REGRESSION ANALYSIS ON SPECIFIC YIELD OF THIN FILM SOLAR PANEL

MUHAMMAD ‘ARASH BIN AZMI

BACHELOR OF ELECTRICAL ENGINEERING WITH HONOURS
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2024

REGRESSION ANALYSIS ON SPECIFIC YIELD OF THIN FILM SOLAR PANEL

MUHAMMAD ‘ARASH BIN AZMI

A report submitted
in partial fulfilment of the requirements for the degree of
Bachelor of Electrical Engineering with Honours

Faculty of Electrical Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2024

DECLARATION

I declare that this thesis entitled "REGRESSION ANALYSIS ON SPECIFIC YIELD OF THIN FILM SOLAR PANEL is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature :

Name : MUHAMMAD ‘ARASH BIN AZMI

Date : 21/6/2024

APPROVAL

I hereby declare that I have checked this report entitled "REGRESSION ANALYSIS ON SPECIFIC YIELD OF THIN FILM SOLAR PANEL", and in my opinion, this project fulfils the partial requirement to be awarded the degree of Bachelor of Electrical Engineering with Honours

Signature	:	
Supervisor Name	:	DR. ARFAH BINTI AHMAD
Date	:	21/6/2024

DEDICATIONS

Dedicated to my beloved parents
Azmi Bin Mohamad & Norli Ayu binti Osman

My supervisor,
Dr. Arfah Binti Ahmad

And all of my family and friends

Thank you

ACKNOWLEDGEMENTS

I would like to begin by expressing my sincere appreciation to all those who provided invaluable assistance during the course of this research. I firmly believe that without their valuable advice and contributions, the completion and progress of this study would not have been possible. I am especially grateful to my dedicated supervisor, Dr. Arfah Binti Ahmad, for her unwavering guidance, direction, and constructive feedback. Her constant encouragement and inquisitiveness have been instrumental in shaping this project and making it what it is today

I would like to express my gratitude to my peers and individuals who have provided assistance throughout this journey. Their valuable advice and perspectives have been a source of motivation for me. Furthermore, I am thankful to my family members for their unwavering support and encouraging words.

ABSTRACT

This project involves conducting a comprehensive regression analysis on the specific yield of thin film solar panels. Utilizing a weather data that represents five independent variables, that are; global irradiance, tilt irradiance, average temperature, average relative humidity, and average windspeed. The study aims to develop a predictive model for the specific yield of thin film solar panels. There are five single regression models and four multiple regression models developed in this study for predicting specific yield. The models' accuracy is measured using mean absolute error (MAE), mean squared error (MSE), and mean percentage error (MPE). The regression model suited the data well, with MAE ranging from 0.56 to 1.01, MSE ranging from 0.71 to 2.13, and MPE ranging from 5.85% to 43.44%. Through regression techniques, the relationships between the independent variables and specific yield are explored, providing valuable insights into the performance of thin film solar technology under varying conditions. The regression model with relative humidity and tilt irradiance is the most accurate multiple regression model for prediction of specific yield by this solar panel at the FTKE, UTeM. The findings of this study are to understand factors influencing specific yield production in thin-film solar panels, contributing to the optimization of their efficiency and overall sustainability.

ABSTRAK

Projek ini melibatkan analisis regresi komprehensif ke atas hasil spesifik panel solar filem nipis. Menggunakan data cuaca yang mewakili lima pembolehubah bebas, iaitu; sinaran global, sinaran condong, suhu purata, kelembapan relatif purata, dan kelajuan angin purata. Kajian ini bertujuan untuk membangunkan model ramalan untuk hasil spesifik panel solar filem nipis. Terdapat lima model regresi tunggal dan empat model regresi pelbagai yang dibangunkan dalam kajian ini untuk meramalkan hasil spesifik. Ketepatan model diukur menggunakan ralat mutlak min (MAE), ralat kuasa dua min (MSE), dan ralat peratusan min (MPE). Model regresi sesuai dengan data dengan baik, dengan MAE antara 0.56 hingga 1.01, MSE antara 0.71 hingga 2.13, dan MPE antara 5.85% hingga 43.44%. Melalui teknik regresi, hubungan antara pembolehubah bebas dan hasil spesifik diterokai, memberikan wawasan berharga tentang prestasi teknologi filem nipis di bawah pelbagai keadaan. Model regresi dengan kelembapan relatif dan sinaran condong adalah model regresi pelbagai yang paling tepat untuk ramalan hasil spesifik oleh panel solar ini di FTKE, UTeM. Penemuan kajian ini adalah untuk memahami faktor-faktor yang mempengaruhi pengeluaran hasil spesifik dalam panel solar filem nipis, menyumbang kepada pengoptimuman kecekapan dan kelestarian keseluruhannya..

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

FTKE	-	Faculty of Electrical Technology and Engineering
UTeM	-	University Technical Malaysia Malacca
PV	-	Photovoltaic
HIT	-	Hetero-Junction With Intrinsic Thin Layer
NAR	-	Nonlinear Autoregressive
ANN	-	Artificial Neural Networks
CSV	-	Comma Separated Value
RMSE	-	Root Mean Squared Error
MAE	-	Mean Absolute Error
MAPE	-	Mean Absolute Percentage Error
MBE	-	Mean Bias Error
n	-	Number of data
μ	-	Mean
x_i	-	Independant Variables
y_i	-	Dependant Variables
r	-	Correlation Coefficient

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

Solar power is a renewable energy source that converts sunlight into electricity. Solar power generates clean and renewable electricity [1]. As a result, solar power is a sustainable and environmentally friendly energy source that can help to prevent climate change. Technology on solar energy generation is getting better, the cost of solar panels and associate equipment has become reasonable. Therefore, solar system installation is more affordable for individuals and businesses. Additionally, solar power can help minimize the dependency on fossil fuels by generating electricity from the sun, resulting in increased green energy usage.

Solar radiation is the result of atoms fusing inside the sun. Part of the energy created by the fusion process heats the chromosphere, the sun's outer layer that is much cooler than the centre of the sun, and the radiation from the chromosphere produces the solar radiation incident on Earth [2]. Aerosols occur when solar radiation enters the earth's atmosphere and part of the incident energy is released through scattering or absorption by air molecules, clouds, and particulate matter. Beam radiation occurs when the radiation from the sun is not reflected on the surface of the PV module while the scattered radiation that reaches the ground is called diffuse radiation. Some radiation may reach reception after reflection from the ground and is known as albedo. Global irradiation can reach 1000 W/m^2 if the weather conditions are good and the sun is in a direct line from the PV module [3]. Global irradiation is the amount of solar radiation on the horizontal surface of the PV module. Figure 1.1 shows the solar radiation in the earth’s atmosphere.

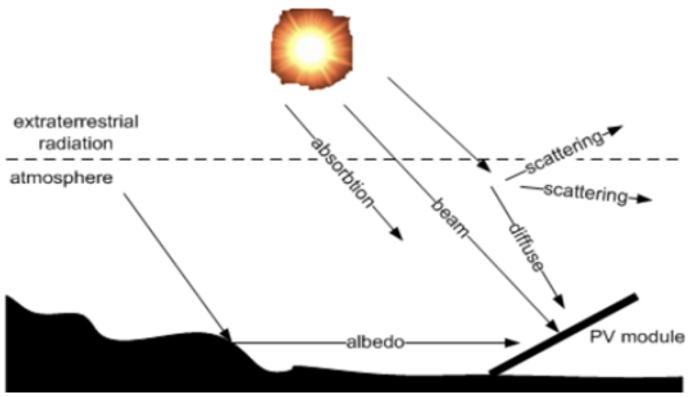


Figure 1.1: Solar radiation on earth's atmosphere [2].

Solar panels use sunlight to produce electricity through photovoltaic cells that convert sunlight into direct current electricity. Thin-film solar panels, among various types of solar panels, have gain attention due to their unique advantages. These panels are distinguished by their thin layers of photovoltaic material deposited onto a substrate, making them flexible, lightweight, and potentially less expensive to manufacture compared to silicon-based panels [4] .

Figure 1.2: Types of solar panel at FTKE, UTeM [5].

There are several types of solar panels at Faculty of Electrical Technology and Engineering (FTKE) that are polycrystalline, monocrystalline, thin film and hetero junction with intrinsic (HIT) solar panels. This project focus on thin film because this thin film solar panel produces the highest electrical energy compared to the polycrystalline, monocrystalline, and HIT. Figure 1.2 show types of solar panel installed at FTKE, UTeM. Weather data for this study is obtained from pyranometer, pyranometer ambient temperature sensor, (Vaisala) Module Backsheet Temperature, Sensor (Campbell Scientific) Ultrasonic Anemometer and rain sensor (EML) installed at FTKE, UTeM [5]. The available on weather data are global irradiance, tilt irradiance, average temperature, relative humidity, panel temperature, and wind speed. Figure 1.3 shows the monitoring system of solar PV Lab at FTKE, UTeM.

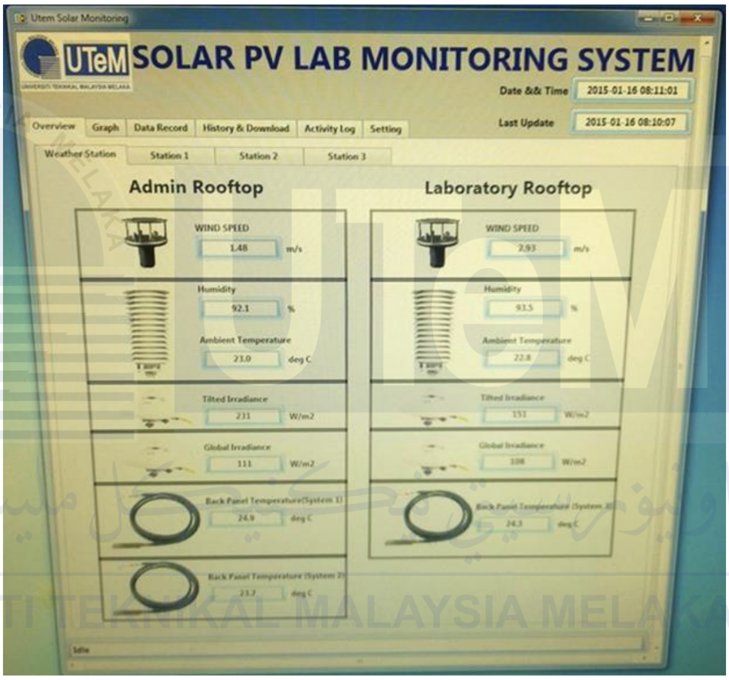


Figure 1.3: Monitoring system at FTKE, UTeM [5].

1.2 Problem Statement

The usage of solar panel technology as a renewable energy production provides a promising revenue for sustainable power generation. However, developing an accurate model for estimating the energy output of solar panel under diverse environmental conditions remain a challenge. Accurate predictions provided by regression models help to reduce costs connected with system design, maintenance, and energy generation. The pyranometer and sensor collects a lot of data. Thus, there is need to determine which variables are significantly affect the electricity generation of the PV module in the FTKE, UTeM. Identifying relevant variables into forecasting models will improves the accuracy of predictions. This study aims to predict the specific yield for the upcoming month using regression analysis techniques. Different types of solar panels will produce various amounts of specific yield. Furthermore, the acquired data contains missing values. Specifically, some data has been missing for some months within the given time frame. To improve the accuracy of the prediction process, it must be done to calculate the missing values for the specific time.

1.3 Objectives

1. To analyse the relationship between specific yield and weather factors through regression analysis techniques.

2. To predict the specific yield by utilizing both linear regression and multiple regression models.

3. To evaluate the precision of the regression models in forecasting specific yield through error measurement analysis.

1.4 Scope of research

This project used combination of weather data and solar data for one year from 1st January 2016 until 31st December 2016. Weather data and solar panel data are collected at different time interval. Weather data is gathered every 1 minute, whereas solar panel data is recorded every 5 minutes. Therefore, the time collected by the weather data needs to be changed into 5-minute intervals to be combined with the data from the solar panel. The original data for weather and solar panel data contains lots missing values and data cleaning is required to solve this problem. Weather data and data from solar panels are obtained from FTKE, UTeM only. There are five weather variables used for this study, which are relative humidity, tilt irradiance, global irradiance, temperature average, and windspeed average. There are four solar panels installed in FTKE, UTeM which are polycrystalline, monocrystalline, thin film and HIT. This research only focuses on thin film solar panels. Furthermore, the model used for forecasting specific yield are linear regression and multiple regression analysis.

The error measurement used to measure the models prediction are mean square error (MSE), mean absolute error (MAE) and mean percentage error (MPE).

1.5 Project Motivation

Solar energy forecasting is important in Malaysia as the country aims to increase renewable energy capacity, especially solar energy, in line with sustainable energy goals [6]. This is because investment in solar energy infrastructure ensures a long-term and sustainable energy supply for future. Besides that, forecasting solar energy assists authorities in making informed decisions about infrastructure development, resource allocation, and future energy projects, optimizing energy planning across the nation. Next, understanding the variables that affect solar panel energy output leads to improved energy planning. Forecasting models derived from regression analysis enable more precise forecasting, assisting in resource allocation and cost-effective energy planning.

CHAPTER 2

LITERATURE REVIEW

This chapter will provide an overview of various forecasting methods used for solar and error measurement analysis that were applied in previous studies.

2.1 Forecasting Method

There are various methods for forecasting solar irradiation. For example, the methods used are probability distribution, artificial neural network, time series analysis, and hybrid model. These methods have been utilized in previous studies and is discussed in following subchapter.

