Philippine Journal of Science
152 (S1): 95-108, Philippine Meteorological Research
ISSN 0031 - 7683
Date Received: 15 Feb 2023

Determining Meteorological Parameters Influencing Photovoltaic Solar Energy Generation in Quezon City Using Machine Learning Algorithms

Lea Angela M. Saure[1]*, Joshua P. Quides[1], Raymond C. Ordinario[1,2], and Rhenish C. Simon[3]*

[1]Department of Physical Sciences, College of Science,
Polytechnic University of the Philippines, Sta. Mesa, Manila 1016 Philippines
[2]Weather Division, Philippine Atmospheric, Geophysical,
and Astronomical Services Administration, Quezon City, Metro Manila 1100 Philippines
[3]Physics Department, College of Science,
De La Salle University, Taft Avenue, Manila 1004 Philippines

One challenge in adapting to energy generation using solar photovoltaic (PV) modules is its variability with changing weather conditions. In this study, we aim to determine the effect of meteorological parameters that have the most effect on the variability of solar energy generation (SEG). Our study is conducted in Quezon City, part of the National Capital Region, Philippines. The maximum temperature, relative humidity, mean temperature, and cloud opacity have the most effect on the variability of the SEG among the eight meteorological parameters that we consider in our study based on the principal component regressor (PCR) and random forest regressor (RFR) machine learning algorithms. The PCR model explains 55.5 and 49.2% variability in SEG of the training and test sets, respectively. On the other hand, the RFR model explains a 77.1% variation of the SEG in the training and 52.7% in the test set. Furthermore, the two models provided comparable predictions of SEG.

Keywords: data analytics, solar photovoltaic cells

INTRODUCTION

Access to electricity is necessary to function in this modern time. Electrical energy powers simple appliances in a household to big machinery in the industrial sector. According to the report of the Department of Energy (2021), the Philippines' energy mix consists of around 62% fossil fuels and 38% renewable energy sources. Continuous use of fossil fuels threatens the country's energy security, as it is subject to volatile prices (Kreps 2020) and depletion (Shafiee and Topal 2009). In the study by Mondal *et al.* (2018), they state that our energy mix should be more diversified by using more renewable energy sources for long-term sustainability.

Solar energy is one of the most promising renewable energy sources (Martínez-Sánchez *et al.* 2022). The most common way to convert solar energy to useful electrical energy is through photovoltaic (PV) modules (Zuniga-Reyes *et al.* 2021). In the present time, installing PV panels for residential and commercial use is more economical due to new government policies, improved solar technology, and relatively lower installation prices (Farias-Rocha *et al.* 2019). However, one challenge the energy industry needs to consider in investing in solar energy is the seasonal

*Corresponding author: lamsaure@iskolarngbayan.pup.edu.ph
rhenish.simon@dlsu.edu.ph

variability of the energy output. Solar energy conversion to useful electrical energy is affected by daily weather conditions. Studies in other countries show that the energy generation of solar PV panels is affected by meteorological parameters such as temperature, wind speed, relative humidity, and cloud opacity. Specifically, in the study by Smith *et al*. (2018), different values of temperature, wind speed, and albedo induce changes in PV power output. While Barbieri *et al*. (2017) show that changes in cloud cover and the type of cloud can result in insufficient solar power generated by a solar PV system. Additionally, Njok and Ogbulezie (2018) show that relative humidity has an inverse relationship with solar PV output. While the Philippines has excellent solar energy industry potential due to our geographical location (Mondal *et al.* 2018), the climate of the Philippines is described by high temperature, high humidity, and abundant rainfall (Tower 1903). The study by Acuzar *et al.* (2017) investigates the effects of weather and climate on renewable energy in the Visayas, Philippines. They are able to make a simulation tool and deduce that temperature affects solar energy generation (SEG). Notably, no studies are showing the relationship between different weather parameters to SEG in Quezon City. In this paper, we present a study on the association of SEG with the meteorological parameters within Quezon City. This contribution can be used as a reference for energy management studies in the National Capital Region. This is also an effort to achieve the UN Sustainable Goal Number 7: affordable and clean energy (UN ESCAP 2021)

We aimed to build statistical models that would determine the meteorological parameters that affect solar PV energy generation the most. To do this, we developed different models to predict the response of the solar generation output to the variation in the weather parameters. We used two machine learning algorithms – namely, the principal component regressor (PCR) and random forest regressor (RFR). Studying how weather affects SEG can help make the use of solar energy more efficient and cost-effective in the Philippines. This paper was organized into five sections. In Section 2, we described the data and the machine learning algorithms we used in detail. We presented the key findings of our study in Section 3 and give an analysis in relation to similar studies in Section 4. Finally, we provided the conclusion in Section 5, highlighting the main findings and significance.

MATERIALS AND METHODS

We used the PCR and the RFR to identify the meteorological parameters that greatly affect the solar generation output. These two machine learning algorithms are computationally simple and easy to interpret (Goh *et al*. 2017). A detailed discussion of the methods were demonstrated in this section starting from data collection.

Data Collection

We used meteorological data from the data archives of the Science Garden Station of the Philippine Atmospheric, Geophysical, and Astronomical Services Administration (PAGASA) located in Quezon City, Philippines from 01 Jan 2010–31 Dec 2021. The data from the Science Garden Station can represent the meteorological condition within Quezon City. However, capturing the local weather condition of a specific location in Quezon City may have had some variations. The meteorological parameters we use were: [1] the daily rainfall in millimeters (mm), which is measured as the accumulated rainfall from the 6-h observations using the manual 8-in rain gauge; [2] the daily minimum and [3] maximum temperatures in degrees Celsius (°C), which represent the lowest and highest recorded air temperature within the day; [4] the mean daily temperature in °C, which is calculated as the average of the highest and lowest temperatures of the day; [5] the daily wind speed in meters per second (m/s), which is determined as the average of the 3-hourly observation within the day alongside [6] the wind direction (degrees), which is the mode of the 3-hourly observations within the day obtained from the aerovane installed at 10 m high; [7] the daily relative humidity, which is measured by taking the average of the lowest wet bulb temperature and the highest dry bulb temperature recorded during the day; and, lastly, [8] the daily cloud amount is the average of available cloud amount observation within the day in oktas or fraction of 8-divided aerodrome observations as seen by the observer within the station, which is the metric for cloud amount.

In addition, we used the SEG data from Ensky Corporation. The SEG data were recorded in terms of kilowatt-hours (kWh). The solar facility is located 7.27 km northeast of the Science Garden Station. The solar PV system was installed on 10 May 2017 and consists of 40 Japan solar polycrystalline PV modules (JS-270P-60) and a 5-kW inverter (SOFAR KTLM 5.0). The complete specification for the panel includes the following: panel efficiency of 16.6%; maximum power output of 270Wp; voltage at maximum power of 31.3V; current at maximum power of 8.63A; open circuit voltage of 38.5V; and dimensions of 1640 mm x 992 mm x 40 mm (H x W x D). This PV system is recommended for residential buildings with high electricity consumption.

