Data Analysis of Sulfur Hexafluoride Levels in American Samoa

### 

Nick Corbin

# Table of Contents

[**Table of Contents**](#_94kzr8qgkpzj) **2**

[**Executive Summary**](#_vju7a1o9x88m) **3**

[**Introduction**](#_tthom0tlgx9r) **3**

[**Time Series Data Analysis**](#_pkzci4h5v7na) **4**

[Data Exploration](#_uxhq99tlcosd) 4

[Trend and Seasonality](#_8mfved7uuab6) 4

[Autocorrelation Function](#_81wqso1umjc2) 10

[Time Series Regression](#_ehgkdg5rgij0) 11

[Exponential Smoothing](#_213j4wp9a6iw) 15

[ARIMA Forecasting](#_l77qw2shrc80) 22

[**Conclusion**](#_i1ishybv9lu1) **26**

[**References**](#_qnmgt9ntps4o) **27**

[Works Cited](#_ozj6bprgy60d) 27

[R Code](#_r4d81rbo85er) 27

[Data Exploration](#_7mx1aspxvcs8) 27

[Trend and Seasonality of TS data](#_36y884o4qth9) 28

[Exponential Smoothing](#_simpyc4mpvrd) 29

[Times Series Regression](#_l3dmm7y3b082) 32

[ARIMA Forecasting](#_lhr51s55tlvl) 35

# Executive Summary

The time series analysis of sulfur hexafluoride levels in American Samoa uses multiple modeling techniques to forecast the future sulfur hexafluoride levels in American Samoa. Using multiple measures of accuracy including MSE, the autocorrelation function, the Box-Ljung test, and the partial autocorrelation function to name a few; of these options, the technique that offered the most accurate forecast was time series regression.

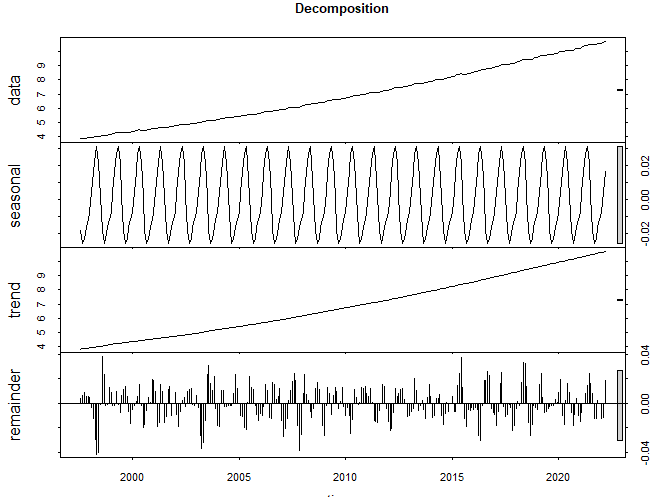
# Introduction

2881.3 miles east of Australia is the island territory of American Samoa. American Samoa is owned by the United States of America. American Samoa’s topography is tropical with lots of mountains due to it being a volcanic island. With a population of 45,443 and a total land area of 199 square kilometers which is slightly larger than Washington D.C, there are not a whole lot of pollutants coming out of American Samoa. Their main export is canned tuna, accounting for 93% of their total exports. Why do we care so much about these exports? Why do we care about the population size? The answer: climate change. According to the American Samoa Environmental Protection Agency (ASEPA), American Samoa could not be in more danger of going underwater. The ASEPA warns “Any rise in ocean level will threaten homes and displace our residences. Any loss of coastal land will disturb the delicate balance of our Territory’s ability to provide resources such as food and local building materials. Most of our local water supply comes from underground freshwater aquifers. Any rise in ocean level will likely compromise our freshwater aquifers via seawater intrusion.” What specifically is contributing to this threat? Sulfur hexafluoride, or in its molecular form,. Sulfur hexafluoride is notable due to its potency. The compound is 10 times more potent than carbon dioxide () and according to the Intergovernmental Panel on Climate Change’s (IPCC) fifth assessment report “It’s estimated that, over a 100-year period, SF6 is 23,500 times more effective at trapping infrared radiation than CO21, meaning that 1 kg of SF6 has the same impact as 23,500 kg of CO2. Once in the atmosphere, it has an atmospheric lifetime of 3,200 years, which means it can accumulate without degrading for millennia to come.” To better track impending climate change, I look to predict future sulfur hexafluoride readings using current sulfur hexafluoride flask readings from the National Oceanic and Atmospheric Administration. I can then monitor how the fight against climate change is going and if any additional measures are needed to deter the rising levels of the compound.

# Time Series Data Analysis

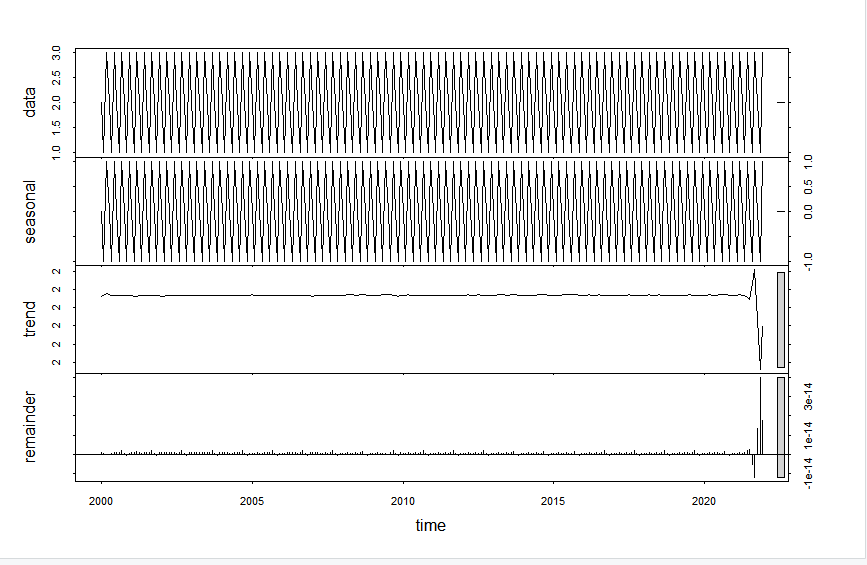
## Data Exploration

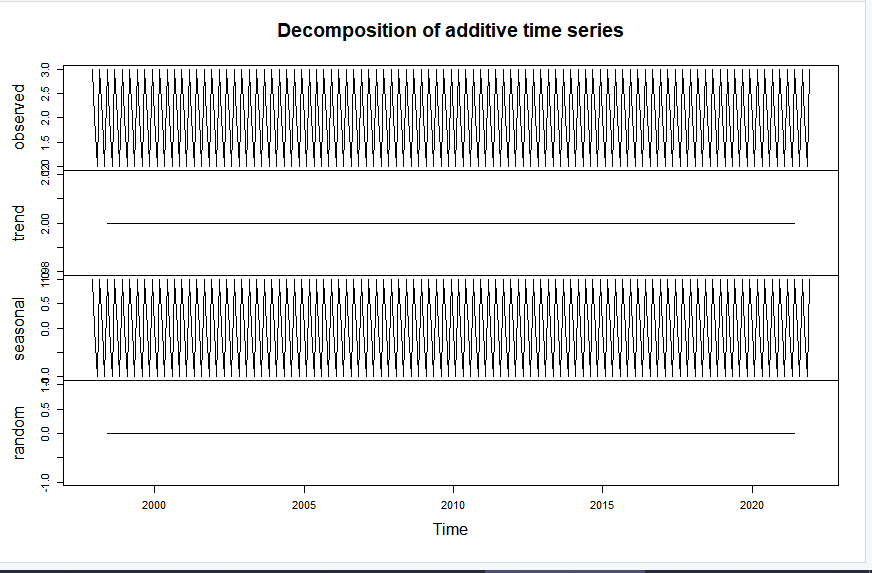
The data set I am analyzing starts in 1997 and continues through 2021. Looking at the decomposition of the time series, we can clearly see that the overall data models an increasing linear dataset. I can see a seasonal component, throughout the years there are clear highs and lows in the concentration of sulfur hexafluoride. The overall trend of the data is almost identical to the original data plot. The remainder section of the decomposition graph shows the randomness in the data, some abnormalities are not explained by the seasonality.



