THE CLIMATE ANTHROPOCENE: A TIME SERIES PERSPECTIVE

SUBMITTED TO: PROF. AMIT MITRA

Divya Gupta
Statistics and Data Science
Department of Mathematics & Statistics
Indian Institute of Technology, Kanpur

Palak Mishra
Statistics and Data Science
Department of Mathematics & Statistics
Indian Institute of Technology, Kanpur

Kumar Kanishk Singh
Statistics and Data Science
Department of Mathematics & Statistics
Indian Institute of Technology, Kanpur
kksingh21@iitk.ac.in

Kundan Kumar
Statistics and Data Science
Department of Mathematics & Statistics
Indian Institute of Technology, Kanpur
kundank21@iitk.ac.in

Siddharth Pathak
Statistics and Data Science
Department of Mathematics & Statistics
Indian Institute of Technology, Kanpur
siddharthp21@iitk.ac.in

November 14, 2023

1 Introduction

While some claim that climate change is the greatest threat facing humanity, others believe that it is a hoax based on dubious science. In this project, we are going to analyze the trends of the global land temperature, ranging from the year 1743 to 2013. To achieve this, we are using two datasets, namely **Global Land Temperature by City**, and **Global Land Temperature by Country**. Further, we are going to do an in-depth analysis of the India dataset for the same. We will perform ARIMA model fitting, residual analysis, validation & forecasting from the collected data. Note that the second dataset i.e. 'Global Land Temperature by Country' composed of data from all countries, we are going to perform the analysis of climate change in India only. The project leverages the open nature of Berkeley Earth's data, including source data and transformation code, promoting transparency and reproducibility in scientific analysis.

2 Motivation

There are plenty of reasons, why climate change should be studied. Here we are mentioning some of them:

- It is recognized as one of the most critical challenges globally. So we need a system to monitor it.

- Temperature data provides clear-cut evidence of global warming, a key aspect of climate change. By examining long-term temperature trends, scientists can identify patterns that indicate a warming climate.
- It helps researchers understand climate patterns and variations. It allows them to identify regions that are experiencing more significant temperature increases or changes in temperature extremes, helping to predict and adapt to potential impacts.
- Temperature variations impact agriculture by affecting growing seasons, crop yields, and the prevalence of pests and diseases.

3 Data Pre-Processing

As mentioned above we used the Berkeley Earth dataset, which is available on Kaggle, for our analysis. The **Global Land Temperature by City** has 8,599,212 rows and 6 columns. There are many missing values from 1743 to 1849, so we use only the data from 1850 to 2013. We have imputed our data for the remaining missing values present in our data. We finally got 6,734,783 rows in total. We got 2,894 rows for the year 1850 and 3490 rows for the year 2013.

We take the data from **Global Land Temperatures by Country** and separate the data for India from it. Then we remove the data for the years before 1850. We split the date **column (dt)** into Year, Month and Day for India data. After that, we checked for NA values in the Average Temperature Column and Imputed those NA values with the method pmm (Predictive Mean Matching). We then converted the data to a **ts** dataset.

