Forecasting CO2 Emissions from the U.S Energy Industry with Renewable Energy's Rapid Growth and a Warming Earth

December 15, 2022

Introduction:

Over the past decade, climate change has become a problem that requires immediate action. Countries across the globe have committed to mitigating greenhouse gas emissions. While many nations have set aggressive emission reduction goals, taking action has become an obstacle. In November of 2022, COP 27 met in Egypt and determined that many countries seem to be failing their emission goals. This either means that the policies enacted to fight climate change were poorly implemented or nations inaccurately predicted their emission trajectories, leading them to assume that weaker policies will work. The United States has been implementing climate change reduction policies over the past two decades to respond to global warming. The energy production industry in the US accounts for ~25% of its annual emissions. Reducing greenhouse gas emissions in the energy industry is imperative to fighting a rise in global temperatures. While this paper does not explore the intricacies of environmental policies, it will explore different methods to predict future CO2 emissions in the energy industry in the United States. It is important to note that the forecasts will be calculated assuming that the US continues on its current trajectory and that no new shocks will cause significant disruption to CO2 output.

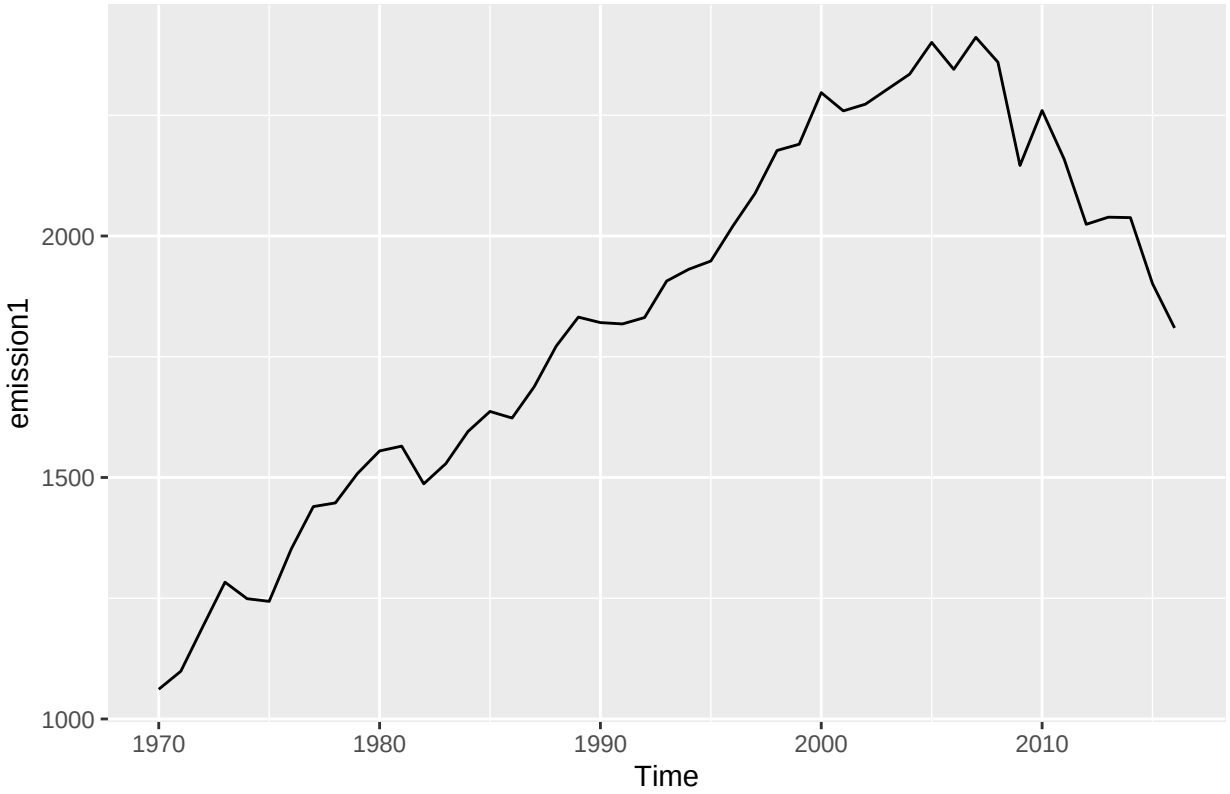
Literature:

Many studies predict future emission levels depending on various scenarios. One model by the US Energy Information Administration predicts that, by 2030, the US energy sector will emit 1,214 MMmt of CO2, a far cry from the net zero emissions the US committed to in the Paris Climate Accord. Interestingly the EIA predicts a major slowdown in the transition to renewables and a major slowdown regarding CO2 emissions from the energy produced by coal. . Natural gas, a cleaner option, is the only greenhouse gas source predicted to grow continuously. While the time series model created by the EIA has far more depth than this study, their projected slowdown in decreasing CO2 emissions in energy production will be challenged (EIA). Papers also look at emissions in the energy sector and use GDP as a proxy with VAR. It has been established at COP27 that GDP and emissions are no longer linked in advanced economies like the US. An environmental economics journal from 2018 found that in many European countries, emissions increase at a much slower pace than GDP; thus, decoupling is in effect (Mikayilov). More advanced economies will increasingly be able to grow at a desirable rate without emitting unwelcome greenhouse gasses in the process. Decoupling GDP and emissions may also be significant in developing nations as their industrialization is less emission-intensive than in previous decades. For CO2 output in energy production in the US, exploring emission forecasts without GDP as a variable has led to a more accurate projection. The main dataset used in the study comes from FRED and shows the annual emissions released by the US energy sector since 1970. While analyzing this dataset, it is important to keep in mind increasing annual energy production from renewable energy sources. The initial hypothesis for this study was that increasing the use of clean energy will have some effect on decreasing annual emissions over time. Another theory that motivated this research was that increasing global temperatures could catalyze the decarbonization of the energy industry. In other words, as the effects of global warming are felt, more nations and industries will change their energy procedures to reduce emissions and halt global warming.

Plotting Emissions:

```
# Plot emissions
em = em[,2]
emission.ts <- ts(data=em, start = 1970, end = 2020)
autoplot(emission.ts)
```


```
emission1 <- window(emission.ts, start = 1970, end = 2016)
emission2 <- window(emission.ts, start = 2017, end = 2020)
autoplot(emission1)
```



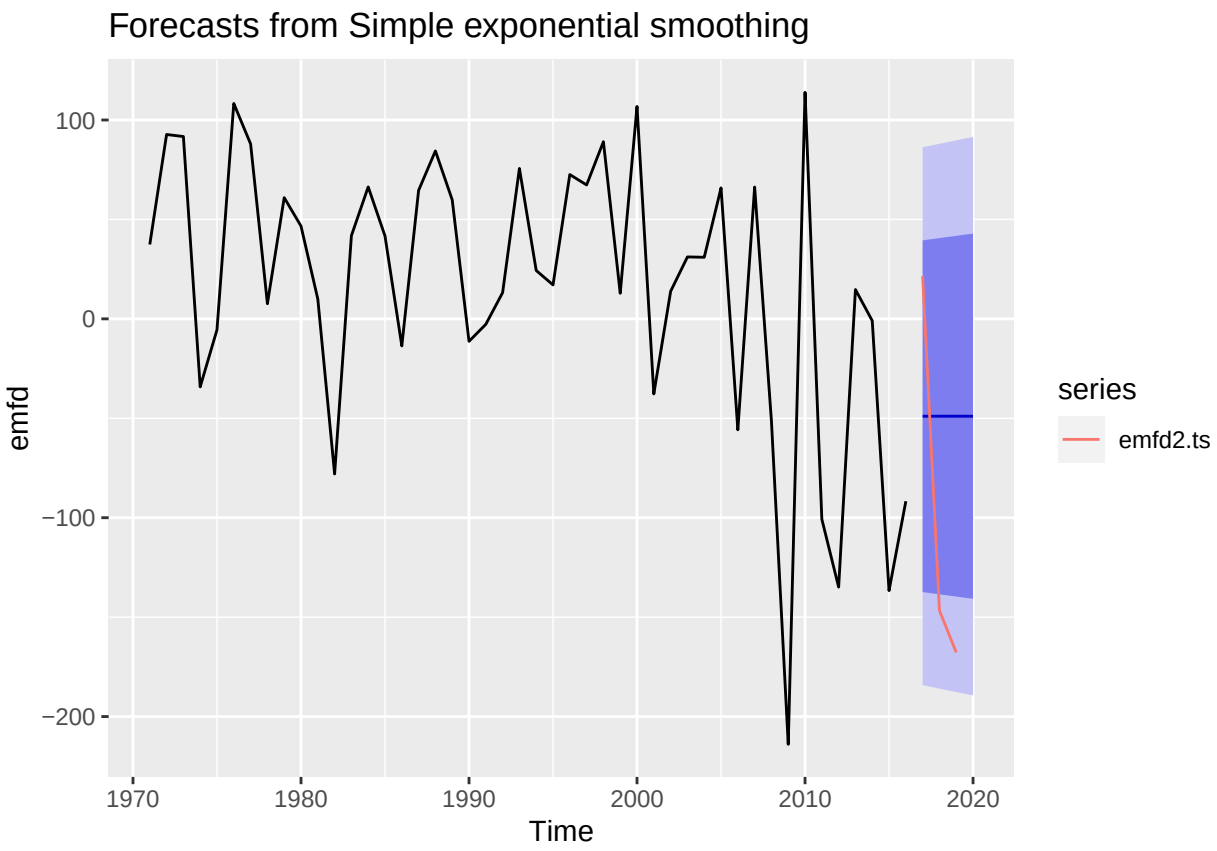
Differencing Data:

```
# Difference data
emfd <- diff(emission1)
autoplot(emfd)
```


```
emfd2 <- diff(emission2)
emfd2.ts <- ts(emfd2, start = 2017)
```

