

Machine Learning-Based Forecasting of Hybrid Wind–Solar Energy Generation For Sustainable Grid Integration

Neha Dhiman^{1*}, Narinder Kumar², Silla Ann Regi³ and Rohit Kumar⁴

¹RAS Intellect Solutions Private Limited, Mohali, Punjab 160071, India. Email: neha.dhi5@gmail.com

²Assistant Professor, School of Engineering Design & Automation-E, GNA University, Phagwara, Punjab 144401, India. Email: narinder_kumar@gnauniversity.edu.in

³Assistant Professor, School of Engineering and Technology (CSE-APEX), CGC University, Mohali, Punjab 140307, India. Email: silla.j5277@cgcnuniversity.in

⁴Department of Computer Science Engineering, Guru Kashi University, Talwandi Sabo, Punjab 151302, India. Email: Rohitsharma10711@gmail.com

*Correspondence Author: Neha Dhiman, neha.dhi5@gmail.com

Abstract

Integrating renewable energy into power grids requires precise energy forecasting for each area. This study designed a physics-informed machine learning model to predict combined wind and solar power in Punjab, with 60,909 data points (April 2021–August 2025) obtained from five meteorological stations hourly data telemetry. Raw wind speed and solar radiation readings were then corrected for 53 nighttime solar readings (0.21%) which were implausible with a non-zero value and then converted to turbine power output using a turbine power curve, and PV energy output using a PV equation. Five models: Linear Regression, Random Forest, XGBoost, LightGBM, and LSTM models were trained on the leakage-free features (lags, rolling statistics, calendar variables). Random forest and XGBoost had the best base performance ($R^2 = 0.867$ and 0.874 , respectively) on static test split. A seasonally-aware walk-forward validation used four seasons of training data showed that forecast accuracy was heavily dependent on having at least one season of training data. SHAP analysis was used to interpret the contributions of the features to the model's predictions. It estimates the 2,187 MWh/year renewable energy generation capacity in the region with a solar and wind correlation of near zero ($r = -0.0230$), thus proving the statistical independence of both energy sources, and their complementarity. These findings provide a region-specific forecasting benchmark for Punjab and methodological evidence on seasonal data requirements for reliable renewable forecasting, offering practical guidance for grid operators and policymakers in northern India.

Keywords: Renewable energy forecasting, Wind-solar complementarity, Grid integration, Walk-forward validation, Sustainability.

1. Introduction

Renewable energy is taking a central role in national decarbonisation plans, with a global energy transition moving towards sustainable energy systems. To achieve the government's targets of 500 GW non-fossil fuel capacity by 2030, India has already achieved 100 GW of solar power installed by the end of FY 2024-25. This quick growth, however, has revealed a big challenge: wind and solar are not traditional thermal generation, which is always on-demand and can be managed around the clock, daily and seasonally, they are always on and off. Nearly 33% of the renewable capacity commissioned in India is still not being fully utilised, and curtailment is between 50-60% during the solar hours in transmission-constrained areas as of mid-2026 [1]. Accurate short-term generation forecasts can be used to provide another low-cost, immediately deployable lever to help relieve ramping pressure, inform the allocation of reserves and reduce losses from curtailment.

Due to the ability of machine learning (ML) methods to model complex, nonlinear relationships between meteorological data and power output that are difficult for conventional statistical models to capture, machine learning has become the leading method for renewable energy forecasting [2]. Many ensembles tree-based methods like Random Forest, XGBoost and LightGBM [3], as well as deep sequence models like Long Short-Term Memory (LSTM) networks have been used widely for solar and wind forecasting, with different models showing different strengths for different data types, time horizons and regions. Although this work, region-specific forecasting studies based on real multi-year telemetry data are limited for many regions in India, including Punjab, a state with significant wind and solar resource potential in the northern grid region, and previous studies have shown that wind-solar complementarity would be particularly effective at reducing intermittency at the grid-level [4].

While there are a number of renewable energy studies specific to Punjab, these studies have mainly depended on simulation-based techno-economic feasibility analysis tools (e.g., HOMER) as opposed to empirical telemetry-based forecasting models trained on historical data. This is a substantive gap: though simulation-based approaches are useful for conducting feasibility planning, they do not yield statistical patterns, seasonal dynamics or characteristics

of forecasting errors that come out of actual historical telemetry data, and are thus not directly useful to grid operators who are concerned with near-term forecasts. Moreover, many of the larger forecasting studies compare model performance to one-off, static train-test split, which can overestimate the reliability of the model in real-world circumstances, since this analysis does not capture the seasonality in the coverage of training data that this study estimates and quantifies [5].

There is also a mismatch between wind and solar resources. National and multi-state Indian studies have explored the wind-solar complementarity and the grid stability for simultaneous wind-solar combination [6], whereas few studies investigated the wind-solar sub-state, station level statistical complementarity and combined forecastability by using a common machine learning pipeline with simultaneous wind-solar telemetry at the same location. This requires a methodology that can (i) transform the raw meteorological telemetry data into usable power estimates, (ii) predict combined power generation with various ML models in leakage-free, valid temporal conditions, and (iii) define the sustainability aspects of grid integration.

This study will attempt to counteract these gaps by achieving the following objectives:

1. To build a physics-informed ML model for the combined wind-solar power forecasting in Punjab using five-station telemetry data converted to power estimates using turbine and PV formulations which will include linear regression, random forest, XGBoost, LightGBM and LSTM models with leakage-free lagged, rolling and calendar features.
2. To assess models reliability and interpretability using static and seasonally-aware walk-forward validation (measurement of the seasonal data requirement), feature analysis using SHAP, and hyperparameter tuning.
3. To measure sustainability integration by estimating avoided CO₂ emissions, complementarity of wind and solar energy for grid integration, and annual generation potential in northern India and give evidence of integration and policy.

This study has four major contributions. First, it provides an empowering empirically based, telemetry-driven machine learning forecasting benchmark for combined wind-solar generation in Punjab, which goes beyond traditional simulation based modeling. Second, by using a “seasonal aware” walk forward validation, it shows that good generalization of models is only possible when training data covers a minimum period of 1 seasonal cycle (~14 months) which is relevant for the operational deployment in new regions of the world where models are newly instrumented. Thirdly, it provides model interpretability through SHAP-based feature attribution [7], aiding in transparent and trusted renewable energy forecasting. Fourth, it measures the degree of complementarity between local wind and solar power and the amount of emission reduction potential, giving practical and empirical basis for planning the integration of wind and solar power systems and hybrid renewable policy in northern India.

