TIME SERIES AND ENSEMBLE MODELS FOR FORECASTING TANZANIAN BANANA CROP YIELD UNDER THE EFFECTS OF CLIMATE CHANGE

Sabas Patrick

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master's in Mathematical and Computer Sciences and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

July, 2024

ABSTRACT

Amid escalating global worries about climate change's impact on agriculture, this study thoroughly explores how climate shifts might affect Tanzania's vital bananas. The study employed a multiple regression model to analyze the correlation between bananas and key climate variables in Tanzania, the results showed gradual decrease in bananas. Additionally, the study utilized two powerful global sensitivity analysis methods, Sobol' Sensitivity Indices and Response Surface Methodology, to comprehensively explore the sensitivity of bananas to climate variables. So, these methods showed that minimum temperature, precipitation and soil moisture have the most impact on bananas and affect the crop's production variability. Furthermore, uncertainty quantification was performed using Monte Carlo simulation, estimating uncertainties in regression model parameters to enhance the reliability of findings, this indicated substantial variability in the predictions. Conversely, the study configured time series models such as Seasonal ARIMA with Exogenous Variables (SARIMAX), State Space (SS), and Long Short-Term Memory (LSTM) to forecast bananas in Tanzania under the effects of climate change. Hence, the study builds predictive frameworks capturing temporal variations and offering glimpses of future trends. Leveraging historical bananas data and relevant climate variables, an ensemble model was formulated using a weighted average approach, revealing a future decrease in bananas. This study combines data analysis and advanced models to explore how climate change affects bananas. Its insights reach beyond farming, impacting stakeholders, policymakers, and farmers alike. By understanding sensitivities, vulnerabilities, and future trends, this research informs decisions for sustainable banana production, enhances food security, and encourages adaptable strategies amidst changing climates.

DECLARATION

I, **Sabas Patrick**, declare to the Senate of Nelson Mandela African Institution of Science and Technology that this dissertation is my own work and it has not been presented and will not be presented to any other Course of Study for a similar or any other Degree Award. I confirm that appropriate credit has been given where reference has been made to the work of others.

Sabas Patrick **(Name and Signature of the Candidate)**	**Date**

The affirmation stated above is verified:

Dr. Silas Mirau **(Name and Signature of Supervisor 1)**	**Date**
Dr. Isambi Mbalawata **(Name and Signature of Supervisor 2)**	**Date**
Dr. Judith Leo **(Name and Signature of Supervisor 3)**	**Date**

COPYRIGHT

CERTIFICATION

The undersigned certifies that they have read and hereby recommend for acceptance by the Nelson Mandela African Institution of Science and Technology a dissertation entitled: *"Time Series and Ensemble Models for Forecasting Tanzanian Banana Crop Yield under the Effects of Climate Change"* in Partial Fulfillment of the Requirements for the Degree of Master's in Mathematical and Computer Sciences and Engineering of the Nelson Mandela African Institution of Science and Technology.

Dr. Silas Mirau **(Name and Signature of Supervisor 1)**	**Date**
Dr. Isambi Mbalawata **(Name and Signature of Supervisor 2)**	**Date**
Dr. Judith Leo **(Name and Signature of Supervisor 3)**	**Date**

ACKNOWLEDGEMENTS

First I thank God, that this was not an easy task to accomplish without the blessings from Him, the Nelson Mandela African Institution of Science and Technology (NM-AIST), as an academic institution, is where I anticipate gaining the knowledge necessary for my scholarly career, the Government of Tanzania, in particular Higher Education Students' Loans Board (HESLB), and the West African Science Service Centre on Climate Change and Adapted Land Use (WASCAL - RUFORUM) Capacity Building in Agriculture at NM-AIST under the RAINCA Project for their financial support in this study, and I am grateful to the Department of Applied Mathematics and Computational Science at NM-AIST for imparting scholars with proficient knowledge and skills through the educational courses offered, enabling eligibility and growth.

I would like to express my appreciation to my supervisors, Dr. Silas Mirau, Dr. Isambi Mbalawata, and Dr. Judith Leo for their invaluable guidance, assessment, and recommendations throughout the course of this dissertation. Heartfelt gratitude is extended to the community and all the scholars at Nelson Mandela African Institution of Science and Technology, especially my fellow classmates in the Applied Mathematics and Computational Science program, for their genuine concern and valuable companionship during my time there. While not overlooking those who have contributed to this study in various ways, I sincerely appreciate their support. May they all be blessed by God as a whole.

DEDICATION

This work is dedicated to my beloved ones; my late father Mr. Patrick Rwiza, we pray for you, rest in peace; my mom Ms. Sylivia Kiita, we love you always; my lovely wife Ms. Atubela Mtafungwa, and our children Aniela, Adeola and Arcangela, I'm proud of you always.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS

ACRONYM	DEFINITION
AIMS	African Institute for Mathematical Sciences
AMCS	Applied Mathematics and Computational Sciences
ANN	Artificial Neural Networks
ARIMA	Auto Regressive Integrated Moving Average
CoCSE	Computational and Communication Sciences and Engineering
CRU	Climatic Research Unit
FAOSTAT	Food and Agriculture Organization Corporate Statistical Database
LSTM	Long Short-Term Memory
MATLAB	Matrix Laboratory and/or Software
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NM-AIST	Nelson Mandela African Institution of Science and Technology
PYTHON	Programming Language and/or Software
RAINCA	Responsible Artificial Intelligence Network for Climate-Action in Africa
RNN	Recurrent Neural Networks
SARIMAX	Seasonal ARIMA with Exogenous Variables
SD	Standard Deviation
SI	Sensitivity Index
SS	State Space
WASCAL	West African Science Service Centre on Climate Change and Adapted Land Use
WCRP	World Climate Research Program

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Among the biggest flowering herbaceous trees is the banana (*Musa spp.*) plant (Ighalo & Adeniyi, 2019; Jayasinghe *et al*., 2022; Lal *et al*., 2017). Ripe bananas are soft fruits that have a lifespan of 5 to 10 days and are ready for use and consumption. The unripe fruit, leaves, inflorescence, stem, and rhizome of the banana plant are also used in numerous ways as vegetables, food, and animal feeds (Jayasinghe *et al*., 2022; Lai & Dzombak, 2020). In the world's top ten crops in terms of productivity, planted area, and calories produced are bananas (Varma & Bebber, 2019). The banana crop is the fourth most significant crop in the world, behind maize, rice, and wheat, for producing food and income for more than 30% of the global population (Lucas & Jomanga, 2021). The banana is beneficial to human health in all areas and has many traditional and medicinal uses. While the leaf of the banana is consumed as a vegetable in many parts of India, the fruit is a great source of nutrients (Lal *et al*., 2017). After Uganda, Tanzania is the second-largest producer of bananas in East Africa (Lucas & Jomanga, 2021). Banana farming plays a crucial role in Tanzania's agricultural sector by contributing to economic growth and food security (Lucas & Jomanga, 2021; Varma & Bebber, 2019).

Global issues like climate change pose serious risks to agricultural systems all over the world (Hoque & Haque, 2016). Tanzania, along with other nations that depend heavily on crop agriculture for both economic stability and subsistence, is especially susceptible to the potential impacts of changing climatic patterns (Mayaya, 2015; Shirima & Lubawa, 2017; URT, 2021). The banana is one of the key crops for Tanzania's agricultural industry, as it is essential to both food security and livelihoods (Lal *et al*., 2017; Lucas & Jomanga, 2021; Varma & Bebber, 2019). Understanding how banana crop yield is impacted by climate change is essential for developing effective adaptation and mitigation strategies that safeguard food production and financial success (Varma & Bebber, 2019).

Although processing bananas has many advantages for science and technology (Jayasinghe *et al*., 2022; Lal *et al*., 2017), it is striking to note that, in spite their vital role in subsistence and commerce, bananas are given relatively little weight in global assessments of the potential effects of climate change on food security and nutrition (Varma & Bebber, 2019). The impacts of climate change on crop production are diverse, and they pose a growing threat to the sustainability and productivity of banana crops (Chowhan *et al*., 2016). The region is confronted with significant hazards to crop yield and overall agricultural productivity as a result of temperature increases, altered rainfall patterns, and an increased incidence of extreme weather events

(Hoque & Haque, 2016). Tanzania is among the countries globally currently coping with the extreme consequences of climate change (Shirima & Lubawa, 2017), farmers face a variety of challenges that impede the growth and development of the agricultural sector (Lokupitiya, 2018). Accurate and trustworthy forecasting models are crucial for ensuring that banana farming is resilient and adaptable to shifting climatic conditions (Varma & Bebber, 2019).

1.2 Statement of the Problem

In particular, when it comes to multidisciplinary studies, Tanzania has seen very little research on the effects of climate change (Kahimba *et al*., 2015). As a result, it is difficult to accurately estimate the possible effects on the region's food security and banana crop productivity (Lucas & Jomanga, 2021). Furthermore, Tanzania's socioeconomic circumstances and geographic location present particular risks and vulnerabilities for a nation so dependent on agriculture, particularly the banana industry (Shirima & Lubawa, 2017). Tanzania is a fascinating and relevant case study to examine the possible effects of climate change on food security and the productivity of a major crop like bananas because of these factors. As a major export and staple food for millions of Tanzanians, bananas directly impact the country's economy and rural communities' well-being (Lucas & Jomanga, 2021).

In light of the unique difficulties brought on by climate change, this study contribute significantly to the existing knowledge gap in the area of banana crop yield forecasting in Tanzania. By pinpointing the possible effects of climate variables on banana crop productivity, we can offer important insights into the sector's vulnerabilities and adaptability (Wood *et al*., 2014). For efficient agricultural planning, resource allocation, and policy-making, banana crop yield forecasting is essential. This research equips farmers, policymakers, and stakeholders with valuable insights to make informed decisions and implement appropriate strategies to mitigate the adverse effects of climate change (Varma & Bebber, 2019).

1.3 Rationale of the study

The implications of this research are wide-ranging, as the findings from the analysis will assist stakeholders, policymakers, and farmers in making decisions about sustainable banana production and food security in Tanzania (Kahimba *et al*., 2015). Effective strategies to reduce negative effects and increase resilience in the face of future climate uncertainty can be developed by understanding how vulnerable banana cultivation is to climate change (Wood *et al*., 2014). The study's findings add to the understanding of how climate change impacts agriculture and emphasize the necessity of taking proactive, flexible measures to safeguard the region's agricultural livelihoods (Lucas & Jomanga, 2021).

In light of Tanzania's increasingly unpredictable climate, this research aims to offer useful guidance for guaranteeing the long-term sustainability and prosperity of banana cultivation in the country through the use of evidence-based data and thorough scientific analysis (Abdoussalami *et al*., 2023; Kahimba *et al*., 2015). Through tackling the urgent issues posed by climate change on banana crop productivity, we can collaborate to guarantee food security, economic stability, and environmental resilience in Tanzania and other regions (Lokupitiya, 2018; Wood *et al*., 2014).

1.4 Research objectives

1.4.1 General objective

This study's primary goal is to utilize time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change.

1.4.2 Specific objectives

To achieve the overarching objective, the study pursued the following specific objectives:

(i) To assess the impact of climate change on banana crop yield in Tanzania.

(ii) To determine the sensitivity of banana crop yield to climate variables.

(iii) To develop time series models that can accurately forecast banana crop yield in Tanzania under the effects of climate change.

(iv) To develop an ensemble model that can improve the accuracy of banana crop yield forecasting.

1.5 Research questions

The study addressed its objectives through proposed and answered research questions:

(i) How has banana crop yield changed over time in Tanzania in response to climate change?

(ii) How do changes in climate variables affect the banana crop yield?

(iii) Which model will best forecast banana crop yield under the effects of climate change?

(iv) Which ensemble method most effectively enhances banana yield forecasting accuracy?

1.6 Significance of the Study

This study holds significant implications for various stakeholders:

(i) The aim is to assist the Tanzanian government, agriculture, academia, research disciplines, non-governmental organizations, scholars, and other stakeholders by raising awareness about the challenges tied to banana crop yield and climate variables. The study emphasizes empowering regional, district, and village agricultural entities to make informed decisions regarding climate change impacts on banana crop yield.

(ii) Additionally, it endeavors to improve practices and strategies to ensure sustainable banana production and food security, not only in Tanzania but globally.

1.7 Delineation of the Study

This study explores the complex relationship between crop yields of bananas in Tanzania and climate change. It seeks to comprehend how this crucial agricultural product is affected by changing climatic patterns. The research takes a two-fold approach. First dimension is correlation analysis; the study examines the relationship between banana crop yield and important climate variables, such as precipitation, soil moisture, minimum temperature, maximum temperature, and relative humidity. A strong multiple regression analysis reveals weak points and sensitive areas in this relationship. The second dimension is forecasting models; time series models such as SARIMAX, SS, and LSTM are used in this study to forecast future banana yields in the context of changing climates. While an ensemble model incorporated to improve prediction accuracy of these forecasting models.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, the study synthesized and analyzed the scholarly contributions that have shaped the trajectory of this research topic. By critically engaging with diverse viewpoints, the study endeavored to refine the research framework and establish a solid foundation upon which the study was unfold.

2.2 Review of Empirical Studies

2.2.1 Correlation Analysis

According to Lucas and Jomanga (2021), Tanzania has the highest per-person consumption rate in the world, ranging from 280 to 500 kg. Despite the crop's significance, in the 1960s, production was 18 t/ha; in 2016, it was 5-7 t/ha. This research used papers review method through online sources in order to give researchers and agriculture extension officers quick access to resources for better banana production. Through the analysis of risks and opportunities, this review evaluated the state of banana production in Tanzania.

Salvacion (2020) examined the impact of climate on banana yield in the Philippines. The study utilized yield data at the provincial level and considered various climatic factors (such as yearly rainfall, wet day frequency, precipitation seasonality, mean annual temperature, temperature seasonality, and mean annual diurnal temperature range) extracted from the CRU dataset spanning 1991 to 2016. Linear trend analysis informed their approach, and they employed multiple regression analysis to evaluate how climate variables affected banana yield at the provincial level. Notably, temperature-related variables, particularly temperature seasonality, exerted a stronger influence on provincial-level banana yield in the Philippines compared to rainfall.

Bhausaheb *et al*. (2023) conducted an investigation to quantify how climate change influenced the yield of bananas in the Thiruvananthapuram district of Kerala, India. They employed a multiple linear regression model specific to the Thiruvananthapuram district to assess the climate change impact on banana production. Quarterly climatic data covering temperature, rainfall, relative humidity, and wind speed over a 31-year span from 1991 to 2021 were used as independent variables, while banana production data from Thiruvananthapuram served as the dependent variable. Multiple regression analysis was applied to ascertain growth trends and variations. Their findings indicated that climate change had a positive impact on banana production in the Thiruvananthapuram district.

Anzures *et al*. (2022) investigated the limitations affecting the initial cultivation of bananas in the Davao Region, Philippines. They sourced climatic parameters (such as mean temperature, rainfall quantity, and relative humidity), instances of Panama disease, and data on governmental assistance from 1990 to 2019 through government organizations. The study employed techniques like multiple regression analysis and various statistical examinations. The primary aim was to provide actionable insights that could equip farmers with readiness to confront the detrimental challenges of climate change impacting their crops.

Varma and Bebber (2019) analyzed how climate change has affected banana plantations around the world. It was intended to statistically match recorded yearly banana production statistics to the current mean temperature and annual total precipitation. They used a modified version of the beta function that was also employed in earlier studies on the physiology of bananas.

Sabiiti *et al*. (2016) used time series moments, correlation analysis, and regression analysis to assess the historical links between variations in rainfall and temperature over Uganda and banana production empirically (1971-2009). The study discovered excellent comparability in moment indices, with some noticeable differences in the values of the banana yields, rainfall, and temperature moment indices.

Hoque and Haque (2016) examined how the climate affects crop production systems and made recommendations for suitable coping mechanisms and adaptation methods for enhancing coastal agriculture for enhanced agricultural output in Bangladesh's coastal saline areas. To accomplish the goals, data were gathered utilizing a pre-tested interview schedule and a thorough survey of 240 randomly chosen sampled respondents. Long-term statistics and research on climate change revealed a tendency of rising temperatures and irregular precipitation.

2.2.2 Forecasting Models

Jayasinghe *et al*. (2022) conducted a comprehensive analysis involving 75 full-text articles published from 1985 to 2021, focusing on mathematical models associated with banana forecasting. The study highlighted the indispensability of mathematical models for strategic planning and predictive purposes. However, it pointed out the relative scarcity of models specifically related to banana crops and the lack of extensive reviews of prior modeling endeavors, underscoring the demand for evidence-driven investigations in this domain. Notably, the 'SIMBA' banana process-based simulation model and ANN emerged as dependable tools for predicting banana plant growth. The review revealed a dearth of comprehensive data concerning mathematical models for banana fiber yield. Furthermore, it was noted that researchers commonly leaned on multiple linear regression models for estimating banana plant growth and fruit yield, which hampers the potential for meaningful comparisons and optimal model selection.

Lai and Dzombak (2020) used the ARIMA methodology (Auto Regressive Integrated Moving Average). On the other hand, this work used local historical measurements to develop a statistical time series forecasting technique for estimating regional temperature and precipitation over the short term (20 years). Forecasts of annual precipitation and temperature were quantitatively compared and derived from a linear trend approach and the ARIMA model. They came to the conclusion that even though it cannot outperform all other techniques for all investigated climate variables, the ARIMA model typically provides more accurate projections, particularly in interval forecasts, and is more trustworthy than other standard statistical techniques.

Meeradevi *et al*. (2022) employed advanced modern data science methods to facilitate proactive prediction of future crop prices and yields. The objective was to establish a user-friendly hybrid decision support system that anticipates both crop prices and yields. This was achieved by combining ARIMA and LSTM forecasting techniques and leveraging historical data. The resulting system offers farmers an enhanced understanding of a crop's market value through graphical representations of forecasted price and yield values.

Bhimavarapu *et al*. (2023) introduced an enhanced optimization function (IOF) aimed at achieving precise predictions. They integrated this improved IOF into the long short-term memory (LSTM) model. Historical data, manually collected from local agricultural departments spanning 1901 to 2000 for training and 2001 to 2020 from government websites of Andhra Pradesh, India, for testing, were employed. The findings highlighted that the proposed IOF in conjunction with LSTM presents a noteworthy advantage in accurately predicting crop yields. The observed reduction in RMSE for the proposed model underscores its superior performance compared to CNN, RNN, and LSTM in predicting crop yields.

Rathod and Mishra (2018) carried out a study to predict the mango and banana yield in Karnataka, a variety of classes of linear and nonlinear, parametric and non-parametric statistical models were used. The problem with linear models is that they are typically expected to have a linear form. Since time series models usually include both linear and nonlinear elements, they are not always completely linear or completely nonlinear. An innovative hybrid model comprising linear and nonlinear models was presented to solve this issue. The SVM, ARIMA, and ANN models were combined in the hybrid model. In comparison to other models, the Vector Regression model performed better during both model creation and model validation.

Chi and Chi (2021) furthered the study of the time series data and illustrated the role of the time series model in the prediction process using long-term records of the monthly global price of bananas from January, 1990 to November, 2020. They chose ANN and ARIMA models. The results demonstrated that compared to the ARIMA model, the ANN model fared better.

Moore and Lobell (2014) undertook research to evaluate the possible efficacy of individual farmer adaptation within Europe. This was achieved by concurrently estimating short-term and long-term response functions, utilizing variations in yield and profit data across different regions and time periods. The study also incorporated 2040 projected yields and farm profits through analysis of a diverse set of 13 climate model simulations. The findings revealed that the pace at which farmers adjust to increasing temperatures constitutes a notable element of uncertainty in the scenario.

Bertsimas and Boussioux (2023) introduced a novel technique to create resilient ensembles of models for time series forecasting. Their methodology employed Adaptive Robust Optimization (ARO) to establish a linear regression ensemble with adaptable model weights that evolve with time. The efficacy of their approach was demonstrated through synthetic experiments and real-world instances like air pollution control, energy consumption prediction, and tropical cyclone intensity forecasting. The outcomes highlighted that their adaptive ensembles consistently outperformed the top individual ensemble member retrospectively, showcasing a reduction in root mean square error by 16-26% and in conditional value at risk by 14-28%, while also displaying improvements compared to competitive ensemble methods.

The reviewed studies and the associated questions raised by various researchers, as discussed in this Chapter Two, are important to this study because both concentrate on crop yields and climate variables. However, this study conducted to go further by utilizing time series and ensemble models to anticipate banana crop yield under climate change in Tanzania because banana crop yield and climate variables fluctuate from time to time. Thus, this study can contribute significantly to the existing critical knowledge gap in the area of banana crop yield forecasting in Tanzania, taking into account the unique obstacles presented by climate change.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Introduction

In this chapter, the study outlines the techniques and steps employed to achieve the research objectives. This includes data description, methodology, correlation analysis, sensitivity analysis, uncertainty quantification, and the models utilized in conducting the study.

3.2 Data Description

The monthly climate variables that the study gathered from multiple sources were converted into annual data for every year during the analysis. The thorough analysis of the effects of these variables on banana crop yield was made easier by this conversion, which let us work with annual averages. The monthly gridded temperature, minimum temperature, and maximum temperature datasets for the reanalysis were provided by the University of East Anglia's Climatic Research Unit (CRU). These datasets were available for free download from the website at (CRU, 2023). The land surface is covered by the CRU dataset version 4.05 (CRU TS 4.05) with a resolution of $0.5° \times 0.5°$ for the years 1961 to 2020.

Several published studies have used CRU dataset to investigate East African precipitation variability by contrasting it with the World Climate Research Program's (WCRP) monthly precipitation dataset from the GPCC (Ongoma *et al*., 2019). The results of the study consistently showed that the CRU dataset was more dependable and efficient for analysis. Additionally, earlier studies effectively used Tanzania's rainfall data from the Climatic Research Unit (CRU) dataset (Mbigi & Xiao, 2021).

The NCEP/NCAR Reanalysis dataset, which can be downloaded from the website at (NCEP/NCAR, 2023), provided the soil moisture and relative humidity data. The soil moisture dataset has a resolution of $0.25° \times 0.25°$, while the relative humidity dataset has a precision of $2.5° \times 2.5°$ (Anwar *et al*., 2019). The average annual banana crop yield statistics used in this study were sourced from the FAOSTAT database, which is available at (FAOSTAT, 2023). The units of measurement for each variable used in this study are displayed in Table 1.

Table 1: The dataset variables used in this study

S/N	Variable	Unit of Measurement
1.	Precipitation	mm
2.	Minimum temperature	°C
3.	Maximum temperature	°C
4.	Relative humidity	%
5.	Soil moisture	fraction
6.	Banana crop yield	(t / ha)

3.3 Methodology

The objective of this research is to investigate any potential vulnerabilities related to climate change for Tanzania's major crop, bananas, and to assess how the changing climate is affecting the production of bananas. This study investigates the relationships between important climate variables, such as soil moisture, minimum and maximum temperatures, relative humidity, and precipitation, and banana crop yield using a robust multiple linear regression model. Using this method, the study determines which climate factors have the greatest influence on the production of bananas (Bhausaheb *et al*., 2023). This study uses two potent global sensitivity analysis techniques, Sobol Sensitivity Indices and the Response Surface Methodology, to increase the knowledge of how do changes in climate variables like temperature, precipitation and relative humidity affect the banana crop yield (Iooss & Lemaître, 2015). The selection of these methodologies was driven by various factors, including the characteristics of the data, the study's objectives, and the resources at hand (Box *et al*., 2015; Hyndman & Athanasopoulos, 2018). With the aid of these techniques, the study was able to evaluate quantitatively the relative significance of each climate variable in influencing variations in banana crop yield (Borgonovo & Plischke, 2016; Rahn *et al*., 2018). Furthermore, this study uses Monte Carlo simulation to address uncertainty in regression model parameter estimates, which improves the robustness of the findings and offers a probabilistic representation of the possible outcomes (Antanasijević *et al*., 2014; Dega *et al*., 2023; Li *et al*., 2016; Rahn *et al*., 2018).

Time series analysis and ensemble modeling have become increasingly effective methods for forecasting agricultural crop yields in recent years (Kamir *et al*., 2020). By using historical data to spot trends, patterns, and seasonality in crop yield, time series analysis makes it possible to create forecasting models (Box *et al*., 2015). On the other hand, ensemble modeling increases accuracy and robustness by combining the advantages of different forecasting models (Bertsimas & Boussioux, 2023). With a particular focus on the effects of climate change, the goal of this study was to forecast Tanzania's banana crop yield using time series and ensemble models. The study aimed to develop forecasting models that capture the dynamics

of banana productivity under changing climate conditions by incorporating pertinent climate variables and historical data on banana crop yield (Pham *et al.*, 2019). Advanced analytical techniques are necessary because conventional forecasting methods frequently fail to capture the complex interactions between climatic variables and crop yield (Bertsimas & Boussioux, 2023; Varma & Bebber, 2019). Time series analysis and ensemble modeling combined can improve forecast accuracy and reliability, which will help the agriculture sector make better decisions (Kourentzes *et al.*, 2014).

Furthermore, under the influence of climate change, the combination of time series and ensemble modeling techniques offers promising opportunities for reliable and accurate forecasting of banana crop yields (Bertsimas & Boussioux, 2023). So, three forecasting models, including SARIMAX, LSTM, and State Space, were utilized to forecast banana yields in the face of climate change. The results of the data exploration and regression analysis guided the selection of these methods (Box *et al.*, 2015; Hyndman & Athanasopoulos, 2018; Jayasinghe *et al.*, 2022). The study then used a weighted average technique to formulate the ensemble model. More specifically, because it is simple to apply in practical settings, the use of weighted linear combinations of different ensemble members has grown in popularity (Bertsimas & Boussioux, 2023). The overall workflow is depicted in Fig. 1's schematic diagram:

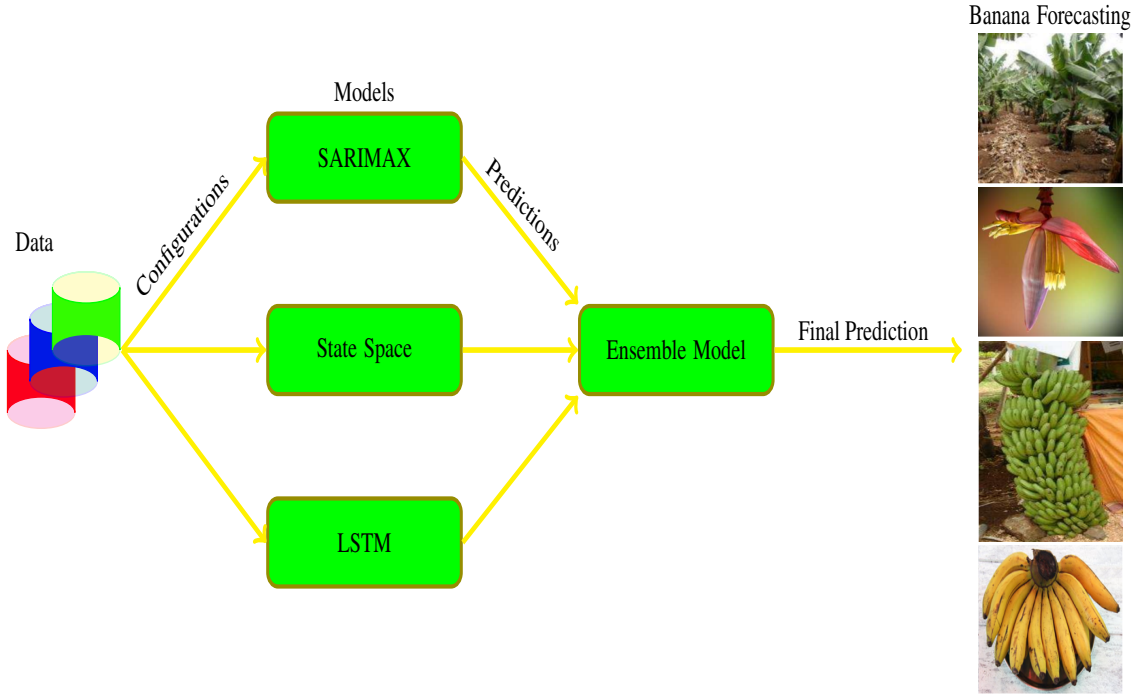


Figure 1: The schematic diagram, a representation of methodology

3.4 Correlation Analysis

3.4.1 Multiple Regression Model

In this work, a regression model to investigate the relationship between the five explanatory variables like precipitation (X_1), soil moisture (X_2), minimum temperature (X_3), maximum temperature (X_4), and relative humidity (X_5) and the yield of banana crops, denoted by the response variable Y was developed. The actual relationship between the response variable and the explanatory variables is shown by the population regression equation (Ngo & La Puente, 2012). But in order to estimate the population regression equation, which is unknown, we must use sampled data (Hanson, 2010; Sagamiko *et al.*, 2020).

