Final Project

Xiaoyuan Mao

2021-04-16

1. Abstract

This project studies how global mean ocean temperature changes from 1880 to 2017. We build a model consisting of a zero-mean lagged relationship and a nonstationary trend. We find that the temperature is increasing around a linear trend and temperature in the current year is only correlated to the temperature in the previous two years. There, we can forecast the future ten years global mean temperature. We also find that there is no significant cycle as the predominant frequency estimator subject to a lot of uncertainties.

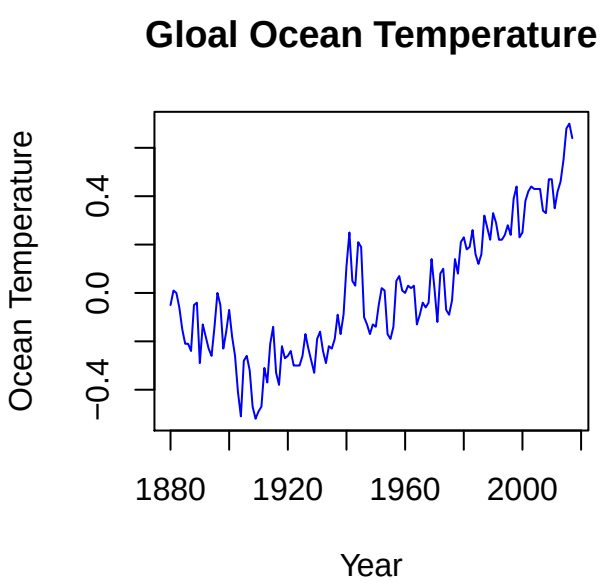
2. Introduction

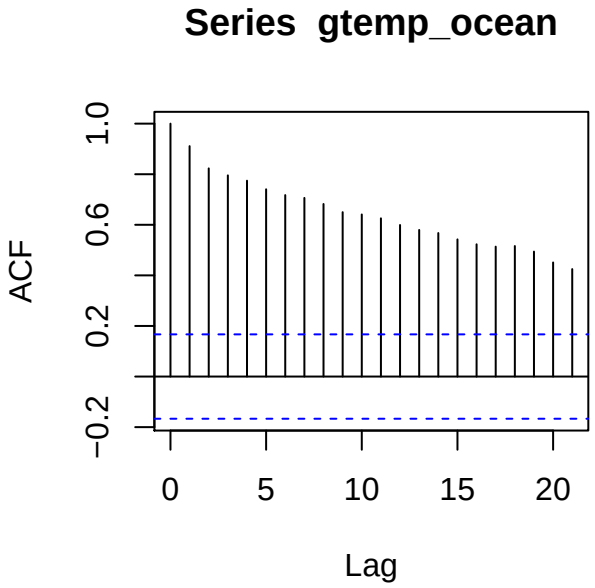
Changes in ocean surface temperature are affecting every aspect of our lives. It's directly related to the ecosystem. Because the ocean surface is in constant interaction with the air, the temperature changes can affect the amount of water vapour and then influence the global climate. Besides, the rise of temperature also increases the lifespan of certain bacteria, which can cause potential health risks. All of them will eventually cause a profound impact on the global economy. Therefore, predicting future temperature changes to control risks is critical.

In this project, we use the annual global surface ocean temperature averaged over the ice-free part of the ocean around the world from 1880 to 2017 to build a model and make a prediction

of the future ten years. We mainly focus on the time-domain approach, where we studied lagged relationships and build a model. We also briefly perform spectral analysis on the data, where we investigate if the predominant frequencies are statistically important.

