# Lake Ontario Monthly Mean Water Level Analysis and Forecasting

Team 3 | 5/20

Nihar R. Panda, Rahul Menon,
Achintya Mishra, Kyu Sung Cho

# Our team

**Nihar**

**Rahul**

**Kyu**

**Achintya**

# Table of contents

# Introduction

**Problem Statement**: Variations in Lake Ontario's water level significantly impact lake management, environmental planning, and community safety by governments and corporations.

**Objective**: Use historical water level data along with *neighboring lakes water level data*, *mean temperature*, *mean net precipitation*, and *mean net evaporation* data to forecast future water levels.

## Lake Size and Volume

LO is smaller in surface area and volume than lake Michigan, Superior, and others, contributing to lower water levels.

## Inflow Outflow Dynamics

LO receives water from Lake Erie and discharges into the St. Lawrence River, with regulated outflow impacting its levels.


## Elevation Difference

LO is situated at a lower elevation compared to the other Great Lakes, with water flowing from the higher elevation lakes down to LO

## Moses-Saunders Dam

The dam is primarily used to regulate water flow and levels to balance various needs such as hydroelectric power generation, navigation, and flood control

# Data Set Description

| Features | Description | Source | Missing Value |
|---|---|---|---|
| **Time Period** | Jan 1980 - Dec 2017 | Link | No |
| **Water Level** | Monthly Mean Water Level of the Great Lakes (in m) | Link | No |
| **Temperature** | Mean Temp of Toronto (°C) | Link | No |
| **Evaporation** | Monthly Net Evaporation (in mm) | Link | No |
| **Precipitation** | Monthly Net Precipitation (in mm) | Link | No |

# Exploratory Data Analysis

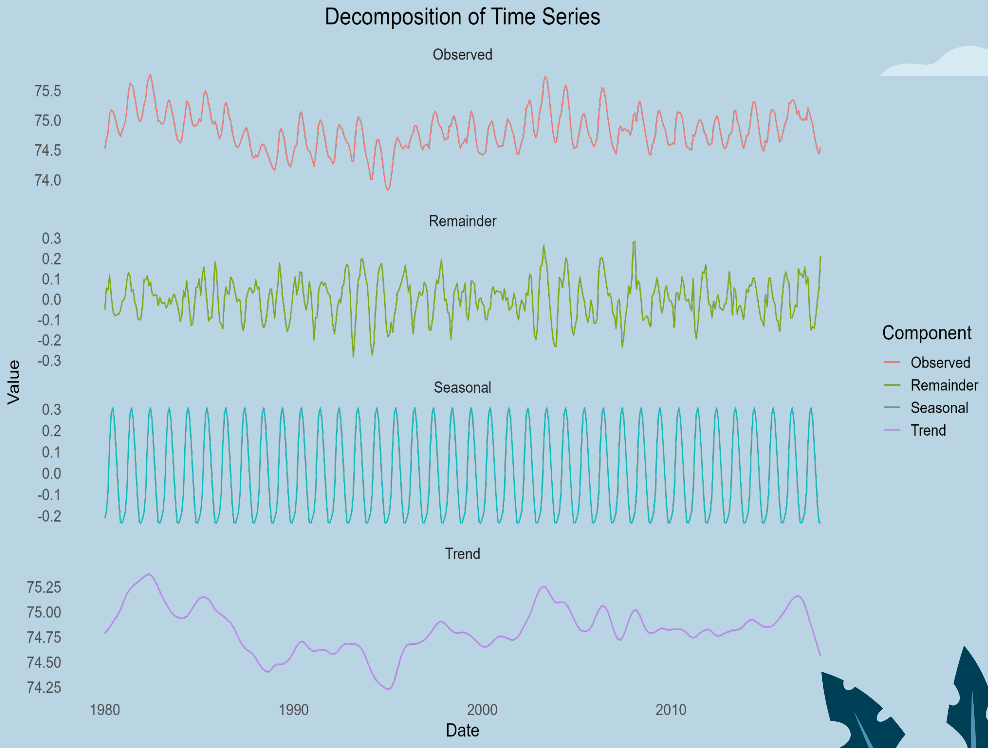
# Water level characteristics

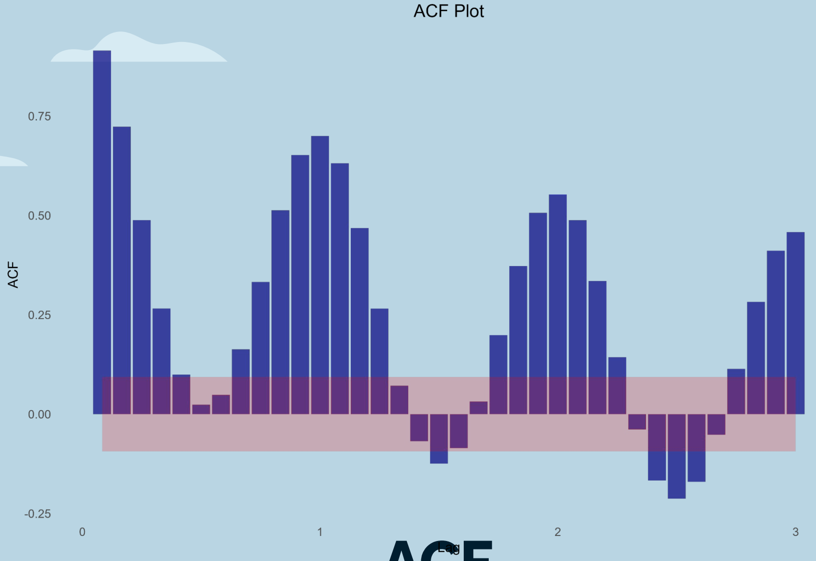
## Trend

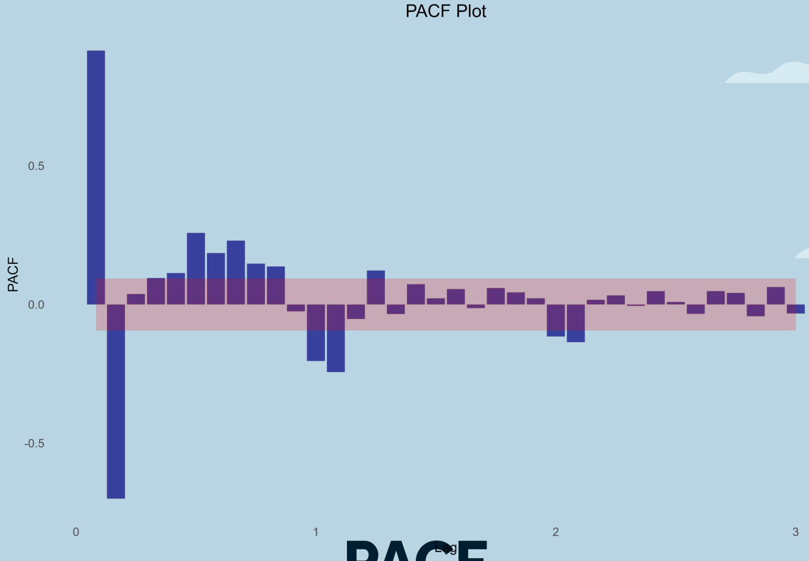
The overall trend appears relatively stable, with no significant long-term increase or decrease in water levels over the entire period.

## Seasonality

The water levels show clear seasonal fluctuations, with regular peaks and troughs indicating cyclical changes within each year.





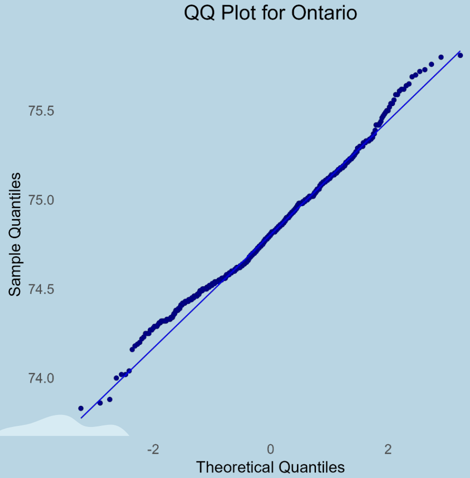

## ACF

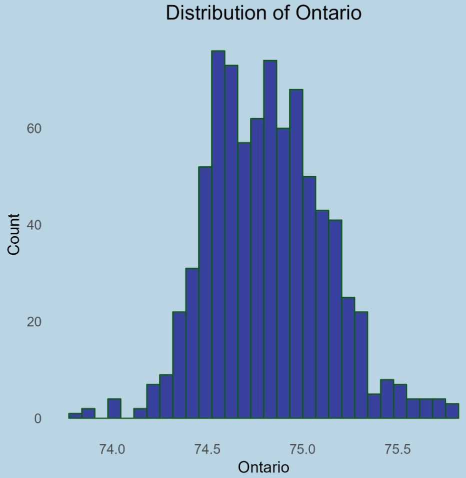
shows strong autocorrelation at the first few lags, with bars extending beyond the red confidence intervals, suggesting non-randomness. The periodic peaks observed suggest seasonality in the data, indicating a repeating structure at certain intervals.

## PACF

shows significant partial autocorrelation at lag 1, highlighting a strong immediate past influence, while subsequent lags are less significant, falling mostly within the confidence intervals.

