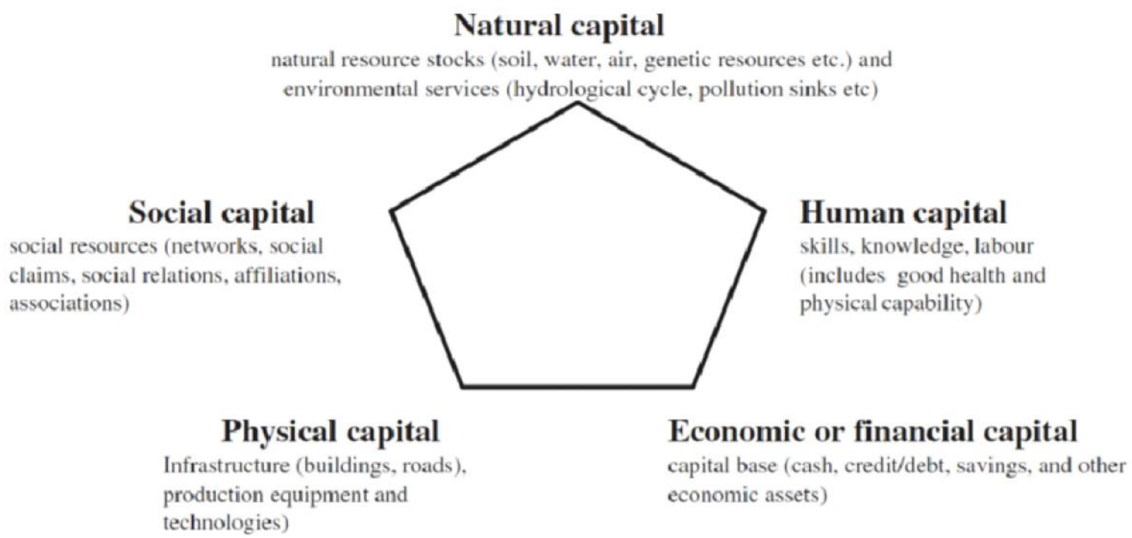

Figure 1-1. The Five capital of the Sustainable Livelihood Approach (Ashley & Carney, 1999)

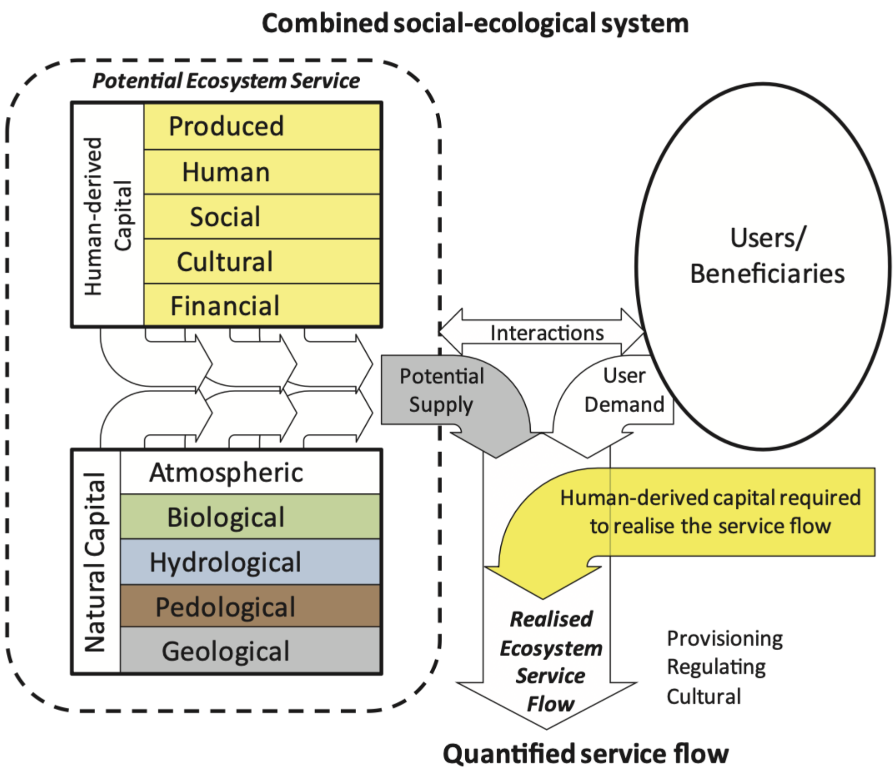

Figure 1-2. The system approach to ecosystem service social-ecological concept (Jones et al., 2016)

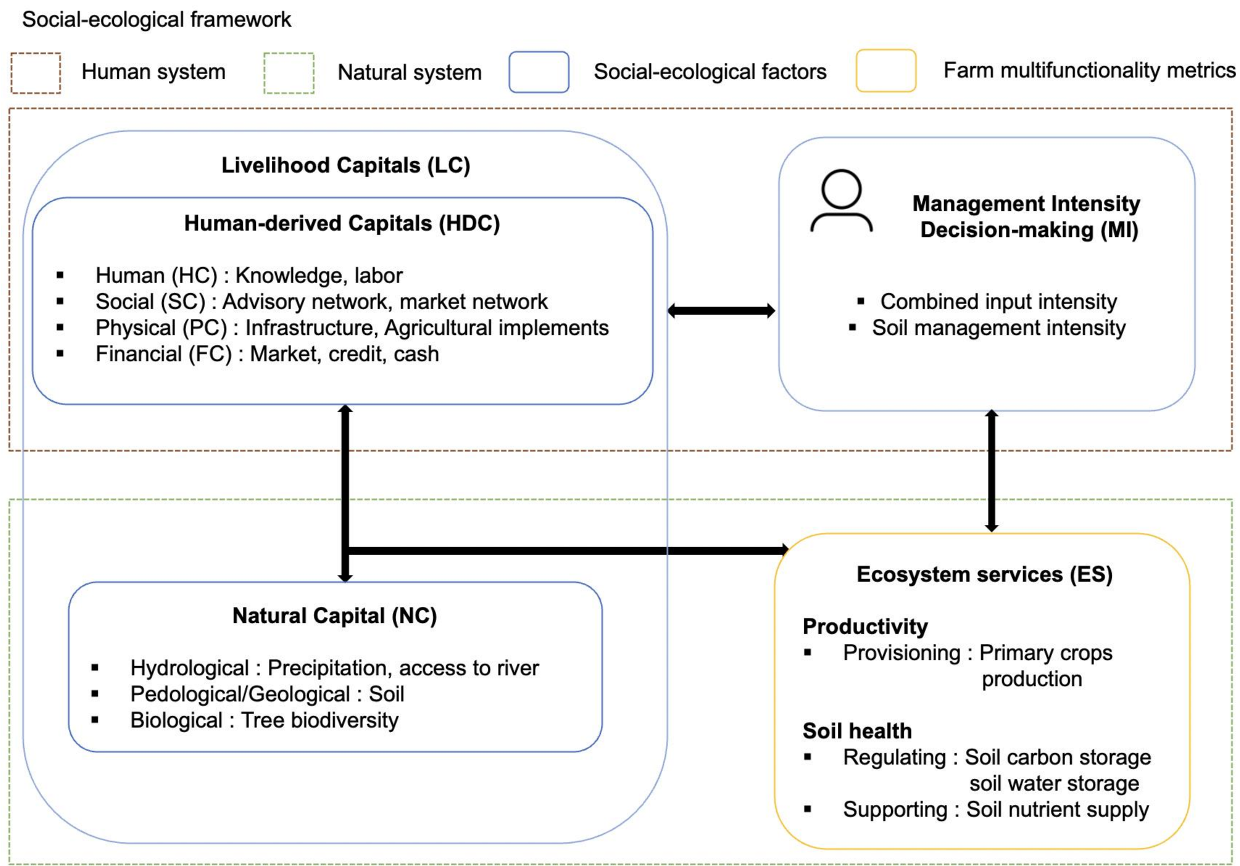

Figure 1-3. Conceptual diagram of the social-ecological framework used in this research.

CHAPTER 2
IDENTIFYING PATTERNS OF RELATIONSHIPS BETWEEN ECOSYSTEM SERVICES AND LIVELIHOODS INTENSIFICATION STRATEGIES IN SMALLHOLDER LANDSCAPES: CASE STUDIES IN SAGCOT TANZANIA AND THE UPPER EWASO NGIRO BASIN KENYA

2.1 Introduction

Sustaining natural capital and ecosystem services while addressing the increase in global food demand is a global grand challenge. Smallholder farmers across Sub-Saharan Africa and Asia (accounting for 80% of the world's farmers) play a key role in the global food system and will face significant pressure to increase food production. (Lowder et al., 2014). In order to increase food production, they will need to employ agricultural intensification strategies able to increase production per unit of land (Godfray et al., 2010; Pretty & Bharucha, 2014; Vanlauwe et al., 2014). These strategies may involve efforts that aim at prioritizing productivity over soil health, which may positively or negatively affect the supply of all ecosystem services on farms, thus limiting the multifunctional benefits of farming (Aryal et al., 2022; Rodríguez et al., 2006). Common synergies and trade-offs demonstrated in the literature are amongst provisioning and regulating services (Aryal et al., 2022). In the case of farm level ecosystem services of interest (Chapter 1) trade-offs relationships are generally observed between crop production and water storage or carbon storage(Bennett et al., 2009; Morizet-Davis et al., 2023; Power, 2010). Understanding interactions between ecosystem services across social-ecological systems continue to be a research priority. Several studies revealed trade-off and synergistic relationships between ecosystem services across a span of provisioning, regulating, and supporting services, and across multiple scales, however they show some inconsistencies as these relationships can be context and scale dependent (Qiu et al., 2018, 2021). Agricultural intensification strategies of farmers are important drivers of farm level ecosystem services (Pretty, 2008; Robertson et al., 2014). Additionally, land use intensity can play a key role in mediating ecosystem services relationships at the landscape and regional scales (Qiu et al., 2021).Understanding the relationships between farm/field level social-ecological factors (i.e., livelihood capitals, management intensity indicators) and ecosystem services, as applied in our framework (Chapter 1) is imperative (Paruelo & Sierra, 2023). Although many studies have analyzed the relationship between livelihood and ecosystem services (Agarwala et al., 2014; J. A. Fisher et al., 2014), very few have linked farming intensity to livelihoods and ecosystem services. This work aims to apply the framework proposed in this dissertation (chapter 1) to assess synergistic/trade-offs relationships between social-ecological factors (i.e., livelihood capitals, management intensity) and ecosystem services (i.e., crop yield, soil carbon storage, soil water storage, soil nutrient supply) in the two smallholder systems proposed, the SAGCOT and the Upper Ewaso Ng'iro regions.

The first objectives of this study is to assess synergistic and trade-off relationships between ecosystem services, livelihood capitals, management intensity across case studies. We hypothesized common ecosystem service relationships studied in the literature such as synergistic relationships between provisioning services and supporting services, trade-off relationships between provisioning and regulating services. We also hypothesized synergistic relationships between farm intensity and provisioning services, trade-offs relationships between farm intensity/livelihood capitals and regulating services. The second objective of this study is to identify patterns of relationships between ecosystem services and livelihood/intensification strategies characterizing each case study, to confirm synergies and trade-offs tested for in the first objective.

2.2 Materials and Methods

2.2.1 Study Regions

The Southern Agricultural Growth Corridor of Tanzania (SAGCOT) is a region selected by the Tanzania government in 2010 for boosting agricultural productivity and ensuring the commercialization of smallholder agriculture. The region has been undergoing significant agricultural transformation to expand 350,000 hectares of agricultural land into production by 2030 in six established development clusters (Scholes et al., 2013) (Figure 2-1).

The Upper Ewaso Ng'iro north catchment in Laikipia (Upper Ewaso) is the largest river basin in Kenya. It drains the northern and northwestern slopes of Mount Kenya, covering an approximated area of 220,000 km^2. The region spans an impressive ecological gradient stretching from forested and humid Mount Kenya across the Laikipia plateau to the dry Samburu plains. Economic development, increasing population, and climatic variability in the region has significantly increased competing water demands amongst various groups (crop producers, pastoralists, ranchers) over the last three decades (Figure 2-1).

2.2.2 Data Collection and Processing

Data on soils, climate biodiversity, household socio-economic condition and assets, and farming practices for the SAGCOT case study were collected by the Vital Signs Project, a large-scale data collection initiative led by Conservation International on agriculture, environment, and human well-being in the region (http://vitalsigns.org/atlas).

The collected data were processed and analyzed by the Innovation Lab for Collaborative Research on Sustainable Intensification Innovation Lab(SIIL) indicator project repository at the University of Florida (https://gitlab.com/gklarenberg/siil-environmental-indicators). The SIIL focused on developing a generic indicator framework across five sustainability domains (productivity; economic, environmental, human condition, and social) and four scales (field, farm, household, and landscape) (Musumba et al., 2021). Overall sampling took place from 2012 to 2016 in landscapes of10 x 10 km located within each of the six SAGCOT development clusters, to account for their unique land use and land cover characteristics, and livelihood and ecological diversity (Figure 1-1). Data were collected at the field and farm scale for 10 e-plots (1m by 1m scale plots that were randomly selected) in each landscape. Soils data were collected for a total of 60 e-plots within the landscapes, and 328 e-plots outside the landscapes (i.e., the semi-natural area). In addition to the environmental data collected from each landscape, three households were surveyed per e-plot, and their respective fields were sampled. A total of 210 household were surveyed and 371 cultivated fields sampled for soil testing. Additional soil properties data were estimated at 30 m resolution, such as depth to bedrock, bulk density, and cation exchange capacity were collected for the central locations of the e-plots from the Innovative Solutions for Digital Agriculture database (https://www.isda-africa.com/isdasoil) (Hengl et al., 2021). The data were accessed using the Africa soil and Agronomy data cube repository on GitLab (https://gitlab.com/openlandmap/africa-soil-and-agronomy-data-cube).The compiled dataset was carefully examined by evaluating missing data and outliers. The number of missing data for the variables varied from 1 to 33, less than 25% of the of the total number of observations. K-nearest neighbor algorithm was used to impute missing values.

The Upper Ewaso Ngiro in Laikipia, Kenya used a different sampling framework than the SAGCOT. Data on livelihood and socio-economic conditions of farmers, farming practices, environmental conditions and food security were collected through a household survey undertaken in 2020 by the Nature Conservancy (TNC Kenya) in conjunction with the Centre for Training and Integrated Research in arid and semi-arid development (CETRAD).The sampling strategy was defined to best represent the social-ecological diversity of the basin. Sampling was informed by the agro-ecological zones and main land use systems of the region, the history of settlement and crop production in the upper and middle watershed, and land use/ land cover changes in recent years. A total of 500 households were surveyed across sub-humid, humid, and semi-arid zones. The survey dataset was also examined for missing data and outliers. The number of missing data for the variables varied from 2 to 109, less than 25% of the of the total number of observations. K-nearest neighbor algorithm was used to impute missing values(Stekhoven & Bühlmann, 2012). Farm fields were not directly sampled due to time constraints. Soil properties data (i.e., depth to bedrock, clay content, bulk density, extractable aluminum, total organic carbon, extractable calcium, Iron, Potassium, Magnesium, CEC, total nitrogen, organic carbon, phosphorous, sulfur, stone, content, zinc extractable, pH, sand content, silt, content, texture, class), all estimated at 30 m resolution were collected for the 500 farm households locations from the Innovative Solutions for Digital Agriculture database (https://www.isda-africa.com/isdasoil) (Hengl et al., 2021). The data were accessed using the Africa soil and Agronomy data cube repository on GitLab (https://gitlab.com/openlandmap/africa-soil-and-agronomy-data-cube).

Summary statistics of livelihood capitals, management intensity of farm households surveyed are provided in Table 2-2 and Table 2-3.

2.2.3 Ecosystem Indicators

Ecosystem services indicators proposed in our framework (Chapter 1) were established for quantifying ecosystem services. We focus on the following across the two case studies: one provisioning service (primary crops grown by farmers in the regions), two regulating services (soil carbon storage, soil water storage), and one supporting service (soil nutrient balance/supply). Maize and rice yields were considered for the SAGCOT, and maize and potato yield for upper Ewaso Ngiro. Soil carbon storage is determined using the “less is better” standardizing function on a soil carbon deficit indicator. The soil carbon deficit indicator is defined as the ability of the soil to store additional carbon and is estimated as the difference between the maximum potential carbon that can be associated with the soil and the current amount of carbon found in the soil. The less the deficit, the higher the ability of the soil to store additional carbon (Barré et al., 2017; Sanchez et al., 2003) (See Appendix A on how the soil carbon deficit function is calculated). Soil nutrient balance in the SAGCOT is determined using the average of nitrogen, phosphorous and potassium partial nutrient budget indicators. Partial nutrient budgets are calculated as the difference between nutrient inputs through fertilizers and nutrient outputs through harvests. Indicators were calculated using the “less is better” standardized function on partial nutrient budgets because all budgets were at a deficit (i.e.,negative). This standardization approach reflects the less the nutrient deficit, the better the nutrient balance ecosystem service.

Soil nutrient supply in the Upper Ewaso Ngiro is determined using the average of total nitrogen and cation exchange capacity indicators. Both indicators were calculated using the “more is better” standardizing function. This standardization approach reflects the higher the total nitrogen and cation exchange capacity, the higher the nutrient supply ecosystem service.

For both case studies, all ecosystem services indicators were averaged at the farm level. Soil properties selected to quantify ecosystem services functions, and summary statistics are provided in Table 2-1. Additional information on the functions used to construct soil health indicators using soil properties collected are provided in Appendix A. The code for retrieving soil properties and calculating soil health indicators can be found in the dissertation repository (Appendix B).

2.2.4 Agricultural Management Intensity Indicators

Agricultural Management intensity measures are determined using two metrics, a combined farm input intensity index and a soil management intensity index. We measure farm input intensity by obtaining an index that uses the discrete input use/no input use responses for decision-making factors such as fertilizer use, improved seeds use, mechanization, and irrigation from the survey, to rank the grouping structure of their combinations. Ranking of inputs grouping structure is based on the inputs relative impact on productivity in the following order: fertilizer use > improved seeds use > mechanization > irrigation > pesticide use. The soil management intensity index is calculated using the discrete practice/no practice responses for decision-making factors such as soil tillage and residue retention practices responses, to rank the grouping structure of the combinations. Ranking of practices is based on their relative impact on soil disturbance in the following order: soil tillage > residue retention practice. The code for the management intensity indexes calculation can be found in the dissertation repositoty (See Appendix B). The indexes and summary statistics are presented in Table 2-3.

2.2.5 Livelihood Capitals Indicators

Households' livelihood capitals were quantified using evaluation scores based on a set of indicators embracing the different types of capitals (Carney, 2003; Solesbury, 2005). These evaluation scores are calculated using indicators selected from the household surveys and/or spatial data products. We selected a total of 25 indicators in the SAGCOT and 21 indicators in the upper Ewaso Ngiro . Capitals are calculated using the entropy-weight method that uses a mathematical model that applies information entropy to assign weights to criteria set in a multi-criteria decision matrix (Lotfi & Fallahnejad, 2010). This approach is objective in weighting indicators while accounting for positive (the bigger the better) or negative direction (the smaller the better). The selected indicators were first standardized to 0-1 scale using the fuzzy logic membership functions (i.e., more is better, less is better), towards eliminating measurement bias in comparing indicator levels. Entropy weights were calculated using the credit model package in R. Evaluation scores for each indicator were determined by multiplying each indicator by its respective weight. Evaluation scores of each capital including natural (NC), human (HC), social (SC), physical (PC), and financial (FC) were calculated by aggregating indicators belonging to the same capital category. Evaluation score for the human-derived capital (HDC) were calculated using the entropy weight method on human, social, and physical capital scores. Summary statistics of livelihood capital indicators are presented in Table 2-2. The code for retrieving livelihood capital indicators and calculating livelihood capital indexes can be found in the dissertation repository (see link in Appendix B).

2.2.6 Analytical Approach

First, we assess the synergistic and trade-off relationships between ecosystem services, livelihood capitals and management intensity indicators, across farm households. We used the Spearman Rank correlation measure to capture the behavior of bivariate correlations as it is more robustness to non-normality and outliers (Gómez-Baggethun et al., 2018; Lee & Lautenbach, 2016; Raudsepp-Hearne et al., 2010). A positive correlation coefficient implies a synergistic relationship between indicators, a negative correlation coefficient implies a trade-off relationship between indicators. A *p*-value<α (significance level) implies a significant relationship between a pair of indicators. In this study we used α=0.05.

Second, we used the Self-organizing map (SOM) algorithm (Kohonen, 1998) to discern clusters of ecosystem services relationships. These ecosystem services clusters (ESC) are defined as bundles of ecosystem services bundles that repeatedly appear together across space and/or time(Raudsepp-Hearne et al., 2010). SOM is an unsupervised artificial neural network learning technique that can reduce the dimensional space of high input data while preserving the topology of the data (Kohonen, 1998). The output of SOM is a vector with similar number of observations and attributes than of the input vector. This feature of SOM is chosen for our analysis as it can capture non-linear relationships between the ecosystem services. Several studies opted for this method for characterizing ecosystem services bundles (Crouzat et al., 2015; Dittrich et al., 2017; Li et al., 2022).We use the package AweSom in R to implement the SOM. Each observation is classified being assigned a "codebook vector" based on a set of best matching unit variables picked by the algorithm (Kohonen, 2013; Wehrens, 2015). Input datasets for the models included a total of m = 5 ecosystem services indicators as attributes in the SAGCOT and the upper Ewaso Ngiro region, and n = 210 farm-households in the SAGCOT and n = 500 farms in the Upper Ewaso Ngiro. We initialize the SOM grid, determining the number of codebook vectors to start with, based on the heuristic rule of m = $5\sqrt{n}$ by setting up a 9 × 8 nodes typology grid for the SAGCOT and an 8 x 12 typology grip for the UENB (Vesanto et al., 1999). Both models were then optimized so that all best matching units could fit the number of codebook vectors provided for the sample. The initial number of nodes were reduced in both models. The SAGCOT SOM model was run for 6 X 6 nodes with a quantization error of 0.0133 and topographic error of 0.067, and the Upper Ewaso SOM model was run for 3 X 1 nodes with a quantization error of 0.0399 and topology error of 0. The Upper Ewaso model resulted in codebook vectors representing the final clusters obtained. The SAGCOT model resulted in several codebook vectors that needed to be further classified under super classes using an optimized number of clusters. We used the K-medoid classification method to obtain contiguous clusters of the vectors based on their similarity (Kassambara, 2017; Varmuza, 1980). The number of clusters was chosen based on the elbow method: the number of clusters that corresponds to the lowest percentage of unexplained variance of the code vectors (Pulkkinen & Nurmi, 2012).

We also use the Self-organizing map (SOM) algorithm to discern clusters of livelihood strategies based on patterns of relationship livelihood capitals and management intensity indicators. Input datasets for the models included a total of m = 4 livelihood capitals and management intensity indicators as attributes in both case studies, and $n = 210$ farm-households in the SAGCOT and $n = 500$ farms in the Upper Ewaso Ngiro. Similar initialization processes implementing the ESC SOM were undertaken. The SAGCOT SOM model was run for 4 X 4 nodes with a quantization error of 0.038 and topographic error of 0.157, and the Upper Ewaso SOM model was run for 4 X 4 nodes with a quantization error of 0.028 and topographic error of 0.044. Both models resulted in several codebook vectors that needed to be further classified under super classes using an optimized number of clusters. We also used the k-Medoids classification method based on the elbow method to obtain contiguous clusters of the vectors. We proceed to characterizing farms based on ecosystem services, livelihood capitals and management intensity indicators contributions to the observed clusters. We then selected the non-parametric tests of Kruskal Wallis one-way-analysis of variance and post-hoc Dunn's test to determine if there is a statistically significant difference between one or more ecosystem services clusters and between one or more livelihood/intensification clusters based on indicators, and to identify exactly which clusters are different in each case. A p-value$<\alpha$ (significance level) implies a significant relationship between one or more clusters. In this study we used $\alpha=0.05$.

Third, we overlapped farms belonging to each livelihood strategy clusters (LSC) and each ecosystem services clusters (ESC) concurrently (Dittrich et al., 2017), to further assess which livelihood/intensification strategy characterizes the given ecosystem services relationships observed. All codes on the application of the analytical framework for each can be found in the dissertation repository (See corresponding link in Appendix B). The analytical framework is illustrated in Figure 2-2.

2.3 Results

2.3.1 Ecosystem Services, Management Intensity, Livelihood Capitals Relationships across Case Studies

In this study, we hypothesized trade-offs relationships for the following pairs of indicator types or indicators: (1) provisioning vs. regulating services, (2) management intensity vs. regulating services, (3) human-derived capitals vs. regulating services. We also hypothesized synergistic relationships for the following pairs of indicators: (1) provisioning vs. supporting services, (2) management intensity vs. provisioning/supporting services, (3) natural capital vs. ecosystem services, (4) human-derived capital and supporting/provisioning ecosystem services.

In the case of provisioning services, while trade-off relationships were not observed between crop production and water storage, they were observed between rice/ maize,/potato production and carbon storage in both case studies (Table 2-4). Significant synergies were observed between crop production and nutrient supply in both case studies as well.

In the case of management intensity indicators (i.e.,combined input intensity and soil management intensity), both trade-offs and synergies were observed between combined input intensity and crop production across studies (Table 2-4). Expected synergies were only confirmed between combined input intensity and rice production, combined input intensity and potato production in the SAGCOT and the Upper Ewaso Ng'iro respectively. Expected trade-offs were confirmed between combined input intensity and carbon storage across both case studies, and between soil management intensity and carbon storage in SAGCOT, and between soil management intensity and water storage in the Upper Ewaso Ng'iro. Expected synergies were confirmed between both management intensity indicators (i.e., combined input intensity, soil management intensity) and nutrient supply (Table 2-4).

In the case of livelihood capitals, expected trade-offs relationships between natural capital and crop production were observed in both case studies, except between natural capital and maize production in the Upper Ewaso Ngiro (Table 2-4). Synergistic relationship between human-derived capital and crop production was only observed in the SAGCOT for rice production. Expected synergies between natural capital and supporting service such as nutrient supply are consistent and significant across both case studies.

Our analysis revealed that although relationships between pairs of ecosystem services varied across the two case studies, especially considering varying crop production relationships, consistent expected relationships were still detected. These relationships are mainly consistent synergies between crop production and nutrient supply, combined input intensity and crop production, consistent trade-offs between natural capital and crop production, and consistent trade-off between human-derived capital and carbon storage.

2.3.2 Ecosystem Service Clusters

2.3.2.1 SAGCOT

Four ecosystem services clusters with varying contribution of services were detected in the SAGCOT region (Figure 2-3), all with significant differences between clusters for one more ecosystem service indicator (Figure 2-4). The distribution of the clusters across the Vital Signs landscapes are summarized in Table 2-4. The resulting clusters had the following characteristics:

ESC1– *High rice/low maize provisioning – low water storage and high nutrient supply farms* (n=25 {12%}).

The average rice yield of these farms is around 6069 kg/ha. These farms are dominant in the Mbarali landscapes (Figure 2-1a).

ESC2 – *Low maize/mid rice provisioning – high carbon storage farms* (n=94 {45%}).

The average maize yield and rice yield of these farms is around 285 kg/ha and 348 kg/ha, respectively. These farms are dominant in the Rufiji and Ihemi Kilolo landscapes (Figure 2-1a).

ESC3 – *High maize/low rice provisioning – high carbon storage farms* (n=50 {24%}).

The average maize yield in these farms is 2076 kg/ha. The average rice yield of these farms is around 266 kg/ha. This type of farm is dominant in the Ludewa landscape (Figure 2-1a).

ESC4 – *Mid maize/low rice provisioning –low carbon storage farms* (n=41 {20%}).

The average maize and rice yields of these farms are around 1135 kg/ha and 314 kg/ha respectively. These farms are dominant in the Kilombero and Sunbawanga landscapes (Figure 2-1a).

2.3.2.2 Upper Ewaso Ng'iro

Three ecosystem services clusters (Figure 2-5) with varying contribution of each ecosystem service were detected with significant differences between one or more group (Kruskal-Wallis test, $p< 0.05$) for potato production and soil carbon storage (Figure 2-6). The distribution of the ecosystem services bundles across the agro-ecological zones of the region are summarized in Table 2-6. The resulting clusters had the following characteristics:

ESC1 – *High potato/high maize provisioning, high carbon storage farms* (n=238 {48%}). The average maize and potato yields in this cluster are 345 kg/ha and 374 kg/ha respectively. This cluster is dominant in the sub-humid to semi-humid landscapes of the region.

ESC2 – *Mid potato/ mid maize provisioning, mid carbon storage farms* (n=115 {23%}). Average maize and potato yield in this cluster are 304 kg/ha and 277 kg/ha respectively. This cluster is dominant in the Sub-humid to semi-humid landscapes of the region.

