Scientific Computing, Modeling & Simulation
Savitribai Phule Pune University

Master of Technology (M.Tech.)
Programme in Modeling and Simulation

Internship Project Report

Solar Irradiance Prediction using Satellite Data and Machine Learning.

Gaurav Prakash Lute
MT2313

Academic Year 2024-25

Scientific Computing, Modeling & Simulation
Savitribai Phule Pune University

Certificate

This is certify that this report titled

Solar Irradiance Prediction using Satellite Data and Machine Learning.

authored by

Gaurav Prakash Lute (MT2313)

describes the project work carried out by the author under our supervision during the period from January 2025 to June 2025. This work represents the project component of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Department of Scientific Computing, Modeling & Simulation, Savitribai Phule Pune University.

Dr.Abhay Kumar Tiwari Founder and CEO
SPATIALTY.AI Bangalore, India

Dr.Deepashri Saraf Program Manager
SPATIALTY.AI Bangalore, India

Dr.Bhalachandra Pujari,Internal Guide
SCMS-SPPU, Pune, India

Prof.Arun Banpurkar, Head
SCMS-SPPU, Pune, India

Scientific Computing, Modeling & Simulation
Savitribai Phule Pune University

Author's Declaration

Statement on Authorship. This document titled

Solar Irradiance Prediction using Satellite Data and Machine Learning.

authored by me is an authentic report of the project work carried out by me as part of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Department of Scientific Computing, Modeling & Simulation, Savitribai Phule Pune University.

Statement on Plagiarism. In writing this report, I have taken reasonable and adequate care to ensure that material borrowed from sources such as books, research papers, the internet, etc., is acknowledged as per accepted academic norms and practices in this regard. I have read and understood the University's policy on plagiarism (`http://unipune.ac.in/administration_files/pdf/Plagiarism_Policy_University_14-5-12.pdf` or the latest version thereof).

Statement on the use of LLMs/AI. I further certify that

1. any consultations with LLMs/AI tools have been duly acknowledged;
2. any material borrowed from LLMs/AI tools (including code, text, images, etc.) has been duly acknowledged; and
3. I understand that I am the one who is responsible for the work and the results presented in this report including any measures of accuracy or quality, etc.

Gaurav Prakash Lute
MT2313

Contents

Acknowledgements

I would like to begin by expressing my heartfelt gratitude to my industrial mentors, **Dr.Abhay Tiwari** and **Dr. Deepashri Saraf**, for their continuous guidance, valuable technical insights, and unwavering support throughout the course of this project. Their mentorship provided a strong bridge between theoretical knowledge and practical application, helping me navigate complex challenges with confidence.

I am equally grateful to my college guide, **Dr.Bhalchandra Pujari**, for his consistent support, expert guidance, and constructive feedback. His encouragement and academic supervision played a pivotal role in shaping the direction and quality of this research.

I would also like to extend my heartfelt thanks to all the other respectful professors for their insightful contributions and valuable input. Additionally, I am grateful to the staff members for their services and the assistance they offered throughout this process.

As part of this project, I made use of Large Language Models (LLMs) such as OpenAI's to assist with tasks like improving the clarity of my writing, better understanding technical concepts, and refining parts of the code.These tools were used carefully and responsibly to support my work, and everything generated through them was thoroughly reviewed and edited to ensure it truly reflected my own understanding and original effort.

Lastly, I would like to express my profound gratitude to my parents, friends, and colleagues for their unwavering support and belief in me. Their encouragement and understanding were vital to the completion of this thesis.

Abstract

This thesis presents a machine learning (ML) approach to predict Global Horizontal Irradiance using NASA POWER satellite data Feb 2016 to Feb 2025, crucial for solar energy planning. Various machine learning and forecasting models Random Forest, XGBoost, LightGBM, Prophet, ARIMA were developed, incorporating engineered features like lagged values and rolling statistics.Evaluated using RMSE and R^2, XGBoost achieved the best performance lowest RMSE, highest R^2, especially with a three-month training window. While ARIMA and Prophet captured long-term trends, they struggled with daily atmospheric variability. The research validates the effectiveness of Machine Learning with satellite data for accurate GHI forecasting, laying groundwork for real time applications and solar panel optimization, thus enhancing reliable solar resource utilization.

Chapter 1

Introduction

This study focuses on understanding how important solar energy reaches the Earth's face, a measure known as Global Vertical Irradiance(GHI)[3]. Directly vaticinating GHI is essential for the effective planning and operation of solar power systems. In this work, particular attention is given to prognosticating GHI for March 2025, a month that plays a crucial part in solar energy product due to seasonal transitions. To make and test the soothsaying models, the study uses long- term environmental data attained from NASA's POWER dataset, which contains diurnal atmospheric records from February 2016 to February 2025. This literal data served as the foundation for relating patterns in solar radiation and training machine literacy models to make dependable prognostications.

1.1 Domain background

In recent decades, the world has witnessed a growing shift toward sustainable and renewable energy sources. Among the various forms of clean energy, solar power has emerged as one of the most promising due to its abundance, low environmental impact, and scalability. However, solar energy generation is inherently variable, heavily influenced by atmospheric conditions such as cloud cover, aerosols, humidity, and temperature. These factors affect the amount of solar irradiance, which is the energy received from the sun on a given surface area over time.

Solar irradiance prediction plays a crucial role in a variety of domains including photovoltaic (PV) system design, energy production forecasting, grid management, and the integration of renewable energy into smart grid systems. Accurate irradiance forecasting allows power producers and grid operators to optimize solar energy use, manage storage, and improve reliability in power delivery.

Traditional methods for irradiance prediction rely on **numerical weather prediction(NWP)**[4] models or ground based meteorological stations. While useful,these methods face limitations such as:

- Sparse spatial coverage in many regions.
- High computational requirements for real time predictions.
- Limited ability to handle complex, non-linear relationships in meteorological data.

In response to these challenges, the integration of **satellite-derived data with machine learning (ML)** techniques has shown great potential. Satellite datasets, such as those from NASA's POWER project, offer global, consistent, and long term atmospheric and radiative data at various temporal and spatial resolutions. These datasets include important variables

like global horizontal irradiance (GHI), cloud cover, surface temperature, relative humidity, and wind speed, all of which influence the irradiance received at the Earth's surface.

Simultaneously, **machine learning algorithms** have advanced significantly and can now uncover hidden patterns in large and complex datasets. ML models like **Random Forests, Gradient Boosting, XGBoost, LightGBM, ARIMA, and Facebook Prophet** have been widely adopted for time series forecasting problems due to their ability to handle multivariate data and capture both linear and non-linear dependencies.

1.1.1 Solar Irradiance

Solar irradiance refers to the quantum of solar energy entered per unit area in the form of electromagnetic radiation within the wavelength range sensible by the measuring device. It is generally expressed in watts per square meter (W/m^2) according to the International System of Units (SI).

Solar irradiance can be measured either in space or at the Earth's surface. In space, it depends on how far the Sun is, changes in the solar cycle, and long-term variations. On Earth, it also depends on factors like the angle of the surface receiving sunlight, the Sun's position in the sky, and rainfall or atmospheric conditions. Solar irradiance also influences how plants grow and how animals behave.

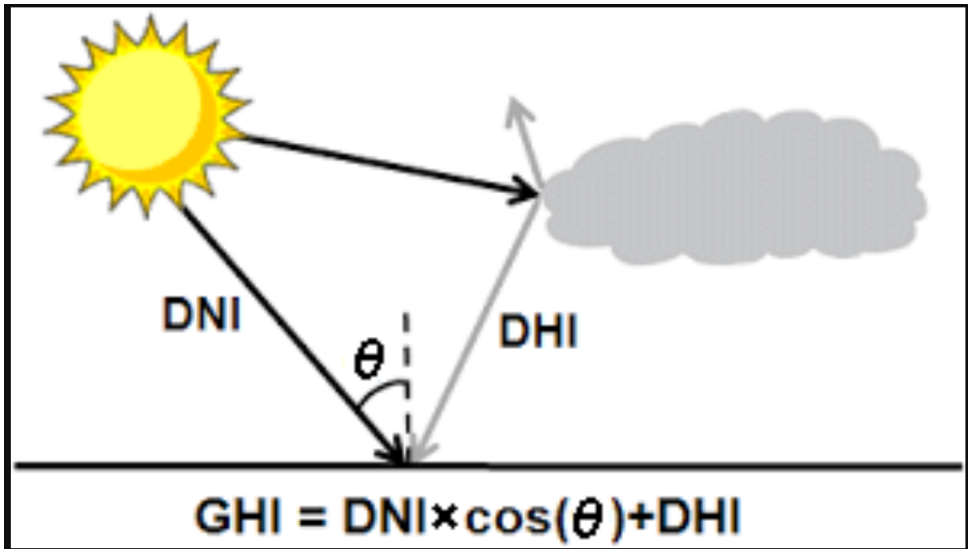

Figure 1.1: Illustration of Global Horizontal Irradiance (GHI) Calculation. This diagram shows GHI as the sum of Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI)

The images provide a comprehensive analysis of Global Horizontal Irradiance (GHI)[1] for solar energy applications. They first explain GHI as the sum of direct (DNI) and diffuse (DHI) sunlight, influenced by the sun's angle. The analysis then explores daily and seasonal patterns in multi-year irradiance data through time series plots and decomposition. While some decomposition plots appear flat (suggesting limited seasonality extraction), others, using manual monthly averages, clearly show seasonal trends and varied residuals.

Further exploration involves correlation heatmaps to identify relationships between GHI and other meteorological features (like temperature, humidity, wind), aiding in feature selection for predictive models. Boxplots are widely used to visualize data distributions and identify outliers across numerous features, from wind and pressure to temperature and various solar fluxes.

Finally, the images showcase the model's predictive power through forecast plots, comparing predicted GHI with actual values for periods like January-February and March 2025. These plots often include performance metrics like RMSE and R^2, quantifying the model's accuracy in capturing real-world irradiance patterns. In essence, the images demonstrate a systematic workflow for solar irradiance forecasting, encompassing fundamental understanding, data preparation, modeling, and evaluation.

1.1.2 Types Of Solar Irradiance

1. **Total Solar Irradiance (TSI):**

   TSI represents the total amount of solar energy received per unit area from all wavelengths of sunlight at the top of Earth's atmosphere. It is measured on a surface that is directly facing the Sun. The average value of TSI at a distance of one astronomical unit (the average distance from Earth to the Sun) is known as the solar constant.

2. **Direct Normal Irradiance (DNI):**

   DNI measures sunlight that comes directly from the Sun, without being scattered or reflected. It is taken on a surface that is always perpendicular (at a right angle) to the Sun rays. This value is affected by atmospheric conditions such as air particles, clouds, water vapor, and the position of the Sun in the sky. The more atmosphere sunlight travels through, the greater the losses due to scattering and absorption.

3. **Diffuse Horizontal Irradiance (DHI):**

   DHI measures sunlight that has been scattered by particles in the atmosphere and reaches the Earth's surface from all directions, excluding the direct beam from the Sun. It is recorded on a flat, horizontal surface. On a perfectly clear day with no atmosphere, DHI would be nearly zero.

4. **Global Horizontal Irradiance (GHI):**

   GHI is the total solar radiation received on a horizontal surface. It includes both direct sunlight and the scattered (diffuse) component. It can be calculated using the formula:

   $$\mathbf{GHI = DHI + DNI \times \cos(z)}$$

   Where z is the solar zenith angle the angle between the Sun and the vertical.

5. **Global Tilted Irradiance (GTI):**

   GTI refers to the total sunlight received on a surface that is tilted at a specific angle and direction (azimuth). This surface can be fixed or adjusted to track the Sun. GTI can be measured directly or estimated using values of GHI, DNI, and DHI. It is particularly important for solar panels, which are often installed at a tilt to capture more sunlight.

6. **Global Normal Irradiance (GNI):**

   GNI is the total amount of solar radiation received on a surface that is directly perpendicular to the Sun's rays, similar to how DNI is measured. However, GNI may include both direct and diffuse components, depending on the context.

This thesis leverages **NASA POWER satellite data (2016–2025)** and multiple ML algorithms to developed simple ML model and forecast**solar irradiance in March 2025**, with a focus on **Global Horizontal Irradiance (GHI)**. By using historical patterns and atmospheric variables, the models aim to generate accurate, short-term forecasts to support solar energy planning and utilization, especially during the transition month of March, when seasonal variability is pronounced.

1.2 Problem statement

Reliable solar irradiance data is vital for the effective design, planning, and operation of solar energy systems. Given the fluctuating and unpredictable nature of solar radiation, accurately forecasting and estimating irradiance is crucial for renewable energy deployment, grid integration, and strategic resource management.

1.2.1 Problem 1: Prediction of Solar Irradiance Using a Parametric Model.

This approach involves developing a simple parametric model based on known physical and meteorological relationships to estimate solar irradiance.To create an interpretable model that quantifies the relationship between meteorological variables and solar irradiance, allowing users to estimate irradiance under various atmospheric conditions.

1.2.2 Problem 2: Forecasting Solar Irradiance for March 2025 Using Historical Data (Time Series Forecasting).

This approach involves the use of historical satellite data February 2016 to February 2025 to **forecast solar irradiance values for March 2025** using machine learning and time series modeling techniques. To build a data-driven, non-parametric model that can accurately forecast future solar irradiance based on temporal patterns and long-term climatic behavior

Chapter 2

Data

2.1 Data description

The dataset used in this study is sourced from the **NASA POWER (Prediction of Worldwide Energy Resources)**[5] project, which provides global satellite derived meteorological and solar data intended for energy applications.The specific dataset selected covers the time period from February 28, 2016 to February 28, 2025, and corresponds to the geographic location of **Pune, India (Latitude: 18.52°N, Longitude: 73.86°E)**. The data is collected at daily resolution, making it suitable for time series forecasting of solar irradiance.

The primary focus of this study is the prediction of Global Horizontal Irradiance (GHI), represented by the feature **ALLSKY_SFC_SW_DWN** the total shortwave solar radiation received at the Earth's surface under all sky conditions. Several other meteorological and atmospheric features are used as predictors in the model.

2.1.1 Modeling Features

Solar Radiation Parameters

- **ALLSKY_SFC_SW_DWN:** All-sky surface downward shortwave radiation (W/m2).This is the total solar radiation reaching the surface under all sky conditions (including clouds).
- **CLRSKY_SFC_SW_DWN:** Clear-sky surface downward shortwave radiation (W/m2). This is the total solar radiation reaching the surface under clear sky conditions (no clouds).
- **ALLSKY_SFC_SW_DNI:** All-sky surface downward direct normal irradiance (W/m2). This is the direct solar radiation reaching the surface perpendicular to the sun's rays.
- **ALLSKY_SFC_SW_DIFF:** All-sky surface downward diffuse shortwave radiation (W/m2). This is the scattered solar radiation reaching the surface from all directions.
- **TOA_SW_DWN:** Top of atmosphere downward shortwave radiation (W/m2). This is the total solar radiation reaching the top of the atmosphere.
- **ALLSKY_SFC_PAR_TOT / CLRSKY_SFC_PAR_TOT:** All-sky surface total photosynthetically active radiation (mol/m2/day). This is the portion of solar radiation that plants use for photosynthesis.
- **ALLSKY_SFC_UVA / ALLSKY_SFC_UVB:** Ultraviolet radiation components A and B.

Temperature and Thermal Features

- **T2M:** Temperature at 2 meters above the surface (°C).
- **TT2MDEW:** Dew point temperature at 2 meters above the surface (°C).
- **T2MWET:** Wet bulb temperature at 2 meters above the surface (°C).

Humidity and Precipitation

- **QV2M:** Specific humidity at 2 meters above the surface (kg/kg).
- **RH2M:** Relative humidity at 2 meters above the surface.
- **PRECTOTCORR:** Total precipitation corrected (mm/day).

Wind and Atmospheric Pressure

- **PS:** Surface pressure (kPa).
- **WS2M:** Wind speed at 2 meters above the surface (m/s).
- **WD2M:** Wind direction at 2 meters above the surface (degrees).
- **WS10M:** Wind speed at 10 meters above the surface (m/s).
- **WD10M:** Wind direction at 10 meters above the surface (degrees).

Time Related Features

- **Year:** Year Of Observation.
- **DOY:** Day of Year (ranging from 1 to 365)

2.2 Exploratory data analysis

The purpose of the EDA is to understand the behavior and relationships between variables and to detect patterns or anomalies that influence the GHI.

