Stat 153 Project: Forecasting Mauna Loa CO_2 Levels

Group Member: Jiayao Li; SID: 3035120687

Group Member: Kara Jia; SID: 3031907464

Group Member: Tiffany Shi; SID: 3036000668

1 Introduction

For the purpose of measuring the condition of atmosphere and global environment, Mauna Loa Observatory in Hawaii has been collecting and measuring data of CO_2 in the atmosphere since March, 1958. The dataset that we will be using in this report is monthly mean carbon dioxide from March, 1958 to October, 2021. In this report, we are going to analyze and model this dataset, in order to forecast a time series for CO_2 measurement. We will set up two signal models to pursue stationarity and model the plausibly-stationary noise from the two signal models. After getting our final model, we will predict CO_2 measurements for the next 10 month. Based on our predictions, we can provide suggestions on possible future actions that have impact on CO_2 in the atmosphere. One thing to notice is that the Mauna Loa data that we are using here are being obtained at an altitude of 3400m in the northern subtropics, and may not be the same as the globally averaged CO_2 concentration at the surface, so there exists limitations in the conclusion.

2 Data Description

This report will analyze and forecast a time series of CO_2 based on the dataset that contains 764 records of CO_2 in the atmosphere. For each record, the dataset contains the following 8 variables:

Variable Name	Description	Type
`Pos`	Year of CO_2 measurement record	Integer
`month`	Month of CO_2 measurement record	Integer
`decimal.date`	Date of CO_2 measurement record in decimal format	Double
`average`	Monthly mean CO_2 measurement (ppm), which are corrected to center of month based on average seasonal cycle	Double
`interpolated`	Interpolated monthly mean CO_2 measurement (ppm) due to missing months	Double
`trend`	Missing days in the month, negative numbers for NA	Double
`ndays`	Standard deviation of days, negative numbers for NA	Double
`X`	Uncertainty of monthly mean CO_2 measurement (ppm), negative numbers for NA	Double

3 Exploratory Data Analysis

Before constructing time series models to forecast the CO_2 measurement, we first need to better understand our dataset. Our dataset starts recording MM CO_2 from March 1958 to October 2021. To make the dataset easier for forecasting, we will only keep data from the years where all twelve months have records for. We will plot the whole time series of monthly mean CO_2 (`$CO_2$`) from Jan. 1959 to Dec. 2020 (Figure 1(a)), and the partial time series from Jan. 2018 to Dec. 2020 (Figure 1(b)), in order to visualize any existing seasonal, trend, and random noises.

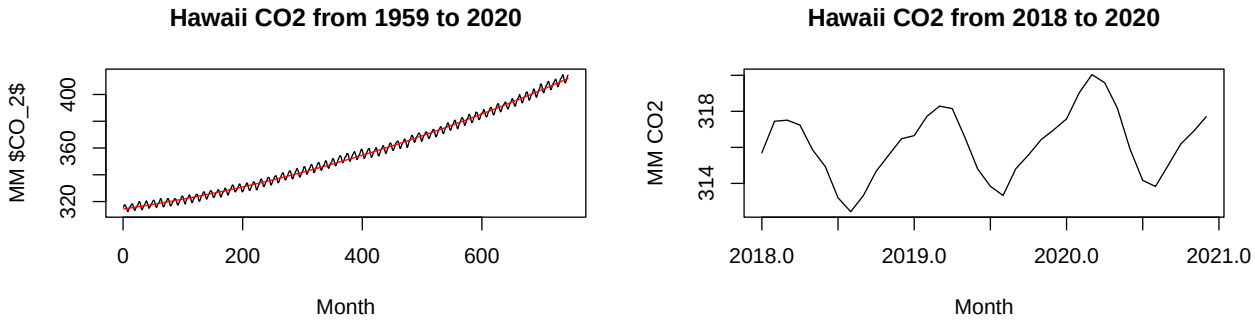


Figure 1: Time Series of Carbon Dioxide Measurements: (a) 1959 − 2020; (b) 2018 − 2020

Using `tso` function, we found no outliers in our time series. Our data does not follow a linear model so a polynomial trendline (square root) was added (Figure 1(a)). There is also an obvious seasonal pattern in the data. A key point is how to remove the seasonality, so next we will plot the ACF and PACF of monthly mean CO_2.

From Figure 2, we can observe that the original data points are all largely correlated with each other but not with itself at a different time point. To pursue stationarity, our first step is to remove the polynomial trendline by applying first differences. The PACF (Figure 2(b)) tells that an appropriate parametric model such as some linear model would help explain the correlation among data.

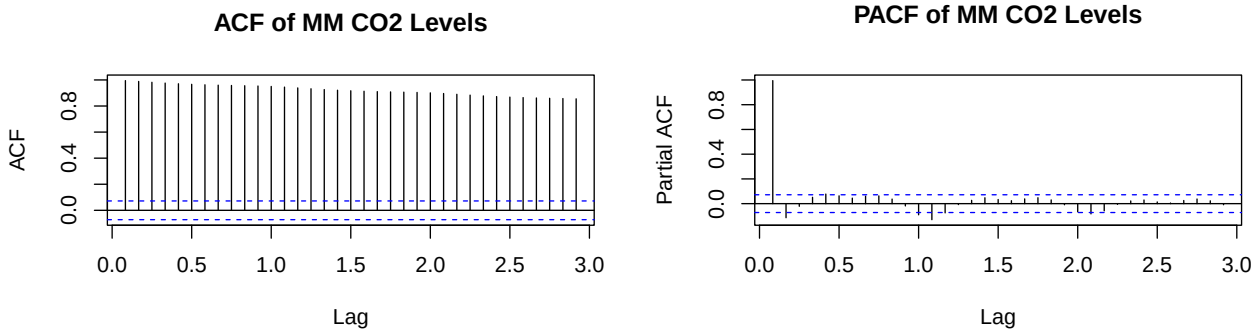


Figure 2: Autocorrelation of MM CO_2 (a) ACF (b) PACF

After first differencing, the our ACF plot confirmed the presence of seasonality of the first differenced series for the non-parametric model. We then wrote a function to test how many differences we should choose with lag of 12 (annual) for seasonal differencing. Using Augmented Dickey-Fuller Test, differences of 4 returned the smallest, significant test statistic.

