Norwegian University
of Life Sciences

**Master's Thesis 2019 30 ECTS**
Faculty of Science and Technology

# A data-driven approach for power loss detection in utility-scale solar power plants

Meron Haile Tesfazion
MSc. Environmental Physics and Renewable Energy

## Preface

John Naisbitt once said; *We are drowning in information but starved for knowledge*. In this thesis, I have been working with five years of data from a utility-scale solar power plant and I hope that I have been able to create actionable knowledge for the power plant operators. I am grateful and humbled to be a part of this research project where I have been given the trust to tackle the problem as I redeemed most optimal. At the time of writing, there has not been published much work within the topic. Therefore, I hope that the work presented in this thesis will sparkle the engagement within the field and for others to build upon this research in the near future.

The work presented would not be possible without several people whose help has been highly valued. First, I would like to express my profound gratitude to my advisor Heidi Samuelsen Nygård, for guidance and help with writing this thesis. I would also like to thank co-advisor Andreas Størdal, for providing me with the data from the power plant, advice, discussions and allowing me to be a part of the ongoing project.

Furthermore, I would like to give my sincere gratitude to Åsmund Skomedal for his help with the management of data from the power plant, interesting discussions and advise. In addition, I would like to thank Oliver Tomic and Kristian Liland for their help within the concept of data processing, uncertainties and machine learning.

Finally, I would like to thank my family for their support and help throughout the years. I would also like to thank my friends for their help and discussions in the past couple of months.

Oslo, 14.05.2019

Meron Haile Tesfazion

**Abstract**

With increased industrial development centred around photovoltaic (PV) production technologies, volume growth and component related cost reduction, operation and maintenance (O&M) of PV systems emerge as one of the most important R&D topics worldwide.

Therefore, the purpose of this thesis is to detect power loss in utility-scale solar power plants and provide power plant operators with information about abnormal system behaviour. Abnormal behaviour is based upon the ratio between measured system behaviour and estimated system behaviour for power current and voltage. The estimated values are calculated by both physical models for PV systems and machine learning models.

Based on the evaluation of the models, non-linear machine learning models have the best predictive ability with solar irradiance and solar cell temperature as explanatory variables. The predictive ability is further increased with the introduction of new variables such as inverter temperature, cosine and sine transformation of the day in the year, the hour of the day and estimations of the sky conditions. As a result, the best performing model has a mean model uncertainty between the average and maximum measurement uncertainty for power, voltage and current.

The estimates from these models are used to detect power loss events on the inverter. The result is that, based upon the chosen threshold values, the models are able to detect possible soiling and other events when multiple strings reduce their performance. Soling tends to accumulate at a higher degree in the middle of the site from east to west. However, the models are not able to detect scenarios where only a couple of strings are disconnected or when the performance of the string slightly decreases trough time. Furthermore, most of the events detected are due to a lower measured current than estimated. Therefore, it is believed that the inverter favours lowering the current of the strings where a possible fault is present instead of lowering the voltage on the array.

The high amount of low current events enables monitoring of each string-pair, where the ratio between the measured and estimated current is evaluated. Based on this evaluation it is possible to detect strings which have a decreasing performance. These strings were then evaluated against infrared images where the top 3 thermal defects could be found based upon the evaluation of the ratio between the measured and estimated current for a chosen inverter.

**Sammendrag**

Gjennom en økt industriell utvikling sentrert rundt solcelleteknologi, volumøkning og komponent relaterte kostnader, har drift og vedlikehold av solcellesystemer vokst frem som et av de viktigste områdene for forskning og utvikling på verdensbasis.

Formålet med denne oppgaven er bruke data fra ett storskala solcelleanlegg for å detektere effekttap på anlegget og å tilby operatørene av anlegget varsler om system oppførsel utenom det vanlige. For å angi hva som er oppførsel utenom det vanlige beregnes forholdet mellom målte verdier fra system og estimerte verdier for hvordan system burde oppføre seg. De estimerte verdiene er beregnet basert på fysiske modeller og maskinlæringsmodeller.

Etter en evaluering av modellene viser det seg at ikke-lineære maskinlæringsmodeller har høyest evne til å estimere de målte verdiene for effekt, strøm og spenning til flere invertere på anlegget. Modellene baserer seg på målt solinnstråling og en estimert temperatur på solcellene. Videre blir det vist at maskinlæringsmodellene gjør det bedre når flere variabler er inkludert, som inverter temperatur, cosinus og sinus transformasjoner av dagen på året, timen på døgnet og andre variabler som beskriver sky-forholdene. Resultatet er at den beste maskinlæringsmodellen har en gjennomsnittlig usikkerhet i intervallet mellom gjennomsnittlig og maksimal måleusikkerhet for strøm, spenning og effekt.

Fysiske modeller og den beste maskinlæringsmodellen blir så brukt for å detektere hendelser som gir effekttap på inverteren. Resultatet, basert på de beregnede terskelverdiene, er at modellene virker å kunne detektere tilsmussing av solcellemodulene og hendelser hvor flere strenger underpresterer i forhold til forventet produksjon. Fra resultatene ser det ut til at tilsmussing av solcellemodulene er høyest midt i anlegget fra øst til vest. Modellene klarer derimot ikke å detektere hendelser hvor et par stringer er frakoblet eller hvor ytelsen til ett eller flere streng-par synker gradvis. Majoriteten av de detekterte hendelser er tilfeller hvor den målte strømmen er lavere enn den estimerte strømmen. Det blir derfor naturlig å tro at inverteren heller vil redusere strømmen for en streng som opplever en feil, enn å redusere spenningen på alle tilkoblede strenger.

Av den grunn blir det mulig å monitorere hvert streng-par på anlegget for å evaluere forholdet mellom målt og estimert strøm. Basert på denne evalueringen er det mulig å detektere streng-par som opplever en nedgang i ytelsen. strengene med en nedsatt ytelse ble vurdert opp mot infrarøde bilder. Resultatet var at de tre streng-parene med høyest termisk defekt også kunne detekteres ved å evaluere forholdet mellom målt og estimert strøm, ved hjelp av metodene presentert i oppgaven.

# Nomenclature

**Symbols**

| | | |
|---|---|---|
| B | Number of decision trees in random forest regressor | - |
| D | Distance matrix | - |
| $E$ | Energy | $J$ |
| $G$ | Irradiance | $W/m^2$ |
| $I$ | Current | $A$ |
| $R$ | Resistance | $\Omega$ |
| S | Apparent Power | $VA$ |
| $T_{cell}$ | Solar cell temperature | ℃ |
| $T_{inv}$ | Inverter temperature | ℃ |
| $T_{ref}$ | Reference temperature | ℃ |
| V | Voltage | $V$ |
| $\nu$ | Frequency | $Hz$ |
| $Wp$ | Watt peak | $W$ |
| $\beta$ | Material property | - |
| $\beta_m$ | Module tilt angle | ° |
| $\beta_s$ | Scaling factor for solar irradiance | - |
| $\Delta G$ | Relative difference in irradiance | - |
| $\delta I$ | Uncertainty in current | $A$ |
| $\delta P$ | Uncertainty in power | $W$ |
| $\delta V$ | Uncertainty in voltage | $V$ |
| $\eta$ | Efficiency | - |
| $\theta_z$ | Solar zenith angle | ° |
| $\theta$ | Angle of incidence | ° |
| $\lambda$ | Penalty term | - |
| $\sigma_{max/min}$ | Maxmin | - |
| $\gamma_I$ | Short circuit temperature coefficient | - |
| $\gamma_P$ | Peak power temperature coefficient | - |
| $\gamma_s$ | Solar azimuth angle | ° |

| | | |
|---|---|---|
| $\gamma_m$ | Module orientation angle | ° |

## Abbreviations

| | |
|---|---|
| AC | Alternate current |
| AM | Air mass |
| $AOI$ | Angle of incidence |
| C | Current channel |
| CR | Current Ratio |
| DC | Direct current |
| FE | Feature engineering |
| GHI | Global horizontal irradiance |
| IEA | International Energy Agency |
| inv | Inverter |
| KNN | K-Nearest neighbours |
| LCOE | Levelised cost of energy |
| PV | Photovoltaic |
| MAE | Mean absolute error |
| MP | Maximum power |
| MPP | Maximum power point |
| MPPT | Maximum power point tracking |
| OLS | Ordinary least squares |
| OM | Operation and maintenance |
| POA | Plane of array |
| PV | Photovoltaic |
| PPC | Power plant controller |
| RCR | Relative current ratio |
| RFR | Random forest regressor |
| SFS | Sequential forward selector |

| | |
|---|---|
| SM | String monitor |
| STC | Standard test conditions |
| tresh | Threshold |
| WS | Weather station |

**Subscripts**

| | |
|---|---|
| cell | Solar cell |
| $d1$ | Diode 1 |
| $d2$ | Diode 2 |
| $f$ | Final |
| $g$ | Band Gap |
| $i$ | Initial |
| $in$ | Input |
| inv | Inverter |
| $m$ | Module |
| *meas* | Measured |
| $MPP$ | Maximum power point |
| $OC$ | Open circuit |
| P | Shunt |
| $ph$ | Photon |
| *pred* | Predicted |
| $ref$ | Standard test conditions |
| SM | String monitor |
| S | Series |
| $SC$ | Short circuit current |

**Constants**

| | | |
|---|---|---|
| h | Planck's constant | $6.626\,069\ \times 10^{-34}\ Js$ |
| $k_B$ | Boltzmann's constant | $1.380\,649\ \times 10^{-23}\ J/K$ |
| q | Elementary charge | $1.602\ \times 10^{-19}\ C$ |

# Table of content

# 1 Introduction

## 1.1 Motivation

As of July 13th, 2018, 195 countries have signed the Paris agreement (Climate Analytics, 2018). The agreement aims to strengthen the global response to the threat of climate change. Furthermore, the parties of the agreement aim to hold the increase in global average temperature to well below 2℃ above pre-industrial levels. However, the parties pursue to limit the temperature increase to 1.5℃ due to the significant reduction in risk and impact of climate change (UNFCCC, 2015). The goals can be made by reduction in greenhouse gas emission whereas the energy industry stands for approximately two-thirds of the total greenhouse emissions (IEA, 2017). According to the International Energy Agency (IEA) the global energy-related $CO_2$ emission grew to a historic high of 32.5 Gigatons in 2017, up 1.4% from the previous year (IEA, 2017). Although, the subsequent year the record was broken yet again as the global energy-related $CO_2$ emission rose another 1.7% to 33.1 Gigatons in 2018 (IEA, 2018).

According to IEA, solar energy is one of the most promising technologies to reduce emissions from the energy sector. Whereas, the clear majority of all Photovoltaic (PV) plants are relatively new. In fact, over half of the installed capacity in 2017 was less than three years old (Solar Power Europe, 2018). The increasing scale of solar energy makes it the fastest-growing power generation source (Solar Power Europe, 2018). As the PV penetration rate increases, investors in PV plants and the broader energy sector strive for accurate energy yield and returns on investment. In addition, aggressive targets for lower levelised costs of electricity (LCOE) are set (IPN, 2018).

With record low cost of PV installations, the operation and maintenance (O&M) related cost becomes increasingly important when determining LCOE and the competitiveness of each company (IPN, 2018). Until now, industrial developments and R&D has been centred around, production technologies, volume growth and component-related cost reduction. Thus, PV O&M emerges as one of the most important R&D topics worldwide (Brehaut, 2016) (Lumby, 2015) (IPN, 2018).

Today the O&M situation for utility-scale PV is based on two main groups of activities; scheduled, periodic maintenance and unscheduled corrective maintenance. Unlike the PV industry, many other industries have moved towards data-driven predictive O&M models, with multiple benefits such as (IPN, 2018):

- Data-driven O&M activities which removes unnecessary periodic O&M activities.
- Avoidance of potentially hazardous situations resulting from damaged or broken equipment.
- Anticipation of failures before they occur, thereby reducing downtime and repair cost.

## 1.2 Objective

Therefore, this thesis aims to use a data-driven approach for power loss detection on a utility-scale power plant operated in Sub-Saharan Africa. Power losses in the power plant can be due to component failure or component performance which gradually falls outside of products specification. Estimates of the power loss will be evaluated by comparison of the expected power and the measured power. The estimates of the expected power will be made by leveraging both machine learning and physical models to mimic the behaviour of the PV system on the inverter level. Subsequently, these representations are used to evaluate measured values against expected values.

In summary, the objective of this thesis is to detect power loss events on utility-scale solar PV systems. For this purpose, the following tasks will be performed:

- Generate models of the estimated DC power, voltage and current with machine learning and physical models.
- Evaluate and compare the performance of the different models.
- Detect various anomalies in the data.
- Create informative features based upon the sky conditions over the power plant.
- Detect and evaluate possible drift in the irradiance measurements.
- Evaluate the estimated power, current and voltage with the measured values to detect possible power loss factors on inverter level. Thereupon, compare the results with the performance of the strings connected to the inverter.

## 1.3 Code

The code in this thesis is written in Python and leverages common machine learning tools such as Scikit-learn. "*Scikit-learn* is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems" (Pedregosa, et al., 2019). In addition, the python module Mlxtend has been utilised for Sequential forward selection made by Sebastian Raschka. Additionally, the python module PYOD has been utilised for KNN outlier detection. Furthermore, to model solar position, irradiance, detection of clear sky conditions and more, the python module *PV-lib* was used (Holmgren, et al., 2018). "*PV-lib* provides a set of function and classes for simulating the performance of photovoltaic energy systems" (Holmgren, et al., 2018).

# 2 Theory

The theory presented in chapter 2.1 and 2.2 is mainly based on Solar Energy - The physics and engineering of photovoltaic conversion technologies and systems (Arno Smets, 2016), Renewable Energy resources (Wier, 2006), Physics of solar energy (Chen, 2011) and PVeducation.org. Additional sources are specified where they occur.

## 2.1 Solar irradiance

Each second approximately $1.73 \cdot 10^{17} W$ reach Earth as radiation from the Sun. The received power is equal to over nine thousand times the average power demand for 2017 (International Energy Agency, 2017). The power released from the sun is released as electromagnetic radiation at a mean distance from Earth of $149.6 \cdot 10^{6} km$. The *solar constant* is the total irradiance at the mean distance between the Earth and the Sun, perpendicular to the direction of the Sun at the edge of Earth's atmosphere. The value of the solar constant is approximately $1361 W/m^2$. The energy from the sun is generated by nuclear fusion. Nuclear fusion can take place due to the temperature-pressure conditions at the centre of the sun. As a result, each second approximately 4 million tons of mass is converted into $3.8 \cdot 10^{26}\ J$. The process of nuclear fusion heats the surface of the Sun, called the photosphere, to a temperature of about $6000 K$. Thus, releasing energy in the form of electromagnetic radiation due to thermal radiation. Although electromagnetic radiation is the major contributor of radiation by the Sun, accounting for 98%, neutrinos carry the remaining energy radiated by the Sun. The energy emitted by the solar surface is radiated as a sphere, thus the absorbed energy by an object decreases as the distance between the Sun and the object increases, which is illustrated in Figure 2.1. Another factor affecting the absorbed radiation intensity at Earth's surface is the distance which the radiation must travel through the atmosphere.

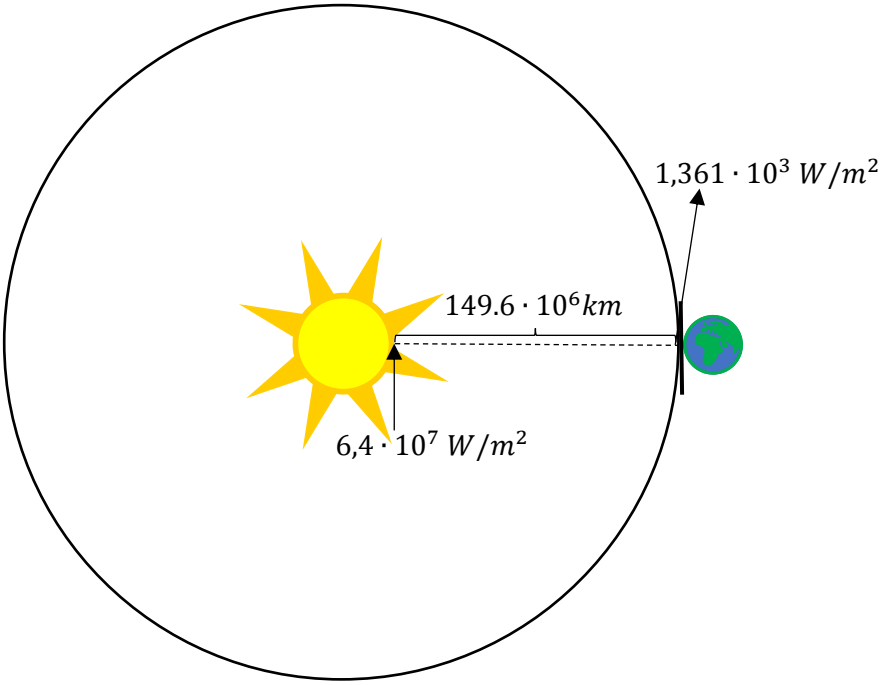



Figure 2.1 illustration of the decrease of solar irradiance from the Sun to the Earth. Due to the distance between the Earth and the Sun, the irradiance is decreased from $6{,}4 \cdot 10^{7}\ W/m^2$ at the photosphere to $1{,}361 \cdot 10^{3}\ W/m^2$ at the edge of Earth's atmosphere.

### 2.1.1 Atmospheric effects

The distance travelled by the radiation through the atmosphere is dictated by the angle of incidence to the atmosphere, together with the height above sea level of the object. The angle of incidence to the atmosphere is called the *solar zenith angle* $\theta_z$ and is illustrated in Figure 2.2. As the radiation travels through the atmosphere it is attenuated due to effects like absorption and scattering by dust particles, air molecules and components such as $CO_2$, $O_3$ and $H_2O$ (water vapour). Therefore, to compare the received radiation at different solar zenith angles the *air mass ratio* is used.

The air mass (AM) ratio is a comparison, at normal pressure, between the path of the radiation at normal incidence passing through the atmosphere and the path of the radiation at solar zenith angle $\theta_z$. When the radiation travels through the atmosphere at normal incidence then a standard mass of atmosphere is encountered by the radiation. While at solar zenith angle $\theta_z$ the mass of the atmosphere is increased in comparison because of the increased path. The air mass ratio is defined as:

$$\mathrm{AM} = \frac{1}{cos\theta_z} \qquad 2.1$$

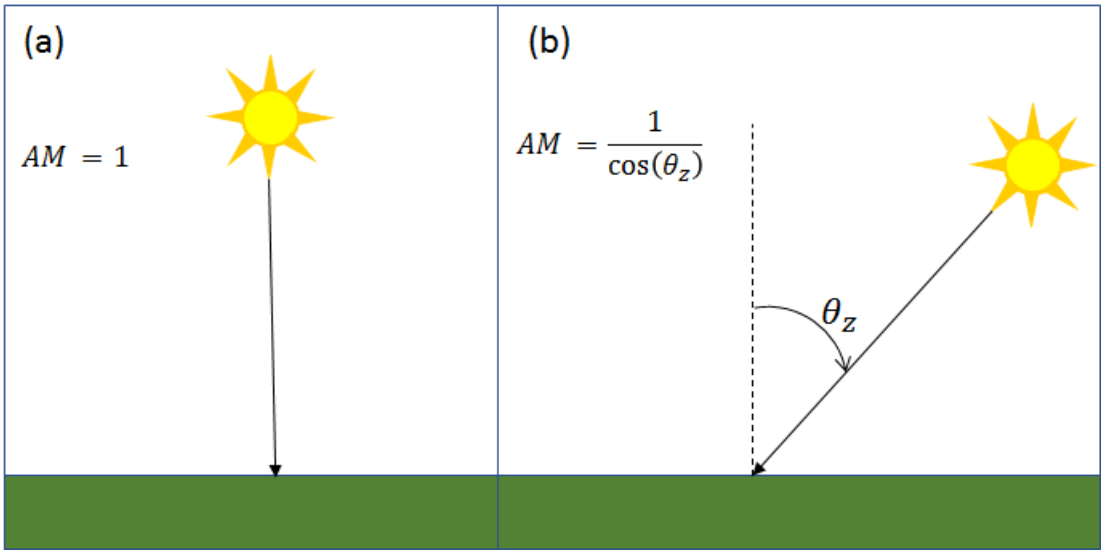



*Figure 2.2 The figure illustrates the relationship between the solar zenith angle and the air mass ratio. As the solar zenith angle increases the air mass ratio increases. Therefore, the sunrays travel through more air mass at lower solar zenith angles which contributes to lower irradiance levels during the morning and evening.*

As the solar zenith angle varies during the day, due to Earths rotations around its axis, so varies the AM. AM0 refers to zero atmosphere irradiance which is the solar spectrum outside the atmosphere. While AM1 is illustrated in Figure 2.2a and explains the air mass ratio when the sunrays are perpendicular to the Earth's surface. A decrease in AM will increase the absorbed energy at the surface of the Earth. Since the AM varies throughout the day, the amount of attenuated radiation also varies and thus varies the absorbed energy by the Earth's surface.

### 2.1.2 Solar spectrum

The relationship between radiation and energy was discovered by Planck and later interpreted by Einstein in 1905. Einstein explains that radiation comes in quanta of energy with the size of

$$E_{ph} = hv \qquad 2.2$$

where $E_{ph}$ is the quanta or photon energy, while $h$ is placks constant and $v$ is the frequency of the light.

Like all matter with a temperature above absolute zero, the Sun also emits thermal radiation. The energy emitted by the sun has a unique spectral distribution almost identical to the one of a black body at the temperature of 6000 K. The spectral distribution at AM0 and AM 1.5 is shown in Figure 2.3. The solar constant is the area under the solar spectrum curve at AM0.

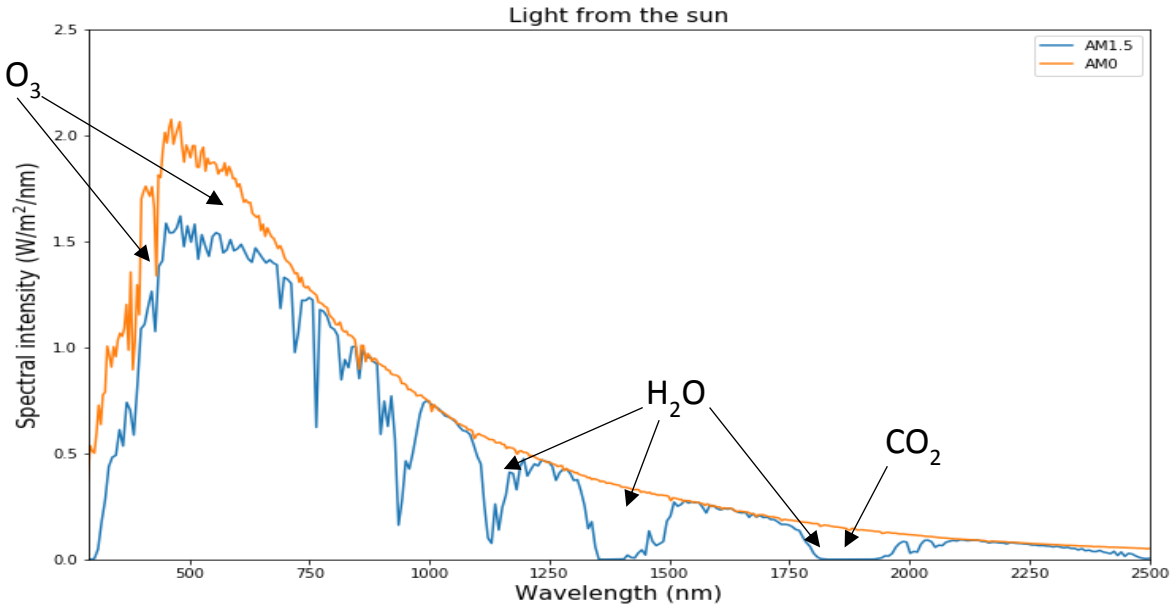



*Figure 2.3 Spectral distribution of the solar radiation. The spectral distribution of the solar radiation at AM1.5 and AM0 illustrated in blue and orange. The decrease in solar radiation from AM1.5 compared to AM0 is due to effects like scattering and absorption by the components marked in the figure. AM1.5 spectrum data from (NRL, 2019) and AM0 spectrum from (Anon., 2019)*

### 2.1.3 Seasonal variation

As described in Chapter 2.1.1, the distance between the Earth and the Sun has an impact on the intensity of solar radiation absorbed by the Earth. Longer distance results in a decrease in intensity, on the other hand, smaller distance increases the intensity. As the Earth travels around the Sun within approximately 365 days, illustrated in Figure 2.4, the distance varies and thus varies the absorbed solar radiation. More importantly, the motion of the Earth around the Sun is tilted by 23.5° towards the equator, which contributes to seasonal effects. Therefore, the northern hemisphere will experience summer around June, while the southern hemisphere experience winter. In the months around June, the northern hemisphere is tilted

towards the Sun and thus, receives solar radiation at a lower solar zenith angle than the southern hemisphere. As a result, the northern hemisphere receives more radiation due to a lower AM. Similarly, the southern hemisphere experience summer around December, while the northern hemisphere experience winter due to the same effect.

