Weather Temperature Forecasting on Jena Climate data

- Time series

Group Members:

Dinesh Chandra Gaddam

Project Overview

Scope of the Project

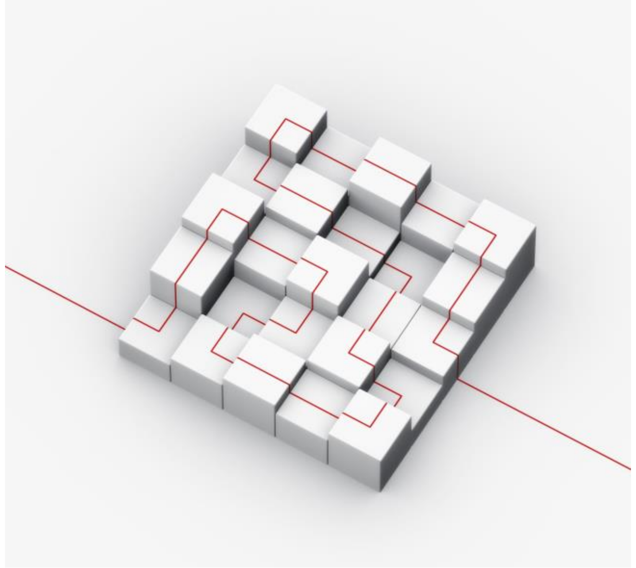
- Develop the best possible model using Arma, ARIMA, SARIMA or Box-Jenkins to forecast Temperature

- Perform all the required analysis on pretrained and post-trained data

Understanding the concepts and underlying methodology of the data

DATA SOURCE

Public Datasets: Use dataset from Kaggle named Jena Climate Data. Data used of period 2009 to 2012.

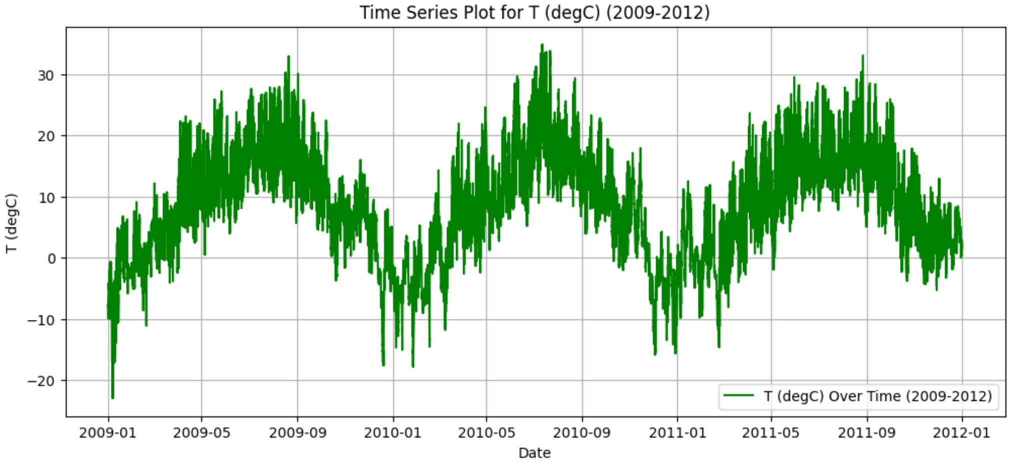


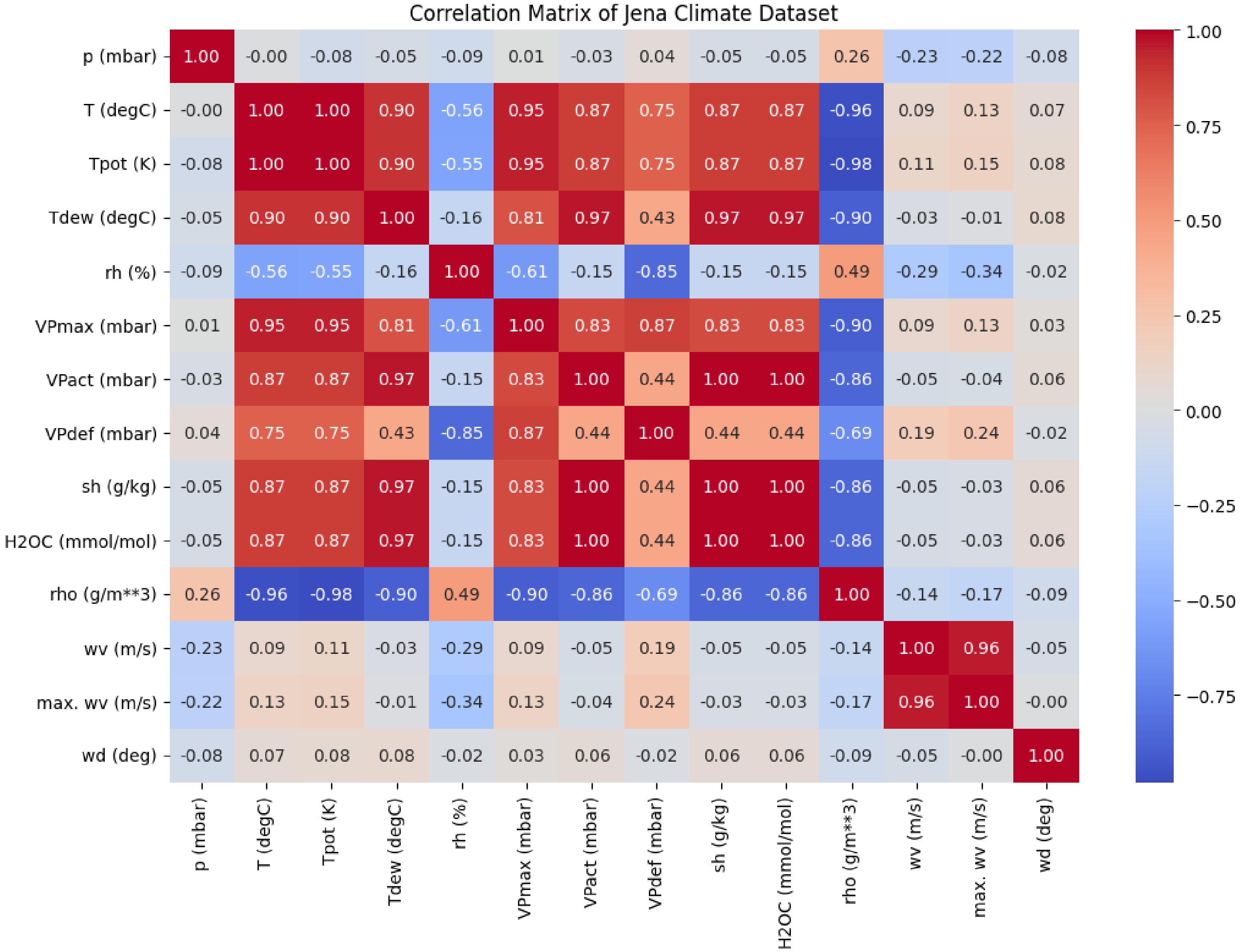
Project content

- Data Collection and Preprocessing
- Stationarity Check and Seasonal Decomposition
- Feature Selection and Base Models
- Model Development
- Model Evaluation
- Forecasting
- Deciding best model
- Summary

Data Collection and Preprocessing

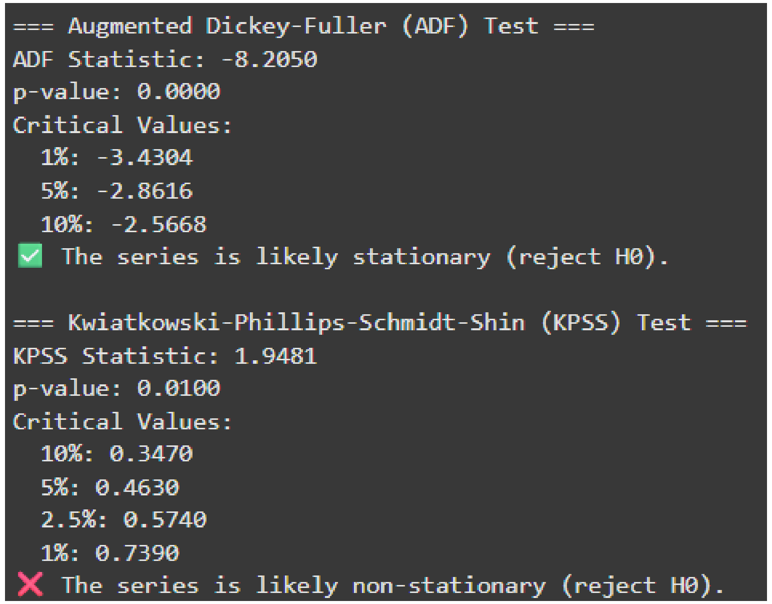
I have Limited Data to 3 years from 2009 to 2011 to make the data flexible to work with





