**Data-Driven Analysis and Forecasting of Solar Power Generation: A Case Study of Plants in Gandikotta and Nasik, India**

Aditya Jethani, Saurav Navdhare, Smit Sutariya

CSE & PDEU, INDIA

CSE & PDEU, INDIA

CSE & PDEU, INDIA

[aditya.ce21@sot.pdpu.ac.in](mailto:aditya.ce21@sot.pdpu.ac.in) ;[saurav.nce21@sot.pdpu.ac.in](mailto:saurav.nce21@sot.pdpu.ac.in) ;[smit.sce21@sot.pdpu.ac.in](mailto:smit.sce21@sot.pdpu.ac.in)



***The rising threat of climate change and costs of conventional energy have increased demand for sustainable alternatives like solar energy. This study analyzes solar energy potential, predictive modeling for forecasting, and develops a recommendation system for solar energy adoption and optimization at two solar power plants in Gandikotta, Andhra Pradesh, and Nasik, Maharashtra.***

***The analysis examines solar irradiance, ambient temperature, and historical power generation data to uncover patterns and trends. Random Forest Regressor and K-Nearest Neighbors models are employed for precise solar energy prediction across varying time horizons.***

***Building on analysis and prediction, a recommendation system provides tailored guidance to stakeholders interested in solar energy. It considers user profiles, location-specific data, financial factors, and technological aspects to match recommendations for optimal solar energy solutions.***

***The project integrates data analysis, predictive modeling, and the recommendation system into a user-friendly platform, empowering stakeholders to make informed decisions about solar energy investments, usage strategies, and sustainability initiatives.***

***Keywords: Solar energy, analysis, visualization, prediction, machine learning, sustainability, India, Gandikotta, Nasik, Random Forest Regressor, K-Nearest Neighbors***



1. **Introduction**

The accelerating shift towards sustainable energy sources is driven by the imperative to combat climate change, the rising costs of fossil fuels, and the desire for increased energy independence. Solar energy, with its limitless potential, zero-emission operation, and rapidly declining costs, is poised to play a transformative role in this transition. Global installed solar photovoltaic (PV) capacity surpassed 1.18 terawatts in 2022, with a record increase of 270 terawatt-hours (TWh) compared to the previous year (IEA, 2023). However, to capitalize on solar energy's benefits, accurate resource assessment, reliable forecasting, and personalized adoption strategies are paramount. Everyday earth receives sunlight above (1366W approx.) This is an unlimited source of energy which is available at no cost. The major benefit of solar energy over other conventional power generators is that the sunlight can be directly converted into solar energy with the use of the smallest photovoltaic (PV) solar cells [6].

Focusing on solar technology, photovoltaics have experienced enormous growth over the last years, amounting to a total installed capacity of around 177 GW worldwide by the end of 2014 (IEA, 2015) and growth is projected to continue at a similar rate in the future [1]. The global capacity of PV power installation increased from 5.1to 16.0 GW in 3 years (2005–2008) and continued to increase from 16.0 GW to 100.0 GW in the next 4 years (2008–2012) [2]. The inherent variability of solar power generation, influenced by factors such as diurnal cycles, cloud cover, and seasonal changes, complicates energy planning and investment decisions.

The project begins with meticulous analysis of key solar energy indicators, including historical solar irradiance, ambient temperature, geographical data, and technological specifications. Statistical techniques and visualizations will be used to uncover patterns, trends, and correlations within these datasets. This analysis phase will lay the foundation for developing robust predictive models.

Machine learning algorithms, such as K-Nearest Neighbors and Random Forests will be explored for their ability to forecast solar energy generation with high precision. These models will consider dynamic weather inputs and other relevant variables to improve forecasting accuracy across various time horizons.

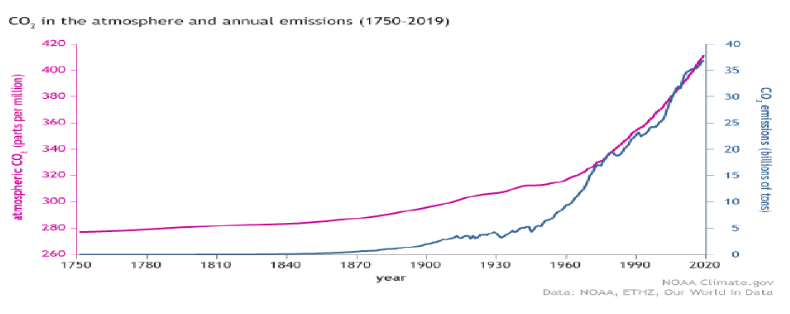
This system will provide stakeholders with personalized recommendations regarding solar energy adoption and system optimization, taking into account factors such as:

* Organizational energy consumption patterns
* Location-specific solar potential and site suitability
* Predictions of solar yield for Plant 1 (near Gandikotta, Andhra Pradesh) and Plant 2 (near Nasik, Maharashtra) for the immediate next day and the next seven days.
* Predictions of irradiation based on ambient and module temperature for both plants for the next immediate day and the next seven days.

This project holds the potential to make significant contributions to the field of solar energy. Ultimately, this project aims to contribute to a future where solar energy becomes a fundamental backbone of a sustainable, reliable, and affordable global energy system.

**II. Literature Review**

Energy can be scientifically defined as the ability or capacity of a system to accomplish work. In our everyday existence, energy manifests itself in numerous forms, including gravitational energy, kinetic or mechanical energy, electrical energy, heat, radiation, wave energy, chemical energy, and nuclear energy. The European Union sets the production target of renewable energy to the domestic reduction of GHG emissions at least 40% by 2030 and80% by 2050 (from a 1990 baseline) [2]. The intensity of solar energy outside the Earth's atmosphere is approximately constant at 1370 W/m², but varies between 0 and 1100 W/m² on Earth [8].



