ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 8 - Due date 03/27/23

Maggie O'Shea

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the project open the first thing you will do is change "Student Name" on line 3 with your name. Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A08_Sp22.Rmd"). Submit this pdf using Sakai.

Set up

Some packages needed for this assignment: `forecast`,`tseries`,`smooth`. Do not forget to load them before running your script, since they are NOT default packages.

```
#Load/install required package here
library(forecast)
library(tseries)
library(smooth)
library(stringr)
library(dplyr)
library(tidyr)
library(lubridate)
library(ggplot2)
library(forecast)
#library(Kendall)
library(tseries)
#library(outliers)
library(tidyverse)
library(smooth)
library(kableExtra)
```

Importing and processing the data set

Consider the data from the file "inflowtimeseries.txt". The data corresponds to the monthly inflow in m^3/s for some hydro power plants in Brazil. You will only use the last column of the data set which represents

one hydro plant in the Amazon river basin. The data span the period from January 1931 to August 2011 and is provided by the Brazilian ISO.

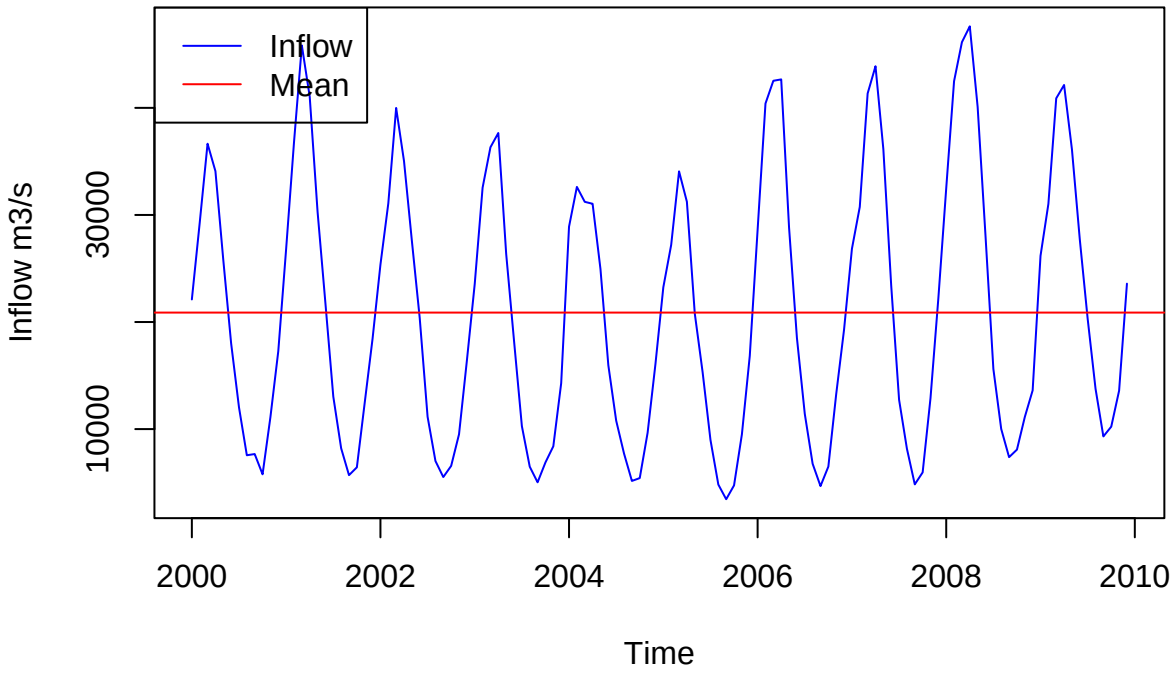
For all parts of the assignment prepare the data set such that the model consider only the data from January 2000 up to December 2009. Leave the year 2010 of data (January 2010 to December 2010) for the out-of-sample analysis. Do **NOT** use data fro 2010 and 2011 for model fitting. You will only use it to compute forecast accuracy of your model.