Data Preparation of Meteorological Data

The raw meteorological data from PAGASA were arranged in an array to construct the matrix $\dot{\mathbf{X}}$. Each

column of this matrix represents a meteorological variable. There were eight meteorological parameters considered in this raw data matrix, giving $\dot{\mathbf{X}}$ a total of eight $(N = 8)$ columns. On the other hand, the rows of the data matrix $\dot{\mathbf{x}}$ represent the days of the measurement, which are arranged chronologically from top to bottom. The total number of daily measurements is $m = 1307$, giving $\dot{\mathbf{X}}$ a dimension of $m \times N$. The element $\dot{x}_j^{(i)}$ of $\dot{\mathbf{X}}$ represents the measurement done on the i th day for the meteorological measurement for parameter $j = \{1, 2, 3 \ldots N\}$. The corresponding meteorological parameter for a particular value for the dummy variable j is given in Table 1.

Table 1. Dummy variables for the meteorological parameters.

Dummy variable (j)	Meteorological parameter
1	Maximum temperature
2	Mean temperature
3	Minimum temperature
4	Rainfall
5	Relative humidity
6	Wind speed
7	Wind direction
8	Cloud opacity

In the data provided by PAGASA, there were some days with missing values that are caused by technical issues that occurred on the day of the measurement. Particularly, there were no meteorological measurements from 25–28 Mar 2021. To fill out the missing data entries, we used a non-negative matrix factorization (NMF) algorithm. Although we could have simply performed a listwise deletion on a row with missing data, we opted to preserve the rows by using data imputation to avoid any biases induced by data deletion and maximize the available data. Data deletion would also put a list of observed data to waste because of one unobserved data (Little and Rubin 2019). Our meteorological data set had a low number of missing data and, thus, listwise deletion did not create a significant deviation. However, we chose to use data imputation in dealing with the missing data to show a holistic approach in case the data had a significant number of missing values. Figure 1 shows the list plot of the scaled and imputed meteorological data in chronological order.

There are alternative techniques available to analyze the data despite having missing values such as masking and adversarial techniques. In our work, we used NMF for data imputation. The NMF algorithm has an innate clustering property (Ding *et al.* 2005), which implies that the data

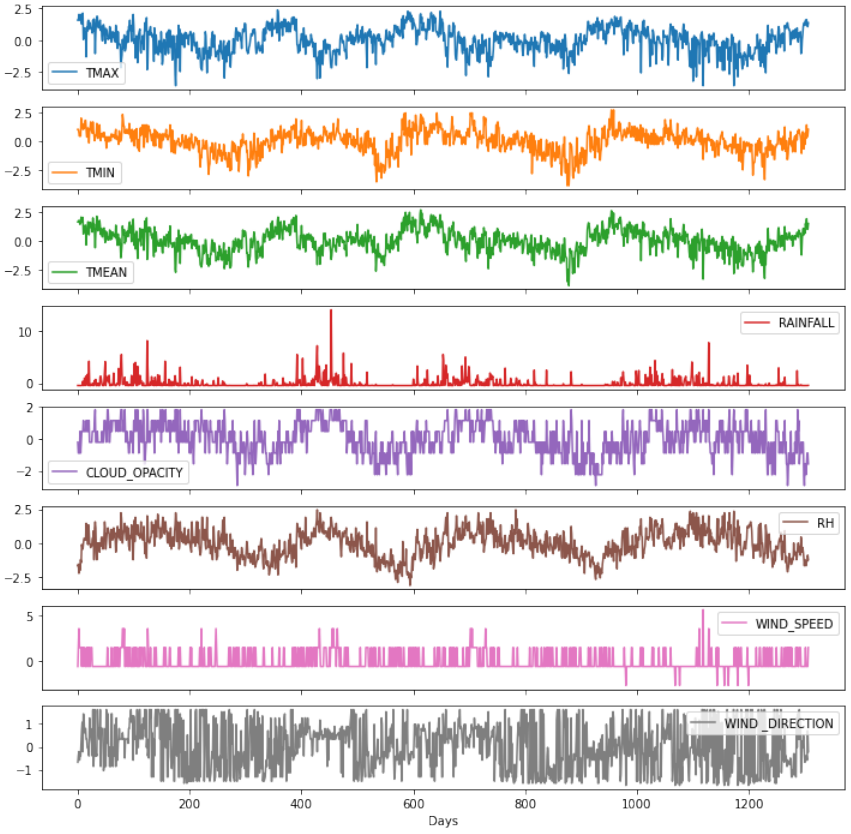

Figure 1. Time series plot of the feature-scaled meteorological (explanatory) variables. The plot contains both observed and imputed data points.

imputation using this algorithm is based on the clusters in the data. In this algorithm, the matrix $\dot{\mathbf{X}}$ was modeled as a product of two matrix factors, as shown in Equation 1:

$$\dot{\mathbf{X}} = \mathbf{U}\mathbf{V}^{\intercal} \tag{1}$$

Matrices $\mathbf{U}$ and $\mathbf{V}$ had dimensions $m \times k$ and $N \times k$, respectively. The k here represents the total number of latent features used in the factorization. The symbol $^{\intercal}$ means matrix transpose. This algorithm exploited the fact that the daily measurements (rows) and meteorological parameters (columns) had similar latent features due to relative seasonality and/or correlation to one another. The matrix $\mathbf{U}$ captured the latent features among the daily measurements, whereas the matrix $\mathbf{V}$ captured the latent features among the meteorological parameters. The best approximation for $\mathbf{U}$ and $\mathbf{V}$ was determined by applying gradient descent to reduce the sum of the square errors (SSE) between raw and predicted values for $\mathbf{x}_j^{(i)}$. The SSE is simply the sum of the square of the elements of the error matrix given by:

$$\mathbf{\Delta} = (\dot{\mathbf{X}} - \mathbf{U}\mathbf{V}^{\intercal}) \odot \mathbf{R} \tag{2}$$

The symbol $\odot$ denotes Hadamard or element-wise matrix multiplication. The matrix $\mathbf{R}$ in the equation above has elements $R_j^{(i)}$:

$$R_j^{(i)} = \begin{cases} 0 \text{ if } \mathrm{x}_j^{(i)} \text{ has value} \\ 1 \text{ if } \mathrm{x}_j^{(i)} \text{ is missing .} \end{cases}$$

Algorithm 1 summarizes the process of the NMF via gradient descent.

Algorithm 1. **Matrix factorization *via* gradient descent**

Input: **Matrix $\dot{\mathbf{x}}$**

Output: **$\mathbf{U}$ and $\mathbf{V}^{\intercal}$.**

Initialize guess for $\mathbf{U}$ and $\mathbf{V}$

$$\partial_U E = -\frac{2}{N}\Delta\mathbf{V}$$

$$\partial_V E = -\frac{2}{N}\Delta^{\intercal}\mathbf{U}$$

$$\mathbf{U} \leftarrow \mathbf{U} + \eta\,\partial_U E$$

$$\mathbf{V} \leftarrow \mathbf{V} + \eta\,\partial_V E$$

Repeat until $\mathrm{SSE} = \sum_i \sum_j \left(\Delta_j^{(i)}\right)^2$ **converges**

After performing NMF, each column of the matrix $\dot{\mathbf{X}}$ was feature scaled. Feature scaling is a standard procedure in data preparation to help the learning algorithm eliminate bias and find the optimization conditions efficiently (Ozsahin *et al.* 2022). Feature scaled observed variables were calculated using z-scores:

$$x_j^{(i)} = \frac{\dot{x}_j^{(i)} - Mean\left[\dot{x}_j^{(i)}\right]}{\text{Standard Deviation}\left[\dot{x}_j^{(i)}\right]}. \tag{3}$$

