PSTAT174: TIME SERIES ANALYSIS

RISING TEMPERATURES IN AUSTIN, TEXAS 2013-2017

Collaboration

Dang, Phong

- In charge of drafting "Application Background Section and Research" as well as "Technical Review and Time Series Methodology".

Souksamlane, Jalen

- In charge of drafting "Data Explorations" and "Conclusion". Carried out coding for initial review of the dataset.

Sun, Rachel

- In charge of drafting "Hypothesis" and "Analysis of Time Series Data (Low Temperatures)". Carried out coding for low temperature analysis and prediction.

Yabut, Lorenzo Pineda

- In charge of drafting "Nature of the Data and the Data Collection Process" and "Analysis of Time Series Data (Diurnal Temperature)". Carried out coding for diurnal temperature range analysis and prediction.

With the submission of this paper, all group members are aware of their contributions and approved this for evaluation.

APPLICATION BACKGROUND

In our project, the usage of time series analysis and modeling applies to environmental science. Known as “environmental statistics” (Yue, 2011), it is the “[usage] of statistics as a means to solve the problems in various environmental fields.” Environmental science, as discussed by environmentalist Yue Rong in *Environmental Statistics* (2011), is a “relatively new field which stemmed from the industrialization in recent human history.” As this science has evolved from its early stages, it has now become a vastly multidisciplinary field that ranges from environmental engineering, sciences, public health, environmental studies, environmental law and economics, urban planning, and studies dealing with regional to global problems. Rong argues that when statistics is involved and applied to an environmental field, one would have to contextualize and characterize the environmental data. In following their logic, Rong argues that statistical data without adequate context can be misleading; therefore, in the field of environmental science, the practice of statistics as utilized by professionals to solve real world environmental problems. Thus, the usage of descriptive statistics is streamlined to describe the essence and characteristics rather than making inferences. Its counterpart, inferential statistics, is then used to test hypothesis, speculate, and predict.

In *Statistics for Environmental Science and Management* (2008), Bryan F. J. Manly argues that the three broad types of situations that environmental scientists and resource managers are interested in are baseline studies, targeted studies, and regular monitoring (2008). Baseline studies is “intended to document the present state of the environment in order to establish future changes.” Targeted studies is “designed to assess the impact of planned events […] or accidents.” Lastly, regular monitoring is “intended to detect trends and changes in important variables.”

As environmental statistics developed, the *United Nations Framework of the Development of Environment Statistics* (2013) established a framework that standardizes the scope of environmental statistics as surveying “biophysical aspects of the environment and those aspects of the socioeconomic system that directly influence and interact with the environment.” Noted was the interconnectedness of environmental, social and economic statistics whereby clear distinctions may not be necessary or possible. Commenting on the interconnectivity, the Department of Economic and Social Affairs of the United Nations stated, “Social and economic statistics that describe processes or activities with a direct impact on, or direction interaction with, the environment are used widely in environment statistics” and “other relevant social and economic statistics, which are not part of environmental statistics, are also required to place environmental issues in context.”

Nature of the Data and the Data Collection Process

The datasets being used are the daily temperature highs, lows, and the diurnal temperature range of the city of Austin, Texas in the United States from December 21, 2013 until July 31, 2017. This dataset is taken from Camp Mabry in Austin Texas (ICAO: KATT), which is a National Weather Service first order weather observation station for the greater Austin metropolitan area, although the dataset did not include the diurnal temperature range explicitly, it was calculated for this exploration. The data is collected by the same methods used by the National Weather Service. Since we are analyzing temperature data, the method for this collection was through Upper Air Sounding, by the use of an Aircraft, andAutomated Surface Observing Systems (ASOS). Upper Air Sounding is when they release a weather balloon into the upper atmosphere taking the air temperature, among other, measurements twice a day. Aircrafts take data from each flight and are used to forecast and improve models as they take temperature by use of an immersion thermometer probe. Lastly and most likely origin of the data comes from the ASOS, which is the primary surface weather observing network, which is constantly updated every minute of every day. It forecasts weather operations and is always taking observations through the network.

The datasets were chosen as representations of the last few years and prediction of future increases or decreases in temperature in the City of Austin. Specifically, the diurnal temperature range which is the daily highest temperature subtracted by the daily lowest temperature and the difference is taken, will be investigated. From "Diurnal Asymmetry To The Observed Global Warming", by Davy, Esau, Chernokulsky, Outten, and Zilitinkevich, published in 2016, this statistic is a better representation of the climate affecting the environment. As smaller ranges would lead to a decrease in vegetation (Alward *et al.,* 1999; Peng *et al.,* 2004, 2013; Zhou *et al.,* 2015) and has led to temperature-related fatalities (Yang *et al,* 2013). The difference between the daily high temperatures and low temperatures, has shown to be an indication of climate change and these statistics warrant investigation.

DATA EXPLORATION

There are two variables of interest within our dataset from kaggle: Temp.high.f and Temp.low.f. Additionally, we will use these two variables to create a new variable, Temp.d.f. Temp.high.f contains the highest recorded temperature per day. Temp.low.f contains the lowest recorded temperature per day. Temp.d.f. will be created by subtracting the data found in Temp.high.f by the data found in Temp.low.f, which will result in the daily diurnal temperature range.

Firstly, here is some basic analysis of the three variables. The mean, median, and mode of Temp.high.f is 80.86, 83, and 86 respectively. The highest recorded Temp.high.f value is 107 and the lowest recorded Temp.high.f value is 32. Temp.high.f also has a variance of 218.05 and standard deviation of 14.76. The mean, median, and mode of Temp.low.f is 59.9, 63, and 75 respectively. The highest recorded Temp.low.f value is 81 and the lowest recorded Temp.low.f value is 19. Temp.low.f also has a variance of 201.37 and standard deviation of 14.19. The mean, median, and mode of Temp.d.f is 20.96, 21, and 21 respectively. The highest Temp.d.f value is 45 and the lowest Temp.d.f value is 2. Temp.d.f also has a variance of 49.79 and standard deviation of 7.05. As we can see, both Temp.high.f and Temp.low.f have very high variance values. This makes sense as the daily temperature can change drastically due to the season changing (i.e. the temperatures recorded in the seasons of winter and fall will be much colder than those recorded during spring and summer). Interestingly, the median and mode of Temp.d.f are both 21 and the mean is very slightly lower than them. This can be indicative of Temp.d.f being close to a perfectly symmetrical distribution.

Secondly, by looking at the untransformed time series plots of the three variables, we might be able to come to some conclusions.