Part I: Preparing the data sets

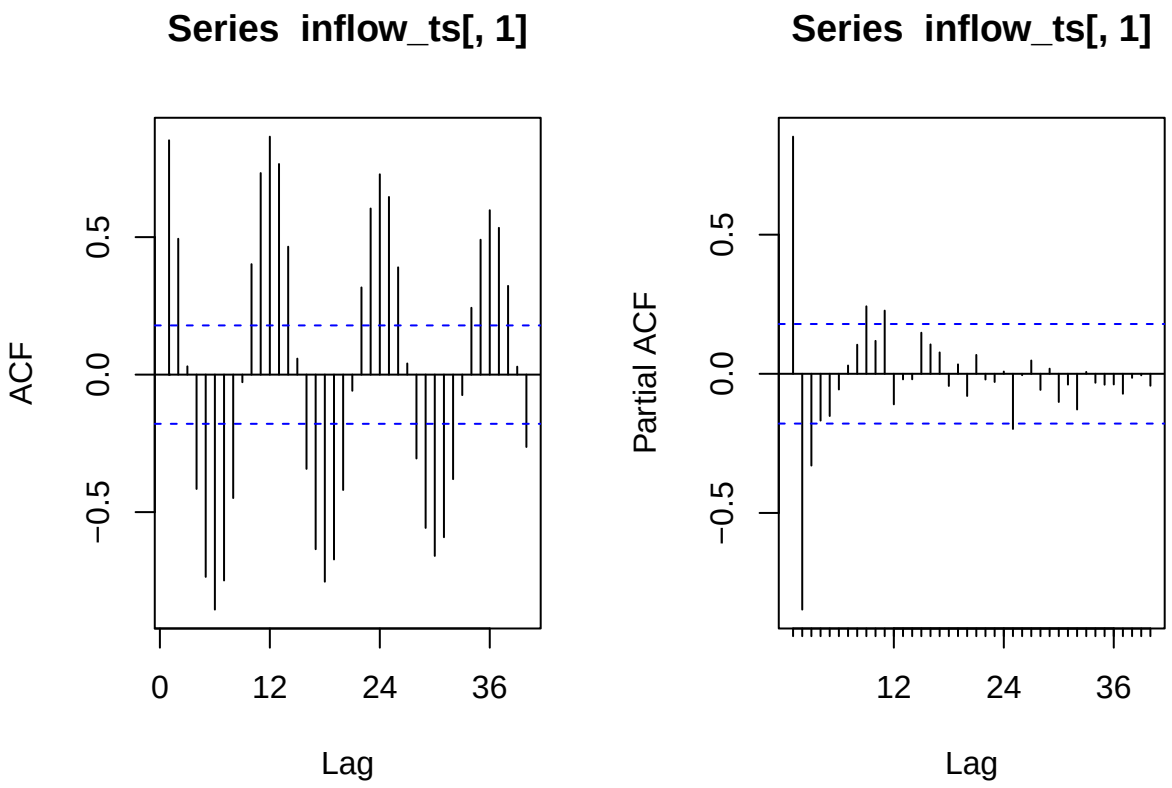
Q1

Read the file into a data frame. Prepare your time series data vector such that observations start in January 2000 and end in December 2009. Make you sure you specify the **start=** and **frequency=** arguments. Plot the time series over time, ACF and PACF.

```
inflow <- read.table("./Data/inflowtimeseries.txt", header=FALSE, sep = "\t")%>%
  mutate("Month" = substr(V1, 1, 8) )%>%
  select(Month, V15)%>%
  rename("Inflow" = V15)%>%
  separate(Month, into = c("Month", "Year"), sep = " (?=[^ ]+$)")%>%
  filter(Year>1999, Year<2010)%>%
  mutate(Date = my(paste0(Month,"-",Year)))%>%
  select(Inflow, Date)

inflow_ts <- ts(inflow, start=c(2000,1),
                          frequency=12)

#Time Series Plots
plot(inflow_ts[,1],type="l",col="blue",
     ylab="Inflow m3/s",xlab="Time",
     main="Monthly Inflow from Amazon River Basin Hydroplant 2000-2009")
abline(h=mean(inflow_ts[,1]),col="red")
legend("topleft",legend=c("Inflow","Mean"),
       lty=c("solid","solid"),col=c("blue","red"))
```



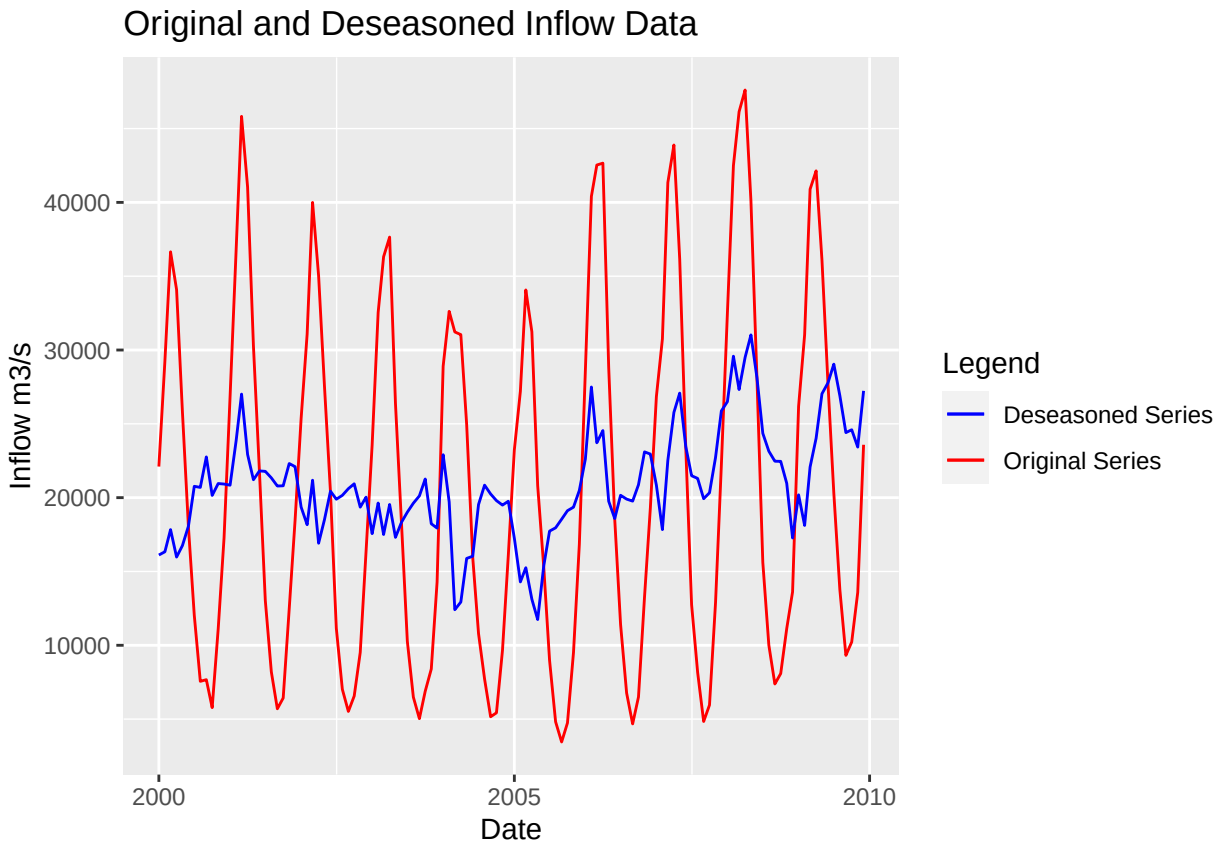

```
#ACF and PACF plots
par(mfrow=c(1,2))

ACF_Plot <- Acf(inflow_ts[,1], lag = 40, plot = TRUE)
PACF_Plot <- Pacf(inflow_ts[,1], lag = 40)
```



