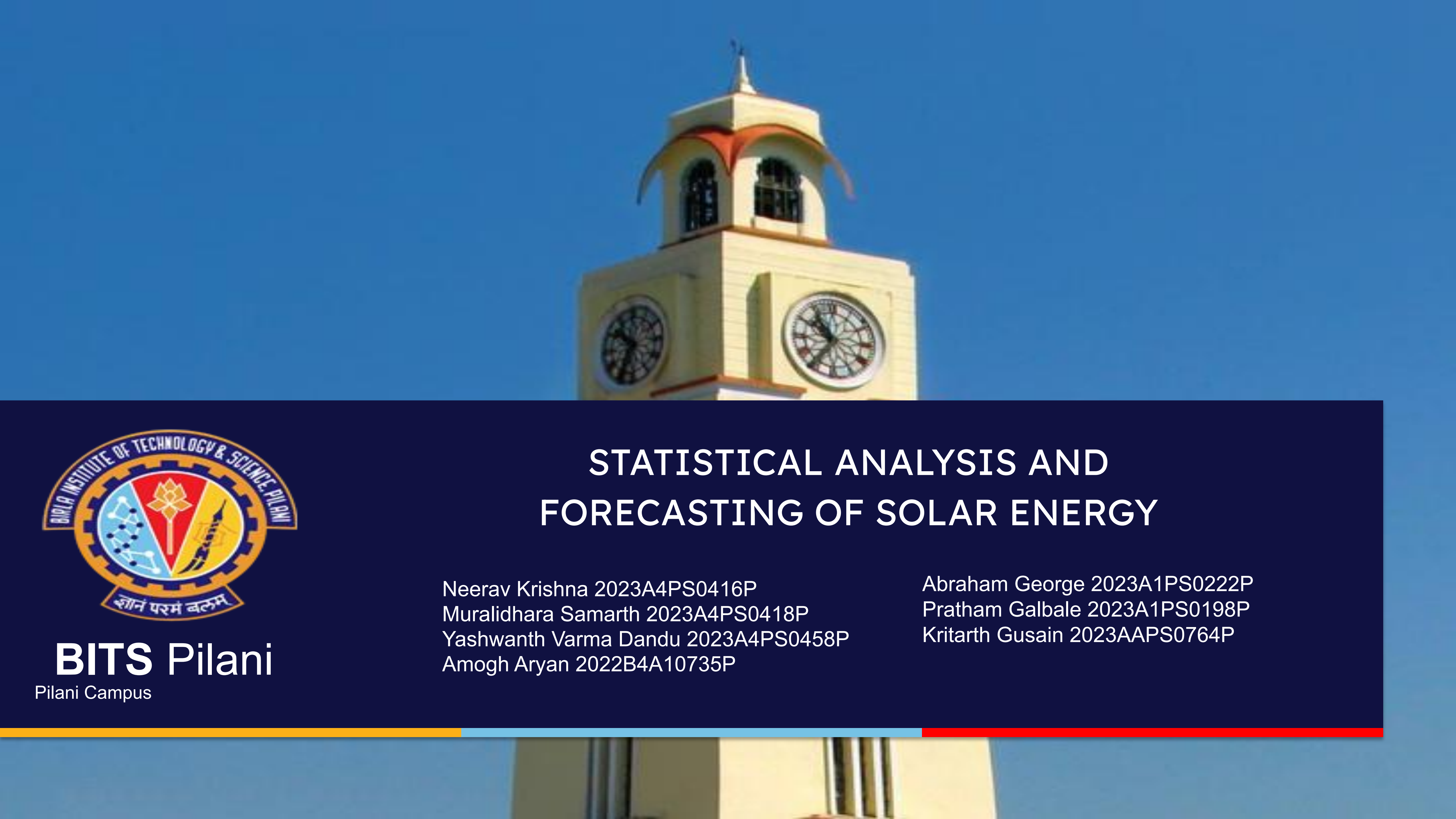


BITS Pilani
Pilani Campus

STATISTICAL ANALYSIS AND FORECASTING OF SOLAR ENERGY

Neerav Krishna 2023A4PS0416P
Muralidhara Samarth 2023A4PS0418P
Yashwanth Varma Dandu 2023A4PS0458P
Amogh Aryan 2022B4A10735P

Abraham George 2023A1PS0222P
Pratham Galbale 2023A1PS0198P
Kritarth Gusain 2023AAPS0764P

Introduction

- Solar energy is an essential source of renewable energy in the modern world. As energy demand continues to rise, **enhancing the reliability and predictability of renewable sources becomes essential for achieving long-term sustainability and energy security.**

- Solar power in India is a fast developing industry as India receives an abundant amount of sunlight throughout the year. The country's solar installed capacity was 36.9 GW as of 30 November 2020. **Rajasthan is one of India's most solar developed states, with its total photovoltaic capacity reaching 2289 MW.**

- However, effective utilization of solar power is inherently limited by meteorological and seasonal factors, such as cloud cover, humidity, temperature fluctuations, and local climatic patterns, that affect the availability of solar radiation. U**nderstanding these variations is therefore crucial for accurate prediction and planning of solar energy generation.**

Why Forecasting?

- In this study, we a**nalyse hourly Global Horizontal Irradiance (GHI) data from 2000 to 2014 obtained from two solar parks located in Rajasthan**. The primary objective of this assignment is to **perform statistical and time-series analysis of the GHI data.** This includes **investigating the distributional characteristics of the dataset, identifying seasonal and trend components, and developing suitable forecasting models.**

- **Why 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 and 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.**

Terms Associated with Solar Power

- **Direct Normal Irradiance(DNI)**: Amount of solar radiation received per unit area by a surface that is always held perpendicular (or normal) to the rays that come in a straight line from the direction of the sun at its current position in the sky.
- **Diffuse Horizontal Irradiance(DHI)**: Solar radiation that does not arrive on a direct path from the sun, but has been scattered by clouds and particles in the atmosphere and comes equally from all directions.
- **The solar zenith angle(Z):** Angle between the sun’s rights and the vertical.
- **Global Horizontal Irradiance (GHI)**: Total amount of shortwave radiation received from above by a surface parallel to the ground.

- The following relation holds between GHI, DNI and DHI: **GHI = DNI ∗ cosZ + DHI**
- Therefore, GHI is a function of DNI, Z and DHI and **takes into account the effect of all these variables.**

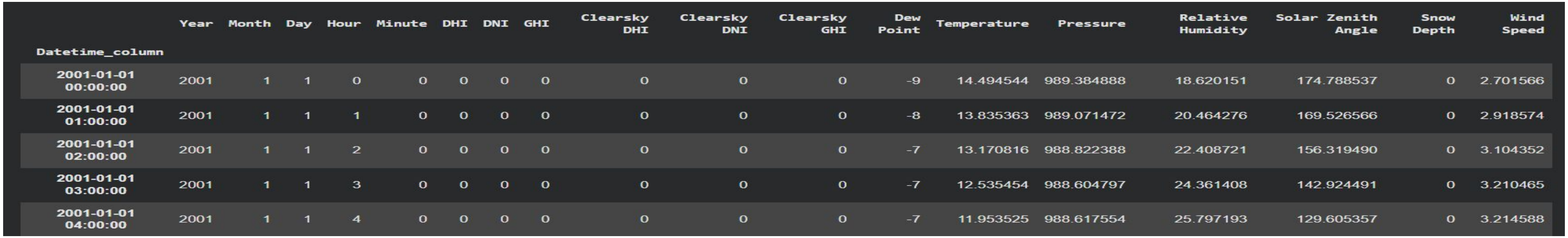
Dataset and Pre-Processing

The implementation of time series analysis was done using Python. The given dataset contains hourly information collected over a period of **15 years(2000-2014) at 2 regions in Rajasthan.** Information about the following attributes is available in the dataset:

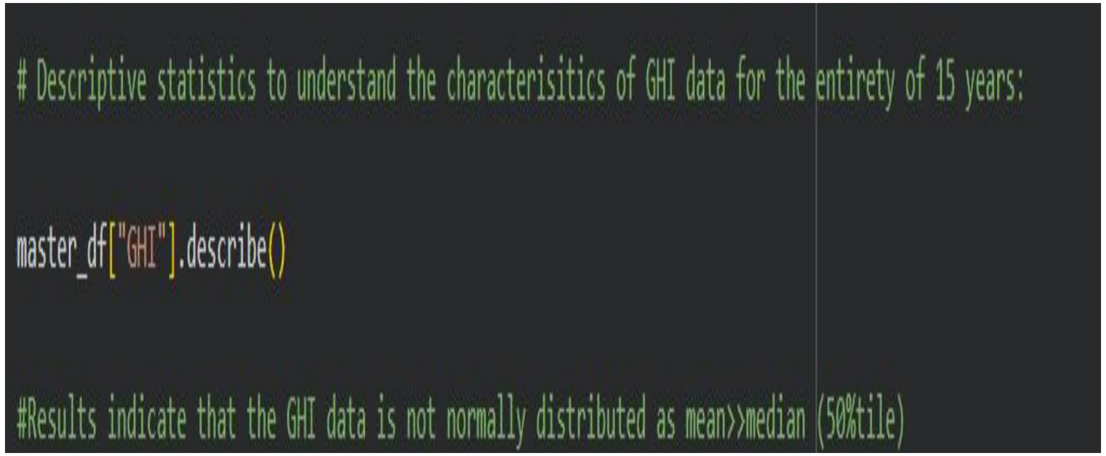
1. Date and Time of measurement.
2. DHI and Clearsky DHI.
3. DNI and Clearsky DNI.
4. GHI and Clearsky GHI.
5. Dew Point.
6. Temperature.
7. Pressure.
8. Relative Humidity.
9. Solar Zenith Angle.
10. Wind Speed.

- We choose to analyse the GHI column as **it represents the total energy available to a typical solar park**, making it the most important metric for forecasting energy generation.

