



INDIANA UNIVERSITY
BLOOMINGTON

SP25-BL-DSCI-D590-10650

**Analyzing Long-Term Climate Trends: Investigating the Evolution
of Temperature, Humidity, and Wind Speed Across Regions.**

Final Project Part 2

Pasan Kamburugamuwa

Indiana University Bloomington

April 06, 2025

1. Description

This dataset provides a comprehensive collection of global temperature records, compiled and cleaned by Berkeley Earth, a research organization affiliated with the Lawrence Berkeley National Laboratory. Unlike simpler datasets, this one accounts for complex historical inconsistencies in temperature measurement, making it one of the most reliable sources for climate trend analysis.

Key Features of the Dataset

1. Data Sources & Cleaning Challenges

- **Historical Measurement Issues:**
 - Early data (pre-1940s) came from mercury thermometers, affected by inconsistent recording times.
 - Airport construction in the 1940s displaced many weather stations.
 - Electronic thermometers introduced in the 1980s had a documented **cooling bias**.
- **Repackaged from Berkeley Earth:**
 - Combines **1.6 billion temperature reports** from 16 archives (NOAA's MLOST, NASA's GISTEMP, HadCrut, etc.).
 - Uses advanced methods to retain data from shorter time series, minimizing discarded observations.
 - Fully transparent: Includes **source data** and **transformation code**.

2. Time Series Decomposition

What kind of data is available in the dataset.

- **Date:** Time series starting from 1750 (land) and 1850 (land & ocean).
- **LandAverageTemperature:** Global average land temperature in Celsius.
- **LandAverageTemperatureUncertainty:** 95% confidence interval for land temperature.
- **LandMaxTemperature & LandMinTemperature:** Global maximum and minimum land temperatures.
- **LandAndOceanAverageTemperature:** Combined global land and ocean temperature.
- **LandAndOceanAverageTemperatureUncertainty:** 95% confidence interval for combined temperature.

	dt	AverageTemperature	AverageTemperatureUncertainty	City	Country	Latitude	Longitude
0	1743-11-01	6.068	1.737	Århus	Denmark	57.05N	10.33E
1	1743-12-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
2	1744-01-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
3	1744-02-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
4	1744-03-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
5	1744-04-01	5.788	3.624	Århus	Denmark	57.05N	10.33E
6	1744-05-01	10.644	1.283	Århus	Denmark	57.05N	10.33E
7	1744-06-01	14.051	1.347	Århus	Denmark	57.05N	10.33E
8	1744-07-01	16.082	1.396	Århus	Denmark	57.05N	10.33E
9	1744-08-01	NaN	NaN	Århus	Denmark	57.05N	10.33E

Diagram: First 10 rows of the dataset

Time series decomposition is a technique used to break down a time series into several components. This helps to understand the underlying structure of the data, useful for forecasting, anomaly detection and modeling.

1. Trends
2. Seasonality
3. Residuals (Noise)

2.1 Trends

This shows the **long-term direction** in the data—whether it's increasing, decreasing, or staying flat.

- Capture the structural change over time.
- Helps visualize gradual patterns like global warming.

From the plot, **There is a clear upward trend- temperature have been rising, especially after the 1900s.**

