A Report
on

Statistical Analysis and Forecasting of Solar Energy for Study Regions in Rajasthan

by
RAO GROUP

of
Applied Statistical Methods

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
PILANI, PILANI CAMPUS
November 2023

Contents

1. 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 purpose of this report is to conduct a comprehensive analysis of solar energy data for two study regions in the state of Rajasthan. The primary objectives include exploring renewable energy resources, obtaining and understanding solar park data, identifying relevant parameters, and forecasting solar energy for future planning.

1.1 Renewable Energy Resources

- **Overview of Renewable Energy in India**

India, with its commitment to sustainable development, has been actively harnessing renewable energy. The Ministry of New and Renewable Energy (MNRE) plays a crucial role in promoting and regulating renewable energy sources. Solar energy, derived from the sun's radiation, stands out as a key focus due to its abundance and potential for widespread application.

- **Importance of Solar Energy**

Solar energy, captured through photovoltaic cells or solar thermal systems, provides clean and inexhaustible power. As a decentralized and scalable source, solar energy can contribute significantly to meeting the energy demands of both urban and rural areas.

1.2 Forecasting

The solar power output in solar plants is dependent on various uncontrollable variables which affect the amount of sunlight falling on the solar panels. Short-term forecasts are valuable for operators in order to make decisions of grid operation, as well as, for electric market operators to make decisions related to supply and demand. Long-term forecasts are useful for energy producers and to negotiate contracts with financial entities or utilities that distribute the generated energy. Thus, accurate forecasting is required so that the resources can be utilized in a way that generates higher power output.

1.3 Data Collection

Solar Park Data from Rajasthan

The provided dataset from 2000 to 2014 includes hourly information from two solar parks in Rajasthan. These solar parks serve as representative sources for studying solar energy patterns in the region. The choice of time span allows for a robust analysis of historical trends and patterns.

1.4 Terms associated with solar power

- **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.

2. Pre - Processing

2.1 Dataset

The given dataset contains hourly information collected over a period of 15 years (2000-2014) at 2 regions in 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

3. Descriptive Statistics

3.1 Correlation

We obtained the feature correlation maps for each region. The plot for park1 and park 2 are shown below:

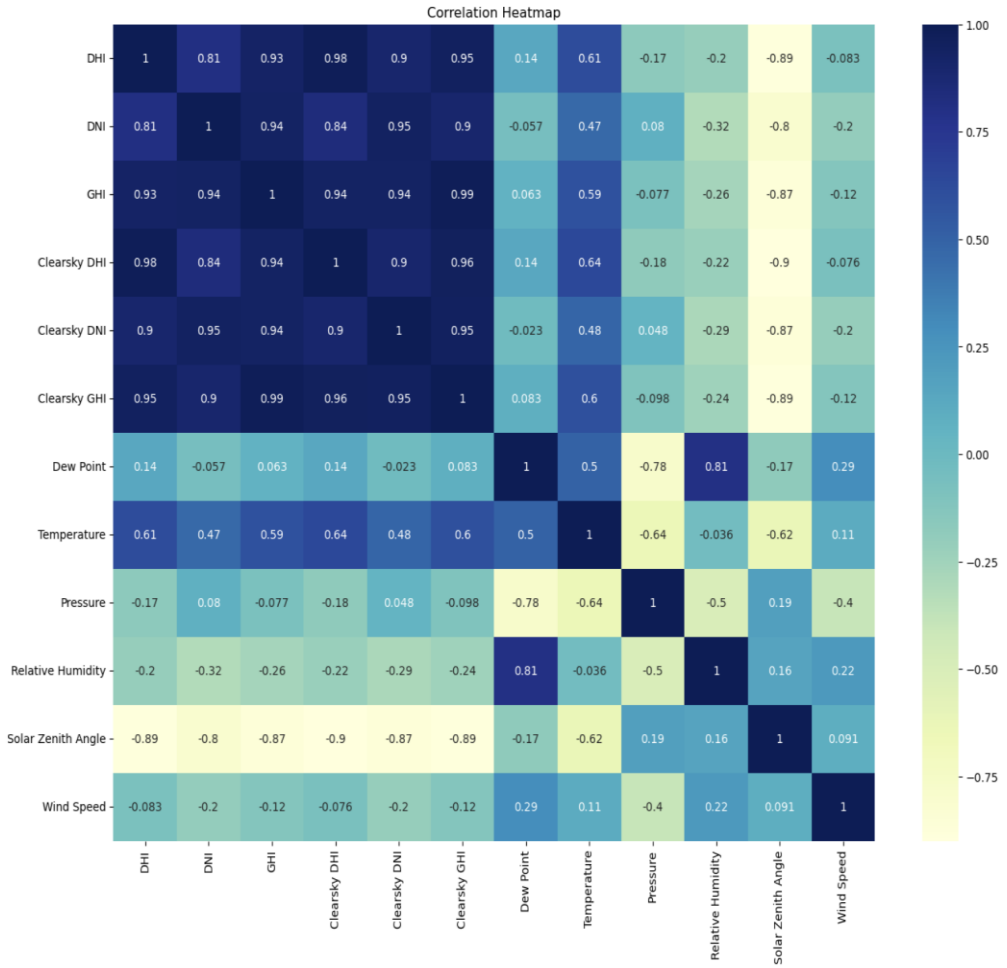


Figure 1(a): Correlation Plot of Park-1

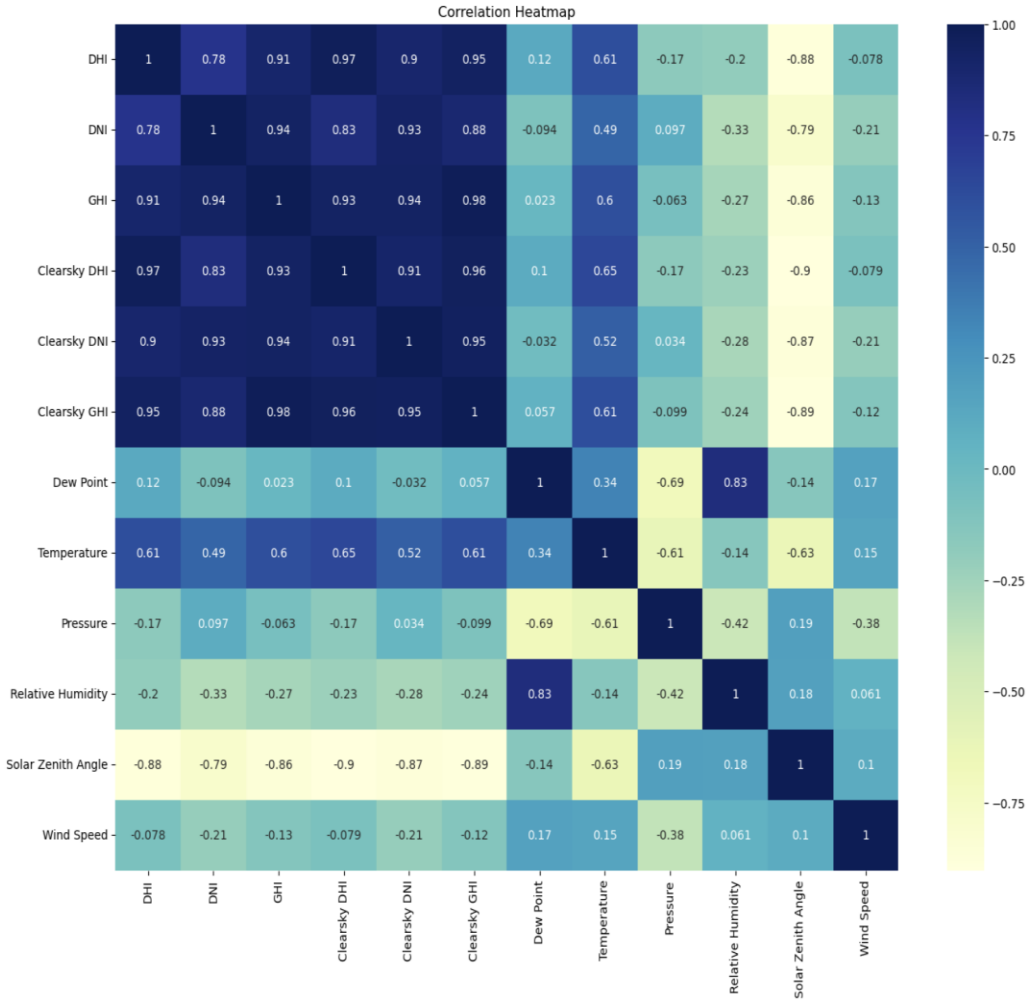


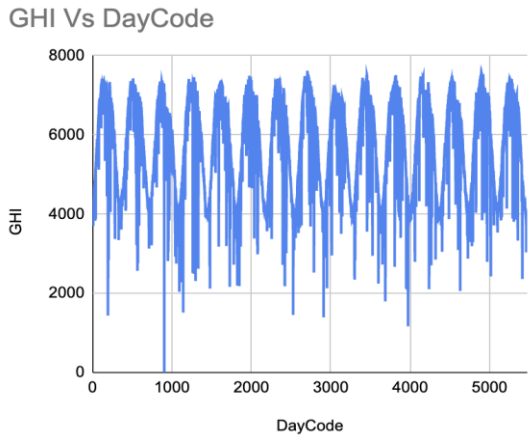
Figure 1(b): Correlation Plot of Park-2

It can be observed that GHI has a highly positive correlation with the DNI and DHI values. This is clear from the relation between these variables mentioned before. GHI is negatively correlated with Zenith Angle which is clear as cosine is a decreasing function in the first quadrant. A moderately positive correlation with temperature is explainable as higher temperatures, to a certain extent, are likely to be caused by higher amounts of solar radiation and thus in turn lead to higher GHI values. The GHI values are made up of both DNI and DHI values and are a good measure relating to power output.