ESC3 – *Low potato/mid maize provisioning, low carbon storage farms* (n=147 {29%}). Average maize and potato yield in this cluster are 269 kg/ha and 236 kg/ha respectively. This cluster is dominant in the semi-humid landscapes of the region.

2.3.3 Livelihood/Intensification Strategy Clusters

2.3.3.1 SAGCOT

Three (LSC1, LSC2, LSC3) livelihood/intensification strategy clusters with varying contribution of natural, human-derived capitals, combined input intensity and soil management intensity were detected (Figure 2- 7), all with significant differences between clusters for one or more indicators (Figure 2-8).

LSC1 – *Mid natural capital, mid human-derived capital, low intensity* n=84 {40%}). Farms in this cluster have levels of natural capital that are significantly lower than the other clusters, and levels of combined input intensity that are significantly higher than one cluster LSC1.

LSC2 – *Low natural capital, low human-derived capital, mid intensity*. (n=50 {24%}). Natural capital is significantly higher than in LSC1and significantly lower than in LSC3.

LSC3 – *High natural capital, high human-derived capital, high intensity* (n=76{36%}). Natural capital is significantly lower in this cluster than LSC2.

2.3.3.2 Upper Ewaso Ng'iro

Three livelihood strategy clusters with varying contribution of natural, human derived capitals, combined input intensity and soil management intensity were detected (Figure 2-9), with significant differences between clusters for combined input intensity, soil management intensity and natural capital indicators (Figure 2-10).

LSC1 – *Mid natural capital, high human-derived capital, high-intensity* (n=236 {51%}). Natural capital in this cluster is significantly higher than in LSC2 and LSC3. Combined input intensity is significantly lower compared to LSC3.

LSC2 – *Low natural capital, mid human-derived capital, mid-intensity* (n=132{26%}). Natural capital in this cluster is significantly low compared to LSC3. Combined input intensity is significantly lower than in LSC3 and significantly higher than LSC1. Soil management intensity is significantly higher in this cluster compared to the other two clusters.

LSC3 - *High natural capital, mid human-derived capital, low-intensity* (n=112 {22%})- Natural capital is significantly high in this cluster compared to the other two clusters. Combined input intensity is significantly lower than LSC2.

2.3.4 Ecosystem Service and Livelihood/Intensification Clusters Overlap

2.3.4.1 SAGCOT

Overlaps of ecosystem service and livelihood/intensification strategy clusters are presented in Table 2-7. The majority of the farmers (40%) undergo LCS1 (*Mid natural capital, mid human-derived capital, low intensity*). These farms mainly overlap with ESC2 (L*ow maize/mid rice provisioning – high carbon storage farms*) and ESC4 (*Mid* *maize/low rice provisioning –low carbon storage farms*). About 37% of farmers undergo LSC3 (*High natural capital, high human-derived capital, high intensity*). The majority of these farms also overlap with ESC2 (*low maize/mid rice provisioning – high carbon storage farms*). About 24% of farmers undergo LSC2 (*Low natural capital, low human-derived capital, mid intensity*). The majority of these farms overlap with ESC1 (*High rice/low maize provisioning – low water storage and high nutrient supply farms)* and ESC2 (*low maize/mid rice provisioning – high carbon storage farms).*

These results provide additional insights on the linkages between intensification strategies and ecosystem service levels. Here, low-intensity farms are associated with lower level of human-derived capital and tend to be the least provisioning.

2.3.4.2 Upper Ewaso Ng'iro

Overlaps of ecosystem service and livelihood/intensification strategy clusters are presented in Table 2-8. The majority of the farmers (51%) undergo LCS1 (*Mid natural capital, high human-derived capital, high-intensity)*. The majority of these farms overlap with ESC1 (*High potato/high maize provisioning, high carbon storage farms*). About 26% of farmers undergo LSC2 (*Low natural capital, mid human-derived capital, mid-intensity*). The majority of these farms also overlap with ESC1 (*High potato/high maize provisioning, high carbon storage farms*). About 22% of farmers undergo LSC3 (*High natural capital, low human-derived capital, low-intensity*). The majority of these farms also overlap with ESC1 and ESC3 (*Low potato/mid maize provisioning, low carbon storage farms*). Here, low-intensity farms are also associated with lower human-derived capital and least provisioning farms.

2.4 Discussion

The results of the synergies and trade-offs analysis of ecosystem services, livelihood capitals, and management intensity indicators confirmed many of the hypothesized relationships between indicators across the two case studies. First, we were able to confirm consistent trade-offs relationships between provisioning (i.e., crop production) and regulating services, mainly carbon storage, and consistent synergistic relationships between provisioning (i.e., crop production) and supporting services, mainly nutrient supply, in both studies. This is in line with many studies in the literature on ecosystem services relationship (Aryal et al., 2022; Lee & Lautenbach, 2016).

Second, the synergistic relationships observed between combined input intensity and crop production for rice and potato, combined input intensity and nutrient supply, support the hypothesis that farm intensification contribute to increasing provisioning and supporting services. Moreover, consistent synergistic relationships between combined input intensity and human-derived capital across case studies demonstrate the important role of livelihood capitals in agricultural intensification (Pretty, 2008).Consistent trade-offs between natural capital and crop production for rice, maize in the SAGCOT and for potatoes in the Upper Ewaso Ng'iro underlines the negative effects of intensification on natural resources (Garcia Alberto, 2020; Gopel et al., 2020; Kremen, 2020).

Results from characterizing farms based on livelihood capitals and management intensity showed consistent patterns of relationships between livelihood capital levels and management intensity levels across case studies. High intensity clusters were characterized by high human-derived capital, while low intensity clusters were

characterized by low human-derived capital. This captures the synergistic relationships observed between combined input intensity and human-derived capital across the case studies. This was the case for LSC3 in the SAGCOT region and LSC1 in the Upper Ewaso Ng'iro region. In overlapping farms with a certain livelihood/intensification strategy with their corresponding ecosystem service clusters, we observed that there is minimal overlap between low-intensity farms and high provisioning farms in both case studies (Table 2-7 and Table 2-8). However, there are considerable overlap between high intensity farms and mid to high provisioning farms (Table 2-7 and Table 2-8). In both case studies, these high provisioning farms are characterized by high nutrient supply in the SAGCOT and high carbon storage in the Upper Ewaso Ng'iro.
Overall, these results were able to support several relationship hypotheses established in this study, mainly confirming that human-derived capital are important drivers of farming intensity and crop production.

2.5 Conclusions

In this chapter, we tested for synergistic and trade-offs relationships between social-ecological factors (i.e., livelihood capitals, management intensity) and ecosystem services (i.e., crop production, soil carbon storage, soil water storage, soil nutrient balance/supply) and defined clusters of ecosystem services relationship and livelihood/intensification strategy, towards gaining better understanding the linkages between agricultural intensification strategies, management intensity, and ecosystem services at the farm level. The synergy/trade-off analysis between all indicators revealed some consistent hypothesized relationships between indicators, mainly supporting the overarching hypothesis that there are synergistic relationships between combined input intensity and provisioning/supporting services, and trade-off relationships between human-derived capital and regulating services.

We detected 4 ecosystem service clusters and 3 livelihood/intensification strategy clusters in the SAGCOT, and 3 ecosystem service clusters and 3 livelihood/intensification strategy clusters in the upper Ewaso Ng'iro. The overlapping of the clusters provided evidence of trade-off linkages between farming intensity and natural capital, and the synergistic linkages between human-derived capital and farming intensity.

This study has several limitations that can lead to future work. First, our indicators for livelihood capitals and ecosystem services measures vary across case studies. Having consistent indicators across case studies would have allowed for a stronger comparative analysis. Additionally, the approach used in selecting ecosystem services indicators included multiple primary crops separately as provisioning services. Focusing on one main crop could have allowed for a more generalized approach to analyzing synergies and trade-offs related to crop production across case studies. Additionally, our approach to measuring management intensity through the combined input index construction is limited in providing important quantitative information on inputs. Although this approach is able to capture faming intensity well in low-input smallholder systems, it may not be applicable for high-input systems.

Future attention on the relationship between farm intensity and ecosystem services over time may be promising to further understanding the effects of agricultural intensification on ecosystem services and natural resources.

1a. SAGCOT, Tanzania

1b. Upper Ewaso Ng'iro basin, Kenya

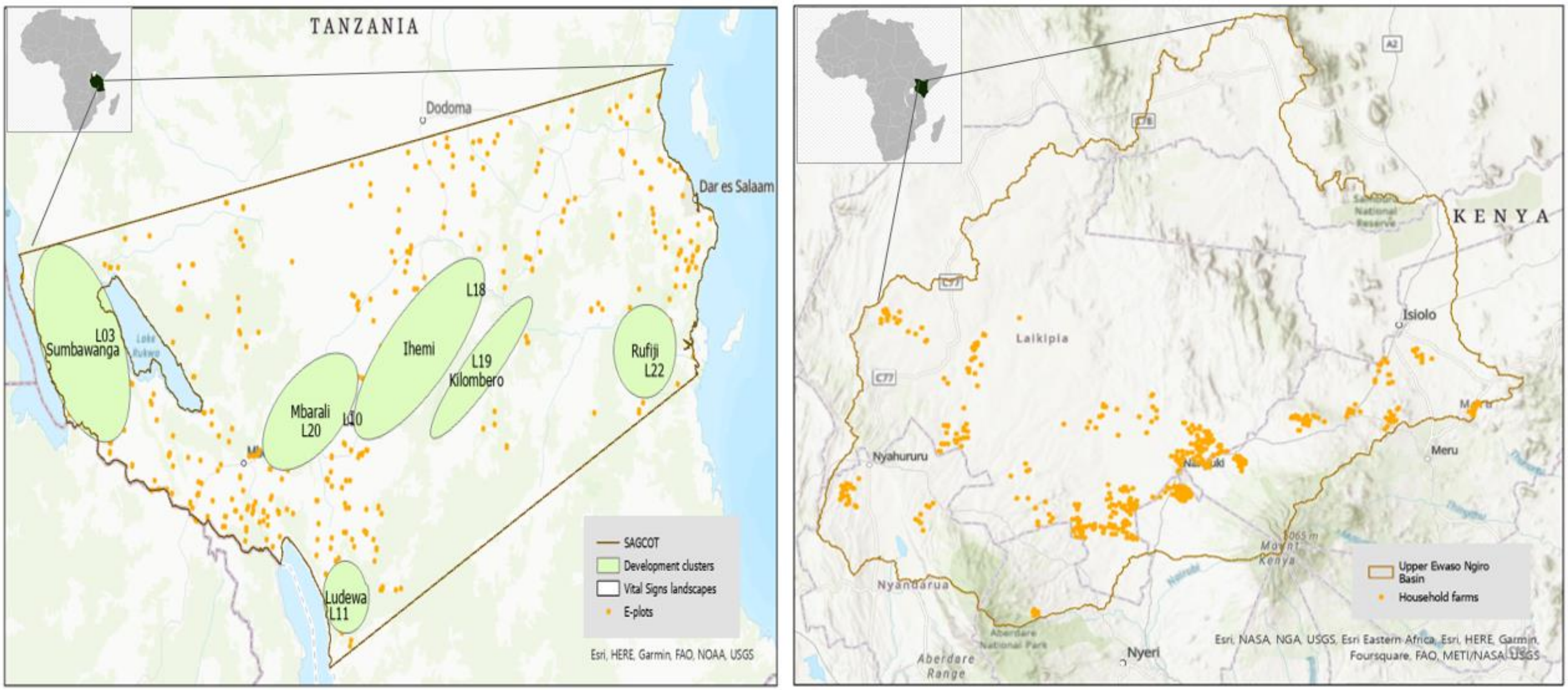

Figure 2-1. Study regions. Yellow dots denote fields sampled.

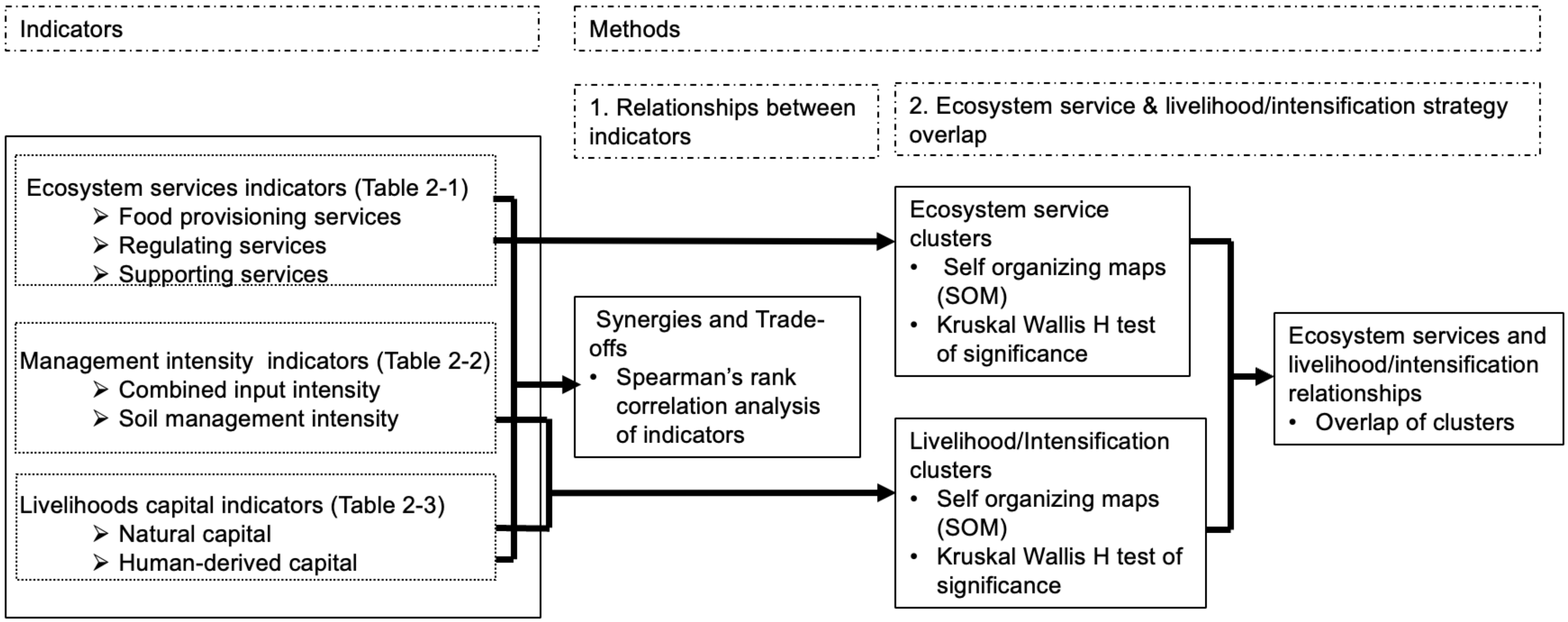

Figure 2-2. Analytical framework.

Table 2-1. Ecosystem services indicators and summary statistics in the two case studies.

Ecosystem services	SAGCOT			Upper Ewaso Ngiro		
	Indicators	**Mean**	**Std. Dev.**	**Indicators**	**Mean**	**Std. Dev.**
Food Provisioning						
Crop production (CP)	[a] Maize yield (kg/ha)	1116.7	849.5	[a] Maize yield (kg/ha)	416.6	591.1
	Rice yield (kg/ha)	2802.8	2387.7	Potato yield (kg/ha)	311.1	828.1
Regulating						
Soil carbon storage (CD)	[b] Acidified carbon (g/kg)	1.14	0.5	[c] Total Inorganic carbon (g/kg)	3.02	0.72
	[b] Clay content (%)	41.3	14.6	[c] Clay content (%)	34.8	2.7
	[b] Silt content (%)	18.8	3.9	[c] Silt content (%)	2.8	1.1
	[b] pH	6.2	0.3	[c] pH	2.1	0.9
Soil water storage (WS)	[b] Clay content (%)	41.3	4.6	[c] Clay content (%)	34.8	2.7
	c Bulk density (kg/m3)	132.2	10.1	[c] Bulk density (kg/m3)	121	5.1
	[c] Depth to bedrock (cm)	19	3.3	[c] Depth to bedrock (cm)	21.2	3.5
Supporting						
Soil nutrient balance/ supply (NS)	[b] Partial N budget (kg/ha)	-3.9	29	[c] Total N (g/kg)	94.6	12.7
	[b] Partial P budget (kg/ha)	-1.81	11.91	[c] Total P (g/kg)	29.06	1.77
	[b] Partial K budget (kg/ha)	-6.06	6.52	[c] Total K (g/kg)	58.72	2.96

[a] Calculated using survey data

[b] Measured directly through field sampling

[c] Measured using 30 m resolution isDA map

Table 2-2. Livelihood capital indicators and summary statistics in the two case studies.

Capital	Asset	SAGCOT, Tanzania				UENB, Kenya			
		Indicator	Measure/Value assignment	Mean	Std. Dev.	Indicator	Measure/Value assignment	Mean	Std. Dev
Natural	cWater	River access	Distance to water bodies (km)	9.82	5.63	Land use border stream or river	No = 1; Yes = 2	1.2	0.42
		Precipitation	Average precipitation from 1981-2016 (mm)	962.43	226.23	Precipitation (mm)	Average precipitation from 1981-2016 (mm)	898.7	216.16
	[c]Trees	Tree species richness	Tree species richness index	7.26	6.78	Tree cover	Percent tree cover	12.16	6.4
		Tree species diversity	Shannon diversity index	1.21	0.86	Tree species diversity	Number of tree species on farm	4.2	2.8
		Forest proximity	Distance to forest (km)	18.82	13.49				
	[b]Soil	Soil essential mineral	% clay content	40.9	18.1	Soil essential mineral	% clay content	34.8	2.7
Human	[c]Human labor	Household dependency ratio	[# of dependents (<15 &>64 years) /# working age people (15-64 years)]	0.51	0.16	Household size	Number of household members	4.07	2.09
		Agricultural labor	Hours of labor per capita for weeding, fertilizing, harvesting	29.4	23.86	Household head age	Age of household head	56.26	14.33
		Gender ratio	Male to Female ratio	1.39	1.08	Household head gender	Female = 1 ; Male = 2	1.77	0.42
	[c] Knowledge	Household head education	Primary levels = 1 ; Secondary levels =2 ; University levels = 3	0.82	0.53	Household head education	Never attended =1; Primary school = 2 ; High school = 3; Tertiary = 4 ; Undergraduate degree = 5 ; Graduate degree = 6	1.94	0.93
		Household literacy rate	Number of educated members/Total household members	0.79	0.2	Agricultural trainings	No = 1; Yes = 2	1.62	0.49
Social	[a] Information	Access to agricultural information	None = 1 ; Media only = 2; People = 3 ; Media & people = 4 ; Media & institutions = 5 ; People & institutions = 6 ; Media, people & institutions = 7	3.27	1.03	Access to agricultural information	None = 1 ; Community groups/Farmers exchange only =2 ; Community & farmers groups = 3	1.5	0.63
	[a] Market	Access to market	No =1 ; Yes = 2	0.16	0.36	Market link support	No = 1; Yes = 2	1.05	0.21

Table 2-2. Continued

Capital	Asset	SAGCOT, Tanzania				UENB, Kenya			
		Indicator	Measure/Value assignment	Mean	Std. Dev.	Indicator	Measure/Value assignment	Mean	Std. Dev
Physical	[c] Equipments	Agricultural implements	Number of imeplements including hand hoe,hand powered sprayer,ox plough,ox seed planter,ox cart, cutlass,machete.	9.07	5.95				
		Farm building	None = 1 ; Silo / shed /storeroom only = 2 ; Silo & shed or silo & storeroom or shed & storeroom = 3 ; Silo, shed and storeroom = 4	1.43	0.61	Agricultural implements (Tractor-pulled tiller,Planter,Sprayer,Harvester)	None owned or leased = 1 ; one or more leased = 2 ; one or more owned = 3 ; one or more owned =4 and leased	2.39	0.64
		Irrigation infrastructe	None = 1 ; Non-mechanized only = 2 ; mechanized only = 3 ; Mechanized and non-mechanized tools = 4	3.09	0.49				
	[c] Communication	Communication tools	None = 1; Radio = 2 ; Mobile = 3 ; Radio and mobile = 4	3.26	1.04	Communication tools	None = 1 ; Radio/television = 2 ; Cell phone and radio/television = 3 ; Internet and cell phone/radio/television = 4	2.1	0.66
		Wireless access	distance to nearest cellular or wi-fi tower						
	[c] Transportation	Transportation means	None = 1 ; Bicycle only = 2 ; Motor cycle only = 3 ; Bicycle & motor cycle = 4 ; Motor vehicle = 5 ; Motor vehicle & bicycle = 6 ; Motor vehicle & motor cycle = 7 ; Motor vehicle, motor cycle and bicycle = 8	1.9	0.79	Transportation means	None =1 ; Walking/foot = 2 ; Bicycle = 3 ; Motorbike = 4; Public motor vehicle = 5	2.53	1.16
		Proximity to main road	Distance to main roads (km)	5.73	4.49				
		Proximity to road	Distance to any road (km)	1.48	1.96	State of road to market	Footpath = 1 ; Seasonal raod =2 ; All weather road not tarmacked = 3 ; All weather road tarmacked = 4	2.89	0.77
		Field to road proximity	Time travel to field (minutes)	13.4					

Table 2-2. Continued

Capital	Asset	SAGCOT, Tanzania Indicator	Measure/Value assignment	Mean	Std. Dev.	UENB, Kenya Indicator	Measure/Value assignment	Mean	Std. Dev
Financial	Housing	House sanitation conditions	No Toilet = 1 ; Unimproved Pit Latrine = 2 ; Improved Pit Latrine = 3 ;Pour Flush = 4	2.04	0.29	House sanitation conditions	None = 1 ; Open pit = 2 ; Enclosed pit = 3 ;Enclosed pour-flush = 4, Enclosed flush = 5	2.93	0.39
		cooking fuel	Firewood = 1 ; Gas Charcoal = 2 ; Paraffin/Kerosene = 3	1.02	0.18	Cooking source	Natural material (wood/sawdust) = 1 ; Charcoal = 2 ; Liquified gas = 3 ; Biogas = 4 ; Liquified gas = 5 ; Electricity from solar cells = 8	1.21	0.72
		Lighting fuel	Firewood = 1 ; Candle = 2 ; Torch/ Lamp Oil = 3 ; Solar = 4 ; Electricity = 5 ; Private Generator = 6	3.17	0.58	Lighting source	None = 1 ; Candle, paraffin wax = 2 ; Liquid fuel [kerosene] = 3 ; LED lamp = 4 ; Unstable voltage electricity from grid = 5 ; Stable voltage electricity from grid = 6 ;Electricity from solar cells/batter-powered source/wind turbine/ small dam = 7 ; Electricity from a generator = 8	6.06	1.38
		Land holding	Small (< 5) = 1 ; Medium (≥ 5 & > 10) = 2 ; Large (>10) = 3 (ha)	1.5	0.69	Land holding	Small (< 5) = 1 ; Medium (≥ 5 & > 10) = 2 ; Large (>10) = 3	1.22	0.52
		House tenure	Rented = 1 ; Employer Provided - Free = 2; Free = 3 ; Owner Occupied = 4	3.85	0.6	Land tenure	Rented for less than 12 months = 1; Leasehold more than years = 2 ; Leasehold less than 5 years = 2 ; Communal = 4 ; Freehold = 5	4.77	0.82
	[c] Finances	Microfinance acces	No = 1; Yes = 2	1.09	0.29	Financial loans access	No = 1 ; Probably not = 2; Probably yes = 3 ; Definitely yes = 4	2.31	1.11
		Wage entry	Number of wage entry sources in a year	0.61	1.06	Debt status	Yes a lot = 1 ; Yes,a moderate amount = 2 ; Yes, a little = 3 ; No = 4	3.26	0.9
		Livestock sales	Number of livestock sold in a year	2.02	3.31	Crop sale point	No sales points = 1 ; Sales point farm gate/others = 2 ; Sales point local business = 3 ; Sales point at local village market = 4 ; Sales point at regional market = 5 ; Sales point outgrower company = 6	2.05	0.96

Table 2-3. Determination of the evaluation scores for input intensity and soil management intensity indexes.

Intensity measures	Evaluation score	SAGCOT Variable assignment (No = 0 ; Yes = 1)	SAGCOT Percentage	Upper Ewaso Ngiro Variable assignment (No = 0 ; Yes = 1)	Upper Ewaso Ngiro Percentage
Combined Input Intensity	1	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	49%	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	6%
	2	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	4%	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	0.80%
	3	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 0	2%	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 0	0.20%
	4	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 1 ; Irrigation = 0, Pesticide use = 0	5%	Fertilizer use = 0 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 1	0.20%
	5	Fertilizer use = 0 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	10%	Fertilizer use = 0 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	2.20%
	6	Fertilizer use = 0 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	4%	Fertilizer use = 0 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	1.60%
	7	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	11%	Fertilizer use = 0 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 1	1%
	8	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	0.50%	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	7%
	9	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 1	0.50%	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	7.00%

Table 2-3. Continued.

		SAGCOT		Upper Ewaso Ngiro	
	Evaluation score	Variable assignment (No = 0 ; Yes = 1)	Percentage	Variable assignment (No = 0 ; Yes = 1)	Percentage
Combined Input Intensity	10	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 1 ; Irrigation = 0, Pesticide use = 0	1%	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 0	1.00%
	11	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 1 ; Irrigation = 0, Pesticide use = 1	9%	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 1	2%
	12	Fertilizer use = 1 ; Improved seeds = 1; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	1%	Fertilizer use = 1 ; Improved seeds = 0 ; Mechanization = 1 ; Irrigation = 0, Pesticide use = 0	0%
	13	Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	1%	Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 0	15%
	14	Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 0	2%	Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 0, Pesticide use = 1	29%
	15	Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 1	0.50%	Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 0	4%
	16	NA		Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 0 ; Irrigation = 1, Pesticide use = 1	19%
	17	NA		Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 1 ; Irrigation = 0, Pesticide use = 1	0.80%
	18	NA		Fertilizer use = 1 ; Improved seeds = 1 ; Mechanization = 1 ; Irrigation = 0, Pesticide use = 1	0.80%

Table 2-3. Continued.

		SAGCOT		Upper Ewaso Ngiro	
Intensity measures	**Evaluation score**	**Variable assignment (No = 0 ; Yes = 1)**	**Percentage**	**Variable assignment (No = 0 ; Yes = 1)**	**Percentage**
Soil Management Intensity	1	Soil tilling = 0 ; No residue retention = 0	30%	Soil tilling = 0 ; No residue retention = 0	15%
	2	Soil tilling = 0 ; No residue retention = 1	12%	Soil tilling = 0 ; No residue retention = 1	48%
	3	Soil tilling = 1 ; No residue retention = 1	51%	Soil tilling = 1 ; No residue retention = 0	9%
	4	Soil tilling = 1 ; No residue retention = 0	6%	Soil tilling = 1 ; No residue retention = 1	27%

Table 2-4. Spearman rank correlations for combination of pairs of indicators of interest across the two case studies. Level of significance: *** p < 0.001, ** p <0.01,* p<0.05.

Indicator pairs	SAGCOT (n=210)	Upper Ewaso (n=500)
Provisioning vs. Regulating services		
Maize production vs. Carbon storage	—0.143*	0.06
Maize production vs. Water storage	0.313***	0.046
Rice production vs. Carbon storage	0.153**	—
Rice production vs. Water storage	0.173*	—
Potato production vs. Carbon storage	—	0.071
Potato production vs.Water storage	—	0.247***
Provisioning vs. Supporting services		
Maize production vs. nutrient supply	0.141*	0.04
Rice production vs. nutrient supply	0.597***	—
Potato production vs. nutrient supply	—	0.155***
Management intensity vs. Provisioning services		
Combined input intensity vs. Maize production	-0.133	0.034
Combined input intensity vs. Rice production	—	—
Combined input intensity vs. Potato production	—	0.141***
Soil management intensity vs. Maize production	—0.085	—0.05
Soil management intensity vs. Rice production	0.146*	—
Soil management intensity vs. Potato production	—	0.058
Management intensity vs. Regulating services		
Combined input intensity vs. Carbon storage	0.182**	0.071
Combined input intensity vs. Water storage	0.061	0.257***
Soil management intensity vs. Carbon storage	0.145*	—0.078
Soil management intensity vs. Water storage	0.038	0.026

Table 2-4. Continued.

