Kristina White
Professor Castruccio
ACMS 60855
25 March 2021

Undergraduate Exercise: Sea Ice Data

This dataset contains measurements of Arctic sea ice extent (measured in square kilometers) from October 1978 to February 2021. These measurements are taken using satellite imaging. Satellite sensors observe the microwaves emitted by the surface of the ice. Since microwave energy passes through clouds, it can be measured year-round (National Aeronautics and Space Administration [NASA], 2008). These measurements began in 1978 with the Scanning Multichannel Microwave Radiometer (SMMR) on the Nimbus-7 satellite (1978-1987) (NASA, 2008). Other satellite technology has been used since then.

The satellite sensors are able to distinguish between ice and ocean water because ocean water emits microwaves differently. Each pixel of the digital picture elements from the satellite observations represents a 25 km X 25 km square, and the amount of ice in each pixel is estimated (NASA, 2008). Sea ice extent is the area of ocean that contains at least some ice. To estimate this quantity, every pixel that meets or exceeds a certain threshold percentage is counted as "ice-covered." The National Snow and Ice Data Center uses a threshold of 15% (NASA, 2008).

The sea ice extent values in the dataset are in the millions of square kilometers, which is what one would expect from measurements of oceanic area. The measurements are for the sum of the 14 regions of the Arctic, or the full Arctic. The minimum value of the time period was 3,552,586 km^2, and the maximum value was 16,072,926 km^2. It is also important to consider the quality of the data. More sophisticated instruments have been put into use since data collection started in 1978. Additionally, data collection has been done by different people over time and has likely also become more automated. While the scientists likely tried to keep the data as uniform as possible, these differences in instruments and collection could affect the quality of the data. The data was measured and collected by sophisticated equipment and qualified scientists, so it should be reliable.

The sea ice extent measurements are important as scientists are worried about the impact of climate change on sea ice. Melting ice could raise ocean levels, posing problems for coastal communities, cities, and countries. Additionally, sea ice is bright and reflects sunlight back into space. When overall warming temperatures melt sea ice, more sunlight is absorbed by the ocean water, raising water temperatures, which begins causes delayed ice growth in the fall and winter and faster melting in the spring, leaving the ocean waters exposed for longer durations and creating a positive feedback loop (National Oceanic and Atmospheric Administration [NOAA], 2021]. This can alter ocean circulation and further affect the global climate (NOAA, 2021).

Figure 1 shows a plot of the data. There is clearly a seasonal variation, and the sea ice extent seems to be generally decreasing from 1978 to 2021. Figures 2 and 3 look at the monthly and annual trends, respectively. Sea ice extent increases and has higher measurements in the winter months and the decreases and has lower measurements in the summer. This makes sense because we would expect some of the ice to melt when its warm, and more of the water to freeze when its cold. The yearly averages for sea ice extent are trending downward. This is to be expected with climate change and the overall increase in the Earth's temperature. The bumps or fluctuations in the sea ice extent averages are likely due to temperature variations from year to year. Some years are warmer or colder than others. The spike in the graph for 2021 is artificial. The yearly average for 2021 is based only on the measurements for January and February, which are winter months when we expect to see greater sea ice extent. We would expect this spike to disappear when the data is collected for the rest of 2021 and the warmer months are included.

