1. Forecasting COVID-19 Cases

The coronavirus pandemic has devastated social, economic, and health systems across the world. In an effort to save lives and ease the burden on hospitals, many businesses have been ordered to close, people have quarantined in their homes, and unnecessary travel has been restricted. It has been important for leaders and decision makers to have an accurate count of cases and deaths as well as forecasts of the number of cases expected in coming days and weeks to accurately plan. We used daily data of confirmed coronavirus cases and deaths in the United States to forecast the number of cases expected in the United States for ten days ahead.

2. Data Description

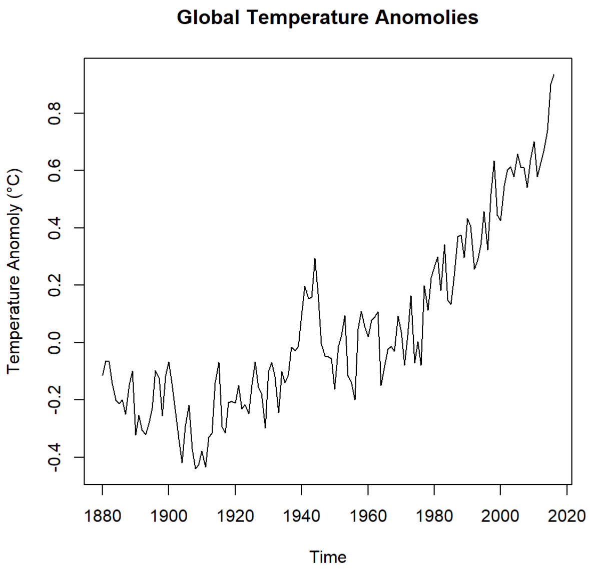
The data used for our analysis is the New York Times data set of the daily total count of confirmed COVID-19 cases and deaths in the United States (https://github.com/nytimes/covid-19-data/blob/master/us.csv).

The start date of this daily data is January 21, 2020, when the first confirmed case was recorded in the US. While the data continues to be updated daily, the end date of the data for our analysis is May 25, 2020.

We also calculated the difference between each day to find the change in cases and deaths each day, which we will refer to as *daily cases*.

