Statistical aspects of Global Warming dynamics

Andrei Moga, Isabel Opris and Teodor Spiridon

December 2024

1 Introduction

The average surface temperature of Earth has been rising over time, mainly due to human activities that contribute to the emission of greenhouse gases like carbon dioxide, methane, and nitrous oxide. This phenomenon is known as global warming. Conceptually, this notion was described by Joseph Fourier in the 1820s as the "Greenhouse Effect" [2]. As a consequence, this warming also leads to more significant shifts in weather patterns, signaled by greater differences in temperatures.

In 1856, Eunice Newton Foote demonstrated that carbon dioxide could significantly raise air temperature, and in 1859 John Tyndall's experiments confirmed that gases such as carbon dioxide and water vapor absorb heat, contributing to atmospheric warming. Building on these findings, Svante Arrhenius in 1896 calculated that doubling atmospheric carbon dioxide could increase global temperatures by approximately 5–6°C. Despite early skepticism, the mid-20th century saw accumulating evidence of rising carbon dioxide levels and global temperatures, leading to the establishment of the Intergovernmental Panel on Climate Change (IPCC) in 1988. The IPCC has since provided comprehensive assessments, reinforcing the scientific recognition that human activities are the primary drivers of recent global warming.

Although the effects of global warming are increasingly evident in our daily lives, there remain communities that continue to deny its existence and impact. This study, therefore, seeks to validate the reality of global warming by analyzing long-term temperature trends and the rising frequency of extreme weather events. Additionally, it will examine the correlation between land and ocean temperatures to determine whether they influence each other.

2 Data

We have utilized data from the Berkeley Earth Surface Temperature Study, a comprehensive dataset compiled by Berkeley Earth in collaboration with the Lawrence Berkeley National Laboratory. This dataset integrates 1.6 billion temperature reports sourced from 16 pre-existing archives. It offers a robust framework for analysis, with the ability to slice data into subsets such as country-specific information.

The "Climate Change: Earth Surface Temperature Data" dataset was taken from Kaggle [1], having a usability score of 7.65. The dataset contains five data files. From these data files, we have used two of them: "GlobalLandTemperaturesByCity.csv" and "GlobalTemperatures.csv." In the following sections, we are going to describe how the data was studied, cleaned, and prepared for its use.

2.1 Data Study

We will begin by studying the previously mentioned files: "GlobalLandTemperaturesByCity.csv" and "GlobalTemperatures.csv".

2.1.1 "GlobalLandTemperaturesByCity.csv" file

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8599212 entries, 0 to 8599211
Data columns (total 7 columns):
 #   Column                         Dtype
---  ------                         -----
 0   dt                             object
 1   AverageTemperature             float64
 2   AverageTemperatureUncertainty  float64
 3   City                           object
 4   Country                        object
 5   Latitude                       object
 6   Longitude                      object
dtypes: float64(2), object(5)
memory usage: 459.2+ MB
```

Data types contained in the "GlobalLandTemperaturesByCity.csv" file are: `dt` refers to the date, `AverageTemperature` refers to the average temperature (in Celsius) taken on the date specified in `dt`, `AverageTemperatureUncertainty` refers to the 95% confidence interval around the `AverageTemperature`, `City` refers to the city from which the temperature was taken, `Country` refers to the country of the specified city and finally, `Latitude` and `Longitude` refer to the coordinates of the city from which the temperature was taken.

We follow up with a preview of the first rows from our file:

```
           dt  AverageTemperature  AverageTemperatureUncertainty   City  \
0  1743-11-01               6.068                          1.737  Århus
1  1743-12-01                 NaN                            NaN  Århus
2  1744-01-01                 NaN                            NaN  Århus
3  1744-02-01                 NaN                            NaN  Århus
4  1744-03-01                 NaN                            NaN  Århus

   Country Latitude Longitude
0  Denmark   57.05N    10.33E
1  Denmark   57.05N    10.33E
2  Denmark   57.05N    10.33E
3  Denmark   57.05N    10.33E
4  Denmark   57.05N    10.33E
```

We check for missing values:

```
dt                                   0
AverageTemperature              364130
AverageTemperatureUncertainty   364130
City                                 0
```

```
Country                                0
Latitude                               0
Longitude                              0
dtype: int64
```

We observe that our file contains 364,130 empty values from both `AverageTemperature` and `AverageTemperatureUn` which we will get rid of during the data cleaning stage.

2.1.2 "GlobalTemperatures.csv" file

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3192 entries, 0 to 3191
Data columns (total 9 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   dt                                         3192 non-null   object
 1   LandAverageTemperature                     3180 non-null   float64
 2   LandAverageTemperatureUncertainty          3180 non-null   float64
 3   LandMaxTemperature                         1992 non-null   float64
 4   LandMaxTemperatureUncertainty              1992 non-null   float64
 5   LandMinTemperature                         1992 non-null   float64
 6   LandMinTemperatureUncertainty              1992 non-null   float64
 7   LandAndOceanAverageTemperature             1992 non-null   float64
 8   LandAndOceanAverageTemperatureUncertainty  1992 non-null   float64
dtypes: float64(8), object(1)
memory usage: 224.6+ KB
```

Data types contained in the "GlobalTemperatures.csv" file are: `dt` refers to the date, `LandAverageTemperature` refers to the global average land temperature in Celsius, `LandAverageTemperatureUncertainty` refers to the 95% confidence interval around the average, `LandMaxTemperature` refers to the global average maximum land temperature in Celsius, `LandMaxTemperatureUncertainty` refers to the 95% confidence interval around the maximum land temperature, `LandMinTemperature` refers to the global average minimum land temperature in Celsius, `LandMinTemperatureUncertainty` refers to the 95% confidence interval around the minimum land temperature, `LandAndOceanAverageTemperature` refers to the global average land and ocean temperature in Celsius, and finally, `LandAndOceanAverageTemperatureUn` refers to the 95% confidence interval around the global average land and ocean temperature.

We follow up with a preview of the first rows from our file:

	dt	LandAverageTemperature	LandAverageTemperatureUncertainty	\
0	1750-01-01	3.034	3.574	
1	1750-02-01	3.083	3.702	
2	1750-03-01	5.626	3.076	
3	1750-04-01	8.490	2.451	
4	1750-05-01	11.573	2.072	

	LandMaxTemperature	LandMaxTemperatureUncertainty	LandMinTemperature	\
0	NaN	NaN	NaN	
1	NaN	NaN	NaN	
2	NaN	NaN	NaN	

```
3                  NaN                            NaN                  NaN
4                  NaN                            NaN                  NaN

   LandMinTemperatureUncertainty  LandAndOceanAverageTemperature  \
0                            NaN                             NaN
1                            NaN                             NaN
2                            NaN                             NaN
3                            NaN                             NaN
4                            NaN                             NaN

   LandAndOceanAverageTemperatureUncertainty
0                                        NaN
1                                        NaN
2                                        NaN
3                                        NaN
4                                        NaN
```

We check for missing values:

```
dt                                          0
LandAverageTemperature                     12
LandAverageTemperatureUncertainty          12
LandMaxTemperature                       1200
LandMaxTemperatureUncertainty            1200
LandMinTemperature                       1200
LandMinTemperatureUncertainty            1200
LandAndOceanAverageTemperature           1200
LandAndOceanAverageTemperatureUncertainty    1200
dtype: int64
```

We observe that our file contains empty values, which we will get rid of during the data cleaning stage.

2.2 Data Cleaning and Preprocessing

During the data cleaning stage, our objective is to organize the datasets by removing unnecessary columns and handle incomplete data. For the "GlobalLandTemperaturesByCity.csv" file, we removed rows with empty values and excluded the "Latitude" and "Longitude" columns, as they are not relevant to our analysis. For the "GlobalTemperatures.csv" file, we retained only the columns `dt`, `LandAverageTemperature`, and `LandAndOceanAverageTemperature`, while eliminating all other columns. Rows containing empty values were also removed to ensure data consistency. These steps refine the datasets, ensuring they are clean and ready for analysis.