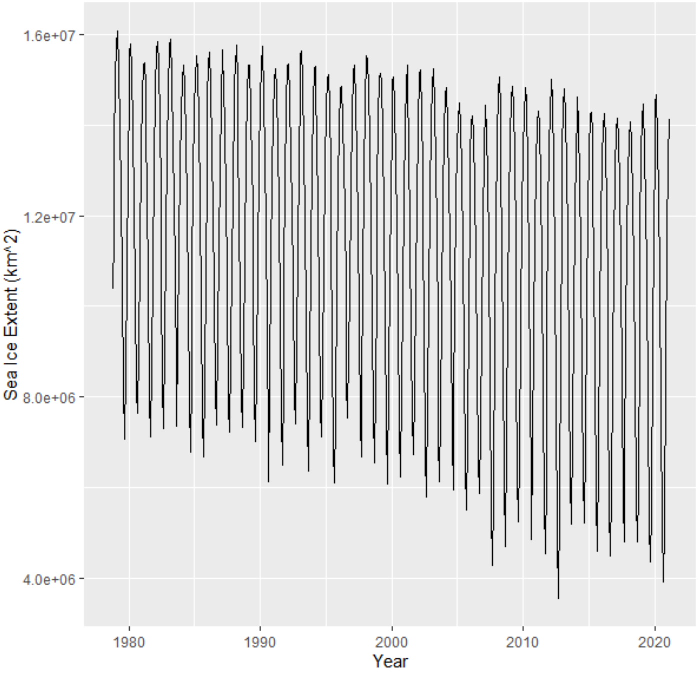


Figure 1 Arctic sea ice extent (km^2) from October 1978 to February 2021

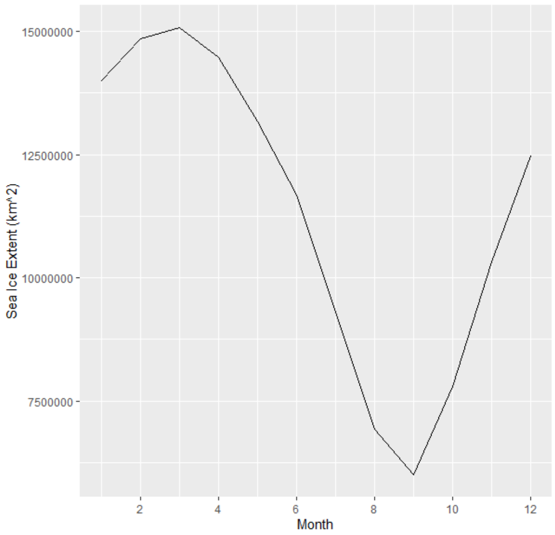


Figure 2 Average monthly trend for Arctic sea ice extent (km^2)

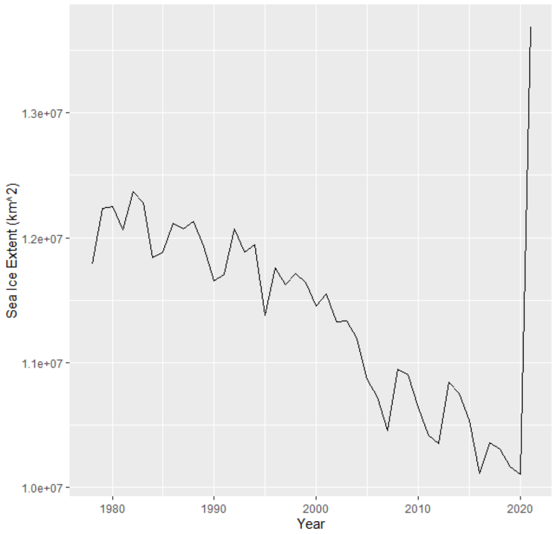


Figure 3 Plot of yearly averages for Arctic sea ice extent (km^2)

I fit a linear model for the data with a seasonal trend. Looking at Figures 1 and 3, a linear model would be reasonable, and looking at Figure 2 a seasonal trend is necessary. I looked at models with 1, 3, and 4 harmonics:

```
mod.h1=tslm(y.ts~X_1+year.ts)

mod.h3=tslm(y.ts~X_3+year.ts)

mod.h4=tslm(y.ts~X_4+year.ts)
```

where year.ts is a vector of the years of observation (1978-2021), and X_1, X_3, and X_4 are matrices containing terms from Fourier series up to orders 1, 3, and 4, respectively. The p-value for the linear term is < 2e-16 for all three models, so the linear trend is significant (using α = 0.05). The p-values for the sine and cosine terms for the first and third harmonics were significant, but the sine and cosine terms for the fourth harmonic were not significant (p-values of 0.08275 and 0.15715, respectively), so I chose the second model using three harmonics. Figure 4 shows the model. The model has an R^2 value of 0.9833 and appears to be a good model. We use this model to predict the trend of Arctic sea ice extent over the next 22 months (March 2021 to December 2022).