4 Data Analysis

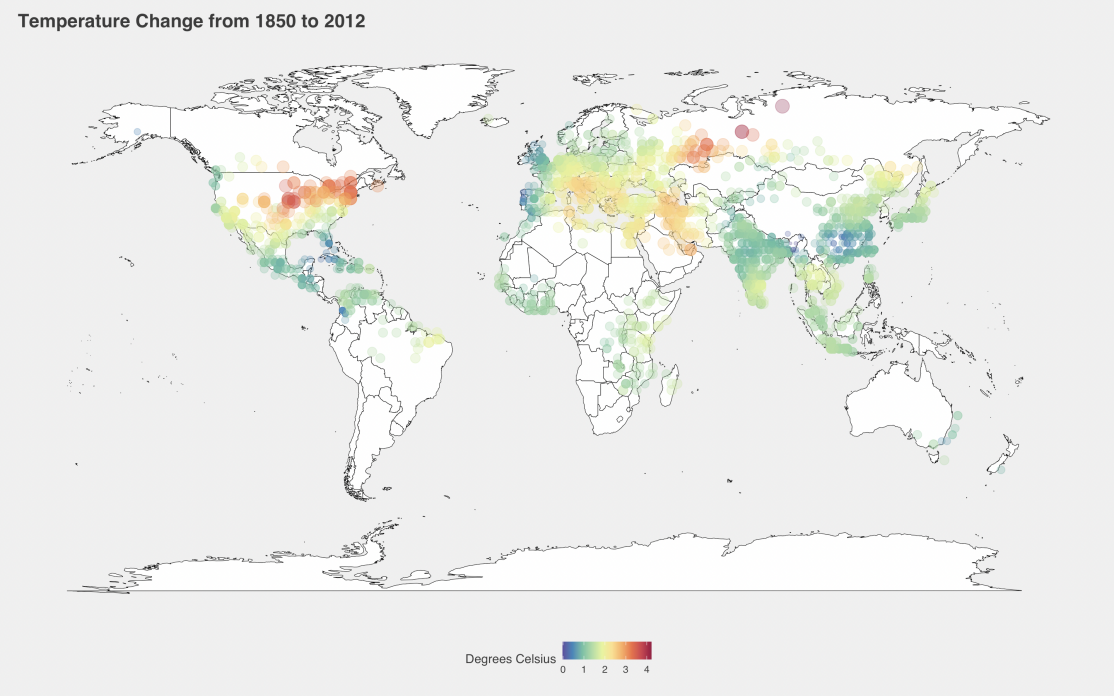

Figure 1: Temperature Change over the Globe

4.1 Temprature Change over the Globe

We use the data on cities to plot a world map and observe the distribution of changes in temperature over the globe. Western countries like US, Russia and some European countries showing red color indicate heavy change in temperature. However, India didn't show any drastic change when the temperatures were compared.

4.2 Temperature Change in India

In our analysis, we opt to analyse climate change in India only. So, we extract Indiaś dataset and preprocess it.

4.2.1 Checking for Trend and Seasonality

We decompose a time series into seasonal, trend and irregular components using moving averages. We can check trends, seasonality and randomness of the data doing this.

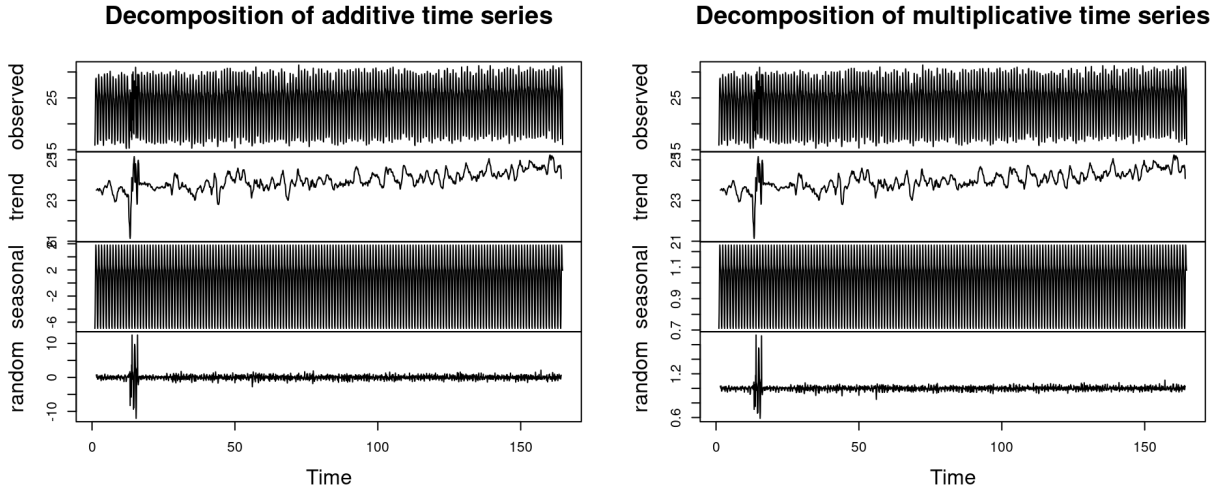

Figure 2

Trend can clearly be seen through the above plots. We see that error term in the multiplicative model has mean 1(approx.) and that of additive model is 0(approx.) So, We are taking additive model for further analysis.