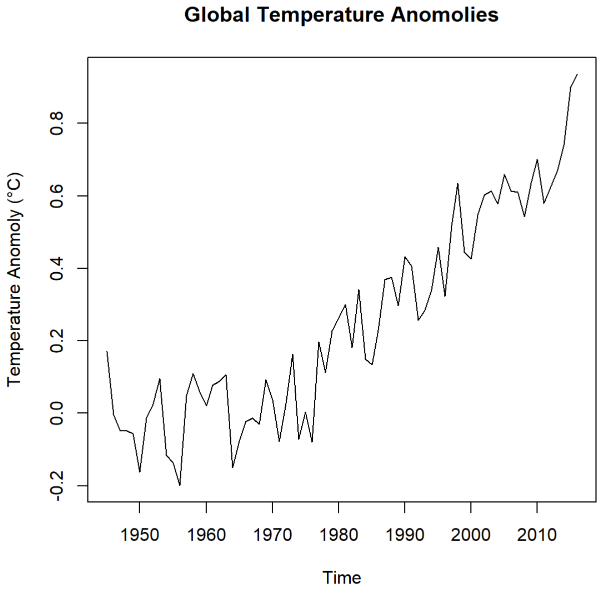
Plot of Data

The full dataset contains data from 1880 to 2020



Narrow Down Data

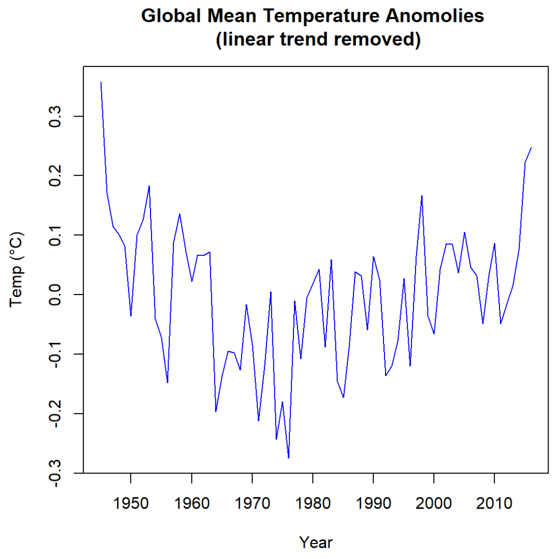
It is clear from the above plot that a positve trend begins to show about midway through the dataset. We decided to focus on the period of time since World War 2 (Starting in 1945). This period was chosen because of the clear trend shown as well as the cultural and environmental significance of the post-war era.

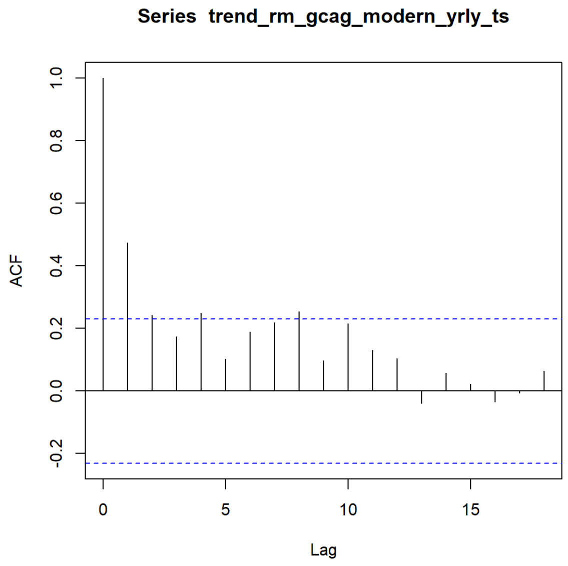


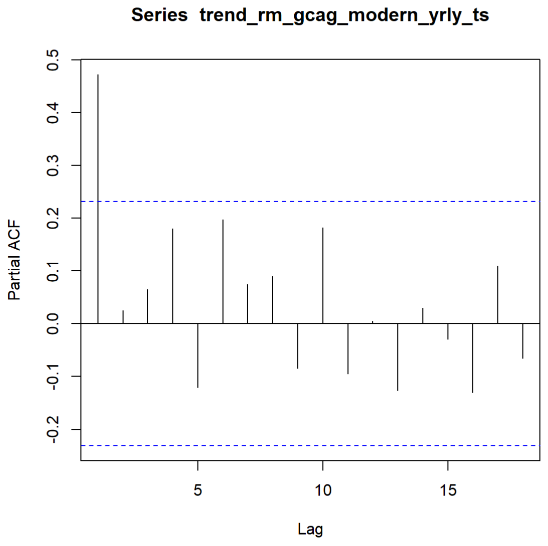
Remove linear trend to get stationary time series

In order to perform ARIMA forecasting, a stationary dataset is needed. The trend shown in the data makes it clear the data is non-stationary and therefore the trend is removed before proceeding.

```
##
## Call:
## lm(formula = Mean ~ Year, data = gcag_modern_yrly_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27469 -0.08838  0.02063  0.07453  0.35821
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.420e+01  1.341e+00  -18.05   <2e-16 ***
## Year         1.235e-02  6.768e-04   18.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1194 on 70 degrees of freedom
## Multiple R-squared:  0.8262, Adjusted R-squared:  0.8237
## F-statistic: 332.7 on 1 and 70 DF,  p-value: < 2.2e-16
```



