



**University Of Connecticut**

**OPIM 5671 – Data Mining and Business Intelligence**

Time Series Forecasting Project Proposal

**Mumbai Precipitation Forecasting for Sustainable Urban Development**

By

Group 5

Lahari Maddula

Pradeepti Dokka

Sai Deepika Bandari

Sanchita Godse

Shuang Ma

## Table of Contents

1. [Executive Summary](#)
2. [Introduction](#)
  - 2.1 Problem Statement
  - 2.2 Data Description
3. [Data Exploration](#)
4. [Time Series Exploration](#)
  - 3.2. Series Value Graph
  - 3.3. Seasonal Cycle Graph
  - 3.4. Seasonal Component Graph
  - 3.5. Cross Correlation Graphs
  - 3.6. Trend Component Graph
  - 3.7. Augmented Dickey-Fuller Unit Root Test
5. [Data Models & Forecasting](#)
6. [Model Comparison](#)
  - 5.1. AIC and SBC Values
  - 5.2. AIC and SBC Comparison for All Models
7. [Business Insights and Recommendations](#)
8. [Conclusion](#)
9. [References](#)

## 1. Executive Summary:

In today's era of climate change and sustainable development, the renewable energy sector is pivotal in mitigating environmental impacts and ensuring a sustainable future. Accurate weather forecasting, especially precipitation forecasting, is vital for optimizing renewable energy generation, agricultural practices, and building energy efficiency. Our project, “Mumbai Precipitation Forecasting for Sustainable Urban Development,” aims to develop a precipitation forecasting model that will assist Mumbai in crafting sustainable solutions such as hydropower generators and more energy-efficient infrastructures.

## 2. Introduction:

Accurate precipitation forecasting is essential for:

**Renewable Energy:** Precipitation forecasts can help energy operators anticipate changes in solar and wind power generation due to weather conditions, enabling efficient energy resource management.

**Precipitation:** forecasts assist in optimizing heating and cooling systems in buildings, reducing energy consumption and costs.

**Agriculture:** Farmers can use precipitation forecasts to plan irrigation and planting schedules, improving crop yields and resource utilization.

### 2.1 Problem Statement:

Mumbai, a densely populated coastal city, experiences significant rainfall variations throughout the year. Accurate precipitation forecasting is vital to ensure sustainable energy generation, building energy efficiency, and agricultural productivity. However, existing forecasting models may not capture the nuances of Mumbai's unique climate.

### 2.2 Data Description:

The Data Set used is sourced from the Kaggle. It has information on the duration of rainfall for the first day of every month from 2000 to 2020. It has 252 observations.

The dataset explored in this report includes the following meteorological variables: specific humidity, relative humidity, temperature, and precipitation.

Meteorological data plays a pivotal role in understanding weather patterns and their impact on various sectors, including renewable energy, agriculture, and building management. The dataset under examination contains a diverse array of meteorological variables, each contributing unique insights into our analysis. In this section, we will delve into the specifics of these variables and the steps taken to prepare the dataset for comprehensive analysis.

The variables present on our dataset are:

1.     **Precipitation:** Rainfall data generally are collected using electronic data loggers that measure the rainfall in 0.01- inch increments every 15 minutes using either a tipping-bucket rain gauge or a collection well gauge.
2.     **Specific humidity:** Specific humidity is a simple percentage measurement of the total water in the atmosphere, the other term for this is the absolute humidity.
3.     **Relative humidity:** Relative humidity on the other hand is the amount of water in the atmosphere as a percentage of the maximum carrying capacity of the air.
4.     **Temperature:** The temperature of the data expressed in numbers. The unit of the measurement is
5.     **Date:** The format is D/M/YYYY and this column was obtained by combining the three columns of the original dataset which had Date, month and the year as three different columns.

- Snapshot of the dataset before cleaning:

Year	Month	Day	Specific Humidity	Relative Humidity	Temperature	Precipitation
2000	1	1	8.06	48.25	23.93	0
2000	2	1	8.73	50.81	25.83	0.11
2000	3	1	8.48	42.88	26.68	0.01
2000	4	1	13.79	55.69	22.49	0.02
2000	5	1	17.4	70.88	19.07	271.14
2000	6	1	19.53	84.19	7.91	313.67
2000	7	1	18.8	88.5	6.67	820.45
2000	8	1	18.86	88.44	7.07	362.38
2000	9	1	18.43	86.12	10.63	97.85
2000	10	1	16.72	78.38	15.38	63.41
2000	11	1	12.02	63.69	17.48	4.37
2000	12	1	7.39	44.56	20.09	11.25
2001	1	1	8.06	45.81	22.94	0
2001	2	1	7.57	41.56	22.7	0
2001	3	1	11.29	55.56	20.97	0.03
2001	4	1	12.27	49.69	22.73	1.57
2001	5	1	16.6	62.44	16.03	29.11
2001	6	1	19.1	81.94	9.83	510.09
2001	7	1	19.23	89	5.98	622.31
2001	8	1	18.92	90.94	5.47	429.62
2001	9	1	18.68	87.69	10.91	155.88
2001	10	1	16.72	80.19	14.76	120.24
2001	11	1	12.51	66.56	18.25	6.15
2001	12	1	9.83	56.44	17.97	0
2002	1	1	8.18	51.06	23.58	0.03
2002	2	1	7.75	39.31	25.44	0.11
2002	3	1	9.77	43.44	23.32	0.47
2002	4	1	12.88	50.12	21.04	2.06
2002	5	1	16.85	60.44	19.38	8.63
2002	6	1	19.17	78.06	13.12	498.92
2002	7	1	18.92	85.12	6.26	126.77
2002	8	1	18.86	89.88	6.24	581.79
2002	9	1	17.88	85.75	12.04	87.71
2002	10	1	15.81	71.75	17.16	17.29
2002	11	1	11.17	57	18.57	2.59

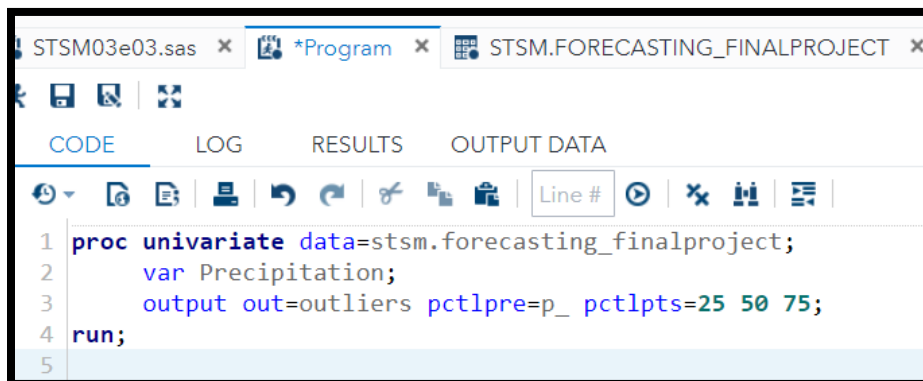
- **Cleaned Dataset:**

1	Date	Specific Humidity	Relative Humidity	Temperature	Precipitation
2	1/1/2000	8.06	48.25	23.93	0
3	2/1/2000	8.73	50.81	25.83	0.11
4	3/1/2000	8.48	42.88	26.68	0.01
5	4/1/2000	13.79	55.69	22.49	0.02
6	5/1/2000	17.4	70.88	19.07	271.14
7	6/1/2000	19.53	84.19	7.91	313.67
8	7/1/2000	18.8	88.5	6.67	820.45
9	8/1/2000	18.86	88.44	7.07	362.38
10	9/1/2000	18.43	86.12	10.63	97.85
11	10/1/2000	16.72	78.38	15.38	63.41
12	11/1/2000	12.02	63.69	17.48	4.37
13	12/1/2000	7.39	44.56	20.09	11.25
14	1/1/2001	8.06	45.81	22.94	0
15	2/1/2001	7.57	41.56	22.7	0
16	3/1/2001	11.29	55.56	20.97	0.03
17	4/1/2001	12.27	49.69	22.73	1.57
18	5/1/2001	16.6	62.44	16.03	29.11

### 3. Data Exploration:

Upon analyzing the dataset provided, we attempted to identify the optimal model. However, the values returned by the model appeared off and didn't seem logical. This led us to re-evaluate our dataset to determine if there was something we might have missed.

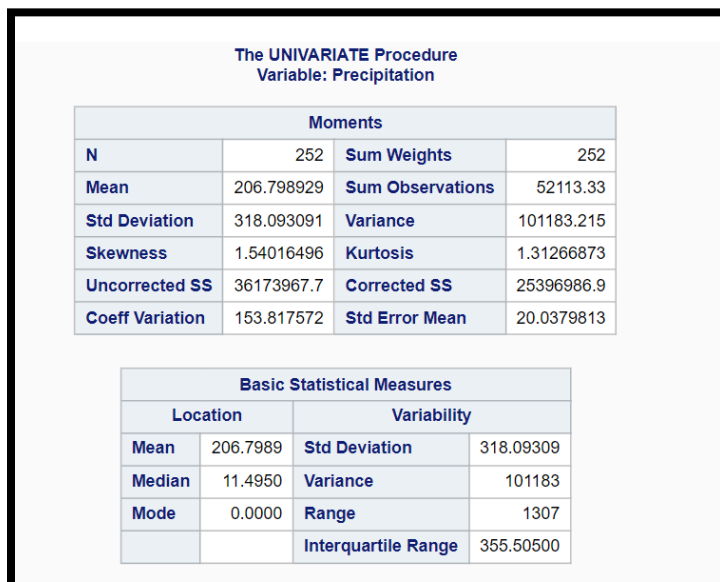
We then performed the Performed Outlier Detection by using the proc univariate function present in SAS. We are using this to generate the outliers and the values for the 25th, 50th and 75th quantile.



```
1 proc univariate data=stsm.forecasting_finalproject;
2   var Precipitation;
3   output out=outliers pctlpre=p_ pctlpts=25 50 75;
4 run;
5
```

#### Results :

The results obtained can be seen as follows:

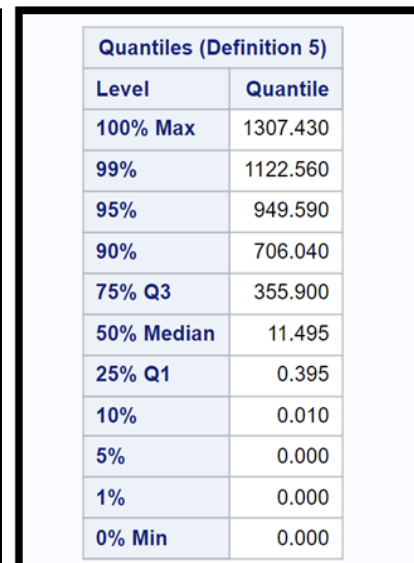


The UNIVARIATE Procedure  
Variable: Precipitation

Moments			
N	252	Sum Weights	252
Mean	206.798929	Sum Observations	52113.33
Std Deviation	318.093091	Variance	101183.215
Skewness	1.54016496	Kurtosis	1.31266873
Uncorrected SS	36173967.7	Corrected SS	25396986.9
Coeff Variation	153.817572	Std Error Mean	20.0379813

Basic Statistical Measures			
Location		Variability	
Mean	206.7989	Std Deviation	318.09309
Median	11.4950	Variance	101183
Mode	0.0000	Range	1307
		Interquartile Range	355.50500



Quantiles (Definition 5)	
Level	Quantile
100% Max	1307.430
99%	1122.560
95%	949.590
90%	706.040
75% Q3	355.900
50% Median	11.495
25% Q1	0.395
10%	0.010
5%	0.000
1%	0.000
0% Min	0.000

