PhD Dissertation Research Proposal

Name of the candidate: Caroline Akinyi OWUOR

Title of Dissertation: Integrating Earth Observation, Artificial Intelligence, and Crop Modelling for Maize Yield Estimation in Smallholder Farming Systems in Kenya

Department: Center for Remote Sensing Applications (CRSA)

Dissertation Committee:

Thesis Director: Pr. Ali El BATTAY,	University Mohammed VI Polytechnic
Thesis Co-Director: Dr. Louise LEROUX,	International Institute of Tropical Agriculture (IITA) and Cirad
Member: Dr. Siyabusa MKUHLANI,	IITA
Member: Dr. Marc CORBEELS,	IITA and Cirad

Academic Year: 2025-2026

Date: June 27, 2025

EXECUTIVE SUMMARY

Accurate crop yield estimation remains a persistent challenge in smallholder farming systems, particularly in Sub-Saharan Africa, where fragmented fields, diverse cropping practices, and limited ground-truth data hinder the scalability and precision of models. This study proposes an integrated framework combining Earth Observation (EO), Artificial Intelligence (AI), and process-based Crop Modelling (CM) to address these limitations and improve yield prediction in data-scarce environments. The research aims to assess the potential of high-resolution EO data for capturing spatial variability in smallholder landscapes, calibrate crop models using region-specific parameters and agroclimatic data, and develop and benchmark AI models for scalable, cross-regional yield prediction. A multi-source data fusion strategy will integrate satellite imagery, weather records, and field observations to support the development of models. Calibration and validation will be conducted using field-level yield data from diverse agroecological zones in Kenya to ensure robustness and transferability of the results. The study aims to develop a validated, scalable framework for yield estimation that can inform climate-resilient agricultural planning, support early warning systems, and enhance digital advisory services tailored to the needs of smallholders. This research will contribute to the broader agenda of food security by advancing context-sensitive, data-driven tools for monitoring crop productivity in under-resourced regions.

Keywords: Yield Estimation, Earth Observation, Artificial Intelligence, Crop Modelling, Smallholder Systems.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
APSIM	Agricultural Production Systems sIMulator
ASTGNN	Attention-based Spatio-Temporal Graph Neural Network
cGANs	Conditional Generative Adversarial Networks
CM	Crop Modeling
CNNs	Convolutional Neural Networks
CWSI	Crop Water Stress Index
DL	Deep Learning
DSSAT	Decision Support System for Agrotechnology Transfer
DST	Dempster-Shafer Theory
EO	Earth Observation
EPIC	Environmental Policy Integrated Climate
EVI	Enhanced Vegetation Index
GCVI	Green Chlorophyll Vegetation Index
GDPR	General Data Protection Regulation
GRADE	Grading of Recommendations Assessment, Development and Evaluation
LAI	Leaf Area Index
LORO-CV	Leave-One-Region-Out Cross-Validation
NDVI	Normalized Difference Vegetation Index
MAE	Mean Absolute Error
ML	Machine Learning
MMAT	Mixed Methods Appraisal Tool
MODIS	Moderate Resolution Imaging Spectroradiometer
R^2	coefficient of determination

ReCI	Red Edge Chlorophyll Index
RMSE	Root Mean Squared Error
RTE	Radiative Transfer Equation
RTT	Radiative Transfer Theory
SAR	Synthetic Aperture Radar
SCYM	Scalable Crop Yield Mapping
SFPS	Sustainable Farming Science Program
SHAP	SHapley Additive exPlanations
SIMPLACE	Scientific Impact Assessment and Modelling Platform for Advanced Crop and Ecosystem Management
SMOGN	Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise
SNAP	Sentinel Application Platform
SPI	Standard Precipitation Index
SSA	Sub-Saharan Africa
SSM	Soil Surface Moisture
TCI	Temperature Condition Index
TVDI	Temperature Vegetation Dryness Index
VAEs	Variational Autoencoders
VCI	Vegetation Condition Index

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

Global food systems are under increasing pressure to sustainably enhance agricultural productivity while adapting to the effects of climate change and mitigating environmental degradation. Sub-Saharan Africa (SSA) is particularly vulnerable due to its reliance on rainfed agriculture, rapid population growth, and exposure to climatic extremes such as drought and erratic rainfall (Foley et al., 2011; Giller et al., 2021). By 2100, SSA's population is projected to exceed 3.1 billion (Vollset et al., 2020), intensifying demands on food production systems already constrained by limited resources, land degradation, and low input use. Smallholder farmers, who produce approximately 85% of the region's food (Li et al., 2022), face persistent yield gaps due to poor soil quality, climate variability, and limited access to technology (Frelat et al., 2016; Giller et al., 2021).

Maize is a key staple crop across SSA, with per capita consumption reaching as high as 450 grams per day in certain regions (Ekpa et al., 2019). In Kenya, it is the most widely cultivated staple crop and is consumed widely across both rural and urban households (KNBS, 2024). However, maize yields in the country have exhibited a statistically significant decline of approximately 0.07 tons per hectare per decade, accompanied by a marked interannual variability. This decline is attributed mainly to rising temperatures, lower rainfall, low fertilizer, pest pressures, and the fragmentation of production systems (Mahuku et al., 2015; Mumo et al., 2018; Nyombi, 2019). The situation is expected to worsen under ongoing climate change, with projections indicating yield reductions of 11% in Eastern Kenya and 7% in Western Kenya by the 2040s if current warming and drying trends continue (Davenport et al., 2018). These challenges highlight the urgent need for spatially detailed, timely yield monitoring systems to inform adaptation strategies and enhance national food security planning.

Traditional yield estimation methods, such as field surveys and farmer-reported data, are limited by high costs, scalability issues, and inconsistencies in data quality (Burke & Lobell, 2017). In response to this, researchers are increasingly integrating Earth Observation (EO), Artificial Intelligence (AI), and Crop Modeling (CM) to enhance yield prediction (Bouras et al., 2023; Burke & Lobell, 2017; Leroux et al., 2019). EO platforms, such as Sentinel and PlanetScope, provide high-resolution, multi-temporal imagery that can be used to derive vegetation indices and biophysical parameters for monitoring crop growth and phenology (Aslan et al., 2024; Swain et al., 2024). These data, when combined with AI models, particularly Machine Learning (ML) and Deep Learning (DL), enable the extraction of complex patterns that relate biophysical indicators to yield outcomes (Boukhris et al., 2024).

Process-based models, such as the Decision Support System for Agrotechnology Transfer (DSSAT), CERES-Maize, and the Agricultural Production Systems sIMulator (APSIM), simulate crop development under varying environmental and management conditions, providing physiological insights that support model calibration and interpretation (Huang et al., 2019; Mkuhlani et al., 2024). These models are especially valuable in data-scarce settings, where their outputs can be used as proxy ground data to train AI models, a strategy increasingly applied in uncalibrated modeling approaches (Burke & Lobell, 2017; Jin, Azzari, Burke, et al., 2017). Integrated approaches that combine EO, AI, and CM, including uncalibrated strategies, have shown strong potential in improving yield estimates across heterogeneous smallholder landscapes by leveraging their respective strengths: scalability, data richness, and biophysical realism (Alimagham et al., 2025).

Despite these advancements, significant methodological and operational gaps persist in yield estimation for smallholder systems, particularly in data-scarce regions. Current data-driven approaches are often constrained by the limited availability of high-quality, georeferenced ground-truth data, which hampers model calibration and scalability (Gbodjo et al., 2021). While institutional agronomic datasets exist, they often provide only general information on crop presence and lack the spatial precision necessary for EO-based modeling. To address these limitations, this study proposes an integrated and scalable framework that combines high-resolution EO data, process-based crop models, and advanced AI techniques for estimating maize yields. Unlike many existing approaches, the framework will utilize crop model outputs as pseudo-ground truth to support yield estimation in data-scarce environments where field observations are sparse. It will also incorporate spatiotemporal data fusion and transfer learning to improve scalability across diverse agroecological regions. To effectively manage uncertainty and synthesize conflicting data inputs, the framework will apply modified Dempster-Shafer Theory (DST) for evidence-based integration. By aligning biophysical realism with data-driven inference, the study will enhance prediction accuracy, improve spatial adaptability, and contribute to more reliable and scalable yield monitoring systems for smallholder agriculture.

1.2 Research Problem

Reliable and scalable crop yield estimation remains a significant challenge in smallholder farming systems, particularly in SSA, where high spatial variability, fragmented fields, and limited access to quality ground-truth data constrain model accuracy (Leroux et al., 2019; Sisheber et al., 2024). The additional complexity of intercropping and diverse management practices further limits the generalizability of existing models across smallholder landscapes (Gómez-Dans et al., 2022). Although simulation-based approaches have been introduced to address data gaps, further research is needed to evaluate their transferability across different agroecological zones and to investigate their integration with ML for large-scale applications

(Sisheber et al., 2024). Many current approaches still fall short in capturing fine-scale spatial variability and lack the advanced integration techniques needed to improve accuracy and operational scalability (Quintero et al., 2025). Addressing these limitations is critical for strengthening agricultural monitoring systems and supporting climate adaptation and food security planning in smallholder-dominated regions.

