Open Access

Journal of Energy and Power Technology

LIDSEN Publishing Inc.

Original Research

Quantification and Comparative Analysis of Environmental Factors to Large-Scale Solar Plant's Energy Performance via Regression and Linear Correlation Models

Jazael Ballina, Yun Ii Go *

School of Engineering and Physical Sciences, Heriot-Watt University Malaysia, No 1 Jalan Venna P5/2, Precinct 5, 62200 Putrajaya, Malaysia; E-Mails: jb2019@hw.ac.uk; y.go@hw.ac.uk

* **Correspondence:** Yun Ii Go; E-Mail: y.go@hw.ac.uk

Academic Editor: Mohammad Jafari

Collection: Optimal Energy Management and Control of Renewable Energy Systems

Journal of Energy and Power Technology
2025, volume 7, issue 1
doi:10.21926/jept.2501001

Received: September 15, 2024
Accepted: December 09, 2024
Published: January 02, 2025

Abstract

Performance degradation, including system deterioration, corrosion, and energy loss in solar PV systems, can be caused by environmental conditions such as high humidity, frequent rainfall, and temperature swings in tropical nations. Over the years, photovoltaic energy has been successfully developed to have low production costs and high efficiency. It has been the main reason that solar energy is one of the fastest-growing renewable energies in the world, with a generation of 821 TWh of electricity since 2021, representing 23% incremental from the previous record. Even though photovoltaic technology is the preferred renewable energy due to the abundant availability of solar energy, one of its challenges is the significant reduction in the performance, power output, and efficiency caused by its installation directly into the open atmosphere and to the environmental phenomena translating into potentially avoidable economic losses. Various factors related to energy degradation are related to ambient temperature, dust particles, water droplets, shading, wind speed, humidity, and other climate parameters, particularly in tropical climate conditions. It quantifies the relationship of the least and most significant measurements using various techniques. These include computation of correlation coefficients, linear regression methods, and statistical

evaluations. This study applies statistical methods such as the Pearson, Kendall, and Spearman correlation coefficients, ARIMA forecasting models, and regression analysis to assess how tropical environmental factors affect solar energy yield. Historical energy data from a large-scale solar plant located in Malaysia is adopted. The selected performance parameters are energy yield and plane of irradiance energy generated by the photovoltaic panels. The results found that humidity and temperature impact PV system performance the most. This work is significant, especially in countries with tropical climates. It provides a reference model with similar weather and energy data that plans to implement large-scale solar power projects as a Nationally Determined Contribution (NDC). Based on the research's findings, mitigating strategies like regular panel cleaning, enhanced monitoring systems, and predictive maintenance plans are recommended to reduce the effects of these climate-induced losses. This work aligns with UN-SDG on clean energy and contributes towards the national net-zero target in meeting the global energy challenge 2030. The beneficiaries include the National Energy Commission, regulators, urban planners, solar plant owners' national utility providers, etc. This paper contributes to improving the operational effectiveness of solar PV systems in tropical countries, supporting Malaysia's efforts to attain its energy transition goals.

Keywords
Environmental; correlations; coefficients; statistics

1. Introduction

According to the International Energy Agency (IEA), solar PV is the second renewable energy that grew from the other technologies in the world by 2020, with a generation of 821 TWh that represents 23% incremental from the previous record. Solar PV technology has emerged as a leading renewable energy source, driven mainly by falling production costs and improvements in panel efficiency. These low-carbon technologies could provide more than 30% of the global energy supply by 2040 [1]. Malaysia is the second-world manufacturer of PV, according to the data from the Malaysian Investment Development Authority (MIDA). It has 250 companies dedicated to the photovoltaic energy market. Due to Malaysia's equatorial location, the country has significant solar energy potential due to the abundant solar irradiance it receives throughout the year. However, its tropical climate brings operational challenges that can affect the performance of PV systems. The government focuses on increasing the contribution with a total of 1350 MW from the large-scale plants commissioned from the LSS program between 2022 and 2023. Malaysia has an objective to install 9 GW of solar energy by 2050 under the task force of the Ministry of Energy and Natural Resources in the country [2].

One of the challenges of solar photovoltaic generation is to avoid the degradation of solar energy due to its installation directly into the open atmosphere and to the environmental phenomena, which significantly reduces performance, power output, and efficiency, translating into potentially avoidable economic losses worldwide. Many factors related to the degradation of energy, combined with ambient temperature, dust particles, water droplets, shading, wind speed, humidity, and other climate parameters, decrease the performance of the panels and affect the energy systems,

particularly in tropical climate conditions. Thus, the objective of this study will cover the analysis of correlation coefficients, linear regression methods, and statistical evaluations, comparing environmental parameters against and the energy yield, performance ratio, and plane of irradiance energy generated by the photovoltaic panels. The results indicated by the coefficients, linear regression, and statistics show the relationship between the environmental parameters and the energy yield, plane of irradiance, and performance ratio, which allows knowing the intensity and direction of the relationship between them by interpreting the positive or negative coefficients obtained.

Numerous studies worldwide evaluating the environmental factors affecting the performance of the PV showed energy losses of 21.4% to 37.5% due to lack of rain by days or months. Several ecological parameters like wind velocity, wind direction, ambient temperature, water drops, and dust particles from different seasonal areas influence the power output day and night. On the other hand, a lower humidity effect between 69 and 75% increases the power output from the PV panels [3]. Many of these studies showed some equations, algorithms, correlation coefficients, and statistics resulting in higher or lower correlation, good or bad relationships between parameters, or simply matching some variables [4]. This study investigates the impact of these environmental factors on PV performance by quantifying their relationships using statistical models and ARIMA forecasting.

1.1 Problem Statement and Aim

In this study, a large-scale solar photovoltaic plant situated in a tropical rainforest climate, with a capacity generation of 49 MW_{ac}, has recorded daily environmental parameters measured by sensors, and it needs to be compared with the energy yield of the plant to detect the key factors that affect the system output. All these factors influence the performance of the photovoltaic panel, in some cases reducing the overall efficiency of the system and generating energy losses [5]. Another consequence would be the damage or deterioration that these factors cause to the photovoltaic panel, accumulation of soiling, dust composition, particle size, etc, affecting the module performance [6]. An analysis of the effect of correlation coefficients, linear regression methods, and statistical evaluations will be applied to determine the most impactful parameters that will help to find the linear relationship, system performance, positive or negative effects on the output, direct proportionality relation, and the evaluation of future prediction methods.

This work aims to compare the data provided, including data acquired for two years measuring the power generation and environmental parameters measured by sensors utilizing correlation coefficients, linear regression methods, and statistical analysis to select the most impactful environmental parameter affecting the performance of the power generation. This work also investigates and compares key environmental parameters to determine which one has the most and least significant impact on the energy produced by a solar panel.

A systematic data storage system has been created to deposit historical energy data from large-scale solar plants. A comparative study was carried out to evaluate the effect of the environmental parameters on the energy produced by the plant. This includes different statistical correlations: Pearson, Kendall, and Spearman, and linear regressions to determine the trends and relationships. The analysis will indicate which relationship will impact the solar performance of the photovoltaic array system. This will be followed by the statistical calculation of median, mode, median, standard

deviation, skewness, and kurtosis for data evaluation. A prediction method termed autoregressive integrated dynamic models (ARIMA) based on seasonal data selection is adopted to analyze the behavior using the environmental parameters. The results can be used to estimate power generation by understanding the effect of panel cleaning and soiling in the large-scale photovoltaic power plant. This contributes to an enhanced solar plant's control and management.

1.2 Environmental Factors

The use of photovoltaic energy is increasing day by day. However, various technical and environmental problems make it challenging to obtain optimal power from photovoltaic panels [7]. Humidity is one of the environmental factors that degrades solar cell operations. The moisture could potentially trigger rust and corrosion on electrical connections. Humidity also deterioarates soiling effect to the modules due to dust accumulations. This could be difficult to be removed even in rains and can lead to heterogeneous soiling, especially in modules with low inclination angles. Humidity can affect the electrical connections causing oxidation or corrosion. In the meantime, humidity could degrade the energy generation of the solar panel due to the reflection or diffraction of the sunrays by the water molecules. Rain is another factor that reduces the amount of solar radiation that reaches the surface of the photovoltaic panel due to cloudiness. On the other side, percipitation remove and clean the dirt deposited on the panel. According to [8], soiling is influenced by many factors including wind speed and direction, humidity, panel angle of inclination and rainfall frequency in a complex manner. Wind can affect the photovoltaic modules continuously, giving rise to a static and dynamic load. It presents dynamic load characteristics to what has been standardized. Wind is one of the complex factors due to its interaction with other factors. Two parameters are involved, they are the wind speed and wind direction. According to [9], the effect of sand and dust can be minimized by refining the anti-reflection coating on glass-encapsulated modules. This exhibits minimal impact on the open-circuit voltage and fill factor of the solar panel. The experiments show the presence of sand and dust caused a reduction in current, leading to lower power output. This is due to the etched surface of the solar panel that degrades the ability to capture solar energy. In this context, the internal structure of the module remained unaffected.