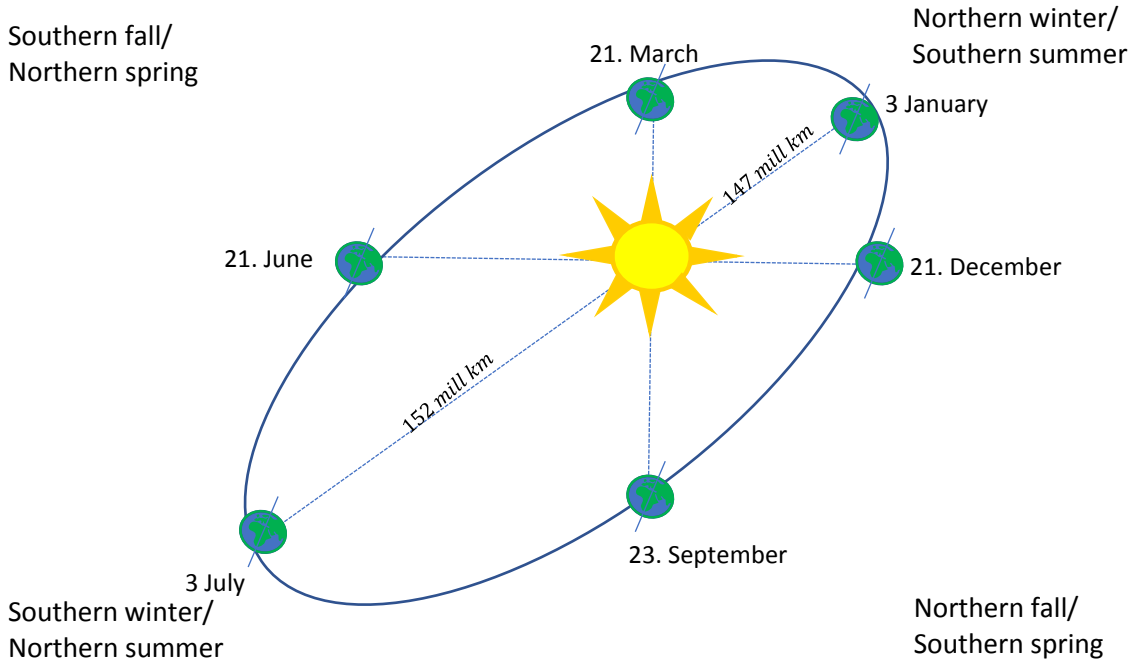



*Figure 2.4 Illustration of Earth's rotation around the Sun. The elliptical form is strongly exaggerated. The figure illustrates the distance between the Earth and the Sun under different periods of the year.*

In addition to the increase in air mass ratio with an increase in zenith angle, the sunrays are also spread over a larger area. Consequently, the intensity of the solar radiation decreases further. As demonstrated in Figure 2.5 the sunrays are spread over a larger area $L_1$ during summer and $L_2$ during winter, where $L_1 < L_2$.

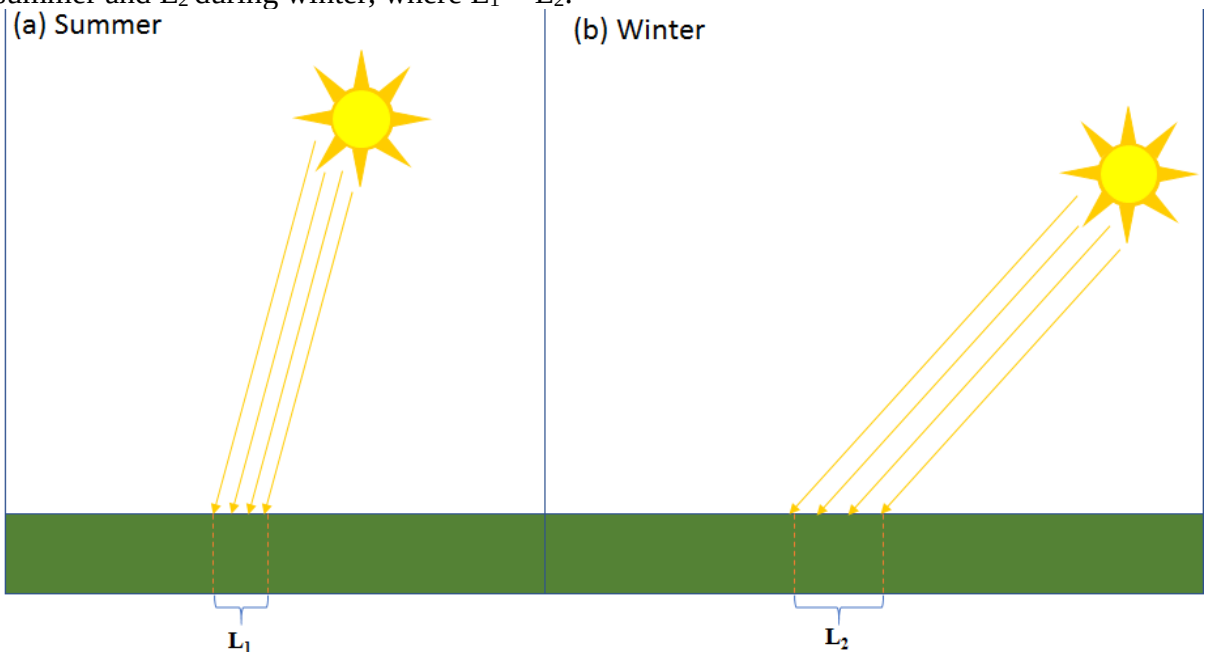



*Figure 2.5 Solar positions during summer figure (a) and winter figure (b): The figure shows how the solar intensity varies due to the motion of the Earth relative to the Sun. Inspired from (Øgaard, 2016)*

## 2.2 Solar cells, modules, strings and arrays

After the solar radiation has transmitted through the atmosphere it can be converted to energy by photovoltaic cells at Earth's surface. The following subchapter describes the conversion from solar radiation to electrical energy in PV systems.

### 2.2.1 Solar cells

A solar cell is composed of semiconductor material, the two main components in a solar cell are the *p-type* material and the *n-type* material. P-type material are positively doped semiconductors, implying a deficit of electrons. Whereas n-type material are negatively doped semiconductors, implying a surplus of electrons. Creating a surplus or deficit of electrons can be accomplished by a process called *doping*. Doping is a process that adds impurities of dopant ions to a semiconductor. When ions of less valency enter the original semiconductor, it becomes an electron acceptor producing a positive (p) type material. On the other side, atoms of greater valency become electron donors, producing a negative (n) type material. Although the explanation of p- and n-type material may indicate that there are two entirely different components, that is not the case. In fact, these material regions appear on the same component and the region where the material properties change is defined as the *p-n junction*.

The p-n junction is created due to diffusion between the n- and p-type material in the region they connect. Eventually, the diffusion will stabilize, reaching a steady state condition. Due to the accumulation of charges of opposite sign an electric field is created. Therefore, the region where the n- and p-type material interconnect becomes a depletion region. The p-type, n-type and p-n junction is illustrated with Figure 2.6 the figure also illustrates the main principles behind the band gap model.

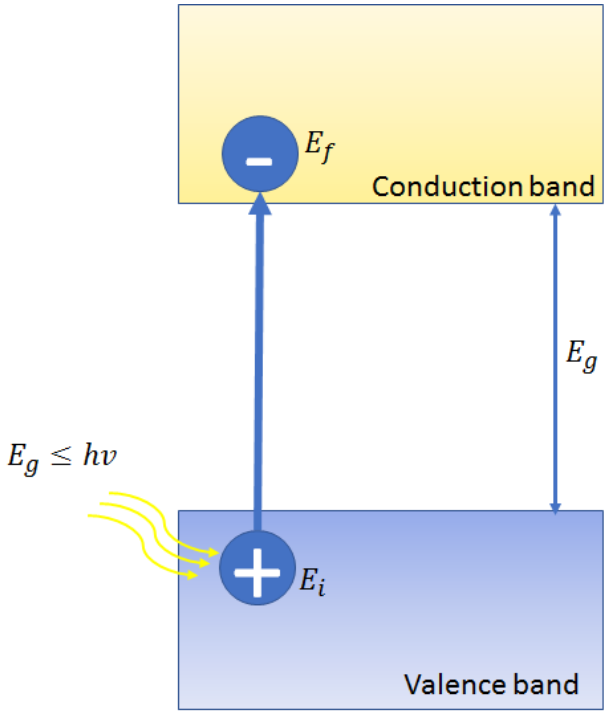



*Figure 2.6 Illustration of the band gap model. The fading blue region illustrates the n-type region while the fading yellow region represents the p-type region. Lastly, the region between those two regions is the p-n junction. The positive circle is a hole, where holes refer to missing electrons in an atom bond and can be regarded to have a positive charge. In contrast, the negative circle is an electron and has a negative charge.*

The band gap model illustrated in Figure 2.6 describes how photons can excite electrons in the valence band into the conduction band. Due to the depletion region in the p-n junction, there is a potential energy difference the electrons must overcome to be excited. The potential energy difference is the minimum energy the photons must have to excite an electron. Electrons will be exited from an initial energy $E_i$ to a higher energy level $E_f$, where the band gap energy $E_g$ is equal to the difference between the final energy and the initial energy. Photons with energy higher than the band gap energy will excite the electrons, as described in the following criteria

$$E_g \leq hv \qquad 2.3$$

The size of the band gap energy depends on the semiconductor material and, as discussed in Chapter 2.2.1.1, the temperature of the material.

Under illumination with photon energy above the band gap, the band gap will increase if it does not exist an external connection between the n-type and p-type material. Due to the illumination, electrons will be excited to the n-type region and thus build up a voltage across the p-n junction. The direction of the voltage is the opposite direction of the original voltage, therefore current is generated to compensate for the electron current. When the two currents reach an equilibrium the *open-circuit* voltage is established, which can be seen from Figure 2.7a).

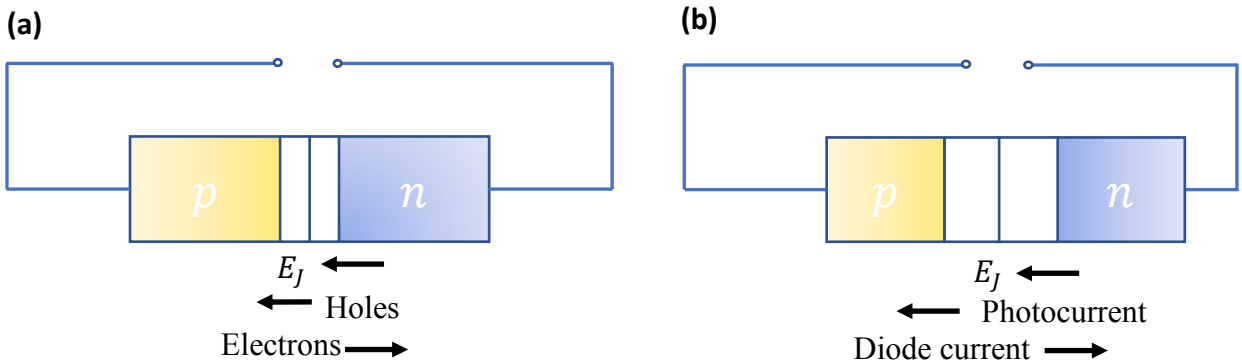



Figure 2.7**:** The solar cell. (a) under open circuit conditions and (b) under short circuit conditions. Reused with permission from (Chen, 2011). $E_J$ is the electric field in the p-n junction.

On the other hand, when the n- and p-type materials are externally connected, under illumination with photon energy above the band gap, an electrical current is generated. Under illumination, negatively charged electrons will travel through the band gap into the n-type region creating a current, which can be seen from Figure 2.7b. Without a load connected the generated current is the *short-circuit* current provided by the solar cell. However, the regular operating conditions of the solar cell is somewhere between open-circuit and short-circuit conditions.

A solar cell regularly operates in a point between open-circuit and short-circuit conditions, in such a scenario an external load is connected. The scenario described is illustrated in Figure 2.8. In the figure, the box to the left represents the solar cell, while the box to the right represents the external load. The figure is also an illustration of the two-diode model which represent the equivalent circuit of a solar cell.

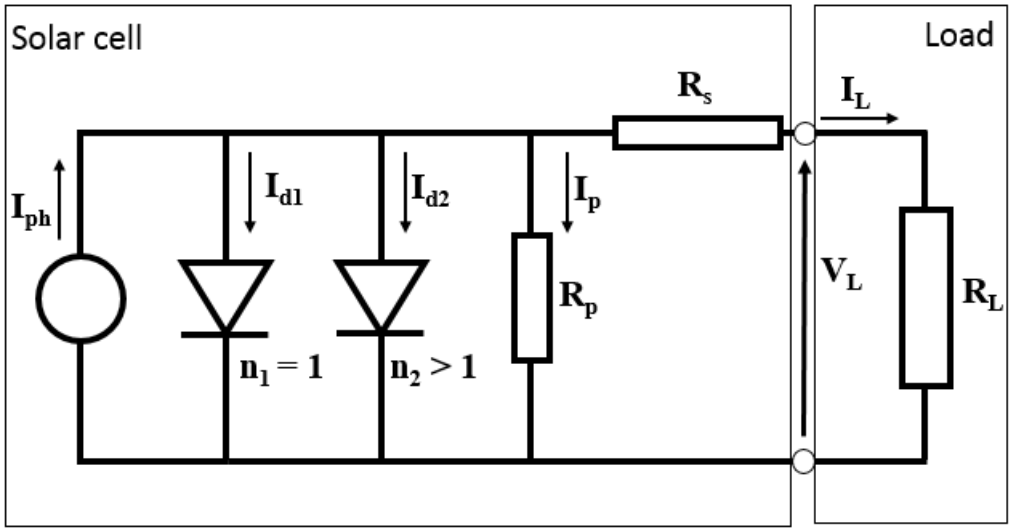



*Figure 2.8 The two-diode model. Illustrates the equivalent circuit of a solar cell in the left box and connection to an external load in the right box. Whereas the solar cell consists of two diodes and two resistances.*

In the figure, diode 1 is a result of the p-n junction generated in the solar cell, in addition a series resistance $R_S$ and a shunt resistance $R_P$ represents the internal resistance in the solar cell. A second diode is used to represent the *recombination* present in the p-n junction. Recombination is the process where electrons recombine with holes in the p-n junction. Therefore, the relationship between current and voltage in the two-diode model can be described by equation 2.5, which is derived by equation 2.4

$$I_L = I_{ph} - I_{d1} - I_{d2} - I_p \tag{2.4}$$

in the equation $I_L$ is the current which passes through the load, $I_{ph}$ is the current generated from the photons, while $I_{d1}$and $I_{d2}$ are the current through diode 1 and diode 2 respectively. Furthermore, the current through the shunt resistance is denoted as $I_p$. The expression is equivalent to

$$I_L = I_{ph} - I_{01}\left\{exp\left[\frac{q(V_L - I_L(R_S + R_L))}{n_1 K_B T_{cell}}\right] - 1\right\} - I_{02}\left\{exp\left[\frac{q(V_L - I_L(R_S + R_L))}{n_2 K_B T_{cell}}\right] - 1\right\} + \frac{V_L - I_L(R_S + R_L)}{R_p} \tag{2.5}$$

where $n_1$ and $n_2$ are the ideality factor of diode 1 and 2, $I_{01}$ and $I_{02}$ are the saturation current for diode 1 and diode 2. Furthermore, $K_B$ is the Boltzmann's constant, $q$ is the elementary charge and $T_{cell}$ is the solar cell temperature. The relationship between the current and the voltage of a solar cell can be further evaluated by the *I-V curve*.

The current-voltage (I-V) curve of a solar cell visualise the output current as a function of output voltage and defines the operating characteristics of the solar cell. As illustrated by Figure 2.9 the current can range from 0 (open circuit) to $I_{sc}$ (short circuit) while the output voltage ranges from $V_{oc}$ (open circuit) to 0 (short circuit). The rectangular area formed by a point on the I-V curve and origin indicates the operating power of the solar cell. Thus, the power output from the solar cell $P_{cell}$ is defined as

$$P_{cell} = I_L \cdot V_L \tag{2.6}$$

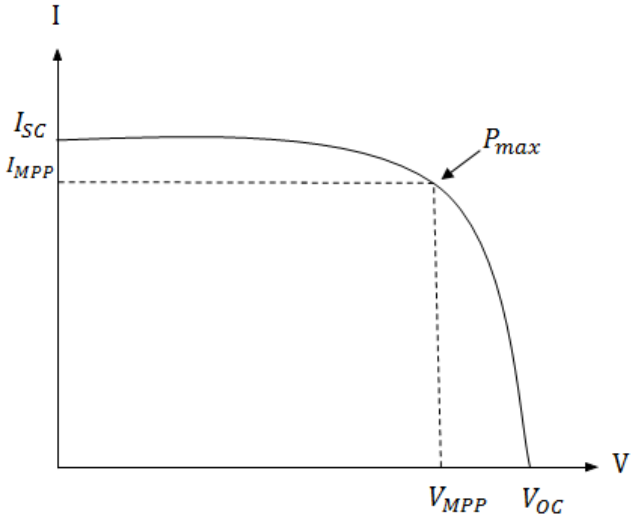



*Figure 2.9 I-V curve: the figure illustrates the I-V curve of a solar cell. The rectangular areal defined from the origin to the operating point is the power output from the solar cell. The ratio between the areal made from MPP and the areal from $V_{oc}$ and $I_{oc}$ is known as the fill factor.*

As a result, the maximum power produced by the solar cell is given by the current and the corresponding voltage which maximises the rectangular area under the I-V curve. The operating point described is known as the *maximum power point* (MPP) for the solar cell. The maximum power output from the solar cell can be used to determine the efficiency of the solar cell. Whereas the efficiency of the solar cell ($\eta$) can be determined as the ratio of the produced power and the irradiance on the cell ($G$) multiplied with the area of the cell $A_{cell}$.

$$\eta = \frac{P_{cell\,MPP}}{P_{in}} = \frac{I_{MPP} \cdot V_{MPP}}{G \cdot A_{cell}} \tag{2.7}$$

In equation 2.7 $P_{cell\,MPP}$ is the maximum power output from the solar cell achieved at the MPP. $I_{MPP}$ *and* $V_{MPP}$ are the corresponding current and voltage when the cell operates at the maximum power point. Lastly, $P_{in}$ is the power input to the cell.

#### *2.2.1.1 Causes of temperature increase and effects*

As indicated by equation 2.7 the efficiency of a solar cell is described as the ratio of the produced power by the cell and the amount of incident solar irradiance. At most, almost 97 % of the incident solar irradiance is absorbed by the solar cell. The rest of the solar irradiance is either reflected or transmitted by the cell. The percentage of absorbed solar irradiance varies with the zenith angle since more irradiance is reflected at higher zenith angles. The absorbed solar irradiance can be characterized into three regions.

The first region is the photons whose energy is less than the band gap energy $hv < E_G$. Absorption of photons within this energy range produces heat and no electricity. Whereas the second region is the amount of solar irradiance which is converted to electricity for photons with energy above the band gap energy. The last region is the amount of photon energy which dissipates as heat for photons with energy above the band gap energy. Due to heat dissipation, the internal temperature of the solar cell might increase.

As a result, the efficiency of the solar cell will decrease according to the traditional linear expression (Evans, 1981).

$$\eta_{cell} = \eta_{ref}\left(1 - \beta_{ref}\left(T_{cell} - T_{ref}\right)\right) \qquad 2.8$$

where $\eta_{cell}$ is the cell efficiency, $\eta_{ref}$ is the efficiency at the reference temperature $T_{ref}$ and solar radiation at $1000\ W/m^2$ (Evans, 1981). Lastly, $T_{cell}$ is the cell temperature and the temperature coefficient $\beta_{ref}$, is a material property. Both the efficiency at the reference temperature and the temperature coefficient are typically provided by the solar cell manufacturer. When the operating temperature of a solar cell increases the band gap in the semiconductor decreases and less energy is needed to excite electrons to the conduction band. The reason for the decrease in band gap energy is the increase in the electron's thermal energy. As a result, lower energy is needed to break the bonds between electrons and their atom. Thus, more electrons can be excited leading to a slight increase in the short-circuit current. Whereas the open circuit voltage will decrease leading to overall lower performance by the solar cell. The effects due to an increase in cell temperature are illustrated in Figure 2.10.

In conclusion, most of the solar irradiance incident on the solar cell is absorbed, however as the zenith angle increases, more solar irradiance is reflected. Furthermore, a major part of the absorbed solar irradiance is dissipated as heat due to energy either over or below the band gap. Consequently, the temperature in the solar cell can increase and the performance of the solar cell decreases. Therefore, to compare the performance of different solar cell technologies a set of standard test condition (STC) has been defined.

### *2.2.2 Standard test conditions*

The standard test conditions are defined by the International Electrotechnical Commission (IEC) in the standard *IEC 60904-3*. The standard describes basic measurements principles and test conditions for estimating the performance of a solar cell or module. The condition dictates an irradiance of $1000 W/m^2$, AM1.5 spectrum (illustrated in Figure 2.3) and a cell temperature of $25℃$. The AM1.5 spectrum is defined as irradiance and spectrum of sunlight

on a clear day incident on a 37° degrees tilted surface with the Sun at an angle of 41.81° above the horizon (Niclas, 2011). Furthermore, the nameplate rating and material properties of a solar module will typically be stated based on the STC.

### 2.2.3 Solar modules, strings and arrays

In utility-scale solar parks, millions of solar cells are installed to generate electricity. Typical values for the peak power produced by utility-scale solar plants can be in the range of kilo to megawatts. To produce this amount of power multiple solar cells can be connected in series to create a *solar module*, illustrated in Figure 2.10. In a solar module where the cells are connected in series, the voltage from the isolated solar cells is added up. Therefore, if one module consists of 4 cells all producing $0.6V$, the output voltage will be $2.4V$.

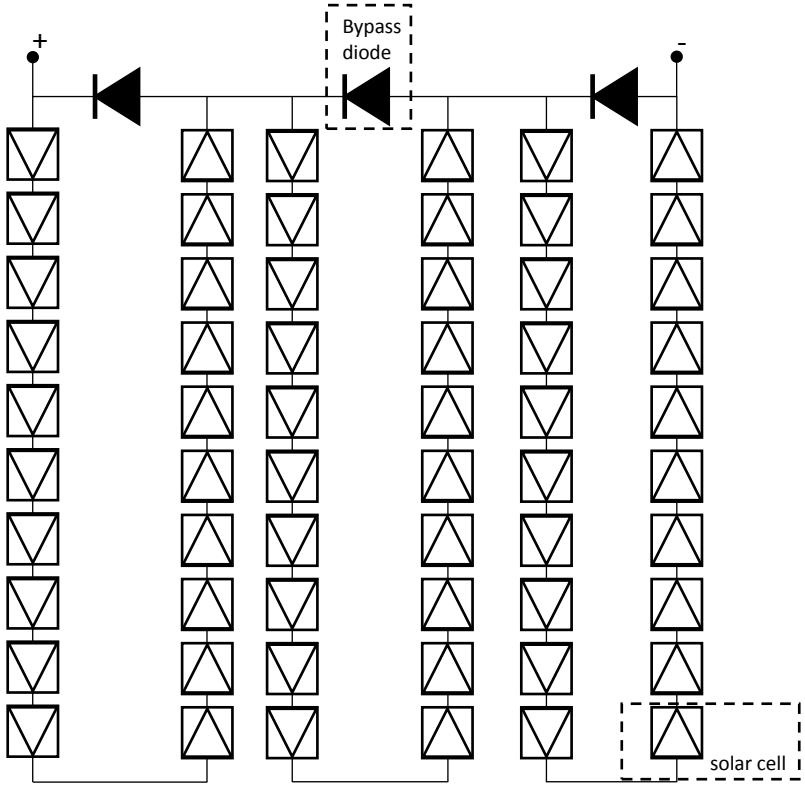



*Figure 2.10 Illustration of composition of a solar module. As for the solar module used in this thesis, the illustration consists of 60 solar cells in a 6x10 formation where there are 3 bypass diodes.*

Unlike voltage, the current does not add up when the cells are connected in series. The solar cell which delivers the least amount of current determines the current passing through each cell. To prevent significant current limitations due to limitations from a single solar cell, one or multiple bypass diodes can be connected in parallel with the solar cells. The cells are then interconnected and fixed within a weatherproof encapsulation which makes up the module. To further scale up the power production *strings* are used.

Strings are multiple solar modules connected in series. In a string, the voltage from each module is aggregated, whereas the current is constant throughout the string and limited by the current output of the lowest-performing module. Ultimately, multiple strings can be connected in parallel, creating an *array*, before the system is connected to a load. The aggregation from solar cells to modules, to strings and arrays are shown in Figure 2.11.

