Prediction of the Mean Ocean Temperature using ARMA model

Wenxi Li

Abstract

This study uses R to stimulate an ARIMA(0,1,2) model to forecast the mean temperature of the ocean for the next 10 years based on the data in the past 133 years. We forecast that the mean ocean temperature drops from the year 2017 to 2018 by 0.031 Celsius degree. After the year 2018, a steady annual increase of about 0.005 degree Celsius in the mean ocean temperature is found. A spectral analysis is also carried out. However, none of the first three predominant spectrum is found to be significant, and we are unable to distinguish between them. Therefore, no specific dominant period of mean temperature changing is found based on the previous 133 years of data.

Introduction

Global warming has become one of the most important tasks for the human to tackle, and one of the most consequential results of global warming is the significant increase in the mean ocean temperature over the past century. The study of the ocean temperature is meaningful and contributive in various ways; the ocean is considered the most effective way to monitor global warming as it is the main heat absorber for the earth.

In this study, we will be mainly focusing on predicting the mean ocean temperature for the next 10 years using the data of mean ocean temperature collected over the past 133 years, which could be found in the R package astsa. The temperature was collected in ice-free open ocean surface in different regions and then get averaged.

Apart from predicting for the next ten years of the mean ocean temperature. We will also try to examine the three most predominant spectrums and their corresponding frequency and periods, which indicates the time interval that captures the trend of variations in the mean ocean temperature.

Statistical Methods

For this study, we are going to use an autoregressive integrated moving average model (ARIMA), which is a model commonly used in analysing time series data, to fit our past observed data, and use it to predict the trend in the future.

Figure 1 shows a sketch of the original data, we observe that there is an increasing trend in the mean temperature from 1880 to 2020. In order to build a model for the data, we have to first make it stationary so that the data shows a constant expectation value of temperature as well as a constant variance over time. Figure 2 shows the data after we took the first difference of the temperature, and it now looks stationary with a constant expectation value over the years and a roughly constant variance.

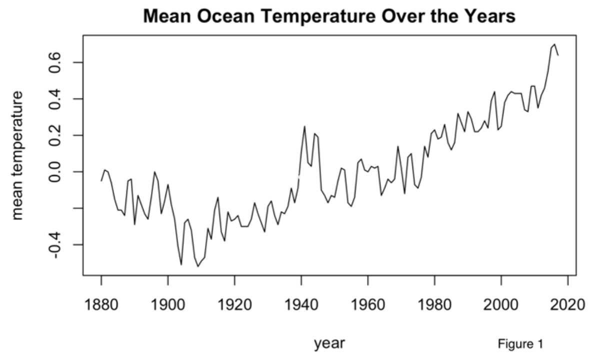

Figure 1

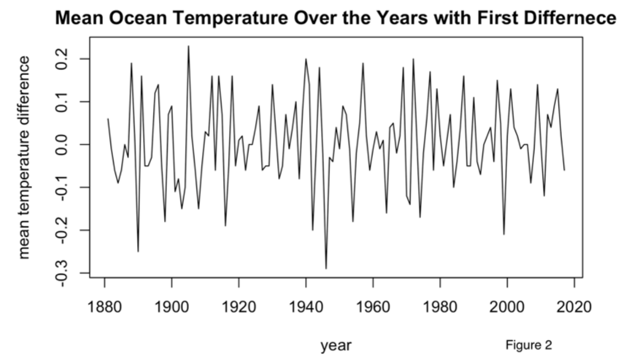

Figure 2

In order to find out which ARIMA model fits the data well, we need to first see its autocorrelation function (ACF) and partial autocorrelation function (PACF). The ACF and PACF graph of the differenced data are shown below in figure 3. We see that both the ACF and PACF is cutting off at lag2. Therefore, we propose two arima models, ARIMA(2,1,0) and ARIMA(0,1,2), to fit and original data, where the 1 indicates we took the first differential of the original data to build the model.