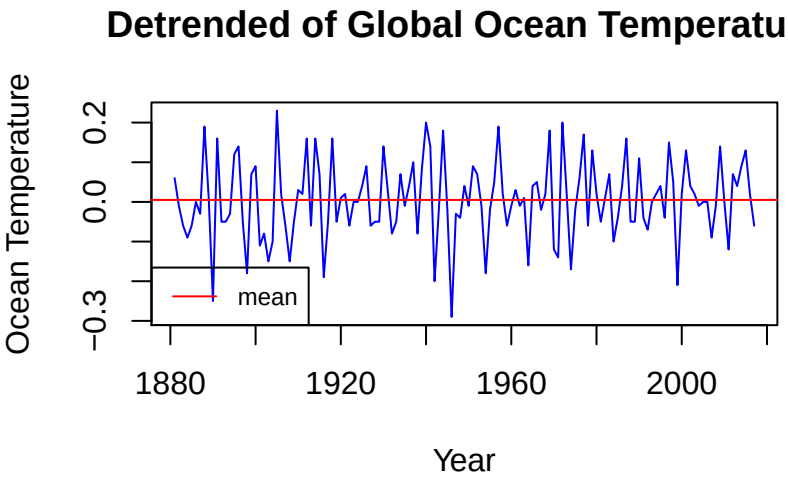
3. Statistical Methods



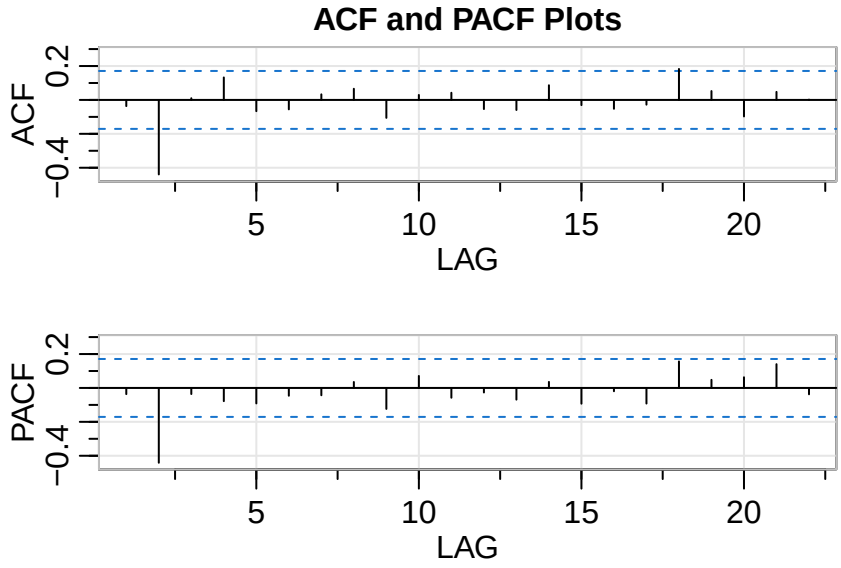


We try to model the data using the time-domain approach. We denote the annual global mean ocean temperature as x_t.

From the graph, We can see that there is some nonstationary trend. Our first step is to detrend. By looking at the sample ACF,$\rho(\hat{h})$, we find a slow decay in $\rho(\hat{h})$ relative to time increment. This is an indication of differencing.



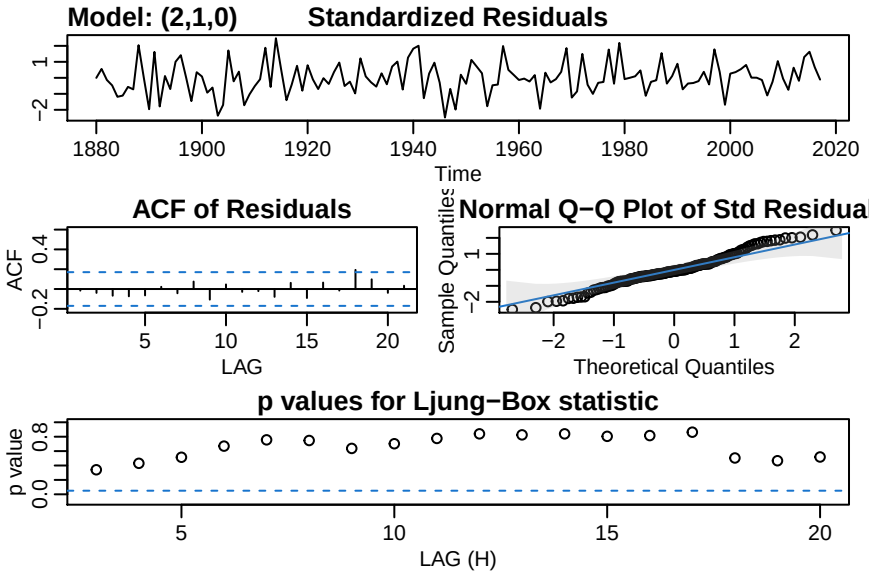
The above figure is the first difference of Global Temperature, ∇x_t. The red line indicates the mean. It appears to be a stable process with approximately zero mean and constant variance. Then, the next step would be identifying the dependence orders. To do that, we need to examine the ACF and PACF plots.