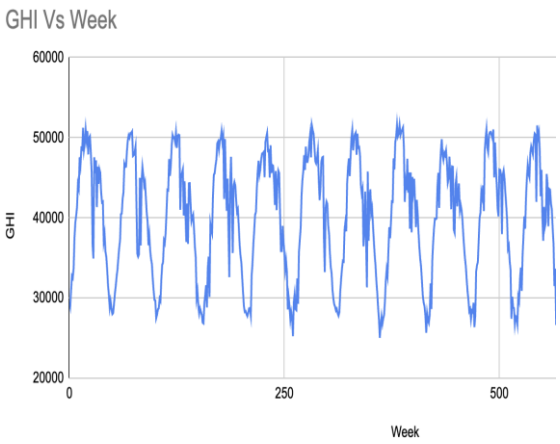
Hence, GHI forecasts can be useful for forecasting solar power output.

3.2 Plotting the Data

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. It is however very clear that some kind of seasonality is in play. It looks like the GHI values at instants of time separated by approximately 52 weeks are very close. As was the case for weekly data, we cannot observe any significant trend from the plot for daily data shown in Figure. We can also observe the existence of seasonality in this figure.

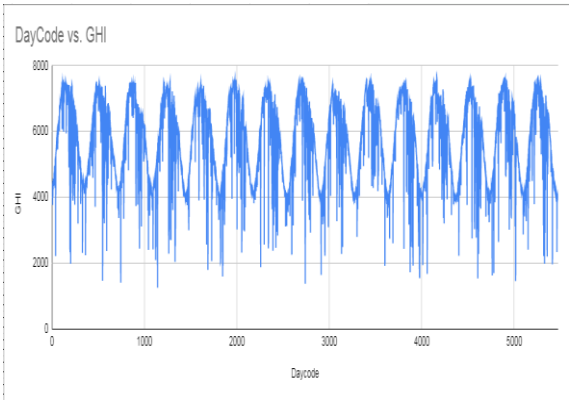


(a) Plot of Daily Data from Park-1

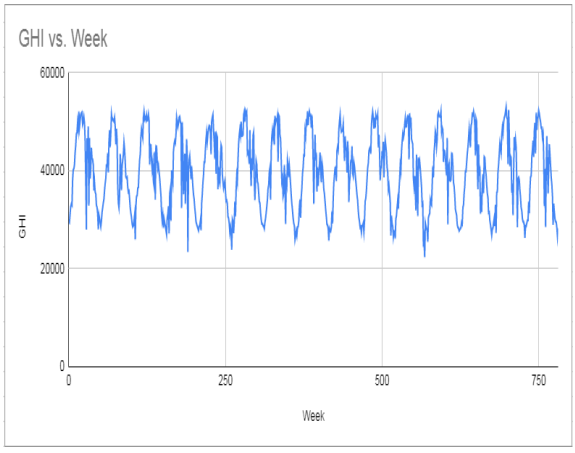


(b) Plot of Weekly Data from Park-1

Figure 2: Park-1 Data Plots



(a) Plot of Daily Data from Park-2



(b) Plot of Weekly Data from Park-2

Figure 2(b): Park-2 Data Plots

From both of the plots, 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.

3.3 Distribution Fitting

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.

3.3.1 Kolmogorov-Smirnov Test

The test is based on the Kolmogorov-Smirnov statistic (D), which represents the maximum vertical deviation between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of the theoretical distribution. $D=\max |Fn(x)-F(x)|$

Here, Fn(x) is the empirical distribution function of the sample, and F(x) is the cumulative distribution function of the theoretical distribution.

The test involves two hypotheses:
Null Hypothesis (H_0): The sample is drawn from the specified distribution.
Alternative Hypothesis (H_1): The sample is not drawn from the specified distribution.

The results obtained are shown here:

	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

3.3.2 Distribution Fit Plot

On seeing the above results, we can clearly say that for KS-test on daily data, as the p-values are smaller than 0.01; at 1% significance level, we can reject the null hypothesis. Thus, we can say that the daily data from each Region is not derived from any of the above listed distributions. On the other hand, KS-test for weekly data shows that the p-values for weekly data from each region tested against the beta-distribution is greater than 0.01. Thus, we cannot reject the null hypothesis. Fig 3(a) and Fig3(b) shows the best fit beta-distribution plot for weekly data from Park-1 and Park-2.

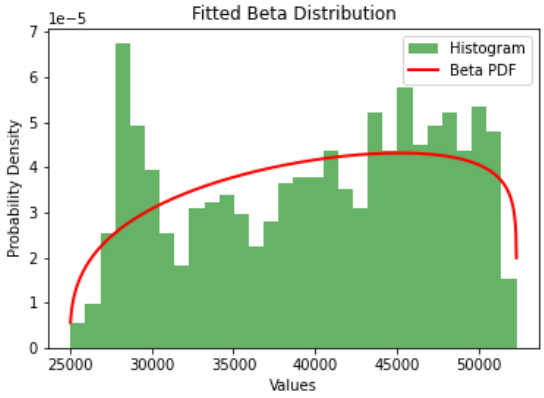


Fig 3(a)

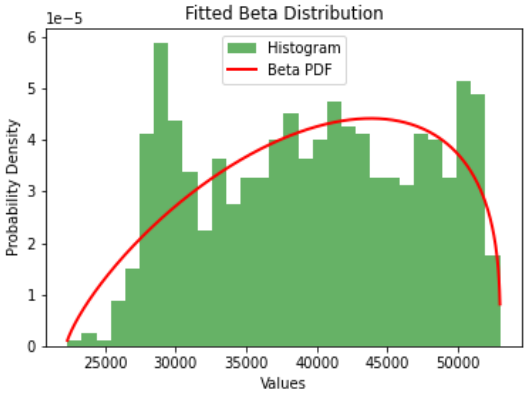


Fig 3(b)

4. Tests for Stationarity

Stationarity in statistics means that the statistical properties of time series like mean, variance and covariance do not vary with time. Normally two tests are used to check a time series for stationarity:

• Augmented Dickey Fuller (ADF)
• Kwiatkowski-Phillips-Schmidt-Shin (KPSS)

4.1 Augmented Dickey Fuller Test (ADF)

One of the most common causes of non-stationarity are unit-roots. A unit root is a stochastic trend in a time series. [Unit root mathematics is quite complex to be mentioned here.]

The ADF test checks for the presence of a unit root. 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	Weekly		Daily	
	ADF Statistic	p-value	ADF statistic	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 (for both daily and weekly data). Thus, it is likely that the series data (from all regions) don't have a unit root. But non-stationarity can be caused by other factors too, thus we conduct another test to confirm that our series is stationary.

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

The hypotheses of the KPSS test are:

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	Weekly		Daily	
	KPSS Statistic	p-value	KPSS statistic	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 cannot reject the null hypothesis for any of the series.

4.3 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. **Interpretation:** 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. **Interpretation:** The series is stationary without a deterministic trend, which is desirable for many time-series analyses.