2.2.1 Performed Univariate Analysis

Distribution of Solar Flux parameters

The below figure 2.1 displays nine histograms, each providing a univariate analysis of a solar flux parameter by showing its frequency distribution observe various shapes.**Unimodal** distributions (like ALLSKY_SFC_SW_DWN, ALLSKY_SFC_SW_DIFF, ALLSKY_SFC_PAR_TOT, ALLSKY_SFC_UVA) with a single peak, indicating a common range for most values;**Bimodal** distributions (such as CLRSKY_SFC_SW_DWN, ALLSKY_SFC_SW_DNI, CLRSKY_SFC_PAR_TOT) with two distinct peaks, suggesting two prevalent conditions or groups within the data. and a U-shaped distribution (TOA_SW_DWN) where extreme values are more frequent than central ones.The ALLSKY_SFC_UVB parameter shows a discrete, quantized distribution with distinct spikes at specific values, implying measurements are recorded in fixed increments. No truly uniform distributions, where frequencies are roughly equal across all values, are present in this dataset.

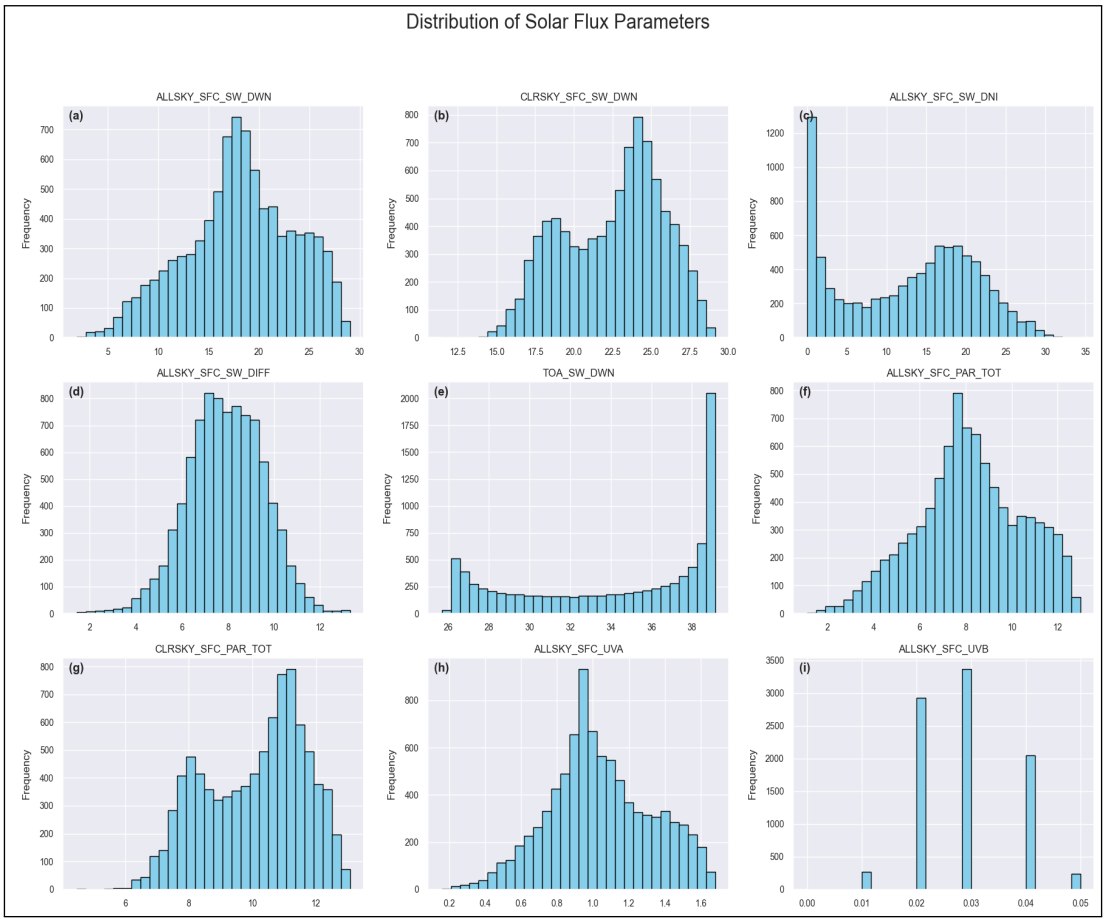

Figure 2.1: Distribution of Solar Flux Parameters. This figure displays histograms for various solar flux measurements. (a) All-Sky Surface Shortwave Downward Radiation. (b) Clear-Sky Surface Shortwave Downward Radiation. (c) All-Sky Surface Shortwave Direct Normal Irradiance. (d) All-Sky Surface Shortwave Diffuse Radiation. (e) Top of Atmosphere Shortwave Downward Radiation. (f) All-Sky Surface Photosynthetically Active Radiation Total. (g) Clear-Sky Surface Photosynthetically Active Radiation Total.(h) All-Sky Surface UVA Radiation. (i) All-Sky Surface UVB Radiation.

Temperature and Thermal IR Flux

The four histograms illustrate the distribution of meteorological parameters: T2M, T2MDEW, and T2MWET, each depicting temperature-related data at a 2-meter height. T2M, shown in the top left, represents air temperature in degrees Celsius, displaying a right-skewed pattern with most values between 15°C and 25°C, peaking near 25°C, and fewer instances above 30°C, suggesting rarer high temperatures. The top right histogram, T2MDEW, indicates dew point temperature, also right-skewed, with a peak around 15°C to 20°C and a wide range from -10°C to 25°C, reflecting significant moisture variability. T2MWET, in the bottom left, shows wet bulb temperature, a heat stress indicator influenced by humidity, with values mostly between 7.5°C and 20°C, peaking at 17.5°C, lower than T2M due to evaporative cooling. All distributions exhibit right-skewness, with T2MDEW showing the broadest spread, highlighting diverse humidity levels, while T2M and T2MWET suggest more consistent temperature patterns, useful for assessing climate, heat stress, and moisture in a specific region or timeframe.

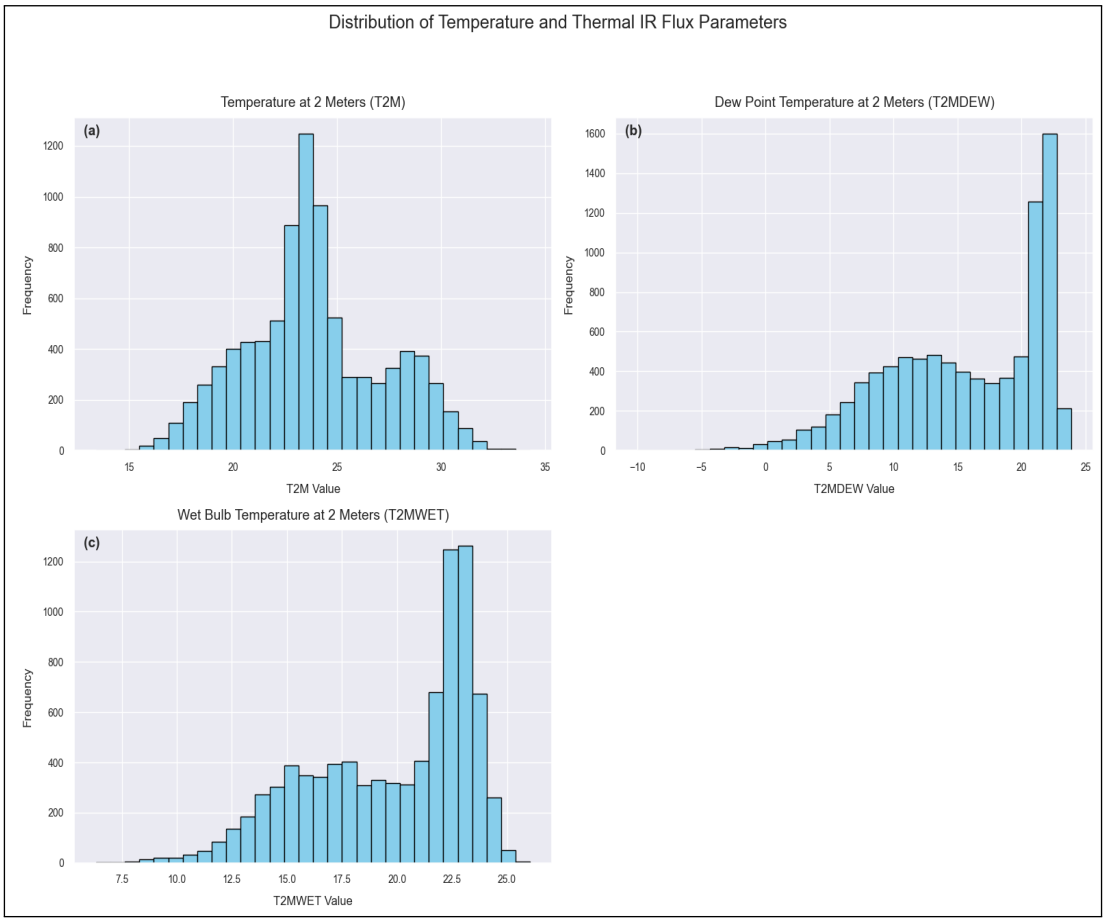

Figure 2.2: Distribution of Temperature and Thermal IR Flux Parameters.This figure presents histograms illustrating the frequency distribution of various temperature-related parameters.(a)Histogram for Temperature at 2 Meters(T2M).(b)Histogram for Dew Point Temperature at 2 Meters(T2MDEW).(c) Histogram for Wet Bulb Temperature at 2 Meters (T2MWET).

Humidity and Precipitation

When analyzing individual data points for humidity and precipitation, we observe distinct patterns.Specific humidity often presents two common levels,possibly reflecting different environmental states like varying seasons or times of day.Relative humidity predominantly remains at very high levels, indicating a consistently moist atmosphere,with lower values being infrequent. For precipitation,the data overwhelmingly shows periods of no rainfall, and when rain does occur,it is typically in small quantities, with heavy downpours being rare. These separate examinations of each variable offer foundational insights into the typical conditions and range of events for humidity and rainfall within the dataset.

Wind and Pressure parameters

The wind and pressure feature histograms (PS, WS2M, WD2M, WS10M, WD10M) from the NASA Power Dataset, as part of the solar irradiance prediction study (Page 8). The PS histogram (surface pressure, kPa) shows a near-normal distribution, peaking around 93.8 kPa, indicating stable pressure conditions. WS2M and WS10M (wind speed at 2 and 10 meters, m/s) both exhibit right-skewed distributions, with WS2M ranging from 0 to 10 m/s (peaking near 3 m/s) and WS10M from 0 to 12 m/s (peaking near 4 m/s), where WS10M shows slightly

higher speeds, as expected at greater height. WD2M and WD10M (wind direction at 2 and 10 meters, degrees) display multimodal distributions with peaks at 50–100° and 250–300°, reflecting dominant wind directions. These patterns align with the study's findings of a strong correlation between WS2M and WS10M (0.99) and WD2M and WD10M (0.97), as well as a moderate negative correlation between PS and wind speeds (-0.67 for WS2M, -0.63 for WS10M), suggesting lower pressure may drive higher winds, potentially affecting solar irradiance through atmospheric dynamics.

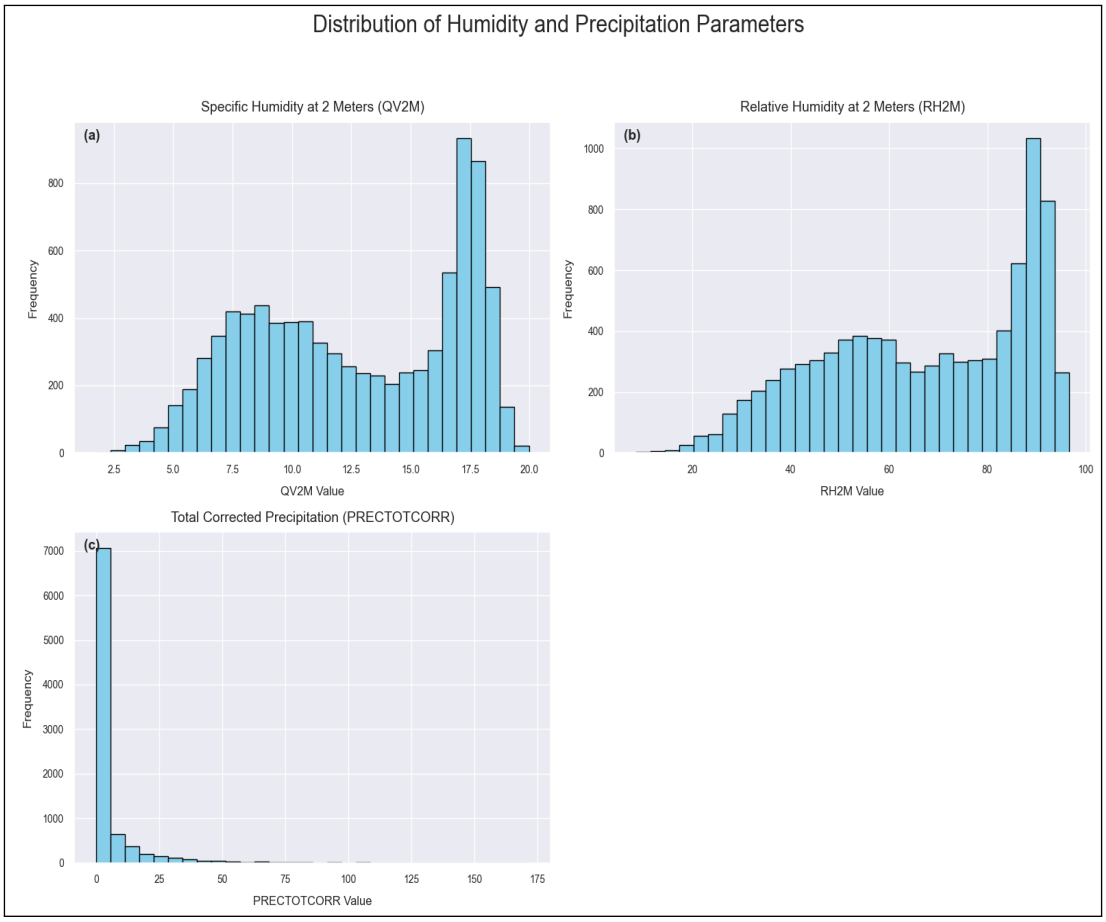

Figure 2.3: This figure presents histograms illustrating the frequency distribution of various humidity and precipitation measurements.(a)Histogram for Specific Humidity at 2 Meters(QV2M).(b)Histogram for Relative Humidity at 2 Meters(RH2M).(c)Histogram for Total Corrected Precipitation (PRECTOTCORR).

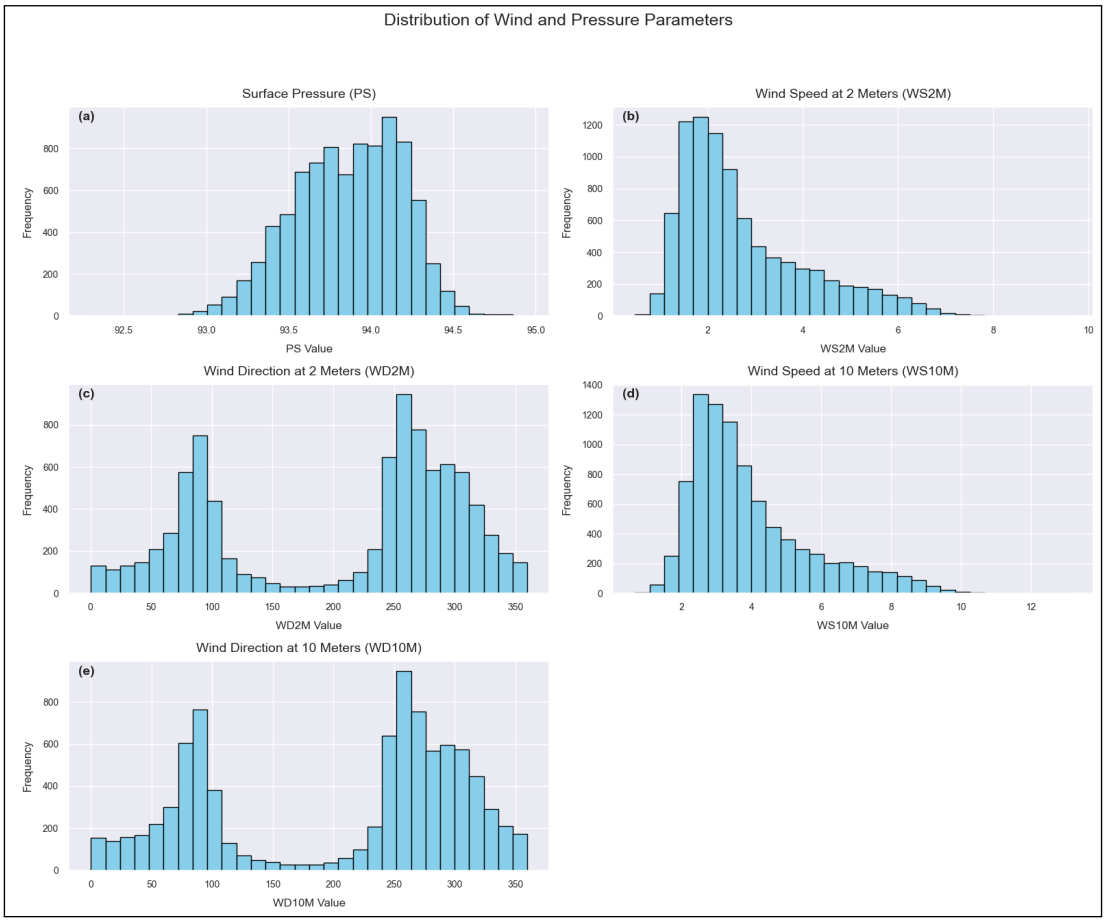

Figure 2.4: This figure displays histograms illustrating the frequency distribution of various wind and pressure measurements.(a) Histogram for Surface Pressure (PS).(b) Histogram for Wind Speed at 2 Meters (WS2M).(c) Histogram for Wind Direction at 2 Meters (WD2M).(d) Histogram for Wind Speed at 10 Meters (WS10M).(e) Histogram for Wind Direction at 10 Meters (WD10M).

