Modelling of photovoltaic system power prediction based on environmental conditions using neural network single and multiple hidden layers

R Azka¹, W Soefian¹, D R Aryani¹,², F H Jufri¹,², and A R Utomo¹,² *

¹ Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
² Electric Power and Energy Studies (EPES), Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia

*E-mail: arutomo@eng.ui.ac.id

Abstract. The solar power plant is an alternative to the provision of environmentally friendly renewable electricity, especially in the tropics, which are sufficiently exposed to the sun throughout the year. However, environmental conditions such as rainfall, solar radiation, or clouds may affect the output power of photovoltaic (PV) systems. These factors make it difficult to know whether PV can meet the needs of the existing load. This research develops a model to predict the output power of a 160 x 285W PV system located in the tropics and has certain environmental conditions. The prediction development is supported by the Python programming language with a single hidden layer and two hidden layers Neural Network, as well as the traditional Multiple Linear Regression tools. The simulation results show that the two hidden layers Neural Network method has a higher level of accuracy compared to the single hidden layer and Multiple Linear Regression as seen from the value of R², MSE, and MAE.

1. Introduction

The total world energy consumption is projected to increase by 40% from 2019 to 2050 [1]. Currently, most of the energy used is still fossil-based energy, such as petroleum and coal, which will be limited in the future. Moreover, there are also disadvantages caused by the energy produced by fossil fuels to the environment [2]. Therefore, the development of technology related to energy production has begun to shift from fossil-based energy towards new and renewable energy, such as solar energy.

The energy potential possessed by solar energy is also enormous, especially in a tropical area that is exposed to the sun throughout the year. For instance, in Indonesia, it is estimated that more than 500 GW and the largest compared to other renewables resources the country has. This case also occurs in other tropical countries, such as Malaysia and Thailand [3][4]. However, the development of the solar power plant is still growing in Southeast Asia [5]. Despite its abundant resources and environmental benefit, solar PV systems rely on atmospheric conditions, such as irradiation, temperature, cloud condition, and others. It is therefore important to be able to predict the power output from the solar power system for planning the operation of the power plant to ensure the reliability and cost-effectiveness of the system. The output parameters of a PV system usually have Standard Test Conditions (STC). However, the actual atmospheric state is rarely the same compared to STC, which results in the output of PV having a significant difference to the rated value [6].
Recently, experiments and literature have tried various methods to estimate the value of solar PV output power. Generally, the methodology to predict solar PV output power can be divided into the physical approach and the statistical approach [7]. Physical modelling estimates solar PV output by focusing on the physical construction of PV cells. Then the output will be matched with the available parameters. However, this method is very complicated and it is difficult to ensure the robustness of predictions due to the diversity of parameters [8]. Besides, this method is usually only concentrated on deterministic predictions, which sometimes fail in evaluating uncertainty from PV power data [9]. On the other hand, the statistical prediction method depends on the historical dataset to develop the model that can be modified to get maximum results.

Modelling with statistical analysis methods can be assisted with the Machine Learning algorithm by employing the weather and PV power output data, modelling can be made from various statistical methods. The statistical analysis can be divided into indirect and direct method [10]. The indirect method uses the historical data of solar radiation which is considered proportional to the PV power generation. For example, a study used support vector regression (SVR) for very short-term forecasting horizons of 5 to 60 minutes ahead in Australia use weather data and power output with data pre-processing [11], the author uses two types of experiments, experiments without weather variables and using weather variables. Another example was proposed by Shi et al, where they classified the day into four types namely (clear-sky, cloudy, foggy, and rainy) and used data at 15-minute intervals [12], it can be seen from the literature that weather is very influential in producing solar PV output power.

On the other hand, the direct method uses the records of the PV output power to develop the prediction model [13]. Artificial Neural Network (ANN) is a method that is often used to make predictions of PV output because of its ability to process variables that are not linear-shaped [12]. Alan J Thomas discussed the impact of the number hidden layer for the prediction of solar PV output, resulting that two hidden layers with fewer datasets are usually better than single hidden layer with more datasets [14]. Multiple linear regression (MLR) can also be used to make predictions of PV output [15]. However, due to its inability to analyze non-linear variables, this method is not as effective as the ANN method and usually produce a lower accuracy [9]. However, MLR can be useful to find the significance between the weather variables and the PV output power by looking at the parameters of the MLR equations.

This paper proposes a methodology by using ANN as core prediction model which is supplemented by MLR algorithm to pre-process the datasets. The MLR algorithm is employed to select weather variables and can be used to find the outliers that can significantly affect the forecasting accuracy. Nevertheless, outliers may also be required for the operation planning of the solar PV system because it represents a real application [16].

The rest of the paper is structured as follows. In section 2, we will explore the dataset, the theoretical basis of the PV itself, the methodology used to make predictions, namely Multiple Linear Regression and Artificial Neural Networks, modelling of each method will also be made in this chapter. Section 3 will discuss the results of the training and predictions of the two methods obtained and their analysis. Section 4 will summarize the main results and conclude this study.

