Forecasting Specific Yield of Thin Film Solar Panel in Malacca Via Regression Analysis

Arfah Ahmad[1*], M. 'A. Azmi[1], A. J. Haja Mohiddeen[2], K. A. Baharin[1], V. Sreeram[3]

[1]Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

[2]Kuliyyah of Engineering, International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur

[3] School of Electrical, Electronic and Computer Engineering, The University of Western Australia, 35 Stirling Highway, Perth WA 6009, Australia

[*]Corresponding author's email: arfah@utem.edu.my

Abstract – *This study employs multiple regression analysis to investigate the relationship between environmental variables and the specific yield of solar photovoltaic (PV) panels installed in Malacca, Malaysia. Several predictive statistical models for specific yield were developed to quantify the influence of environmental parameters. This study using data collected from PV solar panels with environmental parameters that are; relative humidity, tilt irradiance, global irradiance, temperature and wind speed. 365 days dataset of specific yield and environmental parameters were pre-processed and analysed. Single variable and multiple variables regression models were developed to identify the best-fit model for forecasting specific yield of thin film solar panel. The accuracy of the developed models was evaluated using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results from error calculation and graphical analysis reveal that the multiple regression model demonstrating a high predictive accuracy as compared to single variable regression model. The analysis quantifies the relative impact of each variable, providing valuable insights into which factors are most critical for optimizing PV performance in Malacca, Malaysia.*

Keywords: *forecasting, regression analysis, solar panel, specific yield.*

Article History

Received 1 September 2025

Received in revised form 1 October 2025

Accepted 31 October 2025

I. Introduction

Energy is the primary source of economic development and urbanization. It is one of the vital elements in the operation of any modern industrialization to drive economic productivity. Numerous activities and equipment used in daily life require energy as the basic input. Fossil fuels are the main non-renewable energy resources. Fossil energy sources such as oil, coal, and natural gas are depleting over time and are unable to fulfill the increasing energy demands. In addition, the usage of fossil fuels as a source of energy negatively impacts the environment. When fossil fuels are burned, a large amount of carbon dioxide and greenhouse gases are emitted into the air. Therefore, exploring alternative energy sources that are clean and renewable is vital [1],[2].

Renewable energy is produced from renewable sources such as solar energy, wind energy, hydroelectric energy, biomass, and geothermal energy. Solar energy has been recognized as a promising clean and green energy. Solar energy is converted to electrical energy via a photovoltaic (PV) material or thermal process. The PV material will generate an electrical potential when it is exposed to the light. Solar panels harness the heat from the sun and convert it to electrical energy. The solar panel is made of conducting materials such as silicon. Solar panels are usually placed on the top of a building for maximum absorption of solar energy. Solar panels harness the energy from sunlight through the photovoltaic effect, converting it into electricity by facilitating the movement of electrons within silicon-based solar cells. This clean energy source is a crucial component in the transition towards sustainable energy solutions, helping to reduce reliance on fossil fuels. By forecasting the energy produced by a PV system, Return on Investment (ROI) and long-term profit can be estimated. However, forecasting the energy produced by a PV system can be challenging due to non-ideal environmental factors that may affect the performance of a PV system. Improved forecasting methods can be developed by investigating the root cause of the differences between the forecast and actual energy output of a solar PV plant. A variety of methods for forecasting solar exist such as statistical-based method, artificial neural network, machine learning, and hybrid models. Linear regression is the most basic machine learning method that can be used to forecast solar irradiation and subsequently energy produced by the PV system [2].