2.2.2 Bivariate Analysis

Bivariate analysis is a way of studying how two variables are related to each other. It helps us understand whether a change in one variable affects or is connected to a change in the other. This type of analysis is useful for spotting patterns, trends, or relationships between the two variables. Some common techniques used in bivariate analysis include scatter plots, correlation measures, and regression analysis, which help visualize and measure the strength and direction of the relationship. Common methods include scatter plots, correlation coefficients, and regression analysis. For instance, in studies predicting solar irradiance, bivariate analysis can be used to examine how atmospheric factors like temperature or humidity impact solar radiation, thereby uncovering dependencies that enhance predictive model accuracy.

Explain Each of the Scatter Plot

- **Temperature at 2 Meters (T2M) vs. Solar Irradiance.**

  The scatter plot for T2M reveals a fascinating, non-linear relationship. Initially, as temperature rises up to approximately 27^0C, solar irradiance also increases, peaking around 30 MJ/m^2/day. However, beyond this point, higher temperatures (e.g., 35^0C) see a de-

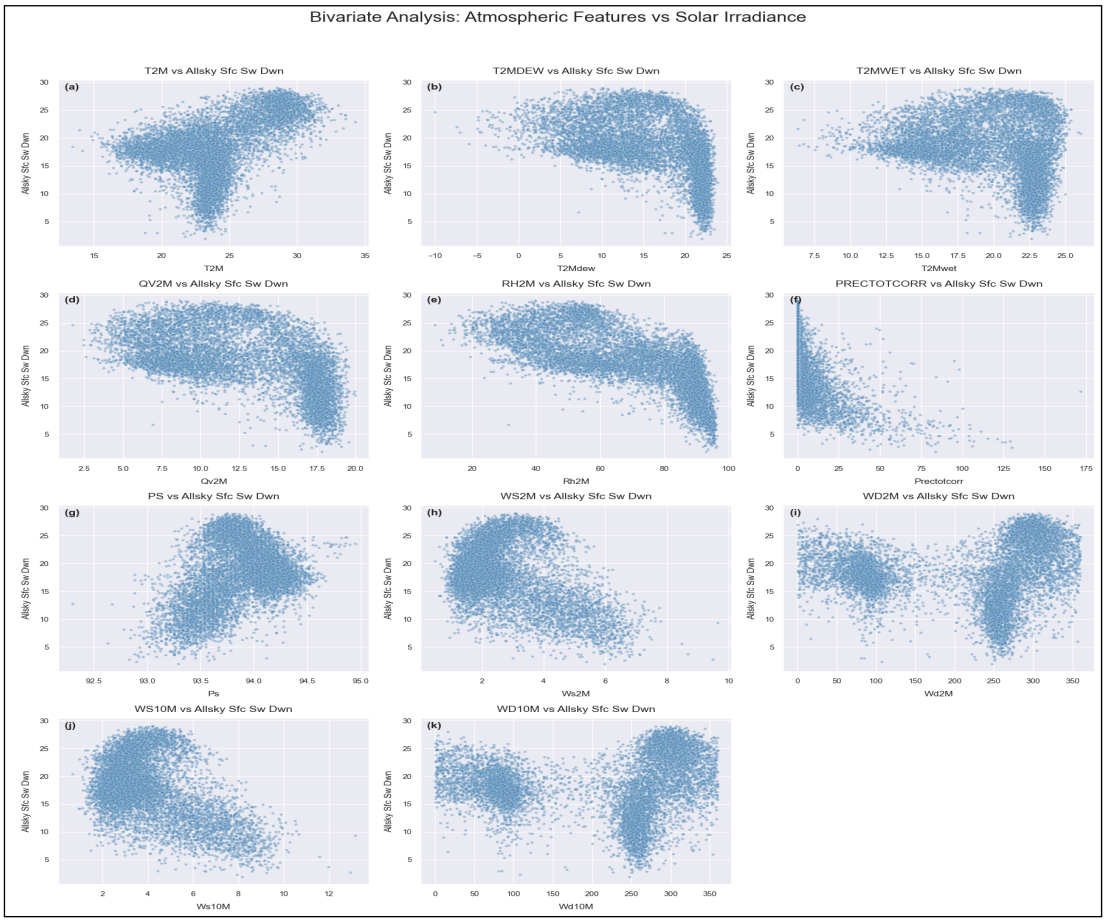

Figure 2.5: Bivariate Analysis of Atmospheric Features vs. All-Sky Surface Shortwave Downward Radiation. This figure displays scatter plots showing the relationship between various atmospheric parameters and solar irradiance(ALLSKY_SFC_SW_DWN).(a) Scatter plot of Temperature at 2 Meters(T2M) vs. ALLSKY_SFC_SW_DWN.(b)Scatter plot of Dew Point Temperature at 2 Meters(T2MDEW) vs. ALLSKY_SFC_SW_DWN.(c) Scatter plot of Wet Bulb Temperature at 2 Meters (T2MWET) vs. ALLSKY_SFC_SW_DWN.(d) Scatter plot of Specific Humidity at 2 Meters (QV2M) vs. ALLSKY_SFC_SW_DWN.(e) Scatter plot of Relative Humidity at 2 Meters (RH2M) vs. ALLSKY_SFC_SW_DWN.(f) Scatter plot of Total Corrected Precipitation (PRECTOTCORR) vs. ALLSKY_SFC_SW_DWN.(g) Scatter plot of Surface Pressure (PS) vs. ALLSKY_SFC_SW_DWN.(h) Scatter plot of Wind Speed at 2 Meters (WS2M) vs. ALLSKY_SFC_SW_DWN.(i) Scatter plot of Wind Direction at 2 Meters (WD2M) vs. ALLSKY_SFC_SW_DWN.(j) Scatter plot of Wind Speed at 10 Meters (WS10M) vs. ALLSKY_SFC_SW_DWN.(k) Scatter plot of Wind Direction at 10 Meters(WD10M) vs. ALLSKY_SFC_SW_DWN.

cline or plateau in irradiance, dropping to about 5 MJ/m^2/day. This suggests that while moderate warmth often correlates with clearer skies and thus more sunlight, extremely high temperatures might be associated with atmospheric conditions like cloud formation or haze, which then diminish solar radiation. This observed non-linearity supports the use of a quadratic term in predictive models to accurately capture this complex interaction.

- **Dew Point Temperature at 2 Meters (T2MDEW) vs. Solar Irradiance**

  A distinct bell-shaped curve characterizes the relationship between T2MDEW and solar irradiance. Irradiance reaches its highest values (25–30 MJ/m^2/day) at moderate dew

point temperatures, typically between 0^0C and 10^0C. In contrast, higher dew points (e.g., 20^0C) correspond to significantly lower irradiance, plummeting to around 5 MJ/m^2/day. This pattern suggests that optimal moisture levels are conducive to clear skies and maximum solar radiation, while elevated dew points indicate increased atmospheric humidity, which can lead to greater cloud cover and reduced solar irradiance. This makes T2MDEW a valuable proxy for humidity in modeling efforts.

- **Wet Bulb Temperature at 2 Meters (T2MWET) vs. Solar Irradiance**

  The scatter plot for T2MWET exhibits a trend similar to that of T2M. Solar irradiance increases with wet bulb temperature up to approximately 17.5^0C, reaching peaks of 30 MJ/m^2/day. Beyond this threshold, at higher wet bulb temperatures (e.g., 25^0C), irradiance tends to plateau or slightly decrease, falling to about 5 MJ/m^2/day. Since wet bulb temperature accounts for both temperature and humidity, this trend implies that very high wet bulb temperatures might signify humid conditions that impede solar radiation from reaching the surface, thus highlighting its relevance for predictive modeling.

- **Specific Humidity at 2 Meters (QV2M) vs. Solar Irradiance**

  A strong negative correlation is evident between QV2M and solar irradiance. As specific humidity increases from 2.5 to 20 kg/kg, solar irradiance sharply declines from 30 MJ/m^2/day to approximately 5 MJ/m^2/day. This robust inverse relationship indicates that greater water vapor content in the atmosphere is strongly associated with increased cloud formation, which in turn scatters solar radiation and significantly reduces the amount of irradiance reaching the ground. To accurately represent this relationship in models, a log transformation or polynomial feature might be beneficial.

- **Relative Humidity at 2 Meters (RH2M) vs. Solar Irradiance**

  The scatter plot for RH2M also shows a pronounced negative trend. At lower relative humidity levels (around 20%), solar irradiance can be as high as 30 MJ/m^2/day. Conversely, at high relative humidity values (85–95%), irradiance drops dramatically to 5 MJ/m^2/day. This strong inverse relationship underscores that air nearing saturation, characteristic of high relative humidity, is directly linked to increased cloudiness and a substantial reduction in solar irradiance. Consequently, RH2M stands out as a powerful predictor in solar irradiance modeling.

- **Precipitation (PRECTOTCORR) vs. Solar Irradiance**

  The relationship between PRECTOTCORR and solar irradiance is markedly negative and exhibits a left-skewed distribution. The majority of data points show zero precipitation correlating with high irradiance (up to 30 MJ/m^2/day). However, as precipitation increases, even to levels like 175 mm/day, solar irradiance sharply drops to near zero. This clearly demonstrates that rainfall directly impedes sunlight by means of clouds and the rain itself. Given the skewness of the data, a log transformation is recommended for this feature in predictive models.

- **Surface Pressure (PS) vs. Solar Irradiance**

  The scatter plot for PS indicates a largely flat relationship with solar irradiance. Irradiance values are broadly scattered across the pressure range (92.5–95 kPa), fluctuating between 5 and 30 MJ/m^2/day without a discernible trend. This suggests that surface pressure has minimal to no direct influence on solar irradiance, justifying its exclusion from predictive models due to its limited predictive utility.

- **Wind Speed at 2 Meters (WS2M) vs. Solar Irradiance**

  A convex pattern emerges in the relationship between WS2M and solar irradiance. Irradiance peaks (around 25–30 MJ/m²/day) at moderate wind speeds (3–6 m/s). Both very low wind speeds (0 m/s) and very high wind speeds (10 m/s) correlate with reduced irradiance. This suggests that moderate winds might help disperse clouds, thereby enhancing solar irradiance. In contrast, stagnant air (low wind) or stormy conditions (high wind) could lead to decreased solar radiation. A polynomial feature is suggested to capture this observed non-linearity.

- **Wind Speed at 10 Meters (WS10M) vs. Solar Irradiance**

  Similar to WS2M, the scatter plot for WS10M displays a convex pattern. Peak solar irradiance is observed at moderate wind speeds (4–6 m/s), with irradiance decreasing at higher wind speeds (e.g., 12 m/s). This parallel trend reinforces the idea that moderate winds at higher altitudes also contribute to clearer skies and improved solar irradiance by clearing cloud cover.

- **Wind Direction at 2 Meters (WD2M) vs. Solar Irradiance**

  The relationship between WD2M and solar irradiance appears complex and cyclic. Irradiance values vary across the full 0–360° range of wind directions, with notable clusters of higher irradiance observed at approximately 50–100° and 250–300°. This complexity suggests that wind direction influences irradiance indirectly through regional weather patterns. To properly incorporate the circular nature of this feature into a model, sine and cosine transformations are recommended.

- **Wind Direction at 10 Meters (WD10M) vs. Solar Irradiance**

  The scatter plot for WD10M mirrors the observations from WD2M, exhibiting a similar cyclic pattern with irradiance clusters in comparable directional ranges. This consistency across different altitudes further emphasizes the need for transformations, such as sine and cosine, to effectively capture the directional effects of wind on solar irradiance in predictive models.

2.2.3 Handle Outlier and Analysis

To effectively use the boxplot distribution for outlier handling in the solar irradiance prediction study, we can apply techniques to identify, analyze, and manage outliers for the solar flux and related features.

This analysis highlights that features such as ALLSKY_SFC_SW_DWN (all-sky surface downward shortwave radiation) and ALLSKY_SFC_SW_DIFF (all-sky surface diffuse shortwave radiation) exhibit a higher percentage of outliers (0.32% and 0.80% respectively),indicating greater variability likely influenced by dynamic atmospheric conditions like cloud cover and storms.

In contrast, parameters like TOA_SW_DWN (top of atmosphere downward shortwave radiation) and ALLSKY_SFC_UVB (all-sky surface downward UVB radiation) demonstrate more stable distributions with no reported outliers. Understanding these extreme values and their frequency is crucial for data preprocessing, as it ensures that unusual observations do not unduly influence the performance of predictive models.

The boxplot visualizations for T2M (temperature at 2 meters), T2MDEW (dew point temperature at 2 meters), and T2MWET (wet bulb temperature at 2 meters) from the NASA Power

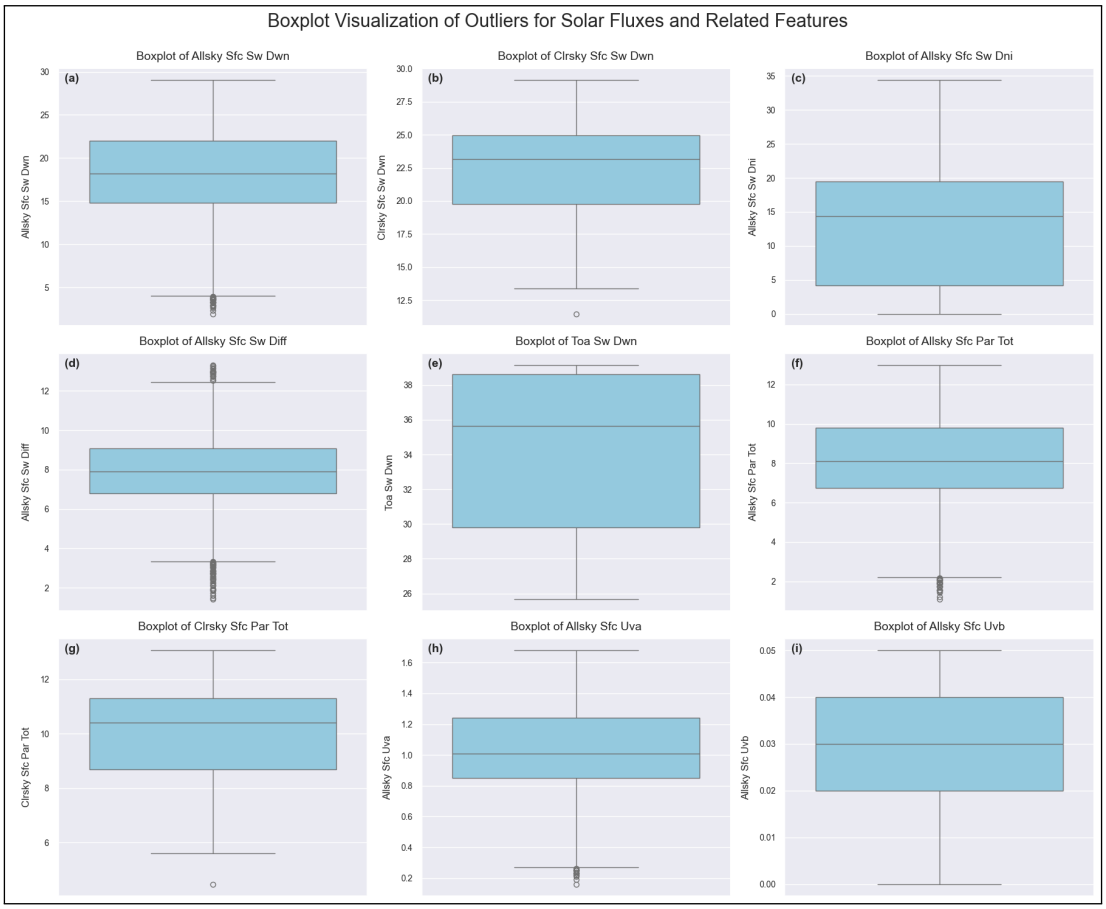

Figure 2.6: . This figure presents box plots illustrating the distribution and outliers for various solar flux parameters.(a) Boxplot for All-Sky Surface Shortwave Downward Radiation (Allsky Sfc Sw Dwn).(b) Boxplot for Clear-Sky Surface Shortwave Downward Radiation (Clrsky Sfc Sw Dwn).(c) Boxplot for All-Sky Surface Shortwave Direct Normal Irradiance (Allsky Sfc Sw Dni).(d) Boxplot for All-Sky Surface Shortwave Diffuse Radiation (Allsky Sfc Sw Diff).(e) Boxplot for Top of Atmosphere Shortwave Downward Radiation (Toa Sw Dwn).(f) Boxplot for All-Sky Surface Photosynthetically Active Radiation Total (Allsky Sfc Par Tot).(g) Boxplot for Clear-Sky Surface Photosynthetically Active Radiation Total (Clrsky Sfc Par Tot).(h) Boxplot for All-Sky Surface UVA Radiation (Allsky Sfc Uva).(i) Boxplot for All-Sky Surface UVB Radiation (Allsky Sfc Uvb).