Indicator pairs	SAGCOT (n=210)	Upper Ewaso (n=500)
Combined input intensity vs. Nutrient supply	0.177*	0.181***
Soil management intensity vs. Nutrient supply	0.037	0.035
Livelihood capitals vs. Management intensity		
Natural capital vs. Combined input intensity	0.118	0.235***
Human-derivec capital vs. Combined input intensity	0.164*	0.011
Livelihood capitals vs. Provisioning services		
Natural capital vs. Maize production	—0.352***	0.002
Natural capital vs. Rice production	—0.234***	—
Natural capital vs. Potato production	—	—0.011
Human-derived capital vs. Maize production	-0.016	
Human-derived capital vs. Rice production	0.007	—
Human-derived capital vs. Potato production	—	—0.101*
Livelihood capitals vs. Regulating services		
Natural capital vs. Carbon storage	0.128	—0.078
Natural capital vs. Water storage	—0.561	0.08
Human-derived capital vs. Carbon storage	0.069	0.026
Human-derived capital vs.Water storage	0.029	—0.036
Livelihood capitals vs. Supporting services		
Natural capital vs. Nutrient supply	—0.387***	0.199***
Human-derived capital vs. Nutrient supply	—0.025	0.063

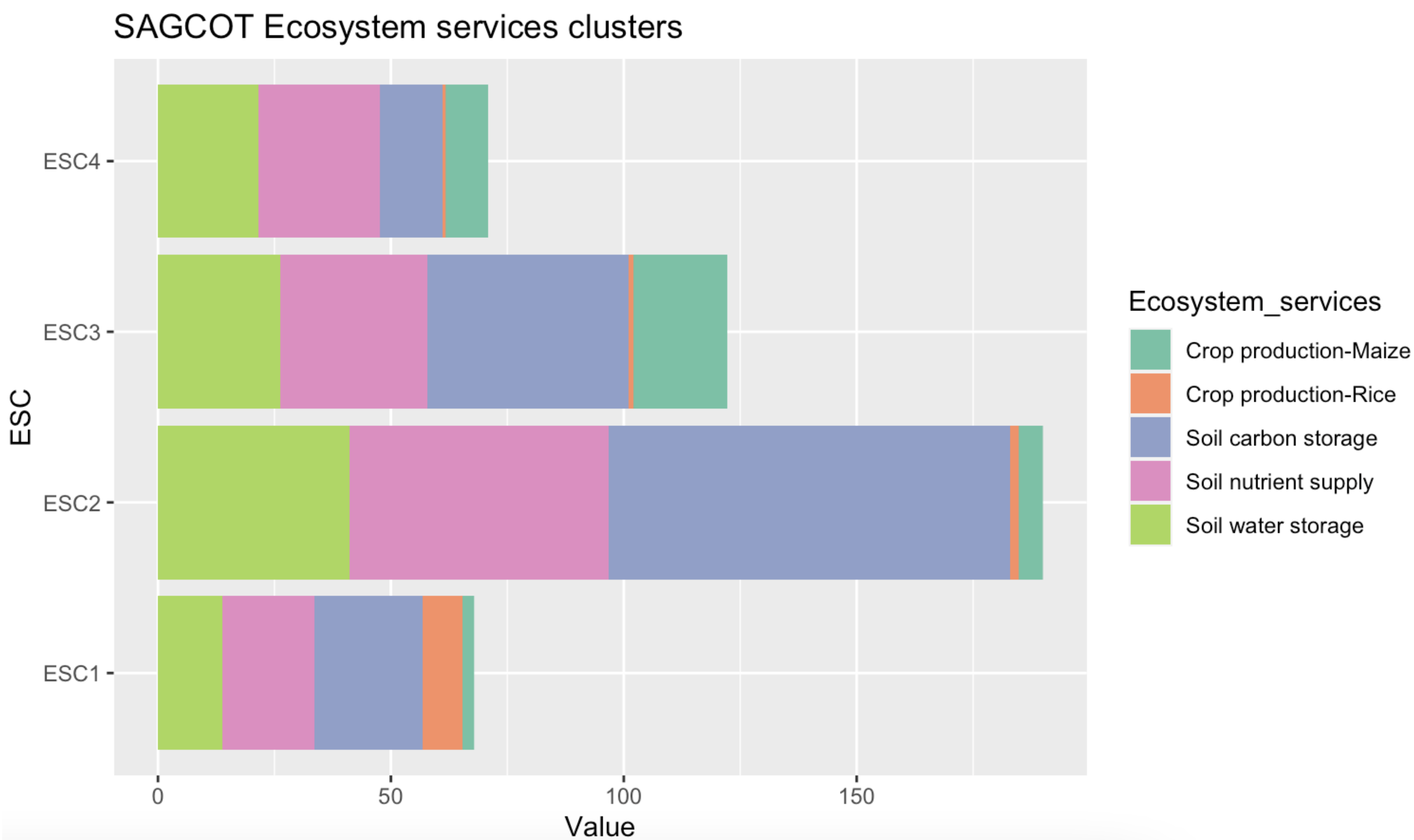

Figure 2-3. Bar plots of the codebook vectors of the SOM showing the relative contribution of ecosystem services to each cluster in the SAGCOT.

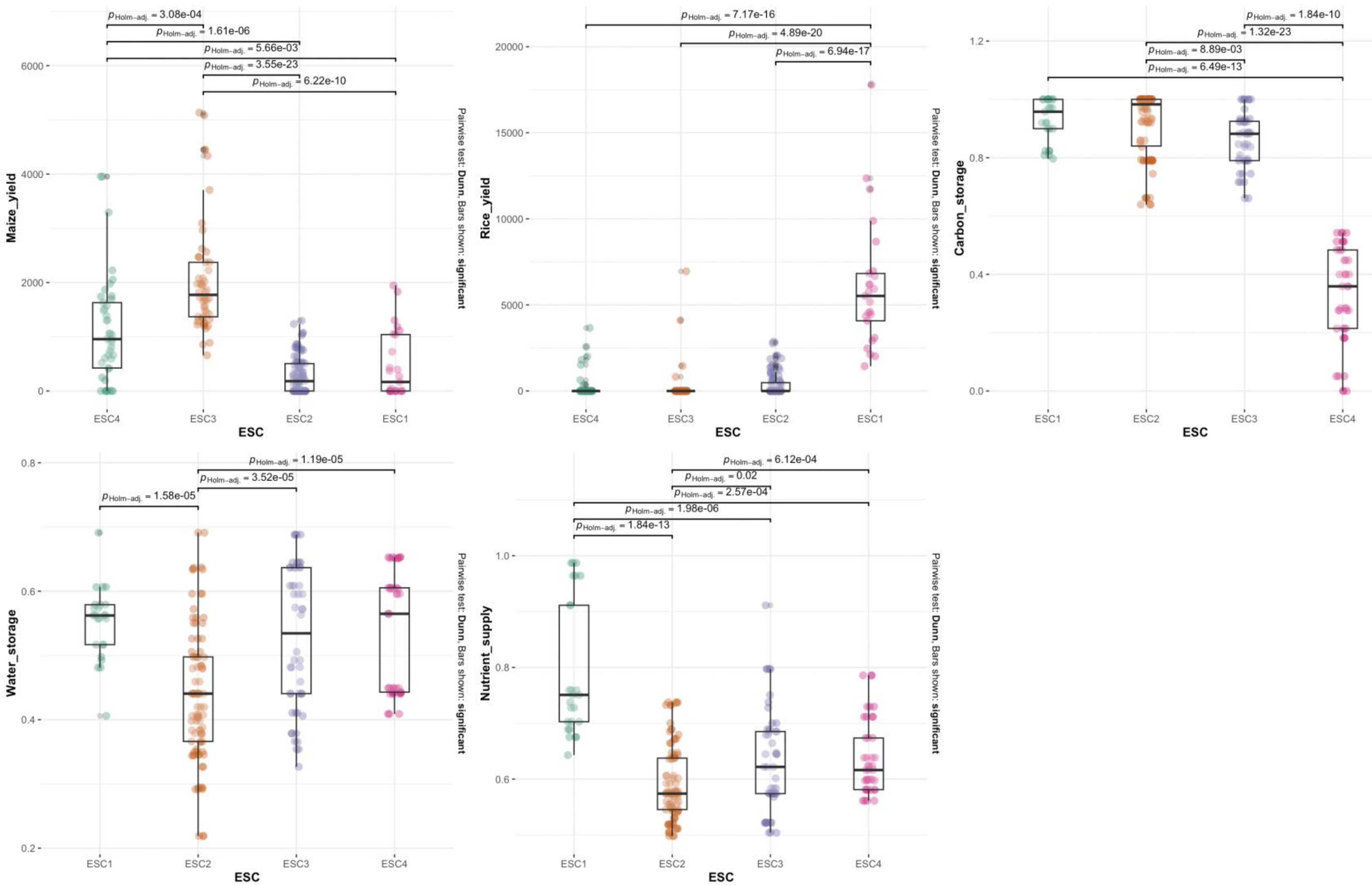

Figure 2-4. Boxplots comparing significant ecosystem services levels in the three clusters derived from the SOM using the original dataset of indicators, in the SAGCOT. Brackets indicate significant differences between two clusters.

Table 2-5. Ecosystem clusters' distribution across the SAGCOT landscapes.

Landscapes	ESC1 [Mid maize/mid rice provisioning – low carbon storage farms]	ESC2 [High maize/high rice provisioning – high nutrient supply and water storage farms]	ESC3 [High maize/low rice provisioning – low soil water storage farms]	ESC4 [Mid rice/low maize provisioning – low nutrient supply, high carbon farms]
L03 -Sunbawanga	0	2	6	22
L10 - Ihemi (Mufundi)	0	27	3	0
L11 -Ludewa	0	3	23	4
L18 - Ihemi (Kilolo)	0	18	9	3
L19 - Kilombero	0	17	1	12
L20 -Mbarali	23	2	5	0
L22 -Rufiji	2	25	3	0

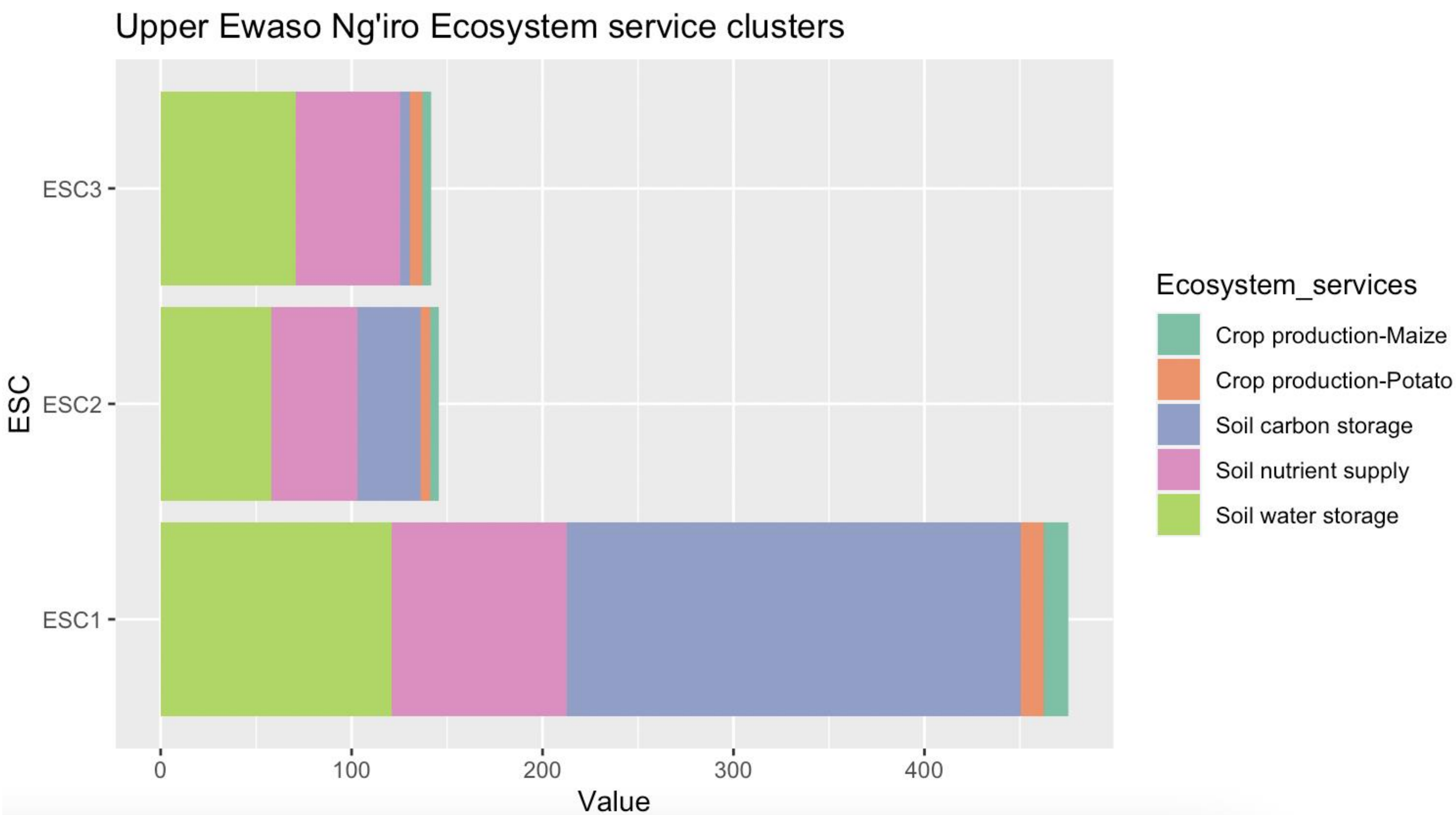

Figure 2-5. Bar plots of the codebook vectors of the SOM showing the relative contribution of ecosystem services to each cluster in the Upper Ewaso Ng’iro.

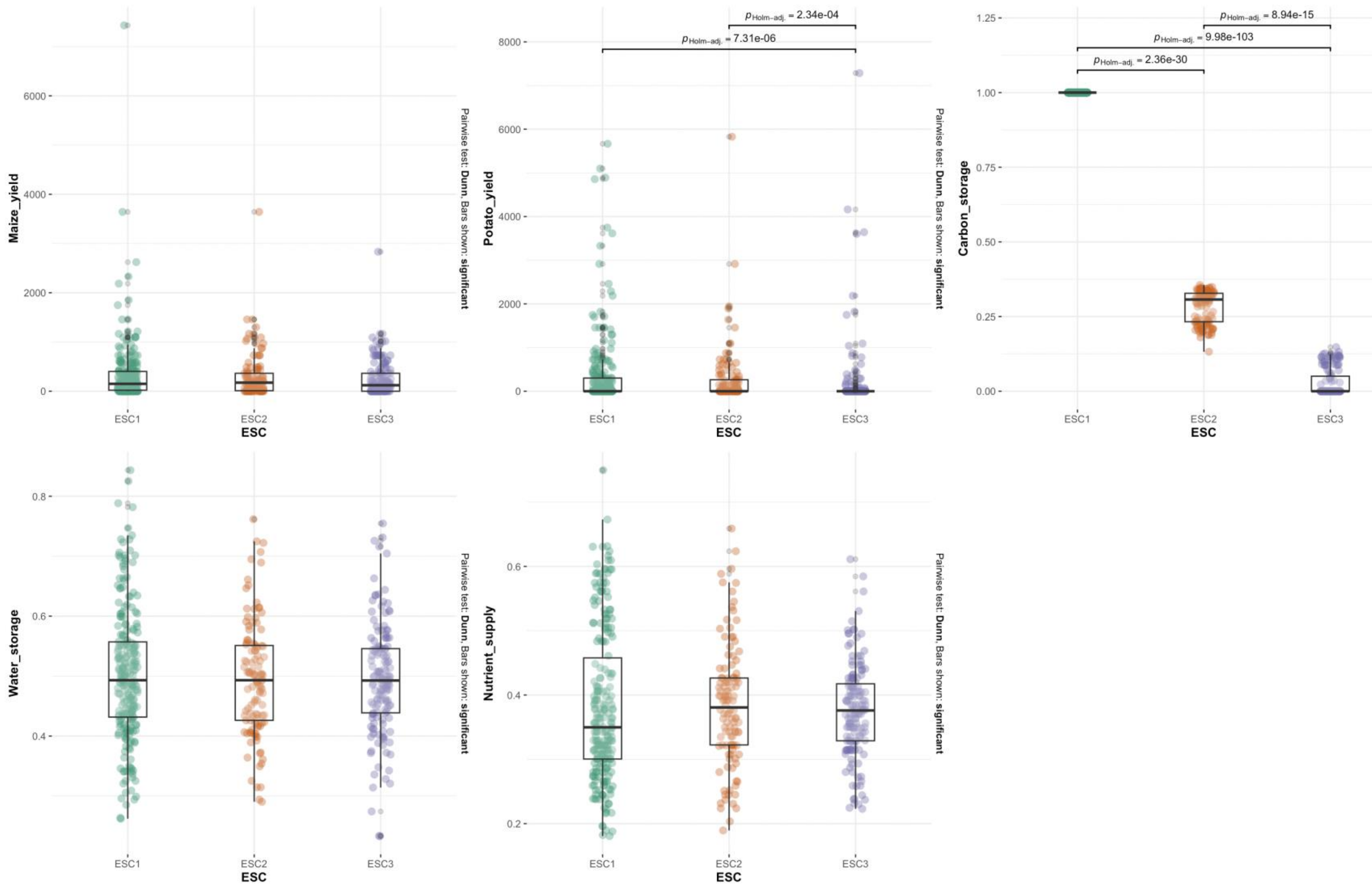

Figure 2-6. Boxplots comparing significant ecosystem services levels in the three clusters derived from the SOM using the original dataset of indicators, in the Upper Ewas Ng'iro. Brackets indicate significant differences between two clusters.

Table 2-6. Ecosystem services clusters distribution across the upper Ewaso Ngiro agro-ecological zones.

Agro-ecological zone	ESC1 [High potato/high maize provisioning, high carbon storage farms]	ESC2 [Mid potato/ mid maize provisioning, mid carbon storage farms]	ESC3 [Low potato/mid maize provisioning, low carbon storage farms]
Sub-humid	52	19	19
Sub-humid to semi-humid	101	57	52
Semi-humid	80	39	74

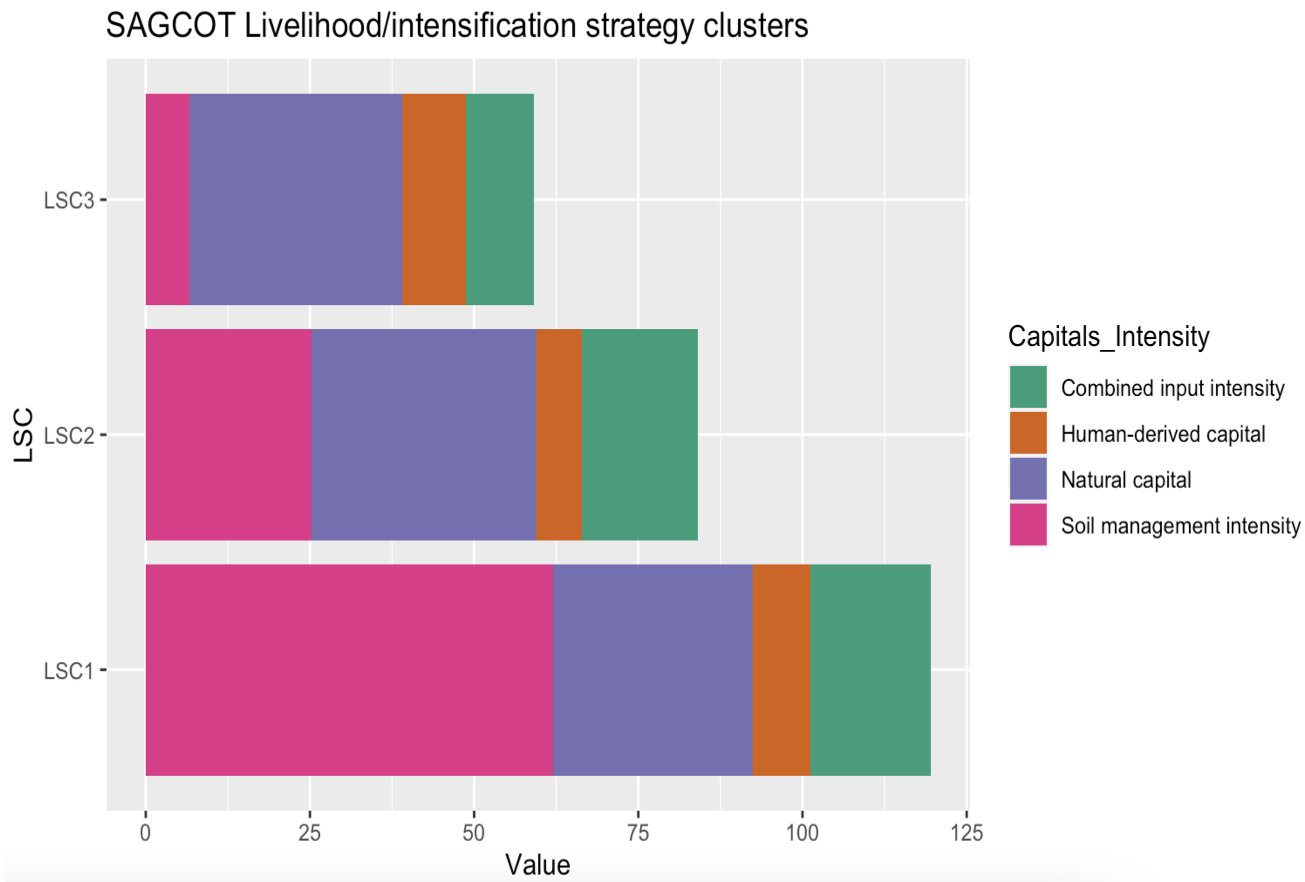

Figure 2-7. Bar plots of the codebook vectors of the SOM showing the relative contribution of livelihood capital and management intensity indicators to each detected livelihood/intensification cluster in the SAGCOT.

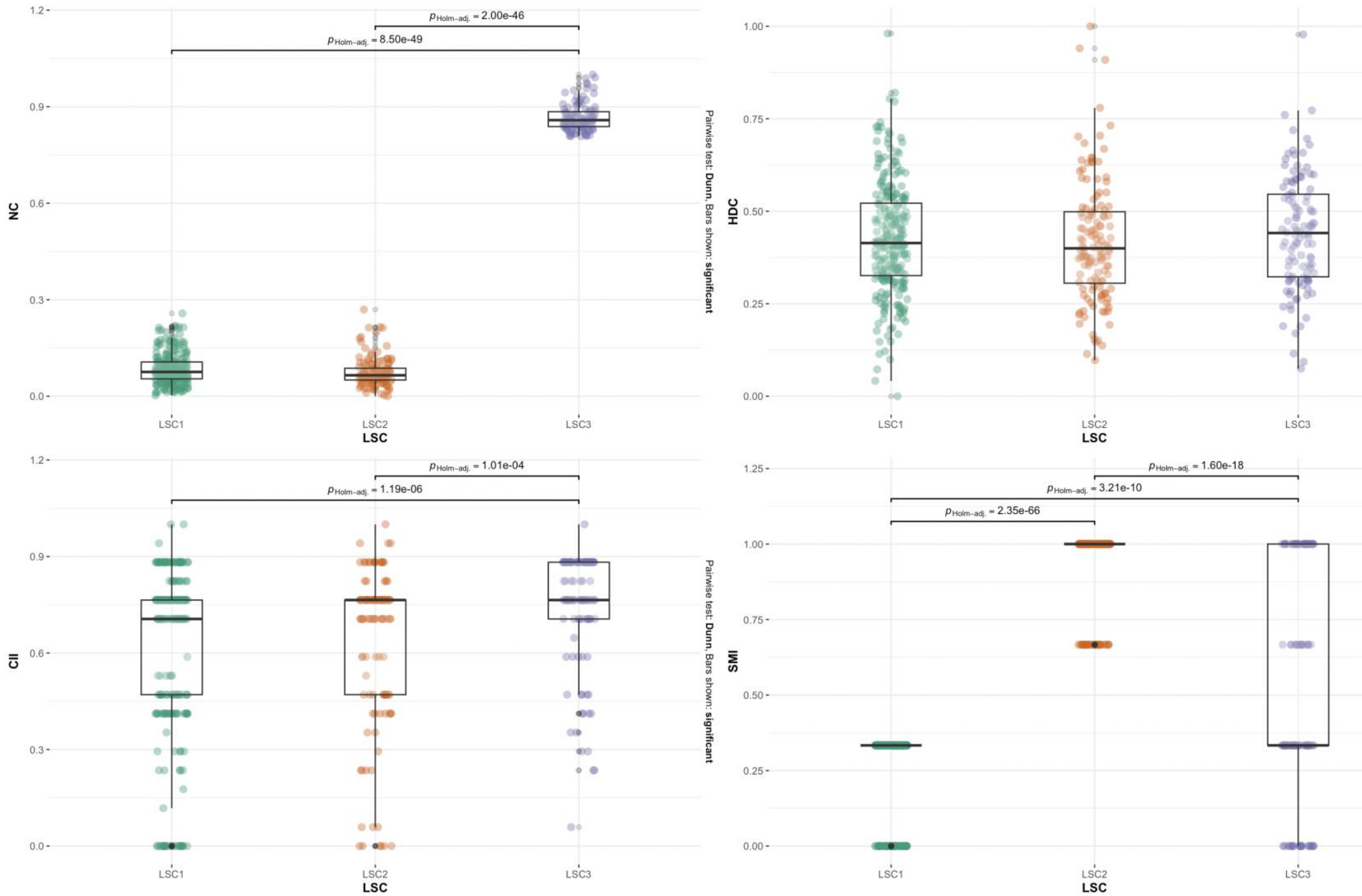

Figure 2-8. Boxplots comparing significant differences in livelihood/intensification strategy clusters for each indicator (i.e., natural capital (NC), human-derived capital (HDC), combined input intensity (CII), soil management intensity (SMI) in the three clusters derived from the SOM using the original dataset of indicators, in the SAGCOT. Brackets indicate significant differences between two clusters.

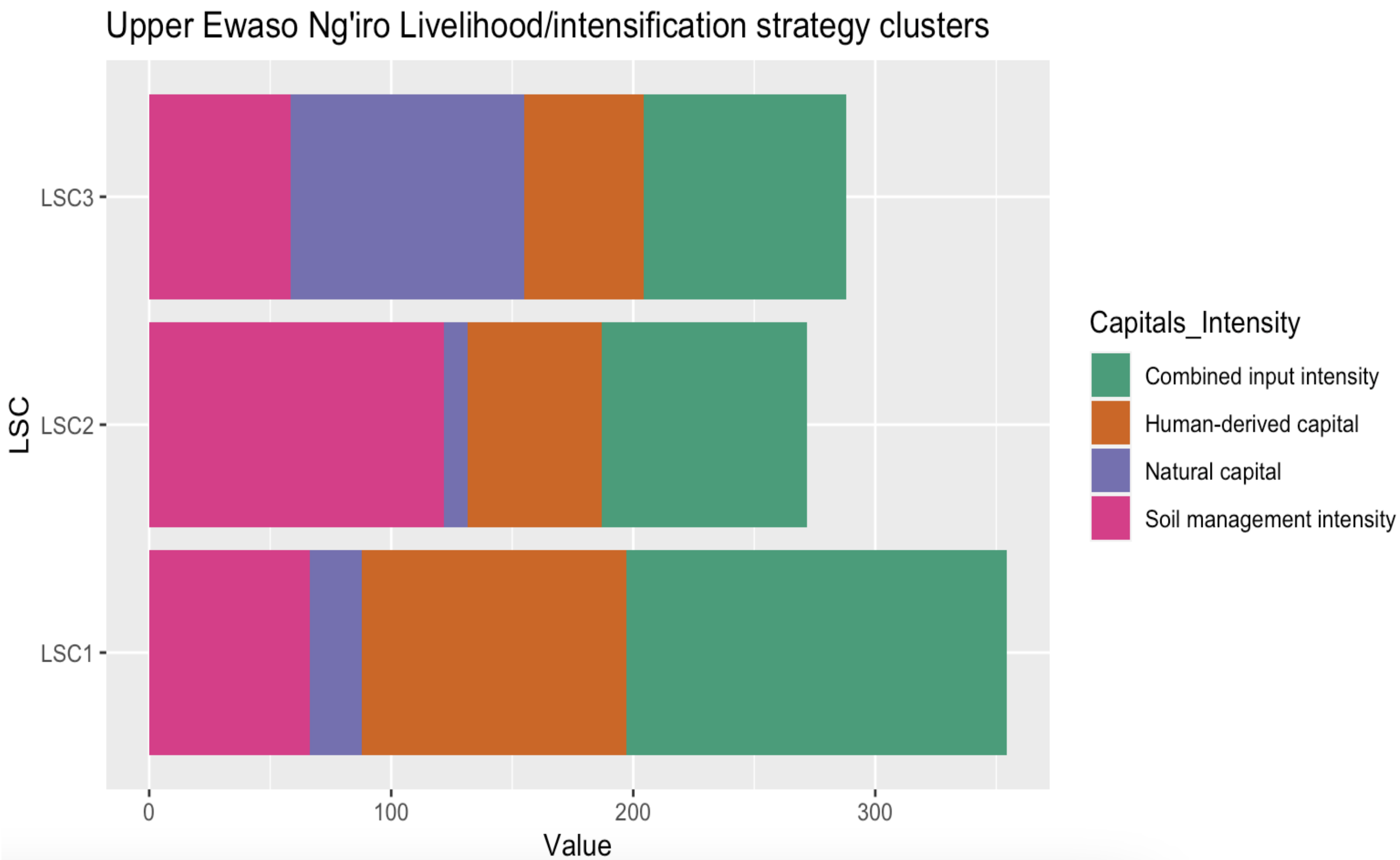

Figure 2-9. Bar plots of the codebook vectors of the SOM showing the relative contribution of livelihood capital and management intensity indicators to each detected livelihood/intensification cluster in the SAGCOT and the Upper Ewaso Ng'iro.

