PROJECT REPORT

ISYE 6402 – TIME SERIES ANALYSIS

TIME SERIES ANALYSIS OF ATLANTA'S CLIMATIC VARIABLES
TEAM NUMBER 1

AHMED RABBANI
SHANSHAN HUAN
TIANZI REN
VATSAL GOYAL

27 / APRIL / 2022

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Study Aims

The main goal of this project is to establish appropriate univariable and multivariable time series models to forecast future temperature levels based on the analysis of characteristics in the past observations. In specific, this report analyses the past 20 year daily and weekly meteorological data in Atlanta, from 2002 to 2022, including temperature, precipitation, wind speed and air pressure time series. ARIMA(X), SARIMA(X) and VAR(X) models are utilized for describing the features in time series and predict future 28 days or 28 weeks temperatures, depending on data aggregation used.

1.2 Motivation and Expectation

Temperature forecasting is one of the most important topics that people want to find the answer for millennia when trying to understand and explore the nature. Temperature is not only a simple variable that tells people how hot or cold it is, but meteorological theories behind these numbers are influencing the growth and reproduction of plants and animals, people's daily life and work and even our earth every day. Therefore, we want to find an appropriate model to explain the meteorological behaviour of temperature time series.

First, we fit ARIMA and SARIMA model to the temperature data itself, which helps us explore the trend and seasonality within the temperature. But as we know, each part of ecosystem does not work separately, the temperature level is also correlated with other meteorological variables. As per the literature, temperatures are not independent of precipitation and in general humidity. When humidity is lower, usually during winters, temperatures generally fluctuate considerably. Whereas, during summer monsoon when humidity levels are high, temperatures tend to fluctuate less dramatically. Likewise, air pressure depends on the temperature of the air and density of the air molecules. Warm air in particular causes air pressure to rise. Lastly, as the wind speed increases, it draws heat from the body, reducing skin temperature and eventually the internal body temperature.

As discussed above, the relationship between temperature and precipitation, air pressure and wind speed might help explain more about the characteristics of temperature that were not addressed in ARIMA and SARIMA. Hence, we will leverage ARIMAX and SARIMAX, which treat other variables as exogenous factors, to find the significance of whether precipitation, air pressure and wind speed can help explain and predict the future temperature. Also, the relationship between factors could be more complex. It is necessary to investigate if there are any lead lag relationships among factors by VAR model, which can be helpful for predicting temperature. Eventually, we will calculate and compare the prediction accuracy for each model to find the more pertaining one that could reflect the relationship between temperature and other weather conditions.

2 DESCRIPTION OF DATASET

We will be primarily working on data acquired from the Python Library called Meteostat. Provided specific time frames and geographical coordinates, the library provides data related to various climatic variables. Data provided is obtained by Meteostat’s bulk data interface which mainly consists of public interfaces, primarily sponsored by governments. Daily data acquired from Meteostat consists of data related to following climatic variables:

- The average air temperature in °C (tavg)
- The daily precipitation total in mm (prcp)
- The average wind speed in km/h (wspd)
- The average sea-level air pressure in hPa (pres)

The data range is from 04/01/2002 to 04/01/2022, containing 61,614 records.[1]

[1] Further information related to API can be found https://dev.meteostat.net.

3 EXPLORATORY DATA ANALYSIS

3.1 Data Aggregation

Daily time series is often used for a more granular analysis, specifically, when the data shows significant daily changes, we would prefer to gauge the variance in these observations. But the temperature time series has a very obvious seasonality, and the 20-year range of the dataset makes us more inclined towards less granular analysis (e.g. weekly time series). In addition, models like SARIMA can only handle a max seasonality of 350 instead of 365. Thus, the weekly time series data is more applicable to our analysis.

We aggregated the average temperature from daily time series into weekly time series, and then plotted this weekly time series to observe its trend, seasonality and stationarity. From time series plot and ACF plots in Figure 1, we can clearly see that the original weekly time series is not stationarity and has obvious seasonality, but there is no clear upward or downward trend.

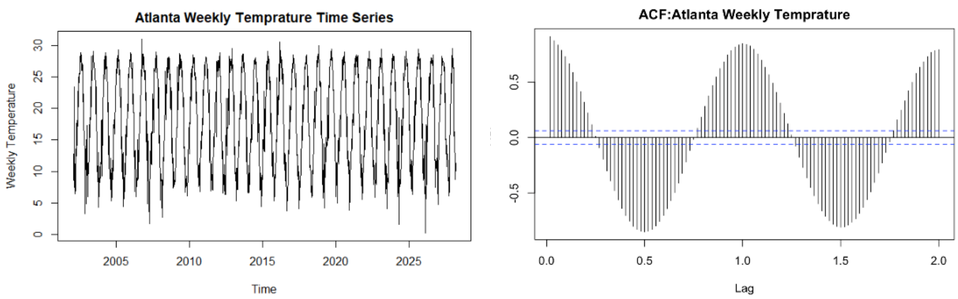

Figure 1 *Atlanta Weekly Temperature Time Series Plot & ACF Plot*

3.2 Correlation Matrix

By implementing the correlation matrix in Table 1, we want to explore the relationship between the four variables (temperature, precipitation, wind speed and air pressure). The negative correlations between temperature and wind speed, temperature and air pressure indicate that temperature will drop when wind speed and air pressure rise. Precipitation and temperature’s correlation is close to 0, demonstrating that they have a weak relationship. Similarly, air pressure and wind speed also have a weak relationship. Rising rainfall will be accompanied by increasing wind speeds as they have a positive correlation.

	temp	prcp	wspd	pres
temp	1.00000000	-0.04016753	-0.343622775	-0.503261169
prcp	-0.04016753	1.00000000	0.150115003	-0.118433395
wspd	-0.34362278	0.15011500	1.00000000	-0.004593475
pres	-0.50326117	-0.11843340	-0.004593475	1.00000000

***Table 1** Correlation Matrix of 4 variables*

Autocorrelation plots are shown in Appendix Figure 11. For autocorrelation, temperature and wind speed ACF plots show non-stationarity while precipitation and air pressure ACF plots resemble white noise. In terms of cross-correlation, we cannot see any lead or lag significant relationships.

3.3 Splines Trend + ANOVA Seasonality

Next, we applied splines regression approach to fit trend to check if trend is present in the weekly temperature time series. We overlapped the original time series with the fitted trend to find a slight decrease in Atlanta's 20-year temperature, accompanied by a slight fluctuation. Since this time series shows clear seasonality, we implemented Anova approach to fit seasonality and combined trend with seasonality. Based on original time series and trend plus seasonality overlapped plots in Figure 2, we can conclude that trend and seasonality capture most features of temperature time series.