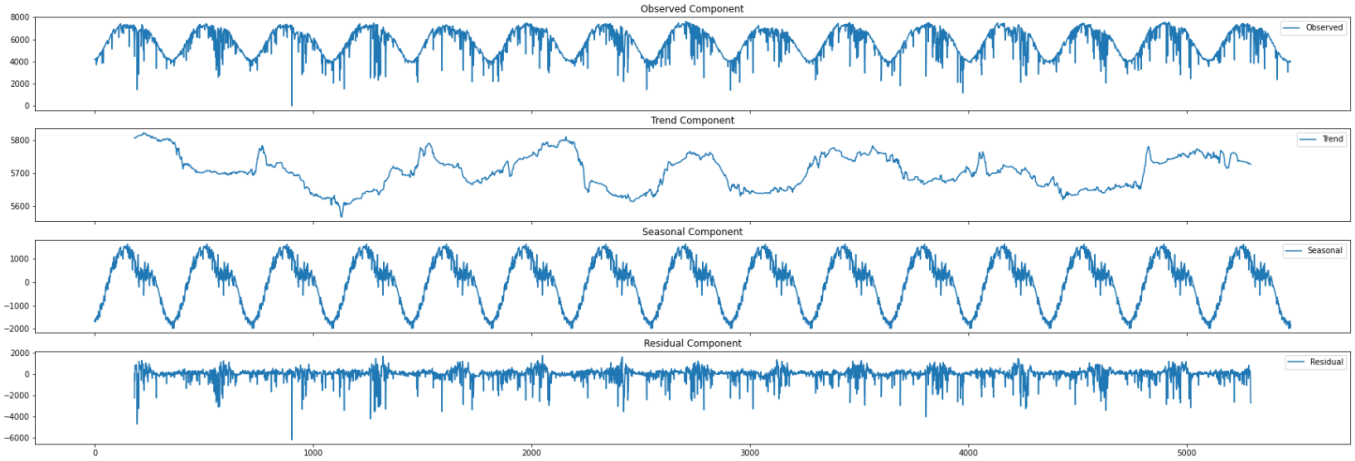
3) **ADF Does Not Reject, KPSS Rejects:** ADF Does not reject the null hypothesis of a unit root but KPSS Rejects the null hypothesis of trend-stationarity. **Interpretation:** The series is non-stationary but has a deterministic trend indicating a need for differencing or detrending.

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. **Interpretation:** The series is non-stationary without a deterministic trend indicating a need for further analysis or transformations.

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.

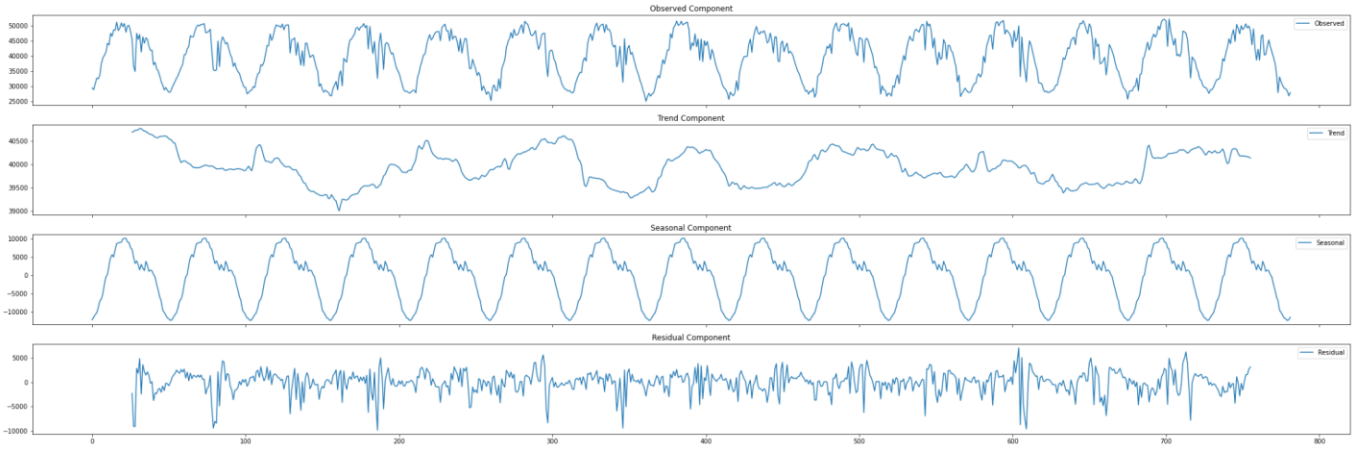
5. Time Series Decomposition

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 is shown below for both weekly and daily data for both regions Park-1 and Park-2.

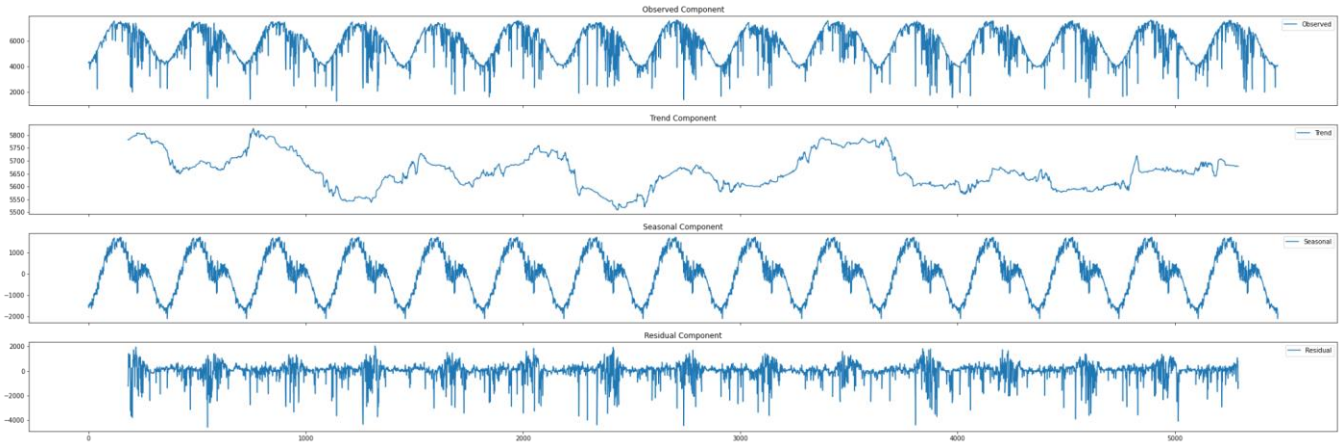


Time Series Decomposition of Daily Data from Park-1

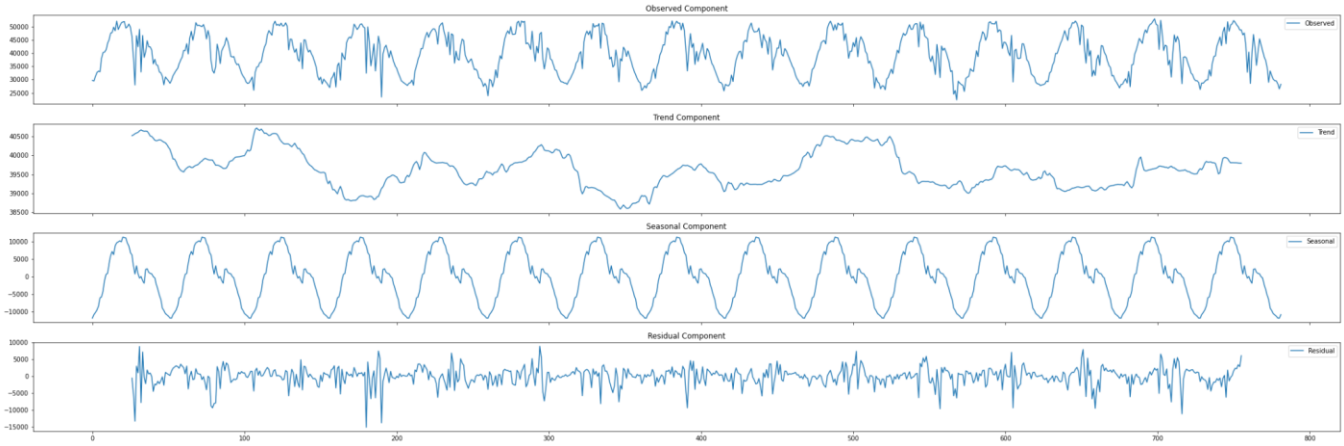
From the above figure for Park-1, it can be seen that there is a seasonality in the daily series and no uniform trend exists. Similar inferences can be drawn about its weekly data.