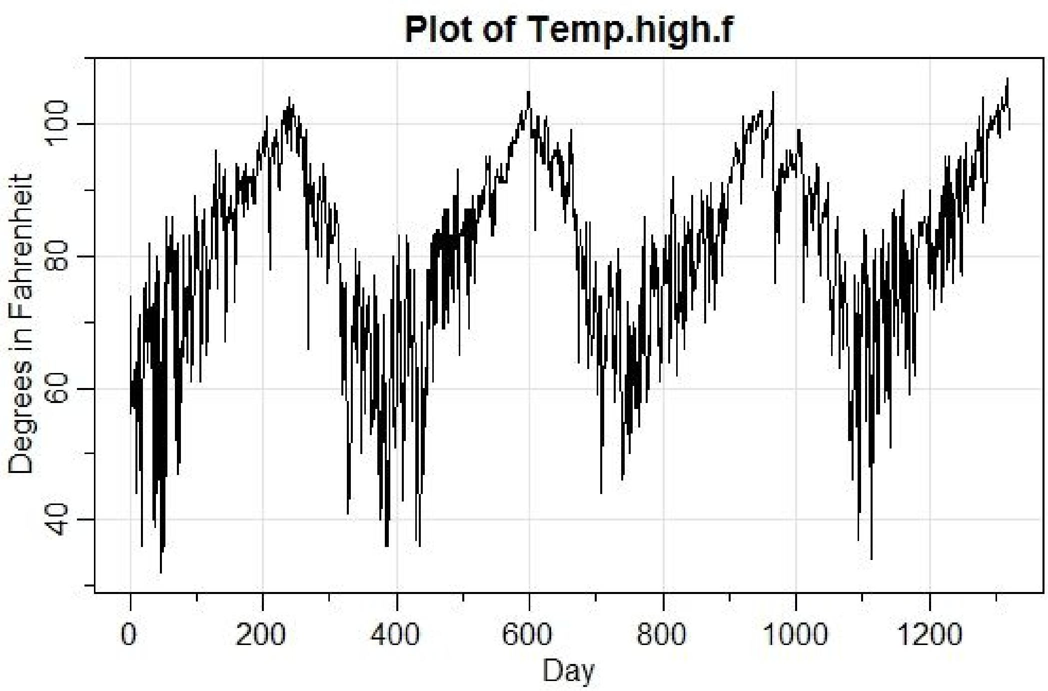

Figure 1

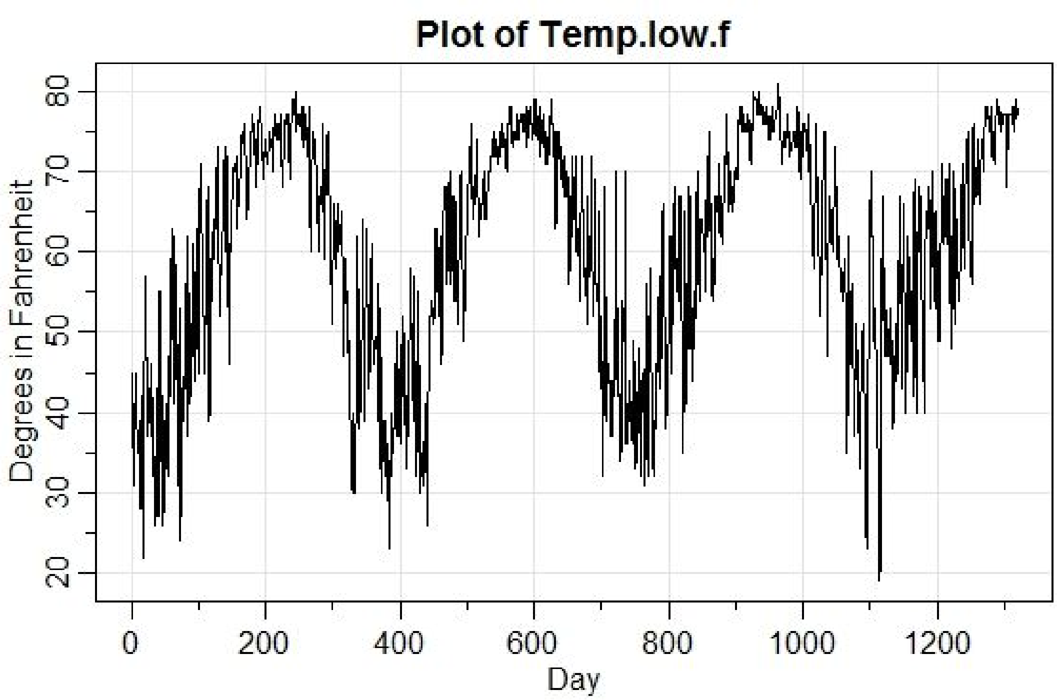

Figure 2

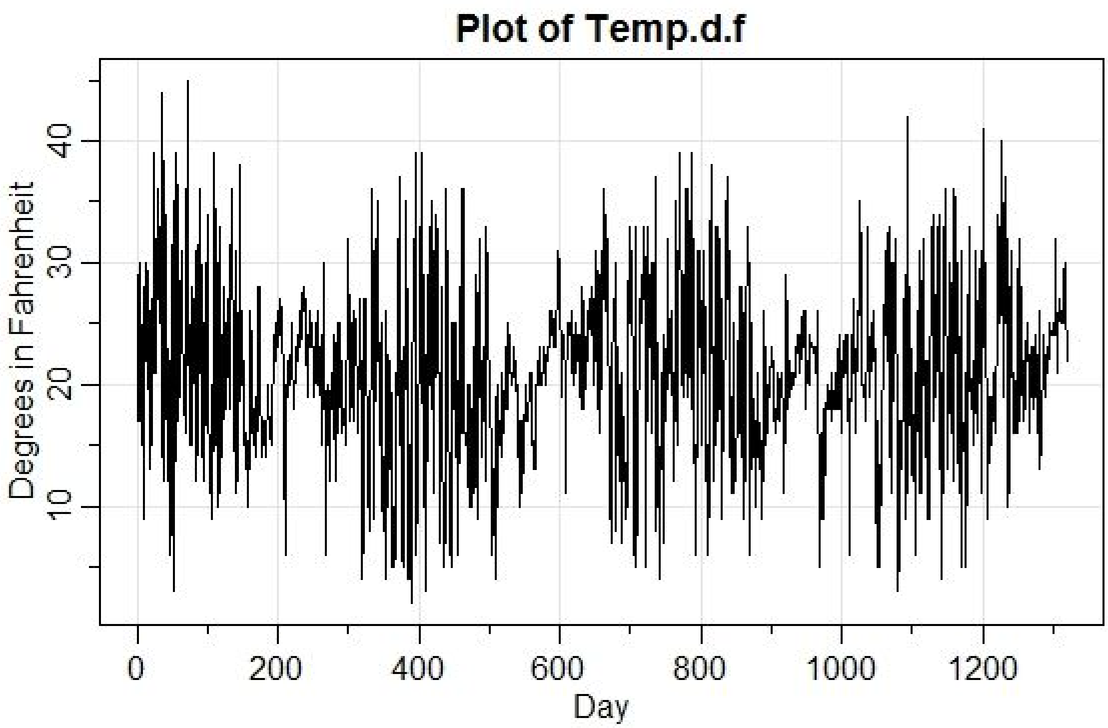

Figure 3

As we can see here, both Temp.high.f and Temp.low.f have heavy implications of seasonality. Seasonality when a time series exhibits predictable changes that recur every calendar year. This is because both plots appears to rise and fall in periods of around every 360 days. Furthermore, the period being around 360 days is indicative of temperature rising and falling throughout the calendar year, thus, showing daily seasonality in both plots. On the other hand, the plot of Temp.d.f is also interesting as it shows stationary behaviour.

HYPOTHESIS

Austin, Texas is notorious for its humid, subtropical weather patterns. As expected, the city's highest average temperatures accumulate around the peak summer time, June through August (Current Results, 2010). August typically ranks as Austin's hottest month, averaging around the high 90-degrees Fahrenheit. Additionally, the city's overall temperature has seen a gradual increase over the last decade. Over the last few years, Austin has experienced multiple days with weather over 100 degrees Fahrenheit. In 2019, the city underwent 19 consecutives days of triple digit temperatures. The last record with such an extensive streak was in 2011 when the bar was set at a whopping 27 days. Unfortunately, scientists are now speculating that 100-degree weather patterns will become more commonplace in the future, starting as early as 2036 (Bradshaw, 2019). Temperature growth Climate change and carbon emissions are the primary reasons for the incredible temperature growth. Perhaps not coincidentally, Texas is home to one of the country's largest economic contributors: the oil and petrochemical industry. The sheer size and power of these industries is one of the many reasons why air pollution is a predominant environmental issue in the state of Texas. Specifically, the city of Austin's air pollution continues to decline with every year that passes. The main contributor: tailpipe emissions. The city's average measure of ozone is anywhere between 63 and 65 parts per billion, depending on the region. According to the National Ambient Air Quality organization, that measurement should always stay below 76 parts per billion in order to maintain a safe living standard (Dugan, 2018). As the number inches closer to the threshold, more and more civilians are experiencing environmental issues such as rising atmospheric temperatures.

From the 2013 - 2017 meteorological dataset gathered from Kaggle, our team hypothesizes that the overall density of low temperatures during winter periods will become less and less frequent, in part due to air pollution and carbon emissions. Consequently, we also hypothesize that the overall gap between daily high and low temperatures will close more and more in the future. From exploratory data analysis and our initial review of the data, we found that when the raw measurements are plotted against time (in this case, the time scale is in terms of days) the plot shows less data points in the 10-degree to 40-degree Fahrenheit range during the winter periods after approximately day 700 (around 2 years after the first day of documentation). Eventually, we speculate that the difference between the highs and lows will become smaller and smaller as temperatures grow hotter and hotter.

TECHNICAL REVIEW AND TIME SERIES METHODOLOGY

