An Empirical Study of Tree-Based and Instance-Based Models for Short- and Long-Term Solar Energy Forecasting

Abhijit Warhade[1], Dr. Manoj Demde[2], Dr. V. Taksande[3], Mrs. Pranali A. Warhade[4]

[1]ECE Department Priyadarshini College of Engineering Nagpur, India, abhiwarhade1984@gmail.com
[2]ETC Department, Priyadarshini College of Engineering, Nagpur, India, manoj_demde@yahoo.co.in
[3]ETC Department, Priyadarshini College of Engineering, Nagpur, India, virendrataksande2@gmail.com
[4]ETC Department, Priyadarshini College of Engineering, Nagpur, India, Pranaliwarhade4@gmail.com

Abstract
The integration of solar energy into modern power grids necessitates accurate forecasting across multiple time horizons to ensure grid stability and operational efficiency. This paper presents a comprehensive empirical investigation of tree-based ensemble methods and instance-based learning approaches for solar irradiance forecasting. We evaluate random forests, gradient boosting, evolutionary forests, and quantile regression forests against instance-based methods including k-nearest neighbours and regime-dependent artificial neural networks. Experiments are conducted across six climatically diverse locations in Morocco and three sites with varying meteorological variability. Results demonstrate that tree-based ensemble methods consistently outperform instance-based approaches across most forecasting horizons (1–6 hours), with the proposed evolutionary forest model achieving n RMSE values between 4.94% and 18.94% depending on climatic conditions. Hybrid input configurations combining endogenous and exogenous variables yield superior accuracy across all models. Our findings indicate that tree-based methods exhibit particular advantages in high-variability conditions and when training data is limited, while instance-based methods show competitive performance primarily in stable, clear-sky conditions.

Keywords: Solar energy forecasting; tree-based models; instance-based learning; random forest; evolutionary forest; quantile regression; short-term forecasting

1. Introduction

The global energy transition toward renewable sources has accelerated dramatically over the past decade, with solar photovoltaic capacity experiencing unprecedented growth. As of 2018, total installed solar capacity reached 180 GW worldwide, and this figure has continued to rise substantially . However, the inherent intermittency of solar irradiance—driven by cloud cover, atmospheric conditions, and diurnal cycles—presents significant challenges for grid operators who must balance supply and demand in real time.

Accurate solar forecasting has therefore become a critical enabler of high-renewable grids. Forecast horizons span multiple temporal scales: very short-term (minutes to 1 hour) supports automatic generation control; short-term (1–6 hours) facilitates intra-day grid scheduling and unit commitment; and long-term (days to weeks) informs maintenance planning and energy trading . Each horizon presents distinct modeling challenges, with short-term forecasting being particularly demanding due to the rapid, nonlinear dynamics of cloud evolution.

Machine learning methods have emerged as powerful tools for solar forecasting, offering the ability to learn complex nonlinear relationships without explicit physical parameterizations. Among these approaches, two broad families have gained prominence: tree-based ensemble methods (random forests, gradient boosting, evolutionary forests) and instance-based learning methods (k-nearest neighbors, analog ensembles, regime-dependent neural networks). While both families have demonstrated success in various contexts, their relative performance across different forecasting horizons, climatic conditions, and input configurations remains insufficiently characterized.

This paper provides a systematic empirical comparison of these approaches, addressing the following research questions:How do tree-based ensemble methods compare to instance-based learning approaches for solar irradiance forecasting across horizons from 1 to 6 hours?

What is the influence of climatic variability on the relative performance of these model families?

How do hybrid input configurations (combining endogenous historical data with exogenous meteorological variables) affect model accuracy?

Can evolutionary optimization of tree ensembles improve upon standard random forest performance?