In Figure 1(a), we have shown how our parametric model fits the time series. We can see that the model fits the trend well but not with the seasonal. First we looked at residuals from the parametric model; they exhibit heteroscedasticity in Figure 3(a) which means we need to apply first differencing. We can also see seasonality from the ACF plot of the residuals in Figure 3(b). We will then apply seasonal differencing with lag 12, 5th difference (same as the CO_2 MM time series) to the parametric model first differenced residuals in Figure 3.

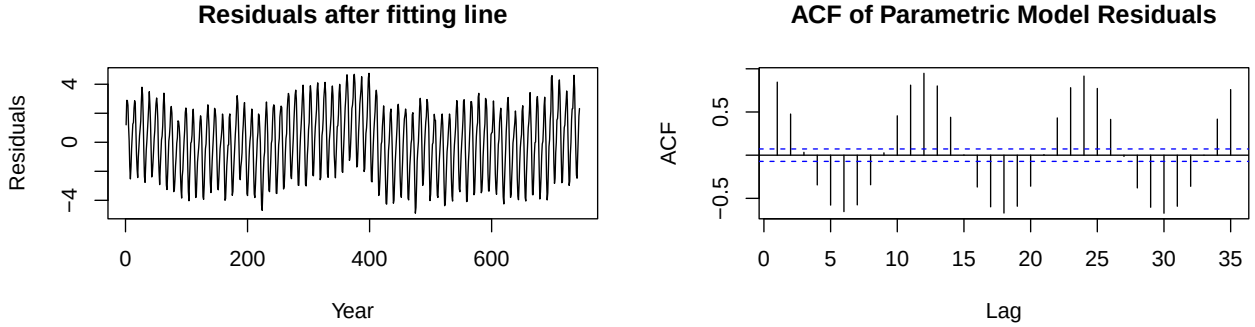


Figure 3: Parametric Model (a) Parametric Model Residuals (b) ACF of Residuals

After pursuing stationarity, we can create models such as SARIMA to forecast future MM CO_2 levels. We can see residuals so far are not stationary yet, so we will keep complementing our model to remove the remaining noises.

4 Models

We will build parametric and non-parametric signal models.

4.1 Parametric Signal Model

First, we consider polynomial model (Figure 1(a)) for removing the trendline using a second degree polynomial The trendline of the parametric model that was used to generate the residuals is as follows:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 I(t^2)$$

Because there is seasonality in the residuals, we will combine the parametric model with SARIMA.

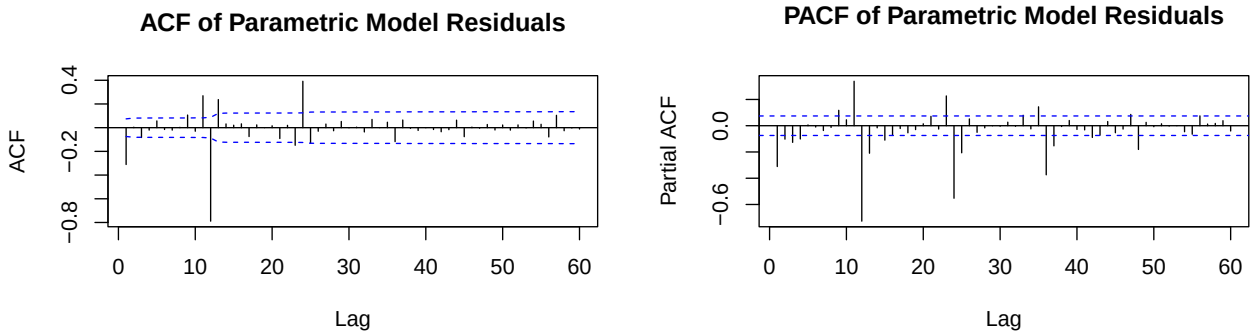


Figure 4: Autocorrelation of Parametric Model Residuals (a) ACF (b) PACF

4.1.1 Polynomial trend + SARIMA(1,0,3)x(1,1,1)[12]

In Figure 4, seasonality is present in both ACF and PACF plots. With a seasonal lag of 12, we observe that the PACF in Figure 4(b) are tailing off seasonally. This means an SARMA model with $Q > 0$. We observe that both the ACF and PACF are tailing off. Using Bartlett's standard errors to plot the confidence intervals, the cutoff for ACF is around $q = 4$ but it not very clear. The PACF has is tailing but it is more obvious that $p = 1$. Since the parameter q is not really obvious, we perform model diagnostics and use AIC, BIC, AICc to find the best model fit, resulting in $q = 3$ and a final model of combining polynomial trend + SARIMA(1,0,3)(1,1,1)[12].

4.1.2 Polynomial trend + SARIMA(1,0,1)x(0,1,1)[12]

Back to the PACF seasonal tailing, we decided on a SARMA model with $Q > 0$ and $P = 0$. Like the previous model, p is kept at 1. With the help of model diagnostics, we decided our q parameter to be 1.

