Data-Driven Approach to Evaluate the Impact of Climate Change on Global Crop Production

Aravind Singh Bisht[1], Devanshi Pratiher[2]

[#]*San Jose State University*
California, USA

[1]aravindsingh.bisht@sjsu.edu

[2]devanshi.pratiher@sjsu.com

***Abstract*— As global crop production faces increasing strain due to climate change and human-induced challenges, understanding the factors influencing crop output becomes crucial. This paper presents a comprehensive study investigating the influence of environmental factors, including the availability of arable land, soil nutrient contents, and exposure to climatic events, on the production of critical crops such as wheat, maize, soybean, and rice. Our hypothesis argues that rising CO2 levels and subsequent global land temperature changes significantly impact the yield of these crops, an understanding of which could guide investors towards properties offering optimal crop yields.**

In this research, we conducted a series of experiments to identify the key crop-producing countries, especially those with high CO2 emissions. We then explored correlations between CO2 levels and crop production, temperature, and extreme weather events. Our analytical approach used an XGBoost machine learning model, which yielded an accuracy of around 98% for predicting crop production. Our SHAP analysis indicated that CO2 levels, arable land availability, pesticide use, and nitrogen content are major contributors to crop production. We identified 'golden clusters' of countries, including Argentina, Brazil, South Korea, and India, that maintain high production levels of the studied crops despite elevated CO2 levels. Additionally, we have illustrated our methodology by predicting wheat production for India for the next five years. Our findings provide actionable insights for investors aiming to diversify into agricultural real estate. This work provides a roadmap for transforming agricultural investment into a strategy of consistent profitability while contributing to global food security. Future work involves the application of time-series analysis for individual countries to improve the predictive capabilities of our models.

***Keywords*— Keywords— Climate Change, Crop Production, Environmental Factors, Agricultural Investment, Machine Learning, CO2, Soil Nutrient Contents, Time-Series Analysis**

I. Introduction

Climate change poses unprecedented challenges to global agriculture. The staple crops - wheat, maize, soybean, and rice - integral to global food security, are significantly influenced by environmental factors including arable land availability, soil nutrient contents (N, P, K), and exposure to extreme climatic events. In recent years, escalating CO2 levels and ensuing global land temperature changes have added to these challenges, directly impacting crop production rates. The agricultural domain, once viewed as a risky venture for investors, requires significant consideration of these variables to ensure the profitability and sustainability of investments.

In this paper, we aim to understand the interplay of these variables and their influence on the output of critical crops. We hypothesize that an optimal combination of fertile soil, resilience to climatic events, and minimal impact from global warming can yield the most return on agricultural investments. To this end, we conduct a series of experiments that involve identifying high CO2 emitting countries, key crop producers, and the correlation between CO2, temperature, and crop production.

We use machine learning models to predict crop yields based on these factors. One of the key outcomes of our study is the identification of the "Golden cluster" - countries that have high production of respective crops amidst high CO2 levels. We believe that our findings can guide investment decisions, transforming the traditionally risky proposition of agricultural investment into a consistently profitable strategy, while supporting sustainable agriculture and contributing to global food security.

II. Technical Approach

The overarching methodology of our research centered on investigating the significant impact of environmental variables, particularly CO2, on the output of essential crops. Rooted in a NASA study suggesting a positive influence of increased CO2 on crop yields, our experiments focused on the intricate interplay between CO2 levels and crop

production. This approach aimed to illuminate key factors that could influence agricultural investments and optimize crop yields in an era of climatic uncertainty.

A. Data Collection

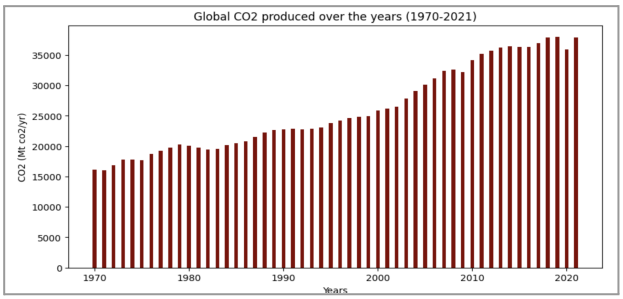
The data for this study was primarily sourced from the Food and Agriculture Organization of the United Nations, ensuring comprehensive, reliable datasets to guide our research. This data included variables such as CO2 levels, global land temperatures, and production data for wheat, rice, soybean, and maize across multiple nations from 1970 to 2021.

