Analyzing Global Temperature Trends and Forecasting Future Climate Change

Yue Zhu (yue_zhu@ucsb.edu)

2025-03-16

Abstract:

Climate change is one of the most pressing global challenges, with rising temperatures impacting ecosystems, economies, and societies. This study analyzes global temperature changes from 1961 to 2019 using a dataset of monthly temperature anomalies from meteorological organizations. To account for the underlying trend, first-order differencing was applied before model selection. Using time series modeling, we identified ARIMA(0,0,1) as the optimal model for differenced data based on the Akaike Information Criterion (AIC) and diagnostic checks. Forecasts suggest a continued rise in global mean temperature, reinforcing the urgency of climate action. Additionally, spectral analysis confirms the long-term warming trend in the original data and reveals a periodic pattern of approximately 3.5 to 4 years in the differenced data. This study also demonstrates that incorporating external regressors in ARIMA models effectively captures multi-year periodic patterns, which traditional seasonal models (defined by within-year cycles) fail to detect. These findings contribute to a deeper understanding of long-term climate variability and can support climate policy and environmental planning.

Introduction

Project Overview and Purpose

Climate change has become one of the most urgent global concerns, affecting ecosystems, human health, and economic stability. Understanding long-term temperature variations is crucial for assessing climate risks and developing effective environmental policies. This project aims to analyze global temperature trends from 1961 to 2019 using a comprehensive dataset of monthly temperature anomalies. The objective is to investigate historical warming patterns, quantify climate variability, and forecast future temperature changes using time series modeling techniques.

Reason for Choosing the Topic and Dataset

The increasing frequency of extreme weather events and rising global temperatures underscore the need for rigorous climate data analysis. The dataset, sourced from meteorological organizations and climate research agencies, provides over five decades of monthly temperature anomalies across various regions. This dataset was chosen for its extensive time span and global coverage, enabling a robust analysis of climate trends and regional variations. By studying this dataset, we aim to contribute to the understanding of climate change and support data-driven decision-making for mitigation strategies.

Previous Studies on the Dataset

Numerous climate studies have utilized similar datasets to analyze global warming trends, model temperature anomalies, and evaluate the impact of greenhouse gas emissions. Prior research has demonstrated a significant upward trend in global temperatures, with notable variations across geographic regions. Researchers have also explored relationships between temperature anomalies and factors such as El Niño events, carbon emissions, and industrialization. A review of trend analysis methods in climate time series data has highlighted various statistical techniques used to detect and attribute changes in climate variables over time, emphasizing the importance of robust methodologies for accurate assessments (Mudelsee, 2019). Additionally, advancements in machine learning techniques, such as the signature method, have shown promise in improving climate time series forecasts, capturing complex temporal dependencies that traditional statistical models may overlook (Arribas-Bel et al., 2023). However, many studies have focused on broad statistical summaries rather than predictive modeling, leaving a gap in forecasting temperature changes with advanced time series methods.

Key Findings from the Data

Preliminary analysis confirms a consistent rise in global temperatures, with anomalies increasing significantly over the decades. The ARIMA(0,1,1) model was identified as the most effective in capturing the data's temporal dependencies, demonstrating a strong predictive capability for future temperature changes. The model forecasts continued warming, reinforcing concerns about ongoing climate change and its implications for environmental policy.

This study contributes to the broader discussion on climate change by providing a data-driven perspective on historical trends and future projections. The findings highlight the urgency of climate action and the importance of incorporating predictive analytics into climate research.

Data

Dataset Overview

The dataset used in this project contains monthly global temperature anomalies spanning from 1961 to 2019. It provides a comprehensive view of long-term climate trends, enabling the analysis of temperature variations over nearly six decades. The dataset includes temperature deviations from historical baselines, which are essential for identifying global warming patterns.

Dataset Details

The dataset spans from 1961 to 2019 and contains monthly temperature records for each year, providing a long-term view of climate trends. Temperature changes are measured in degrees Celsius (°C), representing deviations from historical norms. The dataset consists of 9,656 rows and 66 columns, including location information, month details, and temperature values, making it a comprehensive source for analyzing global temperature variations.

Background of the Dataset

The dataset is publicly available on Kaggle, a widely used platform for data science research. This dataset is part of global climate research efforts and is compiled from major meteorological organizations and climate research groups, including NASA's Goddard Institute for Space Studies (GISS), National Oceanic and Atmospheric Administration (NOAA) and World Meteorological Organization (WMO)

Dataset Display

```r
library(readr)
library(dplyr)
library(tidyr)
library(astsa)
library(zoo)
data <- read_delim("Environment_Temperature_change_E_All_Data_NOFLAG.csv", delim=',')
data
```

```
## # A tibble: 9,656 x 66
##    `Area Code` Area     `Months Code` Months `Element Code` Element Unit   Y1961
##          <dbl> <chr>            <dbl> <chr>           <dbl> <chr>   <chr>  <dbl>
##  1           2 Afghani~          7001 Janua~           7271 Temper~ "\xb~  0.777
##  2           2 Afghani~          7001 Janua~           6078 Standa~ "\xb~  1.95
##  3           2 Afghani~          7002 Febru~           7271 Temper~ "\xb~ -1.74
##  4           2 Afghani~          7002 Febru~           6078 Standa~ "\xb~  2.60
##  5           2 Afghani~          7003 March            7271 Temper~ "\xb~  0.516
##  6           2 Afghani~          7003 March            6078 Standa~ "\xb~  1.51
##  7           2 Afghani~          7004 April            7271 Temper~ "\xb~ -1.71
##  8           2 Afghani~          7004 April            6078 Standa~ "\xb~  1.41
##  9           2 Afghani~          7005 May              7271 Temper~ "\xb~  1.41
## 10           2 Afghani~          7005 May              6078 Standa~ "\xb~  1.23
## # i 9,646 more rows
## # i 58 more variables: Y1962 <dbl>, Y1963 <dbl>, Y1964 <dbl>, Y1965 <dbl>,
## #   Y1966 <dbl>, Y1967 <dbl>, Y1968 <dbl>, Y1969 <dbl>, Y1970 <dbl>,
## #   Y1971 <dbl>, Y1972 <dbl>, Y1973 <dbl>, Y1974 <dbl>, Y1975 <dbl>,
## #   Y1976 <dbl>, Y1977 <dbl>, Y1978 <dbl>, Y1979 <dbl>, Y1980 <dbl>,
## #   Y1981 <dbl>, Y1982 <dbl>, Y1983 <dbl>, Y1984 <dbl>, Y1985 <dbl>,
## #   Y1986 <dbl>, Y1987 <dbl>, Y1988 <dbl>, Y1989 <dbl>, Y1990 <dbl>, ...
```

Methodology

We will apply two methods to analyze the dataset.