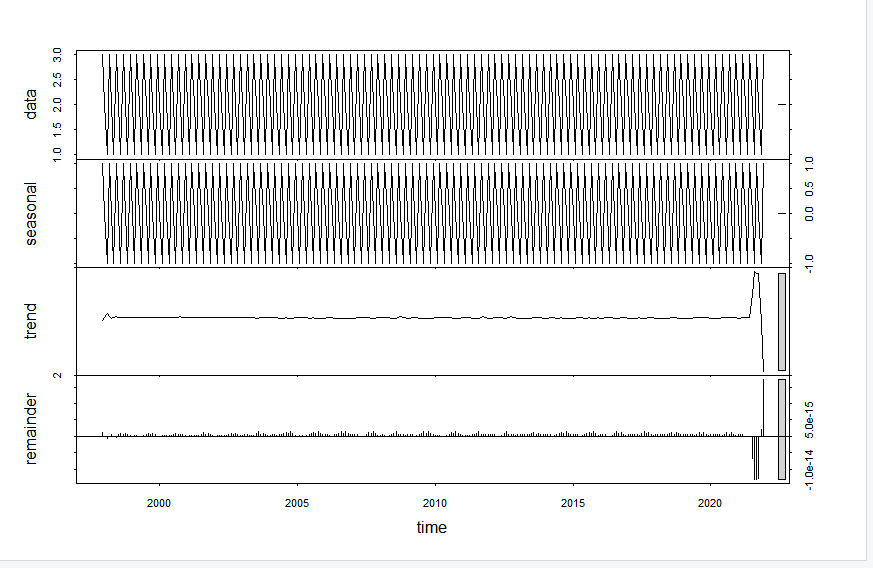
### Trend and Seasonality

Decomposition of time series analysis on the 21st Century

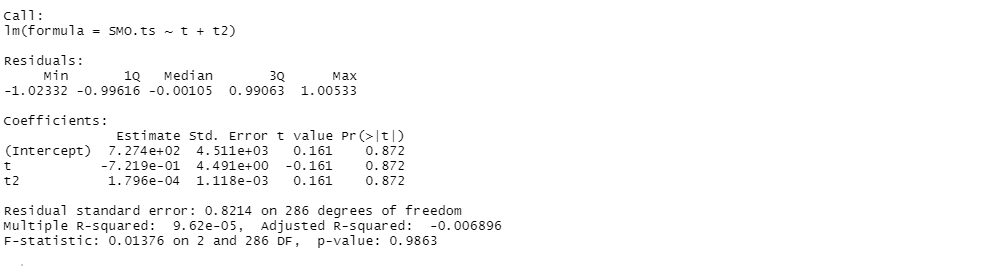


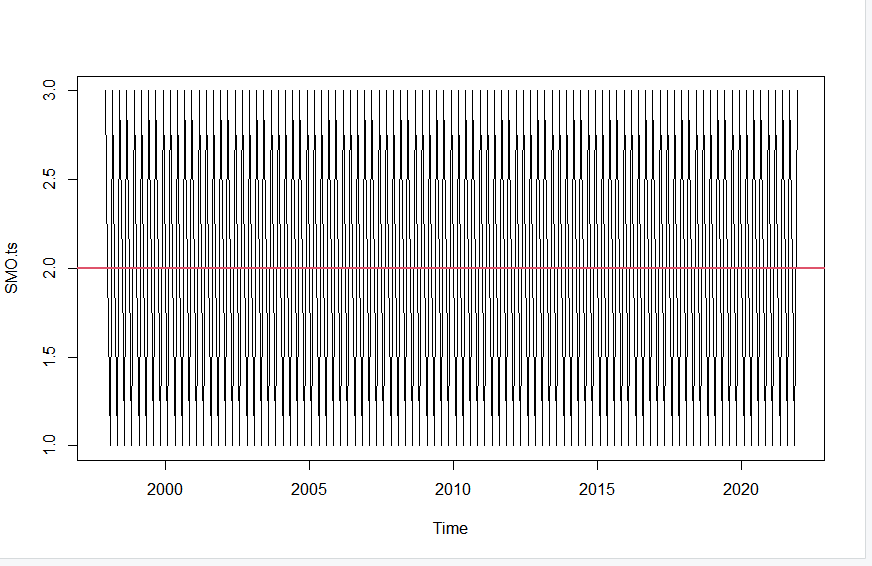


Decomposition plot (21st Century)

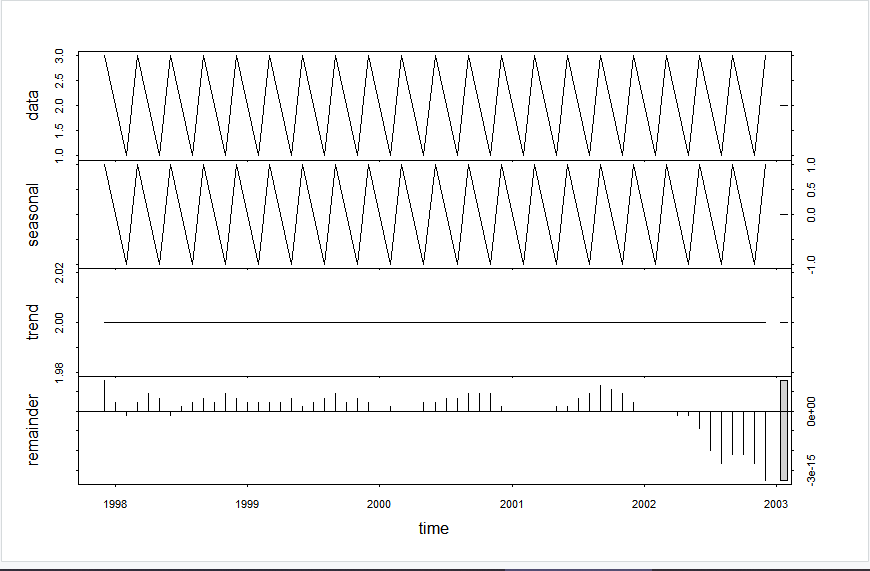


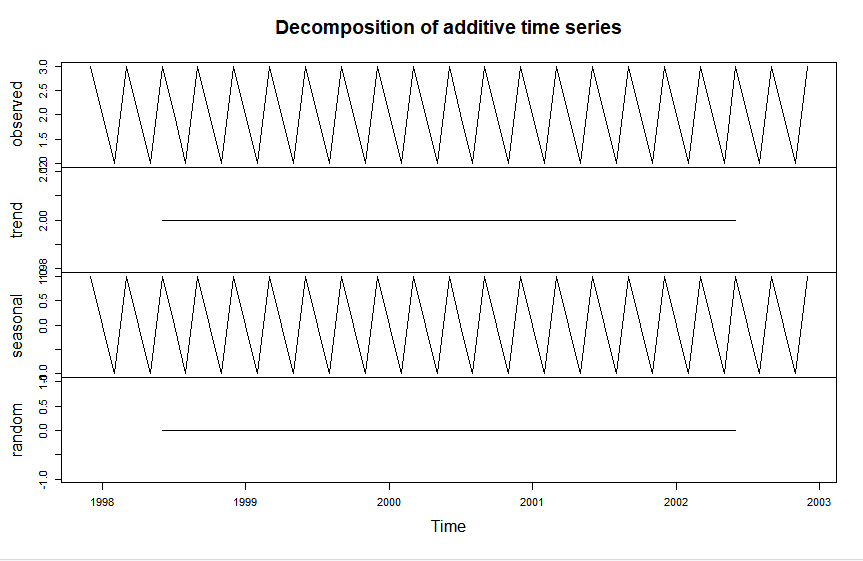
Linear trend



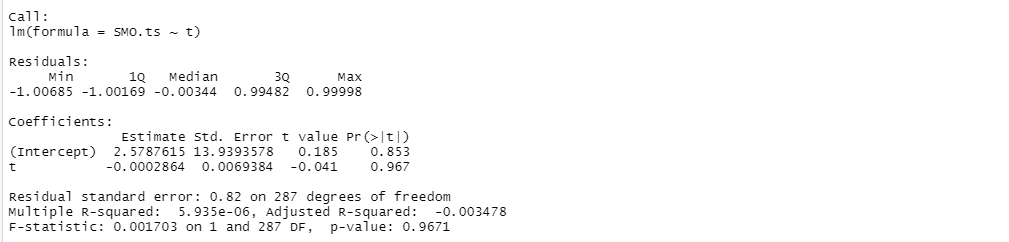


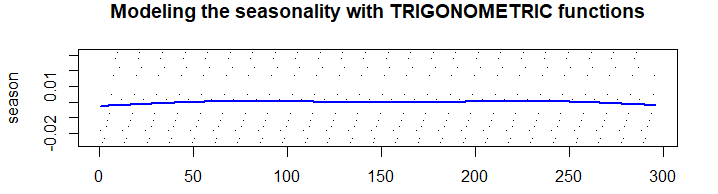
Decomposition of First 5 Years of American Samoa Sulfur Hexafluoride Time Series