1.3 Research Objectives

1.3.1 General Objective

To develop an integrated crop yield estimation framework that combines EO, CM, and AI to improve predictive accuracy, address data scarcity, and assess scalability across diverse agroecological regions.

1.3.2 Specific Objectives

1) Systematically review current methodologies and advancements in using EO, CM, and AI for crop yield estimation among smallholder farmers in SSA.
2) Develop an approach for integrating multisource data from EO and agronomic factors with crop models to enhance the spatial and temporal accuracy of crop yield estimations in data-scarce environments.
3) Evaluate the contribution of data augmentation and synthetic data generation to the robustness and accuracy of AI-based yield prediction models in smallholder farming systems.
4) Examine the scalability and transferability of EO, CM, and AI-based crop yield estimation methods across different agroecological regions through cross-regional validation and identify the exogenous and endogenous factors influencing their generalizability.

1.4 Hypothesis

This study hypothesizes that integrating EO, AI, and CM can enhance crop yield estimation in smallholder farming systems. Specifically, it hypothesizes that:

1) Emerging satellite platforms such as Sentinel and PlanetScope enable improved monitoring of smallholder agriculture.
2) Process-based crop models contribute physiological insights that enhance the scalability of EO data across diverse agroecological zones.
3) Recent advances in AI, including transfer learning and domain adaptation, can effectively bridge data gaps and enhance the integration of EO-CM.

1.5 Research Questions

1) What are the current methodologies and advancements in using EO, CM, and AI for crop yield estimation among smallholder farmers in SSA?
2) How can multisource data from EO and agronomic factors be integrated with crop models to improve the spatial and temporal accuracy of crop yield estimations in smallholder farming systems?
3) How do data augmentation strategies and the use of synthetic data influence the robustness and predictive accuracy of AI-based yield estimation models in smallholder systems?
4) How scalable and transferable are EO, CM, and AI-based crop yield estimation methods across different agroecological regions, and how do exogenous and endogenous factors differently influence their generalizability?

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter examines a body of literature on crop yield estimation, focusing on integrative approaches that combine EO, AI, and CM. It begins by situating the discussion within the broader biophysical and environmental context of maize yield formation before introducing the theoretical frameworks that guide the study's modelling strategy. The remainder of the chapter is organized around key themes aligned with the study's objectives, including data sources, modelling techniques, and the integration of methods to improve the accuracy and scalability of yield prediction.

2.2 Determinants of Maize Yield

Maize yield is governed by a complex interplay of physiological, environmental, genetic, and management factors, whose interactions shape productivity across spatial and temporal scales. The reproductive stage, particularly the period surrounding tasseling and silking, is a crucial determinant of the final yield. During this phase, the plant's growth rate and carbon assimilation capacity directly influence the number of kernels per ear, which in turn determines grain yield, even under moderate stress conditions (Maddonni et al., 2024).

Light availability plays a key role in supporting photosynthesis and biomass accumulation. Optimal canopy traits, including upright leaves, high Leaf Area Index (LAI), and smaller tassels, enhance light-use efficiency and radiation interception, thereby contributing to increased biomass production and yield (Lambertet al., 2014). However, excessive canopy density or low-light conditions can restrict light penetration, limiting photosynthesis and reducing dry matter and nutrient uptake. Variability in yield under such conditions is often linked to shade tolerance, underscoring the importance of canopy architecture (J. Gao et al., 2017).

In tropical rainfed systems, yield is highly sensitive to variations in temperature and precipitation. High temperatures accelerate crop development up to an optimum, beyond which excessive heat shortens the growth cycle and reduces grain filling (Tonnang et al., 2018). Sensitivity is most significant during pre-flowering and flowering, when high temperatures can impair reproductive processes, reduce chlorophyll content, and restrict assimilate production, thereby disrupting pollination and kernel development (Holzkämper et al., 2013; Mu et al., 2022). Precipitation patterns further influence yield by shaping the timing and duration of growth stages. Adequate rainfall extends vegetative growth, while deficits often shorten reproductive phases and restrict grain filling. Water stress during the flowering stage is especially detrimental, often disrupting pollination and reducing kernel set. (Peng et al., 2024).

The interplay among developmental stages, canopy-light interactions, and environmental drivers is summarized in *Figure 1*, providing a visual representation of the processes that influence maize yield.

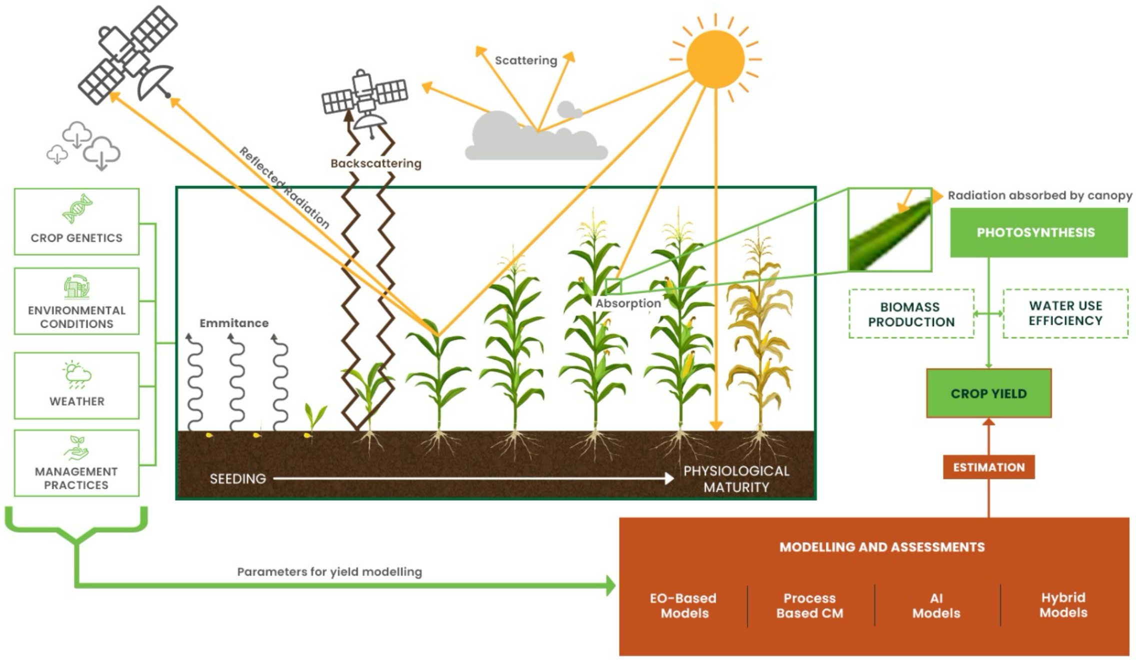

Figure 1: Interaction of crop growth, radiation, and environmental factors driving maize yield

2.3 Theoretical Framework

As outlined in Section 2.2, maize yield is shaped by complex interactions among physiological processes, environmental stressors, canopy-light dynamics, and resource availability. Capturing these multi-dimensional influences requires theoretical frameworks that reflect biological mechanisms, account for radiative behavior, and handle uncertainty in diverse data environments. As a result, this study draws on Ecological Modelling Theory, Radiative Transfer Theory (RTT), and Dempster-Shafer Theory (DST) to guide the integration of crop modelling, EO, and AI in a hybrid yield estimation framework.

2.3.1 Ecological Modelling Theory

Ecological Modelling Theory provides a systems-based perspective for capturing the complex interactions between crops and their environment highlighted in Section 2.2. This theoretical foundation underlies process-based crop models, which simulate key physiological processes, including photosynthesis, respiration, and nutrient uptake, based on environmental drivers (Arhonditsis et al., 2017; Biswal et al., 2020). By translating these biophysical interactions into predictive outputs, such models generate ecologically grounded indicators of crop performance. Within this study, ecological modelling theory supports the integration of crop models into the hybrid framework, enhancing the biological realism and contextual relevance of yield predictions in smallholder farming systems.

2.3.2 Radiative Transfer Theory (RTT)

RTT is central in understanding how electromagnetic radiation interacts with different media through processes such as absorption, scattering, and emission. At the core of this framework is the Radiative Transfer Equation (RTE), which mathematically describes these interactions and is widely employed in remote sensing to interpret surface reflectance and estimate biophysical properties (Khawaja, 2017). In crop monitoring, physically based radiative transfer models, such as PROSAIL, simulate the spectral response of vegetation canopies using structural and biochemical inputs (Dhakar et al., 2022). These models enable the retrieval of key variables that indicate crop condition and yield potential, such as LAI and chlorophyll content (Ishaq et al., 2024). By linking spectral observations to crop characteristics, RTT serves as a theoretical foundation for the EO component of this study, enabling the extraction of yield-relevant information from satellite imagery.