We started our exploratory data analysis with a simple time series plot of the dataset "LowTemp" to look at its characteristics. Knowing that it was temperature data, we suspected from the beginning that it would be seasonal and the plot confirmed our expectations. Because of the seasonal nature of the data, it must also be non-stationary. According to Robert H. Shumway and David S. Stoffer in *Time Series Analysis and Its Application* (2010), "it is necessary for time series data to be stationary so that averaging lagged products over time [...] will be a sensible thing to do. With time series data, it is the dependence between the values of the series that is important to measure." Thus, it is our prerogative to estimate the autocorrelations and estimating dependence would be difficult if the "dependence structure [was] not regular or is changing at every point." To equalize the variability over the time span of the series, we attempted two different and common transformations: log and square root transformation. We then plotted the two transformations of the "LowTemp" time series and realized that it was similarly seasonal like the untransformed data. Since a transformation of the dataset did not work to deseasonalize the data, we decided to differenced the data. We chose to difference the data instead of detrending since we saw that the trend was not fixed. Our goal was to coerce the data into stationarity, and sure enough, after differencing the data once and plotting the time series plot, the data was stationary.

Next we opted to look at the autocorrelation (ACF) plot and the partial autocorrelation (PACF) plot of the different data to acquire more insight on our dataset. According to Shumway and Stoffer, the ACF measures the "linear predictability of the series at time t, say x_t, using only the value x_s" (2010). In our case, the ACF works to help us decide whether a moving average model of order q, MA(q), is an appropriate choice or not. A key characteristic of the MA(q) model is the assumption that the white noise terms are linear combinations that map out the dataset. Since the MA(q) model is a linear combination of white noise, the model is said to be stationary with mean 0. The ACF plot cuts off and becomes 0 after q amount of lags. Similarly, we use the partial autocorrelation (PACF) plot to predict the order of the autoregressive model, AR(p). The ACF dictates the order of dependence of a MA(q) model; alone, it cannot say much about a possible AR(p) or autoregressive moving average, ARMA model. According to Shumway and Stoffer, the tool we need to diagnose our dataset to see if there are any AR qualities is with the partial autocorrelation, because "the correlation between x_s and x_t with the linear effect of everything in the middle' removed." Thus, the PACF works to partial out the dependence of say, x_0, between x_s and x_t, to find the correlation of x_s and x_t. In looking at the ACF and PACF of our dataset, we brainstormed potential candidate models. We chose to work with integrated ARMA, or ARIMA, models since we had differenced our data.

The next step of the process was to look at the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In short, the smaller the AIC the better. The AIC works to penalize larger complicated models that may overfit the data. The BIC on the other hand, works similarly to AIC but with a different penalty term. Both will be used for model selection whereby we would pick the model that produces the lowest AIC and BIC.

After, we will perform multiple diagnostics on our chosen model. First we did the Shapiro-Wilk normality test to see if our residuals were normally distributed. Residuals should be normal as to avoid heteroscedasticity since our model should have consistent predictive ability throughout the range of our observed data. We set our acceptance threshold for the Shapiro-Wilk test to be 0.05. Therefore, any p-value less than that would mean that our residuals were normally distributed. The Ljung-Box test was used to check whether our residuals were independent. The residuals should be uncorrelated, since correlation of residuals would imply that there was extra information in the residuals that could have been used to forecast the model. Again we set our acceptance threshold for the Ljung-Box test to be 0.05. Any p-values under 0.05 would mean that the data was correlated. We also opted to do a histogram and a Q-Q plot of our model's residuals to graphically see their behavior.