*CO2 in atmosphere and annual emission [4]*

The critical challenge is climate change driven by our reliance on fossil fuels like coal and gas. Fossil fuels are solar energy stored chemically over millions of years. Over 80% of the total energy consumed by humans is derived from fossil fuels, however, renewables are the fastest growing source of energy in the world [4]. However, we're depleting these reserves faster than they're being replenished naturally, making them unsustainable. Moreover, burning fossil fuels produces greenhouse gases like carbon dioxide, methane, and nitrous oxide, which accumulate in our atmosphere and oceans. Today, there are around 450 nuclear power plants worldwide. Whereas, in 1941, the first solar cell was invented. Today, there are over 75 solar thermal power stations around the world that have a combined capacity of over 4,810 megawatts, enough to power more than 1.7 million homes during peak hours [4].

Most of the recent studies in this ﬁeld have focused on investigating direct PV power forecasting. Direct forecasting methods can achieve accurate forecasting of PV power generation. Therefore, a comprehen-sive literature review based on recent direct forecasting methods,including model development and optimization, should be conducted for new researchers in this ﬁeld [2].

To harness this power requires the element that is second most abundant on Earth, i.e. silicon. Silicon is found abundantly in sand which undergoes a complex purification process. This is to obtain the 99.999% pure silicon crystals to use in solar cells [8]. The large magnitude of solar power available makes a highly appealing source of electricity. 30% (approx.) solar radiation is back to space while the rest is absorbed by ocean, clouds and land masses [5].

In [9] authors used a k-NN-based method for forecasting the power produced by small-scale PV plants installed in three different regions: SanDiego, Braedstrup and Catania. They concluded that simple techniques such as k-NNs can produce relatively accurate forecasts (the nMAE was in fact 0.96%). Different methods have been investigated in [7] including a grey-box model, NNs, k-nearest neighbors (kNNs), quantile random forest (QRF), SVM, and ensemble of methods (ENS). The application of these techniques gave similar performances showing a MAE close to 5%.

**III. Data Collection and Preprocessing**

**Data Acquisition:**

The data utilized in this research was sourced from open source resources, primarily for [Kaggle](https://www.kaggle.com/datasets/anikannal/solar-power-generation-data/data). This dataset also includes time-stamped entries, which are critical for assessing temporal patterns in solar energy generation. Plant 1 is located near Gandikotta in Andhra Pradesh, while Plant 2 is situated in the vicinity of Nasik, Maharashtra.

**Dataset Description:**

It encompasses detailed measurements of solar irradiance along with ambient and module temperatures collected every 15 minutes over a span of 34 days. The dataset also includes detailed weather data crucial for analyzing the impact of environmental conditions on solar energy output. It has data values for ambient and module temperature, Irradiation and Time Stamps of every 15 minutes. Solar Irradiation measurements are critical for determining the potential solar energy that can be converted into electricity. These data points are particularly important for understanding the variability in energy production related to weather conditions and the time of day.

**Variables Recorded:**

The dataset includes several key measurements:

* Date and Time: The exact timestamp of each measurement, critical for time series analysis.
* Plant ID: A unique identifier that distinguishes between data collected from Plant 1 and Plant 2.
* Source Key: Identifies the specific sensors or monitoring equipment within each plant.
* AC Power: Obtained AC power generated through indirect conversion of energy.
* DC Power: Obtained DC power generated through direct conversion of energy.
* Daily Yield: Measurement of energy yield obtained from each plant for each day.
* Total Yield: Cumulative total of yield generated.
* Ambient Temperature: The temperature of the air surrounding the solar panels, measured in degrees Celsius.
* Module Temperature: The temperature on the surface of the solar panels, measured in degrees Celsius.

**Ethical Considerations and Compliance:**

The collection and use of data followed stringent ethical guidelines, ensuring compliance with data privacy laws and environmental regulations. Necessary permissions were secured from the plant management for the use of this data, with all sensitive information being anonymized to protect proprietary interests.

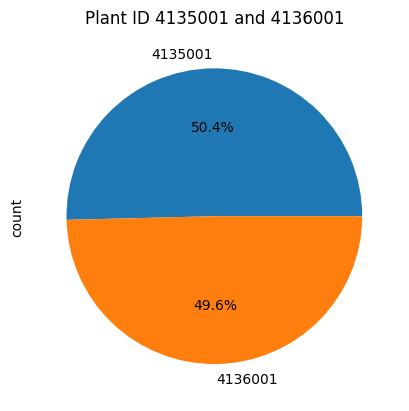
**Limitations of the Dataset:**

The dataset reflects specific operational and environmental conditions of the two distinct locations and may not be generalizable to all solar power plants. The data acquired is limited to information of solar yield and irradiation from 15th May 2020 to 17th June 2020, to which the predictions might change with current time.

**Data Preprocessing:**

Data preprocessing is an essential step in the analysis of complex datasets. It involves transforming raw data into an understandable format suitable for conducting detailed analyses. The preprocessing steps undertaken for this study are crucial to ensure the accuracy, completeness, and usability of the data collected from the two solar power plants. The first step in data preprocessing involved checking the dataset for any missing or incomplete information. Missing data can occur due to various reasons, such as malfunctions in the data recording equipment or transmission errors.

Upon observation, an inference was made that both the plants were actively operating from 6:00 AM till 6:30 PM on each day between 15th May 2020 and 17th June 2020. Naturally, we had to discard other recordings for obtaining quality data analysis.

