University of
South-Eastern Norway

www.usn.no

FMH606 Master's Thesis 2022
Energy and Environmental Technology

Solar Power Electricity Production Correlated to Meteorological Data

Skagerak Arena, Skien, Norway.

OZGUR YALCIN

Faculty of Technology, Natural sciences and Maritime Sciences
Campus Porsgrunn

University of South-Eastern Norway

www.usn.no

Course: FMH606 Master's Thesis, 2022

Title: Solar Energy Production Correlated to Meteorological Data

Number of pages: 124

Keywords: Photovoltaic systems, Forecasting, Artificial neural networks, Linear regression, Weather parameters, Meteorological data, PV power output.

Student:	Ozgur Yalcin
Supervisor:	Assoc. Prof. Kjell-Arne Solli
External partners:	Assoc. Prof. Erik Berge, Meteorologisk Institutt.
	Heine Nygard Riise, Institue for Energy Technology.
	Stig Simonsen, Lede Energi.

University of South-Eastern Norway

www.usn.no

Summary:
Photovoltaic (PV) power production predictions have gained immense popularity in recent years. More and more solar power connects to grids. Considering PV power dependency on various weather parameters, improvements in power prediction possess a massive potential for optimisation and accurate forecasting.

The main objective of this study is to understand weather parameters that affect PV power production. Furthermore, proposing a model to predict power output from historical data is one of the goals. It is expected to the developed PV power prediction model will give insights into forecasting.

PV plant historical data was kindly shared by Lede Energi for Skagerak Arena in Skien. Meteorological data was gathered through meteorological institute frost application programming interface (API) for Gjerpen station which is operated by NIBIO. Air temperature, global horizontal irradiance, wind speed, wind category, relative humidity, and dew point temperature variable in addition to module temperature, and clear sky parameters from pvlib package in python were subject to examination. These variables' impact was investigated on PV power output for a period from 2020 to 2021 on an hourly basis. In particular, weather parameters analysed from 2018 to 2021 to understand changes in climate on a yearly basis. After merging all data, correlation and principal component analysis were performed. Linear regression (LR) and artificial neural networks (ANNs) models were proposed and were tested on various cases.

Models were best performed on consecutive clear sky days with mean absolute error of 2.04 kW, and 1.66 kW for LR and ANN, respectively. ANN did a better job of prediction consecutive clear sky days compared to LR. Furthermore, models were evaluated for a longer testing set period from 2020 to 2021. While the mean absolute error for ANN was 2.41 kW, LR was 2.92 kW. The study indicates that the ANN model's prediction results are slightly improved compared to LR models. Besides, handling different sampling rates within datasets and their impact on the model accuracy were discussed.

As a result of the meteorological variable selection case, it is concluded that while the model run by only irradiance and air temperature values produce sufficient results, the best performance was obtained by adding relative humidity and other sun parameters. In addition, it is found out that all variables that were investigated have an effect on power value predictions during relevant weather variable fluctuation periods.

Preface

I would like to show my deepest appreciation to Assoc Prof. Kjell-Arne Solli for the good advices, support and availability, and constructive comments. His suggestions and flexibility in sharing ideas were precious to me.

I would also like to express my gratitude to Assoc. Prof. Erik Berge for the greatest support, advices, and helping me to access meteorological data. In addition, a special thanks to Heine Nygard Riise for his support, valuable suggestions and fruitful discussions.

I also like to thank Stig Simonsen for helping me to provide power plant values and for his effort in answering my questions throughout the project.

Finally, words cannot express my thankfulness to my parents, Zerrin and Dursun Yalcin for their encouragement and support.

Drammen/Porsgrunn, 18/05/2022

Ozgur Yalcin

Contents

List of Tables

List of Figures

Nomenclature

Symbols	Definition	Units
AC	Alternating current	Ampere
DC	Direct current	Ampere
DHI	Direct horizontal irradiance	W/m2
DNI	Direct normal irradiance	W/m2
GHI	Global horizontal irradiance	W/m2
FF	Fill factor	-
I_{MPP}	Nominal power current	Ampere
I_{SC}	Short circuit current	Ampere
MAE	Mean absolute error	-
MAPE	Mean absolute percentage error	%
MSE	Mean squared error	-
P_{IN}	Incident power	W
P_{MAX}	Power value at maximum	W
P_{MPP}	Nominal power	W
R^2	Variance	-
RMSE	Root mean squared error	-
V_{MPP}	Nominal power voltage	V
V_{OC}	Open circuit voltage	V
WS	Wind speed	m/s

List of abbreviations:

AM	Air mass
ANN	Artificial neural networks
AOD700	Aerosol optical depth
API	Application programming interface
DBN	Deep belief network
GPU	Graphics processing unit
GW	Giga watts
IV1	Inverter 1
IV2	Inverter 2
IV5	Inverter 5
IV7	Inverter 7
LR	Linear regression
LSTM	Long-Short term memory
NOCT	Nominal operating cell temperature
NMOT	Nominal module operating temperature
MET	Metrological
MPPT	Maximum power point tracker
MW	Mega watts
PC	Principal component
PCA	Principal component analysis
PERC	The passivated emitter and rear cell
PERL	Passivated emitter, rear locally doped
POA	Plane of array
PV	Photovoltaic
RELU	Rectified linear unit function
RF	Random forests
RNN	Recurrent neural network
STC	Standard test condition
SVD	Singular value decomposition
SVM	Support vector mechanism
TDS	Thredds Data Server
UTC	Coordinated universal time

1 Introduction

Without a doubt, photovoltaic (PV) systems play an important role in changing the world by its emission free, and quick installation capabilities. In addition to that, the decreasing trend in manufacturing costs makes PV systems more attractive amongst other renewable sources. According to Rystad energy, prices in 2020 which was 0.20$ per watt peak (Wp), jumped to 0.26-0.28$ per Wp in the second half of 2021 due to the material price inflation [1]. Even though the current downward trend has been disrupted by the latest price inflation around the globe, it is expected that PV system installations continue to surge in the coming years. Solar PV systems installed capacity was 843 TWh in 2021 and International Energy Agency (IEA) expects that PV capacity reaches up to 4958 TWh by 2050 [2]. Considering the goals in PV capacity, grid capacity and its optimisation for renewable sources are subject to debate. Lately, grid companies put a great deal of effort into grid optimisation and improvements.

When it comes to PV power production, battery systems have gained popularity due to the lack of power production when the sun is out. That is why, solar power production systems, whether residential or commercial, commonly consist of solar modules with battery systems. While all parts work in harmony, power generation and consumption balance determine the electricity flow either to the grid or batteries. At this level of penetration, prediction of solar energy comes on the scene. Power production companies have integrated forecasting outcomes into their systems. As a result of this integration, it has become possible to get the most out of the sun energy and feed into the grid with the most effective amount. Hence, allocation of power depending on demand and load allows companies to maximize their profits in addition to increased grid security. However, solar power forecasting is a challenging task and may result in economic losses if it is not managed effectively. There are many different approaches for forecasting models such as using only irradiance values or taking into account other meteorological parameters. Models also differ in prediction methods such as regressions, and machine learning algorithms. To reach out higher accuracy and prediction capability of power with minimal errors, forecasting models are continuously in the process of improvement. Different algorithms and model inputs with feature and variable selection are being studied by researchers. Furthermore, the prediction of solar power output not only reduces the probability of power imbalance in the market but also secures the high penetration of PV systems in the grids for extended periods.

Since this study was conducted in Norway, it might be beneficial to provide some information on PV trends in Norway. In 2020, 40 MW of solar panels was installed and the total capacity has reached 160 MW. DNV GL predicts that installed capacity will increase to 1.75 GW by 2040 [3]. Additionally, it is expected that PV will account for only 1% in total electricity production. Relatively low solar radiation compared to central Europe degrades Norway's power business motivation for PV systems. The average daily solar radiation is 2.46 kWh/m^2 and it can reach up to 5.5 kWh/m^2 during summer [4]. Precisely, advances in PV technology and better optimisation with forecasting provide unique opportunities to the national grid.

1.1 Task Description

The main objective of this study is to understand weather parameters that effect PV power production. Furthermore, proposing a model to predict power output from historical data is one of the goals. It is expected to the developed PV power prediction model will give insights into forecasting.

Historical data on the weather observation from MET (The Norwegian Meteorological Institute), and photovoltaic power from Skagerak Arena on an hourly basis will be correlated to electricity production from solar energy. Main data for this task are observations from Gjerpen weather station (Skien) and Skagerak Energilab (Skagerak Arena, Skien). Observed variations in electricity production will be discussed related to the proper operation of the electrical network.

The Norwegian Meteorological Institute, IFE (Institute for Energy Technology), and Skagerak Energi are the project partners of the study. Submitted task description can be found in Appendix A.

1.2 Motivation

Skagerak Arena is home to the first of its type project in Norway. Odd soccer club's Skagerak Arena in Skien, Norway proved that lights can be powered by local renewable sources when the team plays in the evenings. 5,700 square meters of solar modules, with a nominal power of 800 kWp were placed on the rooftop of the stadium by Skagerak Energi. Getting the chance of analysing power data from such a special place, Skagerak Arena, was one of the main motivations.

One of the project partners, IFE, carries on a project called Sunpoint which is a research project for analysing solar power potential. The project aims to increase the estimation energy production of solar power plants in Norway. One of the main study areas of the project is predicting more accurate solar irradiance values for Norway. Thus, localised accurate data will be provided Norwegian solar energy market. By using PV power data and measured irradiance values, this study's outputs can be another valuable input to other projects within Norway.

The main driven motivation throughout the thesis is dealing with large datasets, and having a chance to practise data dealing techniques, statistical methods and machine learning algorithms. Writing python codes except for libraries from the scratch was a challenging part of this study. Furthermore, Norway is one the countries that is located in high latitude. For such a country that has challenging climate, understanding the power of sunlight reaches on the surface and its potential for electricity production is going to be an exciting part of this study.

This thesis is one of a kind for Norway in terms of scaling of power and content of the study such as clear sky algorithms, extensive information meteorological variable – PV data relationship, and meteorological variable input selection.

1.3 Brainstorm for deciding on models

At the beginning of the project, no specific model for historical data analysis and forecasting were stated. The chosen models were decided in the process of literature research. As is well known, PV power production is heavily dependent on the amount of irradiance that solar modules receive. It is also common knowledge that as air temperature increases so as PV cell temperature, the module efficiency decreases. A similar temperature effect is caused by wind speed. The higher wind speed means the lower cell temperature. All fluctuations seem that there is a linear relationship. That is why a simple but effective model such as linear regression may worth to be investigated. However, there are other variables that the relationship with PV power production is not clearly known such as wind direction or humidity. Moreover, different underlying relationships during season changes or the combination of other variables may produce non-linearity in data. In understanding and unrevealing hidden relationships in datasets, more advanced methods are used. For example, artificial neural networks (ANNs) are famous for achieving higher accuracy in time series forecasting [5]. In the literature, it is possible to follow that there is a growing interest in PV power prediction and forecasting by using ANNs [6]. Other supervised machine learning methods such as random forests (RF), and support vector machines (SVMs) are also good fits for predictions [7].

1.4 The Outline of the Thesis

The thesis is structured as follows. In section 2, it is given a general theoretical background on PV systems, and the working principles. In addition, PV cell information is provided based on the cell type that is used in the plant. In the second part of the theoretical background section, short information is available for meteorological variables and measurement methods. In the literature review part, data dealing methods specifically on meteorological variables and PV power values analysed through articles in the literature. Afterward, machine learning algorithms and prediction techniques are explored and model evaluation methods are presented.

In section 3, the PV plant layout and meteorological station are presented. Besides, the data type and time scale are introduced. Data processing methodology is also explained in this section. Moreover, clear sky studies where different methods have been compared, are discussed in this part with examples. It has aimed by clear sky part that the by using measured irradiance values, detecting clear sky days automatically, and filtering out from the datasets easily for future analysis. Correlation analysis is conducted to analyse the relationship within input parameters and principal component analysis is used to investigate the possibility of dimension reduction and a better understanding of feature interdependence.

In section 4, all results including historical data prediction and forecasting are presented with figures and tables.

Section 5 discusses the outcome of the results and gives information on different trials on the way to forecasting steps. In addition, some suggestions for the continuity of this project are stated as future work. And lastly, in section 6, conclusions on all results and discussions are delivered. Some important parts from the code are given in Appendix F.

2 Theoretical Background

In this part, some theoretical knowledge on PV systems, meteorological variable explanations and measurements and literature survey are presented. It is also aimed to give some insights into general concepts of thesis work components to readers.

2.1 Photovoltaic Systems

A typical photovoltaic system consists of three main groups. These are PV modules, inverters and grid or battery connection parts. Several photovoltaic cells are assembled in series or parallel circuits and a module is formed. A photovoltaic panel may include one or several modules together and gets the form ready to install. Moreover, a complete power generation unit is named as an array where PV panels are connected in series or parallel connections. Figure 2.1 describes a compact illustration of a cell, a module, a panel and an array system.

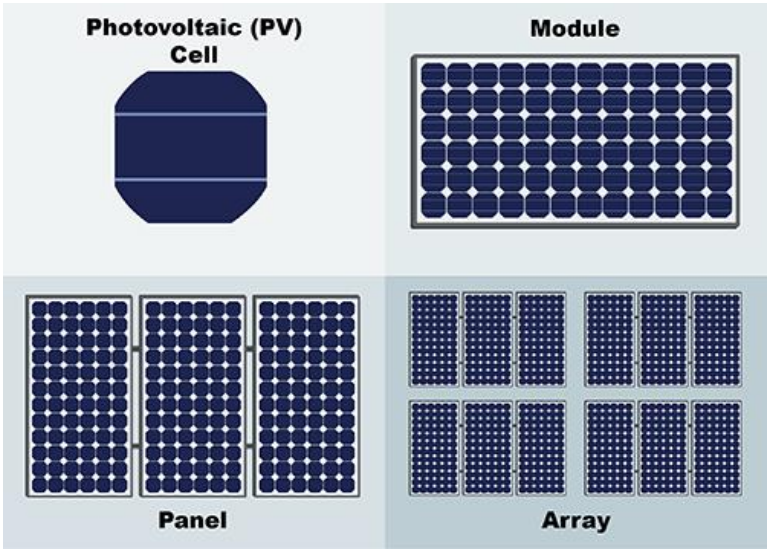


Figure 2.1: Photovoltaic cells, modules, panels and arrays [8].

Unless direct current (DC) is produced from PV modules directly stored battery, it is converted to alternative current (AC) by inverters to be ready last user consumption. However, the electricity goes into several steps before being connected to the grid. First of all, a junction box is installed commonly behind the PV panels to harness electricity from the panel. In addition, the junction box includes bypass diodes to prevent the PV panel from any reverse current due to shadow or darkness. The junction box also includes string fuses is used to protect the wiring from overloading. Next, the electricity should be monitored by a controller. The controller box monitors and tracks the PV generation from the panel and the information feeds up next processes. DC electricity generated from PV panels is converted into alternative current by inverters. In PV systems, inverters are particularly designed for working in the maximum power point (MPP). The inverter is adjusted to get the maximum possible power from PV panels by the MPP tracking system. An electronic circuit adjusts the voltage so that the inverter works at the PC maximum power point. In case of fault status, direct current load switches isolate the inverter from the PV generator in large multi-inverter systems. Lastly, power metering and controlling systems work in harmony to provide electricity for the grid or consumer unit. Cabling is another important part of PV systems as different cable types are

used for module string cables, DC cables and AC cables. Nominal voltage in addition to mechanical and weather determine the specification of cables.

2.1.1 PV Cell

A PV cell absorbs sunlight and converts it to electricity by creating an electrical current. Cell power output heavily depends on solar irradiance, ambient temperature, and other factors that could lead to energy loss. PV Cell technology evolves rapidly to maximize power output and new cell types are coming on the market. While silicon solar cell accounts for the vast majority of the market, other types of cells are becomingly desirable such as crystalline silicon thin-film solar cells, high-efficiency III–V multijunction solar cells, and organic photovoltaics.

A typical semiconductor solar cell consists of the n-type and p-type layers, anti-reflective layers and metal contact. When light hits the surface of a silicon semiconductor solar cell, light triggers electrons in the silicon and results in electron movement from the n-type layer to the p-type layer. Thus, a flow of electricity is created.

In this part, one particular type of solar cell will be explained in detail. One of the advanced solar cell types is the Passivated Emitter and Rear Cell (PERC) which is also installed in Skagerak Arena. PERC cells can be classified as rear passivated cells that differ from traditional solar cells due to the capability of a high amount of light capture. Figure 2.2 compares a conventional cell and a PERC cell. In particular, electrons are not captured by the rear surface and reflected electrons contribute to the current again in the PERC cell [9].

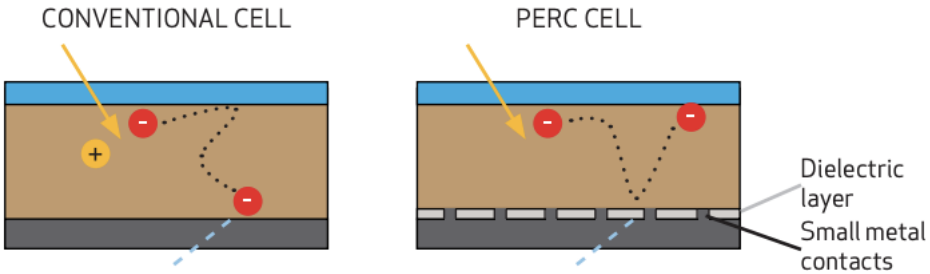


Figure 2.2: Conventional and PERC Cell comparison [9].

Figure 2.3 shows a Passivated Emitter, Rear Locally-doped (PERL) Cell which is one of the most common PERC configurations [10]. While the pyramid surface captures most of the light with its design, rear contact with the oxide surface ensures that getting the most out of the light. The light that is reflected from rear contact reaches the surface of pyramids and goes back into the cell [11]. Thus, the efficiency of the cell may reach up to 25%.

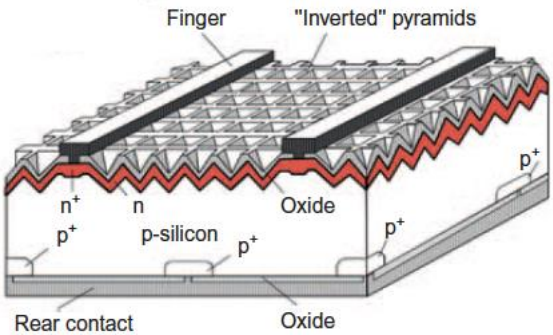


Figure 2.3: PERL Cell (Passivated Emitter, Rear Locally-doped) [10].

PERC cells which are used in Skagerak Arena have a brand name called REC Twinpeak 2 Series Module that has 120 half-cut multi-crystalline cells. Half-cut cell technology aims to reduce power loss by splitting standard square cells into two smaller parts. As a result, internal current decreases by 50% and reduces power loss as well as lowers resistive losses [12]. In addition, multi-crystalline refers to multiple separate crystals forming the cell. It is worth mentioning that in contrast to multi cells, mono cells have a higher energy yield as absorption efficiency is higher [12].

2.1.2 PV Module

When it comes to installing multi-PV modules and connecting to each other, it is important to design module arrays to get the most out of PV power production. Series and parallel module combinations are two ways of installing a PV module system. In series connections, each module's power production occurs at the same current and voltage values add up while in parallel connection, power production occurs at some voltage and current values add up [13]. Figure 2.4 represents an array distribution including series and parallel connections. If a module produces different power from others in a string due to for example partial shading, bypass diodes prevent other modules from failing. The shaded module is not influenced by reverse voltage and does not consume any power from other modules due to bypass diode blocking. Similarly, blocking diodes in series ensures that the current flows only in one direction and the current does not go back to failing strings or modules. Since the current flow in a series connection is determined by the lowest current value, if there is a current reduction in a cell or module, it will result in a loss of power. Similarly, the system voltage is determined by the lowest voltage in a parallel connection and lower voltage results in power loss.

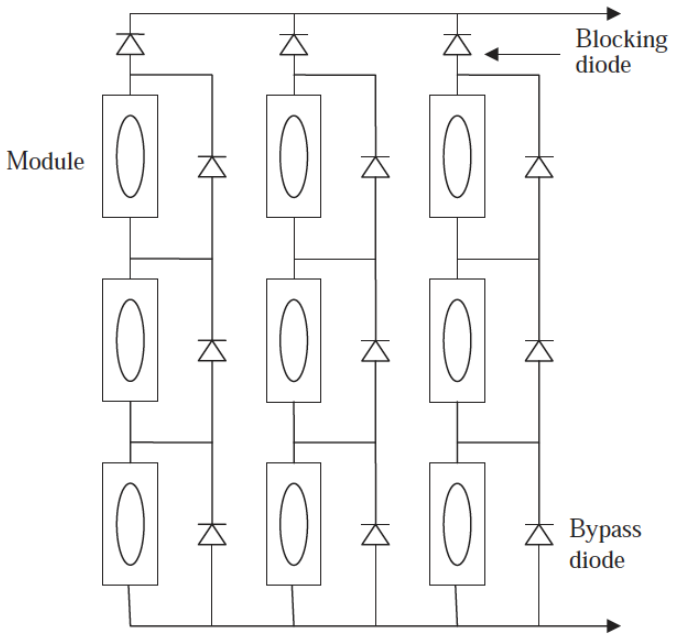


Figure 2.4: Parallel and Series connection of module in a PV system [13].

Shading plays an important role in the energy yield of PV systems. Partial shadowing affects PV output both current and voltage values. Seapan et al. [14] investigated the shading effect on a PV module by analysing maximum voltage V_{mp} and maximum current I_{mp} values. In

Figure 2.5, x_c indicates the ratio of shaded area in each cell. As it is seen clearly from the figure that current and voltage values are reduced caused by the shading.

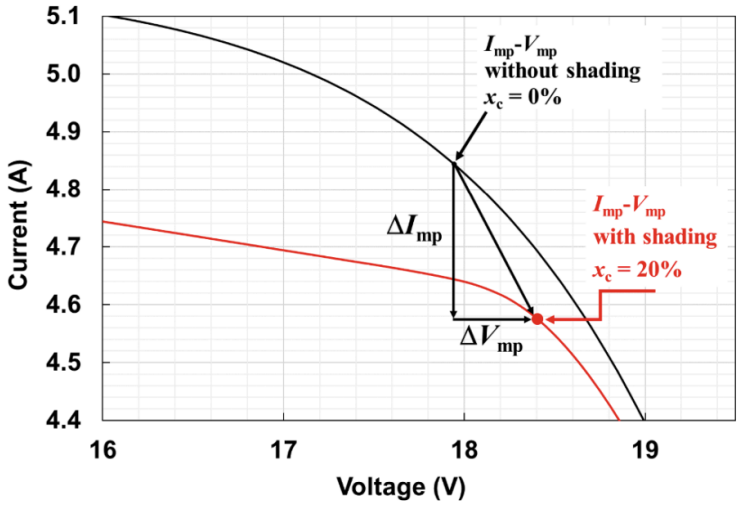


Figure 2.5: Variation of the V_{mp} and I_{mp} caused by the shading [14].

The distinction between series and parallel module connection and its effect on voltage and current values are illustrated best in Figure 2.6. As it is explained above, a series connection of modules which is called a string produces the same amount of voltage and the voltage values are added up. Similarly, the parallel connection of strings produces the same amount of current and is added up. Hence, a PV system capacity reaches up to gigawatts.

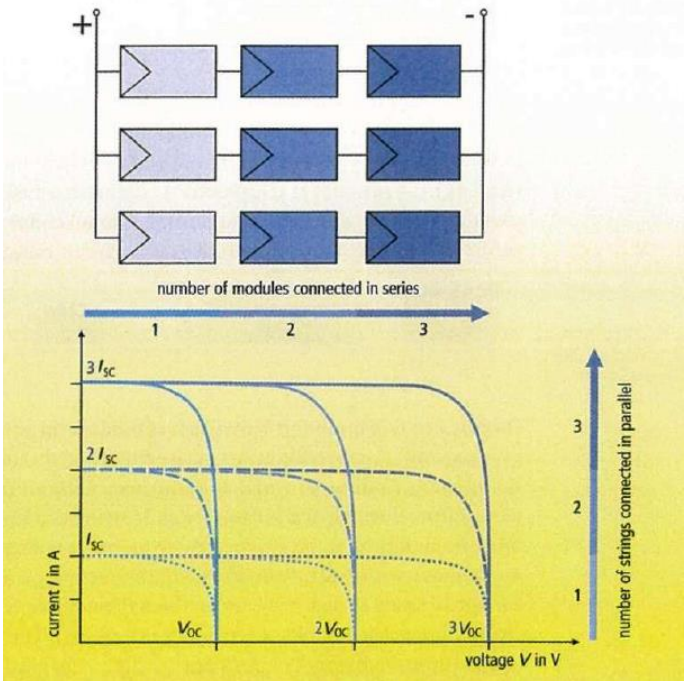


Figure 2.6: Interconnection of PV modules [15].

2.1.3 Electrical Characteristics

In the PV module specification sheet, it is stated that electrical data such as nominal power, open-circuit voltage, or short circuit voltage as well as mechanical data. Nominal power voltage

(V_{MPP}) (V) and nominal power current (I_{MPP}) (A) determines the nominal power (P_{MPP}) (Wp) of the module by Equation (2.1).

$$P = I\,x\,V \tag{2.1}$$

Open circuit voltage V_{OC} (V) is the maximum voltage of the module where the current is zero. In other words, the electrical current does not flow at this point. Short circuit current (I_{SC}) (A) is measured at the zero voltage in the module. It can be said that open-circuit voltage and short circuit current values are the limits of the module and no power is obtained beyond those limits. Solar panel efficiency is calculated by dividing the power value at maximum (P_{MAX}) by incident power (P_{IN}) [16]. Equation (2.2) shows the solar panel efficiency formula.

$$\eta = \frac{P_{MAX}}{P_{IN}} = \frac{FFV_{OC}I_{SC}}{P_{IN}} \tag{2.2}$$

FF is the fill factor and is calculated as in Equation (2.3). The fill factor is the maximum power values of voltage and current divided to open circuit voltage and short circuit current.

$$FF = \frac{V_{mp}I_{mp}}{V_{OC}I_{SC}} \tag{2.3}$$

P_{IN} values are calculated under standard test conditions (STC) where air mass of 1.5, vertical irradiance of 1000 W/m^2, and cell temperature of 25 ºC [16]. When it comes to comparing the panels, having the same STCs ratings does not necessarily mean that panels will produce the same amount of electricity. Panels may have different thermal losses or temperature coefficients and behave differently under low light conditions. Nominal operating cell temperature specifications (NOCT), on the other hand, reflect real world case output. Test values are obtained based on air mass of 1.5, irradiance 800 W/m^2, air temperature 20 ºC, and wind speed 1 m/s. It is important to emphasise that while STC is based on cell temperature, NOCT is air temperature. Nominal specifications can be defined as for modules and named as nominal operating module temperature (NOMT). In the process of the design phase, STCs values are used for sizing. Nominal operating values have a good source for comparing panels that have the same STC rating.

In the panel specification sheets, temperature ratings are listed. It is a scientific fact that electricity output is influenced by irradiance, temperature, and temperature associated with panel/cell cooling effect caused by wind speed. I_{SC}, V_{OC}, and P_{MPP} values are also defined by temperature correction coefficients. For example, nominal power drop in percentage per temperature change.

2.1.4 Grid Connection

Standalone or medium/large scale PV output power is connected to national or local grid networks under certain regulations. Typically, the DC-AC inverter output is connected to the AC circuit breaker to avoid overloads. Theoretical PV output power is always reduced due to losses such as module soiling, shading, DC losses, MPP mismatch error, inverter and AC losses. An electricity meter or an advanced analyser measures the electricity provided to the grid for correct billing and recording. Lastly, transformers readjust the alternating current circuit and PV production is distributed to the grid network.

2.2 Available Meteorological Variables and Measurement

The weather plays an important role in PV power production. Ideal conditions for PV production are receiving high irradiance, cold and windy weather. The sun releases solar radiation that a form of energy and roughly 1361 W/m^2 radiation hits the top of the atmosphere. 30% of this radiation returns to space and the rest reach the surface of the planet [16].

Weather stations are the perfect fit for measuring current weather parameters. Different sensors are capable of measuring different meteorological variables above a certain level of height from the ground. In this study which is subject to analysing PV output correlation to meteorological variables, the station only measures certain variables such as solar irradiation, air temperature, relative humidity, dew point temperature, and wind direction and speed. That is why only these variables are explained in detail.

2.2.1 Solar Irradiation

Irradiance can be defined as the amount of energy from the sun hitting a square meter and having a unit of W/m^2. Global radiation is taken into account for PV power output calculations. Moreover, total downwelling shortwave radiation from the sun includes ultraviolet, visible and infrared light [17]. A pyranometer measures short-wave radiation which is the radiation flux through a horizontal surface. Short-wave radiation has subcategories as downwelling and upwelling short-wave radiation. While downwelling radiation consists of direct solar beams and diffusive components, upwelling radiation only measures light reflection from the surface. That is why downwelling short-wave radiation is responsible for solar cell power production. Specifically, in PV prediction modelling, sun position inputs are possible inputs. Figure 2.7 shows how the sun position is identified with solar elevation angle, azimuth, and zenith.

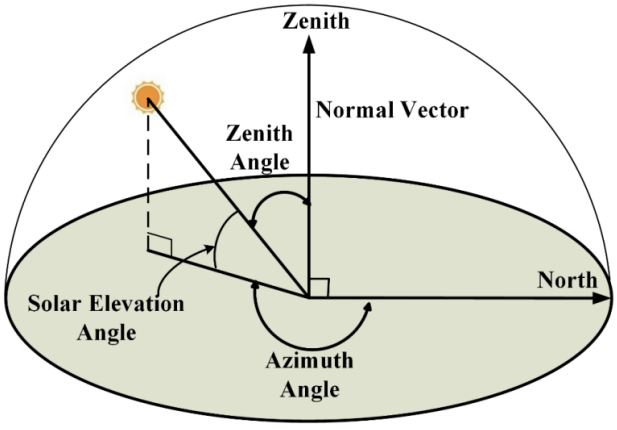


Figure 2.7: Sun position angles with respect to directions [19].

For measuring, CM11 or CM3 type from Kipp&Zonen pyranometers are used in Gjerpen station [18].

2.2.2 Wind Speed and Direction

Wind speed and direction are measured by an anemometer, either 2 m or 10 m above sea level. For wind direction measurement, mechanical parts of the wind vane operate at a 360º angle without stopping. A cup-anemometer-wind-vane pair is generally located at opposite ends of the horizontal bar to avoid a wind tunnel effect [17]. Measurement units are m/s and degree º. In the Gjerpen station, wind speed is the absolute value of the wind speed in the horizontal plane [18].

2.2.3 Air Temperature

Atmospheric temperature is measured by digital sensors. Different type of sensors has different sensitivity measurement. The output voltage is converted to degrees Celsius ºC.

2.2.4 Relative Humidity

Relative Humidity is simply a measure of water content in the air. Relative humidity is one way of measuring atmospheric humidity. The measurement is done by either a traditional psychrometer or thin-film polymers. Thin-film polymers absorb and desorb water throughout the relative humidity changes and the electric circuit is converted to relative humidity percentage [17]. Relative humidity heavily depends on atmospheric temperature and is sensitive to temperature changes. In rainy conditions, relative humidity reaches up to 100%.

2.2.5 Dew Point Temperature

The temperature below which water vapour in a volume of air at constant pressure condenses into liquid water is known as the dew point [20]. The point where air saturation occurs with moisture is called dew point temperature. Dew point temperature is affected by humidity which is also affected by atmospheric temperature. The measurement is done by either a dew point hygrometer or an equation that requires air temperature and humidity values.

2.3 Literature Review

There are two main approaches to estimate solar power production in the literature. While it is becoming popular to use machine learning methods to evaluate solar power production, there are also calculations based on correlation coefficients and proposing an equation for forecasting. When it comes to forecasting by using machine learning algorithms, artificial neural networks (ANN) and support vector machines (SVM) are produced reliable results under varying environmental conditions [21]. Studies on predicting solar PV output can be classified into two main groups; data handling and correlation analysis, model structure and machine learning algorithms. In the literature survey part, researcher's methodology and their findings will be presented under two main subtopics.

2.3.1 Data Handling and Correlation Analysis

Meteorological data and historical PV output data have to be handled differently as they are produced from separate sources. Meteorological data values vary daily and seasonally. That is why zero values, outliers and categorical information such as cloudiness require case specific

data handling. In addition, correlation analysis within the meteorological data holds valuable information for location base analysis. In contrast to meteorological data, PV output data has characteristic variations based on PV module location and inverter variations. Moreover, it is common that low light conditions, snow on module data, before sunrise and after sunshine data are filtered out from the PV output database to measure PV system performance.

N. Maitanova *et al.* [22] predict photovoltaic power based on publicly available weather variables. The study does not take into account solar irradiance values and tries to make a reliable prediction based on other meteorological data such as temperature, wind speed, humidity, precipitation, and cloudiness. In the contrast to other studies, this paper proposes a method to convert raw PV data to adjusted values based on clear-sky condition data and maximum PV power data as a new input to the algorithm. Hence, publicly available weather reports which do not include solar irradiance values can be used for PV power forecast by adjusting historical PV output power by the pvlib clear-sky program. As a result, it is concluded that prediction with solar irradiance values produces accurate outcomes whereas the model without solar irradiance values is still suitable for energy management systems for individual energy production purposes.

L. Hernández *et al.* [23] analyse weather variables and PV power production data. In the process of data pre-processing PCA method is used to remove outliers. Pearson's linear correlation coefficient method is employed to find correlation coefficients between weather variables and electrical power production. The study also calculates seasonal average weather variables and correlations to PV power production as an input for classification algorithms for further studies.

T. AlSkaif *et al.* [24] study 9 different meteorological variables in two different locations. Interdependency of variables is determined by correlation coefficients before moving forward to dimensionality reduction with PCA. PCA results vary in two cases such as some meteorological variables are less correlated to each other. The study concludes that reduced subspace estimation performs well in the linear support mechanism model. As a result, 4 meteorological variables generate similar results compared to 9 meteorological variables. While for one location, humidity, temperature, visibility, and wind speed are important meteorological variables; humidity, visibility, temperature and could cover are valuable for the second location.

There are also some studies for photovoltaic system evaluation for Norway. These studies take into account challenging environmental conditions such as low light and snow. M.B. Øgaard *et al.* [25] evaluate the performance of monitoring algorithms for photovoltaic systems in Norway. To evaluate snow cover on the PV modules, DC voltage variations were investigated. The study found that DC voltage variations increased during partial snow cover. By determining a threshold for DC voltage variations, partial and full snow cover data were removed. In addition, irradiance values below 50 W/m^2 exclude morning and evening effects. In another study, M.B. Øgaard *et al.* [26] investigates the effects that reduce the stability of PV monitoring at high latitude locations such as Norway. Filtering out PV output data by using following criteria is suggested for specific location: 50 W/m^2 < irradiance level < 200 W/m^2, snow depth on the ground < 0m. While below 50 W/m^2 irradiance level represents the low light condition, filtering out data above 200 W/m^2 removes outliers.

G. Kim *et al.* [27] implement the Pearson correlation method to investigate the relation between weather variables and PV power output. After finding correlation coefficients, different model equations are presented and each equation model evaluation is done by mean absolute percentage error (MAPE) and root mean square error (RMSE). It is concluded that humidity has an impact on the accuracy of power prediction where environmental conditions have a low ambient temperature, low irradiance, and high humidity.

2.3.2 Machine Learning Methods

S. Leva *et al.* [28] use artificial neural networks (ANN) to make 24h ahead forecast based on the weather forecast and historical power measurements. A clearness index is proposed based on cloud conditions and provided as input to the algorithm. As a result of the training of the ANN model with historical data, weather forecast data is used as input. One of the highlights is that pre-processing step of historical data has an influence on ANN method accuracy. That is why historical data were used for the training set. After model evaluation, it is proved that solar irradiance is highly correlated to forecast accuracy.

I. Jebli *et al.* [29] investigate four different machine learning methods. Linear regression (LR), random forests (RF), artificial neural networks (ANN) with different weights of the hidden and output nodes, and support vector mechanism (SVM). Two different locations were chosen for the study as Brazil and Morocco. Pearson coefficient analysis was conducted to determine the most relevant meteorological data. The study concluded that ANN produced accurate predictions for both historical and forecasted data. The nonlinearity handling capability of ANN has assumed the reason for leading to better forecasting results among other machine learning algorithms.

X. Wang *et al.* [30] used several machine learning methods to compare each other. Lasso regression, random forests, support vector regression model, and gradient boosting regression model produced promising results. Weather type classification and time correlation were proposed to tackle with overfitting and underfitting problems.

N. Maitanova *et al.* [22] preferred a more advanced machine learning method called Long-Short Term Memory (LSTM). LSTM method is a developed version of recurrent neural network (RNN) with taking into account how long the information should be kept in the layer. In addition, LSTM does a good job of handling time series. The study suggested an LSTM method that handles continuous data input. After the data normalization step, the architecture of the model consists of the following features; five input parameters in the first layer, depending on the data density, two hidden layers with 64 and 32 neurons. The model was trained for 100 epochs. To make better predictions against season change, the model was trained for both cold and hot days in the study. The paper concludes that model accuracy depends on training set size, LSTM network configuration and input features.

Similarly, other studies for solar power forecasting based on weather inputs uses different machine learning algorithms and evaluate results. M. Malvoni *et al.* [31] suggest a hybrid machine learning algorithm called Group Least Square Support Vector Mechanism, M. P. Almeida *et al.* [32] used Quantile Regression Forests to make hourly PV power output

prediction, and L. Li *et al.* [33] adopted Deep Belief Network (DBN) model to build a regression model and to make short-term PV power output forecasting.

2.4 Prediction Methods

In this chapter, it will be given information on the prediction methods used in this study. Linear regression and the ANN model with model evaluation methodology will be explained in this part.

2.4.1 Linear Regression

In prediction applications, linear regression calculates the weighted sum of input features. An intercept or bias term exists as a constant. In case of multiple features are fed into the regression model, it is called multiple linear regression. Training of linear regression and setting model parameters is the starting point of building a model. The training dataset is fed into the regression model and the algorithm learns how to best fits the training dataset. Scatter plots are useful to analyse the data to determine the strength of data relationship with other features. To evaluate model performance, Root Mean Square Error (RMSE) is used which is one of the most common performance indicators [34]. The association of the observed data and variables are evaluated by the correlation coefficient. Standardization or scaling has no big impact on the final performance. Scikitlearn uses Singular Value Decomposition (SVD) linear regression which decomposes the training set matrix into submatrices. The more complex dataset requires improved linear regression models such as Gradient Descent.

2.4.2 Artificial Neural Networks (ANN)

ANN is one of the most efficient methods in prediction applications. Deep learning algorithms are frequently custom made for a specific application [35]. Specifically, ANN methods are commonly used in forecasting studies where non-linearity exists in a database [21]. In general, a neural network consists of three layers that are called the input layer, hidden layer and output layer as it is seen in Figure 2.8. The input layer is the place where raw variables are stored and ready to feed into the network. The actual processing is done through hidden layers. The values are entered into a hidden node and are multiplied by weights. Each node in the network has some weights and a transfer function is responsible for calculating weighted sum of the inputs and the bias. The bias, b, is a scalar vector while the inputs, x, and the weights, w, are vectors.

Scaling of input raw data is important in this step because unscaled data can take large weights and makes the algorithm unstable and increase the error. After computing weighted sums of hidden nodes, the output is sent to the activation functions which deal with non-linearity in the dataset. There are different types of activation functions. One popular function is the rectified linear unit function (RELU). In regression applications, linear activation functions are a good choice. As it is seen in Figure 2.9 that RELU function takes 0 if the value is negative, otherwise the real value is returned. As a result, the activation function decides whether the hidden node should be activated or not. The key limitation of the RELU transfer function is that values from transfer function flow to activation function are negative in the case of large weight updates. Therefore, the output of the activation function will forever be 0 which is called a dry RELU.

Hidden layers are connected to an output layer which represents the prediction of a variable. The flow of variables is from inputs to outputs, so this architecture is called a feedforward neural network. The algorithm first makes a prediction and visits each layer in reverse, and then calculates the error contributed by each connection. In the final step, the algorithm adjusts the connection weights to decrease the error [34]. The loss function which is used during training, is commonly the mean squared error. However, if the training set includes loads of outliers, mean absolute error might be used, instead.

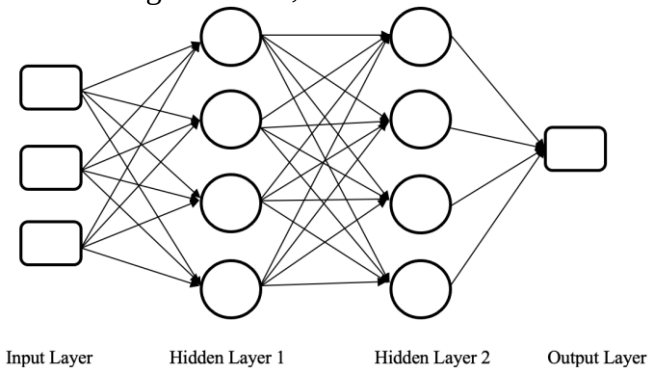


Figure 2.8: An ANN network architecture.

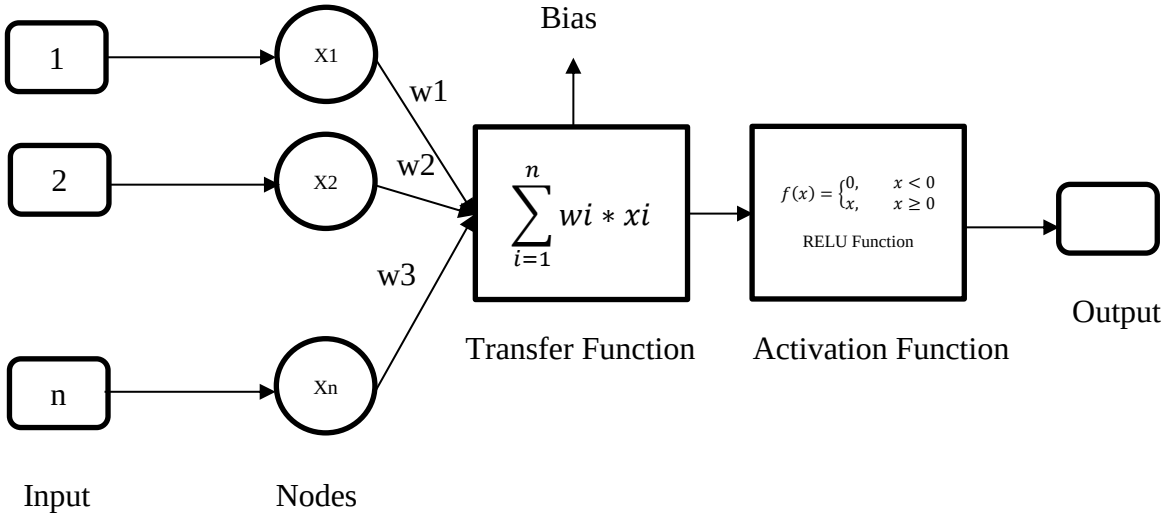


Figure 2.9: ANN network nodes connection with functions.

Once the model is created, the model should be compiled and specified loss function in addition to the optimizer to use. Later, the model is called by the fit method where X_train and y_train sets are introduced. The number of epochs and validation functions are also described in this part. There are other hyperparameters in a neural network such as batch size, learning rate and the number of iterations. Batch size has a big impact on model performance and training time. In addition, the batch size is related to GPUs which process the model efficiently. A large batch can be used but the limit is where training instabilities start. The negative effect of a large batch size can be compensated by adjusting the learning rate. Learning rate determines the updates of weights on the training set. In Keras, the default number for the learning rate is 0.001 [36]. For example, for a given 1000 datasets with 5 batch sizes and 30 epochs, it returns 200 batches in total with 5 samples. The model weights are updated after each batch of 5 samples. Thus, one epoch consists of 200 batches. Each epoch goes through the whole dataset, so 30 epochs go through the dataset 1000 times. That is a total of 200,000 batches for the whole dataset.

