Predicting Future Global Temperature from Greenhouse Gas emissions

Marzieh Rasti

September 2024

Introduction

Global warming, marked by a sustained increase in average temperatures, leads to lasting changes in weather patterns and ecosystems. This project aims to predict the future interplay between greenhouse gas (GHG) emissions and global temperature changes. CO2, CH4, and N2O, identified by the European Commission as the most impactful gases on global warming (European Commission, 2018), are the focus of this study. By examining the relationships between these emissions and temperature fluctuations, the study seeks to provide predictive insights for formulating effective environmental policies.

Climate change, defined by long-term shifts in weather patterns, is significantly affecting life on Earth. Substantial evidence, such as the rise in ocean temperatures and the increased frequency of extreme weather events, underscores the reality of climate change. Addressing global warming is a critical challenge that demands accurate predictions of GHG emissions and temperature impacts, essential for developing strategies to mitigate its effects on a global scale.

Problem Statements

This project focuses on addressing three primary objectives:

1-Quantify the Relationship: Analyze the correlation between CO2, N2O, and CH4 emissions with global temperature variations to understand the relative impact of each GHG on climate change.

2-Trend Analysis: Identify and model trends in both GHG emissions and global temperature changes over the decades to forecast future climate conditions.

3-Predictive Modeling: Develop predictive models to forecast future global temperature changes based on current and hypothetical future GHG emission trends.

Data Source

For this project I identified three main data sources: NOAA, the Berkeley Earth , and Met office climate dashboard. The datasets provided are all publicly available.The dataset includes CO_2, CH_4, N_2O concentrations, and global temperature anomalies . Preprocessing steps included handling missing data, scaling the features, and splitting the dataset into training and test sets for model evaluation.

- **Temperature dataset**

Berkely is provided as a set of csv files,containing a detailed summary of the land-surface average results produced by the Berkeley Averaging method. Temperatures are in Celsius and reported as anomalies relative to the Jan 1951-Dec 1980 average. Uncertainties represent the 95% confidence interval for statistical and spatial undersampling effects.

- **GHG datasets:**

The Global Monitoring Laboratory has measured carbon dioxide and other greenhouse gasses for several decades at a globally distributed network of air sampling sites [Conway, 1994].The dataset includes the global monthly average of GHG.

Monthly nitrous oxide(N2O) concentration in the atmosphere as calculated by WDCGG.

- N2O expressed as a mole fraction in dry air, nanomol/mol, abbreviated as ppb-(1984-2022)
- CO2 expressed as a mole fraction in dry air, micromol/mol, abbreviated as ppm-(1979-2024)
- CH4 expressed as a mole fraction in dry air, nanomol/mol, abbreviated as ppb-(1983-2024)
- Temperature expressed as(°C)-(1880-2023)

Table 1: Final Dataset

	Year	Month	Anomaly	Uncertainty	CO2	CH4	N2O	Actual_Temperature
0	1984	1	0.291	0.055	344.32	1638.79	303.8	12.521
1	1984	2	0.145	0.054	344.82	1638.84	303.8	12.585
2	1984	3	0.292	0.042	344.96	1640.88	303.7	13.352
3	1984	4	0.178	0.070	345.19	1643.99	303.7	14.148
4	1984	5	0.388	0.043	345.33	1643.12	303.7	15.338
...	...	...	...	...	...	...	...	...
931	2022	8	0.830	0.034	414.41	1908.82	335.8	16.610
932	2022	9	0.756	0.034	414.63	1915.52	335.9	15.946
933	2022	10	0.871	0.034	416.14	1919.85	336.1	15.121
934	2022	11	0.654	0.048	417.77	1923.43	336.3	13.884
935	2022	12	0.775	0.050	418.80	1924.69	336.5	13.265

936 rows × 8 columns

Data Wrangling

The datasets are almost clean and without missing values. Below is an overview of the min issues I ran into while cleaning data:

- **Problem 1**: The GHG CSV files contain a lot of commented lines at the beginning, which are not part of the actual data. These comments are causing the ParserError because the CSV parser is expecting data and instead is encountering lines that don't match the expected format. To fix this issue, I want to skip these commented lines when reading the CSV file. The column names are not set properly as well. I had to manually set the column names based on the data file.
- **Problem 2:** Column names include special characters(. ,). I remove all space and characters, change the column name, and remove unnecessary columns.

EDA

In the EDA report, I was able to identify that the dataset will be sufficient to answer the problem statements.

- **Calculation Actual Temperature**

The temperature dataset contains a detailed summary of the land-surface average results produced by the Berkeley Averaging method.Temperatures are in Celsius and reported as anomalies relative to the Jan 1951-Dec 1980 average. Results are based on 40532 time series with 18975001 data points. To calculate the actual temperature from the anomaly, we need two

pieces of information: the anomaly value and the reference temperature. The reference temperature is the average temperature over a specific period, typically a 30-year period, often referred to as the baseline period. This period is used to establish a normal or average temperature for a specific location. The reference temperature can be obtained from historical temperature records or climate datasets. The Berkeley use the the "Estimated Jan 1951–Dec 1980 monthly absolute temperature" refers to the average temperatures for each month during the period from January 1951 to December 1980. These temperatures are estimated values representing the average absolute temperature (in degrees Celsius) for each month over this 30-year period.

Anomaly value represents the deviation of the actual temperature from the reference temperature. It can be either positive or negative, depending on whether the actual temperature is above or below the reference temperature. To calculate the actual temperature, I added the anomaly value to the reference temperature. The picture below shows reference temperature for the database.

Actual temperature = Reference temperature + Anomaly value

```
%
% Estimated Jan 1951–Dec 1980 monthly absolute temperature:
%      Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
%     12.23 12.44 13.06 13.97 14.95 15.67 15.95 15.78 15.19 14.25 13.23 12.49
% +/-  0.03  0.03  0.03  0.03  0.03  0.03  0.03  0.03  0.02  0.03  0.03  0.03
%
%   Using water temperature below sea ice:
%
% Estimated Jan 1951–Dec 1980 monthly absolute temperature:
%      Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
%     12.93 13.20 13.83 14.66 15.47 16.09 16.40 16.32 15.75 14.80 13.80 13.12
% +/-  0.03  0.03  0.02  0.02  0.03  0.03  0.02  0.02  0.02  0.02  0.02  0.02
%
```

Trend Visualization of Features Over Time

To explore the trends of greenhouse gas emissions (CO_2, CH_4, N_2O) and their relationship with global temperature anomalies over time, several visualizations were created(figure 1). These scatter plots provide insights into how the concentrations of these greenhouse gases have evolved and how global temperatures have responded.

