Renewable Energy Focus 50 (2024) 100615

Contents lists available at ScienceDirect

Renewable Energy Focus

journal homepage: www.elsevier.com/locate/ref

Research Paper

A Three-Step Weather Data Approach in Solar Energy Prediction Using Machine Learning

Tolulope Olumuyiwa Falope [a,*], Liyun Lao [a], Dawid Hanak [b]

[a] Energy and Sustainability, School of Water, Energy & Environment, Cranfield University, MK43 0AL, UK
[b] Net Zero Industry Innovation Centre, Teeside University, TS1 3BX, UK

ARTICLE INFO

Keywords:
Solar energy forecasting
Machine learning
Solar radiation
Low-level data fusion
Pearson correlation coefficient

ABSTRACT

Solar energy plays a critical part in lowering CO_2 emissions and other greenhouse gases when integrated into the grid. Higher solar energy penetration is hindered by its intermittency leading to reliability issues. To forecast solar energy production, this study suggests a three-step forecasting method that selects weather variables with a moderate to strong positive correlation to solar radiation using Pearson correlation coefficient analysis. Low-level data fusion is used to combine weather inputs from a reliable local weather station and an on-site weather station, significantly improving the forecasting model's accuracy regardless of the machine learning method used. Weather data was obtained from the Kisanhub Weather Station located in Cranfield University, UK and the meteorological station in Bedford, UK. In addition, PV power supply data was obtained from four solar plants. Using the Regression Learner app in MATLAB, the proposed architecture is tested on a utility scale solar plant (1 MW), showing a 6% and 13% prediction accuracy improvement when compared to solely using data from the on-site and local weather station respectively. It is further validated using data from three residential rooftop solar systems (8 kW, 10.5 kW and 15 kW), achieving root-mean square values of 0.0984, 0.0885, and 0.1425 respectively. The data was pre-processed using both rescaling and list-wise deletion methods. Training and testing data from the 1 MW solar plant was divided into 75% and 25% respectively, while 100% of the residential rooftop solar plants was used for validation.

Introduction

The intermittency of grid-connected renewable energy sources (RES) like solar energy negatively impacts grid reliability, presenting a major challenge to grid managers because energy supply does not always match up with demand [1,2]. To combat this problem, accurate solar forecasting is critical to its integration into the grid.

Solar forecasting involves gathering and analyzing data to predict future solar irradiance or solar power [3]. Data can be historical or in real-time, and can be provided by weather stations, remote sensing (satellite imaging) and so on.

When choosing an overall solar forecasting strategy or approach, there are six main factors to consider. These include the input or raw data's source and format, how the data is pre-processed, data mapping or the forecasting architecture, the forecasting method, the desired output or targeted outcome, and the forecasting time frame.

The first factor deals with the type of weather data, categorized into historical, real-time, and forecast data. In addition, the data source is also important. The data source refers to where all data values used in the model are obtained from. Example of data sources include on-site and local weather outposts, or even software generated data sources like MATLAB.

The second factor revolves around how the raw data is cleaned or pre-processed. Data pre-processing involves operations and techniques that change raw data into quality input data for analysis to obtain a meaningful output [4]. Data pre-processing is critical for effective model performance [5].

The third factor is an architecture that describes the relationship between input variables and output variables [3]. The strength of the prediction model is reliant on the correlation of input variable(s) to the output [6].

The fourth factor is the forecasting method. Solar forecasting methods are the engine room and refer exclusively to the process of converting the input (predictors) into output (response). The forecasting architecture mentioned above may use or combine one or more methods. These methods are classified into physical and statistical (or non-physical) methods. Physical methods predict outputs like solar

* Corresponding author.
E-mail address: t.o.falope@cranfield.ac.uk (T.O. Falope).

https://doi.org/10.1016/j.ref.2024.100615
Received 7 April 2024; Received in revised form 12 May 2024; Accepted 20 August 2024
Available online 23 August 2024
1755-0084/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Nomenclature

Abbreviations

ANFIS	*Adaptive Neuro Fuzzy Inference System*
ANN	*Artificial Neural Network*
ARMAX	*Autoregressive Moving Average with Exogenous Inputs*
FCN	*Fully Convolutional Network*
LLDF	*Low-level Data Fusion*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
ML	*Machine Learning*
MSE	*Mean Squared Error*
NWP	*Numerical Weather Predictions*
PCA	*Principal Component Analysis*
PCC	*Pearson Correlation Coefficient*
PV	*Photovoltaic*
RES	*Renewable Energy Sources*
RLA	*Regression Learner App*
RM	*Rescaling Method*
RMSE	*Root Mean Square Error*
SR	*Solar Radiation*
SVM	*Support Vector Machine*

irradiation by using numerical weather predictions (NWP) models with weather variable inputs like temperature, relative humidity, cloud cover and so on [7,8]. Statistical methods depend purely on historical and/or real-time data to arrive at predicted variable(s) that have a linear correlation with the data over the same time frame. Hybrid models, which offer a higher predicting accuracy, can be created by combining several methods or forecasting algorithms.

The fifth factor is the predicted or output variable, which categorizes forecasting into direct (solar power generated) or indirect (solar irradiance) [6]. Direct forecasting involves forecasting the PV power output directly [9]. Indirect forecasting predicts the solar irradiation first before estimating the PV output using mathematical formulas that describes the specific system under consideration. The solar power and the PV generated refer to the same measure.

The last factor is the forecasting time horizon or timescale. This is the time frame from the present into the future for which an outcome is predicted. Forecasting time horizons are classified into intra-hour or very short-term (5 mins to 60 mins), short-term (1 hour to 6 hours), medium-term (1 day to several days), long-term (week(s) to a year or more). Each of the time horizons is suitable for specific applications. For example, very short-term forecasting is used to achieve responsive control for RE power plants, market clearing, real-time dispatch, and load tracking. Short-term forecasting is used to schedule energy flow from power sources, load-following, and initiating back-up storage. Medium-term forecasting is used for congestion management, market trading and price settlement, predicting reserve requirements and load dispatch. Long-term forecasting is used for contingency analysis, operation management, and resource and maintenance planning [7,10].

Based on the above considerations, several studies have developed varying approaches to forecast solar energy supply with an emphasis on a next day or 24-hour ahead (day-ahead) timescale. As mentioned, one of the key considerations is the data source and type. In Carrera et al. [11], the authors use a combination of forecast and historical weather data from a local weather station, in addition to historical PV generation and sun elevation. Minhas et al. [12] uses historical and forecast data from a local weather station and the historical PV data from the solar plant. Table 1 gives a summary of forecasting architecture for day-ahead PV generated predictions.

As can be seen by the summary given in Table 1, this study proposes a unique combination of historical photovoltaic (PV) data and historical weather data from both an on-site and local weather station to forecast the next-day PV supply. No other study has this data combination. This is significant for a few reasons. The first reason is that the combined sum of data from multiple relevant data sources is better and richer in information than data from a single source; obtaining more effective information and optimizing data combination [19]. The emphasis of the above statement is on "multiple" and "relevant" data because the higher the data quality, the higher the accuracy. While other studies use data from one weather station, this study combines data from two relevant independent sources (local and an on-site weather station) using low-level data fusion (LLDF). LLDF simply combines or concatenates several sources of raw data to produce a new single data sample [20].

The proposed forecasting architecture is a three-step process that predicts PV generation by selecting weather variables with moderate to strong positive correlations to solar radiation (SR) from the two weather stations mentioned above, combined with historical PV generation data obtained from a PV power plant, into machine learning (ML) algorithms from MATLAB's regression learner app (RLA). Data is further pre-processed using List-wise deletion method (LWDM) and rescaling method (RM). SR is the most important predictor to solar power, or PV generated [11,21,22] hence its selection as the reference parameter. To validate this, temperature, which is also closely linked to PV generation [23] will be tested as a reference parameter. Additionally, the correlation between the input and output model variables has a significant impact on how well a forecasting model performs [6]. With the output variable being the PV generated, it is logical to choose inputs that have a strong correlation to PV generation. MATLAB's RLA is a ML toolbox app used to train forecasting algorithms based on known output and input parameters. It selects the best performing out of more than twenty ML algorithms based on the lowest root mean square error (RMSE).

The aim of this study is to investigate the accuracy of the proposed approach and determine its adaptability to both commercial/utility and residential solar plants. Furthermore, the proposed model is benchmarked against other existing forecasting models that use Pearson

Table 1
Summary of day-ahead solar energy forecasting studies

Research Study	Historical Observed Weather Data from Weather Station	Historical Forecast Weather Data	HistoricalObserved Weather Data On-Site	HistoricalObserved PV
This study	✓		✓	✓
Carrera et al. [11]	✓	✓		✓
Minhas et al. [12]			✓	✓
Aghajani et al. [13]		✓		✓
Manjarres et al. [14]	✓	✓		✓
Pawar et al. [15]	✓			✓
Sangrody et al. [16]	✓			✓
Pawar et al. [17]	✓			✓
Feng and Yu [18]				✓

Correlation Coefficient (PCC) analysis to select features for the forecasting model. To achieve this, three forecasting scenarios with different data inputs are considered: (1) data pre-processed with PCC from on-site weather station, (2) data pre-processed with PCC from local weather station, (3) data from both stations that is combined using LLDF and pre-processed with PCC. In addition, the use of temperature as a reference parameter for feature selection and a comparison of the model without PCC is also analysed.

The main contribution of this paper is to assess the proposed three-step forecasting model by evaluating:

i. The ability of LLDF to combine multiple weather data sources to improve solar forecasting accuracy.
ii. The effect of using PCC versus not using PCC in predictor selection for regression models.
iii. The performance of SR as a reference parameter for selecting predictors into learning algorithms.
iv. The performance of the proposed model in both residential and utility-scale solar applications.

The remainder of this paper in organized into five sections. Section 2 provides a literature review, covering data types and sources, PCC-based models, and forecasting architectures. Section 3 gives a description of methods used. The proposed forecasting approach is discussed in Section 4. Section 5 presents and discusses the results. Finally, Section 6 summarizes the conclusions, recommendations, and limitations of the study.

Literature review

The input data can be made up of either historical weather data (observed and/or forecast), real-time data, historical solar energy generation data, or time of day data. While a few studies employ a univariate approach where only the historical PV generation is used [24-27], this research finds that most of the previous studies rely on two sources of data: the historical PV generation and some form of weather data [11-14,16,17,28-53]. This may be because multivariate approaches perform better than univariate [54]. Fewer studies use more than two data sources. This is important because data source is critical to the model performance [49]. In addition, not all site locations have access to both sets of data. Some sites only receive forecast data while others only have observed data.

A further dimension in evaluating the input weather data used is the weather data proximity to the generation plant. The solar energy production of a solar distributed system, solar farm or mini grid depends on the local weather conditions, particularly the SR of that specific location. In fact, no two PV systems show the same performance even if they are identically sized, due to the unique combination of meteorological parameters tied to their geographical locations [6]. In other words, no two locations will have the same correlation of meteorological parameters and PV power generated. Despite the interdependence of local, regional, and global meteorological data, they all offer varying degrees of accuracy. Most research studies use data from local or regional weather stations for their weather analyses. Although these data sources can be very accurate, it is not uncommon that a mismatch exists between them and exact data from the site. Access to an on-site weather station improves the quality of weather data [55]. Site-specific conditions that affect the weather but are not represented in local or regional weather data can be recorded by an on-site weather station. To get the most accurate result, it is best to pair data from a local weather station with data from an on-site weather station where the forecasting system is being deployed [55]. Other options are to exclusively use the data from the on-site weather station. For instance, [28] uses a rooftop PV station located at the solar array site to collect all the data used by the study. [29-32,56] are studies that also use this approach.

The use of observed weather data from a local or regional weather station tends to be more accurate than forecast data because they include more variables which may have a direct correlation to solar generation [49]. For example, Ngoc-Lan Huynh et al [33] use solar radiation and Pedregal et al. [34] use global horizontal irradiation as the only weather parameters inputs into their prediction models. Other studies take a multivariate weather approach. Rodríguez et al. [38] use pressure, humidity, and temperature in their forecast model. Thukral et al. [39] (air temperature, relative humidity, wind, solar radiation), Chen et al. [43] (relative humidity, temperature, global horizontal radiation), Korkmaz [36] (solar radiation, temperature, humidity) and several other studies [35,41,42,57-59] use this multivariate weather approach. Whether using one weather parameter or using several, observed data tends to be more accurate than forecast data. While this is good, using weather data from only one source can be problematic if the source is somehow compromised.

Iyengar et al. [44] uses historical forecast data as a primary predictor and represents another approach used by several studies [45-47]. This data approach has an advantage that is also its Achilles heel. Forecast data needs to be accurate to inspire confidence in the data by the general public, who are the core users. The drawback is that this data then tends to be more conservative due to concerns of inaccuracy and can be a bit vague by giving categorical variables instead of numerical variables. For example, the United Kingdom's national Meteorological Office (the Met Office), gives a forecast in Bedford UK of $<5\%$ as the chance of precipitation instead of the actual percentage figure. Another drawback as mentioned earlier is forecast data tend to have fewer weather parameters than observed data. In the same Bedford weather station, the Met office forecasts seven weather parameters as opposed to reporting twelve parameters as observed data.

Carrera et al. [11], Kim et al. [49], and Abedinia et al. [48] use a hybrid approach, incorporating forecast data with observed data. This method is more accurate [50] because it incorporates the advantages of both sources.

The use of PCC in selecting features for forecasting is a well-established technique. Nematirad and Pahwa [60] use PCC and SR as a reference parameter to select predictors from a meteorological database before passing the data into an artificial neural network with Bayesian optimization. The authors determined that PCC outperformed Principal Component Analysis (PCA) in the selection of features. A similar study, Rodriguez et al. [38] use a government database to obtain weather variables. Using PCC with reference to SR, SR was selected as the only input into the proposed Artificial Neural Network (ANN) model. Jebli et al. [23] use a similar model to the one proposed in this study. Meteorological inputs are pre-processed using PCC, data scaling (RM) and handling of missing data before being trained by four ML algorithms namely Linear Regression, Random Forest, Support Vector Regression and ANN. With weather data obtained from a regional database in Morocco, this forecasting approach is similar to Scenario 2 mentioned in Section 1.