In the equation above, the values used for the calculations for the mean and the standard deviation are only from the jth column. The featured scaled values were used to construct a new matrix $\mathbf{X}$, which is the explanatory matrix. The jth column of matrix $\mathbf{X}$ is given by the vector:

$$\vec{\mathbf{x}}_j = \begin{bmatrix} x_j^{(1)} & x_j^{(2)} & x_j^{(3)} & x_j^{(4)} & \cdots & x_j^{(m)} \end{bmatrix}^{\intercal},$$

where the subscript j is the dummy variable for the meteorological parameters, and the superscript denotes the measurement day. The symbol $^{\intercal}$ means matrix transpose. The vector $\vec{\mathbf{x}}_j$, the explanatory vector corresponding to the explanatory variable j, had a dimension of $m \times 1$, i.e. $\vec{\mathbf{x}}_j \in \mathbb{R}^{m \times 1}$. The matrix $\mathbf{X}$ can be written compactly as an array of explanatory vectors $\vec{\mathbf{x}}_j{}'$s, that is:

$$\mathbf{X} = [\,\vec{\mathbf{x}}_1\ \vec{\mathbf{x}}_2\ \vec{\mathbf{x}}_3 \cdots \vec{\mathbf{x}}_N\,] \in \mathbb{R}^{m \times N}$$

Data Preparation of Solar Data

The SEG data were arranged in an $m \times 1$ array to construct the raw response vector. Just like in the matrix $\mathbf{X}$, the rows were arranged chronologically from the oldest to the latest date.

As seen in Figure 2 (left), the time series plot of the raw response vector exhibited a downward trend. This trend was due to the efficiency degradation of solar PV cells (J. Kim *et al.* 2021). The study by Aboagye *et al.* (2021) in Ghana mentioned that polycrystalline silicon modules had linear degradation with median and mean degradation rates of 1.35 and 1.44%/yr, respectively. The power generated by a PV module is proportional to its efficiency and, thus, efficiency is a multiplicative component of SEG. With this, we model raw SEG data $s^{(i)}$ as the product $\dot{y}^{(i)}l^{(i)}$, where $l^{(i)}$ is the overall trend of the data, and $\dot{y}^{(i)}$ is the component that responses to the variations of the explanatory variables.

The raw SEG data was divided by its linear trend to remove the degradation component. The detrended solar data was then scaled using min-max normalization shown on Figure 2 (right). The resulting detrended and min-max normalized SEG data $y^{(i)}$ was used to construct the final response vector $\vec{\mathbf{y}}$:

$$\vec{\mathbf{y}} = \begin{bmatrix} y^{(1)} & y^{(2)} & y^{(3)} & \cdots & y^{(i)} & \cdots & y^{(m)} \end{bmatrix}^{\intercal}.$$

The rows of matrix $\mathbf{X}$ and the response vector $\vec{\mathbf{y}}$ were matched by date (10 May 2017–09 May 2021). The transformed dataset consisting of the explanatory and response variables was randomly divided into training and test sets in a 70/30 ratio. The training set was used to train the principal component regression and random forest model. The remaining 30% of the data was used for validation.

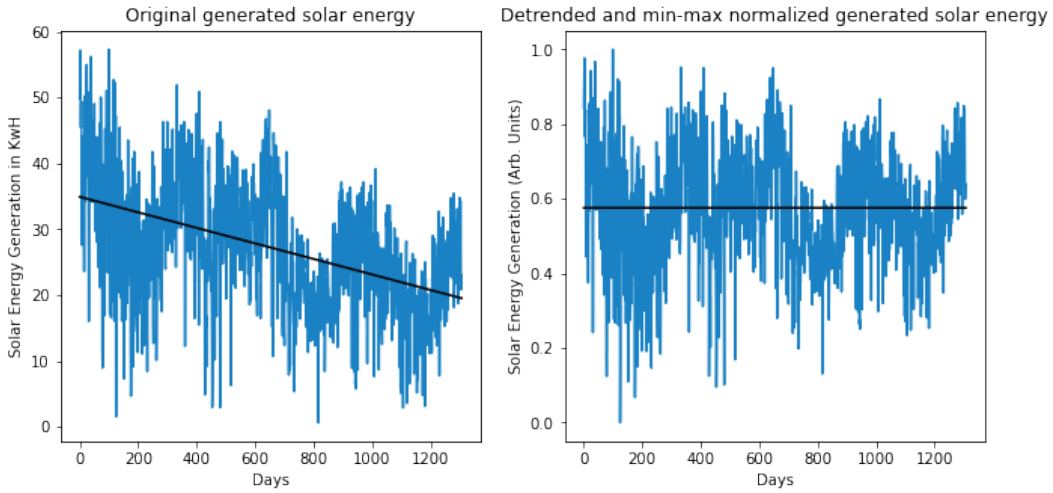

Figure 2. Comparison of the original solar dataset (left) and the detrended, min-max normalized solar dataset (right).

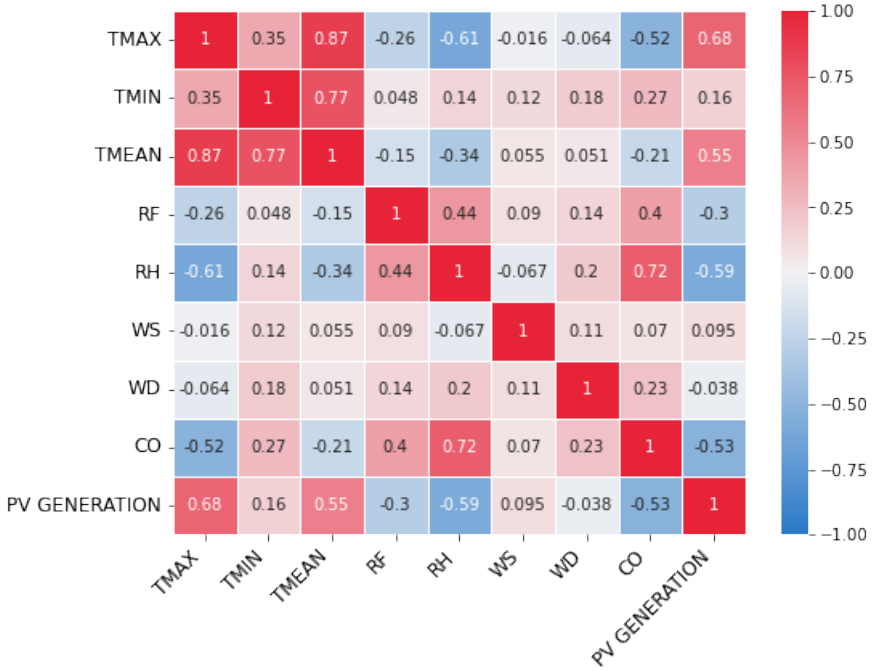

Figure 3. Correlation matrix showing the correlation coefficient between the variables. A value close to 1 indicates a positive linear correlation while a value close to –1 indicates a negative linear correlation.

Principal Component Regression

Instead of directly training the linear regression model with the scaled features, the dataset underwent principal component analysis (PCA) first. The correlation between the variables is summarized in the correlation matrix (Figure 3). Some explanatory variables such as maximum and minimum temperature were highly correlated with each other. Highly correlated response variables induce redundancy in the data and affect the model (Shlens 2014). To address this, we performed dimensionality reduction through PCA.