SARIMA (p, d, q) x (P, D, Q) model

The Autoregressive Integrated Moving Average (ARIMA) model is a powerful time series forecasting technique that captures trends and patterns in sequential data. The modeling process begins with data preprocessing, where stationarity is assessed using statistical tests, and differencing is applied if necessary to remove trends and stabilize variance. Next, the appropriate autoregressive (p) and moving average (q) orders are determined by analyzing Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. The best-fitting model is then selected based on criteria such as the Akaike Information Criterion (AIC), ensuring a balance between complexity and accuracy. After fitting the model, diagnostic checks are performed to validate that the residuals resemble white noise, indicating a well-specified model. Finally, the trained ARIMA model is used to generate forecasts, providing valuable insights into future trends while capturing underlying patterns in the data.

Spectral Analysis

Spectral analysis is a technique used in time series analysis to identify and quantify periodic patterns by transforming the data from the time domain to the frequency domain. It detects dominant cycles by analyzing the strength of different frequencies, allowing researchers to estimate cycle lengths (T = 1/f) and separate trends from seasonal or high-frequency components.

Results

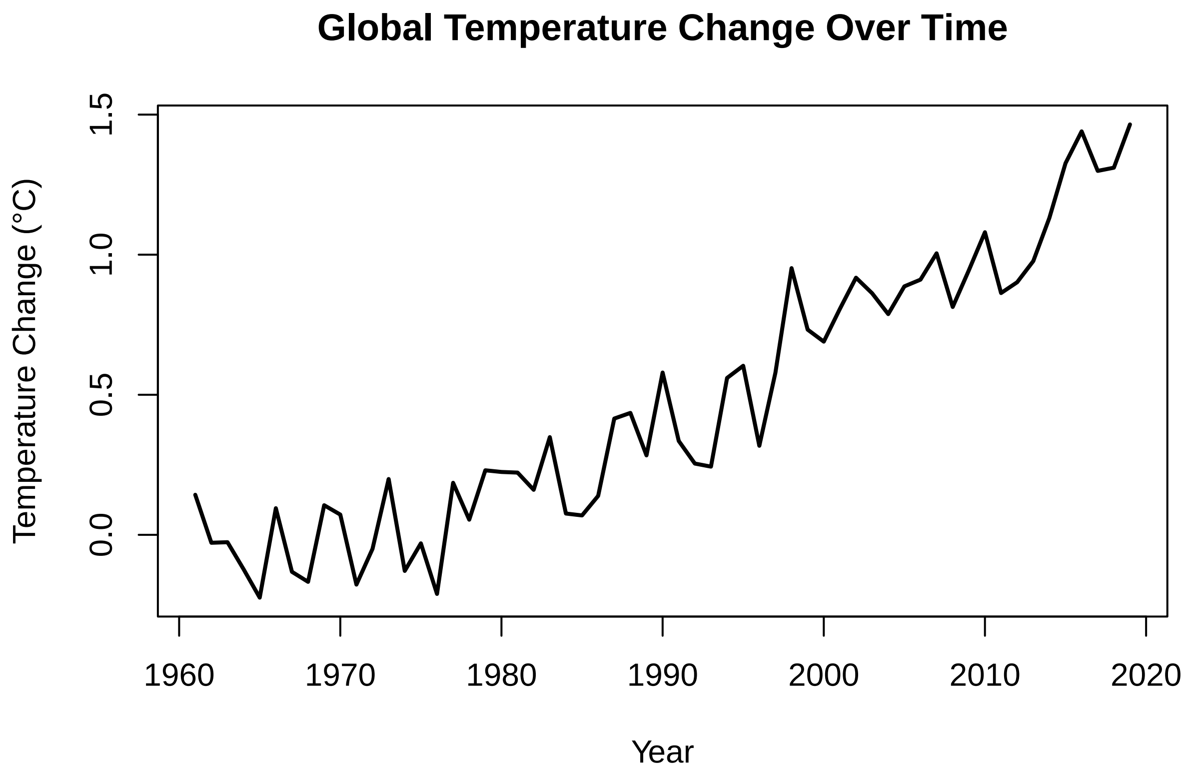
To analyze global temperature change trends, we do not focus on temperature variations in any specific region. Instead, we compute the global average temperature anomaly for each year to obtain an overall trend. This is achieved by first filtering the dataset to include only temperature change records and then grouping the data by geographic regions. For each year, we calculate the mean temperature change across all regions, ensuring that the final dataset represents a global-scale temperature trend rather than localized fluctuations.

```
df <- data %>%
    filter(Element == "Temperature change") %>%
    group_by(Area) %>%
    summarise(across(where(is.numeric), mean, na.rm = TRUE)) %>%
    select(starts_with("Y")) %>%
    summarise(across(everything(), mean, na.rm = TRUE)) %>%
    pivot_longer(cols = everything(), names_to = "ds", values_to = "y") %>%
    mutate(ds = as.integer(sub("Y", "", ds)))

df
```

```
## # A tibble: 59 x 2
##       ds       y
##    <int>   <dbl>
##  1  1961  0.143
##  2  1962 -0.0284
##  3  1963 -0.0263
##  4  1964 -0.123
##  5  1965 -0.224
##  6  1966  0.0951
##  7  1967 -0.132
##  8  1968 -0.168
##  9  1969  0.106
## 10  1970  0.0722
## # i 49 more rows
```

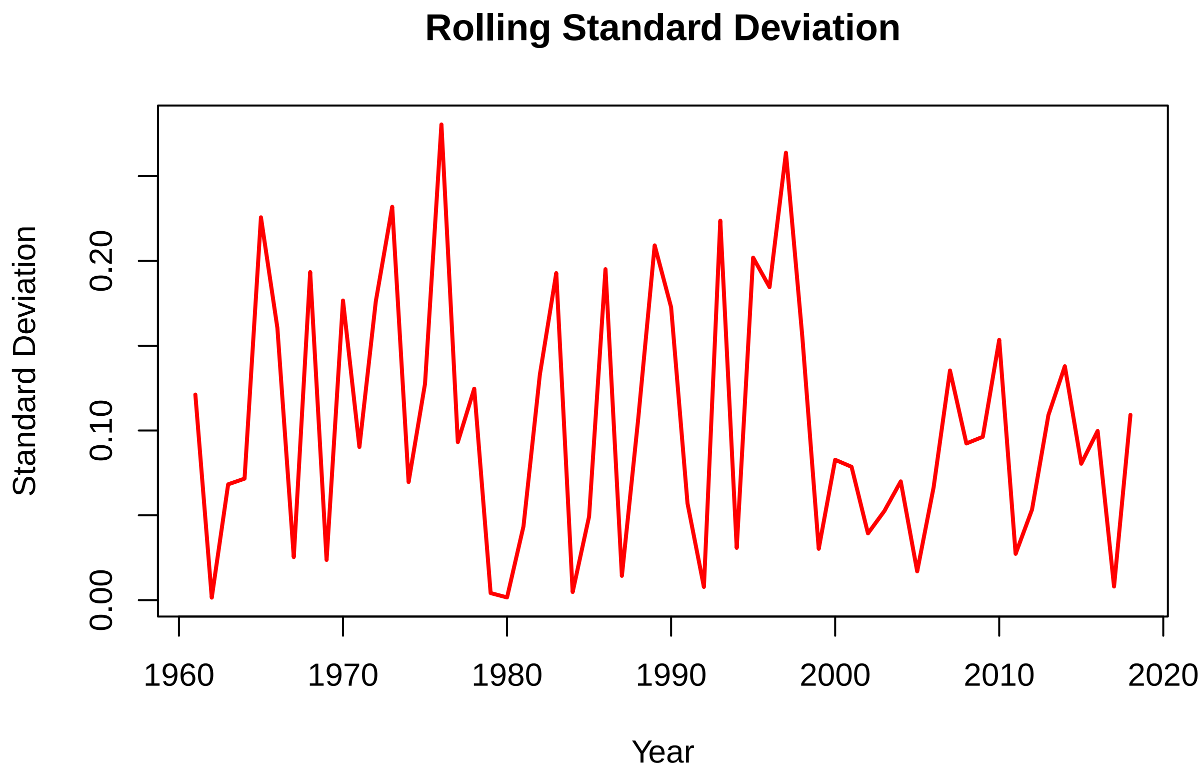
To visualize the overall trend in global temperature changes, we first convert the dataset into a time series object, setting the starting and ending years based on the dataset's range. The time series plot illustrates the annual global temperature anomalies from 1961 to 2019, showing a clear upward trend over time. While short-term fluctuations exist, the overall pattern indicates a significant rise in global temperatures, particularly in recent decades. This initial plot provides insight into long-term warming trends and serves as a foundation for further time series analysis and forecasting.

```
ts_data <- ts(df$y, start = min(df$ds), end = max(df$ds), frequency = 1)

plot.ts(ts_data,
        main = "Global Temperature Change Over Time",
        xlab = "Year",
        ylab = "Temperature Change (°C)",
        lwd = 2)
```