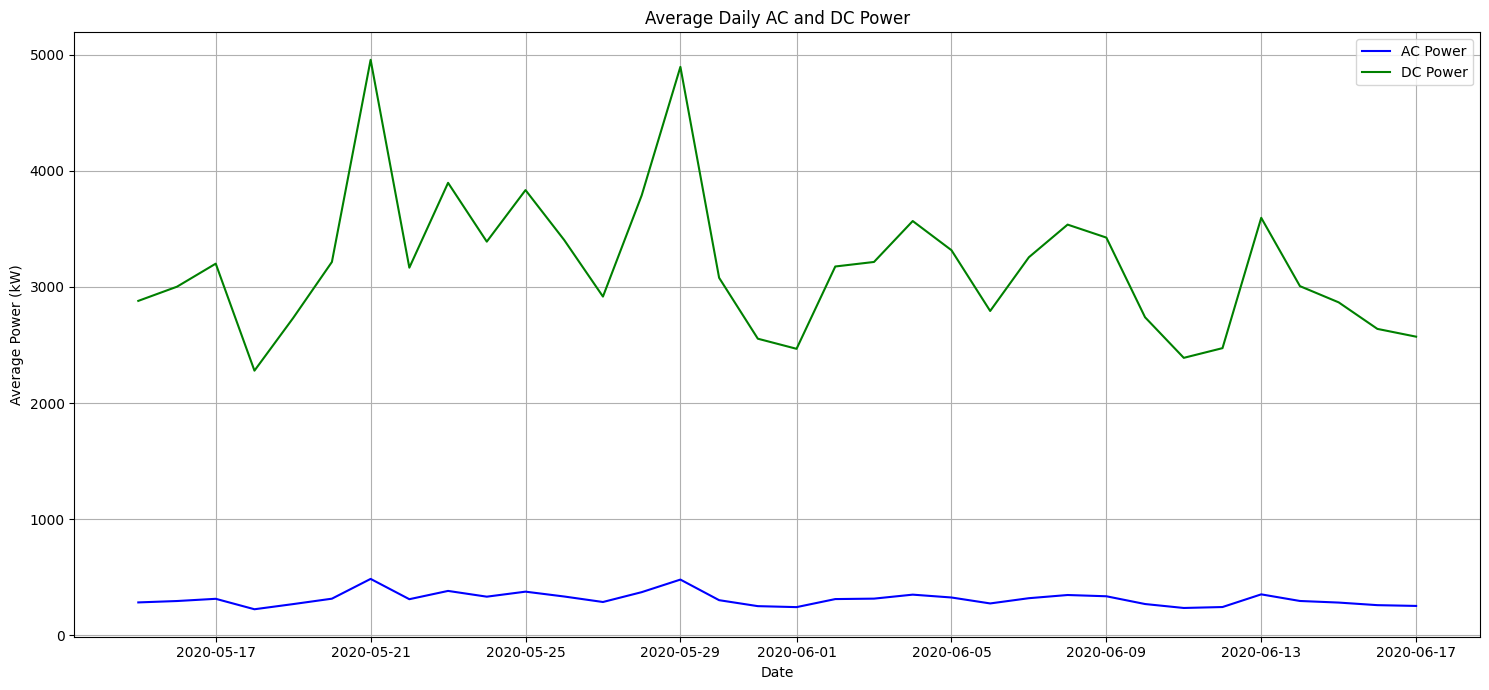


*Plant 1: 4135001 (Gandikotta) and Plant 2: 4136001 (Nasik)*

**IV. Data Analysis**

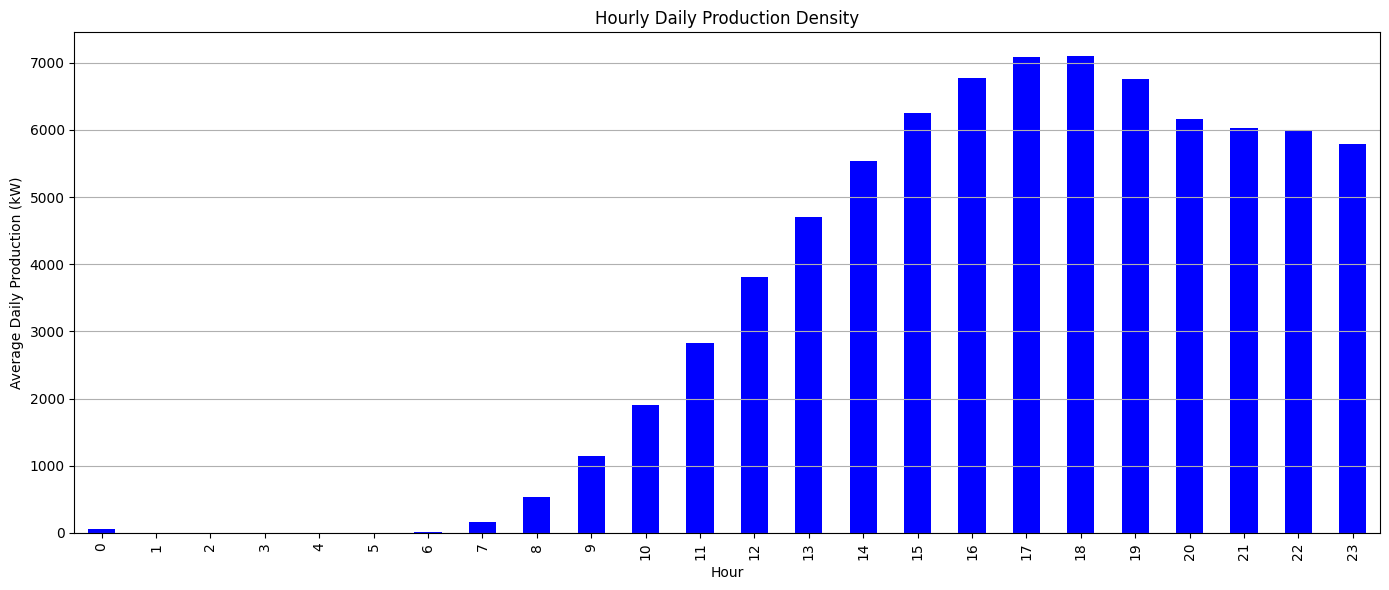
**Plant 1 (ID-4135001, Gandikotta) Power Generation Analysis:**

