International Journal of Applied Mathematics and Numerical Research

Analysis of Multi- faceted ACG-Hybrid model on Effects of Climate Variability on Soybean Productivity using Artificial Intelligence Techniques

Oloo Erick Odhiambo [1*], Erick Okuto [2], Benard Okelo [3], Samuel Oyieke [4]
[1, 2] Department of Applied Statistics, Financial Mathematics and Actuarial Science, Kenya
[3] Department of Pure and Applied Mathematics, Kenya
[4] Department of Economics, Jaramogi Oginga Odinga University of Science and Technology, P. O. Box 210-40601, Bondo, Kenya

* Corresponding Author: **Oloo Erick Odhiambo**

Article Info

Impact Factor (RSIF): 8.62
ISSN (Online): 3107-7110
Volume: 02
Issue: 04
Received: 09-05-2026
Accepted: 11-06-2026
Published: 13-07-2026
Page No: 50-61

Abstract
The Autoregressive Moving Average (ARIMA) model has also been applied successfully for forecasting maize production in Tanzania, as noted in. Nonetheless, ARIMA is inherently limited to univariate time series data and does not adequately address spatial dependencies. To overcome these limitations, machine learning approaches, particularly deep learning techniques such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks, have been applied for spatiotemporal data analysis. Though the CNN-LTSM hybrid model performed fairly well with spatiotemporal data, it failed to capture stationarity and handle the underlying linear structure in the model, leading to higher error metrics. Moreover, developed an ARIMA-CNN-LSTM model to forecast carbon pricing, effectively capturing both linear and non-linear data characteristics, and achieving lower RMSE compared to benchmark models. However, a key drawback of the ARIMA-CNN-LSTM model is the lengthy training time due to the complexity of the LSTM component. In light of these changes, we propose the development of a novel hybrid model, termed the ARIMA-CNN-GRU (ACG)-hybrid model, which seeks to bridge the gaps identified in conventional and existing hybrid deep learning models. This model leverages the strengths of the ARIMA as a foundational model to handle the underlying linear structure in time series data while integrating CNN and Gated Recurrent Unit (GRU) architectures to address complexity and nonlinear residual patterns in unstructured spatiotemporal datasets.

Keywords: Artificial intelligence, Hybrid model, Framework, Soybean, Climate variavility

1. Introduction

The productivity of soybean is dependent on climatic factors such as temperature [1] and [2], precipitation [3], humidity [4], and solar radiation [5] as expressed in the basic yield equation.

$$Y = f\,(T, P, H, S), \qquad (1)$$

where: Y = Soybean yield, T = Temperature, P = Precipitation, H = Humidity, S = Solar radiation.
Research indicates that temperature, precipitation rainfall patterns, and the oc- currence of extreme weather events have adverse effects on soybean growth and productivity [6]. Argentina, the third largest producer of soybeans (approxi- mately 51 million tonnes annually) after the USA and Brazil is dependent on rainfall as its main input in agriculture [7]. The authors of [8] in their study in- dicates that the country lost approximately 31.01% of soybean production during the last three severe droughts of 2009,

2012, and 2018. While assessing the im- pact of climate change on soybean production in Argentina, their study concluded that there was a powerful statistical correlation between rainfall, maximum tem- perature and soybean yields. Rainfall is a strong determinant of yields since 71 million acres of Argentina's farming land purely depend on rain-fed agriculture as indicated by [9]. According to National Bureau of Statistics of China (NBSC) as seen in [10], North- east China is the main producer of soybean in the country, accounting for 45% of national soybean production. The regions is mostly affected by climate change characterized by increase in temperature by 0.3 degrees per decade. The authors of [11] in their research pointed out that fluctuations in temperature and rainfall leads to yield losses. The study highlighted the need for predictive models that can account for these dynamic climate factors. In [12] the study showed how soybean phenology responds to climate warming in northeastern provinces of China, explicitly outlined that temperature is detrimental to the duration of growth of soybean but rainfall's effect on growth duration is positive. In [13], they studied the impact of climate variability on soybean yields in Pampas (central-east Argentina) and noted that minimum temperature has positive impact during seed germination and growth stage. On the contrary, high temperature and excessive rainfall during the period of maturity and harvest have adverse negative impact on the crop. Classical and machine learning models have been widely applied to modern agricultural practices to determine how crop yields are related to climate factors. We outline some of these models in this section. In [14], the authors applied the Pearson's correlation coefficient and regression analysis in their effort to study the relationship between climate variability and maize, millet barley, rice and wheat yields in Lamjung District, Nepal. The computation of a multivariate regression analysis shows the climate anomalies and yield anomalies expressed as;

$$Y = \beta_0 + \beta_1 P + \beta_2 T_{min} + \beta_3 T_{max}.$$

$$\Delta Y = \beta_0 + \beta_1 \Delta P + \beta_2 \Delta T_{min} + \beta_3 \Delta T_{max}. \qquad (2)$$

where, ΔY is the noted change in crop yield; ΔP, ΔT_{min} and ΔT_{max} are the observed changes in precipitation, minimum temperature, and maximum temperature and; $\beta_1, \beta_2, \beta_3$ are their coefficients respectively. β_0 is the non-climate factors. To test the correlation in yield-predictor variables, a correlation coefficient can be used and was applied and the regression results showed low significant relationship between yield and climate variables, in contrast, the non-climatic factors influenced the yield to a significant level. However, the model did not consider all the non- climatic factors. According to [15], non-climatic factors like labor force and financial development significantly and positively affects the cereal yield. We therefore engage fixed effect model to help regulate the unobserved factors. The work of [16] aimed to quantify future environmental implications on cotton and soybean yields in the Southeast United States, applied the fixed-effect model to the study of the effect of climatic variables; daily maximum temperature T_{max}, minimum temperatures T_{min}, and rainfall on cotton and soybean yields from the year 1980 to 2020. The fixed effect model was proposed to control the unobserved variable influencers which had a significant impact in the multiple regression model of the form $Y_{it} = S_i + T_t + \beta X_{it} + \varepsilon_{it}$, where Y_{it} represents the crop yields realized for individual $i = 1, 2,.., N$ at time $t = 1, 2,.., T$, S is the fixed place effects, accounting for all unobserved state-related factors which vary with time and influence crop yields, T is the fixed time effects accounting for human related factors, and X represents the climate variable. β is the parameter associated with the explanatory variable and ε the random error term. The aim is to estimate β the effect on Y_i with varying X_i, when both S_i and T_i are kept constant. To estimate the relationship between dependent and independent variables, Ordinary Least Squares (OLS) is applied.

With minimized sum of squares, the parameter β indicates that the estimated crop yield is mainly dependent on climate variable X_{it}, consisting of rainfall, T_{max} and T_{min} as seen also in [17] and the references therein. The effect of tem- perature and rainfall on the crop yield is non-linear, therefore the squared values of T_{min}, T_{max} and rainfall are computed in the multi-faceted hybrid model alongside climate variables.