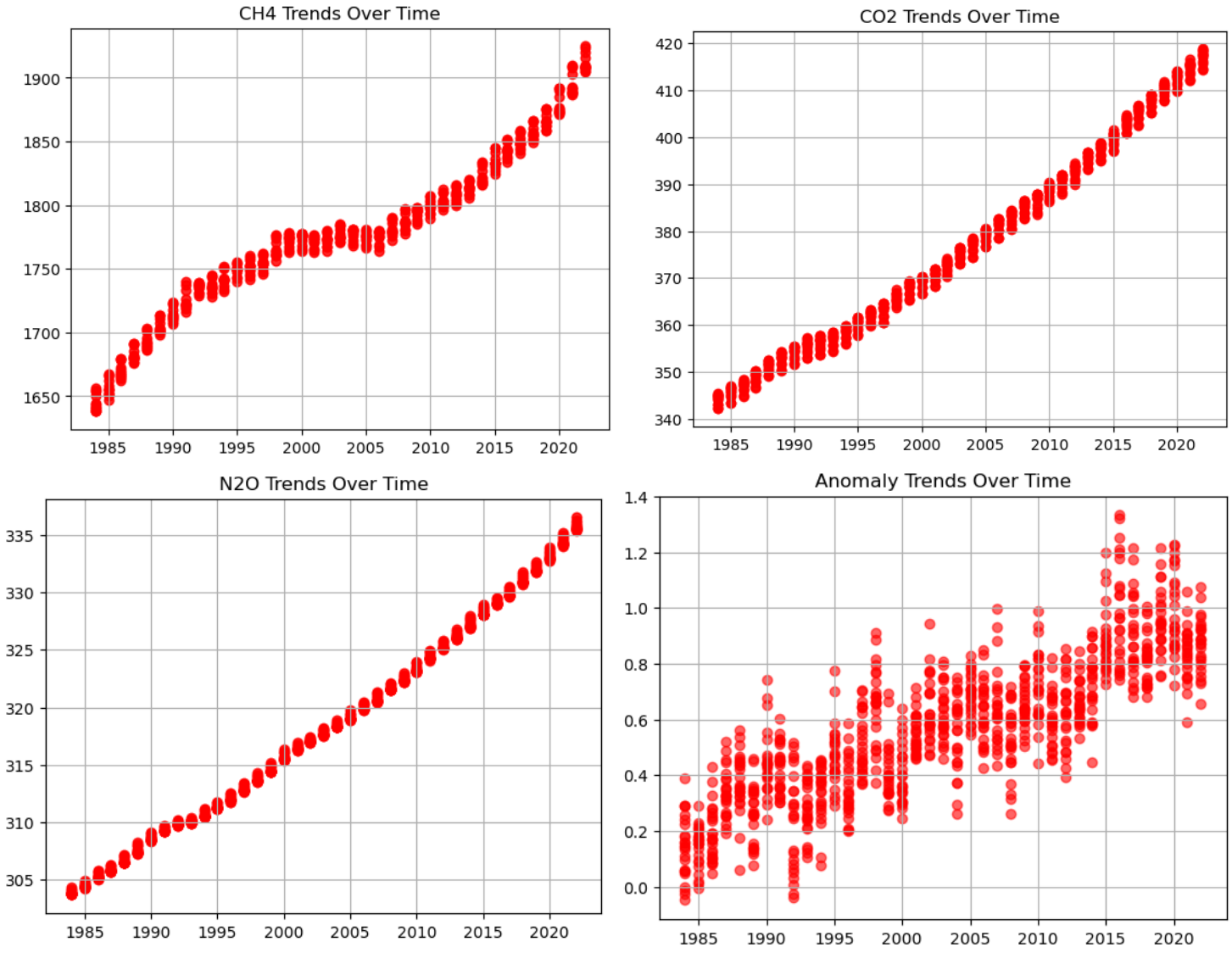


Figure1:GHG and Anomaly Temperature Trends Over Time

The consistent rise in CO_2, CH_4, and N_2O concentrations parallels the upward trend in temperature anomalies, suggesting a direct link between greenhouse gas emissions and global warming.The upward trends in GHG concentrations and temperature anomalies over the past few decades confirm the growing impact of human activities on the climate. These visualizations lay the foundation for further analysis and predictive modeling, which aims to quantify the relationship between GHG emissions and future global temperature changes.

Correlation Analysis

To further examine the relationships between the different GHGs and global temperature anomalies, a correlation analysis was performed using both heat maps and scatter plots.

- **Pearson Correlation**

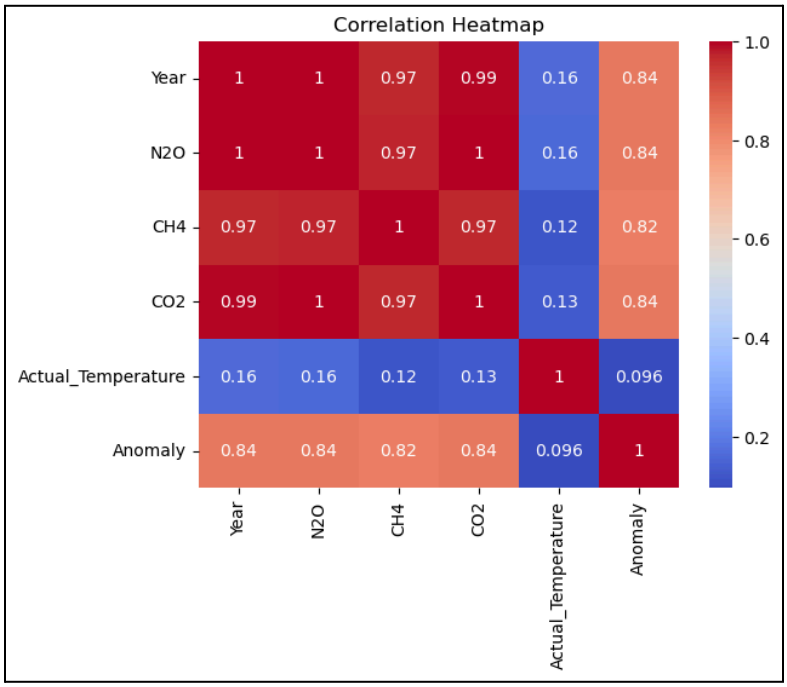


Figure2:Pearson Correlation Heatmap

CO_2 showed a strong positive correlation with both the year (0.99) and temperature anomalies (0.84), confirming the well-established link between increasing CO_2 levels and global warming. **CH_4** and **N_2O** also demonstrated high correlations with temperature anomalies, with correlation coefficients of 0.82 and 0.84, respectively.

Actual Temperature showed a weaker correlation with all GHGs, likely because temperature anomaly data more directly reflects deviations caused by GHG emissions.