The last part of our project involved making future predictions of our dataset. Using the model that we selected and diagnosed, we projected the next few data points. Needless to say, the confidence interval of the projected points spanned a massive range. Personally if this were a scholarly research article we would not include a prediction since the prediction was trivial given how large the confidence interval was. Another point to note is the fact that our prediction is irrelevant given that fact that we are predicting temperature in what would be the past, which have already happened. Regardless, for the purposes of this assignment, we included it for completeness.

ANALYSIS

LOW TEMPERATURES:

Our analysis includes a variety of methods used to achieve the final models for our dataset. First and foremost, all of our coding was done using R. As mentioned before in our Data Exploration section, the time-series plot of our dataset looks seasonal; there is a repeating cycle/pattern that occurs every year. The first dataset that will be analyzed is the daily low temperatures. In order to properly carry out our time-series analysis, this data first has to be differenced in order to remove seasonality. Once the data was deseasonalized, we plotted the ACF and PACF graphs. By plotting these graphs, we can determine our candidate models. The ACF and PACF plots are shown below:

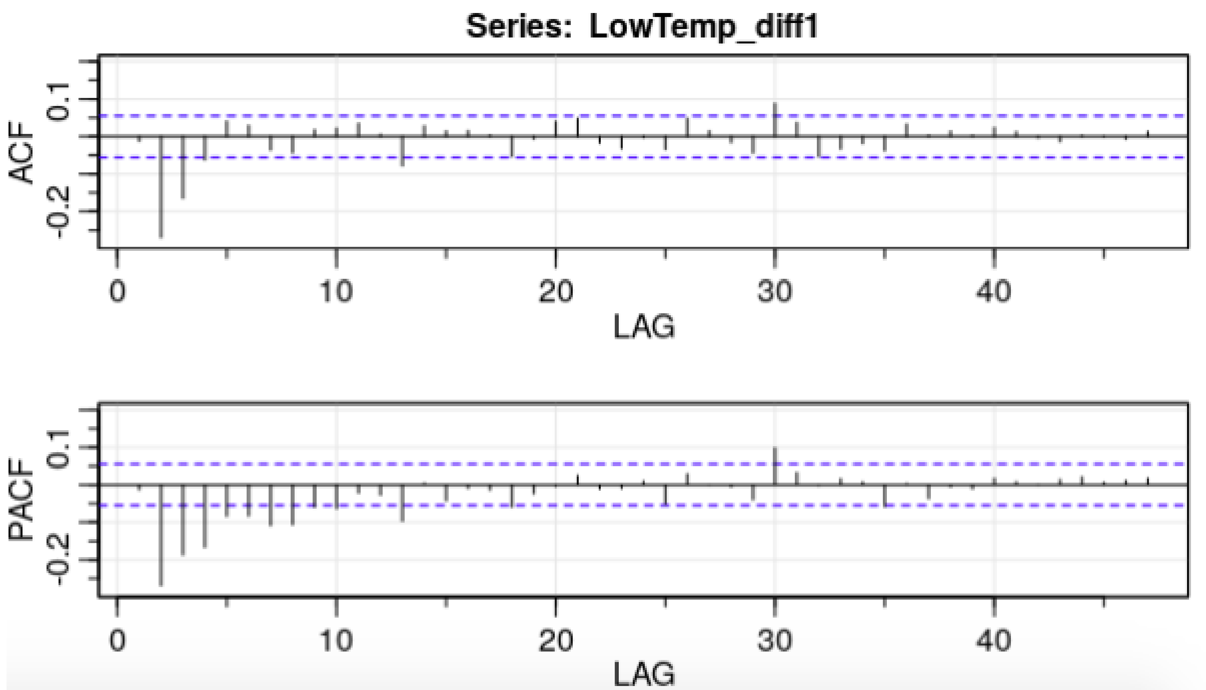

Figure 4

The graph of ACF looks to be significant up until lag 4. Meanwhile, the PACF exhibits a more geometric decay overall. We can deduce that an MA(4) model might fit this dataset the best. Another method of finding a candidate model is using function ar() in R. This function fits an autoregressive (AR) model to a given time series, and determines the best fit by using criterion AIC. Using the said function, we have AR(13) as another possible model. Lastly, we can determine models by consulting with the AIC of every model using a recursive method. By creating nested for-loops and providing ranges for lag p and lag q, we iterated through every possible model between lags 1-10. Afterwards, we grabbed the minimum AIC generated from the for loops; this gives a candidate model for ARMA. Our code returns ARIMA(6,1,7) as the model with the lowest AIC.

Next, we evaluate the three candidate models by reviewing their AIC and BIC values. The model with the lowest AIC and BIC will be our best model fit. In order to determine the AICs and BICs are the candidate models, we must use function sarima(). The parameters of sarima() include: (data, p, d, q, P, D, Q, S)

- p = AR order
- d = difference order
- q = MA order
- P = SAR order (using seasonal data)
- D = difference order (using seasonal data)
- Q = SMA order (using seasonal data)

We have the values for (p,d,q) - what we must determine are (P,D,Q). In order to find these values, we plot the ACF and PACF for the seasonal dataset (before differencing it); in other words, we carry out the same procedure that we did with determining models from ACF and PACF plots except we use the raw data.