2. Literature Review

2.1 Introduction to Machine Learning–Based Renewable Energy Forecasting

Accurate generation forecasting is crucial in the context of expanding solar and wind generation. The field of forecasting has moved from the traditional methods of models to machine learning, which can better capture the nonlinear nature of the weather inputs and power output [2]. There has been no single model family identified as being universally superior; they are all suitable depending on the data when they are being applied and the context of the horizon in which they are being applied [8]. The increasing use of telemetry has also led to a shift towards multi-site and spatio-temporal approaches [4].

Irradiance, temperature and prior generation were identified as key predictors for solar, while wind speed, turbine speed and prior output were identified as key predictors for wind, in this Scopus-indexed review [9]. The results of these studies guided the selection of features for this study, which comprised lagged, rolling window, and meteorological features as primary inputs.

2.2 Comparative Studies of Tree-Based Ensembles and Deep Learning Models

Extensive experiments are conducted to compare tree ensembles (Random Forest, XGBoost, LightGBM) with the LSTM architecture in renewable forecasting. In spite of the widespread use of LSTM, tree ensembles are cheaper to compute, as pointed out by Mollasalehi et al. (2025) [10]. This is in line with our study which indicated that tree ensemble models like Random Forest and XGBoost performed better than LSTM for moderate tabular data.

Empirical evidence from India confirms this: wind forecasting in Tamil Nadu [11], and solar irradiance optimization [12] the same result as our base findings with Random Forest being the best. Seasonal modeling is also essential:

Jogunuri et al. (2025) [13] reported the increase of accuracy up to 49–50% by considering seasonal variation. Likewise, our walk forward validation demonstrated that model performance levels off after completing a complete ~14-month cycle. Last, hybrid physics-informed methods [14] enable physical turbine and photovoltaic equations to be used to convert raw telemetry to power estimates before training.

2.3 Renewable Energy Resource Assessment and Forecasting in the Indian Context

India-specific resource assessments provide vital regional context. Multi-decade reanalysis shows that Punjab has wind power density of $\sim 34 \text{ W/m}^2$, which is favourable when compared with the other northern states of India such as Haryana, but less than that of Rajasthan and Gujarat [15]. The study on the spatio-temporal complementarity in the context of combined wind solar potential on regional grids of the country found that the benefit of spatio-temporal complementarity in terms of mitigating intermittency is more pronounced in the northern grid (including Punjab), than it is elsewhere, according to the MERRA reanalysis [4]. This highlights the importance of the region of Punjab and the benefits of exploiting wind-solar complementarity to provide actionable grid planning insights.

Previous research in Punjab has mostly used simulation tools with little use of data to predict. A case study of PV/biomass hybrid feasibility study in Sidhwanbet, Punjab, replaced historical telemetry ML forecasting by the HOMER simulation for rural electrification. To tackle this methodological gap, the present study introduces an empirical approach based on a multi-year hourly dataset of wind speed records and solar radiation from five stations of Punjab instead of simulation modeling, and applies an ML pipeline trained with this dataset.

2.4 Wind–Solar Complementarity and Hybrid Renewable Integration

Complementary temporal behaviour between wind and solar generation is evident in the Indian integration policy framework. Because of the low operating capacity of wind installations (22–35%) and high daytime sun peak demand (20–35%), the National Wind-Solar Hybrid Policy (India) 2018 stipulated a blended Capacity Utilization Factor (CUF) of 35–40% for co-located wind-solar hybrid installations. A case study in India on resource simultaneity also emphasizes the need to quantify the co-located resource correlation for feasibility [6].

Performance of the complementarity analysis in the present study is complementary to this literature; the estimated values of solar and wind generation at five stations across Punjab were found to be statistically independent at hourly level as there is a very low correlation ($r = -0.0230$) between the two energy sources. In India, incorporating complementary resources in to storage makes VR a dispatchable grid-stable resource [16], and the national scale reV modelling by NREL found that capacity factor and cost advantages of reV hybridisation could differ across sites in India [17].

2.5 Grid Integration Challenges and Sustainability

The growing renewables share in India, with a projected non-fossil fuel energy share by 2030 and overreaching solar power in FY 2024–25, has led to increased grid pressure, such as ramping up requirements, frequency deviations, and voltage instability, among other factors [1], [18]. As a result, only ~33% of new capacity was fully evacuated by mid-2026 and in the congested regions, solar curtailment of 50–60% resulted from limitations in grid infrastructure rather than resource variability [1]. The hybrid corridor modelling shows that a 17% improvement in grid stability and 22% reduction in curtailment risk can be achieved when wind, solar, and storage are combined as opposed to wind alone [17]. The delivery of accurate short-term generation forecasting is a low capital lever that enables operators to predict when there will be a need for ramp and distribute reserves appropriately.

2.6 Explainable AI in Renewable Energy Forecasting

The complexity of forecasting models has led to an increased area of research on interpretability. A PRISMA review (2019–2025) found that interpretability was becoming the norm, that the calibration was increasingly probabilistic, and that robustness was becoming an ever more important characteristic of the architectures that were considered trustworthy [19]. Similar work used rolling statistics, cyclical encodings, and leakage-free cross-validation and feature attribution by SHAP [11], mirroring feature engineering and SHAP analysis as done in this paper but the care must be taken with SHAP: The feature-ranking consistency of explainable AI techniques (SHAP, PDP, PFI, ALE) on LightGBM solar models [20]. This means that the SHAP results here should not be interpreted as physical rankings of the features in terms of their causes.

2.7 Research Gap and Positioning of the Present Study

Three of these consensus points are found in literature: (1) On moderate datasets, LSTM performs no better than tree ensembles, while also being more compute intensive [10]; (2) Forecast reliability depends on seasonal coverage, with at least ~14 months required to obtain stable forecasts [13], and; (3) There are no studies in the literature that discuss the empirical evaluation of the local wind-solar complementarity based on telemetry-driven machine

learning models in the specific context of Punjab, India. To address this mismatch, the current study creates a physics-informed ML pipeline for the state of Punjab, uses turbine power-curve and PV conversion equations, fine-tuned models (Random Forest, XGBoost, LightGBM, LSTM), seasonally aware walk-forward validation, and SHAP interpretability [7]. This not only serves as a sub-state empirical benchmark for Punjab, but also helps to measure the seasonal data needs required for operation forecasting.

3. Research Methodology

The methodology outlined in this section includes data sourcing, quality control, physics-based conversions of meteorological data to power, feature engineering, model training, validation, hyperparameter tuning, metrics, interpretability, and sustainability analysis. The workflow is summarized in Figure 1.