Irradiance is the magnitude that describes the incident solar radiation per unit area and is measured in W/m² [10]. It is defined as the incident radiant power per unit area for any wavelength of the electromagnetic spectrum. It is the magnitude used to quantify the solar radiation that reaches the earth's surface. In units of the international system, it is measured in W/m^2. It is defined as the value of the average energy intensity of an electromagnetic wave at a given point and is calculated as the average value of the Poynting vector. In turn, the module of this vector represents the instantaneous intensity of electromagnetic energy that flows through a unit area perpendicular to the direction of propagation of the electromagnetic wave and whose direction is that of propagation.

$$\boldsymbol{I = \frac{P_{inc}}{A_s}} \qquad \text{(Equation 1)}$$

Where:

- I = Irradiance (Wm^{-2})
- P_{inc} = Power (Watts)

- A_s = Area (m^2)

Irradiance varies throughout the day depending on geographic location and local climate. The instrument used for its measurement is the pyranometer, which is based on exposing a metal sheet with a reflective surface to radiation and next to it, another whose surface is absorbent. The absorbent sheet will be hotter than the reflecting sheet and admitting that the temperature difference is proportional to the radiation received, by measuring the thermal jump the irradiance is therefore determined.

Direct solar irradiance occurs when solar radiation reaches a specific surface of the Earth in the same direction as the sun without changing direction. If the plane is perpendicular to the straight line projected from the solar disk to the Earth, it is called direct normal. It is expressed in W/m^2. Diffuse solar irradiance occurs when solar radiation does not reach the surface in the same direction as the sun due to the atmosphere's molecular dispersion of electromagnetic radiation. It is anisotropic radiation whose value depends on the celestial area of origin. That is, the properties of the atmosphere vary randomly with time. The diffuse component can range from 20% of the global on a clear day to 100% on a cloudy day. It is expressed in W/m^2. The sum of the direct and diffuse components of the radiation is defined as the global reaction.

The reflected solar irradiance is the designated albedo radiation. The amount of radiation depends on the reflection coefficient of the surface, which is called the albedo. It usually supposes a very small contribution, and in some cases, it can be disregarded. Therefore, it is evident that the solar radiation that reaches the Earth's surface will have a different spectral distribution than that existing outside the atmosphere due to absorption and reflection and other factors (altitude of the place, geographic area, etc.) Atmospheric absorption is greater at certain wavelengths. This is known as selective absorption which is a phenomenon influenced by atmospheric factors. For example, very short-wavelength radiation such as ultraviolet (UV) light is absorbed by ozone. While radiation in the infrared (IR) range is attenuated by the presence of water vapor, carbon dioxide, and other gases and particles in the atmosphere.

1.3 ARIMA Models

The ARIMA family of models [11], also called Box-Jenkins models, is an essential forecasting tool and the basis for many fundamental ideas in the time series analysis. Autocovariance and autocorrelation function are the most critical parameters when selecting ARIMA models. The abbreviation ARIMA means Autoregressive Integrated Moving Average Method. The word integrated confuses, but it refers to the differentiation of the data series. A model can be autoregressive if the variable of a period “t” is explained by the observations of itself correlating to previous periods, adding an error term. Autoregressive models are abbreviated with the word AR. The order of the model expresses the number of past observations of the analyzed time series. The three numbers after ARIMA refer to the order of the AR or autoregressive process, the order of differentiation, and the order of the MA or moving average process. A moving average model explains the value of a particular variable in a period “t” based on an independent term and a succession of errors. The order of the model expresses the number of pass observations of the analyzed time series. The three numbers written after ARIMA refer to the order of the AR or autoregressive process, the order of differentiation, and the order of the MA or moving average process. A so-called moving average model explains the value of a particular variable in a period t

as a function of an independent term and a succession of errors corresponding to previous periods, appropriately weighted. The combination of AR, MA, and ARMA models is a special case of ARIMA models.

2. Literature Review

Tripathi, Ray, Aruna, and Prasad [12] studied the effects of humidity on the photovoltaic panel. The experimental arrangement consisted of a mobile horizontal frame support installation, one meter high, to support the panel; above the horizontal frame, a shadow was created to generate any desirable range of solar irradiance within the laboratory. During the evaluation period, the humidity level inside the laboratory was changed using a humidifier, just a water spray. The results are increased humidity levels, solar insolation, and decreased panel power. With an increase of 50.1% from the humidity level, the panel's power output was reduced by 34.2%. In addition, it was found that due to the increase in humidity from 65.4% to 98.2%, the panel temperature was reduced by 11.4%.

The electrical performance of photovoltaic panels is primordial. It depends on two critical parameters: the solar radiation that reacts on the surface of the photovoltaic panel and the temperature of the surface of the photovoltaic panel. The performance of the photovoltaic solar panel is legitimately dependent on the solar radiation on its surface. As the solar radiation in the area increases, the performance of the photovoltaic panel increases. Although the panel is packaged for protection and defensive back isolation to retain heat, the strength of solar radiation entering the panel area is influenced by different natural phenomena, such as the accumulation of dust particles on the surface of the photovoltaic panel, shadow effect on the surface of the PV panel, the proximity of pollutant areas, water drops or vapor, and other climatic components [12].

The temperature of the photovoltaic panel hurts its operation. Significant reductions in photovoltaic panel output have been identified, like the open-circuit voltage, fill factor, and maximum power point, which have been reported with the incremental of the panel temperature, which are measurement parameters of the solar panels [13, 14]. In addition, some studies have reported a slight increase in the short-circuit current of the panel due to the increase in the temperature of the PV panel. Therefore, increasing the temperature of the photovoltaic panel causes a reduction in its general efficiency. Previous studies showed a decrease in the open-circuit voltage of the PV panel with a rate of 0.4%/K due to the increase in panel temperature. Therefore, similarly, the maximum power point and the fill factor decrease at a constant rate of 0.6%/K and 0.2%/K, with increasing PV panel temperature.

In another research carried out by Bahaidarah et al. [15], the results showed a 9% increase in the efficiency of the PV panel when the surface was cooled down and the PV panel temperature was reduced by 20%. In the same way, solar irradiation and the PV panel temperature are greatly affected by environmental parameters, like dust particles, ambient temperature, humidity, wind speed, and ambient temperature. Existing literature reviews and reports that the temperature at the PV panel increases due to the deposition of dust and shadow on the surface of the PV panel. Research related to the performance of photovoltaic panels has been carried out, where it is reported that the performance of photovoltaic panels degrades up to 28.7% due to the 42.1% increase in relative humidity. While in humidity areas, it reduced efficacy to 32.4% when the humidity level was raised to 6% and the PV panel was functioning at 58°C. It was observed that the

interaction between temperature and humidity is negatively correlated. The higher the humidity and the temperature, the efficiency of the equipment decreases, but in a small proportion. This is possible, given the climatic conditions observed in the field experiment, as the temperature is inversely proportional to relative humidity in most cases.

In regions with higher latitudes where the weather is icy, the overall efficiency of solar panels is reduced due to the accumulation of ice and snow on top of the solar panels, such as the layer of ice and snow overloading the solar panel must be eliminated, either by removing snow or melting it, for this reason, a solution is the use of hot air ventilation systems when necessary, to maintain stable efficiencies [16]. In hot and dry desert climates, the efficiency decreases due to wind speed, low humidity, and high temperatures [17]. In the Mediterranean, where temperature and variable climates exist, the efficiency fluctuates depending on the time [18]. In the tropics, the efficiency is more stable and depends more on the generation of the panels [19]. There are differences in the efficiencies compared to urban climates concerning rural climates due to the emission of pollutants.

In the research paper [20], time series is used for forecasting traffic in communication networks, highlighting the use of the ARIMA model. It introduces the statistical models with time series, which allow estimating future forecasts of traffic in modern communication networks, making use of traffic predictability with short-range dependency to carry out more efficient and timely control in an integrated manner at different levels of the network's functional hierarchy. This time series modeling is based on measurements taken of events periodically. The objective of this study is to focus on the series and how the series can be a perfect tool to model the data traffic in networks. This is possible through the Box-Jenkins methodology presented in this article. At the end of this investigation, it was possible to model a WiMAX traffic series of 10 days through an ARIMA time series with a small error.

Ganti et al. [21] analyzed environmental impacts (dust accumulation, water droplets, and partial shade conditions) and improvement of factors affecting the energy utilization of solar PV in the mining industry. A hybrid method (gradient boosting decision tree and sparrow search algorithm) was proposed to improve solar PV system efficiency. The method was implemented in MATLAB/Simulink, and an evaluation of the performance was carried out by comparing current methods. The results demonstrated that the proposed control system was more efficient in tracking performance than existing models.