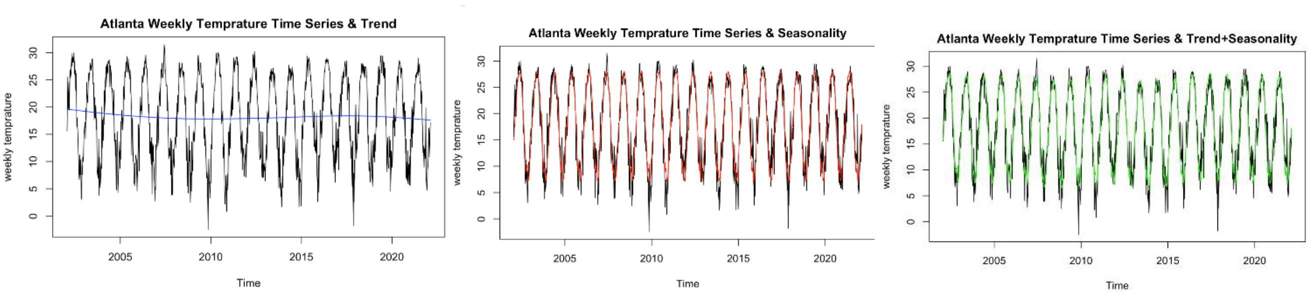

***Figure 2** Atlanta Weekly Data Trend & Seasonality Plot*

Residual analyses are shown in Appendix Figure 12. We plotted residuals, its ACF and PACF. There is high volatility in residuals and ACF plot shows non-stationarity. Thus, we should consider other approaches to model this time series. We performed QQ-plot and histogram for normality test. Both graphs indicate lack of normal distribution in residuals since in QQ-plot the points curve away from the line at each end. We also implemented ADF test for stationarity. P-value is 0.01 which demonstrates we should reject the null hypothesis that residuals are non-stationary.

4 UNIVARIATE ANALYSIS

4.1 ARIMA

After the initial EDA we find out that the time series is non-stationary and hence we move on to modelling the weekly temperature time series using ARIMA model and analysing the residuals to evaluate the goodness of fit. The first step was to divide the weekly temperature dataset into training and testing to evaluate the performance of each subsequent model. For this purpose, we kept the last 28 rows (corresponding to 28 weeks) as the test set while keeping everything before it for training. Once the aggregated data was split in the above manner, we ran it through a pipeline for order selection to select the ARIMA orders (p,d,q) which give us the best fit or the lowest AIC. We ran the order selection for max values of (p,d,q) as (10,2,10) respectively.

The final model with the least AIC came out to be the model corresponding to orders (4,0,3) respectively. The AIC score corresponding to this model was 4782.345 followed closely by the model corresponding to the orders (2,0,5) with an AIC value of 4782.371. Since the selected order has low values for p and q and is a fairly simpler model, we decide to move ahead with it. Thus, we trained a final model with the orders (4,0,3) on the original time series and computed the model residuals to evaluate the fit.

We took the residuals from our ARIMA model and plotted the ACF of residuals and squared residuals as shown in Appendix Figure 13. The residual ACF plot seems to exhibit stationarity since all the values are within the confidence interval. However, it is worth mentioning that there is a cyclic pattern observable in the residual ACF. That cyclic pattern is more evident in the squared residual ACF plot which has a wave like structure and also exhibits non-stationarity as a lot of lag values are outside the confidence interval.

We performed the box test for testing goodness of fit with respect to uncorrelated residuals and squared residuals. Both the hypothesis test returned a very low p-value thus indicating that we should reject the null hypothesis of uncorrelated errors and there is a statistically significant serial correlation between the residual values.

After training the final ARIMA model with the order (4,0,3) we decided to use a multi level approach to generate the predictions in the following manner:

2 Week rolling prediction:

Keeping in mind the business use case of a weather forecast model, it might not be feasible to do the order selection to iteratively forecast single values. Especially if you are working with daily data then that would mean running hundreds of ARIMA iterations every day to find the optimal order and forecast a single value. Hence, we also tried to generate the predictions in a 2 data point rolling window (2 weeks) so that the frequency of order selection for ARIMA halves.

Full prediction from a single model:

We know that time series forecasting techniques do not perform well in long term forecasting, but we tried predicting all the 28 values from a single model to quantify the difference in performance and establish a baseline for comparison with the other two cases as shown in Figure 3.

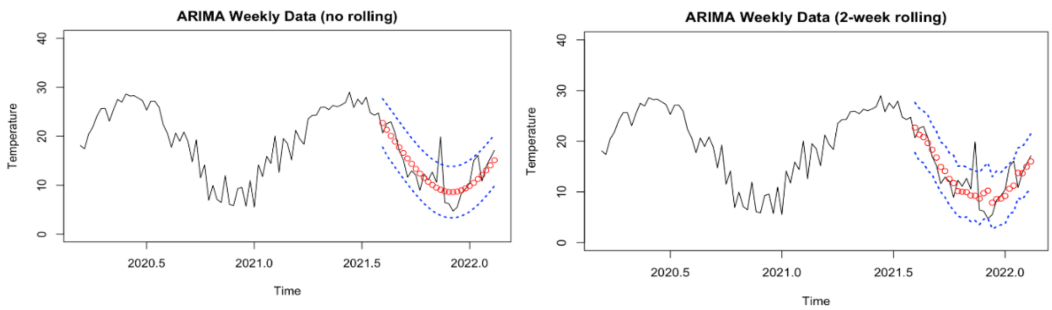

Figure 3 *ARIMA Weekly Data Prediction (No Rolling & 2-Week Rolling)*

S.no	Model	Data	Orders	Prediction Type	MAPE	PM
1	ARIMA	Weekly	(4,0,3)	Flat 28 weeks	0.20	0.37
2	ARIMA	Weekly	(4,0,3)	Rolling 2 weeks	0.21	0.38

Table 2 *ARIMA Model Order Selection & Prediction Accuracy*

We evaluated the performance of the above two prediction approaches in Table 2 with respect to the actual temperature values using the metrics MAPE and PM as they are scale invariant and thus provide means of uniform model comparison. The model with no rolling predictions outperforms the other models in terms of MAPE. The model with a 2-week rolling prediction performs the worse among both the models.

4.2 SARIMA

Since the ARIMA model still had an evident seasonality component in its residuals and squared residuals we decided to fit a SARIMA model to the weekly data to check whether the issue of seasonality can be addressed using SARIMA. Since we have weekly aggregated data points we would expect the data to have a seasonality of 52. Thus keeping the same 28 rows for test set we ran the order selection for SARIMA with max order or ARIMA (p,d,q) as (10,1,10) for the corresponding seasonal order (ps,ds,qs) as (1,0,1) and seasonality of 52. The final SARIMA model with the least AIC came out to be the model corresponding to (6,0,10) for the ARIMA part and (1,0,1) for the seasonal part. The AIC corresponding to this model was 4.71002 hence we trained another model with these orders and looked at the residuals to evaluate its goodness of fit.

Residual analysis for SARIMA on original data is shown in Appendix Figure 14. The ACF plot of the residuals seems to resemble that of a stationary time series as the ACF value is within the confidence interval for all the lags. However, there is still some observable seasonality in the ACF plots which is more evident in the plot of squared residuals. The ACF plot of squared residuals show a lot of deviation from the confidence band and thus has serial correlation. This is also confirmed by the box test which gives out a low value for both residuals and squared residuals thus making us reject the null hypothesis of uncorrelated residuals.

