© MAR 2022 | IRE Journals | Volume 5 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV5I9-1713512

AI-Driven Crop Yield Prediction Models for Smallholder Farmers in Sub-Saharan Africa

ASAMOAH OPPONG ZADOK
Department of Agricultural Economics and Extension, School of Agriculture, College of Agriculture and Natural Science, University of Cape Coast, Ghana.

Abstract- The increasing exposure of smallholder farming systems in Sub-Saharan Africa to climate variability has intensified the need for reliable crop yield prediction methods that can support agricultural planning and food security interventions. Conventional yield estimation approaches based on surveys, statistical aggregation, and process-based models have shown limited capacity to capture non-linear crop–climate interactions and localized yield variability, particularly in data-constrained environments. Recent advances in artificial intelligence and machine learning have introduced new opportunities for modeling complex relationships between climate conditions, environmental factors, and crop performance. The review examines how machine learning models incorporate climate variability, the data sources they rely on, the spatial scales at which predictions are generated, and the extent to which these models align with smallholder decision-making contexts. Through comparative analysis, the review evaluates model performance, validation practices, data dependence, and reported limitations. The findings indicate that machine learning models, particularly tree-based approaches, generally outperform traditional statistical methods in capturing non-linear yield responses to climate variability. However, their practical applicability in smallholder contexts remains constrained by data scarcity, coarse spatial resolution, limited validation under real-world conditions, and weak integration of farmer-level constraints. The paper highlights key gaps in the existing literature and emphasizes the need for data-efficient, scale-appropriate, and decision-relevant modeling approaches to improve the utility of AI-driven yield prediction in climate-sensitive smallholder agricultural systems.

I. INTRODUCTION

In Sub-Saharan Africa, agriculture is still a major source of livelihood, food security, and economic activity. Smallholder farming systems prevail in the region, with most farmers operating small parcels of land, usually in rain-fed and having poor access to modern inputs or tools to make decisions. The Food and Agriculture Organization reports that in the region, a significant portion of food production is attributed to smallholder farms, but crop yields continue to be low and extremely unreliable as compared to the world averages (FAO, 2020). This inconsistency threatens the food security and income of households especially in the rural regions that are very reliant on the seasonal agricultural production.

Climate variability is also one of the major yield instability drivers in Sub-Saharan Africa. The alteration of rainfall timing, distribution, and intensity, as well as rise of temperature, have changed the growing conditions in most parts of the region. According to the reports of the Intergovernmental Panel on Climate Change, there are trends of warming observed in Sub-Saharan Africa which are more than the global average and these trends directly affect crop development, water availability and yield outcomes (IPCC, 2021). In smallholder systems where all rainfall is received due to seasonal patterns, agricultural planning in these contexts is becoming more and more uncertain due to the sensitivity of yield to intra-seasonal patterns of rainfall than to total precipitation on an annual basis (Lobell et al., 2019).

Predictable crop yields are generally considered to be a valuable agricultural decision input. On the farm level, yield predictions can be used to set the date of planting, choice and allocation of crop inputs. In addition to the farm, extension services, market planning, food security interventions are supported by predictions of yields. Nevertheless, the traditional techniques of yield estimation currently employed in Sub-Saharan Africa are mostly based on household survey, crop cutting experiments and statistical aggregation. Although these approaches are necessary to obtain the basic background data, they tend to be expensive, time-intensive, and inappropriate to represent the local yield change across the swiftly evolving climate conditions (Burke and Lobell, 2017).

Artificial intelligence and machine learning processes are also becoming more popular in agricultural studies in response to these constraints. Machine learning models provide the ability to capture non-linear relationships between climate variables, soil properties, and crop responses that are challenging to approximate with other traditional regression-based solutions. According to recent reports, random forests, gradient boosting, and neural networks are the methods that can enhance the accuracy of the yield prediction, especially in the environment with a high degree of climate variability and heterogeneity (Lobell et al., 2019; Jin et al., 2021).

Although this continues to develop, there are unanswered questions as to whether the current machine learning-based yield prediction models can be relevant to the smallholder farming system in Sub-Saharan Africa. Most of the studies are constructed based on information of large commercial farms or areas with high levels of observational networks. Such models typically rely on high-resolution satellite photography, elaborate management information, or large voluminous climate data, which are seldom found with smallholders. Additionally, yield forecasts often are created on regional or national levels, which restricts their usefulness to individual farmers who work in highly localized environments due to soil variations, management, and microclimates (Burke and Lobell, 2017).

The literature also has little focus on the translation of yield predictions into actionable decisions made by the smallholder farmers. Although enhanced predictive accuracy is often reported, less research is done on whether the predictions are consistent with the limitations of the smallholders, including restricted access to inputs, credit, irrigation, or timely information. Consequently, the performance of the technical models compared with the actual decision relevance in the real world has not been adequately addressed.

The current paper aims to answer these questions by critically reviewing the machine learning-based crop yield prediction research related to the smallholder farmers in Sub-Saharan Africa. The review focuses on learning about the application of machine learning models under climate variability conditions, the kind of data needed by these models, and the scales of prediction generated by these models. Three questions within the analysis focus the research on the following: How have machine learning methods been applied to predict crop yields in smallholder-based areas? What modeling and data decisions are relevant to the application in smallholder systems? What are the gaps in ensuring that the yield prediction models are applied in smallholder systems?

This paper will explain the potential and the constraints of AI-based yield prediction in Sub-Saharan Africa by synthesizing recent research on the topic through the prism of smallholder constraints and not through the prism of large-scale production systems. The aim is not merely to evaluate predictive performance, but also to determine design features that would facilitate the process of developing yield prediction methods that will be more appropriate to low-data climatic conditions of smallholder farming.
This paper includes a systematic critical review methodology in lieu of a systematic meta-analysis and focuses on analytical synthesis of current peer-reviewed evidence as opposed to comprehensive quantitative synthesis.

