How Agricultural Food Production Impacts Climate Change

1 Introduction

Climate change is a crucial issue that affects the entire world. It is caused by various natural and human-influenced factors that lead to significant changes in temperature and weather patterns over time. Based on the temperature deviation data provided by NASA, it has been observed that there is a significant increase in temperature deviation. As shown in Figure 1, the temperature deviation has been increasing since 1880 and has rapidly increased in recent years. One of the main causes of this rapid increase is the Industrial Revolution. Since the Industrial Revolution, human activities have released large amounts of greenhouse gas emissions, which have significantly impacted climate change. This suggests that human activities causing greenhouse gas emissions are playing a crucial role in accelerating climate change.

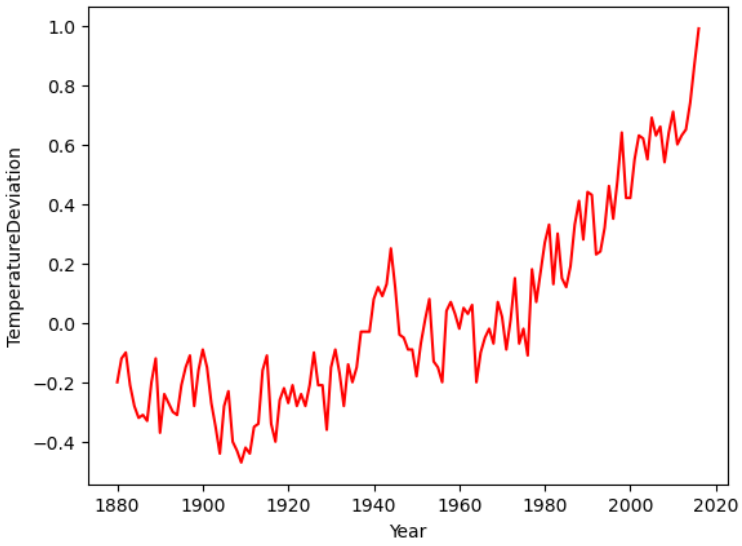


Figure 1: Trend in temperature deviation

Human activities, including food production, transportation, and manufacturing, are the leading causes of greenhouse gas emissions. According to Our World in Data, food production alone accounts for 25% of global greenhouse gas emissions [1]. The processes involved in food production, such as food transportation, food manufacturing, and food packaging, contribute to significant amounts of CO2 emissions, which in turn cause changes in climate.

This report will focus on one of the primary human-influenced causes of climate

change: CO2 emissions from food production. The analysis will involve conducting correlation and time series analyses to determine the relationship between food production and temperature deviation and forecast future temperature deviations based on agricultural food production processes.

2 Data Preprocessing

Data analysis begins with collecting and preprocessing data. Two datasets are chosen for this purpose: Agri-food CO2 emissions and temperature deviation datasets. The agri-food CO2 emissions dataset, obtained from Kaggle, contains 6965 instances from various regions for several years ranging from 1990 to 2020. The dataset includes variables related to the CO2 emissions of different processes involved in agricultural food production, such as $Forestfires$, $On-farmElectricityUse$, and $FertilizerManufacturing$, etc. Apart from CO2 emissions, the dataset also includes variables related to population, such as $Ruralpopulation$, $Urbanpopulation$, and $Totalpopulation$ for males and females. In total, there are 31 variables. On the other hand, the temperature deviation dataset has only two variables: $Year$ and $TemperatureDeviation$. The dataset has 137 instances, representing 137 years from 1880 to 2016.

To begin with, in order to analyse the agri-food CO2 emissions data, the dataset needs to be preprocessed. The first step is to fill in the null values. There are several variables in the dataset that have missing values. Upon closer examination, it can be observed that all variables are missing for certain countries rather than random missing values. For instance, for data with the variable $Area$ as $HolySee$, the variables $Savannafires$ and $Forestfires$ are all missing. However, it is reasonable to assume that fire events cannot occur in the city of the Vatican, so it is evident to assume that these missing values are meant to be 0, indicating that such cases do not exist. Similarly, the variable $Firesinhumidtropicalforests$ is missing for data with the variable $Area$ as $Monaco$ and $SanMarino$. It is evident that these countries do not have tropical forests, so it is reasonable to conclude that they are Missing Data Not at Random (MNAR). Therefore, all missing values are filled with 0, assuming they are missing because the value is 0.

