The present work was completed externally in collaboration with ib vogt GmbH and submitted to the DLR Institute of Solar Research at RWTH Aachen University

Model-based Fault Detection for Grid Connected Photovoltaic Plants from Monitoring Data

Master-Thesis

presented by

Ismail, Yehia Ayman Mohamed Saber

Univ.-Prof. Dr.-Ing. Robert Pitz-Paal[a]
M.Sc. Niklas Blum[a]
M.Sc. Armando Toledo[b]
[a]DLR Institute of Solar Research, RWTH Aachen University, Cologne, Germany
[b]ib vogt GmbH, Berlin, Germany

Berlin, December 2019

Acknowledgements

The present work was completed from June to December 2019 externally in collaboration with ib vogt GmbH and submitted to the DLR Institute of Solar Research at RWTH Aachen University under the supervision of Univ.-Prof. Dr.-Ing. Robert Pitz-Paal.

I would first like to thank Univ.-Prof. Dr.-Ing. Robert Pitz-Paal for giving me the opportunity to conduct the research presented in this work.

I would also like to express my appreciation to my supervisor, M.Sc. Niklas Blum, for his guidance through each stage of the process.

In addition, a thank you to M.Sc. Armando Toledo and M.Sc. Abhishek Gawde, who helped and supported me throughout the realization of this work.

Abstract

The objective of this thesis is to build a model that can be used to characterise the components of a photovoltaic power plant with the final aim of enabling an early and accurate detection of faults and thereby reducing the cost of maintenance and minimizing downtime of the components of the power plant.

Historical data from a monitoring system of a PV power plant in Egypt collected over a span of one year is the basis for building the model used in this thesis: The LFM/MPM Model. It is built based on a combination of the features of two different models: Loss Factor Model (LFM), utilized to determine the normalized parameters, and the Mechanistic Performance Model (MPM), utilized to determine the optimized fitting method, in order to make it capable of accurately fitting the performance measurements (power, voltage, current) of one component (e.g. each inverter and string) to different inputs: Irradiance, module temperature and wind speed.

Based on the results of the fitting, the model can describe the behavior of one component in relationship to each of the inputs. The resulting physical coefficients can then be used to predict the real time optimal output of the different components of the plant from instantaneous input measurements.

A comparison between the predicted real time optimal output and the actual measurements will show if deviations exist. In the case of detection of deviations, the model can use more parameters for further analysis to identify faults. An in-depth analysis of selected components can locate the faults.

Contents

List of Figures

List of Tables

Acronyms

CSV Comma Separated Values.

EPC Engineering, Procurement and Construction.

EPM Empirical Parametric Model.

GHI Global Horizontal Irradiance.

GTI Global Tilted Irradiance.

LFM Loss Factor Model.

MPM Mechanistic Performance Model.

MPPT Maximum Power Point Tracker.

MPR Module Performance Ratio.

MSE Mean Squared Error.

O&M Operation and Maintenance.

RMSE Root Mean Square Error.

STC Standard Test Condition.

1 Introduction

Over the past few years, the PV market and technology have demonstrated rapid growth and have become a mature technology for the production of power from renewable energy sources and a widely-used strategy for the generation of electricity on-site [15].

In recent years, research in the field of photovoltaics (PV) has moved from focusing on increasing efficiency to aiming at increasing the reliability of performance in the field, with a focus on the reliability of the installations and the guaranteed lifetime output. This prioritization of reliability and guaranteed lifetime output is not limited to research activities, but is implemented in the field through the continuous monitoring of the majority of PV installations, either through the inverter or through proprietary monitoring hardware and software. The monitoring data is then used for fault analysis through selected fault detection tools, which allow for the quick identification and accurate quantifying of the factors leading to the failure of mechanisms of a power plant. Performance losses or failures can be a result of several factors, including Maximum Power Point Tracker (MPPT) error, electrical disconnection, wiring losses, shading effects, and faulty equipment. Such failures result in lower output power of the PV system and also lead to degradation of the module's properties. The detection and diagnosis of potential failures at an early stage or even prior to occurrence is crucial for the reduction of the cost related to operation and maintenance and system downtime [12].

Although interest among PV project stakeholders to get a precise estimation of the energy yield of the PV systems is growing, they often encounter significant deviations between the estimated and the measured performance of utility-scale photovoltaic power plants which can be a result of several possible malfunctions. In order to be able to address and solve the malfunctions, stakeholders need the tools that enable them to detect faults in an accurate and timely manner.

The growing need to better understand and improve the performance of the existing and future photovoltaic power plants inspired the topic of this thesis, which was written in cooperation with ib vogt GmbH. ib vogt GmbH is a developer, investor and acting Engineering, Procurement and Construction (EPC) contractor of large-scale PV power plants, which provides performance warranties on realized PV plants under the scope of the EPC contract. In order to increase its competitiveness, and be able to provide its services with the highest levels of accuracy and quality, ib vogt GmbH has a growing interest in better understanding the performance of the large-scale photovoltaic power plants.

The main objective of the thesis is the identification of reasons for occurring deviation between the estimated and the measured performance of the components of one selected power plant using computational and statistical tools. The thesis includes research and identification of the most suitable methodology for accurate estimations of performance based on computational data analytics, which facilitate fault detection and evaluate the performance of utility scale PV plants.

Such accurate PV measurements and performance models are important to understand and enhance the system energy yield. The LFM/MPM Model, built in this thesis, is based on a combination of the features of two different models: LFM, utilized to determine the normalized parameters, and the MPM, utilized to determine the optimized fitting method, in order to make it capable of accurately fitting the performance measurements (power, voltage, current) of one component (e.g. each inverter and string) to different inputs: Irradiance, module temperature and wind speed.

This combination is applied to performance measurements, like monitored large arrays power at maximum point (P_{MP}), current (I_{MP}) and voltage (V_{MP}) at maximum power point. It allows the prediction of PV performance and the validation of the measurements, as well as the identification and quantification of reasons for under-performance. Data filtering was used to enhance the predictive accuracy of the models.

To do this, first a thorough characterization of the performance of the PV power plant is carried out. The monitoring data of existing PV plants, which will be analyzed during the course of the thesis, is provided by ib vogt. For electricity yield measurements, string level monitoring devices, inverter data logger and grid meters are used. Furthermore, meteorological data is collected from three weather stations at different locations at the PV power plant. This data includes the average global irradiation, the diffusive irradiation, the average ambient temperature and the wind velocity. The method and approach used for this thesis for fault detection and system assessment is applicable for similar systems.

The LFM/MPM Model implemented in this thesis to detect faults was built by one year historical measurements of a 64.1 MWp PV plant that is part of 1.86 GWp solar power complex located near the village of Benban, in the desert 650 km south of Cairo, Egypt. In this power complex up to 148,000 MWh of electricity is produced yearly, using a single axis solar tracking system, which is equivalent to the power needed for about 20,000 Egyptian households.

Moreover, the findings of the fault detection algorithm were validated with actual daily reports provided by the operation and maintenance department of the plant operator, which summarize the malfunctions that have to be addressed.

The thesis begins with a *review of the related literature in chapter 2*, which introduces the main components of a PV power plant relevant for this thesis, explains the importance of monitoring a PV power plant and the common faults that can be found, and explores a number of implemented PV system performance models and the analytic approach that was applied to fit the performance models. *Chapter 3 is concerned with data analytics.* It introduces the experimental setup, the model built and used in the thesis and the analytic approach that is implemented. The chapter also describes the data collection and explains how the model is trained. *The fourth chapter presents the findings* of this work. It evaluates the model performance with the analyzed data and shows the accuracy of the estimation resulting from fitting the model with historical data. It also explains how the model is utilized to characterize the modules of the plant to better understand the system behavior, and implements the fault detection procedure and adapts it to the studied PV plant. The results are then validated with daily reports provided by Operation and Maintenance (O&M). *The final chapter of the thesis is the conclusion* summarizing the main findings and a presentation future research which should be considered.

2 Review of Related Literature and Research

This section first introduces the main components of a PV power plant that will be analyzed in the course of this thesis to characterize the plant performance and detect faults. Afterwards, the importance and the reason for monitoring a PV power plant and the common faults found are presented.

Furthermore, a few PV system performance models that were implemented are explained and the analytic approach that was applied to fit the performance models is described.

2.1 Plant Equipment

Usually, photovoltaic systems consist of the following basic components: substation, cables and cable runs, module fasteners/substructure, modules, combiner boxes, centralised or decentralised inverters and monitoring hardware. The most relevant components for the analysis performed in the course of this thesis are described in this sections.

2.1.1 Modules

The sunlight is converted by photovoltaic modules into electrical energy through the photoelectric effect. The so-called PV generator is represented by a solar string, which consists of a series of connected solar module units, which in turn consist of a series of connected solar cells. Figure (1) displays the PV generator's main component parts.

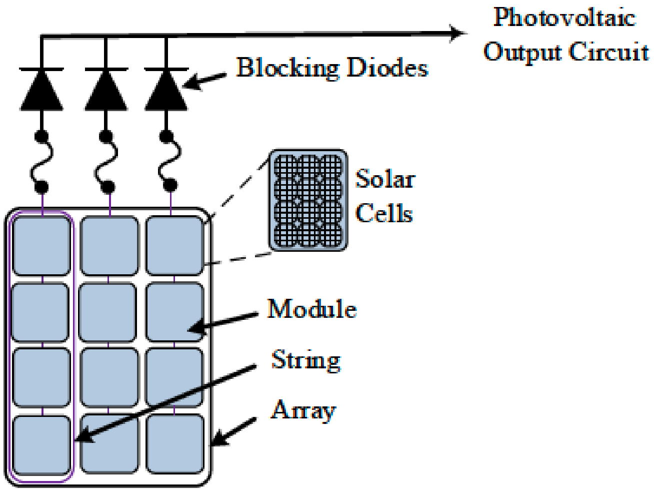


Figure 1: PV generator components [1]

Figure (2) displays the PV module's non-linear I-V and P-V characteristics.

Three crucial points are shown on the curves, namely the short-circuit current I_{SC}, the open-circuit voltage V_{OC} and the MPP. The I-V and P-V curves are both based on the electrical performance of the solar cell and are given on the module data sheet.

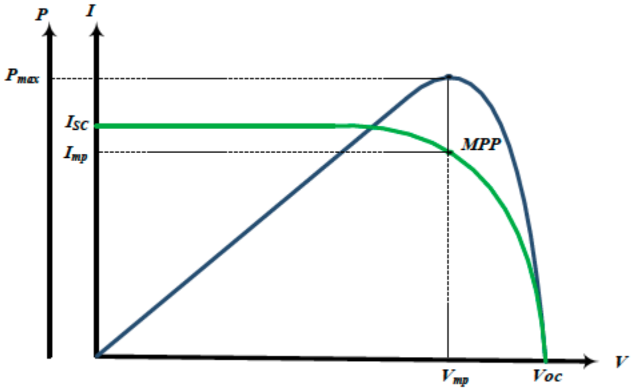


Figure 2: IV Curve of a PV module [1]

The common equivalent model of the solar cell, is presented in Figure (3). The one-diode equivalent circuit is consistent with the following: A current source of photocurrent I_{ph} injection, a single diode reflecting the diffusion phenomenon by I_d, and a shunt resistance R_{sh} limiting the current I_{sh} induced by a solar cell structure manufacturing malfunction.

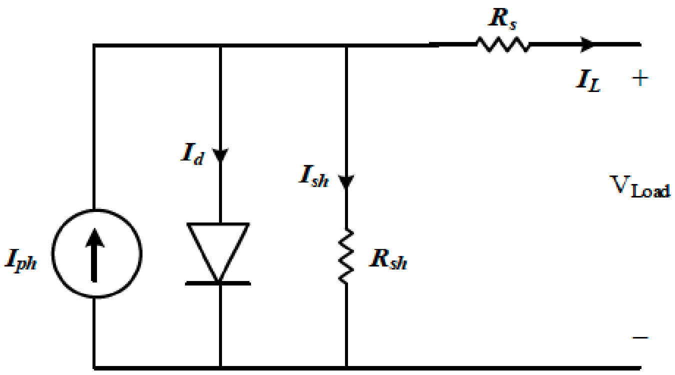


Figure 3: Equivalent circuit for one-diode model [1]

The solar cell generates heat, which lowers the cell efficient and is presented by the series resistance R_s [1].

2.1.2 Inverters

Photovoltaic systems are categorized as centralized or decentralized systems. Several string lines are grouped together in a centralized system – typically up to 100 or more – and then routed to a central inverter, through a combiner box. The described system is identified by a ”small” number of inverters relative to the complete plant.

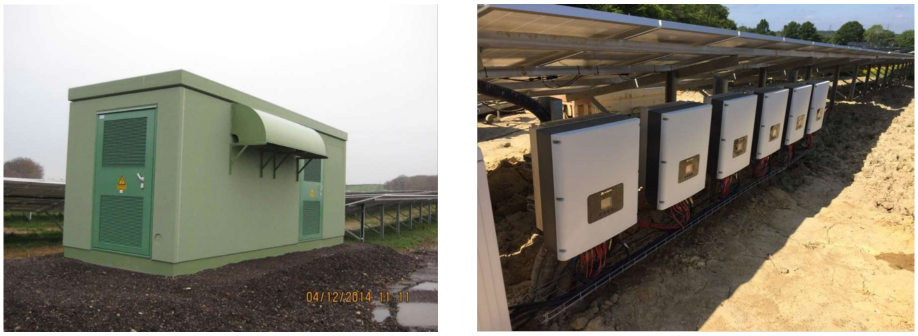


Figure 4: Centralized inverter (left) and string inverters (right) [6]

On the other hand, a small number of strings are separately connected to a string inverter in a decentralized system, which is identified by a rather large number of inverters separated by many sub-distribution system. An example of a centralized and a decentralized inverter is displayed in Figure (4).

The three types of inverter configurations available are central inverters, string inverters and module integrated inverters. Although central inverters are more economic, they have big problems while tracking the maximum power point of the PV array. On the other hand, while the module inverters provide the best option for getting maximum power point for every condition, the are very expensive and complex. Hence, string inverters provide a good trade-off and are usually preferred [8].

2.1.3 Maximum Power Point Tracking

Every inverter is equipped with an MPPT system which insures the maximal energy production from the PV by driving the voltage and current close to MPP.

Cloudy days where the solar energy is delivered with highly fluctuating irradiance cause energy losses due to the non-ideal tracking of the actual position of MPP. On clear-sky days the solar irradiance is changing slowly and is therefore fairly stable which leads to very slow MPP-transitions. In this steady-state the tracking algorithm oscillates around the true MPP and performs the operation of static tracking as shown in Figure (5).

On the other hand, during the rapid irradiance fluctuations caused by passing clouds, the changes happen within very short time frames and this leads to a very high irradiance gradient and this makes the MPP constantly change

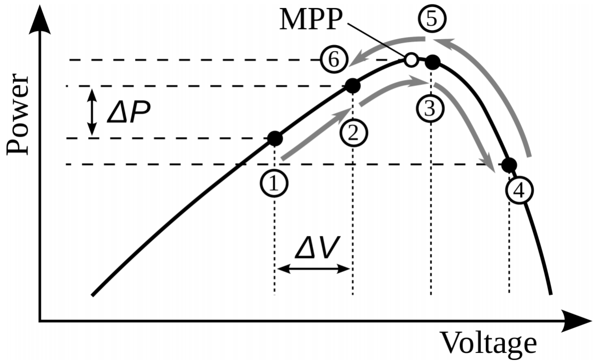


Figure 5: MPPT [13]

positions, that must be followed by a dynamic MPPT. The tracking system is chasing a remote MPP and this leads to a mismatch between the actual operating point of the Inverter and the true MPPT which causes significant energy losses [13].

2.1.4 Combiner Boxes

A combiner box, displayed in Figure (6), is a component that bundles the numerous PV generator strings into one or more main string lines, leading to the inverter.

In addition, the combiner box, together with a central inverter, monitors strings through an integrated module string measuring system, and it provides local voltage protection using overvoltage protection elements.

Where feasible, a combiner box includes a main switch to isolate the inverters, and distinct elements to isolate individual strings, usually as a simple string circuit breaker on the generator array's positive and negative sides.

2.1.5 Tracking

PV modules are typically mounted on a substructure, which holds, and/or supports, the modules and provides both static and dynamic stability. Substructures may be roof-mounted, free field installations, floating, among other forms. Free field substructures are to be classified into fixed-tilt, seasonal-tilt and tracking systems.

In Tracking systems, the substructure design integrates a drive unit, which allows the photovoltaic module to follow the sun to maximize the irradiance received on the module plane, thus increasing the PV generation. Those systems are broadly categorized as single-axis and dual-axis tracking systems.

Single-axis tracking systems have one rotational axis and, almost always, track the position of the sun from east-west. In addition to the east-west track-

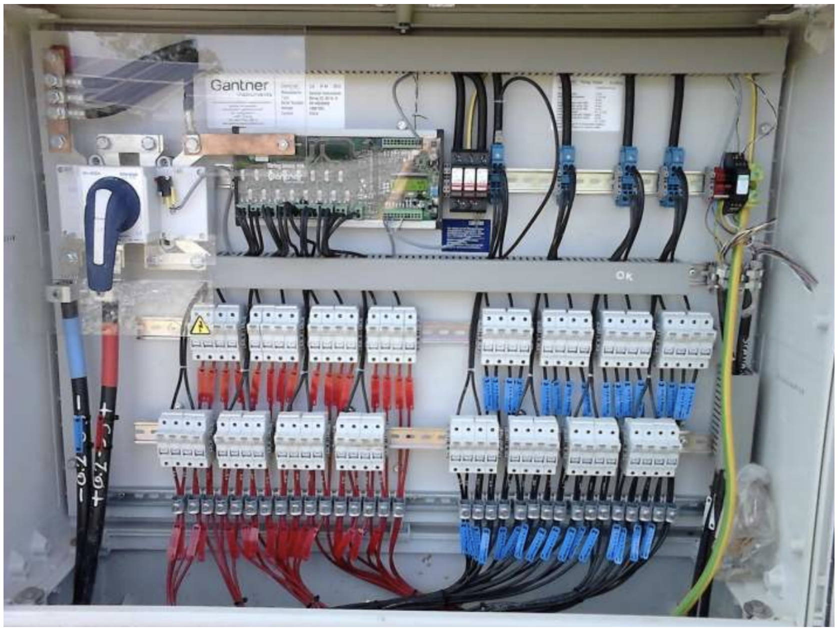


Figure 6: Combiner box [6]

ing, dual-axis tracking systems enable north-south (seasonal) tracking of the sun's position.