2.1 Research Gap

Researchers developed and applied new techniques for modeling time series, proving their effectiveness since they can deal with any data pattern and produce accurate forecasts based on the description of historical patterns in the data. The new proposals for time series analysis were initially applied to pollution problems, epidemiological diseases, and currently to physical and social phenomena. The methods developed were auto-regressive (AR), moving average (MA), integrated (I), and combinations between them (ARMA and ARIMA). The main difference between these models and the classic ones is the stochastic approach that is given to the time series instead of treating them deterministically. Under this approach, the time series is conceived as a set of random time values generated from an unknown process; in other words, the series is conceived as a stochastic process. This approach aims to try to identify the probabilistic model that represents the main characteristics of the behavior of the series. ARIMA models present important advantages

compared with the classic simulation. The main one is the degree of matching that most time series. Different from classical modeling, where a series is fitted to an already established mathematical model, ARIMA models are fitted to a particular series; in this research work, the environmental parameters with higher correlation were selected and simulated to create a forecast and proactive solutions for plan management, idle restoration prevention, and optimum operation.

2.2 Novelty

ARIMA models of classical time series processing methods use concepts for ARIMA modeling derived from theories of old probability and mathematical statistics. The novelty of this study will consist of utilizing different correlations focused on Pearson and Kendal, besides the linear regression methods and statistical properties, to calculate the coefficients that impact solar power generation from a large-scale photovoltaic plant. Another innovative method based on the stochastic process is to perform or develop stationarity autoregression integrated moving average models that have not been addressed in evaluating the environmental parameters that affect the power generation from a photovoltaic plan or system.

3. Methodology

The work plan to develop this study is presented in Figure 1. It includes five phases that cover data gathering, cleaning, and database creation, followed by the basic and advanced plotting with the variables selected by a systematic table to select the variables to be analyzed. The next phase consists of evaluating the environmental parameters against the power generation, plane of array, and irradiation energy, using coefficient correlations to study the relationship between quantitative variables, followed by linear regression methods and statistical analysis. This study will provide information about the intensity and direction of the relation or covariation between all the variables related linearly.

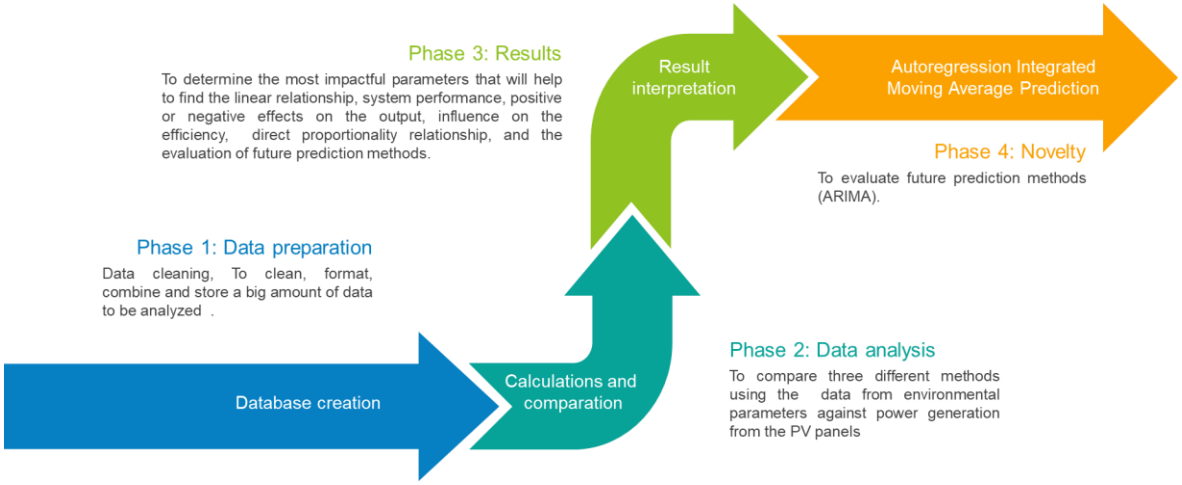


Figure 1 Methodological stages.

3.1 Phase 1: Data Preparation

Phase 1 data preparation focuses on cleaning, structuring, formatting, validation, and storage. All the data and the relevant information for analysis were stored using the software Microsoft

Access and a proper tool to create and display the information. The information selected is the key environmental parameters measured from the Collector Station (CS), Meteorological Monitoring Facility (MMF), SPP interconnection facility data from the Power Quality (PQ) Meter, and forecast weather, as presented in Table 1.

Table 1 Information selected for data preparation and analysis.

Source	Parameter	Data frequency & Observations
Meteorological Monitoring Facility (MMF)	Ambient temperature (deg C)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Average plane of array irradiance from all the MMFs (W/m^2)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Global horizontal irradiance (W/m^2)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Plane of array irradiance (W/m^2)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Plane of array irradiance totalizer (W/m^2)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Wind speed (m/s)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Wind direction (0-360 deg)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Relative Humidity (%)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Rain gauge (mm)	Daily data, 15 minutes interval data
Meteorological Monitoring Facility (MMF)	Atmospheric pressure (mbar)	Daily data, 15 minutes interval data
Power Quality (PQ) Meter	Active Power Reading (MW)	Daily data, 5 minutes interval data
Power Quality (PQ) Meter	Reactive Power Reading (MVAr)	Daily data, 5 minutes interval data
Power Quality (PQ) Meter	Apparent Power Reading (MVA)	Daily data, 5 minutes interval data
Power Quality (PQ) Meter	Active Energy Totalizer (kWh)	Daily data, 5 minutes interval data
Power Quality (PQ) Meter	Reactive Energy Totalizer (kVArh)	Daily data, 5 minutes interval data
Panel cleaning record	Cleaning PV panels by time	Monthly data, 1 string consists of 20 panels
Forecast Weather (data from webpage)	Power generation (Watts)	Daily data, 15 minutes interval data
Forecast Weather (data from webpage)	Global horizontal irradiance (W/m^2)	Daily data, 15 minutes interval data
Forecast Weather (data from webpage)	Temperature (deg C)	Daily data, 15 minutes interval data

After the data was selected and processed, the next step was to create the database via Microsoft Access to store information. This will be connected to the graphical visualizer and final database, which will include data from MMF, PQ meter, and forecast weather, as presented in Figure 2.

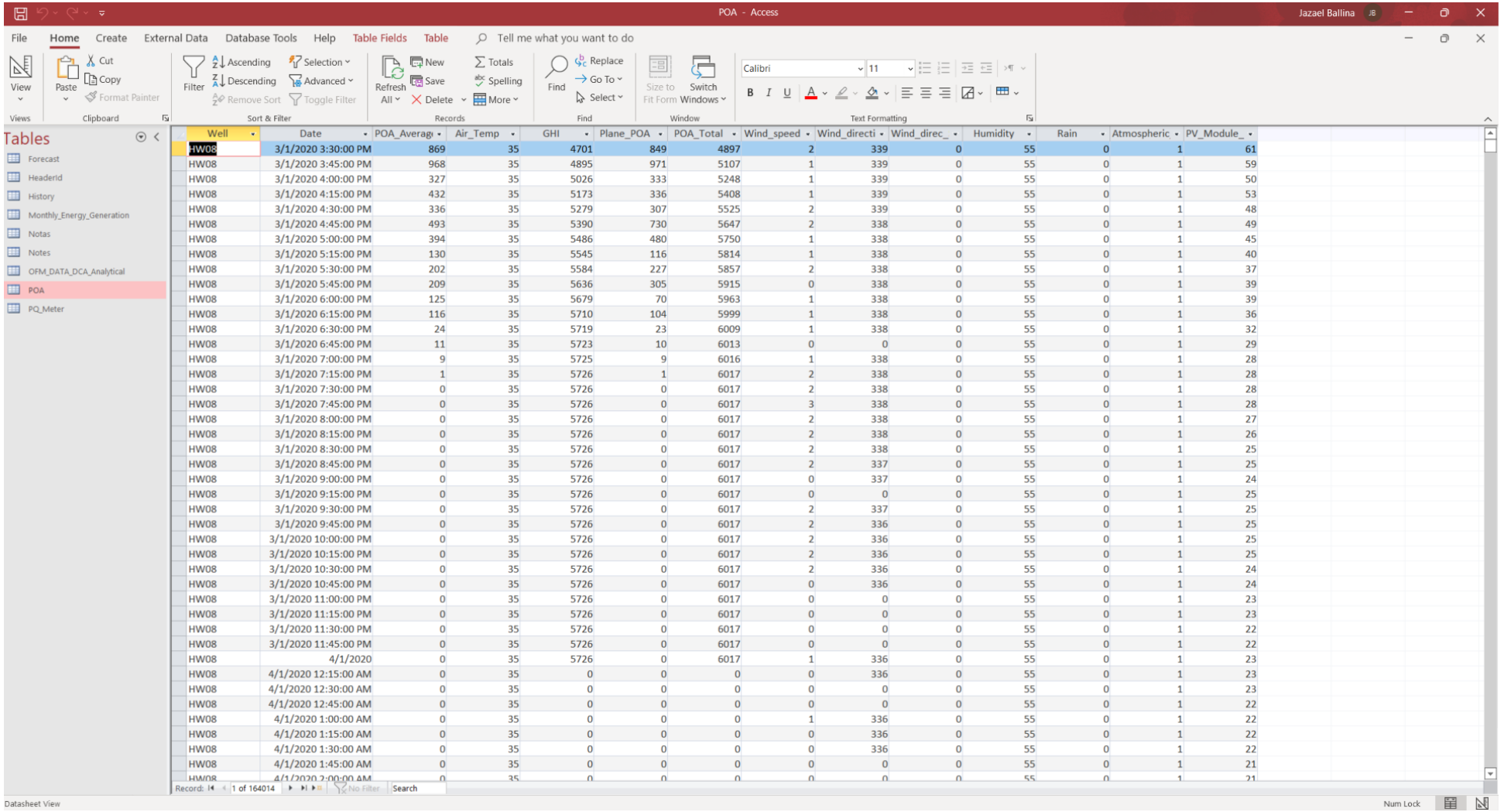


Figure 2 Database created with the data preparation process.