2.1.1 Probability Distribution

Probability distribution is a statistical function that represents all possible and probability values for continuous random variables. In study [7], the authors examine the appropriate probability distribution function to be used in solar radiation forecasting by using Monte Carlo Simulation. The data taken in this research is solar irradiance and three weather conditions, which are summer, rainy, and winter season.

The probability distribution function studied for this research are normal, Rayleigh, Weibull, and log normal distribution. However, not all probability distribution functions are suitable for this research as it will be determined by Root Mean Squared Error (RMSE). RMSE is the standard deviation of the predicted error and is considered to determine the best probability distribution function. Table 2.1 shows the RMSE values for normal, Rayleigh, Weibull, and log distributions against three weather conditions in study [7].

Table 2.1: RMSE values in [7].

Season	Root Mean Square Error			
	Normal	Weibull	Rayleigh	Log Normal
Summer	0.0604	0.0926	0.0625	0.0825
Rainy	0.0591	0.0746	0.0601	0.0646
Winter	0.0771	0.0936	0.073	0.0849

Results from the RMSE values in Table 2.1 shows that the normal distribution is the best distribution for summer and rainy seasons, while the Rayleigh distribution is the best distribution for winter. Based on [7] it is observed that forecasting solar irradiance is very much useful with the prediction of realistic and accurate solar irradiance for every season individually by fitting data into different probability distributions functions.

In study [8], the pyranometer is used to collected data on solar radiation. The data was obtained from the Solar Energy Laboratory of the Mechanical Engineering of Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. The data consists of hourly solar irradiation data from 1995-2008. The author employed descriptive statistical analyses to obtain the desired results. Probability curve fitting was applied to find the best-fitted probability distribution for the various months of the year. The concept of the mean squared error (MSE) was used to estimate the best-fitted probability distribution.

This study also involved the use of standard probability distributions which is the exponential distribution to analyse the data. The exponential distribution for the continuous random variable that represents an interval of time was defined and utilized in the analysis. In conclusion, the study provides a data-driven approach to understand the solar energy potential in the region, and the findings is beneficial for decision-making of solar energy projects in Kumasi, Ghana.

2.1.2 Artificial Neural Networks

Artificial Neural Networks (ANN) is a technique that forms a mathematical model to approximate the relationship between variables and inspired by human brain functions.

Research [9] focuses on forecasting of daily global solar radiation using a Nonlinear Autoregressive (NAR) neural network model. The NAR neural network model is developed and applied to forecast daily global solar radiation, and the results are compared with those obtained from conventional methods such as ARMA modelling.

The results are compared using statistical analysis based on error measurement analysis. From this research, it highlights the implications of renewable energy technologies in desert areas and emphasizes the potential of the NAR model for accurate solar radiation forecasting. It acknowledges the limitations and challenges associated with collecting and forecasting solar radiation data, while underscoring the importance of accurate forecasts for the successful integration of solar power into the existing energy infrastructure.

In study [10], global solar radiation (GSR) using artificial neural networks (ANN) is predicted based on meteorological data. The authors collected data on length of day, air temperature, relative humidity, and sunshine hours from Safiabad station located in Dezful city, Iran. They used this data to train a feedforward ANN using backpropagation algorithm to predict daily GSR. The authors experimented with different input variables and neural network architectures to find the best model for predicting global solar radiation. It can be found that a neural network with three inputs which are length of day, daily mean air temperature and sunshine hours and two hidden layers with logistic sigmoid transfer function and Levenberg-Marquardt training algorithm was the most effective ANN for predicting global solar radiation. Figure 2.1 shows the three-layer feedforward network.

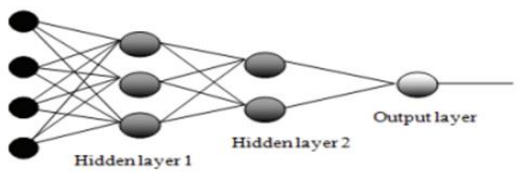


Figure 2.1: Three-layer feedforward network [10].

The results in [10] shows that using the sunshine hours along with daily mean air temperature and relative humidity outperformed the other cases with MPE error of 8.84% and MSE of 0.0041. The MPE and MSE for the case when length of day, daily mean air temperature, and sunshine hours were used as inputs were 9.40% and 0.0044, respectively. For the input length of day, daily mean air temperature, and relative humidity, the MPE and MSE for the case were 14.50% and 0.0089.

2.1.3 Hybrid Models

A hybrid model combines two statistical models to forecast solar irradiation. In study [11], Hargreaves model which estimates daily solar radiation from daily maximum and minimum air temperatures is used in Kelantan. The authors compare the estimated solar radiation values with measured data using statistical indicators such as coefficient of residual mass, root mean squared of error (RMSE), coefficient of determination and percentage error. The results indicate that the Hargreaves model can be used to estimate solar radiation in Kelantan, with a coefficient of residual mass value is 0.09, RMSE is 8.21%, coefficient of determination is 0.8661, and percentage error is 7.98%. These findings suggest that the model estimation is reasonably accurate. The authors conclude that the Hargreaves model can effectively estimate solar radiation in Kelantan, supporting the development of renewable energy strategies in the region. The research findings provide valuable insights into solar characteristics in Eastern Malaysia and offer a practical approach for predicting solar radiation based on readily available meteorological data. This validation of the Hargreaves model in Kelantan adds to the growing body of evidence supporting its utility and accuracy in diverse climatic conditions.

2.1.4 Time Series Analysis

A time series analysis is a mathematical representation used to describe and understand the structure, patterns, and behaviour observed within a time series data. Time series models are designed to capture the dependencies, trends, and potential factors affecting the data over time. The regression model is a type of time series analysis. In study [12], a linear regression model which involves estimating the intercept and slope is used on the sample data. Data analysis involves assessing the relationship between solar radiation and air temperature, with scatter plot showed a positive and roughly linear relationship between the two variables. As a result, the linear regression model is expressed as (2.1),

$$\hat{y} = -81.6415 + 3.6554x \tag{2.1}$$

where $\hat{y}$ is solar radiation, -81.6415 is the intercept, 3.6554 is the slope and x is air temperature. Figure 2.2 shows an example of scatter plot between solar radiation and air temperature in [11].

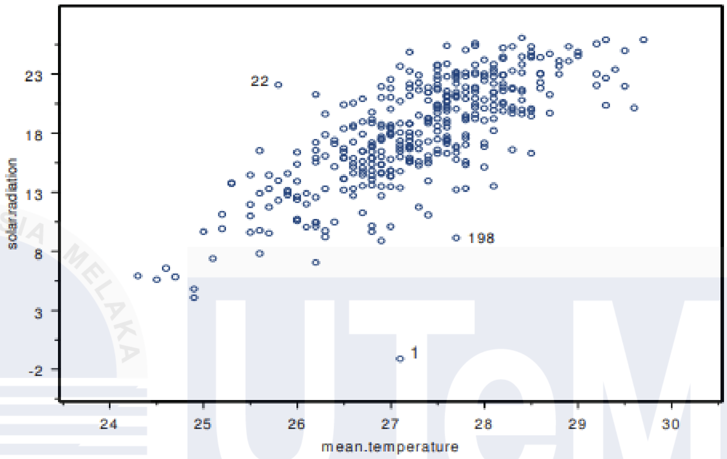


Figure 2.2: The scatter plot of solar radiation versus air temperature [12].

The research also shows the calculation of correlation coefficient and the coefficient of determination, to quantify the relationship between solar radiation and temperature. The authors discuss the model adequacy and diagnostic checks, including residual analysis and normal probability plots, to ensure the reliability of the regression model. The results indicate a strong linear relationship between solar radiation and temperature, with the linear correlation coefficient value ranging from 0.7473 to 0.7780. The coefficient of determination values is about 56% to 61% of the variability in temperature is accounted for by the straight-line fit to solar radiation.

2.1.5 Single and Multiple Parameter modelling

Single-parameter modelling involves analysing the influence of specific meteorological factors, while multiple-parameter modelling integrates many meteorological parameters to improve the accuracy of predictions. In study [13], using Single-Parameter Modelling and Multiple-Parameter Modelling are used to predict global solar irradiance in Malaysia. The study is conducted by analysing various meteorological parameters such as temperature, cloud cover, rain precipitate, relative humidity, wind speed, pressure, and gust speed to develop the regression model.

The authors utilised statistical methods to evaluate the performance of the developed regression models and compare them with existing models. Key performance indicators such RMSE, mean bias error (MBE), and coefficient of determination (R^2) were employed to assess the accuracy and reliability of the regression models. RMSE measured the error between observed and estimated values. RMSE equation is expressed as (2.2),

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \frac{(H_{est} - H_{obs})^2}{n}}{\overline{H_{obs}}}} \times 100\% \tag{2.2}$$

where H_{est} is the estimated global solar irradiance, H_{obs} is the observed global solar irradiance, $\overline{H_{obs}}$ is the averaged observed global solar irradiance, and n is the number of days of estimated global solar irradiance. MBE determined the error of the prediction model. MBE equation is expressed as (2.3),

$$MBE = \frac{\sum_{i=1}^{n} \frac{(H_{est} - H_{obs})^2}{n}}{\overline{H_{obs}}} \times 100\% \tag{2.3}$$

R^2 indicated the model's suitability. R^2 equation is expressed as (2.4),

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(H_{est} - H_{obs})^2}{\sum_{n=1}^{n}(H_{est} - \overline{H_{obs}})^2} \tag{2.4}$$

There have twelves models were evaluated through regression analysis by comparing predicted values with observed data to estimate global solar irradiance based on selected meteorological parameters. As a result of Single-Parameter modelling, humidity had the best RMSE, MBE, and R^2 values, indicate its potential for modelling global sun irradiance. For Multiple-Parameter modelling, Parameter Modelling 12 (PM12) had the lowest RMSE and MBE values. PM12 is combination of five parameters which are temperature, rain, humidity, wind, and pressure. The equation of PM12 is in (2.5),

$$Y = 46327.0656R + 0.0286P + 0.011\,(131.713)^{W} - 12.813T - 3137.320\,(H) - 1831839.527\,(H)^{2} \tag{2.5}$$

Where Y is solar irradiance, R is rain, P is pressure, W is wind, T is temperature and H is humidity. For all the models developed in study [13], Table 2.4 and Table 2.5 show the result of best RMSE, MBE, and R^2.

Table 2.2: RMSE, MBE, and $\boldsymbol{R^2}$ of Single Parameter [13].

PM	Parameter	RMSE (%)	MBE (%)	R^2
1	Temperature	1.7735	0.3540	0.9505
2	Cloud cover	0.6988	−0.1272	0.9923
3	Rain	1.0575	−0.0407	0.9823
4	Humidity	0.4214	0.0001	0.9972
5	Wind	1.9124	0.1089	0.9422
6	Pressure	0.8389	0.3620	0.9889
7	Gust	0.9722	−0.1571	0.9851

Table 2.3: RMSE, MBE, and $\boldsymbol{R^2}$ of Multiple Parameter [13].

PM	RMSE (%)	MBE (%)	R^2
8	1.7431	0.6025	0.9520
9	1.5492	0.3090	0.9621
10	1.2130	0.2679	0.9768
11	2.0228	0.4691	0.9354
12	0.8561	0.2822	0.9884

2.2 Error Measurement

Study [14], discusses the significance of accurate solar photovoltaic (PV) power forecasting for efficient energy management. The study evaluates the effectiveness of deep learning models, specifically ANN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), in predicting one-day-ahead PV power generation. The research utilized three error measurement analysis to evaluate the performance of the predictive models, which are MAE, RMSE, and MAPE. MAE measures the average magnitude of the errors between predicted and actual values, providing insight into the accuracy of the model. A lower MAE value indicates higher accuracy of the predictive model. RMSE calculates the square root of the average of the squared differences between predicted and actual values. It measures the deviation between predicted and actual values, with smaller values indicating better performance. RMSE is sensitive to outliers. MAPE measures the accuracy of the predicted results as a percentage, providing a relative assessment of the predictive model's performance. A smaller MAPE value indicates better predictive accuracy. MAE, RMSE, and MAPE were used as evaluation indices to assess the accuracy of the predicted values against the actual values [14] .

The error measurement results in the study [14] show that the MAPE reflects the average percentage difference between predicted and actual values, with lower values indicating more accurate predictions. In this case, the GRU model achieved the lowest MAPE of 19%, suggesting superior accuracy compared to the ANN and LSTM models. For mean absolute error (MAE), the LSTM model exhibits the lowest MAE of 0.931, indicating the smallest average absolute error among the models. The LSTM model also demonstrated the lowest RMSE of 0.940, signifying the smallest overall error in predicting PV power generation. These results show that the LSTM model is better than the ANN and GRU models at accurately predicting PV power generation one day in advance.

CHAPTER 3

METHODOLOGY

This chapter explains about raw data analysis, the steps of data preprocessing, formula for correlation coefficients, regression model and error measurements.