Time Series Decomposition of Weekly Data from Park-1



Time Series Decomposition of Daily Data from Park-2



Time Series Decomposition of Weekly Data from Park-2

6. Time Series Forecasting

6.1 Important Concepts

6.1.1 Autocorrelation Function (ACF)

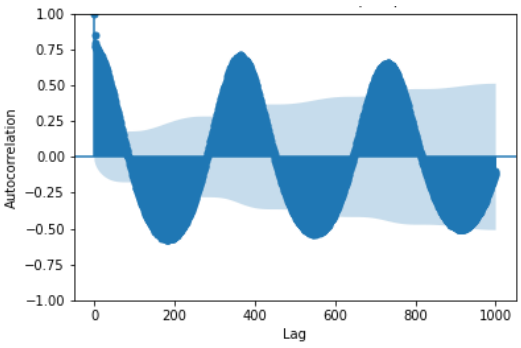
The autocorrelation function (ACF) is a fundamental tool in time series analysis that helps in understanding the temporal dependencies within a sequence of observations. It quantifies the correlation between a time series and its past values at various lags. For a time series X(t), the autocorrelation at lag k, denoted as ACF(k), is computed as the correlation coefficient between the series at time t and the series at time t−k. Mathematically, it is defined as:

$$\frac{Cov(X(t), X(t-k))}{\sqrt{((Var(X(t))(Var(X(t-k)))}}$$

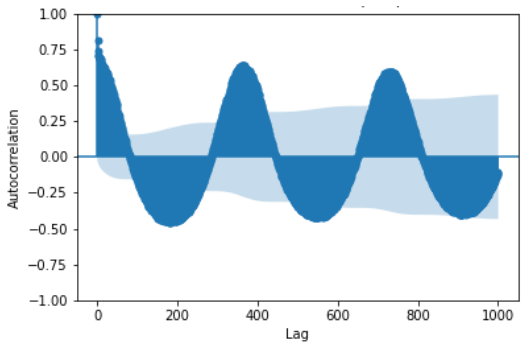
where: Cov(X(t), X(t−k)) is the covariance between X(t) and X(t−k), Var(X(t)) and Var(X(t−k)) are the variances of X(t) and X(t−k) , respectively. If ACF(k) is close to 1, it indicates a strong positive autocorrelation, suggesting that values at time t are positively correlated with values at time t−k. If ACF(k) is close to -1, it indicates a strong negative autocorrelation, suggesting that values at time t are negatively correlated with values at time t−k. If ACF(k) is close to 0, it suggests a weak or no autocorrelation at lag k. The ACF is often visualized using an ACF plot, where the x-axis represents the lags, and the y-axis represents the autocorrelation coefficients. Significant spikes or patterns in the ACF plot can indicate the presence of underlying structures, such as seasonality or trends. ACF plots for daily and weekly data for both regions are shown below.

The maximum significant lags came out to be around 1130 in the respective ACF plots for all the daily GHI value datasets.

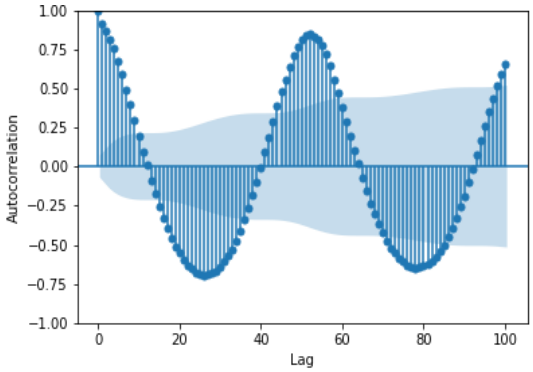
The maximum significant lags came to be around 180 in the respective ACF plots for all the weekly GHI value datasets.



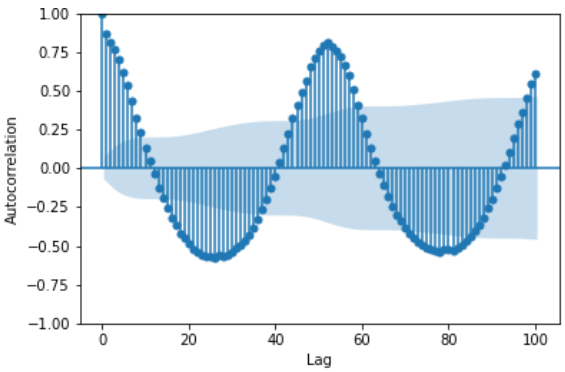
Park-1 Daily ACF Plot



Park-2 Daily ACF Plot



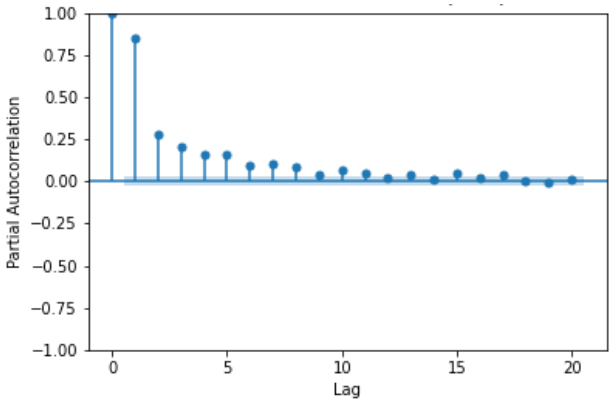
Park-1 Weekly ACF Plot



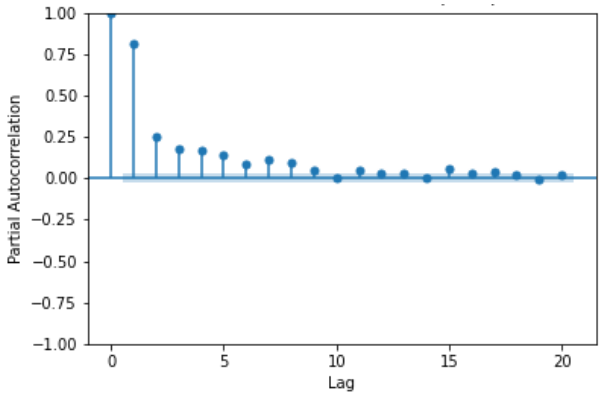
Park-2 Weekly ACF Plot

6.1.2 Partial Autocorrelation Function (PACF)

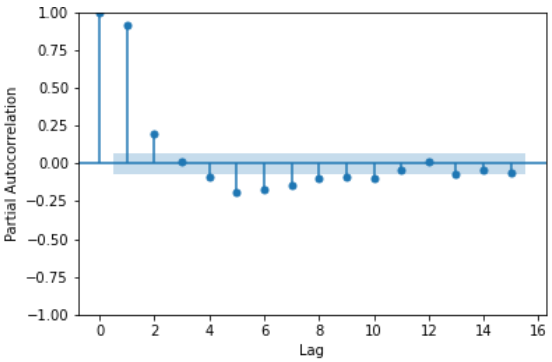
The Partial Autocorrelation Function (PACF) elucidates the partial correlation between a time series and its lagged values. To illustrate, envision a linear regression predicting y(t) from y(t-1), y(t-2), and y(t-3). PACF, in this context, examines the correlation between the unaccounted parts of y(t) and y(t-3), not predicted by y(t-1) and y(t-2). For a time series z(t), denoting the partial autocorrelation at lag k as α(k), it represents the autocorrelation between z(t) and z(t+k), excluding the linear dependence on z(t+1) through z(t+k-1). Mathematically, for k > 1, α(k) is the correlation between z(t) and z(t+k) not explained by lags 1 through k-1, involving a projection operation. The PACF plots for both weekly and daily datasets of GHI values in both regions are provided below. The identified maximum significant lags in the respective PACF plots for daily GHI datasets were around 20. Similarly, for weekly GHI datasets, the maximum significant lags were approximately 10.



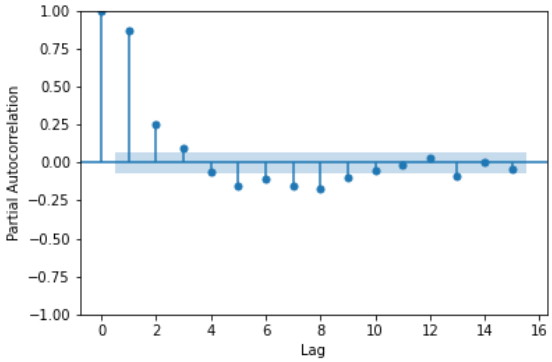
Park-1 Daily PACF Plot



Park-2 Daily PACF Plot



Park-1 Weekly PACF Plot



Park -2 Weekly PACF Plot

6.1.3 Grid Search

Grid Search is a technique designed to identify optimal values for trend-related hyperparameters (p, d, q) and seasonal hyperparameters (P, D, Q, m) in time series models, including AR(p), MA(q), ARMA (p, q), ARIMA (p, d, q), and SARIMA (p, d, q) (P, D, Q, m). The goal is to minimize losses by leveraging the principle of minimizing Akaike Information Criterion (AIC) values for each model. This method involves systematically evaluating various combinations of hyperparameter values to find the set that yields the lowest AIC, indicating the most suitable configuration. Typically, it covers a range of hyperparameter values for p and q, initially derived from PACF and ACF plots, respectively. The chosen model is the one associated with the lowest AIC values, signifying the optimal hyperparameters.