2.3.3 Dempster-Shafer Theory (DST)

The DST, also known as Evidence Theory, is a mathematical approach designed to handle uncertainty and imprecision when integrating information from multiple sources (Skau et al., 2023). Unlike traditional probabilistic models that assign likelihoods to single outcomes, DST distributes belief across sets of possibilities using basic probability assignments, offering a more adaptable framework for situations with incomplete or ambiguous data (Silva & De Almeida-Filho, 2016). This is particularly useful in hybrid yield estimation, where data inputs, ranging from satellite observations and crop simulation outputs to ancillary environmental variables, often vary in accuracy and completeness. By employing belief and plausibility measures, DST can quantify the level of support for different yield-related scenarios and combine evidence through Dempster's Rule. Its relevance to data fusion is well demonstrated in recent applications, including the harmonization of conflicting land-use datasets (Bethuel et al., 2025) and the integration of heterogeneous EO sources for land surface modelling (A. Huang et al., 2022). Within this study, DST forms the theoretical basis for combining diverse data streams to support more reliable and interpretable yield predictions across variable smallholder landscapes.

2.4 Empirical Review of Related Studies

2.4.1 Data Sources for Yield Estimation

Accurate yield estimation in SSA relies on the integration of diverse data sources, typically categorized into remote sensing, *in situ* observations, and integrated datasets. Remote sensing platforms, such as Landsat, the Moderate Resolution Imaging Spectroradiometer (MODIS), Sentinel, and PlanetScope, are widely used to derive key biophysical indicators, crop phenology, and vegetation indices (Sisheber et al., 2024; Zhang & Zhang, 2016). Sentinel-2 and PlanetScope are also effective for capturing spatial detail in smallholder and intercropped systems due to their high spatial and temporal resolution (Bandaru et al., 2022; Swain et al., 2024). Additionally, Synthetic Aperture Radar (SAR) data offer coverage in cloud-prone regions, operating independently of light and weather conditions (Kalecinski et al., 2024), making them well-suited for regions where agriculture is predominantly rainfed.

In situ data, including crop type surveys, LAI measurements, and field yield observations, play a crucial role in calibrating and validating EO and ML models (Gómez-Dans et al., 2022; Tiedeman et al., 2022). In contexts where ground-truth data are limited, crop models have been used to generate pseudo-ground-truth data (Burke & Lobell, 2017; Jin, Azzari, Burke, et al., 2017; Leroux et al., 2019). Integrated datasets, such as EthCT2020 by Blasch et al. (2024) and HarvestStat Africa by Lee et al. (2025), further strengthen modeling efforts by providing georeferenced crop samples and subnational yield statistics across the SSA. These datasets enhance spatial consistency and enable broader applicability of yield prediction models across diverse agroecological zones.

2.4.2 Modelling Approaches for Yield Estimation

Modelling approaches for yield estimation in SSA have integrated EO, ML, and crop simulation models. In a study conducted in Burkina Faso and Somalia, Lee et al. (2024) demonstrated that incorporating static biophysical variables, such as soil properties and cropland extent, into time-series models significantly enhanced the coefficient of determination (R^2) of yield predictions by 0.11. In Kenya, Asalla et al. (2024) combined EO data with a Random Forest model to produce maize yield estimates, achieving a Root Mean Squared Error (RMSE) of 0.203 Mg/ha, which indicates the scalability of ML-based approaches. Similarly, multisensor data fusion in Ethiopia has been shown to reduce yield prediction errors, with relative RMSE values for maize as low as 16% (Sisheber et al., 2024).

Hybrid methodologies that combine probabilistic estimation and crop modelling have also gained traction. In Central Ghana, the integration of the LINTUL5 crop model within the SIMPLACE (Scientific Impact Assessment and Modelling Platform for Advanced Crop and Ecosystem Management) framework, coupled with a multilayer soil water balance model (SLIM), showed that probabilistic assumptions about sowing dates consistently reduced RMSE compared to deterministic approaches (Srivastava et al., 2016). In Burkina Faso, Leroux et al. (2019) applied a Random Forest model calibrated with outputs from the SARRA-O crop model to estimate maize yields using vegetation indices and crop water stress metrics. The model explained up to 46% of the observed yield variability two months before harvest, outperforming traditional regression methods and highlighting the potential of integrated EO-ML-CM frameworks in data-scarce environments.

Building on these advances, recent studies have also explored integrated modelling systems to assess the effects of climate variability on yield outcomes across SSA. For maize in SSA, Alimagham et al. (2025) developed a three-step modelling system that combines Random Forest algorithms with eco-physiological crop processes to simulate sowing dates, phenology, and water-limited potential yields. Applied across the region, this hybrid approach produced gridded yield projections with accuracy comparable to high-input crop models, within 20% deviation in 95% of cases, while requiring significantly less data. In another study, Dale et al. (2017) utilized the AquaCrop-OS model, which is linked to over 120 climate model projections, to assess future trends in maize production. The modelling ensemble revealed robust spatial patterns, projecting yield declines in semi-arid zones of the Sahel and Southern Africa, and potential gains in more humid regions such as East Africa. These applications demonstrate how integrated modelling systems, combining process-based and data-driven components, can enhance yield forecasting and support climate-resilient agricultural planning in SSA.

2.4.3 Validation Techniques in Yield Estimation

Robust validation techniques are critical for ensuring the reliability and generalizability of crop yield estimation models. Cross-validation methods are widely used to assess model performance across different data splits and temporal scenarios. K-fold cross-validation, for instance, enables repeated training and testing across subsets of data, thereby reducing bias and overfitting. Multi-annual cross-validation has also been applied to evaluate model stability across growing seasons, thereby enhancing confidence in long-term yield forecasting (Brandt et al., 2024; Kolipaka & Namburu, 2023).

Ensemble learning approaches further strengthen model validation by combining predictions from multiple algorithms. Techniques such as majority voting and stacking have demonstrated improved accuracy and resilience in yield prediction tasks (Brandt et al., 2024). Performance evaluation typically involves metrics such as R^2, RMSE, and Mean Absolute Error (MAE), which enable comparison across models (Segarra et al., 2022). In data-scarce environments, bootstrap methods have been employed to improve sample representativeness and parameter estimation. Bayesian-enhanced bootstrap techniques have shown potential in extending predictive validity when limited observations are available (Song et al., 2018).

Process-based crop models are usually validated against field measurements. These simulations have shown strong alignment with actual yields, particularly when calibrated using site-specific management and environmental data (Mandapati et al., 2024). Collectively, these validation strategies enhance the credibility of yield prediction models.

2.5 Synthesis and Research Gaps

Hybrid modelling approaches have shown promise in improving crop yield prediction across diverse agricultural landscapes. With notable successes in smallholder-dominated regions of SSA, empirical studies in Kenya, Ethiopia, Ghana, and Burkina Faso have demonstrated that combining geospatial data with ML techniques can produce spatially explicit, scalable yield estimates (Asalla et al., 2024; Lee et al., 2024; Sisheber et al., 2024; Srivastava et al., 2016). Crop models have also played a critical role by simulating crop growth and generating pseudo-ground truth, enabling calibration and validation in data-scarce settings (J. Huang et al., 2019).

Nevertheless, key limitations persist in literature. First, the combined use of phenological indicators and yield outputs remains underutilized, despite the fact that phenology-aware modeling can enhance early-season forecasting and improve the interpretability of ML outputs. Second, while crop models are increasingly used to generate synthetic data, the systematic use of crop simulation outputs as pseudo-ground truth to bootstrap AI models in the absence of high-quality field data is still limited. Additionally, spatial transferability across agroecological zones has not been sufficiently explored. Most yield models are calibrated for specific locations and exhibit limited generalizability due to the lack of techniques such as transfer learning, stratified validation, or domain adaptation.

This study aims to address these gaps by developing an integrated EO–CM–AI framework that combines phenological and yield indicators. This framework will leverage crop model outputs to support data-scarce learning and validation. The model will also be evaluated for generalizability across diverse agroecological contexts. Through this approach, the research will contribute to the development of more robust, interpretable, and scalable yield estimation systems, which are better suited to the heterogeneity and data limitations of smallholder agriculture.

CHAPTER 3: METHODOLOGY

3.1 Research Design and Approach

This study will employ a quantitative, systems-based modelling approach that integrates EO, AI, and CM to predict maize phenological stages and yield in smallholder farming systems. The methodology will be implemented through three interrelated components: (1) EO-based extraction of phenological and biophysical indicators, (2) process-based crop simulation for synthetic data generation and pseudo-ground truthing, and (3) AI-based data fusion and prediction.

