Analysis Report

Methods of Advanced Data Engineering (MADE), SoSe-2024

Redwanul Karim - 23426184

Faculty of Engineering
Friedrich-Alexander University of Erlangen-Nuremberg

1 Introduction

The project aims to explore the possible connection between solar activity - like solar flares, and climate change on Earth. Through the statistical analysis of historical data, our goal is to uncover whether there exists a clear relationship between solar events and the observable shift in global climate patterns. This investigation holds significant importance as it provides insights essential for refining climate models and improving our capacity to forecast and address the consequences of climate change.

2 Research Question

"Is there any relationship between solar activity (Solar Flares) and climate change on Earth?"

3 Data Pipeline

In order to answer our research question, we obtain the final dataset for analysis by implementing an ETL (Extract, Transform, Load) data pipeline. This pipeline systematically downloads, processes, and stores data from various sources, ensuring it is clean, structured, and ready for analysis. The following figure is a conceptual diagram of the Data Pipeline.

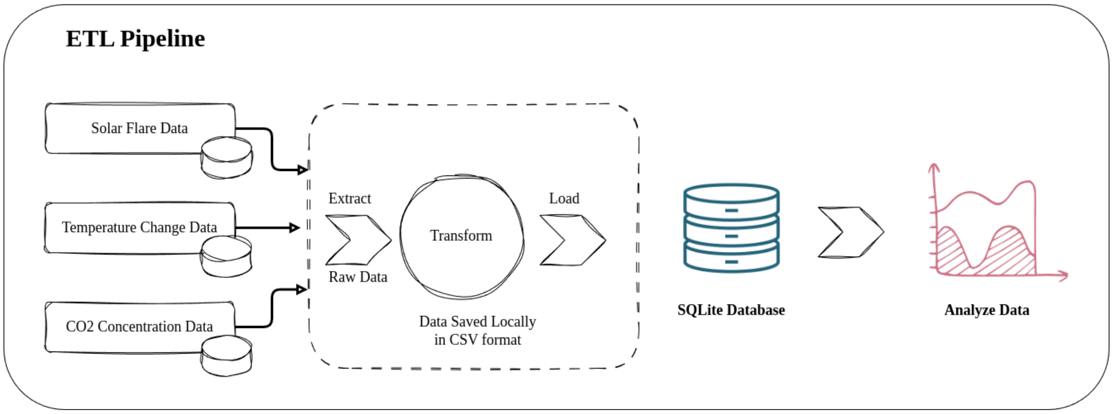

Figure 1: ETL Data Pipeline

4 Data Used

The analysis utilizes data from three primary datasets obtained from the output of a data pipeline depicted in the figure [1] above. Here's a description of data used during analysis:

4.1 CO2 Concentration Data

4.1.1 Features

Feature	Description
Date	Date of the measurement.
CO2_Concentration_PPM	CO2 concentration in parts per million (PPM).

Table 1: CO2 Concentration Data Features

4.1.2 Structure and Meaning

This dataset tracks the concentration of carbon dioxide (CO_2) in Earth's atmosphere over time. CO_2 concentration is a critical indicator of greenhouse gas levels, directly influencing global climate patterns and temperature trends.

4.2 Temperature Change Data

4.2.1 Features

Feature	Description
Date	Date of the measurement.
Temp_Change	Temperature change or anomaly.

Table 2: Temperature Change Data Features

4.2.2 Structure and Meaning

This dataset records changes in temperature over time, providing insights into climate variability and trends. Temperature changes are essential indicators of climate change impacts and can be influenced by various factors, including solar activity.

4.3 Solar Flare Data

4.3.1 Features

Feature	Description
Date	Date of the observation.
FlareNumber	Identifier for the solar flare event.
TOTUSJH	Total unsigned current helicity.
TOTBSQ	Total unsigned magnetic flux.
TOTPOT	Total photospheric magnetic free energy.
TOTUSJZ	Total unsigned current helicity in the photosphere.
ABSNJZH	Absolute value of the net vertical electric current helicity.
SAVNCPP	Sum of the absolute value of net current per polarity.
USFLUX	Net unsigned magnetic flux.

Table 3: Solar Flare Data Features

4.3.2 Structure and Meaning

These parameters are measurements related to solar activity, specifically focusing on magnetic flux and helicity. They are crucial in understanding the dynamics and energy release of solar flares, which can impact Earth's atmosphere and potentially influence climate variables.