To prepare the data for analysis, we focus on converting columns to appropriate data types and creating new features to enhance usability. For both files, specific columns are reformatted, and additional columns are generated to support our analysis.

After completing the cleaning and preprocessing stages, the "GlobalLandTemperaturesByCity.csv" file includes:

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 8235082 entries, 0 to 8599210
Data columns (total 7 columns):
 #   Column                          Dtype
---  ------                          -----
 0   Year                            int32
 1   Season                          string
 2   AverageTemperature              float64
 3   AverageTemperatureUncertainty   float64
 4   LowerBound                      float64
 5   UpperBound                      float64
 6   Continent                       string
dtypes: float64(4), int32(1), string(2)
memory usage: 471.2 MB
```

	Year	Season	AverageTemperature	AverageTemperatureUncertainty	LowerBound	UpperBound	Continent
0	1743	Autumn	6.068	1.737	4.331	7.805	Europe
5	1744	Spring	5.788	3.624	2.164	9.412	Europe
6	1744	Spring	10.644	1.283	9.361	11.927	Europe
7	1744	Summer	14.051	1.347	12.704	15.398	Europe
8	1744	Summer	16.082	1.396	14.686	17.478	Europe

And for the "GlobalTemperatures.csv" file:

```
<class 'pandas.core.frame.DataFrame'>
Index: 1992 entries, 1200 to 3191
Data columns (total 3 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Year                            1992 non-null   int32
 1   LandAverageTemperature          1992 non-null   float64
 2   LandAndOceanAverageTemperature  1992 non-null   float64
dtypes: float64(2), int32(1)
memory usage: 54.5 KB
```

	Year	LandAverageTemperature	LandAndOceanAverageTemperature
1200	1850	0.749	12.833
1201	1850	3.071	13.588
1202	1850	4.954	14.043
1203	1850	7.217	14.667
1204	1850	10.004	15.507

3 Descriptive Statistics

First, we will present some basic descriptive statistics for our datasets. We will only include the relevant columns.

For "GlobalTemperaturesByCity.csv" we have:

	AverageTemperature
count	8.235082e+06
mean	1.672743e+01
std	1.035344e+01
min	-4.270400e+01
25%	1.029900e+01
50%	1.883100e+01
75%	2.521000e+01
max	3.965100e+01

The dataset contains 8,235,082 temperature records, with an average temperature of approximately 16.73°C. The standard deviation is around 10.35°C, indicating a significant range of temperature variation. The lowest recorded temperature is −42.74°C, while the highest is 39.65°C. In terms of distribution, 25% of the recorded temperatures are below 10.29°C, the median temperature is 18.83°C, and 75% of the temperatures fall below 25.21°C. These statistics provide a general overview of the range and distribution of daily temperatures in the dataset.

And for "GlobalTemperatures.csv" we have:

	LandAverageTemperature	LandAndOceanAverageTemperature
count	1992.000000	1992.000000
mean	8.571583	15.212566
std	4.263193	1.274093
min	0.404000	12.475000