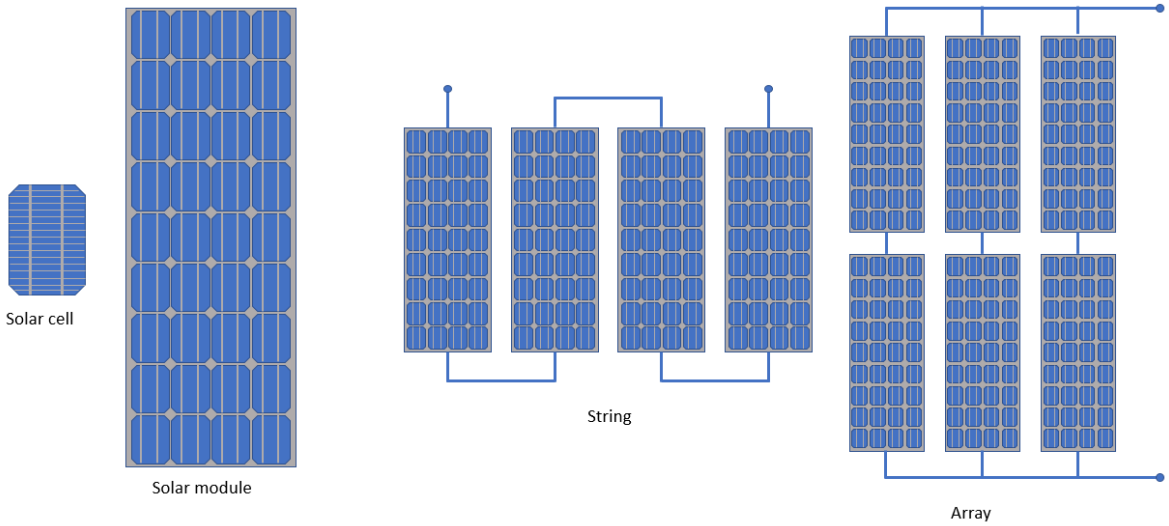



*Figure 2.11 Aggregation from a solar cell to an array. The figure shows the accumulation of multiple solar cells which makes up the solar module. Furthermore, the solar modules can be connected in series which makes up a string. Finally, multiple strings can be connected in parallel and together make up an array. In the solar cell, the vertical lines are referred to as busbars and conduct the electricity within the cell.*

## 2.3 Inverters

In the case of utility-scale solar parks; modules, strings and arrays are typically not directly connected to a load. Arrays are usually connected to an *inverter*. An inverter transforms the DC generated from the PV system into AC which can be used by the grid. A key role of the inverter is to create a voltage curve which complies with the relevant grid conditions. In addition, a maximum power point tracker (MPPT) is usually incorporated within the inverter. Thus, the inverter varies the load connected to the PV system to force the system to operate in the maximum power point as indicated by Figure 2.9. To maximise the performance of the solar cells ideally, there would be one inverter for each module. However, this is not typical since such a setup would be cost inefficient.

While there are many types of inverters, this study will focus on the central inverters illustrated in Figure 2.12. Central inverters are the inverter configuration mainly used in utility-scale solar power plants. Central inverters are popular within the utility-scale solar power production domain due to the lower unit price per MW. In addition, central inverters obtain high efficiency for a broad range of array outputs.

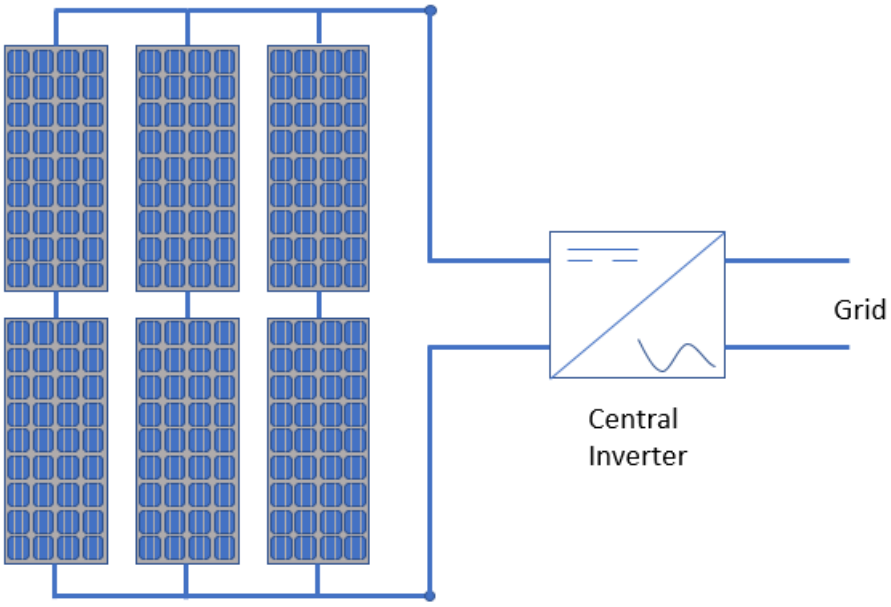



*Figure 2.12: Illustration of the central inverter setup. An array containing strings of modules connected to a central inverter. Other types of inverters are string inverters and module inverters amongst others.*

In contrast to other inverter configurations, central inverters require relatively more DC cabling which increases cable losses. Another downside of the central inverter is due to the common MPPT for the entire array. The strings in the array can be subject to different operating condition due to topography, shading, difference in module rating and hence have different MPPT's. Before the solar arrays can be connected to the grid by means of an inverter, they need to be mounted to the ground or optionally a roof.

## 2.4 Mounting system

To ensure safe operation of the PV system, in addition, to providing ventilation to the modules, a mounting system is required. The system must also be able to endure the weight of the modules and additional loads from weather conditions at the park (Størdal, 2013). Furthermore, the angles necessary to describe the orientation of the PV -module is dictated by the mounting system. Some of the most relevant angles are the solar azimuth angle $\gamma_s$ which is the compass direction from which the sunlight is coming. Together with the angle of incidence $\theta$, the two angles describe the Sun's position relative to the module. The angle of incidence is the angle between the sunray's incident on the module and the line perpendicular to the surface of the module. As described earlier, the solar zenith angle ($\theta_z$) is the angle between the zenith and the line from the Sun's centre. Moreover, to describe the orientation and mounting of the module the tilt angle $\beta_m$ and orientation angle $\gamma_m$ can be used. The tilt angle describes the angle between the module and the horizontal ground. While the orientation angle describes the angle between the perpendicular line from the modules, projected into the ground and the south (Pedersen, 2015).

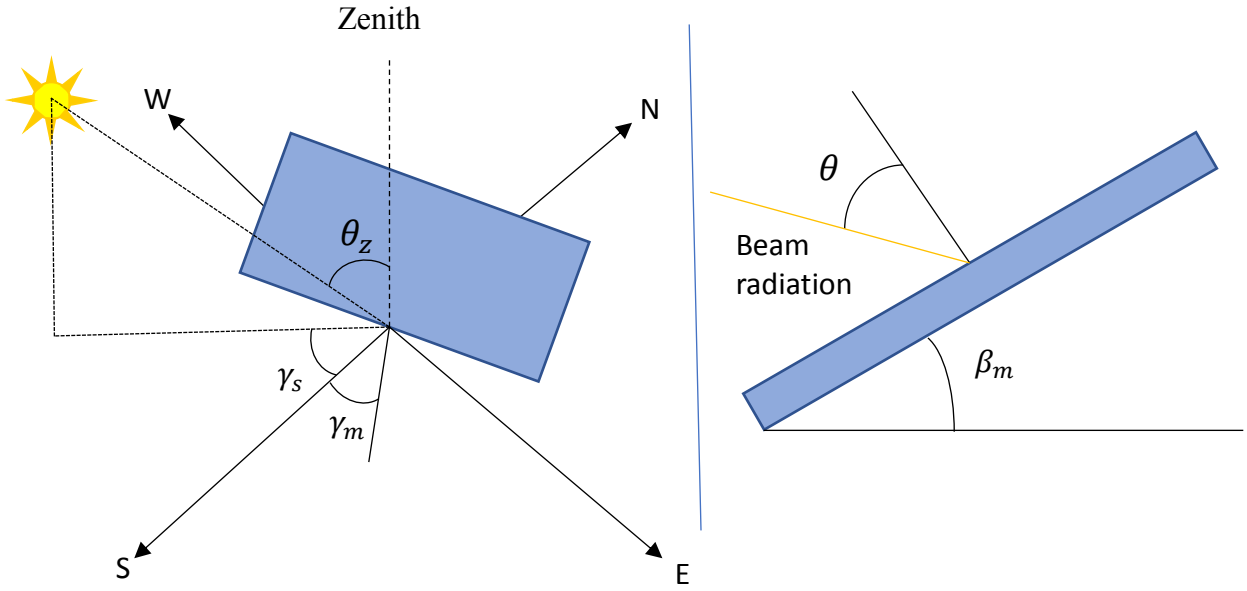



*Figure 2.13 relevant angles describing a PV module orientation.*

The mounting system can be installed with additional tracking devices and as a result, the modules will produce more energy. By means of one-dimensional tracking of the Sun, the modules will obtain a lower angle of incidence. Thus more of the radiation is absorbed. In addition to a lower angle of incidence, two-dimensional tracking of the Sun will provide a lower orientation angle. Consequently, the absorbed solar radiation will increase further. Finally, after modules have been mounted, arrays connected to the inverter, which is subsequently connected to the grid and a Supervisory Control and Data Acquisition (SCADA) -system has been established, the data can be analysed. Thereupon, models which estimate the operating conditions of the site can be established.

## 2.5 PV modelling

Based upon data measured from a PV-site, estimates for voltage, current and power production from the strings or inverters can be made. The following subchapter describes methods to estimates these values.

### 2.5.1 Physical baseline models

The power output from each solar module ($P_{DC}$) can be calculated according to the equation provided by (Dobos, 2014):

$$P_{DC} = \frac{G_{poa}}{1000W/m^2} P_{MP}(1 + \gamma_P \left(T_{cell} - T_{ref}\right)) \quad 2.9$$

Where $G_{poa}$ is the global plane of array irradiance, $1000W/m^2$ is the solar irradiance at STC, $P_{MP}$ is the maximum power at STC and $T_{ref}$ is the cell temperature at STC and $\gamma_P$ ($\gamma_P < 0$) is the Peak Power Temperature Coefficient. Lastly, $T_{cell}$ is the cell temperature. Furthermore, the power of the module decreases with an increase in cell temperature above the reference value which is in accordance with Chapter 2.2.1.1. In the same manner, the module current can be calculated.

The current output from each module ($I_{DC}$) can be calculated using a similar equation (ALQahtani, et al., 2012).

$$I_{DC} = \frac{G_{poa}}{1000W/m^2} I_{MP}(1 + \gamma_I \left(T_{cell} - T_{ref}\right)) \quad 2.10$$

Where $I_{MP}$ is the current at maximum power point under STC, while $\gamma_I$ ($\gamma_I > 0$) is the Short-Circuit Current Temperature coefficient. The positive sign indicates an increase in current when the cell temperature exceeds the reference temperature. However, the relative increase in current is far less than the decrease in voltage. Which is due to the temperature coefficient for voltage. The voltage can be calculated by a combination of the equation for power and current:

$$V_{DC} = P_{DC}/I_{DC} \quad 2.11$$

Lastly to compensate for non-ideal effects, such as cabling losses and more, an ideal factor $\eta_{ideal}$ is multiplied into the equations 2.9-2.11. The ideal factor can be calculated by training the models on a subset of the data with the respective response values. Furthermore, the above equations can be multiplied with the number of strings or modules to represent the system connected to an inverter. Similarly, a representation of the PV system can also be obtained using machine learning applied to the data from a PV system.

## 2.6 Machine learning

Machine learning is a subfield of Artificial intelligence, which involves self-learning algorithms that derives knowledge from data in order to make predictions (Rachka & Mirjalili, 2017). Machine learning is commonly used within topics such as spam filters, text and voice recognition and web search engines (Rachka & Mirjalili, 2017) (Rodrigues, et al., 2018). Within the field of machine learning, there are three different types: supervised learning, unsupervised learning and reinforcement learning. This thesis will focus on supervised learning.

### 2.6.1 Supervised learning

Supervised learning learns a model based on the behaviour of the system from labelled training data, thereby allowing prediction on unseen data using the trained model. Therefore, the response values of the system are known in the training of the model. In contrast to supervised learning, unsupervised learning deals with unlabelled data or data of unknown structure (Rachka & Mirjalili, 2017). Supervised learning can be applied to both regression and classifications tasks. As a result, using explanatory variables and response values of the PV system, a model can be obtained to represent the behaviour of the system. The response values are used to direct the model to the most optimal solution. Similarly, to the physical models represented in Chapter 2.6, measurements of cell temperature and irradiance can be used as explanatory variables and power, current and voltage as response values to create such a representation. However, with access to training data, a crucial part of supervised learning is to obtain a good compromise between bias and variance (Rachka & Mirjalili, 2017).

### 2.6.2 The bias-variance trade-off

Bias and variance can be explained mathematically by expressing the expected prediction error of a regression model. The expected prediction error of a model $\hat{f}(X)$ with given explanatory variables $X = x_0$ and true values $Y$, where $Y = \mathrm{f(x)} + \varepsilon$ and $\mathrm{E}(\varepsilon)=0$ has squared-error losses as (Hastie, et al., 2009);

$$Err(x_0) = Irreducible\ Error + Bias^2 + Variance \qquad 2.12$$

Whereas the first term Irreducible Error, $Var(\varepsilon) = \sigma_\varepsilon^2$, is the variance of the response value around its true mean and can be regarded as the noise in the data (Hastie, et al., 2009). The second term is the squared bias and is the amount by which the average of the predictions differs from the true mean (Hastie, et al., 2009). The last term is the variance, which is the expected squared deviation of $\hat{f}(X)$ around its mean (Hastie, et al., 2009). Access to training data allows the model to memorise the data. Therefore, a complex model can decrease bias and increase the variance by memorising the training data. As a result, the model will overfit the data. The Bias variance trade-off is illustrated in Figure 2.14.

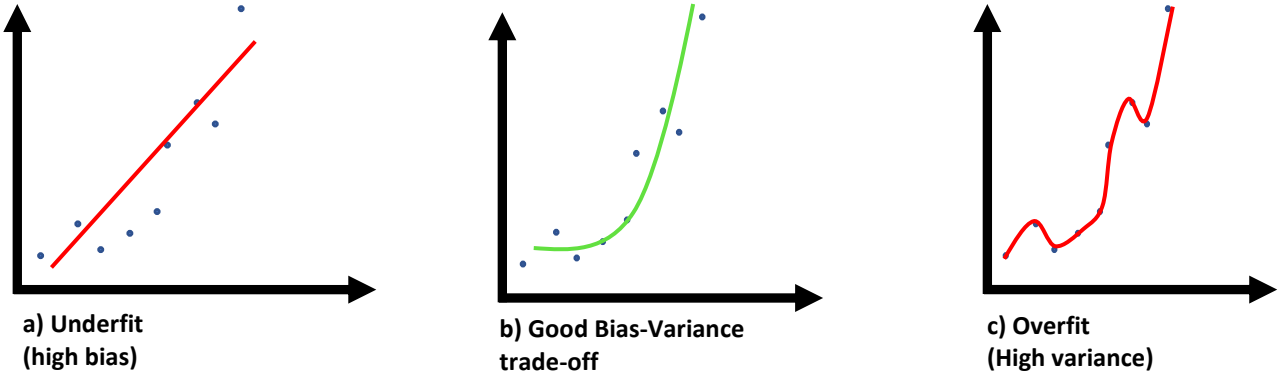



*Figure 2.14: Illustration of the Bias-Variance trade-off. The y-axis is the response value, whereas the x-axis is one of the explanatory variables. Figure a) illustrate the model in red, which has low bias and underfits the data. Figure b has a good bias-variance trade-off and figure c memorise the measurements and has a high variance. Therefore, the model in figure c has a high variance and overfits the data.*

In conclusion, a model with high bias will have a poor performance on the response value due to the lack of complexity to capture the relationship in the training data illustrated in Figure 2.14a (Rachka & Mirjalili, 2017). As a result, the model underfits the data and will not have a good performance on the training data or the test data. In contrast, Figure 2.14c illustrates a complex model which overfits the data. The complex model will have a good performance on the training data but poor performance on test data due to memorisation of training data (Rachka & Mirjalili, 2017). The goal of all machine learning models is to achieve a good bias-variance trade-off illustrated in Figure 2.14b. In practice, the result of applying machine learning models can be the situation in figure c. However, by introducing a form of regularization or change in the parameters of the model, the variance of the model can decrease and the bias slightly increase. As a result, the total model error will decrease.

### 2.7.3 Regression algorithms

There exists a vast amount of regression algorithms linear, tree-based and memory based. Common for all, is their goal of predicting a response value based upon a set of explanatory variables. This chapter will dive further into the detail about the algorithms with the best prediction ability used in this thesis. Therefore, some algorithms are left out, such as the support vector machine. The first and the most relatable algorithm, is linear regression.

**Linear regression**

A linear regression model assumes that the regression function $f(Y|X)$ is linear in the inputs $X_1, X_2, \ldots, X_n$ or that the linear assumption is a reasonable approximation (Hastie, et al., 2009). Therefore, the response value can be estimated as a linear combination of the explanatory variables:

$$\hat{f}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \beta_0 + \sum_{j=1}^{n} X_j \beta_j \qquad 2.13$$

Where $\beta_0$ is the interception at the y-axis, the $\beta_j$'s are unknown regression coefficients and the variables $X_j$ are the explanatory variables. The regression coefficients can be found using

training data and apply the least squares approach to minimize the residual sum of squares (RSS) as follows (Hastie, et al., 2009):

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \hat{f}(X_i)\right)^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{n} X_{i,j}\,\beta_j\right)^2 \quad 2.14$$

Where $i$ is the measurement number from 1 to the total number of measurements (N) and j is the variable number from 1 to n. As a result, the least square solution for $\hat{\beta}$ is:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad 2.15$$

Finally, the response values can be estimated with the equation presented in equation 2.13. However, to obtain better bias-variance trade-off, different *shrinkage methods* can be applied. Shrinkage methods shrinks the regression coefficients by imposing a penalty on their size (Hastie, et al., 2009).

**Ridge**

Ridge regression is a form of shrinkage methods whereas the ridge coefficients minimise a penalised residual sum of squares as follows (Hastie, et al., 2009):

$$\beta^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\,\beta_j \right)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \right\} \quad 2.16$$

Where $\lambda > 0$ is the penalisation parameter, high $\lambda$ indicates a strong penalisation, leading to a greater amount of shrinkage. A low $\lambda$ indicates a low degree of penalisation which allows a higher value of the regression coefficients. Similarly, as to without any penalisation, the ridge regression solutions can be obtained by coefficients that minimise the residuals sum of squares (RSS):

$$RSS(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \quad 2.17$$

Therefore, the ridge regression solution ($\beta^{Ridge}$) can be obtained by:

$$\beta^{Ridge} = (X^T X + \lambda I)^{-1} X^T y \quad 2.18$$

Where $I$ is the $nxn$ identity matrix. In contrast to ridge regression, lasso regression focusses on the absolute value of the regression coefficients.

**Lasso**

Whereas the Ridge regression penalise the squared sum of the coefficients expressed as $\sum_{j=1}^{n} {\beta_j}^2$, Lasso penalise the absolute value of regression coefficients $\sum_{j=1}^{n} |\beta_j|$. Therefore, the regression coefficients obtained by lasso is defined by:

$$\beta^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\, \beta_j \right)^2 + \lambda \sum_{j=1}^{n} |\beta_j| \right\} \qquad 2.18$$

The solution to the lasso regression can be obtained by quadratic programming (Hastie, et al., 2009). In contrast to Ridge regression, Lasso regression can shrink some regression coefficients to zero by increasing the penalisation term $\lambda$. As for ridge, the penalization parameter $\lambda$, dictates the degree of penalization. The impact of ridge regression and lasso regression is further visualised in Figure 2.15.

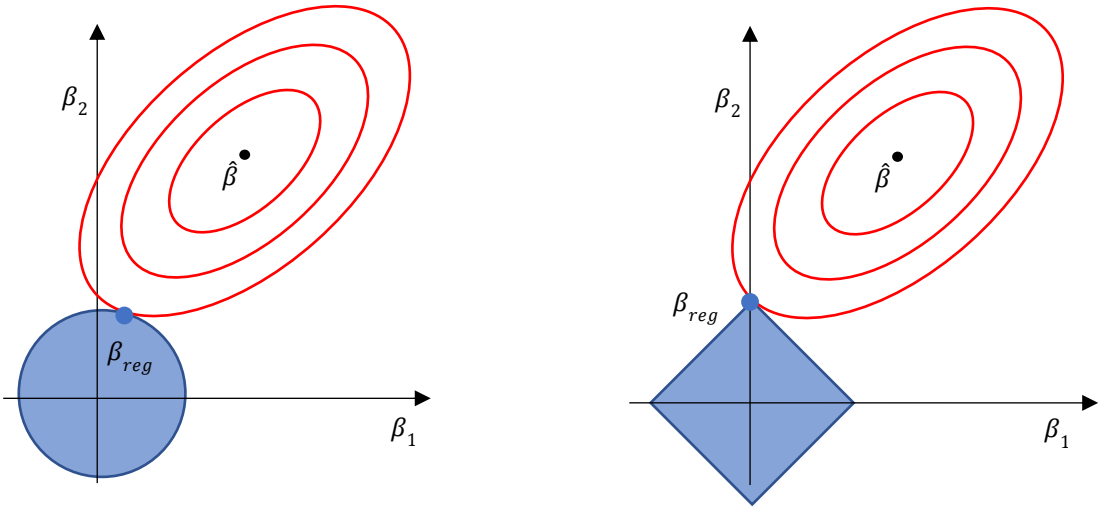



*Figure 2.15 the figure illustrates how the different penalisations affect the model coefficients. Where the ridge regression is illustrated to the left and the lasso regression is illustrated to the right. The red elliptical shapes are the residual sum of squares error function, while the solid blue figures are the constraint regions set by the respective penalisations (Hastie, et al., 2009).*

Both Lasso and Ridge regression assumes that the variables are centred before applied to the model.

**Standardisation**

A form of centring can be made by standardisation. Standardisation can be performed by subtracting the mean value of the features and dividing by the standard deviation of the same feature as shown in equation 2.19.

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x} \qquad 2.19$$

In equation 2.19 $x_{std}^{(i)}$ is the standardised feature number i, $x^{(i)}$ is the original data of feature number i, while $\mu_x \; and \; \sigma_x$ are the mean and standard deviation of the respective feature (Rachka & Mirjalili, 2017). After standardisation, features are centred with a mean equal to zero and a standard deviation equal to one. In addition to Ridge and Lasso regression, other algorithms perform better when standardisation has been applied to the explanatory variables. One of these algorithms is the memory-based regression technique known as K-nearest neighbour regression (KNN regression).

**K-nearest neighbour regression**

KNN is known as a memory based algorithm because the algorithm does not learn a specified function to evaluate future data, but rather memorises the training data instead (Rachka & Mirjalili, 2017). The KNN algorithm is straightforward and can be summarised as follows (Rachka & Mirjalili, 2017):

| **K-Nearest Neighbours regressor algorithm** |
|---|
| 1) Choose the k-number of neighbours to be evaluated when predicting a measurement. |
| 2) Choose the distance metric to evaluate the K-nearest neighbours. |
| 3) Find the K-nearest neighbours of the measurement which is going to be predicted. |
| 4) The predicted value is the mean of its K-nearest neighbours. |

(Rachka & Mirjalili, 2017) (Pedregosa, et al., 2019).

The distance to the neighbouring measurements can be calculated by different metrics such as the Manhattan distance, Euclidean distance or the more general Minkowski distance. Where the parameter p indicates the wanted distance metric, for $p = \langle 0, \infty \rangle$. For arbitrary values of p the Minkowski distance is used

$$D(X,Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \qquad 2.20$$

Where D is the distance matrix, x and y are two separate measurements where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ drawn from the data in X and Y. However, the Minkowski distance is typically used with p being 1 or 2, which correspond to the Manhattan distance and the Euclidean distance respectively. When the number of neighbours and the distance metric has been decided measurements can be predicted in terms of point four in the table above. To evaluate the mean value of the K-nearest neighbour's weights can be applied to the neighbours (Pedregosa, et al., 2019). The weights can be uniform; thus all points in each neighbourhood are weighted by an equal amount. Otherwise, weights can be given as the inverse of their distance. As a result, closer neighbours will have a greater influence on the mean than neighbours which are further away (Pedregosa, et al., 2019).

To sum up, KNN is an algorithm that memorises the training data and evaluates the K-nearest neighbours before the prediction of new data. The user can choose the number of neighbours to evaluate, the distance metric and how weights should be assigned in the prediction processes to find the mean. These parameters are therefore used to tune the bias-variance trade-off. As a result, KNN is a non-linear regression technique and due to the inner workings of the algorithm works well on *interpolated data*[1]. Another technique which works well on interpolated datasets is the Random Forest regressor.

[1] Interpolated data is used in this context to describe data where the values does not gradually change outside an interval. For example, the solar irradiance will never exceed 1361W/m$^2$ but rather circulate between 0 W/m$^2$ and approximately 1250 W/m$^2$.

**Random Forest Regressor - RFR**

As the name might imply, Random Forest regressor is an ensemble of multiple decision trees used for regression. A Random Forest builds a large collection of de-correlated trees and averages their results (Hastie, et al., 2009). The trees are made by iteratively splitting its nodes to obtain a reduction in mean squared error. The root node of the tree represents the subset of data, whereas branches represent a smaller subset of data from the previous node. Finally, the leaves on the tree are the value dedicated to the corresponding branch as illustrated in Figure 2.16.