Several works on solar PV systems forecasting based on regression analysis as presented in [1]-[7]. Studies in [1] and [2] developed a single parameter linear regression model of solar radiation data collected in Northern Malaysia. The linear relationship between average air temperature and solar radiation proves that the linear regression model developed is adequate in forecasting future solar radiation [1],[2]. In another study in [3], the relationship between global irradiance and meteorological parameters were investigated via a single parameter model and multiple parameters regression model. Findings from this study show that relative humidity is identified as the most significant parameter for forecasting global solar irradiance, followed by cloud cover, air pressure, gusts, rain, temperature, and wind speed. The multiple parameters regression model with a combination of temperature, rain, humidity, pressure, and wind speed serves as the best model with the highest accuracy in the prediction of solar irradiance. This study also shows that the regression models developed can estimate monthly global solar irradiance in Malaysia [3]. Additionally, the study in [4] explores the application of the Angstrom linear regression model for monthly average daily global radiation on a horizontal surface for six major climates in Jordan. Interestingly, the coefficients of the Angstrom model were estimated based on sunshine duration and the values of ground albedo in the 6 locations tested in the study [4]. The application of multiple regression models with geodata and weather variables is explained in [5-7]. The geodata such as latitude, longitude, altitude above sea level and average temperature are the parameters tested in 80 locations in Europe and Africa [5]. On the other hand, the study in [6] compared the performance of 7 empirical models with newly proposed multiple regression model that is based on temperature, precipitation, and relative humidity to estimate daily solar radiation in Peru. Meanwhile, a highly formulation model that is based on regression analysis for 12 cities around the world is presented in [7]. It can be concluded that results and findings from studies in [5]-[7] show that the multiple regression models developed are efficient, reliable, and able to improve forecast accuracy on solar radiation.

In like manner, modeling and forecasting PV systems via statistical probability distribution were explored in [8]-[10]. The exponential, Weibull, normal, Rayleigh, log-normal, and Gamma distribution are examples of probability distributions used in the literature to analyze and forecast PV systems data. Findings from previous research proved that, in some cases, some of the chosen distributions are well-fitted to the data. Therefore, statistical probability distributions manage to improve forecasting accuracy without any assumptions and considerations on meteorological parameters [8-10]. Solar radiation on PV systems produces a sequence of data points recorded over consistent intervals of time. Auto Regressive Moving Average (ARMA) and Generalized Auto Regressive Conditional Heteroscedasticity (GARCH) are time series models tested on solar irradiance data [11]. The results in [11] show that the recursive ARMA-GARCH model can perform point forecasts as accurately as machine learning-based techniques. In another study by Alsharif et.al. [12], the Auto Regressive Integrated Moving Average (ARIMA) model is used to predict the daily and monthly solar radiation in Seoul, South Korea. According to the findings of this study, ARIMA (1,1,2) and ARIMA (4,1,1) are the best-fitted time series models for daily and monthly solar radiation, respectively [12].

Although existing statistical based methods have been proven to model and forecast solar data with high accuracy, the search for more accurate model that is based on artificial intelligence (AI) and machine learning approach has gain attention recently. Long-short term memory (LSTM), artificial neural network (ANN), convolution neural network (CNN), random forest, gated recurrent unit (GRU) are examples of machine learning models used in the analysis and prediction of solar PV systems. A study in [13] investigate the application of random forest algorithm for effective day-ahead forecasting using metering data and open-source weather data. Results from this study shows that the random forest algorithm is sufficiently reliable for forecasting solar power plant output with high prediction accuracy [13]. The application of LSTM, ANN and GRU in forecasting one-day-ahead PV power with three groups of weather data were explored in [14]. The three models were applied with results show that LSTM model combined with hybrid weather data is the most reliable model for one- day-ahead PV power forecasting [14]. Another study in [15] discovered that the LSTM can serves as an accurate predictive model to forecast solar energy as compared to other conventional machine learning methods. However, the LSTM-based model required longer data training time. Following this, a study in [16] proposed hybrid model consists of CNN and LSTM in solar power forecasting. Experimental results shows that the hybrid CNN-LSTM is able to forecast power generation and can help to optimize PV plant operation [16]. Recent study by Gokhan Sahin et.al.[17] investigated the surface parameters and environmental factors that affect the energy production on PV solar power plant in Igdir province. This study use feed forward neural network (FFNN) and multiple regression analysis to model the solar power plant efficiency. Findings from this study prove that the FFNN can be used to estimate the efficiency of the solar power plant together with multiple regression for a successful performance of the network [17].

Solar energy is considered as one of the main renewable sources in Malaysia. This is due to the favorable climate because of Malaysia being strategically located close to the equator resulting in high solar irradiation available throughout the year. A series of government initiatives have been introduced throughout the years to boost the application of renewable energy to achieve the 70% renewable energy target by 2050. Net Energy Metering (NEM), Feed-in Tariff (FiT), Large Scale Solar (LSS), and Green Electricity Tariff (GET) are the initiatives by the government to increase the usage of solar panels in Malaysia [18]. However, the amount of energy produced by solar panels can be inconsistent due to weather conditions like cloudy or rainy days. The weather inconsistency makes grid integration a challenging task. Thus, solar forecasting becomes vital to ensure grid stability, reliability, and efficient operation of the power system.