4.2 Non-parametric Signal Model

Next, we fit SARIMA models to the dataset using `sarima` model diagnostics results on our stationary time series.

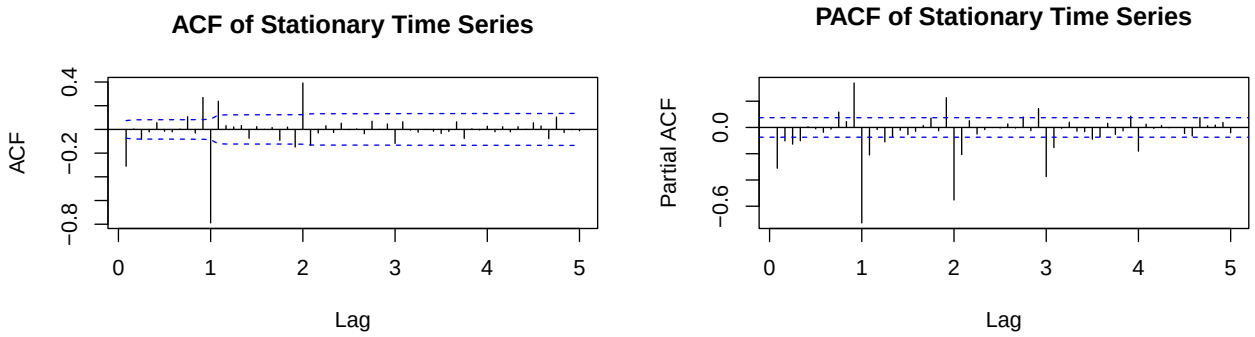


Figure 5: SARIMA Diagnostics (a) ACF (b) PACF

4.3 SARIMA(0,1,1)x(0,1,1)[12]

PACF is tailing off seasonally while ACF is not (Figure 5). We see the seasonal lag 12 for the same reason as in section 4.1. This means a SARMA model with $Q > 0$. For nonseasonal, ACF has a cutoff at q=1; meanwhile, the PACF is tailing off. This concludes a SARIMA(0,1,1)x(0,1,1)[12] model.

4.4 SARIMA(1,1,1)x(0,1,1)[12]

In an another attempt, we used model diagnostics like Q-Q plots to help decide a second non-parametric model. The seasonal lag and parameters are kept the same as the previous model. However this time we tried an SARMA model with $p > 0$ and $q > 0$. With model diagnostics, it achieved stationarity like the previous model.

5 Model Comparison and Selection

After getting four potential models, we first compare their AIC, AICc, and BIC values to have a rough idea about how they fit the dataset.

Table 2: Information Criterion for 4 Models

Model	AIC	AICc	BIC
Polynomial trend + SARIMA(1,0,3)x(1,1,1)[12]	0.5076137	0.5078250	0.5578408
Polynomial trend + SARIMA(1,0,1)x(0,1,1)[12]	0.5052884	0.5054013	0.5429588
SARIMA(0,1,1)x(0,1,1)[12]	0.5260875	0.5261100	0.5449428
SARIMA1(1,1,1)x(0,1,1)[12]	0.5237318	0.5237770	0.5488723

According to Table 2, we can see that the Polynomial trend + SARIMA$(1,0,1)x(0,1,1)_{12}$ model fits the best to our data among the 4 considered models. Overall, the 4 models do not have a very large difference on these 3 criterion. Similarly, Q-Q plot of the residuals for the models show that they are white noise. From the Ljung-Box-Pierce test, we can reject the hypothesis that the ARMA models are causal and invertible.

Now we would like to conduct the cross validation among the 4 models to diagnose their performance on forecasting, and their forecasting performances will be compared through their root mean squared prediction error. We would like to choose the model with the lowest cross validation score to be our final model.

Table 3: RMSE Scores for 4 Models

Model	RMSE_Scores
Polynomial trend + SARIMA(1,0,3)x(1,1,1)[12]	1.7280094
Polynomial trend + SARIMA(1,0,1)x(0,1,1)[12]	1.7312561

Model	RMSE_Scores
SARIMA(0,1,1)x(0,1,1)[12]	0.3855822
SARIMA(1,1,1)x(0,1,1)[12]	0.3899079

According to Table 3, we can see that SARIMA$(0,1,1) \times (0,1,1)_{12}$ model has the lowest (i.e., best) cross validation (RMSE) score. Remember that in Table 2, we notice that Polynomial trend + SARIMA$(1,0,1) \times (0,1,1)_{12}$ model fits best to our data, but it does not have the best prediction performance.

6 Results

In order to forecast CO_2 measurement in the next 10 months, we will use a non-parametric model.

Let X_t be a stationary process defined by SARIMA$(0,1,1) \times (0,1,1)_{12}$, and Wt be white noise with mean zero and finite variance.

$$\nabla_S^1 \nabla^1 X_t = (1 - \Theta_1 B^S)(1 - \theta_1 B) W_t$$

6.1 Estimation of Model Parameters

The following are the estimates of parameters in our final forecast model.

Table 4: Coefficients for Model Parameters

	ar1	ma1	sma1
	0.2011	-0.5548	-0.8623
s.e.	0.0969	0.0831	0.0189

6.2 Prediction

In addition to using SARIMA models for forecasting, we also applied smoothing methods such as simple exponential smoothing and kernel smoothing to stationary MM CO_2 time series (Figure 6). Simple exponential smoothing is a method that emphasizes more recent data points of the time series, which can provide a different perspective than SARIMA.