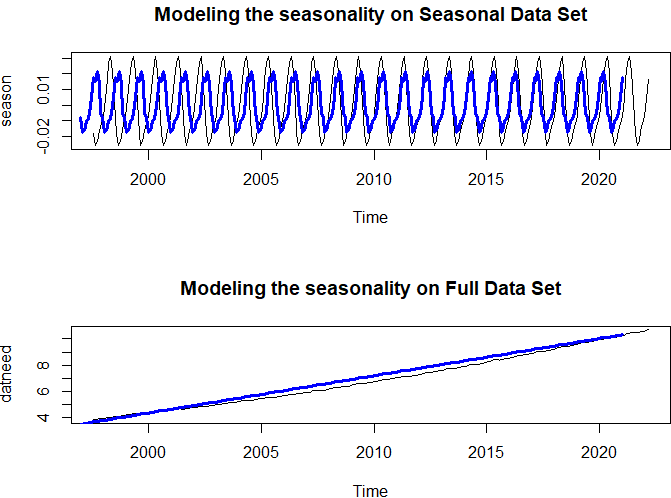




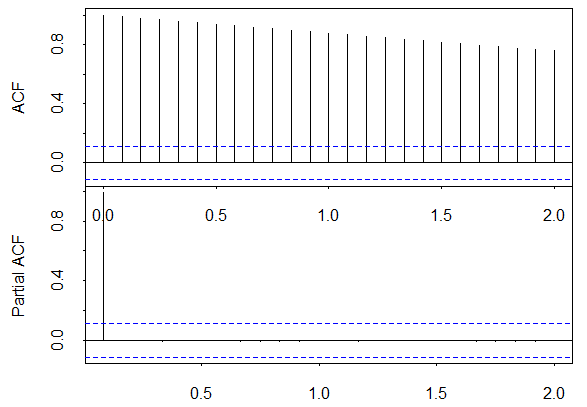
Quadratic trend plot (First 5 years)







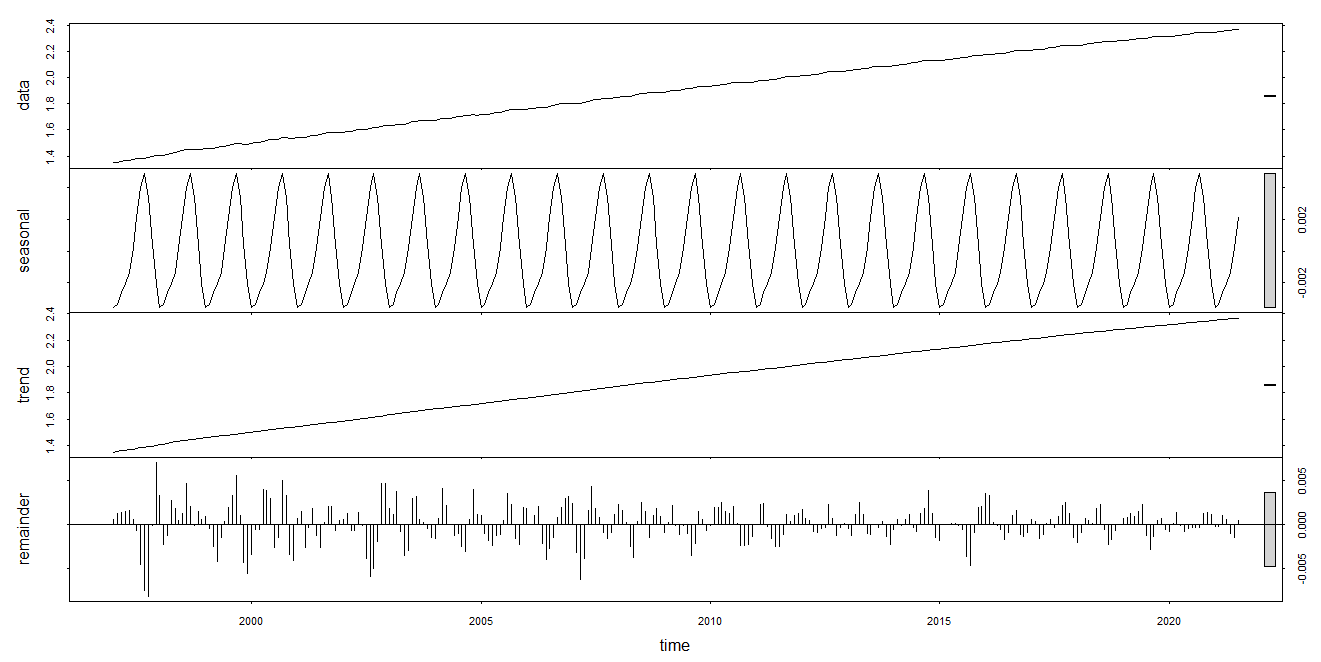
Seasonality is a characteristic of time series data that describes the regular and predictable changes that occur every year. For our data, I have regular spikes and drops throughout the years. I modeled the seasonality using dummy variables, one for each month so twelve in total, after trying to model the seasonality with a trigonometric function. The trigonometric function was ill-fitting and was unable to map our seasonality. While using the dummy variables, I was able to find a model that explained most of the seasonality. After mapping the model onto our complete data set, I see that the seasonality captures most of the variation but is unable to explain the complete dataset. Looking at our time series decomposition for the first five years of the data, I see there is constant seasonality, a constant trend, and after 2002 the remainder begins to get increasingly negative at an alarming rate.



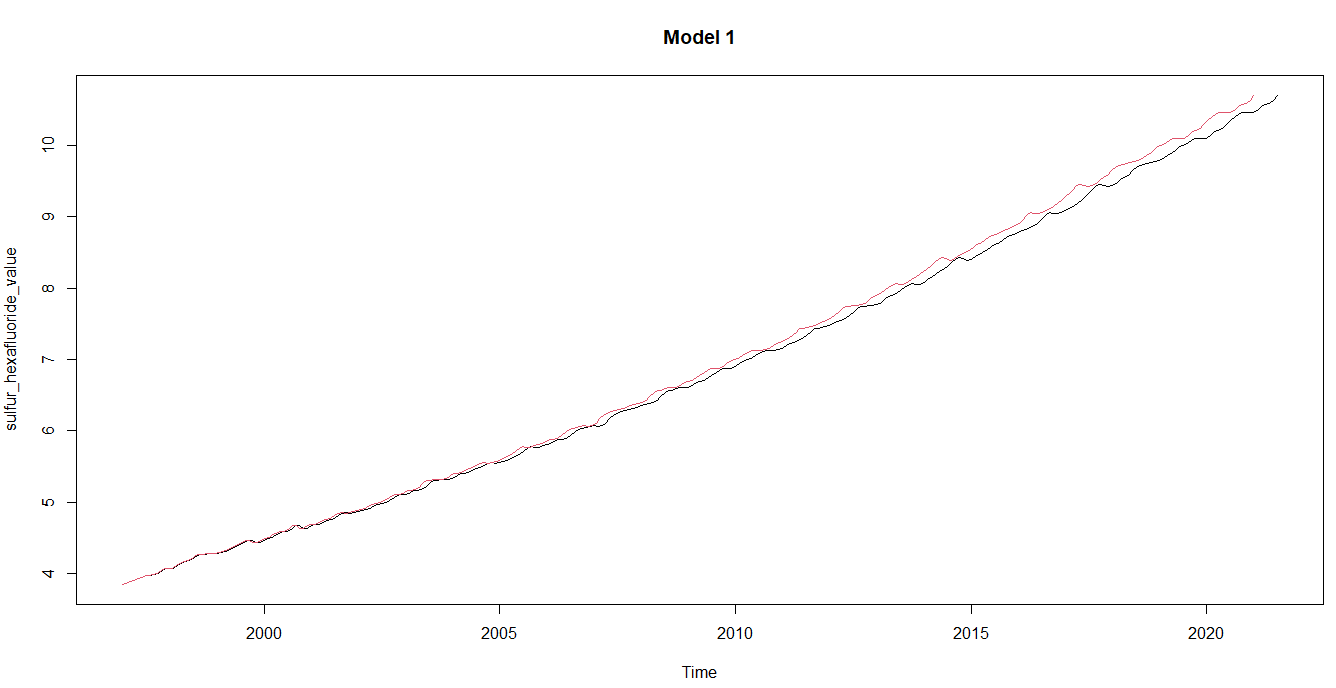
The autocorrelation function shows the correlation between the time series and its own lagged values. This represents a correlation coefficient between the series and the past values. In the graph above, I see a gradual decrease in the correlation. Since there are no spikes in the autocorrelation function graph we can assume that there is little to no seasonality in the data. The partial autocorrelation function gives the partial correlation of a stationary time series with its own lagged values, regressing the values of the time series at all shorter lags. The partial autocorrelation function graph above shows that there is no significant lag after lag(0) and like the autocorrelation function graph, the partial autocorrelation graph suggests that there is little to no seasonality.