Let us consider a dataset of size n observations, where the response variable Y and the p explanatory variables X_i have values in each observation. Let's write Y_i, X_{i1}, X_{i2}, ..., X_{ip} to represent the values for the i^{th} observation (Sagamiko *et al.*, 2020). The multiple regression equation for these values in this context can be shown as follows:

$Y_i = \Psi_0 + \Psi_1 X_{i1} + \Psi_2 X_{i2} + ... + \Psi_p X_{ip} + \varepsilon_i$, here, ($X_{i1}$, X_{i2}, ..., X_{ip}) represent the values of the explanatory variables for the i^{th} observation, and Y_i represents the value of the response variable for the i^{th} observation. The regression model's coefficients are represented by the symbols $\Psi_0, \Psi_1, \Psi_2, \ldots, \Psi_p$, and the error term for the i^{th} observation is ε_i (Sagamiko *et al.*, 2020).

For variable X_j, we can represent the i^{th} observation as X_{ij}, where $j = 1, 2, \ldots, p$, and $i = 1, 2, \ldots, n$. This is helpful in situations where there are more data points (n) than explanatory variables (p), which results in an overdetermined system with equations that are linearly dependent. The following set of equations can be used to represent the population model in this case for all sample observations (Hanson, 2010; Sagamiko *et al.*, 2020):

$$\begin{cases} Y_1 = \Psi_0 + \Psi_1 X_{11} + \Psi_2 X_{12} + \cdots + \Psi_p X_{1p} + \varepsilon_1 \\ Y_2 = \Psi_0 + \Psi_1 X_{21} + \Psi_2 X_{22} + \cdots + \Psi_p X_{2p} + \varepsilon_2 \\ \quad \vdots \\ Y_n = \Psi_0 + \Psi_1 X_{n1} + \Psi_2 X_{n2} + \cdots + \Psi_p X_{np} + \varepsilon_n \end{cases} \tag{3.1}$$

The system of equations (Eq. 3.1) can be expressed more succinctly in matrix form, as demonstrated by (Hanson, 2010; Sagamiko *et al.*, 2020):

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \Psi_0 \\ \Psi_1 \\ \vdots \\ \Psi_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.2}$$

The primary objective of this work's regression analysis is to identify the explanatory variables that significantly affect the yield of banana crops (Rathod & Mishra, 2018). Assuming a linear relationship between the response variable and the explanatory variables, we can quantitatively express the equation as follows (Adejuwon & Agundiminegha, 2019; Sagamiko *et al*., 2020; Salvacion, 2020):

$$Y = \Psi_0 + \Psi_1 X_1 + \Psi_2 X_2 + \Psi_3 X_3 + \Psi_4 X_4 + \Psi_5 X_5 + \varepsilon \tag{3.3}$$

where each explanatory variable's coefficients or parameters are $\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4$, and Ψ_5, and the error term or residual, ε, captures the variability in crop yield that the model is unable to explain.

The most common approach to estimating the population regression equation is to use the least squares method (Ngo & La Puente, 2012). Reducing the squared differences between the response variable's observed values and the corresponding regression model predictions is the aim of this strategy (Hanson, 2010).

The following equation yields the population regression equation's least squares estimator:

$$\Psi = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.4}$$

The response variable's observed values are represented by the vector Y in this equation, the matrix of explanatory variables is denoted by X, and the estimated coefficients of the regression equation are represented by Ψ.

We can use matrix notation to rewrite the multiple regression equation in order to illustrate this. By employing the matrices that were previously established in equation (3.3), we derive the subsequent expression:

$$\mathbf{Y} = \mathbf{X}\Psi + \varepsilon \tag{3.5}$$

The response variable values are represented by the column vector $\mathbf{Y}$, the design matrix is denoted by $\mathbf{X}$, the column vector of coefficients is denoted by Ψ, and the column vector of error terms is denoted by ε.

Using the least squares approach, we estimate the coefficients Ψ by utilizing the following estimator:

$$\hat{\mathbf{\Psi}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.6}$$

Upon entering the multiple regression equation with the estimated coefficients $\hat{\mathbf{\Psi}}$, we obtain:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{\Psi}} \tag{3.7}$$

where $\mathbf{X}$ is the design matrix, $\hat{\mathbf{\Psi}}$ indicates the column vector of estimated coefficients, ε is the column vector of error terms, and $\hat{Y}$ represents the vector of the response variable's predicted values.

With the least squares approach, the estimated population regression equation is thus $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{\Psi}}$. We substitute the estimated coefficients $\hat{\beta}$ into the equation $\Psi = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ in order to derive the equation $\Psi = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. The estimated population regression coefficients Ψ derived from the least squares approach are given by this equation. It is important to keep in mind that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ exists if the design matrix $\mathbf{X}^T\mathbf{X}$ is invertible.

3.5 Sensitivity Analysis

3.5.1 Sobol' Sensitivity Indices

By evaluating individual variables and their interactions based on variance, Sobol' Sensitivity Indices shed light on the relative significance of each climate factor (Iooss & Lemaître, 2015). The study may effectively present Sobol' Sensitivity Indices to thoroughly and in-depthly examine the sensitivity of banana crop yield to climate variables by adhering to pertinent mathematical representations (Raj *et al*., 2019). Considering a mathematical model with one output, Y, and a set of input parameters, $X_1, X_2, \ldots, X_k$ (Todorov *et al*., 2021). For every input parameter, we define the minimum and maximum bounds, X_i^{min} and X_i^{max}, respectively (Owen, 2013). For every input parameter X_i, we produce sets of input samples using a particular sampling strategy, such as Monte Carlo sampling, Sobol' sequence, or Latin Hypercube sampling (Iooss & Lemaître, 2015). Based on all produced sets of input samples, we determine the overall variance of the model output, Y: Total Variance (V(Y)) = Var$(Y^{(j)})$ (Todorov *et al*., 2021).

Without accounting for other parameters, we compute the variance of the model's output that is attributable to each distinct input parameter X_i. The first-order sensitivity index (main effect) for the i^{th} input parameter is given by (Iooss & Lemaître, 2015; Raj *et al*., 2019):

$$S_i = \frac{\text{Var}(E[Y|X_i])}{\text{Var}(Y)} \tag{3.8}$$

where Var$(E[Y|X_i])$ is the variance of the conditional expectation of Y given X_i, and Var(Y) is the total variance.

Then, we estimate the variance of the model output due to the i^{th} input parameter, accounting for all of its interactions with other factors. The total effect sensitivity index for the i^{th} input parameter is provided by (Raj *et al.*, 2019):

$$S_{Ti} = 1 - \frac{\text{Var}(E[Y|X_{\sim i}])}{\text{Var}(Y)} \tag{3.9}$$

where Var$(E[Y|X_{\sim i}])$ is the variance of the conditional expectation of Y given all input parameters other than X_i.

If preferred, Sobol' indices can also be calculated for higher-order interactions or interactions between pairs of input parameters (second-order indices) (Kucherenko & Song, 2016). These indices provide insight into the way different elements interact to influence the variance of the model's output (Raj *et al.*, 2019). Sobol's Sensitivity Indices are used to identify the most significant input parameters and quantify their contributions to the output variance (Iooss & Lemaître, 2015).Sobol's Sensitivity Indices provide a global sensitivity analysis because they account for the interactions between various factors (Raj *et al.*, 2019). In other words, the indices evaluate the significance of every climate variable throughout the whole input space, taking into account how they affect the output when paired with other variables (Todorov *et al.*, 2021).

3.5.2 Response Surface Methodology

For the design and optimization of experiments, Response Surface Methodology (RSM) is a popular statistical and mathematical technique (Yolmeh & Jafari, 2017). By examining the relationships between numerous input variables (factors) and an output of interest, it is frequently used to model and optimize complex systems (Jankovic *et al.*, 2021). When there may be interactions between the factors and a non-linear relationship between the factors and the response, RSM is particularly useful (Reji & Kumar, 2022). The Design of Experiments (DOE) strategy was expanded upon in the 1950s with the development of the Response Surface Methodology by George E. P. Box and K. B. Wilson (Kleijnen, 2014). By maximizing or minimizing the output, RSM aims to determine the optimal input variable settings that will yield the desired response (Reji & Kumar, 2022).

RSM's primary objective is to develop a mathematical model that roughly represents the system's response surface (Nwabueze, 2010). The response variable's variation as a function of the input variables is represented by the response surface (Jankovic *et al.*, 2021). By fitting a math-

ematical model to the experimental data, researchers can investigate the effects of individual factors, the interactions between factors, and the optimal factor settings that yield the desired result (Jou *et al*., 2014). A response surface model's general mathematical representation can be written like this (Kleijnen, 2014; Yolmeh & Jafari, 2017):

$$Y = \theta_0 + \sum(\theta_i X_i) + \sum(\theta_{ij} X_i X_j) + \sum(\theta_{ii} X_i^2) + \varepsilon \tag{3.10}$$

where Y is the response variable (output), the input variables (factors) are $(X_i, and\ \ X_j)$, the regression coefficients are $(\theta_0, \theta_i, \theta_{ij}, and\ \ \theta_{ii})$, and the error term is ε. The response is influenced by various variables and their interactions; the model may contain linear terms $(\theta_i X_i)$, interaction terms $(\theta_{ij} X_i X_j)$, and quadratic terms $(\theta_{ii} X_i^2)$.

Using response surface methodology (RSM), researchers can determine which factors have the greatest impact on response and determine the ideal factor combinations to achieve desired results. It is widely used in many industries, such as engineering, chemistry, and manufacturing, to improve product performance and streamline processes.

3.6 Uncertainty Quantification

3.6.1 Monte Carlo Simulation

Monte Carlo Simulation (MCS) is a powerful numerical technique used in engineering, finance, physics, and computational science to quantify uncertainty (Amin *et al*., 2022; Tahmasebinia *et al*., 2022). When analytical solutions are not practical or difficult to obtain, MCS offers a probabilistic approach to estimate uncertainties and risks associated with complex systems or models (Dega *et al*., 2023). In order to obtain a distribution of the model output, the basic idea of MCS is to propagate samples through the model while sampling from probability distributions of uncertain input parameters (Antanasijević *et al*., 2014; Urquizo *et al*., 2017).

Assume we have a mathematical model with one output, Y, and a set of input parameters, $X_1, X_2, \ldots, X_k$. A probability distribution $P(X_i)$ is linked to each input parameter X_i. The model can be expressed as (Dega *et al*., 2023; Urquizo *et al*., 2017):

$$Y = f(X_1, X_2, \ldots, X_k) \tag{3.11}$$

The following steps are part of the general Monte Carlo simulation procedure: defining the probability distributions for each entered parameter X_i (Dega *et al*., 2023). Depending on the type of uncertainty, common probability distributions include exponential, Gaussian, uniform, and others (Amin *et al*., 2022). The sets of input parameter values are obtained by randomly

selecting samples from the given probability distributions (Dega *et al.*, 2023). The user determines the number of samples, which is indicated by N. We evaluate the model to get the corresponding model output Y^j, where $j = 1, 2, \ldots, N$, for each set of sampled parameter values (Antanasijević *et al.*, 2014). The mean, variance, percentiles, and other relevant statistics are estimated by examining the obtained distribution of model outputs. This data sheds light on the model's output's uncertainty and variability (Urquizo *et al.*, 2017). Thus, one of the most widely used methods for estimating uncertainty is Monte Carlo Simulation because of its versatility and ability to handle complex problems. It is an essential tool for risk assessment and decision-making in a variety of real-world scenarios (Dega *et al.*, 2023).

3.7 Forecasting Models

3.7.1 Seasonal ARIMA (SARIMA) with External Variables, SARIMAX

Among the most widely used and successful time-series models, ARIMA is a classic (Rathod & Mishra, 2018). Because of its linear statistical properties and the widely applied Box-Jenkins approach to model creation, which was developed by Box and Jenkins in the 1970s, the ARIMA model has become extremely popular (Box *et al.*, 2015). $ARIMA(p, d, q)$ is the standard form of the ARIMA model, where p denotes the auto-regressive term order, d denotes the differencing term order, and q denotes the moving average term order (Arunraj *et al.*, 2016; Hyndman & Athanasopoulos, 2018). The $ARIMA(p, d, q)$ model can be expressed mathematically as (Arunraj *et al.*, 2016):

$$\phi_p(B)(1 - B)^d X_t = \mu + \theta_q(B)\varepsilon_t \tag{3.12}$$

where X_t denotes the time-series variable at time t, which is the variable being modeled or predicted, and $\phi_p(B)$ stands for the autoregressive (AR) operator of order p. The differencing operator is represented by the expression $(1 - B)^d$, where d denotes the order of differencing. The equation's constant term, μ, accounts for any deterministic offset or component in the time series. The error term at time t, ε_t, indicates the random or unexplained component of the time-series. $\theta_q(B)$ stands for (MA), the moving average operator of order q.

To account for seasonal variations, the ARIMA model can be extended as SARIMA$(p, d, q)(P, D, Q)_s$, where s is a term that takes the length of the seasonal period into account (Meeradevi *et al.*, 2022; Neog *et al.*, 2022; Raj *et al.*, 2019). It is possible to depict the SARIMA model as (Arunraj *et al.*, 2016):

$$\phi_p(B)\Phi_P(B^S)(1 - B)^d(1 - B^S)^D X_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \tag{3.13}$$

where $\theta_q(B)$ denotes (MA) the seasonal moving average operator of order q, $\phi_p(B)$ denotes

(AR) the seasonal autoregressive operator of order p, $(1-B)^d$ stand for the differencing operator applied d times, $(1-B^S)^D$ indicates the seasonal differencing operator applied D times, and S denotes the seasonal length (e.g., $s=4$ in quarterly data, and $s=12$ in monthly data).

Multi linear regression techniques are used to model the external variables given the SARIMAX$(p,d,q)(P,D,Q)_s$ model, where (X) is the vector of external variables (Arunraj *et al*., 2016). A multiple regression model in this study can be mathematically expressed as follows:

$$Y_t = \Psi_0 + \Psi_1 X_1 + \Psi_2 X_2 + \Psi_3 X_3 + \Psi_4 X_4 + \Psi_5 X_5 + \alpha_t \tag{3.14}$$

where each explanatory variable's coefficients or parameters are $\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4$, and Ψ_5. The error term, or residual, is represented by α_t, and it captures the variability in crop yield that the model is unable to explain. α_t is the error term which can be expressed in the form of SARIMA model as (Arunraj *et al*., 2016):

$$\alpha_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{3.15}$$

We obtain the following equation by substituting Equation (3.15) into Equation (3.14):

$$Y_t = \Psi_0 + \Psi_1 X_1 + \Psi_2 X_2 + \Psi_3 X_3 + \Psi_4 X_4 + \Psi_5 X_5 + \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{3.16}$$

3.7.2 State Space (SS) Model

A mathematical framework for representing time series data is the state space approach (Aoki, 2013). It represents the underlying process as a collection of unobserved states that change over time in accordance with a set of stochastic equations, which produces the observable data (Verma, 2018). Then, using a set of observation equations, the observed data is produced from these unseen states (Suman & Verma, 2017).

State equation:

$$\boldsymbol{x}_t = \boldsymbol{A}_t \boldsymbol{x}_{t-1} + \boldsymbol{B}_t \boldsymbol{\omega}_t \tag{3.17}$$

The $n \times 1$ vector of unobserved states at time t is denoted by $\boldsymbol{x}_t$; the $n \times n$ state transition matrix is represented by $\boldsymbol{A}_t$; the $n \times m$ matrix of state noise is denoted by $\boldsymbol{B}_t$; and the $m \times 1$ vector of state noise at time t is denoted by $\boldsymbol{\omega}_t$.

Observation equation:

$$\boldsymbol{y}_t = \boldsymbol{C}_t \boldsymbol{x}_t + \boldsymbol{\tau}_t \tag{3.18}$$

where the $p \times 1$ vector of observed data at time t is denoted by $\boldsymbol{y}_t$, the $p \times n$ observation matrix is represented by $\boldsymbol{C}_t$, and the $p \times 1$ vector of observation noise at time t is denoted by $\boldsymbol{\tau}_t$.

The state space model assumes that the noise in the observations and the state are independent, with known covariance matrices and zero mean normal distributions for each (Verma, 2018):

$$\boldsymbol{\omega}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t) \quad \text{and} \quad \boldsymbol{\tau}_t \sim N(\mathbf{0}, \boldsymbol{R}_t) \tag{3.19}$$

where the $m \times m$ and $p \times p$ covariance matrices of the state noise and observation noise, respectively, are denoted by the symbols $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$.

The state space technique can be used to create a wide range of time series models, such as ARMA models, ARIMA models, and state space models with non-linear and non-Gaussian state transitions and observation equations (Hooda *et al.*, 2020; Hu *et al.*, 2019; Verma, 2018). When modeling time series data impacted by several external factors, such as climate change, state space models can be very helpful (Cook, 1985; Marolla *et al.*, 2021). By representing the external factors as extra states in the model, they are able to represent the effects of several external factors on the time series. To do this, extra equations that explain the dynamics of the external factors are added (Marolla *et al.*, 2021).

Usually, maximum likelihood estimation or Bayesian techniques are used to estimate the state space model (Newman *et al.*, 2023). Assuming a state space model with state vectors x_t and observations y_t. The likelihood function can be expressed as follows:

$$L(\Theta|y) = f(y_1|\Theta)f(x_1|\Theta)\prod_{t=2}^{T} f(y_t|x_t,\Theta)f(x_t|x_{t-1},\Theta) \tag{3.20}$$

where $y_t|x_t,\Theta)$ and $f(x_t|x_{t-1},\Theta)$ are the conditional densities of the observations and state vectors, respectively, and Θ indicates the parameters of the state space model.

The set of parameters $\hat{\Theta}$ that maximizes the likelihood function is found using the MLE method:

$$\hat{\Theta} = \text{argmax}_{\Theta} L(\Theta|y) \tag{3.21}$$

Furthermore, the posterior distribution of the parameters can be expressed as follows:

$$p(\Theta|y) \propto L(\Theta|y)p(\Theta) \tag{3.22}$$

where $p(\Theta)$ denotes the previous distribution of the parameters and $L(\Theta|y)$ is the likelihood function as previously defined.

Parameter estimation is done using the recursive Bayesian estimation technique known as the Kalman filter algorithm (de Bézenac *et al.*, 2020). To estimate the parameters, the Kalman filter

combines past understanding of the dynamics of the system with the observed data. When new data becomes available, it updates the parameter estimates and optimally integrates the available information (de Bézenac *et al*., 2020; Suman & Verma, 2017).

In order to obtain the predicted values, such as banana crop yield, the forecasts are created by projecting the latent state variables into the future and using the observation equation:

$$\hat{y}T + 1|T = \mathbb{E}[yT + 1|y_{1:T}, \Theta] = \mathbb{E}[f(s_{T+1})|\hat{s}_{T+1|T}, \Theta] \tag{3.23}$$

where $f(s_{T+1})$ is the observation equation connecting the latent state variables to the observed yield, and $\hat{s}T + 1|T$ is the predicted state estimate for time $T + 1$ given the observed data $y1 : T$.

3.7.3 Long Short-Term Memory (LSTM) Model

A specific architecture designed to handle sequential data, especially time series data, is Long Short-Term Memory (LSTM), also referred to as a kind of recurrent neural networks (RNNs) (Meeradevi *et al*., 2022; Tian *et al*., 2021). LSTMs specifically solve the vanishing gradient issue that traditional RNNs find difficult. Long-term dependencies in sequential data are better described by LSTMs because they store and retrieve information efficiently over long periods of time (Reddy *et al*., 2022). The LSTM model has been applied in many domains, including climate forecasting, and essentially provides a useful technique for working with sequential data (Bhimavarapu *et al*., 2023; Meeradevi *et al*., 2022; Tian *et al*., 2021).

The architecture of an LSTM consists of three gates (input, output, and forget) and a memory cell with long-term information retention (Tian *et al*., 2021). The input gate regulates the amount of new data that is added to the memory cell, the output gate regulates the amount of data that is removed from the memory cell, and the forget gate regulates the amount of outdated or superfluous data that is eliminated from the memory cell. These gates control data flow into and out of the memory cell within the LSTM paradigm, facilitating effective information retention and utilization (Bhimavarapu *et al*., 2023; Liu *et al*., 2023).

The following equations are part of the configuration of the LSTM model:

$$Input\ layer:\ y_t = g(W_i * x_t + b_i) \tag{3.24}$$

$$LSTM\ layer:\ h_t = LSTM(h_{t-1},\ y_{t-1}) \tag{3.25}$$

$$Output\ layer:\ y_{t+1} = g(W_o * h_t + b_o) \tag{3.26}$$

Applying an activation function g to the dot product of the weight matrix W_i and the input vector x_t yields the output y_t in the input layer. A bias term b_i is then added. The output of the input layer at time t is represented by this y_t.

The previous hidden state h_{t-1} and the previous input y_{t-1} are passed to the LSTM cell to determine the output h_t of the LSTM layer at time t. Based on these inputs, the LSTM cell generates a new hidden state h_t by updating its internal state.

In the subsequent time step, the activation function g is applied to the dot product of the weight matrix W_o and the hidden state h_t to compute the predicted output y_{t+1}. A bias term b_o is then added.

Prior to using the training data to train the model, the LSTM neuron weights are updated using the backpropagation through time (BPTT) technique. This entails applying the chain rule to calculate the gradients of the loss function with respect to the weights (Sadowski, 2016). By iteratively changing the weights based on the estimated gradients, the LSTM model gains the ability to make more accurate predictions and learns to perform better on the training set (Bhimavarapu *et al.*, 2023).

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial W} \tag{3.27}$$

In this equation, L, W, y, h, and t represents the loss function, weight, output, hidden state, and the time step, respectively. In order to train the LSTM model and maximize its performance on the training set, these variables are crucial.

Additionally, an appropriate optimization method, like Adam, is used to account for the gradient's first and second moments. This improves the LSTM model's training process by enabling the algorithm to modify the learning rate independently for each weight (Bhimavarapu *et al.*, 2023).

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot \frac{\partial L}{\partial W} \tag{3.28}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot \left(\frac{\partial L}{\partial W}\right)^2 \tag{3.29}$$

$$W = W - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{3.30}$$

where W is the weight being updated, α is the learning rate, β_1 and β_2 represent hyperparameters that control the decay rates of the moment estimates, and ϵ stands for a small constant to prevent division by zero. The first and second moment estimates of the gradient are denoted by m_t and v_t, respectively.

Finally, after LSTM model optimization, if its performance reaches the required level of accuracy, it is deployed for use in predictions (Bhimavarapu *et al*., 2023), such as predicting banana crop yield under various climate scenarios.

3.7.4 Ensemble Modeling Approach

Ensemble modeling is a flexible approach that seeks to improve prediction accuracy and reliability by combining the output of multiple models (Bertsimas & Boussioux, 2023). While time series data can be used with ensemble models, their primary objective is to enhance overall model performance, not to specifically address the unique characteristics of time series data (Hao *et al*., 2020).

Time series forecasts can be made more accurate by integrating ensemble modeling with time series modeling (Bertsimas & Boussioux, 2023). An example of a common technique in ensemble modeling is building a varied ensemble of time series models, such as neural network, exponential smoothing, and ARIMA models. The predictions from each of these separate models are then combined using weighted averaging or other techniques (Bayati *et al*., 2020; Kamir *et al*., 2020; Kourentzes *et al*., 2014).

For example, the ensemble model uses a weighted average technique to combine the predictions of each individual model to produce a final prediction. The weighted average approach can be represented mathematically as follows:

$$y = w_1 \times y_1 + w_2 \times y_2 + w_3 \times y_3 + \cdots + w_n \times y_n \tag{3.31}$$

where w_1, w_2, ..., and w_n are the weights assigned to the individual models based on their performance on the training, or validation set, and y is the final predicted value. The individual models' predicted values are y_1, y_2, ..., and y_n, respectively.

Moreover, model performance on the testing set determines the weights assigned, based on the inverse of each model's error or loss (Van Leeuwen *et al*., 2023). In this study, weights were derived from R-squared values, normalized to ensure comparability and sum to one. The final forecast was obtained by taking a weighted average of the predictions, minimizing weaknesses and maximizing the strengths of each model.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, the study delve into the heart of this research, where the culmination of extensive analysis and rigorous modeling efforts is unveiled. It consists the presentation of the empirical outcomes derived from the intricate analysis. This study embark on an interpretive journey that contextualizes the results within the broader landscape of the research objectives.

4.2 Data Exploration Results

Throughout the analysis, the MATLAB (R2021a) and Spyder (Python 3.9) tools were utilized interchangeably. The yearly reanalysis datasets of relative humidity, minimum and maximum temperatures, precipitation, and soil moisture were used in the analysis. The yield of the banana crop was modeled and predicted using the yearly reanalysis datasets of climate variables. Every stage needed for pre-processing and filtering the data was taken into account, including autocorrelation and detrending (non-stationarity and seasonality).

Furthermore, the study acknowledges the unavailability of alternative sources at the subnational level for banana yield data in Tanzania. It also considered a clear distinction between weather and climate, as explained in the subsequent sections. In particular, the average annual (climate variables, and banana crop yield) statistics were used throughout the analysis in this study.

The methods that were chosen all used climate time-series data to forecast Tanzania's banana production yield for the 60 years from 1961 to 2020. The models were trained on the first 80% of the datasets, and their effectiveness was validated on the remaining 20%. To ensure that all variables are on the same scale, the training and validation sets were normalized as needed using the normalize function. This normalization procedure aids in preventing possible problems brought on by variations in the variables' magnitudes. To make processing and handling of the training and validation data easier for the models, they were converted into cell arrays. When building and evaluating models, more flexible and effective data manipulation is possible when converting the data to cell arrays.

As indicated in Table 2, a number of statistical metrics were discovered in each model, indicating how well the best model fit the data was chosen. The models' performance metrics and evaluation outcomes are displayed in this table.

Table 2: Statistical Evaluation Metrics

MODEL	Training Set				Validation Set			
	MSE	MAE	RMSE	R^2	MSE	MAE	RMSE	R^2
SARIMAX	0.3828	0.3650	0.6187	0.8109	4.3797	1.4789	2.0928	0.1825
State Space	0.0105	0.0423	0.1026	0.9948	0.0885	0.2068	0.2974	0.9835
LSTM	0.6200	0.4192	0.7874	0.6991	0.5288	0.6890	0.7272	0.9013

4.3 Regression Analysis and Results

The assumption that the response and the explanatory variables have a linear relationship is supported by this research. Table 3 displays the coefficients from the regression analysis. These coefficients illustrate how, when each explanatory variable is changed by one unit while keeping all other variables constant, the rate of change in banana crop yield is affected by key climate variables. By inputting the values of these regression coefficients into the regression (Eq. 3.3), the study derived the following expression:

$$Y = -22.8320 + 0.0206X_1 - 0.0085X_2 + 4.8328X_3 - 1.6594X_4 - 0.0991X_5 \tag{4.1}$$

With a negative sign indicating a gradual decrease in banana crop yield, the intercept (-22.8320), a constant term, represents the predicted value of Y. This value indicates the baseline or initial level of banana crop yield in the absence of any influence from the explanatory variables. The coefficients (0.0206, -0.0085, 4.8328, -1.6594, -0.0991) attached to the explanatory variables (X_1, X_2, X_3, X_4, and X_5 represent precipitation, soil moisture, minimum temperature, maximum temperature, and relative humidity, respectively) indicate the impact of each variable on the banana crop yield when the other variables are held constant.

By considering statistical measures (like p-values) to determine the significance and impact of each variable in this model while maintaining the context of the data and the specific research question being addressed. Here, a high positive coefficient and a small negative coefficient indicate that the independent variable contributes significantly to banana crop yield. Overall, Table 3 shows that the chosen explanatory variables can account for about 50.2% of the variability in banana crop yield, according to the regression model's R-squared value of 0.502. The model as a whole is statistically significant, as evidenced by the highly significant F-statistic of 10.87 (Prob (F-statistic): 2.89e-07).

Table 3: OLS Regression Results

Model	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)
OLS	0.502	0.455	10.87	2.89e-07

Variable	Coef.	Std. Err.	t-stat	P-value	[0.025, 0.975]
Constant	-22.8320	30.506	-0.748	0.457	[-83.993, 38.329]
X_1	0.0206	0.043	0.478	0.634	[-0.066, 0.107]
X_2	-0.0085	0.007	-1.147	0.257	[-0.023, 0.006]
X_3	4.8328	1.628	2.968	0.004	[1.569, 8.097]
X_4	-1.6594	1.648	-1.007	0.318	[-4.963, 1.644]
X_5	-0.0991	0.069	-1.439	0.156	[-0.237, 0.039]

4.4 Results of Sobol' Sensitivity Indices

Relative weights of each predictor variable in explaining the variation in banana yield are displayed in Table 4, based on the Sobol' Sensitivity Indices from the multiple regression model in (Eq. 3.3). These sensitivity indices quantify the individual effects of each input parameter on the banana crop's yield. Specifically, a value near 1 denotes a substantial impact of the parameter on the variability in banana yield, whereas a value near 0 indicates a negligible effect. Certain values in the multiple regression model may seem higher than 100% because of the interactions between the predictor variables (Raj *et al*., 2019).