# Outlier Detection :

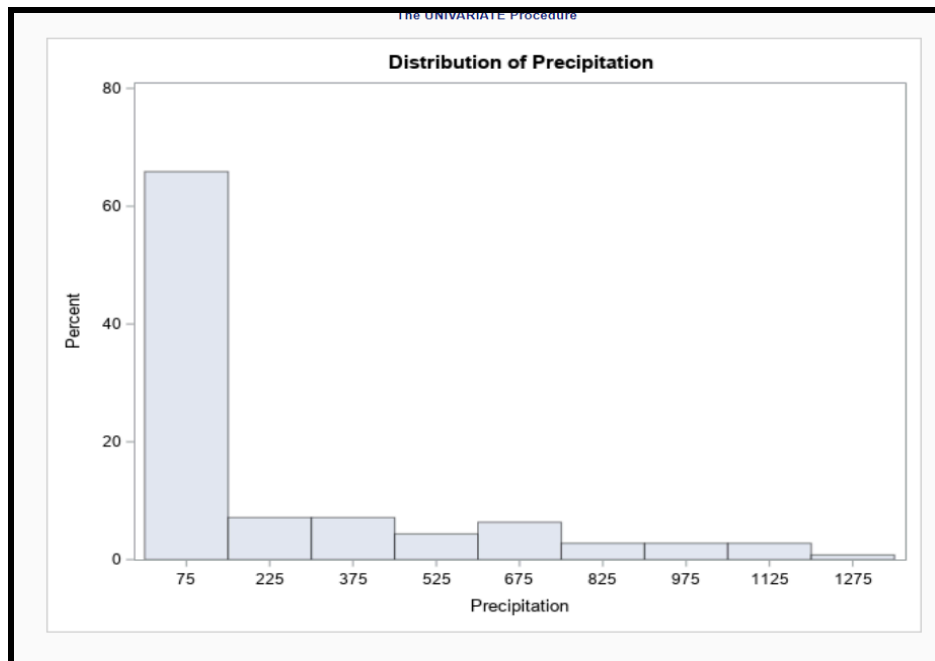
\*Program 1

CODELOGRESULTSOUTPUT DATA

Line #

```
1 data stsm.forecasting_finalproject;
2 set stsm.forecasting_finalproject;
3 /* Assuming your_variable contains the data you want to analyze */
4
5 /* Calculate lower and upper bounds for Precipitation based on quantiles */
6 lower_bound_Precipitation = 0.395; /* 0% quantile */
7 upper_bound_Precipitation = 355.900; /* 100% quantile */
8
9 /* Flag outliers for Precipitation */
10 if Precipitation < lower_bound_Precipitation or Precipitation > upper_bound_Precipitation then
11     outlier_flag_Precipitation = 1;
12 else
13     outlier_flag_Precipitation = 0;
14 run;
15
```

Temperature	Precipitation	lower_bound_Precipitation	upper_bound_Precipitation	outlier_flag_Precipitation
23.93	0	0.395	355.9	1
25.83	0.11	0.395	355.9	1
26.68	0.01	0.395	355.9	1
22.49	0.02	0.395	355.9	1
19.07	271.14	0.395	355.9	0
7.91	313.67	0.395	355.9	0
6.67	820.45	0.395	355.9	1
7.07	362.38	0.395	355.9	1
10.63	97.85	0.395	355.9	0
15.38	63.41	0.395	355.9	0
17.48	4.37	0.395	355.9	0
20.09	11.25	0.395	355.9	0
22.94	0	0.395	355.9	1
22.7	0	0.395	355.9	1
20.97	0.03	0.395	355.9	1
22.73	1.57	0.395	355.9	0



After carrying out outlier detection, we decided to use log transformation. We chose this approach mainly because our data is right-skewed and the majority of the outliers are higher values.

### Log Transformation :

Results:

Precipitation	outlier_flag_Precipitation	log_Precipitation
0	1	0
0.11	1	0.1043600153
0.01	1	0.0099503309
0.02	1	0.0198026273
271.14	0	5.6063166398
313.67	0	5.7515244706
820.45	1	6.7110710714
362.38	1	5.8954491187
97.85	0	4.5936035496
63.41	0	4.1652689006
4.37	0	1.6808279085
11.25	0	2.505525937
0	1	0
0	1	0

**Transformed Data:**



Total rows: 252 Total columns: 9								
	Date	SpecificHumidity	RelativeHumidity	Temperature	Precipitation	outlier_flag_Precipitation	log_Precipitation	
1	01/01/2000	8.06	48.25	23.93	0	1	0	
2	02/01/2000	8.73	50.81	25.83	0.11	1	0.1043600153	
3	03/01/2000	8.48	42.88	26.68	0.01	1	0.0099503309	
4	04/01/2000	13.79	55.69	22.49	0.02	1	0.0198026273	
5	05/01/2000	17.4	70.88	19.07	271.14	0	5.6063166398	
6	06/01/2000	19.53	84.19	7.91	313.67	0	5.7515244706	
7	07/01/2000	18.8	88.5	6.67	820.45	1	6.7110710714	
8	08/01/2000	18.86	88.44	7.07	362.38	1	5.8954491187	
9	09/01/2000	18.43	86.12	10.63	97.85	0	4.5936035496	
10	10/01/2000	16.72	78.38	15.38	63.41	0	4.1652689006	
11	11/01/2000	12.02	63.69	17.48	4.37	0	1.6808279085	
12	12/01/2000	7.39	44.56	20.09	11.25	0	2.505525937	

After the transformation, "log\_precipitation" becomes our target variable.

## 4. Time Series Exploration:

Input:

DATA ANALYSES INFORMATION

DATA

STSM.FORECASTING\_FINALPROJECT

Filter: (none)

ROLES

Dependent variable: (1 item)

log\_Precipitation

Independent variables:

SpecificHumidity

RelativeHumidity

Temperature

Transformations

Variable	Accumulation	Transformation
log_Precipitation	None	None
SpecificHumidity	None	None
RelativeHumidity	None	None
Temperature	None	None

DATA ANALYSES INFORMATION

ANALYSES

SERIES PLOTS

Time Series

Series histogram

Seasonal cycles

STATISTICS

AUTOCORRELATION ANALYSIS

Perform autocorrelation analysis

Select plots to display:

Default plots

Number of lags:

Use default value

CROSS-CORRELATION ANALYSIS

Perform cross-correlation analysis

Plots

Cross-series

Cross-correlation function

Normalized cross-correlation function

Number of lags:

Use default value

DECOMPOSITION ANALYSIS

Perform decomposition analysis

Select plots to display:

Selected plots

Plots

Decomposition panel

Components

Select components:

Seasonal component

Irregular component

"log\_Precipitation" is the dependent variable for our time series analysis. The time identifier is "Date", and the independent variables include "SpecificHumidity", "RelativeHumidity", and "Temperature".

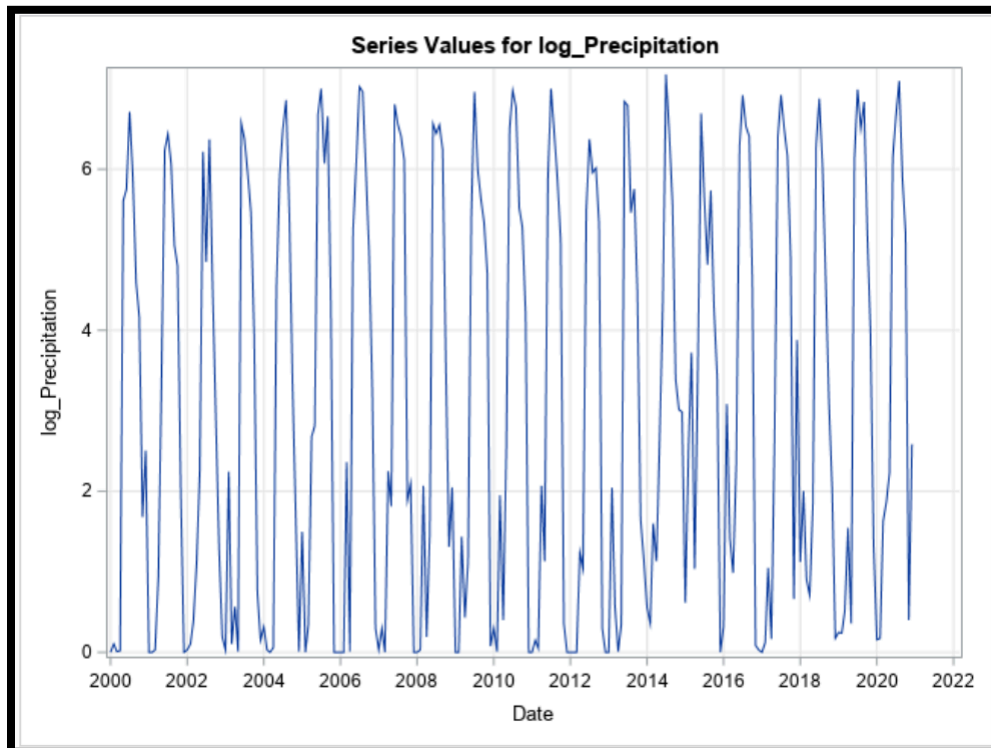
Output:

Input Data Set	
Name	WORK.PREPROCESSEDDATA
Label	
Time ID Variable	Date
Time Interval	MONTH
Length of Seasonal Cycle	12

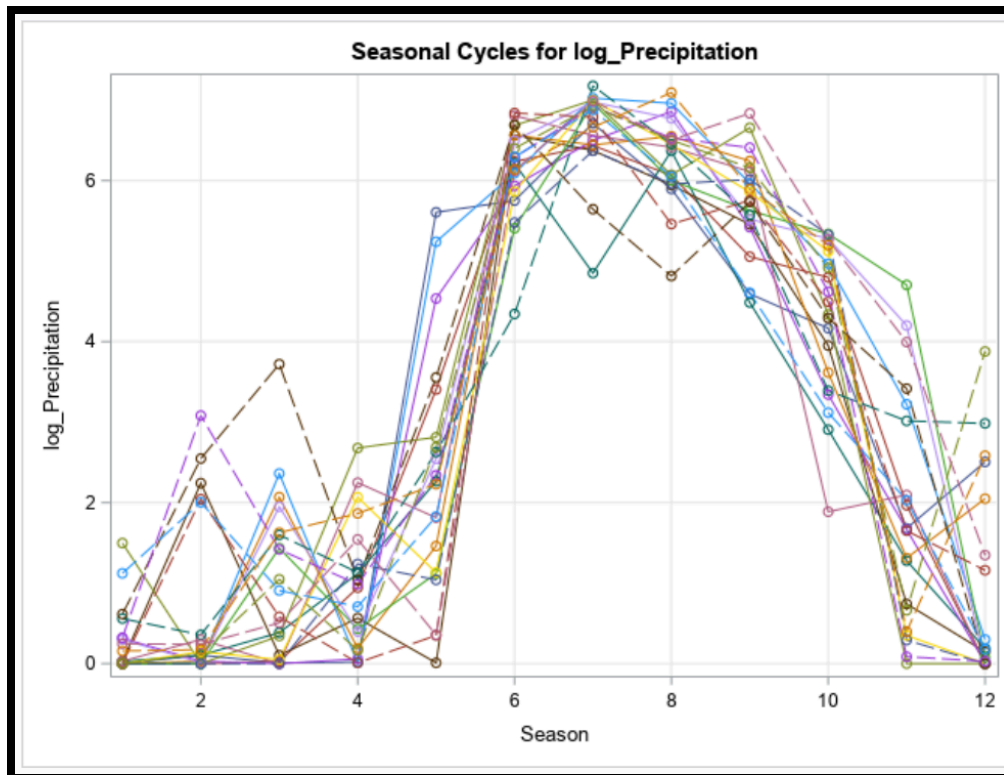
Length of our Seasonal Cycle is 12

Precipitation is plotted over the years in the graphs below:

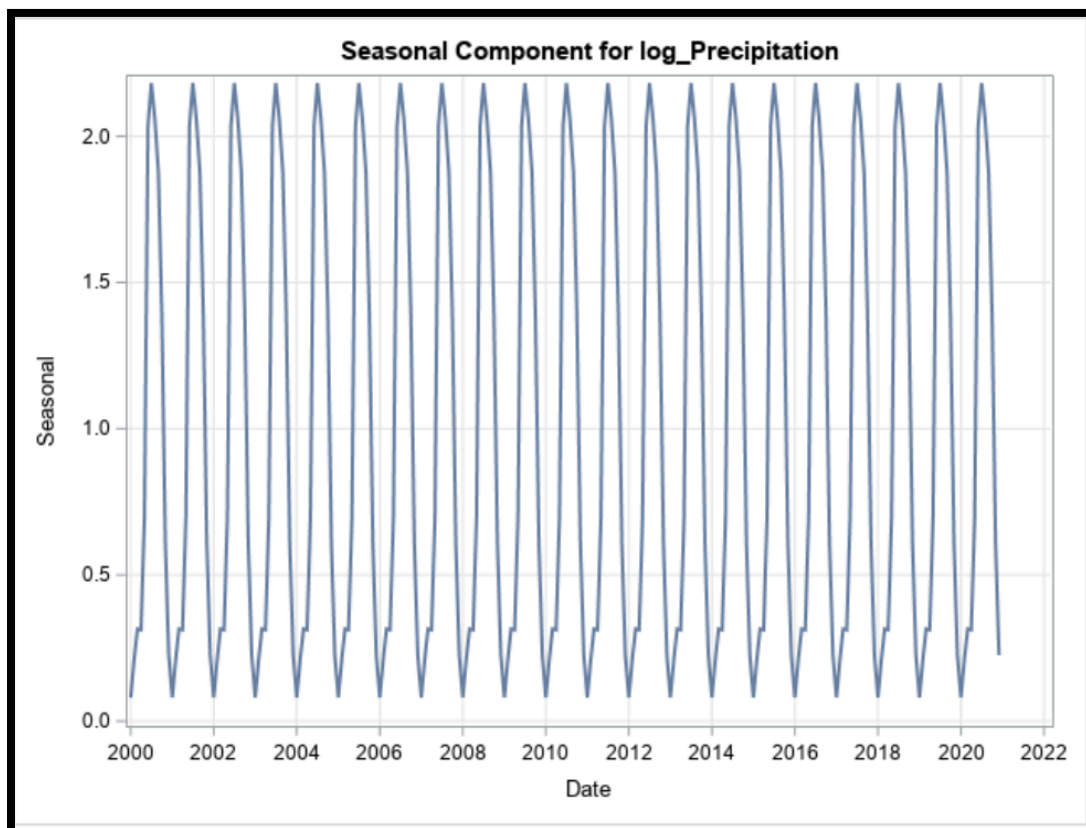
**Series Value Graph for Precipitation:**