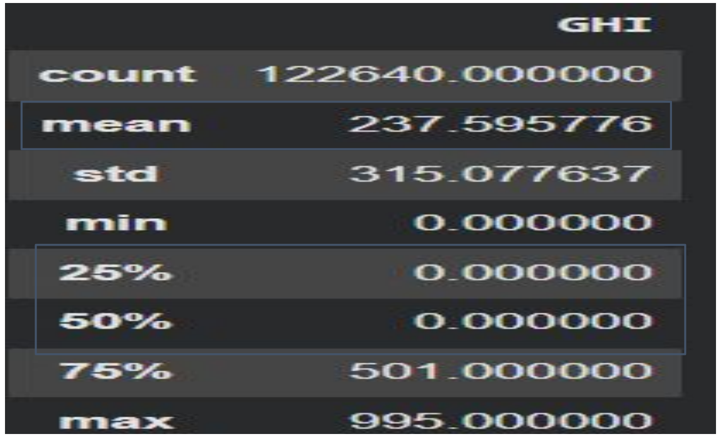
EDA: Descriptive Statistics



	Year	Month	Day	Hour	Minute	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	Temperature	Pressure	Relative Humidity	Solar Zenith Angle	Snow Depth	Wind Speed
Datetime_column																		
2001-01-01 00:00:00	2001	1	1	0	0	0	0	0	0	0	0	-9	14.494544	989.384888	18.620151	174.788537	0	2.701566
2001-01-01 01:00:00	2001	1	1	1	0	0	0	0	0	0	0	-8	13.835363	989.071472	20.464276	169.526566	0	2.918574
2001-01-01 02:00:00	2001	1	1	2	0	0	0	0	0	0	0	-7	13.170816	988.822388	22.408721	156.319490	0	3.104352
2001-01-01 03:00:00	2001	1	1	3	0	0	0	0	0	0	0	-7	12.535454	988.604797	24.361408	142.924491	0	3.210465
2001-01-01 04:00:00	2001	1	1	4	0	0	0	0	0	0	0	-7	11.953525	988.617554	25.797193	129.605357	0	3.214588

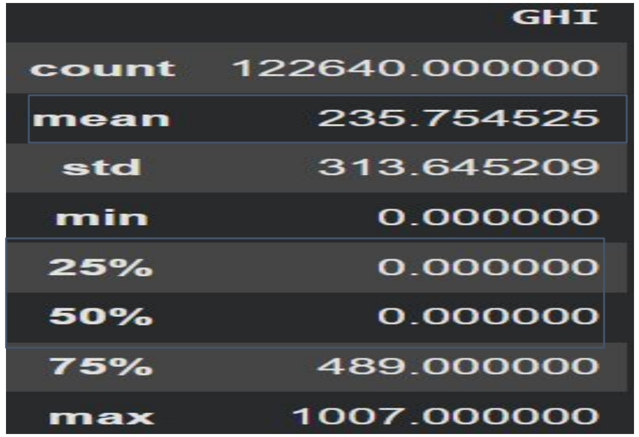


```
# Descriptive statistics to understand the characterisitics of GHI data for the entirety of 15 years:

master_df["GHI"].describe()

#Results indicate that the GHI data is not normally distributed as mean>>median (50%tile)
```



```
                    GHI
count  122640.000000
mean      237.595776
std       315.077637
min         0.000000
25%         0.000000
50%         0.000000
75%       501.000000
max       995.000000
```

RJ1



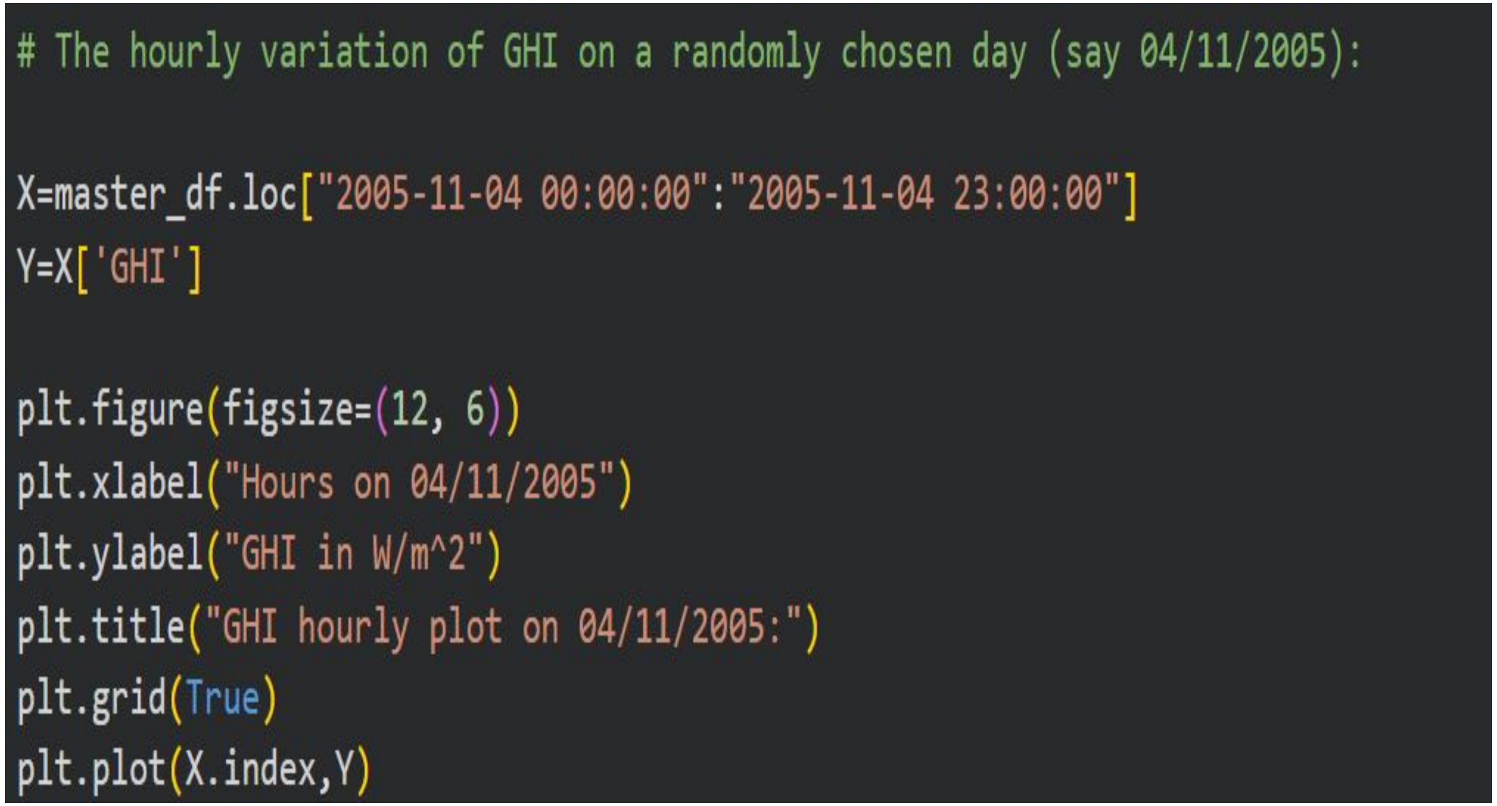
```
                    GHI
count  122640.000000
mean      235.754525
std       313.645209
min         0.000000
25%         0.000000
50%         0.000000
75%       489.000000
max      1007.000000
```

RJ2

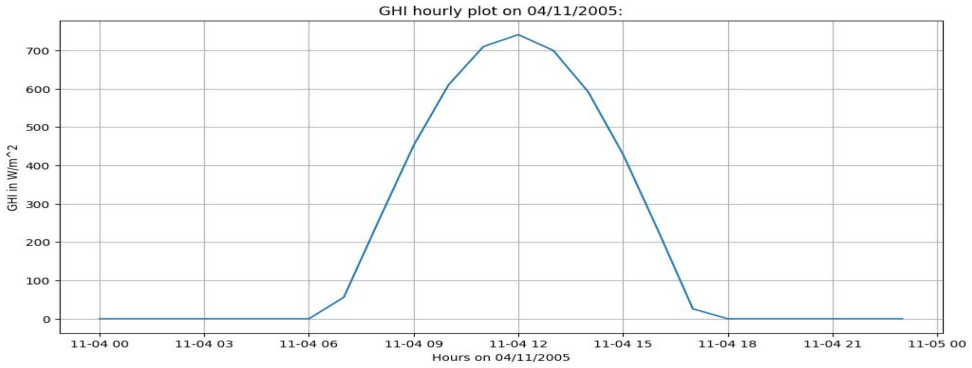
Note that mean>>Q2 and Q1=Q2=0

EDA: Hourly variation of GHI in a day

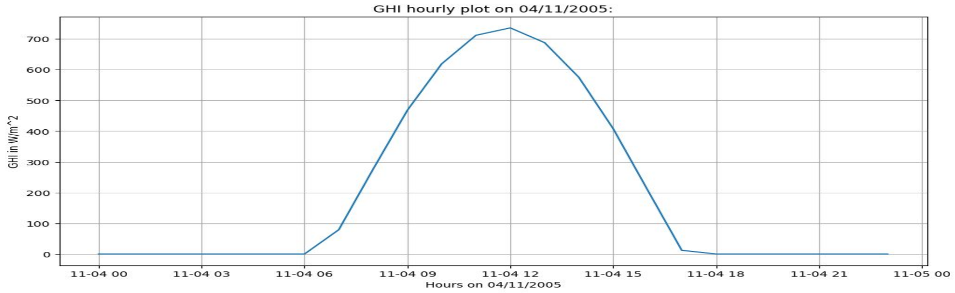
Examining GHI on a randomly selected day:



```
# The hourly variation of GHI on a randomly chosen day (say 04/11/2005):

X=master_df.loc["2005-11-04 00:00:00":"2005-11-04 23:00:00"]
Y=X['GHI']

plt.figure(figsize=(12, 6))
plt.xlabel("Hours on 04/11/2005")
plt.ylabel("GHI in W/m^2")
plt.title("GHI hourly plot on 04/11/2005:")
plt.grid(True)
plt.plot(X.index,Y)
```