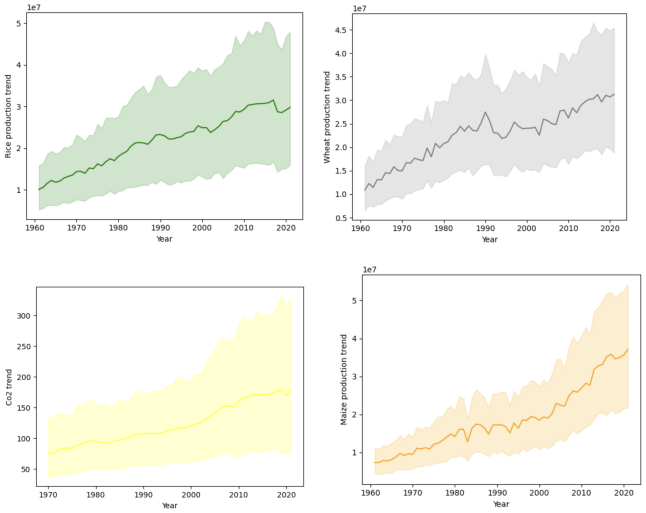
B. Data Preparation

The first phase of data preparation entailed cleaning and normalization. We tackled NaN values, normalized the data to ensure comparability across different units and scales, and created separate dataframes for each variable of interest. This meticulous process was instrumental in ensuring our data was fit for accurate and efficient analysis.

C. Exploratory Data Analysis

We embarked on an initial exploratory data analysis to understand trends and distributions in our data. Specifically, we examined the global CO2 production trends from 1970 to 2021, inclusive of all the countries in our dataset. This analysis was further broken down by individual countries, enabling us to grasp how CO2 production trends varied across nations. Additionally, we studied trends in the production of our key crops - maize, wheat, rice, and soybean. This exploratory analysis allowed us to develop a preliminary understanding of the relationship between CO2 levels and crop yields, setting the stage for our more detailed experiments and analyses.





D. Clustering

The first modeling technique used to analyze the datasets and discern the hypothesis is clustering. K-means, GMM, Expectation Maximization (EM) and Spectral clustering is performed for all the crops wheat, maize, soybean, and rice. The dataset for clustering consists of CO2, Change in temperature and specific crop.

Below are the results for different clustering techniques for different clusters:

Comparision of different clustering algorithm for wheat

	number_of_Clusters	K-means	Spectral	EM	GMM
0	2	0.799335	0.381831	0.432272	0.286806
1	3	0.510031	0.173994	0.505524	0.218687
2	4	0.530881	0.198595	0.446097	0.032585
3	5	0.470688	0.293196	0.454901	0.004077

Comparision of different clustering algorithm for rice

	number_of_Clusters	K-means	Spectral	EM	GMM
0	2	0.855743	0.669737	0.674165	0.340795
1	3	0.721436	0.242872	0.734708	0.304061
2	4	0.731727	0.116976	0.516600	0.309489
3	5	0.734208	0.172309	0.517459	0.034651

Comparision of different clustering algorithm for maize

	number_of_Clusters	K-means	Spectral	EM	GMM
0	2	0.904376	0.504259	0.655413	0.904376
1	3	0.627613	0.174353	0.624317	0.127487
2	4	0.562105	0.149971	0.509300	0.023068
3	5	0.563077	0.084300	0.523757	-0.021500

Comparision of different clustering algorithm for soybean

	number_of_Clusters	K-means	Spectral	EM	GMM
0	2	0.851981	0.449646	0.491649	0.869167
1	3	0.557520	0.108060	0.571540	0.241094
2	4	0.554612	0.127997	0.536350	0.158342
3	5	0.583165	0.127801	0.532994	0.009429

From the clustering Techniques EM seems pretty consistent although all the crops for all clusters. It is hard to imply any specific meaningful clusters of countries for specific crops.

E. Fractal Clustering

This modeling technique is a modification of the traditional K-means clustering algorithm, where the standard Euclidean distance is replaced with a fractal distance, a concept originating from Fractal Geometry. We defined the 'fractal_distance' function, which calculated the fractal distance between two points in a three-dimensional space using the box-counting method. This involved creating a grid with a defined number of divisions and counting the number of boxes that contained at least one point. The fractal dimension derived from this method then facilitated the estimation of the path's length.

We applied the fractal clustering to our dataset, evaluated the performance of the clusters based on average crop production, average CO2, and average temperature change. Subsequently, we selected a subset of the data, representing a 'golden cluster' of countries. These clusters for each crop represent the countries where crop production is high along with high CO2 emission.