**Seasonal Cycle Graph for Precipitation**

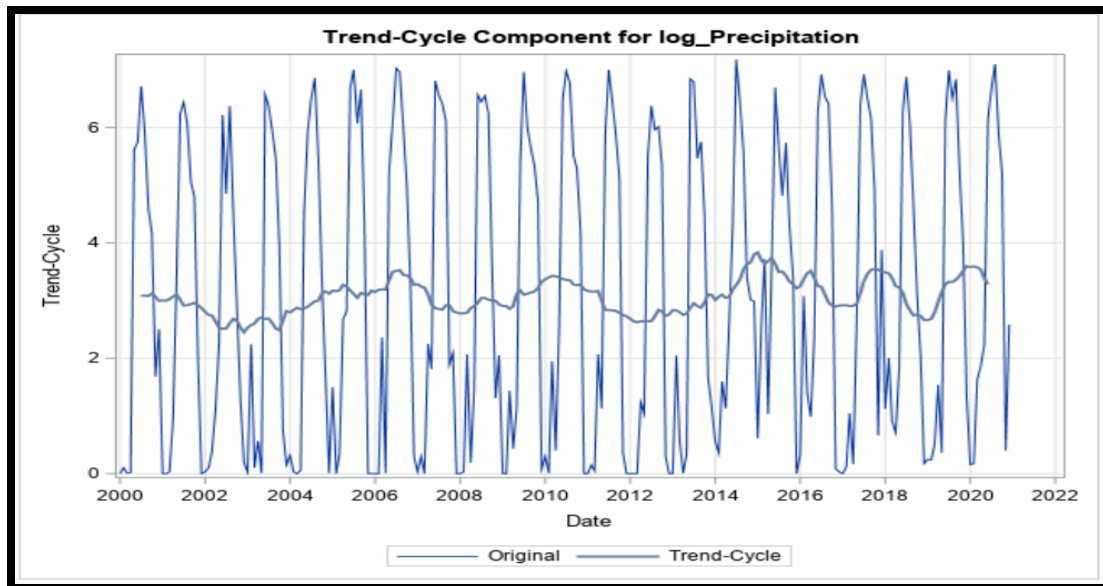


**Seasonal Component Graph:**



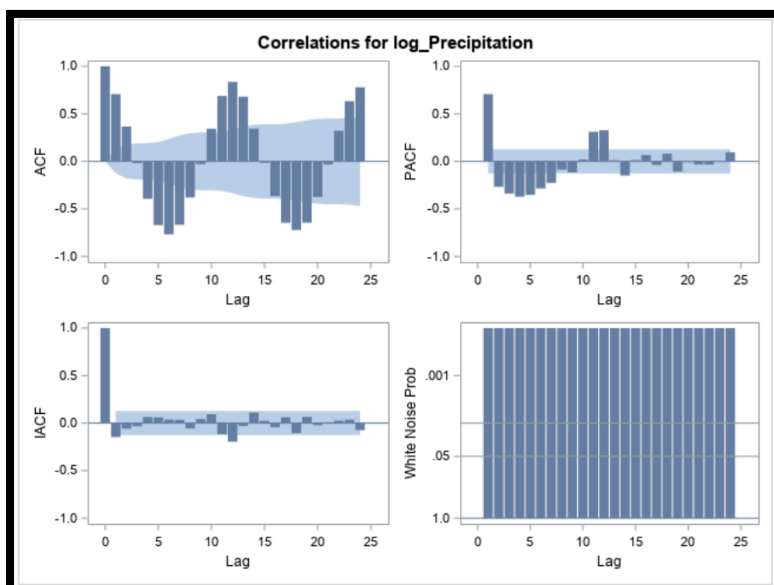
The seasonal plot and seasonal cycle plot clearly reveal seasonality in the series. From the graphs, we can observe significant precipitation around mid-year, specifically in the months of July and August.

### Trend Component Graph:



From the above graph we can observe that there is no particular trend followed for the data.

### Correlation Graphs:



The Auto Correlation (ACF) graph reveals a clear seasonality associated with precipitation. Significant lags can be observed at various intervals in the PACF graph. The White Noise Probability graph notably lacks white noise, suggesting a significant signal that can be harnessed through modeling.

To further check the stationarity ,we have performed Augmented Dickey-Fuller Test

### Augmented Dickey-Fuller Test:

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-30.2520	<.0001	-3.99	<.0001		
	1	-40.9815	<.0001	-4.49	<.0001		
	2	-58.6573	<.0001	-5.07	<.0001		
Single Mean	0	-73.4430	0.0015	-6.57	<.0001	21.57	0.0010
	1	-127.033	0.0001	-7.96	<.0001	31.70	0.0010
	2	-353.498	0.0001	-10.14	<.0001	51.45	0.0010
Trend	0	-73.6188	0.0006	-6.56	<.0001	21.51	0.0010
	1	-127.696	0.0001	-7.96	<.0001	31.66	0.0010
	2	-357.518	0.0001	-10.14	<.0001	51.44	0.0010

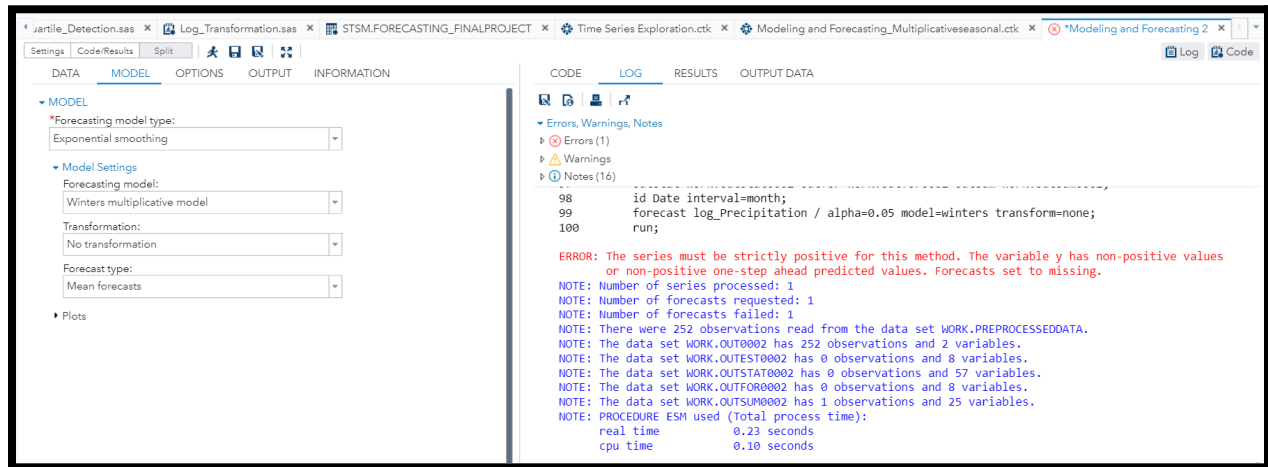
Based on the Augmented Dickey-Fuller Unit Root Test analysis results, we observe that both the Rho and Tau values for Precipitation are less than 0.05. Since our series exhibits a mean value, we employ a single mean check. Consequently, we have sufficient evidence to reject the null hypothesis that our data is non-stationary. As a result, we can confidently conclude that the data is stationary, and there's no necessity to perform differencing on the series.

## 5. Modeling and Forecasting:

### 5.1 Exponential Smoothing Models:

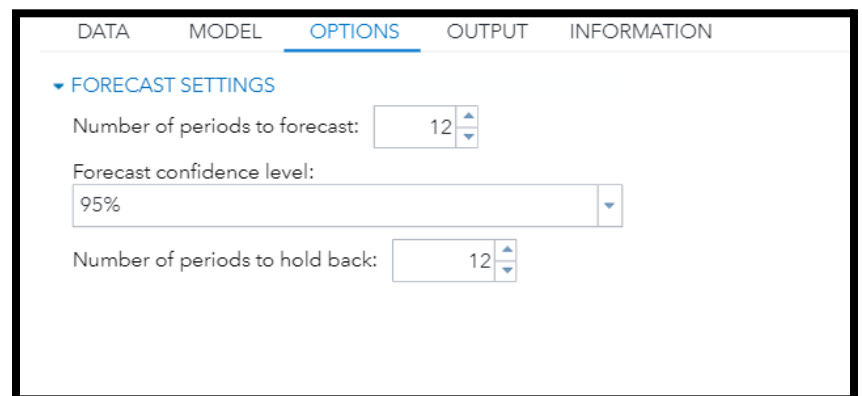
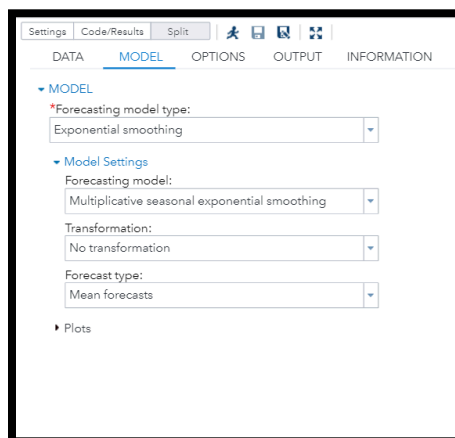
#### 5.1.1 Winters Multiplicative Model:

Given that the data is stationary, exhibits seasonality, and lacks any discernible trend, we initially considered modeling with Winters Multiplicative. However, due to the presence of "0" values in the Precipitation column, this model is not suitable.