Dual-axis tracking systems are able to track the daily and the seasonal movements of the sun with two rotational axes; one along the north-south axis, and another along the east-west one [6].

2.1.6 Monitoring Hardware

To monitor the performance and detect faults in a photovoltaic system, a monitoring tool is usually used, which consists of hardware at the system and software running on a central monitoring portal.

The monitoring hardware, as displayed in Figure (7), is equipped with suitable interfaces to collect relevant data on electrical components, required for monitoring and evaluating the PV system. Electrical components include inverters and module strings, among others. The overall energy production of the complete plant is monitored with energy meters at the grid connection level.

Moreover, the majority of the monitoring systems are equipped with components to register peripheral conditions for the system. An example would be a meteorological station, which measures solar irradiation, temperature, humidity and wind speed at the system's location.

The PV system's characteristics may be monitored and recorded. Examples include the module temperature and the insulation resistance of the PV generator, usually at the inverter.

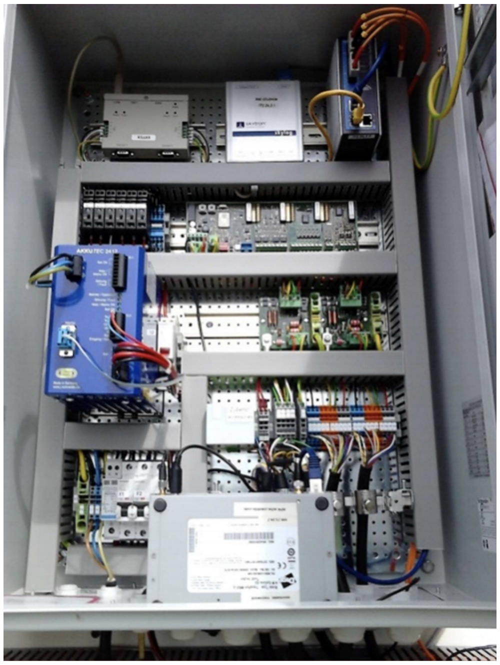


Figure 7: Decentralized monitoring system [6]

One or more data loggers are used to log and evaluate all measured or derived values.

The telecommunications unit is used by both types of the system to send data centrally to an external evaluation medium. It also provides data transmission into the internet via one or more commonly used paths.

For the case of a power cut, the system's individual monitoring stations are commonly equipped with a battery backup system to buffer monitoring. The backup system as well is monitored [6].

2.2 Monitoring

Supervision and monitoring of photovoltaic systems is of high importance with the main purpose of evaluating the performance of the plant, following up on the energy yield and the early detection of system malfunctions. Other reasons for monitoring an expensive and long-term system as a PV plant include the documentation of the performance guarantee, electricity network interaction assessment, system degradation diagnostics and forecasting performance.

In addition to measuring the electrical yield on different levels of the PV plant and the temperature of the modules, it is necessary to collect meteorological data to be able to compare actual monitored production to the estimated production.

For utility scale PV power plants the electricity yield should be measured on different levels of the plant to be able to distinguish for example between faults on the AC-side or the DC-side of the plant. Usually the inverter-integrated measurements alone are not sufficiently precise to analyze the performance of the plant and therefor the power and current on the combiner box level or the string currents should be measured [23].

In addition to that the installment of a pyranometer is recommended for measuring the irradiance in the plane of array. Pyranometers are thermopile sensors based on thermocouple devices.

The location of the sensor should be chosen carefully, because this might affect the accuracy of the readings and with that the performance assessment. Place with near or far shading should be avoided while installing the sensors, even if parts of the plant are affected by shading. Furthermore, it is preferred to have more than one sensor as spread as possible and that way the measurements can be compared and detailed readings can be achieved by eliminating data that is not really representing the actual irradiance. A yearly calibration of the sensors is also of great importance [24].

2.3 Faults

This section defines and introduces the most commonly found faults in photovoltaic arrays. A fault is detected when there's an output power reduction of the PV array compared to the expected output power. These can be caused by faults in a PV module or string of PV modules which may include shading, degradation and corrosion, soiling effect and snow covering, by-pass or shunted diode failure, electrical connections, short circuit, or wiring losses.

Faults in PV arrays can be categorized based on their time characteristic as permanent, incipient, and intermittent. Permanent faults include PV module damages such as short circuit, open-circuit, combiner box faults, and interconnection damage. Incipient faults, on the other hand, can be a result of cells degradation, corrosion, and partial damage in interconnections. It is worth noting that incipient faults can lead to permanent faults. Finally, intermittent faults have temporary effects such as shading, leaf, bird drop, and environmental stress like dust, contamination, snow accumulation, and high humidity. Figure (8) shows the most common types of faults in PV arrays.

The following describes some of the most common faults, that are also relevant to the findings of this thesis, in more detail:

1. Degradation in PV Array

Degradation is an incipient fault that reduces the cell output and may lead to up to 50% power output loss. There are multiple causes for degradation, including the regression of adhesive material between glass and cells, which results in decreasing the light reaching the solar cells and thereby reducing the generated power. Other causes of degradation include delamination, which causes gaps between different subsequent layers of the PV module where the adherence is

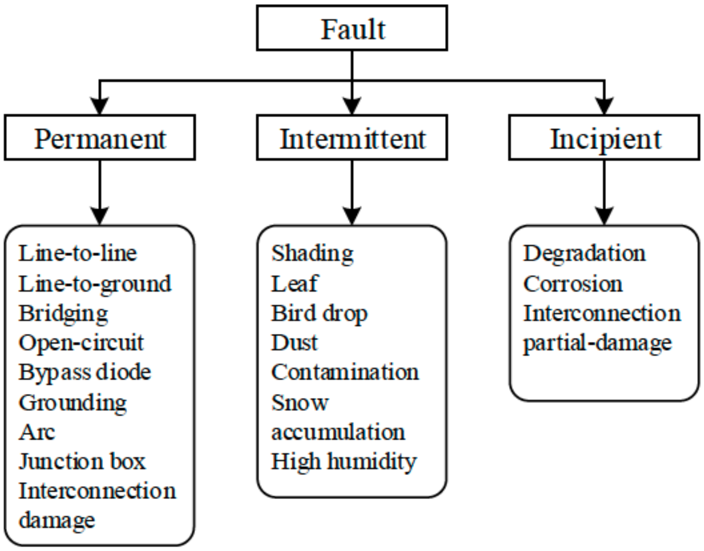


Figure 8: Classification of faults in PV array [1]

lost; as well as the defect in the anti-reflective coating, which reduces the amount of light reaching the cell.

2. Partial Shading Fault

Partial shading is an intermittent fault and refers to covering part of the PV array and thereby reducing output. This can be the result of of passing clouds, smoke, or other temporary effects.

3. Line-to-Line Fault

A line-to-line fault is a permanent fault that involves high fault current or DC arcs between two potential points in the PV array, and causes a reduction in the open-circuit voltage, while the short-circuit current could remain unchanged [26].

4. Open-Circuit Fault

This is another permanent fault that results from disconnection problems in a PV string or more, which are often due to poor soldering in strings interconnections. An open-circuit fault decreases short circuit current and maximum power, whereas open voltage stays close to its normal value.

5. Earth or Grounding Faults

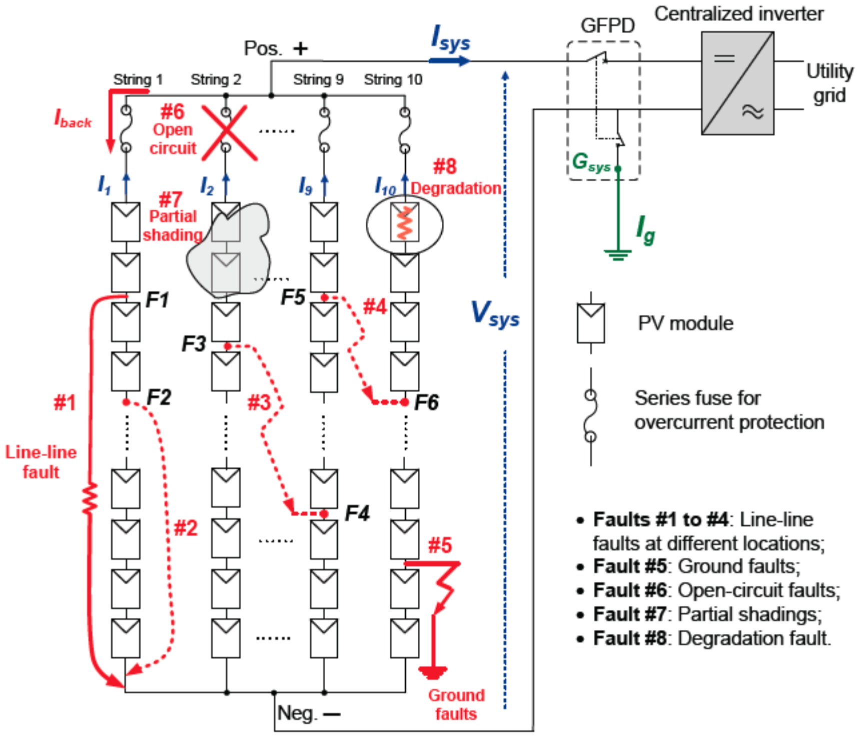


Figure 9: Typical faults in grid-connected PV systems [25]

A ground fault refers to a considerable increase in the current passing through affected conductors, resulting in mismatched currents and changes of the PV array configuration. This type of fault occurs due to an unexpected short-circuited path involving one or more currying current conductors and the ground.

Ground faults can be caused by cable insulation failures and are considered the most common faults in the PV system [1].

Figure (9) visualizes some of the typical faults described above.

2.4 PV System Performance Modelling

When precise PV measurements are available an accurate performance model is vital in understanding and optimizing the energy yield. Since measuring IV curves of single modules is easier than measuring those of strings of arrays, there are more parameters available for individual modules. On the other hand it is often the case that for multi MW_p solar plants the measurements available for the strings either just include the P_{MP} or include also the I_{MP} and the V_{MP}. To make the best use of whichever data parameters are available the model should be adaptable. Therefore the fitting modelling of the Mechanistic Performance Model (MPM) was combined with the normalized Loss Factor Model (LFM) parameters. This section introduces both models in detail and explains the combination of both models.

Furthermore two empirical parametric models are introduced: The David L. King Model and the $\eta(G,T)$ Model.

2.4.1 Loss Factor Model (LFM)

The LFM is a PV module performance model with coefficients relating directly to IV characteristics (see Figure (10)) that allows the characterization of any PV technology by outdoor IV measurements into six independent parameters. These are normalized and physically significant. This leads to the ability of the model to represent technology performance differences and changes over time [20].

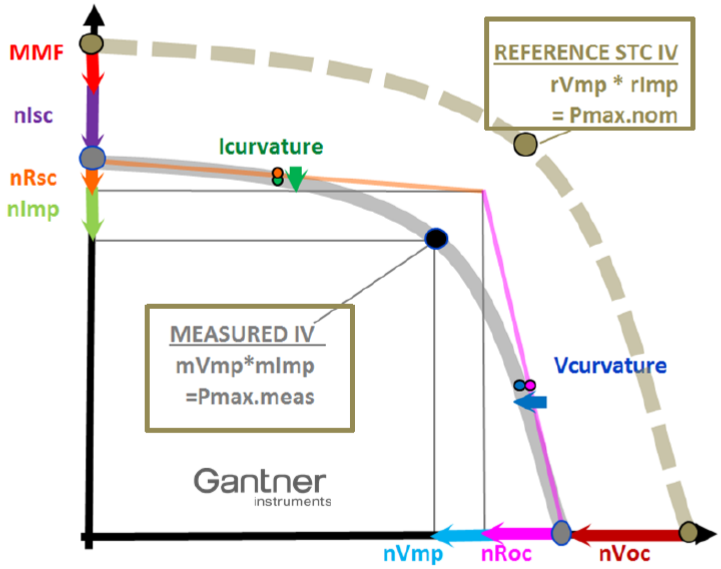


Figure 10: Loss Factor Model Parameters [20]

The LFM consists of a set of normalized parameters which represent each IV curve and of fitting coefficients that describe the variation of the parameters with irradiance in the array plane and module temperature as shown in equation (1).

$$nLFM = C_1 + C_2 * (T_{Mod} - T_{STC}) + C_3 * log_{10}(G_I) + C_4 * G_I \qquad (1)$$

where $nLFM$ represents any normalized parameter to be analyzed with the LFM and G_I is the irradiance, T_{Mod} is the module temperature and T_{STC} is the temperature at Standard Test Condition (STC).

The product of the six LFM parameters result in the normalized efficiency PR_{DC} or the Module Performance Ratio (MPR) as shown in equation (2).

$$\frac{\eta_{measured}}{\eta_{nominal.STC}} = nI_{SC} * nR_{SC} * nI_{MP} * nV_{MP} * nR_{OC} * nV_{OC} \tag{2}$$

where $\eta_{measured}/\eta_{nominal.STC}$ is the efficiency, nI_{SC} is the short-circuit current, nR_{SC} the short-circuit resistance, nI_{MP} and nV_{MP} the current and voltage at MPP, nR_{OC} the open-circuit resistance and nV_{OC} the open-circuit voltage. All the parameters are normalized (prefix = "n") [18].

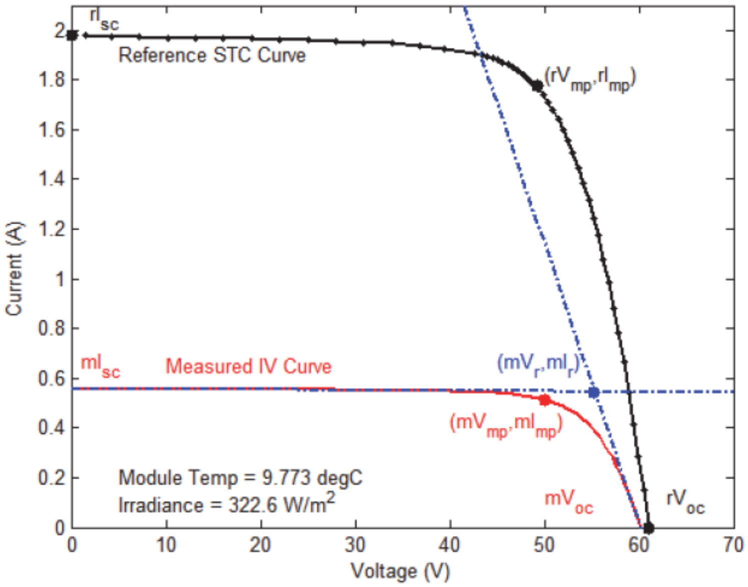


Figure 11: Example measured and reference IV curves showing key points in the electrical coordinate System (V,I) [21]

The main requirement for this model is a measured IV curve (prefix = "m") and a reference IV curve at STC (prefix = "r") shown in Figure (11) that can be calculated with the aid of the module data sheet. Equations (3) to (8) define the normalization of the variables for the LFM. The normalization allows for cross-comparison of different modules or technologies.

$$nI_{SC} = \frac{mI_{SC}}{rI_{SC} * G_I} \tag{3}$$

$$nR_{SC} = \frac{mI_r}{mI_{SC}} \tag{4}$$

$$nI_{MP} = \frac{mI_{MP}}{mI_r} * \frac{rI_{SC}}{rI_{MP}} \tag{5}$$

$$nV_{MP} = \frac{mV_{MP}}{mV_r} * \frac{rV_{OC}}{rV_{MP}} \tag{6}$$

$$nR_{OC} = \frac{mV_r}{mV_{OC}} \tag{7}$$

$$nV_{OC} = \frac{mV_{OC}}{rV_{OC}} \tag{8}$$

mV_r and mI_r represent the coordinates of the intersection point of lines tangent to the ends of the measured IV curve as shown in Figure (11).

Apart from the model being able to predict energy yield over time, it also shows a low bias error. Furthermore, it is capable of fitting a wide variety of modules with different quality. The LFM model can also detect the root causes of degradation and seasonal variation due to the physical meaning of the parameters since they relate directly to the behavior of the key points on the normalized IV curve with changing irradiance. To conclude, the models strengths lies in the quick identification of strange performance patterns through accurate predictions [22].

2.4.2 Mechanistic Performance Model (MPM)

A new, optimized mechanistic performance model was developed by testing 11 different empirical models for outdoor PV monitored data and combining their best features.

The empirical models did not deliver the required accuracy, because of their nonphysical coefficients. Hence, the MPM was proposed with five physical coefficients C1 to C5 defined in equation (9) and explained in table (2). Table (1) shows a comparison between the existing empirical models and the MPM.

Table 1: Empirical vs. Mechanistic Models

Empirical Model	**Mechanistic Model**
not normalized- coefficients values scale with array size	normalized- values independent of array size
nonphysical coefficients	physically significant dependencies are used
not easy to use to compare and contrast arrays of different sizes	easy to validate and compare different sized arrays

$$PR_{DC} = C_1 + C_2 * (T_{Mod} - T_{STC}) + C_3 * log_{10}(G_I) + C_4 * G_I + C_5 * WS \tag{9}$$

where PR_{DC} is the normalized Power on the dc side, T_{Mod} is the measured module temperature (°C), T_{STC} is the STC temperature (°C), G_I is the measured irradiance in the plane of array ($kW/m2$) and WS is the measured wind speed (m/s) [17].

Table 2: Explanation of MPM coefficients

Coefficient	**Dependency**	**Comment**	**Unit**
C_1	Performance Tolerance	Actual/Nominal	%
C_2	Delta T_{Mod}	Temperature Coefficient	$\%/K$
C_3	log_{10}	low light fall	%
C_4	G_I	high light fall	%
C_5	WS	wind speed	$\%/(ms^{-1})$

The MPM is a normalized and optimized model that works well with all PV technologies and can be used to fit outdoor data. The model is very robust and therefore predicts the energy yields with much less variability compared to the empirical models and is able to fit rough data with the least possible errors [16].

2.4.3 LFM/MPM Model

Features from the LFM which is used to define the parameters and the MPM which determines the optimized fitting method have been combined to give an advanced analysis of IV or MPPT data using the same procedure.

The electrical measurements in a PV plant differ based on which monitoring system is available. If a system is installed that sweeps between the short-circuit current and the open-circuit voltage than the IV curve is measured. In other cases the monitoring data just measures the parameters at maximum power point tracking. This leads to a differing number of parameters. Since performance modeling should be able to adapt to the number of parameters available, both the LFM and the MPM were adjusted accordingly.