2.2 Seasonality

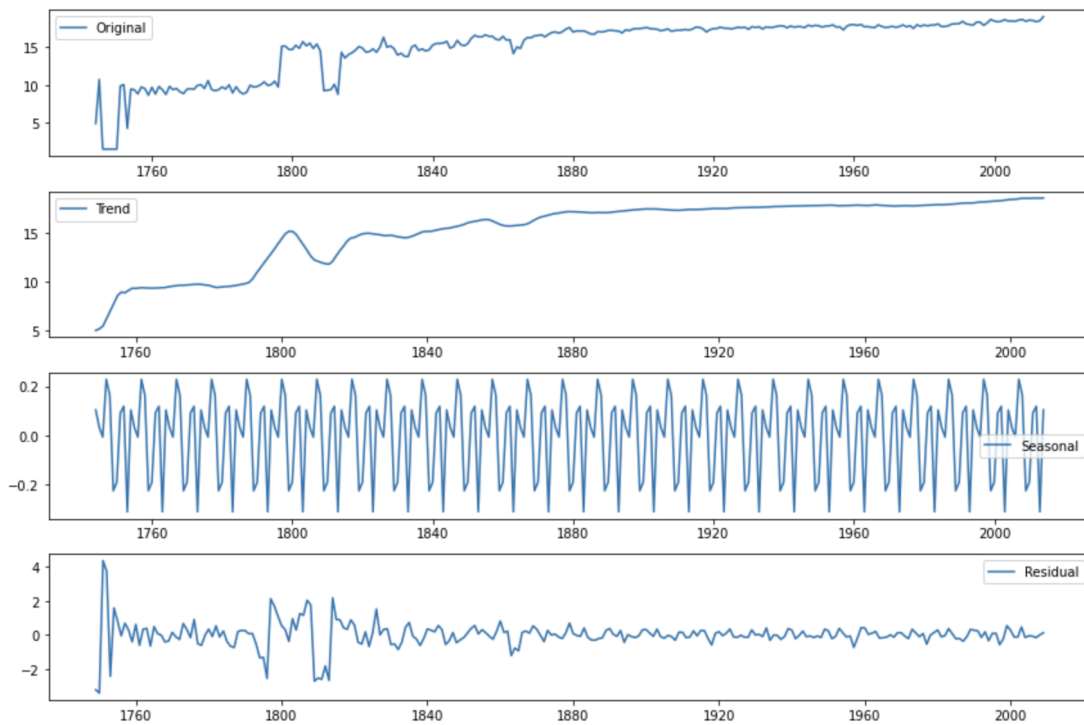
This captures repeating patterns at regular intervals. (eg: monthly, quarterly, yearly). The seasonal variation repeats over the same period.

From the plot, **A wave-like structure repeating consistently, showing annual seasonal fluctuation.**

2.3 Residuals

This is the **leftover part** of the series after removing the trend and seasonality—random variation, errors, or anomalies. This helps out outliers or structural breaks.

From the plot, **Earlier years have more volatility, likely due to less accurate or inconsistent measurements.**

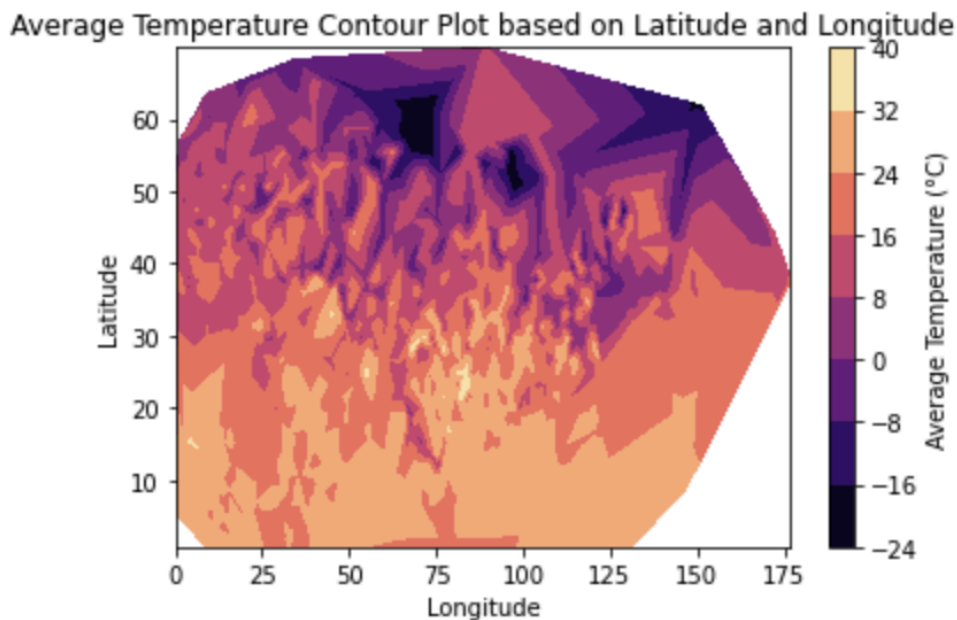


Plot showing the original, trend, seasonality and residual

3. Time series visualization

As we already plotted the trends, seasonality and residuals in the time series decomposition part, let's focus on other factors of the data.

