STA137 Applied Time Series Analysis - Final Project

Name: Aditya Mittal, Student ID: 919336522

Date: December 4th, 2022

View Dataset

```
'data.frame':   172 obs. of  2 variables:
 $ Year   : int  1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 ...
 $ Anomaly: num  -0.436 -0.219 -0.236 -0.245 -0.175 -0.246 -0.334 -0.451 -0.359 -0.248 ...
```

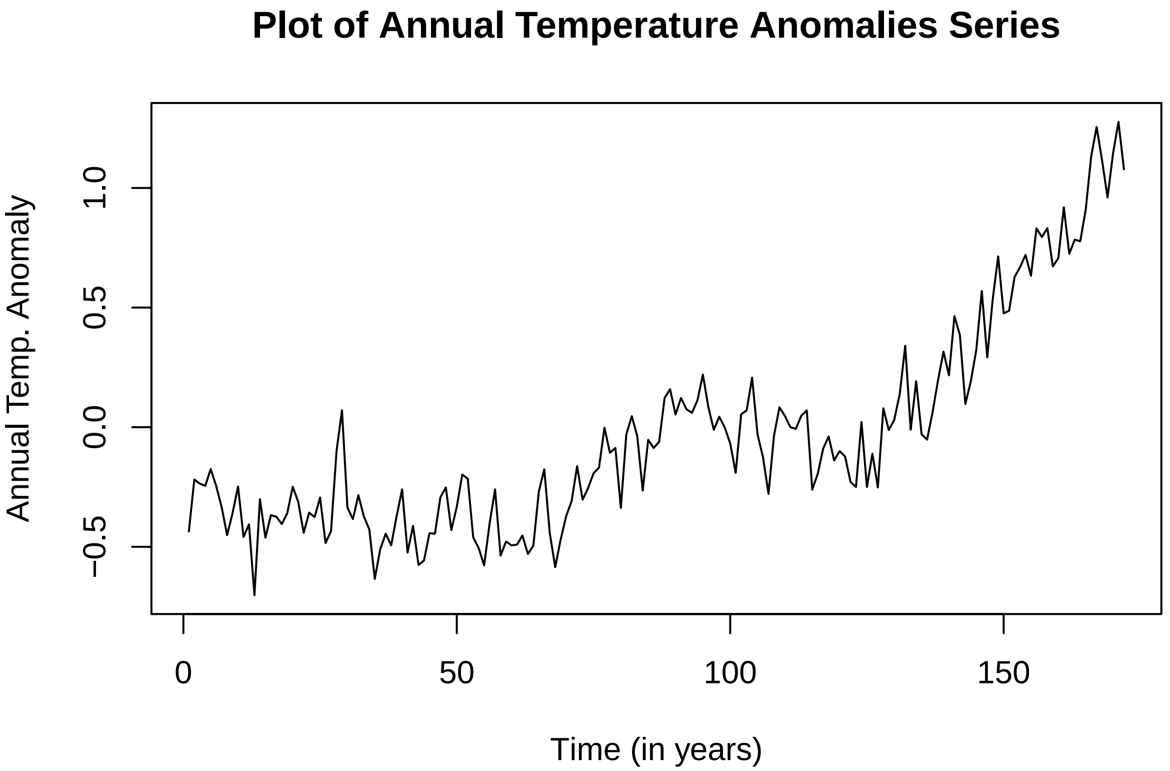
Introduction

Recent years have successively recorded record high annual temperatures. Consequently, further debates surrounding climate change and shifts in temperature trends have become increasingly prominent today. A temperature anomaly is defined as the departure of temperature data from a baseline - one that is typically set as the average of temperatures between a set interval of time. With a baseline annual temperature set between 1850-2021, we can study anomalies to identify trends and the deviations away from the expectation which may signal an overall shift in average annual temperature. The purpose of this project is to study data from the past to analyze potential trends, fit an ARMA/ARIMA time series model, and ultimately create forecasts for annual temperature anomalies on years 2016 to 2021 & compare them with true observed data to test our model's predictions.

The dataset TempNH_1850_2021 contains 172 observations of annual temperature anomalies for the northern hemisphere from year 1850 to 2021. This data can be considered as time series because the data is recorded as successive measurements from the same source over a set time interval and the intent is to track potential changes overtime. For practical purposes, it may be important to analyze this type of time series data with the intent to create future forecasts on upcoming years to help agencies & governments prepare resources accordingly.

Materials and Methodologies

Attached below is the plot for Yt's - the annual temperature anomalies from year 1850 (year 0) -2021 (year 172):



From the plot, there appears to be an increasing trend in the data as the anomalies were negative initially at 1850 and then were reaching approximately 1.0 at year 2021. Starting around year 110, the trend appears to become more positive compared its the prior years. This indicates that anomalies in annual temperatures in recent years have consistently been higher than the set baseline. In terms of variance, the data appears to be approximately equally spread across all the years. No seasonality seems to be present.

We can now begin conducting time series analysis on this data. In this project, we are going to be using two methods and build two separate models to create forecast and compare the results.