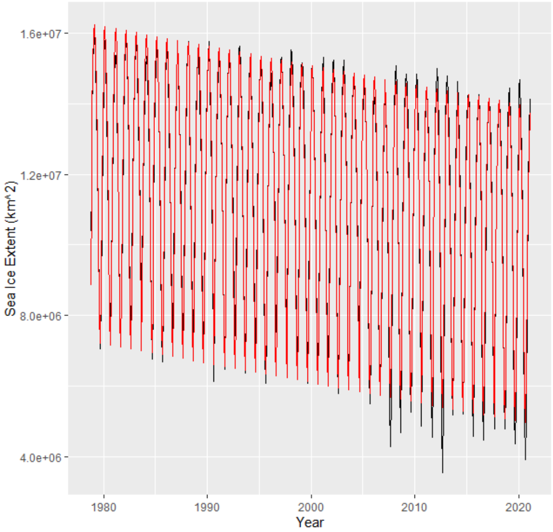


Figure 4 Graph of the fitted values for Arctic sea ice extent model with 3 harmonics

Figure 5 shows the plot of the residuals (observed values – fitted values). Since this is not white noise, there is still structure in the residuals to explain. I fit an ARMA model on the residuals in order to detrend the data. The AR(4) model has φ = (0.7937, -0.0778, -0.0061, -0.1445) and σ^2 = 8.056e+10. Looking at the plot of the autocorrelation function for this model, while the model captures most of the variability in the residuals, we see large spikes outside the 95% confidence interval at lags 12 and 24. This indicates a seasonal affect the model has not captured. We could bring these spikes inside the 95% confidence interval if by refitting the model with seasonal ARMA, but we have not covered that model in class, so I will continue with the non-seasonal ARMA.

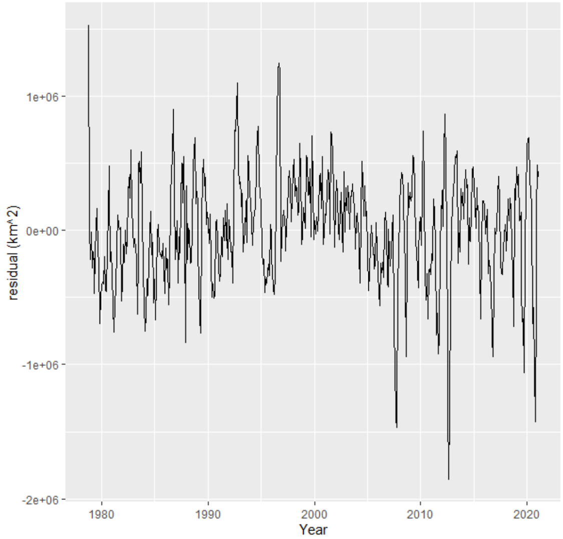


Figure 5 Graph of residuals (km^2)

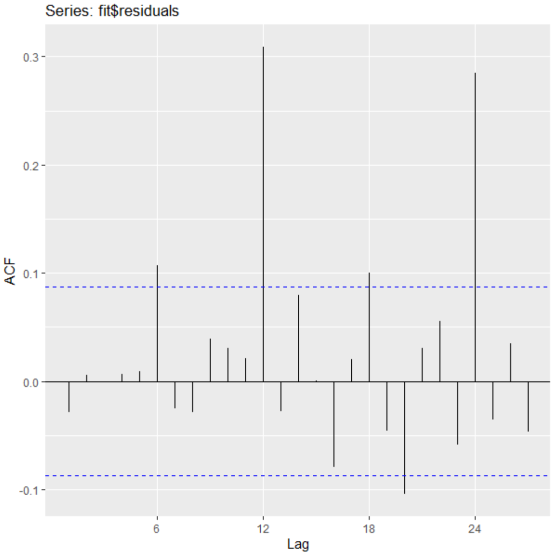


Figure 6 Plot of the autocorrelation function

I created a forecast of the ARMA model over the next 22 months, and added the trend predicted by the linear model into this forecast, so it contains the predictions from the residuals and from the linear trend. These forecasts are shown in Figures 7 and 8, with the 80% confidence interval in dark blue and the 95% confidence interval in light blue. Looking at these figures, these predictions seem to be consistent with the observed trend and seasonality. Since the data came from reputable source and is likely of the best possible quality, these predictions are reliable. However, they do not account for catastrophic events that could affect temperatures and sea ice extent.