Since we were still unable to resolve the issue of uncorrelated errors, we tried to log transform the weekly temperature data to stabilize the variations (if any) and then train a SARIMA model on it to see whether that addresses the problem. Since there were negative data points in our data, the transformation that we used was log(temperature+(8/3)) to convert those values to positive before taking their log. We performed a similar order selection as SARIMA and the final orders of the selected model were (3,0,10) for the ARIMA part and (1,0,1) for the Seasonality part. The AIC corresponding to this model was -0.016 which is much better than all the previous models till now. Residual analysis for SARIMA on log transformed data is shown in Figure 15. Both the ACF plots for residuals and squared residuals seem to show stationarity as both of them are within the confidence band for all the lag values. Despite being stationary there is still some cyclic pattern evident in the ACF plots and the box test also gives out a low value indicating that there is serial correlation among residuals.

Similar to our predictions in ARIMA, we used both the SARIMA and log transformed SARIMA model to predict in a rolling window of size 2 (2 weeks) and in a non-rolling fashion forecasting for 28 data points using a single model. As a result, we obtained four different sets of predictions as shown in Figure 4 and Figure 5. One general observation is that the models with a 2-week rolling window prediction seems to perform visually better than its non-rolling counterpart, as it seems to capture the patterns of the data better. The log transformed SARIMA predictions seem to be more off than the actual values than the SARIMA model.

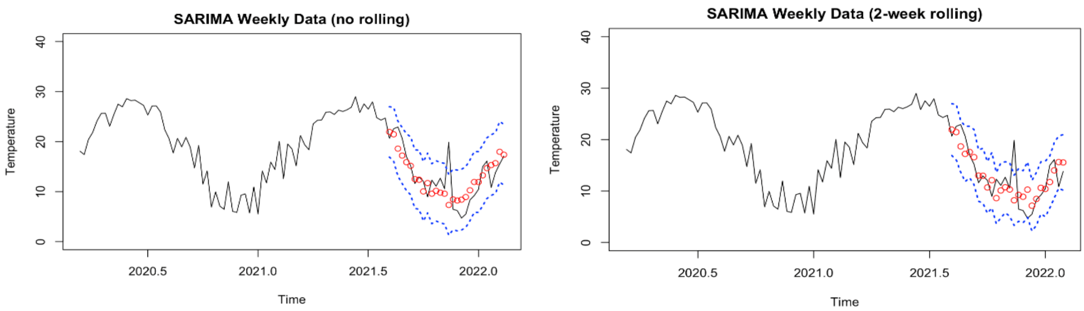

Figure 4 *SARIMA Weekly Data Prediction (No Rolling & 2-Week Rolling)*

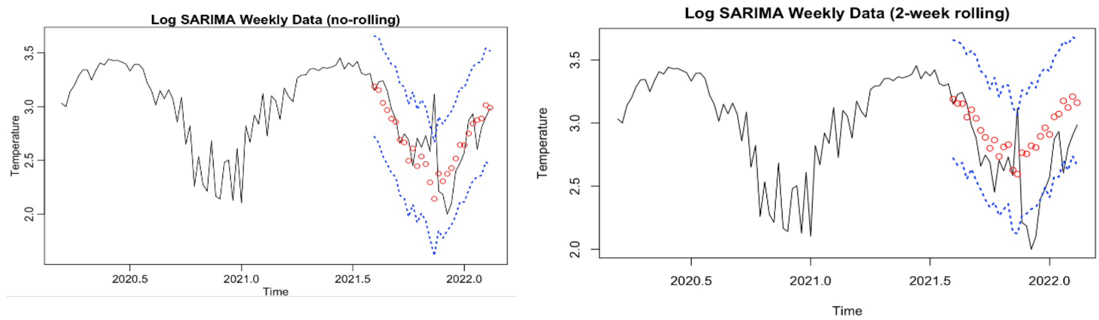

Figure 5 *Log SARIMA Weekly Data Prediction (No Rolling & 2-Week Rolling)*

S.no	Model	Data	Orders	Prediction Type	MAPE	PM
1	SARIMA	Weekly	(6,0,10) & (1,0,1) & 52	Flat 28 weeks	0.21	0.43
2	SARIMA	Weekly	(6,0,10) & (1,0,1) & 52	Rolling 2 weeks	0.22	0.43
3	SARIMA	Log Weekly	(3,0,10) & (1,0,1) & 52	Flat 28 weeks	0.07	0.58
4	SARIMA	Log Weekly	(3,0,10) & (1,0,1) & 52	Rolling 2 weeks	9.11	1.19

Table 3 *SARIMA Model Order Selection & Prediction Accuracy*

We compared the predictions across all the 4 prediction sets by evaluating 2 metrics MAPE and PM in Table 3. It is surprising to see that the non-rolling models seem to perform better than their 2 rolling-window predictions set based on both MAPE and PM. Among the log transformed and normal SARIMA we can see that the normal SARIMA performs better in terms of PM than the log transformed SARIMA but log transformed SARIMA outperforms SARIMA in terms of MAPE. One interesting thing to note here is that both SARIMA models perform worse than the ARIMA model in terms of the PM.

4.3 ARIMAX

Next, we applied on ARIMAX on log transformed weekly data. For model order selection, we used AIC as criterion and selected the model with the smallest AIC value, which has the order (4,1,6). The coefficient estimations are shown in Table 4:

	ARIMAX												
Coefficient	ar1	ar2	ar3	ar4	ma1	ma2	ma3	ma4	ma5	ma6	lagged_prcp	lagged_wspd	lagged_pres
Estimate	0.4177	1.0998	0.4189	-0.9883	-1.0515	-0.9774	0.3149	1.4124	-0.5548	-0.1398	0.019	-0.4257	-11.6287
P Value	0.0059	0.0082	0.0063	0.0063	0.0426	0.0533	0.0332	0.0288	0.0648	0.0352	0.0225	0.0377	2.3041

Table 4 *ARIMAX Model Estimated Coefficients*

As expected, conditional to other variables present in the model, the negative coefficient of wind speed means that high wind speed is more likely to push down the temperature. It is the same for air pressure. A low pressure could lead to a high temperature in the next period. However, the coefficient for precipitation is relatively small in this model and also insignificant, conditional to other variables present in the model. It could be because the precipitation in Atlanta does not vary with season as much as temperature. As shown in Figure 6, the left one is the time series plot for log transformed temperature, the right one is log transformed precipitation that exhibits a cyclical pattern instead of a clear seasonal pattern like temperature series. This makes the precipitation not significant for explaining the temperature level. It leads us to consider the model with seasonality in the next section.

