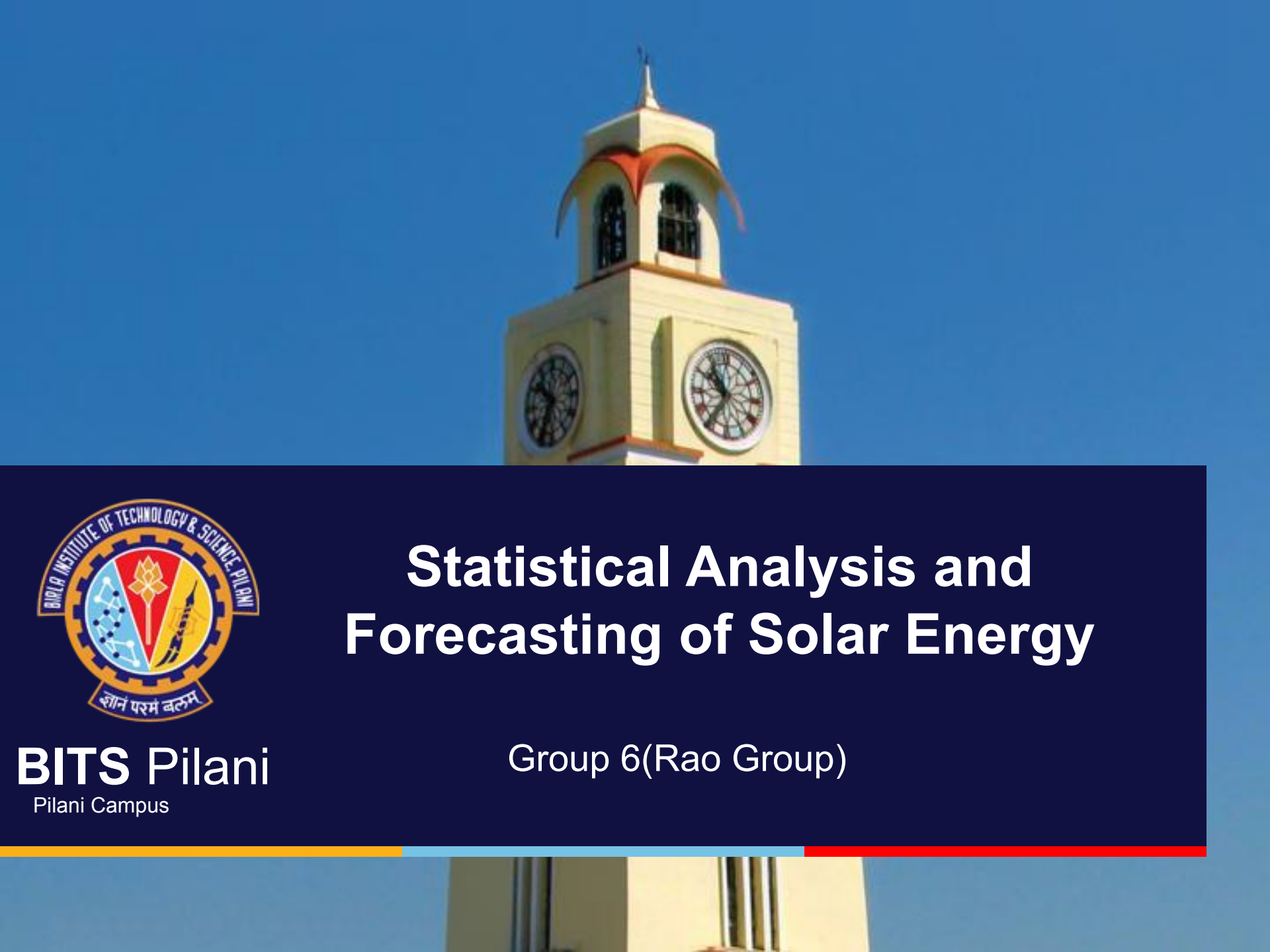


Statistical Analysis and Forecasting of Solar Energy

BITS Pilani
Pilani Campus

Group 6(Rao Group)

Introduction

Solar Energy is a pivotal component of renewable energy resources, offering sustainable and environmentally friendly solutions for power generation. Solar power in India is a fast developing industry as India receives an abundant amount of sunlight throughout the year. The Ministry of New and Renewable Energy (MNRE) also plays a crucial role in promoting and regulating this renewable energy resource.

Though the sunlight received by India is abundant, the solar power output in solar plants depends on various uncontrollable variables which affect the amount of sunlight falling on the solar panels. So we tried to forecast the amount of solar energy a plant would receive next day or next week based on the 15 year data obtained between 2000 to 2014 from 2 solar parks in Rajasthan.

Some key terms

innovate achieve lead

- **DHI (Direct Horizontal Irradiance)** - It is the solar radiation received by the earth's surface in a diffuse manner because of scattering in the atmosphere.
- **DNI (Direct Normal Irradiance)** - It is the solar radiation received on earth's surface perpendicular to the sun's rays. It is the direct sunlight which isn't scattered by the earth's atmosphere.
- **GHI (Global Horizontal Irradiance)** - It is the total radiation received on the horizontal surface, combining direct sunlight and the diffused sunlight because of the atmosphere.

$$GHI = DNI + DHI * \cos Z$$

Here, Z is the solar zenith angle made between the vertical and the diffused Sun ray.

- **Dew Point** - It is the temperature at which the air becomes saturated with moisture which becomes dew at the end.
- **Relative Humidity** - It is the ratio of the current amount of moisture in air to the maximum amount air could hold at the current pressure.