We remove the trend by first order differencing and take a look at the average temperature by month plot for data from 1850-2013 and Seasonality for 2000-2013. We can clearly see that a seasonal component is present.
We can also see that for an year, the temperature increases to a peak which can be considered as the month of May and then decreases.

We can further confirm the presence of trend in the data by looking at the following plots:

We can again see that the month of may has the highest temperature range.

This box plot show temperature steadily has increased. Above 5 years data was taken as a

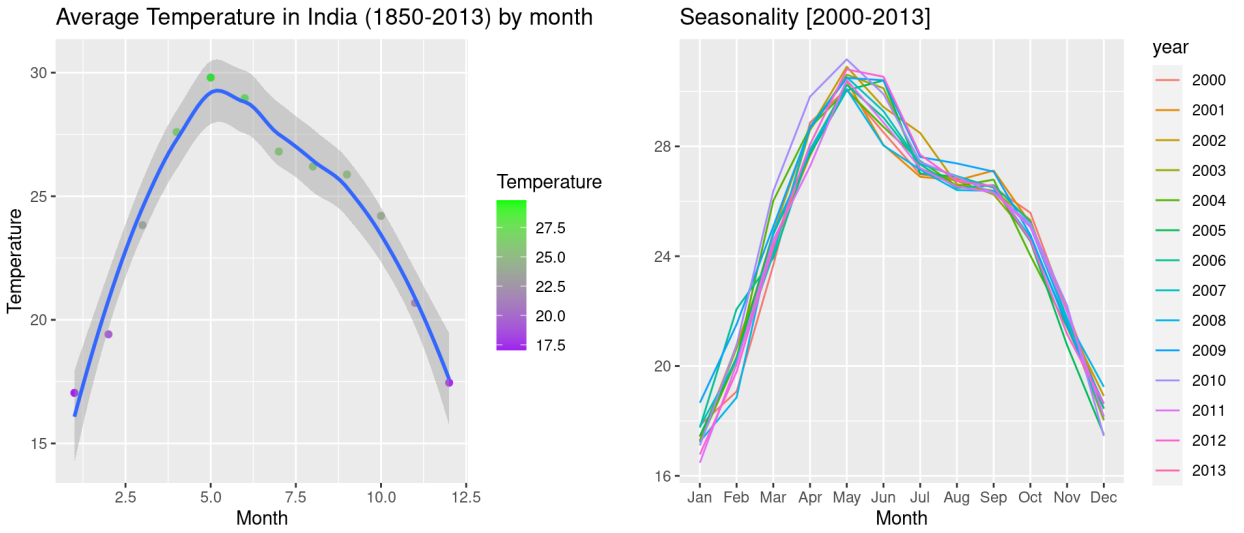

Figure 3

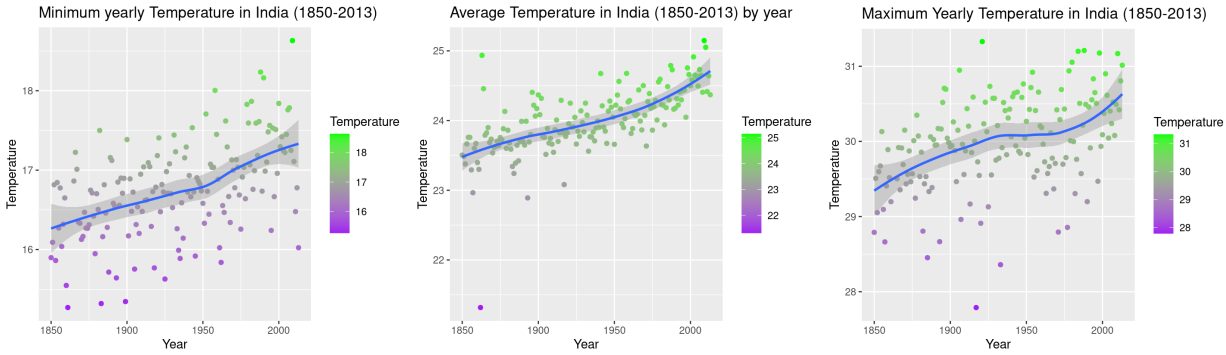

Figure 4

difference of 40 years. It is observable that that both the range and median increase.This further confirms the trend.

