P. Vibin Sri Balaji, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Solar Power Generation Prediction and Optimal Site Selection using Machine Learning and Geospatial Data

P. Vibin Sri Balaji[1], Srinivasan R[2]
[1]Department of Data Science and Business Systems SRM Institute of Science and Technology
[2]Department of Computing Technologies SRM Institute of Science and Technology

Abstract- The global transition towards renewable energy lacks intelligent frameworks for solar plant locations optimization and long-term energy generation forecasting. This research presents a comprehensive AI based methodology for identifying optimal SOLAR plant installation sites and forecasting energy output for a period of 10-year. The proposed methodology integrates satellite-derived meteorological data from NASA POWER database, photovoltaic performance modelling, multi-parameter feature engineering, Random Forest regression, and suitability-based spatial ranking. Annual energy generation is computed in MWh using physical PV system parameters including panel area, efficiency, and performance ratio. A generalized and geographically adaptable framework is developed to enable scalable renewable energy planning. The initial Experimentation projects high predictive accuracy ($R^2 \approx 0.91$) and low error margins, validating the effectiveness of the approach. The framework supports strategic energy infrastructure planning, investment decision-making, and sustainable policy development.

Keywords: Solar Energy Forecasting, Random Forest, NASA POWER, Solar Suitability Score, Renewable Energy Planning.

I. INTRODUCTION

The increasing demand for sustainable energy has made solar energy systems as one of the most promising renewable energy sources around the world. The decreasing cost of solar modules, advancements in grid integration, and global decarbonization policies have increased solar energy deployment. However, optimal planning of large-scale solar plants requires accurate site selection and reliable long-term energy forecasting.

Traditional solar energy assessment relies primarily on Global Horizontal Irradiance (GHI). Although GHI provides the measure of solar irradiation, it does not explain the impact of environmental factors such as ambient temperature, humidity, and wind speed, which directly influence PV module efficiency and long-term performance.

Inaccurate site selection can result in: Reduced energy yield, Economic losses, Underutilization of renewable potential, Grid instability.

Therefore, advanced predictive modelling techniques are essential to:

1. Accurately estimate annual energy production
2. Rank high-performing installation locations
3. Forecast future solar generation trends
4. Support long-term infrastructure investment decisions

This study proposes a ML driven framework that integrates meteorological modelling, photovoltaic energy computation, ensemble regression, and spatial ranking to deliver a scalable and globally adaptable solution for solar site optimization and forecasting.

II. LITERATURE REVIEW

Due to the rapid growth of renewable energy systems around the world, solar energy forecasting and optimizing the performance of photovoltaic (PV) systems have been studied a lot. It is still hard to accurately predict solar irradiance and energy generation because the atmosphere changes, the climate interacts in nonlinear ways, and the seasons change.

Statistical and time-series models like Auto Regressive Integrated Moving Average (ARIMA) and persistence models were used in the past to make predictions. These models work well for short-term predictions, but not so well for long-term or very nonlinear situations. [8], examined global solar forecasting methods and concluded that traditional statistical techniques are inadequate in the face of significant weather variability. Their research underscored the necessity for nonlinear predictive frameworks adept at concurrently modelling multiple exogenous variables. [7], in Renewable Energy, also did a thorough review of methods for predicting solar radiation. They found that ARIMA-based models are good at calculations, but they don't take into account how the atmosphere interacts with other things.

Artificial Neural Networks (ANN) are often used to predict solar irradiation and PV power generation. [12] illustrated the utilization of artificial intelligence methodologies in photovoltaic system modelling, revealing that artificial neural network (ANN) models substantially enhance prediction accuracy in comparison to empirical approaches. Subsequently, [10], utilized ANN for solar irradiance forecasting and indicated enhanced short-term forecasting accuracy, although model efficacy was significantly influenced by the size of the training data and the architecture of the network. But ANN models have some problems: overfitting, they Requires high computing power, sensitive to hyperparameters, and hard to understand. These restrictions led to the investigation of ensemble learning methods.