3.1 Project Flowchart

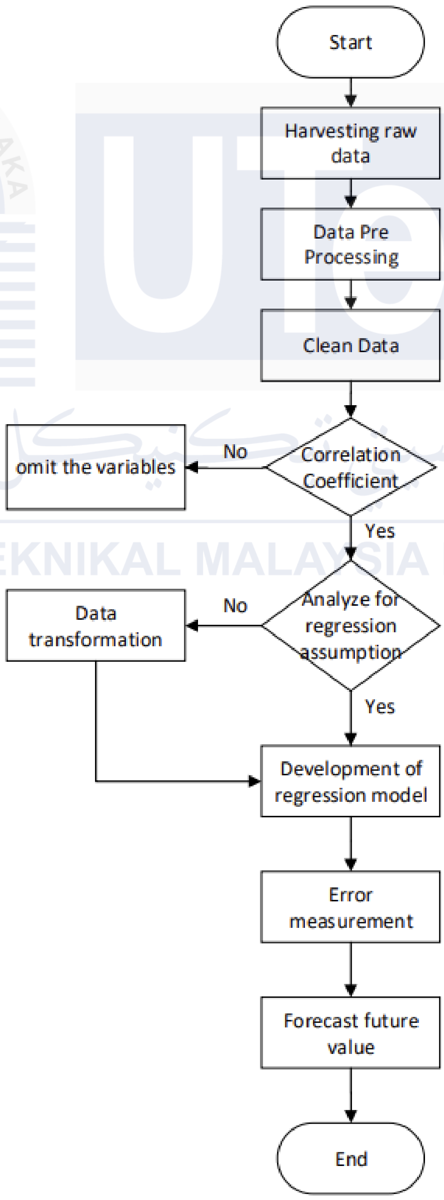


Figure 3.1: Project Flowchart

Figure 3.1 shows the project flowchart. The first step in the project starts with harvesting the raw data. The second step is data pre-processing. There are three steps to be done in this project, which are; data cleaning, merging data, and descriptive statistics of the data. After data pre-processing, the correlation coefficient will be calculated on the clean data. The correlation coefficient is used to quantify the strength and direction of the association between two continuous variables. If the variable is not significant for solar energy, then the variable will be omitted. The next steps is to analyse for regression assumption which are include normality, independence and constant variance assumption. If the data does not meet these assumptions, proceed to the data transformation. Data transformation is the process of modifying the data to improve the fit of the regression model and ensure that the assumptions are met. Transformation techniques such as Min-Max, Z-score, and robust scaling will be done. Following this, the regression model, which are linear regression and multiple linear regression will be developed. After acquiring the regression models, error measurement analysis will be calculated to determine how well it fits the data and clarify the quantity of fit. Finally, the future value will be forecast based on the develop model.

3.2 Harvesting Raw Data

Pyranometers and sensors installed at FTKE collected weather data. The weather data was collected in 2016 from 1st January to 31st December at 12 am until 11:59 pm every day. The reading time is collected every 1-minute interval. Weather data is recorded in Excel file. The weather data collected are global irradiance, tilt irradiance, average temperature, average relative humidity (RH) and average windspeed. Figure 3.2 shows an example weather data for a year collected from the pyranometer.

	Station ID	Data Timestamp	Glob Irrac	Tilt Irrad	Temp Avg	RH Avg	Rain Sum	Panel Tem	WSpd Avg
2	1	2015-12-31 00:01:00 MYT	4	4	25.2	89.8	0	22.2	0.6
3	1	2015-12-31 00:02:00 MYT	4	4	25.2	89.9	0	22.2	0.7
4	1	2015-12-31 00:03:00 MYT	4	4	25.2	90	0	22.2	1.3
5	1	2015-12-31 00:04:00 MYT	4	4	25.2	90	0	22.2	1.2
6	1	2015-12-31 00:05:00 MYT	4	4	25.2	90.1	0	22.2	1.1
7	1	2015-12-31 00:06:00 MYT	4	4	25.2	90.3	0	22.2	1.2
8	1	2015-12-31 00:07:00 MYT	4	4	25.2	90.4	0	22.2	1
9	1	2015-12-31 00:08:00 MYT	4	4	25.2	90.4	0	22.1	1.2
10	1	2015-12-31 00:09:00 MYT	4	4	25.1	90.5	0	22.2	1.2
11	1	2015-12-31 00:10:00 MYT	4	4	25.1	90.5	0	22.1	1.2
12	1	2015-12-31 00:11:00 MYT	4	4	25.1	90.6	0	22.1	1.4
13	1	2015-12-31 00:12:00 MYT	4	4	25.1	90.7	0	22	1.1
14	1	2015-12-31 00:13:00 MYT	4	4	25.1	90.7	0	22.1	1.4
15	1	2015-12-31 00:14:00 MYT	4	4	25.1	90.6	0	22.2	1.5
16	1	2015-12-31 00:15:00 MYT	4	4	25.1	90.6	0	22.2	1.4
17	1	2015-12-31 00:16:00 MYT	4	4	25.1	90.5	0	22.2	1.6
521603	1	2016-12-31 23:43:00 MYT	5	5	24.8	100	0	22.5	0.9
521604	1	2016-12-31 23:44:00 MYT	5	5	24.9	100	0	22.5	0.7
521605	1	2016-12-31 23:45:00 MYT	5	5	24.8	100	0	22.2	0.9
521606	1	2016-12-31 23:46:00 MYT	5	5	24.8	100	0	22.3	0.9
521607	1	2016-12-31 23:47:00 MYT	5	5	24.7	100	0	22.2	0.2
521608	1	2016-12-31 23:48:00 MYT	5	5	24.7	100	0	22.2	0.4
521609	1	2016-12-31 23:49:00 MYT	5	5	24.7	100	0	22.2	0.4
521610	1	2016-12-31 23:50:00 MYT	5	5	24.7	100	0	22.2	0.6
521611	1	2016-12-31 23:51:00 MYT	5	5	24.6	100	0	22.3	0.2
521612	1	2016-12-31 23:52:00 MYT	5	5	24.6	100	0	22.2	0
521613	1	2016-12-31 23:53:00 MYT	5	5	24.6	100	0	22.2	0.3
521614	1	2016-12-31 23:54:00 MYT	5	5	24.6	100	0	22.2	0.5
521615	1	2016-12-31 23:55:00 MYT	5	5	24.6	100	0	22.2	0.7
521616	1	2016-12-31 23:56:00 MYT	5	5	24.7	100	0	22.2	0.4
521617	1	2016-12-31 23:57:00 MYT	5	5	24.7	100	0	22.2	0.1
521618	1	2016-12-31 23:58:00 MYT	5	5	24.7	100	0	22.2	0.4
521619	1	2016-12-31 23:59:00 MYT	5	5	24.7	100	0	22.2	0.2
521620	1	2017-01-01 00:00:00 MYT	5	5	24.7	100	0	22.2	0
521621	1	2017-01-01 00:01:00 MYT	5	5	24.7	100	0	22.2	0.2

Figure 3.2: Weather data collected from pyronometer

Raw data of solar irradiation is collected from solar panels at FTKE. Solar radiation data used in this study is for 1 year from 1st January 2016 to 31st December 2016. The solar radiation data collected in five minutes interval. Figure 3.3 shows solar radiation data in the form of Excel file.

	A	B	C	D	E	F	G	H
1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8
2	CSV-Export	Version: 1.01	Separator: Semicolon					
3								
4		WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C
5		2120211965	2120211965	2120211965	2120211965	2120211965	2120211965	2120211965
6	TimeStamp	A.Ms.Amp	A.Ms.Vol	A.Ms.Watt	Error	E-Total	GridMs.Hz	GridMs.PhV.phsA
7	hh:mm	A	V	W		kWh	Hz	V
8	07:05	0	201.73	0	0: --------	6859.91	50.02	249.38
9	07:10	0	217.46	0	0: --------	6859.91	50.06	249.75
10	07:15	0	222.24	0	0: --------	6859.91	50	249.27
11	07:20	0	227.66	0	0: --------	6859.91	49.96	249.04
12	07:25	0	232.39	0	0: --------	6859.91	50	248.72
13	07:30	0	230.06	0	0: --------	6859.91	49.97	247.69
14	07:35	0	234.45	0	0: --------	6859.91	49.99	246.46
15	07:40	0.21	205.03	40.57	0: --------	6859.92	49.98	244.44
16	07:45	0.34	211.43	70.33	0: --------	6859.92	49.97	243.22
17	07:50	0.35	221.43	77.67	0: --------	6859.92	49.96	242.82
18	07:55	0.41	235.35	94.29	0: --------	6859.93	49.98	241.79
19	08:00	0.48	237.23	112.83	0: --------	6859.93	49.98	240.49
20	08:05	0.76	228.53	172.29	0: --------	6859.94	49.98	239.39
21	08:10	0.89	240.91	213.67	0: --------	6859.96	49.96	240.06
22	08:15	1.19	230.66	274.5	0: --------	6859.97	50	238.67
23	08:20	1.22	230.82	280.17	0: --------	6859.99	49.98	238.97
24	08:25	1.13	237.95	267.43	0: --------	6860.01	49.98	238.54
25	08:30	1.24	250.92	309.25	0: --------	6860.04	49.99	237.48
26	08:35	1.19	250.19	297	0: --------	6860.06	49.99	237.07
27	08:40	1.14	246.72	279.4	0: --------	6860.07	49.96	236.59

	I	J	K	L	M	N	O	P
1	Column9	Column10	Column11	Column12	Column13	Column14	Column15	Column16
2								
3								
4	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C	WRHF0K5C
5	2120211965	2120211965	2120211965	2120211965	2120211965	2120211965	2120211965	2120211965
6	GridMs.PhV.phsB	GridMs.PhV.phsC	Inv.TmpLimStt	InvCtl.Stt	MainModel	Mode	Mt.TotOpTmh	Mt.TotTmh
7	V	V					h	h
8	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10898.95
9	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10899.03
10	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10899.1
11	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10899.17
12	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10899.26
13	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10899.34
14	0	0	1: NoneDrt	1: Off	1: Solar-WR	7: Warten	10439.51	10899.42
15	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10439.58	10899.51
16	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10439.65	10899.59
17	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10439.73	10899.67
18	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10439.83	10899.76
19	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10439.9	10899.84
20	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440	10899.93
21	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.08	10900.01
22	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.16	10900.09
23	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.24	10900.17
24	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.33	10900.26
25	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.41	10900.34
26	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.49	10900.42
27	0	0	1: NoneDrt	1: Off	1: Solar-WR	1: MPP	10440.57	10900.5

Figure 3.3: Solar radiation data in Excel file.

3.3 Data Pre-Processing

Data pre-processing is the initial step in data analysis to improves and refines data quality [15]. Data pre-processing is important because raw data from the original file may contain redundant readings or missing values. Data pre-processing procedure in this study begin with estimation on missing values, followed by merge data and finally the descriptive statistics of the data.

3.3.1 Missing Data

Missing data will affect the accuracy of the forecast model. Missing data from the original data sheet were calculated by taking the average of prior values. There are two methods for determining missing values in small scale missing data and big scale missing data as shown Figure 3.4 and Figure 3.5 respectively. The mean was used to estimate the missing values in the data collection. To solve small scale missing data, the average is calculated based on the data from the previous three readings. For big scale missing data, the mean was calculated based on data from the same time frame on three consecutive days.

1	2016-07-28 02:30:00 MYT	5	5	25	100	0	22.2
1	2016-07-28 02:35:00 MYT	5	4	25	100	0	22.1
1	2016-07-28 02:40:00 MYT	5	4	24.9	100	0	22
1	2016-07-28 02:45:00 MYT	5	5	24.8	100	0	22.2
1	2016-07-28 02:50:00 MYT	5	5			0	22
1	2016-07-28 02:55:00 MYT	5	5	24.9	100	0	22.2
1	2016-07-28 03:00:00 MYT	5	5	25	100	0	22
1	2016-07-28 03:05:00 MYT	5	5	25	100	0	22
1	2016-07-28 03:10:00 MYT	5	5	24.9	100	0	22.2
1	2016-07-28 03:15:00 MYT	4	4	24.8	100	0	22
1	2016-07-28 03:20:00 MYT	4	4	24.6	98.2	0	22
1	2016-07-28 03:25:00 MYT	5	4			0	22
1	2016-07-28 03:30:00 MYT	4	4	24.8	100	0	22
1	2016-07-28 03:35:00 MYT	5	5	24.9	100	0	22.2

2016-07-28 02:30:00 MYT	5	5	25	100	0	22.2
2016-07-28 02:35:00 MYT	5	4	25	100	0	22.1
2016-07-28 02:40:00 MYT	5	4	24.9	100	0	22
2016-07-28 02:45:00 MYT	5	5	24.8	100	0	22.2
2016-07-28 02:50:00 MYT	5	5	24.9	100	0	22
2016-07-28 02:55:00 MYT	5	5	24.9	100	0	22.2
2016-07-28 03:00:00 MYT	5	5	25	100	0	22
2016-07-28 03:05:00 MYT	5	5	25	100	0	22
2016-07-28 03:10:00 MYT	5	5	24.9	100	0	22.2
2016-07-28 03:15:00 MYT	4	4	24.8	100	0	22
2016-07-28 03:20:00 MYT	4	4	24.6	98.2	0	22
2016-07-28 03:25:00 MYT	5	4	24.77	99.4	0	22
2016-07-28 03:30:00 MYT	4	4	24.8	100	0	22
2016-07-28 03:35:00 MYT	5	5	24.9	100	0	22.2

Figure 3.4: Small scale missing data.