For the LFM the normalization has been changed as shown in equations (11) to (13) so that the product of all the normalized parameters equals the PR_{DC} as shown in equation (14) It was also shown that the LFM can analyze fewer parameters as opposed to it previously only being implemented to analyze six independent parameters from the IV curve. Furthermore, the MPM which was originally developed only to fit the PR_{DC} is generalized to fit all the normalized LFM parameters according to equation (10).

$$nLFM = C_1 + C_2 * (T_{Mod} - T_{STC}) + C_3 * log_{10}(G_I) + C_4 * G_I \tag{10}$$

$$PR_{DC} = \frac{mP_{MP}}{rP_{MP} * G_I} \tag{11}$$

$$nI_{DC} = \frac{mI_{MP}}{rI_{MP} * G_I} \tag{12}$$

$$nV_{DC} = \frac{mV_{MP}}{rV_{MP}} \tag{13}$$

$$PR_{DC} = nI_{DC} * nV_{DC} \tag{14}$$

This model allows for the prediction of the optimum PV system output and thereby the validation of the instantaneous measurements in real time, which in return makes the reason for any under-performance or faults to be easily identified and quantified to minimize any downtime errors. By integrating the LFM, seasonal effects and degradation can also be identified and quantified [19].

The LFM/MPM Model requires the use of two weather data sets: these are the irradiance in the plane of array and the wind speed. The module temperature is also used as an input for the model to calculate the output variable. Furthermore, the output variables are needed to fit the curves according to the LFM/MPM Model. These are defined in equations (11 - 13).

For the normalization, the reference voltage and current of the module need to be extracted from the module data sheet.

While Mechanistic Models are based on an understanding of the behavior of a component in a system, Empirical Models are based on direct observations, measurements and large-scale data records.

2.4.4 Empirical Parametric Model (EPM)

The **David L. King Model** is a straight-forward model for predicting array performance and works for all operation conditions.

The model is described in equations (15) and (16) which reflect linear relationships closely related to the fundamental electrical characteristics of cells in the module [10].

$$I_{MP}(E_e, T_c) = E_e * (I_{MPo} + \alpha_{I_{MP}} * (T_c - T_o)) \tag{15}$$

$$V_{MP}(E_e, T_c) = V_{MPo} + C_2 * ln(E_e) + C_3 * (ln(E_e))^2 + \beta_{V_{MP}} * (T_c - T_o) \tag{16}$$

where $I_{MP}(E_e, T_c)$ and $V_{MP}(E_e, T_c)$ the current (A) and the voltage (V) at MPP, E_e the effective irradiance, I_{MPo} and V_{MPo} the current and the voltage at MPP at the reference cell temperature T_o (°C). Furthermore $\alpha_{I_{MP}}$ (A/°C) and $\beta_{V_{MP}}$ (V/°C) are the temperature coefficients for the I_{MP} and the V_{MP} and finally T_c is the temperature of the cells inside the modules (°C).

The parameters required for this model can be easily acquired through outdoor measurements and the performance is related to the cell temperatures and not the module temperatures and thereby compensating for the situation where the modules are not in thermal equilibrium [11].

The **$\boldsymbol{\eta(G, T)}$ Model** represents a simple approach for estimating the grid-connected PV-System MPP performance on the dc side in dependence of irradiance in the array plane and module temperature. First a model for the dependence of the efficiency at MPP operation on the irradiance G is introduced:

$$\eta_{MPP}(G) = a_1 + a_2 * G + a_3 * ln(G) \tag{17}$$

where $\eta_{MPP}(G)$ the efficiency at MPP, G the irradiance in the plane of array and $a_1 - a_3$ device specific coefficients are.

Since the equation represents the efficiency at 25°C another equation (18) is applied that allows modeling of the performance at all operation temperatures.

$$\eta_{MPP}(G,T) = \eta_{MPP}(G, 25°C) * (1 + \alpha * (T - 25°C)) \tag{18}$$

where T is the operation temperature and α is the temperature coefficient.

The model uses the measured power output to fit the parameters and due to its structure simple linear fit procedures may be applied. [2]

2.5 Model Performance Accuracy Metrics

To evaluate the predictive models' output accuracy, two common performance metrics are applied. This section will introduce these metrics, () and Root Mean Square Error (RMSE), alongside the residuals that form the basis for calculating these metrics.

2.5.1 Residuals

The residual measures the deviation between an observed value and the estimated value of a certain quality and is given by:

$$r_t = y_t - \widehat{y_t} \tag{19}$$

where y_t and $\widehat{y_t}$ are the actual measured and the corresponding predicted value by the model.

2.5.2 R Squared (R^2)

In statistical literature, the R^2 measure is referred to as the coefficient of determination. Its value lies between 0 and 1 and indicates in how far a set of predictions fit to the actual measured values, 0 indicating no-fit and 1 being perfect-fit.

However, R^2 is not the optimal measure for assessing the fitting accuracy, as it only explains the proportion of variation in the dependent variable that is explained by the independent variable. Moreover, R^2 may not describe the importance of a variable, because when a new variable is added, the error decreases, hence R^2 always increases when a new variable is added to the model, without any addition information provided by this variable. R^2 is given by:

$$R^2 = 1 - \frac{\sum_{t=1}^{m}(y_t - \widehat{y_t})^2}{\sum_{t=1}^{m}(y_t - mean(Y))^2} \tag{20}$$

where y_t and $\widehat{y_t}$ are the actual measured and the corresponding predicted value by the model and m is the number of samples for the calculation [4]

2.5.3 Root Mean Square Error (RMSE)

Similar to the Mean Absolute Error, Mean Squared Error (MSE) provides a gross idea of the magnitude of error. Units can be converted back to the original units of the output variable, by taking the square root of the mean squared error, which is referred to as the Root Mean Squared Error (RMSE).

RMSE is another measure of accuracy, which can be used to compare forecasting errors of different models for a selected dataset. In general, a lower RMSE indicates lower levels of errors than a high RMSE. However, RMSE is sensitive to outliers because the effect of each error on RMSE is proportional to the size of the squared error. As a consequence, larger errors have a disproportionately large effect on RMSE.

RMSE is given by:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{m}(y_t - \widehat{y_t})^2}{m}} \tag{21}$$

where y_t and $\widehat{y_t}$ are the actual measured and the corresponding predicted value by the model and m is the number of samples for the calculation [7].

2.6 Analytic Approach

2.6.1 Linear Regression

Linear Regression is an algorithm that was developed in the field of statistics and borrowed by machine learning, more specifically the field of predictive modeling. This field is essentially concerned with making the most accurate predictions by minimizing the error of a model.

Due to the simplicity of the representation, the Linear Regression is perceived as an attractive model to understand the linear relationship between input and output numerical variables. The output variable y is either calculated from a single input variable x in the case of a simple linear regression or from a combination of multiple variables in the case of a multiple linear regression.

To each input value or column in a linear equation one scale factor, referred to as a coefficient, is assigned. An additional coefficient, usually called the intercept or the bias coefficient, gives the line an added degree of freedom. Equation (22) shows an example of a model in a simple regression problem with a single x and a single y.

$$y = B_0 + B_1 * x, \tag{22}$$

where B_0 is the bias coefficient and B_1 is the coefficient for input value or column.

A good set of coefficient values are found through implementing a learning technique, after which different input values can be plugged in, in order to predict the output. When a coefficient equals zero, it removes the influence of the input variable on the model and thereby removes it also from the prediction made from the model ($0 * x = 0$).

The equation mentioned above could be plotted as a line in two-dimensions, by plugging in several input values, predicting output values and thereby creating a line as shown in an example in Figure (12). Making the predictions is as simple as solving the equation for a specific set of inputs, since the representation is a linear equation.

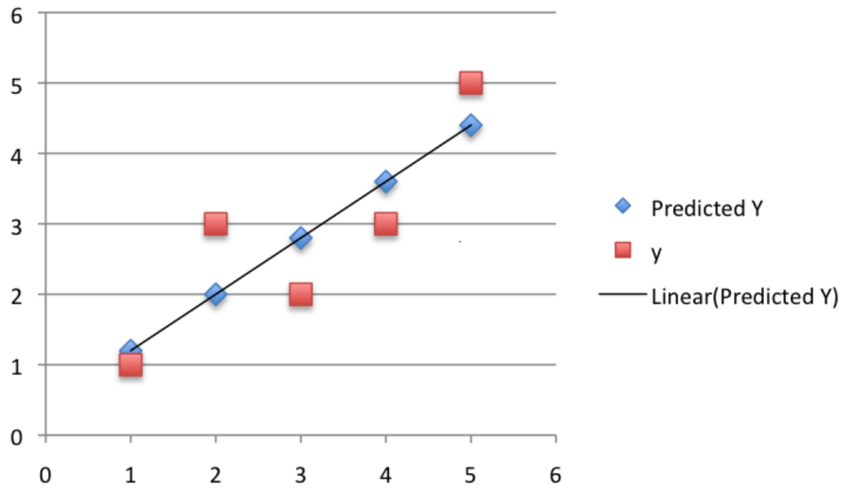


Figure 12: Simple Linear Regression Predictions.

In higher dimensions with more than one input (x), the line is referred to as a plane or a hyper-plane and in such cases, the representation is defined by the equation and the specific values used for the coefficients (B_0 and B_1 in the above example).

Learning a linear regression model means estimating the values of the coefficients used in the representation using available data. This section briefly examines one technique of preparing a linear regression model.

The most commonly used method is the Ordinary Least Squares. Although with simple linear regression with a single input, statistics can be used to estimate the coefficients, requiring the calculation of statistical properties from the data such as means, standard deviations, correlations and co-variance; this method is not really useful in practice . One of the main requirements of this method is the availability of all of the data to traverse and calculate statistics.

Ordinary Least Squares can be used when there is more than one input to estimate the values of the coefficients. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals, meaning that given a regression line through the data, the distance from each data point to the regression line is calculated and squared, and all of the squared errors are summed. This is the quantity that Ordinary Least Squares seeks to minimize.

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients, which requires the availability of all of the data and of enough memory to fit the data and perform matrix operations. It is more likely that you will call a procedure in a linear algebra library. This procedure is very fast to calculate [3].

2.6.2 Non-Linear Regression

Nonlinear regression is a type of regression analysis that models observational data by a function that is a nonlinear combination of the model parameters. The function depends on one or several independent variables, and the data are fitted by a method of successive approximations.

Unlike in linear regression, generally, there is no closed-form expression for the best-fitting parameters. Those are usually determined by applying numerical optimization algorithms.

The Ordinary Least Squares introduced in the preceding section again is presumed to be the best curve fit for a non-linear equation. It is based on an approximation based on a linear model and an optimization of the parameters through successive iterations [14]

3 Data Analysis

This chapter is concerned with the data analytics part of this thesis. First the experimental setup, model and the analytic approach that are used are briefly introduced. The data collection is described, before going through the preparation and the understanding of the data. Afterwards the way the model is trained is explained.

The main purpose of this study is to analyze the behavior of PV plant components and as a result detect faults. Therefor a tool was developed in the programming language Python, which allows the analysis of an inverter with all the combiner boxes and strings connected to it. The tool fits curves using the LFM/MPM Model (2.4.3) for the power, current and voltage of each device and then uses the model to predict data. This data is then compared to the actual measured data and the deviations are saved. The deviations allow faults to be detected.

3.1 Experimental Setup

The structure of the plant plays an important role in the analysis and is therefore introduced in this section.

Historical data were collected from a large-scale photovoltaic power plant that has a DC nominal capacity of 64.1 MW_p, generated by almost 200,000 Trina Solar 320/ 325 W_p 72-cell PV modules spread over an area of 954,000 m^2. The entire system has AC power capping at 50,000 kVA. The geographical coordinates of the plant are given by a latitude of 24.447889°and a longitude of 32.716741°. Figure (13) shows the PV power plant.

Figure 13: PV power plant located in Benban, Egypt [6]

The grid-connected PV plant consists of several major components, including a solar PV array, 80 centralized inverters with MPPT algorithm and electrical connection wirings. The inverters are from the manufacturer Schneider Electric, each with a rated output power of 680 kVA AC.

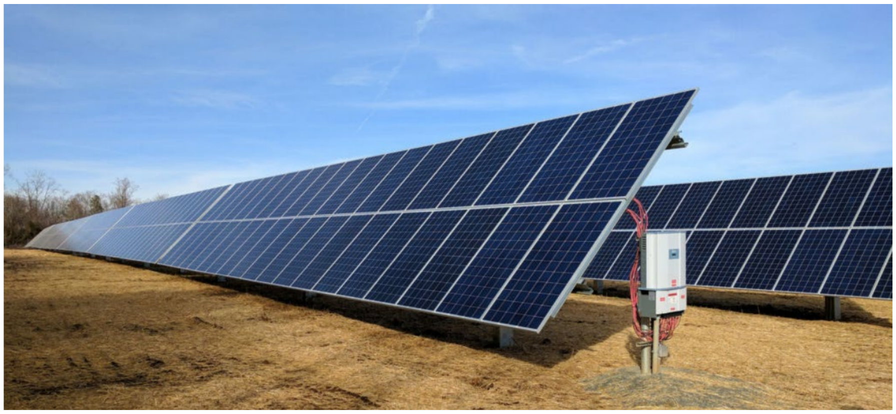


Figure 14: Single Axis Tracker with 2 in Portrait Configuration [6]

20 PV modules, which represent the fundamental building blocks of the PV system, are assembled in series to build a PV string. Then, 24 or in very few cases 12 strings in parallel construct a combiner box. Five of these combiner boxes are connected in parallel to one of the 80 central inverters. An extract from the single line diagram as shown in Figure (15) demonstrates the centralized configuration of the system.

The PV module are mounted on a free field substructure that holds them and provides a static as well as a dynamic stability. A single axis tracking system with two modules arranged in portrait is installed as shown in Figure (14). The single axis tracking system has one rotational axis and tracks the position of the sun from east to west with a rotation limitation between a minimum of -45° and a maximum of 45°.

The actual monitoring system involves two different types of measurements. Built-in electrical measurement channels in system devices (inverters, combiner boxes, additional monitoring devices) and three meteorological stations positioned at strategical locations in the power plant.

Figure (16) shows the different sensors that form the meteorological station. These include a module temperature sensor, a pyranometer in the horizontal plane and one in the array plane, an anemometer and an ambient temperature sensor. The anemometer is an instrument to measure the speed of wind and the pyranometer measures the solar irradiance.

3.2 Model Selection

To achieve a high accuracy for the prediction model three different models were implemented and compared for the course of this thesis.

First the $\eta(G,T)$ model was applied for power, current and voltage, but

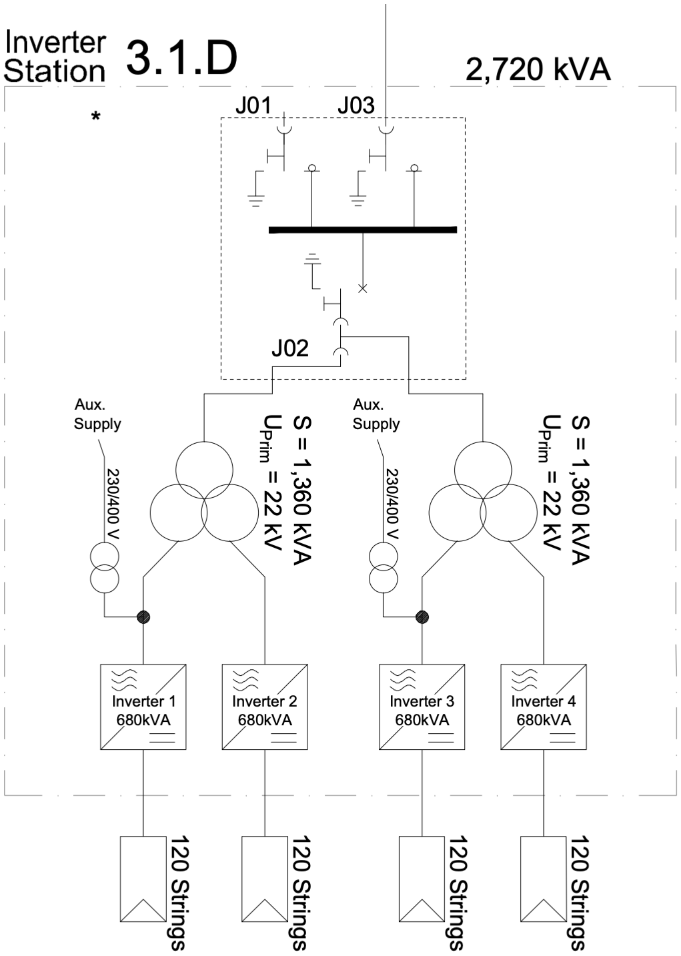


Figure 15: Extract from the single line diagram [6]

only delivered good prediction accuracy for the power. Hence, the equations (15) and (16) from the the David L. King were used for the voltage and current fitting. The reason for implementing these empirical models was their ability to provide a straightforward procedure for fitting the parameters without the need for running many iterations.

The empirical models are partly based on system output dependencies on the irradiation and the module temperature derived for ideal cell characteristics and partly based on direct observations and measurements of large-scale data records. Therefore their implementation did not deliver the required accuracy needed for a precis fault detection procedure, because every plant behaves differently depending on location, weather conditions and plant structure. Furthermore, the empirical models did not help in the understanding of the system

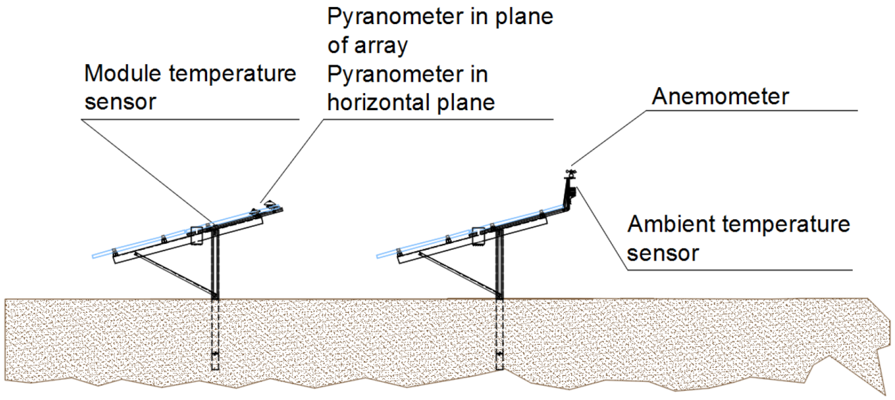

Module temperature sensor
Pyranometer in plane of array
Pyranometer in horizontal plane
Anemometer
Ambient temperature sensor

Figure 16: Example of a meteorological station

behavior and consequently an implementation of a different model was necessary.

Furthermore, as shown in Figure (17) the fitting using the EPM has increasing errors for low irradiance as opposed to the more accurate fit generated by the MPM.