Dataset

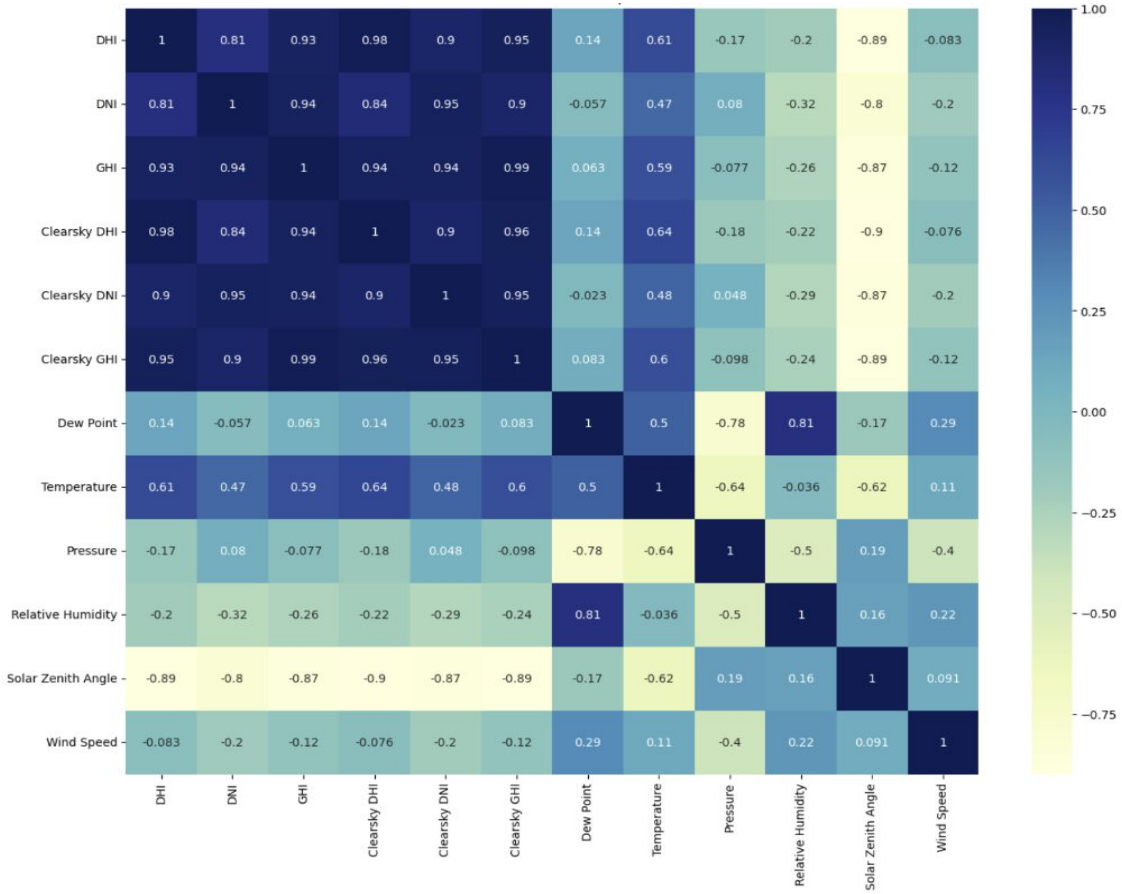
Information collected over a period of 15 years (2000-2014) in 2 solar parks of Rajasthan. The following features are provided in the dataset:

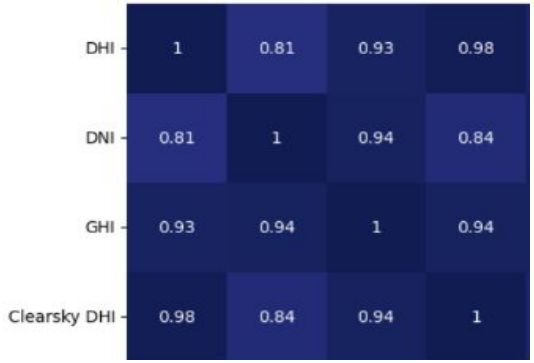
1. Date and time
2. DHI
3. DNI
4. GHI
5. Clearsky DHI
6. Clearsky DNI
7. Clearsky GHI
8. Dew Point
9. Temperature
10. Pressure
11. Relative Humidity
12. Solar Zenith Angle
13. Snow Depth
14. Wind Speed

Year	Month	Day	Hour	Minute	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	Temperature	Pressure	Relative Humidit	Solar Zenith Anç	Snow Depth	Wind Speed
2000	1	1	0	0	0	0	0	0	0	0	-4	13.81143711	982.7498169	27.44786192	174.7559969	0	3.976945162
2000	1	1	1	0	0	0	0	0	0	0	-4	13.11730025	982.670105	29.04885655	169.5442047	0	4.017370701
2000	1	1	2	0	0	0	0	0	0	0	-4	12.43395971	982.4041748	31.21665891	156.3409377	0	3.992325068
2000	1	1	3	0	0	0	0	0	0	0	-3	11.79763165	982.5552979	34.95607459	142.9458642	0	3.981812477
2000	1	1	4	0	0	0	0	0	0	0	-1	11.19548679	982.5581055	41.17753996	129.6257669	0	3.985007524
2000	1	1	5	0	0	0	0	0	0	0	-11	11.13673005	984.3792725	19.56375494	116.4282082	0	3.480460167
2000	1	1	6	0	0	0	0	0	0	0	-10	10.64697883	985.1685791	21.35784818	103.5837831	0	3.523517609
2000	1	1	7	0	0	0	0	0	0	0	-9	11.26520893	985.9255981	22.16891612	91.18164767	0	3.546341181
2000	1	1	8	0	78	306	135	74	354	139	-8	14.27732456	986.4864502	20.1208141	79.45062726	0	3.323988438
2000	1	1	9	0	114	597	331	121	600	339	-6	17.67727827	986.8563232	18.36524012	68.755938	0	2.818876982
2000	1	1	10	0	144	681	488	147	721	512	-4	21.87284825	986.9790649	16.31093616	59.67053571	0	2.604524851
2000	1	1	11	0	151	759	608	162	782	633	-2	26.03696308	986.3904419	14.89185464	53.0401449	0	2.763721228