1	2016-03-04 02:20:00 MYT	4	4			0	22.5	2.8
1	2016-03-04 02:25:00 MYT	4	4			0	22.5	2.5
1	2016-03-04 02:30:00 MYT	4	4			0	22.5	3.3
1	2016-03-04 02:35:00 MYT	4	4			0	22.5	3.4
1	2016-03-04 02:40:00 MYT	4	4			0	22.5	3.4
1	2016-03-04 02:45:00 MYT	4	4			0	22.5	2.7
1	2016-03-04 02:50:00 MYT	4	4			0	22.5	3.5
1	2016-03-04 02:55:00 MYT	4	4			0	22.5	2.5
1	2016-03-04 03:00:00 MYT	4	4			0	22.5	3
1	2016-03-04 03:05:00 MYT	4	4			0	22.4	2.5
1	2016-03-04 03:10:00 MYT	4	4			0	22.2	2.8
1	2016-03-04 03:15:00 MYT	4	4			0	22.2	2.4
1	2016-03-04 03:20:00 MYT	4	4			0	22.4	2.9
1	2016-03-04 03:25:00 MYT	4	4			0	22.2	2.3
1	2016-03-04 03:30:00 MYT	4	4			0	22.2	3.3
1	2016-03-04 03:35:00 MYT	4	4			0	22.5	3.9
1	2016-03-04 03:40:00 MYT	5	5			0	22.7	3.2
1	2016-03-04 03:45:00 MYT	4	4			0	22.7	3.2
1	2016-03-04 03:50:00 MYT	4	4			0	22.9	3.3
1	2016-03-04 03:55:00 MYT	4	4			0	23	3.6
1	2016-03-04 04:00:00 MYT	4	4			0	23	2.8
1	2016-03-04 04:05:00 MYT	4	4			0	23	3.6
1	2016-03-04 04:10:00 MYT	4	4			0	23	3.1
1	2016-03-04 04:15:00 MYT	4	4			0	22.8	2.4

Figure 3.5: Big scale missing data.

3.3.2 Data Merging

For the data merging, weather data is combined with solar radiation data. However, the weather data needs to be changed from a minute interval data reading to a five-minute interval. Figure 3.6 shows step change to five minutes interval. Figure 3.7 shows the output reading time by solar radiation data in five minutes interval. The function used to convert interval time is KuTools. Kutools is an add-in function for Microsoft Excel designed to enhance productivity by providing a wide range of handy tools and utilities. It offers numerous features that simplify complex tasks, saving time and effort for Excel users.

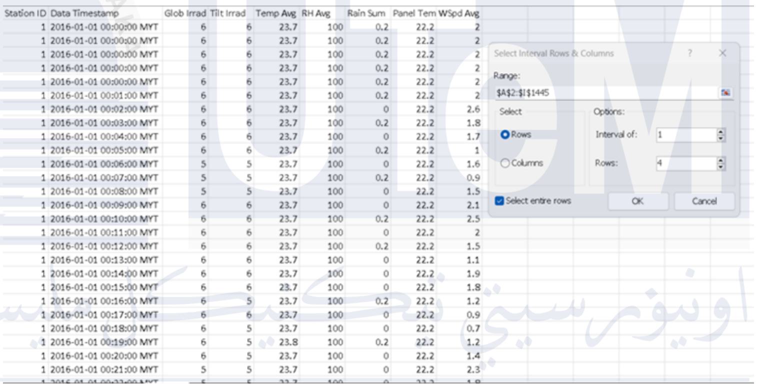


Figure 3.6: Convert data 1-minutes interval to 5-minutes interval by using KuTools function.

Station ID	Data Timestamp	Glob Irrad	Tilt Irrad	Temp Avg	RH Avg	Rain Sum	Panel Tem	WSpd Avg
1	2016-01-01 00:00:00 MYT	6	6	23.7	100	0.2	22.2	2
1	2016-01-01 00:05:00 MYT	6	6	23.7	100	0.2	22.2	1
1	2016-01-01 00:10:00 MYT	6	6	23.7	100	0.2	22.2	2.5
1	2016-01-01 00:15:00 MYT	6	6	23.7	100	0	22.2	1.8
1	2016-01-01 00:20:00 MYT	6	5	23.7	100	0	22.2	1.4
1	2016-01-01 00:25:00 MYT	6	5	23.7	100	0	22.2	1.7
1	2016-01-01 00:30:00 MYT	5	5	23.7	100	0	22.2	2.2
1	2016-01-01 00:35:00 MYT	6	5	23.7	100	0.2	22.2	2.3
1	2016-01-01 00:40:00 MYT	6	6	23.7	100	0	22.2	1.3
1	2016-01-01 00:45:00 MYT	6	6	23.7	100	0	22.2	1.1
1	2016-01-01 00:50:00 MYT	6	5	23.7	100	0	22.2	1.1
1	2016-01-01 00:55:00 MYT	6	6	23.7	100	0	22.2	1.3
1	2016-01-01 01:00:00 MYT	6	6	23.7	100	0.2	22.2	1.8
1	2016-01-01 01:05:00 MYT	6	6	23.6	100	0	22.2	1.9
1	2016-01-01 01:10:00 MYT	5	5	23.6	100	0	22.2	1.2
1	2016-01-01 01:15:00 MYT	6	5	23.7	100	0	22.1	1.2
1	2016-01-01 01:20:00 MYT	6	6	23.7	100	0	22.2	0.4
1	2016-01-01 01:25:00 MYT	5	5	23.7	100	0	22.1	0.6
1	2016-01-01 01:30:00 MYT	6	6	23.6	100	0	22.2	0.5
1	2016-01-01 01:35:00 MYT	6	6	23.6	100	0	22.1	1
1	2016-01-01 01:40:00 MYT	6	6	23.6	100	0	22.2	1.1
1	2016-01-01 01:45:00 MYT	6	6	23.6	100	0	22.2	2.2
1	2016-01-01 01:50:00 MYT	5	5	23.6	100	0	22	1.5
1	2016-01-01 01:55:00 MYT	5	5	23.5	100	0	22	1.7
1	2016-01-01 02:00:00 MYT	6	5	23.5	100	0	22	2
1	2016-01-01 02:05:00 MYT	6	6	23.5	100	0	22	1.6

Figure 3.7: Data converted 5-minutes interval by using KuTools function.

After the weather data is changed to 5 minutes, the solar radiation data is then merged with the weather data. The solar radiation data that has been combined is as shown in Figure 3.8. Following this, the specific yield for thin film solar panels is calculated using the equation (3.1),

$$\text{Specific Yield} = \frac{\text{Total system generated (kWh)}}{\text{Maximum output of PV system (kWp)}} \tag{3.1}$$

THIN FILM SOLAR PANEL ID								21202138	21202138	2120213897	
Station ID	Data Timest	Glob Irrad	Tilt Irrad	Temp Avg	RH Avg	Panel Tem	WSpd Avg	E-Total	E-Total	E-Total	S-Yield
1	2016-01-01	634.00	651.00	3.80	6.90	17.20	7.60	4.83	4.88	4.88	2.34
1	2016-01-02	1194.00	1262.00	7.00	31.80	27.00	6.00	8.78	8.81	8.84	4.24
1	2016-01-03	847.00	897.00	7.40	36.10	26.20	4.30	9.41	9.47	9.51	4.55
1	2016-01-04	1103.00	1221.00	8.20	42.50	33.50	4.90	11.37	11.45	11.48	5.50
1	2016-01-05	1091.00	1185.00	7.70	42.40	37.00	4.50	12.79	12.79	12.87	6.16
1	2016-01-06	1145.00	1219.00	8.40	41.30	37.20	4.90	10.81	10.83	10.87	5.21
1	2016-01-07	1014.00	1066.00	7.20	39.50	30.40	6.10	10.44	10.50	10.54	5.04
1	2016-01-08	1093.00	1183.00	9.40	46.40	35.00	4.40	12.86	12.92	12.96	6.21
1	2016-01-09	1091.00	1189.00	8.10	47.00	29.90	5.60	11.77	11.85	11.88	5.69
1	2016-01-10	1076.00	1145.00	8.60	48.00	34.40	3.50	11.37	11.44	11.48	5.50
1	2016-01-11	1050.00	1084.00	10.00	53.40	36.70	4.50	12.64	12.69	12.75	6.10
1	2016-01-12	1103.00	1175.00	8.40	47.80	35.50	4.10	12.07	12.13	12.17	5.83
1	2016-01-13	1146.00	1211.00	8.90	44.60	39.50	5.00	11.09	11.19	11.21	5.37
1	2016-01-14	1166.00	1214.00	9.20	49.10	36.30	5.50	10.54	10.68	10.67	5.11
1	2016-01-15	1103.00	1120.00	7.80	39.40	41.20	5.00	9.30	9.40	9.40	4.50
1	2016-01-16	1231.00	1223.00	8.60	37.40	41.80	4.70	9.48	9.53	9.58	4.58
1	2016-01-17	1170.00	1223.00	7.40	35.90	37.30	6.10	8.33	8.41	8.43	4.03
1	2016-01-18	1122.00	1157.00	9.40	42.50	44.60	3.90	10.19	10.27	10.30	4.93
1	2016-01-19	1109.00	1170.00	9.00	45.70	34.60	5.10	10.11	10.27	10.26	4.91
1	2016-01-20	1143.00	1251.00	8.50	48.50	37.80	4.10	3.71	4.08	4.01	1.89
1	2016-01-21	1184.00	1317.00	7.20	40.90	29.00	6.40	7.83	7.86	7.90	3.78
1	2016-01-22	1216.00	1307.00	5.90	30.50	29.40	4.80	10.46	10.45	10.54	5.04
1	2016-01-23	1154.00	1159.00	9.20	49.70	31.50	5.30	12.93	13.04	13.10	6.26

Figure 3.8: Merged data of weather variables and solar radiation.

3.3.3 Descriptive Statistic

Descriptive statistics are used to convey quantitative information in an understandable manner. The mean is the average value of a time series over a specified time. The mean is calculated by adding all the values in the dataset and dividing by the total number of values. The mean can be calculated via (3.2),

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.2}$$

where $\bar{x}$ is mean, $\sum_{i=1}^{n} x_i$ is the summing of total variable and n is number of variables.

The standard deviation is a measure of the variability or dispersion of data points within a time series collection. It suggests how much the values deviate from the mean. The standard deviation can be calculated using equation (3.3),

$$Standard\ Deviation = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i}-\bar{x})^2} \quad (3.3)$$

where x_i is the value of the observation data, $\bar{x}$ is the mean value and n is number of the data. In addition, the value of maximum and minimum values will be used for descriptive data. The value of maximum is the larger value in set of data and minimum value is the smallest value in set of data.

3.4 Correlation Coefficient

Correlation coefficient is a measure of the strength of linear relationship between two variables, which are independent variables and dependent variables. Global irradiance, tilt irradiance, temperature average, relative humidity, panel temperature and windspeed average are the independent variables for this study. Specific yield is the dependent variables. The equation of correlation coefficient can be calculated in equation (3.4),

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} {y_i}^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (3.4)$$

where r is correlation coefficient, n is number of set data, x_i is independent variables, y_i is specific yield thin film solar panel, $\sum_{i=1}^{n}$ is sum of data.

Scatter diagrams will show the existence of a linear relationship between variables. The correlation coefficient value ranges from -1 to 1. A correlation coefficient value equal to 1 signifies a perfect positive correlation, meaning that as one variable increases, the other variable increases proportionally and the data points and lie perfectly on a straight line with a positive slope. Meanwhile, when the value of the correlation coefficient is equal to -1, it indicates a perfect negative correlation, whereas one variable increases, the other decreases proportionally and the data points that lie on a straight line with a negative slope. When the value of the correlation coefficient is equal to 0, there is no linear correlation, which means there is no linear relationship between the variables, and the data points appear as a random cloud without any discernible pattern. When the value of the correlation coefficient is close to 1, there is a positive correlation, indicating that as one variable increases, the other tends to increase but not perfectly with the data points showing a general upward trend. When the value of the correlation coefficient is close to -1, there is a negative correlation, suggesting that as one variable increases, the other tends to decrease but not perfectly with the data points showing a general downward trend.

3.5 Regression Assumption

In regression analysis, several key assumptions are required to make sure the results are reliable. The trend line pattern on the residual plot allows one to observe the regression assumption. There are three regression assumptions on the residuals, which are independence, normality, and constant variance.

Regression analysis requires independence among the dataset's observations, ensuring that the residuals (errors) from one observation do not correlate with those from another. The aim for independence assumption is to ensures that the information provided by each data point is unique and not influenced by other data points. Normality in regression analysis means that the residuals, which are differences between observed and predicted values should follow a normal distribution. Ensuring the normality of residuals guarantees that the estimated regression coefficients are unbiased, and that hypothesis tests and confidence intervals are valid. If residuals are not normally distributed, it can lead to incorrect inferences about the relationships between variables, affecting the overall reliability of the regression analysis. Constant variance means that the residuals should have constant variance across all levels of the independent variables. In other words, the spread of the residuals should remain consistent regardless of the predictor's value. Constant variance ensures that the model maintains equal precision across all values of the independent variables, leading to more reliable coefficient estimates. Non-constant variance can cause inefficient estimates and invalid standard errors, affecting hypothesis tests.

The null hypothesis that the coefficient is equal to zero will have no effect and be tested by the *p*-value for each term. The null hypothesis can be rejected if the *p*-value is less than 0.05. In other words, a predictor with a low *p*-value is likely to be a significant addition to your model, as changes in the predictor's value are correlated with changes in the response variable. In contrast, a *p*-value that is significantly larger indicates that the response is not correlated with variations in the predictor.

3.6 Data Normalization

Data normalization is a preprocessing technique employed to standardise the range of independent variables before the application of regression models. Data normalization guarantees consistent scaling to prevent features with larger scales from having an excessive impact on the regression model, which consequently avoids biased on parameter values. There are some techniques of data normalization such as Min-Max, Z- score and robust scaling.

Min-Max Scaling is a data normalization technique that transforms the features of data to lie within a specified range, usually [0, 1] or [-1, 1]. This method is simple and often used when relationships among the data points needs to maintain. Min-Max Scaling is very useful when data does not fit a normal distribution. This technique is effective because it makes no assumptions about the data, making it suitable for a wide range of data distributions. Min-Max scaling can be calculated as in equation (3.5),

$$x' = \frac{x - min\ (x)}{max(x) - min\ (x)} \tag{3.5}$$

where x' is the scaled value, x is the original value in data, min (x) and min (x) is the minimum and maximum value in data.