This integrated design is guided by a pragmatic research philosophy, emphasizing methodological flexibility and the use of tools best suited to address real-world agricultural challenges. Instead of focusing on just one source of knowledge, the framework will employ a combination of mechanistic, observational, and data-driven methods to demonstrate the complexity of smallholder production systems. The CM component is informed by a post-positivist stance, treating crop growth as a function of biophysical processes while acknowledging model uncertainty and contextual variability. The EO component is rooted in a scientific realist paradigm that applies physical laws of RTT to infer crop conditions from remotely sensed data. The AI component reflects a constructivist-pragmatic view, where knowledge emerges from patterns in multisource data and is primarily evaluated through predictive performance.

3.2 Study Area and Data Collection

3.2.1 Study Area

This study will focus on smallholder maize production systems in Kenya, a country characterized by complex topography and agroecological variability that influence localized climate conditions and heterogeneous farming practices (Ondiek et al., 2024).

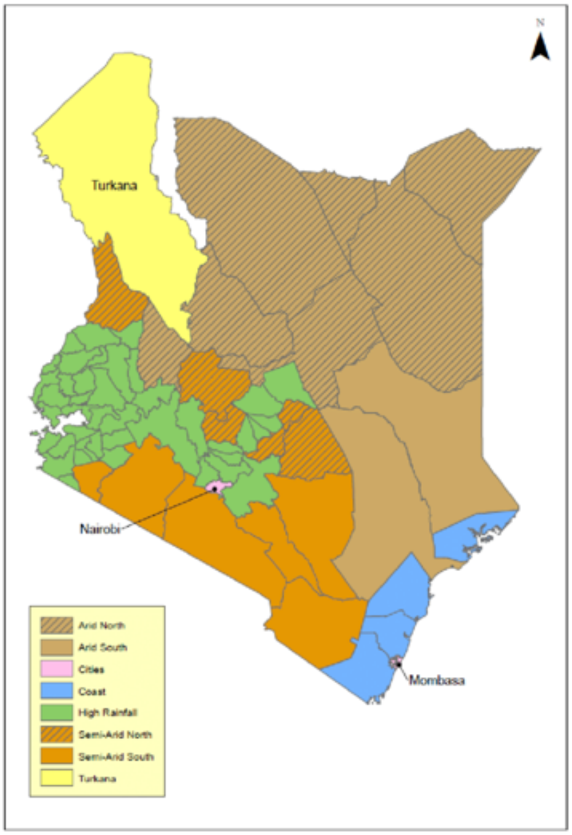

Figure 2: Kenya agroecological zones. Source (Vigani et al., 2019)

3.2.2 Data

Multispectral and thermal imagery from Sentinel-1 SAR, Sentinel-2, and Landsat-8 will be used to derive vegetation, drought, and heat stress indices, which will serve as input features for training the yield estimation models. For this study, NDVI, Enhanced Vegetation Index (EVI), Green Chlorophyll Vegetation Index (GCVI), Red Edge Chlorophyll Index (ReCI), Soil Surface Moisture (SSM), Temperature Condition Index (TCI), CWSI, Temperature Vegetation Dryness Index (TVDI), and LAI will be used (Jin, Azzari, Burke, et al., 2017; Leroux et al., 2019). Simulated variables from process-based crop models will be used to complement EO data. These include phenological stages, biomass accumulation, soil moisture, water and nitrogen stress factors, evapotranspiration, and yield estimates.

Field data for calibrating and validating the models will be obtained through the CGIAR Sustainable Farming Science Program (SFSP) and its partnerships with local institutions. This includes phenological dates, measured grain yield, LAI, soil moisture, and site-specific soil and management data (sowing date, row spacing, fertilizer application, and tillage). Local weather station data will also be incorporated for model initialization and calibration.

3.3 Modeling and Analytical Approach

3.3.1 Objective 1: Systematic Review of Yield Estimation Methodologies

A systematic review will be conducted using the PRISMA 2020 guidelines (Page et al., 2021), drawing on Scopus, Web of Science, and PubMed to identify peer-reviewed literature on crop yield estimation in smallholder systems of SSA using EO, CM, and AI. Searches will employ Boolean-enhanced queries adapted to each database, with deduplication performed using a Python script and reference management conducted in Mendeley. Studies will be included if they present original empirical research or systematic reviews published between 2015 and 2025, focus on EO, CM, and AI-based yield estimation in SSA smallholder contexts, and meet minimum standards of methodological transparency. Data extraction will follow a standardized framework that captures the geographic scope, input data, modeling approaches, validation strategies, and performance metrics. The study quality and risk of bias will be assessed using the Mixed Methods Appraisal Tool (MMAT), and the strength of cumulative evidence will be evaluated through the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework (Hong et al., 2019; Prasad, 2024). Due to expected heterogeneity in methods and outcomes, a narrative synthesis will be conducted, thematically grouping studies by modeling and evaluation approach, data type, and spatial-temporal scale. A complementary bibliometric analysis will be done using PyBibX and Latent Dirichlet Allocation (LDA) topic modeling in Python to identify publication patterns and latent research themes (Asmussen & Møller, 2019; Pereira et al., 2023). Conceptual maps generated in VOSviewer will visualize relationships among methods, data, and outcomes.

3.3.2 Objective 2: Integrated Modeling Framework

This study will implement both calibrated and uncalibrated variants of the Scalable Crop Yield Mapper (SCYM) framework to estimate maize yields. SCYM integrates crop model outputs, EO data, and weather variables within a regression-based pipeline, enabling yield estimation even in the absence of field observations (Lobell et al., 2015). Crop simulations will be conducted using an ensemble approach that combines DSSAT CERES-Maize and APSIM models to represent the physiological diversity of maize responses under varying agroecological and management conditions. Simulations will draw on harmonized weather inputs (ERA5, CHIRPS), soil data (ISRIC SoilGrids), and management parameters. Model outputs will be aggregated using Bayesian model averaging, assigning weights based on historical performance (Y. J. Gao et al., 2021).

Crop model calibration will ensure alignment between simulated and observed crop performance. Phenological parameters will be adjusted using field-observed dates of emergence, flowering, and maturity, which will be calibrated against thermal time derived from ERA5 temperature data. Biomass accumulation and canopy development will be calibrated using field-based biomass measurements and LAI estimates obtained through PROSAIL inversion of Sentinel-2A imagery (Dhakar et al., 2021; Feleke et al., 2021). Grain yield will be adjusted to match the observed harvest data. Calibration will be conducted through iterative adjustment, with EO-derived phenology and canopy metrics used as proxies where ground observations are limited. Simulated variables will be translated into remotely observable proxies highlighted in Section 3.2.2, using crop-specific empirical relationships. These pseudo-observations, paired with simulated yields, will be used to train multitemporal regression models within SCYM. EO inputs will be preprocessed through atmospheric correction, cloud masking, and temporal smoothing. Predictive models will include Random Forests and Gradient Boosted Regression Trees for baseline performance, as well as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for learning spatial and temporal patterns from EO time series.

Yield estimation will apply trained regression models to actual EO and weather observations. In the calibrated SCYM workflow, models will be trained on observed and ensemble-simulated outputs, while in the uncalibrated workflow, pseudo-yields will be derived solely from crop simulations (Jin, Azzari, & Lobell, 2017; Leroux et al., 2019). Evaluation will involve 10-fold cross-validation and spatial-temporal holdout techniques to assess generalizability across seasons and locations. Performance metrics will include RMSE, MAE, R^2, and phenology-specific accuracy scores, with predictive uncertainty quantified through the Monte Carlo dropout approach to support robust interpretation of yield outputs.

3.3.3 Objective 3: Data Augmentation and Synthetic Data Generation

To assess the impact of synthetic data on model robustness and accuracy, this study will apply a combination of deep learning and sample-based augmentation techniques. Conditional generative adversarial networks (cGANs) and variational autoencoders (VAEs) will be used to generate synthetic EO-derived index sequences and yield estimates (Bai et al., 2024; Dhumale et al., 2024). To improve the representation of data-scarce regions, the Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise (SMOGN) will be employed, producing additional samples of the yield distribution while preserving data variability through the injection of Gaussian noise (Li et al., 2025). Temporal augmentation of the EO time series will be performed to enhance sample diversity further using elastic warping and window cropping.