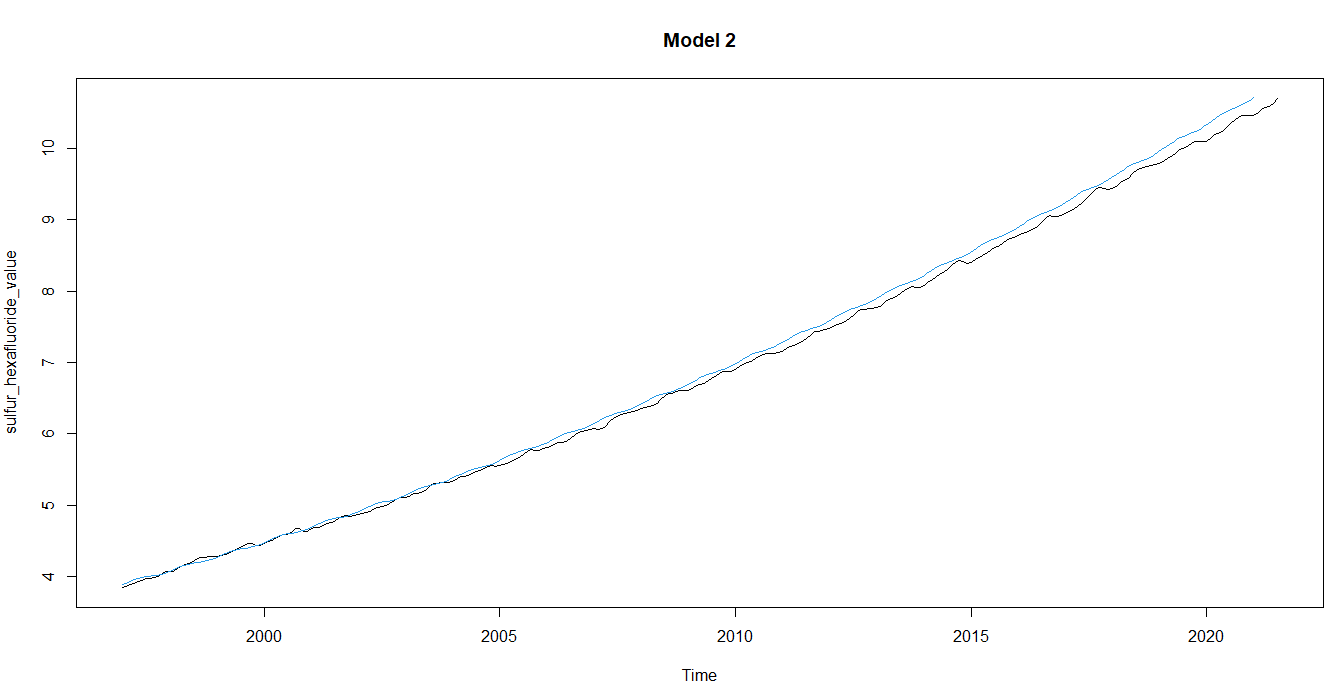
## Time Series Regression

Time Series Regression is used to predict the future using multiple factors. In this case, I utilized different methods of transformation to account for multiple factors, and from there I constructed two models. Before I did anything, I made sure the data was correct by checking the head and tail end of it. Next, I identified the time and cycle of the data. The actual goal of these models is to try and predict what will happen to future values, 10, 20, 40, etc years down the line. Time Series Regression is one of the simpler ways of doing this because it just combines these factors and gives a predictive model. Between the first and second models, I wanted to make sure which factors of the transformed data would be the most significant. Here is what I found.

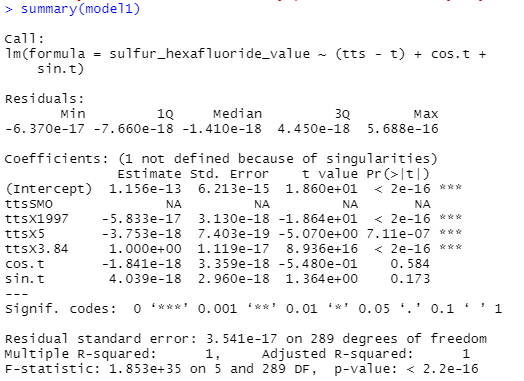


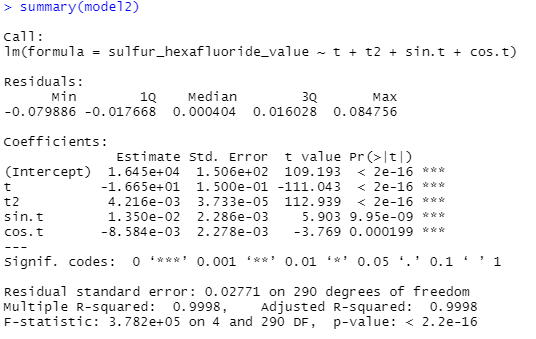
First, I check some trend and seasonality components of the data As depicted by the data, there is clear trend and seasonality, so don’t need to look at log.

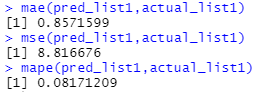




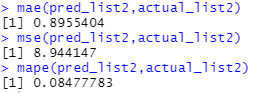
At first glance, these models look pretty similar. A noticeable difference visually includes a sharper measure of seasonality in Model 1 so I will look at numeric data to determine which is more accurate.