The weak correlation observed between actual temperature and greenhouse gases (GHGs) may be influenced by several factors, including data aggregation, non-linear relationships, time lags, the influence of other climate variables, and potential measurement errors. These complexities in the climate system can dilute the direct relationship between GHG concentrations and actual temperature, making the correlation appear weaker.This is why, in climate research, anomalies—rather than actual temperatures—are often used. Anomalies represent deviations

from a long-term average, which helps to remove the effects of seasonal and geographic variability, providing a clearer signal of how GHGs influence temperature changes over time. By focusing on anomalies, we can better capture the underlying trends and relationships, leading to more accurate assessments of the impact of GHGs on global warming.

- **Spearman Correlation**

The Spearman correlation captures both linear and non-linear correlations by ranking the variables and assessing their monotonic relationships.

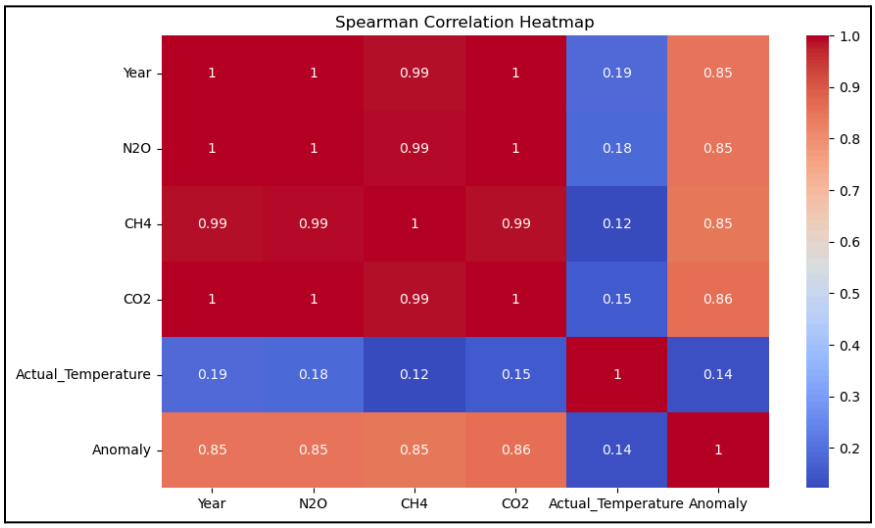
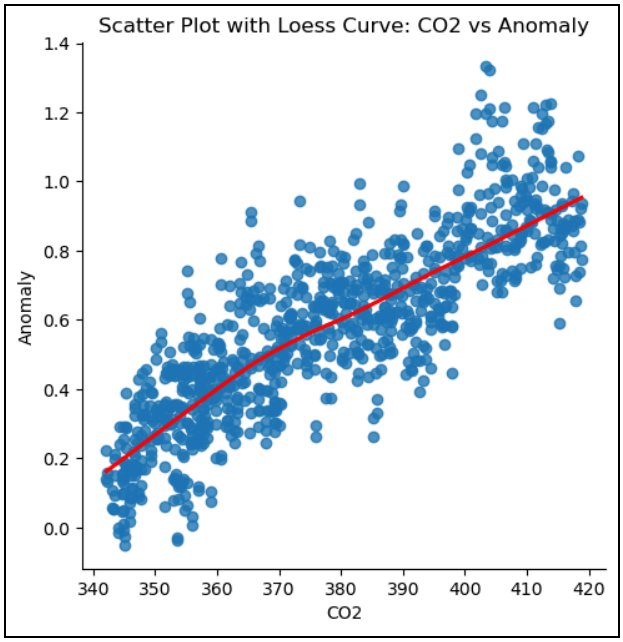


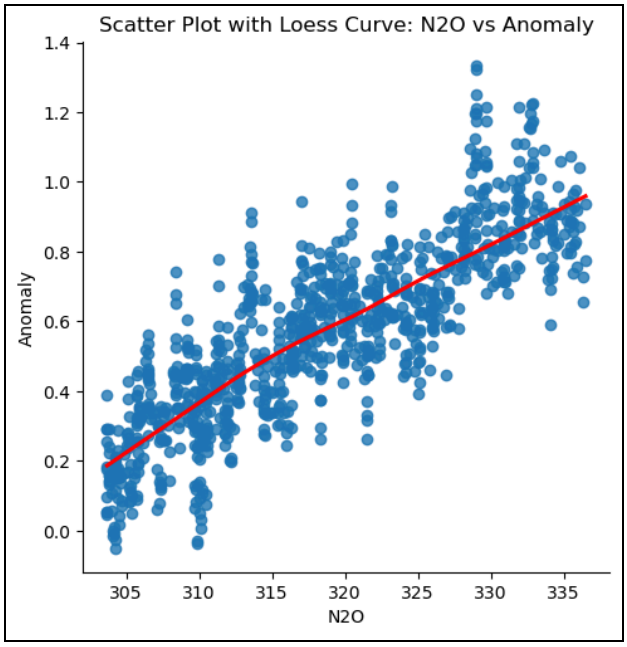
Figure3: Spearman Correlation Heatmap

The Spearman correlation values between GHGs and temperature anomalies were similar to the Pearson correlation values, confirming strong positive monotonic relationships.

- **Scatter Plots with Loess Curves**

  To visualize the non-linear relationships between individual greenhouse gases and temperature anomalies, scatter plots with Loess (figure4) curves were plotted. These curves provide a smoothed line that shows the general trend in the data, capturing any non-linearities.





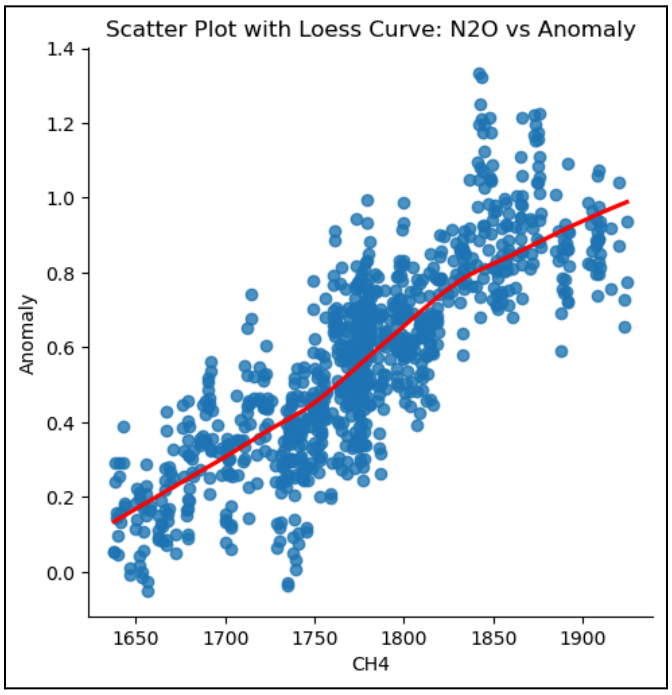


Figure4: Scatter Plots with Loess Curves

The plots indicate a positive relationship between GHG levels and temperature anomalies. As GHG concentrations increase, the temperature anomaly tends to increase as well. This suggests that higher GHG levels are associated with greater deviations from expected temperature patterns. The Loess curve (in red) indicates a non-linear relationship. The increase in anomaly accelerates as GHG levels rise beyond a certain point. This suggests that the impact of GHG on temperature anomalies becomes stronger at higher concentrations, potentially pointing to a compounding effect. Initially, at lower GHG levels, the rise in anomaly is more gradual, then the increase becomes steeper. These findings support the predictive modeling efforts to forecast future temperature changes based on GHG trends.