The focus of this study will be on model development based on regression analysis for forecasting the energy output of PV systems installed in Malacca, Malaysia. Linear regression is the most basic and widely used technique for regression as it models the relationship between the input and output variables while using linear predictor functions. The regression analysis is chosen over more advanced models due to its simplicity, interpretability, computational efficiency, and suitability with the collected data used in this study. Regression analysis are straightforward and easier to implement, train and maintain as compared to other advanced models. Advanced models such as ANN, CNN, LSTM, and random forest model required complex layers in model development, making it difficult to elaborate in term of prediction. In addition, advanced models generally required a large amount of data to perform well and avoid overfitting. Whereas regression analysis often yields reliable and robust results with smaller dataset, provided all the regression assumption are fulfilled. As this study used several environmental variables with one year data of energy yield, the regression analysis will directly explain and interpret the relationship between each input variable and output variable based on the coefficient's values in the model. Therefore, this study will focus on the prediction of energy yield from solar PV systems installed in Melacca via regression analysis with environmental variables as the input variables. Although several studies have been proven to forecast solar energy data with high accuracy, however, there is a concern regarding the usability of those models in prediction at different locations around the world. In addition, different parts of the world experience different climate and environmental effects, which could further influence the solar energy received.

The contributions of this study can be summarized as; (i) Linear regression-based models are implemented to forecast the energy yield from thin film solar PV panels installed in Malacca, Malaysia; (ii) the reliability, accuracy, suitability, and performance of the models are investigated based on the error measurement analysis; (iii) the forecasting energy yield from the developed model for the coming year is analyzed and compared to the observed data collected from solar PV systems in Malacca, Malaysia.

II. Methodology

This section of the paper explained in detail on the solar data collected used in the study and the development of regression model for single variable and multiple variables.

A. *Solar Data*

The Energy and Power System (EPS) Research Group based in the Universiti Teknikal Malaysia, Malacca (UTeM) has installed a grid – connected PV system at the Faculty of Electrical Technology and Engineering, (FTKE) in UTeM along the coordinates 2.3140° N, 102.3200° E. The PV system contains polycrystalline, monocrystalline, thin film, and heterojunction with intrinsic (HIT) PV panels that are connected to the utiliy grid through inverters. The PV systems have been installed around the FTKE area in Malacca, Malaysia. The monocrystalline panels, thin film panels and HIT panels were installed at the rooftop of the administration building, laboratory building and the lecture hall building respectively. Meanwhile, the polycrystalline panels are located at the entrance of FTKE building.

The data collected from the PV solar panels consists of readings of energy generated by the PV system in kWh. The readings are collected at a 5-minutes time interval for a total of 365 days. The environmental data used in this study is collected from pyranometers and sensors installed at the PV system site. The environmental data collected are global irradiance, tilt irradiance, average temperature, average relative humidity, and average wind speed. Each of these readings were recorded at a 1-minute interval throughout a one year duration. Data cleaning procedures such as estimation of missing values, data merging, and filtering were applied to the raw data to reduce error and increase the accuracy of model development. The specific yield for each type of solar panel is given as in (1).

$$Specific\ Yield = \frac{system's\ annual\ average\ energy\ yield\ (kWh)}{system's\ installed\ capacity\ (kWp)} \quad (1)$$

B. *Model Development*

The proposed forecasting models for this study are generated via regression analysis. Regression analysis is a statistical technique for investigating and modeling the relationship between a variable of interest (the response) and a set of related predictor variables. Initially, the

Pearson correlation coefficient, r is calculated to measure the strength of relationship between the two variables. Equation (2) gives the formula to calculate r.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{2}$$

where SS_{xy}, SS_{xx} and SS_{yy} are sum of squares for x, the predictor variable and y, the response variable [19]. The r value ranges from -1 to $+1$. A value of r close to $+1$ signifies a strong positive correlation between the two variables. Meanwhile, when the value of r is close to -1, it indicates a strong negative correlation between predictor variables and response variables. Table I shows the conventional approach to interpret the correlation coefficient scale [20]. If the value of r is equal to 0, the relationship between variables under study can be neglected.