2. Development of forecasting model
2.1. Data description & pre-processing

The PV system used for the experiment was 160 of 285 Watt PV panels with total nominal power of 45.6kW, located in the sunny tropical region throughout the year, the northern latitude of Indonesia, at longitude 100.51° and latitude 13.75° with an altitude of 13 meters above ground level. The PV has a tilted array of 13°, facing south at 0°, and no shade.

The input-output relationship of the prediction model is presented in Figure 1. The input variables consist of environmental data, while the output is the predicted power. The environmental variables are obtained from www.meteoblue.com with variable data taken at each hour [17]. These variables include temperature, humidity, pressure, rainfall, cloud cover, duration of solar radiation, wind speed, and wind direction. Another variables is incorporated to the model, that is solar altitude angle, which can be calculated as follows [18]:

\[
\sin \beta = (\cos L \times \cos \delta \times \cos H) + \sin L \times \sin \delta
\] (1)
where \( L \) is the site latitude, \( \delta \) is the solar declination, and \( H \) is the hour angle. Adjustments are required to calculated \( H \) so the model can be applied for the local time, such as longitude and local time meridian.

The process of the model development is presented in Figure 2. The data is processed by using a computer programming language, Python, to simplify and to speed up the work. The earliest stage of the model development is to preprocess the variable data by analyzing the mean, median, and quartile in accordance with the output power.

Firstly, clean up the data that is needed in making predictions. Then, analyze the data and visualize each variable against the target to show the relationship. To find multicollinearity, Variance Inflation Factor (VIF) is the quotient of the variance in the model many variables with variants from one model with only one variable. Furthermore, what needs to be done is to look for outliers, which can be solved by calculating the Z-Score. Then the data is randomized and normalized in order to differentiate the bias of the model. Data is divided into two types, namely data train and test data, where 80% of the data train is from the whole dataset, and the test data is 20% from the available dataset. Test data serves to check whether the modelling made is good enough or not.

Multiple Linear Regression (MLR) can be used as part of pre-processing data because its linear relationship between the input variables and the output. The general form of the MLR is shown in eq.2.

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e
\]  

where \( y \) is dependent variable such as the PV power generation, \( \beta_0 \) is intercept, \( \beta_n \) is slope coefficients for each explanatory variable, \( x \) is explanatory variables such as weather variables and altitude angle, and \( e \) is error term.

### 2.2. Artificial Neural Network (ANN) model

Artificial Neural Network (ANN) has in fact the same core general form as regression, but ANN has the capability to interpret the non-linear relationship between input and output so that it can result in a better accuracy. ANN is a network of a group of small processing units that are modeled based on
the behavior of human neural networks [19], as shown in Figure 3. Each neuron is interconnected, and the information flows from each of these neurons. Each data is processed through weighted dot operation, added up to form a weighted sum along with a certain bias. Then, it is transferred via an activation function to generate an output. Figure 4 provides a conceptual configuration of a single hidden layer ANN and a double hidden layer ANN.

The ANN model consists of one input layer, one or two hidden layers, and one output layer. In the case of two hidden layers, the first hidden layer uses Rectified Linear Unit (ReLu) and hyperbolic tangent (tanh) as its activation function, and the second hidden layer uses linear activation function. ReLu and tanh allows the model to learn non-linear correlations between the weather variables and PV output. Whereas, the linear activation function ensures that the range of output is mimicking the actual PV output characteristic.

The parameters that represent the ANN model are obtained by using Root Mean Square Propagation (RMSProp) optimization to calculate the momentum in order to avoid pitfall in gradient descent. The learning rate that is used is 0.01 to make sure the expected model even though it will prolong the computation time. The validation data is required to prevent overfitting, which uses as much as 20% of the randomly spread training data.

2.3. Evaluation of forecasting model

The accuracy of the forecasting model is evaluated by using three criteria, such as $R^2$, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). $R^2$ is a statistical calculation that represents the proportion of variance for a dependent variable predicted using the independent variable in the regression model [20]. MAE is the average magnitude of error of a set of predictions regardless of the direction, whereas RMSE is a squared assessment that is rooted.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \bar{y}_i|$$

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

3. Simulation results

MLR is used to create an equation that can show the significance of each variable entered. After making the selection, some variables that are collinear and insignificant are deleted. The results of $R^2$ and the resulting error values also will not change significantly. From the algorithm obtained, the parameter used for modelling are solar altitude angle, temperature, precipitation, total cloud cover, sunshine radiation, and speed of the wind.

The result for ANN two hidden layers, the best $R^2$ is to use the ReLu activation function on the first hidden layer, then use the Linear activation function on the second hidden layer, with 15 epochs, and 200+200 neural. For a single hidden layer, the best $R^2$ obtained using ReLu as activation function, 150 epoch, and 400 neural.

