2023-12-01

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

Patrick, Sabas

Elsevier

https://doi.org/10.1016/j.resenv.2023.100138

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

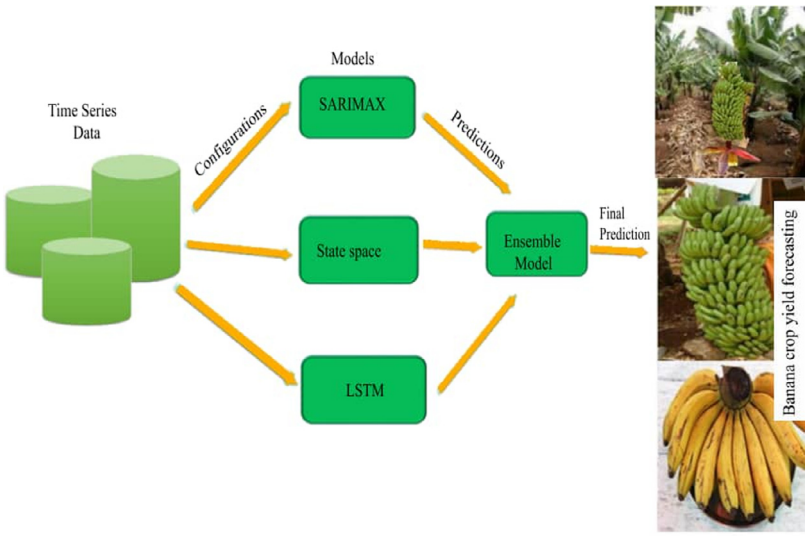


Fig. 1. The schematic diagram, a representation of methodology.

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:

$$Y = -22.8320 + 0.0206X_1 - 0.0085X_2 + 4.8328X_3 - 1.6594X_4 - 0.0991X_5 \tag{28}$$

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

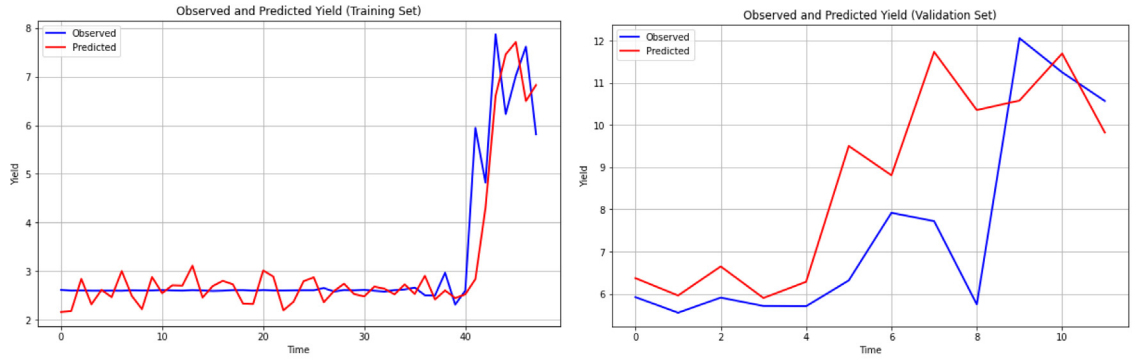
It was done to use the State Space concept. Following that, the State Space model's state vector, state transition matrix, observation matrix, process noise covariance matrix, and measurement noise covariance matrix were defined. These elements are crucial for defining the dynamics and uncertainty properties 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.

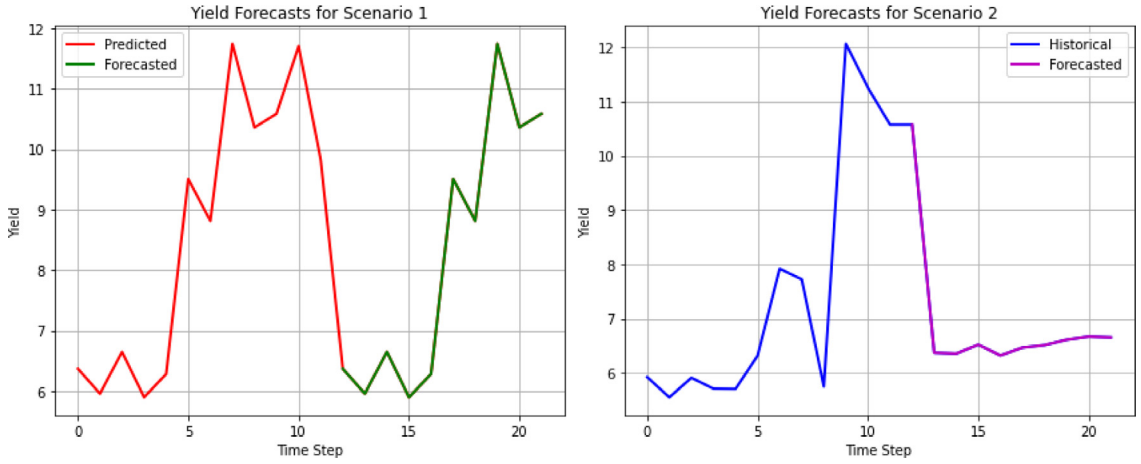
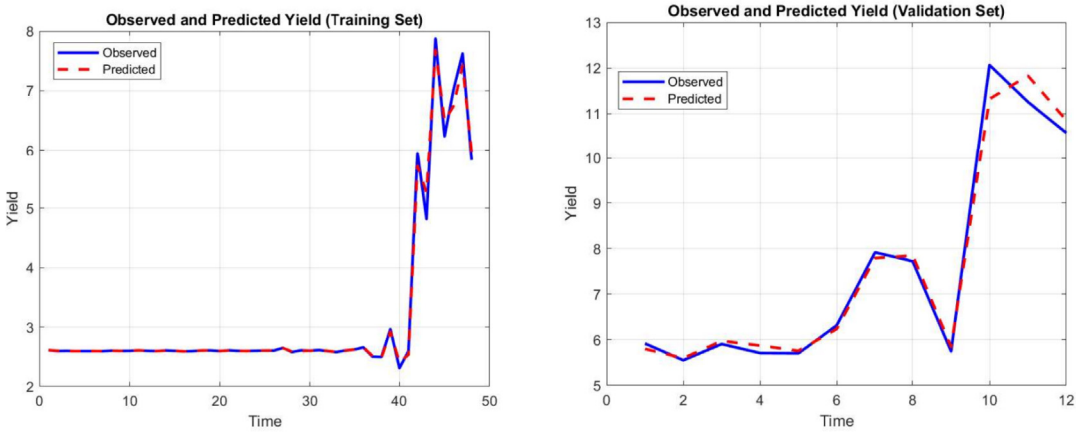