# BoxCox & Shapiro Test



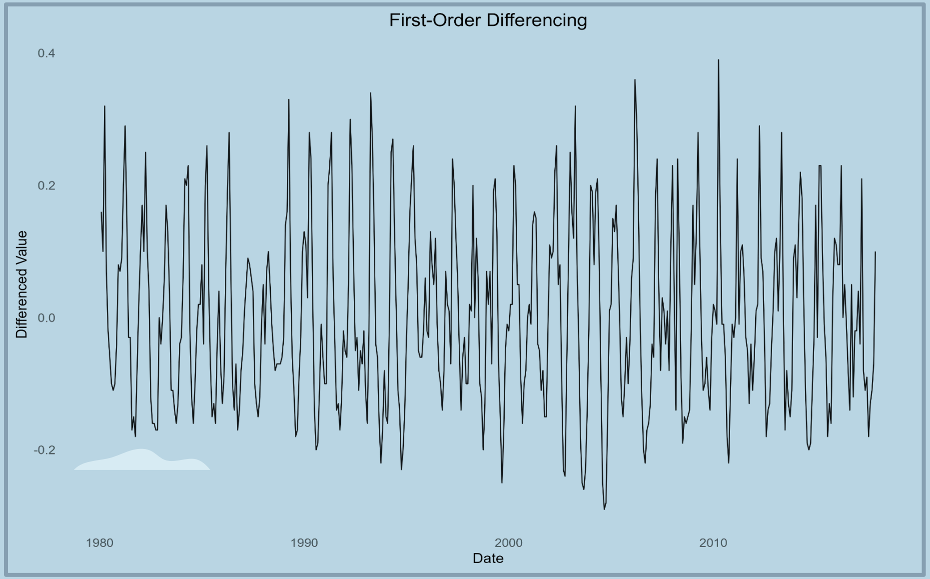



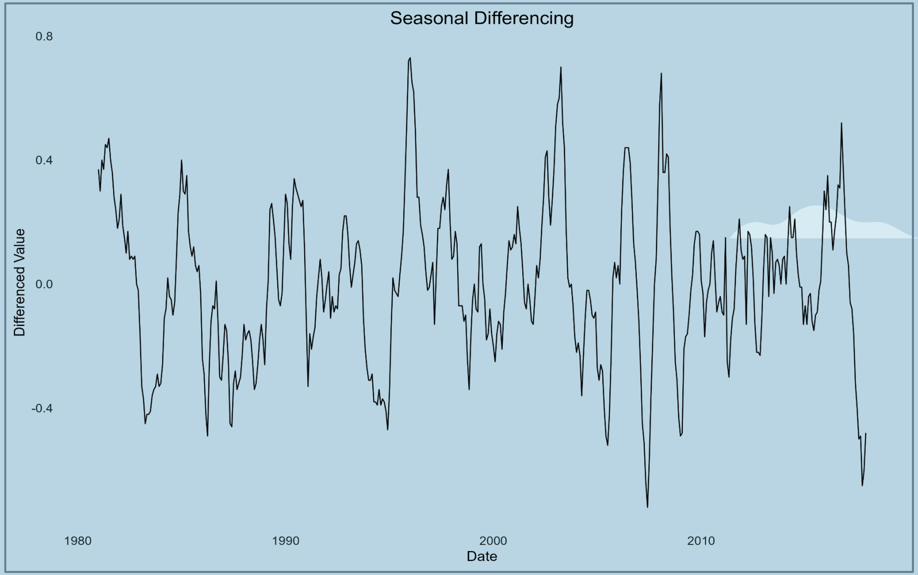

| | |
|---|---|
| **Lambda (BoxCox)** | -1.00 |
| **Shapiro (Normality)** | 0.047 |

The lambda value for the Box-Cox transformation is -1.00. indicating that a reciprocal transformation has been applied to stabilize the variance and make the data more normally distributed.

The Shapiro-Wilk test p-value is 0.047. Since this p-value is less than 0.05, it indicates that the data deviates slightly from a normal distribution.

# Check for Stationarity



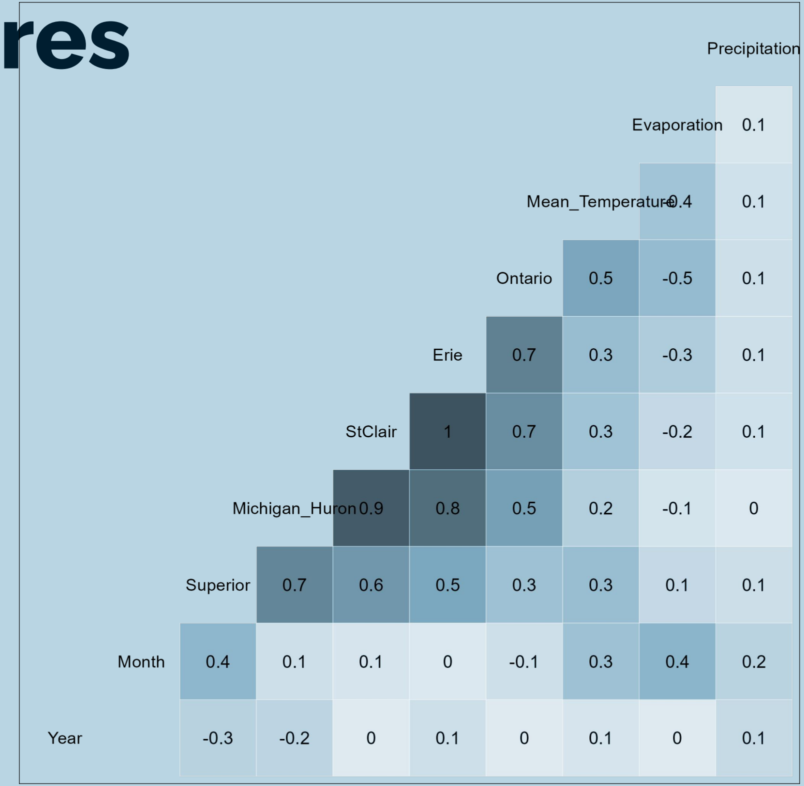

The **KPSS** test for Ontario yielded a p-value of **0.03**, rejecting the null hypothesis and indicating that the data is not stationary. The KPSS test for first-order differencing for trend and seasonality provided a p-value of **0.1**, suggesting that both trend and seasonal differencing are necessary to achieve stationarity.

# Correlation among various features

Negative correlation between mean temperature and evaporation (-0.39), suggesting other factors influence evaporation rates beyond temperature alone.

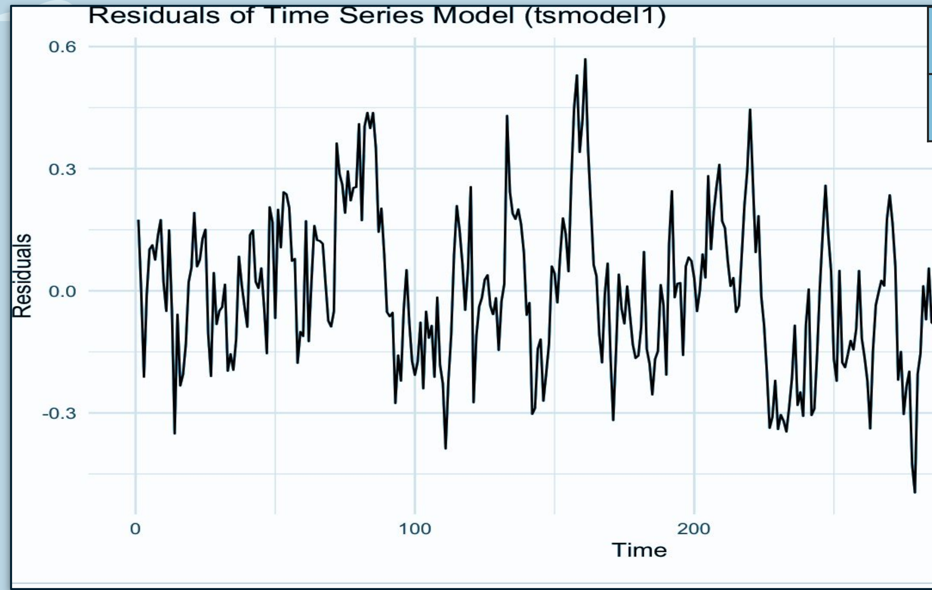
Moderate positive correlation between Lake Ontario water levels and mean temperature (0.5), suggesting higher water levels with higher temperatures.

Moderate negative correlation between Lake Ontario water levels and evaporation (-0.48), indicating higher evaporation rates reduce water levels.



# Models

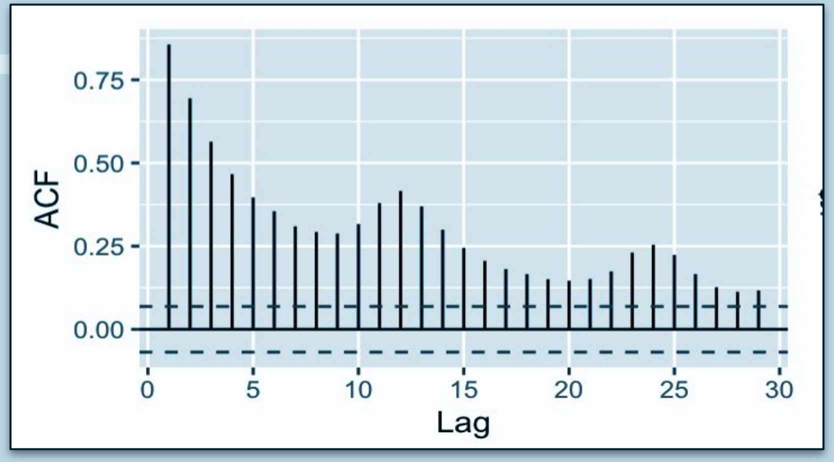
# Time Series Linear Regression

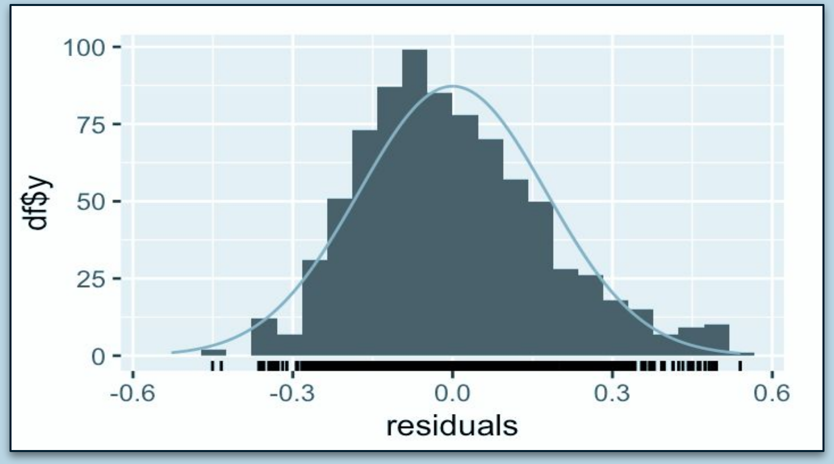



| | |
|---|---|
| R2 | 0.5564 |
| Adj R2 | 0.5536 |

| | |
|---|---|
| Intercept | 0.485669 |
| Superior | 4.26e-05 |
| Michigan Huron | 2.00e-16 |
| St. Claire | 0.661455 |
| Eerie | 2.00e-16 |
| Mean Temperature | 5.56e-07 |
| Evaporation | 6.26e-11 |
| Precipitation | 0.573000 |

**tslm(train_water_level ~ train_superior + train_eerie + train_stclair + train_mich_huron + log(train_temp) + train_evap + log(train_precip))**



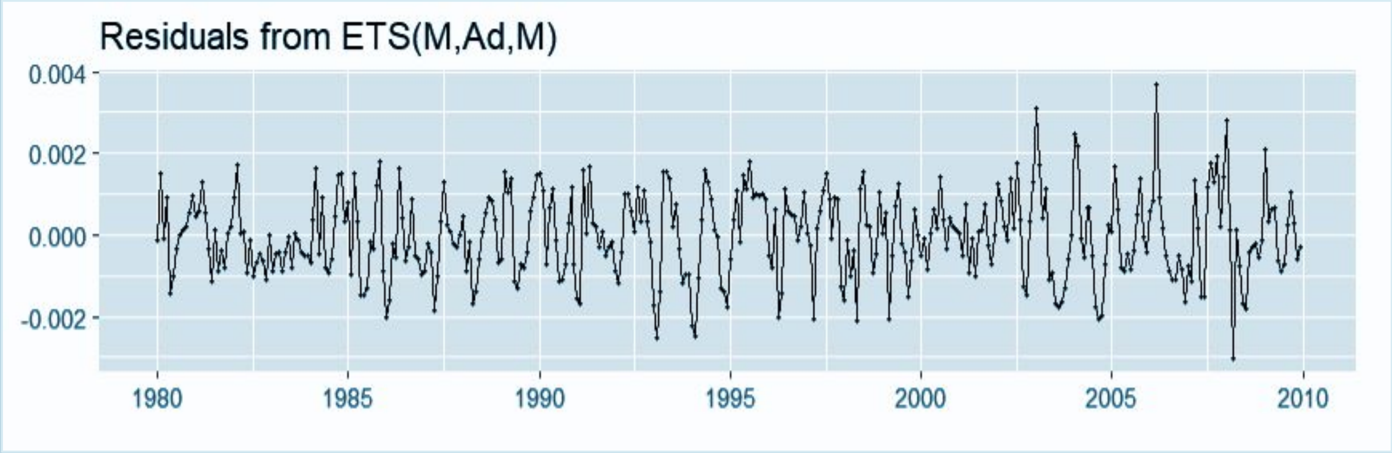




Significant autocorrelation in the residuals suggests that the model may not be adequately capturing all the information in the data.