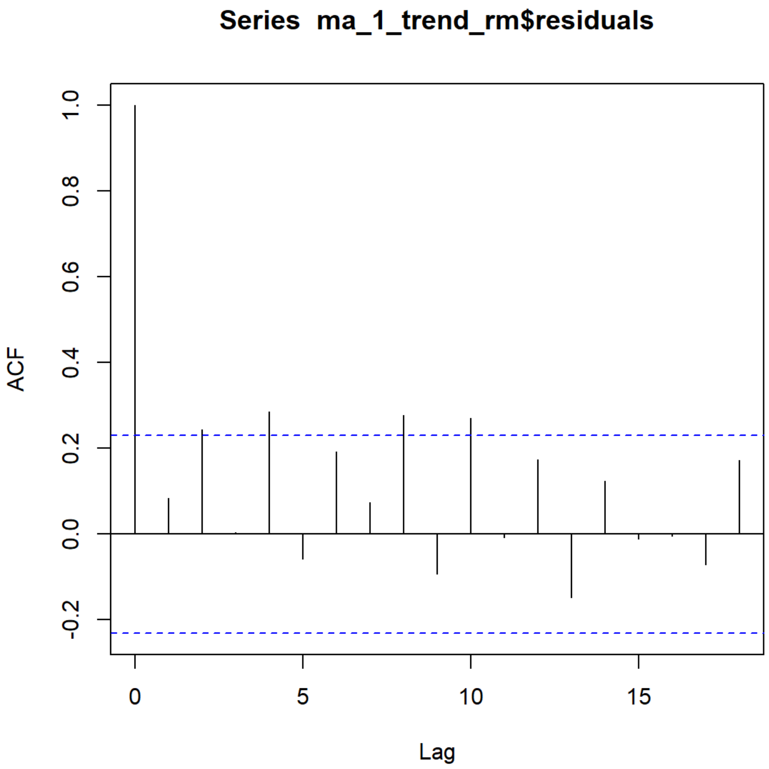




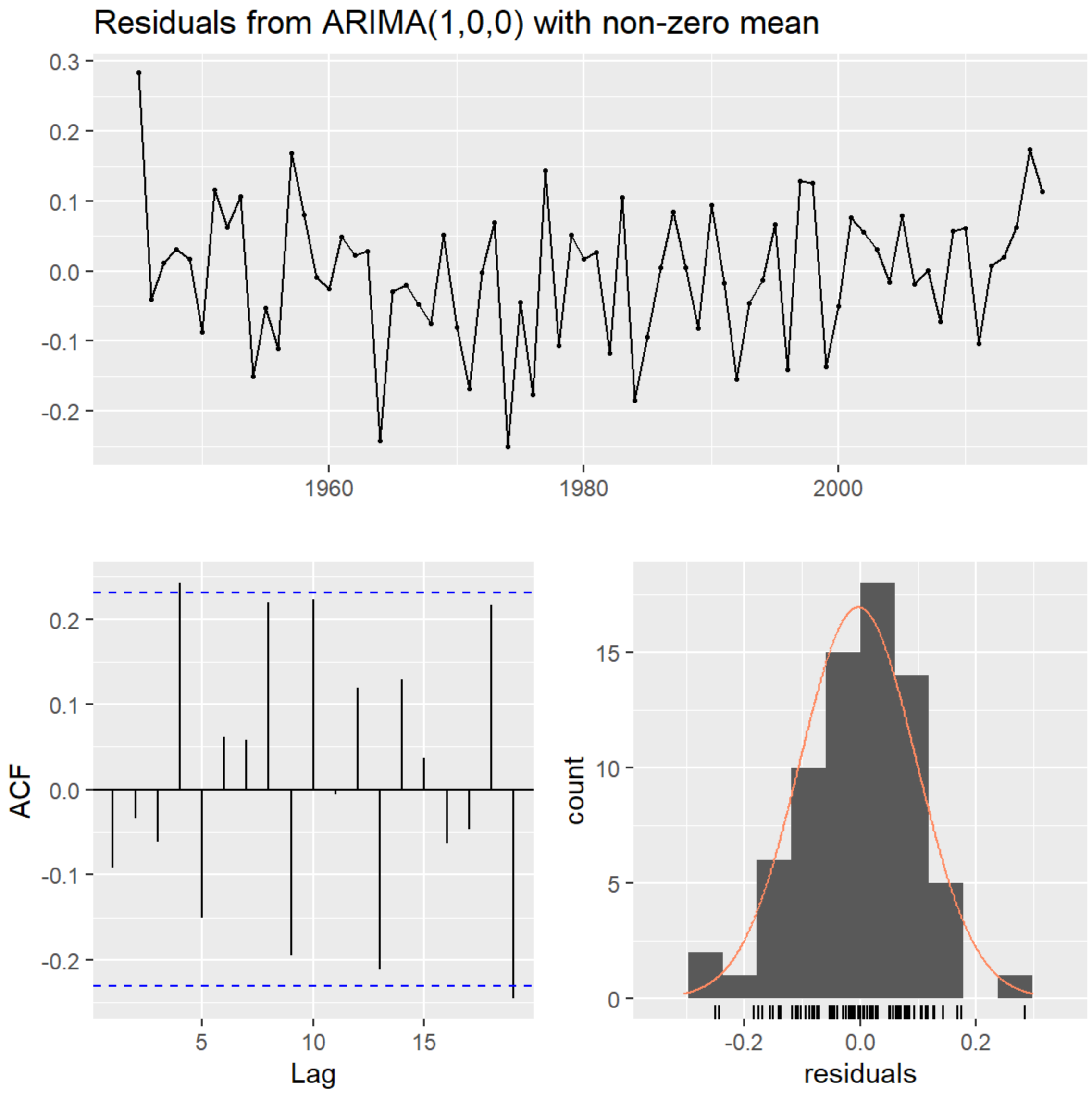
At this point our linear trend appears to be removed resulting in a stationary dataset. Examining the ACF and PACF of the adjusted time series, one sees a significant spike at lag 1 for both correlation functions. This implies an order 1 process explains the dataset.

