Predicting Solar PV Generation Using Weather Station Data

Jamil Al-Nouman[1(✉)] and Abdulmalek Al-Gahmi[2]

[1] Southeast New Mexico College, Carlsbad, NM 88220, USA
jal-nouman@senmc.edu
[2] Weber State University, Ogden, UT 84408, USA
aalgahmi@weber.edu

Abstract. Electric power generated from Solar Photovoltaic (PV) panels is estimated to have increased globally by 22% in 2019, to 720 TWh [5]. It is now considered the third-largest renewable energy technology after wind and hydro powers. The primary reason for this growth is the need to utilize free energy resources that are also environmentally clean. PV-generated power, however, is uncertain and varies from time to time and season to season. Dealing with this uncertainty requires having predictive and forecasting models that accurately estimate generated power from historical data. This paper reports on an in-progress research project that explores weather-related variables such as humidity, temperature, and wind speed and uses them to predict and forecast generated power using a dataset collected over three years by a weather station at Southeast New Mexico College in Carlsbad, New Mexico. Various predictive and forecasting models are built, trained, and evaluated. The goal is to explore these variables and report on what makes a good predictive model and how such a model behaves over time.

Keywords: Photovoltaic modules · PV generation forecasting · Weather station · Solar irradiance

1 Introduction

In 2020 despite the Covid 19 pandemic, solar Photovoltaic (PV) panel installation was up by 23% and the renewable energy market expanded by 45%: the highest growth rate since 1999 [8]. In addition, the US Energy Information Administration projects that solar PV modules installation will reach 46% of all the renewable energy sources by the end of 2022 [2]. Many factors contribute to this growth. Concerns about climate change and energy crises, for instance, have been linked to such substantial growth in solar power generation [1].

PV-generated power, however, suffers from uncertainty and varies from time to time and season to season. The solar irradiance is reliant on several geographic and atmospheric factors. The PV-generated power of a solar panel depends on its location and the weather conditions at that location. Important variables here include but are not limited

J. Al-Nouman and A. Al-Gahmi—These authors contributed equally to this work.

A. K. Ghosh and C. Rixham (Eds.): ASES SOLAR 2022, SPE, pp. 150–160, 2022.
https://doi.org/10.1007/978-3-031-08786-8_18

to temperature, humidity, wind speed, wind direction, and time of day [4, 9–11, 17]. The efficiency of a solar panel is also impacted by the aforementioned weather conditions. For example, as the temperature of the panel increases, its efficiency decreases. Similarly, as the wind speed increases, the panel cools down, and the efficiency increases [4, 10]. Being able to forecast PV generation based on historical data is an important way to deal with such uncertainty. It is also important for better power planning and management [3].

This requires creating and evaluating multiple predictive models. Indeed, many such models have been proposed that use weather-related data to predict solar intensity and/or PV generation. One such model tries to forecast PV generation utilizing site-specific forecasting models trained using data from the National Weather Service (NWS) [15]. Doing so resulted in a 27% improvement over the performance of existing forecasting models. Another model attempted to predict PV generation utilizing a combination of weather and PV system parameters [11]. The effect of wind speed and air velocity, measured by PV panel surface temperature at different angular positions, on the performance of solar PV modules has also been studied [4]. This experiment showed that as the panel's temperature drops, its efficiency and power output increases. Another study investigated how temperature and wind speed affect the PV module efficiency [10].

More studies have looked into the impact of temperature and wind speed on the performance of PV modules [9, 10, 17]. The effect of humidity has also been studied [14]. Humidity creates a minimal sheet or layer of water on the PV module surface and a concentration of water vapor in the air, which reduces solar radiation and causes it to be reflected away from the PV module surface. This leads to decreased PV productivity by 10%-20%. Many machine learning models have been used in these studies such as artificial neural networks (ANN) with mean absolute percentage error (MAPE) [13]; hidden Markov models and support vector machines (SVM) [12]; and an artificial neural network with a self-organizing feature map (SOFM) [19].

This paper 1) explores the relationships between weather-related variables and solar irradiance. It uses data collected over three years using a weather station installed at the campus of Southeast New Mexico College (formerly New Mexico State University - Carlsbad) in Carlsbad, New Mexico. It then 2) utilizes multiple machine learning models to predict solar irradiance given these weather-related variables. Treating the solar irradiance data as a time series, it 3) uses an additive forecasting model with logistic growth to forecast future solar irradiance values given historical ones.