The fixed effect model assumes that the unobserved variables are time-invariant and they are not in any way correlated to the independent variables. Oth- erwise if correlation exists, the results lead to bias in estimates as observed in [18]. This limitation conforms with the findings from the study in [19] which states that fixed effect model accounts for possible bias for time-invariant variables by elimination of all the variations between the variables in estimation. In our study, we applied the Bayesian approach to address the bias due to time- invariant variables.

The work of [20] was concerned with predicting yields of maize soybean and other crops and employed two-dimensional Gaussian function of temperature and pre- cipitation with monthly contribution accumulated over the growing period. The model, with d data points for each year being fitted is as follows;

$$X_d = (Y_d, T_d, P_d) \qquad (3)$$

T_d and P_d are vectors with temperature and precipitation values respectively for each month of the year. This paper takes the following structure: 1. Introduction, 2. Literature Review, 3. Method 4. Results and Discussion, and 5. Conclusion.

2. Literature review

ARIMA model is limited to univariate time series data with one variable measured over time and do not work well with spatial dependencies. Machine learning models were applied as a remedy for inadequacies seen in classical statistical techniques to handle high-dimensional data, capture non-linear relationships, and adapt to complex environmental variables. Machine learning has rapidly developed in the agricultural field, specifically in crop classification, growth monitoring, and crop yield prediction, compared to conventional statistics-based models. The authors of [21] in their study to predict winter wheat yield in Pakistan, employed linear (Random Forest, RF, and Support Vector Machine, SVM) and nonlinear (LASSO and Ridge) machine learning models based on climatological and remote sensing data. Five remote sensing indices (GNDVI, NDVI, EVI, SAVI, and ARVI) were integrated with five climate variables (Tmax, Tmin, Soil moisture, rainfall, and wind speed) alongside the drought index and standardized precipitation evapotranspiration index (SPEI) across the two wheat-season scenarios: full-season mean (FSM) and peak-season mean

(PSM). In both scenarios, RF outperformed other models, achieving the lowest RMSEs, with R2=0.75 and R2=0.78, respectively, suggesting a strong relationship between variables. The linear LASSO also performed fairly well, registering R2=0.74-0.69. Though RF outperformed the other models, it was deduced that it cannot predict beyond the range of training data, and it may change significantly with a small change in the data. The authors of [22] in their comparative analysis of statistical and machine learning for rice yield forecasting in three districts, Raipur, Surguja and Bastar in India, applied five techniques: stepwise multiple linear regression (SMLR), an artificial neural network (ANN), LASSO, an elastic net (ELNET), and ridge regression, to discover the best model for rice yield prediction. ANN performed well in Raipur and Surguja, producing validated R2=1 and R2=0.99, respectively. The main drawback of ANN is the memorization of training data rather than generalizing, and a greater computational burden. Generally, classical machine learning techniques discussed above perform poorly on high-dimensional unstructured datasets. Deep Learning models, an advanced version of machine learning, are preferred over classical machine learning for their capability to handle spatiotemporal data dependencies and extract important features based on data training. Its limitation lies in the longer time taken for data training. According to [23], convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) are the two dominant deep learning architectures applied in Agriculture, especially in the prediction of crop yield. CNN has been used in extracting satellite imagery of crops to help identify crop health and conditions. Authors of [24] in their research on deep learning in agriculture applied CNN to identify plant diseases through the extraction of visible leaf images. LSTM is essential in handling sequential data effectively and has been applied extensively in the field of agriculture to predict future crop yield based on historical data. Gated Recurrent Unit (GRU), a variant of RNN reduces computational complexity and simplifies the structure of LSTM when handling sequential data. Unlike the LSTM, it has a gated mechanism to input or forget certain features with no output gate. In [25], the research considered trained DNN, 1D-CNN and GRU using time series gross primary production (GPP) data for corn and soy beans separately. GRU performed second best after 1D-CNN in yield prediction. According to [26], hybrid deep learning models have emerged as powerful tools in predicting crop yields, demonstrating a significant improvement in performance over traditional statisti- cal methods and single machine learning models. When integrated, CNNs extract satellite images, while LSTMs do temporal analysis of the extracted features. The work of [27] utilized a hybrid CNN-LSTM model in the study of soybean yield prediction in CONUS at the county level. The results of the experiment indicated that the hybrid CNN-LSTM model performed better compared to pure CNN or LSTM models in both levels of the season. In [28] the authors developed and applied a CNN-GRU hybrid model to estimate the county-level wheat yield in the Gaunzhong Plain, China. Though the CNN-GRU hybrid produced better accuracy compared to the single GRU model, when working with limited data availability, it can reduce the model's generalization ability. The ARIMA model is preferred in statistical forecasting due to its robustness, interpretability, and ability to transform non-stationary series into stationary series, eventually outlining the significance of different variables and their underlying relationships. On the other hand, the CNN-GRU model can capture complex nonlinear relationships in data and deal with long-term dependencies and trends in crop productivity data. Leveraging the strengths of each model, therefore, in this study, we develop and apply a novel ACG-hybrid model to analyze the effect of climate factors on soy-bean production. This is a robust model well-suited to handle the nonlinearities, temporal dependencies, and climatic drivers that may influence soybean productivity.

This section gives a summary of research gaps for the formulation and analysis of an ACG-Hybrid model for assessing the effects of climate variability on soybean productivity. It is noted that while individual models ARIMA, CNN, GRU show great potential for complex spatiotemporal forecasting, their particular ap- plication to estimate quantity for the isolated and interactive effects of climate variables on soybean productivity remains underdeveloped. It has been realized that most research activities done focus on prediction accuracy rather than explanatory analysis of climate impacts. The specific gaps are therefore summarized as below:

1. **Hybrid Models Interpretation:** Hybrid DL models that is CNN and GRU are often referred to as black boxes. This means that there is a lack of stud- ies that integrate explainable AI techniques like SHAP, LIME with ACG-architectures to explicitly attribute soybean yield variations to specific cli- mate drivers for instance heat stress during flowering, drought in pod-filling. Therefore, there is need for research that moves beyond correlation to causal inference, using a multi-faceted model to identify critical climate thresholds and windows of sensitivity for soybean.
2. **Integration of multi-scale and heterogeneous data:** Current models often use uniform meteorological data therefore there is a gap that exists in effectively fusing high-resolution, multi-source data like satellite imagery - CNNs for spatial features, daily weather station time series - ARIMA/GRU, soil mois- ture sensors and management practices into a cohesive ACG-integration. Hence, there is need for a multi-faceted hybrid model to harmonize temporal (ARIMA/GRU) and spatial (CNN) feature extraction from disparate data sources with different scales and resolutions.

Handling extreme climate events and non-Linearities. The existing models are trained on historical norms and may fail to accurately capture the impact of compound extreme events for instance concurrent drought and heatwave or non-linear, tipping-point responses of soybean to climate stress. Therefore, there is need for hybrid architectures specifically designed to detect and forecast the yield impact of extreme, low-probability climate sequences.