Therefore, the combination of the LFM/MPM model described in (2.4.3) delivered physical coefficients that can be analyzed and thereby can help in characterizing the system. Moreover, the model allowed for more accurate predictions.

Finally, the fact that all parameters are normalized allows for a comparison between different components.

3.3 Analytic Approach Selection

Because of the simplicity of the linear regression and it's ability to produce curved fits, it was implemented as the first analytic approach for the model introduced in (2.4.3). After observing the fitted data compared to the train set, a bad fit was indicated. Therefore, a more complex approach was required.

Figure (18) displays the bad fit in the case of applying a linear regression model. This is visible by observing the residuals especially for low irradiance. An ordinary least squares regression was implemented with a tool called linear_model.LinearRegression() provided by the open-source scikit learn.

On the other hand, it is shown in Figure (19) how the non-linear model is more robust and results in a noticeable increase in the accuracy of fitting shown through the decrease in residuals. The tool optimize.curve_fit() provided by the open-source software SciPy is used for the non-linear least squares fit of a function, f, to data.

By applying a non-linear approach the RMSE for the example above improves from 1.6% to 1.0%.

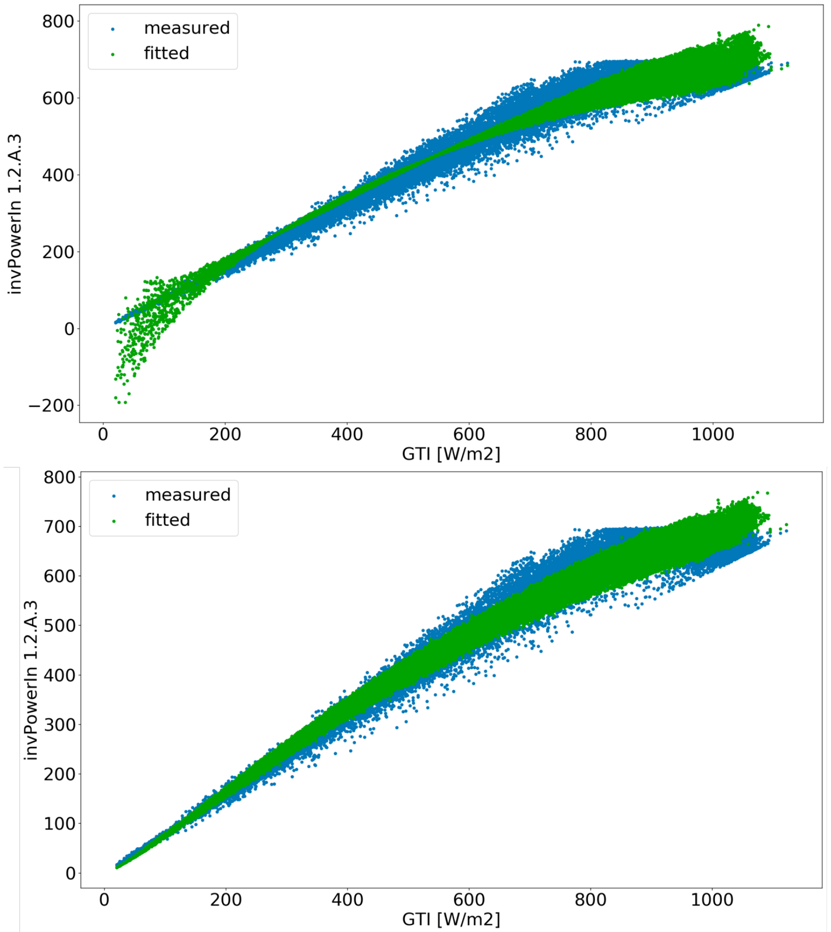


Figure 17: Fitting the inverter DC power with EPM (top) and MPM (bottom)

3.4 Data Collection

To be able to analyze the correlation between weather variables and each component's power, current and voltage generation, a huge amount of electrical measurements and the corresponding weather data is required.

First step of the analysis is collecting the specific system characteristics. This is vital for calculating the performance ratio and the normalized parameters. They consist of the system geographic location and the PV module parameters from manufacturer data sheet, which include the module class $[W_p]$, the reference maximum power point voltage V_{MPP} [V] and current I_{MPP} [A] at STC and the power temperature coefficient t_c [%/°C].

The next step is collecting the monitoring data, which is provided as data packages through a web portal shown in Figure (20) and is managed by a third-

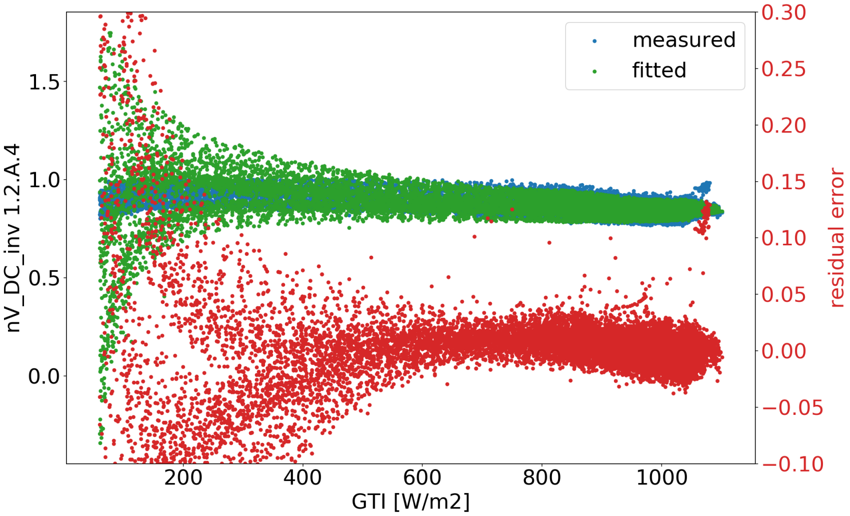


Figure 18: The low accuracy of the linear regression fit represented by the residuals as an example for the normalized voltage of one inverter especially for low irradiance

party, Gantner Instruments. They are specialized in precision, industrial measurement data acquisition and signal conditioning. The data output is received in the file format Comma Separated Values (CSV).

The only analysis the web portal provides is a simple visualization of the monitored parameters over time shown in Figure (21). Very few information of the performance of the plant can be extracted from this and therefore the development of a detailed analysis tool was necessary.

The sampling rate of the measurements is 5 min, covering a period from the 1st of November 2018 till the 31st of October 2019. A long period guarantees a better representation of the system behavior under varying weather conditions. Finally, the data is saved in a dictionary which entails the measurements required for the evaluation process. These include the solar irradiance (Global Horizontal Irradiance (GHI) and Global Tilted Irradiance (GTI)), ambient temperature, PV module temperature, total plant electricity generation at energy meter, AC output power of each inverter, DC input power, voltage and current of each inverter, DC output voltage of each combiner box and DC input current of each combiner box measurement channel.

Furthermore, parameters that are used for calculating the reference values from the module data sheet are collected. These include the number of strings per inverter and the number of strings per combiner box.

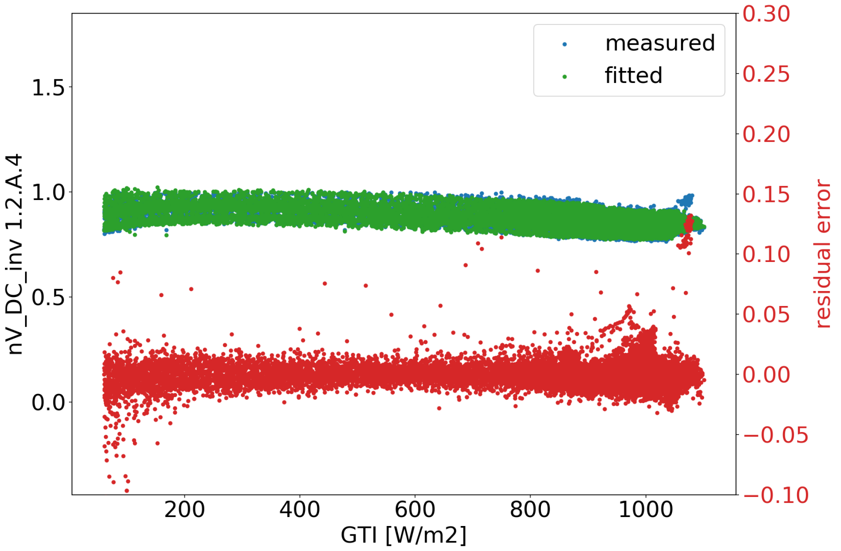


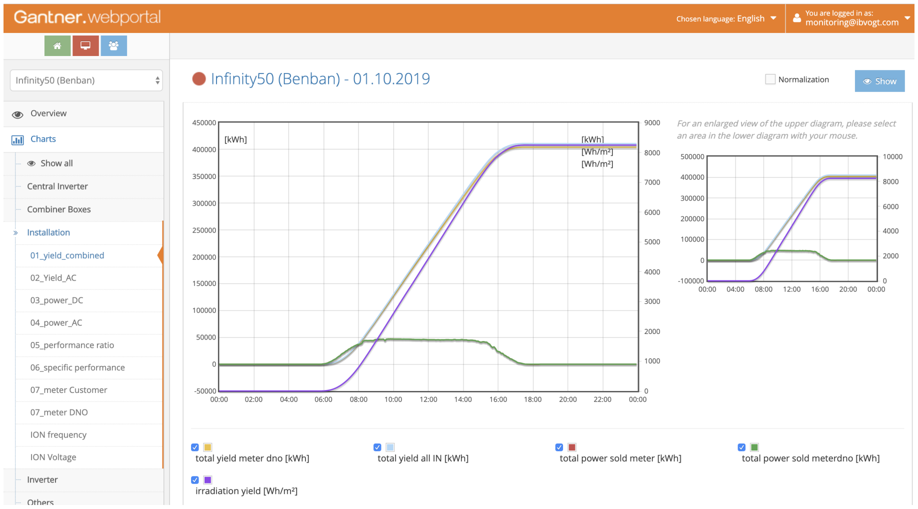
Figure 19: The high accuracy of the non-linear regression fit represented by the residuals as an example for the normalized voltage of one inverter for the complete irradiance spectrum

3.5 Data Preparation

This section introduces the manipulation of the data collected from the monitoring system as a preparation step before the start of the actual analysis. This is crucial and it involves cleaning, transforming, reshaping the data. Its main purpose is to guarantee high consistency and low uncertainty of the analysis results.

The evaluation of the availability of the data is one of the first steps performed in the process of data preparation. Due to the great importance of the influence of the irradiance on the energy generation of the plant, the missing timestamps are determined based on the irradiance dataset and erased for all the datasets.

The irradiance, as the backbone of all measurements, is manipulated by a set of standard rules. For the rules to be applicable to similar datasets, they are represented as functions to automate the application. One of the functions replace the negative readings of the irradiance with a zero. Another function performs the averaging over the three meteorological stations for all weather data and the PV module temperature, except when measurements are detected that are out of range for the particular variable. It compares the readings from the three different weather stations and removes the data that shows strong deviations, if it is only the case for one station, before calculating the mean of the measurements from the remaining stations. That way if one of the readers was not working properly it would not contaminate the data.



For an enlarged view of the upper diagram, please select an area in the lower diagram with your mouse.

Figure 20: Gantner web portal

The power plant is equipped with three different meteorological stations spread as much as possible with the purpose of getting the best possible accuracy and avoiding contaminating the irradiance measurements by a cloud that might be passing and covering only one area of the plant. An example of a cloud passing first over one meteorological station and afterwards passing over another one is shown in Figure (22).

First to get the power output at grid level of the plant the readings from three different feeders are summed. If the readings are negative the positive values are set to zero and the absolute value of the power is taken and else the negative power values are set to zero. The reason this is done is because the energy meters sometimes deliver negative power readings instead of positive ones.

After processing the data that is used as inputs for all the models or for setting filters for all components in the plants now the preparation commences for the different components that will be analyzed.

For the inverter the name is specified and given a string input. Then the tool builds a different data frame for each component, including all combiner boxes and strings connected to the specified inverter, separately. This includes the power on the DC side, the voltage and the current as the main outputs of the different models.

Furthermore, depending on which component is being analyzed more variables are required for calculating the times where there were outages. This is used later for filtering the data. For instance, for the inverter a function has the outages as an output and takes the power on AC side and the power of the combiner boxes connected to this inverter as an input.

For the normalization of DC power, current and voltage more inputs and

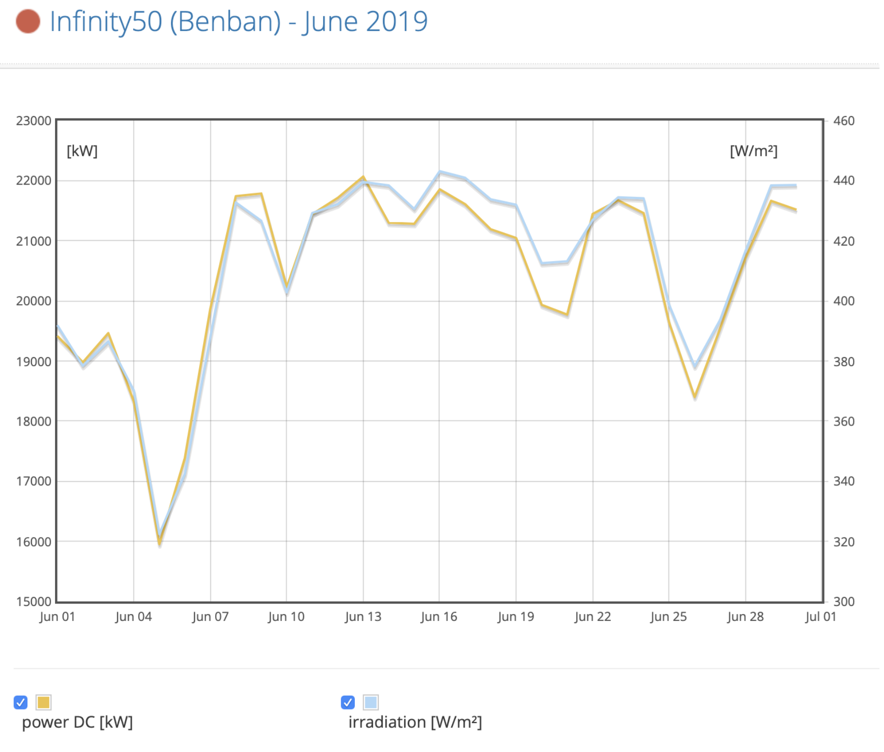


Figure 21: Time-based visualization of the irradiance and the DC power for one month provided by the web portal

calculations are required. The normalization functions are demonstrated below. The power, current and voltage are divided by the reference values at standard test conditions (STC). Furthermore, the power and current are also divided by the irradiance in the plane of array. Since the modules are connected in series, the reference voltage of one module is multiplied by the number of modules connected to one string. The reference voltage at combiner box and inverter level are the same as on string level since all the strings and combiner boxes are connected in parallel. The reference current of a string is equal to the reference current of one module, but is multiplied by the number of strings connected to the studied component, either inverter or combiner box, to determine the reference current at the different levels. The reference power is calculated by multiplying the reference current with the reference voltage. A function calculates the number of strings connected to one component.

To save memory while running the code only the necessary variables for one component are added to the data frame used for the regression model and the predictions later on. It contains the module temperature, irradiance in the plane of array and the wind speed which constitute the model inputs and

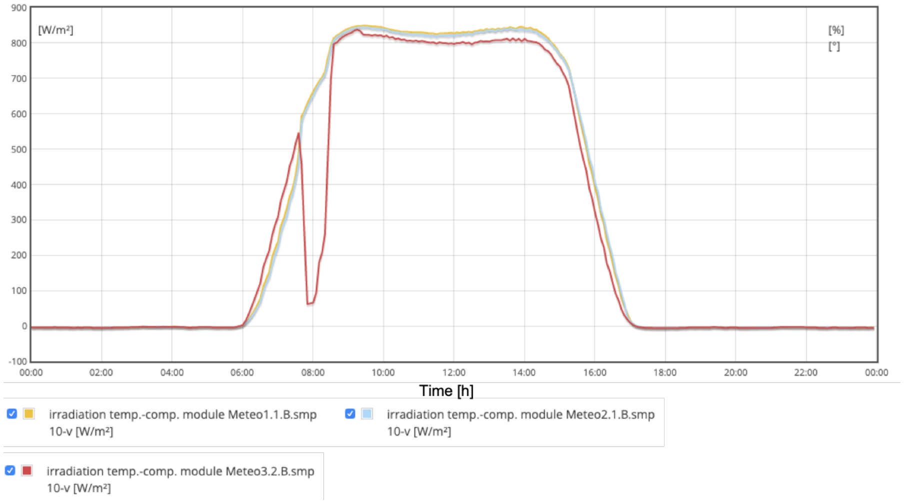


Figure 22: Drops in irradiation readings from single meteo stations caused by the passing of a cloud

the normalized power, current and voltage which constitute the model outputs. There are three different functions for building the data frames for the different types of components in the plant. The weather corrected performance ratio of the plant as a whole is also added to each data frame to be used for filtering later. The weather corrected performance ratio is calculated using formula (. . .).

The data frame is ready to proceed with the preparation. A new function is implemented for filtering the data frame. First, the data frame is copied into a version for fitting the model. This copy is first cleaned for all outages detected before and then the data is filtered for the clear days by eliminating all the data corresponding to cloudy days. Furthermore, another function allows the user to specify the threshold for the data to perform the regression for. This is done by adding filters on many variables for instance removing the data where the overall performance of the plant is either too low or unrealistically high. The data corresponding to low irradiance in the plane of array is also eliminated. The influence the filtering has on the model results are discussed in the Findings section.

3.6 Data Understanding

Before starting the analysis an understanding of the data is of vital importance.

A very fast way of understanding the data is through visualization. This is done with the help of the Matplotlib in Python. It represents an initial analysis for the data which gives a first impression of the system's behavior.