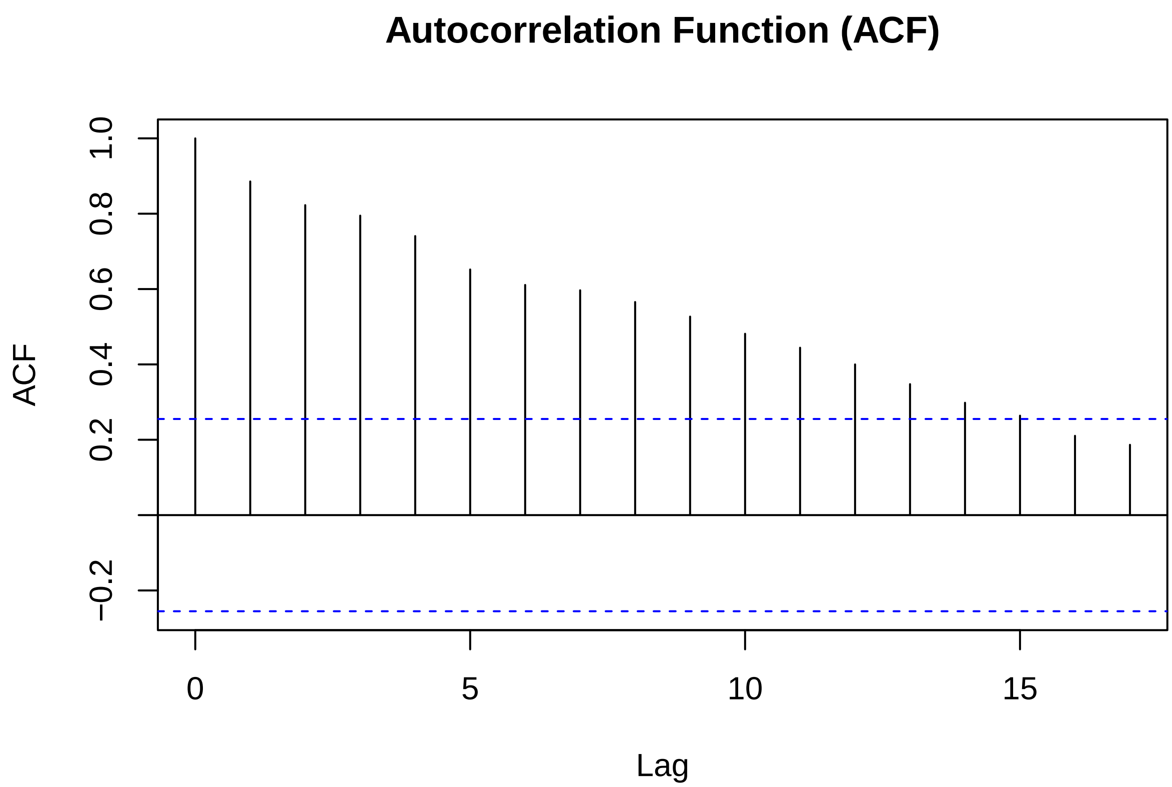
The plot displays the rolling standard deviation over time, with the line color set to red for clarity. The results show that while short-term fluctuations exist, there is no clear increasing trend in variability over time. This indicates that variance remains relatively stable, suggesting that no variance-stabilizing transformation (such as logarithmic or Box-Cox transformation) is necessary before further modeling. This step helps confirm that the dataset is suitable for ARIMA modeling without additional pre-processing.

```
rolling_sd <- rollapply(ts_data, width = 2, FUN = sd)

plot.ts(rolling_sd,
        main = "Rolling Standard Deviation",
        xlab = "Year",
        ylab = "Standard Deviation",
        col = "red",
        lwd = 2)
```