II. LITERATURE REVIEW

The field of crop yield prediction has grown at an accelerated rate in the last ten years due to the growing concerns regarding climate change, food security and agricultural sustainability. In this literature, Sub-Saharan Africa takes a unique status because most of their systems are comprised of smallholder farming, high vulnerability to climate fluctuation, and constrained data. Research in the field of yield prediction in the area is based on a variety of fields, such as agronomy, climate science, remote sensing and machine learning, which makes the literature in this area diverse and disjointed. The current literature review is aimed at peer-reviewed articles investigating crop prediction in the context of climate variability, specifically those that are associated with smallholder farming systems in Sub-Saharan Africa and similar areas. The review has been designed in a way that it attempts to establish the influence of climate variability in determining the smallholder crop production and, then, evaluates the modeling methods and data limitations identified in the recent studies.

2.1 Climate Variability and Smallholder Crop Production

Climate variability is among the most powerful variables that determine the outcome of crop production of smallholder farming systems in Sub-Saharan Africa. Smallholder systems unlike large-scale commercial farms are mainly rain-fed and have minimum access to irrigation infrastructure, climate-resilient inputs or formal risk management systems. This leads to the fact that the changes in weather conditions would very likely be directly translated into the variability of yields, which makes agricultural production one of the most uncertain activities despite season changes. Of special importance is the variability of rainfall. There is a lot of evidence indicating that crop production in Sub-Saharan Africa is more responsive to the timing and distribution of rainfall than total seasonal amounts of rainfall.

A specific role is played by rainfall variability. It has been demonstrated that in Sub-Saharan Africa, the effects of rainfall on crop yields are more responsive to the timing and location of rainfall rather than the amount of precipitation received during a season. Lobell et al. (2019) show that the patterns of rainfall during the season, like late arrival of rains or prolonged dry periods during critical growth periods in crops such as maize and sorghum explain a substantial percentage of yield variation. To an extent that the planting decisions of smallholder farmers are usually made on the signal of early rainfall, this variability exposes them to the production risk in case rainfall patterns do not conform to the historical trends.

These challenges are further aggravated by variation in temperature. It has also been demonstrated that increasing mean temperatures and frequency of heat stress events adversely impact crop development, specifically, by hastening phenological stages and reducing grain-filling intervals. Sultan et al. (2019) report empirical data that historical warming has already led to yield losses, measurable in cereal crops, in West Africa even during years when there was no severe drought. Smallholder systems are particularly susceptible to these effects because the farmers are not able to access the heat-tolerant crop varieties or adaptive technologies.

The relationship between rainfall and the variability of temperature provides another problem. The effects of individual climatic factors on yield outcomes are usually affected not by individual factors, but by their interaction through out the growing season. As an illustration, the sufficient rainfall might fail to counter the loss in the yield due to excessive heat in the flowering seasons. These non-linear interactions make it more difficult to represent yield responses with simple statistical methods and emphasize the importance of model schemes with processes that can demonstrate complex climate crop interactions.

The effectiveness of climatic variability on the smallholder crop production is further enhanced by spatial heterogeneity. The interaction of soil properties, topography and local management practices with weather conditions result in an uneven yield response in small geographical regions. The study by Roudier et al. (2021) demonstrates the difference in yield results of farmers working on similar agro-ecological zones even with equal seasonal climatic conditions. Such heterogeneity makes regional averages of yields less useful and introduces difficulties in yield prediction models that are run at coarse spatial scales.

Socioeconomic constraints also determine how smallholders will be vulnerable to climate variability. Lack of access to credit, inputs and extension services also hinders the adaptive capability of farmers to adverse weather conditions. Smallholders are not in a position to replant, irrigate, or change the amount of inputs when rainfall onset is delayed or there is a mid-season drought. Consequently, climate shocks are more likely to produce more serious and persistent effects on the livelihoods of smallholder farmers than on larger systems of agriculture.

These features, in terms of yield prediction, highlight the fact that climate variability is not only a leading determinant of production but also a key source of uncertainty. Prediction models should take into consideration fine-scale temporal variations, local environmental variations, and non-linear crop responses to climatic stress. Otherwise, it may lead to the generation of forecasts, which would underestimate the risk of yield or would not mirror the conditions that affect the use of smallholder farmers.

Climate variability literature is thus a crucial basis upon which the appropriateness of the machine learning-driven yield prediction models in smallhold agricultural settings can be assessed.

2.2 Traditional Crop Yield Estimation Approaches

In Sub-Saharan Africa, agricultural monitoring and planning has traditionally been based on traditional methods of crop yield estimation. These strategies are based mostly on off-the-field methods of data collection, statistical reporting, and econometric modelling, and have remained guiding national agricultural statistics and food security evaluations. Although these approaches offer valuable baseline data, they are especially limited in smallholder dominated systems where there is both high spatial heterogeneity and rising climate variability.

Household survey based yield estimation is one of the most popular methods. National agricultural surveys and studies on living standards measurement use self-reported yield information on the farmers, usually accompanied by a measure of the size of the lot. Such surveys continue to feature prominently in the official statistics of yields in the region. Nevertheless, several articles report high measurement error in self-reported yields, which is caused by recall bias, rounding, and the inability to estimate harvested amounts, especially in smallholder farmers who manage several plots (Carletto et al., 2015; Gourlay et al., 2019). Even though the survey methodologies have been enhanced with time, these inaccuracies have remained and restrict the accuracy of yield estimates at small scales. Experiments with crop cutting are also applied to enhance the accuracy of the results as they provide the actual yields of the chosen plots directly. In the method a sub-section of a field is picked and the mass is calculated to determine the total output. Although crop cutting is thought to be more precise, reliable as compared to self-reporting, it is resource-intensive, expensive, and hard to implement in vast or isolated regions. Furthermore, small-scale potency and intra-field difference that prevail in the smallholder systems can still be able to create bias when sampling is not designed properly (Desiere and Jolliffe, 2018). This means that crop cutting is normally used sparingly and this limits its application in estimating the yields in a timely or wide manner.