5 Methodology

- **model1** : Defined on ts object **avg_time**, storing yearly mean average temperature of India
- **model2** : Defined for ts object ts object **monthly _avg_time**, considering monthly average temperature of India from 1850 to 2013

5.1 Trend Estimation

In order to check for Trend, We have used **Mann-Kendall test** (though we can easily see trends from plot). With Null Hypothesis: there is no trend and alternate: Trend is present.
Testing on **avg_time** gives p-value < 0.05, hence trend is present. So we first checked for differencing order=1, we finally get p-value >0.05, hence now our series is detrended with order of differencing = 1
Now Testing on monthly_avg_time, we get p-value<0.05, hence our series has a trend component. Using differencing of order 1, we have obtained a detrended series

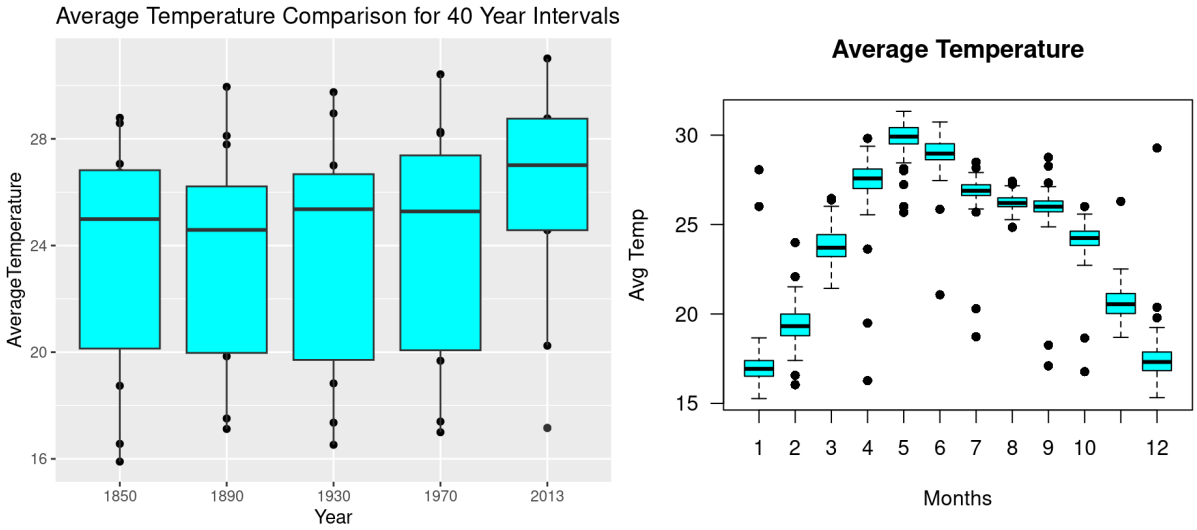

Figure 5

5.2 Stationarity Estimation

In order to check stationarity we have used **augmented Dickey-Fuller test (ADF)** test with h_0: Non-stationary and h_1: Stationary
For **avg_time**, We are getting p-value<0.05 for the detrended series i.e. d=1.
For **monthly_avg_time**, we are getting p-value<0.05, which means we have stationary series for our detrended series

5.3 Randomness Estimation

We will be using **Turning point test** to get randomness with the alternative hypothesis: non-randomness.
Both avg_time and monthly_avg_time have p-value <0.05, so our both are non-random.

6 Forecasting

6.1 Model

We need to fit a model on our data to be able to forecast future values. It is clear that the time series used to fit model1 is not stationary. When we fit ARIMA model, we get $p = 3, d = 1, q = 2$.

Plots have been attached to show the fit. For the non-stationary time series, all the lines in ACF graph are above the blue line.
For non-stationary time series, most of the lines fit between the blue lines in both ACF and PACF graph.