Six training datasets will be constructed: a baseline dataset using observed data; three datasets individually augmented with cGANs, VAEs, and SMOGN; a hybrid dataset combining all three methods; and a sixth incorporating pseudo-observations from the uncalibrated SCYM workflow outlined in Section 3.3.2. Identical CNNs and RNNs will be trained on each dataset to isolate the effects of augmentation. Model performance will be assessed using RMSE, MAE, and R^2 under 10-fold cross-validation and spatial-temporal holdout schemes, allowing for a comparative evaluation against SCYM-based models. Predictive uncertainty will also be quantified using a Monte Carlo dropout approach. To assess the statistical significance of performance differences, the Friedman test, followed by pairwise Wilcoxon signed-rank tests with Holm correction, will be employed (Satpathi et al., 2023). Ablation experiments will further examine the individual and combined contributions of each augmentation strategy to improving model stability, generalization, and predictive accuracy in smallholder contexts.

3.3.4 Cross-Regional Validation and Transferability Assessment

This study will adopt a transfer learning framework to assess the cross-regional scalability of deep learning-based yield estimation models developed in Objective 3. Pretrained models initially developed on maize yield data from western Kenya will be fine-tuned and evaluated on eastern regions that differ in agroecological characteristics, as seen in *Figure 2*. In parallel, NASA's Presto, a lightweight transformer architecture pretrained on large-scale EO time series (Tseng et al., 2023), will be fine-tuned using smallholder datasets from target domains. A comparative analysis will be conducted to evaluate the generalization performance of these models under data-scarce conditions across spatial holdout and leave-one-region-out cross-validation (LORO-CV) protocols.

To explain model performance across domains, the analysis will distinguish between exogenous factors, including climate regime, soil type, topography, and crop phenology, and endogenous factors, such as model architecture, hyperparameter choices, internal crop model assumptions, and the quality or resolution of EO inputs. SHapley Additive exPlanations (SHAP) and feature ablation analyses will be employed to quantify the contribution of each factor to model accuracy and error propagation. This dual-layered evaluation will help determine whether performance limitations arise primarily from agro-environmental variability or from constraints within the modeling pipeline itself. The findings will provide evidence on how and under what conditions pretrained and transferable models can support robust yield estimation across heterogeneous smallholder farming systems.

The model development and evaluation processes will be implemented using Python and R, with geospatial preprocessing conducted on Google Earth Engine and the Sentinel Application Platform (SNAP). Crop model integration and simulation will be done using DSSAT and APSIM platforms. Spatial analyses and visualization will be complemented using QGIS.

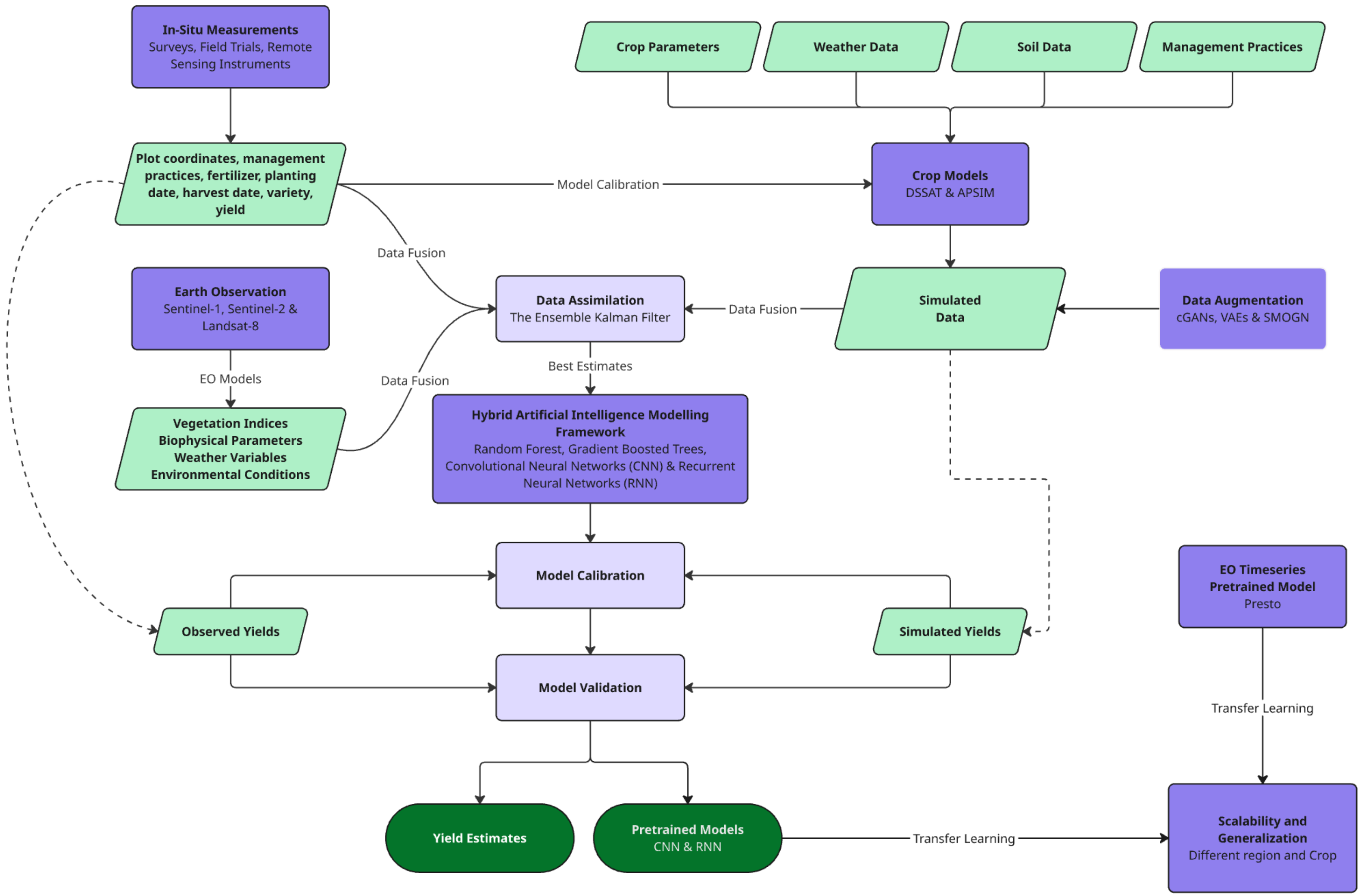

Figure 3: Hybrid modelling framework integrating Earth Observation, crop simulations, and Artificial Intelligence for yield estimation

3.4 Ethical Considerations

This study will adhere to ethical standards in the collection, use, and analysis of agricultural and geospatial data in smallholder farming systems.

3.4.1 Data Privacy and Security

This study will use remote sensing data and farmer-contributed information. Identifiable data will be anonymized to protect the privacy of individual farms. For ground-truth data or surveys, informed consent will be obtained, and all collected information will be securely stored and used solely for research purposes. Data handling will comply with the General Data Protection Regulation (GDPR) and relevant local governance frameworks.

3.4.2 Bias and Fairness

AI models are highly dependent on the quality and representativeness of training data. Incomplete or biased datasets may lead to unfair predictions that disadvantage specific farming communities or regions. To ensure equitable performance across agroecological zones, this study will utilize diverse, multi-source datasets and apply bias detection and mitigation strategies throughout the model development process.

3.4.3 Regulatory and Ethical Compliance

All data collection, storage, processing, and sharing will comply with national and international data governance standards. Institutional ethical approval will be obtained, and the rights of participants will be upheld. Furthermore, any third-party data will be used in accordance with relevant licensing agreements and data-sharing policies.

3.5 Expected Findings and Outputs

The study is expected to generate high-resolution, spatially explicit maize yield predictions across heterogeneous smallholder systems in Kenya. Methodologically, it will contribute a scalable, hybrid framework for yield estimation in data-scarce smallholder farming contexts, aiming to advance current approaches in agricultural monitoring. The findings will inform evidence-based decision-making for climate adaptation, input targeting, and policy design in smallholder agriculture. Results will be disseminated through peer-reviewed journal articles, conference and workshop presentations, and open-access platforms, including integration into the dissemination channels of the CGIAR SFSP and the Center for Remote Sensing Applications department of the University Mohammed VI Polytechnic, to support broader scientific and practical uptake.