Historical climate data have also been used extensively in estimating yields of crops using statistical and econometric models. Such models are typically associated with linear or semi-linear regression models of the relationship between yield results and the amount of rainfall, mean temperature, or growing degree days. These methods have played a leading role in the initial studies of climate impact and continue to play a significant role in policy-based studies. Their assumptions are however very limiting when it comes to smallholder settings. Linear models are not always suitable to understand threshold effects, non-linear crop responses, and climate stress-related interactions between climate variables (Lobell and Burke, 2018).

Another conventional methodology is process-based crop models. These are models of growth simulating crop growth using physiological processes with the addition of climate, soil as well as management parameters. Process-based models are useful in understanding crop-climate interactions but the models need in-depth input data which is not often accessible to smallholder systems in Sub-Saharan Africa. The limited field observations and uncertainties regarding the management practices complicate further the calibration and validation, making them less predictable at the local scales (Challinor et al., 2018).

In all these conventional methods, scale has been an issue. The estimates of yields are usually generated at district, regional or national levels which is the summation of the survey data or model results. Such estimates help in planning on the macro-level, but they hide a lot of within-region disparity, which individual smallholder farmers go through. Burke and Lobell (2017) demonstrate that localized yield losses and variability can be misinformed by spatial aggregation, which does not allow using the traditional estimates to make farm-level decisions.

Another important constraint is timeliness. The survey-based and crop cutting systems tend to record the yield months after the harvest, making them less useful in providing in-season decision support. When climate changes quickly over a growing season, there is little opportunity to respond in an adaptive manner using delayed yield information. This time delay makes the conventional methods even more restrictive as climate uncertainty grows. Collectively, the literature emphasizes that the conventional techniques of crop yield estimation, though they form the basis, are ill-equipped with the requirements of smallholder farming in the modern climate conditions. They are limited in their ability to measure non-linear effects of climate, use coarse spatial resolution, are subject to measurement error, and are delayed in reporting. These constraints have raised a high incentive to develop alternative methodologies, such as machine learning-driven models, that will be able to combine various sources of data and give more dynamic, timely predictions of yields. It is also necessary to learn the advantages and limitations of the traditional approaches to assess the degree to which the newer AI-based approaches can be considered as the true progress, as opposed to the enhancement of the previous ones.

2.3 Machine Learning Applications in Crop Yield Prediction

While climate variability shapes yield outcomes, the ability to predict these outcomes depends largely on the modeling approaches used, particularly the growing application of machine learning techniques. Machine learning-based crop yield prediction has experienced an increasingly fast rate of growth over recent years, and this is mainly due to the weaknesses of traditional statistical and process-based methods in contexts of climate variability and data heterogeneity. Specifically, machine learning techniques are of interest to agricultural systems since they can learn non-linear and complicated connections among climate variables, soil characteristics, and crop reactions without making strong assumptions on the functional form.

One of the most common methods used has been tree-based machine learning models. Random forests and gradient boosting machines are commonly adopted because of their highest capability to work with mixed data types, missing values, and interaction effects between predictors. As demonstrated by Lobell et al. (2019), tree-based models are more effective compared to linear regression in predicting maize yields in fluctuating rainfall conditions, especially in climate-stress years. Such models can model threshold effects, including the loss of yield past-critical temperatures or moisture contents that cannot be easily modeled in traditional models.

Various models combine climate forecasts with remote sensors in order to enhance predictive accuracy. Normalized Difference Vegetation Index is one of the commonly used satellite-derived vegetation indices that can be used as a proxy of crop condition in the growing season. Jin et al. (2021) show how the combination of climate variables and remote sensing inputs enhances the yield prediction accuracy in the Eastern Africa smallholder dominated areas. They also state though that model performance suffers in the case when the satellite data are missing or corrupt, which underlines the weakness of their method in tropical areas due to the constant cloud cover.

Multilayer perceptrons and recurrent neural networks are also neural network models that were used to produce prediction tasks. The models are especially appropriate in models of climate and vegetation capturing of temporal dynamics. According to You et al. (2020), neural network-based models have the potential to outperform simpler machine learning models trained on large and multi-year-long datasets. Nevertheless, their functionality is greatly dependent on data size and regularity. When the historical data are sparse and lumpy, as is the case with smallholder settings, neural networks tend to overfit and have lower generalizability.

In addition to model choice, the model feature selection and model data representation are more important determinants of performance. The lack of uniformity in the summarization of climate variables in studies varies, where seasonal averages of certain variables are used, or intra-seasonal variables such as the date of rainfall onset or heat stress index are considered. The most recent studies consistently indicate that the model with fine-scale time analysis is superior to those based on the aggregate climate metrics, especially in rain-fed models, in which a crop reaction is very dependent on the brief climate patterns (Lobell et al., 2019).

Although there are reported improvements in predictive accuracy, a number of limitations arise throughout literature. To begin with, most models of yield prediction that make use of machine learning are

© MAR 2022 | IRE Journals | Volume 5 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV5I9-1713512

trained and tested on a regional or national level. Though these models might work well at aggregate level, they do not tend to reflect the localities of conditions facing the individual smallholder farmers. Burke and Lobell (2017) note that because of spatial aggregation, within-region variation is likely to be hidden, thus making predictions that are inaccurately related to farm-level results.

Second, there is a significant difference in model evaluation practices. Other researches use random cross-validation which may exaggerate predictive performance on heterogeneous agricultural landscapes. Spatial or temporal cross-validation that represents a more strict test of model transferability is not so commonly used. This brings questions of the strength of claimed accuracy measures in deploying models outside the environment they were trained on (Jin et al., 2021). Last, not all of the research clearly investigates the way machine learning-driven yield predictions can help decision-making among the smallholder farmers. Although the increased accuracy is usually emphasized, little is said about converting the predictions to actionable advice in the context of the limitations of access to inputs, credit or extension services. Consequently, the usefulness of the technical complexity of many models is dubious.