Test ARMA Models

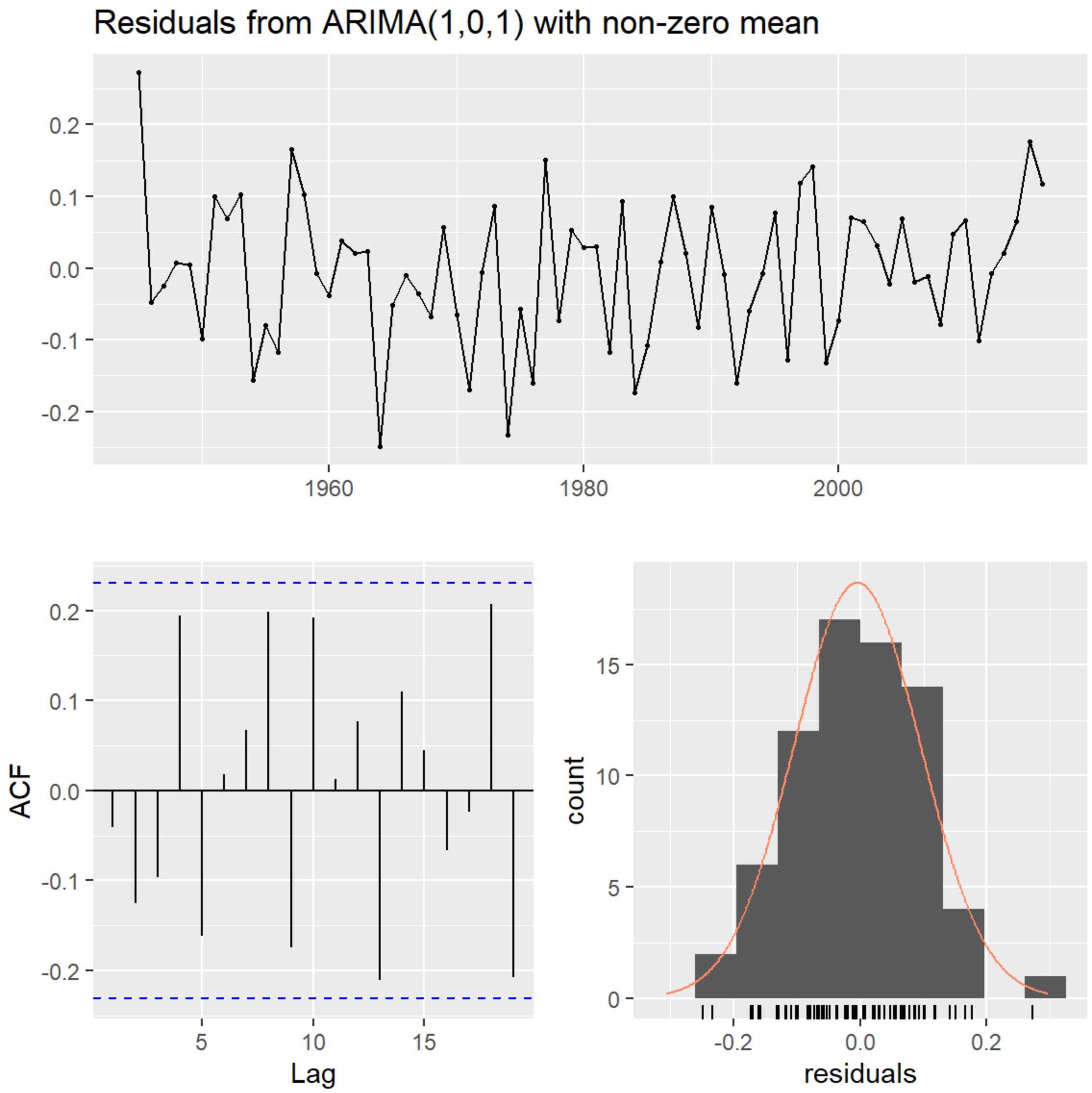
For modeling the time series, all three types of order 1 ARMA models were considered. After plotting the residuals of the models, it was clear the MA(1) model did not result in a white-noise error term. However the other two models, AR(1) and ARMA(1,1), both showed residuals which implied a white-noise error term. These two models were chosen for further validation.