1. **Spatial transferability and generalization:** The existing models trained on one region often perform poorly in new geographies with different soils, cultivars, or climate regimes. There is limited research on transfer learning or meta- learning approaches within the ACG paradigm to create adaptable models for data-scarce regions. Therefore, a formulation that allows the model's learned spatial (CNN) and temporal (GRU) features to generalize across different agricultural zones is very crucial.

2. **Incorporation of agronomic processes dynamics:** Pure data-driven hybrids may violate biophysical principles. A significant gap is the tight coupling of process-based crop models with the statistical/DL model, allowing the hybrid to be constrained by known plant physiology. So, there is need to formulate a hybrid physically informed or process-guided model where the CNN-GRU learns residuals or corrects biases of a mechanistic model, with ARIMA handling stochastic noise.
3. **Optimization of hybrid architecture and data flow:** The optimal fusion strat- egy for ARIMA, CNN, and GRU components is not established. Questions remain: Should ARIMA pre-process residuals for CNN-GRU? Should CNNs and GRUs run in parallel or series? How is long-term trend (ARIMA) best separated from short-term spatial patterns (CNN) and sequential dependencies (GRU)? This calls for comparative studies on different hybrid schematics and their efficacy for capturing climate-yield relationships.
4. **Focus on non-yield productivity metrics:** Most existing models have overwhelming focus on final yield prediction. Gaps exist in using such models to assess climate effects on other productivity aspects like protein/oil con- tent, seed quality, or water-use efficiency, which are crucial for market value and resilience. Therefore, there is need to come up with multi-task learning models that predict both yield and quality metrics from climate inputs.

The primary gap is not merely building an ACG-Hybrid model for soybean yield forecasting, but rather engineering and interpreting it as a comprehensive diagnostic tool for climate impact assessment. Future research should focus on enhancing model interpretability, integrating diverse data sources, improving generalizabil- ity, and embedding agricultural knowledge to transform the hybrid model from a powerful predictor into a credible, actionable tool for climate adaptation planning in soybean production and also for policy making information.

3. Methodology

In this section, we review the basic concepts that are key to our study employing a multi-faceted hybrid approach to the study of the effects of climate variability on soybean yield.

Definition 1 [6]: ARIMA is a statistical model that has AutoRegressive component of order p; Moving Average of order q and an Integrating part (I) of order d. It is best applied on time series data. The autoregression and moving average components are used to describe the stochastic process.

Definition 2 [8]: Recurrent Neural Network (RNN) is a type of machine learning model designed to analyze temporal and time series data.

Definition 3 [15]: Convolutional Neural Network (CNN) is a type of deep learning model used primarily for extraction and recognition of image from the spatial data.

Definition 4 [22]: Long Short-Term Memory (LSTM) is a family of recurrent neural network designed to handle long term dependencies in sequential data. Consists of memory cell structure controlled by the three gates; the input, the forget and the output gate.

Defnition 5 [14]: Gated Recurrent Unit (GRU) is a type of RNN designed to handle long term dependencies in sequential data. It differs with LSTM in terms of architecture, with fewer parameters and controlled with two gates; updated and reset gate.

4. Results and Discussion

Model formulation requires rigorous mathematical formulation using mathematical skills and statistical techniques while incorporating deep learning techniques also as seen in the earlier sections. At this point we give the ACG-Hybrid model formulation summary in the table below.

Table 1: ACG-Hybrid model integration components

Component	Purpose
ARIMA	Capture linear trends and climate effects
CNN	Extract spatial-temporal features
GRU	Model long-term dependencies
ACG-Hybrid	Combine linear and nonlinear components

Table 1 shows the components of AG-hybrid model. Since the development of a new model requires mathematical rigour, it is important to consider certain mathematical properties of the developed model. The first case is the bias-variance decomposition. The main idea is that the total error of a model can be mathematically decomposed into three interpretable components, that is, *Total Error* = *Bias*^{2} + *Variance* + *Irreducible Error*. This helps a lot in understanding the source of prediction error, whereby the expected prediction error is hence decomposed as seen in the equation below.

$$E[(Y_t - \hat{Y}t)^{2]} = \text{Bias}^2(\hat{Y}t) + \text{Var}(\hat{Y}t) + \sigma \quad (4)$$

The work of [12] utilized the ARIMA, Exponential Smoothing (ETS), and NNAR to model data; forecasting area, production, and productivity of wheat, paddy, maize, jowar, and cotton crops for the period 1980-2020. The best model for both training and testing was selected based on computation of metrics; mean absolute error (MAE), mean absolute percentage error (MAPE), mean absolute scale error (MASE), and root mean square error (RMSE). The results indicated that ARIMA (0, 1, 0) was the best for wheat growing area and ARIMA (0, 1, 1) was the best for wheat productivity, while ETS (M, A, N) was the best for wheat production.

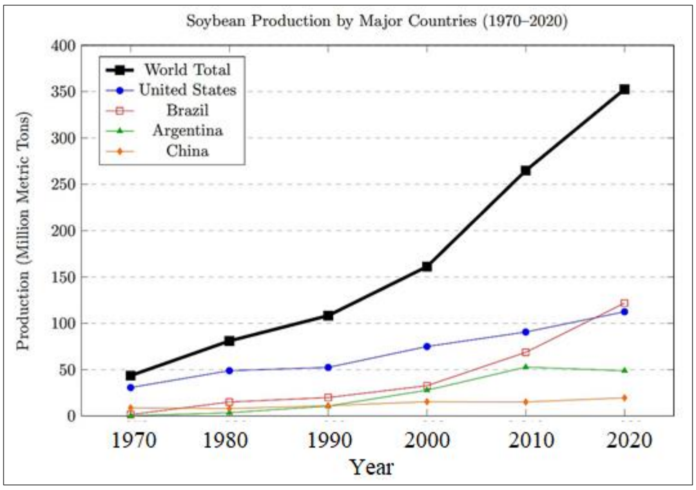

Fig 1: Line chart showing soybean production growth of major producing countries from 1970 to 2020.

From Figure 1, the United States and Brazil have shown stable soybean production from 1970 to 2020. China, ranked fourth in production, is also a heavy consumer, relying on imports. Generally, the four countries are major soybean producers, contributing significantly to global production.

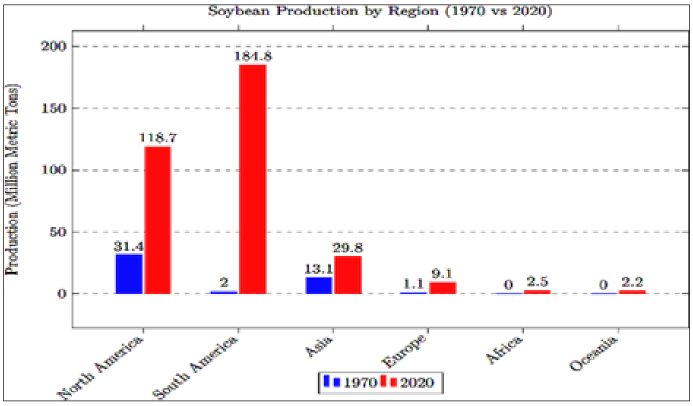

Fig 2: Comparative bar chart showing regional soybean production in 1970 and 2020.