RJ1

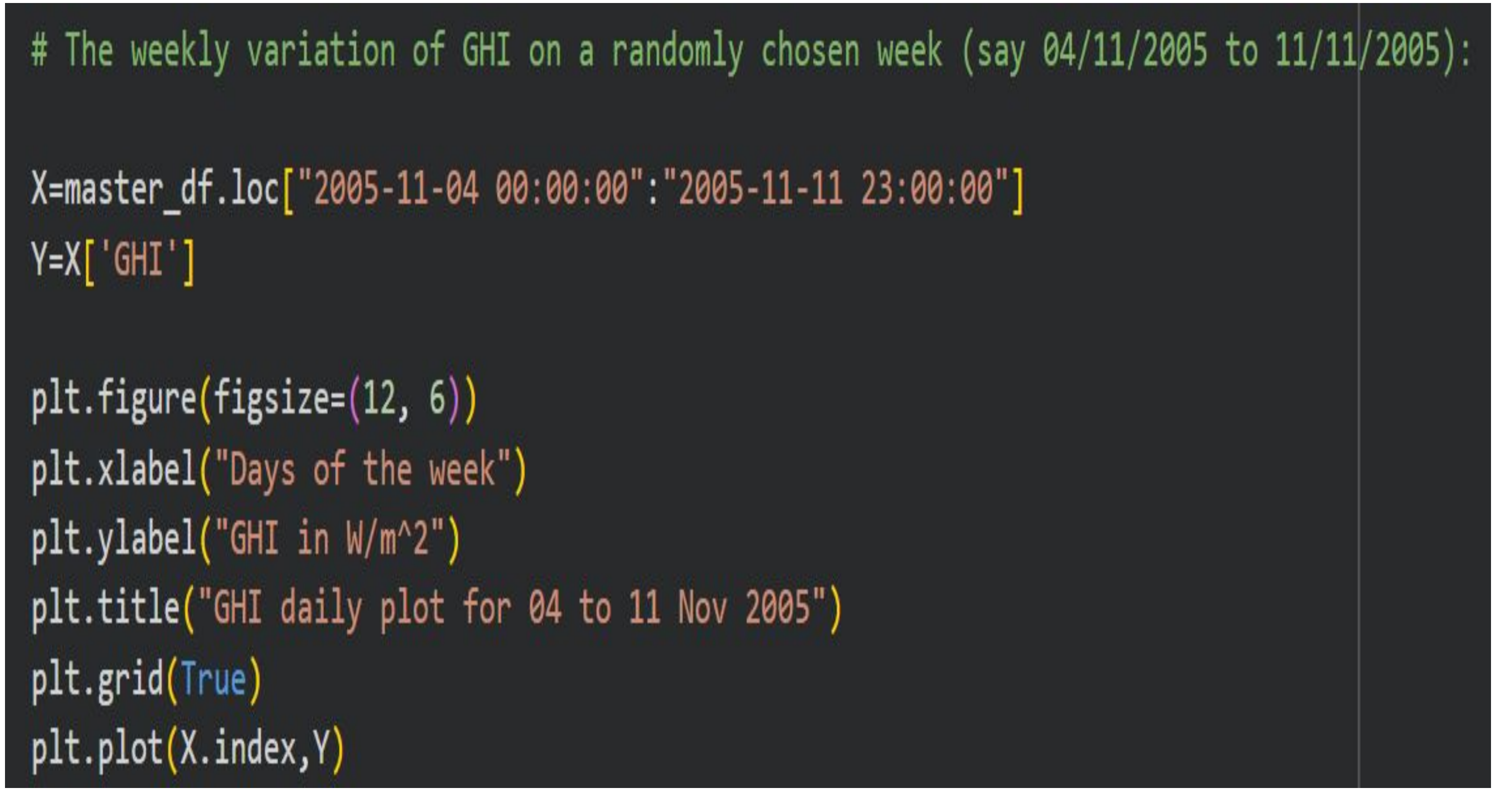


RJ2

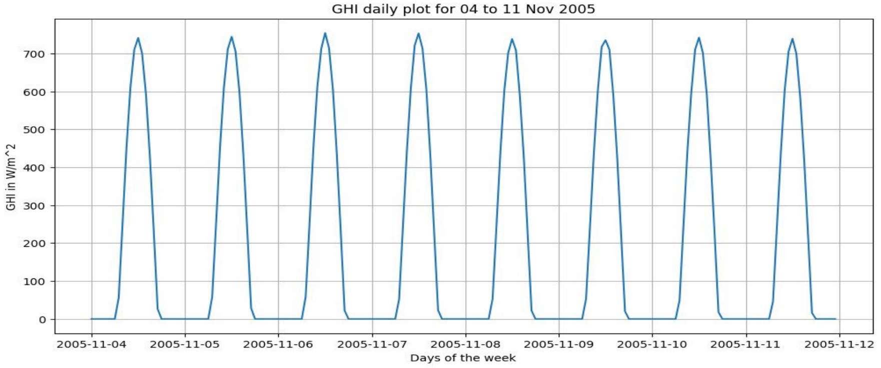
We see that GHI peeks exactly noon and is the least during the dusk and dawn

EDA: Daily variation of GHI in a week

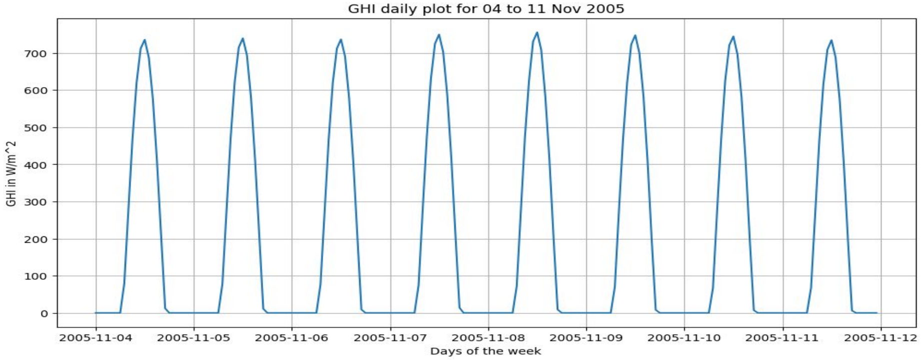
Examining GHI during a randomly selected week:



```
# The weekly variation of GHI on a randomly chosen week (say 04/11/2005 to 11/11/2005):

X=master_df.loc["2005-11-04 00:00:00":"2005-11-11 23:00:00"]
Y=X['GHI']

plt.figure(figsize=(12, 6))
plt.xlabel("Days of the week")
plt.ylabel("GHI in W/m^2")
plt.title("GHI daily plot for 04 to 11 Nov 2005")
plt.grid(True)
plt.plot(X.index,Y)
```



RJ1

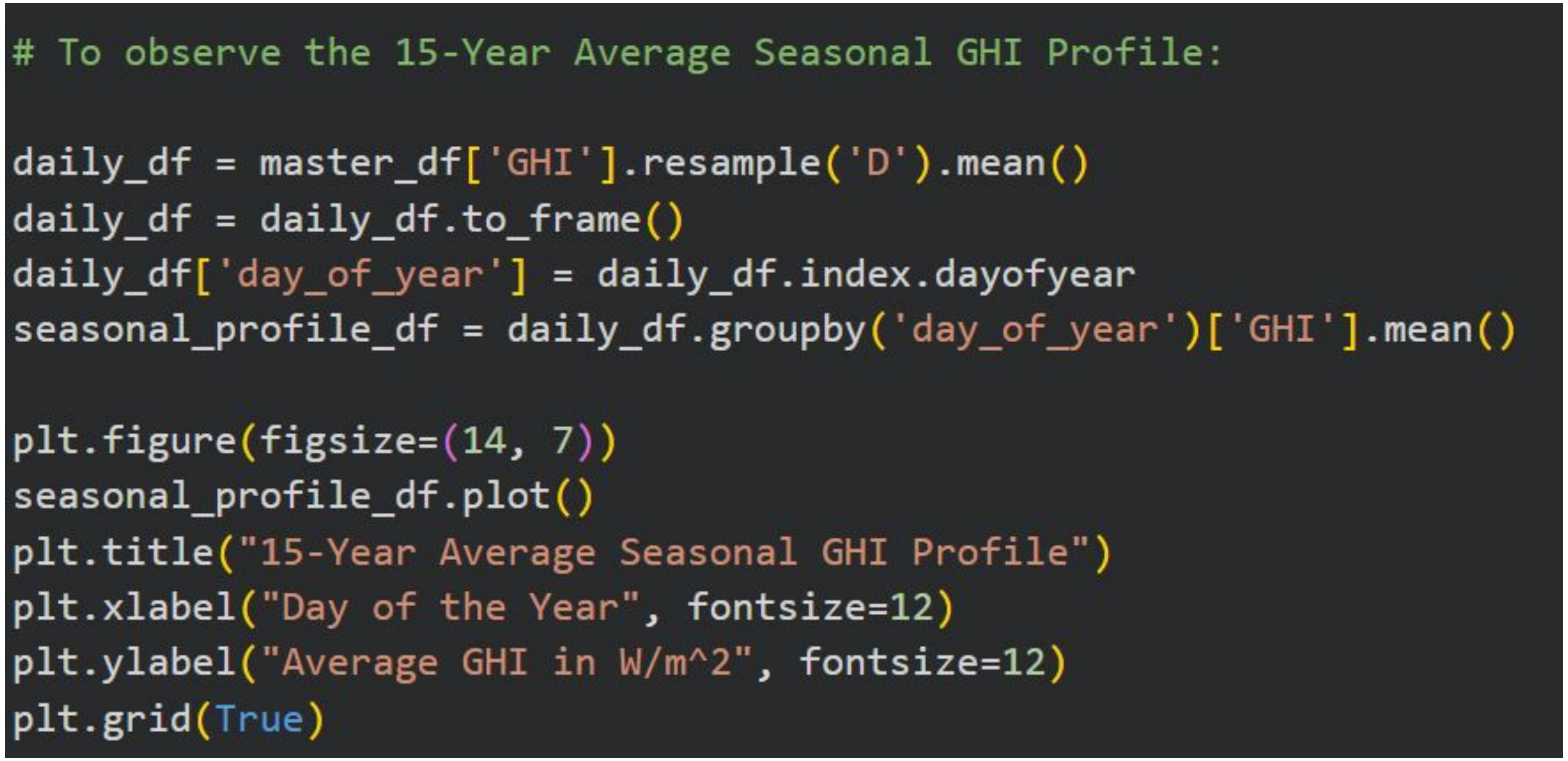


We observe a seasonal pattern of peaks that seem to occur everyday around 12pm

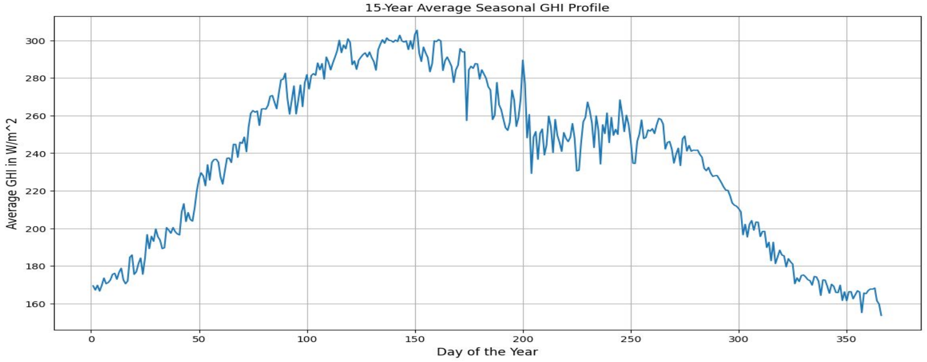
RJ2

EDA: Daily variation of GHI in a year

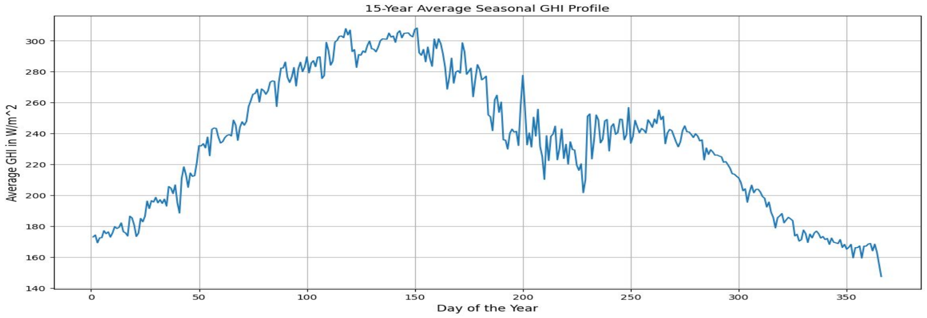
Examining averaged out daily GHI during a year:



```
# To observe the 15-Year Average Seasonal GHI Profile:

daily_df = master_df['GHI'].resample('D').mean()
daily_df = daily_df.to_frame()
daily_df['day_of_year'] = daily_df.index.dayofyear
seasonal_profile_df = daily_df.groupby('day_of_year')['GHI'].mean()

plt.figure(figsize=(14, 7))
seasonal_profile_df.plot()
plt.title("15-Year Average Seasonal GHI Profile")
plt.xlabel("Day of the Year", fontsize=12)
plt.ylabel("Average GHI in W/m^2", fontsize=12)
plt.grid(True)
```