The connection between the application and Microsoft Access was created for the data visualization. The final plotting covered all environmental parameters and combined them with the power generation to create the scatter plots, as shown in Figure 3.

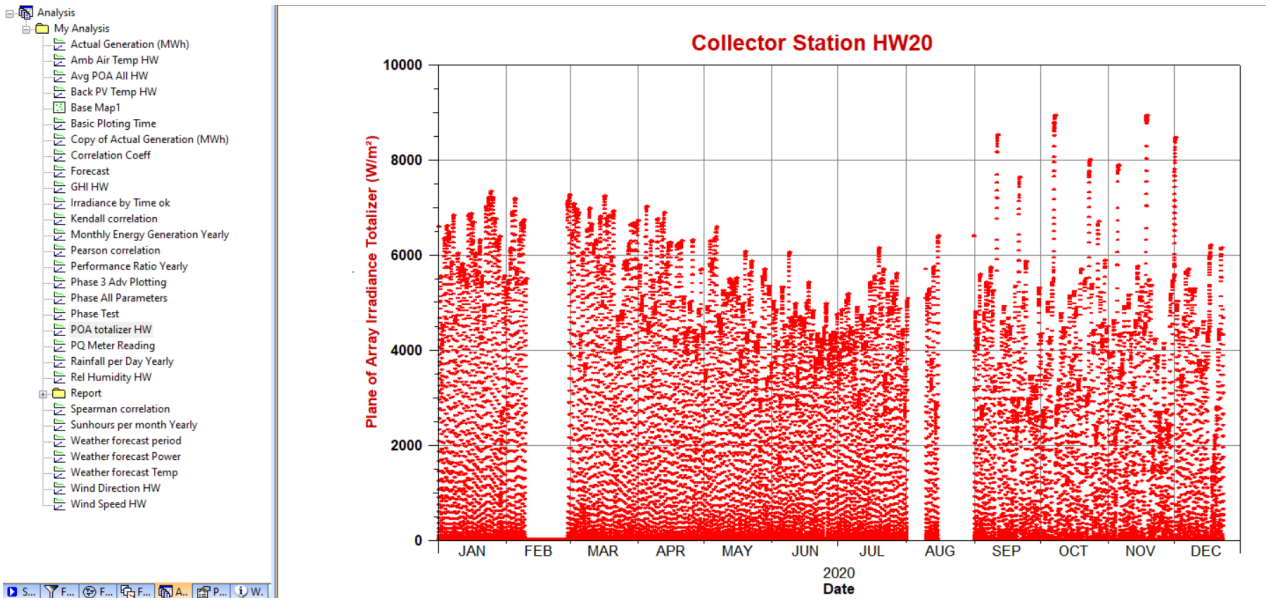


Figure 3 Data visualization stage.

3.2 Phase 2: Data Analysis

Statistical analysis is adopted via histograms to observe the data distribution. It measures how frequently each value appears in every interval within a set of values. It also observes the intervals of selected values that appeared more frequently within the data set. It validates the data plotted and calculates the selected variables to form the correlation matrix, as presented in Figure 4. It demonstrates the use of statistics and distribution to select the variables for data analysis. For this study, descriptive statistics are adopted, including quantitative (number), qualitative (categorical), discrete (exact values), ordinal (order), and variability (std deviation).

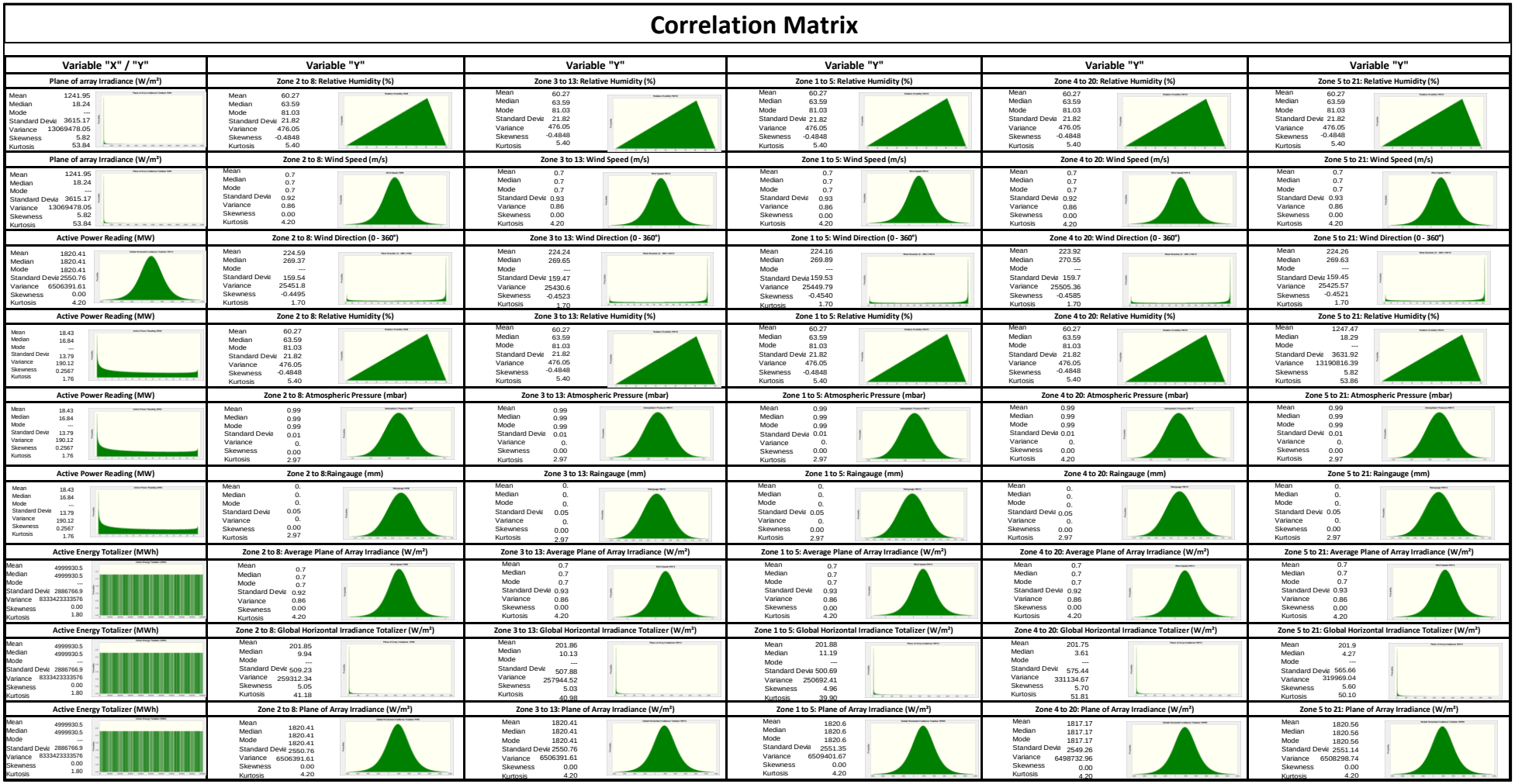


Figure 4 Data analysis and selection for evaluation.

The next step in the data analysis was to filter the data using Microsoft Excel. The first stage of screening was carried out via the correlation function to develop the correlation matrix. Variables that exhibit a certain level of correlation, which varied between -1 and +1, are selected and highlighted, as presented in Figure 5. The values shaded in blue and red are selected to compute their Pearson, Kendall, and Spearman correlations.