6.1.1 Model1

We now fit a model using **auto. arima()**. This model will fit multiple ARIMA models on our data and select the best model on the basis of AIC.

As we can see our both acf and pacf plots are tailed-off.Hence it is arima model. The best model obtained is ARIMA(3,1,2). The 1 in the center denotes the seasonality observed. It has the least

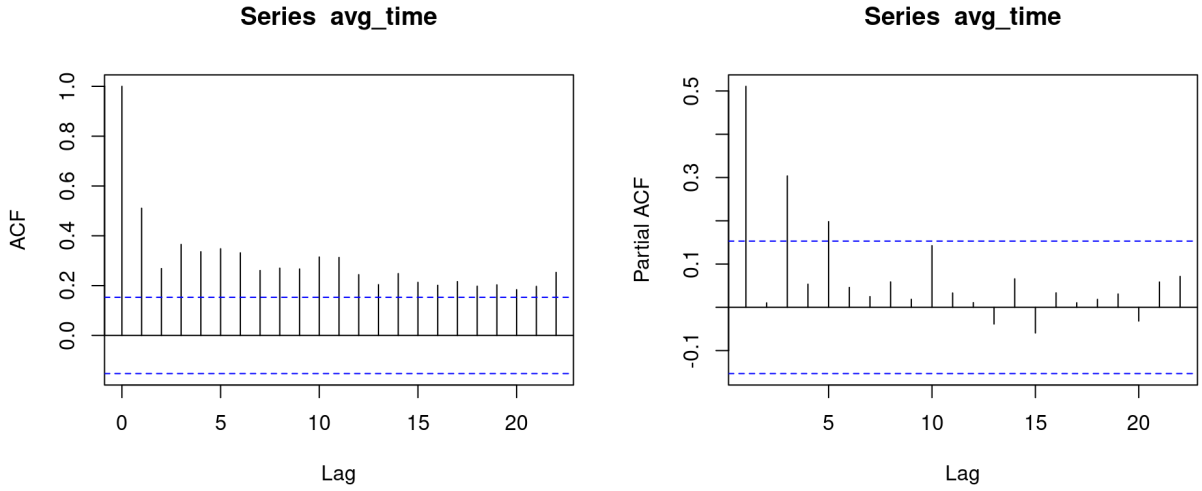

Figure 6: ACF and PACF plots for non-stationary time series

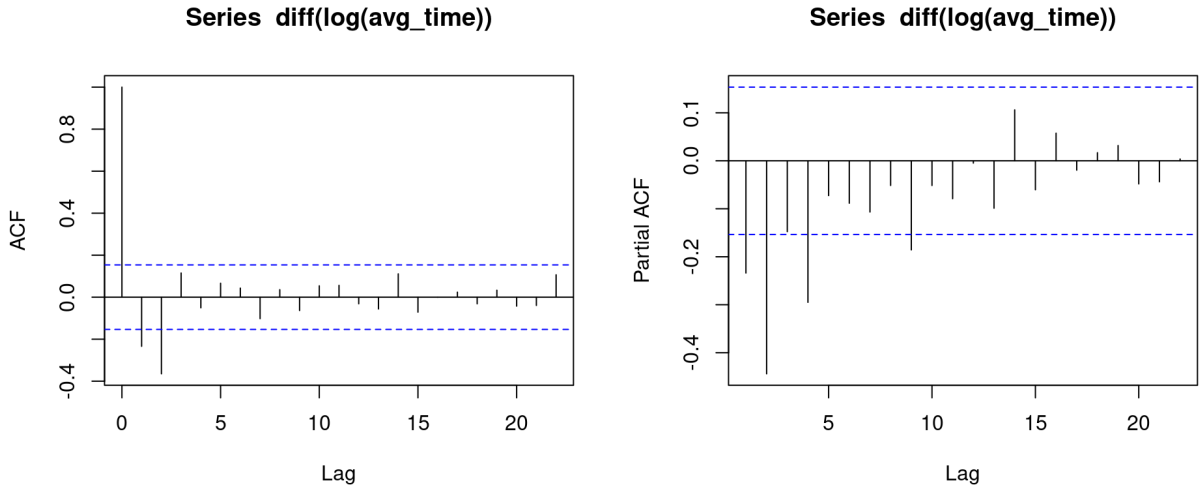

Figure 7: ACF and PACF plots for stationary time series

AIC and BIC Values and maximum log-likelihood confirming it to be the most suitable model.

