Impact of Temperature and Humidity on Photovoltaic System Energy Output in Northern Queensland

Wei, Quek (n10503196)

IFN704 Assessment 3, due 11:59pm Friday 30 October 2020

Executive Summary

Due to falling technology price and availability of government subsidy, the use of solar photovoltaic (PV) systems to generate electricity has become widespread among the households in Australia. However, its power output is often characterised by inconsistency and unreliability. Key environmental determinants that influence the power throughput of a solar photovoltaic system include geographic location, solar orientation, and weather condition, such as temperature and humidity, for the installation site. This study examined the effect of weather conditions, specifically temperature and relative humidity conditions, on the energy output of 16 residential solar PV systems installed in Northern Queensland. A linear regression model was fitted to assess the linearity of temperature, humidity and solar irradiance to PV energy produced. The results indicated that all three environmental variables explained 84.6% variability of PV energy produced during 2019. These findings could be used as a starting point of a more comprehensive PV performance modelling which would assist in refining the solar energy planning process that determines the suitability of solar PV installation sites in Northern Queensland.

Introduction

Queensland has an ambitious plan of meeting 50% of the state's energy needs through renewable resources by 2030 [1]. To achieve this target, one of the pathways is transitioning to harvesting solar energy through a distributed renewable solar energy power system. Given the falling technology cost, and no-interest loans, grants and rebates offered from both state and federal governments, grid-connected roof-top solar PV systems have been adopted increasingly across many households in Queensland. With over 580,000 solar systems installed and connected, there is more than 4,000 megawatts of solar power capacity - the largest power station in Queensland [2].

Contrary to common perception, solar PV systems are capable of generating energy not only in warm and sunny conditions but also through rain, clouds and even snow. The source of generating energy in solar PV systems is solar irradiation, not thermal heat. Nevertheless, its energy yield is highly influenced by the ambient weather conditions which it is operating in. Weather data, such as temperature, relative humidity, solar radiation intensity and wind speed, are commonly considered by engineers when estimating energy production of a particular solar PV installation site [3].

In this study, the power generation performance of solar PV systems installed on household roof-top were investigated under Northern Queensland meteorological conditions. By combining the datasets

obtained from Solar Analytics and Bureau of Meteorology, together with solar irradiance data calculated from PVLIB – a Python library from Sandia National Laboratory in US, the correlations between temperature, relative humidity and energy generated were analysed. This new knowledge could be incorporated into future planning and development of solar PV installation sites which would help to better answer questions such as "*To what extent weather conditions affect the power output of solar PV systems at any given site?"* and "*How can we predict solar power yields at a given location?"*.

Literature Review

Australia has some of the best solar resources in the world – approximately 58 million petajoules (PJ) of annual solar radiation, which is 10,000 times of Australia's annual energy consumption [4]. Queensland, being labelled as the "sunshine state" of Australia, is located within the high solar radiation area, as shown in Figure 1. Queensland is the leader in embracing solar energy. According to *Clean Energy Regulator* [5], more than 700,000 small scale solar energy generation units (<100kW) have been installed across the state as at 22 July 2020. As solar energy consumption is growing at a rate faster than other renewable energy sources in Queensland, an accurate and reliable assessment of solar energy yield from these small-scale generation units is becoming exigent because of its complex consequences on the planning and development of the overall electricity generation.

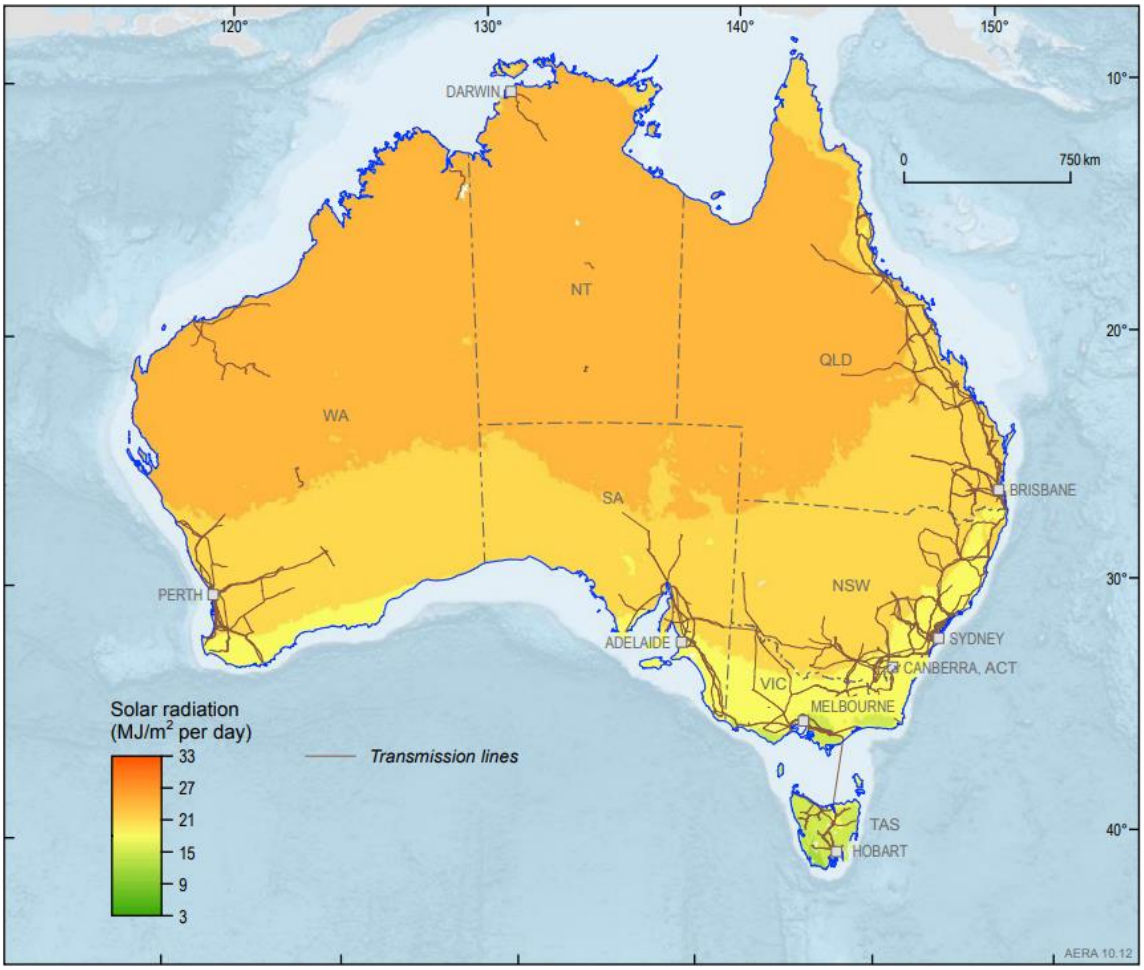


Figure 1. Australia annual average solar radiation (MJ/m2) [4]

Photovoltaic (PV) electricity generation is one of the commonly used technology to harvest energy from solar irradiation. Also known as solar cells [6], they are to be connected in series or parallel to form a PV array to generate electricity from solar irradiation. As solar PV systems are exposed to outdoor environment, the efficiency of energy conversion is therefore determined by several factors [3]. These includes light reflectivity, tilt angle, dust accumulation on the cells [7, 8] and cell temperature [3, 9, 10]. Most solar PV modules can only operate optimally under certain specific conditions known as standard test conditions (STC), which specified an operating temperature of 25^{o}C, relative humidity of 45%, solar radiation intensity of 1,000 W/m2 and air mass of 1.5 [11, p. 2]. However, these conditions basically do not exist in most part of the world.

Cell temperature is probably the most important factor to influence the efficiency of solar PV system to generate energy [3, 9, 10]. There are several research studies [6, 9, 10, 12, 13] themed at examining the influence of both temperature and relative humidity on solar cell temperature. Cell temperature increases means a reduction of the band gap of the semiconductor components containing in solar PV cells and a decrease in voltage [13]. All these research studies show that an elevated air temperature will result a higher PV cell temperature and thus a lower energy conversion rate. Humidity is another important determinant in controlling cell temperature. Whilst humidity acts as a cooling agent on the solar cell, it can impact the reception of solar radiation by presence of water vapour particles in the air. [6, 14].

Therefore, due to the sensitivity of operating conditions to solar cells, its power output estimation is often a complex problem. Rightfully as Kawajiri et. al. [12] pointed out, suitability of PV module type needs to be assessed for each region in consideration of its weather and seasoning variations throughout the year. Furthermore, given that most of the studies were conducted on either one or two PV installation sites or on a large-scale solar farm, there is a need to conduct further research to consider a broader sample size covering small-scale solar generation units. There is also a need to further evaluate the nature of the correlation of temperature, humidity and solar PV energy output in the context of Northern Queensland. This study will be valuable to provide solar power planner and engineers with an enhanced understanding of the impacts of temperature and humidity to energy yield and thus determine the optimal environment to solar generation in Northern Queensland.