Descriptive statistics

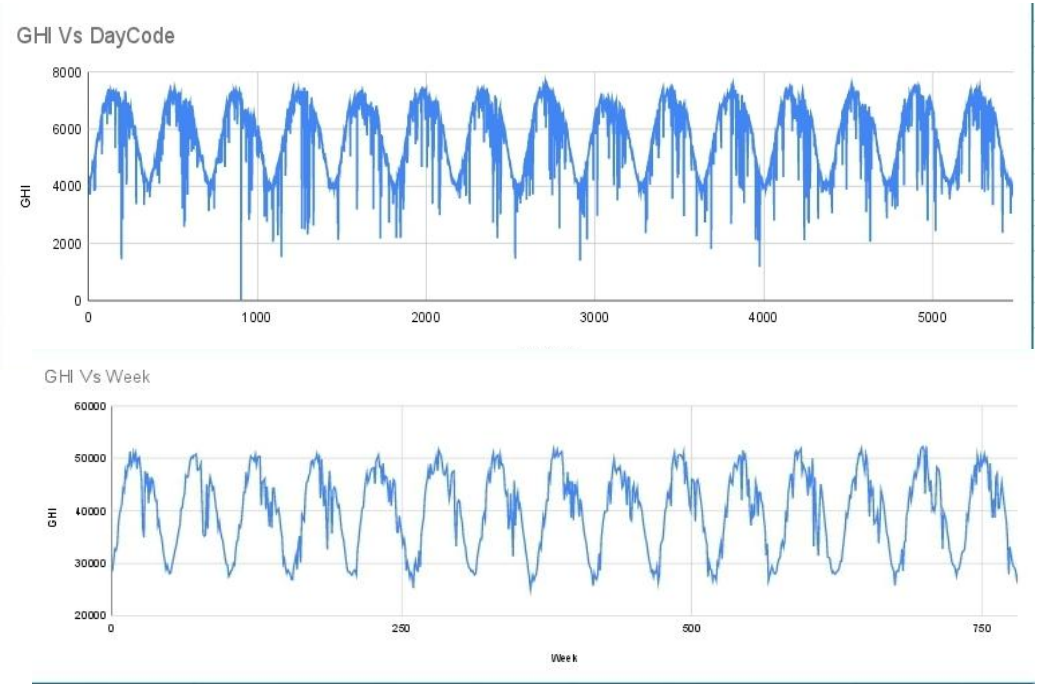
Let's look at the Correlation plots(Heat map) for both the solar park 1:





We can see that GHI has very high correlation with DHI and DNI values.

Seasonality Park 1



Looking at Figures, we can observe that there exists no trend in weekly data large enough to be visible to the eye. It is possible that a very small trend does exist but we will test for the existence of a trend in a later section.

We can make a rough inference that the data is **seasonal** and values of GHI are similar after every gap of one year. However, concrete tests need to be conducted to verify the stationarity (inexistence of trend) of data.

Distribution Plotting

Distribution fitting is the process of identifying a curve that is best fit to the series of data points. Generally used as an aid in visualization. They can also be used to summarize the relationships among two or more variables. We first used the KS test to check if our series could be derived from any of the commonly known distributions and then if successful, we plotted the distribution fits.

Kolmogorov -Smirnov Test

Kolmogorov-Smirnov statistic represents the maximum vertical deviation between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of the theoretical distribution.

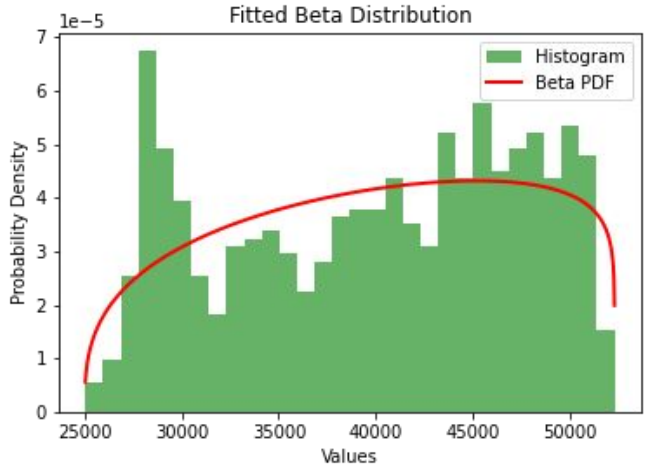
$D=\max |Fn(x)-F(x)|$

This test involves two hypotheses:

Null Hypothesis (Ho): The sample is drawn from the specified distribution.

Alternative Hypothesis (Ha): The sample is not drawn from the specified distribution.

Beta Distribution Plot



Park - 1 Weekly Data

	Weibull Min p-value	Weibull Max p-value	Normal p-value	Gamma p-value	Exponential p-value	Lognormal p-value	Beta p-value
Park-1	1.04E-32	0	2.51E-42	4.42E-45	0	4.43E-45	1.45E-22
Park -2	1.03E-21	0	1.31E-25	3.04E-26	0	5.80E-28	8.58E-17

Results of Daily data from Park-1 and Park-2

	Weibull Min p-value	Weibull Max p-value	Normal p-value	Gamma p-value	Exponential p-value	Lognormal p-value	Beta p-value
Park -1	7.24E-05	0	1.65E-05	8.49E-06	3.60E-27	0	4.37E-02
Park -2	1.04E-03	0	1.28E-03	1.37E-03	1.13E-38	0	8.48E-02

Results of Weekly data from Park-1 and Park-2

At **1%** significance level, we can reject the null hypothesis for daily data, but for the weekly distribution, we can see that beta distribution p-value is greater than 0.01. So we fail to reject Ho and conclude that weekly data will follow beta distribution and daily data won't follow any specified distribution.

Tests For Stationarity

innovate achieve lead

1) Augmented Dickey Fuller Test (ADF)

This test checks for the presence of a unit root. A unit root is a stochastic trend in a time series.

The hypotheses for this test are as follows:

H0: The series has a unit root.

Ha: The series has no unit root.

The results obtained on conducting the ADF test are shown below:

Region	ADF Statistic (Weekly)	p-value	ADF Statistic (Daily)	p-value
Park-1	-9.4898369	3.6779325e-16	-5.4193085	3.0848507e-06
Park-2	-9.0754760	4.1906491e-15	-5.4203636	3.0691081e-06

As the p-values are less than 0.01, we can reject the null hypothesis for each of the regions. But non-stationarity can be caused by other factors too, thus we conduct another test to confirm that our series is stationary.

Tests For Stationarity

innovate achieve lead

2) Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS)

The hypothesis of the KPSS test is:

H0: The series is stationary

Ha: The series is not stationary

The 1% critical value for this test is known to be 0.739. On conducting this test, we get the following results:

Region	KPSS statistic (weekly)	p-value	KPSS statistic (daily)	p-value
Park-1	0.0114219	0.1	0.0213104	0.1
Park-2	0.0188151	0.1	0.0336456	0.1

We can see above that the test statistics are less than the 1% critical value for each of the 2 regions for both daily and weekly data. So, we can reject the null hypothesis for any of the series.