The remainder of this paper is organized as follows. Section 2 describes the approach taken to satisfy the above three tasks. Section 3 presents the obtained results, and Sect. 4 concludes this paper and discusses future work.

2 Approach

The main goal of this paper is to show how weather conditions affect solar irradiance, which, in turn, affect power output. As stated in the last section, this paper focuses on three tasks. The first task explores how weather-related variables such as humidity, temperature, and wind speed affect solar irradiance. Unlike the studies cited in the last section, this paper also considers the impact that meteorological seasons have over these variables. Understanding how these variables affect solar irradiance is critical to being able to perform the second task. The second task involves the use of machine learning models to predict solar irradiance. As will be shown later in this paper, solar irradiance is also a good predictor of voltage and power output. The third task looks at the solar irradiance data as a time series and employs a forecasting model to forecast future values given historical ones. For brevity, we only forecast solar irradiance values. The same approach can, however, be applied to other variables such as voltage and power outputs.

The data used in this paper comes from a Campbell Scientific CR6 series weather station installed outside the campus of Southeast New Mexico College in Carlsbad, New Mexico since January 2019. This station is equipped with a solar PV panel and a data logger. The solar panel faces true south, with a fixed tilt of 30° angle for the whole year. This station collects information about the following variables:

- **timestamp** which is the date and time of when the data example or record is added to the data file.
- **relative humidity** as a percent (%). This is also related to the dew point temperature, also measured by the station.
- **temperatures** in °C. The station supports three temperatures: ambient, panel, and dew point. The panel and ambient temperatures are very similar, while the dew point temperature is more related to humidity. This paper uses only the panel temperature; this is the temperature of the surface of the PV panel.
- **wind speed** measured in m/s. Wind direction is also supported but not used in this paper.
- **solar irradiance**, which according to the National Renewable Energy Laboratory (NREL) is the “incident flux of radiant power per unit area expressed in W/m2” [18]. This is a key variable for the analysis and models of this paper. It is also different from irradiation, which is irradiance integrated over time and expressed in kWh/m2.
- **voltage output** measured in Volts.

The station reads these variables every minute. These readings, however, are not recorded directly to the data file. They are combined (averaged) and recorded every fifteen minutes. In addition to this fifteen-minute frequency data file, the station also provides hourly, daily, and monthly data files. Only the fifteen-minute and hourly data files are used in this paper. In addition to the above variables, three calculated variables are utilized: the power output of the PV module, its power input, and its efficiency. The power output is calculated based on the formula:

$$P_{out} = \frac{V^2}{R}$$

where V (in Volts) is the voltage output and R (in Ω) is the load connected to the module (≈19Ω). We opted to calculate the power output this way, instead of descaling the solar irradiance by an arbitrary factor, because the load is known, and we have direct measurement of the voltage output. The power input is the same as solar irradiance adjusted for the area of the PV panel in (m2). In other words:

$$P_{in} = Irradiance \times Area$$

The efficiency of the PV panel is the calculated using the formula:

$$Efficiency = \eta = \frac{P_{out}}{P_{in}} \tag{1}$$

Two additional variables are extracted from the timestamp variable: time of day and meteorological season. These variables tell how solar irradiance changes from time to time and season to season.

As is typical with data files, the files produced by this station require cleanup and preprocessing. Some examples have `inf` and `NAN` solar irradiance values. These examples are removed from the dataset. For prediction, all data examples with negative irradiance or negative humidity values are also removed. The daily time points are extracted from timestamps, converted to numeric, and used as an input variable. The data examples outside the time period of 5:00 am to 9:00 pm, where solar irradiance is around zero, are removed from the dataset. Moreover, the data examples where the calculated efficiency is more than 100% are removed. For time series forecasting, only the timestamp and solar irradiance variables are required. In addition, only the data examples where solar irradiance is `inf` or `NAN` are removed. This is to avoid disturbing the time series or adding gaps to it.

This paper uses data collected over a period of more than 3 years (January 2019 to April 2022). The first exploration task is done using the Pandas and NumPy Python packages. The second prediction task uses Scikit-learn [7]. While there are many machine learning models to use, we only report the results of four models: linear regression, k-nearest neighbors, decision tree, and random forest. Not all machine learning models are the same, however. Some models such as linear regression or decision trees are simpler than others such as radial basis function networks (RBF) or artificial neural networks (ANN) [6]. Simpler models tend to be easier to understand and produce results that are easier to visualize and interpret. This is why the aforementioned four models were chosen by this paper; they are all simple models to understand, interpret and visualize. Finally, the Python Prophet package [16] is used for the third forecasting task. The next section elaborates more on these models and presents their obtained results.