Data Series:	Active Power Reading MW	MVAr	Apparent Power Reading MVA	Active Energy Totalizer KWh	Reactive Energy Totalizer KVArh	Average Plane of Array Irradiance from 5 MMFs	Ambient Air Temperature	Global Horizontal Irradiance Totalizer	Plane of Array Irradiance	Plane of Array Irradiance Totalizer	Wind Speed	Wind Direction (0 - 360°)	Relative Humidity	Raingauge	Atmospheric Pressure	Back of Module Temperature
Distribution:	20.91	0.72	21.41	4999799.5	5000086.5	205.51	27.27	2233.75	209.09	2168.91	1.07	117.9	85.72	0.	0.98	31.47
Best Fit:	Weibull	Logistic	Beta	Discrete Uniform	Discrete Uniform	Gamma	Lognormal	Min Extreme	Gamma	Min Extreme	Beta	Logistic	Beta	Beta	Uniform	Lognormal
Rank Methods:																
Anderson-Darling	476.4564	296.1384	373.5551			1817.5160	390.8744	1591.6175	1810.5952	1551.7717	########	745.4294	278.5317	8644.7762	8611.7936	########
Chi-Square				3228.0181	170348.8306											
P-Value:	0.000	0.000	---	0.000	0.000	0.000	0.000	0.000	0.000	0.000	---	0.000	---	---	0.000	0.000

Correlations:	Active Power Reading MW	Reactive Power Reading MVAr	Apparent Power Reading MVA	Active Energy Totalizer KWh	Reactive Energy Totalizer KVArh	Average Plane of Array Irradiance from 5 MMFs	Ambient Air Temperature (deg C) HW8	Global Horizontal Irradiance Totalizer (W/m^2) HW8	Plane of Array Irradiance (W/m^2) HW8	Plane of Array Irradiance Totalizer (W/m^2) HW8	Wind Speed (m/s) HW8	Wind Direction (0 - 360°) HW8	Relative Humidity (%) HW8	Raingauge (mm) HW8	Atmospheric Pressure (mbar) HW8	Back of Module Temperature (deg C) HW8
Active Power Reading (MW)	1															
Reactive Power Reading (MVAr)	-0.6	1														
Apparent Power Reading (MVA)	1	-0.6	1													
Active Energy Totalizer (KWh)	-0.009	-0.040	-0.013	1												
Reactive Energy Totalizer (KVArh)	0.041	-0.112	0.054	0.095	1											
Average Plane of Array Irradiance from 5 MMFs	0.002	-0.013	0.000	-0.006	0.031	1										
Ambient Air Temperature (deg C) HW8	0.018	0.093	0.022	0.035	-0.042	0.654	1									
Global Horizontal Irradiance Totalizer (W/m^2) HW8	0.012	0.041	0.008	0.006	0.020	0.292	0.584	1								
Plane of Array Irradiance (W/m^2) HW8	0.002	-0.014	0.001	-0.004	0.030	0.992	0.655	0.291	1							
Plane of Array Irradiance Totalizer (W/m2) HW8	0.013	0.050	0.011	0.008	0.031	0.296	0.590	0.997	0.295	1						
Wind Speed (m/s) HW8	0.000	0.191	0.011	0.032	-0.025	0.404	0.564	0.446	0.403	0.452	1					
Wind Direction (0 - 360°) HW8	0.002	0.156	0.004	-0.004	-0.067	0.332	0.469	0.320	0.332	0.323	0.639	1				
Relative Humidity (%) HW8	-0.025	-0.097	-0.045	-0.048	-0.043	-0.679	-0.861	-0.527	-0.679	-0.537	-0.587	-0.429	1			
Raingauge (mm) HW8	-0.011	0.008	-0.014	0.006	-0.040	-0.055	-0.092	0.070	-0.054	0.066	0.058	0.015	0.119	1		
Atmospheric Pressure (mbar) HW8	-0.009	0.062	0.009	0.031	0.018	0.034	-0.004	-0.026	0.034	-0.022	0.036	0.023	-0.044	-0.017	1	
Back of Module Temperature (deg C) HW8	0.008	0.032	0.007	0.011	-0.027	0.827	0.916	0.520	0.829	0.524	0.508	0.431	-0.801	-0.074	0.012	1

Figure 5 First filtering for variable selection to calculate coefficient correlations.

4. Result and Discussion

The monthly actual generation was plotted as part of the analysis from the large-scale solar photovoltaic plant, as shown in Figure 6. The performance ratio or PR is a factor that determines the quality of a photovoltaic installation and is one of the most important magnitudes to be considered. The performance ratio expresses the relationship between the actual energy performance and the theoretically possible energy performance [13]. It indicates what proportion of the energy is available after having deducted the energy losses and the consumption. It is not possible to reach a real value of 100% because, during the operation of the photovoltaic installation, there are always unavoidable losses. For example, thermal losses due to heating of the modules, losses due to the Joule effect in the wiring, and losses due to soiling [22]. However, it is possible to reach a PR of up to 85% for inefficient photovoltaic installations. The performance of a photovoltaic installation can be evaluated via the performance ratio over time as shown in Figure 7.

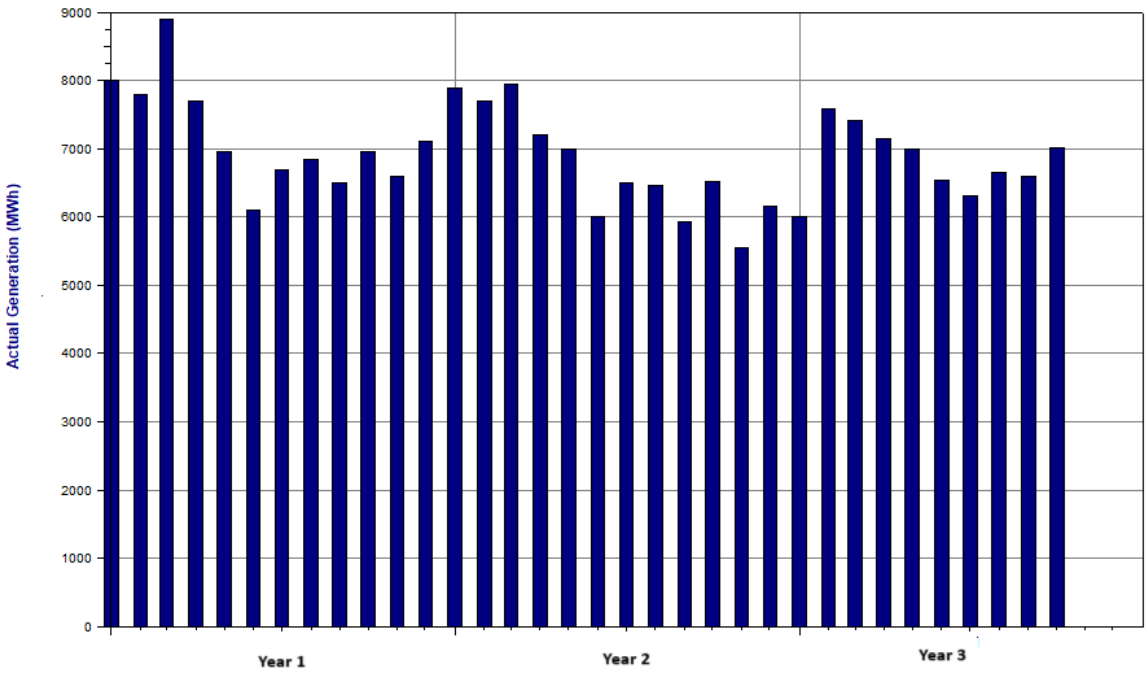


Figure 6 Monthly energy generated for three years.

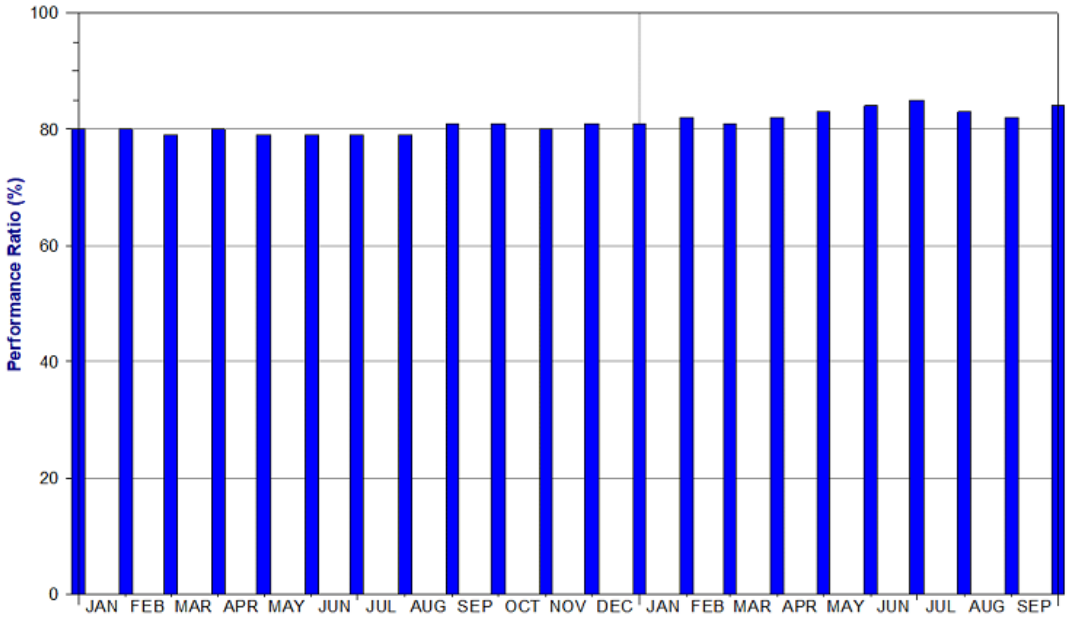


Figure 7 Performance ratio in each month.