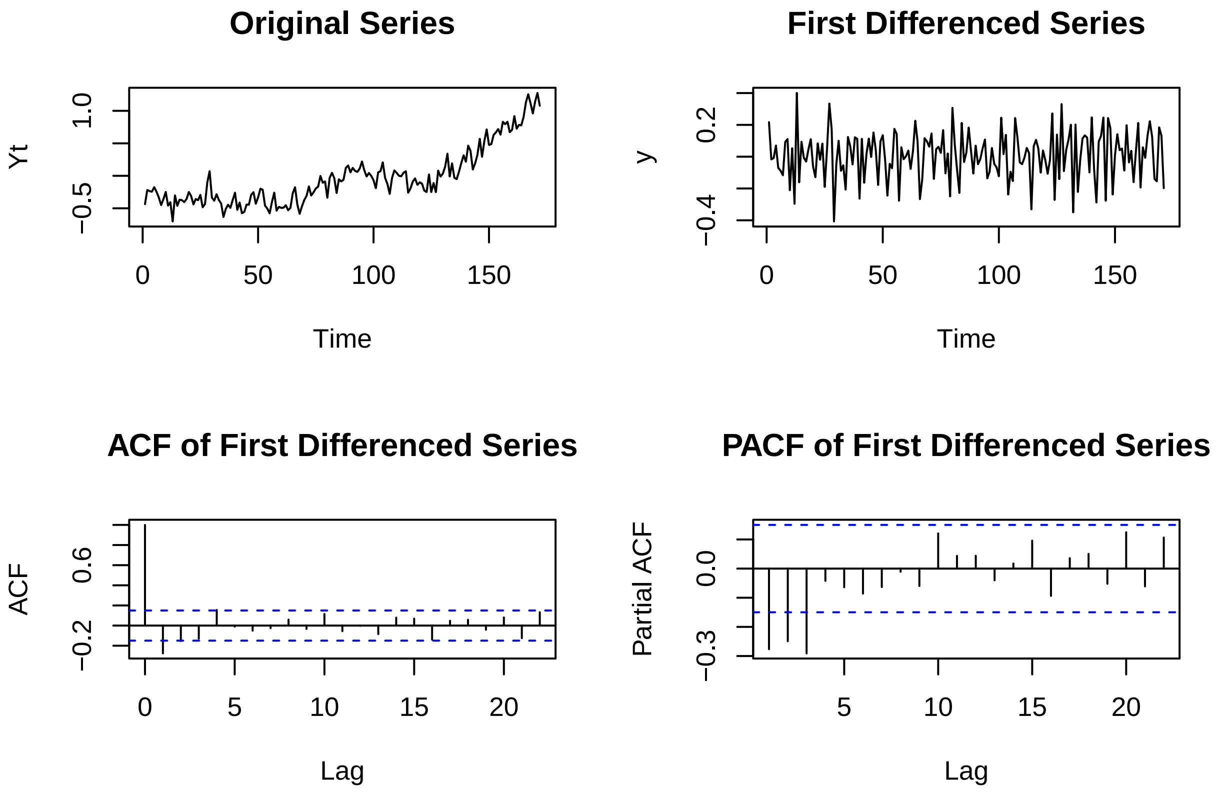
- **Method 1:** We will determine if the series is stationary or not and find its order difference to achieve stationarity. *Note:* The series is not stationary so will conduct all subsequent analysis and forecasts on the stationary differenced series. We will then build our model using ARIMA(p, 1, q) and test its residuals using ACF plot and Ljung-Box test. Once we've found a statistically sound model, we will refit the model on all data except the last 6 years. We will then create forecasts for the excluded data from year 2015 to 2021 and compare our predictions with the true observed values.
- **Method 2:** We will first estimate the trend using spline, plot the rough, and build and ARMA(p,q) model based on all data except the last 6 years. We will then guess the trend and forecast the rough for year 2015 to 2021 and add them together to create our new forecasts. The intent is to check whether both procedures differ in the forecasts significantly, if at all.

We will display the forecasts under "Results" section and discuss them further under "Conclusion."

Method One: Finding the first order difference of Yt

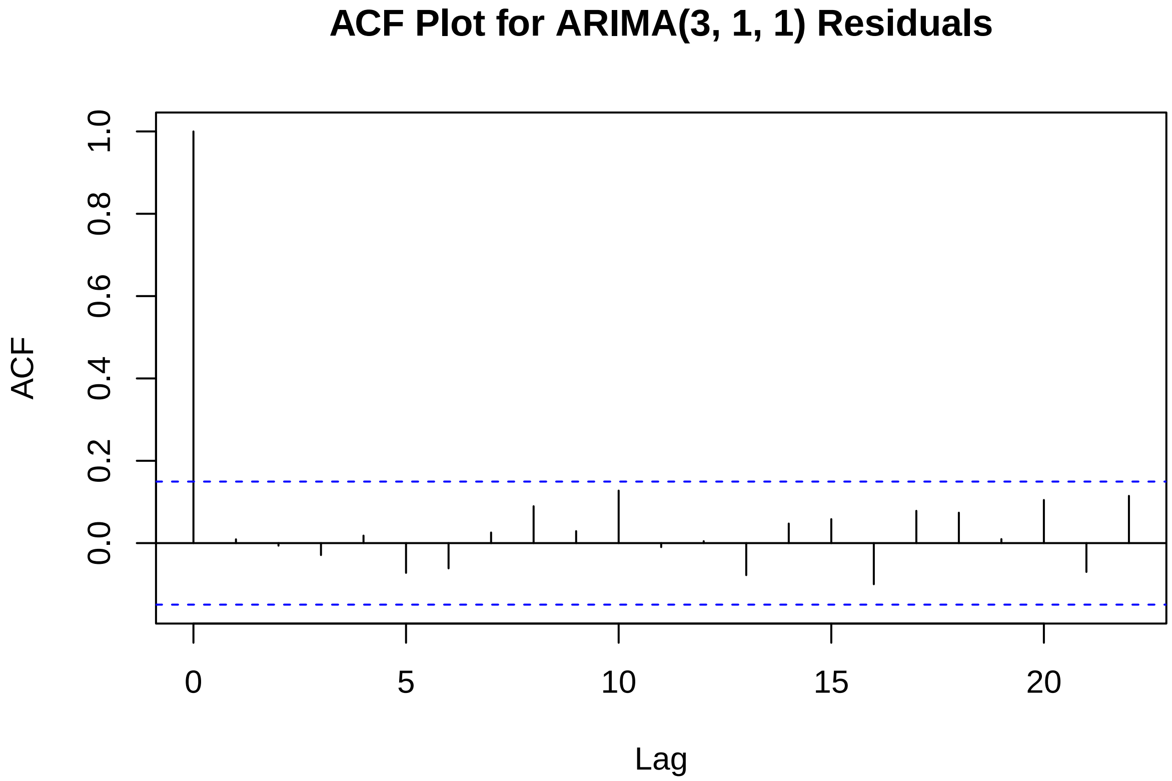
The annual temperature anomalies time series can be separated into two components: smooth (trend) and rough - in equation form: $Y_t = m_t + X_t$. Since the data has a trend, the time series cannot be considered to be stationary because the expected value does not equal a constant μ for each point at time t.