**Average Daily AC and DC Power:**



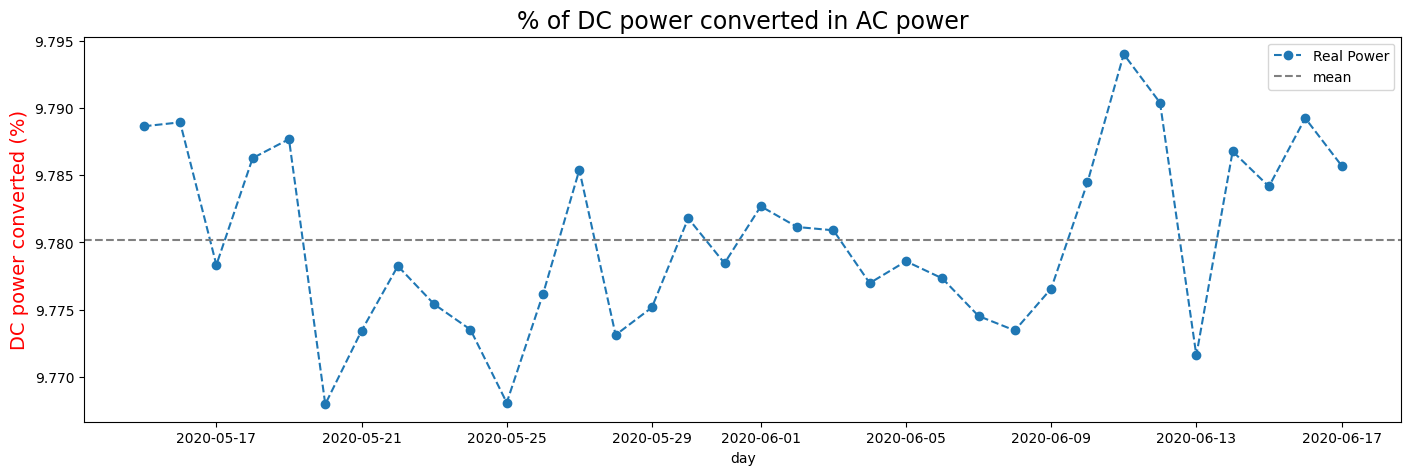
The line graph displays the average daily AC and DC power. AC power is represented by the blue line, while DC power is indicated by the green line. Both types of power exhibit fluctuations over time, making it challenging to discern a general trend.

**Hourly Daily Production Density:**

****

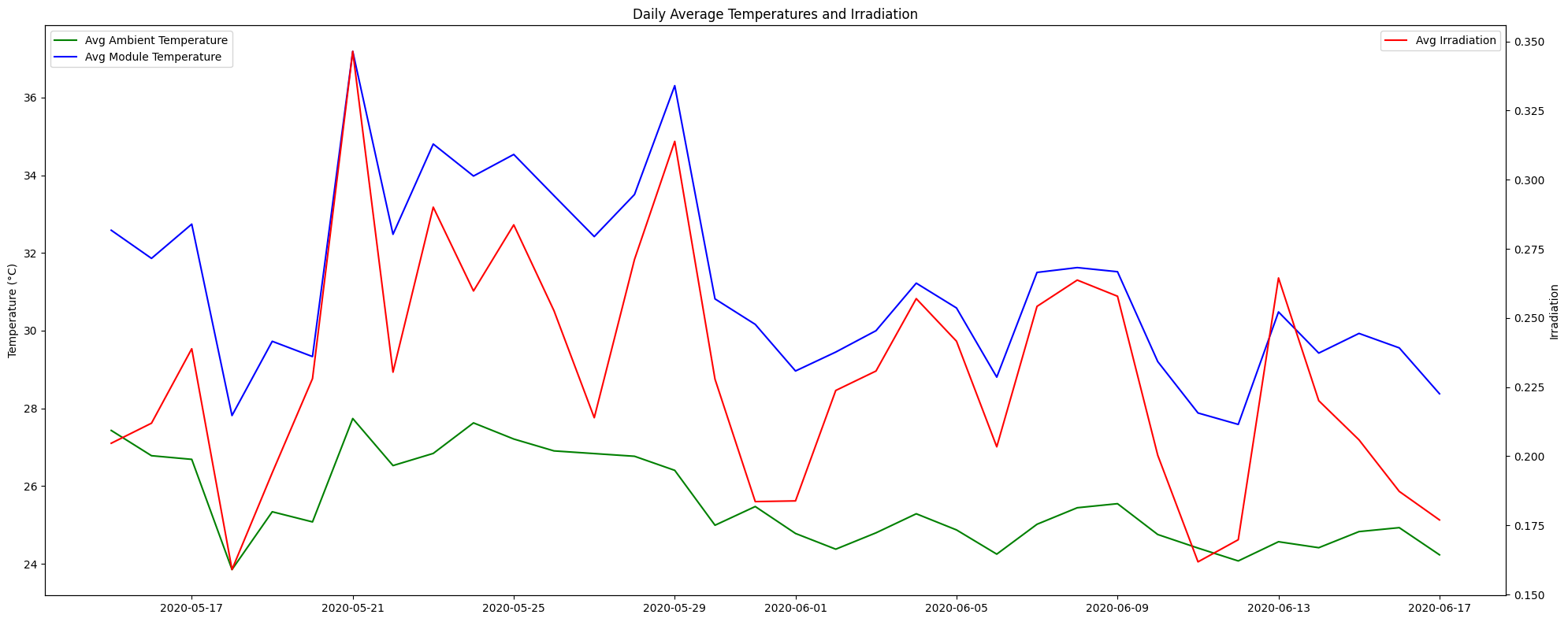
The bar graph shows the average daily energy production by hour. It provides a clear view of how daily production varies at specific hours.