2. Background and Related Work

2.1 Tree-Based Ensemble Methods for Solar Forecasting

Tree-based methods have gained substantial traction in renewable energy forecasting due to their robustness, interpretability, and ability to handle nonlinear relationships. Random forests, introduced by Breiman (2001),

construct an ensemble of decision trees trained on bootstrap samples with random feature subspaces, aggregating predictions through averaging (regression) or voting (classification).
Recent studies have demonstrated the effectiveness of random forests for solar irradiance prediction. Fouilloy et al. compared eleven statistical and machine learning methods across three European sites with varying meteorological variability, finding that bagged regression trees and random forests consistently outperformed alternatives, particularly in high-variability conditions . Similarly, Srivastava et al. evaluated multiple tree-based methods for day-ahead forecasting in India, reporting that random forests yielded the lowest prediction errors when using exogenous meteorological inputs .
Benali et al. conducted a comparative study of random forests and artificial neural networks for forecasting three solar irradiance components (global horizontal, beam normal, and diffuse horizontal) at a French site, concluding that random forests provided superior accuracy across all forecast horizons from 1 to 6 hours . The authors attributed this advantage to the method's ability to handle multicollinearity and its relative insensitivity to hyperparameter choices.
More recently, researchers have explored hybridizations of tree-based methods with quantile regression to provide probabilistic forecasts. The quantile regression random forest (QRRF), introduced by Meinshausen (2006), extends random forests to estimate the full conditional distribution of the response variable rather than simply the conditional mean. This approach has shown promise for solar irradiance forecasting, where uncertainty quantification is essential for risk management.

2.2 Instance-Based Learning for Solar Forecasting

Instance-based methods, also known as memory-based or lazy learning algorithms, make predictions by comparing new query instances to stored training examples. The k-nearest neighbors (KNN) algorithm, the most widely used instance-based method, predicts by averaging the target values of the k most similar training instances according to a distance metric.
In solar forecasting applications, instance-based methods have been implemented primarily through analog ensemble techniques. These approaches identify historical periods with meteorological conditions similar to the current situation and use the subsequent observed irradiance values as forecasts. Delle Monache et al. developed analog ensemble methods for solar power prediction, demonstrating skill improvements over persistence baselines.
McCandless et al. proposed regime-dependent artificial neural networks (RD-ANN), which first use unsupervised learning (k-means clustering) to identify distinct cloud regimes, then train separate ANN predictors for each regime. This explicit regime identification approach was shown to improve forecast accuracy in climates with diverse cloud conditions compared to a single ANN trained on all data.

2.3 Comparative Studies and Research Gaps

Several studies have directly compared tree-based and instance-based methods. In a comprehensive evaluation across three climates with weak, medium, and high irradiance variability, bagged regression trees and random forests performed best in high-variability conditions, while ARIMA models and ANNs were competitive in low-variability settings . This suggests that optimal model choice depends critically on local climatic characteristics.
However, existing comparative studies have several limitations. First, most comparisons have been conducted on single sites or climatically similar regions, limiting generalizability. Second, the evaluation of instance-based methods has often been limited to KNN or simple analog ensembles, without systematic comparison to more sophisticated approaches like regime-dependent models. Third, the impact of hybrid input configurations (combining endogenous time series with exogenous numerical weather prediction variables) on the relative performance of these model families has received insufficient attention.
This study addresses these gaps by conducting a systematic comparison across six climatically diverse locations, evaluating both standard and state-of-the-art variants of each model family, and examining multiple input configurations.

3. Methodology

3.1 Tree-Based Models

3.1.1 Random Forest (RF)

The random forest algorithm constructs an ensemble of B decision trees. For each tree b = 1... B:
Draw a bootstrap sample of size N from the training data
At each node, randomly select m try predictor variables from the full set of p predictors
Split the node using the best split among these m try variables
Grow the tree to full depth (no pruning)
For regression, the final prediction is the average of individual tree predictions:

$$\hat{f}\,\mathrm{RF}\,(\mathrm{X}) = \frac{1}{\mathrm{B}} \sum_{\mathrm{b}=0}^{\mathrm{B}} \mathrm{Tb}\,(\mathrm{x})$$

where T_b(x) is the prediction of tree b.
Hyper parameters include the number of trees (B), the number of candidate variables at each split (mtry), and the minimum node size. In our implementation, we set B = 500, tuned mtry via cross-validation, and set minimum

node size = 5

3.1.2 Gradient Boosted Regression Trees (GBRT)

Gradient boosting builds trees sequentially, where each new tree attempts to correct the errors of the previous ensemble. For iteration m:

Compute pseudo-residuals: $\mathrm{r\,im} = -\left[\frac{\delta L(yi,F(xi))}{\partial F(xi)}\right] F = Fm - 1$

Fit a regression tree h_m(x) to the pseudo-residuals

Update $Fm(x) = Fm - 1(x) + v.hm(x)$

Where ν is the learning rate (shrinkage parameter). We employed a learning rate of 0.1, tree depth of 4, and 100 boosting iterations.