RJ1

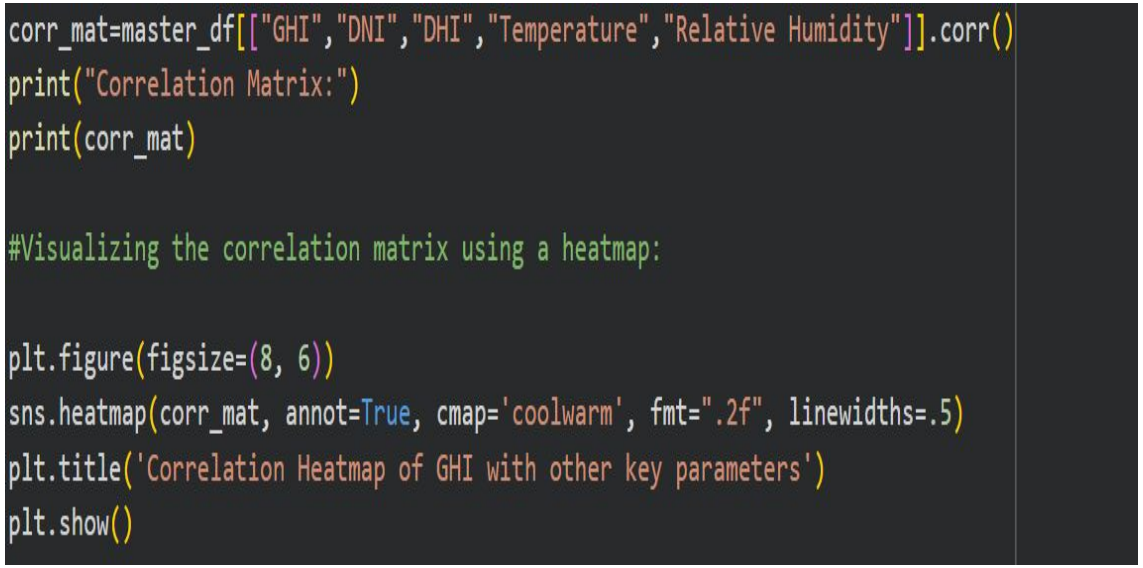


RJ2

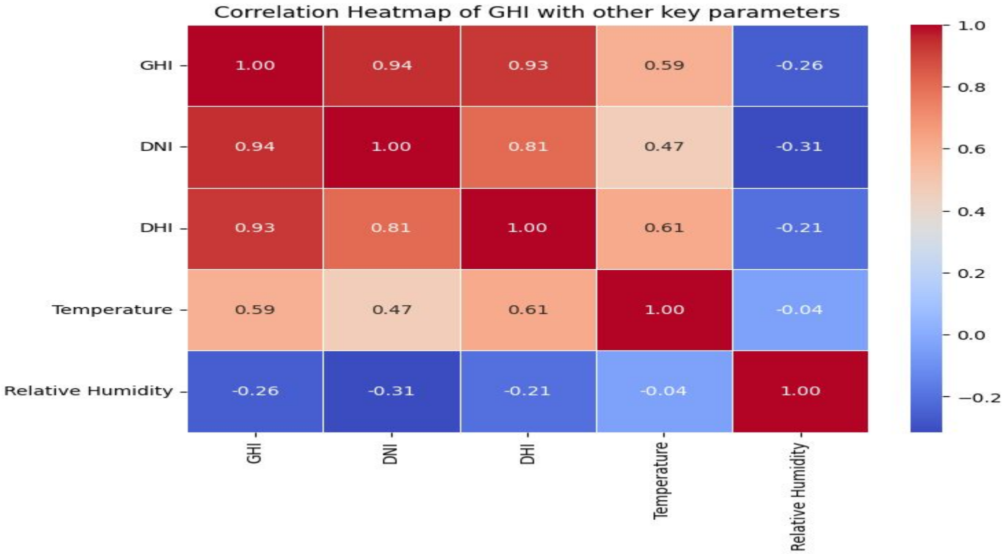
The GHI peaks during the summer days and slowly reduces on the onset of monsoon (sharp drop)

Relation between GHI and other parameters

Evaluated by the **correlation matrix** between GHI other parameters like temperature, relative humidity, DNI among others. Visualised using a **heatmap**.



```
corr_mat=master_df[["GHI","DNI","DHI","Temperature","Relative Humidity"]].corr()
print("Correlation Matrix:")
print(corr_mat)

#Visualizing the correlation matrix using a heatmap:

plt.figure(figsize=(8, 6))
sns.heatmap(corr_mat, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Correlation Heatmap of GHI with other key parameters')
plt.show()
```



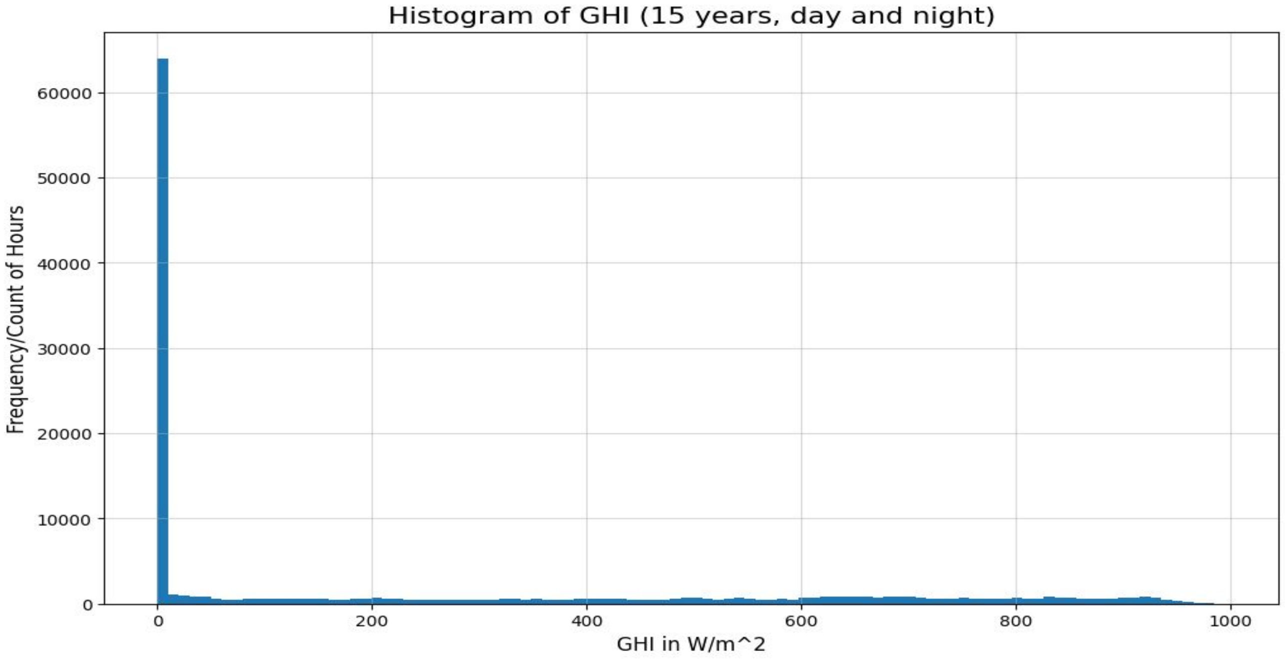
Relation between GHI and other parameters

Key Inferences:

1. GHI shows **strong positive linear relationship with DNI and DHI** as **all three are different measures of irradiance**.
2. GHI also has p**ositive correlation with temperature** as the hotter a day is, the more likely a high irradiance is obtained.
3. GHI has a **negative linear relationship** with relative humidity because this is might be an indication of the amount of clouds present which will definitely reduce the irradiance incident on a solar panel

Is the GHI data normal?

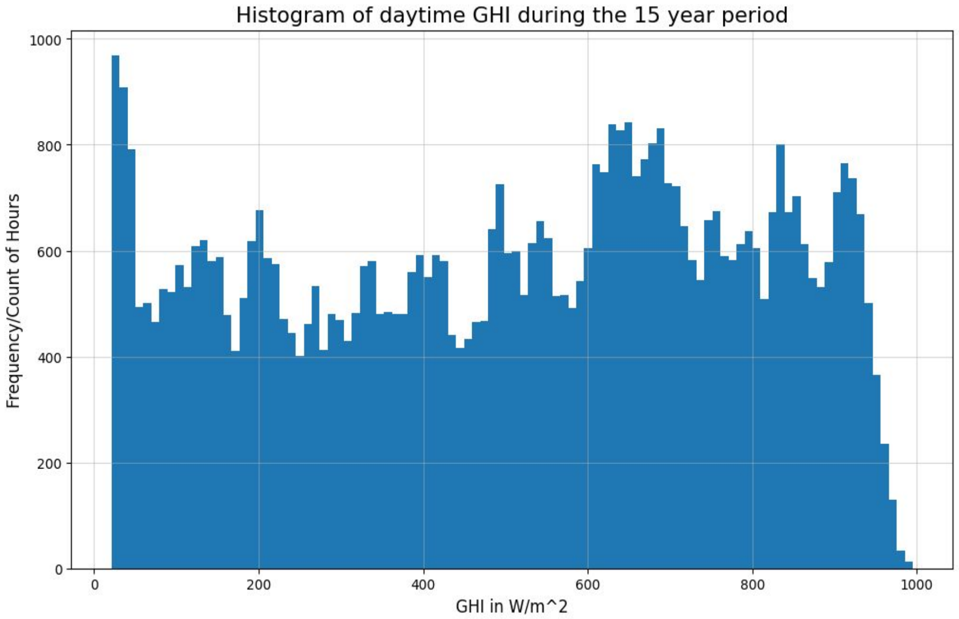
Plotting the histogram:



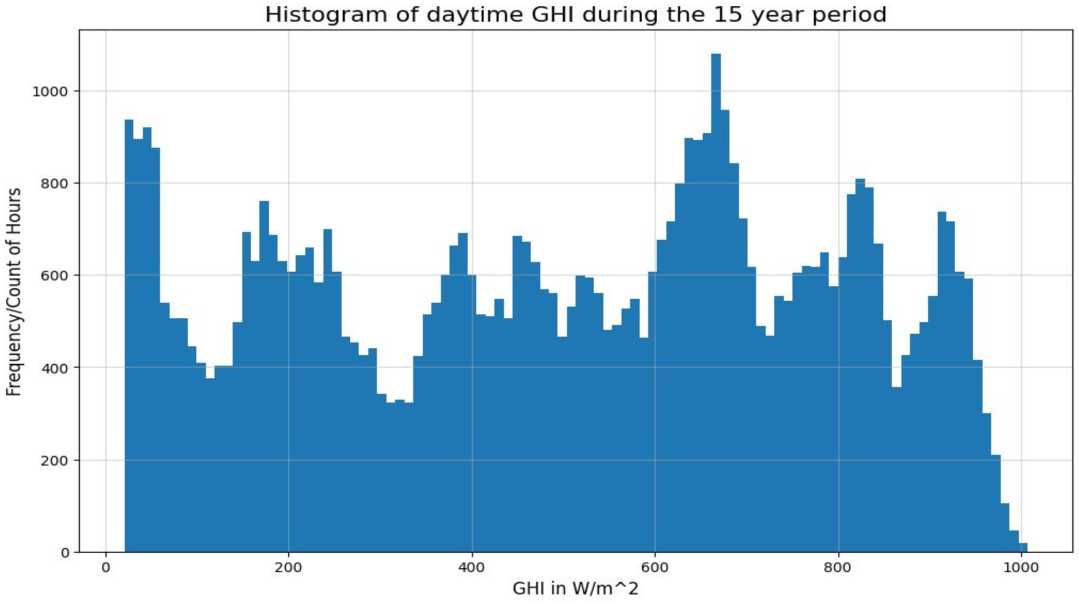
Heavy positively skewed histogram as it includes night time data

Is the GHI data normal?