By plotting the GHI and the GTI as displayed in Figure (23) the effect of having a tracking system is visible. The amount of irradiance over the course of

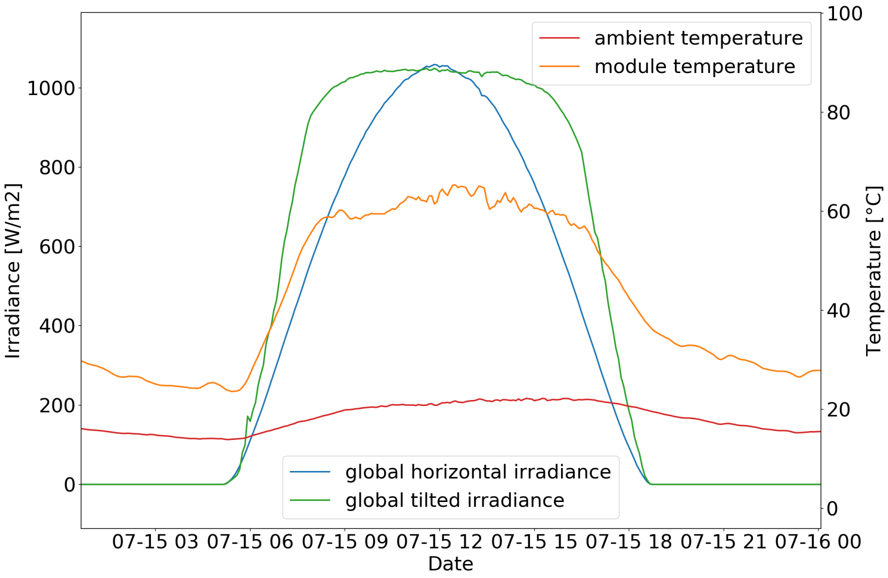


Figure 23: GHI and GTI, ambient and module temperature for a typical day

the day that can be transferred by the PV modules into electricity in the case of a tilted pyranometer that follows the sun position is much higher. Furthermore, the increase of the module temperature with higher ambient temperatures and irradiance can also be seen in the figure.

To get a first overview of the performance of the plant the weather corrected performance ratio is calculated and plotted for a complete year. This is shown in Figure (24) and the seasonality affecting the performance of the plant throughout the year is observable.

Another benefit from visualizing the data is understanding the correlation between the variables. Moreover, by implementing a colour scale to the plot the simultaneous influence of two variables on a third variable can be seen. Figures (25) and (26) show the DC voltage of one inverter plotted over the irradiance in the array plane and the module temperature.

Figure (25) shows the linear dependency of the voltage on the module temperature and how the voltage decreases with increasing temperatures. Meanwhile Figure (26) displays the minimal dependency of the voltage on the irradiance and it can also be seen that this small decrease of voltage with increasing irradiance is mainly due to the correlation between the heating up of the modules with increasing irradiance.

Finally, Figure (27 the linear increase of the current with higher irradiance and the negligible dependency of the current on the module temperature.

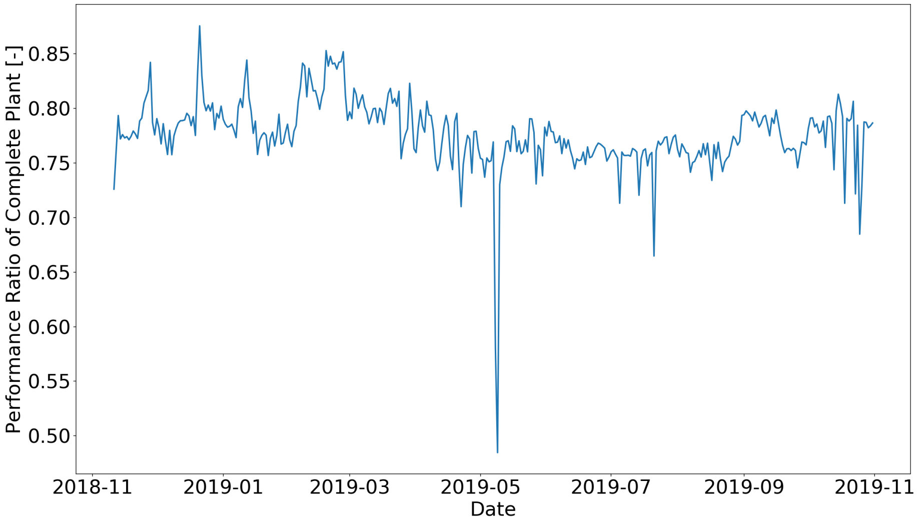


Figure 24: Weather corrected performance ration of the complete year over one year

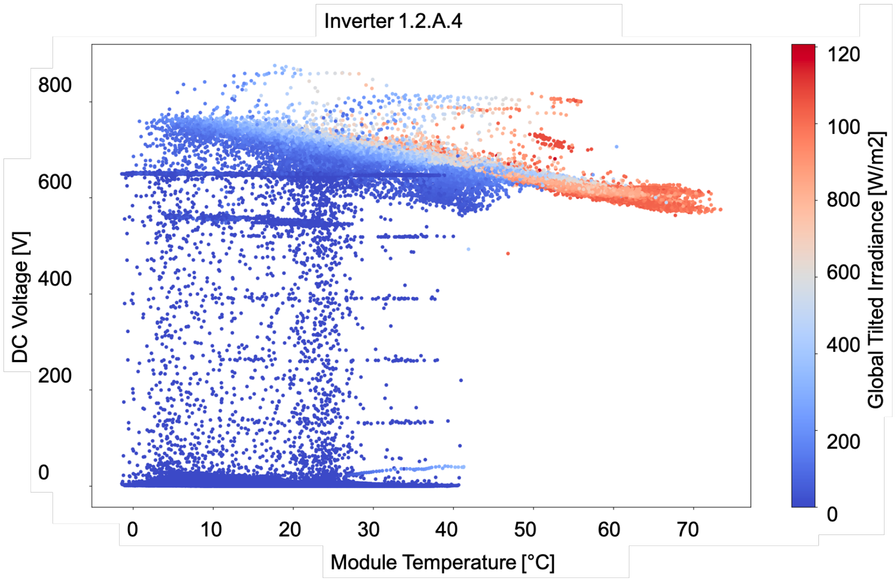


Figure 25: DC voltage of one inverter plotted over the module temperature and the GTI(color scale)

3.7 Model Training

First step for training the model is splitting the aquired datasets into two subsets. The first subset is used to train the model and the second one is used to

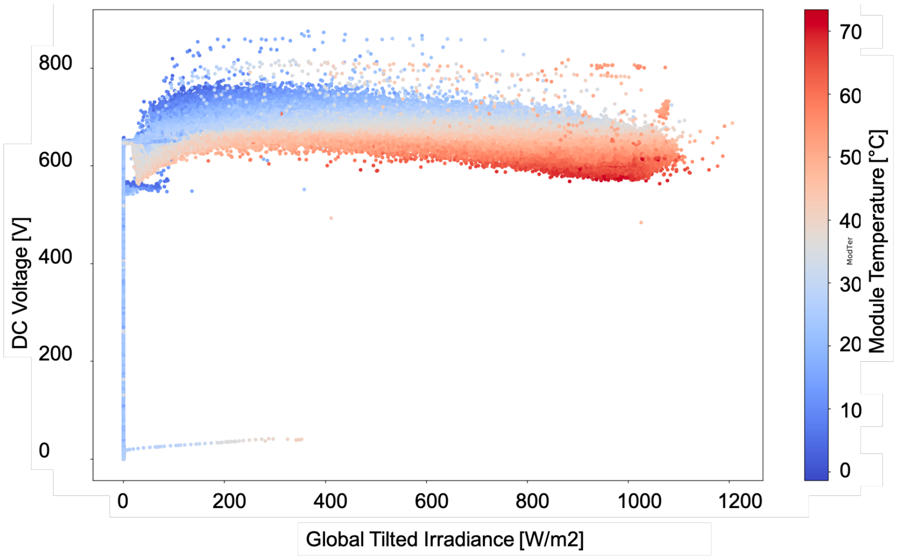


Figure 26: DC voltage of one inverter plotted over the GTI and the module temperature(color scale)

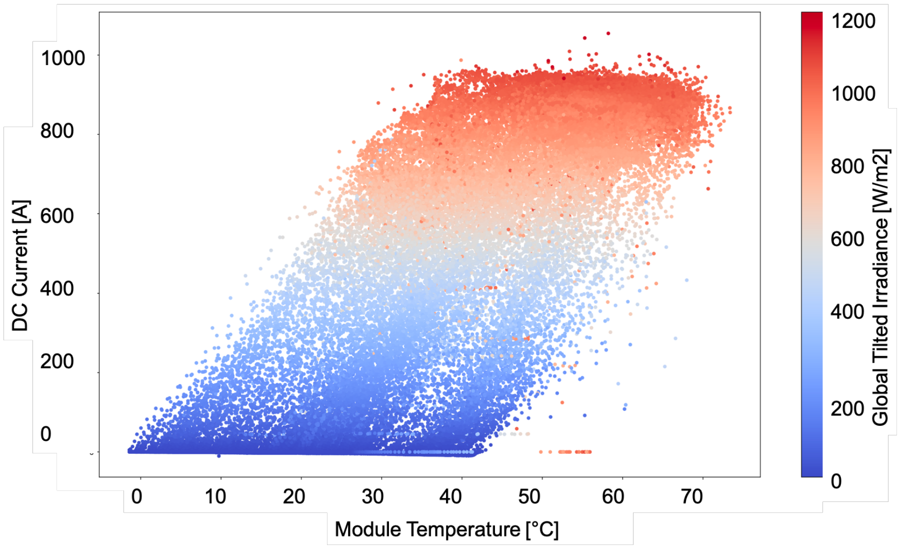


Figure 27: DC current of one inverter plotted over the module temperature and the GTI (color scale)

test how accurate the model is fitting the data and that relates to how good the model is able to predict the output for new sets of inputs. The sklearn library in python provides a function named train_test_split that randomly spits the data according to a specified ratio for training and testing the model.

During the course of this thesis 33% of the measurements were randomly selected as the test set for the year analyzed in a uniform distribution manner. Different model training conditions were then used and tested in order to develop an effective model with high accuracy based entirely on the data acquired.

Next curve fitting is applied to the train data with an optimization based on the least-squares minimization. This implementation fits the MPM equation defined in (9) with a function provided by the open-source software SciPy called optimize.curve_fit.

3.8 Fault Detection and Identification Procedure

In this section, the procedure for finding and classifying a few examples of the many faults detectable with the developed tool are introduced.

The main purpose of the tool is to assess when and where a fault occurred in the PV plant and identify the type of the fault. Because of the low impact for instance one faulty string out of 10,000 would have on the overall performance of the complete plant the analysis is applied to one inverter with all its combiner boxes and strings. By running the tool for each inverter separately a detailed fault allocation is possible. For this thesis one inverter is considered as an example to show the effectiveness of the procedure.

The real time measured irradiance, module temperature and wind speed are used as inputs to the LFM/MPM model for each component to determine its optimal expected DC power, current and voltage. Moreover, to be able to conclude information about the inverters the model is also implemented for the power of the AC side of the inverters.

To first decide whether or not a fault occurred at a certain time, a reference is taken based on the absolute value of the difference between the measured and the simulated PR_{DC} of the analyzed inverter as shown in equation (23). A deviation of zero or close to zero means no fault has occurred. The high accuracy of the model was discussed above, but since the overall accuracy of the procedure also depends on the accuracy and the sensors and components, a tolerance is needed. A threshold of 0.07 was set to account for system losses, measurement noise and losses.

$$|PR_{DC\_meas} - PR_{DC\_sim}| > 0.07 \tag{23}$$

Now that a fault has been detected the tool proceeds with the identification of the fault type. This requires the calculation of ratios between the estimated and the measured power, voltage and current to allow for conditioning and locating of the faults. These include the DC current ratio, the DC voltage ratio and the DC-AC power ratio of the inverter and given by equations (24) to (26).

$$R_c = \frac{I_{DC\_sim}}{I_{DC\_meas}} \tag{24}$$

$$R_v = \frac{V_{DC\_sim}}{V_{DC\_meas}} \tag{25}$$

$$RP_{DC\_AC} = \frac{(\frac{P_{DC\_sim}}{P_{DC\_meas}})}{(\frac{P_{AC\_sim}}{P_{AC\_meas}})} \tag{26}$$

By analyzing the ratios mentioned above a distinction between the different fault types is possible. A RP_{DC_AC} less than one would indicate that the measured power on the AC side of the inverter is reduced, suggesting that either the inverter is broken or a cable on the AC side was damaged.

On the other hand a value greater than one implies a fault on the DC side and can be specified by further analyzing R_c and R_v. If the fault occurred in the PV array it is indicated by both ratios being greater than one. The fault could be due to partial shadowing of the modules, aging or an MPPT error. The case where both ratios are less than one is an indication of an optimal behavior of both the voltage and the current which means the fault detection delivered a false alarm. If only the R_c delivers a value less than one it could be due to a string breakdown or disconnection. Meanwhile, if only the R_v is less than one then it can suggest a breakdown of modules or the disconnection of a module in a string.

To further identify the exact location of the faulty strings the calculation of the DC current ratio as given by equation (27) of each individual string is required.

$$Rc_i = \frac{I_{DC\_meas_i}}{I_{DC\_sim_i}} \tag{27}$$

Figure (28) displays a flowchart that demonstrates the fault identification and location procedure [5].

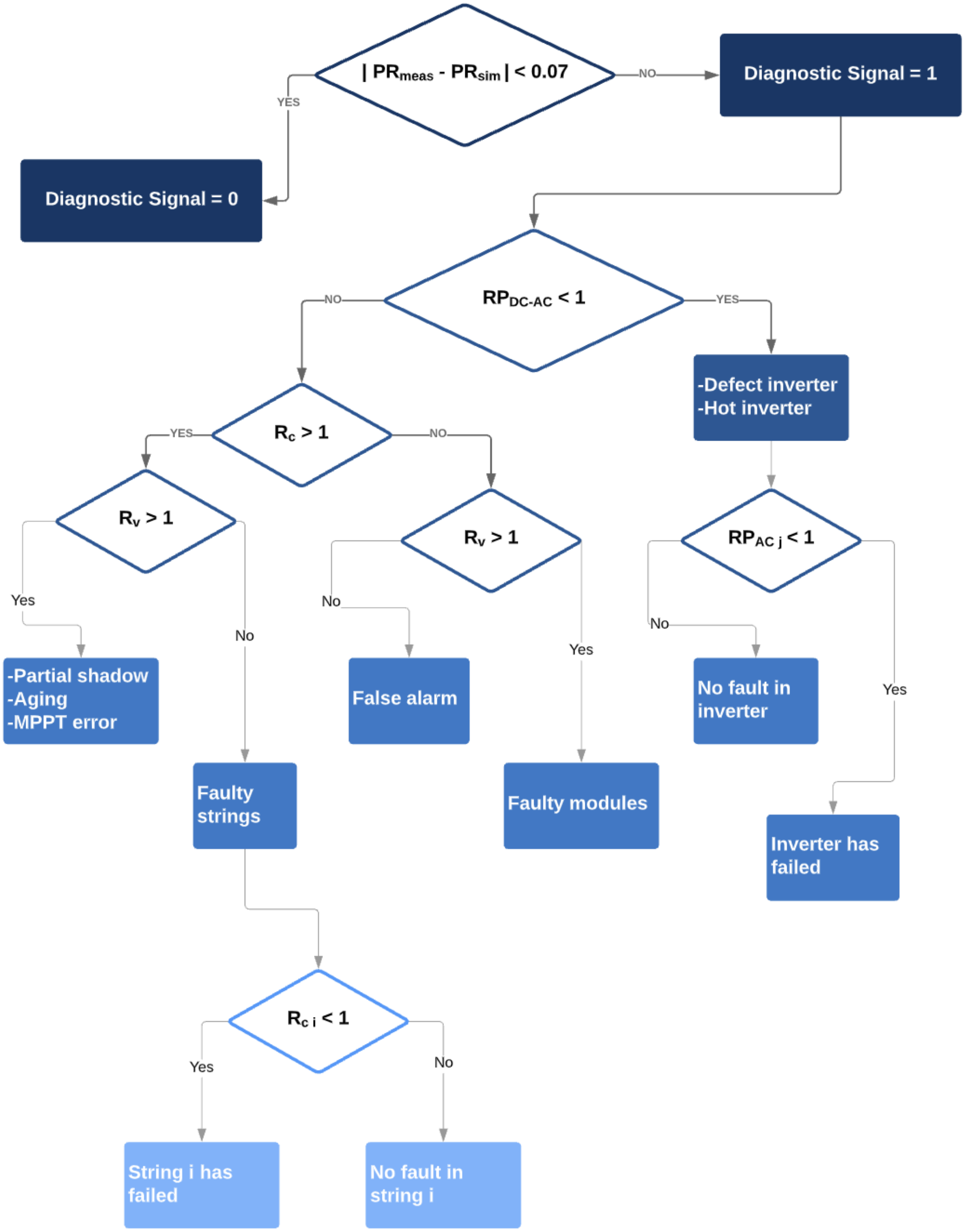


Figure 28: Flowchart of the fault detection and identification procedure

4 Findings

In this section the findings of this work are introduced. First by evaluating the model performance with the analyzed data and showing the accuracy of the estimation resulting from fitting the model with historical data.

Furthermore, the second part of this section explains how the model is utilized to characterize the modules of the plant (or any other component) to better understand the system behavior. By doing so the influence of seasonal effects over the year can be identified dependencies of the system behavior on the irradiance and module temperature can be quantified.

Finally, the fault detection procedure introduced in (3.8) is implemented and adapted to the studied PV plant. The results are then validated with daily reports provided by O&M.

4.1 Model Evaluation

In this section, the prediction performance of the model is evaluated.

4.1.1 Model Fit Robustness

In this subsection the robustness of the fit resulting from the model is introduced.

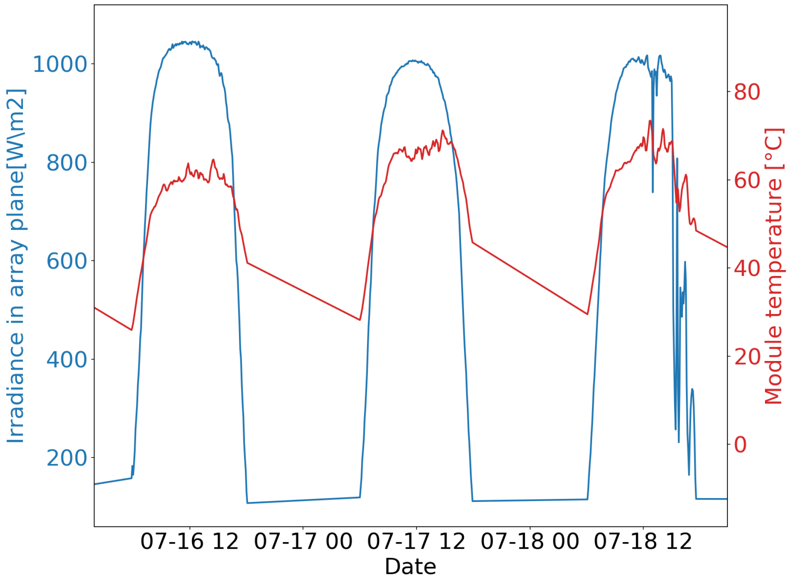


Figure 29: Irradiance in array plane and module temperature for variable weather, a clear day and a cloudy day.