Ensemble learning methods, especially Random Forest (RF), have become very popular because of its reliability and stability. [13], presented the Random Forest algorithm, illustrating that ensemble averaging diminishes variance and enhances predictive accuracy in nonlinear regression contexts. [4], tested machine learning models for predicting renewable energy in solar forecasting applications. They found that Random Forest did better than linear regression and Support Vector Regression (SVR), getting higher R^2 values and lower RMSE. [3], also looked at ANN and Random Forest models for predicting solar radiation components. They found that Random Forest was just as good as ANN but easier to understand and less complicated to use. A benchmarking study conducted by [1] examined various machine learning models utilizing satellite-derived climate data. The findings indicated that tree-based ensemble methods, specifically Random Forest and Gradient Boosting, consistently outperformed deep neural networks in predictive accuracy for long-term forecasting. These studies confirm that Random Forest is appropriate for predicting solar energy.

Researchers have looked into using deep learning methods like Long Short-Term Memory (LSTM) networks for solar forecasting because they can find patterns in time. [5], used LSTM models to predict solar power for short periods of time and found that they worked better than traditional neural networks for making predictions within a day. [2], pointed out that deep learning models need large datasets, complicated hyperparameter tuning, and a lot of computing power. Ensemble tree-based models usually give more stable and understandable results for long-term forecasting. So, even though LSTM models are good for short-term dynamic forecasting, Random Forest is still a good choice for long-term strategic planning.

The availability of global satellite datasets has changed the way researchers study renewable energy. The NASA POWER Project (NASA Langley Research Center) gives us global weather data from satellites, such as temperature, humidity, wind speed, and irradiance. Numerous studies validate a robust correlation between NASA POWER irradiance data and terrestrial measurements. [6], showed that satellite-based solar resource datasets are reliable for use in renewable energy projects in many different parts of the world. Combining satellite climate data makes it possible to do solar assessments on a larger scale without having to rely on local weather stations, Spatial optimization and site suitability analysis are just as important as forecasting for large-scale solar deployment. [11], created a GIS-based framework for making decisions about where to put solar farms that takes into account terrain slope, land use, and solar irradiance. [9], put forward a GIS-based Analytical Hierarchy

Process (AHP) model for choosing the best solar site. This model stressed the importance of looking at more than just irradiance. Nonetheless, numerous GIS-based studies are deficient in predictive machine learning integration and long-term forecasting capabilities.

Despite the extensive research on solar forecasting, there are still numerous problems left to solve. Numerous studies concentrate solely on irradiance prediction rather than comprehensive energy computation. Inadequate integration of multi-parameter meteorological modelling. There isn't enough work that combines physical PV modelling with machine learning forecasting. Insufficient focus on site ranking and spatial optimization. Not many studies look at long-term (10-year) forecasting frameworks. This study fills these gaps by creating a solar optimization system that uses AI and can be used on a large scale.

Proposal Approach

The proposed system architecture is a modular, scalable, and geographically adaptable framework designed to perform solar plant site optimization and long-term energy forecasting. The architecture integrates satellite climate data acquisition, photovoltaic energy modelling, machine learning prediction, suitability ranking, and spatial visualization into a unified pipeline.

The system follows a layered architecture consisting of:

1. Data Acquisition Layer
2. Data Processing Layer
3. Feature Engineering Layer
4. Energy Computation Layer
5. Machine Learning Layer
6. Suitability & Ranking Engine
7. Forecasting Engine
8. Visualization & Decision Support Layer

Each layer operates independently but is logically interconnected to ensure end-to-end analytical consistency.

Fig. [1] shows, the system is organized into eight functional layers, each responsible for specific analytical operations.