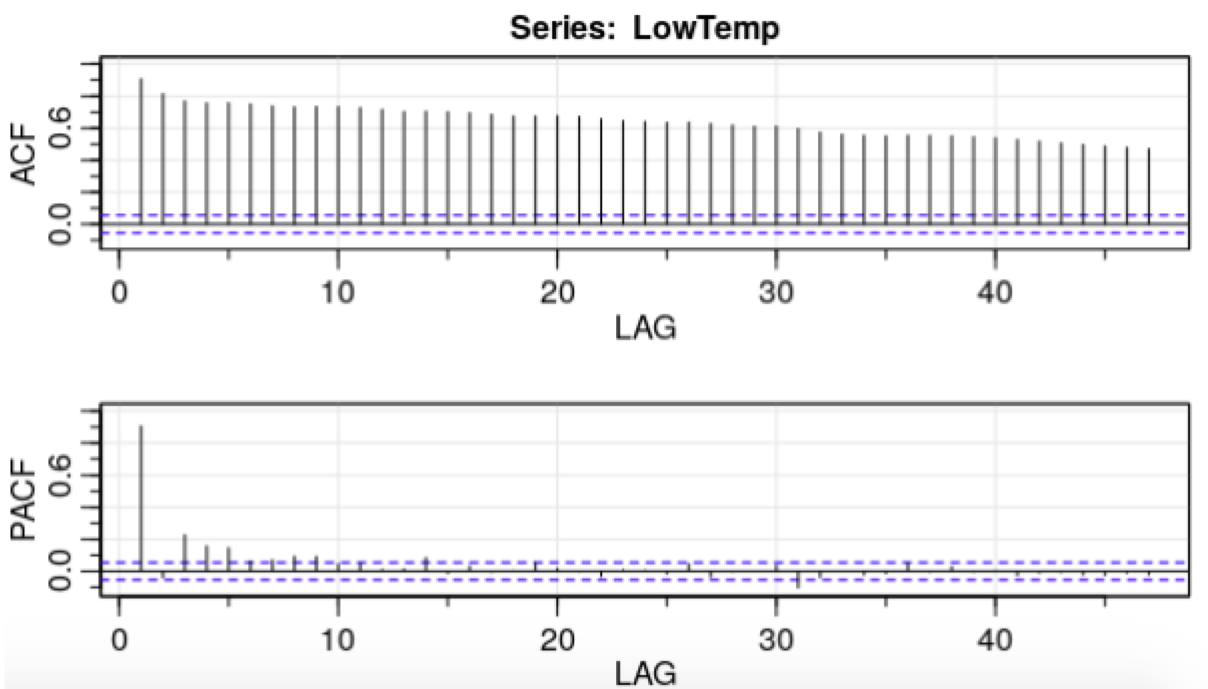

Figure 5

In Figure 5, we plotted the ACF and PACF of the raw dataset. We see that ACF has a geometric decay while PACF is significant up until lag 1. From this, we conclude that MA(1) is our best model for the seasonal dataset. Hence, our values for (P,D,Q) is (1,0,0) (since we didn't difference the data, D = 0).

Using the function sarima(), we found that the following AIC and BIC values for the three candidate models:

- MA(4): AIC = 6.300727 and BIC = 6.328276
- AR(13): AIC= 6.436217 and BIC = 6.499186
- ARMA(6,7): AIC = 6.298166 and BIC = 6.361135

ARMA(6,7) has the lowest AIC, but MA(4) has the lowest BIC. Therefore, we can eliminate AR(13) as a possible model. Next, we test the normality and independence of models MA(4) and ARMA(6,7). In order to test normality, we use the shapiro test. Using the function shapiro.test() in R, we found that neither of the models follow a normal distribution. Next, we test for independence; function box.test() can be used to carry out the hypothesis testing for that. The p-value obtained for MA(4) is 0.6131 and the value obtained for ARMA(6,7) is 0.9972. Using our default alpha value 0.05, when the p-value is above 0.05 the model is independent. Otherwise, it is not independent. Both p-values of MA(4) and ARMA(6,7) are well above our default alpha, so instead we will select the model with the higher p-value over the second choice. As a result, ARMA(6,7) is our best fit model.

Lastly, we use function sarima.for() to predict values beyond our time-scope. Figure 6 shows a 95% confidence interval for low temperature values, as well as predicted measurements 10 days after the last day of data collection.

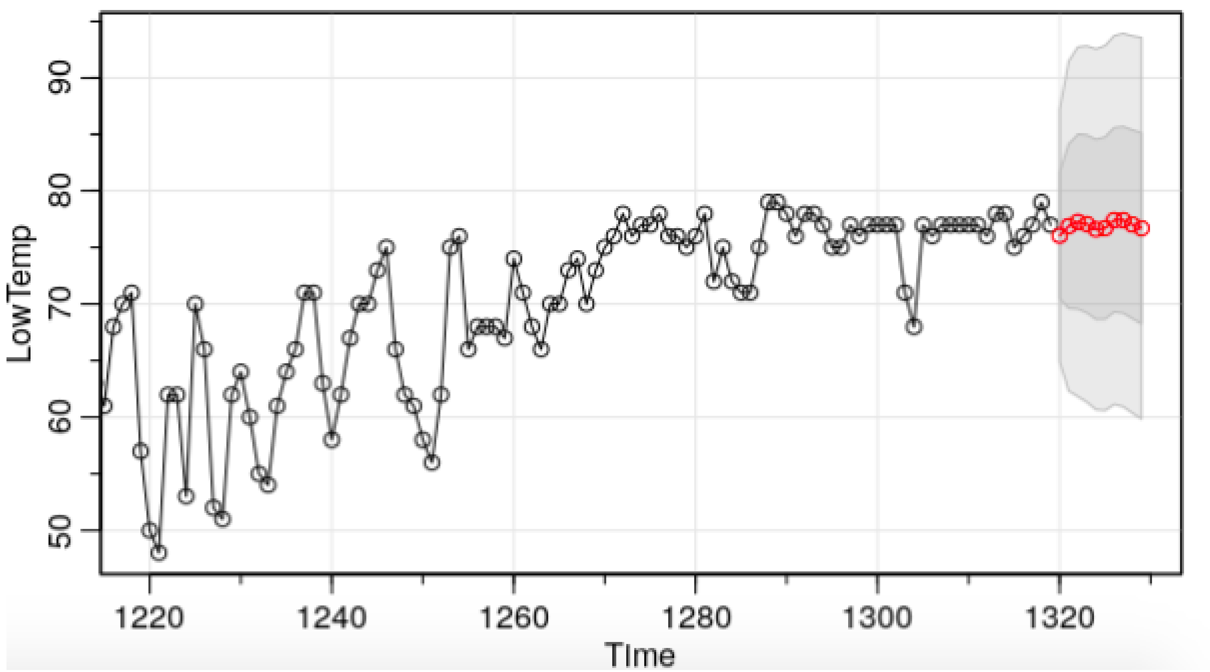

Figure 6

According to the graph and our predictions, we see an upward trend for low temperatures. Theoretically, this means that the diurnal temperature range should be decreasing as well.

Additionally, looking at the raw periodogram for low temperatures might help support the idea that low temperatures are trending upwards. Figure 7 shows the periodogram for the raw low temperature dataset.