6.2 Step 1: Data Pre-processing

The daily data series and weekly data series were split to create training datasets and testing datasets, respectively. The final errors are computed for the test data, respectively. After this pre-processing step, we had 4 datasets (2 daily and 2 weekly) that we split into training and test data. The number of data points in each of them are**:**

	Daily Dataset for all regions	Weekly Dataset for all regions
Number of Training Data Points	4378	625
Number of Testing Data Points	1091	156

6.3 Step 2: Hyperparameter Evaluation for each model

We conducted hyperparameter optimization for our time series forecasting models through various methods. ACF and PACF plots were employed to identify significant lags, determining the maximum values of q and p, respectively. Subsequently, grid search was utilized to find the values of p and q that minimize the Akaike Information Criterion (AIC) for different scenarios: p for autoregressive (AR) models with q=0, q for moving average (MA) models with p=0, and (p, q) for autoregressive moving average (ARMA) models where both p and q could vary.

In the case of the Autoregressive Integrated Moving Average (ARIMA) model, as our data was stationary, the differencing parameter d was deemed unnecessary, supporting our choice of d=0. However, due to the computational complexity of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, we encountered challenges in performing hyperparameter optimization for this model. Consequently, the process was not completed.

6.4 Step 3: Models for forecasting

6.4.1 Autoregressive (AR) models

An autoregressive (AR) model forecasts future behaviour based on past behaviour data. This type of analysis is used when there is a correlation between the time series values and their preceding and succeeding values. Autoregressive modelling uses only past data to predict future behavior. Linear regression is carried out on the data from the current series based on one or more past values of the same series. AR models are linear regression models where the outcome variable (Y) at some point of time is directly related to the predictor variable (X). In AR models, Y depends on X and previous values for Y, which is different from simple linear regression. The model is mathematically represented by the equation:

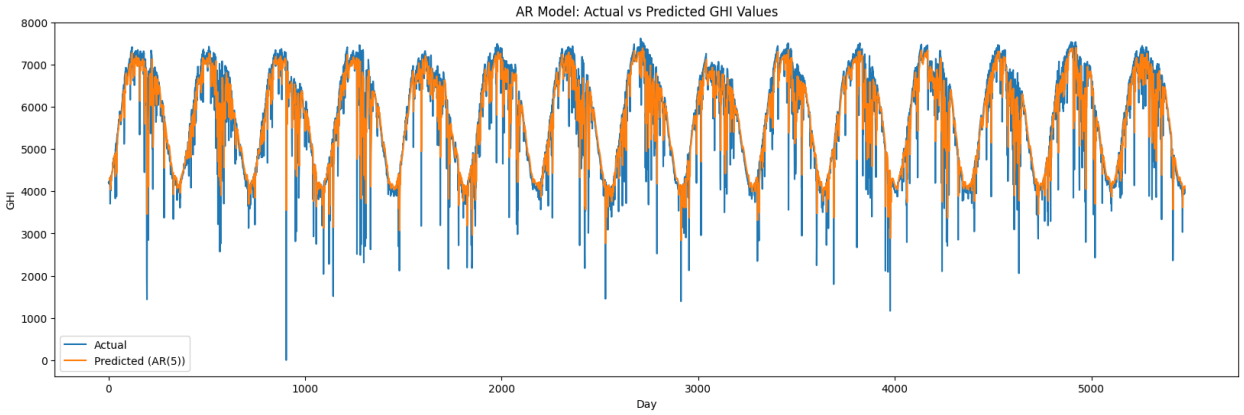
$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

Here, ϕ1 ,…,ϕp represent the model parameters, c is a constant, and e(t) denotes white noise.

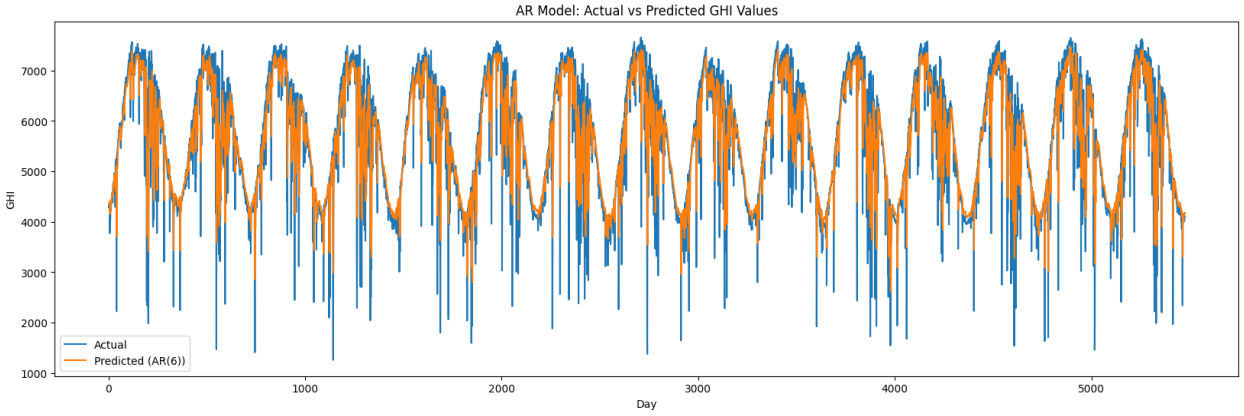
The AR graphs along with tables for both daily and weekly data for both regions are shown below:

AR model results for daily data

Region	Hyperparameters (p)	MAPE	MAE
Park-1	5	6.86%	1008.33
Park -2	6	8.62%	1068.43



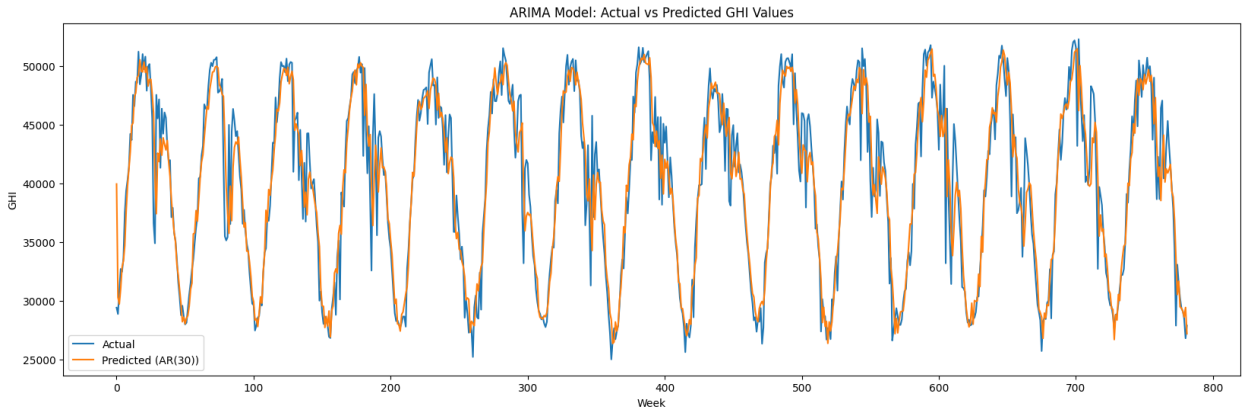
Park-1 Daily AR Plot



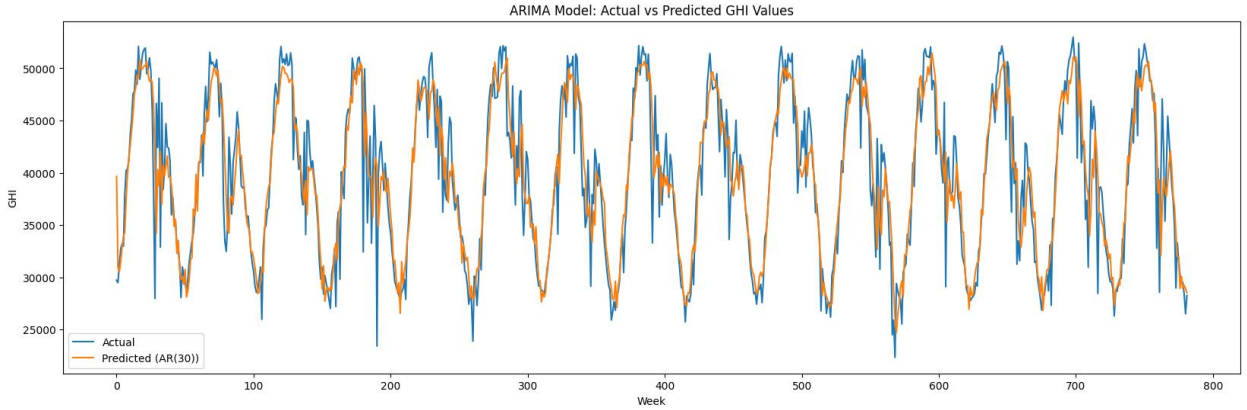
Park-2 Daily AR Plot

AR model results for weekly data

Region	Hyperparameters (p)	MAPE	MAE
Park-1	30	4.72%	1864.52
Park-2	30	5.95%	2273.54



Park-1 Weekly AR Plot



Park-2 Weekly AR Plot

6.4.2 Moving Average (MA) models

In time series analysis, the moving-average model (MA model), also known as moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable is cross-correlated with a non-identical to itself random-variable. Together with the autoregressive (AR) model, the moving-average model is a special case and key component of the more general ARMA and ARIMA models of time series.