Feature Scaling

After splitting data into training and test sets, Given that the features (CO2, CH4, N2O, and Temperature) have different units and ranges, I should standardize or normalize the data.Given that the features (especially CO2 and CH4) are not perfectly normally distributed and have skewness, **standardization** would generally be the better option.

Feature Selection

Since the GHGs (CO_2, CH_4, N_2O) have strong correlations with temperature anomalies, it makes sense to include these variables as key features in the models.

Machine learning models

This is a regression problem in supervised learning. Due to the presence of multicollinearity between features, I have chosen models that are robust to multicollinearity and more complex models that can be useful in capturing subtle relationships and interactions between the features. The non-linear patterns observed in the Loess plots further support the use of flexible models capable of handling non-linearity, such as Random Forest, Gradient Boosting, and Polynomial Regression. Additionally, tree-based models like Random Forest and XGBoost are inherently robust to multicollinearity as they handle feature selection implicitly during the model-building process.

For these reasons, the models chosen for this project include:

1. Random Forest
2. XGBoost Regression
3. Time Series Models
4. Polynomial Regression

Hyperparameter Tuning

GridSearchCV was used for hyperparameter tuning in XGBoost regression model and Cross-validation was used to assess different polynomial degrees (1 to 5), and degree 3 was selected as the best-performing option based on the R^2 score.

Evaluation Metrics

Mean Squared Error (MSE), R^2 Score, and Cross-Validation Scores was used to evaluate Model performance.

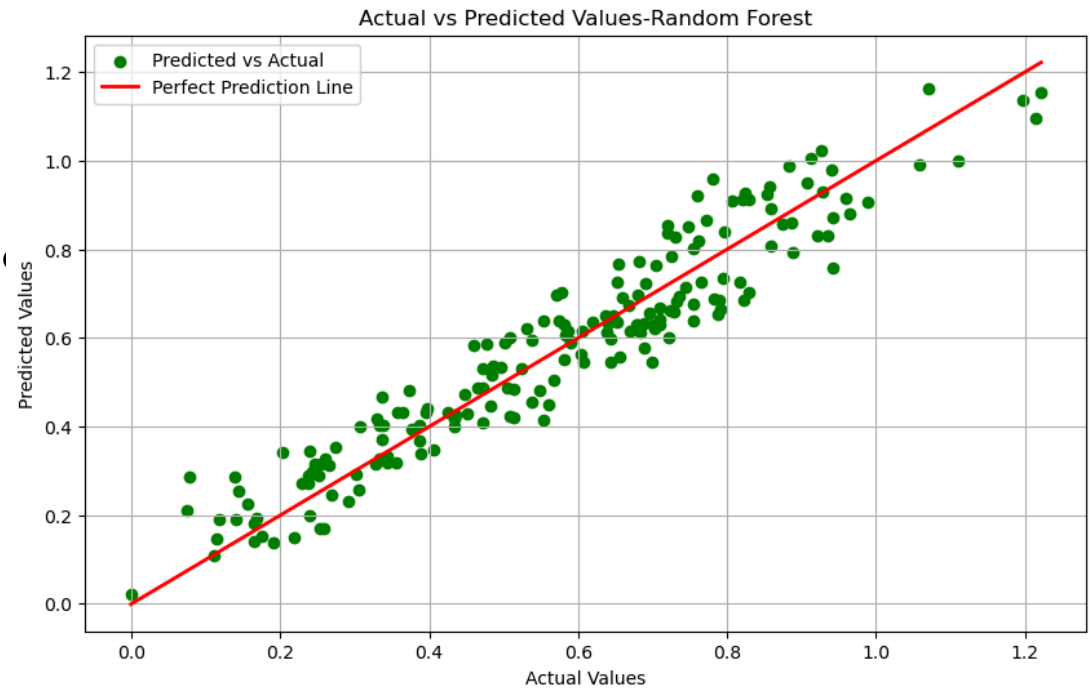


Figure5: Evaluation Random Forest Model based on the Actual vs predicted values

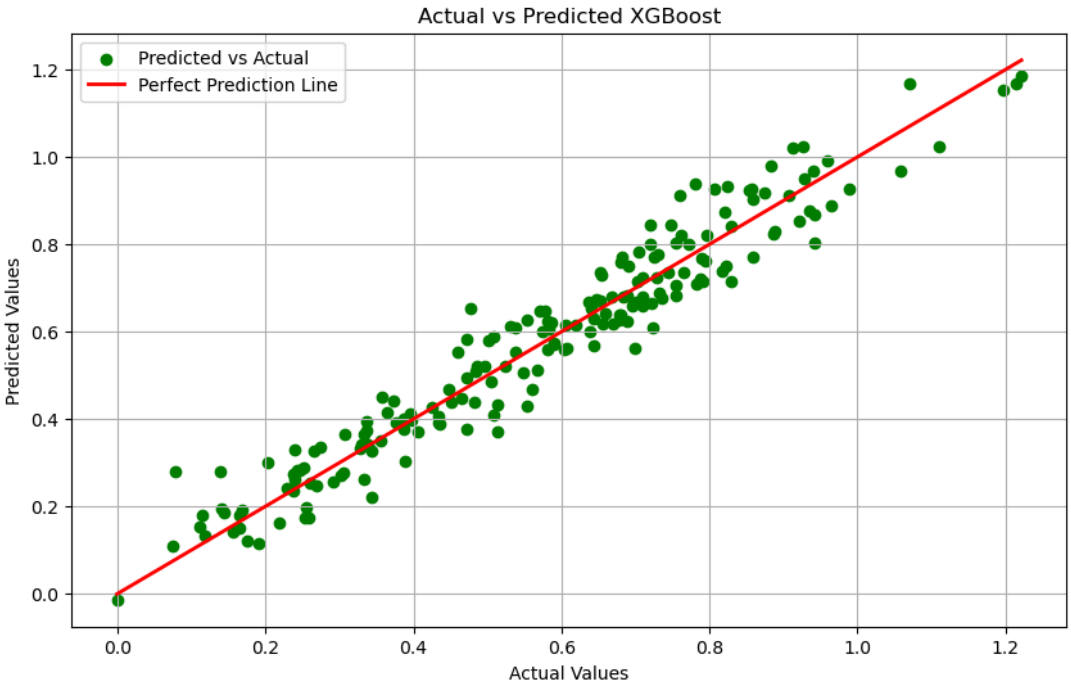


Figure6: Evaluation XGBoost Regression Model based on the Actual vs predicted values