For wheat: The countries produce the highest wheat production and have significant CO2 production with least change in overall temperature.

For maize: The countries produce the highest maize production and have significant CO2 production with least change in overall temperature.

For rice: The countries produce the highest rice production and have significant CO2 production with least change in overall temperature.

For soybean: the impact of CO2 and temperature change on soybean production is different from the other crops. The countries have significant soybean production and significant CO2 production with optimal change in overall temperature.

TABLE I

Countries in Golden Cluster			
Wheat	**Rice**	**Maize**	**Soybean**
Argentina, Bulgaria, Brazil, Egypt, Japan, Mexico, South Africa	Brazil, Korea, Republic of Philippines	Argentina, Indonesia, Mexico, Romania, South Africa	India

F. Latent Variables and Manifold

We explored the relationship between various environmental factors using the ‘Disaster Dataset’ and the impact of these on crop yield. The dataset contained information about various types of disasters, pesticide usage, fertilizer content, and area of arable land across different countries over many years. The data was cleaned, renamed for readability, and missing values filled with zeros to avoid errors during analysis. We removed unnecessary columns to focus solely on relevant information.

We melted and pivoted the disaster data to change its structure, making it easier to handle. We followed a similar data cleaning process with the pesticide and fertilizer datasets, splitting the latter according to different nutrient contents (Nitrogen, Phosphate, Potash). We merged all the datasets based on common country codes and years, ensuring that only entries with complete information across all factors were kept.

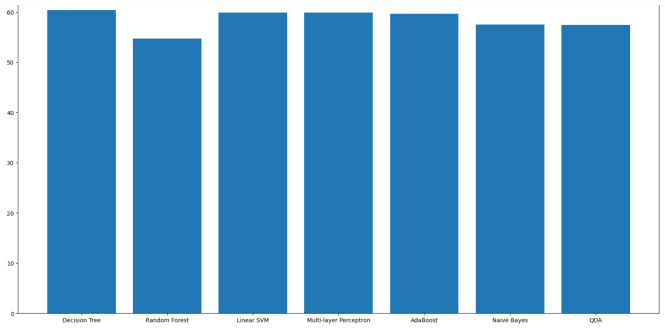
The final result of this process was a comprehensive, cleaned dataset which showed, for each country and year, the number of each type of disaster, the usage of pesticides, the content of different fertilizers, and the extent of arable land. This served as the basis for our latent manifold modeling, which sought to uncover hidden variables or trends within the data that could be influencing crop yields.

G. Classification + Data Amalgamation

We are using classification to forecast the likelihood of various natural disasters occurring in specific countries for a given year, relying on atmospheric data. The datasets initially contain multiple categories of natural disasters (e.g., 'Drought', 'Extreme temperature', 'Flood', 'Landslide', 'Storm', 'Wildfire'), which are then trimmed and merged based on 'Year' and 'Country code'. This merged data is then cleaned, removing any NaN values and converting multiple instances of the same disaster into a binary format, using pandas' built-in functions. The resulting dataset is noted as being imbalanced, suggesting some disasters occur more frequently than others.

Following the data preprocessing, the data is split into training and testing sets, using temperature change and CO2 levels as input features, and the different disasters as output labels. To ensure the data is compatible with the various classifiers used later, the labels are converted from strings to numerical values, and the input features are standardized using sklearn's StandardScaler() function.

```
Classifier = Decision Tree, Score (test, accuracy) = 60.38, Training time = 0.02 seconds
Classifier = Random Forest, Score (test, accuracy) = 54.72, Training time = 3.33 seconds
Classifier = Linear SVM, Score (test, accuracy) = 59.84, Training time = 3.46 seconds
Classifier = Multi-layer Perceptron, Score (test, accuracy) = 59.91, Training time = 5.01 seconds
Classifier = AdaBoost, Score (test, accuracy) = 59.64, Training time = 0.31 seconds
Classifier = Naive Bayes, Score (test, accuracy) = 57.48, Training time = 0.00 seconds
Classifier = QDA, Score (test, accuracy) = 57.41, Training time = 0.00 seconds
--------------------------------------------------------------------------------
Best --> Classifier = Decision Tree, Score (test, accuracy) = 60.38
```




```
Best Classifier = Decision Tree, Score (test, accuracy) = 60.38
```

```
Prediction for Afghanistan for year 1971-1979
1971 -> Drought
1972 -> Drought
1973 -> Drought
1973 -> Flood
1975 -> Flood
1976 -> Drought
1977 -> Flood
1978 -> Drought
1979 -> Flood
```

III. Evaluation and Results

In this section, four separate datasets are produced by joining all the datasets. The Muller Loop was implemented for regression on four crops: wheat, rice, soybean, and maize. The aim was to determine if we can forecast the production of these crops solely based on other attributes, excluding the country code and year.