We could see that PACF is cutting off at lag 2, which means a non-zero correlation between the current and the previous two values that cannot be explained by all shorter lags. Therefore, the order of dependence for autoregression(AR) is 2 on the first difference, ∇x_t. ACF

on the other hand merely represents a correlation between current and lagged value. It's cutting off at lag 2, which indicates the order of dependence for moving average(MA) is 2 on the first difference, ∇x_t. Now, we add back the linear trend to construct the Auto Regressive Integrated Moving Average model for our data of interest, ocean temperature data x_t. Therefore, we are proposing two candidate models for x_t. The first one is ARIMA(2, 1, 0), which is equivalent to AR(2) on ∇x_t, and the second is ARIMA(0, 1, 2), which is equivalent to MA(2) on ∇x_t.

Now, let's examine the two model seperately. First, we look at ARIMA(2, 1, 0): $x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$.



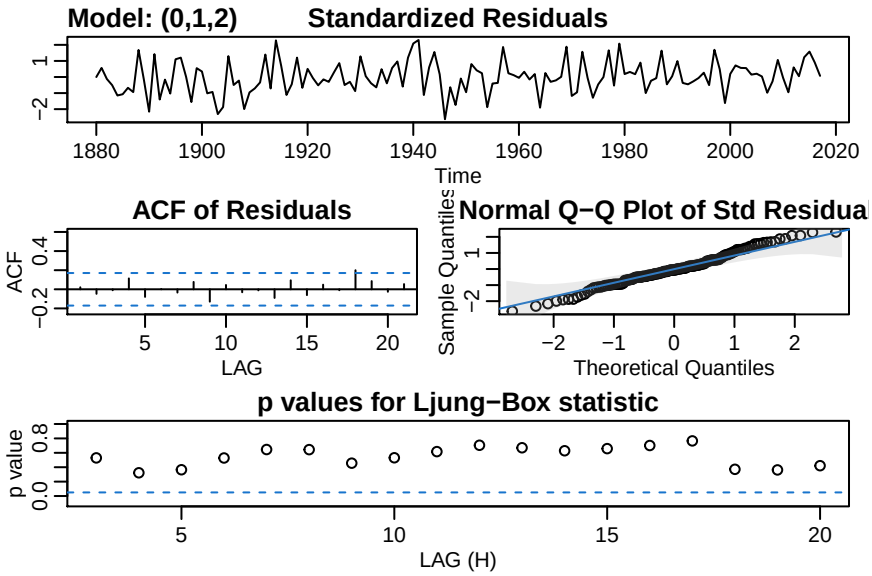

```
##          Estimate     SE t.value p.value
## ar1       -0.0519 0.0767 -0.6762  0.5001
## ar2       -0.4375 0.0761 -5.7522  0.0000
## constant   0.0051 0.0051  0.9905  0.3237

##
## sigma^2: 0.00785963

## degrees of freedom: 134
```

ar1, ar2 are the regression coefficients for ϕ_1, ϕ_2 in our equation model. From the table, we are 95% confident that ar2 is non-zero, but not ar1. Now we look at the diagnostic plot. The standardlized residual plot, ACF of residual and Ljung-Box indicate randomness and constant variance. The Q-Q plot represent normality. All assumptions hold except for a few outliers.

Then, we look at ARIMA(0, 1, 2): $x_t = \mu + \theta_1 w_{t-1} + \theta_2 w_{t-2} + w_t$.




```
##          Estimate     SE t.value p.value
## ma1       -0.1204 0.0739 -1.6291  0.1056
## ma2       -0.4400 0.0703 -6.2556  0.0000
## constant   0.0050 0.0034  1.4682  0.1444

##
## sigma^2: 0.00785963

## degrees of freedom: 134
```

ma1, ma2 are the regression coefficients for θ_1, θ_2 in our equation model. From the table, we are 95% confident that ma2 is non-zero, and 90% confident that ma1 is non-zero. Now we look at the diagnostic plot. The standardlized residual plot, ACF of residual and Ljung-Box