As the time series is not stationary, we can find the first order difference $X_t = Y_t$ - $Y_t - 1$. ACF and PACF plots for the differenced series are also attached below.



The first difference series appears to be stationary. It appears to have equal variance between different Xt's and has a constant mean $\mu = 0$. The ACF plot of the series shows there is significant correlation at lag 1 which cuts off after and can be considered negligible as its values falls within CI bounds +- $1.96/sqrt(n)$. The PACF plot shows some significant partial autocorrelation that cuts off after lag 3 and can be considered negligible after as the PACF values fall within the confidence interval bands. Using both ACF and PACF plots, through preliminary identification, we may choose to fit the data using ARIMA model arima(p = 3, d = 1, q = 1).

We should test if the residuals of our preliminary arima model are i.i.d to see if its a good fit for our data. Attached below is the ACF plot of residuals and the results from Box-Ljung test for i.i.d:


ACF Plot for ARIMA(3, 1, 1) Residuals

```
    Box-Ljung test

data:  mod$fit$residuals
X-squared = 6.591, df = 10, p-value = 0.7634
```

From the ACF plot of the residuals, it appears the model fit is stationary as all ACF values are within the bounds of the CI and can be considered negligible. The ACF values cut off after lag 0.

Hypothesis testing using Ljung-Box:

- H0 : p(1) = = p(h) = 0
- H1 : at least one of p(1),. . . ,p(h) is nonzero

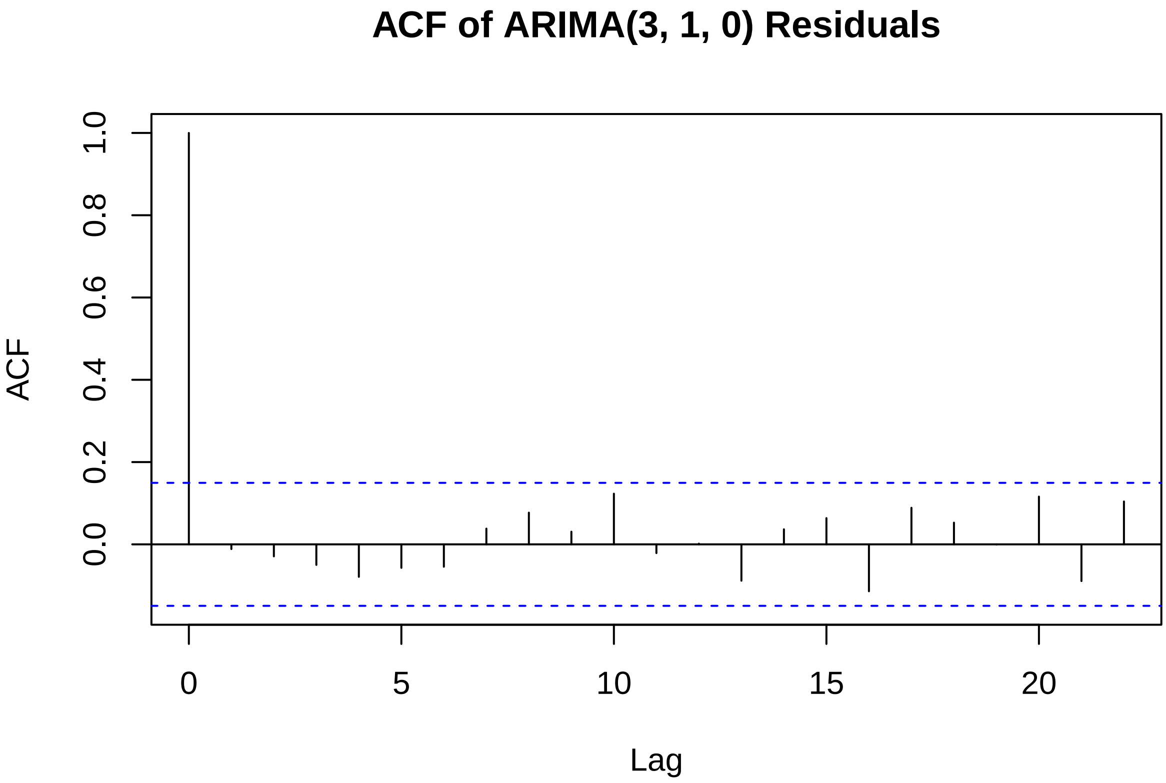
We'd fail to reject the null hypothesis of Ljung-Box test as p-value is 0.7634 is greater than all α levels (α = 0.01, 0.05, 0.01). We can state that the residuals of our model are i.i.d and that we may fit arima(3,1,1) for this data.