After removing the night time GHI values (which all are 0):



RJ1

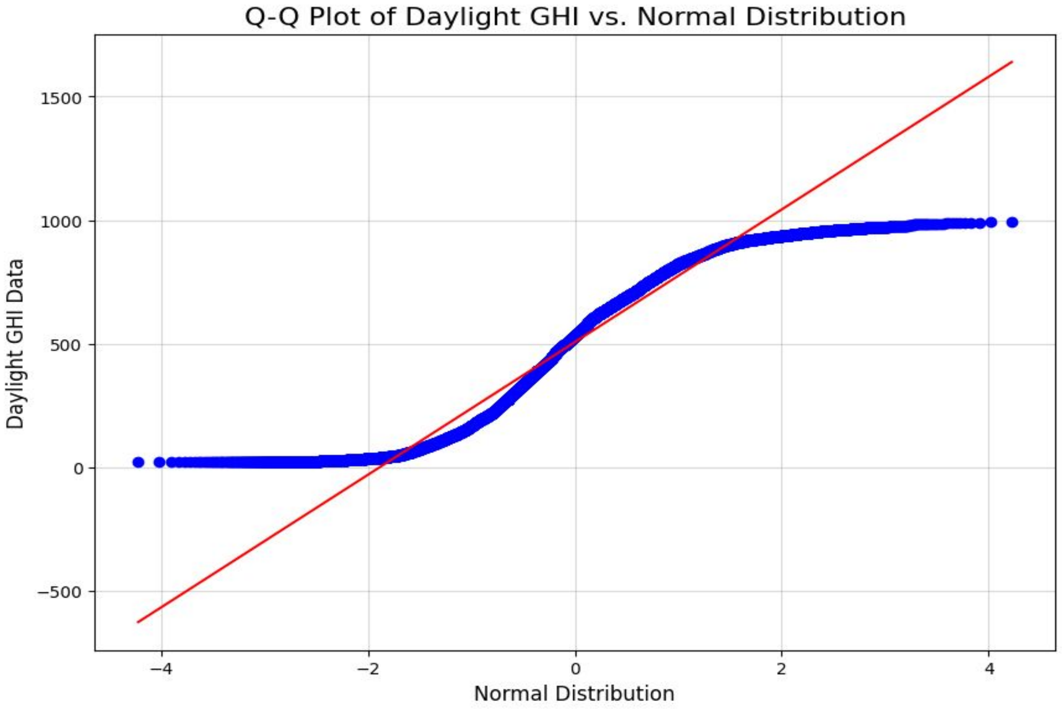


RJ2

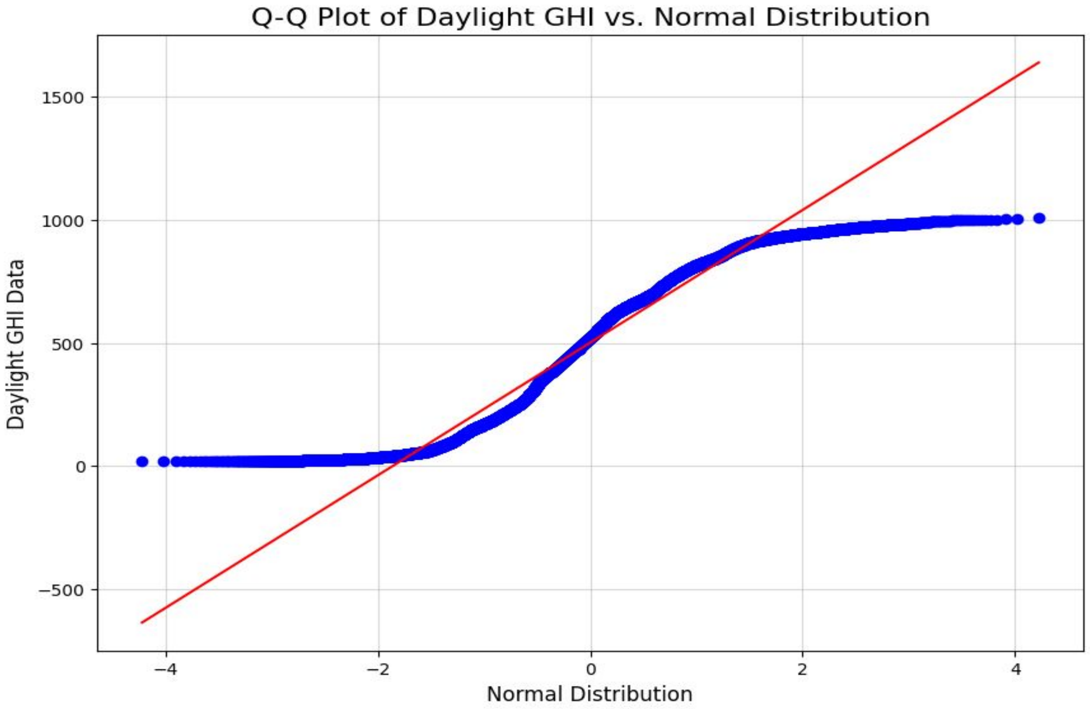
Day time distribution of GHI does not follow a normal, distibution. It has multiple different peaks, therefore can be considered a **complex multimodal distribution**

Is the GHI data normal?

The Q-Q plots for both RJ 1 and RJ 2 indicate the same results:



RJ1



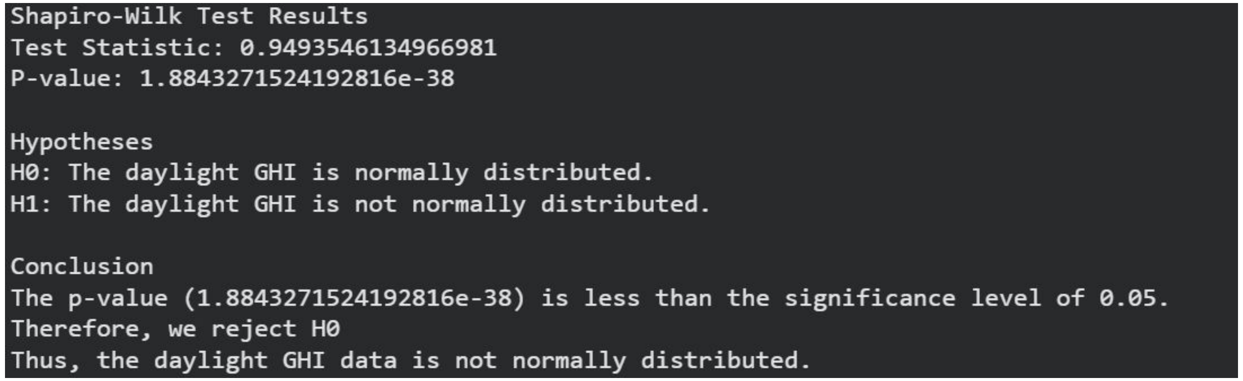
RJ2

Is the GHI data normal?

The Shapiro Wilk test (with alpha=0.05) was performed on a set of 5000 random datapoints to confirm the non normality of day time GHI data.

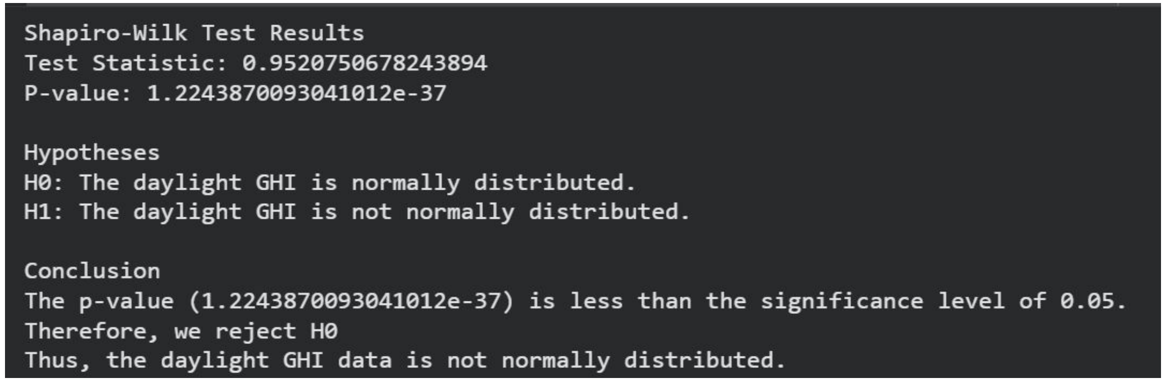
H0: The daylight GHI is normally distributed

H1: The daylight GHI is not normally distributed



```
Shapiro-Wilk Test Results
Test Statistic: 0.9493546134966981
P-value: 1.8843271524192816e-38

Hypotheses
H0: The daylight GHI is normally distributed.
H1: The daylight GHI is not normally distributed.

Conclusion
The p-value (1.8843271524192816e-38) is less than the significance level of 0.05.
Therefore, we reject H0
Thus, the daylight GHI data is not normally distributed.
```