From Figure 2, South America has experienced an upsurge of soybean production from the year 1970 to 2020, and is currently the leading soybean producer globally, followed closely by North America. Asia comes distant third. Africa and Oceania are the least contributors to global production, and therefore, they heavily rely on imports for their domestic consumption.

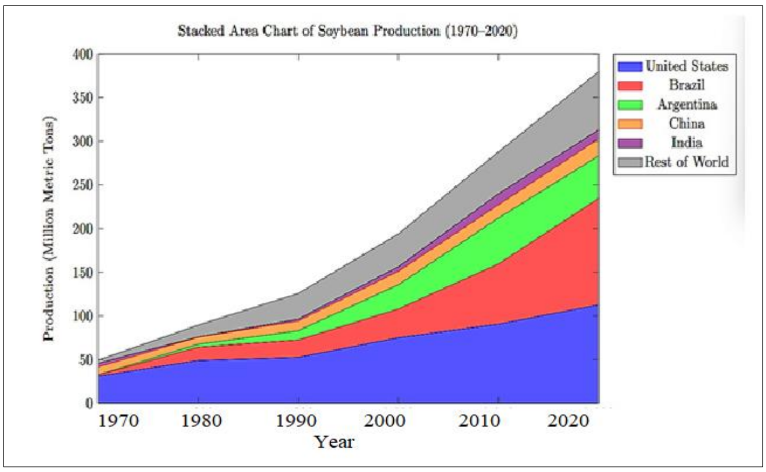

Fig 3: Stacked area chart showing the contribution of major countries to the total soybean production over time.

From Figure 2, it is clear that over the years the United States has offered stable soybean production. In 1970, Brazil's yield was very low; from 1990 onwards, it experienced an upsurge, surpassing the United States in 2020. Equally, Argentina's yield production has increased over the years due to its advanced agricultural techniques.

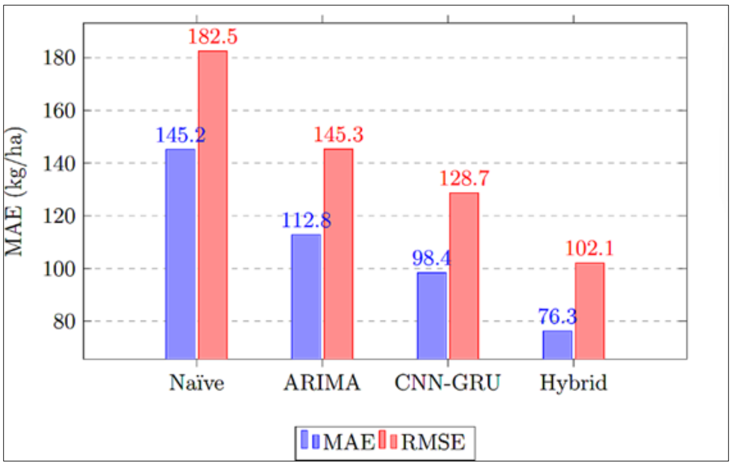

Fig 4: Forecast accuracy comparison of different models

From Figure 4, the baseline model has lower accuracy, especially in volatile spatiotemporal conditions. The hybrid ACG model performed better than ARIMA and CNN-GRU models by 32.4% and 22.5% in MAE, respectively and 29.7% and 20.7% in RMSE, respectively.

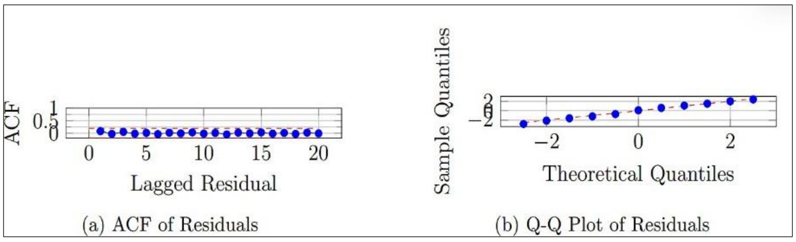

Fig 5: Residual diagnostic plots

Figure 5(a) shows that lagged residuals are white noise and that there is no autocorrelation in soybean yield data. Figure 5(b) shows that residuals are normally distributed.

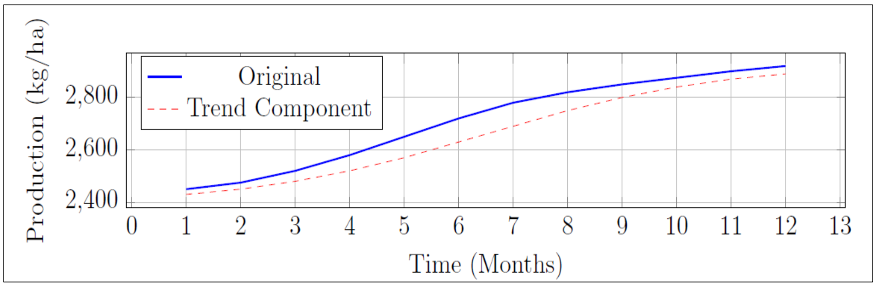

Fig 6: Time series decomposition of soybean production

Long-term trend shows on average a rising soybean production curve due to technological advancements and agricultural management improvement globally as seen in Figure 6.

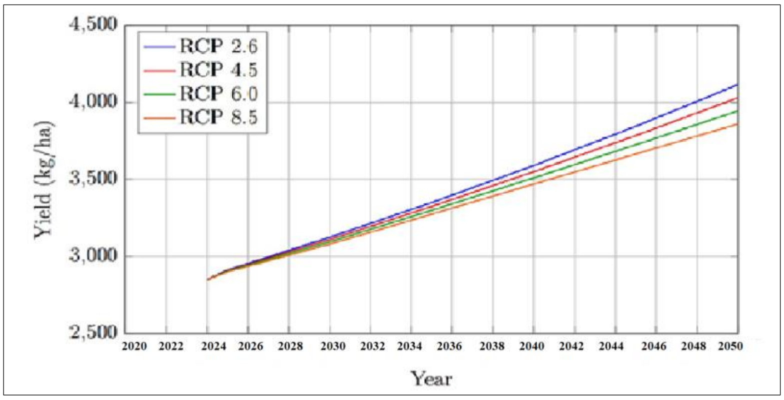

Fig 7: Global soybean yield projections under different climate scenarios (2024-2050)

Figure 7 shows that under RCP 2.6 with stringent mitigation measures, the 2050 projected soybean productivity is higher than that of any other representative concentration pathway. The climate scenario RCP 8.5 assumes higher emissions and a significant temperature increase without adaptation, leading to greater crop yield losses and the lowest projected yields. RCP 4.5 demonstrates moderate emissions.