3.1 Average temperature contour plot based on latitude and longitude.



Plot: Average temperature contour plot based on latitude and longitude

Colder areas are generally located at higher latitudes (top of the plot), which makes sense because those are closer to the poles.

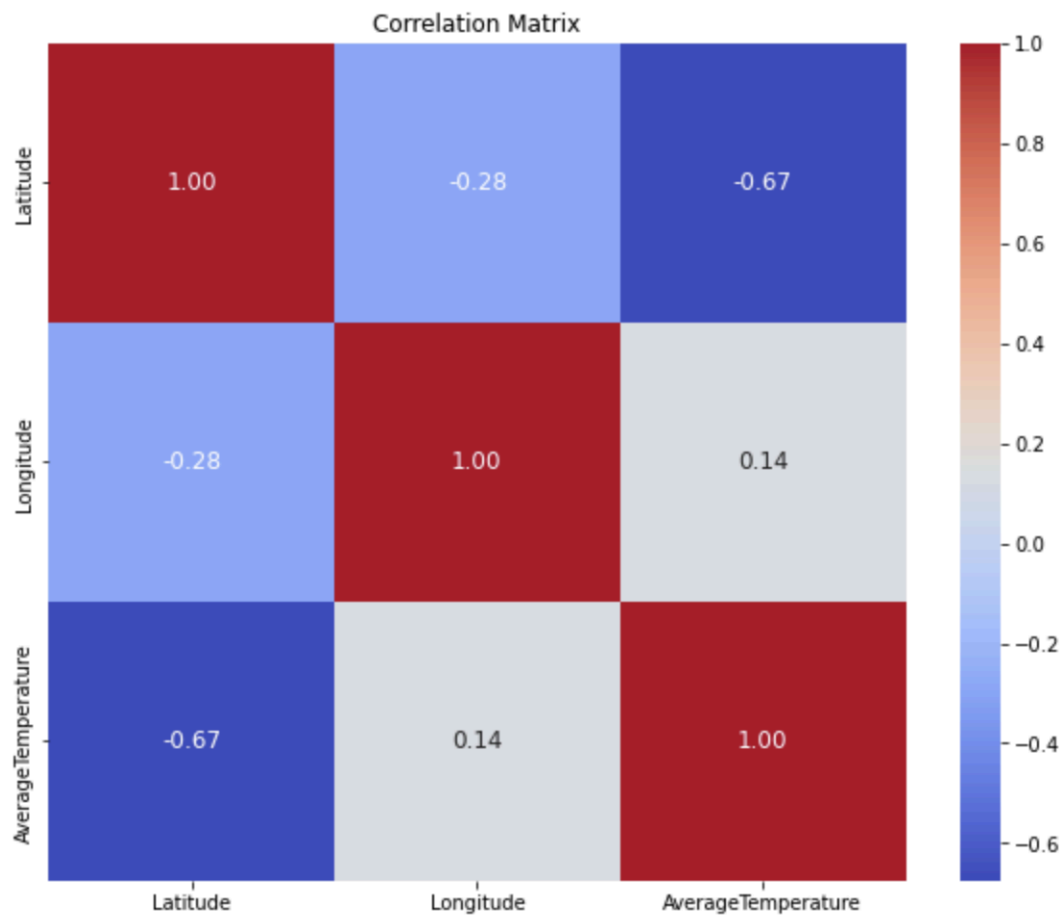
Equatorial regions (lower latitudes) are shown in lighter colors, indicating warmer average temperatures

This diagram is super useful in climatology and geospatial analysis because it shows how temperature varies spatially.

3.2 Average temperature contour plot based on latitude and longitude.

This image is a **correlation matrix heatmap**, and it visually represents the strength and direction of relationships between three variables:

- **Latitude**
- **Longitude**
- **Average Temperature**



Plot: Correlation Matrix

Latitude vs Average Temperature:

- **Correlation: -0.67**
- **Strong negative correlation** → As latitude increases (i.e., you move farther from the equator toward the poles), average temperature decreases.