3 Results

3.1 Data Exploration

First, we explore how humidity, temperature, and wind speed relate to solar irradiance per season. Figure 1 does not show a clear relationship between humidity and solar irradiance except for the fact that large humidity values seem to correlate with lower irradiance values, which suggests the existence of a negative relationship. There is no clear relationship between wind speed and solar irradiance either. There seems to be, however, a positive relationship between temperature and solar irradiance. In addition, summer and spring data points coincide together while fall/winter points do the same. Dividing seasons into these two groups is a better fit for the actual weather pattern of Carlsbad, New Mexico, where it is more like two seasons than four. Figure 2 shows the same data as above but for two season groups: spring/summer and fall/winter.

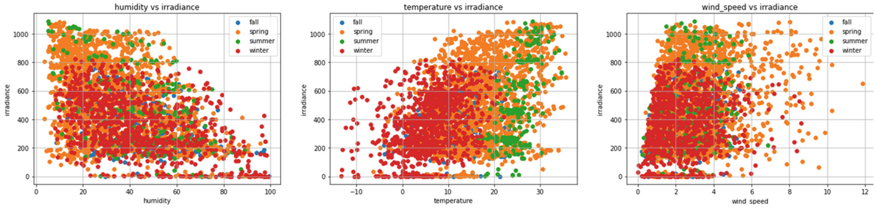

Fig. 1. Humidity, temperature, and wind speed vs solar irradiance per meteorological season

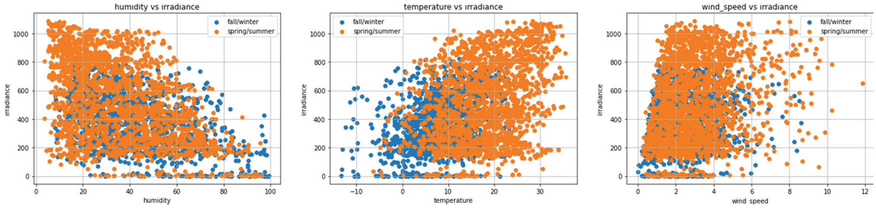

Fig. 2. Humidity, temperature, and wind speed vs solar irradiance using two season groups

Figure 3 depicts how these variables behave on average over a 16-h (5:00AM to 9:00PM) period for the two-season groups. Humidity, as expected, is high in the morning before it decreases throughout the day. Temperature and wind speed, on the other hand, increase throughout the day before they decrease at night.

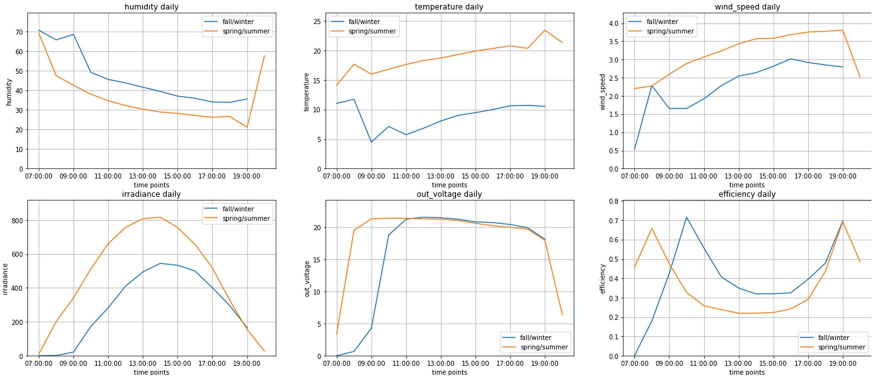

Fig. 3. Humidity, temperature, wind speed, solar irradiance, output voltage, and efficiency averaged over a 5:00 AM to 9:00 PM time period.

Figure 3 also shows how solar irradiance, voltage output, and efficiency behave on average throughout the same time period. The solar irradiance curves are bell-shaped. The spring/summer curve is earlier, wider, and higher than the fall/winter curve. It corresponds to longer days with lower humidity curves, and higher temperature and wind speed curves. The spring/summer voltage output curve is earlier than and peaks and flattens at the same level as the fall/winter curve. The efficiency curves are, however, interesting. They increase in the morning before they change direction and decrease. At midday they bottom out at about 22% for spring/summer and 32% for fall/winter. This can be attributed to the interplay between temperature and wind speed and the effect of that on solar irradiance. Higher temperatures decrease the efficiency of the PV module while stronger winds cool the PV module down and increase efficiency.