6.1.2 Model2

In model2, for monthly_avg_time, we are getting ARIMA(2,1,0), order of differentiation is 1 as we have get from trend estimation. Also we can see pacf has cut-off after 2 lags, which verifies our results.

6.2 Validation

After fitting the model, we need to check the goodness-of-fit of the model. Box-Ljung test is used to check the fit of the model. The Box-Ljung test rejects the null hypothesis (indicating that the model has significant lack of fit) if $Q > \mathcal{X}^2_{1-\alpha,h}$ where $\mathcal{X}^2_{1-\alpha,h}$ is the chi-square distribution table

```
Series: log(avg_time)
ARIMA(3,1,2)

Coefficients:
          ar1      ar2     ar3      ma1      ma2
      -0.0799  -0.0263  0.0130  -0.4217  -0.4156
s.e.   0.3769   0.1591  0.1184   0.3694   0.3189

sigma^2 = 0.0001825:  log likelihood = 472.27
AIC=-932.53   AICc=-931.99   BIC=-913.97
```

Figure 8

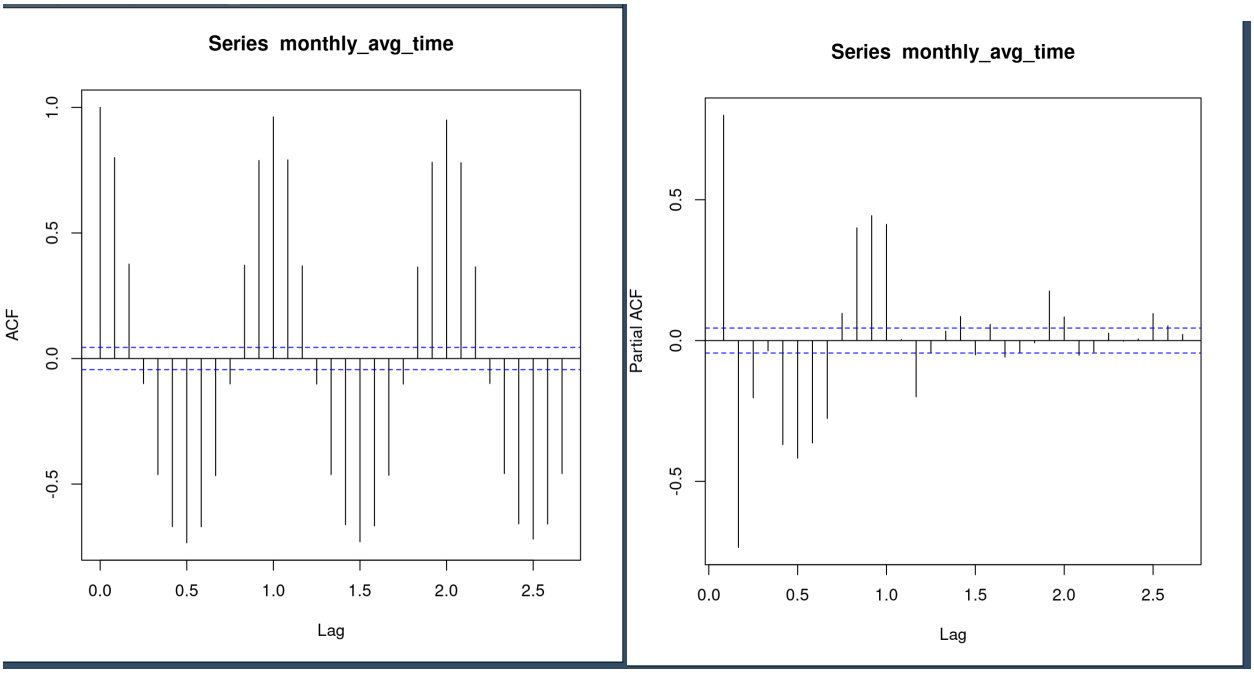

Figure 9: ACF and PACF monthly average time

value with h degrees of freedom and significance level α. Because the test is applied to residuals, the degrees of freedom must account for the estimated model parameters so that h=m-p-q, where p and q indicate the number of parameters from the model fit to the data. We are showing the process and results of validation below:

```
Box.test(model1$residuals, lag=5, type="Ljung-Box")
Box.test(model1$residuals, lag=10, type="Ljung-Box")
Box.test(model1$residuals, lag=15, type="Ljung-Box")
```