Table 4: Sobol' Sensitivity Indices

Parameter	Sensitivity Index	Ranking
const	1.0	-
Precipitation	0.979	2
Soil moisture	1.270	5
Minimum temp	0.953	1
Maximum temp	0.981	3
Relative humid	1.055	4

Table 5 presents the second-order sensitivity indices that quantify the cumulative impact of input parameter pairs on banana crop yield. Whereas a lower value denotes a weaker interaction, a higher value shows that the two parameters interact significantly and affect the output. The derived Sobol' Sensitivity Indices and Second-Order Sobol' Sensitivity Indices, respectively, show the effects of each input parameter and their interactions on the model's output (banana crop yield). The values shed light on how important each variable is in relation to the others and how they interact with one another to affect yield variability.

Table 5: Second-Order Sobol' Sensitivity Indices

Parameter Pair	SI	Parameter Pair	SI
(const, Precipitation)	0.995	(Precipitation, Relative humid)	0.981
(const, Soil moisture)	0.796	(Soil moisture, Minimum temp)	0.727
(const, Minimum temp)	0.571	(Soil moisture, Maximum temp)	0.783
(const, Maximum temp)	0.833	(Soil moisture, Relative humid)	0.935
(const, Relative humid)	0.938	(Minimum temp, Maximum temp)	0.601
(Precipitation, Soil moist)	0.778	(Minimum temp, Relative humid)	0.756
(Precipitation, Min temp)	0.969	(Maximum temp, Relative humid)	0.874

Moreover, bar plots can be used to visually represent Sobol' Sensitivity Indices, as shown in Fig. 2. These graphical representations demonstrate the relative contributions of each climate variable to the variation in banana crop yield. However, interaction plots (Fig. 2) can also be used to illustrate how variable interactions affect the result. The values of Sobol's Sensitivity Indices for each climate variable are shown in a bar plot, with higher values indicating greater influence on the output variance. The results demonstrate how the relative importance of each input parameter and how their pairwise interactions impact crop yield of bananas.

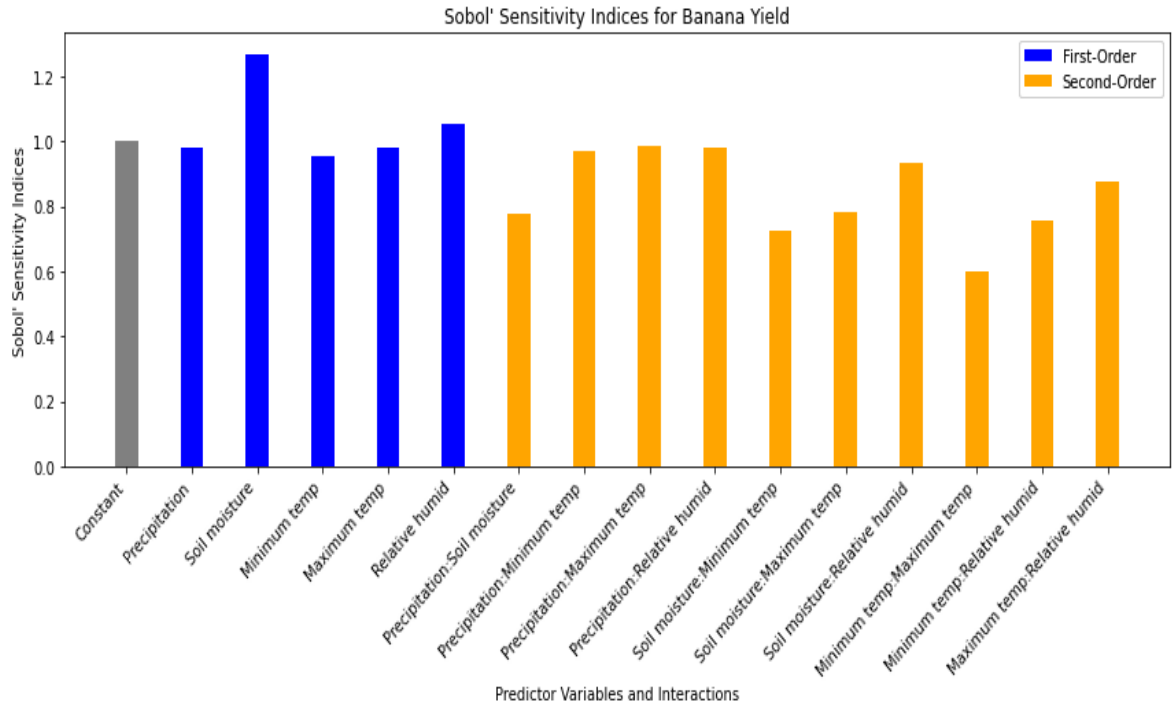


Figure 2: The plot of Sobol' Sensitivity Indices for Banana Crop Yield

Overall, the Sobol' Sensitivity Indices help us identify the climate variables that most influence banana yield and how their effects relate to the crop's variability in production. In order to use the model as a decision-making tool for agricultural practices and climate change adaptation plans, it is imperative to understand its sensitivity to different input elements. By knowing these

sensitivity indices, stakeholders, legislators, and farmers can make well-informed decisions to mitigate and adapt to potential effects of climate change on Tanzania's banana output. The conclusions from this analysis highlight the need of proactive and flexible actions in the face of a climatically uncertain future and contribute to the understanding of how agriculture is impacted by climate change.

4.5 Results of Response Surface Methodology

The study anticipated that the datasets used in this work would have complex relationships that would not be adequately captured by only linear, non-linear, and non-monotonic models, so the study considered the Response Surface Methodology. It should be noted that the Response Surface Methodology is an approximation technique that assumes linear, interaction, and quadratic relationships between the explanatory and response variables. Precipitation, soil moisture content, and minimum temperature are the three most crucial input parameters for banana yield, according to Table 6 results. Significant individual interaction effects are also seen for relative humidity, soil moisture, and precipitation.

Table 6: Sensitivity Indices based on Response Surface Methodology

Parameter	**Sensitivity Index**
Precipitation	1.0525
Interaction Effect for Precipitation	1.1101
Soil moisture	1.2880
Interaction Effect for Soil moisture	1.2223
Minimum temp	0.4189
Interaction Effect for Minimum temp	0.2029
Maximum temp	0.6257
Interaction Effect for Maximum temp	0.4436
Relative humid	0.7657
Interaction Effect for Relative humid	1.2646

The interaction effect between soil moisture and precipitation, on the other hand, is the strongest, as Table 7 demonstrates. This indicates that when soil moisture is high, the impact of precipitation on banana yield is increased. Although it is not as strong as the interaction between precipitation and soil moisture, the relationship between precipitation and minimum temperature is still significant. This indicates that when the minimum temperature is low, the impact of precipitation on banana yield is increased. Of the three interaction effects, the relationship between relative humidity and precipitation is the weakest, despite still being significant. This indicates that a high relative humidity does not considerably increase the impact of precipitation on banana yield.

Table 7: Interaction Effects based on Response Surface Methodology

Parameter Pair	Interaction Effect
Precipitation:Soil moisture	1.1101
Interaction Effect for Precipitation:Soil moisture	1.4703
Precipitation:Minimum temp	1.2223
Interaction Effect for Precipitation:Minimum temp	0.0216
Precipitation:Maximum temp	0.2029
Interaction Effect for Precipitation:Maximum temp	0.7875
Precipitation:Relative humid	0.4436
Interaction Effect for Precipitation:Relative humid	0.2899
Soil moisture:Minimum temp	1.2646
Interaction Effect for Soil moisture:Minimum temp	0.1881
Soil moisture:Maximum temp	1.4703
Soil moisture:Relative humid	0.0216
Minimum temp:Maximum temp	0.7875
Minimum temp:Relative humid	0.2899
Maximum temp:Relative humid	0.1881

For a more thorough examination, the Response Surface Methodology results were visually represented in Fig. 3. Overall, the results of the sensitivity analysis show that soil moisture, minimum temperature, and precipitation are the three most important input variables for banana yield. The way in which these parameters interact can also have a big effect on banana yield.

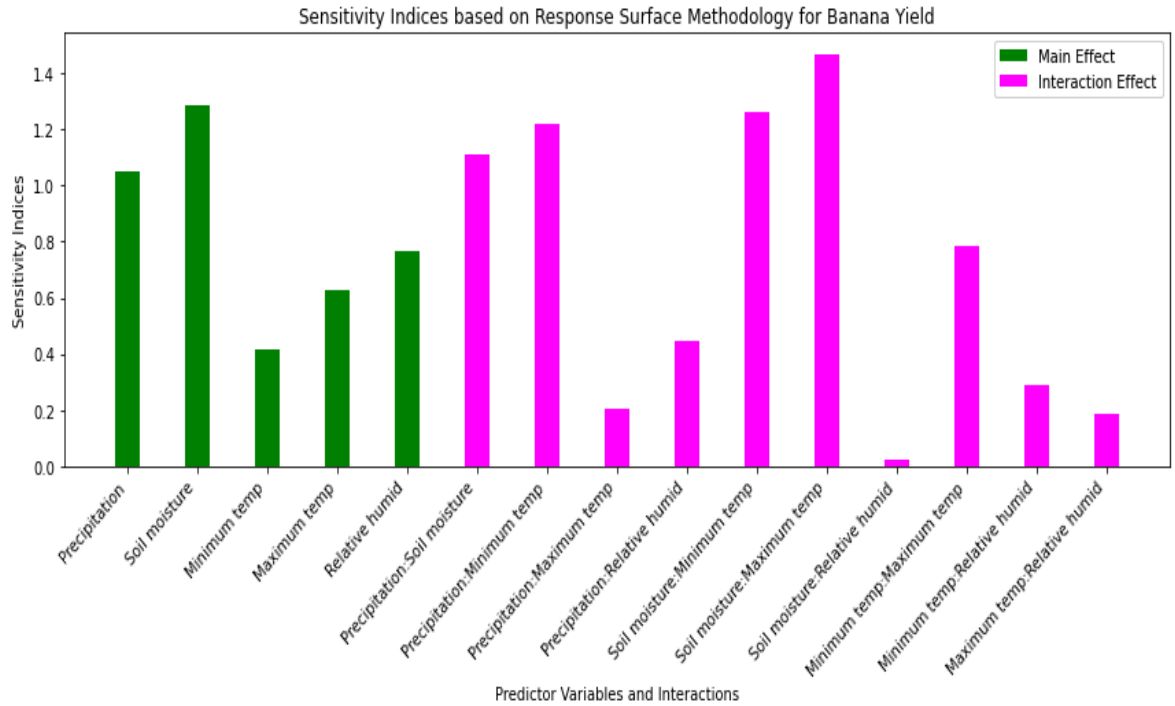


Figure 3: The plot of Sensitivity Indices based on Response Surface Methodology for Banana Crop Yield

4.6 Results of Monte Carlo Simulation

In order to evaluate the variability and confidence in the banana yield predictions based on the predictor variables of precipitation, soil moisture, minimum temperature, maximum temperature, and relative humidity, the study performed uncertainty quantification using Monte Carlo simulation in this study. The datasets were fitted with a multiple regression model, and 1000 simulations were used to assess the degree of uncertainty. The study performed the Monte Carlo convergence analysis by calculating the standard deviation of the predicted yields for a range of iterations (from 1 to the total number of simulations).

The curve in Fig. 4 illustrates how the standard deviation changes as the number of simulations increases. By doing so, we can verify whether the Monte Carlo simulation has reached a sufficient number of iterations to produce stable and trustworthy results and check for convergence. Here, the plot demonstrates how the standard deviation stabilizes as the number of simulations rises, signifying the completion of convergence.

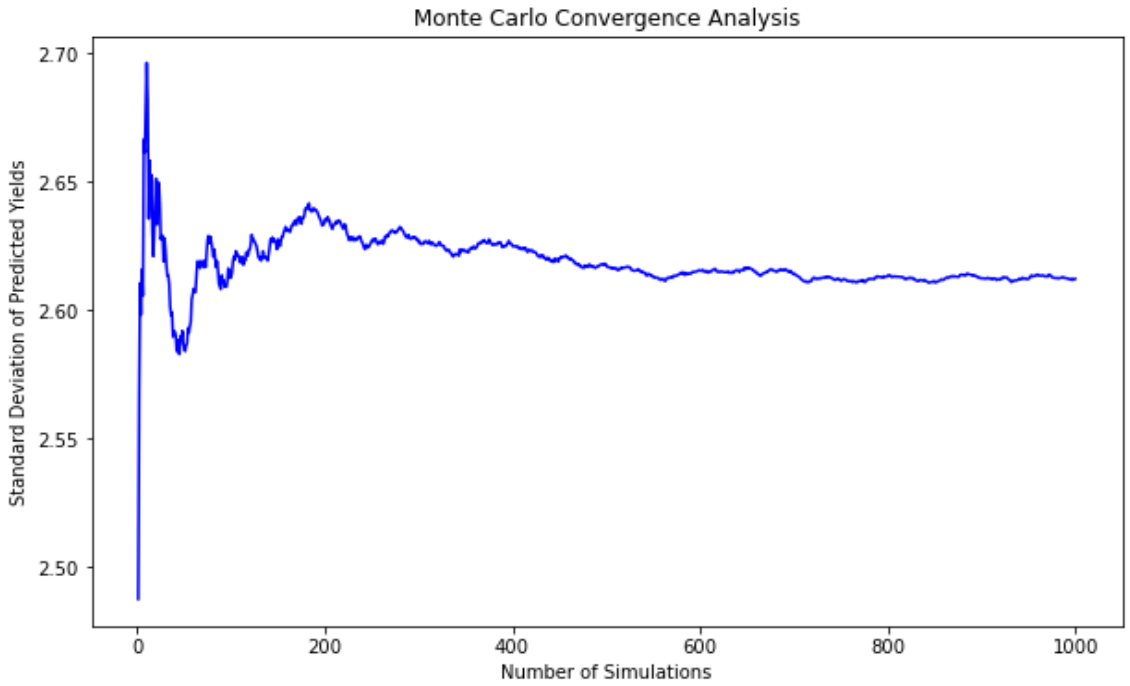


Figure 4: The plot of Monte Carlo convergence analysis

Key findings from the uncertainty quantification are presented in Table 8. These include: With regard to precipitation, the mean predicted yield of bananas was discovered to be 5.44, exhibiting a standard deviation of 1.99. The 95% confidence interval encompassed a range of 1.55 to 9.14, signifying significant fluctuations in the forecasts. In terms of soil moisture, the model predicted a mean yield of 2.51 with a standard deviation of 1.97; the 95% confidence interval covered a range from -1.47 to 6.40, indicating a significant degree of uncertainty in the results.

Minimum Temperature: The 95% confidence interval covered a significant range, from -1.23 to 6.38, and the predicted mean yield was 2.69 with a standard deviation of 1.95. Maximum Temperature: The model indicated a mean yield of 0.82 with a standard deviation of 1.97. The 95% confidence interval covered a wide range of values, from -3.11 to 4.78, suggesting a significant degree of uncertainty. In terms of relative humidity, the 95% confidence interval encompassed a range from -1.42 to 6.07, indicating significant variability in the predictions. The mean predicted yield was 2.25, with a standard deviation of 1.93.

Table 8: Summary of Uncertainty Quantification Results

Predictor Variable	Mean Prediction	SD	95% CI (Low)	95% CI (High)
Precipitation	5.437753	1.985648	1.552227	9.143635
Soil Moisture	2.507884	1.968617	-1.471693	6.401481
Minimum Temp	2.688415	1.951703	-1.226391	6.381435
Maximum Temp	0.817874	1.968334	-3.109263	4.778339
Relative Humid	2.249391	1.925037	-1.420177	6.074196

These quantification results of uncertainty offer important new information about how robust and dependable the model's predictions are. It is clear that a large portion of the uncertainty in the yield estimates is accounted for by the predictor variables, particularly minimum temperature, soil moisture and precipitation. These discoveries should be taken into account by scholars and decision-makers as they analyze the data and draw conclusions from the model's predictions. In order to lower uncertainty and improve prediction accuracy, more research and data collection may be required.

A comprehensive residual analysis was then carried out to assess the validity of the model. Plots were used to analyze residuals, or the differences between actual and predicted values, as shown in Fig. 5. The distribution of residuals in respect to the predicted values is displayed in the residual plot. In order to identify differences between the actual and predicted values, a horizontal dashed line was placed at zero. The residual distribution is revealed by the histogram. A density histogram shows the residuals' concentration in the vicinity of zero. To determine whether the residuals had a normal distribution, the Q-Q plot was employed. Deviations from normalcy were found by comparing sample quantiles with theoretical quantiles.

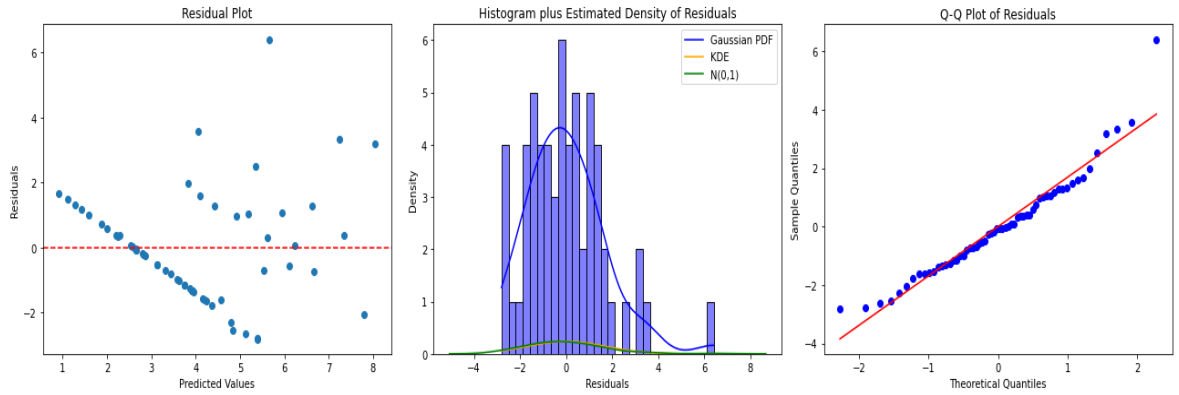


Figure 5: The residuals were analyzed using these plots to check for normality and patterns

Furthermore, the normality assumption of the residuals was formally tested using the Shapiro-Wilk test. The degree to which the residuals' distribution resembles a normal distribution is indicated by the test statistic (0.9446). A better fit to normalcy is indicated by a value that is closer to 1. The p-value (0.0088) represents the probability of observing the test statistic assuming that the residuals are normally distributed. A smaller p-value indicates that the residuals significantly deviate from a normal distribution.

The results indicate that although the multiple linear regression model predicts banana yield reasonably well, there are signs of non-normality in the residuals according to the Shapiro-Wilk test and graphical analysis. This suggests that there may be unrecognized sources of variability or possible limitations in the model. These results highlight how important it is to evaluate and apply the regression model's predictions while carefully examining model assumptions and accounting for uncertainty. To improve the accuracy and robustness of the model, more research and development may be required.

In addition, the yield of banana crops and the historical temporal trends of the climate variables piqued the study's interest. The findings presented in Fig. 6 demonstrate that, over the course of the study, precipitation, soil moisture, and relative humidity all tended to decrease while maximum and minimum temperatures showed increasing trends. The current state of the climate is supported by this analysis (Field & Barros, 2014). The climate variables gathered for this study were thought to have an effect on the yield of banana crops.

The time series plot of banana crop yield (Fig. 6), viewed from a different angle, shows some intriguing trends. Seasonality is evident in the 60-year dataset (1961–2020), with consistent upward and downward trends over time. Furthermore, there is variation in the size of these variations. Moreover, a distinct pattern emerges from the data, suggesting a cumulative rise in the yield of bananas.

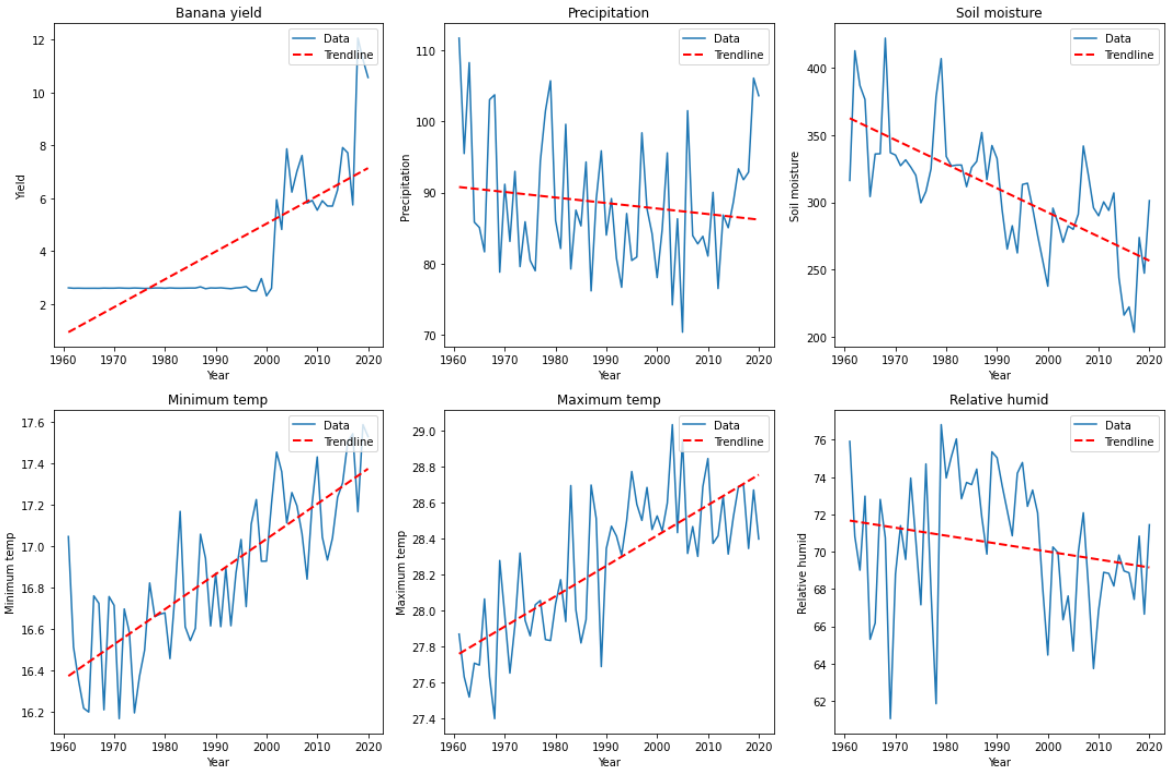


Figure 6: The plot illustrates historical temporal trends of climate variables and banana yield from 1961 to 2020

As shown in Fig. 7, the study computed the correlation matrix between the input variables and the target variable yield and then displayed the correlation matrix as a heatmap. All of the input variables can be chosen for modeling, or the uncertainty quantification process, since they are all correlated with yield according to the correlation matrix (Yan *et al.*, 2013).

As was previously mentioned, further data analysis is possible by examining the plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF), as shown in Fig. 8. It is clear from the ACF plots that there are substantial autocorrelations at various lags that fall outside of the 95% confidence interval. This suggests a strong relationship between the climate variables or current banana yield and their historical values at these various specific lags.

Similarly, significant partial autocorrelations that fall outside of the 95% confidence interval are seen in the PACF plots at various lags (Fig. 8). Even after taking into consideration the effects of other lags, this suggests a strong relationship between the current yield, or climate variables, and their past values at these various lags. Eventually, the strategies for uncertainty quantification and sensitivity analysis can be further informed by the insights obtained from the ACF and PACF plots.

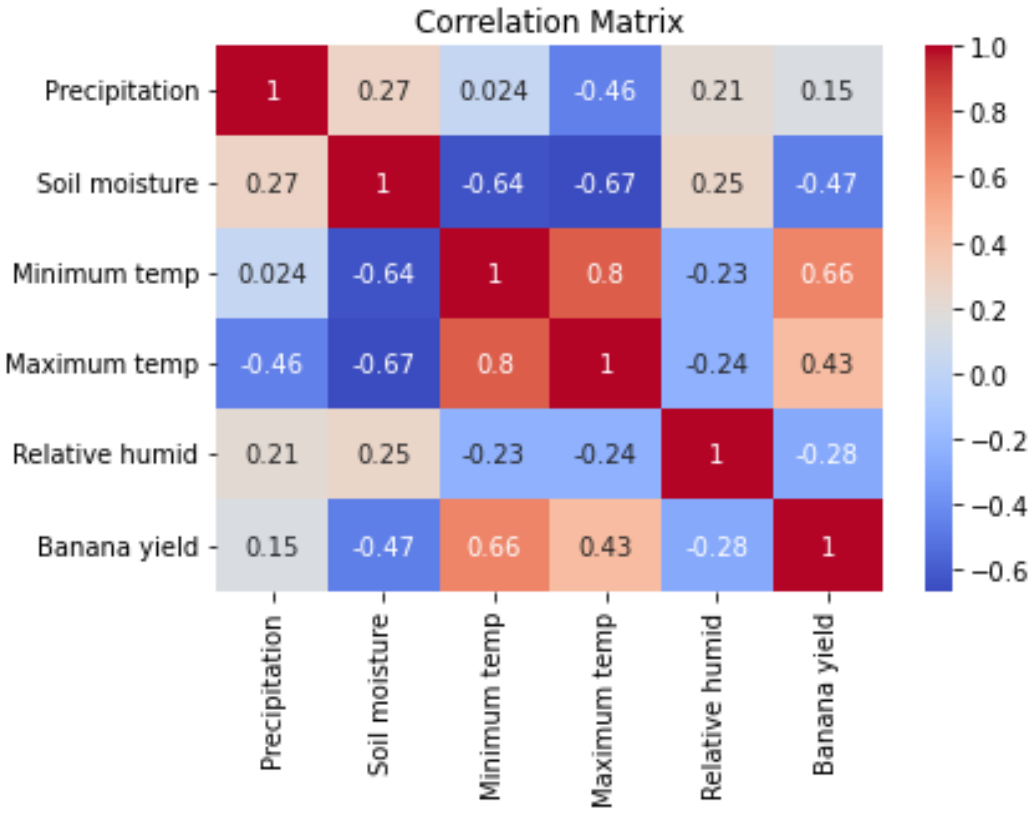


Figure 7: Correlation matrix of the datasets

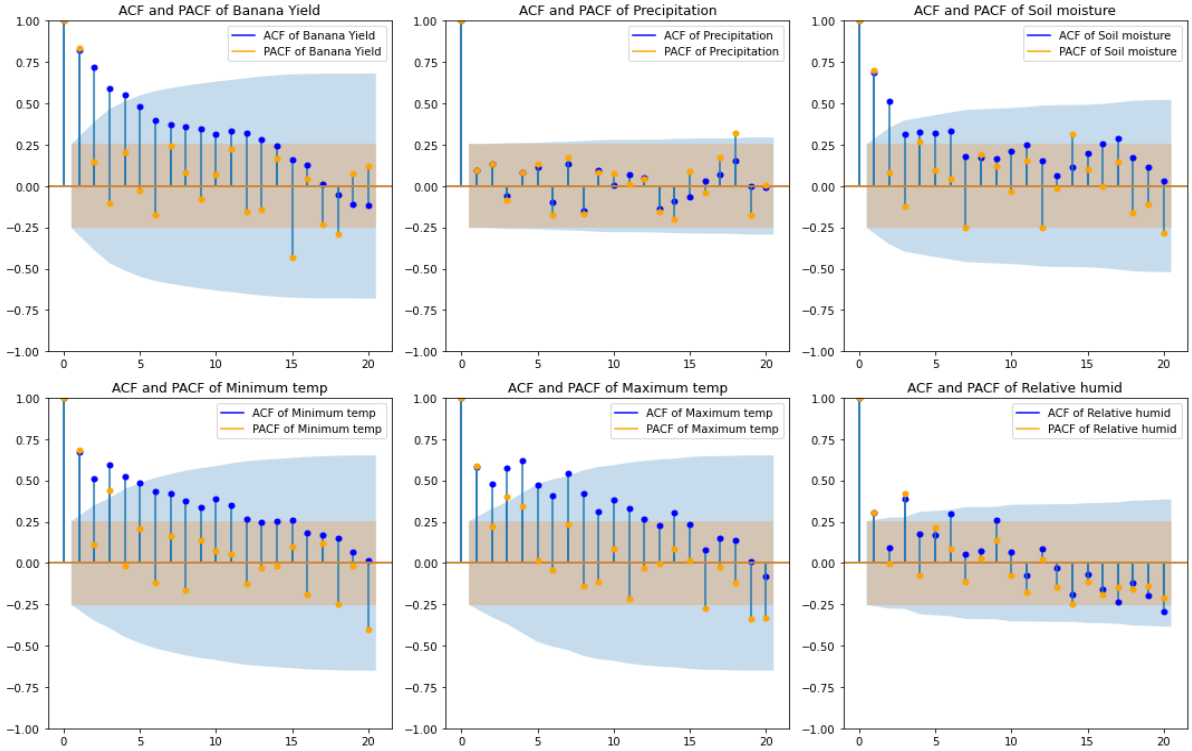


Figure 8: The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots

4.7 Results of SARIMAX model

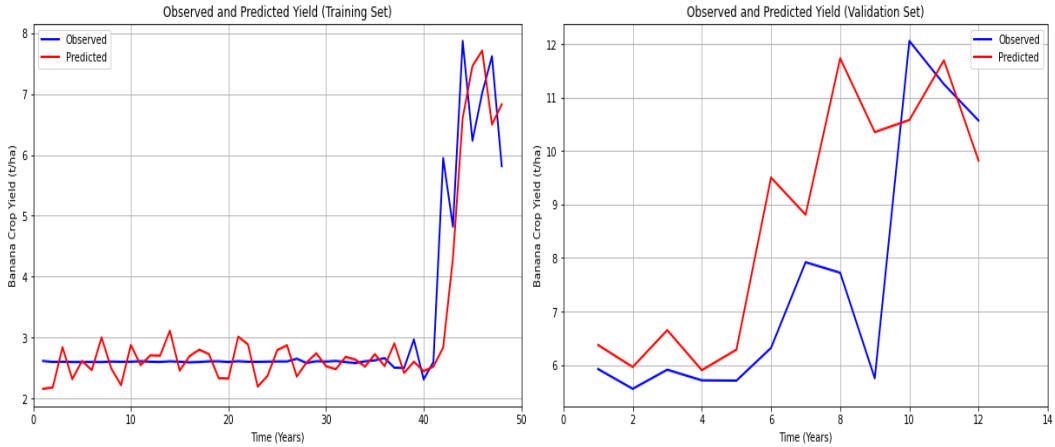
The SARIMAX model for banana crop yield was set up. The recommended SARIMAX models, derived from the data exploration findings, were: SARIMAX$(0,1,1)(0,1,1)_{12}$, SARIMAX$(0,1,1)(0,1,0)_{12}$, SARIMAX$(0,1,2)(0,1,1)_{12}$, and SARIMAX$(0,1,2)(0,1,0)_{12}$. The model met all of the requirements of the Box-Jenkins technique, including model fitting, which included parameter estimation (estimates are shown in Table 9), model identification (SARIMAX$(0,1,2)(0,1,0)_{12}$) model was selected), and diagnostic checking.