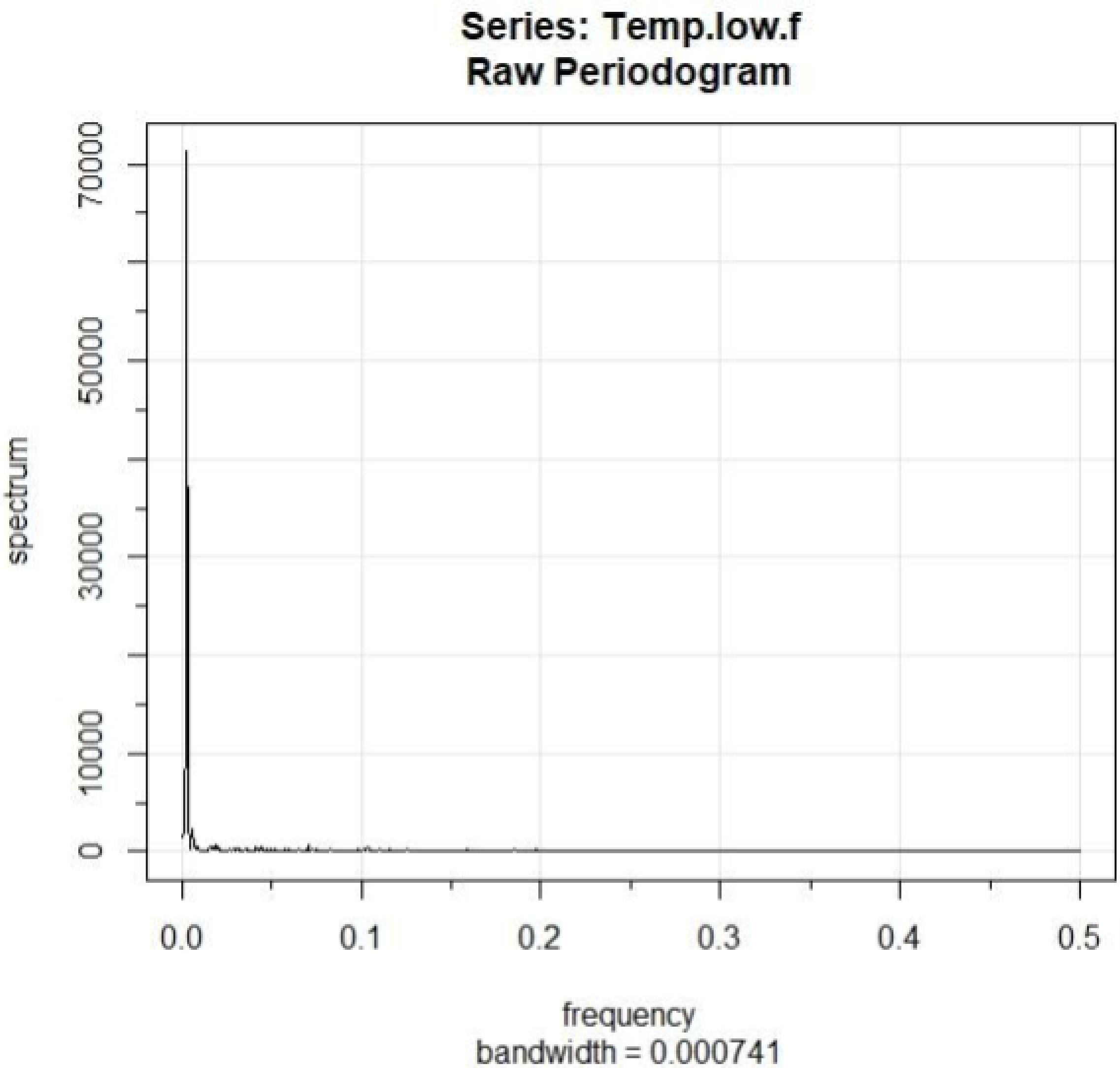

Figure 7

As you can see here, the periodogram peaks once very close to 0. This indicates that there is a strong sinusoidal signal within the low temperature data. In other words, this heavily implies that there is a cycle recurring within our data set, which can be explained by seasonality. By zooming into the plot, we came to the conclusion that the peak occurs at a frequency of 0.00296. To get the length of the cycle we divide 1 by the frequency of the peak which comes out to about 338 days. This signifies that temperatures are beginning to rise before the calendar year ends. Thus, the temperature during the year is trending upwards.

DIURNAL TEMPERATURE RANGE:

The time series of the diurnal temperature appeared to be somewhat stationarity but it still included seasonality. Thus, the time series must be differenced once. The differenced plot is shown in Figure 8.

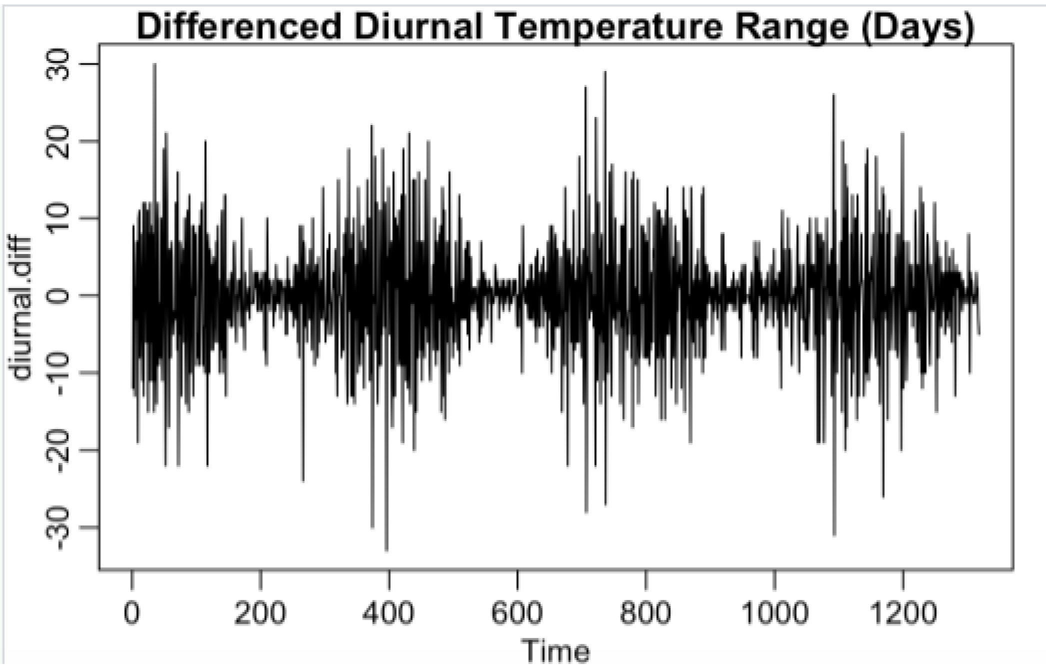

Figure 8

The ACF and PACF of this differenced time series can be shown below.