Simple Exponential Smoothing:

```
fit1 <- ses(emfd, h=4)
autoplot(fit1) + autolayer(emfd2.ts)
```



ARIMA Modeling:

```
fitaicc <- auto.arima(emission1, seasonal=FALSE, ic = c("aicc"), test = c("adf"))
fitbic <- auto.arima(emission1, seasonal=FALSE, ic=c("bic"))
summary(fitaicc)
```

```
## Series: emission1
## ARIMA(0,1,0) with drift
##
## Coefficients:
##         drift
##       16.2647
## s.e.  10.4403
##
## sigma^2 = 5125:  log likelihood = -261.23
## AIC=526.46   AICc=526.74   BIC=530.12
##
## Training set error measures:
##                      ME     RMSE      MAE      MPE     MAPE      MASE      ACF1
## Training set 0.02223776 70.05257 52.73163 0.158594 2.917847 0.9028256 0.1056568
```

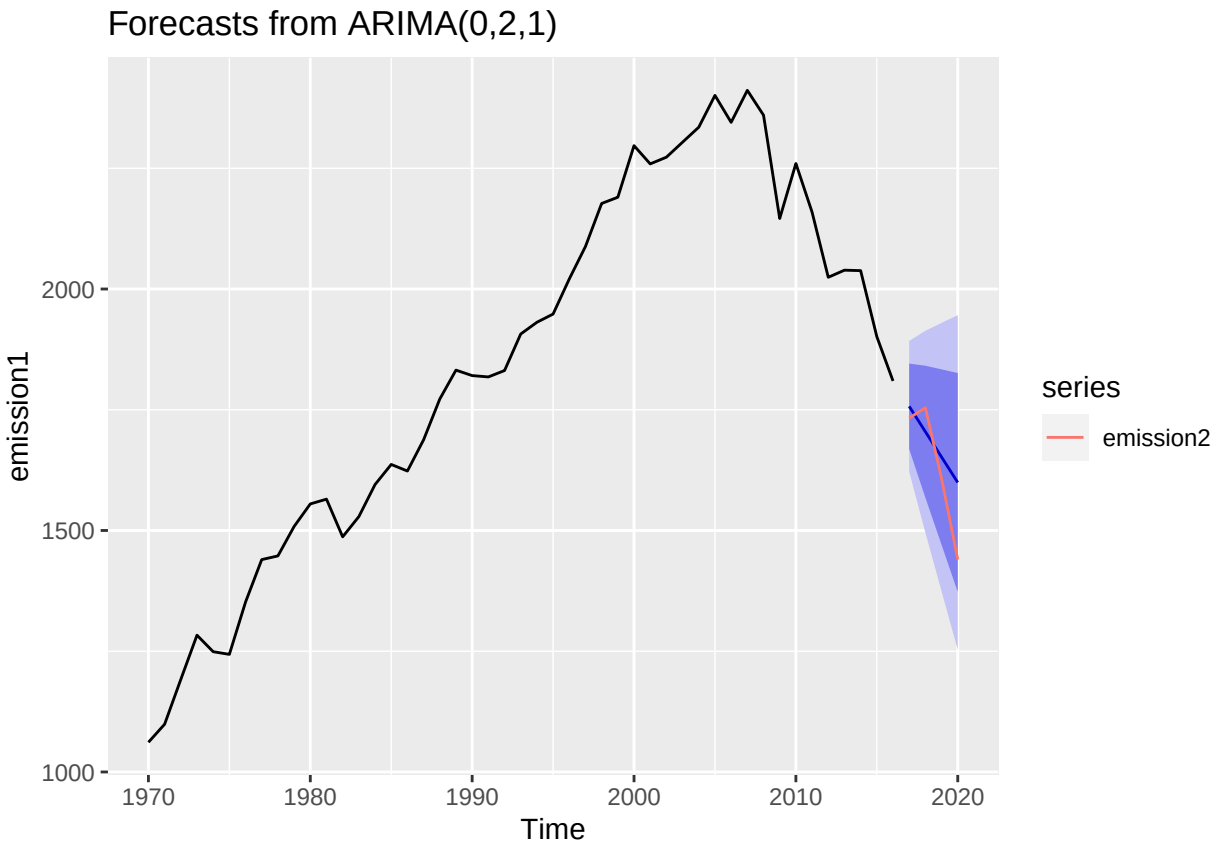
```
summary(fitbic)
```

```
## Series: emission1
## ARIMA(0,2,1)
##
## Coefficients:
##           ma1
##       -0.8233
## s.e.   0.0763
##
## sigma^2 = 4765:  log likelihood = -254.47
## AIC=512.93   AICc=513.22   BIC=516.55
##
## Training set error measures:
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set -12.39249 66.78955 50.91479 -0.6585782 2.778234 0.8717192
##                     ACF1
## Training set -0.1790633
```

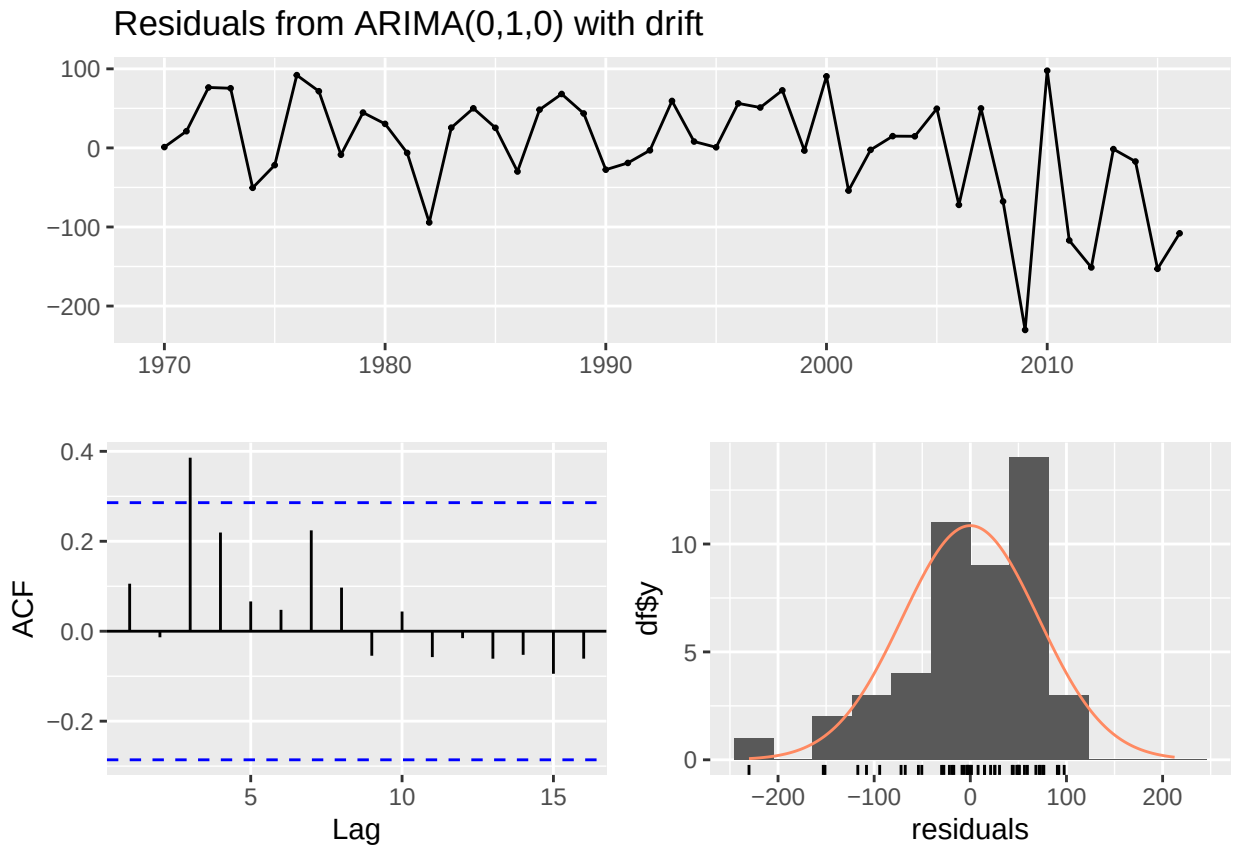
Differentiating the data once did not make the data stationary. However, after differencing the data twice, an Augmented Dickey-Fuller Test checked for a stationary process. The p-value was less than 0.01, so the null hypothesis could be rejected suggesting that the twice differenced dataset was stationary at a 5% significance level.