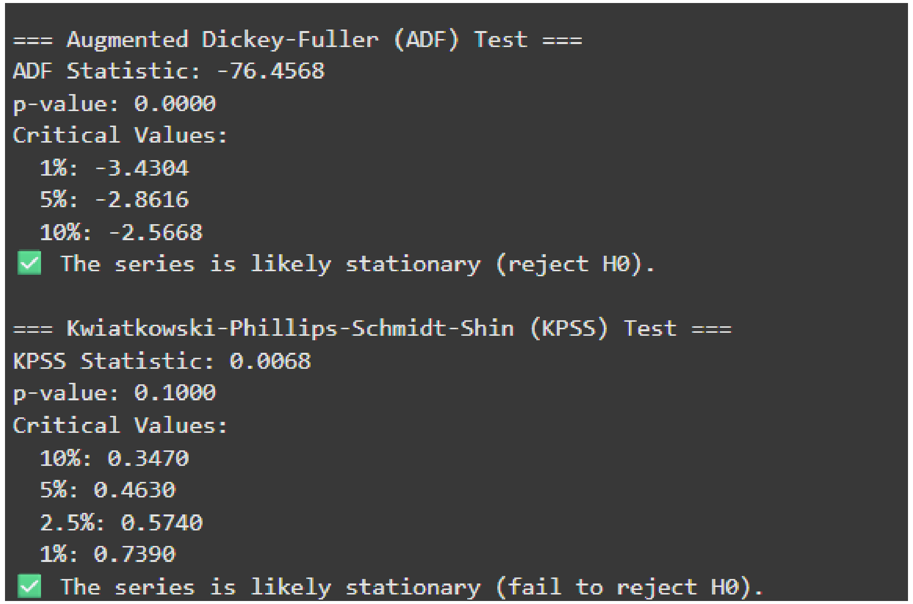
Stationarity Check and Seasonal Decomposition

Before differencing

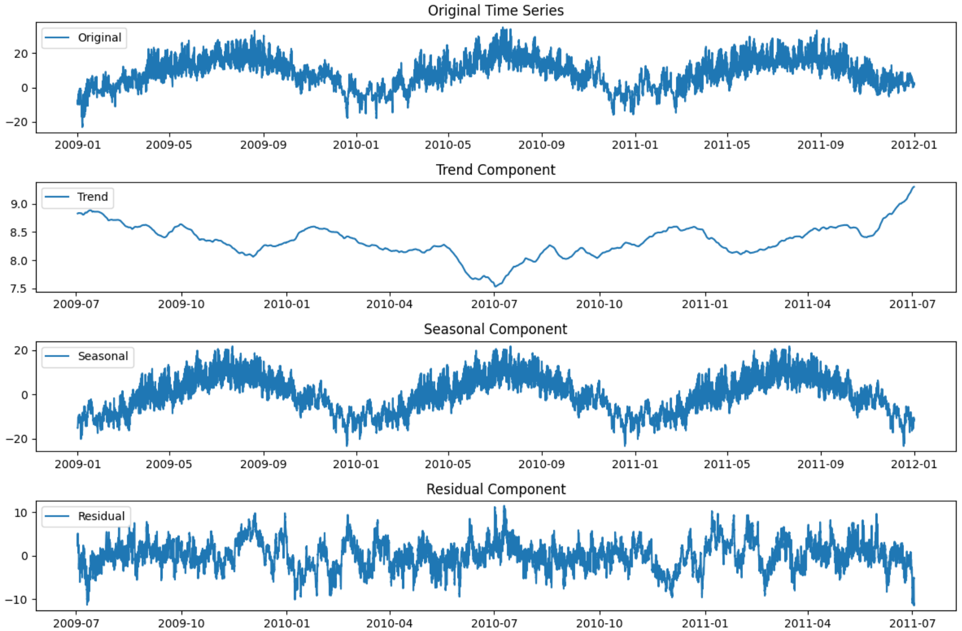


```
=== Augmented Dickey-Fuller (ADF) Test ===
ADF Statistic: -8.2050
p-value: 0.0000
Critical Values:
  1%: -3.4304
  5%: -2.8616
  10%: -2.5668
✅ The series is likely stationary (reject H0).

=== Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test ===
KPSS Statistic: 1.9481
p-value: 0.0100
Critical Values:
  10%: 0.3470
  5%: 0.4630
  2.5%: 0.5740
  1%: 0.7390
❌ The series is likely non-stationary (reject H0).
```

After Differencing



```
=== Augmented Dickey-Fuller (ADF) Test ===
ADF Statistic: -76.4568
p-value: 0.0000
Critical Values:
  1%: -3.4304
  5%: -2.8616
  10%: -2.5668
✅ The series is likely stationary (reject H0).

=== Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test ===
KPSS Statistic: 0.0068
p-value: 0.1000
Critical Values:
  10%: 0.3470
  5%: 0.4630
  2.5%: 0.5740
  1%: 0.7390
✅ The series is likely stationary (fail to reject H0).
```


Original Time Series

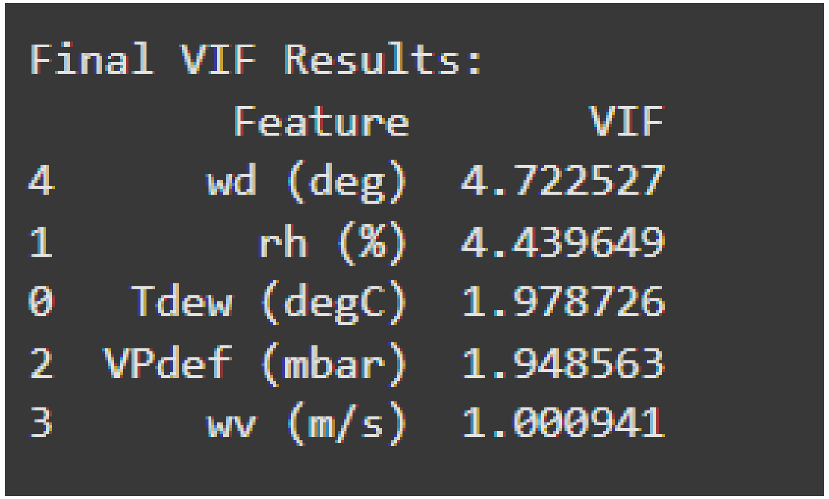
Trend Component

Seasonal Component

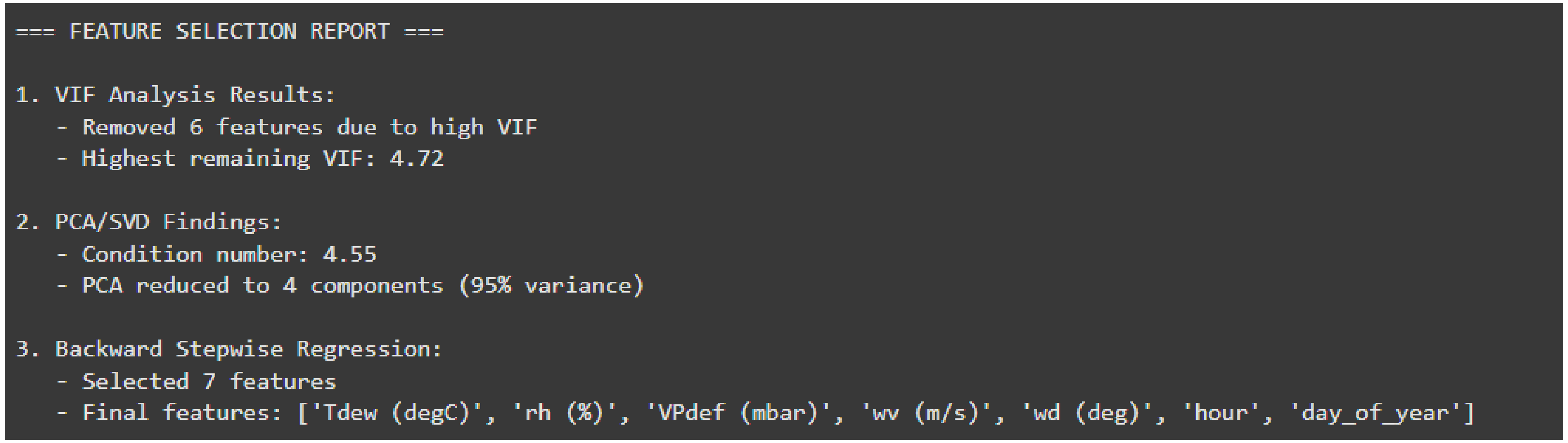
Residual Component

The Data is Seasonal: with 87% seasonality

Feature Selection and Base Models

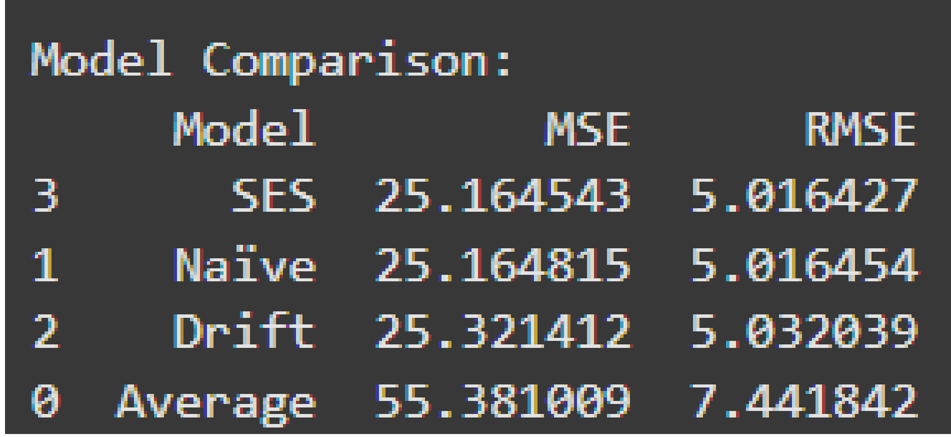


```
Final VIF Results:
          Feature       VIF
4        wd (deg)  4.722527
1          rh (%)  4.439649
0    Tdew (degC)  1.978726
2   VPdef (mbar)  1.948563
3        wv (m/s)  1.000941
```