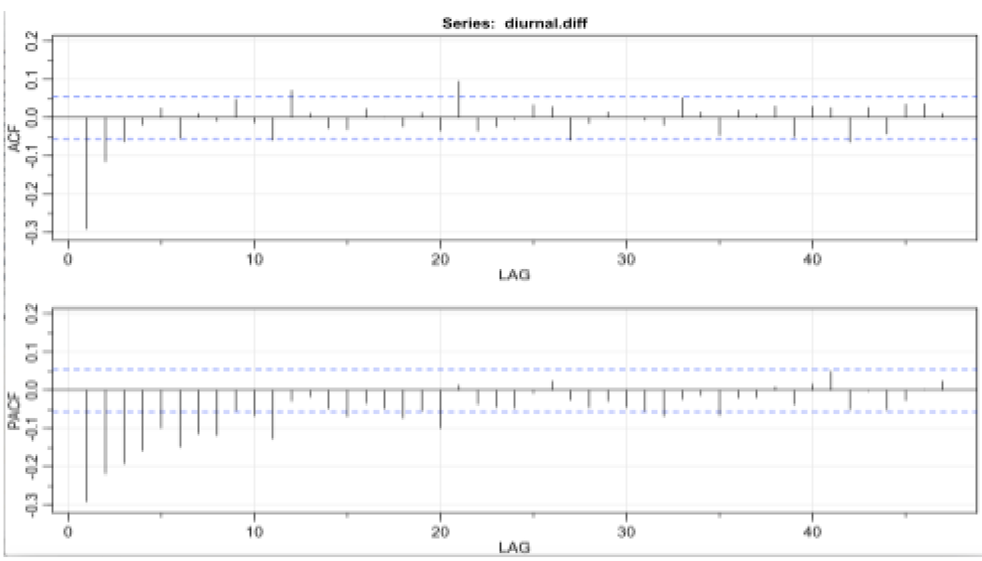

Figure 9

The ACF appears to cut off after a significant value on lag 3. The PACF seems geometrically decay. This could suggest a MA(3) model. Further, it also could be interpreted that the ACF of this figure is also a geometric decay. These would suggest candidate models appropriate for the data set. (AR model could be included here as another candidate model but I don't see it being viable) To check for the ARMA process we will fit it through the same method used in the previous dataset, and it suggested an ARMA(6,1,6) model for the differenced data set.

From these candidate models, one should be selected to be the final model to predict and evaluate the diurnal temperature dataset. Through the use of their AIC and BIC, as well as normality and independence of each candidate model, as done previously.

In order to fit these models appropriately we must consider the natural seasonality of temperature datasets. Again, the values for P and Q must be determined in order to fit the final SARIMA model. The ACF and PACF of diurnal temperature indicated that the seasonality of it should follow an AR(3) model for its P.

Using the function sarima(), the candidate model had the following AIC and BIC values:

- MA(3): AIC = 6.732 and BIC = 6.76
- ARMA(6,6): AIC = 6.637 and BIC = 6.7

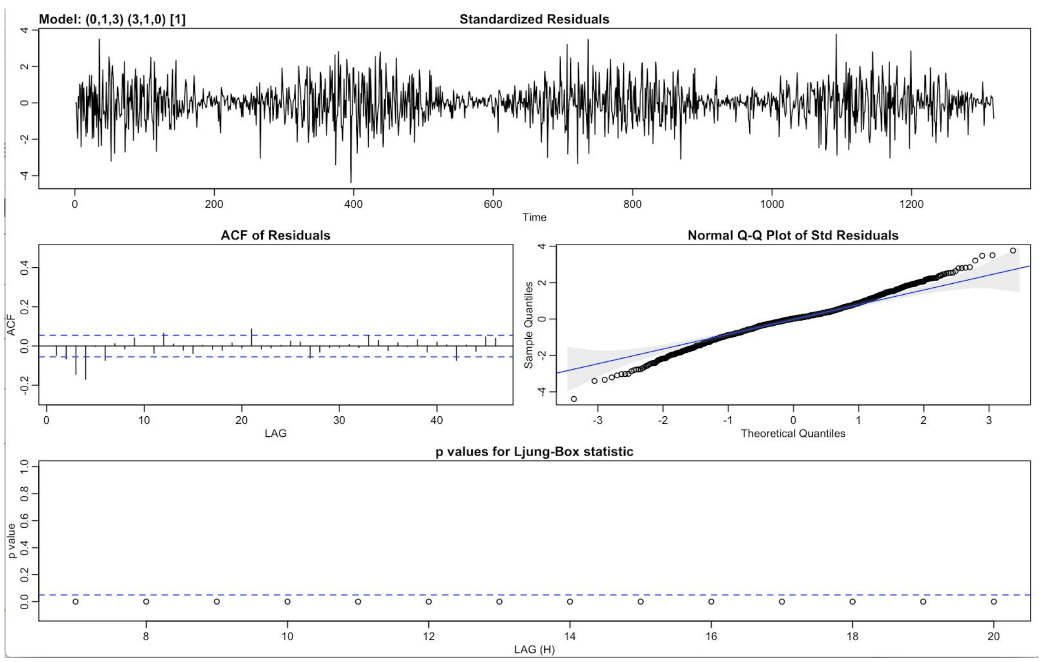

Figure 10

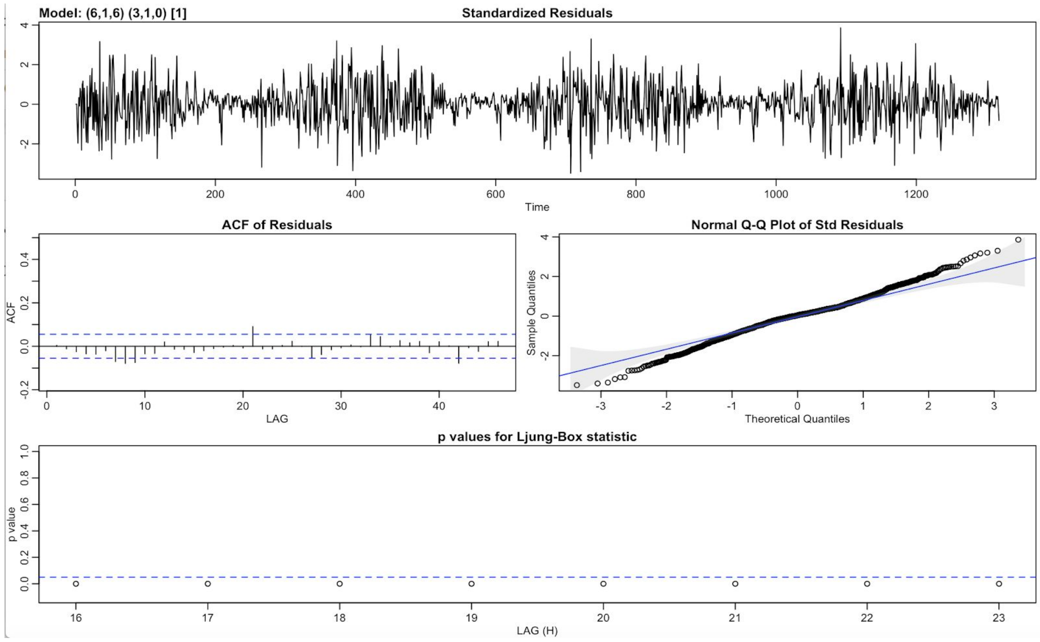

Figure 11

Looking at the plots the sarima() function provides, the SARIMA(6,1,6)(3,1,0) has the better AIC and BIC. This is selected as our prediction model.