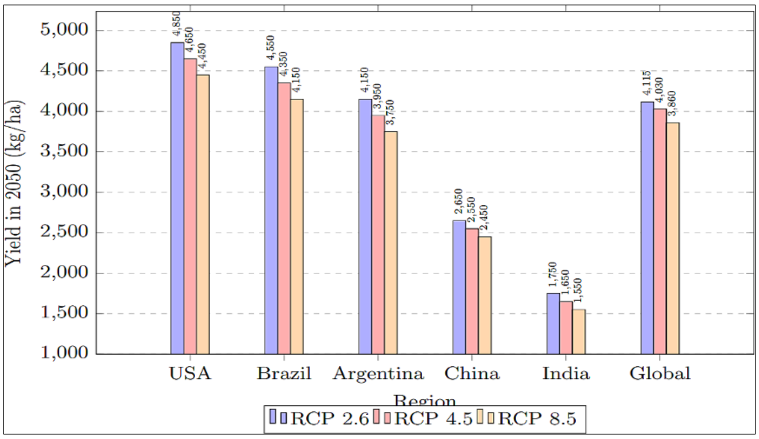

Fig 8: Regional yield projections for 2050 under different climate scenarios

Figure 8 shows that under the RCP 2.6 climate scenario, the 2050 projected yield productivity is highest in all major soybean-producing countries. High temperatures in RCP 8.5 severely affect yield productivity, projecting lower values. Globally, RCP 2.6 with lower greenhouse gas emissions is the most conducive climate scenario for soybean production, while RCP is the intermediate condition with moderate emissions.

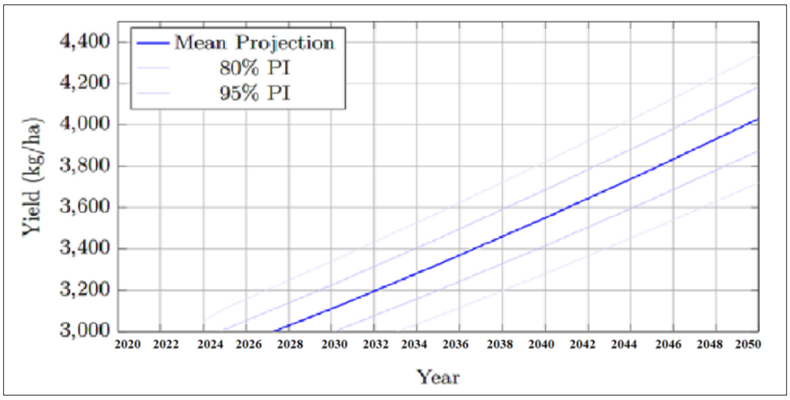

Fig 9: Prediction intervals for global soybean yield (RCP 4.5 Scenario)

From Figure 9, the RCP 4.5 experiences moderate emissions of greenhouse gas and a moderate decline in yield productivity. Under stabilized conditions with a slight increase in temperature neutralized by increased CO_2 fertilization effect, the mean projected value is realized. The 80% projection interval offers more precise estimates to the actual yield, while the 95% PI offers a wide range within which the actual projected yield is expected to fall, ensuring a high confidence interval.

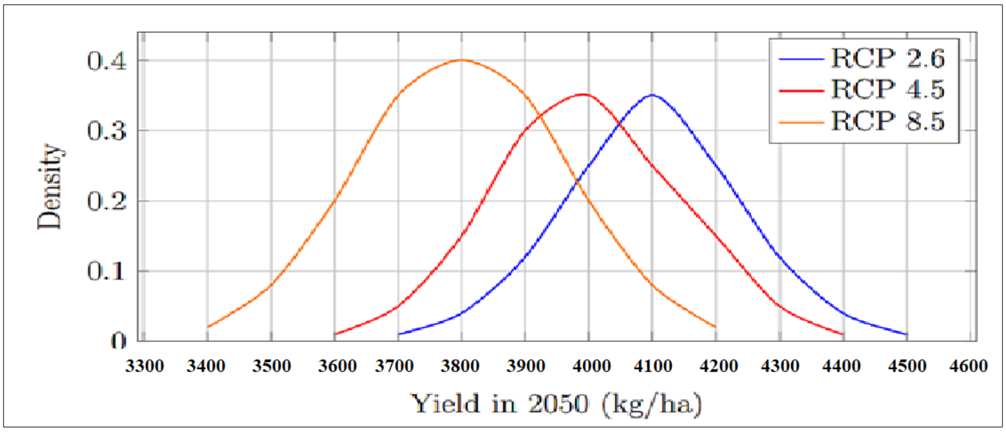

Fig 10: Probability distributions of 2050 yield projections from Monte Carlo simulations

From Figure 10, the projected soybean yield production under different RCPs takes on a normal curve, with varied means. The projected yield productivity under RCP 2.6, with lower emissions and lower plant density, is the best globally, with a projected mean average value of 4100kg/ha in 2050. The RCP 8.5 condition with high plant density is most severe to crop production, resulting in yield losses with an average projected mean value of 3800kg/ha. The RCP 4.5 condition, averaged at 4000 t/ha productivity, is the most stabilized and ideal condition with moderate increase in temperature, with lowered precipitation and increased atmospheric CO2.

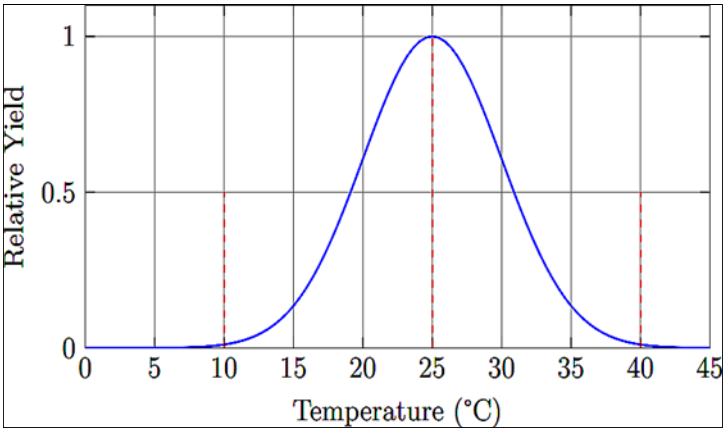

Fig 11: Temperature impact function on soybean yield

The maximum potential soybean yield is realized with an optimal temperature threshold of 25^0C as seen in Figure 11. Increased temperature results in yield losses. Extreme heat >35^0C results in drastically reduced yield production.