Keras measures the loss and training time per sample including accuracy for both the training set and the validation set [36]. If the training loss decreases, it means that the model performs well.

Overfitting is one of the common problems in ANN algorithms. One way to analyse overfitting is to evaluate the performance of the training set on the test set. If training set performance is much higher than on the test set, there is a possibility that the model is overfitting on the training set. The good fitting can be described as training and test loss plot decreases to a point of stability until to reach a small gap between the plots. Further training will likely result in overfitting, again.

In contrast to good fitting, the model may have an unrepresentative training set. It means that the model training set does not provide enough information to the model. Unrepresentative data results in loss curve as while training loss decreases, validation loss stops decreasing and stays linear so the gap between the plots increases. Besides, if the validation set is not representative, then the plot becomes noisy on the validation curve. In some cases, the model predicts the test dataset easier. In the loss curve, it will be seen as test loss significantly low compared to training loss.

2.4.3 Model Performance Evaluation

The accuracy of the model is critical and the model should produce reliable results. In the event of forecasting, it is expected from a model that prediction accuracy should be above a certain level. In addition, using similar model performance evaluation criteria makes models comparable with other models. The commonly used model evaluation models are Mean Square Error (MSE), Root Mean Square Error (RMSE), Normalized Root Mean Square Error (nRMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and variance R^2. Calculation methods are shown in Equations (2.4), (2.5), (2.6), (2.7), and (2.8), respectively. These models are available in scikit-learn under the regression metrics functions.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{predicted} - y_{test})^2 \tag{2.4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{predicted} - y_{test})^2} \tag{2.5}$$

$$nRMSE\ \% = \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{predicted} - y_{test})^2)}\right) x100/y_{test_{max}} \tag{2.6}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_{predicted} - y_{test}\right| \tag{2.7}$$

$$MAPE\ \% = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_{predicted} - y_{test}\right|}{y_{test}}\ x100\% \tag{2.8}$$

In these equations, $y_{predicted}$ is the output class and y_{test} is the input value. While N represents the total number of data, $y_{test,\ max}$ describes the maximum value of power values. In addition,

R^2 (coefficient of determination) or regression coefficient function takes 1.0 which is the best possible prediction model.

Another way of evaluating model performance is learning curves. Figure 2.10 describes the relation between underfitting and overfitting with error evolution. Both linear regression and ANNs benefit from learning curves to evaluate the model performance.

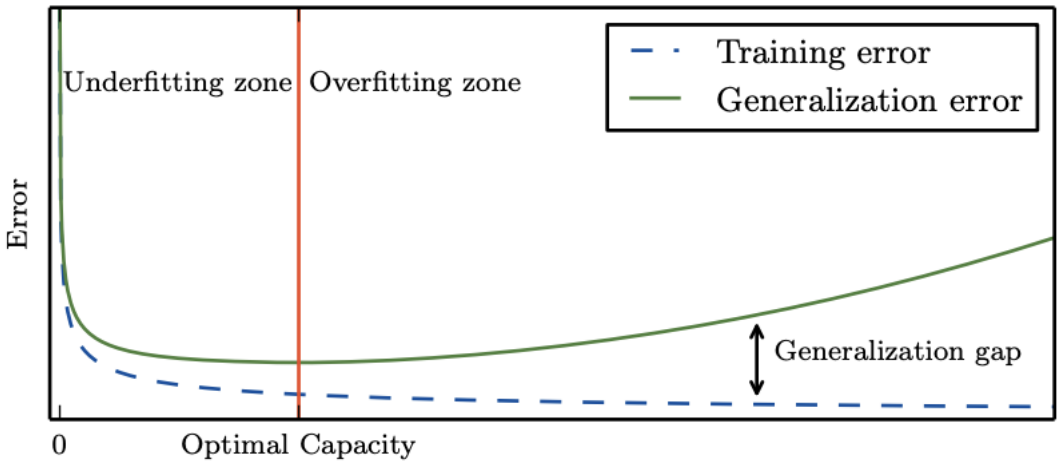


Figure 2.10: A typical learning curve [37].

When training and generalization/validation/test error is both high, the model underfits. The more model learns, the more training error decreases. In overfitting zones, the generalization gap becomes higher. Furthermore, the main issue with overfitting is that the model loses its ability to make a good prediction since it is too much suited for training data. Test error decreases and to some degree flattens out, then it begins increasing again. In addition, the model does not learn from training data if the training error/loss is a flat line or noisy. Good fit occurs where test and training loss decrease together and it reaches stability with a small gap between each plot at some point.

3 Methodology

3.1 PV Plant Layout and Meteorological Station

In this part, extensive information will be given about the PV power plant such as PV modules, layout, and specifications. The second subsection will include information on the meteorological station.

3.1.1 PV Plant Specifications

The PV plant was built on the roof of the football stadium of Odds Ballklubb, Skagerak Arena and is located in Skien, Vestfold and Telemark County, Norway in 2019. The solar modules with a battery system are installed for storing electricity in addition to supplying power for internal usage in the stadium at nights and to the national grid. Figure 3.1 shows the stadium and solar modules on top of the roofs. Modules are only installed in the South, West and East direction. There are no PV modules on the north side of the roof. Shading is not a question in this plant since all 3 directions are in the open environment. Table 3.1 describes the plant's overall specifications.

Figure 3.1: Skagerak Arena stadium layout.

The detailed module specification is given in Appendix B. REC Twinpeak 2 series modules are being used in the plant with two types of nominal power output 295 and 300 Wp. 300 Wp panel types are only used on the south direction rooftop.

Table 3.1: PV plant specifications.

Installed power PV (kWp)	840
Area (m^2)	5330
Modules (pcs)	3230
Inverter power (kW)	675
Production in a normal year (MWh)	660
Specific Performance in a normal year (kWh/kWp.year)	786
Energy storage battery capacity (kWh)	1000

Table 3.2 describes solar module specifications. Solar cell type is 120 half-cut multi-crystalline PERC cells that information was given in the theory part. Two different nominal power types of modules are used in the plant. One has 295 Wp and other type has 300 Wp.

Table 3.2: Solar module selected features.

Nominal Power - PMPP (Wp)	295	300
Nominal Power Voltage - VMPP (V)	32.3	32.5
Open Circuit Voltage - VOC (V)	39.0	39.2
Panel Efficiency (%)	17.7	18.0

Panel efficiency values are measured at standard test conditions which are air mass at AM 1.5, irradiance at 1000 W/m², and temperature at 25°C.

Even though roof slopes are the same at 8º, azimuth values vary for different layouts. Table 3.3 shows descriptive information on slope and azimuth values for all directions. Surface azimuth input data is vital due to the PV plant layout. Surface azimuth values are explained as panel azimuth from the north which means the azimuth convention is defined as degrees east of north. The built algorithm in python, pvlib, assumes north as 0 degrees. North takes 0 degrees, south 180 degrees, east 90 degrees and west 270 degrees. However, the PV plant document accepts the south direction as 0 degrees and values take a negative sign from the north, east, and south directions.

Table 3.3: Module slope, azimuth angels and area with respect to directions.

	Slope	Azimuth	Area (m^2)
South Tribune	8	-20	1425
West Tribune	8	70	2146
East Tribune	8	-110	2146
North Tribune	8	160	-

That is why PV plant document values were adjusted as 0 degrees for the north. By determining east direction as - 90º and the west direction as + 90º, the azimuth values take negative values in the east and south directions. Figure 3.2 describes the layout of tribunes based on azimuth degrees. The adjusted azimuth variables based on 0-degree north direction are shown on the right-hand side.

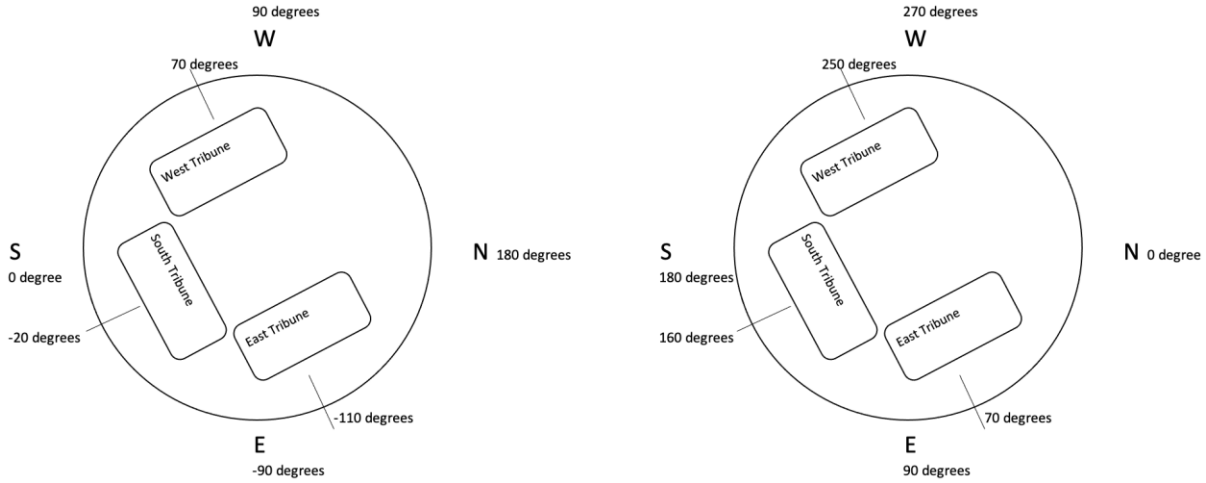


Figure 3.2: The original azimuth angles with layout (left) and adjusted azimuth angles (right).

Table 3.4 shows an updated version of Table 3.3 based on 0-degree north direction. These values were used as input in the get_total_irradiance function.

Table 3.4: Azimuth degrees based on north direction.

	Slope	Azimuth (0 degrees South)	Azimuth (0 degrees North)	Area (m^2)
South Tribune	8	-20	160	1425
West Tribune	8	70	250	2146
East Tribune	8	-110	70	2146

Table 3.5 describes the inverter's selected features. The detailed inverter information is given in Appendix B.

Table 3.5: AC/DC Inverter Specification.

Absolute maximum DC input voltage ($V_{max,abs}$)	1000 V
Rated DC input voltage (V_{dcr})	620 V
Rated DC input power (P_{dcr})	102 000 W
Number of independent MPPT	6
Maximum DC input current for each MPPT (I_{dcmax})	36 A
Maximum AC output power (P_{acmax} @cosφ=1)	100 000 W
Maximum efficiency (η_{max})	98.4%

Since the available power output is in AC which is after the inverter, an inverter connection was also investigated. Each inverter serves a different number of module strings. For example, Inverters 1-2-3 are connected to east direction panels. While 410 panels are linked to inverter 1 with 6 MPPT, 400 panels are connected to inverter 2 with 6 MPPT, and 210 panels are bound to inverter 3 with 3 MPPT. East and west directions are identical in terms of inverters, connections and the number of panels. 2 inverters serve to south direction with 444 and 216 panels for inverters 7 and 8, respectively. Figure 3.3 illustrates the panel layout based on directions.

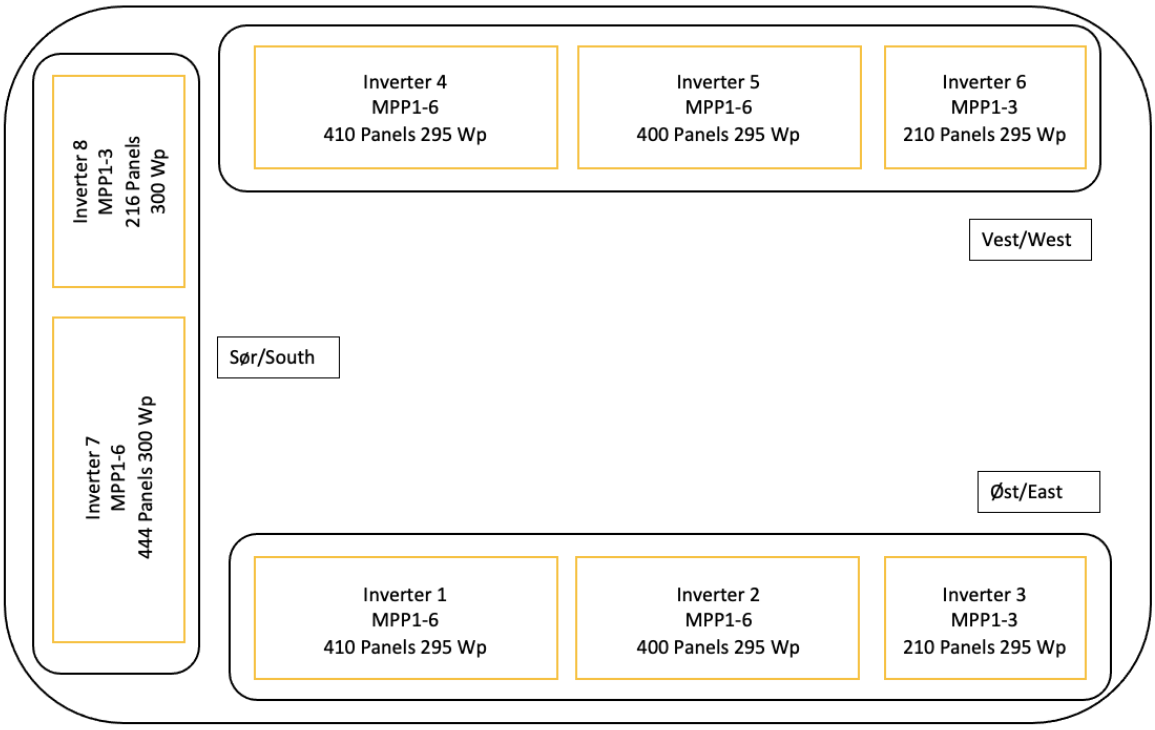


Figure 3.3: An illustration for PV plant layout inverter – module connections with respect to direction.

String connections are different for each group of inverters. Figure 3.4 shows a string connection for inverter 4 on the west side. For example, the inverter 4 has 410 panels but 20 strings in total. Each string has either 20 or 21 panels. Since inverter output AC power is analysed in this study, there will be no further investigation on detailed string connections.

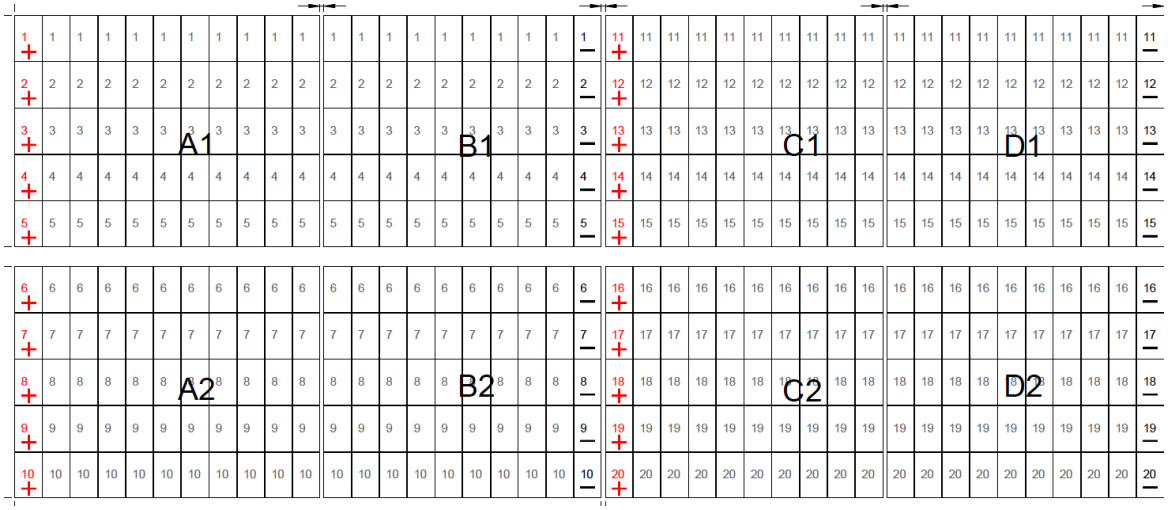


Figure 3.4: String connection example for modules that are connected to Inverter 4.

In general, 4 strings are connected to one MPPT, and some MPPT slots are empty due to matching the total number of string numbers. Each inverter has 5 or 6 MPPTs with DC/DC type. To sum up, module string connections to the inverter with MPPTs are illustrated in Figure 3.5.

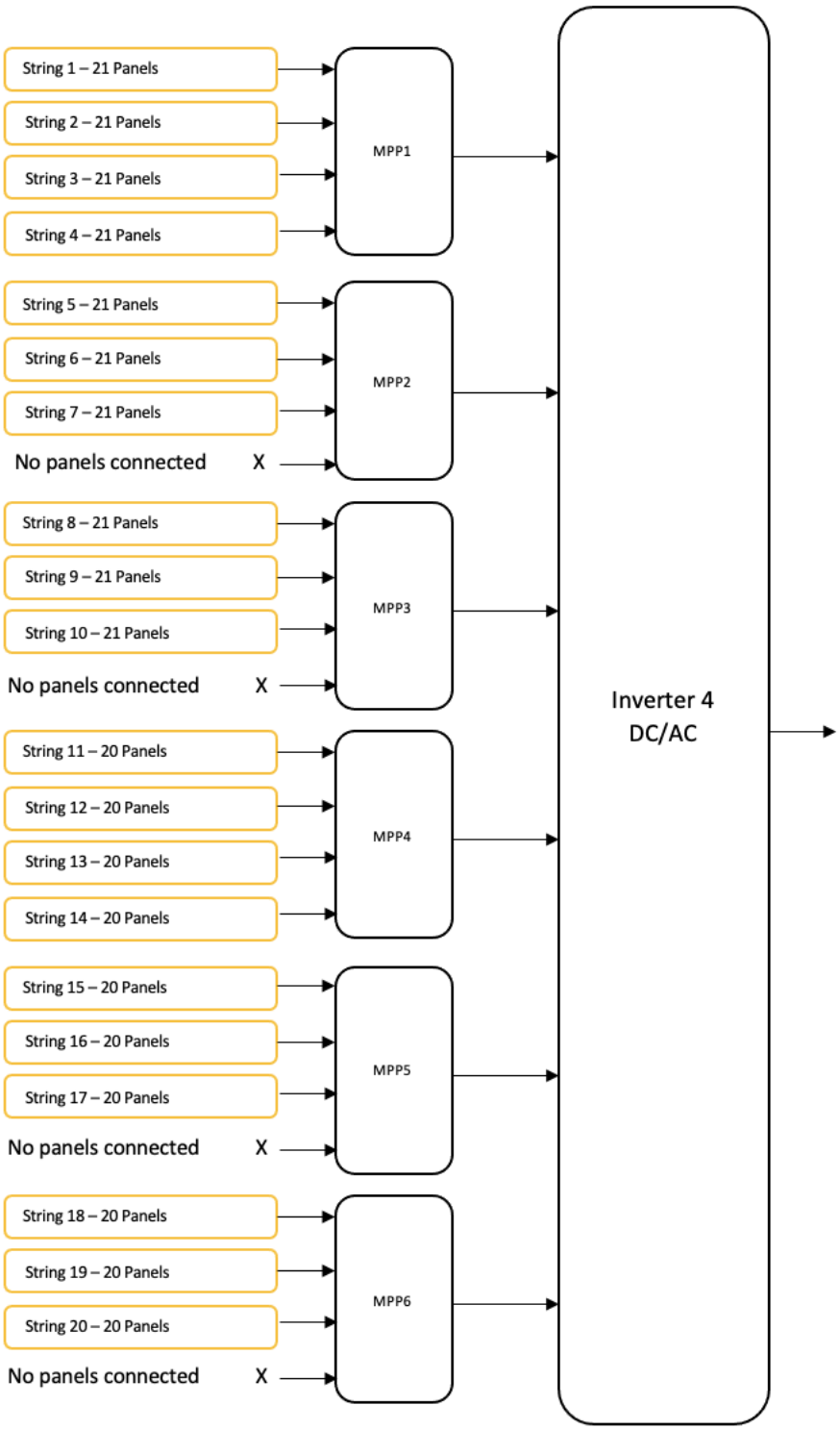


Figure 3.5: Module string connections to MPPs and the inverter.

3.1.2 Meteorological Station

Gjerpen meteorological station is positioned approximately 1.87 km further from Skagerak Arena where the PV plant is located. In Figure 3.6, the station location is pointed out by a red pin. The station's detailed information about latitude, longitude, and altitude is given in Table 3.6. It is possible to say that the station measured meteorological variable represent perfectly the environment around the PV plant. Station variables were obtained by the station code from the free access meteorologisk institutt frost API system. Historical weather and climate data with quality control parameters were accessed by python codes.

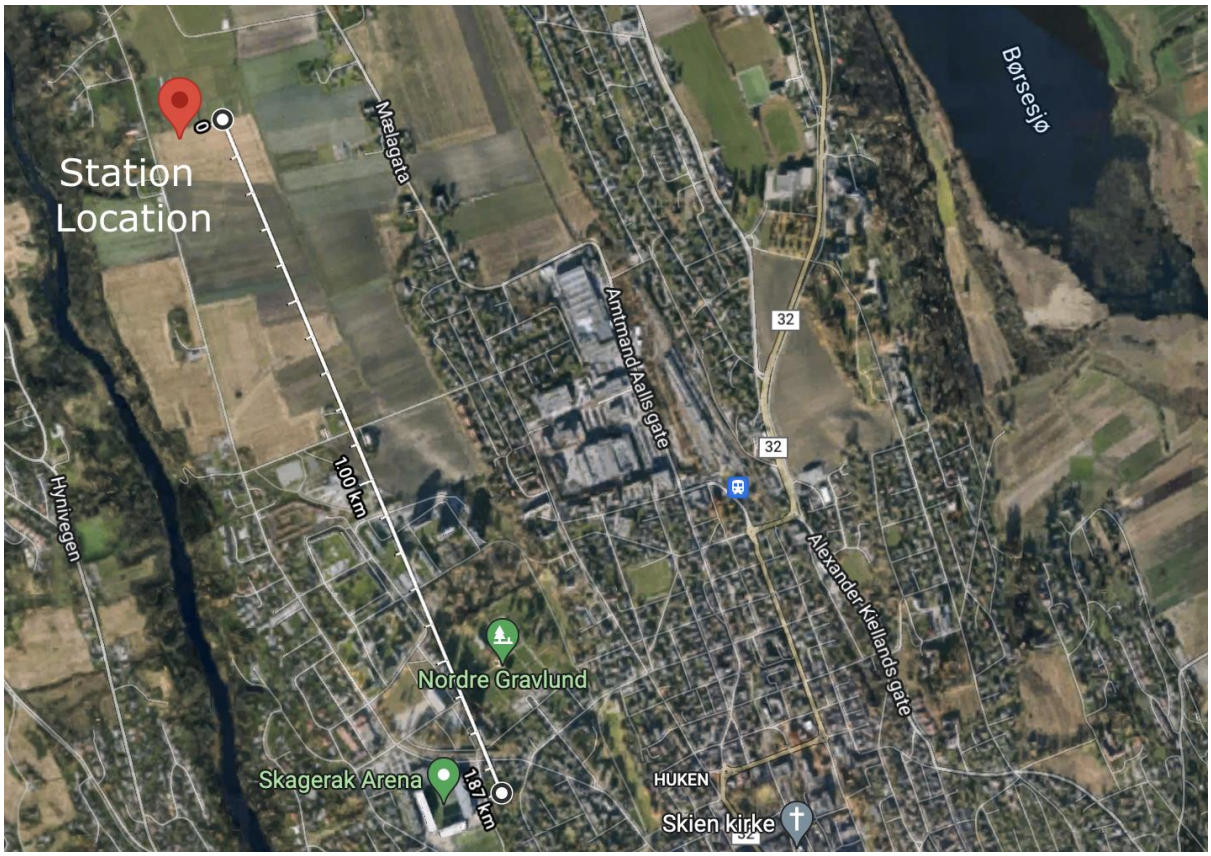


Figure 3.6: MET station and stadium location.

Table 3.6: MET Station information.

Station Code in the MET internal system	SN30330
Owner	NIBIO
Latitude	59.22684º
Longitude	9.57805º
Altitude	41 m

3.2 Data Gathering and Data Pre-processing

Meteorological data was retrieved from frost database, PV data was delivered from Lede Energi, and clear sky including solar position data was accessed through pvlib-python package. The general methodology for meteorological data and PV data dealing flow diagram is shown in Figure 3.7.

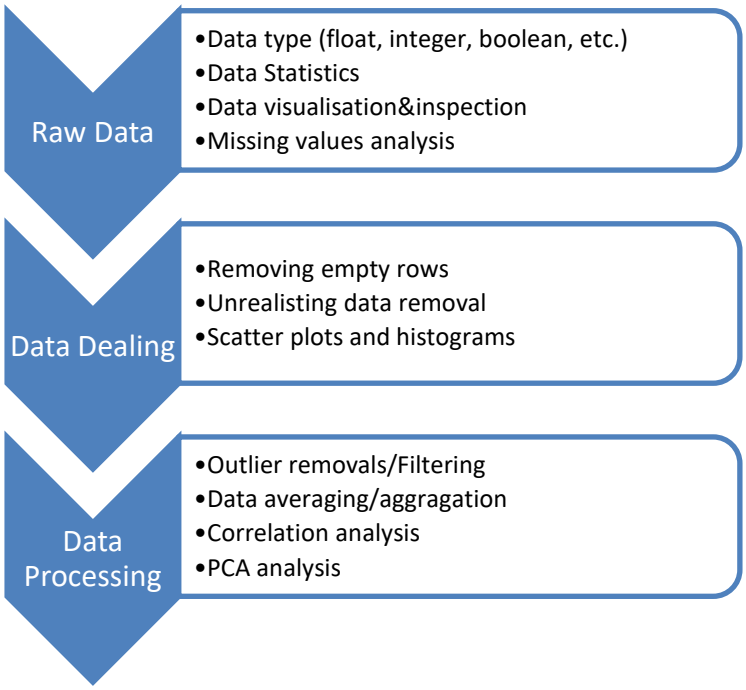


Figure 3.7: Data processing flow diagram.

The sampling rate and time interval vary a lot for different data sets. For example, the meteorological data sampling rate is 1 hour, and the pvlib database is 1 min whereas PV data is 10 mins. Therefore, the limitation for further analysis was meteorological data. The other variables had to be averaged to match with meteorological data. In addition, the resampling function in pandas was used to fit all sampling rates in an equal time interval. Measurement can represent separate time intervals. For example, meteorological data 10:00 data describes the observation from 09:01 to 10:00, however, PV data after resampling function 09:00 data represents from 09:00 to 09:59. To avoid this conflict, time shifting was applied. Besides, Pvlib data is available for 1 min sampling rate and 10:00 represents the observation at exactly 10:00. It was needed to a value to symbolise from 09:00 to 10:00 and this was achieved by taking the value at 09:30. Table 3.7 shows the sampling rate and time with changes in the time interval. Pvlib Data consists of Solar Position Elements (zenith, azimuth, elevation, equation_of_time), ClearSky (ghi, dni, dhi), POA (poa_global), erbs (dni_generated, dhi_generated).

Table 3.7: Time resolution for each data set.

	Meteorological Data	PV Data	Pvlib Data
Sampling Rate	1 hour	10 mins	1 min
Measurement	09:01-10:00 as 10:00	Resampling function assingns 09:00-09:59 as 09:00	10:00 as 10:00
Time Change	No change	PV Data shifted 1 hour further so 10:00 represents 09:00 - 09:50	subtracted 30 mins so 10:00 represents 09:30

The presence of daylight-saving time in the dataset leads to a time lag between variables. To avoid any confusion in data, all time dependent data is gathered based on Universal Time Coordinated (UTC). Hence, the data does not affect by daylight saving changes throughout the year. That is why it is important to keep in mind adding +01:00 or +02:00 hours depending on the selected day when comparing the results with local time-based values for Norway.

3.2.1 Meteorological Data

Meteorological data was accessed through a free source meteorology institute frost database by a python script. Measuring interval is available in the frost database based on hour, day and month. Hourly values are the average value for the first hour after the stated measurement time as it was explained at the beginning of the chapter. Available variables are listed in Table 3.8. However, some critical parameters for solar power output are not recorded by the station such as precipitation, cloudiness and air mass. Nearby stations also have no record for these variables. Mean hourly values of variables are used for the study.

Table 3.8: Available meteorological variables.

Variable name and unit	**Variable name in the frost database**	**Sampling rate and year periods**
Average Temperature at 2m (°C)	mean(air temperature PT1H)	Hourly mean average 2018-2021
Dew Point Temperature at 2m (°C)	dew_point_temperature	Hourly mean average 2018-2021
Relative Humidity at 2m (%)	mean(relative_humidity PT1H)	Hourly mean average 2018-2021
Wind Speed at 2m (m/s)	wind_speed	Hourly mean average 2018-2021
Wind Direction at 2m (°)	mean(wind_from_direction PT1H)	Hourly mean average From May 2019 to 2021
Global Horizontal Radiation (W/m^2)	mean (surface_downwelling_shortwave_flux_in_air PT1H)	Hourly mean average 2018-2021

Figure 3.8 shows raw meteorological data from January 2018 to December 2021. There are some periods where the station has no record due to being out of order. Especially, there is huge data loss for the period from January 2020 to April 2020 for irradiance and relative humidity values. Furthermore, the data loss exists in June 2018 for all variables except dew point temperature. It is evident that there are also some outliers. There are some peak values that disassociate from the pattern in wind speed. Moreover, air and dew point temperatures roughly are in harmony with irradiance. It is difficult to make a comment on relative humidity but wind speed is relatively higher during summertime. Wind direction data will be presented in a different chapter as further processing was applied to the data.

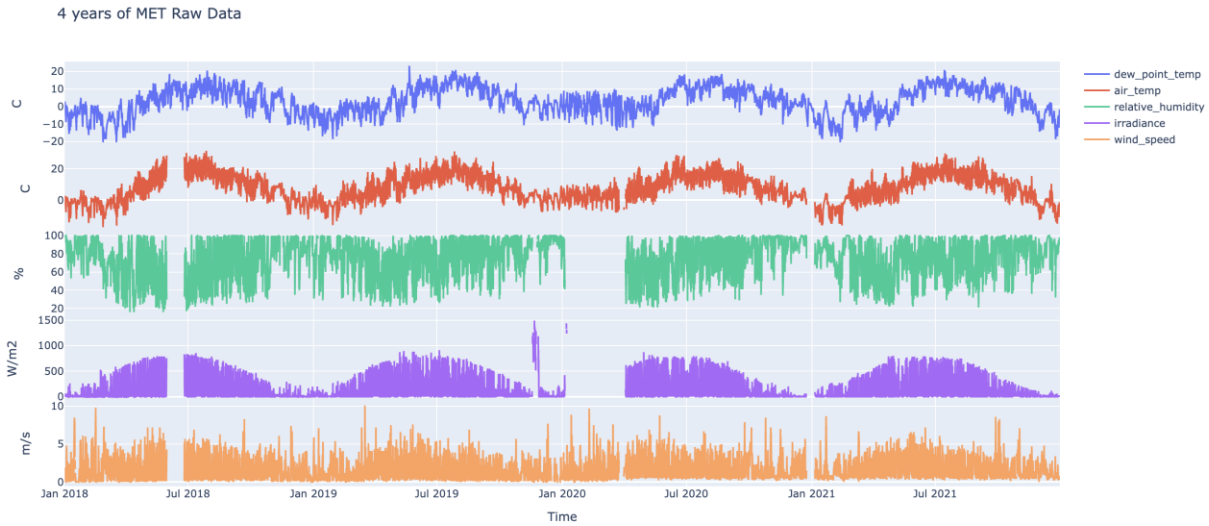


Figure 3.8: Meteorological variables raw data against time.

3.2.1.1 Wind Direction Category

Wind direction is only available from May 2020 and it is recorded in 360 degrees. Using the raw wind direction in degrees, which stores 360 different values, causes data uncertainty and reduces relationships within other variables. That is why this data was converted into 4 main categories as north, east, west, and south. While Table 3.9 shows directions and corresponding values in degrees, the figure illustrates a compass with directions.

Table 3.9: Wind directions and corresponding degrees.

Direction	Explanation	Degrees
N	North	349-011
NNE	North-Northeast	012-033
NE	Northeast	034-056
ENE	East-Northeast	057-078
E	East	079-101
ESE	East-Southeast	102-123
SE	Southeast	124-146
SSE	South-Southeast	147-168
S	South	169-191
SSW	South-Southwest	192-213

SW	Southwest	214-236
WSW	West-Southwest	237-258
W	West	259-281
WNW	West-Northwest	282-303
NW	Northwest	304-326
NNW	North-Northwest	327-348

For example, north direction limits are assumed from northeast to northwest. Northeast limits are 034º-056º while northwest limits are 304º-326º. Thus, north direction limits were determined as 327º-056º. Likewise, the same approach was applied to calculate other directions. Table 3.10 shows the 4 main direction categories of the table above. Corresponding degree numbers are also stated.

Table 3.10: Categorical wind direction data with adjusted degrees.

Category	Direction	Degrees
1	North	327-56
2	East	56-146
3	South	146-236
4	West	236-326

3.2.2 Pvlib Data

Pvlib is a comprehensive package for simulating PV energy applications. It is available on python and provides reliable and open applications for PV systems with libraries and functions. This chapter consists of two subchapters as clear sky data and plane of array irradiance data.

3.2.2.1 Clear Sky Data

Free clear sky data is available on the internet in various platforms and pvlib library on python. In addition, there are different models of calculating clear sky global horizontal irradiance (ghi), direct horizontal irradiance (dhi), and direct normal irradiance (dni) values. In this study, the pvlib library is used and different models were compared to pick the best model. The best model can be defined in a way that covers most of the measured irradiance values and produces clear sky days successfully by using a clear sky day detection algorithm. There are three common clear sky calculation methods in the pvlib library. These are Perez-Ineichen, Haurwitz and simplified Solis method. Each method is differentiated by various weather parameters. All equations use the linke turbidity factor as default or user input. Moreover, factor values are influenced by atmospheric absorption and scattering of the solar radiation. Linke turbidity is a function of aerosol particles and water vapour in the atmosphere. Aerosol particles are relative to the dry and clean atmosphere and absorption by the water vapour changes the optical thickness of the atmosphere. Thus, a larger linke turbidity factor refers to a reduction in the radiation by the clear sky atmosphere [38].

Three methods were chosen and compared. These are Perez-Ineichen, Haurwitz, and Solis method. Method short explanations are given below.

Ineichen: Uses default climatological turbidity values, and produces good results with fewer input requirements [39].
Haurwitz: The model has the best performance in terms of average monthly error among models which require only the zenith angle. The relationship between cloudiness, air mass and cloud density are the parameters in the equation [39].
Solis: The Simple Solis clear sky model is based on RTM and the Lambert–Beer relation to estimating irradiance. The model is a simplified version in order to reduce the computational requirements. The model requires predictable water vapour and aerosol optical depth (AOD700) as the main input parameters [39].

Figure 3.9 shows Perez-Ineichen, Haurwitz, Solis method global horizontal values (ghi) comparison with measured ghi values for 2021. The red line which represents the Haurwitz method has the highest average irradiance values and the Solis method is the second highest. Perez-Ineichen values are not visible in the graph, so a closer look for a day was plotted.

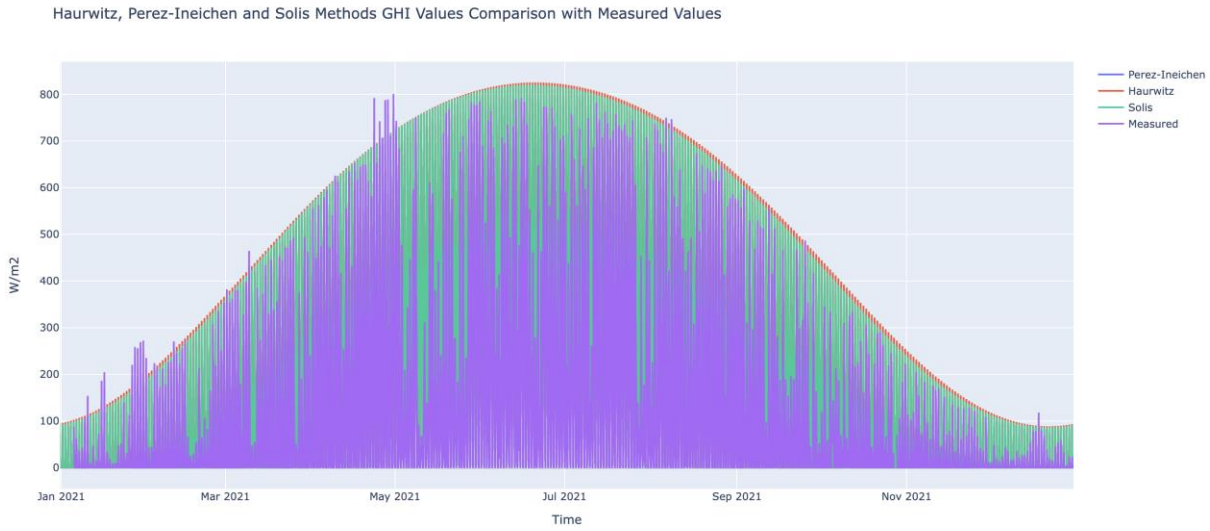


Figure 3.9: Perez-Ineichen, Haurwitz, Solis methods and measured irradiance values against time.

Figure 3.10 shows clear sky day calculation method results for a clear sky day on 24th June 2021. Perez-Ineichen method values are lower than Haurwitz and Solis methods.

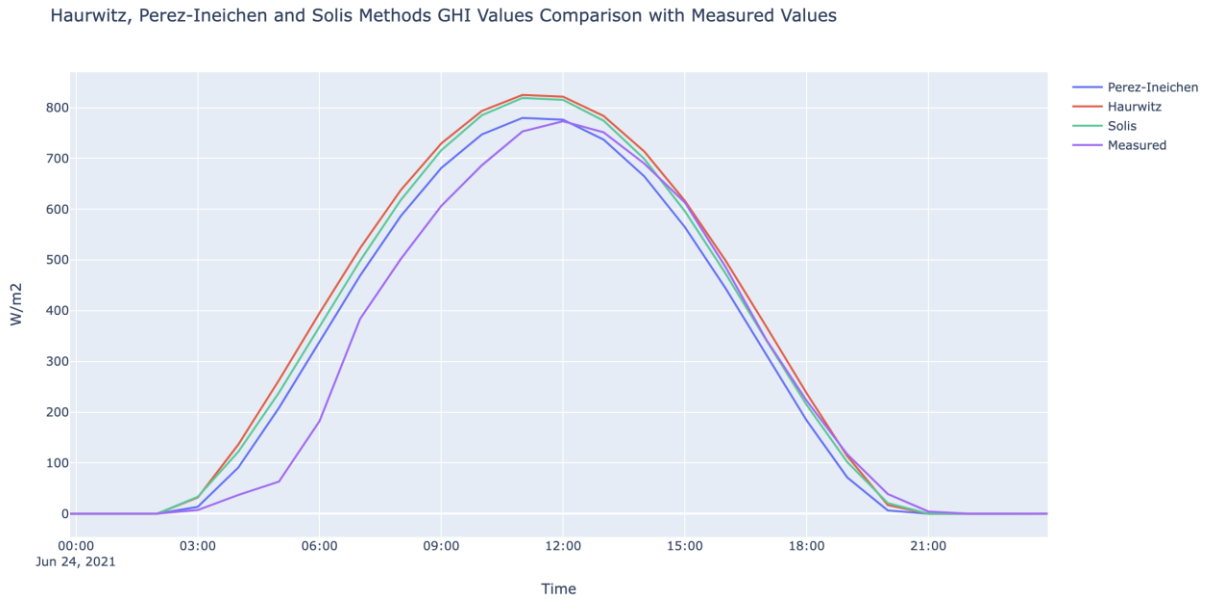


Figure 3.10: Clearsky methods and measured irradiance values for a selected clear sky day.

There are some days that measured actual irradiance values exceed the theoretical clear sky irradiance values. These time periods were dug deeper. One might think that these values might be measurement errors or indicate air parameters change. It is concluded that these peak values are not measurement errors because PV production values also peaked in the same period. Figure 3.11 shows a specific time period where measured irradiance values exceed the calculated values with emphasised the time interval red lines. It is known that changes in meteorological variables and atmospheric conditions such as water content, albedo, and aerosol have an effect on measured irradiance values. That is why it was decided to compare meteorological variables for that specific time interval.

Figure 3.11: The period of exceeding calculated clear sky values of measured irradiance.

Figure 3.12 represents the corresponding time interval of exceeded irradiance values with emphasising two red lines. The examination was done only with available meteorological variables. The start and end date of plotting were kept long to help to compare how

meteorological conditions changed before and after the time interval. When two figures are compared, there are two variables that have different trends than before and after. Wind speed was quite low and relative humidity was high even though the sun shows up. Water content in the atmosphere was high and wind speed was low. There is a possibility that these findings might have affected measured irradiance values which resulted in high values than calculated values. As a result, the measurement device recorded high numbers and it was assumed that numbers reflect the real situation.

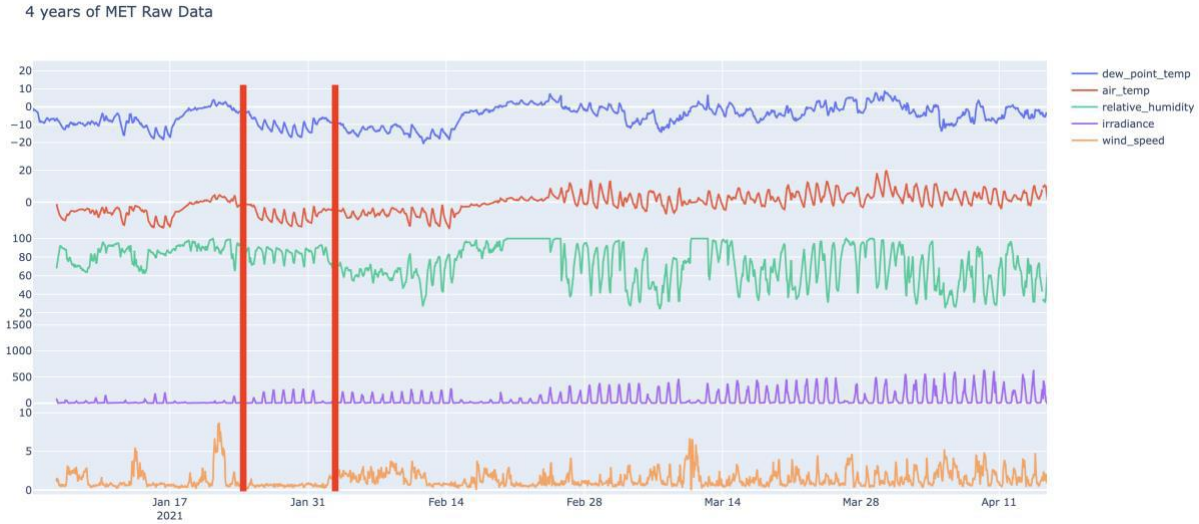


Figure 3.12: Meteorological variables for exceeding time interval.

To evaluate each model and compare it with measured irradiance values, correlation analysis was performed. By doing so, it was aimed to select a suitable model for location specific irradiance values. A consecutive clear sky data was captured between 22nd and 25th July 2021. Correlation analysis was done within these dates by using global horizontal irradiance values (ghi) and the results are presented in Table 3.11.

Table 3.11: Correlation analysis results for clear sky and measured irradiance.