Then, the agri-food CO2 emissions dataset is preprocessed by year to analyse climate change over time. This involves using the *groupby* function to group the dataset by the variable $Year$ and calculating the sum of each variable for each year. This grouped dataset provides the total CO2 emissions of each agricultural food production process for each year.

Next, standardisation is applied to the agri-food CO2 emissions dataset. The

dataset represents global CO2 emissions, which results in large emission values and variable scales that vary greatly. For instance, the median value of CO2 emissions for the $Forestfires$ variable is approximately 200,000, while the median value for the $IPPU$ (Industrial Processes and Product Use) variable is around 3,000,000. Therefore, the variables need to be scaled. To accomplish this, the $StandardScaler$ function of $sklearn$ is applied to all variables except for the variable $Year$, as it is an indicator for time series.

Several variables are removed from the agri-food CO2 emissions dataset due to many missing values. These variables include $CropResidues$, $ManureappliedtoSoils$, $ManureManagement$, and $On-farmenergy$, each containing around 1,000 missing values. Additionally, some variables not related to CO2 emissions are removed from the analysis, such as $Forestland$, $NetForestconversion$, $Ruralpopulation$, etc. This is because the report exclusively focuses on the CO2 emissions of agricultural food production.

Lastly, the preprocessed agri-food CO2 emissions dataset is combined with the temperature deviation dataset. The $merge$ function is used with options of $on =$ '$Year$' and $how =$ '$inner$'. Since the variable $TemperatureDeviation$ covers data from 1880 to 2016, and the agri-food CO2 emissions dataset covers data from 1990 to 2020, the datasets are combined to create a new dataset that includes selected CO2 emissions variables from the previous step and the $TemperatureDeviation$ variable from 1990 to 2016.

3 Data Analysis

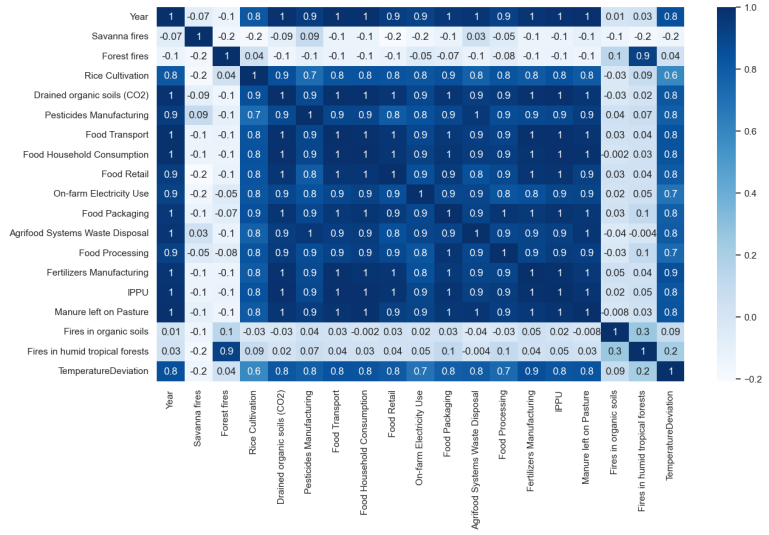
Two different analyses will be conducted on the preprocessed dataset: correlation analysis and time series analysis. The analyses are performed using Python.