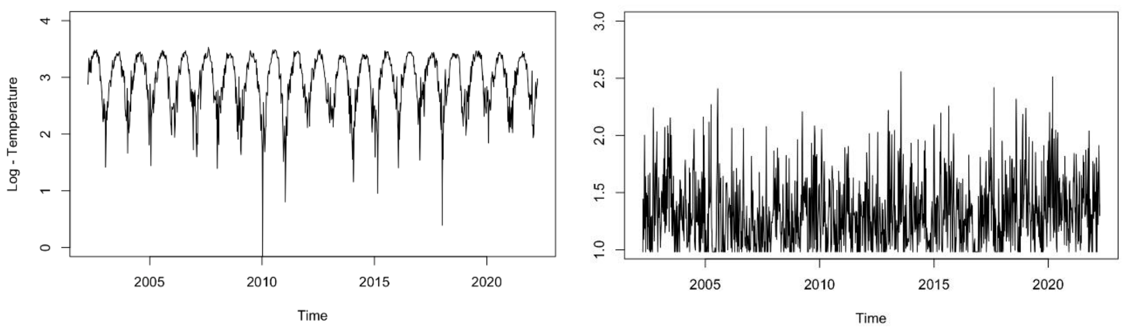

Figure 6 *(a) Log Temperature (b)Precipitation Time Series*

In addition to AIC criterion, we also utilized residual analysis to evaluate the ARIMAX model fitting as shown in Appendix Figure 16. The residual plot exhibits that the mean of residual process is constant, but there is a strong seasonality pattern shown after fitting. It is confirmed by the ACF plot, though most lags of autocorrelation are within the confidence band, yet the autocorrelation series fluctuates seasonally. And the p-value of correlation tests are less than 0.01, which means we should reject the null hypothesis of uncorrelation. Therefore, though the exogenous factors partially help explain more about temperature, it is proved by the residual analysis that the model fitting for ARIMAX is not perfect, and it is necessary to involve seasonality component into our model.

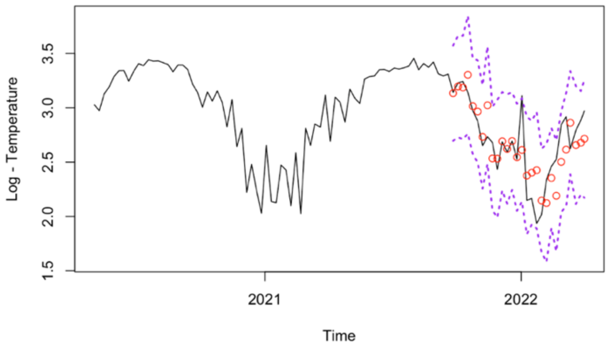

***Figure 7** ARIMAX Model Prediction*

For model comparison, two-week rolling predictions for 28 weeks are conducted as shown in Figure 7. That is adding two weeks of data to train our model and then predicting two weeks ahead each time, then repeating forward for 28 weeks. We can see that the predicted value (red points) and confidence band (purple dash line) are moving with the true value, which indicates that the ARIMAX model captures some characteristics in the temperature series. Also, we use MAPE and PM as prediction accuracy measures here, they are 0.07 and 0.39 for ARIMAX forecasting, respectively.

4.4 SARIMAX

Considering the complexity of the SARIMAX model and the costly computation, we did the order selection on log transformed weekly temperature data, based on the order that was already selected in the ARIMAX model. Specifically, we start from ARIMAX(4,1,6) and add seasonality order, then do order selection locally for the SARIMAX. In the end, the model with order (4,0,6) - (1,0,1) is selected due to the lowest AIC value. After fitting the SARIMAX model, the coefficient estimates are as shown in Table 5.

	SARIMAX															
Coefficient	ar1	ar2	ar3	ar4	ma1	ma2	ma3	ma4	ma5	ma6	sar1	sma1	intercept	lagged_prcp	lagged_wspd	lagged_pres
Estimate	0.1188	1.6922	0.1172	-0.9844	0.2393	-1.5621	-0.6501	0.723	0.2598	0.0985	0.9998	-0.9928	71.4669	0.0156	-0.3319	-9.784
P Value	0.0089	0.01	0.0077	0.0079	NaN	0.0231	NaN	0.0305	0.0102	0.0185	0.0003	0.0054	16.3081	0.0227	0.0397	2.3498

Table 5 *SARIMAX Model Estimated Coefficients*

From the results, we find that the coefficients for ma1 and ma3 are not significant, conditional to other coefficients present in the model. It means that there exists a bit of overfit for our model by adding seasonal factors and therefore, removing these redundant components from the full model could help with computational cost. Still the coefficients for wind speed and air pressure are negative, indicating that previous week's wind speed and air pressure could have a negative influence on the next week's temperature, conditional to other variables present in the model. The coefficient for precipitation is not significant in model anymore, since the p-value is larger than 0.01. It is stated that when we involve seasonality into our model, the precipitation from the previous week could not help explaining the temperature level in the next period. It is not a determinant for weekly prediction, which quite makes sense since we think precipitation has more of an "instant" effect for temperature in Atlanta and it might be useful when doing daily level interpretation or prediction.

To evaluate the SARIMAX model fitting goodness, we did residual analysis as shown in Appendix Figure 17. The residual process plot shows, though still some seasonal peaks appear, less seasonality pattern than that for ARIMAX model. Also, less seasonality autocorrelation is exhibited in the ACF plot and most are close to zero except several outliers are out of confidence band, which means the SARIMAX captures some seasonal factors in this series. The residual analysis indicates that the SARIMAX fitting is better than the ARIMAX.

Similarly, we did the two-week rolling predictions for the next 28 weeks and the results are shown Figure 8. The actual values are within the 95% confidence band (blue dash line), except one abnormal point is outside. Mean Absolute Percentage Error (MAPE) and Precision Measure (PM) are used here as prediction accuracy measures. They are 0.06 and 0.38 respectively, and are both smaller than that of ARIMAX and SARIMA. The prediction values are close to the true value, which means the model captures most of trend and seasonality patterns, and exogenous variables also help in forecasting the temperature levels.

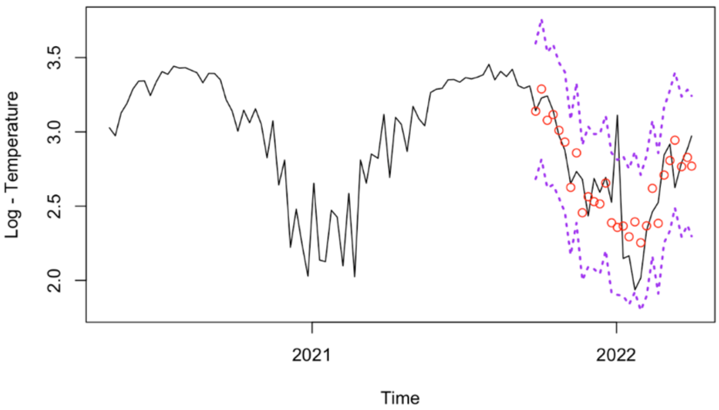

Figure 8 *SARIMAX Model Prediction*

5 MULTIVARIATE ANALYSIS

5.1 VAR – Weekly Data

In this section we applied VAR on log differenced weekly data. We had to take one order difference of all the time series in order to maintain stationarity. Time series and ACF plots for original and differenced time series are shown in Appendix Figure 18. For VAR order selection, we used both AIC and BIC criteria in VAR select formulation in R studio. Since BIC penalizes model complexity further than AIC does, we chose the recommended order with respect to BIC, which was 3. Result of VAR model with order 3 are shown in Table 6. Coefficients which are found to be statistically significant for a significant level of 5% are highlighted in red.