In a review of forecasting architectures, Kim et al. [49] propose a two-stage architecture that uses an auxiliary model in the first stage to identify the relationship between forecast and auxiliary variables in the weather data. In the second stage, the model receives the output from the first to forecast the PV energy. This results in an improvement in the forecasting model accuracy. Kyliashkina et al. [50] suggest an alternative strategy in which forecast data is adjusted using the observed data rather than the auxiliary variables. Abedinia et al. [48] selects both historical and forecast data for solar irradiation, weather temperature and cell temperature; three weather parameters with strong correlations to PV generation. However, local conditions specific to the site are not reflected in the forecast.

Looking at the previous studies including those listed in Table 1, it is clear to see that no study uses the unique three-data source approach proposed by this study which combines historical observed data from both an on-site and local weather station, along with the historical PV data using LLDF. It is intuitive that the level of accuracy of the weather data in this approach increases as it gets closer to the power generating

site. In other words, the proximity of the weather station to the PV site is directly proportional to the accuracy of the weather data and must be considered when forecasting [38]. The reason for using the most accurate weather data is that the forecasting model's performance depends on the accuracy of the database [38].

In addition, several studies are examined to determine the application of LLDF and other techniques used in enhancing and combining raw data to build the forecasting model. Site-adaptation techniques have been developed to improve satellite-derived weather data by using in-situ data from meteorological stations [22]. But these methods rely on complex ML and deep learning algorithms to transform satellite images into solar irradiation values as an example. Jesus et al. [61] uses data fusion based on a hybrid of two deep neural networks namely Fully Convolutional Network (FCN) and Long Short-Term Memory (LSTM) network, to fuse the output of two individual forecast models known as Autoregressive Moving Average with Exogenous Inputs (ARMAX) and Adaptive Neuro Fuzzy Inference System (ANFIS) to predict short-term PV output. In another short-term prediction study, Wang et al. [19] uses multi-source data fusion to fuse NWP, PV output data and remote satellite sensing data with a Support Vector Machine (SVM) algorithm. Results showed that the data fused SVM model performed better than the regular SVM model more than 95% of the time. None of the highlighted studies use LLDF to combine data from an on-site weather station and a trustworthy local weather station to improve the forecasting data used in the model.

Methods

The data and procedure used to create the forecasting model are described in this section. The source and type of data are described in Section 3.1. Section 3.2 outlines the pre-processing methods used to clean the data. The error equations used to assess the model performance are described in Section 3.3. The methodology and flow chart of the RLA in MATLAB are highlighted in Section 3.4.

Data description

In this case study, the Kisanhub weather station (KH075) [62] at Cranfield University, recorded data every fifteen minutes on nine weather parameters, including solar irradiation. Historical PV data from four PV plants located in Cranfield University representing the spectrum of distributed generation customers is obtained from Cranfield's online energy manager platform [63]. The aim is to investigate the proposed architecture's performance in both commercial and residential applications. The specification and data profiles of the locations are listed in Table 2 and 3. 75% of the data from the 1 MW solar farm was used for training the model while the remaining 25% was used for testing. 100% of the data from the three residential solar arrays was used to validate the model. Upon request, The Met Office station located in Bedford, the station closest to Cranfield, provided hourly data for twelve weather parameters. Missing data points were simply deleted from the data sample.

The varied data sample periods of each PV plant in Table 2 and 3 were due to data availability on the energy manager portal.

Table 2
PV system specifications of 1 MW solar Farm in Cranfield University

Location	1 MW Solar Farm, Cranfield University
PV Generator Output	999.78 kWp
PV Generator Surface	5,858.4 m^2
Number of PV Modules	3508
Type of PV Modules	Hanwha Q cells: 285 W
Number of Inverters	35
Type of Inverters	Fronius Symo: 10 kW to 17.5 kW
Data Sample Period	January 2019 – December 2020
Training (% of data sample)	75%
Testing (% of data sample)	25%

Table 3
PV system specification of three roof-top solar arrays in Cranfield University

Building No.	B052A	B146 FAAM	B252
PV Generator Output	10.45 kWp	15 kWp	8.26 kWp
Number of PV Modules	35	60	28
Type of PV Modules	BYD: 275 W	Vikram: 250 W	Q.antum: 295 W
Number of Inverters	1	1	1
Type of Inverter	Solis: 12 kW	PVI: 12.5 kW	Solar Edge: 10 kW
Data Sample Period	Jan. 2020 – Dec. 2020	July 2018 – Sept. 2021	Dec. 2019 – Oct. 2021

Data pre-processing

List-wise deletion method

Missing data is almost certain because the physical properties of measuring devices are prone to wear and tear and other mechanical malfunctions. In addition, human error and faulty communications can cause these devices to either omit data or give wrong measurements, resulting in data bias [64]. LWDM is one of the simplest ways of handling missing data [65]. It deletes all measured values corresponding to a specific timestamp if the value of a single variable is missing. The advantage of this method is that the original data set remains 100% authentic without any value estimation. The disadvantage is that the sample size is reduced and the contribution from variables with available values is made redundant by one missing measurement or faulty device. However, this is compensated for if the sample size is large enough.

Low-level data fusion

The aim of data fusion is to improve the overall accuracy of a system or model by integrating data from multiple sources. Data fusion can be categorized into low-level, mid-level or intermediate, and high-level fusion. LLDF or measurement level fusion [66] combines more than one source of raw data to produce new raw data [20]. In this study, LLDF is used to combine PV data, and weather data from an on-site and local weather stations.

Pearson correlation coefficient analysis

The PCC function [67] given in Eq. (1), is a feature selection technique used to select relevant features for model prediction while ignoring redundant variables. In this study, PCC is used to determine which weather variables have a moderate or strong positive correlation to SR. Results are displayed in a correlation matrix [68].

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \quad (1)$$

Where r is the correlation coefficient, x and y represent the values of the x-variable and y-variable in the sample respectively, $\bar{x}$ and $\bar{y}$ are the mean of the x-variable values and the y-variable values respectively.

The correlation coefficients, with a range of values between -1 to +1, indicate how strong the linear relationship between two variables is. A perfect negative correlation is equivalent to -1 on the scale and a perfect positive is equivalent to +1. This study is interested in weather parameters with a moderate to strong positive correlation to SR, which is equivalent to having a PCC greater than 0.4 [69]. The implication is that having a strong positive linear relationship to the most significant predictor of PV generation (SR) suggests the weather parameter is also a good predictor or feature. In a comparison of reference parameters, PCC is also used to select predictors with reference to temperature.

Rescaling method

Different weather parameters have varying units, scales and dimensions that can form a bias in determining the forecasting model. A weather parameter with a disproportionate number of data points can significantly influence a model by its sample size even though it has a weak correlation to the output variable. For this reason, it is imperative to normalize all variables to the same order of magnitude and weight [70]. RM or the min-max nominalization method is a popular and widely used method for normalizing data [71]. In this study, the intervals [a b] for rescaling were selected as [0 1]. The RM equation [72] is given in Eq. (2).

$$X_{rescaled} = a + \left[\frac{X - X_{min}}{X_{max} - X_{min}}\right](b - a) \tag{2}$$

Where $X_{rescaled}$ is the normalized data, X represents the values of the x-variable in the sample, X_{min} and X_{max} are the minimum and maximum values of the x-variable in the sample respectively.

Model performance evaluation

To evaluate the performance of the model under considered scenarios, four performance statistical error functions [73] and the coefficient of determination, R^2 [74], were used. The mean squared error (MSE) shown in Eq. (3) measures the average of the square of the errors. The mean absolute error (MAE) described in Eq. (4) measures the average magnitude of the errors. Eq. (5), which is the RMSE, is simply the square root of the MSE. R^2 in Eq. (6) is the proportion of the change in the dependent variable that can be predicted due to the corresponding change in independent variable(s). When evaluating how well the underlying models fit the data, this coefficient of determination is typically used because it is well-defined for linear regression models.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 = RMSE^2 \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(\hat{x}_i - x_i)| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}{\sum_{i=1}^{N}(\hat{x}_i - \hat{x}_{average})^2} \tag{6}$$

Where N is the number of data points, $\hat{x}_i$ are the predicted values and x_i are the observed values.

Regression learner app methodology and flow chart

After pre-processing, the data is then trained using the RLA; an application in the ML Toolbox in MATLAB. Regression analysis is used to determine the relationship and future predictions of a dependent variable (the PV generated) from an independent variable(s) (weather parameters). The RLA is a useful tool because it allows the user to train and validate a model using more than twenty different forecasting algorithms. It has also been used successfully in numerous ML applications that require forecasting [75-80]. After training, the user can compare the algorithms side by side based on their errors and choose the best model. The app identifies the best forecasting algorithm based on the lowest RMSE and highlights the model in the results display. Another advantage of the RLA is that multiple algorithms can be run simultaneously through a "parallel" function which allows the user to view the results in a shorter time frame than running each algorithm one after the other. To initiate the model training process via the app, the response (dependent variable) and predictors (dependent variables) must be selected. For this case study, the correlated weather data are the predictors while the response or output is the PV generation. Figure 1 highlights the flow chart of the RLA.

Methodology

The justification for employing the combined data from both weather stations is that weather stations capture weather parameters differently (different methodologies) and often times don't capture the same type or number of parameters. The premise in this section supports the assertion made above. The three-step forecasting architecture used to correlate input data with energy supply is described in Section 4.1. In order to evaluate the efficiency of the forecasting architecture, Section 4.2 presents a case study of three scenarios with different data inputs.

Forecasting architecture

To correct the missing data, LWDM was used in pre-processing to remove incomplete or missing data values from the two weather stations and PV plant. In the proposed approach, the SR measured by the Cranfield weather station is used as the reference weather parameter to select predictors from both weather stations. The selection of SR as the reference was made since it is the most significant predictor of PV generation. In step 1, the PCC function is used to calculate the correlation matrix of weather parameters reported by the on-site weather station in Cranfield. The weather parameters with moderate to strong positive correlations to SR are then selected as predictors for the forecasting model. By being strongly correlated to SR, this implies that these weather parameters will also have a strong linear relationship with PV generation and therefore are good predictors. In step 2, the correlation matrix is determined for the recorded weather parameters from the Bedford station. Again, the weather parameters with moderate to strong positive correlations to SR are selected as predictors. In step 3, the predictors from steps 1 and 2 and the historical PV generation supplied by the PV power plant meter (response) are combined using LLDF and are selected as inputs into the RLA. The input data is further pre-processed by using Eq. (2), before being separated into 75% training data set and 25% for model testing. The RLA flowchart presented in Figure 1 trains and test different forecasting algorithms and identifies the best forecasting model based on the lowest RMSE. Figure 2 shows an illustration of the forecasting architecture.

Case study description

Three scenarios are examined, and the outcomes are compared, to gauge the effectiveness of the forecasting architecture.

a. Scenario 1: Weather data measured by the Cranfield weather station with moderate to strong positive correlation to SR are selected as predictors to PV generated via PCC.
b. Scenario 2: Weather data measured by the nearby Bedford weather station with moderate to strong positive correlation to the measured SR at the Cranfield weather station are selected as predictors to PV generated via PCC. This scenario represents existing PCC-based scenarios mentioned in Section 2.
c. Scenario 3: SR-correlated weather data from both stations are combined using LLDF and used as predictors to PV generated.

The above scenarios will investigate whether using LLDF increases the forecasting model's accuracy and therefore is a valid component of the proposed forecasting architecture.