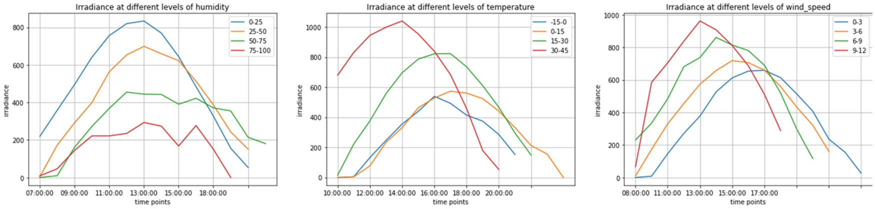

Fig. 4. Solar irradiance at different levels of humidity, temperature, and wind speed

In addition, we can have a better picture of how humidity, temperature, and wind speed affect solar irradiance by drawing solar irradiance curves at different values of these variables. To make this manageable, we divide the range of these variables into four levels and draw a solar irradiance curve at each level. Figure 4 shows these curves. It is clear from this figure that higher levels of humidity correspond to lower solar irradiance, and higher levels of temperature and wind speed correspond to higher solar irradiance.

In summary, the humidity, temperature, and wind speed variables affect solar irradiance and the efficiency of the PV module. Next, we see if these weather-related variables are good predictors of solar irradiance. We also evaluate how good a predictor solar irradiance is of the voltage output.

Table 1. Performance of multiple solar irradiance-predicting models with different dataset configurations (Hmd = Humidity, Tmp = Temperature, Wnd = Wind Speed) tables.

Predictor	Hmd	Tmp	Wnd	Hmd & Tmp	Hmd & Wnd	Tmp & Wnd	All
Linear regression	0.27	0.22	0.10	0.31	0.29	0.23	0.31
K-Nearest neighbors	0.55	0.62	0.29	0.67	0.56	0.59	0.70
Decision tree	0.65	0.70	0.45	0.70	0.63	0.67	0.73
Random forest	0.47	0.50	0.40	0.49	0.47	0.48	0.46

3.2 Making Predictions

As mentioned before, this paper uses four machine learning regression models for their simplicity, understandability, and interpretability. These models are linear regression, k-nearest neighbors, decision tree, and random forest. Before the results of these four models are presented, we briefly describe them. Linear regression is a parametric model for learning the parameters of a linear equation of the form:

$$y = f(X) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m$$

where $x_1, x_1, \cdots, x_m$ are the input variables, y is the output variable, and $w_0, w_1, \cdots, w_m$ are the parameters learned by the model. The goal is to assign these parameters values so as to minimize the mean square error (MSE). In addition to its simplicity, linear regression is robust against overfitting, which is the process of overexposing the model to the training data. Once trained, the equation above is used to calculate the output values of unseen data examples.

The k-nearest neighbors is a non-parametric model that predicts output values based on the mean of the values of its nearest neighboring points. The number of neighbors is a parameter used to control for overfitting.

Decision tree is another non-parametric model that uses the data to learn a set of if-then-else rules, which effectively divide the input space into multiple partitions. It is a simple model to understand, interpret, and visualize. It can, however, easily overfit the data, and we use the maximum depth parameter to constrain the tree and guard against overfitting.

Random forest is an ensemble method that applies a technique called bagging or bootstrap aggregation to decision trees. The idea here is to take multiple bootstrap samples from the given dataset and to train a decision tree using each sample. It also uses a technique called feature bagging, which requires that only a randomly selected subset of the input variables is considered at each node during the construction of the decision trees. The predicted value of an unseen example is the average of all the output values predicted by all the trained trees.

The described four models are used in this paper to see if humidity, temperature, and wind speed can accurately predict solar irradiance. We trained these models using all combinations of these variables as input and solar irradiance as output. All these models resulted in very low R^2 scores (0.0–0.13). This is because one important variable is missing from these models. That variable is the time of day, which is needed to tell the models when solar irradiance is low and when it is high. Once added to these models, the scores improved significantly.

Table 1 summarizes the performance of these models. As can be seen from this table, the best performance happens when all three variables (humidity, temperature, and wind speed) plus time of day are considered. The best performing models with scores of .70 and .73 are the k-nearest neighbors (k = 3) and the decision tree (max-depth = 5), respectively. The linear regression model did not perform well, which suggests a non-linear relationship between input and output variables.