5 Analysis

5.1 Determining the Common Time Window

To integrate the Solar Flare Data, CO2 Concentration Data, and Temperature Change Data, we first determined the common time window across all datasets, which is: 2010-05-01 to 2019-12-31.

5.2 Calculate Feature Importance

To identify the top 5 significant features [Figure-2] from the Solar Flare dataset, we utilized the Random Forest Regressor:

- **Features Identified:** 'TOTBSQ', 'TOTPOT', 'ABSNJZH', 'SAVNCPP', and 'USFLUX'.

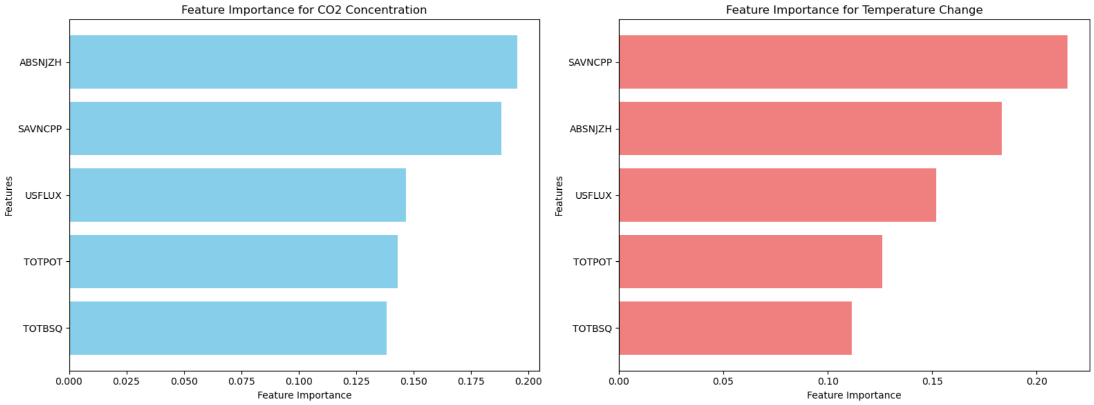

Figure 2: Five most important feature of Solar Flare data in relation to Temperature Change and CO_2 Concentration on Earth

5.3 Hypothesis Testing

We performed hypothesis testing using t-tests to examine the influence of Solar Flare Activity on Temperature Change and CO2 Concentration:

5.3.1 Temperature Change

The null hypothesis H_0 tested whether there is no significant difference in mean temperature change between periods with different levels of Solar Flare Activity.

$$t_{temp} = \frac{\bar{X}_{high} - \bar{X}_{low}}{\sqrt{\frac{s^2_{high}}{n_{high}} + \frac{s^2_{low}}{n_{low}}}}$$

where $\bar{X}_{high}, \bar{X}_{low}$ are the sample means of temperature change during periods of high and low Solar Flare Activity, s_{high}, s_{low} are the corresponding sample standard deviations, and n_{high}, n_{low} are the sample sizes.

5.3.2 CO2 Concentration

Similarly, we tested the null hypothesis H_0 for CO2 concentration:

$$t_{CO2} = \frac{\bar{X}_{high} - \bar{X}_{low}}{\sqrt{\frac{s^2_{high}}{n_{high}} + \frac{s^2_{low}}{n_{low}}}}$$

where $\bar{X}_{high}, \bar{X}_{low}$ are the sample means of CO2 concentration during periods of high and low Solar Flare Activity, s_{high}, s_{low} are the corresponding sample standard deviations, and n_{high}, n_{low} are the sample sizes.

5.3.3 Interpretation

For both temperature change and CO2 concentration, we obtained significant p-values ($p < 0.05$), indicating strong evidence against the null hypothesis. This suggests that Solar Flare Activity is statistically associated with changes in both temperature and CO2 concentration. The results imply that periods of increased solar activity may influence Earth's climate variables.