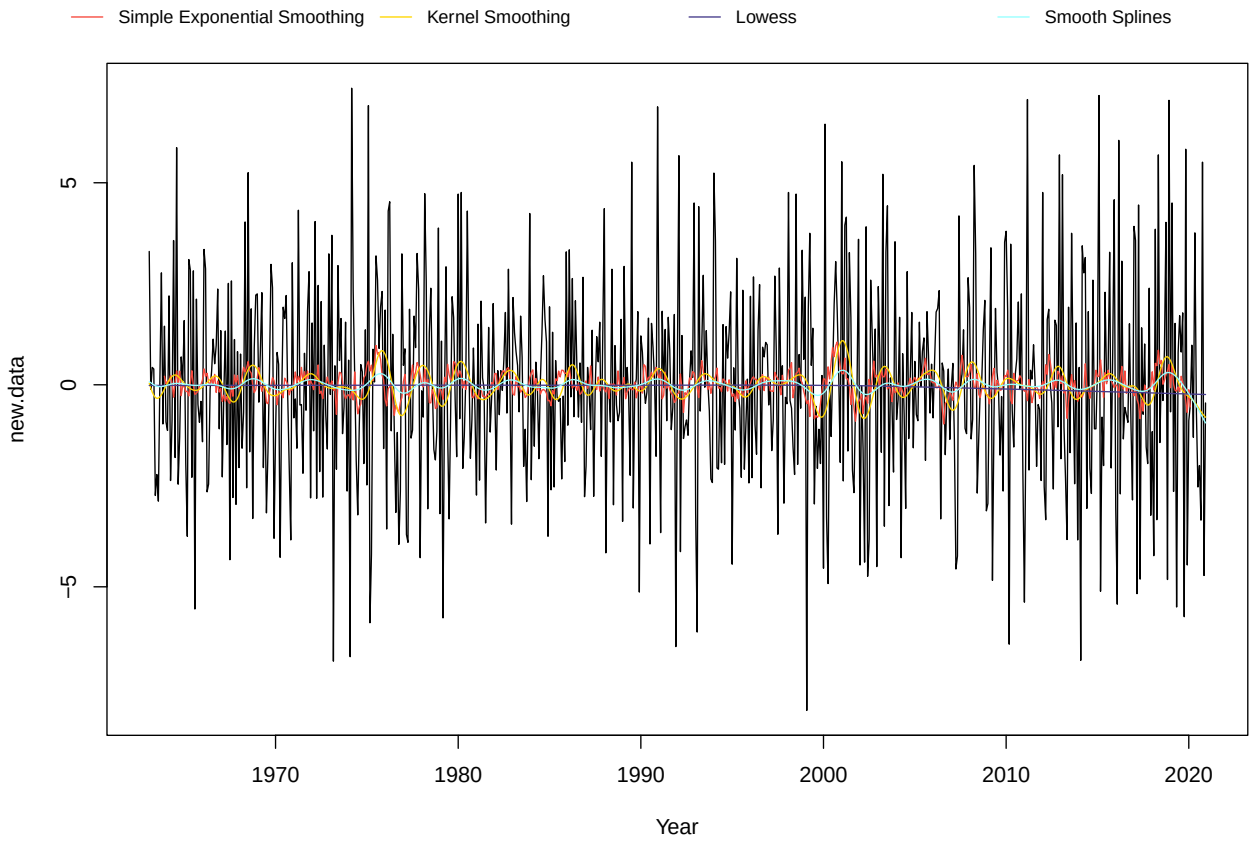


Figure 6: Smoothing Methods on Stationary MM CO2 Series

The next plot below (Figure 7) shows the forecasted values of CO_2 measurements for the next 10 months using SARIMA.