Statistical techniques such as *Z*-score normalization are employed to convert data into a standard normal distribution with a mean of 0 and a standard deviation of 1. When the *Z*-score is positive, a data point is above the mean, and when the *Z*-score is negative, it is considered below the mean. *Z*-score normalization can be described as in equation (3.6),

$$z = \frac{x - \mu}{\sigma} \tag{3.6}$$

where z is the standardized value, x is the original data, μ is mean and σ is standard deviation.

Robust scaling is a data normalization approach that is highly successful for datasets that include outliers. The median provides a more consistent assessment of central tendency in the presence of outliers, in contrast to the mean. The stability of the median makes it a preferable option in datasets that contain extreme values. The scaled value can be calculated as in equation (3.7)

$$x' = \frac{x - median}{IQR} \tag{3.7}$$

where x' is scaled value, x is original data and IQR is the interquartile range.

3.7 Development of Regression Model

Regression models that will be used in this project are linear regression and multiple linear regression. The linear regression equation can be described as in equation (3.8),

$$Y_i = A + B_1 x_1 \tag{3.8}$$

where Y_i is specific yield, A is Y-intercept, x_1 is the highly correlated independent variables, that is the weather data and B_1 is the slope. The equation of multiple linear regression is as in equation (3.9),

$$Y_i = A + B_1 x_1 + B_2 x_2 + \cdots B_n x_n \tag{3.9}$$

where Y_i is specific yield from thin film solar panel, A is Y-intercept, $B_1, B_2, \dots, B_n$ with are the slope and $x_1, x_2, \dots, x_n$ are the significant weather variable.

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.

3.8 Error Measurement

Error measurement is important in regression analysis because it determines the predictive accuracy of a regression model. There are several formula that have been used for calculating the performance and accuracy of regression models. Error measurements that implemented in this study are the mean absolute error (MAE), mean square error (MSE), and mean percentage error (MPE).

3.8.1 Mean Absolute Error (MAE)

The MAE calculates the average number of the errors in the forecast model. It computes the average distance between predicted and true values, with distance defined as the absolute difference between predicted values and the true values. The MAE equation is expressed as in equation (3.10),

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \widehat{Y}_i| \tag{3.10}$$

where $\widehat{Y}_i$ is predicted value, Y_i is actual value and n is number of data.

3.8.2 Mean Square Error (MSE)

Mean Square Error (MSE) is a standard statistic for calculating the average squared difference between expected values and actual values. The MSE calculates the sum of the squared differences between each expected and actual value, then divides the amount by the total number of data points. The equation is expressed as in (3.11),

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 \tag{3.11}$$

where$\widehat{Y}_i$ is predicted value, Y_i is actual value and n is number selected data.

3.8.3 Mean Percentage Error

The MPE is a formula that analyses the predictive model's accuracy by calculating the average percentage difference between the predicted and actual values. It is particularly effective in situations where the relative scale of errors is more significant than their absolute size. MPE can be calculated as in equation (3.12),

$$MPE = |\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_i})| \times 100 \qquad (3.12)$$

where $\widehat{Y}_i$ is predicted value, Y_i is actual value and n is number of selected data.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Results

Correlation coefficients were calculated for each independent variable. There are 365 total of days from 1st February until 31st December 2016. The data was collected between 7:00 am until 7:00 pm. Result from the analysis is discussed in the next subchapter.

4.2 Daily Trend of Thin Film Solar Panel

Figure 4.1 shows the daily trend of specific yield from thin film solar panel in January. The graph shows specific yield vs days. The maximum specific yield value for January 2016 is 6.26 kwh/kwp on 23 January and the minimum value is 1.89 kwh/kwp, that is on 20 January.

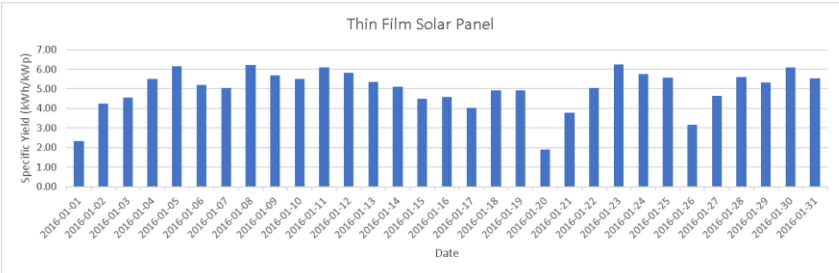


Figure 4.1: Daily specific yield trend for January

Figure 4.2 shows the daily trend of specific yield in February using a pictograph. The graph shows that the specific yield was lower on cloudy days. This is because the clouds block the sun's rays from reaching the solar panel. It can be concluded that the maximum value of specific yield produced for the month of February is 7.64 kWh/kWp on February 9 and the minimum value is 2.11 kWh/kWp on February 19.

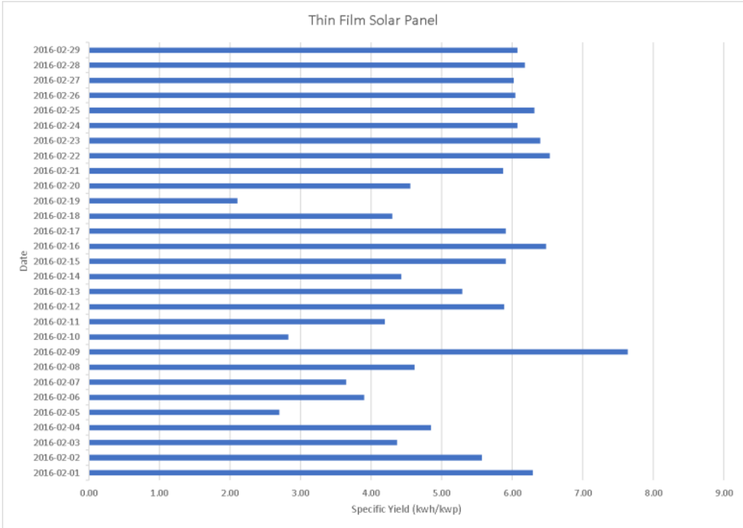


Figure 4.2: Daily trend of specific yield in February

Figure 4.3 shows the daily specific yield trend in August using a line graph. The graph shows a clear daily cycle in specific yield, with a peak in the middle of the day and lower values in the morning and evening. This is because the sun's irradiance is strongest at midday when it is directly overhead. The maximum value for the specific yield produced is 5.17 kWh/kWp on August 23 and the minimum value is 1.70 kWh/kWp on August 3.

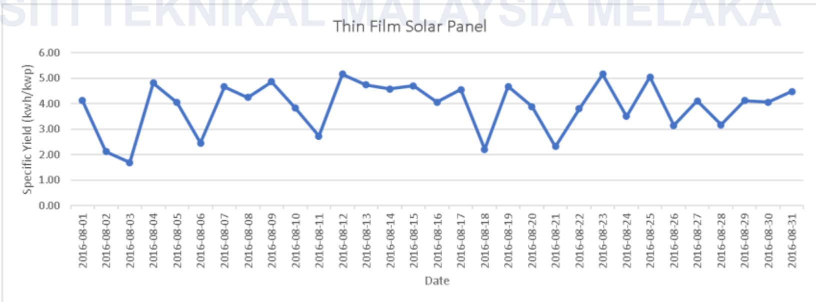


Figure 4.3: Daily specific yield trend in August.

In conclusion, the specific yield flow graph shows that the specific yield is fluctuate differently every month. The highest specific yield was around 7.64 kWh/kWp on February 9 and the lowest was around 0.34 kWh/kWp on December 12. Figure 4.4 shows daily specific yield trend for one year by using a line graph. However, the graph shows that from Jun 2016 to month July 2016 the pattern of specific yield produced is constant which is 3.79 kWh/kWp due to missing values of specific yield during that period reason to malfunction of system. Therefore, missing values are estimated by using average calculation.

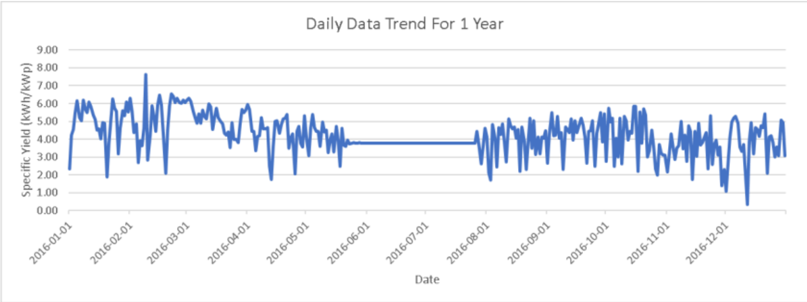


Figure 4.4: Daily specific yield trend for one year.

4.3 Descriptive Statistic

Table 4.1 shows the descriptive statistics for the whole one year. The value of maximum, minimum, mean and standard deviation for each independent variable and dependent variable are presented in Table 4.1. The standard deviation shows how far each value lies from the mean. A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

Table 4.1: Descriptive statistic for one year.

Descriptive Statistic	Global Irradiance $(\mathbf{Wh/m^2})$	Tilt Irradiance $(\mathbf{Wh/m^2})$	Temp Average (°C)	Relative Humidity (%rh)	Panel Temperature (°C)	Windspeed Average (m/s)	Specific Yield (kWh/kWp)
Max	1340.00	1409.00	25.80	63.50	73.70	11.40	7.64
Min	433.00	423.00	2.50	6.90	8.60	1.83	0.34
Mean	1044.36	1026.08	8.08	39.77	29.05	4.35	4.24
Standard Deviation	166.48	181.84	2.12	8.88	9.27	1.59	1.07

From Table 4.1, the maximum value of global irradiance is 1340 Wh/m^2 and the minimum value is 433 Whm^2. For the mean of global irradiance, it is 1044.36 Wh/m^2 and the value of standard deviation is 166.48 Whm^2. The maximum value of tilt irradiance is 1409 Wh/m^2 and minimum value is 423 Wh/m^2. The mean value of tilt irradiance is 1026.08 Wh/m^2 and the value of standard deviation is 181.84

Wh/m^2. The maximum temperature average is 25.80 °C and the minimum is 2.50 °C. The mean of temperature average is 1026.08 °C and the standard deviation is 2.12 °C. The maximum value of relative humidity is 63.50 %rh and the minimum value is 6.90 %rh. For the mean of relative humidity, it is 39.77 %rh with standard deviation of 8.88 %rh. The maximum panel temperature is 73.70 °C and the minimum is 8.60 °C. The average panel temperature is 29.05 °C and the standard deviation is 9.27 °C. The maximum value of windspeed average is 11.40 m/s, the minimum value is 1.83 m/s and the mean is 4.35 m/s with standard deviation of 1.59 m/s. Global irradiance, tilt irradiance, temperature average, relative humidity, panel temperature and windspeed average are independent variables for this project. The specific yield is dependent variables. The maximum specific yield is 7.64 kWh/kWp and minimum value is 0.34 kWh/kWp. The mean of specific yield is 4.24 kWh/kWp with standard deviation of 1.07 kWh/kWp.

4.4 Correlation Coefficient

This subchapter presents the correlation coefficient for each variable used in this study.

4.4.1 Single Variable

Correlation coefficient for linear regression of single variable is as shown Table 4.2. The highest correlation coefficient value is on relative humidity, which is 0.5848 (positive linear) and follow by tilt irradiance, global irradiance and temperature average. The lowest correlation coefficient is for that windspeed average, which is 0.2835. Relative humidity shows the highest correlation, which means that there is a strong positive correlation between the variables and specific yield. In other words, as relative humidity average increases, specific yield also tends to increase.

Table 4.2: Correlation coefficient for single variable

Single Variable	Correlation Coefficient ($\boldsymbol{r}$)
Relative Humidity (RH)	**0.5848**
Tilt Irradiance (TI)	0.4814
Global Irradiance (GI)	0.4715
Temperature Average (TA)	0.3316
Wind Speed Average (WSpd Ave)	0.2852

4.4.2 Multiple Variable

From table 4.3, it is shown the highly correlation coefficient variables to specific yield are combination of relative humidity, tilt irradiance, global irradiance, temperature average and windspeed average which is 0.6568. The lowest correlation coefficient variables to the specific yield is 0.6398, which are combination of relative humidity and tilt irradiance.

Table 4.3: Correlation coefficient of multiple variable

Multiple Variable	Correlation Coefficient (r)
Relative Humidity and Tilt Irradiance	0.6398
Relative Humidity,Tilt Irradiance and Global Irradiance	0.6478
Relative Humidity,Tilt Irradiance, Global Irradiance and Temperature Average	0.6486
Relative Humidity,Tilt Irradiance, Global Irradiance, Temperature Average and Windspeed Average	**0.6568**

4.5 Regression Assumption

Regression assumptions are important requirements for the validity and reliability of regression models. Regression assumption will determine the suitability of the data with the regression models. Figures 4.5 to Figure 4.8 show the regression assumption on multiple models, which uses relative humidity and tilt irradiance as a variable. Figure 4.5 shows normality probability plot for normality assumptions.

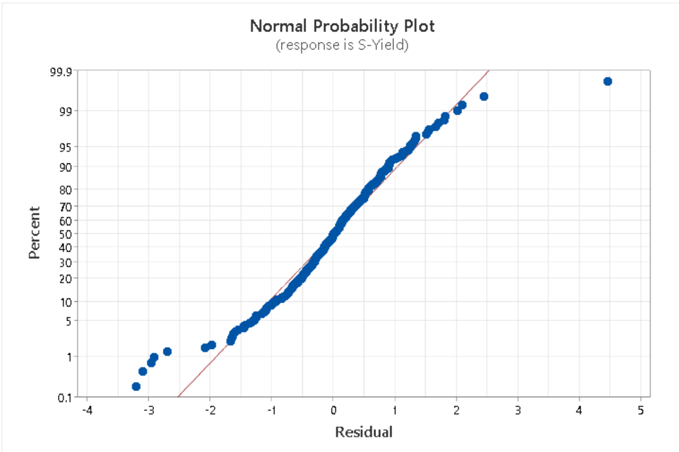


Figure 4.5: Normal Probabilty Plot.