Predicting voltage output from solar irradiance is also investigated. The same models, as before, are used, and Table 2 shows their obtained scores. These scores indicate that solar irradiance is a good predictor of voltage output. It is also a good predictor of power output, because power output is calculated using voltage output. The linear regression model performed the worst here as well, which again suggests a non-linear relationship between these variables.

Table 2. Performance of multiple output voltage-predicting models.

Predictor	R-Squared
Linear regression	0.13
K-Nearest neighbors	0.92
Decision tree	0.96
Random forest	0.94

3.3 Forecasting Solar Irradiance

Lastly, forecasting future solar irradiance values given historical ones is investigated. This is motivated by the fact that solar irradiance is a time series with a pattern that repeats, with minor changes, every day. It also changes slightly from one season to another. Only the results of forecasting solar irradiance are presented in this paper. The same approach could be used for other variables such as voltage and power output.

To this end, an additive forecasting model that takes the form:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

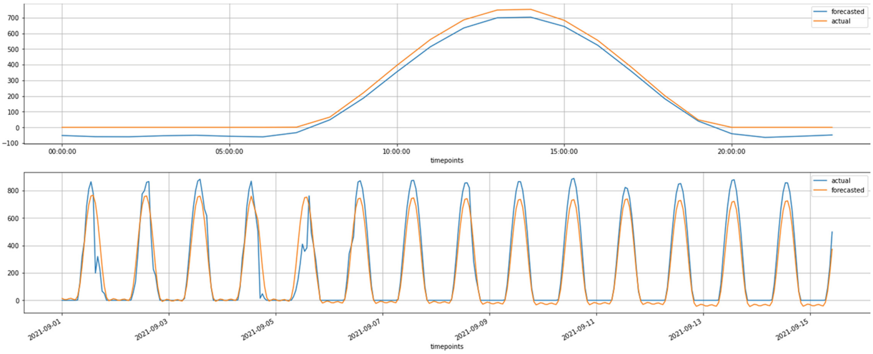

Fig. 5. Forecasted values vs actual values

where $g(t)$ is a growth function, $s(t)$ is a seasonality function, $h(t)$ is a holiday function, and ϵ_t is an error term [16]. The growth function $g(t)$ represents the trend of the data. The seasonality function $s(t)$, which uses a Fourier series as a function of time, represents the periodic changes, and the holiday function $h(t)$ represents the effect of holidays on the timeline. The error term ϵ_t represents any changes not captured by the model. In this paper, a logistic growth function with a saturating minimum of 0 and maximum of 1200 is utilized. The seasonality was automatically detected from the dataset. No holidays were specified.

The resulting model is a simple one with a single variable: solar irradiance. Figure 5 shows the results of such a model and compares forecasted values to actual ones. The top graph shows the forecasted future values against the actual ones averaged over a day. The bottom graph shows both values over a two-week period, immediately after the time series that the model was trained on. As Fig. 5 shows, both values are close with some peak differences. Parameter tuning and/or utilizing a more complex forecasting model is likely to improve these results. These improvements, however, are left as future work.

4 Future Work and Concluding Remarks

The scope of this paper is determined by the data provided by the weather station, which only supports a few weather-related variables (humidity, temperature, and wind speed) in addition to solar irradiance and voltage output. As has been shown, these weather-related variables impact solar irradiance, which, in turn, impacts generated power output. As cited studies have shown, these are not the only relevant variables. Missing from the dataset used in this paper are variables such as sky cloud cover, air quality, and precipitation potential, to name a few. These variables are not tracked by the weather station of this paper. There is a need to improve the performance of the models presented in this paper by augmenting the data collected by this weather station with outside datasets such as the ones provided by the Carlsbad Environmental Monitoring & Research Center and/or the National Weather Service (NWS). Additional predictive models could then be evaluated using the combined data.

In addition, forecasting models can also be used to predict the future solar irradiance, voltage, and/or power output values given historical data. The forecasting model presented in the paper is a simple one with a single variable. More complex forecasting models that utilize a time series with more than a single variable are yet to be investigated.

The weather conditions at the locations where solar panels are installed add uncertainty to their generated power output, which also varies from time to time and season to season. Predictive and forecasting models such as the ones presented in this paper, are useful tools to cope with this uncertainty. They are also important for better power planning and management.