REFERENCES

Alimagham, S., van Loon, M. P., Ramirez-Villegas, J., Berghuijs, H. N. C., Rosenstock, T. S., & van Ittersum, M. K. (2025). Integrating crop models and machine learning for projecting climate change impacts on crops in data-limited environments. *Agricultural Systems*, *228*. https://doi.org/10.1016/j.agsy.2025.104367

Arhonditsis, G., Recknagel, F., & Joehnk, K. (2017). Process-based modeling of nutrient cycles and food-web dynamics. In *Ecological Informatics Data Management and Knowledge Discovery Third Edition*. https://doi.org/10.1007/978-3-319-59928-1_10

Asalla, I., Frimpong, D. B., & Nakalembe, C. (2024). Field-Scale Yield Estimation Using Remote Sensing Time-Aggregate Variables. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 9716–9720. https://doi.org/10.1109/IGARSS53475.2024.10642941

Aslan, M. F., Sabanci, K., & Aslan, B. (2024). Artificial Intelligence Techniques in Crop Yield Estimation Based on Sentinel-2 Data: A Comprehensive Survey. *Sustainability (Switzerland)*, *16*(18). https://doi.org/10.3390/su16188277

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, *6*(1), 93. https://doi.org/10.1186/s40537-019-0255-7

Bai, Y., Chen, D., & Zhang, Z. (2024). A Conditional Generative Adversarial Networks-Augmented Case-Based Reasoning Framework for Crop Yield Predictions with Time-Series Remote Sensing Data. *Ceur Workshop Proceedings*, *3708*, 192–205.

Bandaru, V., Chirumamilla, P., Skakun, S., Lasko, K., & Yampracha, S. (2022). Application of Geo-CropSim framework for rainfed sugarcane yield assessment in Thailand. In *Remote Sensing of Agriculture and Land Cover/Land Use Changes in South and Southeast Asian Countries*. https://doi.org/10.1007/978-3-030-92365-5_22

Bethuel, C., Arvor, D., Corpetti, T., Helie Julia and Descals, A., Gaveau, D., Cheron-Bessou, C., Gignoux, J., & Corgne, S. (2025). Applying the Dempster-Shafer Fusion Theory to Combine Independent Land-Use Maps: A Case Study on the Mapping of Oil Palm Plantations in Sumatra, Indonesia. *REMOTE SENSING*, *17*(2). https://doi.org/10.3390/rs17020234

Biswal, A., Chakraborty, A., & Murthy, C. S. (2020). Spatialization of Crop Growth Simulation Model Using Remote Sensing. In *Geospatial Technologies for Crops and Soils*. https://doi.org/10.1007/978-981-15-6864-0_4

Blasch, G., Alemayehu, Y., Lesne, L., Wolter, J., Taymans, M., Tesfaye, T., Negash, T., Andulalem, M., Gutu, K., Debela, M., Defourny, P., & Hodson, D. P. (2024). Ethiopian Crop Type 2020 (EthCT2020) dataset: Crop type data for environmental and agricultural remote sensing applications in complex Ethiopian smallholder wheat-based farming systems (Meher season 2020/21). *Data in Brief*, *54*. https://doi.org/10.1016/j.dib.2024.110427

Boukhris, A., Jilali, A., & Sadiq, A. (2024). Satellite imagery, big data, IoT and deep learning techniques for wheat yield prediction in Morocco. *Results in Control and Optimization*, *17*. https://doi.org/10.1016/j.rico.2024.100489

Bouras, E. L. H., Thapa, S., Olsson, P. O., Albertsson, J., & Eklundh, L. (2023). Assimilation of Sentinel-2 data into a Crop Growth Model for High-resolution Crop Yield Mapping. *AGU Fall Meeting Abstracts*, *2023*, GC44E-04.

Brandt, P., Beyer, F., Borrmann, P., Möller, M., & Gerighausen, H. (2024). Ensemble learning-based crop yield estimation: a scalable approach for supporting agricultural statistics. *GIScience and Remote Sensing*, *61*(1). https://doi.org/10.1080/15481603.2024.2367808

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proceedings of the National Academy of Sciences*, *114*(9), 2189–2194. https://doi.org/10.1073/pnas.1616919114

Dale, A., Fant, C., Strzepek, K., Lickley, M., & Solomon, S. (2017). Climate model uncertainty in impact assessments for agriculture: A multi-ensemble case study on maize in sub-Saharan Africa. *Earth's Future*, *5*(3), 337–353. https://doi.org/10.1002/2017EF000539

Davenport, F., Funk, C., & Galu, G. (2018). How will East African maize yields respond to climate change and can agricultural development mitigate this response? *Climatic Change*, *147*(3–4), 491–506. https://doi.org/10.1007/s10584-018-2149-7

Dhakar, R., Sehgal, V. K., Chakraborty, D., Sahoo, R. N., & Mukherjee, J. (2021). Field scale wheat LAI retrieval from multispectral Sentinel 2A-MSI and LandSat 8-OLI imagery: effect of atmospheric correction, image resolutions and inversion techniques. *GEOCARTO INTERNATIONAL*, *36*(18), 2044–2064. https://doi.org/10.1080/10106049.2019.1687591

Dhakar, R., Sehgal, V. K., Chakraborty, D., Sahoo, R. N., Mukherjee, J., Ines, A. V. M., Kumar, S. N., Shirsath, P. B., & Roy, S. B. (2022). Field scale spatial wheat yield forecasting system under limited field data availability by integrating crop simulation model with weather forecast and satellite remote sensing. *Agricultural Systems*, *195*, 103299. https://doi.org/10.1016/J.AGSY.2021.103299

Dhumale, Y., Bamnote, G., Kale, R., Sawale, G., Chaudhari, A., & Karwa, R. (2024). Generative Modeling Techniques for Simulating Rare Agricultural Events on Prediction of Wheat Yield Production. *2024 2nd Dmiher International Conference on Artificial Intelligence in Healthcare Education and Industry Idicaiei 2024*. https://doi.org/10.1109/IDICAIEI61867.2024.10842716

Ekpa, O., Fogliano, V., Linnemann, A. R., Palacios-Rojas, N., & Kruseman, G. (2019). Sub-Saharan African Maize-Based Foods - Processing Practices, Challenges and Opportunities. *Food Reviews International*, *35*(7), 609–639. https://doi.org/10.1080/87559129.2019.1588290

Feleke, H. G., Savage, M., & Tesfaye, K. (2021). Calibration and validation of APSIM-Maize, DSSAT CERES-Maize and AquaCrop models for Ethiopian tropical environments. *South African Journal of Plant and Soil*, *38*, 36–51. https://doi.org/10.1080/02571862.2020.1837271

Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., Mueller, N. D., O'Connell, C., Ray, D. K., West, P. C., Balzer, C., Bennett, E. M., Carpenter, S. R., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., … Zaks, D. P. M. (2011). Solutions for a cultivated planet. *Nature*, *478*(7369), 337–342. https://doi.org/10.1038/nature10452

Frelat, R., Lopez-Ridaura, S., Giller, K. E., Herrero, M., Douxchamps, S., Djurfeldt, A. A., Erenstein, O., Henderson, B., Kassie, M., Paul, B. K., Rigolot, C., Ritzema, R. S., Rodriguez, D., Van Asten, P. J. A., & Van Wijk, M. T. (2016). Drivers of household food availability in sub-Saharan Africa based on big data from small farms. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(2), 458–463. https://doi.org/10.1073/PNAS.1518384112

Gao, J., Shi, J., Dong, S., Liu, P., Zhao, B., & Zhang, J. (2017). Effect of different light intensities on root characteristics and grain yield of summer maize (Zea Mays L.). *Scientia Agricultura Sinica*, *50*(11), 2104–2113. https://doi.org/10.3864/j.issn.0578-1752.2017.11.016

Gao, Y. J., Wallach, D., Hasegawa, T., Tang, L., Zhang, R. Y., Asseng, S., Kahveci, T., Liu, L. L., He, J. Q., & Hoogenboom, G. (2021). Evaluation of crop model prediction and uncertainty using Bayesian parameter estimation and Bayesian model averaging. *AGRICULTURAL AND FOREST METEOROLOGY*, *311*. https://doi.org/10.1016/j.agrformet.2021.108686

Gbodjo, Y. J. E., Ienco, D., & Leroux, L. (2021). Benchmarking statistical modelling approaches with multi-source remote sensing data for millet yield monitoring: a case study of the groundnut basin in central Senegal. *International Journal of Remote Sensing*, *42*(24), 9277–9300. https://doi.org/10.1080/01431161.2021.1993465

Giller, K. E., Delaune, T., Silva, J. V., Descheemaeker, K., van de Ven, G., Schut, A. G. T., van Wijk, M., Hammond, J., Hochman, Z., Taulya, G., Chikowo, R., Narayanan, S., Kishore, A., Bresciani, F., Teixeira, H. M., Andersson, J. A., & van Ittersum, M. K. (2021). The future of farming: Who will produce our food? *Food Security*, *13*(5), 1073–1099. https://doi.org/10.1007/s12571-021-01184-6

Gómez-Dans, J. L., Lewis, P. E., Yin, F., Asare, K., Lamptey, P., Aidoo, K. K. Y., Maccarthy, D. S., Ma, H., Wu, Q., Addi, M., Huang, J., & Li, X. (2022). Location, biophysical and agronomic parameters for croplands in northern Ghana. *Earth System Science Data*, *14*(12), 5387–5410. https://doi.org/10.5194/essd-14-5387-2022