	Haurwitz_ghi	Solis_ghi	Ineichen_ghi	Measured_ghi
Haurwitz_ghi	1.0000	0.9996	0.9983	0.9847
Solis_ghi	0.9996	1.0000	0.9994	0.9865
Ineichen_ghi	0.9983	0.9994	1.0000	0.9873
Measured_ghi	0.9847	0.9865	0.9873	1.0000

The highest correlation is obtained by Perez-Ineichen method. Figure 3.13 shows Perez-Ineichen and measured irradiance values graph for the year of 2021.

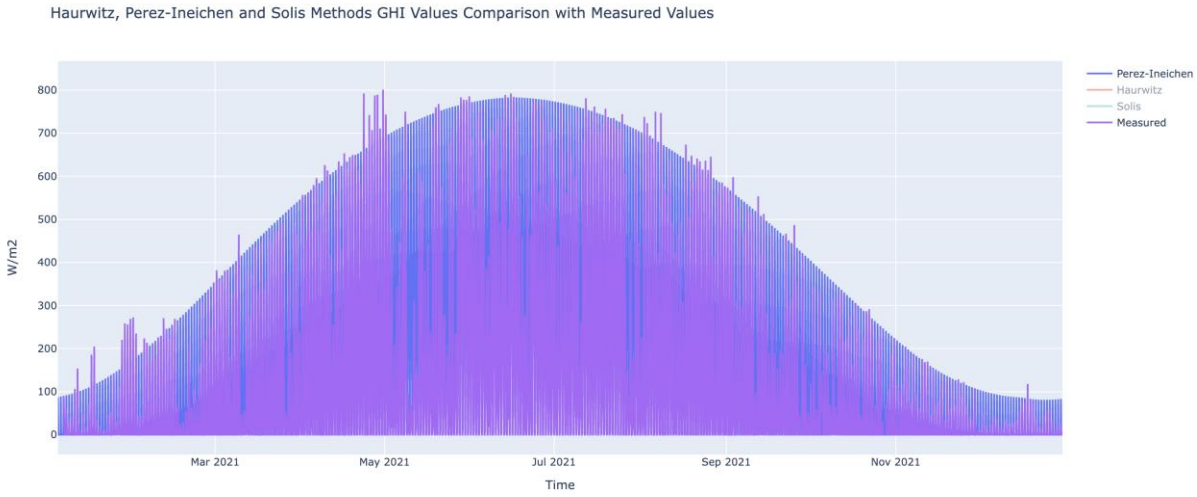


Figure 3.13: Perez-Ineichen and measured irradiance values graph for the year of 2021.

However, Ineichen- Perez clear sky data comparison with actual irradiance values throughout the year of 2021 does not overlap well. There are some peak measured irradiance values. The comments on this matter are available under the discussion section. For further analysis with PV data, Ineichen- Perez values were chosen.

Figure 3.14 shows irradiance components of Ineichen- Perez model on a selected day. Direct normal irradiance (dni), direct horizontal irradiance (dhi), and global horizontal irradiance (ghi) values were plotted with default algorithm settings.

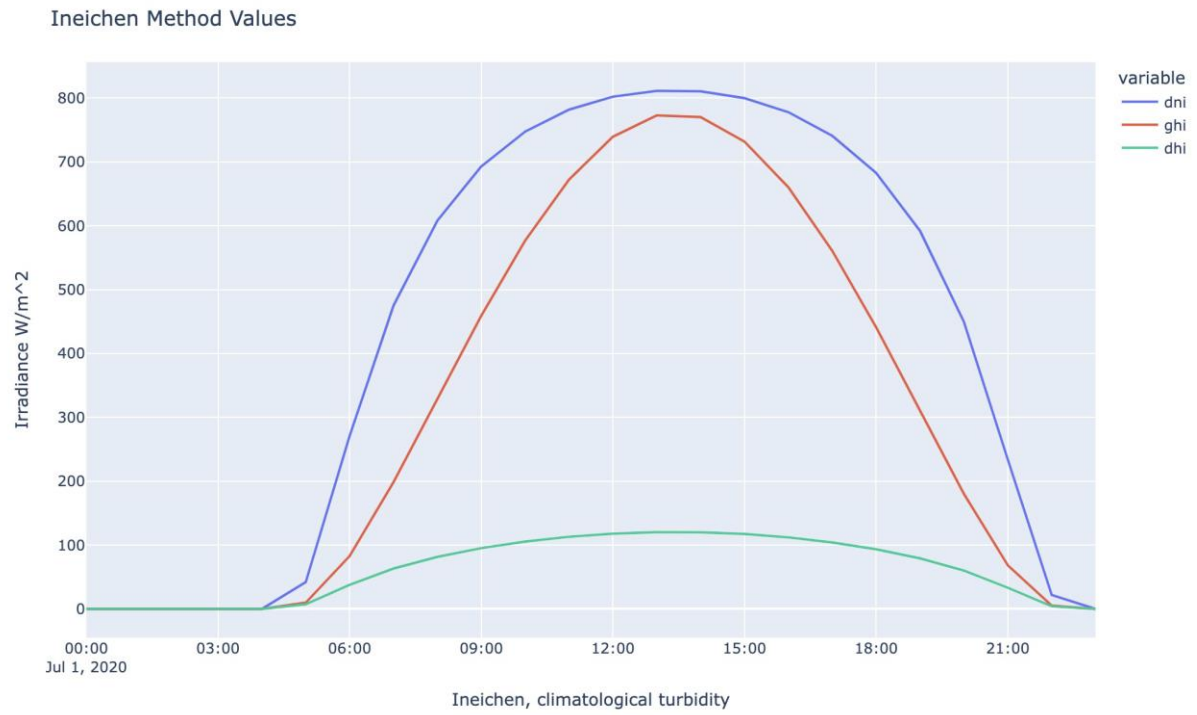


Figure 3.14: Irradiance components of Ineichen- Perez model on a selected day

Detection of clear sky days with the pvlib algorithm is challenging because it requires minute based data. However, measured irradiance values are on an hourly based. There are other numerical ways of detecting of clear sky in literature which also require minute-based data. In

order to overcome this problem, hourly base irradiance values were converted to minute-based data. The polynomial interpolation method was used with a power factor of 3. detect_clearsky algorithm in the pvlib library produces Boolean results. These Boolean results were converted to 0 and 1 integer values which 0 refers to cloudy and 1 refers to clear sky. The produced minute-based results were again transformed into hourly based outputs. Changing the time frame twice, even though the approach and method are correct, produces some errors. However, most of the errors were removed in the process of resampling to hourly data by the forward filling method.

Figure 3.15 demonstrates the clear sky detection algorithm result for a clear sky and a cloudy day with measured and clear sky irradiance included. The data in the figure is minute based. False values refer to cloudy time and true values are clear sky time. As it is seen from the graph that clear sky was detected on 24th June 2021 with cloud detection in the early morning which measured irradiance values proved. The rest of the day was cloud clear. The next day measured irradiance values were below clear sky irradiance values with some variation. Before sunset, a clear sky was detected. The clear sky line was interrupted on 26th June 2021 before sunset even though measured irradiance and clear sky data overlaps. The algorithm detects wrong due to resampling. This problem was eliminated when the data was inversed hourly based. Another reason for the wrong detection is that while measured hourly based irradiance values represent measurement from 09:00 – 10:00 as 10:00. Time shifting correction was included in the algorithm to reduce wrong detection errors.

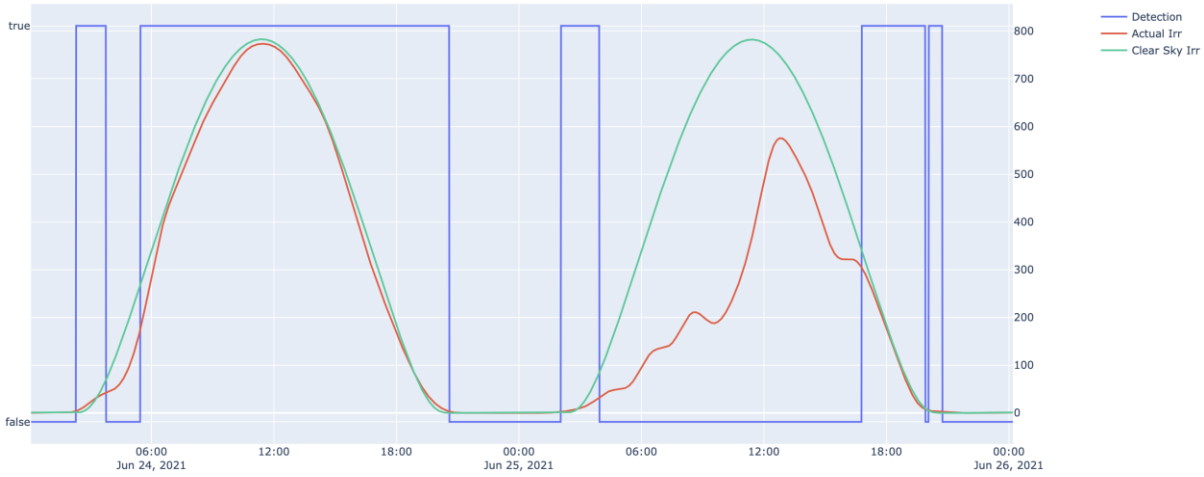


Figure 3.15: Clear sky detection algorithm result.

Figure 3.16 displays clear sky detection minute-based data with transformed hourly based data. For the purpose of eliminating some detection errors due to data transformation, a conservative approach was chosen to resampling data on an hourly base. Hence, some errors were removed at the cost of some clear sky data loss. For example, there was a clear sky in the early morning on 24th June 2021 from 02:19 to 03:36 UTC. When this time interval was converted to 1 hour sampling, 03:00 was classified as the clear sky and 2:00 was recorded as cloudy or no irradiance.

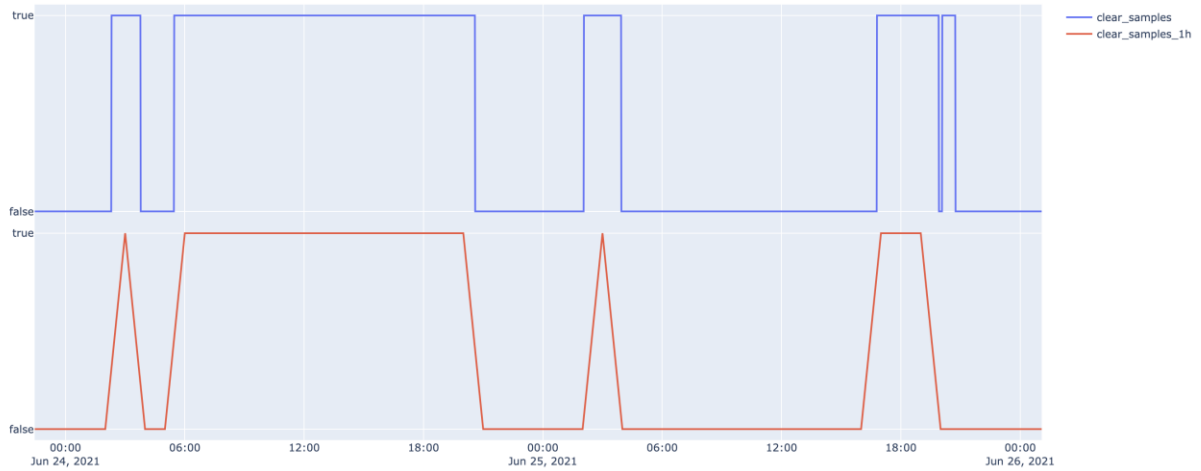


Figure 3.16: Clear sky detection minute-based data (blue) with transformed hourly based data (red).

3.2.2.2 Plane of Array Irradiance Data

Since PV power output is dependent on the plane of array irradiance (POA) data, global horizontal irradiance measured values are required to be converted to the plane of array irradiance values. POA values can easily be calculated by the get_total_irradiance algorithm in pvlib, however, the solver requires measured dni and dhi which are not available. There are models for estimating dni and dhi values from ghi. Analytical approaches require measured dni and dhi values to calculate POA values. One model is the erbs model in the pvlib library. The erbs model uses diffuse fraction to estimate dhi values and dni values are calculated by an equation that uses a zenith angle [39]. By feeding the get_total_irradiance function with solar azimuth and zenith, surface tilt and azimuth, ghi, and generated dni and dhi values, POA values for the south, east, and west directions were obtained. Figure 3.17 shows monthly total irradiance values based on direction. While the south and west directions receive higher irradiance compared to measured global horizontal irradiance, the east direction catches low irradiance.

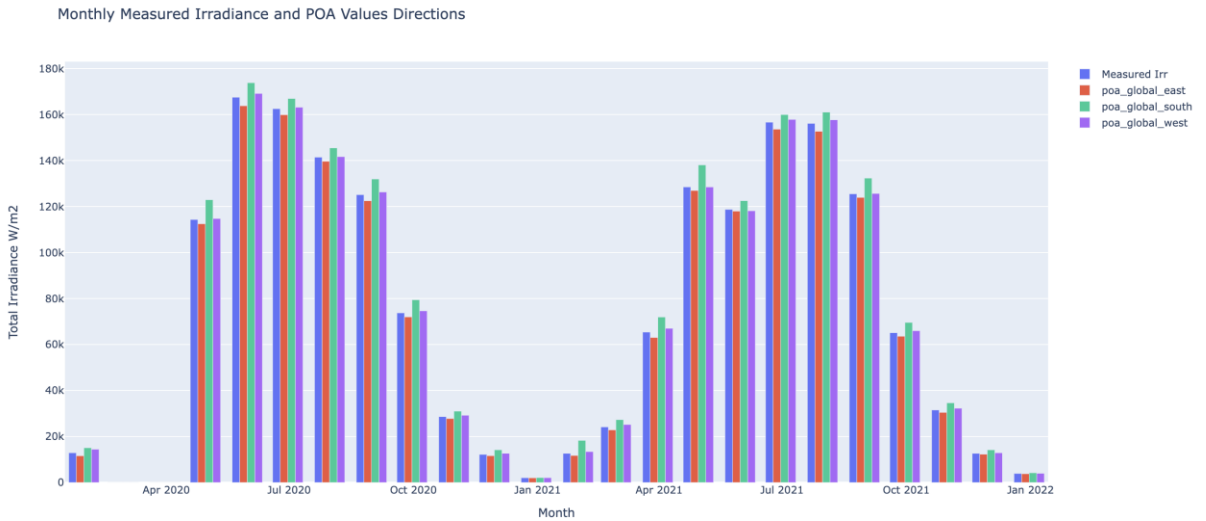


Figure 3.17: Monthly total irradiance values based on direction.

Figure 3.18 shows the plane of array irradiance for the east direction compared with measured ghi values which is assigned as actual on the graph. Global horizontal irradiance is slightly higher than the plane of array irradiance values for each direction throughout the year.

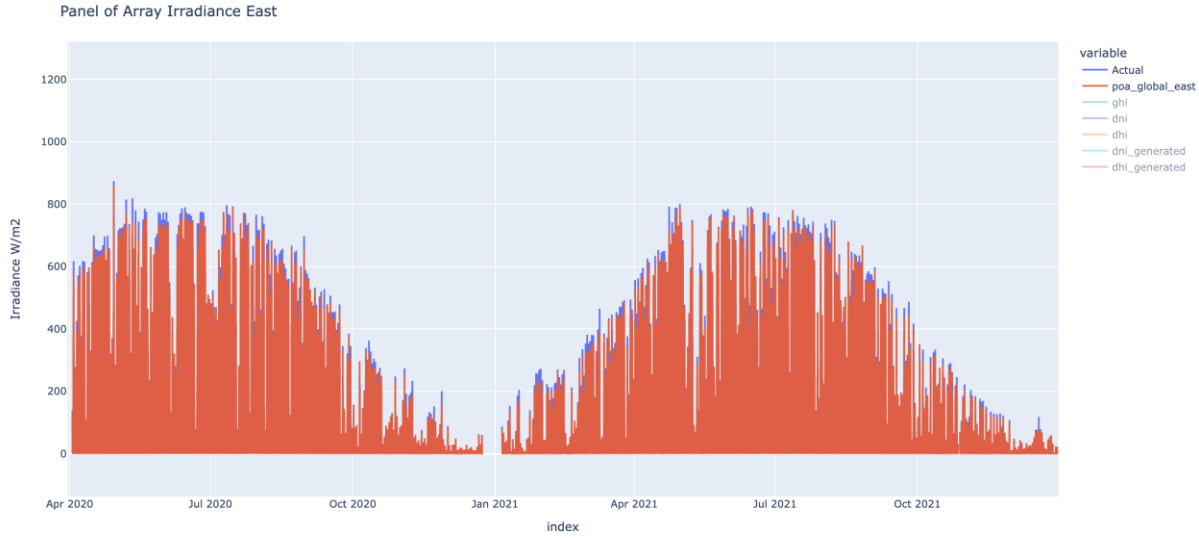


Figure 3.18: Plane of array irradiance (red) and measured irradiance (blue) against time.

3.2.3 PV Plant Data

A comprehensive signal list for the PV plant was provided by Lede Energi. PV power out analysis is based on inverter AC power output. Inverter status values were also used to examine whether the inverter was functioning well or not. Status values are Boolean type whereas power values are float in kW. Data covers only two years, 2020 and 2021. Figure 3.19 shows IV1-8 power values separately, and total_AC_power_IV_on represents the sum of inverter power values which are only in operation together. That means if all inverters produce record power values at the same time, it will be summed up. Otherwise, there will be no record for total ac power. Total AC power represents plant total power production on high at its best times where all inverters are in operation together. As it is seen from Figure 3.19 that there are some periods inverters were not in operation. That is why not all inverters are valued to be analysed. Inverter 2, Inverter 5, and Inverter 7 values are going to be further analysed as representations for each direction. Inverter 2 and 5 are identical which means the same number of modules, 400, are connected with the same module type, 295 Wp. Inverter 7, however, is powered by 444 panels and a 300 Wp module type. Pre-processing results will be shown in the results chapter.

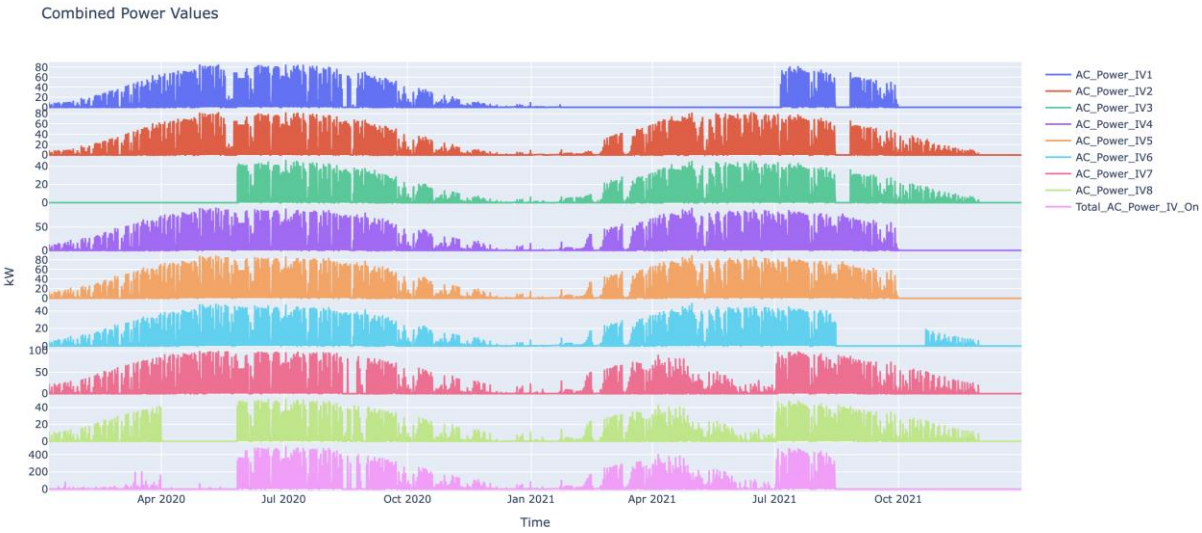


Figure 3.19: Inverter based AC power values against time.

PV power data has gone through a bunch of pre-processing. As it is seen from the graph and detailed investigation of data, there are zero power outputs which either represent an inverter/module problem or night time values. Firstly, zero values were removed. By using solar elevation data from the pvlib library, data before sunrise and after sunshine were cleared out from the dataset. Ultimately, the PV power dataset consists of only values which cover the time the sun's presence without fault data.

Module temperature is another important parameter in PV power production. Solar cells have an optimum operating temperature range. Above a certain level, PV module power output decreases due to low voltage. Wind speed, air temperature, and module materials help to maintain heat balance on PV cells and PV modules. Therefore, module temperature is a crucial parameter to watch, however, the parameter is not available for the plant. Instead, module temperature values were generated since pvlib.sapm_module function in pvlib provides this parameter by feeding the function with the plane of array irradiance, wind speed, air temperature and coefficients for the module type. Figure 3.20 shows module temperature with air temperature on the same scale, and power output of IV2 in kW for different days in July 2021.

Figure 3.20: Module temperature (red), air temperature (green) on the right y-axis, IV2 AC power (blue) values on the left y-axis.

Data filtering for PV data is another data processing step. Even though the prediction of PV power output will be under any circumstances such as cloudy days, low irradiance and low irradiation, measuring PV data quality and filtering corrupt and inaccurate data may increase the model accuracy. This idea will be investigated. Therefore, low elevation and irradiance values which cause possible noise in PV data, were determined for further processing. One way to measure PV data quality is to compare actual values with predicted power output values which are estimated by a general power output formula shown in Equation (3.1).

$$E = A * \eta * I * L \tag{3.1}$$

where A is total solar panel area (m^2), η is panel efficiency, I is irradiance (W/m^2), L is loss factor, and E is expected power output (W). One estimation example was done for IV2 connected panels which have 1.67*400 m^2 area, 17.7% assumed panel efficiency, 0.75 assumed loss factor and irradiance values throughout the year. By dividing expected power output by actual power output, a performance index value is obtained. Figure 3.21 shows elevation and irradiance values plotting with performance index values as data points. Yellow points correspond to the lowest performance values which are assumed as the lowest 10% of performance values. As it is seen clearly from the graph, yellow dots are gathered where elevation is below 9º and irradiance around 100 W/m^2. Obviously, there are also a significant number of data points has good performance index within these values. However, it is possible to comment that low elevation outweighs low irradiance. As a result, the prediction model will be analysed above 9º elevation during the analysis.

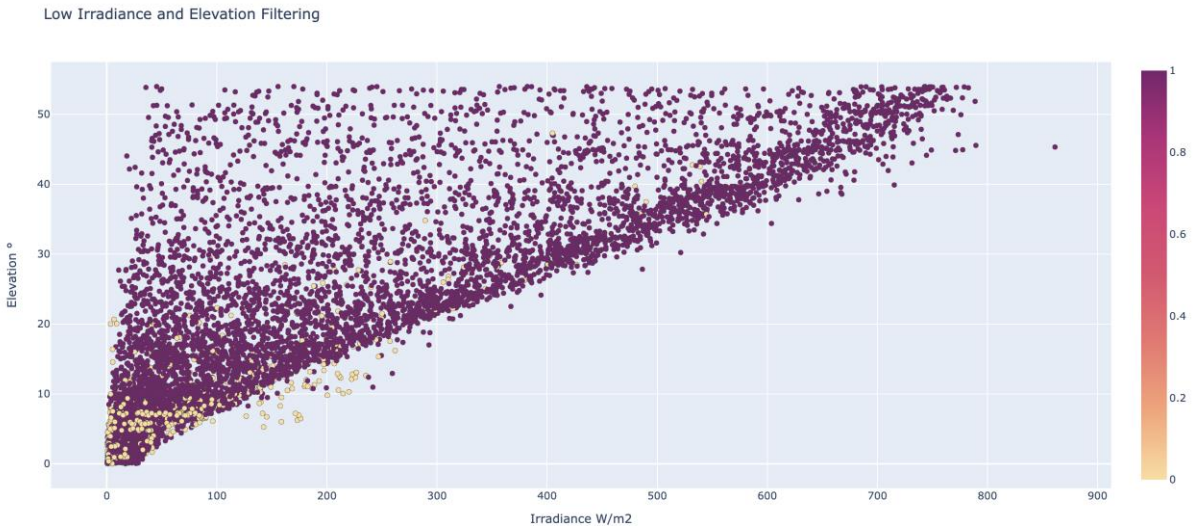


Figure 3.21: Elevation and irradiance with coloured performance index.

At the end of the data gathering and pre-processing phase, all features are ready to further processing and analysis. Figure 3.22 shows the whole process of data dealing methodology at a glance.

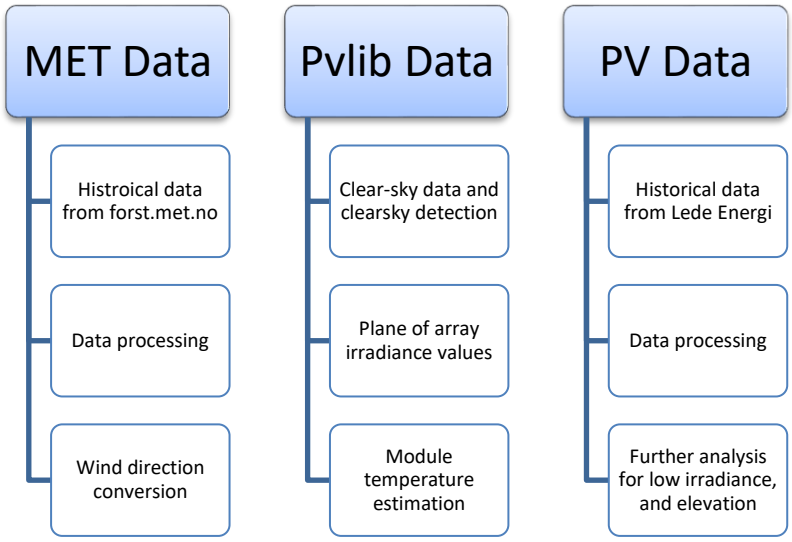


Figure 3.22: Data dealing methodology for each data set.

3.3 Correlation Analysis

Correlation analysis is a method to evaluate the linear relationship between variables. The Pearson correlation analysis is by far the most common method for correlation analysis [40]. Correlation values vary from -1 to 1 and positive values indicate variables are positively correlated. For example, if a negative correlation exists, it means that while one variable increases, the other variable decreases. The more correlation values close to -1, the more correlation gets stronger. The correlation becomes weaker close to zero values. The power of Pearson correlation analysis is that it is not influenced by the variable regardless of dependent or independent [41]. In this study, pandas.DataFrame.corr() in python was used for calculating Pearson correlation coefficients.

3.4 Principal Component Analysis

Dealing with a great number of features and using all features in further data analysis can be excessive or can cause uncertainty in the model. Principal component analysis is a powerful method that is used in data analysis. The method is able to handle feature reduction or outlier detection. By maximizing the number of variations in the features, it creates orthogonal components that is based on orthogonal decomposition [41]. Eigenvectors and eigenvalues are produced based on covariance matrix. The output class of the matrix is linear relation within the input variables. The highest variance obtained is held in the first principal component and other components are produced with a decreasing variance score. Moreover, scree plots and visualization of values help to understand the output classes. Scree plot describes eigenvalues and the explanation ratio of components. In this study, sklearn.decomposition PCA was used to perform PCA analysis. However, there is one point in PCA analysis that is critical to making analysis effectively and that is scaling. Scaling transforms the data and creates a new set of data which is in the same range by keeping variance information in the dataset. Thus, each feature contributes equally to the analysis. There are two common scaling methods. These are standard scaling and minmax scaling which is also called normalization.

Standard Scaler is a function in Scikitlearn and aims to remove the mean and scales the data to unit variance. Thus, mean average becomes zero and the standard deviation one. However, standard scaling is influenced by outliers in the dataset. In order to make all feature set mean average zero, outliers take extreme values and data distribution does not reflect the main dataset. Therefore, standard scaling is useful where data distribution normal or Gaussian.

If the feature has a skewed distribution, then, MinMax scaling will keep the shape of the dataset as it distributes the values for a given range such as [0, 1]. The spaces between each feature are maintained, and the information and shape of the dataset are mainly preserved. This method is useful for regression and neural network methods. Since outliers were removed from the dataset, standard scaler was used for PCA analysis.

3.5 Prediction Methods

In this study, two prediction methods were investigated. Linear regression (LR) is one of the basic statistical approaches to various problems that assume a linear relationship between inputs and outputs. Furthermore, more advanced and improved methods have been developed in recent years. One of them is Artificial Neural Networks (ANN) which is able to handle non-linear relationships between inputs and outputs. Sklearn library linear_model was used for linear regression and Keras API was used for ANN which works on TensorFlow 2 [42]. TensorFlow is an open-source library and Keras is a high-level neural network library that runs on TensorFlow. As it was stated in the theory part, it is essential to use scaling in the ANN model. Importantly, the dataset should be split into training and test sets before models are used. While LR has a simple methodology, the ANN model requires input parameters. Table 3.12 shows input parameters that were used for ANN model. Model parameters are subject to change to improve model accuracy. Selected parameters are indicated as trial in the table.

Table 3.12: ANN network model parameters and inputs.

Model Parameters	**Inputs**	**Model Parameters**	**Inputs**
Number of Inputs	Trial	Optimizer	Adam
Training/Test size	0.25	Batch size	Trial
Model Type	Sequential/ Feedforward	Learning rate	Trial
Nodes	Trial	Epoch	Trial
Dense layer	3	Loss	MSE
Hidden layer	1	Performance	MAE, MSE, RMSE, R^2
Output nodes	1		
Activation function	RELU		
Kernel	Glorot-uniform		

The methodology of prediction models is summarized in Figure 3.23. Specifically, reverse scaling of scaled data in the ANN model is done before visualization.

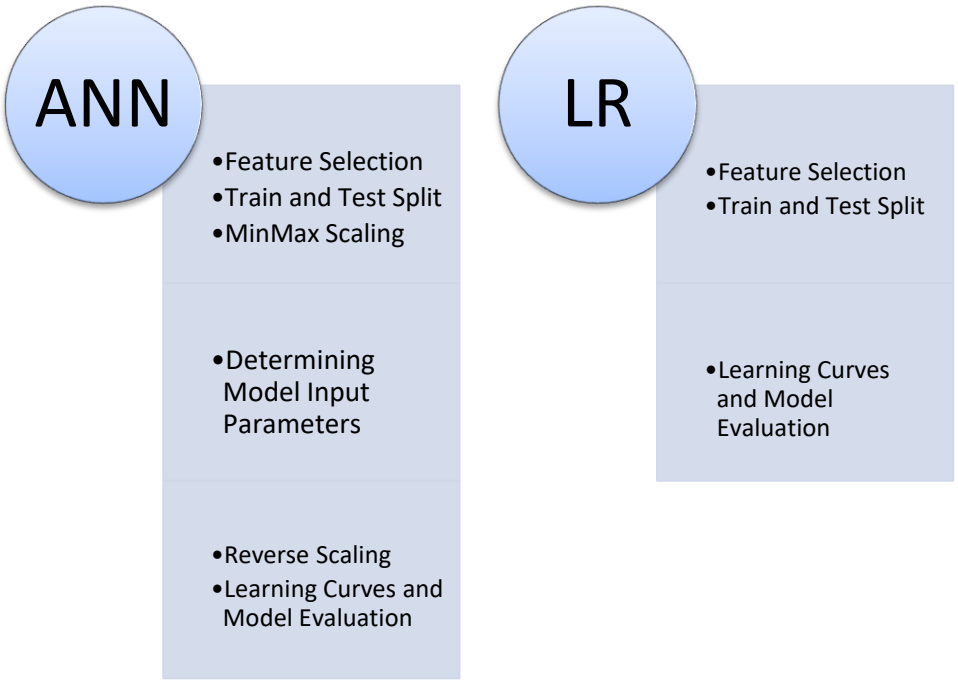


Figure 3.23: ANN network and linear regression model methodology.

3.6 Pvlib and other Python Libraries

Python and its libraries were preferred as the programming language. It will be emphasised information on some specific functions in this part.
pvlib.irradiance.get_total_irradiance: This function produces the plane of array irradiance and its beam by introducing surface tilt, surface azimuth, solar zenith, solar azimuth, dni, ghi and dhi values. Poa_global, poa_direct, poa_diffuse variables are created. Isotropic sky diffuse model is selected by default [39].
pvlib.solarposition.get_solarposition: This function produces zenith, elevation, and azimuth of the solar position in addition to the equation of time. The function takes time, altitude and longitude variables as an input [39].
pvlib.irradiance.erbs: This function estimates DNI and DHI values from global horizontal irradiance, GHI by using the Erbs model [39].
pvlib.get_clearsky: The function calculates clear sky irradiance values for a given time interval and location. The default calculation method is Ineicnen [39].
pvlib.clearsky.detect_clearsky: The algorithm determines clear sky times based on measured GHI values. Measured and clear sky irradiance values are inputs in addition to window length which is the length of the sliding time window in minutes. The best performance was captured by determining window length is 8 [39].
pvlib.temperature.sapm_module: The function calculates module back surface temperature by Sandia Array Performance Model. Plane of array irradiance, wind speed, air temperature and a and b coefficients are inputs for the algorithm as shown in Equation (3.2).

$$T_m = E * exp(a + bxWS) + T_a \quad (3.2)$$

where E is irradiance, WS is wind speed, T_a is air temperature. a and b coefficients take different values based on module type (glass/polymer) or mounting type (open rack/close roof/insulated back). For glass/polymer module and open rack mounting, a takes -3.56 and b takes -0.075 as inputs [39].
sklearn.metrics: The function is used for model evaluation such as RMSE, and MAE. It is important to note that MAPE in sckitlearn does not return a percentage value [42]. That is why a function was written to calculate the percentage output.

3.7 Case Studies

It is possible to make different types of predictions with large datasets. Without a doubt, the overall aim is to keep PV power at a minimum error. Since this study also investigates how PV power output varies with other meteorological data, different types of case studies were planned. For example, one case study evaluates how models behave without measured irradiance data but calculated clear sky data in the event of having no access to measured irradiance values. Another case study tries to explore each meteorological variable impact on PV power output. Table 3.13 shows planned case studies for both ANN and LR.

Table 3.13: Planned case studies for PV power output analysis

	Case Study	Information
1	Model performance on clear sky days	Consecutive prediction
2	Model performance on clear sky days	Consecutive prediction without measured irradiance values
3	Training and tests set on a yearly basis	Training with 2020 data and test with 2021 data
4	Training with the 2020/2021 data set	Shorter test set with longer training data.
5	MET variable selection	Each meteorological variable is examined
6	Prediction with forecasted data	1 day ahead PV power output prediction

4 Results

This chapter was divided into 3 main sections. Data inspection and pre-processing are started with meteorological data. Later, data were combined with PV power data. The merged dataset is used to make PV power output predictions. According to the methodology of this work, the results for data processing are represented in this part and the results are supported by data visualisation. The output of this study for power output prediction trails and model evaluation results are introduced at the end.

4.1 Meteorological Data

4 years of meteorological raw data is shown in Table 4.1. Raw data represents raw values without being processed. The data type explanation was given in Appendix C. Count refers to the total number of available data in dataset. The method of calculation mean and standard deviation is as mathematical standards. Minimum and maximum values show minimum and maximum values in the relevant column. Percentiles of data are categorized as 25%, 50%, and 75% of data. 50% values also represent the median value of the related column.

Table 4.1: Raw meteorological data statistics.

	dew_point_temp	air_temp	relative_humidity	irradiance	wind_speed
count	34806	33776	31036	31776	33622
mean	3.06	7.61	74.84	115.69	1.49
std	7.65	8.51	20.83	198.76	1.26
min	-20.4	-17.2	16	-6.8	0
25%	-2.5	1.1	60	0	0.6
50%	3	7.3	79	5.2	1.1
75%	9.2	14.1	93	144.3	2.1
max	23.2	31.5	100	1491	10.1

As it is seen from the data statistics, the number of data varies. The total number of missing values is shown in Table 4.2.

Table 4.2: The number of missing values of 4 years of meteorological data.

Missing Values	Count
dew_point_temp	139
air_temp	1169
relative_humidity	3909
irradiance	3169
wind_speed	1323

Missing values were dropped from the dataset in addition to unrealistic minus irradiance values. In the end, the processed data summary is shown in Table 4.3.

Table 4.3: Processed data statistics.

	dew_point_temp	air_temp	relative_humidity	irradiance	wind_speed
count	29311	29311	29311	29311	29311
mean	3.57	8.55	74.59	121.23	1.53
std	7.76	8.54	20.98	195.36	1.25
min	-20.4	-17.2	16	0	0
25%	-2.2	2	59	0	0.6
50%	3.9	8.9	79	11.1	1.1
75%	9.9	15	93	163.05	2.1
max	20.9	31.5	100	1186	10.1

The total number of values dropped to 29311 from the minimum count in the raw dataset. The reason is that distribution of missing values is not homogenous. For example, while there is a value in the air temperature column for a specific row, there is no value for corresponding irradiance values. In this case, the whole row is deleted to keep the same amount of data in each column.

Another data inspection step is histogram figures of variables. Thus, it is possible to get an overview of data distribution. Figure 4.1 shows histograms of each variable.

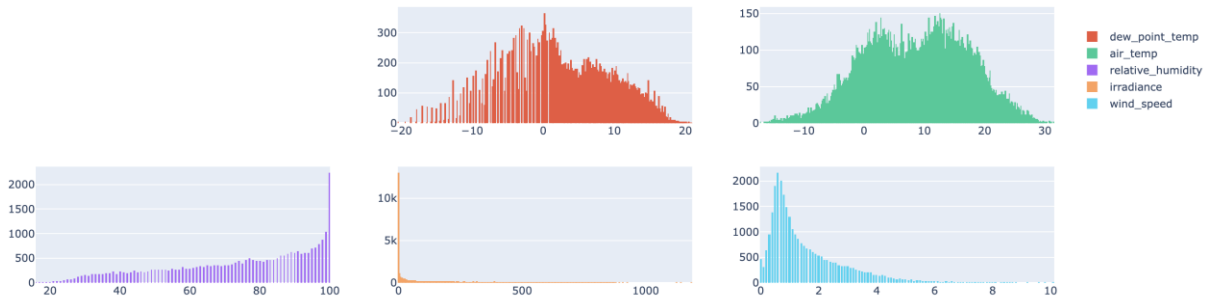


Figure 4.1: Histograms of meteorological variables.

While wind speed has a skewed distribution, dew point and air temperature are close to normal distribution. Irradiance and relative humidity histograms are heavily under the impact of night and rainy/cloudy times.

Another type of data inspection step is producing scatter plots. There are some outliers were observed in Figure 4.2. For example, some irradiance values are above 1000 W/m2. When it was investigated these values, there was a time that the meteorological station was out of order. The specific 10 days from 18/11/2019 to 28/11/2019 were removed from the database. Wind speeds above 9.2 m/s were also excluded.

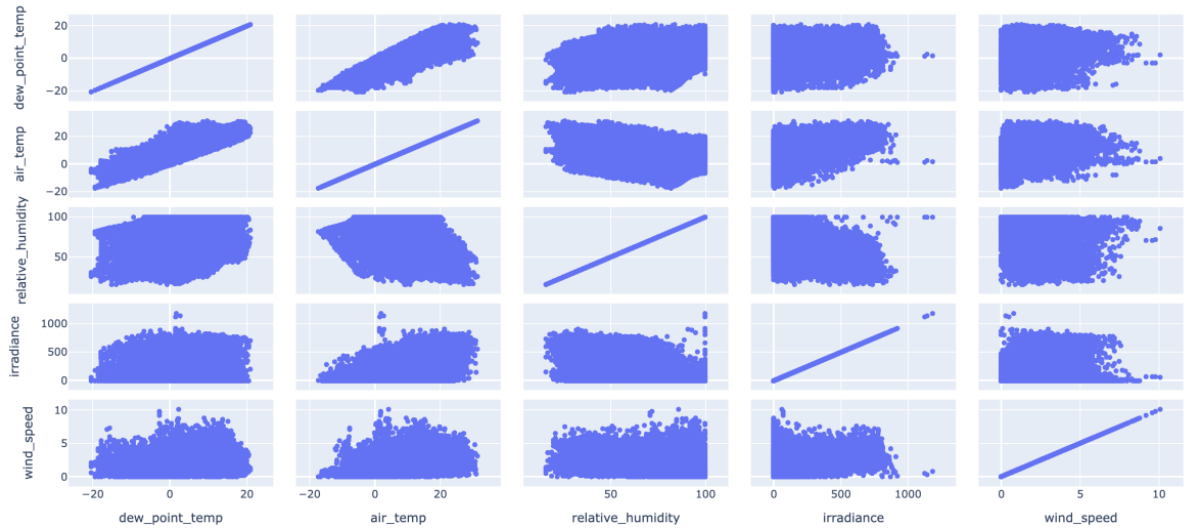


Figure 4.2: Scatter plots of meteorological variables.

Scatter plots after outliers removed are represented in Appendix C. Before investigate the dataset further, the last statistics of variables are shown in Table 4.4.

Table 4.4: Processed data statistics.

	dew_point_temp	air_temp	relative_humidity	irradiance	wind_speed
count	29267	29267	29267	29267	29267
mean	3.58	8.56	74.57	120.93	1.53
std	7.77	8.54	20.99	194.74	1.25
min	-20.4	-17.2	16	0	0
25%	-2.2	2	59	0	0.6
50%	3.9	8.9	79	11.1	1.1
75%	9.9	15	93	162.6	2.1
max	20.9	31.5	100	912	8.8

Figure 4.3 shows 4 years of processed meteorological data in a line graph after outliers were removed. From time to time the station was out of order. For example, at the beginning of 2020, July 2018 and the beginning of 2021 there were no records at the station. The missing periods including removed periods are connected by a line in the graph and they are not representing real values.

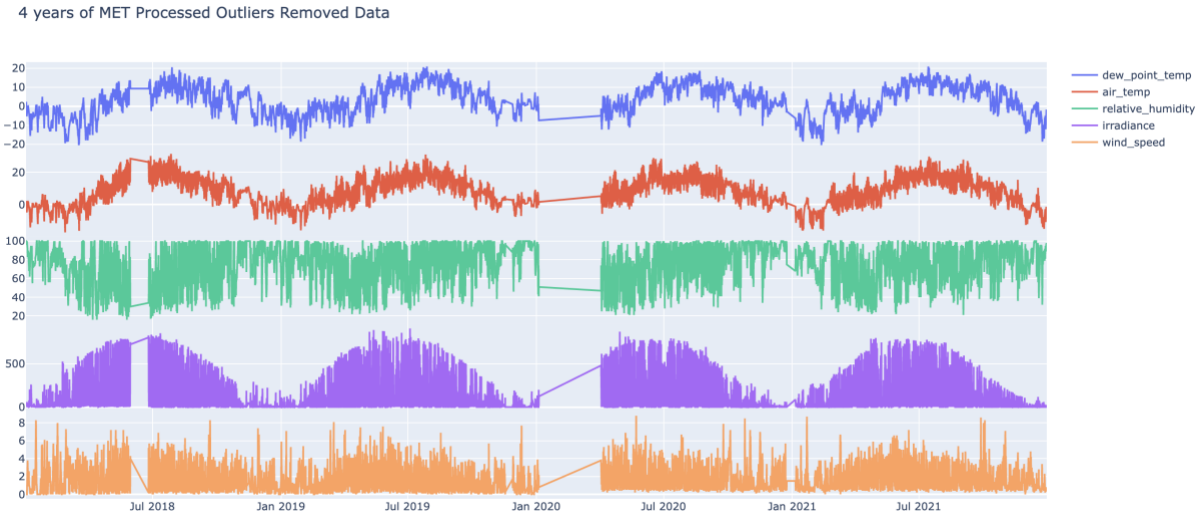


Figure 4.3: Meteorological data plotting against time after outliers removed.

Since all data was pre-processed, it is ready to investigate further. To understand how variables are correlated to each other, the correlation matrix was used. Figure 4.4 shows the correlation matrix of meteorological variables. The highest positive correlation is among air temperature and irradiance with 0.82, whereas the lowest positive correlation is among dew point temperature and wind speed. Likewise, the highest negative correlation is observed between relative humidity and air temperature, however, the lowest negative correlation is observed in two different variables as relative humidity – air temperature and relative humidity – wind speed.