The next step is the calculation of the correlation factors and plotting them in a scatter plot. The main objective of this calculation was to see the trends of data and the relationship between variables. If the correlation between two variables is close to zero, it indicates that these variables do not present any relationship between them. On the other hand, when the correlation increases and approaches one, the variables show a strong relationship between them. When the correlation between two variables is equal to one, the variables behave the same and follow the same trend.

The correlation results of the solar photovoltaic system are observed, it can be concluded that environmental parameters that have a direct relationship over the power generation are the ambient temperature and back module temperature in strong correlation values of Pearson correlation of 0.95, Kendall of 0.70 and Spearman of 0.84 and wind speed and wind direction have a moderate correlation with values of Pearson correlation of 0.47, Kendall of 0.45 and Spearman of 0.60, and the inverse relationship over the power generation is the humidity with strong negative correlation with values of Pearson correlation of 0.79, Kendall of 69 and Spearman of 0.71 as shown in Figure 8.

Correlation Matrix

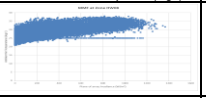
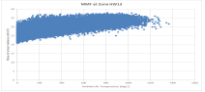
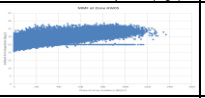
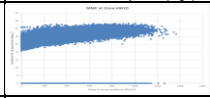
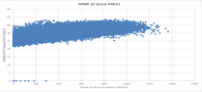
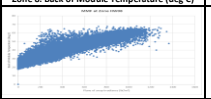
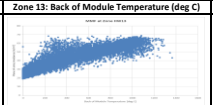
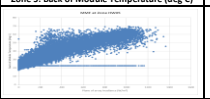
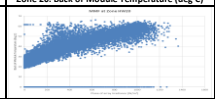
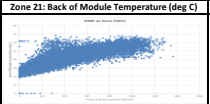
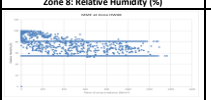
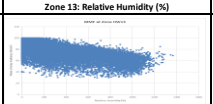
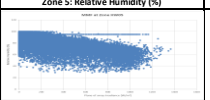
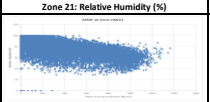
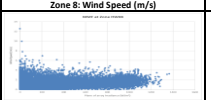
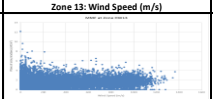
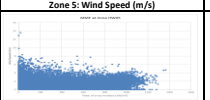
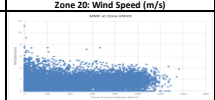
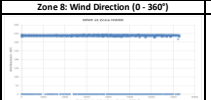
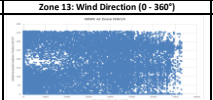
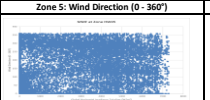
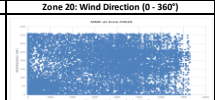
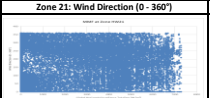
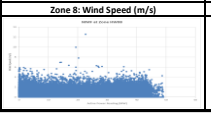
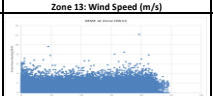
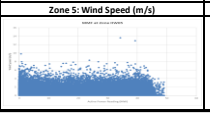
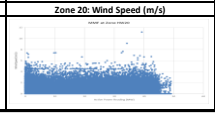
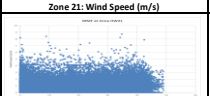
X/Y	Zone 8: Ambient Air Temperature (deg C)	Correlation Results	Zone 13: Ambient Air Temperature (deg C)	Correlation Results	Zone 5: Ambient Air Temperature (deg C)	Correlation Results	Zone 20: Ambient Air Temperature (deg C)	Correlation Results	Zone 21: Ambient Air Temperature (deg C)	Correlation Results
Plane of array Irradiance (W/m²)		Pearson (r): 0.7902 Kendall's (tau): 0.5037 Spearman rank: 0.6777		Pearson (r): 0.8034 Kendall's (tau): 0.5149 Spearman rank: 0.6923		Pearson (r): 0.7890 Kendall's (tau): 0.4999 Spearman rank: 0.6733		Pearson (r): 0.3695 Kendall's (tau): 0.4548 Spearman rank: 0.6086		Pearson (r): 0.7783 Kendall's (tau): 0.4928 Spearman rank: 0.6669
X/Y	Zone 8: Back of Module Temperature (deg C)	Correlation Results	Zone 13: Back of Module Temperature (deg C)	Correlation Results	Zone 5: Back of Module Temperature (deg C)	Correlation Results	Zone 20: Back of Module Temperature (deg C)	Correlation Results	Zone 21: Back of Module Temperature (deg C)	Correlation Results
Plane of array Irradiance (W/m²)		Pearson (r): 0.9516 Kendall's (tau): 0.7009 Spearman rank: 0.8446		Pearson (r): 0.9493 Kendall's (tau): 0.6939 Spearman rank: 0.8381		Pearson (r): 0.9332 Kendall's (tau): 0.6874 Spearman rank: 0.8278		Pearson (r): 0.7438 Kendall's (tau): 0.5981 Spearman rank: 0.7249		Pearson (r): 0.9319 Kendall's (tau): 0.6823 Spearman rank: 0.8289
X/Y	Zone 8: Relative Humidity (%)	Correlation Results	Zone 13: Relative Humidity (%)	Correlation Results	Zone 5: Relative Humidity (%)	Correlation Results	Zone 20: Relative Humidity (%)	Correlation Results	Zone 21: Relative Humidity (%)	Correlation Results
Plane of array Irradiance (W/m²)		Pearson (r): -0.0618 Kendall's (tau): -0.0502 Spearman rank: -0.0588		Pearson (r): -0.7906 Kendall's (tau): -0.5394 Spearman rank: -0.7123		Pearson (r): -0.7712 Kendall's (tau): -0.5189 Spearman rank: -0.6901		Pearson (r): -0.3683 Kendall's (tau): -0.5189 Spearman rank: -0.5615		Pearson (r): -0.7552 Kendall's (tau): -0.5166 Spearman rank: -0.6825
X/Y	Zone 8: Wind Speed (m/s)	Correlation Results	Zone 13: Wind Speed (m/s)	Correlation Results	Zone 5: Wind Speed (m/s)	Correlation Results	Zone 20: Wind Speed (m/s)	Correlation Results	Zone 21: Wind Speed (m/s)	Correlation Results
Plane of array Irradiance (W/m²)		Pearson (r): 0.4736 Kendall's (tau): 0.4570 Spearman rank: 0.6094		Pearson (r): 0.4108 Kendall's (tau): 0.3921 Spearman rank: 0.5312		Pearson (r): 0.3452 Kendall's (tau): 0.3474 Spearman rank: 0.4560		Pearson (r): 0.3589 Kendall's (tau): 0.3499 Spearman rank: 0.4529		Pearson (r): 0.3880 Kendall's (tau): 0.2990 Spearman rank: 0.4170
X/Y	Zone 8: Wind Direction (0 - 360°)	Correlation Results	Zone 13: Wind Direction (0 - 360°)	Correlation Results	Zone 5: Wind Direction (0 - 360°)	Correlation Results	Zone 20: Wind Direction (0 - 360°)	Correlation Results	Zone 21: Wind Direction (0 - 360°)	Correlation Results
Global Horizontal Irradiance Totalizer (W/m²)		Pearson (r): 0.2993 Kendall's (tau): 0.2484 Spearman rank: 0.3513		Pearson (r): 0.2000 Kendall's (tau): 0.1840 Spearman rank: 0.2549		Pearson (r): 0.2597 Kendall's (tau): 0.2376 Spearman rank: 0.3297		Pearson (r): 0.2192 Kendall's (tau): 0.2173 Spearman rank: 0.2902		Pearson (r): 0.2818 Kendall's (tau): 0.2160 Spearman rank: 0.3219
X/Y	Zone 8: Wind Speed (m/s)	Correlation Results	Zone 13: Wind Speed (m/s)	Correlation Results	Zone 5: Wind Speed (m/s)	Correlation Results	Zone 20: Wind Speed (m/s)	Correlation Results	Zone 21: Wind Speed (m/s)	Correlation Results
Active Power Reading (MW)		Pearson (r): 0.0030 Kendall's (tau): 0.0031 Spearman rank: 0.0045		Pearson (r): 0.0155 Kendall's (tau): 0.0096 Spearman rank: 0.0140		Pearson (r): 0.0155 Kendall's (tau): 0.0151 Spearman rank: 0.0212		Pearson (r): 0.0113 Kendall's (tau): 0.0066 Spearman rank: 0.0092		Pearson (r): 0.0023 Kendall's (tau): 0.0041 Spearman rank: 0.0058

Figure 8 Pearson, Kendall's, and Spearman's correlation results.