Table 9: Estimated Parameters for SARIMAX$(0,1,2)(0,1,0)_{12}$ Model

R^2**(Training Set)**		R^2**(Validation Set)**		**AIC**	**BIC**
0.8109		0.1825		91.469	103.911
Variable	**Coef.**	**Std. Err.**	**t-stat**	**P-value**	**[0.025, 0.975]**
X_1	0.0166	0.032	0.510	0.610	[-0.047, 0.080]
X_2	-0.0016	0.006	-0.243	0.808	[-0.014, 0.011]
X_3	-0.1055	1.421	-0.074	0.941	[-2.891, 2.680]
X_4	0.0635	1.253	0.051	0.960	[-2.392, 2.519]
X_5	0.0110	0.068	0.162	0.872	[-0.122, 0.144]
ma.L1	-0.3858	0.177	-2.181	0.029	[-0.733, -0.039]
ma.L2	0.5397	0.184	2.939	0.003	[0.180, 0.900]
sigma2	0.4894	0.188	2.604	0.009	[0.121, 0.858]

As shown in Fig. 9(a), the predicted crop yields in the training set for the first 40 years closely match the observed crop yields. The aforementioned observation suggests that the model is effectively detecting the fundamental patterns present in the data. For the first four years, the validation set's predicted crop yields and observed crop yields closely match, demonstrating the model's strong performance on unobserved data. The fact that the predictions and actual values match indicates that the model can satisfactorily generalize to new data points.

On the other hand, there is a noticeable difference between the predicted and observed crop yields from 4 to 8 years, as shown in Fig. 9(b). However, the model performs well over a period of 8 to 10 years, suggesting that it may be able to identify the underlying patterns in the validation data during that time.



(a) Training Set **(b) Validation Set**

Figure 9: The Observed and Predicted Banana crop yield for the SARIMAX model

Lastly, the model forecasts yields for the following ten time steps. The last values of 9.8245 and 10.5738 from the validation set were utilized as the starting points for scenarios 1 and 2, respectively. The model then predicts the subsequent value iteratively using the preceding prediction. The following are the forecasted yields for Scenario 1:

6.3705, 5.9595, 6.6489, 5.9003, 6.2839, 9.5054, 8.8109, 11.7376, 10.3568, 10.5829. The crop yields for Scenario 1 over a forecast horizon of ten time steps are represented by these values. Scenario 1's predicted yields point to a pattern of varying values. The yields begin at 6.37, drop to 5.96, rise to 6.65, and then drop once more to 5.90. The yields that follow fluctuate even more, peaking at 11.74 before leveling off between 10.36 and 10.58. For Scenario 2, the forecasted yields are as follows:

10.5738, 6.3705, 6.3542, 6.519, 6.3192, 6.4653, 6.5113, 6.6108, 6.6679, 6.6528. The predicted yields in Scenario 2 follow a different pattern. The model's starting value was set to 10.5738, the final observed value from the validation set. But the later forecasted yields deviate from this starting point and progressively decline. The yields exhibit a consistent downward trend, with a range of 6.32 to 6.67. In general, the model forecasts that crop yields in Scenario 1 and Scenario 2 are somewhat lower. The plots in Fig. 10 below provide a thorough analysis of the model's performance by showing the predicted and forecasted yields visually:

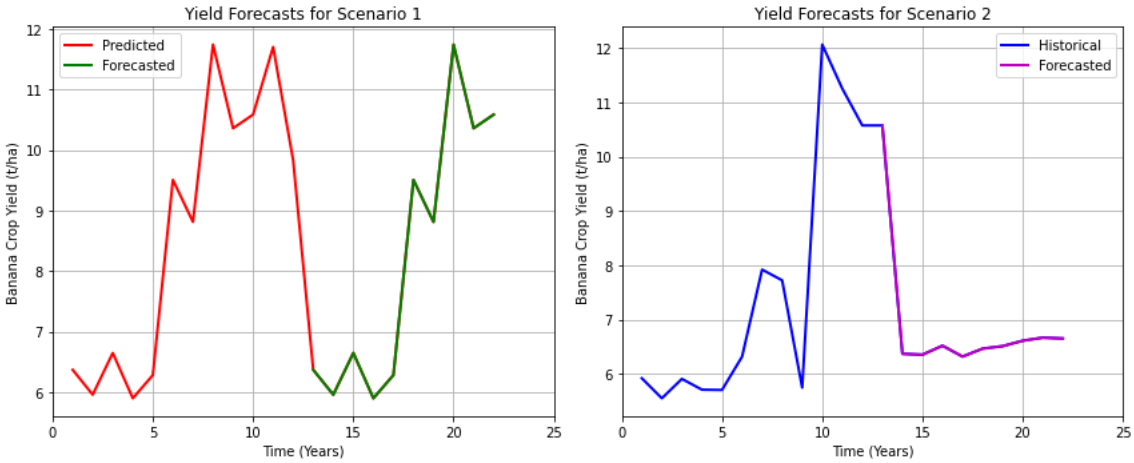


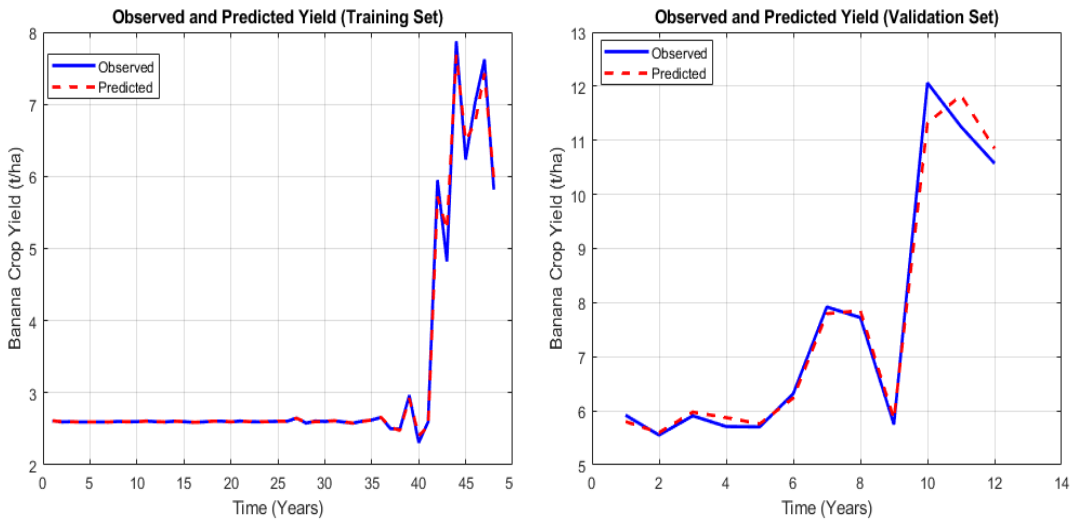
Figure 10: The plot of banana crop yield forecasting for the SARIMAX model

4.8 Results of State Space (SS) Model

The State Space concept was used in its execution. The state vector, state transition matrix, observation matrix, process noise covariance matrix, and measurement noise covariance matrix of the State Space model were then defined. These components are essential for specifying the model's dynamics and uncertainty characteristics. The state covariance matrix was created using the identity matrix, and arrays were initialized to store the results. Table 2 also displays the outcomes of the parameter estimate for the state space model based on the Kalman filter algorithm.

With low prediction errors (MSE and MAE), a small standard deviation of errors (RMSE), and a high proportion of explained variability (R-squared), the SS model generally exhibits a strong match to the training set of data. When the model is used on the validation set, however, its performance is marginally worse, with somewhat higher prediction errors and a somewhat lower coefficient of determination. To validate the model's performance, the training and validation plots in Fig. 11 compare the yields of the crops as predicted and as observed.

There are some differences between the observed and predicted yields in Fig. 11; the model determines the trend of the observed yields. The model's predictions are not as accurate for the validation set as they were for the training set, as evidenced by the red dashed line, which displays some deviations and discrepancies from the blue line. Plots show that the crop yields are accurately predicted by the state space model, especially for the training set. The model demonstrates a strong ability to accurately represent the observed yields' trends and fluctuations.



(a) Training Set **(b) Validation Set**

Figure 11: The Observed and Predicted Banana crop yield for the SS model

Using the final states of 10.8487 and 10.5738 from the predicted yield for scenario 1 and the true yield values for scenario 2 as the initial states, the SS model then forecasts for the following 10 time steps. The following are the forecasted yields for Scenario 1:

7.8510, -0.0261, -15.7032, -42.9991, -86.7473, -152.9127, -248.7089, -382.7142, -564.9889, -807.1921.These forecasted yields represent the crop yields predicted for the next 10 time steps based on the SS model and using the final predicted and actual yield as input. The forecasted yields exhibit a declining trend, with the magnitudes growing increasingly negative over time. The negative values point to a possible decline in crop yields over time due to unfavorable conditions or other factors influencing crop growth. For Scenario 2, the forecasted yields are as follows:

5.9203, 5.5512, 5.9086, 5.7096, 5.7043, 6.3168, 7.9205, 7.7233, 5.7479, 12.0627. These forecasted yields indicate the crop yields that are anticipated over the course of the next 10 time steps, based on the state space model and the final true yield as input. Although there are some fluctuations, there is no clear trend in the forecasted yields. The values show comparatively constant or stable crop yields over time because they vary within a small range. Plots showing the predicted and forecasted yields visually are shown below (Fig. 12), enabling a thorough examination of the model's performance.

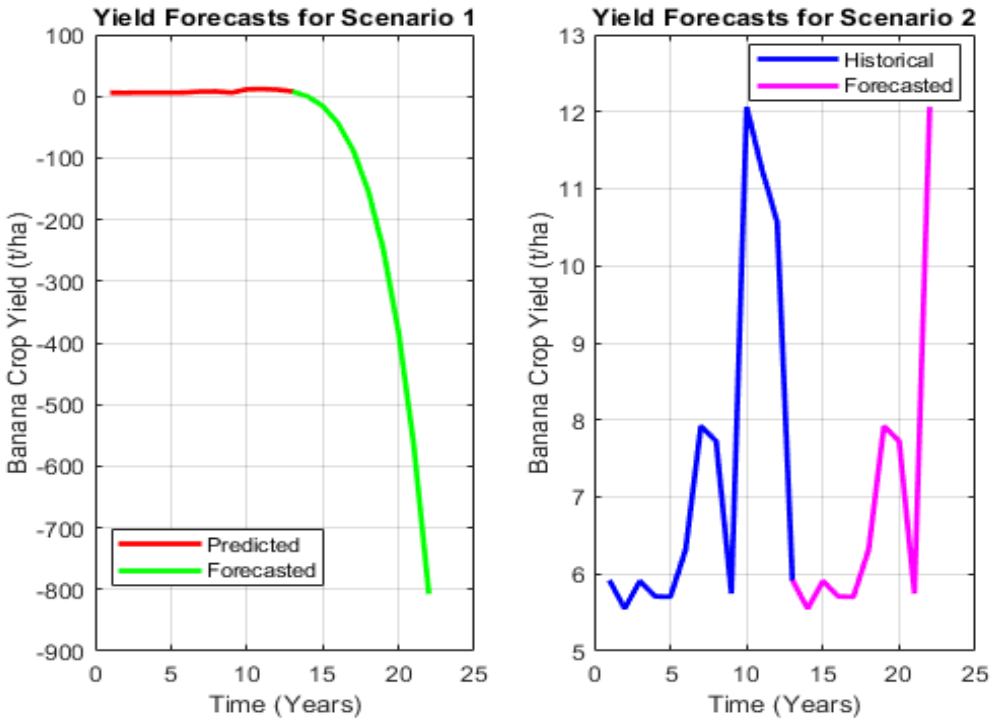
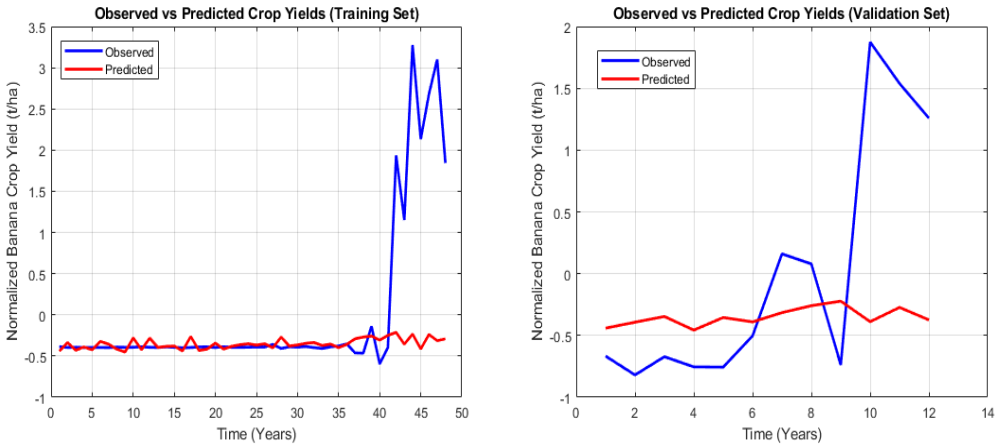


Figure 12: The plot of banana crop yield forecasting for the State Space model

4.9 Results of LSTM Model

The LSTM model was set up. After generating the sequences and labels required for the LSTM model, the model was built and trained. Then predictions were made, and the inverse transform was used to return the scaled predictions to their initial form. The LSTM model performs well on the given dataset, according to the R-squared value for the training data. Note that Table 2 displays the MSE, MAE, RMSE, and R-squared values for every model used in this investigation. These metrics shed light on each model's accuracy and performance. Overall, the evaluation results show that the LSTM network used in this analysis fits the data well, as shown by the high coefficient of determination (R-squared) and low errors (MSE, MAE, RMSE).

Crop yields for the training and validation sets were predicted by the trained LSTM network. The model's fit was assessed by comparing the predicted and actual yields. To assess the model's fit, the actual yields and predicted yields were contrasted. In order to compare the trends and performance of the training and validation sets visually, the observed and predicted crop yields were plotted, as Fig. 13 illustrates.



(a) Training Set **(b) Validation Set**

Figure 13: The Observed and Predicted Banana crop yield for the LSTM model

As shown in Fig. 13(a), the predicted crop yields in the training data for the first 40 years closely match the observed crop yields. The aforementioned observation suggests that the model is effectively detecting the fundamental patterns present in the data. Conversely, the significant discrepancy between the predicted and observed crop yields after the 40s raises the possibility that the model is predicting low values of yields in the training set.

On the other hand, Fig. 13(b) shows that for the first ten years, the predicted crop yields agree quite well with the observed crop yields, suggesting that the model works well with unknown data. This agreement between predicted and observed values shows that the model can reasonably generalize to new data points. However, after ten years, there is a significant difference between the predicted and actual crop yields, indicating small values of yields in the validation set.

Lastly, the model forecasts yields for the following 10 time steps. The model iteratively predicts the subsequent value based on the previous prediction. The last predicted and true values, which are -0.3610 and 10.5738 from the validation set, were used as the initial inputs for scenario 1 and scenario 2, respectively. The following are the forecasted yields for Scenario 1:

-0.2636, -0.0950, -0.3842, -0.1814, -0.2485, -0.2660, -0.2304, -0.2616, -0.2469, -0.2483. The crop yields for Scenario 1 over a forecast horizon of 10 time steps are represented by these values. These values represent the predicted yields in relation to the range of yields seen in the validation data since they are normalized yields. The negative values indicate that the predicted yields are expected to be lower than the validation dataset's average yield. For the anticipated time steps in Scenario 1, the model forecasts comparatively low crop yields.For Scenario 2, the forecasted yields are as follows:

-1.2502, -0.3620, -0.4858, -0.6064, -0.5217, -0.3284, -0.2751, -0.2657, -0.2618, -0.2626. The crop yields for Scenario 2 over a forecast horizon of 10 time steps are represented by these values. The range of values is -0.2618 to -1.2502. These values show the predicted yields in relation to the validation data and are normalized yields, just like in Scenario 1. In comparison to Scenario 1, the model predicts even lower crop yields in Scenario 2. In general, the model predicts that crop yields in Scenario 1 and Scenario 2 will be lower. The plots (Fig. 14) that illustrate the predicted and forecasted yields are shown below:

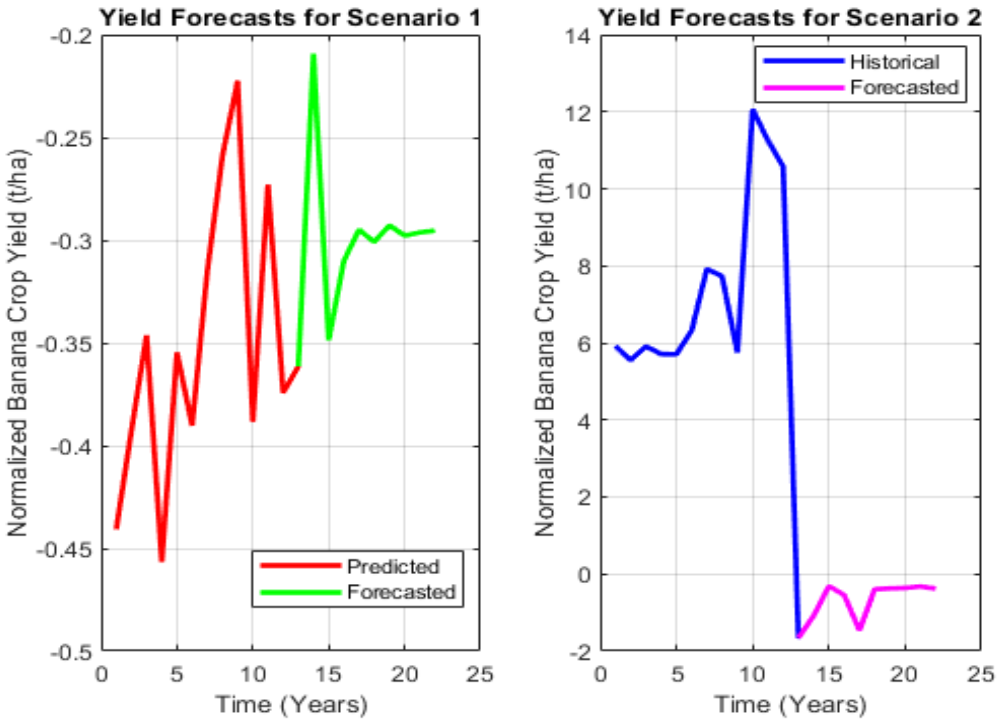


Figure 14: The plot of banana crop yield forecasting for the LSTM model

4.10 Results of Ensemble model

As previously mentioned, after the SARIMAX, State Space, and LSTM models have been trained and assessed on the datasets, we can move forward with figuring out the weights and getting the final predicted values in order to create an ensemble model for predicting banana crop yield. Based on how each model performed on the validation set, the study calculated its weights. Taking into account each model's degree of data fit, the study arrived at the weights. A useful metric for assessing the overall goodness of fit and the degree to which the model captures the variability in the data is the R-squared (Coefficient of Determination). Table 2 shows the R-squared values for the SARIMAX, State Space, and LSTM models, which are, respectively, 0.1825, 0.9835, and 0.9013.

When calculating the final predicted value, the ensemble model is represented as follows:

$$y = 0.7204 \times 9.8245 + 0.1337 \times 10.8487 + 0.1459 \times -0.3610 \rightarrow y = 8.4754. \quad (4.2)$$

The LSTM, State Space, and SARIMAX models have normalized weights of 0.1459, 0.1337, and 0.7204, respectively. The final estimated values are -0.3610, 10.8487, and 9.8245 from the State Space, SLTM, and SARIMAX models, respectively. The outputs from the individual SARIMAX, State Space, and LSTM models are combined to yield the final predicted value of 8.4754 for the ensemble model. The individual models each have different predictions, State Space predicts the highest value of 10.8487, followed by SARIMAX with 9.8245 and LSTM with -0.3610. To get the final predicted value, the ensemble model uses weights that are optimized during validation to combine the predictions of these individual models. Because it considers the advantages and disadvantages of the individual models to produce a more accurate prediction, the ensemble model can thus be thought of as a more reliable and accurate model (Bertsimas & Boussioux, 2023).

As shown in Table 10, the final step involved assessing the ensemble model's performance using pertinent metrics. This involved evaluating the efficacy of the ensemble model by contrasting its performance with that of the individual models. Therefore, the R-squared values of SARIMAX, State Space, LSTM, and Ensemble models are, in order, 0.1825, 0.9835, 0.9013, and 0.9999. Here a gauge of each model's effectiveness using the validation data is provided by these R-squared values. A model's ability to predict actual values is demonstrated by a higher R-squared (coefficient of determination) value.

In this case, it appears that the ensemble model performed better on the validation data because it had the highest R-squared value of any other model. The SARIMAX model performs the worst out of all the models, as seen by its lowest R-squared value. R-squared values for the LSTM and State Space models are somewhat similar, with the State Space model outperforming the LSTM model.

Table 10: Evaluation Metrics for the Ensemble Model

Metric	Value
Mean Squared Error (MSE)	8.35788099957876e-10
Mean Absolute Error (MAE)	2.8029999999290567e-05
Root Mean Squared Error (RMSE)	5.2945999999290567e-05
R-squared	0.9999

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

To sum up, this dissertation offers a thorough comprehension of the complex relationship between Tanzania's banana crop yield and climate change. The study used a range of analytical techniques and tools to evaluate how climate variables affected banana production, highlighting vulnerabilities and directing adaptive measures.

Examining Tanzania's banana crop yield's sensitivity to climate variables was the primary goal of the study. The study clarified the complex relationships between these variables by employing a multiple regression model and advanced global sensitivity analysis techniques, including Sobol' Sensitivity Indices and Response Surface Methodology. Using Monte Carlo simulation to quantify uncertainty strengthened the findings' dependability. This analysis has yielded invaluable knowledge that will help farmers, policymakers, and stakeholders make critical decisions.

However, the study adopted a different strategy, concentrating on the configuration and forecasting of banana crop yield while taking climate change into account. It combined powerful statistical tools like correlation analysis, time series analysis, and ensemble modeling to accurately forecast banana production in Tanzania, taking into account the influence of climate and the complexities of bananas. The results demonstrated how crucial it is to take climate change into account when predicting banana crop yields. Actionable policy recommendations and strategies to safeguard and improve Tanzania's banana production are based on these insights. Promoting climate-resilient practices, data-driven decision-making, infrastructure investments, policy flexibility, knowledge dissemination, and continuous research all depend on cooperation between researchers, policymakers, and farmers.

Overall, the study emphasizes how important it is for climate factors to continue influencing Tanzania's banana production in the future. It highlights the necessity of taking proactive steps to lessen the effects of climate change and advance sustainable agriculture. This dissertation adds to the larger endeavor of guaranteeing food security and prosperity for Tanzania's agricultural communities in the face of changing climate challenges by bringing attention to these issues and offering practical suggestions.

5.2 Limitations

While providing insightful information about the relationship between climate and yield, the research highlights the complexity of the issue and the need for adaptive strategies. Therefore, challenges in this study include handling uncertainties and the dynamic nature of climate. The research encourages informed decision-making for sustainable banana production and food security. Despite its strengths, the study acknowledges limitations in the inclusivity of climate variables and potential weaknesses in forecasting accuracy. Though, the study does admit that not all relevant climate variables were covered. Nonetheless, the study assert the relevance of these crucial climate variables to this research. Unfortunately, the study relies on low-quality and non-credible FAOSTAT data due to the unavailability of alternative sources at the subnational level for banana yield data. Such data is essential for establishing accurate forecasting models, which would enhance the analysis and findings of this study. In essence, this study contributes to understanding the climate-yield nexus and emphasizes the importance of proactive adaptation in the face of changing environments.

5.3 Recommendations

This study revealed that Tanzania's banana crop yield has been impacted by climate change, providing insights into potential vulnerabilities in light of these important climate variables at hand. In Particular, *"National food production is projected to decrease by 8 –13 per cent by 2050 due to increased heat stress, drying, erosion, and flood damage, as well as post-harvest loss"*, as quoted from the *"National Climate Change Response Strategy 2021-2026"* (URT, 2021). The knowledge gained from this research provides a vital basis for practical policy suggestions and tactics to protect and improve Tanzania's banana crop in the face of climate change. For this research to have practical implications, ties to the community must be established. This calls for urgent cooperation between researchers, policymakers, and farmers in order to put the following strategies into action:

(i) *Climate-Resilient Farming Practices:* Encourage and facilitate the adoption of climate-resilient farming practices among banana growers. This includes promoting drought-resistant banana varieties, optimizing irrigation systems, and disseminating information on weather-smart farming techniques.

(ii) *Data-Driven Decision Making:* Foster data-driven decision-making processes by making the forecasting models and climate data from this study readily accessible to farmers and local agricultural authorities. Empower them to make informed choices on planting schedules, resource allocation, and risk mitigation. Exploring opportunities to discover

alternative sources of subnational-level data on banana yields is crucial. This effort aims to enhance the quality of data used in similar studies, addressing the limitations imposed by relying on poor-quality and non-credible FAOSTAT data.

(iii) *Investment in Infrastructure:* Invest in critical infrastructure, such as improved water management systems and post-harvest facilities, to minimize yield losses due to climate-induced disruptions. Strengthening the agricultural supply chain is essential for ensuring that bananas reach markets efficiently.

(iv) *Policy Flexibility:* Develop adaptive agricultural policies that can respond to evolving climate conditions. Policymakers should be prepared to adjust policies in response to changing climate realities and emerging challenges.

(v) *Knowledge Dissemination:* Conduct workshops, training programs, and outreach activities for farmers and stakeholders to raise awareness about climate change hazards and the measures available for mitigation and adaptation. Building local knowledge and capacity is essential for resilience.

Ultimately, Tanzania can strengthen its banana production industry against the disruptive effects of climate change by putting these recommendations into practice. By working together, stakeholders can ensure food security and prosperity for the country's agricultural communities by promoting sustainable banana production.

5.4 Future Work

To encourage further research endeavors that expand upon the findings of this study. Future research should consider additional climate variables, management factors, improving the analysis by setting up the models at the subnational level and finding an alternative source of data on banana yields, incorporate artificial intelligence (AI) and machine learning techniques, and assess the prospective impacts of climate change on banana crop yield using advanced impact assessment and climate modeling approaches.