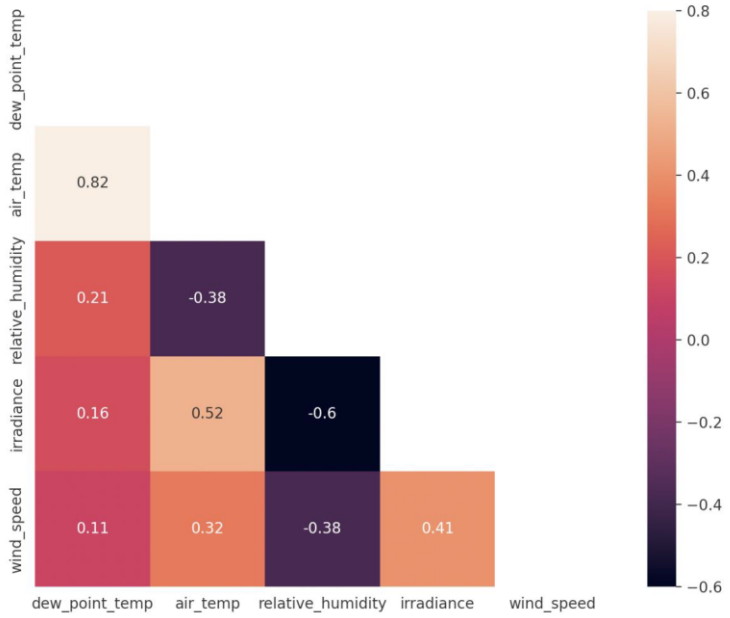


Figure 4.4: Correlation matrix of meteorological variables.

The next step is performing PCA analysis. 3 PCA components explain 92% of the total variance in the dataset with 49% PC1, 29% PC2, and 14% PC3. The scree plot of variance explanation is shown in Figure 4.5.

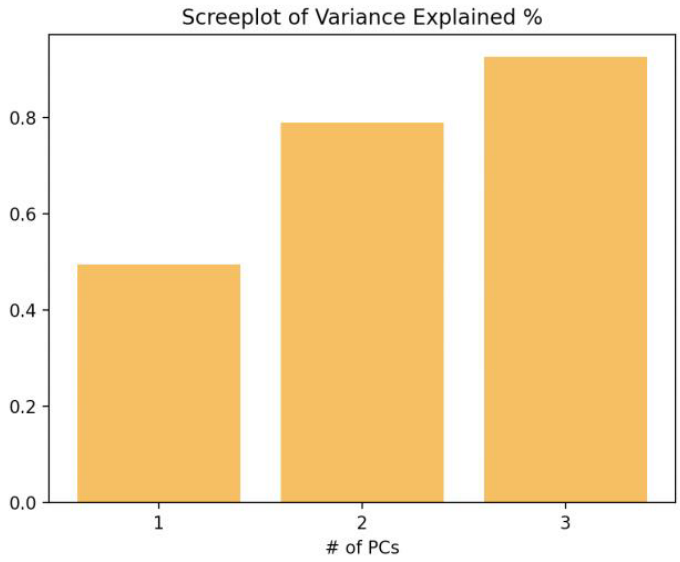


Figure 4.5: Scree plot of 3 components PCA analysis.

Each PCA component results are shown in scatter plots with respect to other components. In Figure 4.6, principal components 1-2 and principal components 1-3 are represented. A detailed representation for PC1-2 is in Appendix D with indicated colours which refer to irradiance values.

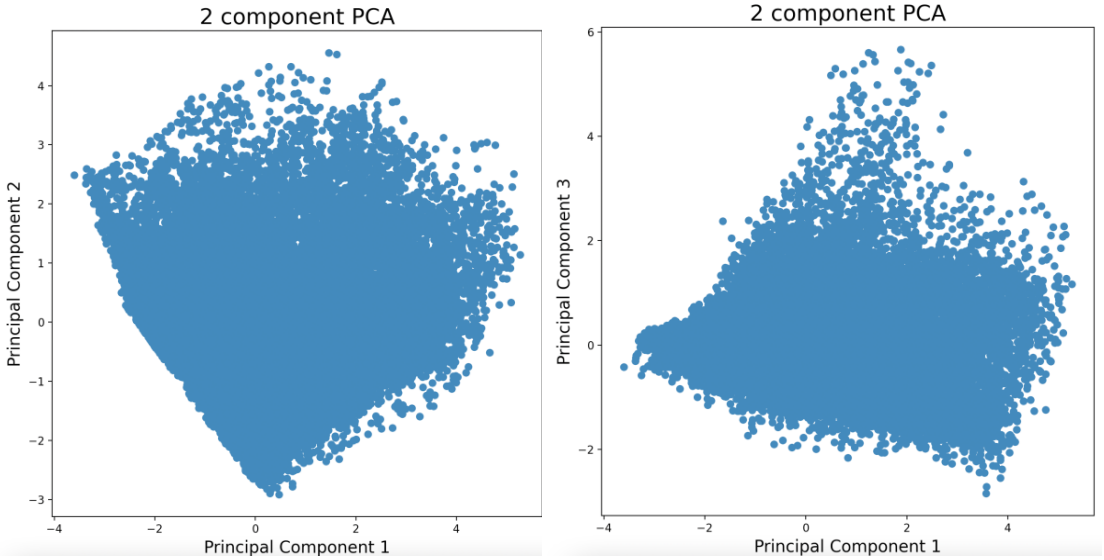


Figure 4.6: Principal components plotting PC1-2 (left), PC1-3 (right).

PCA loadings are shown in Figure 4.7 and Figure 4.8. In Figure 4.7, while irradiance and relative humidity have a high contribution to PC1 and PC2, the negative correlation in irradiance and relative humidity together with wind speed is hidden in these loadings. It is observed that irradiance and wind speed are positively correlated. Although air temperature contributes more to PC1, dew point temperature influences PC2 more than PC1.

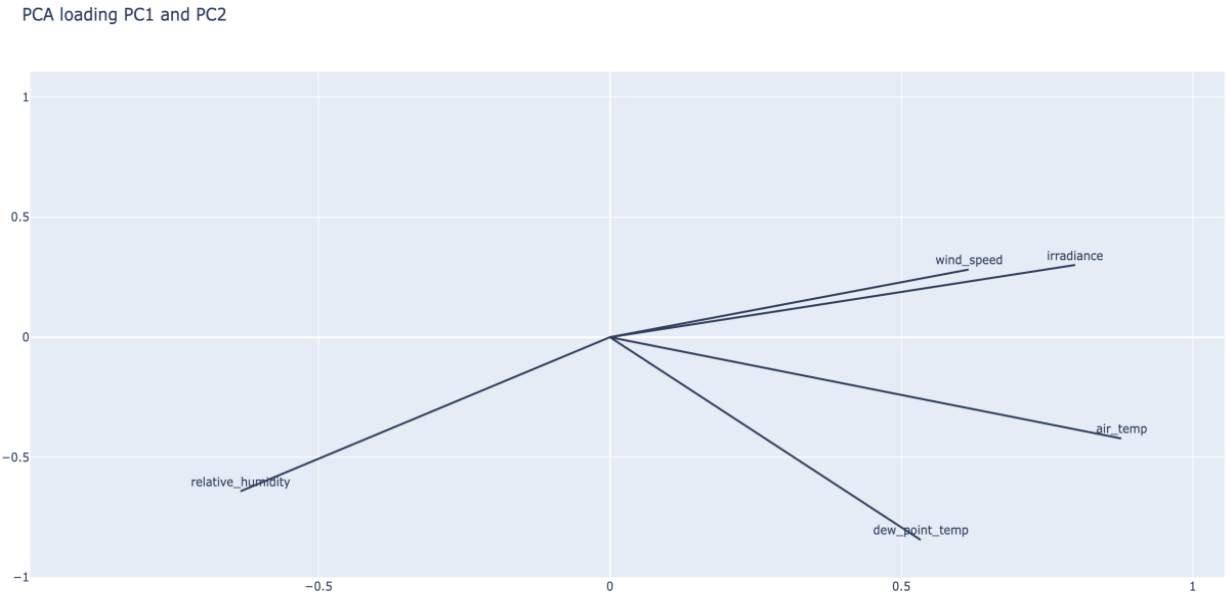


Figure 4.7: PCA loadings for PC1 and PC2.

In contrast to PC1 and PC2 loadings, the highest contribution comes from wind speed and relative humidity to PC1 and PC3, shown in Figure 4.8. Dew point temperature, air temperature and irradiance did not capture by PC3. It should be noted that PC1 and PC3 only explain 63% of total variations.

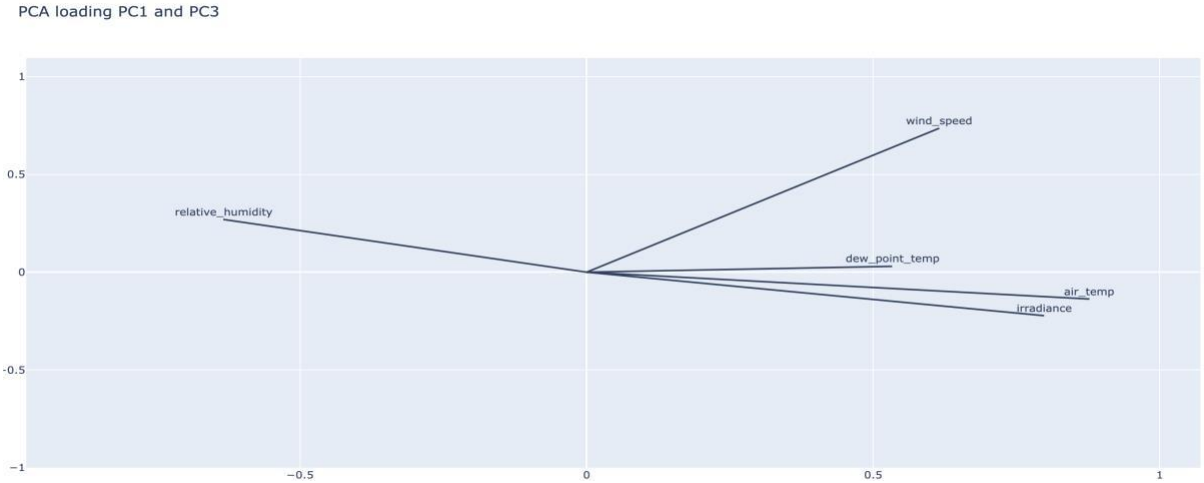


Figure 4.8: PCA loadings for PC1 and PC3.

It is clear that different meteorological variables are captured differently in PCA loadings. 3 PCA loadings represent 92% of the total variance in 4 years period dataset. During the PCA analysis, when components were plotted by using a scatter plot, a linear boundary was observed in the PC1 and PC2 plots. This finding was investigated further and it turned out that limits in meteorological conditions cause this type of boundary. More broadly, the variables accumulated near the linear boundary belong to low irradiance values or 100% relative humidity.

Wind direction data is only available after April 2020. That is why the data from 2020 to 2021 was investigated separately from 4 years of data. Figure 4.9 reveals the histogram of variables

for 2020-2021. At the same time, 4 years of meteorological and 2020/2021-year meteorological situation is going to be compared.

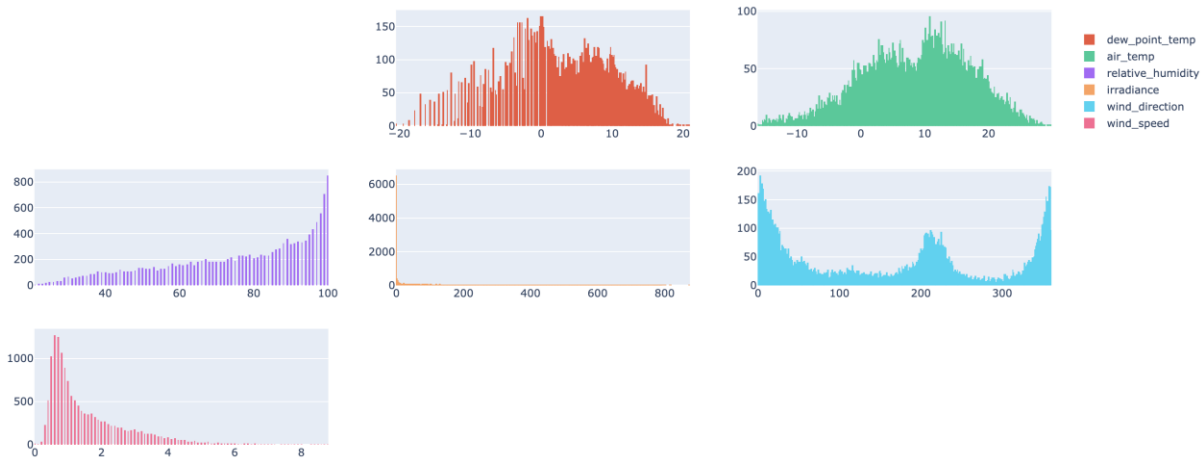


Figure 4.9: Histograms of meteorological variables from 2020 to 2021.

Wind direction is recorded in degrees. That is why 360 different values have dominated the wind direction data. In order to produce meaningful results from wind direction data, this variable was categorized as 4 main directions: north, south, east and west. The methodology was explained under the methodology section. It was aimed that with 4 main direction categories, a better correlation with other meteorological variables and in PCA analysis would be achieved.

Table 4.5 shows 2020–2021-year data meteorological variables statistics. Wind direction data recording starts from April 2020, that is why the number of available data points is the lowest.

Table 4.5: 2020–2021-year data meteorological variables statistics.

	dew_point_temp	air_temp	relative_humidity	irradiance	wind_direction	wind_speed
count	17384	16946	15037	15045	14968	16949
mean	3.37	7.8	75.85	115.78	163.32	1.57
std	7.53	8.26	20.2	191.16	124.87	1.24
min	-20.4	-16.1	21	0	0	0
25%	-1.9	1.8	61	0	31	0.7
50%	3.8	7.6	81	7.9	179	1.1
75%	9.3	13.9	94	154.8	252	2.1
max	20.9	29.8	100	1438	360	9.7

The mean value of wind direction is 163º and corresponds to the south-southeast direction. After pre-processing of this data, the new statics are shown in Table 4.6.

Table 4.6: Processed data statistics.

	dew_point_temp	air_temp	relative_humidity	irradiance	wind_speed	wind_category
count	14944	14944	14944	14944	14944	14944
mean	3.89	8.54	75.85	115.58	1.59	1.94
std	7.79	8.42	20.21	189.55	1.23	1.07
min	-20.4	-16.1	21	0	0	1
25%	-1.4	2.7	61	0	0.7	1
50%	5	9.3	81	7.9	1.1	1
75%	10	14.7	94	155.7	2.2	3
max	20.9	29.8	100	874	8.8	4

The frequency of wind direction data is shown in Figure 4.10. As it is seen clearly, wind category 3 is the highest frequency after wind category 1.

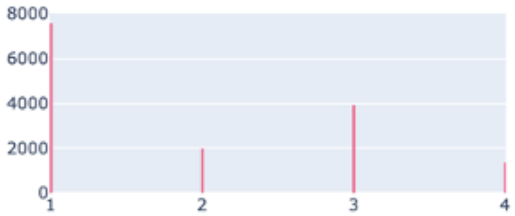


Figure 4.10: Wind direction categorical data histogram.

Figure 4.11 shows the correlation plot comparison after wind direction in degrees categorised as 4 main directions. While the figure on the left-hand side consists of the wind_category variable which states categorised wind direction as 4, the figure b on the right-hand side shows correlation of wind_direction variable in degrees with other meteorological data. More robust results were obtained for wind direction. One spectacular difference is relative humidity and wind category/direction have no correlation. Wind direction data which is numerical has no correlation with irradiance, however, a relationship was captured by categorical wind data. Air and dew temperature correlations against wind category slightly increased.

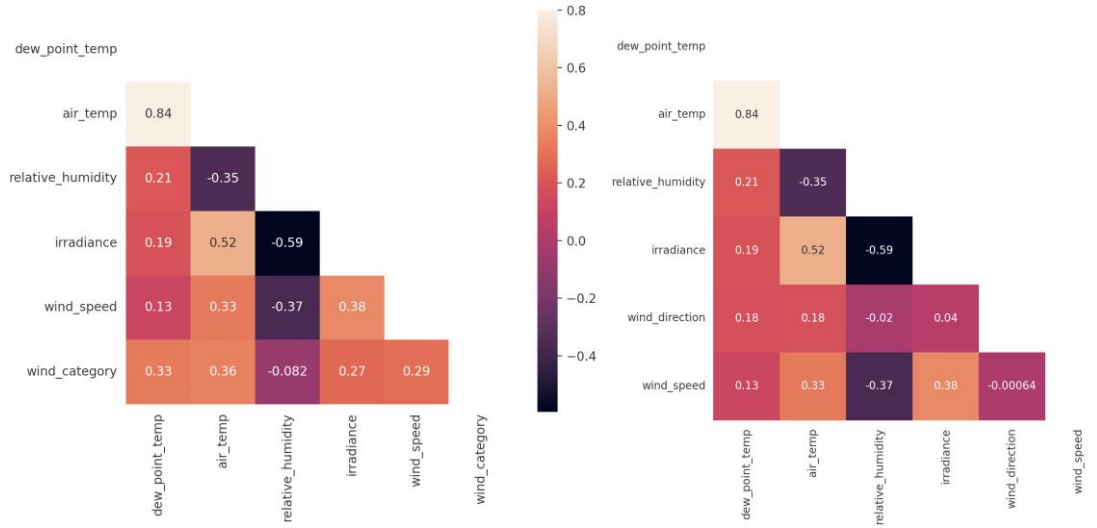


Figure 4.11: Correlation matrixes including wind category (left) and wind direction numerical values (right).

Once more, PCA analysis were repeated. 3 PCA components explain 84% of the total variance in the dataset with 45% PC1, 25% PC2, and 14% PC3. The scree plot of variance explanation is shown in Figure 4.12. Since the wind_category variable was added to PCA analysis, the explanation dropped to 84% from 92%.

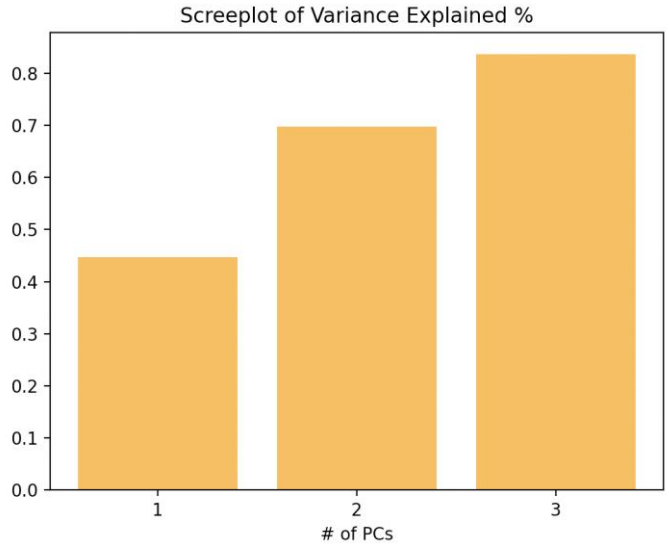


Figure 4.12: Scree plot for data including wind direction.

PC1-2 and PC1-3 scatter plots are shown in Figure 4.13. Likewise, first PCA analysis, the distribution is similar. However, since categorical information was added such as wind_category, we capture wind category information in PC1 and PC3. PCA loadings will explain the relation of variables further.

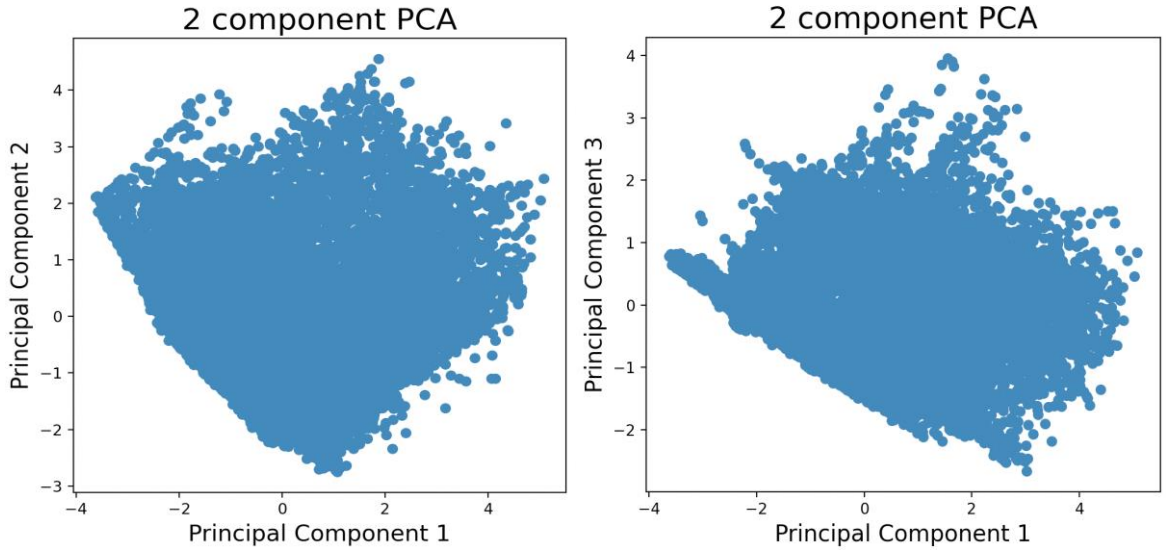


Figure 4.13: PCA scatter plots with PC1-2 (left) and PC1-3 (right).

Figure 4.14 describes PC1 and PC2 loadings. The wind category contributes to both PC1 and PC2. In addition, wind category and air temperature are positively correlated. This information was also gathered in correlation analysis.

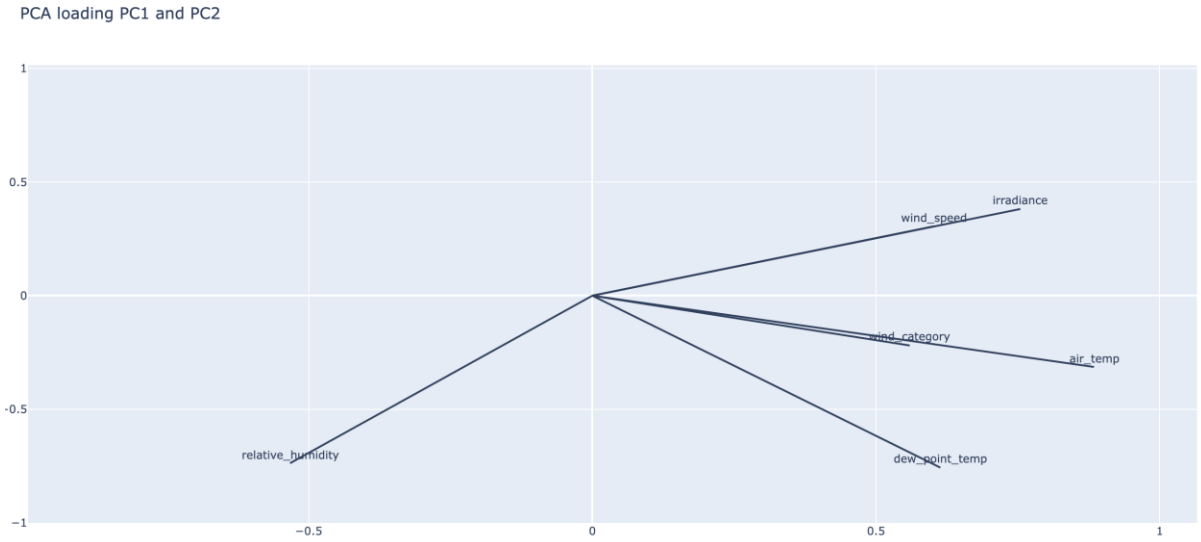


Figure 4.14: PCA loadings for PC1-2.

Figure 4.15 describes PC1 and PC3 loadings. PC1 and PC3 explain only 39% of the dataset. Air temperature, dew point temperature and irradiance are positively correlated. Furthermore, wind speed and wind category variables were captured as positive correlation in PC1 and PC3.

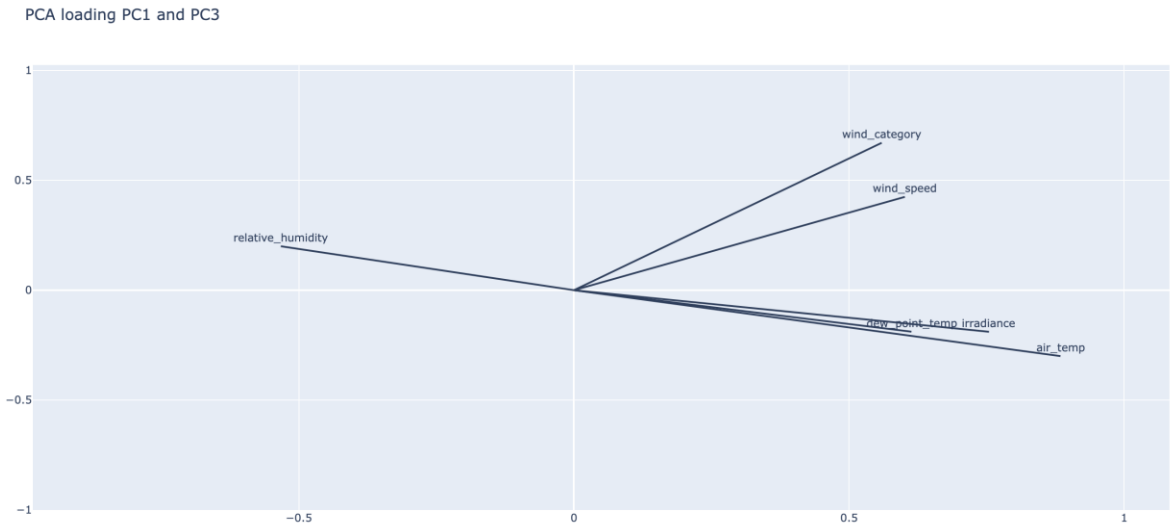


Figure 4.15: PCA loadings for PC1-3.

As a result, the meteorological variable analysis included wind category does a better job with categorised meteorological variable, however, PCA analysis does not produce clear wind category relation, yet.

Wind direction data was analysed in detail to understand the relation with other meteorological variables. Table 4.7 shows how meteorological variable mean values changes regarding wind direction category.

Table 4.7: Meteorological data mean variables with respect to wind category.

Wind Category	dew_point_temp	air_temp	relative_humidity	irradiance	wind_speed
1	1.32	5.46	77.90	54.73	1.11
2	3.61	8.79	72.94	155.41	1.97
3	8.59	13.91	73.35	221.07	2.51
4	5.01	9.75	75.62	92.3	1.04

Since the location of the station is in the south of Norway, the wind from the north should decrease the weather temperature. In addition, seasonal changes affect the direction of the wind. The lowest mean values of air temperature, irradiance and dew point temperature were recorded when the wind comes from the north. In other words, this indicates the winter season. In contrast to the north direction, the highest mean values of air temperature, dew point temperature and irradiance were recorded when the wind comes from the south. This direction refers to the south and summer season. For east and west directions, it is difficult to make a conclusion. While the east direction had higher irradiance than the west, air temperature and dew point temperature were slightly higher. The highest wind speed was recorded in the west direction. However, wind speed measurement below 1 m/s can be affected by local turbulence at 10m. Making correlations below and around 1 m/s measurement is not very reliable. Statistically, wind from the southwest is frequent in summer for the south of Norway and wind speed could be higher daily due to solar rotation in the afternoon.

MET data was analysed seasonally to understand the weather changes in detail. The following radar charts belong from autumn 2019 to 2021. Figure 4.16 shows the seasonal change in wind speed, and relative humidity. Winter and autumn seasons had frequently higher relative humidity.

Figure 4.16: Radar charts for seasonal meteorological variables, wind speed (left), relative humidity (right).

Figure 4.17 refers to dew point temperature in blue colour and air temperature in orange colour. On the right-hand side, the figure describes irradiance values. Summer seasons recorded the highest air temperature and dew point temperature. Irradiance values in 2020 spring and summer had very close numbers. One reason is that there was data loss in the 2020 spring season for irradiance values. The rest of the data which corresponds to the spring season belonged to largely the end of the season before the summer season began.

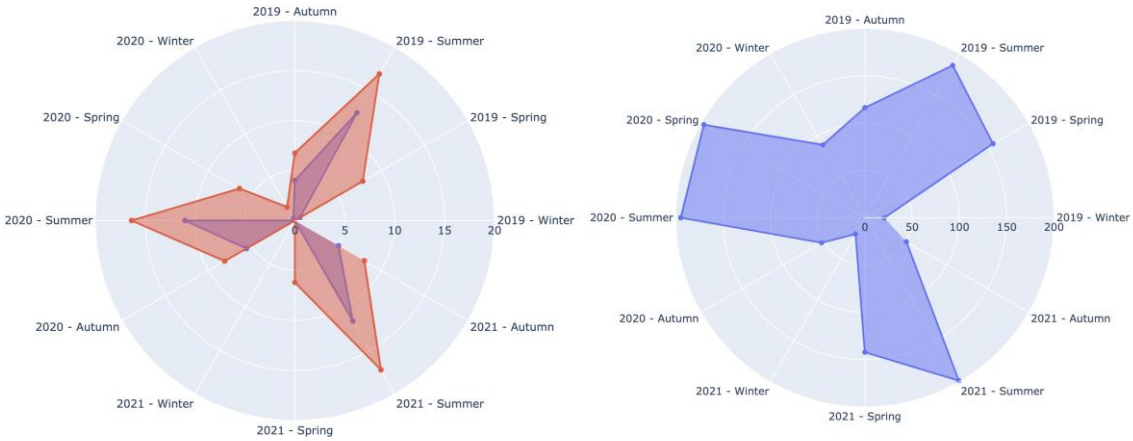


Figure 4.17: Radar charts for seasonal meteorological variables, dew point (blue) and air temperature (orange) (left), irradiance (right).

One question was raised during the analysis of relative humidity. As stated in PCA analysis, there is a linear boundary in PCA scatter plots. If it is rainy, relative humidity hits 100% and cannot rise further. The frequency of 100% relative humidity is also high. Instead of using relative humidity, values were converted to absolute humidity. Figure 4.18 shows the correlation matrix of absolute humidity and relative humidity values correlated with other meteorological variables for 4 years period from 2018 to 2021. Relative humidity has a strong correlation with irradiance which actually represents rainy days or sunny days due to cloudiness. However, we lose irradiance correlation with absolute humidity. Instead, absolute humidity has a strong correlation with air and dew point temperature due to the calculation method which uses air and dew point temperature.

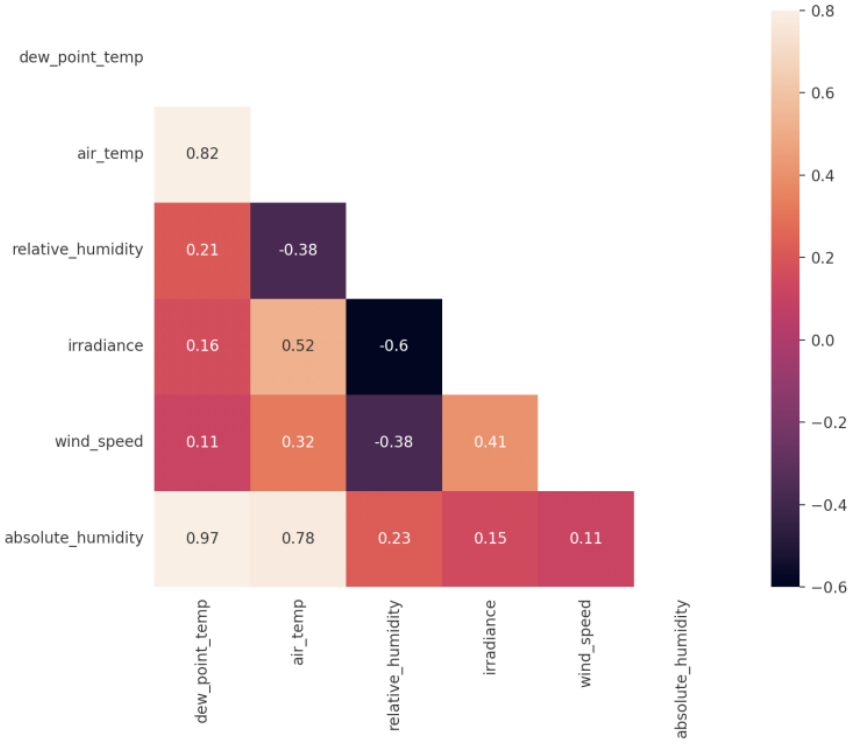


Figure 4.18: Correlation matrix included absolute humidity

PCA analysis was conducted again and PCA loadings are shown in Figure 4.19 and Figure 4.20. Absolute humidity was captured on the same side in PC1 and PC2 with a positive correlation.

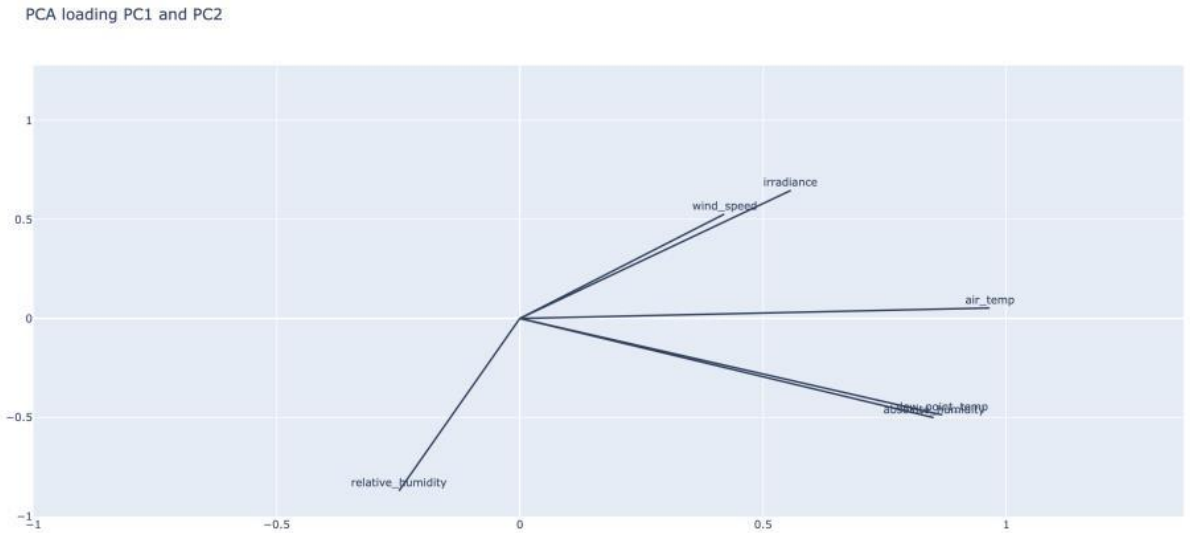


Figure 4.19: PCA loadings of PC1-2 included absolute humidity.

In contrast relative humidity, absolute humidity has no impact on PC3. This result is undesirable. There is a risk of losing one variable's footprint in PCA analysis.

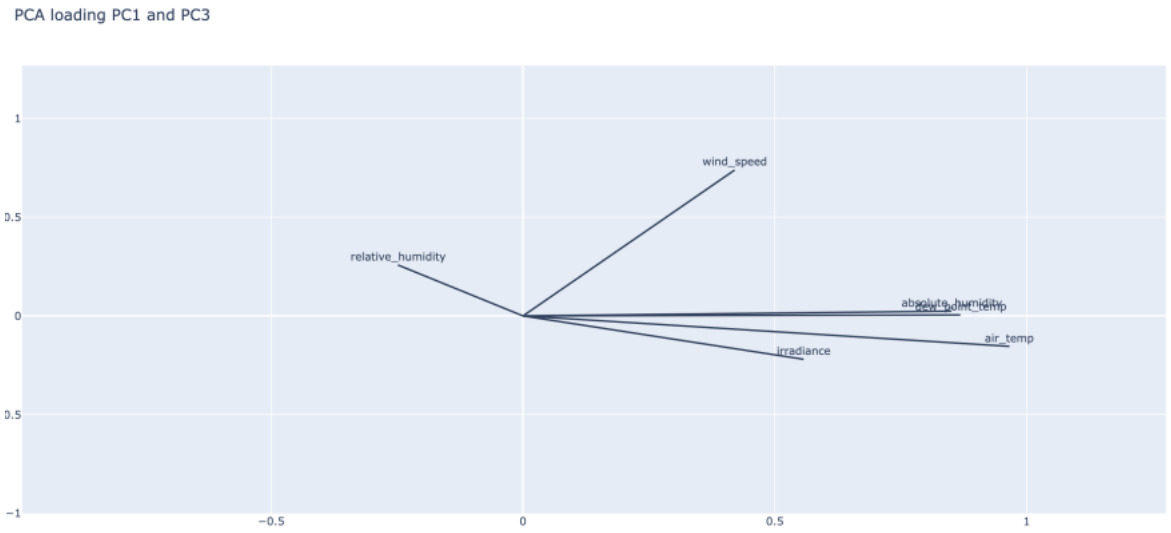


Figure 4.20: PCA loadings of PC1-3 included absolute humidity.

One conclusion would be that if absolute humidity values were used, dew point temperature values would be excluded from the database. Since precipitation and cloudiness are not available at nearby stations, relative humidity is going to be used in further analyses to have an idea about precipitation and cloudiness. Wind direction was only available from 2020 April. To observe if any relation exists in absolute humidity with wind direction, the correlation plot was produced for the 2021 year of data to represent one whole year. Figure 4.21 shows the correlation plot including wind category and absolute humidity records. The highest positive correlation was observed for wind category with absolute humidity.

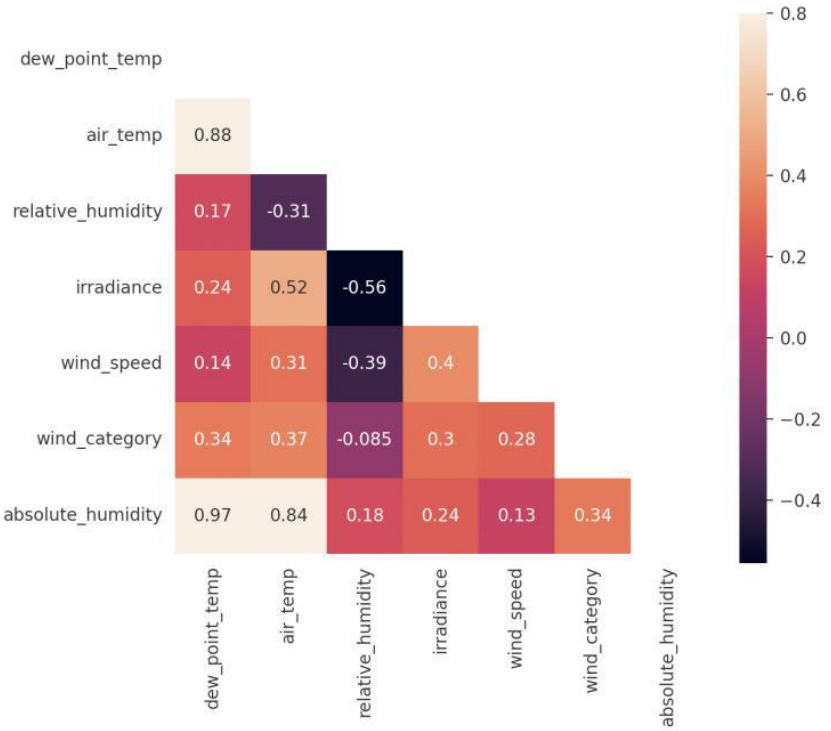


Figure 4.21: Correlation matrix with absolute humidity and wind category for 2021.

4.2 PV power analysis results

PV data analysis with other meteorological variables included starts from 4th April 2020 until the end of 2021 due to the availability of wind category variable. Power values are only available from the beginning of 2020. That is why the first 3 months in 2020 for power values are out of the investigation. Figure 4.22 shows IV2 power values with irradiance values for the analysed period. Layout and slope cause receiving less irradiance than global horizontal irradiance valu1es on east direction panels which is shown in the graph as measured irradiance. The information for receiving irradiance based on direction had been shown in the methodology section. Power values with irradiance plotting for other selected inverters IV5 and IV7 were given in Appendix D.

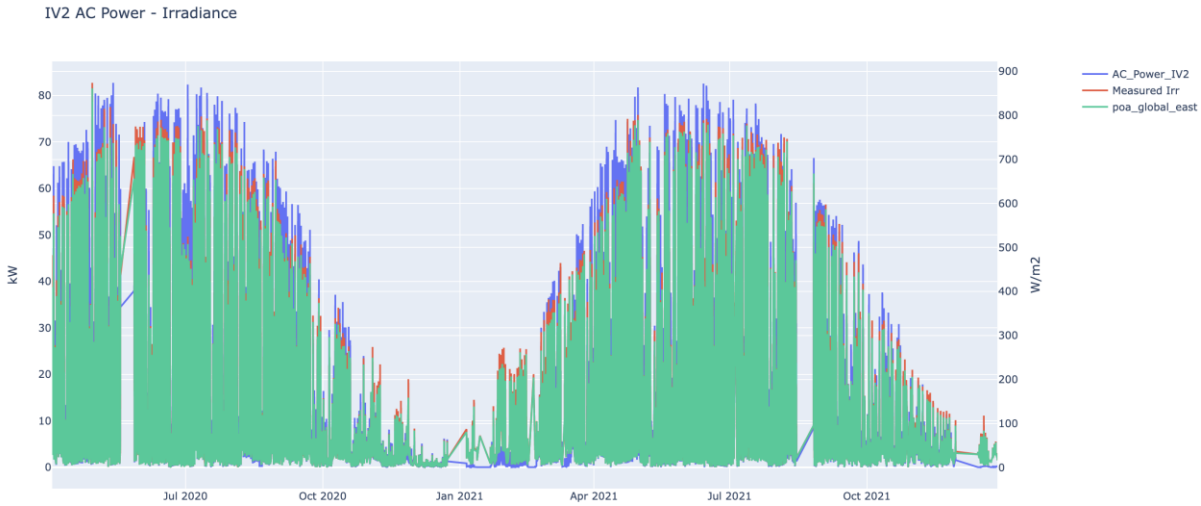


Figure 4.22: Power values for IV2 and measured irradiance values in addition to POA irradiance.

Figure 4.23 illustrates a comprehensive plot with all meteorological values and irradiance values. In the dataset, if one variable has no record for a period of time, the corresponding values for other variables were also deleted to have the same number of rows in each column.

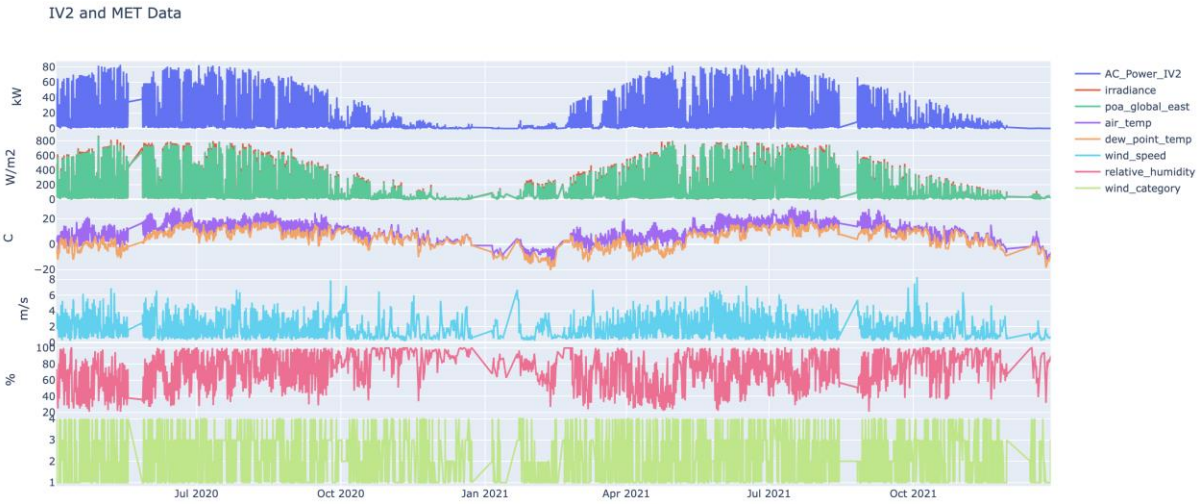


Figure 4.23: IV2 AC power values with meteorological variables.