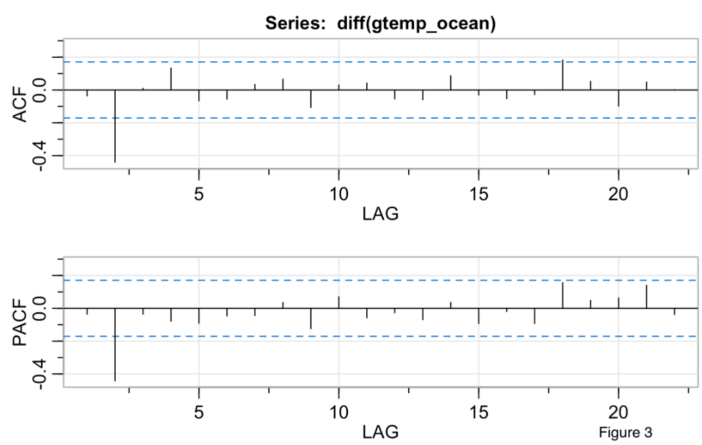

Figure 3

In order to see which one of the two models we are going to use, we first do four of the diagnostic check. Figure 4 shows the diagnostic check for model ARIMA(2,1,0) and Figure 5 shows the diagnostic check for model ARIMA(0,1,2). For it to be a valid model, we would need the standardized residuals to show no obvious pattern, no major outliers, which is exactly what we observe for both two models. For the ACF of the residuals, we want the residuals to have no significant autocorrelations. As we see in the two figures below, both of the models have almost all of their residuals to be insignificant at the 5% level except for lag 18. But we could still conclude that the assumption of the randomness of the residuals is not violated. We would also need to check if the residuals follow a normal distribution. From the normal Q-Q plot of the two models, we could observe that the distribution of residuals for model ARIMA(0,1,2) is closer to the Normal Q-Q distribution line than those for the model ARIMA(2,1,0). But ARIMA(2,1,0) is still an acceptable model, given that most of the residuals follow a normal distribution, except for few outliers at the two tails. The last diagnostic check is to make sure all the residuals are independent. To do that, we check the Ljung-Box statistic, and we want the statistic for all the lags to be insignificant, indicating that the residuals are independent. We see that the p-values for the Ljung-Box statistic for all the lags for both the models are above 0.05, which indicates all the residuals are independent.

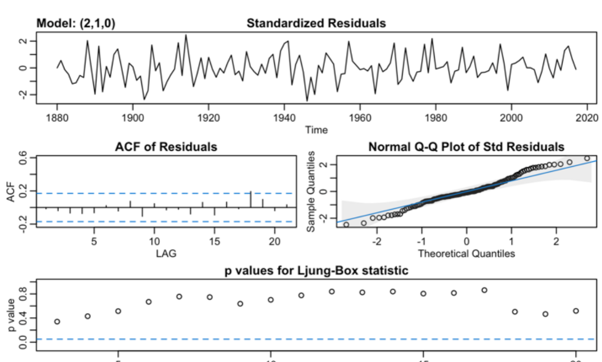

Figure 4

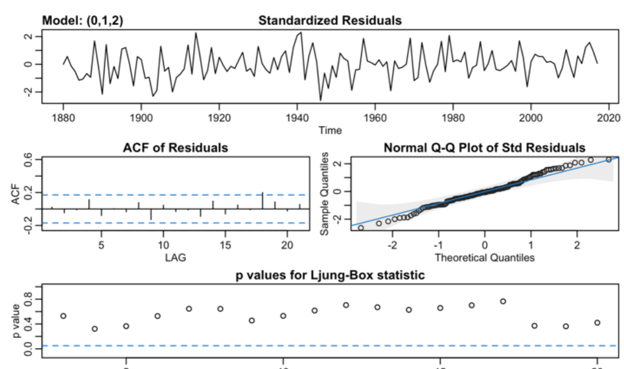

Figure 5

As discussed above, both of the models pass the four diagnostic tests. We could further check the AIC and BIC values for the two models to see which model fits the data better.

Table1: AIC/BIC for both models

	ARIMA(0,1,2)	ARIMA(2,1,0)

BIC	-1.862222	-1.861379
AIC	-1.947477	-1.946634

AIC and BIC values are calculated and shown in Table 1. Since both the BIC and AIC values for ARIMA(2,1,0) are greater than those for ARIMA(0,1,2), and the residual distribution for the latter model also shows a closer pattern to the normal distribution, we select the model ARIMA(0,1,2). The final model we have is:

$$X_t = \mu + w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2}$$

Where μ is the mean ocean temperature over the years based on the model, w_t is the white noise at time t. θ_1 and θ_2 are the parameters for the corresponding white noise.

Results

Table 2 summarise the estimates for the model ARIMA(0,1,2). The estimates for ma1, ma2 and the mean of the temperature estimated are -0.1204, -0.4400, and 0.005 respectively. The p-values indicate that the parameter for MA2 is significant while MA1 and the constant are insignificant. This indicates that there's no enough evidence that the mean temperature of the ocean over the years is 0.005 degree celcius, and there's also lack of evidence that the future ocean temperature depends on the white noise of the previous year.

As the model we have is an MA2 model, we could forecast the future mean temperature of the ocean as a function of the mean temperature (represents as the constant in table 2) and the white noise from the previous two years, i.e., $X_t = 0.0005 + w_t - 0.1204w_{t-1} - 0.4400w_{t-2}$

Table 2: Main results for ARIMA(0,1,2)

	Estimate	Standard Error	P-value
MA1	-0.1204	0.0739	0.1056
MA2	-0.4400	0.0703	0.0000
constant	0.0050	0.0034	0.1444

We then could use the model above to make predictions for the mean temperature of the ocean for the next ten years.

Figure 6 shown below illustrates the temperature trend forecasted by the model, in which we clearly see an increasing trend after the year 2018. The only exception occurs at the time from 2017 to 2018 where we observe a decrease of 0.031 degree Celsius (year 2017 has a recorded mean ocean temperature of 0.64 degree Celsius).

As summarised in Table 3, we see an increase of 0.002 degrees Celsius in the mean temperature from 2018 to 2019. Then after the year 2019, the mean temperature has an annually increase in the mean ocean temperature of 0.005 degree Celsius.