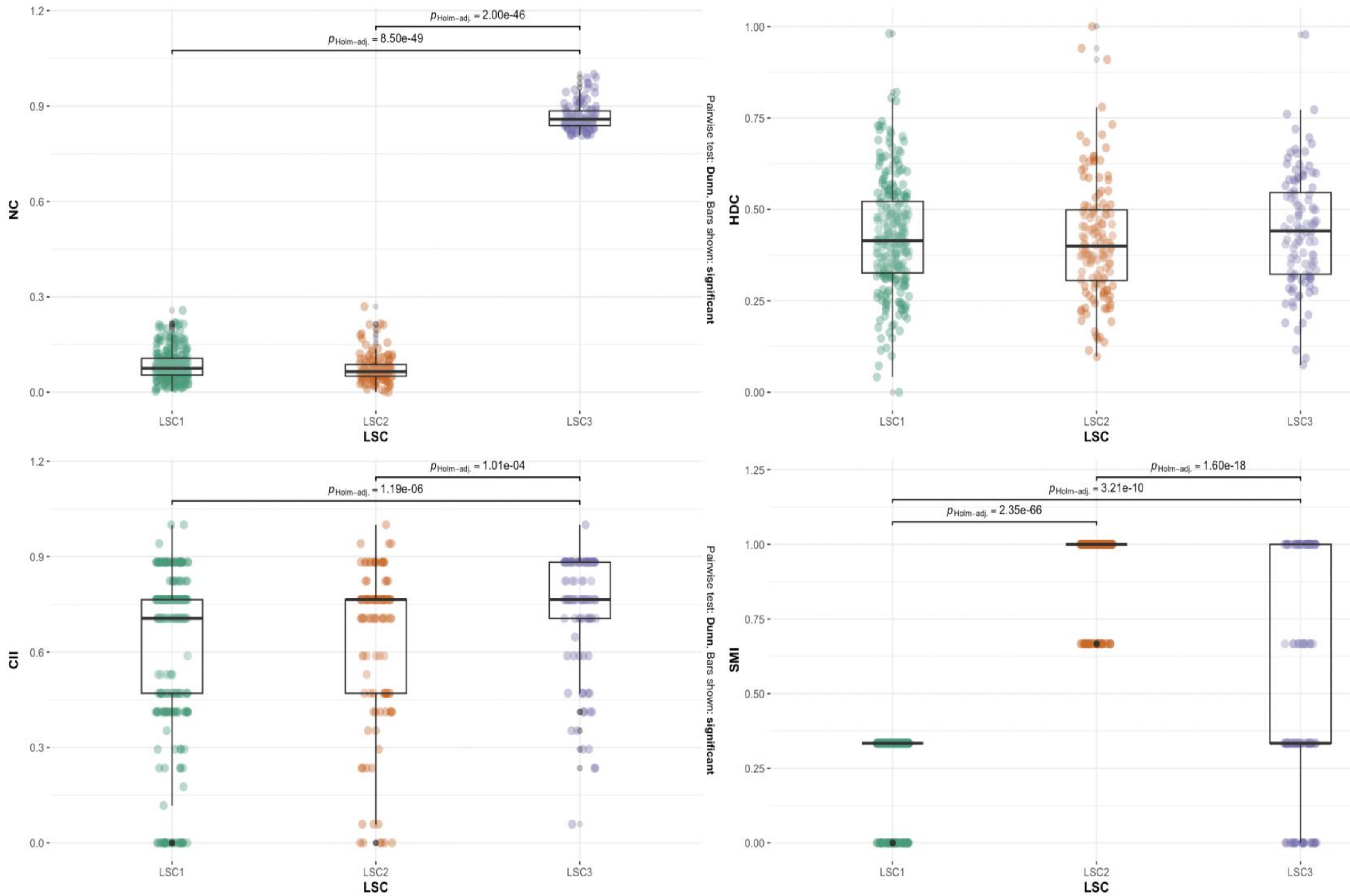

Figure 2-10. Boxplots comparing significant differences in livelihood/intensification strategy clusters for each indicator (i.e., natural capital (NC), human-derived capital (HDC), combined input intensity (CII), soil management intensity(SMI)) in the three clusters derived from the SOM using the original dataset of indicators, in the Upper Ewaso Ng'iro. Brackets indicate significant differences between two clusters.

Table 2-7. Livelihood/intensification strategies (LSC) and ecosystem services clusters (ESC) overlap in the SAGCOT.

LSC/ESC	**LSC1** [Mid natural capital, mid human-derived capital, low intensity]	**LSC2** [Low natural capital, low human-derived capital, mid intensity]	**LSC3** [High natural capital, high human-derived capital, high intensity]	**Total**
ESC1 [High rice/low maize provisioning – low water storage and high nutrient supply farms]	2%	9%	1%	12%
ESC2 [low maize/mid rice provisioning – high carbon storage farms]	16%	3%	26%	45%
ESC3 [High maize/low rice provisioning – high carbon storage farms]	9%	9%	6%	24%
ESC4 [Mid maize/low rice provisioning –low carbon storage farms]	13%	3%	4%	20%
Total	40%	24%	37%	101%

Table 2-8. Livelihood/intensification strategies (LSC) and ecosystem services clusters (ESC) overlap in the Upper Ewaso Ngiro.

LC/ESC	**LSC1** [Mid natural capital, high human-derived capital, high-intensity]	**LSC2** [Low natural capital, mid human-derived capital, mid-intensity]	**LSC3** [High natural capital, mid human-derived capital, low-intensity]	**Total**
ESC1 [High potato/high maize provisioning, high carbon storage farms]	26%	12%	9%	47%
ESC2 [Mid potato/ mid maize provisioning, mid carbon storage farms]	12%	6%	5%	23%
ESC3 [Low potato/mid maize provisioning, low carbon storage farms]	13%	8%	8%	29%
Total	51%	26%	22%	100%

CHAPTER 3

INTEGRATED MODELING OF SMALLHOLDER LIVELIHOOD CAPITALS, MANAGEMENT INTENSITY, AND SOIL ECOSYSTEM SERVICES INDICATORS TO PREDICT FARM PRODUCTIVITY: CASE STUDY OF SAGCOT TANZANIA

3.1 Introduction

Increasing global food and crop demand up to 110 % by 2050 (Alexandratos & Bruinsma, 2012; Tilman et al., 2011) will place significant pressure on smallholder farmers, who contribute to 80 % of food production across Asia and Sub-Saharan Africa (SSA) and make up for a significant portion of the world's poor (Lowder et al., 2014, 2016). Staple crops such as maize, wheat, rice, sorghum, and millet, will have to increase by 80% in many countries aiming for food self-sufficiency in the region (Jayne et al., 2018; Van Ittersum et al., 2016). However, cereal yields in SSA have stagnated around 1.5 t/ha compared to global yields since 2000 (Jayne & Sanchez, 2021). Yield gap analyses for the region pointed some key social-ecological factors of the framework developed in this dissertation (Chapter 1) explaining smallholders' productivity gap. These factors include lack of inputs, education, physical infrastructure, labor, financial markets, access to land, poor farm management practices, poor soil fertility, and climate variability (Dzanku et al., 2015; Giller et al., 2006; Hillocks, 2014; Jayne et al., 2010; Raimi et al., 2017; Tittonell & Giller, 2013). Smallholders' efforts to intensify their land will need to encompass overcoming the intertwined social-ecological challenges that contribute to hindering agricultural productivity. Understanding the role of social-ecological factors in smallholder productivity is imperative for identifying sustainable agricultural intensification strategies that can best improve food security and livelihoods in the region (Rasmussen et al., 2018; Zimmerer et al., 2015). While process-based crop models can effectively measure yield variability across varying biophysical conditions and management practices (Jones et al., 2003; Vanuytrecht et al., 2014), and have been widely applied in the region (Folberth et al., 2013; Sileshi et al., 2010; Webber et al., 2014), they are often limited in capturing micro-scale human factors (i.e., landscape and farm household livelihood and decision-making factors) of farming systems that affect yield. Although farming systems of SSA are widely studied towards understanding and characterizing smallholder heterogeneity (Alvarez et al., 2018; Chikowo et al., 2014; Kihoro et al., 2021; Tittonell et al., 2010), these studies are rarely centered around yield variability. Household and agricultural survey data used to characterize smallholder heterogeneity could potentially have as much relevant explanatory power as biophysical factors over yield variability in smallholder farms, if techniques for capturing inherent non-linear dynamics in social -ecological systems are applied to these datasets (Biggs et al., 2021; Levin et al., 2013). In spite of the important role of social-ecological factors in smallholder productivity, very few studies have considered modeling yield productivity with a combination of social-ecological and biophysical indicators (Banerjee et al., 2014; Dutta et al., 2020; Liang, 2023). Modeling such relationships is possible with new machine learning (ML) algorithms, as they are powerful prediction techniques that are efficient at identifying non-linear patterns in large unstructured datasets containing explanatory variables of various sources at various scales, and of various type (i.e., categorical, continuous, discrete) for prediction (Dangeti, 2017; James et al., 2017). Moreover, ML model reliability can be effectively tested with tools like Global Sensitivity Analysis (GSA) and uncertainty analysis aim at quantifying the importance of model inputs and their interactions with respect to model output. The computational cost required for applying and testing can be reduced remarkably using these approaches (Saltelli et al., 2008). Coupled GSA and ML techniques represent efficient dimension-reduction (factor importance) techniques also able to increase the explainability of the “black-box” characteristic of ML approaches (Muñoz-Carpena et al., 2023)

This study aims to contribute to the literature on the drivers of yield variability in smallholder systems investigating the ability of livelihood capital, management intensity and soil health indicators, to predict the productivity levels of smallholders in the Southern Agricultural Growth Corridor of Tanzania (SAGCOT) Tanzania, using ML algorithms and GSA variance-based high-dimensional factor importance evaluation. The objectives of this study are to (1) develop a predictive ML model able to uncover linkages between livelihood capitals, management intensity, soil health, and maize yield ; and (2) Couple the ML model with GSA to identify key drivers of productivity (Maize yield) in smallholder systems in the SAGCOT.

3.2 Materials and Methods

3.2.1 Study region

This work is conducted for agricultural landscapes located within six development clusters/landscapes of the SAGCOT, a region (Fig. 3-1) selected by the Tanzanian government under a public-private partnership to boost agricultural productivity and ensuring the commercialization of smallholder agriculture(Mendelsohn et al., 2014). The corridor covers about 300,000 square kilometers and the program reaches about 450,000 smallholder farming households around the development clusters where agricultural activities and investments are concentrated (Fig. 3-1). Each of these clusters/landscapes are characterized by distinct agro-ecological conditions (Mendelsohn et al., 2014; Reuben et al., 2017). Maize and rice are the dominant primary crops grown by smallholders in the region. Main characteristics and average primary crop yields of each landscape are presented in Table 3-1.

3.2.2 Methodological framework

The methods used in this chapter followed the workflow described in Figure 3-1. It encompasses 4 main steps described in the sections below: 1) data compilation 2) data preparation and curation 3) Data filtering and variable selection through multicollinearity analysis 4) machine learning model development 5) Feature importance assessment.

3.2.3 Data Collection and Feature Selection

Data on household socio-economic conditions, farm management, and soil properties were collected for 210 farming households in the agricultural landscapes of the SAGCOT region, by the Vital Signs project (Mendelsohn et al., 2014), a large-scale data collection initiative led by Conservation International on agriculture, environment, and human well-being in the region. Soils data were analyzed by the Sustainable Intensification Innovation Lab (SIIL), which applied an indicator framework to calculate soil health indicators (i.e., soil carbon deficit indicator, soil fertility indicator, soil Nitrogen, Phosphorous, and Potassium partial nutrient budgets) of farms. Additional soils data (i.e., bulk density, porosity) estimated at 30 m resolution were gathered from the Innovative Solutions for Digital Agriculture database, isDA (https://www.isda-africa.com/isdasoil) to calculate soil water storage and soil nutrient supply indicators at the e-plot scale (randomly selected 1m by 1m plots, each linked to 10 surrounding households) (Scholes et al., 2013) .

The selection of candidate explanatory variables as livelihood capitals, management intensity and soil ecosystem services indicators, were informed by the work conducted in Chapter 2 of this dissertation. The list of the 22 candidate explanatory variables and their measurements are provided in Table 3-2. The final dataset included data for 152 farming household farms producing maize in the region.

3.2.4 Data Processing

Missing values for selected features were assessed. The number of missing data for all selected features was less than 25% of the of the total number of observations. K-nearest neighbor algorithm was used to impute missing values (Kumar et al., 2023). Boxplots were used to visualized outliers. Outliers were not removed to preserve the structure of the data. Variance Inflation factor (VIF) was used to detect multicollinearity amongst variables. Only variables with a VIF < 10 were retained for modeling (O'Brien, 2007). The final dataset excluded the following predictor variables from the initial candidate features in Table 3-2: precipitation, gender ratio, literacy rate, access to microfinancing indicators, soil management intensity, resulting in 12 features used in the ML development. The data processing steps can be found in the dissertation's github repository (See Appendix B).

The target variable "yield class" was determined by classifying maize yield of farms under three discrete classes for modeling: [*Low productivity - class 0*] include yield values below or equal to 1.5 t/ha, which is the average maize yield observed in Sub-Saharan Africa over the last decade ; [*Mid productivity- class 1*] include values ranging above 1.5 t/ha and below 2.5 t/ha. The latter is considered an "attainable yield" as the average yield observed in the region when farms are assisted. [*High productivity - class 3*] include values of and above 2.5 t/ha (Jayne & Sanchez, 2021). The ratio of the number of households in the three productivity classes from highest to lowest were 104:35:13 (Figure 3-2). The distribution of the productivity of farm households across the clusters is presented in Table 3-3.

Although the skewed distribution of the classes could affect the performance of ML algorithms, particularly for predicting minority classes (Wang & Yao, 2012), we expect tree-ensemble algorithms to work best with imbalanced data as demonstrated in several studies (Khoshgoftaar et al., 2007; Zhang et al., 2022). The performance of these models is mainly evaluated based on prediction accuracies on the test dataset as described below.

3.2.5 Machine Learning Model Development

The final dataset of n=152 (farms producing maize) observations was used to build ML models aim at predicting the yield class target. The data was split into a training subset (75%) for training the models, and a test subset (25%) for testing using stratified 10-fold cross-validation (Hastie et al., 2006). Three ML models were tested on their ability to predict yield classes accurately: a baseline logistic regression (LR), and two tree-ensemble methods, Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Logistic regression had been commonly used for classification in many studies prior to recent expansion of more powerful classification algorithms such as tree-based algorithms. It uses an optimized linear boundary in which a nonlinear logistic function is integrated to separate feature classes (Hastie et al., 2006). Tree-based ML algorithms have gained a lot of popularity in predictive modeling. They are non-parametric supervised learning methods that use trees as building blocks to construct powerful predictive models that usually outperform logistic regression models(Hastie et al., 2006). They involve separating the predictor space into several simple regions based on splitting rules arranged in a tree. These tree-based algorithms include decision trees, random forest, gradient boosting and XGBoost. Here we focus on a category of tree-based algorithms ‘ensemble methods’ which incorporate many decision trees instead of a single tree to improve the predictive performance of models. Random Forest and XGBoost algorithms are categorized under ensemble trees(Dangeti, 2017; James et al., 2017). Random Forest (Breiman, 2001; James et al., 2017) is constructed using bootstrap training samples of the dataset (i.e., bagging technique) where unique random subsets of predictors are responsible for splitting the data at each node leading to an outcome. The algorithm predicts the final output based on the majority votes of predictions (Breiman, 2001; James et al., 2017). XGBoost (Chen & Guestrin, 2016) is constructed using a boosting technique where trees are optimized iteratively using previous weak learners. All ML models were calibrated with hyperparameters to reduce overfitting (i.e., allowing the model to generalize and not fitting too closely to the training set). Hyperparameters for the LR model included penalty and solvers, hyperparameters for the RF and XGBoost model included the maximum depth of trees, maximum leaf nodes, the number of trees, the number of features to account for in each split, the minimum samples split).

Model performance was evaluated using accuracy, precision, and recall scores metrics. The accuracy score measures the overall number of correct predictions of the models. It is calculated as the ratio of the number of correctly predicted observations to the total number of observations in the sample. The precision score measures how well the model predicts each class. It is the percentage of predicted classes of target variables (low productivity, mid productivity, high productivity) that are correctly assigned by the machine learning algorithms. The recall measures the number of unique class predictions of the model. It is the percentage of actual classes of target variables that are correctly assigned by the machine learning algorithms. These scores are informed by a 3 -by 3 confusion matrices comparing the number of observations that the model can predict correctly and incorrectly for each productivity class for the train, test and full datasets. Diagonal outputs of the confusion matrix denote the classes of target variables that are classified correctly by the machine learning model, whereas the off-diagonal elements represent the classes that are incorrectly classified by the model. An accuracy score above 0.7 on the test dataset denotes a robust model. The best performing model with the highest accuracy, precision and recall score was retained for further analyses. The model's code can be find in the dissertation's github repository (See Appendix B).

3.2.6 Feature Importance

The relative importance of features on productivity was assessed with the best performing model, comparing the entropy/information gain method to the Global Sensitivity Analysis (GSA) based method. GSA is considered a more robust method applicable for complex system models, as it accounts for higher order interactions amongst features (Saltelli et al., 2008; Sobol', 1993). The Sobol GSA was implemented as follows: (1) Sobol input sequences were generated using the empirical probability distribution functions of the full dataset; (2) model simulations were executed for each feature samples; and (3) the variance of the model probability outputs is then decomposed into fractions attributed to single (direct) or combined (higher order) influential features' effects. The first-order sensitivity indices (i.e., percent variance explained by the model's features), S_i are calculated by dividing the fraction of the model output attributed to the variation of a single feature X_i to the total variance of the model output. The second-order indices (i.e., percent variance explained by the model's feature interactions), S_{ij} are calculated dividing the fraction of the model output attributed to the variation of a combined pair of features X_i to the total variance of the model output. The total order sensitivity indices, S_{Ti} are calculated as the sum of all effects of the variation of a factor Xi (direct) with its interactions of all orders with the other factors. A sample size of N=736000 was determined for performing a reliable Sobol Sensitivity analysis as suggested for models of high number of inputs (Saltelli, 2002). Input distribution assignment, Sobol sampling, and sensitivity analysis were conducted using Python functions from the following repository : (https://github.com/AlvaroCarmonaCabrero/dissertation)

3.3 Results

3.3.1 Model Performance

The performance of all three machine learning algorithms used to predict yield class is shown in Table 3-3 in terms of accuracy, precision, recall and F1 scores. XGBoost had the best training (0.91) and testing accuracy (0.82) over the Random Forest and Logistic regression algorithms. The XGBoost F_1 score of 0.80, and recall score of 0.90 for testing infers that the model correctly predicts 80% of the time and is correct about productivity classes 90% of the time. The confusion matrices of the models are shown in Figure 3-2 to demonstrate correctly and incorrectly predicted classes. Classes 0,1,2 in the figure denotes low productivity, mid productivity, and high productivity classes respectively.The misclassification rates for the low productivity, mid productivity and high productivity classes in the test set are 14%, 29%, and 0% respectively.

3.3.2 Feature Importance

The feature importance of the most predictive XGBOOST model is presented in Figure 3-4 reports the entropy/information gain feature importance results. Features are arranged according to their relative contribution to the model's prediction process, where only 12 of the original 23 features are found important and are presented in the Figure. The leading five influential features were ranked as follow: potassium partial nutrient budget (K_PNB)> phosphorous partial nutrient budget (P_PNB)> distance to roads> household dependency ratio > nitrogen partial nutrient budget (N_PNB)>.

Figure 3-5 reports the Sobol GSA – based feature importance results. Features are arranged according to their first order indices (i.e., percent variance explained by the model's features). The leading influential features were ranked as follows: K_PNB > P_PNB > labor > N_PNB > distance to river. K_PNB is the only feature explaining more than 10% of the variance by itself (direct effect). The model revealed to be highly interactive as the sum of first order indexes < 0.6 (Table 3-3). Features that explained more than 10% of the variance when interacting with one feature include K_PNB and P_PNB (Table 3-4). Features that explained more than 5% of the variance when interacting with one feature include distance to river, labor, distance to roads, gender ratio, N_PNB (Table 3-5).

3.4 Discussion

The results on the performance of the models show that tree-ensemble (i.e., XGBOOST and Random Forest) outperformed the logistic regression baseline model. Both resulted in higher acceptable accuracy. The XGBoost accuracy score on the test was higher (0.82) than the Random Forest accuracy score (0.70). These models revealed to be more powerful at capturing non-linear dynamics between social-ecological factors and productivity outcomes, compared to the logistic regression model, as expected. The few studies that have used similar modeling approaches also had acceptable accuracy for tree-ensemble models (Banerjee et al., 2014; Dutta et al., 2020; Liang, 2023)

Results from the analysis of feature importance shows a difference between the ranking outputs of the entropy/information gain method compared to the ranking outputs of the variance-based method. Consistent ranking orders across the two methods was only applicable for the leading two influential features which are K_PNB and P_PNB. Many studies using the entropy/information gain method to assess feature importance usually denote feature with a factor importance > 0.05 as most important for feature selection (Dutta et al., 2020). A factor importance > 0.05 is observed for all features, suggesting that all features are important to the model's outputs for feature selection. A previous study (Carmona-Cabrero, 2022; Carmona-Cabrero et al., 2023) found that the variance-based approach provides a more robust factor importance ranking than conventional ML feature importance metrics such as the entropy/information gain, particularly for highly interactive models, as it able to quantify hidden interactions between features. Our results from the variance-based approach for assessing feature importance revealed a highly interactive model with all features explaining large fractions of the output variance. These outputs support the message from the outputs of the entropy/information gain outputs, showing the important role of all factors in the model. Feature evaluation approaches involving both the entropy/information gain and variance-based methods are complementary in informing feature selection for either improving the prediction accuracy of the machine learning model or for further analyses evaluating features' behavioral input ranges for scenario modeling.

3.5 Conclusions

Narrowing the maize yield gap in smallholder systems of SSA through sustainable intensification requires understanding the various socio-ecological factors affecting productivity. In this chapter, we propose powerful machine learning modeling approaches able to detect complex non-linear relationships between social-ecological factors and productivity in smallholder systems based on a literature guided benchmarks for maize production in Sub-Saharan Africa. We focused on testing three machine learning models (i.e., Logistic regression, random forest and XGBoost) for their ability to predict the productivity level of smallholders in the SAGCOT region case study, using livelihood capitals, management intensity, and soil health explanatory factors. As expected, amongst the three models, XGBoost outperformed the random forest model and the logistic regression model, having the highest accuracy on training, test, and full datasets. The XGBoost GSA—based factor importance identified soil health indicators (i.e., Potassium partial nutrient budget (K_PNB), Phosphorous partial nutrient budget (P_PNB), a physical capital indicator (i.e., distance to roads), a human capital indicator (i.e., household dependency ratio), and Nitrogen partial nutrient balance(N_PNB)) as most influential factors. However, in terms of importance, all features had a factor importance above 0.05 (Figure 3-5) showing their important role in the model. When comparing the GSA-based factor importance results to the entropy/information gain factor importance results, we found that both methods are able to report the two leading two influential features (i.e., i.e., Potassium partial nutrient budget, Phosphorous partial nutrient budget) similarly, however ranks the other features in a different order. The ability of the GSA-based method to detect the highly interactive nature of the model provided a more robust assessment of feature importance. Most importantly, the outputs of the two methods revealed the important role of all factors in the model's output due to feature interactions.

While several studies used machine learning algorithms to predict crop yield have been they mainly focus on biophysical indicators as explanatory variables(Crane-Droesch, 2018; Ranjani et al., 2021; Rashid et al., 2021). Studies that use a combination of social-economic, management, and biophysical factors to predict crop production, and achieving robust models with acceptable accuracy rare in the literature(Banerjee et al., 2014; Dutta et al., 2020). The work accomplished in this chapter led to a classification model of acceptable predictive skills (accuracy score > 0.70). Finally, the assessment of feature importance through GSA adds a new dimension to our model by providing additional information on interaction between features, allowing for a more informed approach to feature selection towards the improvement and use of the model.

One limitation of the proposed model is the assumption of homogeneity within the established productivity classes. Although a regression approach to modeling yield directly would have been more explicit, our classification approach to modeling productivity is an informed and effective way to capture the challenge of yield variability in smallholder systems and allowing the model to better identify patterns in the data. This approach did not limit our understanding of patterns of relationships between social-ecological explanatory features and productivity levels, as we were able to further conduct GSA with the probabilistic outputs of the model. Future research on such approaches to modeling should further focus on coupling classification machine learning algorithms with GSA with a particular focus on the continuous probabilistic outputs of models to allow for more comprehensive approach to assessing feature importance.

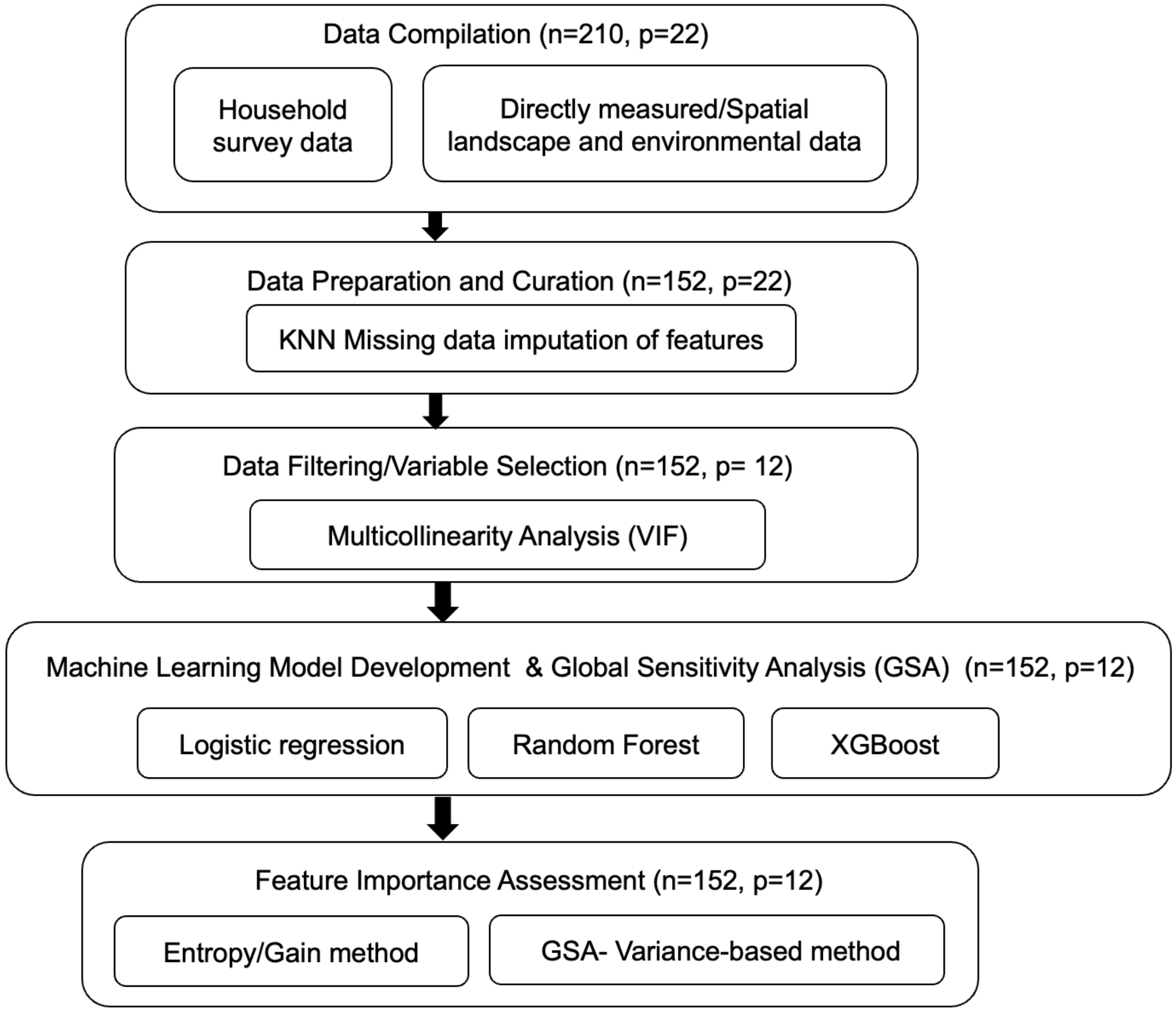

Figure 3-1. Methodological framework employed in this study.

