Forecasting approach for solar power based on weather parameters (Case study: East Kalimantan)

P Hasanah1*, S A Wiradinata2, and M Azka1

1Actuarial Science Study Program- Kalimantan Institute of Technology, Balikpapan-Indonesia
2Statistics Study Program- Kalimantan Institute of Technology, Balikpapan-Indonesia
primadina@lecturer.itk.ac.id

Abstract. Solar Energy is the most popular among several clean energies. As a tropical country, Indonesia has big opportunity to develop solar power, particularly in East Kalimantan which spans around the equator. Solar energy generation, however, is influenced by weather parameters which give uncertain values of the amount of the captured energy. Therefore, this research is conducted to overcome the effect of weather towards solar energy. The aim of this research is to examine the model for sun power forecasting based on the data. The Artificial Neural Network (ANN) and Multiple Linear Regression have taken as the approach models to determine energy forecasting. This study used five input variables; temperature, precipitation level, humidity, wind speed, and surface pressure, while the solar radiation was taken as the output variable. Moreover, the daily solar power and weather data from East Kalimantan has been taken along the period of 27th July 2018 – 28th July 2021. The result of this study showed that the RMSE of ANN was slightly similar with the multiple linear regression methods which were calculated by 160.26 and 160.46 respectively. However, the ANN is preferable to use in the solar energy forecasting since the tendency of nonlinearity of the climate data. Keywords: ANN, Regression, Solar Energy.

1. Introduction
The security of energy becomes one of the major concerns through the global development. Many approaches have conducted by the government to realize energy security in the future. Particularly, in Indonesia which is known for its numerous energy resources- notably coal, oil and gas, geothermal- yet it is facing energy crisis. Therefore, the government of Indonesia has established some regulations to overcome the future energy’s challenge that were Law No.30 of 2007 on Energy, Law No.30 of 2008 on Electricity, and Minister of Energy and Mineral Resources (MEMR) Regulation No. 50 of 2019 on the Utilization of Renewable Energy Sources for Electric Supply [1]. Regarding these regulations, the government of East Kalimantan supported by releasing Governor Regulation No. 27 of 2019 on General Electricity Plan for East Kalimantan. This regulation was arranged for guiding the development electricity security in East Kalimantan [2].

According to the International Renewable Energy Agency, the data shown that Indonesia has big opportunity to develop sustainable energy from solar photovoltaic, hydropower, bioenergy, geothermal, ocean wave power and wind [3]. Under the direction of the National Energy Policy of 2014 (Kebijakan Energi Nasional/KEN), the level of utilization of electrical energy from renewable sources is expected by 25% of total national energy consumption on 2025, while the utilization of fossil fuels will be decreased moderately by the same year [1]. Under this regulation, the Referral of Region Spatial Plans (RTRW) of East Kalimantan directed that the Province of East Kalimantan will produce future energy with high added value of environmentally friendly to improve their energy security. One of the leading
energy sources that will be developed is solar power. Based on the Regional Electricity General Plan (Rencana Umum Ketenagalistrikan/ RUKD) of East Kalimantan, solar energy will be projected as a source of electricity in East Kalimantan in the future [4].

However, the implementation of solar energy in Indonesia should deal with numerous problems ranging from the technology of installation to the cost of the project. Especially in East Kalimantan, the transmission of solar-energy electricity that will be generated into the grid should deal with the grid quality itself [5]. Moreover, the anomalous weather condition in Equator Areas will affect the solar energy generation. According to [6–8], the climate parameters have been causing the uncertain value of solar power that influenced to the cost of productions. Therefore, to measure how much the potential solar energy can be captured in East Kalimantan, a forecasting approach is required to predict the power.

Some weather parameters have argued to influence the solar irradiance. Air temperature at 2m, air relative humidity at 2m, water vapor, wind speed, and dust loading have analyzed by [6] to assess the impact on sun power potential in United Arab Emirates which is has large scales of desert. Some similar variables were also used by [9,10] to perform sun energy forecasting using European Data with some added investigation of geographical parameter’s impact to the sun power. Meanwhile, [11] also considered these weather measures to examine the solar generation over some regions in Indonesia.

This study compared two most popular approach to investigate the effect of weather condition to solar radiation, namely Artificial Neural Network (ANN) and Multiple Linear Regression. These methods have used among several researchers and perform adequate prediction of solar radiation. The ANN is supposed to be better to model the nonlinear data, while the regression is well known as an easy way to carry on causality problems. Since the weather parameters is assumed to bring impact to the solar power generation, both these approximations will be analyzed in this research.