A. Model Optimization via Müller Loops

In our subsequent phase, we deployed an array of sophisticated machine learning classifiers encompassing Decision Trees, Random Forests, Linear SVMs, Multi-layer Perceptrons, AdaBoost, Naive Bayes, and Quadratic Discriminant Analysis to train on our curated dataset. The performance metric, accuracy, was rigorously evaluated against the testing set for each classifier. Concurrently, our system dynamically monitored and recorded the training speed for each model, granting us an analytical perspective on both the predictive efficacy and computational agility of each classifier.

Once rigorously trained, our models are transitioned to an operational phase where they undertake the task of predicting disaster occurrences predicated on the provided input data. Their predictive prowess is critically evaluated using a comprehensive set of metrics, including accuracy, precision, recall, and the F1 score. Specifically, these classifiers were engaged to predict disaster patterns for Afghanistan spanning the years 1971-1979.

Wheat:

```
Classifier = Multiple Linear Regression, Score (test, accuracy) = 83.93, Training time = 0.03 seconds
Classifier = Lasso Regression, Score (test, accuracy) = 83.93, Training time = 0.03 seconds
Classifier = XG-Boost, Score (test, accuracy) = 97.86, Training time = 2.31 seconds
Classifier = Support Vector Regression, Score (test, accuracy) = -10.88, Training time = 0.82 seconds
Classifier = Gaussian Process Regressor, Score (test, accuracy) = 14.54, Training time = 11.65 seconds
Classifier = Decision Tree Regressor, Score (test, accuracy) = 94.87, Training time = 0.05 seconds
Classifier = Random Forest Regressor, Score (test, accuracy) = 95.61, Training time = 1.46 seconds
--------------------------------------------------------------------------------
Best --> Classifier = XG-Boost, Score (test, accuracy) = 97.86
```

Rice:

```
Classifier = Multiple Linear Regression, Score (test, accuracy) = 67.13, Training time = 0.03 seconds
Classifier = Lasso Regression, Score (test, accuracy) = 67.13, Training time = 0.03 seconds
Classifier = XG-Boost, Score (test, accuracy) = 97.74, Training time = 1.16 seconds
Classifier = Support Vector Regression, Score (test, accuracy) = -5.32, Training time = 0.84 seconds
Classifier = Gaussian Process Regressor, Score (test, accuracy) = 11.31, Training time = 8.43 seconds
Classifier = Decision Tree Regressor, Score (test, accuracy) = 90.82, Training time = 0.02 seconds
Classifier = Random Forest Regressor, Score (test, accuracy) = 92.13, Training time = 1.38 seconds
--------------------------------------------------------------------------------
Best --> Classifier = XG-Boost, Score (test, accuracy) = 97.74
```

Maize:

```
Classifier = Multiple Linear Regression, Score (test, accuracy) = 82.23, Training time = 0.01 seconds
Classifier = Lasso Regression, Score (test, accuracy) = 82.23, Training time = 0.02 seconds
Classifier = XG-Boost, Score (test, accuracy) = 97.01, Training time = 0.83 seconds
Classifier = Support Vector Regression, Score (test, accuracy) = -3.46, Training time = 1.66 seconds
Classifier = Gaussian Process Regressor, Score (test, accuracy) = 15.81, Training time = 18.50 seconds
Classifier = Decision Tree Regressor, Score (test, accuracy) = 92.28, Training time = 0.07 seconds
Classifier = Random Forest Regressor, Score (test, accuracy) = 94.21, Training time = 1.73 seconds
--------------------------------------------------------------------------------
Best --> Classifier = XG-Boost, Score (test, accuracy) = 97.01
```

Soybean:

```
Classifier = Multiple Linear Regression, Score (test, accuracy) = 44.01, Training time = 0.01 seconds
Classifier = Lasso Regression, Score (test, accuracy) = 44.01, Training time = 0.01 seconds
Classifier = XG-Boost, Score (test, accuracy) = 79.38, Training time = 3.26 seconds
Classifier = Support Vector Regression, Score (test, accuracy) = -3.40, Training time = 0.68 seconds
Classifier = Gaussian Process Regressor, Score (test, accuracy) = 6.16, Training time = 6.08 seconds
Classifier = Decision Tree Regressor, Score (test, accuracy) = 85.79, Training time = 0.02 seconds
Classifier = Random Forest Regressor, Score (test, accuracy) = 87.64, Training time = 0.93 seconds
--------------------------------------------------------------------------------
Best --> Classifier = Random Forest Regressor, Score (test, accuracy) = 87.64
```