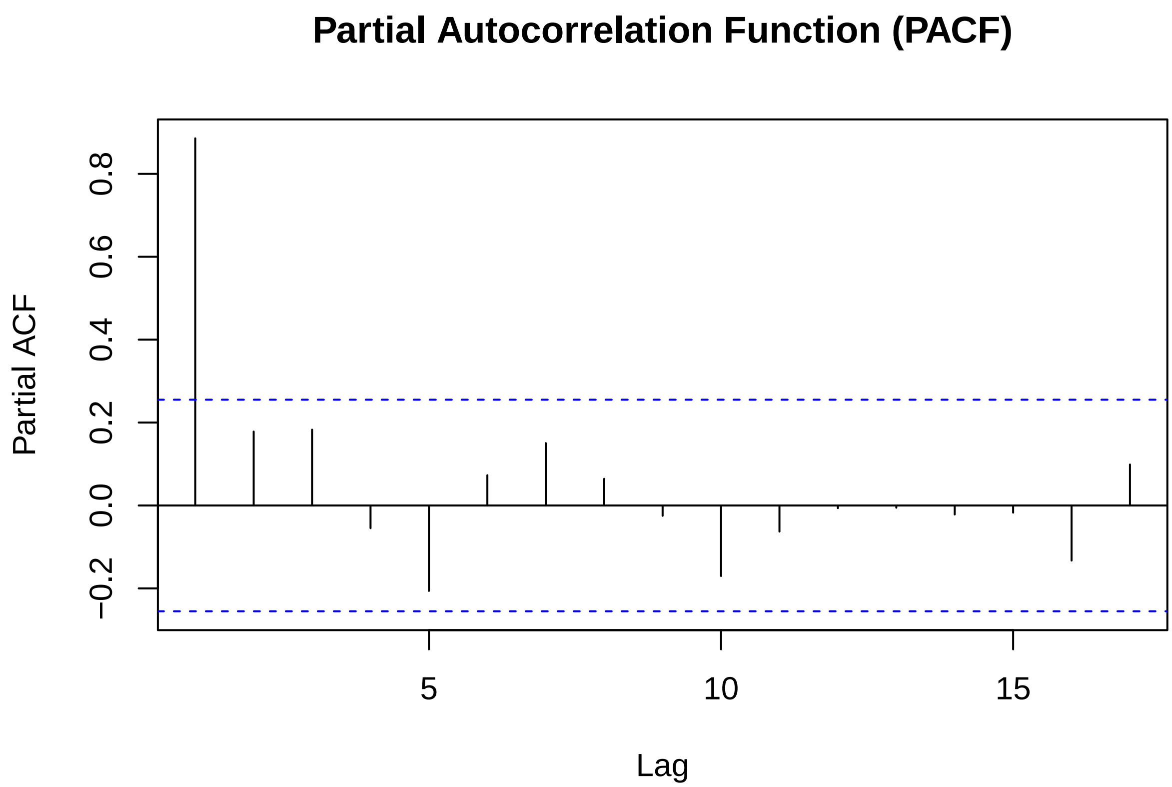


The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are used to analyze the dependency structure of the time series data. The ACF plot shows a slow decay rather than cutting off at a specific lag, indicating the presence of a trend and non-stationarity in the data. The PACF plot helps identify the autoregressive components by showing significant lags with direct correlations. Since the ACF suggests a non-stationary series, we need to apply differencing to remove trends and stabilize the time series before proceeding with ARIMA modeling.

```
acf(ts_data, main = "Autocorrelation Function (ACF)")
```

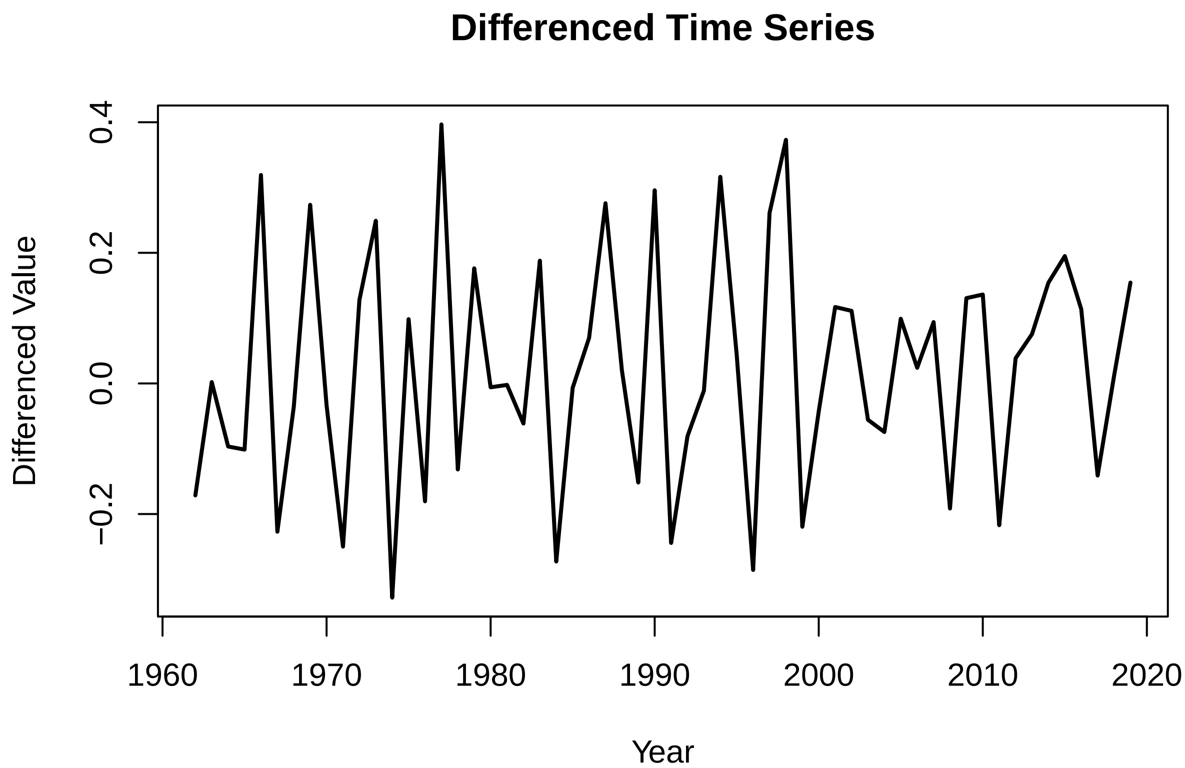



```r
pacf(ts_data, main = "Partial Autocorrelation Function (PACF)")
```

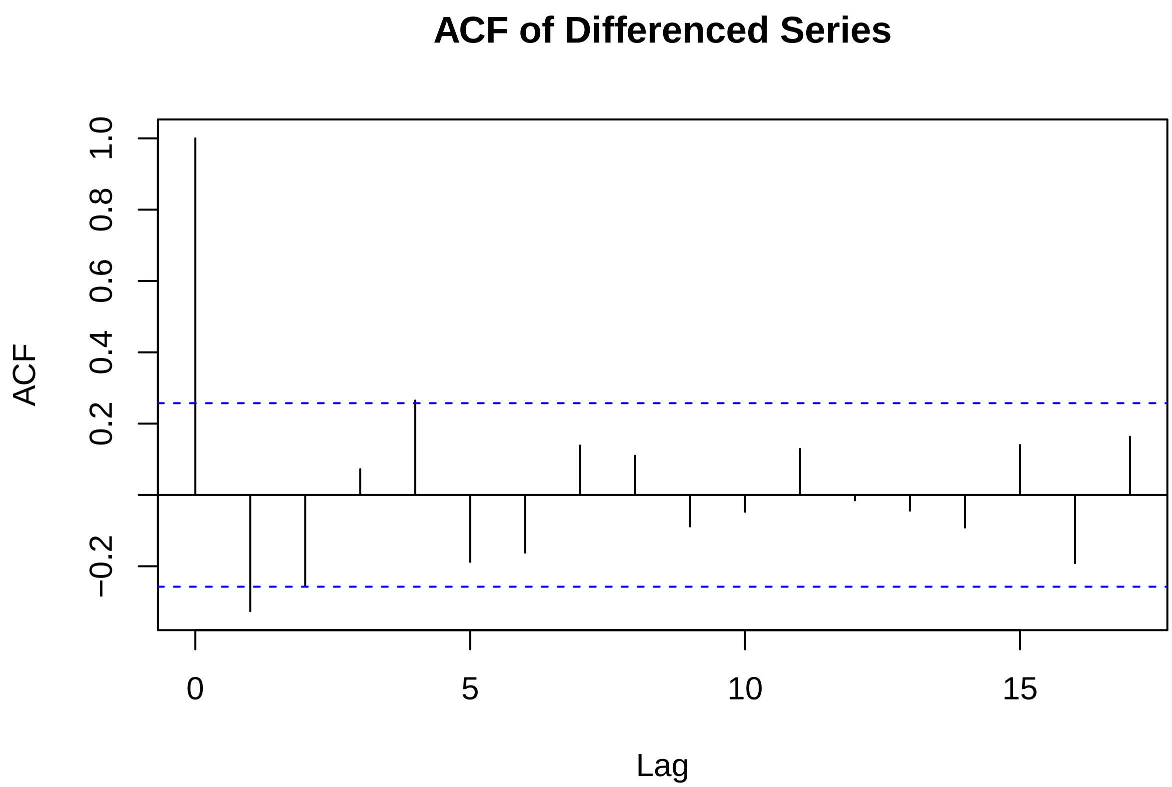


To address the non-stationarity identified in the previous ACF/PACF analysis, first-order differencing was applied to the original time series data.