PCA uses a symmetric covariance matrix to determine the principal components (PCs). One fast way to determine the PCs is by solving the eigenvectors of the covariance matrix given by:

$$\mathbf{C_x} = \frac{1}{m-1}\mathbf{X}^\intercal\mathbf{X}\,. \tag{4}$$

The eigenvalue of an eigenvector of the covariance matrix determines the amount of spread in the data it captures. The eigenvector with the largest eigenvalue is called the first principal component $\vec{\mathbf{P}}_1$. The second principal component $\vec{\mathbf{P}}_2$ is the eigenvector that has the second highest eigenvalue, whereas the third principal component $\vec{\mathbf{P}}_3$ has the third highest eigenvalue, and so on.

The covariance matrix of the training set had a total of eight PCs. The eigenvalues of the PCs were arranged in descending order; the eigenvalues are 0.3614, 0.2539, 0.1233, 0.1068, 0.0932, 0.0308, 0.0255, and 0.0004. To reduce the dimension of the analysis, we only used the first five PCs, which captured around 94% of the variance of the original training set. We re-expressed the explanatory matrix using PCs as basis. This process was compactly written as:

$$\mathbf{P}_j = \mathbf{X}\,\vec{\mathbf{p}}_j, j = \{1, 2, 3 \ldots N\}. \tag{5}$$

The $\mathbf{P}_j$'s were used to construct the design matrix for the regression analysis given by:

$$\mathbf{P} = [\mathbf{P}_0 \quad \mathbf{P}_1 \quad \mathbf{P}_2 \quad \mathbf{P}_3 \quad \mathbf{P}_4 \quad \mathbf{P}_5]$$

where $\mathbf{P}_0$ is a column vector of ones in $\mathbb{R}^{m \times 1}$.

Assuming a simple linear relationship between the meteorological variables and the SEG, we can write a matrix equation for the predicted response vector as:

$$\hat{\mathbf{Y}} = \mathbf{P}\vec{\boldsymbol{\Theta}} \tag{6}$$

where $\vec{\boldsymbol{\Theta}} = [\theta_0\ \theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5]^\intercal$ is the parameter vector. The parameter vector was numerically solved using the normal equation given in Equation 7 (Kroese et al. 2019):

$$\boldsymbol{\Theta} = (\mathbf{P}^\intercal\mathbf{P})^{-1}\mathbf{P}^\intercal\mathbf{Y} \tag{7}$$

After training the principal component regression model, the PCs were transformed into their original coefficient. This is to easily determine the correlation between the solar energy produced and each meteorological parameter.

Random Forest

Ensemble learning methods combine predictions from multiple algorithms to make an accurate prediction. Random forest is an example of ensemble learning, where it consists of an ensemble of decision trees. In building the decision trees, random forest resamples the dataset by bootstrap aggregation or also known as the bagging method. Bootstrap aggregation is an algorithm used to ensure that the samples are unique to each other. The main idea of random forest is to execute bootstrap aggregation in the dataset to construct trees using only a subset of features (Kroese et al. 2019).

To achieve optimal performance, random forest requires tuning of several hyperparameters. While widely used machine learning libraries such as Sci-kit Learn provide default values for these hyperparameters, these defaults may not always produce the best results (Villegas-Mier et al. 2022). As such, we conducted tuning of these hyperparameters to improve model performance. The five hyperparameters we considered are listed below:

[a] `n_estimators`: the number of trees. More decision trees make the results more robust. However, after a certain number of trees, adding more trees would not result in observable changes in accuracy.

[b] `max_features`: the number of maximum features provided to construct each tree in a random forest. This parameter makes the random forest model handle correlation and redundancy in the dataset, as it decorrelates the trees along with bootstrap aggregation.

[c] `max_depth`: the maximum depth of the tree. Increasing the depth causes complexity in the trees; if not tuned, this may cause overfitting in the decision tree.

[d] `min_samples`: the minimum number of samples needed to split one node. Similar to maximum depth, decreasing the minimum number of samples may cause overfitting in the decision tree.

[e] `criterion`: the function used to measure the quality of the split.

To determine the best hyperparameters that generalize both training and test set, we used a Sci-kit Learn tool (version 1.0.2) called RandomizedSearchCV. Sci-kit Learn is a Python package that offers efficient versions of numerous commonly used algorithms. (VanderPlas 2016). In here, the internal five-fold cross-validation was used to find the best parameters of the model. The parameter was composed of `n_estimators` (in the range of 1–1000 trees), `max_features` ($\sqrt{N}$ and $\log_2 N$, wherein N is the total number of features in $\mathbf{X}$), `max_depth` (in the range of 1–1000), `min_samples` (in the range of 2–50), and lastly criterion (mean absolute error, mean squared error, and friedman_mse).

The best_score_params is the set of parameters that give the best generalization of the training set, which was `n_estimators` = 396, `max_features` = $\log_2 N$, `max_depth` = 927, `min_samples` = 15, and `criterion` = mean absolute error.

Algorithm 2 summarizes the steps done in training the random forest.

Algorithm 2. Random forest construction (Kroese *et al.* 2019)

Input: Training set $\tau = \{(x_j, y_j)\}_{j=1}^{m}$,

Output: Ensemble of trees = $\{g_{\mathcal{T}_b^*}\}_{b=1}^{B}$, wherein $B = 396$ (n_estimators)

Generate bootstrapped training sets $\{\mathcal{T}_1^*, \mathcal{T}_2^*, \ldots, \mathcal{T}_B^*\}$

for $b = 1$ *to* B *do*

Randomly select features, without replacement.
Using only these features, train a decision tree $g_{\mathcal{T}_b^*}$ using best_score_params as hyperparameters.

return $\{g_{\mathcal{T}_b^*}\}_{b=1}^{B}$

The average prediction of all the trees gives the final prediction of the random forest model, which is computed as:

$$\hat{\mathbf{Y}} = \frac{1}{B}\sum_{b=1}^{B} g_{\mathcal{T}_b^*}(x_j) \tag{8}$$

Lastly, random forest has a variable importance tool that was measured by how much each variable decreased the model's mean absolute error. Using the `feature_importance` tool of Sci-kit library, the meteorological variables were ranked based on how much it affects the SEG.

Model Testing

We used two model evaluation metrics to test the PCR and RFR models. The root mean squared error (RMSE) is a standard deviation of how far prediction values are from the actual values. It is computed using:

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{\left(y^{(i)} - y_p^{(i)}\right)^2}{m}} \tag{9}$$

wherein m is the number of data points. The $y^{(i)}$ is the ith element of the response data whereas $y_p^{(i)}$ is the ith element of the predicted response value. An RMSE value close to 0 indicates that the predicted response of the model is close to the actual response.

We also used the r^2 or coefficient of determination. It is a measure that determines the proportion of variance in the data described by a particular predictor or feature variable. It shows how well the regression line fits into the dataset. The higher the r^2 value is, the better the model fits the data. The resulting value for r^2 is the amount of the explained variability on the response variable by the model. Its value ranges from 0–1 in the decimal representation:

$$r^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}} = \frac{\sum_{i=1}^{m}\left(y^{(i)} - y_p^{(i)}\right)^2}{\sum_{i=1}^{m}\left(y^{i} - \hat{y}_p\right)^2}, \tag{10}$$

wherein $\sum_{i=1}^{m}\left(y^{(i)} - y_p^{(i)}\right)^2$ is the total sum of squares, and $\sum_{i=1}^{m}\left(y^{i} - \hat{y}_p\right)^2$ is the residual sum of squares. An r^2 value close to 1 indicates that the model was able to account most of the variance of the response variable.