Although the correlations show how the variables are related, it is important to visualize this relationship or trend through graphs. Now the correlation coefficients have been carried out, and the statistical analysis of these data has been carried out, it is known what data should be used, the quantity of these, how they are distributed, how they are related to each other, etc. The following analysis in the methodology is to conduct the linear regression methods. Linear regression evaluation refers to finding a linear relationship between parameters or variables. The main idea of the algorithm is to get a line that best fits the data. The best-fit line is a line that matches most of the points from the data with the minimum error. The average distance of all the points to the line is the total error of the model. The results of the linear regression calculations are described in the correlation matrix below, where the results perfectly match those variables where the Pearson, Kendall's, and Spearman's coefficient correlations are positive or negative. Figure 9 shows the strong, medium, and weak coefficient correlations calculated using the Linear regression methods.

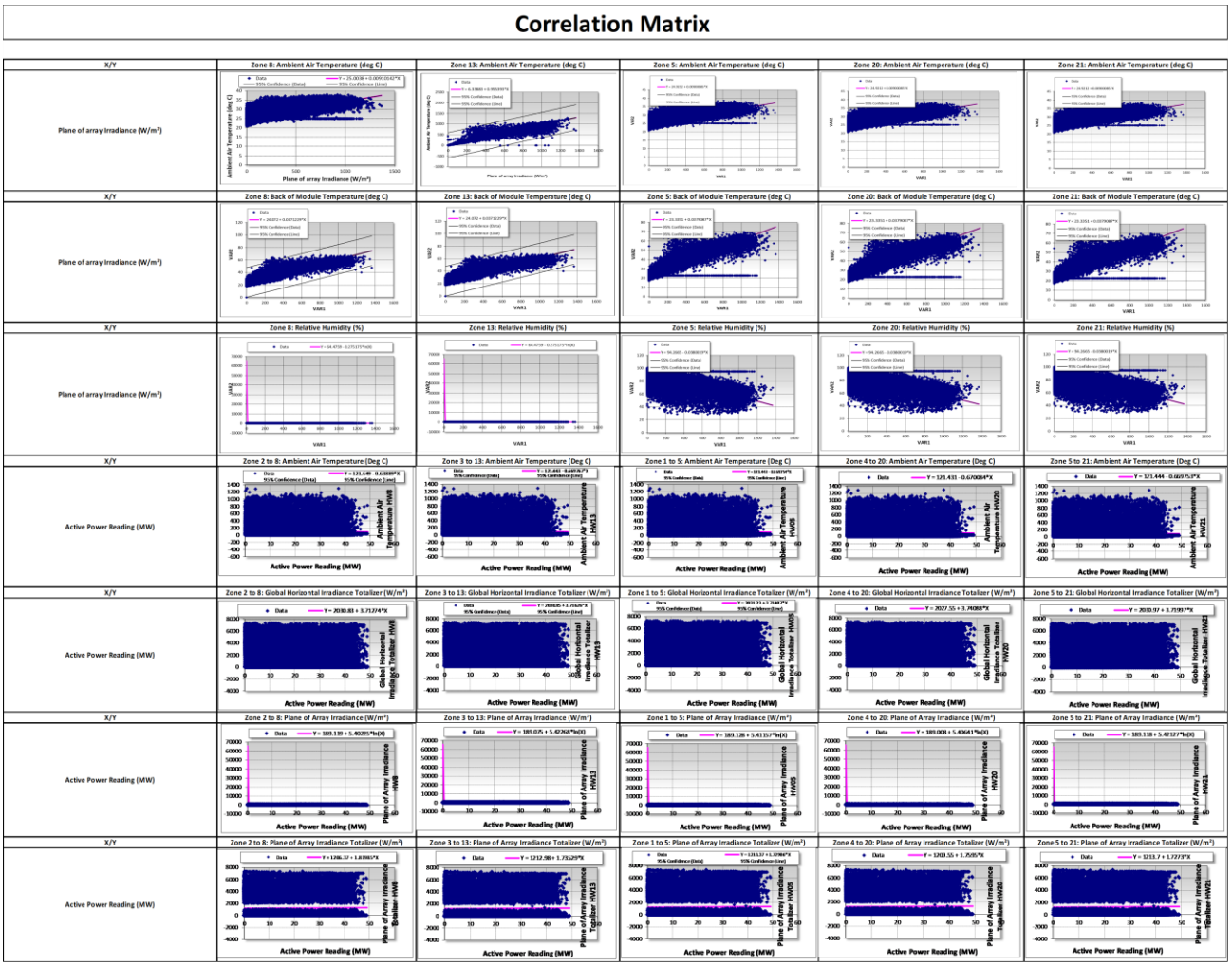


Figure 9 Linear regression results for the variables selected.

Figure 10 shows the actual generation power versus monthly sun hours and rainfall days per month. From observation, power generation is higher from January to March, whereas it is lower from June to August.

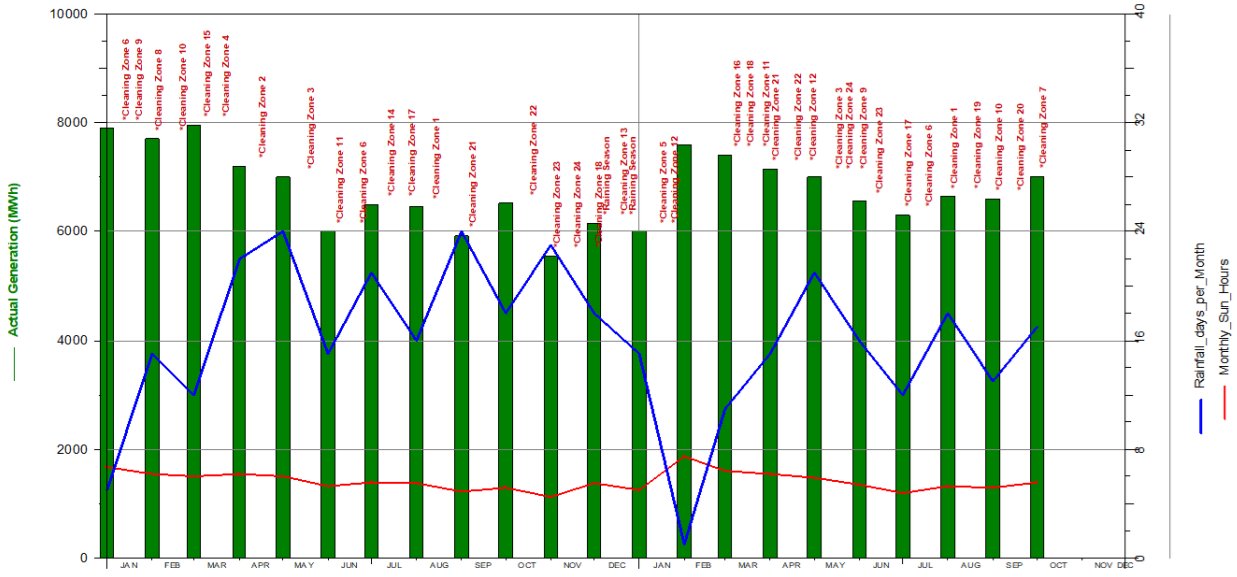


Figure 10 Actual energy generation versus monthly sun hours and rainfall days in each month.

Figure 11 shows the SARIMA prediction simulation for the back module temperature, observing a lower standard deviation compared with the mean, a seasonality autodetected method, a lower calculation of error of 2.5%, and the best-selected information criterion of Akaike with a value of 1.85.