We now have an initial working model for our data. However, fitting an ARIMA(3,1,1) model means we are estimating (3 + 1 + 1) = 5 parameters, which is a lot. Based on the principle of parsimony, we'd like to select a model that's contains lesser parameters but still fits our data well. More specifically, we are going to fit 16 models: 0 <= p <= 3 and 0 <= q <= 3 where p is the AR order and q is the MA order to find smaller model. We are going to use the AIC criterion to select the model the that has the smallest AIC value.

	q = 0	q = 1	q = 2	q = 3
p = 0	-0.9229414	-1.094005	-1.119878	-1.108348
p = 1	-0.9922384	-1.114102	-1.108237	-1.109690
p = 2	-1.0455561	-1.114930	-1.112556	-1.114580
p = 3	-1.1233958	-1.116518	-1.110751	-1.103335

Based on this table of AIC values, we can fit an ARIMA(3, 1, 0) model as it has the smallest AIC value of -1.1233958 and it estimates one less parameter than our initial model. *Note:* The model we chose through graphical methods initially has the 3rd smallest AIC value.

Let's see if the residuals of ARIMA(3,1,0) model are i.i.d and using ACF plot and the Ljung-Box test:



```
    Box-Ljung test

data:  mod1$fit$residuals
X-squared = 7.1741, df = 10, p-value = 0.7089
```

From the ACF plot of the residuals, it appears the model fit is stationary as all ACF values are within the bounds of the CI and can be considered negligible.

Hypothesis testing using Ljung-Box:

- H0 : p(1) = = p(h) = 0

- H1 : at least one of p(1),. . . ,p(h) is nonzero

We'd fail to reject the null hypothesis of Ljung-Box test as p-value is 0.7089 is greater than all α levels ($\alpha =$ 0.01, 0.05, 0.01). We can state that the residuals of our model are i.i.d and we can fit arima(3,1,0) for this data.

Thus, we may select ARIMA(3,1,0) as our final model.

Here are the coefficients for our parameters of the ARIMA model:

	x
ar1	-0.4233537
ar2	-0.3557308
ar3	-0.2947135
constant	0.0085370

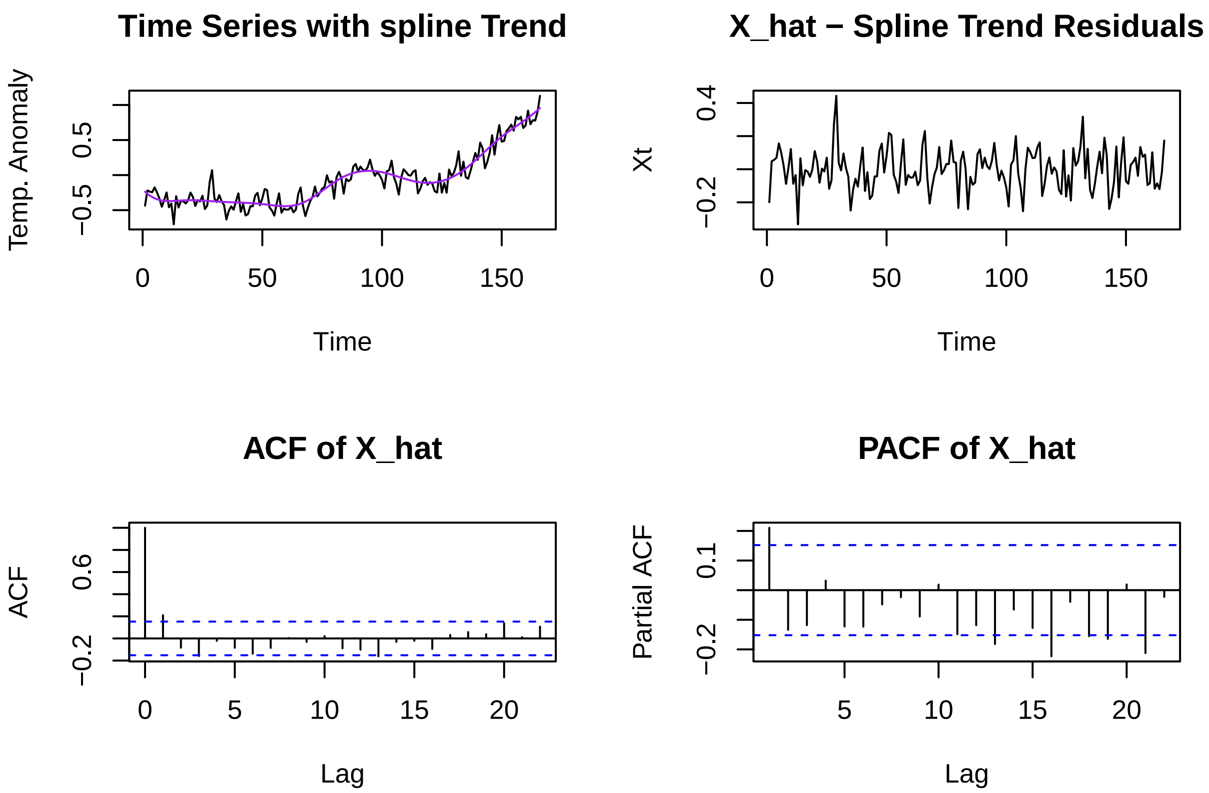
Here are the standard errors for our parameters of the ARIMA model:

	x
ar1	0.0735117
ar2	0.0755855
ar3	0.0735093
constant	0.0049636

We are now ready to create forecasts based on the first differenced series.

