Impact of Climate Change on Crop Yield Prediction

XXXX

December 2, 2024

Abstract

Agriculture plays a vital role in ensuring global food security; however, climate change poses considerable challenges to agricultural productivity. The objective of this research is to forecast the effects of climate change on crop yields through the application of both statistical and process-based modeling approaches. The investigation examines the interactions among climatic factors, agricultural practices, and crop yields. It employs exploratory data analysis (EDA), machine learning algorithms, and time series forecasting methodologies. It encompasses variables on temperature, precipitation emissions, soil health, and many more across regions and types of crops. This is a dataset on temperature and precipitation, ascertaining that this is the prime driver in crop yield. Precisely, an integrated approach using Random Forest Regression hybridizes with XGBoost and GRU models in enhancing predictive precision for outcomes. These findings will be very important to policymakers, researchers, and agricultural practitioners as they put forward climate-resilient strategies to ensure food security under changing future climatic conditions.

1 INTRODUCTION

Agriculture is one of the most crucial ways to guarantee human existence. It is a source of food, jobs, and economic stability worldwide. However, climate change seriously affects the productivity of agriculture. Rising temperatures, changes in rainfall patterns, and increased frequency of extreme weather events are risking growing conditions and, thereby, food security. Knowing such impacts is very important for the future of agriculture.

The global demand for food is increasing. This trend is driven by the growth of population, urbanization, and a change in dietary choices. The Global Harvest Initiative states that 2017 in order for the demand of the crop to be met, then agricultural productivity must increase to at least 1.75

To address these questions, researchers have been focusing on the understanding of how climatic factors—that is, temperature, precipitation, and carbon dioxide levels—impact agricultural production. Three main approaches are followed: field experimentation, process-based modeling, and statistical deduction. All three have their relative strengths and weaknesses. The former, for example, provides experientially authentic observations but suffers from scale limitations and cost-related constraints.

Process-based models simulate crop growth but require large amounts of data. Statistical models look at the relationship between past climate data and crop yields. They provide quick predictions but are poor at extrapolation and mechanistic understanding. More recently, there have been hybrid approaches that incorporated both process-oriented and statistical frameworks, taking advantage of advances in machine learning and explainable artificial intelligence (Hu et al., 2023). Such methods can offer the opportunity to fuse heterogeneous data and offer researchers more advanced insights. For example, machine learning models can be very good at predicting agricultural yield but usually are opaque. Recently, there has been an increased development of physics-informed and explainable AI strategies (Zhang et al., 2023). This study builds upon these developments. It seeks to make spatial and temporal forecasts of climate change impacts on crop yields. In integrating statistical and process-based models, we hope to fill in the knowledge gaps that currently exist. The results will be of great importance to policymakers, scientists, and farmers alike. These will help in developing climate-resilient crops, informing agricultural policies, ensuring food security, and guaranteeing economic stability.

2 DATA

Dataset provides a detailed overview that how climate factors influence crop yields across different countries, regions, and years. It is designed to analyze the relationships between environmental variables, agricultural practices and crop production. he information would be used for predictive modeling and policy development. The key elements of the dataset are outlined below.

2.1 Target Variable: Crop Yield

The main outcome variable is **Crop_Yield_MT_per_HA**, which measures crop productivity in metric tons per hectare. The dataset includes data for various crops:

- **Staples**: Corn, Wheat.
- **Specialty Crops**: Coffee, Sugarcane.

The yield data covers multiple years, including both historical years (e.g., 1998, 2001) and recent years (e.g., 2024). The data is aggregated at both national and regional levels. Some regions include *West Bengal (India)*, *North China*, *Île-de-France (France)*, and *the Prairies (Canada).*

2.2 Predictors

The dataset contains both static and dynamic variables. These are grouped into three categories: climatic variables, agricultural practices, and socioeconomic factors.

2.2.1 Climatic Variables

Climatic data highlights the impact of weather and extreme events on crop yields. Key predictors include:

- **Average_Temperature_C**: The annual mean temperature (C), which reflects thermal conditions during the growing season.
- **Total_Precipitation_mm**: The total annual precipitation (in millimeters), which captures rainfall variability.

- **Extreme_Weather_Events**: The frequency of extreme events such as droughts, storms, and floods, indicating climatic risks.