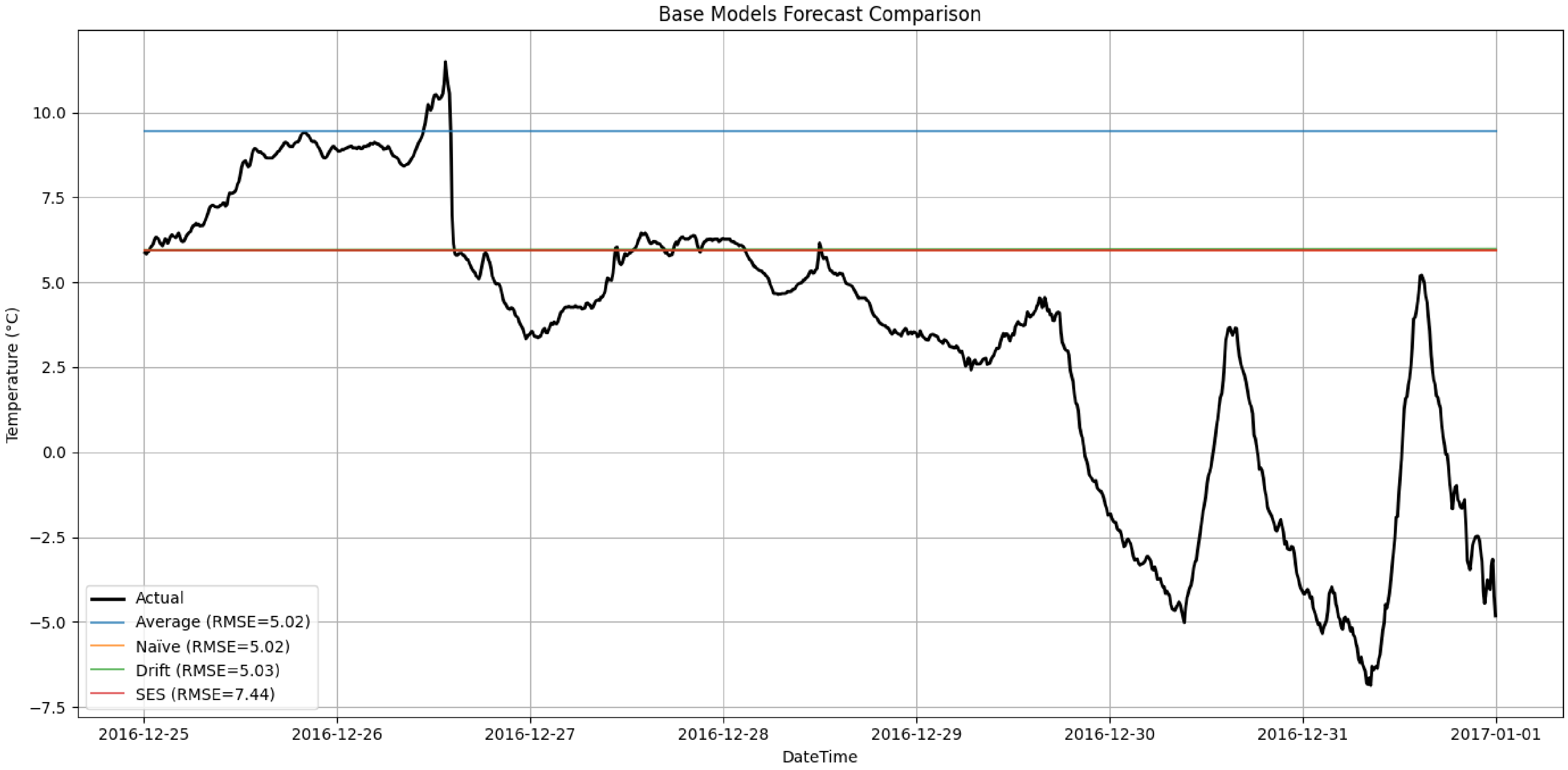


```
=== FEATURE SELECTION REPORT ===

1. VIF Analysis Results:
   - Removed 6 features due to high VIF
   - Highest remaining VIF: 4.72

2. PCA/SVD Findings:
   - Condition number: 4.55
   - PCA reduced to 4 components (95% variance)

3. Backward Stepwise Regression:
   - Selected 7 features
   - Final features: ['Tdew (degC)', 'rh (%)', 'VPdef (mbar)', 'wv (m/s)', 'wd (deg)', 'hour', 'day_of_year']
```