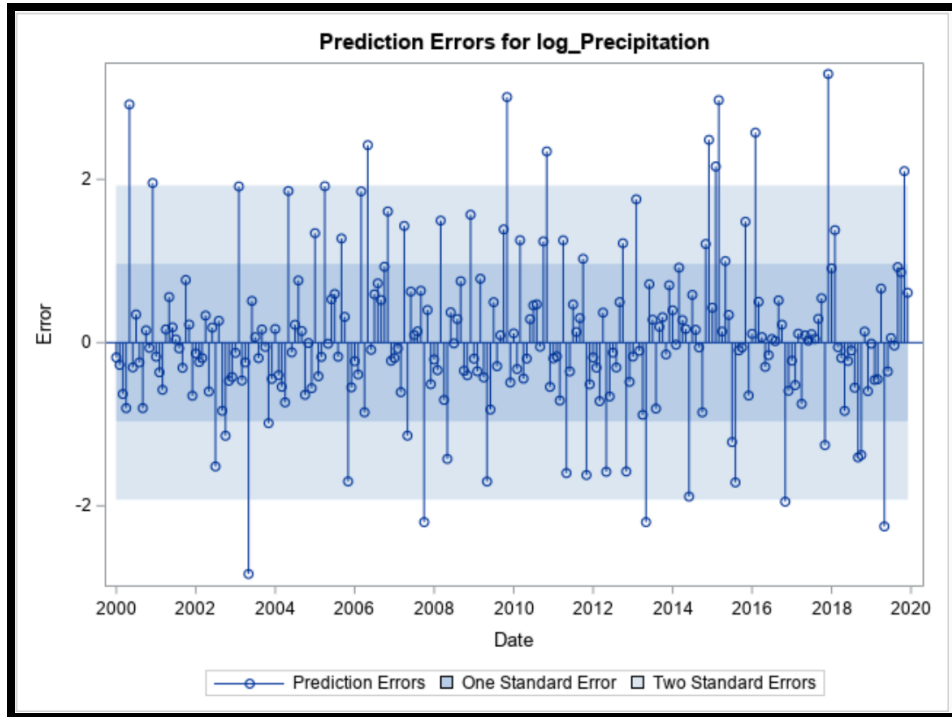


### 5.1.2 Multiplicative Seasonal Exponential Smoothing Model:

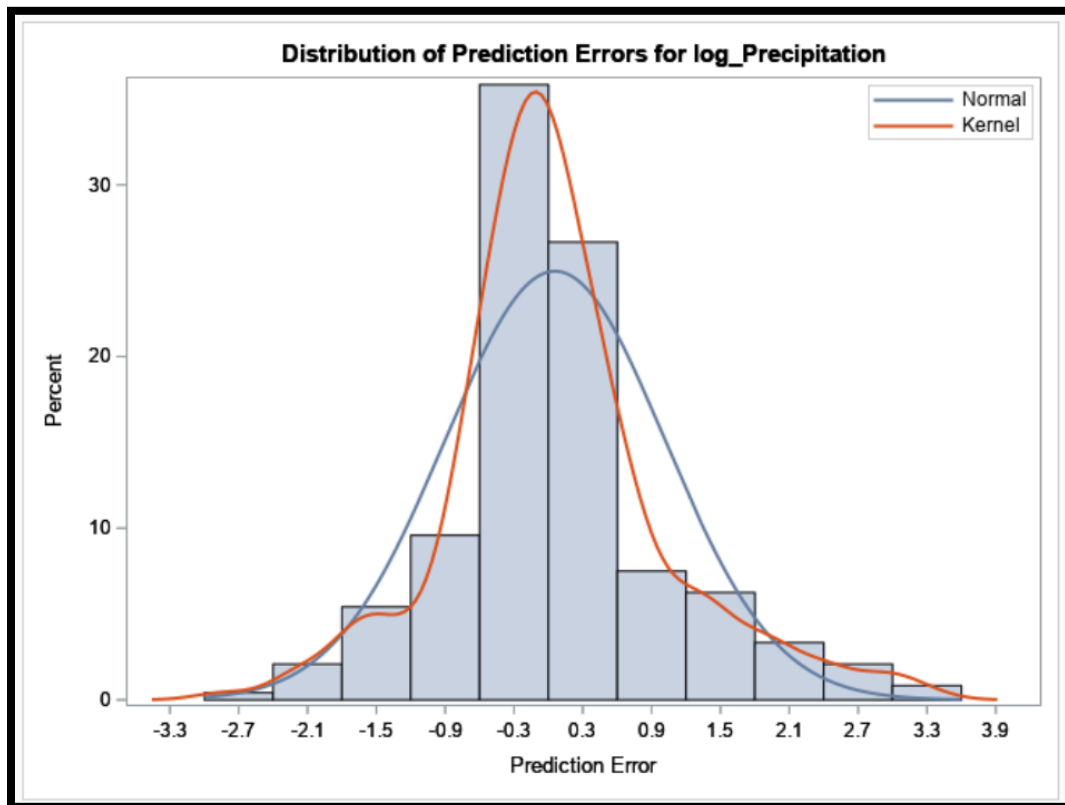
We then proceeded to use the Multiplicative Seasonal Exponential Smoothing Model, following the steps outlined below in SAS

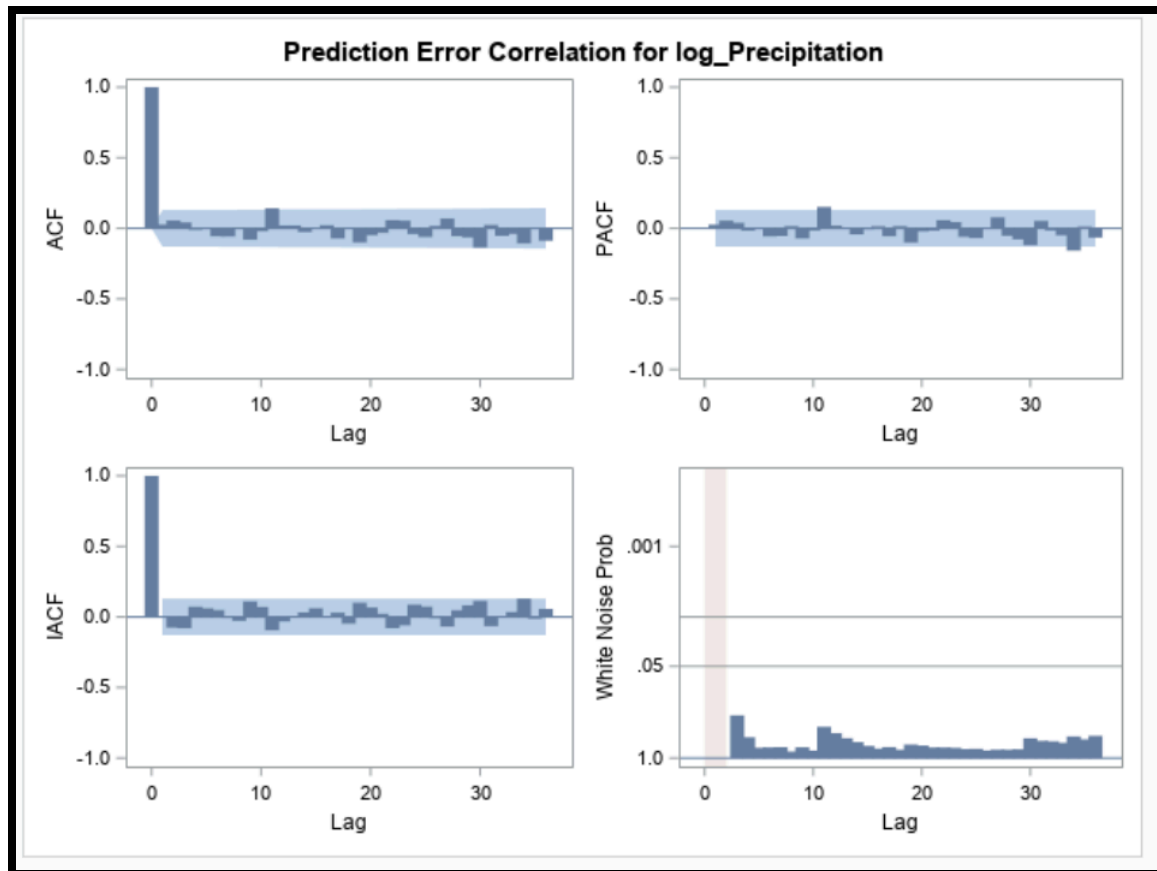


We used "log\_Precipitation" as the dependent variable and "Date" as the Time ID. Our approach was to forecast for a complete season while retaining data from one entire season for validation.



We see that there are errors that stand outside the one standard and two standard errors.





From the ACF and PACF graphs presented, it's clear that there aren't any significant lags. Furthermore, based on the White Noise Probability Test, the residuals appear to be distributed almost like white noise, indicating that most of the signal has been extracted by the model.

## Results:

Total rows: 2 Total columns: 57						Rows 1-2	
	MSE	RMSE	MAE	AIC	SBC		
1	0.9182887231	0.9582738247	0.6793647701	-16.45842191	-9.497144068		
2	0.7180731069	0.8473919441	0.6613728111	-3.974206739	-3.974206739		

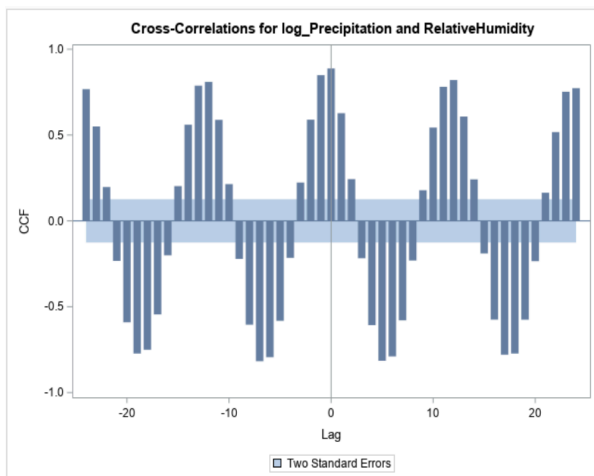
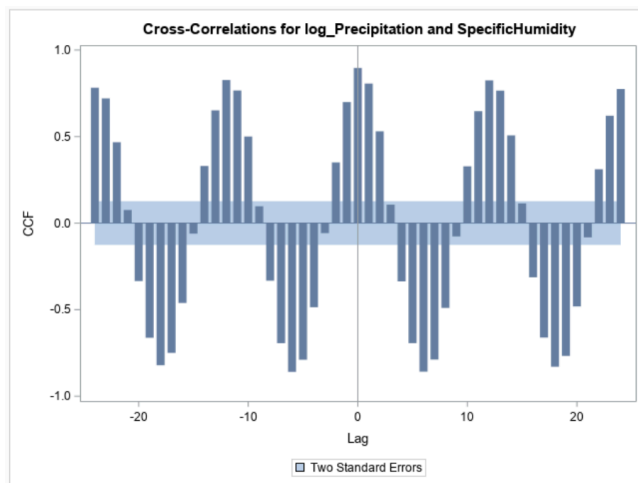
Based on the provided statistics, the low MAPE and RMSE values for both the forecast and fit suggest that we are making precise predictions regarding future precipitation.

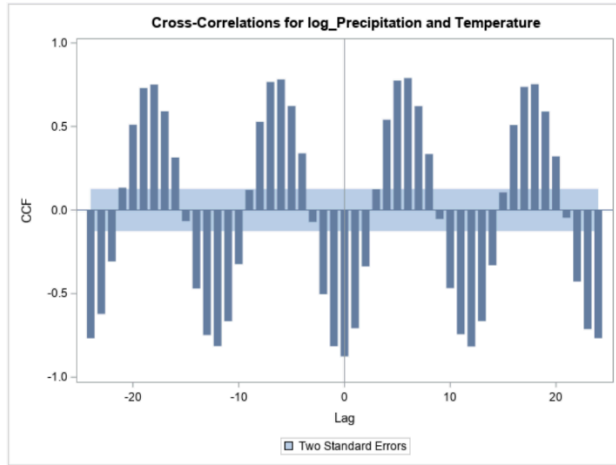


Similarly, the low AIC and SBC values indicate an efficient balance between the model's fit and complexity.

Given that our dataset contains 3 independent variables, it was essential to understand their relationship and potential influence on the target variable. To achieve this, we undertook a cross-correlation analysis. This method allows us to explore any significant associations or lags between the independent variables and the target variable. By examining these relationships, we can better ascertain the impact of each independent variable on our target and adjust our modeling strategy accordingly. To generate these graphs, we conducted a Cross Correlation Analysis using SAS.

### Cross Correlation Graphs:





Above correlations graphs indicate that the precipitation has significant dependence on all the variables i.e., Specific Humidity, Relative Humidity, Temperature at lag 0, lag 5 and lag 10 respectively.

## 5.2 Pre-Whitening:

As part of our analysis, we are performing prewhitening, an essential preliminary step, especially in the context of ARMA (AutoRegressive Moving Average) modeling, before delving into the ARMA model development. Its primary purpose is to eliminate the autocorrelation structure present within a time series. Autocorrelation, where data points at different time steps are correlated, can complicate the analysis and modeling process. By prewhitening the data, this correlation is effectively removed, rendering the data more amenable to modeling. The transformed data becomes akin to white noise, a stationary series with no autocorrelation, simplifying the modeling process. Prewhitening also enhances the stability of parameter estimation, improves model performance, and aids in diagnostic checks by ensuring that residuals exhibit white noise properties, aligning with the assumptions underlying ARMA modeling. In essence, prewhitening is a critical step that paves the way for more accurate, reliable, and interpretable ARMA model results and predictions, and we perform it diligently before embarking on ARMA model development.