Method 2: Estimating Trend

In this method, we first estimate the trend using a spline function for all data *except the last 6 years*. We will also plot the estimate of the spline trend with the data and model the rough.



The spline trend appears to fit our model well as the points seem to fluctuate around the purple estimated trend line. The rough can be calculated by finding the residuals of the spline trend and it appears to be stationary with mean 0 and has an approximately constant variance across the all different values of time t. The ACF plot shows that the significance in autocorrelation cuts off after lag 1, suggesting we may fit an MA(1) model. The PACF plot is harder to interpret as there are several PACF values that pass the CI bands at lag 15 and around lag 20, suggesting an AR(p) model itself may not be suitable due to needing a large # of parameters. We will fit an ARMA(p,q) model:

Again, we are now going to fit 16 models: 0 <= p <= 3 and 0 <= q <= 3 where p is the AR order and q is the MA order to find our ARMA(p,q) model based on smallest AIC value.

	q = 0	q = 1	q = 2	q = 3
p = 0	-228.5292	-235.6150	-233.6283	-238.1587
p = 1	-234.2372	-233.6198	-233.7616	-254.8977
p = 2	-235.4474	-256.2489	-234.9504	-253.3076
p = 3	-235.8595	-234.6004	-254.0339	-251.3119

Based on the AIC table, we can fit an ARMA(2,1) model as it has the smallest AIC value at -256.2489. *Note:* We are not required to perform analysis of residuals to test for model fit.

Here are the coefficients for our parameters of the ARMA model:

	x
ar1	1.0896909

	x
ar2	-0.3410504
ma1	-0.9999996
intercept	-0.0000074

Here are the standard errors for our parameters of the ARMA model:

	x
ar1	0.0733904
ar2	0.0736744
ma1	0.0152180
intercept	0.0006703

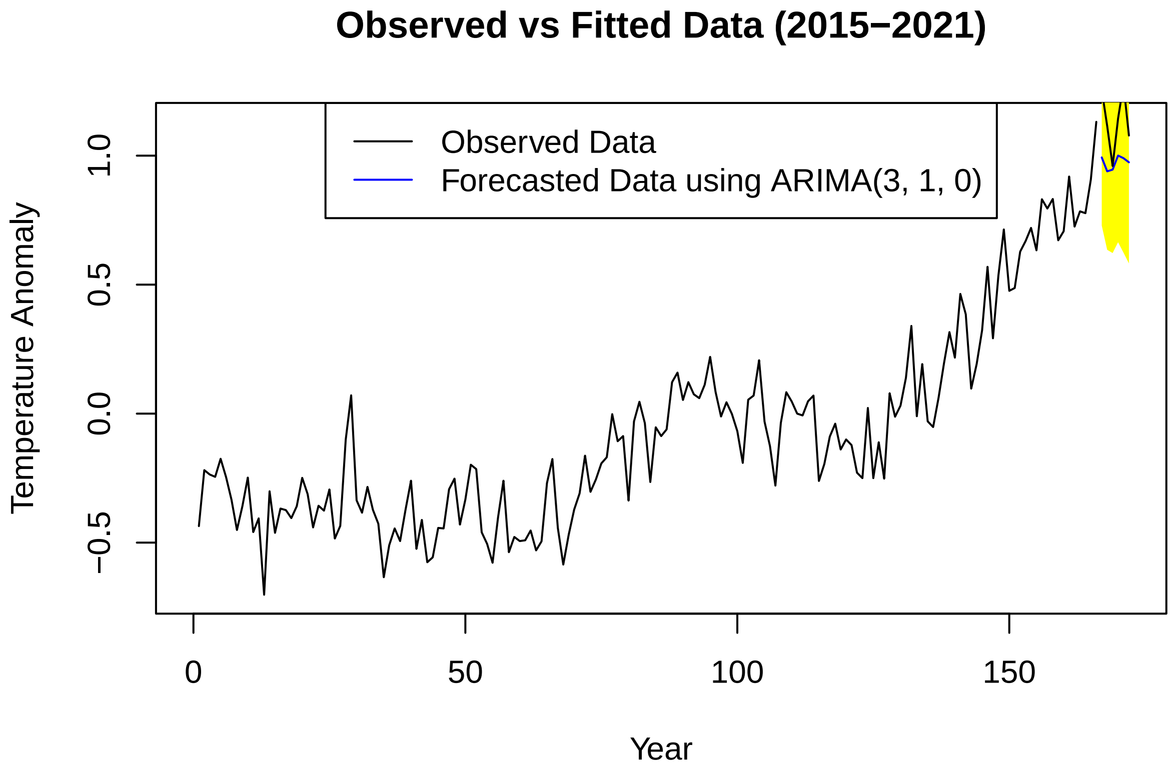
We are now ready to create forecasts for our data using this method.

Results

Method One: First order difference of Yt.

Using the first order difference of series, we have fit our final model as ARIMA(3, 1, 0). It is important to see how well the model is at forecasting future data - so we will refit this ARIMA(3, 1, 0) model on all observations except the last 6 years - 2016 to 2021 - to create forecasts of the following years and compare them with the true recorded observations. *Note:* Our model will fit the same AR and MA orders but will have different parameter estimates.