Approach

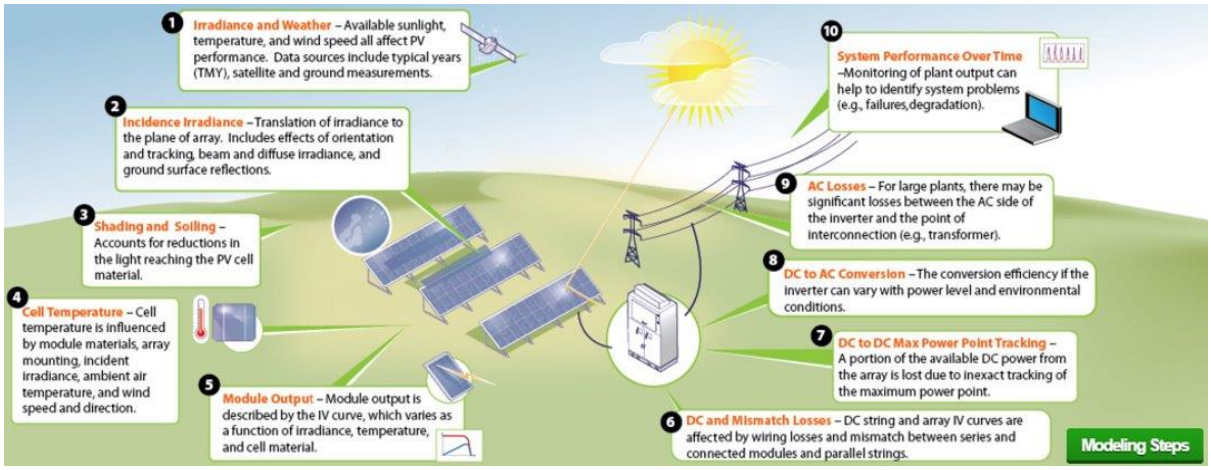


Figure 2. Research in PV Performance Modelling, source: https://pvpmc.sandia.gov/ [15]

Studying weather and irradiance data is the first step of setting up a PV performance model [15]. Past weather data are typically used to assess the design and suitability of a proposed systems of a given location. Figure 2 shows the modelling methodology used by U.S. Sandia National Laboratories, an industry and national laboratory collaborative to improve PV performance modelling. This study would form the basis of providing weather and atmospheric observations in step 1 in the modelling steps in this modelling methodology.

Data Gathering

Solar Analytics Data

The primary PV solar data were supplied by Solar Analytics [16]. The dataset contains residential solar power generation and voltage measurements from 1,000 residential Australian customer for period ranging from Jan 2019 to Dec 2019. Each site is provided with information of postcode and the state in which the site is located. For this study, the data were filtered to Northern Queensland installation sites only. Here, Northern Queensland is defined as the part of the Australian state of Queensland located north of Tropic of Capricorn, which is the southernmost latitude (23.436806°S) where the sun can be seen directly overhead.

One of the main characteristics of this Solar Analytics dataset is that it is in 5-minute sampling rate for each site, making it valuable to understand trends of each metrics in high frequency. In this study, we are particularly interested in the metric `'energy_(Wh)'`, which is the total PV energy produced during the 5 minutes in watt-hour. The data dictionary for this dataset is provided in Appendix A.

Weather Data

The weather data was sourced from Australian Government agency - Bureau of Meteorology (BOM). 227 Queensland installation sites were identified based on the installation sites' state provided in the site details file from Solar Analytics. These sites' postcodes are then search online to obtain its geo-coordinates, followed by entering these geo-coordinates into the BOM's climate data website to find the nearest weather-station. Five weather-stations located in Northern Queensland were identified, namely Cairns, Townsville, Proserpine, Mackay and St Lawrence. To match the 5-minute frequency in Solar Analytics dataset, a data service request was submitted through to the BOM

website, requesting temperature and relative humidity data in 1-minute interval for these 5 weather-stations. Subsequently, BOM provided an FTP link to download these requested data, which contain the maximum, minimum and mean readings of both temperature and relative humidity in 1-minute interval from Jan 2019 to Dec 2019.

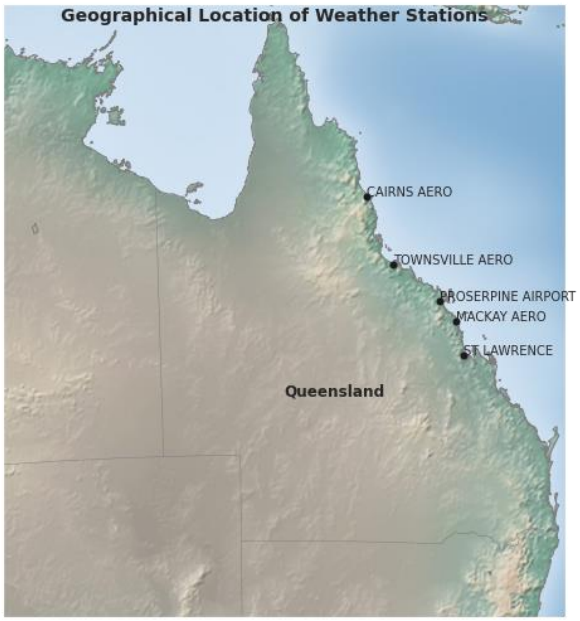


Figure 3. Geographical locations of weather-stations selected

Solar Irradiance Data

To analyse PV energy generated correctly, solar irradiance data must be sought. Australian Bureau of Meteorology provides solar data including global, diffuse, direct and terrestrial irradiance. However, it is only limited to certain weather-stations which none of them are in Northern Queensland. Cairns was one of these weather-stations, but the data was only available from 1997 to 2004. Daily solar exposure, which data is available on BOM's website for free, is not suitable as it is not in the same frequency as Solar Analytics data. Through a telephone conversation with BOM, they highlighted that solar exposure provided is only an estimate derived from satellite images. As for the case of global and diffuse solar irradiance, the readings are obtained through a ground-based instrument called "pyranometers". Pyranometers need to be calibrated regularly as they are not absolute devices. The pyranometer sensitivity may change over time due to its exposure to radiation, and the deterioration of the black paint on the device. Due to the unavailability of the required data and the potential inaccuracy readings from the instrument, the idea of getting solar irradiance data from BOM was abandoned.

PVLIB-Python modelling package [17] was found during a literature search on Google scholar. Initially developed in MATLAB and later ported to Python in 2015, it provides a set of functions for simulating the performance of PV systems. The reasons this package is selected as the data source for solar irradiance data in this study are: (1) it is a credible data source as the package was developed by Sandia National Laboratory which is supported by U.S. Department of Energy's Energy Efficiency and Renewable Energy; (2) it has been cited 87 times, according to Google Scholar [17]; and (3) it is developed and maintained on GitHub with contributors from national laboratories, private industries and academia. Through exploring the package, a specific function in the package

that returns several irradiance values for any given specific timestamp and geographic location is found to be useful for this study. The definitions of these irradiances can be found in Appendix C. Another interesting function provided in the package is that it is able to calculate plane of array (POA) irradiance for any given timestamp, geo-coordinates, panels tilt angle, site's solar orientation, shading etc., which could be helpful for extending this study if site information become available.

Data Preparation and Cleaning

All collected data are pre-processed before used. Datasets are checked for consistency, cleaned, and formatted appropriately. Firstly, by using the site information provided, Queensland's postcodes are selected and searched for its geographical location using Google Map. A list of 15 Northern Queensland postcodes is produced.

Downloaded Solar Analytics extracts are then loaded chronologically into a single dataframe as Solar Analytics only provides one extract per month. The dataframe is filtered to sites that reside in these 15 postcodes. 16 installation sites are found to have fulfilled this condition.

Next, weather data supplied by BOM were loaded. BOM supplied one extract for one weather-station. Because the weather data has a higher frequency than Solar Analytics data, each metrics are aggregated by calculating the mean temperature and humidity into 5-minute interval.

Lastly, to obtain solar irradiance data, the geo-coordinates of the five weather-stations were used to generate various irradiance data from the PVLIB function in a frequency of 1-minute, from Jan to Dec 2019. The solar irradiance data were summed and aggregated into 5-minute frequency to match the frequency of Solar Analytics data.

Data Completeness

To assess the quality of the weather data gathered and loaded, both columns `'Quality of air temperature'` and `'Quality of relative humidity'` are been examined. According to the data dictionary supplied by BOM, these columns indicate whether the specific measurements have been quality controlled or not. Figure 4 shows the number of no quality control reading counts for temperature and humidity by weather-stations.