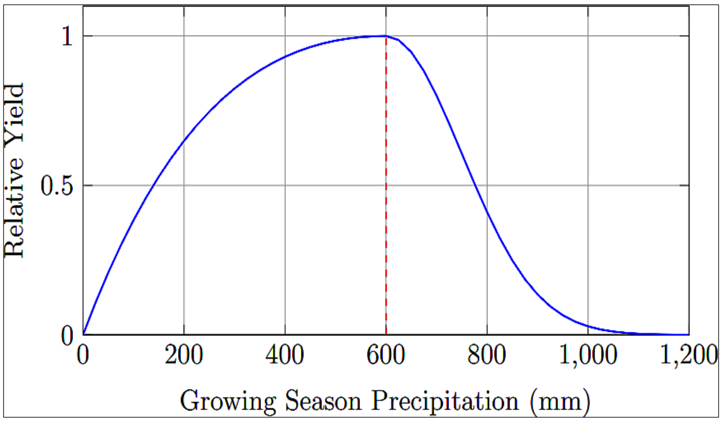

Fig 12: Precipitation impact function on soybean yield

Soybean yield production increases gradually with an increase in precipitation up to an optimal value of 600mm as seen in Figure 12. Heavy rainfall above 600mm results into declined production and may lead to heavy losses.

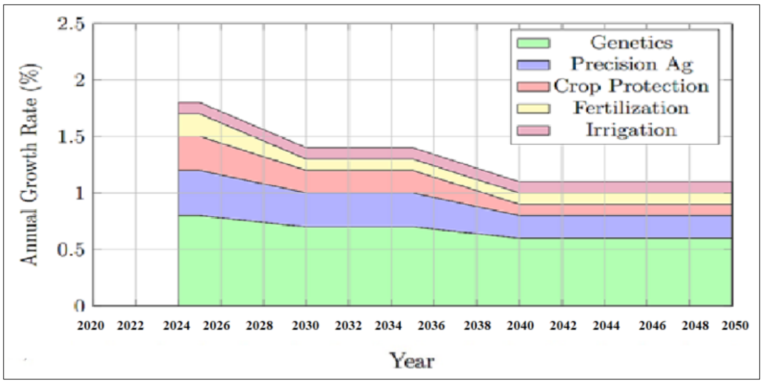

Fig 13: Decomposition of technological growth contributions

Figure 13 shows that technological practices have a significant influence on soybean productivity. Genetic improvement through breeding programs enhances yields, and therefore, it is the most significant of all other factors throughout crop production and projection. Precision agriculture through digital adoption also has a significant impact on yields. Crop protection, fertilization and irrigation all impact positively on soybean yield production.

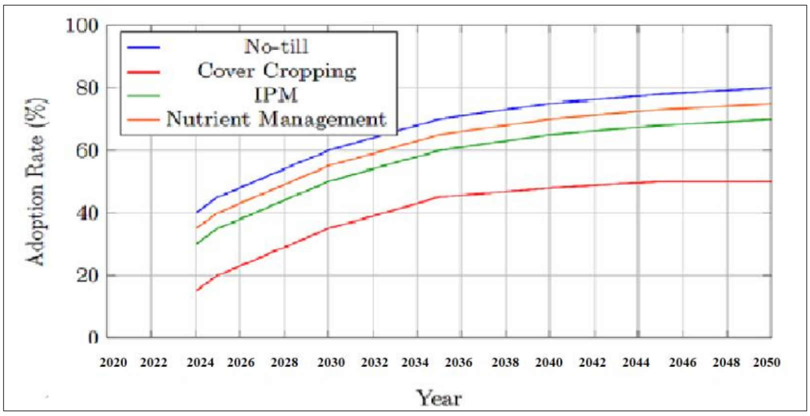

Fig 14: Adoption rates of key management practices (2024-2050)

Management practices have a positive influence on the soybean yield output as seen in Figure 14. No-till has the highest potential yield increase due to improved soil structure and increased moisture retention. Nutrient management, Integrated pest management and cover cropping all have significant impact on yield production and projection.

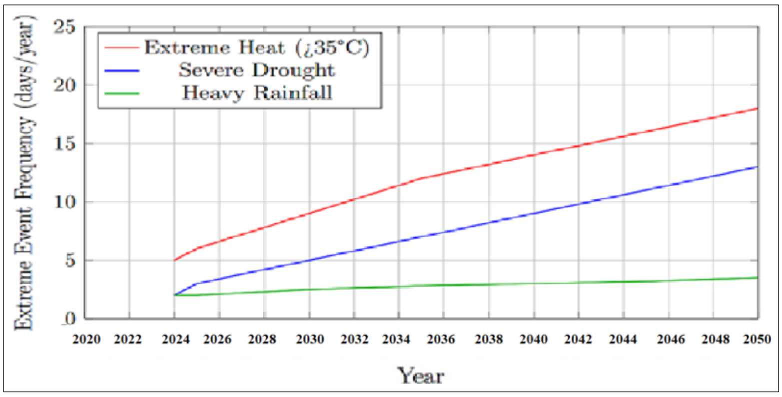

Fig 15: Projected increase in extreme event frequency (2024-2050)

The frequency of occurrence of extreme events increases with time. From Figure 15, we note that the increase factors of heavy rainfall, severe drought and extreme heat are 3.0x, 2.0x and 1.5x, respectively.

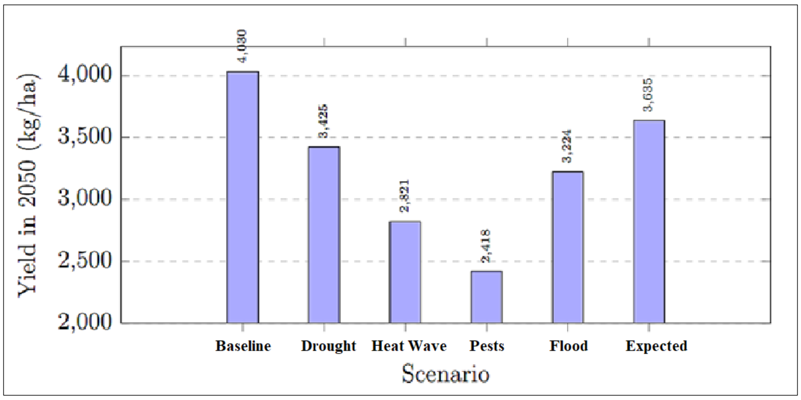

Fig 16: Risk-adjusted yield projections for 2050 (RCP 4.5)

The projected soybean yield in 2050 is likely to be affected by extreme events, resulting in yield losses as seen in Figure 16. Pest outbreaks cause severe damage to the crop, resulting in the lowest yield, followed by heat waves and floods.