**Ljung-Box Test**
**p-value = 2.2e-16**

P-value is less than 0.05, hence, we are rejecting the null hypothesis which states that the residuals have autocorrelation.

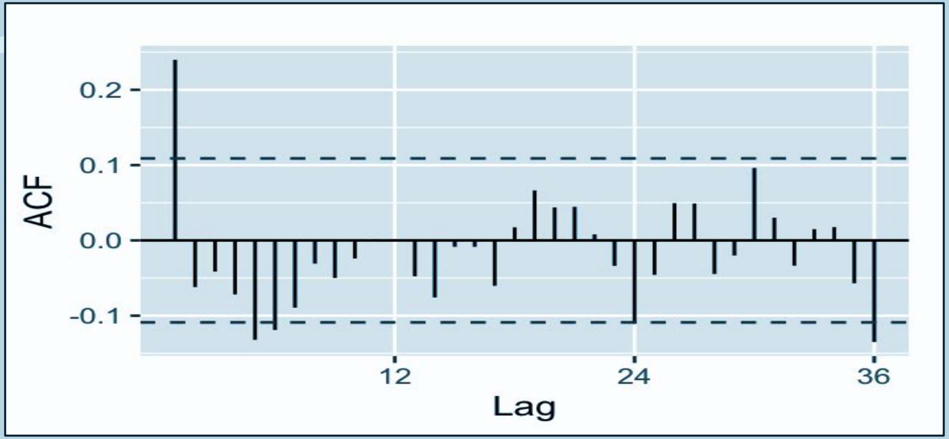
# Exponential Smoothing




| Model | M,Ad,M |
|---|---|
| AICc | 314.16 |
| BIC | 382.11 |

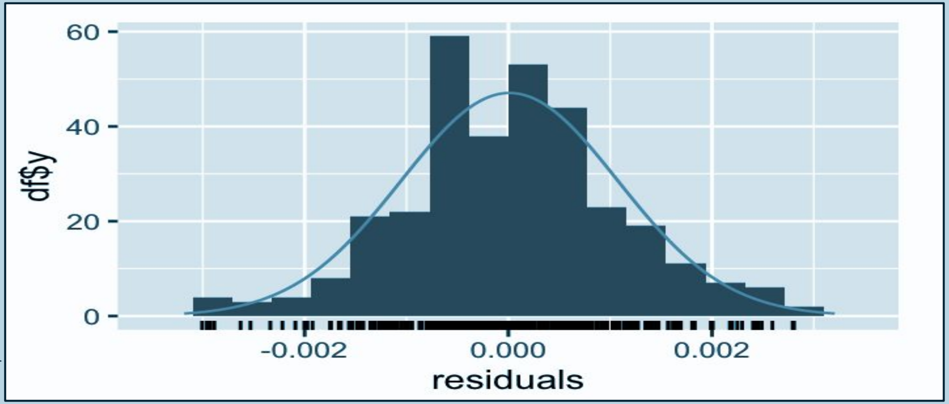
**BEST Model: ETS(M, Ad,M)**

Residual plot for our ETS model showcases that residuals are oscillating around mean indicating that if we use this model for forecasting there won't be any bias. We can also observe that AICc is quite high meaning our model is performing quite well.

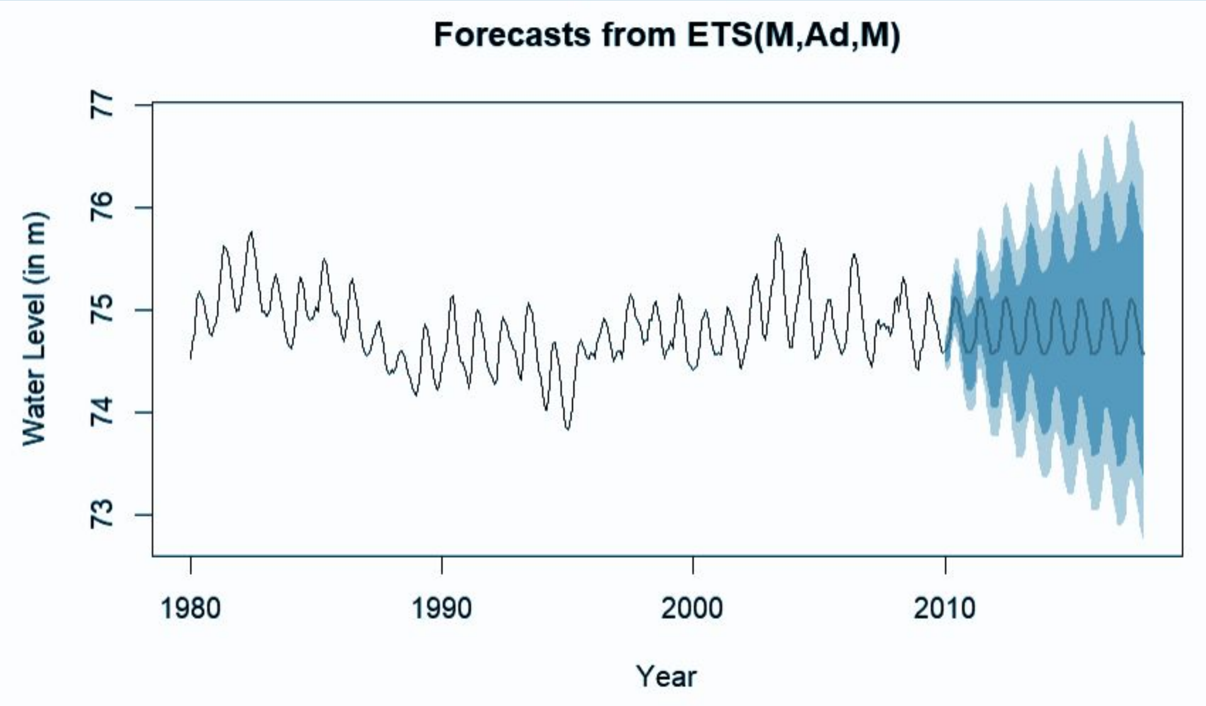



ETS model cannot capture the all underlying patterns in the data as residuals are not behaving as white noise.

**Ljung-Box Test**
**p-value = 0.00013**




The residuals are approximately normally distributed, as expected under the assumption of a well-fitting model.




| Metric | Train | Test |
|---|---|---|
| **RMSE** | 0.077 | 0.177 |
| **MAE** | 0.063 | 0.129 |
| **MAPE** | 0.083 | 0.173 |
| **MASE** | 0.293 | 0.611 |

The above plot clearly shows forecast value for the next 8 years. Looking at the metrics the RMSE and MAE value is very low indicating a good prediction on the forecast.

# SARIMA

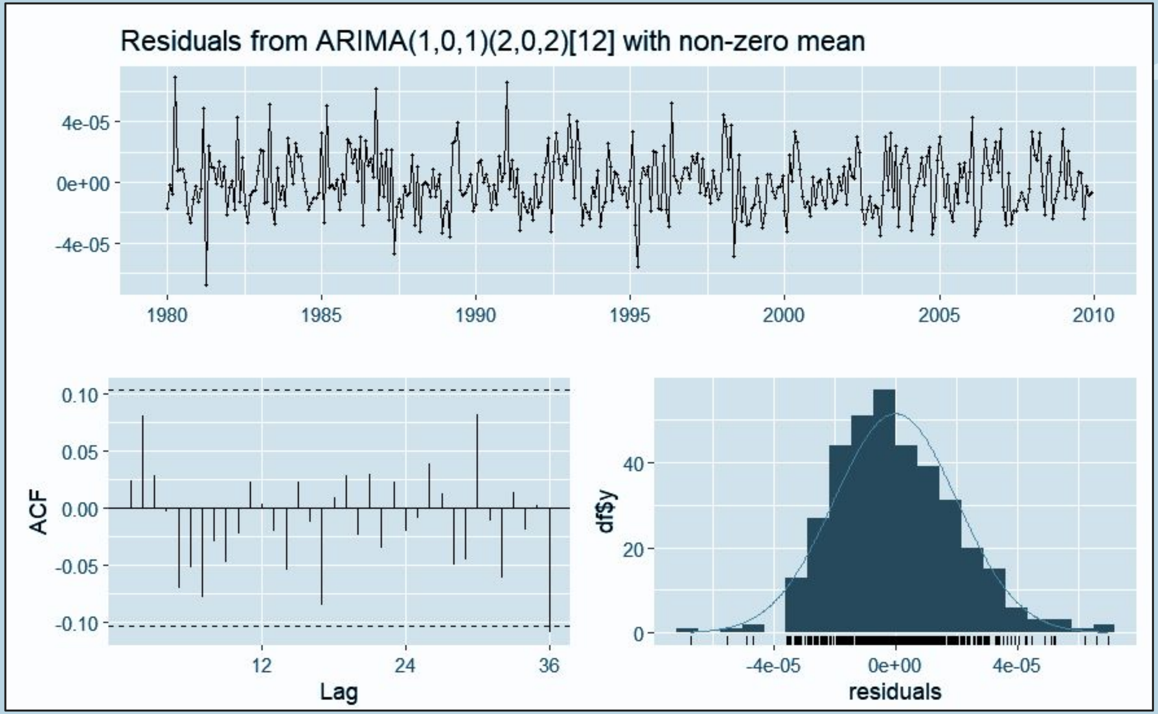
Auto.arima gave ARIMA(2,0,1)(2,0,2) as the best model. The model parameters are changed around the best model to check whether the AICc is constant by decreasing the model complexity.