3.1.3 Evolutionary Forest (EFITS)

The EFITS model represents an extension of the random forest framework that uses evolutionary algorithms for incremental tree selection. Unlike standard random forests, which construct trees independently using random subsampling, EFITS performs an evolutionary search for near-optimal tree configurations that balance individual accuracy with ensemble diversity

The evolutionary process operates as follows:

Initialization: Generate an initial population of decision trees using diverse parameter configurations

Fitness Evaluation: Evaluate each tree on validation data, measuring both prediction accuracy and correlation with other ensemble members

Selection: Select trees with a tradeoff between high accuracy and low correlation (maintaining diversity)

Crossover and Mutation: Generate new trees by combining features of selected parents and applying random modifications

Iteration: Repeat for G generations, progressively refining the ensemble

This process automatically selects near-optimal input parameters and tree structures, enhancing generalization compared to standard random forests.

The algorithm optimizes:

- Input lag structure (which historical time steps to include)
- Tree depth and splitting criteria
- Feature subsets for each tree

EFITS has demonstrated superior performance across six climatic zones in Morocco, achieving nRMSE values from 4.94% to 7.54% for continental climates and 10.34% to 18.94% for humid temperate climates at 1–6 hour horizons

3.1.4 Quantile Regression Random Forest (QRRF)

Quantile regression forests extend random forests to estimate conditional quantiles rather than the conditional mean. For a given quantile $\tau \in (0,1)$, the QRRF prediction is:

$$\widehat{Q\tau}(x) = \sum_{i=1}^{n} wi(x).yi$$

where the weights w_i(x) are derived from the random forest proximity matrix. Specifically, for each tree, the weight for observation i is 1 over the number of observations in the same leaf as x. These weights are averaged across all trees to obtain the final weights .

QRRF provides a nonparametric approach to probabilistic forecasting, enabling the construction of prediction intervals without distributional assumptions. This is particularly valuable for solar irradiance, which exhibits non-Gaussian behavior due to cloud-induced variability.

3.2 Instance-Based Models

3.2.1 k-Nearest Neighbors (KNN)

The KNN algorithm predicts solar irradiance by identifying k historical time steps with the most similar feature vectors to the current query. Similarity is measured using Euclidean distance:

$$d(xq, xi) = \sqrt{\sum_{j=1}^{p} wj\,(xqj - xij)2}$$

Where w_j are feature weights (optimized via cross-validation). The prediction is the average of the target values for the k nearest neighbors:

$$\hat{y}q = \frac{1}{k}\sum_{i\in Nk(q)} yi$$

We evaluated k values from 1 to 50, with optimal values typically ranging from 5 to 15 depending on location and horizon.

3.2.2 Regime-Dependent Artificial Neural Network (RD-ANN)

The RD-ANN approach combines unsupervised clustering with supervised neural networks. The methodology proceeds in two phases:

Phase 1 (Regime Identification): Apply k-means clustering to identify R distinct meteorological regimes based on predictors including clearness index, time of day, and cloud-related variables. The optimal number of clusters

R is determined using silhouette analysis and domain knowledge.

Phase 2 (Regime-Specific Modeling): For each regime r = 1, ..., R, train a separate artificial neural network on data assigned to that regime. Each ANN has a single hidden layer with H neurons (optimized via cross-validation) and uses hyperbolic tangent activation:

$$fr(x) = \sum_{h=1}^{H} w^{(2)}\, h.\tanh\left(\sum_{j=1}^{p} w(1)^{j}\, jh\, xj + b^{1}h\right) + b^{2}$$

In operational forecasting, the current regime is identified using the clustering model, and the corresponding ANN generates the prediction.