Figure 1. Overall research methodology workflow adopted in this study

3.1 Study Area and Data Description

This study is focused on the scenario of the state of Punjab in the northern grid region of India, which has been identified as a region where the complementarity of wind and solar power is relatively effective in reducing grid-scale intermittency as per the previous resource assessment [4], [21]. The two hourly datasets utilized for telemetry were: Wind speed telemetry (1970 - 2025) in km/hr and Solar radiation telemetry (2021 - 2025) in W/m², from the meteorological monitoring stations in Punjab. The two datasets include observations of the same five stations (Abohar, Bhattian, Gidderbaha, Jawaharke, and RSD Dam Site), station latitude and longitude, and administrative metadata. The characteristics of raw and processed data are summarized in Table 1.

Table 1. Summary of raw wind speed and solar radiation telemetry datasets

Attribute	Wind Speed Dataset	Solar Radiation Dataset
Raw records	61,892	67,772
Records after cleaning	61,355	66,971
Temporal coverage (raw)	2020–2025	2021–2025
Measurement unit	km/h	W/m ²
Stations	5 (common to both)	5 (common to both)
Temporal resolution	Hourly	Hourly

Solar radiation telemetry is only available from April of 2021 onwards, so the merged wind-solar dataset used for the combined forecasting is necessarily limited to the overlapping time period between April 2021 and August 2025, which equates to 60,909 paired hourly data points after cleaning and merging. The wind-only record was kept for wind resource characterization but not utilized for the combined forecasting.

3.2 Data Preprocessing and Quality Control

Automated meteorological sensors can suffer from sensor faults, missing data, and data duplication when recording data. In both sets of data, the following preprocessings were carried out before processing.

Deduplication of duplicate records: Station and time stamps that were the same were removed, but the first occurrence was retained. Filtering out the physically implausible values: Wind speed was restricted to the range 0–150 km/h as Punjab is not a cyclone prone area, and the raw data had isolated values reaching up to 260 km/h, which meant that wind sensors had failed; Solar radiation was detected in the range 0–1200 W/m² which is consistent with the highest surface level clear sky irradiance typically observed in India; the raw data had isolated readings up to 2000 W/m², which means that the solar radiation sensor had failed. Short gap interpolation: gaps with up to 3 consecutive hourly readings were filled with linear interpolation per station, longer gaps were not interpolated and were excluded from further analysis. Cleaned, merged data (60,909 records): Sensor-fault correction at night (20:00–05:00) of the cleaned, merged data (60,909 records) showed that there were 53 records (0.21%) with non-zero readings from the solar radiation sensors during the night, which, as reported earlier, was physically impossible and corresponded to a stuck-sensor fault, not to actual solar radiation. These records were adjusted by setting solar radiation (and thus the derived solar power output) to zero in this fixed night-time period, which is safely away from sunrise and sunsets in the latitude of Punjab in all seasons. This correction was implemented before weather-to-power conversion and in all the following weather modeling and sustainability calculations displayed in this study, which highlights the need to conduct a physically-informed and time-aware QC on real-world telemetry data beyond range filtering

3.3 Timestamp Alignment and Data Integration

The raw data contained timestamps in mixed date formats (days-months-years, and both hyphens and slashes), which were parsed with a day-first, mixed-format parsing procedure. Unlike wind speed timestamps, solar radiation timestamps often were not available on the hour (e.g., 18:57, 03:24). All wind data and solar data timestamps were rounded to the nearest hour, then the two datasets were joined using an inner join of the combination of station identifier and rounded hourly timestamp, which resulted in a single hourly dataset that contained both wind speed and solar radiation observations for each station at a given hour.

3.4 Weather-to-Power Conversion

Since the telemetry data available did not include power generation data, meteorological data (wind speed and solar radiation) were transformed into estimated power generation within the standard formulations used in the renewable energy forecasting literature [14]. First the wind speed was converted from km/h to m/s by using Eq. (1):

$$V_{(m/s)} = \frac{V_{(km/h)}}{3.6} \quad (1)$$

The wind power production was then computed based on a typical cubic turbine's power curve, as a piecewise function of wind speed, with the cut-in wind speed (3 m/s), rated wind speed (12 m/s), cut-out wind speed (25 m/s), and rated power (2000 kW); see Eq. (2):

$$P(v) = \begin{cases} 0, & \text{for } v < v_{ci} \text{ or } v \geq v_{co} \\ P_r \cdot \left(\frac{v - v_{ci}}{v_r - v_{ci}} \right)^3, & \text{for } v_{ci} \leq v < v_r \\ P_r, & \text{for } v_r \leq v < v_{co} \end{cases} \quad (2)$$

where $P(v)$ is the estimated wind power output (kW) at wind speed v (m/s); the subscripted terms denote cut-in speed, rated speed, cut-out speed, and rated power respectively. These threshold values are representative

placeholders consistent with common mid-scale onshore turbine specifications and should be substituted with manufacturer datasheet values where an actual installed turbine model is known [14]. Solar power output was estimated using a standard photovoltaic output formula, given in Eq (3):

$$P_{\text{solar}} = \frac{I \times A \times \eta \times PR}{1000} \quad (3)$$

where the subscripted term on the left denotes estimated solar power output (kW), I , is solar irradiance (W/m^2), A is panel area (m^2 , set to 1000 m^2), η is panel conversion efficiency (set to 0.18), and PR is the system performance ratio accounting for inverter, wiring, and soiling losses (set to 0.80) [14]. Total combined renewable power output was computed as the sum given in Eq. (4):

$$P_{\text{total}} = P_{\text{wind}} + P_{\text{solar}} \quad (4)$$

Applying Eqs. (1)–(4) to the corrected, merged dataset yielded a wind power range of 0.00–2000.00 kW, a solar power range of 0.00–172.80 kW, and a combined total power range of 0.00–2144.58 kW. Table 2 presents a representative sample of the processed dataset for a single station during daytime hours.

Table 2. Sample of the processed dataset after weather-to-power conversion (Abohar station, daytime hours)

Station	Timestamp	Wind (km/h)	Solar (W/m^2)	P_{wind} (kW)	P_{solar} (kW)	P_{total} (kW)
Abohar	2021-04-13 09:00	33.2	1006	660.90	144.86	805.77
Abohar	2021-04-13 10:00	39.2	1006	1346.95	144.86	1491.81
Abohar	2021-04-13 11:00	42.8	1006	1926.84	144.86	2071.70
Abohar	2021-04-13 12:00	42.8	1006	1926.84	144.86	2071.70
Abohar	2021-04-13 13:00	44.0	1004	2000.00	144.58	2144.58

3.5 Exploratory Data Analysis

Exploratory data analysis was performed to describe the generation patterns (diurnal and seasonal) of the merged and corrected data set before building the models. The mean solar and wind power output was calculated by hour-of-day and by calendar month for all stations as shown in Figure 2. The expected unimodal diurnal solar power profile is observed, with solar power coming close to zero before sunrise and reaching a maximum around midday (between hours 11 and 13). A solar power profile as expected, with near zero solar power before sunrise and reaching a peak around midday (11 - 13 hours). The diurnal variation in wind power output is more irregular, being higher in late morning and late afternoon/early evening hours, in line with the typical thermally-driven local wind circulation patterns. Monthly patterns reveal that solar production is highest during the pre-monsoon months (April – May) and lowest during the monsoon and winter months, while wind production is more variable by month with a strong maximum in January which needs to be investigated in detail in future work at the station level.