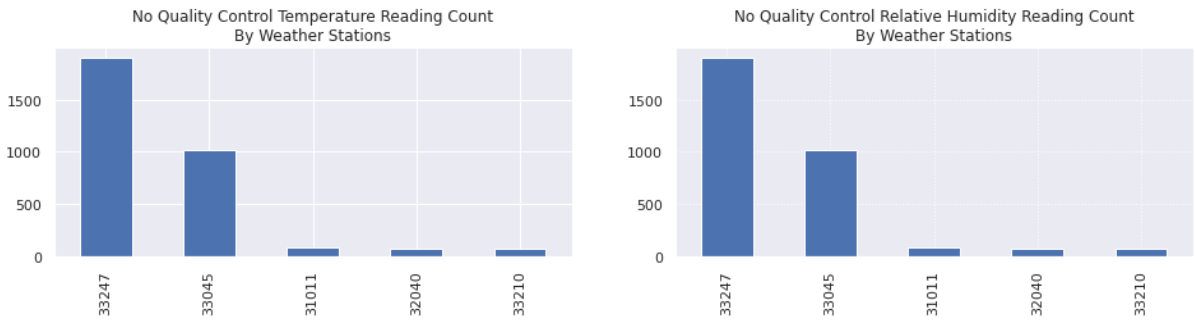


Figure 4. No quality control temperature and relative humidity reading count by weather stations

To illustrate some of these data quality issue, an example of these incomplete readings is selected here. Figure 5 shows a Cairns' weather-station record timestamped at 02/04/2019 16:04 which has no quality status with incomplete temperature and relative humidity readings.

	hd	Station Number	Year Month Day Hour Minutes in YYYY	MM	DD	HH24	MI format in Local standard time	Air Temperature in degrees Celsius	Quality of air temperature	Air temperature (1-minute maximum) in degrees Celsius	Quality of air temperature (1-minute maximum)	Air temperature (1-minute minimum) in degrees Celsius	Quality of air temperature (1-minute minimum)	Relative humidity in percentage %	Quality of relative humidity
132001	hd	31011	2019	4	2	16	0	25.8	Y	26.1	Y	25.8	Y	82	Y
132002	hd	31011	2019	4	2	16	1	25.6	Y	25.8	Y	25.6	Y	84	Y
132003	hd	31011	2019	4	2	16	2	25.5	Y	25.6	Y	25.5	Y	83	Y
132004	hd	31011	2019	4	2	16	3	25.5	Y	25.5	Y	25.5	Y	84	Y
132005	hd	31011	2019	4	2	16	4								
132006	hd	31011	2019	4	2	16	5	25.4	Y	25.5	Y	25.4	Y	85	Y

Figure 5. A weather record with incomplete readings

To correct these data, a technique of interpolation was adopted. Before interpolation starts, readings with quality status of "S" were removed as well, as this tag means “suspicious”, according to the data dictionary. These readings were corrected by interpolation as well. Missing values were then filled using interpolate linear method. This is deemed to be appropriate as the interval between records are considered small (1-minute interval). Figure 6 shows the record has a temperature of 25.45°C and a humidity of 84.5% after interpolation, which is derived by averaging readings from the preceding and succeeding records.

	hd	Station Number	Year Month Day Hour Minutes in YYYY	MM	DD	HH24	MI format in Local standard time	Air Temperature in degrees Celsius	Quality of air temperature	Air temperature (1-minute maximum) in degrees Celsius	Quality of air temperature (1-minute maximum)	Air temperature (1-minute minimum) in degrees Celsius	Quality of air temperature (1-minute minimum)	Relative humidity in percentage %	Quality of relative humidity
132001	hd	31011	2019	4	2	16	0	25.800000	Y	26.1	Y	25.8	Y	82.0	Y
132002	hd	31011	2019	4	2	16	1	25.600000	Y	25.8	Y	25.6	Y	84.0	Y
132003	hd	31011	2019	4	2	16	2	25.500000	Y	25.6	Y	25.5	Y	83.0	Y
132004	hd	31011	2019	4	2	16	3	25.500000	Y	25.5	Y	25.5	Y	84.0	Y
132005	hd	31011	2019	4	2	16	4	25.450000						84.5	
132006	hd	31011	2019	4	2	16	5	25.400000	Y	25.5	Y	25.4	Y	85.0	Y

Figure 6. A weather record with readings populated by interpolation

It is important that all weather records with missing or suspicious readings are correctly filled as missing readings would affect the accuracy of the aggregated value at the later part of this data preparation process.

For Solar Analytics data, it is of our view that it is inappropriate to apply any interpolation technique to create any missing record, as there are no specific patterns observed from the energy generated data, or any literature to suggest on how to fix this. Figure 7 shows an example of missing records found in Solar Analytics data.

site_id t_stamp_utc	46273932	92071696	131097238	133852058	173994775	186331009	230589246	241271752	264445722	315622138	327993590	373739408
2019-12-31 23:10:00+00:00	NaN	322.1428	178.0150	159.4733	338.3992	210.9289	-0.2292	330.7553	268.4158	357.3733	338.6703	NaN
2019-12-31 23:15:00+00:00	NaN	327.0872	252.1283	164.1964	341.8956	215.8147	-0.2272	332.1239	281.9211	237.1422	341.7394	NaN
2019-12-31 23:20:00+00:00	NaN	334.7617	319.7208	168.3283	346.3425	221.8678	-0.2275	334.2981	287.1581	207.0600	346.7994	NaN
2019-12-31 23:25:00+00:00	NaN	341.4083	341.9753	172.9736	350.4211	215.2486	-0.2278	335.3983	290.1258	335.4594	351.7992	NaN

Figure 7. Incomplete Solar Analytics data

Duplicate and Null Value Check

Duplicate check had also been performed to ensure no duplicate data exist in all the three datasets. Null value data are also been checked. There were some null value data in the solar analytics datasets, as shown in Figure 8. However, due to the low percentage of null value of each site, it is of our view that the impact of these missing data would be minimal. Duplicate records, which mainly occurred on 17 Apr, 30 Dec and 31 Dec data, were removed.

site_id	Null-Value Percentage
46273932	0.39
315622138	0.01
327993590	0.01
373739408	0.39
558829996	0.01
577341905	0.01
689045440	0.02
758907147	0.02
1150449095	0.02
1292510239	0.01
1323929011	0.39
1451036406	0.01
1883623070	0.01

Figure 8. Missing data (%) by installation site

Other Data Quality Check

Several reasonable data checks had been performed such as checking the number of sites per month, identifying any outliers, etc. This is to ensure the final dataset is of quality and ready for the next stage.

Removal of record with negative energy generated value

As this study is related to energy generated, keeping negative energy generated values would not be useful and may add unwanted noise to the data analysis. Moreover, most of these values fall in a range of 0 to -10. For this reason, records with negative energy generated values are removed.

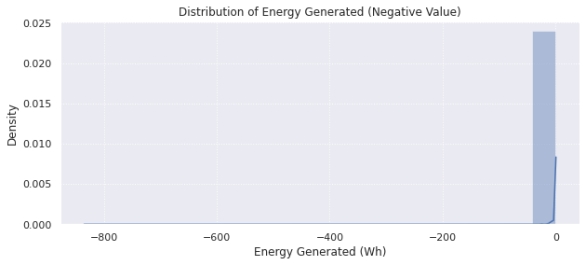


Figure 9. Distribution of Energy Generated (Negative Value)

Unused Attributes

As this study is related to PV energy generated, two attributes in the Solar Analytics data are found to be not useful and dropped, namely `'voltage_max_(V)'` and `'voltage_min_(V)'`.

Aggregating weather and solar irradiance data

Cleaned weather and irradiance data are aggregated to match the same frequency as Solar Analytics dataset. For weather data, both temperature and humidity readings were averaged into 5-minute

interval. For solar irradiance data, irradiance values were summed and aggregated into a 5-minute frequency.

Merge Solar Analytics, weather and solar irradiance datasets

After ensuring all datasets have the same frequency, they are merged into one single dataframe by matching the common keys. To merge site-based Solar Analytics dataset with weather-station-based weather and solar irradiance datasets, a table containing the mappings of postcode and weather-station, is used to join sites to weather-station.

Linear Regression

Multicollinearity

In linear regression, the interpretation of a coefficient is that it represents the mean change in the dependent variable for each unit change in an independent variable where other independent variables remain constant. Hence, if independent variables are correlated, it means a change in one of them would have an impact in the other variable. Multicollinearity makes coefficient unreliable and reduces its precision.