Feature Selection and Base Models

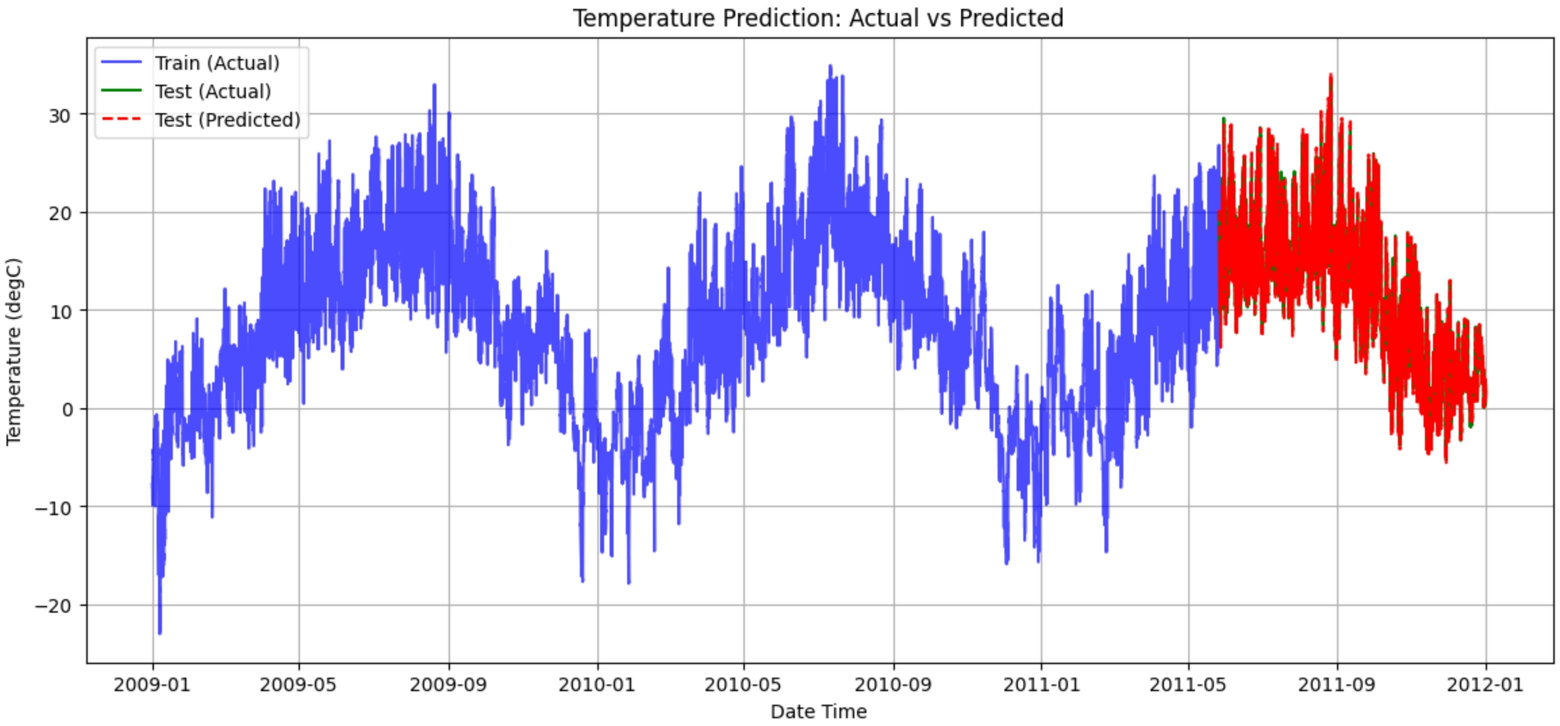


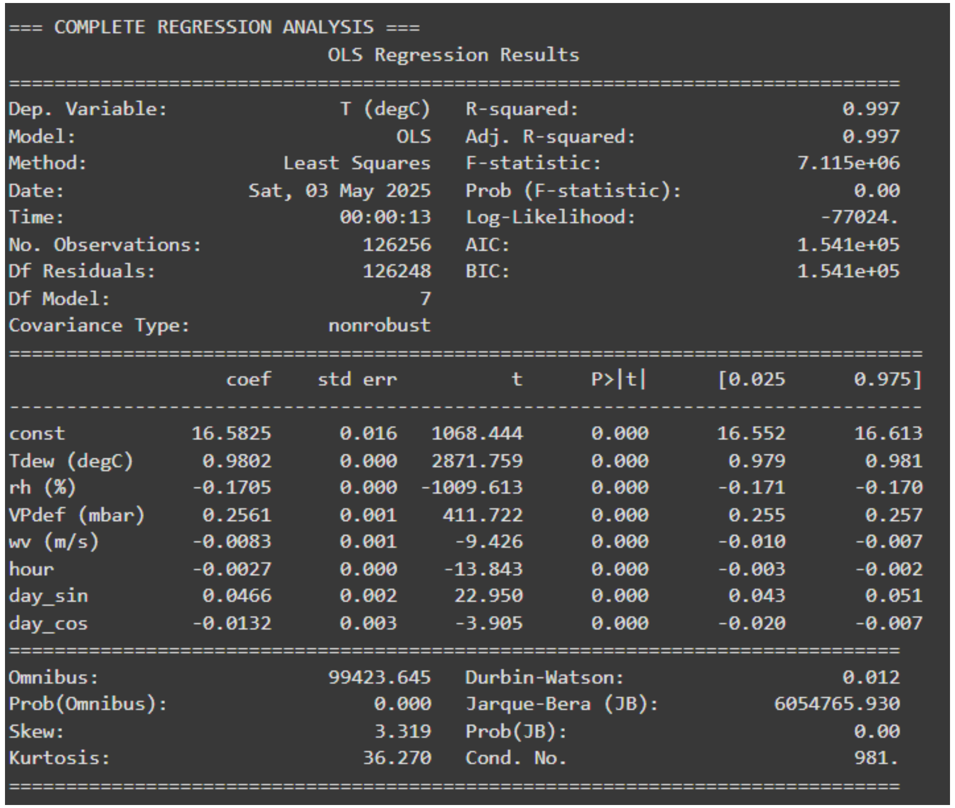
```
Model Comparison:
      Model        MSE      RMSE
3       SES  25.164543  5.016427
1     Naïve  25.164815  5.016454
2     Drift  25.321412  5.032039
0   Average  55.381009  7.441842
```



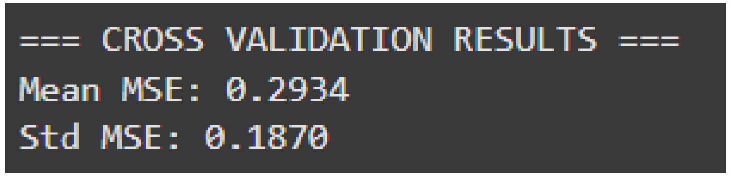
Model Development

Multiple Linear Regression



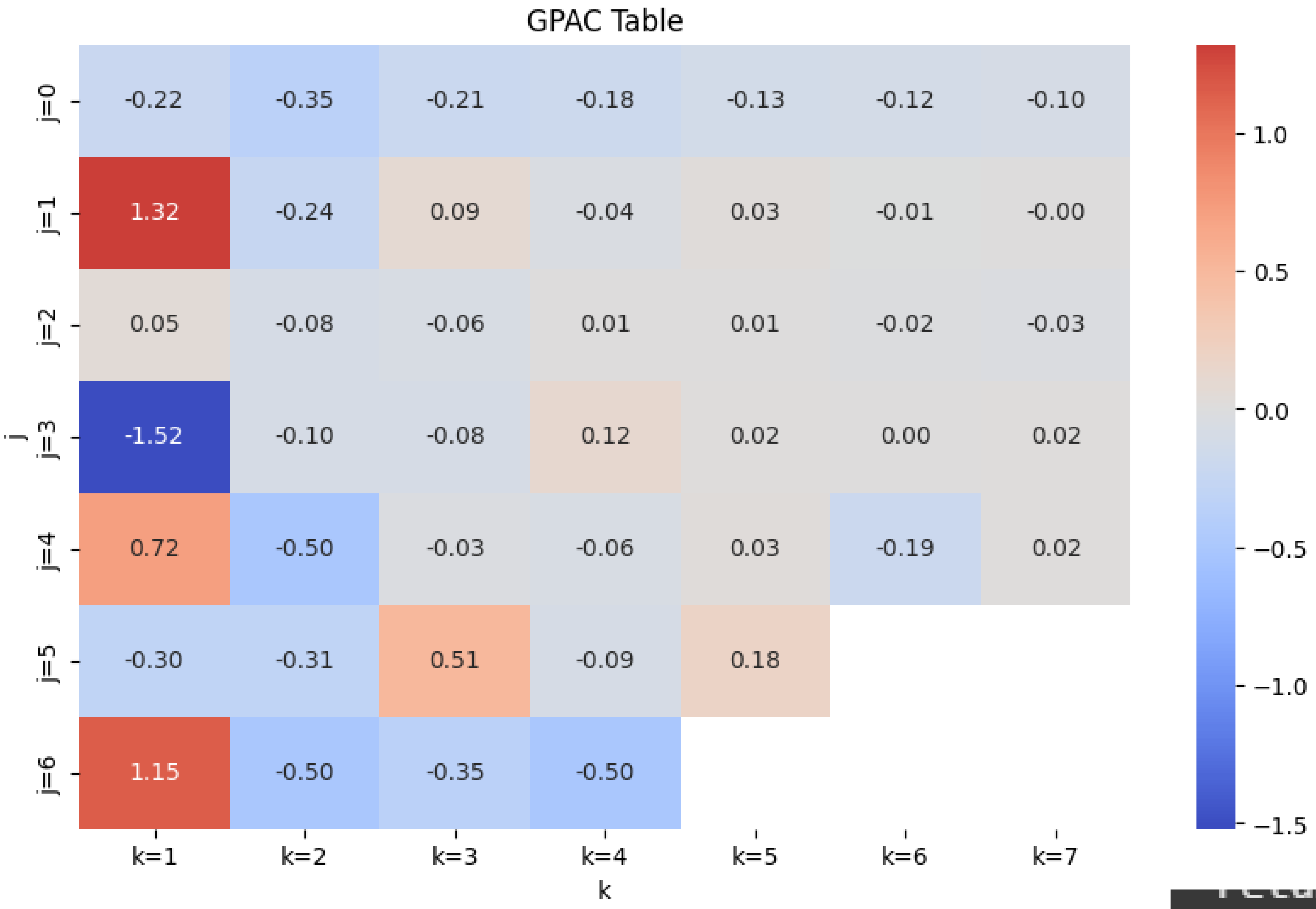


```
=== COMPLETE REGRESSION ANALYSIS ===
                            OLS Regression Results
==============================================================================
Dep. Variable:              T (degC)   R-squared:                       0.997
Model:                           OLS   Adj. R-squared:                  0.997
Method:                Least Squares   F-statistic:                 7.115e+06
Date:               Sat, 03 May 2025   Prob (F-statistic):               0.00
Time:                       00:00:13   Log-Likelihood:                -77024.
No. Observations:             126256   AIC:                         1.541e+05
Df Residuals:                 126248   BIC:                         1.541e+05
Df Model:                          7
Covariance Type:           nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          16.5825      0.016   1068.444      0.000      16.552      16.613
Tdew (degC)     0.9802      0.000   2871.759      0.000       0.979       0.981
rh (%)         -0.1705      0.000  -1009.613      0.000      -0.171      -0.170
VPdef (mbar)    0.2561      0.001    411.722      0.000       0.255       0.257
wv (m/s)       -0.0083      0.001     -9.426      0.000      -0.010      -0.007
hour           -0.0027      0.000    -13.843      0.000      -0.003      -0.002
day_sin         0.0466      0.002     22.950      0.000       0.043       0.051
day_cos        -0.0132      0.003     -3.905      0.000      -0.020      -0.007
==============================================================================
Omnibus:                   99423.645   Durbin-Watson:                   0.012
Prob(Omnibus):                 0.000   Jarque-Bera (JB):          6054765.930
Skew:                          3.319   Prob(JB):                         0.00
Kurtosis:                     36.270   Cond. No.                         981.
==============================================================================
```



```
=== CROSS VALIDATION RESULTS ===
Mean MSE: 0.2934
Std MSE: 0.1870
```