2.2.2 Agricultural Practices

This category includes variables related to management practices that affect productivity:

- **Irrigation_Access_%**: The percentage of cropland with access to irrigation infrastructure.
- **Pesticide_Use_KG_per_HA**: The application rate of pesticides, indicating pest control efforts.
- **Fertilizer_Use_KG_per_HA**: The rate of fertilizer use, which serves as a proxy for nutrient management.

2.2.3 Soil and Environmental Quality

This category includes variables that reflect soil health and environmental conditions:

- **Soil_Health_Index**: A composite metric that assesses soil quality for crop growth.
- **CO2_Emissions_MT**: The annual carbon dioxide emissions (in metric tons), which indicate the environmental impact of agricultural activities.

3 METHODOLOGY AND RESULTS

3.1 Exploratory Data Analysis (EDA)

3.1.1 Data Loading and Cleaning

Loading the Dataset. We loaded the dataset. It provides detailed data on factors affecting agricultural performance under changing climate conditions. The dataset includes:

- **10,000 rows** Representing observations across years, regions, and crop types.
- **15 columns** Covering metrics such as temperature, precipitation, CO_2 emissions, crop yield, economic impact, and adaptation strategies.

Data Cleaning.

- **Categorical Column Optimization:** To improve processing efficiency, we converted these columns to categorical data types:
  - *Country*
  - *Region*
  - *Crop_Type*
  - *Adaptation_Strategies*
- **Handling Missing Values:** A review of the dataset confirmed no missing values. This ensured all rows and columns were fully populated.
- **Duplicate Rows:** The dataset was examined for duplicates, but none were found. This verified the uniqueness of the data.
- **Outlier Detection:** Key numerical metrics were analyzed for potential outliers. The findings were:
  - *Average Temperature (°C):* Ranged from -4.99 to 35. This range is realistic for global temperature data.
  - *Total Precipitation (mm):* Spanned 200.15 to 2999.67, consistent with climatic variations across regions.
  - *Economic Impact (Million USD):* Values ranged from $47.84M$ to $2346.47M$, reflecting economic diversity in agriculture.

Table 1: Correlation Matrix

	Crop_Yield	Average_T	Total_Prec	CO2_Emiss	Fertilizer	Irrigation_Access
Crop_Yield	1.00	0.26	0.03	-0.09	0.01	-0.00
Average_T	0.26	1.00	0.01	-0.00	-0.01	-0.01
Total_Prec	0.03	0.01	1.00	-0.01	-0.03	-0.01
CO2_Emiss	-0.09	-0.00	-0.01	1.00	-0.02	0.00
Fertilizer	0.01	-0.01	-0.03	-0.02	1.00	0.01
Irrigation_Access	-0.00	-0.01	-0.01	0.00	0.01	1.00

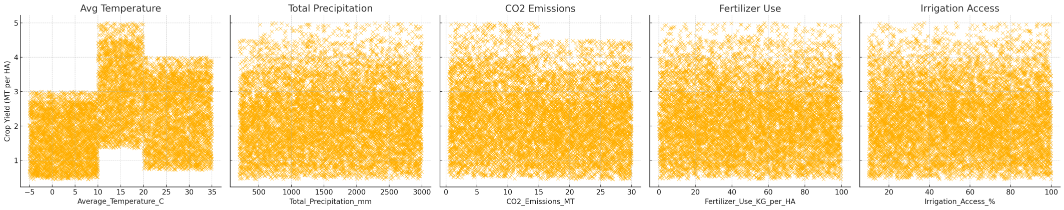

Figure 1: Summary: Relationship Between Crop Yield and Key

3.1.2 Correlation Analysis

Average temperature emerges as the most influential variable, underscoring the dominant role of climatic conditions in determining crop yield. Other variables, such as precipitation, CO_2 emissions, fertilizer use, and irrigation access, exhibit weak or negligible direct relationships with crop yield.

3.1.3 Data Visualization

1. Crop Yield Distribution Across Years

Visualization: A box plot displays crop yield distributions from 1990 to 2024.

Crop yields have remained relatively stable across years, with slight variations in the median. Notably, years such as 2013 and 2017 exhibited slightly higher median yields. Outliers in the distribution indicate extreme yield values in certain years, suggesting isolated instances of unusually high or low production.