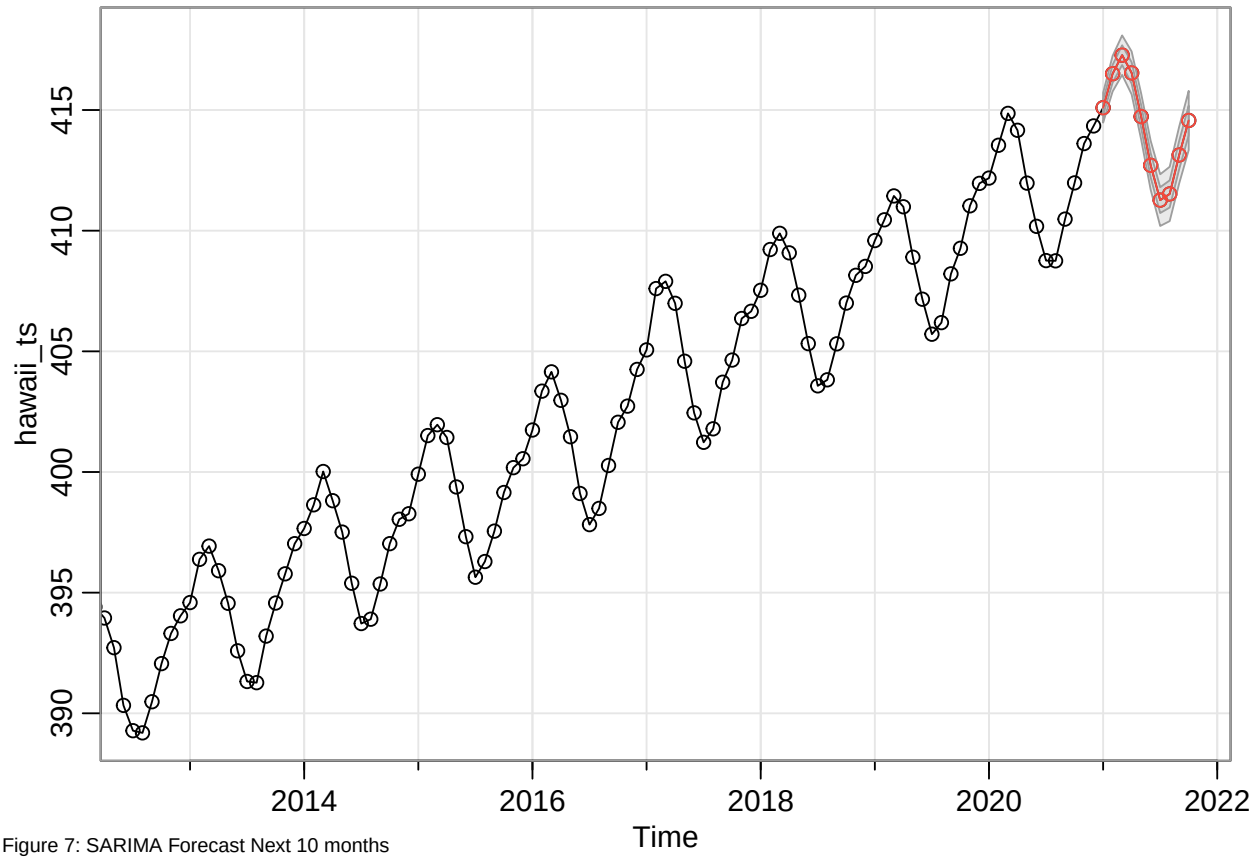


Figure 7: SARIMA Forecast Next 10 months

The results of simple exponential smoothing forecast are shown below (Figure 8). Figure 8(a) shows the result of using `ses` function to forecast the