In summary, this paper explored the relationships between weather-related variables and solar irradiance. Such exploration is critical to understanding how these variables interact with and affect the PV-generated power. It then utilized multiple machine learning models to predict solar irradiance given these weather-related variables. The performance of these models varies from one model to another, which suggests a complex relationship between input and output variables that some models capture better than others. The fact that the linear regression model performs the worst suggests a non-linear relationship between these variables. Finally, the paper utilized an additive forecasting model with logistic growth to forecast future solar irradiance values given historical ones. This model yields good results that can still be further improved by tuning and adding more variables.

Acknowledgements. The HSI Grant Services at Southeast New Mexico College (formerly New Mexico State University - Carlsbad) provided the weather station utilized in this study. The authors would like to acknowledge this support.

References

1. Is the energy crisis bad for climate change investors? (2021). https://www.schroders.com/en/us/insights/equities/is-the-energy-crisis-bad-for-climate-change-investors/
2. Solar power will account for nearly half of new U.S. electric generating capacity in 2022 (2022). https://www.eia.gov/todayinenergy/detail.php?id=50818
3. Ahmed, A., Khalid, M.: A review on the selected applications of forecasting models in renewable power systems. Renew. Sustain. Energy Rev. **100**, 9–21 (2019). https://www.sciencedirect.com/science/article/pii/S1364032118306932
4. Ali, M., et al.: Performance investigation of air velocity effects on PV modules under controlled conditions. International Journal of Photoenergy 2017 (2017)
5. Bahar, H., Bo jek, P.: Tracking solar PV 2020 (2020). https://www.iea.org/reports/tracking-solar-pv-2020
6. Bishop, C.M.: Pattern Recognition and Machine Learning. ISS, Springer, New York (2006). https://doi.org/10.1007/978-0-387-45528-0
7. Buitinck, L., et al.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122 (2013)
8. Cappell, B.: Renewable energy growth rate up 45 ‘new normal’ (2021). https://www.npr.org/2021/05/11/995849954/renewable-energy-capacity-jumped- 45-worldwide-in-2020-iea-sees-new-normal
9. Fesharaki, V.J., Dehghani, M., Fesharaki, J.J., Tavasoli, H.: The effect of temperature on photovoltaic cell efficiency. In: Proceedings of the 1stInternational Conference on Emerging Trends in Energy Conservation–ETEC, Tehran, Iran, pp. 20–21 (2011)
10. Khan, A.B.: Effect of temperature and wind speed on efficiency of PV module (2020). https://doi.org/10.13140/RG.2.2.28345.93282
11. Kumar, A., Rizwan, M., Nangia, U.: Artificial neural network based model for short term solar radiation forecasting considering aerosol index. In: 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), pp. 212–217 (2018). https://doi.org/10.1109/ICPEICES.2018.8897290
12. Li, J., Ward, J.K., Tong, J., Collins, L., Platt, G.: Machine learning for solar irradiance forecasting of photovoltaic system. Renew. Energy **90**, 542–553 (2016). https://doi.org/10.1016/j.renene.2015.12.069, https://www.sciencedirect.com/science/article/pii/S0960148115305747
13. Munir, M.A., Khattak, A., Imran, K., Ulasyar, A., Khan, A.: Solar pv generation forecast model based on the most effective weather parameters. In: 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1–5 (2019). https://doi.org/10.1109/ICECCE47252.2019.8940664
14. Panjwani, M., Narejo, G.: Effect of humidity on the efficiency of solar cell (photo-voltaic). Int. J. Eng. Res. Gen. Sci. **2**, 499–503 (2014)
15. Sharma, N., Sharma, P., Irwin, D., Shenoy, P.: Predicting solar generation from weather forecasts using machine learning. In: 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 528–533. IEEE (2011)
16. Taylor, S.J., Letham, B.: Forecasting at scale. Am. Stat. **72**(1), 37–45 (2018)
17. Veldhuis, A., Nobre, A., Reindl, T., Rüther, R., Reinders, A.H.: The influence of wind on the temperature of PV modules in tropical environments, evaluated on an hourly basis. In: 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), pp. 0824–0829. IEEE (2013)

18. Walker, A., Desai, J.: Understanding solar photovoltaic system performance (2021). https://www.energy.gov/sites/default/files/2022-02/understanding-solar-photo-voltaic-system-performance.pdf
19. Yousif, J.H., Kazem, H.A., Boland, J.: Predictive models for photovoltaic electricity production in hot weather conditions. Energies **10**(7) (2017). https://www.mdpi.com/1996-1073/10/7/971