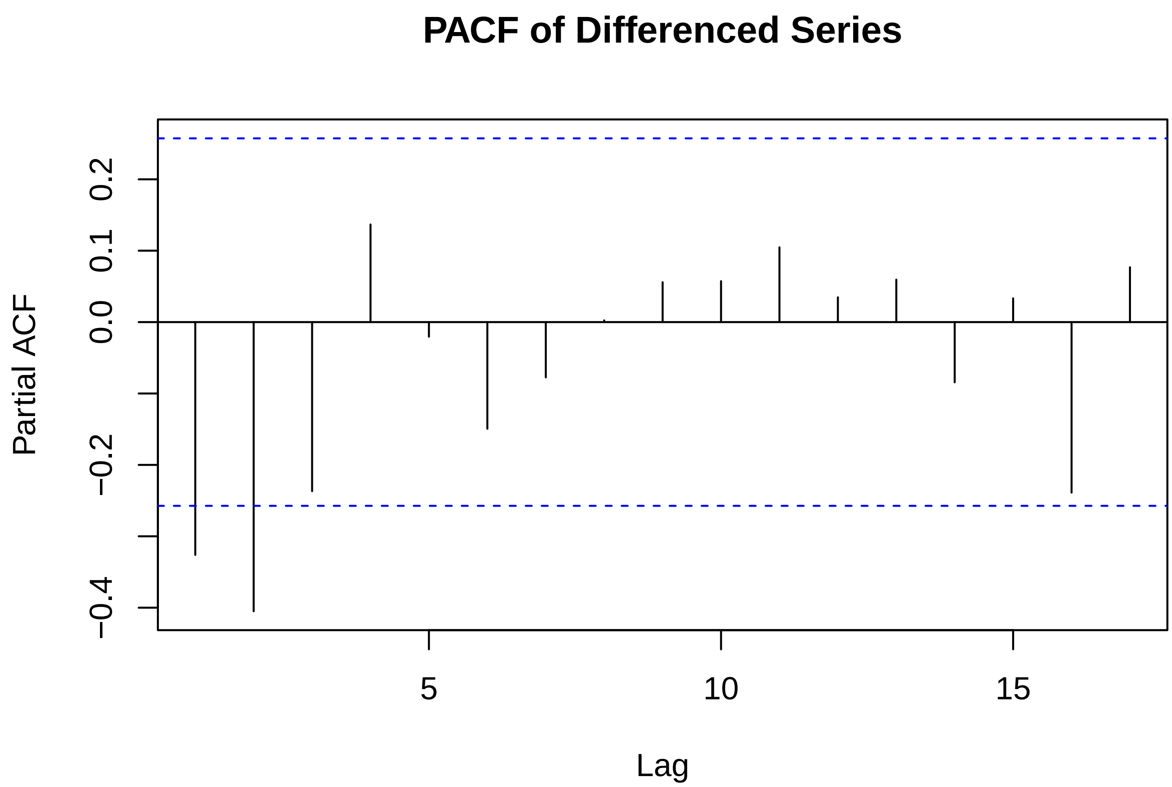
```
ts_diff <- diff(ts_data)

plot.ts(ts_diff,
        main = "Differenced Time Series",
        xlab = "Year",
        ylab = "Differenced Value",
        lwd = 2)
```




```r
acf(ts_diff, main = "ACF of Differenced Series")
```




```
pacf(ts_diff, main = "PACF of Differenced Series")
```



After applying first-order differencing to the time series, we examine the ACF and PACF plots to determine whether the transformation successfully removed trends and made the series stationary. Since no clear pattern remains in the differenced series, this confirms that d = 1 is an appropriate choice for the ARIMA model on the original data, ensuring that the data meets the stationarity requirement.

Additionally, we assess whether the data exhibits seasonality by looking for recurring patterns over the years. Since no obvious seasonal cycles appear in the dataset, a regular ARIMA model is more suitable than a seasonal ARIMA (SARIMA) model, simplifying the modeling process by focusing only on short-term and long-term dependencies without a seasonal component.

To determine the appropriate values for the autoregressive (p) and moving average (q) components, we used the auto.arima function from the forecast package, which conducts a search over possible models within the order constraints provided using AIC score and selects the best one for the given time series data.

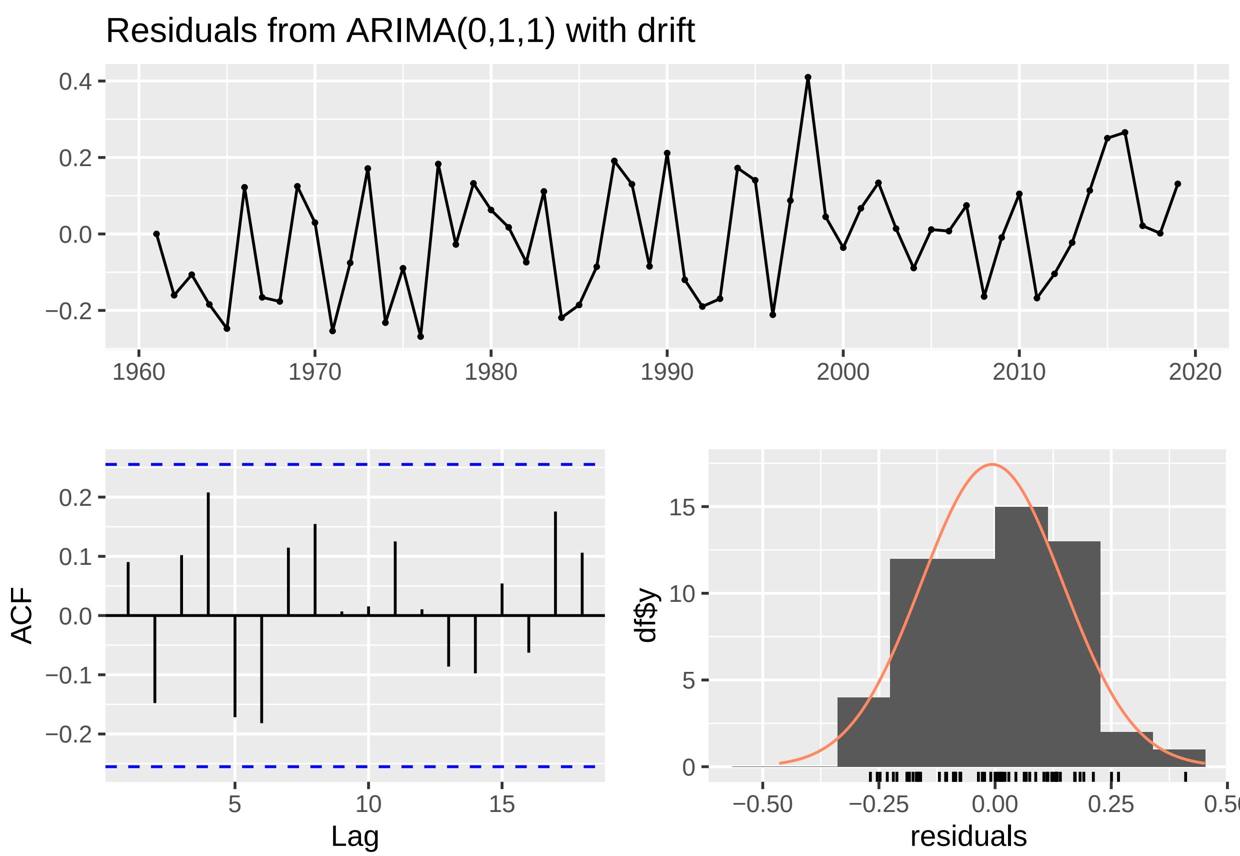
```r
library(forecast)

a <- auto.arima(ts_data, seasonal=FALSE, stepwise=FALSE)
a
```

```
## Series: ts_data
## ARIMA(0,1,1) with drift
##
## Coefficients:
##           ma1   drift
##       -0.7046  0.0247
## s.e.   0.1119  0.0062
##
```

```
## sigma^2 = 0.02418:  log likelihood = 26.32
## AIC=-46.64   AICc=-46.2   BIC=-40.46
```