TABLE I
INTERPRETATION OF CORRELATION COEFFICIENT [20]

Absolute Magnitude of the Correlation Coefficient, r	Interpretation
0.00 – 0.09	Negligible correlation
0.10 – 0.39	Weak correlation
0.40 – 0.69	Moderate correlation
0.70 – 0.89	Strong correlation
0.90 – 1.00	Very strong correlation

A linear regression model with one predictor variable is given in (3);

$$y = \beta_0 + \beta_1 x_1 \tag{3}$$

where y is the response variable and x_1 is the predictor variable. In general, y might be related to several predictor variables given as $x_1, x_2, \ldots, x_k$. Therefore, the multiple linear regression model is given as in (4).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \tag{4}$$

To estimate the unknown parameters of $\beta_0, \beta_1, \beta_2, \ldots \beta_k$ in the regression model, the least square estimation procedure was applied to the data. The least-square procedure is the process of finding the best-fitting curve or line for a set of data points by reducing the sum of squares of the residual from the curve. Therefore, β_0 and β_1 are the least squares estimators of the intercept and slope of the curve or line. For a linear regression model in (3), the estimation of β_0 and β_1 via least square estimation are given as in (5) and (6):

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} \tag{5}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{6}$$

SS_{xy} and SS_{xx} in equation (5) are sum of squares for x and y, while $\bar{x}$ and $\bar{y}$ in (6) are the mean for x and y respectively [19]. When the regression model is fitted to the data, it is important to determine the goodness of fit of the model. For this purpose, the coefficient of determination, R^2 is calculated for the single linear regression model [19]. R^2 is calculated by using the equation in (7):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

where $\hat{y}$ is the predicted response variable from the regression model, y is the observed specific yield from the PV solar system in Malacca and $\bar{y}$ is the average of specific yield from the data. In addition, the adjusted coefficient of determination, $\bar{R}^2$ is calculated for the multiple linear regression model. The $\bar{R}^2$ is used for multiple variables model to provide a more reliable measure of the model's fitness by penalizing the inclusion of unnecessary predictor variables. The value of R^2 is always increase or remain the same when a new predictor variable is added to the regression model, even if the predictor variable is not correlated to the respond variable. However, the $\bar{R}^2$ value will increase only if the new predictor variable is significantly contributed to the regression model [19]. Equation (8) is the formula used to calculated $\bar{R}^2$. The lowest possible value for both R^2 and $\bar{R}^2$ are 0 and the highest possible value is 1. A value that is closer to 1 implies that the regression model can predict the response variable with a high accuracy.

$$\bar{R}^2 = 1 - \left[(1 - R^2) \times \frac{n-1}{n-k-1}\right] \tag{8}$$

where n is the number of observation (sample size) and k is the number of predictor variables in the model.

In this study, the developed regression model will be used to forecast the monthly specific yield of a PV solar panel. The performance of the model gain will be assessed via error measurement analysis. Mean absolute error (MAE) and root mean square error (RMSE) were calculated to determine the deviations between observed values and estimated values from the model. Low values on MAE and RMSE indicate that the developed model accurately forecast the specific yield. Equations (9) and (10) are mathematical formula of MAE and RMSE respectively with n refer to the number of days for the estimated specific yield.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

III. Results and Discussion

Input data for this study is collected from four types of

PV solar panels that are polycrystalline, monocrystalline, thin film, and HIT panels. Table II display the descriptive statistics of each PV panels for a one-year duration. Descriptive statistics analysis provides insights into features and patterns of the data. With reference to Table 1I, the highest maximum value of specific yield is by thin film solar panel, that is 7.64 kWh/kWp, followed by HIT, monocrystalline and polycrystalline. Hence, thin film solar panels produced the highest energy yield for every peak power of the PV system. Meanwhile, polycrystalline had the lowest minimum value of specific yield (0.32 kWh/kWp), followed by HIT, monocrystalline and thin film solar panels. Therefore, polycrystalline PV panels produced the lowest energy yield for every peak power of the PV system.