In general, the literature reveals that machine learning methods will provide a significant improvement in crop yield prediction in the conditions of climate variability. Simultaneously, they cannot be so effective in smallholder settings not just because of the performance of their models, but also because of the presence of data, spatial resolution, rigor of validation, and correspondence to real-world decision making. These are some of the points of consideration when assessing the issue of whether AI-based yield prediction models can transition out of the experimental success to the practical significance of the model to the smallholder farmers in Sub-Saharan Africa.

2.4 Data Constraints in Sub-Saharan African Agriculture

One of the most lingering limitations to crop yield analysis in Sub-Saharan Africa is the quality and data availability. These limitations determine not only the conventional methods of yield estimation but also the quality and practicality of machine learning-based systems. In smallholder agrarian systems, the lack of data is due to a conglomeration of infrastructure, environmental, and institutional factors, which leads to the sparse, incomplete, and skewed datasets both spatially and through time.

A significant constraint is related to climate data. The density of weather stations in most of Sub-Saharan Africa is one of the lowest in the world hence significant gaps in ground-based measurements of rainfall, temperature, and other important variables. Consequently, numerous research works use the gridded climate products of satellite measurements or climate reanalysis models. Although these datasets enhance the spatial coverage, they tend to be biased at the local scales especially in areas where the topography is complicated or convective rainy systems are present (Dinku et al., 2018). In the case of smallholder farms, which can be as small as a few hectares, such discrepancies can have a very strong impact on the accuracy of yield predictions. Farm level agricultural data are also limited. Smallholder farmers seldom keep records of planting date, use of inputs, crop type and management methods which are systematic. Surveys, which yield the data on yields, are also susceptible to measurement error and recall error. Gourlay et al. (2019) provide evidence that any error in the farmer-reported production data may significantly misrepresent the yield estimates, especially when the plots are small or when the harvest is intermittent across periods. They limit the quality of training data available to machine learning models and make it more difficult to validate the models. Gap filling Remote sensing data are often utilized to offset gaps in ground-based measurements. A satellite data gives a uniform spatial area and repeated exposures during the growing season. Its applicability by smallholders, however, is limited by a number of factors. The overcast in tropical areas causes unavailability of optical imagery at crucial times of crop development. Also, the spatial resolution of most widely used satellite goods is usually too low to represent small, heterogeneous plots common in smallholder systems (Burke and Lobell, 2017). Such constraints add noise and uncertainty to model inputs and decreases predictability. There is also the issue of data integration. Climate, soil, remote sensing, and yield information are usually measured in varying spatial and temporal scales, with dissimilar methodologies. The process of harmonizing these datasets assumes and entails aggregation processes, which can obscure local variation. According to Jin et al. (2021), when the data resolution is wrong, the model can overestimate its performance when validating the model but hides strengths that may only be revealed once the model is deployed in the field. Data constraints are further enhanced by institutional and logistical constraints. In Sub-Saharan Africa, there are numerous and continuous agricultural data collection initiatives that lead to short time series and poor coverage. Monetary constraint and lack of technical expertise limit long term monitoring and data preservation. These circumstances are a disadvantage to the models that depend on long-term history or regular revision. These data constraints are significant as far as yield prediction is concerned. Models whose training demand large amounts of high-resolution and multi-source data can work well in an experimental context, but poorly when deployed in a low-data environment. Literature is focusing more on the necessity of data-efficient modeling strategies that may work in settings of missing, noisy, or uncertain inputs. These limitations are thus the key factors that need to be understood to assess how relevant machine learning-based models of yield prediction are to smallholders farmers in Sub-Saharan Africa.

2.5 Gaps in Existing Literature

In spite of the fact that there has been a lot of research on crop yield prediction, the available literature presents a number of gaps that restrict the applicability of machine learning-based models to smallholder farmers in Sub-Saharan Africa. These gaps are not limited to data and modeling methods, but also to the conceptualization, assessment and connection between yield prediction and real world decision-making. One of the key gaps is the incompatibility of model development conditions and smallholder reality. Most machine learning models are trained and tested on datasets based on large scale agricultural systems or areas with relatively high levels of observation infrastructure. Although these models can be very predictive in controlled settings, their assumptions regarding the availability of data, management uniformity, and the scale of space are not applicable in most of the smallholder settings (Burke and Lobell, 2017). Subsequently, the amount of reported accuracy improvement might not be converted into quality estimates when models are implemented within data-limited, heterogeneous farming systems. The other technicality is a gap in the scale at which yield predictions are generated and assessed. The creation of regional or national predictions in a large percentage of studies is an output of available data and policy-seeking goals. Nonetheless, decision making among smallholder farmers occurs at the plot or farm level which is highly affected by localized soil conditions, practices in farm management, and microclimates. The literature does not often discuss how regional forecasting can be scaled down in such a manner that it does not lose significant variability or could be used in personal decision-making (Jin et al., 2021). This is one of the most significant obstacles to a realistic adoption as it is a scale mismatch. Another limitation is the performance of models. Most studies make use of random cross-validation method that intermixes observations both in space and time. Although these methods are applicable to benchmarking models, they may overestimate future performance in heterogeneous agricultural environments. Much less often, spatial or temporal cross-validation procedures are used to offer a more rigorous test of model transferability (Lobell et al., 2019). Due to this fact, the reported performance measures might not be a measure of the performance of models in new locations or seasons. A similar challenge is that very little has been done on uncertainty and risk communication. Studies on yield prediction are usually aimed at point estimates of the expected yield, and less on the uncertainty ranges, or measures of confidence. In the case of smallholder farmers who are exposed to high risk of climate, the knowledge of uncertainty is equally important as knowledge of what is likely to happen. The literature offers a comparatively small amount of information on the way that prediction uncertainty ought to be quantified, conveyed, or involved into the decision-making procedures (Roudier et al., 2021). Another significant disconnect between agronomic experience and farm constraints and model design is also present. Although machine learning models are outstanding in recognizing patterns, there is a lot of literature that considers yield prediction as a technical activity and very little attention has been laid on how farmers react to information in reality. Limits to access to inputs, availability of labor, credit conditions amongst other constraints are not often added to modeling projects or assessment criteria. Such a blank undermines the relationship between predictive accuracy and practical utility. Lastly, there is no prospective and participatory assessment in the literature. A majority of studies on yield prediction evaluate the model performance based on historical data in a retrospective manner. Few also study the performance of predictions in the real time, nor the interpretation or utilization of prediction outputs by the farmers.In the absence of such evaluation, it is hard to determine whether machine learning-based systems of predicting the yield can in a meaningful way support the smallholder decision-making in the presence of climate uncertainty (Challinor et al., 2018). Combined, these voids point to the fact that machine learning-based predictive advancement has left gains ahead of adapting models to the context of smallholder agriculture. To overcome these flaws, a change of focus is necessary, moving beyond the maximization of predictive accuracy to the design, analysis, and implementation of data-efficient, scale-relevant, transparent, and decision-relevant models. These gaps were identified and described, which will serve as the reasoning of the analytical focus of this paper and the necessity of more context-dependent methods of AI-based yield prediction in Sub-Saharan Africa.