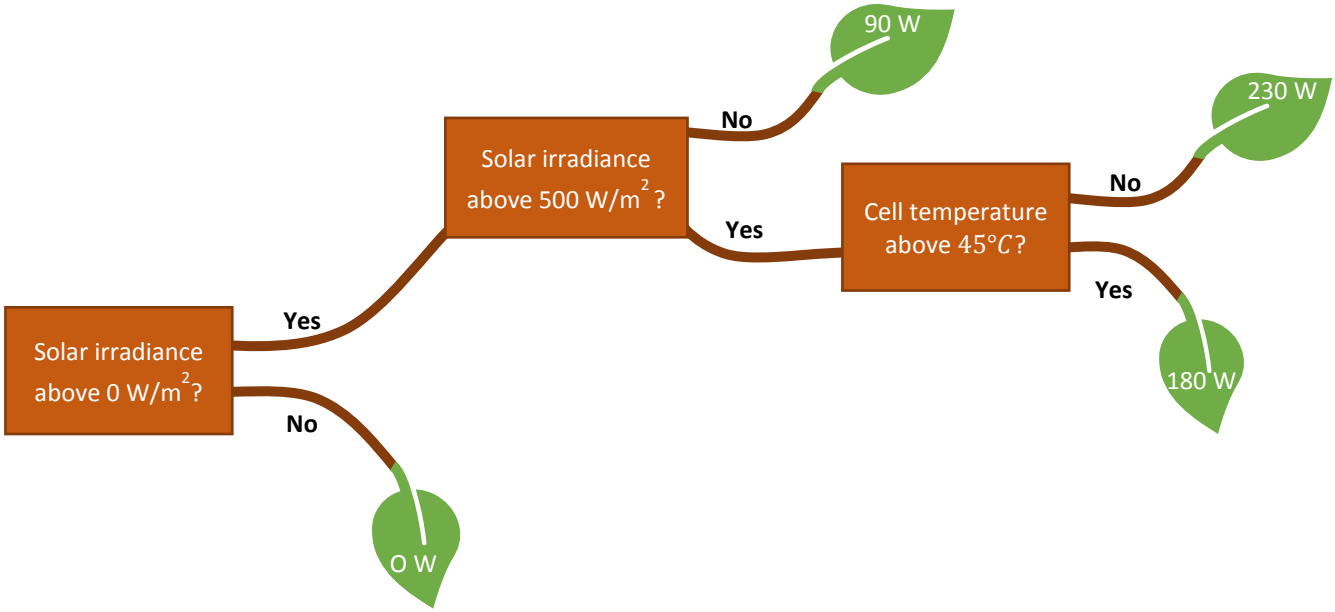



*Figure 2.16 Illustration of decisions and outputs of a decision tree. A decision tree is based upon multiple queries to group the data. Prediction of measurements is therefore based upon the mean value of the measurements in a leaf. In the figure, the decision tree is applied to data from a solar module, where the explanatory variables are solar irradiance and cell temperature. Furthermore, the response value is the produced power from the module.*

Nodes are split based upon the highest decrease in mean squared error (MSE) as indicated by equation 2.21 (Rachka & Mirjalili, 2017)

$$MSE(t) = \frac{1}{N_t} \sum_{i \in D_t} \left(y^i - \hat{y}_t\right)^2 \qquad 2.21$$

Where $N_t$ is the number of training samples at node t, $D_t$ is the training subset at node t, $y^i$ is the true response value and $\hat{y}_t$ is the predicted response value. $\hat{y}_t$ is calculated as the sample mean in accordance with equation 2.22

$$\hat{y}_t = \frac{1}{N_t} \sum_{i \in D_t} y^i \qquad 2.22$$

In a decision tree, the maximum depth indicates the maximum number of nodes made from the root node to the leaf node and can be used to tune the bias-variance trade-off. The tree will continue to generate leaves until the maximum depth is reached, the number of measurements in the node is less than a user-defined threshold or that the splits do not decrease the MSE above a certain threshold value (Pedregosa, et al., 2019).

In a Random Forest Regressor, multiple decision trees make up the model. Therefore, the algorithm behind Random Forest regression can be summarised as follows:

**Random Forest regressor algorithm**

1: For b=1:B, where B is the total number of decision trees.
   a. Draw a bootstrap sample Z* of size Q from the training data
   b. Grow a random forest tree $T_b$ to the bootstrapped data, which is an ordinary decision tree, as described on the previous page, applied to a random subset of the data. Recursively repeat the following step for each node until the limitations for leaf generation is achieved.

      i. Select m variables at random from the N variables in the data
      ii. Pick the best variable/split-point among the m variables, to achieve the maximum decrease in MSE.
      iii. Split the node into two new nodes.

2. The output is the ensemble of trees $\{T_b\}_1^B$ whereas the prediction of a measurement is the average prediction across the randomly generated decision trees such as:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad 2.23$$

(Hastie, et al., 2009)

Where $\hat{f}_{rf}^{B}$ is the random forest model, B is the number of decision trees and x is a measurement. The random forest model inherits the parameters used in the decision trees to tune the bias-variance trade-off. Also, the number of estimators (trees) in the model (B) can be used to tune the bias the bias-variance trade-off. As a result, the RFR usually has better performance than individual decision trees, in addition to being less sensitive to outliers in the dataset and less sensitive to non-informative features (Rachka & Mirjalili, 2017).

**Section summary**

To summarise, four different machine learning models have been introduced together with the concept of bias and variance. The models are Ridge, Lasso, K-nearest neighbour (KNN) Regressor and Random forest regressor (RFR). With each of the algorithm's, parameters can be tuned to vary the bias-variance trade-off. Common for all the models is that they do not know the physical behaviour of the system, but rather mimic the system based upon explanatory variables and response values. Furthermore, Ridge and Lasso are linear, whereas KNN and RFR are non-linear. These models will be applied to the data gathered from the PV system. Subsequently, these models will mimic the behaviour of the system and finally be used to detect power losses on the system.

### 2.7.4 Variable selection

Before the models will be used to detect power loss, a decision about which variables who are to be used as input to the models must be made. Although variables with information about solar irradiance and solar cell temperature are crucial for the evaluation of the PV system, other variables may increase the predictive power of the model. Therefore, a technique to evaluate the importance of each variable is crucial. Indeed, there exist many different techniques to evaluate the importance of a variable. Some of them are based upon the models already introduced, such as evaluation of the coefficients to a ridge model. In this technique, variables with small or zero coefficients are regarded as non-informative (Rachka & Mirjalili, 2017). Another approach is the random forest feature importance which evaluates variables based on their average reduction in MSE due to node-split, accumulated for all the decision trees (Hastie, et al., 2009) (Rachka & Mirjalili, 2017).

Still, these two methods have some drawbacks. The ridge evaluation is based upon the assumption that a linear relationship between the explanatory variables and the response value exists, which is not necessarily the case. Furthermore, the random forest feature importance values have proven to not always give an accurate picture of importance (Parr, et al., 2018). In fact, the mechanism for computing variable importance is biased and inflates the importance of continuous or high cardinality categorical variables (Parr, et al., 2018). High cardinality categorical variables are categorical variables with many unique values. Therefore, an alternative approach is a greedy search algorithm to decide which variables to use in the models.

**Sequential Feature selector**

A greedy search algorithm makes locally optimal choices at each stage of a search problem, thus yield a suboptimal solution to the problem in contrast to the exhaustive search algorithm (Rachka & Mirjalili, 2017). However, due to the efficient solution and decrease in computation time, a greedy search algorithm is often feasible (Rachka & Mirjalili, 2017). Sequential feature selection aims to reduce the dimensionality of the feature space with a minimum decrease in performance of the model. One of the sequential feature selector algorithms is the sequential forward selection (SFS). SFS achieves the aim of reduction in the feature space by the inclusion of the features whose presence improves the performance of the model the most. The algorithm starts with zero features and includes one feature at the time until the chosen threshold of features (p) is reached, where $p \leq$ original number of features. A pseudo code is provided, adapted from (Raschka, 2018)

**SFS algorithm**

1: Input $Y = \{y_1, y_2, \dots, y_d\}$, The whole d-dimensional feature set is input
2: Initialization: $X_k = \emptyset, k = 0$. Initialise the algorithm with an empty set $\emptyset$. Therefore, $k$=0 where $k$ is the size of the subset.
3: Inclusion
$x^+ = arg\ max\ J(x_k + x)$, where $x \in Y - X_k$. Finds a new feature within the original feature space that maximises the function J. J is a chosen objective function and can be the MSE score of the model. The chosen $x^+$improves the objective function the most and is an element within the original set Y subtracted by the already chosen variables $X_K$.
$X_{k+1} = X_k + x^+$. Adds the additional feature $x^+\ to\ the\ feature\ subspace\ X_k$
$k = k + 1$
If k==p:
  Break
Else:
  Repeat step 3
4: Output: $X_k = \{x_j \mid j = 1, 2, \dots, k;\ x_j \in Y\}, where\ k = (0,1,2,\dots,d)$

(Raschka, 2018)

In summary, the model will use the explanatory variables whose presence results in the best model performance. As a result, the chosen variables are not based upon any assumptions which can be associated with other techniques for feature selection presented. Furthermore, the chosen variables may differ between the machine learning models. As an example, KNN can use solar irradiance, cell temp and the hour of the day as the most optimal variables according to SFS. On the other hand, Lasso may use irradiance and the inverter temperature as the most optimal variables according to SFS. Finally, when the models are trained with the most optimal variables, they can be used to detect power losses on the PV system.

## 2.8 Power loss factors

When accurate and reliable representations of the PV system have been established, they can be used to detect power production losses in the PV system. Although several factors for power loss can strike in a PV system, the ones represented in this thesis has been established as the most common in the utility-scale power plant of interest;

### 2.8.4 Module faults

**Delamination**

Solar modules consist of multiple solar cells interconnected and fixed within a weatherproof encapsulation. The most common method for encapsulation is lamination (Dìez-Mediavilla, et al., 2014). Delamination is defined as the breaking of the ties between the layers of material in a laminate (Dìez-Mediavilla, et al., 2014). Delamination can occur both on the front- and back-side of the solar module. Delamination on the front-side of the module can increase the reflection of light. Furthermore, weather and dust can penetrate the module structure. On the other hand, delamination on the backside of the module complicates heat transfer on the backside of the module. Thus, increases the possibility of hot spots (Dìez-Mediavilla, et al., 2014).

**Partial shading**

Partial shading is when a part of a string receives substantially less illumination. The solar cells which receive less illumination will reduce the overall current in the module and the bypass diode can kick in. Thus, the performance of the module will decrease because of two reasons.

1) The shaded cell(s) will receive less irradiation and hence generate less current.

2) The losses of other cells connected in series will increase as they limit their current output to that of the shaded cell. (Rmaprabha & Dr. Mathur, 2009).

Therefore, partial shading can result in more damage such as diodes in reverse bias. As a result, the cell can experience severe damage due to overheating (Ramaprabha & Mathur, 2008). A result of overheating can be hot spots.

**Hot spot**

A hot spot can occur when a solar cell generates less current than the remaining cells in the series. A solar cell can generate less current than the remaining cells due to damage, total or partial shading amongst other reasons (Herrmann, et al., 1997). As a result, the current from the other cells can still pass through the limited cell and it becomes reverse biased. When the cell operates in reverse bias conditions, it dissipates power in the form of heat (Herrmann, et al., 1997). The heat dissipation can increase the cell temperature to above 150℃. An example of a hot spot in a solar module is illustrated in Figure 2.17. Hot spots can be detected by Thermography.

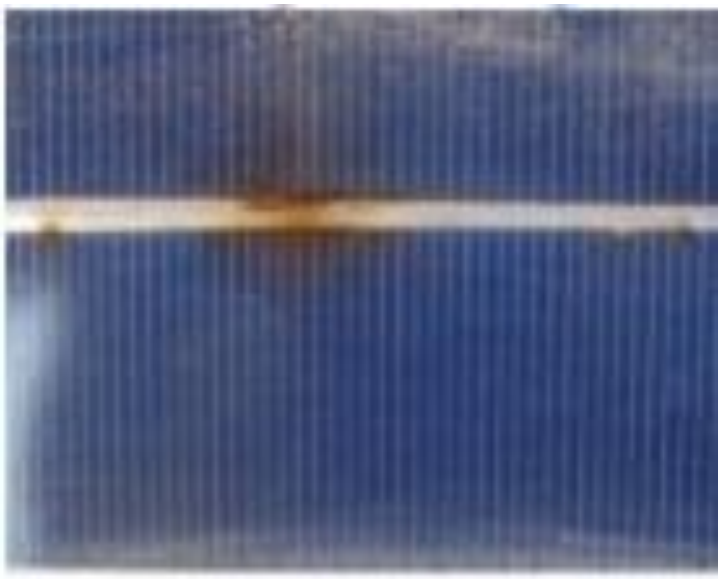


*Figure 2.17 Illustration of a hot spot on a solar module. Reused with permission (Dìez-Mediavilla, et al., 2014)*

**Thermography**

Since all objects with a temperature above absolute zero emit thermal radiation, thermography aims to detect the radiation and calculate the temperature of the object. The amount of thermal radiation is dictated by the surface and temperature of the object (Havens & Sharp, 2016). Based upon a small range of the infrared spectrum absorbed from the object, thermography can measure the temperature of the object and soundings. Thereupon an image can be created based on that information (Vollmer & Møllman, 2010). As a result, thermal images can be used to detect temperature differences in a PV module. If a module contains a cell with an increase in temperature compared to the rest of the module, a possible explanation could be that a hot spot has happened.

### Breakage and cracks

Hot spots, mechanical stress from snow, wind and during installation and transport can lead to breakage and cracks in solar cells (Dìez-Mediavilla, et al., 2014). Furthermore, the orientation of the crack is of utmost importance in the evaluation of the potential impact on power output. Cracks in parallel with the busbars are shown to be the most frequent crack orientation (Kajari-Schroder, et al., 2011). These forms of cracks are also shown to have a high potential impact on power output. In addition, cracks in parallel can accelerate cell ageing and cause other defects such as delamination and corrosion (Dìez-Mediavilla, et al., 2014) (Kajari-Schroder, et al., 2011).

*Figure 2.18 Illustration of a broken solar cell. Reused with permission (Dìez-Mediavilla, et al., 2014)*

### Degradation

Degradation of solar cells can be split into two, fast degradation and long degradation. Fast degradation of solar panels is due to exposure of light (Sopori, et al., 2012) and occurs approximately within 72 hours after light exposing. Hence it is known as Light-induced degradation (LID). The total effects from LID have an absolute value of 0.5% after an experimental test on over 25 solar cells (Sopori, et al., 2012). In contrast, long degradation has a longer time span with a typical degradation rate of 0.5% per year (Jordan, 2011). However, the degradation rate for solar modules varies between approximately 0% to 3% depending on the climate, year of manufacturing, material and mounting. As indicated in a literature study by Jordan *et al* in *Compendium of photovoltaic degradation rates*, the degradation rate has a median value across multiple studies of 0.5% per year[2] in hot and humid climate (Jordan, et al., 2016).

[2] Values for multiple measurements

### 2.8.5 Disconnected inverter

Since the inverter controls the MPP tracking of the solar modules in the connected array, periods when the inverter is disconnected result in zero produced power from the array. Disconnected inverters can occur due to a communication error between the inverter station and the grid, unstable conditions in the grid or cable faults – which will turn off the inverter for safety reasons. Periods where the inverter is disconnected can be characterised by zero power input to the inverter while the irradiance is above a certain threshold value. However, zero power input to the inverters can also happen due to PPC -curtailment.

### 2.8.6 Curtailment

**PPC -curtailment**

PPC- Curtailment is a reduction in the output power due to power plant control (PPC) interventions. When curtailment occurs the power plant controller could generate a higher amount of power given the available resources, which are typically on an involuntary basis (Bird, et al., 2014). Due to transmission congestion, lack of transmission access, excess generation during low loads periods, voltage- or interconnection issues the transmission system operator intervenes (Bird, et al., 2014).

Although there are many ways to act, a common solution is to intervene in the MPPT of the inverter. Since the goal is to reduce the DC power from the inverters, the operating point on the inverter is guided away from the MPP. PPC interventions can be found in the data due to adjustments in the active power target. Moreover, another form for curtailment is due to the system limitations to extract all the resources in a given period.

**Maximum apparent power clipping**

Maximum apparent power clipping occurs when the maximum apparent power reaches the nominal apparent power of the inverter. Situations where these kinds of limitations happen are usually in peak DC-power production periods or when the PV system consumes or generates high amounts of reactive power. As a result, there is high apparent power from the PV system which may be over the threshold value.

**Voltage Out of Range -VOoR**

Voltage out of range is considered due to its high impact in a hot climate, which are the operating conditions for many PV systems. VOoR refers to scenarios when the accumulated voltage generated from the modules are out of the operating range of the inverter. Since the inverter has manufacture specified lower and upper voltage operating range, deviation either below or above will result in suboptimal MPP tracking. As described in Chapter 2.2.1.1, an increase in cell temperature will result in lower generated voltage. Therefore, when the cell temperature is relatively high, the voltage from the array could be lower than the operating range of the inverter. As a result, the inverter is not able to track the MPP.

### 2.8.7 Soiling

Soling can occur due to the accumulation of dust, dirt, pollen and other containment on solar modules. Wind can lift dust from the ground into the air which is later dropped onto the modules. Thus, decreasing the absorbed solar irradiance which leads to a decrease in power output from the modules. An evaluation of PV systems for dry seasons in 2005 concluded with an average efficiency decline of 0.2% per day without rainfall in dry climates (Kimber, et al., 2006). However, the impact of soling is highly dependent on the properties of the dust and the local environment (Mani & Pillai, 2010). The local environment consists of site-specific factors such as surface finishes, orientation and height of the installation (Mani & Pillai, 2010). Besides, the surface type of the ground and surrounding weather conditions also has an impact on soiling levels in PV systems (Mani & Pillai, 2010).

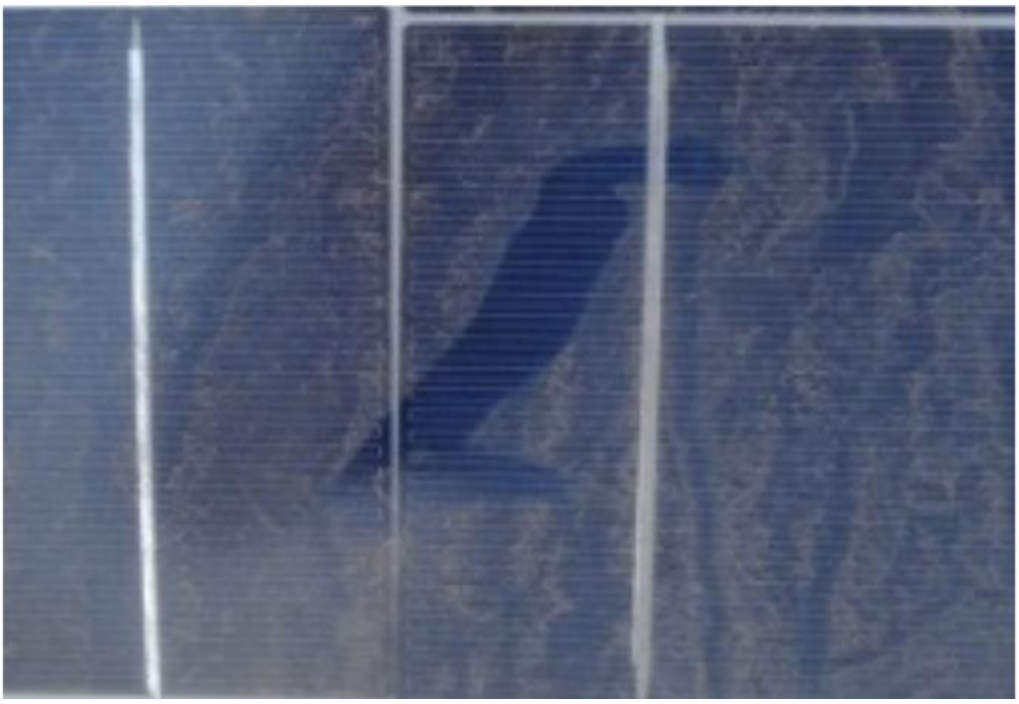


*Figure 2.19 illustration of soiling on solar cells. Reused with permission (Dìez-Mediavilla, et al., 2014)*

# 3 Event detection

This chapter discusses the methodology used within a data-driven approach to detect power losses in utility-scale PV systems. The chapter includes an introduction to the data used in this work, pre-processing and methodology.

## 3.1 Problem-solving approach

To detect power losses in utility-scale PV power plants a problem-solving methodology illustrated in Figure 3.1 and Figure 3.2, was developed and used. Key points of the proposed method are to:

1. Generate models which estimate produced DC – power, voltage and current on inverter level under normal operations. (Illustrated in Figure 3.1a).
2. Evaluate the ratio between predicted power and measured power to decide if an event is present. (Illustrated in Figure 3.1b).
3. Evaluate the ratio between estimated current and voltage with measured current and voltage to classify possible events. (Illustrated in Figure 3.2).

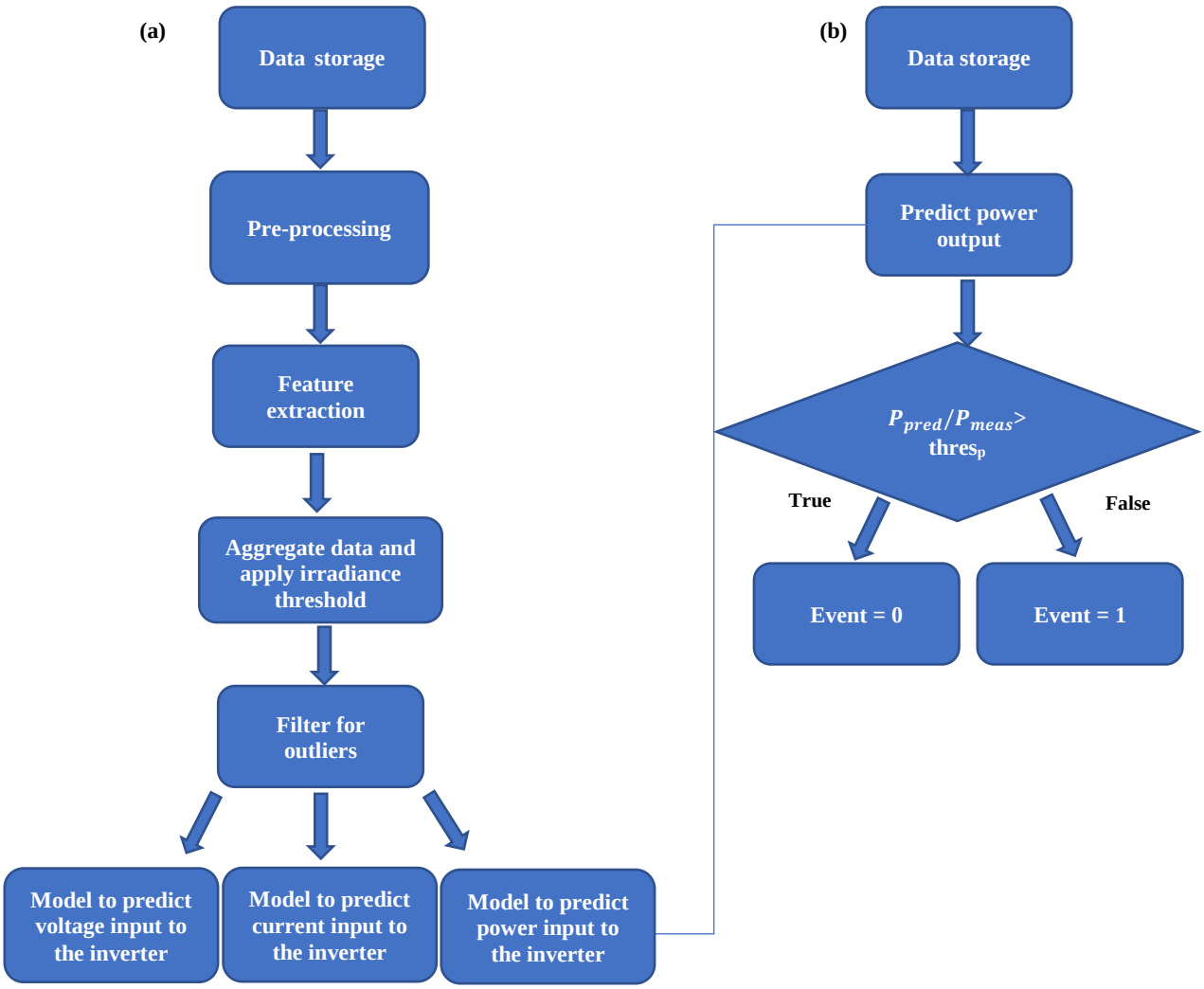



*Figure 3.1 Part 1 and 2 of the proposed method. (a) Steps included to generate a healthy evaluation of the PV system. (b) Evaluation of whether the PV system operates under normal conditions. $thres_p$ is the threshold value for event detection and is calculated as the sum of the model uncertainty and the maximum measurement uncertainty*

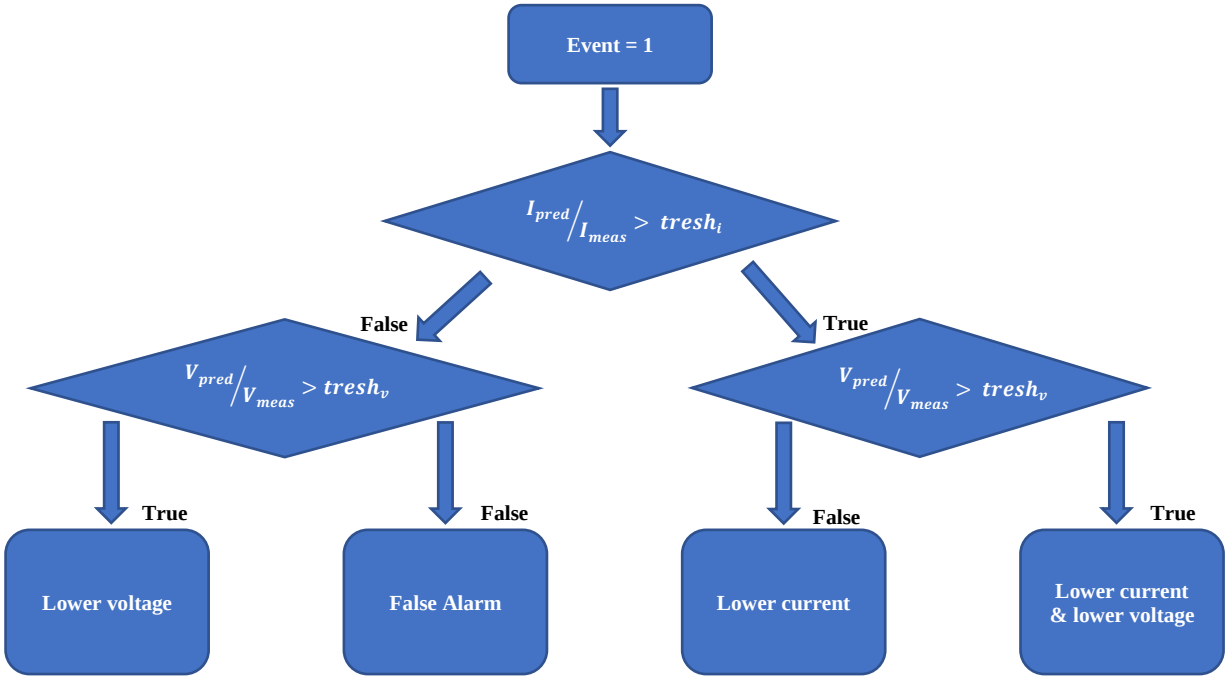



*Figure 3.2 Part 3 of the proposed method. When an event has been detected, it is then classified according to the flowchart. Flowchart adapted from (Chouder & Silvestre, 2010). In the work from Chouder & Silvestre, each classification would result in possible faults such as partial shading, faulty modules in string and faulty string*

The first part of the suggested method is to generate a representation of the PV system under *healthy conditions*. Healthy conditions represent how the PV system operates under conditions where approximately no faults are present. To represent the PV system under healthy conditions, data from the power plant is gathered and filtered out for solar irradiance above a decided threshold value. Furthermore, relevant features are extracted and created for each timestep. Subsequently, the measurements are aggregated to 1-hour intervals and filtered for abnormalities. Finally, the data is used to create models that estimate the input DC-power, voltage and current to each specific inverter. The described workflow is illustrated in Figure 3.1a). The model which estimates the input DC - power is then used to decide whether an event is present.