3.1 Correlation Analysis

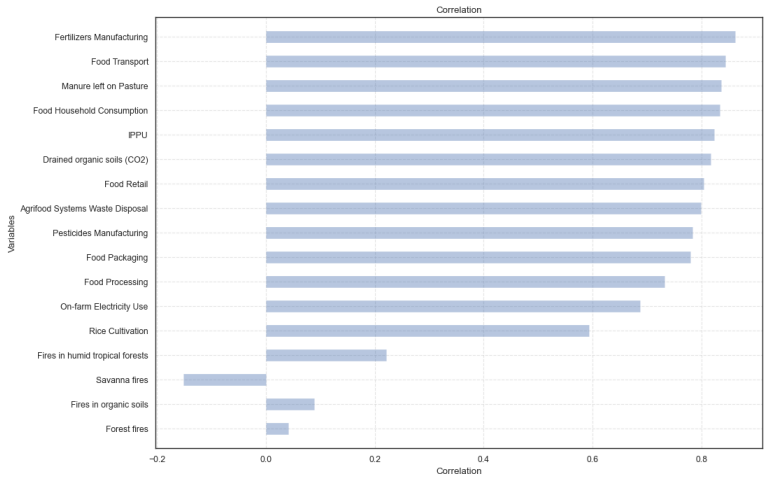
Firstly, correlation analysis is conducted to analyse the relationship between agricultural food production processes and temperature deviation. Correlation analysis is a statistical method that helps discover any relationship between variables and how strong they are. The correlation value ranges from -1 to 1, and the absolute value close to 1 indicates a stronger relationship. In addition, the signs of correlations indicate the direction of the relationship. A positive correlation indicates that both variables increase with each other, while a negative correlation means that as one variable decreases, the other variable increases. There are three different ways to calculate correlations: Pearson correlation, Spearman correlation,

and Kendall correlation. Pearson correlation is the most widely used correlation analysis measure for analysing linear relationships between two variables.

In Figure 2, it is evident that most of the variables are strongly correlated with each other. However, some variables, such as $Savannafires$, $Forestfires$, $Firesinorganicsoils$, and $Firesinhumidtropicalforests$, have low correlations, indicating that they are special events that do not occur frequently. This means that CO2 emissions for these variables do not follow any trend and are not correlated with the other variables. Additionally, by visualising correlations between the variables and the $TemperatureDeviation$ variable in decreasing order of absolute values, it can be concluded that most variables have highly positive correlations of around 0.8 with the $TemperatureDeviation$ variable.



(a) Correlation matrix



(b) Correlations with TemperatureDeviation

Figure 2: Correlation analysis

The next step in correlation analysis is to conduct a correlation test. This test is used to determine whether there is a significant correlation between variables. The null hypothesis is set as $p = 0$ to indicate that there is no linear relationship between the two variables. The alternative hypothesis, on the other hand, states that there is a linear relationship between the two variables. To perform this test, the *pearsonr* function in *scipy* is used to determine the p-value for each variable. Based on the p-value, it can be determined whether the null hypothesis of no correlation can be rejected at a 5% significance level. If the p-value is less than 0.005, it can be concluded that there is a linear relationship between the two variables, rejecting the null hypothesis. The variables $Savannafires$, $Forestfires$, $Firesinorganicsoils$, and $Firesinhumidtropicalforests$ are the only ones with p-values greater than 0.05, indicating that they do not have a linear relationship with the variable $TemperatureDeviation$. Additionally, these variables have the lowest absolute values of correlations with the $TemperatureDeviation$ variable, which further supports this conclusion.

In summary, the correlation analysis shows that there is a strong positive correlation between most of the variables and the $TemperatureDeviation$ variable. These highly positive correlations suggest that agricultural food production has a significant impact on the temperature deviation.

3.2 Time Series Analysis

Secondly, a time series analysis is conducted. Time series analysis is a method that analyses a sequence of data that changes over time, providing some insights and identifying any trend, cycle, or seasonal variance. Time series analysis can be utilised in various ways, such as classification, curve fitting, and forecasting. In this report, a Vector Autoregressive (VAR) model is applied to forecast the temperature deviation. A VAR model is a multivariate time series that analyses dynamic relationships, considering their associations with other variables. This model is flexible and easy to use for multivariate time series analysis.