Dataset provide critical insights into the distribution and presence of outliers within these temperature and thermal infrared flux features, as detailed on page 11 of the solar irradiance prediction study. While T2M generally shows a stable temperature range with a small percentage of outliers (0.28%) indicating rare extremes, T2MDEW exhibits wider variability in moisture levels, with a very small number of outliers (0.06%) representing exceptionally dry or humid conditions. T2MWET, which combines temperature and humidity effects, demonstrates moderate variability with a similarly low outlier percentage (0.09%), pointing to infrequent extreme heat stress. The presence of these outliers, though few, highlights the occasional occurrence of extreme atmospheric conditions that can significantly impact solar irradiance predictions.

The boxplot visualizations for specific humidity (QV2M), relative humidity (RH2M), and precipitation (PRECTOTCORR) reveal distinct data distributions and outlier patterns. QV2M exhibits a concentrated range of values around its median with no significant outliers, indicat-

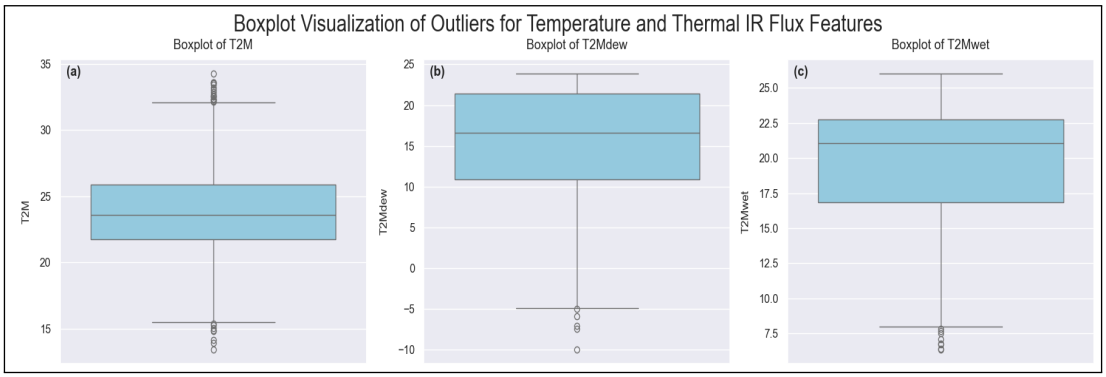

Figure 2.7: This figure displays box plots illustrating the distribution and outliers for various temperature and thermal infrared flux parameters.(a) Boxplot for Temperature at 2 Meters (T2M).(b) Boxplot for Dew Point Temperature at 2 Meters (T2MDEW).(c) Boxplot for Wet Bulb Temperature at 2 Meters(T2MWET).

ing a consistent spread of specific humidity within expected bounds. Similarly, RH2M shows a

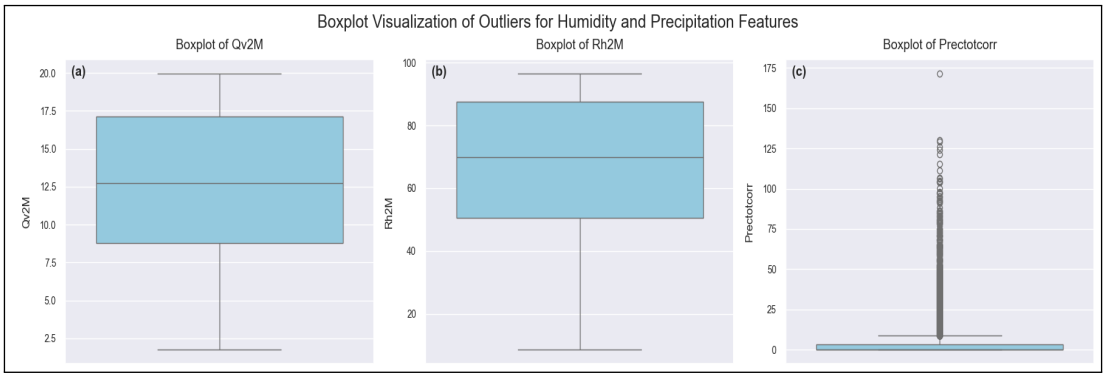

Figure 2.8: Boxplot Visualization of Outliers for Humidity and Precipitation Features. This figure presents box plots illustrating the distribution and potential outliers for various humidity and precipitation parameters.(a) Boxplot for Specific Humidity at 2 Meters (Qv2M).(b) Boxplot for Relative Humidity at 2 Meters (Rh2M).(c) Boxplot for Total Corrected Precipitation (Pretotcorr).

broad yet outlier-free distribution, signifying that a wide spectrum of relative humidity conditions, from moderately dry to very humid, are common occurrences. In stark contrast, PRECTOTCORR is severely right-skewed, with the vast majority of data points indicating zero or near-zero precipitation, but a substantial number of high-value outliers. This highlights that while dry periods dominate, rainfall, when it occurs, can be highly variable and intense, registering as statistical outliers relative to the prevailing conditions.

The boxplot analysis of wind and pressure features reveals distinct distributional patterns and outlier characteristics. Surface pressure (PS) demonstrates remarkably low variability, with a tightly concentrated interquartile range and a few notable outliers indicating infrequent extreme high or low pressure events. Conversely, both wind speed at 2 meters (WS2M) and 10 meters (WS10M) exhibit small interquartile ranges for typical values, but a significant number of outliers extending to much higher speeds, suggesting frequent occurrences of strong winds or gusts. In stark contrast, wind direction at both 2 meters (WD2M) and 10 meters (WD10M) shows broad, uniform distributions across all degrees, with no apparent outliers. This indicates

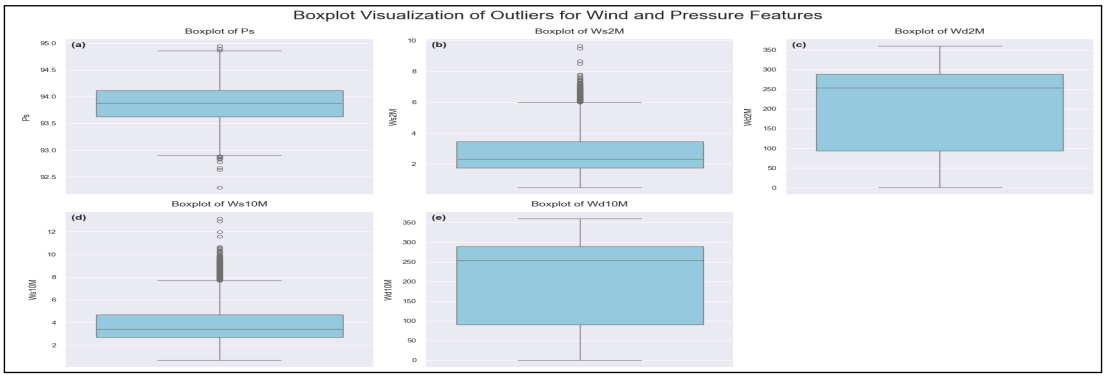

Figure 2.9: Boxplot Visualization of Outliers for Wind and Pressure Features. This figure presents box plots illustrating the distribution and potential outliers for various wind and pressure parameters.(a) Boxplot for Surface Pressure (Ps).(b) Boxplot for Wind Speed at 2 Meters (Ws2M).(c) Boxplot for Wind Direction at 2 Meters (Wd2M).(d) Boxplot for Wind Speed at 10 Meters (Ws10M).(e) Boxplot for Wind Direction at 10 Meters (Wd10M).

that wind direction is highly variable, with all directions being regularly observed rather than any specific direction being anomalous. In essence, while pressure remains stable with rare extremes, wind speeds frequently deviate to higher values, and wind directions are consistently diverse.

2.2.4 Time Series Data Analysis

The chart clearly illustrates the prominent seasonal variations in daily solar irradiance, alongside considerable short term fluctuations attributable to meteorological conditions. This combination of a regular annual cycle and unpredictable daily changes underscores the intricate character of solar irradiance data, emphasizing the necessity for models capable of encompassing both extensive cyclical trends and immediate atmospheric impacts.

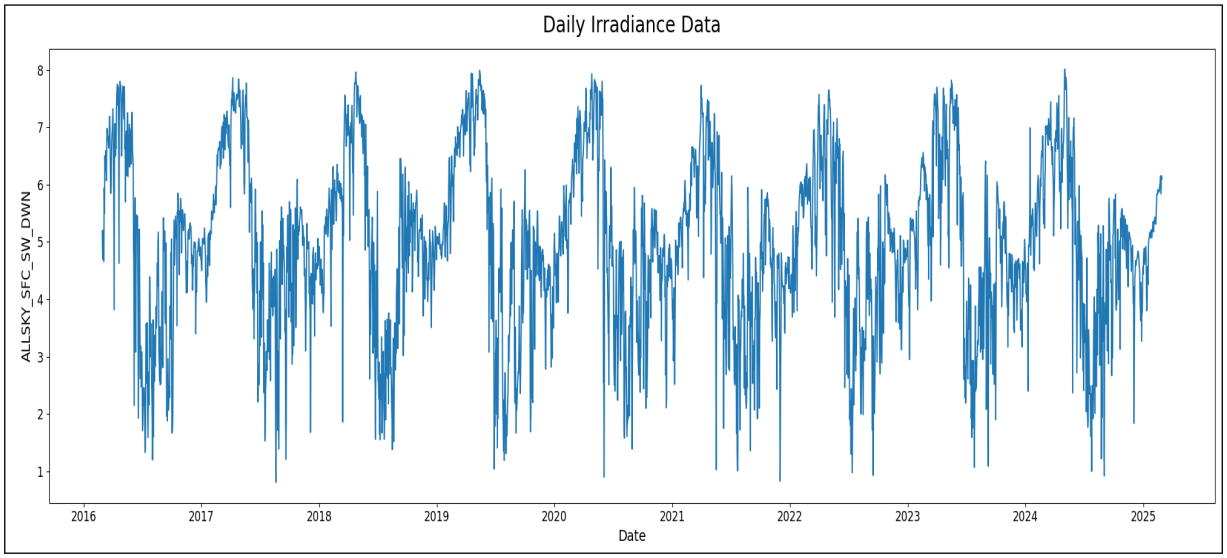

Figure 2.10: Daily Irradiance Data.This line plot illustrates the daily variation of all-sky surface downward shortwave radiation over time, showing seasonal patterns and overall trends.

Seasonal decompose Multiplicative model

Time series decomposition is a method for breaking down a series, Y_t into its core constituent patterns. The enduring Trend(T_t), predictable Seasonality(S_t),and random Residuals(R_t).These elements can be integrated through an Additive Model, $Y_t = T_t + S_t + R_t$,which is appropriate when the scale of seasonal shifts remains uniform,or a Multiplicative Model, $Y_t = T_t * S_t * R_t$ which is preferred when seasonal variation grows in proportion to the series overall magnitude.The selection between these models depends on whether the seasonal fluctuations intensity changes with the series average level.

The seasonal decomposition shown in the image does not seem to have worked as expected. Although the original time series clearly shows a strong yearly pattern, both the Seasonal and Residual components appear as flat lines around the value of 1.00. This outcome suggests that something may have gone wrong during the decomposition process.

One possible reason could be an incorrect seasonal period—perhaps the daily data was not set to a 365 day cycle as required.It could also be that the algorithm chosen for decomposition is not well-suited to handle this type of complex seasonality, even though the dataset spans nine years.

In a properly applied multiplicative decomposition, we would expect the Seasonal component to show a clear, repeating cycle fluctuating around 1.00, and the Residual to reflect random variations, also centered near 1.00. Since both components appear flat in this case, it indicates that the decomposition method was not able to correctly isolate the seasonal structure or the irregular patterns in the data.

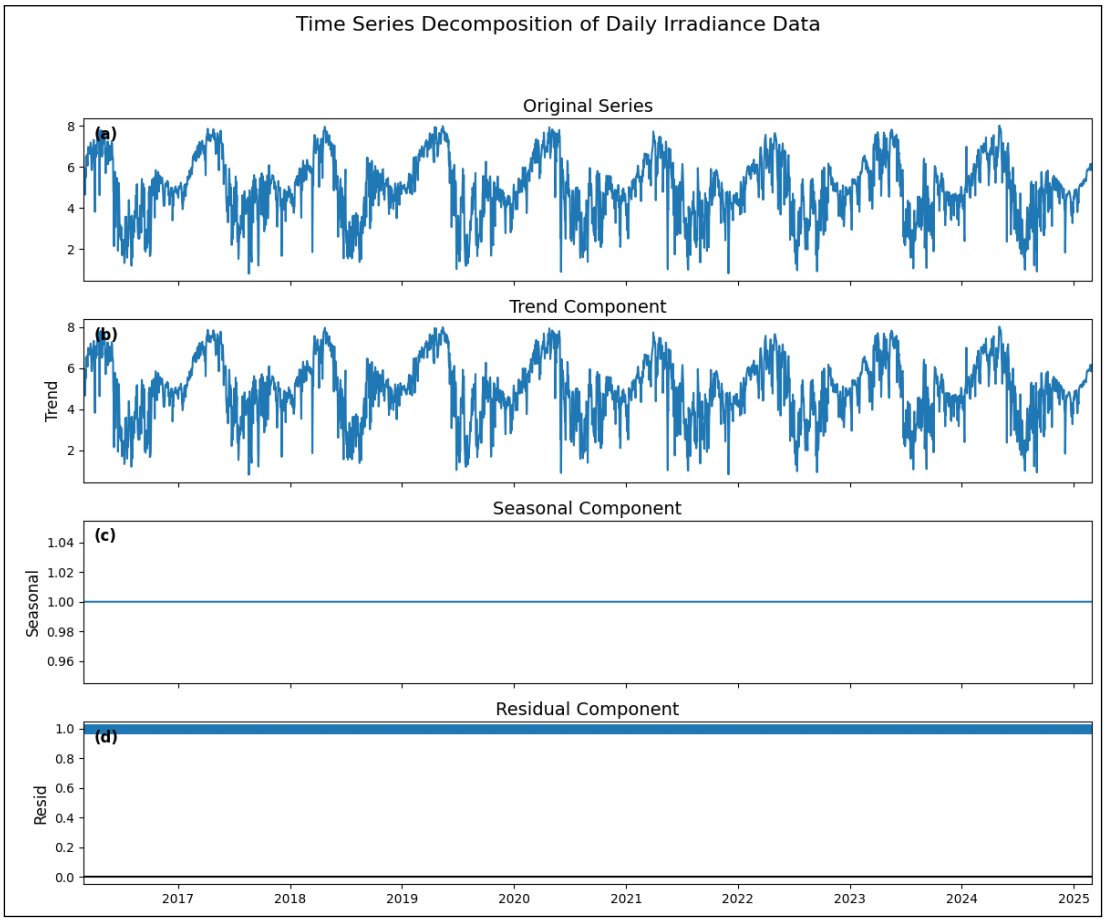

Figure 2.11: (a) Original Series: Displays the raw daily irradiance data over time.(b) Trend Component: Shows the long-term trend of the irradiance data.(c) Seasonal Component: Appears as a flat line at 1.0, indicating that a seasonal period of 1 was used, which effectively extracts no explicit seasonality.(d) Residual Component: Also appears flat, reflecting that no variation is attributed to random noise after accounting for the trend and the non-existent seasonal component with period 1.