To indicate whether an event is present the ratio between predicted power production and measured power production are analysed. Since the predicted power production represents the PV system under healthy condition, deviations from these values can indicate the presence of an event. Therefore, the model can be used to evaluate future data to decide if an event is present at each given timestep. This workflow is illustrated in Figure 3.1b). If the ratio between predicted power output ($P_{pred}$) and measured power output ($P_{meas}$) exceeds a specific threshold ($tresh_{\mathrm{P}}$) an event has occurred.

If an event has occurred, it can be classified by analysing the measured voltage ($V_{meas}$) and current ($I_{meas}$) with the predicted voltage ($V_{pred}$) and current($I_{pred}$). The event can be classified by establishing whether the PV system experience a decrease or increase in current and voltage. This workflow is illustrated in Figure 3.2. All the threshold values in the method are calculated as the maximum measurement uncertainty plus the model uncertainty. As an example the threshold value for current is the uncertainty of the model plus the maximum measurement uncertainty for current. To generate the models and classify events data from a PV system has been used.

## 3.2 About the data

The data presented are from a utility-scale PV power plant operated in Sub-Sharan Africa in the period from 24.02.2014 to 24.02.2019. Different parameters monitoring the power plant has been stored, where the resolution is on an inverter level for most of the parameters. However, current measurements can be obtained on a string level, the setup is visualised in Figure 3.3. For weather data, such as module temperature, environment temperature, wind speed and irradiance, measurements were made at 4 different weather stations. Whereas measurements of rain, wind direction and relative humidity were only performed at one location. The time resolution can be extracted at 1-minute intervals, although this thesis uses 10-minute averages. 10-minute averages were chosen to balance computation time and the relevant information gathered. In addition, 1-minute intervals are rarely seen in the literature. Therefore, it could be difficult to compare results where different intervals are used.

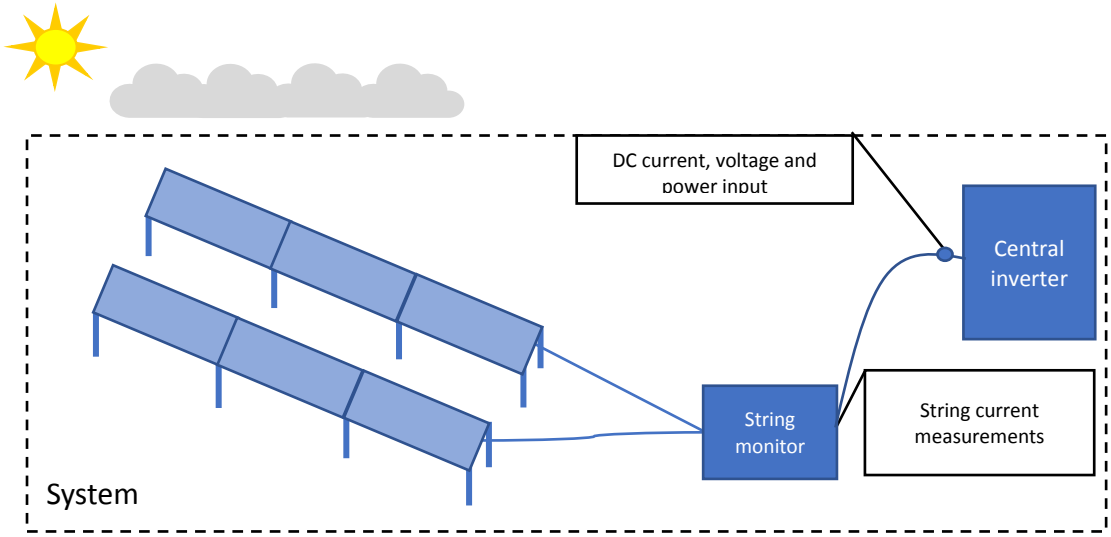



*Figure 3.3 Illustration of the monitored system. Sixteen strings are first connected pairwise into one String Monitor (SM) and then 9-10 SM's are connected to one central inverter. Although only two strings are shown in the figure, over 150 strings are connected to one inverter trough SM's.*

### 3.2.1 Site

The site in Sub-Saharan Africa includes over 300 000 modules distributed on over 80 inverter stations spread across an area of over 100 hectares. The inverter stations are located pairwise. These pairwise inverters are then connected to a transformer station which is connected to the power grid via a control station. In this thesis, 12 inverters have been picked, where 8 are located near a weather station and the remaining 4 are spread across the site. 8 of the inverters are visualised in Appendix A. For each inverter, the input DC power voltage and current are measured. In addition, values for the produced current on a string level is monitored for two strings connected a sin the figure. Over 150 strings are connected to one inverter. In each string, there are 24 modules, each consist of 60 solar cells connected in series. The modules are fixed at a tilt angle of 30 degrees facing north. Next to four of the transformer stations lies a weather station spread across the site.

The weather stations contain a solar module, pyranometer and a reference cell to gather measurements of the environment temperature and module backside temperature. From the pyranometer, information about horizontal and incline radiation is gathered. In addition, the reference cell also measures inclined radiation. An important part of the measurements obtained at the site is how reliable they are.

### 3.2.2 Uncertainties

The reliability of a measurement is connected to the uncertainty of the measuring device. The uncertainty of a measurement can be viewed as a doubt about the validity of the measurement. Whereas uncertainty is defined as a

> "*parameter, associated with the result of a measurement, that characterize the dispersion of the values that could reasonably be attributed to the measurand"*- (Joint Committee for Guides in Metrology, 2008)

Thus, the uncertainty is an interval around the measured value where the true value lies with a given probability. The concept of uncertainty is further illustrated in Figure 3.4.

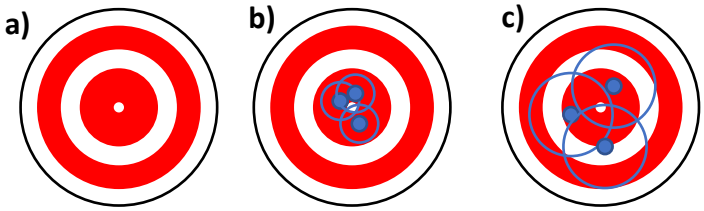



*Figure 3.4 illustration of the true value, marked as the smallest white circle and the measured values are the blue circles. Figure a) illustrates the position of the true value. Figure b illustrates the position of the true value relative to three measured values marked in blue, where their uncertainty is illustrated with a circle around the measured value. Lastly, figure c illustrates the true value relative to three measured values with a higher relative uncertainty compared to the values in figure b.*

Measurement uncertainty is highly dependent on the measurement device at hand. The name of the pyranometer located on the site is MS-802 provided by EKO. With accordance to ISO 9060:2018, the pyranometers are Secondary standard. The pyranometer has a response time of under 5 seconds for 95% of the measurements and operating temperature ranging from $-40℃\ to\ 80℃$. ( EKO Instruments , 19) Furthermore, the maximum uncertainty of the device for total hourly radiation is 3% and 2% for daily totals. (Hinckley, 2017). The pyranometer on the site measures both the horizontal and incline radiation. Although the reference cell located on the weather stations measures the same parameters there are several advantages obtained by the pyranometer.

In contrast to the reference cell, pyranometers are not specified at STC and therefore provides better accuracy under real-world conditions. In addition, the reference cell suffers more from pollution than pyranometers, which is because of the hemispherical dome of a pyranometer in contrast to the flat surface of a reference cell. Another advantage obtained from the pyrometer is that the pyranometer integrates the measurements over time. As a result, sudden changes due to birds, planes or passing of a small cloud will not create spikes or dips in the data (Kipp & Zonen, u.d.). For these reasons the measurement from the pyranometer was used and not the reference cell. Each weather station also measures a reference module temperature.

As for the uncertainty in the measurements of the module temperature, the exact value is not available, although it is assumed to be relatively higher than for the pyranometers. As well as the instrument is not calibrated at all under the measurement period. However, the uncertainties of the measurements from the specific inverter and the string monitors used are available and provided in Table 3-1. Since the power input to the inverter is calculated by multiplying the measurements of current and voltage the rule of propagation of errors is used. By assuming that the uncertainties in voltage ($\delta V$) and current ($\delta I$) are independent and random, the uncertainty in the calculated power ($\delta P$) value is (Taylor, 1997):

$$\frac{\delta P}{|P|} = \sqrt{\left(\frac{\delta V}{|V|}\right)^2 + \left(\frac{\delta I}{|I|}\right)^2} \qquad 3.1$$

*Table 3-1 Uncertainties of measurements from the inverter and string monitor*

| **parameter** | **Typical deviation** | **Maximum deviation** |
|---|---|---|
| DC current | $\pm$1.5% | $\pm$3.0% |
| DC voltage | $\pm$0.5% | $\pm$1.0% |
| DC power[3] | $\pm$1.6% | $\pm$3.2% |
| AC power | $\pm$2.0% | $\pm$3.0% |
| DC current String monitor | NA | $\pm$1.0% |

To summarize, the measurement values generated both from the inverter and the weather stations have a relatively low uncertainty, besides the module temperature where the uncertainties are not available. Furthermore, measurements from the pyranometer have been used instead of measurements from the reference cell. Due to the advantages obtained by pyranometers in comparison with the reference cell. The frequency between each measurement is 10 minutes starting from 24.02.2014 to 24.02.2019. Although the data sampling rate was 10 minutes, there are times where either the values are missing or are unrealistic. For this reason, amongst others, data pre-processing steps were used.

[3] Calculated by equation 3.1

## 3.3 Pre-processing

A common saying within the field of computer science is "*garbage in, garbage out*", emphasising the importance of informative data as input to a model. Since the models created in this thesis approximate the reality, informative data need to be provided. In any task which includes data analysis, the pre-processing step will significantly affect the results. Therefore, a detailed overview of the pre-processing steps is provided so that the results can be replicated. Although the availability of the data seems generally high, the first step was to correct for missing or unrealistic values.

### 3.3.4 Imputation

First, missing values and unrealistic values were detected. Missing and unrealistic values only applied for solar irradiance, environment temperature and module temperature amongst the used parameters in this thesis. For module temperature and environment temperature, this only occurred for one weather station at a time. Therefore, the median value from the other weather stations was used. As for the pyranometer measurements, some days had multiple unrealistic zero values in the middle of the day while the inverters produced power. Due to multiple unrealistic values on 9.05.2017 and 10.05.2017, these days were removed from the analysis. Otherwise, the median value from the other weather stations was imputed. By imputing a new value, the measurements will not be lost and can be used to evaluate the performance of the PV system. Subsequently, the continuous measurements for solar irradiance were used to evaluate the sky condition for each measurement.

### 3.3.5 Sky conditions

Secondly, although over a million parameters are monitored and stored for the specific site, none of them directly explains the amount of clouds in the sky. Since the measurements of solar irradiance are provided from weather stations located across the plant, clouds can generate noise. Some of the panels can experience less irradiance due to clouds while the weather stations are cloudless or vice versa. Therefore, to indicate whether clouds are present or not an implementation provided by PV-lib (Holmgren, et al., 2018) has been used.

The implementation is based upon the algorithm provided by Matthew J. Reno & Clifford W. Hansen (Reno & Hanse, 2016). In contrast to many other techniques for detection of clear sky periods, the algorithm provided by Reno & Hanse only requires information about global horizontal irradiance (GHI) (Reno & Hanse, 2016). The algorithm compares the measured GHI with the GHI of a selected model of clear sky solar irradiance. In the comparison, five criteria must be fulfilled for the period to be defined as clear sky. If one or more of the criteria is not fulfilled, the period will be classified as cloudy. The criteria include comparisons of mean values, maximum values, line lengths of irradiance vs time, the standard deviation of rate of change and maximum difference between changes in the two different measures of solar irradiance (Reno & Hanse, 2016). The comparison is done over a user-defined interval, in addition to user-defined thresholds for deviations in all the five criteria. Furthermore, the user can choose the clear sky irradiance model. The one chosen in this thesis is provided by Ineichen & Perez (Perez, et al., 2002) extracted from the python module PV-lib. The models estimate the clear sky irradiance based on the location and the time. The model from Ineichen & Perez was chosen since the model error has a low dependency of the day and year in contrast to other models for clear sky irradiance (Reno, et al., 2012). Also, the model is commonly used in the research community.

In the comparison between the measured irradiance and the one provided by Ineichen & Perez, the chosen window length is 1 hour. Therefore, the method evaluates six subsequent measurements to evaluate the sky condition at one timestamp. The mean difference threshold in this period is set to 30 W/m$^2$ and the difference between maximum values of 30 W/m$^2$. The maximum absolute difference in line length is set to be 100. Furthermore, the threshold value for the standard deviation of the rate of change is 7.5 Hz and the difference in changes between the two is set to be 7.5 W/m$^2$. In short, the output of the model will be boolean values of ones and zeros for each timestep. Where 1 (True) indicate clear skies and 0 (False) indicates cloudy skies. A cloudy day is visualised in Figure 3.5.

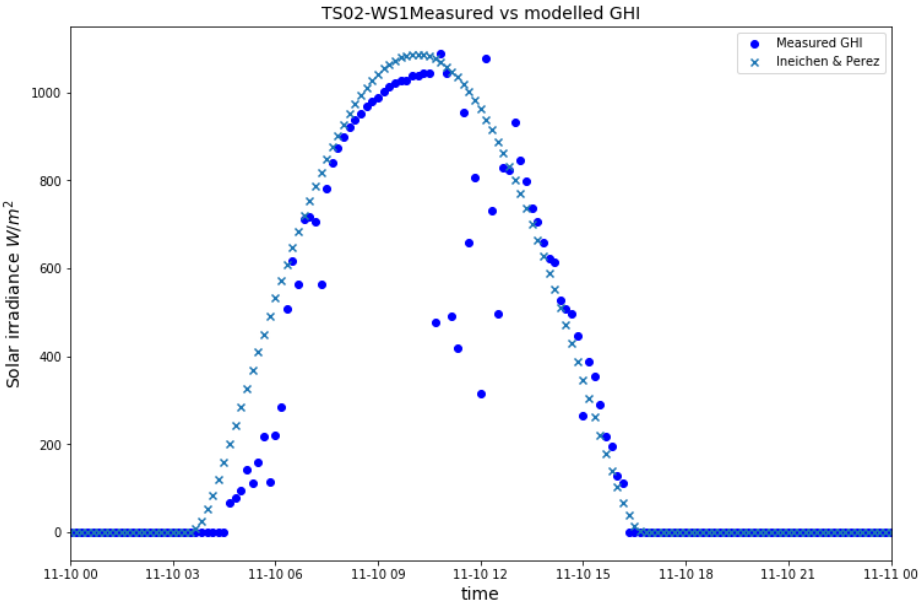



Figure 3.5 Measured vs modelled GHI. The figure is from 10.11.2014 where the points marked in x are the modelled clear sky values provided by Ineichen & Perez. Furthermore, the points marked in blue circles are the measured GHI. The y-axis is the solar irradiance and the x-axis is time labelled as month-day hour. TS02-WS1 is the name of weather station 1.

Another method to indicate the amount of clouds on the site is to compare the difference between solar irradiance from the four different weather stations on the site. The main idea is to measure the variety of solar irradiance on the site which indicates partial shadowing due to small clouds across the area. If small clouds are present, then the power input to the inverters may be lower than expected and can give further information about the conditions on the site. Two different techniques have been used, the first one calculates the standard deviation ($\sigma_{std}$) for each time step according to equation 3.2.

$$\sigma_{std} = \sqrt{\frac{\sum_{i=1}^{i=N}(X_i - \bar{X})}{N-1}} \qquad 3.2$$

In the equation, N is the total number of weather stations, $\bar{X}$ is the mean solar irradiance at the specific timestep and $X_i$ is the solar irradiance for weather station $i$. Another technique measures the deviation from the minimum and the maximum value for a given timestep across the different weather stations. Another measure of variance ($\sigma_{max/min}$) is therefore calculated in accordance with equation 3.3, $\sigma_{\max/min}$ is later referred to as *maxmin*.

$$\sigma_{maxmin} = \frac{max(X) - min(X)}{max(X)} \quad 3.3$$

Likewise, the same evaluation can be applied for the current in each string monitor connected to the inverter. The standard deviation is calculated as the standard deviation across all the currents from each string monitor into the inverter. In equation 3.2, the $\bar{X}$ will be the mean current from each string-pair, whereas $X_i$ is the current from string-pair number $i$. In this scenario, N is the number of strings-pairs. As for the technique described in equation 3.3, max(X) is the maximum current across all string-pairs for one timestep. Whereas min(X) is the minimum current across all string-pairs for the same timestep. These new features can be used to describe the variance in solar irradiance across the strings connected to the inverter. However, variance in measured solar radiation can also occur due to pyranometer drifting.

### 3.3.6 Drifting

Drifting may occur due to changes in the properties of the pyranometer. In the middle of the pyranometer, there is a black-painted ceramic disk which aims to absorb as much of the incoming photons as possible to measure the solar irradiance (Myers, 2013). The black disk inside the pyranometer dome will therefore be warmed up when radiated upon. In between the two glass domes illustrated in Figure 3.6 there is a gas which reduces the conductivity between the environment and the black disk. Thus, limiting the measurement error due to environmental temperature. Furthermore, the black disk is connected to a thermoelectric element which converts the thermal energy to electric energy (Myers, 2013). Therefore, the measurements of voltage will be proportional to the amount of radiation received by the black disk. These initial coefficients for generated voltage as a function of solar radiation were in the order of $7\mu V/Wm^{-2}$ for the pyranometer on the site. However, due to drifting these values can change over a long period of time.


*Figure 3.6 Picture of the MS802 pyranometer used at the site provided by EKO. The picture shows the black plate inside the two hemispherical domes of glass. In-between the two domes of glass there is a gas which reduces the conductivity, thus reducing the error in the measurements due to the environment temperature. Image copied by EKO with permission gained on 10.04.2019.*

Drifting occurs due to solar radiation from the Sun. When the black disk absorbs the UV-radiation, the properties of the disk can change over time (Olsen, 2019). As a result, the coefficient for generated voltage as a function of solar irradiance can change. Thus, the measurements from the pyranometer will not be stable over a long period and will drift as illustrated in Figure 3.7 To find the amount of drifting for the pyranometers on the site a method proposed by (Ødgaard, et al., 2018) was used.

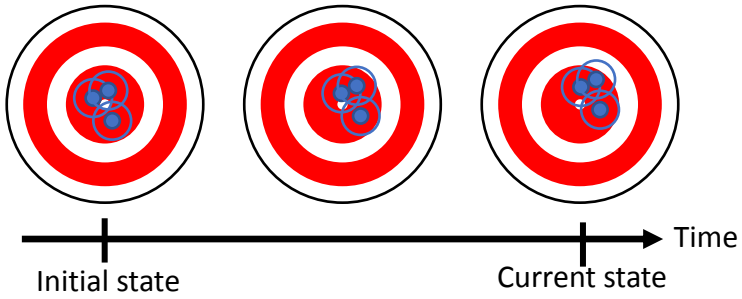



*Figure 3.7 illustration of the effects of drifting on the measured values. Figure 3.4 is further used to illustrate how the measured values deviate due to drifting. As illustrated the measured values are subject to a systematic error.*

First, timestamps that are not classified as clear sky conditions are removed according to the method proposed by (Reno & Hanse, 2016). Then the global horizontal irradiance (GHI) is estimated by the Ineichen & Perez model (Perez, et al., 2002). Furthermore, the GHI is transposed to find the estimated plane of array irradiance ($G_{POA}$) by calculating diffuse ($G_d$), reflected ($G_g$) and beam ($G_b$) irradiance in the plane according to equation 3.4 (Laboratories, Sandia National, 2018).

$$G_{POA} = G_b + G_g + G_d \tag{3.4}$$

The beam irradiance was calculated by equation 3.5

$$G_b = DNI * cos(\theta) \tag{3.5}$$

where the Direct Normal Irradiance (DNI) is also obtained from the model provided by Ineichen & Perez. The DNI is the solar irradiance received by a unit perpendicular to the sunrays. While the angle of incidence ($\theta$) is calculated with PV-lib's angle of incidence class, with values for the solar position from the *get_solarposition* class in PV-lib. To find the solar azimuth and zenith values the default method nrel numpy were used which uses an implementation of the NRL SPA algorithm (Reda, 2004). The implementation calculates the solar zenith and azimuth angles with an uncertainty of $\pm 0.0003$ degrees (Reda, 2004). The estimates are based upon the date, time and location on Earth (Reda, 2004). Furthermore, the reflected irradiance was estimated based on get_ground_diffuse implemented in PV-lib with an albedo value of 0.2 due to a combination of grass and sand on the site. The method calculates the diffuse irradiance based upon the horizontal irradiance, albedo and tilt angle. Lastly, the diffuse irradiance was also calculated with an implementation in PV-lib named *get_sky_diffuse*. The implementation calculates the diffuse irradiance based upon the surface tilt, orientation angle solar azimuth and solar zenith angles. In addition to values of solar irradiance from the model provided by Ineichen & Perez.

Finally, the relative difference ($\Delta G$) between the measured ($G_{meas}$) and modelled irradiance ($G_{POA}$) is given by (Ødgaard, et al., 2018):

$$\Delta G = (G_{meas} - G_{POA})/G_{meas} \quad 3.6$$

To evaluate how the measured solar irradiance deviates relative to the model on a year to year basis, a scaling factor ($\beta_s$) is estimated each year. The scaling factor is calculated with a least squares approach and minimizes the Root Mean Squared Error (RMSE) between the measured and clear sky irradiance (Ødgaard, et al., 2018):

$$RMSE(\beta_s) = \sqrt{\frac{\sum_{i=1}^{n}(\beta_s G_{meas} - G_{POA})^2}{n}} \quad 3.7$$

Where n is the total number of samples and $RMSE(\beta_s)$ is the root mean squared error associated with the given scaling factor $\beta_s$. The year to year drift was calculated by dividing the scaling factor from one year by the scaling factor for the previous year. If the calculated drift is over a threshold value, the pyranometers should be calibrated. However, calibration of pyranometers is usually done in accordance with the calibration interval provided by the pyranometer manufacturer.