We will now perform a prediction using the SARIMA(6,1,6)(3,1,0) model, predicting the next 10 days. Again through the use of sarima.for(), the figure shows a 95% confidence interval for the diurnal temperature range, as well as predicted measurements for the 10 days after the last day of data collection.

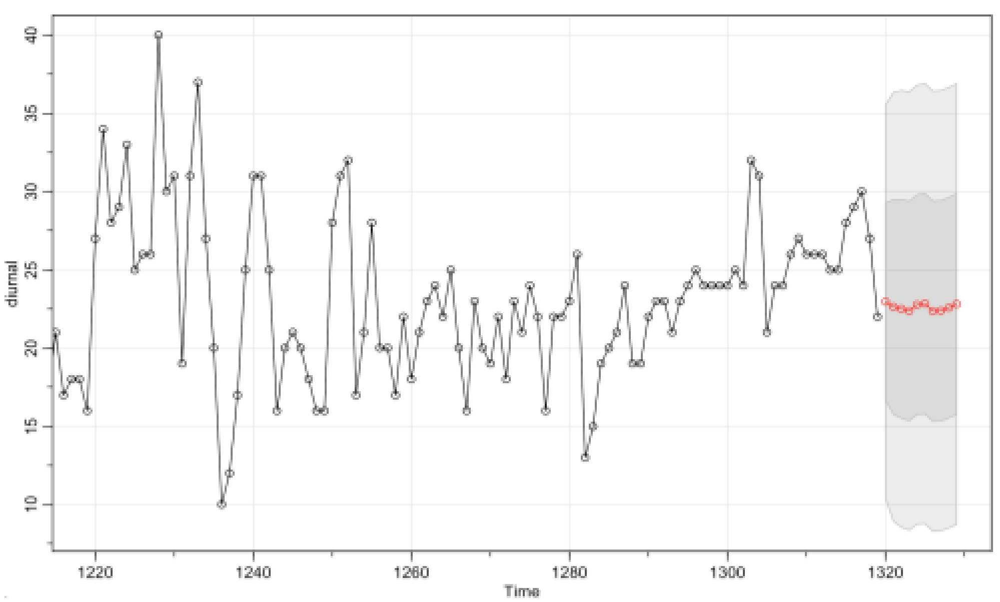

Figure 12

The prediction predicts the model should be hovering around a difference 22.5. Comparing this to the total average of the diurnal temperatures 20.96, which is higher than the average.

In summation, the model selected for the diurnal temperature range is not completely normal and may not be the ideal model to predict future data, as seen by the QQ-plot of the SARIMA model. It is, although, independent as the p-value for its residuals are less than 0.05.

Conclusion

Based on our analysis, we conclude that our hypothesis could be wrong. The daily low temperature is slowly rising over the years yet the diurnal temperature from the model is predicted to be higher than the average. This could stem from multiple possibilities as to why the diurnal temperature is actually increasing. Since the low temperatures are rising, high temperatures would be rising at a faster rate, which was not initially seen by the time series plots. Another possibility is that the model could be inaccurately predicting the trend as the confidence intervals are wide. There is always more possibilities on selecting the models as the selected models may not be ideal. It does, however, present a model that explores the current time series data available and allows for future estimation to be conducted. The low temperatures increasing is still a cause for concern as it is warmer at night, which could affect certain flora and fauna that need lower temperatures at night to operate. This might be due to the rise in greenhouse gas emissions happening around the world. As the concentration of greenhouse gases in the earth's atmosphere rises, the easier it becomes for the atmosphere to trap heat radiating from Earth to space (climate.nasa.gov). This phenomenon is known as global warming. Additionally, “a stronger greenhouse effect will warm the oceans and partially melt glaciers and other ice, increasing sea level. Ocean water also will expand if it warms, contributing further to sea level rise” (climate.nasa.gov). Along with the temperature slowly getting warmer, sea levels are also increasing. This poses a problem as people that live near the ocean might have to start moving inwards to avoid being flooded. Possible solutions to global warming include using renewable energy, sustainable transportation, and recycling” (https://solarimpulse.com/). In the end, temperatures rising can be explained by global warming and the best way to combat it is to reduce greenhouse gas emissions.

Works Cited

Austin Temperatures: Averages by Month. (n.d.). Retrieved December 10, 2019, from https://www.currentresults.com/Weather/Texas/Places/austin-temperatures-by-month-average.php.

Bradshaw, K. (2019, August 24). August's 19-day stretch of triple-digit temperatures is among Austin's longest. Retrieved December 10, 2019, from https://www.statesman.com/news/20190823/augusts-19-day-stretch-of-triple-digit-temperatures-is-among-austins-longest.

Climate Change: Vital Signs of the Planet. (n.d.). Retrieved from https://climate.nasa.gov/.

Dugan, A. (2018, May 4). Austin's air quality is gradually getting worse. Retrieved December 10, 2019, from https://www.kxan.com/news/local/austin/austins-air-quality-is-gradually-getting-worse/.

Impulse, S. (n.d.). 1000 profitable solutions for the environment. Retrieved from https://solarimpulse.com/.

Manly, B. F. J. (2008, October 21). *Statistics for Environmental Science and Management*. New York, NY: CRC.

Rong, Y. (2011). Environmental Statistics. *Environmental Forensics*, 12(3), 189-190. doi:10.1080/15275922.2011.599263

Shumway, R. H. and Stoffer, D. S. (2016, September). *Time Series Analysis and Its Applications*. Pittsburgh, PA: Springer.

United Nations, F.M. (United Nations Publication 2013). *Framework For The Development of Environment Statistics* (FDES No.92). New York, NY: United Nations Publication.