Hence, correlations between variables were examined due the reasons stated. It is also important to determine which variable has a higher correlation with the dependent variable. Variables that have a high correlation to another variable are removed if it has a lower correlation to the dependent variable. `'ghi'`, `'dni'` and `'dhi'` all have very correlations to each other and ultimately `'ghi'` is retained due to its higher correlation with the dependent variable `'energy_generated'`

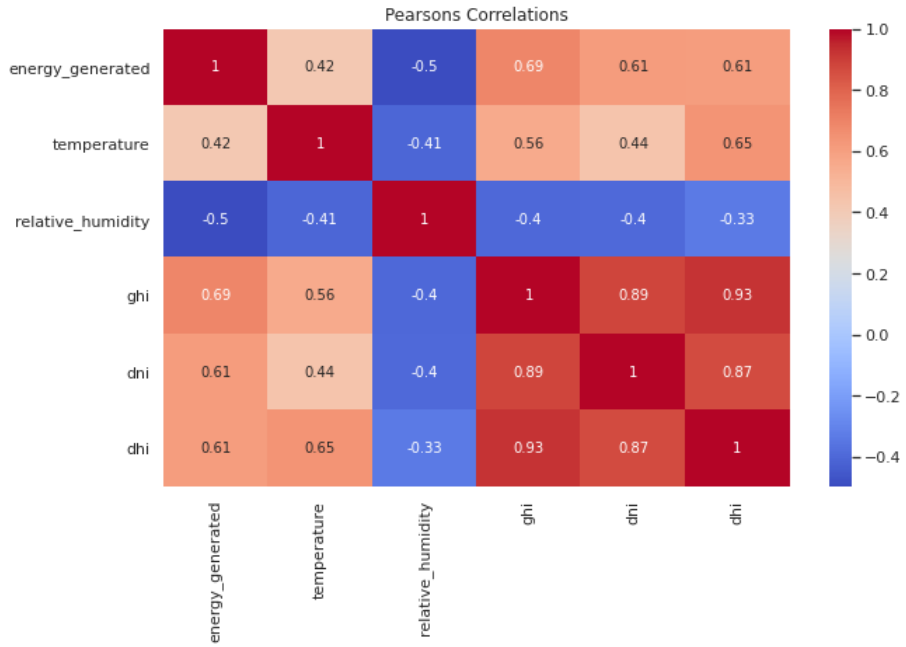


Figure 10. Correlation Matrix

Linear Regression Model

The next step after inspecting correlation is fitting a linear regression model. This is a statistical approach to examine the relationships between temperature, humidity, irradiance and PV energy generated. Characterise by its ease of interpreting result, linear regression is the simplest and

probably one of most widely used statistical modelling techniques. It determines if and to what extent several dependent variables impact the variability of a response variable.

For this study, a linear regression model including all the terms are fitted first. It is used as the baseline for further model fitted. Subsequent models fitted are to consider various interactions and term transformations. The performance of these models is evaluated using two metrics – R-squared value and Root Mean Squared Error (RMSE).

To ensure the model fitted do not have biased estimates, diagnostics plots are used to check whether the model fitted based assumptions for linear regression to yield robust result. This includes checking for linear relationships between independent and dependent variables, observations are independent, homoscedasticity check and normally distributed residuals.

Findings

Exploration Data Analysis

Average Daily Energy Output by Site

Solar PV systems have different capacity. Therefore, it is important to look at the daily energy output of each site. From Figure 11, it is observed that there is a huge difference between each site in term of average daily energy output. This suggests the capacity difference in these solar PV systems. This needs to be observed when fitting the linear regression model to account for site as one of the terms.

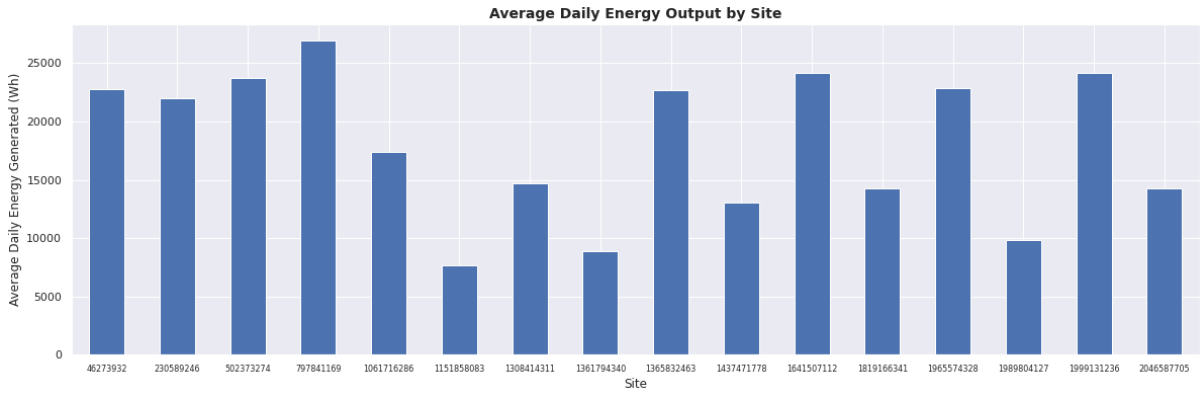


Figure 11. Average daily energy output by site

Distribution of 5-min Energy Generated (Wh) by Site

Build on the previous data exploration, the data were further explored into the data points of each site. There are some sites that are able to reach maximum capacity over a period of time. For example, in Figure 12, site 1641507112 which is located in Cairns has recorded many data points within its highest energy generated in.

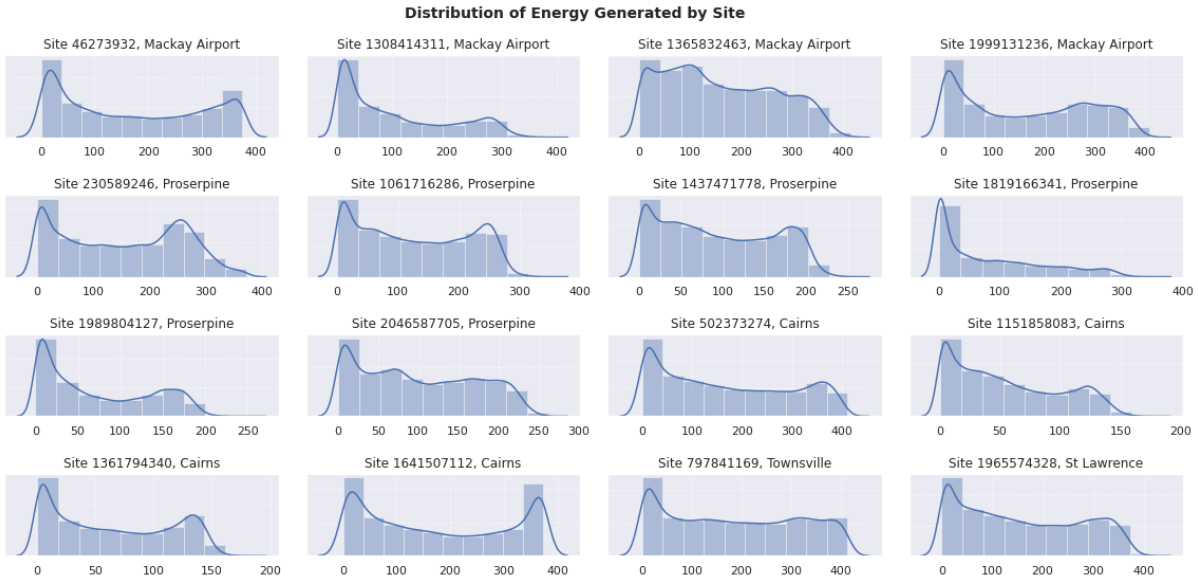


Figure 12. Average daily energy output by site

Distribution of Irradiance by Weather Stations

Figure 13 shows the three different irradiance values throughout the year of 2019. Notice that while both GHI and DHI show a dip during cooler months, DNI is not. This might be due to the fact that there are insufficient inputs, namely the incident angle between the collection plane and sun. They must be known for DNI to be calculated correctly.