REFERENCES

Abdoussalami, A., Hu, Z., Islam, A. R. M. T., & Wu, Z. (2023). Climate change and its impacts on banana production: A systematic analysis. *Environment, Development and Sustainability*, 1–30.

Adejuwon, J., & Agundiminegha, Y. (2019). Impact of climate variability on cassava yield in the humid forest agro-ecological zone of Nigeria. *Journal of Applied Sciences and Environmental Management*, *23*(5), 903–908.

Amin, M., Qasim, M., Afzal, S., & Naveed, K. (2022). New ridge estimators in the inverse Gaussian regression: Monte Carlo simulation and application to chemical data. *Communications in Statistics-Simulation and Computation*, *51*(10), 6170–6187.

Antanasijević, D., Pocajt, V., Perić-Grujić, A., & Ristić, M. (2014). Modelling of dissolved oxygen in the Danube river using artificial neural networks and Monte Carlo simulation uncertainty analysis. *Journal of Hydrology*, *519*, 1895–1907.

Anwar, S. A., Zakey, A., Robaa, S., & Abdel Wahab, M. (2019). The influence of two land-surface hydrology schemes on the regional climate of Africa using the RegCM4 model. *Theoretical and Applied Climatology*, *136*, 1535–1548.

Anzures, A. F., Hipolito, K., & Pestolante, K. (2022). Constraints in the primary production of bananas in the Davao Region, Philippines. *International Journal of Social and Management Studies*, *3*(1), 1–31.

Aoki, M. (2013). *State space modeling of time series*. Springer Science & Business Media.

Arunraj, N. S., Ahrens, D., & Fernandes, M. (2016). Application of SARIMAX model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems*, *7*(2), 1–21.

Bayati, A., Nguyen, K. K., & Cheriet, M. (2020). Gaussian process regression ensemble model for network traffic prediction. *IEEE Access*, *8*, 176540–176554.

Bertsimas, D., & Boussioux, L. (2023). Ensemble modeling for time series forecasting: An adaptive robust optimization approach. *arXiv preprint arXiv:2304.04308*.

Bhausaheb, T. A., Lazarus, T. P., Vijayan, A., Sathayan, A. R., & Joseph, B. (2023). Impact of climate change on banana production in Thiruvananthapuram District of Kerala, India. *Asian Journal of Agricultural Extension, Economics & Sociology*, *41*(3), 114–123.

Bhimavarapu, U., Battineni, G., & Chintalapudi, N. (2023). Improved optimization algorithm in LSTM to predict crop yield. *Computers*, *12*(1), 10.

Borgonovo, E., & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, *248*(3), 869–887.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.

Chi, Y. N., & Chi, O. (2021). Modeling and forecasting of monthly global price of bananas using seasonal ARIMA and multilayer perceptron neural network. *Econometrics/Ekonometria*, *25*(3).

Chowhan, S., Ghosh, S. R., Chowhan, T., Hasan, M. M., & Roni, M. S. (2016). Climate change and crop production challenges: An overview. *Research in Agriculture Livestock and Fisheries*, *3*(2), 251–269.

Cook, E. R. (1985). *A time series analysis approach to tree ring standardization* [Doctoral dissertation, University of Arizona Tucson].

CRU. (2023). The monthly gridded temperature, minimum temperature, and maximum temperature datasets for the reanalysis were provided by the University of East Anglia's Climatic Research Unit (CRU). *Homepage: https://data.ceda.ac.uk/badc/cru/data/cru_ts/cru_ts_4.05 (online)*.

de Bézenac, E., Rangapuram, S. S., Benidis, K., Bohlke-Schneider, M., Kurle, R., Stella, L., Hasson, H., Gallinari, P., & Januschowski, T. (2020). Normalizing Kalman filters for multivariate time series analysis. *Advances in Neural Information Processing Systems*, *33*, 2995–3007.

Dega, S., Dietrich, P., Schrön, M., & Paasche, H. (2023). Probabilistic prediction by means of the propagation of response variable uncertainty through a Monte Carlo approach in regression random forest: Application to soil moisture regionalization. *Frontiers in Environmental Science*, *11*, 53.

FAOSTAT. (2023). The average annual banana crop yield statistics were sourced from the FAOSTAT database. *Homepage: https://www.fao.org/faostat/en/#data/QCL (online)*.

Field, C. B., & Barros, V. R. (2014). *Climate change 2014–impacts, adaptation and vulnerability: Regional aspects*. Cambridge University Press.

Hanson, T. (2010). *Multiple regression*. Springer Science & Business Media.

Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, *43*(4), 549–558.

Hooda, E., Verma, U., & Hooda, B. (2020). ARIMA and State-Space models for sugarcane (saccharum officinarum) yield forecasting in Northern agro-climatic zone of Haryana. *Journal of Applied and Natural Science*, *12*(1), 53–58.

Hoque, M., & Haque, M. (2016). Impact of climate change on crop production and adaptation practices in coastal saline areas of Bangladesh. *International Journal of Applied Research*, *2*(1), 10–19.

Hu, Y., Liu, S., Lu, H., & Zhang, H. (2019). Remaining useful life model and assessment of mechanical products: A brief review and a note on the state space model method. *Chinese Journal of Mechanical Engineering*, *32*, 1–20.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Ighalo, J. O., & Adeniyi, A. G. (2019). Thermodynamic modelling and temperature sensitivity analysis of banana (musa spp.) waste pyrolysis. *SN Applied Sciences*, *1*(9), 1–9.

Iooss, B., & Lemaître, P. (2015). A review on global sensitivity analysis methods. *Uncertainty management in simulation-optimization of complex systems: Algorithms and applications*, 101–122.

Jankovic, A., Chaudhary, G., & Goia, F. (2021). Designing the design of experiments (DOE)–An investigation on the influence of different factorial designs on the characterization of complex systems. *Energy and Buildings*, *250*, 111298.

Jayasinghe, S., Ranawana, C., Liyanage, I., & Kaliyadasa, P. (2022). Growth and yield estimation of banana through mathematical modelling: A systematic review. *The Journal of Agricultural Science*, 1–58.

Jou, Y. T., Lin, W. T., Lee, W. C., & Yeh, T. M. (2014). Integrating the taguchi method and response surface methodology for process parameter optimization of the injection molding. *Applied Mathematics & Information Sciences*, *8*(3), 1277.

Kahimba, F., Sife, A., Maliondo, S., Mpeta, E., & Olson, J. (2015). Climate change and food security in Tanzania: Analysis of current knowledge and research gaps. *Tanzania Journal of Agricultural Sciences*, *14*(1).

Kamir, E., Waldner, F., & Hochman, Z. (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, 124–135.

Kleijnen, J. P. (2014). Response surface methodology. In *Handbook of simulation optimization* (pp. 81–104). Springer.

Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, *41*(9), 4235–4244.

Kucherenko, S., & Song, S. (2016). Derivative-based global sensitivity measures and their link with Sobol' sensitivity indices. *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, 455–469.

Lai, Y., & Dzombak, D. A. (2020). Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation. *Weather and Forecasting*, *35*(3), 959–976.

Lal, N., Sahu, N., Shuirkar, G., Jayswal, D. K., & Chack, S. (2017). Banana: Awesome fruit crop for society.

Li, W., Lin, G., & Li, B. (2016). Inverse regression-based uncertainty quantification algorithms for high-dimensional models: Theory and practice. *Journal of Computational Physics*, *321*, 259–278.

Liu, F., Jiang, X., & Wu, Z. (2023). Attention mechanism-combined LSTM for grain yield prediction in China using multi-source satellite imagery. *Sustainability*, *15*(12), 9210.

Lokupitiya, E. (2018). Book of abstracts of 2nd international conference on climate change 2018 (ICCC 2018). Climate Change Conference. Colombo, Sri Lanka: The international institute of knowledge management (TIIKM).

Lucas, S. S., & Jomanga, K. E. (2021). The status of banana production in Tanzania: A review of threats and opportunities.

Marolla, F., Henden, J. A., Fuglei, E., Pedersen, Å. Ø., Itkin, M., & Ims, R. A. (2021). Iterative model predictions for wildlife populations impacted by rapid climate change. *Global Change Biology*, *27*(8), 1547–1559.

Mayaya, H. K. (2015). *Community adaptation and mitigation strategies to climate change in semi-arid areas of Dodoma region, Tanzania* [Doctoral dissertation, School of Environmental Studies in Partial Fulfilment of the Requirements.].

Mbigi, D., & Xiao, Z. (2021). Analysis of rainfall variability for the October to December over Tanzania on different timescales during 1951–2015. *International Journal of Climatology*, *41*(14), 6183–6204.

Meeradevi, Yasaswi, I., Mundada, M. R., Sarika, D., & Shetty, H. (2022). Hybrid decision support system framework for enhancing crop productivity using machine learning. *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, 57–66.

Moore, F. C., & Lobell, D. B. (2014). Adaptation potential of European agriculture in response to climate change. *Nature Climate Change*, *4*(7), 610–614.

NCEP/NCAR. (2023). The soil moisture and relative humidity data were obtained from the NCEP/NCAR Reanalysis dataset. *Homepage: https://psl.noaa.gov/data/gridded/reanalysis/ (online)*.

Neog, B., Gogoi, B., & Patowary, A. (2022). Development of hybrid time series models for forecasting autumn rice using ARIMAX-ANN and ARIMAX-SVM. *Annals of Forest Research*, *65*(1), 9119–9133.

Newman, K., King, R., Elvira, V., de Valpine, P., McCrea, R. S., & Morgan, B. J. (2023). State-space models for ecological time-series data: Practical model-fitting. *Methods in Ecology and Evolution*, *14*(1), 26–42.

Ngo, T. H. D., & La Puente, C. (2012). The steps to follow in a multiple regression analysis. *Proceedings of the SAS Global forum*, 22–25.

Nwabueze, T. U. (2010). Basic steps in adapting response surface methodology as mathematical modelling for bioprocess optimisation in the food systems. *International Journal of Food Science & Technology*, *45*(9), 1768–1776.

Ongoma, V., Chen, H., & Gao, C. (2019). Evaluation of CMIP5 twentieth century rainfall simulation over the equatorial East Africa. *Theoretical and Applied Climatology*, *135*(3-4), 893–910.

Owen, A. B. (2013). Better estimation of small Sobol' sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *23*(2), 1–17.

Pham, Y., Reardon-Smith, K., Mushtaq, S., & Cockfield, G. (2019). The impact of climate change and variability on coffee production: A systematic review. *Climatic Change*, *156*, 609–630.

Rahn, E., Vaast, P., Läderach, P., Van Asten, P., Jassogne, L., & Ghazoul, J. (2018). Exploring adaptation strategies of coffee production to climate change using a process-based model. *Ecological Modelling*, *371*, 76–89.

Raj, E. E., Ramesh, K., & Rajkumar, R. (2019). Modelling the impact of agrometeorological variables on regional tea yield variability in South Indian tea-growing regions: 1981-2015. *Cogent Food & Agriculture*, *5*(1), 1581457.

Rathod, S., & Mishra, G. (2018). Statistical models for forecasting mango and banana yield of Karnataka, India. *Journal of Agricultural Science and Technology*, *20*(4), 803–816.

Reddy, M. P., Mathur, A. K., Jain, R. K., Agarwal, S. K., & Singh, S. (2022). Climate change and weather variability in crop modelling: Evidence from rice yield trials in India using LSTM model.

Reji, M., & Kumar, R. (2022). Response surface methodology (RSM): An overview to analyze multivariate data. *Indian Journal of Microbiology Research*, *9*, 241–248.

Sabiiti, G., Ininda, J. M., Ogallo, L., Opijah, F., Nimusiima, A., Otieno, G., Ddumba, S. D., Nanteza, J., & Basalirwa, C. (2016). Empirical relationships between banana yields and climate variability over Uganda. *Journal of Environmental and Agricultural Sciences*, *7*, 3–13.

Sadowski, P. (2016). Notes on backpropagation. *Homepage: https : // www . ics . uci . edu / pjsadows/notes.pdf (online)*.

Sagamiko, T., Shaban, N., & Mbalawata, I. (2020). Sensitivity analysis and uncertainty parameter quantification in a regression model: The case of deforestation in Tanzania. *Tanzania Journal of Science*, *46*(3), 673–683.

Salvacion, A. R. (2020). Effect of climate on provincial-level banana yield in the Philippines. *Information Processing in Agriculture*, *7*(1), 50–57.

Shirima, A. O., & Lubawa, G. (2017). Farm based adaptation strategies to climate change among smallholder farmers in Manyoni District, Tanzania. *International Journal of Research in Social Sciences*, *7*(7), 1–22.

Suman, S., & Verma, U. (2017). State space modelling and forecasting of sugarcane yield in Haryana, India. *Journal of Applied and Natural Science*, *9*(4), 2036–2042.

Tahmasebinia, F., Jiang, R., Sepasgozar, S., Wei, J., Ding, Y., & Ma, H. (2022). Implementation of BIM energy analysis and Monte Carlo simulation for estimating building energy performance based on regression approach: A case study. *Buildings*, *12*(4), 449.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., & Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong plain, PR China. *Agricultural and Forest Meteorology*, *310*, 108629.

Todorov, V., Dimov, I., Ostromsky, T., Apostolov, S., Georgieva, R., Dimitrov, Y., & Zlatev, Z. (2021). Advanced stochastic approaches for Sobol' sensitivity indices evaluation. *Neural Computing and Applications*, *33*, 1999–2014.

Urquizo, J., Calderón, C., & James, P. (2017). Using a local framework combining principal component regression and Monte Carlo simulation for uncertainty and sensitivity analysis of a domestic energy model in sub-city areas. *Energies*, *10*(12), 1986.

URT. (2021). United Republic of Tanzania (URT). National Climate Change Response Strategy (2021-2026). Vice President's Office, Division of Environment, Government Printer, Dodoma, Tanzania.

Van Leeuwen, S. M., Lenhart, H. J., Prins, T. C., Blauw, A., Desmit, X., Fernand, L., Friedland, R., Kerimoglu, O., Lacroix, G., & Van Der Linden, A. (2023). Deriving pre-eutrophic conditions from an ensemble model approach for the North-West European seas. *Frontiers in Marine Science*, *10*, 1129951.

Varma, V., & Bebber, D. P. (2019). Climate change impacts on banana yields around the world. *Nature Climate Change*, *9*(10), 752–757.

Verma, S. (2018). Modeling and forecasting maize yield of India using ARIMA and state space models. *Journal of Pharmacognosy and Phytochemistry*, *7*(5), 1695–1700.

Wood, S. A., Jina, A. S., Jain, M., Kristjanson, P., & DeFries, R. S. (2014). Smallholder farmer cropping decisions related to climate variability across multiple regions. *Global Environmental Change*, *25*, 163–172.

Yan, X., Guo, J., Liu, S., Cheng, X., & Wang, Y. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. *Proceedings of the 2013 SIAM International Conference on Data Mining*, 749–757.

Yolmeh, M., & Jafari, S. M. (2017). Applications of response surface methodology in the food industry processes. *Food and Bioprocess Technology*, *10*(3), 413–433.

APPENDICES

Appendix 1

Regression Analysis PYTHON Codes

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import
    variance_inflation_factor

# Load the dataset
data = pd.read_csv('Banana1.csv')

# Extract the relevant columns
df = data[['precipitation', 'soilmoisture', 'minimumtemperature', '
    maximumtemperature', 'humidity', 'Yield']]

# Separate the predictor variables (X) and the response variable (Y)
X = df[['precipitation', 'soilmoisture', 'minimumtemperature', '
    maximumtemperature', 'humidity']]
Y = df['Yield']

# Add a constant column to the predictor variables (for the intercept
     term)
X = sm.add_constant(X)

# Fit the multiple regression model
model = sm.OLS(Y, X).fit()

# Calculate the VIF for each predictor variable
vif = pd.DataFrame()
vif["Variable"] = X.columns
vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X
    .shape[1])]

# Print the model summary and VIF
print(model.summary())
print("\nVariance Inflation Factor (VIF):")
print(vif)
```

Appendix 2

SARIMAX Model PYTHON Codes

```
import pandas as pd
import numpy as np
from statsmodels.tsa.statespace.sarimax import SARIMAX

# Load the Banana1 dataset
data = pd.read_csv('Banana1.csv')
Yield = data.iloc[:, 6].values

# Split the data into training and validation sets (80% for training)
train_ratio = 0.8
n = len(Yield)
train_samples = round(train_ratio * n)

Yield_train = Yield[:train_samples]
Yield_val = Yield[train_samples:]

# Define SARIMAX model parameters
order = (0, 1, 2) # SARIMA order (p, d, q)
seasonal_order = (0, 1, 0, 12) # Seasonal order (P, D, Q, S), S = 12,
    S = 4 for monthly and quarterly time series data respectively.
exog_train = data.iloc[:train_samples, 1:6].values # Exogenous
   variables for training set
exog_val = data.iloc[train_samples:, 1:6].values # Exogenous
   variables for validation set

# Fit the SARIMAX model
model = SARIMAX(Yield_train, exog=exog_train, order=order,
   seasonal_order=seasonal_order)
model_fit = model.fit(disp=False)

# Display the estimated parameters with summary statistics
print('Estimated Parameters:')
print(model_fit.summary().tables[1])

# Plot diagnostics (ACF, PACF, etc.)
model_fit.plot_diagnostics(figsize=(10, 8))

# Perform predictions
Yield_pred_train = model_fit.predict(start=0, end=train_samples-1,
   exog=exog_train)
```

```
    Yield_pred_val = model_fit.predict(start=train_samples, end=n-1, exog
        =exog_val)


    # Compute model fit measures for training set
    mse_train = np.mean((Yield_train - Yield_pred_train) ** 2) # Mean
        Squared Error
    mae_train = np.mean(np.abs(Yield_train - Yield_pred_train)) # Mean
        Absolute Error
    rmse_train = np.sqrt(mse_train) # Root Mean Squared Error

    # Compute R-squared for training set
    ss_total_train = np.sum((Yield_train - np.mean(Yield_train)) ** 2) #
        Total sum of squares
    ss_residual_train = np.sum((Yield_train - Yield_pred_train) ** 2) #
        Residual sum of squares
    r_squared_train = 1 - (ss_residual_train / ss_total_train) #
        Coefficient of Determination

    # Compute model fit measures for validation set
    mse_val = np.mean((Yield_val - Yield_pred_val) ** 2) # Mean Squared
        Error
    mae_val = np.mean(np.abs(Yield_val - Yield_pred_val)) # Mean Absolute
         Error
    rmse_val = np.sqrt(mse_val) # Root Mean Squared Error

    # Compute R-squared for validation set
    ss_total_val = np.sum((Yield_val - np.mean(Yield_val)) ** 2) # Total
        sum of squares
    ss_residual_val = np.sum((Yield_val - Yield_pred_val) ** 2) #
        Residual sum of squares
    r_squared_val = 1 - (ss_residual_val / ss_total_val) # Coefficient of
         Determination

    # Display the model fit measures for training set
    print('Training Set:')
    print('Mean Squared Error (MSE):', mse_train)
    print('Mean Absolute Error (MAE):', mae_train)
    print('Root Mean Squared Error (RMSE):', rmse_train)
    print('Coefficient of Determination (R-squared):', r_squared_train)

    # Display the model fit measures for validation set
    print('Validation Set:')
    print('Mean Squared Error (MSE):', mse_val)
```

```
print('Mean Absolute Error (MAE):', mae_val)
print('Root Mean Squared Error (RMSE):', rmse_val)
print('Coefficient of Determination (R-squared):', r_squared_val)

# Plot the observed and predicted Yield for training set
import matplotlib.pyplot as plt

# Insert a dummy point at the beginning of the data
Yield_train_dummy = np.insert(Yield_train, 0, np.nan)
Yield_pred_train_dummy = np.insert(Yield_pred_train, 0, np.nan)

plt.figure(figsize=(10, 6))
plt.plot(Yield_train_dummy, 'b-', linewidth=2, label='Observed')
plt.plot(Yield_pred_train_dummy, 'r-', linewidth=2, label='Predicted'
    )
plt.xlabel('Time (Years)')
plt.ylabel('Banana Crop Yield (t/ha)')
plt.title('Observed and Predicted Yield (Training Set)')
plt.legend()
plt.grid(True)
plt.xlim([0, 50])

# Set x-axis ticks starting from 0 to 50 with a step of 10
plt.xticks(np.arange(0, 51, step=10))

plt.show()

# Plot the observed and predicted Yield for validation set
import matplotlib.pyplot as plt

# Insert a dummy point at the beginning of the data
Yield_val_dummy = np.insert(Yield_val, 0, np.nan)
Yield_pred_val_dummy = np.insert(Yield_pred_val, 0, np.nan)

plt.figure(figsize=(10, 6))
plt.plot(Yield_val_dummy, 'b-', linewidth=2, label='Observed')
plt.plot(Yield_pred_val_dummy, 'r-', linewidth=2, label='Predicted')
plt.xlabel('Time (Years)')
plt.ylabel('Banana Crop Yield (t/ha)')
plt.title('Observed and Predicted Yield (Validation Set)')
plt.legend()
plt.grid(True)
plt.xlim([0, 14])
```

```
# Set x-axis ticks starting from 0 to 14 with a step of 2
plt.xticks(np.arange(0, 15, step=2))

plt.show()

# Display the final predicted yield
final_predicted_yield = Yield_pred_val[-1]
print('Final Predicted Yield:', final_predicted_yield)

# Display the final true yield
final_true_yield = Yield_val[-1]
print('Final True Yield:', final_true_yield)

# Insert a dummy point at the beginning of the data for forecast
    scenarios
Yield_val_dummy = np.insert(Yield_val, 0, np.nan)

# Perform model forecasts
n_forecast = 10
exog_forecast = data.iloc[train_samples:, 1:6].values[:n_forecast]

# Forecast using scenario 1 (using predicted Yield)
forecast_scenario1 = model_fit.get_forecast(steps=n_forecast, exog=
    exog_forecast)
Yield_forecast_scenario1 = forecast_scenario1.predicted_mean

# Forecast using scenario 2 (using true Yield)
Yield_forecast_scenario2 = [Yield_val_dummy[-1]] # Initialize with
    the last value from the validation set

# Generate forecasts recursively
for i in range(1, n_forecast):
exog_step = exog_forecast[i-1].reshape(1, -1) # Exogenous variables
    for each forecasted time step
forecast_step = model_fit.get_forecast(steps=1, exog=exog_step)
yield_step = forecast_step.predicted_mean[0] # Forecasted yield for
    the current time step
Yield_forecast_scenario2.append(yield_step)

# Print the forecasts
print('Forecasted Yield (Scenario 1):')
print(Yield_forecast_scenario1)

print('Forecasted Yield (Scenario 2):')
```

```
print(Yield_forecast_scenario2)

# Generate time vector for plotting
time_train = np.arange(1, train_samples + 1)
time_val = np.arange(train_samples, n + 1)
time_forecast = np.arange(n + 1, n + n_forecast + 1)

# Plot yield forecasts for Scenario 1
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
plt.plot(np.concatenate([Yield_pred_val_dummy,
    Yield_forecast_scenario1]), 'r', linewidth=2)
plt.plot(np.arange(len(Yield_pred_val_dummy), len(
    Yield_pred_val_dummy) + n_forecast), Yield_forecast_scenario1, 'g-
    ', linewidth=2)
plt.title('Yield Forecasts for Scenario 1')
plt.xlabel('Time (Years)')
plt.ylabel('Banana Crop Yield (t/ha)')
plt.legend(['Predicted', 'Forecasted'])
plt.grid(True)
plt.xlim([0, 25])

# Set x-axis ticks starting from 0 to 25 with a step of 5
plt.xticks(np.arange(0, 26, step=5))

# Plot yield forecasts for Scenario 2
plt.subplot(1, 2, 2)
plt.plot(np.concatenate([Yield_val_dummy, Yield_forecast_scenario2]),
     'b', linewidth=2)
plt.plot(np.arange(len(Yield_val_dummy), len(Yield_val_dummy) +
    n_forecast), Yield_forecast_scenario2, 'm-', linewidth=2)
plt.title('Yield Forecasts for Scenario 2')
plt.xlabel('Time (Years)')
plt.ylabel('Banana Crop Yield (t/ha)')
plt.legend(['Historical', 'Forecasted'])
plt.grid(True)
plt.xlim([0, 25])

# Set x-axis ticks starting from 0 to 25 with a step of 5
plt.xticks(np.arange(0, 26, step=5))

plt.tight_layout()
plt.show()
```

Appendix 3

State Space Model MATLAB Codes

```
clc
clear
close all
% Load the Banana1 dataset
data = readtable('Banana1.csv');
Yield = data{:, 7};
precipitation = data{:, 2};
soil_moisture = data{:, 3};
min_temp = data{:, 4};
max_temp = data{:, 5};
humidity = data{:, 6};
% Split the data into training and validation sets (80% for training)
train_ratio = 0.8;
n = length(Yield);
train_samples = round(train_ratio * n);
Yield_train = Yield(1:train_samples);
Yield_val = Yield(train_samples+1:end);

% Define the state vector x
x = [Yield_train(1); 0; 0; 0; 0; 0];
% Define the state transition matrix A
delta_t = 1;
A = [1, 1, 0, 0, 0, 0;
0, 1, 1, 0, 0, 0;
0, 0, 1, 1, 0, 0;
0, 0, 0, 1, 1, 0;
0, 0, 0, 0, 1, 1;
0, 0, 0, 0, 0, 1];
% Define the observation matrix C
C = [1, 0, 0, 0, 0, 0];
% Define the process noise covariance matrix Q and the measurement
    noise covariance matrix R
sigma_Yield_slope = 0.1;
sigma_precipitation = 0.1;
sigma_soil_moisture = 0.1;
sigma_min_temp = 0.1;
sigma_max_temp = 0.1;
sigma_humidity = 0.1;
Q = diag([sigma_Yield_slope^2, sigma_precipitation^2,
    sigma_soil_moisture^2, sigma_min_temp^2, sigma_max_temp^2,
```

```
    sigma_humidity^2]);
sigma_Yield = 1;
R = sigma_Yield^2;

% Initialize the state covariance matrix P to the identity matrix.
P = eye(6);
% Initialize arrays to store results
Yield_pred_train = zeros(train_samples, 1);
Yield_pred_val = zeros(n - train_samples, 1);
% Loop over training time steps
for t = 1:train_samples
% Prediction step
x_predict = A * x;
P_predict = A * P * A' + Q;
% Update step
K = P_predict * C' * inv(C * P_predict * C' + R);
x_update = x_predict + K * (Yield_train(t) - C * x_predict);
P_update = (eye(6) - K * C) * P_predict;
% Save predicted Yield
Yield_pred_train(t) = C * x_update;
% Update state and covariance matrices
x = x_update;
P = P_update;
end

% Loop over validation time steps
for t = 1:(n - train_samples)
% Prediction step
x_predict = A * x;
P_predict = A * P * A' + Q;
% Update step
K = P_predict * C' * inv(C * P_predict * C' + R);
x_update = x_predict + K * (Yield_val(t) - C * x_predict);
P_update = (eye(6) - K * C) * P_predict;
% Save predicted Yield
Yield_pred_val(t) = C * x_update;
% Save Q values for train_samples+1, train_samples+2, ..., n
if t > 1
x_diff = x_update - A * x;
Q_vals(train_samples + t - 1, :) = diag(x_diff * x_diff');
end
% Update state and covariance matrices
x = x_update;
P = P_update;
```

```
    end
    % Estimate process noise covariance matrix Q
    Q_est = mean(Q_vals, 1);

    % Print estimated parameters
    disp(['Estimated sigma_Yield_slope: ', num2str(sqrt(Q_est(1)))]);
    disp(['Estimated sigma_precipitation: ', num2str(sqrt(Q_est(2)))]);
    disp(['Estimated sigma_soil_moisture: ', num2str(sqrt(Q_est(3)))]);
    disp(['Estimated sigma_min_temp: ', num2str(sqrt(Q_est(4)))]);
    disp(['Estimated sigma_max_temp: ', num2str(sqrt(Q_est(5)))]);
    disp(['Estimated sigma_humidity: ', num2str(sqrt(Q_est(6)))]);

    % Save estimated parameters to file
    state_space_model_params = struct('A', A, 'C', C, 'Q', Q_est, 'R', R)
        ;
    save('state_space_model_params.mat', 'state_space_model_params');

    % Compute model fit measures for training set
    mse_train = mean((Yield_train - Yield_pred_train).^2); % Mean Squared
         Error
    mae_train = mean(abs(Yield_train - Yield_pred_train)); % Mean
        Absolute Error
    rmse_train = sqrt(mse_train); % Root Mean Squared Error

    % Compute R-squared for training set
    ss_total_train = sum((Yield_train - mean(Yield_train)).^2); % Total
        sum of squares
    ss_residual_train = sum((Yield_train - Yield_pred_train).^2); %
        Residual sum of squares
    r_squared_train = 1 - (ss_residual_train / ss_total_train); %
        Coefficient of Determination