Figure (29) shows three different days that represent a clear day and two cloudy days to examine the robustness of the model for predicting under differing weather conditions. Plotted are the irradiance in the array plane and the module temperature.

For this analysis the measured power on the DC side of one inverter is normalized to create the parameter PR_{DC} and this parameter is fitted with the MPM and de-normalized to show the modeled power vs the measured power. Many filters are applied to build a good model. These filters include fitting with $GTI > 100 W/m^2$ and removing data where the weather corrected PR is higher than 0.93 or lower than 0.7. Furthermore, the PR_{DC} is filtered for measurements above 1 and below 0.7. This leads to mean R^2 of 0.995 and a RMSE of 0.9%. By applying this data a fraction of 20% of the complete dataset is eliminated.

Figure (30) shows the measured vs the predicted power for the same three days shown in Figure (29). Additionally the residual is plotted to see how the model behaves under different weather conditions.

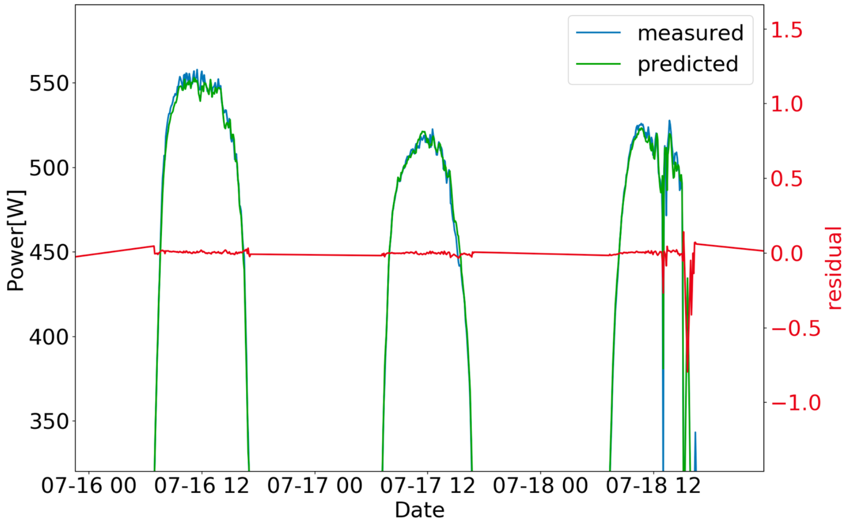


Figure 30: Measured vs modeled inverter power on DC side and residual error

The model shows very accurate behaviour with errors usually below 3%. The only discrepancies are during quickly changing irradiance periods. This is because of the large size of the array and with the irradiance sensors being very small, the clouds do not move over the modules and the sensors at the same time.

Furthermore to evaluate the model robustness it is used to fit two other normalized parameters, which are the nI_{DC} and the nV_{DC}. Figure (31) shows the measured and the fitted parameters for one month plotted over GTI and the calculated residuals. The parameter nV_{DC} shows the lowest errors and the most accurate fit. nI_{DC} is usually the most scattered and may benefit from soiling, angle of incidence, shading and spectrum corrections.

Alongside this the coefficients that are shown in table (3) are used to plot the different fits against GTI with each curve representing a different T_{Mod}. The lines that can be observe in the figure represent the visualization of the fits with increasing GTI. nI_{DC} has clearly the least dependence on the model temperature. This can also be derived from the low temperature coefficient shown in table (3).

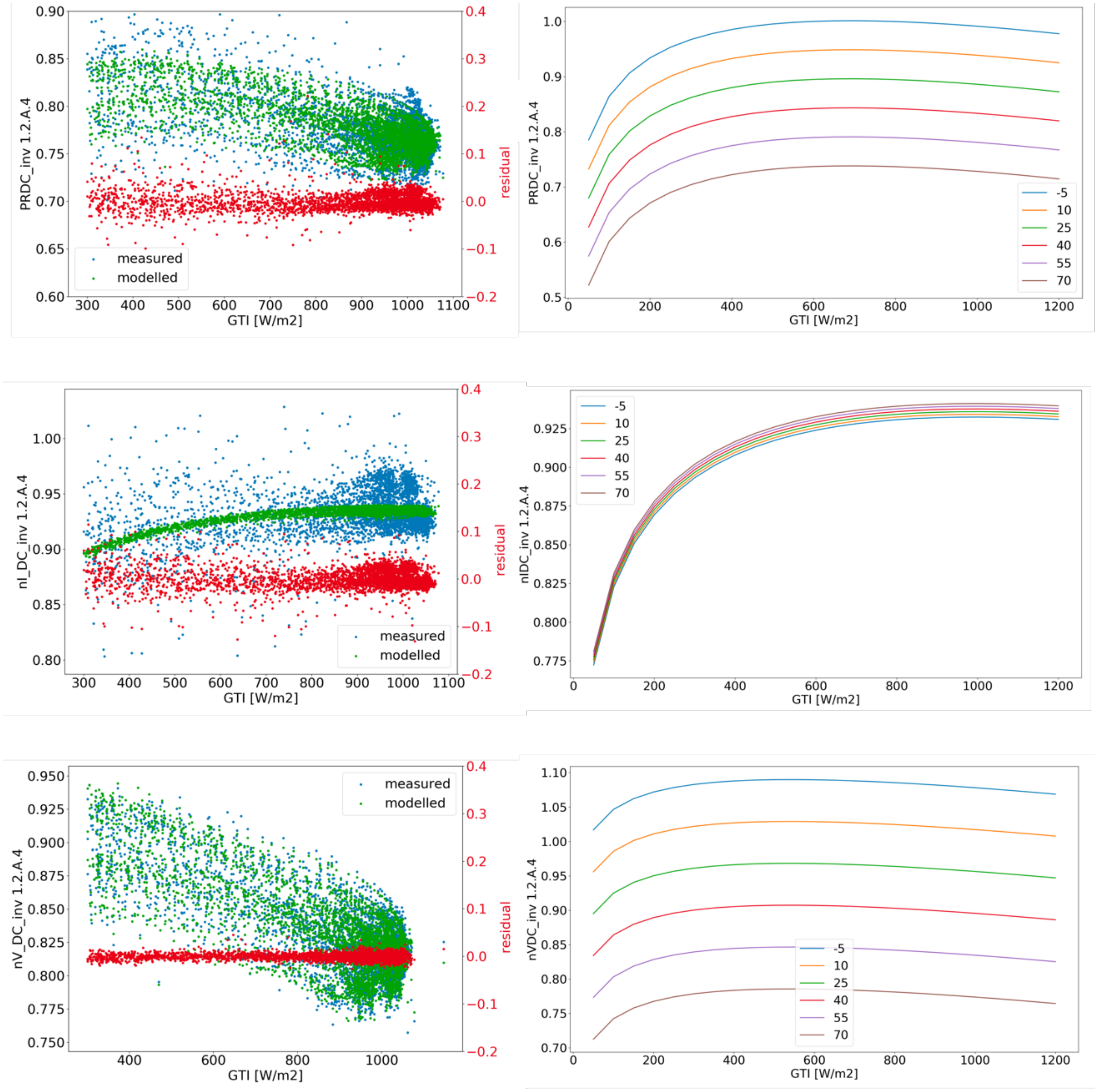


Figure 31: Power, current and voltage fits and residuals (left) Parameter fits against GTI for different T_{Mod} (right)

Table (3) gives a summary of the normalized coefficients used in Figure (31) for the three different parameters PR_{DC}, nI_{DC} and nV_{DC}. Furthermore, it also compares the $RMSE$ and the R^2. While all parameters show very good $RMSE$ and R^2 the fit for the normalized voltage results in the highest scores.

Table 3: Fitting coefficients, R^2 and RMSE for all the parameters

	PR_{DC}	nI_{DC}	nV_{DC}
C_1	107.51%	101.22%	106.02%
C_2	-0.41%	-0.02%	-0.42%
C_3	30.40%	17.41%	12.60%
C_4	-17.03%	-7.22%	-9.48%
C_5	0.11%	0.07%	-0.07%
$RMSE$	1.53%	1.81%	0.82%
R^2	99.14%	99.23%	99.77%

To evaluate the accuracy of the model in predicting future data, the model is fitted using measurements from January to May and then the inverter power on the DC side output for the month of June is calculated with the fitted model.

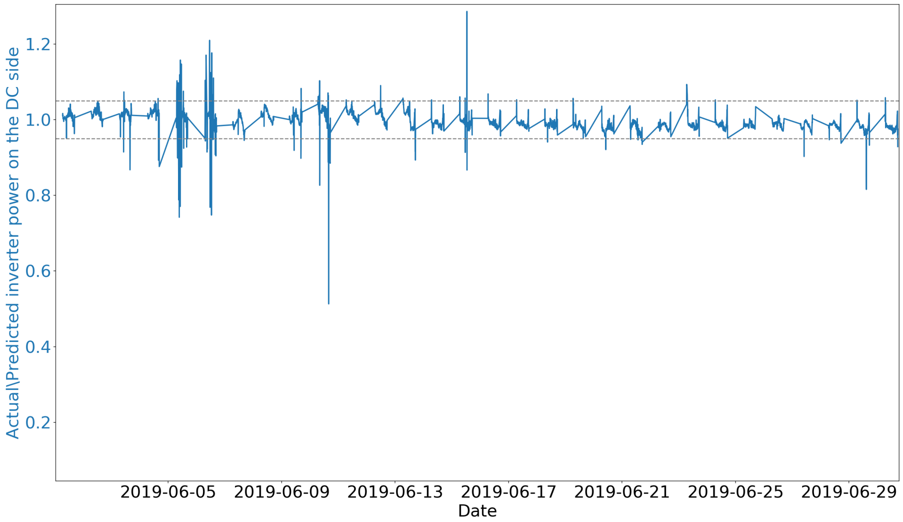


Figure 32: Actual/predicted inverter power on DC side

The actual measurements of June are then divided by the predicted data to evaluate the accuracy of the model. Figure (32) shows that the model is very accurate in predicting for future months. A threshold of 95% to 105% is met with very few outliers that are due to high weather variability. Accurate predictions are impossible for these outliers, because the accuracy of the GTI

measurements is very low and the model uses these as input to predict the output power.

4.1.2 Influence of Data Filtering and Weighting with G_i

Given that it was proven that the model fits and predicts the data well, the next step would be to examine the dependency of the model accuracy on data filtering.

First a filter is applied that removes the data points where the component being analyzed is completely out. This is done by running the power of the component through a function that detects whenever the output is below a threshold of $10W$ and filtering out all the measurements that occur at that time. By removing this data the RMSE for the PR_{DC} improves from 3.23% to 2.71%. Meanwhile for the nI_{DC} it reduces from 3.81% to 3.14% and for the nV_{DC} from 1.03% to 0.83%. This is due to the fact that measurements taken when this specific component was out do not represent the system behaviour and by that reduce the model accuracy.

The second filter applied is an irradiance condition filter. Figure (33) shows an example of how the residuals increase exponentially for the low irradiance points for the PR_{DC} in the case of not applying any filter to the irradiance before creating the fit. This again is due to the difficulty of prediction for low irradiance, because of the inaccuracy of measurements during times with high weather variability.

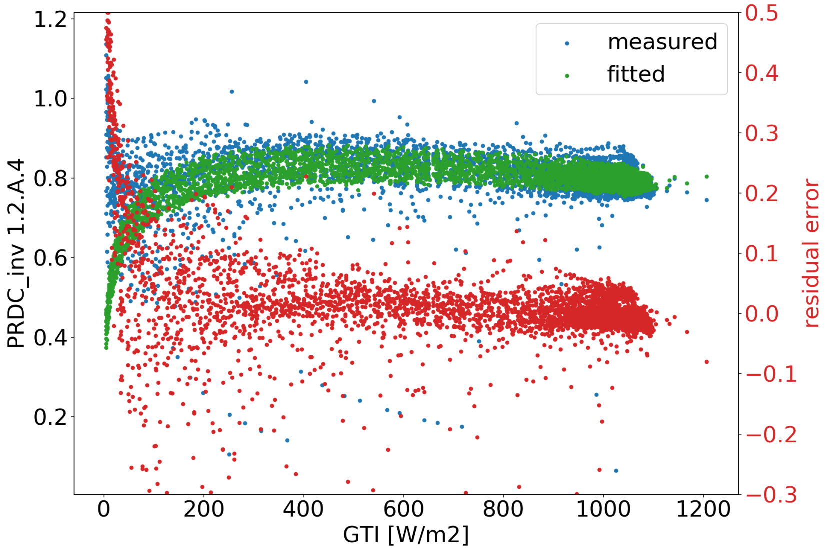


Figure 33: The influence of not applying a low irradiance condition filter on the residuals.

Therefore, it is of great importance to apply a low irradiation condition filter.

Table (4) summarizes the effects of low irradiance condition filters for all the parameters on the RMSE.

Table 4: Influence of low irradiance condition filters on the RMSE of the different parameters

	PR_{DC}	nI_{DC}	nV_{DC}
no Filter	3.23%	3.81%	1.02%
$GTI > 20W/m^2$	2.78%	3.22%	0.86%
$GTI > 50W/m^2$	2.47%	2.95%	0.84%
$GTI > 60W/m^2$	2.97%	3.39%	0.92%

For each parameter a different low irradiance condition is applied depending on the trade-off between getting rid of high residuals and not loosing much data.

The filter that leads to the greatest enhancement of the RMSE is the oultlier filter which is also applied in the course of this thesis. Since the effect of each error on RMSE is proportional to the size of the squared error, larger errors have a disproportionately large effect on RMSE. Therefor, RMSE is sensitive to outliers.

By only including data points, where $0.65 < PR_{DC} < 1$, $0.8 < nI_{DC} < 1.03$ and $0.7 < nV_{DC} < 1$ the RMSE shown in the first row of table (4) reduce to 1.58% for PR_{DC}, to 1.74% for nI_{DC} and to 0.63% for nV_{DC}. Figure (34) shows the high residuals caused by the outliers in the measurements for the nV_{DC}.

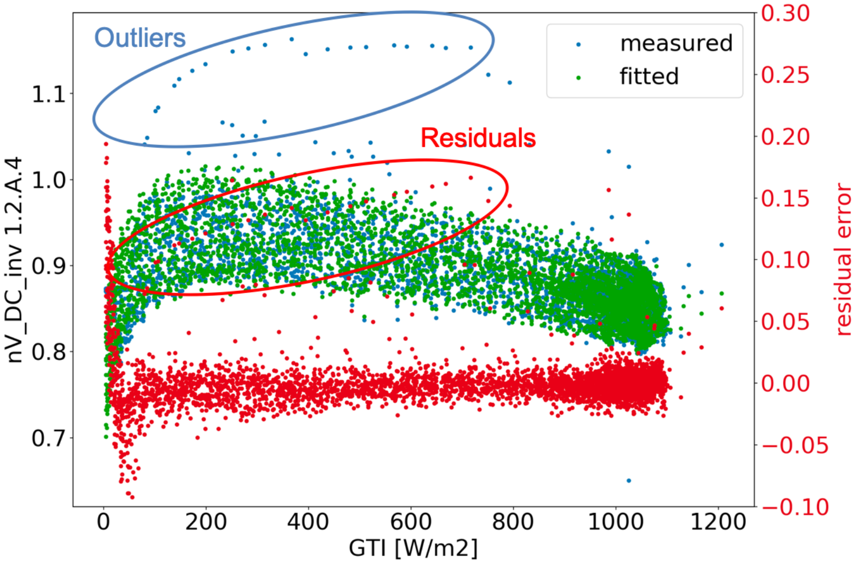


Figure 34: Very high residuals caused by outliers in the measurements of nV_{DC}.

Another filter applied is based on the weather corrected performance ratio of the whole plant, which is calculated with the power measurements at the meter level. The meter is at the connection point to the grid. The reason for adding this filter is to eliminate all the measurements where the local grid experienced an outage or when the plant had tracking problems. Data that coincides with a PR above one was removed for all parameters, since this is unrealistic and can only mean the pyranometers were not working properly and were providing false irradiance measurements due to soiling of the sensors or irregular clouds leading to uneven irradiance measured by the sensors. Furthermore, for the PR_{DC} removing measurements where the PR is below 0.7 lead to a reduction of the RMSE.

When analyzing inverters with higher capacities where inverter clipping would take place, it was important to add another filter to remove the data at the clipping point. Inverter clipping plays a decisive role in regulating the amount of power generated by the PV array. According to the capacity of transmission lines in the geographical area the inverters cut off the excess power to adjust to the PV plant output. That way some potential electricity generation will be lost, especially on sunnier days of the week [9].

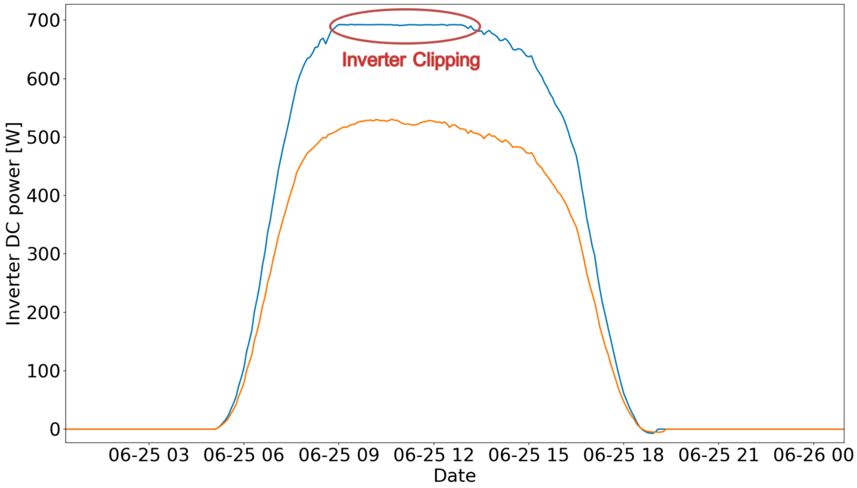


Figure 35: Inverter Clipping

Data collected at times where inverter clipping took place do not represent the actual physical behaviour of the PV plant and need to be filtered out while training the model. Figure (35) shows a typical DC power graph of an inverter that is limiting the system in comparison to another inverter with a lower capacity where no clipping takes place.

One of the main approaches for cleaning the data before fitting the model is to filter for clear days. The purpose of this is to minimize as much inconsistencies as possible to simulate the best possible physical behavior of the inverters. This is due to the behavior of the MPP-Tracking system that was introduced in

(2.1.3).