### 3.3.7 Aggregation and solar irradiance threshold

As one of the final pre-processing steps, the measurements were aggregated to one-hour averages. This was performed to decrease the variability in the dataset. In addition, aggregating measurements have proven to have a positive effect on modelling of PV systems as shown in (De Benedetti, et al., 2018), (Betti, et al., 2017) and (Platon, et al., 2015). However, it should be emphasized that aggregating measurements to hourly averages increase the probability of losing relevant information. On the other hand, smaller time intervals for aggregating measurements can increase the probability of including noise in the measurements. Therefore, one should experiment with the size of the time interval of the aggregated averages before deployment of methods which uses aggregated data. Due to common practice within the field on PV-analysis one-hour averages were chosen in this work. Another common practice is to introduce a cut-off value for solar irradiance.

Lastly, due to significantly low measurement accuracy for the pyranometers at low irradiance values (Myers, 2013), measurements when the solar irradiance was less than $50 W/m^2$ were removed. This cut off value applies for the solar irradiance which is used for the specific analysis. In some scenarios, the solar irradiance is measured from the weather stations closest to the inverter and in others, it is the average irradiance across all the weather stations. In the analysis other parameters are used in addition to the solar irradiance, some are measured while others are generated through feature engineering.

## 3.4 Feature engineering

To provide necessary information to the models *feature engineering* is used. Feature engineering is the process of applying hardcoded transformations to the data (Chollet, 2018). The typical benefits of feature engineering are to increase the feature space and excel the learning process of a model. As a result, the predictive power of the model can increase. As indicated by equation 2.8, one of the most useful parameters when analysing the performance of a solar cell is the cell temperature.

### 3.4.4 Cell temperature

The four different weather stations on the site measure the solar modules back-surface temperatures. Therefore, to estimate the solar cell temperature equation 3.8 is applied. The equation assumes one-dimensional thermal heat conduction through the module materials behind the cell. The cell temperature $(T_c)$ can, therefore, be calculated using the measured back-surface temperature $(T_m)$ in addition to a predetermined temperature difference between the cell and the back surface as shown in equation 3.8 (King, et al., 2004)

$$T_c = T_m + \frac{G_{poa}}{1000\ W/m^2} \cdot \Delta T \qquad 3.8$$

Where $G_{poa}$ is the measured solar irradiance on the module and $\Delta T$ is the temperature difference between the cell and the module back surface at an irradiance level of $1000 w/m^2$. The value for $\Delta T$ has been found empirically in (King, et al., 2004) to be $3°C$ for open rack mounted glass-cell-glass and glass-cell-polymer sheet modules. In the calculations, the median module temperature from all weather stations is used. Whereas the solar irradiance is from either the closest weather station or the average across all depending on the inverter. Since the cell temperature is a linear combination of other parameters used it will not increase the feature space. However, the parameter is a useful input into the physical models. In addition to the cell temperature, features which contain information about the season and time of day are generated.

### 3.4.5 Time-dependent features

To provide further information to the models, multiple features were generated to indicates the time of the measurement. For each measurement, the day of the year was expressed as sin(day) and cos(day) transformations as follows:

$$\sin\left(\frac{dayofyear \times 2 \times \pi}{365}\right),\ \cos\left(\frac{dayofyear \times 2 \times \pi}{365}\right)$$

As a result, two new features were generated. Similar features were also proposed by Mike Green and Eyal Brill in their report "*Fault Prediction Using Clustering Algorithms*" (Green & Eyal, 2017). In addition, the hour of the day of each measurement is also added as a new feature. Lastly, the time in seconds since the first measurement was generated.

## 3.5 Outlier detection

A crucial point of the proposed method is to represent a healthy version of the PV system. To create such a model measurement representing an unhealthy PV plant must be filtered out. An unhealthy PV plant operates under abnormal conditions where one or more faults are present. As PV plants are assumed to mainly operates under healthy conditions, deviation from normal operations can represent an unhealthy system. Such deviations are therefore anomalies or outliers. According to Douglas M Hawkins, an intuitive definition of outliers is

> "*an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*" (Hawkins, 1980).

In accordance with the definition provided by Hawkins, an outlier would be a measurement that compared with other measurement seems likely to be generated by a different mechanism. Since PV systems mainly operate under healthy conditions, the "*different mechanism*" could be under faulty operations. To find outliers for each inverter or string multiple different techniques have been tested with varying degree of success, such as *distance to the K-nearest neighbour, residuals from Ordinary Least Squares (OLS), Isolation forest* and *Local outlier factor.* The techniques which had a high degree of outlier detection will be discussed further.

### 3.5.4 Distance to K-nearest neighbour.

When measurements are placed in a hyperdimensional space the distance to the neighbouring measurements can indicate whether the measurement is an outlier (Angiulli, 2002). The distance can be calculated according to the Euclidean distance, that is Minkowski distance with p = 2;

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2} \quad 3.9$$

Where D is the distance matrix, x and y are two separate measurements where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ drawn from the data in X and Y. Since multiple outliers could be located relatively close and therefore be indicated as inliers, the average distance between the *k*-nearest neighbours was used. Where *k* is a hyperparameter which is tuned to the specific use-case and dictates the number of neighbours to evaluate. To evaluate whether an observation is an outlier the mean, max or min distance to the K-nearest neighbours could be used to evaluate outlier degree. Features used in this technique for outlier detection was the produced power, current or voltage input for the specific inverter and the solar irradiance from the nearest weather stations or the average across the weather stations. To compensate for the different range between the input to the inverter and solar irradiance standardisation was performed. For example, the power input ranges from $0 - 0.86\ MW$ while solar irradiance has a range from $0 - {\sim}1160 W/m^2$. As a result, a variation in power, voltage, current and solar irradiance will have equal weight in the calculations of the Euclidean distance. The python module PYOD has been utilized for KNN outlier detection (Zhao, et al., 2019).

### 3.5.5 Residuals from Ordinary least squares

Another method for outlier detection utilises the relationship between solar irradiance and produced power has been introduced in (Betti, et al., 2017). In this case, the authors applied an Ordinary Least Squares (OLS) regression to predict the produced power based on solar irradiance for the given measurement. Furthermore, the model was used to detect outliers according to the furthest points from fitting. Outliers were classified based on the following condition:

$$\left|\mathrm{P} - \mathrm{G_{poa}} * \mathrm{m} + \mathrm{b}\right| > \mathrm{thr} * \left(\mathrm{G_{poa}} * \mathrm{m} + \mathrm{b}\right) \qquad 3.11a$$

Where P is the measured power, $\mathrm{G_{poa}}$ is the plane of array solar irradiance, while m and b are computed by an OLS approach and are respectively, the slope and the interception of the model. $thr$ is the desired threshold (Betti, et al., 2017). To get an indication of the degree of which an observation is an outlier, the technique was modified as shown in 3.11b.

$$Outlier\ degree\ = |(P\ -\ P_{pred})|\ /\ P_{pred} \qquad 3.11b$$

From 3.11a $\mathrm{G_{poa}} * m + b$ is replaced with $P_{pred}$. The outlier degree value should be low in absolute value for inliers and the absolute value increases as the measurements deviate from normal operation. After outliers were filtered out from the dataset, the data were then used as input to models to estimate the produced power, current and voltage at that specific time interval.

## 3.6 Prediction of power production, generated voltage and generated current

After previous pre-processing, feature engineering and outlier rejection, the data is ready to be implemented into the models presented in Chapter 2.6 and 2.7. The physical baseline model is based upon the theory presented in 2.6 The temperature coefficient for the modules on the site is presented in Table 3-2.

*Table 3-2 temperature coefficient for the modules used on the site*

| **parameter** | **Value** |
|---|---|
| $\gamma_P$ | $-0.47\%/°C$ |
| $\gamma_I$ | $0.045\%/°C$ |

By comparison of $\gamma_P$ and $\gamma_I$ the power drops relatively faster than the increase in current for high cell temperatures. As a result, the theory represented in 2.2.1.1 holds. Furthermore, from the Machine learning models represented in 2.7.3, the following *hyperparameters* were used as part of the model evaluation. In contrast to regular parameters, such as regression coefficient, hyperparameters are optimised separately (Rachka & Mirjalili, 2017). Examples of hyperparameters are the regularisation parameter in a Ridge and Lasso regression or the k-number of neighbours in *K-Nearest Neighbours* (Rachka & Mirjalili, 2017). The optimisation of hyperparameters is often performed as a part of the model evaluation. Where the hyperparameters corresponding to the best model in the evaluation process are the most optimal ones.

*Table 3-3 Hyperparameters for Random Forest regressor*

| **parameter** | **Values** |
|---|---|
| Number of estimators ($B$) | [5 10 50 100 200 300 400] |
| Maximum depth | [5 10 25 50 100] |

*Table 3-4 Hyperparameters for K-Nearest neighbours regressor*

| **parameter** | **Values** |
|---|---|
| Number of neighbours ($k$) | [5 10 25 50 100 200] |
| Weights | [‘Uniform’, ‘distance’] |

*Table 3-5 Hyperparameters for Lasso and Ridge regression*

| **parameter** | **Values** |
|---|---|
| Alpha $(1/\lambda)$ | [0.0001 0.001 0.01 10 100 1000 ] |

## 3.7 Model evaluation

The data will be used to establish a framework to detect and classify events on utility-scale PV systems. To ensure that the models and methods *generalise* the data will be split into training, validation and test sets. Generalisation means that the specified models or methods will perform well on unseen data (Chollet, 2018).

### 3.7.4 Validation

**Simple hold out validation**

To test the method, data from 24.02.2017 to 24.02.2019 has been held out of model training. Such an approach is known as *Simple hold-out validation* and consists of setting aside some fraction of the data as test data. Subsequently, the data that were not left out will be used to train the models and choose the appropriate thresholds in the method. The concept of simple hold-out validation is visualised in Figure 3.8.

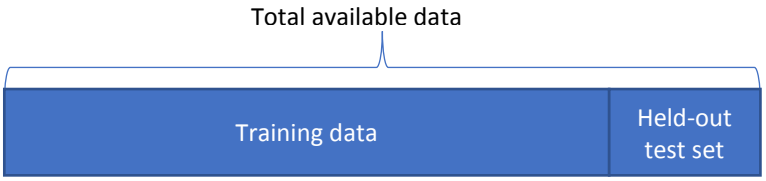



*Figure 3.8 visualisation of the simple hold out method. Test data is left out of the training and tuning of the algorithm to prevent information leakage, to simulate how the method would behave on unseen data. Information leakage happens due to the tuning of the algorithm according to the held-out data. As a result, some information about the held-out data leaks into the algorithm.*

Therefore, the training data will consist of measurements that will be used to train the different models and their respective hyperparameters in addition to estimating the model error. Furthermore, to tune the hyperparameters *K-fold cross-validation* has been applied.

**K-fold validation**

Different models applied to the training data can be optimised by *K-fold* cross-validation. K-fold cross-validation is a technique regularly used to evaluate the performance of mathematical or statistical models. The approach consists of splitting the data into K partitions of equal size (Chollet, 2018). For partition number *i*, where *i* is in the range of 1 to K, the model is validated on partition *i*, while trained on the remaining K – 1 partition. The method is illustrated in Figure 3.9 and can be summarised by the following steps:

1. Shuffle the data
2. Partition the shuffled data into K folds
3. For i in 1: K
    a. Train the model on all the folds except fold i
    b. Validate the model on fold i
4. The performance of the model is the average score across the different validation folds

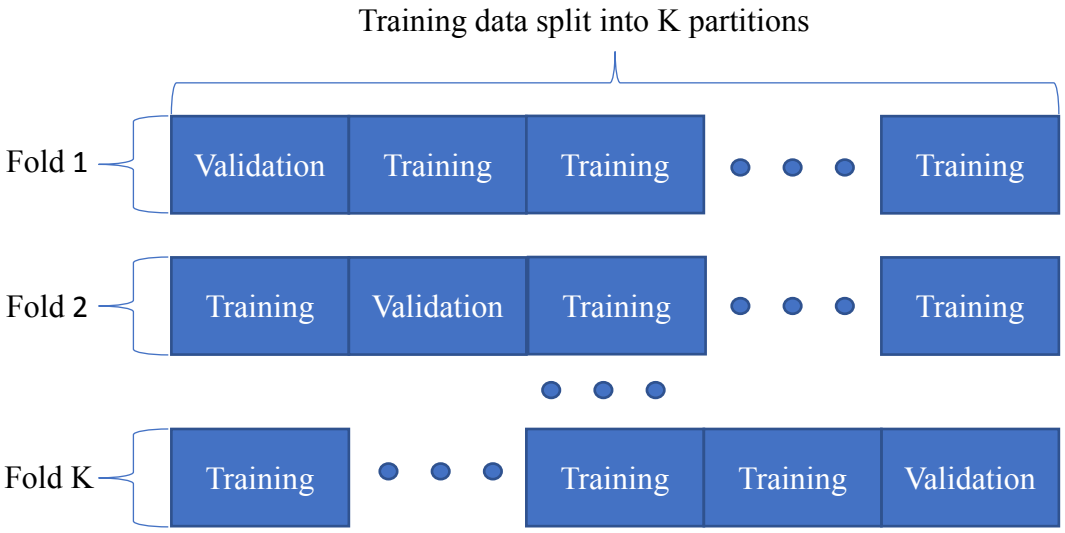



*Figure 3.9 Visualisation of K-fold cross-validation. To ensure that seasonal and yearly variation does not influence the predictive power of the model in each fold, the data is shuffled before the k fold validation.*

K-fold cross-validation will be applied to all different combinations of hyperparameters and features to search for the optimal model. As a result, the combination of hyperparameters which has the best average score from K-fold validation will be the best combination of hyperparameters for the specific model applied to the data. However, the K-fold cross-validation error estimate has shown to give a significantly biased estimate of the true error (Varma & Richard, 2006) (Cawley & Talbot, 2011). Thus, nested cross-validation has been used to compare the prediction error between different models.

**Nested cross-validation**

Nested cross-validation consists of applying the K-fold cross-validation search for optimum hyperparameters and features as part of the model training in addition to evaluate the performance on new unseen data (Varma & Richard, 2006). In other words, the model will be tuned on a specific set of data and subsequently be tested at new unseen data to evaluate the performance of the model over multiple folds. In principle, this can be achieved by applying K-fold validation whereas a search for optimum hyperparameters and features with a new K-fold validation is done on the training data. Therefore, the inner loop, which consists of the model, K-fold validation and hyperparameter and feature tuning can be viewed as a wrapper algorithm around the model. Next, a K-fold validation is applied to the wrapper algorithm. Nested cross-validation is visualised in Figure 3.10 and can be summarised by the following steps:

1. Shuffle the data
2. Partition the shuffled data into K folds
3. For k in 1: K
    a. Train the wrapper on all the folds except fold k. The training will consist of K-fold cross-validation on the training data, where the search for optimal hyperparameters and features takes place.
    b. Train the optimal model on all folds expect fold k.
    c. Validate the model on fold k
4. The performance of the model is the average score across the different validation folds

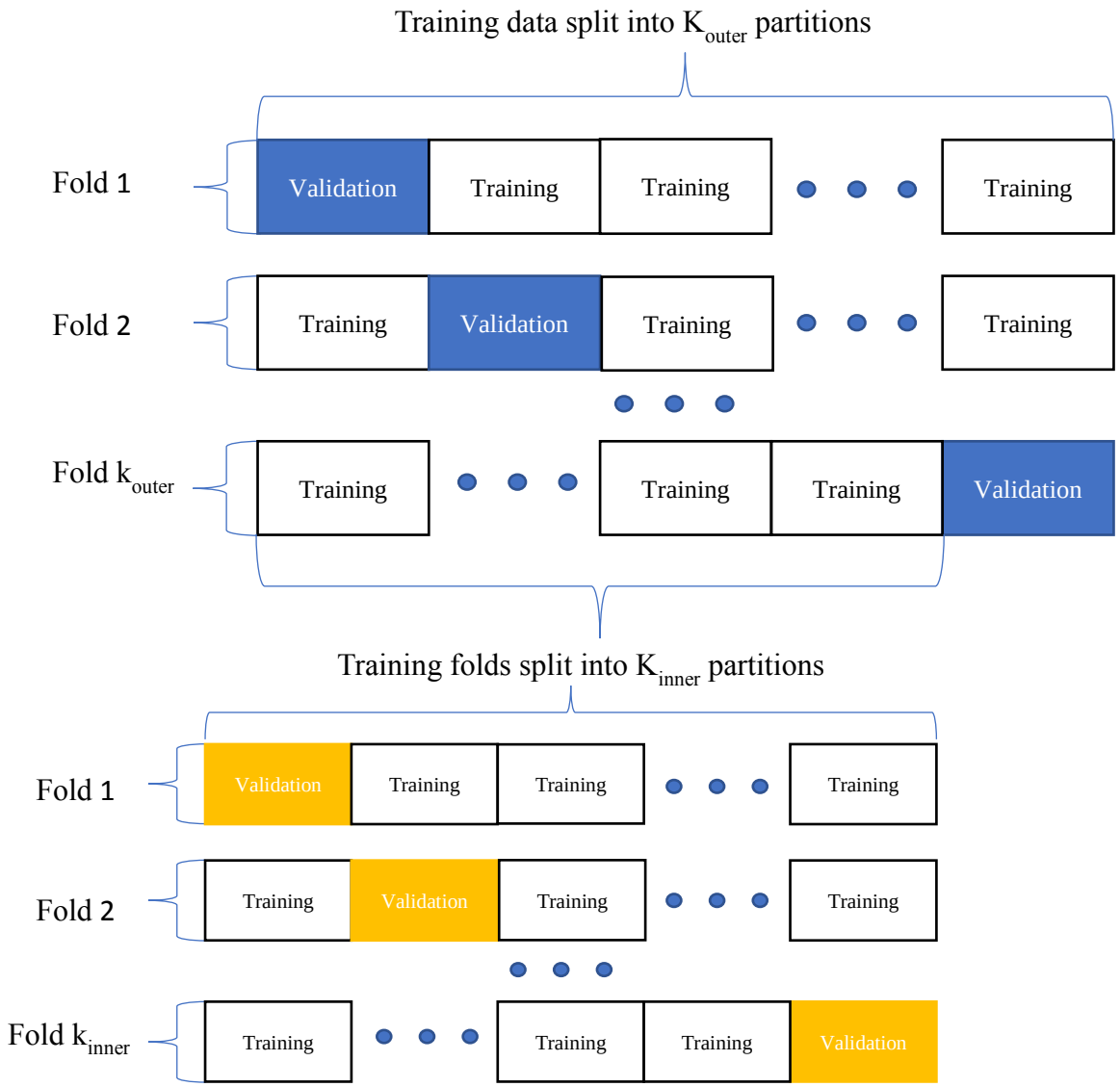



*Figure 3.10 illustration of the methodology behind nested cross-validation. There is multiple K-fold cross-validation, one in the outer loop, where the best model from the inner cross-validation is validated. In addition, the models hyperparameters and features are tuned in the inner loop.*

To sum up, data from 24.02.2017 to 24.02.2019 will be used to test the whole methodology, whereas data from 24.02.2014 to 24.02.2017 will be used to train the proposed model and evaluate the model uncertainty. The models will be evaluated by nested cross-validation since it has been proven to obtain a more accurate estimation of true model error. Furthermore, K-fold cross-validation will be used to obtain the best hyperparameters from the highest scoring model. The evaluation of the different models or the best hyperparameters is done by comparing the results obtained from a decided performance metric.

### 3.7.5 Performance metric

In the evaluation of regression models, there is a variety of scoring metric which can be used. The one used in this thesis is mean absolute error (MAE);

$$MAE(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad 3.16$$

Where y is the true value, $\hat{y}$ is the predicted value and $N$ are the total number of predicted samples. Although there is a multiple of different evaluation techniques, such as mean squared error, median absolute error and $R^2$ score, MAE was chosen since its value is highly relatable to the real-world application. A summary of the model evaluation process is further described as a *trial run*.

### 3.7.6 Trial run

The trial run is a walk-through of the different steps in the evaluation process and is included due to the numerous steps in the evaluation. The data is split into five folds, visualised in the outermost K-fold cross-validation ($K_{outer}$) in Figure 3.11. This fold is where the wrapper model is trained on all but one fold at a time, whereas the cross-validated model score is the average score from all folds. Furthermore, the training process consists of a search for optimal hyperparameters and their corresponding best features. The search is done by K-fold cross-validation ($K_{inner}$) for all the different hyperparameters where the most optimal model has the best average score across all folds. However, the most optimal features are found by applying SFS with K-fold cross-validation at the training data in the $K_{inner}$ fold, marked as $K_{SFS}$.

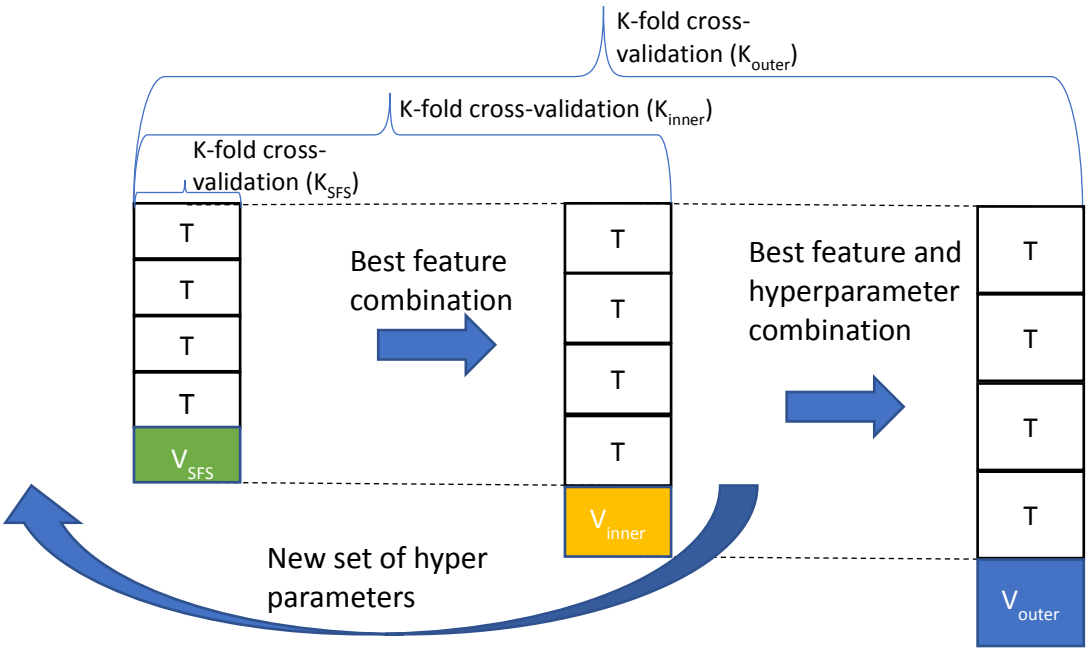



*Figure 3.11 Trial run. A visualisation of the model evaluation process. From the left the model is trained according to the SFS algorithm and validated on the green validation set. Subsequently the best feature combination is found, whereas the model is trained with these features and evaluated on a new set of data (yellow validation set). The combination of hyperparameters and features with best average score from this search is thereupon used in the outer cross-validation and validated on the blue validation set.*

### 3.7.7 Model uncertainty

After the best model is found an evaluation of the uncertainty in upcoming results is determined based upon K-fold validation on data from 24.02.2014 to 24.02.2017. In this scenario, the best performing model is trained on four folds and tested on a final fold. When the model has been tested on all folds unbiased predictions in the period are made. Then the relative absolute error was found for each measurement $(i)$ by equation 3.17, where the mean absolute error of all measurements was used as the mean model error $(\bar{\varepsilon})$.

$$\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3.17}$$

Thereupon, the standard error $(\bar{\sigma_{\varepsilon}})$ was found from the same measurements as described in equation 3.18 (Taylor, 1997).

$$\sigma_{\bar{\varepsilon}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}}$$

$$\bar{\sigma_{\bar{\varepsilon}}} = \frac{\sigma_{\bar{\varepsilon}}}{\sqrt{N}} \tag{3.18}$$

In the equation $\bar{\sigma_{\bar{\varepsilon}}}$ is the uncertainty in the model error. The model uncertainty can be estimated as

$$\varepsilon = \bar{\varepsilon} \pm z\bar{\sigma_{\bar{\varepsilon}}} \tag{3.19}$$

Where z is based upon the confidence interval chosen and the observed distribution. If the model error is normally distributed, then 99.7% of the samples will lie within 3 $\bar{\sigma_{\bar{\varepsilon}}}$ (Montgomery & Runger, 2014). Although many samples are (or approximately) normally distrusted for large N, this is not necessarily the situation. However, if the data does not follow a normal distribution at least 88.9% of the samples will lie within 3 $\bar{\sigma_{\bar{\varepsilon}}}$ according to Chebyshev's inequality (Montgomery & Runger, 2014). Therefore, the model uncertainty can be estimated to be equal to the equation in 3.19 with a Z value of 3 (De Benedetti, et al., 2018).