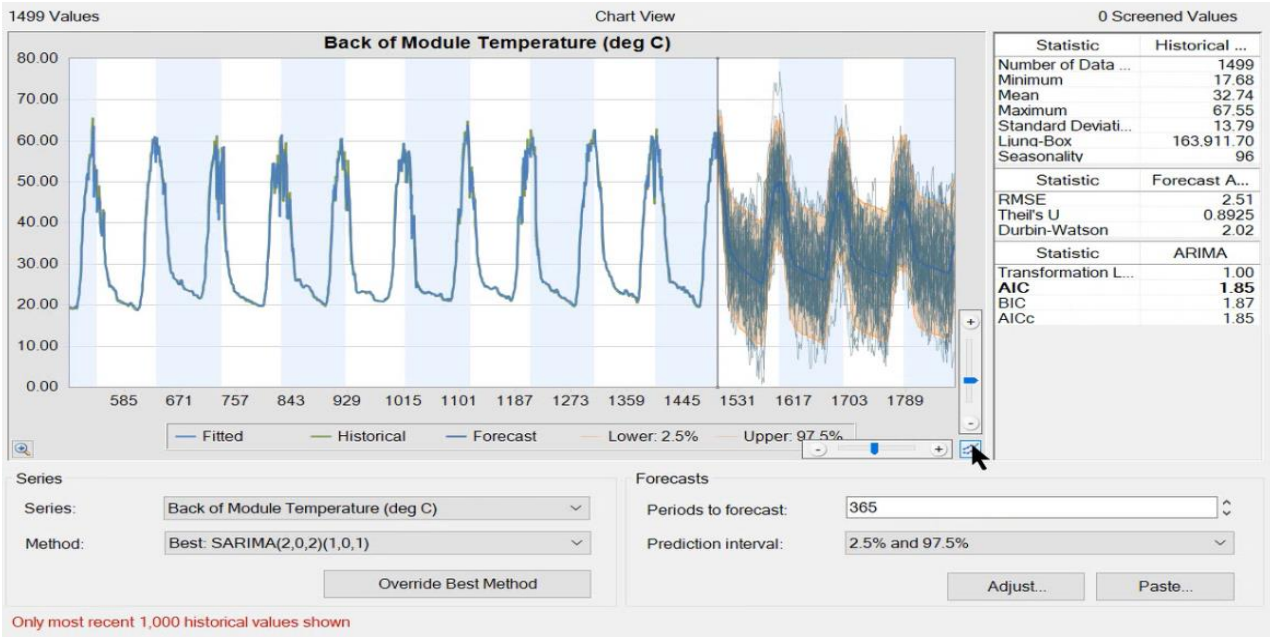


Figure 11 SARIMA model simulation for back module temperature.

5. Conclusion

Various research on optimal solar system sizing [23-24], solar plant-grid integration [25] and energy storage designs and safety for large scale solar [26] have been carried out to facilitate

national energy transition in meeting world energy challenges. This paper evaluated the environmental factors that impact solar PV performance in Malaysian weather conditions. This research provided quantification and comparative analysis of environmental factors to large-scale solar plants' energy performance via regression and linear correlation models. The environmental parameters that have a direct relationship over the power generation are the ambient temperature and back module temperature in strong correlation, and wind speed and wind direction have a moderate correlation, and the inverse relationship over the power generation is the humidity with a strong correlation and rain with moderate correlation. The findings from this paper demonstrated that humidity and ambient temperature were significant factors that affected the energy yield of solar PV systems, while moderate wind speeds aided in the dispersion of dust. Therefore, the large-scale solar PV plant must take the necessary steps in surveillance to monitor all these parameters to improve power generation in the future. Data analytics, average calculation, moving average methods, use of statistics, or the use of buckets to remove and clean unnecessary data, where the solution is to divide all the data into variables and classified by groups, each data group must have similar values than the other parameters with the help of programming language (python). To run prediction models based on regression methods, statistics, and correlations to have a proactive plan for cleaning the LSS PV plant. This can improve plant management, enhancement of power energy production, and detect idle panel restoration. It has also been observed that the places with the highest irradiation usually register more significant losses. ARIMA model forecasts support proactive maintenance, allowing plant operators to enhance performance under changing conditions. It is recommended to apply for future analysis SARIMA models to predict and forecast seasonal parameters, considering environmental parameters like temperature, wind speed, direction, humidity, and irradiance to predict soiling effects, shading, and degradation as a contingency proactive plan. Recommendations such as regular cleaning, real-time monitoring, and targeted maintenance strategies contribute to the country's renewable energy goals. Future research can be carried out using advanced forecasting methods and adaptable maintenance systems to further enhance the efficiency of solar PV systems applied to weather conditions from various geographical regions.

Author Contributions

Jazael Ballina, writing original manuscript, methodology, analysis, simulation; Yun Ii Go, review and supervision.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Bojek P. Solar PV – Analysis [Internet]. Paris, France: IEA; 2021. Available from: https://www.iea.org/reports/solar-pv.
2. IRENA. Solar energy [Internet]. Abu Dhabi: IRENA; 2019. Available from: https://www.irena.org/solar.

3. Yaghoubirad M, Azizi N, Ahmadi A, Zarei Z, Moosavian SF. Performance assessment of a solar PV module for different climate classifications based on energy, exergy, economic and environmental parameters. Energy Rep. 2022; 8: 15712-15728.
4. Javed W, Guo B, Figgis B. Modeling of photovoltaic soiling loss as a function of environmental variables. Sol Energy. 2017; 157: 397-407.
5. Costa SC, Kazmerski LL, Diniz AS. Impact of soiling on Si and CdTe PV modules: Case study in different Brazil climate zones. Energy Convers Manag X. 2021; 10: 100084.
6. Valerino M, Ratnaparkhi A, Ghoroi C, Bergin M. Seasonal photovoltaic soiling: Analysis of size and composition of deposited particulate matter. Sol Energy. 2021; 227: 44-55.
7. Syafiq A, Pandey AK, Adzman NN, Abd Rahim N. Advances in approaches and methods for self-cleaning of solar photovoltaic panels. Sol Energy. 2018; 162: 597-619.
8. Ballestrín J, Polo J, Martín-Chivelet N, Barbero J, Carra E, Alonso-Montesinos J, et al. Soiling forecasting of solar plants: A combined heuristic approach and autoregressive model. Energy. 2022; 239: 122442.
9. Shi C, Yu B, Liu D, Wu Y, Li P, Chen G, et al. Effect of high-velocity sand and dust on the performance of crystalline silicon photovoltaic modules. Sol Energy. 2020; 206: 390-395.
10. Zhu H, Blackborow P. Understanding Radiance (Brightness), Irradiance, and Radiant Flux [Internet]. Woburn, MA: Energetiq Technology, Inc.; 2011. Available from: https://www.energetiq.com/technote-understanding-radiance-brightness-irradiance-radiant-flux.
11. ORACLE. Oracle® Hyperion planning predictive planning in smart view user's guide [Internet]. Austin, TX: Oracle Corporation; 2019. Available from: https://docs.oracle.com/cd/E57185_01/CBPPU/PRHist_ARIMA_intro.htm#CBPPU-pp_user_book_217.
12. Tripathi AK, Ray S, Aruna M, Prasad S. Evaluation of solar PV panel performance under humid atmosphere. Mater Today Proc. 2021; 45: 5916-5920.
13. Kazem HA, Chaichan MT, Al-Waeli AH, Sopian K. Effect of dust and cleaning methods on mono and polycrystalline solar photovoltaic performance: An indoor experimental study. Sol Energy. 2022; 236: 626-643.
14. Laseinde OT, Ramere MD. Efficiency Improvement in polycrystalline solar panel using thermal control water spraying cooling. Procedia Comput Sci. 2021; 180: 239-248.
15. Bahaidarah H, Subhan A, Gandhidasan P, Rehman S. Performance evaluation of a PV (photovoltaic) module by back surface water cooling for hot climatic conditions. Energy. 2013; 59: 445-453.
16. ur Rehman H, Hirvonen J, Sirén K. A long-term performance analysis of three different configurations for community-sized solar heating systems in high latitudes. Renew Energy. 2017; 113: 479-493.
17. Ali AH, Zeid HA, AlFadhli HM. Energy performance, environmental impact, and cost assessments of a photovoltaic plant under Kuwait climate condition. Sustain Energy Technol Assess. 2017; 22: 25-33.
18. Malvoni M, De Giorgi MG, Congedo PM. Study of degradation of a grid connected photovoltaic system. Energy Procedia. 2017; 126: 644-650.

19. Ogbomo OO, Amalu EH, Ekere NN, Olagbegi PO. Effect of operating temperature on degradation of solder joints in crystalline silicon photovoltaic modules for improved reliability in hot climates. Sol Energy. 2018; 170: 682-693.
20. Ferro R, Hernández C, Puerta G. Rating Prediction in a Platform IPTV through an ARIMA Model. Int J Eng Technol. 2016; 7: 2018-2029.
21. Ganti PK, Naik H, Barada MK. Environmental impact analysis and enhancement of factors affecting the photovoltaic (PV) energy utilization in mining industry by sparrow search optimization based gradient boosting decision tree approach. Energy. 2022; 244: 122561.
22. Hussain N, Shahzad N, Yousaf T, Waqas A, Javed AH, Khan S, et al. Designing of homemade soiling station to explore soiling loss effects on PV modules. Sol Energy. 2021; 225: 624-633.
23. Thadani HL, Go YI. Large-scale Solar System Design, Optimal Sizing and Techno-Economic-Environmental Assessment. Sustain Energy Res. 2023; 10: 11.
24. Alias ND, Go YI. Decommissioning Platforms to Offshore Solar system: Road to Green Hydrogen Production from Seawater. Renew Energy Focus. 2023; 46: 136-155.
25. Mohanan M, Go YI. Optimized power system management scheme for LSS PV grid integration in Malaysia using reactive power compensation technique. Glob Chall. 2020; 4: 1900093.
26. Moa EH, Go YI. Large-scale Energy Storage System: Safety and Risk Assessment. Sustain Energy Res. 2023; 10: 13.