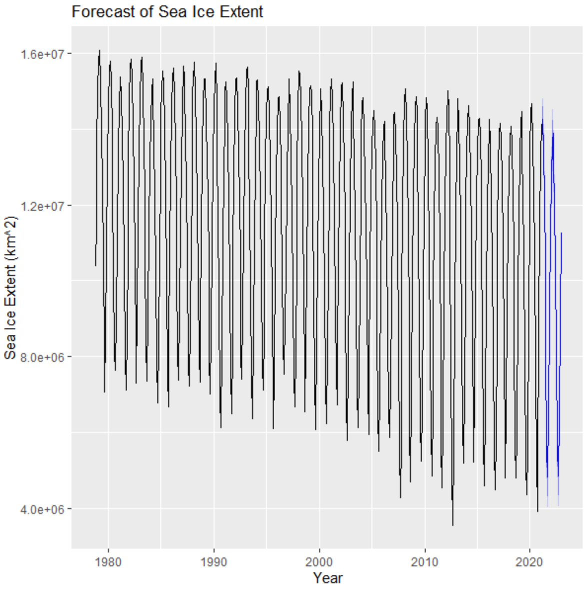


Figure 7 Forecast for the next 22 months of Arctic sea ice extent

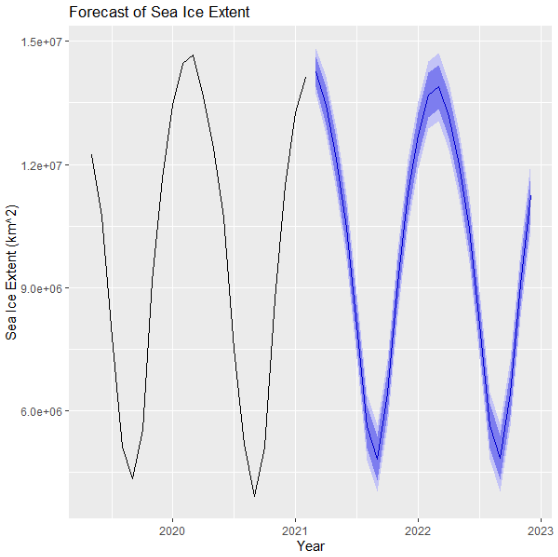


Figure 8 Forecast of Arctic sea ice extent, zoomed in

Finally, this analysis looked at data collected via satellite on Arctic sea ice extent from October 1978 to February 2021. Sea ice extent is a measure of the area of ocean (in square kilometers) that contains at least some ice. Sea ice extent has a seasonal pattern where it increases in the fall and winter as temperatures cool and decreases in the spring and summer when warmer temperatures melt the ice. There has been a decreasing trend in Arctic sea ice extent since 1978, likely because of warming global temperatures due to climate change. The predictions for March 2021 through December 2022 continue to follow these patterns, assuming there are no catastrophic events that could alter global climate or temperature, such as nuclear war. Keeping track of sea ice extent is important because if too much ice melts, more ocean water will be exposed and will absorb solar energy due to its darker color. This increases temperatures and causes further melting, which becomes a cycle. This can disrupt normal ocean circulation, which can change the global climate, making sea ice an important factor in climate change.

Graduate Exercise: Forecast Simulation Study

When we have time series data, we want to be able to make a prediction about what is going to happen in the future. We do not predict events in the past that have already happened, it is harder to predict what will happen farther into the future. This is because, for example, the weather right now is likely a good indicator of the weather in the next 30 minutes, but not the next 30 days. We use forecasting to make these predictions with the best possible use of the information we already have.

However, in order to produce a forecast, we must first fit a model. This can be done using a variety of model selection and parameter estimation techniques. AIC, AICc, and BIC can be used when we do not know the order of the model, and methods like MLE and Yule-Walker can be used to estimate parameters when the model order is known. The consequence of fitting the model before performing the forecast is, we are assuming we know the true parameters of the model when we do not. When we use a model to produce a forecast, we are assuming that the estimated order and/or parameters of the model are the true parameters of the data, or the observed data exactly follow this model. Since this is not the case with real datasets, we should consider the effect of making these assumptions on our forecasts.

To that end, I have conducted a simulation study on ARMA models. I ran simulations on an AR(1) model with $\varphi=0.6$ and an ARMA(2, 2) model with $\varphi= (0.6,0.2)$ and $\theta=(0.8,-0.1)$. These models are stationary, and the ARMA(2, 2) model is invertible. For each, I simulated $X_1,\ldots,X_n$ and fit the same model for $X_1,\ldots,X_{n-1}$, and computed at 95% confidence interval for X_n. I used values of n from the sequence 6, 11, …, 101. For each value of n, I ran 500 simulations from the AR(1) model and 100 simulations from the ARMA(2, 2) model generating the data, fitting the model for $X_1,\ldots,X_{n-1}$, and forecasting and computing a 95% confidence interval for X_n. After 500 simulations, I found the mean percentage of predictions that fell within the confidence interval. The plots of these averages are included below.