Longitude vs Average Temperature:

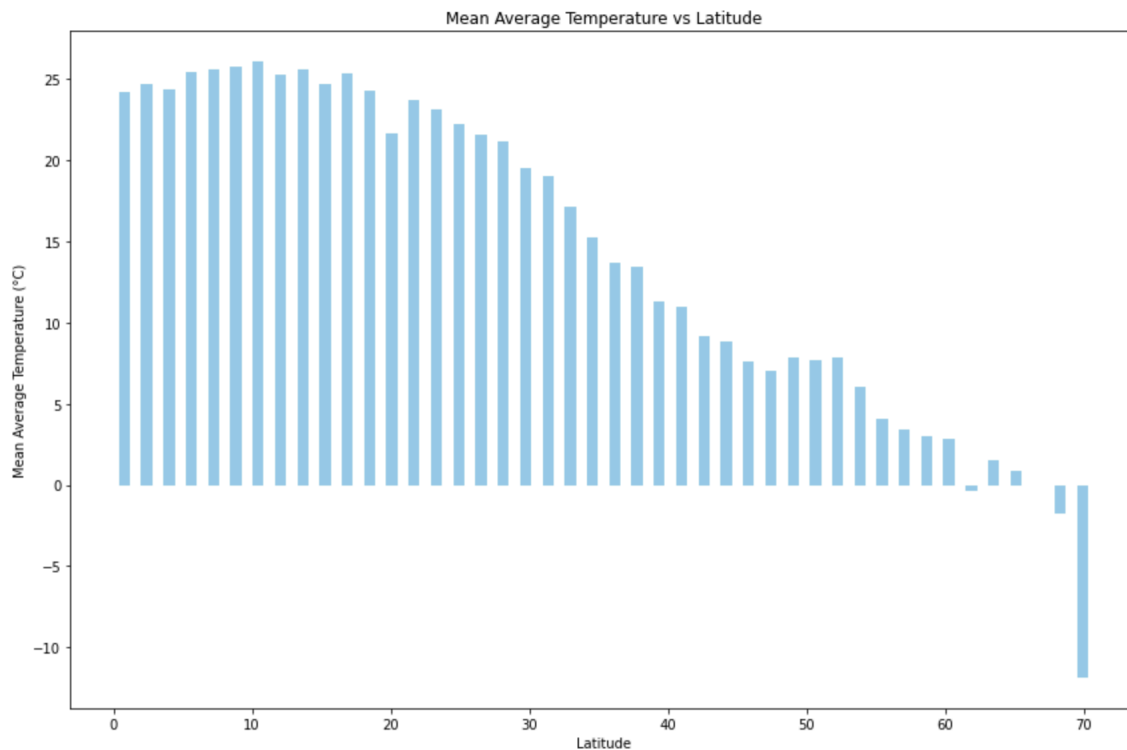
- **Correlation: 0.14**
- Weak positive correlation → Longitude has **almost no meaningful impact** on temperature in this dataset.

Latitude vs Longitude:

- **Correlation: -0.28**
- Slight negative correlation → There may be some spatial clustering, but it's not strong.

Latitude is the strongest geographical factor influencing **average temperature**. **Longitude** shows **minimal correlation** with temperature, meaning that **east-west position doesn't strongly affect** global temperature patterns