25%	4.430000	14.047000
50%	8.850500	15.251000
75%	12.858500	16.396250
max	15.482000	17.611000

The summary statistics for the `LandAverageTemperature` and `LandAndOceanAverageTemperature` columns show that both have 1,992 entries. The average land temperature is 8.57°C, while the combined land and ocean temperature is 15.21°C. The land temperature has more variability with a standard deviation of 4.26°C, compared to 1.27°C for the land and ocean temperature. The lowest recorded land temperature is 0.40°C, while the land and ocean temperature's minimum is 12.48°C. The median land temperature is 8.85°C, and for the combined temperature, it's 15.25°C. The highest recorded land temperature is 15.48°C, while the maximum for land and ocean is 17.61°C, indicating slightly higher extremes in the combined data.

3.1 Line Plots

For a comprehensive understanding of the issue, we have displayed the following plots.

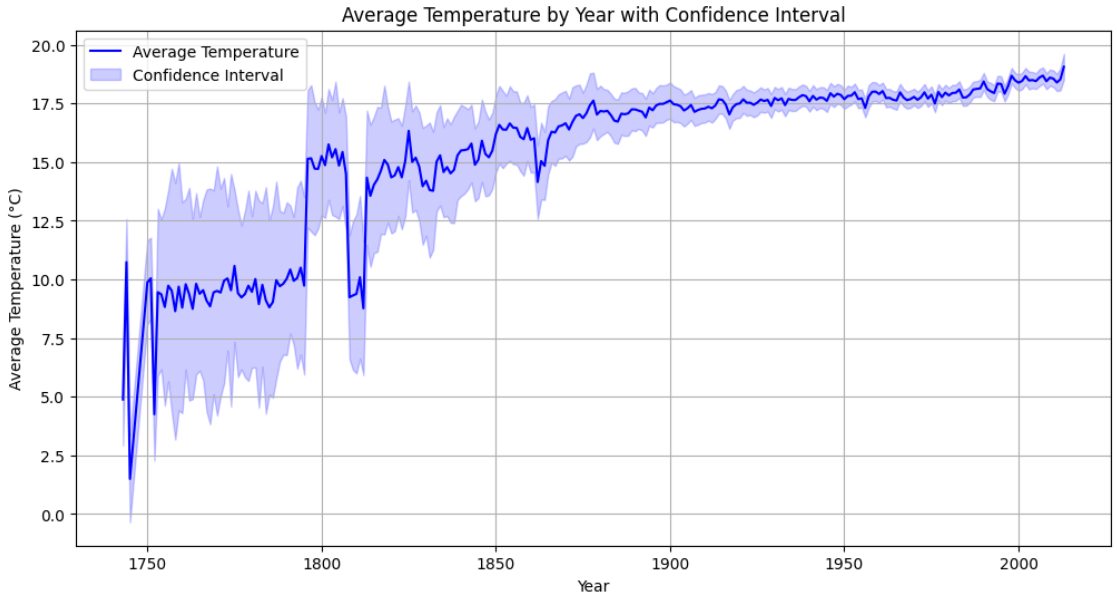

Figure 1: Yearly Average Temperature with Confidence Interval

The plot shows a clear upward trend in average temperatures over the years, particularly from the 1800s onwards, indicating global warming. In earlier periods, the temperatures are relatively stable, but fluctuations are more noticeable around the 18th and 19th centuries. The confidence interval, which shows the degree of uncertainty or variability in the temperature data, is represented by the blue shaded area. This interval narrows as we move closer to the 20th century, demonstrating that the data gets more dependable over time, but in the early years it is wider, suggesting greater uncertainty in the measurements.

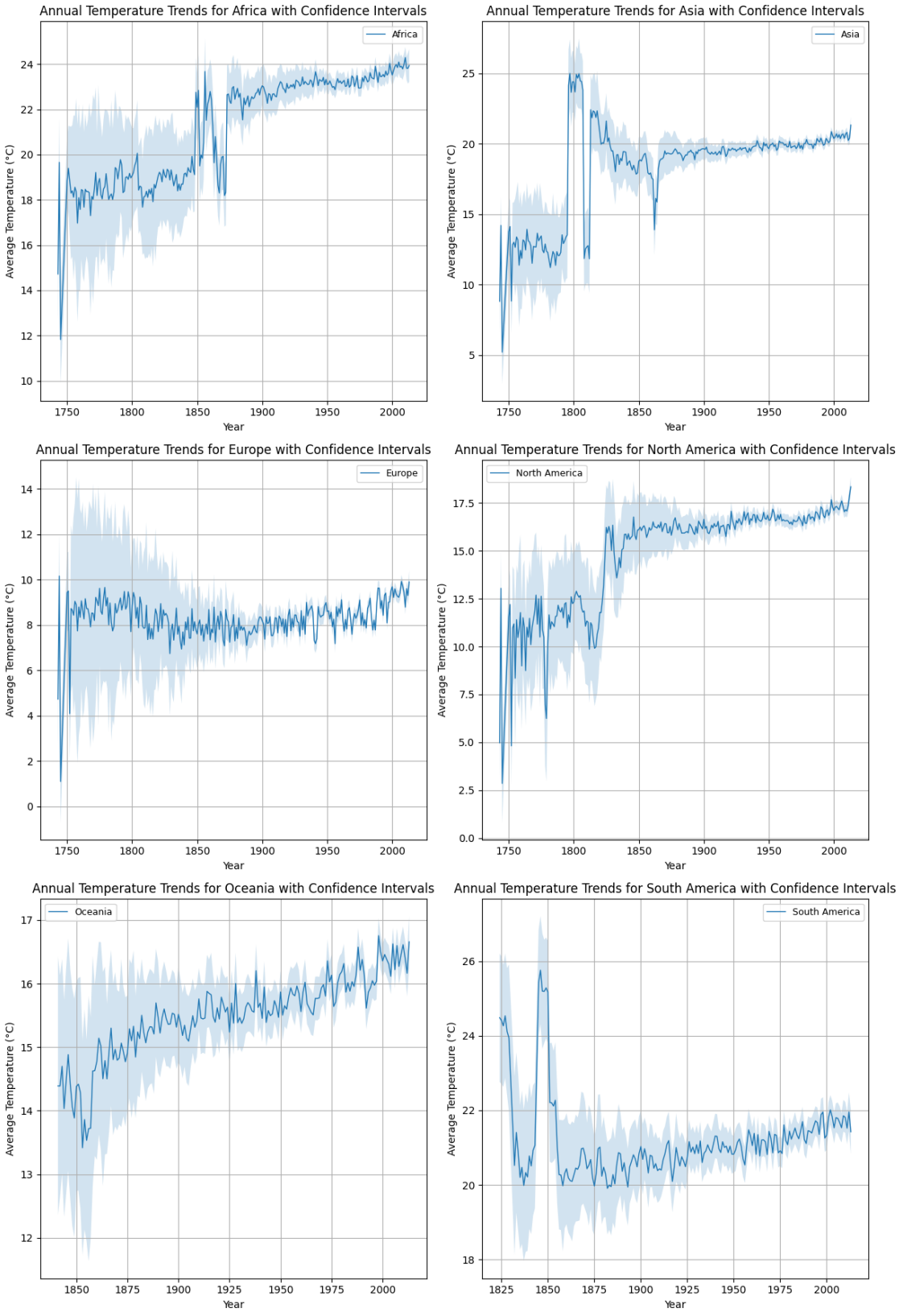

Figure 2: Average Temperatures by Continent with Confidence Intervals

The plots demonstrate a clear global warming trend across all continents, with average temperatures steadily increasing over time. Due to uncertainty, early data—especially before 1850—shows wider confidence intervals, whereas contemporary records after 1900 have narrower intervals and are more dependable.

Particularly after the mid-20th century, Africa and Asia show notable warming trends, with Asia's rise being especially steep. Europe and North America show strong and consistent warming, with Europe benefiting from historically more reliable data. Because of its oceanic climate, Oceania has consistently rising temperatures but lower averages. South America also warms steadily, though less dramatically than other continents.

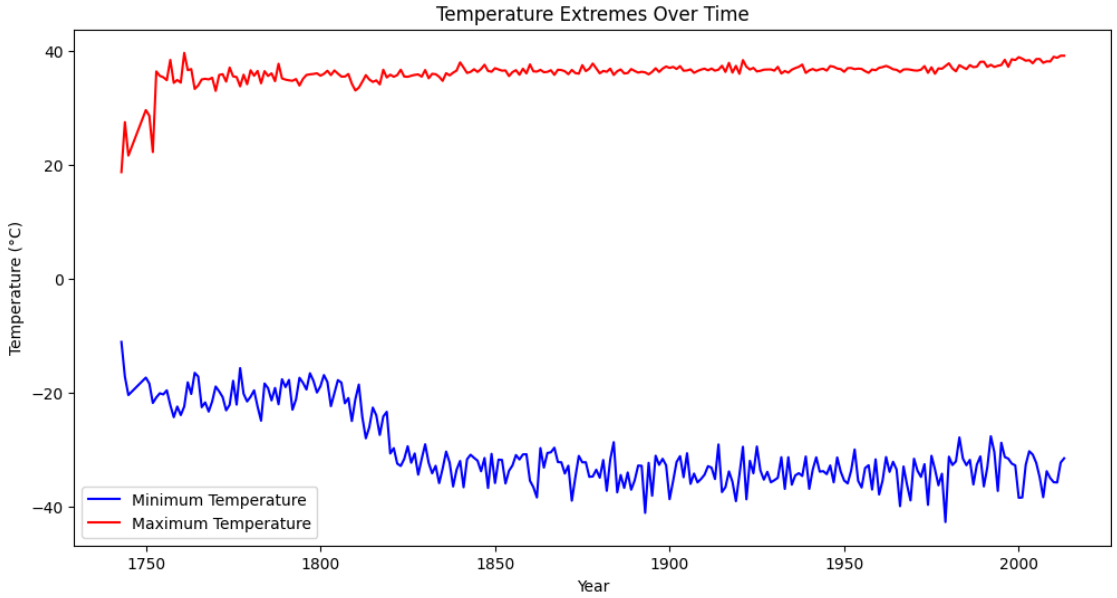

Figure 3: Minimum and Maximum Temperature Extremes Over Time

This plot shows the trends of temperature extremes—minimum and maximum temperatures—over time. The red line represents maximum temperatures, which remain relatively stable around 40°C after an initial increase during the earlier years of the dataset. This implies that there has not been any discernible variation or long-term patterns in maximum temperatures over time. From the earliest times until roughly the 19th century, the blue line, which represents minimum temperatures, clearly shows a downward trend. In more recent years, it stabilizes at lower values of about −40°C. This indicates that minimum temperatures have historically experienced greater fluctuations and a long-term cooling trend before stabilizing.