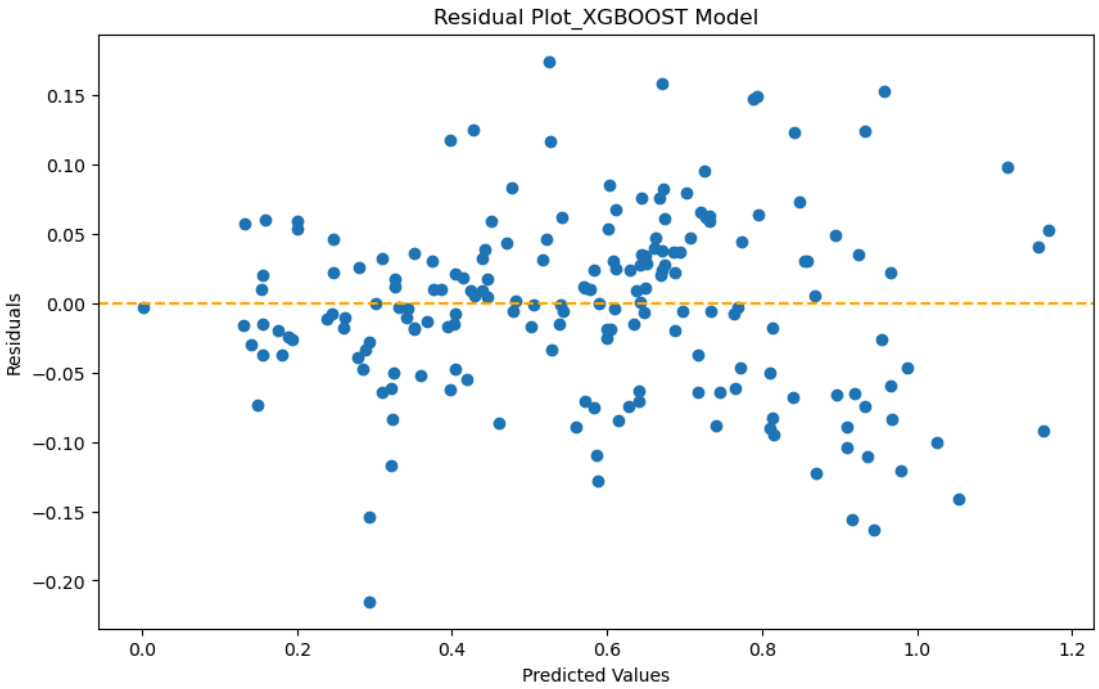


Figure7: Residual Analysis-XGBoost Regression

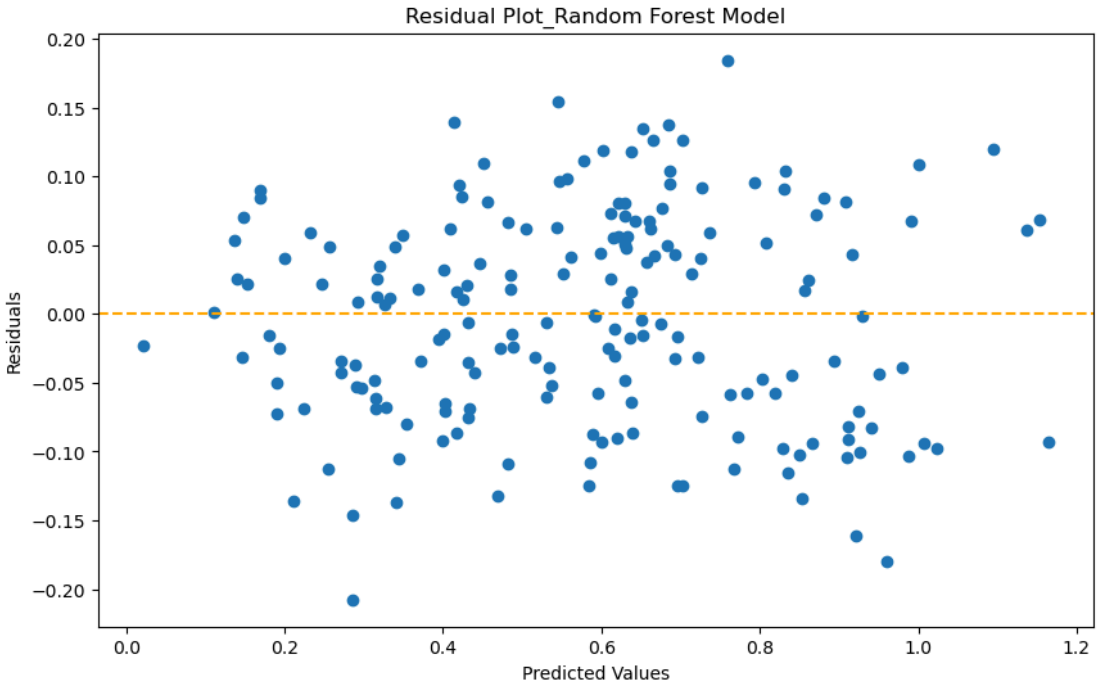


Figure8: Residual Analysis-Random Forest

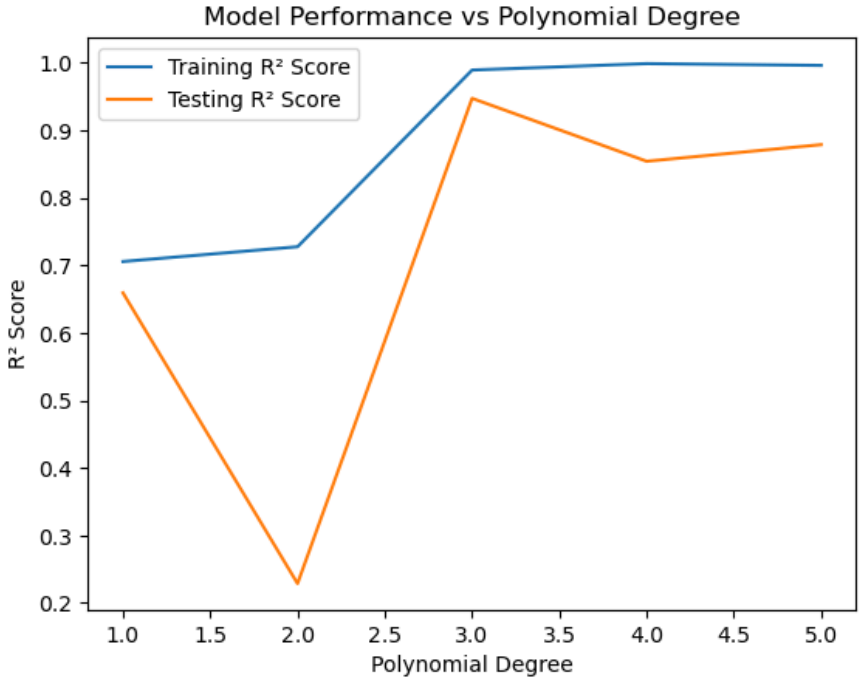


Figure9: Cross-Validation to Choose the Optimal Degree

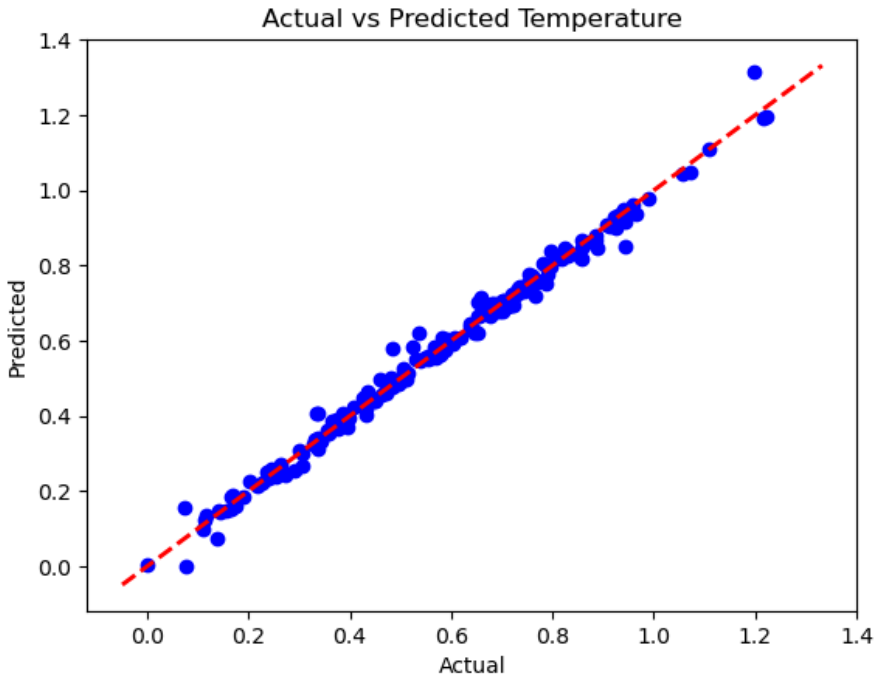


Figure10: Evaluation Polynomial Model based on the Actual vs predicted values