Temperature		
Coefficient	Estimate	P Value
tavg.l1	-0.445	0.000
prcp.l1	-0.033	0.159
wspd.l1	-0.151	0.001
pres.l1	2.882	0.266
tavg.l2	-0.219	0.000
prcp.l2	-0.006	0.839
wspd.l2	-0.069	0.171
pres.l2	-0.279	0.921
tavg.l3	-0.062	0.064
prcp.l3	0.013	0.571
wspd.l3	-0.029	0.501
pres.l3	-1.006	0.698
const	0.000	0.983

Precipitation		
Coefficient	Estimate	P Value
tavg.l1	0.073	0.108
prcp.l1	-0.745	0.000
wspd.l1	-0.003	0.955
pres.l1	2.938	0.394
tavg.l2	0.038	0.430
prcp.l2	-0.495	0.000
wspd.l2	-0.009	0.893
pres.l2	-2.150	0.567
tavg.l3	0.002	0.960
prcp.l3	-0.285	0.000
wspd.l3	-0.013	0.828
pres.l3	-4.984	0.148
const	0.001	0.935

Wind Speed		
Coefficient	Estimate	P Value
tavg.l1	0.059	0.027
prcp.l1	-0.006	0.762
wspd.l1	-0.681	0.000
pres.l1	-1.959	0.333
tavg.l2	0.035	0.216
prcp.l2	-0.016	0.454
wspd.l2	-0.388	0.000
pres.l2	-3.037	0.168
tavg.l3	0.009	0.731
prcp.l3	0.009	0.644
wspd.l3	-0.220	0.000
pres.l3	0.059	0.977
const	-0.001	0.831

Air Pressure		
Coefficient	Estimate	P Value
tavg.l1	0.000	0.291
prcp.l1	0.001	0.009
wspd.l1	0.001	0.013
pres.l1	-0.585	0.000
tavg.l2	0.000	0.542
prcp.l2	0.000	0.237
wspd.l2	0.001	0.054
pres.l2	-0.384	0.000
tavg.l3	-0.001	0.064
prcp.l3	0.000	0.796
wspd.l3	0.001	0.123
pres.l3	-0.208	0.000
const	0.000	0.970

Table 6 *VAR Model Estimated Coefficients on Weekly Data*

As per the results above, lagged one temperature, wind speed and lagged two temperatures effect future temperature values. On first glance, it seems that very few variables effect future weekly temperature values. We were expecting more coefficients to be statistically significant however, given that this is weekly data, due to aggregation, it is possible that impact of certain coefficients got diluted. We explored this intuition further by doing VAR on daily data in subsequent sections.

In order to evaluate goodness of fit for our fitted var model we did residual analysis using hypothesis testing. We did normality test for which the null hypothesis was that the residuals are normally distributed. The test returned a low p value indicating that we reject our null hypothesis and conclude that residuals are plausibly not normally distributed. We used arch test where the null hypothesis was that the residuals have constant variance. The test returned a low p value indicating that we reject our null hypothesis and conclude that residuals plausibly do not have constant variance. Lastly, to check for uncorrelated residuals we used serial test where the null hypothesis was the residuals are uncorrelated. This test also returned a low p value indicating that we reject our null hypothesis and conclude that residuals are plausibly not uncorrelated.

In order to evaluate the different variables selected by various variable selection techniques, we evaluated the variables selected by restricted var and stepwise regression with backward propagation and three steps. Results for restricted VAR and stepwise regression are shown in Table 7 and 8 respectively, selected regression coefficients are marked with X.

Restricted VAR				
Coefficient	Temperature	Precipitation	Wind Speed	Air Pressure
tavg.l1	X		X	
prcp.l1	X	X		X
wspd.l1			X	X
pres.l1				X
tavg.l2	X			
prcp.l2		X		
wspd.l2			X	X
pres.l2				X
tavg.l3				
prcp.l3		X		
wspd.l3			X	X
pres.l3				X

Table 7 *Restricted VAR Variables Selection*

Stepwise Regression				
Coefficient	Temperature	Precipitation	Wind Speed	Air Pressure
tavg.l1	X	X	X	X
prcp.l1	X	X		X
wspd.l1	X		X	X
pres.l1	X	X	X	X
tavg.l2	X	X	X	
prcp.l2		X	X	X
wspd.l2	X	X	X	X
pres.l2		X	X	X
tavg.l3	X		X	X
prcp.l3	X	X	X	
wspd.l3	X	X	X	X
pres.l3	X	X		X

Table 8 *Stepwise Regression Variables Selection*

In general, restricted VAR has mainly selected statistically significant or nearly statistically significant predictors from the unrestricted VAR model. Whereas the predictors excluded by step wise regression at the end of the third step are neither statistically significant in restricted VAR nor in unrestricted VAR.

In order to evaluate which meteorological indicators can help in explaining or predicting other indicators we did Granger Causality Analysis. Results of the analysis are shown in Table 9:

Granger Causality				
Indicators	Temperature	Precipitation	Wind Speed	Air Pressure
Temperature				
Precipitation				X
Wind Speed	X			X
Air Pressure				

Table 9 *Granger Causality Analysis*

As per the results, changes in wind speed do effect changes in temperature. Whereas changes in wind speed and precipitation do effect changes in air pressure. However, as per the results there is no granger causality between precipitation, air pressure and temperature. This is counter intuitive to our initial expectations since one would expect there to be some granger causality between precipitation in particular and temperature. One possible explanation for this result could be that in Atlanta usually it rains once or twice in a week where as temperature is recorded every day. When we aggregate these values on a weekly basis, precipitation value becomes even smaller due to lower number of observations. Additionally, by intuition we know that precipitation is more likely to effect temperature on the same day or at max next day. It is unlikely for precipitation to affect the average temperature for the entire subsequent week. Therefore, in order to substantiate our intuition, we repeated our VAR analysis on daily data in subsequent sections.

We did two weeks rolling predictions for 28 weeks. Resulting plot for last 100 observations, along with predictions in red and confidence bands in purple are shown in Figure 9. The prediction measures for these predictions were 0.82 and 1.15 for mean absolute percentage error and precision measure respectively. As per the prediction plot, we can see that predictions follow the overall behaviour of actual data in terms of trend and seasonality.

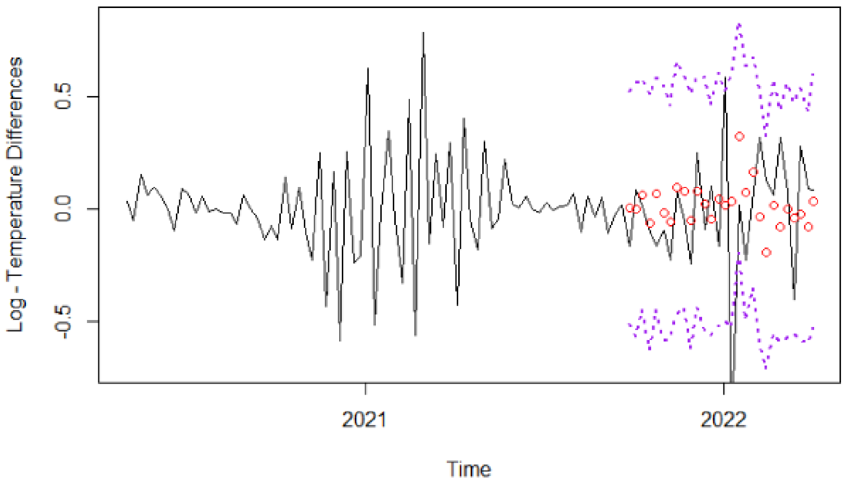

Figure 9 *VAR Model Prediction on Weekly Data*

5.2 VAR – Daily Data

In order to follow our intuition that precipitation and wind speed will have more impact on daily temperature than on weekly temperature, we repeated VAR analysis on log differenced daily data. Result of VAR model with order 3 are shown are in Table 10. Coefficients which are found to be statistically significant for a significant level of 5% are highlighted in red. Granger causality results and predictions are shown in Table 11 and Figure 10 respectively. Predictions had a precision measure of 0.80.