Results and discussion

Results from the case study and methods given in the previous

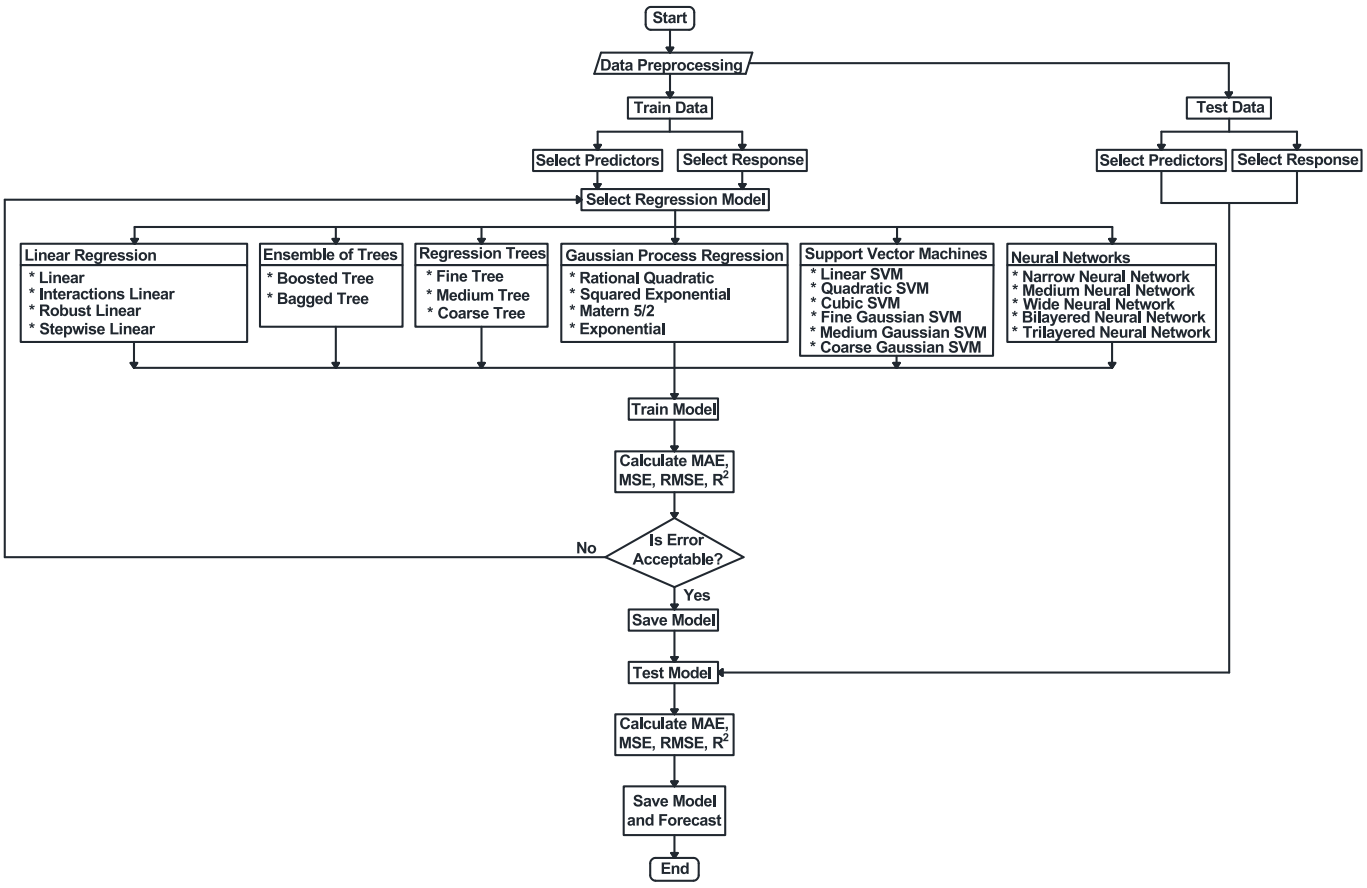


Fig. 1. Flow Chart of the Regression Learner App Methodology

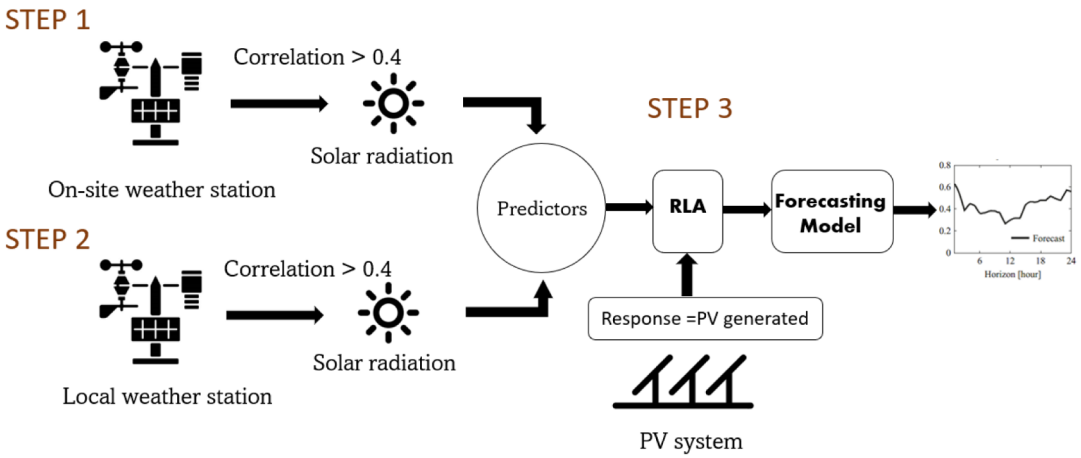


Fig. 2. Proposed three-step forecasting architecture

sections are summarized here. The outcome of the correlation matrix of the on-site meteorological variables is shown in Section 5.1. The correlation matrix of the weather variables captured by the Bedford weather station is shown in Section 5.2. Based on RMSE and R^2 values for the three case scenarios mentioned above, Section 5.3 shows which prediction model is the most accurate. Additionally, a prediction for the following day is used to demonstrate how well the model performs. Section 5.4 shows results validating LLDF and SR as the reference parameter, while the validation of the proposed model in residential rooftop applications is shown in Section 5.5.

Cranfield correlation matrix

According to the first step of the proposed forecasting architecture, weather parameters from the on-site weather station that have moderate to strong positive correlations to SR are selected based on coefficients above 0.4. Figure 3 shows a graphical representation of the weather parameters from the Kisanhub Weather Station in 2019.

Using the PCC function, the correlation matrix and plots are obtained. Table 4 and Figure 4 show the results. In Figure 4, the scatterplots show the relationship between a pair of variables and a trendline whose slop is equivalent to the correlation coefficient (boldly

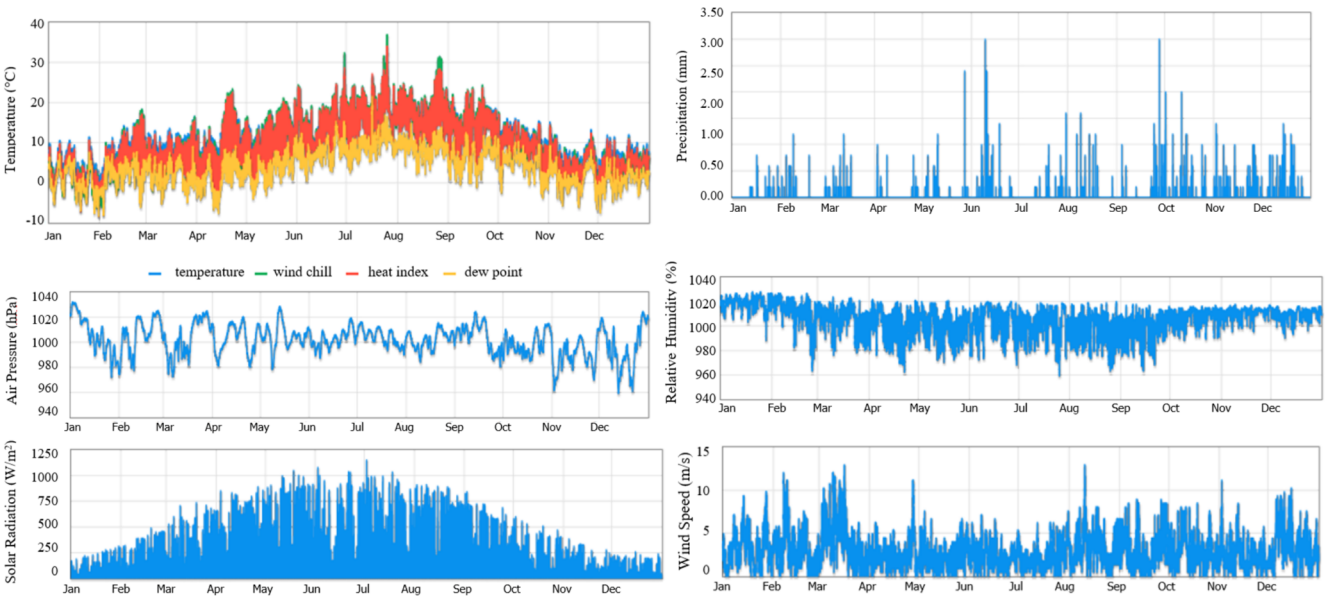


Fig. 3. 2019 Weather parameters from the Kisanhum Weather Station in Cranfield University [62]

Table 4
Cranfield correlation matrix

	Solar Radiation
Solar Radiation	1
Wind Speed	0.1515
Relative Humidity	-0.5910
Pressure	0.1483
Precipitation	-0.0677
Dew Point	0.2164
Heat Index	**0.5039**
Temperature	**0.5184**
Wind Chill	**0.4993**

highlighted in each subplot). The major diagonal are subplots containing the distribution of each variable as a histogram. Table 4 corroborates the values of the correlation coefficients in Figure 4 and shows the three weather parameters, Heat Index, Temperature and Wind Chill, with coefficients above 0.4.

This step is crucial in finding the predictors (SR included) or input variables into the forecasting algorithm for Scenario 1 and one-half of the predictors for Scenario 3. The selection of temperature-related parameters is consistent with findings from literature. SR has a strong linear relationship to outdoor surface air temperature [81,82]. Heat index is also called apparent temperature [83] and wind chill can also be expressed as wind chill temperature; a measure of human sensation to cold weather temperature [84]. All three weather parameters are related

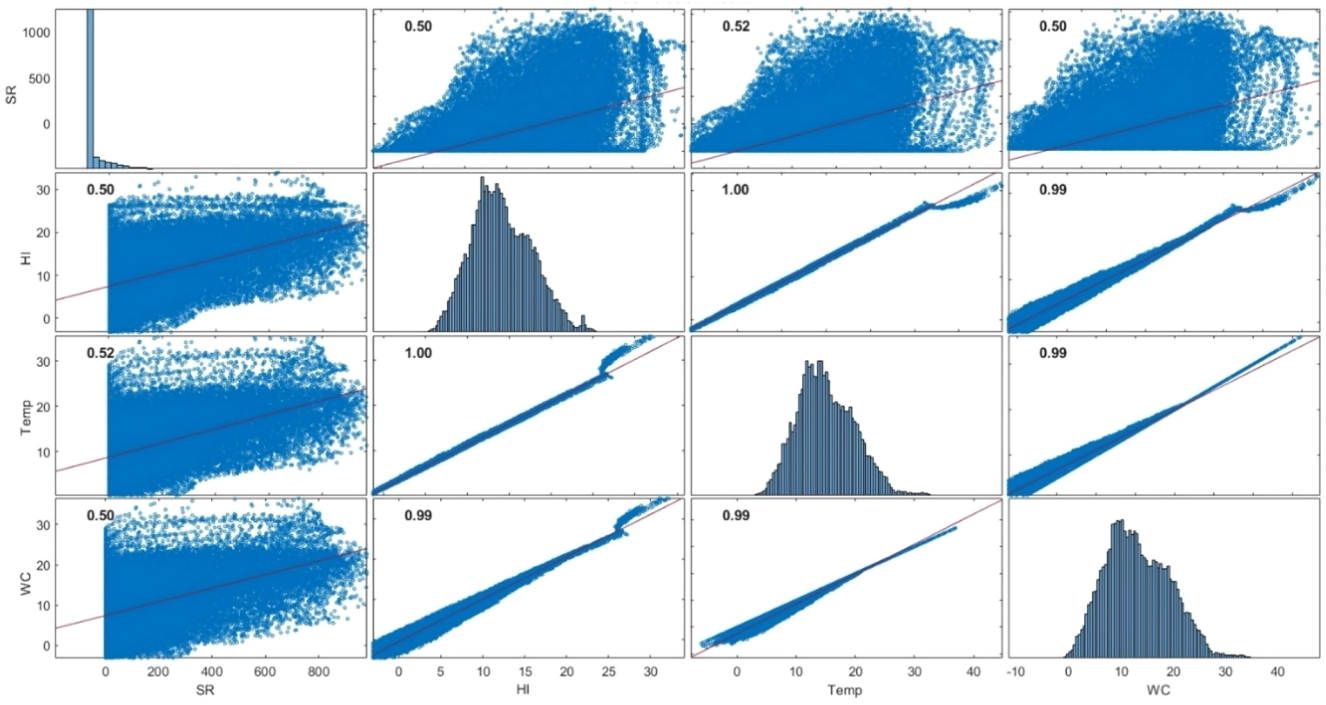


Fig. 4. Cranfield correlation plot Note: SR = Solar radiation, HI = Heat Index, Temp = Temperature, WC = Wind Chill

to the measurement of outdoor temperature.

Bedford correlation matrix

Similarly, from the Bedford station, the PCC function is used to obtain the correlation matrix where the Hourly Temperature, Hourly Wet Bulb Temperature and Hourly Global Radiation all had moderate to strong positive coefficients to SR according. Results are displayed in Table 5 and Figure 5.

In this step, the input variables for Scenario 2 and the other half of Scenario 3 are identified. Parameters related to temperature are once again selected. In a similar assessment of twenty-four weather parameters, Fraihat et al. [85] used a correlation coefficient benchmark of 0.5 to identify which parameter had a strong correlation to SR. Temperature and radiation were among eight out of the twenty-four parameters that fit this criterion. This is consistent with the analysis here. Scenario 2 is similar to a majority of the existing methods that use PCC as described in Section 2. The data source is either a nearby or local weather station, or a regional/national weather database.

Performance of forecasting architecture

The weather parameters highlighted in Section 5.1 and Section 5.2 are the predictors that become the input into the RLA as shown in the block-diagram in Figure 6. Table 6 shows the selected weather inputs of the three scenarios, with Scenario 3 being the proposed forecasting architecture.

Table 6 shows a summary of the training input parameters for each scenario when using the RLA, with the historical PV energy selected as the response for all of them.

According to Table 6, Scenario 1 has four predictors, Scenario 2 has three, while Scenario 3, which combines predictors from Scenarios 1 and 2 has seven. Each scenario is analyzed separately using the RLA flowchart in Figure 1. The RLA trains over twenty forecasting algorithms and identifies the best performing based on the lowest RMSE. To give an example of this, Scenario 3 is analyzed, and the results are shown in Table 7 (see Appendix A, Table A1 and A2 for similar analysis on Scenarios 1 and 2 respectively).

Table 7 shows that Boosted Trees had the lowest RMSE values for the testing phase. This model is then used to analyze Scenarios' 1 and 2 and the results are compared to the result in Table 7 and are displayed in Table 8.

According to Table 8, Scenario 3 had the lowest RMSE values for both training (0.1042) and testing (0.0859), a 6% and 13% improvement on testing from Scenario 1 and 2 respectively. This RMSE was also the lowest compared to any of the forecasting algorithms used for all the scenarios. This proves that by combining strong predictors from more than one source using LLDF, the accuracy of the forecast is improved. This is likely because forecasting depends on the number of relevant inputs in the forecasting architecture that correlate strongly with the forecasting output. The 13% increase is also an improvement on existing PCC studies such as that mentioned in Jebil et al. [23]. Figure 7 shows a graphical illustration of the data points between the predicted and the actual responses for the three scenarios. The trendline of Scenario 3 is more accurate than the other scenarios because its R-square value of 0.84 is the closest to 1. Further confirmation of this is how the data points are packed closer to the trendline in Scenarios 3 than the other scenarios.