The equation describing this model is:

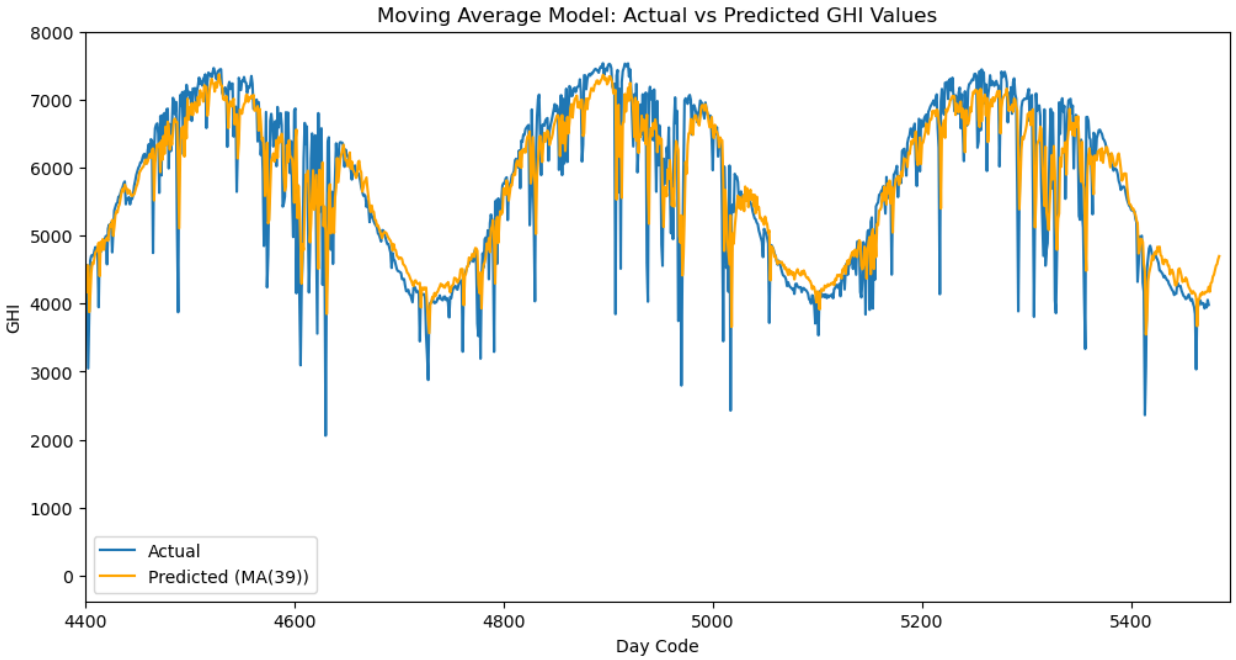
$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

where μ is the mean of the series, θ(1),...,θ(q) are the parameters of the model and the e(t), e(t−1), ..., e(t−q) are white noise error terms. The value of q is called the order of the MA model. Thus, a moving-average model is conceptually a linear regression of the current value of the series against current and previous (observed) white noise error terms or random shocks. The random shocks at each point are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with location at zero and constant scale.

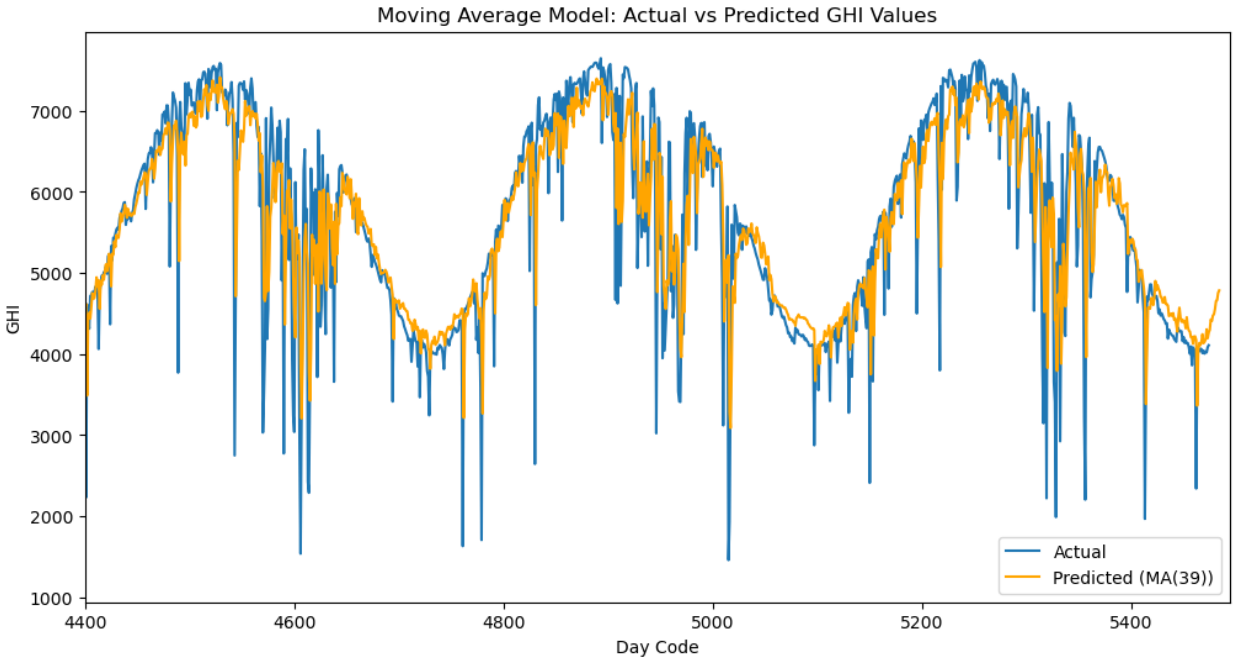
The MA graphs along with tables for both daily and weekly data for both the regions are shown below:

MA model results for daily data

Region	Hyperparameters (q)	MAPE	MAE
Park-1	37	10.24%	355.21
Park-2	39	8.91%	415.49