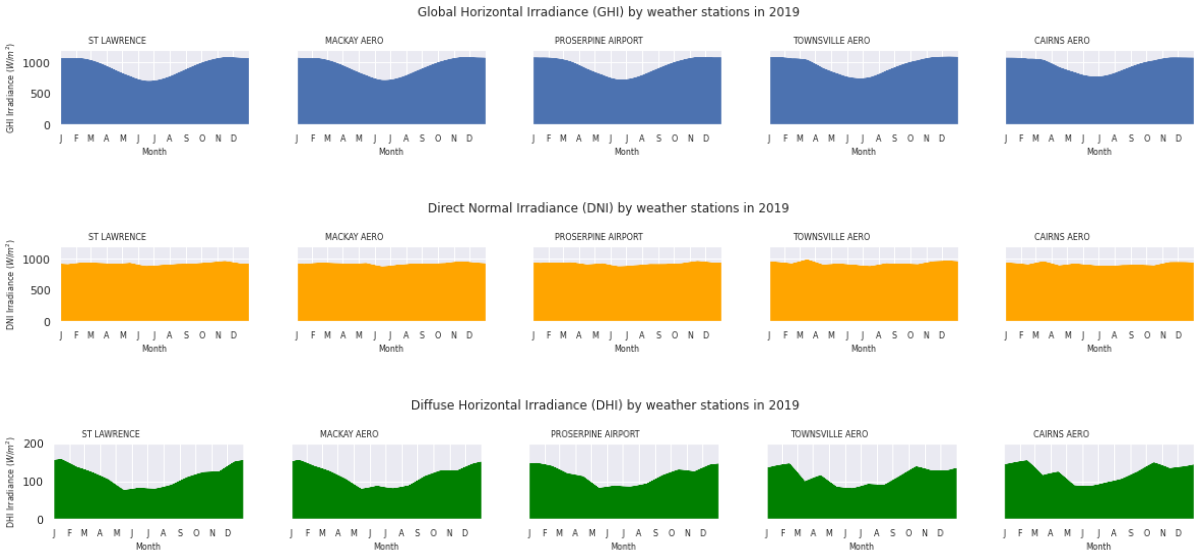


Figure 13. Distribution of irradiance by weather-stations

Temperature distribution with time for seasons in 2019

The weather conditions in Northern Queensland has four distinct seasons. The spring season starts from September to November, followed by summer season from December to February. Autumn season officially starts from beginning of March to the end of May, and winter from June to August. In this study, as we have data for the entire year 2019, it would be insightful to explore the variables in each season.

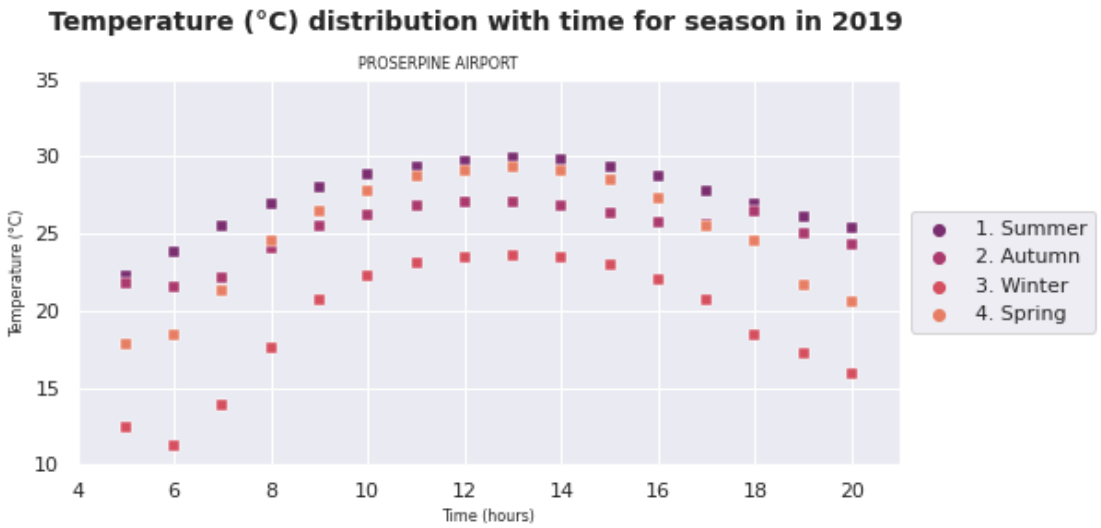


Figure 14. Temperature distribution with time for seasons in 2019

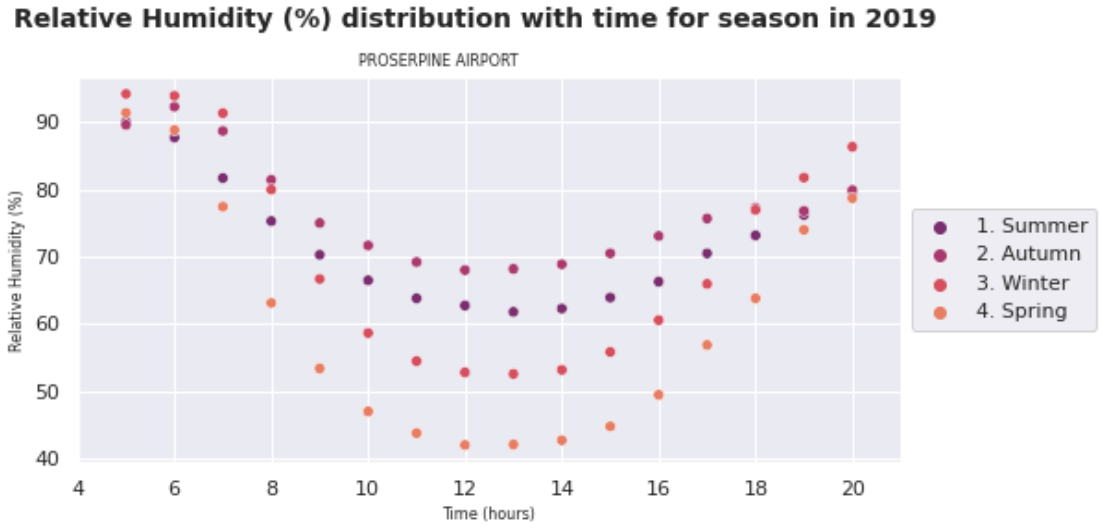


Figure 15. Relative humidity distribution with time for seasons in 2019

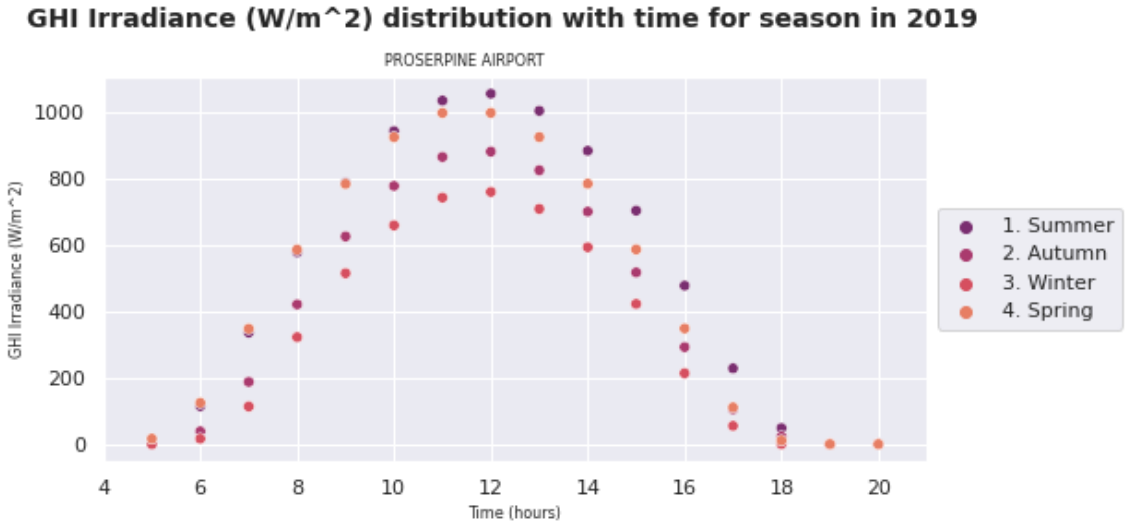


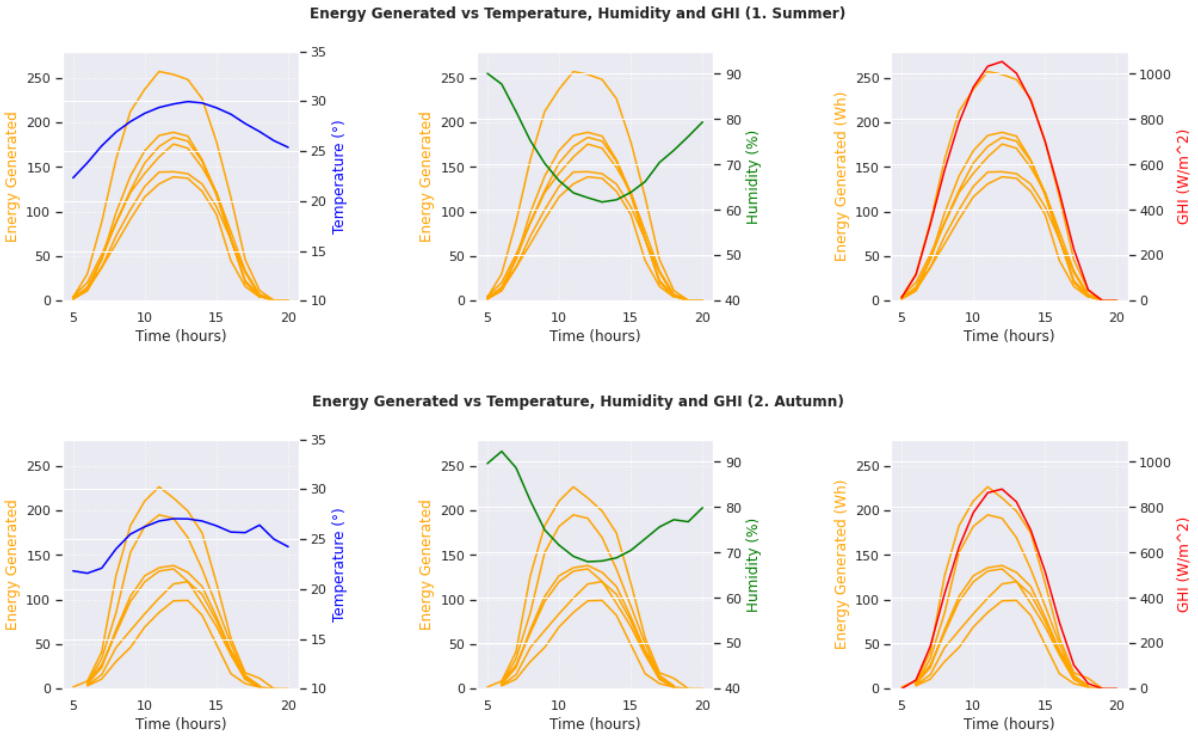
Figure 16. GHI irradiance distribution with time for seasons in 2019