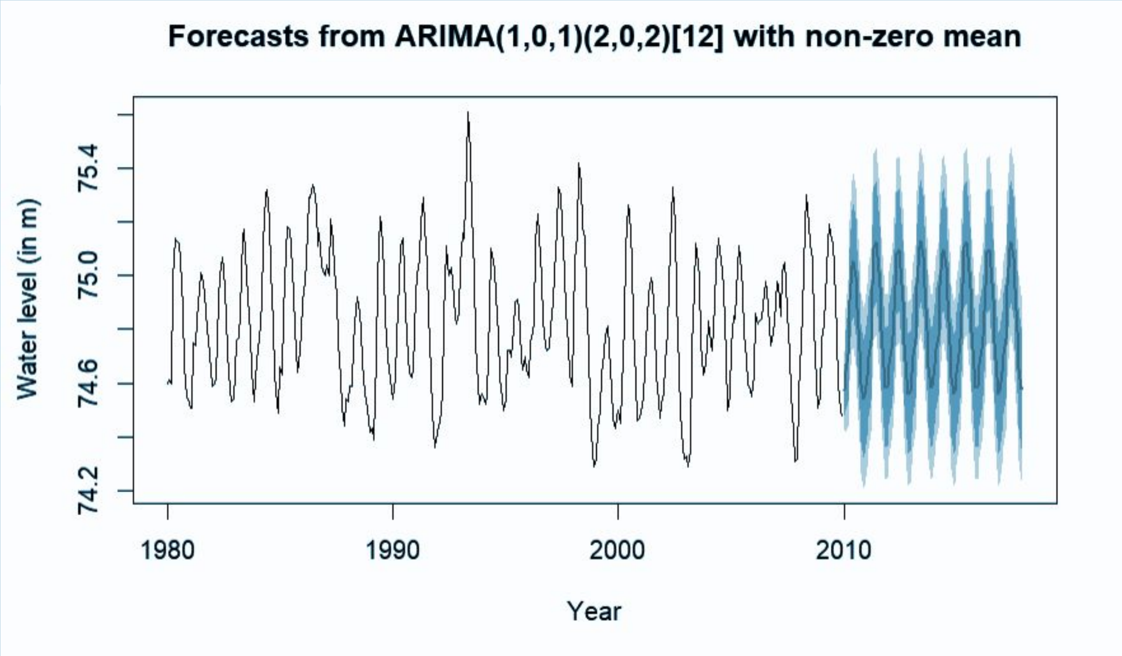
The AICc for ARIMA(2,0,1)(2,0,2) is better as compared to ARIMA(1,0,1)(2,0,2) however, BIC is lower for the later model which shows that the less complex model is better. **Hence, the best model is ARIMA(1,0,1)(2,0,2)[12].**

| Model (ARIMA) | AICc | BIC |
| --- | --- | --- |
| **(1,0,1)(2,0,2)[12]** | -6706.117 | -6675.439 |
| **(2,0,1)(2,0,2)[12]** | -6707.434 | -6672.973 |

**Ljung-Box Test**
**p-value = 0.6193**

The residuals are unbiased and the distribution of the residuals looks like normal distribution. The ACF plot shows that there isn't any autocorrelation in the residuals.






| Metric | Train | Test |
|---|---|---|
| RMSE | 0.073 | 0.171 |
| MAE | 0.057 | 0.136 |
| MAPE | 0.073 | 0.182 |
| MASE | 0.258 | 0.642 |

The forecasts from ARIMA plot shows forecast value for the next 8 years. Looking at the metrics the RMSE and MAE value is low indicating a good prediction on the forecast.

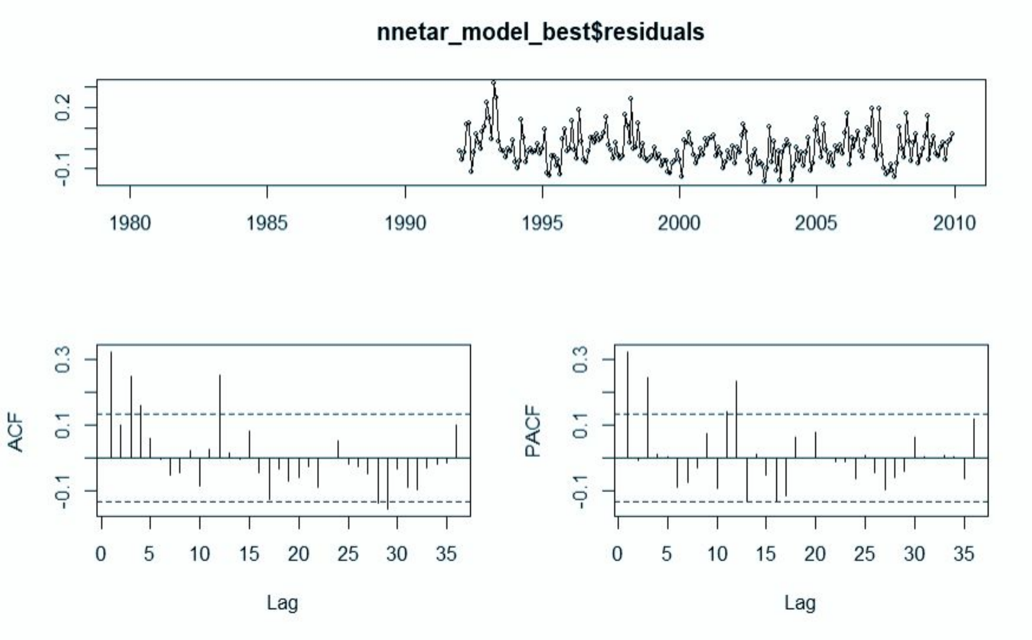
# NNAR

Best parameters for NNAR are chosen by grid search optimization. The best parameters are **size = 11, repeats = 20, P=12 and decay =1.**

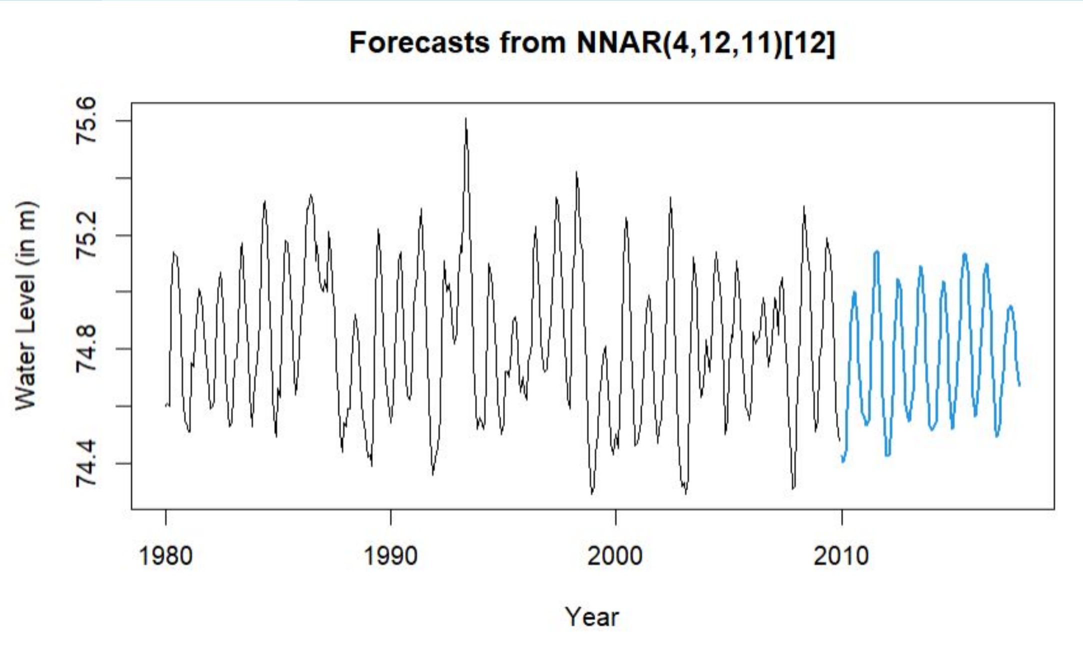
**BEST MODEL: NNAR(4,12,11)[12]**

The residual plot shows that the residuals are oscillating around 0. However, the ACF and PACF plot shows the presence of autocorrelation in the residuals.

**Ljung-Box Test p-value is close to 0** which reject the null hypothesis stating that there exist autocorrelation.



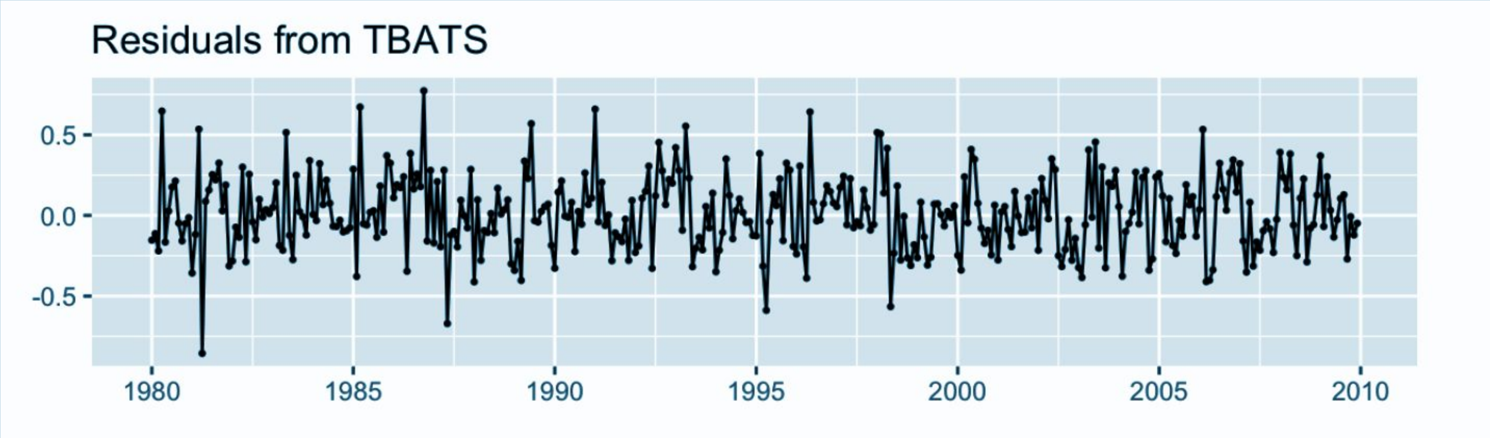

```
nnetar(train_water_level, P=12, size=11, repeats=20, lambda="auto", decay=1)
```




| Metric | Train | Test |
|---|---|---|
| RMSE | 0.082 | 0.243 |
| MAE | 0.063 | 0.170 |
| MAPE | 0.077 | 0.202 |
| MASE | 0.302 | 0.791 |

The forecasts from NNAR model shows forecast value for the next 8 years. Looking at the metrics the RMSE and MAE value is low indicating a good prediction on the forecast.