- Significant residuals for MA(1) model



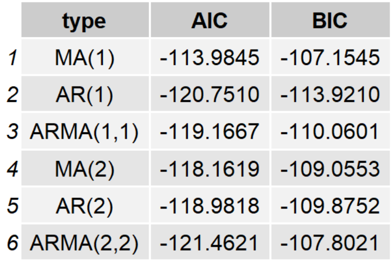
```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0) with non-zero mean
## Q* = 19.508, df = 8, p-value = 0.01237
##
## Model df: 2.   Total lags used: 10
```


Residuals from ARIMA(1,0,1) with non-zero mean

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1) with non-zero mean
## Q* = 16.507, df = 7, p-value = 0.02087
##
## Model df: 3.   Total lags used: 10
```

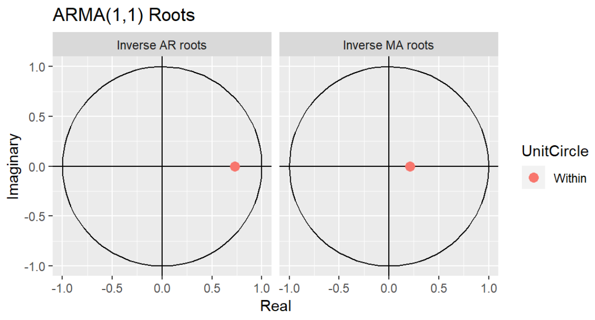
- The ARMA(1,1) residuals are least significant of the three models which implies the best fit.

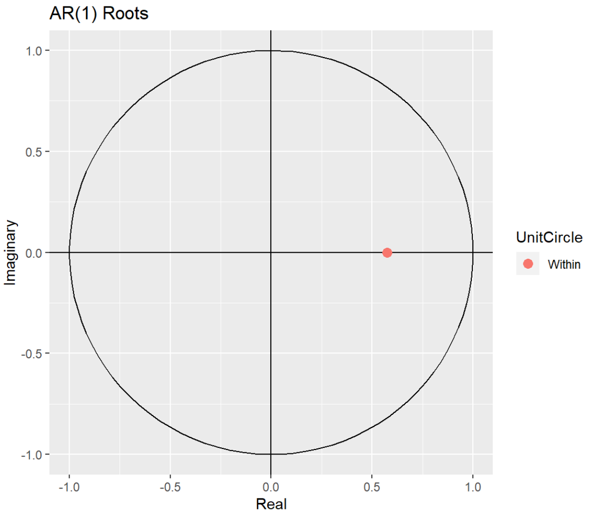
The validation process continued by testing higher order fits and also testing Information Criteria (AIC/BIC). As seen in the Information Criteria Table below, the best (lowest) AIC came from the ARMA(2,2) model, followed closely by AR(1) and ARMA(1,1) respectively. When looking at BIC, the best (highest) was again ARMA(2,2), followed by ARMA(1,1) and AR(1) respectively. Higher order fits are expected to have better IC values but risk over-fitting the data. Based on this fact, as well as the correlation functions not implying a 2nd order process, the ARMA(2,2) model was rejected. The other two best models were both close in fit quality (residuals) and information criteria and were chosen for further validation.



	type	AIC	BIC
1	MA(1)	-113.9845	-107.1545
2	AR(1)	-120.7510	-113.9210
3	ARMA(1,1)	-119.1667	-110.0601
4	MA(2)	-118.1619	-109.0553
5	AR(2)	-118.9818	-109.8752
6	ARMA(2,2)	-121.4621	-107.8021

Check for invertability of models





Both of the chosen models are invertable and we can continue testing both models.

Optimal Forecast

The estimation sample chosen for the two forecasts were years 1945 - 2016. This left 4 years of out-of-sample data to use as the forecasting sample. The forecasts were performed and errors were compared. ARMA(1,1) has lower MAE (0.096) than AR(1) (0.125), but need to be checked if errors are statistically different.

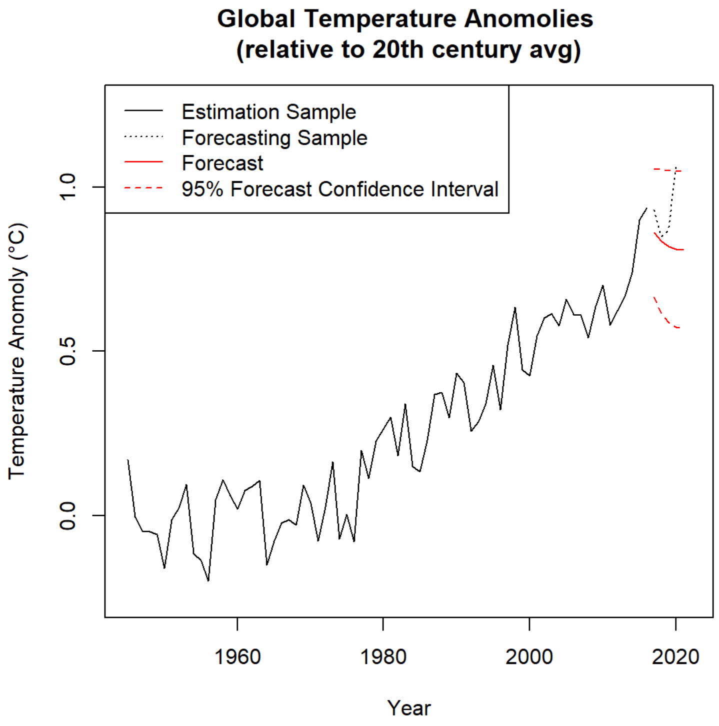
```
##

##  One Sample t-test

##

## data:  (abs(gcag_modern_yrly_with_ar_1_forecast_df$Mean[73:76] - gcag_out_of_sampl
e$Mean)) -     (abs(gcag_modern_yrly_with_forecast_df$Mean[73:76] - gcag_out_of_sampl
e$Mean))