Conclusion about Stationarity

When these tests are applied together, there are four possible cases:

1) Both ADF and KPSS Reject the Null Hypothesis: ADF Rejects the null hypothesis of a unit root (indicating stationarity) and KPSS Rejects the null hypothesis of trend-stationarity. So, we can interpret that the series is stationary around a deterministic trend, indicating the presence of a long-term equilibrium.

2) ADF Rejects, KPSS Does Not Reject: ADF Rejects the null hypothesis of a unit root but KPSS Does not reject the null hypothesis of trend-stationarity.

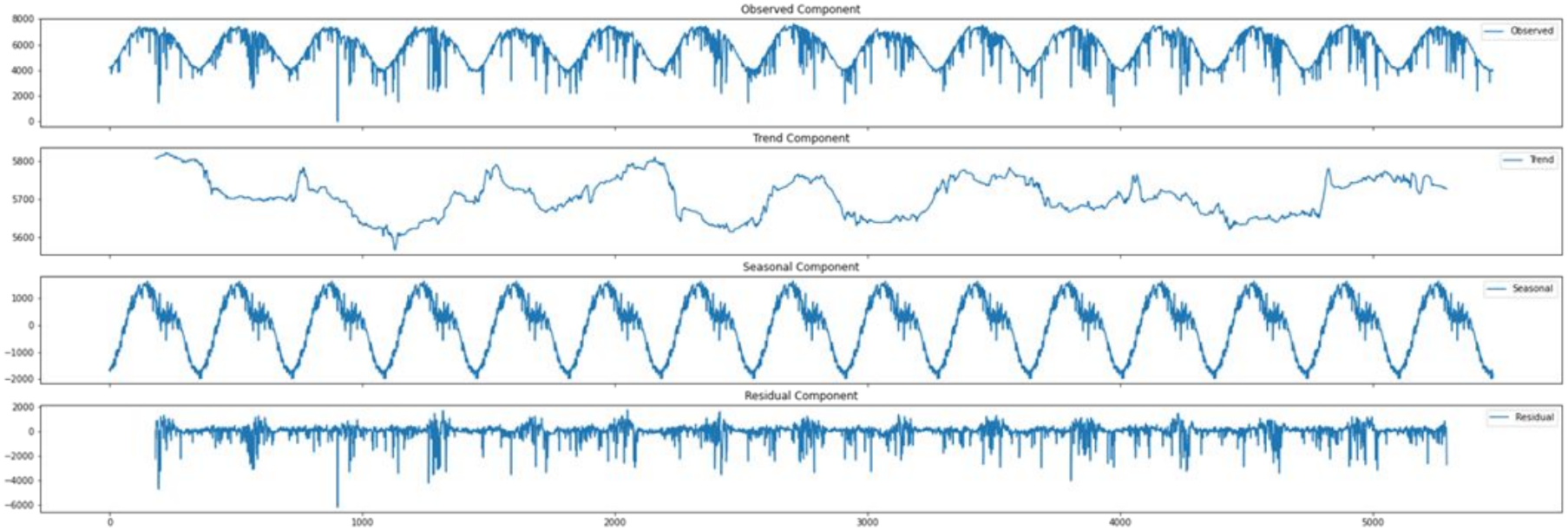
3) ADF Does Not Reject, KPSS Rejects: ADF Does not reject the null hypothesis of a unit root but KPSS Rejects the null hypothesis of trend-stationarity.

4) Both ADF and KPSS Do Not Reject: ADF Does not reject the null hypothesis of a unit root and KPSS Does not reject the null hypothesis of trend-stationarity.

For our data, we can thus conclude that all of the series (weekly and daily for each region) are **stationary** as both the tests are arriving at this conclusion.

Time Series Decomposition

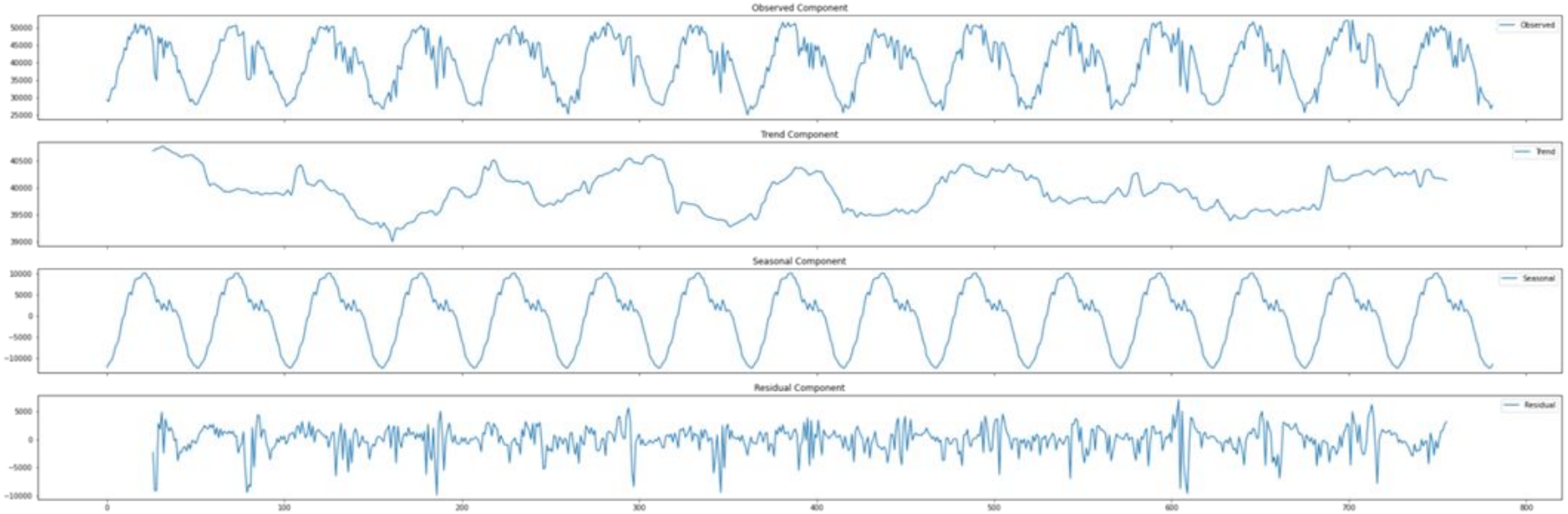
Time series decomposition is the process of breaking down a time series into its underlying components, typically trend, seasonality, and residual, to better understand the patterns and variations within the data.