Temperature		
Coefficient	Estimate	P Value
tavg.l1	-0.082	0.000
prcp.l1	-0.008	0.028
wspd.l1	-0.186	0.000
pres.l1	6.216	0.000
tavg.l2	-0.204	0.000
prcp.l2	0.000	0.958
wspd.l2	-0.062	0.000
pres.l2	7.179	0.000
tavg.l3	-0.067	0.000
prcp.l3	0.003	0.477
wspd.l3	-0.070	0.000
pres.l3	4.951	0.000
const	0.000	0.993

Precipitation		
Coefficient	Estimate	P Value
tavg.l1	0.318	0.000
prcp.l1	-0.611	0.000
wspd.l1	-0.016	0.548
pres.l1	-3.245	0.328
tavg.l2	0.189	0.000
prcp.l2	-0.410	0.000
wspd.l2	-0.001	0.968
pres.l2	20.110	0.000
tavg.l3	0.225	0.000
prcp.l3	-0.198	0.000
wspd.l3	-0.059	0.026
pres.l3	13.400	0.000
const	0.000	0.999

Wind Speed		
Coefficient	Estimate	P Value
tavg.l1	0.077	0.000
prcp.l1	0.033	0.000
wspd.l1	-0.419	0.000
pres.l1	-25.600	0.000
tavg.l2	0.098	0.000
prcp.l2	0.007	0.294
wspd.l2	-0.339	0.000
pres.l2	3.723	0.013
tavg.l3	0.086	0.000
prcp.l3	0.010	0.081
wspd.l3	-0.222	0.000
pres.l3	-12.270	0.000
const	0.000	0.975

Air Pressure		
Coefficient	Estimate	P Value
tavg.l1	-0.001	0.000
prcp.l1	0.000	0.000
wspd.l1	0.002	0.000
pres.l1	0.177	0.000
tavg.l2	-0.001	0.000
prcp.l2	0.000	0.086
wspd.l2	0.000	0.012
pres.l2	-0.363	0.000
tavg.l3	-0.001	0.000
prcp.l3	0.000	0.023
wspd.l3	0.000	0.000
pres.l3	-0.102	0.000
const	0.000	0.983

Table 10 *VAR Model Estimated Coefficients on Daily Data*

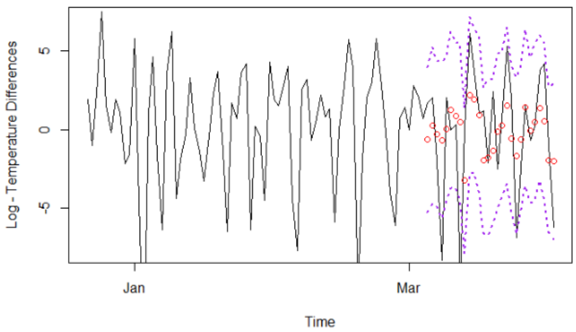

Figure 10 *VAR Model Prediction on Daily Data*

Granger Causality				
Indicators	Temperature	Precipitation	Wind Speed	Air Pressure
Temperature		X	X	X
Precipitation	X		X	X
Wind Speed	X			X
Air Pressure	X	X	X	

Table 11 *Granger Causality Analysis on Daily Data*

As per the results, almost all of the lagged and lead factors are statistically significant in predicting daily average temperature. As per granger causality we can see that there is a dual granger causality between temperature and precipitation, wind speed and air pressure, implying that all these factors have lead lag relationships and granger cause each other. These results are more in line with the principles of water cycle and psychometry. Based on above results, it would make more sense to include other factors as endogenous variables to explain daily average temperature via VAR analysis.

5.3 VARX – Weekly & Daily Data

Although we proved earlier that there is a lead lag relationship between temperature and other factors, we wanted to further explore the possible impact that other factors can have on daily temperature in exogenous capacity. Therefore, we applied VARX on daily and weekly log difference temperature values. In both cases, we kept wind speed and air pressure as endogenous variables and precipitation as exogenous variable. Resulting model and granger causality tables are shown below.

Temperature - Weekly Data		
Coefficient	Estimate	P Value
tavg.l1	-0.450	0.000
wspd.l1	-0.153	0.000
pres.l1	2.828	0.271
tavg.l2	-0.221	0.000
wspd.l2	-0.065	0.193
pres.l2	-0.288	0.918
tavg.l3	-0.059	0.078
wspd.l3	-0.020	0.643
pres.l3	-1.154	0.652
const	0.000	0.986
prcp	0.044	0.019

Temperature - Daily Data		
Coefficient	Estimate	P Value
tavg.l1	-0.089	0.000
wspd.l1	-0.185	0.000
pres.l1	5.941	0.000
tavg.l2	-0.208	0.000
wspd.l2	-0.059	0.000
pres.l2	6.213	0.000
tavg.l3	-0.069	0.000
wspd.l3	-0.068	0.000
pres.l3	4.798	0.000
const	0.000	0.994
prcp	0.019	0.000

Table 12 VARX Model Estimated Coefficients

Granger Causality - Weekly Data	
Indicators	Temperature
Temperature	
Precipitation	X
Wind Speed	X
Air Pressure	

Granger Causality - Daily Data	
Indicators	Temperature
Temperature	
Precipitation	X
Wind Speed	X
Air Pressure	X

***Table 13** VARX Model Granger Causality*

As expected, similar to VAR analysis, other meteorological factors are found to have a stronger influence on temperature in daily data analysis as compared to weekly data analysis. Meanwhile, precipitation is found to be statistically significant both in exogeneous and endogenous capacity, given other variables are present in the model.

6 DISCUSSION & CONCLUSION

For consistency, we compared the prediction accuracy of two-week rolling prediction for each model with log weekly data as shown in Table 14. As per the table, all the models accounting for exogenous variables have better prediction accuracies than their corresponding models without exogenous factors. This indicates that temperature level is correlated with other weather conditions in addition to being dependent on lagged temperature values.

SARIMAX has the least Mean Absolute Percentage Error (MAPE) value, which makes a lot of sense since SARIMAX model considers both the exogenous factors and the seasonality exhibited in temperature time series. Although, the VAR and VARX model help explain the relationships between the temperature and other meteorological measures, the prediction accuracies for these models are not as good as that of SARIMAX. It might be because the differencing of time series done in VAR/VARX might have failed to remove all forms of seasonality completely and hence that may have affected the prediction measures. Although SARIMAX considers both exogenous factors and seasonality in temperature series itself, it does not capture the feedback relation with other meteorological variables beyond the first lead. This can affect the explanatory power of the SARIMAX model, since as per the Granger Causality analysis from VAR, we saw that lead values of other factors, with order greater than 1 do effect changes in temperature time series. Hence, there is a tradeoff when it comes to selecting the best model between SARIMAX and VAR for predictions, but given the intuition that seasonality is present in the model, it would make more sense to consider SARIMAX for temperature predictions.