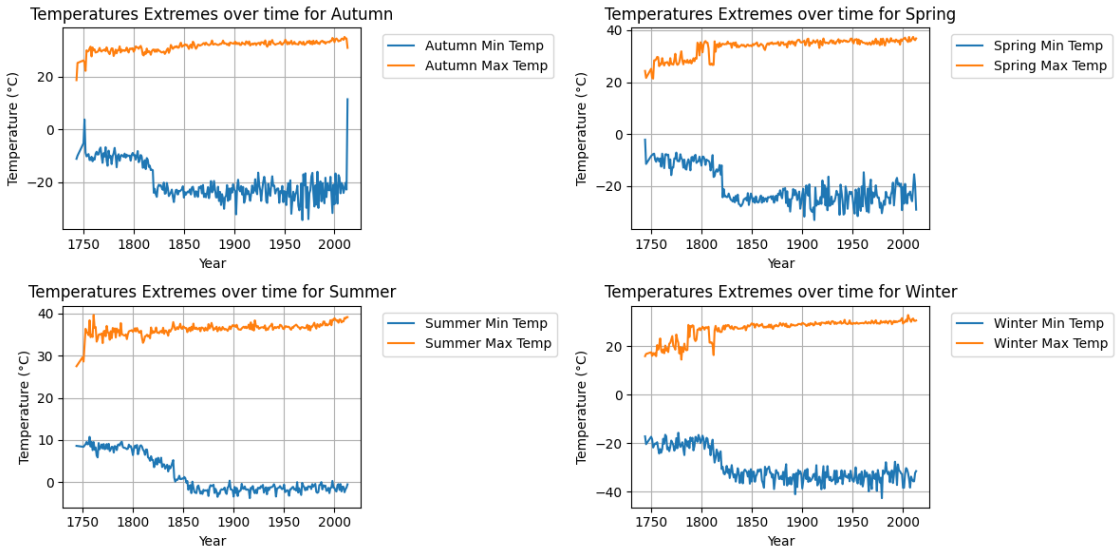

Figure 4: Seasonal Minimum and Maximum Temperature Extremes

The lowest and maximum temperature extremes for each season are displayed in these plots over time. After the 1750s, maximum temperatures stabilize with only slight variations, while minimum temperatures trend downward before leveling off at lower levels. Winter exhibits the most extreme variability in minimum temperatures, reaching as low as $-40°C$, while summer shows the narrowest range. Spring and autumn have moderate trends, stabilizing in the early 20th century. These patterns imply that minimum temperatures, especially during the winter, have been more erratic in the past but have leveled off in the last few centuries.

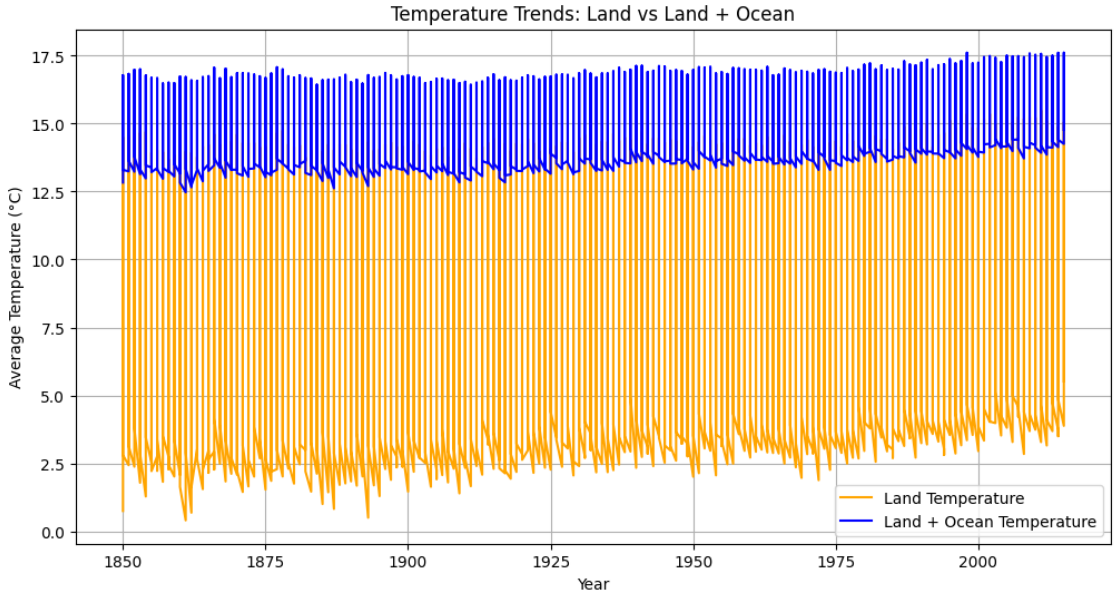

Figure 5: Land and Land + Ocean Temperatures Over Time

The graph shows a clear rise in temperatures over time, with lower values between 1850 and 1900 and

higher values by the 2000s, indicating global warming. Land temperatures are increasing faster than ocean temperatures, evident from the narrowing gap between land and land + ocean lines. While the blue line stays smoother, reflecting the moderating influence of the ocean, seasonal variations are more noticeable in land temperatures, with sharp spikes between summer and winter. These trends highlight the impact of global warming, including rising sea levels and more extreme weather events.

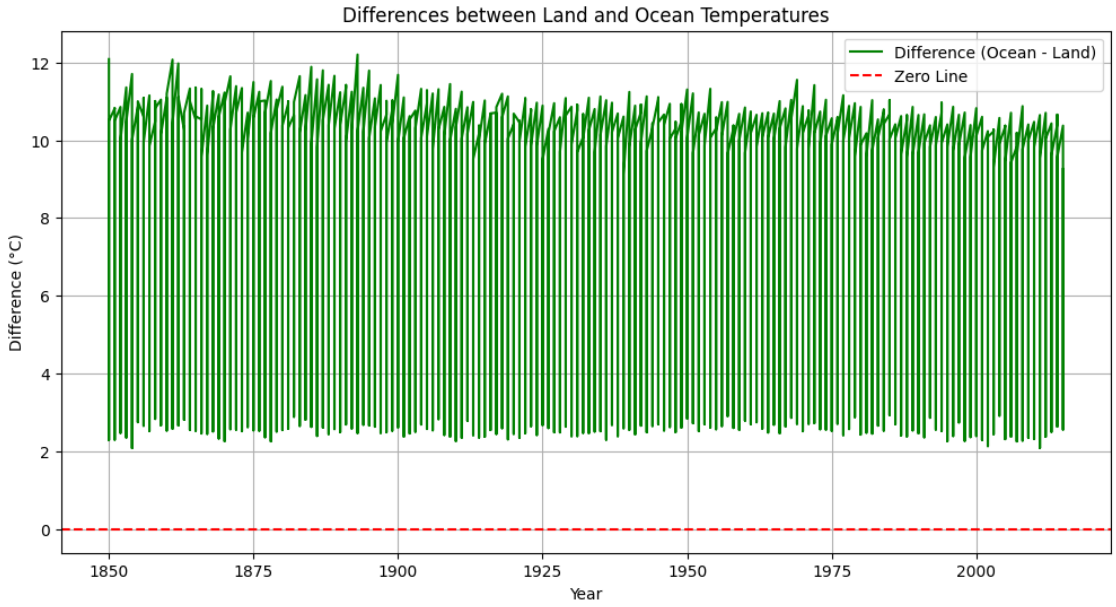

Figure 6: Difference Between Land and Land + Ocean Temperatures

The plot demonstrates that oceans consistently moderate global temperatures, with the green line showing that global (land + ocean) temperatures are always higher than land-only temperatures. The difference ranges between 2°C and 12°C, with a slight decrease in recent years, indicating that land is warming faster than oceans due to climate change. Oceans absorb and release heat more slowly, reducing temperature fluctuations, while the green line's spikes reflect more extreme seasonal variations on land compared to the oceans' stability. This consistent trend highlights the oceans' stabilizing role and underscores the faster warming of land as a clear sign of global warming.

3.2 Histograms

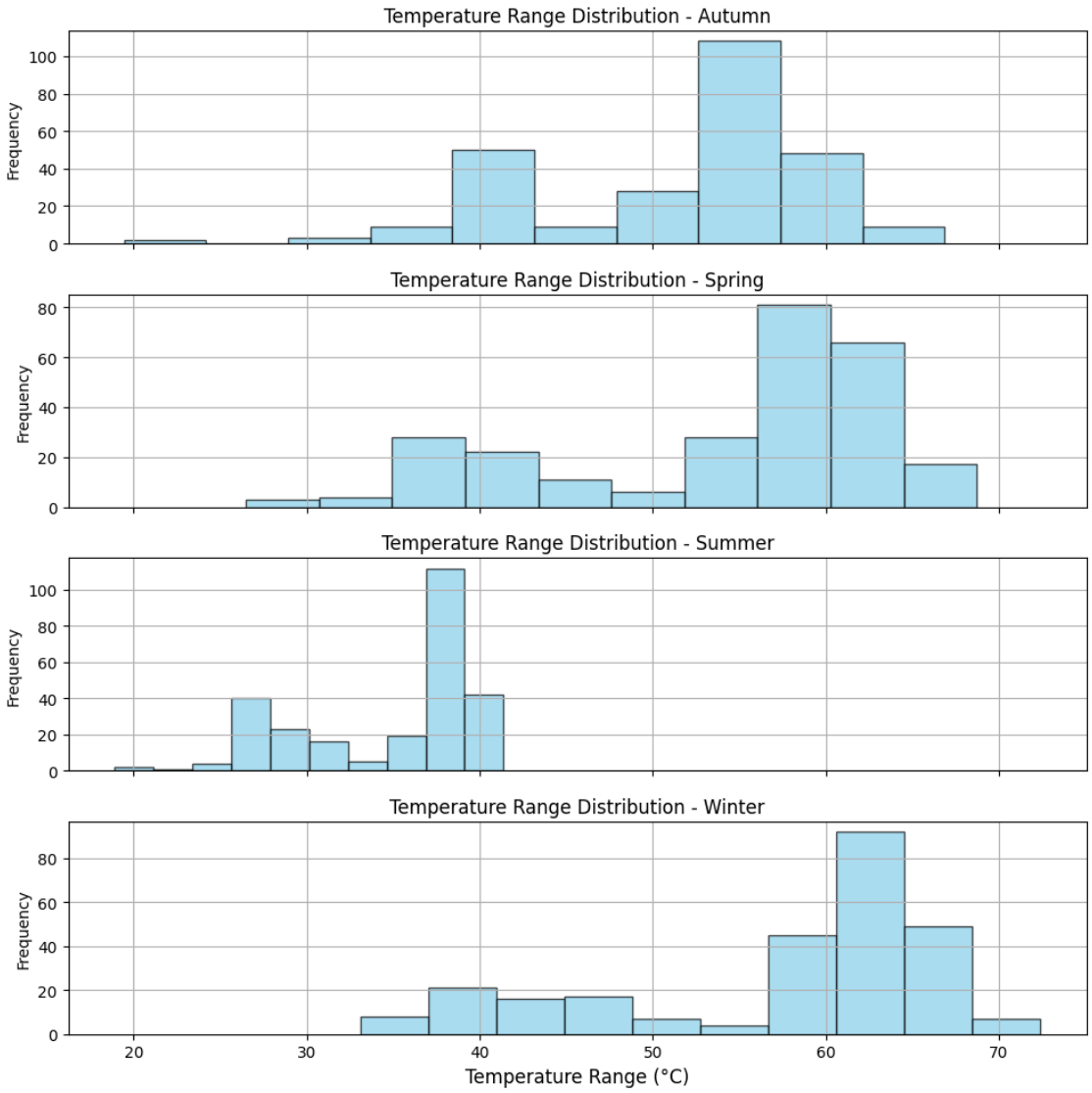

Figure 7: Histograms of Temperature Ranges by Season