An ARIMA was used to forecast CO2 emissions from the US energy industry. The once differentiated data contained a unit route, so the dataset was differenced twice. Running an auto.arima on this dataset, the optimal model is found to be an ARIMA (2,2,1) shown in Figure 5.

```
fcst1 <- forecast(fitbic, h=4)
autoplot(fcst1) + autolayer(emission2)
```

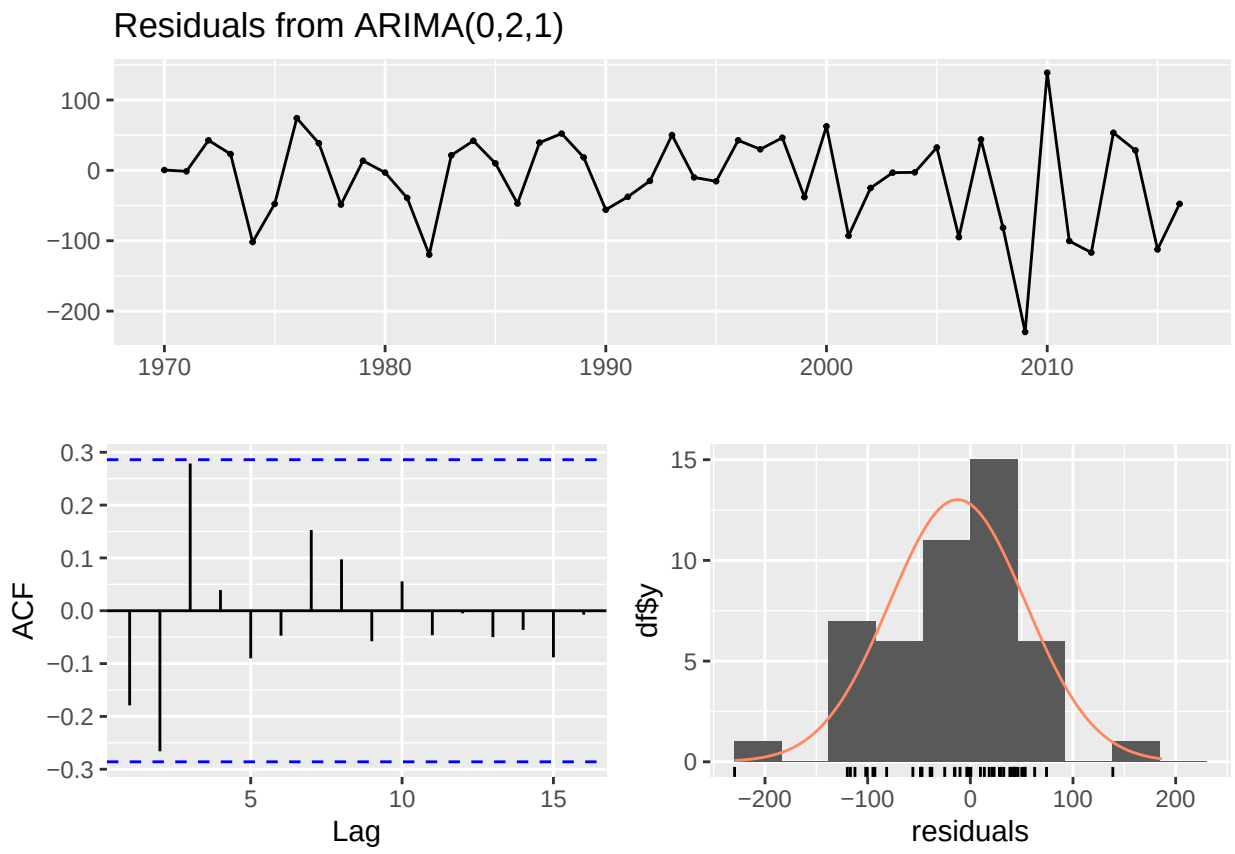



```
checkresiduals(fitaicc)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0) with drift
## Q* = 14.939, df = 9, p-value = 0.09263
##
## Model df: 0.   Total lags used: 9
```

```
checkresiduals(fitbic)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,2,1)
## Q* = 12.046, df = 8, p-value = 0.1491
##
## Model df: 1.   Total lags used: 9
```

```
accuracy(fcst1, emission2)
```

```
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set -12.39249 66.78955 50.91479 -0.6585782 2.778234 0.8717192
## Test set     -44.68497 87.34695 69.52511 -3.1061888 4.522226 1.1903489
##                    ACF1 Theil's U
## Training set -0.1790633        NA
## Test set      0.0849536 0.7924802
```

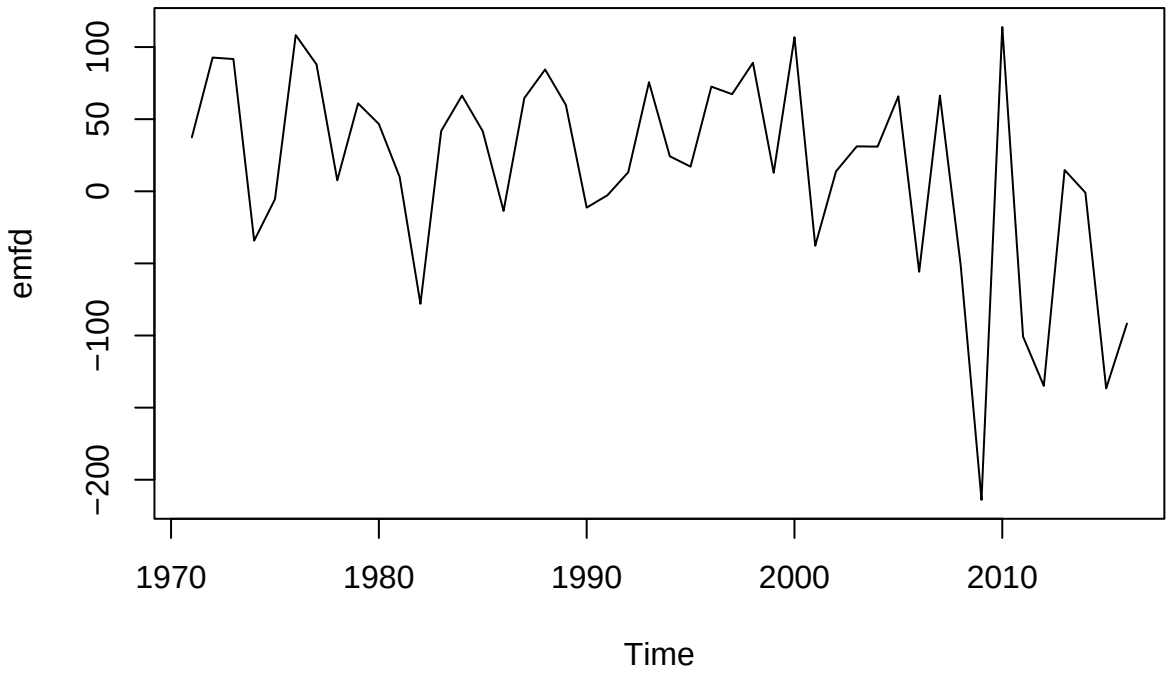
After checking the residuals, it can be concluded that the best Arima model for forecasting CO2 emissions from the US energy industry is an ARIMA (2,2,1). It had a lower RMSE than the other ARIMAs provided.