Table 5
Bedford correlation matrix

Solar Radiation	
Solar Radiation	1
Hourly Pressure at Mean Sea Level	0.1237
Hourly Dewpoint Temperature	0.1796
Hourly Rainfall Total	-0.0567
Hourly Relative Humidity	-0.6463
Hourly Mean Wind Direction	-0.0470
Hourly Mean Windspeed	0.1275
Hourly Maximum Gust	0.1688
Hourly Temperature	**0.5383**
Hourly Wet Bulb Temperature	**0.4067**
Hourly Global Radiation	**0.9367**
Hourly Total Cloud Cover	-0.1631
Hourly Visibility	0.1779

Table 8 and Figure 7 shows the model's performance in learning how to predict the PV generation. To investigate its forecasting ability further, we evaluate the prediction of a randomly selected day and determine how close it is to the actual PV generated. Figure 8 presents a comparison of the actual and the predicted amount of energy supplied from the PV under Scenarios 1 to 3. It is clear to see that Scenario 3 (Figure 8c) provides the best approximation of the actual PV generated than any of the other 2 scenarios. This is because the PV energy supplied in Scenario 1 (Figure 8a) appears to be underestimated between 0 and 12 hrs, and the afternoon peak between 16-18 hrs is also not represented. In Scenario 2 (Figure 8b), the night and morning energy supply (0-12 hrs) are relatively well represented. However, there is a substantial dip in the amount of energy predicted between 12 and 13 hrs. There is evidence that the model predicts a small energy supply peak between 16-18 hrs. In Scenario 3, the prediction for the night and morning energy supply are in good agreement with the actual data. There is a small dip in the forecasted energy supply between 12 and 13 hrs, however, it is of a much smaller magnitude than that observed in Scenario 2. The afternoon peak in energy supply is small.

Beyond visually comparing the performance of the scenarios in Figure 8, further insight can be provided by estimating the area under the predicted curve of each scenario and comparing it to the area under the actual curve. The scenario that has the closest approximation of its predicted curve to the actual curve has the better performance and accuracy. This area is simply the amount of energy supplied (normalized) per day. For a simple approximation of this area, the Trapezoidal Rule is used. This rule approximates the area under the graph by dividing it into smaller trapezoids. The total area under the curve is equal to the sum of the area of the trapezoids under the graph. The Trapezoidal Rule is shown in Eq. (7) [86].

$$\int_a^b f(x) \approx \left(\frac{b-a}{2}\right)(f(a)+f(b)) \tag{7}$$

In Eq. (7), limits a and b refer to the hours on the x axis in Figure 8. This represents the width of the trapezoid. Values *f(a)* and *f(b)* represent the corresponding energy supplied (normalized) at hour 'a' and hour 'b' respectively. Eq. (7) was applied to the three scenarios (See Appendix B for calculation) to compare their approximation of the actual energy supplied. The results show that the area under the curve of the actual energy supplied is approximately 4.969. Ranking the scenario's based on the closest approximation to the actual area, Scenario 3 is closest with an area of 5.115, followed by Scenario 2 with 4.424 and Scenario 1 with 2.906. This result and the order of performance validates the test result in Table 8 in which Scenario 3 (0.0859) outperforms the other scenarios, and Scenario 2 (0.0987) outperforms Scenario 1 (0.0917).

A further proof of the versatility of the proposed forecasting architecture can be made with a longer forecast. Section 1 references the different types of forecasting time horizons. Figure 9 shows how the proposed forecasting architecture of Scenario 3 is applied to a day-ahead short-term forecast. Here, a random week in 2019 is selected to show the performance of the forecasting architecture when applied to a medium-term forecast. The graph shows a good enough approximation, which means that the proposed architecture is flexible and can be used for predicting longer time horizons.

A closer look at Figure 9 shows some lagging of the predicted to the actual curve, along with a few misplaced approximations of PV peaks

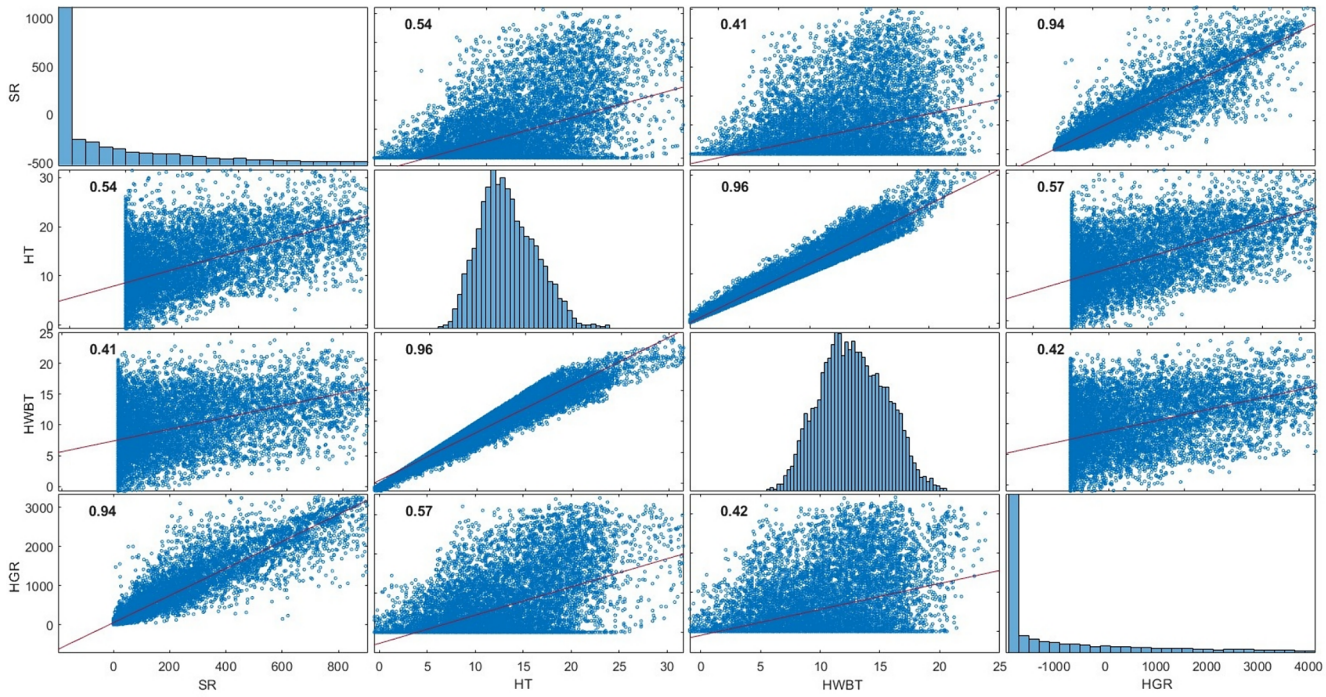


Fig. 5. Bedford correlation plot Note: SR = Solar radiation, HT = Hourly Temperature, HWBT = Hourly Wet bulb temperature, HGR = Hourly Global radiation

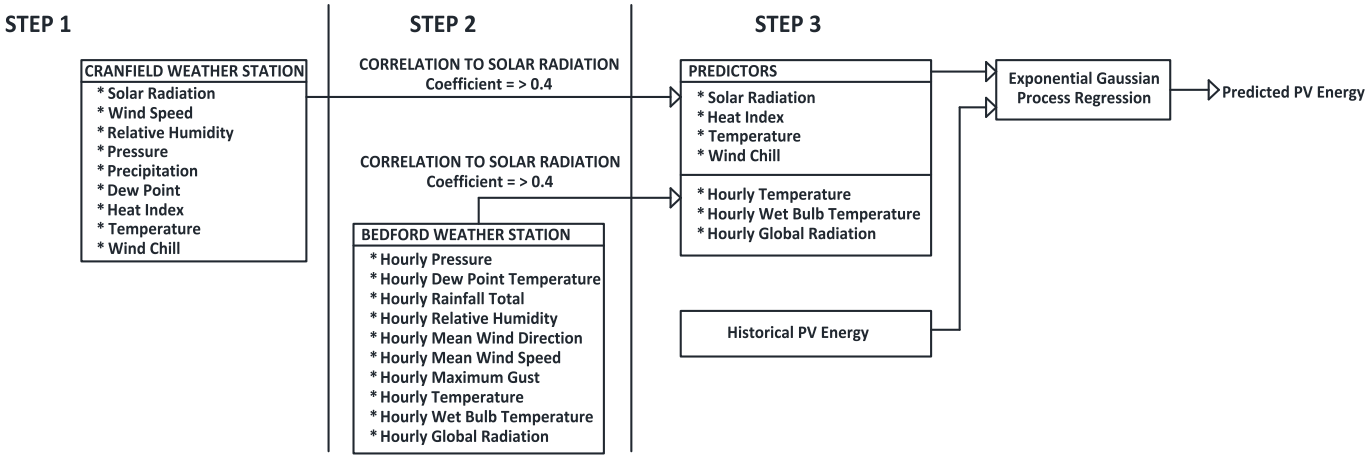


Fig. 6. Case Study : Three-step forecasting architecture block diagram

Table 6
RLA model training input parameters

Training Input	Scenario 1	Scenario 2	Scenario 3
Predictors	*Solar Radiation *Heat Index *Temperature *Wind Chill	*Hourly Temperature *Hourly Wet bulb Temperature *Hourly Global Radiation	*Solar Radiation *Heat Index *Temperature *Wind Chill *Hourly Temperature *Hourly Wet bulb Temperature *Hourly Global Radiation
Response	PV energy generated (all scenarios)		

(between days 1 and 29). A possible explanation is the effect of noise in the pre-processed data. Improving data pre-processing, for example by identifying and removing outliers, may lead to a closer approximation of the actual PV generated. In addition, Scenario 3 outperformed the other scenarios because it had more data points (by having more variables). Perhaps this bias could be investigated if the analysis required equal number of data points for each scenario.

Validation of PPC with SR and LLDF methods

To further investigate the effect of PCC and the selection of SR as a reference parameter, we consider a correlation matrix with reference to PV generated. According to Table 9, SR and global radiation have the highest correlation to PV generation confirming literature findings that SR is the most important predictor to PV. Apart from radiation, the hourly temperature (0.4497) has the highest positive correlation to PV generation and will be tested as a substitute to SR in predictor selection.

Four additional scenarios are considered using the same Boosted

Table 7
Performance comparison of regression models in RLA For Scenario 3

Model	Train MAE	MSE	RMSE	R^2	Test MAE	MSE	RMSE	R^2
LINEAR REGRESSION								
Linear	0.0724	0.0124	0.1114	0.81	0.0975	0.0138	0.1176	0.57
Interactions Linear	0.0645	0.0109	0.1044	0.83	0.0904	0.0167	0.1293	0.49
Robust Linear	0.0741	0.0191	0.1381	0.71	0.0489	0.0090	0.0950	0.72
Stepwise Linear	0.0646	0.0109	0.1045	0.83	0.0889	0.0162	0.1273	0.50
REGRESSION TREES								
Fine Tree	0.0339	0.0044	0.0664	0.93	0.0495	0.0116	0.1079	0.64
Medium Tree	0.0474	0.0081	0.0901	0.88	0.0451	0.0098	0.0992	0.70
Coarse Tree	0.0560	0.0104	0.1019	0.84	0.0432	0.0087	0.0932	0.73
SUPPORT VECTOR MACHINES								
Linear SVM	0.0676	0.0129	0.1135	0.80	0.0707	0.0097	0.0985	0.70
Quadratic SVM	0.0598	0.0111	0.1054	0.83	0.0871	0.0178	0.1335	0.45
Cubic SVM	0.0573	0.0010	0.0999	0.85	0.0914	0.0192	0.1384	0.41
Fine Gaussian	0.0433	0.0068	0.0828	0.90	0.0914	0.0137	0.1171	0.58
Medium Gaussian	0.0561	0.0098	0.0992	0.85	0.0608	0.0102	0.1008	0.69
Coarse Gaussian	0.0644	0.0119	0.1092	0.82	0.0504	0.0074	0.0862	0.77
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	0.0007	3.05e-06	0.0017	1.00	0.0948	0.0213	0.1459	0.34
Squared Exponential	0.0013	8.71e-06	0.0030	1.00	0.1411	0.0447	0.2115	-0.38
Matern 5/2	0.0004	9.30e-07	0.0010	1.00	0.0966	0.2194	0.1481	0.33
Exponential	5.41e-05	1.20e-08	0.0001	1.00	0.0639	0.0111	0.1054	0.66
ENSEMBLES OF TREES								
Boosted Trees	0.0571	0.0109	0.1042	0.82	0.0448	0.0074	0.0859	0.84
Bagged Trees	0.0413	0.0059	0.0765	0.91	0.0436	0.0082	0.0906	0.75
NEURAL NETWORK								
Narrow Neural	0.0561	0.0010	0.0995	0.85	0.0803	0.0146	0.1209	0.55
Medium Neural	0.0535	0.0086	0.0928	0.87	0.1109	0.0206	0.1436	0.37
Wide Neural	0.0491	0.0071	0.0841	0.89	0.1284	0.0295	0.1718	0.09
Bilayered Neural	0.0528	0.0091	0.0956	0.86	0.0631	0.0107	0.1036	0.67
Trilayered Neural	0.0491	0.0081	0.0899	0.88	0.0940	0.0235	0.1534	0.28

Table 8
Performance comparison of case study scenarios using Boosted Trees

Scenario	Test MAE	MSE	RMSE	R^2	Train MAE	MSE	RMSE	R^2
Scenario 1	0.0615	0.0118	0.1087	0.82	0.0441	0.0084	0.0917	0.74
Scenario 2	0.0663	0.0128	0.1133	0.80	0.0465	0.0097	0.0987	0.70
Scenario 3	0.0571	0.0109	0.1042	0.82	0.0448	0.0074	0.0859	0.84

Tree algorithm in Table 8, with the full results shown in Appendix A, Table A4 to A6. The summary of the results is displayed in Table 10:

a. Scenario 4: Weather data without PCC from the Cranfield weather station selected as predictors to PV generated.
b. Scenario 5: Weather data without PCC from the Bedford weather station selected as predictors to PV generated.
c. Scenario 6: The weather data from Scenarios 4 and 5 are combined using LLDF and are used as predictors to PV generated.
d. Scenario 7: Temperature-correlated weather data from both stations are combined using LLDF and used as predictors to PV generated.