These histograms depict the distribution of temperature ranges across the four seasons: autumn, spring, summer, and winter. Temperature ranges in the fall are fairly evenly distributed, with the majority of values falling within the mid-range. In wider temperature ranges, spring exhibits a similar pattern, but with a slightly higher frequency. When compared to other seasons, the temperature range distribution in the summer is smaller and less variable. With peaks centered around higher ranges, winter displays the widest distribution of temperature ranges, indicating more severe cold conditions than the other seasons. This seasonal breakdown highlights that temperature variability is most pronounced in winter and least in summer.

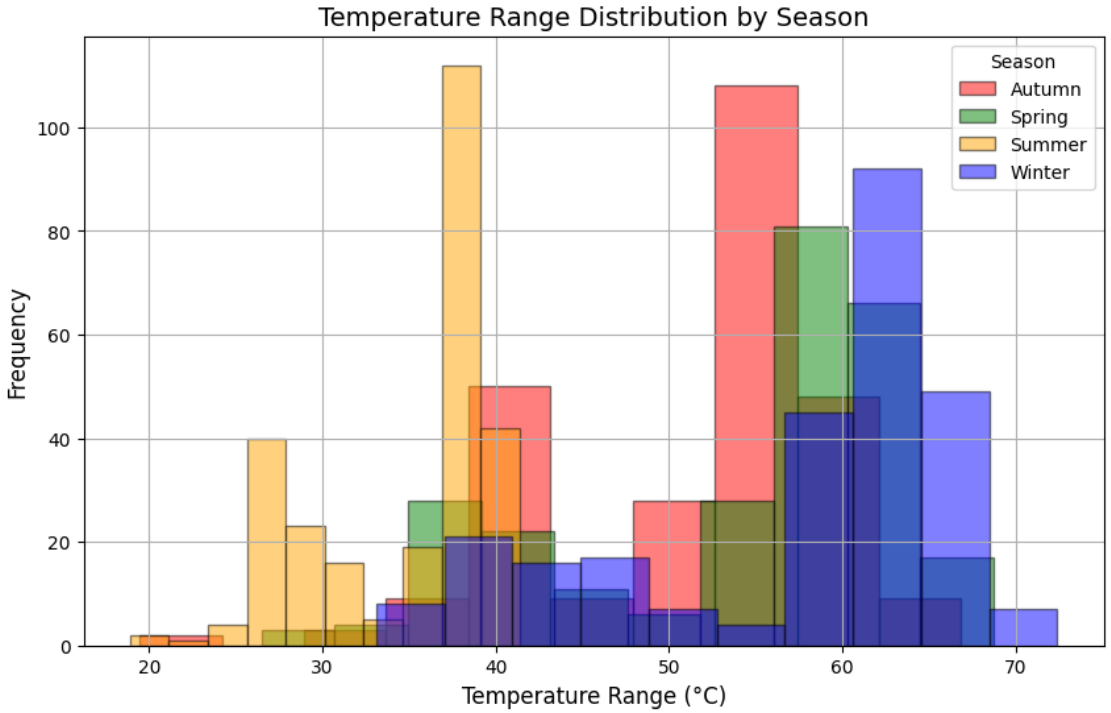

Figure 8: Combined Histogram of Temperature Ranges Across Seasons

This combined histogram compares the temperature distributions across seasons. Autumn and spring have overlapping distributions, with most temperatures ranging from 40°C to 55°C, showing moderate variability. Summer has a narrower range around 40°C, indicating less variation. Winter shows a wider spread, with temperatures from 50°C to 70°C, reflecting extreme variability. Compared to individual seasonal histograms, this view confirms that winter has the widest range, while summer has the least, with autumn and spring showing intermediate variability.

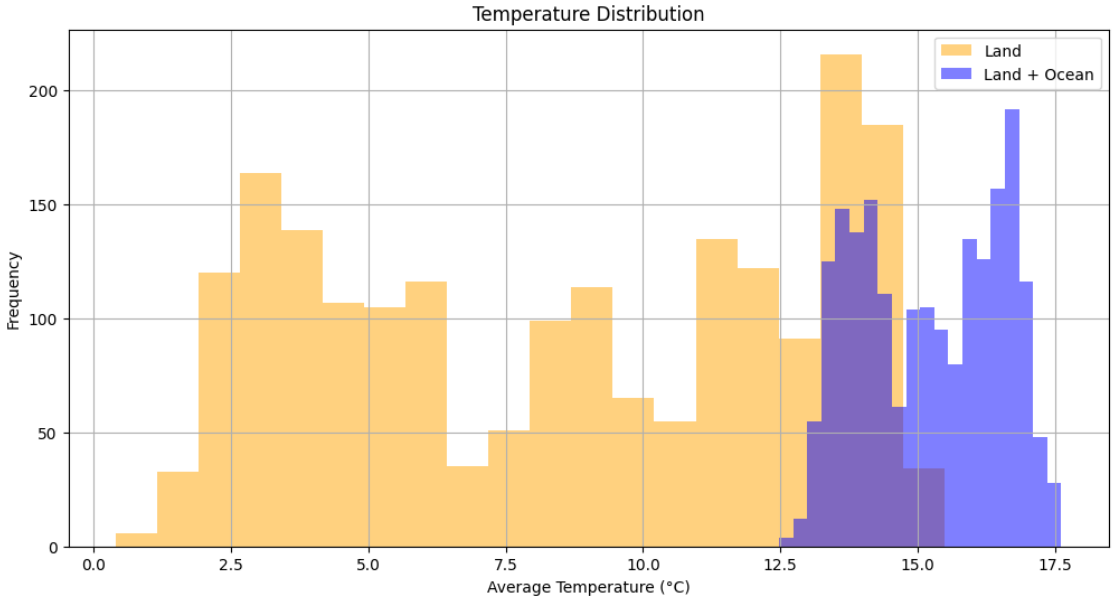

Figure 9: Histogram Comparing Land and Land + Ocean Temperatures

This histogram compares the distribution of land temperatures with combined land and ocean temperatures. With a range of almost 0°C to 15°C, land temperatures exhibit more variability. In contrast, land + ocean temperatures are more concentrated between 12.5°C and 17.5°C, reflecting the ocean's moderating effect. The overlap highlights common temperature ranges between 12.5°C and 15°C. In contrast to the ocean-influenced dataset, the land-only data exhibits more extreme low temperatures, suggesting greater variability.

3.3 Box Plot

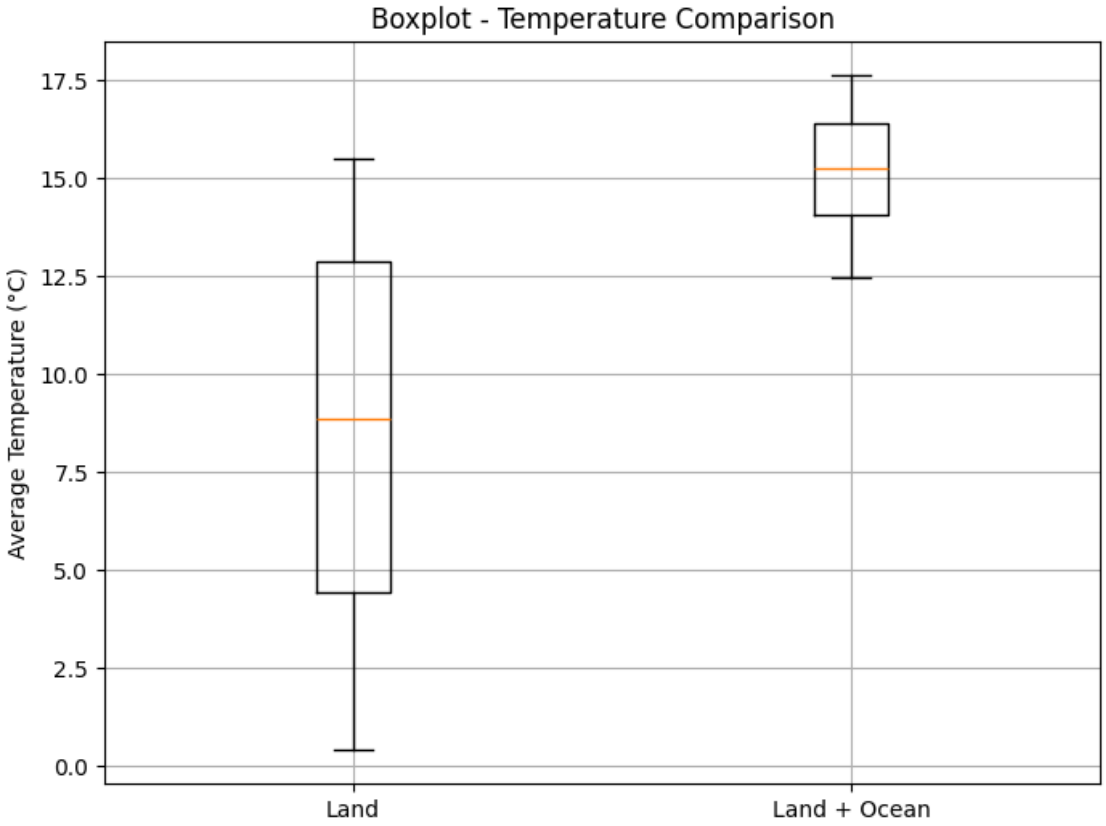

Figure 10: Box Plot of Land vs. Land + Ocean Temperatures

This boxplot compares the distributions of average land temperatures and land + ocean temperatures. The range of land temperatures is greater, with a lower median of about 9°C and notable variability ranging from close to 0°C to roughly 15°C. On the other hand, because of the moderating effect of the oceans, land + ocean temperatures have a narrower range and a higher median around 15°C, indicating greater stability. The plot highlights the broader variability of land temperatures compared to the more consistent global averages.

3.4 Frequency Polygons

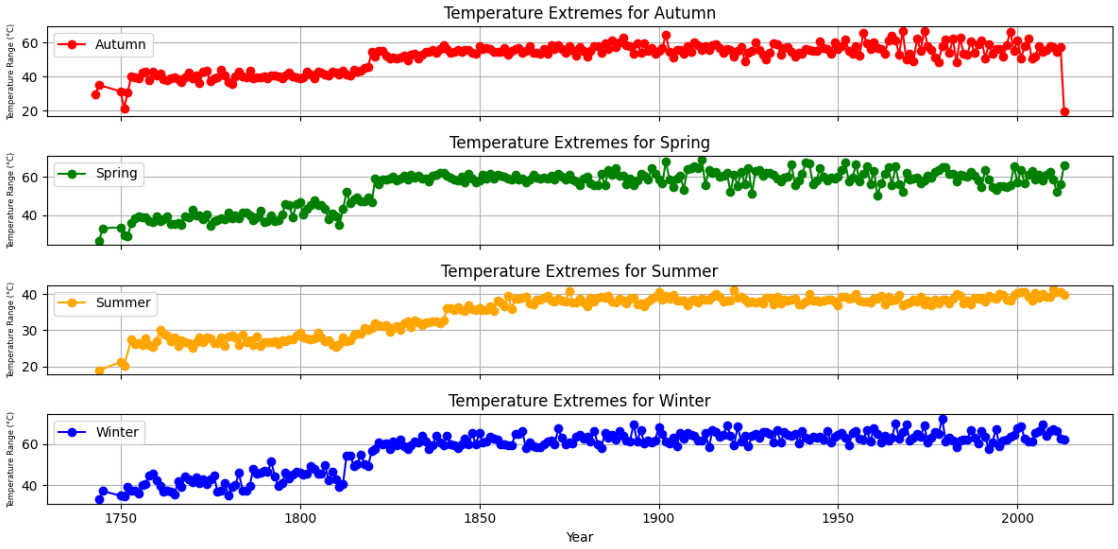

Figure 11: Frequency Polygons of Temperature Extremes by Season

The frequency polygons show that winter has the largest and most variable temperature range, increasing from below 30°C to around 60°C over time. Summer maintains the most consistent and narrow range, staying near 40°C. Autumn and spring show moderate increases, stabilizing around 50°C. These patterns highlight winter's greater extremes compared to the other seasons. We can also observe a steady increase of temperature extremes over time, proving their frequency.

3.5 Regression Lines