### 5.2.1 Prewhitening (Temperature SARIMA (2,0,2 3,0,3)):

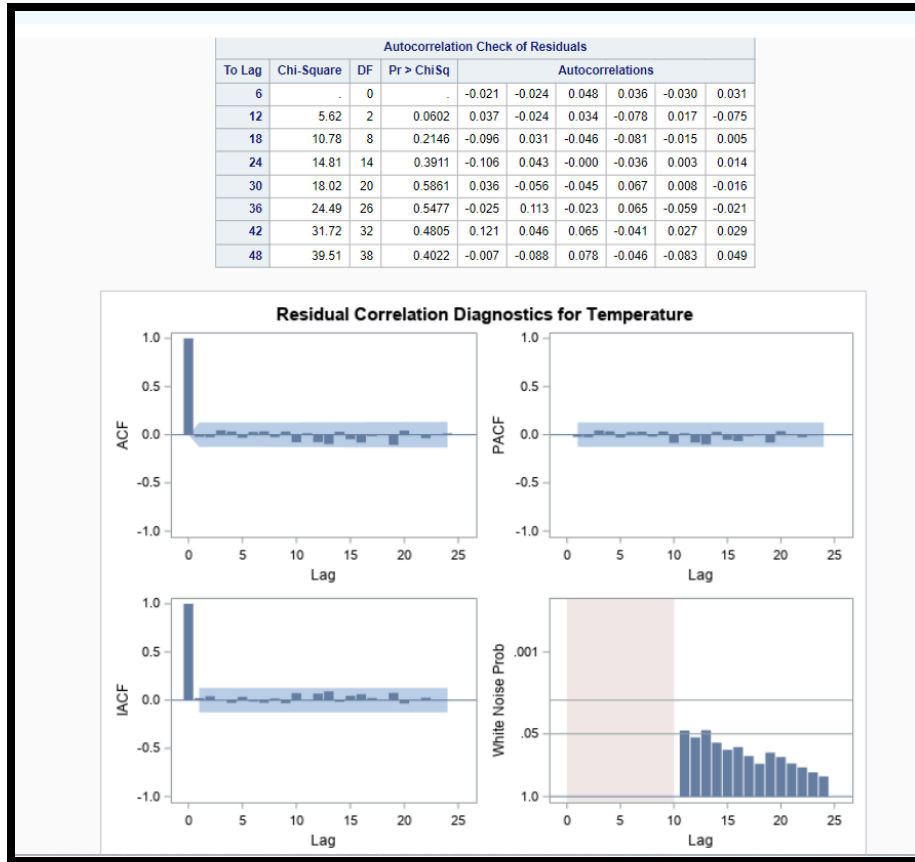
In our analytical process, we initially applied the SARIMA (2,0,2 3,0,3) model to the temperature data, given its evident sinusoidal seasonality. The model's output, particularly the p,q values and the reduced significant lags observed in both the

ACF and PACF, were instrumental. Leveraging this insight, we then applied the same SARIMA model to our primary dependent variable, Precipitation, while using temperature as an independent variable. SARIMA, standing for Seasonal Autoregressive Integrated Moving Average, is particularly suited for time series data with clear seasonality. Given that both temperature and precipitation patterns can have distinct seasonal components, using SARIMA allowed us to effectively capture the underlying time series structure and the seasonal intricacies. Thus, by modeling temperature first and then applying the findings to precipitation, we aimed to harness the seasonality in temperature to yield more precise precipitation forecasts.

### **Prewhitening Model for Temperature(2,0,2 3,0,3)**

```
proc sort data=STSM.FORECASTING_FINALPROJECT out=Work.preProcessedData;
    by Date;
run;

proc arima data=Work.preProcessedData plots
    (only)=(series(corr crosscorr) residual(corr normal)
        forecast(forecastonly));
    identify var=Temperature;
    estimate p=(1 2) (12 24 36) q=(1 2) (12 24 36) method=ML;
    forecast lead=12 back=0 alpha=0.05 id=Date interval=month;
    outlier;
run;
quit;
```



The residuals of the model appear to have white noise characteristics over the majority of lags, indicating that the model has captured the majority of the information and structure from the data. However, certain lags, particularly those between 12 and 48, may still contain significant autocorrelation that should be examined further.

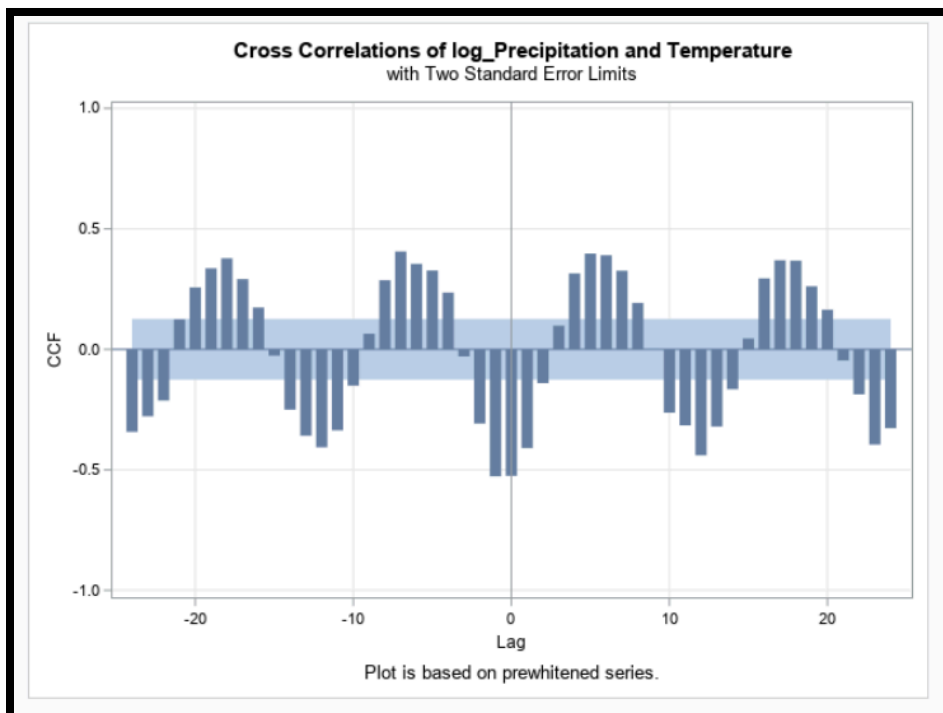
## Code:

```
ods noproctitle;
ods graphics / imagemap=on;

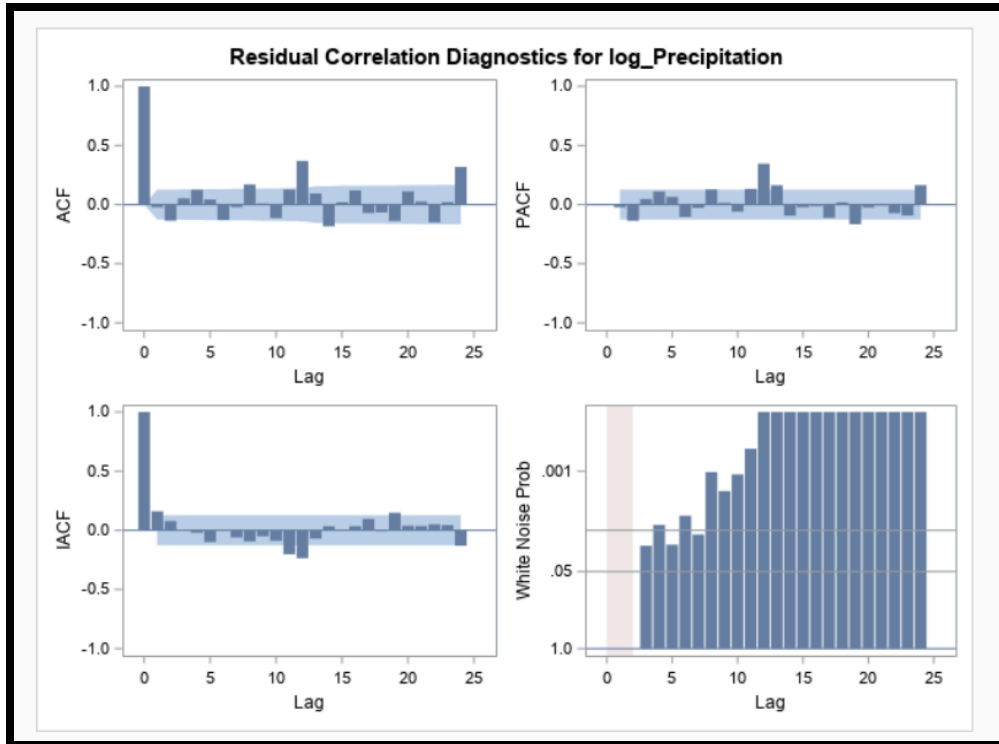
proc sort data=STSM.FORECASTING_FINALPROJECT out=Work.preProcessedData;
  by Date;
run;

proc arima data=Work.preProcessedData plots
  (only)=(series(corr crosscorr) residual(corr normal)
    forecast(forecastonly));
  identify var=Temperature;
  estimate p=(1 2) (12 24 36) q=(1 2) (12 24 36) method=ML;
  identify var=log_Precipitation crosscorr=(Temperature);
  estimate p=(5) q=(5)
  input=(Temperature) method=ML;
  forecast lead=12 back=0 alpha=0.05 id=Date interval=month;
  outlier;
run;
quit;

proc delete data=Work.preProcessedData;
run;
```



After prewhitening, from the above graph, we say that the prewhitening reduced the dependency of the past values of temperature on the precipitation. We see the major significance at lag 0 alone.



Based on the ACF and PACF plots, the residuals for log\_Precipitation show no significant autocorrelation, indicating that the model well represents the underlying pattern. The White Noise Probability plot, on the other hand, reveals probable non-white noise at specific lags.

### 5.2.2 Prewhitening (Relative Humidity (1,0,1 2,0,2)):

Similar to the Temperature data, here we applied the SARIMA (1,0,1 2,0,2) model to the Relative Humidity data, given its evident sinusoidal seasonality. The model's output, particularly the p,q values and the reduced significant lags observed in both the ACF and PACF, were instrumental. Leveraging this insight, we then applied the same SARIMA model to our primary dependent variable, Precipitation, while using Relative Humidity as an independent variable. Thus, by modeling Relative Humidity first and then applying the findings to precipitation, we aimed to harness the seasonality in Relative Humidity to yield more precise precipitation forecasts.

▼ MODEL

\*Forecasting model type:

ARIMA

▼ Model Settings

▼ ARIMA

Autoregressive order (p): 1

Differencing order (d): 0

Moving average order (q): 1

▼ Seasonal ARIMA

Autoregressive order (P): 2

Differencing order (D): 0

Moving average order (Q): 2

☒ Include intercept in model

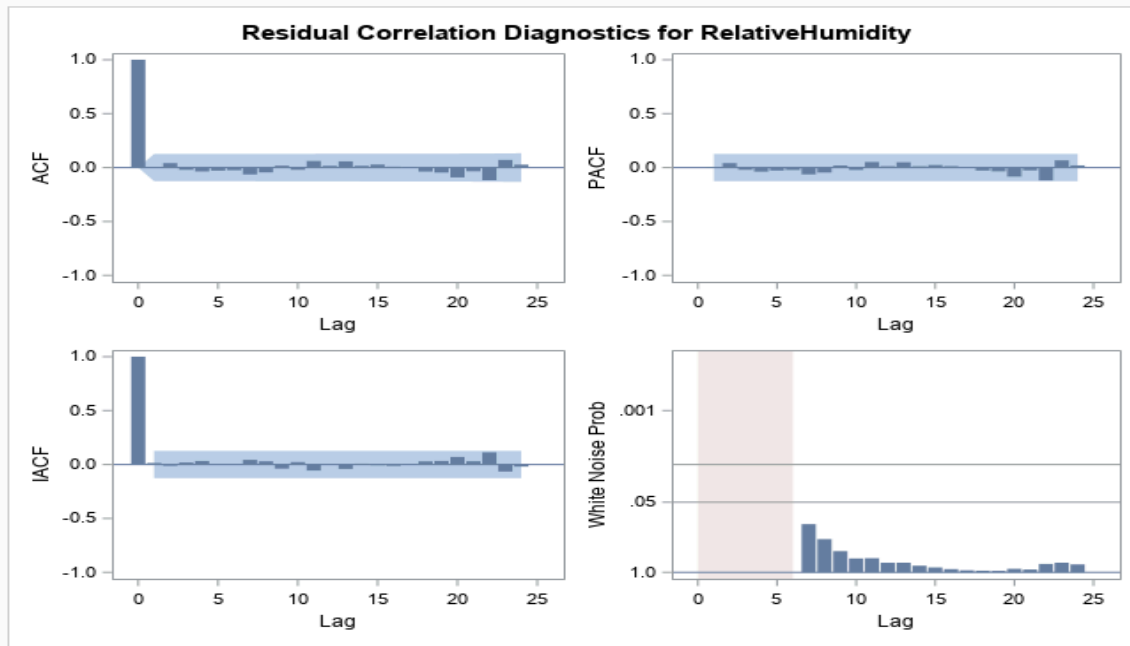
► Plots

```

21
22 proc arima data=Work.preProcessedData plots
23     (only)=(series(corr crosscorr) residual(corr normal)
24         forecast(forecastonly));
25     identify var=RelativeHumidity;
26     estimate p=(1) (12 24) q=(1) (12 24) method=ML;
27     forecast lead=12 back=0 alpha=0.05 id=Date interval=month;
28     outlier;
29     run;
30 quit;
31

```

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.001	0.043	-0.018	-0.034	-0.027	-0.023
12	4.01	6	0.6753	-0.063	-0.043	0.019	-0.019	0.062	0.018
18	5.61	12	0.9347	0.058	0.017	0.028	0.009	-0.002	-0.036
24	14.14	18	0.7198	-0.046	-0.091	-0.034	-0.116	0.071	0.027
30	19.98	24	0.6977	0.102	-0.090	-0.001	-0.024	0.040	0.001
36	24.45	30	0.7515	-0.036	0.002	-0.046	0.011	-0.083	0.069
42	32.83	36	0.6204	0.068	0.025	0.094	-0.006	0.061	-0.099
48	34.49	42	0.7882	-0.008	-0.060	-0.021	-0.007	0.034	0.005



The residuals of the model appear to have white noise characteristics over the majority of lags, indicating that the model has captured the majority of the information and structure from the data.