To study the effect of seasonality, the values of temperature, humidity and GHI are averaged by hour of a day in a season. Here, Proserpine weather-station is selected for exploration. The temperature gap between summer and winter is about 10°C during the night-time and 6°C at noon time (Figure 14). However, it is completely different for humidity. In Figure 15, while the humidity stays above 90% early in the morning regardless of season, the humidity drops drastically at noon during spring to as low as 40%. This pattern may need to be further explored to see whether it has an impact on the PV system. Figure 16 shows that irradiance value could reach above 1,000 W/m^2 for a typical day in summer, whereas during winter, its peak is 20% lower than the summer's peak.

Energy Generated vs Temperature, Humidity and GHI by Seasons

Again, built on the previous data exploration, energy generated by each site in Proserpine is investigated. Figure 17 represents the recorded PV energy generated of the 6 sites installed at Proserpine group by seasons. The figure shows the significant differences between different seasons. When the plots are compared, the following were observed:

1- Although irradiance value is the highest during summer, none of the installation recorded its maximum energy generated in 2019 in this season;
2- Surprisingly some site recorded its lowest energy generated in this season, not winter, despite the 3 variables observed are relatively stable in autumn;
3- In winter, despite recorded the lowest irradiance value, sites had managed to produced more energy than in autumn;
4- Finally, PV systems are found to be most productive in spring, which has recorded the lowest humidity of the year. All systems are capable to reach a maximum average of at least 150 Wh. This result supports the findings of the literature that were reviewed in this study: the presence of water vapour in the air can impact the reception of solar radiation by PV system.



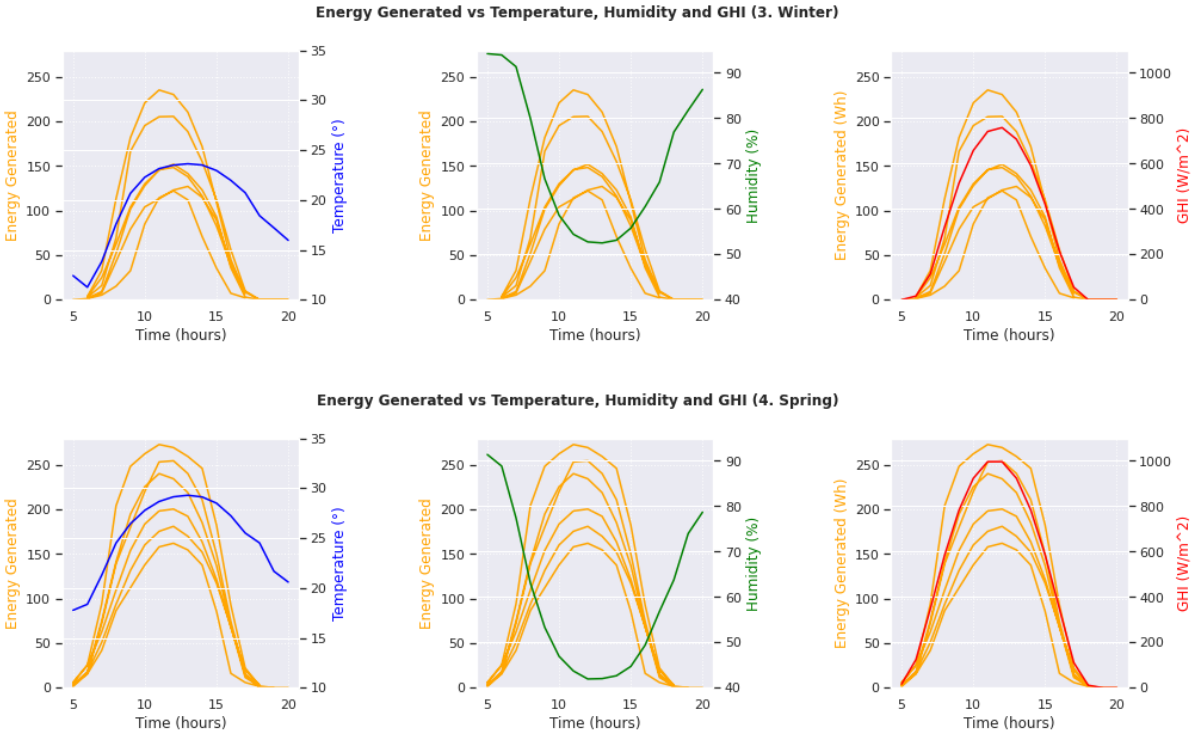


Figure 17. Energy Generated vs Temperature, Humidity and GHI by Seasons

Linear Regression Models

Several linear regression models have been fitted which include various transformations and polynomials on the independent variables. Only models with all terms satisfying p-value < 0.05 are presented here. The performance of the models fitted is summarised below:

#	Formula	R^2	RMSE
1	energy_generated~ temperature + relative_humidity + ghi + C(site_id)-1	0.688	61.18
2	energy_generated~ temperature * relative_humidity * ghi + C(site_id)-1	0.713	58.66
3	energy_generated~ temperature * ghi + relative_humidity * ghi + C(site_id)-1	0.708	59.21
4	energy_generated~ temperature+temperature_squared + relative_humidity + ghi + C(site_id)-1	0.689	61.09
5	energy_generated~ temperature + rh_squared+relative_humidity + ghi + C(site_id)-1	0.689	61.04
6	energy_generated~ temperature + relative_humidity + ghi_squared+ghi + C(site_id)-1	0.688	61.15
7	energy_generated~ temperature_squared + relative_humidity + ghi + C(site_id)-1	0.688	61.19
8	energy_generated~ temperature_squared * relative_humidity * ghi + C(site_id)-1	0.714	58.60
9	energy_generated~ temperature * rh_squared * ghi + C(site_id)-1	0.880	58.05
10	energy_generated~ temperature * relative_humidity * ghi_squared + C(site_id)-1	0.688	61.15
11	energy_generated~ temperature_cube * rh_squared*relative_humidity * ghi + C(site_id)-1	0.722	57.75
12	energy_generated~ temperature * rh_cube*rh_squared * relative_humidity * ghi + C(site_id)-1	0.575	71.39
13	energy_generated~ temperature * relative_humidity * ghi_cube * ghi_squared * ghi + C(site_id)-1	0.567	72.04
14	energy_generated~ temperature_squared + rh_squared + ghi_squared + C(site_id)-1	0.673	62.66
15	energy_generated~ temperature_cube + rh_squared + ghi_squared + C(site_id)-1	0.673	62.66
16	energy_generated~ temperature_squared + rh_cube + ghi_squared + C(site_id)-1	0.671	62.83
17	**energy_generated~ temperature_squared + rh_squared + ghi_cube + C(site_id)-1**	**0.846**	**65.74**

Table 1. Summary of linear regression models fitted

The performance of models is evaluated using two metrics – R-squared value and Root Mean Squared Error (RMSE). Ideally, lower RMSE and higher R-squared values are indicative of a good model.

It is important to note that all models fitted have some issues in satisfying linear regression's assumption of homoscedasticity, which is defined as the variance of residuals is the same for any value of independent variables.

As the value of R^2 suggested, model 9 would be an ideal model. However, its diagnostics plot of residuals vs various terms, as shown in Figure 18, is not considered ideal. The plot residual vs ghi is showing a cone shape instead of randomly scattered.