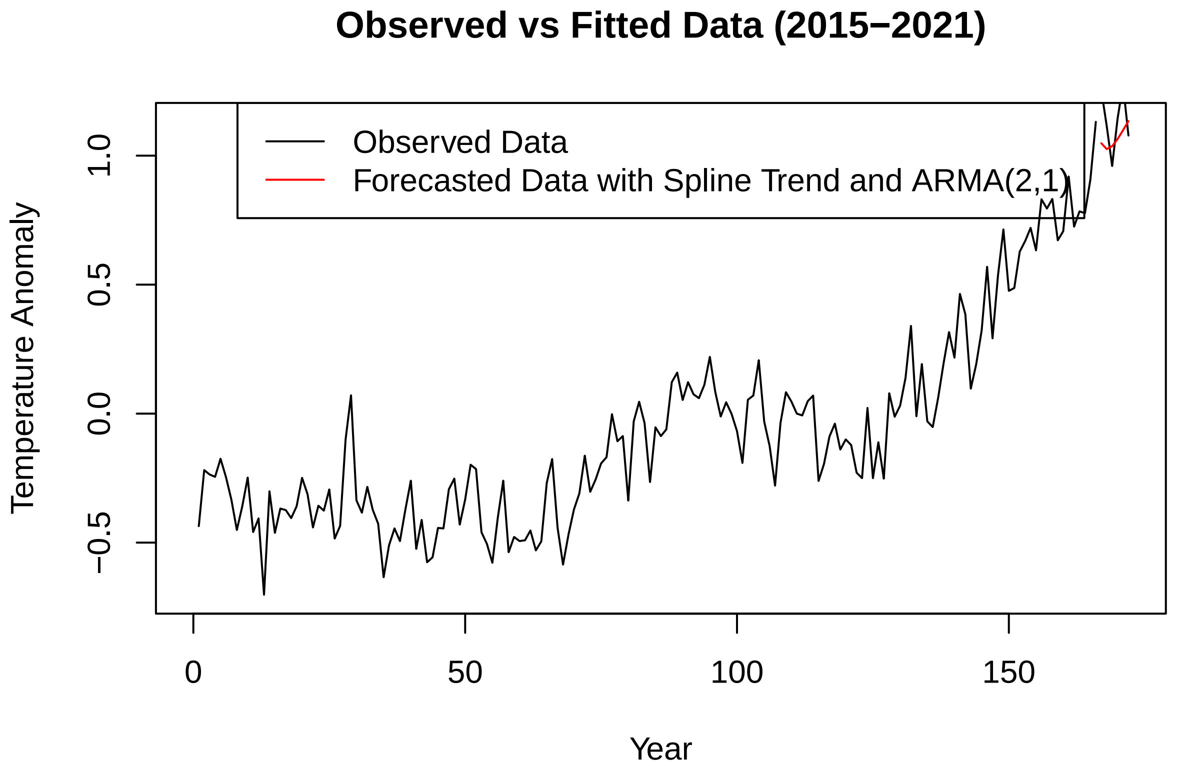
Attached below is the plot for observed vs. forecasted data against time (6 years).



The observed values (in black) all fall within the yellow prediction interval band from the ARIMA model. This indicates that our model is a good fit for this time series data. Our forecasted data (as shown by the blue line) follows the same pattern of increasing and decreasing fluctuations of temperature anomalies as the recorded data - although the forecasts appear to be a lot more conservative in their variances than the actual observed data.

Method 2: Guess spline trend & forecast the rough

Having estimated the trend using spline, we have fit our final model for the rough as ARMA(2, 1). We will using the function approxExtrap to guess the trend from 2016 to 2021, and forecast the data for rough using usual time series methods. The sum of the two results will provide the forecasts which may be compare with the recorded data. We will check these forecasts with actual observed data *and* with compare our results with method 1.



The forecasted data seems to be fairly accurate in predicting our data when compared to the actual recorded values from 2016-2021. The fitted data seems to lie close to the of the observed data. Again, the forecasts appear to be a lot more conservative in their variances than the actual observed data. In this method, the forecasts appear to be following a slightly different pattern of fluctuations than the true values - there is a dip in temp. anomaly for year 2017 and then a steady increase after for the forecasted data (meanwhile the observed data fluctuates both up and down). We can be more confident in the accuracy of our predictions if we were to add prediction interval bands to account for any deviance. Overall, this model fit is also good in predicting the data for our time series.