**Percentage of DC power converted to AC power:**



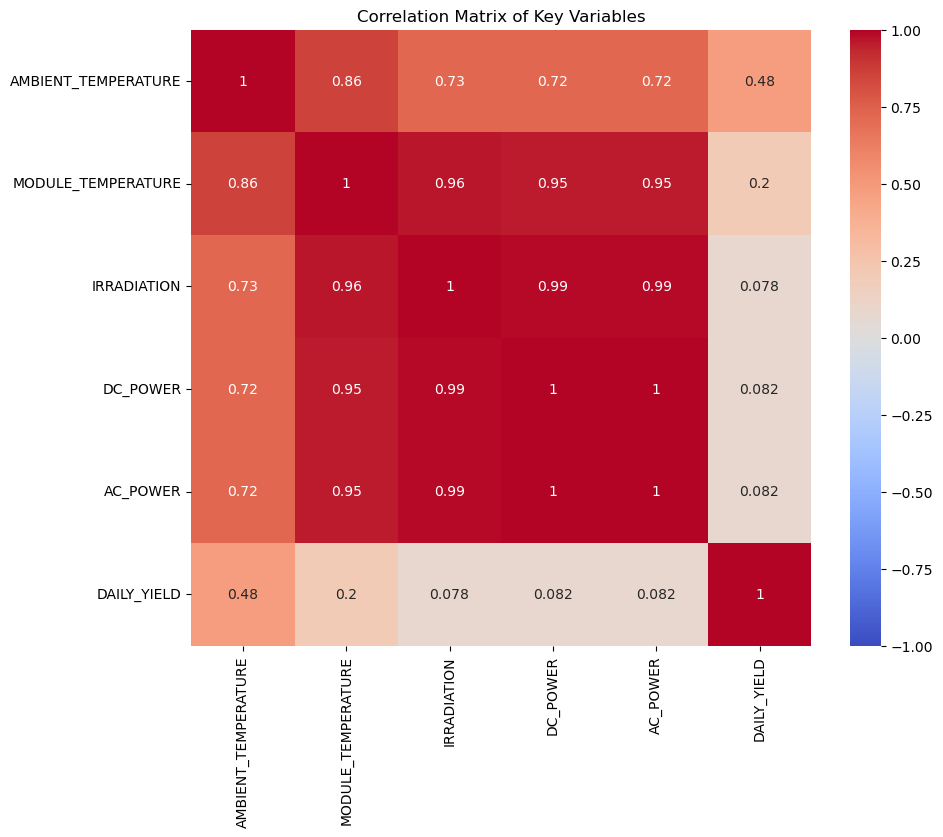
**Plant 1 (ID-4135001, Gandikotta) Weather Analysis:**

**Daily Average Temperatures and Irradiation:**

****

This chart displays the daily average ambient and module temperatures along with irradiation levels. The green and blue lines represent temperature values, while the red line represents irradiation levels. The chart aims to visualize the daily variations in temperature and irradiation levels.

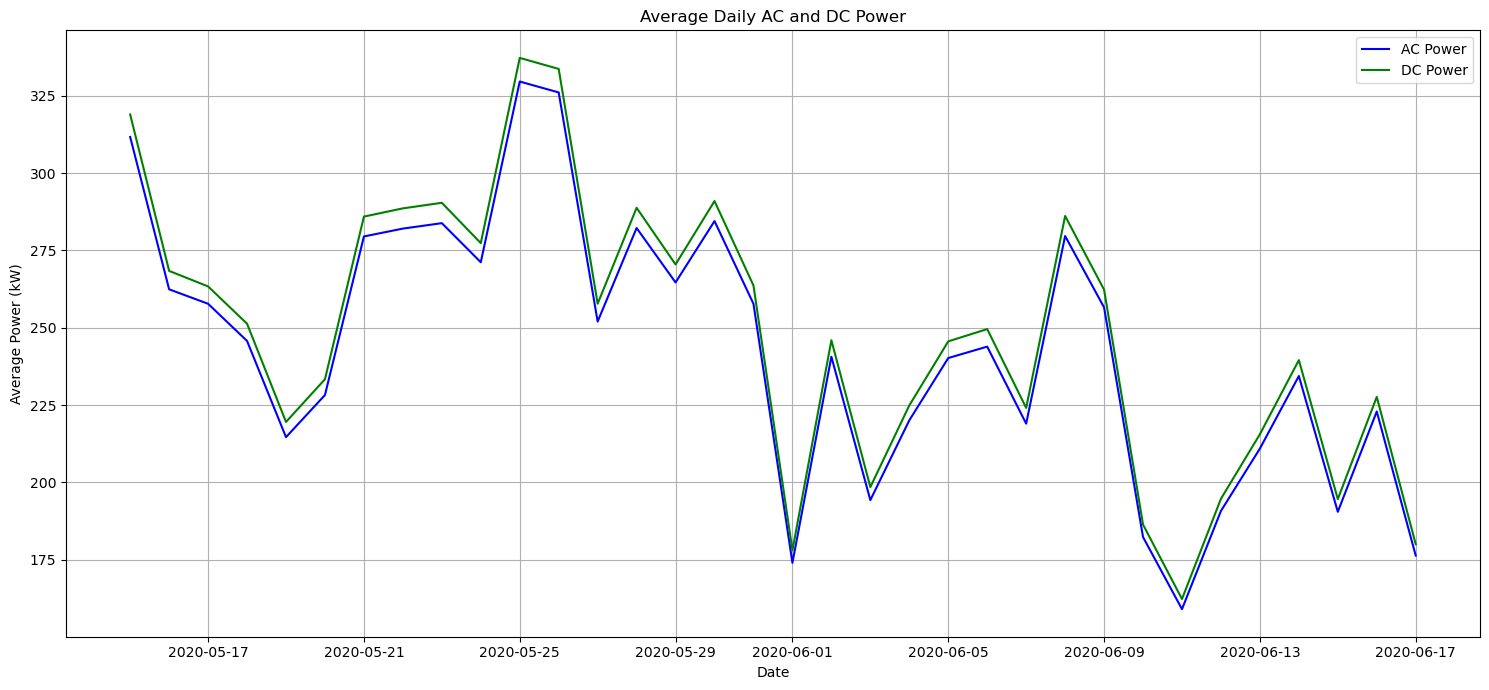
**Correlation Matrix of Key Variables:**