next 10 months. `holt` function was used for forecasting the next 5 years since `ses` does not perform well with data with a long-term trend (Figure 8(b))

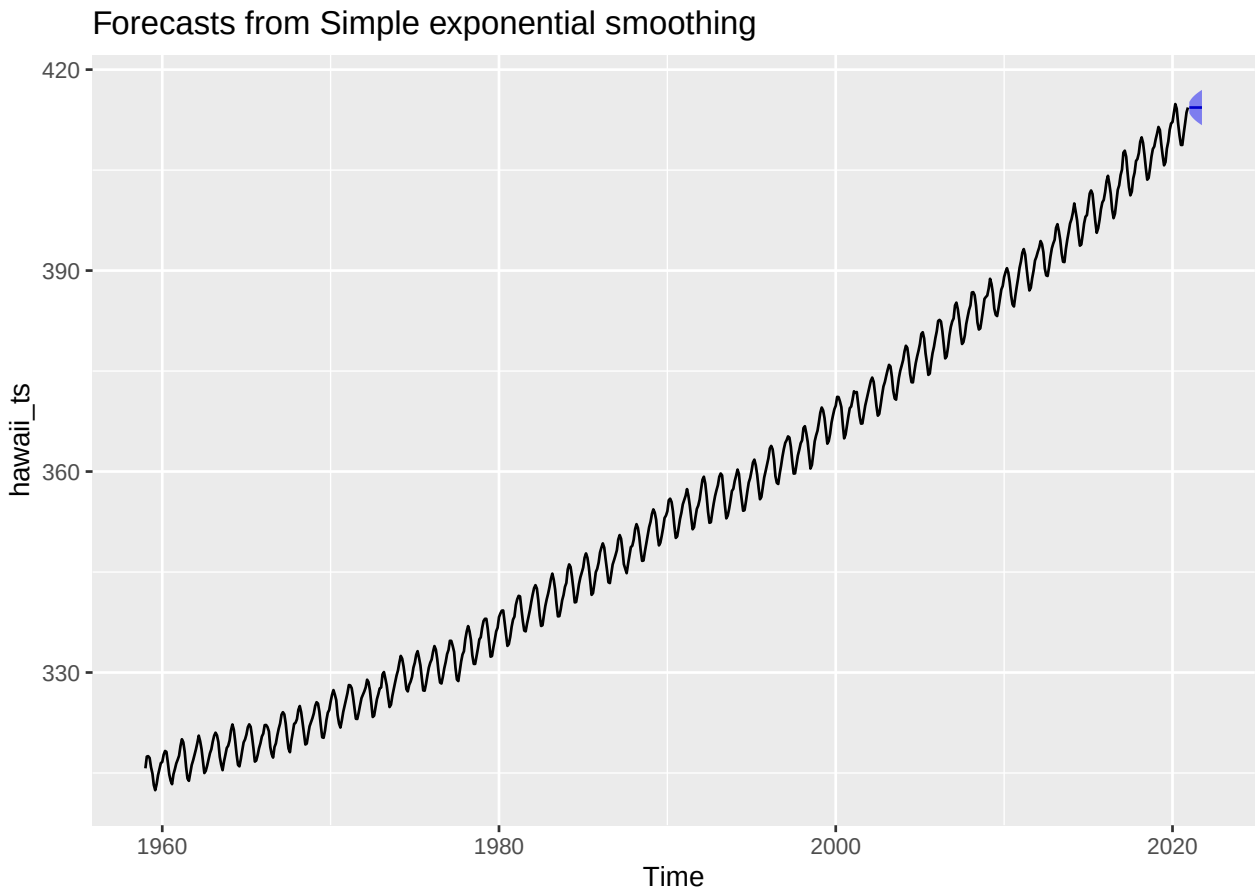
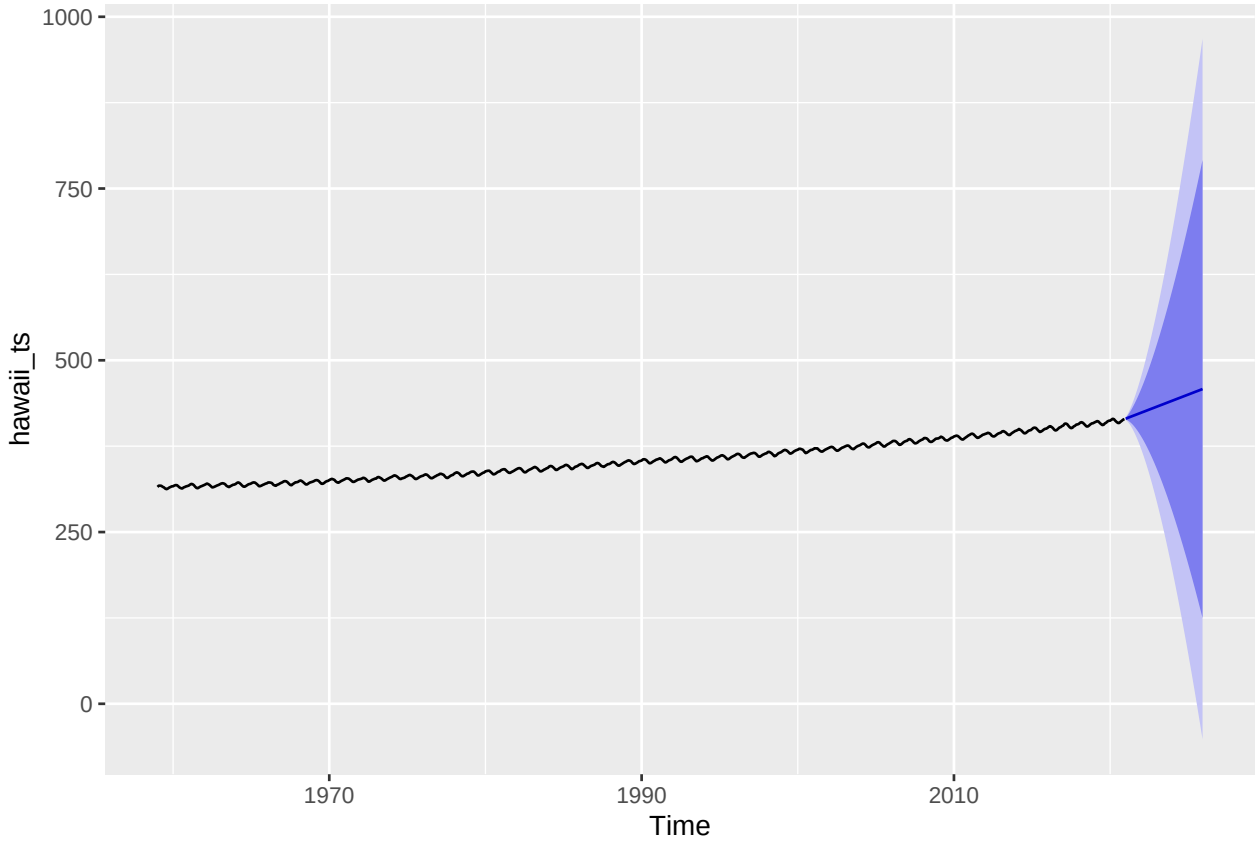