2. Method
This section defines the methodology proposed to develop analytical models of solar energy potential. The popular methods, Multiple Linear Regression and Artificial Neural Network will be elaborated in this section.

2.1. Artificial Neural Network
An ANN is a system of connected computers whose objective is to replicate the fabric of neurons in the brain. A human brain is capable of perception thanks to biological neurons that absorb multi-sensory information from the outside world, change it to electrical signals capable of travelling through our nervous system, and transmit the same to the brain for our understanding.

The structure of ANN is shown in Figure 1, the ANN’s dept indicates the number of layers. While the ANN width indicates the number of units in the layer

![Figure 1. The General Structure of ANN](image-url)

The Neural Network method which consists of Multilayer Perceptron is commonly referred to as a feedforward neural network. The important step in the ANN is to build the structure of the model and calculate the weight of each node in every layer. This weight will impact the output value of the model. This is the illustration of the weight relation in a structure of ANN.
Mathematically, the output value is influenced by the weight of input and hidden layer, whereas
\[ y_{in} = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_mx_m \] (1)

From the Eq.1 the net input, \( y_{in} \), represents the accumulation of the scalar product of the weight \( (w_i) \) and the input vector \( (X_i) \). Meanwhile, \( i = 1,2,\ldots,m \) denotes the number of the input variables. In addition, the output can be calculated by applying the activation function over the net input which defines the output of a neuron in terms of a local induced field. Activation functions are a single line of code that gives the neural nets non-linearity and expressiveness[12].

\[ Y = F(y_{in}) \] (2)

The steps to build the ANN methods is illustrated as the flowchart below [9].

![Flowchart](image)

**Figure 3. Stages to build an ANN**

The ANN accuracy of the model is measured by calculating the Mean Square Error (MSE) or Root Mean square Error (RMSE) by Eq.3 as follows,

\[ MSE = \frac{1}{n} \left( \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right) \] (3)

Where \( y_i \) denotes the real test data, and \( \hat{y}_i \) is the prediction value of output variable.

### 2.2. Multiple Linear Regression

The Multiple linear regression, also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Given several independent variables \( X_i \), the dependent variable \( Y \) is formulated as,

\[ Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n \] (4)

Where \( \beta_i \) represents the slope coefficient for each independent variable. This method used the coefficient determination \( R^2 \) to measure how much of the variation in outcome can be explained by the variation in the independent variables. Furthermore, the MSE or RMSE can be applied also to measure the accuracy.
3. Data and Variables

3.1. Data Source
This research used the data from the province of East Kalimantan which was observed at the latitude of 0.538659 and the longitude at 116.419389. The data is derived from the website of Nasa SSE and Solcast [13] starting from 27 July 2018 to 29 July 2021. The solar and weather data used in this study was measured at 11-12 AM which was assumed as the highest point of solar radiation in East Kalimantan.

3.2. Data Variables
The objective of this research is to approximate the solar forecast from the daily data using two approach models. The dependent variable is Solar Power which is measured as the global horizontal irradiance (GHI). GHI (W/m²) is the amount of terrestrial irradiance falling on a surface horizontal to the surface of the earth, whereas solar irradiance is the power per unit area received from the sun in the form of electromagnetic radiation as measured in the wavelength range of the measuring instrument[14]. The data of GHI is given in the time series plot, and the heterogeneity description of the data is illustrated in the box plot.

![Figure 4. Hourly data of global horizontal irradiance](image)

From Figure 4, the GHI data performed the seasonal pattern each day, while the boxplot is illustrated that the data of GHI was heavy in the right tail which it has quite wide range at about 992 point. In this research the prediction of solar potential was assumed to be measured at the average of peak hour each day, that is at 11.00-12.00 AM. Therefore, the data in this period has chosen to build the estimation model.

In addition, the independent variables used in this research are described both in Table 1 and Figure 5. Among five independent variables, the precipitation variables figure out the high range of data with high value of the upper outlier. In contrast, the other variable appeared the difference pattern.

| Variable                        | Unit  |
|---------------------------------|-------|
| Air Temperature                 | °C    |
| Precipitation Total             | mm    |
| Wind Speed range at 50 m        | m/s   |
| Surface Pressure                | Pa    |
| Specific Humidity at 2 m        | g/Kg  |
In the first step, it will advantageous to measure the coefficient of correlation between the variables. Since the model will be built to estimate the causality prediction among the independent and dependent variables. The correlation matrices between these variables are provided in Figure 6 which it showed that only temperature have had the quite strong correlation with the GHI and the other variables have had lower correlation. However, the ANN model is supposed to give a best approach although the data had nonlinear relationship.