6.3 Forecasting

We have forecast for next ten years from the fitted model $2013 - 2023$.

```
forecast1 = forecast(model1, level=c(95),h = 10)
plot(forecast1) ## the trend continues as the avg temperature continues to
print(forecast1)
forecast_data = as.data.frame(forecast1)
```

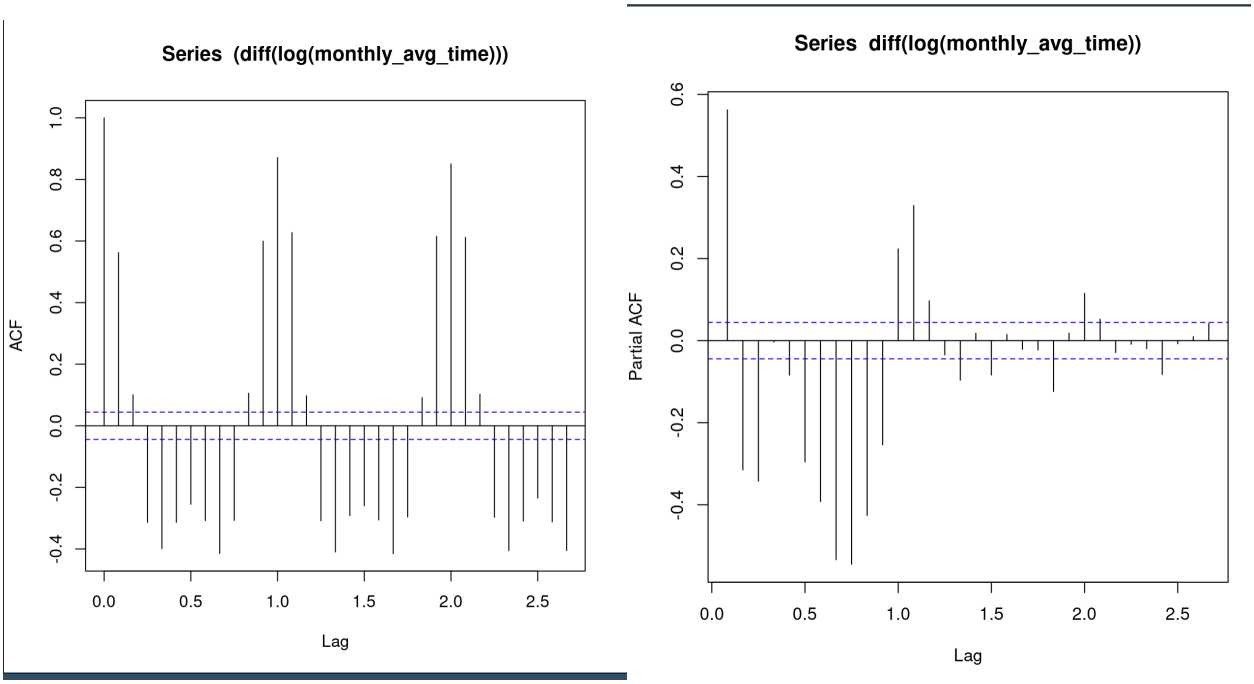

Figure 10: Difference ACF and PACF monthly average time

```
Series: monthly_avg_time
ARIMA(2,0,1)(2,1,0)[12]

Coefficients:
          ar1     ar2     ma1    sar1     sar2
      -0.3046  0.2546  0.5543  -0.543  -0.2812
s.e.   0.1071  0.0296  0.1089   0.022   0.0218

sigma^2 = 0.8413:  log likelihood = -2591.5
AIC=5194.99   AICc=5195.04   BIC=5228.43
>
```