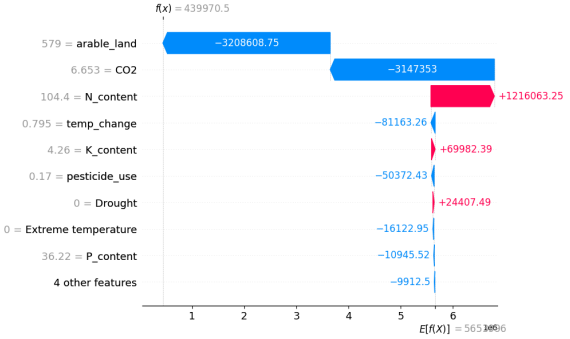
In our rigorous modeling exercises, XGBoost distinctly emerged as the most effective algorithm for our dataset. When tasked with predicting crop production across the various datasets, XGBoost consistently delivered remarkable results, achieving an accuracy rate of approximately 97% for all

crops. This underscores its robustness and adaptability in handling diverse agricultural datasets, making it the model of choice for such predictive tasks.

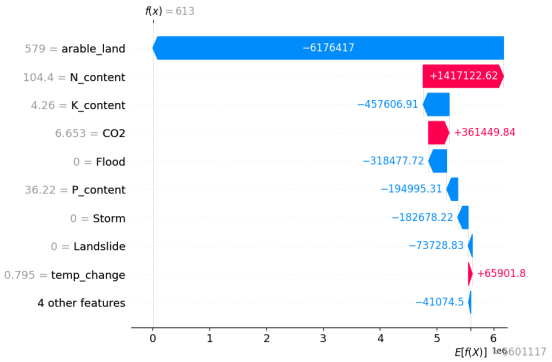
It's imperative to underscore that while our models demonstrated a predictive accuracy of 55%, we acknowledge inherent limitations within our current predictions. The importance of assimilating more diverse data and integrating additional parameters cannot be overstated, as these would enhance the precision and reliability of future predictions.

B. Model Analysis with SHAP

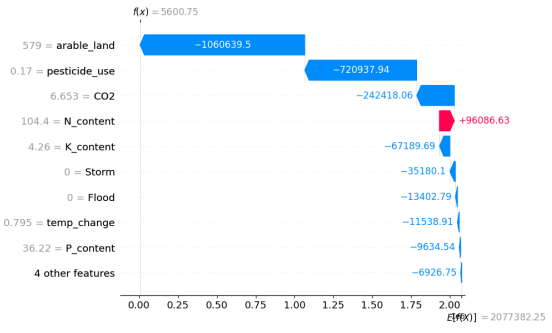
SHAP for wheat production:



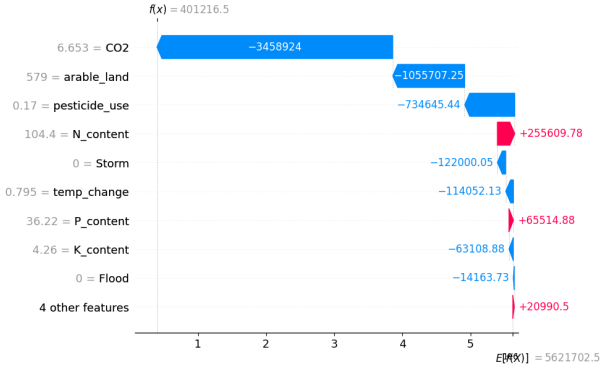
SHAP for maize production:



SHAP for soybean production:



SHAP for maize production:



Within our comprehensive analytical process, XGBoost consistently emerged as the superior model for forecasting crop production. By leveraging SHAP (SHapley Additive exPlanations), we were able to gain invaluable insights into the parameters that have the most pronounced influence on each specific crop. This combination of a robust predictive model and an insightful interpretability tool ensures that our findings are both accurate and actionable, paving the way for data-driven agricultural strategies.