Figure 4.24 illustrates correlation analysis for inverter 2 with other methodology variables. Correlation coefficients are shown in the figure itself. While darkest colours indicate the highest negative correlation, lightest colours describe the highest positive correlation. The highest correlation was obtained among the plane of array irradiance and irradiance values. Power output correlation with the plane of array irradiance is higher than horizontal irradiance values as expected. The second highest correlation with power output is module temperature with a 0.77 coefficient. Module temperature is a generated value that takes into account three different dependent variables, air temperature, wind speed, and the plane of array irradiance. The third highest correlation is with air temperature at 0.47. On the other hand, the highest negative correlation was observed among power output and relative humidity. Power output and solar position parameters are not going to be discussed as the main focus is weather parameters. It is important to note that only above zero sun elevation values have been taken into account which represents the periods when the PV plant is ready to produce power. Correlation analysis results for inverters 5 and 7 were given in Appendix D.

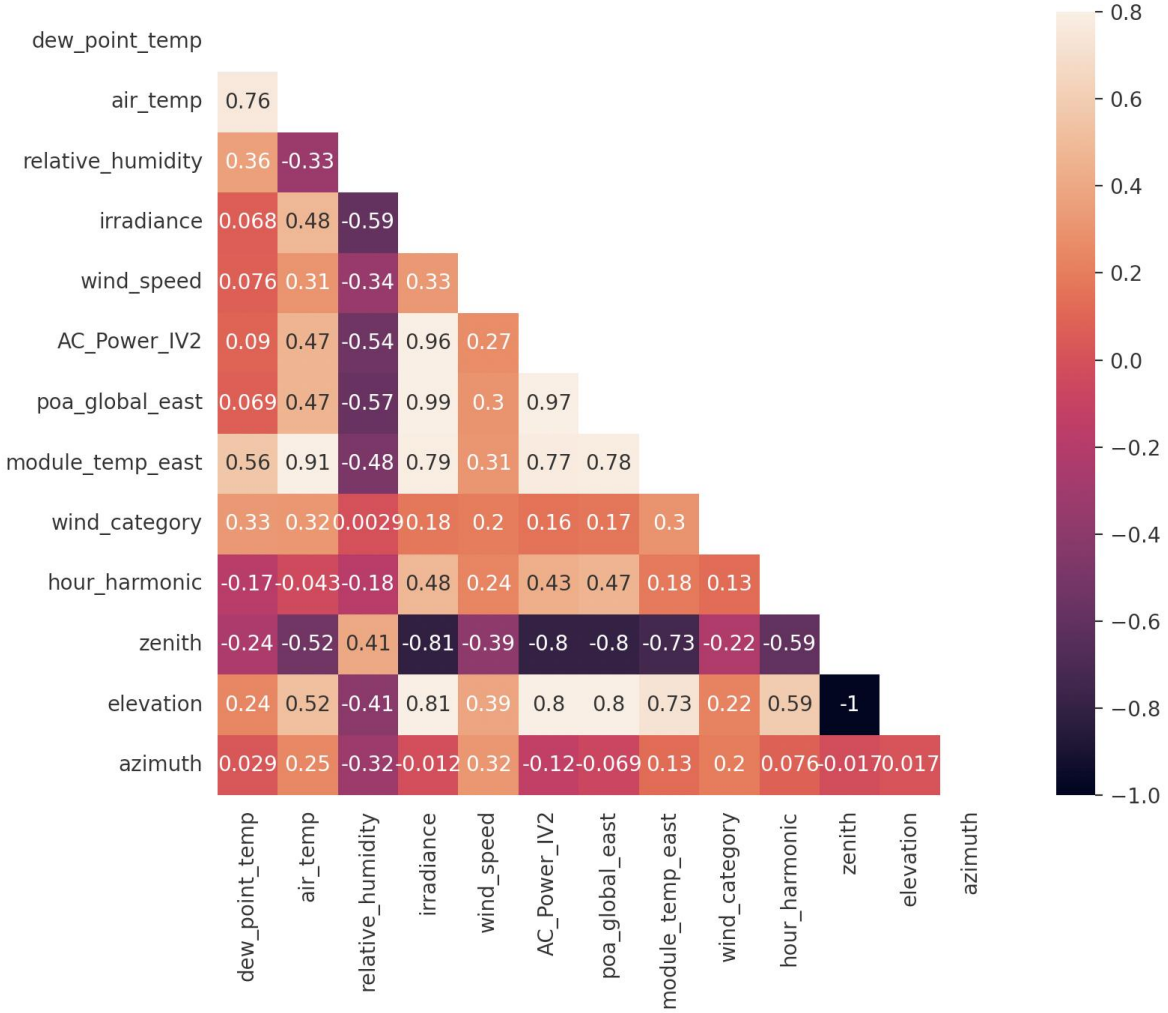


Figure 4.24: IV2 correlation analysis with other meteorological variables included.

Figure 4.25 introduces inverter 2 power output and the plane of array irradiance value graph. Dots were coloured by the corresponding air temperature. Clearly, low irradiance leads to lower PV power. It is also possible to conclude that higher PV values were recorded when air temperature was relatively low. Yellows values which belong to higher air temperature, gathered mostly at the bottom of the trend line.

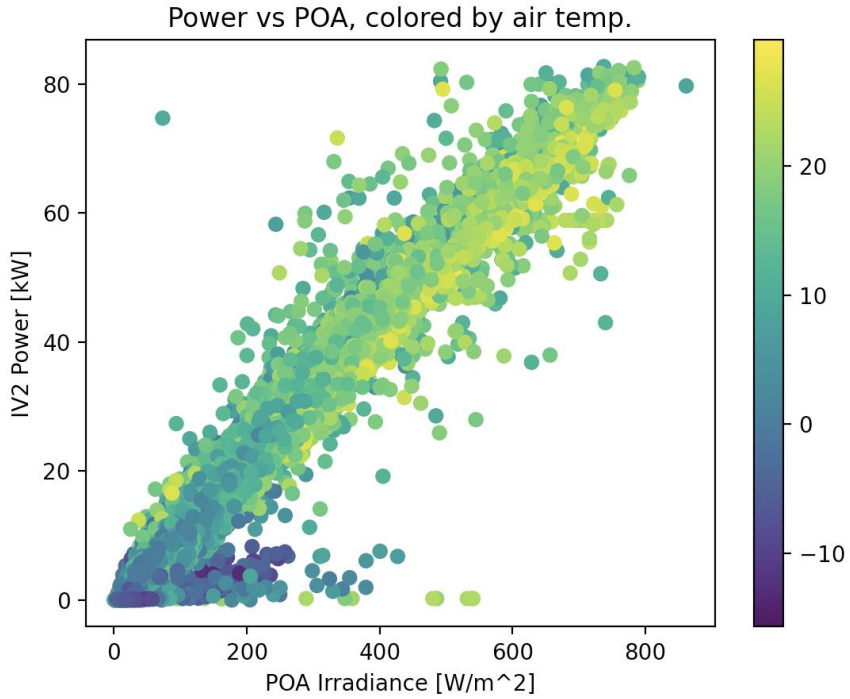


Figure 4.25: Inverter 2 power output and plane of array irradiance coloured with air temperature.

In the methodology section, the PV power output filtering method had been discussed. A performance index calculation method had also been introduced. What can be concluded from the analysis is that low elevation leads to a drop in performance index. Among the values, the bottom 10% of the performance index values were assigned as 0 and other values as 1. On the left-hand side of Figure 4.26, dots were coloured based on elevation values. The lowest PV output values were recorded in the event of low elevation and irradiance, obviously. On the right-hand side of the same figure, PV power values were filtered out based on performance index 0 and the plot was reproduced. Thus, high irradiance but low power values are eliminated. It is aimed that by introducing a performance index and filtering low performance values, noise in PV power would be eliminated. Thus, cleaned power output data may reduce the error in predictions.

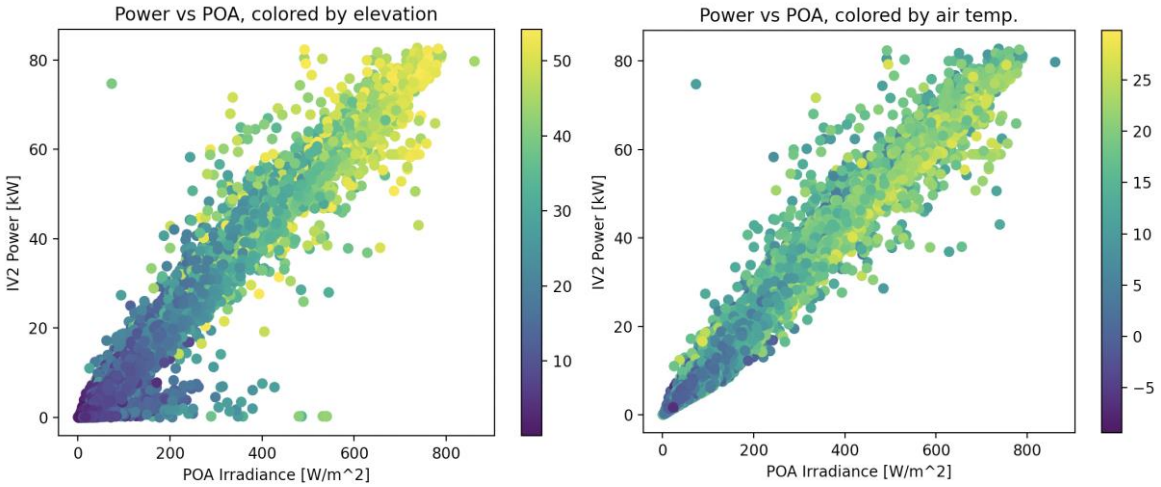


Figure 4.26: Inverter 2 power values and POA values coloured by elevation (left) and PV performance filtered graph coloured with air temperature (right).

Having had PV data, PCA analysis can be expanded with PV power data. Since the number of input variables increased, 4 PCs were determined. Figure 4.27 describes the scree plot for PCA analysis with 4 components. PC1 explains 50%, PC2 16%, PC3 11%, and PC4 9% variance in the dataset with a total of 85.44% explanation.

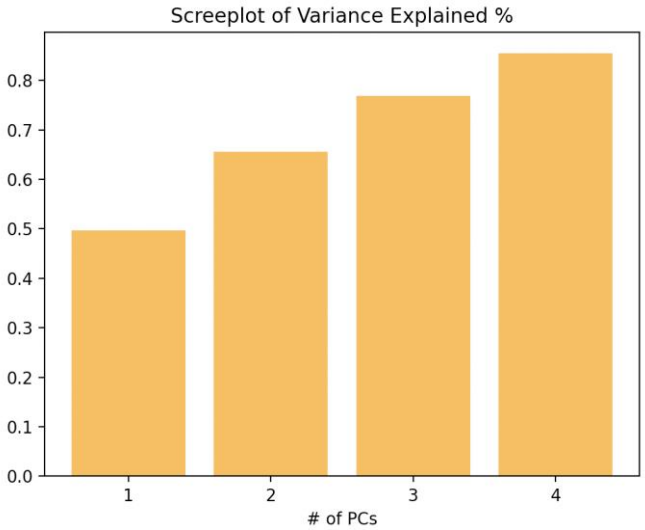


Figure 4.27: Scree plot of PCA analysis including PV power values and sun parameters.

Figure 4.28 illustrates PC scatter plots. With additional inputs, the distribution of variables is much more homogeneous. That is why variance explanations have dropped.

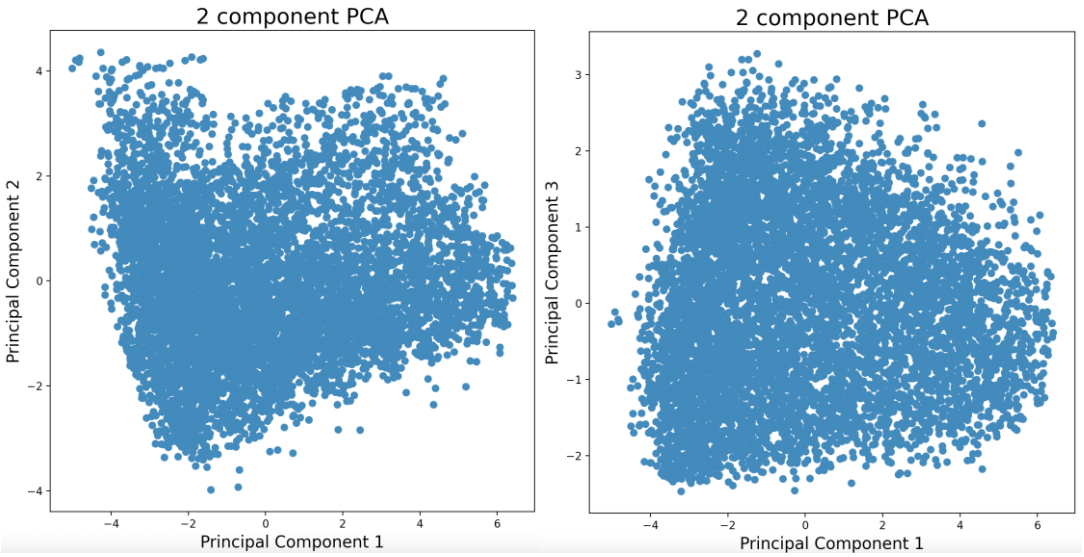


Figure 4.28: PCA components with PC1-2 (left), PC1-3 (right).

In PCA loadings, irradiance related parameters such as elevation, module temperature and POA irradiance overlap with PV power. Figure 4.29 and Figure 4.30 explain PC1-2 and PC1-3 loadings.

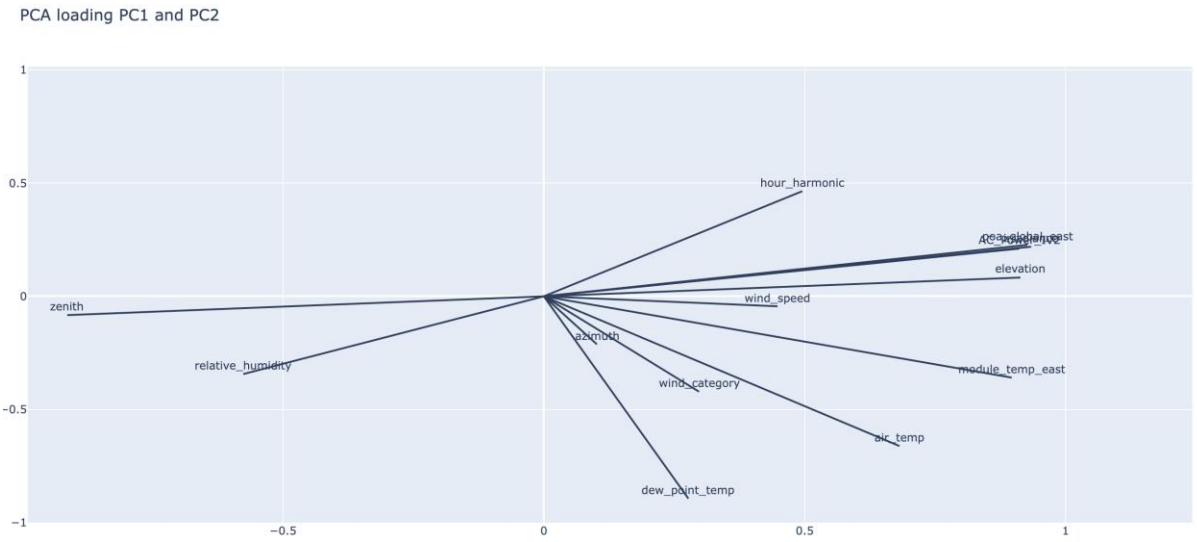


Figure 4.29: PCA loadings for PC1-2 with PV power values and sun parameters included.

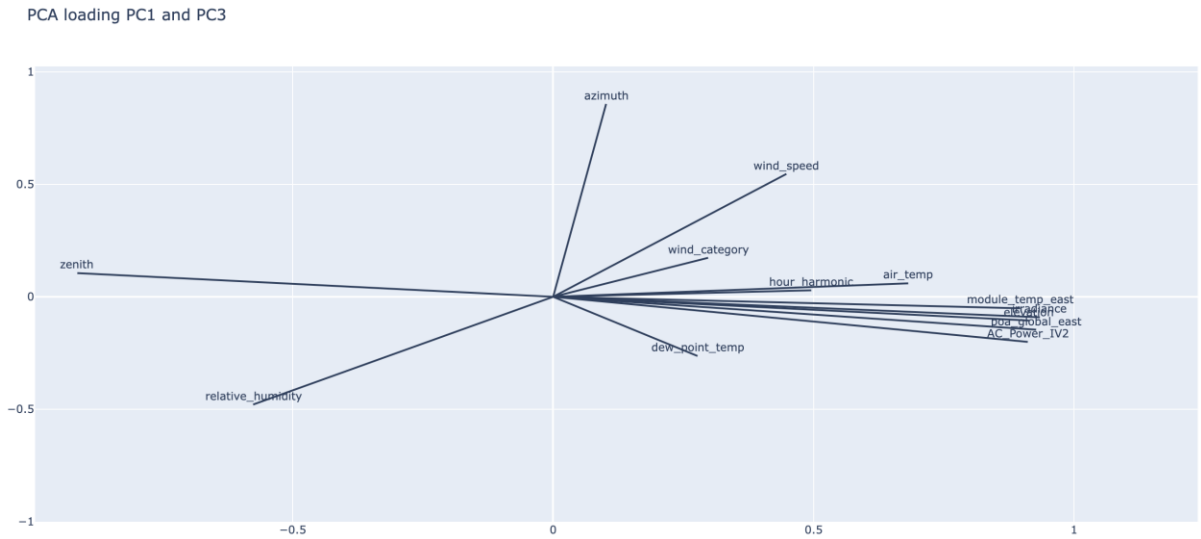


Figure 4.30: PCA loadings for PC1-3 with PV power values and sun parameters included.

4.3 PV Power output prediction case study results

In the methodology chapter, planned case studies were announced. In this part of the result chapter, each case study will be represented for both linear regression (LR) and artificial neural networks (ANNs). Different plotting styles were used. For example, short term LR graphs were plotted as scatter plots while ANN graphs were plotted as a line with value markers. The reason is that in the linear regression model, graphs also show corresponding time while ANN graphs have only index numbers. Hence, it was avoided non-value period within night time which has made plotting readable. Since ANN graphs will be introduced together with LR graphs, it is possible to identify the ANN graph time scale by comparing LR graphs.

4.3.1 Model performance on clear sky days

Consecutive prediction aims to short term prediction with relatively similar weather variables in one period. Thus, smaller data sets can achieve sufficient results and less computation time and resources are required. In this section, short term model performance on clear sky was evaluated.

Consecutive clear sky days were observed between 22nd July 2021 and 24th July 2021. The model is trained with data from 12th July 2021 to 21st July 2021. Table 4.8 shows training and test set mean and standard deviation results for each variable.

Table 4.8: Training and test set mean and standard deviation for each variable.

Variables	Mean Values - Training	Std - Training	Mean Values - Test	Std - Test
dew_point_temp	12.58	4.07	12.6	1.62
air_temp	21.91	3.96	21.93	3.92
relative_humidity	57.13	14.7	57.41	15.67
poa_global_east	364.61	238.3	354.87	241.36
wind_speed	1.79	1.03	2.09	1.12
wind_category	2.25	1.05	2.72	0.77
module_temp	30.86	8.51	30.46	8.90
hour_harmonic	0.36	0.51	0.32	0.51
zenith	59.72	16.16	62.17	16.23
elevation	30.27	16.16	27.82	16.25
azimuth	177.96	82.37	188.21	84.03

PV power training data mean value is 38.14 kW, and standard deviation is 24.61 kW. In Appendix E, the historical PV power output with irradiance values plot for the period when the model was evaluated is accessible.

Figure 4.31 describes the result of the linear regression model. In each graph, model error results were printed on top of the plot. Mean absolute error (MAE) is 2.04 kW, mean squared error (MSE) is 5.99 kW, root mean square error (RMSE) is 2.45 kW, and variance is 0.99. For the next result graphs, the same error representation approach will be used. While the LR model indicates linear regression, the ANN model shows artificial neural network model on the plotting.

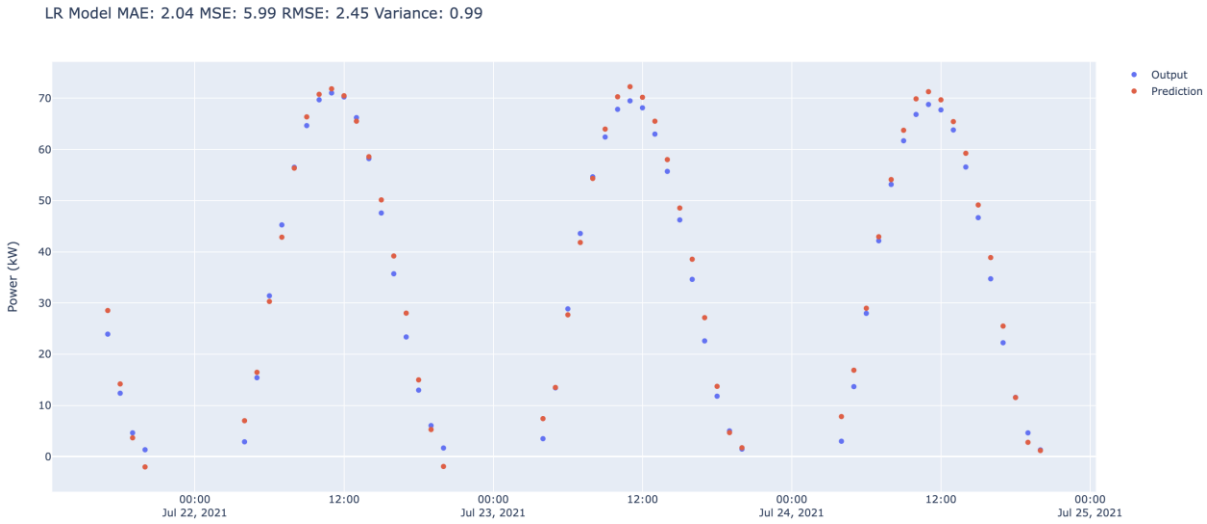


Figure 4.31: LR model prediction for clear sky days.

The LR learning curve explains the model stopped learning after 70 training data. The more training data feeds into the model, the model no longer learns.

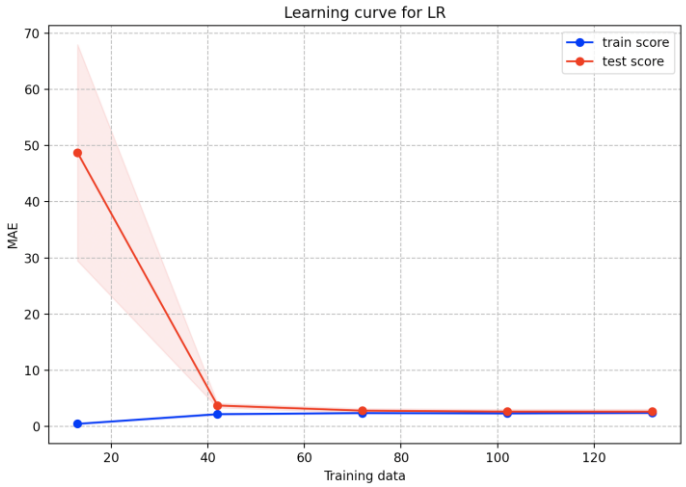


Figure 4.32: LR model learning curve for clear sky days.

The simulation was repeated for the ANN model. Figure 4.33 shows the ANN prediction plot which achieves relatively better prediction with fewer errors.

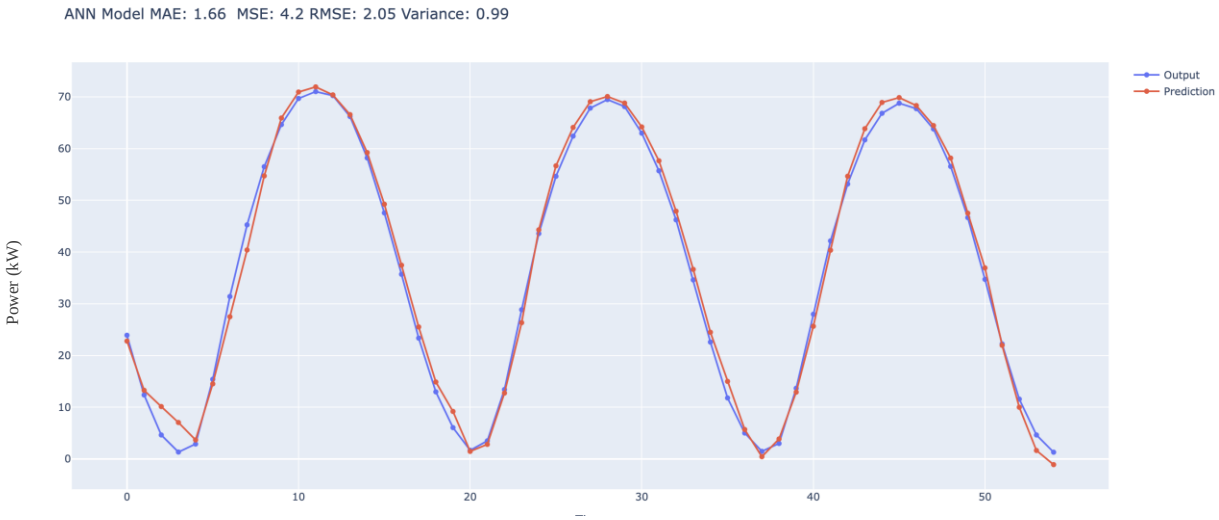


Figure 4.33: ANN regression prediction for clear sky days

In the ANN model, different model configurations were used due to the small data set. It is seen from the RMSE error graph that the model achieves its best performance with 500 epochs.

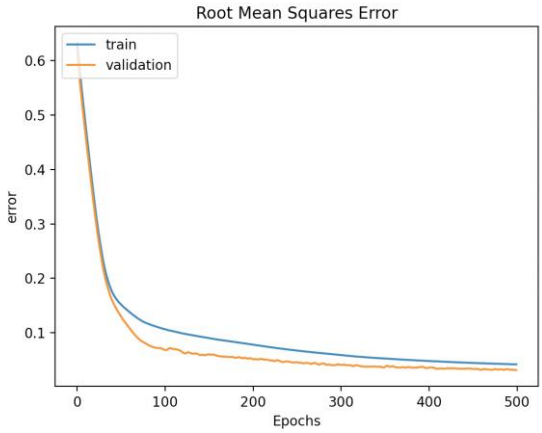


Figure 4.34: ANN learning curve for clear sky days.

4.3.2 Model performance on clear sky days without measured irradiance input

It is not always possible to access the latest measured irradiance values for PV power output prediction. That is why the model performance was evaluated without measured irradiance values. Since module temperature input was derived from measured irradiance values, this variable was also excluded. Instead, calculated clear sky irradiance values were fed into the model. Figure 4.35 illustrates LR model results without measured irradiance values but calculated clear sky irradiance values. Higher errors were obtained compared to prediction with measured irradiance values.

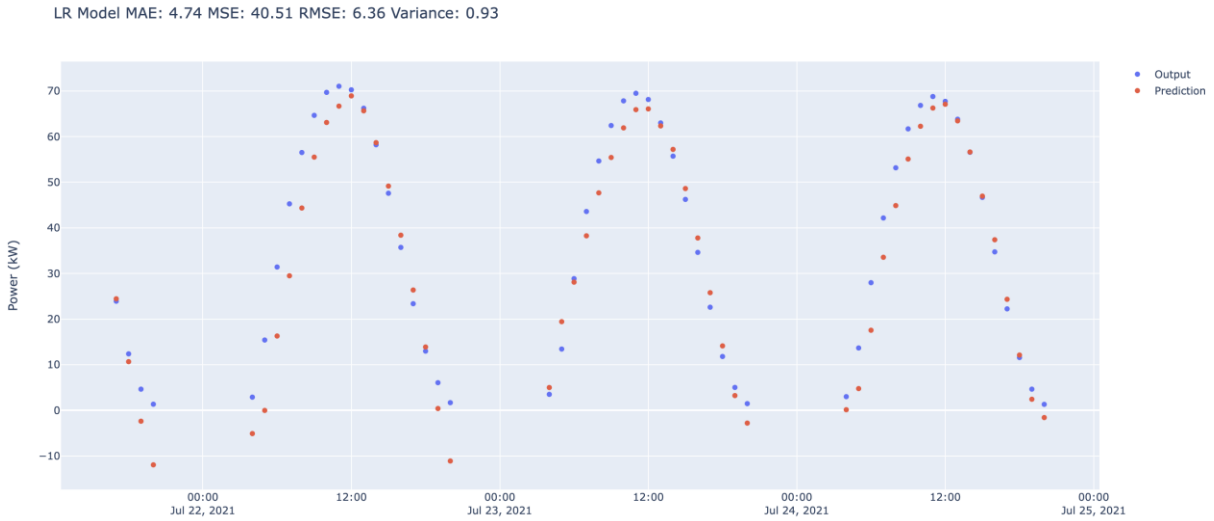


Figure 4.35: LR model prediction for clear sky days without measured irradiance values.

ANN model achieves the same prediction with lower error but with a low performance compared to measured irradiance values included. The model completely fails without any irradiance values either clear sky or measured irradiance. The figure is given in Appendix E.

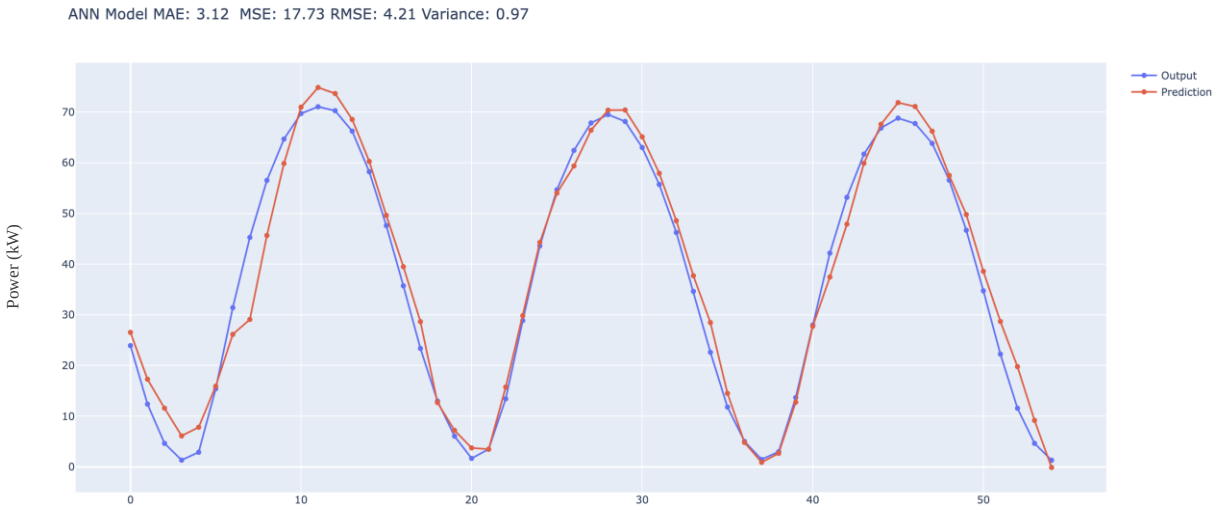


Figure 4.36: ANN model prediction for clear sky days without measured irradiance values.

4.3.3 Year based training and test sets

The PV plant has been in operation since 2020. It is aimed in this part to evaluate prediction performance for 2021 with the 2020 year of data training. In the methodology section, it was observed that 4 years of meteorological variable correlations are not quite different from 2021-year data. Hence it can be concluded that the prediction of PV power from 2021 historical data may be a reliable source for future predictions in 2022. The models were trained by 2020 historical PV power output and meteorological variables and tested in 2021. The training period starts from 1st April 2020 when the wind direction variable started to be recorded until 01 April 2021. On the test set side, December was not included as there are only a few PV power data

available. Table 4.9 includes training and test sets mean and standard deviation values. Power values training set mean value is 24.94 kW and standard deviation is 22.99 kW.

Table 4.9: Training and test data variable statistics.

Variables	Mean Values - Training	Std - Training	Mean Values - Test	Std - Test
dew_point_temp	4.68	7.4	9.54	5.31
air_temp	11.7	7.3	15.09	6.7
relative_humidity	65.9	21.93	72.2	18.45
poa_global_east	233.4	210.08	203.15	195.91
wind_speed	2.01	1.31	1.8	1.17
wind_category	2.23	1.03	2.32	0.98
module_temp	17.3	10.42	20.02	10.18
hour_harmonic	0.5	0.46	0.55	0.43
zenith	65.87	15.2	67.42	14.6
elevation	24.12	15.2	22.58	14.6
azimuth	179.63	70.34	180.32	70.41

Figure 4.37 illustrates 2021 year of data prediction for LR. As it is seen from the figure, predictions fall below zero on some days. Furthermore, there are some days when the model was not able to capture peak values.

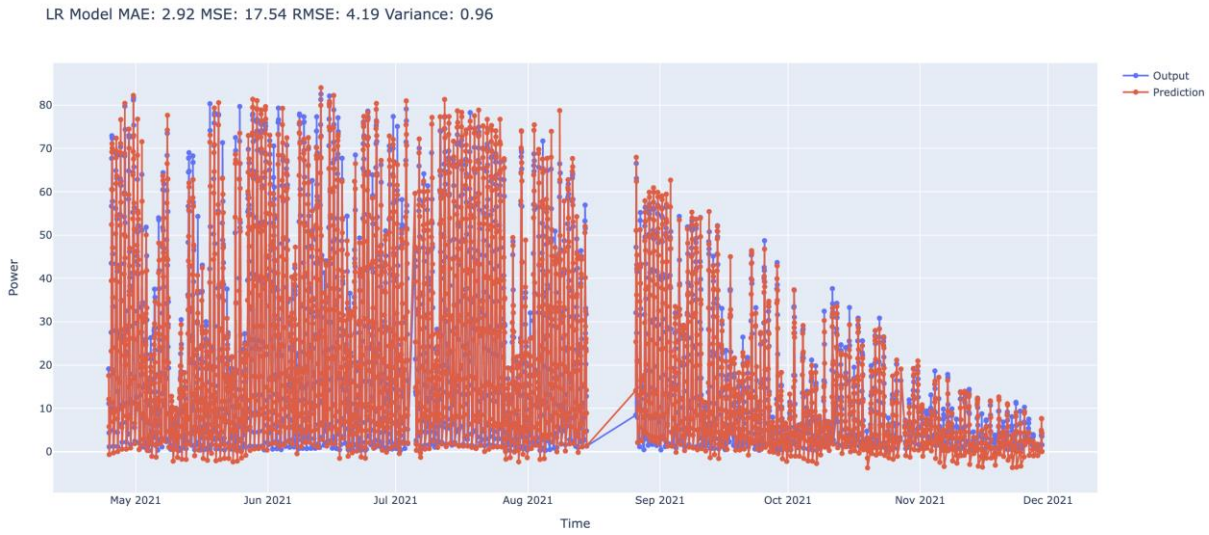


Figure 4.37: LR model prediction for 2021-year data from April.

For the same period, the ANN model did a better job of capturing peak values. Predictions rarely fell below zero as it is seen in Figure 4.38. As a result, ANN has low errors overall. Different ANN configurations achieved better results. For this prediction, 100 epochs, 64 and 32 dense were provided to the model.

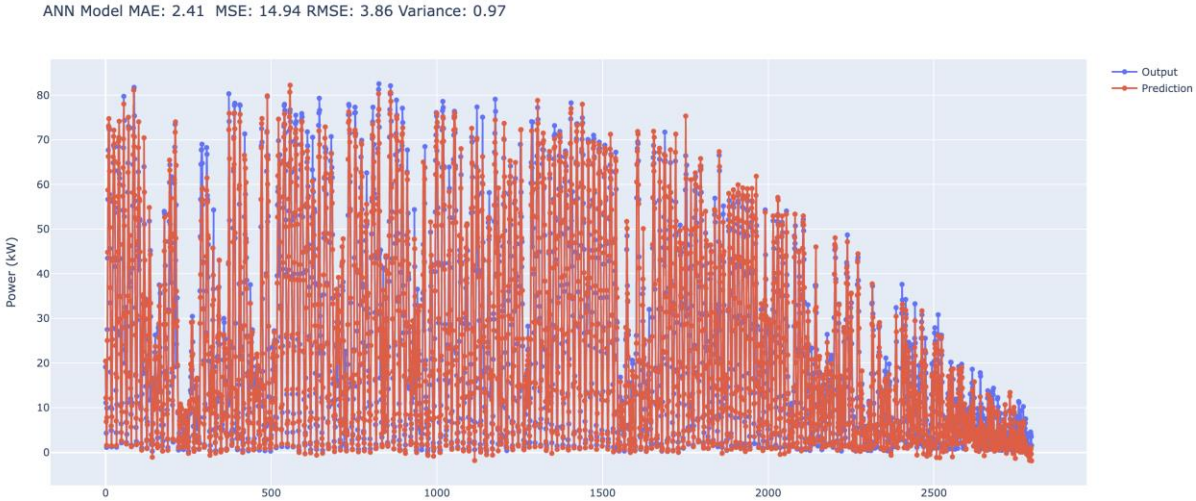


Figure 4.38: ANN model prediction for 2021-year data from April.

The learning curve for LR shows that as the model predicts better with more training sets, training errors get higher values indicating that the data set gets complicated as it is observed in Figure 4.39.

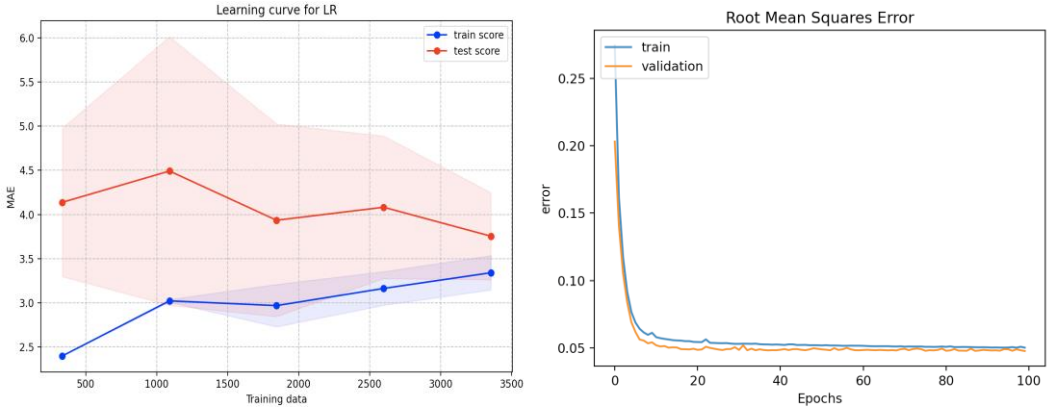


Figure 4.39: Learning curves for LR (left) and ANN (right).

4.3.4 Training on 2020/2021 data and meteorological variable selection

In this section, it is aimed to train the model with as much as possible data and test them during different periods to observe the overall performance of models. One of the expectations was to evaluate model performance on the large dataset by letting the models learn from more datasets. In addition, this case study will show how the model might perform on forecasted meteorological data. Furthermore, the meteorological variable's impact on the overall result was examined.

The models were trained from 01/04/2020 to 21/07/2021 and tested on consecutive clear sky days. By doing this, model performance was compared with the trial in the 4.3.1 chapter. Figure

4.40 shows ANN model performance. Since the model has more training data, the prediction was easy and the model overfitted on the test set. The model was run by 50 epochs.

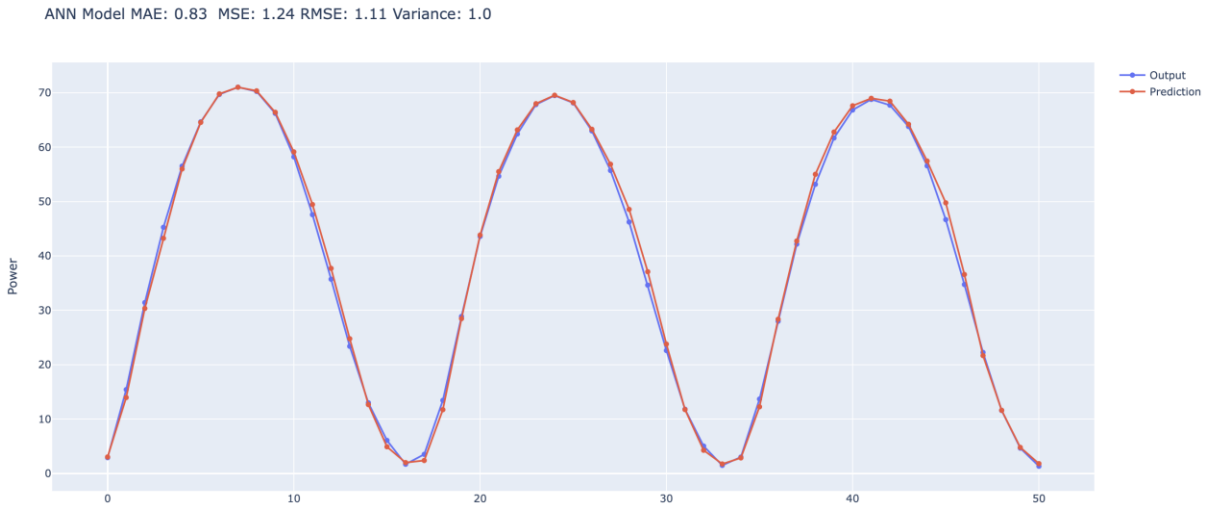


Figure 4.40: ANN model output training with 2020/2021 and testing on consecutive clear sky days.

However, LR produced poor results on the same test set as it is shown in Figure 4.41.

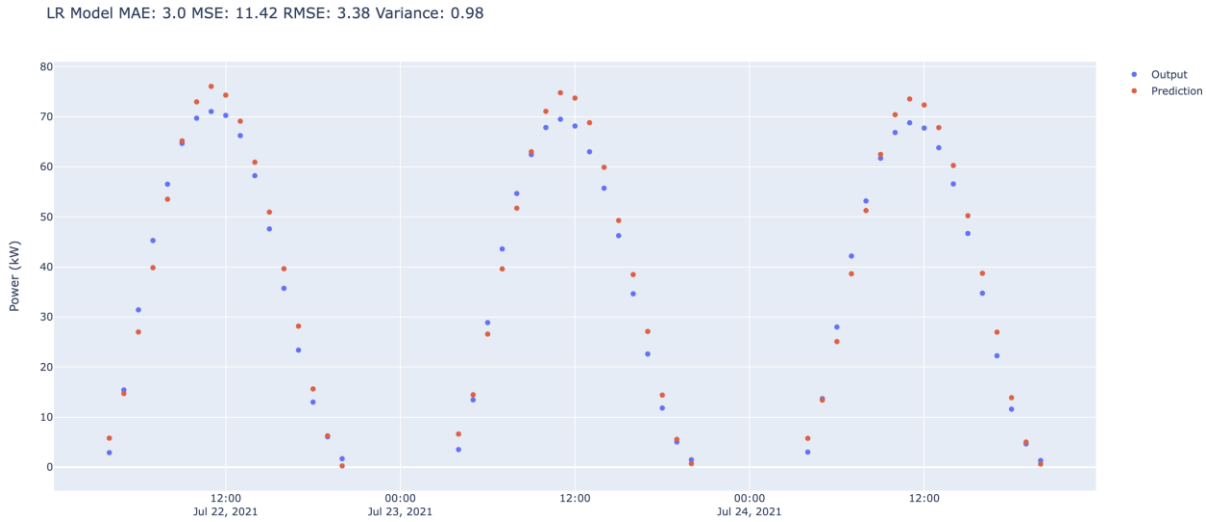


Figure 4.41: LR model output training with 2020/2021 and testing on consecutive clear sky days.