# 4 Results & discussion

This Chapter will present the results based upon the methodology of event detection presented in Chapter 3 applied onto 12 inverters and discuss the results. The inverters have a name from 1-12, where the chosen inverters are pairwise close to the same weather stations which applies for inverter 1-2, 3-4, 5-6, 7-8. Therefore, these inverters will use solar irradiance from the closest weather station. However, inverter 9-12 are randomly distributed across the site and will use the average irradiance from all weather stations as a measure of solar irradiance. A crucial point in the methodology is to attain a reliable evaluation of the PV system visualised in Figure 3.3 (Chapter 3.2). Thus, an important part is to obtain information describing the environmental conditions at the site. Although solar irradiance, humidity, rain intensity and module temperature can be measured, other environmental conditions need to be engineered. As seen in Figure 3.3 (Chapter 3.2), clouds can be a part of the environmental conditions.

## 4.1 Sky conditions

The sky conditions at the site are evaluated with different techniques described in Chapter 3.3.2. The first technique is the detection of clear skies and is applied to 10-minute data. Subsequently, the values are aggregated to 1-hour averages. As a result, the values are in the range of 0 (clouds) to 1 (clear sky). Furthermore, a visual evaluation can be made by inspecting the Power- Irradiance plot illustrated in Figure 4.1.

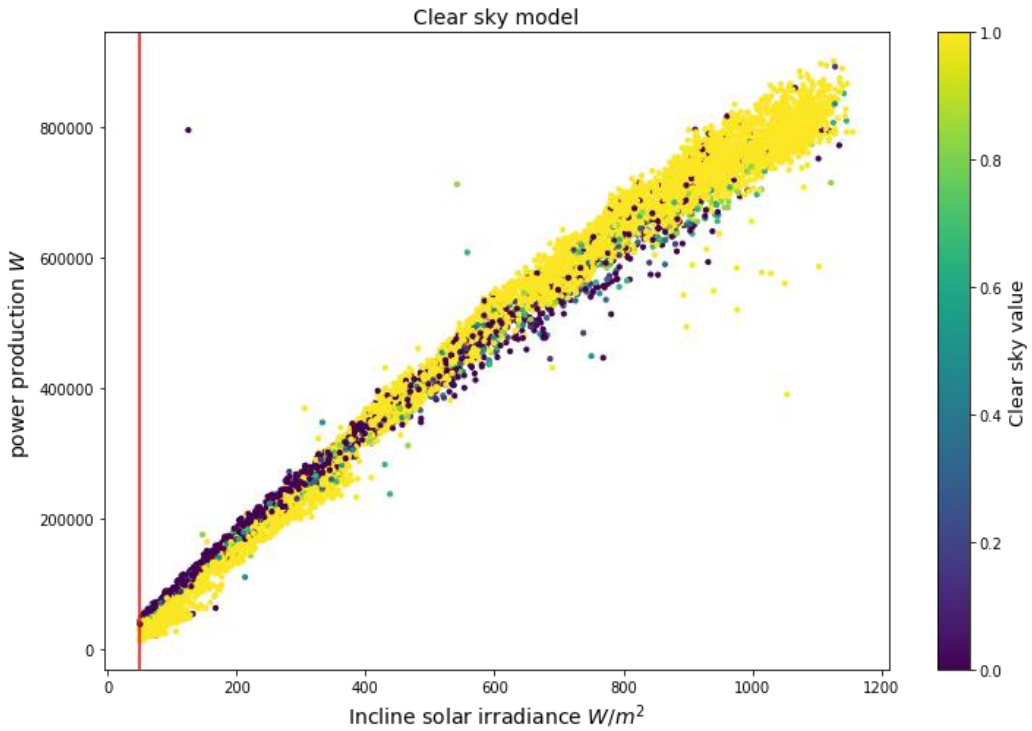



*Figure 4.1 Detect clear sky model illustrated by the color of hourly power-irradiance data of an inverter which is near a weather station. The x-axis is the incline solar irradiance [W/m$^2$], the y-axis is the power production [W] and the color is the value of the clear sky detection. The figure is based upon data from 24.02.2014 to 24.02.2016 and disconnected inverter faults are filtered out. The red line illustrates the irradiance threshold at 50W/m$^2$.*

Furthermore, it is also interesting to see how the algorithm perform on cloudy days, as illustrated in Figure 3.5 (Chapter 3.3). Therefore, Figure 4.2 illustrates the algorithm applied to the same data as the one represented in Figure 3.5 (Chapter 3.3).

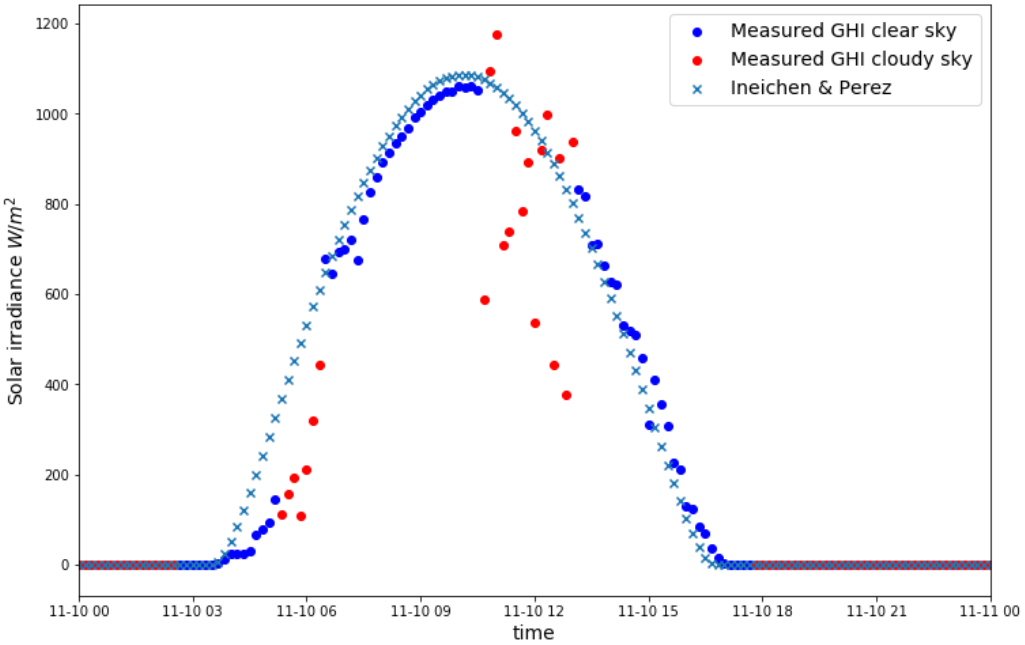



*Figure 4.2 Measured vs modelled GHI. The figure illustrates samples labelled as clear sky marked in blue and samples which are labelled as cloudy marked in red, the points marked with x are the modelled clear sky values provided by the Ineichen & Perez algorithm.*

The figure shows that the detection technique properly detects measurements as cloudy if there is a certain deviation from the measured solar irradiance. In addition, there are some points around 13:00 which has a low deviation although they are detected as cloudy. The algorithm has assessed these points as cloudy due to the deviations in the subsequent measurements. Furthermore, to evaluate how clouds can affect the whole site four additionally techniques were applied.

In addition to the detection of clear skies, four other methods were used to indicate the sky condition for a specific time. All of these were also evaluated in the power-irradiance curve illustrated in Figure 4.3. The figure illustrates the Power-irradiance plot for an inverter station, where the irradiance is measured from the closest weather station. The colour values indicate the amount of clouds in the sky. Furthermore, the plots to the left have been calculated with the standard deviation in equation 3.3, whereas the figures to the right have been calculated with equation 3.4. Besides, the upper figures have used current measurements and the lower figures have used solar irradiance measurements to indicate the sky conditions.

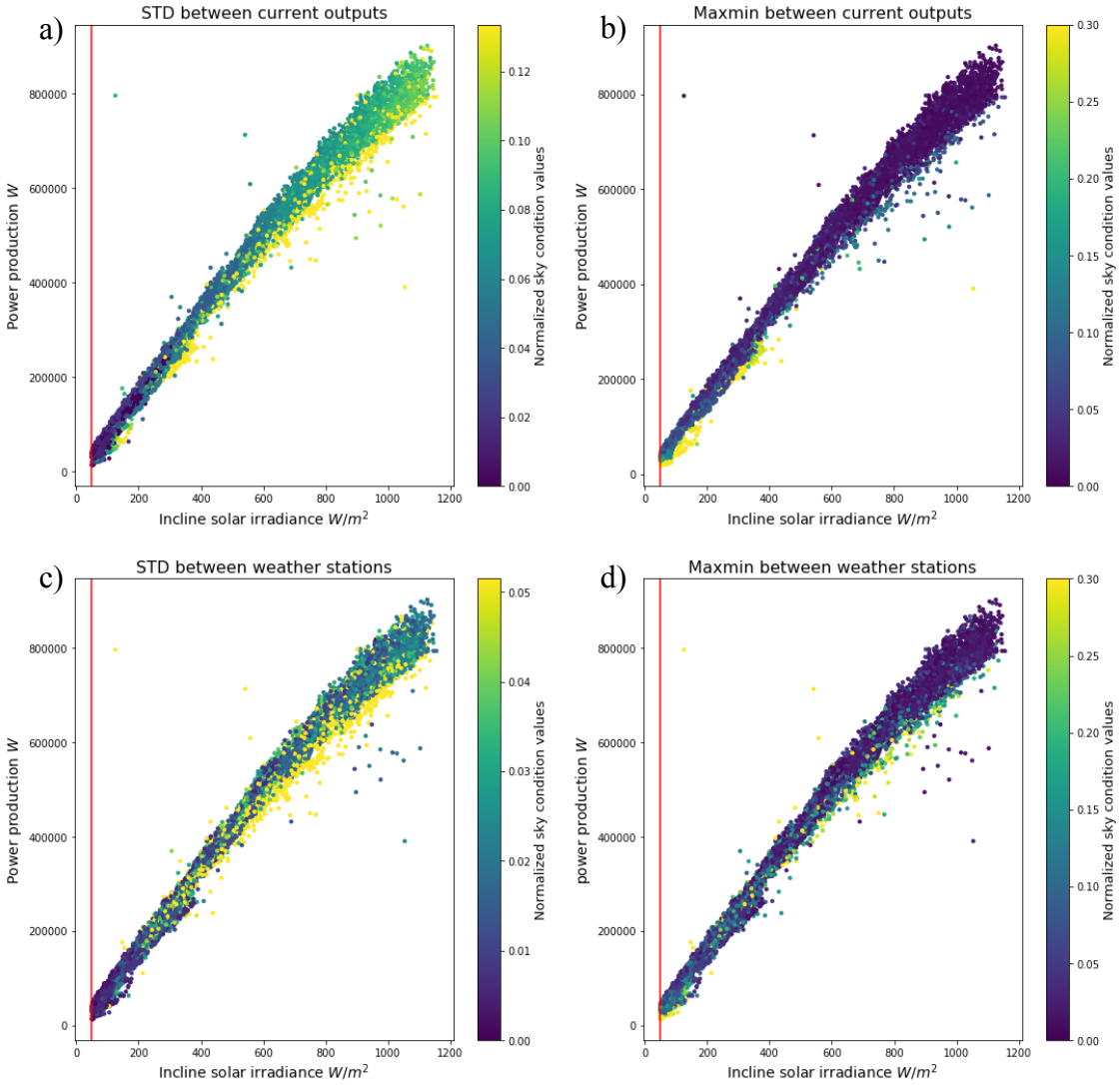



*Figure 4.3 Illustration of power-irradiance relationship for a specific inverter when four of the techniques are applied to measure the sky conditions. The X-axis is the solar irradiance measured from the closest weather stations, the y-axis is the DC-power input to the inverter and the colour of each point represent the amount of clouds in the sky. Dark blue indicates clear skies and yellow indicates cloudy conditions. The figure is based upon data from 24.02.2014 to 24.02.2016 and disconnected inverter faults are filtered out. The red line visualises the solar irradiance threshold value of 50/m².*

By comparison of all inverters, there are some differences between the power-irradiance curves for evaluations made with the nearest weather station in contrast to the average between all the weather stations. The difference is more variation between the solar irradiance and power production for inverters with irradiance values as the average from all the weather stations at the site. Moreover, the values from the four different methods applied to a cloudy day are shown in Figure 4.3. As can be seen from the figure, all the measurements of sky conditions react at the time around 12 and 9. Although one would expect near-zero values for the maxmin irradiance at night-time, the offset for each pyranometer is different which is the reason for the non-zero value. The offset value is in the range $[1.8 \cdot 10^{-6} W/m^{2}, 6 \cdot 10^{-7} W/m^{2}]$. The offset value is also a motivation for the solar irradiance threshold value, thus eliminating error in maxmin irradiance.

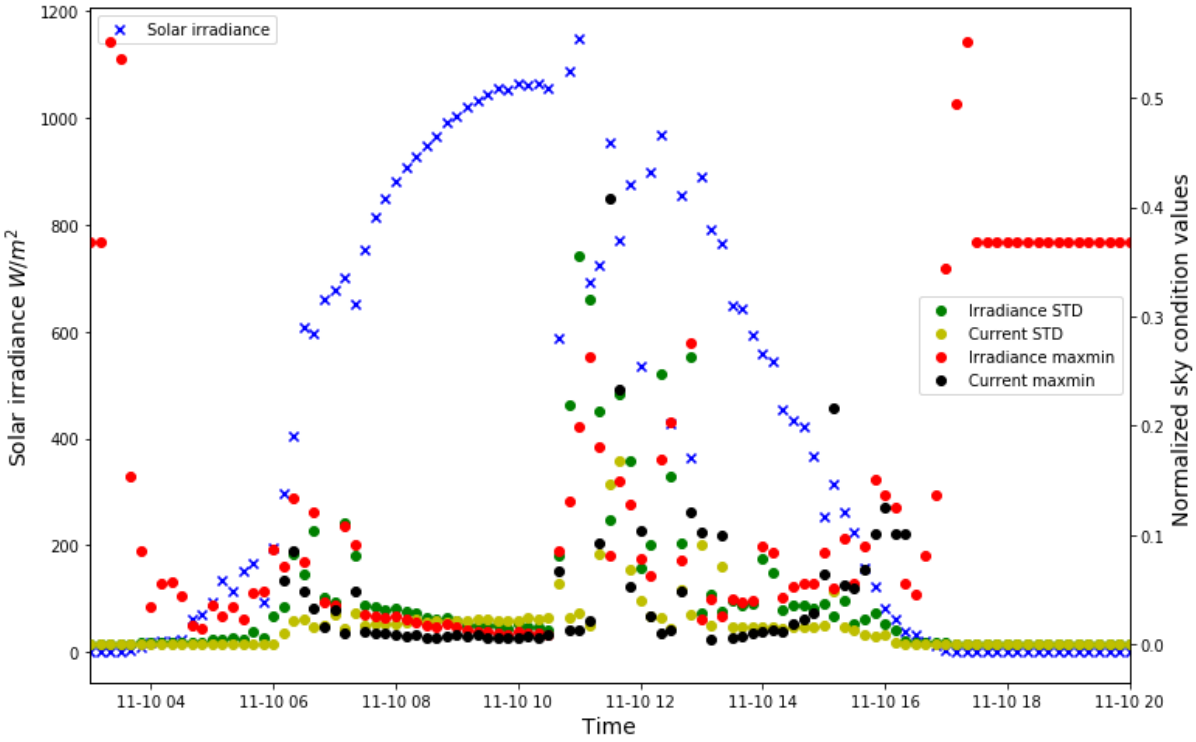



*Figure 4.4 Values for the four different methods to evaluate sky conditions applied on the date of 10.11.2014. In addition, the solar irradiance is also illustrated marked in blue x's. All values for sky conditions has been normalized before visualisation. The x-axis is the time, the first y-axis is the solar irradiance from the closest weather station, furthermore the second y-axis the normalized values for sky conditions. STD stands for standard deviation.*

In essence, five different methods to evaluate the sky conditions have been applied to the data and visually evaluated against the power-irradiance plot. In addition, all the methods were evaluated on a cloudy day in November 2014. The result is that the methods react to changes in the irradiance curve that could suggest cloudy conditions illustrated in Figure 4.2 and Figure 4.4. Furthermore, many of the lower points on the power-irradiance curve in Figure 4.1 and Figure 4.3 has been evaluated as cloudy with a yellow colour value. Although, for weather stations who use the average solar irradiance much of the variation in power-irradiance could be due to variation in the solar irradiance across the site. With over 300 000 solar modules in the park, spread across over 100 hectares, the mean value of solar irradiance from 4 weather stations does not reflect the variation in irradiance across the park. This is the main reason for evaluating the sky conditions at the site.

However, the methods which involve evaluation of the current in each string can be biased towards faults. These methods can evaluate a period as cloudy due to a faulty string. Thus, methods which involve evaluation of the string current are left out due to this bias. In addition, as can be seen from Figure 3.3, the current measurements are within the system to be evaluated. Therefore, it seems likely that these evaluations would navigate the models towards an unhealthy representation of the PV system. This can be done by generating a relationship between high standard deviation in current with lower current and power input to the inverter. Subsequently, after the clear sky periods have been detected, an evaluation of drifting within the pyranometers was performed.

## 4.2 Pyranometer drifting

Pyranometer drifting was found by the methods described in Chapter 3.3.6, where the relative difference between measured irradiance and calculated irradiance is illustrated in Figure 4.5. Due to lower residuals between the measured irradiance and calculated irradiance at higher irradiances, irradiance values of above 900 W/m$^2$ were used. In addition, only measurements where the detect clear sky value is 1 was used.

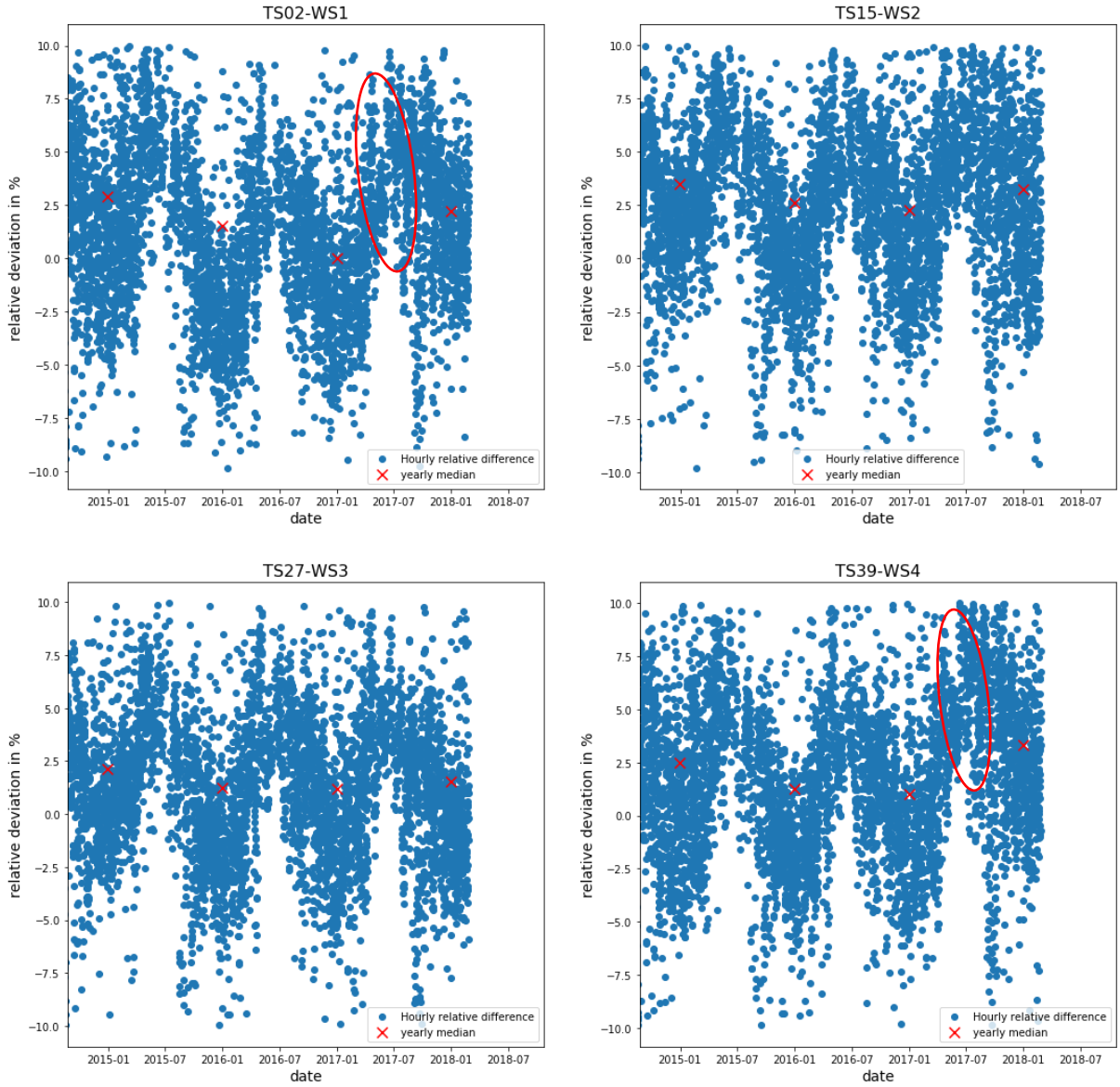



*Figure 4.5 Illustration of the relative difference between measured incline solar irradiance and calculated incline solar irradiance calculated by equation 3.6 multiplied with 100%. In the figure relative deviations with an absolute value above 10% has been filtered out. The measured incline solar irradiance is found by equation 3.4. Red ellipses represent periods of sudden change in the relative deviations.*

It seems to be a steady decrease in the period between 24.02.2014 to around mid-2017. As a result, the drift from 2014 to 2016 was calculated in accordance with equation 3.7 for each year. Thereupon, the coefficient from one year was divided by the coefficient from the previous year. Finally, the average of these values was used as the drift from each weather station presented in Table 4-1.

*Table 4-1: Yearly drift values in the incline irradiance sensors*

| Station | Calculated | Calibration report |
|---|---|---|
| WS1 | -1.1% | -0.03% |
| WS2 | -0.3% | 0% |
| WS3 | -0.2% | 0.24% |
| WS4 | -0.4% | -0.44% |

A reason for the sudden change in drift in the period between February and June 2017 could be that a calibration or renewal of the pyranometers had happened. As it turns out, the pyranometers were calibrated on 9. and 10. of May 2017. The calibration report included yearly averaged drift values which are presented in Table 4-1. The results from the calibration report are different from the calculated values found in this thesis. A challenge with the chosen approach is the assumption of stable atmospheric conditions at the site. However, it could be changes in air pollution, with respect to transmission and scattering, which is the reason for all the negative calculated drift values (Øgaard, 2016). In fact, the concept of global dimming and brightening is based upon changes in air pollution over the years and witnessed across the globe and could be the reason for the negative trend calculated (iac.ethz.ch, 2019).

Although calculations of the yearly drift have been made, the values are well within the uncertainty values of the measuring device. Therefore, it is difficult to conclude whether the drift from calculations is real. Furthermore, the average drift from the calibration report is based upon the average drift from installation. Whereas the drift values calculated are from a shorter time domain. As a result, if the drift is not evenly distributed throughout the years inaccurate values would be calculated. To accurately calculate a yearly average drift with absolute value in the range from 0% to 0.44% based upon measurements with 2% uncertainty would be difficult within a three years interval.

On the other hand, it is possible to detect changes in the overall trend and thereby estimate a period where a change in the measurements has occurred, visualised by red ellipses in Figure 4.5. The trend change in the figure coincides with the calibration date from the calibration report. After the sky conditions have been evaluated and drifting detected, a sufficient evaluation of the site conditions has been established.

## 4.3 Outliers

Before outlier detection was performed, a filter for common limitations and faults such as curtailments and disconnected inverters was applied. The results in this chapter are visualised for inverter 3. Firstly, disconnected inverters were found in measurements with solar irradiance above the threshold value of 50W/m$^2$ and power input to the inverter under 1000W for 10-minute averages. The times where inverter 3 is disconnected is illustrated in Figure 4.6 marked with x's.