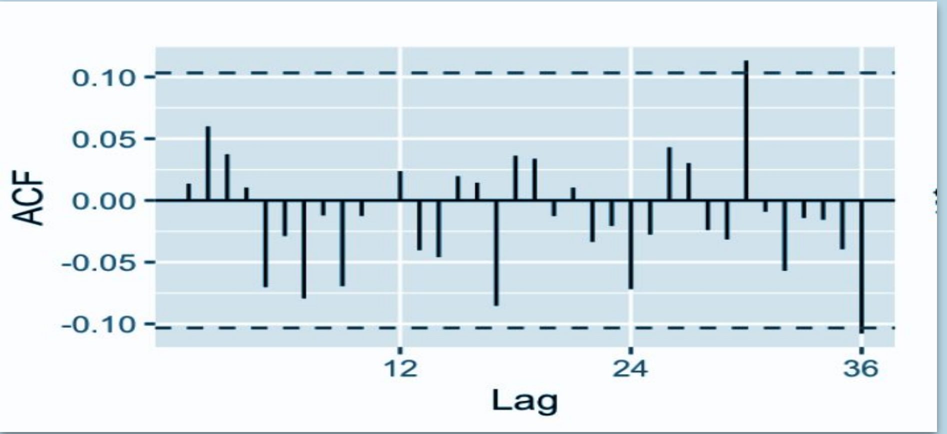
# T-BATS

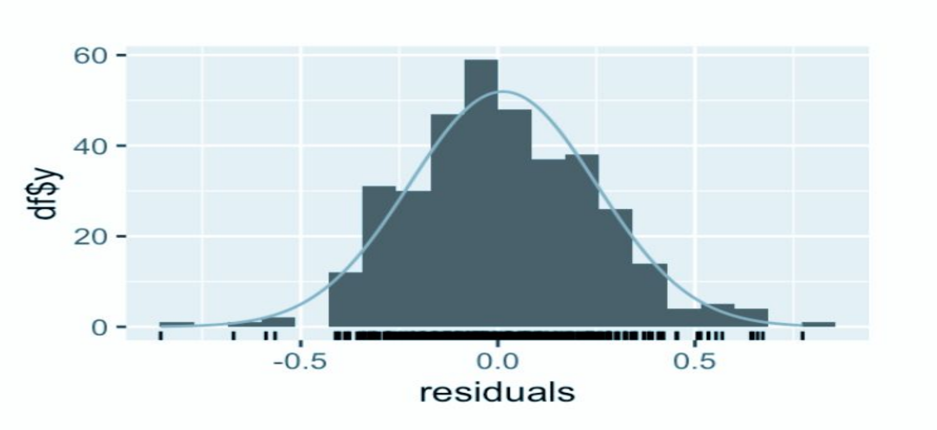



The residuals plot from the TBATS model shows random distribution around zero, indicating the model effectively captures the underlying structure without significant autocorrelation.

**Best Model: TBATS(0.004, {3,3}, 0.952, {<12,5>})**

```
tbats(train_water_level, use.box.cox = TRUE, seasonal.periods = 12)
```

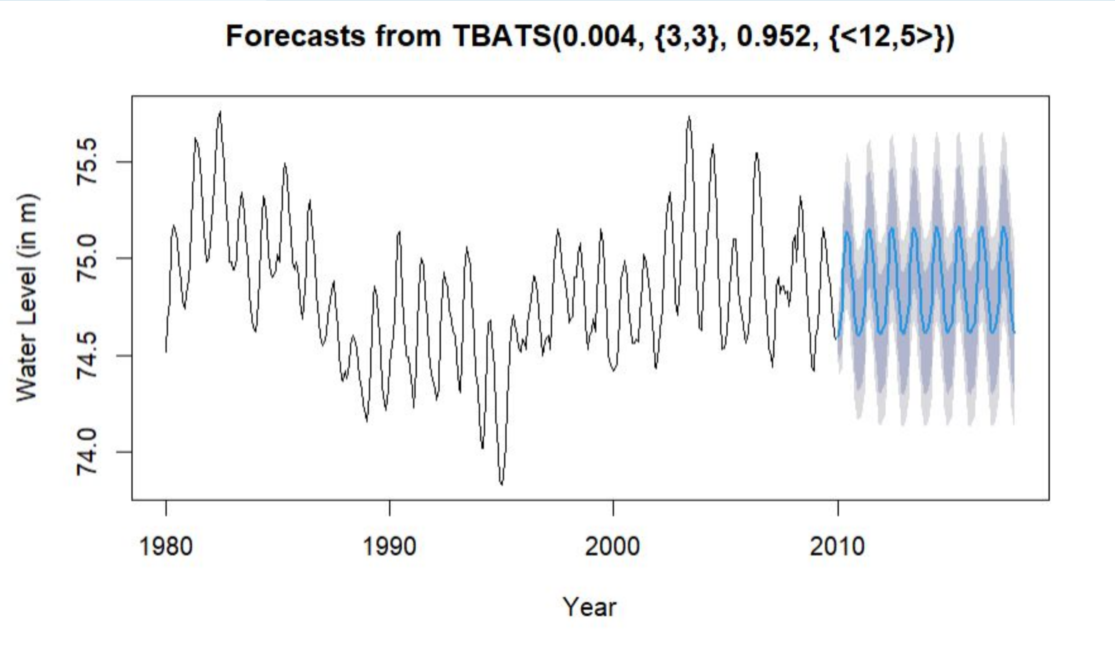



Most lags fall within the 95% confidence interval, indicating that the residuals are **normally distributed** and have **little to no autocorrelation**

**Ljung-Box Test**
**p-value = 0.7156**

The **p-value** of **0.7156** suggests we cannot reject the null hypothesis, meaning the residuals have **no significant autocorrelation**

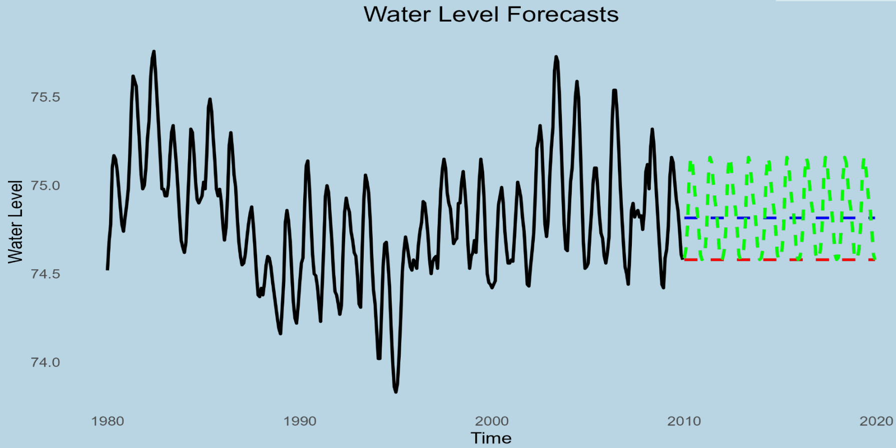



| Metric | Train | Test |
|---|---|---|
| RMSE | 0.066 | 0.161 |
| MAE | 0.052 | 0.123 |
| MAPE | 0.069 | 0.164 |
| MASE | 0.245 | 0.581 |

The forecasts from T-BATS plot shows forecast value for the next 8 years. Looking at the metrics the RMSE and MAE value is very low (close to 0) indicating a good prediction on the forecast.

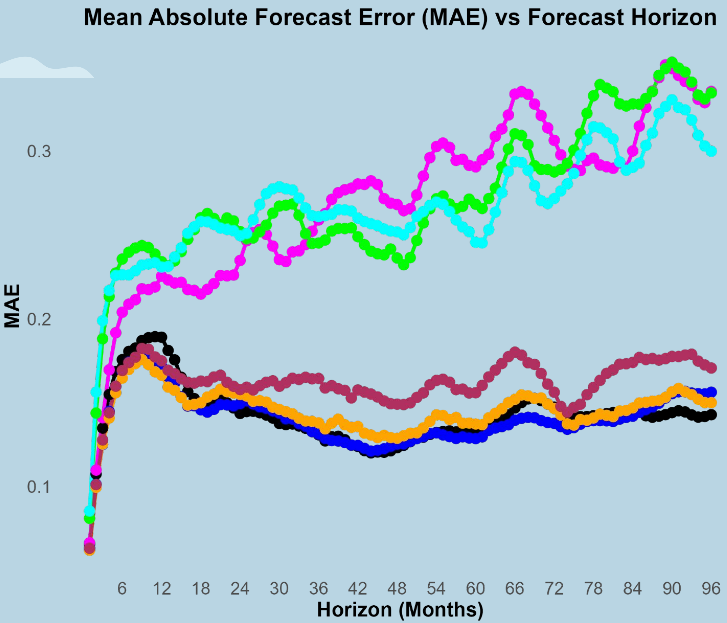
# Benchmark Forecast

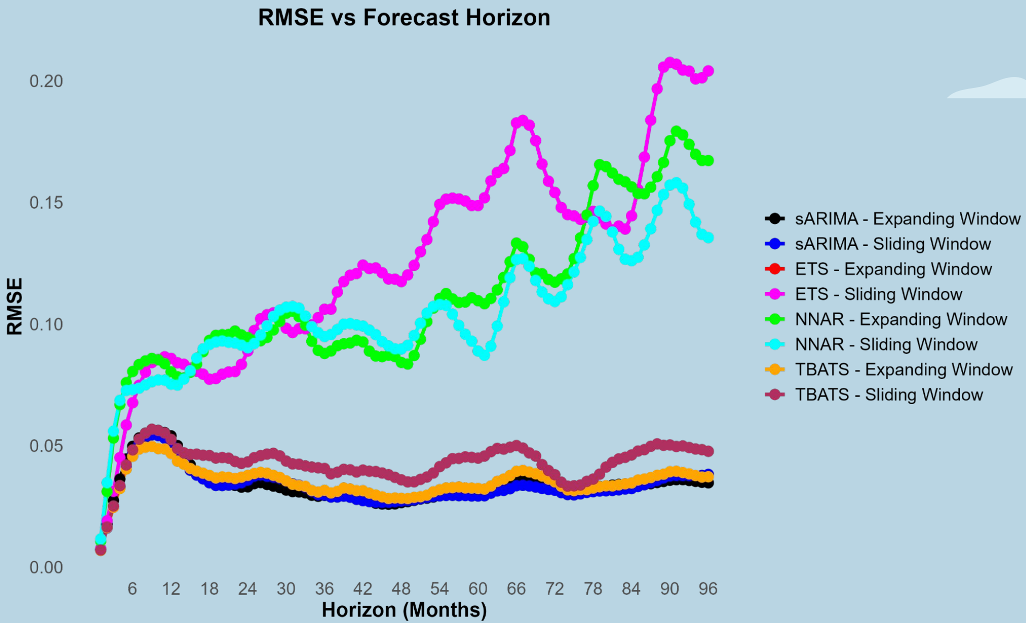
| | Mean | | Naïve | | Seasonal Naïve | |
|---|---|---|---|---|---|---|
| **Metric** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| **RMSE** | 0.348 | 0.248 | 0.137 | 0.378 | 0.265 | 0.167 |
| **MAE** | 0.277 | 0.214 | 0.113 | 0.306 | 0.212 | 0.127 |
| **MAPE** | 0.370 | 0.286 | 0.151 | 0.407 | 0.284 | 0.169 |




Seasonal Naive, Naive and Mean forecasts are used as benchmark forecast and the forecast plot has been shown above. The seasonal naive forecast accuracy measures are better compared to mean and naive forecasts because of presence of annual seasonality in the data.