The trial was done on fluctuating PV power output days. Figure 4.42 shows the days are going to be tested with PV power and POA values.

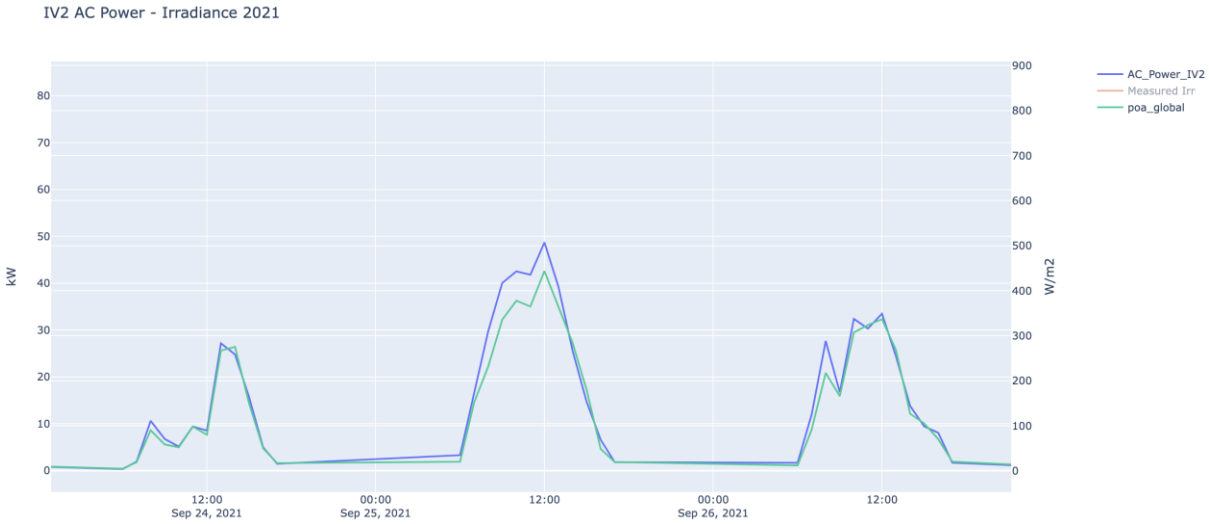


Figure 4.42: IV2 and poa_global values for selected days.

Results are shown in Figure 4.43 and Figure 4.44 for ANN and LR models. ANN model was run by 50 epochs, and 64 batches. As a result, while ANN predicts slightly better than the LR model, both models failed to capture the power value drop on 27th September 2022 at 11:30. Some comments have been made on this issue in the discussion section.

Figure 4.43: ANN model output training with 2020/2021 and testing on fluctuating power output days.

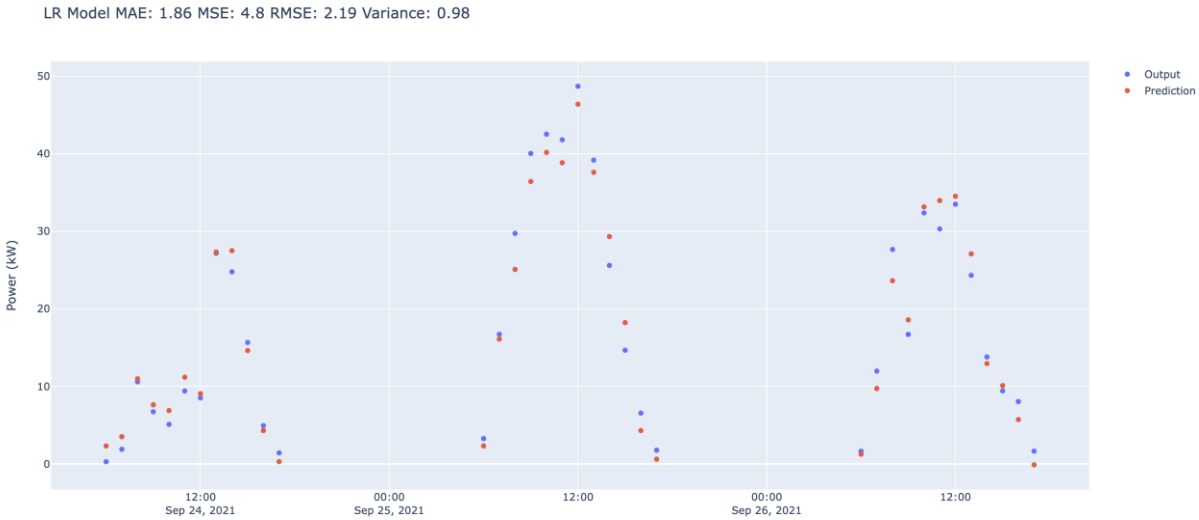


Figure 4.44: LR model output training with 2020/2021 and testing on fluctuating power output days.

In most scenarios not all meteorological variables are accessible. For example, without wind parameters, it is still possible to predict power outputs. In this part, it is explored how meteorological variables affect PV power output prediction. All trials were done on the same test set period which is from 28th August 2021 to 20th October 2021. The training period starts from April 2020 until the beginning of the test set. Figure 4.45 illustrates prediction results for the ANN model with all parameters included. For the ANN model, with all parameters included, prediction results for the test set were given in the first row in Table 4.10. Likewise, Figure 4.46 describes prediction results for the LR model with all parameters included and prediction errors were given in Table 4.10.

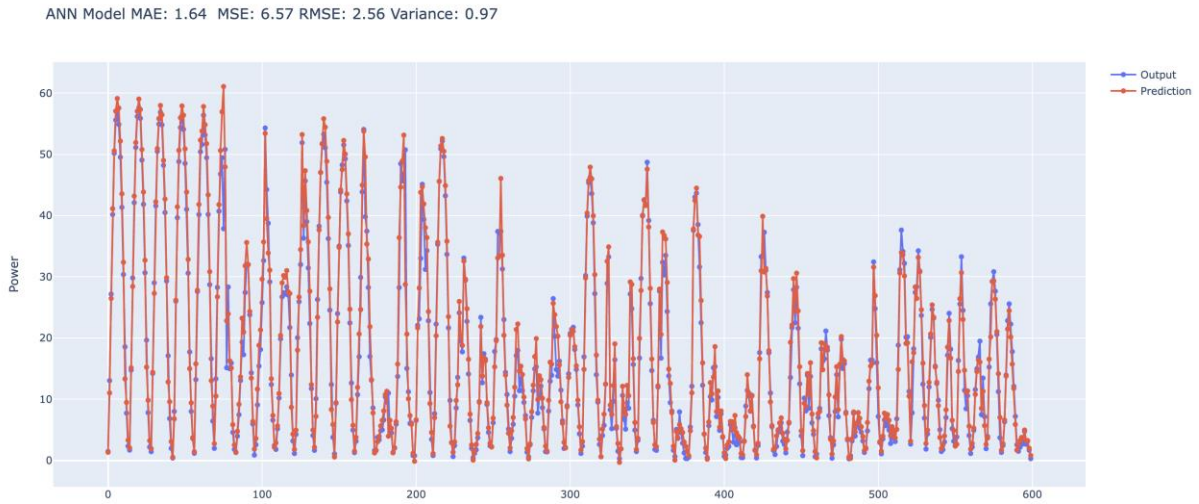


Figure 4.45: ANN model prediction results with all parameters included.

The best predictions with the ANN model were obtained with hyperparameters in which 64 and 32 dense, 32 batch sizes and 80 epochs.

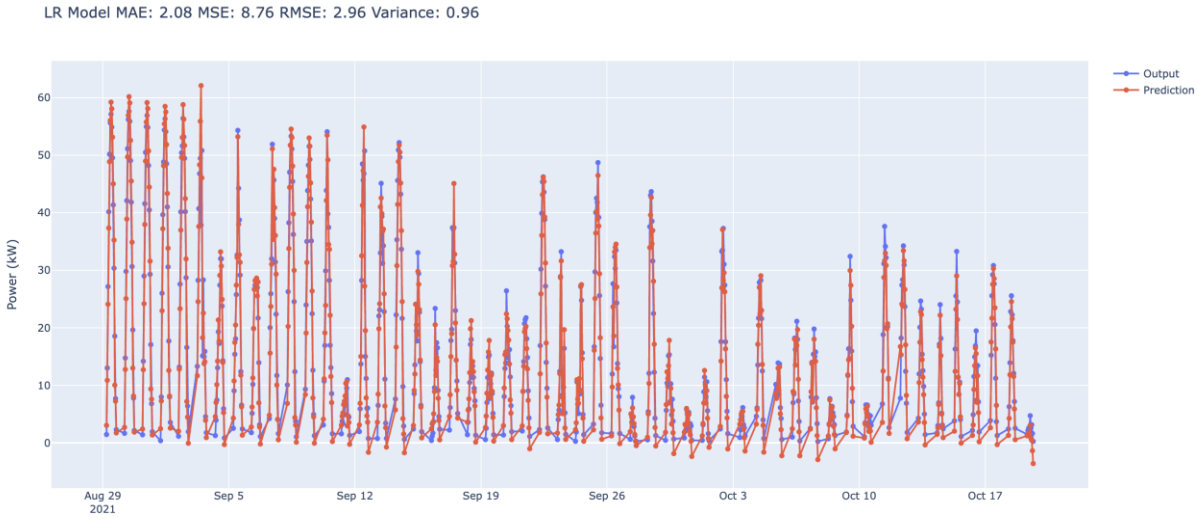


Figure 4.46: LR model prediction results with all parameters included.

The learning curves for LR and ANN in Figure 4.47 describe that while the LR model continues to learn as more training data is fed in, the ANN model with given parameters makes an easier prediction on the test set.

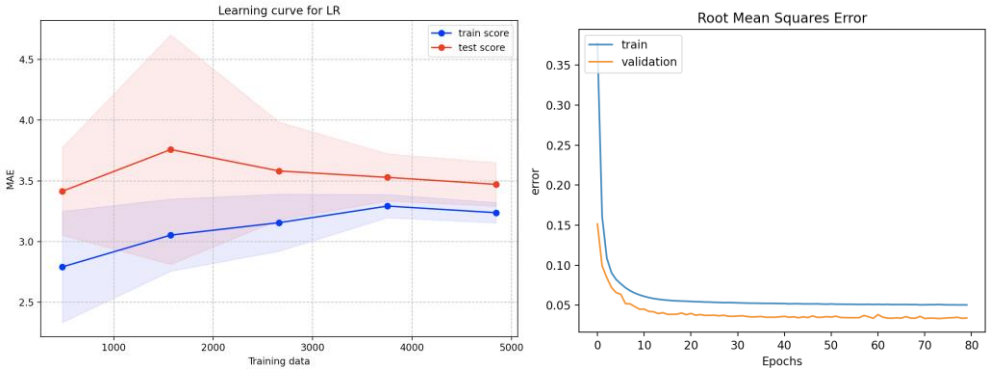


Figure 4.47: Learning curves for LR (left) and ANN (right) with all parameters included.

More broadly explanations for Table 4.10 is that excluded variables indicate all other parameters included but only the stated variable was not taken into account. There are other cases such as only the selected variable's effect was explored. ANN parameters were simplified in the process of trial of only POA irradiance with 10 epochs and 16 and 8 dense.

Table 4.10: Meteorological variable selection for ANN model's error and variance values.

Models	ANN Model Results				LR Model Results			
Errors / Variance	MAE*	MSE*	RMSE*	R2	MAE*	MSE*	RMSE*	R2
All Parameters	1.64	6.57	2.56	0.97	2.08	8.76	2.96	0.96
wind_speed excluded	1.78	7.17	2.68	0.97	2.04	8.64	2.94	0.96
wind_speed and wind_category excluded	1.65	7	2.65	0.97	2.04	8.64	2.94	0.96
relative_humidity excluded	1.82	7.66	2.77	0.97	2.09	8.82	2.97	0.96
air and module_temperature excluded	1.74	7.05	2.65	0.97	2.12	8.92	2.99	0.96
Only air_temp, POA irradiance, and sun parameters included	1.63	6.66	2.58	0.97	2.11	8.86	2.98	0.96
Only air_temp, and POA irradiance included	1.8	9.14	3.02	0.96	1.97	9.51	3.08	0.96
Only air_temp, POA irradiance, and relative_humidity included	1.78	9.61	2.93	0.96	1.96	9.27	3.04	0.96
Only POA irradiance included	1.95	10.25	3.2	0.96	1.9	9.32	3.05	0.96
Only MET variables included, POA irradiance and sun parameters excluded	10.24	184.77	13.59	0.23	11.6	206.3	14.26	0.14
Only MET variables and sun parameters included but POA irradiance excluded	5.3	54.6	7.39	0.77	7.31	82.13	9.06	0.66

* Values are in kW.

Due to small differences between errors for different trials in the same model, it is difficult to assess variable effects on the model output. That is why some cases were dug in with a closer look.

4.3.4.1 Wind Direction Effect

Figure 4.48 demonstrates metrological variables and PV power output for a selected time scale. There was one specific date, 18th July, for wind direction was on 1 category label which represents wind from the north. This information has shown in the red box in the figure. In the corresponding period, PV power output peaked as it can be seen in Figure 4.48 and Figure 4.49. The trained period from 2020 to July 2021 was tested from July 7th to 29th.

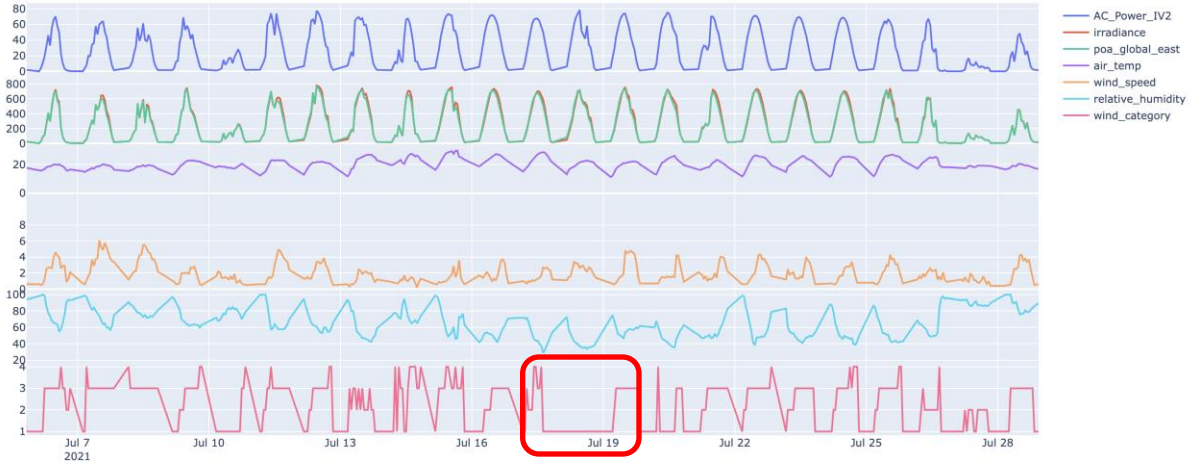


Figure 4.48: Meteorological variables and power output for wind direction analysis.

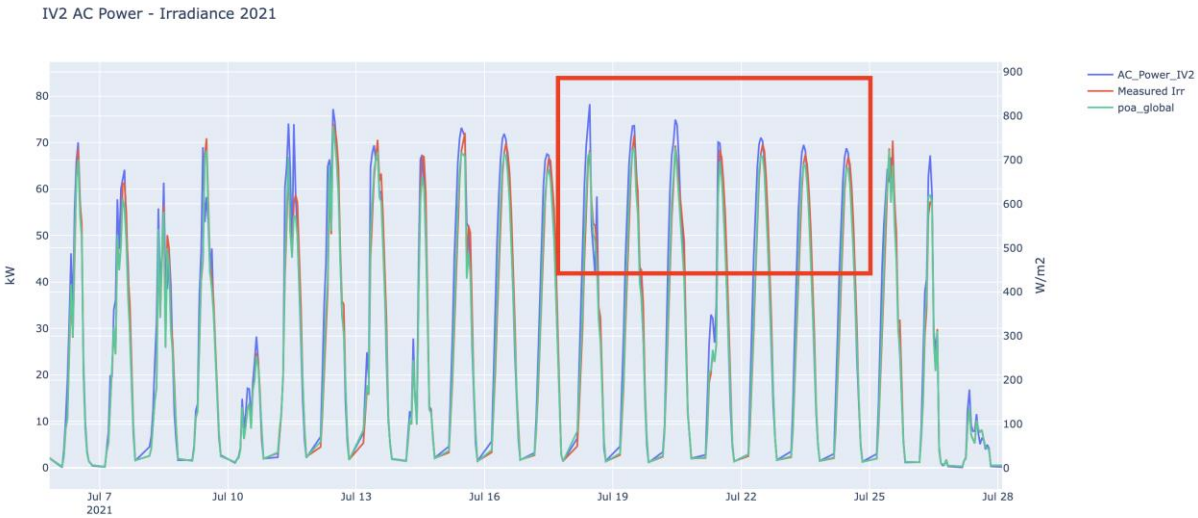


Figure 4.49: PV power output and irradiance values for wind direction effect analysis.

When the wind direction category was excluded from inputs of the model, ANN and LR models fails to catch the peak power output value. A comparison for the ANN model before all inputs were included for the given period and after the wind direction variable was excluded from the input is given in Figure 4.50 and Figure 4.51. Clearly, after wind direction data was eliminated, prediction falls for the peak value which was achieved at a lower number. The difference is emphasised with red circles.

Figure 4.50: ANN model result with all variables included within wind direction analysis period.

Figure 4.51: ANN model result without wind direction variable.

The same relation was observed for IV5 and IV7 which are on different roofs with different layouts. Figure 4.52 and Figure 4.53 describes IV5 and IV7 result, respectively. In addition, the LR model has also detected the same difference.

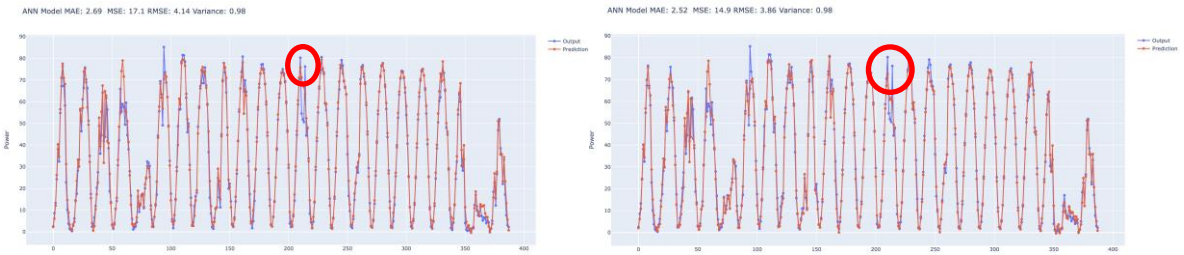


Figure 4.52: ANN model results comparisons with (left) and without (right) wind direction variable for IV5.

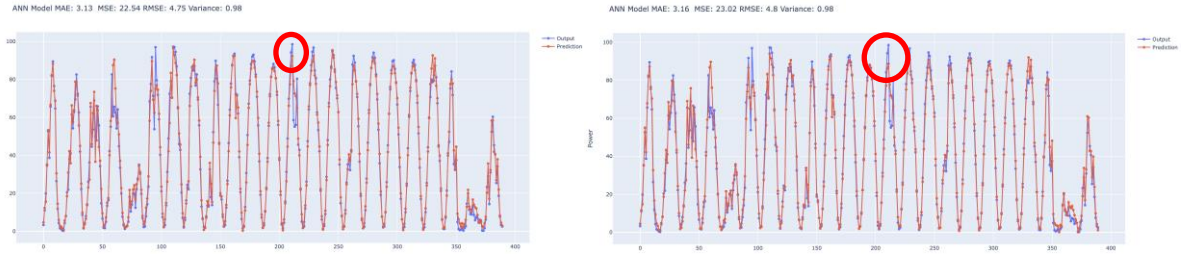


Figure 4.53: ANN model results comparisons with (left) and without (right) wind direction variable for IV7.

4.3.4.2 Relative Humidity Effect

The relative humidity effect on the model was explored for the period from 29th May to 19th June. Figure 4.54 illustrates the period that relative humidity has an increasing trend for a certain period and it is shown with a red box. When relative humidity variable was eliminated from the database, the ANN model predicted power values at a lower value for the time where relative humidity was relatively high between 9th and 10th June. Likewise, similar prediction results were obtained with the LR model.

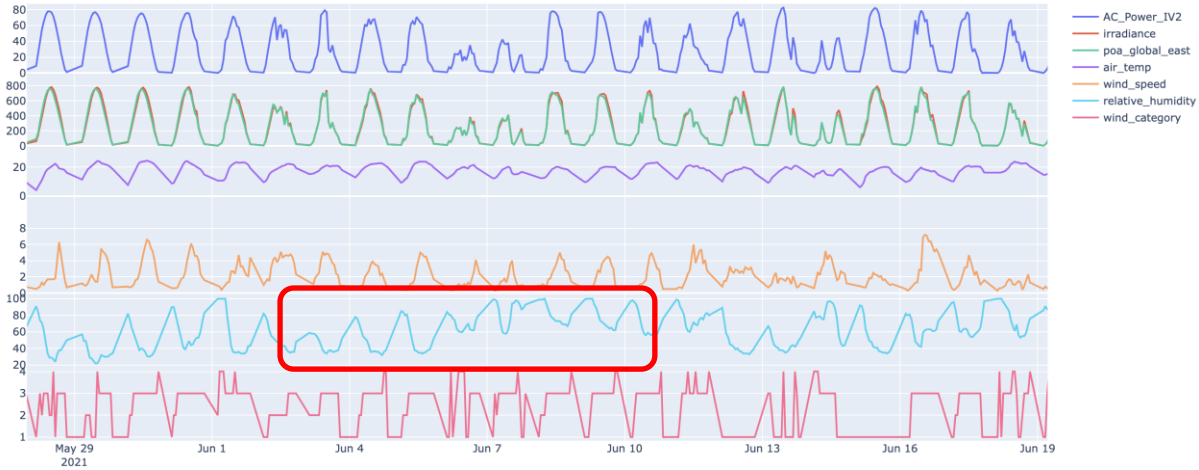


Figure 4.54: Meteorological variables and power output for relative humidity analysis.

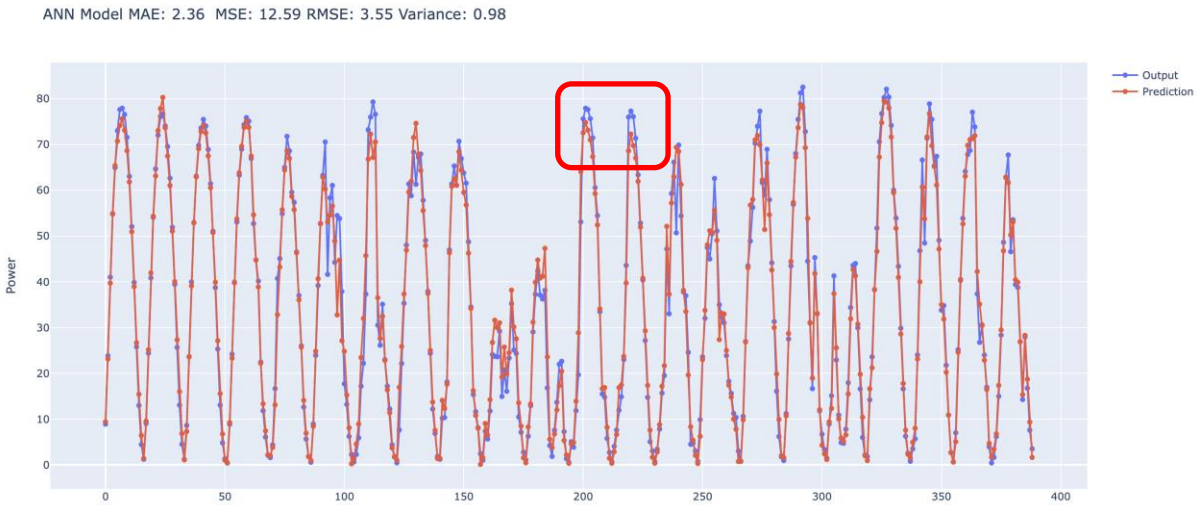


Figure 4.55: ANN model result with all variables included within relative humidity analysis period.

Figure 4.56: ANN model result without relative humidity variable.

4.3.4.3 Dew Point Temperature Effect

Relative humidity effect on the model was explored for the period from 5th May to 24th June. Figure 4.57 describes the period that dew point temperature fluctuations specifically from July 13th to 15th.

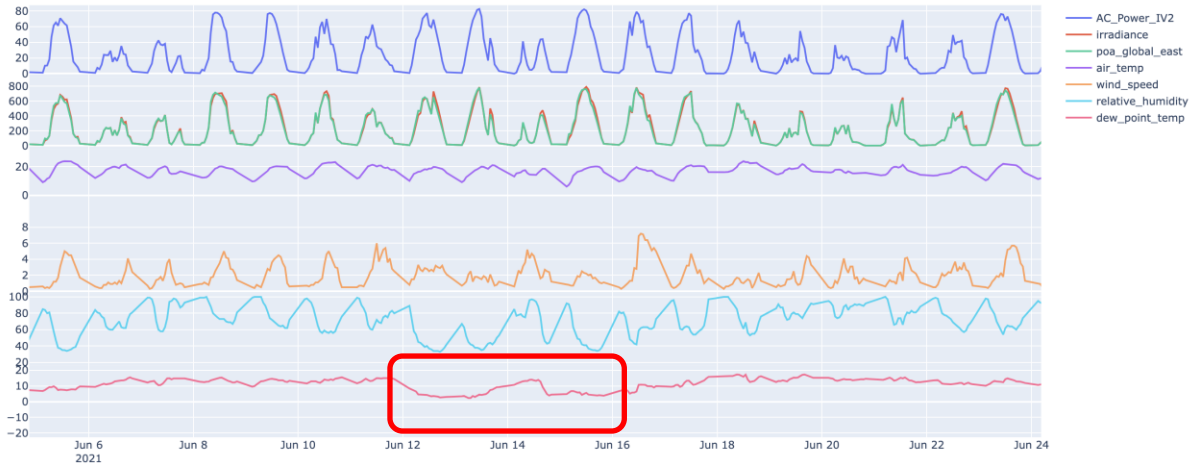


Figure 4.57: Meteorological variables and power output for dew point temperature analysis.

When dew point temperature variable was eliminated from the database, predictions were the same and did not change for ANN and LR models. Figure 4.58 and Figure 4.59 describe ANN model predictions with a red box emphasised for dew point temperature fluctuations period.

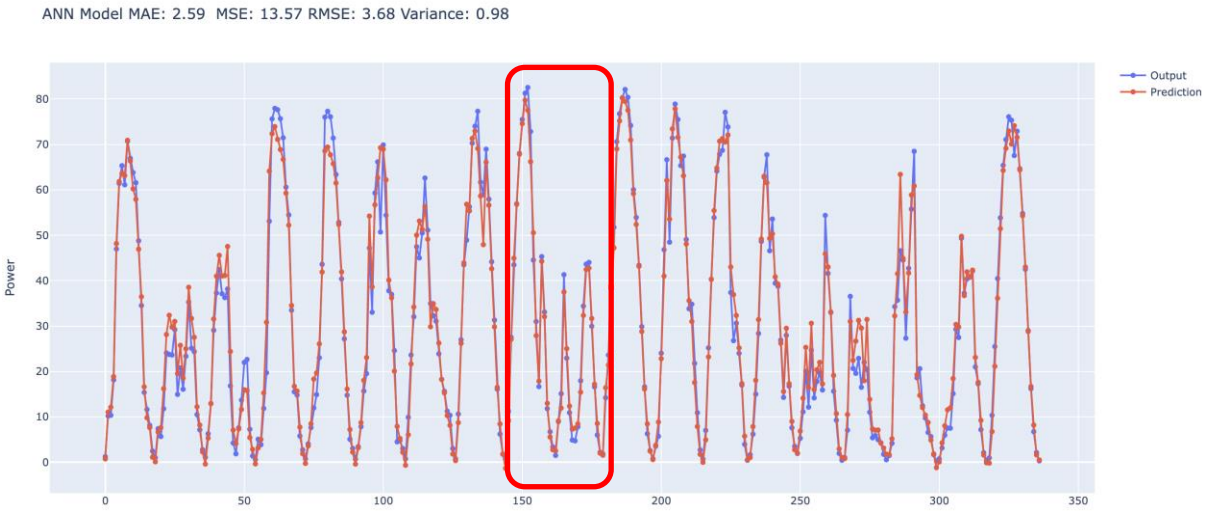


Figure 4.58: ANN model result with all variables included within dew point temperature analysis period.

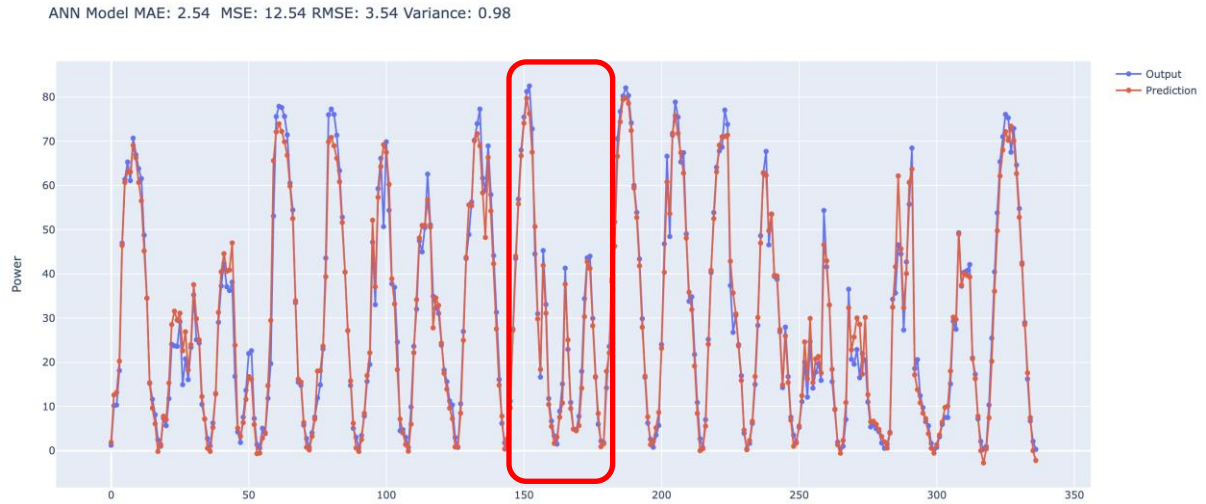


Figure 4.59: ANN model result without dew point temperature variable.

4.3.4.4 Wind Speed Effect

Figure 4.60 illustrates PV power output and meteorological variables for wind speed analysis and the red box focuses on the period when there were wind speed fluctuations from June 13th to 17th. Since the module temperature variable is a function of wind speed, this variable was excluded before the wind speed variable was eliminated from the database. Predictions without module temperature are shown in Figure 4.61. The model was run without wind speed and module temperature; however, the effect of wind speed was not observed in Figure 4.62.

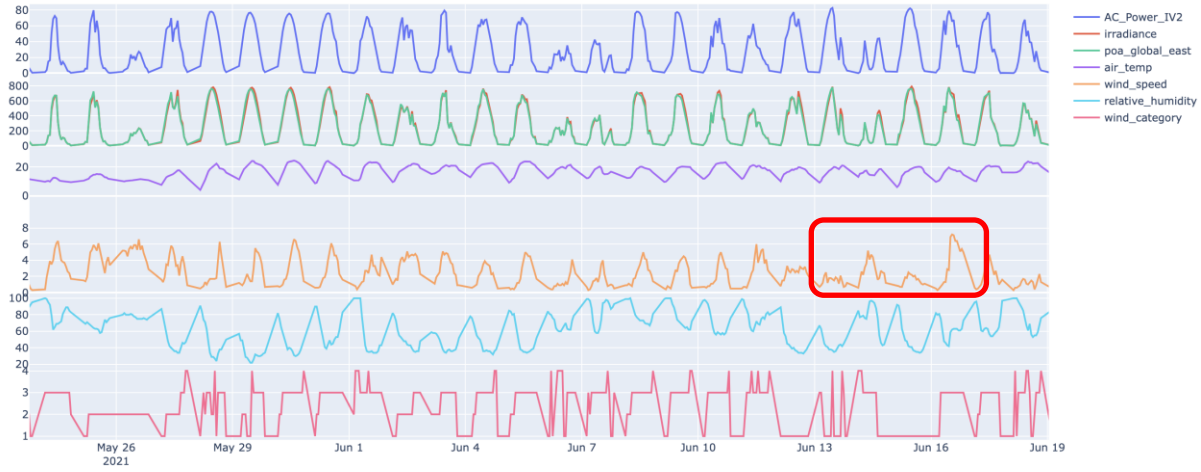


Figure 4.60: Meteorological variables and power output for wind speed analysis.

Figure 4.61: ANN model result with all variables included except module temperature.

Figure 4.62: ANN model result with all variables included except module temperature and wind speed.

Clearly, this result was unexpected. It is a scientific phenomenon that wind speed is an important parameter for PV power output. A different time period was investigated for the wind speed case. First, a period when there is an obvious impact of wind speed on power output was determined. Figure 4.63 shows the plane of array and IV2 power output line graph on 11th July 2021. The second peak in power value occurred at low irradiance value compared to the first peak. Figure 4.64 illustrates the same period in two parallel red lines with other weather variables and emphasised wind speed in the red box. A steep increase was observed for the wind speed during the second peak occurrence in power output.

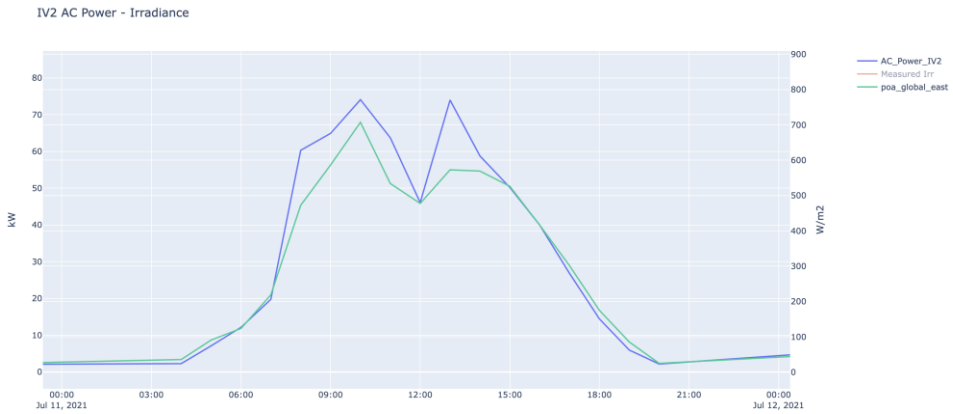


Figure 4.63: IV2 Power output and the plane of array irradiance for the wind speed case.

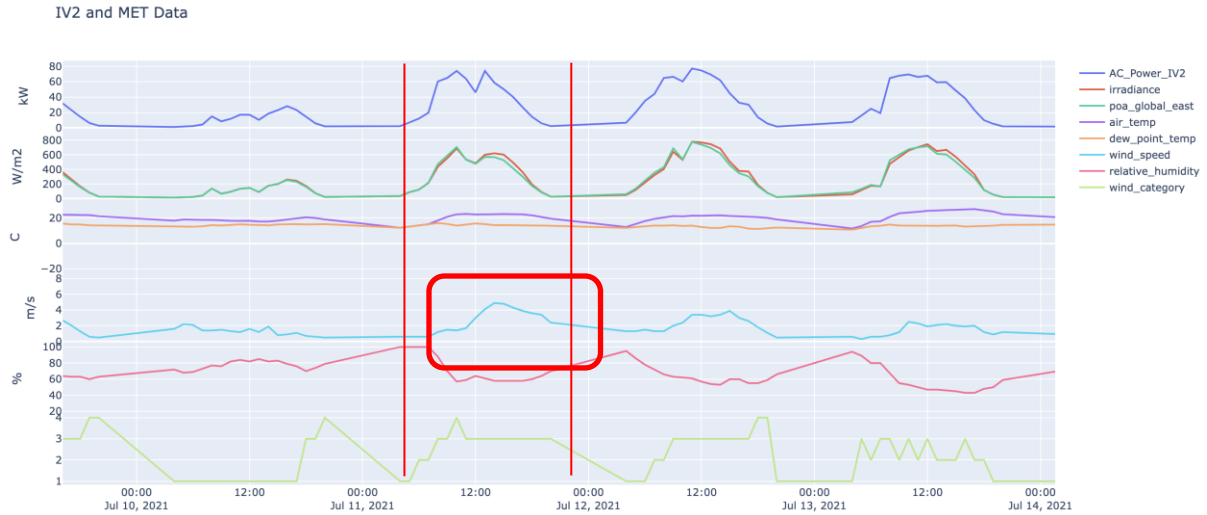


Figure 4.64: IV2 and weather parameters for the wind speed case.

When the model was trained with the same time period as previous examples, the second peak was not captured. This time the model was trained for a shorter time period and results were shown in Figure 4.65 with wind speed included and Figure 4.66 with excluded wind speed parameter. The wind speed included graph predicts higher values for the rest of day compared to Figure 4.66. The red line exceeds the blue line. Model accuracy is low for the wind speed excluded case. Longer training periods make the model count on irradiance values more.

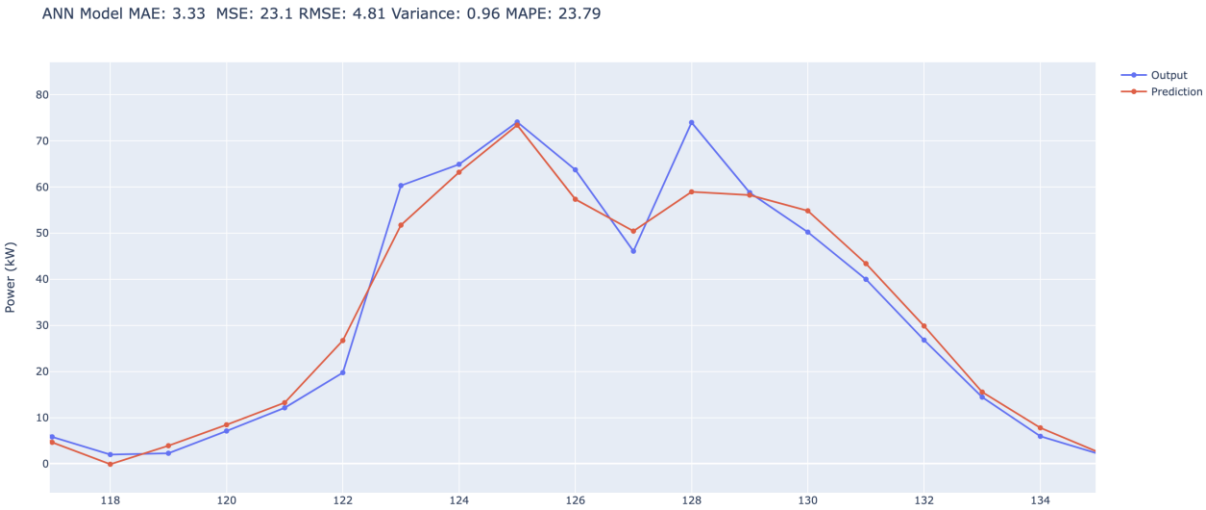


Figure 4.65: ANN results with wind speed included.

Figure 4.66: ANN results without wind speed parameter.

4.3.5 Predictions with forecasted meteorological data on a clear sky day

ANN and LR models were trained and tested on historical data. Moreover, the accuracy of the models was evaluated. The limit of the models has been assessed and the possible lowest error numbers have given an idea about how models would behave on forecasted metrological data. Even though forecasting was not front-and-centre of this study, it is now possible to test the model on forecasted metrological data. Forecasted meteorological data was accessed through THREDDS Data Server (TDS) operated by met.no. Unfortunately, forecasted irradiance values are not available. Instead, clear sky calculated data was used for the prediction day and adjusted to the plane of array irradiance. Besides, PV power data was requested from Lede Energi for the corresponding period. Module temperature was not fed into the model as it is a function of irradiance values. Models were trained by one year before in the same month with 30 days of data and prediction was held on 25th April 2022. Table 4.11 describes the training and test data mean and standard deviation values.

Table 4.11: Training and test data variable statistics for forecasting analysis.

Variables	Mean Values - Training	Std - Training	Mean Values - Test	Std - Test
dew_point_temp	-3.6	3.43	0.36	1.49
air_temp	7.3	4.56	13.2	3.84
relative_humidity	48.7	17.9	43.9	10.1
Clearksy data adjusted to poa_global_east	269.2	189.2	370.1	222.3
wind_speed	2.2	1.16	2.0	0.57
wind_category	2.0	1.03	3.3	0.69
zenith	66.2	12.2	63.6	13.7
elevation	23.8	12.2	26.3	13.7
azimuth	180.6	65.2	182.6	71.2

Power values training set mean value is 30.9 kW and standard deviation is 21.5 kW. Figure 4.67 and Figure 4.68 show ANN and LR results, respectively. The ANN model produced lower error than the LR model and the trend fitted with output values. In the LR model, the model predicted higher power values after midday. One reason is that using clear sky values caused the linear regression model predicts higher values. ANN was able to learn more from historical values.

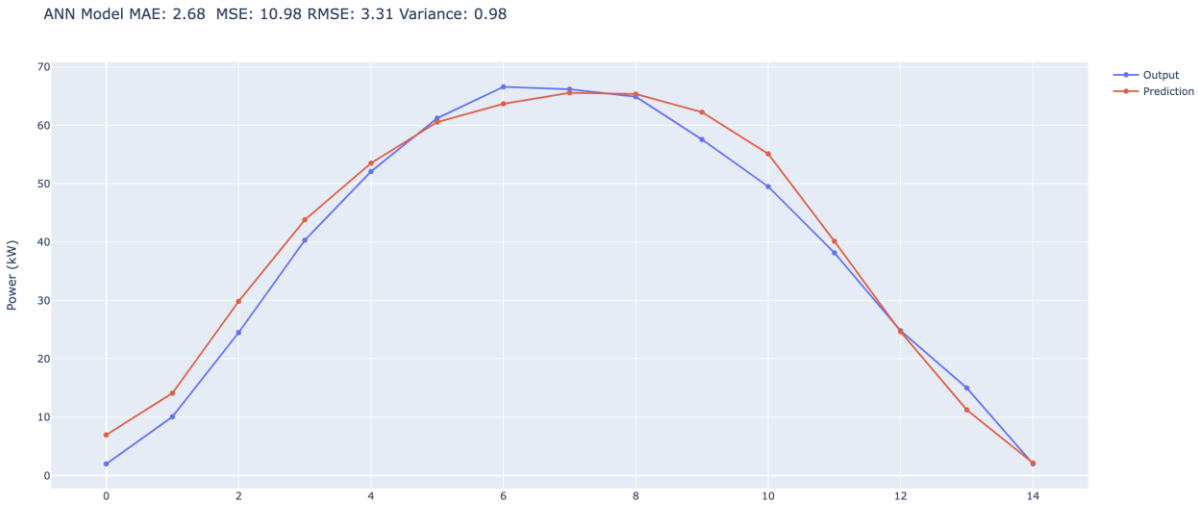


Figure 4.67: ANN prediction results for forecasting analysis on a clear sky day.