The lowest value of n, n = 6, saw the lowest average percentage of forecasts falling within the 95% confidence interval for both orders of ARMA models. As n increased, in general this percentage trended higher. The fluctuations seen in the graphs are due to random variation and would decrease with increased simulations, though this does get more computationally expensive.

Assuming we know the true parameters of the model when forecasting does seem to yield accurate forecasts, according to the simulation study, for time series that are sufficiently long. These time series saw their forecasted values falling in the 95% confidence interval close to the expected 95% of the time in the simulation study. It makes sense that longer time series lead to better forecasts than shorter time series because they contain more information.

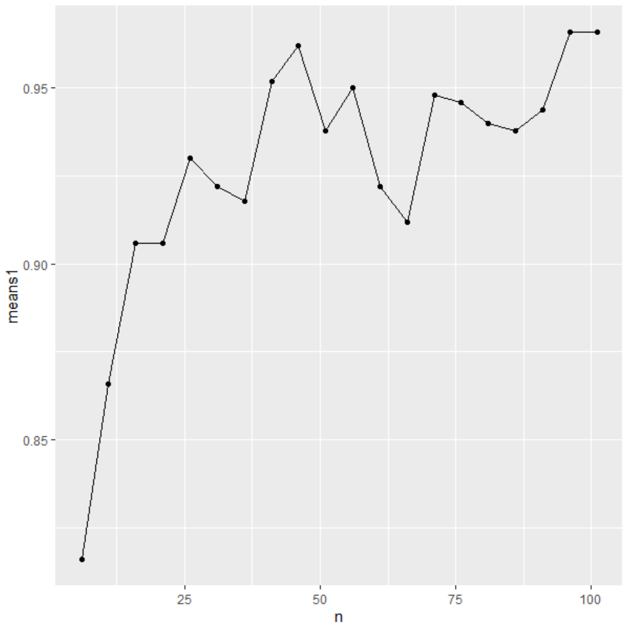


Figure 9 Plot of average percent of predictions of X_n that lie in the 95% confidence interval for AR(1) model

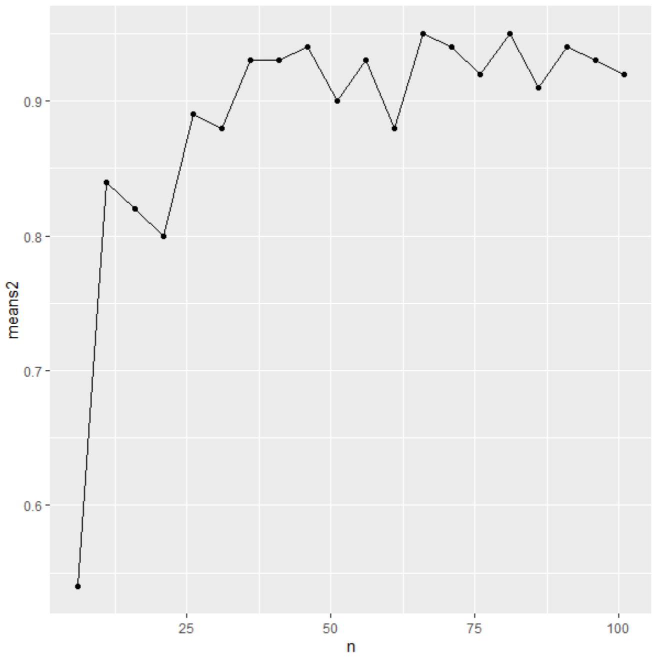


Figure 10 Plot of average percent of predictions of X_n that lie in the 95% confidence interval for ARMA(2, 2) model

However, it is important to note I used time series for this simulation study that were generated from specified models. Whenever we do forecasts, we used a fitted model that we are assuming is the true model. However, if we are working with a real-world dataset, such as the Arctic Sea Ice dataset used in the first part of this report, we use model selection to fit a model to the data. This data was not generated from a specific model, so there may be more noise in real data than in data whose parameters were pre-specified.