Time Series Decomposition of Daily Data from Park-1

Time Series Decomposition

From these graph plots for Park-1, it can be seen that there is a seasonality in the daily series and no uniform trend exists. Similar inferences can be drawn about its weekly data.



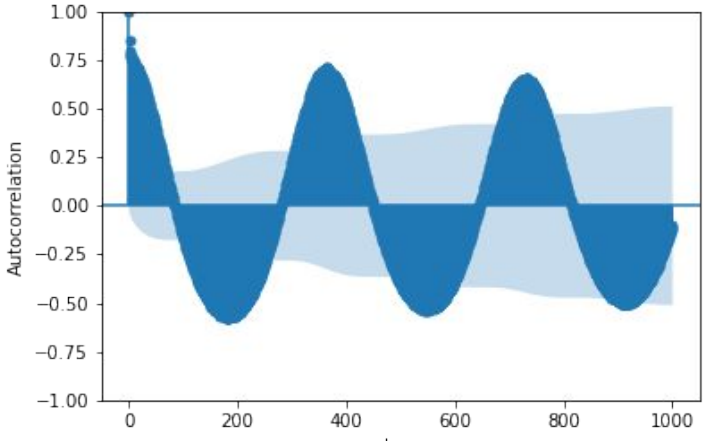
Time Series Decomposition of Weekly Data from Park-1

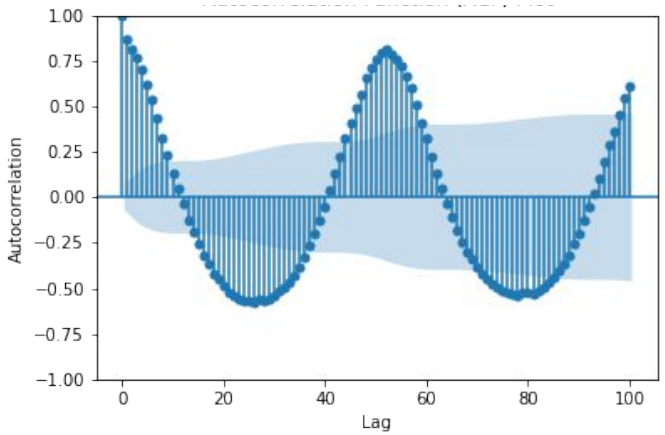
Autocorrelation Function(ACF)

ACF plot is a bar chart of coefficients of correlation between a time series and the lagged values.

It explains how the present value of a given time series is correlated with the past (1-unit past, 2-unit past, ..., n-unit past) values.

Blue spikes on an ACF plot above are the error bands, denotes the maximum significant lag that also represents the upper limit of the hyperparameter q and anything within these bars is not statistically significant.



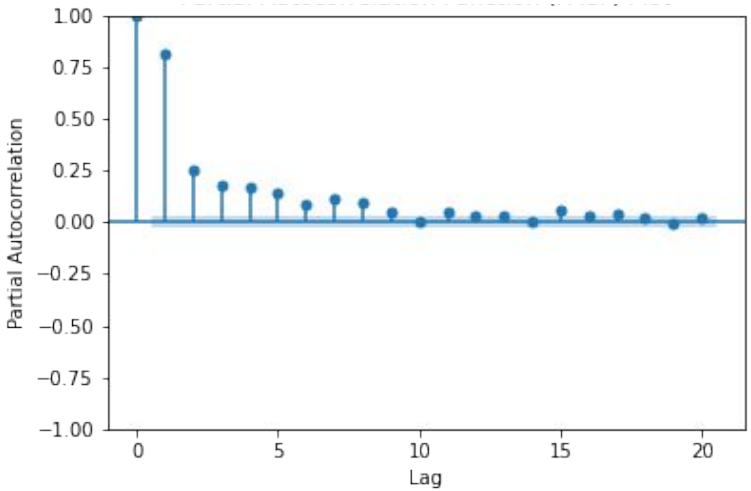


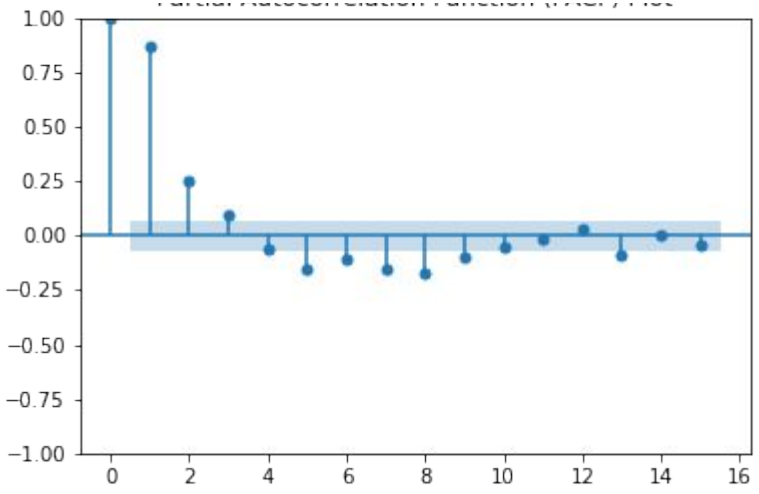
Partial Autocorrelation Function (PACF)

Partial autocorrelation is a statistical measure used to quantify the degree of correlation between a particular observation in a time series and its past values, while accounting for the influence of intermediate observations at other time points.

Unlike total autocorrelation, which measures the overall correlation with any lag, partial autocorrelation isolates and identifies the unique contribution of a specific lag to the correlation structure of the time series.

In essence, it helps to reveal the direct relationship between a given time point and its lagged values, excluding the indirect effects mediated by the intervening observations.





Hyperparameter Evaluation

We have used the ACF plots to determine the maximum possible value of q and PACF plots to determine the maximum possible value of p using the number of significant lags in the plot.