```
The Standard Error: 0.0014
Slope (temperature increase per year): 0.0375°C/year
P-value: 2.1439e-80
R-squared value: 0.7441
```

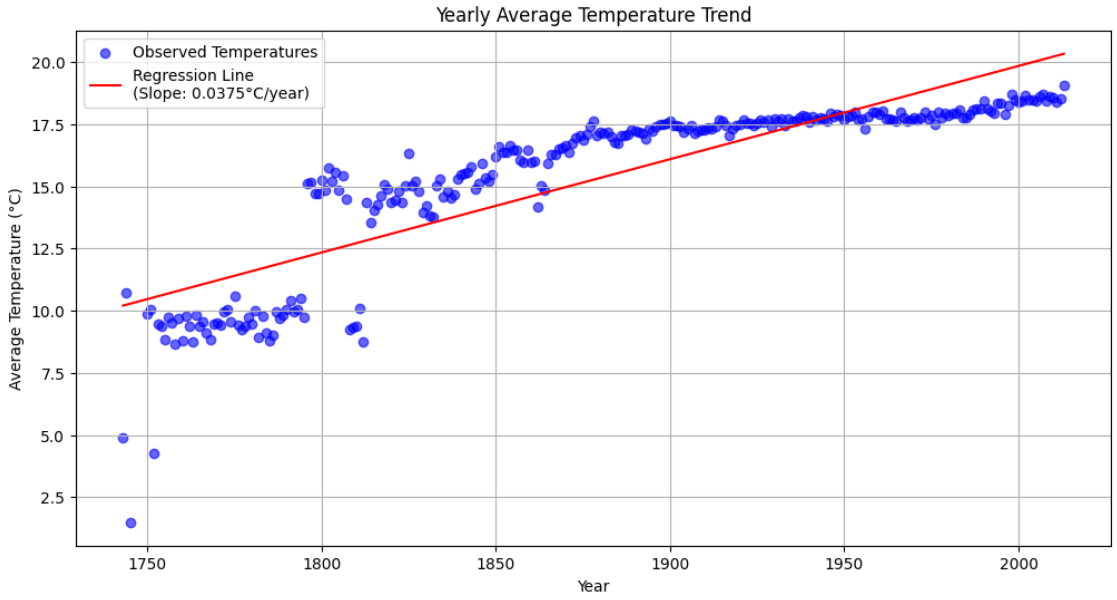

Figure 12: Linear Regression of Average Temperature Over Time

With a regression slope of 0.0375°C annually, the plot clearly displays an upward trend in annual average temperatures, suggesting a consistent rise over time. The p-value of 2.14×10^{-80} confirms this trend is statistically significant, and the R-squared value of 0.7441 shows that 74.41% of the temperature variation is explained by the trend. This offers compelling proof of a steady and substantial increase in global temperatures brought on by climate change.

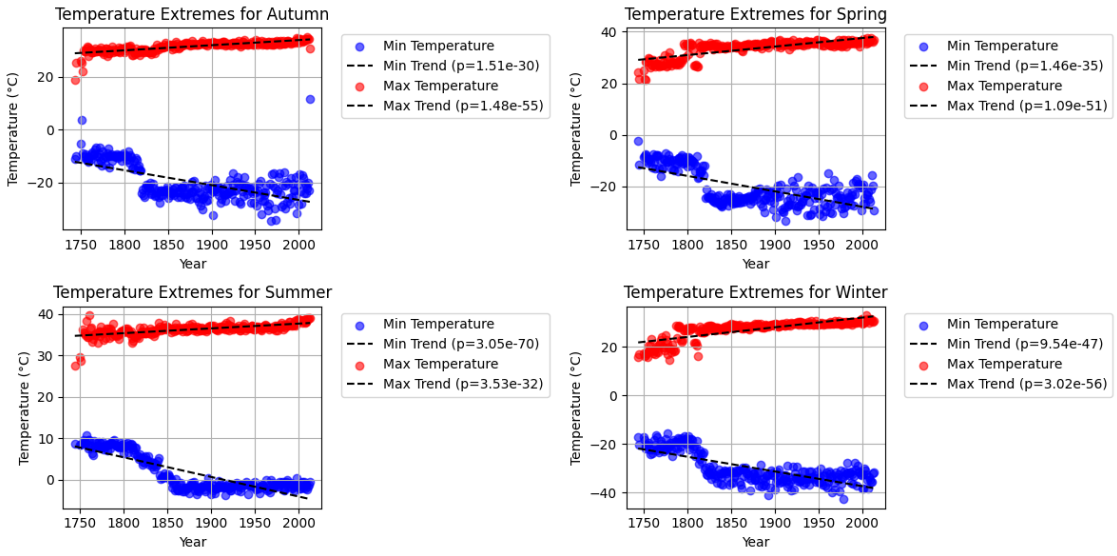

Figure 13: Seasonal Regression Lines for Temperature Extremes

	Min Temp Slope	Min Temp P-value	Min Temp R-value	Max Temp Slope	\
Autumn	-0.056080	1.511042e-30	-0.627674	0.019404	
Spring	-0.059876	1.456370e-35	-0.666707	0.032971	
Summer	-0.047357	3.046442e-70	-0.835992	0.011547	