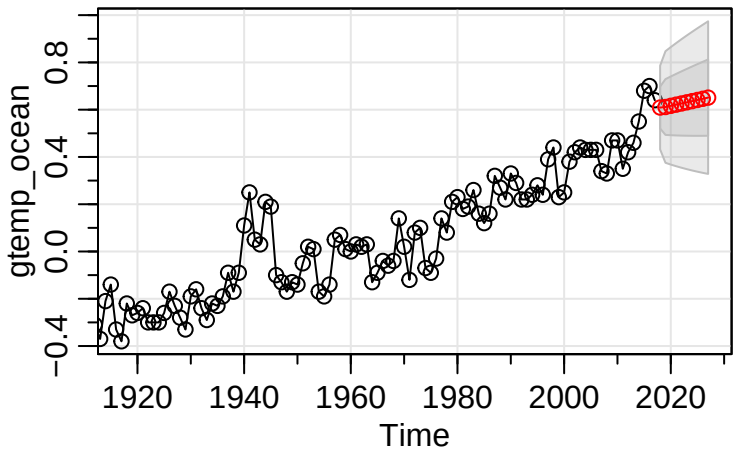
indicate randomness and constant variance. The Q-Q plot represent normality.

All of our assumptions hold. Until now, our second model ARIMA(0,1,2) appears to be better because there is no strong indicator that our first coefficient for ARIMA(2,1,0) is 0. To confirm this, we compare the estimators of prediction error: AIC, AICc, and BIC. As the name suggests, we want the smaller estimators the better.

```
##                       AIC      AICc       BIC
## ARIMA (2,1,0) -1.946634 -1.945317 -1.861379
## ARIMA (0,1,2) -1.947477 -1.946160 -1.862222
```

The above table is the estimators of prediction error, and we can see it confirms our preference. Model ARIMA(0,1,2) has smaller estimators and therefore fits our data better.

Now, we are finally ready to predict the global ocean temperature for the years 2018-2027 using our selected model ARIMA(0,1,2).

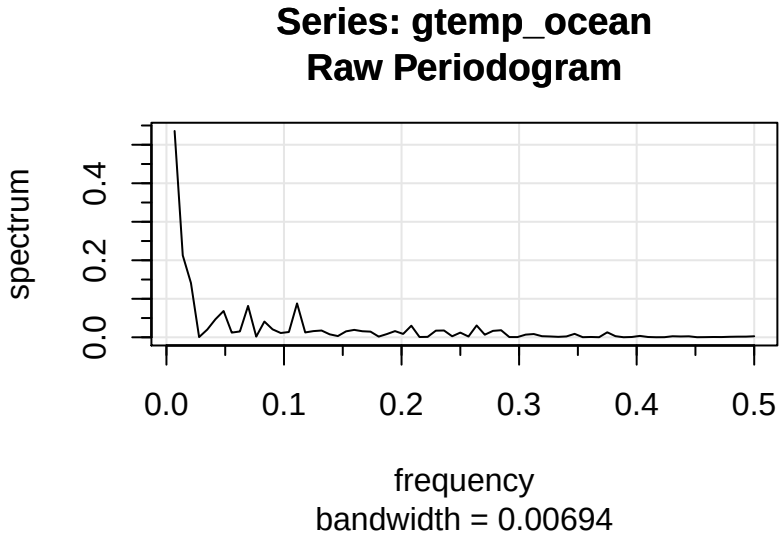



```
##    Prediction PI.95..Lower.Bound PI.95..Upper.Bound
## 1   0.6091037          0.4354498          0.7827576
## 2   0.6113172          0.3800405          0.8425939
## 3   0.6162881          0.3727364          0.8598397
## 4   0.6212589          0.3660220          0.8764958
## 5   0.6262297          0.3598196          0.8926399
```

```
## 6   0.6312006          0.3540673          0.9083338
## 7   0.6361714          0.3487147          0.9236281
## 8   0.6411422          0.3437203          0.9385642
## 9   0.6461131          0.3390491          0.9531771
## 10  0.6510839          0.3346715          0.9674963
```

The red points in the above graph are our prediction, where the two grey areas represent 90% and 95% prediction interval respectively. The table shows the precise number for the 95% prediction interval. From the graph, our prediction seems reasonable. Thus, there is no obvious indicator for overfitting.

Now let's perform spectral analyses to inspect cycles.