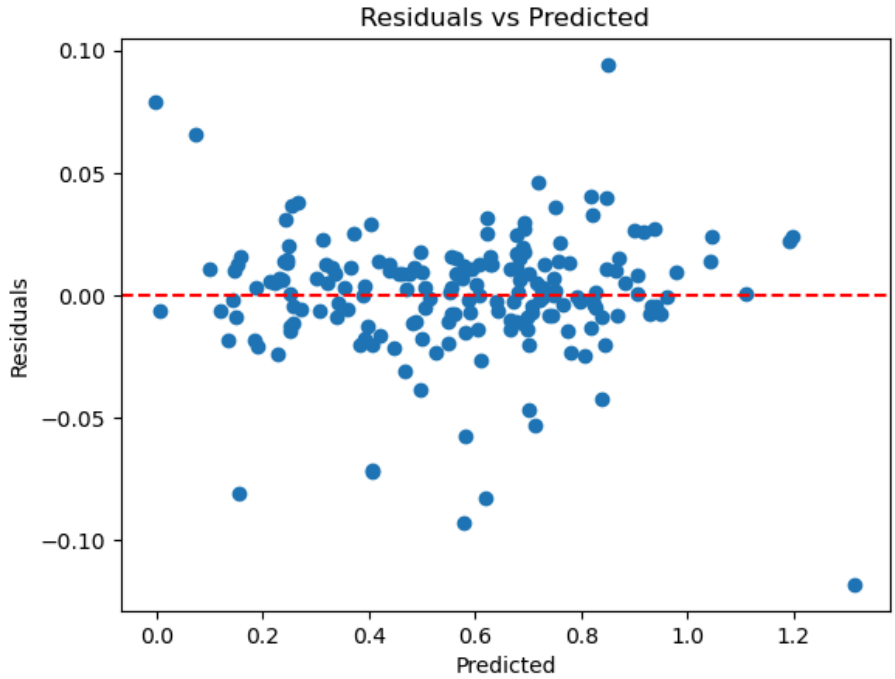


Figure11: Residual Analysis-Polynomial Model

Comparison and Model Selection

I applied different ML models and evaluated their performances using cross-validation for both the training and test data. Table 2 shows the results of the scores.

Table 2: Model performance metrics

MODEL	MSE	R^2
Random Forest	0.005865	0.909581
XGBoost	0.004220	0.934952
Polynomial Regression	0.000637	0.990175