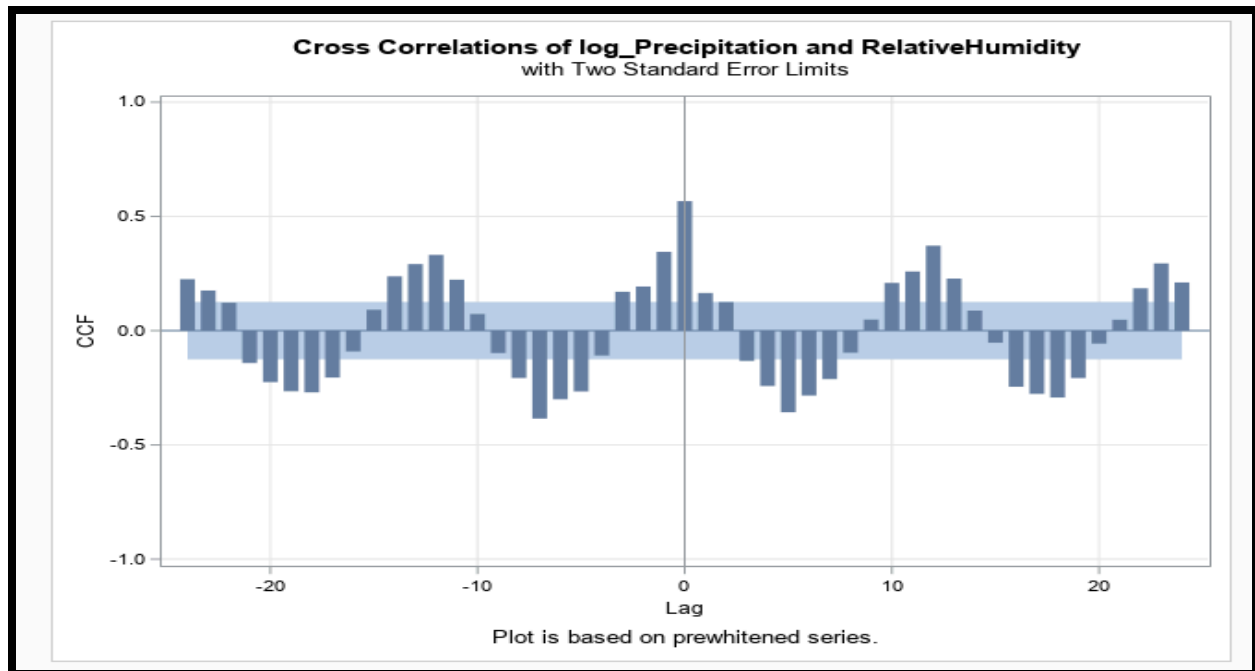
**Code:**

```

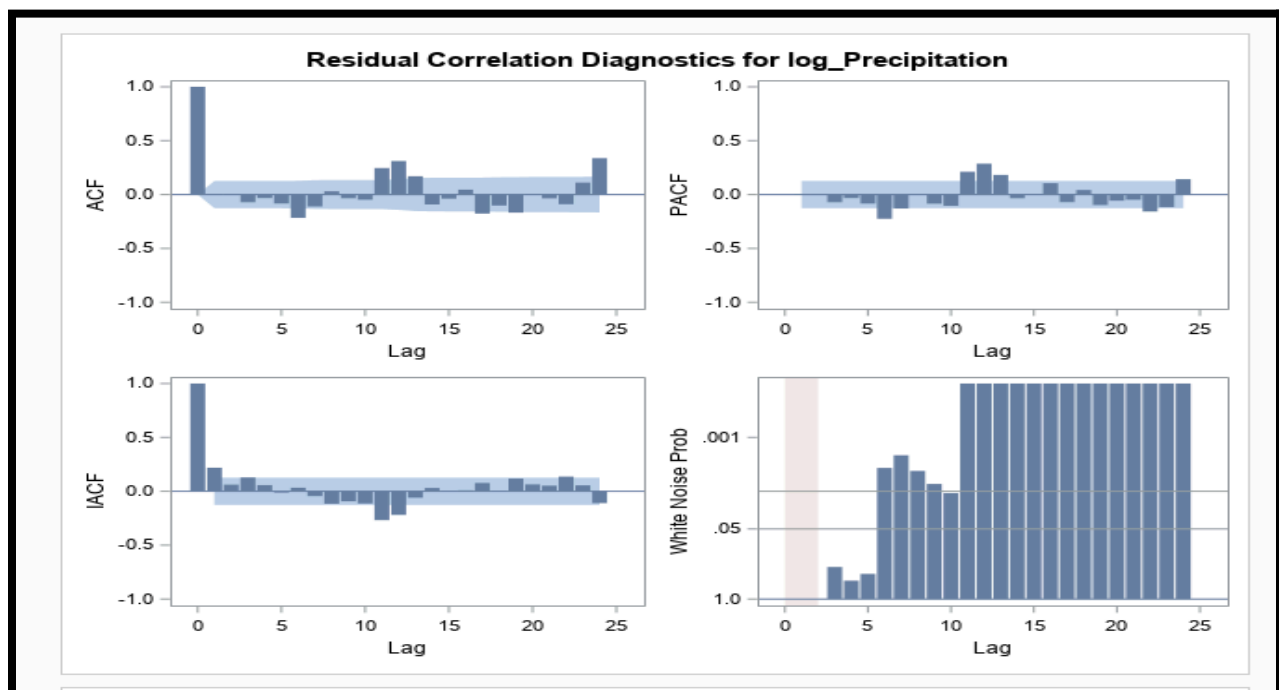
22 proc arima data=Work.preProcessedData plots
23   (only)=(series(corr crosscorr) residual(corr normal)
24     forecast(forecastonly));
25   identify var=RelativeHumidity;
26   estimate p=(1) (12 24) q=(1) (12 24) method=ML;
27   identify var=log_Precipitation crosscorr=(RelativeHumidity);
28   estimate p=(1) q=(1) input=(RelativeHumidity) method=ML;
29   forecast lead=12 back=0 alpha=0.05 id=Date interval=month;
30   outlier;
31   run;
32 quit;

```





After prewhitening from the above graph, we say that the prewhitening reduced the dependency of the past values of Relative Humidity on the precipitation and we see a significant lag 0.



Based on the ACF and PACF plots, the residuals for log\_Precipitation show no significant autocorrelation, indicating that the model well represents the underlying

pattern. The White Noise Probability plot, on the other hand, reveals probable non-white noise at specific lags.

### 5.2.3 Prewhitening (Specific Humidity (2,0,1 1,0,1)):

Same as the above two independent variables, we applied the SARIMA (2,0,1 1,0,1) model to the Specific Humidity data, given its evident sinusoidal seasonality. The model's output, particularly the p,q values and the reduced significant lags observed in both the ACF and PACF, were instrumental. Leveraging this insight, we then applied the same SARIMA model to our primary dependent variable, Precipitation, while using Specific Humidity as an independent variable. Thus, by modeling Specific Humidity first and then applying the findings to precipitation, we aimed to harness the seasonality in Specific Humidity to yield more precise precipitation forecasts.

▼ MODEL

\*Forecasting model type: ARIMA

▼ Model Settings

▼ ARIMA

Autoregressive order (p): 2

Differencing order (d): 0

Moving average order (q): 1

▼ Seasonal ARIMA

Autoregressive order (P): 1

Differencing order (D): 0

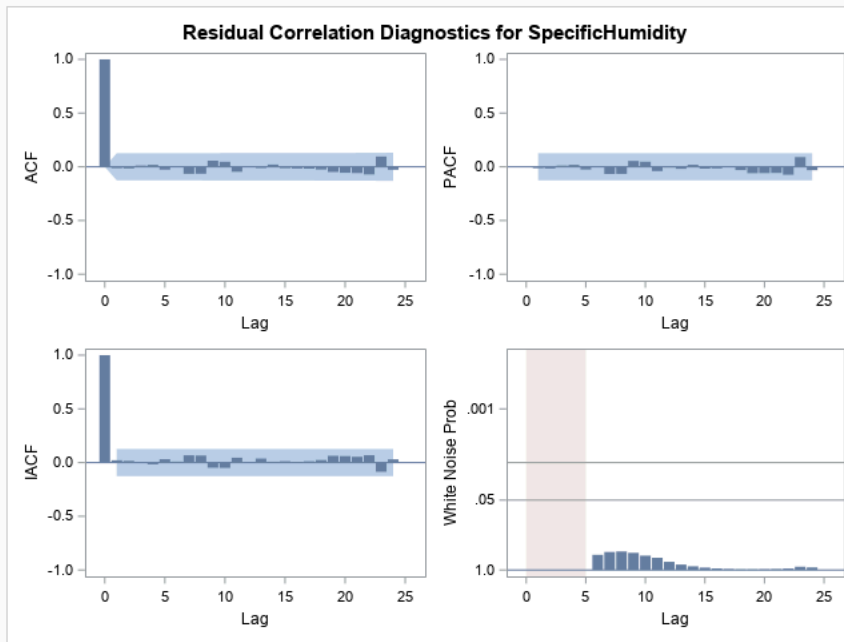
Moving average order (Q): 1

☒ Include intercept in model

► Plots

```
22 proc arima data=Work.preProcessedData plots
23     (only)=(series(corr crosscorr) residual(corr normal)
24         forecast(forecastonly));
25     identify var=SpecificHumidity;
26     estimate p=(1 2) (12) q=(1) (12) method=ML;
27     forecast lead=12 back=0 alpha=0.05 id=Date interval=month;
28     outlier;
29     run;
30 quit;
31
```

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.41	1	0.5244	-0.006	-0.008	0.018	0.025	-0.021	0.010
12	4.41	7	0.7319	-0.061	-0.060	0.061	0.050	-0.041	0.001
18	4.85	13	0.9783	-0.007	0.025	-0.010	-0.012	-0.015	-0.023
24	11.16	19	0.9184	-0.046	-0.053	-0.056	-0.069	0.097	-0.025
30	14.44	25	0.9535	0.059	-0.062	0.046	-0.019	0.030	0.029
36	15.02	31	0.9930	-0.031	-0.016	0.009	-0.028	0.001	-0.007
42	19.56	37	0.9917	-0.044	0.031	0.081	0.026	0.038	-0.060
48	22.22	43	0.9983	0.030	-0.038	0.006	-0.000	-0.014	0.077



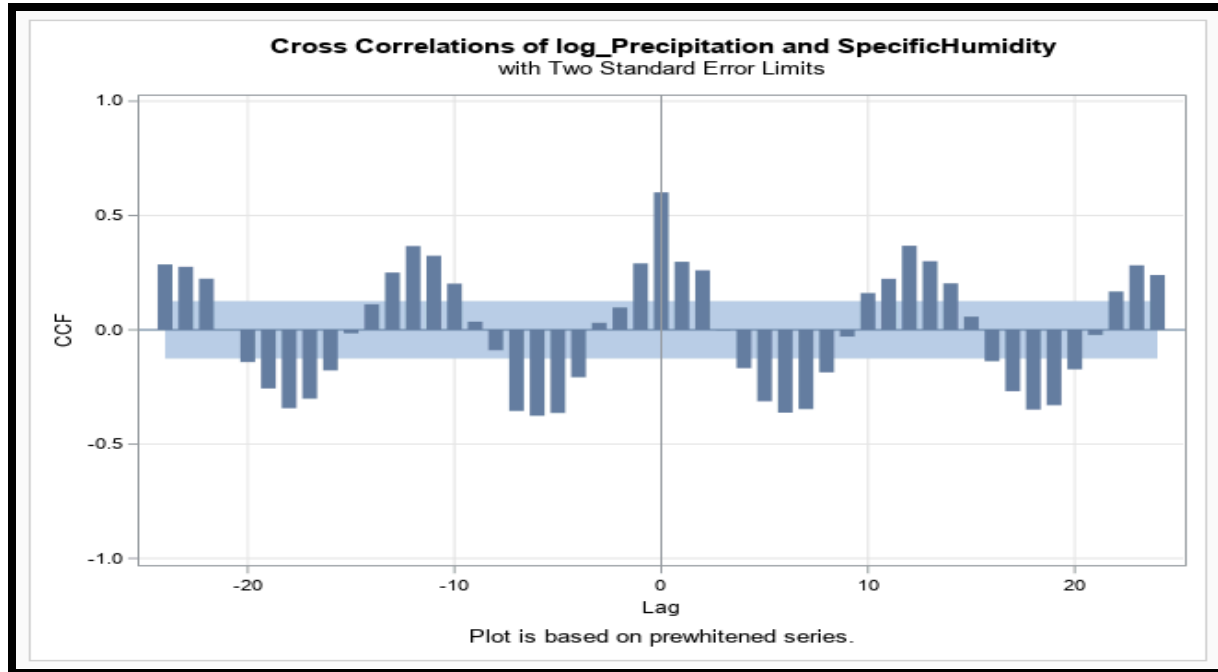
The residuals of the model appear to have white noise characteristics over the majority of lags, indicating that the model has captured the majority of the information and structure from the data.