Table 10 shows that Scenario 3 remains the best performing model with a combined use of LLDF and PCC. Scenarios 1 to 3 outperform Scenario 7, confirming SR is better as a reference parameter in features than temperature. Additionally, the importance of PCC for feature selection is validated as Scenarios 4 to 6 (without PCC) perform relatively poorly to Scenarios 1 to 3 (with PCC). Out of the scenarios without PPC, Scenario 6 with LLDF performs the best, confirming the positive impact of LLDF on forecasting accuracy.

Validation of forecasting architecture

To validate the model, we apply the forecasting architecture in predicting PV generation of the three rooftop solar arrays described in Section 3.1. The results are displayed in Table 11.

In Table 11, the results show that the proposed architecture maintains its accuracy in predicting PV generation as evident by the low RMSE values. Another important fact is that the proposed architecture made the predictions across three different PV system sizes and data time periods. By confirming its accuracy for both commercial and residential applications, the models can be universally adopted for different distributed generation clients using different historical weather data as inputs into the model. However, the R squared values shown here (0.52 to 0.68) indicate that the goodness of fit to the model can be improved. Higher R squared values closer to 1, suggest that the movement of the predicted PV values is dependent and explained by the movement of the actual PV values. In other words, though the predicted values in Table 11 sufficiently approximate the actual values, they could be better. As mentioned above, including other pre-processing techniques could help improve the accuracy.

Conclusion

The proposed three-step forecasting architecture that uses low-level data fusion to combine weather parameters from both an on-site and a nearby local weather station outperforms models using a singular data source. Pearson Correlation Coefficient was used to identify weather variables with moderate to strong positive correlations to solar radiation. Combining useful data from the two weather stations increased the number of strong predictors in the forecasting model and improved its performance. In addition, raw meteorological data was pre-processed

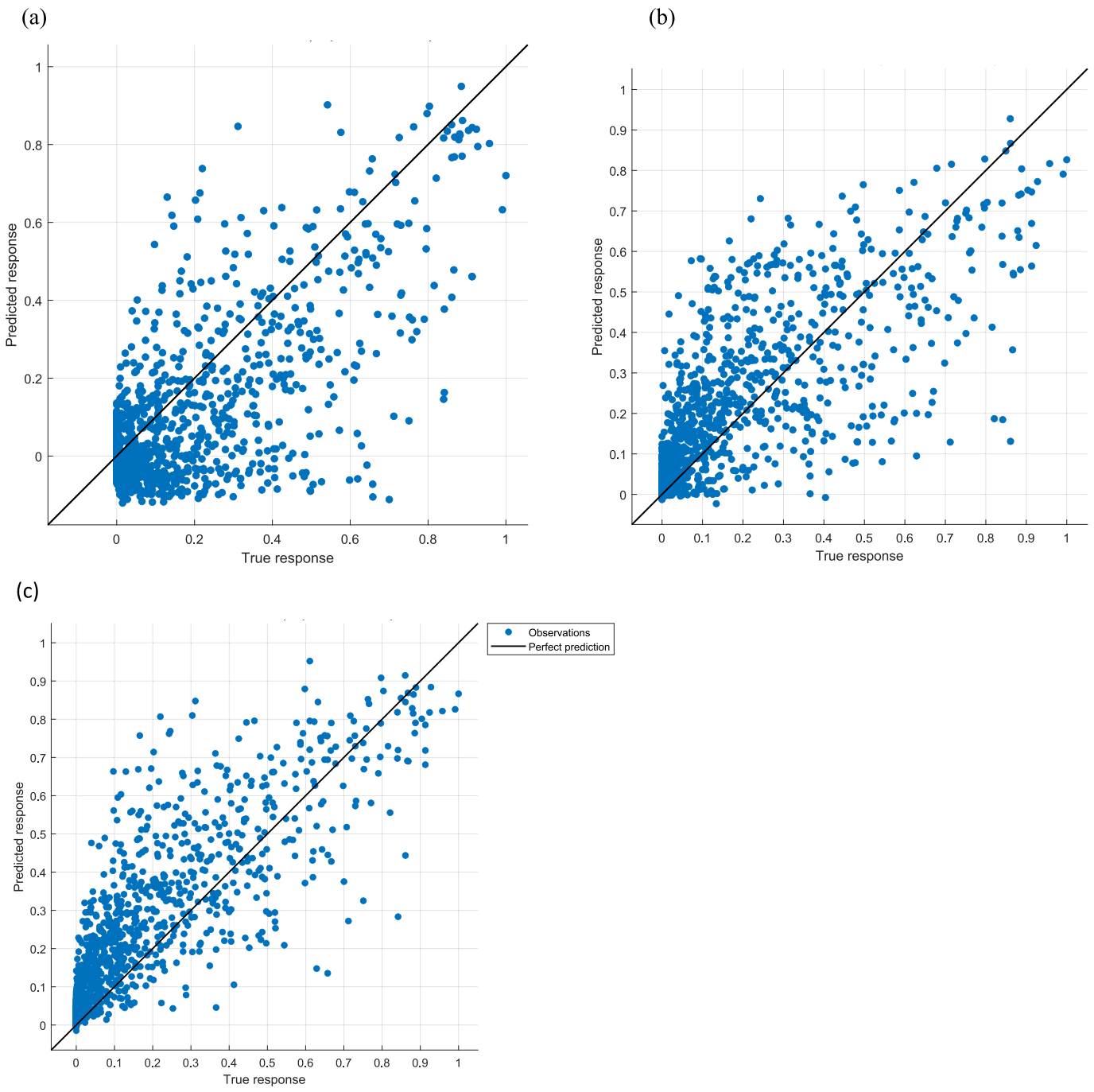


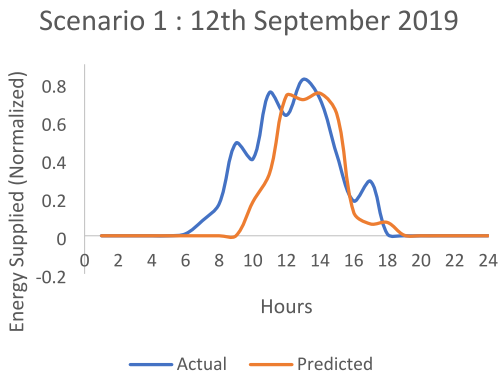
Fig. 7. Model performance of Boosted Trees ((a) Scenario 1 (b) Scenario 2 (c) Scenario 3)

using the Rescaling and List-wise deletion methods, which proved to be effective in increasing the accuracy of the model. The proposed architecture is tested on a utility scale solar plant (1 MW), showing a 6% and 13% prediction accuracy improvement when compared to solely using data from the on-site and local weather station respectively. It is further validated using data from three residential rooftop solar systems (8 kW, 10.5 kW and 15 kW), achieving root-mean square values of 0.0984, 0.0885, and 0.1425 respectively. The low root-mean square values, a measure of the predicted PV to actual PV generation, proves that the model can be adopted for different PV plant sizes. In determining the reference parameter for selecting forecasting features, solar radiation achieved an RMSE of 0.0859 compared to temperature with 0.1132 when the Boosted Tree algorithm was used. Solar radiation had the highest correlation to PV generation, confirming it as the most important feature for forecasting solar energy. Finally, in an analysis of three scenarios without the use of Pearson Correlation Coefficient in selecting features, combining weather data from Cranfield and Bedford using low-level data fusion outperformed data solely from Cranfield by 19%, and data from Bedford by 9%. In scenarios with or without PCC, low-level data fusion significantly improved the forecasting.

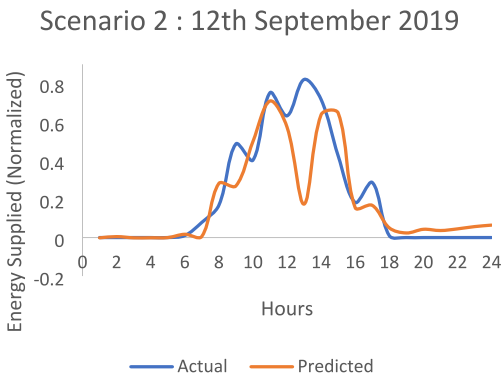
The impact of this study for utility-scale PV plants, solar farms, and other large PV installations could be to encourage grid owners and operators to install on-site weather stations for forecasting purposes. It is best to pair data from a local weather station with data from an on-site weather station where the forecasting system is being deployed.

The limitation of the forecasting architecture is that it only works if the local weather station and the on-site PV plant measure the reference weather parameter, which in this case is solar radiation. In the absence

(a)



(b)



(c)

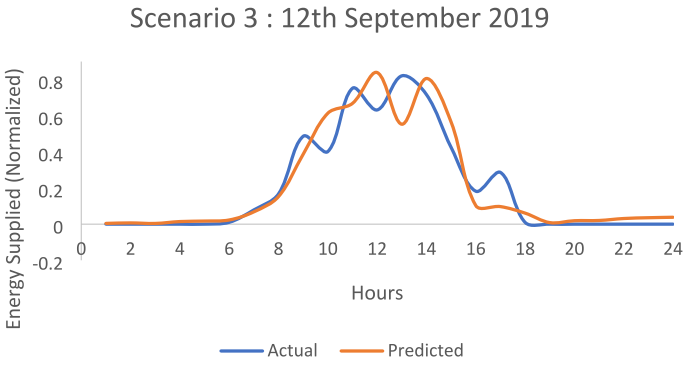


Fig. 8. Actual and Predicted PV Generation on 12th September 2019 ((a) Scenario 1 (b) Scenario 2 (c) Scenario 3)

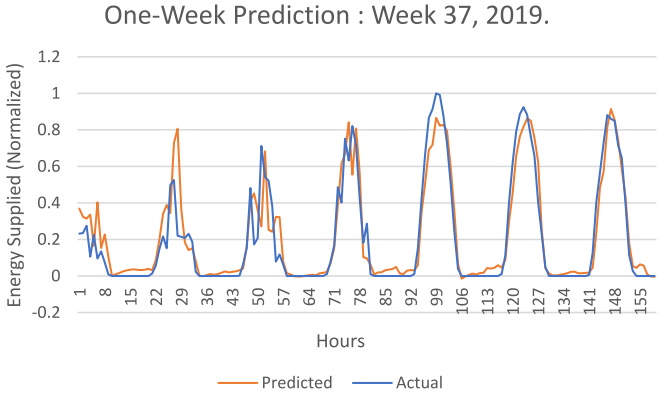


Fig. 9. Actual and Predicted PV Generation for Scenario 3 on the 37th Week in 2019.

of this, an alternate reference parameter with strong correlation to PV generation needs to be determined. Additionally, the scope of the study was limited to data within the UK. To test the robustness of the forecasting architecture, other PV plant in different locations outside the UK could be considered. In addition, only weather parameters with strong positive correlations to the reference parameter were considered. In a future study, the impact of weather parameters with strong negative correlations should also be analyzed. Furthermore, the optimization of the Regression learner App methodology can be investigated by exploring the impact of other validation methods like Cross-validation and Holdout validation to see if they directly improve the accuracy of the prediction model. Finally, an investigation into other pre-processing methods may help reduce the noise in the final output while future studies may consider this approach for forecasting other weather dependent renewable energy sources like wind and tidal energy.

Table 9
Bedford and Cranfield correlation matrix to PV Generation

	PV Generation
PV Generation	1
Solar Radiation	0.8668
Hourly Pressure at Mean Sea Level	0.1152
Hourly Dewpoint Temperature	0.1266
Hourly Rainfall Total	-0.0511
Hourly Relative Humidity	-0.5119
Hourly Mean Wind Direction	-0.0940
Hourly Mean Windspeed	0.0534
Hourly Maximum Gust	0.0741
Hourly Temperature	0.4497
Hourly Wet Bulb Temperature	0.3545
Hourly Global Radiation	0.8569
Hourly Total Cloud Cover	-0.2104
Hourly Visibility	0.0524
Wind Speed	0.0690
Relative Humidity	-0.5158
Pressure	0.1206
Precipitation	-0.0534
Dew Point	0.1253
Heat Index	0.4129
Temperature	0.4241
Wind Chill	0.4087

Table 10
Performance comparison of case study scenarios using Boosted Trees

Scenario	Test MAE	MSE	RMSE	R^2	Train MAE	MSE	RMSE	R^2
Scenario 1	0.0615	0.0118	0.1087	0.82	0.0441	0.0084	0.0917	0.74
Scenario 2	0.0663	0.0128	0.1133	0.80	0.0465	0.0097	0.0987	0.70
Scenario 3	0.0571	0.0109	0.1042	0.82	0.0448	0.0074	0.0859	0.84
Scenario 4	0.0571	0.0104	0.1018	0.86	0.0689	0.0160	0.1266	0.78
Scenario 5	0.0561	0.0101	0.1003	0.87	0.0623	0.0136	0.1165	0.81
Scenario 6	0.0535	0.0093	0.0963	0.88	0.0588	0.0113	0.1065	0.84
Scenario 7	0.0607	0.0112	0.1057	0.85	0.0609	0.0128	0.1132	0.82

Table 11
Validation results of proposed architecture using residential solar arrays

BUILDINGS	MODEL	RMSE	MSE	MAE	R^2
B146	Boosted Trees	0.0885	0.0078	0.0499	0.68
B252	**MODEL**	**RMSE**	**MSE**	**MAE**	**R^2**
	Boosted Trees	0.0984	0.0097	0.0547	0.67
B052A	**MODEL**	**RMSE**	**MSE**	**MAE**	**R^2**
	Boosted Trees	0.1425	0.0203	0.0811	0.52

Funding

This research was supported by the Petroleum Technology Development Fund (PTDF) [PTDF/ED/OSS/PHD/TOF/1945/20].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A

.