Figure 4.5 shows a reasonable fit for the data, as the residuals approximate a normal distribution. This validation of the model's assumptions improves confidence in its predictive accuracy and reliability. Nevertheless, the minor deviations observed are not significant enough to undermine the overall fit of the model. Figure 4.6 shows scale the error plot for independence assumption.

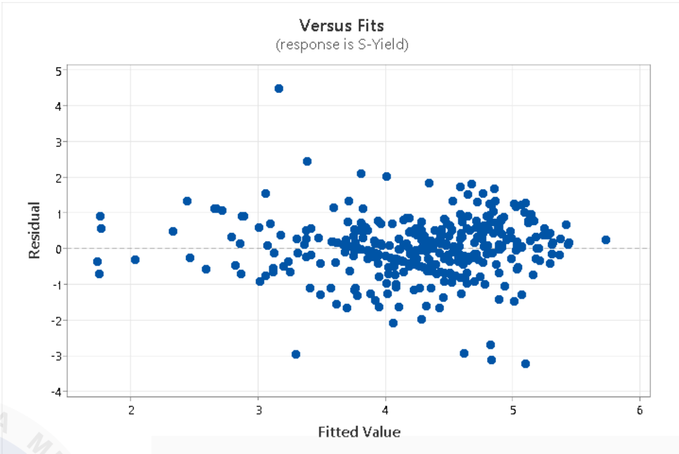


Figure 4.6: Plot on independance assumptions.

The random distribution of residuals around zero in Figure 4.6 supports the regression model's assumptions and shows that it can accurately predict specific yield based on the given predictors. This validation indicates that the model is appropriately specified and captures the underlying relationship between the variables without systematic errors. However, while the model looks good enough to address the outliers it could further refine its accuracy. Figure 4.7 shows residuals vs. observation order plot for constant variance assumption.

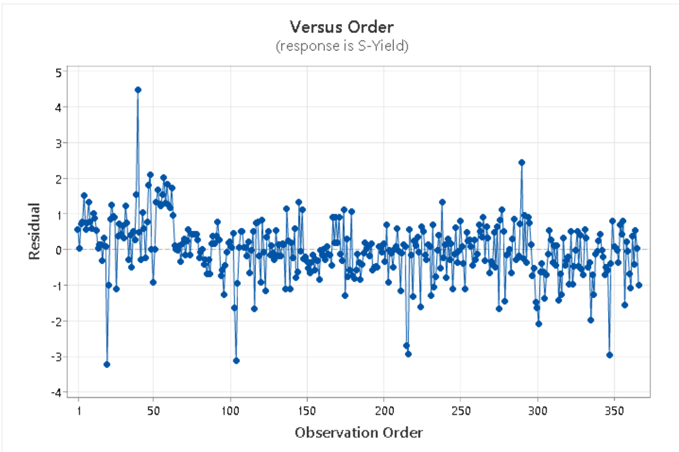


Figure 4.7: Residuals vs observation order plot

Figure 4.7 shows that the random distribution of residuals around zero confirms the regression model's validity, indicating that the residuals are independent and there is no autocorrelation. Figure 4.8 shows the histogram of residual.

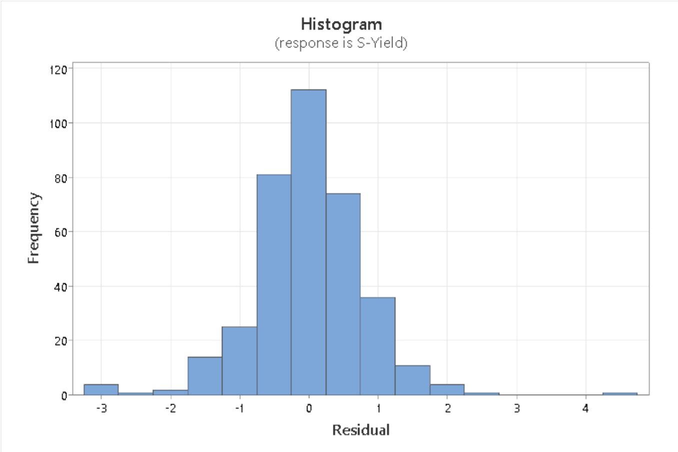


Figure 4.8: Histogram of residual

Figure 4.8 shows the regression model's histogram of residuals shows an approximately normal distribution that is centred around zero, thereby confirming the model's suitability and supporting the assumption of normally distributed errors.

4.6 Development Regression Model

There are nine models developed to forecast the specific yield of Thin Film solar panel. Five models are based on single variable, and four models are based on combination of multiple variables. Table 4.4 shows the regression equation for single model. Single variable model means each model uses only one variable. For M1, the model uses relative humidity (RH), M2 is Tilt Irradiance (GA), M3 is global irradiance (GI) and M4 is temperature average (TA) and M5 is windspeed average (Wspd Avg).

Table 4.4: Single Model Regression

Single Model	Regression Equation
M1- Relative Humidity (RH)	$Specific\ Yield = 1.437 + 0.0704\ RH$
M2- Tilt Irradiance (TI)	$Specific\ Yield = 1.335 + 0.0028\ TI$
M3- Global Irradiance(GI)	$Specific\ Yield = 1.077 + 0.0030\ GI$
M4- Temperature Average (TA)	$Specific\ Yield = 2.885 + 0.1672\ TA$
M5- Windspeed Average (Wspd Avg)	$Specific\ Yield = 3.402 + 0.1917\ Wspd\ Avg$

The multiple-regression models are shown in Table 4.5. Multiple regression models are those that use a combination of multiple variables. For instance, multiple regression model for M6 model is based on relative humidity and tilt irradiance. After that, the M7 model comprises relative humidity, tilt irradiance, and global irradiance. The multiple variables for M8 are determined by the following factors relative humidity, tilt irradiance, global irradiance and temperature average. Lastly, combination of multiple variables of M9 are relative humidity, tilt irradiance, global irradiance, temperature average and windspeed average.

Table 4.5: Multiple Model Regression

Multiple Model	Regression Equation
M6- RH, TI	$Specific\ Yield = 0.290 + 0.0559RH + 0.0017TI$
M7- RH, TI, GI	$Specific\ Yield = 0.518 + 0.0613RH + 0.0032TI - 0.0020GI$
M8- RH, TI, GI , TA	$Specific\ Yield = 0.557 + 0.0643RH + 0.0032TI - 0.0019GI - 0.0206TA$
M9- RH, TI, GI, TA, Wspd Ave	$Specific\ Yield = 0.390 + 0.06234RH + 0.0025TI - 0.0013GI - 0.0147TA + 0.0782Wspd\ Avg$

4.7 Error Measurement Analysis

Error measurement functions is used to evaluate how accurate a model prediction match with the actual data. There are three types of error measurement is used for single variable model and multiple variable models, which are MAE, MSE and MPE.

4.7.1 Single Variable Model

Table 4.6 shows error measurements for single-variable regression models that predict a specific outcome, each of which uses a different independent variable The most accurate to predict future value of single variable model is based on relative humidity (M1). It has the lowest MAE (0.64) and MSE (0.78) with a moderate MPE (12.35%), which suggests a high level of predictive accuracy than other linear regression model. The model that uses the wind speed average (M5), on the other hand, is the worst. It has the highest MAE (1.01), MSE (2.13), and MPE (37.75%), that means the predictions are further away from the actual values as shown in Table 4.6.

Table 4.6: Error Measurement Single Variable

Model	MAE	MSE	MPE(%)
M1- RH	**0.64**	**0.78**	**12.35**
M2- Tilt Irrad	0.86	1.40	43.44
M3- Global Irrad	0.80	1.26	21.40
M4- Temp Ave	0.86	1.34	10.81
M5- Windspeed Ave	1.01	2.13	37.75

4.7.2 Multiple Variable Model

Table 4.7 shows the error measurement for multiple-variable regression models for multiple regression models. The best and most accurate model is M6, which is based on relative humidity and tilt irradiance. It shows the highest performance, with the lowest errors (MAE = 0.56, MSE = 0.71 and MPE = 5.85%). When extra variables like GI, TA, and Wspd Avg are added to models, the errors go up and the predictions are further away from the actual values. However, the combination variables which is relative humidity and tilt irradiance are the most accurate.

Table 4.7: Error Measurement Multiple Variable Model

Model	MAE	MSE	MPE (%)
M6- RH, TI	**0.56**	**0.71**	**5.85**
M7- RH, TI, GI	0.58	0.72	15.54
M8- RH, TI, GI, TA	0.59	0.72	15.27
M9- RH, TI, GI, TA, Wspd Avg	0.58	0.73	15.34

The M1 model (Specific Yield = 1.437 + 0.0704RH) and the M6 model (Specific Yield = 0.290 + 0.0559RH + 0.0017TI) have the lowest error measurement values for single-variable and multiple-variable models. After making a comparison between M1 and M6, the M6 model has the lowest error measurement value, which means it is suitable for predicting actual specific yield in 2017.

.

4.8 Prediction Future Value of Specific Yield

The line graphs show the specific yield in kilowatt (kW) and various independent variables which are relative humidity, tilt irradiance, global irradiance, temperature average, and windspeed average.

4.8.1 Single Variable Model

Figure 4.9 shows the prediction of specific yield based on M1 model, that is relative humidity as independent variable. The equation of the model is Specific Yield=1.437 + 0.0704RH. Both lines show a similar pattern, with the predicted values are close to the actual specific yield. The model predicts specific yield reasonably well based on relative humidity.

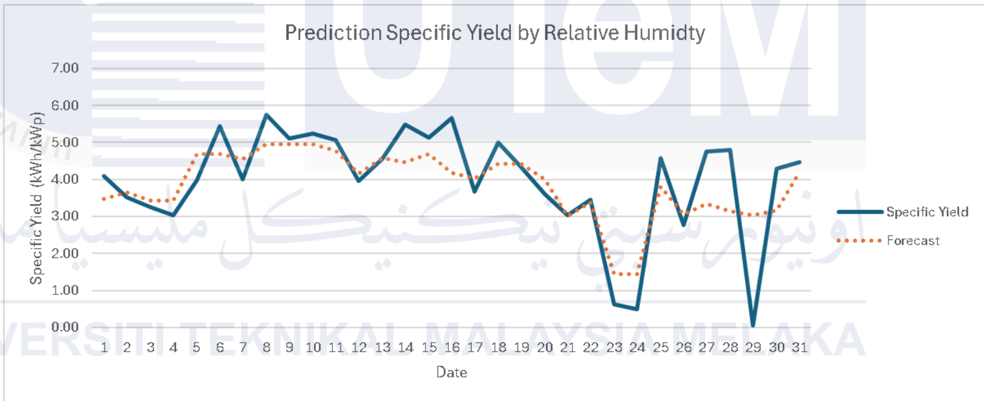


Figure 4.9: Prediction specific yield by M1.

Figure 4.10 shows the prediction of specific yield based on M2 model, that is tilt irradiance as independent variable. The equation of the model is Specific Yield=1.335 + 0.0029TI. The lines generally follow a similar trend, indicating that the model captures the relationship between predicted value specific yield by tilt irradiance and actual specific yield to some extent though there are some deviations observed.

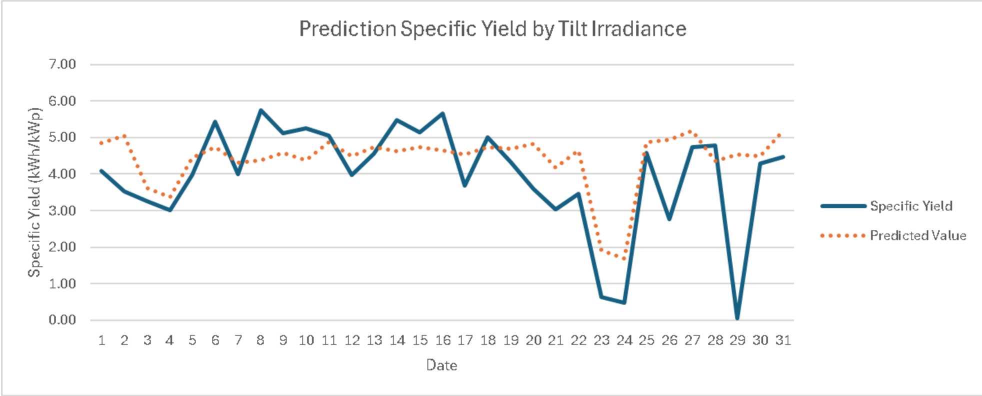


Figure 4.10: Prediction specific yield by M2 model.

Figure 4.11 shows the prediction of specific yield based on M3 model, that is global irradiance as independent variable. The equation of the model is Specific Yield=1.077 + 0.0030GI. The figure shows that some of the predicted value is near to the measured and some values is far from the measured values.