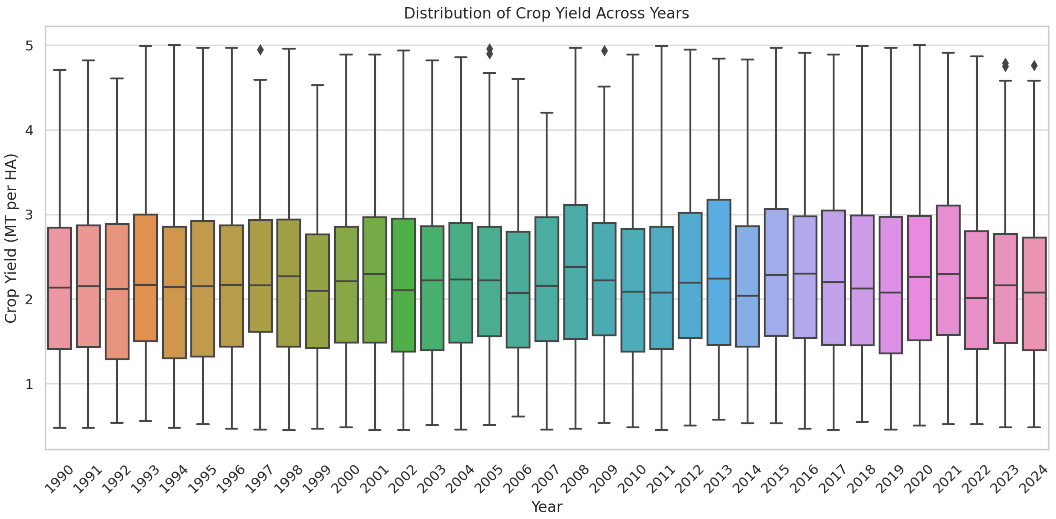

Figure 2: Distribution Of Crop Yield Across Years

2. Crop Yield Distribution Across Regions

Visualization: A box plot illustrates crop yield distribution across various regions.

Regions such as ”Patagonia” and ”Prairies” consistently show higher and more stable yields compared to others. In contrast, regions like ”British Columbia” and ”Volga” exhibit broader variability, reflecting diverse environmental conditions and agricultural practices. These regional differences underscore the significant influence of geographic and environmental factors on crop production.

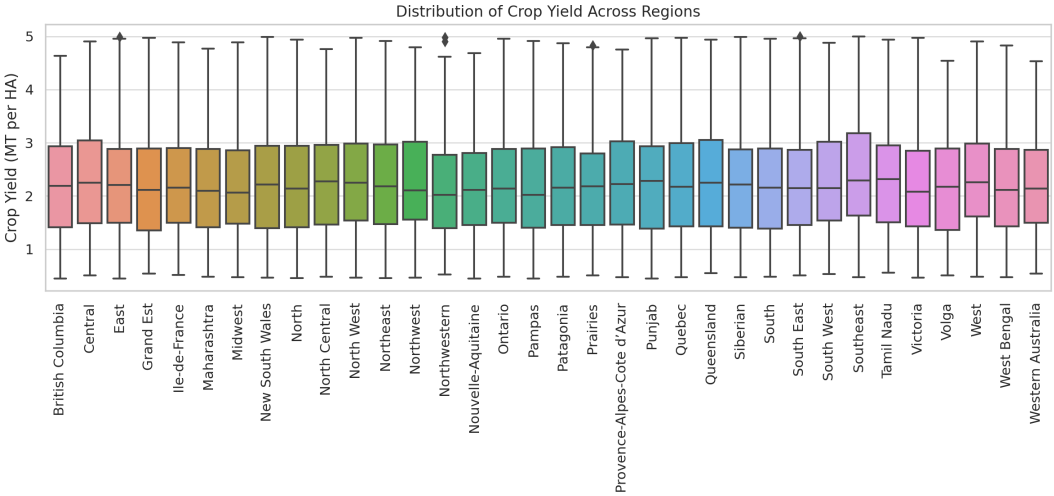

Figure 3: Distribution Of Crop Yield Across Regions

3. Crop Yield Distribution Across Crop Types

Visualization: A box plot highlights yield distributions among various crop types.

Crops such as ”Wheat” and ”Rice” display relatively higher median yields, indicating their robustness under current agricultural conditions. Conversely, crops like ”Barley” and ”Coffee” show moderate to low median yields but exhibit greater variability. This variability likely reflects differences in crop resilience, input requirements, and sensitivity to environmental conditions.