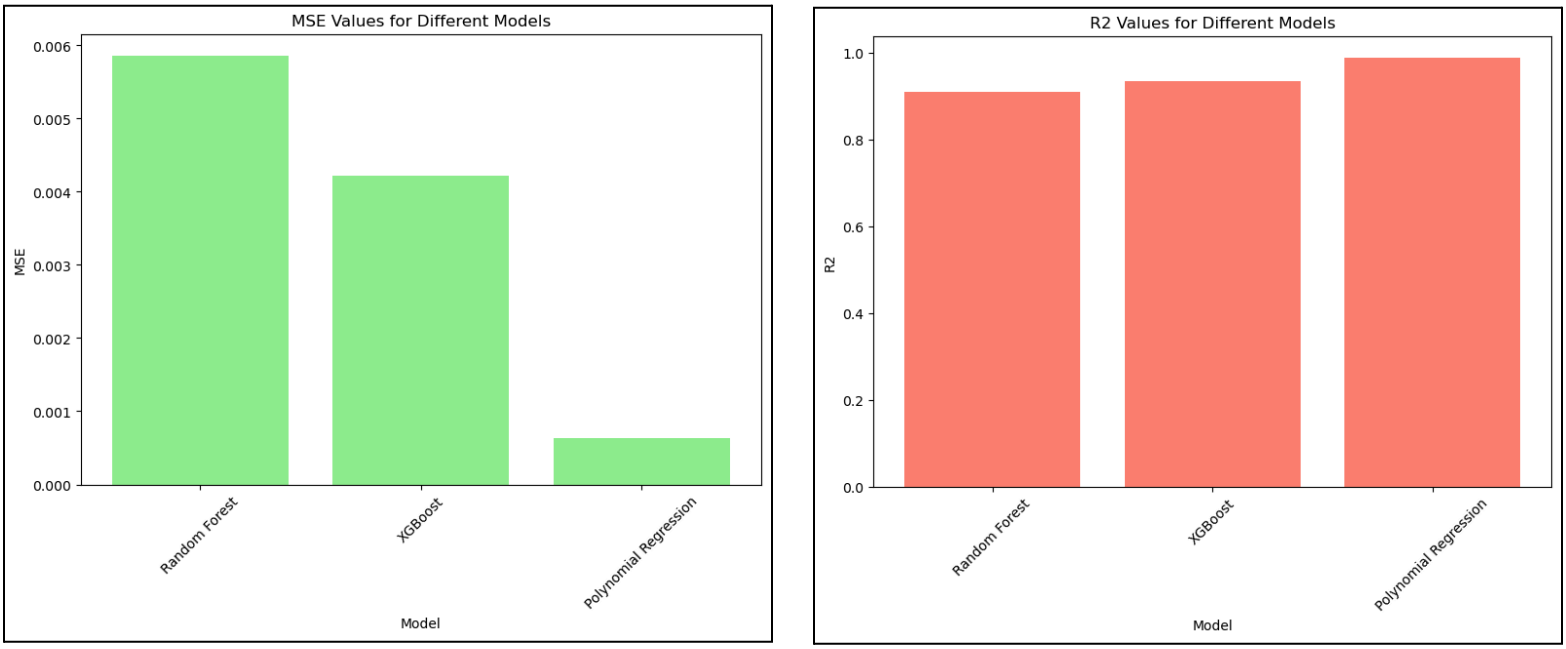


Figure12: Evaluation Model performance

Based on the results of comparison models, the **polynomial** model has higher R^2 scores on both the training and test sets are very close to each other (0.99 on both), which indicates that the model generalizes well and is not overfitting.

Time Series Models

Time series models were applied to identify and model trends in both GHG emissions and global temperature changes over the decades to forecast future climate conditions. Given the seasonality present in the data, SARIMA was chosen as a suitable time series forecasting technique for estimating data with seasonal patterns. However, the results indicated that the SARIMA model did not perform well in predicting temperature anomalies (as shown in Plot 3). To improve the model, I tested the SARIMAX model, which incorporates exogenous variables like greenhouse gas emissions to account for external influences on the response variable (temperature anomalies). These exogenous variables were expected to enhance the forecast's accuracy by providing information that SARIMA alone could not capture. Unfortunately, the SARIMAX model also failed to yield accurate predictions on the dataset. As a result, I decided to implement Polynomial Regression as the final model, which was more appropriate given the non-linear relationship between GHG emissions and temperature anomalies.