Seasonal decompose Additive model

Interestingly, the additive decomposition method used here seems to have captured most of the annual cycle within the Trend component. The trend closely follows the overall shape of the original time series, even though there is no clear long-term increase or decrease. As a result, both the Seasonal and Residual components appear as nearly flat lines centered around zero. This suggests that the model interpreted the repeating yearly pattern as part of the trend rather than separating it out as a distinct seasonal effect. It also means that there is very little random noise or unexplained variation left in the data, as indicated by the minimal activity in the residual plot.

Another Seasonal decompose Additive Model by Hand Calculation

This visualization displays the additive seasonal decomposition of the "ALLSKY_SFC_SW_DWN" time series, covering from late 2016 to early 2025. The top panel, showing the original data, exhibits a clear annual cycle with generally higher values mid-year and lower values at the year start and end, despite some inherent noisiness. The "Trend" component, depicted in the

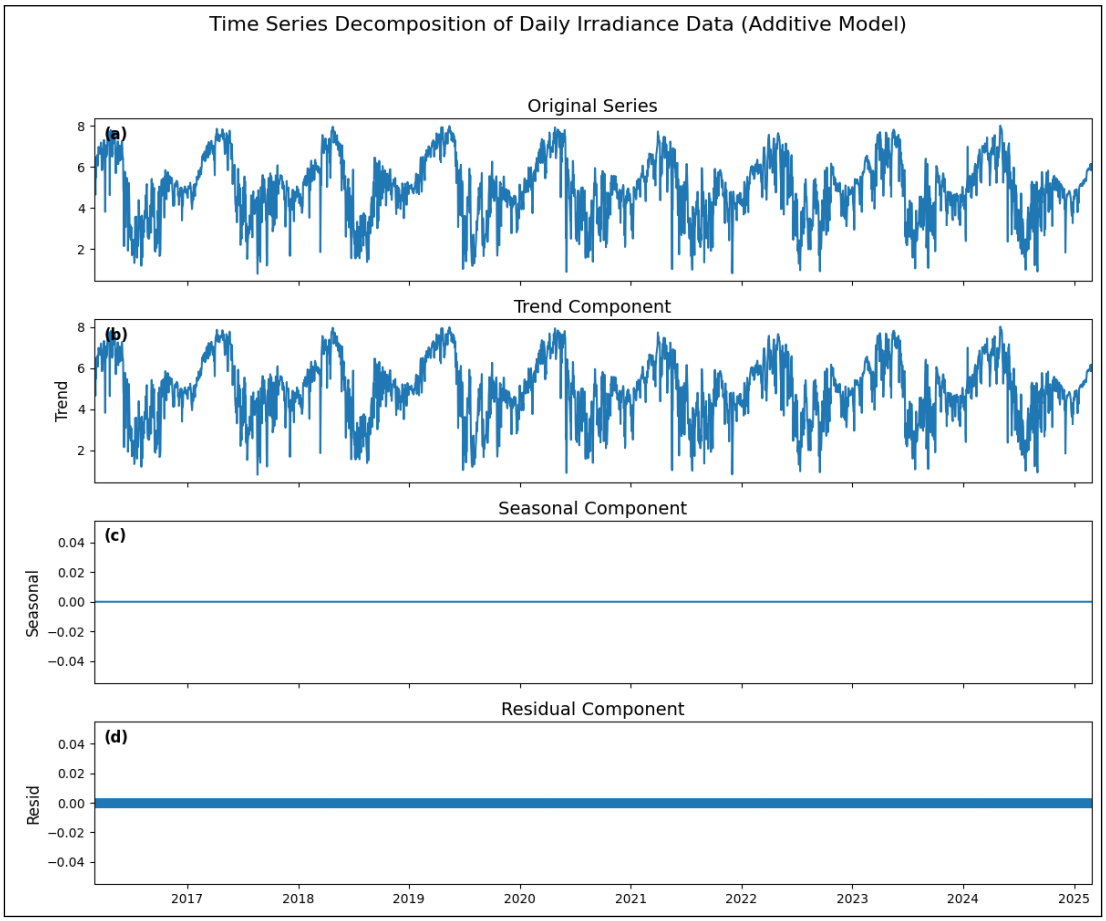

Figure 2.12: This figure illustrates the decomposition of the daily irradiance time series into its individual components using an additive model.(a) Original Series: Displays the raw daily irradiance data.(b) Trend Component: Shows the underlying long-term trend extracted from the data.(c) Seasonal Component: Appears as a flat line at 0.0, indicating that no explicit seasonality was extracted by the decomposition (likely due to a period of 1 being specified in the decomposition, or absence of seasonality).(d) Residual Component: Also appears flat around 0.0, representing the remaining noise after the trend and seasonal components were accounted for.

second panel, effectively captures this prominent yearly fluctuation as an underlying pattern, indicating the data overall direction without a consistent long-term increase or decrease. Below it, the "Seasonality" panel distinctly highlights recurring positive and negative deviations from the trend, representing the predictable annual variations that consistently influence the data.Finally, the bottom "Residual" panel shows the remaining unexplained variation as fluctuations around zero, suggesting that the model has successfully isolated the systematic patterns both the overarching annual trend and the specific seasonal effects leaving behind only random noise.

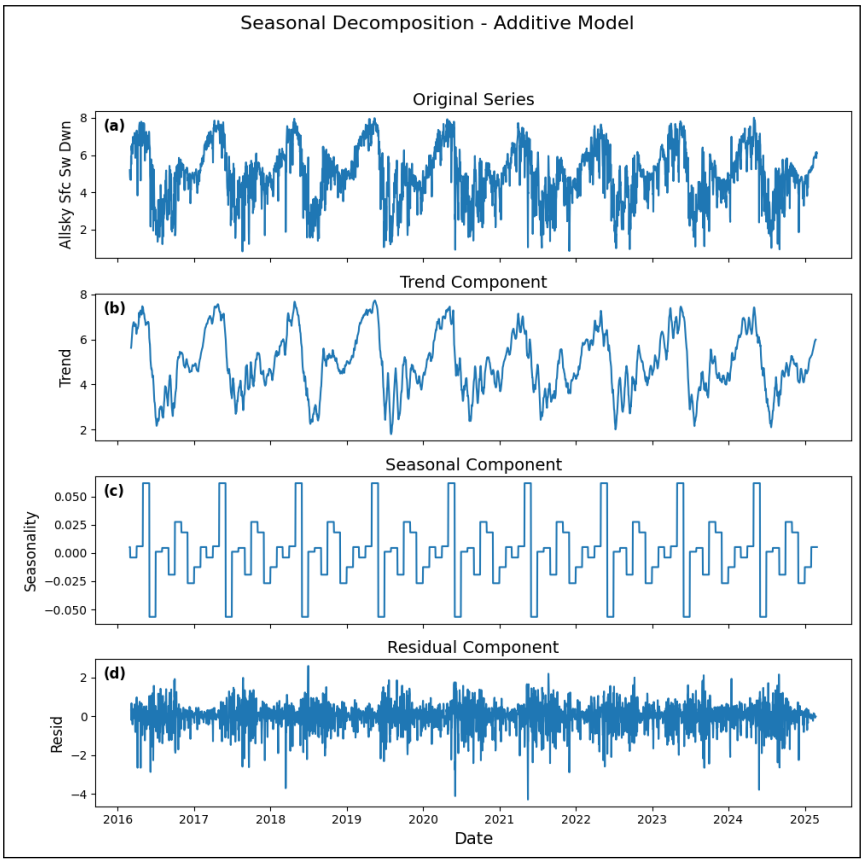

Figure 2.13: Seasonal Decomposition - Additive Model. This figure illustrates the manual decomposition of daily irradiance data into its constituent components using an additive model.(a)**Original Series:**Displays the raw time series data for 'Solar Irradiance'.(b) **Trend Component:**Shows the long-term trend extracted using a rolling mean.(c)**Seasonal Component:** Presents the estimated seasonal pattern, calculated as the mean detrended value for each month, resulting in a step-like appearance.(d)**Residual Component:** Represents the remaining irregular or random fluctuations after subtracting the trend and seasonal components.

2.3 Data preprocessing and feature Engineering

Data preprocessing prepares raw data for machine learning by cleaning, transforming, and organizing it.Feature engineering, often part of preprocessing, takes this a step further by creating new, more informative variables or improving existing ones. This is like turning basic ingredients into a gourmet meal, making the data more "intelligent" for the model.

2.3.1 Transform Wind direction Feature

Wind direction is a circular (cyclic) variable that ranges from 0^0 to 360^0, where 0^0 and 360^0 represent the same direction (North). Treating such features as linear in machine learning models can mislead the learning algorithm, as the numerical difference between 1^0 and 359^0 is large 358, but in reality, these directions are very close.

To properly encode the cyclic nature of wind direction data, the features WD2M (wind direction at 2 meters) and WD10M (at 10 meters) were transformed using sine and cosine functions.This technique maps the angle onto a unit circle, preserving the circular continuity of the data.

The transformation was performed using the following equations:

$$\mathrm{WD}_{\sin} = \sin\left(\frac{2\pi \times \mathrm{Wind\_Direction}}{360}\right) \tag{2.1}$$

$$\mathrm{WD}_{\cos} = \cos\left(\frac{2\pi \times \mathrm{Wind\_Direction}}{360}\right) \tag{2.2}$$

2.3.2 Logarithmic Transformation of Precipitation Feature

The variable PRECTOTCORR, which represents the corrected total daily precipitation (in mm/day), exhibits a **highly right-skewed distribution**. This means that while most days have little or no rainfall, there are a few days with extremely high precipitation, leading to outliers and distributional imbalance.Such skewness can negatively impact the performance of machine learning models, especially those sensitive to feature scale and variance.

$$\mathrm{PRECTOTCORR}_{\log} = \log\left(1 + \mathrm{PRECTOTCORR}\right) \tag{2.3}$$

2.3.3 Correlation Analysis

Correlation is a statistical measure that expresses the degree to which two variables change together. In other words, it quantifies the strength and direction of a linear relationship between two continuous variables.

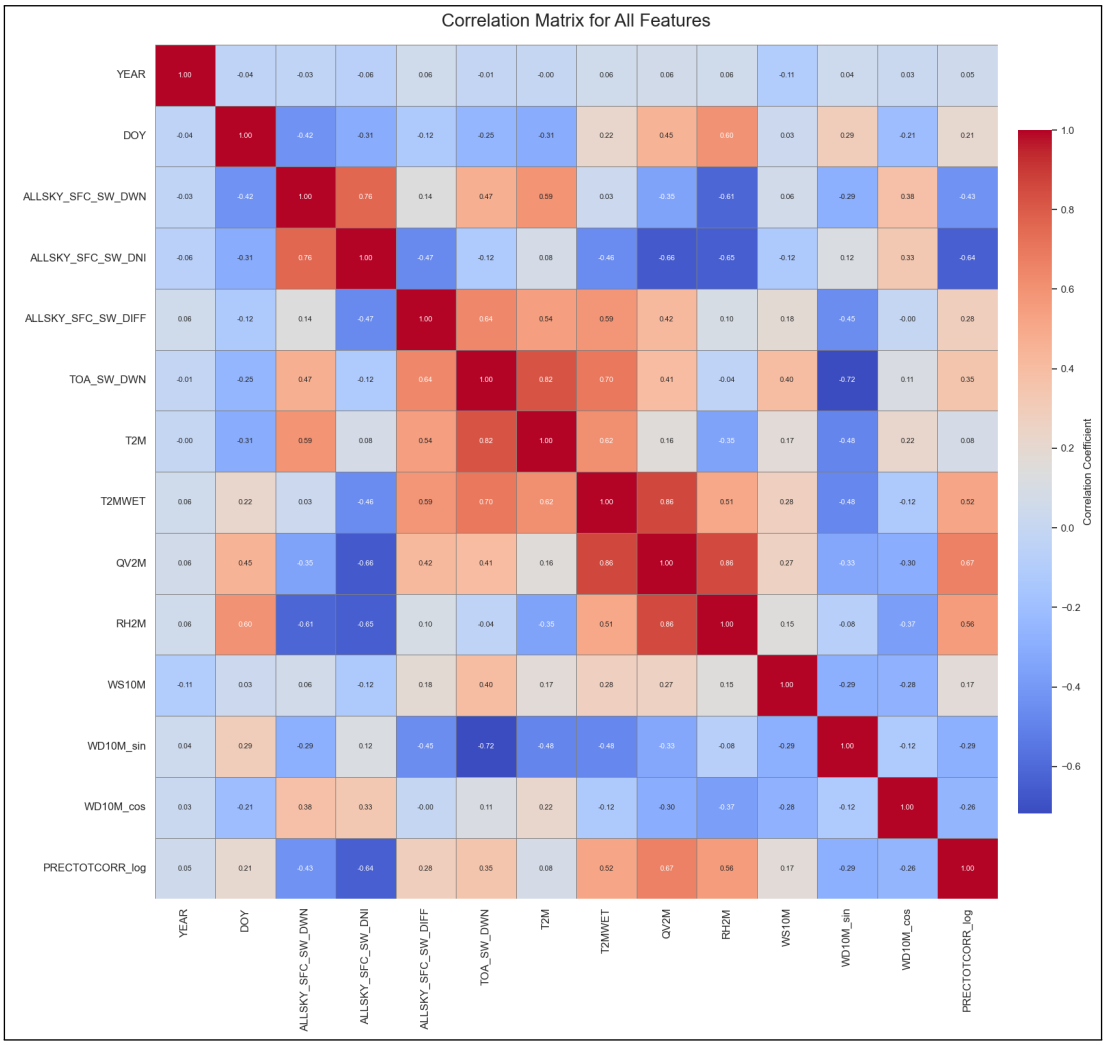

Figure 2.14: Correlation matrix for all features

The most common measure of correlation is the Pearson correlation coefficient,denoted by r, which ranges from 1 to +1:

- +1: Perfect positive correlation — as one variable increases, the other increases proportionally.
- 0: No linear relationship between the two variables.
- 1: Perfect negative correlation — as one variable increases, the other decreases proportionally.

In the preprocessing phase, certain features were excluded from the final dataset based on their low predictive importance, high redundancy, or minimal contribution to the forecasting task. The goal of this step was to improve model performance by reducing noise, lowering computational complexity, and preventing multicollinearity.

2.3.4 Time series feature Engineering

To capture temporal dependencies in solar irradiance patterns, the dataset was enhanced with lag features a commonly used technique in time series forecasting. Lag features allow the model

to learn from historical values of the target variable by introducing previous observations as predictors.

Lag Based Features

- lag_1: GHI value from the previous day.
- lag_2: GHI value from two days prior.
- lag_7: GHI value from the same day one week earlier.
- lag_365: GHI value from the same day one year earlier.

This type of feature engineering effectively converts a univariate time series problem into a supervised learning problem, where each row represents a day, and the features represent past values.

Chapter 3

Models and methods

In this chapter, we describe the machine learning and statistical methods used to forecast Global Horizontal Irradiance (GHI) based on satellite-derived environmental features. Multiple models were developed and evaluated, each chosen for its ability to capture different aspects of temporal and non-linear patterns in solar irradiance data.

The goal was to compare these models in terms of forecast accuracy, computational efficiency, and interpretability, with a focus on predicting GHI values for March 2025.

3.1 Classical Machine Learning Models

3.1.1 Random Forest Model

The Random Forest Regressor is an ensemble learning method used for regression tasks. It builds a collection of decision trees during training and combines their outputs to produce more accurate and stable predictions than individual models. Random Forest is particularly well suited for datasets with complex, non-linear relationships and high dimensionality, such as those involving atmospheric and solar parameters.

A single decision tree is prone to overfitting, especially when trained on noisy data.To overcome this, Random Forest creates a "forest" of decision trees,each trained on slightly different data, and then averages their predictions.This ensemble approach reduces variance, improves generalization,and mitigates the risk of overfitting.

Algorithm of Steps

1. **Bootstrap Sampling**
    - From the original training dataset, the algorithm draws multiple random subsets (with replacement).
    - Each subset is used to train one individual decision tree.
    - This is called bagging (Bootstrap Aggregating).
2. **Building Decision Trees**
    - Each decision tree is trained using a different bootstrap sample.
    - At each split in a tree, only a random subset of features is considered (not all features).
    - This introduces further randomness, ensuring that individual trees are diverse and less correlated.