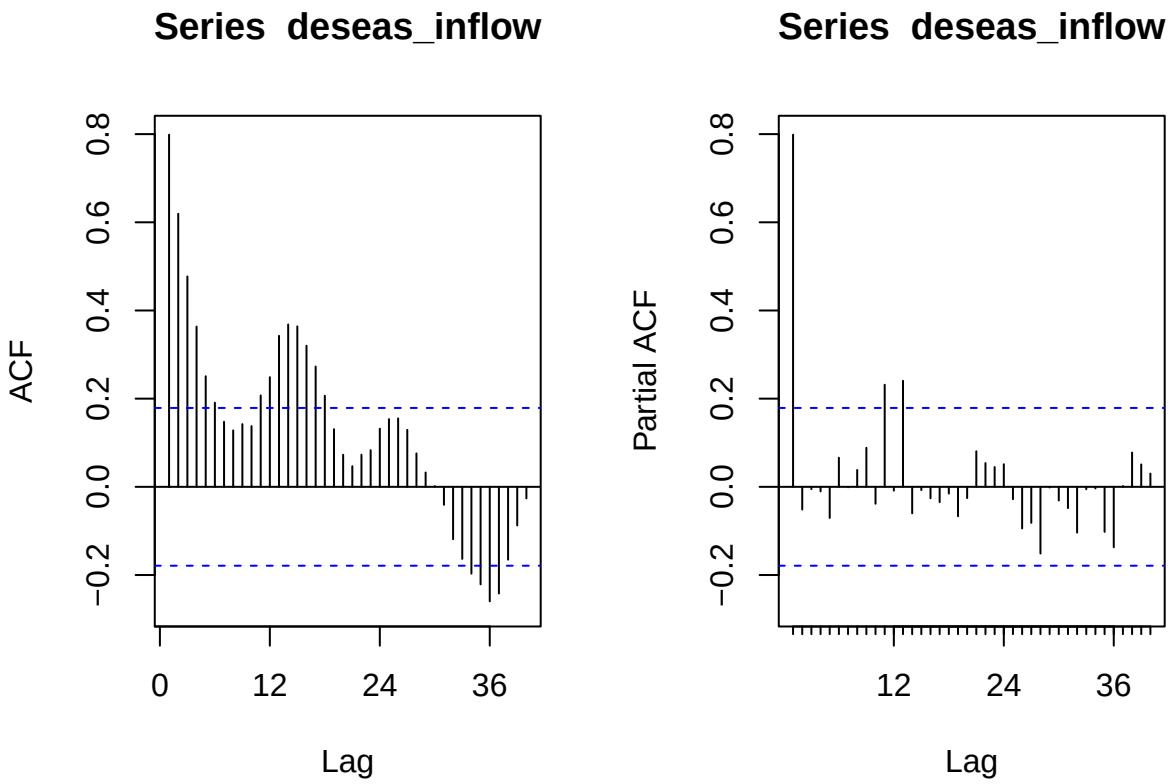
Q2

Using the *decompose*() or *stl*() and the *seasadj*() functions create a series without the seasonal component, i.e., a deseasonalized inflow series. Plot the deseasonalized series and original series together using ggplot, make sure your plot includes a legend. Plot ACF and PACF for the deaseasonalized series. Compare with the plots obtained in Q1.

```
decompose_inflow <- decompose(inflow_ts[,'Inflow'])
#plot(decompose_inflow)
deseas_inflow <- seasadj(decompose_inflow)

ggplot(inflow, aes(x=Date, y=Inflow))+
  geom_line(aes(color = "Original Series"))+
  geom_line(aes(x=Date, y=deseas_inflow, color = "Deseasoned Series"))+
  labs(x = "Date",
         y = "Inflow m3/s",
         color = "Legend",
       title = "Original and Deseasoned Inflow Data")+
  scale_color_manual(values = c("blue","red"))
```




```
par(mfrow=c(1,2))
ACF_Plot_deseas <- Acf(deseas_inflow, lag = 40, plot = TRUE)
PACF_Plot_deseas <- Pacf(deseas_inflow, lag = 40)
```



Q2 Answer: The ACF with deseasoned data still appears to have seasonality. Though the original ACF moves from positive to negative with each season which doesn't happen in the deseasoned ACF, there is still an increase and decrease occurring at lag 12, 24, etc. The declining trend, however, in the ACF is more visible in the deseasoned ACF. The PACFs look more similar, however, rather than 2 high correlations at 0 and 1 this is only at lag 0 in the deseasoned PACF. The correlations are also much smaller in the deseasoned PACF with the exception of lag 11 and 13.