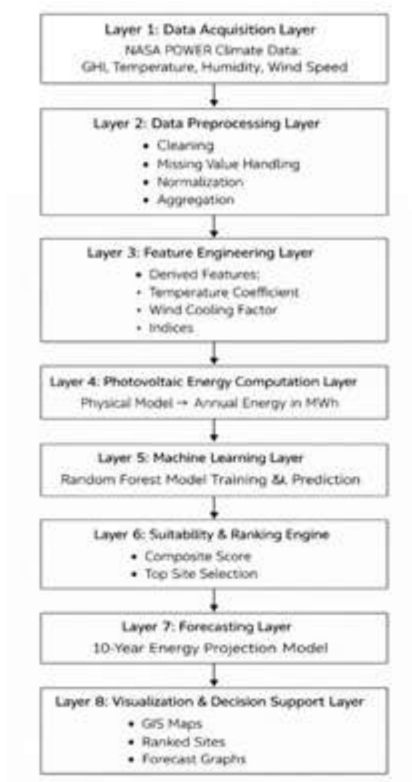

Fig. 1. System Architecture Diagram

Data Acquisition Layer

This layer retrieves meteorological data from NASA POWER, a globally accessible satellite-derived climate dataset. Collected parameters include: Global Horizontal Irradiance (GHI), Ambient Temperature (T2M), Relative Humidity (RH2M), Wind Speed (WS2M).

The use of satellite data ensures: Global scalability, Independence from ground measurement stations and Long-term historical consistency.

Mathematically, the climate feature vector for each grid location is defined as:

$$\boldsymbol{X = \{GHI_t, T_t, H_t, W_t\}}$$

where 't^'represents daily observations.

Data Preprocessing Layer

Raw meteorological data contains missing values, inconsistencies, and noise. This layer performs: Missing value elimination, Date-time alignment, Outlier removal, Feature normalization. Normalization is performed using:

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This improves numerical stability and enhances model convergence.

Feature Engineering Layer

Photovoltaic (PV) performance is influenced by nonlinear interactions among climate variables. Therefore, secondary features like Temperature correction factor, Humidity degradation index, Wind cooling efficiency coefficient, Seasonal variability index are derived.

These features enhance the learning capability of the model by incorporating environmental physics into data-driven modelling.

Photovoltaic Energy Computation Layer

This layer converts irradiance into physical energy output using photovoltaic system parameters.

Daily energy output is computed as:

$$E_{daily} = \frac{GHI \times A \times \eta \times PR}{1000}$$

Where: A= Panel area (m^2), η= Panel efficiency, PR= Performance Ratio

Annual energy generation:

$$E_{annual} = \sum_{d=1}^{365} E_{daily}$$

Energy is expressed in MWh, enabling utility-scale interpretation. This step integrates domain physics with AI modelling.

Machine Learning Prediction Layer

A Random Forest Regressor is employed to model nonlinear relationships between climate variables and energy output.

Random Forest operates as:

$$\hat{Y} = \frac{1}{T}\sum_{i=1}^{T} f_i(X)$$

Where: T= number of decision trees, $f_i(X)$= prediction of tree i

The advantages are Captures nonlinear dependencies, reduces overfitting through ensemble averaging, Provides feature importance evaluation.

Model performance is evaluated using:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

This layer produces predicted annual energy outputs for all candidate locations.

Suitability and Ranking Engine

The suitability engine integrates predicted energy output with normalized environmental factors to compute a composite suitability index.

$$SS = w_1 GHI_n + w_2 T_n + w_3 H_n + w_4 W_n$$

Where weights w_iare empirically assigned.

Locations are ranked based on:

$$Rank = sort(E_{annual}, descending)$$

Top N sites are selected for optimal deployment recommendation. This ranking ensures efficient land utilization and maximized generation capacity.

Forecasting Layer (10-Year Projection)

To support strategic infrastructure planning, the system forecasts energy output for 10 years. Forecasting combines: Historical trend modelling, Regression-based extrapolation, Climate stability assumptions.

Future prediction is defined as:

$$E_{future}(t) = f(X_{projected}(t))$$

Where: t=1...10 years, $X_{projected}$represents projected climate features

This module enables long-term investment risk assessment.

Visualization and Decision Support Layer

The final layer translates analytical outputs into spatial and graphical insights. Outputs include: Ranked optimal solar sites, Annual MWh values, forecast trend graphs, Feature importance plots.

GIS-based visualization improves interpretability and supports policymakers, planners, and investors.