In a VAR model, each variable is regressed on its own lagged values and the lagged values of the other variables. The model with k variables and p lags can be defined as:

$$\mathbf{y}(t) = \mathbf{a} + \mathbf{w}_1 \cdot \mathbf{y}(t-1) + \cdots + \mathbf{w}_p \cdot \mathbf{y}(t-p) + \mathbf{e}, \tag{1}$$

where $\mathbf{a}$ is a vector representing a constant term with length k, $\mathbf{a}_i$'s are $k \times k$ coefficient matrices for $i = 1, \cdots, p$, and $\mathbf{a}$ is a vector indicating an error term with

length k. This also can be represented as a matrix form as follows:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_k(t) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} + \begin{bmatrix} w1_{1,1} & \cdots & w1_{1,k} \\ w1_{2,1} & \cdots & w1_{2,k} \\ \vdots & \cdots & \vdots \\ w1_{k,1} & \cdots & w1_{k,k} \end{bmatrix} \begin{bmatrix} y_1(t-1) \\ y_2(t-1) \\ \vdots \\ y_k(t-1) \end{bmatrix} + \cdots + \begin{bmatrix} wp_{1,1} & \cdots & wp_{1,k} \\ wp_{2,1} & \cdots & wp_{2,k} \\ \vdots & \cdots & \vdots \\ wp_{k,1} & \cdots & wp_{k,k} \end{bmatrix} \begin{bmatrix} y_1(t-p) \\ y_2(t-p) \\ \vdots \\ y_k(t-p) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}, \tag{2}$$

where $\mathbf{y}_i$'s, $\mathbf{a}$, and $\mathbf{e}$ are $k \times k$ dimensional and $\mathbf{w}_i$'s are $k \times k$ dimensional. For example, if a VAR model with $k = 3$ and $p = 3$ is evaluated, the equation will be defined as follows:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} + \begin{bmatrix} w1_{1,1} & w1_{1,2} & w1_{1,3} \\ w1_{2,1} & w1_{2,2} & w1_{2,3} \\ w1_{3,1} & w1_{3,2} & w1_{3,3} \end{bmatrix} \begin{bmatrix} y_1(t-1) \\ y_2(t-1) \\ y_3(t-1) \end{bmatrix} + \cdots + \begin{bmatrix} w3_{1,1} & w3_{1,2} & w3_{1,3} \\ w3_{2,1} & w3_{2,2} & w3_{2,3} \\ w3_{3,1} & w3_{3,2} & w3_{3,3} \end{bmatrix} \begin{bmatrix} y_1(t-3) \\ y_2(t-3) \\ y_3(t-3) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}. \tag{3}$$

To begin with, two variables with larger scales are selected from the agri-food CO2 emissions dataset: $AgrifoodSystemsWasteDisposal$ and $IPPU$. The median values of these two variables are around 1.5 million and 3 million, respectively, which are 4-6 times larger than the median values of other variables. This indicates that any variations in these selected variables will significantly affect the total amount of CO2 emissions, leading to changes in climate.

The VAR model relies on the assumption that the properties of a time series do not depend on time, which is known as stationary. In order to satisfy this assumption, differentiation is applied to each variable by the function $diff$ in *pandas*, making it stationary. The augmented Dickey-Fuller (ADF) test is then used to examine the stationary of each variable. The ADF test is the most widely used statistical test for this purpose. In this test, the null hypothesis is that the series is non-stationary, while the alternative hypothesis is that the series is stationary. The p-values obtained from the ADF test are analysed to determine whether to reject the

null hypothesis. A p-value of less than 0.05 indicates that the series is stationary and significant. Figure 3 shows how the p-value of each variable changes after differentiation. Based on the ADF test results, it is evident that the variables $AgrifoodSystemsWasteDisposal$ and $TemperatureDeviation$ are stationary, and the p-value of the variable $IPPU$ has improved after differentiation.

```
# Stationarity
adf_ASWD = adfuller(df['Agrifood Systems Waste Disposal'])
adf_IPPU = adfuller(df['IPPU'])
adf_TD = adfuller(df['TemperatureDeviation'])
print("p-value for Agrifood Systems Waste Disposal: {}".format(adf_ASWD[1]))
print("p-value for IPPU: {}".format(adf_IPPU[1]))
print("p-value for TemperatureDeviation: {}".format(adf_TD[1]))
```

```
p-value for Agrifood Systems Waste Disposal: 0.00419553630045962
p-value for IPPU: 0.9936132536929111
p-value for TemperatureDeviation: 0.8531533506108475
```