The higher the spectrum is, the more dominant the frquency is. The first three predominant frequencies are:

```
## frequency    period  spectrum
##    0.0069  144.0000    0.5357
```

```
## frequency    period  spectrum
##    0.0139   72.0000    0.2127
```

```
## frequency    period  spectrum
##    0.0208   48.0000    0.1410
```

We use the 95%-confidence interval to determine Whether the predominant periods are significant enough.

```
##   Dominant.Freq Spectrum Lower.Bound Upper.Bound
## 1        0.0069   0.5357      0.1452     21.1590
## 2        0.0139   0.2127      0.0577      8.4012
## 3        0.0208   0.1410      0.0382      5.5692
```

The 95%-confidence intervals for frequencies as estimators are wide, therefore, it's susceptible to uncertainties. Thus, we fail to establish significance of peak. To further exploiting and investigating frequency estimators of the global mean ocean temperature, a reduction of variance is needed.

4. Results

Even a small change in global mean ocean temperature is closely associated with several severe risks. Therefore, it is desired to model and predict global ocean temperature. To forecast the near future, we study the annual average ocean temperature from 1880 to 2017.

The data is a time series. Exploring time series and autocorrelation function plot, we note that it's not a stationary process. After differencing, the data appears station. Then, we determine its dependence orders of an ARIMA model by looking at the cutoff lag for ACF and PACF plots. ACF plot of the detrended data cutoff at lag 2, so we propose ARIMA(0,1,2): $x_t = \mu + \theta_1 w_{t-1} + \theta_2 w_{t-2} + w_t$ for original mean ocean temperature. PACF plot of the detrended data cutoff at lag 2, so we propose ARIMA(2,1,0): $x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$ for original mean ocean temperature.

We need to examine both models to pick the best one. So we fit the data to both models. The

p-value for each parameter indicates whether the parameters are significantly different from 0. For a well-fitted model, residuals need to be random and normal with constant variance. The residual for fitting data are independently and normally distributed around 0 as the standardized residual plot shows no apparent pattern. Points in Ljung-Box and ACF all lie within the reasonable level, which indicates randomness and independence. Most points in the Q-Q plot lies on the diagonal line except for some outliers. This means normality assumption holds except for a few outliers. Both two model's residuals satisfy our assumptions. However, one parameter of model ARIMA(2,1,0), ϕ_1, is not statistically significant. This means ARIMA(2,1,0) might not be the best model. At last, we examine estimators of prediction error: AIC, AICc, BIC. All three estimators for ARIMA(0,1,2) are slightly smaller, indicating a better fit. Therefore, we select ARIMA(0,1,2). With this model, we forecast the mean ocean temperature for the future 10 years. Our forecast and its 95% prediction interval seem to follow the trend and seem reasonable. Thus, there is no evidence for overfitting.

The model we derive indicates that the data is a moving average of order 2 process around an increasing linear trend. In our scientific background, this means that global ocean warming is an overall trend, where each year has some variant that depends solely on the previous two years. Therefore, in the big picture, the ocean temperature will eventually increase. We need to be prepared for the impacts and take close look at the last two years for prediction.

We also perform spectral analysis on the data and find the three periods with the highest spectrum: 144(=1/0.0069) years cycle, 72(=1/0.0139) years cycle and 48(=1/0.0208) years cycle. Their 95% confidence interval is extremely wide. This means that the frequency estimator is susceptible to huge uncertainties. So, it is not useful. We can smooth the spectrum to increase certainty and narrow the confidence interval. This way, it's possible to reach another statistically significant model regressing on sinusoids.

5. Discussion

In this project, we try to model the annual global mean ocean temperature from 1880 to 2017 and predict the temperature for the future 10 years. Our model is a relationship around an increasing linear trend. The relationship is that the temperature for every year is dependent only on the temperature in the previous two years, with some error terms and average 0. This means that the gloable annual ocean temperature is overall linearly increasing with some variant. Using this model, we forecast the annual temperature mean for the years 2018-2027. We also derive the prediction interval in which we are 95% confident to capture the real temperature.

The model fits well overall. But there are still some outliers. The prediction interval is also relatively wide. Thus, there might be some hidden periodic cycles that we fail to capture. In addition, global temperature changes last throughout the earth's history. Comparing to the earth's age, our data over 137 years is little. Therefore, we might not be able to capture the overall trend of global change in the long term because of lack of data.

Reference

- United States Environmental Protection Agency. Climate Change Indicators: Sea Surface Temperature. Retrived April 15, 2021 from https://www.epa.gov/climate-indicators/climate-change-indicators-sea-surface-temperature#:~:text=Sea%20surface%20temperature
- Shumway, R., Stoffer, D. 2016. Time Series Analysis and Its Applications with R Examples.