3.2 Mean average temperature vs latitude



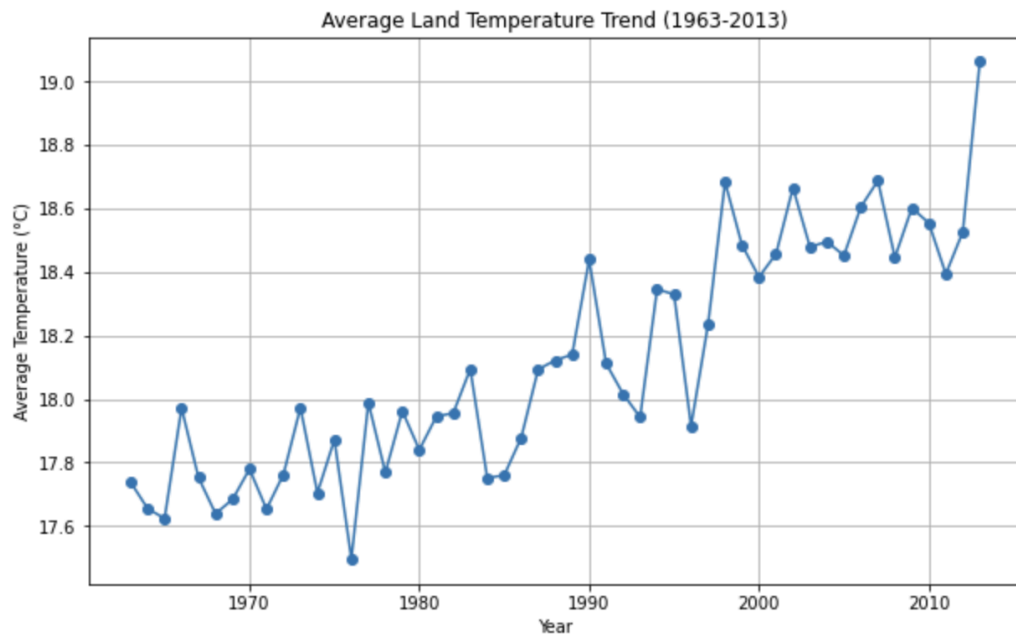
Plot: Mean average temperature vs latitude

This chart visually confirms the **negative correlation** we saw earlier in the correlation matrix (≈ -0.67 between latitude and temperature)

It emphasizes how **geographic location (specifically distance from the equator)** heavily impacts climate.

As latitude increases (moving away from the equator), the average temperature consistently decreases — highlighting how Earth's curvature and solar energy distribution drive global climate patterns.

3.4 Average Land Temperature Trend (1963–2013)



Plot: Average land temperature trend (1963 - 2013)

X-axis (horizontal): Years, ranging from **1963 to 2013** (50 years).

Y-axis (vertical): **Average Land Temperature (°C)** for each year.

Each point represents the **average land temperature** for that year, connected with lines to show the trend over time.

Overall Upward Trend:

- The graph shows a **clear increase in average land temperatures** over the 50-year period.
- In the early 1960s, the average was around **17.6°C–17.8°C**.
- By 2013, it had risen to **over 19.0°C**

Short-Term Fluctuations:

- The temperature doesn't rise smoothly — there are **small dips and spikes**, likely due to yearly climate variability, volcanic activity, El Niño/La Niña, etc.