Holzkämper, A., Calanca, P., & Fuhrer, J. (2013). Identifying climatic limitations to grain maize yield potentials using a suitability evaluation approach. *Agricultural and Forest Meteorology*, *168*, 149–159. https://doi.org/10.1016/j.agrformet.2012.09.004

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. C., & Vedel, I. (2019). Improving the content validity of the mixed methods appraisal tool: a modified e-Delphi study. *Journal of Clinical Epidemiology*, *111*, 49-59.e1. https://doi.org/10.1016/J.JCLINEPI.2019.03.008

Huang, A., Shen, R., Li, Y., Han, H., Di, W., & Hagan, D. F. T. (2022). A Methodology to Generate Integrated Land Cover Data for Land Surface Model by Improving Dempster-Shafer Theory. *Remote Sensing*, *14*(4). https://doi.org/10.3390/rs14040972

Huang, J., Gómez-Dans, J. L., Huang, H., Ma, H., Wu, Q., Lewis, P. E., Liang, S., Chen, Z., Xue, J.-H., Wu, Y., Wang, J., & Xie, X. (2019). Assimilation of remote sensing into crop growth models: Current status and perspectives. *Agricultural and Forest Meteorology*, *276–277*. https://doi.org/10.1016/j.agrformet.2019.06.008

Ishaq, R. A. F., Zhou, G., Tian, C., Jiang, H., Tan, Y., Jing, G., & Obaid-ur-Rehman. (2024). A Systematic Review of Radiative Transfer Models for Crop Yield Prediction and Crop Traits Retrieval. *Remote Sensing*, *16*(1). https://doi.org/10.3390/rs16010121

Jin, Z., Azzari, G., Burke, M., Aston, S., & Lobell, D. B. (2017). Mapping smallholder yield heterogeneity at multiple scales in eastern Africa. *Remote Sensing*, *9*(9). https://doi.org/10.3390/rs9090931

Jin, Z., Azzari, G., & Lobell, D. B. (2017). Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. *Agricultural and Forest Meteorology*, *247*, 207–220. https://doi.org/10.1016/J.AGRFORMET.2017.08.001

Kalecinski, N. I., Skakun, S., Torbick, N., Huang, X., Franch, B., Roger, J.-C., & Vermote, E. (2024). Crop yield estimation at different growing stages using a synergy of SAR and optical remote sensing data. *Science of Remote Sensing*, *10*. https://doi.org/10.1016/j.srs.2024.100153

Khawaja, H. A. (2017). Solution of pure scattering radiation transport equation (RTE) using finite difference method (FDM). In *Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics: Vol. 10269 LNCS*. https://doi.org/10.1007/978-3-319-59126-1_41

KNBS. (2024). *NATIONAL AGRICULTURE PRODUCTION REPORT*.

Kolipaka, V. R. R., & Namburu, A. (2023). K-Fold Validation of Multi Models for Crop Yield Prediction with Improved Sparse Data Clustering Process. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(10s), 454–463.

Lambert, R. J., Mumm, R. H., & Mansfield, B. D. (2014). Effect of leaf area on maize productivity. *Maydica*, *59*(1), 58–63.

Lee, D., Anderson, W., Chen, X., Davenport, F., Shukla, S., Sahajpal, R., Budde, M., Rowland, J., Verdin, J., You, L., Justice, C., & Meyer, C. (2025). HarvestStat Africa – Harmonized Subnational Crop Statistics for Sub-Saharan Africa. *Scientific Data*, *12*(1). https://doi.org/10.1038/s41597-025-05001-z

Lee, D., Davenport, F., Shukla, S., Husak, G., Funk, C., & Verdin, J. (2024). Contrasting performance of panel and time-series data models for subnational crop forecasting in Sub-Saharan Africa. *Agricultural and Forest Meteorology*, *359*. https://doi.org/10.1016/j.agrformet.2024.110213

Leroux, L., Castets, M., Baron, C., Escorihuela, M.-J., Bégué, A., & Lo Seen, D. (2019). Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. *European Journal of Agronomy*, *108*, 11–26. https://doi.org/10.1016/j.eja.2019.04.007

Li, C., Chimimba, E. G., Kambombe, O., Brown, L. A., Chibarabada, T. P., Lu, Y., Anghileri, D., Ngongondo, C., Sheffield, J., & Dash, J. (2022). Maize Yield Estimation in Intercropped Smallholder Fields Using Satellite Data in Southern Malawi. *Remote Sensing*, *14*(10). https://doi.org/10.3390/rs14102458

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324–333. https://doi.org/10.1016/J.RSE.2015.04.021

Maddonni, G. A., Cárcova, J., Otegui, M. E., & Slafer, G. A. (2024). Recent Research on Maize Grain Yield in Argentina. In *Physiological Bases for Maize Improvement*. https://doi.org/10.1201/9781003578499-11

Mahuku, G., Kimunye, J. N., Prasanna, B. M., Lockhart, B. E., Wanjala, B., Wangai, A., Jones, M. W., Stewart, L. R., Cassone, B. J., Redinbaugh, M. G., Asea, G., & Pappu, H. R. (2015). Maize lethal necrosis (MLN), an emerging threat to maize-based food security in sub-Saharan Africa. *Phytopathology*, *105*(7), 956–965. https://doi.org/10.1094/PHYTO-12-14-0367-FI

Mandapati, R., Gumma, M. K., Metuku, D. R., & Maitra, S. (2024). Field-level rice yield estimations under different farm practices using the crop simulation model for better yield. *Plant Science Today*, *11*(1), 234–240. https://doi.org/10.14719/pst.2690

Mkuhlani, S., Bendito, E. G., Tofa, A. I., Aliyu, K. T., Shehu, B. M., Kreye, C., & Chemura, A. (2024). Spatial and temporal distribution of optimal maize sowing dates in Nigeria. *PLOS ONE*, *19*(5), e0300427-. https://doi.org/10.1371/journal.pone.0300427

Mu, X., Ma, Z., Zhang, L., Fu, J., Ding, Y., Xia, L., Zhang, F., Zhang, J., Qi, J., Zhao, X., Tang, B., & Liu, T. (2022). Responses of photosynthetic fluorescence characteristics, pollination, and yield components of maize cultivars to high temperature during flowering. *Chinese Journal of Eco Agriculture*, *30*(1), 57–71. https://doi.org/10.12357/cjea.20210313

Mumo, L., Yu, J., & Fang, K. (2018). Assessing Impacts of Seasonal Climate Variability on Maize Yield in Kenya. *International Journal of Plant Production*, *12*(4), 297–307. https://doi.org/10.1007/s42106-018-0027-x

Nyombi, K. (2019). Soil data importance in guiding maize intensification and yield gap estimations in East Africa. *Archives of Agronomy and Soil Science*, *65*(6), 809–821. https://doi.org/10.1080/03650340.2018.1528497

Ondiek, R. A., Saber, M., & Abdel-Fattah, M. (2024). Spatial–Temporal Analysis of Impacts of Climate Variability on Maize Yield in Kenya. *Agriculture Switzerland*, *14*(8). https://doi.org/10.3390/agriculture14081415

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Peng, Q., Shen, R., Li, X., Sun, Q., Liu, Y., Huang, J., & Yuan, W. (2024). Spatial and Temporal Patterns of Maize Phenology in China From 2001 to 2020. *Journal of Geophysical Research Biogeosciences*, *129*(10). https://doi.org/10.1029/2024JG008095

Pereira, V., Basilio, M., & Tarjano, C. (2023). *pyBibX -- A Python Library for Bibliometric and Scientometric Analysis Powered with Artificial Intelligence Tools*. https://doi.org/10.48550/arXiv.2304.14516

Prasad, M. (2024). Introduction to the GRADE tool for rating certainty in evidence and recommendations. *Clinical Epidemiology and Global Health*, *25*, 101484. https://doi.org/10.1016/J.CEGH.2023.101484

Quintero, D., Mishra, V., Limaye, A. S., Van Abel, N., Bright Ross, J., & Rashid, A. (2025). Bayesian calibration of management practices for a crop model implemented in a subsistence farming region. *European Journal of Agronomy*, *164*. https://doi.org/10.1016/j.eja.2025.127524

Satpathi, A., Setiya, P., Das, B., Nain, A. S., Jha, P. K., Singh, S., & Singh, S. (2023). Comparative Analysis of Statistical and Machine Learning Techniques for Rice Yield Forecasting for Chhattisgarh, India. *Sustainability*, *15*(3). https://doi.org/10.3390/su15032786

Segarra, J., Araus, J. L., & Kefauver, S. C. (2022). Farming and Earth Observation: Sentinel-2 data to estimate within-field wheat grain yield. *International Journal of Applied Earth Observation and Geoinformation*, *107*. https://doi.org/10.1016/j.jag.2022.102697

Silva, L. G. D. O., & De Almeida-Filho, A. T. (2016). A multicriteria approach for analysis of conflicts in evidence theory. *Information Sciences*, *346–347*, 275–285. https://doi.org/10.1016/j.ins.2016.01.080

Sisheber, B., Marshall, M., Mengistu, D., & Nelson, A. (2024). Assimilation of Earth Observation Data for Crop Yield Estimation in Smallholder Agricultural Systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 557–572. https://doi.org/10.1109/JSTARS.2023.3329237

Skau, E., Truong, D. P., Sentz, K., Armstrong, C., & Gerts, D. (2023). Open World Dempster-Shafer Using Complementary Sets. *Proceedings of Machine Learning Research*, *215*, 438–449.