Structural Break Analysis:

To check for any major changes in the structure of the original dataset on emissions, a structural break test was run. The results shows a structural change in the dataset in 2007. The overall trend of CO2 Emissions

from the US Energy Sector changed after 2007 and became negative, representing a decreasing trend in CO2 emissions.

```
plot(emfd)
```




```
fs.emission <- Fstats(emfd ~ 1)
plot(fs.emission)
sctest(fs.emission)
```

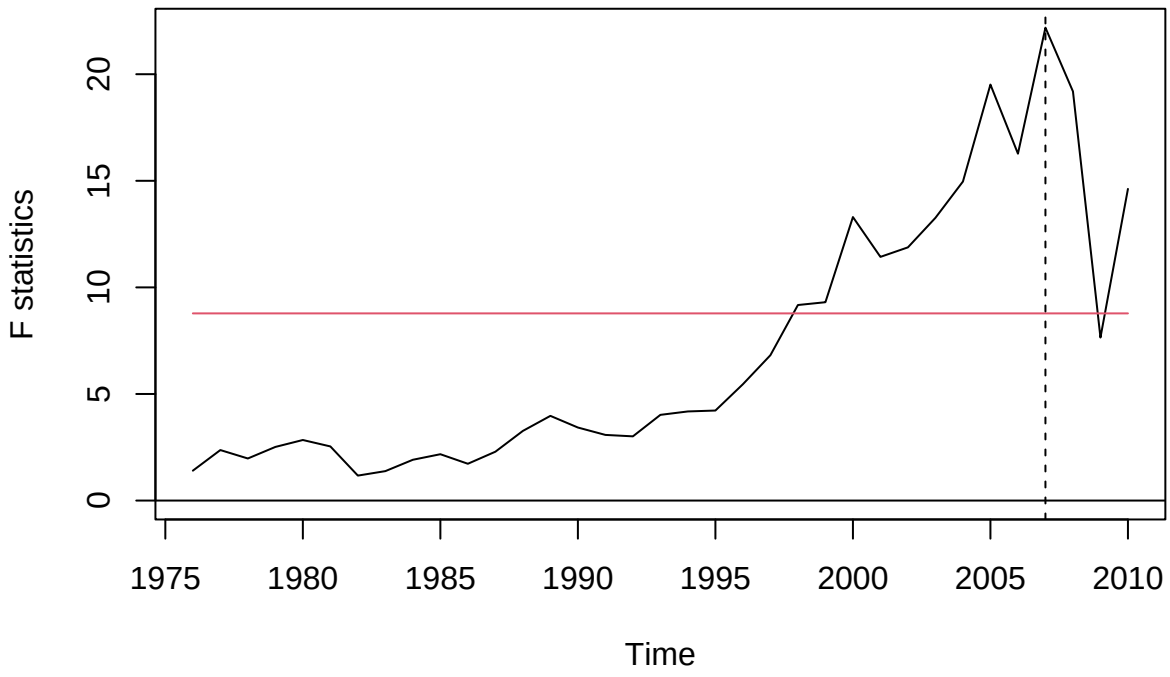
```
##
##  supF test
##
## data:  fs.emission
## sup.F = 22.184, p-value = 8.307e-05
```

```
breakpoints(fs.emission)
```

```
##
##   Optimal 2-segment partition:
##
## Call:
## breakpoints.Fstats(obj = fs.emission)
##
## Breakpoints at observation number:
```

```
## 37
##
## Corresponding to breakdates:
## 2007
```

```
lines(breakpoints(fs.emission))
```

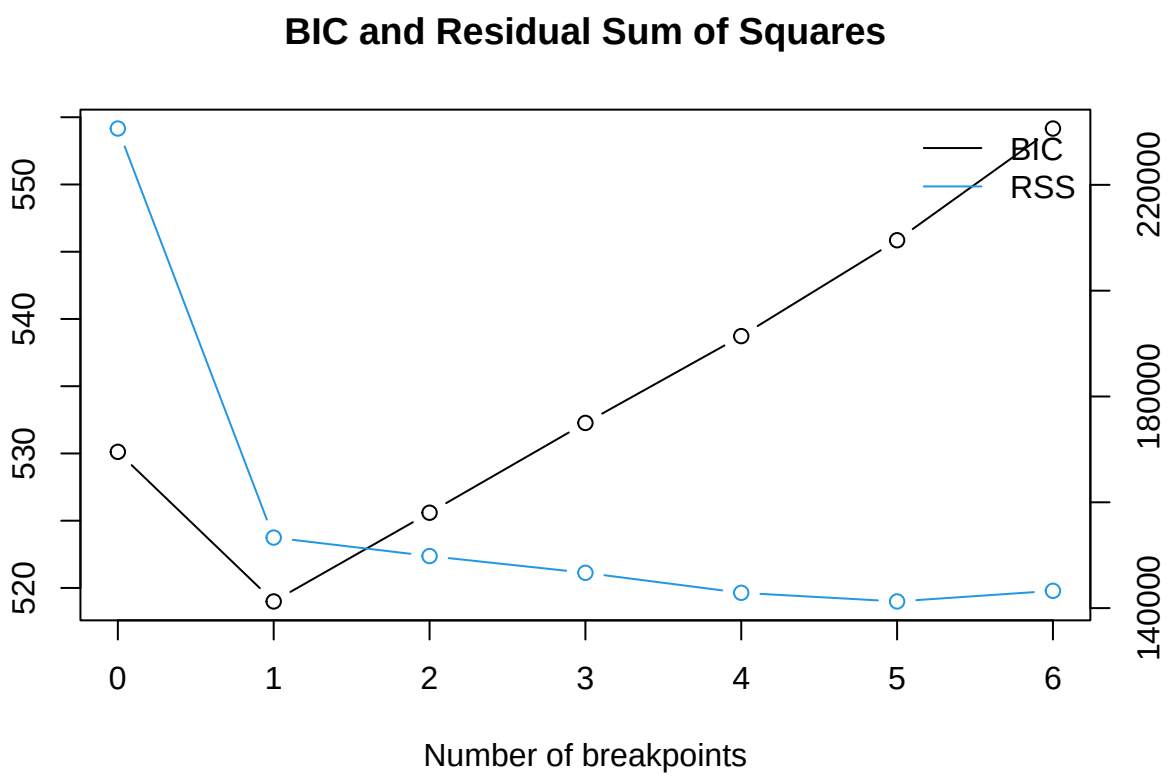



```
bp.emiss <- breakpoints(emfd ~ 1)
summary(bp.emiss)
```

```
##
##   Optimal (m+1)-segment partition:
##
## Call:
## breakpoints.formula(formula = emfd ~ 1)
##
## Breakpoints at observation number:
##
## m = 1                    37
## m = 2                30  37
## m = 3          22    30  37
## m = 4   10     22    30  37
## m = 5   10 16  22    30  37
## m = 6   7  13  19 25 31  37
##
```

```
## Corresponding to breakdates:
##
## m = 1                                2007
## m = 2                           2000 2007
## m = 3                1992       2000 2007
## m = 4   1980         1992       2000 2007
## m = 5   1980 1986    1992       2000 2007
## m = 6   1977 1983 1989 1995 2001 2007
##
## Fit:
##
## m   0        1        2        3        4        5        6
## RSS 230644.9 153336.3 149840.6 146676.9 142894.0 141264.1 143267.3
## BIC    530.1    519.0    525.6    532.3    538.7    545.9    554.2
```

```
plot(bp.emiss)
```




```
breakpoints(bp.emiss)
```