Part II: Forecasting with ARIMA models and its variations

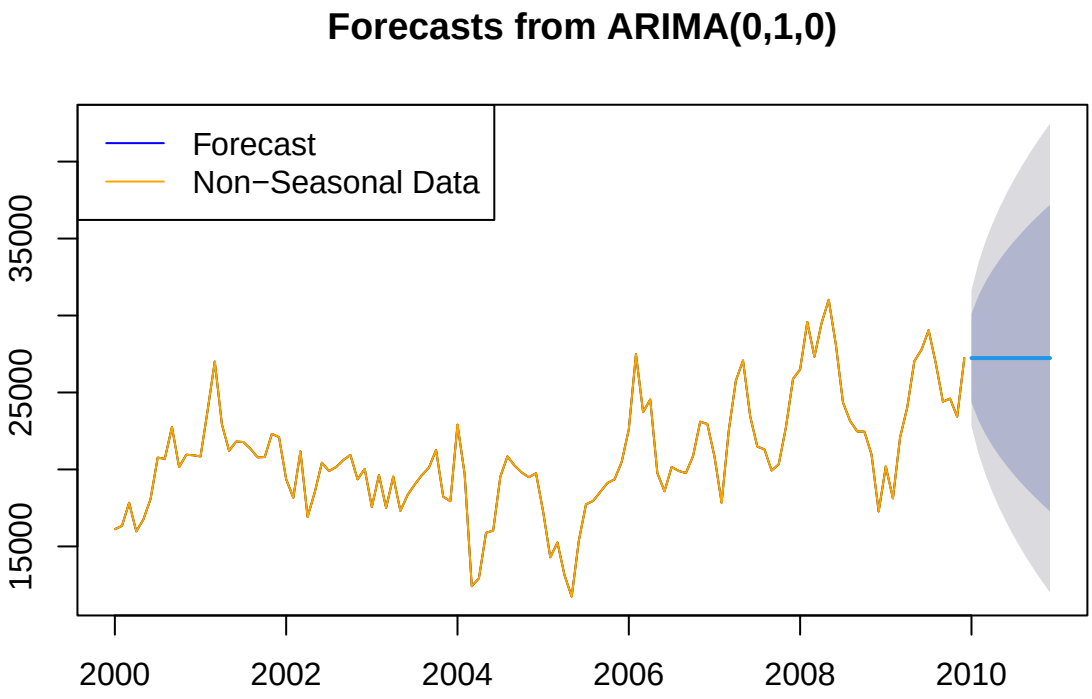
Q3

Fit a non-seasonal ARIMA(p, d, q) model using the auto.arima() function to the non-seasonal data. Forecast 12 months ahead of time using the $forecast()$ function. Plot your forecasting results and **further include on the plot the last year of non-seasonal data to compare with forecasted values** (similar to the plot on the lesson file for M10).

```
#ARIMA on deseasonal data
ARIMA_autofit <- auto.arima(deseas_inflow, max.D = 0, max.P = 0, max.Q = 0)
print(ARIMA_autofit)
```

```
## Series: deseas_inflow
## ARIMA(0,1,0)
##
## sigma^2 = 5031380:  log likelihood = -1087.01
## AIC=2176.02   AICc=2176.05   BIC=2178.8
```