Table A1
Performance Comparison of Regression Models in RLA For Scenario 1

Model	Train MAE	MSE	RMSE	R^2	Test MAE	MSE	RMSE	R^2
LINEAR REGRESSION								
Linear	0.0772	0.0155	0.1244	0.76	0.0562	0.0093	0.0967	0.71
Interactions Linear	0.0722	0.0143	0.1197	0.78	0.0955	0.0145	0.1206	0.55
Robust Linear	0.0735	0.0185	0.1360	0.72	0.0688	0.0123	0.1110	0.62
Stepwise Linear	0.0721	0.0143	0.1198	0.78	0.0920	0.0141	0.1188	0.57
REGRESSION TREES								
Fine Tree	0.0412	0.0062	0.0787	0.91	0.0486	0.0120	0.1095	0.63
Medium Tree	0.0541	0.0098	0.0990	0.85	0.0458	0.0100	0.1002	0.69
Coarse Tree	0.0599	0.0117	0.1082	0.82	0.0442	0.0091	0.0955	0.72
SUPPORT VECTOR MACHINES								
Linear SVM	0.0711	0.0158	0.1256	0.76	0.0673	0.0109	0.1046	0.66
Quadratic SVM	0.0678	0.0138	0.1174	0.79	0.0870	0.0149	0.1221	0.54
Cubic SVM	0.1181	0.0279	0.1671	0.57	0.1294	0.0469	0.2165	-0.44
Fine Gaussian	0.0603	0.0113	0.1065	0.83	0.0744	0.0134	0.1157	0.59
Medium Gaussian	0.0648	0.0128	0.1133	0.80	0.0672	0.0114	0.1070	0.65
Coarse Gaussian	0.0692	0.0140	0.1185	0.79	0.0478	0.0081	0.0901	0.75
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	0.0068	0.0002	0.0147	1.00	0.0613	0.0124	0.1115	0.62
Squared Exponential	0.0625	0.0117	0.108	0.82	0.1008	0.0239	0.1547	0.26
Matern 5/2	0.0608	0.0112	0.1057	0.83	0.0997	0.0230	0.1518	0.29
Exponential	3.03e-4	1.22e-6	0.0011	1.00	0.0744	0.0166	0.1288	0.49
ENSEMBLES OF TREES								
Boosted Trees	0.0615	0.0118	0.1087	0.82	0.0441	0.0084	0.0917	0.74
Bagged Trees	0.0502	0.0081	0.0898	0.88	0.0531	0.0093	0.0964	0.71
NEURAL NETWORK								
Narrow Neural	0.0643	0.0124	0.1112	0.81	0.0780	0.0128	0.1133	0.61
Medium Neural	0.0616	0.0116	0.1078	0.82	0.1031	0.0234	0.1529	0.28
Wide Neural	0.0596	0.0109	0.1045	0.83	0.1224	0.0327	0.1808	-0.01
Bilayered Neural	0.0621	0.0119	0.1089	0.82	0.0556	0.0108	0.1040	0.67
Trilayered Neural	0.0611	0.0115	0.1070	0.82	0.0957	0.0202	0.1422	0.38

Table A2
Performance Comparison of Regression Models in RLA For Scenario 2

	Train				Test			
Model	**MAE**	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
LINEAR REGRESSION								
Linear	0.0780	0.0147	0.1213	0.78	0.0607	0.0010	0.0100	0.69
Interactions Linear	0.0739	0.0141	0.1189	0.78	0.0559	0.0097	0.0984	0.70
Robust Linear	0.0740	0.0167	0.1302	0.74	0.0465	0.0010	0.0998	0.69
Stepwise Linear	0.0739	0.0141	0.1189	0.78	0.0559	0.0097	0.0984	0.70
REGRESSION TREES								
Fine Tree	0.0429	0.0066	0.0811	0.90	0.0556	0.0145	0.1204	0.55
Medium Tree	0.0562	0.0103	0.1017	0.84	0.0505	0.0119	0.1092	0.63
Coarse Tree	0.0638	0.0126	0.1122	0.81	0.0483	0.0108	0.1037	0.67
SUPPORT VECTOR MACHINES								
Linear SVM	0.0737	0.0150	0..1226	0.77	0.0518	0.0097	0.0985	0.70
Quadratic SVM	0.0713	0.0145	0.1203	0.78	0.0500	0.0095	0.0975	0.71
Cubic SVM	0.0660	0.0133	0.1155	0.80	0.0487	0.0099	0.0997	0.69
Fine Gaussian	0.0611	0.0113	0.1064	0.83	0.0559	0.0108	0.1038	0.67
Medium Gaussian	0.0664	0.0129	0.1136	0.80	0.0497	0.0098	0.0991	0.70
Coarse Gaussian	0.0717	0.0142	0.1192	0.78	0.0503	0.0092	0.0999	0.72
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	0.0020	2.577e-05	0.0051	1.00	0.0597	0.0146	0.1209	0.55
Squared Exponential	0.0639	0.0123	0.1110	0.81	0.0495	0.0100	0.1001	0.69
Matern 5/2	0.0576	0.0104	0.1018	0.84	0.0498	0.0103	0.1015	0.68
Exponential	3.20e-4	3.81e-6	0.0020	1.00	0.0530	0.0118	0.1088	0.64
ENSEMBLES OF TREES								
Boosted Trees	0.0663	0.0128	0.1133	0.80	0.0465	0.0097	0.0987	0.70
Bagged Trees	0.0573	0.0094	0.0968	0.86	0.0522	0.0101	0.1007	0.69
NEURAL NETWORK								
Narrow Neural	0.0652	0.0128	0.1130	0.80	0.0489	0.0097	0.0985	0.70
Medium Neural	0.0643	0.0124	0.1113	0.81	0.0517	0/0103	0.1015	0.68
Wide Neural	0.0617	0.0115	0.1073	0.82	0.0523	0.0106	0.1031	0.67
Bilayered Neural	0.0630	0.0122	0.1105	0.81	0.0496	0.0105	0.1023	0.68
Trilayered Neural	0.0626	0.0120	0.1096	0.82	0.0507	0.0106	0.1030	0.67

Table A3
Performance comparison of regression models in RLA For Scenario 4

	Train				Test			
Model	**MAE**	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
LINEAR REGRESSION								
Linear	0.0762	0.0160	0.1267	0.79	0.0812	0.0194	0.1394	0.73
Interactions Linear	0.0682	0.0137	0.1171	0.82	0.0750	0.0167	0.1292	0.77
Robust Linear	0.0699	0.0175	0.1321	0.77	0.0772	0.0219	0.1481	0.70
Stepwise Linear	0.0688	0.0134	0.1178	0.82	0.0751	0.0168	0.1294	0.77
REGRESSION TREES								
Fine Tree	0.0373	0.0057	0.0755	0.92	0.0781	0.0237	0.1539	0.67
Medium Tree	0.0506	0.0096	0.0980	0.87	0.0734	0.0196	0.1398	0.73
Coarse Tree	0.0603	0.0128	0.1130	0.83	0.0718	0.0184	0.1355	0.75
SUPPORT VECTOR MACHINES								
Linear SVM	0.0712	0.0164	0.1280	0.78	0.0764	0.0200	0.1416	0.72
Quadratic SVM	0.0657	0.0144	0.1201	0.81	0.0721	0.0177	0.1330	0.76
Cubic SVM	0.0593	0.0127	0.1127	0.83	0.0690	0.0176	0.1326	0.76
Fine Gaussian	0.0401	0.0063	0.0791	0.92	0.0837	0.0205	0.1431	0.72
Medium Gaussian	0.0594	0.0123	0.1110	0.84	0.0693	0.0174	0.1321	0.76
Coarse Gaussian	0.0707	0.0155	0.1244	0.80	0.0763	0.0187	0.1367	0.74
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	0.0002	1.33e-07	0.0004	1.00	0.0718	0.0177	0.1329	0.76
Squared Exponential	0.0536	0.0089	0.0943	0.88	0.0694	0.0156	0.1249	0.79
Matern 5/2	0.0397	0.0050	0.0707	0.93	0.0704	0.0164	0.1279	0.77
Exponential	5.57e-05	9.73e-09	9.86e-05	1.00	0.0686	0.0162	0.1274	0.78
ENSEMBLES OF TREES								
Boosted Trees	0.0571	0.0104	0.1018	0.86	0.0689	0.0160	0.1266	0.78
Bagged Trees	0.0427	0.0068	0.0824	0.91	0.0665	0.0162	0.1273	0.78
NEURAL NETWORK								
Narrow Neural	0.0628	0.0116	0.1076	0.85	0.0736	0.0168	0.1296	0.77
Medium Neural	0.0638	0.0104	0.1021	0.86	0.0794	0.0164	0.1282	0.77
Wide Neural	0.0524	0.0066	0.0815	0.91	0.0867	0.0208	0.1443	0.71
Bilayered Neural	0.0560	0.0100	0.1003	0.87	0.0734	0.0176	0.1328	0.76
Trilayered Neural	0.0551	0.0093	0.0963	0.88	0.0758	0.0189	0.1374	0.74

Table A4
Performance comparison of regression models in RLA For Scenario 5

	Train				Test			
Model	**MAE**	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
LINEAR REGRESSION								
Linear	0.0772	0.0154	0.1241	0.80	0.0756	0.0162	0.1274	0.77
Interactions Linear	0.0671	0.0123	0.1110	0.84	0.0683	0.0142	0.1191	0.80
Robust Linear	0.0732	0.0164	0.1282	0.79	0.0735	0.0180	0.1342	0.74
Stepwise Linear	0.0681	0.0128	0.1131	0.83	0.0682	0.0142	0.1193	0.80
REGRESSION TREES								
Fine Tree	0.0340	0.0049	0.0698	0.94	0.0654	0.0168	0.1296	0.76
Medium Tree	0.0469	0.0087	0.0932	0.89	0.0624	0.0157	0.1252	0.78
Coarse Tree	0.0574	0.0120	0.1096	0.84	0.0646	0.0163	0.1278	0.77
SUPPORT VECTOR MACHINES								
Linear SVM	0.0744	0.0160	0.1263	0.79	0.0734	0.0172	0.1312	0.76
Quadratic SVM	0.0661	0.0135	0.1162	0.82	0.0656	0.0146	0.1207	0.79
Cubic SVM	0.0562	0.0108	0.1039	0.86	0.0607	0.0133	0.1154	0.81
Fine Gaussian	0.0437	0.0074	0.0862	0.90	0.1019	0.0241	0.1553	0.66
Medium Gaussian	0.0565	0.0108	0.1040	0.86	0.0614	0.0134	0.1156	0.81
Coarse Gaussian	0.0740	0.0154	0.1240	0.80	0.0730	0.0164	0.1282	0.77
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	6.61e-05	1.49e-08	0.0001	1.00	0.0570	0.0104	0.1019	0.85
Squared Exponential	0.0577	0.0097	0.0984	0.87	0.0614	0.0121	0.1098	0.83
Matern 5/2	0.0456	0.0062	0.0789	0.92	0.0573	0.0106	0.1028	0.85
Exponential	4.85e-05	7.01e-09	8.37e-05	1.00	0.0557	0.0102	0.1009	0.86
ENSEMBLES OF TREES								
Boosted Trees	0.0561	0.0101	0.1003	0.87	0.0623	0.0136	0.1165	0.81
Bagged Trees	00396	0.0058	0.0761	0.92	0.0571	0.0128	0.1130	0.82
NEURAL NETWORK								
Narrow Neural	0.0616	0.0107	0.1035	0.86	0.0658	0.0136	0.1167	0.81
Medium Neural	0.0597	0.0088	0.0938	0.89	0.0696	0.0129	0.1136	0.82
Wide Neural	0.0459	0.0043	0.0656	0.94	0.0822	0.0149	0.1221	0.79
Bilayered Neural	0.0532	0.0091	0.0956	0.88	0.0612	0.0134	0.1155	0.81
Trilayered Neural	0.0517	0.0089	0.0942	0.88	0.0605	0.0138	0.1173	0.81

Table A5
Performance comparison of regression models in RLA For Scenario 6

	Train				Test			
Model	**MAE**	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
LINEAR REGRESSION								
Linear	0.0722	0.0137	0.1171	0.82	0.0731	0.0141	0.1187	0.80
Interactions Linear	0.0607	0.0101	0.1003	0.87	0.0698	0.0163	0.1277	0.77
Robust Linear	0.0702	0.0178	0.1333	0.77	0.0723	0.0194	0.1392	0.73
Stepwise Linear	0.0618	0.0110	0.1049	0.86	0.0634	0.0111	0.1054	0.85
REGRESSION TREES								
Fine Tree	0.0303	0.0039	0.0627	0.95	0.0717	0.0191	0.1382	0.73
Medium Tree	0.0466	0.0085	0.0920	0.89	0.0612	0.0145	0.1205	0.80
Coarse Tree	0.0568	0.0118	0.1086	0.85	0.0638	0.0149	0.1222	0.79
SUPPORT VECTOR MACHINES								
Linear SVM	0.0684	0.0145	0.1204	0.81	0.0694	0.0152	0.1234	0.79
Quadratic SVM	0.0572	0.0115	0.1072	0.85	0.0604	0.0121	0.1100	0.83
Cubic SVM	0.0472	0.0086	0.0927	0.89	0.0560	0.0010	0.0987	0.86
Fine Gaussian	0.0444	0.0079	0.0886	0.90	0.1133	0.0295	0.1718	0.59
Medium Gaussian	0.0511	0.0097	0.0986	0.87	0.0566	0.0107	0.1033	0.85
Coarse Gaussian	0.0670	0.0136	0.1166	0.82	0.0676	0.0139	0.1178	0.81
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	8.36e-05	1.87e-08	0.0001	1.00	0.0578	0.0104	0.1018	0.86
Squared Exponential	0.0001	5.15e-08	0.0002	1.00	0.0683	0.0132	0.1150	0.82
Matern 5/2	8.66e-05	2.00e-08	0.0001	1.00	0.0599	0.0110	0.1048	0.85
Exponential								
ENSEMBLES OF TREES								
Boosted Trees	0.0535	0.0093	0.0963	0.88	0.0588	0.0113	0.1065	0.84
Bagged Trees	0.0358	0.0052	0.0721	0.93	0.0554	0.0114	0.1070	0.84
NEURAL NETWORK								
Narrow Neural	0.0560	0.0105	0.1023	0.86	0.0612	0.0111	0.1054	0.85
Medium Neural	0.0553	0.0074	0.0861	0.90	0.0717	00122	0.1104	0.83
Wide Neural	0.0373	0.0026	0.0514	0.97	0.0928	0.0182	0.1349	0.75
Bilayered Neural	0.0479	0.0077	0.0875	0.90	0.0610	0.0132	0.1151	0.82
Trilayered Neural	0.0481	0.0078	0.0885	0.90	0.0612	0.0125	0.1118	0.83