Table 1. Key gaps in machine learning-based crop yield prediction literature for smallholder systems

Identified Gap	Evidence from Literature	Implication for Smallholder Farming
Scale mismatch between predictions and decision-making	Most models generate regional or national predictions (Burke & Lobell, 2017; Jin et al., 2021)	Limited usefulness for plot-level farm decisions
Heavy reliance on high-resolution data	High-performing models depend on dense climate and satellite data (You et al., 2020)	Poor transferability to data-scarce environments
Limited validation under real-world conditions	Predominant use of random cross-validation (Lobell et al., 2019)	Overestimation of predictive performance
Weak integration of farmer constraints	Minimal consideration of socioeconomic and input constraints (Roudier et al., 2021)	Predictions may not translate into actionable decisions
Limited treatment of uncertainty	Focus on point estimates rather than risk ranges	Reduced usefulness under climate uncertainty

III. METHODOLOGY

This study adopts a structured critical review approach, emphasizing analytical synthesis of selected peer-reviewed studies rather than exhaustive systematic coverage or quantitative meta-analysis. The review-based methodology is suitable because the literature on the topic remains very fragmented, and the methods and types of data used along with the modes of evaluation are very different in different studies. The methodology will be such that it will be transparent in selecting the studies, consistent in its appraisal and clear in its evidence synthesis.

3.1 Literature Search Strategy

Systematic searches of three scholarly databases Web of Science, Scopus, and Google Scholar were used to identify peer-reviewed studies. The selection of these databases was due to the fact that it was broad-based and covered interdisciplinary research on agriculture, climate science, remote sensing, and machine learning. The search was performed through key words such as crop yield prediction, machine learning, climate variability, smallholder agriculture, and Sub-Saharan Africa as a combination.

3.2 Inclusion and Exclusion criteria

The studies that should be included in the review were to satisfy the following criteria. First, the analysis used machine learning or artificial intelligence methods that were close to the prediction of crop yield. Second, the analysis was based on actual agricultural data such as climate data, remote sensing data, or survey data or a combination of both. Third, the research was based on Sub-Saharan Africa or on those areas where the farming systems resembled that of littleholders in scale, availability of data and production status. Research was not considered studies that were very much theoretical, based on simulation data only or when it was carried out on a controlled experimental station without any connection to the smallholder production systems. Studies that used machine learning to classify crops or map them, but not their yield, were also not included. These were some of the criteria used to make sure that the studied reviewed articles were both relevant and contextual.

3.3 Study Selection and Scope

The original search of the database indicated a high number of records in the various fields. The first screening was done to rid titles and abstracts of duplicates and obviously irrelevant studies. Then full-text screening was performed to ascertain the eligibility according to the inclusion and exclusion criteria. The latter group of studies reviewed is a narrow research subsample of the literature that deals with yield prediction in a climate variability setting, within smallholder or smallholder-relevant contexts. The review does not focus on a comprehensive coverage, but instead its emphasis is on the depth of analysis. The focus will be on the studies, which specifically work with the data constraint, spatial scale or the sensitivity of climate because these are the components of smallholder farming systems in Sub-Saharan Africa. Upon the screening of the title, reviewing the abstract, and the full-text, 12 peer-reviewed studies were included in the analysis and became the basis of the review.

3.4 Analysis Framework and Comparison of the study

A comparative framework was used to analyze the selected works in a structured manner. All studies have been considered in the context of four major dimensions, i.e., model type, data inputs, the validation strategy, and limitations reported. Types of models were classified into loose categories such as tree-based models, neural network-based models, and hybrid models. The different data inputs were categorized as to the origin of the data i.e. climate data, remote sensing indicators, soil characteristics, or yield records taken in surveys. Validation practices were evaluated to identify random, spatial, or temporal validation practices used in studies. Much consideration was taken into regard with how validation option influenced claims of predictive performance and model transferability. Instead of attempting to directly compare absolute accuracy measures across studies, which can differ based on crop, region and dataset, the analysis is done based on relative performance trends and common methodological issues.

3.5 Conceptual Review Workflow

Figure 1 enhances transparency by explicitly illustrating how study selection, data inputs, modeling approaches, and evaluation criteria are linked within the review process. The figure shows relationships between climate data, agricultural observations and machine learning models in the reviewed studies, and the relations between yield prediction outputs and farm-level decision situations. It also draws to attention issues where data constraints and scale mismatches bring ambiguity. This abstract presentation gives a clear picture of the review rationale and explains how the evidence of various studies is rationalized.