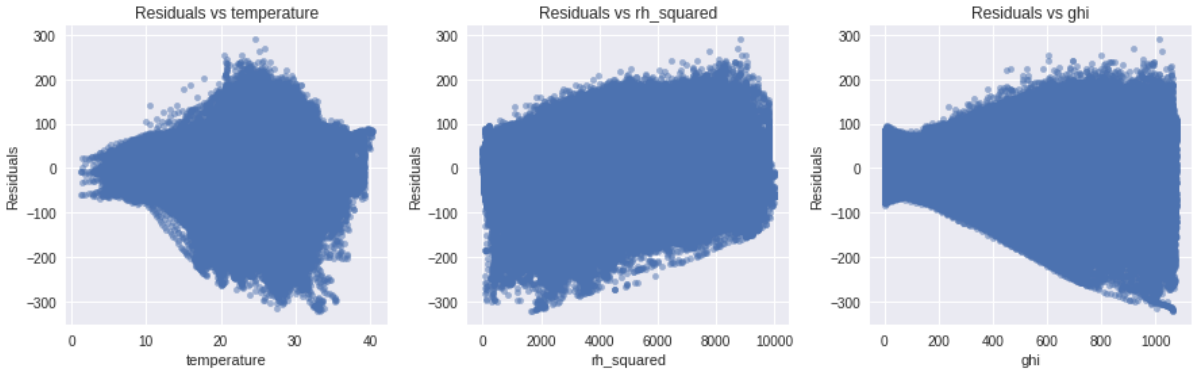


Figure 18. Model 9: Residual vs independent variables plots

Ultimately, after examining all the diagnostic plots together with R^2 and RMSE values, Model 17 is selected. Even though it has a higher RMSE value in comparison to some of the other models, it has the second highest R^2 value and better diagnostics plots. The model summary and its diagnostics plots are shown in Figure 19.

OLS Regression Results

Dep. Variable:	energy_generated	R-squared (uncentered):	0.846
Model:	OLS	Adj. R-squared (uncentered):	0.846
Method:	Least Squares	F-statistic:	2.413e+05
Date:	Sun, 25 Oct 2020	Prob (F-statistic):	0.00
Time:	06:57:18	Log-Likelihood:	-4.6596e+06
No. Observations:	831369	AIC:	9.319e+06
Df Residuals:	831350	BIC:	9.319e+06
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
C(site_id)[46273932]	192.9303	0.482	400.359	0.000	191.986	193.875
C(site_id)[230589246]	178.8652	0.481	372.019	0.000	177.923	179.808
C(site_id)[502373274]	189.9941	0.496	383.369	0.000	189.023	190.965
C(site_id)[797841169]	198.0696	0.481	411.829	0.000	197.127	199.012
C(site_id)[1061716286]	148.2691	0.483	306.865	0.000	147.322	149.216
C(site_id)[1151858083]	76.1600	0.497	153.268	0.000	75.186	77.134
C(site_id)[1308414311]	132.8245	0.479	277.227	0.000	131.885	133.764
C(site_id)[1361794340]	84.6928	0.499	169.776	0.000	83.715	85.670
C(site_id)[1365832463]	188.7005	0.478	394.804	0.000	187.764	189.637
C(site_id)[1437471778]	116.6497	0.483	241.698	0.000	115.704	117.596
C(site_id)[1641507112]	196.5356	0.499	393.911	0.000	195.558	197.514
C(site_id)[1819166341]	124.1852	0.462	269.030	0.000	123.280	125.090
C(site_id)[1965574328]	169.7380	0.460	368.889	0.000	168.836	170.640
C(site_id)[1989804127]	93.2920	0.485	192.445	0.000	92.342	94.242
C(site_id)[1999131236]	197.0784	0.475	414.845	0.000	196.147	198.009
C(site_id)[2046587705]	125.0252	0.480	260.223	0.000	124.084	125.967
temperature_squared	-0.0018	0.000	-3.879	0.000	-0.003	-0.001
rh_squared	-0.0167	3.75e-05	-444.857	0.000	-0.017	-0.017

```
ghi_cube                  1.568e-07   2.44e-10     642.618       0.000     1.56e-07     1.57e-07
==============================================================================
Omnibus:                     2813.621   Durbin-Watson:                   1.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2846.615
Skew:                          -0.140   Prob(JB):                         0.00
Kurtosis:                       3.059   Cond. No.                     1.09e+10
==============================================================================
```

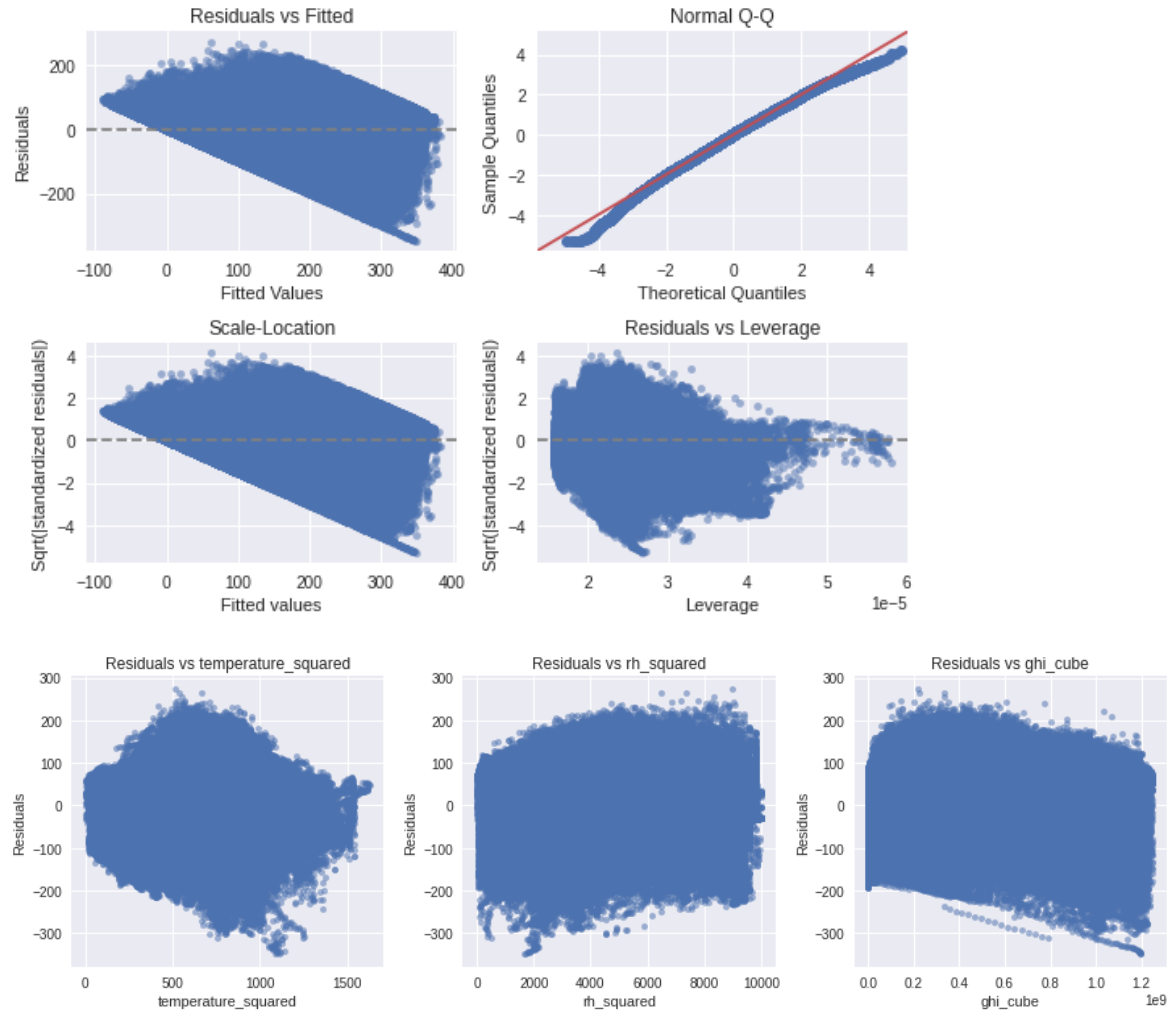


Figure 19. Model 17: model summary and diagnostics plots

Reflection

In this study, a great amount of time was spent on gathering and preparing the data. One of the rationales in handling these data is to preserve the high frequency data points in the Solar Analytics datasets if possible and to find the weather and irradiance dataset that could match this frequency. Australian Government Bureau of Meteorology has been contacted and they found to be very helpful in explaining some of the meteorological terms especially around solar data. Irradiance dataset were later added into this study after receiving project supervisor's feedback on the project proposal. Many literatures reviewed had included irradiance data. It is also through literatures reviewed after the project proposal submission that PVLIB was found to be useful in generating high frequency irradiance data.

Extra steps had also been taken to clean and prepare the data. This proves to be time-consuming due the large volume of data and computational constraints. Also, due the computational

constraints, the scope of this project was changed from 227 Queensland sites to just 16 sites in Northern Queensland.