Assuming we know the true parameters of the model can still give us accurate forecasts, according to the simulation study. While forecasts from short time series will be less accurate, longer time series that contain more information should fall within a 95% confidence interval about 95% f the time. So, producing forecasts as if we know the true parameters should work decently if we fit a good model. In practice, we do not have data ore-generated to follow a specific model like we did in the simulation study. There may be more noise to capture in such data, or it may be more difficult to fit a model to the data. However, with good data and a sufficient amount of it, we should be able to find a good model and produce a reliable forecast. The forecast is only as reliable as the data, so if the data are not good, the forecast will not be either.

Appendix A: Arctic Sea Ice Extent Model with Quadratic Trend

I also fit a model for Arctic sea ice extent using a quadratic model with three harmonics. The p-value for the quadratic terms was 8.23e-05, so it is significant at $\alpha = 0.05$. The R^2 value is 0.9838, which is bit higher than 0.9833 for the linear model. However, this term does add complexity to the model, and more formal model selection methods could help to determine which model we should use.

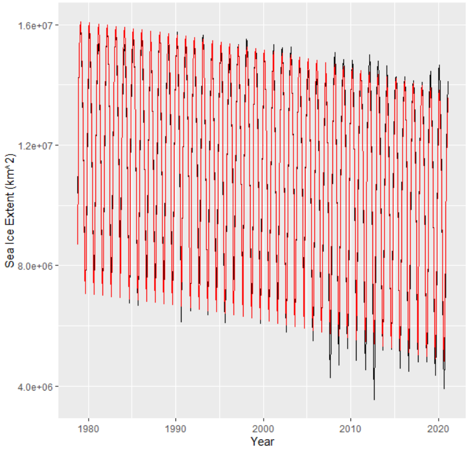


Figure 11 Fitted values for Arctic sea ice extent (km^2) from quadratic model

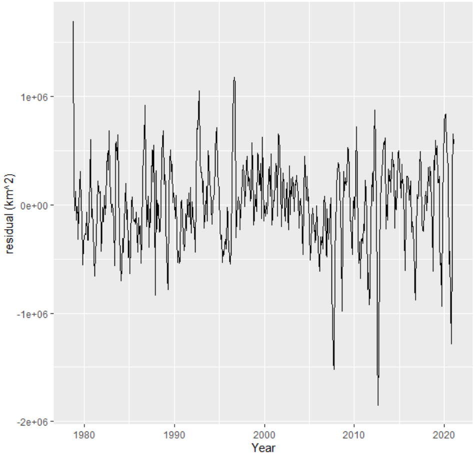


Figure 12 Plot of residuals from quadratic model

The model fits the data well, and the residuals have structure that still needs to be explained, much like for the linear model. I fit an AR(4) with $\varphi = (0.7937, -0.0778, -0.0061, -0.1445)$ model to the residuals. Looking at the autocorrelation function, we see spikes outside the 95% confidence interval at lags 12 and 24, suggesting a seasonal affect that could be explained with a seasonal ARMA model.

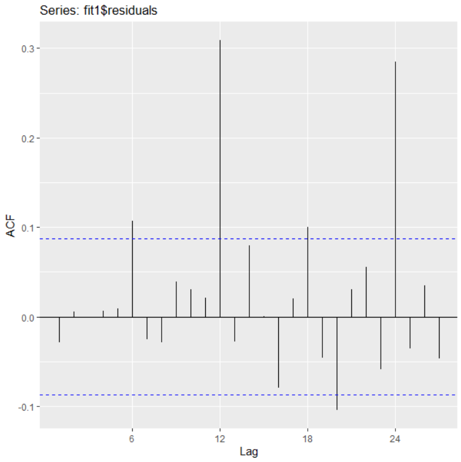
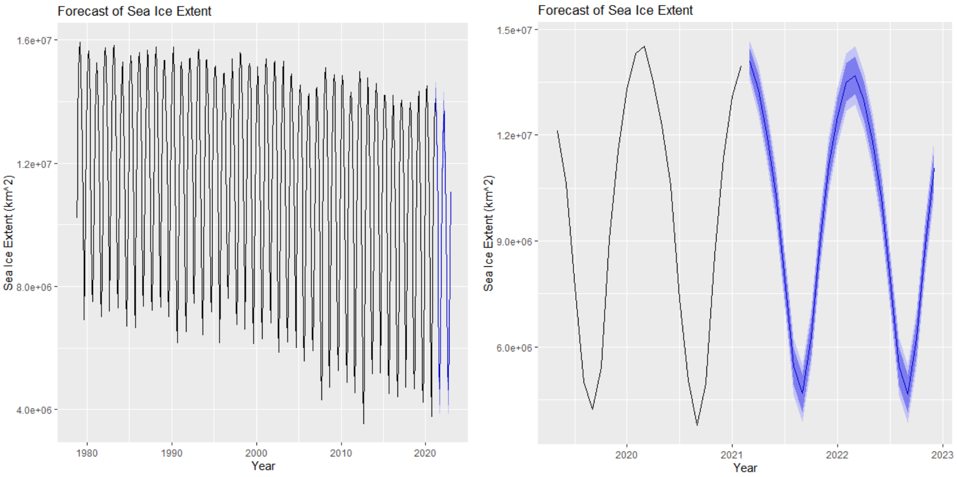