RJ1

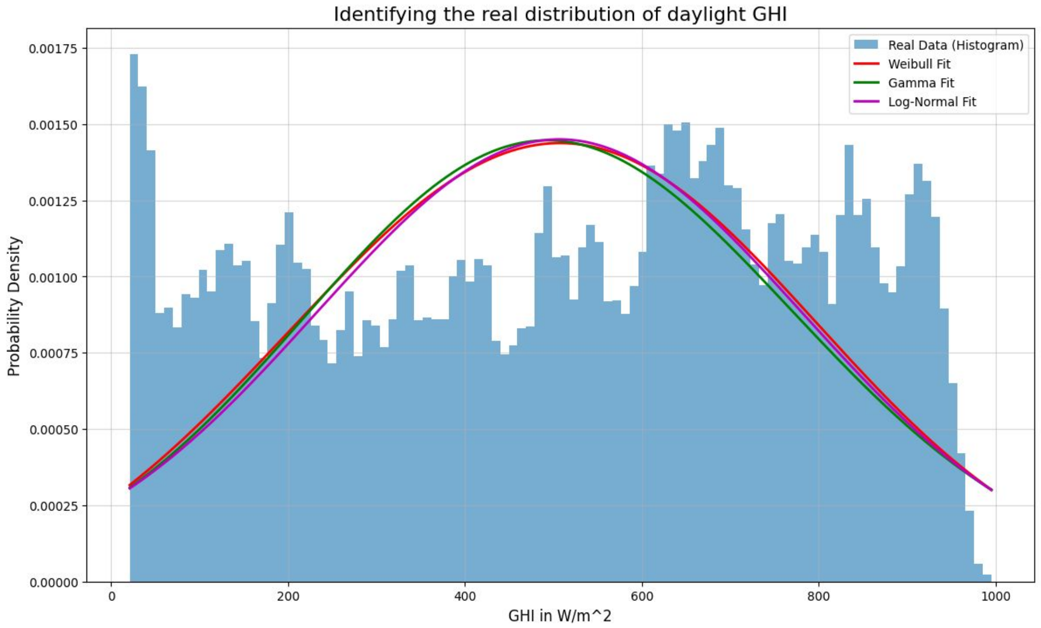


```
Shapiro-Wilk Test Results
Test Statistic: 0.9520750678243894
P-value: 1.2243870093041012e-37

Hypotheses
H0: The daylight GHI is normally distributed.
H1: The daylight GHI is not normally distributed.

Conclusion
The p-value (1.2243870093041012e-37) is less than the significance level of 0.05.
Therefore, we reject H0
Thus, the daylight GHI data is not normally distributed.
```

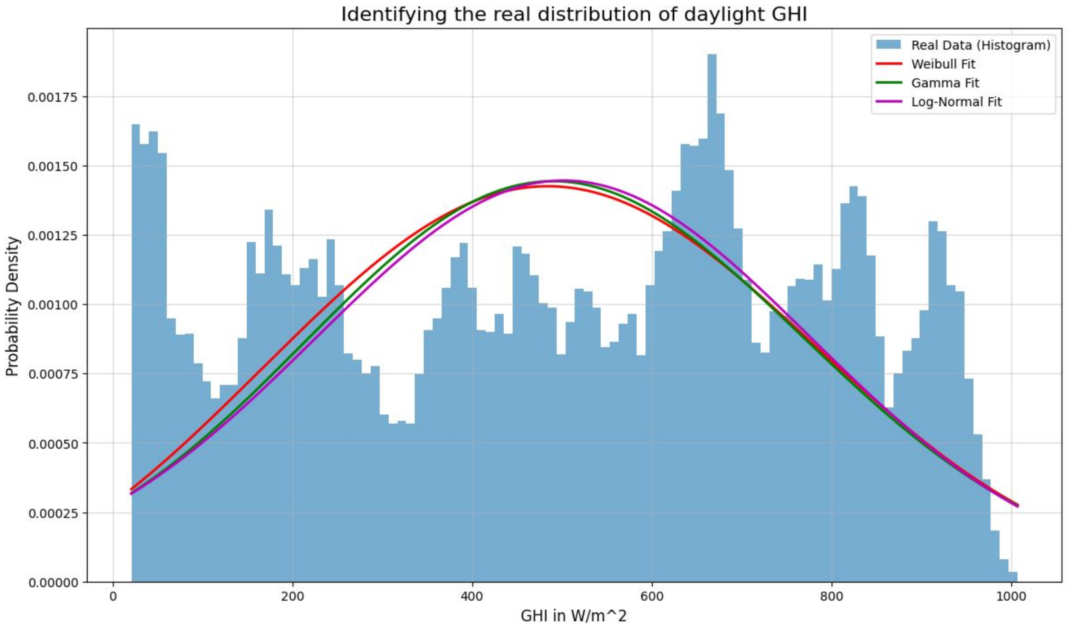
RJ2

What distribution does GHI follow?

We try to visually see which distribution that fits the histogram of daytime GHI. We choose three distributions (Weibull, Gamma and Log-normal) for this visual analysis.



RJ1



RJ2

What distribution does GHI follow?

Key inferences from the plots:

1. Daylight GHI does not seem to follow any of the standard distributions.
2. Rather seems to be a mixture/combination of several different distributions.
3. Implies that, none of the standard distributions seem to fit the complex real world daylight GHI data.

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)

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:

   H_0: The series has no unit root.

   H_a: The series has a unit root.

Tests for Stationarity

For Rajasthan-1 for daily data:

```
Results of Dickey-Fuller(ADF) Test:
Test Statistic              -5.219990
p-value                      0.000008
```

For Rajasthan-2 for daily data:

```
Results of Dickey-Fuller(ADF) Test:
Test Statistic              -5.101528
p-value                      0.000014
```

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(KPSS Test).

Tests for Stationarity

2. Kwiatkowski-Phillips-Schmidt-Shin Test(KPSS): The hypotheses of the KPSS test are:

H_0: The series is stationary

H_a: The series is not stationary.

The 1% critical value for this test is known to be 0.739.

```
Results of KPSS Test:
```

For Rajasthan-1 for daily data:

```
p-value                     0.03364
```

For Rajasthan-2 for daily data:

```
p-value                     0.04890
```

The test statistic was less than the 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. 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.

Tests for Stationarity

Conclusion about Stationarity From the combination of ADF and KPSS tests, four cases can arise:

Case 1: Both tests conclude that the series is stationary. In this case, we can conclude that the series is stationary.

Case 2: KPSS indicates stationarity and ADF does not. Here, the conclusion would be that the series is trend stationary. Trend needs to be removed to make the series strict stationary.

Case 3: ADF indicates stationarity and KPSS does not. This indicates that the series is difference stationary. Differencing needs to be done to make it stationary.

Case 4: Both tests conclude that the series is not stationary. In this case, we can conclude that the series is not stationary.

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

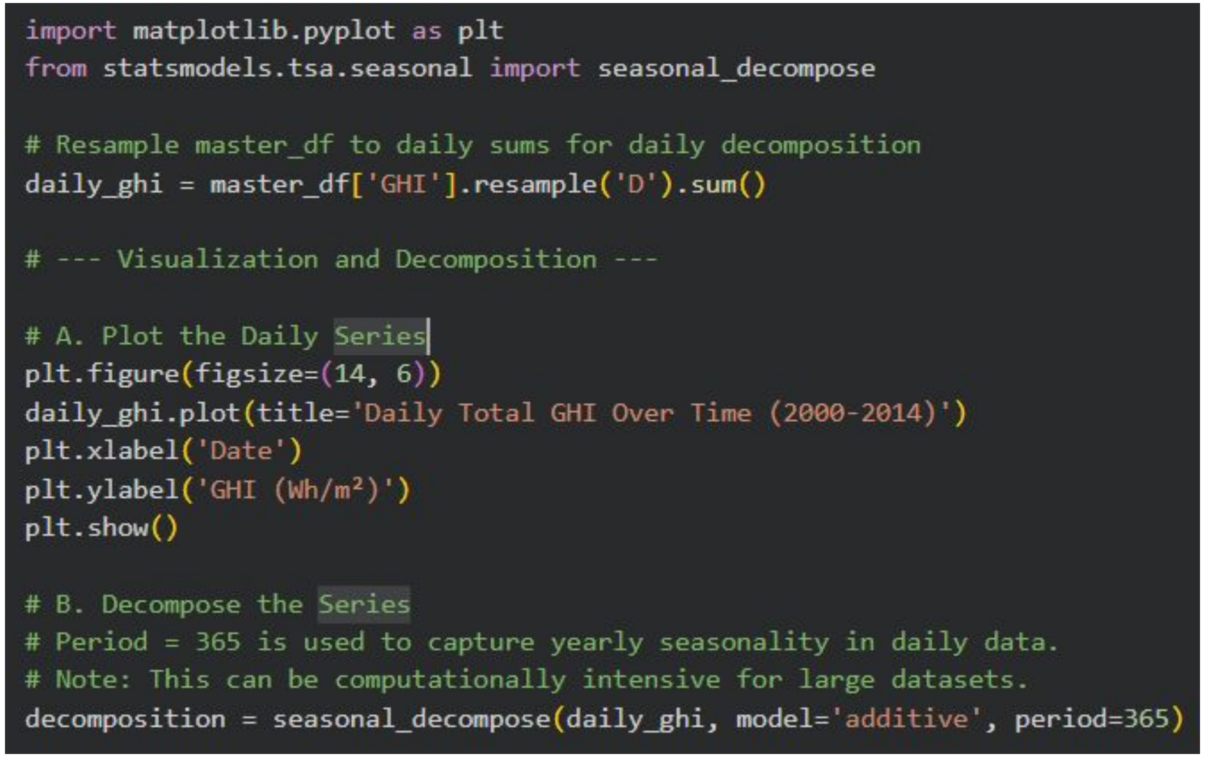
The time series analysis serves as the foundation for selecting and tuning appropriate forecasting models. This analysis focuses on the Daily Average Global Horizontal Irradiance (GHI) data from 2000 to 2014, obtained by resampling the original hourly data.

The Time Series Decomposition was decomposed into its three primary components—Trend, Seasonality, and Residuals(noise), using an additive model with an annual period of 365. The additive model was used here as the magnitude of seasonal fluctuation is not varying much.

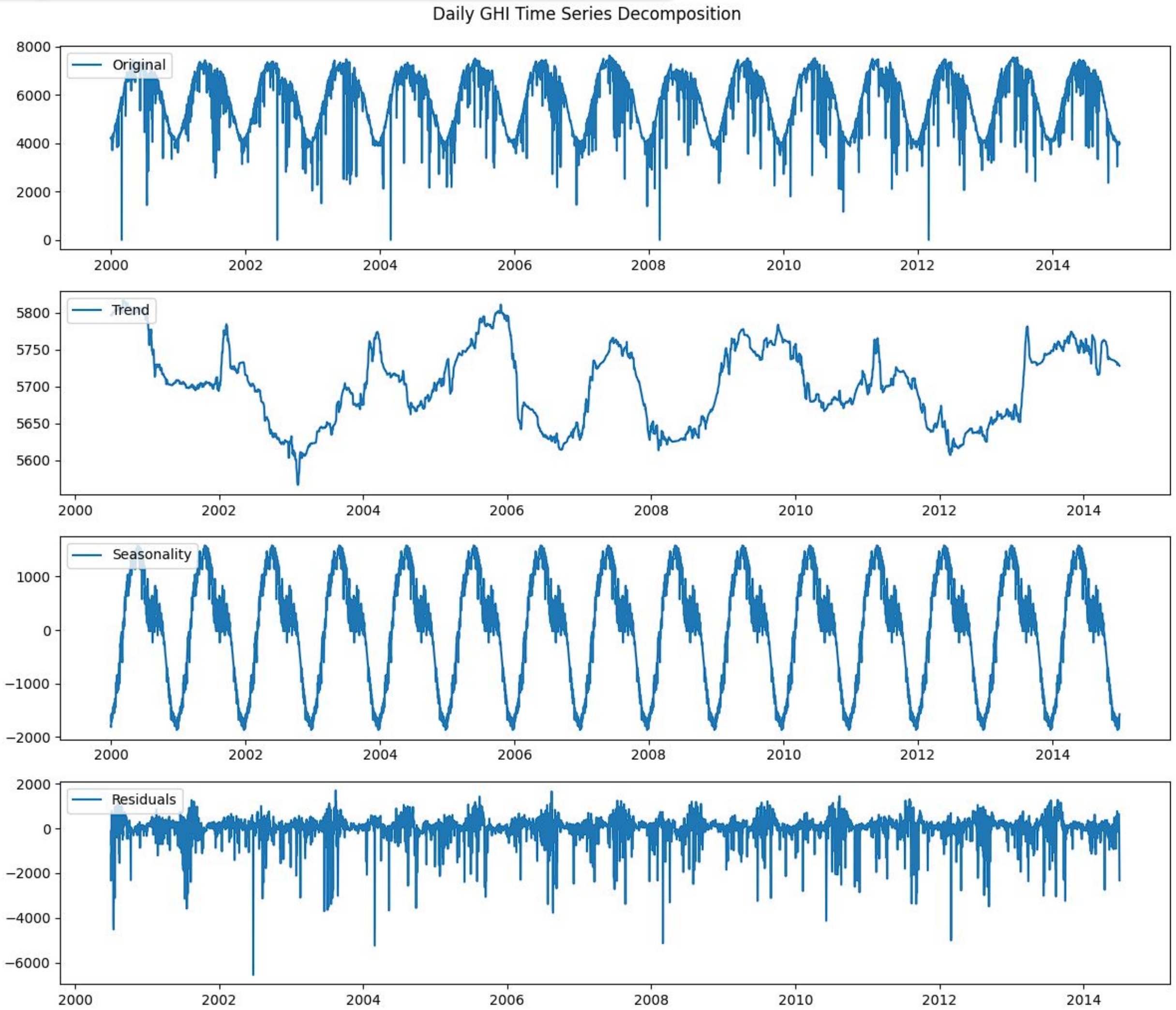
It can be seen that there is a seasonality in the daily series and no uniform trend exists. Similar inferences can be drawn about weekly data