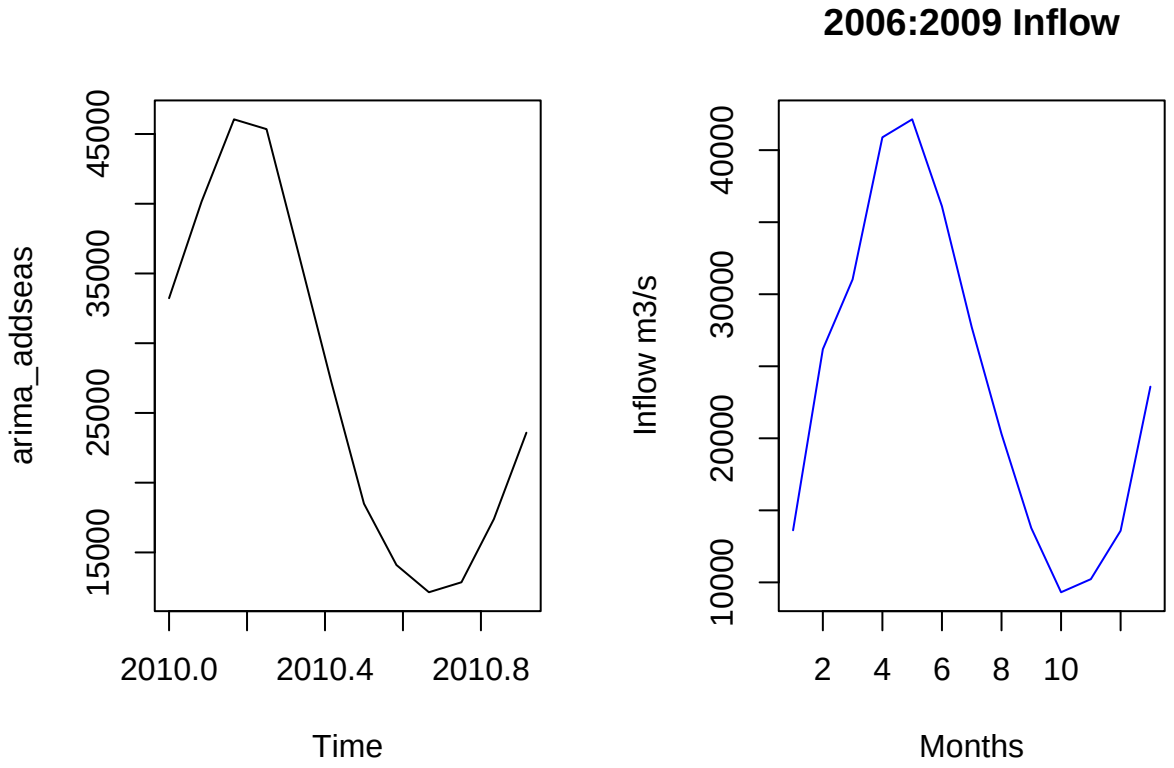
```
ARIMA_forecast <- forecast(object = ARIMA_autofit, h = 12)
plot(ARIMA_forecast)
lines(deseas_inflow, col="orange")
legend("topleft",legend=c("Forecast","Non-Seasonal Data"),
       lty=c("solid","solid"),col=c("blue","orange"))
```



Q4

Put the seasonality back on your forecasted values and compare with the original seasonal data values. *Hint* : One way to do it is by summing the last year of the seasonal component from your decompose object to the forecasted series.

```
#Add seasonal component
oneyear <- decompose_inflow$seasonal[1:12]
arima_addseas <- ARIMA_forecast$mean+oneyear

# Compare with original seasonal data
par(mfrow=c(1,2))
plot(arima_addseas)
plot(inflow_ts[108:120,1],type="l",col="blue",
     ylab="Inflow m3/s",xlab="Months",
     main="2006:2009 Inflow")
```



> Answer: When comparing these, they look very similar. The forecast looks to have more clean transitions season-to-season which is unsurprising given the modeling done to obtain this plot as opposed to the real observations shown in the second plot.

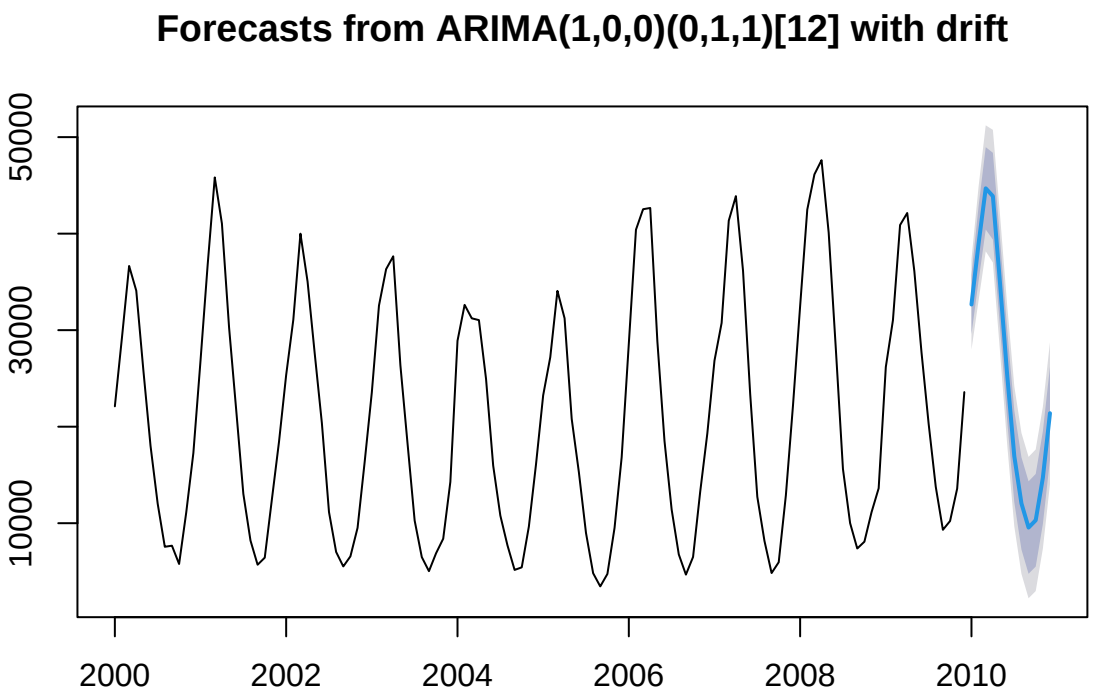
Q5

Repeat Q3 for the original data, but now fit a seasonal ARIMA$(p,d,q)x(P,D,Q)_{12}$ also using the auto.arima().

```
SARIMA_autofit <- auto.arima(inflow_ts[,1])
print(SARIMA_autofit)
```

```
## Series: inflow_ts[, 1]
## ARIMA(1,0,0)(0,1,1)[12] with drift
##
## Coefficients:
##          ar1     sma1    drift
##       0.7739  -0.8724  54.7963
## s.e.  0.0614   0.1767  25.8402
##
## sigma^2 = 5566545:  log likelihood = -999.07
## AIC=2006.14   AICc=2006.52   BIC=2016.86
```

```
SARIMA_forecast <- forecast(object = SARIMA_autofit, h = 12)
plot(SARIMA_forecast)
```