Table A6
Performance comparison of regression models in RLA For Scenario 7

Model	Train MAE	MSE	RMSE	R^2	Test MAE	MSE	RMSE	R^2
LINEAR REGRESSION								
Linear	0.0770	0.0149	0.1220	0.80	0.0740	0.0144	0.1199	0.80
Interactions Linear	0.0663	0.0125	0.1119	0.84	0.0648	0.0130	0.1141	0.82
Robust Linear	0.0757	0.0207	0.1440	0.73	0.0728	0.0198	0.1406	0.73
Stepwise Linear	0.0669	0.0129	0.1137	0.83	0.0635	0.0126	0.1122	0.83
REGRESSION TREES								
Fine Tree	0.0352	0.0053	0.0727	0.93	0.0176	0.0661	0.1328	0.6
Medium Tree	0.0503	0.0098	0.0991	0.87	0.0655	0.0184	0.1355	0.75
Coarse Tree	0.0624	0.0134	0.1159	0.82	0.0632	0.0154	0.1243	0.79
SUPPORT VECTOR MACHINES								
Linear SVM	0.0725	0.0158	0.1255	0.79	0.0696	0.0150	0.1224	0.79
Quadratic SVM	0.0628	0.0134	0.1157	0.82	0.0616	0.0135	0.1164	0.81
Cubic SVM	0.0588	0.0115	0.1070	0.85	0.0607	0.0134	0.1156	0.81
Fine Gaussian	0.0376	0.0056	0.0749	0.93	0.0637	0.0138	0.1174	0.81
Medium Gaussian	0.0569	0.0116	0.1076	0.85	0.0581	0.0122	0.1104	0.83
Coarse Gaussian	0.0696	0.0148	0.1216	0.81	0.0665	0.0143	0.1194	0.80
GAUSSIAN PROCESS REGRESSION								
Rational Quadratic	0.0047	0.0001	0.0110	1.00	0.0580	0.0129	0.1134	0.82
Squared Exponential	0.0063	0.0002	0.0141	1.00	0.0701	0.0171	0.1309	0.76
Matern 5/2	0.0031	6.78e-05	0.0082	1.00	0.0622	0.0144	0.1198	0.80
Exponential	4.83e-05	8.75e-09	9.35e-05	1.00	0.0541	0.0117	0.1081	0.84
ENSEMBLES OF TREES								
Boosted Trees	0.0607	0.0112	0.1057	0.85	0.0609	0.0128	0.1132	0.82
Bagged Trees	0.0421	0.0067	0.0816	0.91	0.0560	0.0126	0.1122	0.83
NEURAL NETWORK								
Narrow Neural	0.0590	0.0112	0.1061	0.85	0.0584	0.0122	0.1104	0.83
Medium Neural	0.0579	0.0096	0.0982	0.87	0.0637	0.0135	0.1162	0.81
Wide Neural	0.0473	0.0059	0.0769	0.92	0.0702	0.0154	0.1243	0.79
Bilayered Neural	0.0521	0.0092	0.0960	0.88	0.0611	0.0137	0.1171	0.81
Trilayered Neural	0.0511	0.0089	0.0941	0.88	0.0613	0.0137	0.1173	0.81

Appendix B

.

Table B1
Trapezoidal Rule Calculation of Areas Under the Curves of Figure 8

Hours	Actual	Scenario 1	Scenario 2	Scenario 3
0	0	-0.06196	0.001677	0.005502
1	0	-0.05747	0.001677	0.00539
2	0	-0.0543	-0.00131	0.009166
3	0	-0.05265	-0.00066	0.01572
4	0.005156	-0.05272	0.008427	0.020084
5	0.044569	-0.06125	0.011387	0.045674
6	0.123169	-0.08926	0.143359	0.110453
7	0.326946	-0.09722	0.274514	0.269908
8	0.444135	0.045696	0.383478	0.500645
9	0.577094	0.252907	0.604034	0.641404
10	0.692264	0.532067	0.641365	0.753875
11	0.726599	0.724883	0.373679	0.696112
12	0.7689	0.731178	0.406016	0.679695
13	0.570539	0.692602	0.642644	0.685248
14	0.303218	0.380112	0.400898	0.333426
15	0.234184	0.092457	0.161005	0.099877
16	0.147901	0.065957	0.108772	0.079057
17	0.004814	0.036532	0.037098	0.034738
18	0	-0.00446	0.033382	0.013739
19	0	-0.01649	0.03967	0.019668
20	0	-0.02783	0.040755	0.025797
21	0	-0.03587	0.050936	0.0334
22	0	-0.03719	0.060723	0.03688
23	-	-	-	-
Approximate Area	**4.969**	**2.906**	**4.424**	**5.115**

References

[1] R. Banik, A. Biswas, Improving Solar PV Prediction Performance with RF-CatBoost Ensemble: A Robust and Complementary Approach, Renewable Energy Focus 46 (2023) 207–221, https://doi.org/10.1016/J.REF.2023.06.009.

[2] B.O. Abisoye, Y. Sun, W. Zenghui, A survey of artificial intelligence methods for renewable energy forecasting: Methodologies and insights, Renewable Energy Focus 48 (2024) 100529, https://doi.org/10.1016/J.REF.2023.100529.

[3] B. Li, J. Zhang, A review on the integration of probabilistic solar forecasting in power systems, Solar Energy 210 (2020) 68–86, https://doi.org/10.1016/J.SOLENER.2020.07.066.

[4] A. Tawakuli, B. Havers, V. Gulisano, D. Kaiser, T. Engel, Survey:Time-series data preprocessing: A survey and an empirical analysis, Journal of Engineering Research (2024), https://doi.org/10.1016/J.JER.2024.02.018.

[5] E. Oluwasakin, T. Torku, S. Tingting, A. Yinusa, S. Hamdan, S. Poudel, et al., Minimization of high computational cost in data preprocessing and modeling using MPI4Py, Machine Learning with Applications 13 (2023) 100483, https://doi.org/10.1016/J.MLWA.2023.100483.

[6] U.K. Das, K.S. Tey, M. Seyedmahmoudian, S. Mekhilef, M.Y.I. Idris, W. van Deventer, et al., Forecasting of photovoltaic power generation and model optimization: A review, Renewable and Sustainable Energy Reviews 81 (2018) 912–928, https://doi.org/10.1016/J.RSER.2017.08.017.

[7] T. Tian, I. Chernyakhovskiy, Forecasting Wind and Solar Generation: Improving System Operations, Greening the Grid, National Renewable Energy Laboratory, 2016.

[8] Meenal R, Binu · D, Ramya · K C, Prawin ·, Michael A, Vinoth Kumar · K, et al. Weather Forecasting for Renewable Energy System: A Review. Archives of Computational Methods in Engineering 2021;1:3. doi: 10.1007/s11831-021-09695-3.

[9] Gupta P, Singh R. International Journal of Sustainable Engineering ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsue20 PV power forecasting based on data-driven models: a review Priya Gupta & Rhythm Singh PV power forecasting based on data-driven models: a review 2021. doi: 10.1080/19397038.2021.1986590.

[10] J. Ma, X. Ma, A review of forecasting algorithms and energy management strategies for microgrids, Systems Science & Control Engineering: An Open Access Journal 6 (2018) 237–248, https://doi.org/10.1080/21642583.2018.1480979.

[11] B. Carrera, M.-K. Sim, J.-Y. Jung, PVHybNet: a hybrid framework for predicting photovoltaic power generation using both weather forecast and observation data; PVHybNet: a hybrid framework for predicting photovoltaic power generation using both weather forecast and observation data, The Institution of Engineering and Technology 14 (2020) 2192–2201, https://doi.org/10.1049/iet-rpg.2018.6174.

[12] Minhas DM, Khalid RR, Frey G. Load control for supply-demand balancing under Renewable Energy forecasting. 2017 IEEE 2nd International Conference on Direct Current Microgrids, ICDCM 2017 2017:365–70. doi: 10.1109/ICDCM.2017.8001071.

[13] G.R. Aghajani, H.A. Shayanfar, H. Shayeghi, Demand side management in a smart micro-grid in the presence of renewable generation and demand response, Energy 126 (2017) 622–637, https://doi.org/10.1016/J.ENERGY.2017.03.051.

[14] Manjarres D, Alonso R, Gil-Lopez S, Landa-Torres I. Solar Energy Forecasting and Optimization System for Efficient Renewable Energy Integration. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2017;10691 LNAI:1–12. doi: 10.1007/978-3-319-71643-5_1.

[15] P. Pawar, M. TarunKumar, K.P. Vittal, An IoT based Intelligent Smart Energy Management System with accurate forecasting and load strategy for renewable generation, Measurement 152 (2020) 107187, https://doi.org/10.1016/J.MEASUREMENT.2019.107187.

[16] H. Sangrody, N. Zhou, Z. Zhang, Similarity-Based Models for Day-Ahead Solar PV Generation Forecasting, IEEE Access 8 (2020) 104469–104478, https://doi.org/10.1109/ACCESS.2020.2999903.

[17] P. Pawar, N. Mithulananthan, R.MQ. Solar, P.V. Power Forecasting Using Modified, SVR with Gauss-Newton Method. Proceedings -,, IEEE 2nd Global Power, Energy and Communication Conference, GPECOM 2020 (2020) (2020) 226–231, https://doi.org/10.1109/GPECOM49333.2020.9247935.

[18] H. Feng, C. Yu, A novel hybrid model for short-term prediction of PV power based on KS-CEEMDAN-SE-LSTM, Renewable Energy Focus 47 (2023) 100497, https://doi.org/10.1016/J.REF.2023.100497.

[19] Wang B, Che J, Wang B, Feng S. A Solar Power Prediction Using Support Vector Machines Based on Multi-source Data Fusion. 2018 International Conference on Power System Technology, POWERCON 2018 - Proceedings 2019:4573–7. doi: 10.1109/POWERCON.2018.8601672.

[20] K. Rammelkamp, S. Schröder, S. Kubitza, D.S. Vogt, S. Frohmann, P.B. Hansen, et al., Low-level LIBS and Raman data fusion in the context of in situ Mars exploration, Journal of Raman Spectroscopy 51 (2020) 1682–1701, https://doi.org/10.1002/JRS.5615.

[21] Kuo PH, Huang CJ. A Green Energy Application in Energy Management Systems by an Artificial Intelligence-Based Solar Radiation Forecasting Model. Energies 2018, Vol 11, Page 819 2018;11:819. doi: 10.3390/EN11040819.

[22] G. Narvaez, L.F. Giraldo, M. Bressan, A. Pantoja, Machine learning for site-adaptation and solar radiation forecasting, Renew Energy 167 (2021) 333–342, https://doi.org/10.1016/J.RENENE.2020.11.089.

[23] I. Jebli, F.Z. Belouadha, M.I. Kabbaj, A. Tilioua, Prediction of solar energy guided by pearson correlation using machine learning, Energy 224 (2021) 120109, https://doi.org/10.1016/J.ENERGY.2021.120109.

[24] V. Kushwaha, N.M. Pindoriya, A SARIMA-RVFL hybrid model assisted by wavelet decomposition for very short-term solar PV power generation forecast | Elsevier Enhanced Reader, Elsevier Ltd 140 (2019) 124–139, https://doi.org/10.1016/j.renene.2019.03.020.

[25] R. Anaadumba, Q. Liu, B.D. Marah, F.M. Nakoty, X. Liu, Y. Zhang, A renewable energy forecasting and control approach to secured edge-level efficiency in a distributed micro-grid, Cybersecurity 4 (2021), https://doi.org/10.1186/s42400-020-00065-3.

[26] M. Rana, A. Rahman, Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling, Sustainable Energy, Grids and Networks 21 (2020), https://doi.org/10.1016/J.SEGAN.2019.100286.

[27] D. Sadeghi, A. Golshanfard, S. Eslami, K. Rahbar, R. Kari, Improving PV power plant forecast accuracy: A hybrid deep learning approach compared across short, medium, and long-term horizons, Renewable Energy Focus 45 (2023) 242–258, https://doi.org/10.1016/J.REF.2023.04.010.

[28] S. Sarp, M. Kuzlu, U. Cali, O. Guler, An Interpretable Solar Photovoltaic Power Generation Forecasting Approach Using An Explainable Artificial Intelligence Tool, in: IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2021, pp. 1–5, https://doi.org/10.1109/ISGT49243.2021.9372263.

[29] H. Aprillia, H.-T. Yang, C.-M. Huang, Short-Term Photovoltaic Power Forecasting Using a Convolutional Neural Network-Salp Swarm Algorithm, Energies (Basel) 13 (2020) 1–20, https://doi.org/10.3390/en13081879.

[30] D. Cannizzaro, A. Aliberti, L. Bottaccioli, E. Macii, A. Acquaviva, E. Patti, Solar radiation forecasting based on convolutional neural network and ensemble learning, Expert Syst Appl 181 (2021), https://doi.org/10.1016/j.eswa.2021.115167.

[31] N.Q. Nguyen, L.D. Bui, B.V. Doan, E.R. Sanseverino, D.D. Cara, Q.D. Nguyen, A new method for forecasting energy output of a large-scale solar power plant based on long short-term memory networks a case study in Vietnam, Electric Power Systems Research (2021) 199, https://doi.org/10.1016/j.epsr.2021.107427.

[32] C. Pan, J. Tan, D. Feng, Prediction intervals estimation of solar generation based on gated recurrent unit and kernel density estimation, Neurocomputing 453 (2021) 552–562, https://doi.org/10.1016/j.neucom.2020.10.027.

[33] A. Ngoc-Lan Huynh, R.C. Deo, M. Ali, S. Abdulla, N. Raj, Novel short-term solar radiation hybrid model: Long short-term memory network integrated with robust local mean decomposition, Appl Energy (2021) 298, https://doi.org/10.1016/j.apenergy.2021.117193.

[34] D.J. Pedregal, J.R. Trapero, Adjusted combination of moving averages: A forecasting system for medium-term solar irradiance, Appl Energy (2021) 298, https://doi.org/10.1016/j.apenergy.2021.117155.