**Code:**

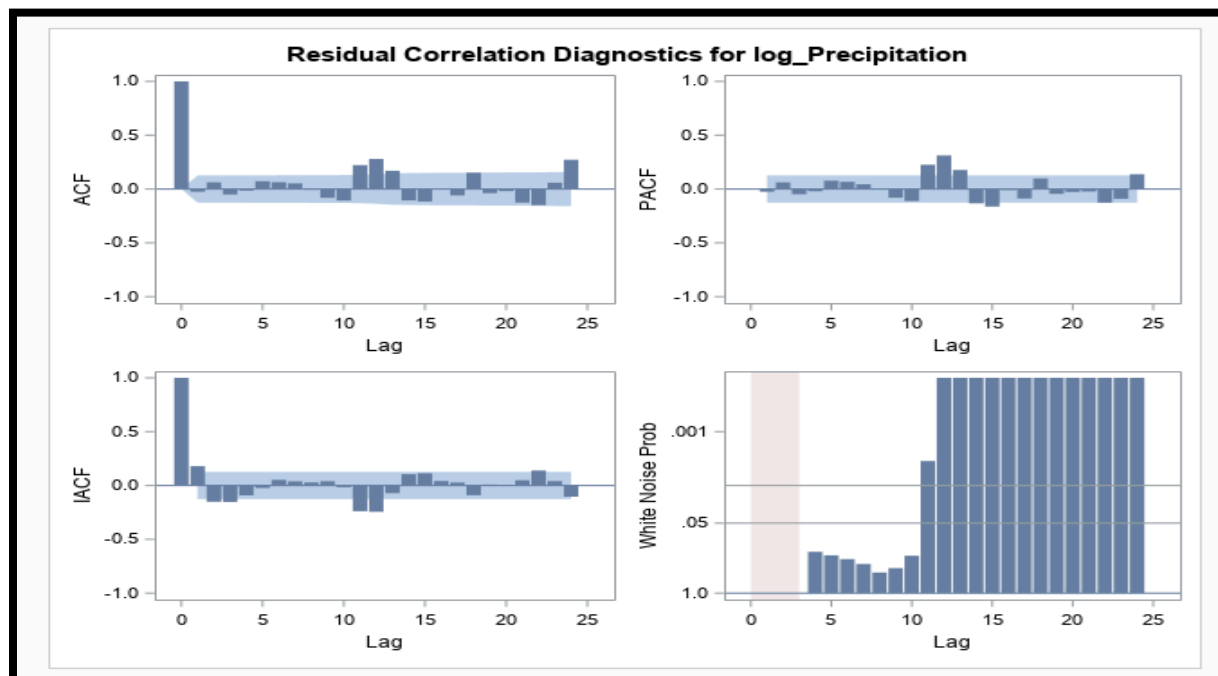
```

18 proc sort data=STSM.FORECASTING_FINALPROJECT out=Work.preProcessedData;
19     by Date;
20 run;
21
22 proc arima data=Work.preProcessedData plots
23     (only)=(series(corr crosscorr) residual(corr normal)
24         forecast(forecastonly));
25     identify var=SpecificHumidity;
26     estimate p=(1 2) (12) q=(1) (12) method=ML;
27     identify var=log_Precipitation crosscorr=(SpecificHumidity);
28     estimate p=(1 2) (12) q=(1) (12) input=(SpecificHumidity) method=ML;
29     forecast lead=12 back=0 alpha=0.05 id=Date interval=month;
30     outlier;
31     run;
32 quit;
33

```



After prewhitening from the above graph, we say that the prewhitening reduced the dependency of the past values of Specific Humidity on the precipitation and we see a significant lag 0.



Based on the ACF and PACF plots, the residuals for log\_Precipitation show no significant autocorrelation, indicating that the model well represents the underlying

pattern. The White Noise Probability plot, on the other hand, reveals probable non-white noise at specific lags.

After prewhitening for all the independent variables, with the prewhitened graphs, we can observe that all independent variables have a significant impact at lag0.

### 5.3 ARIMAX Modeling:

For the ARIMAX modeling, the number of forecasts is 12 and also the number of holdout sets is set to 12 to show us seasonality for at least one cycle.

#### 5.3.1 ARIMAX(2,0,2) (1,2,1) Model (SEASONAL):

In the ARIMAX model, we have set p and q to be 2 each. The dependent variable is Precipitation, Date serves as the Time Identifier, and all three independent variables have been included. The screenshots showing the process is as follows:

- Selecting the model and setting the p and q values. As, there is also seasonality which could be seen from the cross correlation graphs, we also added the seasonal values which are 1 for p, the differing order as 2 and the value of q as 1 respectively.

The screenshot displays the configuration interface for an ARIMAX model. It is organized into several sections:

- MODEL**: A dropdown menu for "Forecasting model type" is set to "ARIMAX".
- Model Settings**:
  - ARIMA**:
    - Autoregressive order (p): 2
    - Differencing order (d): 0
    - Moving average order (q): 2
  - Seasonal ARIMA**:
    - Autoregressive order (P): 1
    - Differencing order (D): 2
    - Moving average order (Q): 1
- Independent Variables**: A list of variables to include in the model, containing "SpecificHumidity", "RelativeHumidity", and "Temperature".
- Include intercept in model**: A checkbox that is checked.

- This picture shows the forecast setting and the outlier detection settings.

▼ FORECAST SETTINGS

Number of periods to forecast:

Forecast confidence level:

Number of periods to hold back:

▼ OUTLIER DETECTION

☒ Perform outlier detection

- The output of the above model can be seen as follows where we can see the autocorrelation check for the white noise and correlation of each independent variable.

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9.90	6	0.1290	-0.093	0.104	0.111	-0.049	0.072	-0.056
12	113.18	12	<.0001	-0.025	0.044	-0.102	-0.122	0.096	-0.623
18	123.47	18	<.0001	0.094	-0.068	-0.094	0.082	-0.039	0.118
24	137.39	24	<.0001	-0.033	-0.034	0.082	0.149	-0.057	0.152

Variable SpecificHumidity has been differenced.

Correlation of log_Precipitation and SpecificHumidity	
Period(s) of Differencing	12,12
Variance of input =	4.695027
Number of Observations	228
Observation(s) eliminated by differencing	24

Variable RelativeHumidity has been differenced.

Correlation of log_Precipitation and RelativeHumidity	
Period(s) of Differencing	12,12
Variance of input =	134.9642
Number of Observations	228
Observation(s) eliminated by differencing	24

Variable Temperature has been differenced.

Correlation of log_Precipitation and Temperature	
Period(s) of Differencing	12,12
Variance of input =	16.44721
Number of Observations	228
Observation(s) eliminated by differencing	24

- Now the below picture shows the AIC, SBC and other values along with the minimum likelihood estimations.

ARIMA Estimation Optimization Summary							
Estimation Method	Maximum Likelihood						
Parameters Estimated	10						
Termination Criteria	Maximum Relative Change in Estimates						
Iteration Stopping Value	0.001						
Criteria Value	128.7123						
Maximum Absolute Value of Gradient	108.7527						
R-Square Change from Last Iteration	0.299518						
Objective Function	Log Gaussian Likelihood						
Objective Function Value	-383.324						
Marquardt's Lambda Coefficient	0.00001						
Numerical Derivative Perturbation Delta	0.001						
Iterations	9						
Warning Message	Estimates may not have converged.						

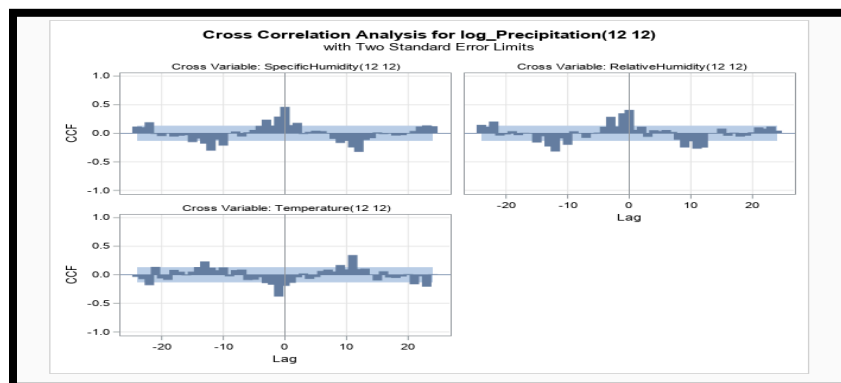
  

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag	Variable	Shift
MU	0.0048640	0.0084821	0.57	0.5683	0	log_Precipitation	0
MA1,1	-1.34119	0.31577	-4.25	<.0001	1	log_Precipitation	0
MA1,2	-0.43998	0.30584	-1.44	0.1503	2	log_Precipitation	0
MA2,1	0.99968	30.56514	0.03	0.9739	12	log_Precipitation	0
AR1,1	-1.44117	0.28952	-4.98	<.0001	1	log_Precipitation	0
AR1,2	-0.57838	0.27617	-2.09	0.0382	2	log_Precipitation	0
AR2,1	-0.27684	0.07286	-3.79	0.0001	12	log_Precipitation	0
NUM1	0.40581	0.11884	3.41	0.0008	0	SpecificHumidity	0
NUM2	-0.0000498	0.02285	-0.00	0.9983	0	RelativeHumidity	0
NUM3	0.01653	0.03839	0.43	0.6687	0	Temperature	0

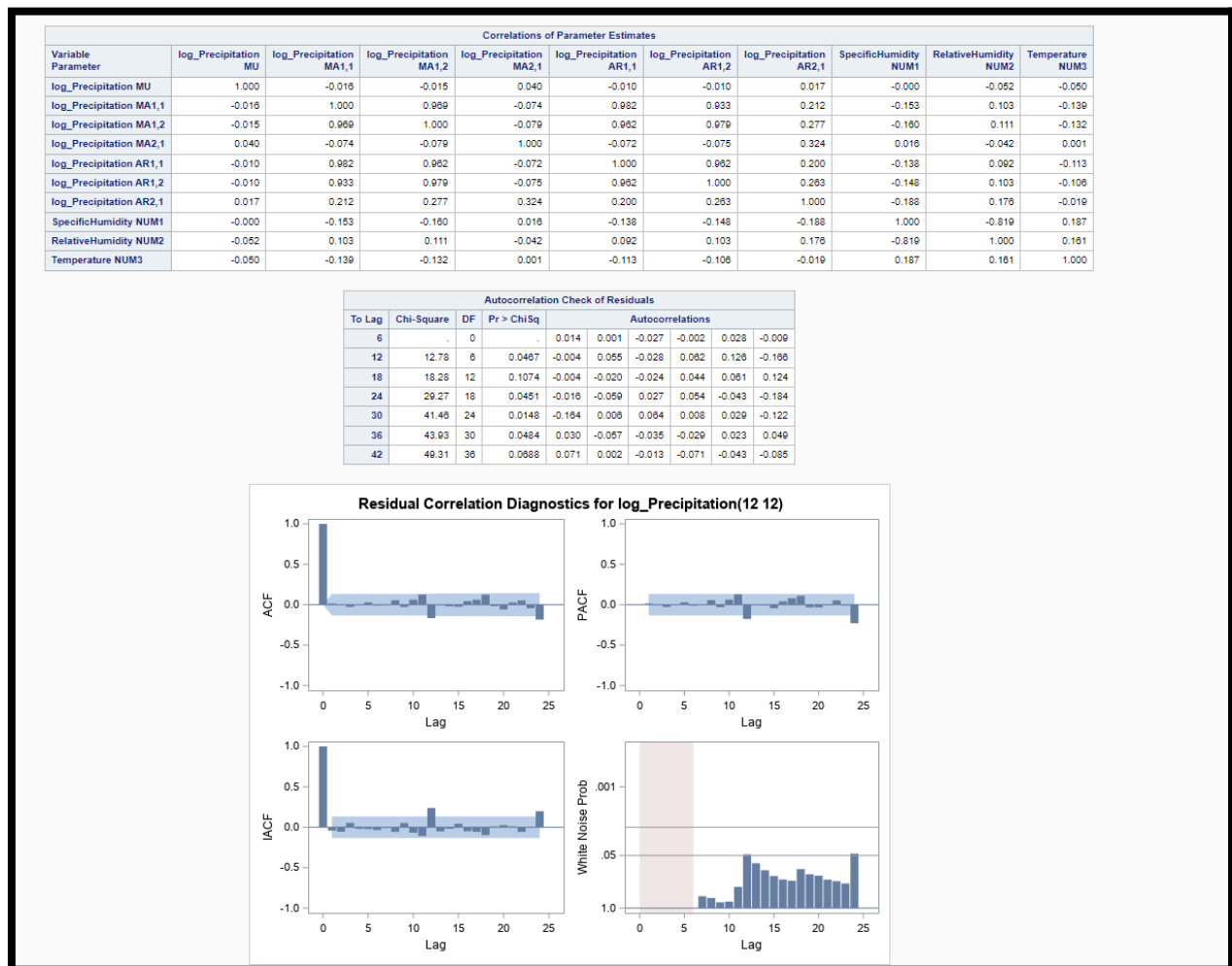
  

Constant Estimate	0.018735
Variance Estimate	1.229224
Std Error Estimate	1.108704
AIC	749.6481
SBC	780.9416
Number of Residuals	228

- Below picture shows the cross correlation analysis for our target variable which is log\_precipitation. We can see all the cross correlations with the independent variables which shows that there are some significant spikes at few intervals.