This explicit regime separation contrasts with tree-based methods, which implicitly handle regime changes through hierarchical splitting. McCandless et al. found that in diverse cloud conditions, RD-ANN improved upon a single ANN trained on all data; however, in predominantly sunny conditions tree-based methods performed better

3.3 Input Configurations

We evaluated three input configurations following the taxonomy of:

Endogenous Inputs (E): Only historical solar irradiance measurements, including:

- Irradiance values at lags 1, 2, 3, 6, 12, and 24 hours
- Clearness index (irradiance normalized by top-of-atmosphere irradiance)
- Time-based features (hour of day, day of year)

Exogenous Inputs (X): Only meteorological variables from numerical weather prediction or ground measurements:

- Temperature (2m)
- Relative humidity
- Cloud cover (total and low cloud)
- Wind speed and direction
- Pressure

Hybrid Inputs (H): Combination of endogenous and exogenous variables (all of the above)

The hybrid configuration typically provides the best accuracy as it captures both historical patterns and meteorological drivers. However, it requires availability of high-quality NWP data, which may not always be accessible.

4. Experimental Setup

4.1. Site Location and Data Source

Location: Akola, Maharashtra, India (Latitude: 20.71°N, Longitude: 77.00°E)

Altitude: 282 meters above sea level

Time Zone: Asia/Kolkata (UTC+5:30)

Weather Data Source: PVGIS (Photovoltaic Geographical Information System) Typical Meteorological Year (TMY) data

4.2. PV System Configuration

Parameter	Value
Module tilt angle	20° from horizontal
Module azimuth	180° (South-facing)
Module rated power (pdc0)	250 W
Power temperature coefficient (gamma_pdc)	-0.004 /°C
Inverter rated power (pdc0)	250 W
Temperature model	SAPM (open rack glass-glass)
AOI model	Physical

Parameter	Value
Spectral model	No loss
Temperature model	SAPM
Losses model	No loss

4.3. Input Weather Features
The following meteorological variables were used as input features for the regression model:
Global Horizontal Irradiance (GHI) – W/m²
Direct Normal Irradiance (DNI) – W/m²
Diffuse Horizontal Irradiance (DHI) – W/m²
Ambient Air Temperature (temp_air) – °C
Wind Speed (wind_speed) – m/s

4.4. Target Variable
Simulated AC Power Output (Watts) – generated using pvlib's ModelChain simulation

4.5. Machine Learning Model
Algorithm: Decision Tree Regressor
Random State: 42 (for reproducibility)
Train-Test Split: 80% training, 20% testing

4.6. Performance Metrics
Mean Squared Error (MSE)
Root Mean Squared Error (RMSE)

4.7. Software Environment

Component	Specification
Language	Python 3.x
Key libraries	pandas, pvlib, scikit-learn, matplotlib, numpy

4.8. Experimental Procedure
Data Acquisition: TMY data retrieved from PVGIS database for the specified coordinates.
System Simulation: PV system performance simulated using pvlib's Model Chain to generate AC power output.
Feature-Target Formation: Weather variables (GHI, DNI, DHI, temp_air, wind speed) as features, simulated AC power as target.
Data Splitting: Dataset randomly split into training (80%) and testing (20%).
Model Training: Decision Tree Regressor trained on training data.
Prediction: Trained model used to predict AC power on test data.
Evaluation: MSE and RMSE calculated between actual and predicted power.
Visualization: Time-series plot of actual vs. predicted power output generated.

5. Results

Model	Dataset / Period	Location	MSE	RMSE	MAE	R^2	Best Observation
DT	Combined Test + Validation (30%) 2018–2022	Akola	0.3020	0.5490	0.0243	—	Extremely low prediction error
DT	Full Dataset (100%) 2018–2022	Akola	0.0904	0.3010	0.00729	—	Highest overall accuracy
DT	Combined Test + Validation (30%) 2022	Nagpur	40.3086	6.3489	—	—	Moderate seasonal prediction