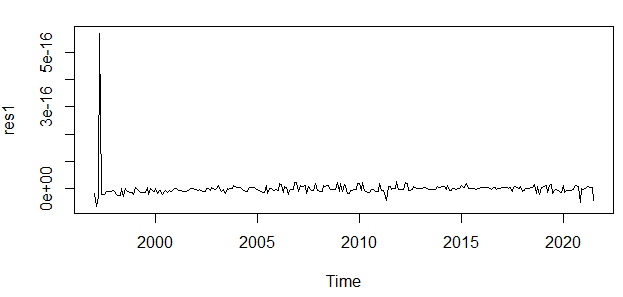




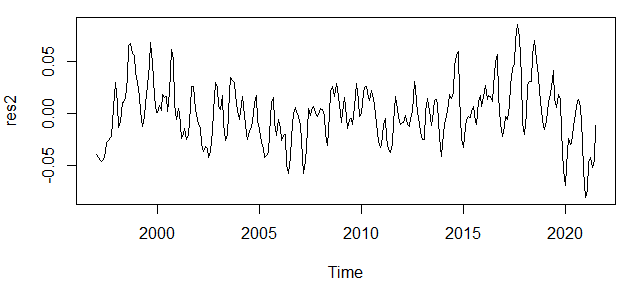
Above: MAE, MSE, MAPE for Model 1



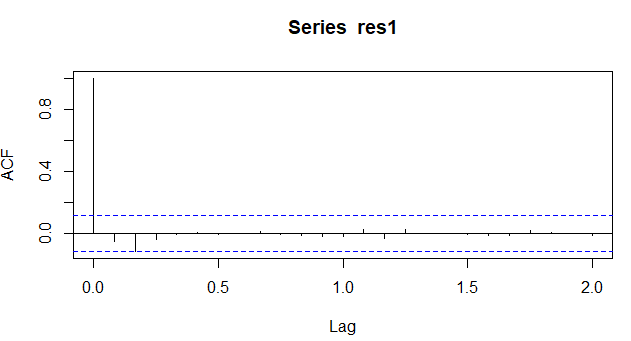
Above: MAE, MSE, MAPE for Model 2

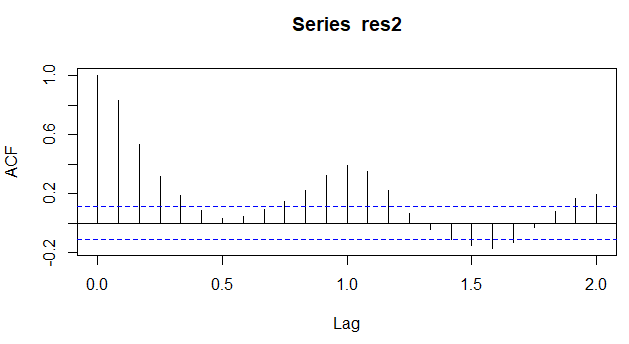


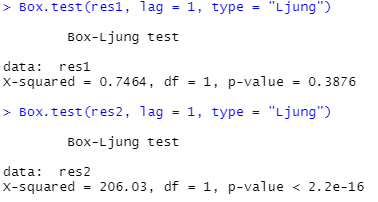
Above: Residuals Plot Model 1



Above: Residuals Plot Model 2

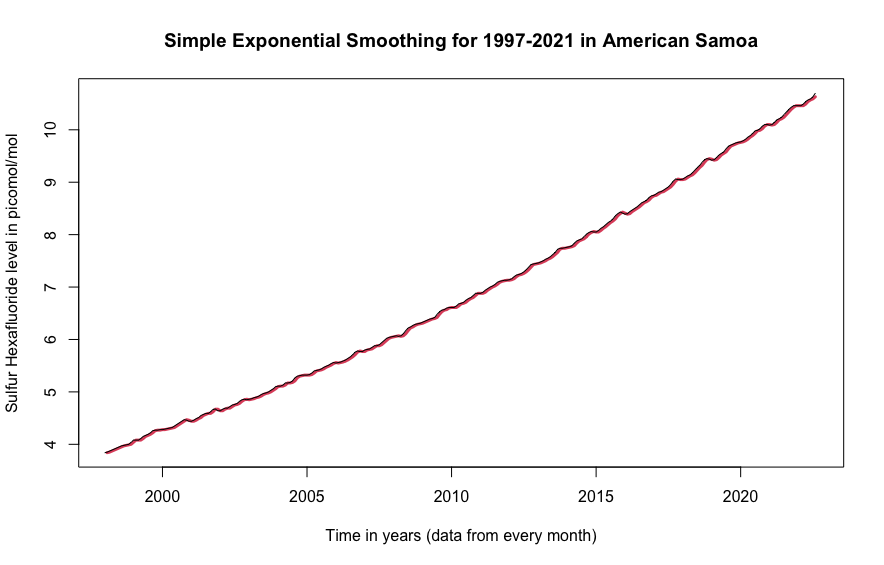


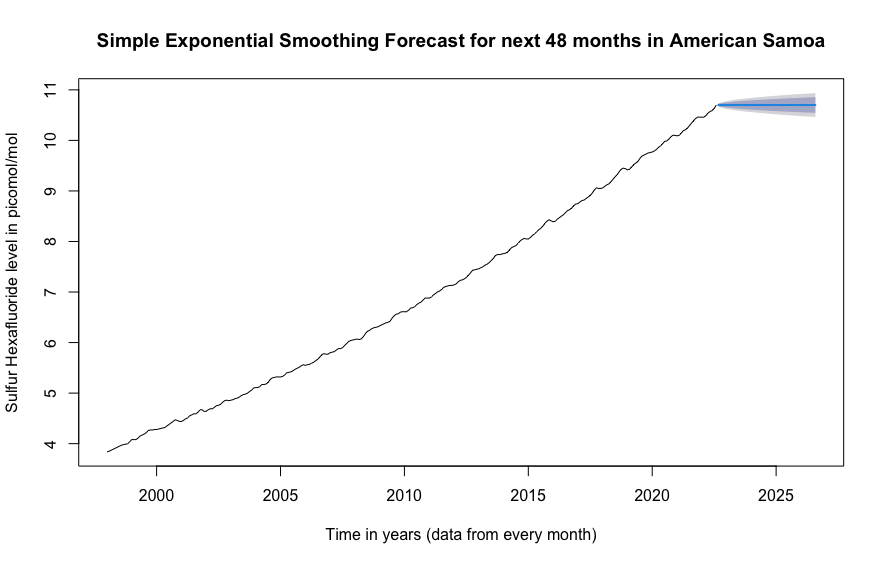


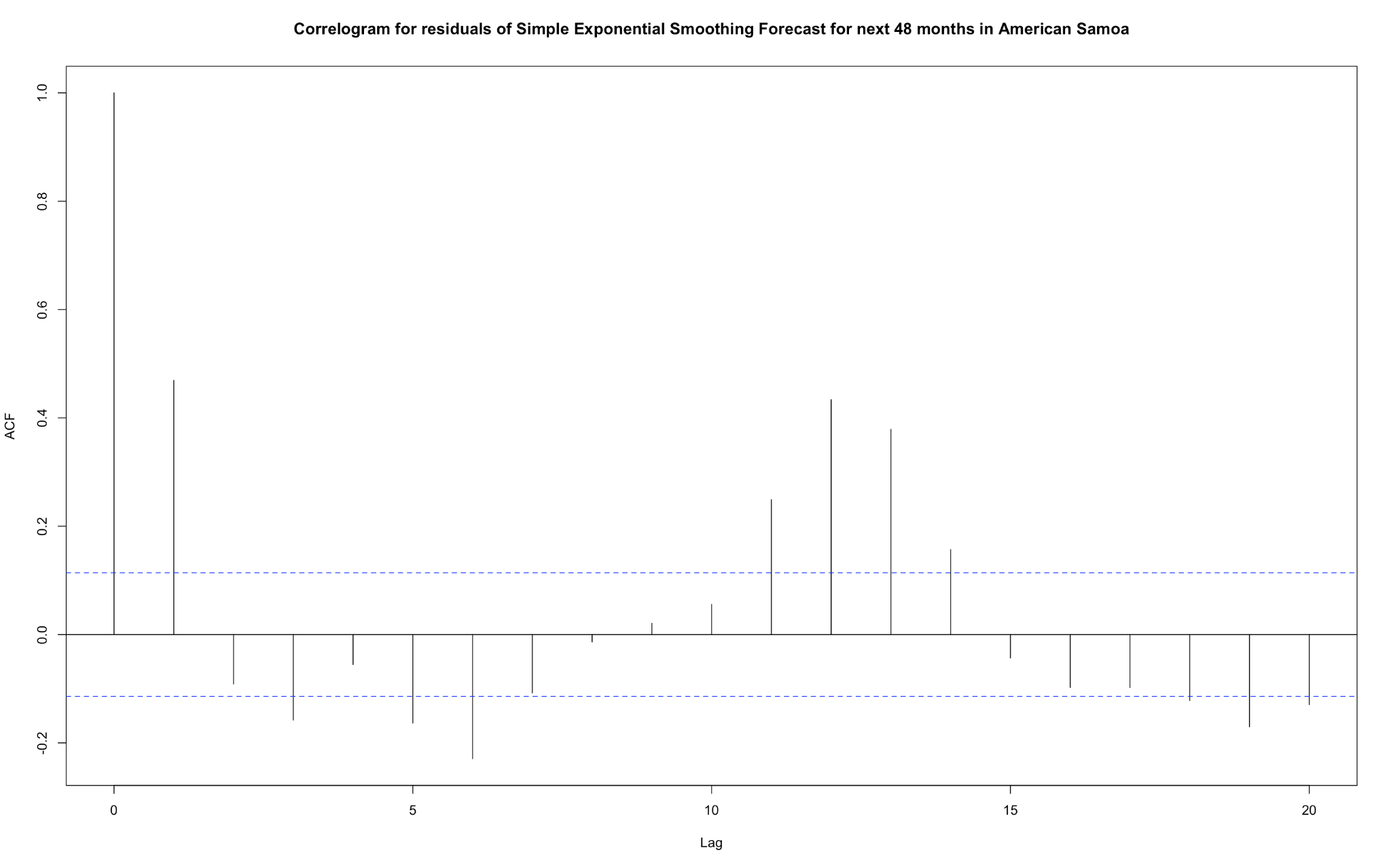


First thing of note, Model 1 has an R-Squared value of 1, which is the best possible value. Model 2 however, has an R-Squared value of .9998 which is still really good but not as near perfect as Model 1 is. The MAE and MAPE of Model 1 are also both smaller then Model 2’s, meaning that its acceptable accuracy is higher. The MSE of Model 1 is also smaller than Model 2 which is better. This is because when there is a smaller MSE, it indicates that the data is dispersed more closely around the mean. For the residuals plot however, Model 2 shows more wider white noise, but Model 1 also shows a greater amount of white noise, just over a smaller interval of the residuals. The ACF for Model 1 is also closer to the blue lines compared to the ACF for Model 2, which is better. For the Box-Ljung test, Model 1 did not provide a significant value meaning there is no autocorrelation between lags 1-20, making the data suitable for modeling. Model 2, failed. Overall, in almost every single metric Model 1 is the best fit.