## t = 5.1709, df = 3, p-value = 0.01403

## alternative hypothesis: true mean is not equal to 0

## 95 percent confidence interval:

##  0.01103578 0.04636123

## sample estimates:

## mean of x

## 0.0286985
```

The T test shows the model MAEs are statistically significantly different so ARMA(1,1) is our optimal model.

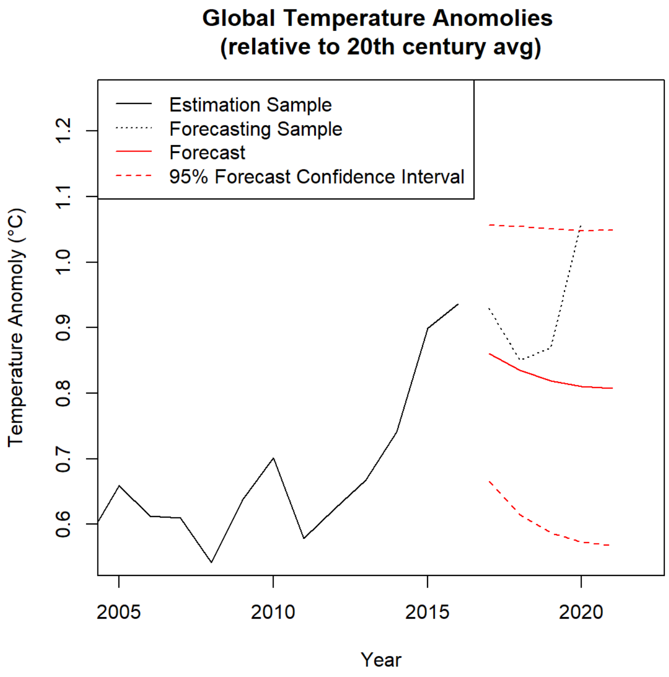
Loss Function

The forecast loss function for global temperature change was chosen to be an absolute symmetric loss function. Environmental agencies rely on forecasting to better prepare for the impact of global warming. Underpredicting temperature change suggests cooler temperature. If the observed temperature is warmer, our society and ecosystem could face larger consequences than anticipated. Overpredicting temperature change indicates rising global temperature, which also have a large penalty. There could be concerns for rising sea level, poor crop productions, and extinction of wildlife resulting in panic and over-commiting resources to fighting the warming climate. The two outcomes were decided to be equally severe and thus a symmetric loss function was chosen.

Plot forecasts



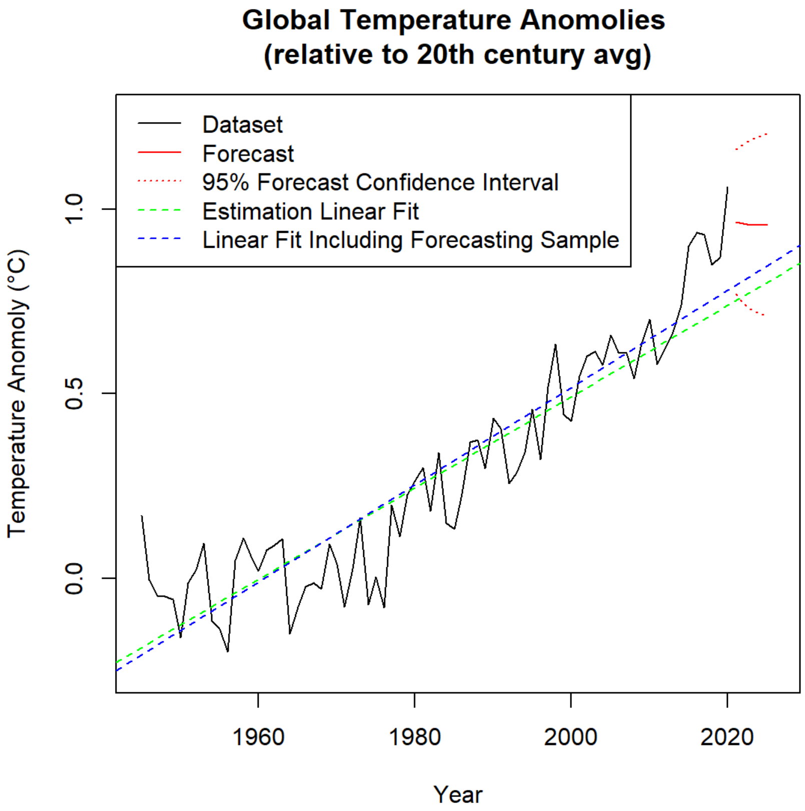
Zoomed plot



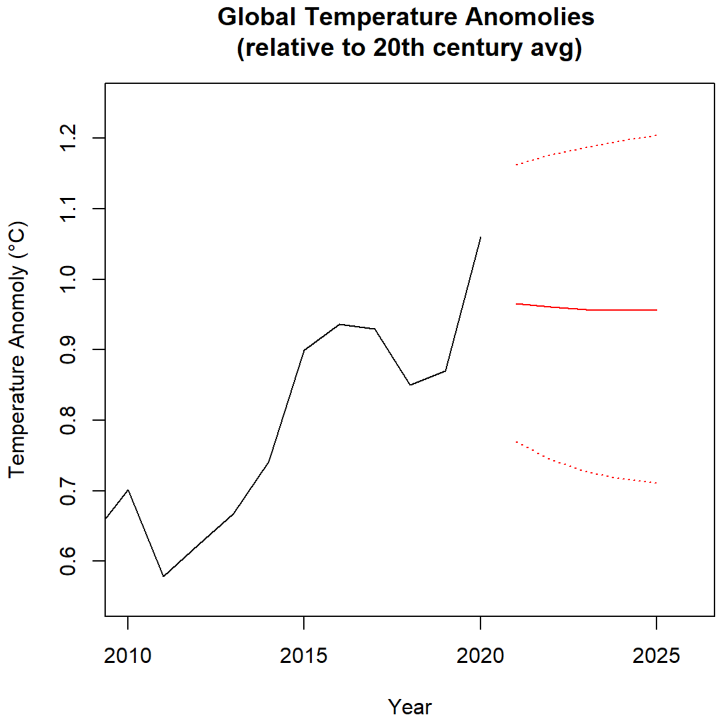
As seen in the above plot of the forecast, the first few steps of the predicted values are fairly close to the realized value of the forecasting sample. The 4th and 5th predictions however do not predict the large positive shock observed in the 2019 and 2020 realizations. It is expected that larger time horizons will have weaker predictions so this is still a good forecast. The entire forecasting sample lies within the positve 95% boundary which implies our model is consitently underpredicting the shocks.

Predicting the next few years

As seen above, the ARMA(1,1) model provided the optimal forecast with MAE of roughly 0.1°C. For the final analysis, the forecasting sample used previously is incorporated into the model in order to provide a forecast for the near future.



Zoom in on forecasting region



Results and Further Analysis

The results of the forecast for years 2021 through 2025 are shown above. This forecast predicts a small negative shock for the near future, with the anomoly falling slightly below 1.0°C. The confidence interval shows the realized temperature anomoly will remain roughly between 0.8 and 1.2°C with 95% confidence for the next 5 years. We saw our forecast errors were all negative when analyzing the performance of the model on the forecasting sample, so it is likely the future realizations will also skew to the upper part of the interval.

It appears the best-fit linear trend for the entire estimation sample is underpredicting the actual current trend. This is likely one of the main causes of the underprediction seen in the forecasts. A future analysis could use a more complex trend fit or use a differenced time-series in order to remove the trend. These methods could result in a better forecast by more accurately fitting the modern increasing trend of the data.

In conclusion, global temperatures are extremely likely to remain above 1.0°C above the 20th century average for the next 5 years. It is also likely the trend of warming has increased over the past 75years, meaning the rate of warming is increasing.