Despite several attempts of fitting linear regression model, the diagnostics plots suggested that there is a presence of bias estimates. To improve this issue, one possible area to explore is to expand the dataset to include more variables. According to the PV Performance Modelling Collaborative (PVPMC), windspeed and air pressure could play vital roles which impact PV energy generated. Additional details, such as site's geo-coordinates, panels tilt angle, solar orientation, shading, etc. could also be included as inputs to calculate a more accurate version of solar irradiance value – plane of array (POA) to improve the accuracy of this model.

Still, the performance of each PV system can be is very sensitive to site factors and these new features must also be carefully assessed across different sites using the appropriate feature selection technique as many of these terms might not have any correlation to PV energy generated.

To further improve the statistical models built in this study, other statistical models, such as Generalised Linear Model (GLM) or Weighted Least Squares (WLS) Regression could be the next statistical models to be considered as they would have reduced the effects of heteroscedasticity which is present in most of the models fitted in this study.

In conclusion, the effect of temperature and relative humidity on PV solar systems was studied. During exploratory data analysis, it was found that PV system can generate more energy in a low humidity environment. Several linear regression models were developed in order to find the relationships to PV energy generated. These models were developed using different combinations of the three terms with polynomials and transformations considered. A model was selected out of the 17 models presented in this study. However, from this study we would not attempt to quantify the variability from this model as it still has some issues in meeting the homoscedasticity assumption which suggests the model is less robust.

Afterall, "*all models are wrong, but some models are useful*". (George E.P. Box)

References

[1] Queensland Renewable Energy Expert Panel, "Credible pathways to a 50% renewable energy target for Queensland," 2016. [Online]. Available: https://www.dnrme.qld.gov.au/__data/assets/pdf_file/0018/1259010/qreep-renewable-energy-target-report.pdf. [Accessed 10 August 2020].

[2] Department of Natural Resources, Mines and Energy, Queensland Government, "Achieving our renewable energy targets," 2 April 2020. [Online]. Available: https://www.dnrme.qld.gov.au/energy/initiatives/achieving-our-renewable-energy-targets. [Accessed 10 August 2020].

[3] S. Kurtz, D. Townsend, C. Whitaker, A. Maish, R. Hulstrom and K. Emery, "Outdoor rating conditions for photovoltaic modules and systems," *Solar Energy Materials and Solar Cells,* vol. 62(4), pp. 379-391, 2000.

[4] Geoscience Australia, Bureau of Resources and Energy Economics, Australian Bureau of Agricultural and Resource Economics, "Australian Energy Resource Assessment (Second Edition)," Commonwealth of Australia (Geoscience Australia), Canberra, 2014.

[5] Clean Energy Regulator, "Postcode data for small-scale installations," 22 07 2020. [Online]. Available: http://www.cleanenergyregulator.gov.au/RET/Forms-and-resources/Postcode-data-for-small-scale-installations#SGU--Solar-Deemed. [Accessed 15 08 2020].

[6] M. T. Chaichan and H. A. Kazem, "Experimental analysis of solar intensity on photovoltaic in hot and humid weather conditions," *International Journal of Scientific & Engineering Research,* vol. 7, no. 3, pp. 91-96, 2016.

[7] H. A. Kazem, M. T. Chaichan, S. A. Saif, A. A. Dawood, S. A. Salim, A. A. Rashid and A. A. Alwaeli, "Experimental Investigation of Dust Type Effect on Photovoltaics Systems in North Region, Oman," *International Journal of Scientific & Engineering Research,* vol. 6, no. 7, pp. 293-298, 2015.

[8] A. Sayyah, M. N. Horenstein and M. K. Mazumder, "Energy yield loss caused by dust deposition on photovoltaic panels," *Solar Energy,* vol. 107, pp. 576-604, 2014.

[9] S. Dubey, J. N. Sarvaiya and B. Seshadri, "Temperature Depedent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the Word - A Review," *Energy Procedia,* vol. 33, pp. 311-321, 2013.

[10] M. Bayrakci, Y. Choi and J. R. S. Brownson, "Temperature Dependent Power Modeling of Photovoltaics," *Energy Procedia,* vol. 57, pp. 745 - 754, 2014.

[11] E44.09 Solar, Geothermal and Other Alternative Energy Sources (Sponsoring Committee), "Standard Test Methods for Electrical Performance of Nonconcentrator Terrestrial Photovoltaic Modules and Arrays Using Reference Cells," 2012. [Online]. Available: https://doi.org/10.1520/E1036-12. [Accessed 15 08 2020].

[12] K. Kawajiri, T. Oozeki and Y. Genchi, "Effect of Temperature on PV Potential in the World," *Environmental Science & Technology,* vol. 45, pp. 9030-9035, 2011.

[13] W. Liao, Y. Heo and S. Xu, "Evaluation of Temperature Dependent Models for PV Yield Prediction," in *4th Building Simulation and Optimization Conference*, Cambridge, UK, 2018.

[14] H. A. Kazem, M. T. Chaichan, I. M. Al-Shezawi, H. S. Al-Saidi, H. S. Al-Rubkhi, J. K. Al-Sinani and A. H. A. Al-Waeli, "Effect of Humidity on the PV Performance in Oman," *Asian Transactions on Engineering,* vol. 02, no. 04, pp. 29-32, 2012.

[15] Sandia National Laboratories, "PV Performance Modeling Collaborative," [Online]. Available: https://pvpmc.sandia.gov/.

[16] Solar Analytics Pty Ltd., "2019 Australian Voltage and PV Generation as part of ARENA, Sydney [Data set]," 2020. [Online]. Available: http://www.solaranalytics.com/data.

[17] R. W. Andrews, J. S. Stein, C. Hansen and D. Riley, "Introduction to the open source PV LIB for python Photovoltaic system modelling package," *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC),* pp. 0170-0174, 2014.

[18] J. W. Tukey, Exploratory Data Analysis, Pearson, 1977.

Appendices

A. Solar Analytics Data Dictionary

This data set ranges from 1st Jan 2019 - 31st Dec 2019 in 5 minute sampling rate. The data set includes 1000 residential sites across Australia from the Solar Analytics data set (randomly selected from a subset that has passed through our privacy controls).

Column Name	Description
t_stamp_utc	The time stamp in UTC
site_id	The ID of this site
energy_(Wh)	The PV energy produced during the 5 minutes in watt-hour
voltage_max_(V)	The maximum voltage during the 5 minutes in volt
voltage_min_(V)	The minimum voltage during the 5 minutes in volt

B. BOM Climate Data Dictionary

Byte Location	Byte Size	Description
1-2	2	Record identifier - hd
4-9	6	Bureau of Meteorology Station Number.
11-26	16	Year Month Day Hours Minutes in YYYY,MM,DD,HH24,MI format in Local standard time
28-32	5	Air Temperature in degrees Celsius
34	1	Quality of Air Temperature
36-40	5	Air temperature (1-minute maximum) in degrees Celsius
42	1	Quality of Air Temperature (1 minute maximum)
44-48	5	Air temperature (1-minute minimum) in degrees Celsius
50	1	Quality of Air Temperature (1 minute minimum)
52-54	3	Relative humidity in percentage %
56	1	Quality of Relative humidity
58-60	3	Relative humidity (1 minute maximum) in percentage %
62	1	Quality of relative humidity (1 minute maximum)
64-66	3	Relative humidity (1 minute minimum) in percentage %
68	1	Quality of Relative humidity (1 minute minimum)
70	1	# symbol - end of record indicator

C. Solar Irradiance Definitions

Irradiance	Description
Global Horizontal Irradiance	Global Horizontal Irradiance (GHI) is the amount of terrestrial irradiance falling on a surface horizontal to the surface of the earth. If GHI cannot be measured directly, it may be calculated from direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) using the following equation: GHI = DHI + DNI x $\cos(\theta z)$ In some PV systems, GHI is measured and a model, such as the DISC or DIRINT models, is used to estimate the DNI or DHI.
Direct Normal Irradiance	Direct Normal Irradiance (DNI) may be measured directly via an absolute cavity radiometer. If direct measurements of DNI are not available, DNI may be calculated

	via co-planar measurements of the diffuse and total radiation by devices with a 180° field of view (the incident angle between the collection plane and sun must also be known). If co-planar diffuse and total radiation measurements are unavailable, models have been developed to estimate DNI from the global horizontal irradiance (GHI) and other environmental factors.
Diffuse Horizontal Irradiance	Diffuse horizontal irradiance (DHI) is the terrestrial irradiance received by a horizontal surface which has been scattered or diffused by the atmosphere. It is the component of global horizontal irradiance which does not come from the beam of the sun (where “beam” is a 5° field of view concentric around the sun). If diffuse horizontal irradiance is not measured directly, it may be calculated in a fashion similar to global horizontal irradiance

D. Python Codes