    % Compute model fit measures for validation set
    mse_val = mean((Yield_val - Yield_pred_val).^2); % Mean Squared Error
    mae_val = mean(abs(Yield_val - Yield_pred_val)); % Mean Absolute
        Error
    rmse_val = sqrt(mse_val); % Root Mean Squared Error

    % Compute R-squared for validation set
    ss_total_val = sum((Yield_val - mean(Yield_val)).^2); % Total sum of
        squares
    ss_residual_val = sum((Yield_val - Yield_pred_val).^2); % Residual
        sum of squares
    r_squared_val = 1 - (ss_residual_val / ss_total_val); % Coefficient
```

```
    of Determination

% Display the model fit measures for training set
disp('Training Set:');
disp(['Mean Squared Error (MSE): ', num2str(mse_train)]);
disp(['Mean Absolute Error (MAE): ', num2str(mae_train)]);
disp(['Root Mean Squared Error (RMSE): ', num2str(rmse_train)]);
disp(['Coefficient of Determination (R-squared): ', num2str(
    r_squared_train)]);

% Display the model fit measures for validation set
disp('Validation Set:');
disp(['Mean Squared Error (MSE): ', num2str(mse_val)]);
disp(['Mean Absolute Error (MAE): ', num2str(mae_val)]);
disp(['Root Mean Squared Error (RMSE): ', num2str(rmse_val)]);
disp(['Coefficient of Determination (R-squared): ', num2str(
    r_squared_val)]);

% Plot the observed and predicted Yield for training set
figure;
plot(Yield_train, 'b-', 'LineWidth', 2);
hold on;
plot(Yield_pred_train, 'r--', 'LineWidth', 2);
hold off;
xlabel('Time (Years)');
ylabel('Banana Crop Yield (t/ha)');
title('Observed and Predicted Yield (Training Set)');
legend('Observed', 'Predicted');
grid on;
xlim([0,50]);

% Plot the observed and predicted Yield for validation set
figure;
plot(Yield_val, 'b-', 'LineWidth', 2);
hold on;
plot(Yield_pred_val, 'r--', 'LineWidth', 2);
hold off;
xlabel('Time (Years)');
ylabel('Banana Crop Yield (t/ha)');
title('Observed and Predicted Yield (Validation Set)');
legend('Observed', 'Predicted');
grid on;
xlim([0,14]);
```

```
% Perform model forecasts
x_forecast = x; % Initial state for forecasting

% Loop over forecast time steps
for t = 1:10
% Prediction step
x_forecast = A * x_forecast;

% Save forecasted Yield for scenario 1 (using predicted Yield)
Yield_forecast_scenario1(t) = C * x_forecast;

% Save forecasted Yield for scenario 2 (using true Yield)
x_true = A * [Yield(t + train_samples); 0; 0; 0; 0; 0];
Yield_forecast_scenario2(t) = C * x_true;
end

% Display the forecasted yields for scenario 1
disp('Forecasted Yields for Scenario 1:');
disp(Yield_forecast_scenario1);

% Display the forecasted yields for scenario 2
disp('Forecasted Yields for Scenario 2:');
disp(Yield_forecast_scenario2);

% Generate time vector for plotting
time_train = 1:train_samples;
time_val = (train_samples+1):n;

% Plot yield forecasts for Scenario 1
figure;
subplot(1, 2, 1);
time_pred = 1:length(Yield_pred_val);
time_forecast1 = length(Yield_pred_val) + (1:10);

% Combine the predicted and forecasted Yield data
combined_data_scenario1 = [Yield_pred_val; Yield_forecast_scenario1
    '];
combined_time_scenario1 = [time_pred, time_forecast1];

% Define colors
color_combined_scenario1 = 'r'; % Red for combined
color_forecasted = 'g'; % Green for forecasted

% Plot the combined data with red color and forecasted with green
```

```
plot(combined_time_scenario1, combined_data_scenario1,
    color_combined_scenario1, 'LineWidth', 2);
hold on;
plot(time_forecast1, Yield_forecast_scenario1, color_forecasted, '
    LineWidth', 2);
hold off;
title('Yield Forecasts for Scenario 1');
xlabel('Time (Years)');
ylabel('Banana Crop Yield (t/ha)');
legend('Predicted', 'Forecasted');
grid on;
xlim([0,25]);

% Plot yield forecasts for Scenario 2
subplot(1, 2, 2);
time_val = 1:length(Yield_val);
time_forecast2 = length(Yield_val) + (1:10);

% Combine the historical and forecasted Yield data
combined_data_scenario2 = [Yield_val; Yield_forecast_scenario2'];
combined_time_scenario2 = [time_val, time_forecast2];

% Define colors
color_combined_scenario2 = 'b'; % Blue for combined
color_forecasted = 'm'; % Magenta for forecasted

% Plot the combined data with blue color and forecasted with magenta
plot(combined_time_scenario2, combined_data_scenario2,
    color_combined_scenario2, 'LineWidth', 2);
hold on;
plot(time_forecast2, Yield_forecast_scenario2, color_forecasted, '
    LineWidth', 2);
hold off;
title('Yield Forecasts for Scenario 2');
xlabel('Time (Years)');
ylabel('Banana Crop Yield (t/ha)');
legend('Historical', 'Forecasted');
grid on;
xlim([0,25]);
```

Appendix 4

LSTM Model MATLAB Codes

```
clc
clear
close all
% Load the data from Banana1.csv file
data = readmatrix('Banana1.csv');

% Separate the data into predictor and response variables
X = data(:, 2:end-2); % predictor variables (excluding year and yield
    )
X(:, 5) = data(:, end-1); % add relative humidity as the fifth column
     of X
y = data(:, end); % response variable (banana crop yield)

% Split the data into training and testing sets
numObs = length(y);
numTrain = round(0.8*numObs); % 80% for training
XTrain = X(1:numTrain, :);
yTrain = y(1:numTrain);
XTest = X(numTrain+1:end, :);
yTest = y(numTrain+1:end);

% Data scaling
XTrain_scaled = normalize(XTrain);
yTrain_scaled = normalize(yTrain);
XTest_scaled = normalize(XTest);
yTest_scaled = normalize(yTest);

% Convert the training data to cell arrays
X_train_cell = num2cell(XTrain_scaled', 1);
Y_train_cell = num2cell(yTrain_scaled');

% Define the LSTM layers
num_hidden_units = 50;
layers = [
sequenceInputLayer(size(XTrain_scaled, 2), 'Name', 'input')
lstmLayer(num_hidden_units, 'Name', 'lstm')
fullyConnectedLayer(1, 'Name', 'fc')
regressionLayer('Name', 'output')
];
```

```
% Define the training options
options = trainingOptions('adam', ...
'MaxEpochs', 100, ...
'MiniBatchSize', 32, ...
'SequenceLength', 'longest', ...
'Shuffle', 'never', ...
'GradientThreshold', 1, ...
'Verbose', true);

% Train the LSTM network
net = trainNetwork(X_train_cell, Y_train_cell, layers, options);

% Convert the testing data to cell arrays
X_test_cell = num2cell(XTest_scaled', 1);
Y_test_cell = num2cell(yTest_scaled');

% Perform model fit evaluation on the training set
Yield_pred_train = predict(net, X_train_cell);
Yield_pred_train = cell2mat(Yield_pred_train')';
mse_train = mean((yTrain_scaled - Yield_pred_train).^2, 'all');
mae_train = mean(abs(yTrain_scaled - Yield_pred_train), 'all');
rmse_train = sqrt(mse_train);
ss_residual_train = sum((yTrain_scaled - Yield_pred_train).^2, 'all')
    ;
ss_total_train = sum((yTrain_scaled - mean(yTrain_scaled)).^2, 'all')
    ;
r_squared_train = 1 - (ss_residual_train / ss_total_train);

% Perform model fit evaluation on the validation set
Yield_pred_val = predict(net, X_test_cell);
Yield_pred_val = cell2mat(Yield_pred_val')';
mse_val = mean((yTest_scaled - Yield_pred_val).^2, 'all');
mae_val = mean(abs(yTest_scaled - Yield_pred_val), 'all');
rmse_val = sqrt(mse_val);
ss_residual_val = sum((yTest_scaled - Yield_pred_val).^2, 'all');
ss_total_val = sum((yTest_scaled - mean(yTest_scaled)).^2, 'all');
r_squared_val = 1 - (ss_residual_val / ss_total_val);

% Display the model fit measures for the training set
disp('Training Set:');
disp(['Mean Squared Error (MSE): ', num2str(mse_train)]);
disp(['Mean Absolute Error (MAE): ', num2str(mae_train)]);
disp(['Root Mean Squared Error (RMSE): ', num2str(rmse_train)]);
disp(['Coefficient of Determination (R-squared): ', num2str(
```

```
    r_squared_train)]);
% Display the model fit measures for the validation set
disp('Validation Set:');
disp(['Mean Squared Error (MSE): ', num2str(mse_val)]);
disp(['Mean Absolute Error (MAE): ', num2str(mae_val)]);
disp(['Root Mean Squared Error (RMSE): ', num2str(rmse_val)]);
disp(['Coefficient of Determination (R-squared): ', num2str(
    r_squared_val)]);

% Plot observed vs predicted crop yields for the training set
figure;
plot(yTrain_scaled, 'b', 'LineWidth', 2);
hold on;
plot(Yield_pred_train, 'r-', 'LineWidth', 2);
hold off;
title('Observed vs Predicted Crop Yields (Training Set)');
xlabel('Time (Years)');
ylabel('Normalized Banana Crop Yield (t/ha)');
legend('Observed', 'Predicted');
grid on;
xlim([0,50]);
% Plot observed vs predicted crop yields for the validation set
figure;
plot(yTest_scaled, 'b', 'LineWidth', 2);
hold on;
plot(Yield_pred_val, 'r-', 'LineWidth', 2);
hold off;
title('Observed vs Predicted Crop Yields (Validation Set)');
xlabel('Time (Years)');
ylabel('Normalized Banana Crop Yield (t/ha)');
legend('Observed', 'Predicted');
grid on;
xlim([0,14]);

% Display the final predicted value from the validation set
disp('Final Predicted Value from Validation Set:');
disp(Yield_pred_val(end));
% Display the final true value from the validation set
disp('Final True Value from Validation Set:');
disp(yTest(end));

% Perform model forecasts
last_value_scenario1 = Yield_pred_val(end); % Final predicted value
    from the validation set (Scenario 1)
```

```
last_value_scenario2 = yTest(end); % FTV from the vs (Scenario 2)

% Prepare the input for the forecast steps
X_forecast_scenario1 = [X_test_cell{end}(2:end, :);
    last_value_scenario1];
X_forecast_scenario2 = [X_test_cell{end}(2:end, :);
    last_value_scenario2];

% Initialize arrays to store the forecasted yields
Yield_forecast_scenario1 = zeros(10, 1);
Yield_forecast_scenario2 = zeros(10, 1);

% Loop over forecast time steps
for t = 1:10
% Perform model forecast for scenario 1
forecasted_value_scenario1 = predict(net, X_forecast_scenario1);
Yield_forecast_scenario1(t) = forecasted_value_scenario1;
X_forecast_scenario1 = [X_forecast_scenario1(2:end, :);
    forecasted_value_scenario1];

% Perform model forecast for scenario 2
forecasted_value_scenario2 = predict(net, X_forecast_scenario2);
Yield_forecast_scenario2(t) = forecasted_value_scenario2;
X_forecast_scenario2 = [X_forecast_scenario2(2:end, :);
    forecasted_value_scenario2];
end

% Display the forecasted yields for scenario 1
disp('Forecasted Yields for Scenario 1:');
disp(Yield_forecast_scenario1);
% Display the forecasted yields for scenario 2
disp('Forecasted Yields for Scenario 2:');
disp(Yield_forecast_scenario2);

% Plot yield forecasts for Scenario 1
figure;
subplot(1, 2, 1);
plot([Yield_pred_val; Yield_forecast_scenario1], 'r', 'LineWidth', 2)
    ;
hold on;
plot(length(Yield_pred_val)+1:length(Yield_pred_val)+10,
    Yield_forecast_scenario1, 'g-', 'LineWidth', 2);
hold off;
title('Yield Forecasts for Scenario 1');
```

```
    xlabel('Time (Years)');
    ylabel('Normalized Banana Crop Yield (t/ha)');
    legend('Predicted', 'Forecasted');
    grid on;
    xlim([0,25]);