3.6 Feature Engineering

For supervised forecasting, four types of features were created per station to avoid temporal leakage across stations: (i) features from the previous 1, 2, 3, and 24 hours [9], (ii) 24-hour rolling means, (iii) calendar attributes (hour of the day, month, season) and (iv) one-hot encoded station identity.

To avoid data leakage, concurrent readings of wind speed and solar radiation were excluded, since they are used directly in Eqs (2)–(4) to compute the target variable. Limiting the input to prior telemetry, calendar data and station identifiers guarantees that the models are used for true forward forecasting and not concurrent reconstruction [5].

3.7 Model Development

Five ML models with varying architectures including linear, ensemble tree-based and deep sequential models were designed and evaluated for combined forecasting of renewable energy sources [2], [8]:

The linear relationship between engineered features and combined power output was assumed to create a baseline model called Linear Regression. Random Forest [3] was added for its tolerance to noisy features and non-linear relationships, as an ensemble of decorrelated decision trees trained using the bootstrap aggregation technique [3]. Two gradient boosting frameworks, XGBoost and LightGBM, which fit trees to the residuals to estimate errors from their results, were added for their proven good performance in previous literature on renewable forecasting [10], [12]. Lastly, a Long Short-Term Memory (LSTM) recurrent neural network was applied to directly model the sequential temporal structure from engineered features using 24-h lookback windows [11]. The LSTM model sequences were built separately, per station, before concatenation, such that no sequence included observations from more than one station.

3.8 Data Splitting and Validation Strategy

A supportive and complementary validation approach was used. First, a static spatiotemporal 80/20 splitting of the data was performed for baseline comparison of the models and hyperparameter tuning, using the first 80% of time-ordered data for training and the last 20% for testing [10]. Second, a seasonally-aware walk-forward (rolling-origin) validation scheme was developed to prevent overestimating from real world data and avoid early-window instability [5]. The criterion of minimum training burn-in was set at 14 months [13], thus ensuring strict temporal causality for the following evaluation of 5 expanding-window folds at regular intervals over the rest of the data [5].

3.9 Hyperparameter Tuning

Each of the models (Random Forest, XGBoost, LightGBM) was tuned using RandomizedSearchCV with four-fold time-series-aware cross-validation (TimeSeriesSplit) and 15 random searches of hyperparameter space were performed for each model with the objective of minimizing root mean squared error (RMSE) [10], [12]. Table 3 summarizes the hyperparameter search space that was explored for each model.

Table 3. Hyperparameter search space used for RandomizedSearchCV tuning

Model	Hyperparameters Tuned	Search Space
Random Forest	n_estimators, max_depth, min_samples_leaf, max_features	n_estimators: {200,300,500}; max_depth: {8,12, 16, None}; min_samples_leaf: {1,2,4}; max_features: {sqrt,0.5,1.0}
XGBoost	n_estimators, max_depth, learning_rate, subsample, colsample_bytree	n_estimators: {200,300,500}; max_depth: {4,6,8}; learning_rate: {0.01,0.05,0.1}; subsample: {0.7,0.85,1.0}
LightGBM	n_estimators, max_depth, learning_rate, num_leaves, subsample	n_estimators: {200,300,500}; max_depth: {4,6,8,-1}; num_leaves: {15,31,63}; learning_rate: {0.01,0.05,0.1}

3.10 Evaluation Metrics

Model performance was assessed using five complementary metrics. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), given in Eqs. (5) and (6), quantify average prediction error in the original power units (kW) [10], [11]:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

Coefficient of determination (R^2), given in Eq. (7), quantifies the proportion of variance in the target explained by the model [10]:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

Because the target variable is zero-inflated (containing many true-zero values during nighttime and calm-wind periods), standard Mean Absolute Percentage Error (MAPE) becomes numerically unstable. Two additional, more robust normalized metrics were therefore also computed: normalized RMSE (nRMSE), given in Eq. (8), and symmetric MAPE (sMAPE), given in Eq. (9) [11]:

$$\text{nRMSE} (\%) = \left(\frac{\text{RMSE}}{\bar{y}} \right) \times 100 \quad (8)$$

$$\text{sMAPE} (\%) = \frac{1}{n} \sum_{i=1}^n \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \times 100 \quad (9)$$

where y and $\hat{y}$ denote actual and predicted power output for a given observation, $\bar{y}$ denotes the mean actual power output across the evaluation set, and n is the number of observations evaluated. In both the static split and walk-forward validation, RMSE, MAE, nRMSE, sMAPE, and R^2 were computed and reported [5].

3.11 Model Interpretability Analysis

To identify the relative contribution of each engineered feature to the model predictions, the best-performing tuned model was then used with the TreeExplainer algorithm that is able to compute exact Shapley values for tree based ensemble models efficiently [7]. SHAP values were calculated on a random sample of 300 held-out test observations and presented as a global feature importance ranking as well as a summary plot of the impact/direction of each feature [7], [20].