Fig. 3. The plot of banana crop yield forecasting for the SARIMAX model.

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

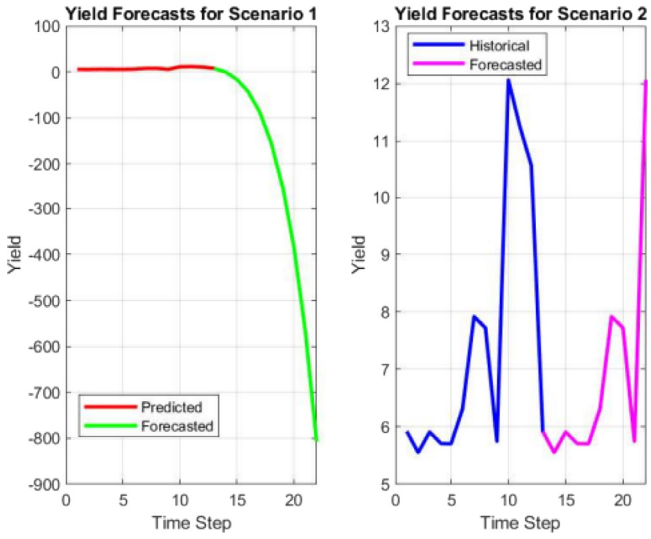
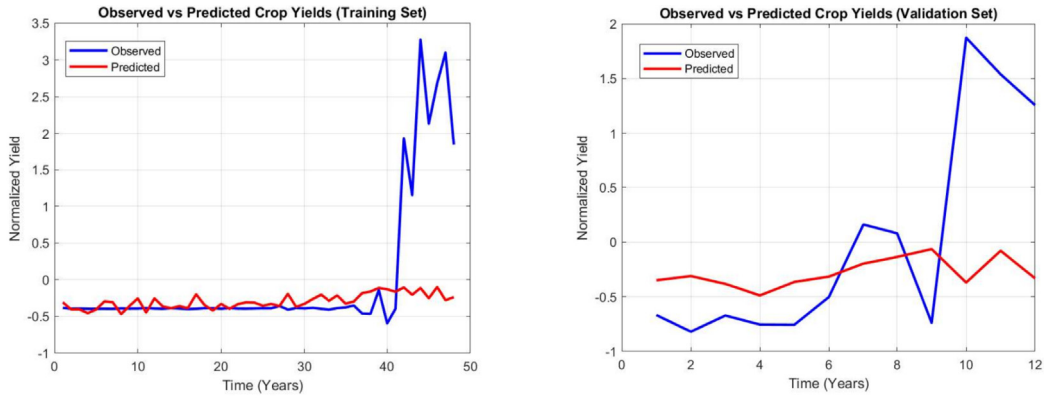


Fig. 5. The plot of banana crop yield forecasting for the State Space model.

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:

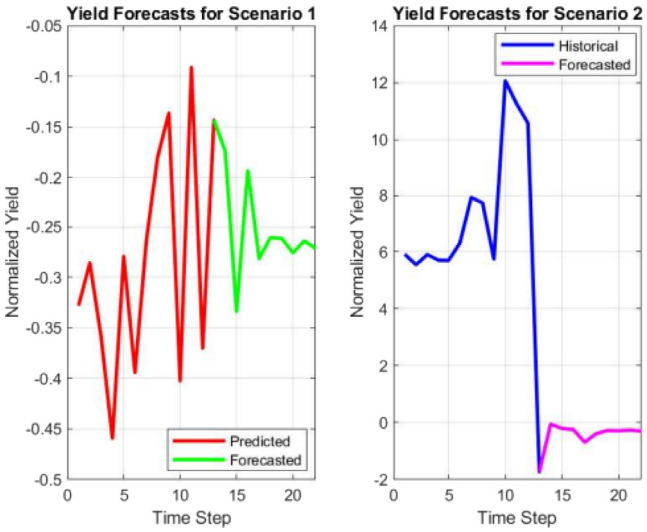


Fig. 7. The plot of banana crop yield forecasting for the LSTM model.

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.