(a) Before differentiation

```
# After differentiation
adf_ASWD_diff = adfuller(df_stationary['Agrifood Systems Waste Disposal'])
adf_IPPU_diff = adfuller(df_stationary['IPPU'])
adf_TD_diff = adfuller(df_stationary['TemperatureDeviation'])
print("p-value for Agrifood Systems Waste Disposal: {}".format(adf_ASWD_diff[1]))
print("p-value for IPPU: {}".format(adf_IPPU_diff[1]))
print("p-value for TemperatureDeviation: {}".format(adf_TD_diff[1]))
```

```
p-value for Agrifood Systems Waste Disposal: 0.000572844647297303
p-value for IPPU: 0.5017051639984066
p-value for TemperatureDeviation: 2.4215502370478316e-06
```

(b) After differentiation

Figure 3: ADF test

A VAR model is trained using three preprocessed variables, including $TemperatureDeviation$. The VAR order is selected based on the Akaike Information Criterion (AIC) values of various trained models with different VAR orders, which are determined by dividing the dataset into training and test datasets. The model with the lowest AIC value is observed to have a VAR order of three, as shown in Figure 4. Therefore, the model is defined as a linear function of three past lags of itself and other variables. The following equation refers to the formula for the $TemperatureDeviation$ variable.

$$\begin{aligned} x_{TD}(t) = 0.095 &- 0.44 \times x_{ASWD}(t-1) + 0.10 \times x_{IPPU}(t-1) - 0.60 \times x_{TD}(t-1) \\ &+ 0.18 \times x_{ASWD}(t-2) + 0.021 \times x_{IPPU}(t-2) - 0.53 \times x_{TD}(t-2) \\ &+ 0.12 \times x_{ASWD}(t-3) - 0.41 \times x_{IPPU}(t-3) - 0.39 \times x_{TD}(t-3) + e \end{aligned}$$

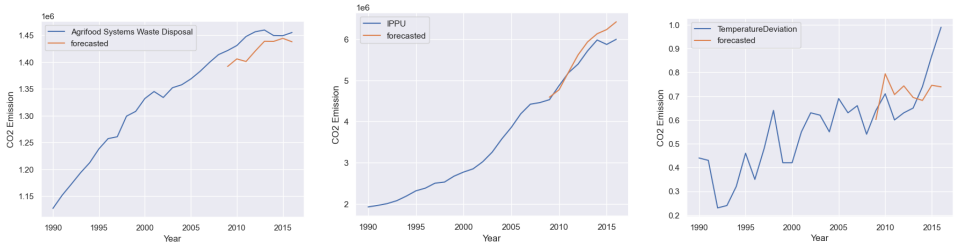
To assess the performance of the model, three different graphs are plotted for

VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	37.35	37.49*	1.666e+16*	37.35
1	37.45	38.02	1.921e+16	37.45
2	37.27	38.26	1.934e+16	37.26
3	37.09*	38.51	2.936e+16	37.08*

Figure 4: VAR order selection

each variable. These graphs show the original time series data from 1990 to 2016 and the forecasted data generated by the model from 2009 to 2016. As shown in Figure 5, the forecasted trends exhibit similarities to the original trends in terms of their characteristics and graph shape, indicating that the model is performing well. Therefore, it can be concluded that the model is reliable.



(a) Waste Disposal (b) IPPU (c) TemperatureDeviation

Figure 5: Model evaluation

Based on the previously trained model, the temperature deviation is forecasted for the next ten years, starting in 2017. As shown in Figure 6, the analysis indicates that there will be a significant increase in temperature deviation between 2017 and 2026.

4 Conclusion

The temperature deviation data shows that it has been increasing over time and has rapidly increased in recent years. This led to an investigation into the causes and effects of climate change. In this report, focusing on one of the primary factors of climate change, CO2 emissions of food production, analysis is conducted based on the data on agricultural food production and temperature deviation.