© MAR 2022 | IRE Journals | Volume 5 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV5I9-1713512

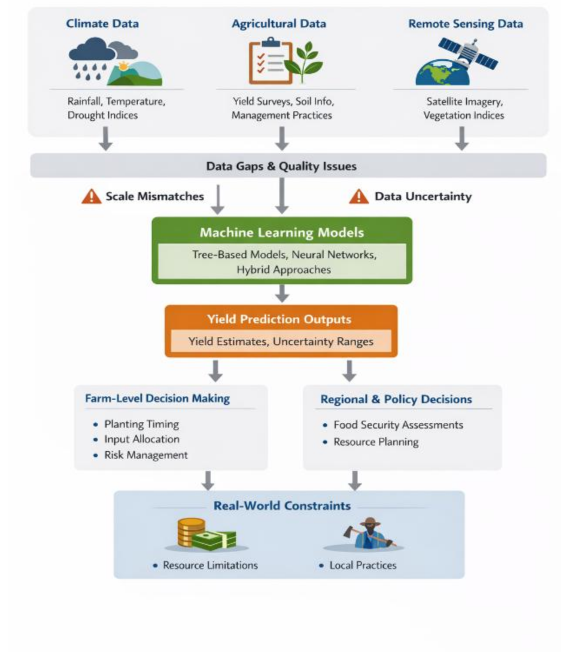

Fig 1: *Conceptual workflow for machine learning-based crop yield prediction in smallholder farming systems.*

The figure illustrates the relationship between climate, agricultural, and remote sensing data; machine learning model development; yield prediction outputs; and decision-making contexts, while highlighting key data and scale constraints relevant to smallholder agriculture.

3.6 Methodological Limitations

A number of limitations are worth considering. To begin with, the review is based on published findings and does not require reanalysis of raw data. This limits the capability of making direct comparisons of model performance between studies. Second, the literature can be affected by publication bias, with researches that demonstrate high performance having larger chances of being published. Third, it emphasizes a particular time period and thereby loses any prior work in the foundations but this sacrifice was needed to preserve analytical consistency. Although these exist, the systematic approach of the review offers a very stringent framework of assessing the applicability of machine learning-based yield prediction models to a smallholder farming system in Sub-Saharan Africa.

IV. RESULTS AND EVALUATION

This paragraph is an empirical synthesis of the results in the reviewed studies with an emphasis on predictive results, sensitivity to climate variability, data reliance, and applicability to smallholder agriculture systems. Instead of introducing new experimental findings, the section critically analyses trends and findings based on existing machine learning-based yield prediction studies.

4.1 Overview of Reviewed Studies

The last batch of the reviewed literature covers a variety of countries in Sub-Saharan Africa and similar smallholder-dominated systems, with a considerable focus on the staple crops, such as maize, sorghum, and mixed cereal systems. The majority of the studies use both climatic variables and either survey-based yield data or the satellite-based indicators of crop condition. The machine learning methods used are tree-based models, neural networks and integration of two or more sources of data. Table 2 gives a comparative analysis of the machine learning-based crop yield prediction literature, where variations in the modeling systems, data needs, validation methods, and constraints are shown in relation to the application in smallholder farming systems

Table 2. Summary of selected machine learning-based crop yield prediction studies relevant to smallholder farming systems

Study	Region and Crop Focus	Modeling Approach	Data Sources	Validation Strategy	Main Findings	Key Limitations
Lobell et al. (2019)	Sub-Saharan Africa; maize and cereals	Random forest; regression	Climate data; satellite indices; survey yields	Cross-validation	ML models reduced prediction error relative to linear	Limited farm-level validation; performance

Study	Region and Crop Focus	Modeling Approach	Data Sources	Validation Strategy	Main Findings	Key Limitations
					regression, especially under climate stress	varied across regions
Burke and Lobell (2017)	Sub-Saharan Africa; mixed crops	Statistical and satellite-based analysis	Remote sensing; climate data	Spatial aggregation analysis	Spatial heterogeneity strongly influenced yield outcomes	Coarse spatial resolution masked plot-level variability
Jin et al. (2021)	Eastern Africa; maize	Tree-based ML models	Climate data; high-resolution satellite imagery; yield observations	Spatial and temporal validation	Combining climate and satellite data improved accuracy	Reduced reliability under cloud cover; scale mismatch
You et al. (2020)	Global datasets including smallholder regions	Deep Gaussian process	Time-series remote sensing data	Cross-validation	Deep learning captured temporal yield dynamics effectively	High data requirements; limited suitability for data-scarce systems
Sultan et al. (2019)	West Africa; cereal crops	Crop–climate modeling and statistical analysis	Climate records; crop model outputs	Model comparison	Historical warming contributed to yield losses	Limited integration with ML; coarse spatial scale

4.2 Predictive Performance of Machine Learning Models

In the literature reviewed, the machine learning models are always superior in crop yield prediction as compared to traditional regression-based methods. Some of the reported improvements consist of a decrease in root mean squared error by about 10 to 20 percent compared to linear models and especially in conditions of high climate variability (Lobell et al., 2019). Random forests and gradient boosting machines, tree-based models, prove to be able to maintain their consistent performance under a variety of environmental conditions and are often recognized as the safest and surest approach in data-limited environments. Models based on neural network demonstrate high-performance in research that has large, high-quality data. According to You et al. (2020) better predictive accuracy is found when deep learning models are established on multi-year remote sensing time series. Nonetheless, performance improvement is less pronounced in smallholder situations where data is sparse or inconsistent. These results imply that the level of complexity of these models does not determine high performance especially when the training data is scarce.