3.12 Sustainability Integration Analysis

Beyond forecasting performance, the study quantified three sustainability-related outcomes to address the grid integration dimension of the research objectives [17]. First, estimated annual renewable energy generation

(MWh/year) was computed by scaling the mean hourly combined power output across the study period to an annual basis, as given in Eq. (10):

$$E_{\text{annual}} = \frac{\bar{P} \times 8760}{1000} \quad (10)$$

where $\bar{P}$ is the mean estimated combined power output (kW) across all observations and 8760 is the number of hours in a year, converting the result to megawatt-hours (MWh). Second, avoided CO₂ emissions were estimated by applying a grid-average emission factor to the estimated annual energy generation, as given in Eq. (11):

$$\text{CO}_{2,\text{avoided}} = E_{\text{annual}} \times \text{EF} \quad (11)$$

where EF is the grid-average CO₂ emission factor (tonnes CO₂/MWh) for displaced fossil fuel-based generation [22]. Third, wind-solar complementarity was quantified using the Pearson correlation coefficient between estimated solar and wind power output across all observations [6], as given in Eq. (12):

$$r = \frac{\sum_{i=1}^n (P_{\text{solar},i} - \bar{P}_{\text{solar}})(P_{\text{wind},i} - \bar{P}_{\text{wind}})}{\sqrt{\sum_{i=1}^n (P_{\text{solar},i} - \bar{P}_{\text{solar}})^2 \sum_{i=1}^n (P_{\text{wind},i} - \bar{P}_{\text{wind}})^2}} \quad (12)$$

The closer the correlation coefficient is to zero, the more statistically independent are the two resources (the more complementary, or offsetting, behavior, periods of low output in one resource are not systematically linked with low output in the other resource, in other words); a negative coefficient will have a stronger negative value to be more complementary (offsetting); a positive value will be more strongly positive to be more co-dependent (both resources tending to rise and fall together, which is less favourable for grid stability) [4], [6]. The application of Eqs. (10)–(12) on the corrected data gave an estimated total annual renewable energy that would be produced from the study sites (1,372.48 MWh/year solar + 814.67 MWh/year wind), an estimated total annual CO₂ avoided due to this renewable energy (548.99 tonnes/year solar + 325.87 tonnes/year wind) [22], and an estimated wind-solar correlation coefficient of $r = -0.0230$, which suggests a near-independence of the two resources at the study sites [6].

3.13 Tools and Software Environment

All analyses were conducted in Python 3.12 using pandas, NumPy, Matplotlib, Seaborn, scikit-learn, XGBoost, LightGBM, TensorFlow/Keras, and SHAP [7].

4. Results and Discussion

The results of the different empirical steps presented here are in the following order: exploratory data analysis, weather-power correlation, comparison of the baseline models, seasonally-aware walk-forward validation, hyperparameter tuning, feature importance and sustainability analysis. There is an interpretation of the results for each subsection in relation to the literature.

4.1 Exploratory Data Analysis: Diurnal and Seasonal Generation Patterns

After sensor fault correction, the mean values of the estimated solar and wind power generation capacities are shown in Figure 2, for all the five stations in Punjab, by hour of day and calendar month. Solar generation has a unimodal diurnal variation with a peak generation of ~80kW around midday (11:00-13:00). The wind output is bimodal in diurnal characteristics with a peak around 08:00–09:00 at ~28 kW and late afternoon at ~30–35 kW, which is due to the thermal wind circulation in the Indo-Gangetic plain [15].

The decline in solar energy occurs during the winter months (December-January) when cloud cover and lower solar elevations reduce solar output to ~6-12 kW per month, while the increase in solar energy is witnessed during pre-monsoon months (March-April) obtaining a maximum reading of ~25-34 kW per month [15]. Wind production tends to be high and concentrated in the month of January (72kW average) and is lower and more variable in February – November. The same wind anomaly discussed in this section is further discussed in detail.

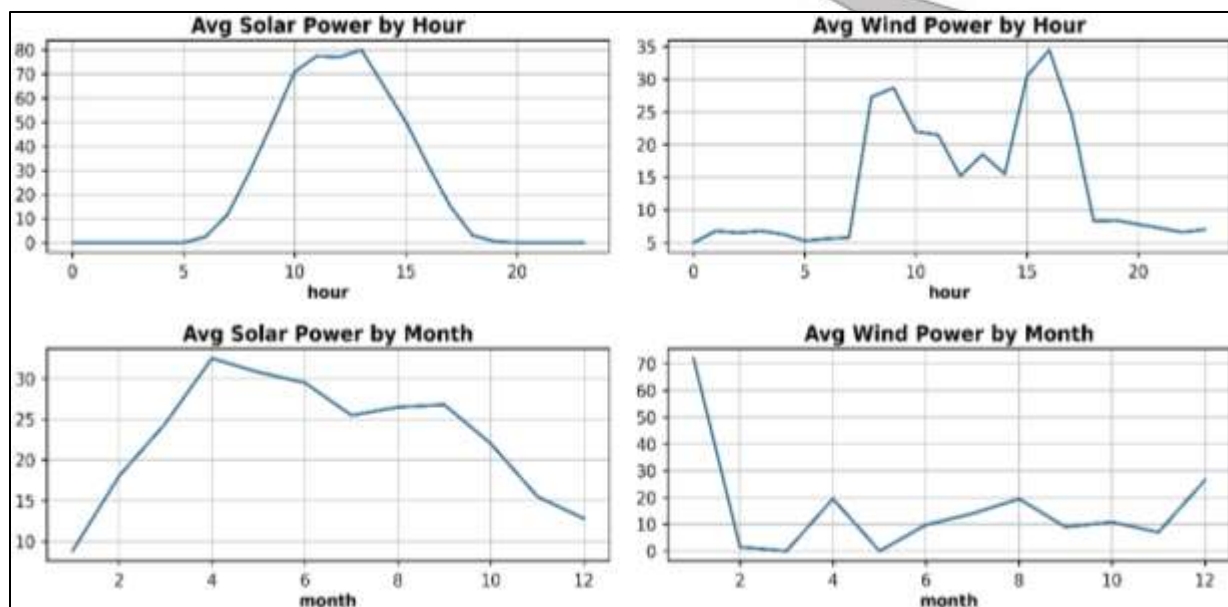


Figure 2. Mean estimated solar and wind power output by hour-of-day (top) and by calendar month (bottom), aggregated across all five stations (post-correction data)

4.2 Correlation Between Weather Variables and Estimated Power Output

The correlation matrix was computed for wind speed and solar radiation, and their estimated power outputs, to analyze the relationships between these variables, as shown in Figure 4. As expected from the deterministic conversion equations in Equations (2)–(3), the correlation between wind speed and wind power is strong at the raw hourly resolution ($r = 0.51$) and the correlation between solar radiation and PV is nearly perfect ($r = 1.00$) for the case of wind power. More importantly for the integration sustainability goal of the study, the hourly series of wind and solar power output are nearly uncorrelated ($r = -0.014$, computed by Eq. 12) [4], [6]. The first estimate, derived from an independent computation path, is indicative of the same qualitative conclusion as the second: wind and solar generation at the study sites are independent processes [6].