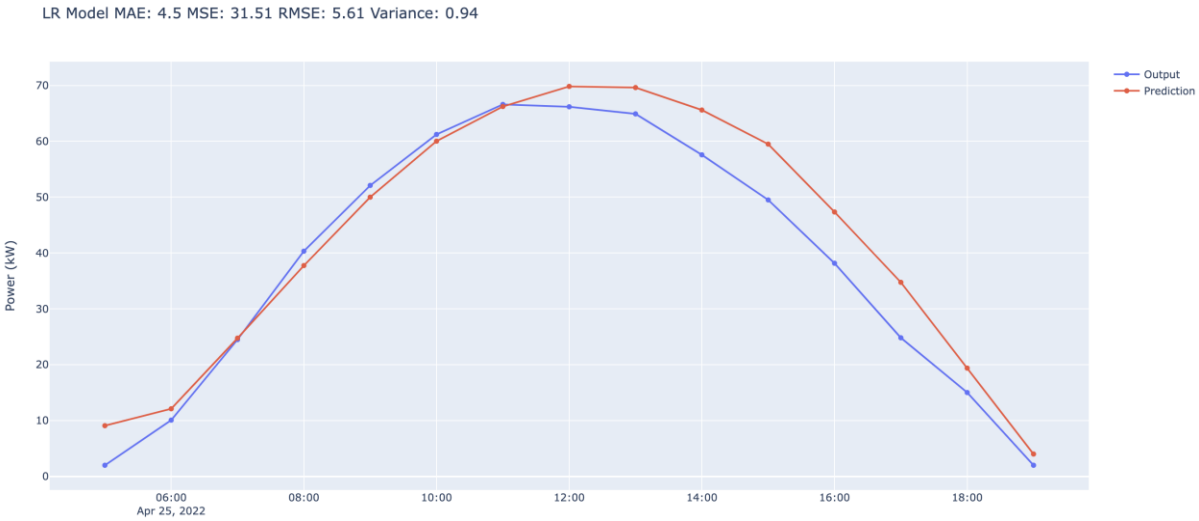


Figure 4.68: LR prediction results for forecasting analysis on a clear sky day.

Prediction on a forecasted cloudy was also planned. Alternatively, cloudiness percent and clear sky values would be used to predict the possible amount of irradiance that modules can absorb. However, it was noticed that the forecasted cloudiness percent on the TDS did not match with the real scenario. That is why the cloudy day forecasting was not presented here.

5 Discussion

Since this study consists of three main parts as data dealing part, clear sky studies, and modelling, the discussion section is also divided into three parts in line with structure.

5.1 Meteorological and PV power datasets

Data dealing comes first for setting up an accurate model. That is why the common data processing techniques were applied to the datasets. Pre-processing of data can result in a huge amount of data loss due to excessive filtering. In this study, while the most of data were tried to be kept, the data was cleared out from outliers, fault status, and unrealistic values. Specifically for PV data, zero values and night periods were removed before modelling. It is important to keep in mind that noise in data still can hold valuable information, and disregarding noise in data might reduce the model performance. At the end of data processing for 4 years of meteorological data, available irradiance values dropped 8.5% which is acceptable. Differently, PV power data was received mainly filtered based on inverter's operation period from the company's database. However, this data was also filtered in terms of night periods when the sun is out. As a result, 12.5% of PV power data filtered out from the database. Considering night periods in PV dataset and out of order period in the MET station, it is fair to say that the data count after processing is sufficient. Moreover, not all PV data does consist of decent values. Irradiance values and PV power output usually go hand in hand. As is seen from Figure 4.25, that is not the case every time. Some power values do not proportionally rise with respect to irradiance. As it was explained in detail in the methodology section, after a rough performance index proposal, power values are becoming degenerate below 9º elevation. If a performance index filtering is applied to data, only decent power values are obtained. However, feeding filtering power values into models increases model erroneous. That is why PV data filtering was halted after removing night periods. It is aimed to predict power values for the complete cycle in a day. Removing some PV power data disrupts models to predict low PV power values.

PCA analysis provided a great deal amount of information about datasets. Thus, it became possible to observe the data distribution and outliers in data. In addition, PCA loadings are in compliance with the correlation matrix. One of the aims of performing PCA analysis was dimension reduction. However, considering three and four components PCA only explains 92% and 85% of variations in datasets. To achieve a higher explanation more components will be required and the dimension reduction goal becomes unreasonable. It is already known that irradiance and power values have a correlation above 95%. This study is actually after the other 5%. As a result, PC values were not fed into the model.

The correlation matrix proves that irradiance and power values have the highest correlation. Sun elevation and air temperature are also important inputs for the model. Wind speed has a higher correlation than wind category. There is no correlation between dew point temperature and power values. In relation to this information, dew point temperature effect could not detect in the variable case model evaluation. Correlation matrixes among other selected inverters slightly differ. One reason is that each inverter was not in operation for different periods. That

is why each inverter's value was correlated with meteorological variables at different times. It is concluded that the difference is insignificant and the effect on the model was negligible.

Detailed meteorological variable investigation gives some insights into seasonal and yearly changes. It is possible to conclude that training models with only one year of meteorological data can still produce similar accuracy based on close correlation coefficients. However, correlations for each different year in the same season may differ.

5.2 Clear sky studies

The model's performance on clear sky days is expected to be at its best value. To prove this idea, a clear sky days study was performed. However, 2 years of data which consists of more than 7000 rows make it difficult to capture clear sky days easily. On the way to clear sky days analysis, it was noticed that measured irradiance values exceed calculated clear sky days and those values are not measurement errors since they were checked with PV power output at the corresponding period. Atmospheric conditions have an impact on irradiance values that reach the earth's surface. If wind speed is low, and relative humidity is high, there is a possibility that measured global horizontal irradiance values will exceed calculated clear sky values.

Different clear sky calculation methods are accessible in the literature and some of them are available on pvlib. To choose the appropriate clear sky model, correlation analysis between measured global horizontal irradiance values and different clear sky model values was performed. The highest correlation is obtained by Perez-Ineichen method among the other two models. As is seen from Figure 3.13, while other methods predict higher values for irradiance values, Perez-Ineichen is the closest model to measured irradiance values. Nevertheless, calculated values frequently exceed the measured values. This problem can be eliminated by parameter change in the model. It is possible to feed different linke turbidity or air mass values but the model run by default values. These values may differ geographically and unfortunately, air mass and linke turbidity values do not exist for Gjerpen station. As a result, clear sky values can be adjusted better to be in line with measured irradiance values.

Pvlib detect_clearsky() function works only with 1min time resolution. Measured irradiance values had to be transformed from 1h to 1min resolution. By using interpolation, values were produced but this rough estimation for 1min measured irradiance values, the algorithm from time to time fails. Some adjustments within the algorithm have been made such as resolution adjustment and different interpolation methods and the best possible output was obtained. Hence, the effort to find clear sky days within the data has become lower. Clear sky values are important as these values are used to produce the plane of irradiance values. That is why an extensive investigation was performed on clear sky irradiance values.

5.3 Prediction and model evaluations

5.3.1 Time Resolution Problem

Three different datasets were used from three different sources. As given information in Table 3.7, PV values have 10 min resolution. 10 mins PV values fluctuate a lot especially on cloudy

days as it is seen in Appendix G. This data averaged to on an hourly basis. In general, averaging is one way of dealing outliers in datasets in data processing steps. One advantage of averaging PV data to 1 hour is having much more smooth data. However, when a new dataset from another source is combined with hourly PV power data, in this case irradiance, the time scale problem arises. For example, in the event of irradiance value drop for a specific period, no changes were captured in PV power data. In the second figure in Appendix G, one example of this situation was illustrated. There might be two reasons that cause the problem, station location might be cloudy and plant location was cloudiness or irradiance measurement disrupted for couple of minutes for some reasons. Whatever the reason, this situation frequently occurred and it is quite difficult to filtering out such periods from the data. One biggest disadvantage keeping such situations in the dataset, they cause prediction fault and increase the model error. One report presented form Sandia National Laboratories proves that the higher time resolution means higher prediction erroneous [43]. In the same report, it is also concluded that reducing the weather parameters interval from one hour to 15 minutes generally results in an error drop in energy by a factor of 10.

In the literature, it is possible to find articles that obtain lower prediction errors for historical data analysis and forecasting. For some cases MAPE error numbers were produced exclusively to make comparisons with specific articles. One paper conducted in Cyprus, found 4.7% mean absolute percentage error on historical dataset for a period of 170 days compared to 25% MAPE for a 90 days period in our study [44]. For consecutive clear sky days study in chapter 4.3.4, MAPE was 14%. Another study for day ahead forecasting found 10.06% MAPE on a clear sky day [45]. A different study conducted research for one year period with 5 min resolution and found R^2 92.2 [46]. Even though this study's outputs for clear sky days evaluation is close to the literature, time resolution problem leads a great amount of erroneous on the output. It is important to check the article's evaluation method. The reason is some evaluations were made based on scaled inputs and outputs. This study uses unscaled values in other words, actual power values for error calculations.

One might argue that instead of averaging PV values to an hour basis, meteorological variables could have converted to 10 mins resolution by interpolation. This idea would not solve the problem for overlapping. An example is shown in the second page of Appendix G. It is notable that observation values do not overlap between irradiance and PV power output. What it means that values do not reflect the same time interval. That is why interpolation of meteorological variables would lead the same issue, again.

5.3.2 Model input selection

Apart from PV power values, it was observed that 11 different inputs have different impacts on power values. In addition, the selection of training period dramatically changes model accuracy. That is why models were evaluated for different periods with different training periods such as training for a short period, or one whole year of training. It was also benefited from K-fold validation. When it comes to predicting clear sky days, the training period is not as important as predicting cloudy days. The best model accuracy values always were obtained with clear sky days predictions. Time resolution issue leads to high cloudy days prediction errors.

Sun parameters such as azimuth, and elevation cause a rise in the model accuracy. These parameters follow a pattern on a daily basis. Thus, predictions catch the power values pattern easily. Air temperature and the plane of array irradiance values are vital for the model. It was possible to obtain low MAE values. In contrast, relative humidity which has the highest negative correlation with power values has almost no impact on the model where the plane of array, air temperature, and relative humidity are only inputs. However, with all variables included and only relative humidity excluded, ANN prediction accuracy increases. LR prediction accuracy does not change. Thus, ANN was able to capture non-linearities in the relative humidity-power values relationship.

It was observed during the simulations that the longer training period is chosen, the more models become dependent on irradiance values. Impact of other weather variables decreases.

5.3.3 Model Evaluations

Learning curves was used in addition to statistical analysing tools to evaluate model performance. It is important to mention that ANN models were run by trial-and-error approach. Model trials were suspended where the best ANN outputs were obtained.

Training for long periods and predicting short periods require simple model parameters. For example, 50 epochs were sufficient for clear sky day power value predictions training with 2020/2021 data. Whereas more epochs were required in case of using a smaller training dataset.

In general, the training and test set ratio basically manipulates data complexity. The reason is that learning capabilities from the data are limited where a lower train/test ratio is applied. Higher node numbers were used in the ANN model and the LR model still needed more data to learn.

Even though squared error methods are widely used as a comparison method in the literature, mean absolute errors were the main comparison criteria between LR and ANN as having relatively high error numbers due to time resolution issues in this study. By definition, squared error methods produce higher numbers and the sensitivity dramatically drops for this dataset. MAPE is, however, used to make this study comparable with some articles in the literature.

5.4 Future Work Discussions

This comprehensive study still has a huge potential to make accurate predictions. Averaging power is overestimated and leads to higher errors in predictions. Once inevitable issues due to time resolution eliminated, models can produce more sensitive results in terms of capturing meteorological variable effects. Since power values are recorded at 10min resolution, frequent sampling rate recording values are required for irradiance, air temperature, and relative humidity including wind direction data. Thus, it will be possible to capture irradiance and power output fluctuations at the same time for better training. Furthermore, it was expected to observe wind direction effect for each layout, separately. By achieving higher accuracy, this effect may be explored deeper for IV5 and IV7.

An optimisation study for ANN hyperparameters is required to make model reliable throughout the year. Thus, a real-time PV power forecasting system can be built and work in harmony for a grid optimisation.

Depending on forecasting horizons such as short term (less than 1 day) or medium term (1-3 days ahead), there is a trade-off between selling electricity to the market and storage of batteries. In general market electricity prices are at the lowest from 11:00 to 16:00 depending on the season, however, PV power output is at its highest. This relation is known as the duck curve. Without a doubt, battery systems are one solution to sell PV power to the market when prices are high such as in early mornings or in the evenings. Nevertheless, building and operating feasible systems are challenging. On the one hand, accurate short term forecasting outputs may suggest a period when expected PV power is high out of the lowest electricity prices period. Thus, the grid connected period can be planned. In addition, the observed differences in PV power output for the same weather conditions hide unique potential to detect problems and keep the system in operation at the highest efficiency all the time. On the other hand, medium term forecast contributes to power system management in addition to planning maintenance activities from a broader perspective. In summary, for future work, building a model and training with a shorter time resolution such as 10min to increase the model accuracy, and forecasting with a 1hour time resolution for grid planning scenario should be investigated.

One might demand using trained models in this study for a different PV module, layout, or plant to predict PV power output. However, PV power output is not only the function of weather parameters. Firstly, inverter type and its efficiency are one parameter that has not been discussed in this study. Typically, for a given DC input, the inverter converts to AC power for only a certain amount. The efficiency of inverters not only changes during the day but also takes different values for each inverter type. Since this study uses the output of inverters which is AC, the results might comparable for only the same type of inverter. Secondly, the module type is another parameter that affects PV power output. In the PV module specification sheet, the temperature coefficients of PV modules are stated as a linear equation. For example, the module types that are subject to this analysis have a -0.36%/ºC for P_{MPP}. It describes that every 1 ºC rise in module temperature, results in a 0.36% drop in power. The nominal module operating temperature has given as 44.6 ± 2 ºC. These values vary for different manufacturers. That is why the models are fit for the same module types. In Skagerak Arena, there are different types of modules with 300 Wp in the south direction. However, since models are trained by each inverter's historical data separately, PV power output values are already influenced by module specifications. Alternatively, an efficiency factor can be calculated by taking into account measured module temperature and power loss/rise percentage. Thus, PV power values would be normalised and models can be used for different types of modules. Thirdly, tilt angle and direction (azimuth) are other parameters. Measured global horizontal values have been converted to the plane of array irradiance values by taking into account tilt angle and azimuth values. Therefore, it is possible to use these models feeding with adjusted irradiance values in the process of predictions. Moreover, cabling systems also affect power output. Considering various parameters that effects using the model in other PV systems, further adjustments have to be made. In particular, building models by training with DC instead of AC power can eliminate the effect of inverters and most some cabling differences. Thus, the newly established model including the temperature coefficient factor would serve different PV systems.

6 Conclusion

This study provides thorough research in the PV power prediction domain focusing on time series prediction using LR and ANN models. The results show that ANN models are relatively better in the prediction of PV power values than LR models.

It is concluded from the meteorological variable study that yearly changes within meteorological variables will not have a big impact on PV power prediction estimations according to correlation analysis. Even though PCA analysis provides some information on data and relationships, using PCA components does not produce improved results for predictions due to low explanation. PV power data filtering based on performance resulted in an accuracy drop due to data loss for low elevation periods. There is a slight difference between the correlations and power values with respect to different layouts and other meteorological variables. One of the reasons that cause this difference is that each inverter has a different operation time.

For the specified location, Ineichen-Perez clear sky method delivered better results for correlation with measured irradiance values with a 0.9873 correlation coefficient. It is possible to produce more correlated results by taking into account observed air mass or linke turbidity variables.

It is concluded that air temperature and the plane of array irradiance parameters are vital to predicting PV power values. Relative humidity is another important parameter to reach better accuracy, especially for ANN models. However, clear relation could not be detected with the LR model for relative humidity variable. The best performance was obtained by adding relative humidity and other sun parameters to air temperature and the plane of array irradiance parameters. In the detailed model parameters study, wind direction is another parameter to make predictions closer to real outputs. Furthermore, it is found out that all variables that were investigated have an effect on power value predictions during relevant variable fluctuation periods. The time resolution issue which causes swings in power values and weather parameters leads to an accuracy drop for both models.

The study was mainly conducted on IV2 values. Correlations with other weather parameters are quite similar for other selected inverters, IV5 and IV7. Using trained models produce more accurate predictions for similar inverter-module types.

Clear sky prediction accuracies are comparable with literature and it is practical to use clear sky irradiance values instead of measured irradiance values for the prediction of clear sky days including other meteorological variables. Moreover, by using forecasted meteorological data for a day ahead forecasting on a clear sky day produced comparable results to historical data predictions.

References

[1] D. Dixon. "Most of 2022's solar PV projects risk delay or cancelation due to soaring material and shipping costs." Rystad Energy, 2022. https://www.rystadenergy.com/newsevents/news/press-releases/most-of-2022s-solar-PV-projects-risk-delay-or-cancelation-due-to-soaring-material-and-shipping-costs/#:~:text=Driven%20by%20core%20component%20price,50%25%20increase%20in%20a%20year. (accessed 5 May, 2022).

[2] C. Philibert, "Solar Energy Perspectives," International Energy Agency, France, 2021. [Online]. Available: https://iea.blob.core.windows.net/assets/2b3c53f4-1c8f-478c-a4fa-a98597cde27b/SolarEnergyPerspectives.pdf

[3] S. Alvik, "Energy Transition Norway 2020," DNV GL AS, 2020. [Online]. Available: https://www.norskindustri.no/siteassets/dokumenter/rapporter-og-brosjyrer/energy-transiton-norway-2020.pdf

[4] Y. Xue, C. M. Lindkvist, and A. Temeljotov-Salaj, "Barriers and potential solutions to the diffusion of solar photovoltaics from the public-private-people partnership perspective – Case study of Norway," *Renewable and Sustainable Energy Reviews,* vol. 137, p. 110636, 2021/03/01/ 2021, doi: doi.org/10.1016/j.rser.2020.110636.

[5] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting,* vol. 14, no. 1, pp. 35-62, 1998, doi: doi.org/10.1016/S0169-2070(97)00044-7.

[6] P. Gupta and R. Singh, "PV power forecasting based on data-driven models: a review," *International Journal of Sustainable Engineering,* vol. 14, no. 6, pp. 1733-1755, 2021, doi: doi.org/10.1080/19397038.2021.1986590.

[7] H. F. Hamann, "A Multi-scale, Multi-Model, Machine-Learning Solar Forecasting Technology," IBM, Yorktown Heights, NY (United States). Thomas J. Watson Research Center, United States, 2017. [Online]. Available: https://www.osti.gov/biblio/1395344

[8] S. Shields. "Cells, Modules & Arrays." The Florida Solar Energy Center, 2014. http://www.fsec.ucf.edu/en/consumer/solar_Electricity/basics/cells_modules_arrays.htm (accessed February 24th, 2022).

[9] REC Solar Pte. Ltd. "Maximizing cell performance How REC's use of Passivated Emitter Rear Cell technology improves the capture of light and optimizes cell performance." REC, 2019. https://www.recgroup.com/sites/default/files/documents/whitepaper_perc.pdf?t=1643363188 (accessed January, 28, 2022).

[10] M. A. Green, "Chapter I-2-B - High-Efficiency Silicon Solar Cell Concepts," in *McEvoy's Handbook of Photovoltaics (Third Edition),* S. A. Kalogirou Ed.: Academic Press, 2018, pp. 95-128.

[11] M. A. Green, "The Passivated Emitter and Rear Cell (PERC): From conception to mass production," *Solar Energy Materials and Solar Cells,* vol. 143, pp. 190-197, 2015, doi: doi.org/10.1016/j.solmat.2015.06.055.

[12] REC Solar Pte. Ltd. "REC Twin Design: Innovative design to raise the power of solar panels even higher!" REC. https://www.recgroup.com/sites/default/files/documents/whitepaper_twinpeak_technology.pdf?t=1643363188 (accessed January, 28, 2022).

[13] S. Dubey, G. N. Tiwari, *Fundamentals of Photovoltaic Modules and Their Applications*. Cambridge, UK: The Royal Society of Chemistry, , 2010, p. 422.

[14] M. Seapan, Y. Hishikawa, M. Yoshita, and K. Okajima, "Detection of shading effect by using the current and voltage at maximum power point of crystalline silicon PV modules," *Solar Energy*, vol. 211, pp. 1365-1372, 2020, doi: doi.org/10.1016/j.solener.2020.10.078.

[15] D. Sonnenenergie, *Planning and Installing Photovoltaic Systems: A Guide for Installers, Architects and Engineers*. Earthscan, 2008.

[16] A. Luque and S. Hegedus, *Handbook of Photovoltaic Science and Engineering*. Wiley, 2011.

[17] E. F. Bradley, "Bounday layer (athmospheric) and air pollution | Observational Techniques In Situ," in *Encyclopedia of Atmospheric Sciences (Second Edition)*, G. R. North, J. Pyle, and F. Zhang Eds. Oxford: Academic Press, 2015, pp. 274-283.

[18] NIBIO. "Instrumenter og målinger." NIBIO. https://lmt.nibio.no/information/5/ (accessed May 4th, 2022).

[19] H. Y. Cheng, C. C. Yu, K. C. Hsu, C. C. Chan, M. H. Tseng, and C. L. Lin, "Estimating Solar Irradiance on Tilted Surface with Arbitrary Orientations and Tilt Angles," *Energies*, vol. 12, no. 8, p. 1427, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/8/1427.

[20] R. Kent, "Chapter 4 - Services," in *Energy Management in Plastics Processing (Third Edition)*, R. Kent Ed.: Elsevier, 2018, pp. 105-210.

[21] U. K. Das *et al.*, "Forecasting of photovoltaic power generation and model optimization: A review," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 912-928, 2018, doi: doi.org/10.1016/j.rser.2017.08.017.

[22] N. Maitanova *et al.*, "A Machine Learning Approach to Low-Cost Photovoltaic Power Prediction Based on Publicly Available Weather Reports," *Energies*, vol. 13, no. 3, p. 735, 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/3/735.

[23] L. Hernández *et al.*, "A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework," (in eng), *Sensors (Basel, Switzerland)*, vol. 12, no. 9, pp. 11571-11591, 2012, doi: doi.org/10.3390/s120911571.

[24] T. AlSkaif, S. Dev, L. Visser, M. Hossari, and W. van Sark, "A systematic analysis of meteorological variables for PV output power estimation," *Renewable Energy*, vol. 153, pp. 12-22, 2020, doi: /doi.org/10.1016/j.renene.2020.01.150.

[25] M.B. Øgaard, A. Skomendal, J.H. Selj, "Performance Evaluation of Monitoring Algorithms for Photovoltaic Systems," presented at the 36th European Photovoltaic Solar Energy Conference and Exhibition, 2019.

[26] M. B. Øgaard, H. N. Riise, H. Haug, S. Sartori, and J. H. Selj, "Photovoltaic system monitoring for high latitude locations," *Solar Energy*, vol. 207, pp. 1045-1054, 2020, doi: doi.org/10.1016/j.solener.2020.07.043.

[27] G. G. Kim *et al.*, "Prediction Model for PV Performance With Correlation Analysis of Environmental Variables," *IEEE Journal of Photovoltaics*, vol. 9, no. 3, pp. 832-841, 2019, doi: doi.org/10.1109/JPHOTOV.2019.2898521.

[28] S. Leva, A. Dolara, F. Grimaccia, M. Mussetta, and E. Ogliari, "Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power," *Mathematics and Computers in Simulation*, vol. 131, pp. 88-100, 2017, doi: doi.org/10.1016/j.matcom.2015.05.010.

[29] I. Jebli, F. Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," *Energy*, vol. 224, p. 120109, 2021, doi: doi.org/10.1016/j.energy.2021.120109.

[30] X. Wang, Y. Sun, D. Luo, and J. Peng, "Comparative study of machine learning approaches for predicting short-term photovoltaic power output based on weather type classification," *Energy*, vol. 240, p. 122733, 2022, doi: doi.org/10.1016/j.energy.2021.122733.

[31] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Forecasting of PV Power Generation using weather input data-preprocessing techniques," *Energy Procedia*, vol. 126, pp. 651-658, 2017, doi: doi.org/10.1016/j.egypro.2017.08.293.

[32] M. P. Almeida, O. Perpiñán, and L. Narvarte, "PV power forecast using a nonparametric PV model," *Solar Energy*, vol. 115, pp. 354-368, 2015, doi: doi.org/10.1016/j.solener.2015.03.006.

[33] L. L. Li, P. Cheng, H. C. Lin, and H. Dong, "Short-term output power forecasting of photovoltaic systems based on the deep belief net," *Advances in Mechanical Engineering*, vol. 9, no. 9, p. 1687814017715983, 2017, doi: doi.org/10.1177/1687814017715983.

[34] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc, 2017.

[35] A. C. Muller, and S. Guido, "Introduction to machine learning with Python : a guide for data scientists,", O'Reilly Media, Inc. 2017.

[36] F. Chollet, "Keras", https://github.com/fchollet/keras, 2015.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[38] J. Angles, L. Ménard, O. Bauer, C. Rigollier, and L. Wald, "A climatological database of the Linke turbidity factor," in *ISES Solar World Congress 1999*, Jerusalem, Israel, vol. 1, pp. 432-434, 1999.

[39] W. Holmgren, C. Hansen, and M. Mikofski, "pvlib python: a python package for modeling solar energy systems," *Journal of Open Source Software*, vol. 3, p. 884, 2018, doi: doi.org/10.21105/joss.00884.

[40] D. Nettleton, "Chapter 6 - Selection of Variables and Factor Derivation," in *Commercial Data Mining*, D. Nettleton Ed. Boston: Morgan Kaufmann, 2014, pp. 79-104.

[41] S. Boslaugh, "Statistics in a nutshell", O'Reilly Media, Inc., 2012.

[42] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.

[43] S. N. Laboratories, "Effect of Time Scale on Analysis of PV System Performance," Energy Systems Integration Group, USA, 2012. Accessed: 7th May. [Online]. Available: https://www.esig.energy/wiki-main-page/effect-of-time-scale-on-analysis-of-pv-system-performance/#cite_note-1

[44] S. Theocharides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, and G. E. Georghiou, "Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing," *Applied Energy*, vol. 268, p. 115023, 2020, doi: doi.org/10.1016/j.apenergy.2020.115023.

[45] M. Ding, L. Wang, and R. Bi, "An ANN-based Approach for Forecasting the Power Output of Photovoltaic System," *Procedia Environmental Sciences*, vol. 11, pp. 1308-1315, 2011, doi: doi.org/10.1016/j.proenv.2011.12.196.

[46] D. O'Leary and J. Kubby, "Feature Selection and ANN Solar Power Prediction," *Journal of Renewable Energy*, vol. 2017, p. 2437387, 2017, doi: doi.org/10.1155/2017/2437387.

Appendices

Appendix A: Task Description

FMH606 Master's Thesis

Title: Solar power electricity production correlated to meteorological data

USN supervisor: Kjell-Arne Solli

External partner: The Norwegian Meteorological Institute (Erik Berge, Associate Professor - Meteorology og Oceanography at UiO). Support from IFE (NFR-project Sunpoint) and Skagerak Energi.

Task Background: Renewable energy sources will gain importance as profitability and acceptance will become in favour of existing energy sources based on thermal power plants ('fossil fuel') and limited hydropower potential. Wind and solar energy sources are inherently unstable and weather dependent with respect to predictable production. It is of interest to gain improved knowledge on solar power potential for Norway, as is the aim of the NFR-project Sunpoint.

Task description:
Historical data on weather observations (The Norwegian Meteorological Institute), and forecasts from Meteorologist Weather Processor (MWP) on an hourly basis, shall be correlated to electricity production from solar energy. Correlations will be valuable input for the development of solar electricity production models. Main data sources for this task are observations from Gjerpen weather station (Skien) and Skagerak Energilab (Skagerak Arena, Skien). If time permits, supplemental data from campus Porsgrunn is planned.
Observed variations in electricity production shall be discussed related to the proper operation of the electrical network and the grid as well as benefit from weather observations and forecasts.
The study should satisfy the following steps: A literature review of relevant methods and results, effective way of data handling and analysis including pre-processing methods, correlation analysis, developing a model, and evaluation of the model on historical and forecast data.
The task calls for skills in mathematical correlation and modelling of large database data sets.

Student category: EET

Is the task suitable for online students (not present at the campus)? Yes

Supervision:
As a general rule, the student is entitled to 15-20 hours of supervision. This includes necessary time for the supervisor to prepare for supervision meetings (reading material to be discussed, etc).

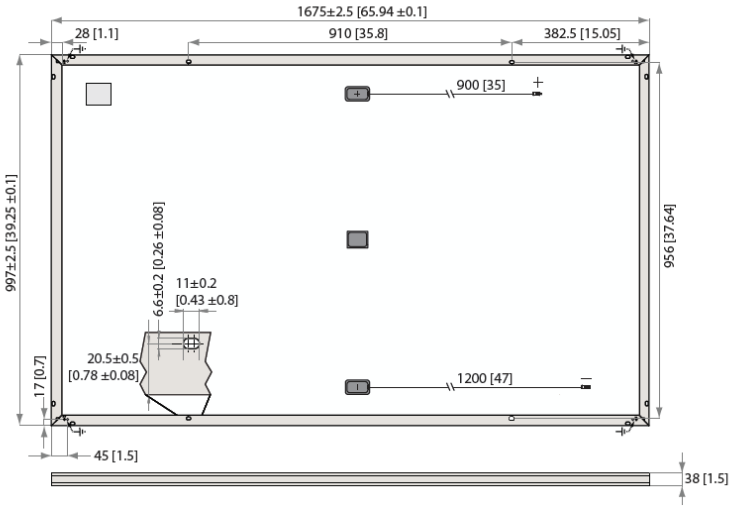
Signatures:
Supervisor (date and signature):

Student (write clearly in all capitalized letters): OZGUR YALCIN
Student (date and signature): 26/01/2022

Appendix B: PV Plant Information

1. Solar Module Specification Sheet

REC TWINPEAK 2 SERIES



Measurements in mm [in]

18.0% EFFICIENCY

10 YEAR PRODUCT WARRANTY

25 YEAR LINEAR POWER OUTPUT WARRANTY

GENERAL DATA

Cell type:	120 half-cut multicrystalline PERC cells 6 strings of 20 cells in series
Glass:	3.2 mm solar glass with anti-reflection surface treatment
Backsheet:	Highly resistant polyester polyolefin construction
Frame:	Anodized aluminum (silver / black)
Junction box:	3-part, 3 bypass diodes, IP67 rated in accordance with IEC 62790
Cable:	4 mm² solar cable, 0.9 m + 1.2 m in accordance with EN 50618
Connectors:	Stäubli MC4 PV-KBT4/PV-KST4 (4 mm²) Tonglin TL-Cable01S-FR (4 mm²) in accordance with IEC 62852, IP68 only when connected
Origin:	Made in Singapore

ELECTRICAL DATA @ STC

Product code*: RECxxxTP2						
Nominal Power - P_{MPP} (Wp)	275	280	285	290	295	300
Watt Class Sorting - (W)	-0/+5	-0/+5	-0/+5	-0/+5	-0/+5	-0/+5
Nominal Power Voltage - V_{MPP} (V)	31.5	31.7	31.9	32.1	32.3	32.5
Nominal Power Current - I_{MPP} (A)	8.74	8.84	8.95	9.05	9.14	9.24
Open Circuit Voltage - V_{OC} (V)	38.2	38.4	38.6	38.8	39.0	39.2
Short Circuit Current - I_{SC} (A)	9.52	9.61	9.66	9.71	9.76	9.82
Panel Efficiency (%)	16.5	16.8	17.1	17.4	17.7	18.0

Values at standard test conditions (STC: air mass AM 1.5, irradiance 1000 W/m², temperature 25°C), based on a production spread with a tolerance of V_{OC} & I_{SC} ±3% within one watt class. At a low irradiance of 200 W/m² at least 95% of the STC module efficiency will be achieved.
*Where xxx indicates the nominal power class (P_{MPP}) at STC indicated above, and can be followed by the suffix BLK for black framed modules.

MAXIMUM RATINGS

Operational temperature:	-40 ... +85°C
Maximum system voltage:	1000 V
Design load (+): snow Maximum test load (+):	367 kg/m² (3600 Pa)* 550 kg/m² (5400 Pa)
Design load (-): wind Maximum test load (-):	163 kg/m² (1600 Pa)* 244 kg/m² (2400 Pa)
Max series fuse rating:	25 A
Max reverse current:	25 A

*Safety factor 1.5

ELECTRICAL DATA @ NMOT

Product code*: RECxxxTP2						
Nominal Power - P_{MPP} (Wp)	206	210	214	218	223	226
Nominal Power Voltage - V_{MPP} (V)	29.2	29.4	29.6	29.8	30.0	30.1
Nominal Power Current - I_{MPP} (A)	7.07	7.15	7.24	7.32	7.43	7.51
Open Circuit Voltage - V_{OC} (V)	35.4	35.6	35.8	36.0	36.2	36.3
Short Circuit Current - I_{SC} (A)	7.52	7.59	7.68	7.75	7.85	7.91

Nominal module operating temperature (NMOT: air mass AM 1.5, irradiance 800 W/m², temperature 20°C, windspeed 1 m/s).
*Where xxx indicates the nominal power class (P_{MPP}) at STC indicated above, and can be followed by the suffix BLK for black framed modules.

TEMPERATURE RATINGS*

Nominal Module Operating Temperature:	44.6°C (±2°C)
Temperature coefficient of P_{MPP}:	-0.36 %/°C
Temperature coefficient of V_{OC}:	-0.30 %/°C
Temperature coefficient of I_{SC}:	0.066 %/°C

*The temperature coefficients stated are linear values

CERTIFICATIONS

IEC 61215, IEC 61730 & UL 1703; MCS 005, IEC 62804 (PID)
IEC 62716 (Ammonia Resistance), IEC 60068-2-68 (Blowing Sand)
IEC 61701 (Salt Mist level 6), UNI 8457/9174 (Class A), ISO 11925-2 (Class E)
ISO 9001: 2015, ISO 14001: 2004, OHSAS 18001: 2007

WARRANTY

10 year product warranty
25 year linear power output warranty
(max. degression in performance of 0.7% p.a.)
See warranty conditions for further details.

MECHANICAL DATA

Dimensions:	1675 x 997 x 38 mm
Area:	1.67 m²
Weight:	18.5 kg

Specifications subject to change without notice.
Ref: NE-05-07-07 Rev-G.2 11.17

2. Inverter Specification Sheet

SOLAR INVERTERS

ABB string inverters
PVS-100/120-TL

01

—
01
PVS-100/120-TL
three-phase outdoor
string inverter

The PVS-100/120-TL is ABB's cloud connected three-phase string solution for cost efficient decentralized photovoltaic systems for both ground mounted and large commercial applications.

This completely new platform, for extreme high power string inverters with power ratings up to 120 kW, maximizes the ROI for decentralized ground mounted and large rooftop applications. With six MPPT energy harvesting is optimized even in shading situations.

Extreme power with high integration level
The extreme high power module up to 120 kW saves installation resources as less units are required. Due to its compact size further savings are generated in logistics and in maintenance. Thanks to the integrated DC/AC disconnection, 24 string connections, fuses and surge protection no additional boxes are required.

Ease of installation
The horizontal and vertical mounting possibility creates flexibility for both ground mounted and rooftop installations. Covers are equipped with hinges and locks that are fast to open and reduce the risk of damaging the chassis and interior components when commissioning and performing maintenance actions.

Standard wireless access from any mobile device makes the configuration of inverter and plant easier and faster. Improved user experience thanks to a build in User Interface (UI) enables access to advanced inverter configuration settings.

The installer mobile APP, available for Android/IOS devices, further simplifies multi-inverter installations.

The design supports both copper and aluminum cabling even up to 185 mm² cross section to minimize the energy losses.

Fast system integration
Industry standard Modbus/SUNSPEC protocol enables fast system integration. Two ethernet ports enable fast and future proof communication for PV plants.

ABB plant portfolio integration
Monitoring your assets is made easy as every inverter is capable to connect to ABB plant portfolio manager to secure your assets and profitability in long term.

Design flexibility and shade tolerance
The double stage conversion topology and six MPPT guarantee maximum flexibility for the system design on rooftops or hilly ground.
With this technological choice energy harvesting is optimized even in shading situations.

Highlights
- 6 independent MPPT
- Transformerless inverter
- 120 kW for 480 Vac and 100 kW for 400 Vac
- Wi-Fi as standard for configuration
- Two ethernet ports for plant level communication
- Large set of specific grid codes available which can be selected directly in the field
- Double stage topology for a wide input range
- Both vertical and horizontal installation
- Separate wiring compartment for fast swap and replacement
- IP66 Environmental protection
- Maximum efficiency up to 98.9%

ABB string inverters

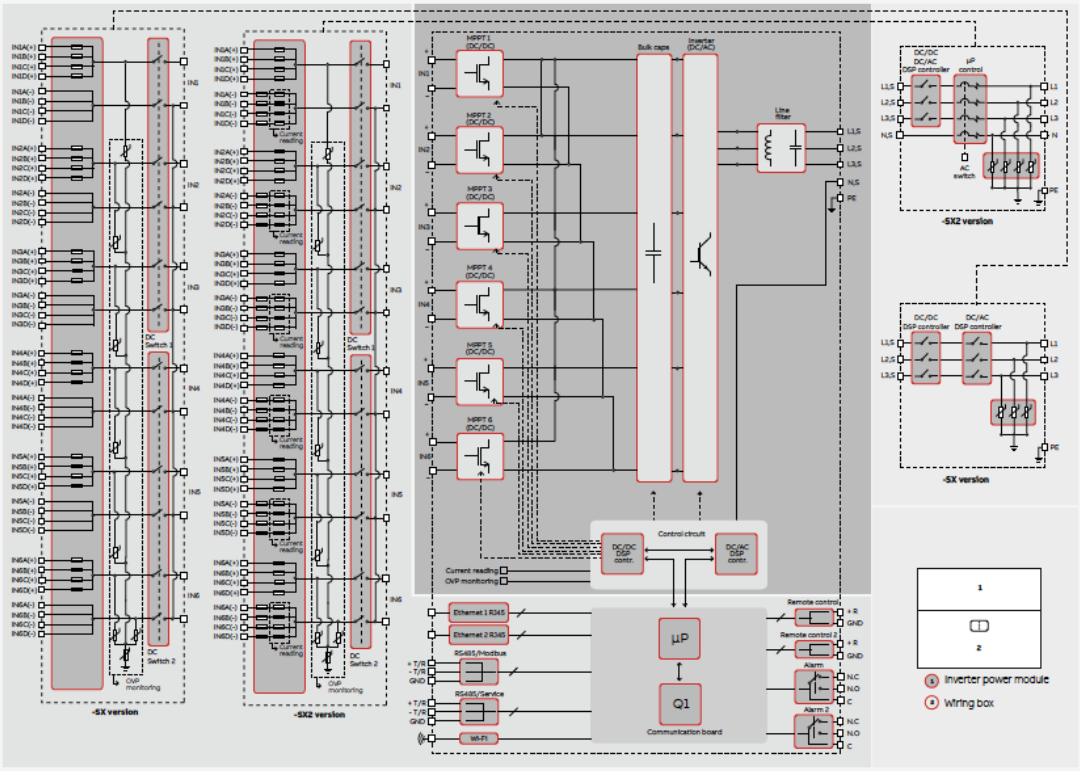
PVS-100/120-TL

100 to 120 kW

Technical data and types

Type code	PVS-100-TL	PVS-120-TL
Input side		
Absolute maximum DC input voltage ($V_{max,abs}$)	1000V	
Start-up DC input voltage (V_{start})	420V (400...500 V)	
Operating DC input voltage range (V_{dcmin}...V_{dcmax})	360...1000 V	
Rated DC input voltage (V_{dcr})	620V	720V
Rated DC input power (P_{dcr})	102 000W	123 000W
Number of independent MPPT	6	
MPPT input DC voltage range at ($V_{MPPTmin}$...$V_{MPPTmax}$) at P_{acr}	480...850V	570...850V
Maximum DC input power for each MPPT ($P_{MPPT,max}$)	17500 W [480V≤V_{MPPT}≤850V	20500 W [570V≤V_{MPPT}≤850V
Maximum DC input current for each MPPT (I_{dcmax})	36 A	
Maximum input short circuit current (I_{scmax}) for each MPPT	50 A [1]	
Number of DC input pairs for each MPPT	4	
DC connection type	PV quick fit connector [2]	
Input protection		
Reverse polarity protection	Yes, from limited current source	
Input over voltage protection for each MPPT - replaceable surge arrester	Type 2 with monitoring	
Photovoltaic array isolation control	as per IEC62109	
DC switch rating for each MPPT	50 A / 1000 V	
Fuse rating (versions with fuses)	15 A / 1000 V [3]	
String current monitoring	SX2: (24ch) Individual string current monitoring; SX: (6ch) Input current monitoring per MPPT	
Output side		
AC Grid connection type	Three phase 3W+PE or 4W+PE	
Rated AC power (P_{acr} @cosΦ=1)	100 000 W	120 000 W
Maximum AC output power (P_{acmax} @cosΦ=1)	100 000 W	120 000 W
Maximum apparent power (S_{max})	100 000 VA	120 000 VA
Rated AC grid voltage ($V_{ac,r}$)	400 V	480 V
AC voltage range	320...480 V [4]	384...576 [3]
Maximum AC output current ($I_{ac,max}$)	145 A	
Rated output frequency (f_r)	50 Hz / 60 Hz	
Output frequency range (f_{min}...f_{max})	45...55 Hz / 55...65 Hz [5]	
Nominal power factor and adjustable range	> 0.995, 0...1 inductive/capacitive with maximum S_{max}	
Total current harmonic distortion	< 3%	
Maximum AC cable	185mm2 Aluminum and copper	
AC connection type	Provided bar for lug connections M10, single core cable glands 4xM40 and M25, multi core cable gland M63 as option	
Output protection		
Anti-islanding protection	According to local standard	
Maximum external AC overcurrent protection	225 A	
Output overvoltage protection - replaceable surge protection device	Type 2 with monitoring	
Operating performance		
Maximum efficiency (η_{max})	98.4%	98.9%
Weighted efficiency (EURO)	98.2%	98.6%
Communication		
Embedded communication interfaces	1x RS485, 2x Ethernet (RJ45), WLAN (IEEE802.11 b/g/n @ 2,4 GHz)	
User interface	4 LEDs, Web User Interface	
Communication protocol	Modbus RTU/TCP (Sunspec compliant)	
Commissioning tool	Web User Interface, Mobile APP/APP for plant level	
Remote monitoring services	Aurora Vision® monitoring portal	
Advanced features	Embedded logging, direct telemetry data transferring to ABB cloud	
Environmental		
Ambient temperature range	-25...+60°C /-13...140°F with derating above 40°C / 104 °F	

ABB PVS-100/120-TL string inverter block diagram



Technical data and types

Type code	PVS-100-TL	PVS-120-TL
Relative humidity	4%...100% condensing	
Sound pressure level, typical	68dB(A) @ 1m	
Maximum operating altitude without derating	2000 m / 6560 ft	
Physical		
Environmental protection rating	IP 66 (IP54 for cooling section)	
Cooling	Forced air	
Dimension (H x W x D)	869x1086x419 mm / 34.2" x 42.8" x 16.5"	
Weight	70kg / 154 lbs for power module ; ~55kg / 121 lbs for wiring box Overall max 125 kg / 276 lbs	
Mounting system	Mounting bracket vertical & horizontal support	
Safety		
Isolation level	Transformerless	
Marking & EMC	CE conformity according to LV and EMC directives	
Safety	IEC/EN 62109-1, IEC/EN 62109-2	
Grid standard (check your sales channel for availability)	CEI 0-16, CEI 0-21, IEC 61727, IEC 62116, IEC 60068, IEC 61683, JORDAN IRR-DCC-MV, AS/NZS4777.2, VDE-AR-N 4105, VDE V 0-126-1-1, VFR 2014, Belg C10-C11, UK59/3, P.O. 12.3, ITC-BT-40, EN50438 Generic +Ireland, CLC-TS 50549-1/2	
Available products variants		
Inverter power module	PVS-100-TL-POWERMODULE-400	PVS-120-TL-POWERMODULE-480
24 ch quick input connections + fuses (both poles) + DC switches + individual string current monitoring (ch 24) + AC breaker + surge arresters Type 2, (DC & AC)	WB-SX2-PVS-100-TL	WB-SX2-PVS-120-TL
24 ch quick input connections + fuses (single pole) + DC switches + input current monitoring per MPPT (ch 6) + surge arresters Type 2 (DC & AC)	WB-SX-PVS-100-TL	WB-SX-PVS-120-TL
Optional available		
Support for multi core AC cable M63 + M25 (PE)	AC output panel M63 for wiring box	