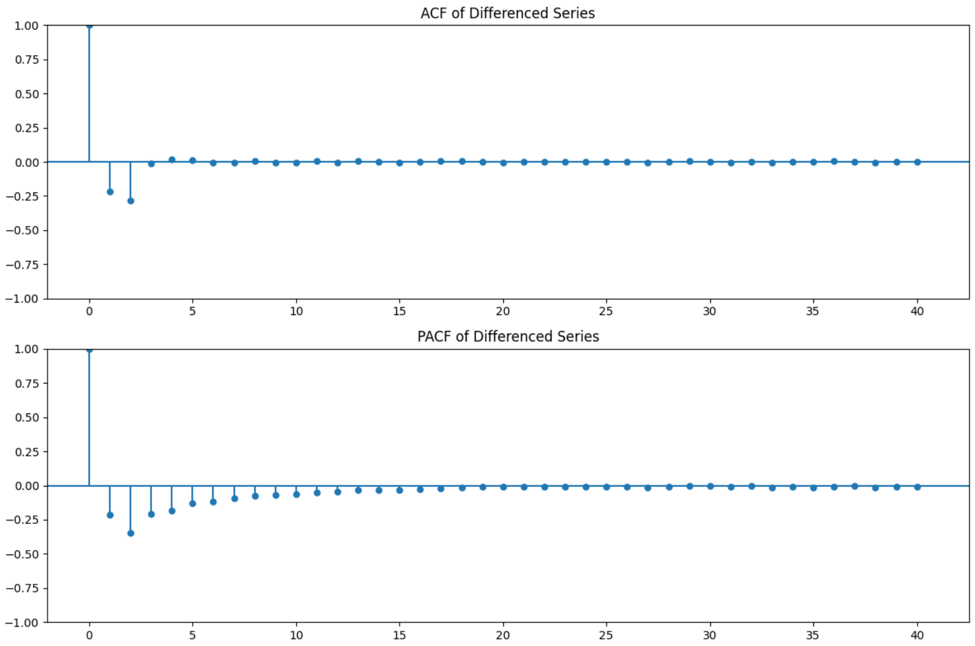
Model Development



GPAC Table

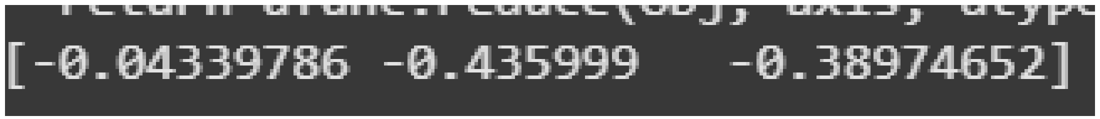
j \ k	k=1	k=2	k=3	k=4	k=5	k=6	k=7
j=0	-0.22	-0.35	-0.21	-0.18	-0.13	-0.12	-0.10
j=1	1.32	-0.24	0.09	-0.04	0.03	-0.01	-0.00
j=2	0.05	-0.08	-0.06	0.01	0.01	-0.02	-0.03
j=3	-1.52	-0.10	-0.08	0.12	0.02	0.00	0.02
j=4	0.72	-0.50	-0.03	-0.06	0.03	-0.19	0.02
j=5	-0.30	-0.31	0.51	-0.09	0.18		
j=6	1.15	-0.50	-0.35	-0.50			

j
k
1.0
0.5
0.0
−0.5
−1.0
−1.5

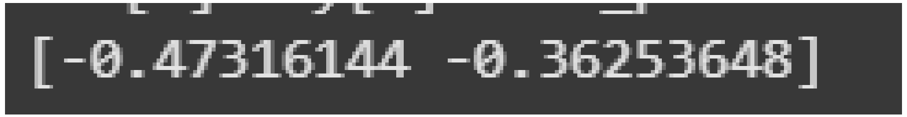


ACF of Differenced Series

PACF of Differenced Series



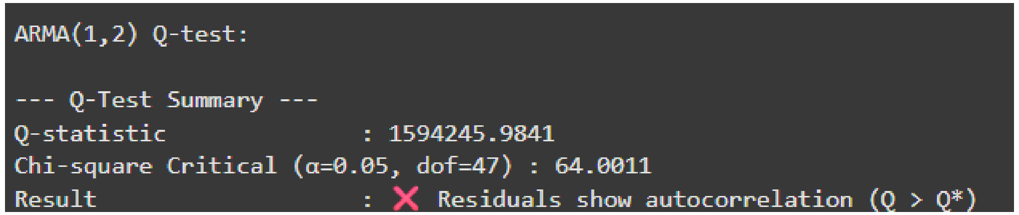
```
return ufunc.reduce(obj, axis, dtype
[-0.04339786 -0.435999   -0.38974652]
```



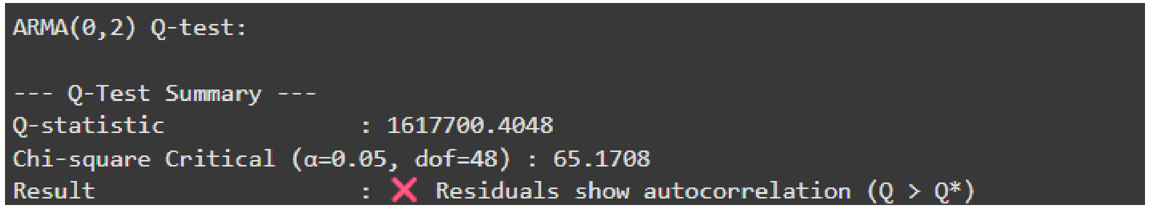
```
[-0.47316144 -0.36253648]
```

GPAC, PACF and ACF of the Stationary data, expecting order of AR: 0 or 2 and MA order of 2

ARMA observations

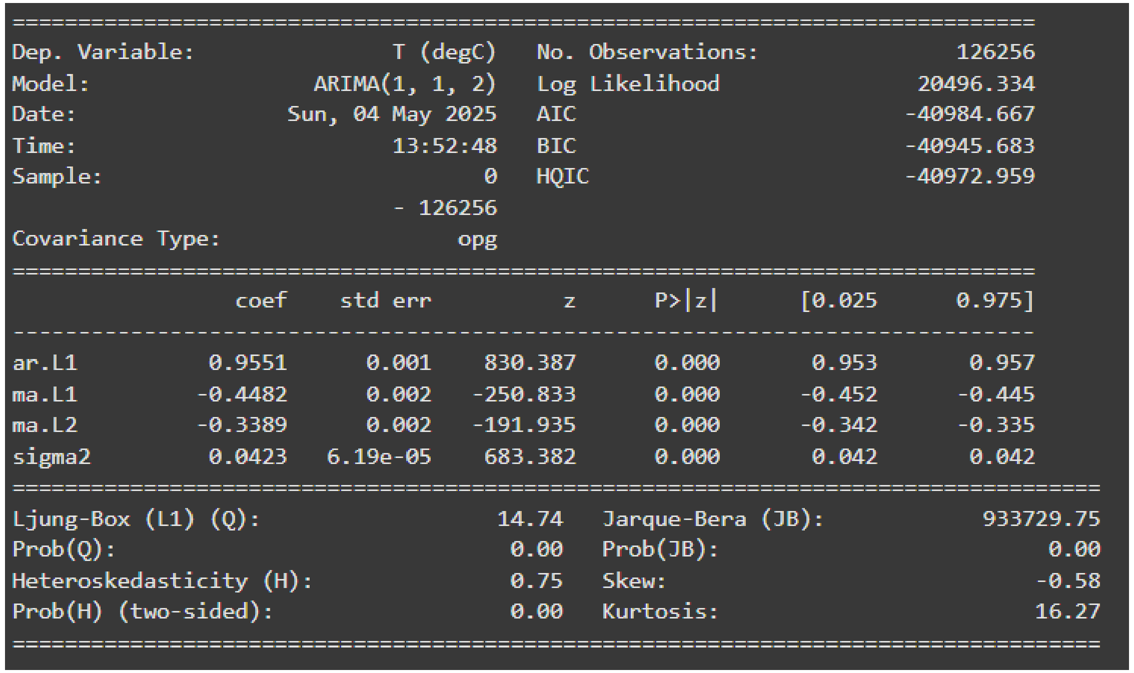


```
ARMA(1,2) Q-test:

--- Q-Test Summary ---
Q-statistic                    : 1594245.9841
Chi-square Critical (α=0.05, dof=47) : 64.0011
Result                         : ❌ Residuals show autocorrelation (Q > Q*)
```

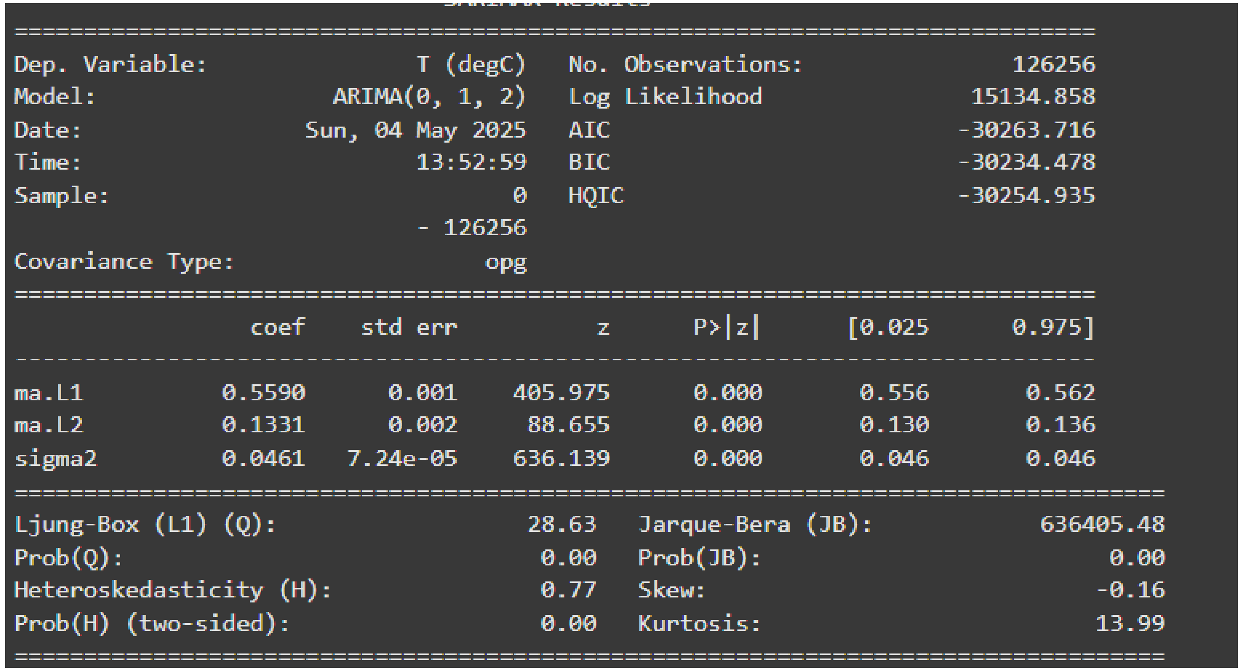


```
ARMA(0,2) Q-test:

--- Q-Test Summary ---
Q-statistic                    : 1617700.4048
Chi-square Critical (α=0.05, dof=48) : 65.1708
Result                         : ❌ Residuals show autocorrelation (Q > Q*)
```