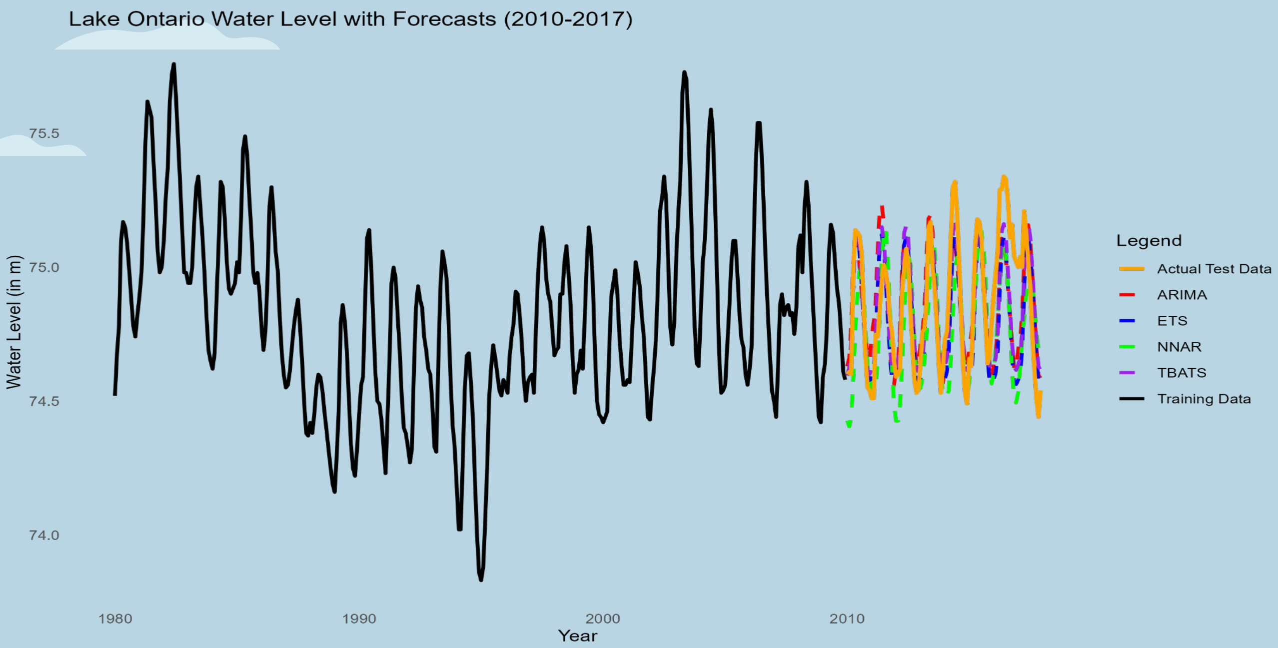
# Time Series Cross Validation




Both the plots show that errors for sARIMA and TBATS are consistent over different horizons compared to the other models for both the expanding and sliding window.

# Conclusion




| Test | ARIMA | TBATS |
|---|---|---|
| RMSE | 0.171 | 0.161 |
| MAE | 0.136 | 0.123 |
| MAPE | 0.182 | 0.164 |

From the models tested, **ARIMA** and **TBATS** are providing the best forecast when compared to other models with respect to actual test data. Even though the **test metric of TBATS** yield the best model, the **ARIMA** model would more robust when taking **model complexity** into consideration.

# Future Scope

1

## Water Consumption

Incorporate  water consumption and human intervention related data from lake ontario.

2

## Advanced Models

Advanced modelling techniques like LSTM and ensemble methods can be explored.

3

## Inflow/Outflow

Monitor inflow rates into Lake Ontario to gain insights into changes in water levels.

4

## Wind Impact

Consider the impact of wind, which can cause temporary changes in water levels through wave action.

# Thank You

# Individual Contributions

| Member | Contribution |
|---|---|
| Nihar R. Panda | sARIMA, Time Series Cross Validation, VAR, Regression with ARIMA errors |
| Rahul Menon | Business Value, NNAR,EDA, Conclusion |
| Kyu Sung Cho | EDA, TBATS, Benchmark Models |
| Achintya Mishra | Time Series Linear Regression, ETS, Future Scope |

# Appendix

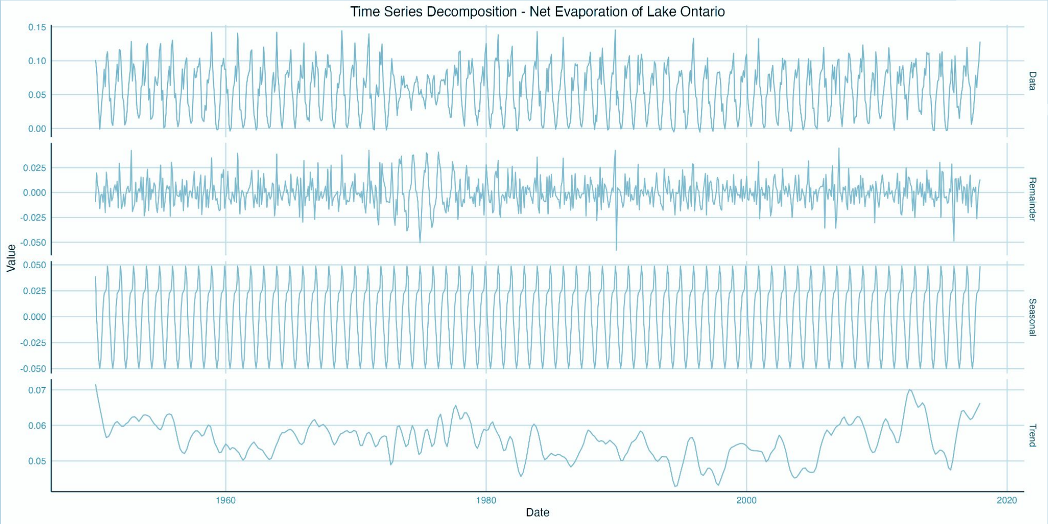
# Appendix : Evaporation characteristics

## Seasonality

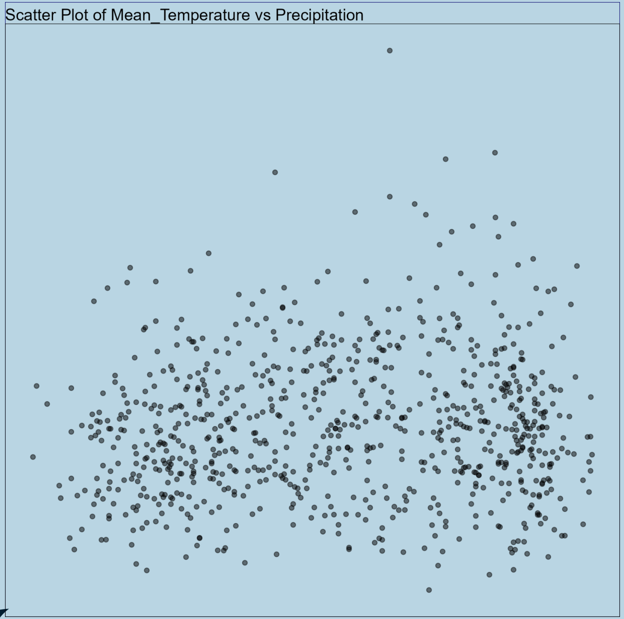
The Toronto's evaporation plots show a significant seasonal pattern in the evaporation of Lake Ontario, with regular cycles corresponding to the annual climatic changes

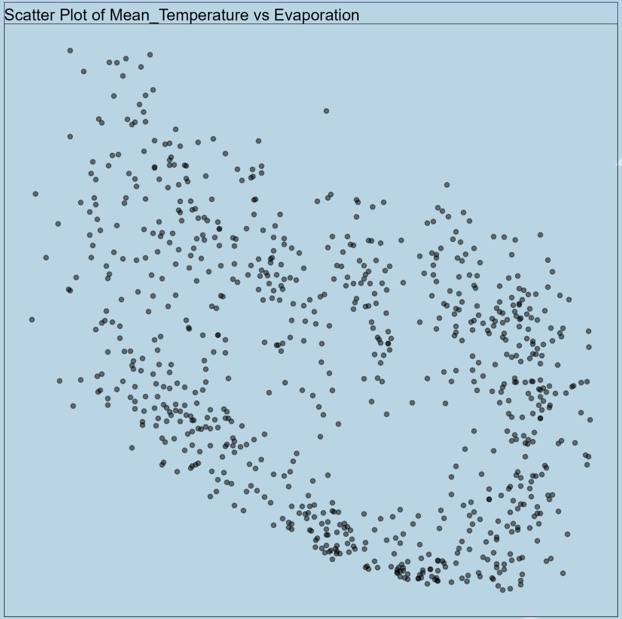
## Trend

The trend in evaporation is relatively stable over the long term, with minor fluctuations.