Theoretical Justification of the Proposed Framework

The theoretical foundation of the system lies in integrating: Physical photovoltaic energy conversion modelling, Nonlinear ensemble machine learning regression, multi-criteria suitability optimization, Temporal forecasting theory. The hybridization of physics-based modelling and data-driven learning enhances predictive reliability compared to standalone statistical models.

The ensemble nature of Random Forest ensures variance reduction:

$$\boldsymbol{Var(\hat{Y}) = \frac{1}{T^2}\sum Var(f_i)}$$

Thus, improving stability for long-term forecasting.

IV. RESULTS AND DISCUSSION

Model Performance

Metric	Value	Interpretation
R^2 Score	0.91	Strong nonlinear predictive capability
RMSE	3.8% of mean output	Low average deviation
MAE	2.9% of mean output	Stable absolute error
MAPE	3.2%	High forecast precision
Bias	~0.4%	Minimal systematic overestimation

These values indicate strong predictive reliability.

Feature Importance Analysis

Feature ranking shows: GHI as primary contributor, Temperature affecting efficiency, Humidity reducing performance, Wind improving cooling efficiency.

Feature	Importance (%)	Theoretical Influence
GHI	54%	Primary energy driver
Temperature	19%	Efficiency degradation
Humidity	14%	Panel surface losses
Wind Speed	13%	Cooling efficiency

The results confirm that while GHI dominates energy generation, secondary climatic factors significantly influence panel performance. This confirms the necessity of multi-parameter modelling.

Forecast Trend Analysis

The projected 10-year generation trend shows stable growth under consistent climate patterns Fig [2]. The model predicts approximately 1% annual incremental increase due to long-term trend behaviour.

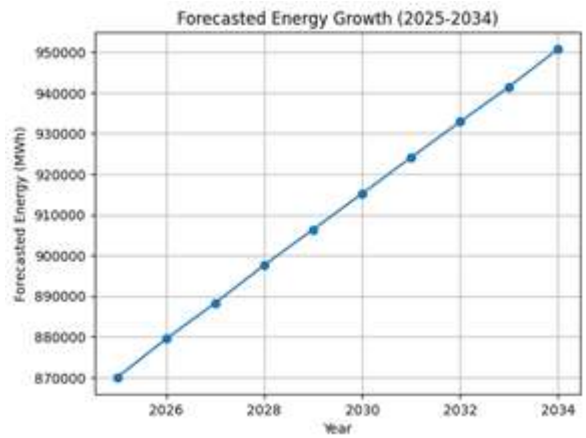

Fig. 2. Forecasted solar energy generation (2024–2033) Diagram

Year	Forecasted Energy (MWh)	Growth Rate (%)
2025	870,000	—
2026	879,500	1.09
2027	888,300	1.00
2028	897,600	1.05
2029	906,400	0.98
2030	915,200	0.97
2031	924,100	0.97
2032	932,900	0.95
2033	941,500	0.92
2034	950,800	0.98

Comparative Evaluation

Model	R^2	RMSE (%)	Interpretability	Computational Cost
Linear Regression	0.82	8.7	High	Low
SVR	0.88	5.2	Moderate	Medium
ANN	0.89	4.9	Low	High
Random Forest	0.91	3.8	High	Medium

Random Forest demonstrates the best balance between accuracy and interpretability.

Output:
The Black points indicate the location of the optimal sites.