3. **Tree Growth**

   - Trees in a Random Forest are typically grown deep and unpruned, capturing a wide range of patterns.
   - Each tree outputs a numerical prediction (in regression tasks,like predicting GHI).

4. **Aggregating Results**

   - For regression, the final prediction is the average of the predictions from all individual trees:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i \tag{3.1}$$

   Where $\hat{y}_i$ is the prediction from $ı^{th}$ tree and N is the number of trees.

Algorithm Flowchart

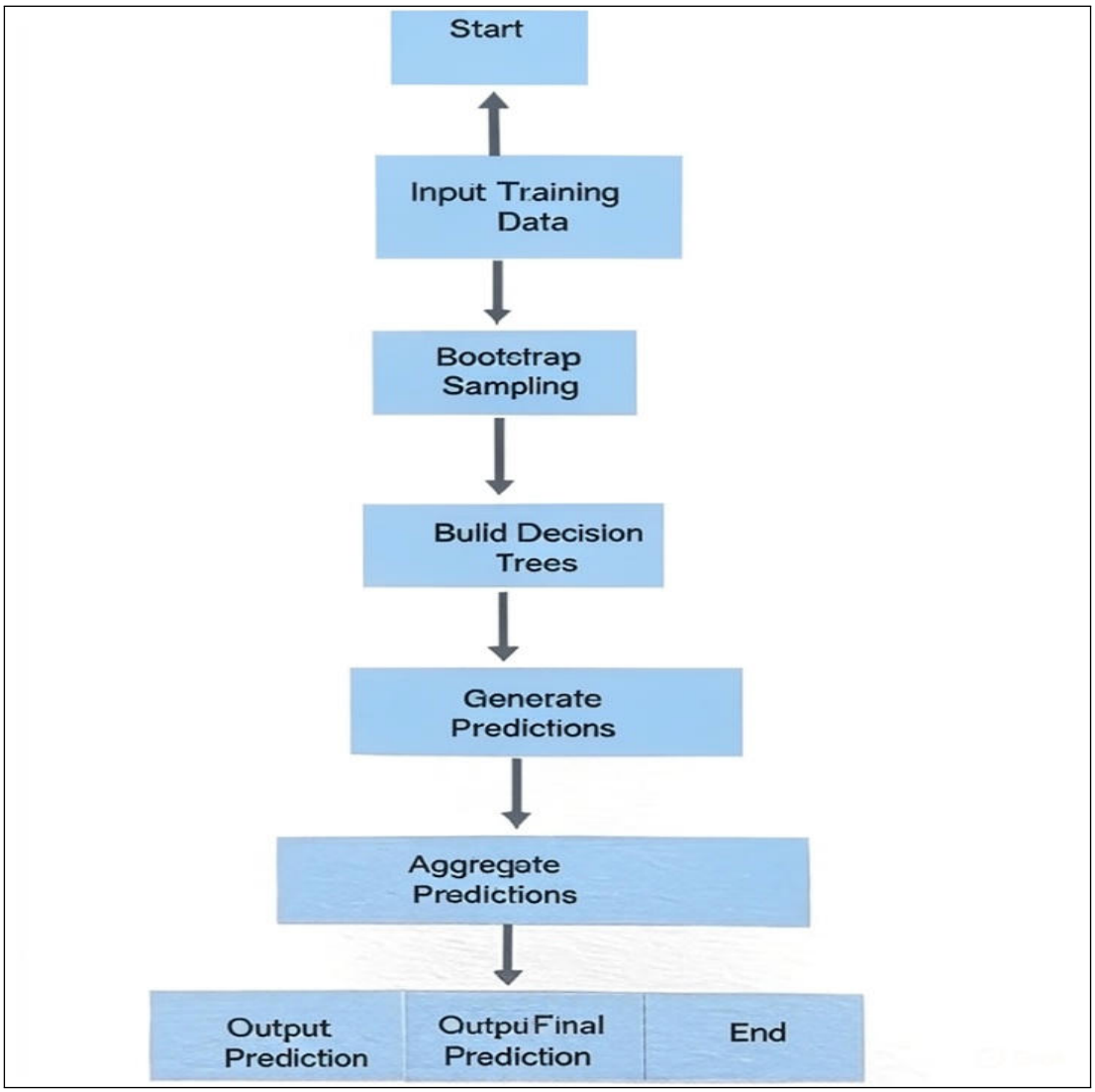

Figure 3.1: Random Forest Architecture. This flowchart illustrates the typical steps involved in building and using a Random Forest model.

3.1.2 Gradient Boosting

Gradient Boosting Regression is a powerful ensemble machine learning algorithm that builds predictive models in a stage wise manner by combining multiple weak learners (typically decision trees) into a strong learner. Unlike Random Forest, which builds trees in parallel, Gradient Boosting constructs trees sequentially, with each tree attempting to correct the errors made by the previous one.This method is particularly effective for regression problems with non-linear relationships, such as forecasting solar irradiance using atmospheric variables.

Gradient Boosting works by training one tree at a time, where each tree tries to correct the errors made by the previous one. The process starts with an initial prediction, and the difference between actual and predicted values (called residuals) is calculated. These residuals are then used to train the next tree. This cycle continues, with each new tree improving the overall prediction. In the end, all tree outputs are combined to give a final, more accurate result.

Algorithm of Steps

1. **Initialize the Model**

   Start with a simple model (usually predicting the mean of the target variable) as the initial prediction $F_0(x)$.

2. **Compute Residuals**

   Calculate the residuals (errors) between the true values and the current predictions:

   $$ri = y_i - F_{m-1}(x_i) \tag{3.2}$$

   where $F_{m-1}(x_i)$ is the prediction from the previous stage.

3. **Fit a Weak Learner**

   Train a new regression tree to predict the residuals(i.e., errors from Step 2).

4. **Update the model**

   Update the prediction by adding a scaled version of the new tree's prediction to the previous prediction.

   $$F_m(x) = F_{m-1}(x) + \eta.h_m(x) \tag{3.3}$$

   where:

   - $h_m(x)$ is the prediction from the new tree.
   - η is the learning rate ($0 < \eta < 1$) that control the contribution of each tree.

5. **Repeat**

   Repeat steps 2–4 for a predefined number of iterations (trees) or until the loss converges.

Algorithm Flowchart

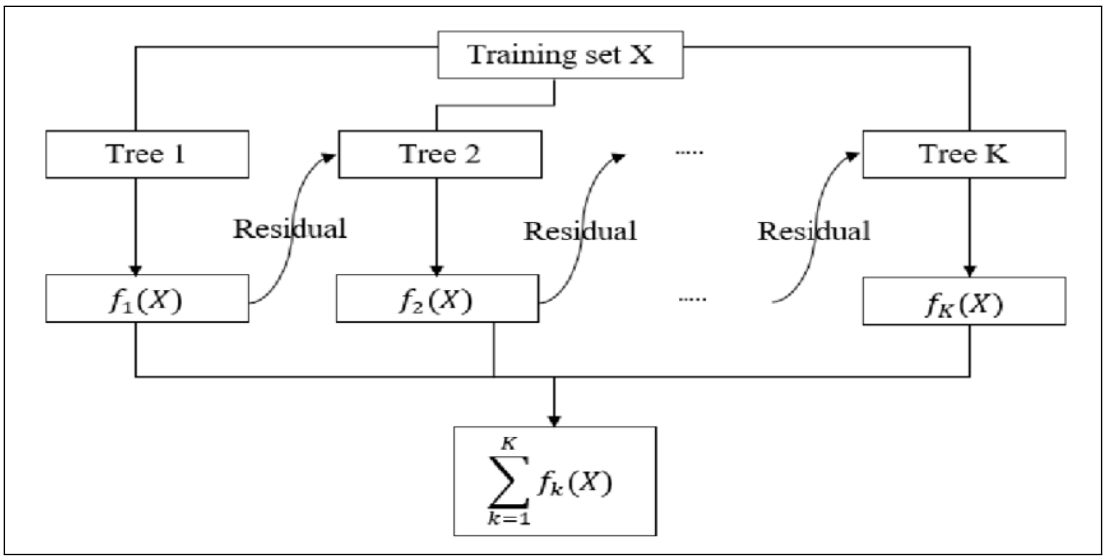

Figure 3.2: Flowchart of Gradient Boosting Regression Algorithm. This diagram illustrates the iterative process of Gradient Boosting, where an ensemble of weak learners (trees) are sequentially built. Each subsequent tree focuses on predicting the residuals (errors) from the previous trees, and their outputs are combined to form the final prediction.

3.2 Time Series Forecasting Models

The problem statement highlights the need to forecast[2] solar irradiance using satellite imagery and atmospheric parameters, addressing the limitations of traditional meteorological models. XGBoost and LightGBM aligns well with this problem because:

- It supports multi-feature, tabular data that includes cloud cover, temperature, humidity, wind, and radiation parameters.
- It can handle time-dependent patterns using engineered features like lags and seasonal encoding, even though it is not inherently a time series model.
- It performs robustly with non-linear, high-dimensional data such as the NASA POWER dataset with multiple correlated features.
- It can capture interactions between variables like temperature and humidity, which affect irradiance but are hard to model with traditional linear techniques.

3.2.1 XGBoost Model with Lag Features

XGBoost (Extreme Gradient Boosting) is a high-performance, scalable implementation of the gradient boosting algorithm. It has gained popularity for structured (tabular) data problems due to its ability to model complex, non-linear relationships and efficiently handle large datasets.

In the context of this project, where the objective is to predict daily Global Horizontal Irradiance (GHI) for March 2025 using satellite-derived meteorological data, XGBoost was an ideal choice due to the following reasons:

How Algorithm Works

1. **Initialized the model**

   Start with a simple model (usually predicting the mean of the target variable) as the initial prediction$F_0(x)$.

2. **Compute Residuals (Errors)**

   Calculate the difference between the actual value and the predicted value. These residuals represent the portion of the data that the model has not yet explained.

$$r_i = y_i - \hat{y}_i \tag{3.4}$$

3. **Build a Decision Tree to Fit Residuals**

   - A shallow regression tree is trained to predict the residuals from Step 2.
   - Each leaf node in the tree represents a prediction that helps correct the original model's output.
   - **Gradient Descent Optimization**
     - Instead of fitting residuals directly, XGBoost minimizes a loss function (e.g., Mean Squared Error for regression) using gradient descent.
     - The first-order derivative (gradient) of the loss function.The second-order derivative (Hessian) to capture curvature.
   - **Update the model**

     The predictions of the new tree are scaled by a learning rate η and added to the existing model.

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{3.5}$$

     Here, $F_m(x)$ is the updated prediction, and $h_m(x)$is the output from the m-th tree.

Algorithm Flowchart

The diagram illustrates the working of XGBoost, a powerful and optimized version of gradient boosting. It begins with the training dataset, and at each iteration, a new decision tree is built left by the previous trees. Each tree is uniquely defined by its parameters$(\theta_1, \theta_2 ..... \theta_k)$ and the model continuously improves by learning from these residuals. The tree splits are guided by a regularized objective function, which balances prediction accuracy and model complexity. This helps in preventing overfitting and ensures better generalization. After several rounds of learning, the final prediction is obtained by summing the outputs of all individual trees.

While XGBoost is a boosting algorithm that builds an ensemble of decision trees, it is fundamentally inspired by the principles of gradient descent. In gradient descent, the model updates its parameters step by step in the direction that reduces the loss. Similarly, XGBoost adds new trees in a sequential manner, where each new tree acts like a "gradient step" learning to correct the mistakes of the previous ones. However, instead of adjusting numeric weights, XGBoost builds trees as functional approximations to minimize the loss. This makes XGBoost a tree-based implementation of gradient boosted learning, where the optimization is performed over a space of functions rather than individual parameters.

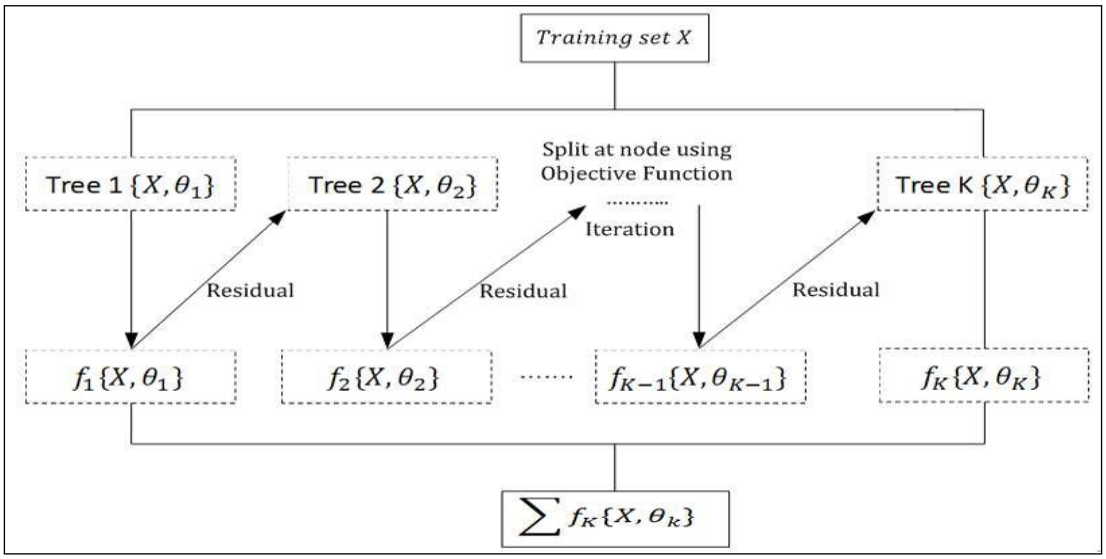

Figure 3.3: Workflow of the XGBoost algorithm. Each decision tree is trained sequentially, learning from the residuals of the previous trees. The model optimizes a regularized objective function at each node split, and the final prediction is the sum of outputs from all trees.

3.2.2 Light-GBM (Light Gradient Boosting Machine) model

LightGBM is a highly efficient and fast implementation of the gradient boosting framework, developed by Microsoft. It is designed to provide high accuracy with faster training time and lower memory usage compared to traditional boosting algorithms such as XGBoost. LightGBM is particularly well-suited for large-scale datasets and supports advanced features such as histogram-based learning and leaf-wise tree growth.

In this project, LightGBM was used to forecast Global Horizontal Irradiance (GHI) using multivariate satellite-derived meteorological data and time-series features. Its ability to handle high-dimensional, non-linear, and sparse data makes it a strong choice for solar irradiance prediction.

LightGBM Algorithm Works

1. **Initialize the Model**

   The model starts by making an initial prediction, often the mean of the target variable in the case of regression.This prediction is the baseline from which errors will be computed.

2. **Calculate Gradients and Hessians**

   The gradient (first derivative of the loss function), representing the direction of the error.The Hessian (second derivative), representing the curvature of the loss function.These derivatives are used to optimize the objective function during tree construction.

3. **Histogram Binning**

   - Instead of using raw continuous feature values, LightGBM groups values into discrete bins (e.g., 256 bins).
   - This histogram-based technique reduces computation by simplifying the process of finding the best split in each node.
   - It also reduces memory usage and allows faster training on large datasets.

4. **Tree Building (Leaf-wise Growth)**
   - Unlike level-wise tree growth (used in Random Forest and XGBoost), LightGBM grows trees leaf-wise.
   - In each iteration, the algorithm finds the leaf node with the maximum loss reduction and splits it.
   - This creates asymmetrical trees that can grow deep where necessary, leading to better accuracy.
5. **Repeat for Multiple Iterations**

   Predictions from all trees are aggregated (summed) to produce the final output.

$$\hat{y} = F_0(x) + \sum_{i=1}^{M} \eta \cdot f_i(x) \tag{3.6}$$

Where:
$F_0(x)$ is the initial prediction,
$f_i(x)$ is the prediction from the i-th tree,
η is the learning rate.

3.2.3 ARIMA (Autoregressive Integrated Moving Average)

The **ARIMA** model is a widely used statistical time series forecasting technique[2] that is particularly effective for modeling data that shows temporal dependence and trend. The acronym ARIMA stands for:

- **AR: Autoregression** A model that uses the dependent relationship between an observation and a number of lagged observations.
- **I: Integrated** Involves differencing the observations to make the time series stationary (i.e., constant mean and variance over time).
- **MA: Moving Average** Models the relationship between an observation and a residual error from a moving average model applied to lagged forecast errors.

In this thesis, ARIMA was applied to forecast **Global Horizontal Irradiance (GHI)** using daily historical solar radiation data from the NASA POWER dataset, covering February 2016 to February 2025.

ARIMA Model work **ARIMA** works in three main stages: **Identification**, **Estimation**, and **Forecasting**.