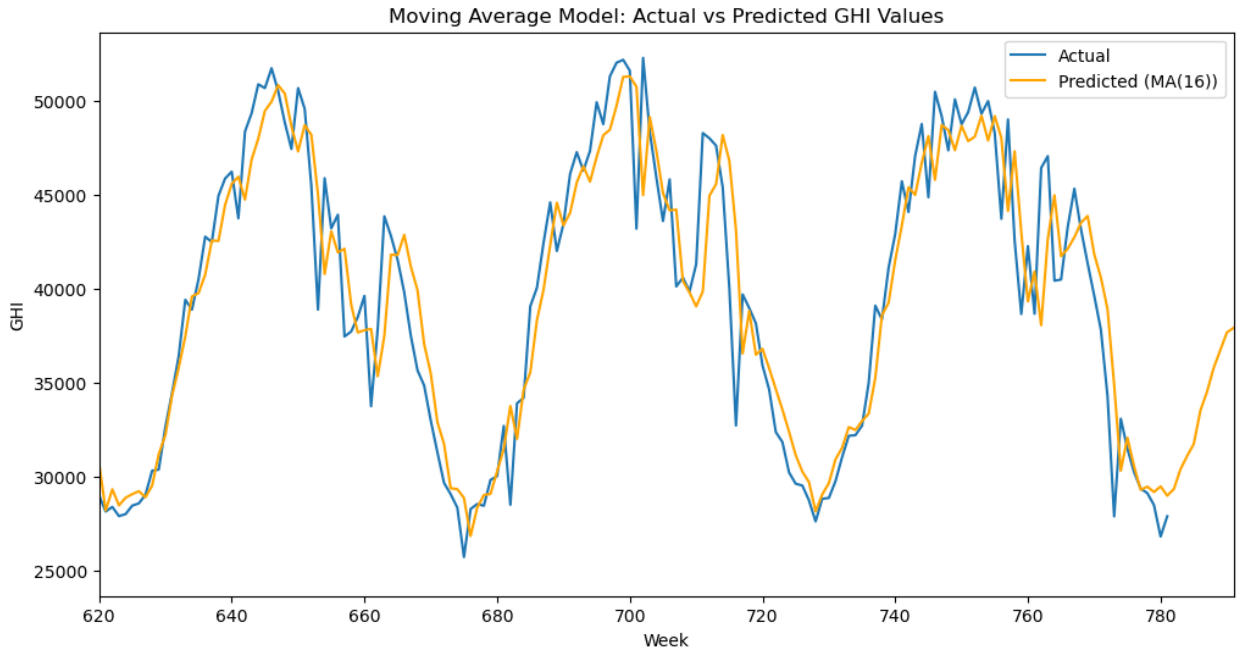
Park-1 Daily MA Plot



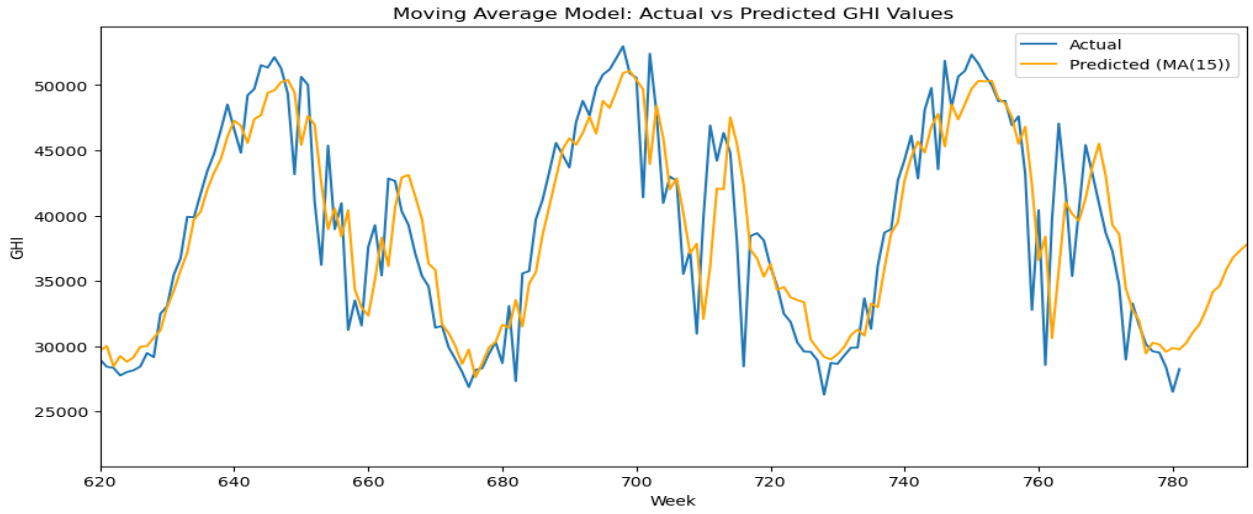
Park-2 Daily MA Plot

MA model results for weekly data

Region	Hyperparameters (q)	MAPE	MAE
Park-1	16	5.430%	2132.42
Park-2	15	6.720%	2560.17



Park-1 Weekly MA Plot



Park-2 Weekly MA Plot

6.4.3 Autoregressive Moving Average (ARMA) models

The Autoregressive Moving Average model combines the above two approaches to generate a model that can describe a weakly stationary time series in terms of two polynomials one with p autoregressive terms and the other with q moving average terms. The equation describing this model is:

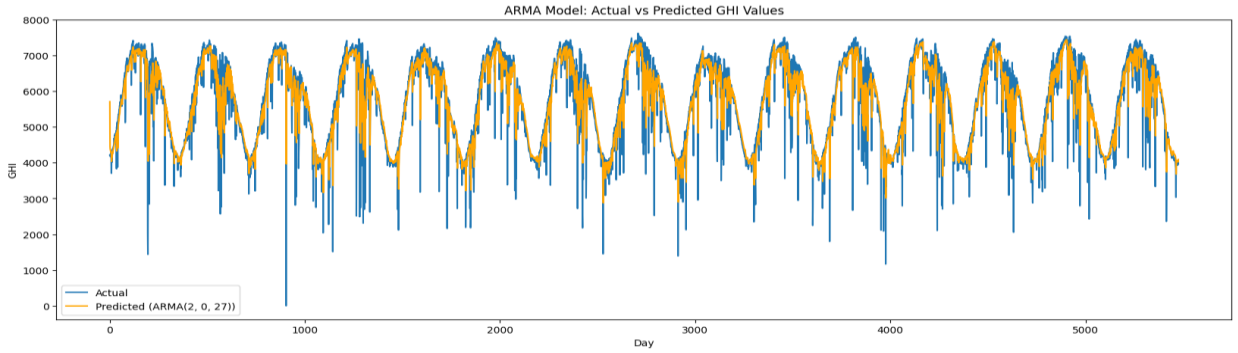
$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where $\varphi_1,\ldots,\varphi_p$ are the coefficients of the autoregressive polynomial, c is a constant, $\theta_1,\ldots,\theta_q$ are the coefficients of the moving average polynomial and they are white noise error terms.

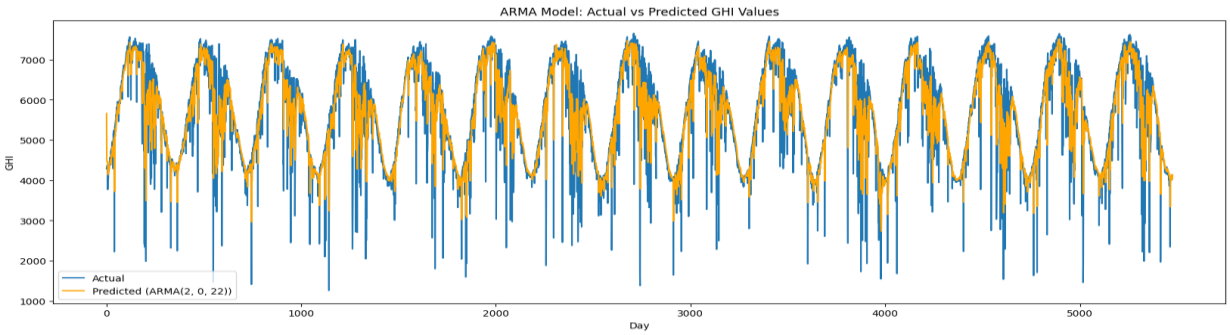
The results of ARMA models for both weekly and daily data for both the regions with tuned hyperparameters are shown below:

ARMA model results for daily data

Region	Hyperparameters (p,q)	MAPE	MAE
Park-1	2, 27	6.88%	354.81
Park-2	2, 22	8.82%	397.87