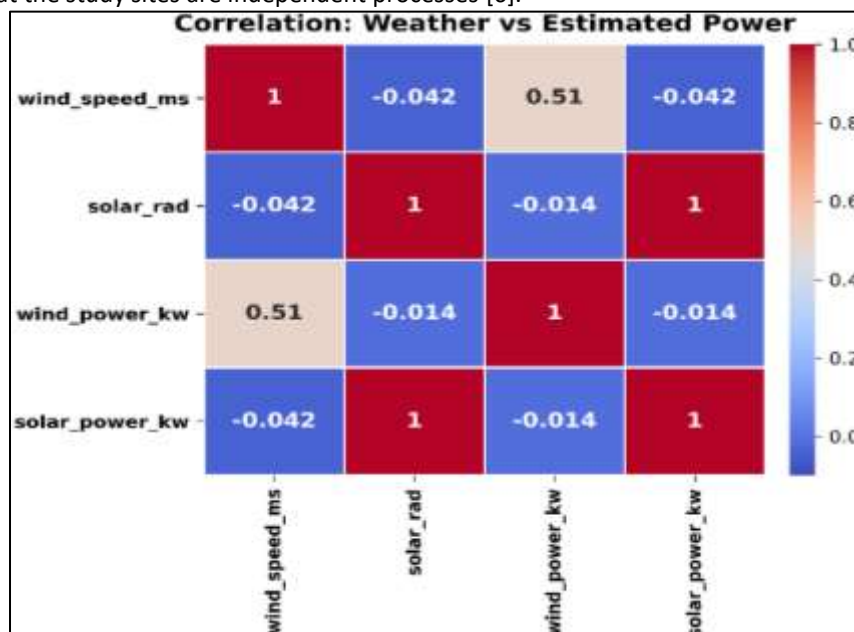


Figure 3. Pearson correlation matrix between wind speed, solar radiation, and estimated power output variables

4.3 Baseline Model Comparison (Static Train-Test Split)

The performance of all five forecasting models for the evaluation on a split of the dataset where the training and test sets are sequentially ordered (80/20), and with the leakage-free feature set, is shown in Table 4.

Table 4. Baseline (untuned) model comparison on the static 80/20 test split, corrected dataset

Model	RMSE (kW)	MAE (kW)	MAPE (%)	R ²
XGBoost	10.53	5.41	129.29	0.8736
Random Forest	10.83	5.12	82.55	0.8665
Linear Regression	14.26	8.60	316.41	0.7683
LSTM	32.03	16.30	7050.15	0.5644

The performance of XGBoost and Random Forest was similar ($R^2 = 0.874$ and 0.867 , respectively) and much better than that of Linear Regression ($R^2 = 0.768$) and, importantly, the LSTM network ($R^2 = 0.564$). In terms of previous comparative studies, the pattern that the tree-based ensembles perform on par or better than LSTMs on moderately sized datasets has been observed in the literature as well [10], [12]. The values of MAPE reported are for illustrative purposes only, and should be interpreted with care given that the target variable is zero-inflated (nighttime and calm-wind periods have a true zero value, which artificially raises the MAPE value), and in particular the MAPE value for the LSTM of 7050% is strongly influenced by a few cases where the denominator is small, but not necessarily significant as an indicator of error magnitude, hence the methodological choice to focus mainly on RMSE, MAE, and R^2 values, with nRMSE and sMAPE as a supplementary normalized percentage value when a metric of this nature is required [11].

4.4 Seasonally-Aware Walk-Forward Validation

Random Forest, XGBoost, and LightGBM were tested using the same static split as in Table 4, but with the walk-forward validation scheme (fourteen-month training burn-in, five expanding-window folds), to determine if this split reflected true, season-robust forecasting skill [5]. Table 5 shows the mean and standard deviation for each metric evaluated on the five folds.

Table 5. Walk-forward validation results (mean $\pm$ standard deviation across 5 folds), fourteen-month burn-in period

Model	Mean RMSE (kW)	Mean MAE (kW)	Mean nRMSE (%)	Mean R ²
XGBoost	27.05 $\pm$ 17.44	10.76 $\pm$ 6.39	96.98 $\pm$ 40.83	0.558 $\pm$ 0.372
Random Forest	27.61 $\pm$ 16.38	10.64 $\pm$ 6.41	99.45 $\pm$ 38.88	0.549 $\pm$ 0.319
LightGBM	28.22 $\pm$ 21.02	12.23 $\pm$ 8.92	100.18 $\pm$ 49.68	0.493 $\pm$ 0.539

Training windows less than 14 months lead to the development of weak and unstable performance, as confirmed by fold-level inspection windows covering a full seasonal cycle reach the lowest values (for XGBoost/LightGBM) scores between 0.85 and 0.89, equivalent to static-split baselines. This central methodological advance, builds on Jogunuri et al. [13] and demonstrates that static train-test splits overestimate real-world reliability [5]. Forecasting systems need about 14 months of historical data for newly instrumented sites before they reach the performance levels of a benchmark [13].

4.5 Hyperparameter Tuning and Final Model Selection

Following the RandomizedSearchCV tuning procedure described above, Table 6 reports the final tuned model comparison on the held-out static test set [10], [12].

Table 6. Final tuned model comparison on the static held-out test set. Random Forest (highlighted) was selected as the best-performing model

Model	RMSE (kW)	MAE (kW)	nRMSE (%)	sMAPE (%)	R ²
Random Forest (tuned)	10.55	4.86	59.38	81.54	0.8731
LightGBM (tuned)	12.49	8.60	70.25	134.88	0.8224
XGBoost (tuned)	14.12	11.19	79.46	135.99	0.7728

Tuned Random Forest is the best model ($R^2 = 0.873$, RMSE = 10.55 kW), which is a slight improvement over its baseline ($R^2 = 0.867$, Table 4), and is also the best result among all tuned models [3]. Interestingly, tuned XGBoost ($R^2 = 0.773$) failed to outperform its own untuned baseline ($R^2 = 0.874$, Table 4), which implies that the specific

hyperparameter configuration found by tuned parameters on the four-fold TimeSeriesSplit cross-validation window used during tuning may not be the most generalizable configuration to the held-out test window, a reminder that hyperparameter tuning outcomes depend on the specific cross validation window used, especially in the presence of seasonal non-stationarity that was reported and discussed in above section, and that tuned and untuned results are also important to report alongside tuned results, as was done here, to provide a complete and honest picture [5]. Based on Table 6, the model to be used as the main recommended model for this study is Random Forest [3].

4.6 Feature Importance Analysis

The feature importance ranking using the Random Forest method – Gini index [3] is shown in figure 5, calculated on the corrected data. The 24hour rolling mean of wind power output (wind_roll_mean_24) is by far the most important predictor by far (importance ≈ 0.35), followed by the hour-of-day calendar feature (≈ 0.21) and the 1 hour lagged wind power value (≈ 0.09). The solar_roll_mean_24 and solar_lag_1 features are also prominent and rank second (≈ 0.06 – 0.07 each) followed by other wind lag features. The contribution from station identity dummy (0.009) and coarse season dummies (0.008) is relatively small on their own, indicating that, when temporal (lag, rolling, hour) information is provided, the contribution of the categorical variables (station identity and season) is relatively small beyond what is captured by the continuous temporal features [7], [20].