****

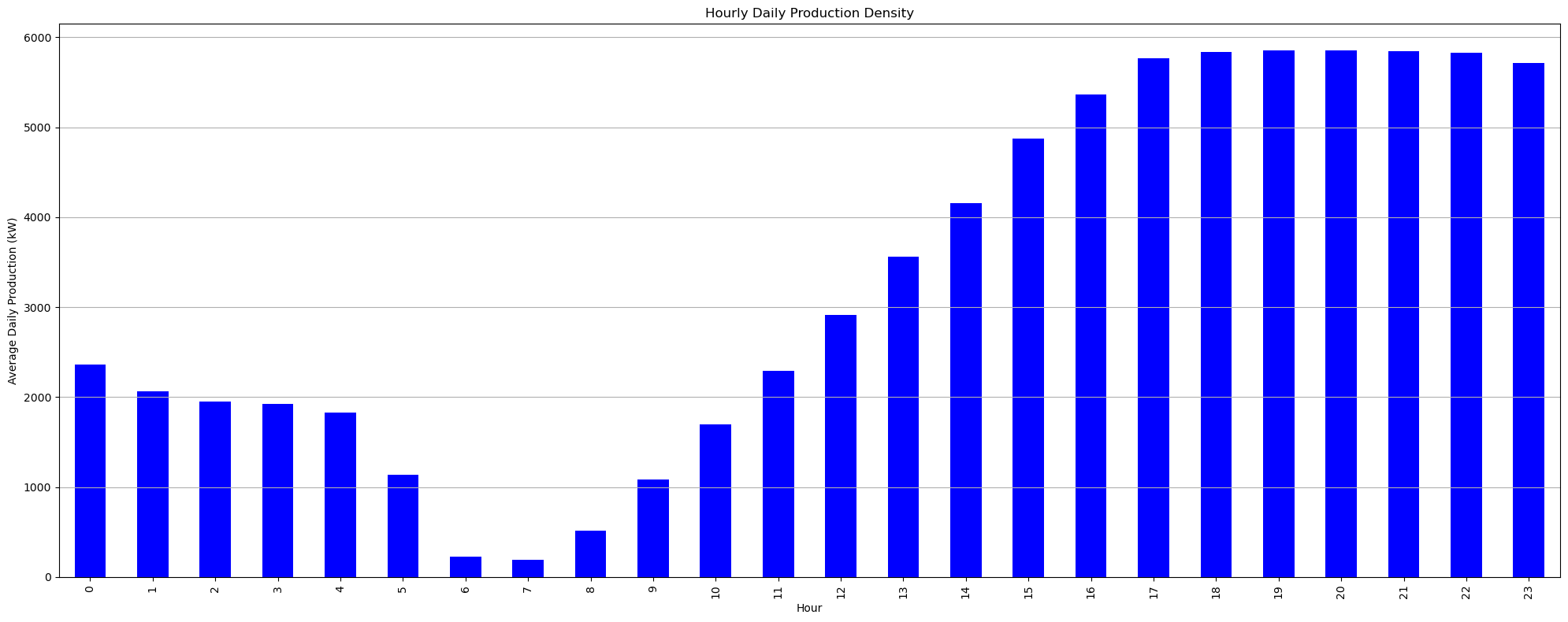
This heatmap shows the correlation among key variables. For instance, there is a high correlation between 'MODULE\_TEMPERATURE' and 'DC\_POWER'. Likewise, 'IRRADIATION' also shows a high correlation with energy production (both AC and DC power).