S.no	Model	Data	Orders	Prediction Type	MAPE	PM
1	ARIMA	Weekly	(4,0,3)	Flat 28 weeks	0.20	0.37
2	ARIMA	Weekly	(4,0,3)	Rolling 2 weeks	0.21	0.38
3	SARIMA	Weekly	(6,0,10) & (1,0,1) & 52	Flat 28 weeks	0.21	0.43
4	SARIMA	Weekly	(6,0,10) & (1,0,1) & 52	Rolling 2 weeks	0.22	0.43
5	SARIMA	Log Weekly	(3,0,10) & (1,0,1) & 52	Flat 28 weeks	0.07	0.58
6	SARIMA	Log Weekly	(3,0,10) & (1,0,1) & 52	Rolling 2 weeks	9.11	1.19
7	ARIMAX	Log Weekly	(4,1,6)	Rolling 2 weeks	0.07	0.39
8	SARIMAX	Log Weekly	(4,0,6) & (1,0,1)	Rolling 2 weeks	0.06	0.38
9	VAR	Log Weekly Differenced	3	Rolling 2 weeks	0.82	1.15
10	VAR	Log Daily Differenced	3	Rolling 2 weeks	-	0.80
11	VARX	Log Weekly Differenced	3	Rolling 2 weeks	-	1.11
12	VARX	Log Daily Differenced	3	Rolling 2 weeks	-	0.80