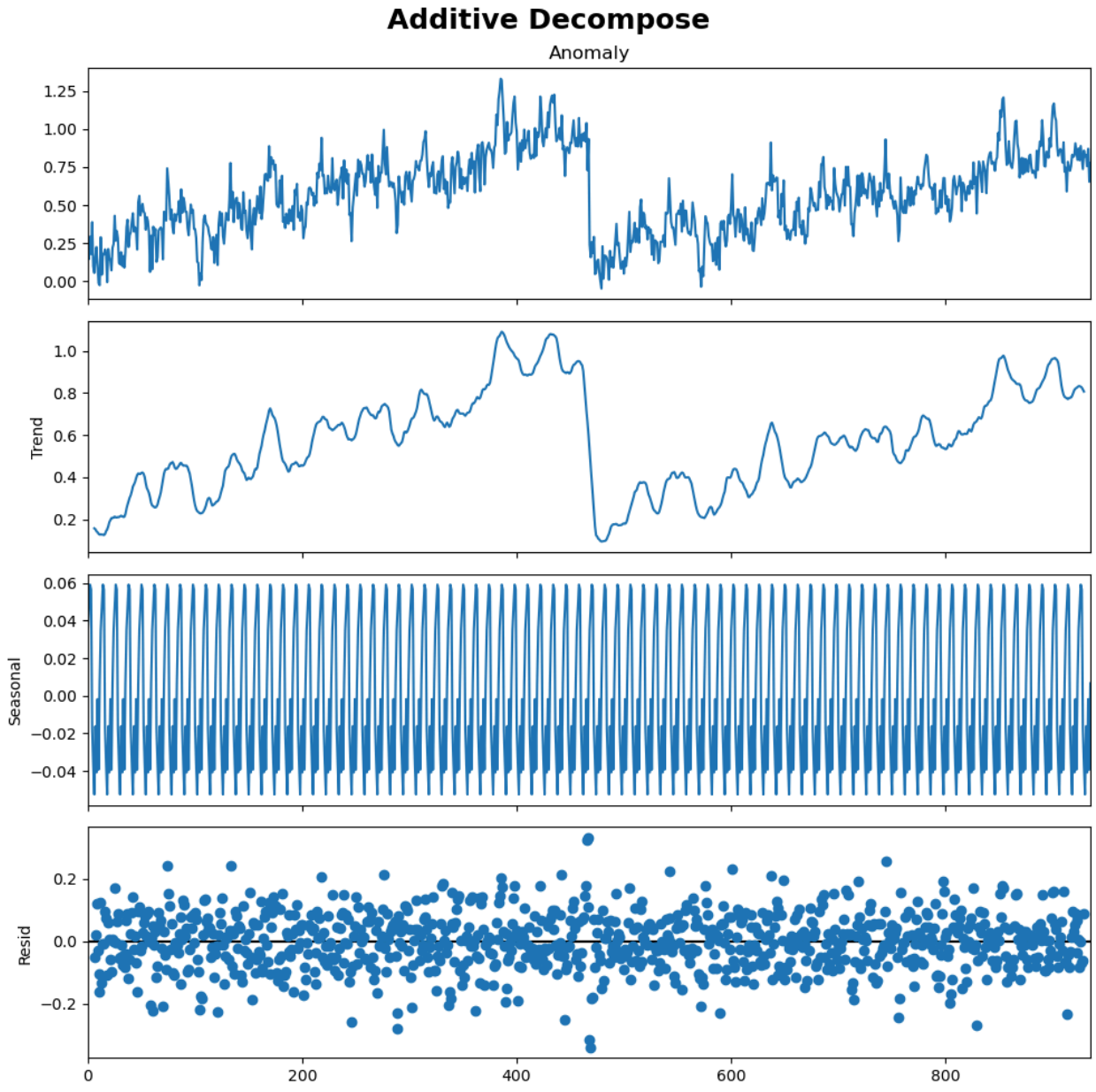


Figure13: Time series Decomposition

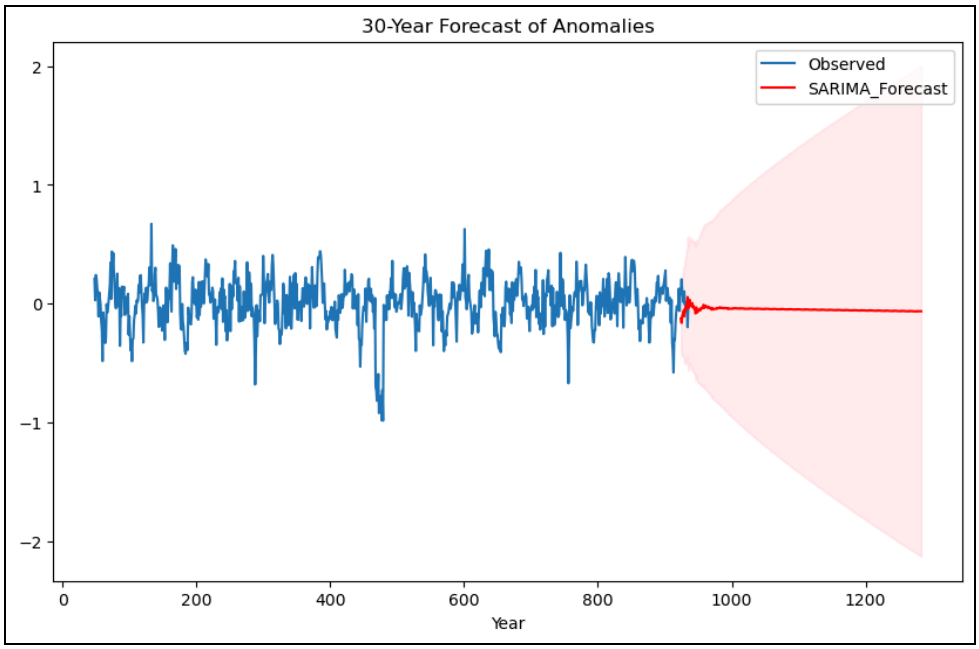


Figure14: Forecasting Future Values-SARIMA

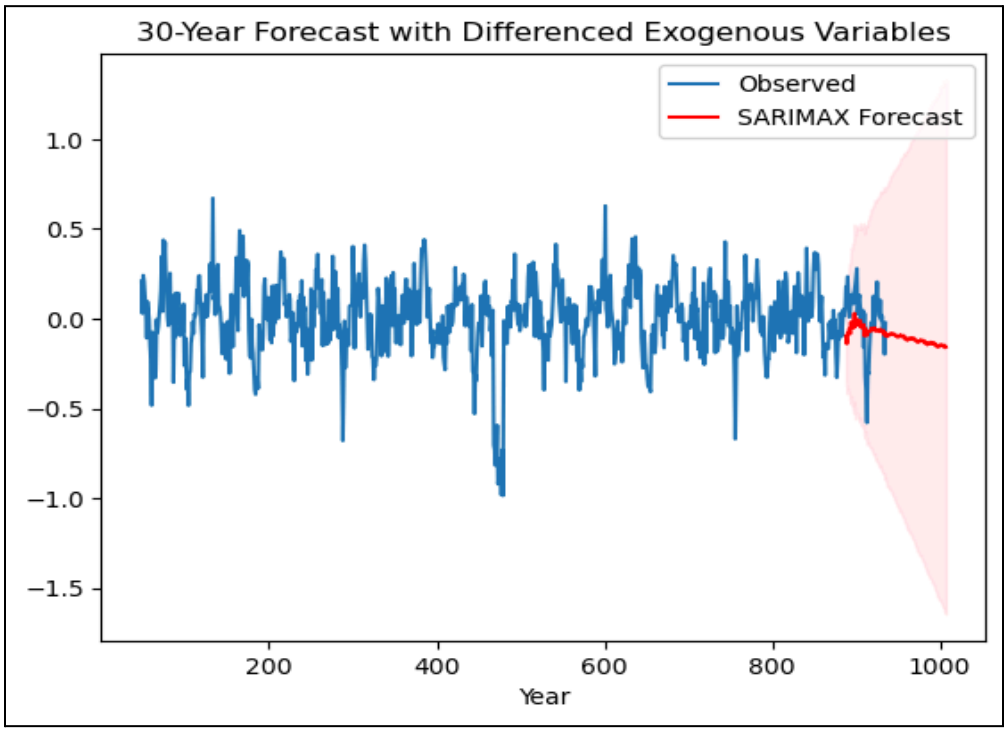


Figure15: Forecasting Future Values-SARIMA

Conclusion

The goal of this project was to predict future global temperature changes by analyzing the relationship between greenhouse gas (GHG) emissions—specifically CO2, CH4, and N2O—and temperature anomalies. The initial stages involved conducting a comprehensive correlation analysis to determine the strength and nature of the relationships between these variables. The analysis revealed a significant positive correlation between CO2 and temperature anomalies, with CH4 and N2O also showing moderate correlations. This indicated that increasing GHG emissions were directly linked to rising global temperatures. Building upon this analysis, time series models such as SARIMA and SARIMAX were employed to capture temporal trends and seasonality in the data. However, these models did not perform well in predicting temperature anomalies. The inclusion of exogenous variables in the SARIMAX model slightly improved performance, but it still fell short of providing accurate predictions, likely due to the non-linear and complex interactions between GHGs and temperature.

To address these challenges, Polynomial Regression was used as the final predictive model. This approach captured the non-linear relationship between GHG emissions and temperature anomalies more effectively. The Polynomial Regression model significantly outperformed the time series models, with a higher R-squared value and lower MAE and RMSE scores, reflecting better alignment between predicted and observed temperature anomalies.

While the project successfully demonstrated the strong correlation between GHG emissions and temperature changes, the model results also highlighted the complexity of predicting global temperature based solely on these variables. There is room for future improvement by incorporating additional climatic factors, refining feature engineering, and exploring advanced machine learning methods such as neural networks.

In summary, this project confirmed the critical role of GHG emissions in driving global temperature increases. Although Polynomial Regression provided reasonable predictive power, further refinements—such as adding more variables, conducting scenario analysis, and utilizing higher-resolution data—would enhance the model's accuracy and provide more actionable insights for environmental policy-making and climate change mitigation efforts.

Future Improvement

1- **Use Higher Frequency Data:** using daily or monthly data instead of annual data could help capture more detailed patterns in emissions and temperature fluctuations. This could provide a finer-grained model capable of making more accurate short-term predictions.

2- **Obtain More Data:** extend the dataset to include more years or other sources of historical climate data. A more extensive dataset could improve model robustness and predictive accuracy.

3- **Incorporate Additional Variables:** including other greenhouse gases or climate factors (such as aerosols, solar radiation, or ocean temperatures) that may influence global temperature changes. This would provide a more comprehensive model for predicting temperature.

4- **Explore Deep Learning Models:** Considering implementing recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks, which are powerful for modeling temporal sequences and can capture long-term dependencies in time series data.

5- **Sensitivity Analysis:** Conducting sensitivity analysis to understand how changes in specific GHGs (e.g., CO2, CH4) affect temperature predictions. This could help prioritize which variables have the most significant impact on global temperature.