    % Plot yield forecasts for Scenario 2
    subplot(1, 2, 2);
    plot([yTest; Yield_forecast_scenario2], 'b', 'LineWidth', 2);
    hold on;
    plot(length(yTest)+1:length(yTest)+10, Yield_forecast_scenario2, 'm-'
        , 'LineWidth', 2);
    hold off;
    title('Yield Forecasts for Scenario 2');
    xlabel('Time (Years)');
    ylabel('Normalized Banana Crop Yield (t/ha)');
    legend('Historical', 'Forecasted');
    grid on;
    xlim([0,25]);
```

RESEARCH OUTPUTS

(i) Published paper

Patrick, S., Silas, M., Mbalawata, I., & Leo, J. (2023). Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change. *Resources, Environment and Sustainability*. Homepage: https://www.sciencedirect.com/science/article/pii/S2666916123000312?via%3Dihub

(ii) Poster presentation

PUBLISHED PAPER

Resources, Environment and Sustainability 14 (2023) 100138

Contents lists available at ScienceDirect

Resources, Environment and Sustainability

journal homepage: www.elsevier.com/locate/resenv

Research article

Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change

Sabas Patrick [a,*], Silas Mirau [a], Isambi Mbalawata [b], Judith Leo [a]

[a] Department of Applied Mathematics and Computational Sciences, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania
[b] African Institute for Mathematical Sciences, Kigali, Rwanda

ARTICLE INFO

Keywords:
Time series
Ensemble
Modeling
Forecasting
Banana crop yield
Climate change

ABSTRACT

Banana cultivation plays a pivotal role in Tanzania's agricultural landscape and food security. Precisely forecasting banana crop yield is essential for resource optimization, market stability, and informed policymaking, particularly in the face of climate change. This study employed time series and ensemble models to forecast banana crop yield in Tanzania, offering crucial insights into future production trends. We utilized Seasonal ARIMA with Exogenous Variables (SARIMAX), State Space (SS), and Long Short-Term Memory (LSTM) models, chosen based on regression analysis and data exploration. Leveraging historical banana yield data (1961–2020) and relevant climate variables, we formulated an ensemble model using a weighted average approach. Our findings underscore the potential of time series and ensemble models for accurate banana crop yield forecasting. Statistical evaluation metrics validate their effectiveness in capturing temporal variations and delivering reliable predictions. This research advances agricultural forecasting by demonstrating the successful application of these models in Tanzania. It emphasizes the importance of considering temporal dynamics and relevant factors for precise predictions. Policymakers, farmers, and stakeholders can leverage this study's outcomes to make informed decisions on resource allocation, market planning, and agricultural policies. Ultimately, our research bolsters sustainable banana production and enhances food security in Tanzania.

1. Introduction

One of the largest herbaceous flowering trees is the banana (*Musa spp.*) plant (Ighalo and Adeniyi, 2019; Lal et al., 2017). Although the unripe fruit, leaves, inflorescence, stem, and rhizome of the banana plant are also utilized in many ways as vegetables, food, and animal feeds, the ripe banana is a soft fruit with a lifespan of 5 to 10 days that is suitable for use and consumption (Jayasinghe et al., 2022; Lai and Dzombak, 2020). Bananas rank among the top 10 crops in the world in terms of yield, area cultivated, and calories produced (Varma and Bebber, 2019). After maize, rice, and wheat, the fourth most important crop for providing food and money to more than 30% of the world's population is the banana crop (Lucas and Jomanga, 2021). Tanzania produces the second-largest amount of bananas in East Africa, behind Uganda (Lucas and Jomanga, 2021). Banana cultivation plays a vital role in Tanzania's agricultural sector, contributing significantly to both food security and economic growth (Lucas and Jomanga, 2021; Varma and Bebber, 2019). The banana has excellent medicinal and traditional advantages for human health and is useful in all sections of the body. The fruit of the banana is a great nutritional supplement, while the leaf is eaten in different parts of India in various ways as a vegetable (Lal et al., 2017).

The biggest worldwide problem of the century is thought to be climate change (Hoque and Haque, 2016). While there are numerous benefits of banana processing for science and technology (Lal et al., 2017). It is surprising to observe that despite their critical importance for subsistence and trade, bananas receive insufficient consideration in worldwide evaluations of how climate change can effect nutritional and food security (Varma and Bebber, 2019). The climate change has a variety of effects on crop production, the productivity and sustainability of banana crops are increasingly challenged by the effects of climate change (Chowhan et al., 2016). The region faces substantial risks to crop yield and overall agricultural productivity due to the effects of rising temperatures, changing rainfall patterns, and a higher frequency of extreme weather events (Hoque and Haque, 2016). Tanzania is one of the nations in the world now dealing with the severe effects of climate change (Omambia and Gu, 2010; Shirima and Lubawa, 2017; Mayaya, 2015). Tanzania's farm owners face a number of difficulties similar to other emerging nations throughout the world that hinder the expansion and development of the agricultural industry (Lokupitiya, 2018). To ensure the resilience and adaptability of banana cultivation to changing climatic conditions, accurate and reliable forecasting models are essential (Varma and Bebber, 2019).

* Corresponding author.
E-mail address: patricks@nm-aist.ac.tz (S. Patrick).

https://doi.org/10.1016/j.resenv.2023.100138
Received 8 August 2023; Received in revised form 8 October 2023; Accepted 8 October 2023
Available online 10 October 2023

Tanzania has seen limited research on the impacts of climate change, especially in the area of transdisciplinary studies (Kahimba et al., 2015; Abdoussalami et al., 2023). Consequently, it is challenging to fully gauge the potential impacts on food security and the productivity of the banana crop in the region (Lucas and Jomanga, 2021). Furthermore, Tanzania, as a country heavily reliant on agriculture, especially the banana sector, faces unique challenges and vulnerabilities due to its socio-economic conditions and geographical location (Omambia and Gu, 2010; Shirima and Lubawa, 2017; Mayaya, 2015). These factors make Tanzania an interesting and pertinent case study to explore the potential impacts of climate change on both food security and the productivity of a significant crop like bananas. As a staple food for millions of Tanzanians and a significant export commodity, the success and resilience of the banana crop directly influence the well-being of both rural communities and the national economy (Lucas and Jomanga, 2021). Thus, this research addresses a critical knowledge gap in the field of banana crop yield forecasting in Tanzania, considering the specific challenges posed by climate change. By identifying the potential impacts of climate variables on banana crop yield, we can provide valuable insights into the vulnerabilities and adaptive capacities of the sector (Wood et al., 2014). Forecasting banana crop yield is crucial for effective agricultural planning, resource allocation, and policy-making. By providing valuable insights, this research empowers farmers, policymakers, and stakeholders to make informed decisions and adopt suitable strategies in order to counteract the detrimental consequences of climate change (Varma and Bebber, 2019).

In recent years, time series analysis and ensemble modeling have emerged as powerful tools for forecasting agricultural crop yields (Kamir et al., 2020). Time series analysis leverages historical data to identify patterns, trends, and seasonality in crop yield, enabling the development of predictive models (Box et al., 2015). Contrarily, ensemble modeling utilizes the strengths of various forecasting models to increase accuracy and robustness (Bertsimas and Boussioux, 2023). This study aimed to utilize time series and ensemble models to forecast banana crop yield in Tanzania, specifically focusing on the effects of climate change. By incorporating historical banana crop yield data and relevant climate variables, we seek to develop forecasting models that capture the dynamics of banana productivity under changing climatic conditions (Pham et al., 2019). Conventional forecasting methods often struggle to capture the intricate interactions between climatic variables and crop yield, highlighting the need to employ sophisticated analytical techniques (Varma and Bebber, 2019; Bertsimas and Boussioux, 2023). By combining time series analysis and ensemble modeling, we can increase the forecasts' precision and dependability, resulting in better decision-making in the agriculture industry (Kourentzes et al., 2014). Moreover, the combination of time series and ensemble modeling techniques offers promising opportunities for accurate and robust banana crop yield forecasting under the influence of climate change (Bertsimas and Boussioux, 2023).

2. Materials and methods

2.1. Data description

In our analysis, we transformed the monthly climate variables, obtained from various sources, into yearly data for each year. This conversion allowed us to work with annual averages and facilitate our comprehensive assessment of the impact of these variables on banana crop yield. The Climatic Research Unit (CRU) at the University of East Anglia provided the monthly gridded data for precipitation, minimum temperature, and maximum temperature for the reanalysis, these datasets were freely downloaded from the following website: https://data.ceda.ac.uk/badc/cru/data/cru_ts/cru_ts_4.05. The CRU dataset version 4.05 (CRU TS 4.05) for a period of 1961–2020, these data cover the land surface at 0.5° × 0.5° resolution. Numerous published papers have utilized this dataset to examine precipitation variability in East Africa, comparing it with the GPCC monthly precipitation dataset provided by the World Climate Research Program-WCRP (Ongoma et al., 2019). The research findings consistently demonstrated that the CRU dataset proved to be more effective and reliable in the analysis. Furthermore, previous researchers successfully used CRU dataset rainfall in Tanzania (Mbigi and Xiao, 2021). The soil moisture and relative humidity data were acquired from the NCEP/NCAR Reanalysis dataset, which was downloaded from the following website: https://psl.noaa.gov/data/gridded/reanalysis/. The relative humidity dataset has a precision of 2.5° × 2.5° while the soil moisture dataset has a resolution of 0.25° × 0.25° (Anwar et al., 2019). The FAOSTAT database, which can be accessed at https://www.fao.org/faostat/en/#data/QCL, provided the study's average annual banana crop yield statistics (see Table 1).

Table 1
Dataset variables used in this study.

N	Variable	Unit of measurement
1.	Precipitation	mm
2.	Minimum temperature	°C
3.	Maximum temperature	°C
4.	Relative humidity	%
5.	Soil moisture	Fraction
6.	Banana crop yield	(t/ha)

2.2. Methodology

This study delves into the intricate relationship between climate change and Tanzanian banana crop yield. It aims to understand how shifting climate patterns impact this essential agricultural output. In order to obtain the addressed objective of this study, the researchers takes a two-fold approach. The first approach is **Correlation Analysis**; the study investigates how key climate variables, including precipitation, soil moisture, temperature extremes, and relative humidity, relate to banana crop yield. A robust multiple regression model uncovers valuable insights within this connection, indeed the multiple regression model used to identify the significance of key climate variables at hand. However, the study acknowledges that not all pertinent climate variables were included. However, we believe that these key climate variables are reasonable factors for this study.

The second approach is **Forecasting Models**; to predict future banana yields amid changing climates, the study employs time series models like SARIMAX, SS, and LSTM. These models capture temporal nuances and yield trends. The choice of these approaches was based on the regression analysis, and data exploration results (Jayasinghe et al., 2022; Hyndman and Athanasopoulos, 2018; Box et al., 2015). Thereafter, we formulated the ensemble model using a weighted average approach. An ensemble model combines historical yield data and relevant climate variables to enhance prediction accuracy. Specifically, the use of weighted linear combinations of various ensemble members has gained popularity because of its ease of implementation in real-world applications (Bertsimas and Boussioux, 2023).

Generally, this paper aims to provide valuable insights into the climate–yield interaction, considering the second dimension (i.e forecasting models) results and discussion. While not overlooking correlation analysis approach (i.e multiple regression model). The schematic diagram in Fig. 1 indicates the flow of the whole work:

2.2.1. Multiple regression model

In this work, the regression model shows a relationship between the yield of the banana crop, denoted by the response variable Y, and five explanatory variables: precipitation (X_1), soil moisture (X_2), minimum temperature (X_3), maximum temperature (X_4), and relative humidity (X_5) (Bhausaheb et al., 2023; Anzures et al., 2022). The population regression equation, in particular, depicts the actual

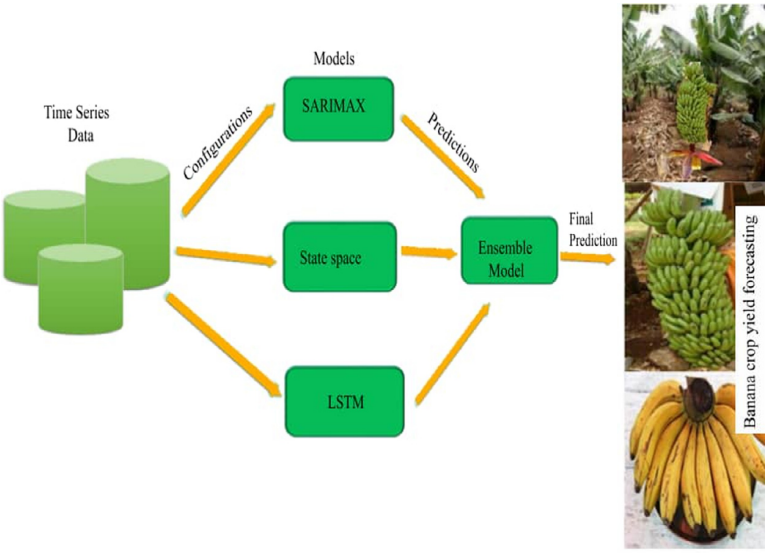


Fig. 1. The schematic diagram, a representation of methodology.

connection between the explanatory variables and the response variable (Ngo and La Puente, 2012). However, as the population regression equation remains unknown, we need to estimate it based on sampled data (Sagamiko et al., 2020; Hanson, 2010).

Let us consider a sample of n observations, each containing values for both the response variable Y and p explanatory variables X_i. We can represent the values for the ith observation as Y_i, X_{i1}, X_{i2}, …, X_{ip} (Sagamiko et al., 2020). Thus, the multiple regression equation for these values is given by: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$, where Y_i represents the value of the response variable for the ith observation, and (X_{i1}, X_{i2}, …, X_{ip}) represents the values of the explanatory variables for the ith observation. The coefficients of the regression model are denoted by $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, and the term ε_i represents the error term for the ith observation (Sagamiko et al., 2020).

If we have more data points (n) than explanatory variables (p), forming an overdetermined system with linearly dependent equations, we can represent the ith observation of variable X_j as X_{ij}, where $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$. In this case, the population model for all observations of the sample can be expressed as the following system of equations (Sagamiko et al., 2020; Hanson, 2010):

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1 \\ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \varepsilon_2 \\ \quad \vdots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \varepsilon_n \end{cases} \tag{1}$$

The system of Eqs. (1) can be represented in matrix notation as follows (Sagamiko et al., 2020; Hanson, 2010):

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2}$$

The primary goal of regression analysis is to select the explanatory variables that have a significant impact on the yield (Rathod and Mishra, 2018). In light of the assumption that the response and explanatory variables have a linear connection, we can express the equation mathematically as Sagamiko et al. (2020), Adejuwon and Agundiminegha (2019) and Salvacion (2020):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \tag{3}$$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, and β_5 are the coefficients or parameters associated with each explanatory variable, and ε represents the error term or residual, capturing the variability in the crop yield that is not explained by the model.

The most common way to estimate the population regression equation is to use least squares (Ngo and La Puente, 2012). A technique called least squares seeks to reduce the squared disparities between the response variable's observed values and those predicted by the regression model (Hanson, 2010).

The least squares estimator of the population regression equation is given by the following equation:

$$\beta = (X^T X)^{-1} X^T Y \tag{4}$$

where β is the estimated coefficients of the regression equation, X is the matrix of explanatory variables, and Y represents the vector of observed values of the response variable.

To prove, we rewrite the multiple regression equation in matrix notation. Using the matrices defined earlier in Eq. (3), we have:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{5}$$

where $\mathbf{Y}$ stand for the column vector of response variable values, $\mathbf{X}$ represents the design matrix, β denotes the column vector of coefficients, and ε is the column vector of error terms.

To estimate the coefficients β, using the least squares method. The estimator is given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{6}$$

Substituting the estimated coefficients $\hat{\beta}$ into the multiple regression equation, we get:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \tag{7}$$

Here, $\hat{\mathbf{Y}}$ represents the predicted values of the response variable based on the estimated coefficients. Therefore, the estimated population regression equation using least squares is $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$. To obtain the equation $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, we substitute the estimated coefficients $\hat{\beta}$ into the equation $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. This equation gives the estimated population regression coefficients β based on the least squares method. Please note that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ exists if the design matrix $\mathbf{X}^T\mathbf{X}$ is invertible.

2.2.2. Seasonal ARIMA (SARIMA) with exogenous variables

ARIMA, one of the most popular and effective time-series models, is one of the classics (Rathod and Mishra, 2018). The ARIMA model has gained considerable popularity because of its linear statistical characteristics and the commonly used Box–Jenkins approach for model

creation created by Box and Jenkins in the 1970 (Box et al., 2015). The ARIMA model's standard form is then written as $ARIMA(p,d,q)$ where the letters p stand for the auto-regressive term order, d for the differencing term order, and q for the moving average term order (Arunraj et al., 2016; Hyndman and Athanasopoulos, 2018). Mathematically, the $ARIMA(p,d,q)$ model can be expressed as Arunraj et al. (2016):

$$\phi_p(B)(1-B)^d X_t = \mu + \theta_q(B)\varepsilon_t \tag{8}$$

where $\phi_p(B)$ stand for the autoregressive (AR) operator of order p, $(1-B)^d$ stand for the differencing operator, where d represents the order of differencing, X_t stand for the time-series variable at time t, which is the variable being modeled or predicted, μ is a constant term in the equation, accounts for any deterministic component or offset in the time series, $\theta_q(B)$ stand for (MA) the moving average operator of order q, and ε_t is the error term at time t, which denotes the random or unexplained component of the time-series.

The ARIMA model can be expanded as SARIMA$(p,d,q)(P,D,Q)_s$ to accommodate seasonal variations, where s is a term that considers the length of the seasonal period (Neog et al., 2022; Meeradevi et al., 2022; Raj et al., 2019). The SARIMA model can be represented as Arunraj et al. (2016):

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D X_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \tag{9}$$

where $\phi_p(B)$ stand for (AR) the seasonal autoregressive operator of order p, $\theta_q(B)$ stand for (MA) the seasonal moving average operator of order q, $(1-B)^d$ represents the differencing operator applied d times, $(1-B^S)^D$ denotes the seasonal differencing operator applied D times, and S stand for the seasonal length (say, $s=4$ in quarterly data, and $s=12$ in monthly data).

Given the SARIMAX$(p,d,q)(P,D,Q)_s$ model, where (X) is the vector of external variables, the multi linear regression techniques are used to model the external variables (Arunraj et al., 2016). In this study, we can express a multiple regression model mathematically as:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + w_t \tag{10}$$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, and β_5 are the coefficients or parameters associated with each explanatory variable, and w_t represents the error term or residual, capturing the variability in the crop yield that is not explained by the model. The error term w_t can be expressed in the form of SARIMA model as Arunraj et al. (2016):

$$w_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{11}$$

By inserting Eq. (11) into Eq. (10), we derive the subsequent equation:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{12}$$

2.2.3. State space (SS) model

The state space approach is a mathematical framework used for modeling time series data (Aoki, 2013). It models the underlying process that generates the observed data as a set of unobserved states that evolve over time according to a set of stochastic equations (Verma, 2018). The observed data is then generated from these unobserved states through a set of observation equations (Suman and Verma, 2017).

State equation:

$$\boldsymbol{x}_t = \boldsymbol{F}_t \boldsymbol{x}_{t-1} + \boldsymbol{G}_t \boldsymbol{w}_t \tag{13}$$

where $\boldsymbol{x}_t$ is the $n\times 1$ vector of unobserved states at time t, $\boldsymbol{F}_t$ is the $n\times n$ state transition matrix, $\boldsymbol{G}_t$ is the $n\times m$ matrix of state noise, and $\boldsymbol{w}_t$ is the $m\times 1$ vector of state noise at time t.

Observation equation:

$$\boldsymbol{y}_t = \boldsymbol{H}_t \boldsymbol{x}_t + \boldsymbol{v}_t \tag{14}$$

where $\boldsymbol{y}_t$ is the $p\times 1$ vector of observed data at time t, $\boldsymbol{H}_t$ is the $p\times n$ observation matrix, and $\boldsymbol{v}_t$ is the $p\times 1$ vector of observation noise at time t.

The state space model presupposes that the noise in the state and the noise in the observations are independent, both of which have known covariance matrices and a normal distribution with a mean of zero (Verma, 2018):

$$\boldsymbol{w}_t \sim N(\boldsymbol{0}, \boldsymbol{Q}_t) \quad \text{and} \quad \boldsymbol{v}_t \sim N(\boldsymbol{0}, \boldsymbol{R}_t) \tag{15}$$

where $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$ are the $m\times m$ and $p\times p$ covariance matrices of the state noise and observation noise, respectively.

A variety of time series models, including ARMA models, ARIMA models, and state space models with non-linear and non-Gaussian state transitions and observation equations, can be created using the state space technique (Hu et al., 2019; Verma, 2018; Hooda et al., 2020). State Space models can be very useful in modeling time series data affected by multiple external factors such as climate change (Cook, 1985; Marolla et al., 2021). They can capture the effects of multiple external factors on the time series by modeling the external factors as additional states in the model. This is done by including additional equations that describe the dynamics of the external factors (Marolla et al., 2021).

The state space model is typically estimated using maximum likelihood estimation or Bayesian methods (Newman et al., 2023). Given a state space model with observations y_t and state vectors x_t. We can express the likelihood function as follows:

$$L(\theta|y) = f(y_1|\theta)f(x_1|\theta)\prod_{t=2}^{T} f(y_t|x_t,\theta)f(x_t|x_{t-1},\theta) \tag{16}$$

where θ denotes the parameters of the state space model, and $f(y_t|x_t,\theta)$ and $f(x_t|x_{t-1},\theta)$ are the conditional densities of the observations and state vectors, respectively.

The MLE method involves finding the set of parameters $\hat{\theta}$ that maximizes the likelihood function:

$$\hat{\theta} = \text{argmax}_\theta L(\theta|y) \tag{17}$$

Also, we can express the posterior distribution of the parameters as follows:

$$p(\theta|y) \propto L(\theta|y)p(\theta) \tag{18}$$

where $L(\theta|y)$ is the likelihood function as defined above, and $p(\theta)$ represents the prior distribution of the parameters.

The Kalman filter algorithm, which is a recursive Bayesian estimation method is used in parameter estimation (de Bézenac et al., 2020). The Kalman filter combines prior knowledge about the system dynamics with the observed data to estimate the parameters. It optimally incorporates the available information and updates the parameter estimates as new data becomes available (de Bézenac et al., 2020; Suman and Verma, 2017).

Hence, the forecasts are generated by projecting the latent state variables into the future and using the observation equation to obtain the predicted values, say banana crop yield:

$$\hat{y}T+1|T = \mathbb{E}[yT+1|y_{1:T},\theta] = \mathbb{E}[f(s_{T+1})|\hat{s}_{T+1|T},\theta] \tag{19}$$

where $\hat{s}T+1|T$ is the predicted state estimate for time $T+1$ given the observed data $y1:T$, and $f(s_{T+1})$ is the observation equation relating the latent state variables to the observed yield.

2.2.4. Long short-term memory (LSTM) model

Long Short-Term Memory (LSTM), often known as a type of recurrent neural networks (RNNs), is a specialized architecture created to manage sequential data, particularly time series data (Tian et al., 2021; Meeradevi et al., 2022). Traditional RNNs struggle with the vanishing gradient problem, which is especially addressed by LSTMs. LSTMs are able to better describe long-term dependencies in sequential data by

efficiently storing and retrieving information over extended periods of time (Reddy et al., 2022). The LSTM model essentially offers a practical method for working with sequential data and has found use in a variety of fields, including climate forecasting (Bhimavarapu et al., 2023; Tian et al., 2021; Meeradevi et al., 2022).

An LSTM's architecture consists of a memory cell that can retain information for extended periods, along with three gates (input, output, and forget) (Tian et al., 2021). The input gate controls how much fresh information is introduced to the memory cell, the output gate controls how much information is taken out of the memory cell, and the forget gate controls how much old or unnecessary information is removed from the memory cell. Within the LSTM paradigm, these gates regulate the flow of data into and out of the memory cell, enabling efficient information retention and usage (Liu et al., 2023; Bhimavarapu et al., 2023).

LSTM model configuration includes the following equations:

$$\textbf{Input layer}: y_t = g(W_i * x_t + b_i) \tag{20}$$

$$\textbf{LSTM layer}: h_t = LSTM(h_{t-1}, y_{t-1}) \tag{21}$$

$$\textbf{Output layer}: y_{t+1} = g(W_o * h_t + b_o) \tag{22}$$

In the input layer, the output y_t is obtained by applying an activation function g to the dot product of the weight matrix W_i and the input vector x_t, followed by the addition of a bias term b_i. This y_t represents the output of the input layer at time t.

The LSTM layer's output h_t at time t is determined by passing the previous hidden state h_{t-1} and the previous input y_{t-1} to the LSTM cell. The LSTM cell updates its internal state based on these inputs, generating a new hidden state h_t.

For the next time step, the predicted output y_{t+1} is calculated by applying the activation function g to the dot product of the weight matrix W_o and the hidden state h_t, then adding a bias term b_o.

The backpropagation through time (BPTT) method is used to update the LSTM neuron weights before training the model with the training data. This involves computing the gradients of the loss function with respect to the weights using the chain rule (Sadowski, 2016). The LSTM model learns to improve its performance on the training data by iteratively modifying the weights based on the estimated gradients, increasing its capacity for precise prediction (Bhimavarapu et al., 2023).

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial W} \tag{23}$$

In the equation, L represents the loss function, W denotes the weight, y is the output, h corresponds to the hidden state, and t indicates the time step. These variables play essential roles in the process of training the LSTM model and optimizing its performance on the training data.

Furthermore, the first and second moments of the gradient are taken into account using an appropriate optimization technique, such as Adam. This allows the algorithm to adapt the learning rate independently for each weight, enhancing the training process of the LSTM model (Bhimavarapu et al., 2023).

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot \frac{\partial L}{\partial W} \tag{24}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot \left(\frac{\partial L}{\partial W}\right)^2 \tag{25}$$

$$W = W - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{26}$$

where m_t and v_t are the first and second moment estimates of the gradient, respectively, and W is the weight being updated, α is the learning rate, β_1 and β_2 represents hyperparameters that control the decay rates of the moment estimates, and ϵ stand for a small constant to prevent division by zero.

Finally, the LSTM model is optimized and if its performance met the desired level of accuracy, it deployed for use in predictions (Bhimavarapu et al., 2023), for example predicting banana crop yield under different climate scenarios.

2.2.5. Ensemble modeling approach

By mixing the results of various models, ensemble modeling is a flexible strategy that aims to increase prediction accuracy and reliability (Bertsimas and Boussioux, 2023). Although ensemble models can be used with a variety of data sources, including time series data, their main goal is to improve overall model performance rather than focusing especially on the special properties of time series data (Hao et al., 2020).

Ensemble modeling can be effectively integrated with time series modeling to enhance the accuracy of time series forecasts (Bertsimas and Boussioux, 2023). As an illustration, a common approach in ensemble modeling involves constructing a diverse ensemble of time series models, which may include ARIMA, exponential smoothing, and neural network models. Subsequently, the predictions from these individual models are combined using techniques like weighted averaging or other methods (Kourentzes et al., 2014; Bayati et al., 2020; Kamir et al., 2020). This allows for the utilization of the unique strengths of each individual model while compensating for their respective weaknesses, leading to more precise and reliable overall forecasts (Moore and Lobell, 2014).

For instance, the ensemble model involves aggregating the predictions of individual models to derive a final prediction using a weighted average approach. The mathematical representation of the weighted average approach can be expressed as follows:

$$y = w_1 \times y_1 + w_2 \times y_2 + w_3 \times y_3 + \cdots + w_n \times y_n \tag{27}$$

where y is the final predicted value, $y_1, y_2, \ldots, y_n$ are the predicted values of the individual models, respectively, and $w_1, w_2, \ldots, w_n$ are the weights assigned to the individual models based on their performance on the training, or validation set.

Furthermore, the weights of the individual models are determined based on their performance on the testing set. Based on the inverse of each model's error or loss, the weights are assigned. To ensure that the weights add up to 1, they are normalized, and the normalized weights are then used in the ensemble model to combine the predictions of the individual models (Van Leeuwen et al., 2023). For instance, in this paper, the weights assigned to each model were derived from their R-squared values, which indicate the proportion of variance in the observed banana crop yield that is explained by each model.

The process of converting R-squared values to normalized weights involved the following steps. Determining weights, the ratio of 1 to each R-squared value is used. Normalization, we divide each weight by the sum of all weights obtained across the models used to ensure that the weights are comparable and would sum up to 1. Assigning weights, the normalized R-squared values are then used as weights to determine the contribution of each model to the final forecast. Final forecast, the ensemble forecast is generated by taking the weighted average of the predictions from the individual models. This approach allow us to leverage the strengths of each model and mitigate potential weaknesses.

3. Results and discussion

3.1. Data exploration results

The analysis relies on the yearly reanalysis datasets of precipitation, soil moisture, minimum temperature, maximum temperature, and relative humidity. These datasets were utilized for modeling and forecasting the banana crop yield. All necessary steps required for data

Table 2
Statistical evaluation metrics.

Model	Training set				Validation set			
	MSE	MAE	RMSE	R-Squared	MSE	MAE	RMSE	R-Squared
SARIMAX	0.3828	0.3650	0.6187	0.8109	4.3797	1.4789	2.0928	0.1825
State space	0.0105	0.0423	0.1026	0.9948	0.0885	0.2068	0.2974	0.9835
LSTM	0.6200	0.4192	0.7874	0.6991	0.5288	0.6890	0.7272	0.9013

pre-processing, and filtering were considered, including detrending (non-stationarity, and seasonality), and autocorrelation. In this study, the collected climate variables were believed to impact banana crop yield under a robust multiple regression analysis.

The MATLAB, and PYTHON tools were used interchangeably throughout the analysis. The all selected methods performed by using climate time-series data to predict the production of banana yield in Tanzania for a period of time from 1961 to 2020. The first 80% of the datasets were used to train the models, while the final 20% were used to test and assess how well they worked. The normalize function was used to normalize the training and testing sets as necessary to make sure that all variables are on a similar scale. This normalization process helps in avoiding potential issues caused by differing magnitudes among the variables. The training and testing data were transformed into cell arrays to facilitate their processing and handling within the models. Converting the data to cell arrays allows for more flexible and efficient data manipulation during the model building and evaluation processes. Various statistical metrics were found in each model as shown in Table 2, which signify the performance for the selection of the best model that fit the data. This table showcases the performance metrics and evaluation results obtained from the models.

3.2. Regression analysis and results

Our research supports the assumption that there is a linear relationship between the explanatory variables and the response. The regression coefficients shown in Table 3 show how key climate variables affect the rate of change in banana crop production when each explanatory variable changes by one unit while all other explanatory variables remain constant. By plugging the values of the regression coefficients from Table 3 into the regression equation, we may obtain the following expression:

$$Y = -22.8320 + 0.0206X_1 - 0.0085X_2 + 4.8328X_3 - 1.6594X_4 - 0.0991X_5 \quad (28)$$

The constant term (−22.8320) is the predicted value of Y when none of the independent variables have an effect, and the negative sign indicates the gradual decrease in banana crop yield. Based on the p-values as presented in Table 3, only minimum temperature has a significant positive impact on the yield, while the other external variables (precipitation, soil moisture, maximum temperature, and relative humidity), and the intercept does not significantly impact the banana crop yield. However, we applied the stepwise regression technique, and all the explanatory variables were selected to be significant.