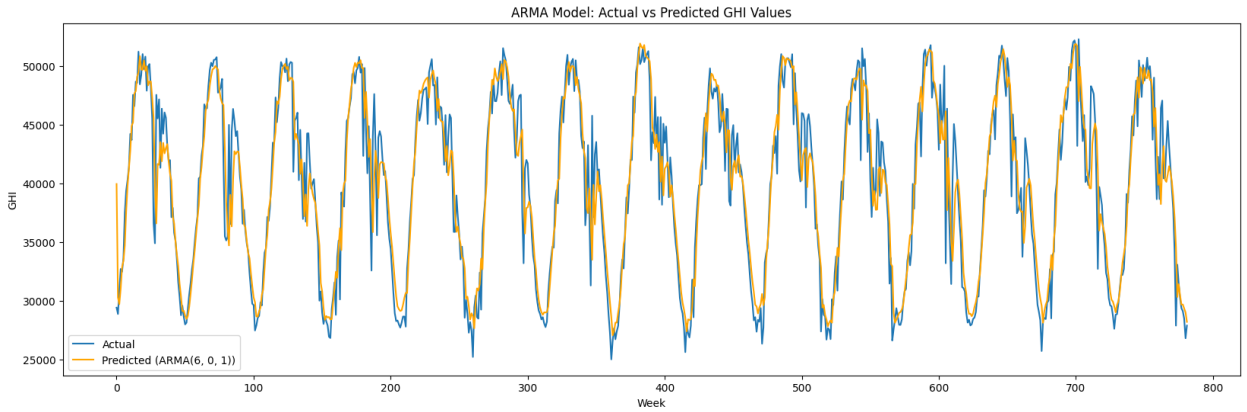
Park-1 Daily ARMA Plot



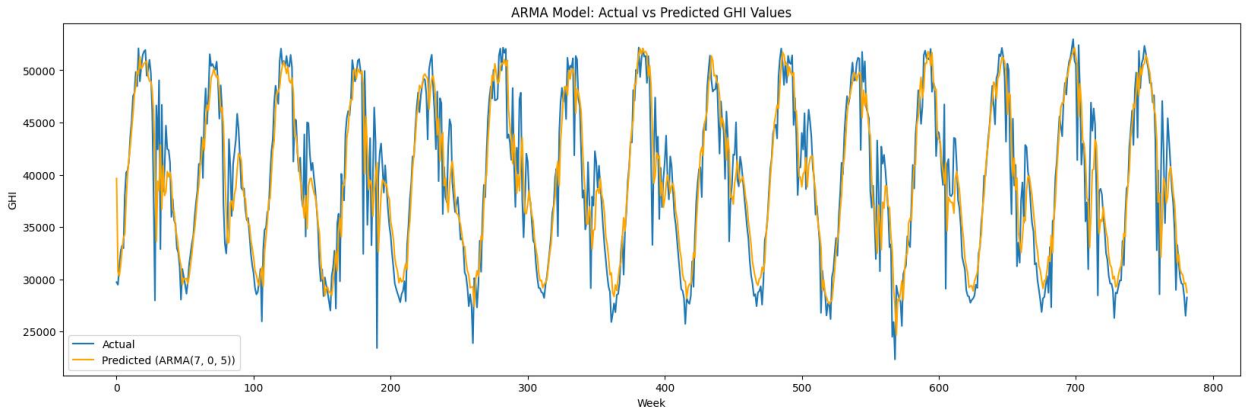
Park-2 Daily ARMA Plot

ARMA model results for weekly data

Region	Hyperparameters (p,q)	MAPE	MAE
Park-1	6, 1	4.87%	1916.27
Park-2	7, 5	6.72%	2508.53



Park-1 Weekly ARMA Plot



Park-2 Weekly ARMA Plot

6.4.4 Autoregressive Integrated Moving Average (ARIMA) models

An ARIMA (Auto Regressive Integrated Moving Average) model is a time series forecasting technique that combines autoregression (AR), differencing (I), and moving averages (MA). The autoregressive component captures the relationship between the current observation and its past values, the integrated component involves differencing to achieve stationarity, and the moving average component models short-term fluctuations and noise. 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. ARIMA models are widely used for forecasting in diverse fields, and the selection of appropriate parameter values involves analyzing autocorrelation and partial autocorrelation functions.

Since our time series was inherently stationary, necessitating no differencing (d=0), the ARIMA models were effectively reduced to ARMA models. Consequently, we have not presented separate results for ARIMA models, as they align with the outcomes derived from ARMA modeling.

6.4.5 Seasonal Autoregressive Integrated Moving Average (SARIMA) models

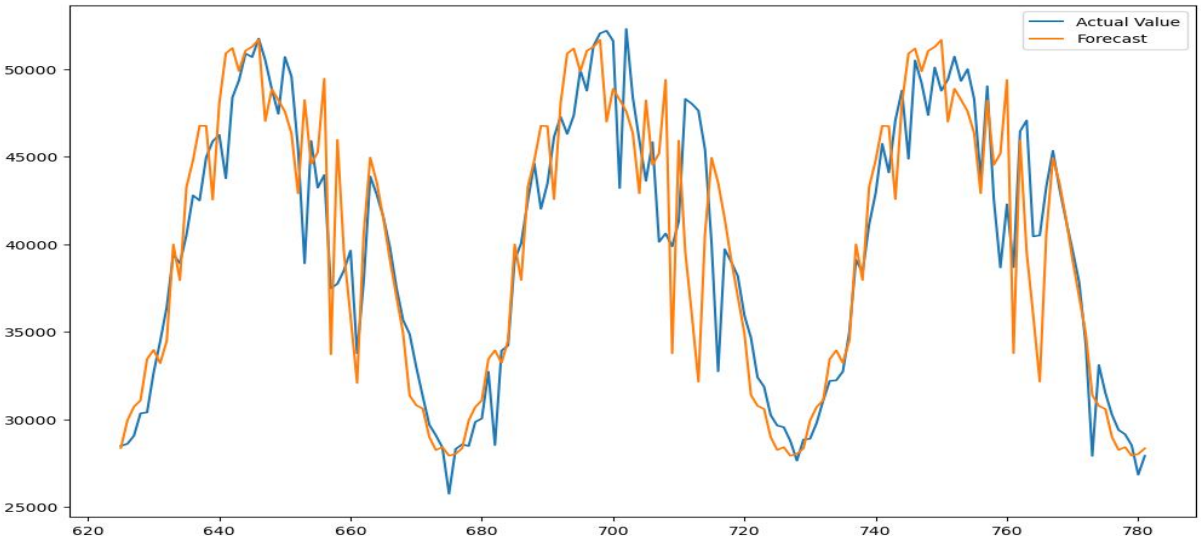
The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends ARIMA to accommodate univariate time series data with a seasonal component. It introduces three additional seasonal hyperparameters (P, D, Q) and a seasonality parameter (m) alongside the existing ARIMA hyperparameters (p, d, q). Specifically, P represents the seasonal autoregressive order, D is the seasonal difference order, Q denotes the seasonal moving average order, and m represents the number of time steps in the seasonal data. A complete SARIMA model is defined by the hyperparameters (p, d, q) (P, D, Q, m), where p, d, and q retain their meanings from the ARIMA model.

Due to the computational intensity of the SARIMA model, its application to daily data was unfeasible. Small hyperparameters were employed for the weekly data.

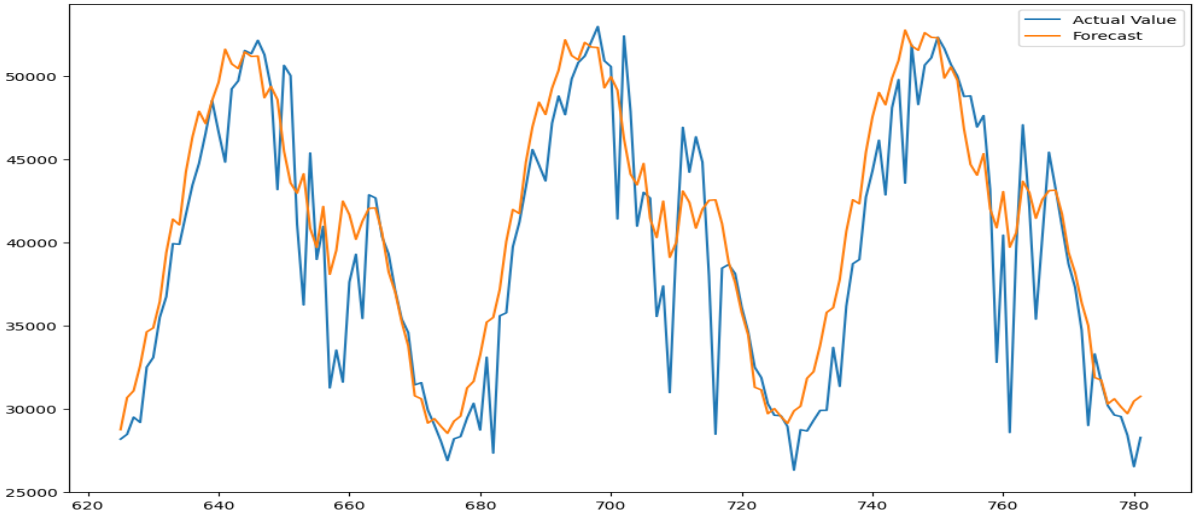
The results of the SARIMA model (for weekly data) are as follows:

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-2	(1,1,1) (1,1,1,52)	6.984%	2560.393



Park-1 SARIMA Plot



Park-2 SARIMA Plot

7. Conclusions

ARMA model gave the best result (lowest MAPE value) for daily data. For weekly, ARMA gave best fit for park 1 and SARIMA for park 2.

As the daily data tends to have much more random variation as compared to weekly data, weekly forecasting was more accurate as compared to daily forecasting.

8. References

- Wikipedia - AR, MA, ARMA, ARIMA, SARIMA
- https://www.linkedin.com/pulse/time-series-analysis-short-introduction-#:~:text=Autocorrelation%20function%20(ACF)%20and%20Partial,Moving%20Average%20(MA)%20models.
- www.mnre.gov.in
- https://www.kaggle.com/code/ryanholbrook/forecasting-with-machine-learning
- https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/#:~:text=There%20are%20various%20statistical%20tests,unit%20root%20in%20the%20data.

Appendices

Please refer to this link for codes:
https://drive.google.com/drive/folders/1iVBR_EuuGmHHJ4CxL0xG8A4gEDu-CD82?usp=drive_link