Then, we used grid search to determine the values of p and q that minimize the value of AIC for the models for every scenario, i.e., p , p for AR where q = 0, q for MA where p = 0 and (p, q) for ARMA.

AIC stands for Akaike Information Criteria, and it's a statistical measure that we can use to compare different models for their relative quality. It measures the quality of the model in terms of its goodness-of-fit to the data, its simplicity, and how much it relies on the tuning parameters.

To determine the value of d for ARIMA model, but since our data is stationary, there is no need for differencing and hence, d=0 should give best results.

Auto Regressive Model(AR(p))

- The model forecasts future values based on past values of the data.
- This model is applied when there is a correlation between the time series values and their preceding and succeeding values.

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

- p here represents the no of lagged values.

Let's take an example to breakdown the above equation: Suppose we want to build an model AR(1)

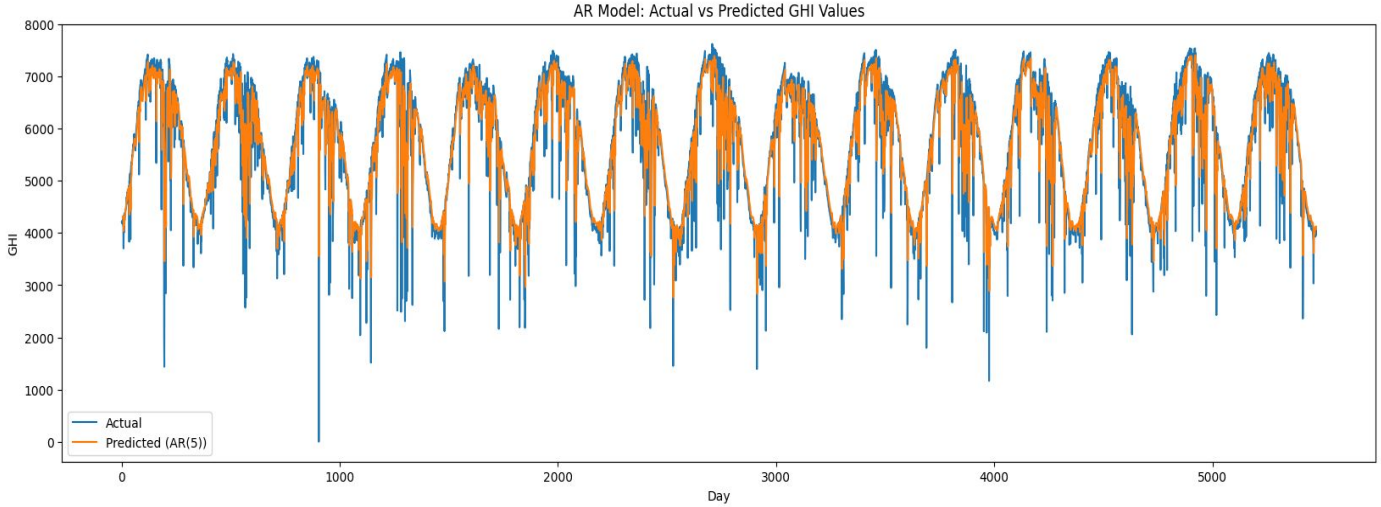
$$Xt = C + \phi 1X(t-1) + \epsilon t$$

X(t-1) represents the previous period's value of x.. say 't' represent today, then 't-1' represents last week's value

The coefficient $\phi 1$ represents the future portion of the previous value. This coefficient is maintained between -1 and 1.

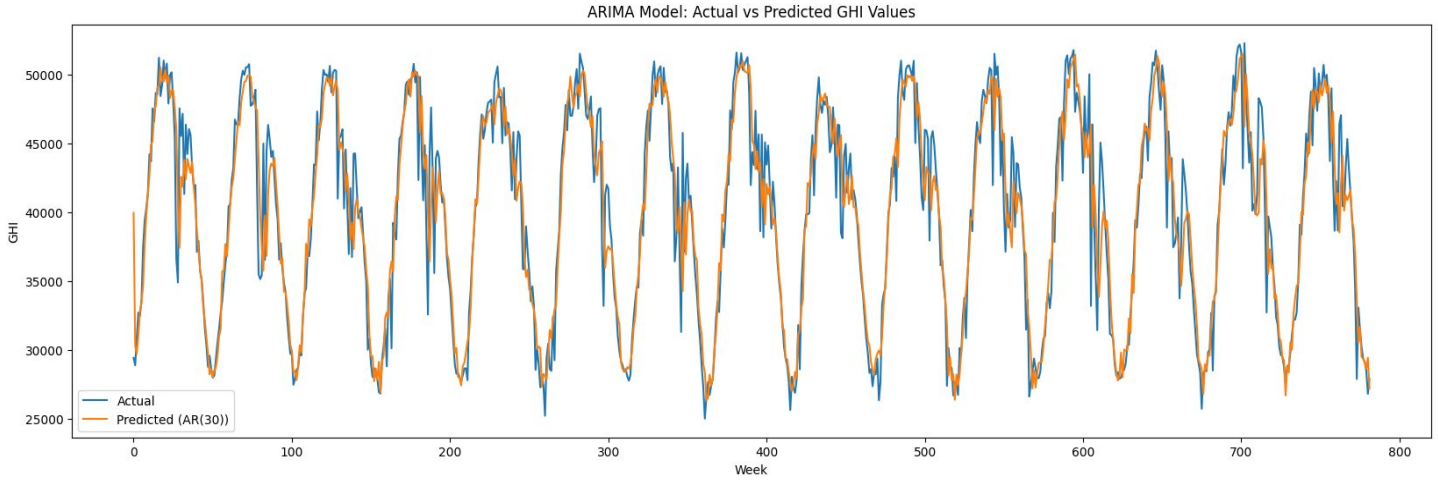
ϵt represents the residual which is the difference between prediction and actual value. ($\epsilon t = yt - \hat{y}t$)

AR model for GHI daily and weekly



This plot here represents the actual vs predicted daily GHI values for one of the region in rajasthan

The MAPE for this model is 6.86%. This represents that the model's prediction is off by 6.86% from the actual observed values.