**Plant 2 (ID- 4136001, Nasik) Power Generation Analysis:**

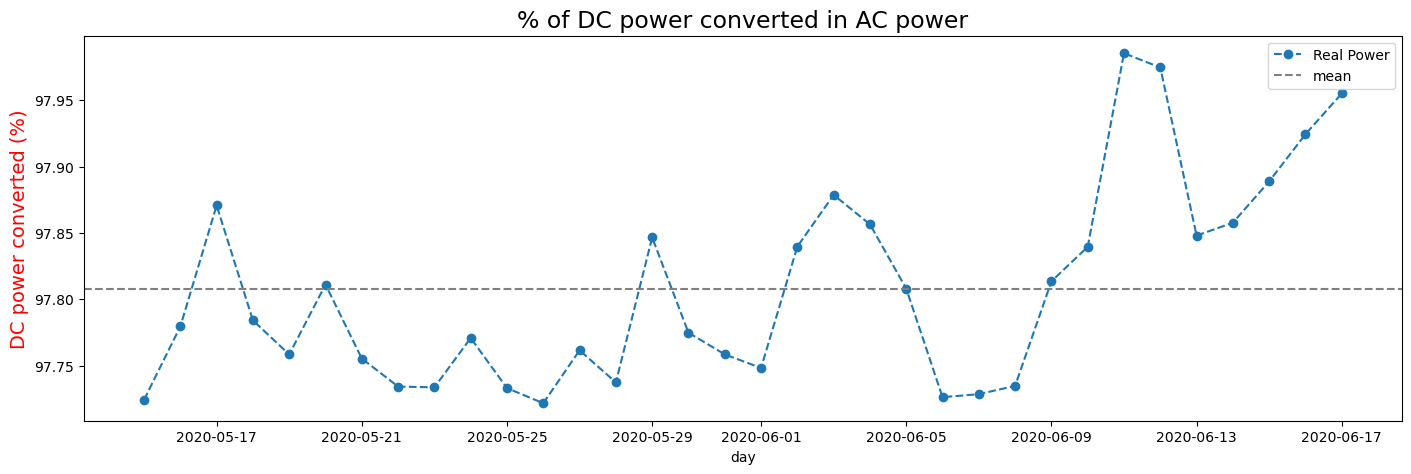
**Average Daily AC and DC Power:**

****

**Hourly Daily Production Density:**

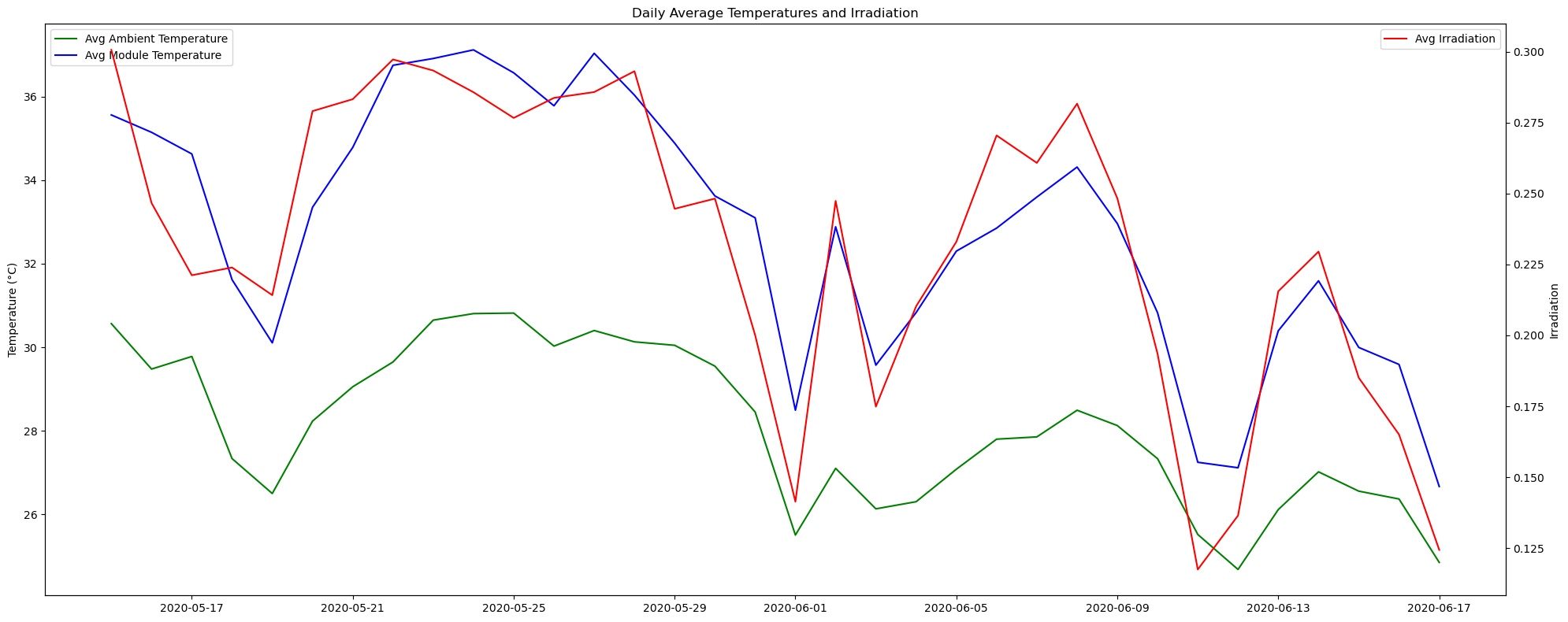
****

**Percentage of DC power converted to AC power:**

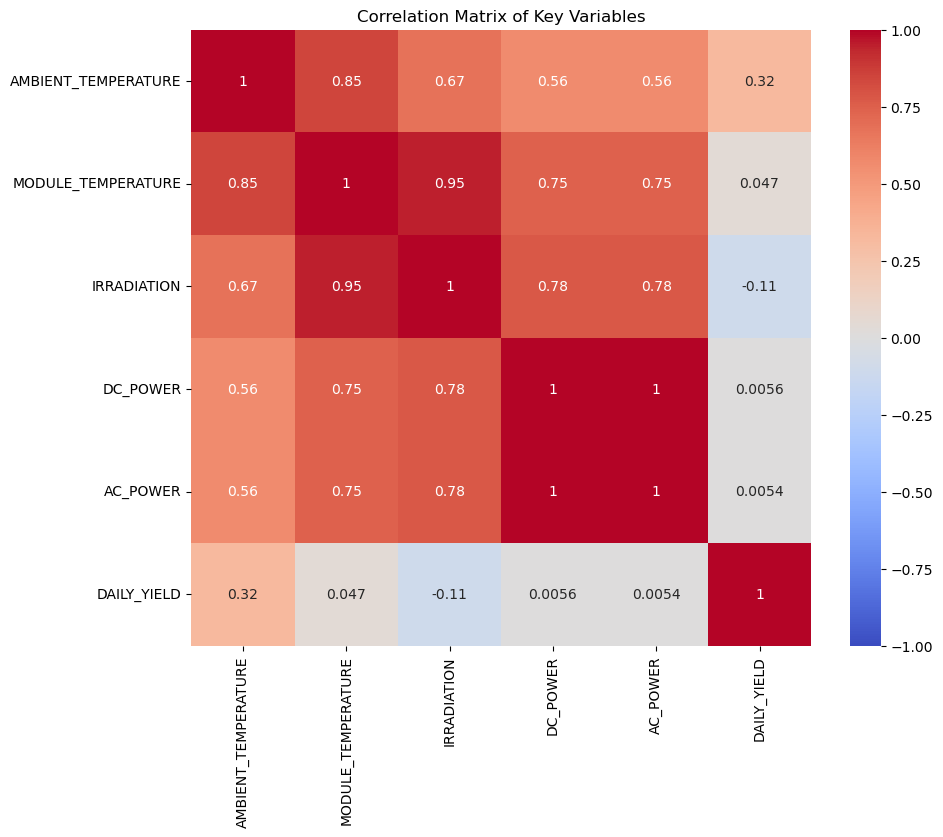
****

**Plant 2 (ID-4136001, Nasik) Weather Analysis:**

**Daily Average Temperatures and Irradiation:**

****

**Correlation Matrix of Key Variables:**

****

**Concluding Statement:**

The alignment between the analyses conducted for Plant 1 and Plant 2 reinforces the consistency and reliability of the derived insights. This alignment emphasizes the potential for applying comparable strategies and interventions to optimize renewable energy production efficiency across diverse settings

**V. Prediction Modelling**

**Model Development:**

This study employed two predictive models to forecast daily solar energy yield based on historical data: RandomForestRegressor and KNeighborsRegressor. These models were selected for their ability to handle non-linear data relationships typical in environmental data sets, providing robust predictive capabilities suitable for this research.