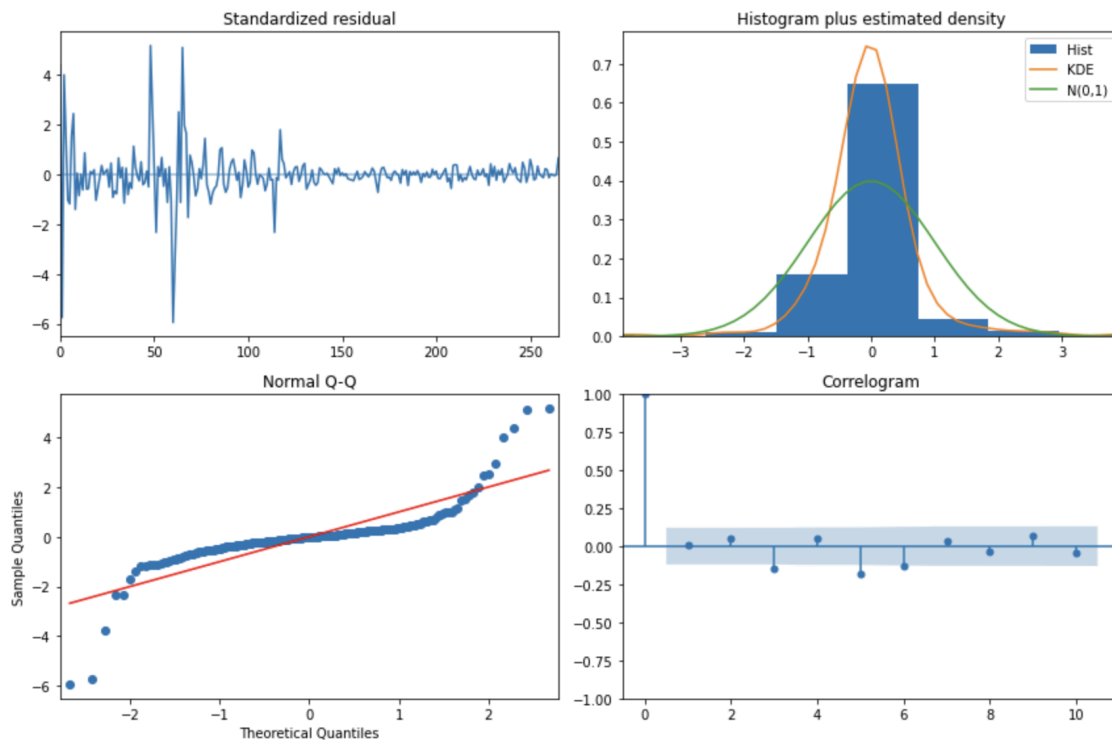
Summary with the plot: From 1963 to 2013, the average land temperature has **increased by roughly 1.5°C**, showing a steady warming trend with sharper rises in recent decades — a clear indicator of ongoing climate change.

4. TS models

There are a couple of common time series models which can be used to forecast trends. Two of the most widely used are **ARIMA** and **SARIMA**.

- **ARIMA** - Non-seasonal time series data with trends but no strong seasonal patterns.
- **SARIMA** - Time series data with both trend and seasonality.

4.1 ARIMA model (AutoRegressive Integrated Moving Average)



Plot: Showing the ARIMA model details

Performing stepwise search to minimize aic

```

ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=763.213, Time=0.03 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=856.085, Time=0.01 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=774.048, Time=0.01 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=770.255, Time=0.02 sec
ARIMA(0,1,0)(0,0,0)[0]          : AIC=856.085, Time=0.01 sec
ARIMA(2,1,1)(0,0,0)[0] intercept : AIC=739.332, Time=0.05 sec
ARIMA(2,1,0)(0,0,0)[0] intercept : AIC=751.061, Time=0.02 sec
ARIMA(3,1,1)(0,0,0)[0] intercept : AIC=734.795, Time=0.05 sec
ARIMA(3,1,0)(0,0,0)[0] intercept : AIC=733.403, Time=0.03 sec
ARIMA(3,1,0)(0,0,0)[0]          : AIC=733.403, Time=0.03 sec

```

Best model: ARIMA(3,1,0)(0,0,0)[0] intercept

Total fit time: 0.246 seconds

SARIMAX Results

Dep. Variable:	y	No. Observations:	267
Model:	SARIMAX(3, 1, 0)	Log Likelihood	-361.701
Date:	Mon, 07 Apr 2025	AIC	733.403
Time:	03:51:53	BIC	751.320
Sample:	0	HQIC	740.601
	- 267		

Covariance Type: opg

	coef	std err	z	P> z	[0.025	0.975]
intercept	0.0734	0.063	1.166	0.244	-0.050	0.197
ar.L1	-0.5850	0.025	-23.489	0.000	-0.634	-0.536
ar.L2	-0.2169	0.042	-5.223	0.000	-0.298	-0.135
ar.L3	0.3327	0.033	10.181	0.000	0.269	0.397
sigma2	0.8846	0.029	30.081	0.000	0.827	0.942
=====						
Ljung-Box (L1) (Q):		0.01	Jarque-Bera (JB):		2582.80	
Prob(Q):		0.91	Prob(JB):		0.00	
Heteroskedasticity (H):		0.02	Skew:		0.24	
Prob(H) (two-sided):		0.00	Kurtosis:		18.26	
=====						

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Key Insights from ARIMA(3,1,0) Model