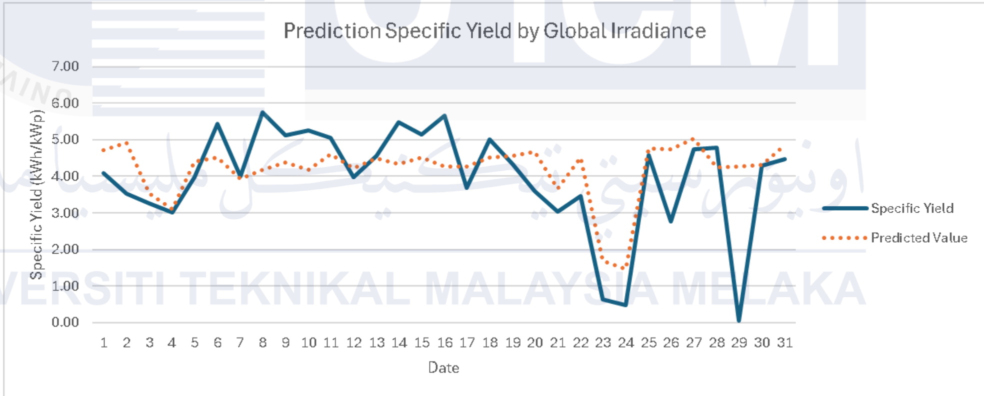


Figure 4.11: Prediction specific yield by M3.

Figure 4.12 shows the prediction of specific yield based on M4 model, that is temperature average as independent variable. The equation of the model is Specific Yield=2.885 + 0.1672TA. From the Figure 4.12, the predicted values show less variability compared to the actual specific yield. It looks likes temperature average is not enough to correctly predict specific yield because the predicted values are far from the actual specific yield.

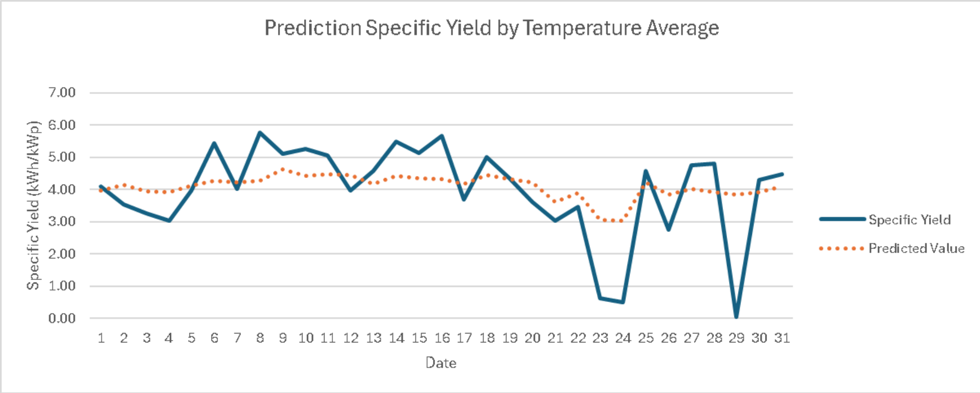


Figure 4.12: Prediction on specific yield by M4 model.

Figure 4.13 shows the prediction of specific yield based on M5 model, that is windspeed average as independent variable. The equation of the model is Specific Yield=3.402 + 0.1917Wspd Avg. The graph shows predicted values is further away from than the actual specific yield. Hence this model is not suitable for future value specific yield at FTKE, UteM.

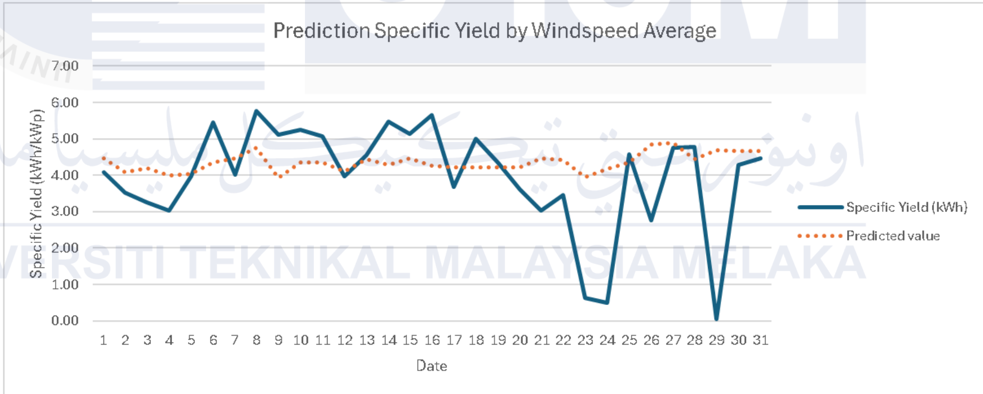


Figure 4.13: Prediction specific yield by M5 model.

4.8.2 Multiple-Variable Model

Figure 4.14 shows the prediction of specific yield based on (M6) model, which are relative humidity and tilt irradiance. The equation of this model is Specific Yield = 0.290 + 0.0559RH + 0.0017TI. The figure shows predicted values generally follow the trend of the actual specific yield, that means this model predicts accurately.

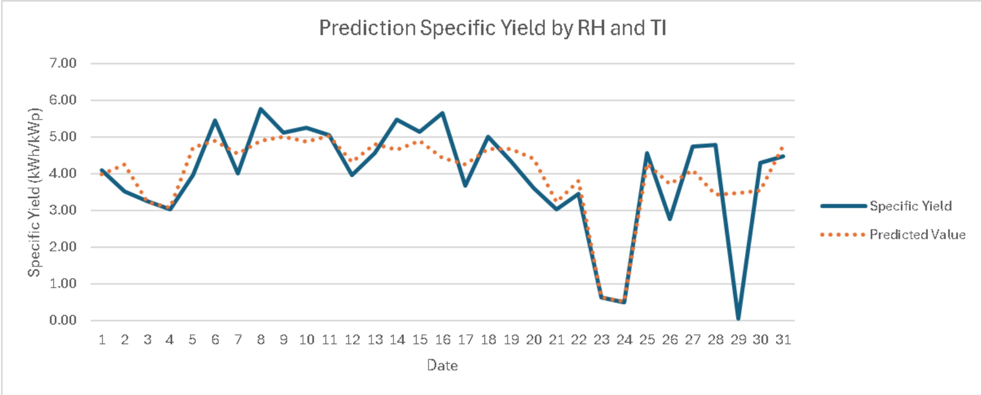


Figure 4.14: Prediction of specific yield by the M6 model.

Figure 4.15 shows the prediction of specific yield based on (M7) model, which are relative humidity (RH), tilt irradiance (TI) and global irradiance (GI). The equation of this model is Specific Yield=0.518 + 0.0613RH + 0.0032TI – 0.0020GI. The figure shows that the predicted value is close to the measured value, but not as good as showing RH and TI.

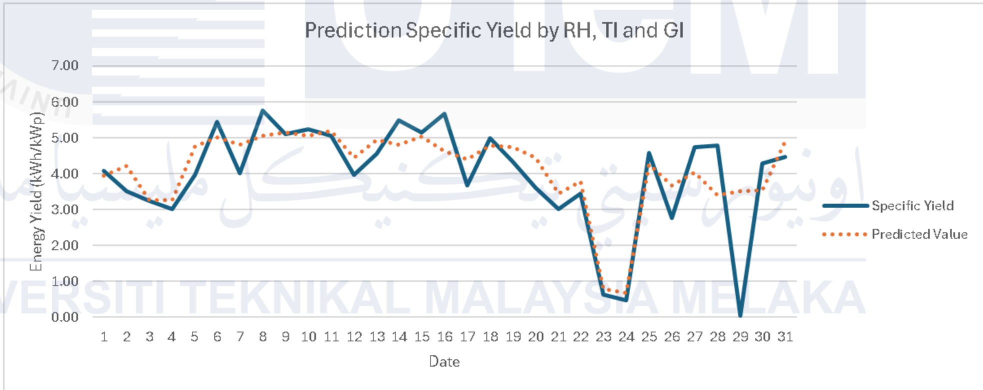


Figure 4.15: Prediction specific yield by M7 model.

Figure 4.16 shows the prediction of specific yield based on (M8) model, which are relative humidity, tilt irradiance, global irradiance and temperature average. The equation of this model Specific Yield=0.557 + 0.0643RH + 0.0032TI – 0.0019GI – 0.0206TA. The figure shows that some of the predicted value is close to the measured value, and some value is far from the measured value.

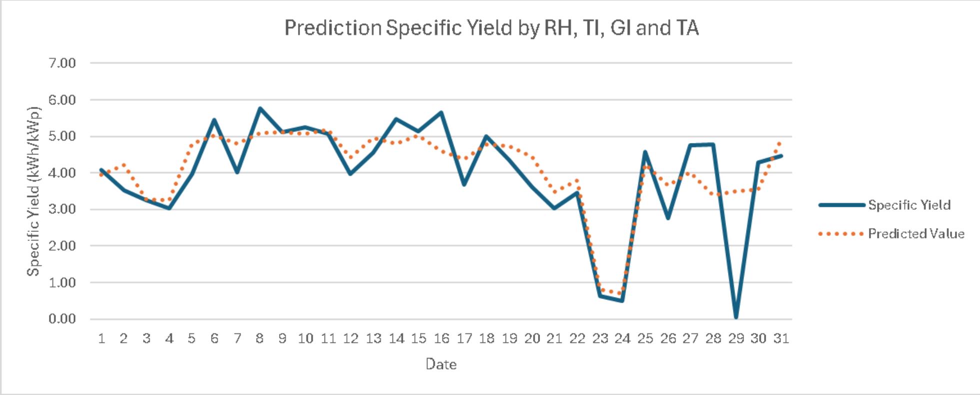


Figure 4.16: Prediction specific yield by M8 model.

Figure 4.17 shows the prediction of specific yield based on M9 model, which are relative humidity, tilt irradiance, global irradiance and temperature average. The equation of this model is Specific Yield= 0.390 + 0.0623RH + 0.0025TI – 0.0013GI – 0.0147TA + 0.0782Wspd Avg. From Figure 4.17, the predicted values show less variability compared to the actual specific yield.

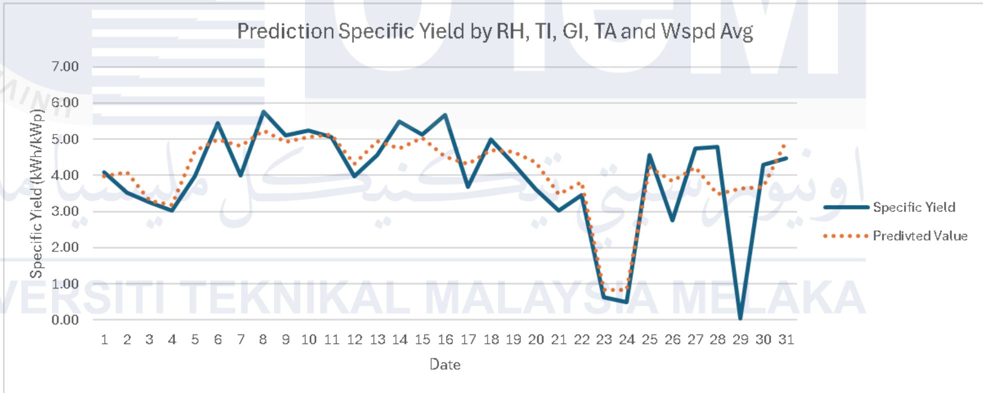


Figure 4.17: Prediction Specific Yield by M9 model.

Therefore, to know which model is the best model for predicting specific yield, error measurements such as MAE, MSE, and MPE need to be compared. When comparing the error between the models, the best model to predict specific yield in 2017 for one month for a single variable is M1, which is relative humidity as a variable because it has a low measurement error value. This model's equation is Specific Yield = 1.437 + 0.0704 RH. For multiple variables, M6 uses relative humidity and tilt irradiance as variables because it has a low measurement error value. The equation of this model is Specific Yield = 0.290 + 0.0704RH – 0.0017TI. According to the comparison of the M1 and M6 models, the M6 model is the most appropriate for use in this specific yield prediction because it has the lowest measurement error value and a graph that is close to the actual specific yield.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

In conclusion, for linear regression, the best fit model is M1, which is Specific Yield = 1.437 + 0.0704RH. This model can be used to predict the specific yield of thin film solar panels if relative humidity is used as a variable. The reason is the lowest error compared to other models. The MAE is 0.64, MSE is 0.78, and MPE is 12.35%. Otherwise, for multiple regression, the best model is Model 6, which is Specific Yield = 0.290 + 0.0559RH + 0.0017TI. This model can forecast the specific yield by using relative humidity and tilt irradiance as variables. The MAE is 0.56, MSE is 0.71, and MPE is 5.85%. The best model to predict specific yield is Model 6 because the measurement error is the lowest than other variables, which means it can predict specific yield accurately.

5.2 Future Works

In future work, conducting a long-term performance analysis of thin film solar panels can be valuable for assessing their durability and sustainability. This involves monitoring specific yield over an extended period to identify any degradation patterns or changes in performance over time. Long-term data could contribute to the development of maintenance strategies and improve the overall reliability of thin film solar technology. Additionally, validating the regression model through field testing and real-world applications is essential to ensure its practical applicability. Collaborating with solar energy installations and collecting data from diverse geographical locations and varying climate conditions will enhance the model's reliability and broaden its scope of relevance.

Furthermore, conducting a comprehensive cost-benefit analysis can provide valuable insights into the economic feasibility of using thin film solar panels under different conditions. Integrating cost-related variables into the regression model will assist in evaluating the economic viability of thin film solar technology and guide decision-making processes.

CHAPTER 6

REFERENCES

[1] M. Y. ERTEN and H. AYDİLEK, "Solar Power Prediction using Regression Models," *Uluslararası Muhendislik Arastirma ve Gelistirme Dergisi*, vol. 14, no. 3, pp. 1–1, Dec. 2022, doi: 10.29137/umagd.1100957.