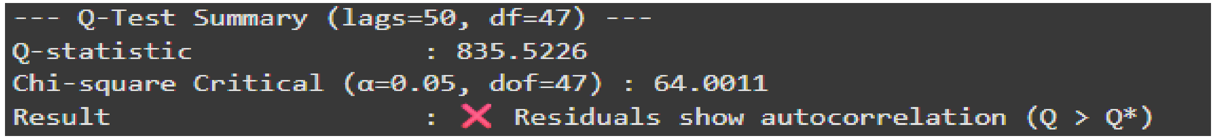
ARIMA Observations



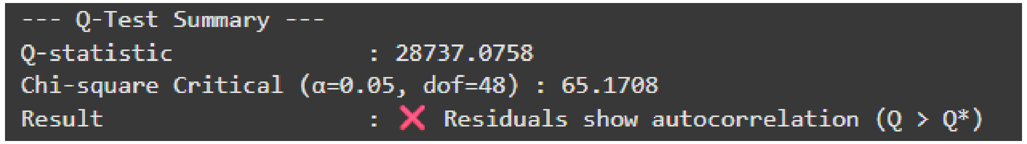
```
==============================================================================
Dep. Variable:               T (degC)   No. Observations:               126256
Model:                 ARIMA(1, 1, 2)   Log Likelihood               20496.334
Date:                Sun, 04 May 2025   AIC                         -40984.667
Time:                        13:52:48   BIC                         -40945.683
Sample:                             0   HQIC                        -40972.959
                             - 126256
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9551      0.001    830.387      0.000       0.953       0.957
ma.L1         -0.4482      0.002   -250.833      0.000      -0.452      -0.445
ma.L2         -0.3389      0.002   -191.935      0.000      -0.342      -0.335
sigma2         0.0423   6.19e-05    683.382      0.000       0.042       0.042
===================================================================================
Ljung-Box (L1) (Q):                  14.74   Jarque-Bera (JB):            933729.75
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.75   Skew:                            -0.58
Prob(H) (two-sided):                  0.00   Kurtosis:                        16.27
===================================================================================
```



```
                               SARIMAX Results
==============================================================================
Dep. Variable:               T (degC)   No. Observations:               126256
Model:                 ARIMA(0, 1, 2)   Log Likelihood               15134.858
Date:                Sun, 04 May 2025   AIC                         -30263.716
Time:                        13:52:59   BIC                         -30234.478
Sample:                             0   HQIC                        -30254.935
                             - 126256
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.5590      0.001    405.975      0.000       0.556       0.562
ma.L2          0.1331      0.002     88.655      0.000       0.130       0.136
sigma2         0.0461   7.24e-05    636.139      0.000       0.046       0.046
===================================================================================
Ljung-Box (L1) (Q):                  28.63   Jarque-Bera (JB):            636405.48
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.77   Skew:                            -0.16
Prob(H) (two-sided):                  0.00   Kurtosis:                        13.99
===================================================================================
```

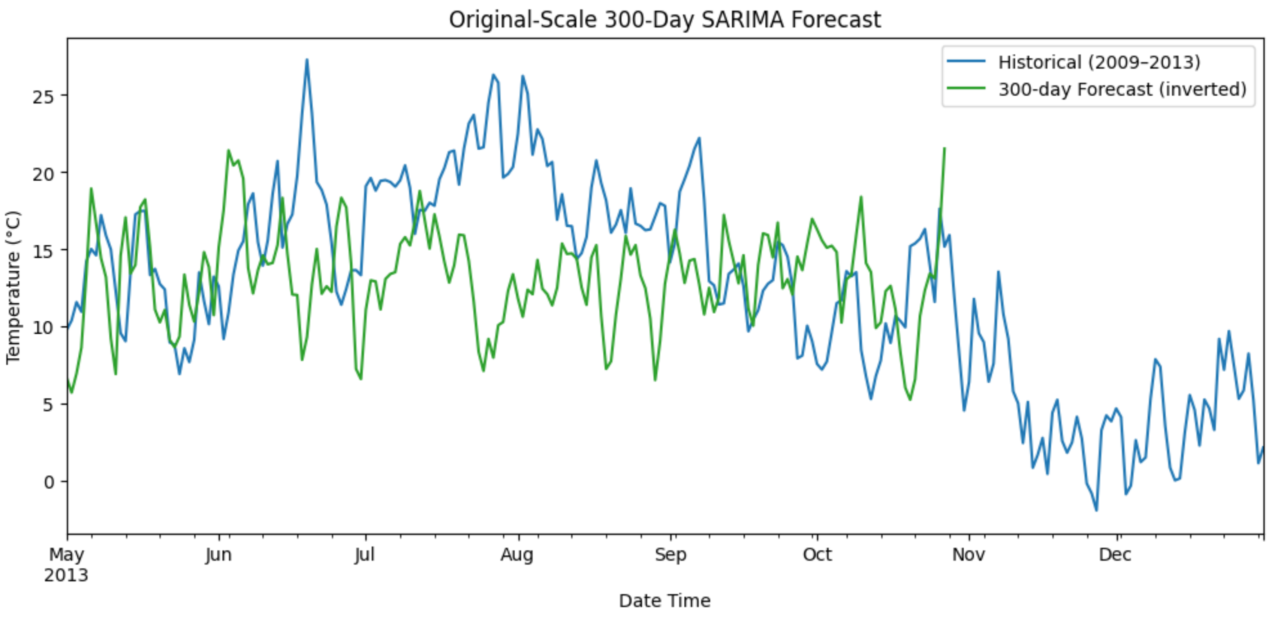


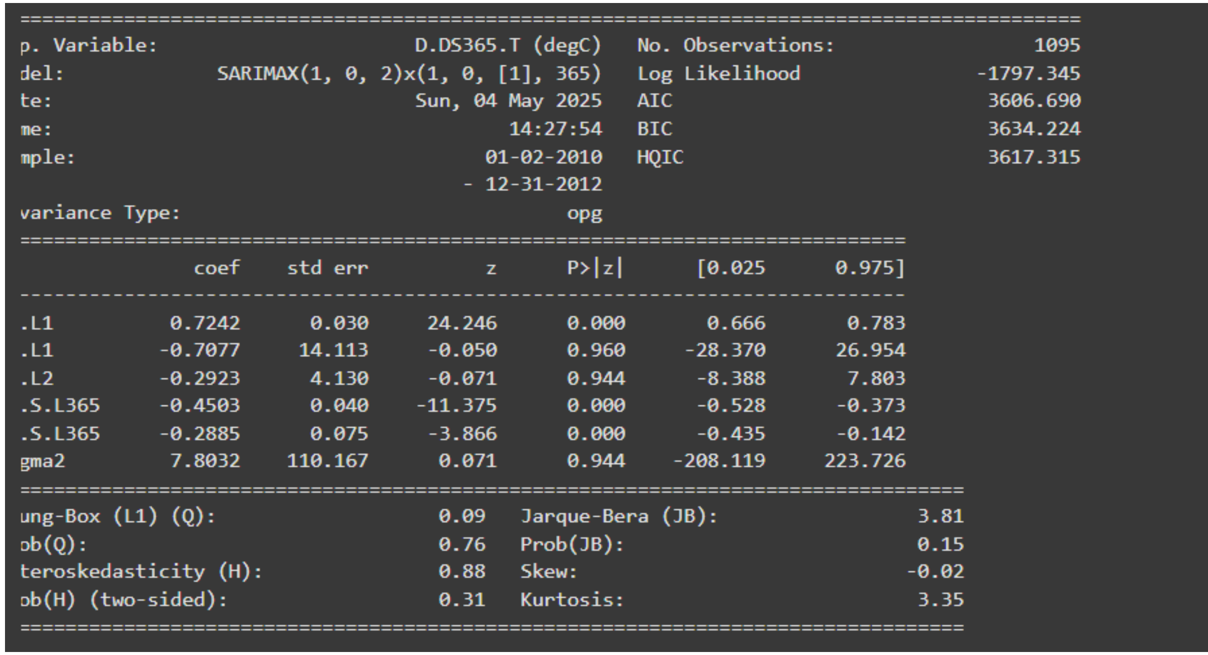
```
--- Q-Test Summary (lags=50, df=47) ---
Q-statistic                    : 835.5226
Chi-square Critical (α=0.05, dof=47) : 64.0011
Result                         : ❌ Residuals show autocorrelation (Q > Q*)
```



```
--- Q-Test Summary ---
Q-statistic                    : 28737.0758
Chi-square Critical (α=0.05, dof=48) : 65.1708
Result                         : ❌ Residuals show autocorrelation (Q > Q*)
```

SARIMA observations for its best model..