1) Maximum number of opening 5 under overloading
2) Please refer to the document "String inverters – Product manual appendix" available at www.abb.com/solarinverters for information on the quick-fit connector brand and model used in the inverter
3) Maximum fuse size supported 20 A. Additionally one string input per MPPT supports 32 A fuse sizes for connecting two strings per input
4) The AC voltage range may vary depending on specific country grid standard
5) Frequency range may vary depending on specific country grid standard
Remark. Features not specifically listed in the present data sheet are not included in the product

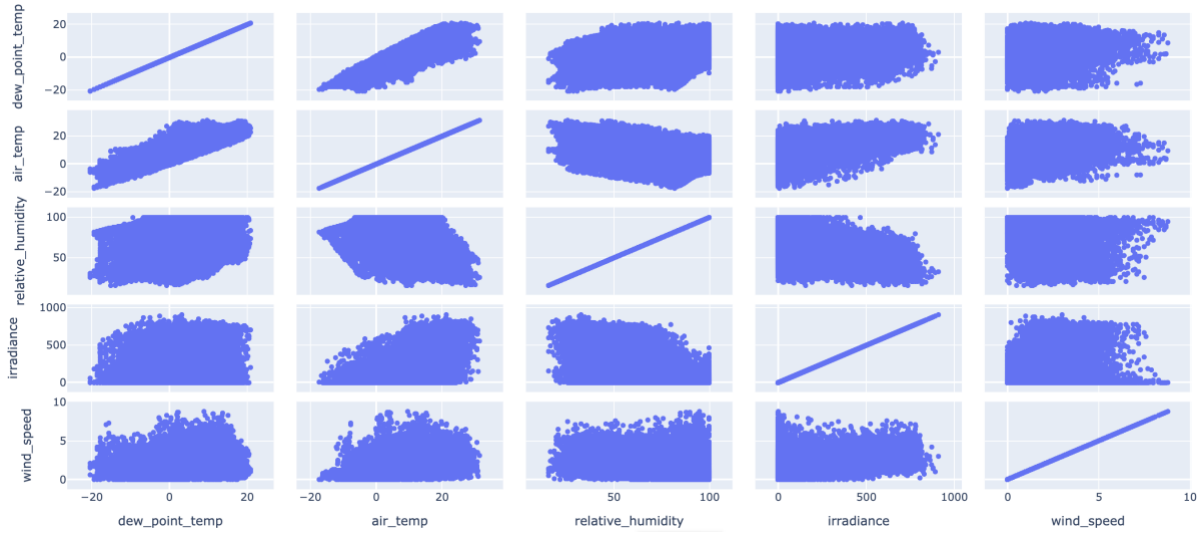
Appendix C: Data

1. Data Type

Data_Type:

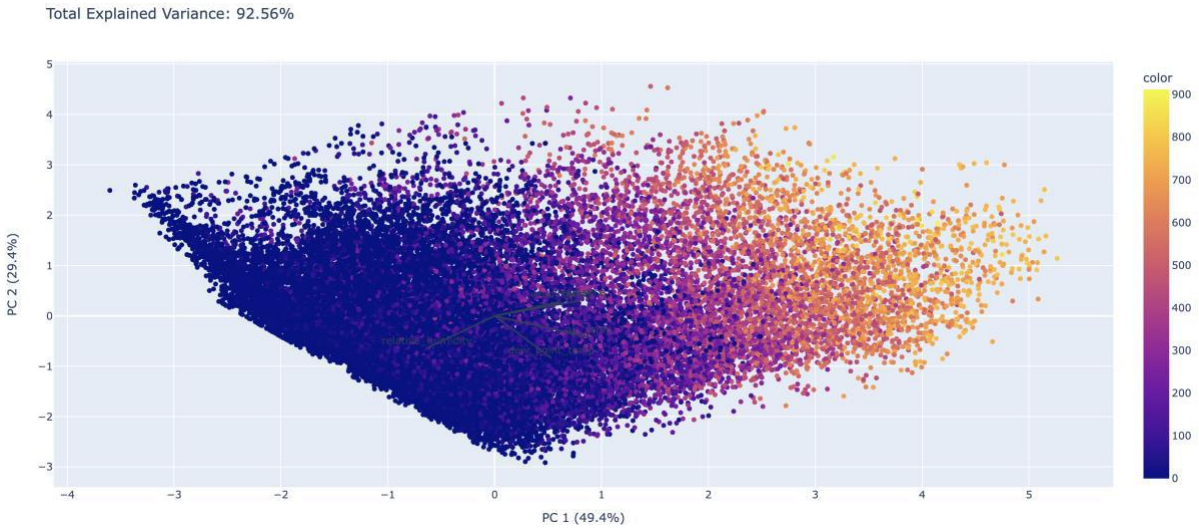
referenceTime	object
dew_point_temp	float64
air_temp	float64
relative_humidity	float64
irradiance	float64
wind_speed	float64

2. Scatter_Matrix_Outliers_Removed

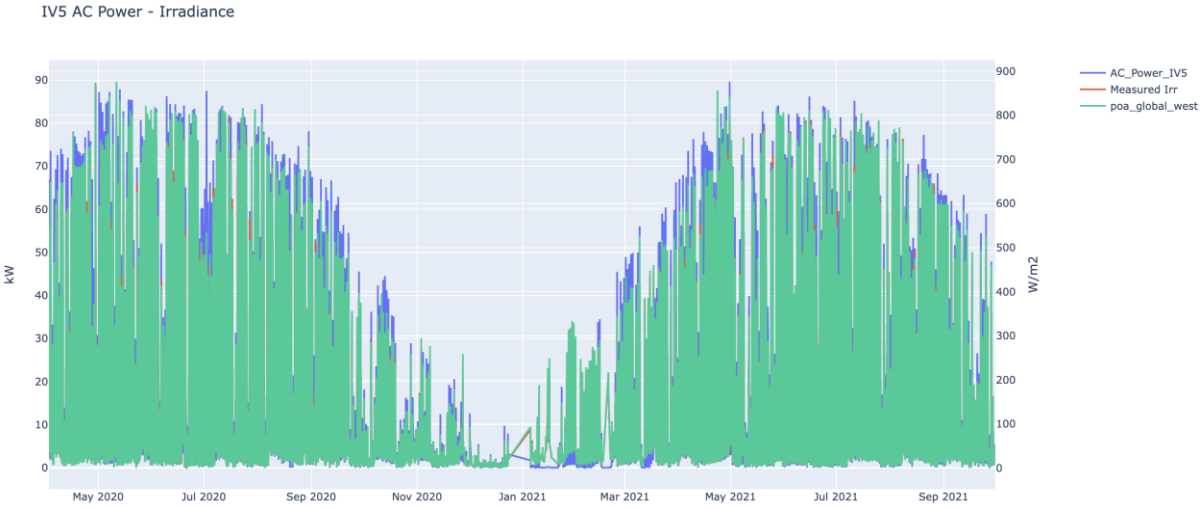


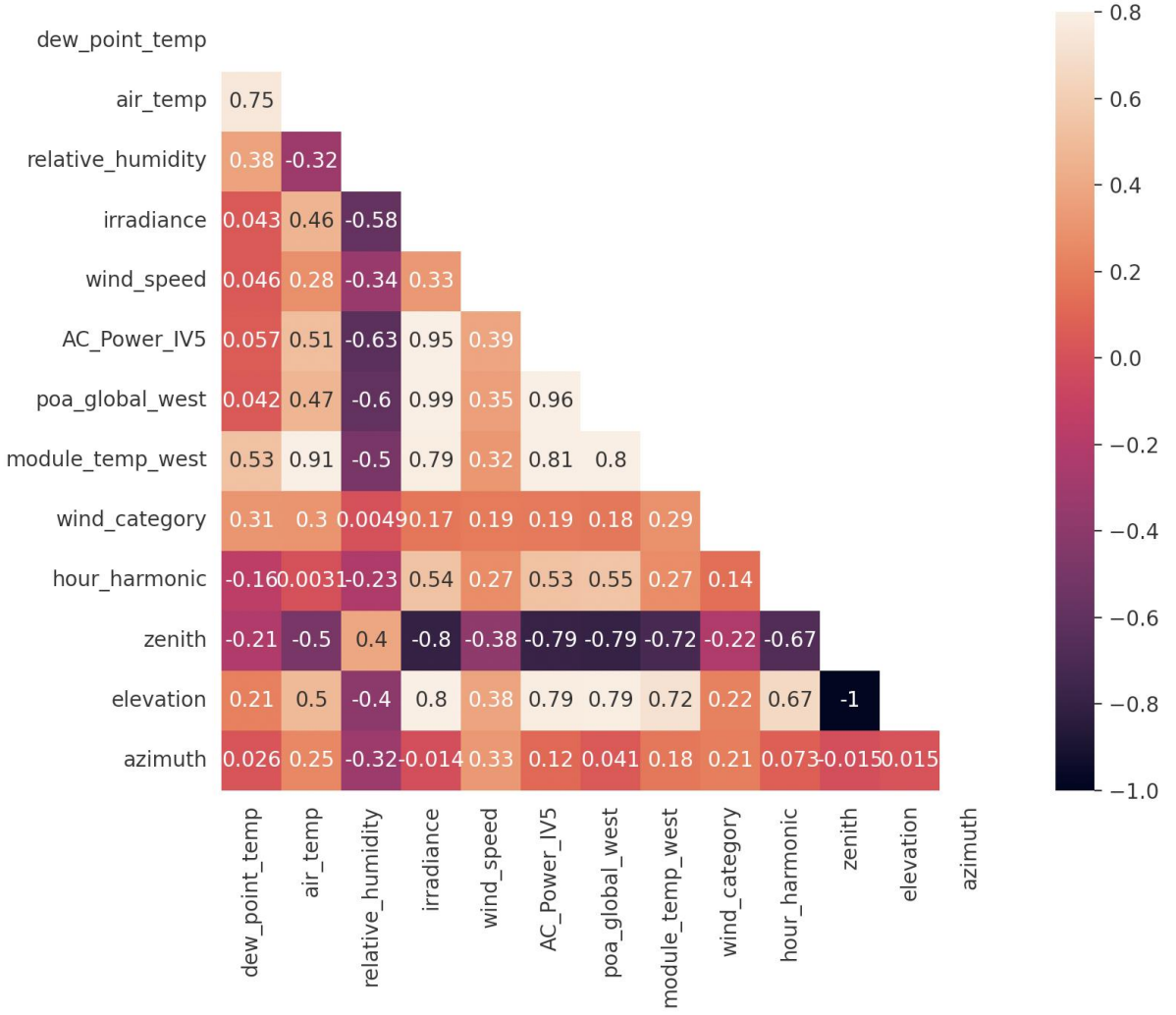
Appendix D: Results Data

1. PCA

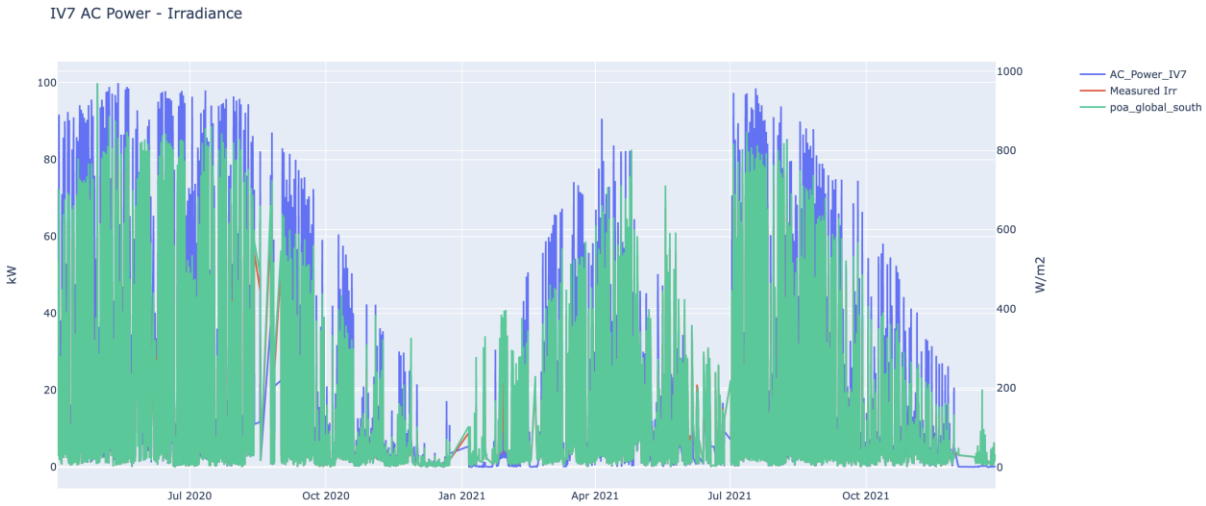


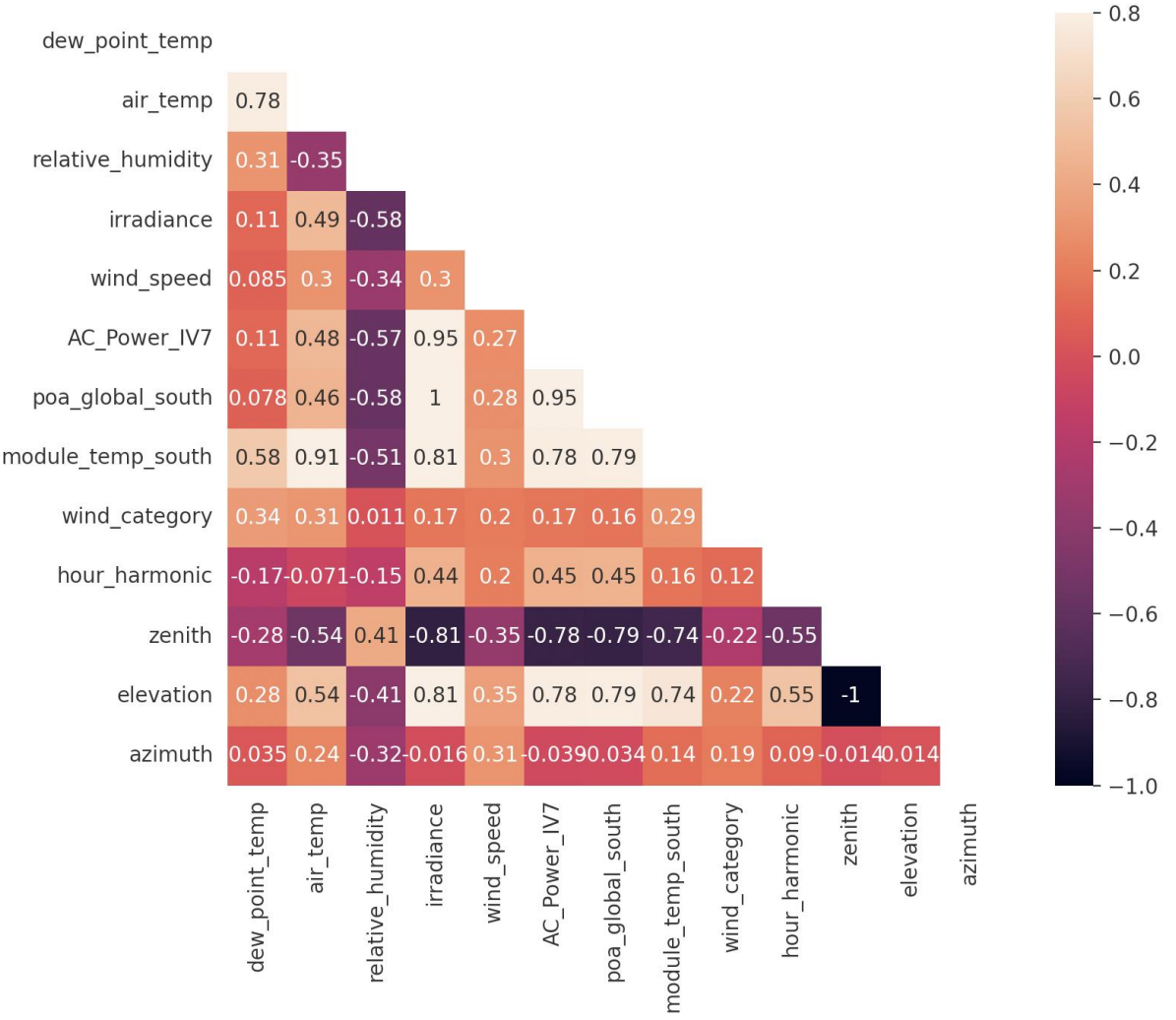
2. IV5 Analysis



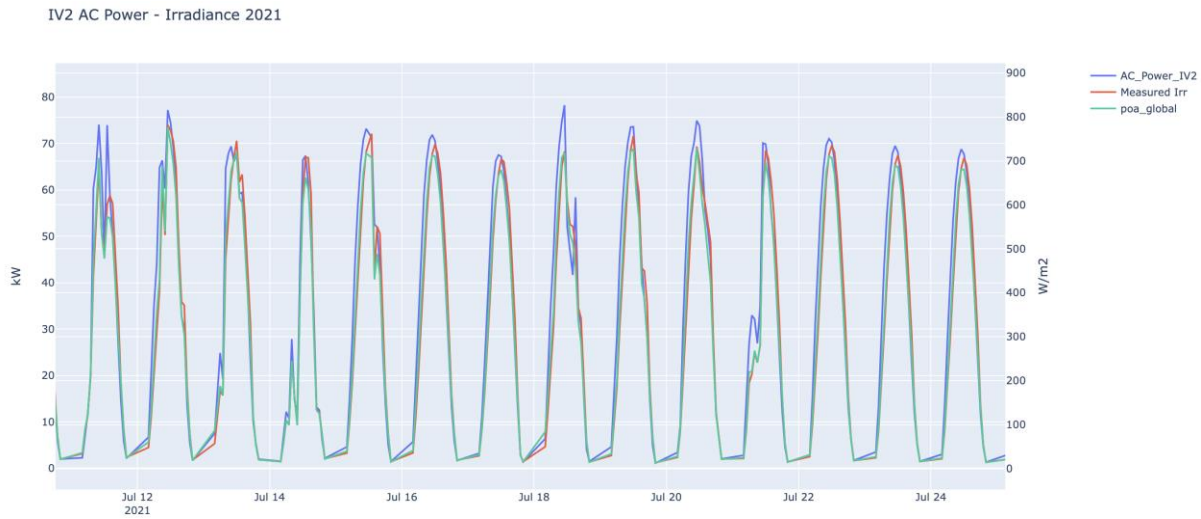


3. IV7 Analysis



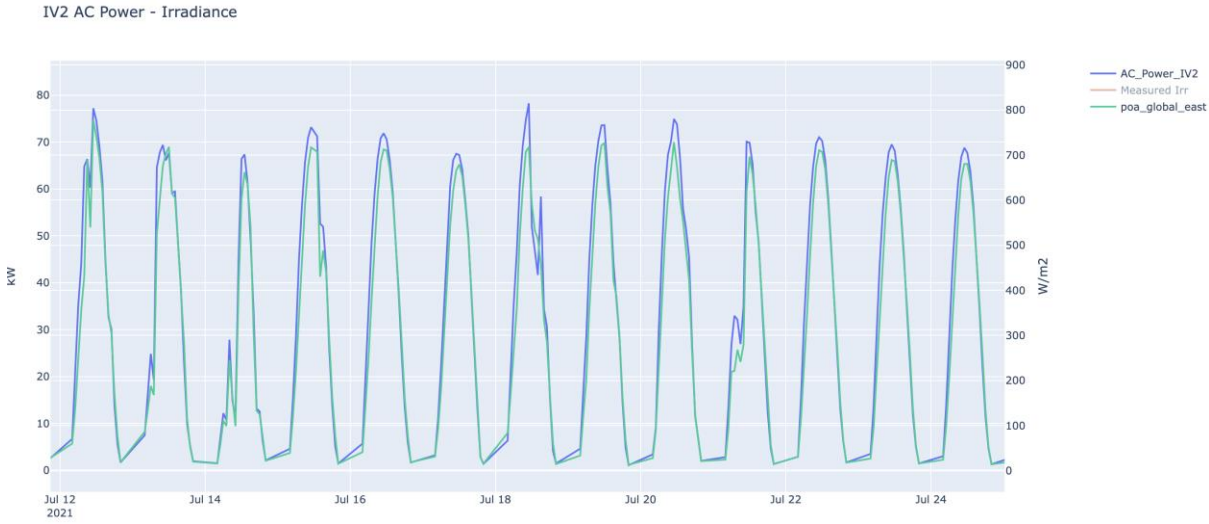


4. IV2 Power - Irradiance

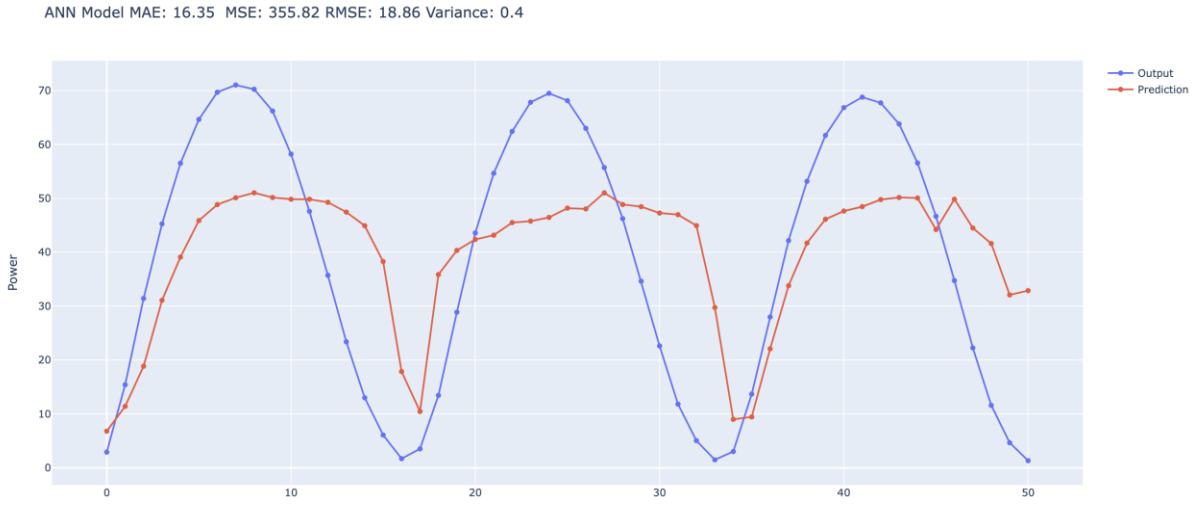


Appendix E: Prediction

1. Historical PV output and irradiance for clear sky day case study.

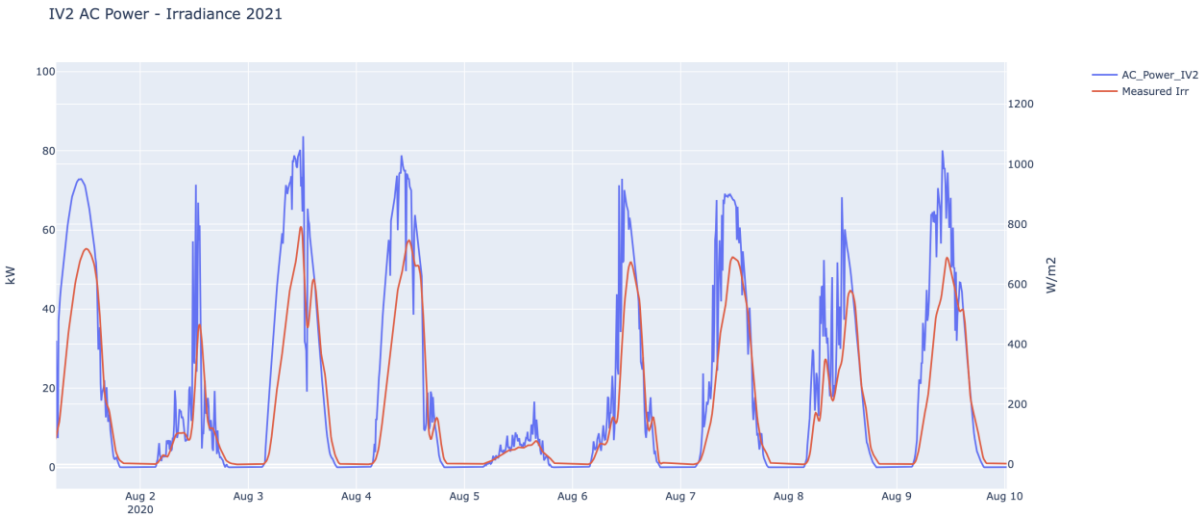


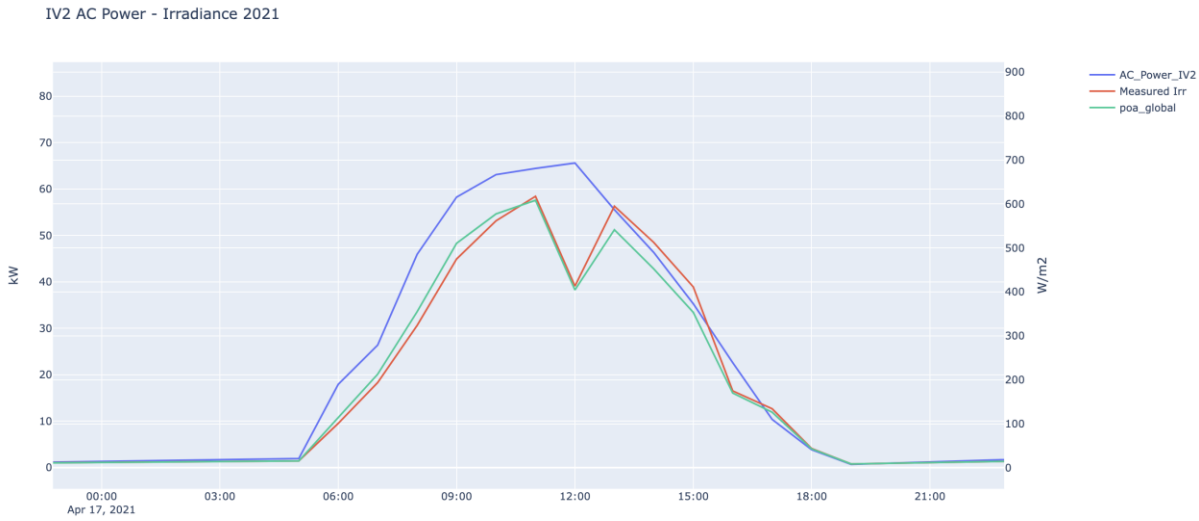
2. Model without irradiance values.



Appendix G: Discussion Section

1. Time Resolution Problem





2. Time Resolution Problem on Excel Representation

Raw PV power values:

47654	2021-08-06 10:00:00+00:00	80.69	1	76.02
47655	2021-08-06 10:10:00+00:00	51.97	1	49.37
47656	2021-08-06 10:20:00+00:00	53.53	1	51.87
47657	2021-08-06 10:30:00+00:00	63.11	1	62.57
47658	2021-08-06 10:40:00+00:00	78.01	1	74.72
47659	2021-08-06 10:50:00+00:00	80.26	1	76.88
47660	2021-08-06 11:00:00+00:00	49.76	1	50.49
47661	2021-08-06 11:10:00+00:00	64.04	1	61.2
47662	2021-08-06 11:20:00+00:00	75.11	1	71.66
47663	2021-08-06 11:30:00+00:00	70.26	1	67.32
47664	2021-08-06 11:40:00+00:00	39.2	1	37.25
47665	2021-08-06 11:50:00+00:00	49.69	1	48.5
47666	2021-08-06 12:00:00+00:00	54.39	1	52.42

1h average: 65.24

1h average: 56.07

The hourly based PV power values:

3572	2021-08-06 10:00:00+00:00	50.39	1	48.59
3573	2021-08-06 11:00:00+00:00	67.93	1	65.24
3574	2021-08-06 12:00:00+00:00	58.01	1	56.07

Irradiance Values are in the green box.

31429	2021-08-06 10:00:00+00:00	14.2	18.7	75	431
31430	2021-08-06 11:00:00+00:00	14.1	20.4	67	647.4
31431	2021-08-06 12:00:00+00:00	11.5	21.5	53	749.9
31432	2021-08-06 13:00:00+00:00	11	21.5	51	505.9

Appendix F: Python Codes

1. ANN Code

```
ANN_Sim = True

dataset =
pd.read_csv('/Users/PycharmProjects/pythonProject/Master_Thesis/PV/Combined_MET_PV
_data.csv')

print('Before Filtering', len(dataset.index))
dataset = dataset[dataset['AC_Power_IV2'] != 0]
#dataset = dataset[dataset['IV2_status'] >= 1]
#dataset = dataset[(dataset['irradiance'] >= 50)]
print('Low_Irradiance_Filter', len(dataset.index))
#dataset = dataset[(dataset['elevation'] >= 10)]
print('Elevation_Filter', len(dataset.index))
#dataset = dataset[dataset['Clear_Sky_Detection'] != 0]
print('Clear_Sky_Days', len(dataset.index))
#dataset.to_csv('dataset.csv')
#print(dataset)

df = dataset[['referenceTime',
              'dew_point_temp',
              'air_temp',
              'relative_humidity',
              'poa_global_east',
              'wind_speed',
              'wind_category',
              'module_temp_east',
              'hour_harmonic',
              'zenith',
              'elevation',
              'azimuth',
              'AC_Power_IV2']]

# timerange
select_training = (df['referenceTime'] >= '2020-04-01 00:00:00+00:00') &
(df['referenceTime'] < '2021-07-21 00:00:00+00:00')
df_training = df.loc[select_training]
#print(df_training)
print('traning count', len(df_training))

select_test = (df['referenceTime'] >= '2021-07-22 00:00:00+00:00') & (df['referenceTime'] <
'2021-07-25 00:00:00+00:00')
df_test = df.loc[select_test]
#print(df_test)
```

```
print('test count', len(df_test))

ratio = len(df_test)/len(df_training)
print('ratio=', ratio)

tf.random.set_seed(1234)

df_training.drop(columns=['referenceTime'], inplace=True)
df_test.drop(columns=['referenceTime'], inplace=True)

X_train = df_training.iloc[:, :-1].values  # does not take Power
y_train = df_training.iloc[:, -1].values  # only takes Power

X_test = df_test.iloc[:, :-1].values  # does not take Power
y_test = df_test.iloc[:, -1].values  # only takes Power

print(X_train)
print(y_train)
y_train = np.reshape(y_train, (-1, 1))
print(X_train.shape, y_train.shape)

print(X_test)
print(y_test)
y_test = np.reshape(y_test, (-1,1))
print(X_test.shape, y_test.shape)

print(X_train.shape[1])

print('X_train mean values', np.mean(X_train, axis=0))
print('X_train std values', np.std(X_train, axis=0, dtype=np.float32))
print('X_test mean values', np.mean(X_test, axis=0))
print('X_test std values', np.std(X_test, axis=0, dtype=np.float32))
print('y_train mean values', np.mean(y_train, axis=0))
print('y_train std values', np.std(y_train, axis=0, dtype=np.float32))

if (ANN_Sim):
    # scaling
    sc_X = MinMaxScaler()
    X_train = sc_X.fit_transform(X_train)
    X_test = sc_X.transform(X_test)

    sc_y = MinMaxScaler()
    y_train = sc_y.fit_transform(y_train)
    y_test = sc_y.transform(y_test)


    # defining accuracy of the function
    def model_input(n_layers, n_activation, kernels):
        model = tf.keras.models.Sequential()
        for i, nodes in enumerate(n_layers):
```

```
        if i == 0:
            model.add(Dense(nodes, kernel_initializer=kernels, activation=n_activation,
input_dim=X_train.shape[1]))
            #model.add(Dropout(0.1))
        else:
            model.add(Dense(nodes, activation=n_activation, kernel_initializer=kernels,
input_dim=X_train.shape[1]))
            #model.add(Dropout(0.1))

    model.add(Dense(1))
    optimizer = tf.keras.optimizers.Adam(learning_rate=0.0001)
    model.compile(loss='mse',
                  optimizer=optimizer,

metrics=[tf.keras.metrics.RootMeanSquaredError(name="root_mean_squared_error",
dtype=None)])
    return model


seq_ANN = model_input([32, 16], 'relu', 'glorot_uniform')
print(seq_ANN.summary())

hist = seq_ANN.fit(X_train, y_train, batch_size=32, validation_data=(X_test, y_test),
epochs=80, verbose=2)

pd.DataFrame(hist.history).plot(figsize=(8, 5))
plt.grid(True)
plt.gca().set_ylim(0, 0.4)  # set the vertical range to [0-1]
plt.show()

plt.plot(hist.history['root_mean_squared_error'])
plt.plot(hist.history['val_root_mean_squared_error'])
plt.title('Root Mean Squares Error')
plt.xlabel('Epochs')
plt.ylabel('error')
plt.legend(['train', 'validation'], loc='upper left')
plt.show()

print(seq_ANN.evaluate(X_train, y_train))

y_pred = seq_ANN.predict(X_test)  # get model predictions (scaled inputs here)
y_pred_orig = sc_y.inverse_transform(y_pred)  # unscale the predictions
y_test_orig = sc_y.inverse_transform(y_test)  # unscale the true test outcomes


def mean_absolute_percentage(y_test_orig, y_pred_orig):
    mape = np.mean(np.abs((y_test_orig - y_pred_orig) / y_test_orig)) * 100
    return mape
```

```
    rmse_ANN = round(mean_squared_error(y_test_orig, y_pred_orig, squared=False), 2)
    mse_ANN = round(mean_squared_error(y_test_orig, y_pred_orig, squared=True), 2)
    r2_ANN = round(r2_score(y_test_orig, y_pred_orig), 2)
    mae_ANN = round(mean_absolute_error(y_test_orig, y_pred_orig), 2)
    mape_ANN = round(mean_absolute_percentage(y_test_orig, y_pred_orig), 2)

    print('RMSE (Mean Squared Error):        ', mse_ANN)
    print('RMSE (Root Mean Squared Error):   ', rmse_ANN)
    print('Mean Absolute Error:              ', mae_ANN)
    print('Mean Absolute Percantage Error:   ', mape_ANN)
    print('R2 Score:                         ', r2_ANN)

    classifier = KerasClassifier(build_fn=model_input,
                                 batch_size=10,
                                 nb_epoch=100)

    accuracies = cross_val_score(
        estimator=classifier,
        X=X_train,
        y=y_train,
        cv=10
    )

    mean = accuracies.mean()
    variance = accuracies.std()
    print(f'K cross mean {mean}')
    print(f'K cross variance {variance}')


    train_pred = seq_ANN.predict(X_train)  # get model predictions (scaled inputs here)
    train_pred_orig = sc_y.inverse_transform(train_pred)  # unscale the predictions
    y_train_orig = sc_y.inverse_transform(y_train)  # unscale the true train outcomes

    print('Root Mean Squared Error Real Values Train', mean_squared_error(train_pred_orig,
y_train_orig, squared=False))
    print('R2 Score Train Values', r2_score(train_pred_orig, y_train_orig))

    np.concatenate((train_pred_orig, y_train_orig), 1)
    np.concatenate((y_pred_orig, y_test_orig), 1)

    fig = go.Figure()
    fig.add_trace(
        go.Scatter(x=results.index, y=results['Real Solar Power Produced'], name='Output',
mode='lines+markers'))
    fig.add_trace(
        go.Scatter(x=results.index, y=results['Predicted Solar Power'], name='Prediction',
mode='lines+markers'))
    fig.update_layout(title=f'ANN Model MAE: {mae_ANN}  MSE: {mse_ANN} RMSE:
{rmse_ANN} Variance: {r2_ANN}',
                      xaxis_title='Time',
```

```
                yaxis_title='Power')
fig.update_yaxes(title_text='Power (kW)')
fig.show()
```

2. LR Code

```
if(linear_regression_sim):


    y_train = y_train.reshape((-1,))

    cv = KFold(n_splits=10, random_state=1, shuffle=True)

    regr = linear_model.LinearRegression()
    # Train the model using the training sets
    regr.fit(X_train, y_train)
    # Make predictions using the testing set
    y_pred = regr.predict(X_test)


    print('Coefficients: \n', regr.coef_)
    # The mean squared error
    print("Mean squared error: %.2f"
          % mean_squared_error(y_pred, y_test))
    # Explained variance score: 1 is perfect prediction
    print('Variance score: %.2f' % r2_score(y_pred, y_test))
    MSE = round(mean_squared_error(y_test, y_pred), 2)
    #r2_score = round(r2_score(y_test, y_pred), 2)

    def mean_absolute_percentage(y_test_orig, y_pred_orig):
        mape = np.mean(np.abs((y_test_orig - y_pred_orig) / y_test_orig)) * 100
        return mape


    rmse_LR = round(mean_squared_error(y_test, y_pred, squared=False), 2)
    mse_LR = round(mean_squared_error(y_test, y_pred, squared=True), 2)
    r2_LR = round(r2_score(y_test, y_pred), 2)
    mae_LR = round(mean_absolute_error(y_test, y_pred), 2)
    mape_LR = round(mean_absolute_percentage(y_test, y_pred), 2)

    scores = cross_val_score(regr, X=X_train, y=y_train, scoring='neg_mean_absolute_error',
                             cv=cv, n_jobs=-1)
    print('Cross Validation accuracy scores: %s' % scores)
    print('Cross Validation accuracy: %.3f +/- %.3f' % (np.mean(scores), np.std(scores)))


    y_test = y_test.flatten()
    X_time = X_time.sort_values()

    fig = go.Figure()
    fig.add_trace(go.Scatter(x=X_time, y=y_test, name='Output', mode='lines+markers'))
    fig.add_trace(go.Scatter(x=X_time, y=y_pred, name='Prediction', mode='lines+markers'))
    fig.update_layout(title=f'LR Model MAE: {mae_LR} MSE: {mse_LR} RMSE:
{rmse_LR} Variance: {r2_LR}',
```

```
                xaxis_title='Time',
                yaxis_title='Power')
    fig.update_yaxes(title_text='Power (kW)')
    fig.show()

    def plot_learning_curves(regr, X, y):

        train_errors, val_errors = [], []
        for m in range(1, len(X_train)):
            regr.fit(X_train[:m], y_train[:m])
            y_train_predict = regr.predict(X_train[:m])
            y_val_predict = regr.predict(X_test)
            train_errors.append(mean_squared_error(y_train[:m], y_train_predict))
            val_errors.append(mean_squared_error(y_test, y_val_predict))
        plt.plot(np.sqrt(train_errors), "r-+", linewidth=2, label="train")
        plt.plot(np.sqrt(val_errors), "b-", linewidth=3, label="val")
    plot_learning_curves(regr, X, y)
    plt.show()

    # Learning Curve

    def plot_learning_curve2(train_sizes, train_scores, test_scores, title, alpha=0.1):
        train_scores = -train_scores
        test_scores = -test_scores
        train_mean = np.mean(train_scores, axis=1)
        train_std = np.std(train_scores, axis=1)
        test_mean = np.mean(test_scores, axis=1)
        test_std = np.std(test_scores, axis=1)
        plt.plot(train_sizes, train_mean, label='train score', color='blue', marker='o')
        plt.fill_between(train_sizes, train_mean + train_std,
                         train_mean - train_std, color='blue', alpha=alpha)
        plt.plot(train_sizes, test_mean, label='test score', color='red', marker='o')
        plt.fill_between(train_sizes, test_mean + test_std, test_mean - test_std, color='red',
alpha=alpha)
        plt.title(title)
        plt.xlabel('Training data')
        plt.ylabel(r'MAE')
        plt.grid(ls='--')
        plt.legend(loc='best')
        plt.show()

    plt.figure(figsize=(9, 6))
    train_sizes, train_scores, test_scores = learning_curve(regr, X=X_train, y=y_train,
                                                            cv=5, scoring='neg_mean_absolute_error')
    fig_lr = plot_learning_curve2(train_sizes, train_scores, test_scores, title='Learning curve
for LR')
```

3. Correlation and PCA Analysis

```
if (correlation_plot):
    mask = np.zeros_like(df[features].corr())
    mask[np.triu_indices_from(mask)] = True
    with sns.axes_style("white"):
        f, ax = plt.subplots(figsize=(9, 7))
        ax = sns.heatmap(df[features].corr(), mask=mask, vmax=.8, square=True, annot=True)
    plt.tight_layout()
    plt.show()

if (PCA_analysis):

    X = df.loc[:, features].values
    X = StandardScaler().fit_transform(X)
    print(np.mean(X),np.std(X))

    pca = PCA(n_components=4)
    principalComponents = pca.fit_transform(X)
    principal_df = pd.DataFrame(data=principalComponents, columns=['Principal Component
1', 'Principal Component 2', 'Principal Component 3', 'Principal Component 4'])
    print(pca.explained_variance_ratio_.round(2))

    fig1 = plt.figure(figsize=(8, 8))
    ax = fig1.add_subplot(1, 1, 1)
    ax.set_xlabel('Principal Component 1', fontsize=15)
    ax.set_ylabel('Principal Component 2', fontsize=15)
    ax.set_title('2 component PCA', fontsize=20)
    plt.scatter(principal_df['Principal Component 1'], principal_df['Principal Component 2'])
    plt.show()

    fig2 = plt.figure(figsize=(8, 8))
    ax2 = fig2.add_subplot(1, 1, 1)
    ax2.set_xlabel('Principal Component 1', fontsize=15)
    ax2.set_ylabel('Principal Component 3', fontsize=15)
    ax2.set_title('2 component PCA', fontsize=20)
    plt.scatter(principal_df['Principal Component 1'], principal_df['Principal Component 3'])
    plt.show()

    variance_exp_cumsum = pca.explained_variance_ratio_.cumsum()
    fig, axes = plt.subplots(1, 1)
    plt.bar(range(1, 1+pca.n_components), variance_exp_cumsum, color='#FFB13F')
    plt.xticks(range(1, 1+pca.n_components))
    plt.title('Screeplot of Variance Explained %')
    plt.xlabel('# of PCs')
    plt.show()
```

4. A Plotting Example

```
if(plot_features_2yaxis):

    fig = make_subplots(specs=[[{"secondary_y": True}]])

    fig.add_trace(go.Scatter(
        x=df['referenceTime'],
        y=df['AC_Power_IV2'],
        name='AC_Power_IV2'),
        secondary_y=False,
        )
    # add line / trace 2 to figure
    fig.add_trace(go.Scatter(
        x=df['referenceTime'],
        y=df['irradiance'],
        name='Measured Irr'),
        secondary_y=True,
        )
    fig.add_trace(go.Scatter(
        x=df['referenceTime'],
        y=df['poa_global_east'],
        name='poa_global_east'),
        secondary_y=True,
        )
    fig.update_layout(title_text='IV2 AC Power - Irradiance')
    fig.update_xaxes(title_text='Time')
    fig.update_yaxes(title_text='kW', secondary_y=False)
    fig.update_yaxes(title_text='W/m2', secondary_y=True)
    plotly.offline.plot(fig, filename='ClearSky_Irradiance_Power ' + '.html')
    fig.show()
```