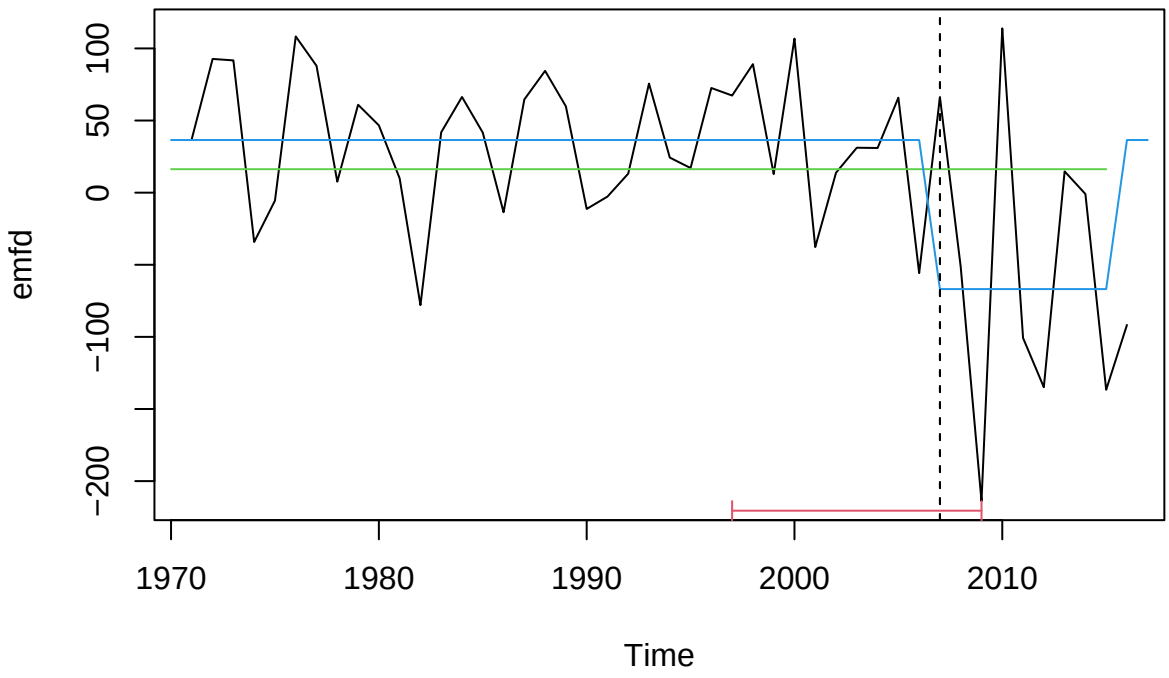
```
##
##   Optimal 2-segment partition:
##
## Call:
## breakpoints.breakpointsfull(obj = bp.emiss)
##
```

```
## Breakpoints at observation number:
## 37
##
## Corresponding to breakdates:
## 2007
```

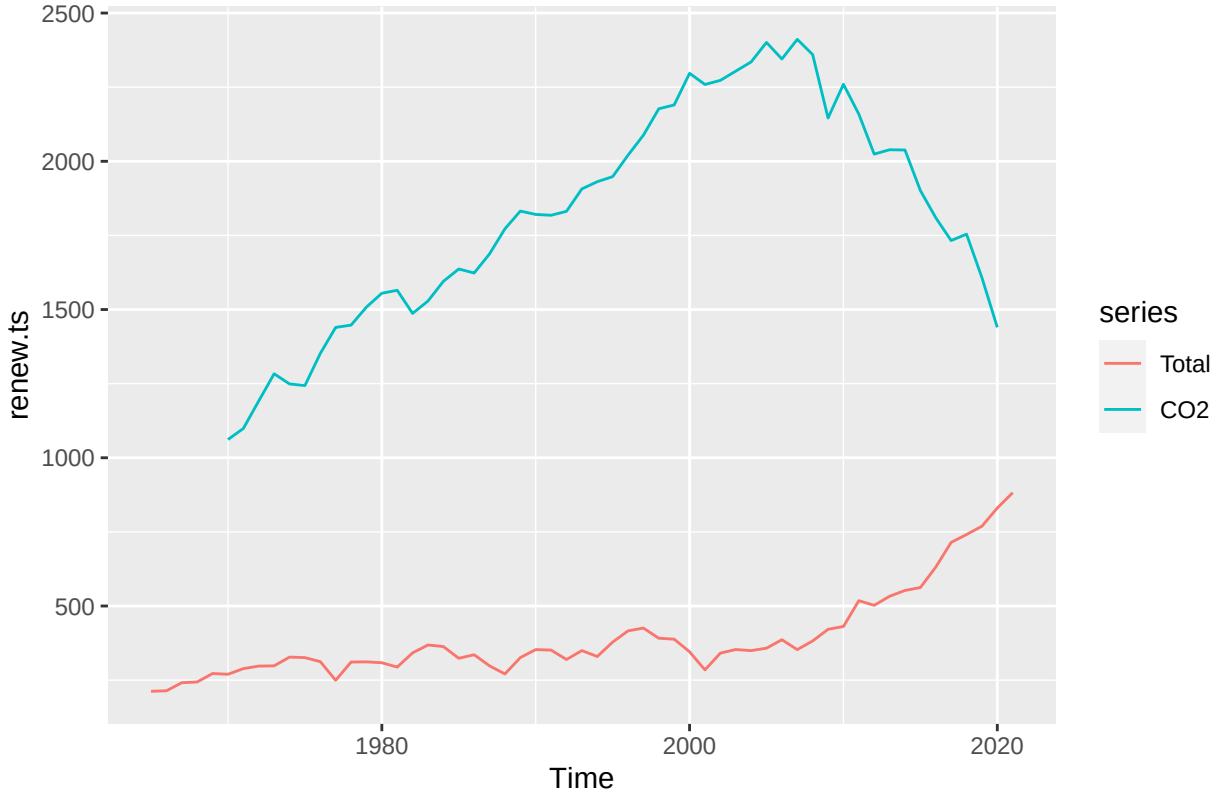
```
fm0 <- lm(emfd ~ 1)
fm1 <- lm(emfd ~ breakfactor(bp.emiss, breaks = 1))
plot(emfd)
lines(ts(fitted(fm0), start = 1970), col = 3)
lines(ts(fitted(fm1), start = 1970, end = 2017), col = 4)
lines(bp.emiss)
ci.emiss <- confint(bp.emiss)
ci.emiss
```

```
##
##   Confidence intervals for breakpoints
##   of optimal 2-segment partition:
##
## Call:
## confint.breakpointsfull(object = bp.emiss)
##
## Breakpoints at observation number:
##   2.5 % breakpoints 97.5 %
## 1    27          37     39
##
## Corresponding to breakdates:
##   2.5 % breakpoints 97.5 %
## 1  1997        2007   2009
```

```
lines(ci.emiss)
```




```
renew <- data.frame(renew[,c(8,9)])
renew.ts <- ts(data=renew, start = 1965, end = 2021)
autoplot(renew.ts)
```




```
renew1 <- window(renew.ts, start = 1970, end = 2016)
renew2 <- window(renew.ts, start = 2017, end = 2021)
```

VAR Modeling:

```
VARselect(renew1, lag.max=5, type="const")
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      1      1      1      1
##
## $criteria
##                  1           2           3           4           5
## AIC(n) 1.556091e+01 1.568351e+01 1.572678e+01 1.581457e+01 1.584536e+01
## HQ(n)  1.565190e+01 1.583516e+01 1.593909e+01 1.608753e+01 1.617899e+01
## SC(n)  1.580915e+01 1.609724e+01 1.630600e+01 1.655928e+01 1.675557e+01
## FPE(n) 5.730998e+06 6.490060e+06 6.804230e+06 7.482368e+06 7.805580e+06
```