TABLE II
DESCRIPTIVE STATISTICS OF THE DATA

Measurement (kWh/kWp)	Poly	Mono	HIT	Thin Film
Maximum	6.66	7.02	7.13	**7.64**
Minimum	**0.32**	0.34	0.33	0.34
Mean	3.78	3.66	3.98	**4.24**
Std. Dev.	**0.83**	1.16	0.93	1.07

The results in Table II shows that thin film solar is the most efficient PV panel in hot and humid climates of Malacca as compared to other type of solar panels. Additionally, thin film solar panels have the highest mean values of specific yield followed by HIT, polycrystalline and monocrystalline. It is evident that the thin film PV panels are the most efficient for generating power in the installed area. Polycrystalline has the lowest standard deviation, followed by HIT, monocrystalline and thin film solar panels. The low standard deviation value shows that the specific yield data is clustered around the mean value. Results in Table 1I proof that thin film solar panels produce the highest specific yield as compared to other types of solar panels used in the study. Henceforth, for the purpose of developing a forecasting model, the rest of this paper will focus on model development.

A. Single Parameter Model

In this section, the effect of each weather parameter on the specific yield of thin film solar panels are discussed based on the value of r, R^2 and error analysis. There are five regression models developed for the different single parameter of weather variables. The weather variables are global irradiance, tilt irradiance, average temperature, average relative humidity, and average windspeed. Single parameter models mean each model uses only one predictor variable as the parameter. The value of correlation coefficient of each weather variable is as shown in Table III. The highest value of r is on relative humidity (0.5848), follow by tilt irradiance, global irradiance, and temperature average. The lowest of r is on windspeed average, which is 0.2835 (weak correlation). A high value of r for relative humidity means that there is a moderate positive correlation between the variable and the specific yield. In other words, as relative humidity average increases, specific yield of thin film solar panel also tends to increase. Based on the value of r in Table III, it is evidence that each weather variable is significantly correlate with the specific yield produce by thin film solar panel.

TABLE III
CORRELATION COEFFICIENT FOR SINGLE VARIABLE

Single Variable	Correlation Coefficient, r
Relative Humidity (RH)	**0.585**
Tilt Irradiance (TI)	0.481
Global Irradiance (GI)	0.472
Temperature Average (TA)	0.332
Wind Speed Average (WSA)	0.285

Following this, the regression models and the associated R^2 value, MAE and RMSE for each model is shown in Table IV. For M1, the model uses relative humidity, M2 is on tilt irradiance, M3 for global irradiance, M4 uses temperature average as the predictor variable and M5 is based on windspeed average.

TABLE IV
SINGLE PARAMETER MODEL

Single Parameter Model	R^2	MAE	RMSE
M1: $SY = 1.437 + 0.0704\ RH$	**0.342**	**0.64**	**0.88**
M2: $SY = 1.335 + 0.0028\ TI$	0.232	0.86	1.18
M3: $SY = 1.077 + 0.0030\ GI$	0.222	0.80	1.12
M4: $SY = 2.885 + 0.1672\ TA$	0.110	0.86	1.16
M5: $SY = 3.402 + 0.1917\ WSA$	0.081	1.01	1.46

SY is specific yield.

The least-square estimation gives the values of β_0 and β_1 for each model as presented in Table IV. The value of R^2, MAE and RMSE illustrates the suitability and performance of each single parameter model in modelling specific yield of thin film solar panel. The results shows that relative humidity (M1) exhibits the best R^2 and the least MAE and RMSE as compared to other models. Therefore, it means that relative humidity is strongly applicable for modelling specific yield of thin film solar panel. This result is supported with the common modelling method, wherein humidity is one of the most conventional and practical parameters used in modelling global solar irradiance [2]. Results in Table IV shows that the single parameter model with wind speed average gives the lowest R^2 (0.081) and the highest error, which are 1.01 (MAE) and 2.13 (RMSE) as compared to other models. This finding is comprehensible with the lowest r value (0.2852), which means that there is a very low positive correlation exists between wind speed average and specific yield.

Although the model formulation with one independent variable appears to be a straightforward, the addition of many independent variables introduces a new idea in the interpretation of the regression coefficient. Multiple regression produces results when each variable is changed one at a time while the values of the others variable remain constant. This is contrasts with conducting numerous simple linear regressions for each of these variables, where each regression models ignores what may be happening with the other variables. The coefficient associated with each independent variable in multiple regression should reflect the average change in the response variable associated with changes in the independent variable, while other variables stay constant. The results in Table IV shows that relative humidity, tilt irradiance, global irradiance, and temperature average give values of R^2, MAE and RMSE within the average values, which implies that combining these variables may add to the accuracy of the specific yield prediction model. Therefore, multiple parameters model is considered in the development of specific yield forecasting model.