Forecasts from Simple exponential smoothing
420
390
360
330
hawaii_ts
1960
1980
2000
2020
Time

Figure 8(b): Simple Exponential Smoothing Forecast Next 5 Years



7 Conclusion

Our forecast model suggests that in the next 10 months, the value for CO_2 measurements will follow the existing trend and seasonality. It will keep going up to some peak, then following the seasonal cycle, go down for some months. Based on our forecast for the next 5 years, there is also an increasing trend in CO_2 levels. The main takeaway from the results of our analysis is the overall trend for CO_2 is going up. Our result may offer some preliminary evidence on climate change in Mauna Loa, Hawaii. On a larger scale, combined with current climate research, there may be increasing need for policy changes concerning climate change.

8 Appendix

```
library(astsa)
library(dplyr)
library(tsoutliers)
library(stats)
library(TSA)
library(forecast)
library(astsa)
library(tseries)
library(gridExtra)

# Load data
dat <- read.csv("/Users/kyjurou/Downloads/co2_mm_mlo.csv")

# Exploratory Data Analysis
hawaii_co2 <- dat %>%
    select(decimal.date, average)
hawaii_ts <- ts(hawaii_co2$average, frequency = 12,
    start = c(1959, 1), end = c(2020, 12))

s18_20 <- ts(hawaii_co2$average, frequency = 12,
    start = c(2018, 1), end = c(2020, 12))
```

```
par(mfrow = c(1, 2))
num.hawaii <- as.numeric(hawaii_ts)
tme <- 1:length(hawaii_ts)
lin.mod <- lm(num.hawaii ~ 1 + tme + poly(tme,
    2))

plot(num.hawaii, main = "Hawaii CO2 from 1959 to 2020",
    xlab = "Month", ylab = "MM $CO_2$", type = "l")
points(tme, lin.mod$fitted, type = "l", col = "red")

plot.ts(s18_20, main = "Hawaii $CO_2$ from 2018 to 2020",
    xlab = "Month", ylab = "MM $CO_2$")

mtext(side = 1, line = -1, adj = 0, cex = 1,
    text = "Figure 1: Time Series of Carbon Dioxide Measurements: (a) 1959 - 2020; (b) 2018 - 2020",
    outer = T)
```

```
# Non-parametric model

adf.test(hawaii_ts)

## First differencing
ndiffs(hawaii_ts)
d1 <- diff(hawaii_ts)
acf(d1)
mean(d1)

## Seasonal differencing, etc
adf.find <- function(series) {
    adf <- 0
```

```
    for (i in 1:11) {
        adf[i] <- adf.test(diff(diff(series,
            12, i), 1))$statistic
    }
    s.diff <- which.min(adf)
    sprintf("seasonal diff: %.0f", s.diff)
}
adf.find(d1)
d2 <- diff(diff(d1, 12, 3))
acf2(d2)
mean(d2)

# Parametric model

adf.find(resid)  #has seasonality
station.resid <- diff(resid, 12, 4)
acf2(station.resid, max.lag = 60)
station.resid <- diff(station.resid)
acf2(station.resid)
mean(station.resid)
sd(station.resid)
```

```
par(mfrow = c(1, 2))
acf(hawaii_ts, lag.max = 35, plot = T, main = "ACF of MM CO2 Levels",
    na.action = na.pass)  #seasonality is still present
pacf(hawaii_ts, lag.max = 35, plot = T, main = "PACF of MM CO2 Levels",
    na.action = na.pass)  #seasonality is still present

mtext(side = 1, line = -1, adj = 0, cex = 1,
    text = "Figure 2: Autocorrelation of MM $CO_2$ (a) ACF (b) PACF",
    outer = T)
```

```
par(mfrow = c(1, 2))

plot(tme, lin.mod$residuals, type = "l",
    xlab = "Year", ylab = "Residuals", main = "Residuals after fitting line")
acf(resid, lag.max = 35, plot = T, main = "ACF of Parametric Model Residuals",
    na.action = na.pass)  #seasonality is still present

mtext(side = 1, line = -1, adj = 0, cex = 1,
    text = "Figure 3: Parametric Model (a)  Parametric Model Residuals (b) ACF of Residuals ",
    outer = T)
```

```
par(mfrow = c(1, 2))
acf(station.resid, lag.max = 60, plot = T,
    main = "ACF of Parametric Model Residuals",
    na.action = na.pass, ci.type = "ma")  #seasonality is still present
pacf(station.resid, lag.max = 60, plot = T,
    main = "PACF of Parametric Model Residuals",
    na.action = na.pass)

mtext(side = 1, line = -1, adj = 0, cex = 1,
    text = "Figure 4: Autocorrelation of Parametric Model Residuals (a) ACF (b) PACF ",
    outer = T)
```