Lastly, we conduct a paired t-test to compare the performance of two models in estimating SEG. The paired t-test is a statistical test that is used to determine whether there is a significant difference between two sets of observations that are related (Kim 2015). This test allowed us to evaluate whether the two models capture the same behavior of the predicted response values and helps us to determine which model performs better in predicting SEG.

RESULTS

In this study, we used PCR and RFR to determine the meteorological parameters that have the most effect on SEG. The resulting models are presented in this section.

Principal Component Regression Model

The principal component regression model was formulated using five PCs. The resulting parameter vector is:

$$\vec{\boldsymbol{\Theta}} = [\,1.000 \quad -0.067 \quad -0.005 \quad 0.019 \quad -0.010 \quad 0.019]^{\intercal}$$

Transforming back **P** in terms of the original coefficient **x**, the resulting mathematical model is given by:

$$\hat{\mathbf{Y}} = \mathbf{x}\vec{\boldsymbol{\beta}}, \tag{12}$$

where $\mathbf{x} = [1 \;\; x_1 \;\; x_2 \;\; x_3 \;\; x_4 \;\; x_5 \;\; x_6 \;\; x_7 \;\; x_8]$

$\vec{\boldsymbol{\beta}} = [0.577 \quad 0.039\; 0.029 \quad 0.004\; -0.007 \quad -0.037\; 0.012 \quad 0.004 \; -0.034]^{\intercal}$.

In above, x_1 is the maximum temperature, x_2 is the mean temperature, x_3 is the minimum temperature, x_4 is the rainfall, x_5 is the relative humidity, x_6 is windspeed, x_7 is the wind direction, and x_8 is the cloud opacity. According to the coefficients of the model, temperature (maximum, minimum, and mean), windspeed, and wind direction were positively correlated with solar generation, whereas rainfall, relative humidity, and cloud opacity showed a negative correlation. The PCR had an r^2 of 0.5551 for the training set and 0.4918 for the test set, whereas its RMSE values were 0.1050 and 0.1090 for the training and test set, respectively.

Figure 4 (left) shows the predicted response versus the actual response. The red line serves as a reference for the perfect match between the predicted and actual response.

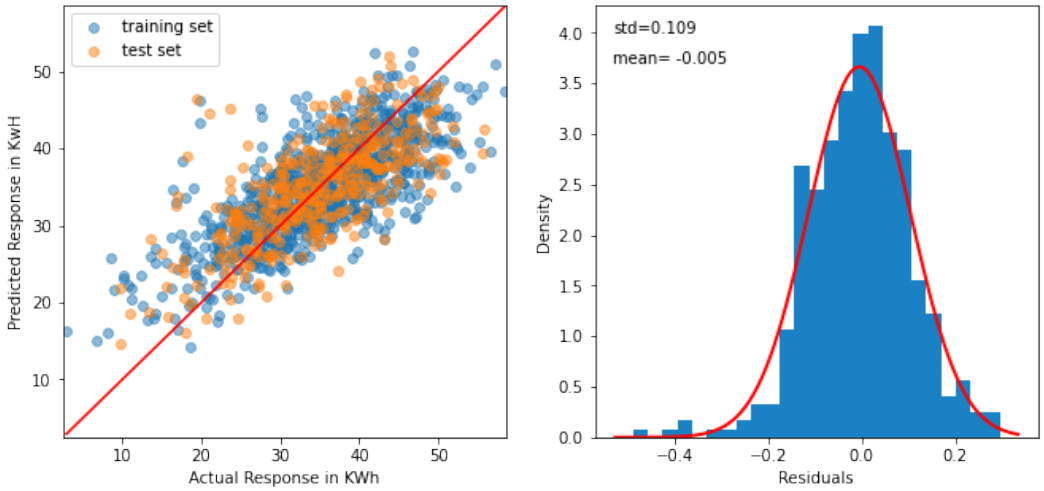

Figure 4. [Left] predicted response vs. response of solar PV generation, and [right] residual plot using PCR model. The residual plot shows that the prediction error is normally distributed about the predicted value of by the PCR model.

When a data point lies exactly on the red line, it implies that the predicted response is equal to the actual response for that particular data point. The points above the red line are overestimates, whereas the points below are underestimates. Residual is the difference between the predicted response and the actual response. The residual plot shown in Figure 4 (right) showed normally distributed residuals. The standard deviation was 0.109, whereas the mean was –0.005.

The absolute value of the magnitude of the coefficients

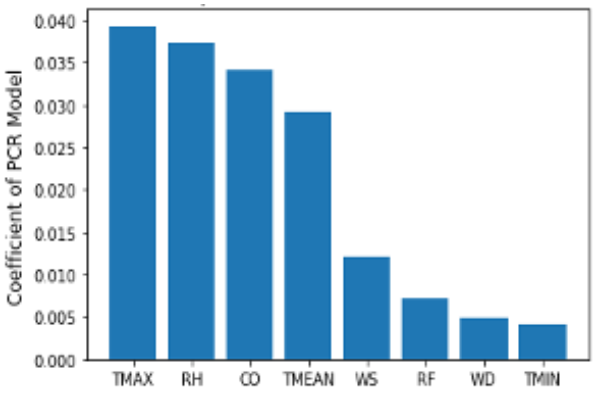

Figure 5. Feature importance ranking using the coefficients of the principal component regressor model. The top four parameters that affects the response of the solar energy generated according to the PCR model are the maximum temperature, relative humidity, cloud opacity, and mean temperature.

in the PCR model indicates the relative importance of each predictor variable. Based on Equation 12, the importance of each feature in descending order was maximum temperature, relative humidity, cloud opacity, mean temperature, windspeed, rainfall, wind direction, and minimum temperature. Figure 5 summarizes feature importance based on the PCR algorithm.

Random Forest Regression Model

The random forest model had r^2 values of 0.7705 on the training set and 0.5270 on the test set. The RMSE values were 0.0754 for the training set and 0.1051 for the test set. The red line shown in Figure 6 (left) illustrates the perfect match between the predicted response and the actual response. The dots represent the ordered pair, actual response, and predicted response values. The same behavior was seen in the residual plot of the RFR model in Figure 6 (right), where residuals were normally distributed and cluster around 0. The mean value of the residual was –0.004, whereas its standard deviation was 0.105.

Using the feature importance algorithm of random forest, the maximum temperature had the highest impact on SEG, with an importance value of 27.94%. The second was relative humidity with 21.09% importance. The rest in order were the mean temperature with 19.82%, cloud opacity with 8.53%, minimum temperature with 7.39%, rainfall with 7.37%, wind direction with 5.79%, and wind speed with 2.04% importance. Figure 7 summarizes feature importance based on the random forest algorithm.