6 Result

The hypothesis testing (t-tests) conducted to explore the relationship between solar flare events and climate change on Earth yielded statistically significant p-values for both CO_2 concentration and temperature change. However, upon deeper examination of the data distribution and the autocorrelation of variables, significant violations of the underlying assumptions of normality and homogeneity of variances were observed. These violations raise substantial concerns

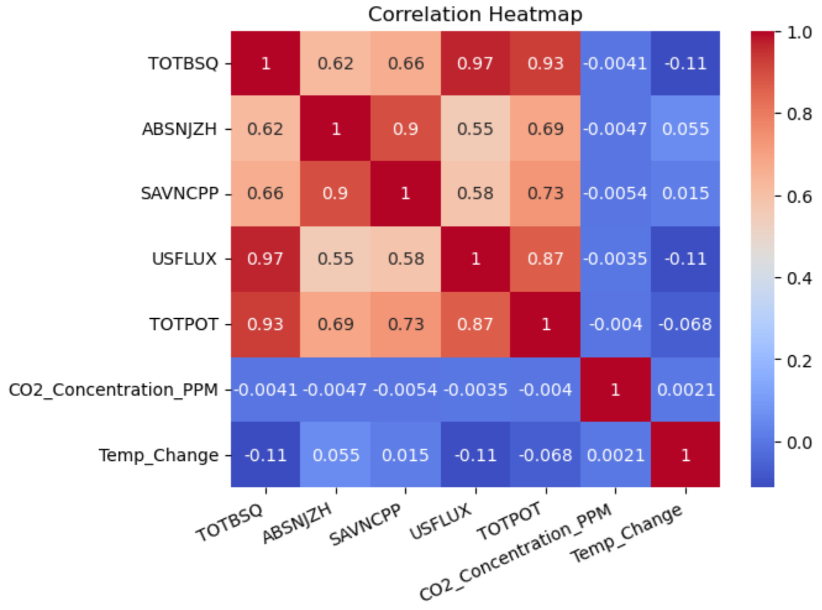

Figure 3: Correlation Coefficient Heat Map of the data features

regarding the reliability and validity of our hypothesis testing results. The correlation coefficient heat-map also suggests a similar concern regarding the hypothesis testing as depicted in the Figure-3.

The assumption of normality is crucial for t-tests, as it ensures that the sample means and variances accurately reflect the population parameters. Our analysis indicated that the CO_2 concentration and temperature change data did not meet the requirements of normal distribution, potentially biasing our hypothesis testing outcomes. Moreover, the presence of autocorrelation among the data points further complicates the interpretation of results, as it implies that observations are not independent, a fundamental assumption of classical statistical tests.

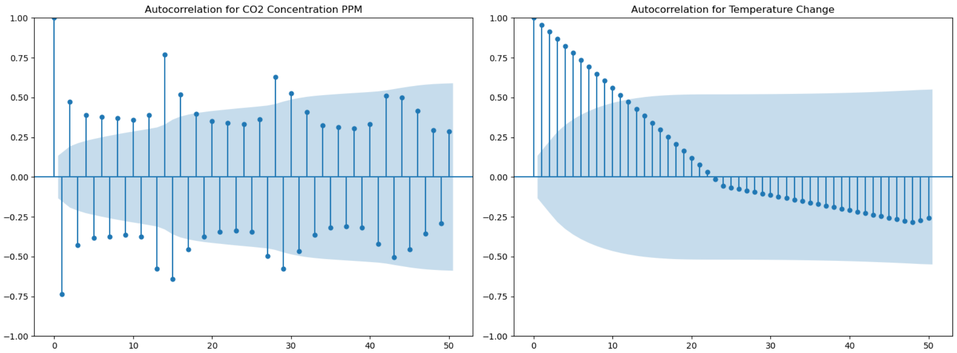

Figure 4: Auto-correlation plot of Temperature Change and CO_2 Concentration data

Therefore, while the obtained p-values suggest a statistically significant relationship between solar flare events and changes in CO_2 concentration and temperature, these findings should be interpreted with caution. The observed violations of normality and autocorrelation indicate that our results may not accurately reflect true causal relationships. At this stage, it is not possible to definitively conclude that solar flare activity influences climate change on Earth based solely on the results of our hypothesis testing.

7 Limitations and Future Direction

While hypothesis testing offers useful information about statistical correlations, it does not prove cause and effect. Additionally, the violations of assumptions like normality and equal variances mean we should be careful with our interpretations. We need to explore domain-specific machine learning algorithms to handle the complexities of non-normally distributed data in order to better understand the influence of solar activity and climate change on Earth.