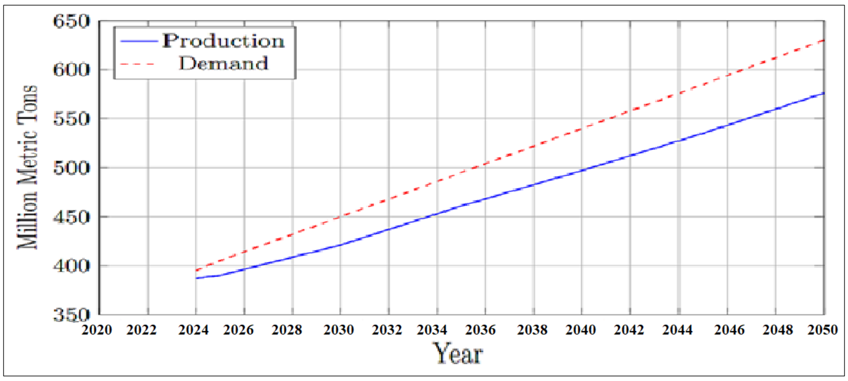

Fig 17: Global soybean production vs demand projections (2024-2050)

Globally, the amount of soybean yield produced in million metric tons is way below the demand as seen in Figure 17. To meet the demand gap, the acreage under soybean farming needs to be increased, especially in Africa and Asia.

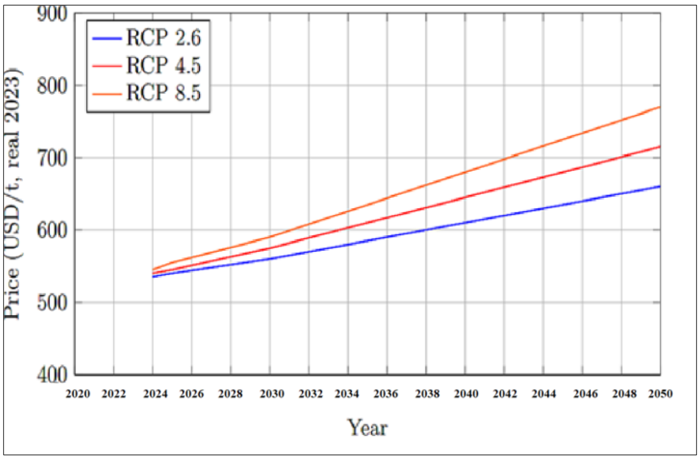

Fig 18: Soybean price projections under different climate scenarios

The price of soybeans is determined by the amount of production as seen in Figure 18. The RCP 2.6 with stringent mitigation measures realizes bumper harvests, resulting in a drop-in price. Under business as usual, RCP 8.5 with high emissions, yield losses are very high and therefore higher prices due to low supply.

As much as RMSE indicates better predictive accuracy in data, [11] pointed out that it is prone to errors in estimation since it is highly affected by outliers. In this work, we envisage to utilize sensitivity and specificity as significant metrics in evaluating the performance of the multifaceted ACG-hybrid model that we intend to develop.

In an attempt to forecast the Indonesia export, from January 1998 to Decem- ber 2019, [7] merged two commonly used time series algorithms (ARIMA and LSTM) and developed a hybrid ARIMA-LSTM model. Having understood that each model has its own strength and weaknesses, time series data was decomposed into trend, seasonality and residuals, and ARIMA model applied to the trend com- ponent due to its linear while LSTM applied on seasonality and residual parts. The developed hybrid (ARIMA(0, 1, 2), LSTM 4000epochs-LSTM 2500epochs) model performed optimally well by registering lowest MAPE and RMSE of as compared to individual SARIMAX and LSTM-1500epochs.

A number of classical statistical techniques have been applied in predicting crop yield production. The study of [23] applied ARIMA(0,0,1) model to predict soybean productivity. The results indicated that by 2028, the average soybean yield will raise due to more positive effect of global warming. Ensemble learning methods like random forests have equally been used in crop yield prediction.

1. Activation saturation given by Percentage($|a| > 0.9$) < 20% for sigmoid/tanh.
2. Dead neuron detection given by Percentage($|a| < 0.01$) < 5% for ReLU.
3. Weight distribution where weight histogram should be approximately normal with mean near 0.

The time series cross-validation procedures therefore can be given via Algorithm 1 as shown below.

Algorithm 1: Time series rolling window cross-validation

Step	Procedure
1	Initialize k = 5 folds
2	For *i* = 1 to *k* do
3	Training set: [1, T − (k − i + 1) × h]
4	Validation set: [T − (k − i + 1) × h + 1, T − (k − i) × h]
5	Test set: [T − (k − i) × h + 1, T − (k − i − 1) × h]
6	Train the model on the training set
7	Validate the model on the validation set
8	Compute performance metrics on the test set
9	Compute the average metrics across all folds

The algorithm for blocked cross-validation is also given as is given in Alogrith 2 as shown below:

Algorithm 2: Blocked cross-validation for time series

Step	Procedure
1	Divide the time series into B blocks, each of length L
2	For *b* = 1 to *B* do
3	Hold out block *b* as the test set
4	Use blocks 1, …, b − 1, b + 1, …, B as the training set
5	Allow gaps between the training and test sets to prevent data leakage

The diagnostic validation which occurs in Phase 3 of hybrid model validation as a requirement has been done. We have carried out stability analysis whereby recursive estimation has been done together with CUSUM tests. We have also done structural break detection, that is, Chow and Bai-Perron tests. With regards to neural network diagnostics we carried out gradient analysis and activation checks. Lastly, cross-validation has been done to determine time-series specific procedures. The practicality of the model should also be validated.

In their research, the authors of [14] used random forests to predict crop yield based on climatic data. The model did not perform quite well especially in handling complex nonlinear relationships resulting in model over-fitting. The major undoing of classical statistical methods is its dependency on feature engineering which is a time-consuming process that need to be altered whenever the problem or the dataset changes.

From the above results on hypothesis testing, we draw the following interpreta- tions and give practical implications in this work. We begin with the statistical interpretations as follows:

1. **H_{01} Rejected:** The ACG-Hybrid model shows statistically significant dif- ferences from individual models ($p < 0.0001$).
2. **H_{02} Rejected:** The ACG-Hybrid model provides statistically significant improvements in all error metrics: RMSE reduction of 29.7% ($p < 0.0001$); MAE reduction of32.4% ($p < 0.0001$); and MAPE reduction of 32.4% ($p < 0.0001$).
3. **H_{03} Rejected:** All components contribute significantly: ARIMA: 28.4% of variance explained; CNN: 19.7% of variance explained; GRU: 24.2% of variance explained; and Interactions: 25.9% of variance explained.

When assessed practically, the significance assessment results are summarized in the table below.

Table 2: Practical significance assessment

Metric	Improvement	Effect Size	Practical Impact
MAE Reduction	36.5 kg/ha	Large (d=0.76)	Farm planning
MAPE Reduction	1.59%	Large (d=0.71)	Yield forecasting
Prediction Lead Time	+2 months	-	Earlier decisions
Uncertainty Reduction	32%	-	Risk management

It is clear from Table 2 that certain limitations are observed. Forecast errors are assumed independent, For normality, error distributions approximate normality.