Comparison of the Models

The corresponding absolute values of the coefficients of the PCR model explain how much the parameters affect the response. The feature importance algorithm of the random forest also tells us the importance percentage of

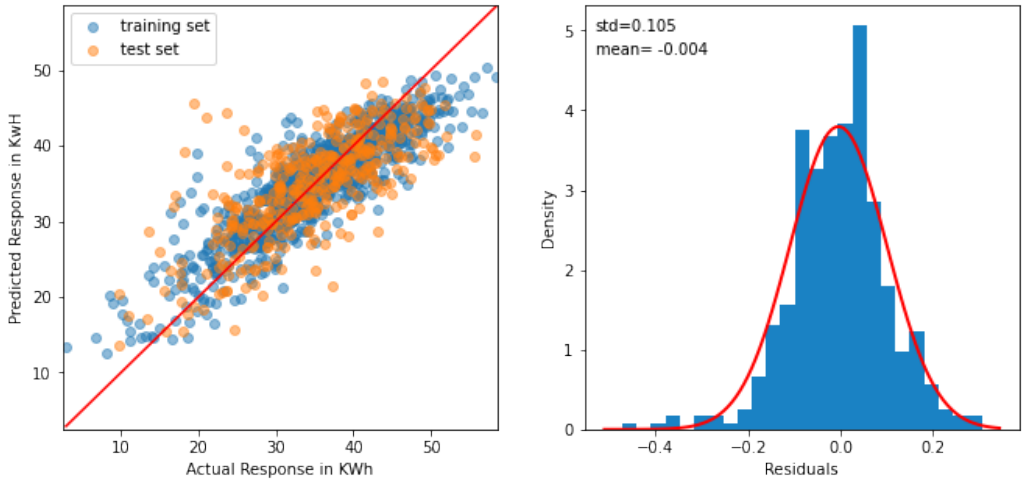

Figure 6. [Left] predicted response vs. response of solar PV generation, and (right) residual plot using RFR Model. The residual plot shows that the prediction error is normally distributed about the predicted value of by the RFR model.

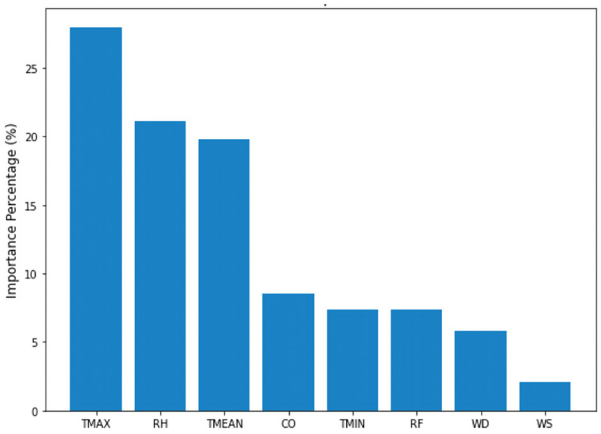

Figure 7. Feature importance ranking using the random forest regressor (RFR) model. The top four parameters that affect the response of the solar energy generated according to the RFR model are the maximum temperature, relative humidity, mean temperature, and cloud opacity.

Table 2. Variable importance ranked in descending order.

Correlation matrix	RFR	PCR
Maximum temperature	Maximum temperature	Maximum temperature
Mean temperature	Relative humidity	Relative humidity
Relative humidity	Mean temperature	Cloud opacity
Cloud opacity	Cloud opacity	Mean temperature
Rainfall	Minimum temperature	Wind speed
Minimum temperature	Rainfall	Rainfall
Windspeed	Wind direction	Wind direction
Wind direction	Wind speed	Minimum temperature

each parameter. Lastly, the correlation matrix (Figure 3) suggests the linear relationship between a pair of variables (Asuero et al. 2006). In this case, solar PV generation was paired with each of the meteorological parameters. Table 2 compares the ranking of each meteorological parameter according to its importance.

As observed, the overall order of the feature importance of the variables differed in the three methods. However, it is notable that maximum temperature, mean temperature, relative humidity, and cloud opacity were consistently the top 4 most important features. The inclusion of these four features in the top rankings suggests that they have a substantial influence on SEG in Quezon City, Philippines.

By analyzing the correlation between meteorological data and SEG, two models were formulated. Table 3 compares the performance of the PCR and random forest model.

Table 3. Calculated metrics of the two models.

Model	r^2		RMSE	
	Training set	Test set	Training set	Test set
PCR	0.555	0.492	0.105	0.109
Random forest	0.771	0.527	0.075	0.105

Overall, the random forest model had a higher r^2, which was able to account for 52.70% of solar generation in the test set compared to the PCR with only 49.20%. The r^2 difference between the two models was 3.50%. The paired t-test distribution plot of the two models is seen in Figure 8; the overlapping regions suggest that the difference in predicted values between the two models was relatively small compared to the variation of the differences. Furthermore, the resulting p-value was 0.7166, which is greater than the commonly used significance level of

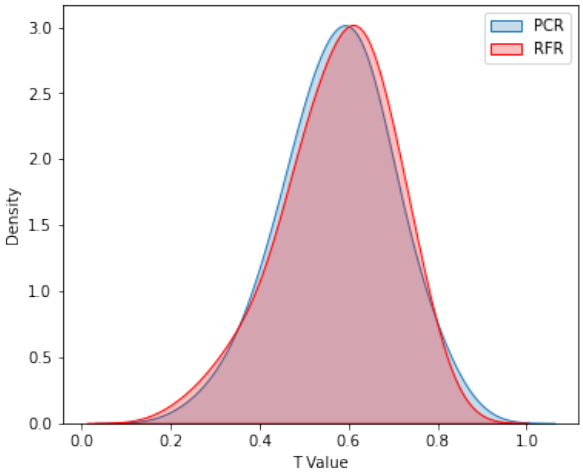

Figure 8. Paired t-test distribution plot comparing PCR and RFR models. Here, density represents the estimated probability density of the t-values. Although the RFR, compared to PCR, is better by 3.50% in predicting SEG in the test set, the t-test reveals that this difference is not statistically significant.

0.05. This indicates that there was no significant difference between the predicted values generated by the two models.

DISCUSSION

Weather and climate dynamics have an impact on SEG. For users of solar panels, having an insight to how the variation in the meteorological parameters affects the SEG output will help them manage energy usage.

This section provides an overview of various studies that investigated the correlation between meteorological parameters and SEG in different countries. Existing studies used different methods to find out how meteorological parameters affect the output of solar PV cells. In the paper of Wu et al. (2022), they reviewed multiple solar power forecasting studies. One of their key findings is that the majority of the models use irradiation, atmospheric temperature, and wind speed as input parameters. Xia (2021) used correlation analysis to study the relationship between some meteorological parameters to PV generation efficiency in China. Their results show that cloud cover and relative humidity are found to have a negative correlation to power output. While temperature decreases the efficiency of solar cells, it has a positive correlation with power output due to the high correlation with radiation. Solar panels capture solar radiation and convert it into useful power. In the study by Mousavi et al. (2015) in Iran, they determined that windspeed greatly affects solar radiation followed by temperature, whereas humidity has the lowest influence. Y.S. Kim et al. (2021) studied the correlation between meteorological parameters and power generated by a solar power plant in Samcheonpo, Korea. Using multilinear regression analysis, they determined that solar irradiation has the most significant impact followed by relative humidity. Alskaif et al. (2020) used genetic programming to see the correlation between weather parameters and PV output power in the Netherlands and the United States. Using different locations, their results show that feature importance depends on the method used and the climate zone.