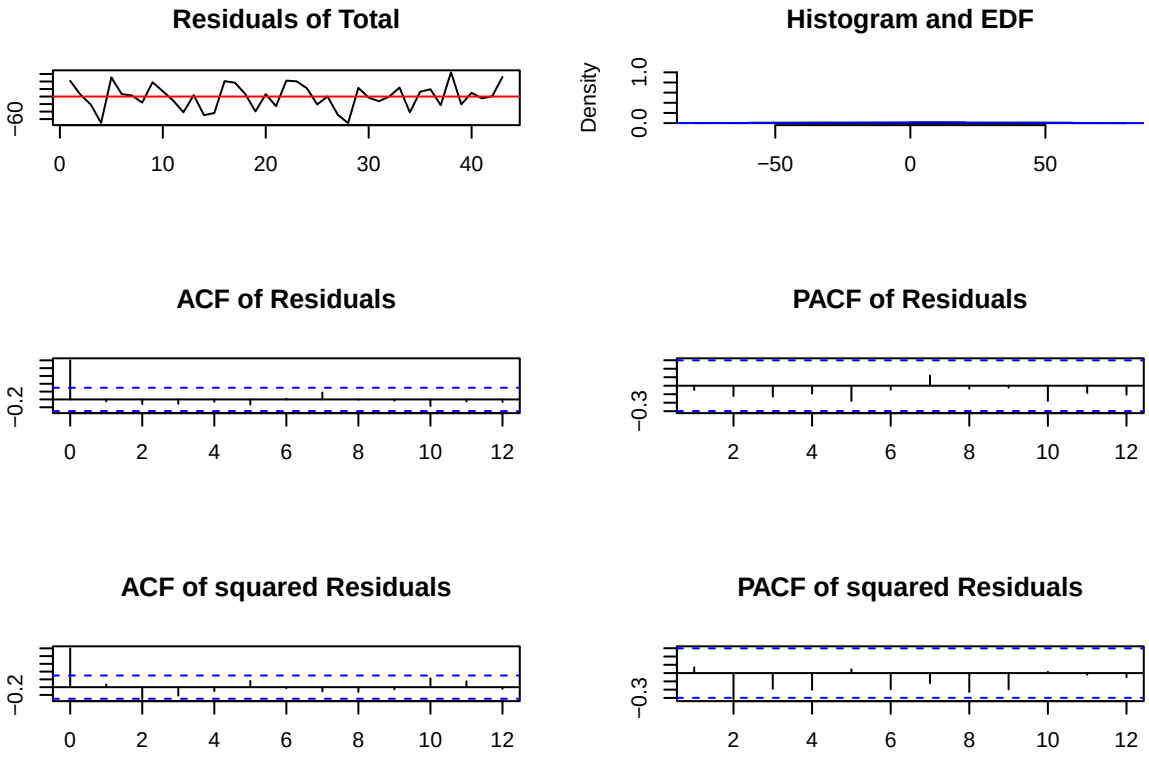
```
var <- VAR(renew1, p=4)
summary(var)
```

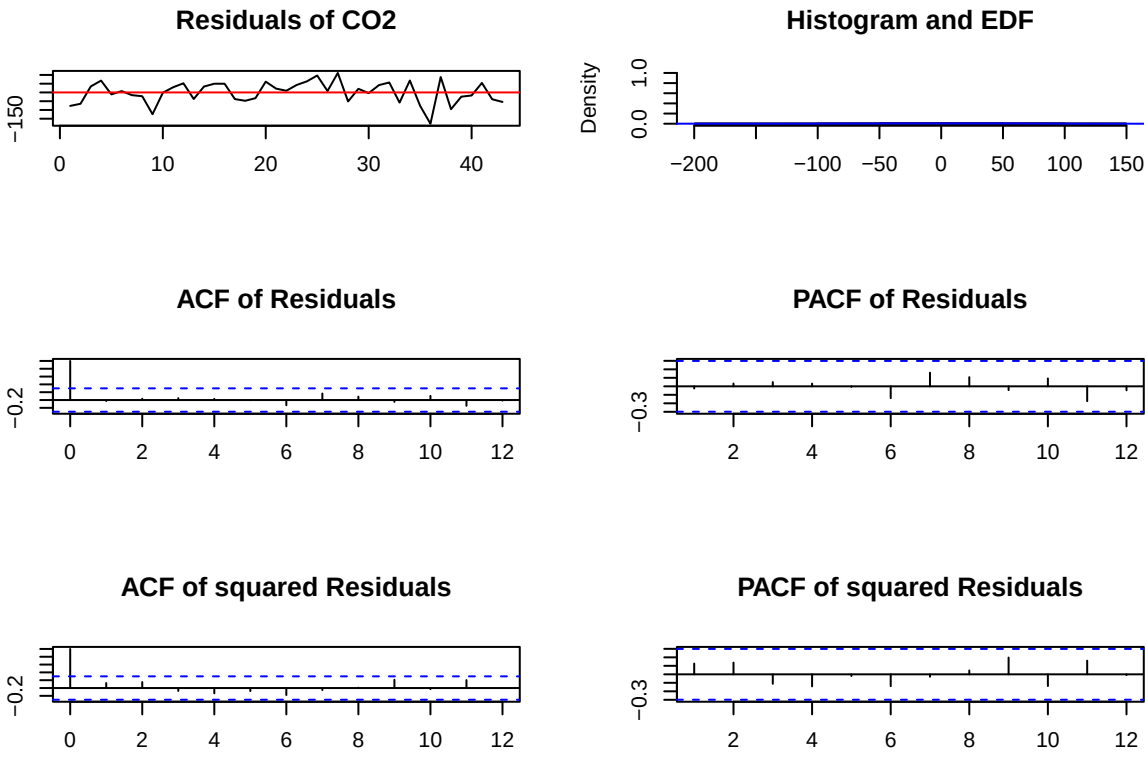
```
##
## VAR Estimation Results:
## =========================
```

```
## Endogenous variables: Total, CO2
## Deterministic variables: const
## Sample size: 43
## Log Likelihood: -443.925
## Roots of the characteristic polynomial:
## 1.011 1.011 0.7476 0.7476 0.4204 0.3113 0.3113 0.1658
## Call:
## VAR(y = renew1, p = 4)
##
##
## Estimation results for equation Total:
## ======================================
## Total = Total.l1 + CO2.l1 + Total.l2 + CO2.l2 + Total.l3 + CO2.l3 + Total.l4 + CO2.l4 + const
##
##           Estimate Std. Error t value Pr(>|t|)
## Total.l1  0.744870   0.198157   3.759 0.000642 ***
## CO2.l1   -0.117867   0.099443  -1.185 0.244126
## Total.l2  0.047760   0.238264   0.200 0.842323
## CO2.l2    0.002432   0.124893   0.019 0.984577
## Total.l3  0.115778   0.239737   0.483 0.632235
## CO2.l3    0.045897   0.126707   0.362 0.719420
## Total.l4  0.042245   0.200439   0.211 0.834332
## CO2.l4    0.076556   0.102468   0.747 0.460128
## const    21.547576  50.376569   0.428 0.671545
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 36.97 on 34 degrees of freedom
## Multiple R-Squared: 0.8403,  Adjusted R-squared: 0.8027
## F-statistic: 22.35 on 8 and 34 DF,  p-value: 2.006e-11
##
##
## Estimation results for equation CO2:
## ====================================
## CO2 = Total.l1 + CO2.l1 + Total.l2 + CO2.l2 + Total.l3 + CO2.l3 + Total.l4 + CO2.l4 + const
##
##           Estimate Std. Error t value Pr(>|t|)
## Total.l1  -0.13502    0.36669  -0.368   0.7150
## CO2.l1     0.94418    0.18402   5.131 1.16e-05 ***
## Total.l2   0.03826    0.44090   0.087   0.9314
## CO2.l2    -0.09202    0.23111  -0.398   0.6930
## Total.l3  -0.07739    0.44363  -0.174   0.8625
## CO2.l3     0.46799    0.23447   1.996   0.0540 .
## Total.l4  -0.13899    0.37091  -0.375   0.7102
## CO2.l4    -0.34332    0.18962  -1.811   0.0790 .
## const    162.76110   93.22116   1.746   0.0898 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 68.41 on 34 degrees of freedom
## Multiple R-Squared: 0.967,   Adjusted R-squared: 0.9592
## F-statistic: 124.5 on 8 and 34 DF,  p-value: < 2.2e-16
```

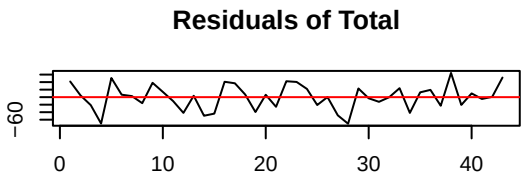
```
##
##
##
## Covariance matrix of residuals:
##       Total   CO2
## Total  1367 -1146
## CO2   -1146  4681
##
## Correlation matrix of residuals:
##         Total     CO2
## Total  1.0000 -0.4531
## CO2   -0.4531  1.0000
```

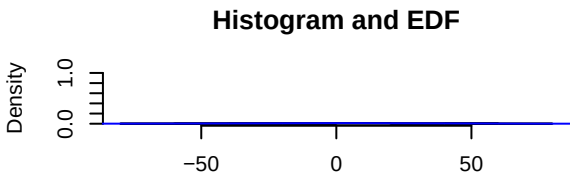
```
serial.test1 <- serial.test(var, lags.pt=5, type="PT.asymptotic")
serial.test2 <- serial.test(var, lags.pt=10, type="PT.asymptotic")
serial.test3 <- serial.test(var, lags.pt=15, type="PT.asymptotic")
plot(serial.test1)
```