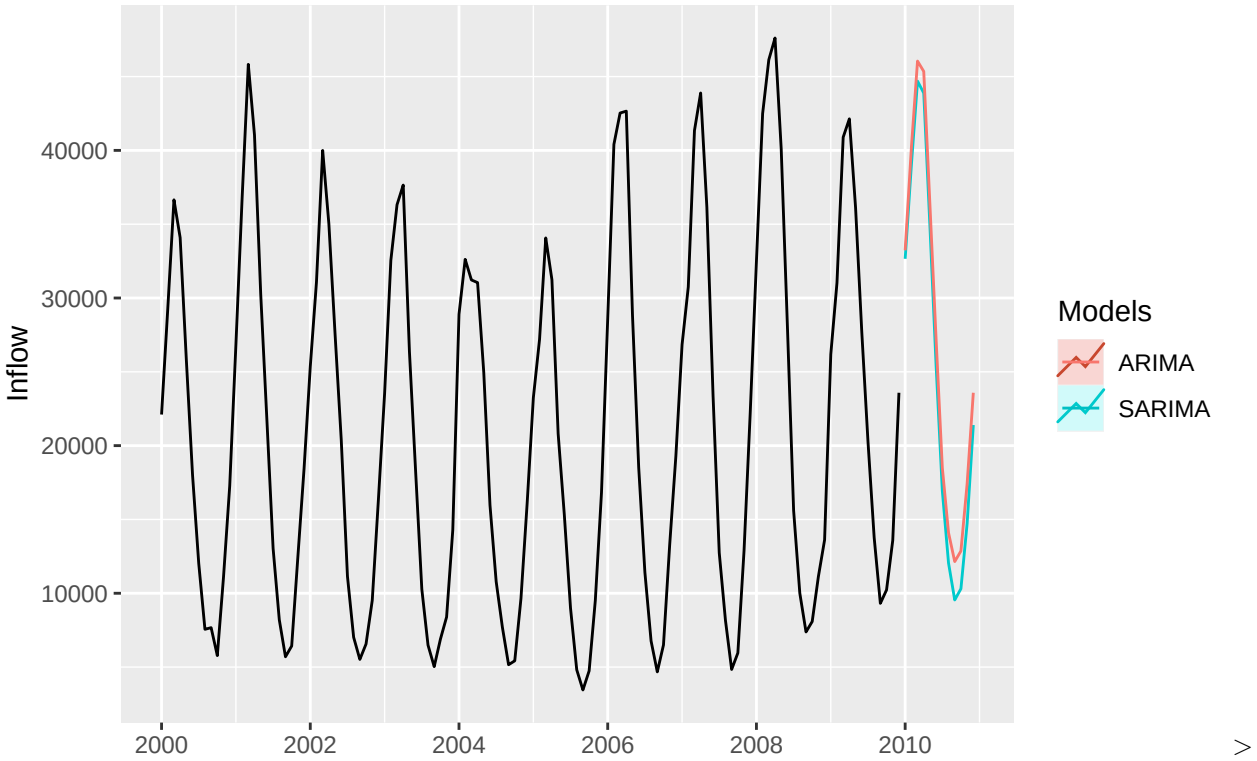


Q6

Compare the plots from Q4 and Q5 using the autoplot() function.

```
autoplot(inflow_ts[,1]) +
    autolayer(SARIMA_forecast,series="SARIMA",PI=FALSE) +
    autolayer(arima_addseas,series="ARIMA",PI=FALSE) +
    ylab("Inflow") +
    xlab("") +
    labs(col="Models")
```

```
## Warning: Ignoring unknown parameters: PI
```



Answer: The SARIMA and the ARIMA (with seasonality added back in) look very similar, though the SARIMA predicts lower inflow in the high season and less in the low season as well.