4.3 Sensitivity to Climate Variability

The main advantage of machine learning methods is that they can model non-linear correlation between crops and climate variability. Results with intra-seasonal climate predictors, e.g. the timing of the onset of rainfall or the extent of heat stress, invariably show greater predictive skill than results with seasonal averages. This sensitivity is of special concern to rain-fed smallholder systems, in which yield outcomes may be disproportionately impacted by short-term climate shocks. Lobell et al. (2019) show that machine learning models have lower prediction error in years of drought than traditional ones and are more robust to climate stress. Nonetheless, the effect of different regions and crops differs, meaning that climate sensitivity should be modeled on a case-by-case basis.

4.4 Data Dependence and Model Robustness

In the studies reviewed, it is observed that model performance is highly dependent on the quality and the availability of data. Models based on a combination of climate information and vegetation indices estimated by satellites are typically more accurate than models that utilize climate variables (Jin et al., 2021). Meanwhile, the utilization of remote sensing presents weaknesses of the cloud cover, spatial resolution, and data gaps. Neural networks seem less robust to missing or noisy input but tree-based models appear more robust to those. Such strength renders tree-based approaches especially appealing to smallholder settings, in which the absence of data is a standard.

4.5 Spatial Scale and Farm-Level Relevance

One of the recurring weaknesses throughout the literature is the spatial scale, in which predictions are made. Majority of studies make yield estimates at the district, regional or the national level, according to the resolution of available data. Although these projections are helpful in policy planning and food security observations, they frequently do not reflect variability at an individual level of the smallholder farmers. Burke and Lobell (2017) demonstrate that spatial aggregation may conceal local yields losses and variability and, as a result, predicts that are not well correlated with farm-specific results. Experiments that have tried to validate at the farm level have found that the accuracy would be lower because of the heterogeneity within the fields and differences in management. These results highlight the difficulty in transforming high aggregate accuracy into actionable information to smallholders.

4.6 Evaluation Practices and Transferability

There is a wide range of model evaluation practices in studies. Most of them are based on random cross-validation and may exaggerate predictive performance in non-homogenous landscapes. Less used are spatial or temporal validation methods which offer a more realistic test of model transferability. As Jin et al. (2021) point out, algorithms that are effective when evaluated on random validation do not necessarily work on new areas or times of the year. This is because of the absence of rigorous assessment that restricts the assurance of the extent of generalizability of the reported results. It is hard to evaluate the performance of machine learning based yield prediction models in the conditions of real-life deployment without more powerful validation practices.

4.7 Synthesis of Key Findings

Collectively, the discussed findings suggest that machine learning-based yield prediction models can provide significant benefits over the conventional ones, especially in their capability to explain climate-induced yield variation. These gains are however unequally spread and heavily depend on the availability of data, spatial scale and rigor of validation. Although technical performance is mainly highlighted, as a smallholder farmer, practical relevance is still limited due to inappropriate scale and limited integration with decision making contexts. Figure 1 presented in the Methodology section illustrates the conceptual interaction of the data inputs, machine learning models, and yield predictions among the studies reviewed. The findings below help identify areas in this workflow in which uncertainty and data constraints have the greatest impact on the results of the models.

V. DISCUSSION

In this review, a synthesis of findings shows the potential of as well as the constraints of machine learning-based models of predicting crop yields when used in the context of smallholder agricultural systems in Sub-Saharan Africa. Although the recent results prove that predictive accuracy is better than using the conventional statistical methods, the findings also indicate that those structural issues restrict the feasibility of these models to the smallholder decision-making.

Among the most cohesive to the reviewed studies is the fact that machine learning models are more appropriate than linear methods when considering the ability of the model to capture non-linear responses of crops to climate variability. This is specifically so in smallholder systems which are rain-fed, where yield performance is very sensitive to rainfall timing, heat stress and short-term weather shocks. In particular, tree-based models seem to be flexible and robust enough, as they are not only better at working in heterogeneous environments but also with partial information. These results are consistent with the bigger evidence that crop to climate interactions require non-linear modeling methods to capture the changing climate variability (Lobell and Burke, 2018).

Simultaneously, the discussion helps to understand that the improvement in predictive performance is not necessarily converted into actual practical utility of smallholder farmers. The main problem is that there has been an ongoing discrepancy between the spatial resolution of model outputs and the scale of decision making by farmers. Most yield prediction models are run at regional or district levels, which is the amount of available data that has been resolved. Nonetheless, smallholder farmers have small parcels which vary in soil characteristics, type of crops and practices. Consequently, even models having good aggregate performance can be of little use at the farm level. This scale immatch has been found over and over in the literature to act as an obstacle to efficient agricultural decision support in smallholder settings (Burke and Lobell, 2017).

The dependence of data also determines the relevance of machine learning-based yield prediction models. The most accurate studies are based on the dense climate observations, high-resolution satellite images, or long-term time series. The latter data conditions are hardly achieved in a large part of Sub-Saharan Africa. Where remote sensing data is employed to correct the gaps in ground data, there are problems of cloud cover and coarse spatial resolution which diminish reliability, especially in small plots. These results indicate that the effectiveness of machine learning models in lab experiments could be exaggerated in practice under the conditions of implementation in the real world.

Weaknesses in the current evaluation practices are also indicated in the discussion. Most studies use random cross-validation, which has the propensity to overinflate performance in terms of mixing across space and time. In non-homogenous agricultural areas, this kind of validation offers scanty information on model transferability. The comparative lack of studies that use spatial or temporal validation is a cause of concern that needs to understand how the models will behave when used in different regions or new seasons. This weakness applies particularly to smallholder systems, where variations in climate between years are large, and past trends might not be reliable indicators of future trends (Jin et al., 2021).