In our initial study in 2022, we used only five weather parameters from Quezon City (maximum temperature, mean temperature, minimum temperature, rainfall, and relative humidity). We developed a PCR model using data from 2017–2019 and obtained an r^2 of 0.153. This study's calculation added three more parameters and a longer timeframe of dataset, which resulted in obtaining a higher r^2 value, which was 0.555. Notably, the feature importance also changed. Table 4 summarizes the key findings of related studies.

The variations observed in the findings across the mentioned studies highlight the complex nature of the relationship between meteorological parameters and SEG. Our study contributes to the understanding of this relationship by examining a specific dataset from Quezon City and by using a polycrystalline type of solar PV module. We determined the top four parameters that affect solar PV generation, the most which were identified consistently by the PCR and RFR. These features are maximum temperature, relative humidity, mean temperature, and cloud opacity.

The Earth's climate system is a complex interaction of various factors that influence how sunlight interacts with the atmosphere and the Earth's surface. One of the key relationships lies in the association between sunlight and temperature. Sunlight serves as the primary source of heat for the Earth. Changes in the amount of sunlight reaching the Earth's surface can have notable effects on temperatures across different time scales (Haigh 2002). Relative humidity serves as a metric for the amount of water vapor present in the air. Water vapor has different effects on sunlight across different regions of the electromagnetic spectrum. In the visible (VIS) region, water vapor can cause scattering of sunlight. Whereas in the ultraviolet (UV) and infrared (IR) regions, it can lead to absorption. When sunlight encounters water droplets present in the atmosphere, they may undergo reflection, refraction, or diffraction (Mekhilef *et al.* 2012). Also, water vapor is the biggest absorber of sunlight in the infrared region (Guechi *et al.* 2011). Factors like clouds, water vapor, and aerosols also contribute to sunlight attenuation. This affects the amount of sunlight that reaches the Earth's surface. Cloud opacity, for instance, can obstruct

Table 4. Comparison of results with related studies.

Author/ year	Location/ timeframe	Method	Solar PV module used	Meteorological parameters	Key findings
Saure *et al*. (2022)	Quezon City, Philippines (2017–2019)	Principal component regression	Polycrystalline	Temperature (mean, maximum, and minimum), relative humidity	The weather parameters that have the most influence on solar PV generation, ranked in descending order, are maximum temperature, relative humidity, rainfall, mean temperature, and minimum temperature
Ziane *et al.* (2021)	Saharan, Algeria (2017–2018)	Random forest regression	Polycrystalline (YL245P-29P)	Windspeed, temperature, humidity, atmospheric pressure, radiation intensity	SEG is mostly affected by radiation intensity and temperature
Kayri *et al*. (2017)	Batman, Turkey (2014–2016)	Random forest, linear regression, and artificial neural network	N/A	Global radiation, temperature, wind speed, wind direction, relative humidity, solar elevation angle	Global radiation and temperature are consistently the parameters that affect the generation of solar PV panel the most
Bahanni *et al.* (2022)	Beni Mellal and El Jadida, Morroco (2017)	Correlation analysis	Polycrystalline, monocrystalline, amorphous	Irradiance, temperature, wind speed, relative humidity	Irradiance and ambient temperature have the highest correlation with power generated output
Pasion (2019)	United States (2015–2017)	Multivariate linear regression and random forest	Polycrystalline (ALEKO 25-watt) and monocrystalline (Renogy 50-watt)	Cloud ceiling, ambient temperature, humidity, wind speed	Cloud ceiling affects solar panel energy generation the most, whereas wind speed is the least

or diffuse sunlight, thereby reducing the solar irradiance received by the solar PV module (Creayla *et al.* 2017).

The interaction between sunlight with temperature, relative humidity, and cloud opacity can be the reason why these meteorological parameters affect SEG the most. It is worthy to note that errors in our findings may be attributed to the fact that the meteorological parameters we used are taken from the nearest weather station that may not give the actual localized weather condition in the location of PV system being analyzed. Despite this, the models have been able to explain a substantial portion of the variation in generated energy, with explained variances ranging from 50–70%. This suggests that even with data limitations, the selected meteorological parameters still hold a significant influence over solar PV generation in Quezon City.

CONCLUSION AND OUTLOOK

Understanding how meteorological parameters affect SEG is important for making the use of solar energy more efficient. We determined the correlation of meteorological parameters to SEG in Quezon City using machine learning algorithms. Our results showed that the maximum temperature, mean temperature, relative humidity, and cloud opacity are the meteorological parameters that have the most effect on the electricity generation of solar panels in Quezon City. Two prediction models were also developed by analyzing the correlation between the variables. The PCR model can account for 55.5% of the variance in SEG of the training set and 49.2% of the test set. Meanwhile, the RFR model can explain 77.1% of the variance in SEG of the training set and 52.7% of the test set. The paired t-test comparing the prediction accuracy of the PCR and RFR models showed that the difference between their accuracy was not statistically significant, indicating that the models show similar patterns in predicting SEG. Notably, the PCR and RFR models determined the same four meteorological parameters that affect SEG the most.

This study contributes to our country's ongoing efforts toward clean and accessible energy by improving our understanding of SEG. In our study, we analyzed the response of a solar PV system with the meteorological data which can provide insights on their dynamics. However, our results may not be directly applicable to a different PV module system in Quezon City. Different types of solar panels have different degradation rates, efficiencies, and materials, which may affect their performance under different weather conditions. The methods used in this study may be used as a framework for future research to investigate the relationship between meteorological parameters and SEG for a variety of PV module systems. In addition, future research could also expand the scope by using data from various cities in the Philippines, as

well as including other meteorological parameters such as albedo and solar irradiance.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Climate and Agrometeorology Division of PAGASA for providing the meteorological data, as well as Ensky Corporation for the solar data. We would also like to extend our sincerest appreciation to the members of the Polytechnic University of the Philippines' Computational Physics Group for their insights and suggestions that contributed to the improvement of the study.

STATEMENT ON CONFLICT OF INTEREST

The authors declare no conflict of interest.

NOMENCLATURE

[m] maximum number of rows (daily measurements)

[N] total number of columns (meteorological parameters)

[i,j] dummy variable for row, column

[$x_j^{(i)}$] element of matrix **X** in its ith row and jth column

[$\dot{\mathbf{X}}$] raw meteorological data

[**X**] meteorological data matrix without missing data (explanatory matrix)

[$\vec{x}_j$] column vector of matrix **X**

[**Y**] response vector (SEG data)

[$\widehat{\mathbf{Y}}$] predicted response vector

[$\vec{\mathbf{p}}_j$] jth principal component (for $j = \{1, 2, 3 \ldots N\}$, only)

[**P**] design matrix for principal component regression

REFERENCES

ABOAGYE B, GYAMFI S, OFOSU E A, DJORDJEVIC S. 2021. Degradation analysis of installed solar photovoltaic (PV) modules under outdoor conditions in Ghana. Energy Reports (7): 6921–6931.

ACUZAR AMA, ARGUELLES IPE, ELISAN JCS, GOBENCIONG JKD, SORIANO AM, ROCAMORA JMB. 2017. Effects of weather and climate on renewable energy resources in a distributed generation system simulated in Visayas, Philippines. In: 2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM). p. 1–6.

ALSKAIF T, DEV S, VISSER L, HOSSARI M, VAN SARK W. 2020. A systematic analysis of meteorological variables for PV output power estimation. Renewable Energy 153: 12–22.