Part III: Forecasting with Other Models

Q7

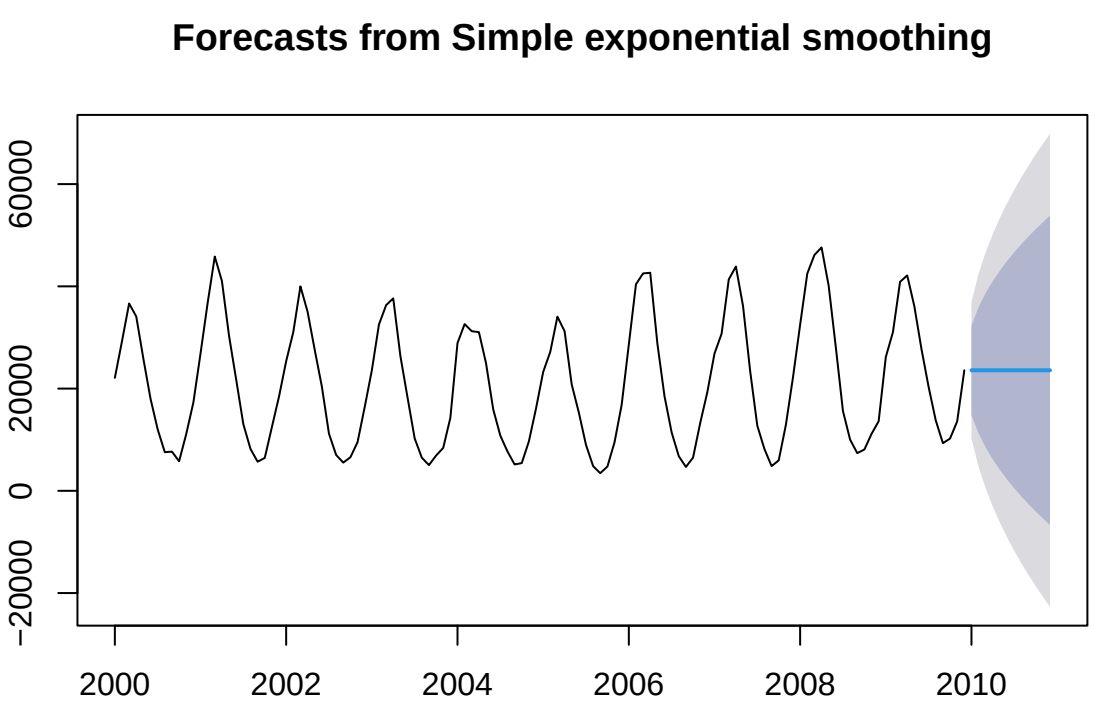
Fit an exponential smooth model to the original time series using the function *ses*() from package `forecast`. Note that this function automatically do the forecast. Do not forget to set the arguments: silent=FALSE and holdout=FALSE, so that the plot is produced and the forecast is for the year of 2010.

```
#Simple exponential smoothing on original data
inflow_seas_fit <- ses(y = inflow_ts[,1], h = 12, holdout = FALSE, silent = FALSE)
summary(inflow_seas_fit)
```

```
##
## Forecast method: Simple exponential smoothing
##
## Model Information:
## Simple exponential smoothing
##
## Call:
##  ses(y = inflow_ts[, 1], h = 12, holdout = FALSE, silent = FALSE)
##
##   Smoothing parameters:
##     alpha = 0.9999
##
##   Initial states:
##     l = 21885.2463
```

```
##
##   sigma:  6813.76
##
##      AIC     AICc      BIC
## 2696.890 2697.097 2705.252
##
## Error measures:
##                     ME    RMSE      MAE      MPE     MAPE     MASE      ACF1
## Training set 14.14103 6756.74 5787.733 -7.09394 34.05251 1.758853 0.7154525
##
## Forecasts:
##          Point Forecast      Lo 80    Hi 80      Lo 95    Hi 95
## Jan 2010          23582 14849.8153 32314.19  10227.276 36936.72
## Feb 2010          23582 11233.4433 35930.56   4696.512 42467.49
## Mar 2010          23582  8458.4205 38705.58    452.481 46711.52
## Apr 2010          23582  6118.9402 41045.06  -3125.445 50289.45
## May 2010          23582  4057.8032 43106.20  -6277.682 53441.68
## Jun 2010          23582  2194.3852 44969.62  -9127.534 56291.53
## Jul 2010          23582   480.7907 46683.21 -11748.251 58912.25
## Aug 2010          23582 -1114.1874 48278.19 -14187.559 61351.56
## Sep 2010          23582 -2612.2260 49776.23 -16478.612 63642.61
## Oct 2010          23582 -4029.1079 51193.11 -18645.546 65809.55
## Nov 2010          23582 -5376.7480 52540.75 -20706.583 67870.58
## Dec 2010          23582 -6664.4029 53828.40 -22675.882 69839.88
```

```
plot(inflow_seas_fit)
```