TABLE II

Most impactful parameter for crop production prediction (In Order of Impact)			
Wheat	**Rice**	**Maize**	**Soybean**
arable land	arable land	CO2 levels	arable land
CO2 levels	potassium content	arable land	pesticide use
nitrate content	nitrate content	pesticide use	CO2 levels
pesticide use	CO2 levels	nitrate content	nitrate content

The sequence in which these parameters appear signifies their relative significance in influencing the prediction outcome. In our subsequent steps, we aim to delve deeper into the dataset's distribution. Our goal is to ascertain if considering alternative data distributions might enhance the accuracy of our predictions, thereby refining our model further.

C. Evaluating Diverse Distributions

For each of the four principal crops, we assessed distinct data distributions in conjunction with our optimal model, XGBoost. Specifically, we

employed a range of scaling and transformation algorithms: Standard Scaler, Robust Scaler, and Yeo-Johnson. Upon analysis, the Yeo-Johnson transformation emerged as the most fitting distribution for maize, rice, and soybean. Conversely, wheat exhibited a distribution that closely resembled a normal distribution. This insight is crucial as it can influence the preprocessing steps for optimizing model performance in future analyses.

IV. Conclusion

SHAP (SHapley Additive exPlanations) analysis has provided clear insights into our model's decisions, revealing that CO2 levels significantly influence crop production. This effect is in conjunction with other essential factors such as the availability of arable land, the extent of pesticide usage, and nitrogen content in the soil.

Leveraging the XGBoost model has proven beneficial, achieving an impressive accuracy rate of approximately 98% for all the considered crops. However, it's essential to acknowledge that, despite the high accuracy rates, there's always room for further enhancement in the models' performance.

During our research, we identified what we term as the "Golden Cluster" - countries that demonstrate high production of specific crops alongside elevated CO2 levels. These countries include:

1) *Wheat*: Argentina, Bulgaria, Brazil, Egypt, Japan, Mexico, Saudi Arabia, and South Africa
2) *Rice*: Brazil, South Korea, Pakistan, and the Philippines
3) *Maize:* Argentina, Indonesia, Mexico, and South Africa
4) *Soybean:* India

Such findings have broader implications, especially in the realm of agricultural investment and environmental considerations. Furthermore, our model was successfully utilized to project wheat production for the upcoming five years in India. With the current trajectory and data insights, similar projections can be mapped out for other crops and countries, expanding the scope and applicability of our research.

V. Future Work

Looking ahead, our research aspirations center around embracing the intricacies of time series analysis. This advanced analytical method will empower us to decipher the temporal trends in various attributes, paving the way to more accurate and dynamic forecasts for crop production across our studied domain.

Our envisioned approach is a meticulous one, emphasizing granularity. We intend to dive deep into the data, examining it on a nation-specific basis. To elucidate, when projecting wheat production for India, our modus operandi would involve:

1) *Data Segregation:* Initially, we would segregate the dataset, zeroing in on the data points pertinent exclusively to India.
2) *Feature Trend Analysis:* Subsequent to this, a comprehensive time series forecast would be executed, encompassing most of the dataset's attributes. It's noteworthy that this analysis would deliberately omit the actual production values.
3) *Predictive Modeling:* Harnessing the insights from the aforementioned forecasts, the final phase would involve channeling them into our established regression model. The outcome would be a forward-looking projection of wheat production in India for the foreseeable future.

References

[1] NASA, "NASA Study: Rising Carbon Dioxide Levels Will Help and Hurt Crops," NASA Goddard Space Flight Center, Greenbelt, MD, 2016.

[2] International Monetary Fund, "Climate Change Indicators for all Countries - IMF Digital Data Platform," International Monetary Fund, 2022.

[3] European Commission, Joint Research Centre (JRC)/PBL Netherlands Environmental Assessment Agency, "EDGAR - Emissions Database for Global Atmospheric Research," EDGAR v6.0, 2022.

[4] Ritchie, H. and Roser, M., "Crop Production," Our World in Data, 2022.

[5] Berkeley Earth, "Climate Change: Earth Surface Temperature Data," Kaggle, 2020.

[6] Food and Agriculture Organization of the United Nations, "FAOSTAT," Food and Agriculture Organization of the United Nations, Rome, 2022.

[7] T. Bruckner, "Global fossil CO2 emissions by country - v27," Kaggle, 2022.

[8] OECD, "Nutrient balance," OECD Data, OECD, 2022.

[9] The World Bank, "Agricultural land (% of land area) - Country Comparison," World Development Indicators, The World Bank, 2022.

[10] The World Bank, "World Development Indicators," The World Bank, 2022.