Figure 13 Plot of the autocorrelation function



Figures 14 and 15 Forecast for Arctic sea ice extent from March 2021-December 2022 using quadratic trend

Figures 14 and 15 show the forecasts Arctic sea ice extent from March 2021 to December 2022 using the trend from the quadratic model. The 80% confidence interval can be seen in light blue in Figure 15, and the 95% confidence interval can be seen in dark blue. The data is of the highest possible quality, and the quadratic and seasonal trends are significant, so the forecast should be reliable.

Appendix B: Code for Undergraduate Exercise

```
library(tidyverse)
library(forecast)

load("SeaIce.Rdata")

gp=autoplot(y.ts)+xlab("Year")+ylab("Sea Ice Extent (km^2)")
gp

#exploratory analysis: monthly profiles, yearly averages
min(y.ts)
max(y.ts)

data.monthly=aggregate(c(y.ts), list(month = cycle(y.ts)), mean)
data.annual=aggregate(c(y.ts), list(year = floor(time(y.ts))), mean)

data.monthly=ts(data.monthly[,2],start=1)
autoplot(data.monthly)+xlab("Month")+ylab("Sea Ice Extent (km^2)")

data.annual=ts(data.annual[,2],start=data.annual[1,1])
autoplot(data.annual)+xlab("data.annual")+ylab("Sea Ice Extent (km^2)")+xlab("Year")
year=time(data.annual)

# analyzing the trend: there is seasonality, we need the harmonics

X_1=fourier(y.ts, K=1)
X_3=fourier(y.ts, K=3)
```

```
X_4=fourier(y.ts, K=4)

year.ts=floor(time(y.ts))

mod.h1=tslm(y.ts~X_1+year.ts)
mod.h3=tslm(y.ts~X_3+year.ts)
mod.h4=tslm(y.ts~X_4+year.ts)
summary(mod.h1)
summary(mod.h3)
summary(mod.h4)

res.ts=y.ts-mod.h3$fitted.values

# fitted values
df=data.frame(X=time(y.ts),Y=mod.h3$fitted.values)
gp+geom_line(data=df,aes(X,Y),color="red")

# residuals
autoplot(res.ts)+xlab("Year")+ylab("residual (km^2)")

# predicting for all 2021 and 2022, total 23 months
# let's start with the trend

mn.hor=22

year.ts=c(kronecker(2021,rep(1,10)),kronecker(2022,rep(1,12)))
trend.predict=forecast(mod.h3,data.frame(fourier(y.ts, K = 3, h =mn.hor),year.ts))
```

```
# model selection and forecasting
fit=auto.arima(res.ts,d=0,seasonal=FALSE)
fit

ggAcf(fit$residuals)

forc=forecast(fit,h=mn.hor)

forc$x=forc$x+mod.h3$fitted.values
forc$mean=forc$mean+trend.predict$mean
forc$lower=forc$lower+trend.predict$mean
forc$upper=forc$upper+trend.predict$mean

# plotting the final results
autoplot(forc)+xlab("Year")+ylab("Sea Ice Extent (km^2)")+ggtitle("Forecast of Sea Ice Extent")
autoplot(forc,include=100)+xlab("Year")+ylab("Sea Ice Extent (km^2)")+ggtitle("Forecast of Sea Ice Extent")
autoplot(forc,include=mn.hor)+xlab("Year")+ylab("Sea Ice Extent (km^2)")+ggtitle("Forecast of Sea Ice Extent")



year.ts=floor(time(y.ts))
year.ts2=year.ts^2

mod.h=tslm(y.ts~X_3+year.ts+year.ts2)
summary(mod.h)

res.ts1=y.ts-mod.h$fitted.values
```

```
# fitted values
df=data.frame(X=time(y.ts),Y=mod.h$fitted.values)
gp+geom_line(data=df,aes(X,Y),color="red")

# residuals
autoplot(res.ts1)+xlab("Year")+ylab("residual (km^2)")

# predicting for all 2021 and 2022, total 23 months
# let's start with the trend

mn.hor=22

year.ts=c(kronecker(2021,rep(1,10)),kronecker(2022,rep(1,12)))
year.ts2=c(kronecker(2021^2,rep(1,10)),kronecker(2022^2,rep(1,12)))
trend.predict1=forecast(mod.h,data.frame(fourier(y.ts, K = 3, h =mn.hor),year.ts, year.ts2))

# model selection and forecasting
fit1=auto.arima(res.ts,d=0,seasonal=FALSE)
fit1

ggAcf(fit1$residuals)

forc1=forecast(fit1,h=mn.hor)

forc1$x=forc1$x+mod.h$fitted.values
forc1$mean=forc1$mean+trend.predict1$mean
forc1$lower=forc1$lower+trend.predict1$mean
```

```
forc1$upper=forc1$upper+trend.predict1$mean


# plotting the final results

autoplot(forc1)+xlab("Year")+ylab("Sea Ice Extent (km^2)")+ggtitle("Forecast of Sea Ice Extent")

autoplot(forc1,include=100)+xlab("Year")+ylab("Sea Ice Extent (km^2)")+ggtitle("Forecast of Sea Ice Extent")

autoplot(forc1,include=mn.hor)+xlab("Year")+ylab("Sea Ice Extent (km^2)")+ggtitle("Forecast of Sea Ice Extent")
```