This plot here represents the actual vs predicted weekly GHI values for one of the region in rajasthan

The MAPE for this comes out to be 4.72%

Moving Average Model(MA(q))

- Rather than taking past values for forecasting, MA model takes past forecast residuals/errors for forecasting.

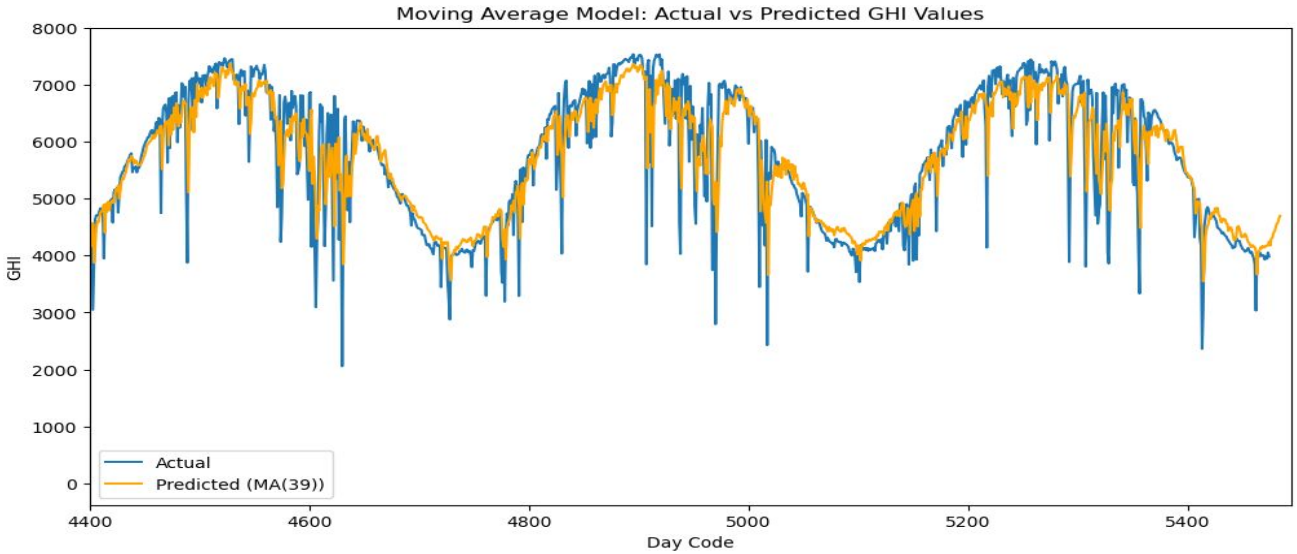
$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

- Let's take an example to breakdown the above equation: Suppose we want to build an model MA(1)

$$X_t = c + \theta_1 \epsilon_{t-1} + \epsilon_t$$

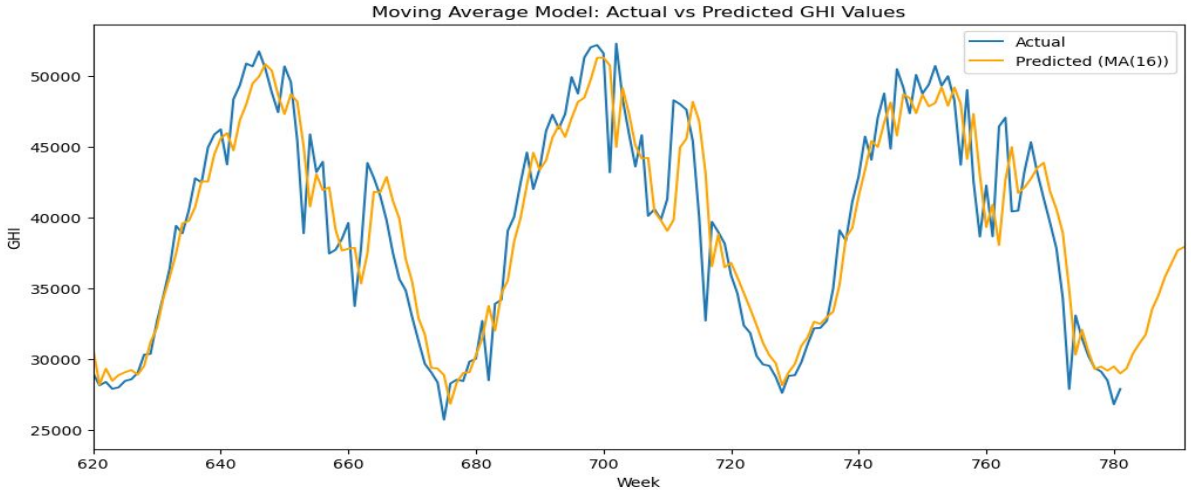
- Xt represents the value we have to forecast say today's value.
- θ1 represents the coefficient of the future portion of the previous residual.(In AR it was previous value). Here also the value of this coefficient is taken to between -1 and 1.
- ϵt and ϵ(t-1) which represent the residuals of the current and the previous period.

MA model for GHI daily and weekly



This plot here represents the actual vs predicted daily GHi values for one of the regions in Rajasthan.

The MAPE for this comes out to be 10.24%



This plot here represents the actual vs predicted weekly GHi values for one of the regions in Rajasthan.