We next generated the diagnostic plots for this model.

```r
checkresiduals(a)
```

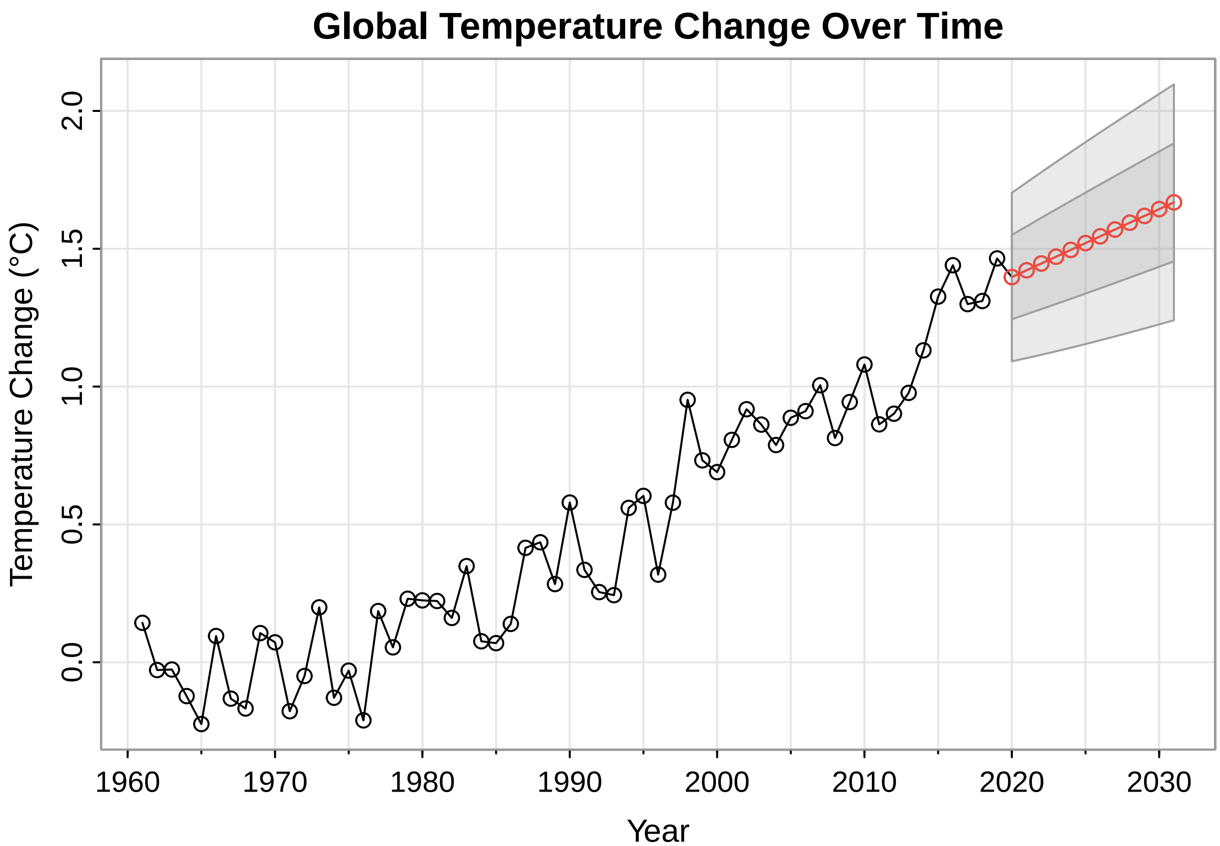



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1) with drift
## Q* = 12.199, df = 9, p-value = 0.2023
##
## Model df: 1.   Total lags used: 10
```

From the Ljung-Box test for normality of residuals, we can see that the p-value is not significant, indicating that the residuals are normally distributed, which meets the assumption of ARIMA models. From the residuals plot and residuals ACF plot, we can see that the residuals fluctuates randomly with no obvious pattern, suggesting that the residuals behave like white noise, which suggests that the model assumptions are likely met.

To forecast future global temperature changes, we apply the ARIMA(0,1,1) model to predict the next 12 years. The sarima.for() function generates forecasts based on historical data, incorporating three autoregressive terms, one differencing step, and one moving average term. The forecasted values provide insights into the expected trajectory of global temperature anomalies, helping to assess long-term climate trends.

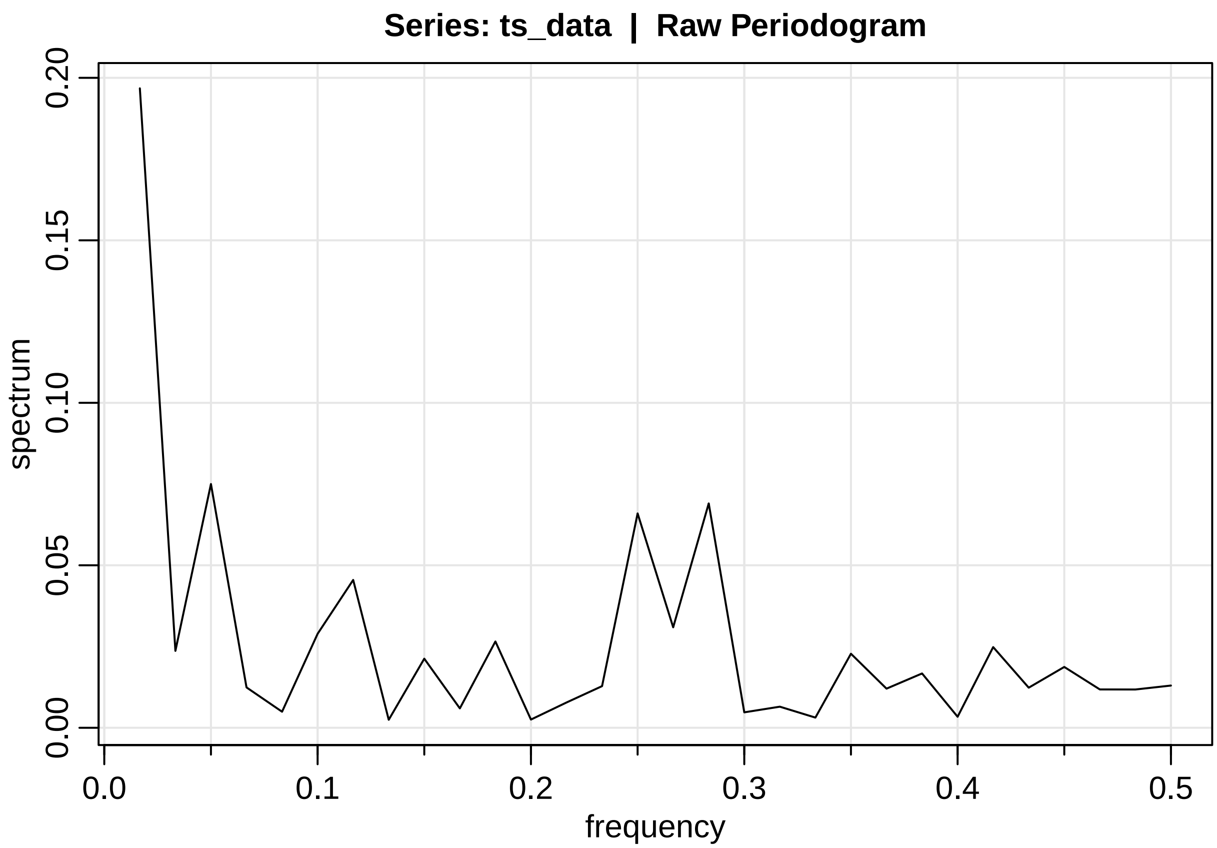
```r
sarima.for(ts_data, n.ahead = 12, p = 0, d = 1, q = 1,
           main = "Global Temperature Change Over Time",
           xlab = "Year",
           ylab = "Temperature Change (°C)")
```




```