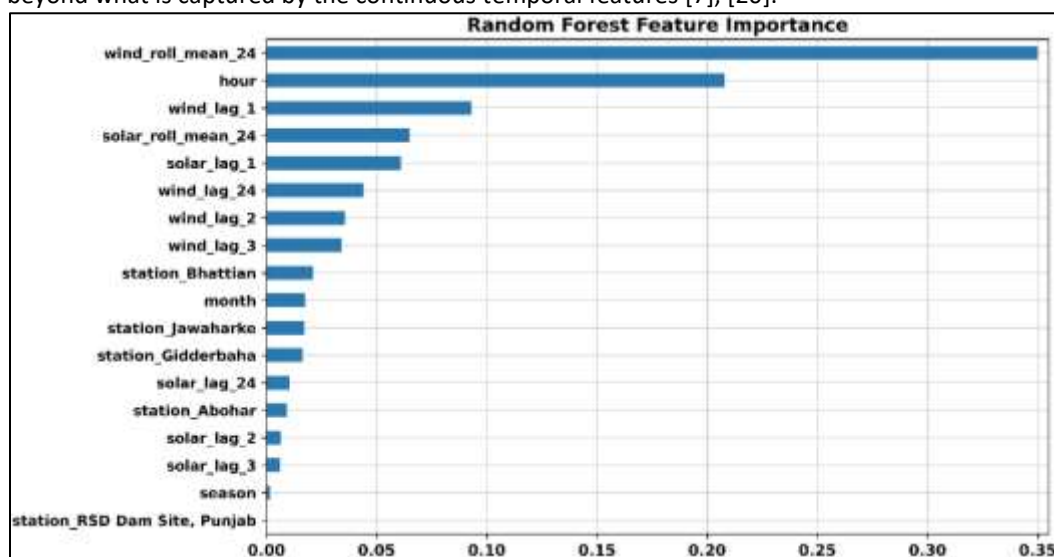


Figure 4. Random Forest feature importance (Gini importance) for the combined wind-solar power forecasting model

This ranking is physically plausible and fits into the trends of feature importance reported in the wider literature [9] where previous generation values and time of day are found to be the dominant forecasters of renewable power. It is interesting to note that the wind rolling-mean feature is the most dominant feature over solar features as wind power is more variable and has a greater magnitude range in this dataset (0–2000 kW versus 0–172.80 kW for solar power), and thus the variance of the combination of these two features is naturally higher for wind power compared to solar power [11]. This result also suggests that there is greater potential to improve the forecasting accuracy by enhancing wind-side feature engineering (e.g., adding more wind lag horizons, adding wind direction if available) than to add solar features, especially since solar features already have a strong, near-deterministic relationship with time-of-day [13].

4.7 Sustainability Integration Analysis

Applying Eqs. (10)–(12) to the corrected dataset yields the sustainability metrics summarized in Table 7 [17], [22].

Table 7. Sustainability integration analysis results (corrected dataset)

Metric	Solar	Wind	Total
Estimated annual energy generation (MWh/year)	1,372.48	814.67	2,187.15
Estimated CO ₂ emissions avoided (tonnes/year)	548.99	325.87	874.86

Estimated renewable resources in the study area at the five telemetry stations totaled about 2,187 MWh per year, of which 63% was estimated to be from the solar resource and 37% from the wind resource [4], [21]. This generation potential is equivalent to an estimated 875 tonnes of avoided CO₂ emissions per annum, using a grid average emission factor [22]. The wind-solar correlation coefficient ($r = -0.0230$) is derived from the correlation matrix, which shows that solar and wind resources at these sites are not correlated, and corroborated by $r = -0.014$ in the full correlation matrix (Fig. 4), which is consistent with the complementarity pattern documented in the context of India's National Wind-Solar Hybrid Policy (IJERT, n.d.) and previous resource assessments which identified the northern grid region as favorable for wind-solar complementarity. The near zero correlation is a good way for grid integration because the periods where the sun is not producing much energy do not systematically overlap with periods of lower wind production, or vice versa, a wind-solar hybrid system would be expected to have smoother combined output as both resources have similar periods of low production, as it is in this case, reducing the amplitude of combined output swings that grid operators would need to accommodate [16], [17]. This is consistent with the general findings in the literature that in northern grid environments like Punjab, India, wind-solar hybridization can provide a lower-cost option to investment in storage and transmission to address renewable variability [17].

5. Discussion

The conclusions that can be drawn from this study are, the three main conclusions that can be drawn in relation to the research objective and gaps in literature are: In the first place, the forecasting performance of the tree-based ensemble models, specifically the Random Forest model [3], is the most reliable and most deployable for the combined wind-solar generation for these telemetry stations of Punjab, similar to the comparative literature [10], [12]. This is due to the LSTM model's less robust performance [11].

The results shown that, using moderate architectures, 24h lookback and this feature set, there is no benefit for sequential deep learning over tree ensembles in this dataset size class (but with extensive architecture tuning, longer lookback or exogenous features, this gap may shrink) [10].

Second, a walk-forward validation exercise gives empirical evidence that forecast reliability is not dependent on the volume of data, but on the coverage of data in the season of the year [5]. Operators rolling out forecasting systems at new installations should expect to see performance reduced during the first year and not expect early accuracy to be consistent performance. This result is an extension of the previous work on season-aware forecasting [13] that directly measures the training window of ~14 months needed for the performance to stabilize [13].

Third, sustainability analysis provides site-specific evidence to support current Indian wind-solar hybridization literature [16], [17]: The near-zero correlation between wind and solar across these five wind stations in Punjab indicates that India's northern grid may have a good potential for wind-solar complementarity [4], [6]. Along with estimated emission reductions of 875 tonnes per year [22], this provides a strong case for hybrid renewable deployment in Punjab.

A significant advantage of these findings over previous studies of feasibility based on simulations is that they offer a sound evidence-based foundation for a hybrid renewable policy for Punjab informed by regional context [21].