B. Multiple Parameter Model

Multiple parameter models are those that use a combination of multiple weather variables. In this study, four multiple parameter models were proposed as shown in Table V. For instance, multiple parameter model of M6 uses the relative humidity and tilt irradiance as the predictor variables. The M7 model comprises relative humidity, tilt irradiance, and global irradiance. The multiple parameter model on M8 were developed by the relative humidity, tilt irradiance, global irradiance and temperature average. The predictor variables for M9 model are based on combination of relative humidity, tilt irradiance, global irradiance, temperature average and windspeed average. For each model, the coefficient of $\beta_0, \beta_1, \beta_2, \beta_{3,} \beta_4$ and β_5 were determine by using least-square estimation. Table V presents the value of each coefficient of $\beta_0, \beta_1, \beta_2, \beta_{3,} \beta_4$ and β_5 for the multiple parameter model. For each of this model, the associate R^2, $\bar{R}^2$, MAE and RMSE is shown in Table VI.

TABLE V
MULTIPLE PARAMETER MODEL

Regression Equation for Multiple Parameter Model
M6: $SY = 0.290 + 0.0559RH + 0.0017TI$
M7: $SY = 0.518 + 0.0613RH + 0.0032TI - 0.0020GI$
M8: $SY = 0.557 + 0.0643RH + 0.0032TI - 0.0019GI - 0.0206TA$
M9: $SY = 0.390 + 0.06234RH + 0.0025TI - 0.0013GI - 0.0147TA + 0.0782WSA$

SY is specific yield.

Based on the results in Table VI, the M6 model gives the best fit, with the highest R^2 (0.491), and the lowest error values with MAE of 0.50 and RMSE is 0.71. The value of R^2 for M7 model is the lowest (0.420) follow by M8 model (0.421) and M9 model (0.431). It is evident that M7 model gives the lowest fit as compared to other models. The inclusion of tilt irradiance and global irradiance are not the best combination in predicting specific yield of thin film solar panel in Malacca, Malaysia.

TABLE VI
MULTIPLE PARAMETER MODEL

Multiple Parameter Model	R^2	$\bar{R}^2$	MAE	RMSE
M6	**0.491**	**0.482**	**0.50**	**0.71**
M7	0.420	0.415	0.58	0.84
M8	0.421	0.414	0.59	0.88
M9	0.431	0.423	0.57	0.86

Observing the values of R^2, $\bar{R}^2$ and error values in Table VI, it can be concluded that adding extra variables such as global irradiance, temperature average, and wind speed average to the model, the errors tend to increase, and the model fitness drop slowly. However, combining all these variables for multiple parameter model is more significant as compared to the single parameter model. Comparing Table IV and Table VI, it clearly shows that the values of R^2 and $\bar{R}^2$ for multiple parameter model are higher than the single parameter model. Together with this, the MAE and RMSE for multiple parameter model are also lower than the single parameter model.

Based on the results gain from Table IV and Table VI, the best fit for single parameter model is M1, that is consist of relative humidity as the predictor variable. Whereas combination of humidity and tilt irradiance as the predictor variables (M6) gives a better fit than M1 model with minimal forecasting error. M6 model gives the best fit to the data by 49.10%, whereas M1 model fit the data by 34.20%. Therefore, the humidity and tilt irradiance are the most significant predictor variables for forecasting specific yield of thin film solar panel in Melacca, Malaysia. Results of this study demonstrate that the linear regression analysis with single parameter and multiple parameters exhibit a reliable predictive capability in estimation of specific yield for thin film solar panel.

C. Forecasting of Specific Yield

Additional data on specific yield and weather variables were collected from January 1st until January 31st. These data were used to validate both M1 and M6 models in forecasting future specific yield of thin film solar panel. Fig. 1 shows the prediction of specific yield based on M1 model over 31 days period. Based on Fig. 1, it is clearly shown that the forecast values of specific yield exhibit a pattern that is fairly close to the actual specific yield of thin film solar panel.