```
Winter       -0.060642      9.537971e-47         -0.736623         0.039999
```

```
    Max Temp P-value  Max Temp R-value
Autumn      1.477335e-55          0.779490
Spring      1.088215e-51          0.761769
Summer      3.534887e-32          0.642853
Winter      3.023242e-56          0.782484
```

The plots show seasonal trends in minimum and maximum temperatures. Minimum temperatures decline sharply, particularly in winter (-0.060642°C/year), before stabilizing, while maximum temperatures remain relatively stable with slight increases. In autumn, spring, and summer, minimum temperatures also decrease, with trends of -0.056080°C/year, -0.059876°C/year, and -0.047357°C/year, respectively. Winter shows the largest increase in maximum temperatures (0.039999°C/year). Every trend is statistically significant (p-values < 0.05), showing that the effects of climate change are reflected in an increase in seasonal extremes, particularly during the colder seasons.

```
Linear Regression for Land Temperatures:
Slope (yearly increase): 0.0085°C/year
Intercept: -7.92

Linear Regression for Land + Ocean Temperatures :
Slope (yearly increase): 0.0054°C/year
Intercept: 4.87
```

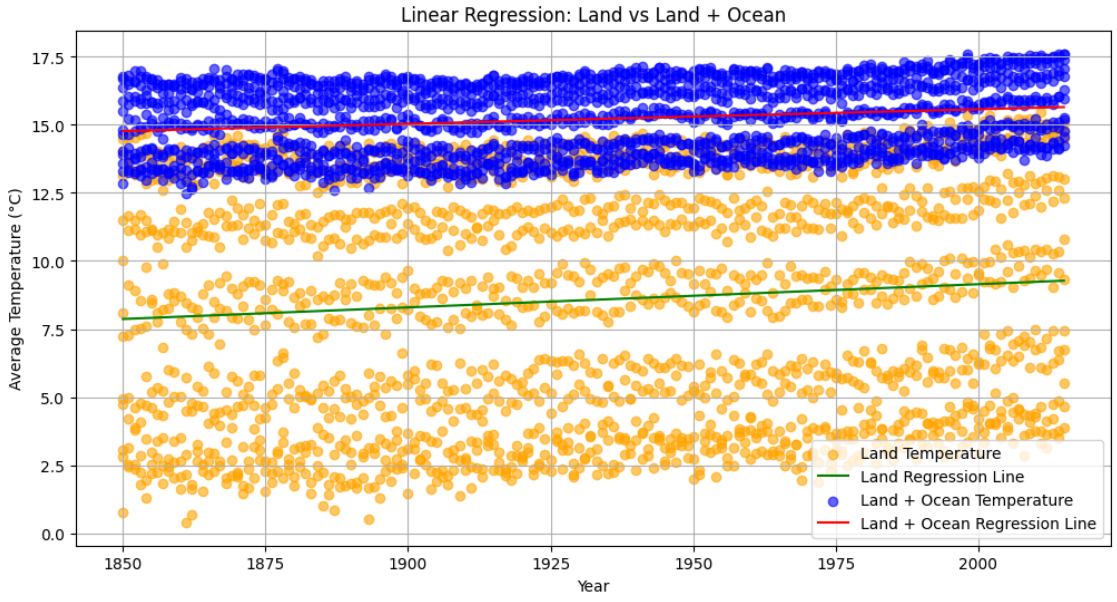

Figure 14: Regression Lines Comparing Land and Land + Ocean Temperatures

The graph illustrates clear evidence of global warming, with both land and land + ocean temperatures showing upward trends over time. The green regression line for land temperatures has a steeper slope compared to the red line for land + ocean temperatures, indicating that land is warming faster than the global average. The scatter points show greater variability for land temperatures,

reflecting seasonal changes and more extreme fluctuations, while the global temperatures are more stable due to the moderating effect of oceans.

4 Inferences

4.1 Confidence Intervals

4.1.1 Theoretical Foundations

Definition A confidence interval (CI) is a range of values that is likely to contain the true population parameter (e.g., mean) with a certain level of confidence, typically 95%.

Equation

$$\text{CI} = \bar{X} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

Where

- $\bar{X}$: Sample mean.
- Z: Z-score corresponding to the desired confidence level (e.g., 1.96 for 95%).
- σ: Population standard deviation (or sample standard deviation for smaller datasets).
- n: Sample size.

Key Components

- **Lower Bound**: The smallest value in the interval, calculated as $\bar{X} - Z \cdot \frac{\sigma}{\sqrt{n}}$.
- **Upper Bound**: The largest value in the interval, calculated as $\bar{X} + Z \cdot \frac{\sigma}{\sqrt{n}}$.

Purpose Confidence intervals provide an estimate of the uncertainty in a measurement. For example, if the CI for the mean temperature is 15°C to 17°C at 95% confidence, it means we are 95% confident the true mean lies within this range. They are widely used in statistics to quantify the precision of estimates and to account for sampling variability.

4.1.2 Interpretation

The plot shows the yearly average temperature trend alongside its confidence interval, represented by the shaded blue area. The solid blue line represents the average temperature, which displays a clear upward trend over time, indicating a long-term increase in global temperatures. The shaded confidence interval captures the uncertainty around the average temperature measurements, with narrower intervals in more recent years reflecting improved measurement precision or reduced variability. The earlier years, especially before the 1800s, have wider confidence intervals due to greater uncertainty in historical temperature data.

The plots display annual temperature trends for different continents, each with a shaded confidence interval indicating the uncertainty in temperature measurements. Across all continents, a general upward trend in average temperatures is evident, consistent with global warming. Africa and North America show significant increases in average temperatures, with relatively narrow confidence intervals in recent years, reflecting higher measurement reliability. Asia and South America exhibit

noticeable warming trends, though confidence intervals are wider for earlier years due to greater uncertainty in historical data. Europe and Oceania also show steady warming, with Europe having narrower intervals, suggesting more precise measurements. The shaded regions, wider in earlier years for all continents, highlight the variability and uncertainty in older records.

4.2 Correlation

4.2.1 Theoretical Foundations

Equation

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where

- r: Pearson correlation coefficient (ranges from -1 to 1).
- X_i: Values of the first variable X.
- Y_i: Values of the second variable Y.
- $\bar{X}$: Mean of the first variable X.
- $\bar{Y}$: Mean of the second variable Y.

Value of r

- $r = 1$: Perfect positive correlation (as X increases, Y increases proportionally).
- $r = -1$: Perfect negative correlation (as X increases, Y decreases proportionally).
- $r = 0$: No linear correlation between X and Y.

Strength and Direction

- The closer r is to -1 or 1, the stronger the relationship.
- Positive r values indicate direct relationships, while negative r values indicate inverse relationships.

Purpose

- Evaluate the strength of the linear relationship between two variables.
- Understand how closely two variables are related. For example, a high r between land and ocean temperatures indicates they move together predictably.

4.2.2 Interpretation