1.Identification

First, the time series is examined for stationarity using plots and statistical tests like the **Augmented Dickey-Fuller (ADF) test**. If the data is not stationary, it is differenced (subtracting current values from previous ones) until stationarity is achieved. The number of differencing operations is the 'd' parameter in ARIMA.

2. Model Parameterization

ARIMA uses three main parameters:

- **p (AR order)**: Number of lag observations included in the model.
- **d (Integration order)**: Number of times the data needs to be differenced to achieve stationarity.
- **q (MA order)**: Number of lagged forecast errors in the prediction equation.

The parameters 'p' and 'q' are typically chosen using.

- **ACF (Autocorrelation Function)** plots
- **PACF (Partial Autocorrelation Function)** plots

3. Model Fitting and Forecasting

Once 'p', 'd', and 'q' are defined, the model is trained using historical data. The model fits a linear equation that incorporates both past observations and past errors. Forecasts are generated by projecting the model into future time steps, such as predicting GHI for March 2025.

4.Mathematical Representation

An **ARIMA(p, d, q)** model can be expressed as:

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)\epsilon_t \tag{3.7}$$

Where:

- y_t: Actual value at time t
- $\phi_p(B)$: Autoregressive (AR) terms (based on lagged values)
- $\theta_q(B)$: Moving Average (MA) terms (based on lagged errors)
- B: Backshift operator (e.g., $By_t = y_{t-1}$)
- ϵ_t: White noise error term

5.ARIMA Model Flowchart

The flowchart illustrates the systematic process of developing an ARIMA (AutoRegressive Integrated Moving Average) model for time series forecasting. The first step involves plotting the time series and analyzing its Autocorrelation Function (ACF) and Partial ACF (PACF) to understand the underlying patterns and dependencies. A critical check is performed to determine if the series is stationary, meaning it has a constant mean and variance over time. If the series is not stationary, regular and seasonal differencing is applied to stabilize it.

Once the data is made stationary, the next step is model selection, where appropriate values for the ARIMA parameters (p, d, q) are chosen based on the behavior observed in the ACF/PACF plots and domain understanding. The selected model is then fit to the data by estimating the parameters. After this, a diagnostic check is conducted to see if the residuals are uncorrelated that is, they should behave like white noise. If not, the model may be underfitting and requires adjustments, such as adding more parameters or modifying its structure. If the residuals are uncorrelated, a further check is performed to ensure that all estimated parameters

are statistically significant. Any parameter that lacks significance is removed from the model, and the estimation process is repeated.

This iterative loop continues until a model with uncorrelated residuals and significant parameters is achieved. At this point, the finalized ARIMA model can be confidently used to forecast future values. This structured, feedback driven approach ensures that the final model is both robust and statistically sound for time series prediction.

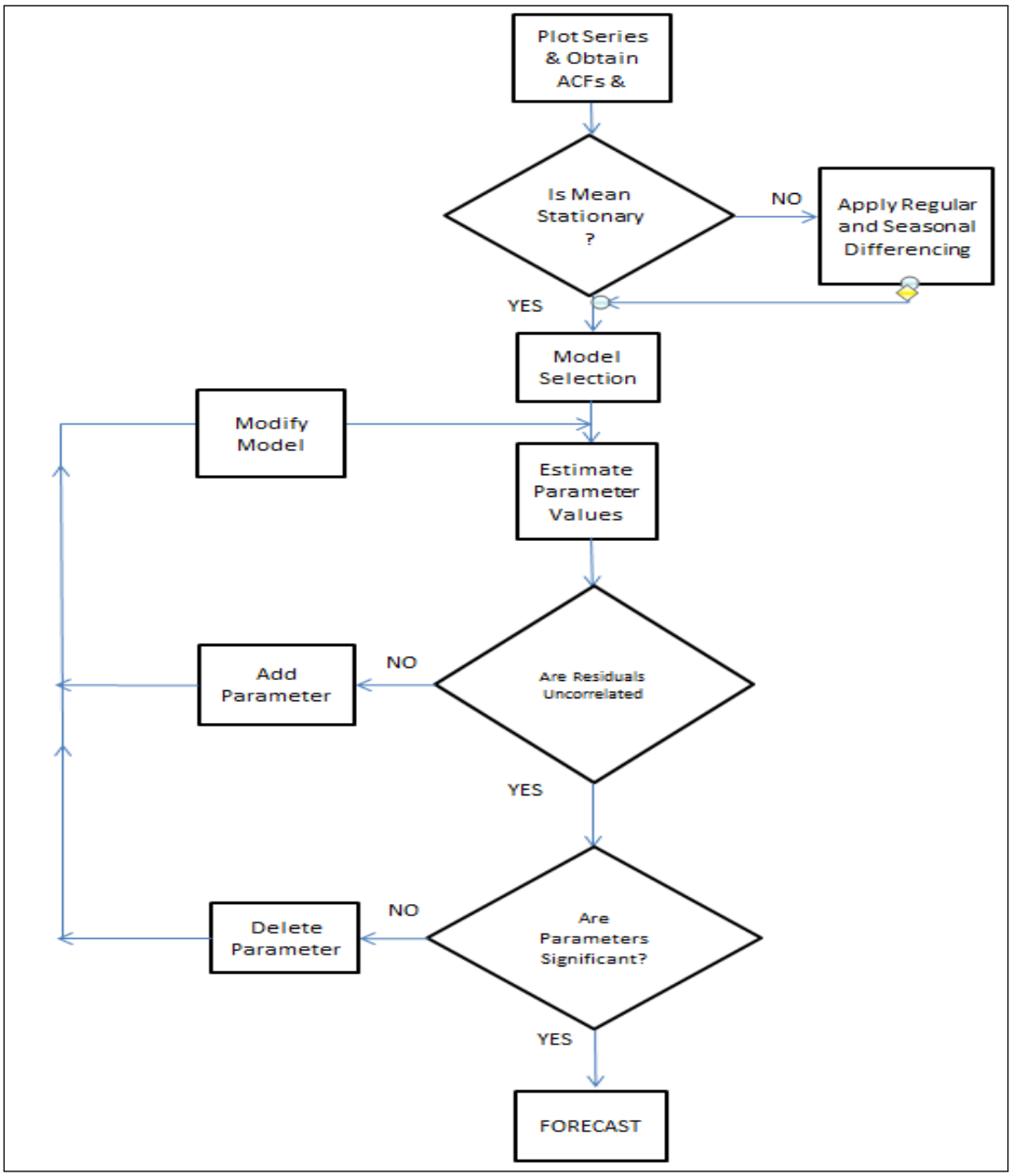

Figure 3.4: Flowchart illustrating the ARIMA modeling process. The steps include checking for stationarity, applying differencing if needed, selecting and estimating model parameters, and validating residuals. The process is repeated until a statistically sound model is achieved, which is then used for forecasting.

3.2.4 Facebook Prophet Model

Facebook Prophet[6] is an open source time series forecasting tool developed by Meta that is designed to handle time series data with strong seasonal effects, missing values, and outliers. It is particularly well suited for forecasting applications in business, meteorology, and energy where data exhibits yearly, weekly, and daily seasonality as is the case with solar irradiance.Its ability to model trend and seasonality independently made it a strong choice for this task.

The prediction of Global Horizontal Irradiance (GHI) is strongly seasonal and affected by cyclical weather patterns, especially across months like March, which is a transition period in solar irradiance behavior.

- Automatically models yearly seasonality, which is dominant in solar data.
- Handles missing values and outliers, which are common in satellite datasets.
- Allows users to view and interpret individual components (trend, seasonality).
- Requires minimal data preprocessing, such as no need for differencing or stationarity adjustments.

Prophet Model Works

Facebook Prophet is an open-source time series forecasting algorithm developed by Meta (formerly Facebook) for producing accurate forecasts of time series data that exhibit strong seasonal patterns and missing values. It is specifically designed for business and environmental applications where time series have **trend, seasonality, and holidays or abrupt changes.**

Prophet is built on the premise that many real world time series can be decomposed into an additive model of the following form:

$$y(t) = g(t) + s(t) + h(t) + \zeta_t \tag{3.8}$$

Where:

- $y(t)$: Observed value of t(e.g daily Solar irradiance).
- $g(t)$: Trend component (long-term changes).
- $s(t)$: Seasonal component(daily, weekly, yearly).
- $h(t)$: Holiday effect (optional).
- ζ_t: Random error term (noise).

Prophet Model Flowchart

The diagram illustrates the internal workflow of the Facebook Prophet forecasting model, which is designed to handle time series data by decomposing it into separate, interpretable components. The process begins with the input data, which includes time-stamped observations of the variable to be predicted. Prophet then breaks this data down into three main components: trend, seasonality, and holiday effects. The trend component captures the long-term progression of the data, whether it is increasing, decreasing, or exhibiting structural breaks. This is especially useful when the data shows sudden shifts or non-linear growth.

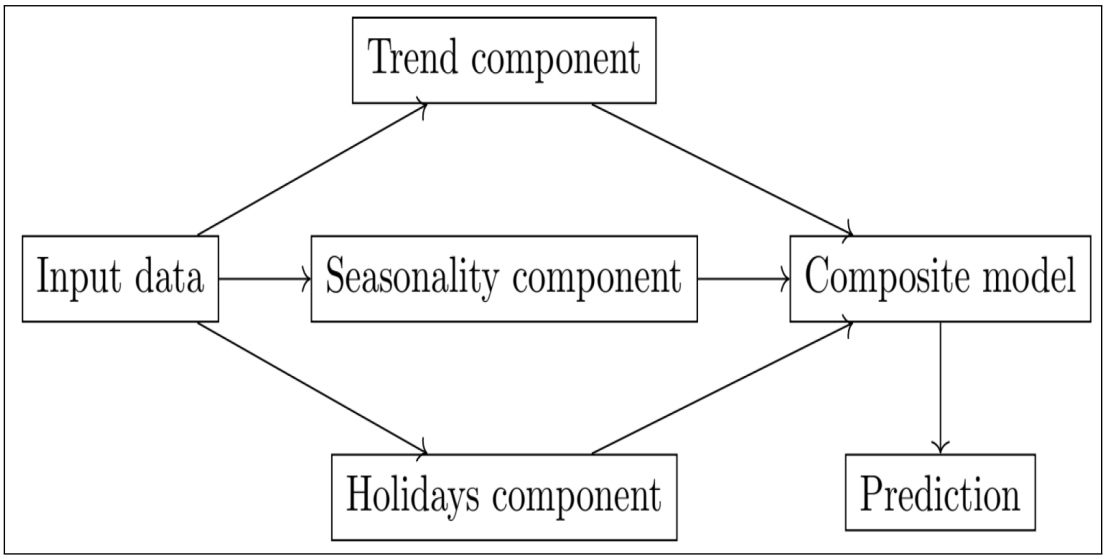

Figure 3.5: Workflow of the Prophet forecasting model. The input time series is decomposed into trend, seasonality, and holiday components. These are combined in a composite model to generate accurate and interpretable predictions.

The seasonality component models recurring patterns at regular intervals, such as daily, weekly, or yearly cycles. These patterns are learned using Fourier series and are essential for capturing the natural fluctuations present in many real-world datasets. Additionally, Prophet allows users to incorporate a holidays component, which accounts for the influence of special events or holidays that can temporarily disrupt regular trends and seasonal patterns. Once all three components are modeled, they are combined into a composite model that synthesizes the overall behavior of the time series. This composite model is then used to generate predictions for future time points. By separately modeling each source of variation, Prophet provides both flexibility and transparency in forecasting, making it a powerful tool for interpretable time series analysis.

Chapter 4

Results

This chapter presents the evaluation and comparison of the models developed to forecast Global Horizontal Irradiance (GHI) using satellite-derived data from the NASA POWER dataset. The models were primarily tested on their ability to forecast solar irradiance for March 2025, a month characterized by seasonal transitions in solar intensity. Performance is assessed using standard regression metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). This section explains the results of classical machine learning models, followed by the forecasting results.

4.1 Method 1: Classical Machine Learning Model

In this method, a classical machine learning model was developed to predict solar irradiance using meteorological parameters. The model utilizes historical data features such as temperature, humidity, wind speed, cloud cover, and solar flux components obtained from the NASA POWER dataset.

4.1.1 Result1 : PCA and Random Forest Model.

Train Result:

- Average MSE: 0.24
- Average R^2 Score: 0.99

Test Result:

- Average MSE: 1.66
- Average R^2 Score: 0.91

The PCA + Random Forest model generalizes well and captures important patterns in the data. Slight overfitting is observed very high train R^2 Score vs. slightly lower test R^2 Score, but overall, the model performs effectively for GHI prediction.

4.1.2 Result2 : Using K-fold cross validation technique to find results.

Random Forest Model Evaluation:

- Average MSE: 0.39
- Average R^2 Score: 0.98

- The Random Forest model showed strong performance with a low average MSE and a high R ^{2} Score, indicating it accurately captured the variance in the data. This suggests the model is reliable and effective for solar irradiance prediction.

Gradient Boosting Model Evaluation:

- Average MSE: 0.43
- Average R^2 Score: 0.98
- The Gradient Boosting model achieved an average MSE and an R^2 Score, indicating high predictive accuracy. It performs comparably to Random Forest, effectively modeling the complex patterns in the data.

Linear Regression Model Evaluation:

- Average MSE: 0.67
- Average R^2 Score: 0.96
- The Linear Regression model, with an MSE and R^2 Score performs slightly worse than both Random Forest and Gradient Boosting models. While still accurate, it is less effective at capturing nonlinear patterns in the data compared to the ensemble methods.

4.2 Method 2 : Forecast Model(using lag and rolling mean feature).

XgBoost Model Result

- RMSE :- 0.192
- R^2 Score :- 0.891

XG Boost Model Jan to Feb 2025

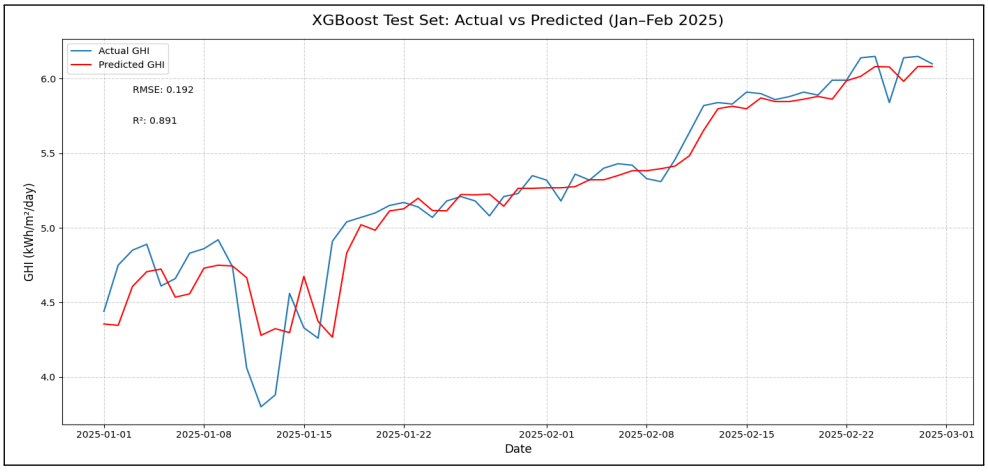

Figure 4.1: XGBoost Test Set: Actual vs Predicted GHI (Jan-Feb 2025). This plot compares the actual daily Global Horizontal Irradiance (GHI) values with those predicted by the XGBoost model over the specified test period.