Figure 3-2. The Southern Agricultural Growth Corridor of Tanzania (SAGCOT).The areas circled in the map correspond to the agricultural development clusters of the region for which data was collected.

Table 3-1. Main characteristics of the SAGCOT clusters/landscapes and average crop yield.

Lansdscape	Main characteristics	Average maize yield (kg/ha)	Average rice yield (kg/ha)
Highland - Maize producing			
L03 -Sunbawanga	• 900-1000 mm rainfall • Only landscape with protected areas	1392.7	_
L10 - Ihemi (Mufundi)	• 1300-1600 mm rainfall • Woodlands and wooded grasslands	611.9	_
L11 -Ludewa	• 1000-1300 mm rainfall • Large forest reserves • Sparsely populated	1944	_
L18 - Ihemi (Kilolo)	• 1300-1600 mm rainfall	678.8	_
Lowland - Rice producing			
L19 - Kilombero	• Valley/Floodplain • Small streams	391.4	905
L20 -Mbarali	• Large river tributaries • Fertile soils	952.5	5095
L22 -Rufiji	• 950-1300 mm rainfall • woodlands,grasslands and swamps • Extensive area of mangrove forest	330.3	1023

Table 3-2. Candidate explanatory variables for modeling. In light grey font are features that were eliminated in the VIF reduction dimension step.

Number of features	Features	Measure	Units
Natural capital			
1	Precipitation	[a] Average precipitation from 1981-2016 (mm)	mm
2	River access	[a] Distance to inland water bodies	km
3	Soil surface condition	[a] Slope	%
4	Tree richness	c Tree species richness index	
5	Tree diversity	c Shannon diversity index	
6	Distance to forest	a Distance to nearest forest	
	Human capital		
6	Household dependency ratio	[c] Number of dependents (<15 &>64 years) / Number of working aged people (15-64 years)	_
7	Human labor	[c] labor hours per capita for weeding, fertilizing, harvesting	Hours
8	Household literacy rate	[c] Number of educated members/Total household members	_
9	Gender ratio	[c] Male to Female ratio	_
Social capital			
10	Access to advisory services	[c] None = 1 ; Media only = 2; People = 3 ; Media & people = 4 ; Media & institutions = 5 ; People & institutions = 6 ; Media, people & institutions = 7	_
11	Access to market	No =1 ; Yes = 2	_
Physical capital			
12	Access to wireless communication	[a] distance to nearest cellular or wi-fi tower	km
13	Distance to road	[c] Distance to main roads (km)	km

Number of features	Features	Measure	Units
Financial capital			
14	Access to microfinancing	[c] No = 1; Yes = 2	_
15	Wage entry	[c] Number of wage entry sources in a year	_
Management intensity			
16	Input Intensity Index	[c] No inputs=1;Irrigation only =2; Mechanization only = 3; improved seeds only = 4 ; Fertilizer only =5; Fertilizer + irrigation = 6; Fertilizer + mechanization =7; Fertilizer + improved seeds =8; Fertilizer +improved seeds + irrigation = 9; Fertilizer + improved seeds + mechanization =10; Fertilizer+ improved seeds + mechanization +irrigation =11	_
17	Soil Management Intensity Index	[c] No till + residue retention = 1; No till only = 2; Till + residue retention = 3 ; Till only = 4	_
Soil health			
18	Nitrogen partial nutrient budget	N added - N removed	kg/ha/yr.
19	Phosphorous partial nutrient budget	P added - P removed	kg/ha/yr.
20	Potassium partial nutrient budget	K added - P removed	kg/ha/yr.
21	Soil carbon deficit indicator	[b] Soil Carbon Capacity	_
22	Soil water storage	Soil water storage indicator	_

[a] Estimated using 250 m spatial product

b Measured using 30 m isDA soil properties data

c Measured using directly field sampled data

d Measured using household survey data

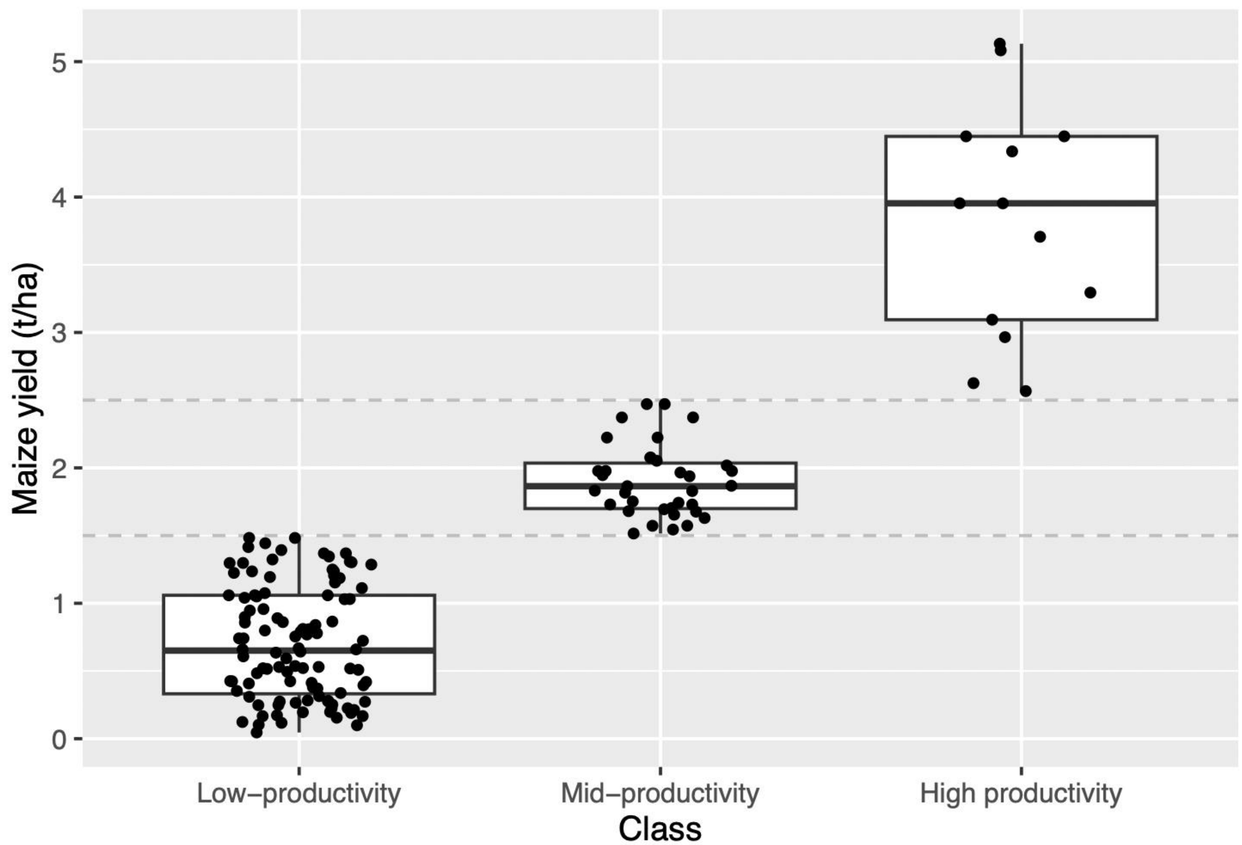

Figure 3-3. Distribution of smallholder maize yield variability across the established productivity classes in the SAGCOT.

Table 3-3. Distribution of maize yield productivity farms across the landscapes of SAGCOT.

Landscape	Low productivity	Mid productivity	High productivity
Highlands			
L03 -Sunbawanga	19	8	3
L10 - Ihemi (Mufundi)	27	1	1
L11 -Ludewa	10	15	5
L18 - Ihemi (Kilolo)	15	4	1
Lowlands			
L19 - Kilombero	11	3	0
L20 -Mbarali	14	2	3
L22 -Rufiji	8	2	0

Table 3-4. Productivity level prediction accuracy of machine learning models.

Model	Dataset	Precision	Recall	F_1	Accuracy
Logistic regression	Training (n=114)	0.768	0.789	0.769	0.79
	Test (n=38)	0.609	0.632	0.585	**0.68**
	Full (n=152)	0.728	0.750	0.725	0.75
Random Forest	Training (n=114)	0.890	0.886	0.875	0.89
	Test (n=38)	0.660	0.737	0.681	**0.78**
	Full (n=152)	0.854	0.849	0.832	0.85
XGBoost	Training (n=114)	0.923	0.921	0.916	0.90
	Test (n=38)	0.81	0.82	0.80	**0.82**
	Full (n=152)	0.90	0.90	0.89	0.88

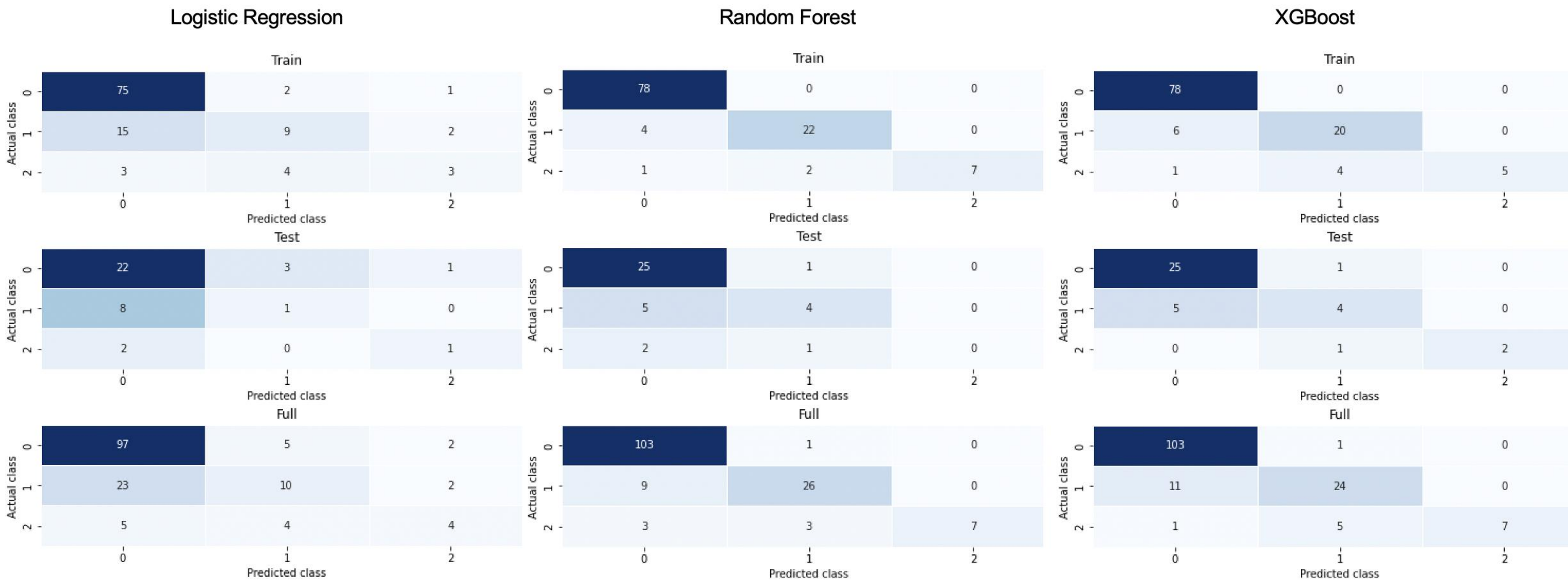

Figure 3-4. Confusion matrix for training, test and full datasets showing the performance of the machine learning algorithms for productivity level prediction. Numbers 0, 1 and 2 denotes low (<1.5t/ha), mid (1.5-2.5 t/ha), and high (>2.5t/ha) Maize productivity classes. Values in diagonal indicate number of correctly predicted productivity class.

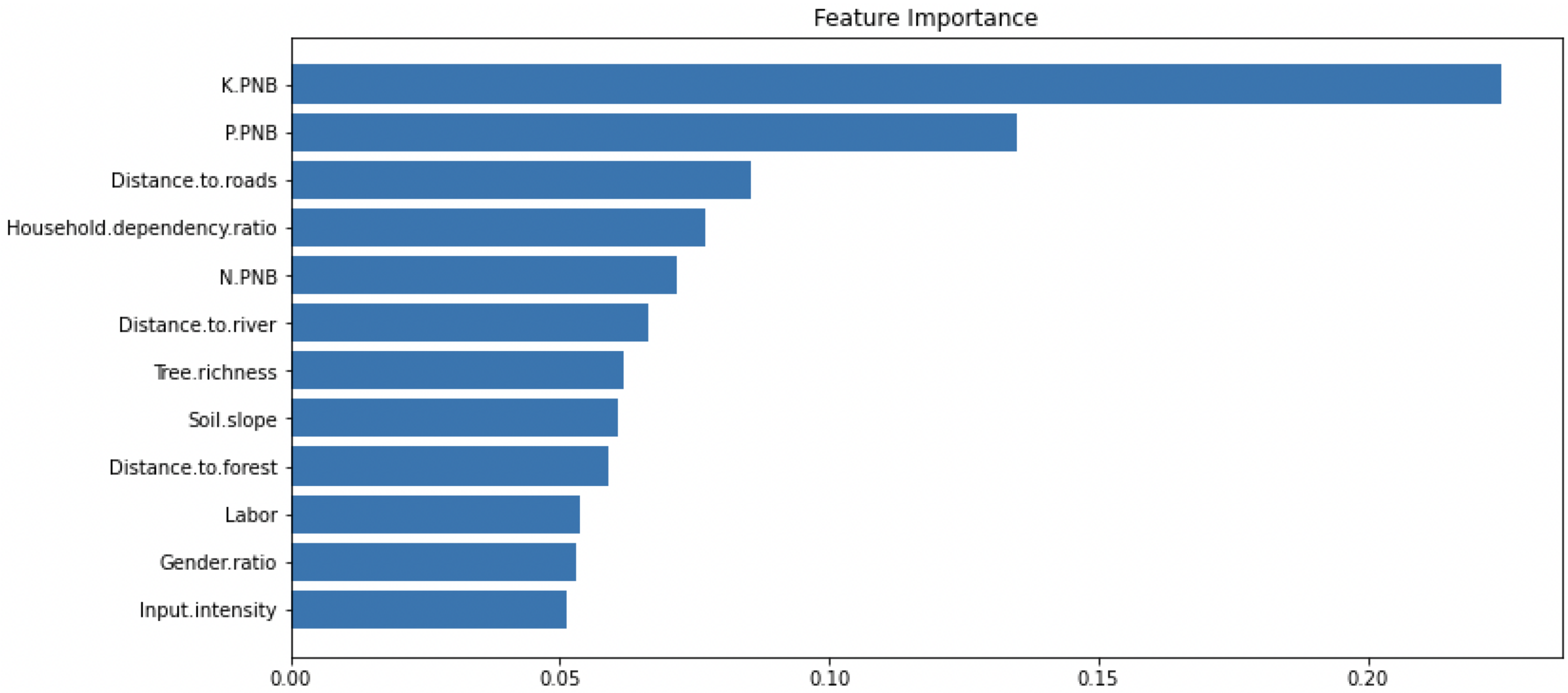

Figure 3-5. Information gain feature importance for the best performing XGBoost ML model on the full dataset in descending order. Potassium partial nutrient budget, Phosphorous partial nutrient budget, Nitrogen partial nutrient budget, are denoted as K.PNB,P.PNB,N.PNB respectively.

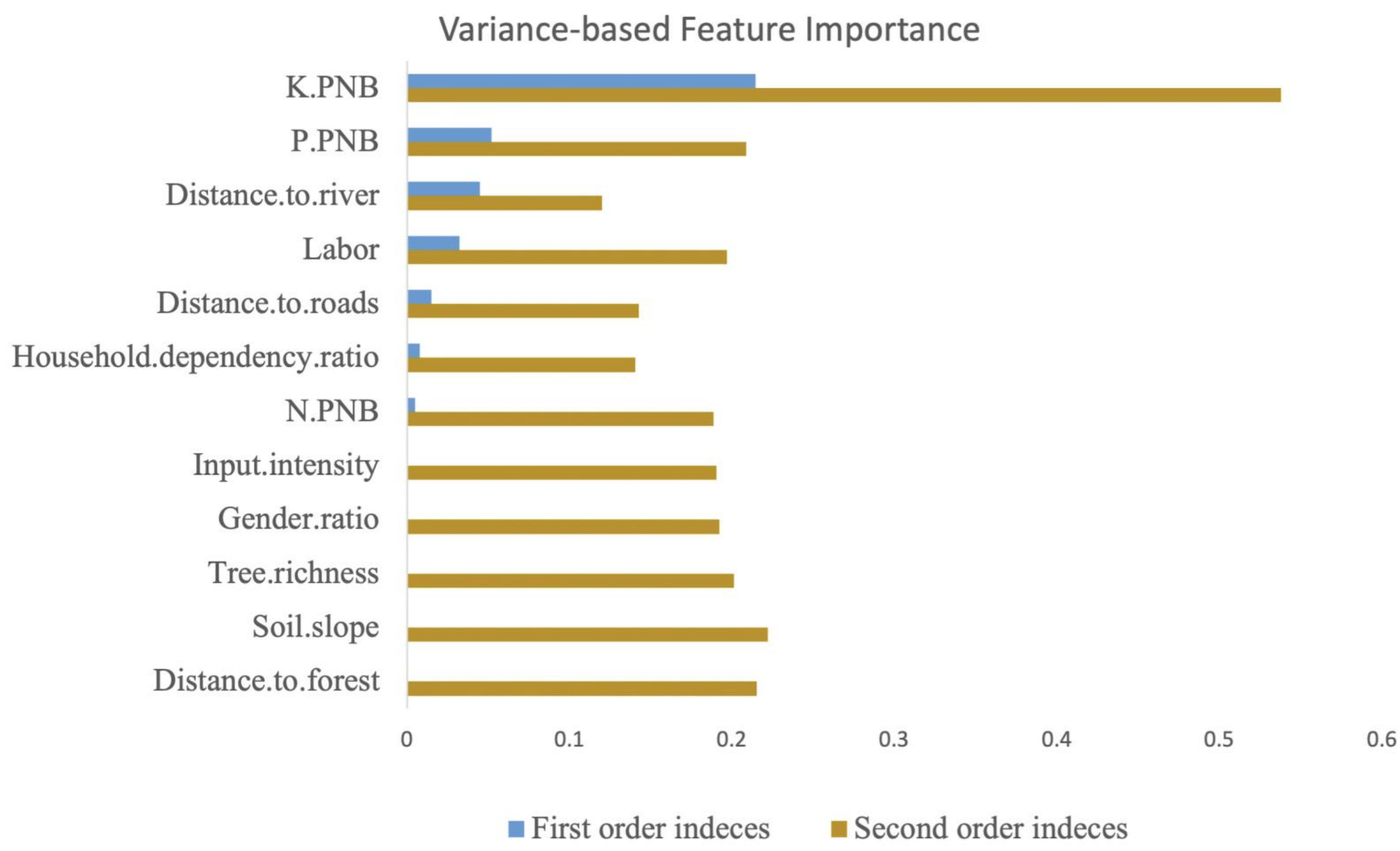

Figure 3-6. Variance-based feature importance of the best XGBoost ML model. Potassium partial nutrient budget, Phosphorous partial nutrient budget, Nitrogen partial nutrient budget, are denoted as K.PNB,P.PNB,N.PNB respectively.

Table 3-5. GSA results for the Sobol sample. First order indexes (S_i), total order indexes (S_{Ti}), interactions ($S_{Ti} - S_i$),and second order indexes (S_{ij}) to quantify the variance of productivity levels probability explained by each factor individually and interactively.

Feature	**First order indeces** (S_i)	**Total order indeces** (S_{Ti})	**Interactions** (S_{ij})
Distance.to.forest	0	0.215354506	0.215354506
Soil.slope	0	0.222096417	0.222096417
Tree.richness	0	0.201249803	0.201249803
Gender.ratio	0	0.192364175	0.192364175
Input.intensity	0	0.190299573	0.190299573
N.PNB	0.005162038	0.193745035	0.188582998
Household.dependency.ratio	0.008044791	0.148510593	0.140465802
Distance.to.roads	0.015069266	0.157663264	0.142593997
Labor	0.032531537	0.229422496	0.196890959
Distance.to.river	0.044968948	0.164797683	0.119828735
P.PNB	0.052193003	0.260897887	0.208704884
K.PNB	0.214322934	0.752110589	0.537787656
Sum of indexes	**0.372292516**	***2.928512021***	***2.556219505***

Feature Interactions	**Second order indeces** (S_{ij})
Distance.to.river X Distance.to.forest	0
Distance.to.river X Tree.richness	0
Distance.to.river X Gender.ratio	0
Distance.to.river X Household.dependency.ratio	0
Distance.to.river X Distance.to.roads	0
Distance.to.river X Input.intensity	0
Distance.to.river X Soil.slope	0.00042917
Distance.to.roads X N.PNB	0.003763976

Table 3-5. Continued

Feature Interactions	Second order indeces (S_{ij})
Distance.to.river X N.PNB	0.004466366
Household.dependency.ratio X Distance.to.roads	0.005160806
Household.dependency.ratio X Labor	0.005448588
Labor X Distance.to.roads	0.005723094
Gender.ratio X Distance.to.roads	0.007912257
Distance.to.river X Labor	0.009070693
Soil.slope X Distance.to.roads	0.009416653
Distance.to.roads X Input.intensity	0.009762059
Gender.ratio X Labor	0.009830873
N.PNB X Input.intensity	0.010349837
Distance.to.forest X Labor	0.010887058
Household.dependency.ratio X Input.intensity	0.010969061
Household.dependency.ratio X P.PNB	0.012031468
Gender.ratio X Household.dependency.ratio	0.014083001
Tree.richness X Distance.to.roads	0.01462553
Soil.slope X Household.dependency.ratio	0.015415198
Distance.to.river X P.PNB	0.015695761
Household.dependency.ratio X N.PNB	0.015780928
Labor X N.PNB	0.016056552
Labor X Input.intensity	0.016188508
Gender.ratio X Input.intensity	0.016689255
Distance.to.forest X Input.intensity	0.01678913
Soil.slope X Labor	0.01696293
Distance.to.forest X Distance.to.roads	0.017086727
Distance.to.forest X Household.dependency.ratio	0.01738135
Soil.slope X Gender.ratio	0.017678988

Table 3-5. Continued

Feature Interactions	Second order indeces (S_{ij})
Soil.slope X Input.intensity	0.01793604
Tree.richness X Input.intensity	0.01794181
Distance.to.forest X N.PNB	0.018460845
Gender.ratio X P.PNB	0.018575111
Tree.richness X Household.dependency.ratio	0.019410801
N.PNB X P.PNB	0.019586653
Distance.to.roads X P.PNB	0.019986557
Distance.to.forest X Soil.slope	0.020532074
Tree.richness X Labor	0.020871114
Gender.ratio X N.PNB	0.020879372
Labor X P.PNB	0.021071696
Tree.richness X N.PNB	0.022285212
Tree.richness X Gender.ratio	0.022782957
Distance.to.forest X Tree.richness	0.023024761
Distance.to.forest X Gender.ratio	0.023024885
Soil.slope X Tree.richness	0.024290001
Distance.to.forest X P.PNB	0.024314851
Tree.richness X P.PNB	0.025618871
Soil.slope X P.PNB	0.025631169
Soil.slope X N.PNB	0.026125364
P.PNB X Input.intensity	0.026192745
Household.dependency.ratio X K.PNB	0.030233189
Tree.richness X K.PNB	0.034688747
Distance.to.forest X K.PNB	0.043852826
K.PNB X Input.intensity	0.047481128
Soil.slope X K.PNB	0.04767883

Table 3-5. Continued

Feature Interactions	**Second order indeces** (S_{ij})
N.PNB X K.PNB	0.050414545
Gender.ratio X K.PNB	0.054990477
Distance.to.roads X K.PNB	0.058643409
Labor X K.PNB	0.087465878
Distance.to.river X K.PNB	0.090166744
P.PNB X K.PNB	0.119349702

CHAPTER 4
IDENTIFYING EFFECTIVE INTERVENTIONS FOR INCREASING PRODUCTIVITY IN SMALLHOLDER SYSTEMS IN THE SAGCOT REGION OF TANZANIA

4.1 Background

Agricultural productivity models such as the ML model developed in chapter 2 of this dissertation, can inform agricultural scientists and policymakers in dealing with the complex challenges of Food Security. Integrated approaches to testing ML model reliability through global sensitivity and uncertainty analysis (GSUA) are key for providing practical guidance on model use. The scope of GSUA is not only to identify the important input factors that drive output uncertainty but also to identify the ranges of these input factors that are responsible for the model realization in acceptable ranges for system management (Saltelli et al., 2004).This is carried out by Monte Carlo Filtering (Gordon et al., 1993; Kitagawa, 1996). In Monte Carlo filtering, a set of constraints that targets the desired characteristics of the model realization (e.g., an acceptable range of outputs, a threshold value, or a ceiling value as set by ecosystem managers or stakeholders) has to be defined. Based on the results of the uncertainty analysis obtained in chapter 3, Monte Carlo Filtering is then performed on the model's input factors to define management-favorable outputs, in our case an increase in Maize yield among smallholders in the SAGCOT. This chapter builds on Chapter 3 ML modeling and input factor importance analysis results to identify intervention strategies that are effective for lowering the probability of low Maize productivity outcomes using global sensitivity and uncertainty analysis and Monte Carlo Filtering.

4.2 Methods

4.2.1 Methodological workflow

The methods used in this chapter followed the workflow described in Figure 4-1. It consists of 4 main steps described in the sections below: 1) machine learning model prediction with the original sample 2) pre-intervention uncertainty analysis with Sobol sample 3) design of intervention strategies through Monte Carlo filtering 4) post-intervention uncertainty analysis with Sobol sample.

4.2.2 Quality of Sobol sampling and ML predictions before intervention

As presented in Chapter 3, the ML XGBoost model includes a total of 12 important input factors (Fig. 4-2) to predict the maize productivity of a smallholder farm. Applying the model to all 152 households in the dataset allows to calculate the probability that households will be below the low productivity threshold (<1.5t/ha).

Input and output uncertainties in this model are likely to arise from various sources: accuracy flaws of survey data, the inherent variability of the landscapes and the households characterizing the sample, and the assumption that productivity is homogeneous across classes. Here we test if the predictions generated by the larger Sobol sample of the input factors is consistent with the probabilities of the original input data set. This is important so that the results from the Monte Carlo filtering and intervention analysis are relevant to the original pre-intervention conditions of the region, and not an artifact of the large sample that might contain input factors that are not representative of the dataset. For this, boxplots of the observed yield classes are compared to those predicted by the ML model based on the Sobol sample.

While the Sobol sample results is expected to have wider ranges, we are looking for similarity in the proportions of the low, mid and high yield classes. The Sobol uncertainty analysis was conducted for a sample size of N=736000 (see details of Sobol GSA implementation in chapter 3). The code for the uncertainty analysis can be accessed in the dissertation's github repository (See Appendix B). These results are used for the intervention analysis based on Monte Carlo Filtering as described below.