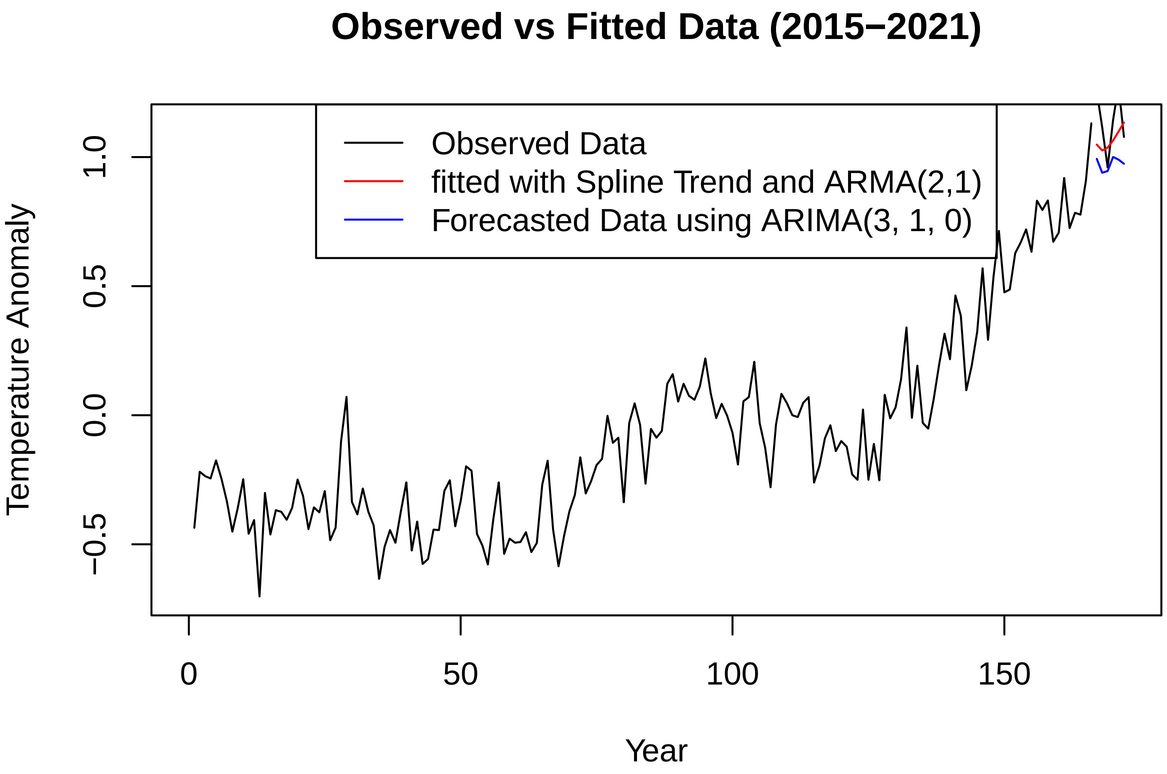
Conclusion

Using temperature anomaly data, our goal was to analyze potential trends & fit an ARMA/ARIMA time series model, ultimately creating forecasts from year 2016 to 2021 and compare them with true observed data to test our model's predictions.

We used two methods to build our model:

- Finding the first order difference $X_t = Y_t$ - $Y_t - 1$ to get a new stationary series and building an ARIMA model through lowest AIC criterion. Through this method, we fitted an ARIMA(3, 1, 0) model and conducted residual analysis using ACF plot and Ljung-Box test to make sure the model was good fit for our data.
- Estimating the trend through spline all data *except the last 6 years* and plotting the rough. Through this method, we fit an ARMA(2, 1) model and created our forecasts by summing our trend and rough predictions.

Once we had both models fitted, we forecasted our data for year 2016 to 2021. Let's compare our results from both methods on the same plot:



Both methods seem to be fairly accurate in forecasting the data as all of the forecasts lie on close to the true recorded data. It is important to note that the forecasts for ARIMA model seem to follow the fluctuations as the recorded data as opposed to forecasts from method 2, which had slightly different forecasts. Overall, both methods resulted in slightly different forecasts and appear to be suitable this time series data. This is to be expected as we used different methodologies to calculate our estimates and build our models. To create future forecasts, I would suggest using method 1 by finding the first order difference and fitting an ARIMA model. Having the ability to create prediction bands allows us to account for deviance and provides more meaningful interpretation.

Concluding Remarks: Potential Improvements

In this project, there may be several ways we can improve our methodologies and analyses:

- By employing a possible box-cox transformation on our initial data, we can reduce the overall changes in variance (have equal variance) across the trend on different values of time t before fitting the ARMA/ARIMA model.
- Model Selection: We can choose to confirm our findings or potentially find another, smaller model through cross validating with the BIC and AICc criterion as well. We may have been able to fit a smaller model that is still statistically sound and that requires us to estimate lesser parameters.
- We should create prediction interval bands when creating the forecasts for method 2 with trend estimation. Ultimately, the exact forecasted values will not be the same as the recorded values so it is important to account for the variance that may naturally occur. We would be more confident in our interpretations and comparisons of both methods if we had prediction bands available.