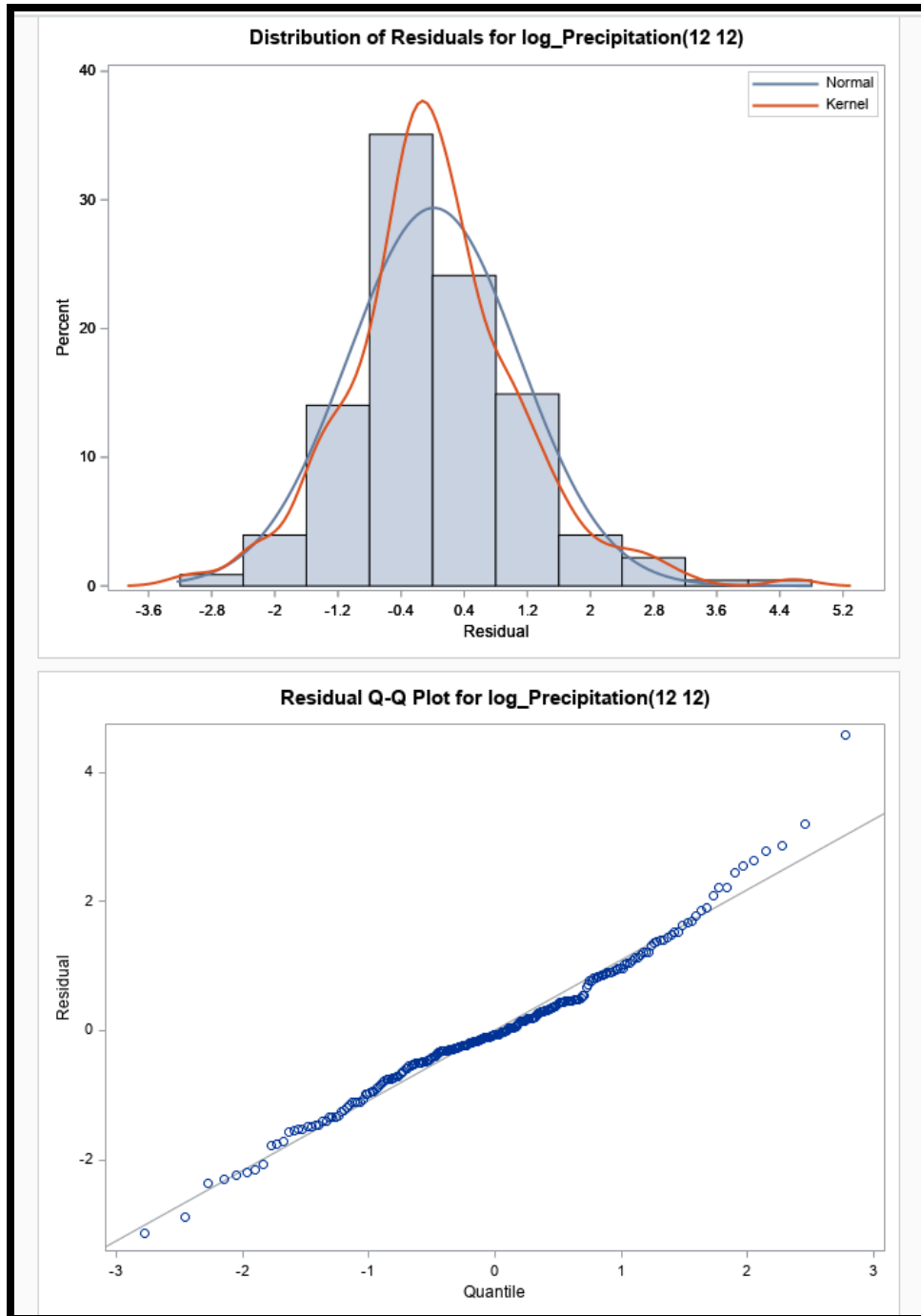


- This snapshot shows the residual correlation for our target variable. Here we could see that for the ACF, IACF and the PACF there are not many spikes and almost every variable in the 95% confidence interval.
- The white noise probability indicates that there is no white noise making this the best ARIMAX model. The lags at spike 12 and 24 are also significant with values of 0.045 and 0.047 which are less than 0.05.



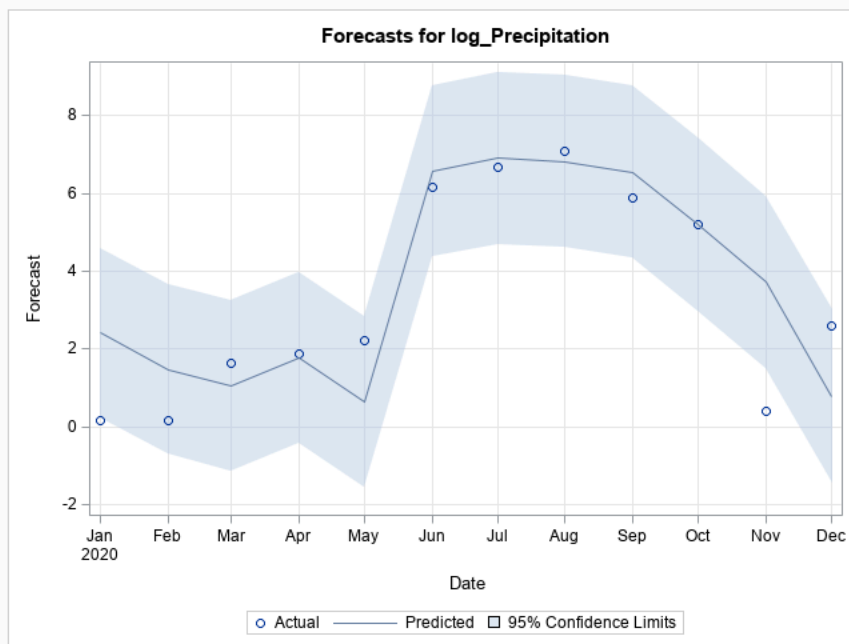
- The below snapshot shows how the distribution of residuals are for the target variable. It indicates that it is a normal distribution which is slightly skewed towards the right side.





- The forecast graph shows that values for the “log\_precipitation” variable are significant lying in the 95% confidence interval.

Forecasts for variable log_Precipitation						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
241	2.4175	1.1087	0.2445	4.5905	0.1570	-2.2805
242	1.4783	1.1142	-0.7056	3.6621	0.1740	-1.3043
243	1.0483	1.1142	-1.1356	3.2321	1.6273	0.5790
244	1.7615	1.1156	-0.4251	3.9480	1.8656	0.1042
245	0.6344	1.1187	-1.5582	2.8270	2.2289	1.5946
246	6.5738	1.1221	4.3745	8.7732	6.1393	-0.4346
247	6.8981	1.1249	4.6935	9.1028	6.6620	-0.2361
248	6.8135	1.1266	4.6054	9.0216	7.0935	0.2800
249	6.5340	1.1275	4.3242	8.7438	5.8925	-0.6415
250	5.1965	1.1278	2.9881	7.4070	5.1995	0.0029
251	3.7186	1.1279	1.5079	5.9292	0.3988	-3.3198
252	0.7821	1.1279	-1.4286	2.9927	2.5825	1.8004



- This screenshot shows the forecasting equation for the best model obtained which is ARIMAX(2,0,2)(1,2,1).

Model for variable log_Precipitation	
Estimated Intercept	0.004864
Period(s) of Differencing	12,12

Autoregressive Factors	
Factor 1:	1 + 1.44117 B <sup>(1)</sup> + 0.57838 B <sup>(2)</sup>
Factor 2:	1 + 0.27564 B <sup>(12)</sup>

Moving Average Factors	
Factor 1:	1 + 1.34119 B <sup>(1)</sup> + 0.43996 B <sup>(2)</sup>
Factor 2:	1 - 0.99968 B <sup>(12)</sup>

- To find the accuracy values for this model, we ran a macro on the forecast values obtained by the reverse logarithmic regression.
- **After Running Macro:** These are the values we see for this model

Series	Model	Holdback Periods	MAE	MSE	RMSE
log_precipitation	work.out1	12	0.71456	1.10274	1.05011

Constant Estimate	-0.10728
Variance Estimate	0.806756
Std Error Estimate	0.898196
AIC	689.3173
SBC	724.6116
Number of Residuals	252

## 6. Model comparison

Comparing the values of our two models, we see the following results:

MODEL	AIC	SBC	RMSE	MAPE
ARIMAX	689.3	724.6	1.05	135.09
Multiplicative Seasonal	-16.4	-9.4	0.84	84.92

Seeing these values we can confirm that “Multiplicative Seasonal Exponential Smoothing Model” is the best model when taking into account of

1. **Accuracy:** The lower the values of AIC, SBC, RMSE, MAPE the better the model is. So we can infer that the Exponential smoothing model is the best.

2. **Parsimonious:** Based on the principle of parsimony and the measurements supplied, the Multiplicative Seasonal model would be the optimal model for predicting in this circumstance. In comparison to the ARIMAX model, it gives a more economical and precise solution.
3. **White noise:** The white noise of the Multiplicative smoothing indicates that it is a better model than the ARIMAX for this model.

## **7. Business Insights and Recommendations:**

The study set out on a goal to create an accurate precipitation forecasting model adapted for Mumbai, acknowledging the city's distinct climatic quirks and the significant implications for a variety of business sectors:

1. **Transportation** - Accurate forecasting allows firms to plan transportation and logistics routes based on anticipated weather conditions. Anticipating severe rain can assist logistics organizations in rerouting cargo, avoiding delays, and ensuring on-time deliveries.

Recommendation: To improve efficiency, logistics companies should incorporate weather forecasting models into their routing software.

2. **Energy and Utilities** - The capacity to accurately estimate rainfall is extremely valuable in the energy sector. Power generation firms, particularly those that use hydropower, can maximize energy production by preparing for future water influxes or predicting drought periods.

Recommendation: Power companies, particularly those that rely on hydropower, should incorporate precipitation projections into their power generation plans.

3. **Construction and Infrastructure** - Accurate humidity and precipitation data is extremely beneficial to construction organizations. Understanding these weather criteria assists for better construction schedule planning. For example, high humidity levels may extend the curing time for concrete, whereas low humidity levels may alter drying periods for materials such as paint and plaster.

Recommendation: Construction businesses should use advanced forecasting techniques to better plan and alter schedules based on weather forecasts.

## **8. Conclusions:**

The study explored deep into Mumbai's distinct environment, utilizing data exploration and modeling approaches to deliver significant insights. Advanced ARIMAX models with prewhitening techniques were used to identify and treat strong seasonal patterns. The strong impact of climatic variables such as specific humidity, relative humidity, and temperature on precipitation highlighted the relevance of these variables.

This study demonstrates the effectiveness of data-driven analytical approaches in addressing complicated forecasting problems. It emphasizes the vast economic consequences across varied sectors, emphasizing the advantage organizations acquire when armed with precise forecasting tools in decision-making, operations optimization, and boosting sustainability and resilience.

## **9. References**

The references here are the dataset link which is obtained from kaggle. The dataset link is as follows

- Dataset link:  
<https://www.kaggle.com/datasets/poojag718/rainfall-timeseries-data>
- SAS help center:  
<https://support.sas.com/en/knowledge-base.html>
- Time Series Modeling Essentials Course Notes