The MAPE for this comes out to be 5.430%

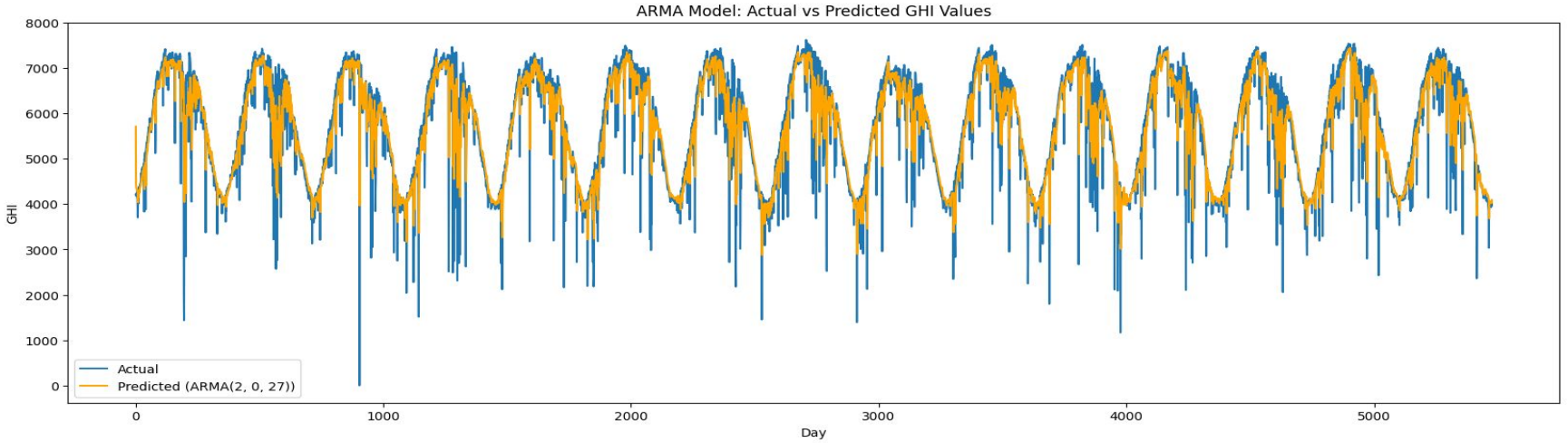
ARMA

AR models capture the relationship between the current value and its past values, assuming that the series exhibits some inertia or memory. MA models capture the impact of past white noise (random) terms on the current value, accounting for short-term fluctuations.

ARMA models allow for the simultaneous consideration of both types of dependencies.

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Region	Hyperparameters (p,q)	MAPE	MAE
Park-1	2, 27	6.88%	354.81



ARIMA

The model is denoted as ARIMA (p, d, q), where p, d, and q are the order parameters for autoregression, differencing, and moving averages, respectively.

It is a time series forecasting technique that combines autoregression (AR), differencing (I), and moving averages (MA).

The integrated component involves differencing to achieve stationarity. Differencing involves computing the difference between consecutive observations to make the series more stationary.

Since our time series was inherently **stationary**, necessitating **no differencing (d=0)**, the ARIMA models were effectively reduced to ARMA models.

SARIMA

Seasonal AutoRegressive Integrated Moving Average (SARIMA) is a time series forecasting method that extends the ARIMA (AutoRegressive Integrated Moving Average) model to incorporate seasonality.

SARIMA model is particularly useful for time series data that exhibit seasonality. It is denoted as **SARIMA (p, d, q) (P, D, Q, s)**

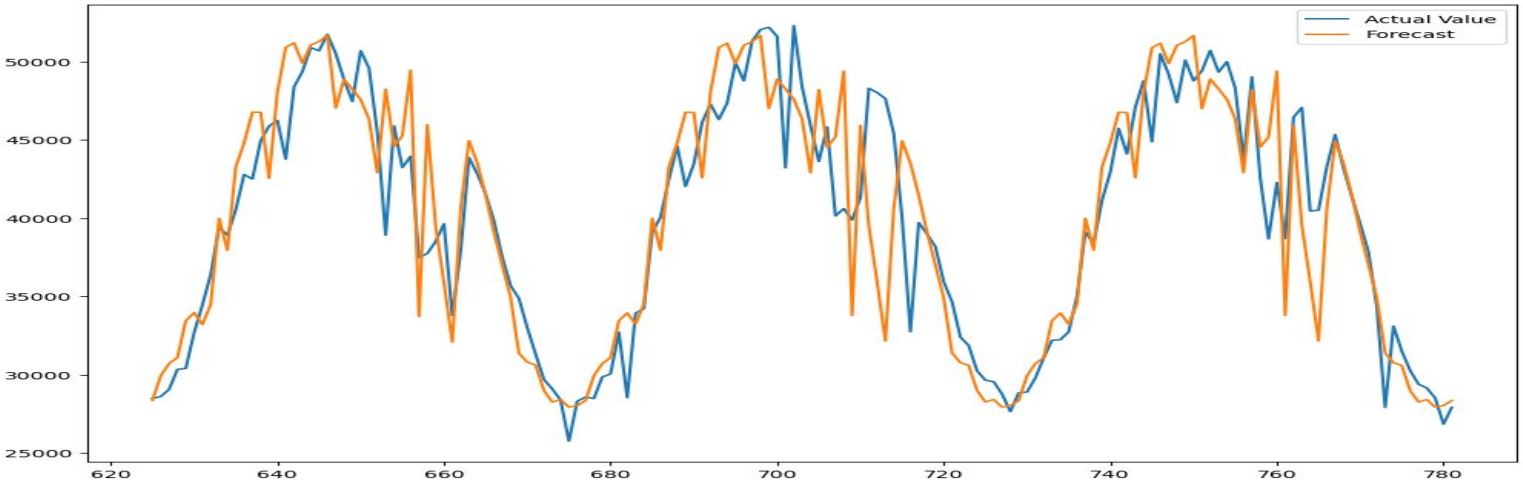
Parameters:

- **s (Seasonal Periodicity):** This parameter defines the length of the seasonal cycle. For example, if the data has a yearly seasonality, s would be set to 12 for monthly data or 4 for quarterly data.
- **P (Seasonal AutoRegressive Order):** This parameter is similar to the autoregressive order but applies to the seasonal component of the time series. It represents the number of seasonal lag observations included in the model.
- **D (Seasonal Integrated Order):** Like the integrated order for the non-seasonal component, D represents the number of differences needed to make the seasonal component stationary. A value of 1 means a first-order seasonal difference, and so on.
- **Q (Seasonal Moving Average Order):** This parameter is similar to the non-seasonal moving average order but applies to the seasonal component. It represents the number of seasonal lagged forecast errors included in the model.

In SARIMA, the seasonal component helps capture patterns that repeat over a fixed period. The non-seasonal components (p, d, q) address trends and other patterns within each season, while the seasonal components (P, D, Q, s) address the seasonal variations.

SARIMA model results for weekly data

Region	Hyperparameters (p,d,q) (P, D, Q, m)	MAPE	MAE
Park-1	(1,0,1) (1,1,1,52)	5.927%	2397.875



PARK 1

Practical Model

Refer to the link -

https://drive.google.com/drive/folders/1iVBR_EuuGmHHJ4CxL0xG8A4gEDu-CD82?usp=drive_link

THANK YOU