Correlation and time series analysis are conducted to explore the relationship

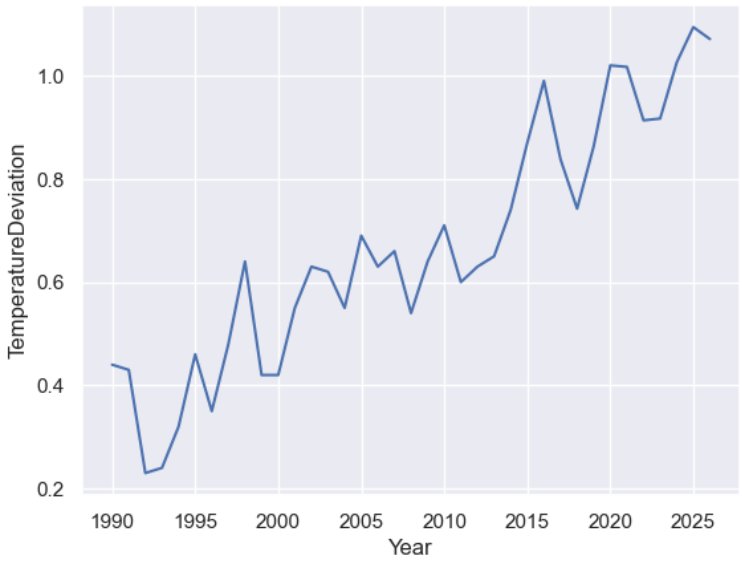


Figure 6: Forcasting the temperature deviation

between agricultural food production and temperature deviation and forecast the future temperature deviation. The selected datasets are preprocessed for reliable analysis. Through correlation analysis, the findings indicate that agricultural food production contributes significantly to climate change, as there are strong positive correlations between the variables from the agri-food CO emissions dataset and the TemperatureDeviation variable. Furthermore, based on the time series analysis, it is predicted that the temperature deviation will increase rapidly based on two major CO2 emission factors. Therefore, reducing CO2 emissions from the agricultural food production process, focusing on the factors that contribute the most to CO2 emissions, would help slow down climate change.

However, it should be noted that the report has a limitation due to an insufficient amount of data. Since the report deals with global data on an annual basis, it is difficult to obtain a larger dataset. This is due to the past limitations of technology, which resulted in limited records. It is challenging to obtain data from many years ago. Moreover, dividing the dataset into smaller training and test datasets reduces its size, resulting in a bad forecasting model with insignificant coefficients. Additionally, due to the small size of the dataset, the maximum VAR order the model could try is three. If data from previous years or in monthly units instead of annual is available, it would help improve the forecasting model and provide more reliable insights.

Ultimately, based on the long-term trends and fluctuations forecasted, it is important to undertake further work to minimise the impact of greenhouse gas emissions on climate change. This requires joint efforts not just at the local level,

but globally as well, ensure that the effects of climate change can be mitigated in the future.

References

[1] Ritchie, H. 2019. *Food production is responsible for one-quarter of the world's greenhouse gas emissions*. Our World in Data. [Online]. [Accessed 25 April 2024]. Available from: `https://ourworldindata.org/food-ghg-emissions`.

[2] Mersmann, K., et al. 2023. *Global Temperature Anomalies from 1880 to 2022*. NASA. [Online]. [Accessed 27 April 2024]. Available from: `https://svs.gsfc.nasa.gov/5060/#:~:text=Continuing%20the%20planet's%20long%2Dterm,Space%20Studies%20(GISS)%20reported`.

[3] *What Is Climate Change?*. United Nations. [Online]. [Accessed 27 April 2024]. Available from: `https://www.un.org/en/climatechange/what-is-climate-change`.

[4] Bello, A.L. 2023. *Agri-food CO2 emission dataset*. Kaggle. [Online]. [Accessed on 28 April 2024]. Available from: `https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml/data`.

[5] Gogtay, N.J., and Thatte, U.M. 2017. *Principles of Correlation Analysis*. The Journal of the Association of Physicians of India. 65, pp.78-81.

[6] Zivot, E., Wang, J. 2023. *Vector Autoregressive Models for Multivariate Time Series*. In: *Modeling Financial Times Series with S-Plus*. New York: Springer, pp.393-427.

[7] Kirchgassner, G., et al. 2013. *Introduction to Modern Time Series Analysis*. 2nd ed. Berlin: Springer, pp.127-153.