![Boxplot of Independent variables](image1)

**Figure 5.** Boxplot of Independent variables.

**Figure 6.** Coefficient of Correlation matrix

4. **Modelling Process**
   In general, the process of modelling is shown in this flowchart:

![Flowchart diagram for modelling process](image2)

**Figure 7.** Flowchart diagram for modelling process
4.1. Data Preparation
For every forecasting method, data preparation is the essential step. In this stage, the data is prepared to be modelled both in Artificial Neural Network and Multiple Linear Regression. First of all, the normalization is taken to scaling the data. This process is important before training the data because it ensures the data to be within certain range or scale then this can guarantee stable convergence of weight and biases. In this study, min-max normalization was taken to the data since it is the simplest alternative to transform the data into range (0, 1) [15]. The histogram of the scaling data is given in Figure 8, where it draws the distribution of the dataset. These histograms have shown that some variable have skewed distribution, however both of Wind Speed and Specific humidity were supposed to follow normal distribution.

![Histogram for Scaled Variables](image1)

**Figure 8. Histogram for Scaled Variables**

After the normalization process, the data has divided into training and testing data with the proportion of 85% and 15%, respectively. The resampling method has applied to divide between training and testing. In the end, the training data will be used to model ANN and the regression.

4.2. Artificial Neural Network Model
The ANN model has the input layer, a hidden layer, and the output layer. The input layer consists of five nodes that are the independent variables. To decide the hidden layer, several things have considered. To begin with, the number of nodes in hidden layer estimated by 2/3 of the size of the input variables. Furthermore, it should be less than twice of the number nodes in input layer. And the last is the size of hidden layer should be between the size of input and output layer[16]. Using these assumptions, in this research, several models were analyzed in the same proportion of training and testing data. To estimate the weight for each node in the model and the accuracy of the model, a package of neuralnet in R has used. These models illustrated in the Table 2 below.

| The Structure of ANN | MSE      | RMSE    |
|----------------------|----------|---------|
| One Hidden layer- 2 nodes | 26769.35 | 163.613 |
| One Hidden Layer- 3 nodes | 25682.56 | 160.26  |
| One Hidden Layer- 4 nodes | 29125.1  | 170.66  |
| Two Hidden Layer – 2 and 1 nodes | 26138.24 | 161.673 |
| Two Hidden Layer – 3 and 1 nodes | 26101.4  | 161.56  |
| Two Hidden Layer – 3 and 2 nodes | 29501.05 | 171.76  |
| Two Hidden Layer – 3 and 3 nodes | 27968.77 | 167.24  |
These networks modelled by the train data which consists of 85% sample data, while feedforward neural network was chosen to illustrate the process. Feedforward neural network is a network which is not recursive. Neurons in the input layer were only connected to hidden layer then it was forwarded to the output layer. Since the neural network did not set as classification, the activation function should not be applied to the output neurons yet the linear output was settled. Moreover, the backpropagation has carried out to measure the accuracy of the model. The accuracy metrics MSE and RMSE have used to analyze the performance of the model.

From table 2, it can be seen the comparison of the MSE and RMSE for each model. It can be inferred that when the nodes in this single layer was decreased, the MSE increased. Furthermore, when the second hidden layer was added to the structure, the MSE was got higher. Therefore, it was decided to choose the model which has the lowest of error, that is the model with one hidden layer and three nodes. Hence, the architecture of ANN is determined in Figure 9.

The quantity of weight in every node of the ANN’s structure is given in Figure 10. Note that the name of variables in this graph represent the normalize variables.

\[ GHI = -0.6345 \times \text{input 1} + 0.48703 \times \text{input 2} + 0.58825 \times \text{input 3} + 0.63905 \times \text{bias} \]  (5)
Then, the value of each input is sums of the weight from the input variables.

The comparation of this model with the testing data is depicted in the Figure 11, which the horizontal axes determined the test data and vertical axes showed the ANN prediction. From this figure could be concluded that most of the point did not close to the linier line which can inform us that the data has still not trained the model well. Moreover, the result was also supported by the time series plot in Figure 12.

![Figure 11. A Comparation between test data and the prediction](image)

![Figure 12. Time Series Plot between test data and the prediction result](image)

To conclude the result of ANN prediction, it can be inferred from Figure 12 that the distance between the test data and the prediction result was clear. It has been supposing that this condition triggered by the high variance condition in the dataset and also the nonlinear relationship among the variables. As it was mentioned in table 2, the MSE and RMSE was calculated at 25682.56 and 160.26, respectively.