[2] M. Masri, R. Badlishah, M. Irwanto, and H. Alam, "Solar Radiation Potential as Energy Source of Photovoltaic Powered Uninterrupted Power Supply in Perlis, Northern Malaysia," *IOSR Journal of Electrical and Electronics Engineering*, vol. 9, no. 6, pp. 31–36, 2014, doi: 10.9790/1676-09623136.

[3] Q. Yang Thee, "The Performance Analysis of a Three-Phase Grid-Tied Photovoltaic System in a Tropical Area," *GSTF Journal of Engineering Technology*, vol. 2, no. 4, Apr. 2014, doi: 10.5176/2251-3701_2.4.94.

[4] M. A. Green, K. Emery, Y. Hishikawa, W. Warta, and E. D. Dunlop, "Solar cell efficiency tables (version 43)," *Progress in Photovoltaics: Research and Applications*, vol. 22, no. 1, pp. 1–9, Jan. 2014, doi: 10.1002/pip.2452.

[5] Assoc Prof Dr. Gan Chin Kim, "Energy and Power System Research Group," https://ceria.utem.edu.my/research-asset-inventory/pvs-research-lab.html, Dic 6, 2014.

[6] A. A. F. Husain, M. H. A. Phesal, M. Z. A. A. Kadir, U. A. U. Amirulddin, and A. H. J. Junaidi, "A decade of transitioning malaysia toward a high-solar pv energy penetration nation," *Sustainability (Switzerland)*, vol. 13, no. 17. MDPI, Sep. 01, 2021. doi: 10.3390/su13179959.

[7] S. Dukkipati, V. Sankar, and P. Srinivasa Varma, "Forecasting of Solar Irradiance using Probability Distributions for a PV System: A Case Study," 2019.

[8] K. B. Gyamfi and S. Appiah, "Probability Distributional Analysis of Hourly Solar Irradiation in Kumasi-Ghana Article in," *International Journal of Business and Social Research*, 2013, doi: 10.18533/ijbsr.v3i3.57.

[9] K. Gairaa, F. Chellali, S. Benkaciali, Y. Messlem, and K. Abdallah, "Daily global solar radiation forecasting over a desert area using NAR neural networks comparison with conventional methods," in *2015 International Conference on Renewable Energy Research and Applications, ICRERA 2015*, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 567–571. doi: 10.1109/ICRERA.2015.7418477.

[10] A. Ghanbarzadeh, A. R. Noghrehabadi, E. Assareh, and M. A. Behrang, "Solar radiation forecasting based on meteorological data using artificial neural

networks," in *IEEE International Conference on Industrial Informatics (INDIN)*, 2009, pp. 227–231. doi: 10.1109/INDIN.2009.5195808.

[11] Y. M. Irwan, I. Daut, I. Safwati, M. Irwanto, N. Gomesh, and M. Fitra, "An estimation of solar characteristic in Kelantan using Hargreaves model," in *Energy Procedia*, Elsevier Ltd, 2013, pp. 473–478. doi: 10.1016/j.egypro.2013.07.053.

[12] S. Ibrahim, I. Daut, Y. M. Irwan, M. Irwanto, N. Gomesh, and Z. Farhana, "Linear regression model in estimating solar radiation in perlis," in *Energy Procedia*, Elsevier BV, 2012, pp. 1402–1412. doi: 10.1016/j.egypro.2012.05.156.

[13] H. Ahmed Kutty, M. H. Masral, and P. Rajendran, "Regression model to predict global solar irradiance in Malaysia," *International Journal of Photoenergy*, vol. 2015, 2015, doi: 10.1155/2015/347023.

[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 (Switzerland)*, vol. 12, no. 15, Aug. 2022, doi: 10.3390/app12157529.

[15] Institute of Electrical and Electronics Engineers and Manav Rachna International Institute of Research and Studies, *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing : trends, prespectives and prospects : COMITCON-2019 : 14th-16th February, 2019*.

APPENDICES A

MINITAB OUTPUT FOR M1 MODEL

Regression Equation

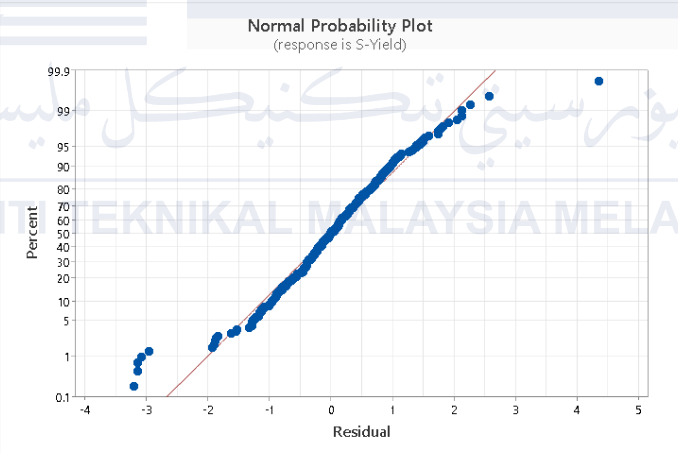
S-Yield = 1.437 + 0.07035 RH Avg

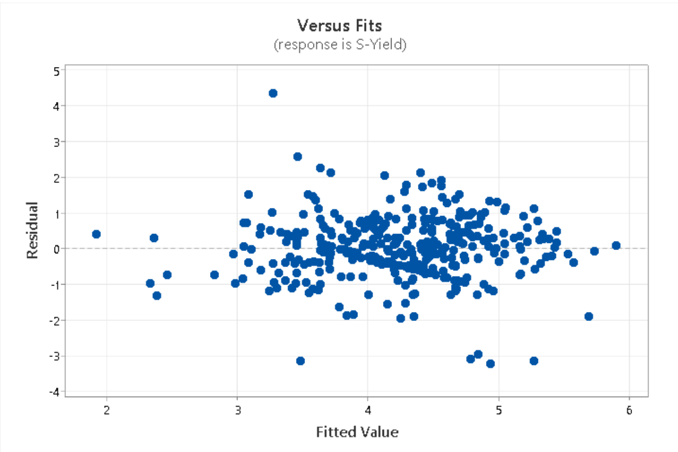
Coefficients

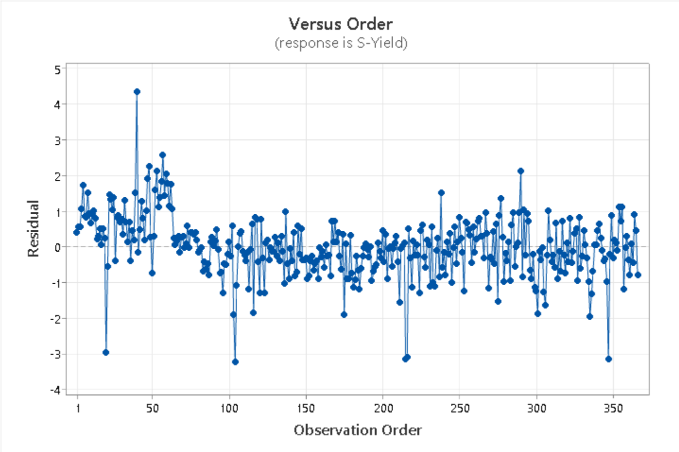
Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	1.437	0.208	(1.027, 1.847)	6.90	0.000	
RH Avg	0.07035	0.00512	(0.06029, 0.08041)	13.75	0.000	1.00

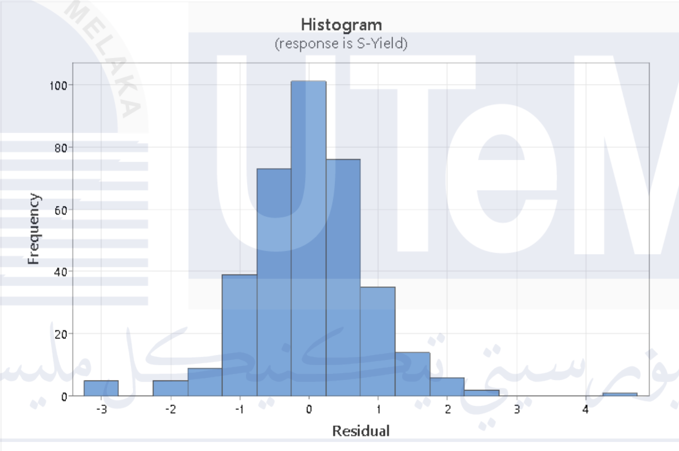
Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.867321	34.20%	34.02%	277.077	33.41%	938.53	950.17






Versus Order
(response is S-Yield)
Residual
Observation Order
5
4
3
2
1
0
-1
-2
-3
-4
1
50
100
150
200
250
300
350


Histogram
(response is S-Yield)
Frequency
Residual
100
80
60
40
20
0
-3
-2
-1
0
1
2
3
4

APPENDICES B

MINITAB OUTPUT FOR M6 MODEL

Regression Equation

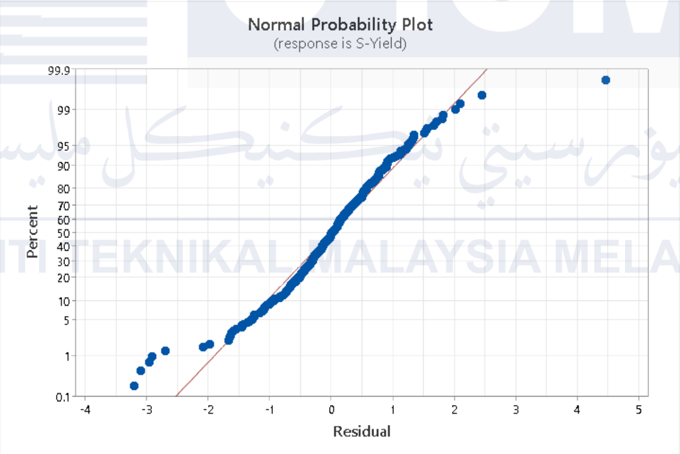
S-Yield = 0.290 + 0.05588 RH Avg + 0.001679 Tilt Irrad

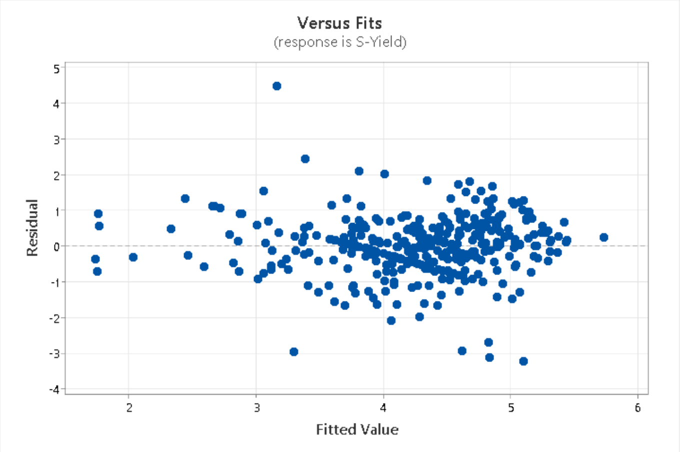
Coefficients

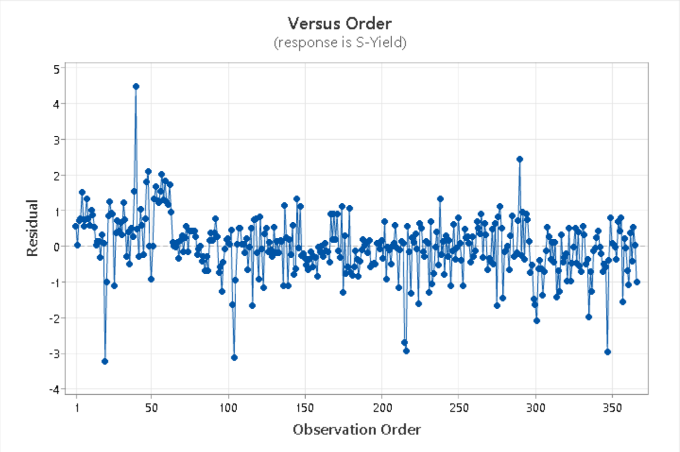
Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0.290	0.266	(-0.234, 0.814)	1.09	0.277	
RH Avg	0.05588	0.00535	(0.04536, 0.06640)	10.45	0.000	1.22
Tilt Irrad	0.001679	0.000261	(0.001166, 0.002193)	6.43	0.000	1.22

Model Summary

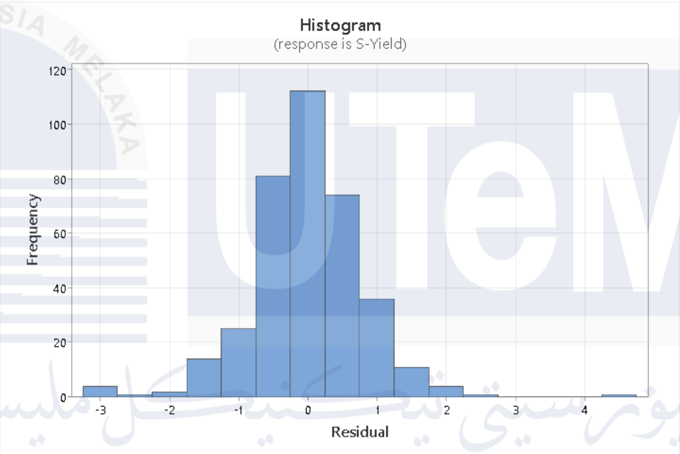
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.822900	40.93%	40.60%	250.359	39.84%	901.08	916.58







Versus Order
(response is S-Yield)
Residual
Observation Order



Histogram
(response is S-Yield)
Frequency
Residual