[35] T. Ahmad, D. Zhang, C. Huang, Methodological framework for short-and medium-term energy, solar and wind power forecasting with stochastic-based machine learning approach to monetary and energy policy applications, Energy (2021) 231, https://doi.org/10.1016/j.energy.2021.120911.

[36] D. Korkmaz, SolarNet: A hybrid reliable model based on convolutional neural network and variational mode decomposition for hourly photovoltaic power forecasting, Appl Energy (2021) 300, https://doi.org/10.1016/j.apenergy.2021.117410.

[37] F. Najibi, D. Apostolopoulou, E. Alonso, Enhanced performance Gaussian process regression for probabilistic short-term solar output forecast, International Journal of Electrical Power & Energy Systems 130 (2021) 106916, https://doi.org/10.1016/J.IJEPES.2021.106916.

[38] F. Rodríguez, A. Fleetwood, A. Galarza, L. Fontán, Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control, Renew Energy 126 (2018) 855–864, https://doi.org/10.1016/J.RENENE.2018.03.070.

[39] Thukral MK. Solar Power Output Prediction Using Multilayered Feedforward Neural Network: A Case Study of Jaipur, IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC) 2020 (2020) 1–6, https://doi.org/10.1109/iSSSC50941.2020.9358821.

[40] U. Munawar, Z. Wang, A Framework of Using Machine Learning Approaches for Short-Term Solar Power Forecasting, Journal of Electrical Engineering & Technology 15 (2020) 561–569, https://doi.org/10.1007/s42835-020-00346-4.

[41] H. Yakoubi, Y. el Mghouchi, N. Abdou, A. Hajou, A. Khellouki, Correlating clearness index with cloud cover and other meteorological parameters for forecasting the global solar radiation over Morocco, Optik (Stuttg) (2021) 242, https://doi.org/10.1016/j.ijleo.2021.167145.

[42] H. Zang, L. Cheng, T. Ding, K.W. Cheung, Z. Liang, Z. Wei, et al., Hybrid method for short-term photovoltaic power forecasting based on deep convolutional neural network, IET Gener Transm Distrib 12 (2018) 4557–4567, https://doi.org/10.1049/iet-gtd.2018.5847.

[43] B. Chen, P. Lin, Y. Lai, S. Cheng, Z. Chen, L. Wu, Very-Short-Term Power Prediction for PV Power Plants Using a Simple and Effective RCC-LSTM Model Based on Short Term Multivariate Historical Datasets, Electronics (Basel) (2020) 9, https://doi.org/10.3390/electronics9020289.

[44] Iyengar S, Sharma N, Irwin D, Shenoy P, Ramamritham K. SolarCast-A Cloud-based Black Box Solar Predictor for Smart Homes. BuildSys '14: 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, 2014, p. 40–9. doi: 10.1145/2674061.2674071.

[45] J.R. Andrade, R.J. Bessa, Improving Renewable Energy Forecasting With a Grid of Numerical Weather Predictions. IEEE Trans Sustain, Energy (2017). 8th:1571–80, https://doi.org/10.1109/TSTE.2017.2694340.

[46] S. Leva, A. Dolara, F. Grimaccia, M. Mussetta, E. Ogliari, Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power, Math

Comput Simul 131 (2017) 88–100, https://doi.org/10.1016/J.MATCOM.2015.05.010.
[47] C. Persson, P. Bacher, T. Shiga, H. Madsen, Multi-site solar power forecasting using gradient boosted regression trees, Solar Energy 150 (2017) 423–436, https://doi.org/10.1016/J.SOLENER.2017.04.066.
[48] O. Abedinia, D. Raisz, N. Amjady, IET Renewable Power Generation Effective prediction model for Hungarian small-scale solar power output, The Institution of Engineering and Technology 11 (2017) 1648–1658, https://doi.org/10.1049/iet-rpg.2017.0165.
[49] S.-G. Kim, J.-Y. Jung, M.K. Sim, A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning, Sustainability (2019), https://doi.org/10.3390/su11051501.
[50] Kyliashkina IA, Eroshenko SA, Shelyug S. Intelligent Systems as a Tool for Predicting Electrical Energy and Power Generation. 2019 60th International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), 2019, p. 1–5. doi: 10.1109/ITMS47855.2019.8940721.
[51] H.A. Kazem, J.H. Yousif, M.T. Chaichan, A.H.A. Al-Waeli, K. Sopian, Long-term power forecasting using FRNN and PCA models for calculating output parameters in solar photovoltaic generation, Heliyon 8 (2022) e08803.
[52] P. Chaudhary, M. Rizwan, Energy management supporting high penetration of solar photovoltaic generation for smart grid using solar forecasts and pumped hydro storage system, Renew Energy 118 (2018) 928–946, https://doi.org/10.1016/J.RENENE.2017.10.113.
[53] O. Elma, A. Taşcıkaraoğlu, A. Tahir İnce, U.S. Selamoğulları, Implementation of a dynamic energy management system using real time pricing and local renewable energy generation forecasts, Energy 134 (2017) 206–220, https://doi.org/10.1016/J.ENERGY.2017.06.011.
[54] Scabbia G, Sanfilippo A, Bachour D, Perez-Astudillo D. Exogenous Parameters in Solar Forecasting. Conference Record of the IEEE Photovoltaic Specialists Conference 2020;2020-June:0894–6. doi: 10.1109/PVSC45281.2020.9300800.
[55] A. Moreno-Muñoz (Ed.), Large scale grid integration of renewable energy sources, Institution of Engineering and Technology, 2017.
[56] B. Gu, H. Shen, X. Lei, H. Hu, X. Liu, Forecasting and uncertainty analysis of day-ahead photovoltaic power using a novel forecasting method, Appl Energy (2021) 299, https://doi.org/10.1016/j.apenergy.2021.117291.
[57] L.L. Li, S.Y. Wen, M.L. Tseng, C.S. Wang, Renewable energy prediction: A novel short-term prediction model of photovoltaic output power, J Clean Prod 228 (2019) 359–375, https://doi.org/10.1016/J.JCLEPRO.2019.04.331.
[58] J. Wang, P. Li, R. Ran, Y. Che, Y. Zhou, A Short-Term Photovoltaic Power Prediction Model Based on the Gradient Boost Decision Tree, Applied Sciences 8 (2018) 1–14, https://doi.org/10.3390/app8050689.
[59] F. Najibi, D. Apostolopoulou, E. Alonso, Enhanced performance Gaussian process regression for probabilistic short-term solar output forecast, International Journal of Electrical Power and Energy Systems (2021) 130, https://doi.org/10.1016/j.ijepes.2021.106916.
[60] R. Nematirad, P.A. Solar, R.F.U.A.N.N.C.F. Selection, IEEE Kansas Power and Energy Conference, KPEC 2022 (2022) 2022, https://doi.org/10.1109/KPEC54747.2022.9814765.
[61] de Jesus DAR, Mandal P, Velez-Reyes M, Chakraborty S, Senjyu T. Data Fusion Based Hybrid Deep Neural Network Method for Solar PV Power Forecasting. 51st North American Power Symposium, NAPS 2019 2019. doi: 10.1109/NAPS46351.2019.9000331.
[62] Cranfield University. Cranfield University's Living Laboratory and Urban Observatory - Data. Living Laboratory 2021. https://www.livinglab.ac.uk/data.html (accessed February 21, 2022).
[63] Cranfield University. Energy Manager Live 2022. https://cranfield.energymanagerlive.com/ (accessed February 21, 2022).
[64] M. Pazhoohesh, M. Sadegh Javadi, M. Gheisari, S. Aziz, Dealing with Missing Data in the Smart Buildings using Innovative Imputation Techniques; Dealing with Missing Data in the Smart Buildings using Innovative Imputation Techniques, in: IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society, 2021, https://doi.org/10.1109/IECON48115.2021.9612650.
[65] F. Bashir, H.L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, Neurocomputing 276 (2018) 23–30, https://doi.org/10.1016/J.NEUCOM.2017.03.097.
[66] C. Robert, W. Jessep, J.J. Sutton, T.M. Hicks, M. Loeffen, M. Farouk, et al., Evaluating low- mid- and high-level fusion strategies for combining Raman and infrared spectroscopy for quality assessment of red meat, Food Chem 361 (2021) 130154, https://doi.org/10.1016/J.FOODCHEM.2021.130154.
[67] Microsoft. PEARSON function. Microsoft Support 2021. https://support.microsoft.com/en-us/office/pearson-function-0c3e30fc-e5af-49c4-808a-3ef66e034c18 (accessed October 15, 2021).
[68] Microsoft. Use the Analysis ToolPak to perform complex data analysis - Office Support. Microsoft 2021. https://support.microsoft.com/en-us/office/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6 (accessed July 15, 2021).
[69] L. Qin, W. Huang, Z. Guo, Y. Zhou, B. Zhang, Topology Identification Method of Low-voltage Distribution Network Based on Improved Pearson Correlation Coefficient Method (2022) 1–6, https://doi.org/10.1109/CIYCEE53554.2021.9676722.
[70] K. Qian, X. Wang, Y. Yuan, Research on regional short-term power load forecasting model and case analysis, Processes 9 (2021), https://doi.org/10.3390/PR9091617.
[71] Herwanto HW, Handayani AN, Wibawa AP, Chandrika KL, Arai K. Comparison of Min-Max, Z-Score and Decimal Scaling Normalization for Zoning Feature Extraction on Javanese Character Recognition. 7th International Conference on Electrical, Electonics and Information Engineering (ICEEE) 2021:1–3. doi: 10.1109/ICEEIE52663.2021.9616665.
[72] Normalize data - MATLAB normalize - MathWorks United Kingdom n.d. https://uk.mathworks.com/help/matlab/ref/double.normalize.html (accessed October 30, 2021).
[73] M. Ghofrani, M. Alolayan, Time Series and Renewable Energy Forecasting, Time Series Analysis and Applications (2017), https://doi.org/10.5772/INTECHOPEN.70845.
[74] B. Zazoum, Solar photovoltaic power prediction using different machine learning methods, Energy Reports 8 (2022) 19–25, https://doi.org/10.1016/J.EGYR.2021.11.183.
[75] S. Ekici, F. Unal, U. Ozleyen, Comparison of different regression models to estimate fault location on hybrid power systems, IET Generation, Transmission and Distribution 13 (2019) 4756–4765, https://doi.org/10.1049/IET-GTD.2018.6213.
[76] S.K. Sunori, P.B. Negi, A. Mittal, S. Arora, S. Sharma, S. Maurya, et al., Prediction of pH of Soil Using Linear Regression, Decision Tree, GPR, and Gaussian SVM, in: IEMECON 2021–10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks, 2021, https://doi.org/10.1109/IEMECON53809.2021.9689193.
[77] Ortigoza G, Zapata U. Covid-19 Projections: A Simple Machine Learning Approach. 2021 IEEE International Conference on Engineering Veracruz, ICEV 2021 2021. doi: 10.1109/ICEV52951.2021.9632647.
[78] M. Madhukumar, A. Sebastian, X. Liang, M. Jamil, M.N.S.K. Shabbir, Regression Model-Based Short-Term Load Forecasting for University Campus Load, IEEE Access 10 (2022) 8891–8905, https://doi.org/10.1109/ACCESS.2022.3144206.
[79] P.S. Popolo, A.M. Johnson, Relating Cepstral Peak Prominence to Cyclical Parameters of Vocal Fold Vibration from High-Speed Videoendoscopy Using Machine Learning: A Pilot Study, Journal of Voice 35 (2021) 703–716, https://doi.org/10.1016/J.JVOICE.2020.01.026.
[80] dela Cruz MAA, Legaspi KMC, Marcelino RMD, Rosete JR, Sangalang DAR, Suarez CG, et al. Joint Gait Kinematic and Kinetic Analysis using Inertial Measurement Units and Plantar Pressure Sensor System. 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2019 2019. doi: 10.1109/HNICEM48295.2019.9072701.
[81] I. Daut, M.I. Yusoff, S. Ibrahim, M. Irwanto, G. Nsurface, Relationship between the Solar Radiation and Surface Temperature in Perlis, Trans Tech Publications, Adv Mat Res, Durnten-Zurich, 2012, pp. 143–147.
[82] K. Shang, Y. Liu, D. Wang, T. Li, Dynamic relationship between outdoor air temperature and solar radiation, Tumu Jianzhu Yu Huanjing Gongcheng/Journal of Civil, Architectural and Environmental Engineering 37 (2015) 116–121, https://doi.org/10.11835/J.ISSN.1674-4764.2015.05.017.
[83] K.F. Lateef, J. Paul, Z. Jahan, Weather Parameters and Heat Index Prediction Applying Deep Neural Networks, in: TENSYMP 2021–2021 IEEE Region 10 Symposium, 2021, https://doi.org/10.1109/TENSYMP52854.2021.9550852.
[84] M.E. Howarth, N.F. Laird, Intraseasonal variations of winter wind chill temperatures across Canada and the United States, J Appl Meteorol Climatol 56 (2017) 2951–2962, https://doi.org/10.1175/JAMC-D-16-0394.1.
[85] Fraihat H, Almbaideen AA, Al-Odienat A, Al-Naami B, de Fazio R, Visconti P. Solar Radiation Forecasting by Pearson Correlation Using LSTM Neural Network and ANFIS Method: Application in the West-Central Jordan. Future Internet 2022, Vol 14, Page 79 2022;14:79. doi: 10.3390/FI14030079.
[86] N.M.M. Altwallbah, M.A.M. Radzi, N. Azis, S. Shafie, M.A.A.M. Zainuri, New perturb and observe algorithm based on trapezoidal rule: Uniform and partial shading conditions, Energy Convers Manag 264 (2022) 115738, https://doi.org/10.1016/J.ENCONMAN.2022.115738.

https://dspace.org/

A three-step weather data approach in solar energy prediction using machine learning

Falope, Tolulope Olumuyiwa; Lao, Liyun; Hanak, Dawid P.

2024-09-01
Elsevier
https://doi.org/10.1016/j.ref.2024.100615
Falope TO, Lao L, Hanak D. (2024) A three-step weather data approach in solar energy prediction using machine learning. Renewable Energy Focus, Volume 50, September 2024, Article number 100615

Downloaded from a DSpace repository. https://dspace.org/