4.2.3 Monte Carlo Filtering

The Monte Carlo Filtering (MCF) technique uses Monte Carlo sampling methods to identify range of input interventions that corresponds to an output behavior of interest (Gordon et al., 1993; Kitagawa, 1996; Kroese et al., 2011; Sarkar, 2003) . It was employed on the Sobol sample outputs to filter for input variability range corresponding to lower probability score threshold for low productivity. Since the initial Maize yields in the original dataset were low, with 85% of the initial smallholders below the 1.5t/ha threshold, we selected as an intervention threshold, i.e., the behavioral outcome, to decrease the number of households with low yields to less than 1/3 of the total or $p_{low-productivity} < 0.33$. Based on the full Sobol simulation set, ML model yield class results were then classified as being favorable or "behavioral (B)" (mid- high-yields, $p_{low-productivity} < 0.33$) and unfavorable or " non behavioural ($\overline{B}$)" (low yields, $p_{low-productivity} > 0.33$) realizations. Two subsets of each input factor, X_i, corresponding to the two realizations were defined as X_i/B and $X_i/\overline{B}$, consisting of m and n elements, respectively, where m + n equals the total number of simulations N. In order to check the separation of distributions, X_i/B and $X_i/\overline{B}$, a two-sided Kolmogorov-Smirnov (K-S) test was performed for each input factor under a null hypothesis that the distribution of the subsets producing B realization is identical to the one producing $\overline{\mathrm{B}}$. The hypotheses used are as follows:

$$H_o: f_m\left(\frac{X_i}{\mathrm{B}}\right) = f_n\left(\frac{X_i}{\overline{\mathrm{B}}}\right) \quad (4\text{-}1)$$

$$H_1: f_m\left(\frac{X_i}{\mathrm{B}}\right) \neq f_n\left(\frac{X_i}{\overline{\mathrm{B}}}\right) \quad (4\text{-}2)$$

where $f_m(X_i|B)$ and $f_n(X_i/\overline{\mathrm{B}})$ are the probability distribution functions (pdfs) of input factor, Xi, belonging to B and $\overline{\mathrm{B}}$ realizations, respectively. The K-S test statistic is given as:

$$d_{m,n}(X_i) = sup_y \left\| F_m\left(\frac{X_i}{\mathrm{B}}\right) - F_n\left(\frac{X_i}{\overline{\mathrm{B}}}\right) \right\| \quad (4\text{-}3)$$

where $F_m(X_i|B)$ and $F_n(X_i/\overline{\mathrm{B}})$ are the empirical cumulative probability cdfs of the two input factor subsets corresponding to B and $\overline{\mathrm{B}}$, respectively. The test statistic, $d_{m,n}$, measures the largest vertical distance between the empirical cdfs of X_i/B and $X_i/\overline{\mathrm{B}}$. A low level (i.e., smaller p-value) implies a significant difference between $f_m(X_i|B)$ and $f_n(X_i/\overline{\mathrm{B}})$ (i.e., larger $d_{m,n}$) suggesting that X_i is a key factor in producing the defined behavior for the model while a high level (i.e., higher p-value, smaller $d_{m,n}$) suggests that any value in the predefined range of the input, X_i, is likely to fall either in B or $\overline{\mathrm{B}}$ (Saltelli et al., 2004). One caveat of this method however is that no higher-order analysis is performed, i.e., if interactions are present, no attempt is made to search for interaction structure to identify particular combinations of filtered inputs resulting in the desired output(Saltelli et al., 2004). The 12 Input factors that were found significant in the previous GSA (Chapter 3) were selected for interventions. The MCF was performed in R software. See Appendix B to access the code for the Monte Carlo Filtering method in the dissertation's github repository.

4.3 Results

4.3.1 Quality of Sobol Sampling and ML Predictions

Figure 4-2 shows the distribution of the probability scores for the three productivity classes obtained from the uncertainty analysis for the original dataset compared to the Sobol sample. Based on the Sobol sample probability outcome, about 85% of the observations exceeded the probability score threshold in the low-productivity sample against 76% in the original sample, 11% of the observations exceeded the threshold in the mid-productivity sobol sample against 19% in the original dataset, and 3% exceeded the threshold in the high-productivity sobol sample against 5% in the original dataset. The median probability score of the low-productivity sample is 0.89 in the Sobol sample compared to 0.95 in the original sample. The median probability score of the mid-productivity sample is 0.59 in the Sobol sample compared to 0.74 in the original sample. The median probability score of the high-productivity sample is 0.23 in the Sobol sample compared to 0.58 in the original sample. The low-productivity sample had the least variance. Since the model has a lower misclassification rate for the low-productivity class than the mid productivity class, and the lowest variance for the low-productivity class (Chapter 3), the results around the change in low productivity households are expected to be robust and further support centering our intervention goal on lowering probability scores of the low-productivity below 0.33, allowing for an increase of probability scores of the mid and high productivity classes.

4.3.2 Design and Evaluation of Intervention Scenarios to Increase Maize productivity.

Monte Carlo Filtering (MCF) is used to quantitatively guide the identification of effective interventions able to reduce the probability of low-productivity classification by the model. The MCF was performed with all input factors of the model to capture the interactions detected from the GSA-based feature importance assessment. The MCF separated the distribution and ranges of input factors resulting in probability scores above and below the threshold probability value for low-productivity ($p_{\text{low-productivity}}$ <0.33). The Kolmogorov-Smirnov test was used to determine if there are significant differences ($p < 0.05$) between behavioral and non-behavioral distributions. Significant differences were confirmed for 9 out of the 12 input factors included in the model, making them suitable for intervention.The factors that were not significant were gender ratio, distance to forest, and tree richness. The results of the MCF are summarized in Appendix B. Based on the MCF, smallholders' low productivity classification is minimal under nutrient management conditions where potassium partial nutrient budget (K_PNB) ≤ -6 kg/ha/yr., phosphorous partial nutrient budget (P_PNB)≤ -7 kg/ha/yr., and nitrogen partial nutrient budget (N_PNB) ≤ -23 kg/ha/yr., landscape conditions where distance to river ≥ 11.7 km, distance to roads ≥ 4 km, soil percent slope ≥ 2.1%, household conditions where labor ≥ 28 hours/day, and household dependency ratio ≤ 0.49. In the case where any of these levels exceed or are below these thresholds, the households are at risk of low productivity, exceeding the $p > 0.33$ threshold.

These threshold values were used to inform new range of values for each input factor (Table 4-1) to evaluate the performance of potential intervention strategies able to reduce risk of low productivity and move more smallholders' farms into the mid and high productivity classes. The input ranges for all soil health factors including landscape and household factors, were refined with their respective behavioral range values. The input ranges for these factors were refined using with ranges between acceptable levels of nutrient budgets for nitrogen [50;50)], phosphorous [-20,20], and potassium [-50,50] kg/ha/yr in the region, determined by the Vital Signs project, and the behavioral ranges of input factors obtained from the MCF.

An uncertainty analysis was performed using refined input ranges of the behavioral distributions for each input factor (Table 4-1). We first evaluated the performance on each landscape intervention separately on the model's low-productivity probability output, using the refined input factor ranges for landscape physical capital factors. The median probability score for the entire Sobol sample decreased from 0.86 to 0.84 when 'distance to river' is refined, to 0.83 when 'distance to road' is refined. All landscape intervention median values were above the threshold of 0.33 (Figure 4-4), showing that landscape physical capital factors alone had no significant impact on reducing the probability score for low productivity as single or combined interventions.

In the case of household human and physical factors, the median probability score for the entire Sobol sample decreased from 0.86 to 0.71 when labor is refined, to 0.80 when household dependency ratio is refined and did not decrease when input intensity is refined. All household interventions median values were above the threshold value of 0.33, demonstrating that household factors alone had no significant impact on reducing the probability of low productivity of smallholders as single or combined interventions (Figure 4-5).

In the case of soil health factors, the median probability score for the entire Sobol sample decreased from 0.86 to 0.84 when soil slope is refined, to 0.56 when Potassium partial nutrient budget (K_PNB) is refined, to 0.69 when Phosphorous partial nutrient budget (P_PNB) is refined, to 0.74 when Nitrogen partial nutrient budget(N_PNB) is refined. These soil health factors are demonstrated as most effective in reducing the probability of low productivity for smallholders(Figure 4-6).

The poor performance evaluation of each intervention strategy individually on reducing the probability of low productivity confirms the need to combine interventions using the results from the feature importance evaluation (Table 4-2), which show the interactive effects amongst input factors as a guide. The outcomes of the soil health interventions (Figure 4-6) and the interactions amongst K_PNB and P_PNB accounting for more than 10% of the model's output variance, demonstrate that centering intervention strategies around K_PNB, P_PNB, are more likely to have the greatest impact on reducing the probability of low productivity amongst smallholders in the SAGCOT.

We designed the first intervention strategy (IS1) which consisted of refining K_PNB, P_PNB, N_PNB combined, by adjusting partial nutrient budgets to an acceptable range considering critical and the MCF thresholds established (Table 4-1). We decided to include N_PNB in this intervention, although it isn't a leading influential factor, to reflect the reality of NPK mineral fertilizer availability in Sub-Saharan Africa. Moreover, this capture the relevant interactions between N_PNB and K_PNB wich account for 5% of the model's output variance (Table 4-2). Adjusting partial nutrient balances of nitrogen, phosphorus and potassium hypothetically reflect an increase in nutrient gains through the use of larger quantities of NPK fertilizers by smallholders in the non-behavioral partial nutrient budget ranges, able to reduce the deficit observed. In a practical sense, such intervention consist of providing smallholders with increased access to mineral fertilizers. Alternative interventions that are likely to be less cost-effective could also include proving smallholder with enhanced efficiency fertilizers, and promoting the use of more organic sources of fertilizers with higher concentration of Potassium, considering the leading importance of Potassium partial nutrient budget in the model.

The second intervention strategy (IS2) considered consisted of refining soil slope for smallholders in the non-behavioral range, to reflect a positive change in soil health conditions that the 'soil slope' indicator may capture, as the behavioral range for soil slope suggests that an increase in soil slope corresponds to higher productivity classification. We choose soil slope this potential next best strategy as it has the highest total second order index after K_PNB (Table 4-2). We hypothesized that this positive change would reflect higher productivity and soil fertility in the landscapes positioned in the highlands compared to the landscapes positioned iin the lowlands of the region, which tend to be less productive. When combined with nutrient gains, this intervention could considerably increase productivity.

The third intervention strategy (IS3) considered accounted for all household capitals interventions together in addition to addressing soil health. We chose this strategy next because the total interactions amongst household factors were more important (S_i -S_{Ti} = 0.043) than total interactions amongst landscape factors (S_i -S_{Ti} =0.01). This strategy should captures the use of additional inputs including seeds, mechanization towards intensifying farming via input intensity, increase in hired labor and labor capacity by the household.

The fourth intervention strategy (IS4) finally accounts for all landscape level infrastructure interventions in addition to the previous ones. The behavioral range for the distance to river intervention suggests that an increase in distance to a nearby waterbody corresponds to higher productivity classification. This factor reflects a positive advantage of farms situated further from waterbodies. Farms located nearest to water bodies in the region tend to be in the lowland less productive areas, while farms further from the water bodies tend to be in the more productive areas. This intervention could represent an improvement in access to water and irrigation infrastructures more likely to be concentrated in the more productive areas. We also adjusted the distance to roads range for smallholders in the non-behavioral range by reducing the distance between their farms and the main road to reflect an improvement in road infrastructure.

4.4 Discussion

Nutrient depletion is a serious challenge in soils managed by smallholders in SSA. Over decades, smallholders' cropping management have led to large quantities of nutrients being removed from their soils without sufficient replenishment, causing soils in the regions to have negative nutrient balances, with more nitrogen and potassium getting depleted over phosphorous (Chianu et al., 2012; Henao & Baanante, 1999; Sanchez, 2002; Smaling et al., 2015). Potassium depletion rates are even more alarming (Goulding et al., 2020). In the case of the smallholder sample of the SAGCOT region, median partial nutrient depletion rates per hectare of cultivated land are higher for Nitrogen (-8 kg) over Potassium (-5 kg) and Phosphorous (-3 kg). We designed four interventions strategies around remediating nutrient imbalance of soils, each combined with specific combinations of household and landscape factors, in order to identify the most effective strategy able to move low productivity smallholders to high productivity levels.

Figure 4-7 shows the results of the four intervention strategies. The median probability score for low productivity for the entire sobol sample decreased from 0.86 to 0.21 with the first intervention strategy (IS1), to 0.186 with the second intervention strategy (IS2), to 0.09 with the third intervention strategy (IS3), and to 0.06 with the fourth intervention strategy. All intervention median values were below the threshold of 0.33 (Figure 4- 7), showing that these strategies are all effective for increasing productivity of smallholders in the SAGCOT. IS3 and IS4 are the most effective out of the four intervention strategies, with IS4 being the only strategy able to completely shift all low-productivity smallholders into higher productivity levels, more spefically 89% into the mid-productivity range and 11% of smallholders in the high productivity range. However, we observed a trade-off between higher producity classes for IS3 and IS4. For both interventions, the number of high productivity farms decrease as the number of mid-productivity farms increase (Table 4-3). These results suggest that interventions around household level human and physical capital, and large-scale infrastructural projects have diminishing returns. Undergoing IS3 and IS4 may be more aspirational in a pro-poor development context, where agricultural development projects may aim to focus on providing access to labor and infrastruturre to the lower productivity farmers. However considering the large financial and time investments of such projects, they may not be at the forefront of advisable strategies aim at addressing urgent and more inclusive food security challenges that benefit an entire region or country. Such dilemma might raise questions on the cost and social acceptability of landscape interventions. If large-scale interventions such as IS3 and IS4 are not promoted, than low-productivity marginalized farmers continue to be at risk. This could further deepens socio-economic divide between low and higher productivity farmers, reinforcing a productivity trap in the long term, where cost-effective interventions keep failing to address poverty and food security issues rooted in agricultural development.

These results present important decision support tools that can guide the planning of sustainable intensification strategies for smallholder communities, towards improving productivity and livelihood conditions simultaneously. Agricultural development programs centered around increasing mineral fertilizer uptake and promoting sustainable cropping management practices to prevent further nutrient loss from farm fields, must be accompanied with household and landscape level interventions that contribute to increasing physical and human capital of households, for low-productivity farms to reach their full potential. Household human and physical interventions include allowing smallholders to meet labor requirements for their farms based on the labor structure and the dependency ratio of the household. Physical capital interventions at the landscape scale include improving rural roads infrastructure to help with increasing accessibility to mineral fertilizers and markets and providing irrigation infrastructure through reliable water sources. These interventions prove to be all together effective for lifting farmers out of low-productivity, and balancing the social, and economic objectives of sustainable intensification in smallholder agriculture.

4.5 Conclusions

Identifying effective interventions for sustainable intensification of smallholder systems in SSA requires a comprehensive understanding the drivers of productivity. In this study we build on a productivity model able to predict productivity levels using social-ecological and soil health factors, to design effective interventions for increasing productivity of smallholders. The uncertainty analysis of the ML based on a large Sobol sample set confirmed that the ML is able to reproduce the observed distribution of Maize yields in Tanzania's SAGCOT region, where the majority smallholders produce less than 1.5 t/ha Maize, classifying their farms as low productivity farms. Through Monte Carlo filtering, we found that smallholders can move up to higher productivity levels, within acceptable behavioral ranges for Nitrogen, Phosphorous and Potassium partial nutrient budgets, and behavioral ranges for landscape physical and natural capital, along with household human and physical capital factors. We designed practical intervention strategies that focus on tackling nutrient imbalance, faced by many smallholders in Africa, by combining several factor interventions. These interventions were proved to be effective through post-intervention uncertainty analyses, with the more integrated strategy which included nutrient imbalance remediation, improvement in soil health, increase in household physical and human capitals such as labor and input use, and improvements of landscape water and road infrastructure, able to shift 100% of smallholders in the SAGCOT from low productivity to mid and high productivity levels. These results can assist decision-support systems on sustainable intensification interventions in smallholder systems.

We were able to demonstrate that strategies for increasing productivity in smallholder systems cannot be addressed using a “one size fits all” approach. This is supported by the results of the variance-based feature importance assesment which is quantified by the Sobol GSA of our highly interactive ML maize productivity model. As a result, a multi-solution approach consisting of intervening across the entire range of livelihood capitals (i.e., natural, human, social, physical, financial) proved to be more effective. Our integrated socio-ecological approach to modeling crop yield proves to be effective for understanding yield variability across ranges of biophysical and socio-economic gradients of smallholder farms. Even if those biophysical gradients are well - understood at the landscape or regional scales, a social ecological approach to modeling productivity add value to biophysical crop and water simulation models in capturing farm/field scale hidden social-ecological patterns of farming that affect productivity. Potential limitations of this modeling approach are that more biophysical factors used to parameterize crop models could be integrated as input factors for more equal contributions of biophysical and socio-economic factors. Additionally, in our model, data gathered at the landscape scale were associated to multiple farms in an e-plot. A more homogeneous approach to scale can help the model capture important granular level patterns that may impact the model differently. Future work on this study should evaluate the cost-benefit analysis of intervention strategies.

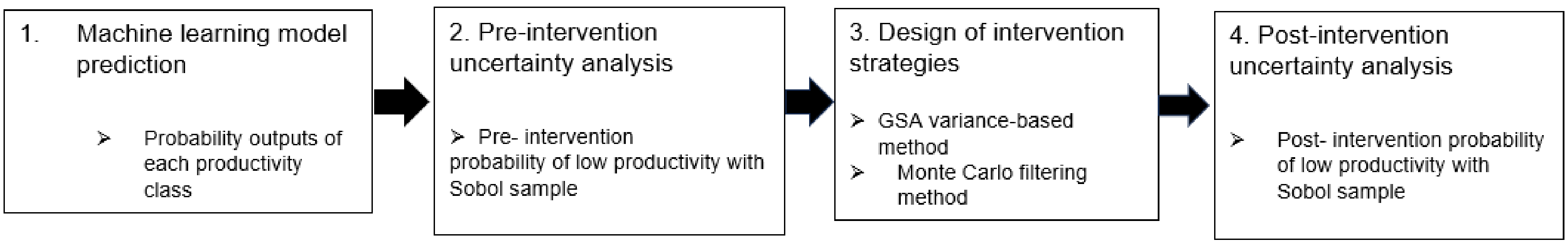

Figure 4-1. Methodological workflow used in this study.

Table 4-1. Range of the model's input factors from the original dataset pre-intervention, and the intervention ranges for significant input factors (Kolmogorov-Smirnov test, α=0.05) for reducing low Maize yield probability.

Model parameters	Pre-intervention range	Acceptable range	Range of intervention
Input factors			
Landscape physical capital			
Distance to roads (km)	[0 - 11.45]	_	[0 - 3.7]
Distance to river (km)	[0.25 - 22.2]	_	[11.27 - 22]
Distance to forest (km)	[0.5 - 45.1]	_	_
Soil percent slope	[0.1 - 13.1]	_	[2.5 - 13]
Tree richness	[0.1 - 29]	_	_
Household human and physical capitals			
Gender ratio	[0.1 - 7]	_	_
Household dependency ratio	[0.2 - 1]	_	[0.17 - 0.49]
Labor (hours)	[0.3 - 95.6]	_	[28 - 95]
Input intensity index	[1-14]	_	[2.4-15]
Soil health			
Nitrogen partial nutrient budget (kg/ha/yr)	[-142.7 - 178.5]	[-20,20]	[-23,-20]
Phosphorous partial nutrient budget (kg/ha/yr)	[-42.3 - 180.0]	[-5,5]	[-42,-6]
Potassium partial nutrient budget (kg/ha/yr)	[-40.3 - 44.5]	[-20,20]	[-20,-7]
Outputs			
Probability of low productivity class	[0.02 - 0.9]	_	_
Probability of mid productivity class	[0.01 - 0.9]	_	_
Probability of high productivity class	[0.008 - 0.5]	_	_

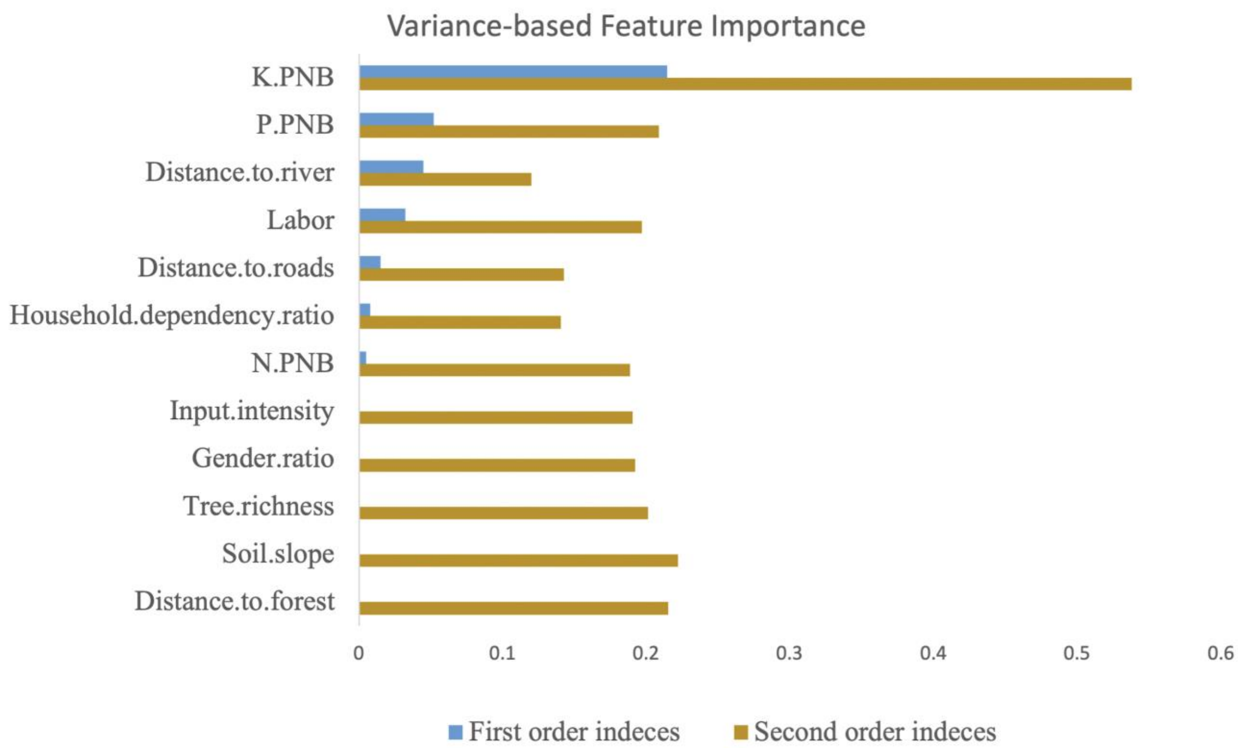

Figure 4-2. Variance-based feature importance of model's input factors.

Table 4-2. GSA results for the Sobol sample. First order indexes (S_i), total order indexes (S_{Ti}), interactions ($S_{Ti} - S_i$),and second order indexes (S_{ij}) to quantify the variance of productivity levels probability explained by each factor individually and interactively.

Feature	**First order indeces** (S_i)	**Total order indeces** (S_{Ti})	**Interactions** (S_{ij})
Distance.to.forest	0	0.215354506	0.215354506
Soil.slope	0	0.222096417	0.222096417
Tree.richness	0	0.201249803	0.201249803
Gender.ratio	0	0.192364175	0.192364175
Input.intensity	0	0.190299573	0.190299573
N.PNB	0.005162038	0.193745035	0.188582998
Household.dependency.ratio	0.008044791	0.148510593	0.140465802
Distance.to.roads	0.015069266	0.157663264	0.142593997
Labor	0.032531537	0.229422496	0.196890959
Distance.to.river	0.044968948	0.164797683	0.119828735
P.PNB	0.052193003	0.260897887	0.208704884
K.PNB	0.214322934	0.752110589	0.537787656
Sum of indexes	**0.372292516**	***2.928512021***	***2.556219505***

Feature Interactions	**Second order indeces** (S_{ij})
Distance.to.river X Distance.to.forest	0
Distance.to.river X Tree.richness	0
Distance.to.river X Gender.ratio	0
Distance.to.river X Household.dependency.ratio	0
Distance.to.river X Distance.to.roads	0
Distance.to.river X Input.intensity	0
Distance.to.river X Soil.slope	0.00042917
Distance.to.roads X N.PNB	0.003763976

Table 4-2. Continued

Feature Interactions	**Second order indeces** (S_{ij})
Distance.to.river X N.PNB	0.004466366
Household.dependency.ratio X Distance.to.roads	0.005160806
Household.dependency.ratio X Labor	0.005448588
Labor X Distance.to.roads	0.005723094
Gender.ratio X Distance.to.roads	0.007912257
Distance.to.river X Labor	0.009070693
Soil.slope X Distance.to.roads	0.009416653
Distance.to.roads X Input.intensity	0.009762059
Gender.ratio X Labor	0.009830873
N.PNB X Input.intensity	0.010349837
Distance.to.forest X Labor	0.010887058
Household.dependency.ratio X Input.intensity	0.010969061
Household.dependency.ratio X P.PNB	0.012031468
Gender.ratio X Household.dependency.ratio	0.014083001
Tree.richness X Distance.to.roads	0.01462553
Soil.slope X Household.dependency.ratio	0.015415198
Distance.to.river X P.PNB	0.015695761
Household.dependency.ratio X N.PNB	0.015780928
Labor X N.PNB	0.016056552
Labor X Input.intensity	0.016188508
Gender.ratio X Input.intensity	0.016689255
Distance.to.forest X Input.intensity	0.01678913
Soil.slope X Labor	0.01696293
Distance.to.forest X Distance.to.roads	0.017086727
Distance.to.forest X Household.dependency.ratio	0.01738135
Soil.slope X Gender.ratio	0.017678988

Table 4-2. Continued

Feature Interactions	**Second order indeces** (S_{ij})
Soil.slope X Input.intensity	0.01793604
Tree.richness X Input.intensity	0.01794181
Distance.to.forest X N.PNB	0.018460845
Gender.ratio X P.PNB	0.018575111
Tree.richness X Household.dependency.ratio	0.019410801
N.PNB X P.PNB	0.019586653
Distance.to.roads X P.PNB	0.019986557
Distance.to.forest X Soil.slope	0.020532074
Tree.richness X Labor	0.020871114
Gender.ratio X N.PNB	0.020879372
Labor X P.PNB	0.021071696
Tree.richness X N.PNB	0.022285212
Tree.richness X Gender.ratio	0.022782957
Distance.to.forest X Tree.richness	0.023024761
Distance.to.forest X Gender.ratio	0.023024885
Soil.slope X Tree.richness	0.024290001
Distance.to.forest X P.PNB	0.024314851
Tree.richness X P.PNB	0.025618871
Soil.slope X P.PNB	0.025631169
Soil.slope X N.PNB	0.026125364
P.PNB X Input.intensity	0.026192745
Household.dependency.ratio X K.PNB	0.030233189
Tree.richness X K.PNB	0.034688747
Distance.to.forest X K.PNB	0.043852826
K.PNB X Input.intensity	0.047481128

Table 4-2. Continued

Feature Interactions	**Second order indeces** (S_{ij})
Soil.slope X K.PNB	0.04767883
N.PNB X K.PNB	0.050414545
Gender.ratio X K.PNB	0.054990477
Distance.to.roads X K.PNB	0.058643409
Labor X K.PNB	0.087465878
Distance.to.river X K.PNB	0.090166744
P.PNB X K.PNB	0.119349702

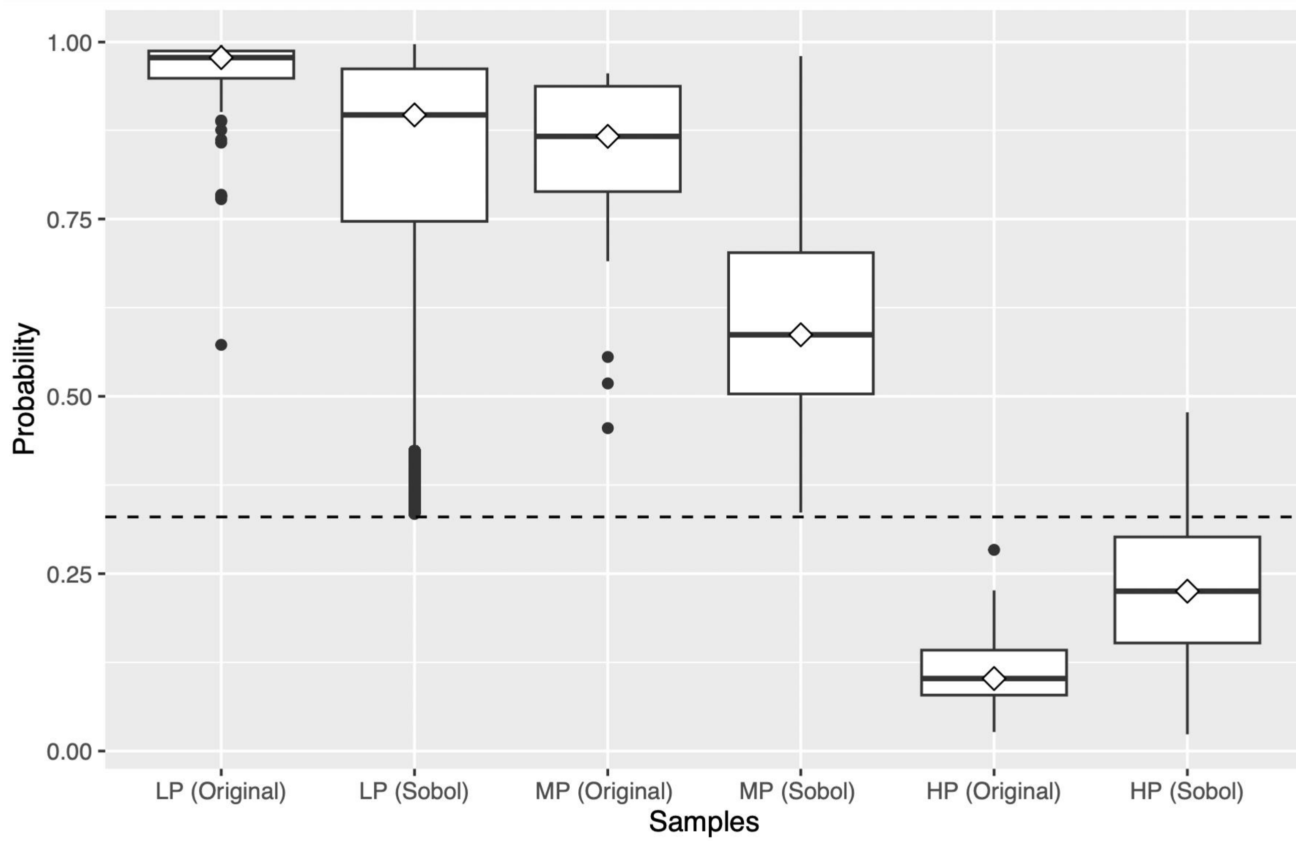

Figure 4-3 Boxplots showing the median probabilities of each productivity class from the original sample against the Sobol sample Uncertainty Analysis. LP denotes the low-productivity class, MP denotes the mi-productivity class, and HP denotes the high-productivity class.