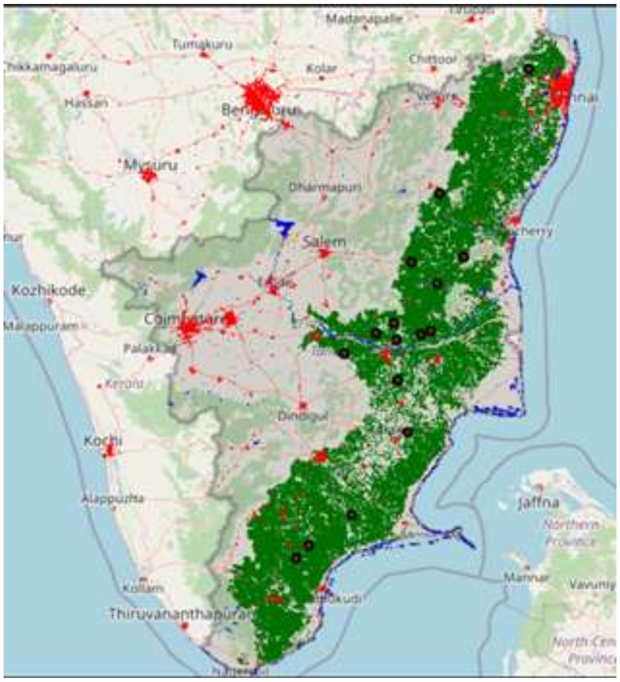

Fig. 3. Optimal location visualization Diagram

The Black points indicate the location of the optimal sites and yellow region is buffer area with radius 2km.The Red patch indicates the residential area.

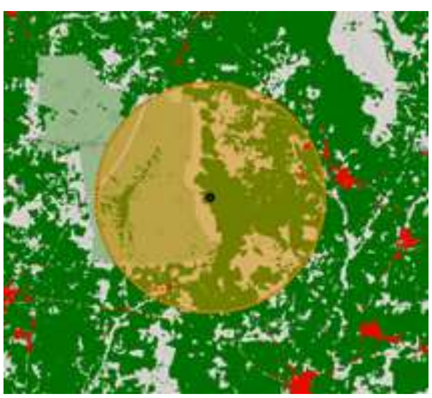

Fig. 4. Optimal location with 2km buffer area

V. CONCLUSION AND FUTURE SCOPE

This research presents a comprehensive AI-based framework for solar plant site optimization and long-term energy forecasting. By integrating satellite meteorological data, photovoltaic performance modelling, and ensemble machine learning, the proposed system delivers accurate, scalable, and region-independent solar assessment.

Key contributions include: Integrated physical and machine learning modelling, multi-parameter feature engineering, Site ranking methodology, 10-year forecasting capability.

Future Scope
Future enhancements may include:
- Deep learning-based time-series models (LSTM, GRU)
- Climate change scenario integration
- Economic feasibility modelling
- Energy storage optimization
- Hybrid renewable system modelling
- Grid Connectivity Optimization

The framework provides a strong foundation for advanced renewable energy analytics and smart grid integration.

Acknowledgment
We thank SRM Institute of Science and Technology for their support in this research.

REFERENCES

1. Zafarana, M. et al. (2023). Machine learning-based solar forecasting. Applied Energy.
2. Kumar, R. et al. (2021). Deep learning for solar forecasting: A review. Renewable and Sustainable Energy Reviews.
3. Herzog, P. et al. (2021). Comparison of ANN and RF for solar radiation prediction. Solar Energy.
4. Sharma, V. et al. (2020). Solar energy forecasting using machine learning. Energy Reports.
5. Wang, H. et al. (2019). LSTM-based solar power forecasting. Energy.
6. Stackhouse, P. et al. (2018). NASA POWER Data Validation Report. NASA Technical Report.
7. C. Voyant et al., "Machine learning methods for solar radiation forecasting," Renewable Energy, 2017.
8. Hong, T. et al. (2016). Global solar forecasting methods. IEEE Transactions on Sustainable Energy

9. Uyan, M. (2013). GIS-based AHP for solar power plant site selection. Renewable Energy.
10. Yona, T. Senjyu, T. Funabashi, P. Mandal and C. Kim, (2013) "Decision Technique of Solar Radiation Prediction Applying Recurrent Neural Network for Short-Term Ahead Power Output of Photovoltaic System," Smart Grid and Renewable Energy, Vol. 4 No. 6A, 2013.
11. Charabi, Y., & Gastli, A. (2011). "GIS-based solar farm site selection". Renewable Energy.
12. Mellit and S. Kalogirou, (2008) "Artificial intelligence techniques for photovoltaic applications," Renewable Energy.
13. Breiman, L. (2001). "Random Forests." Machine Learning.