Table 14 *All Models Prediction Accuracy Summary*

7 APPENDIX

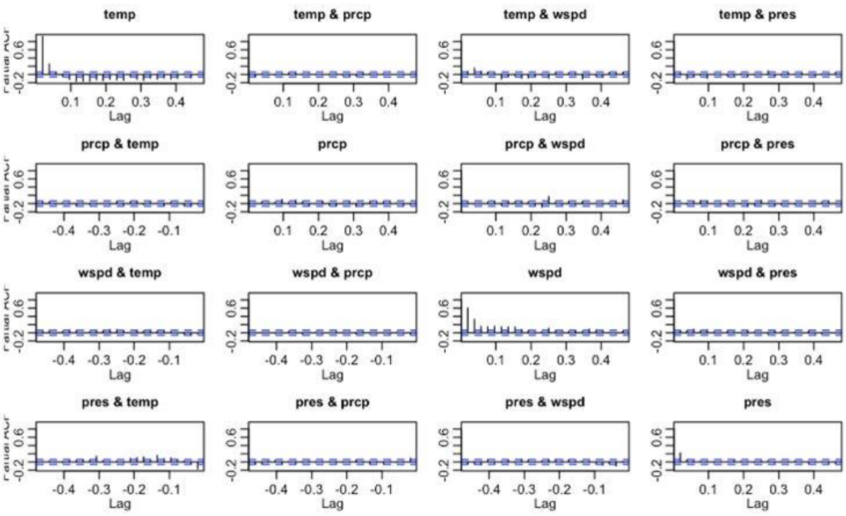

Figure 11 *4 Variables Auto-correlation & Cross-correlation ACF Plots*

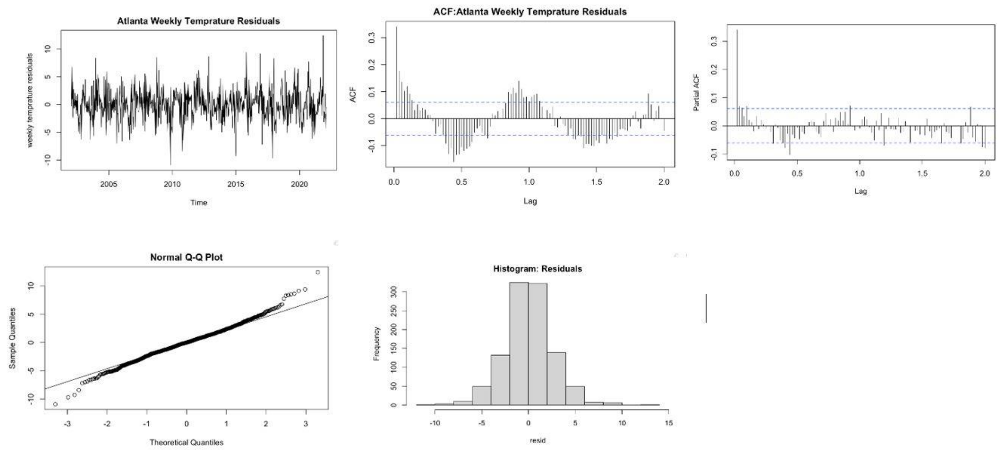

Figure 12 *Weekly Data Residuals Analysis & Normality Testing*

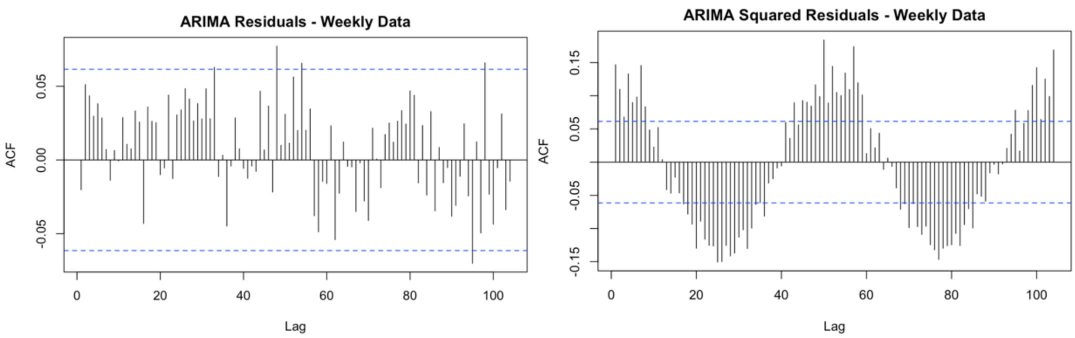

Figure 13 *ARIMA Residuals & Squared Residuals Analysis*

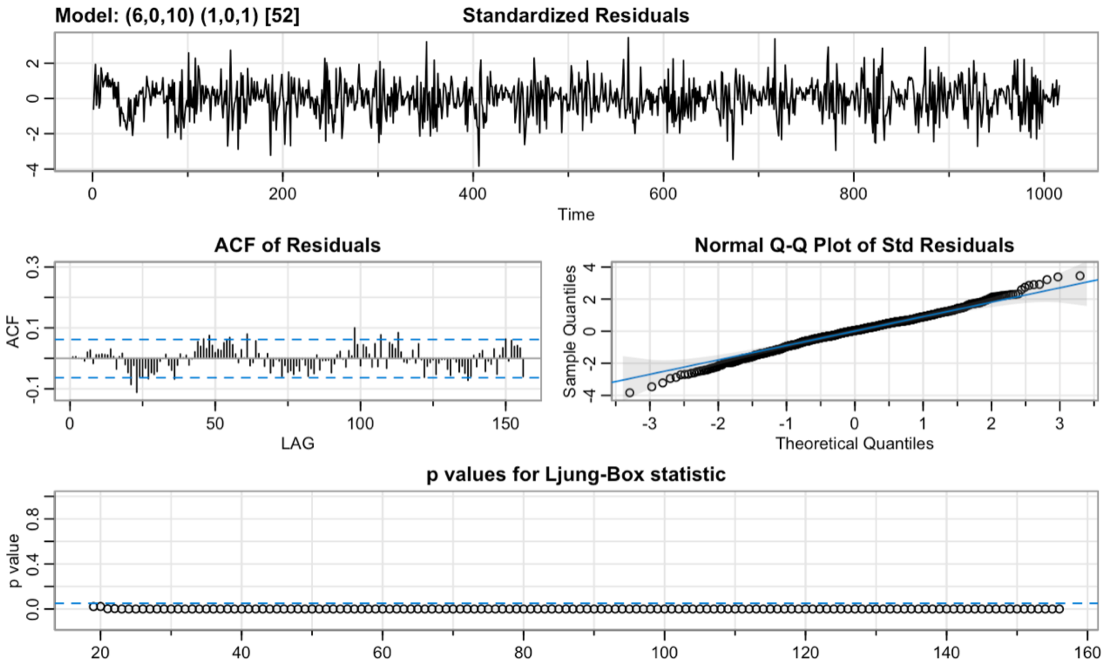

Figure 14 *SARIMA Residuals Analysis on Original Data*

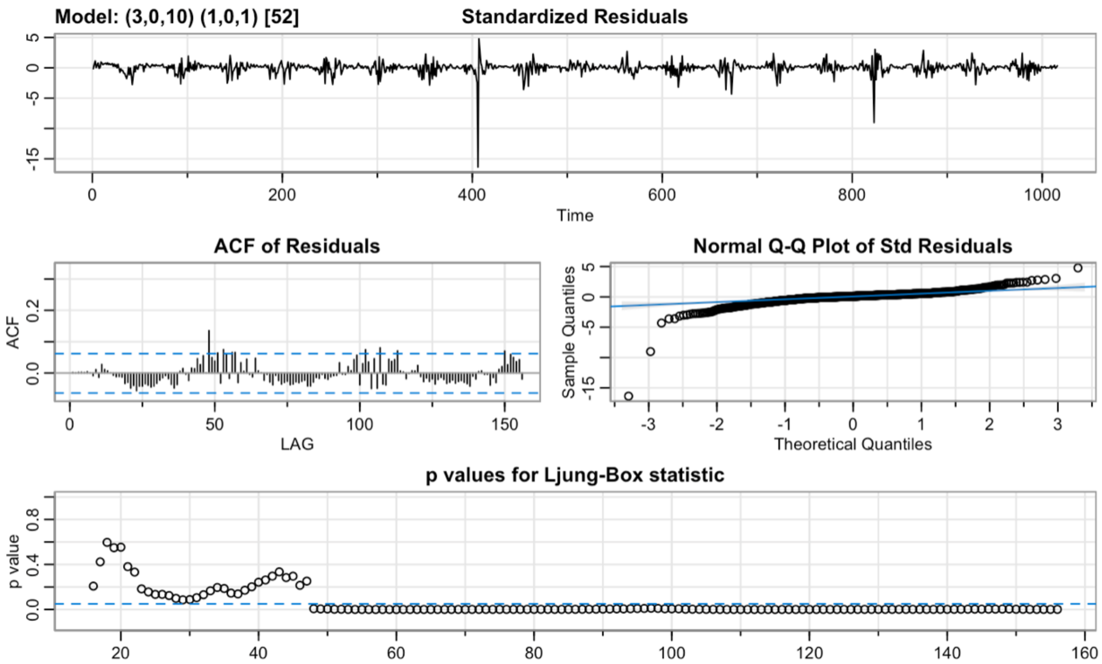

Figure 15 *SARIMA Residuals Analysis on Log Transformed Data*

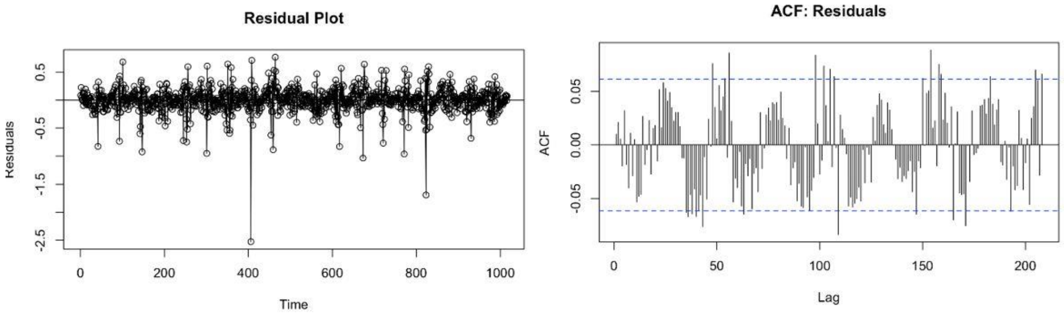

Figure 16 *ARIMAX Residuals Analysis*

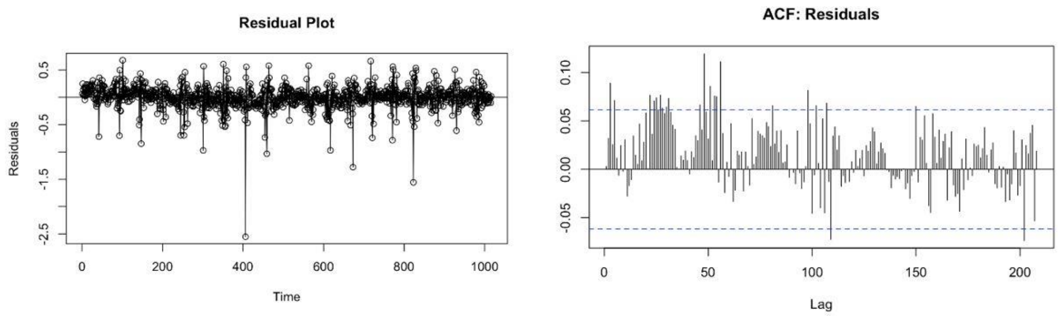

Figure 17 *SARIMAX Residuals Analysis*

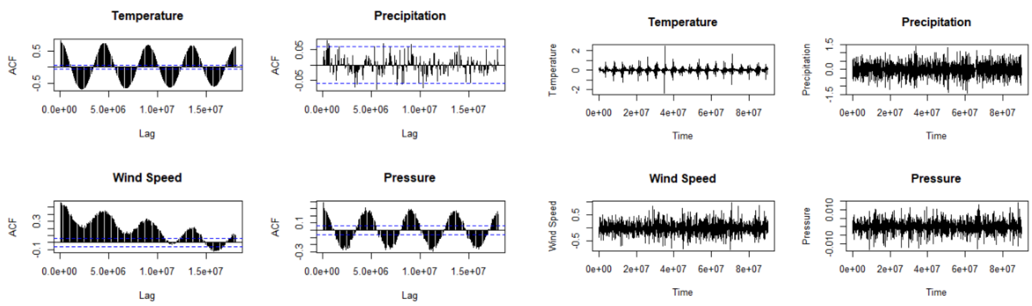

Figure 18 (a) *ACF of Log Time Series (b) ACF of Log Differenced Time Series*