Appendix C: Code for Graduate Exercise

```
ar.par=c(0.6)
min(Mod(polyroot(c(1,-ar.par))))

pred.isin=matrix(NaN,nrow=nsim)
p = 1
q = 0

nsim = 500
pred.isin=matrix(NaN,nrow=nsim)
means1 <- c()
n = seq(6, 101, 5)
for(j in 1:length(n)){
for (i in 1:nsim){
  x=arima.sim(list(ar = ar.par),sd = sqrt(1),n = n[j])
  xpres=x[n[j]]
  x=x[1:(n[j]-1)]
  mod=arima(x,order=c(p,0,q),include.mean = FALSE,method="ML")
  forc=forecast(mod,h=1)
  pred.isin[i]=(forc$lower[2] <= xpres & forc$upper[2] >= xpres )
}
  means1[j] <- mean(pred.isin)
}



ar.par=c(0.6,0.2)
ma.par=c(0.8,-0.1)
```

```
p=length(ar.par)
q=length(ma.par)
c(min(Mod(polyroot(c(1,-ar.par)))),
  min(Mod(polyroot(c(1,ma.par)))))

nsim=100
pred.isin=matrix(NaN,nrow=nsim)
means2 <- c()
for(j in 1:length(n)){
for (i in 1:nsim){
  x=arima.sim(list(ar = ar.par,ma=ma.par),sd = sqrt(1),n = n[j])
  xpres=x[n[j]]
  x=x[1:(n[j]-1)]
  mod=arima(x,order=c(p,0,q),include.mean = FALSE,method="ML")
  forc=forecast(mod,h=1)
  pred.isin[i]=(forc$lower[2] <= xpres & forc$upper[2] >= xpres )
}
  means2[j] <- mean(pred.isin)
}


ggplot(data = NULL, aes(x=n, y=means1))+
  geom_point()+
  geom_line()

ggplot(data = NULL, aes(x=n, y=means2))+
  geom_point()+
  geom_line()
```

References

National Aeronautics and Space Administration. (2008, September 16). *Monitoring Sea Ice*. Earth Observatory. https://earthobservatory.nasa.gov/features/SeaIce/page2.php

National Oceanic and Atmospheric Administration. (2021, February 26). *How does sea ice affect global climate?* National Ocean Service. https://oceanservice.noaa.gov/facts/sea-ice-climate.html