For most of the month, the actual specific yield closely follows the forecasted values, with some days performing slightly better and other days slightly worse. Additionally, some of the prediction data points are lower than the actual specific yiled, for instance on January 13th until January 16th. On Jan 26th until January 31st, the prediction points are diverted from the observed specific yield. There are significant drops in the actual specific yield on days 23, 27, and 30, where the performance falls far below the forecast. Such drops could be due to factors like adverse weather conditions such as, heavy clouds or rain, or shading. In particular, this deviation is supported by the error values on M1 model as presented in Table IV. In general, it is concluded that M1 model can predicts specific yield with reasonably well based on relative humidity as the predictor variable.

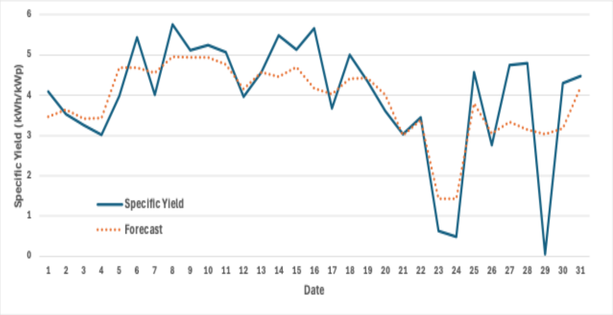

Fig. 1. Prediction of specific yield by using M1 model.

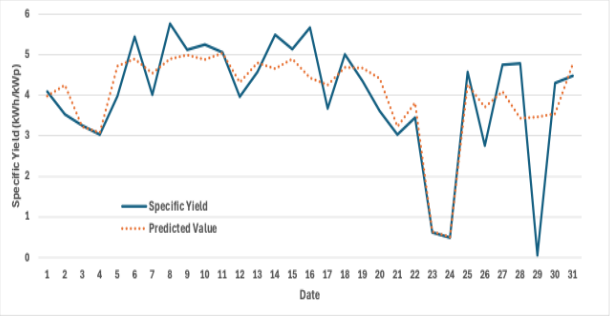

Fig. 2. Prediction of specific yield by using M6 model.

Fig. 2 shows the prediction of specific yield based on M6 model, which consists of combination on relative humidity and tilt irradiance. The line plot of forecast specific yield is nearly close to the actual specific yield, indicating that the predictive model is reasonably effective. Both the actual and predicted values show significant daily fluctuations. For many days, such as days 7-9 and 12-14, the actual specific yield closely matches the predicted value. In addition, the line plot in Fig. 2 shows that the prediction on January 22nd until January 25th is the same as the observed data. This suggests that the M6 model is successfully forecasted the specific yield almost accurate as the recorded data.

IV. Conclusion

Findings from this study show that the linear regression analysis successfully captured a strong relationship between the predictor variables and the specific yield of solar panels in Malacca, Malaysia. For single variable model, the relative humidity serves as the most significant factor in forecasting specific yield of a thin film panel. The combination of relative humidity and tilt irradiance produce a linear regression model with better predicted values than the single variable model. Results of this study shows that the multiple variable models give better accuracy in forecasting future values as compared to the single variable model. Specific yield is a performance metric used in the solar energy industry to measure the energy generated by a solar power plant over a period of time, normalized by its installed capacity. A higher specific yield value generally indicates better performance, as it means the system is producing more energy relative to its size.

Acknowledgements

The authors wish to acknowledge the Ministry of Education, Malaysia and Universiti Teknikal Malaysia Melaka, for supporting this research financially. The authors are grateful to the Energy and Power Systems (EPS) research group for making their solar panels data available, which was essential for this research.

Author Contributions

The authors contribution on this paper are as follows; Author 1: Supervision and writing full paper; Author 2: Data analysis and draft preparation. Author 3: Conceptualization, editing and review. Author 4: Data collection. Author 5: final review on the paper, funding acquisition, project administration.

Conflict of Interest

The authors declare no conflict of interest in the publication process of the research article.

References

[1] S. Ibrahim, I. Daud, Y.M. Irwan, M. Irwanto, N. Gomesh, and Z. Farhana, “Linear Regression Model in Estimating Solar Radiation in Perlis”, *Energy Procedia*, Vol. 18, pp. 1402-1412, 2012.