4.3. **Multiple Linear Regression**

The purpose of using regression model was to propose new model that could be compared with the accuracy from the ANN model. Regression is popular toward approximation of causality problems. It has been used in widely range of research. In the regression model, the linear assumption is considered to be fulfilled at the first stage. However, from the correlation matrix showed above in Figure 6, it can be concluded that the linear assumption among the variable did not match with the dataset.

In this research, the regression model was applied to the data train which has been scaled using min-max normalization. The dependent variable of GHI has estimated by the five independent variables. The result of multiple linear regression for this structure is given Table 3.
Table 3. Regression Coefficient

|                  | Estimate | P-Value   |
|------------------|----------|-----------|
| Intercept        | 0.41867  | <2e-16    |
| Temperature      | 0.50426  | <2e-16    |
| Precipitation Total | -0.05278 | 0.2580    |
| Surface Pressure | 0.07149  | 0.0564    |
| Wind Speed       | -0.03006 | 0.3787    |
| Specific Humidity| -0.07704 | 0.0836    |

It was clear that only temperature that has significant impact to the dependent variable, while the others was not. This result has supported by the information of determinant coefficient and the MSE or RMSE of the model. It has $R^2 = 0.2207$, $MSE = 25747.24$ and $RMSE = 160.46$. Meanwhile the prediction of regression model is depicted in this Figure 13.

![Prediction Vs Test Data for Regression](image)

Figure 13. Prediction Vs Test Data for Regression

To sum up this approximation, the independent variables could not fit to describe the independent variable since the relation among these two variables were nonlinear. Therefore, the regression result was not suitable to this case. In addition, since the RMSE of the regression was slightly higher than the ANN, it was concluded that the ANN still can perform the best result.

5. Conclusion

In this research, both ANN and regression represent a quite similar accuracy with the RMSE of 160.26 and 160.46, respectively. This result could be happened because the data of GHI which represents the solar power had a high range and also weak linear correlation with the independent variables. However, in case of climate or weather conditions, the high variance of the data could not be neglected. Especially in the province of East Kalimantan which is located in the Equator and the climate condition relatively unpredictable each day. Therefore, to model this data, the outlier value cannot be eliminated since it also has an important meaning. From these results, the ANN model is preferable to be taken as the prediction model in the future because it supposed to give best approximation in case of nonlinear forecasting. In the next research, the number of data can be added to train the model as well as taking more input variables. Furthermore, the other method such as univariate forecasting can be tried also to model the solar energy. In case of solar energy potential in the future, the temperature variable could be the best indicator to measure the strength of the sun power to generate the electricity.
Acknowledgments
Authors wishing to acknowledge assistance or encouragement from colleagues and also the LPPM of Kalimantan Institute of Technology who have provided fund for this research by a grant in contract number 3571/IT10.II/PPM.04/2021. Hopefully this research could contribute to the solar energy problems especially in East Kalimantan.

References
[1] Maulidia M, Dargusch P, Ashworth P and Ardiansyah F 2019 Renew. Sustain. Energy Rev. 101 231–47
[2] Pemerintah Provinsi and Kalimantan Timur 2019 Pemprov Kaltim
[3] Udin U 2020 Int. J. Energy Econ. Policy 10 233–7
[4] Karjoko L, Winarno D W, Rosidah Z N and Handayani I G A K R 2020 Int. J. Innov. Creat. Chang. 11 259–69
[5] Yudha S W and Tjahjono B 2019 Energies 12 1–19
[6] Gherboudj I and Ghedira H 2016 Renew. Sustain. Energy Rev. 55 1210–24
[7] Vaderobli A, Parikh D and Diwekar U 2020 Energies 13
[8] Meybodi M A and Beath A C 2016 Renew. Energy 93 510–24
[9] Abuella M and Chowdhury B 2015 North Am. Power Symp. NAPS 2015
[10] Abuella M and Chowdhury B 2017 Solar power forecasting using support vector regression arXiv
[11] Prastawa A and Dalimi R 2020 J. Energi dan Lingkung. 10
[12] Yolcu U, Egrioglu E and Aladag C H 2013 Decis. Support Syst. 54 1340–7
[13] Solcast 2019 Global Solar Irradiance Data and PV System Power Output Data
[14] Sengupta M, Habte A, Stoffel T, Perez R, Myers D, Gueymard C and Philippe Blanc 6 and Stefan Wilbert 2017 Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Second Edition
[15] Moreira J M, de Carvalho A C P L F and Horváth T 2018 A General Introduction to Data Analytics
[16] Karsoliya S 2012 Int. J. Eng. Trends Technol. 3 714–7