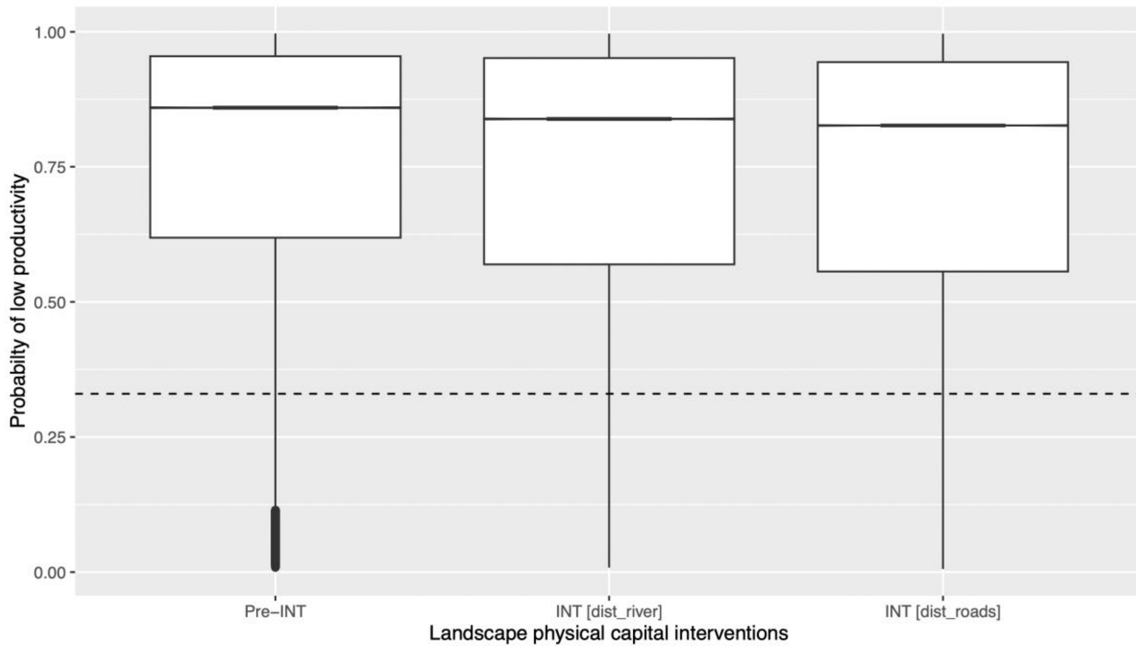

Figure 4-4. Boxplots showing the distribution of the overall Sobol sample before and after each landscape interventions. The dashed horizontal line represents the behavioral threshold ($p_{low\text{-}yields}$ <0.33).

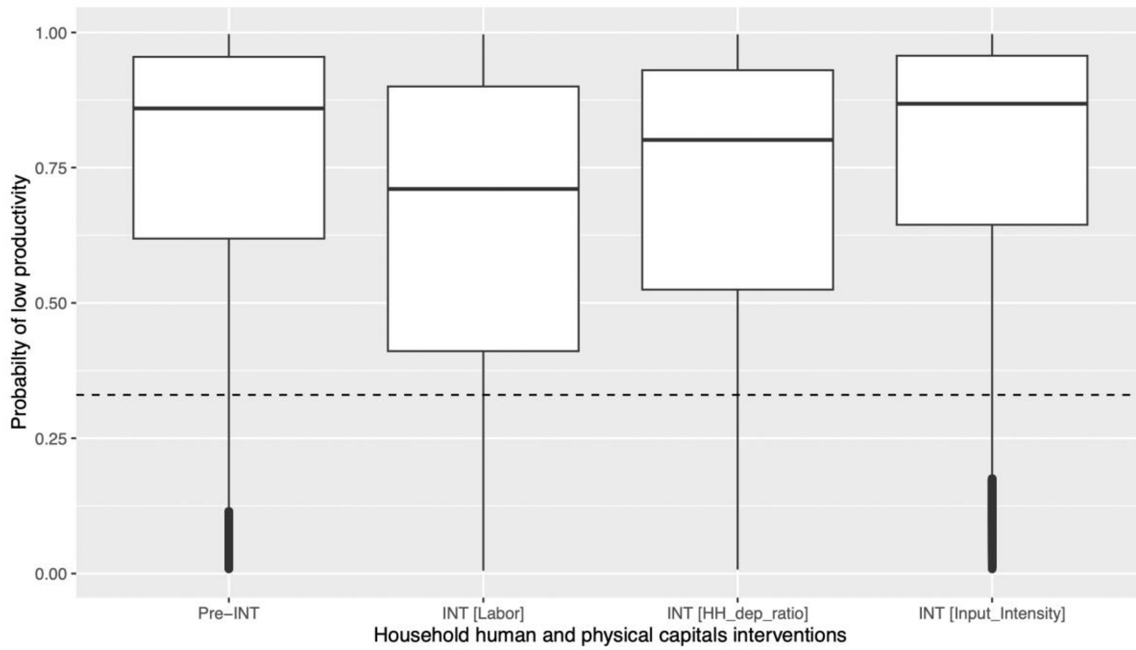

Figure 4-5. Boxplots showing the distribution of the overall Sobol sample before and after each household intervention. The dashed horizontal line represents the behavioral threshold ($p_{low\text{-}yields}$ <0.33).

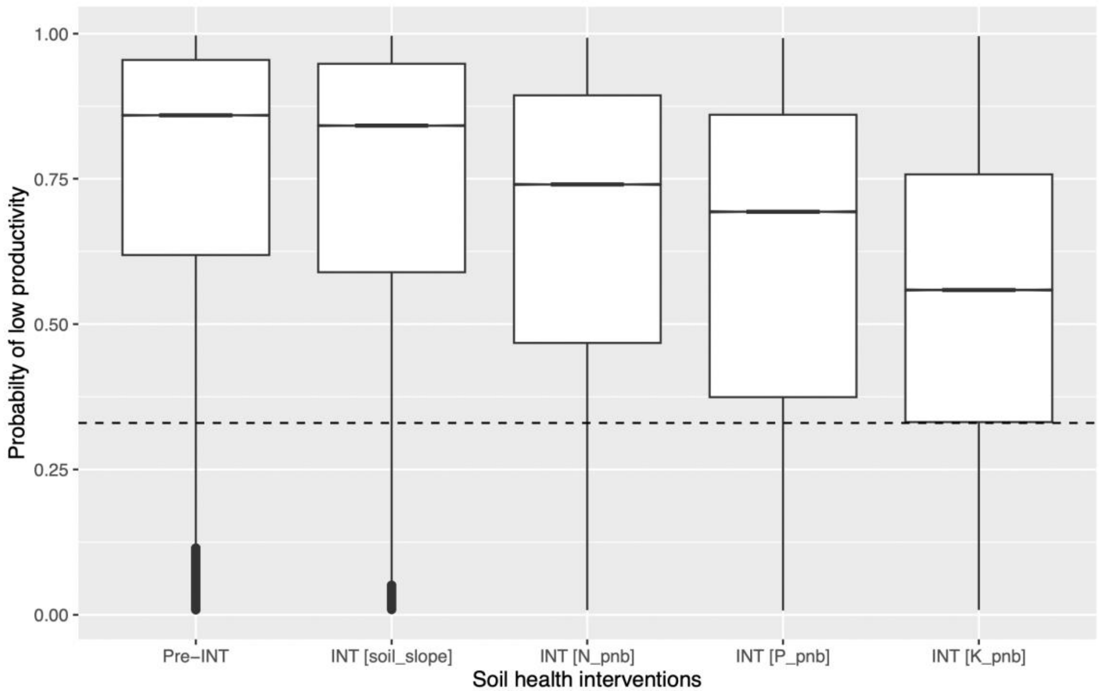

Figure 4-6. Boxplots showing the distribution of the overall Sobol sample before and after each soil health interventions. The dashed horizontal line represents the behavioral threshold ($p_{low\text{-}yields}$ <0.33).

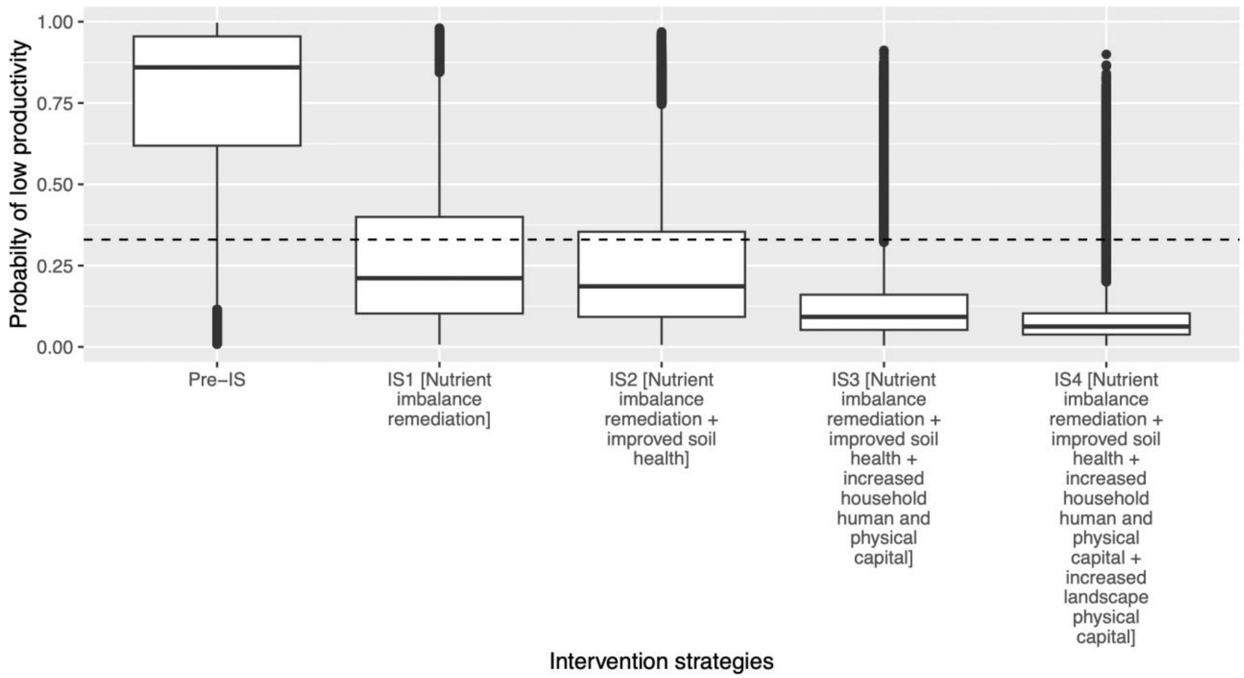

Figure 4-7. Boxplots showing the distribution of the overall Sobol sample pre - intervention and after intervention strategies**.** The dashed horizontal line represents the behavioral threshold ($p_{low\text{-}yields}$ <0.33).

Table 4-3. Final probability outcomes of the model before and after each intervention strategy.

Interventions strategies	Model probability outcomes		
	Low productivity	Mid productivity	High productivity
Pre-intervention	86%	11%	3%
IS1: Soil nutrient imbalance remediation	24%	50%	26%
IS2: Soil nutrient imbalance remediation + improved soil health	19%	56%	25%
IS3: Soil nutrient imbalance remediation + improved soil health + increase in household human and physical capital	2%	85%	13%
IS4: Soil nutrient imbalance remediation + improved soil health + increase in household human and physical capital + improved physical infrastructure capital	0%	89%	11%

CHAPTER 5
GENERAL CONCLUSIONS

This research investigates the linkages and dynamics of social-ecological factors and farm multifunctionality metrics of agricultural systems at the farm scale, specifically the case of smallholder systems of Sub-Saharan Africa, using alternative modeling approaches that aim to capture hidden/non-linear relationships in these systems.

In Chapter 1, a modeling framework for linking social-ecological factors of farming to productivity metrics was proposed. Suitable social-ecological indicators of livelihood capitals, management intensity decision-making were established, along with suitable indicators of farm multifunctionality integrating crop yield and soil health indicators (i.e., soil carbon, water storage and nutrient supply) were also established.

Chapter 2 focused on (1) Assessing synergies and trade-offs between ecosystem services clusters, capital levels, farm input intensity and soil management intensity (2) Identifying patterns of dissimilar ecosystem service levels of farms (i.e., ecosystem service clusters) and dissimilar livelihood/intensification strategies (i.e., livelihood strategy clusters) towards Linking ecosystem service levels with livelihood/intensification strategies. These analyses offer new insights on drivers of productivity in smallholder systems of SSA, while capturing the variability in farm productivity. This analysis approach was able to confirm consistent synergistic relationships between higher levels of human derived capitals and combined input intensity and linking high human-derived capital and combined input intensity to higher productivity across case studies.

In Chapter 3, three ML models (i.e., logistic regression as baseline, random forest, and XGBoost) were tested on their ability to accurately predict productivity classes based on maize yield using livelihood capitals, management intensity, soil health indicators, and guided by the following benchmark for productivity levels in smallholder systems in SSA: low productivity when maize yield is below 1.5 t/ha, mid-productivity when maize yield between 1.5 and 2.5 t/ha, and high productivity when maize yield is above 2.5 t/ha. Overall, the two tree-ensemble ML models were able to accurately identify patterns of social -ecological and soil health indicators explaining the three-productivity levels detected based on maize yield variability.

In Chapter 4, the best performing model (i.e., XGBoost) was selected for identify key factors driving the model by comparing a low dimensional factor importance method (i.e., entropy/information gain) to a high dimensional factor importance method (variance-based method) of feature importance evaluation. A robust Global Sensitivity and Uncertainty Analysis (GSAU) was performed on the model's features and outputs to test for uncertainties, and Monte Carlo filtering was used to inform on intervention strategies on explanatory variable able to boost the productivity of smallholders. Results showed that soil health indicators, interactively with household labor, and landscape level physical capitals such access to roads and access to water, are key factors to interven for boost all low-productivity farms of the region to higher levels of productivity.

The work accomplished in this dissertation were able to successfully answer the research questions posed: 1) What are the synergistic/ trade-offs relationships between social-ecological factors (i.e., livelihood capitals, management intensity) and ecosystem services (i.e., crop yield, soil carbon and water storage, soil nutrient supply) in smallholder systems of SSA? 2) Can social-ecological factors (i.e., livelihood capitals, management intensity), and soil health indicators (i.e., soil carbon and water storage, soil nutrient supply) be good predictors of productivity (i.e., crop yield) ? If so, how can these relationships be modeled to inform the design of intervention strategies able to boost farm productivity in smallholder systems of SSA?

The modeling framework proposed provided methodological approaches able to capture hidden nonlinear relationships often not tested for between metrics of livelihood/intensification strategies and biophysical metrics of farm multifunctionality.

Challenges and limitations in accomplishing this work involve data curation and processing. Reported survey data come with bias, outliers, and missing values. While missing value imputation help achieve complete datasets for analyses, there are no way to identify and solve response bias, also removing outliers may remove natural variability of the data. Moreover, field and landscape level/ high resolution spatial data products for biophysical data are rare. The first case study (SAGCOT) is data rich as various landscape level data were already made available from previous data collection efforts. However, in the case of the Upper Ewaso Ngiro basin, more data could be collected to further integrate landscape level natural capital indicators in our models.

In the case of chapter 2, we decided to use a classification model able to account for yield variability instead of using yield values as outputs. Although this approach helps understand yield variability, a regression model would have supported the benchmarking of productivity levels better.

Future directions of this work involve exploring both modeling approaches of the framework (i.e.,unsupervised clustering, and supervised machine learning) with increased dimensionality, to account for additional productivity and social-ecological metrics that could be relevant for our research questions. In both cases in chapter 1 and 2 , it will increase the chance of finding new relevant patterns of relationships driving smallholder agricultural systems. Moreover, in this work we did not incorporate time as a factor in any part our analysis, as none of the modeling approaches facilitates the use time series data, nor our data provided any notion of time that could inform on how long intensification strategies have been adopted on a farm for example, or previous levels of ecosystem services over the years. Exploring the coupling of data-driven models such as the one developed in this dissertation with agent-based modeling to generate time series outputs towards helping us better understanding changes in social-ecological dynamics of agricultural systems is a good way forward relevant to this research topic.

APPENDIX
A QUANTIFICATION OF ECOSYTEM SERVICES INDICATORS USING SOIL FUNCTIONS

1. Soil carbon storage ecosystem service indicator

We use a standardized soil carbon deficit indicator. It is defined as the ability of the soil to store additional carbon and is estimated as the difference between the maximum potential carbon storage that can be associated with the soil and the current amount of carbon found in the soil. The less the deficit, the higher the ability of the soil to store additional carbon.

The following functions are used to calculate the soil carbon deficit indicator

Carbon capacity is calculated in function of soil type. The values are then standardized for comparison. The soil carbon capacity was calculated as follows {Palm:2005un}:

$$soil\ C\ capacity\ (\%) = \frac{acidified\ carbon}{e^{1.333+0.00994\cdot clay\ +\ 0.00699\cdot silt\ -\ 0.156(0.923\cdot pH-0.6)}} \cdot \left(\frac{10}{7}\right)^{-0.58} \cdot 100$$

Acidified carbon represents the carbon in the topsoil. It is calculated as the % by weight from acid treated samples in the SAGCOT case study. In the case of the Upper Ewaso Ngiro case study, the total inorganic carbon is calculated subtracting organic carbon from total carbon. *clay* and *silt* are the fractions of clay and silt (% by volume) and *pH* is the pH value of the soil. The subtraction of 0.6 is a correction for using soil pH as opposed to pH_{KCl} (Sanchez, 2019). The denominator in the first factor in the formula represents reference soil carbon (at 7 cm), and the second factor in the formula represents correction factor for the average depth of our samples (10 cm).

The soil carbon deficit indicator is then assigned as a value between 0 and 1. Carbon capacity < 50% gets assigned 0, carbon capacity > 80% gets assigned 1, and everything in between gets a value between 0 and 1 based on linear interpolation from 50% to 80%.

We then standardized the carbon deficit indicator to reflect carbon storage as an ecosystem services, with the "less is better standardization function. This standardization approach reflects the less the carbon deficit indicator, the higher the carbon storage ecosystem service. The "less is better" standardization function

$$Soil\ carbon\ \ storage\ indicator = \frac{\text{Soil carbon deficit indicator} - \ \max(\text{Soil carbon deficit indicator})}{\max(\text{Soil carbon deficit indicator}) - \min\ (\text{Soil carbon deficit indicator})}$$

2. Soil water storage indicator

The Soil water storage indicator was determined using clay content, bulk density, and depth to bedrock properties. These properties were standardized using fuzzy logic membership functions. Clay content and depth to rock properties were standardized using the "more is better" function, Bulk density is standardized using the optimal function with an optimal value of 130 kg/m^3. The water storage indicator score was calculated as the arithmetic mean of the standardized properties.The calculations of the soil water storage indicator are as follows:

$$Clay\ content\ indicator = \frac{\text{clay content} - \ \max(clay\ content\ )}{\max(clay\ content) - \min\ (clay\ content)}$$

$$Depth\ to\ rock\ indicator\ = \frac{\text{depth to rock} - \ \text{max(depth to rock)}}{\text{max(depth to rock)} - \text{min (depth to rock)}}$$

$$Bulk\ density\ indicator\ = \frac{(Bulk\ density - 130) - (130 - \text{min}\ (Bulk\ density)}{(\text{max}(Bulk\ density) - 130) - (\text{min}(Bulk\ density) - 130)}$$

$$\text{Soil water storage indicator}\ = \frac{Clay\ content\ indicator + Depth\ to\ rock\ indicator + Bulk\ density\ indictator}{3}$$

3. Soil nutrient balance in the SAGCOT

In the case of SAGCOT, a soil nutrient budget indicator was determined using a total partial nutrient budget indicator based on the nutrient (nitrogen, phosphorous, and potassium) removal and addition calculations at the farm level. Nutrient addition and removal were calculated as follows:

Nutrient addition

Nutrients added to the field by organic and inorganic fertilizer were calculated based on the reported addition of fertilizer in the survey. Values for the nutrients added by fertilizers are shown in Table 1. For carbon from organic fertilizer (animal manure), the value of 47% of weight was used.

Table A-1. Value for nutrients added by each type of fertilizer.

Inputs	N	P	K	P205	K20
Animal Manure	0.012	0.002	0.007	0.005	0.009
Di-Ammonium Phosphate (DAP)	0.18	0.201	0	0.46	0
Urea	0.46	0	0	0	0
Triple Super Phosphate (TSP)	0	0.201	0	0.46	0
Calcium Ammonium Nitrate (CAN)	0.27	0	0	0	0
Sulphate of Ammonium (SA)	0.27	0	0	0	0
Nitrogen Phosphate Potassium (NPK)	0.15	0.065	0.124	0.15	0.15
Rock Phosphate (MRP) [Minjingu]	0	0.144	0	0.33	0

If crop residue was added or left on the fields, the added carbon was calculated as above (47% of the weight). For the addition of nitrogen, phosphorous and potassium from this source, this was only calculated if the crop residue was added from an external source. For crop residue left on the fields, these nutrients are not an addition, they are simply being recycled back into the system. Carbon on the other hand is an addition by the creation of biomass. Since the “application” of crop residue was only reported as a yes/no question in the survey, we used the yield and harvest indices to estimate the amount of crop residue from different crops (Smil, 1999)(Table 2).

Table A- 2. Harvest indices; crop residue as the ratio of the harvested product to total dry shoot matter (Smil, 1999).

Crop residue group	**Harvest index**
Cereals	0.4
Sugar crops	0.56
Roots	0.4
Vegetables	0.38
Fruits	0.38
Legumes	0.49
Oil crops	0.52
Onions	0.83
Other crops	0.28

Nutrient removal

A number of reference tables were used in the calculation of nutrient removal. The table below shows the nutrient removal factors we used (Lesschen et al., 2004) to calculate nitrogen, phosphorous and potassium removal from fields for all harvested products such as cereals, legumes, fruits, and vegetables. Harvest and yield of all crops were calculated at the field, farm, and landscape scales, and were summarized in Figure 1. Carbon removal was calculated taking carbon as 47% of the weight of the harvested product.

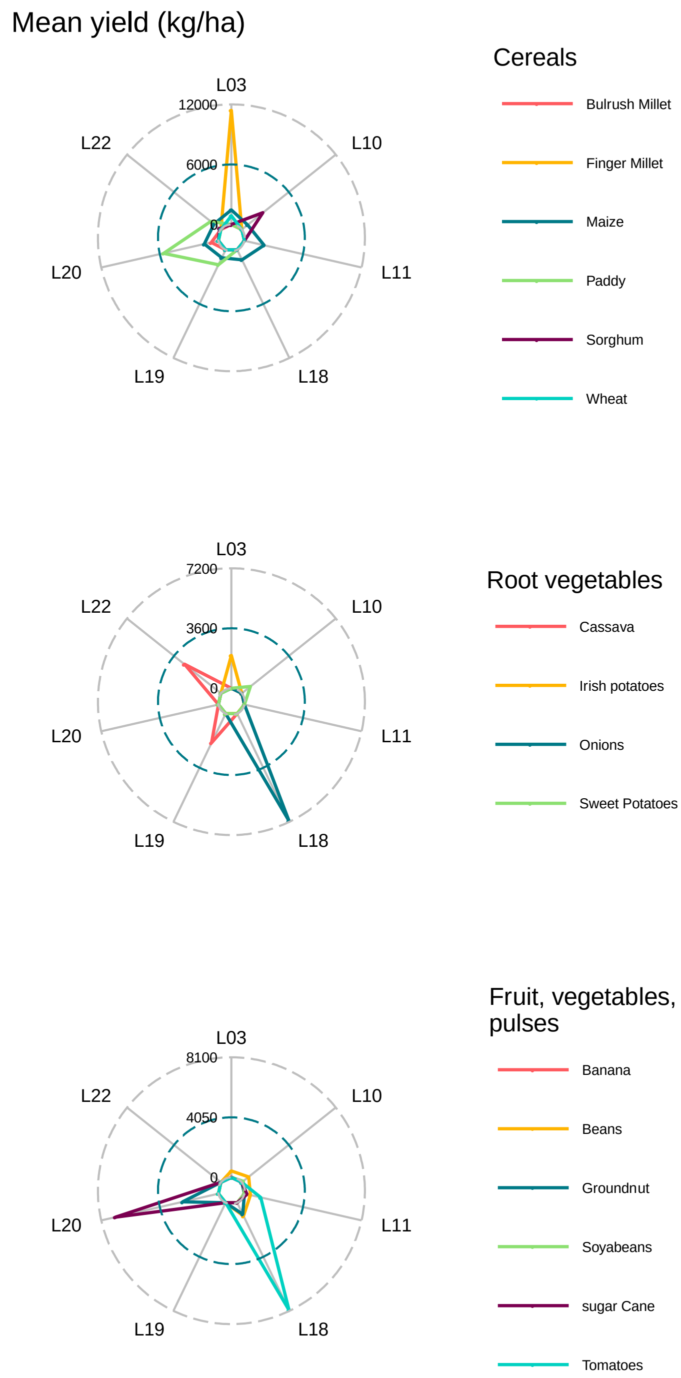

Figure A-1.Mean yield (kg/ha) of various crops in different landscapes.

Table A-3. Nutrient removal of nitrogen (N), phosphorous (P) and potassium (K) in kg/metric ton for harvested product, as well as crop residue (Lesschen et al., 2004).

Crops	N (harvested)	P (harvested)	K (harvested)	N (crop residue)	P (crop residue)	K (crop residue)
Banana	1.2	0.3	4.5	1.6	0.3	11.9
Barley	15.5	2.8	6	7	1	21
Cassava	4.2	0.5	4.3	4.6	0.9	1.4
Cereals other	16.7	4.4	4.8	10.9	2.3	38.6
Citrus	1.8	0.2	2.3	0.6	0.2	4.4
Cocoa	40	8.5	19.3	19.9	4.7	33.3
Coconut	61	7.2	9.8	27	5.7	25.3
Coffee	35	2.6	16.8	4.3	3.8	9.3
Cotton	18.7	9.7	9	13.9	6	29.8
Fibers	5	0.4	6	2.1	0.7	9
Fruits other	2	0.2	2	1.8	0.2	4.9
Groundnut	37.2	6	8.2	15.9	2.4	14.9
Maize	16.8	4.1	4.8	9.7	1.9	21.4
Millet	19.2	6	5.4	20.4	4	59.8
Oil crops other	2.6	0.5	4.4	0.3	0.6	5.4
Oil palm	2.9	0.7	4.1	3.7	0.6	3.3
Plantain	0.7	0.1	3.4	1.2	0.3	6.4
Potato	4.4	1.3	6.9	2.3	0.7	4.5
Pulses	20	3.4	11.1	10.4	1	13.1
Rice	11.6	3.4	3.4	11.3	2.3	35.8
Roots other	4.6	0.3	2.9	1.9	0.5	3.1
Rubber	6.9	1.2	4.6	1	0.2	4
Sesame	30	6.1	6.8	15	5.4	21.1
Sorghum	14.5	5.5	3.8	10.8	4.6	29.2
Soybean	62.1	10.9	20	17.6	3	14.4
Sugar cane	0.6	0.2	1.2	0.3	0.3	0.3
Sunflower	24	3.5	5.5	23	3.2	41.3
Sweet potato	4.8	0.8	7.3	2.1	1.2	3.3
Tea	35	3.8	13.4	0.1	0	0
Tobacco	56	8.2	72.7	0.1	0	0.2
Vegetables	9	0.9	2.6	3.2	1.4	7.8
Wheat	22.3	4.3	5.8	4.3	1.8	26.7

Nutrient budget

Nutrient budgets for each nutrient (nitrogen, phosphorous and potassium) were calculated for each field in both kg and kg/ha, and were aggregated at the farm level. Positive values indicate a net addition of nutrients to the field, negative values a net removal.