Song, T., Ma, P., Zhou, Y., Fang, K., & Yang, M. (2018). A model validation method with bootstrap approach and Bayes estimation for small sample. *30th European Modeling and Simulation Symposium, EMSS 2018*, 74–80.

Srivastava, A. K., Mboh, C. M., Gaiser, T., Webber, H., & Ewert, F. (2016). Effect of sowing date distributions on simulation of maize yields at regional scale - A case study in Central Ghana, West Africa. *Agricultural Systems*, *147*, 10–23. https://doi.org/10.1016/j.agsy.2016.05.012

Swain, K. C., Singha, C., & Pradhan, B. (2024). Estimating Total Rice Biomass and Crop Yield at Field Scale Using PlanetScope Imagery Through Hybrid Machine Learning Models. *Earth Systems and Environment*, *8*(4), 1713–1731. https://doi.org/10.1007/s41748-024-00481-2

Tiedeman, K., Chamberlin, J., Kosmowski, F., Ayalew, H., Sida, T., & Hijmans, R. J. (2022). Field Data Collection Methods Strongly Affect Satellite-Based Crop Yield Estimation. *Remote Sensing*, *14*(9). https://doi.org/10.3390/rs14091995

Tonnang, H. E. Z., Makumbi, D., & Craufurd, P. (2018). Methodological approach for predicting and mapping the phenological adaptation of tropical maize (Zea mays L.) using multi-environment trials. *Plant Methods*, *14*(1). https://doi.org/10.1186/s13007-018-0375-7

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., & Kerner, H. (2023). *Lightweight, Pre-trained Transformers for Remote Sensing Timeseries*. http://arxiv.org/abs/2304.14065

Vigani, M., Dudu, H., Ferrari, E., & Mainar, A. (2019). *Estimation of food demand parameters in Kenya A Quadratic Almost Ideal Demand System (QUAIDS) approach*. https://doi.org/10.2760/479781

Vollset, S. E., Goren, E., Yuan, C. W., Cao, J., Smith, A. E., Hsiao, T., Bisignano, C., Azhar, G. S., Castro, E., Chalek, J., Dolgert, A. J., Frank, T., Fukutaki, K., Hay, S. I., Lozano, R., Mokdad, A. H., Nandakumar, V., Pierce, M., Pletcher, M., … Murray, C. J. L. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study. *Lancet (London, England)*, *396*(10258), 1285–1306. https://doi.org/10.1016/S0140-6736(20)30677-2

Zhang, X., & Zhang, Q. (2016). Monitoring interannual variation in global crop yield using long-term AVHRR and MODIS observations. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 191–205. https://doi.org/10.1016/j.isprsjprs.2016.02.010

APPENDIX A: WORK PLAN

PhD Workplan

Legend: Primary Research Activities | Event-Dependent Activities

Start Date: 04/11/2024

Expected End Date: 30/09/2028

Milestone description	Start	Months
PRE-DOCTORAL STUDIES & RESEARCH FOUNDATIONS		
Completion of Predoctoral Coursework	04/11/2024	8
Research Proposal Development and Defence	02/01/2025	7
Systematic Literature Review on EO, CM, and AI Applications	15/05/2025	6
ELLIS Summer School (AI for Earth and Climate Sciences)	01/09/2025	1 wk
METHOD DEVELOPMENT		
Methodology Protocol Development and Approval	08/09/2025	2
Data Acquisition, Model Design and Calibration	01/12/2025	7
Drafting and Submission of Initial Manuscript	01/06/2026	6
Scientific Dissemination (UM6P Doctoral Days & International Conferences)		
SCALABILITY ANALYSIS & AI BENCHMARKING		
AI Model Validation & Benchmarking	01/12/2026	4
Scalability & Transferability Analysis	01/03/2027	4
Drafting and Submission of Scalability & Model Comparison Manuscript	01/06/2027	6
Dissemination of Results (UM6P, IITA, International Forums)		
SYNTHESIS AND THESIS COMPLETION		
Synthesis of Research Findings Across All Objectives	01/12/2027	4
Engagement in Final Knowledge-Sharing and Outreach Activities		
Doctoral Thesis Writing and Defence	01/11/2025	35

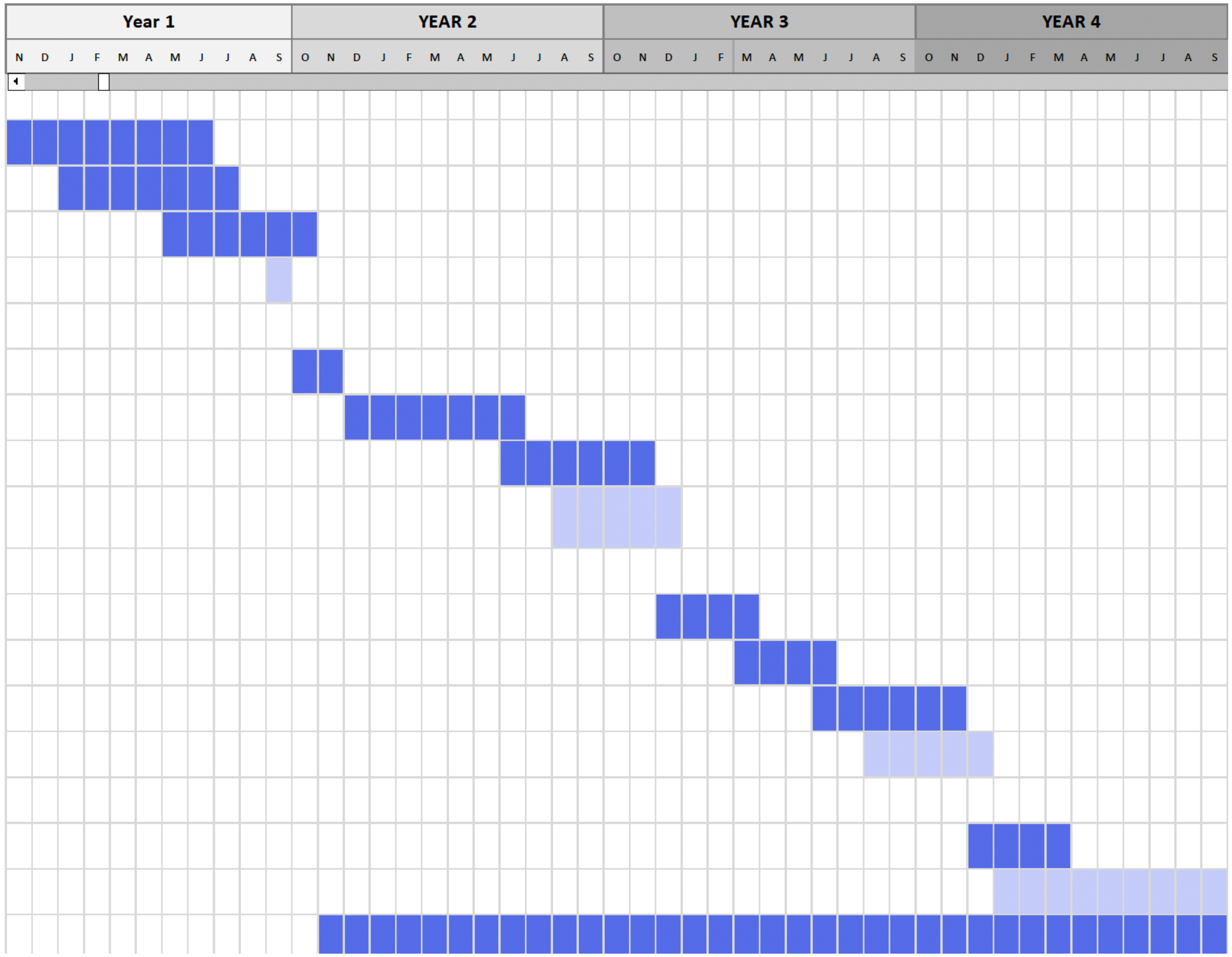

Figure 4: PhD research timeline outlining major activities and milestones from 2024 to 2028