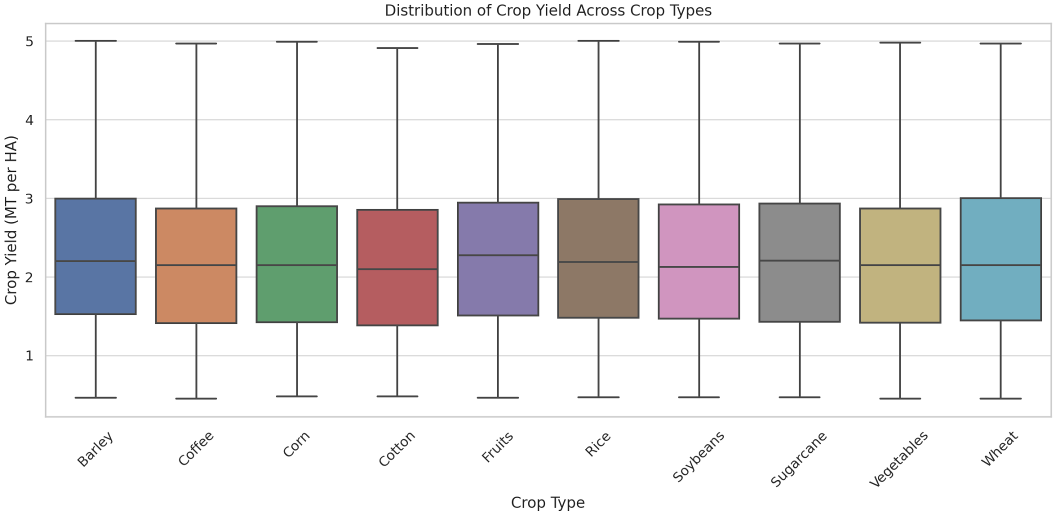

Figure 4: Distribution Of Crop Yield Across Crop Types

4. Average Temperature Trends Over the Years

Visualization: A line plot shows the trend of average temperatures from 1990 to 2024.

Temperature trends reveal significant fluctuations over the years, with notable peaks in the early 2000s and mid-2010s. A gradual upward trend is observed, indicating consistent warming over time. This pattern reflects the ongoing impact of climate change on global agricultural environments.

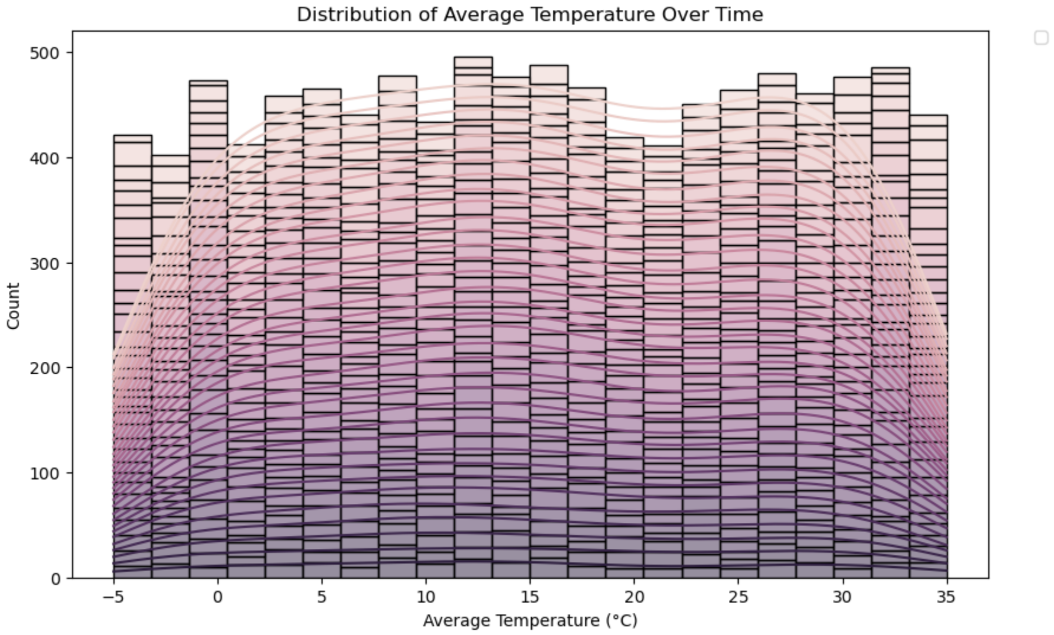

Figure 5: Distribution of Average Temperature Over Time

3.2 Model Implementation

3.2.1 Linear Regression

Linear Regression is used as a baseline model to establish a simple relationship between climate variables and crop yield. The model aims to predict crop yield one season ahead based on independent variables such as temperature, rainfall, and soil moisture.