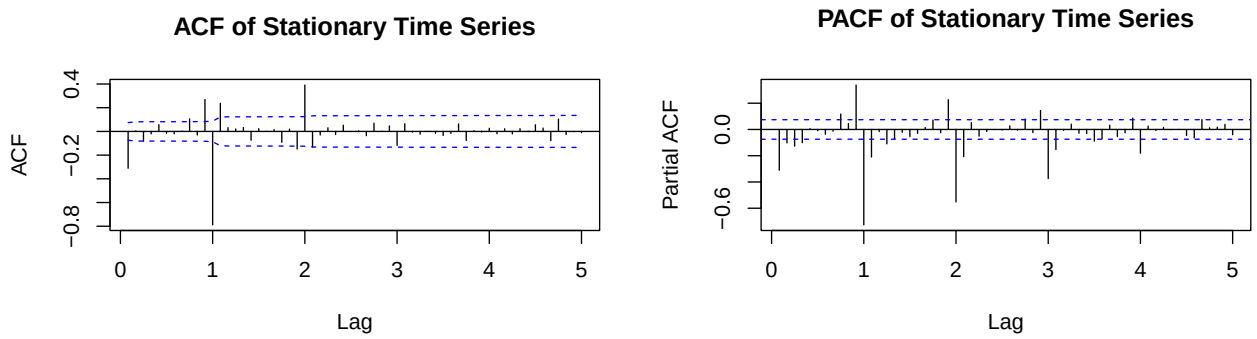


Figure 5: SARIMA Diagnostics (a) ACF (b) PACF

```
# Residuals of parametric model +
# SARIMA

acf2(station.resid, max.lag = 60)
auto.arima(resid)

sarima(resid, p = 1, d = 0, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic .509,bic .547
sarima(resid, p = 1, d = 0, q = 0, P = 1,
    D = 1, Q = 1, S = 12)  #aic .591,bic .623
sarima(resid, p = 2, d = 0, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic .5078,bic .546 #good
sarima(resid, p = 1, d = 0, q = 3, P = 1,
    D = 1, Q = 1, S = 12)  #aic .5077,bic .5579

sarima(resid, p = 1, d = 0, q = 1, P = 1,
    D = 1, Q = 2, S = 12)  #aic .511,bic .555
sarima(resid, p = 0, d = 0, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic 1.25
sarima(resid, p = 1, d = 1, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic .516,bic .548

sarima(resid, p = 1, d = 1, q = 1, P = 1,
    D = 1, Q = 2, S = 12)  #aic .519
sarima(resid, p = 0, d = 1, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic .521,bic .547
sarima(resid, p = 2, d = 0, q = 1, P = 0,
    D = 1, Q = 1, S = 12)  #aic .505,bic .542


# Non-parametric model
d2 <- diff(diff(hawaii_ts, 12, 4))
acf2(d2)
auto.arima(hawaii_ts)

sarima(hawaii_ts, p = 1, d = 0, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic. 536, bic .573

sarima(hawaii_ts, p = 0, d = 1, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic .52879, bic .5539
sarima(hawaii_ts, p = 1, d = 1, q = 1, P = 1,
    D = 1, Q = 1, S = 12)  #aic .526, bic .558
```

```
sarima(hawaii_ts, p = 1, d = 1, q = 1, P = 1,
    D = 1, Q = 2, S = 12)  #aic .529

sarima(hawaii_ts, p = 1, d = 1, q = 1, P = 2,
    D = 1, Q = 1, S = 12)  #aic .5288, bic.566
sarima(hawaii_ts, p = 1, d = 1, q = 1, P = 0,
    D = 1, Q = 1, S = 12)  #aic .5237, bic.5488 #good
sarima(hawaii_ts, p = 0, d = 1, q = 1, P = 0,
    D = 1, Q = 1, S = 12)  #aic .526,bic .544   #good
sarima(hawaii_ts, p = 2, d = 1, q = 1, P = 0,
    D = 1, Q = 1, S = 12)  #aic .524, .54



M1 <- sarima(resid, p = 1, d = 0, q = 3,
    P = 1, D = 1, Q = 1, S = 12)
M2 <- sarima(resid, p = 2, d = 0, q = 1,
    P = 0, D = 1, Q = 1, S = 12)
M3 <- sarima(hawaii_ts, p = 0, d = 1, q = 1,
    P = 0, D = 1, Q = 1, S = 12)
M4 <- sarima(hawaii_ts, p = 1, d = 1, q = 1,
    P = 0, D = 1, Q = 1, S = 12)
```

```
# Cross-validation, forecast past 10
# years since 2021
rmse = matrix(NA, nrow = 11, ncol = 4)  # forecasting out 10 different times, with 4 models
for (i in 1:11) {
    ## Split train/test
    train.test.split.point = 504 + 12 * (i -
        1)  # last point of train
    train1 = hawaii_ts[1:train.test.split.point]
    test1 = hawaii_ts[(train.test.split.point +
        1):(train.test.split.point + 12)]
    model = lm(hawaii_ts ~ 1 + time(hawaii_ts) +
        sqrt(time(hawaii_ts)))
    resid1 = model$residuals
    train.resid = resid1[1:train.test.split.point]
    test.resid = resid1[(train.test.split.point +
        1):(train.test.split.point + 12)]
    ## Fit
    model1 = sarima.for(resid1, 12, 1, 0,
        3, 1, 1, 1, 12)
    model2 = sarima.for(resid1, 12, 2, 0,
        1, 0, 1, 1, 12)
    model3 = sarima.for(train1, 12, 0, 1,
        1, 0, 1, 1, 12)
    model4 = sarima.for(train1, 12, 1, 1,
        1, 0, 1, 1, 12)
    ## Test
    rmse[i, 1] = sqrt(mean((test.resid -
        model1$pred)^2))
    rmse[i, 2] = sqrt(mean((test.resid -
        model2$pred)^2))
    rmse[i, 3] = sqrt(mean((test1 - model3$pred)^2))
```

```
    rmse[i, 4] = sqrt(mean((test1 - model4$pred)^2))

}
rmse <- data.frame(rmse)
rownames(rmse) <- as.character(2011:2021)
colnames(rmse) <- c("Test RMSE: model1",
    "Test RMSE: model2", "Test RMSE: model3",
    "Test RMSE: model4")
# average test rmse over 10 years
cv1 = mean(rmse[, 1])
cv2 = mean(rmse[, 2])
cv3 = mean(rmse[, 3])
cv4 = mean(rmse[, 4])
```

```
arima(hawaii_ts, order = c(0, 1, 1), seasonal = list(order = c(0,
    1, 1), period = 12))
```

```
par(xpd=TRUE)
new.data <- d2
#Exponential smoothing
alpha <- .9
mt <- stats::filter(new.data, filter = c(0,alpha^((1:20)-1)*(1-alpha)),sides = 2,method='con')
plot(new.data, type = "l",xlab = "Year")
lines(mt, type = "l",col=2)
#Kernel smooth
lines(ksmooth(time(new.data), new.data, kernel = "normal", bandwidth = 1), col = "gold")
#Lowess
lines(lowess(new.data), col = "slateblue4")
#Smooth splines
lines(smooth.spline(time(new.data), new.data, spar = 0.5), col = "darkslategray1" )


legend(x = "top",horiz = TRUE, bty = "n", inset=c(-.3), -.1), legend=c("Simple Exponential Smoothing",

mtext(side=1, line=-1, adj=0, cex=1, text= "Figure 6: Smoothing Methods on Stationarity MM CO2 Series",
```

```
sarima.for(hawaii_ts, n.ahead = 10, p = 1,
    d = 1, q = 1, P = 1, D = 1, Q = 1, S = 12)
mtext(side = 1, line = -3, adj = 0, cex = 0.5,
    text = "Figure 7: SARIMA Forecast Next 10 months",
    outer = T)
```

```
# Simple Exponential Smoothing Forecast
ses.forecast <- ses(hawaii_ts, h = 10, level = 0.5,
    initial = "simple", "Figure 8(a): Simple Exponential Smoothing Forecast Next 10 Months")
autoplot(ses.forecast)
holt.fore <- holt(hawaii_ts, h = 60)
autoplot(holt.fore, main = "Figure 8(b): Simple Exponential Smoothing Forecast Next 5 Years")
```