Code Appendix

```
knitr::opts_chunk$set(message = FALSE, echo = FALSE, warning = FALSE, comment=NA, fig.align="center")

# read & view dataset
TempNH_1850_2021 <- read.csv("~/Desktop/Fall Quarter/STA 137/final proj/TempNH_1850_2021.csv")
str(TempNH_1850_2021)
Yt <- TempNH_1850_2021$Anomaly
# plot the time series
plot.ts(Yt, main = "Plot of Annual Temperature Anomalies Series", xlab = "Time (in years)", ylab = "Annu
par(mfrow = c(2,2))
plot.ts(Yt, main = "Original Series") # plot original series
Yt<-unlist(Yt)
y<-diff(Yt,1)
plot.ts(y, main = "First Differenced Series") # plot for first difference series
acf(y, main = "ACF of First Differenced Series") # acf plot for first difference series
pacf(y, main = "PACF of First Differenced Series") # pacf plot for first difference series
library(astsa)
mod = sarima(Yt,p=3,d=1,q=1,details=FALSE) # fit arima model arima(3,1,1)
acf(mod$fit$residuals, main = "ACF Plot for ARIMA(3, 1, 1) Residuals")
Box.test(mod$fit$residuals, lag=10, type='Ljung-Box')
AIC<-matrix(0,4,4) # AIC values matrix for 16 models of 0 <= p <= 3 and 0 <= q <= 3
for (i in 1:4){
  for (j in 1:4){
    AIC[i,j]<-sarima(Yt,p=i-1,d=1,q=j-1,details=FALSE)$AIC
  }
}
colnames(AIC) = c("q = 0", "q = 1", "q = 2", "q = 3")
rownames(AIC) = c("p = 0", "p = 1", "p = 2", "p = 3")
AIC <- as.table(AIC)
knitr::kable(AIC)
mod1 = sarima(Yt,p=3,d=1,q=0,details=FALSE)
acf(mod1$fit$residuals, main = "ACF of ARIMA(3, 1, 0) Residuals")
Box.test(mod1$fit$residuals, lag=10, type='Ljung-Box')
knitr::kable(mod1$fit$coef)
knitr::kable(sqrt(diag(mod1$fit$var.coef)))
#split data
n <- length(Yt)
xnew <- Yt[1:(n-6)]
xlast <- Yt[(n-5):n]

# function for spline (TA FUNCTION)
trend_spline=function(y, lam){
  n=length(y)
  p=length(lam)
  rsq=rep(0, p)
  y=sapply(y,as.numeric)
  tm=seq(1/n, 1, by=1/n)
  xx=cbind(tm, tm^2, tm^3)
  knot=seq(.1, .9, by=.1)
  m=length(knot)
  for (j in 1:m) {
    u=pmax(tm-knot[j], 0); u=u^3
    xx=cbind(xx,u)
  }
  for (i in 1:p){
    if (lam[i]==0){
      ytran=log(y)
    } else {
      ytran=(y^lam[i]-1)/lam[i]
    }
    ft=lm(ytran~xx)
    res=ft$resid; sse=sum(res^2)
    ssto=(n-1)*var(ytran)
    rsq[i]=1-sse/ssto
  }
  ii=which.max(rsq); lamopt=lam[ii]
  if (lamopt==0) {
    ytran=log(y)
  } else {
    ytran=y^lamopt
  }
  ft=lm(ytran~xx);
  best_ft=step(ft, trace=0)
  fit=best_ft$fitted; res=best_ft$resid
  result=list(ytrans=ytran, fitted=fit, residual=res, rsq=rsq, lamopt=lamopt)
  return(result)
}

tm = 1:166
splinetrnd=trend_spline(xnew, 1)
roughspline = splinetrnd$residual
par(mfrow = c(2,2))
plot(tm, xnew, type="l", lty=1, xlab="Time", ylab="Temp. Anomaly", main="Time Series with spline Trend"
points(tm, splinetrnd$fitted, type="l", lty=1, col = "purple")
plot.ts(roughspline, main = "X_hat - Spline Trend Residuals", ylab = "Xt")
acf(roughspline, main = "ACF of X_hat")
pacf(roughspline, main = "PACF of X_hat")
par(mfrow = c(1,1))
AICmat<-matrix(0,4,4)
for (i in 0:3){
  for (j in 0:3){
    AICmat[i+1,j+1]<-arima(roughspline,order = c(i, 0, j))$aic
  }
}
colnames(AICmat) = c("q = 0", "q = 1", "q = 2", "q = 3")
rownames(AICmat) = c("p = 0", "p = 1", "p = 2", "p = 3")
AICmat = as.table(AICmat)
knitr::kable(AICmat)
mod_arma21 <- arima(roughspline,order=c(2,0,1))
knitr::kable(mod_arma21$coef)
knitr::kable(sqrt(diag(mod_arma21$var.coef)))
n <- length(Yt)
xnew <- Yt[1:(n-6)]
xlast <- Yt[(n-5):n]
model1 <- arima(xnew,order = c(3,1,0))
```

```
h <- 6
m <- n - h
fcast <- predict(model1, n.ahead=h)
upper <- fcast$pred+1.96*fcast$se
lower <- fcast$pred-1.96*fcast$se
plot.ts(xnew, xlim = c(0,n), xlab = "Year", ylab = "Temperature Anomaly", main = "Observed vs Fitted Dat
polygon(x=c(m+1:h,m+h:1), y=c(upper,rev(lower)), col='yellow', border=NA)
lines(x=m+(1:h), y=fcast$pred,col='blue')
lines(x=m+(1:h), y=xlast,col='black')
legend("top", legend = c("Observed Data","Forecasted Data using ARIMA(3, 1, 0)"), lty=c(1, 1), col = c(
library(png)
library(jpeg)
library(Hmisc)
trend = approxExtrap(tm, splinetrnd$fitted, xout = c(167:172))
model1 <- arima(roughspline,order = c(2,0,1))
h <- 6
m <- n - h
fcast2 <- predict(model1, n.ahead=h)
Forecasted_values = trend$y + fcast2$pred
plot.ts(xnew, xlim = c(0,n), xlab = "Year", ylab = "Temperature Anomaly", main = "Observed vs Fitted Dat
lines(x=m+(1:h), y=xlast,col='black')
lines(x=m+(1:h), y=Forecasted_values,col='red')
legend("top", legend = c("Observed Data","Forecasted Data with Spline Trend and ARMA(2,1)"), lty=c(1, 1
plot.ts(xnew, xlim = c(0,n), xlab = "Year", ylab = "Temperature Anomaly", main = "Observed vs Fitted Dat
lines(x=m+(1:h), y=xlast,col='black')
lines(x=m+(1:h), y=Forecasted_values,col='red')
lines(x=m+(1:h), y=fcast$pred,col='blue')
legend("top", legend = c("Observed Data","fitted with Spline Trend and ARMA(2,1)", "Forecasted Data usin
```