In order to identify the clear-sky days a function from the python library pvlib is used. This function implements an algorithm to detect the clear and cloudy points of a time series by analyzing the GHI. The detection is based on comparing the actual measured time series with an expected clear-sky time series as shown in Figure (36). The result of the comparison is a score between one and zero with a higher score corresponding to a clearer sky. By setting a threshold all the clear days can be specified.

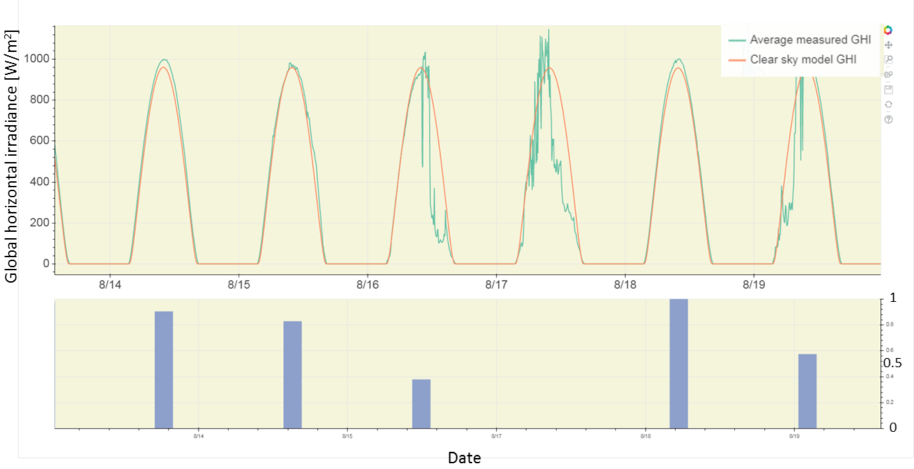


Figure 36: Clear Sky Detection Function.

As a result removing the cloudy days before fitting the model improves the scores and minimizes the root mean square error (RMSE) for the linear model as shown in table (5). Though for the non-linear model this effect of the filtering is negligible.

Table 5: Scores and RMSE improvements for linear and non-linear models by adding a Clear-Sky Days filter

	non-linear		**linear**	
	no filter	filter	no filter	filter
Score	0.9918	0.9921	0.7305	0.8254
RMSE	1.6%	1.4%	2.3%	1.6%

The reason for this is that the linear model is more sensible to fluctuations while the non-linear model is more robust. As a conclusion filtering for clear-sky days would only make sense if a linear model is implemented.

In the course of this thesis the model is weighted with the irradiance in the plane of array as shown in equation (28).

$$PR_{DC}*G_I = (C_1+C_2*(T_{Mod}-T_{STC})+C_3*log_{10}(G_I)+C_4*G_I+C_5*WS)*G_I \tag{28}$$

For instance, multiplying the PR_{DC} with G_I means that the model is fitting the PR_{DC}, while giving the values at high irradiance more weight, as the high irradiance PR_{DC} is more important to energy yield than is low irradiance PR_{DC}.

Simultaneously, multiplying the PR_{DC} with G_I results in the $nP = \frac{P_{max}}{P_{ref}}$, which can also be seen as fitting the nP rather than the PR. This statistically weights to the field performance and means the anomalous scatter at very low light levels is eliminated; similarly multiplying nI_{DC} by G_I results in something meaningful.

However nV_{DC} does not include G_I in the denominator, so multiplying it by G_I does not come up with a meaningful value, although it is useful in the mathematics.

It is common in the industry to statistically weight parameters by G_I, such as V and T_{Mod} as these are used when averaging performance over whole days. As V at night is 0 and T_{Mod} at night is irrelevant it is still useful to do this even though $V * G_I$ and $T_{Mod} * G_I$ do not mean anything themselves.

Figure (37) shows the effect weighting has on fitting the different parameters. The same data (top) is fitted once without (middle) and once with weighting (bottom). The greatest influence is seen on the nI_{DC} for low irradiance in the case of weighting the fitted curve drops as opposed to the increase visible if no weighting is applied.

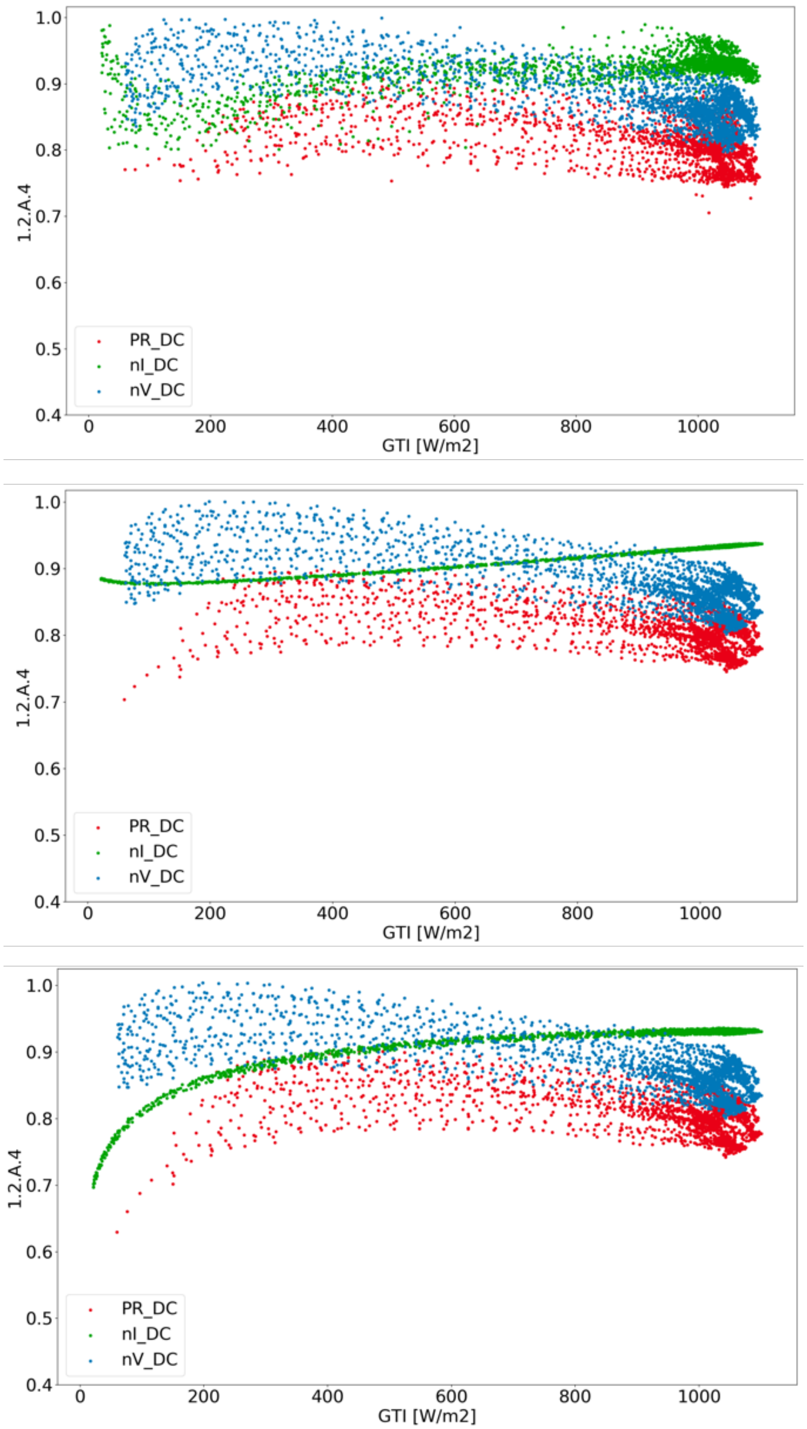


Figure 37: Data for fitting (top), Fitted curves without (middle) and with weighting (bottom).

The positive effect the weighting has on the RMSE and the R^2 of all the parameters is summarized in table (6).

Table 6: Scores and RMSE improvements for weighting the model with G_I

	not weighted	**weighted**
	PR_{DC}	
R^2	81.05%	99.49%
RMSE	1.48%	1.28%
	nI_{DC}	
R^2	43.90%	99.75%
RMSE	1.93%	1.34%
	nV_{DC}	
R^2	97.49%	99.95%
RMSE	0.65%	0.53%

4.2 Fault Detection

Once an accurate performance model is calibrated, it is valuable to examine a component's model coefficients and loss factors and conclude something about the condition or health of the analyzed component.

Furthermore, the model allows for real time validation of measurements by predicting the optimal output and comparing it to the instantaneous readings.

Since it is easier to measure the IV curves of single modules than for strings of arrays, it is often the case for multi MW_p solar plants that the measurements available for the strings either just include the P_{MP} or include also the I_{MP} and the V_{MP}.

For the course of this thesis no IV curve measurement were taken on site and consequently only the power, current and voltage at maximum power point of each component in the plant could be analyzed separately. These components include all the inverters, combiner boxes and strings that make up the PV plant.

After the conformation of the model fitting accuracy in the preceding section, the fits can be utilized for the module characterization analysis and the prediction of the optimum output behavior.

4.2.1 Loss Factor Analysis

This section presents analysis results based on the LFM/MPM Model. The model allows an evaluation of the different irradiance levels and temperature behavior of PV modules.

To have the the best representation of the modules, the measurements from one string are taken as an example.

Since the generated electricity depends on the measured output power of the array, it is calculated by multiplying its measured voltage and current for the strings. And because of the great influence the irradiance has on the output power, the fitted power of one string is evaluated first against the irradiance.

The dependency on the irradiance is shown in Figure (38). The PR_{DC} shows poor low and high irradiance behavior. It is hard to identify the root of such behavior from observing only one parameter. Therefor, by adding more parameters to the analysis the reason for this behavior can be identified. Figure (38) also shows the fits for the nI_{DC} and the nV_{DC} and it becomes clear that the poor low irradiance of the PR_{DC} is caused by the nI_{DC}, while the poor high irradiance behavior is caused by the nV_{DC}.

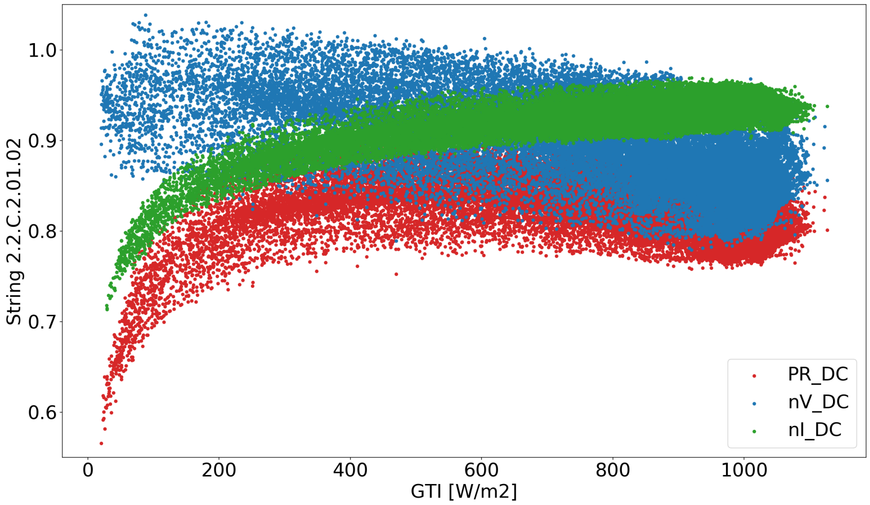


Figure 38: Fitted PR_{DC}, nI_{DC} and nV_{DC} plotted over GTI

To further understand this behavior, it is important to also analyze the effect of the module temperature on the parameters. First this is done by looking at each parameter over the irradiance as in Figure (38) with additionally adding a color scale for the module temperature as a third axis as shown in Figure (39). This shows that the module temperature is responsible for the broad scatter of the fitted curves of both the PR_{DC} and the nV_{DC}, while the nI_{DC} shows a narrow scatter, due to the low dependence of the current on the temperature.

Moreover, on the right side of Figure (39) the parameters are plotted over the module temperature with a color scale for the irradiance as a third axis. This displays a linear dependency of all the parameters on the module temperature except in the case of very low irradiance. While the increase of module temperature causes a slight increase in the current, it causes an obvious decrease for both the voltage and the power.

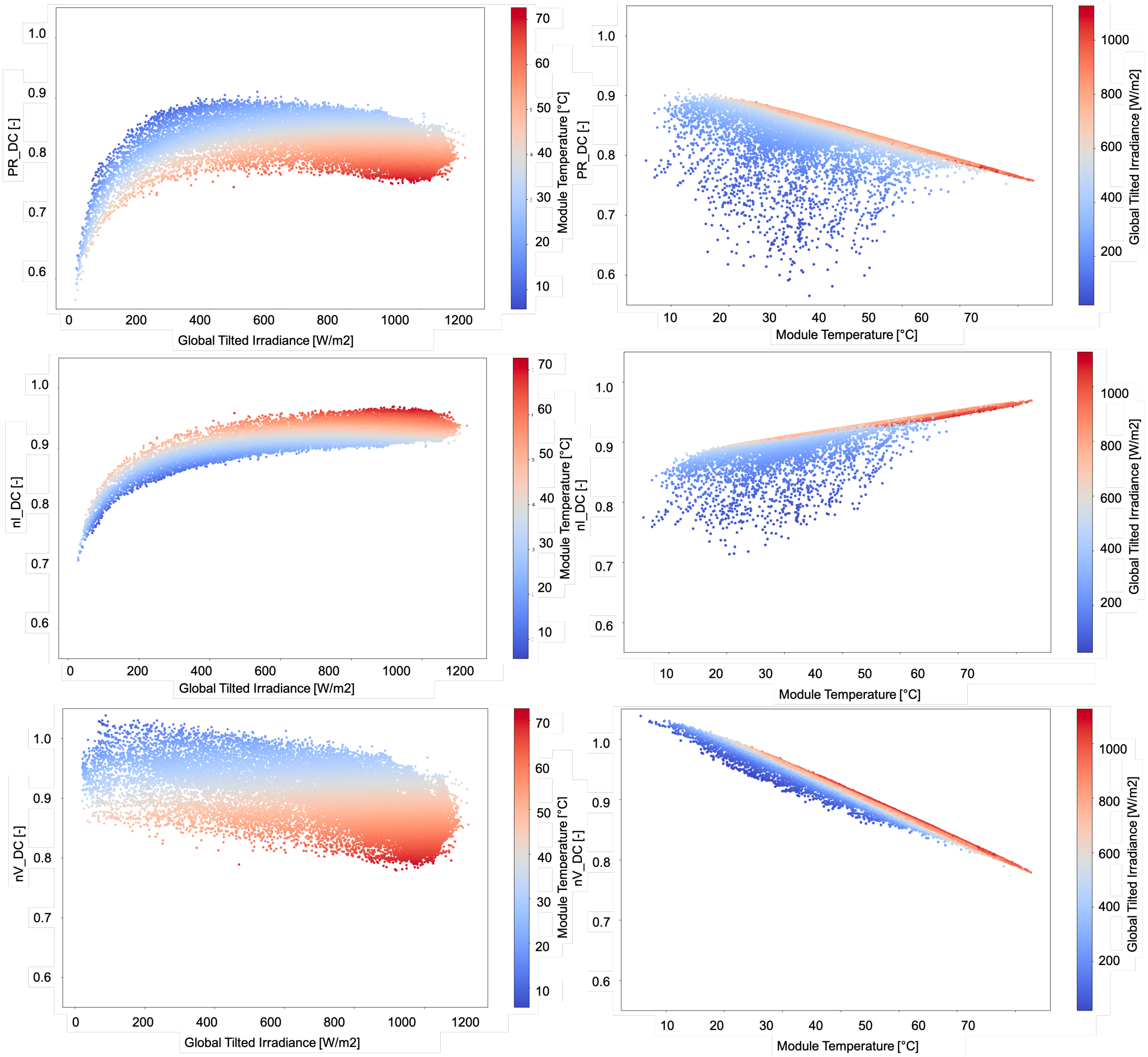


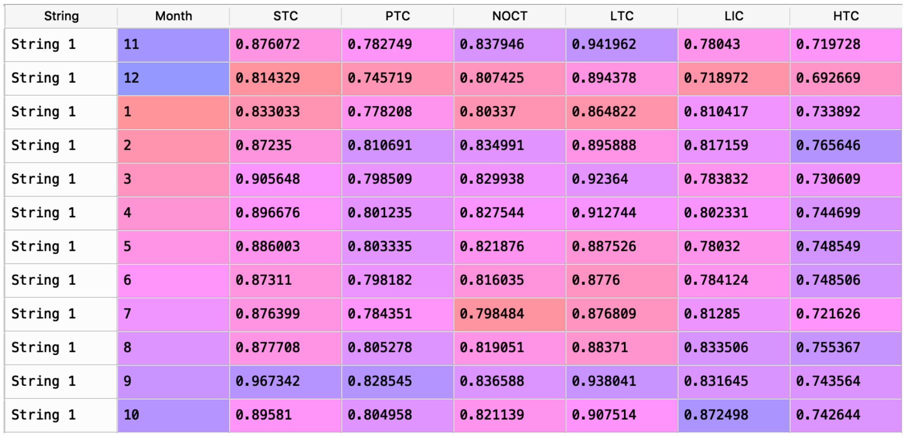
Figure 39: PR_{DC}, nI_{DC} and nV_{DC} against irradiance with a color scale based on the module temperature (left) and against module temperature with a color scale based on the irradiance (right)

An observation of the variation of the loss factors over time allows for the evaluation of the degradation of the modules and components and the monitoring of the influence of the seasonality over the year on the system behavior.

For this study monthly data is fitted and by using IEC 61853 test conditions that are defined in table (7) the output of the analyzed component can be study under varying weather conditions. STC stands for Standard Test Conditions, PTC for PVUSA, NOCT for nominal operating cell temperature, LTC for low temperature, LIC for low irradiance and HTC for high temperature.