Figure 11

We get the following value of the forecasts:

```
> forecast_data
     Point Forecast    Lo 95    Hi 95
2014       3.207142 3.180661 3.233622
2015       3.204606 3.175018 3.234193
2016       3.205098 3.175400 3.234796
2017       3.205041 3.175062 3.235021
2018       3.205000 3.174743 3.235257
2019       3.205011 3.174503 3.235519
2020       3.205011 3.174249 3.235772
2021       3.205010 3.173996 3.236023
2022       3.205010 3.173747 3.236273
2023       3.205010 3.173500 3.236520
```

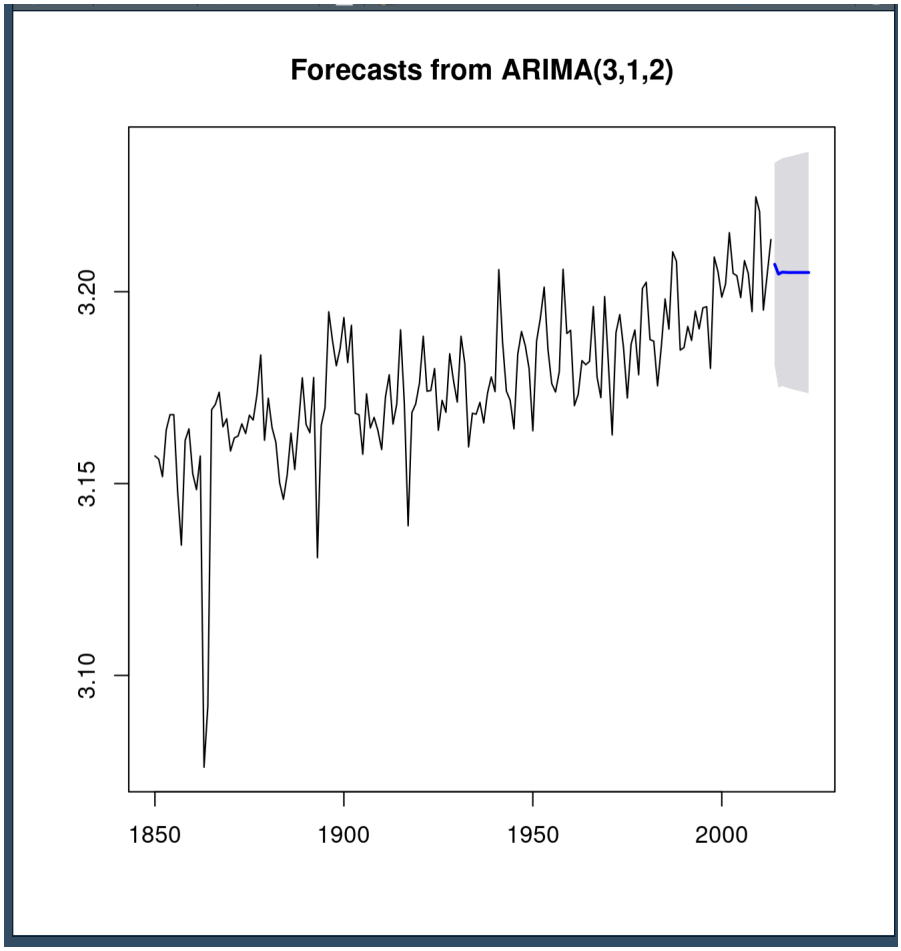

Figure 12: Forecast

7 Conclusion

We analysed the given data and were able to observe the trend and seasonality present in it. We observed that over the years, the temperatures are gradually rising. We also observed that every year the temperatures rose until around the month of May, then they started decreasing. We then fit a ARIMA(3,1,2) model to forecast future temperatures. We conclude that in the future years, the temperatures will gradually rise.

We can clearly see that global warming is indeed a real phenomenon. It is quite alarming to see the rise in temperatures over the years. It is crucial to take drastic measures in order to create a sustainable future for our future generations.

8 References

- https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data
- https://hello.iitk.ac.in/mth442asem12324//lectures