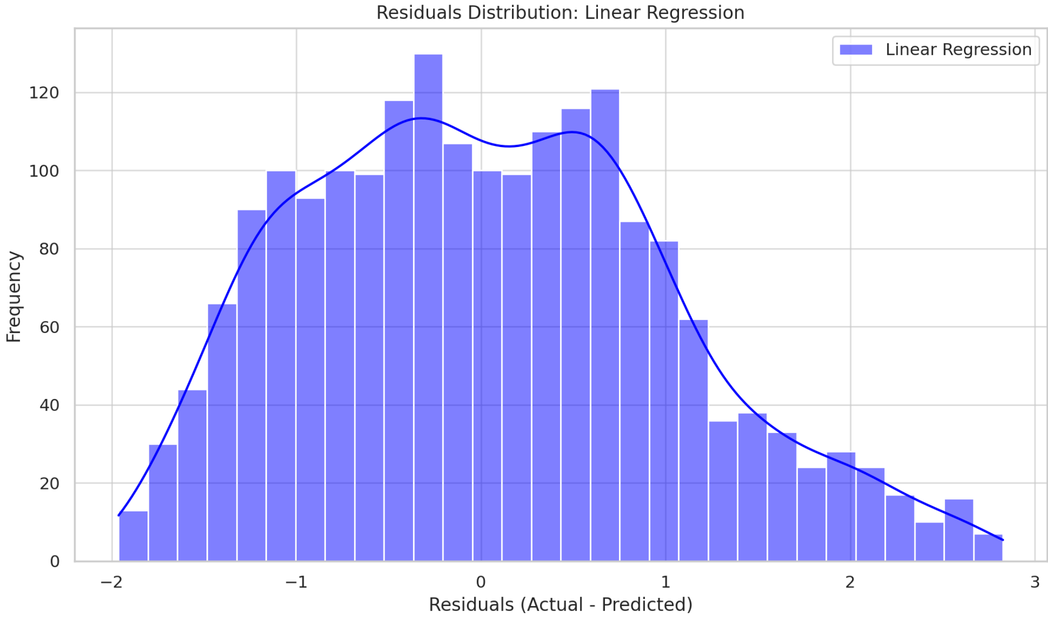

Figure 6: Residuals Distribution: Linear Regressio

Metrics Linear Regression:

- Mean Squared Error (MSE): 0.996
- Root Mean Squared Error (RMSE): 0.998

Residuals Analysis

- Residuals show a wider spread, indicating more frequent prediction errors.
- The distribution is roughly symmetric but less centered around zero.

3.2.2 Random Forest Regression

Random Forest Regression leverages an ensemble of decision trees to model complex, non-linear relationships between climate factors and crop yields.Random Forest captures non-linear interactions between variables and provides better performance when dealing with complex relationships.