The 95% prediction interval is also suggesting a continuing rise in the mean temperature; we could see that the upper bound of the prediction interval increases at a rate that is significantly greater than the decreasing rate of the lower bound, except for the time from 2018 to 2019.

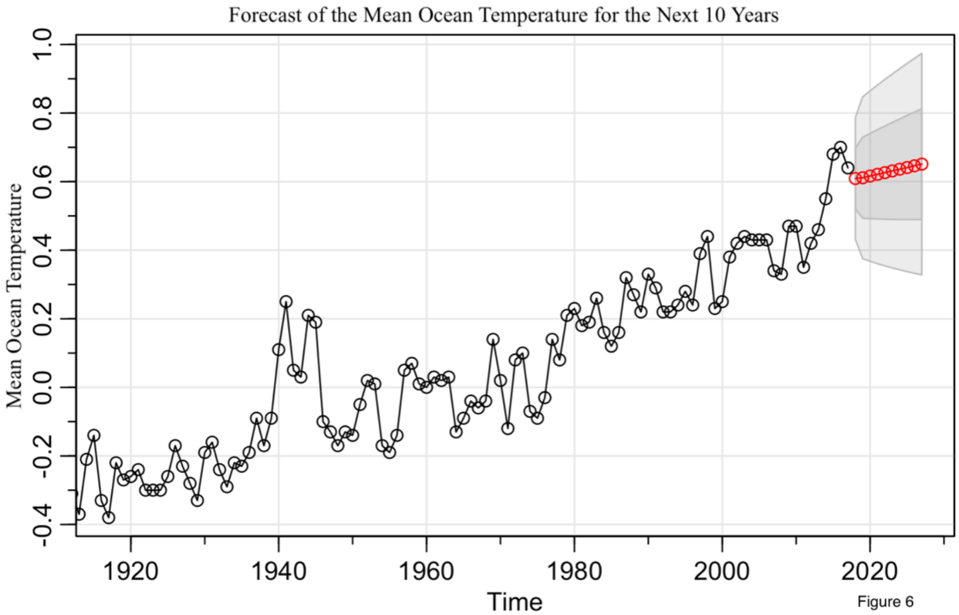

Table 3: Prediction of the mean ocean temperature of the next ten years

Year of Prediction	Prediction of the mean ocean temperature	Lower bound of 95% Prediction Interval	Upper bound of 95% Prediction Interval
2018	0.6091037	0.4354498	0.7827576
2019	0.6113172	0.3800405	0.8425939
2020	0.6162881	0.3727364	0.8598397
2021	0.6212589	0.3660220	0.8764958
2022	0.6262297	0.3598196	0.8926399
2023	0.6312006	0. 3540673	0.9083338
2024	0.6361714	0.3487187	0.9236281
2025	0.6411422	0.3437203	0.9385642
2026	0.6461131	0.3390941	0.9531771

2027	0.6510839	0.3346715	0.9674963

Next, by doing the periodogram analysis, we could find the top three predominant frequencies as well as the top three predominant periods. We place the spectrum in descending order as shown in table 4. We find out that the first three dominant periods are 144 years, 72 years, and 48 years. This indicates that the change in mean temperature of the ocean follows a periodic behavior; the mean temperature shows a similar trend every 144, 72, or 48 years. Out of these periods, 144 years turns out to be the most dominant period, with the other two less dominant periods of 72 and 48 years.

Table 4: First three predominant frequency and periods

	Frequency	Period	Spectrum
1	0.0069	144.0000	0.5357
2	0.0139	72.0000	0.2127
3	0.0208	48.0000	0.1410

We could further test the significance of these spectrums. As shown in table 5, we notice that the 95 confident intervals of the spectrum are extremely wide for all three periods. We could see that the spectrum for the first dominant periods (0.5357) lies in the 95% confident interval of the second (0.0577,8.4012) and the third dominant spectrum (0.0382, 5.5692). And the second dominant spectrum (0.2127) lies in the 95% confidence interval of the third confidence interval (0.0382, 5.5692). So, we are unable to conclude anything about the significance of the spectrum and therefore, the dominant periods we concluded before are not significant.

Table 5: 95% confident interval for Spectrum of the three predominant Periods

	Dominant Periods	Spectrum	Lower	Upper
1	144.0000	0.5357	0.1452	21.1590
2	72.0000	0.2127	0.0577	8.4012
3	48.0000	0.1410	0.0382	5.5692

Discussion

As previously mentioned in the results part, we were unable to find a specific period that captures the change in the mean ocean temperature. This may be due to the lack of data in the study so that no dominant period is showing up. After collecting more data in the future, we are more possible to find one that illustrates the pattern of the variations in the mean ocean temperature.

Also, we do see that in the diagnostic tests, there're some outliers of the residuals that don't follow a perfect normal distribution and the autocovariance at lag 18 exceeds the limit. So the model (ARMA(0,1,2)) that we chose might not be perfect, but it is still a reasonably acceptable one. Nevertheless, we could try for several different other possible ARMA models and compare them to find a better one that fits the data, and the forecasts of the temperature would be more accurate.

Reference

1. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/
2. Package astsa. David Stoffer. https://github.com/nickpoison/astsa