In [3] the authors applied the hybrid approach of machine learning (SVM) and deep learning (LSTM,RNN) to predict crop yield production under both climate and soil parameters.The hybrid model outperformed the Artificial Neural Net- work(RNN) and Random Forest algorithms in predictive accuracy. According to [6] in their study to predict rice productivity based on LSTM model, compared the performance of Support Vector Regression (SVR) and LSTM and noted that LSTM had better predictive power than SVR model. GRU model simplifies the structure of LSTM with fewer parameters within its gated system. Research indicates that the hybrid CNN-LSTM model produced more reliable crop predictive results than pure CNN or LSTM models. Similarly, a study done on hybrid CNN-GRU produced better predictive accuracy compared to GRU model. The developed ACG-hybrid model shall be used to project future soybeans pro- ductivity based on historical data.

5. Conclusion

The Autoregressive Mov- ing Average (ARIMA) model has also been applied successfully for forecasting maize production in Tanzania, as noted in [80]. Nonetheless, ARIMA is inherently limited to univariate time series data and does not adequately address spatial dependencies. To overcome these limitations, machine learning approaches, par- ticularly deep learning techniques such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks, have been applied for spatiotem- poral data analysis. Though the CNN-LTSM hybrid model performed fairly well with spatiotemporal data, it failed to capture station- arity and handle the underlying linear structure in the model, leading to higher error metrics. Moreover, [92] developed an ARIMA-CNN-LSTM model to forecast carbon pricing, effectively capturing both linear and non-linear data characteris- tics, and achieving lower RMSE compared to benchmark models. However, a key drawback of the ARIMA-CNN-LSTM model is the lengthy training time due to the complexity of the LSTM component. In light of these changes, we propose the de- velopment of a novel hybrid model, termed the ACG-hybrid model, which seeks to bridge the gaps identified in conventional and existing hybrid deep learning mod- els. This model leverages the strengths of the ARIMA as a foundational model to handle the underlying linear structure in time series data while integrating CNN and Gated Recurrent Unit (GRU) architectures to address complexity and nonlin- ear residual patterns in unstructured spatiotemporal datasets. Our multi-faceted hybrid model aims to enhance the predictive accuracy of soybean productivity in Kenya in comparison to various regions worldwide, considering the impacts of climate variability more effectively.

References

1. Poincaré H. Analysis situs. J École Polytechnique. 1895;2:1-121. doi:10.1007/978-1-4614-2407-9_10.
2. Brouwer L. Beweis der Invarianz der Dimensionszahl. Math Ann. 1911;70:161-165. Available from: http://eudml.org/doc/158491
3. Baranochnikov I, Slepaczuk R. A comparison of LSTM and GRU architectures with the novel walk-forward approach to algorithmic investment strategy. IJISET. 2022;18:67-78.
4. Bauer E, Kohavi R. An empirical comparison of voting

classification algorithms: bagging, boosting, and variants. Mach Learn. 1999;36:105-139.
5. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157-166.
6. Beck HE, Atkinson PM, Huttich C, *et al*. Global evaluation of four AVHRR-NDVI data sets: inter-comparison and assessment against Landsat imagery. Remote Sens Environ. 2011;115(10):2547-2563.
7. Box GEP, Jenkins GM. Time series analysis: forecasting and control. New York: Holden-Day; 1976.
8. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
9. Burrell AL, Evans JP, De Kauwe MG. Detecting dryland degradation using time series segmentation and residual trend analysis. Remote Sens Environ. 2017;197:43-57.
10. Betti E. Sopra gli spazi di un numero qualunque di dimensioni. Ann Mat Pura Appl. 1871;4:140-158. doi:10.1007/BF02420029.
11. Alexander JW. A proof of the invariance of certain constants of analysis situs. Trans Am Math Soc. 1915;16(2):148-154. doi:10.2307/1988715.
12. Noether E. Ableitung der Elementarteilertheorie aus der Gruppentheorie. Jber Dtsch Math Ver. 1926;36:104-114.
13. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. doi:10.1038/nature14539.
15. Peter B. Statistical data analysis of hybrid models. J Mach Learn Res. 2015;16(1):77-102. doi:10.48550/arXiv.1207.6437.
16. Keith J, Sean M, Peter N, Alain C, Karl J, Alan C, *et al*. A unified statistical approach for determining significant climate variability factors for soybean productivity. Hum Brain Mapp. 1996;4(1):58-73.
17. Indicative Team. Topological data analysis [Internet]. Indicative by mParticle; 2023 [cited 2023 Aug 2[3]]. Available from: https://www.indicative.com/resource/topological-data-analysis/
18. Mehmet E, Esra A, Ahmed E. Data analysis of networks: methods and applications. Appl Netw Sci. 2019;4:61. doi:10.1007/s41109-019-0179-3.
19. Nganyewou T, Khomh F. Reliable future projection tools of crop production [Internet]. 2022. doi:10.48550/arXiv.2211.01535.
20. IndustriusCFO. 7 benefits of using big data [Internet]. IndustriusCFO; 2023 [cited 2023 Jul]. Available from: http://www.industriuscfo.com/7-benefits-using-big-data/
21. Lexico. Definition: Lexico powered by Oxford Dictionary [Internet]. Oxford University Press; 2023 [cited 2023 Aug]. Available from: http://www.lexico.com/definition/
22. Herbert E, John H. Computational statistics: an introduction. New York: Springer; 2010. doi:10.1007/978-3-540-33259-6_7.
23. Xia K, Li Z, Mu L. Multiscale persistent functions for biomolecular structure characterization. Bull Math Biol. 2018;80(1):1-31. doi:10.1007/s11538-017-0362-6.
24. Adcock A, Carlsson E, Carlsson G. The topology of data [Internet]. Homol Homotopy Appl. 2016;18(1):381-402. doi:10.48550/arXiv.1304.0530.
25. Chevyrev I, Nanda V, Oberhauser H. Persistence paths and signature methods for topological data analysis. IEEE Trans Pattern Anal Mach Intell. 2020;42(1):192-202. doi:10.48550/arXiv.1806.00381.
26. Okelo B, Onyango A. Persistent homology and artificial intelligence analysis of COVID-19 in topological spaces. DS J Digit Sci Technol. 2023;2(3):1-8. doi:10.59232/DST-V2I3P101.
27. Hajij M, Assiri B, Rosen P. Parallel mapper. In: Advances in intelligent systems and computing. Cham: Springer; 2020. p.717-731. doi:10.48550/arXiv.1712.03660.
28. Zobaa A, Bihl T. Security methods for critical infrastructure communications. In: Big data analytics in future power systems. Boca Raton: CRC Press; 2018. p.85-106. doi:10.1201/9781315105499.

How to Cite This Article

Odhiambo OE, Okuto E, Okelo B, Oyieke S. Analysis of multi-faceted ACG-hybrid model on effects of climate variability on soybean productivity using artificial intelligence techniques. Int J Appl Math Numer Res. 2026;2(4):50-61.