The other key problem has to do with the low incorporation of farmer constraints and decision environments in model design and evaluation. Majority of yield prediction research evaluates success largely based on accuracy metrics, and little on how the prediction can be utilized in situations where farmers may be constrained due to either access to inputs, availability of labor and risk associated with finances. Even correct predictions without ignoring this aspect will not help in taking any meaningful action. The literature indicates that yield prediction systems to serve smallholders should no longer be developed around technical performance, but should include interpretability, communication of uncertainty, and relevance to actual decision-making schedules (Roudier et al., 2021).

Combined, the discussion suggests that machine learning-based yield prediction models can be discussed as a significant development in the field of agricultural analytics, yet their present state does not quite correspond to the real-life conditions of smallholder farming in Sub-Saharan Africa. The further development will probably rely more on enhancing data efficiency and reducing the complexity of models, rather than increasing them, matching prediction scales and farm scale needs, enhancing validation practice and integrating socioeconomic constraints into model analysis. These problems need to be solved to ensure AI-based yield prediction can transition beyond technical achievements to effective influence in the smallholder farming sectors.

VI. CONCLUSION

In this paper, a critical analysis of machine learning-based models of crop yield prediction in the context of smallholder farmers in Sub-Saharan Africa. The review examined how these models consider climatic variability, data limitations, and realities of a smallholder farming system.

Evidence incorporated in the current review suggests that machine learning strategies tend to be more effective than conventional statistical techniques at predicting crop yields especially in situations when climate stresses and environment heterogeneity is at play. Random forests models and other tree based models show better ability to model non-linear relationships among climate variables and crop responses. These strengths underscore how machine learning can be used to improve yield forecasting in those regions where climate variability is the primary cause of production risk.

Simultaneously, the review also demonstrates significant weaknesses that restrict the relevance of existing models to the situation of smallholders. Poor spatial resolution, lack of data, and the use of aggregated inputs are still an issue. Predictions of yield at regional or national levels often result in many studies which are not useful in making decision at the level of the farms. Moreover, there is an overstatement of model performance associated with evaluation practices which have limited spatial or temporal validation and this makes the transferability of model across locations and seasons questionable.

There is also general disconnect between technical model development and practical decision support as evidenced in the findings. Little research directly examines the compatibility between yield projections and the constraints associated with smallholder farmers such as access to inputs, credit and access to timely information. Unless more focus is placed on uncertainty communication, interpretability and decision relevance, it is unlikely that any enhancement in predictive accuracy will result in a significant on-farm impact.

Altogether, this review has demonstrated that the focus on model performance metrics and their use should be shifted to consider the context-based design and evaluation. Future studies must focus on data-effective methods, a smaller spatial resolution, more rigorous validation plans, and enhanced incorporation of socioeconomic limitations. To effectively apply machine learning to predict yields as part of climate-resilient smallholder agriculture, these problems should be addressed in Sub-Saharan Africa.

REFERENCES

[1] 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 of the United States of America*, 114(9), 2189–2194. https://doi.org/10.1073/pnas.1616919114

[2] Carletto, C., Gourlay, S., & Winters, P. (2015). From guesstimates to GPStimates: Land area measurement and implications for agricultural analysis. *Journal of African Economies*, 24(5), 593–628. https://doi.org/10.1093/jae/ejv011

[3] Challinor, A. J., Watson, J., Lobell, D. B., Howden, S. M., Smith, D. R., & Chhetri, N. (2018). A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change*, 8(7), 628–635. https://doi.org/10.1038/s41558-018-0208-x

[4] Dinku, T., Cousin, R., Ceccato, P., Connor, S. J., & Thomson, M. C. (2018). Validation of satellite rainfall products over East Africa's complex topography. *International Journal of Remote Sensing*, 39(13), 4140–4168. https://doi.org/10.1080/01431161.2018.1430399

[5] Food and Agriculture Organization of the United Nations. (2020). *FAOSTAT statistical database*. https://www.fao.org/faostat

[6] Gourlay, S., Kilic, T., & Lobell, D. B. (2019). Could the debate be over? Errors in farmer-reported production and their implications for the inverse scale–productivity relationship in Uganda. *Journal of Development Economics*, 141, 102376. https://doi.org/10.1016/j.jdeveco.2019.102376

[7] Intergovernmental Panel on Climate Change. (2021). *Climate change 2021: The physical science basis*. Cambridge University Press. https://www.ipcc.ch/report/ar6/wg1/

[8] Jin, Z., Azzari, G., Burke, M., Aston, S., & Lobell, D. B. (2021). Mapping smallholder yield heterogeneity at scale in Eastern Africa. *Remote Sensing of Environment*, 253, 112170. https://doi.org/10.1016/j.rse.2020.112170

[9] Lobell, D. B., & Burke, M. (2018). Environmental impacts on crop yield. In D. B. Lobell & M. Burke (Eds.), *Climate change and food security* (pp. 31–54). Springer. https://doi.org/10.1007/978-94-024-1442-5_2

[10] Lobell, D. B., Azzari, G., Marshall, B., Gourlay, S., Jin, Z., Kilic, T., & Murray, S. (2019). Eyes in the sky, boots on the ground: Assessing satellite- and ground-based approaches to crop

yield measurement. *Remote Sensing of Environment*, 216, 582–595. https://doi.org/10.1016/j.rse.2018.07.010

[11] Roudier, P., Muller, B., D'Aquino, P., Roncoli, C., Soumaré, M. A., Batté, L., & Sultan, B. (2021). The role of climate forecasts in smallholder agriculture: Lessons from West Africa. *Climate Services*, 22, 100232. https://doi.org/10.1016/j.cliser.2021.100232

[12] Sultan, B., Defrance, D., & Iizumi, T. (2019). Evidence of crop production losses in West Africa due to historical global warming. *Science Advances*, 5(2), eaav0537. https://doi.org/10.1126/sciadv.aav0537

[13] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2020). Deep Gaussian process for crop yield prediction based on remote sensing data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01), 455–462. https://doi.org/10.1609/aaai.v34i01.5407