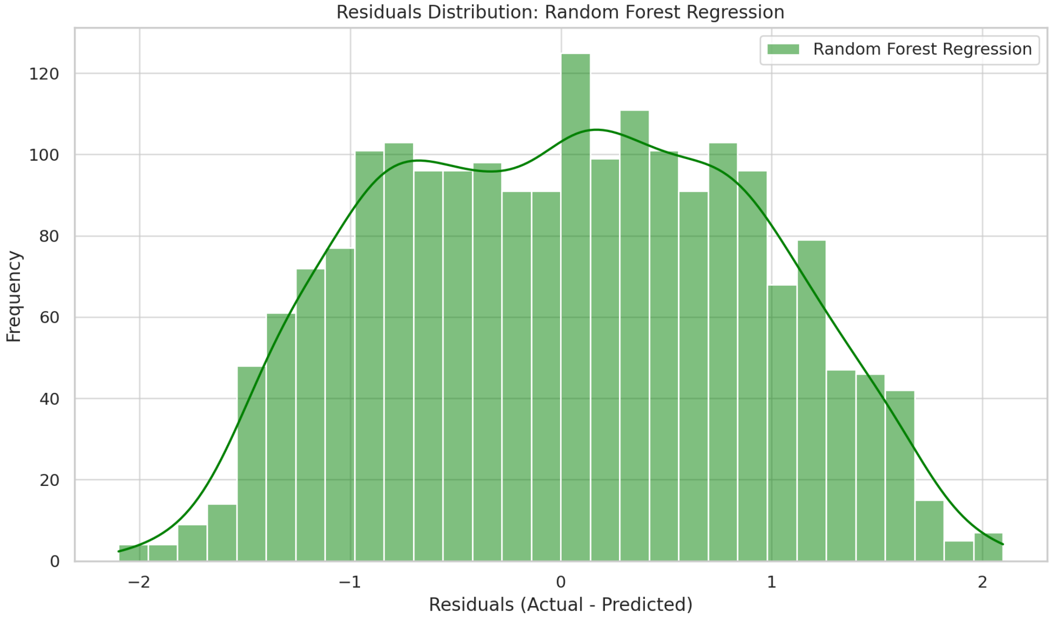

Figure 7: Residuals Distribution: Random Forest Regression

Metrics Random Forest Regression:

- Mean Squared Error (MSE): 0.750
- Root Mean Squared Error (RMSE): 0.866

Residuals Analysis

- Residuals are more tightly distributed, suggesting better predictive accuracy.
- The distribution is more concentrated around zero, indicating fewer significant errors.

Conclusion

- The Random Forest Regression model significantly outperforms Linear Regression.
- Its ability to model non-linear relationships likely contributes to its superior performance.

3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a tree-based gradient boosting algorithm known for its efficiency and accuracy. It is particularly effective for capturing non-linear relationships and complex feature interactions.

Model Training and Hyperparameter Optimization

We employ grid search to optimize the following hyperparameters of the XGBoost model:

- *learning_rate*: Controls the step size; a smaller value increases model stability.
- *max_depth*: The depth of the decision tree, controlling model complexity.
- *n_estimators*: The number of decision trees, affecting model predictive capability and the risk of overfitting.

Results Analysis

The model outputs the following evaluation metrics:

- Best Parameters: 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50

Performance Metrics:

- MSE: 0.7176
- RMSE: 0.8471
- R^2: 0.3201

3.4 Time Series Forecasting

we used a GRU (Gated Recurrent Unit) model to predict crop yields based on historical climate data. GRU is an efficient recurrent neural network (RNN) that handles time series data well. It is similar to LSTM but is more computationally efficient.

3.4.1 Data Preparation

Before we trained the model, we prepared the data. First, we applied **MinMaxScaler** to normalize the features. This rescaling helps the model learn more efficiently. We used the following features for prediction:

- Temperature
- Precipitation
- CO2 emissions
- Pesticide use
- Fertilizer use

The target variable was the crop yield. We then used a **sliding window** approach to format the data for time series forecasting. The `look_back` was set to **12 months**, meaning the model used the past 12 months to predict the crop yield for the next month.

3.4.2 Model Architecture

We used a **GRU model** for this task. GRU models are effective for time series data and work faster than LSTM models.

The architecture is as follows:

- **Input Layer**: The input shape is (`look_back`, `number_of_features`).
- **GRU Layer 1**: This layer has 128 units and outputs sequences to the next layer.
- **Dropout Layer 1**: We applied a dropout rate of 0.3 to reduce overfitting.
- **GRU Layer 2**: This layer has 64 units.
- **Dropout Layer 2**: We applied a second dropout layer with a rate of 0.2.
- **Dense Layer**: This layer has 32 units with ReLU activation.
- **Output Layer**: The final layer has one unit to predict the crop yield.

3.4.3 Model Training

We trained the model for **200 epochs** with a **batch size of 32**. The **Adam optimizer** and **mean squared error (MSE)** loss function were used. MSE is a common metric for regression tasks like crop yield prediction.

The training progress showed that both the training and validation loss decreased over time, indicating that the model was learning well.

3.4.4 Results Visualization

To visualize the model's performance, we plotted the actual crop yields and the predicted crop yields. The plot below shows how closely the predicted values match the actual values for the test data. In this plot:

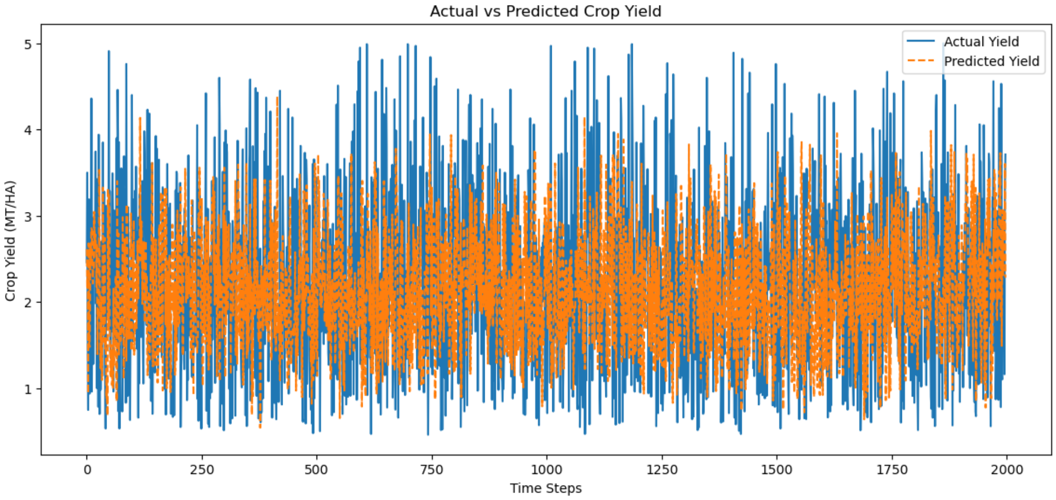

Figure 8: Actual vs Predicted Crop Yield

- The **solid line** represents the **actual crop yields**.
- The **dashed line** represents the **predicted crop yields**.

The plot shows that the model captured the general trend of crop yields, though some deviations occurred, which is expected in forecasting tasks.

3.5 Implementation of the Streamlit App

3.5.1 Introduction

The Streamlit application that renders the prediction of crop yield dependent on different input parameters was built to find how climate change affects agricultural yields. The app employs two machine learning models:

1. A **Pre-trained Joblib Model** predicting crop yield based on historical trends.
2. A **GPU Model** which uses time series data to estimate future crop yield for different climate scenarios.

This enables users to input climate related factors like temperature, precipitation, Co2 emission levels, pesticide application and fertilizer application. The models estimate what the crop yield will be within a given year which can assist in the evaluation of how climate change will alter agricultural output.

3.5.2 Streamlit App Setup

The app was developed on Streamlit, a powerful framework that allows the development of advanced interactive web applications.The following steps describe the key components of the app:

- **Input Features**:The user is able to interact with some sliders and input fields to edit the prediction model. Such include:
  - **Year**: This figure determines the year under which the prediction will be done (range 1990 to 2030).

  - **Average Temperature (°C)**: This parameter is the average temperature for the particular year chosen.
  - **Total Precipitation (mm)**: This parameter gives the average precipitation that is to be received in the selected year.
  - **CO2 Emissions (MT)**: This estimates the volume of CO2 emissions that are in metric tons.
  - **Pesticide Use (KG/HA)**: Measurement of the amount of fumigant (Kg/ha) applied during the planting season of crops
  - **Fertilizer Use (KG/HA)**: Fertilizers (NPK, Urea, Mukkaru, etc.) applied in kilograms per hectare.

- **Model Selection**: The user can select the prediction model:
  - **Pre-trained Joblib Model**: A regression model that was trained on historical climate data to predict crop yield.
  - **GPU Model**: A deep learning model that uses time series data for forecasting future crop yield.

3.5.3 App Functionality

- **Data Input and Preprocessing**: The application has a sidebar for users to input the data. After inputting.Information is subsequently saved in a Pandas DataFrame in which the user can view a summary of the information.

- Once features have the GPU Model selected, the input ones are scaled adopting the MinMaxScaler in order to match the preprocessing performed on the model during the training. This scaling step is essential for the GPU Model to make the right predictions;

- **Model Prediction**: When the user presses the **"Predict Crop Yield"** button, the app loads the appropriate model:
  - For the **Joblib Model**, it loads the pre-trained model using **joblib** and makes the prediction based on the input features.
  - For the **GPU Model**, the app loads the pre-trained GPU Model using **Keras** and predicts the crop yield based on the scaled input features.

- The app then displays the predicted crop yield in **metric tons per hectare (MT/HA)**.

- **User Interface**: The app is designed to be simple and intuitive. The user can easily input climate data, select the model, and view the results.
  - The **sidebar** houses all the input fields, model selection options, and a prediction button.
  - The **main area** displays the **user input summary**, including a table with the entered data, and the **predicted crop yield** after the model makes the prediction.