Time Series Decomposition

Fig. Time Series Decomposition of Daily Data from Rajasthan-1



```
import matplotlib.pyplot as plt
from statsmodels.tsa.seasonal import seasonal_decompose

# Resample master_df to daily sums for daily decomposition
daily_ghi = master_df['GHI'].resample('D').sum()

# --- Visualization and Decomposition ---

# A. Plot the Daily Series
plt.figure(figsize=(14, 6))
daily_ghi.plot(title='Daily Total GHI Over Time (2000-2014)')
plt.xlabel('Date')
plt.ylabel('GHI (Wh/m²)')
plt.show()

# B. Decompose the Series
# Period = 365 is used to capture yearly seasonality in daily data.
# Note: This can be computationally intensive for large datasets.
decomposition = seasonal_decompose(daily_ghi, model='additive', period=365)
```



Time Series Forecasting

There are several time series models that we have considered for forecasting.
These include:

1. **AR (p)**
2. **MA (q)**
3. **ARMA (p, q)**
4. **ARIMA (p, d, q)**
5. **SARIMA (p, d, q) (P, D, Q, m)**

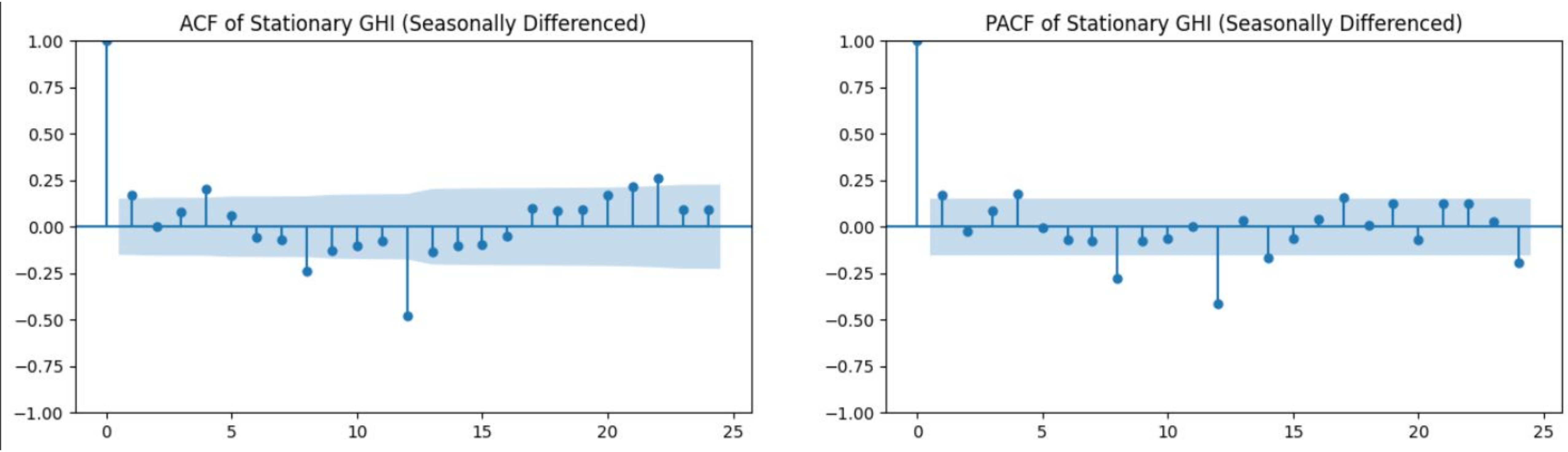
Our approach involves a comparison of the performance of all of these models on various evaluation criteria. These include accuracies and various error terms like **Mean Absolute Percentage Error(MAPE), Mean Absolute Error(MAE), Mean Square Error(MSE) and Root Mean Square Error(RMSE).**

Important Concepts

1. **Autocorrelation Function(ACF)**: ACF plot is a bar chart of coefficients of correlation between a time series and the lagged values. ACF 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. The y-axis expresses the correlation coefficient in the ACF plot, whereas the x-axis mentions the number of lags. The last ACF factor greater than the specified threshold value denotes the maximum significant lag that also represents the upper limit of the hyperparameter q.

2. **Partial Autocorrelation Function (PACF):** The partial autocorrelation function explains the partial correlation between the series and the lagged values. In simple terms, PACF can be explained using a linear regression where we predict y(t) from y(t-1), y(t-2), and y(t-3). In PACF, we correlate the “parts” of y(t) and y(t-3) that are not predicted by y(t-1) and y(t-2). The last PACF factor greater than the specified threshold value denotes the maximum significant lag that also represents the upper limit of the hyperparameter p.

ACF and PACF plots for Rajasthan-1

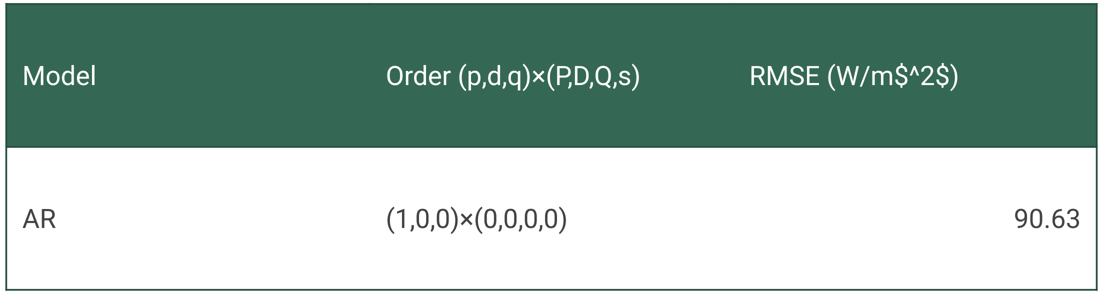


Models for Forecasting

1. **Autoregressive (AR) models:**
   The Autoregressive models implicitly assume that the future values will be based on the past values and predict in accordance with a relationship between them. The equation describing this model is:

$$X_t = C + \sum_{i=1}^{p} \phi_t X_{t\text{-}i} + \epsilon_t$$

   AR forecast quickly flattens out due to the lack of differencing and seasonal terms. Its high RMSE of 90.63 W/m^2 shows it fails to predict any of the GHI's dynamic movements across the year.



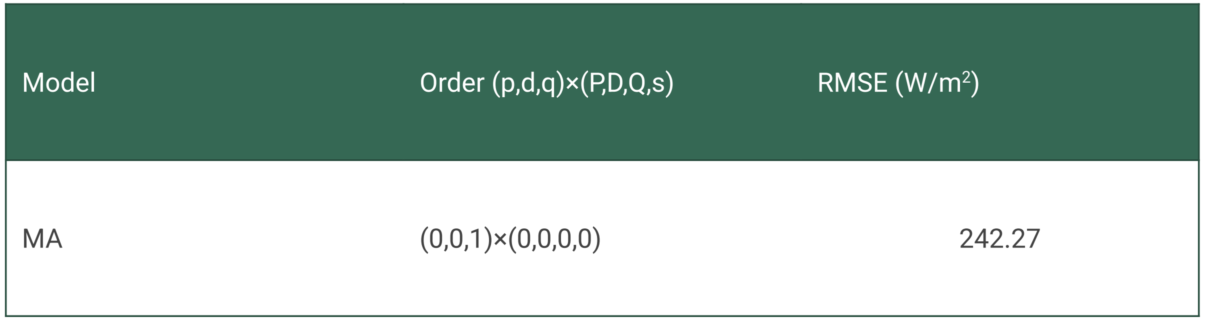
Model	Order (p,d,q)×(P,D,Q,s)	RMSE (W/m^2)
AR	(1,0,0)×(0,0,0,0)	90.63

Models for Forecasting

2. Moving Averages (MA):

The MA model assumes the current value is dependent on a linear combination of the current and past error terms (q=1) and assumes stationarity (d=0). The value of q is called the order of the MA model.

The MA model performed the worst, with a substantial RMSE of 242.27 W/m^2. The forecast line immediately drops to near zero and stays there for most of the year. In a simple MA(1) model, the long-term forecast tends toward the series mean, but for GHI, this mean is far from the seasonal peaks, leading to catastrophic errors.



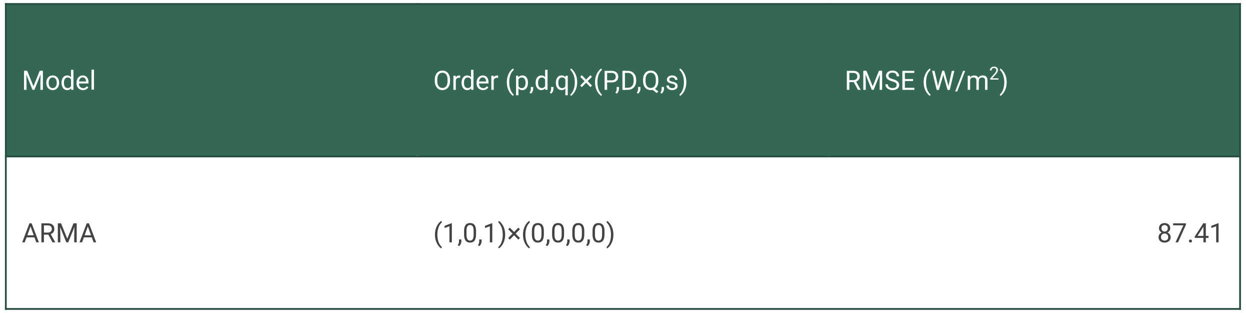
Model	Order (p,d,q)×(P,D,Q,s)	RMSE (W/m^2)
MA	(0,0,1)×(0,0,0,0)	242.27

Models for Forecasting

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.