[2] S. Ibrahim, I. Daud, Y.M. Irwan, M. Irwanto, N. Gomesh, and A.R.N. Razliana, “An Estimation of Solar Radiation using Robust Linear Regression Method”, *Energy Procedia*, Vol. 18, pp. 1413-1420, 2012.

[3] H.A. Kutty, M.H. Masral and P. Rajendran, “Regression Model to Predict Global Solar Irradiance in Malaysia”, *International Journal of Photoenergy*, Vol. 15, pp. 1-7, 2015.

[4] A.A. Badran, and B.F. Dwaykat, “Prediction of Solar Radiation for the Major Climates of Jordan: A Regression Model”, *Journal of Ecological Engineering*, Vol. 19, No. 2, pp. 24-38, 2018.

[5] A. Bocca, L. Bergamasco et. al., “Multiple Regression Method for Fast Estimation of Solar Irradiation and Photovoltaic Energy

Potentials over Europe and Africa," *Energies,* Vol. 11, No. 12, pp. 3477-3494, 2018.

[6] B. Mohammadi and M. Roozbeh, "Performance Analysis of Daily Global Solar Radiation Models in Peru by Regression Analysis", *Atmosphere*, Vol. 12, No. 3, pp.389-417, 2021.

[7] D.R. Arumugham and P. Rajendran, "Modelling Global Solar Irradiance for Any Location on Earth Through Regression Analysis using High-Resolution Data", *Renewable Energy*, Vol. 180, pp. 1114-1123, 2021.

[8] Y.S. Arthur, K.B. Gyamfi and S.K. Appiah, "Probability Distribution Analysis of Hourly Solar Irradiation in Kumasi-Ghana", *International Journal of Business and Social Research,* Vol. 3, No. 3, pp.63-75, March 2013.

[9] S. Dukkipati, V. Sankar and S.P. Varma, "Forecasting of Solar Irradiance using Probability Distributions for a PV System: A Case Study", *International Journal of Renewable Research*, Vol. 9, No. 2, pp. 741-748, June 2019.

[10]S.M. Ibrahim et. al., "Statistical Analysis of Solar Energy Potential and Energy Yield in Kano, Nigeria", *International Journal of Information Processing and Communication*, Vol. 9, No. 1&2, pp. 184-198, May 2020.

[11]M. David, F. Ramahatana, P.J. Trombe, and P. Lauret, "Probabilistic Forecasting of the Solar Irradiance with Recursive ARIMA and GARCH Models", *Solar Energy,* No. 133, pp. 55-72, 2016.

[12]M.H. Alsharif, M.K. Younes and J. Kim, "Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea", *Symmetry,* Vol. 11, pp. 2-17, 2019.

[13]A. Khalyasmaa et.al., "Prediction of Solar Power Generation Based on Random Forest Regressor Model," *2019 International Multi-Conference on Engineering, Computer and Information Sciences* (SIBIRCON), Novosibirsk, Russia, pp. 0780-0785, 2019.

[14]W.C. Kuo, C.H. Chen, S.H. Hua, and C.C. Wang, "Assessment of Different Deep Learning Methods of Power Generation Forecasting for Solar PV System", *Applied Sciences*, Vol. 12, pp. 7529-7545, 2022.

[15]M.J. Nur Liyana e.al., "Investigating the Power of LSTM-Based Model in Solar Energy Forecasting", *Processes*, Vol. 11, pp. 1382-1403, 2023.

[16]S.C. Lim, J.H. Huh, S.H. Hong, C.Y. Park and J.C. Kim, "Solar Power Forecasting using CNN-LSTM Hybrid Model," Energies, Vol. 15, pp. 8233-8250, 2022.

[17]G. Sahin, G. Isik and G.J.H.M. van Sark, "Predictive Modelling of PV Solar Power Plant Efficiency Considering Weather Conditions: A Comparative Analysis of Artificial Neural Network and Multiple Linear Regression, *Energy Reports,* Vol. 10, pp. 2837-2849, 2023.

[18]Renewable Energy Malaysia – Renewable Energy Malaysia," SEDA Malaysia, 2023. https://www.seda.gov.my/reportal/

[19]D.C.Montgomery, E.A.Peck, G.G.Vining, Introduction to Linear Regression Analysis (Fifth Edition), Wiley Series, 2017.

[20]P.Schober and L.A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation", Anesthesia & Analgesia, Vol. 126, No. 5, pp.1763-1768, 2018.