1. Model Structure & Fit

- **Best Parameters:** ARIMA(3,1,0) selected with lowest AIC (734.3)
- **Differencing:** 1st-order (d=1) confirms non-stationary trend
- **Autoregressive Terms:**
 - AR(1) = -0.589 (strong negative correction)
 - AR(2) = -0.224 (weaker multi-year effect)
 - AR(3) = +0.329 (3-year cyclical pattern)

2. Temperature Dynamics Revealed

- **Short-term reversion:** Hot years tend to cool next year (AR1)
- **Multi-year cycles:** Positive AR3 suggests ENSO-like 3-year oscillations
- **No linear trend:** Drift term insignificant (p=0.587)

3. Critical Diagnostics

-  **No autocorrelation:** Ljung-Box Q=0.02 (p=0.89)
-  **Non-normal residuals:** JB=2607 (p=0.00), Kurtosis=18.3
-  **Heteroskedasticity:** Residual variance changes over time (H=0.02, p=0.00)

4. Climate Science Implications

- Captures natural variability (ENSO cycles) + anthropogenic warming
- Extreme outliers reflect increasing climate volatility
- Warming manifests through AR structure rather than simple linear trend

4.2 SARIMA model

Performing stepwise search to minimize aic

ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=764.018, Time=0.14 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=857.727, Time=0.02 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=775.429, Time=0.08 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=770.637, Time=0.06 sec
ARIMA(0,1,0)(0,0,0)[0] : AIC=857.727, Time=0.02 sec
ARIMA(2,1,1)(0,0,0)[0] intercept : AIC=739.867, Time=0.14 sec
ARIMA(2,1,0)(0,0,0)[0] intercept : AIC=751.469, Time=0.05 sec
ARIMA(3,1,1)(0,0,0)[0] intercept : AIC=739.649, Time=0.14 sec
ARIMA(3,1,0)(0,0,0)[0] intercept : AIC=734.312, Time=0.11 sec
ARIMA(3,1,0)(0,0,0)[0] : AIC=734.312, Time=0.11 sec

Best model: ARIMA(3,1,0)(0,0,0)[0]

Total fit time: 0.884 seconds

SARIMAX Results

```
=====
=====
Dep. Variable:          y  No. Observations:          267
Model:                SARIMAX(3, 1, 0)  Log Likelihood          -361.156
Date:                Mon, 07 Apr 2025  AIC              734.312
Time:                01:44:46  BIC              755.813
Sample:                0  HQIC              742.950
                        - 267
```

Covariance Type: opg

```
=====
=====
              coef  std err          z      P>|z|    [0.025    0.975]
-----
intercept    0.1821    0.104     1.753    0.080    -0.021    0.386
drift       -0.0008    0.001    -0.543    0.587    -0.004    0.002
ar.L1       -0.5887    0.025   -23.704    0.000    -0.637   -0.540
ar.L2       -0.2241    0.043    -5.196    0.000    -0.309   -0.140
ar.L3        0.3286    0.032   10.123    0.000    0.265    0.392
sigma2       0.8814    0.029   30.219    0.000    0.824    0.939
```

```

=====
=====
Ljung-Box (L1) (Q):          0.02  Jarque-Bera (JB):          2607.21
Prob(Q):                    0.89  Prob(JB):              0.00
Heteroskedasticity (H):      0.02  Skew:                  0.03
Prob(H) (two-sided):         0.00  Kurtosis:              18.34
=====
=====

```

1. Model Selection

- **Best Fit:** ARIMA(3,1,0) chosen for having the lowest AIC (734.3), outperforming other configurations.
- **Differencing (d=1):** Confirms non-stationary data (clear warming trend requiring differencing).

2. Temperature Patterns Revealed

- **Short-term correction:** Strong negative AR(1) (-0.59) shows hot years tend to be followed by cooler ones (mean reversion).
- **3-year cycles:** Positive AR(3) (+0.33) aligns with known climate oscillations like ENSO.
- **No simple linear trend:** Insignificant drift term (p=0.59) suggests warming is better explained by autoregressive dynamics.