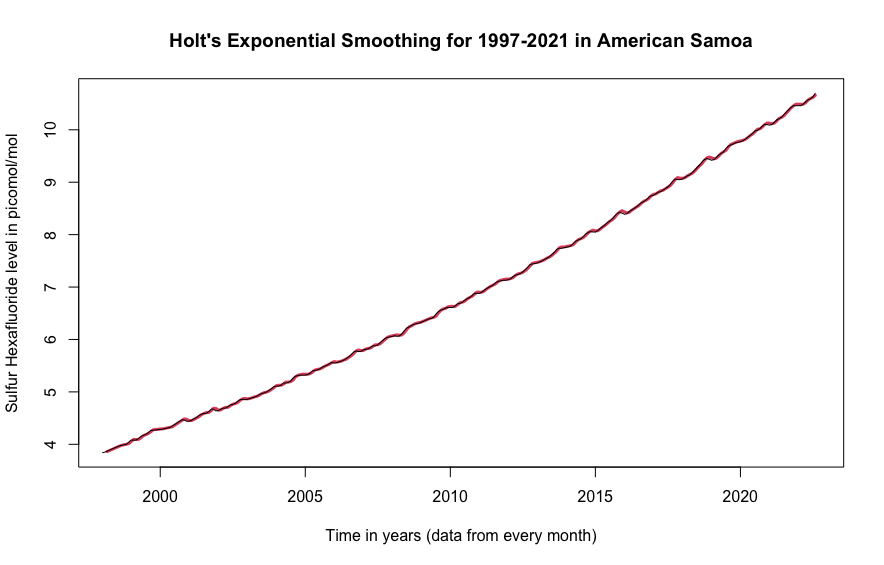
## Exponential Smoothing

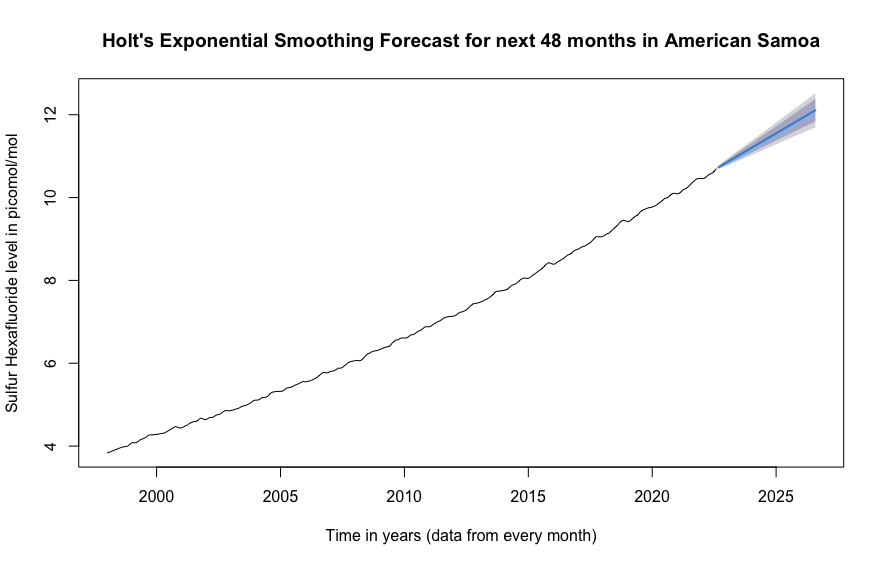
Smoothing is also known as averaging and usually helps in seeing patterns in time series. The smoothing method deals with two different types of moving averages, equally weighted and exponentially weighted. For this data, I tried three different types of smoothing: simple exponential smoothing, Holt’s exponential smoothing, and Holt-Winters exponential smoothing. Before beginning the actual calculation, I know which models are most ideal for our set. Looking at our decomposition from the data exploration section I see that seasonality is present, we have a constant level, and an additive model. I can predict that simple exponential smoothing likely will not suffice since it does not account for seasonality well. I can predict that Holt’s exponential smoothing also will not perform well because, like simple exponential smoothing, it does not account for seasonality well. I predict that the Holt-Winters exponential smoothing will have the most success of these three smoothing methods since it does factor in trend and seasonality in its method. Beginning with simple exponential smoothing:

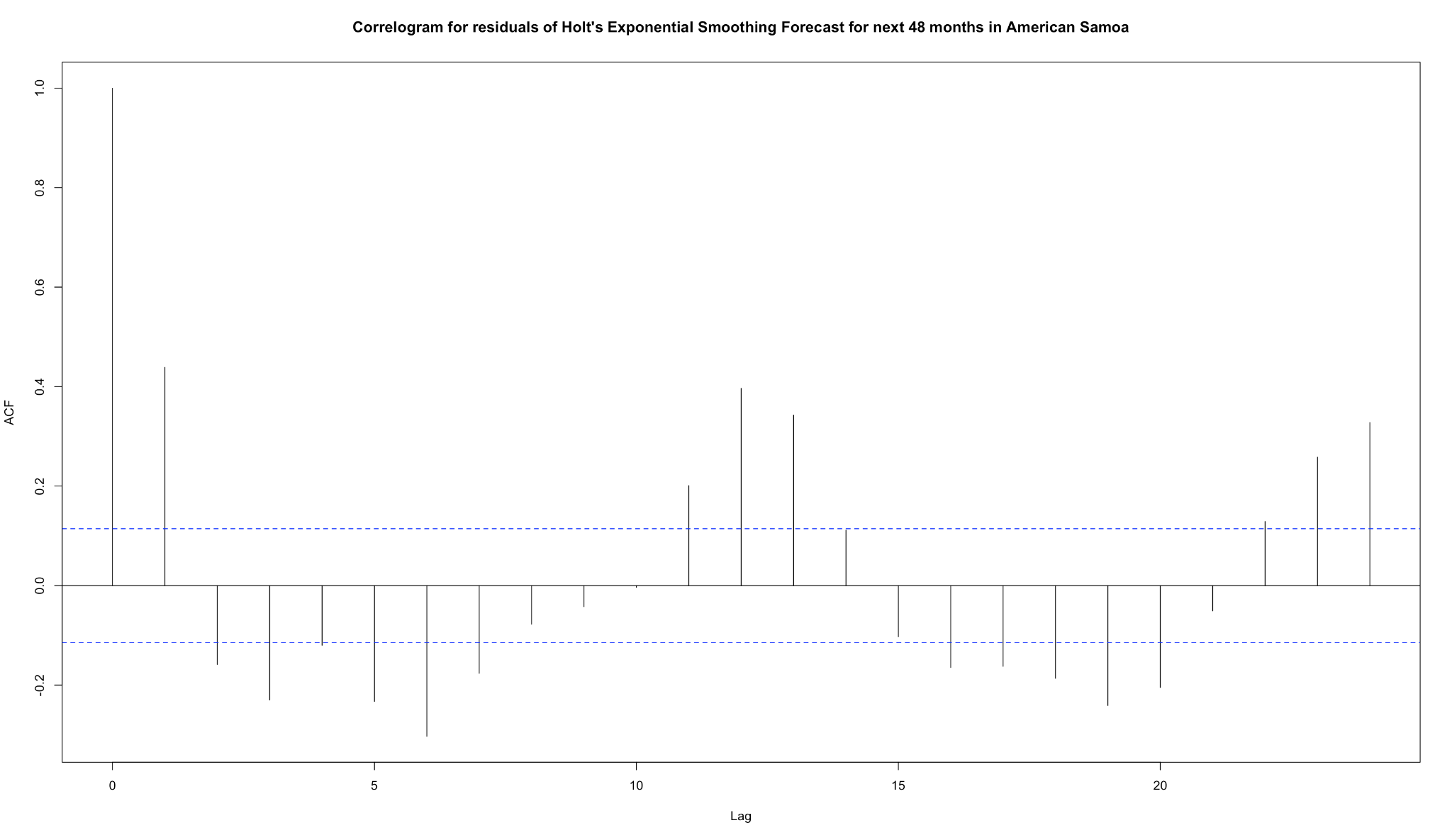




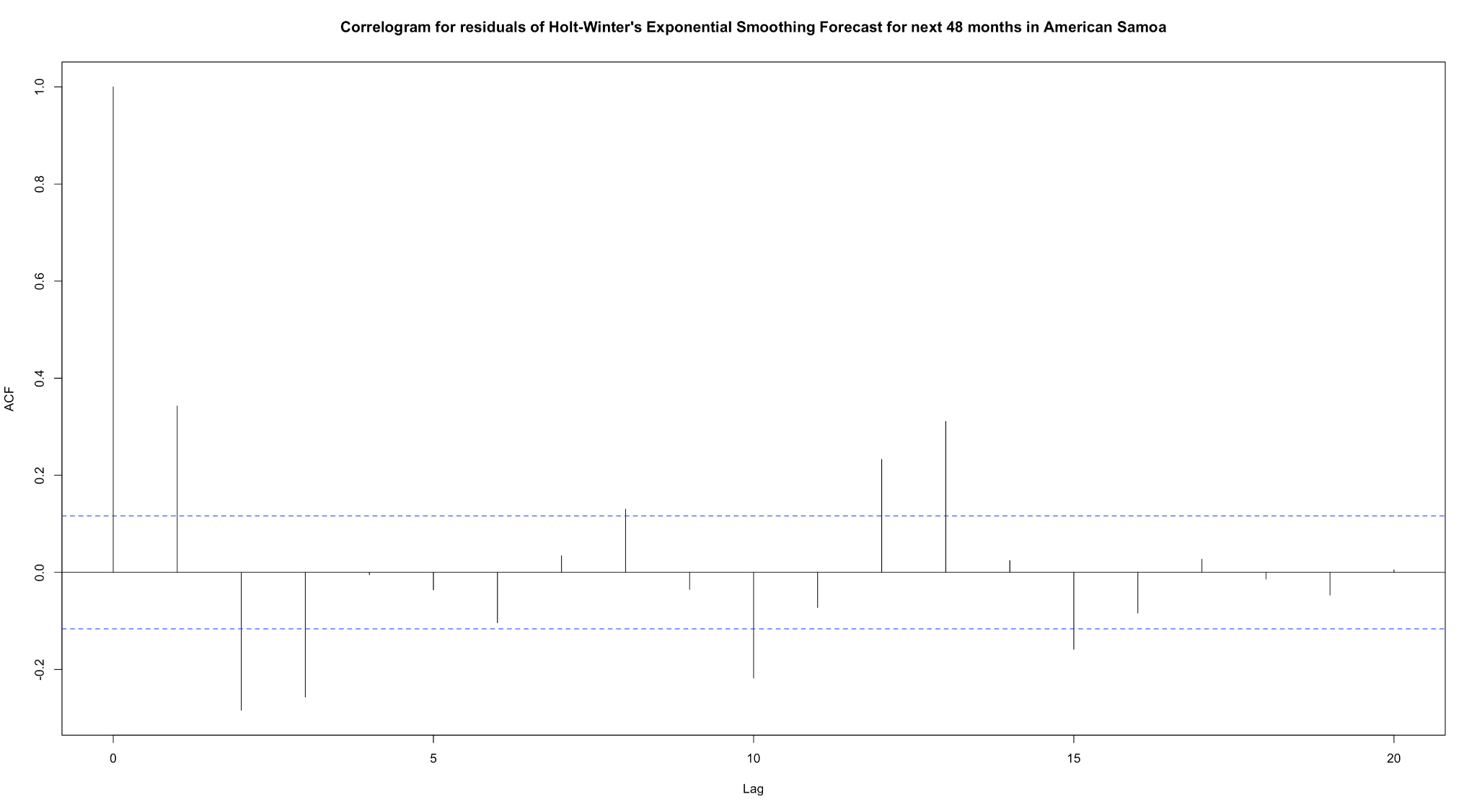
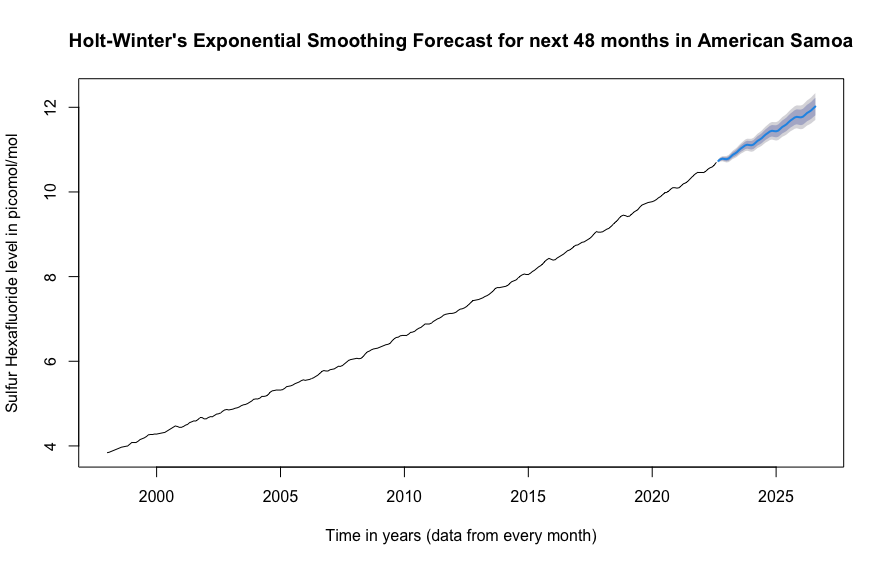
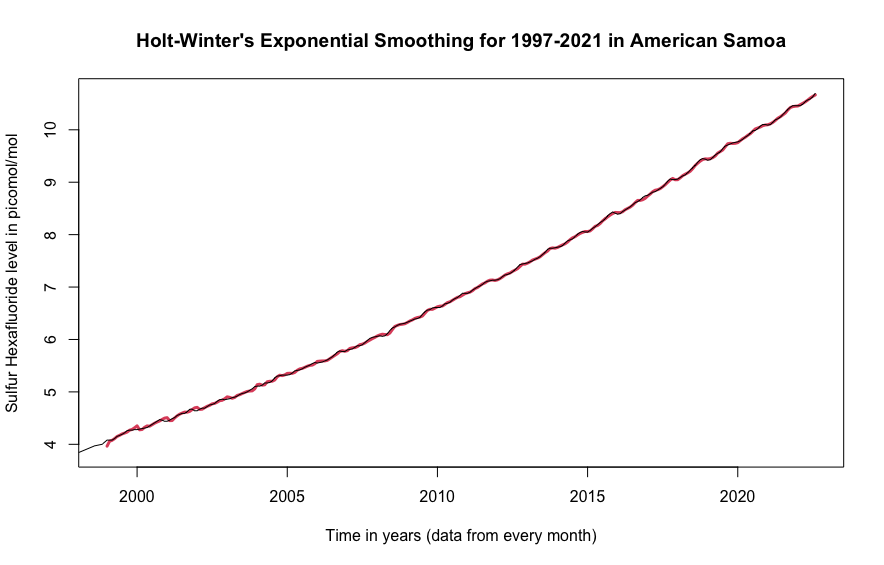
I have three figures above, the model for simple exponential smoothing for our data, the forecast of the next 48 months for our data using the simple exponential smoothing model, and the correlogram for the autocorrelation of the residuals for the forecast. In the output for the HoltWinters() function, I see that there is a high alpha value indicating the estimate of the level at the current time point is based upon recent observations. Strictly looking at the appearance of the model it looks usable. The model is incredibly close to the actual data from the look. Once I look at the forecast and correlogram, I immediately find out this model is actually useless. From a mere look at the forecast, I can make an educated guess that the trend is not accounted for in the prediction. From looking at the correlogram, I see that there is significantly large lag at lags 1, 6, 11, 12, and 13, with a minor lag at lag 19. This isn’t necessarily enough to deem this model unfit. To solidify that this model is not a good fit I turn to the Box-Ljung test. Its p-value is the probability of getting a value as large or larger than that observed under the null hypothesis that the true residuals are independent, where a significant value represents the presence of autocorrelation and a non-significant value represents the contrary. The Box-Ljung’s value gets larger as the sample auto-correlations of the residuals get larger, meaning the lower the value the more accurate fit. Using this test, I get an value of 260.8 and I get a p-value of 0, indicating there is high evidence of non-zero autocorrelation in the in-sample forecast errors at lags 1-20, therefore this is not an adequate model. Measuring the MSE, MAE, and MAPE, I find values of 8.3280008, 1.20716, and 0.1195208. The MAPE of 11.95% is a solid number, but I look to get a better one. The next model, Holt’s exponential smoothing:





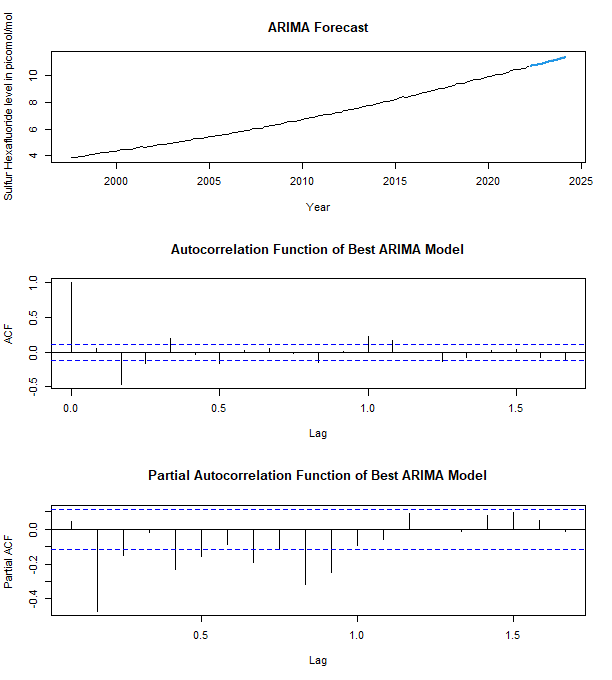


I have three figures above, the model for Holt’s exponential smoothing for our data, the forecast of the next 48 months for our data using Holt’s exponential smoothing model, and the correlogram for the autocorrelation of the residuals for the forecast. Like our first smoothing model, to an amateur strictly looking at the appearance of the model it looks usable. In the output for the HoltWinters() function, I see that there is an extremely high alpha value indicating the estimate of the level at the current time point is based upon recent observations, and I see that I have an extremely low beta value indicating that the slope is based on recent and not recent observations. Using the eye test on the forecast, it seems plausible that the true future values could be within the confidence interval cone, however, with a quick glance at the correlogram of the residuals for this forecast, this model is not usable. There are even more significantly large lags at multiple points, notably lags 1, 6, 12, 13, 19, 23, and 24. To further assure that this model is not useful in predicting the future sulfur hexafluoride levels, I check the Box-Ljung test and find an value of 304.56, larger than our previous smoothing model, and a p-value of 0 again, indicating there is high evidence of non-zero autocorrelation in the in-sample forecast errors at lags 1-20, therefore this is not an adequate model. Looking at the MSE, MAE, and MAPE again, I measure values of 8.537286, 1.05166, and 0.1021833. The MAPE and MSE in this model are better than the Simple Exponential Smoothing, however, the MSE is worse, which is surprising considering this model at least accounts for the trend. To the final exponential smoothing model, Holt-Winter’s exponential smoothing:



I again have three figures above, the model for Holt-Winter’s exponential smoothing for our data, the forecast of the next 48 months for our data using Holt-Winter’s exponential smoothing model, and the correlogram for the autocorrelation of the residuals for the forecast. Similar to the two previous models, to the naked eye the model is almost the same as the real data. The output for the HoltWinters() function shows a relatively high alpha value (though lower than the other two models) indicating the estimate of the level at the current time point is based upon recent observations, an extremely low beta value indicating that the slope is based on recent and not recent observations, and an extremely high gamma value indicating the estimate of the seasonal component at the current time point is based upon recent observations. Out of all three exponential smoothing methods, the forecast with this model looks the best by far. Our prediction that Holt-Winter’s method would produce the best results of the three smoothing methods was correct, however, though the forecast appears to represent the data well, the correlogram once again proves us wrong. This correlogram is not quite like the others though, it does have large lag spikes at lags 2, 3, 4, 10, and 13, but the values of these lags are smaller than those of the previous correlograms. There is potential in this model since this correlogram isn’t too extreme. Box-Ljung p-value: 0 again. value: 197.96. Though this model is the best fit by value of the smoothing models, our p-value indicates that there is high evidence of non-zero autocorrelation in the in-sample forecast errors at lags 1-20. MSE, MAE, MAPE: 8.563392, 1.060363, 0.1029592. The worst MSE among the smoothing models, the second worst MAE, and the second worst MAPE. The correlogram is not as inaccurate, the is the best of the three models, but the smoothing method of forecasting does not provide a good forecast for this time series.