Table 7: Definitions of IEC 61853 test conditions

	STC	**PTC**	**NOCT**	**LTC**	**LIC**	**HTC**
$G_I(kW/m^2)$	1	1	0.8	0.5	0.2	1
$T_{Amb}(°C)$	-	20	20	-	-	-
$T_{Mod}(°C)$	25	∼55	∼47	15	25	75
$WS(ms^{-1})$	0	1	1	0	0	0
$Tilt(°)$	-	-	45	-	-	-
AM	1.5	1.5	1.5	1.5	1.5	1.5



String	Month	STC	PTC	NOCT	LTC	LIC	HTC
String 1	11	0.876072	0.782749	0.837946	0.941962	0.78043	0.719728
String 1	12	0.814329	0.745719	0.807425	0.894378	0.718972	0.692669
String 1	1	0.833033	0.778208	0.80337	0.864822	0.810417	0.733892
String 1	2	0.87235	0.810691	0.834991	0.895888	0.817159	0.765646
String 1	3	0.905648	0.798509	0.829938	0.92364	0.783832	0.730609
String 1	4	0.896676	0.801235	0.827544	0.912744	0.802331	0.744699
String 1	5	0.886003	0.803335	0.821876	0.887526	0.78032	0.748549
String 1	6	0.87311	0.798182	0.816035	0.8776	0.784124	0.748506
String 1	7	0.876399	0.784351	0.798484	0.876809	0.81285	0.721626
String 1	8	0.877708	0.805278	0.819051	0.88371	0.833506	0.755367
String 1	9	0.967342	0.828545	0.836588	0.938041	0.831645	0.743564
String 1	10	0.89581	0.804958	0.821139	0.907514	0.872498	0.742644

Figure 40: PR_{DC} estimates for one string based on IEC 61853 test conditions for each month of the year

Figure (40) displays the variation of the PR_{DC} of one example string over the year with varying weather data inputs for the model. The values shown represent the optimal behavior of the studied string based on the fitted model.

To be able to detect degradation 13 months of are required at least to be able to compare the behavior under similar weather conditions. Furthermore, by collecting data over many years a degradation become easily quantifiable with the application of this model. For the course of this thesis only one full year is analyzed.

The strong variation of the performance for the different months becomes clear and shows how strongly the performance is correlated with the seasonality over the year.

4.2.2 Fault Identification and Location

After characterizing the modules, the coefficients used by the model can be utilized to predict the optimal output of any component in the plant in real-time to validate the power, current and voltage measurements instantaneously. That would facilitate the detection of faults or under-performance and help in identifying the cause.

For this step the tool introduced in (3.8) is applied to an inverter with all the connected combiner boxes and strings.

Since the number of strings connected to the inverter is 124 and this makes the impact a faulty string would have on the inverter performance negligibly small the fault detection procedure is run separately for each combiner box connected to the inverter. Figure (41) shows the predicted versus the estimated DC power on the inverter level for three days in October and how almost identical the curves are.

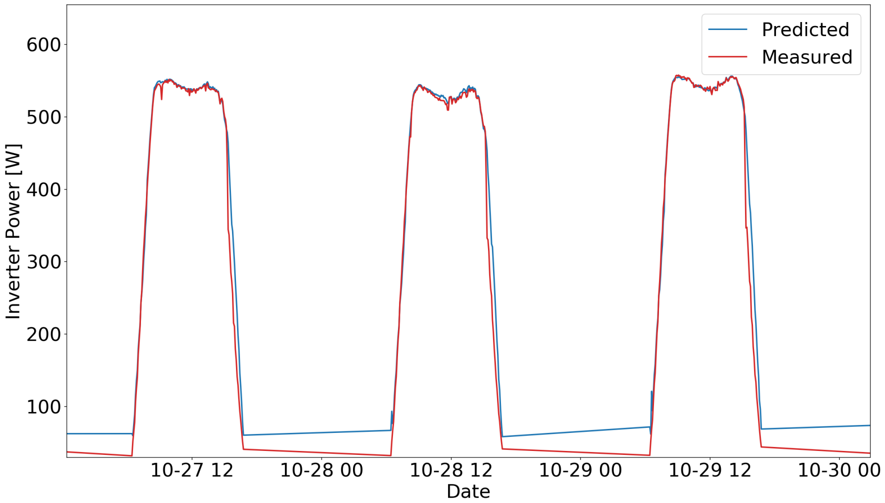


Figure 41: Inverter DC power measured and estimated results for three days in October 2019

Therefore, the reference PR_{DC} deviation will be based on one single combiner box proceeding with the analysis. Due to the high accuracy delivered by the implemented model the threshold of 0.07 for detecting a deviation between the measured and estimated PR_{DC} of the combiner boxes needed to be lowered to 0.02.

Figure (42) displays the deviations for three days in October with two days exceeding the 0.02 threshold. Once the threshold is exceeded a diagnostic signal is set to one, otherwise to zero.

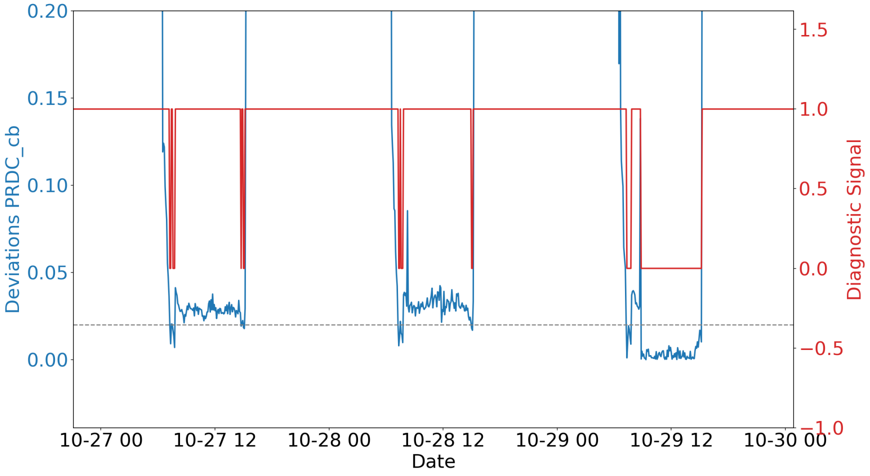


Figure 42: Combiner box PR_{DC} deviations between estimated and measured for two faulty days and one one day that is only partly faulty in October 2019

The analysis is continued with the times corresponding to a diagnostic signal equal to one.

After running fitting the PR_{DC}, nI_{DC} and nV_{DC} for each component the ratios are calculated and added to a DataFrame. Figure (43) shows the evolution of the RP_{DC_AC} of the inverter and the R_c and R_v of the combiner box.

The next step is to identify the type of fault based on the flowchart as shown in Figure (28). By running the conditions for each row of the DataFrame containing all the calculated ratios the type of fault occurring in each 5 min measurements can be classified separately.

Finally to identify the exact location of the string where a fault occurred, the DC current ratio is calculated for each string and added as an extra column to the data frame. Since for the analysed two strings are connected to one measurement channel, a DC current ration of 0.5 indicates one of the two string broke down or is disconnected. This is shown in Figure (44).

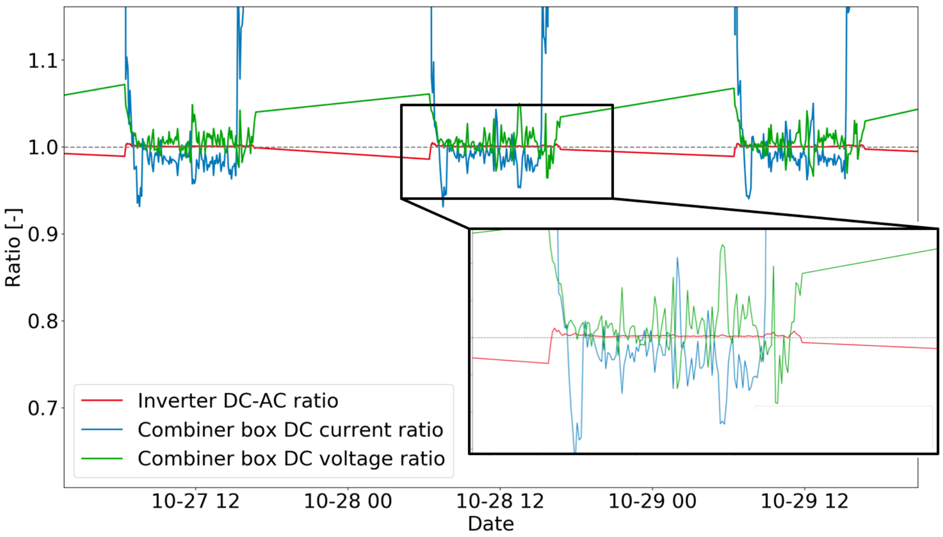


Figure 43: Combiner box R_c and R_v and inverter RP_{DC_AC} evolution for three days in October

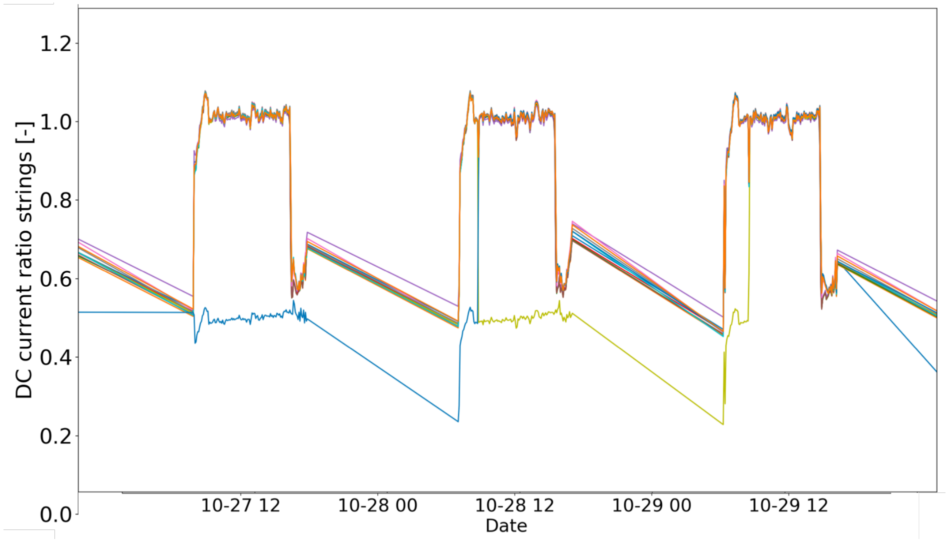


Figure 44: String DC current ratio for three days in October

4.2.3 Validation of Fault Detection Procedure

This section introduces the verification of the performance of the applied tool.

By comparing the results from the fault identification procedure with the daily reports provided by the operation and maintenance team a verification is possible. The daily reports contain a documentation of each string fuse that was replaced at a specific day. After locating the faulty string, the documentation of the fuse replacement can be an indicator if the resulting faulty string was in fact broken or not.

Only three faulty strings were detected during the month of October for the studied inverter, which was also confirmed by the O&M daily reports. The identification procedure is able to detect the faulty strings a day ahead of the documented fuse replacement, which could lead to a faster intervention and avoidance of unnecessary downtime.

5 Conclusion and Further Research

This section will provide a summary of the main findings and discuss future research which should be considered.

5.1 Conclusion

In this thesis, a model was built with the main objective of enabling an early and accurate detection of faults in a PV plant. The model was built based on a combination of two models existing in the literature, it was implemented in this thesis and it was evaluated based on real data from a recently built large-scale power plant in Egypt. Unlike the common practice of detecting faults based on comparing the actual measurements with estimations by theoretical models, such as the one-diode-model, the model implemented in this thesis optimizes accuracy of fault detection by using historical data for fitting, which leads to a more accurate representation of the system behavior.

The model implemented in this thesis has the following outcomes: The first outcome of the model is to accurately characterise the components (e.g. inverters and strings) of the PV plant. The second outcome is to accurately predict the optimal output of the components of the plant. A comparison between the optimal outputs of the model and the measured value allows for the detection of deviations, if they exist. Further analysis by the model of the deviations can identify and locate faults that may exist in a plant.

The model is applied to one inverter with all the connected combiner boxes and strings and fits the performance measurements of each component based on historical data after filtering for outliers, very low irradiance and weighting with the irradiance in the plane of array. The discussed procedure confirms optimum output behavior, or else identifies faults by implementing a fault detection procedure, which allows for quick real-time fault detection, identification and location and thereby can lead to minimizing downtime. Moreover, seasonal effects of the PR_{DC} are identified by evaluating its variation over time. The automatic real time detection procedure was able to detect faulty strings very quickly.

Daily reports from the O&M department in Egypt which specified which strings broke and had to be replaced were used to validate the results of the procedure, which revealed that only three faulty strings were detected during the month of October for the studied inverter.

The identification procedure was able to detect the faulty strings a day ahead of the documented fuse replacement, which could lead to a faster intervention and avoidance of unnecessary downtime.

It can therefore be concluded that the model built and implemented in this thesis has achieved its objective of characterising the components of a photovoltaic power plant and thereby enabling an early and accurate detection of faults and consequently reducing the cost of maintenance and minimizing downtime of the components of the power plant.

With further investigation and research the fault detection tools provided by this model can enable PV operators to better understand the PV power plant, detect faults in an accurate and timely manner, and allow them to address and solve the malfunctions and increase the reliability of the installations and ensure the guaranteed lifetime output.

5.2 Further Research

For the course of this thesis no IV curve measurement were taken on site and consequently only the power, current and voltage at maximum power point of each component in the plant could be analyzed separately.

By measuring high quality IV curves more parameters can be analysed which in return would provide more detailed classification of the malfunctions or under performance of the system's components.

One of the main advantages of the LFM/MPM is that the resulting coefficients from the fits represent physical coefficients. In further research their values could be compared to the data sheet values of the modules, which would allow for a study of the variations of the coefficients over the year.

To be able to detect degradation at least 13 months of data are required to compare the performance evolution under similar weather conditions, which was not present at the studied plant. Furthermore, by analyzing data from several years the degradation can also be quantified.

Finally, by correcting the performance measurements for soiling and temperature a more accurate optimal behavior of the plant can be estimated.

References

[1] Kais Abdulmawjood, Shady S. Refaat, and Walid G. Morsi. Detection and prediction of faults in photovoltaic arrays: A review. In *Proceedings - 2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering, CPE-POWERENG 2018*, pages 1–8. Institute of Electrical and Electronics Engineers Inc., jun 2018.

[2] H. G. Beyer, G. Heilscher, and S. Bofinger. A robust model for the MPP performance of different types of PV-modules applied for the performance check of grid connected systems. Technical report, 2004.

[3] J. Brownlee. Master Machine Learning Algorithms Discover How They Work and Implement Them From Scratch i Master Machine Learning Algorithms. Technical report, 2016.

[4] Jason Brownlee. Machine Learning Mastery With Python Understand Your Data, Create Accurate Models and Work Projects End-To-End. Technical report, 2016.

[5] W. Chine, A. Mellit, A. Massi Pavan, and S. A. Kalogirou. Fault detection method for grid-connected photovoltaic plants. *Renewable Energy*, 66:99–110, jun 2014.

[6] G. Schulze and M. Töpfer. Operation and maintenance manual - ib vogt. Technical report, 2017.

[7] Elyes Garoudja, Fouzi Harrou, Ying Sun, Kamel Kara, Aissa Chouder, and Santiago Silvestre. Statistical fault detection in photovoltaic systems. *Solar Energy*, 150:485–499, 2017.

[8] Nuri Gokmen, Engin Karatepe, Santiago Silvestre, Berk Celik, and Pablo Ortega. An efficient fault diagnosis method for PV systems based on operating voltage-window. *Energy Conversion and Management*, 73:350–360, 2013.

[9] M. Green, E. Brill, B. Jones, and J. Dore. *Improving Efficiency of PV Systems Using Statistical Performance Monitoring.* 2017.

[10] D L King, W E Boyson, and J A Kratochvill. Photovoltaic Array Performance Model. Technical report, Andia National Laboratories, 2004.

[11] David L. King. Photovoltaic module and array performance characterization methods for all system operating conditions. pages 347–368. AIP Publishing, may 2008.

[12] Andreas Livera, Alexander Phinikarides, George Makrides, Juergen Sutterlueti, and George E Georghiou. Advanced Failure Detection Algorithms and Performance Decision Classification for Grid-Connected PV Systems. Technical report, 2017.

[13] W. Marańda and M. Piotrowicz. Efficiency of maximum power point tracking in photovoltaic system under variable solar irradiance. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 62(4):713–721, 2014.

[14] M. R. Osborne. Nonlinear least squares — the Levenberg algorithm revisited. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 19(3):343–357, jun 1976.

[15] Radu Platon, Jacques Martel, Norris Woodruff, and Tak Y. Chau. Online Fault Detection in PV Systems. *IEEE Transactions on Sustainable Energy*, 6(4):1200–1207, oct 2015.

[16] S. Ransome and J. Sutterlueti. How to Choose the best Empirical Model for Optimum Energy Yield Predictions. In *2017 IEEE 44th Photovoltaic Specialist Conference, PVSC 2017*, pages 1–3. Institute of Electrical and Electronics Engineers Inc., 2017.

[17] Steve Ransome and Juergen Sutterlueti. Optimised fitting of indoor (e . g . IEC 61853 matrix) and outdoor PV measurements for diagnostics and energy yield predictions. Technical report, 2017.

[18] Steve Ransome and Juergen Sutterlueti. Optimum Use of the Loss Factor Model (LFM) for Improved PV Performance Modelling. Technical report, 2017.

[19] Steve Ransome and Juergen Sutterlueti. Adaptable PV Performance Modelling for Industrial Needs. Technical report, 2018.

[20] Stefan Sellner, Jürgen Sutterlüti, Ludwig Schreier, and Steve Ransome. Advanced PV module performance characterization and validation using the novel Loss Factors Model. In *Conference Record of the IEEE Photovoltaic Specialists Conference*, pages 2938–2943, 2012.

[21] Joshua S Stein, Juergen Sutterlueti, Steve Ransome, Clifford W. Hansen, and Bruce H. King. Outdoor PV Performance Evaluation of Three Different Models: Single-Diode, SAPM and Loss Factor Model. 2013.

[22] Ransome S Hansen C W King B H Stein J Suttrelueti J. Outdoor Performance Evaluation of Three Different Models:single-diode, SAPM and Loss Factor Model. SAND Report 2013-7913C. Technical report, 2013.

[23] Cristina Ventura and Giuseppe Marco Tina. Utility scale photovoltaic plant indices and models for on-line monitoring and fault detection purposes. *Electric Power Systems Research*, 136:43–56, jul 2016.

[24] A. Woyte and M. Richter. Monitoring of PV systems: good practices and systematic analysis. 2013.

[25] Ye Zhao, Brad Lehman, Roy Ball, Jerry Mosesian, and Jean Francois De Palma. Outlier detection rules for fault detection in solar photovoltaic arrays. In *Conference Proceedings - IEEE Applied Power Electronics Conference and Exposition - APEC*, pages 2913–2920, 2013.

[26] Ye Zhao, Ling Yang, Brad Lehman, Jean François De Palma, Jerry Mosesian, and Robert Lyons. Decision tree-based fault detection and classification in solar photovoltaic arrays. In *Conference Proceedings - IEEE Applied Power Electronics Conference and Exposition - APEC*, pages 93–99, 2012.