```
Pearson Correlation Coefficient: 0.9881
```

A correlation of 0.9881 is very close to 1, meaning that there is a nearly perfect linear relationship between land temperatures and global temperatures. When land temperatures increase, global temperatures (land + ocean) also increase in a predictable and proportional manner.

4.3 Linear Regression

4.3.1 Theoretical Foundations

Equation

$$Y = mX + b$$

Where

- Y: Dependent variable (outcome being predicted or explained).
- X: Independent variable (predictor or explanatory variable).
- m: Slope, indicating how much Y changes for each unit increase in X.
- b: Intercept, the value of Y when $X = 0$.

Core Ideas

1. **Linearity**: Assumes a straight-line relationship between X and Y.
2. **Goodness of Fit (R^2)**: Measures how much of the variability in Y is explained by X.

Purpose Linear regression is used to:

- **Predict** outcomes (Y) from known values of X.
- **Quantify Trends** (e.g., rate of change through the slope m).
- **Explain Variability** in the data using R^2.

Linear regression is simple, interpretable, and widely used for understanding and predicting trends.

4.3.2 Interpretation

Our first regression chart illustrates the yearly average temperature trend over time, with observed temperatures shown as **blue scatter points** and the fitted regression line in **red**. The upward slope of the regression line, calculated as 0.0375°C per year, indicates a consistent rise in average temperatures, demonstrating the impact of long-term global warming.

- The **standard error** of 0.0014 quantifies the accuracy of the slope estimate, suggesting a high level of precision.
- The slope reflects an annual rate of temperature increase, translating to approximately 3.75°C over a century.
- The **R-squared value** of 0.7441 indicates that about 74.4% of the variability in the temperature data is explained by the linear model, showing a strong fit.

The second regression plots illustrate seasonal trends in temperature extremes for **autumn**, **spring**, **summer**, and **winter**.

- *Minimum temperatures* show a clear **decreasing trend** over time, with negative slopes (e.g., −0.056080°C/year for autumn and −0.060642°C/year for winter), indicating colder extremes historically.

- *Maximum temperatures* show a smaller **increasing trend**, with positive slopes (e.g., 0.019404°C/year for autumn and 0.039999°C/year for winter).
- All trends are statistically significant, with extremely small p-values, highlighting shifts in seasonal temperature extremes as minimums become colder and maximums gradually warm, likely due to long-term climate patterns.

The last regression analysis compares **land temperatures** (orange points) and **global temperatures** (blue points, land + ocean) over time.

- The regression lines, shown in **green** for land and **red** for global temperatures, quantify the trends. Land temperatures have a slope of 0.0085°C/year, indicating a faster warming rate compared to the global slope of 0.0054°C/year.
- This shows that land is warming more rapidly than the global average.
- The consistent vertical gap between the lines highlights the ocean's moderating influence on global temperatures.

These trends provide clear evidence of climate change, with land exhibiting greater variability compared to the more stable global temperatures.

4.4 Hypothesis Testing (P-value)

4.4.1 Theoretical Foundations

Definition The p-value is a probability that measures the strength of evidence against the null hypothesis (H_0). It represents the likelihood of observing the given data, or something more extreme, if the null hypothesis is true.

$$\text{p-value} = P(\text{Test Statistic} \mid H_0)$$

Where

- H_0: Null hypothesis, typically assuming no effect or no relationship exists.
- Test Statistic: A value derived from sample data, such as a t-score or z-score.

Significance Levels Common thresholds for interpreting p-values include:

- $p < 0.05$: Reject H_0, the result is statistically significant.
- $p \geq 0.05$: Fail to reject H_0, insufficient evidence to claim significance.

Purpose The p-value helps quantify whether observed data could occur by random chance under H_0. A low p-value indicates that the observed result is unlikely due to chance, suggesting a significant effect or relationship.

4.4.2 Interpretation

The **p-value**, at 2.14×10^{-80}, is extraordinarily small, providing overwhelming evidence that the trend is statistically significant and not due to random chance.

5 Conclusions and Future Work

In this study, we have analyzed comprehensive temperature datasets to validate the reality of global warming. Our findings indicate a clear and statistically significant upward trend in global temperatures over time. The analyses show that:

- Average temperatures are increasing globally, with land temperatures rising faster than ocean temperatures.
- There is a nearly perfect positive correlation between land and ocean temperatures.
- Seasonal analyses reveal that temperature extremes are becoming more pronounced, especially minimum temperatures in winter.
- The p-values from our regression analyses are extremely low, confirming the statistical significance of the observed trends.

These results reinforce the scientific consensus that global warming is a real and ongoing phenomenon, primarily driven by human activities.

5.1 Future Work

Future research could focus on:

- Investigating the impact of global warming on specific regions or ecosystems.
- Analyzing other climate variables such as precipitation, sea level rise, and extreme weather events.
- Assessing the effectiveness of mitigation strategies and policies aimed at reducing greenhouse gas emissions.
- Incorporating more recent data to monitor ongoing trends and update models.

References

[1] Berkeley Earth: "Climate Change: Earth Surface Temperature Data"; Link: https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv

[2] James R. Fleming: "Joseph Fourier, the 'greenhouse effect', and the quest for a universal theory of terrestrial temperatures"; Link: https://www.sciencedirect.com/science/article/abs/pii/S0160932799012107

[3] Shapiro, Maura: "Eunice Newton Foote's nearly forgotten discovery."; *Physics Today*; AIP Publishing.

[4] Jaffe, Michael B: "Infrared Measurement of Carbon Dioxide in the Human Breath: 'Breathe-Through' Devices from Tyndall to the Present Day."; *Anesthesia & Analgesia*.

[5] Arrhenius, Gustaf, Karin Caldwell, and Svante Wold. "A Tribute to the Memory of Svante Arrhenius (1859–1927): A Scientist Ahead of His Time.";

[6] Gao, Yun, Xiang Gao, and Xiaohua Zhang. "The 2 °C Global Temperature Target and the Evolution of the Long-Term Goal of Addressing Climate Change—From the United Nations Framework Convention on Climate Change to the Paris Agreement."; *Advances in Climate Change Research*.

Project link: https://github.com/mogugugugugu/Statistical-Study.git

.