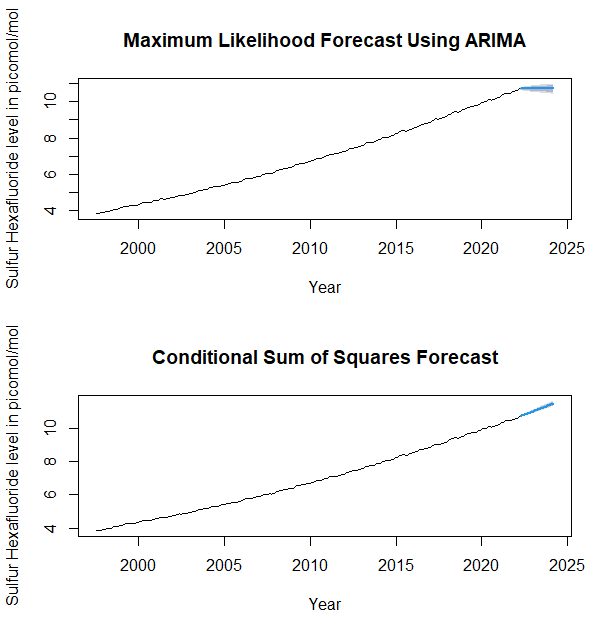
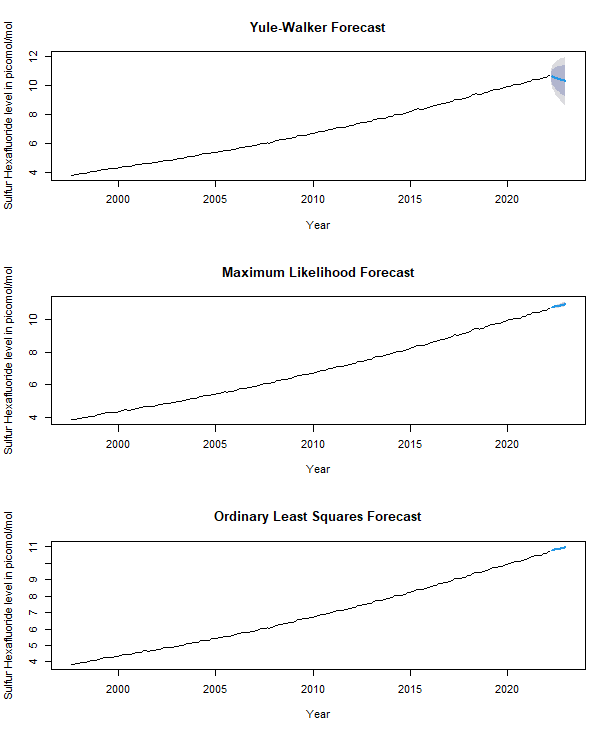
## ARIMA Forecasting

An autoregressive integrated moving average, also known as ARIMA, is another method that I can use to create a forecast model for the data. This method focuses on past values and their lagged averages to smooth/predict future values. Since ARIMA is defined for stationary time series, if the data set is not stationary, the time series must be differenced until I obtain a stationary time series. By using the auto.arima() function and analyzing the data, I came up with 3 possible candidate models to best forecast future values. From there, I analyzed the results of candidate models and chose the option with the highest log likelihood value (874.95) and lowest aic value (-1733.91). The data I have regarding our best model is as follows:

I can observe that the forecasting line follows similarly to the data leading up to it. Additionally, its variance (0.0001494) is also relatively small, meaning that the predictions that this model makes do not appear to vary significantly, making the forecasting values seem more accurate.

Despite how accurate it may look and its quality of having the best log-likelihood and aic values of the candidate models, this model still proves to be unusable after observing the autocorrelation function and partial autocorrelation function graphs. I can see that the lag values 2, 3, 4, 6, and 12 on the autocorrelation graph indicate non-zero autocorrelation in the in-sample forecast errors. Similarly, there are several values on the partial autocorrelation graph that indicate that there is non-zero partial autocorrelation in the in-sample forecast errors. I also conducted the Box-Ljung test for this model and found the value to be 10.411, giving us a p-value of .001253, confirming our assertions based on the ACF and PACF graphs. Even our best ARIMA model gave us an MSE of 9.578, an MAE of 1.214, and a MAPE of 0.111, which is not nearly as good as some of our other models, such as TSR. As a result, I must conclude that the ARIMA forecasting method is inadequate for predicting future sulfur hexafluoride levels.

Other methods of parameter estimation were also explored, but none of them yielded a model that was actually usable. I created fitted forecasting models using the AR method with Yule-Walker, MLE, and Ordinary Least Squares. Additionally, I fitted AR using ARIMA with Maximum Likelihood and Conditional Sum of Squares. The first three models using just AR were inadequate because the Yule-Walker method clearly did not follow the trend of the data set, and the ML and OLS methods had orders of 12 and 21 respectively, which proved to be impractical and unreliable for forecasting. Additionally, fitting AR with ARIMA did not provide viable results. Maximum Likelihood forecasting encountered a similar problem with Yule-Walker, where the forecasted values clearly do not follow the trend line. While the CSS method does not have any immediately apparent problems, I can see that the log-likelihood value is worse than our original candidate ARIMA model, therefore none of these models proved to be appropriate for forecasting. These extra models are appended below:



| Name | Box-Ljung | MSE | MAE | MAPE |
| --- | --- | --- | --- | --- |
| Simple Exponential Smoothing | 0 | 8.328008 | 1.20716 | 0.1195208 |
| Holt’s Exponential Smoothing | 0 | 8.537286 | 1.05166 | 0.1021833 |
| Holt-Winter’s Exponential Smoothing | 0 | 8.563392 | 1.060363 | 0.1029591 |
| TSR Model 1 | 0.3876 | 8.816676 | 0.8571599 | 0.08171209 |
| TSR Model 2 | 0 | 8.944147 | 0.8955404 | 0.08477783 |
| ARIMA(0,1,5)(1,1,2)[12] | 0 | 9.547458 | 1.204553 | 0.1110157 |
| ARIMA(2,2,5) | 0 | 9.649437 | 1.229294 | 0.1128462 |
| ARIMA(5,2,2) | 0 | 9.577753 | 1.214058 | 0.111757 |

# 

# Conclusion

After exhausting all methods of model creation and forecasting, I determined Model 1 of time series regression best forecasts the future sulfur hexafluoride levels in American Samoa. The model features the lowest MSE, the lowest MAPE, and the lowest MAE, passes the Box-Ljung test in that there is no autocorrelation among values from lags 1-20, and has the best ACF compared to all other models in the report. I conclude that the levels of sulfur hexafluoride will steadily increase for years to come. This has grave implications for the future of the island territory. Sulfur hexafluoride’s rapid consumption of the ozone is exactly what the ASEPA warned us of, and consequently, there can be rising sea levels around the island territory that can wipe out natural habitats, small towns, dilute freshwater aquifers, and destroy most basic civilization across the island. To combat these rising levels there are numerous deterring measures that can be taken, including basic recycling, modifying the canned tuna factories on the island so they produce less harmful greenhouse gasses, etc. The ASEPA was ahead of the curve in their warning according to our model and there is still time to turn around the future of American Samoa.

# 

# References

## Works Cited

*American Samoa imports and exports*. CountryReports. (n.d.). Retrieved December 11, 2022, from https://www.countryreports.org/country/AmericanSamoa/imports-exports.htm

*Fifth assessment report*. IPCC. (n.d.). Retrieved December 11, 2022, from https://www.ipcc.ch/assessment-report/ar5/

*What is SF6? Sulphur hexafluoride explained*. National Grid Group. (n.d.). Retrieved December 11, 2022, from https://www.nationalgrid.com/stories/energy-explained/what-is-sf6-sulphur-hexafluoride-explained#:~:text=What%20is%20SF6%20greenhouse%20gas,alter%20its%20state%20and%20effectiveness