ASUERO AG, SAYAGO A, GONZÁLEZ AG. 2006. The Correlation Coefficient: an Overview, Critical Reviews in Analytical Chemistry 36(1): 41–59.

BAHANNI C, ADAR M, BOULMRHARJ S, KHAIDAR M, MABROUKI M. 2022. Performance comparison and impact of weather conditions on different photovoltaic modules in two different cities. Indones J Electr Eng Comput Sci 25(3): 1275–1286.

BARBIERI F, RAJAKARUNA S, GHOSH A. 2017. Very short-term photovoltaic power forecasting with cloud modeling: a review. Renewable and Sustainable Energy Reviews 75: 242–263.

CREAYLA CMC, GARCIA FCC, MACABEBE EQB. 2017. Next day power forecast model using smart hybrid energy monitoring system and meteorological data. In: 2016 IEEE International Symposium on Robotics and Intelligent Sensors IRIS 2016; 17–20 Dec 2016; Tokyo, Japan: Procedia Computer Science 105: 256–263.

DING C, HE X, SIMON HD. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of the 2005 SIAM international conference on data mining. Society for Industrial and Applied Mathematics. p. 606–610.

[DOE] Department of Energy. 2021. Philippine Energy Situationer. Taguig City, Philippines. Retrieved from https://www.doe.gov.ph/sites/default/files/pdf/energy_statistics/doe-pes-kes-2021.pdf?withshield=1

FARIAS-ROCHA AP, HASSAN KMK, MALIMATA JRR, SÁNCHEZ-CUBEDO GA, ROJAS-SOLÓRZANO LR. 2019. Solar photovoltaic policy review and economic analysis for on-grid residential installations in the Philippines. Journal of Cleaner Production 223: 45–56.

GUECHI A, CHEGAAR M, MERABET EA. 2011. The effect of water vapor on the performance of solar cells. Physics Procedia (21): 108–114

GOH GB, HODAS NO, VISHNU A. 2017. Deep learning for computational chemistry. Journal of Computational Chemistry 38(16): 1291–1307.

HAIGH JD. 2003. The effects of solar variability on the Earth's climate. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences 361(1802): 95–111.

KAYRI M, KAYRI I, GENCOGLU MT. 2017. The performance comparison of multiple linear regression, random forest, and artificial neural network by using photovoltaic and atmospheric data. In: 2017 14th International Conference on Engineering of Modern Electric Systems (EMES): IEEE. p. 1–4.

KIM J, RABELO M, PADI SP, YOUSUF H, CHO EC, YI J. 2021. A review of the degradation of photovoltaic modules for life expectancy. Energies 14(14): 4278

KIM TK. 2015. T test as a parametric statistic. Korean Journal of Anesthesiology 68(6): 540–546.

KIM YS, JOO HY, KIM JW, JEONG SY, MOON JH. 2021. Use of a big data analysis in regression of solar power generation on meteorological variables for a Korean solar power plant. Applied Sciences 11(4): 1776

KREPS B. 2020. The Rising Costs of Fossil-fuel Extraction: an Energy Crisis That Will Not Go Away. American Journal of Economics and Sociology 79(3): 695–717.

KROESE DP, BOTEV ZI, TAIMRE T, VAISMAN R. 2019. Data science and machine learning: mathematical and statistical methods. Chapman and Hall/CRC.

LITTLE RJ, RUBIN DB. 2019. Statistical analysis with missing data (Vol. 793). John Wiley & Sons. p. 4.

MARTÍNEZ-SÁNCHEZ RA, RODRIGUEZ-RESENDIZ J, ÁLVAREZ-ALVARADO J M, MACÍAS-SOCARRÁS I. 2022. Solar Energy-based Future Perspective for Organic Rankine Cycle Applications. Micromachines 13: 944.

MEKHILEF S, SAIDUR R, KAMALISARVESTANI M. 2012. Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. Renewable and Sustainable Energy Reviews 16(5): 2920–2925.

MONDAL MAH, ROSEGRANT M, RINGLER C, PRADESHA A, VALMONTE-SANTOS R. 2018. The Philippines energy future and low-carbon development strategies. Energy 147: 142–154.

MOUSAVI SM, MOSTAFAVI ES, JAAFARI A, JAAFARI A, HOSSEINPOUR F. 2015. Using measured daily meteorological parameters to predict daily solar radiation. Measurement (76): 148–155.

NJOK A O, OGBULEZIE JC. 2018. The effect of relative humidity and temperature on polycrystalline solar panels installed close to a river. Physical Science International Journal 20(4): 1–11.

OZSAHIN DU, TAIWO MUSTAPHA M, MUBARAK AS, SAID AMEEN Z, UZUN B. 2022. Impact of feature scaling on machine learning models for the diagnosis of diabetes. In: 2022 International Conference on Artificial Intelligence in Everything (AIE); Lefkosa, Cyprus: IEEE. p. 87–94.

PASION CK. 2019. Modeling Power Output of Horizontal Solar Panels Using Multivariate Linear Regression and Random Forest Machine Learning [MS Thesis]. Air Force Institute of Technology. p. 2348.

SAURE LA, DAHIPON RT, ASUNCION VA, SIMON RC. 2022. Regression analysis of solar power generation with meteorological variables. In: Proceedings of the Samahang Pisika ng Pilipinas 40 [SPP-2022-PB-19].

SHAFIEE S, TOPAL E. 2009. When will fossil fuel reserves be diminished? Energy Policy 37(1): 181–189.

SHLENS J. 2014. A tutorial on principal component analysis. arXiv preprint [arXiv:1404.1100].

SMITH NJ. 2018. A Study of the Sensitivity of Solar Power Generation to Varying Weather Conditions [MS Thesis]. The University of North Dakota.

TOWER WS. 1903. The Climate of the Philippines. Bulletin of the American Geographical Society 35(3): 253–260.

[UN ESCAP] United Nations Economic and Social Commission for Asia and the Pacific. 2021. Asia and the Pacific's Progress towards Sustainable Development Goal 7. Bangkok, Thailand.

VANDERPLAS J. 2016. Python data science handbook: essential tools for working with data. O'Reilly Media, Inc.

VILLEGAS-MIER CG, RODRIGUEZ-RESENDIZ J, ÁLVAREZ-ALVARADO JM, JIMÉNEZ-HERNÁNDEZ H, ODRY Á. 2022. Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours. Micromachines 13(9): 1406.

WU YK, HUANG CL, PHAN QT, LI YY. 2022. Completed Review of Various Solar Power Forecasting Techniques Considering Different Viewpoints. Energies 15(9): 3320.

XIA Q. 2021. Study on The Relationship between Meteorological Factors and Photovoltaic Power Generation Efficiency and Influence Mechanism. In: IOP Conference Series: Earth and Environmental Science Sci.

ZIANE A, NECAIBIA A, SAHOUANE N, DABOU R, MOSTEFAOUI M, BOURAIOU A, KHELIFI S, ROUABHIA A, BLAL M. 2021. Photovoltaic output power performance assessment and forecasting: Impact of meteorological variables. Solar Energy 220: 745–757.

ZUNIGA-REYES MA, ROBLES-OCAMPO JB, SEVILLA-CAMACHO PY, RODRÍGUEZ-RESÉNDIZ J, LASTRES-DANGUILLECOURT O, CONDE-DÍAZ JE. 2021. Photovoltaic failure detection based on string-inverter voltage and current signals. IEEE Access 9: 39939–39954.