We then standardized the final partial nutrient budget indicator to reflect nutrient deficits observed for all observations in the data, using the “less is better” fuzzy logic membership function, meaning less nutrient deficit is better.The standardization of the final partial nutrient budget indicator is calculated as follows:

Since the values obtained for nutrient budgets were negative, they were standardized using the “less is better” standardization function to calculate a nutrient, phosphorous and potassium budget indicator.

The final nutrient balance indicator was calculated where the three values for the standardized nitrogen, phosphorous and potassium indicators, are divided by 3 to give a value from 0 to 1. This reflects overall less nutrient deficit is better.

$$Nutrient\ balance\ indicator = \frac{\begin{array}{c}\text{Nitrogen partial nutrient budget indicator} + \\ \text{Potasisum partial nutrient budget indicator} + \text{Phosphorous partial nutrient budget indicator}\end{array}}{3}$$

4. Soil nutrient supply indicator in the Upper Ewaso Ng’iro

In the case of the Upper Ewaso Ng’iro, a soil nutrient supply indicator was determined using total Nitrogen (Total N) along with soil Cation Exchange Capacity (CEC). Each property was standardized using its respective fuzzy logic membership function. Nitrogen partial budget or total Nitrogen was standardized using the “more is better” function, CEC was also standardized using the more is better function. The nutrient supply indicator score was calculated using the arithmetic mean of the standardized properties. The calculations of the soil water storage indicator are as follows:

$$Nitrogen\ indicator = \frac{\text{Total N} - \max(\text{Total N})}{\max(\text{Total N}) - \min(\text{Total N})}$$

$$CEC\ indicator = \frac{\text{CEC} - \max(\text{CEC})}{\max(\text{CEC}) - \min(\text{CEC})}$$

$$Nutrient\ supply\ indicator = \frac{Nitrogen\ indicator + CEC\ indicator}{2}$$

B DATA AND CODE REPOSITORIES

CHAPTER 2 CODES

https://gitlab.com/tvenort/siil-environmental-indicators

https://github.com/tvenort/Dissertation_Chapter2

CHAPTER 3 CODES

https://github.com/tvenort/Dissertation_Chapter3-4

CHAPTER 4 CODES

https://github.com/tvenort/Dissertation_Chapter3-4

REFERENCES

Agarwala, M., Atkinson, G., Fry, B. P., Homewood, K., Mourato, S., Rowcliffe, J. M., Wallace, G., & Milner-Gulland, E. J. (2014). Assessing the relationship between human well-being and ecosystem services: A review of frameworks. In *Conservation and Society* (Vol. 12, Issue 4). https://doi.org/10.4103/0972-4923.155592

Aryal, K., Maraseni, T., & Apan, A. (2022). How much do we know about trade-offs in ecosystem services? A systematic review of empirical research observations. In *Science of the Total Environment* (Vol. 806). https://doi.org/10.1016/j.scitotenv.2021.151229

Ashley, C., & Carney, D. (1999). Sustainable livelihoods: Lessons from early experience. *Development*.

Banerjee, H., Goswami, R., Chakraborty, S., Dutta, S., Majumdar, K., Satyanarayana, T., Jat, M. L., & Zingore, S. (2014). Understanding biophysical and socio-economic determinants of maize (Zea mays L.) yield variability in eastern India. *NJAS - Wageningen Journal of Life Sciences*, *70*. https://doi.org/10.1016/j.njas.2014.08.001

Barré, P., Angers, D., Basile-Doelsch, I., Bispo, A., Cécillon, L., Chenu, C., Chevallier, T., Derrien, D., Eglin, T., & Pellerin, S. (2017). Ideas and perspectives: Can we use the soil carbon saturation deficit to quantitatively assess the soil carbon storage potential, or should we explore other strategies? *Biogeosciences Discussions*, *September*.

Bennett, E. M., Peterson, G. D., & Gordon, L. J. (2009). Understanding relationships among multiple ecosystem services. *Ecology Letters*, *12*(12). https://doi.org/10.1111/j.1461-0248.2009.01387.x

Biggs, R., de Vos, A., Preiser, R., Clements, H., Maciejewski, K., & Schlüter, M. (2021). The routledge handbook of research methods for social-ecological systems. In *The Routledge Handbook of Research Methods for Social-Ecological Systems*. https://doi.org/10.4324/9781003021339

Bradshaw, C. J. A., & Di Minin, E. (2019). Socio-economic predictors of environmental performance among African nations. *Scientific Reports*. https://doi.org/10.1038/s41598-019-45762-3

Breiman, L. (2001). (impo)Random forests(book). *Machine Learning*.

Carmona-Cabrero, A. (2022). *Coupling global sensitivity analysis with machine learning and agent-based models to disentangle complex system drivers: application to human refugee migration*.

Carmona-Cabrero, A., Muñoz-Carpena, R., & Muneepeerakul, R. (2023). Determining factor importance in scarce and noisy data using tree ensembles and global sensitivity analysis . *Under Review*.

Carney, D. (2003). Sustainable Livelihoods Approaches : Progress and Possibilities for Change. *Secretary*, *2008*. https://doi.org/ISBN 1 86192 491 7

Cassman, K. G., & Grassini, P. (2020). A global perspective on sustainable intensification research. *Nature Sustainability*, *3*(4), 262–268. https://doi.org/10.1038/s41893-020-0507-8

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939785

Chianu, J. N., Chianu, J. N., & Mairura, F. (2012). Mineral fertilizers in the farming systems of sub-Saharan Africa. A review. In *Agronomy for Sustainable Development* (Vol. 32, Issue 2). https://doi.org/10.1007/s13593-011-0050-0

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11). https://doi.org/10.1088/1748-9326/aae159

Crouzat, E., Mouchet, M., Turkelboom, F., Byczek, C., Meersmans, J., Berger, F., Verkerk, P. J., & Lavorel, S. (2015). Assessing bundles of ecosystem services from regional to landscape scale: Insights from the French Alps. *Journal of Applied Ecology*, *52*(5), 1145–1155. https://doi.org/10.1111/1365-2664.12502

Dangeti, P. (2017). Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. In *Packt Publishing*.

Dittrich, A., Seppelt, R., Václavík, T., & Cord, A. F. (2017). Integrating ecosystem service bundles and socio-environmental conditions – A national scale analysis from Germany. In *Ecosystem Services* (Vol. 28). https://doi.org/10.1016/j.ecoser.2017.08.007

Dutta, S., Chakraborty, S., Goswami, R., Banerjee, H., Majumdar, K., Li, B., & Jat, M. L. (2020). Maize yield in smallholder agriculture system-An approach integrating socioeconomic and crop management factors. *PLoS ONE*, *15*(2). https://doi.org/10.1371/journal.pone.0229100

Eissler, S., Ader, D., Huot, S., Brown, S., Bates, R., & Gill, T. (2021). Wild gardening as a sustainable intensification strategy in northwest Cambodian smallholder systems. *Journal of Agriculture, Food Systems, and Community Development*. https://doi.org/10.5304/jafscd.2021.103.006

Erenstein, O., Jaleta, M., Sonder, K., Mottaleb, K., & Prasanna, B. M. (2022). Global maize production, consumption and trade: trends and R&D implications. In *Food Security* (Vol. 14, Issue 5). https://doi.org/10.1007/s12571-022-01288-7

Fisher, B., Turner, K., Zylstra, M., Brouwer, R., de Groot, R., Farber, S., Ferraro, P., Green, R., Hadley, D., Harlow, J., Jefferiss, P., Kirkby, C., Morling, P., Mowatt, S., Naidoo, R., Paavola, J., Strassburg, B., Yu, D., & Balmford, A. (2008). Ecosystem services and economic theory: Integration for policy-relevant research. *Ecological Applications*, *18*(8). https://doi.org/10.1890/07-1537.1

Fisher, J. A., Patenaude, G., Giri, K., Lewis, K., Meir, P., Pinho, P., Rounsevell, M. D. A., & Williams, M. (2014). Understanding the relationships between ecosystem services and poverty alleviation: A conceptual framework. *Ecosystem Services*, *7*. https://doi.org/10.1016/j.ecoser.2013.08.002

Folberth, C., Yang, H., Gaiser, T., Abbaspour, K. C., & Schulin, R. (2013). Modeling maize yield responses to improvement in nutrient, water and cultivar inputs in sub-Saharan Africa. *Agricultural Systems*, *119*. https://doi.org/10.1016/j.agsy.2013.04.002

Fortmann, L. (2011). Sustainable intensification: Increasing productivity in African food and agricultural systems. In *Experimental Agriculture* (Vol. 48, Issue 01).

Garbach, K., Milder, J. C., DeClerck, F. A. J., Montenegro de Wit, M., Driscoll, L., & Gemmill-Herren, B. (2017). Examining multi-functionality for crop yield and ecosystem services in five systems of agroecological intensification. *International Journal of Agricultural Sustainability*, *15*(1). https://doi.org/10.1080/14735903.2016.1174810

Garcia Alberto. (2020). *The Environmental Impacts of Agricultural Intensification*.

Godfray, H. C. J. (2015). The debate over sustainable intensification. *Food Security*, *7*(2). https://doi.org/10.1007/s12571-015-0424-2

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. In *Science*. https://doi.org/10.1126/science.1185383

Gómez-Baggethun, E., Barton, D. N., Berry, P., Dunford, R., & Harrison, P. A. (2018). Concepts and Methods in Ecosystem Services Valuation. In *Routledge Handbook of Ecosystem Services*. https://doi.org/10.4324/9781315775302-9

Gopel, J., Schungel, J., Stuch, B., & Schaldach, R. (2020). Assessing the effects of agricultural intensification on natural habitats and biodiversity in Southern Amazonia. *PLoS ONE*, *15*(11 November). https://doi.org/10.1371/journal.pone.0225914

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian Bayesian state estimation. *IEE Proceedings, Part F: Radar and Signal Processing*, *140*(2). https://doi.org/10.1049/ip-f-2.1993.0015

Goulding, K., Murrell, T. S., Mikkelsen, R. L., Rosolem, C., Johnston, J., Wang, H., & Alfaro, M. A. (2020). Outputs: Potassium losses from agricultural systems. In *Improving Potassium Recommendations for Agricultural Crops*. https://doi.org/10.1007/978-3-030-59197-7_3

Hammond, J., van Wijk, M., Teufel, N., Mekonnen, K., & Thorne, P. (2021). Assessing smallholder sustainable intensification in the Ethiopian highlands. *Agricultural Systems*, *194*. https://doi.org/10.1016/j.agsy.2021.103266

Hastie, T., Tibshirani, R., James, G., & Witten, D. (2006). An Introduction to Statistical Learning, Springer Texts. In *Springer Texts* (Vol. 102).

Henao, J., & Baanante, C. (1999). Estimating Rates of Nutrient Depletion in Soils of Agricultural Lands of Africa. In *International Fertilizer Development Center*.

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J. M., Chamberlin, J., Silatsa, F. B. T., … Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-85639-y

Holden, S. T. (2018). Fertilizer and sustainable intensification in Sub-Saharan Africa. In *Global Food Security*. https://doi.org/10.1016/j.gfs.2018.07.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. Springer.

Jayne, T. S., Chamberlin, J., & Benfica, R. (2018). Africa's Unfolding Economic Transformation. In *Journal of Development Studies*. https://doi.org/10.1080/00220388.2018.1430774

Jayne, T. S., & Sanchez, P. A. (2021). Agricultural productivity must improve in sub-Saharan Africa. In *Science* (Vol. 372, Issue 6546). https://doi.org/10.1126/science.abf5413

Jones, L., Norton, L., Austin, Z., Browne, A. L., Donovan, D., Emmett, B. A., Grabowski, Z. J., Howard, D. C., Jones, J. P. G., Kenter, J. O., Manley, W., Morris, C., Robinson, D. A., Short, C., Siriwardena, G. M., Stevens, C. J., Storkey, J., Waters, R. D., & Willis, G. F. (2016). Stocks and flows of natural and human-derived capital in ecosystem services. *Land Use Policy*, *52*, 151–162. https://doi.org/10.1016/j.landusepol.2015.12.014

Kassambara, A. (2017). Multivariate Analysis I: Practical Guide To Cluster Analysis in R. Unsupervised Machine Learning. *Taylor & Francis Group*, 188.

Khoshgoftaar, T. M., Golawala, M., & Van Hulse, J. (2007). An empirical study of learning from imbalanced data using random forest. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2*. https://doi.org/10.1109/ICTAI.2007.46

Kitagawa, G. (1996). Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, *5*(1). https://doi.org/10.2307/1390750

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1–3). https://doi.org/10.1016/S0925-2312(98)00030-7

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*. https://doi.org/10.1016/j.neunet.2012.09.018

Kremen, C. (2020). Ecological intensification and diversification approaches to maintain biodiversity, ecosystem services and food production in a changing world. In *Emerging Topics in Life Sciences* (Vol. 4, Issue 2). https://doi.org/10.1042/ETLS20190205

Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). Handbook of Monte Carlo Methods. In *Handbook of Monte Carlo Methods*. https://doi.org/10.1002/9781118014967

Kumar, M., Kaul, S., Sethi, S., & Jain, S. (2023). Framework to Impute Missing Values in Datasets. *Lecture Notes in Electrical Engineering*, *968*. https://doi.org/10.1007/978-981-19-7346-8_17

Kuyah, S., Sileshi, G. W., Nkurunziza, L., Chirinda, N., Ndayisaba, P. C., Dimobe, K., & Öborn, I. (2021). Innovative agronomic practices for sustainable intensification in sub-Saharan Africa. A review. In *Agronomy for Sustainable Development* (Vol. 41, Issue 2). https://doi.org/10.1007/s13593-021-00673-4

Lee, H., & Lautenbach, S. (2016). A quantitative review of relationships between ecosystem services. In *Ecological Indicators* (Vol. 66). https://doi.org/10.1016/j.ecolind.2016.02.004

Lesschen, J. P., Stoorvogel, J. J., & Smalling, E. M. A. (2004). *Scaling soil nutrient balances. Enabling mesolevel applications for African realities*.

Levin, S., Xepapadeas, T., Crépin, A. S., Norberg, J., De Zeeuw, A., Folke, C., Hughes, T., Arrow, K., Barrett, S., Daily, G., Ehrlich, P., Kautsky, N., Mäler, K. G., Polasky, S., Troell, M., Vincent, J. R., & Walker, B. (2013). Social-ecological systems as complex adaptive systems: Modeling and policy implications. *Environment and Development Economics*, *18*(2). https://doi.org/10.1017/S1355770X12000460

Li, S., Zhao, Y., Xiao, W., Yellishetty, M., & Yang, D. (2022). Identifying ecosystem service bundles and the spatiotemporal characteristics of trade-offs and synergies in coal mining areas with a high groundwater table. *Science of the Total Environment*, *807*. https://doi.org/10.1016/j.scitotenv.2021.151036

Liang, J. , L. H. , L. N. , Y. Q. , & L. L. (2023). Analysis and Prediction of the Impact of Socio-Economic and Meteorological Factors on Rapeseed Yield Based on Machine Learning. *Agronomy*, *13*(7).

Lotfi, F. H., & Fallahnejad, R. (2010). Imprecise shannon's entropy and multi attribute decision making. *Entropy*, *12*(1). https://doi.org/10.3390/e12010053

Lowder, S. K., Skoet, J., & Raney, T. (2016). The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide. *World Development*, *87*. https://doi.org/10.1016/j.worlddev.2015.10.041

Lowder, S. K., Skoet, J., & Singh, S. (2014). What do we really know about the number and distribution of farms and family farms in the world ? *ESA Working Paper*, *14–02*(14).

Mendelsohn, J., Robertson Tony, & Jarvis Alice. (2014). *Tanzania: The measure of a land*.

Morizet-Davis, J., Marting Vidaurre, N. A., Reinmuth, E., Rezaei-Chiyaneh, E., Schlecht, V., Schmidt, S., Singh, K., Vargas-Carpintero, R., Wagner, M., & von Cossel, M. (2023). Ecosystem Services at the Farm Level—Overview, Synergies, Trade-Offs, and Stakeholder Analysis. In *Global Challenges*. https://doi.org/10.1002/gch2.202200225

Mouratiadou, I., Latka, C., van der Hilst, F., Müller, C., Berges, R., Bodirsky, B. L., Ewert, F., Faye, B., Heckelei, T., Hoffmann, M., Lehtonen, H., Lorite, I. J., Nendel, C., Palosuo, T., Rodríguez, A., Rötter, R. P., Ruiz-Ramos, M., Stella, T., Webber, H., & Wicke, B. (2021). Quantifying sustainable intensification of agriculture: The contribution of metrics and modelling. *Ecological Indicators*, *129*. https://doi.org/10.1016/j.ecolind.2021.107870

Musumba, M., Grabowski, P., Palm, C., & Snapp, S. (2021). Guide for the Sustainable Intensification Assessment Framework. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3906994

Musumba, M., Palm, C., Grabowski, P., & Snapp, S. S. (2017). *A Framework for Selecting and Analyzing Indicators of Sustainable intensification*. 1–46. http://www.k-state.edu/siil/documents/docs_siframework/Guide for SI Assessment Framework - 10.24.17.pdf

Nin Pratt, A. (2015). Agricultural Intensification in Africa: A Regional Analysis. *SSRN Electronic Journal*, *March*. https://doi.org/10.2139/ssrn.2591567

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*(5). https://doi.org/10.1007/s11135-006-9018-6

Paruelo, J. M., & Sierra, M. (2023). Sustainable intensification and ecosystem services: how to connect them in agricultural systems of southern South America. *Journal of Environmental Studies and Sciences*, *13*(1). https://doi.org/10.1007/s13412-022-00791-9

Power, A. G. (2010). Ecosystem services and agriculture: Tradeoffs and synergies. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 365, Issue 1554). https://doi.org/10.1098/rstb.2010.0143

Pretty, J. (2008). Agricultural sustainability: Concepts, principles and evidence. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 363, Issue 1491). https://doi.org/10.1098/rstb.2007.2163

Pretty, J., & Bharucha, Z. P. (2014). Sustainable intensification in agricultural systems. In *Annals of Botany*. https://doi.org/10.1093/aob/mcu205

Pretty, J., Toulmin, C., & Williams, S. (2011). Sustainable intensification in African agriculture. *International Journal of Agricultural Sustainability*. https://doi.org/10.3763/ijas.2010.0583

Pulkkinen, T., & Nurmi, P. (2012). AWESOM: Automatic discrete partitioning of indoor spaces for wifi fingerprinting. *Lecture Notes in Computer Science*

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *7319 LNCS*. https://doi.org/10.1007/978-3-642-31205-2_17

Qiu, J., Carpenter, S. R., Booth, E. G., Motew, M., Zipper, S. C., Kucharik, C. J., Loheide, S. P., & Turner, M. G. (2018). Understanding relationships among ecosystem services across spatial scales and over time. *Environmental Research Letters*, *13*(5). https://doi.org/10.1088/1748-9326/aabb87

Qiu, J., Queiroz, C., Bennett, E. M., Cord, A. F., Crouzat, E., Lavorel, S., Maes, J., Meacham, M., Norström, A. V., Peterson, G. D., Seppelt, R., & Turner, M. G. (2021). Land-use intensity mediates ecosystem service tradeoffs across regional social-ecological systems. *Ecosystems and People*, *17*(1). https://doi.org/10.1080/26395916.2021.1925743

Rahman, N. A., Larbi, A., Kotu, B., Kizito, F., & Hoeschle-Zeledon, I. (2020). Evaluating sustainable intensification of groundnut production in northern ghana using the sustainable intensification assessment framework approach. *Sustainability (Switzerland)*, *12*(15). https://doi.org/10.3390/SU12155970

Ranjani, J., Kalaiselvi, V. K. G., Sheela, A., Deepika Sree, D., & Janaki, G. (2021). Crop Yield Prediction Using Machine Learning Algorithm. *Proceedings of the 2021 4th International Conference on Computing and Communications Technologies, ICCCT 2021*. https://doi.org/10.1109/ICCCT53315.2021.9711853

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction. In *IEEE Access* (Vol. 9). https://doi.org/10.1109/ACCESS.2021.3075159

Rasmussen, L. V., Coolsaet, B., Martin, A., Mertz, O., Pascual, U., Corbera, E., Dawson, N., Fisher, J. A., Franks, P., & Ryan, C. M. (2018). Social-ecological outcomes of agricultural intensification. In *Nature Sustainability* (Vol. 1, Issue 6). https://doi.org/10.1038/s41893-018-0070-8

Raudsepp-Hearne, C., Peterson, G. D., & Bennett, E. M. (2010). Ecosystem service bundles for analyzing tradeoffs in diverse landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11). https://doi.org/10.1073/pnas.0907284107

Reuben, M. J. K., Japhet, J. K., Agnes, S., Felix, K., Anna, S., & Winfred, M. (2017). Land fragmentation, agricultural productivity and implications for agricultural investments in the Southern Agricultural Growth Corridor of Tanzania (SAGCOT) region, Tanzania. *Journal of Development and Agricultural Economics*, *9*(2). https://doi.org/10.5897/jdae2016.0797

Robertson, G. P., Gross, K. L., Hamilton, S. K., Landis, D. A., Schmidt, T. M., Snapp, S. S., & Swinton, S. M. (2014). Farming for ecosystem services: An ecological approach to production agriculture. In *BioScience* (Vol. 64, Issue 5). https://doi.org/10.1093/biosci/biu037

Rodríguez, J. P., Beard, T. D., Bennett, E. M., Cumming, G. S., Cork, S. J., Agard, J., Dobson, A. P., & Peterson, G. D. (2006). Trade-offs across space, time, and ecosystem services. *Ecology and Society*, *11*(1). https://doi.org/10.5751/ES-01667-110128

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2). https://doi.org/10.1016/S0010-4655(02)00280-1

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global sensitivity analysis: The primer. In *Global Sensitivity Analysis: The Primer*. https://doi.org/10.1002/9780470725184

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models. In: Probability and Statistics Series. In *Analyzing Uncertainty in Civil Engineering*.

Sanchez, P. A. (2002). Soil fertility and hunger in Africa. In *Science* (Vol. 295, Issue 5562). https://doi.org/10.1126/science.1065256

Sanchez, P. A. (2019). Properties and Management of Soils in the Tropics. In *Properties and Management of Soils in the Tropics*. https://doi.org/10.1017/9781316809785

Sanchez, P. A., Palm, C. A., & Buol, S. W. (2003). Fertility capability soil classification: A tool to help assess soil quality in the tropics. *Geoderma*, *114*(3–4). https://doi.org/10.1016/S0016-7061(03)00040-5

Sarkar, P. (2003). Sequential Monte Carlo Methods in Practice. *Technometrics*, *45*(1). https://doi.org/10.1198/tech.2003.s23

Scholes, B., Palm, C., & Andelman, S. (2013). *Sampling Frame for the Vital Signs Global Monitoring System*.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1). https://doi.org/10.1177/1536867X20909688

Scoones, I. (1998). Sustainable rural livelihoods: a framework for analysis. *IDS Working Paper*, *72*.

Sileshi, G., Akinnifesi, F. K., Debusho, L. K., Beedy, T., Ajayi, O. C., & Mong'omba, S. (2010). Variation in maize yield gaps with plant nutrient inputs, soil type and climate across sub-Saharan Africa. In *Field Crops Research* (Vol. 116, Issues 1–2). https://doi.org/10.1016/j.fcr.2009.11.014

Smaling, E. M. A., Nandwa, S. M., & Janssen, B. H. (2015). Soil Fertility in Africa Is at Stake. In *Replenishing Soil Fertility in Africa*. https://doi.org/10.2136/sssaspecpub51.c2

Smil, V. (1999). Crop Residues: Agriculture's Largest Harvest - Crop residues incorporate more than half of the world's agricultural phytomass. *BioScience*, *49*(4).

Smith, A., Snapp, S., Chikowo, R., Thorne, P., Bekunda, M., & Glover, J. (2017). Measuring sustainable intensification in smallholder agroecosystems: A review. In *Global Food Security* (Vol. 12). https://doi.org/10.1016/j.gfs.2016.11.002

Sobol', I. M. (1993). Sensitivity Estimates for Nonlinear Mathematical Models. In *Mathematical Modeling and Computational experiment* (Vol. 1).

Solesbury, W. (2005). Sustainable livelihoods: a case study of the evolution of DFID policy. In *Bridging Research and Policy in Development*. https://doi.org/10.3362/9781780444598.006

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1). https://doi.org/10.1093/bioinformatics/btr597

Tully, K., Sullivan, C., Weil, R., & Sanchez, P. (2015). The State of soil degradation in sub-Saharan Africa: Baselines, trajectories, and solutions. In *Sustainability (Switzerland)* (Vol. 7, Issue 6). https://doi.org/10.3390/su7066523

USDA. (2022). World agricultural production, Circular Series. In *Foreign Agricultural Service, United states Department of Agriculture.*

Van Ittersum, M. K., Van Bussel, L. G. J., Wolf, J., Grassini, P., Van Wart, J., Guilpart, N., Claessens, L., De Groot, H., Wiebe, K., Mason-D'Croz, D., Yang, H., Boogaard, H., Van Oort, P. A. J., Van Loon, M. P., Saito, K., Adimo, O., Adjei-Nsiah, S., Agali, A., Bala, A., … Cassman, K. G. (2016). Can sub-Saharan Africa feed itself? *Proceedings of the National Academy of Sciences of the United States of America*, *113*(52), 14964–14969. https://doi.org/10.1073/pnas.1610359113

Vanlauwe, B., Coyne, D., Gockowski, J., Hauser, S., Huising, J., Masso, C., Nziguheba, G., Schut, M., & Van Asten, P. (2014). Sustainable intensification and the African smallholder farmer. In *Current Opinion in Environmental Sustainability* (Vol. 8). https://doi.org/10.1016/j.cosust.2014.06.001

Vanlauwe, B., Six, J., Sanginga, N., & Adesina, A. A. (2015). Soil fertility decline at the base of rural poverty in sub-Saharan Africa. In *Nature Plants* (Vol. 1). https://doi.org/10.1038/nplants.2015.101

Varmuza, K. (1980). *Clustering Methods* (Issue May 2014). https://doi.org/10.1007/978-3-642-93155-0_7

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (1999). Self-organizing map in Matlab : the SOM Toolbox. *Proceedings of the Matlab DSP Conference*.

Wang, S., & Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(4). https://doi.org/10.1109/TSMCB.2012.2187280

Webber, H., Gaiser, T., & Ewert, F. (2014). What role can crop models play in supporting climate change adaptation decisions to enhance food security in Sub-Saharan Africa? In *Agricultural Systems* (Vol. 127). https://doi.org/10.1016/j.agsy.2013.12.006

Wehrens, R. (2015). Package ‘ kohonen .’ *R Topics Documented*.

Yuan, S., Linquist, B. A., Wilson, L. T., Cassman, K. G., Stuart, A. M., Pede, V., Miro, B., Saito, K., Agustiani, N., Aristya, V. E., Krisnadi, L. Y., Zanon, A. J., Heinemann, A. B., Carracelas, G., Subash, N., Brahmanand, P. S., Li, T., Peng, S., & Grassini, P. (2021). Sustainable intensification for a larger global rice bowl. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-27424-z

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6). https://doi.org/10.1177/15501329221106935

Zimmerer, K. S., Carney, J. A., & Vanek, S. J. (2015). Sustainable smallholder intensification in global change? Pivotal spatial interactions, gendered livelihoods, and agrobiodiversity. In *Current Opinion in Environmental Sustainability* (Vol. 14). https://doi.org/10.1016/j.cosust.2015.03.004

BIOGRAPHICAL SKETCH

Taisha Venort was born in the U.S and grew up in Haiti. She moved to the U.S. in 2008 after graduating from high school, to pursue her bachelor's education in Environmental and Ecological Engineering at Purdue University. She obtained her master's degree in the Ecological Sciences and Engineering program at Purdue in 2017 during which she obtained a Borlaug Fellowship to pursue her research on Farm biogas in Kenya. In 2019, she received a McKnight Fellowship at the University of Florida to pursue her doctorate degree in Agricultural and Biological Engineering. She had been a graduate research assistant working under the guidance of Dr. Cheryl Palm and Dr. Rafael Muñoz-Carpena. Her research focuses on agricultural intensification and natural-human systems, with a keen interest in smallholder systems.