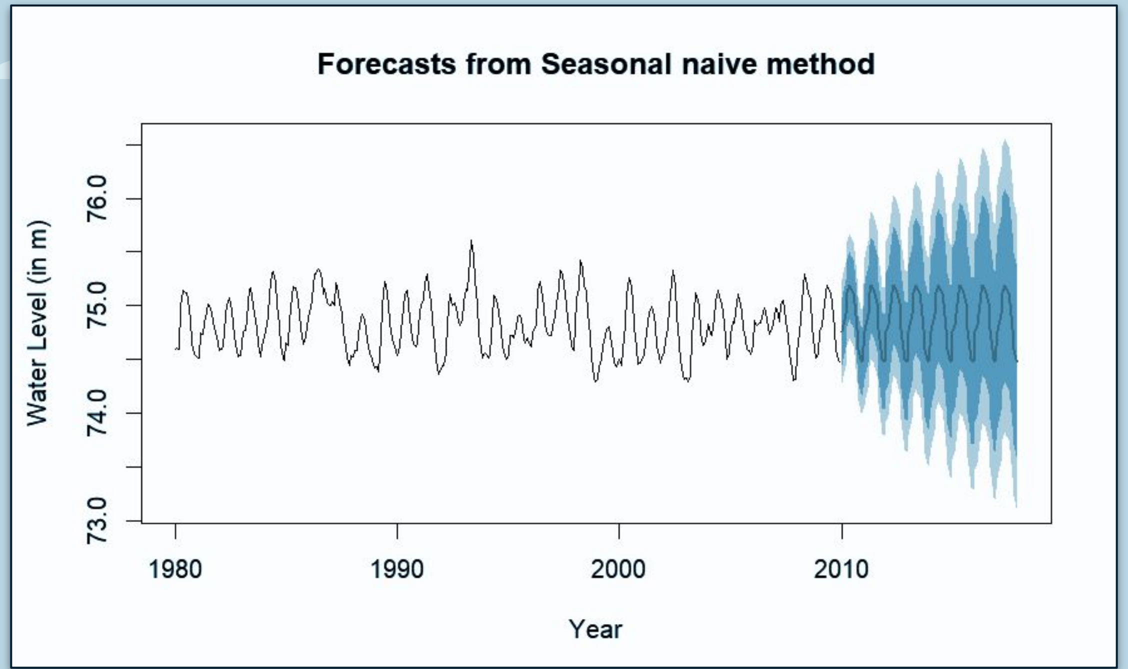


# Appendix: Correlation Check



# Appendix: Seasonal Naive Forecast




| Metric | Train | Test |
| --- | --- | --- |
| RMSE | 0.247 | 0.243 |
| MAE | 0.191 | 0.185 |
| MAPE | 0.256 | 0.247 |
| MASE | 1.000 | 0.965 |

The forecasts from seasonal naive shows forecast value for the next 8 years. Looking at the metrics the RMSE and MAE value is low indicating a good prediction on the forecast.

# Appendix: Regression with ARIMA

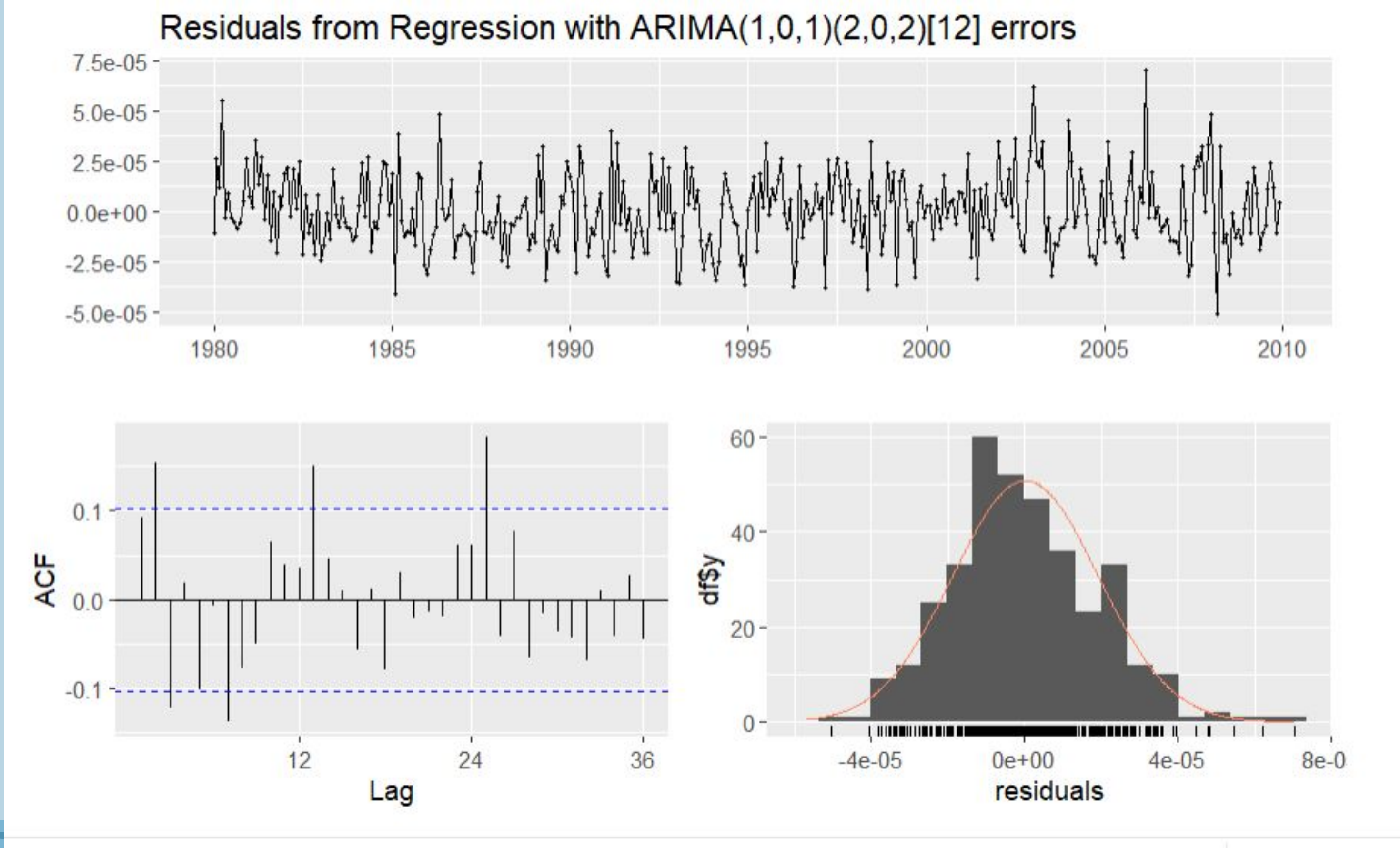
```
Series: train_water_level
Regression with ARIMA(1,0,1)(2,0,2)[12] errors
Box Cox transformation: lambda= -0.8999268

Coefficients:
         ar1     ma1     sar1    sar2    sma1     sma2  intercept  Superior
Michigan.Huron  St..Clair  Erie
      0.9280  0.4361  0.1711  0.8205  0.0692  -0.8678     1.0751     1e-04
0          0      0
s.e.  0.0107  0.0354  0.0440  0.0339  0.0598      NaN        NaN       NaN
NaN          0    NaN
      Mean.Temperature  Evaporation  Precipitation
                     0            0              0
s.e.               NaN          NaN            NaN

sigma^2 = 3.734e-10:  log likelihood = 3387.49
AIC=-6744.98   AICc=-6743.59   BIC=-6686.69

Training set error measures:
                      ME       RMSE        MAE         MPE       MAPE      MASE
ACF1
Training set 0.00865315 0.06887902 0.05476553 0.001067983 0.07319132 0.2578945
0.09087591
```