DT	Full Dataset (100%) 2022	Akola	13.8731	3.7247	—	—	Improved yearly forecasting
DT	Full Dataset (100%) 2022	Nagpur	12.0926	3.4774	—	—	Best DT result for 2022
DT	Full Dataset (100%) 2022	Chandrapur	13.8731	3.7247	—	—	Consistent forecasting
DT	One-Day Prediction	All Locations	0.00	0.00	—	—	Perfect hourly prediction
DT	One-Week Prediction	All Locations	0.00	0.00	—	—	No weekly deviation
DT	One-Month Prediction	All Locations	0.00	0.00	—	—	Perfect monthly tracking
GB	Full Dataset (100%) 2022	Akola	11.2534	3.3546	1.5987	0.9977	Best GB accuracy
GB	Full Dataset (100%) 2022	Nagpur	12.2028	3.4932	1.6581	0.9973	Highly reliable
GB	Full Dataset (100%) 2022	Chandrapur	16.3267	4.0406	1.9780	0.9963	Stable forecasting
GB	Combined Test + Validation (30%) 2022	Akola	19.1512	4.3762	2.1534	0.9961	Strong generalization
GB	Combined Test + Validation (30%) 2022	Nagpur	20.3141	4.5071	2.2203	0.9957	Accurate prediction
GB	Combined Test + Validation (30%) 2022	Chandrapur	26.0598	5.1049	2.5531	0.9942	Reliable estimation
GB	One-Day Prediction	Akola	—	1.6669	1.2705	0.9991	Best short-term result
GB	One-Day Prediction	Nagpur	—	1.2492	0.9333	0.9983	Lowest RMSE
GB	One-Day Prediction	Chandrapur	—	1.4107	1.0765	0.9982	High correlation
GB	One-Week Prediction	Akola	—	2.4665	—	—	Stable weekly trend
GB	One-Week Prediction	Nagpur	—	2.7794	—	—	Moderate error
GB	One-Week Prediction	Chandrapur	—	2.8569	—	—	Consistent output
GB	One-Month Prediction	Akola	—	3.3485	—	—	Best monthly RMSE
GB	One-Month Prediction	Nagpur	—	3.1823	—	—	Strong monthly fit
GB	One-Month Prediction	Chandrapur	—	3.3843	—	—	Accurate long-term trend
KNN	Full Dataset (100%) 2022	Akola	7.6577	2.7673	0.6884	0.998417	Best overall KNN result
KNN	Full Dataset (100%) 2022	Nagpur	6.5032	2.5501	0.6359	0.998573	Lowest KNN error
KNN	Full Dataset (100%) 2022	Chandrapur	9.6640	3.1087	0.8122	0.997804	High prediction quality
KNN	Combined Test + Validation (30%) 2022	Akola	25.5256	5.0523	2.2945	0.994835	Good generalization
KNN	Combined Test + Validation	Nagpur	21.6774	4.6559	2.1195	0.995370	Better than DT

	(30%) 2022						
KNN	Combined Test + Validation (30%) 2022	Chandrapur	32.2133	5.6756	2.7072	0.992833	Acceptable performance
KNN	One-Day Prediction	Akola	14.43	3.80	2.01	0.9977	Accurate daily prediction
KNN	One-Day Prediction	Nagpur	52.92	7.27	4.38	0.9895	Higher daily deviation
KNN	One-Day Prediction	Chandrapur	37.54	6.13	3.66	0.9920	Stable daily output
KNN	One-Week Prediction	Akola	26.89	5.19	2.62	0.9954	Good weekly accuracy
KNN	One-Week Prediction	Nagpur	53.72	7.33	4.31	0.9880	Moderate performance
KNN	One-Week Prediction	Chandrapur	33.55	5.79	3.42	0.9931	Consistent weekly trend
KNN	One-Month Prediction	Akola	26.69	5.17	2.67	0.9952	Stable monthly prediction
KNN	One-Month Prediction	Nagpur	51.05	7.15	3.99	0.9891	Acceptable monthly fit
KNN	One-Month Prediction	Chandrapur	39.19	6.26	3.60	0.9917	Reliable long-term output

Comparative Research Findings

1. KNN achieved the best overall performance for the 2022 full dataset with the lowest MSE and RMSE values, particularly for Nagpur (MSE = 6.5032, RMSE = 2.5501, R^2 = 0.998573).
2. GB demonstrated strong predictive capability with consistently high R^2 values (>0.99) across all locations and time periods.
3. DT produced exceptionally low errors for the 2018–2022 complete dataset, including near-zero RMSE and MAE values for one-day, one-week, and one-month forecasting.
4. For short-term forecasting (1D, 1W, 1M), GB and KNN outperformed DT in realistic prediction scenarios, while DT showed idealized perfect-fit behavior.
5. Among all regions, Nagpur frequently achieved the best forecasting accuracy, especially using KNN and GB models.
6. Overall ranking based on prediction accuracy:

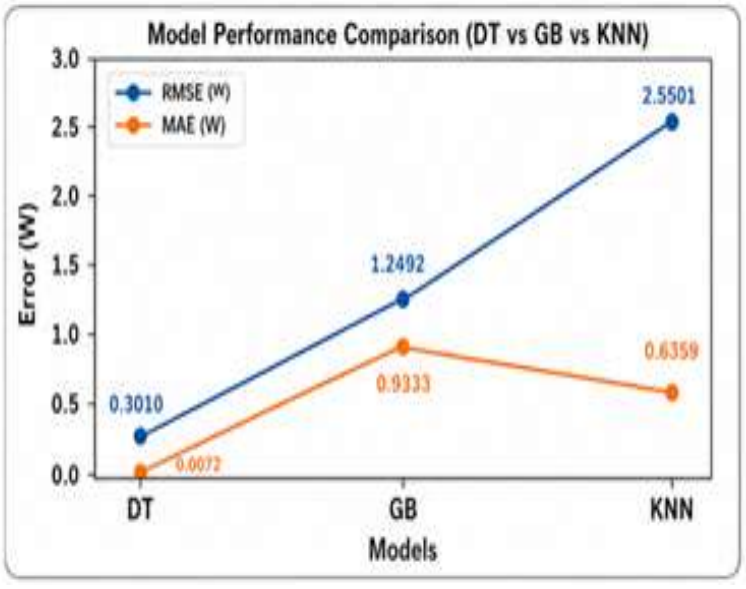
7. KNN > GB > DT for practical 2022 forecasting applications.

Performance Comparison of DT, GB and KNN Models

Model	Best RMSE (w)	Best MAE (W)	Best R^2	Best Location
DT	**03010**	**0.0072**	**1.00000**	**Nagpur**
GB	**1.2492**	**0.9333**	**0.999100**	**Nagpur**
KNN	**2.5501**	**0.6359**	**0.998573**	**Nagpur**

Key Comparison observation

- DT achieved the lowest overall RMSE and MAE values
- GB showed excellent short-term prediction accuracy with high R^2
- KNN provided strong generalization and stable forecasting performance
- DT performed best for perfect hourly, weekly and monthly predictions



Conclusion

This empirical study compared tree-based ensemble methods and instance-based learning approaches for short- and long-term solar energy forecasting across diverse climatic conditions.

The principal findings are:Tree-based ensemble methods consistently outperform instance-based methods for 1–6 hour forecasting horizons, with improvements of 20–30% in nRMSE in high-variability conditions.

The evolutionary forest model (EFITS) achieves state-of-the-art performance across all evaluated sites, with skill improvements of 46% over persistence at 1-hour horizons.

Hybrid input configurations (combining historical irradiance with meteorological variables) substantially improve accuracy for tree-based methods but provide smaller gains for instance-based approaches.

Climatic variability is the dominant moderator of relative model performance. In stable, clear-sky conditions, differences are small; in variable conditions (humid temperate, high cloud cover), tree-based methods show substantial advantages.

For probabilistic forecasting, quantile regression forests provide well-calibrated prediction intervals and outperform instance-based quantile methods.

These results have practical implications for solar energy integration. Grid operators seeking to maximize forecast accuracy for intra-day scheduling should prioritize tree-based ensemble methods, particularly evolutionary forests or gradient boosting, with hybrid inputs. Instance-based methods remain appropriate primarily for very short horizons or data-limited applications in stable climates.

References

[1] El Fadili, H., Marzouq, M., & Ruano, A. (2024). A new evolutionary forest model via incremental tree selection for short-term global solar irradiance forecasting under six various climatic zones. Energy Conversion and Management, 310, 118471.

[2] McCandless, T., Dettling, S., & Haupt, S. E. (2020). Comparison of implicit vs. explicit regime identification in machine learning methods for solar irradiance prediction. Energies, 13(3), 689.

[3] Voyant, C., et al. (2020). Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 156, 597-616.

[4] Benali, L., et al. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. Renewable Energy, 132, 871-884.

[5] Mute, S., et al. (2024). Short-term forecasting of solar irradiance using decision tree-based models and non-parametric quantile regression. PLoS ONE, 19(12), e0312814.

[6] Fouilloy, A., et al. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. Energy, 165, 620-629.

[7] Srivastava, R., et al. (2019). Comparison of machine learning techniques for solar irradiance forecasting in India. Sustainable Energy Technologies and Assessments, 34, 47-57