In general, the regression model's R-squared value of 0.502 shows that the chosen explanatory variables can account for about 50.2% of the variation in banana crop yield. The F-statistic of 10.87 is statistically significant (Prob (F-statistic): 2.89e−07), indicating that the model as a whole is significant. On the other hand, the condition number is large, 4.43e−04. This observation may suggest the presence of significant multicollinearity or other numerical issues in the model. To overcome the multicollinearity doubt, Variance Inflation Factor ($VIF < 10$) test was done and the values are indicated in Table 3, showing that multicollinearity was not an issue among the external variables used in the analysis.

3.3. Results of SARIMAX model

The Banana crop yield SARIMAX model was configured. Based on the data exploration results, the suggested SARIMAX models were SARIMAX$(0,1,1)(0,1,1)_{12}$, SARIMAX$(0,1,1)(0,1,0)_{12}$, SARIMAX$(0,1,2)(0,1,1)_{12}$, and SARIMAX$(0,1,2)(0,1,0)_{12}$. The model complied with the Box–Jenkins technique, including model fitting, which comprised model identification, here (SARIMAX$(0,1,2)(0,1,0)_{12}$) model was selected, parameter estimation, estimates are indicated in Table 4, and diagnostic checking.

In the training set, the predicted crop yields for the first 40 years closely align with the observed crop yields, as depicted in Fig. 2(a). This observation indicates that the model is successfully identifying the underlying patterns in the data. In the validation set, the predicted crop yields closely match the observed crop yields for the first 4 years, indicating that the model is performing well on unseen data. This alignment between predictions and actual values suggests that the model's generalization capability is satisfactory for new data points. However, Fig. 2(b) reveals a notable discrepancy between the observed and predicted crop yields from 4 to 8 years. Nevertheless, the model does well between 8 and 10 years, indicating that it could be able to successfully capture the underlying patterns in the validation data throughout that time.

Finally, the model forecasting future yields for the next 10 time steps. The last values, which are 9.8245 and 10.5738 from the validation set used as the initial inputs for scenario 1 and scenario 2 respectively, and then the model iteratively predicts the next value based on the previous prediction. The forecasted yields for Scenario 1 are as follows:

6.3705, 5.9595, 6.6489, 5.9003, 6.2839, 9.5054, 8.8109, 11.7376, 10.3568, 10.5829. These values represent the forecasted crop yields for Scenario 1 over a forecast horizon of 10 time steps. The forecasted yields for Scenario 1 suggest a pattern of fluctuating values. The yields start at 6.37, decrease to 5.96, increase to 6.65, then fall again to 5.90. The subsequent yields show further variation, reaching a peak of 11.74 and then stabilizing around 10.36 and 10.58. The forecasted yields for Scenario 2 are as follows:

10.5738, 6.3705, 6.3542, 6.519, 6.3192, 6.4653, 6.5113, 6.6108, 6.6679, 6.6528. In Scenario 2, the forecasted yields exhibit a different pattern. The final observed value from the validation set, which equals 10.5738, was used as the model's initial value. However, the subsequent forecasted yields diverge from this initial value and gradually decrease. The yields range from 6.32 to 6.67, showing a consistent downward trend. Generally, the model predicts slightly lower crop yields in both Scenario 1 and Scenario 2. For a comprehensive analysis of the model's performance, Fig. 3 below are the plots providing visual representations of the predicted and forecasted yields:

3.4. Results of state space (SS) model

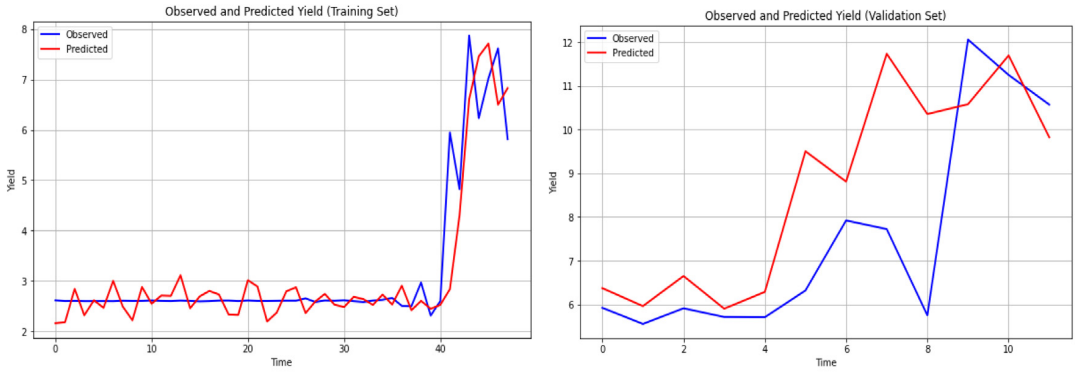
It was done to use the State Space concept. Following that, the State Space model's state vector, state transition matrix, observation matrix, process noise covariance matrix, and measurement noise covariance matrix were defined. These elements are crucial for defining the dynamics and uncertainty properties of the model. The identity matrix was used to construct the state covariance matrix, and arrays were initialized to hold the outcomes. The results of the Kalman filter

Table 3
OLS regression results.

Model	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)
OLS	0.502	0.455	10.87	2.89e−07

Variable	Coef.	Std. Err.	t-stat	P-value	[0.025, 0.975]	VIF
Constant	−22.8320	30.506	−0.748	0.457	[−83.993, 38.329]	–
X_1	0.0206	0.043	0.478	0.634	[−0.066, 0.107]	2.9606
X_2	−0.0085	0.007	−1.147	0.257	[−0.023, 0.006]	1.9909
X_3	4.8328	1.628	2.968	0.004	[1.569, 8.097]	6.7376
X_4	−1.6594	1.648	−1.007	0.318	[−4.963, 1.644]	7.6477
X_5	−0.0991	0.069	−1.439	0.156	[−0.237, 0.039]	1.1402

More model information			
Method:	Least squares	AIC:	245.4
No. observations:	60	BIC:	257.9
Df residuals:	54	Kurtosis:	5.039
Df model:	5	Skewness:	1.000
Covariance type:	nonrobust	Jarque–Bera (JB):	20.380
Durbin–Watson:	1.192	Prob(JB):	3.75e−05
Cond. No.:	4.43e+04	Omnibus:	15.590
Log-Likelihood:	−116.68	Prob(Omnibus):	0.000



(a) Training Set (b) Validation Set

Fig. 2. The observed and predicted banana crop yield for the SARIMAX model.

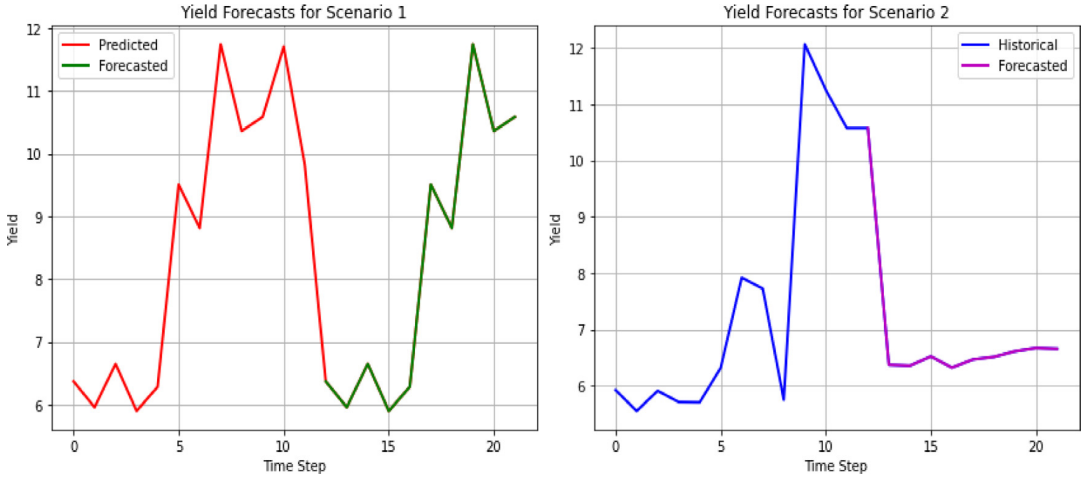
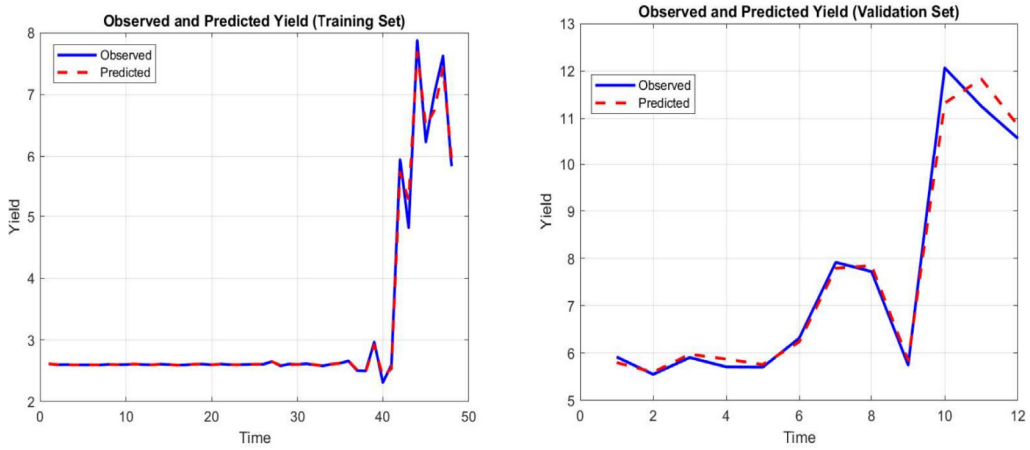


Fig. 3. The plot of banana crop yield forecasting for the SARIMAX model.

algorithm-based parameter estimate for the state space model are also shown in Table 2.

Generally speaking, the SS model shows a strong match to the training set of data, with low prediction errors (MSE and MAE), a small standard deviation of errors (RMSE), and a high proportion of explained variability (R-squared). However, the model's performance is slightly reduced when applied to the validation set, with slightly higher prediction errors and a slightly lower coefficient of determination. Fig. 4, are training and validation plots, plotted to compare the observed and predicted crops yield for validating the model performance. The trend of the observed yields is determined by the model, there are some deviations between the observed and predicted yields. The red dashed line shows some discrepancies and variations from the blue line, indicating that the model's predictions are not as accurate for the validation set as they were for the training set. The plots demonstrate that the state space model performs well in predicting the crop yields, particularly for the training set. The model shows a strong ability to capture the trends and fluctuations in the observed yields.



(a) Training Set (b) Validation Set

Fig. 4. The observed and predicted banana crop yield for the SS model.

Table 4
Estimated parameters for SARIMAX$(0,1,2)(0,1,0)_{12}$ model.

$R^2(Training\,Set)$		$R^2(Testing\,Set)$		AIC	BIC
0.8109		0.1825		91.469	103.911
Variable	Coef.	Std. Err.	t-stat	P-value	[0.025, 0.975]
X_1	0.0166	0.032	0.510	0.610	[−0.047, 0.080]
X_2	−0.0016	0.006	−0.243	0.808	[−0.014, 0.011]
X_3	−0.1055	1.421	−0.074	0.941	[−2.891, 2.680]
X_4	0.0635	1.253	0.051	0.960	[−2.392, 2.519]
X_5	0.0110	0.068	0.162	0.872	[−0.122, 0.144]
ma.L1	−0.3858	0.177	−2.181	0.029	[−0.733, −0.039]
ma.L2	0.5397	0.184	2.939	0.003	[0.180, 0.900]
sigma2	0.4894	0.188	2.604	0.009	[0.121, 0.858]

At the end, the SS model performs forecasts for the next 10 time steps using the final states, which are 10.8487, and 10.5738 from the predicted yield for scenario 1 and true yield values for scenario 2 as the initial states respectively. The forecasted yields for Scenario 1 are as follows:

7.8510, −0.0261, −15.7032, −42.9991, −86.7473, −152.9127, −248.7089, −382.7142, −564.9889, −807.1921. Based on the SS model and using the final expected yield as input, these forecasted yields represent the crop yields predicted for the following 10 time steps. The forecasted yields show a decreasing trend, with the magnitudes becoming more negative as time progresses. The negative values suggest a decrease in crop yields over time, indicating potentially unfavorable conditions or factors affecting crop growth. The forecasted yields for Scenario 2 are as follows:

5.9203, 5.5512, 5.9086, 5.7096, 5.7043, 6.3168, 7.9205, 7.7233, 5.7479, 12.0627. Based on the state space model and the final true yield as input, these forecasted yields show the crop yields that are expected throughout the course of the next 10 time steps. The forecasted yields show some fluctuations but do not exhibit a clear trend. The values vary within a relatively narrow range, suggesting relatively stable or consistent crop yields over time. Below (Fig. 5) are the plots providing visual representations of the predicted and forecasted yields, allowing for a comprehensive analysis of the model's performance.

3.5. Results of LSTM model

The LSTM model also was configured. The sequences and labels necessary for the LSTM model were generated, followed by constructing and training the model. Subsequently, predictions were made, and the scaled predictions were reverted back to their original form using the inverse transform. The R-squared value for the training data indicates

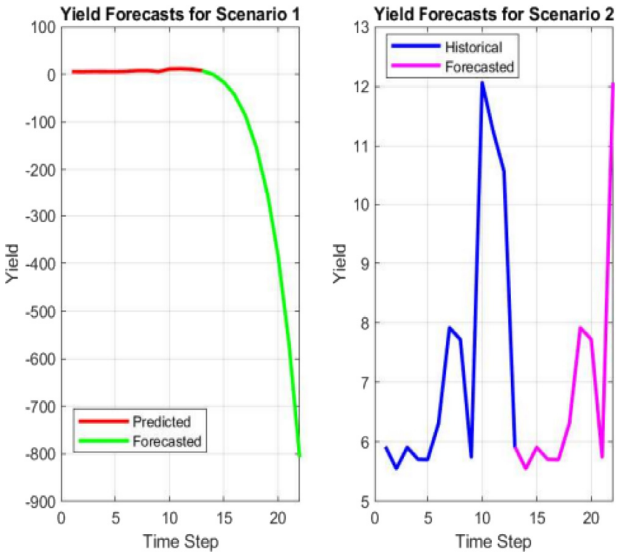
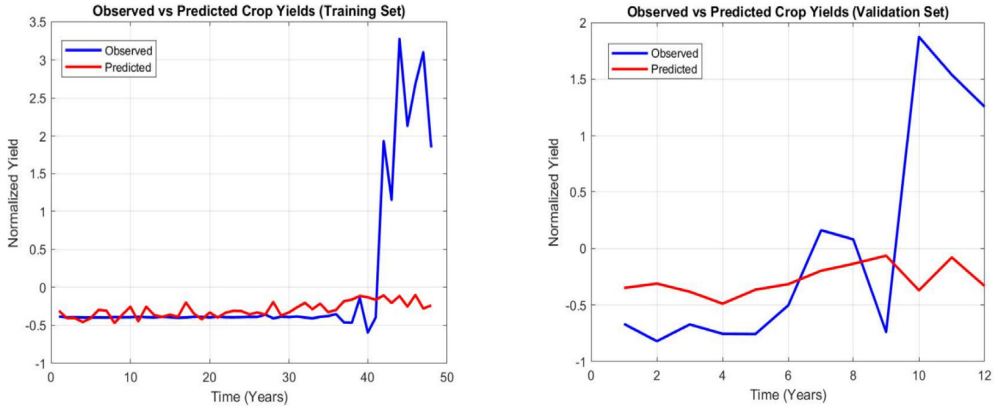


Fig. 5. The plot of banana crop yield forecasting for the State Space model.

that the LSTM model exhibits good performance on the provided dataset. The MSE, MAE, RMSE, and R-squared values for each model employed in this study are shown in Table 2. These metrics provide insights into the performance and accuracy of each model. Overall, the evaluation results indicate that the LSTM network in this analysis provide a good fit to the data, as evidenced by low errors (MSE, MAE, RMSE) and high coefficient of determination (R-squared).

The trained LSTM network was used to predict crop yields for both the training and validation sets. The predicted yields were compared with the actual yields to evaluate the model's fit. The model fit evaluation metrics provided insights into how well the LSTM network performs in predicting crop yields. Hence, the observed and predicted crop yields were plotted for both the training and validation sets to visually compare their trends and performance, as shown in Fig. 6.

In the training data, the predicted crop yields for the first 40 years closely align with the observed crop yields, as depicted in Fig. 6(a). This observation indicates that the model is successfully identifying the underlying patterns in the data. The considerable difference between the observed and anticipated crop yields after 40 s, on the other hand, suggests that the correlations in the training set may not be accurately captured by the model. However, Fig. 6(b) demonstrates that the predicted crop yields closely align with the observed crop yields for the initial 10 years, indicating that the model performs well



(a) Training Set (b) Validation Set

Fig. 6. The observed and predicted banana crop yield for the LSTM model.

on unseen data. This alignment between predictions and actual values suggests that the model has a satisfactory generalization capability for new data points. The actual and predicted crop yields, however, show a significant difference after 10 years, suggesting that the model may be having trouble capturing the underlying trends in the validation data. This disparity suggests that the model's predictive accuracy diminishes for the later time periods of the validation set.

Finally, the model forecasting future yields for the next 10 time steps. The last predicted and true values, which are −0.3610 and 10.5738 from the validation set used as the initial inputs for scenario 1 and scenario 2 respectively, and then the model iteratively predicts the next value based on the previous prediction. The forecasted yields for Scenario 1 are as follows:

−0.2636, −0.0950, −0.3842, −0.1814, −0.2485, −0.2660, −0.2304, −0.2616, −0.2469, −0.2483. These values represent the forecasted crop yields for Scenario 1 over a forecast horizon of 10 time steps. Since these values are normalized yields, they indicate the predicted yields relative to the range of yields observed in the validation data. The negative values represent the predicted yields being lower than the average yield in the validation dataset. In Scenario 1, the model predicts relatively low crop yields for the forecasted time steps. The forecasted yields for Scenario 2 are as follows:

−1.2502, −0.3620, −0.4858, −0.6064, −0.5217, −0.3284, −0.2751, −0.2657, −0.2618, −0.2626. These values represent the forecasted crop yields for Scenario 2 over a forecast horizon of 10 time steps. The values range from −1.2502 to −0.2618. Similar to Scenario 1, these values represent normalized yields and indicate the predicted yields relative to the validation data. In Scenario 2, the model predicts even lower crop yields compared to Scenario 1. Generally, the model predicts lower crop yields in both Scenario 1 and Scenario 2. Below are the plots (Fig. 7) providing visual representations of the predicted and forecasted yields:

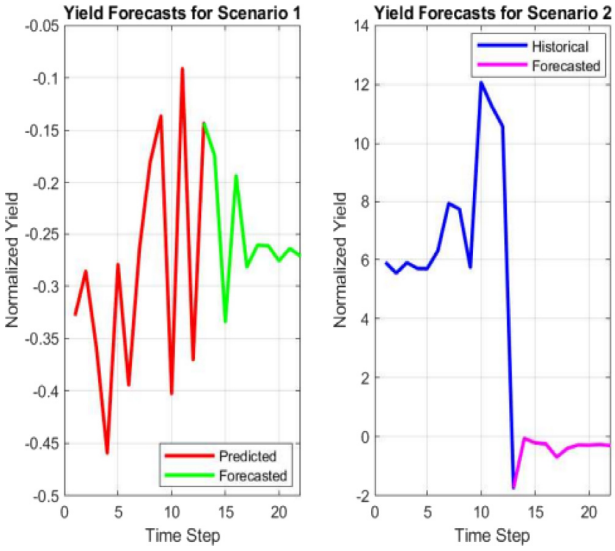


Fig. 7. The plot of banana crop yield forecasting for the LSTM model.

3.6. Results of ensemble model

As we discussed before, once we have trained and evaluated the SARIMAX, State Space, and LSTM models on the datasets, we can proceed with determining the weights, and obtaining final predicted values steps to formulate ensemble model for forecasting banana crop yield. We determined the weights of the individual models based on their performance on the validation set. We determined the weights depending on how well each model fit the data. The R-squared (Coefficient of Determination) represents a valuable metric for evaluating the overall goodness of fit and the extent to which the model captures the variability in the data. The R-squared values of the SARIMAX, State Space, and LSTM models are 0.1825, 0.9835, and 0.9013 respectively, as indicated in Table 2.

In computing the final predicted value, the ensemble model represented as:

$$y = 0.7204\times 9.8245 + 0.1337\times 10.8487 + 0.1459\times -0.3610 \rightarrow y = 8.4754. \quad (29)$$

The normalized weights of the SARIMAX, State Space, and LSTM models are 0.7204, 0.1337, and 0.1459 respectively. The final predicted values from the SARIMAX, State Space, and SLTM models are 9.8245, 10.8487, and −0.3610 respectively. The ensemble model's final predicted value of 8.4754 is the result of combining the outputs from the individual SARIMAX, State Space, and LSTM models. The individual models predict different values, with State Space predicting the highest value of 10.8487, followed by SARIMAX with 9.8245, and LSTM with −0.3610. The ensemble model combines the predictions of these individual models using weights that are optimized during validation to give the final predicted value. Therefore, the ensemble model can be seen as a more robust and accurate model as it takes into account the strengths and weaknesses of the individual models to provide a more accurate prediction (Bertsimas and Boussioux, 2023).

The last phase involved evaluating the ensemble model's performance using relevant metrics, as presented in Table 5. This entailed comparing the ensemble model's performance with that of the individual models to gauge its effectiveness. Thus, the R-squared of SARIMAX, State Space, LSTM, and Ensemble models are 0.1825, 0.9835, 0.9013, and 0.9999999999891197 respectively. The R-squared values provide

Table 5
Evaluation metrics for the ensemble model.

Metric	Value
Mean Squared Error (MSE)	8.35788099957876e−10
Mean Absolute Error (MAE)	2.8029999999290567e−05
Root Mean Squared Error (RMSE)	5.2945999999290567e−05
R-squared	0.9999999999891197

a measure of how well each of the models has performed on the validation data. A higher R-squared value indicates that the model is better at predicting the actual values.

In this instance, the ensemble model achieved the highest R-squared value compared to all other models, suggesting that it outperformed the other models on the validation data. The SARIMAX model has the lowest R-squared value, indicating that it is the worst performer among all the models. The LSTM and State Space models have slightly similar R-squared values, with the State Space model performing slightly better than the LSTM model.

4. Conclusion

This study focuses on the configuration and forecasting of banana crop yield in Tanzania, considering the impact of climate change. In particular, this study delves into the intricate relationship between climate change and Tanzanian banana crop yield. It aims to understand how changing climatic conditions might impact agricultural outcomes, especially in the context of a country like Tanzania. In pursuit of the addressed objective of this study, the researchers takes a two-fold approach, including correlation analysis and forecasting models. A robust multiple regression model uncovers valuable insights within this connection. Time series analysis and ensemble modeling techniques are employed to develop accurate forecasting models that incorporate climate variables and capture the dynamics of banana production in Tanzania. The findings emphasize the significance of accounting for climate change in banana crop yield forecasting. By examining the correlations between climatic variables and banana crop yield, the models provide vital insight into the potential impacts of climate change on banana production.

In light of these key climate variables at hand, this study revealed that Tanzania's banana crop yield has been impacted by climate change, offering insights into potential vulnerabilities. The insights gleaned from this study offer a critical foundation for actionable policy recommendations and strategies to safeguard and enhance banana production in Tanzania amidst the challenges posed by climate change. It is imperative that policymakers, researchers, and farmers collaborate to implement the following measures: climate-resilient practices, data-driven decision-making, infrastructure investment, policy flexibility, knowledge dissemination, and continued research. By implementing these recommendations, Tanzania can fortify its banana production sector against the disruptive effects of climate change. Together, stakeholders can work towards sustainable banana production, ensuring food security and prosperity for the nation's agricultural communities.

Utilizing time series analysis techniques like SARIMAX, State Space, and LSTM helps identify relevant patterns and trends in historical datasets, forming the foundation for robust forecasting models. The ensemble modeling approach further enhances the accuracy and reliability of predictions by combining multiple individual models, while the integration of climate variables improves the precision of forecasts. Understanding the specific climatic factors influencing banana crop yield can inform decisions related to agricultural practices, resource allocation, and policy planning.

Future research can build upon these findings by incorporating additional variables and employing machine learning techniques for even more accurate predictions. The prospective effects of climate change on Tanzania's banana crop yield can also be assessed with the help of impact assessment and climate modeling approaches. Eventually, it is possible to successfully raise knowledge about the hazards posed by climate change to the region's banana crop output by planning workshops and outreach activities for farmers and stakeholders.

CRediT authorship contribution statement

Sabas Patrick: Conceptualization, Methodology, Writing – original draft. **Silas Mirau:** Manuscript – review & editing, Insightful ideas and suggestions. **Isambi Mbalawata:** Methodology, Critical feedback. **Judith Leo:** Financial support, Supervised a research project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledges the funding received from West African Science Service Centre on Climate Change and Adapted Land Use (WASCAL - RUFORUM) Capacity Building in Agriculture at The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania (NM-AIST) under RAINCA Project.

References

Abdoussalami, Andlia, Hu, Zhenghua, Islam, Abu Reza Md Towfiqul, Wu, Zhurong, 2023. Climate change and its impacts on banana production: a systematic analysis. Environ. Dev. Sustain. 1–30.

Adejuwon, J.O., Agundiminegha, Y.G., 2019. Impact of climate variability on cassava yield in the humid forest agro-ecological zone of Nigeria. J. Appl. Sci. Environ. Manage. 23 (5), 903–908.

Anwar, Samy A., Zakey, A.S., Robaa, S.M., Abdel Wahab, M.M., 2019. The influence of two land-surface hydrology schemes on the regional climate of africa using the RegCM4 model. Theor. Appl. Climatol. 136, 1535–1548.

Anzures, Arielle Francis, Hipolito, Kristina, Pestolante, Katherine, et al., 2022. Constraints in the primary production of bananas in the davao region, Philippines. Int. J. Soc. Manage. Stud. 3 (1), 1–31.

Aoki, Masanao, 2013. State Space Modeling of Time Series. Springer Science & Business Media.

Arunraj, Nari Sivanandam, Ahrens, Diane, Fernandes, Michael, 2016. Application of SARIMAX model to forecast daily sales in food retail industry. Int. J. Oper. Res. Inf. Syst. (IJORIS) 7 (2), 1–21.

Bayati, Abdolkhalegh, Nguyen, Kim-Khoa, Cheriet, Mohamed, 2020. Gaussian process regression ensemble model for network traffic prediction. IEEE Access 8, 176540–176554.

Bertsimas, Dimitris, Boussioux, Leonard, 2023. Ensemble modeling for time series forecasting: an adaptive robust optimization approach. arXiv preprint arXiv:2304.04308.

de Bézenac, Emmanuel, Rangapuram, Syama Sundar, Benidis, Konstantinos, Bohlke-Schneider, Michael, Kurle, Richard, Stella, Lorenzo, Hasson, Hilaf, Gallinari, Patrick, Januschowski, Tim, 2020. Normalizing kalman filters for multivariate time series analysis. Adv. Neural Inf. Process. Syst. 33, 2995–3007.

Bhausaheb, Takale Asmita, Lazarus, T Paul, Vijayan, Aswathy, Sathayan, Archana R, Joseph, Brigit, 2023. Impact of climate change on banana production in thiruvananthapuram district of kerala, India. Asian J. Agric. Extens. Econ. Sociol. 41 (3), 114–123.

Bhimavarapu, Usharani, Battineni, Gopi, Chintalapudi, Nalini, 2023. Improved optimization algorithm in LSTM to predict crop yield. Computers 12 (1), 10.

Box, George EP, Jenkins, Gwilym M, Reinsel, Gregory C, Ljung, Greta M, 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Chowhan, Sushan, Ghosh, Shapla Rani, Chowhan, Tushar, Hasan, Md Mahmudul, Roni, Md Shyduzzaman, 2016. Climate change and crop production challenges: An overview. Res. Agric. Livest. Fish. 3 (2), 251–269.

Cook, Edward Roger, 1985. A time series analysis approach to tree ring standardization (Ph.D. thesis). University of Arizona Tucson.

Hanson, Timothy, 2010. Multiple regression.

Hao, Tianxiao, Elith, Jane, Lahoz-Monfort, José J, Guillera-Arroita, Gurutzeta, 2020. Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. Ecography 43 (4), 549–558.

Hooda, Ekta, Verma, Urmil, Hooda, B.K., 2020. ARIMA and state-space models for sugarcane (saccharum officinarum) yield forecasting in northern agro-climatic zone of haryana. J. Appl. Nat. Sci. 12 (1), 53–58.

Hoque, M.Z., Haque, M.E., 2016. Impact of climate change on crop production and adaptation practices in coastal saline areas of Bangladesh. Int. J. Appl. Res. 2 (1), 10–19.

Hu, Yawei, Liu, Shujie, Lu, Huitian, Zhang, Hongchao, 2019. Remaining useful life model and assessment of mechanical products: a brief review and a note on the state space model method. Chin. J. Mech. Eng. 32, 1–20.

Hyndman, Rob J., Athanasopoulos, George, 2018. Forecasting: Principles and Practice. OTexts.

Ighalo, Joshua O., Adeniyi, Adewale George, 2019. Thermodynamic modelling and temperature sensitivity analysis of banana (musa spp.) waste pyrolysis. SN Appl. Sci. 1 (9), 1–9.

Jayasinghe, S.L., Ranawana, C.J.K., Liyanage, I.C., Kaliyadasa, P.E., 2022. Growth and yield estimation of banana through mathematical modelling: A systematic review. J. Agric. Sci. 1–58.

Kahimba, FC, Sife, AS, Maliondo, SMS, Mpeta, EJ, Olson, Jennifer, 2015. Climate change and food security in tanzania: Analysis of current knowledge and research gaps. Tanzan. J. Agric. Sci. 14 (1).

Kamir, Elisa, Waldner, François, Hochman, Zvi, 2020. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J. Photogramm. Remote Sens. 160, 124–135.

Kourentzes, Nikolaos, Barrow, Devon K., Crone, Sven F., 2014. Neural network ensemble operators for time series forecasting. Expert Syst. Appl. 41 (9), 4235–4244.

Lai, Yuchuan, Dzombak, David A., 2020. Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation. Weather Forecast. 35 (3), 959–976.

Lal, Narayan, Sahu, Nisha, Shuirkar, Govind, Jayswal, Dalit Kumar, Chack, Sonbeer, 2017. Banana: Awesome fruit crop for society.

Liu, Fan, Jiang, Xiangtao, Wu, Zhenyu, 2023. Attention mechanism-combined LSTM for grain yield prediction in China using multi-source satellite imagery. Sustainability 15 (12), 9210.

Lokupitiya, Erandathie, 2018. Book of abstracts of 2nd international conference on climate change 2018 (ICCC 2018). Climate change conference. Colombo, Sri Lanka: The international institute of knowledge management (TIIKM).

Lucas, Shija Shilunga, Jomanga, Kennedy Elisha, 2021. The status of banana production in tanzania; a review of threats and opportunities.

Marolla, Filippo, Henden, John-André, Fuglei, Eva, Pedersen, Åshild Ø, Itkin, Mikhail, Ims, Rolf A, 2021. Iterative model predictions for wildlife populations impacted by rapid climate change. Global Change Biol. 27 (8), 1547–1559.

Mayaya, Hozen K., 2015. Community adaptation and mitigation strategies to climate change in semi-arid areas of dodoma region, tanzania (Ph.D. thesis). SCHOOL OF ENVIRONMENTAL STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS ....

Mbigi, Dickson, Xiao, Ziniu, 2021. Analysis of rainfall variability for the october to december over tanzania on different timescales during 1951–2015. Int. J. Climatol. 41 (14), 6183–6204.

Meeradevi, Yasaswi, IGS, Mundada, Monica R, Sarika, D, Shetty, Harshita, 2022. Hybrid decision support system framework for enhancing crop productivity using machine learning. In: Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021. Springer, pp. 57–66.

Moore, Frances C., Lobell, David B., 2014. Adaptation potential of European agriculture in response to climate change. Nature Clim. Change 4 (7), 610–614.

Neog, Borsha, Gogoi, Bipin, Patowary, A.N., 2022. Development of hybrid time series models for forecasting autumn rice using arimax-ann and arimax-svm.. Ann. for. Res 65 (1), 9119–9133.

Newman, Ken, King, Ruth, Elvira, Víctor, de Valpine, Perry, McCrea, Rachel S, Morgan, Byron JT, 2023. State-space models for ecological time-series data: Practical model-fitting. Methods Ecol. Evol. 14 (1), 26–42.

Ngo, Theresa Hoang Diem, La Puente, C.A., 2012. The steps to follow in a multiple regression analysis. In: Proceedings of the SAS Global Forum. Citeseer, pp. 22–25.

Omambia, Ceven Shemsanga, Gu, Yansheng, 2010. The cost of climate change in tanzania: impacts and adaptations. J. Am. Sci. 6 (3), 182–196.

Ongoma, Victor, Chen, Haishan, Gao, Chujie, 2019. Evaluation of CMIP5 twentieth century rainfall simulation over the equatorial east africa. Theor. Appl. Climatol. 135 (3–4), 893–910.

Pham, Yen, Reardon-Smith, Kathryn, Mushtaq, Shahbaz, Cockfield, Geoff, 2019. The impact of climate change and variability on coffee production: a systematic review. Clim. Change 156, 609–630.

Raj, Esack Edwin, Ramesh, K.V., Rajkumar, Rajagobal, 2019. Modelling the impact of agrometeorological variables on regional tea yield variability in south Indian tea-growing regions: 1981–2015. Cogent Food Agric. 5 (1), 1581457.

Rathod, S., Mishra, G.C., 2018. Statistical models for forecasting mango and banana yield of karnataka, India. J. Agric. Sci. Technol. 20 (4), 803–816.

Reddy, Mallidi PSR, Mathur, Ayush K, Jain, Rohit K, Agarwal, Sandip K, Singh, Sriramjee, 2022. Climate change and weather variability in crop modelling: Evidence from rice yield trials in India using LSTM model.

Sadowski, Peter, 2016. Notes on backpropagation. homepage: https://www.ics.uci.edu/pjsadows/notes.pdf (online).

Sagamiko, Thadei, Shaban, Nyimvua, Mbalawata, Isambi, 2020. Sensitivity analysis and uncertainty parameter quantification in a regression model: The case of deforestation in tanzania. Tanzan. J. Sci. 46 (3), 673–683.

Salvacion, Arnold R., 2020. Effect of climate on provincial-level banana yield in the Philippines. Inf. Process. Agric. 7 (1), 50–57.

Shirima, Andrew Omari, Lubawa, Galinoma, 2017. Farm based adaptation strategies to climate change among smallholder farmers in manyoni district, tanzania. Int. J. Res. Soc. Sci. 7 (7), 1–22.

Suman, Suman, Verma, Urmil, 2017. State space modelling and forecasting of sugarcane yield in haryana, India. J. Appl. Nat. Sci. 9 (4), 2036–2042.

Tian, Huiren, Wang, Pengxin, Tansey, Kevin, Zhang, Jingqi, Zhang, Shuyu, Li, Hongmei, 2021. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, PR China. Agricult. Forest Meteorol. 310, 108629.

Van Leeuwen, Sonja M, Lenhart, Hermann-J, Prins, Theo C, Blauw, Anouk, Desmit, Xavier, Fernand, Liam, Friedland, Rene, Kerimoglu, Onur, Lacroix, Genevieve, Van Der Linden, Annelotte, et al., 2023. Deriving pre-eutrophic conditions from an ensemble model approach for the north-west European seas. Front. Mar. Sci. 10, 1129951.

Varma, Varun, Bebber, Daniel P., 2019. Climate change impacts on banana yields around the world. Nat. Clim. Change 9 (10), 752–757.

Verma, Suman, 2018. Modeling and forecasting maize yield of India using ARIMA and state space models. J. Pharm. Phytochem. 7 (5), 1695–1700.

Wood, Stephen A, Jina, Amir S, Jain, Meha, Kristjanson, Patti, DeFries, Ruth S, 2014. Smallholder farmer cropping decisions related to climate variability across multiple regions. Global Environ. Change 25, 163–172.

POSTER PRESENTATION

FORECASTING TANZANIAN BANANAS UNDER THE EFFECTS OF CLIMATE CHANGE

Sabas Patrick[1], Silas Mirau[1], Isambi Mbalawata[2], Judith Leo[1]

[1]Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania.
[2]African Institute for Mathematical Sciences, Kigali, Rwanda.

Introduction

Banana is a significant crop globally, used for food, income, and health benefits. It's among the top crops worldwide in terms of productivity and plays a crucial role in food security, particularly in Tanzania [1]. Climate change poses serious risks to agricultural systems. These risks include altered temperature and rainfall patterns, as well as extreme weather events. This is especially true in countries like Tanzania that are heavily dependent on crop agriculture [1, 2]. Tanzania faces significant challenges due to climate change, impacting agricultural productivity. Farmers encounter obstacles hindering the growth of the agricultural sector [3, 4]. Accurate forecasting models are crucial for ensuring resilience and adaptability in banana farming amid shifting climatic conditions [5].

Research objectives

This study's primary goal is to utilize time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change.
The study pursued the following specific objectives:

(i) To assess the impact of climate change on banana crop yield in Tanzania.

(ii) To determine the sensitivity of banana crop yield to climate variables.

(iii) To develop time series models that can accurately forecast banana crop yield in Tanzania under the effects of climate change.

(iv) To develop an ensemble model that can improve the accuracy of banana crop yield forecasting.

Materials and Methods

This study seeks to comprehend how banana crop yield, as a crucial agricultural product is affected by changing climatic patterns. It takes a two-fold approach: Correlation analysis to examine the relationship between bananas and important climate variables and forecasting models to forecast future banana yields in the context of changing climates. While an ensemble model incorporated to improve prediction accuracy of these forecasting models. The schematic diagram in Fig. 1 indicates the flow of the whole work:

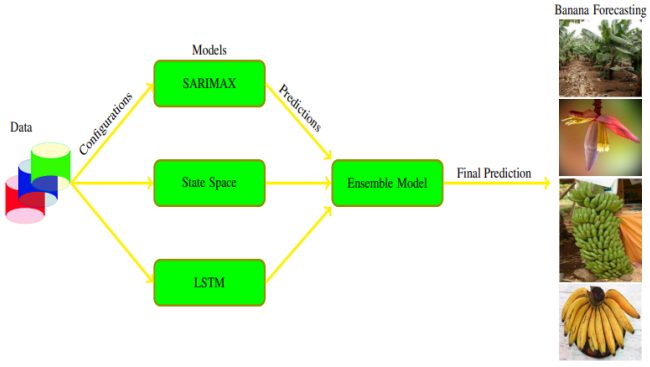


Fig. 1: The schematic diagram, a representation of methodology.

The study quantitatively expressed the multiple regression model's equation as follows:

$$Y = \Psi_0 + \Psi_1 X_1 + \Psi_2 X_2 + \Psi_3 X_3 + \Psi_4 X_4 + \Psi_5 X_5 + \varepsilon \quad (1)$$

The weighted average approach for the ensemble model is represented mathematically as follows:

$$y = w_1 \times y_1 + w_2 \times y_2 + w_3 \times y_3 + \cdots + w_n \times y_n \quad (2)$$

Results

In this study, the dataset spanning 60 years, from 1961 to 2020. The first 80% of the dataset was used to train the models, while the final 20% were used to validate the models. The MATLAB (R2021a), and Spyder (Python 3.9) tools were used interchangeably throughout the analysis.

The study derived the following expression for the multiple regression model:

$$Y = -22.8320 + 0.0206X_1 - 0.0085X_2 + 4.8328X_3 - 1.6594X_4 - 0.0991X_5 \quad (3)$$

When calculating the final predicted value, the ensemble model is represented as follows:

$$y = 0.7204 \times 9.8245 + 0.1337 \times 10.8487 + 0.1459 \times -0.3610 \rightarrow y = 8.4754. \quad (4)$$

A high positive coefficient and a small negative coefficient in Eq. (3) indicate that the variable has a significant impact on banana yields. The R^2 of SARIMAX, State Space, LSTM, and Ensemble models are 0.1825, 0.9835, 0.9013, and 0.9999999999891197 respectively. A higher R^2 value indicates that the model is better at predicting the actual values.

Conclusion

In conclusion, this study revealed that Tanzania's banana crop yield has been impacted by climate change. The results showed gradual decrease in bananas while showing that minimum temperature, precipitation and soil moisture have the most impact on bananas and affect the crop's production variability. This necessitates imperative collaboration among policymakers, researchers, and farmers to implement the following measures [6]:

(i) *Climate-Resilient Farming Practices:* Promoting drought-resistant banana varieties, and optimizing irrigation systems.

(ii) *Data-Driven Decision Making:* Exploring opportunities to discover alternative sources of subnational-level data on banana yields is crucial.

(iii) *Investment in Infrastructure:* Improved water management systems, and post-harvest facilities.

(iv) *Policy Flexibility:* Policymakers should be prepared to adjust policies in response to changing climate realities and emerging challenges.

(v) *Knowledge Dissemination:* Workshops, training programs, and outreach activities for stakeholders to raise awareness.

Limitations

The challenges in this study include: Handling uncertainties and the dynamic nature of climate, unavailability of alternative sources at the subnational level for banana yield data, and the study acknowledges limitations in the inclusivity of climate variables.

Future Work

Future research should consider: Additional climate variables and management factors, improving the analysis by setting up the models at the subnational level and finding an alternative source of data on banana yields, and incorporate artificial intelligence (AI) and machine learning techniques

Acknowledgments

We acknowledge the financial support provided by the West African Science Service Centre on Climate Change and Adapted Land Use (WASCAL - RUFORUM) Capacity Building in Agriculture at NM-AIST through the RAINCA Project.

Contact information:
Sabas Patrick
patricks@mn-aist.go.tz

References

[1] Shija Shilunga Lucas and Kennedy Elisha Jomanga. The status of banana production in tanzania; a review of threats and opportunities. 2021.

[2] Varun Varma and Daniel P Bebber. Climate change impacts on banana yields around the world. *Nature climate change*, 9(10):752–757, 2019.

[3] SL Jayasinghe, CJK Ranawana, IC Liyanage, and PE Kaliyadasa. Growth and yield estimation of banana through mathematical modelling: A systematic review. *The Journal of Agricultural Science*, pages 1–58, 2022.

[4] Andlia Abdoussalami, Zhenghua Hu, Abu Reza Md Towfiqul Islam, and Zhurong Wu. Climate change and its impacts on banana production: a systematic analysis. *Environment, Development and Sustainability*, pages 1–30, 2023.

[5] MZ Hoque and ME Haque. Impact of climate change on crop production and adaptation practices in coastal saline areas of bangladesh. *Int J Appl Res*, 2(1):10–19, 2016.

[6] URT. United republic of tanzania (urt). national climate change response strategy (2021-2026). vice presidentâs office, division of environment, government printer, dodoma, tanzania. 2021.