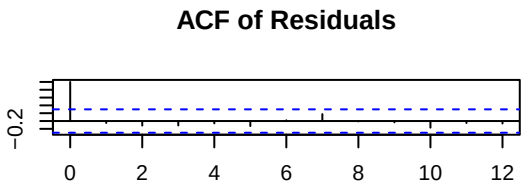


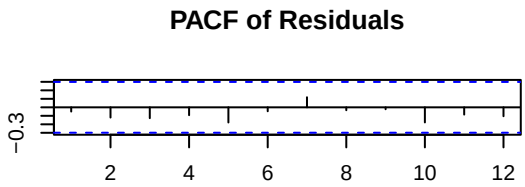


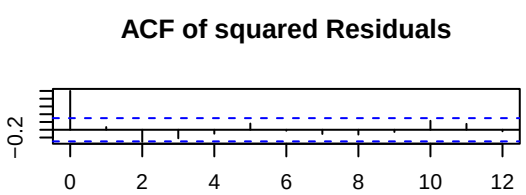

```
plot(serial.test2)
```

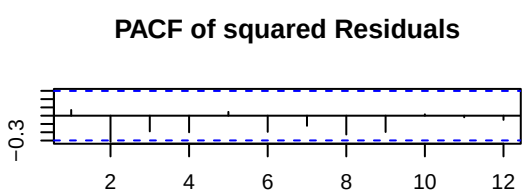

Residuals of Total
−60
0
10
20
30
40

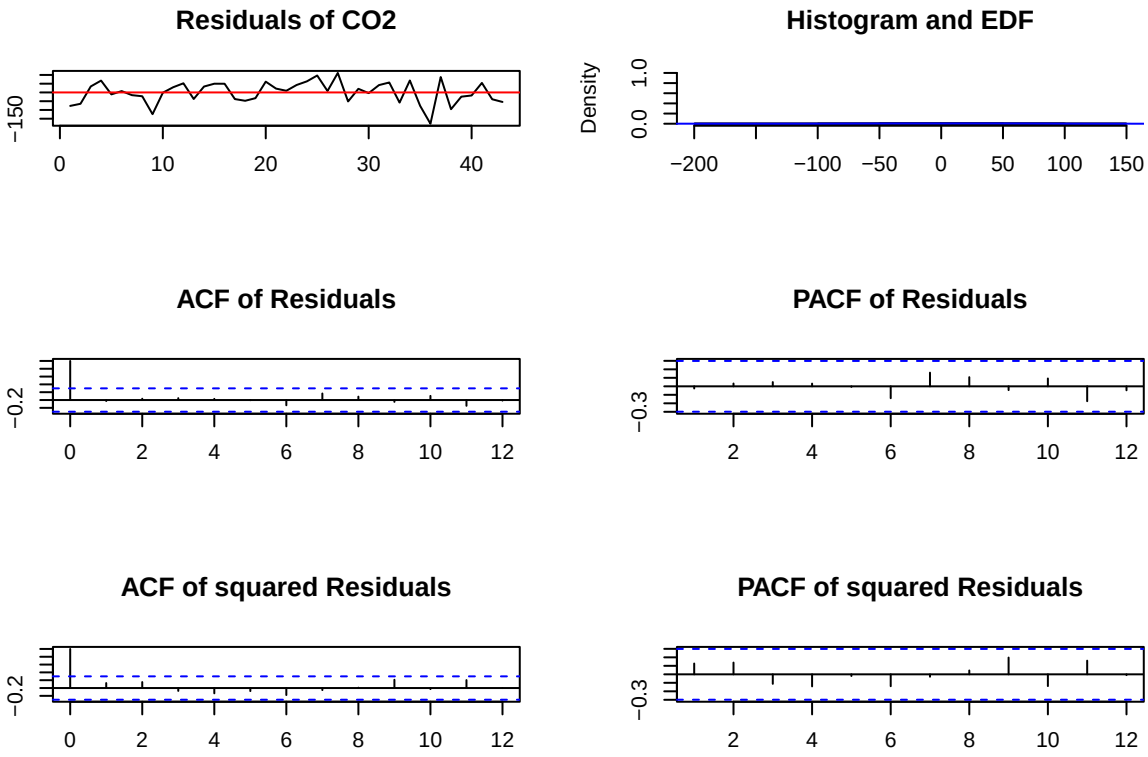

Histogram and EDF
Density
0.0
1.0
−50
0
50


ACF of Residuals
−0.2
0
2
4
6
8
10
12

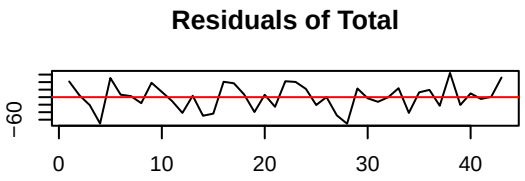

PACF of Residuals
−0.3
2
4
6
8
10
12

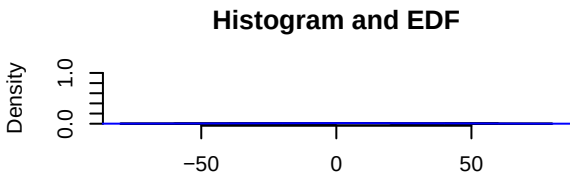

ACF of squared Residuals
−0.2
0
2
4
6
8
10
12

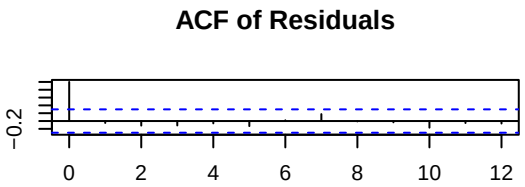

PACF of squared Residuals
−0.3
2
4
6
8
10
12

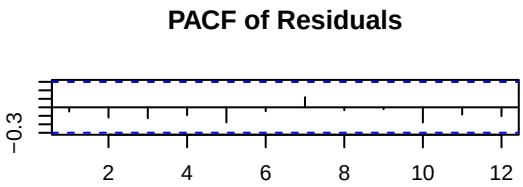


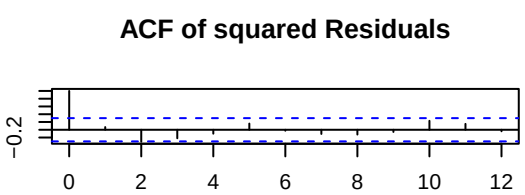

```
plot(serial.test3)
```

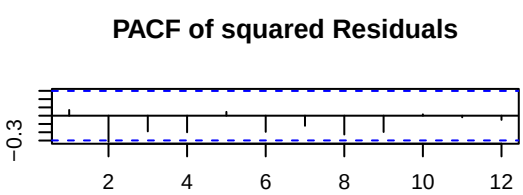

Residuals of Total
−60
0
10
20
30
40

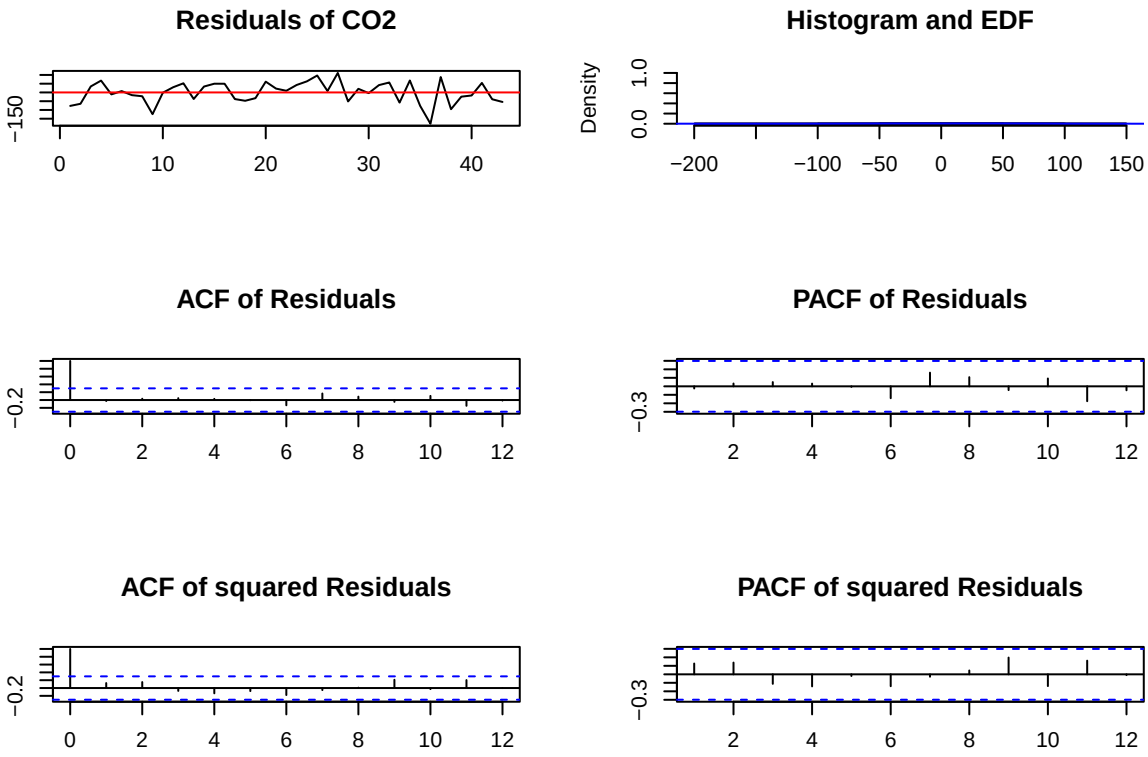

Histogram and EDF
Density
0.0
1.0
−50
0
50


ACF of Residuals
−0.2
0
2
4
6
8
10
12


PACF of Residuals
−0.3
2
4
6
8
10
12

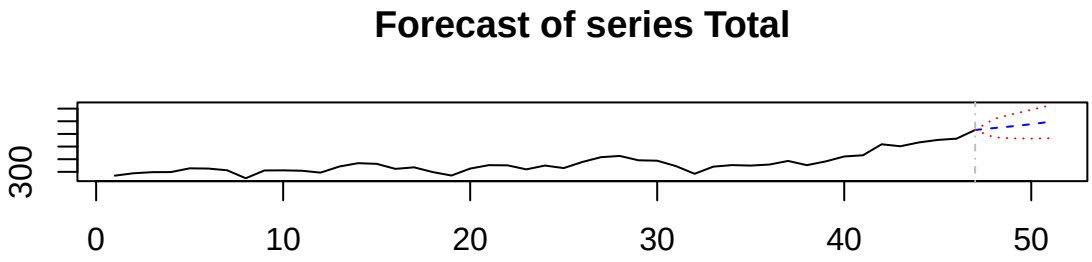

ACF of squared Residuals
−0.2
0
2
4
6
8
10
12

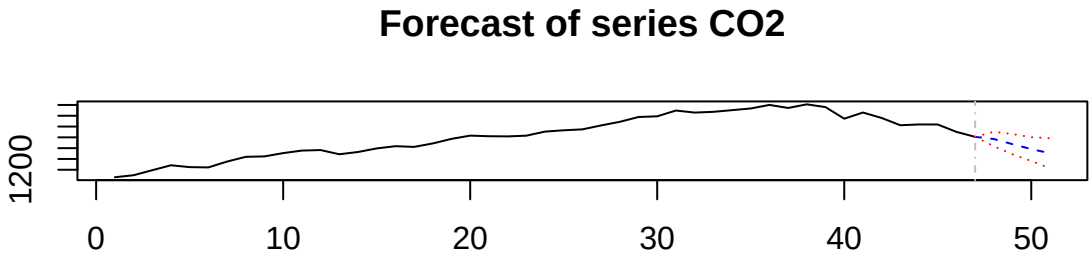

PACF of squared Residuals
−0.3
2
4
6
8
10
12



After testing different models RMSE, BIC, and AIC, the optimal forecasting model ARIMA (2,2,1) was chosen. Since the dataset on emissions only had 50 observations, it was decided that only four observations could be accurately predicted.

```
pred <- predict(var, n.ahead = 4, ci = 0.95)
plot(pred)
```






```
TOT.ts <- ts(pred$fcst$Total, start=2017)
CO2.ts <- ts(pred$fcst$CO2, start=2017)
Total2 <- ts(renew$Total, start=2017)
accuracy(TOT.ts, Total2)
```

```
##                  ME    RMSE      MAE       MPE     MAPE       ACF1 Theil's U
## Test set -441.9555 442.022 441.9555 -194.7324 194.7324 -0.5614136   27.1083
```

```
accuracy(CO2.ts, emission2)
```

```
##                  ME     RMSE      MAE        MPE    MAPE       ACF1 Theil's U
## Test set 0.3524246 56.70491 52.11196 -0.1617923 3.21644 -0.2072653 0.4747584
```

```
# Causality Tests
caus.cons1 <- causality(var, cause = "Total", boot=FALSE)
caus.cons1
```

```
## $Granger
##
##  Granger causality H0: Total do not Granger-cause CO2
##
## data:  VAR object var
## F-Test = 0.3627, df1 = 4, df2 = 68, p-value = 0.8343
##
```

```
##
## $Instant
##
##  H0: No instantaneous causality between: Total and CO2
##
## data:  VAR object var
## Chi-squared = 7.3251, df = 1, p-value = 0.0068
```

```
caus.cons2 <- causality(var, cause = "CO2", boot=FALSE)
caus.cons2
```

```
## $Granger
##
##  Granger causality H0: CO2 do not Granger-cause Total
##
## data:  VAR object var
## F-Test = 0.70486, df1 = 4, df2 = 68, p-value = 0.5914
##
##
## $Instant
##
##  H0: No instantaneous causality between: CO2 and Total
##
## data:  VAR object var
## Chi-squared = 7.3251, df = 1, p-value = 0.0068
```

Conclusion:

The most accurate model found was the ARIMA (2,2,1) model. In this analysis, only historical US CO2 energy emissions are modeled. Since the test period is only 4 years (2017-2020), though, it may be advisable to take the RMSE with a grain of salt. By only looking at historical CO2 trends, moving averages, and shocks, it may not reflect the increasing importance of global warming on economic actors and the recently burgeoning renewables industry. Despite this, it is still a more optimistic forecast than the EIA's.