**Feature Selection and Data Split:**

The dataset was prepared by first transforming the DATE\_TIME column into a datetime format, facilitating the extraction of additional temporal features (day, month, hour). Features used in the prediction models included DC power, AC power, and the extracted time features, which were hypothesized to influence daily solar yield significantly. The dataset was split into training (70%) and testing (30%) sets, ensuring a substantial amount of data for both training the models and evaluating their performance.

**Model Training:**

The RandomForestRegressor and KNeighborsRegressor models were trained on the following features:

* DC\_POWER: The DC output of the solar panels.
* AC\_POWER: The AC output of the solar panels.
* day, month, hour: Temporal features derived from the DATE\_TIME field.

The following features were selected based on their potential impact on irradiation:

* AMBIENT\_TEMPERATURE: The surrounding air temperature, which can influence solar panel performance.
* MODULE\_TEMPERATURE: The temperature of the solar panels themselves, which is directly correlated with their efficiency.
* Temporal Features (day, month, hour): Time-related features to capture seasonal and daily patterns in irradiation.

Each model was configured with specific parameters:

* RandomForestRegressor: Set with 100 estimators to provide a balance between training speed and model complexity.
* KNeighborsRegressor: Utilized with 5 neighbors, the default setting, which was tested to ensure it provided the most reliable predictions for this particular dataset.

**Model Evaluation for Power Prediction:**

The performance of each model was quantified using three key metrics: Root Mean Squared Error, Mean Absolute Error, R-squared

Random Forest Regressor:

* RMSE: 232.06
* MAE: 134.63
* R2: 0.992

K Nearest Neighbour Regressor:

* RMSE: 1393.89
* MAE: 1001.21
* R2: 0.728

**Model Evaluation for Weather Prediction:**

Random Forest Regressor:

* RMSE: 0.084
* MAE: 0.059
* R-squared: 0.916

K Nearest Neighbour Regressor:

* RMSE: 0.086
* MAE: 0.059
* R-squared: 0.913

**Prediction for immediate next day:**

Power Yield:

* RandomForestRegressor: Predicted a daily yield of 2999.81 kWh.
* KNeighborsRegressor: Predicted a daily yield of 4934.65 kWh.

Irradiance:

* RandomForestRegressor: 0.43 kW/m²
* KNeighborsRegressor: 0.49 kW/m²

These predictions are indicative of the models' utility in forecasting daily power yield and solar irradiation, which is critical for planning and optimizing solar energy production.

**VI. Conclusion and Discussion**

The research analyzed solar power generation and weather data from solar plants in Gandikotta and Nasik. It applied machine learning models RandomForestRegressor and KNeighborsRegressor to forecast the next day's solar energy output and irradiation levels. The RandomForestRegressor demonstrated superior performance, particularly in predicting daily solar energy yield, achieving an R-squared value near 0.992, indicating very high accuracy. The KNeighborsRegressor showed comparatively lower performance.

The study also explored the models' predictive capabilities in estimating solar irradiation based on weather data. Both models performed robustly, with RandomForest slightly outperforming KNeighbors, confirming their effectiveness in utilizing ambient and module temperatures, along with temporal features, to accurately predict short-term solar irradiation.

The research shows how data analysis and predictive modeling can better manage solar power plants and improve solar energy predictability, supporting solar integration into grids for reliable, sustainable renewable energy.

References

1. [Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. Solar Energy, 136, 78-111.](https://www.sciencedirect.com/science/article/abs/pii/S0038092X1630250X)
2. [Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., ... & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews, 81, 912-928.](https://www.researchgate.net/publication/319003667_Forecasting_of_Photovoltaic_Power_Generation_and_Model_Optimization)
3. [Mellit, A., Pavan, A. M., Ogliari, E., Leva, S., & Lughi, V. (2020). Advanced methods for photovoltaic output power forecasting: A review. Applied Sciences, 10(2), 487.](https://www.researchgate.net/publication/338510612_Advanced_Methods_for_Photovoltaic_Output_Power_Forecasting_A_Review)
4. [Kafle, Mandip & Rai, Prabhat & Gurung, Nelson & Magar, Sandip. (2020). An analysis on Solar Energy. 10.13140/RG.2.2.11454.31043/1.](https://www.researchgate.net/publication/346059245_An_analysis_on_Solar_Energy)
5. [Shaikh, Mohd Rizwan & Shaikh, Sirajuddin & Waghmare, Santosh & Labade, Suvarna & Tekale, Anil. (2017). A Review Paper on Electricity Generation from Solar Energy. International Journal for Research in Applied Science and Engineering Technology. 887. 10.22214/ijraset.2017.9272.](https://www.researchgate.net/publication/320226399_A_Review_Paper_on_Electricity_Generation_from_Solar_Energy)
6. [Zhang, Yue & Beaudin, Marc & Taheri, Raouf & Zareipour, Hamidreza & Wood, David. (2015). Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators. IEEE Transactions on Smart Grid. 6. 1-1. 10.1109/TSG.2015.2397003.](https://www.researchgate.net/publication/276489042_Day-Ahead_Power_Output_Forecasting_for_Small-Scale_Solar_Photovoltaic_Electricity_Generators/citation/download)
7. [Gigoni, L.; Betti, A.; Crisostomi, E.; Franco, A.; Tucci, M.; Bizzarri, F.; Mucci, D. Day-ahead hourly forecasting of power generation from photovoltaic plants. IEEE Trans. Sustain. Energy 2017, 9, 831–842](https://arxiv.org/pdf/1903.06800.pdf)
8. [Namazov, Rashad & Popal, X. (2023). The role of solar panels in energy production. 2022.](https://www.researchgate.net/publication/368510806_The_role_of_solar_panels_in_energy_production)