XG Boost forcast Model March 2025

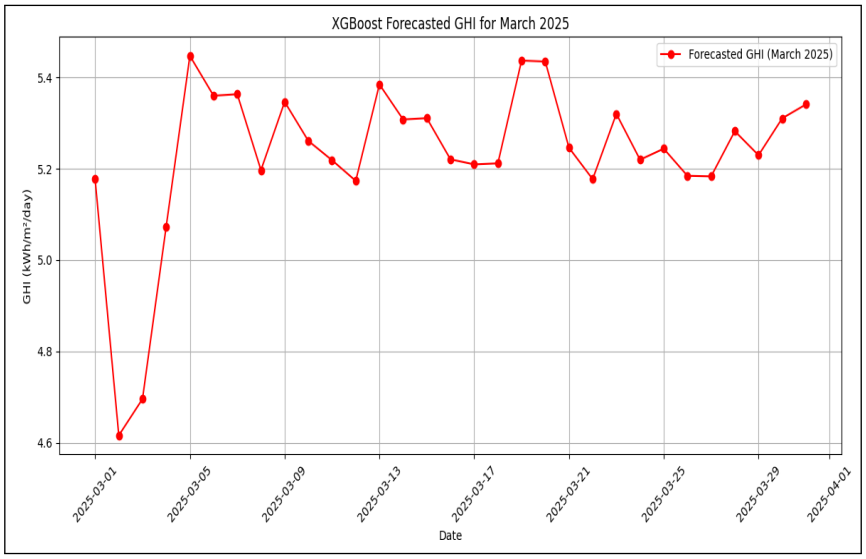

Figure 4.2: Xgboost Forecasted for March 2025

LighGBM Model

- RMSE:- 0.05
- R^2 Score :- 0.856

Lightgbm actual vs forcast results

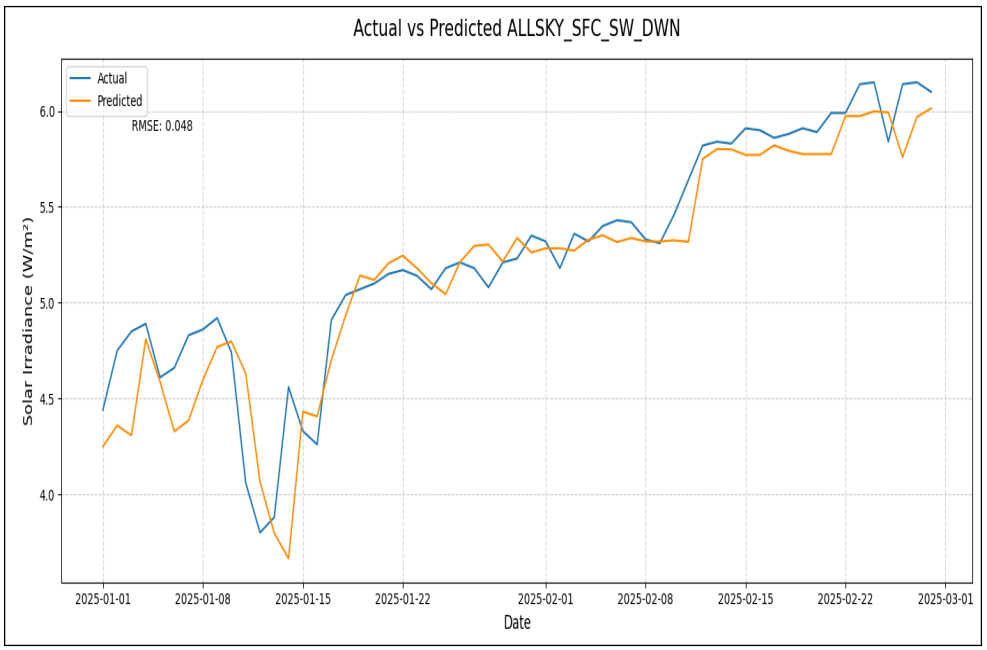

Figure 4.3: This plot compares the actual daily Solar irradiance values with the corresponding predicted values from the model for the period of January-February 2025.

Lightgbm actual vs forcast for March 2025

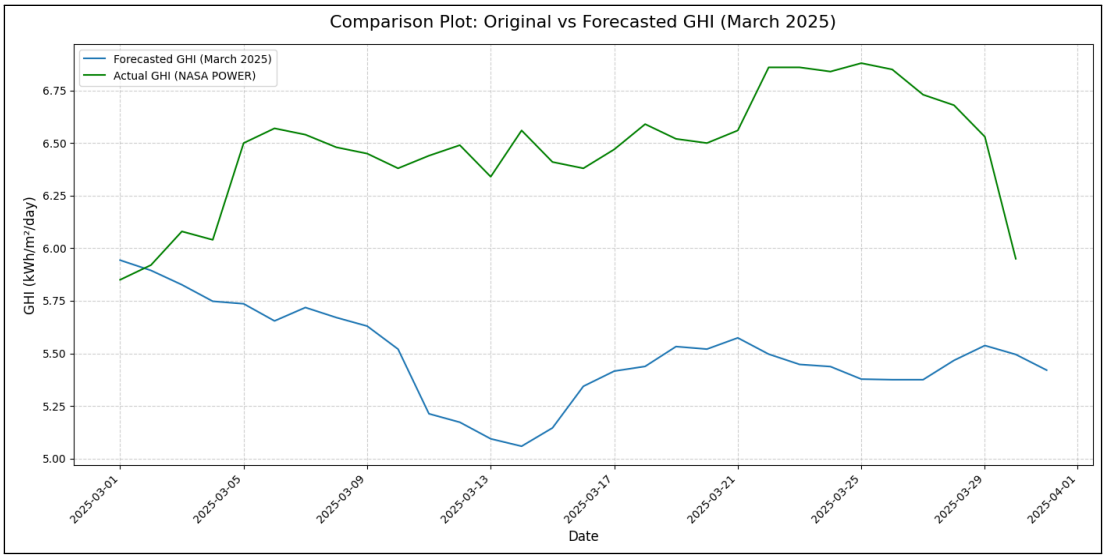

Figure 4.4: Comparison Plot: Original vs Forecasted GHI (March 2025). This figure compares the daily GHI values forecasted by the model with the actual GHI data obtained from NASA POWER for the month of March 2025.

ARIMA model Result

- RMSE :- 0.035
- MAE :- 0.035
- The above ARIMA model not perform well result.

ARIMA model forcast result for March 2025

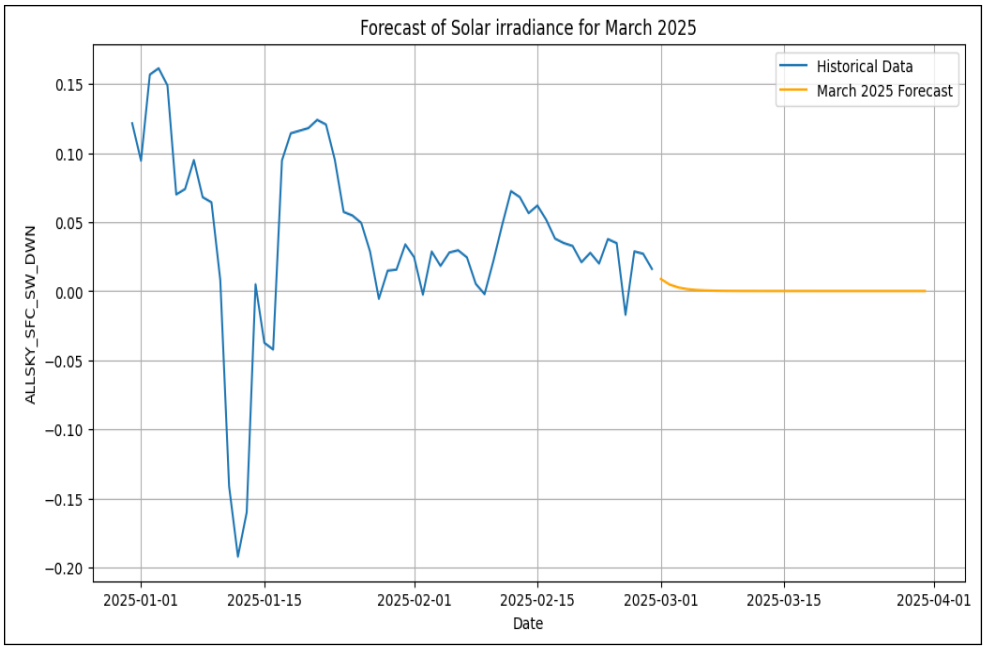

Figure 4.5: ARIMA model forcast result for March 2025

4.3 Method 3: Facebook Prophet Model.

Facebook Prophet is a time series forecasting algorithm developed by Meta. It decomposes time series into trend, seasonality, and residuals using an additive model.

$$y(t) = g(t) + s(t) + h(t) + \zeta_t \tag{4.1}$$

In this approach involves training a model using historical daily GHI data, specifically the ALLSKY_SFC_SW_DWN series, spanning from February 2016 to February 2025. The objective is to predict GHI values for the upcoming month, from March 1 to March 31, 2025. The model's output will include not only the predicted GHI values but also their corresponding 95% confidence intervals, providing a measure of the forecast's uncertainty.

Model Evaluation

- MAPE(Mean absolute percentage error) :- 0.03
- Mean absolute error :- 0.17

Prophet Model actual vs forcast result for Jan-Feb2025

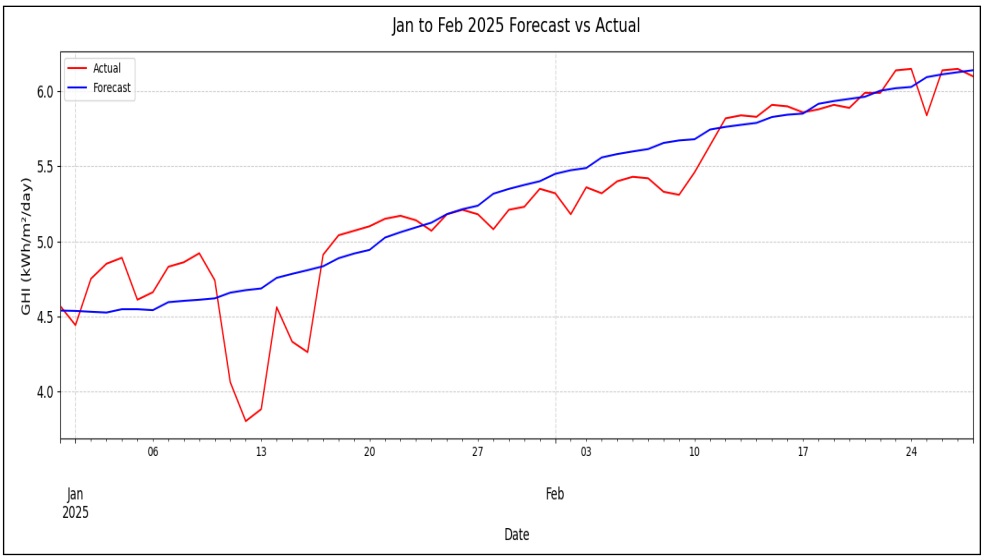

Figure 4.6: Jan to Feb 2025 Forecast vs Actual GHI. This plot compares the actual daily Global Horizontal Irradiance(GHI) with the forecasted GHI for the period of January to February 2025.

Prophet Model actual vs forcast result for March 2025

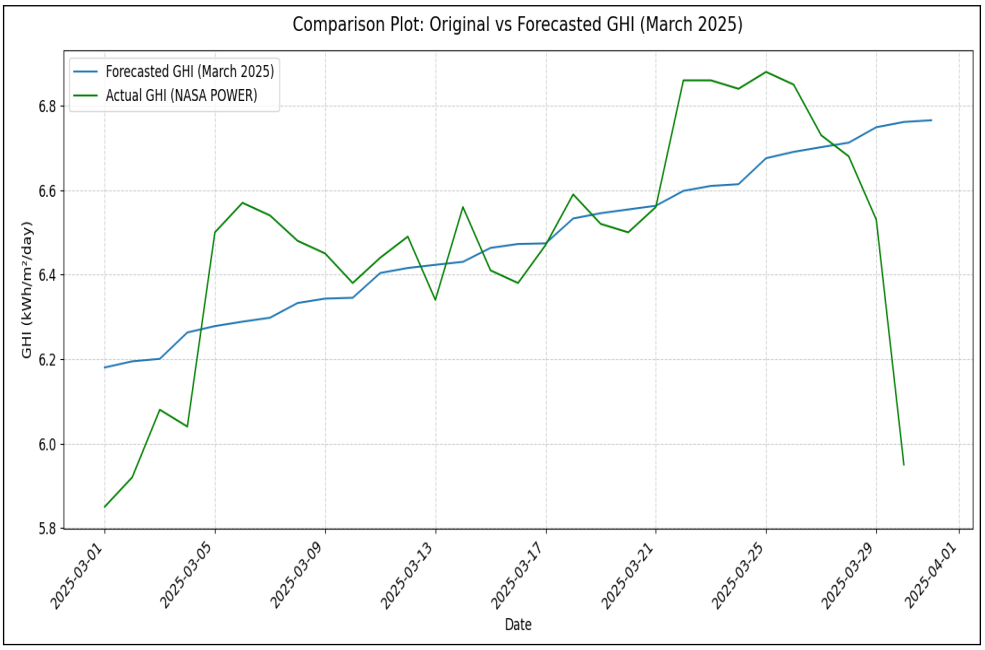

Figure 4.7: Comparison Plot: Original vs Forecasted GHI (March 2025). This figure compares the daily GHI values forecasted by the model with the actual GHI data obtained from NASA POWER for the month of March 2025.

Chapter 5

Conclusion

5.1 Summary of work

This thesis successfully demonstrates that classical machine learning and time series forecasting models particularly Random Forest and Gradient Boosting XGBoost are highly effective for predicting solar irradiance using satellite meteorological data. By utilizing ten years of data from the NASA POWER database, the models were trained to forecast Global Horizontal Irradiance GHI for March 2025, a crucial period for solar energy planning and optimization.

Through the integration of lag features at various time intervals and time based encodings, the machine learning models were able to accurately capture both seasonal trends and short term fluctuations in solar irradiance. Among the tested models, XGBoost outperformed others, including LightGBM, achieving the lowest RMSE and highest R^2 score, demonstrating its ability to model complex, nonlinear patterns in time series data. It was observed that the XGBoost model delivered more stable and informative results when trained on three months of data, as compared to a single month.

In traditional statistical models such as ARIMA and Prophet offered valuable insights into overall trends and seasonality. However, they lacked the flexibility to handle multivariate inputs and were less effective at capturing daily variations in irradiance. While useful for baseline comparison and long term analysis.

Overall, this research highlights the potential of combining satellite data with machine learning techniques to produce reliable solar irradiance forecasts. The modeling framework developed in this study provides a solid foundation for future advancements, including real time solar forecasting, multi source satellite integration, and the implementation of deep learning-based hybrid models.

5.2 Possible future Scope

- A practical future extension involves using forecasted irradiance values to guide solar panel optimization. By linking predicted solar radiation with panel orientation and tilt optimization models, it would be possible to recommend optimal installation parameters for maximum energy yield. This would significantly benefit both small-scale rooftop systems and large-scale solar farms.
- The models developed particularly XGBoost, ARIAM and prophet model performed effectively in predicting Global Horizontal Irradiance (GHI) for March 2025. These models

leveraged engineered features such as lag variables and temporal encodings to capture both seasonal patterns and short term variations.

- XGBoost, in particular, achieved the best performance and showed enhanced accuracy when trained on three months of data instead of just one, highlighting the benefits of using a broader historical context.
- The current study was limited to a single month and location. To improve the generalizability and robustness of the results, future research could involve testing the model across multiple months, seasons, and geographical regions.
- This would help assess its adaptability to different weather patterns, including extremes such as monsoons or winter cloud cover.
- Another area for enhancement lies in the integration of additional satellite data sources. While the NASA POWER dataset provided reliable atmospheric parameters, combining it with higher resolution datasets like NSRDB, GOES, or Solcast could improve both spatial and temporal prediction accuracy.
- The current framework can also be extended by incorporating deep learning models. Techniques such as Long Short-Term Memory (LSTM), GRU, or Time GAN architectures can better capture complex temporal sequences and might outperform classical models in long-range forecasting tasks. These models are especially promising when working with larger datasets or when aiming to include image-based inputs, such as cloud cover maps.
- Additionally, deploying the prediction model as a web-based tool or API would allow real-time access to irradiance forecasts. Such systems could be valuable for solar energy operators, planners, and researchers needing actionable insights for energy management and grid balancing.

In conclusion, while the current analysis addressed its core objectives, there are numerous opportunities for improvement and extension. These enhancements can significantly increase the predictive power, real world usability, and scientific contribution of solar irradiance forecasting systems.

Bibliography

[1] M. K. Boutahir, Y. Farhaoui, M. Azrour, I. Zeroual, and A. Allaoui. Effect of feature selection on the prediction of direct normal irradiance. *Big Data Mining and Analytics*, 5:309 – 317, 12 2022.

[2] C. Chatfield. *The Analysis of Time Series: An Introduction, Sixth Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2003.

[3] A. A. Imam, A. Abusorrah, M. M. A. Seedahmed, and M. Marzband. Accurate forecasting of global horizontal irradiance in saudi arabia: A comparative study of machine learning predictive models and feature selection techniques. *Mathematics*, 12(16), 2024.

[4] V. E. Larson. Chapter 12 - forecasting solar irradiance with numerical weather prediction models. In J. Kleissl, editor, *Solar Energy Forecasting and Resource Assessment*, pages 299–318. Academic Press, Boston, 2013.

[5] N. P. of Worldwide Energy Resources (POWER) Project. Nasa power-access-viewer.

[6] Z. Z. Oo and S. Phyu. Time series prediction based on facebook prophet: A case study, temperature forecasting in myintkyina. *International Journal of Applied Mathematics Electronics and Computers*, 8:263–267, 2020.