![Figure 5. Scatter plot for ANN one hidden layer (a) before & (b) after eliminating outliers 2 hidden layer (c) before & (d) after eliminating outliers](image-url)
For a comparison between the method of neural network one hidden layer and two hidden layer, it can be seen from the table below.

| Method          | ANN one hidden layer | ANN two hidden layer |
|-----------------|----------------------|----------------------|
|                 | before               | after                | Before               | After               |
| \( R^2 \)       | 0.895                | 0.934                | 0.908                | 0.941               |
| RMSE            | 3585.4               | 2667.5               | 3042.6               | 2657.6              |
| MAE             | 2045.5               | 1561.9               | 1727.9               | 1477.9              |

After seeing these results, it is time to try modelling the predicted actual output power. The prediction is made in two days from a dataset that has never been used from both the training data and the test data at the same location and the value of the power output that has also been collected every hour using neural network one hidden layer and two hidden layers. The data set contains 48 data from 48 hours. And the results obtained are as follows.

![Figure 6](image)

It appears that the two hidden layers' neural network has a higher accuracy value, also with no significant difference. Both have the same number of neurons, but on a neural network, two hidden layers are able to have more parameters. This happens because with onetime training data in two hidden layers is the same as dozens of times training in one hidden layer. Thus, one hidden layer has far more training model than two hidden layers.

4. Conclusion
An accurate and robust PV output power prediction is considered as one of the crucial factors to enable a higher penetration of PV system into the grid. This paper proposes a prediction model based on ANN algorithm, and compares the significance of using single hidden layer or two hidden layers in the ANN model. This paper also suggests a method to pre-process the datasets in order to improve the accuracy of the prediction model by looking at the relationship between variables through the MLR analysis. As a result, the double hidden layer ANN model provides a lightly better accuracy compared to the single hidden layer ANN model. Moreover, eliminating the outliers delivers even a better accuracy for both single and double hidden layer ANN model. However, it is worth to note that the elimination of outliers might not be recommended in the case of the prediction model is used for the operational planning.

References
[1] Administration U S E I 2019 International energy outlook 2019 with projections to 2050 Choice Rev. Online 85
[2] Said S A M, Hassan G, Walwil H M, and Al-Aqeeli N 2018 The effect of environmental factors and dust accumulation on photovoltaic modules and dust-accumulation mitigation strategies Renew. Sustain. Energy Rev. 82 743–760
[3] Suphahitanukool C, Hunsacharoonroj I, Usapein P, Khedari J, Waewsak J, and Hirunalbh J 2018 An evaluation of economic potential solar photovoltaic farm in thailand: case study of polycrystalline silicon and amorphous silicon thin film Int. J. Energy Econ. Policy 8 33–41

[4] Belhamadia A, Mansor M, and Younis M A 2013 Assessment of wind and solar energy potentials in malaysia CEAT 2013 - 2013 IEEE Conf. Clean Energy Technol. 152–157

[5] D. Arinaldo J C A and P S Indonesia clean energy outlook 2018.

[6] Ramsami P and Oree V 2015 A hybrid method for forecasting the energy output of photovoltaic systems Energy Convers. Manag. 95 406–413

[7] Wang J, Ran R, and Zhou Y 2017 A short-term photovoltaic power prediction model based on an fos-elm algorithm Appl. Sci. 7

[8] Chen B, Lin P, Lai Y, Cheng S, Chen Z, and Wu L 2020 Very-short-term power prediction for pv power plants using a simple and effective rcc-lstm model based on short term multivariate historical datasets Electron. 9 289

[9] Khandakar A et al. 2019 Machine learning based photovoltaics (pv) power prediction using different environmental parameters of qatar Energies 12 2782

[10] Zhang G, Wang X, and Du Z 2015 Research on the prediction of solar energy generation based on measured environmental data Int. J. u- e-Service, Sci. Technol. 8 385–402

[11] Malvoni M, De Giorgi M G, and Congedo P M 2017 Forecasting of pv power generation using weather input data-preprocessing techniques Energy Procedia 126 651–658

[12] Van Tai D 2019 Solar photovoltaic power output forecasting using machine learning technique J. Phys. Conf. Ser. 1327 0120051

[13] U. R, A. J A M and M T A P 2016 Smart baseline models for solar irradiation forecasting Energy Convers Manag. 108 539–548

[14] Thomas A J, Petridis M, Walters S D, Gheytassi S M, and Morgan R E 2017 Two hidden layers are usually better than one Commun. Comput. Inf. Sci. 744 279–290

[15] Clack C T M 2017 Modeling solar irradiance and solar pv power output to create a resource assessment using linear multiple multivariate regression J. Appl. Meteorol. Climatol. 56 109–125

[16] Weisent J, Seaver W, Odoi A, and Rohrbach B 2014 The importance of climatic factors and outliers in predicting regional monthly campylobacteriosis risk in georgia, usa Int. J. Biometeorol. 58 1865–1878

[17] meteoblue.com Weather history download 2020

[18] Jufri F H and Jung J 2017 Photovoltaic generation forecasting using artificial neural networks model with input variables and model parameters selection algorithm in korea Int. J. Mach. Learn. Comput. 7 156–161

[19] Fogel D B, Liu D, and Keller J M Introduction and single-layer neural networks Fundamentals of Computational Intelligence, 2016 5–34

[20] Murray J M 2017 Variance decomposition in regression 1–7

Acknowledgments
This research is supported by research grant of PUTI (Publikasi Terindeks International) Prosiding 2020 from Universitas Indonesia (UI).