## $pred
## Time Series:
## Start = 2020
## End = 2031
## Frequency = 1
##  [1] 1.397223 1.421874 1.446525 1.471176 1.495827 1.520478 1.545129 1.569780
##  [9] 1.594430 1.619081 1.643732 1.668383
##
## $se
## Time Series:
## Start = 2020
## End = 2031
## Frequency = 1
##  [1] 0.1527903 0.1593179 0.1655883 0.1716298 0.1774658 0.1831158 0.1885967
##  [8] 0.1939227 0.1991063 0.2041583 0.2090883 0.2139047
```

Next, spectral diagrams were plotted for both the original data and differenced data to inspect potential periodic patterns. The mvspec() function was used to transform time-series data from the time domain to the frequency domain. The plot for the original data shows a very high peak at frequency = 0.0167 (acquired by extracting the frequency with maximum spectrum value), which is a relatively low one. This indicates that there is a long-term trend in the data. From the time series plot before, we confirm that this trend is increasing.

```
library(astsa)

ori_spec <- mvspec(ts_data, log='n')
```

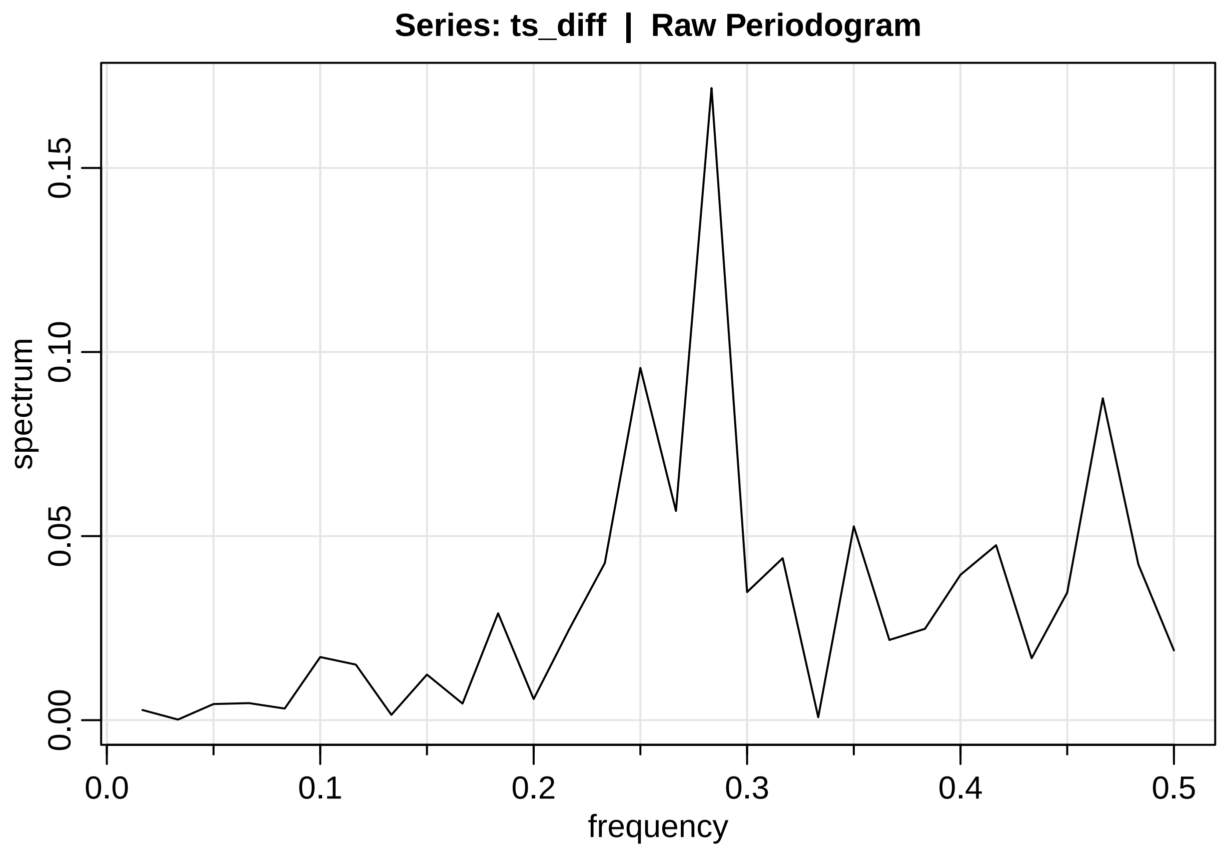



```
ori_spec$freq[which.max(ori_spec$spec)]
```

```
## [1] 0.01666667
```

The spectral diagram for the difference data shows the following few points. First, the long-term trend present in the original data was effectively removed by taking the difference (the peak at lower frequency range disappears). Second, the emerging dominant peak at frequency 0.283 (acquired by extracting the frequency with maximum spectrum value) may suggest that the differenced data contains a periodic pattern that warrants further investigation.