3.5.4 Conclusion

The Streamlit application is able to merge climate data inputs with forecasts and models to make estimates of the agricultural output against climate changes. Not only does it cater for agricultural productivity predictors, it presents users the options of either regression models or GPU models predicting weather. Such a tool can be used to conduct further research where decisions can be made on how to project the effects of climate change on agricultural yield in the future.

3.6 Carbon Emission

Artificial intelligence in recent years has greatly evolved however the efficiency of such systems is hindered by the vast amount of resources required for training them. This increased energy demand and carbon footprint negatively impact the environment. In an attempt to solve this issue, this report employs the CodeCarbon tool to assess the energy and carbon emissions of a GRU (Gated Recurrent Unit) model.

3.6.1 CodeCarbon Tool Overview

CodeCarbon is a free-access application specifically aimed at estimating the power and carbon cost of training practices of machine learning models. Such a system achieves this by coupling with databases on carbon intensity of electricity grids around the world to provide more accurate carbon emissions figures due to the energy consumption throughout model training.

3.6.2 Implementation Steps

In this research, CodeCarbon was used and incorporated into the process of undertaking the GRU model. The steps include the following:

1.Installation and Initialization: The EmissionsTracker was recorded and introduced with the intention of keeping track of the energy spent during the period of training. The tracker commences prior to the inception of training and ceases towards the end of the training.

2.Model Training: The GRU model was trained with 200 epochs, with a batch size of 32. Training also began by issuing the tracker.start() command which was later repossessed by using tracker.stop() after training had completed.

3.New version: Energy and Carbon Emission Calculation: CodeCarbon automatically assesses energy usage and carbon emissions by utilizing local electricity grid data. You can access the total carbon emissions.

Results and Analysis: From the output of CodeCarbon monitoring, we gathered the following information:

- Carbon Emission Rate: The carbon emission rate of the model during training was 0.006328 g CO2eq/s, which translates to an estimated annual emission of 199.56 kg CO2eq/year.
- RAM Energy Consumption: The RAM consumed 0.002581 kWh during training, with an overall power consumption of 5.90 W.
- CPU Energy Consumption: The CPUs used a total of 0.018601 kWh during training, with a total power consumption of 42.5 W.
- Total Energy Consumption: The cumulative energy consumed throughout the entire model training process was 0.021181 kWh.

Conclusion: Incorporating CodeCarbon into the model training process has offered a detailed perspective on the energy consumption and carbon emissions linked to training a GRU model. The findings emphasize the environmental implications of machine learning model training and the necessity of monitoring and reducing energy usage.

4 CONCLUSION

The study reports the importance of climate change on agricultural productivity. The analysis using models such as Random Forest Regression, XGBoost, and Graphical Representation of Uncertainty (the abbreviation for a recurrent neural network, sometimes also seen as GRU) shows how temperature and precipitation as well as agricultural practices, such as irrigation and fertilizer usage, were impacting crop-yield results. XGBoost and Random Forest Regression were found to be superior to linear models in the analysis-none of the linear models managed well due to the ability of the latter models to capture the highly complicated and non-linear relationships. The GRU model, concerned with time series forecasting, efficiently predicted future crop yields through a historical basis, consequently improving the accuracy of the predictions. The proposition launches the opportunity to devise adaptive agricultural practices and policies geared toward ensuring food security under a changing climate.

5 BIBLIOGRAPHIESS

[1] Ray, D.K., Gerber, J.S., MacDonald, G.K., West, P.C. 2015. ”Climate variation explains a third of global crop yield variability.” Nature Communications, 6:5989. https://doi.org/10.1038/ncomms6989

[2] Najafi, E., Devineni, N., Khanbilvardi, R.M., Kogan, F. 2018. ”Understanding the changes in global crop yields through changes in climate and technology.” Earth’s Future, 6(5): 661-681. https://doi.org/10.1002/2017EF000690

[3] Hu, T., Zhang, X., Bohrer, G., Liu, Y., Zhou, Y., Martin, J., Li, Y., Zhao, K. 2023. ”Crop yield prediction via explainable AI and interpretable machine learning: dangers of black box models for evaluating climate change impacts on crop yield.” Agricultural and Forest Meteorology, 336:109458. https://doi.org/10.1016/j.agrformet.2023.109458

[4] Zhang, N., Zhou, X., Kang, M., Hu, B.G., Heuvelink, E., Marcelis, L.F. 2023. ”Machine learning versus crop growth models: an ally, not a rival.” AoB Plants, 15(2): plac061. https://doi.org/10.1093/aobpla/plac061