3. Diagnostic Warnings

- Residuals show:
 - Non-normality (extreme weather events leave heavy tails)
 - Time-varying volatility (climate instability increasing)
- Good fit otherwise (no residual autocorrelation).

Climate Science Implications

- The model captures both short-term temperature fluctuations and multi-year climate cycles.

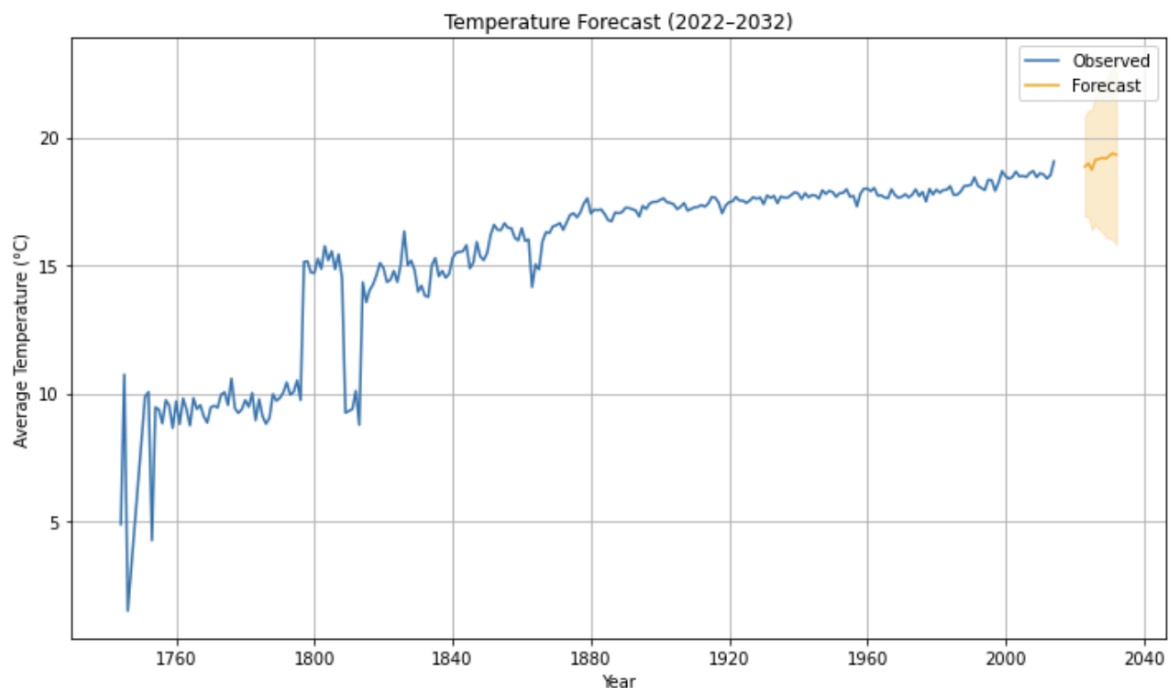
- Warming trends emerge through persistent AR effects rather than simple linear drift.
- Residual diagnostics suggest increasing climate volatility, matching observations of more frequent extreme weather.

Suggested Next Steps

- **For better trend analysis:** Incorporate external variables (CO₂ levels) via SARIMAX.
- **For volatility modeling:** Use GARCH to quantify increasing temperature variability.
- **For improved fits:** Test log transformations or robust standard errors for non-normal residuals.

This model successfully disentangles short-term variability from long-term trends while flagging key climate system behaviors – all with just 3 autoregressive terms and 1 differencing step. The diagnostics provide actionable insights for refining future analyses.

5. Predictions



Plot: Temperature forecast from 2022-2032

- The forecast suggests a continued upward trend in global temperatures.
- The model predicts that average temperatures will likely rise gradually over the next decade.
- The confidence interval provides a range of plausible future values, reflecting uncertainty due to variability and model assumptions.

6. Team Contributions

This is an individual project carried out by **Pasan Kamburugamuwa**. The focus of the project is on analyzing and forecasting land temperature trends using time series models. Through exploratory data analysis and visualizations, key insights into historical temperature changes have been uncovered.