For comparison, 2030 will be used as a reference point since that is when the US committed to being carbon neutral. The EIA predicts that the US energy market will emit 1,214 billion tons of CO2 by 2030. On the other hand, the ARIMA model predicts emissions of 1,074 billion tons, 11% less than the EIA's prediction. It is still a far cry from the US's international commitments. However, it is possible that the VAR model is a better predictor despite its higher RMSE.

Despite being less accurate than the ARIMA model in the test period, it is possible that the VAR model will better predict the trajectory of energy CO2 emissions in the US. The two granger-causes used for CO2 emissions are renewables output in the energy sector and changes in global temperatures. Both are statistically significant granger causes of CO2 emissions when jointly used in a VAR model. This surprisingly implies that US energy emissions have been responsive to rising global temperatures since 1970. The validity of the VAR forecast seems to be reliant on this historical relationship which may warrant further investigation.

That said, global temperatures have become a more prevalent issue in both US government policy and corporate social responsibility. Since CO2 emissions are no longer coupled with GDP growth and renewable technologies are becoming more efficient and less costly, economic agents in the US have the option to minimize negative externalities without necessarily hurting their productivity growth. Assuming economic actors are rational, these agents should increasingly transition to renewable energy. In other words, if economic theory is to be trusted, the invisible hand may be a major proponent of decreasing emissions. This

may especially be the case when the negative externalities from CO2 become increasingly tangible through greater and more frequent natural disasters and climate anomalies. This relationship is modeled through global temperatures as a one degree increase year on year leads to an increase in the frequency of dramatic weather events (EPA). Furthermore, renewables also seem to have a type of Moore's Law as there seems to be consistent proportional increases in efficiency regarding renewable energy production (Blankenhorn).

If one believes that the relationship between global temperatures, renewable output, and CO2 energy output is true, there is hope that US energy emissions will be 200 billion tonnes by 2030. This is a whopping 16% of the EIA's forecasts. The EIA study does not seem to factor in this iteration of Moore's law while the VAR model uses renewable output as a proxy, which may be the biggest reason for this discrepancy between this paper's VAR model and the EIA study. Considering market trends, greater government involvement, and increasing urgency to decrease emissions, CO2 emissions in energy production may end up being more responsive to changes in global temperatures than what the VAR model, based on yearly data from 1970 to 2017, predicts.

The biggest impediment to this analysis is that the ARIMA and VAR forecasts differ so drastically. While this analysis argues that the EIA probably overestimated CO2 forecasts, the extent of that overestimation is difficult to discern. This is also, by no means, a study that attempts to temper people's passion for protecting the environment. Conversely, it should be interpreted as reaping the fruits of environmental work, activism, and research. While there is still a lot more to do, it may also be important to appreciate what has been done.