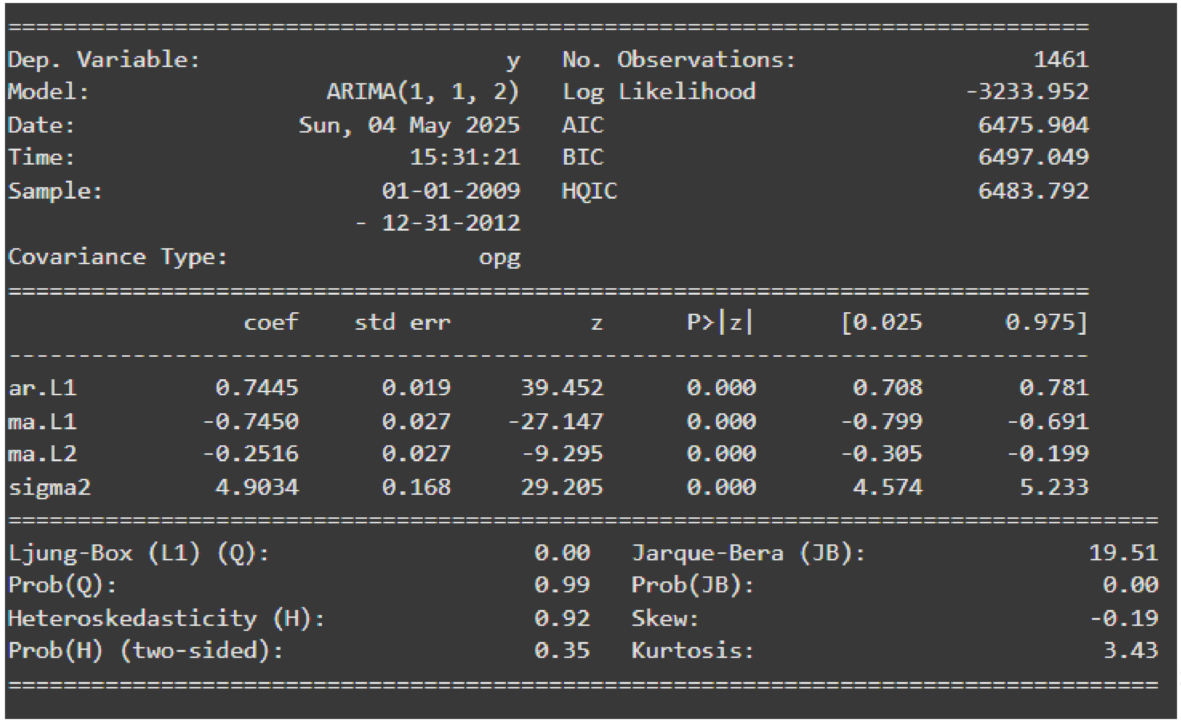




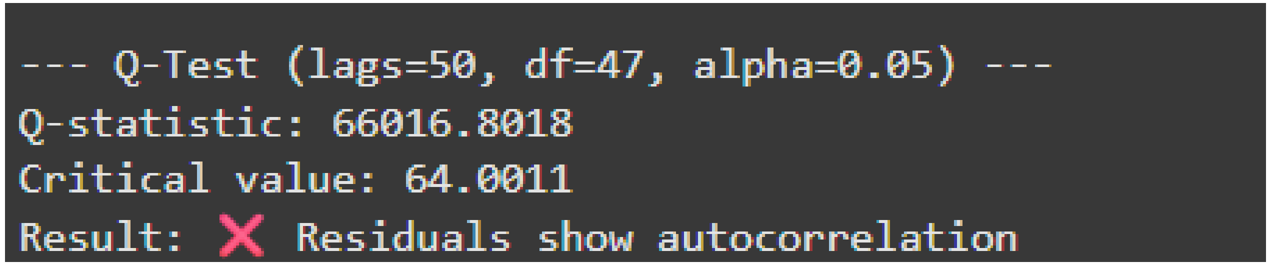
```
=============================================================================================
p. Variable:                       D.DS365.T (degC)   No. Observations:                  1095
del:              SARIMAX(1, 0, 2)x(1, 0, [1], 365)   Log Likelihood                -1797.345
te:                                Sun, 04 May 2025   AIC                            3606.690
me:                                        14:27:54   BIC                            3634.224
mple:                                    01-02-2010   HQIC                           3617.315
                                       - 12-31-2012
variance Type:                                  opg
==============================================================================
               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
.L1          0.7242      0.030     24.246      0.000       0.666       0.783
.L1         -0.7077     14.113     -0.050      0.960     -28.370      26.954
.L2         -0.2923      4.130     -0.071      0.944      -8.388       7.803
.S.L365     -0.4503      0.040    -11.375      0.000      -0.528      -0.373
.S.L365     -0.2885      0.075     -3.866      0.000      -0.435      -0.142
gma2         7.8032    110.167      0.071      0.944    -208.119     223.726
===================================================================================
ung-Box (L1) (Q):                    0.09   Jarque-Bera (JB):                  3.81
ob(Q):                               0.76   Prob(JB):                          0.15
teroskedasticity (H):                0.88   Skew:                             -0.02
ob(H) (two-sided):                   0.31   Kurtosis:                          3.35
===================================================================================
```

Observation: the model captures the seasonality but there is almost 0 trend in my data. So the model was unable to access the right movement. To solve this issue I used **ARIMA(0,1,2) on Seasonally-Adjusted Data + Seasonal Recomposition Forecast**

ARIMA(0,1,2) on Seasonally-Adjusted Data + Seasonal Recomposition Forecast

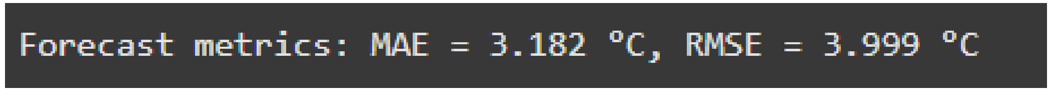


```
==============================================================================
Dep. Variable:                      y   No. Observations:                 1461
Model:                 ARIMA(1, 1, 2)   Log Likelihood               -3233.952
Date:                Sun, 04 May 2025   AIC                           6475.904
Time:                        15:31:21   BIC                           6497.049
Sample:                    01-01-2009   HQIC                          6483.792
                         - 12-31-2012
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7445      0.019     39.452      0.000       0.708       0.781
ma.L1         -0.7450      0.027    -27.147      0.000      -0.799      -0.691
ma.L2         -0.2516      0.027     -9.295      0.000      -0.305      -0.199
sigma2         4.9034      0.168     29.205      0.000       4.574       5.233
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                19.51
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.92   Skew:                            -0.19
Prob(H) (two-sided):                  0.35   Kurtosis:                         3.43
===================================================================================
```

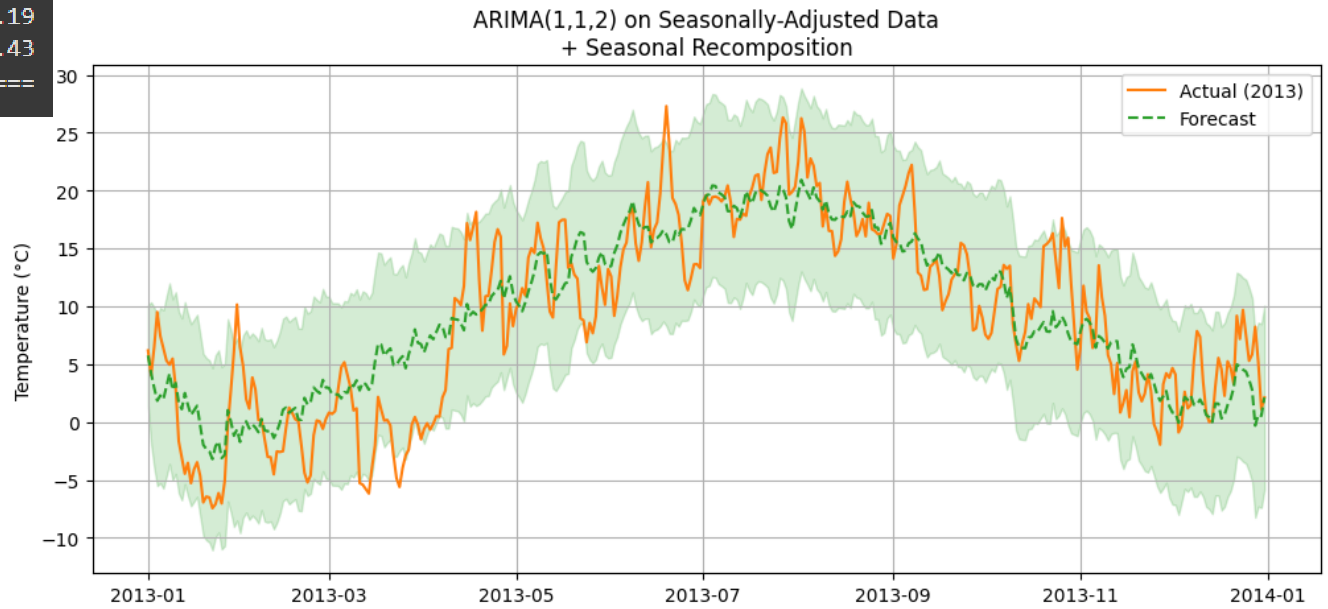


```
--- Q-Test (lags=50, df=47, alpha=0.05) ---
Q-statistic: 66016.8018
Critical value: 64.0011
Result: ❌ Residuals show autocorrelation
```