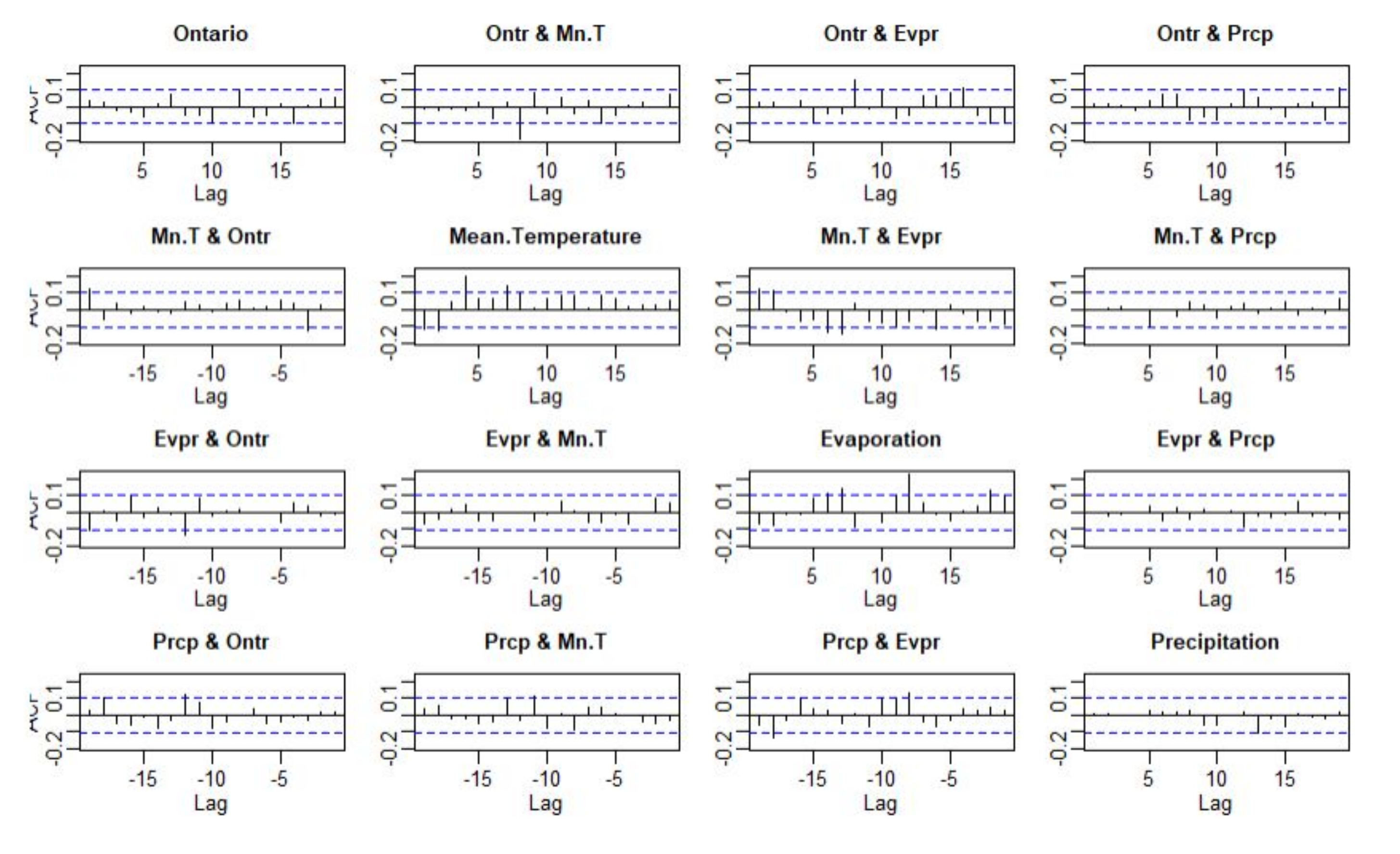
# Appendix: Regression with ARIMA

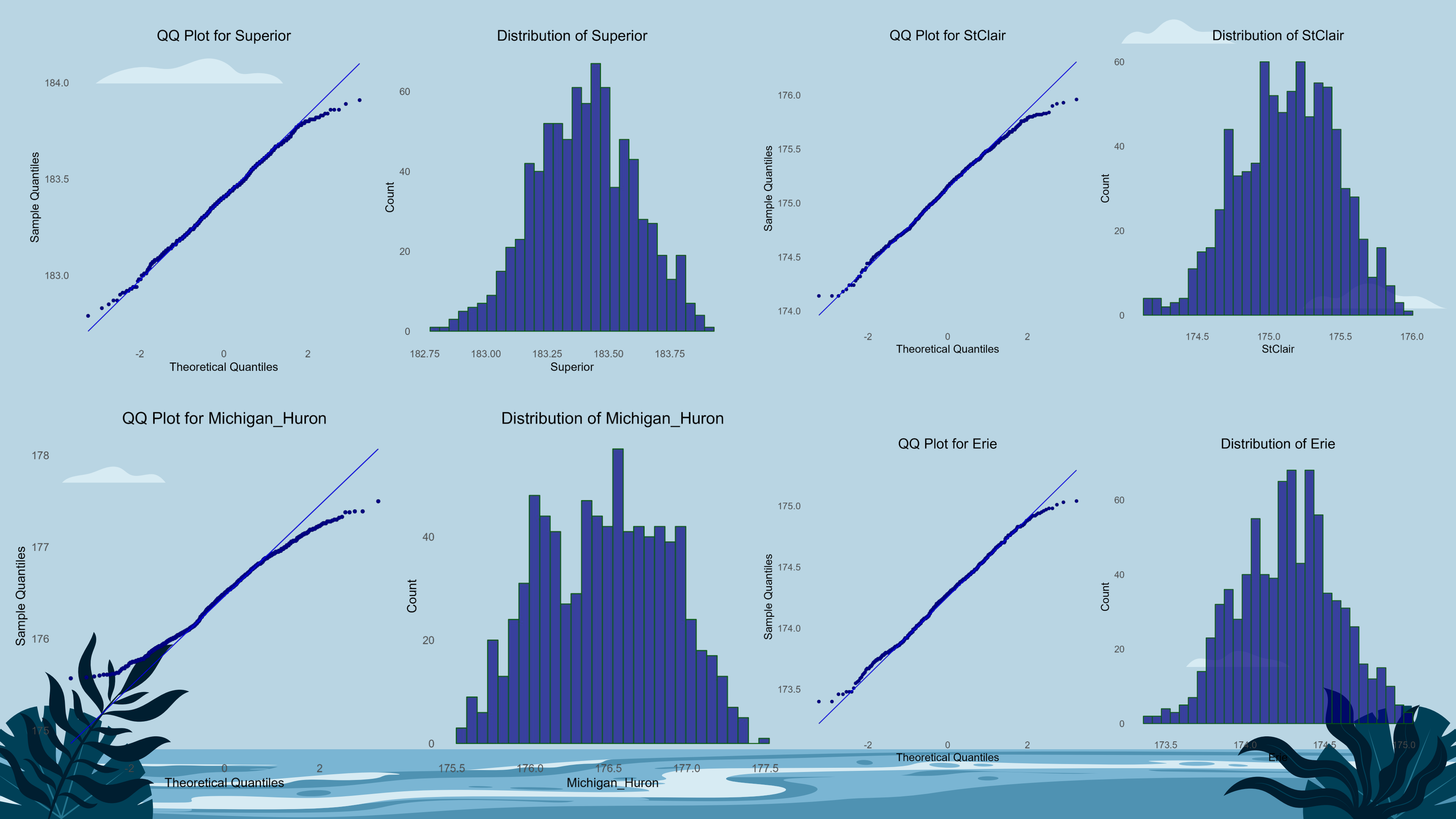


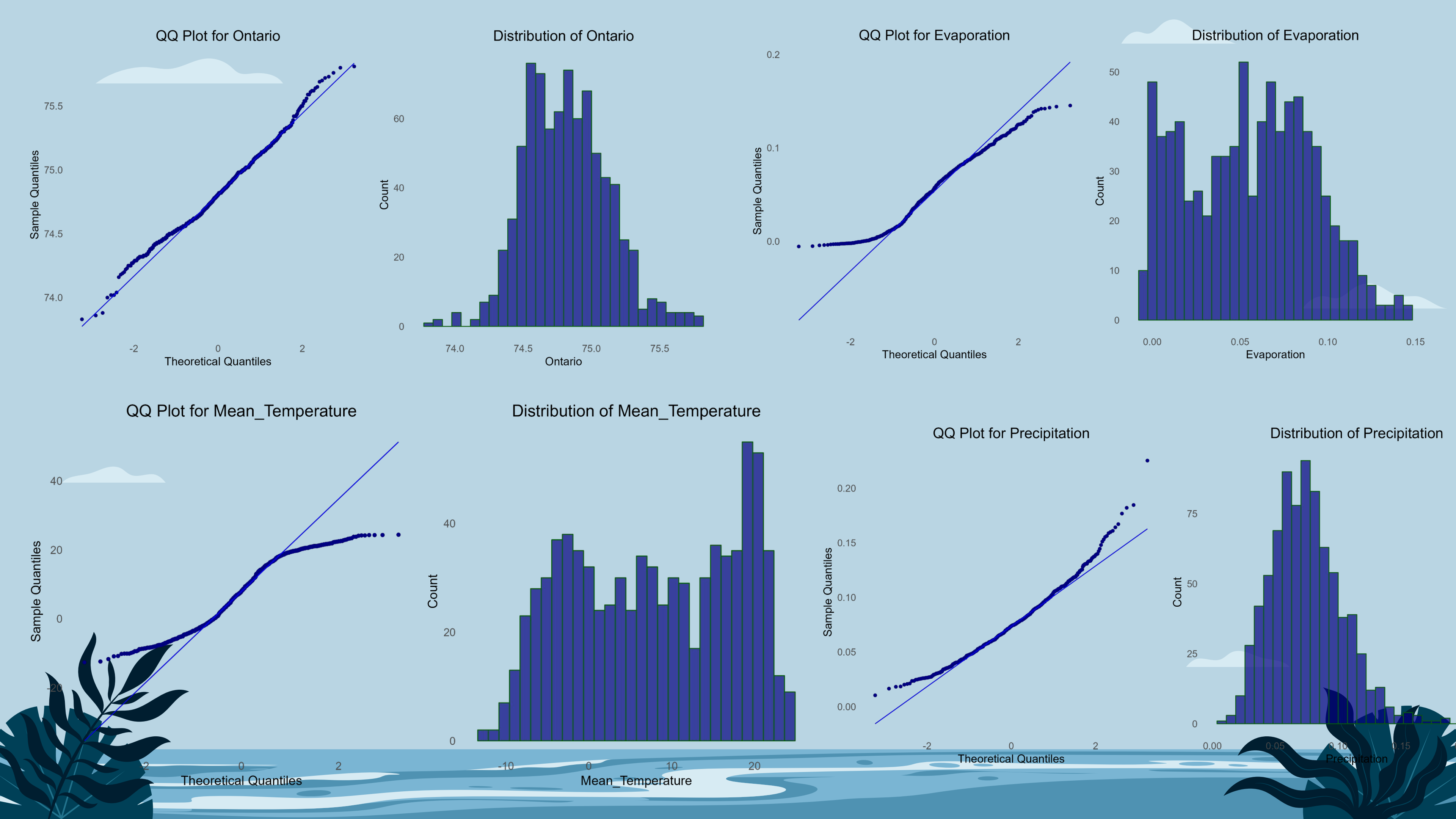
# Appendix: VAR Coefficients

```
$Ontario
                         Estimate   Std. Error     t value     Pr(>|t|)
Ontario.l1           1.060812e+00 5.950340e-02 17.8277577 8.148075e-51
Mean.Temperature.l1  1.520782e-06 4.334328e-07  3.5086919 5.101395e-04
Evaporation.l1       6.463521e-07 3.740447e-07  1.7280073 8.488433e-02
Precipitation.l1     1.168261e-05 1.239757e-06  9.4233034 6.598253e-19
Ontario.l2          -2.546570e-01 7.989556e-02 -3.1873731 1.567348e-03
Mean.Temperature.l2 -3.183272e-06 5.610639e-07 -5.6736359 2.969281e-08
Evaporation.l2      -9.303335e-07 3.965919e-07 -2.3458206 1.955259e-02
Precipitation.l2    -1.561687e-06 1.419158e-06 -1.1004325 2.719132e-01
Ontario.l3           2.045011e-02 5.039662e-02  0.4057834 6.851541e-01
Mean.Temperature.l3 -2.061234e-07 4.458502e-07 -0.4623152 6.441476e-01
Evaporation.l3       6.992916e-07 3.306515e-07  2.1148904 3.515799e-02
Precipitation.l3    -5.646024e-07 1.300029e-06 -0.4342998 6.643430e-01
const                1.886539e-01 2.817087e-02  6.6967729 8.697059e-11
```

# Appendix: VAR Residuals ACF plot




QQ Plot for Superior
184.0
183.5
183.0
Sample Quantiles
-2
0
2
Theoretical Quantiles
Distribution of Superior
60
40
20
0
Count
182.75
183.00
183.25
183.50
183.75
Superior
QQ Plot for StClair
176.0
175.5
175.0
174.5
174.0
Sample Quantiles
-2
0
2
Theoretical Quantiles
Distribution of StClair
60
40
20
0
Count
174.5
175.0
175.5
176.0
StClair
QQ Plot for Michigan_Huron
178
177
176
175
Sample Quantiles
-2
0
2
Theoretical Quantiles
Distribution of Michigan_Huron
40
20
0
Count
175.5
176.0
176.5
177.0
177.5
Michigan_Huron
QQ Plot for Erie
175.0
174.5
174.0
173.5
Sample Quantiles
-2
0
2
Theoretical Quantiles
Distribution of Erie
60
40
20
0
Count
173.5
174.0
174.5
175.0
Erie


QQ Plot for Ontario
Sample Quantiles
Theoretical Quantiles
Distribution of Ontario
Count
Ontario
QQ Plot for Evaporation
Sample Quantiles
Theoretical Quantiles
Distribution of Evaporation
Count
Evaporation
QQ Plot for Mean_Temperature
Sample Quantiles
Theoretical Quantiles
Distribution of Mean_Temperature
Count
Mean_Temperature
QQ Plot for Precipitation
Sample Quantiles
Theoretical Quantiles
Distribution of Precipitation
Count
Precipitation