```
diff_spec <- mvspec(ts_diff, log='n')
```




```r
diff_spec_data <- data.frame(
  Frequency = diff_spec$freq,
  SpectralDensity = diff_spec$spec
)
diff_spec_data[order(-diff_spec_data$SpectralDensity), ]
```

```
##     Frequency SpectralDensity
## 17 0.28333333    0.1716941325
## 15 0.25000000    0.0956912184
## 28 0.46666667    0.0874239425
## 16 0.26666667    0.0568261089
## 21 0.35000000    0.0526430701
## 25 0.41666667    0.0475208847
## 19 0.31666667    0.0439999838
## 14 0.23333333    0.0426622752
## 29 0.48333333    0.0423008718
## 24 0.40000000    0.0394916855
## 18 0.30000000    0.0347957045
## 27 0.45000000    0.0346562689
## 11 0.18333333    0.0290262852
## 23 0.38333333    0.0248107322
## 13 0.21666667    0.0247140305
## 22 0.36666667    0.0217852638
## 30 0.50000000    0.0189619615
## 6  0.10000000    0.0171455633
```

```
## 26 0.43333333    0.0168239812
## 7  0.11666667    0.0150934127
## 9  0.15000000    0.0123838443
## 12 0.20000000    0.0057314071
## 4  0.06666667    0.0046264073
## 10 0.16666667    0.0044990935
## 3  0.05000000    0.0043803479
## 5  0.08333333    0.0031560549
## 1  0.01666667    0.0027638330
## 8  0.13333333    0.0014407098
## 20 0.33333333    0.0008015041
## 2  0.03333333    0.0001585603
```

Since periodic pattern was identified in the differenced data, it is meaningful to try to fit the pattern using time series models. We first tried to fit the pattern by setting seasonal=TRUE in auto.arima for the differenced data.

```r
library(forecast)
auto.arima(ts_diff, seasonal=TRUE)
```

```
## Series: ts_diff
## ARIMA(0,0,1) with non-zero mean
##
## Coefficients:
##           ma1    mean
##       -0.7046  0.0247
## s.e.   0.1119  0.0062
##
## sigma^2 = 0.02418:  log likelihood = 26.32
## AIC=-46.64   AICc=-46.2   BIC=-40.46
```

However, as can be seen from the result, the best model does not include a seasonal component, indicating that the periodic pattern was not detected. A possible explanation is that since the periodic pattern's unit is years, which is not what SARIMA models are designed to fit. Instead, SARIMA models are designed for seasonality that occurs within a single year.

To capture this multiple year periodic pattern, we created a new time series data based on the original one but with frequency > 1. In this way, we are pretending that the data has a unit time length of several years. The period was determined by calculating the reciprocal of the dominant frequency in the spectral diagram. Since the dominant frequency is 0.283, which means the corresponding period is $1/2.8333 \approx 3.5$, we set the frequency parameter to be the closest integer, 4.

```r
ts_seasonal <- ts(ts_diff, frequency=4)
```

Since ARIMA models are not designed for this kind of periodic pattern, we decided to incorporate external regressors in auto.arima to fit the pattern. The external regressors help the arima model fitting by incorporating additional information beyond the time series itself. This was done by using K terms from a fourier series comprised of sine and cosine functions. Since K must be an integer less than or equal to half of the frequency of the time series, we experimented with K = 1 and 2.

Note that when including external regressors to fit the periodic pattern, we turned off the seasonal switch of the auto.arima function because the periodic pattern is already captured by the external regressors so SARIMA's seasonal differencing may overcorrect or distort the pattern.

```r
auto.arima(ts_seasonal, seasonal=FALSE, stepwise=FALSE, xreg=fourier(ts_seasonal, K=1))
```

```
## Series: ts_seasonal
## Regression with ARIMA(0,0,1) errors
##
## Coefficients:
##           ma1  intercept    S1-4    C1-4
##       -0.6819     0.0245  0.0728  0.0485
## s.e.   0.1162     0.0063  0.0322  0.0323
##
## sigma^2 = 0.02232:  log likelihood = 29.73
## AIC=-49.45   AICc=-48.3   BIC=-39.15
```

```r
auto.arima(ts_seasonal, seasonal=FALSE, stepwise=FALSE, xreg=fourier(ts_seasonal, K=2))
```

```
## Series: ts_seasonal
## Regression with ARIMA(0,0,1) errors
##
## Coefficients:
##           ma1  intercept    S1-4    C1-4    C2-4
##       -0.6809     0.0245  0.0734  0.0490  0.0255
## s.e.   0.1168     0.0063  0.0321  0.0321  0.0315
##
## sigma^2 = 0.02249:  log likelihood = 30.05
## AIC=-48.11   AICc=-46.46   BIC=-35.74
```

As can be seen from the results, both models work better by giving a lower AIC score than a non-seasonal ARIMA model. However, since the AIC and BIC scores for K = 1 model is lower, it is preferred. This indicates that incorporating external regressors can help fit the periodic pattern in the data and such pattern does exist, although slight, in the data. This means that the global mean temperature changes may have a periodic pattern of roughly 4 years.

Conclusions

This study used spectral analysis and time series modeling to examine variations in global temperatures between 1961 and 2019. The results supported a long-term warming trend, and projections showed that global mean temperatures would continue to rise. The model that best captured temperature changes was found to be ARIMA(0,1,1), and its reliability was validated by diagnostic tests. Traditional seasonal ARIMA models, which are intended for within-year seasonality, were unable to capture the periodic pattern of roughly 3.5 to 4 years that was shown by spectral analysis. Rather, this long-term periodicity was effectively captured by adding Fourier terms as external regressors in an ARIMA model, indicating that global temperature anomalies show faint multi-year cycles. These findings underline the necessity of further study to enhance climate forecasting models and the significance of sophisticated time series approaches in climate analysis. Policymakers can create data-driven plans to lessen the effects of climate change by having a thorough understanding of both long-term trends and sporadic changes.

Reference

1. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2024). forecast: Forecasting functions for time series and linear models. R package version 8.23.0, https://pkg.robjhyndman.com/forecast/.

2. Shumway, R. H., & Stoffer, D. S. (2025). Time Series Analysis and its applications. In Springer texts in statistics. https://doi.org/10.1007/978-3-031-70584-7
3. Shumway, R. H., & Stoffer, D. S. (2019). Time Series: A Data Analysis approach using R. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9780429273285
4. Temperature change statistics 1961–2023. (2024). https://doi.org/10.4060/cd0082en
5. Arribas-Bel, D., Signer, J., & Wang, H. (2023). Improved climate time series forecasts by machine learning and signature methods. Ecological Informatics, 73, 101996. https://doi.org/10.1016/j.ecoinf.2023.101996
6. Mudelsee, M. (2019). Trend analysis of climate time series: A review of methods. Earth-Science Reviews, 190, 310-322. https://doi.org/10.1016/j.earscirev.2018.12.005