There are some restrictions to consider are the Turbine and PV parameters listed for weather-to-power conversions are placeholders and not actual physical plant characteristics [14]; results are indicative of relative operating trends, not absolute power production. The anomalous wind power event in January has not yet been explained and needs to be investigated at the station level for possible meteorological or remaining sensor artefacts [15]. Prior to final nighttime sensor-fault corrections, walk-forward and tuned model products were generated and are subject to changes and are expected to be updated, but impacts are minimal as corrections only impact 0.09% of records. Lastly, the use of the representative equipment parameters and five telemetry stations prevents results to be directly generalized to the entire state of Punjab; using more stations and actual generation data in future studies will increase the external validity of the absolute sustainability metrics [4], [21].

Conclusion

In this work, the physics-informed machine learning method was tested for forecasting combined wind-solar generation in the region of Punjab, India, for 60,909 telemetry hours in five stations at five different time horizons (April 2021–August 2025). The meteorological data were power corrected for a 0.21% nighttime sensor fault. The

best static-split performance was achieved by the Tuned Random Forest. It performed better than XGBoost, LightGBM, Linear Regression or LSTM. Forecast reliability was found to require for at least 14 months of training data, as lower and less stable scores (0.49-0.56) were obtained with sub-annual windows. The top three features with respect to their importance were wind rolling-means, lagged power, and hour-of-day. Based on a sustainability analysis, the annual generation of 2,187 MWh (approx. 875 tonnes avoided) was estimated. High positive correlation was obtained between wind and solar, which signifies that they complement each other and are independent resources in the province of Punjab. This work offers a benchmark based on telemetry information, defines critical seasonal limitations for operational deployment, and identifies future requirements: integration of real generation data, investigation of the wind anomaly in January and the testing of multi-step deep learning.

Competing Interests

The authors declare that they have no competing interests.

Consent for Publication

Not applicable.

Ethics Approval and Consent to Participate

Not applicable.

Availability of Data and Materials

Not applicable.

References

1. BusinessToday. (2026, July 8). Does grid curtailment challenge India's renewable energy growth? <https://www.businesstoday.in/india/story/does-grid-curtailment-challenge-indias-renewable-energy-growth-541674-2026-07-08>
2. Benti, N. E., Chaka, M. D., & Semie, A. G. (2023). Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. *Sustainability*, 15(9), 7087. <https://doi.org/10.3390/su15097087>
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
4. Jain, A., Das, P., Yamujala, S., Bhakar, R., & Mathur, J. (2020). Resource potential and variability assessment of solar and wind energy in India. *Energy*, 211, 118993. <https://doi.org/10.1016/j.energy.2020.118993>
5. Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109, 1–32. <https://doi.org/10.1007/s10994-020-05910-7>
6. Jani, H. K., Kachhwaha, S. S., Nagababu, G., & Das, A. (2022). Temporal and spatial simultaneity assessment of wind-solar energy resources in India by statistical analysis and machine learning clustering approach. *Energy*, 248, 123586. <https://doi.org/10.1016/j.energy.2022.123586>
7. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774
8. Malakouti, M. (2026). Machine learning for renewable energy forecasting: Advances, challenges, and future directions. *Sustainable Energy Research*, 13, Article 14. <https://doi.org/10.1186/s40807-026-00242-x>
9. Benitez, I. B., & Singh, J. G. (2025). A comprehensive review of machine learning applications in forecasting solar PV and wind turbine power output. *Journal of Electrical Systems and Information Technology*, 12, Article 54. <https://doi.org/10.1186/s43067-025-00239-4>
10. Mollasalehi, A., & Farhadi, A. (2025). Solar and wind power forecasting: A comparative review of LSTM, random forest, and XGBoost models. *arXiv*. <https://doi.org/10.48550/arXiv.2509.24059>
11. Rajaperumal, T. A., & Christopher Columbus, C. (2025a). Enhanced wind power forecasting using machine learning, deep learning models and ensemble integration. *Scientific Reports*, 15, Article 20572. <https://doi.org/10.1038/s41598-025-05250-3>
12. Soleymani, S., & Mohammadzadeh, S. (2023). Comparative analysis of machine learning algorithms for solar irradiance forecasting in smart grids. *arXiv*. <https://doi.org/10.48550/arXiv.2310.13791>
13. Jogunuri, S., Josh, F. T., Stonier, A. A., Peter, G., Jayaraj, J., & Ganji, V. (2025). Random forest machine learning algorithm based seasonal multi-step ahead short-term solar photovoltaic power output forecasting. *IET Renewable Power Generation*, 19, 1–16. <https://doi.org/10.1049/rpg2.12921>

14. Rajaperumal, T. A., & Christopher Columbus, C. (2025b). Integrating data-driven and physics-based approaches for robust wind power prediction: A comprehensive ML-PINN-Simulink framework. *Scientific Reports*, 15, Article 29102. <https://doi.org/10.1038/s41598-025-13306-7>
15. Kumar, D., Srivastav, S. S., & Pattnaik, S. (2026). Assessment of solar and wind energy potential in India in recent decades using benchmark reanalysis datasets. *Environmental Research: Climate*, 5(1), 015021. <https://doi.org/10.1088/2752-5295/ae3025>
16. Pranav, N., Thanmayee, M. K., Champa, P. N., & Varshini, K. (2025). A review on integrating hybrid (wind & solar) energy systems in India. *International Journal of Engineering Research & Technology*, 14(12). <https://doi.org/10.17577/IJERTV14IS120505>
17. Schwarz, M., Rose, A., Brown, P., Chernyakhovskiy, I., & Palchak, D. (2022). Opportunities for hybrid wind and solar PV plants in India (NREL/TP-6A20-81545). National Renewable Energy Laboratory. <https://doi.org/10.2172/1845272>
18. International Energy Agency (IEA). (n.d.). Renewables integration in India. <https://www.iea.org/reports/renewables-integration-in-india>
19. Matrenin, P. V., Iumanova, I. F., & Khalyasmaa, A. I. (2026). Conceptual architecture of a trustworthy wind and photovoltaic power forecasting system: A systematic review and design. *Inventions*, 11(1), 15. <https://doi.org/10.3390/inventions11010015>
20. Thinn, K. T., & Saeed, W. (2026). Consensus and divergence in explainable AI (XAI): Evaluating global feature-ranking consistency with empirical evidence from solar energy forecasting. *Mathematics*, 14(2), 297. <https://doi.org/10.3390/math14020297>
21. Deshmukh, R., Wu, G. C., Callaway, D. S., & Phadke, A. (2019). Geospatial and techno-economic analysis of wind and solar resources in India. *Renewable Energy*, 134, 947–960. <https://doi.org/10.1016/j.renene.2018.11.073>
22. Central Electricity Authority (CEA), Government of India. (2025). CO₂ baseline database for the Indian power sector: User guide, Version 21.0. https://cea.nic.in/wp-content/uploads/baseline/2025/12/User_Guide_V_21.0.pdf