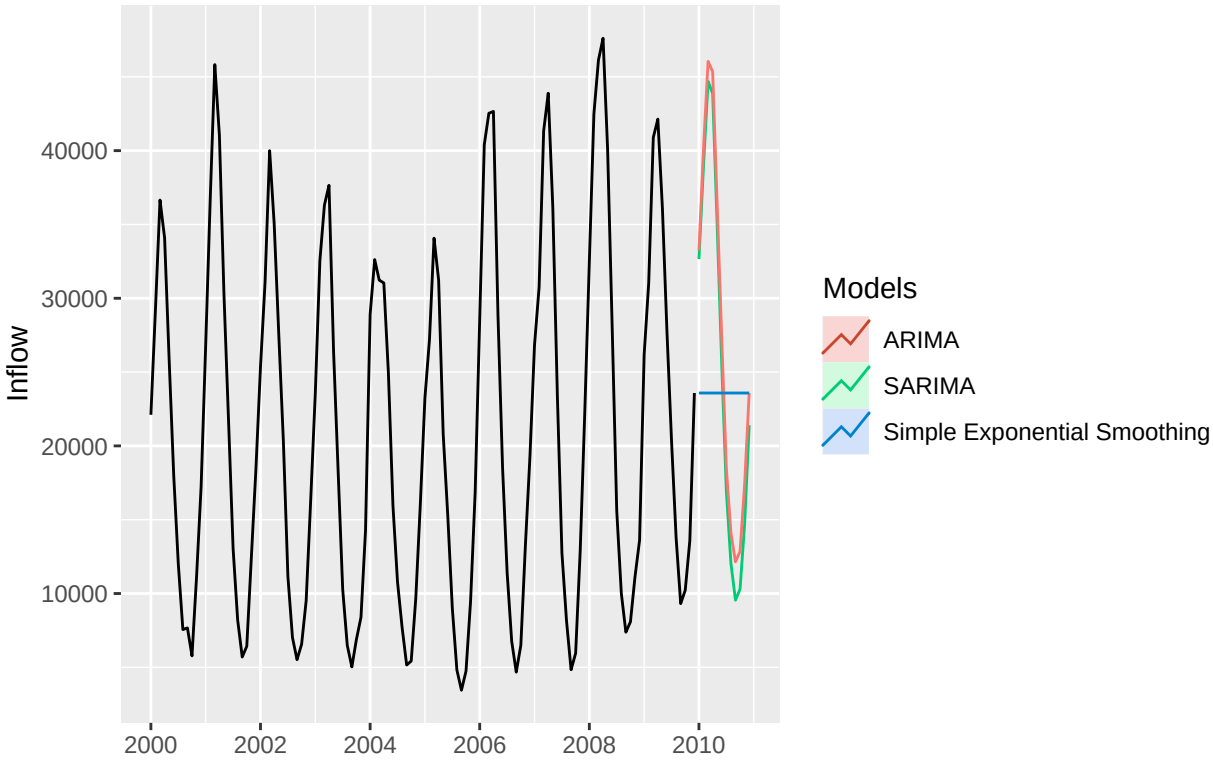

```
#Best alpha = 0.9999
```

Part IV: Checking Forecast Accuracy

Q8

Make one plot with the complete original seasonal historical data (Jan 2000 to Dec 2010). Now add the forecasts from each of the developed models in parts Q4, Q5, Q7 and Q8. You can do it using the autoplot() combined with autolayer(). If everything is correct in terms of time line, the forecasted lines should appear only in the final year. If you decide to use ggplot() you will need to create a data frame with all the series will need to plot. Remember to use a different color for each model and add a legend in the end to tell which forecast lines corresponds to each model.

```
ts_inflow_full <- ts(
  inflow_ts[,"Inflow"],
  start=c(year(inflow$Date[1]),month(inflow$Date[1])),
  frequency=12)

autoplot(ts_inflow_full) +
    autolayer(SARIMA_forecast,series="SARIMA",PI=FALSE) +
    autolayer(arima_addseas,series="ARIMA",PI=FALSE) +
    autolayer(inflow_seas_fit,series="Simple Exponential Smoothing",PI=FALSE) +
    ylab("Inflow") +
    xlab("") +
    labs(col="Models")
```



Q9

From the plot in Q9 which model or model(s) are leading to the better forecasts? Explain your answer. Hint: Think about which models are doing a better job forecasting the high and low inflow months for example.

Table 1: Forecast Accuracy for Seasonal Data

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
SARIMA	-80.63461	2206.972	1611.593	-2.60900	8.99301	0.48975	0.07464
ARIMA	92.80632	2233.708	1750.409	-0.20111	8.64088	0.53194	-0.01906
SEAS	14.14103	6756.740	5787.733	-7.09394	34.05251	1.75885	0.71545

Answer: Without running diagnostic tests and only based on the plots, it is clear that the SARIMA and ARIMA are stronger models given their ability to capture the important elements of seasonality. Between these two, however, it's difficult to determine which one is a better forecast especially because they are very similar. The SARIMA may offer more conservative estimates, forecasting lower inflow in the low season and lower inflow in the high season which may be useful if such a conservative approach is appealing to the hydropower plant.

Q10

Now compute the following forecast metrics we learned in class: RMSE and MAPE, for all the models you plotted in part Q9. You can do this by hand since your have forecasted and observed values for the year of 2010. Or you can use R function *accuracy*() from package "forecast" to do it. Build and a table with the results and highlight the model with the lowest MAPE. Does the lowest MAPE corresponds match your answer for part Q10?

```
SARIMA_scores <- accuracy(SARIMA_forecast)
ARIMA_scores <- accuracy(ARIMA_forecast)
SEAS_scores <- accuracy(inflow_seas_fit)
#create data frame
scores <- as.data.frame(rbind(SARIMA_scores, ARIMA_scores, SEAS_scores))
row.names(scores) <- c("SARIMA", "ARIMA","SEAS")

#choose model with lowest RMSE
best_model_index <- which.min(scores[,"RMSE"])
cat("The best model by RMSE is:", row.names(scores[best_model_index,]))
```

```
## The best model by RMSE is: SARIMA
```

```
kbl(scores,
      caption = "Forecast Accuracy for Seasonal Data",
      digits = array(5,ncol(scores))) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  #highlight model with lowest RMSE
  kable_styling(latex_options="striped", stripe_index = which.min(scores[,"RMSE"]))
```

The best model by RMSE was the SARIMA model which does match my answer from Q10.