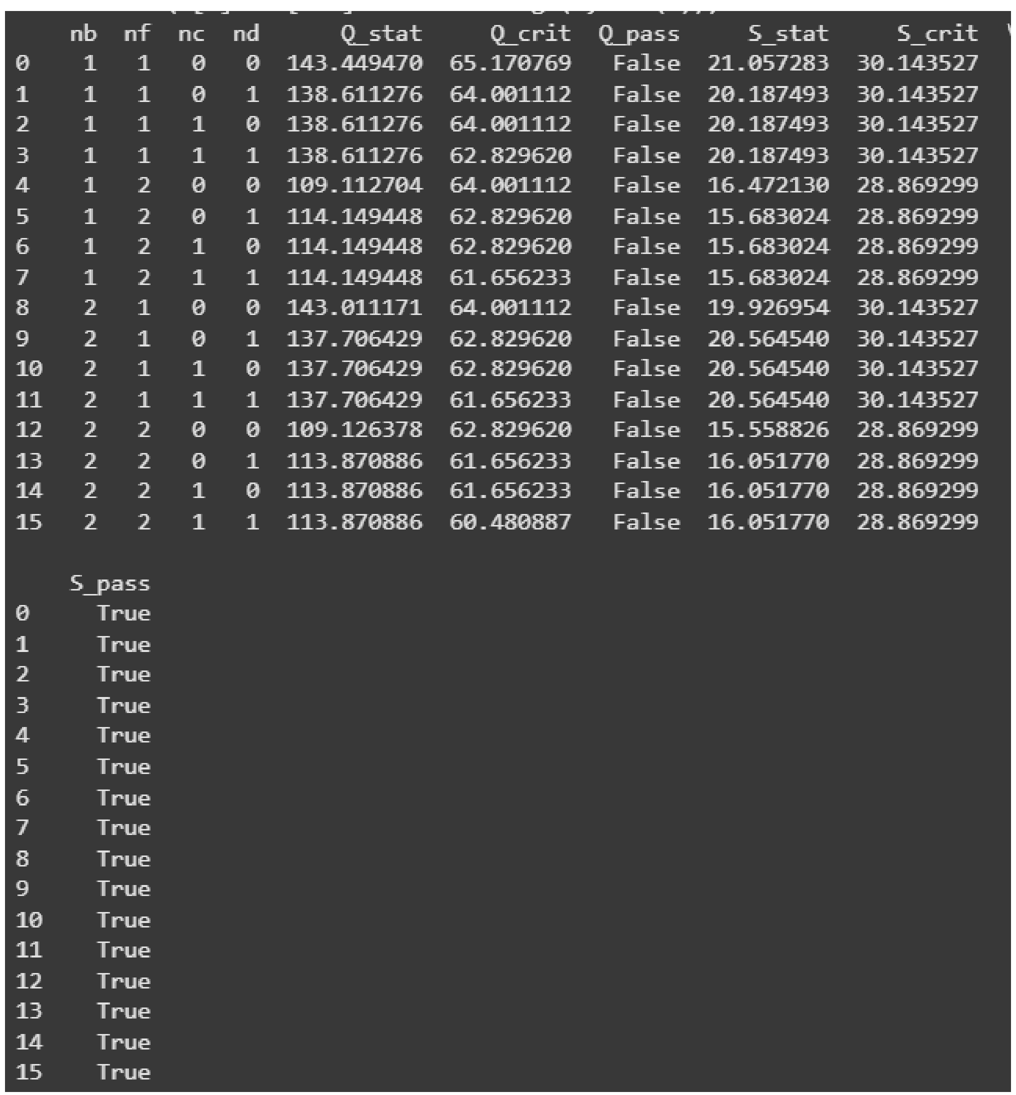
365 step forecast of the resampled data=>



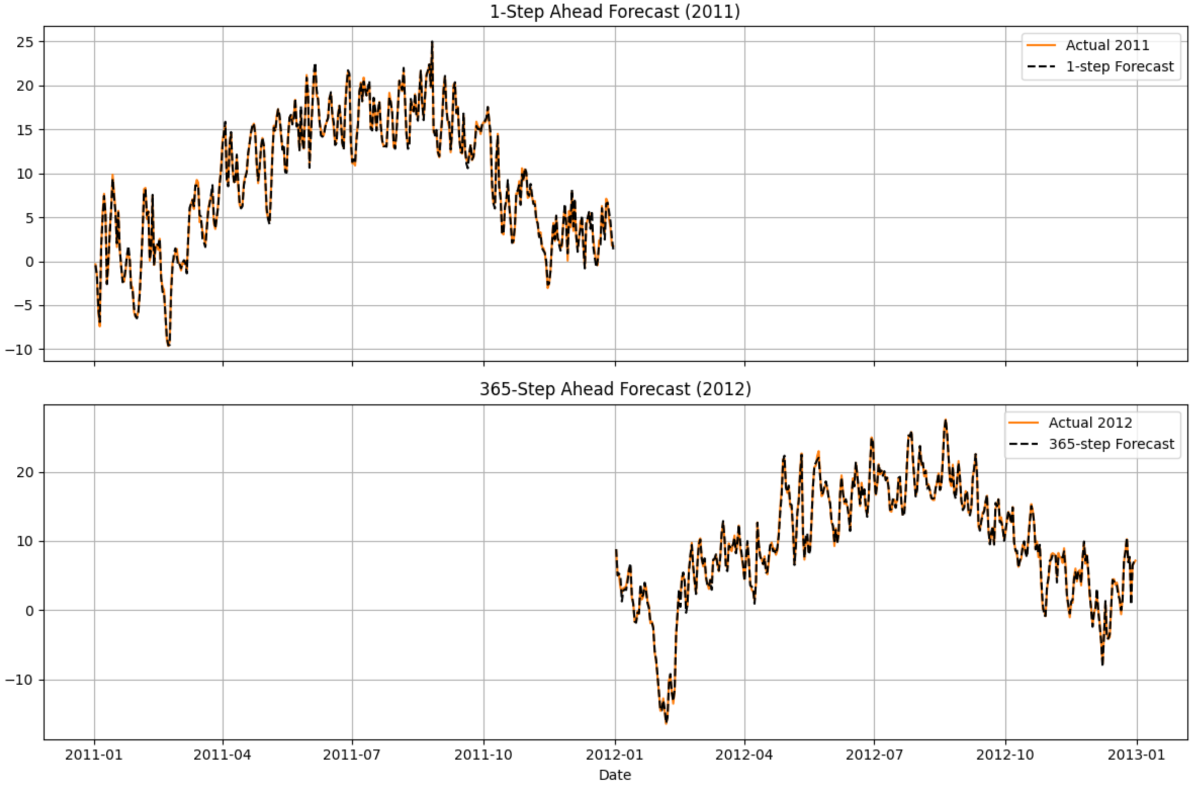
```
Forecast metrics: MAE = 3.182 °C, RMSE = 3.999 °C
```



Box-Jenkins



```
    nb  nf  nc  nd     Q_stat     Q_crit  Q_pass     S_stat     S_crit
0    1   1   0   0 143.449470  65.170769   False  21.057283  30.143527
1    1   1   0   1 138.611276  64.001112   False  20.187493  30.143527
2    1   1   1   0 138.611276  64.001112   False  20.187493  30.143527
3    1   1   1   1 138.611276  62.829620   False  20.187493  30.143527
4    1   2   0   0 109.112704  64.001112   False  16.472130  28.869299
5    1   2   0   1 114.149448  62.829620   False  15.683024  28.869299
6    1   2   1   0 114.149448  62.829620   False  15.683024  28.869299
7    1   2   1   1 114.149448  61.656233   False  15.683024  28.869299
8    2   1   0   0 143.011171  64.001112   False  19.926954  30.143527
9    2   1   0   1 137.706429  62.829620   False  20.564540  30.143527
10   2   1   1   0 137.706429  62.829620   False  20.564540  30.143527
11   2   1   1   1 137.706429  61.656233   False  20.564540  30.143527
12   2   2   0   0 109.126378  62.829620   False  15.558826  28.869299
13   2   2   0   1 113.870886  61.656233   False  16.051770  28.869299
14   2   2   1   0 113.870886  61.656233   False  16.051770  28.869299
15   2   2   1   1 113.870886  60.480887   False  16.051770  28.869299

    S_pass
0     True
1     True
2     True
3     True
4     True
5     True
6     True
7     True
8     True
9     True
10    True
11    True
12    True
13    True
14    True
15    True
```



Q-test: Q=143.0, crit=64.0, df=47 -> ❌ autocorrelation

S-test: S=19.9, crit=30.1, df=19 -> ✅ G(q) accurate

Summary

- The Models have been built for ARMA, ARIMA, SARIMA, BOX-Jenkins etc:- al from all of them Box-Jenkins Has performed the best..

- I have derived the order from GPAC, PACF and ACF plots by feeding them the stationary data..

- The model received a accuracy of RMSE of 0.360 for the normalised data after making it stationary...

References

Notes and Assignments

THANK YOU!