With an RMSE of 87.41 W/m^2, the ARMA forecast is effectively a flat line. Since the GHI series is non-stationary and highly seasonal, the ARMA model immediately stabilizes to a constant value close to the mean of the training data or the start of the test period. This demonstrates its complete inadequacy for modeling data with strong annual cycles.



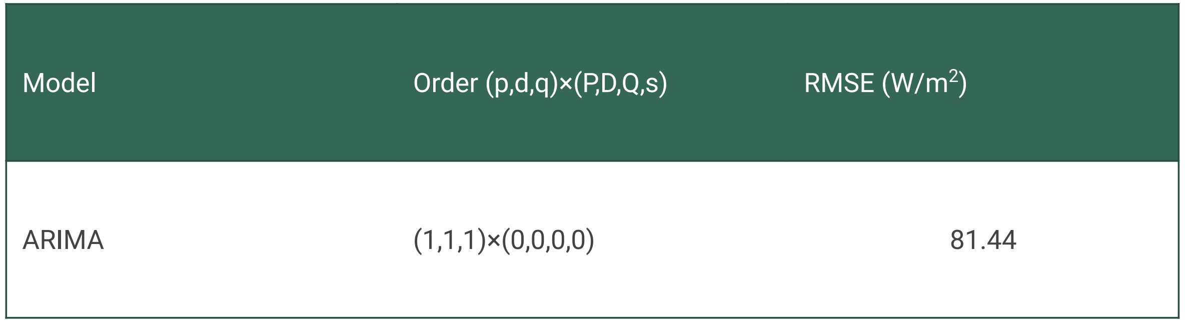
Model	Order (p,d,q)×(P,D,Q,s)	RMSE (W/m^2)
ARMA	(1,0,1)×(0,0,0,0)	87.41

Models for Forecasting

4. **ARIMA (Autoregressive Integrated Moving Average)**

   The ARIMA model accounts for non-stationarity through non-seasonal differencing (d=1) but lacks a dedicated seasonal component. Each ARIMA model uses three hyperparameters (p,d,q), of which p and q are similar to the ARMA model and d represents the number of times the data needs to be differenced to produce a stationary output.

   The RMSE of 81.44 W/m^2 is significantly higher than that of the SARIMA model. The forecast shows a slight downward slope but fails to capture the rising GHI through spring, the summer peak, or the steep decline in autumn. It acts primarily as a persistence forecast based on the last observed value after differencing, quickly missing the dominant annual pattern.



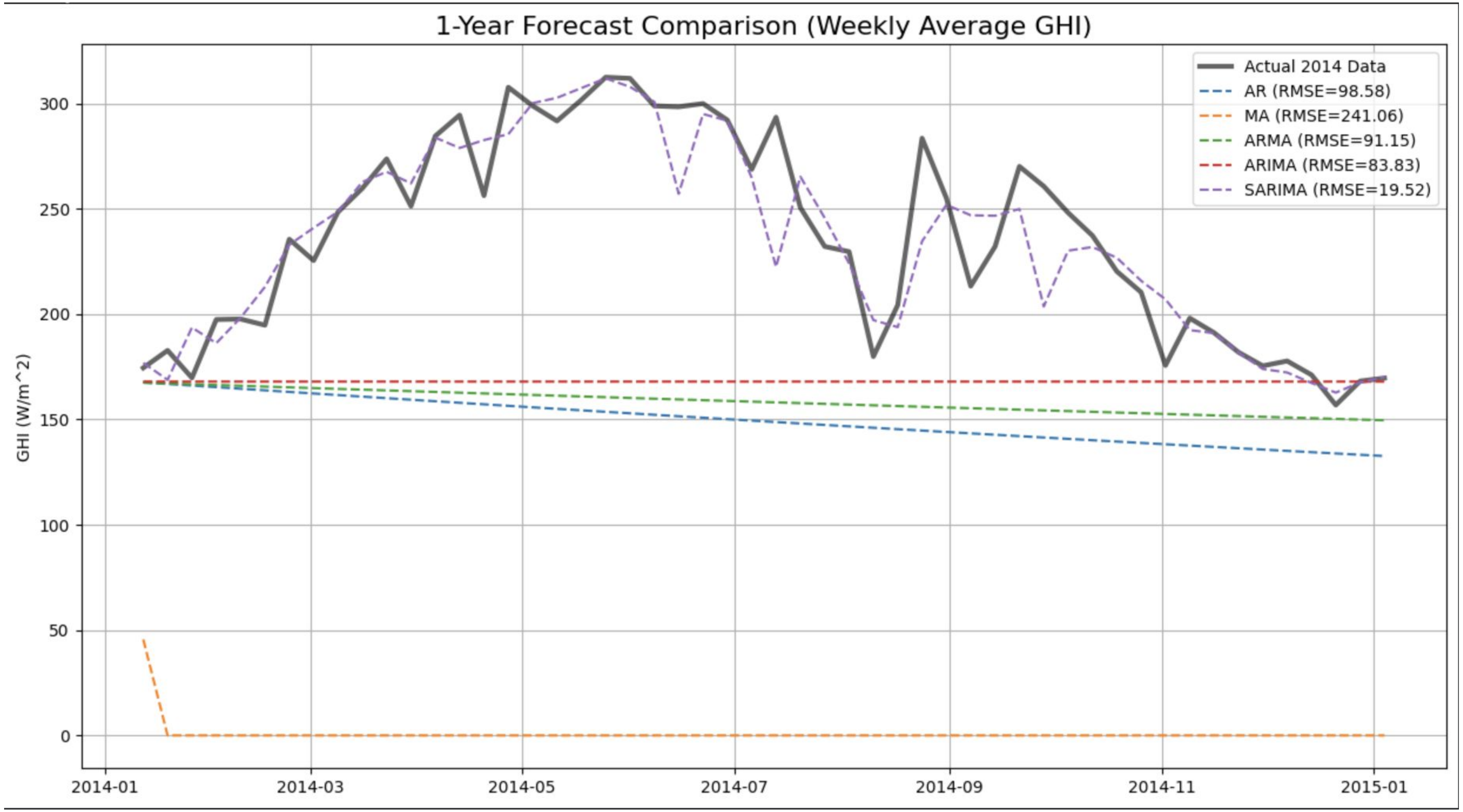
Model	Order (p,d,q)×(P,D,Q,s)	RMSE (W/m^2)
ARIMA	(1,1,1)×(0,0,0,0)	81.44

Models for Forecasting

5. **Seasonal Autoregressive Integrated Moving Average model (SARIMA):**
   The SARIMA model explicitly includes components to account for the strong annual seasonality (s=52 weeks) and non-stationarity of the GHI data, making it the most appropriate model type.
   The SARIMA model's forecast successfully captures both the overall trend (the annual solar cycle) and the week-to-week variations. Visually, the SARIMA forecast line closely tracks the actual GHI data, confirming the importance of the seasonal parameters (P=1, D=1, s=52) for this type of data.

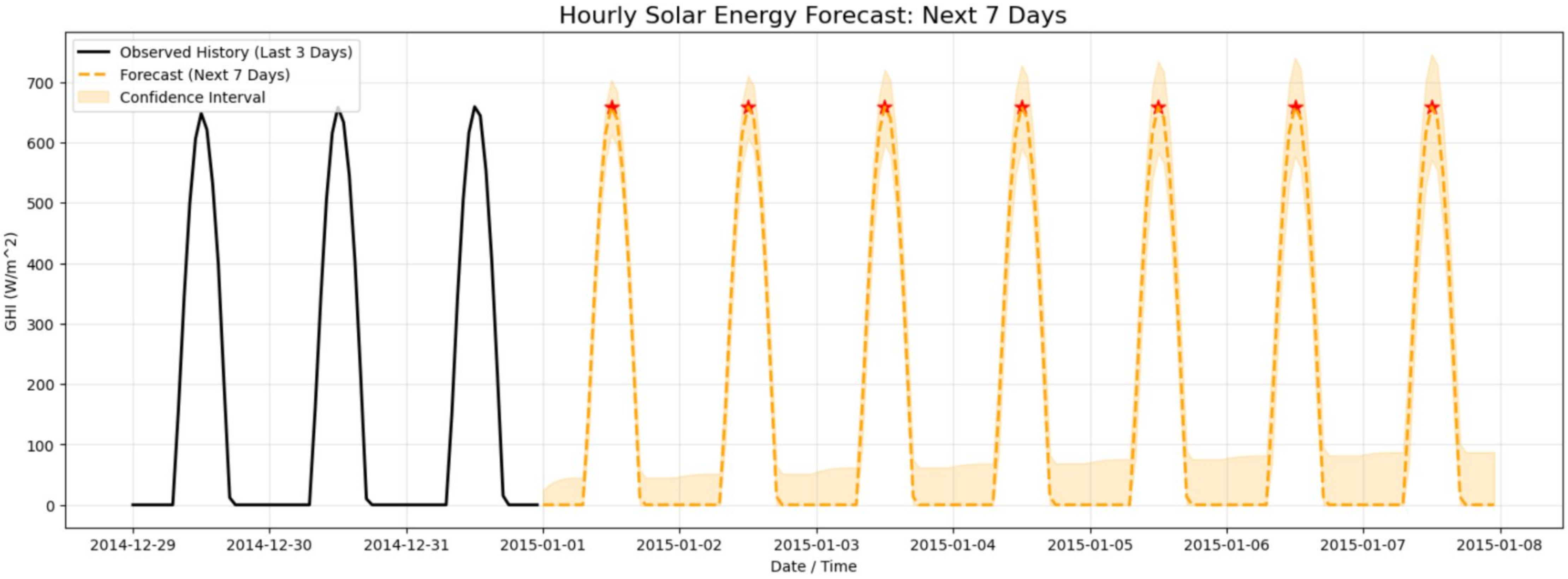
Model	Order (p,d,q)×(P,D,Q,s)	RMSE (W/m^2)
SARIMA	(1,1,1)×(1,1,0,52)	16.47

Model Selection



We compared five time series models : AR, MA, ARMA, ARIMA and SARIMA

Forecasting Results: Next week (Hourly)



The final model was retrained on the most recent data (last 60 days) to generate a high precision operational forecast for the next 7 days (168 hours)

Conclusion

- This study successfully analyzed 15 years of solar data for two regions in Rajasthan. The analysis confirmed that while Rajasthan has immense solar potential, it is subject to significant seasonal variance, particularly during the Monsoon.

- Weekly forecasting was found to be more accurate (with any model) as compared to daily forecasting. This is because daily data tends to have much more random variation as compared to weekly data.

- The **SARIMA** model proved to be the most effective tool for forecasting, capable of capturing both the daily "bell curve" and the annual seasonal shifts. The generated hourly forecasts for the upcoming week provide actionable insights for grid management and energy dispatch planning.

Thank You