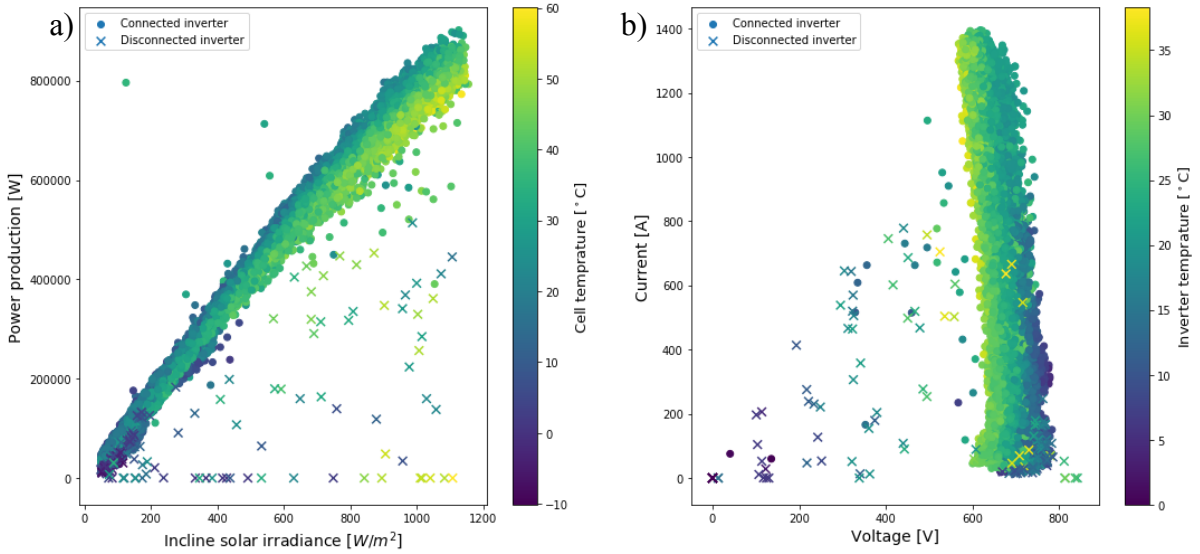



Figure 4.6 Illustration of the times where the inverter is disconnected in the training data. Disconnected inverters strike due to communication errors, cable faults and PPC-curtailments. In the figures disconnected inverters are marked with an x and applies for hourly averages. If an inverter has been disconnected at some point within the hourly average, then the measurement is labelled as disconnected inverter. a) shows the Power-Irradiance curve while b) is I-V curve.

A common reason for disconnected inverters is PPC-curtailments. However, the curtailment does not need to limit the inverter to a zero-power production. Therefore, yet a filter for PPC-curtailments is applied and illustrated in Figure 4.7.

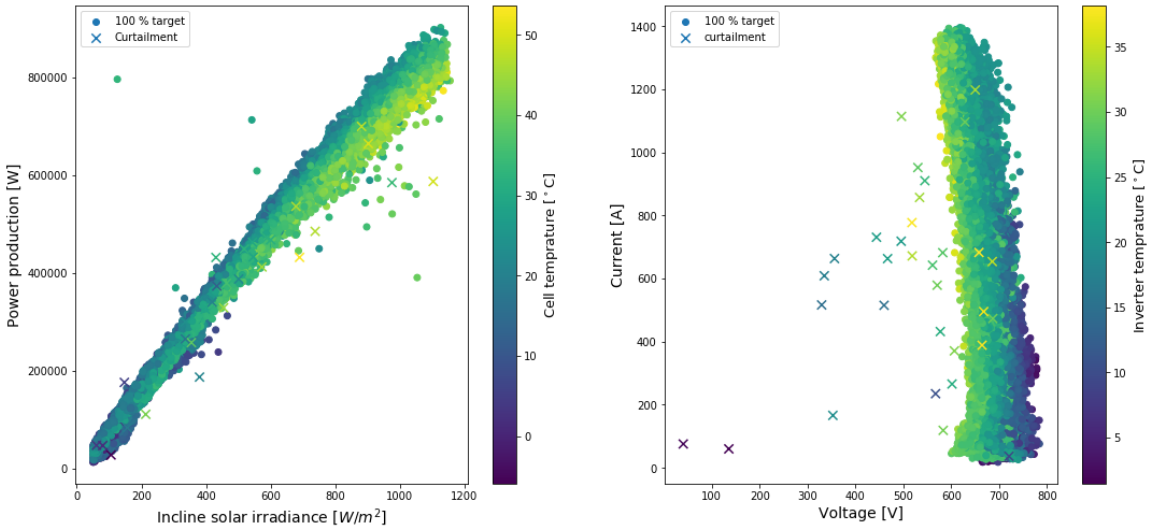



*Figure 4.7 PPC curtailments. PPC curtailments are marked with an x and applies for hourly averages. The figure is for inverter number 8, but the same tendencies applies for all inverters.*

The last limitation is clipping due to a maximum amount of apparent power. From the inverter datasheet, it was found that the maximum apparent power is 880 kVA and 800 kVA for respectively an inverter temperature at 25℃ and 50℃. Furthermore, an assumption of linear behaviour was made and a function of the maximum apparent power ($S_{max}$) with respect to inverter temperature ($T_{inv}$) was found to be

$$S_{max}(T_{inv}) = 967.5kVA - 3.2kVA/℃ \cdot T_{inv}$$

This relationship is illustrated in Figure 4.8 as a black line. In the figure, maximum apparent power clipping is detected as measurements where the apparent power is above 1.02 of the rated power. The value of 1.02 is chosen based upon the mean apperent power uncertainties presented in Chapter 3.2.2.

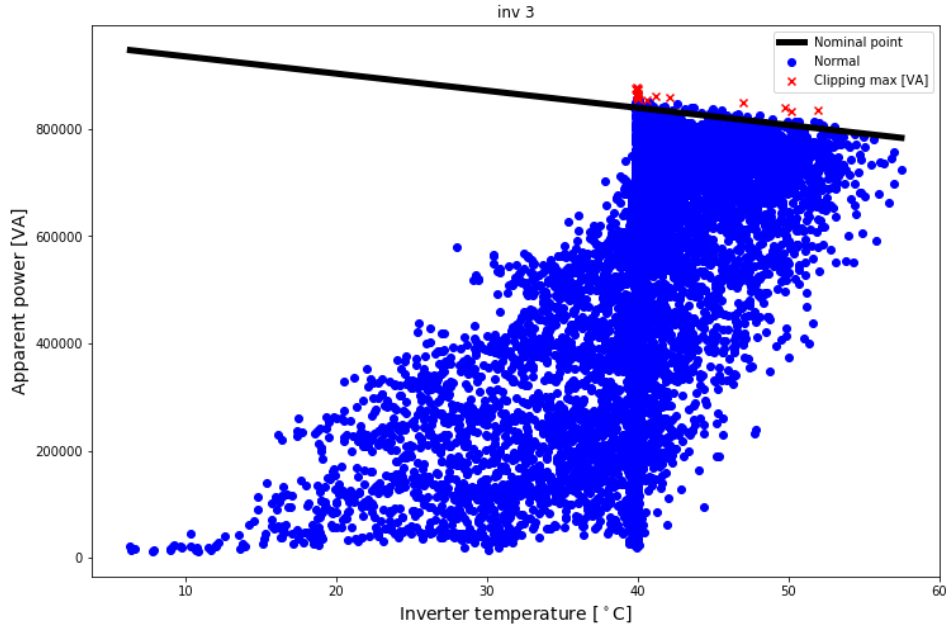



*Figure 4.8 Inverter clipping marked in red x's. The x-axis is the inverter interior temperature and the y-axis is the apparent power. The picture is filtered for disconnected inverters and PPC-curtailment. The blue circles indicate normal operation, whereas the red x's marks situation where the inverter has been clipped. Moreover, the black line is the nominal apparent power due to the measured inverter temperatures. Only the points 2% above the nominal operating point are defined as clipped measurements.*

After these forms for limitations were filtered out, outliers were detected as described in Chapter 3.5, the methods were applied to both the I-V curve of the inverter and the power-irradiance curve. The outlier rejection was performed on the whole training data (24.02.2014 to 24.02.2017). The result of KNN-outlier detection applied to the I-V curve of the inverter is illustrated in Figure 4.9, with the outlier degree as colour values. KNN outlier detection was used with 15 neighbours and the mean distance between the neighbours as the outlier degree.

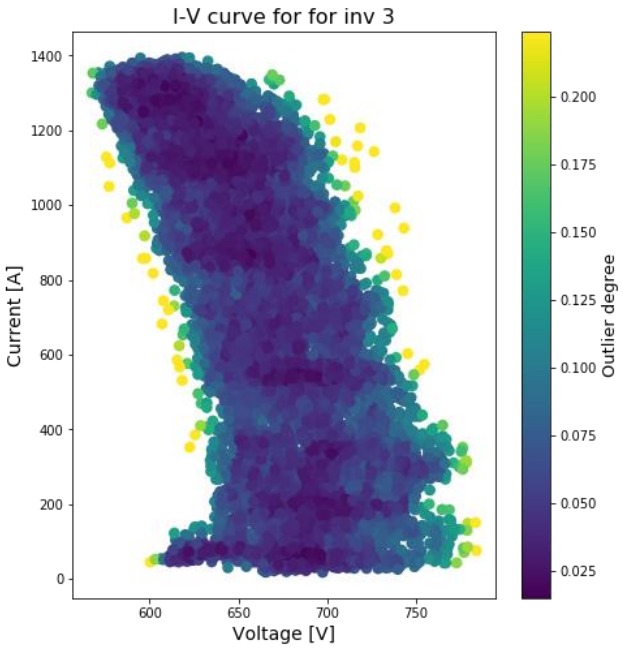



Figure 4.9 Illustration of the I-V plot of inverter 3. The colours illustrate the outlier degree for each measurement detected with KNN. The colours have a maximum value equal to the 99.7 percentage quantile. The y-axis is current, while the x-axis is the voltage.

Although only outlier detection for inverter 3 is visualised the same results were also obtained with the other inverters. The values with an outlier degree equal to the 99.7 percentage quantile, marked in yellow in the figure are some of the measurements which are rejected before the model training. Furthermore, outlier detection was also performed in the evaluation of the Power-Irradiance curve for all the inverters. The results for outlier detection with OLS, KNN and isolation forest for inverter 3 is illustrated in Figure 4.10. Although isolation forest did not perform well, the results are visualised for comparison purposes. In these figures, outlier degrees in the 99% quantile of abnormality are the maximum colour value.

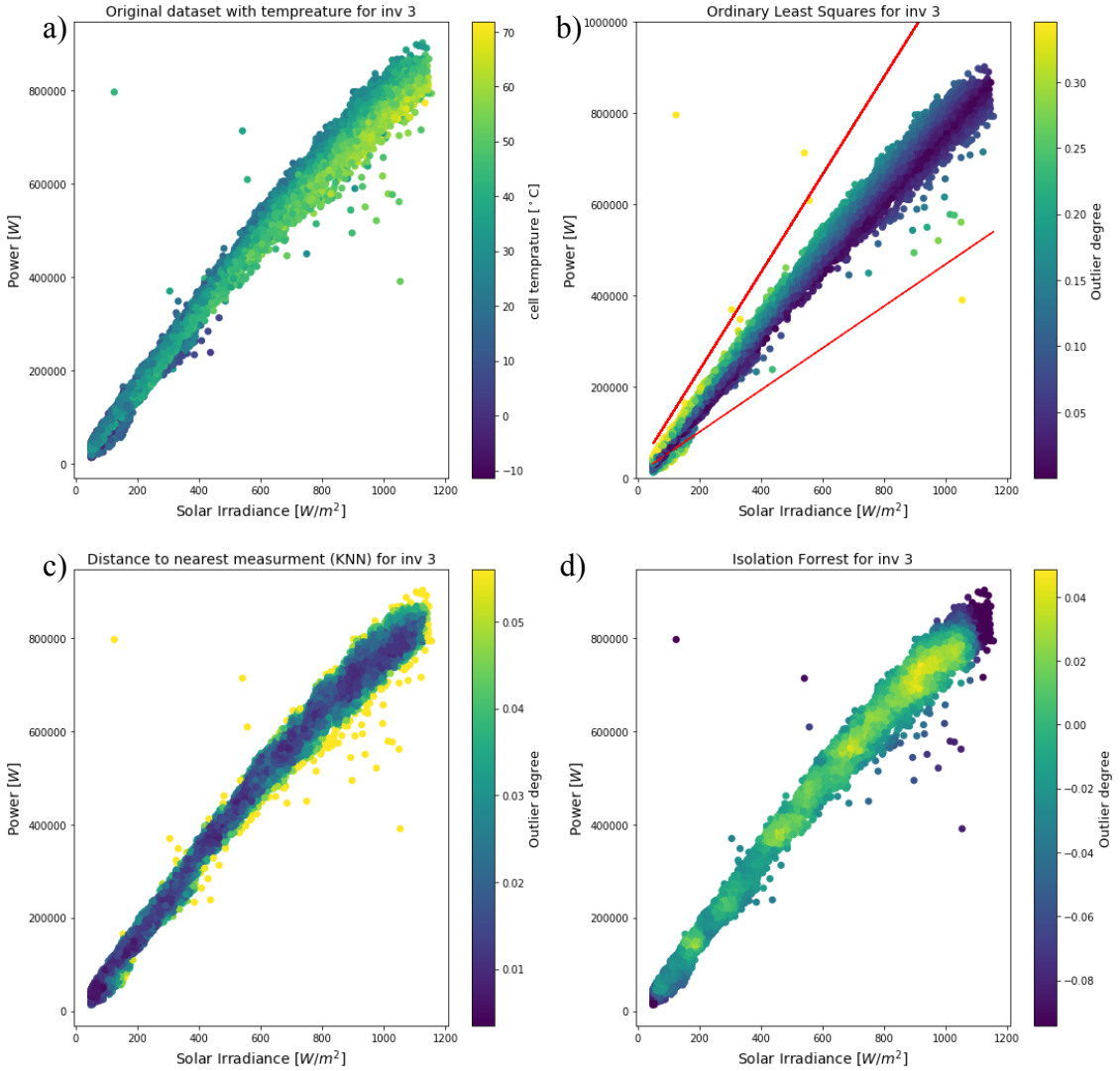



*Figure 4.10 Power -Irradiance plot for inverter 3. Common for all figures is that the y-axis is the power input to the inverter and the x axis is the measured solar irradiance from the closest weather station. In a) the colour values is the cell temperature whereas for the rest it is the outlier degree calculated with the respective detection algorithms. Furthermore, the two red lines in the OLS plot represent 0.4 outlier degree. The maximum colour value is the 99% quantile for KNN and OLS detection and 1% for isolation forest due to lower outlier degree indicates outliers.*

As seen from the figure, OLS-outlier detection is more aggressive towards outliers at low irradiance values in contrast to KNN. Furthermore, KNN can closer detect outliers at the lower side of the power-irradiance curve. Therefore, the detection algorithm going forward is KNN for both Power-Irradiance outliers in addition to I-V outliers. The outliers were also evaluated through time and the result was that the outliers detected seem spread through the period. However, it could be seen small groups of clusters, which could indicate a fault which has been quickly fixed.

By comparison of Figure 4.3 and Figure 4.10, it can be seen that many of the outliers detected with KNN are situations where the techniques for sky conditions evaluate the time as cloudy. Therefore, the filter for outliers could decrease the information regarding cloudy measurements in the training of the models. Another downside of the outlier detection performed in this work is the fact that it is applied to all the data. A more proper appliance would be to reject outliers in the test data based on the training data to prohibit information leakage. However, due to the low percentage of outliers ($\leq 1.3\%$) it will not affect the result significantly. Moreover, this feature is not currently available at scikit-learn but it is something the developers work on in time of writing of this thesis (Authors of scikit-learn, 2019). With this in mind, it was considered unnecessary to make a proper module for outlier rejection.

Furthermore, rejection of outliers in the data is necessary for two reasons. First, it is helpful to filter out abnormal behaviour of the PV system. Thus, enabling the models to learn the healthy behaviour of the system. In addition, the rejection of measurements increases the certainty of the model error by reducing the size of the error. Thereby, a more accurate description of the model error can be established. Moreover, when the chosen number of outliers were rejected, models can be built to mimic the PV system.

## 4.4 Model evaluation

Firstly, an evaluation was made upon measurements of only module temperature and irradiance. The solar module temperature is converted to solar cell temperature in accordance with equation 3.8. Subsequently, the models were evaluated after the introduction of drifting found in Table 4-1. Thereupon, the models were evaluated after implementation of feature engineering.

### 4.4.4 Irradiance and solar cell temperature

The five different models represented in the theory have been applied to the data where power, current and voltage are the response values. Detailed results after nested cross-validation for estimated power, voltage and current is available in appendix B. Where the value in each cell is based upon the mean MAE from the outer cross-validation in addition to the uncertainty calculated as the standard error of the MAE in the cross-validations in the outer fold. Nested cross-validation was applied with 5-folds and a subsequent hyperparameter tuning with 5-folds. Furthermore, nested cross-validation was applied without SFS, due to few explanatory variables.

The results from nested cross-validation can be summarized as error plots where the mean value is the mean MAE for all inverter and the error is the mean of the standard error in MAE for each inverter. Although Figure 4.11 only represent power, the same trend can be seen for current and voltage in appendix B. However, for current and voltage, the physical model's performance relatively worse than the others. As seen from the figure, the baseline has the worst performance and slightly better is the ridge and lasso models. Similarly, the performance of RFR and KNN is close to each other and has a higher performance than the linear models.

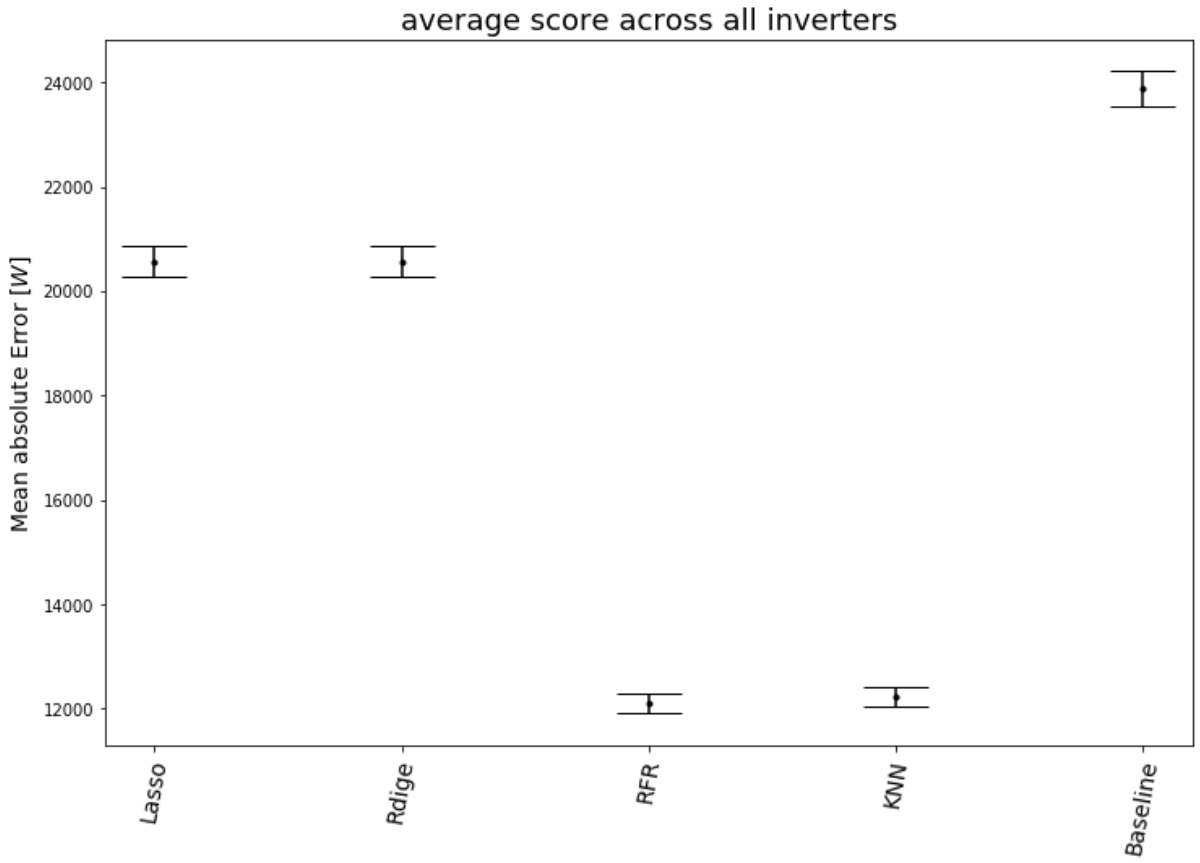



Figure 4.11 Average score after nested cross-validation for each model across the different inverters. The baseline has the lowest performance, close to Ridge and Lasso which has an equal performance. RFR and KNN has the highest performance amongst the models.

The results show an increase in predictive ability by the introduction of non-linear models such as RFR and KNN. Although the coefficients in the baseline models have been scaled with the ideal factor, they are not the most optimal ones. Due to the better performance of other linear models such as Ridge and Lasso. However, in contrast to Ridge and Lasso the coefficients of the Baseline models have the explanatory variables irradiance and irradiance times the cell temperature. Furthermore, the coefficients in the baseline model is based upon the conditions provided by the manufacturer. Yet, PV modules are known to perform differently according to the location, time of day and season of the year which is a drawback in the baseline models. With this established, it is interesting to see how the sensor drifting in the solar irradiance measurements affect the predictive power of the models.

### 4.4.5 Drifting

To evaluate the effects of drifting on the model performance, the average drift values for each year from Table 4-1 has been used. The result for nested cross-validation with the models applied to power is presented in appendix B for drift values from own calculation and the calibration report. The tables in appendix B are the mean absolute error after nested cross-validation without SFS applied for each inverter and the uncertainty of the value as the standard error between all the folds in nested cross-validation. The input to the models is the solar irradiance and solar cell temperature. Whereas the solar irradiance is multiplied by the average drift raised to the power of the number of years since 2014. For example, for an average drift value of 0.5% in 2016, the irradiance is multiplied by $(1+0.005)^2$. The average score for each model is further visualised by the error-plot in Figure 4.12. The blue plots are the results from drift based upon own calculation, the red plots are the result from drift based upon the reported drift values and the black plots is the results without compensation for drift.

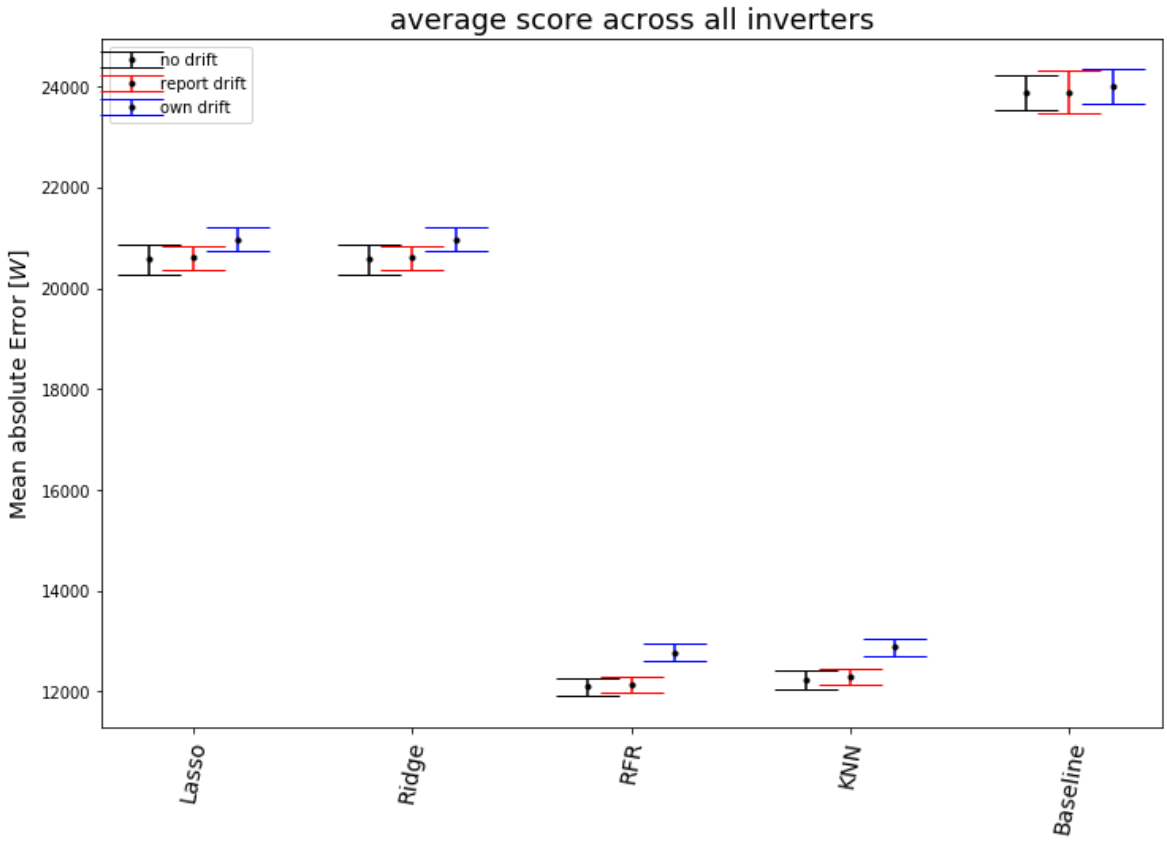



*Figure 4.12 Illustration of the mean MAE across all inverters and the mean standard error as the errors in the plot for the five different models applied to the inverter power input. For all models three scores has been found for irradiance. The first is without calibration for drift visualised in black. Secondly, for irradiance calibration based upon own calculated drift in blue. Lastly, for irradiance calibration based upon reported drift in red.*

From the figure and the tables in appendix B, there is little difference between the three evaluations. However, for the models based upon machine learning the standard error of the model's performance decreases. Furthermore, the same tendencies can be seen for both current and voltage in appendix B. The result confirms the suspicion that the wrong drift values were found with the approach presented in Chapter 3.3.3 since the models are more stable with the reported drift values. Therefore, in further evaluation, drifting is compensated for in the irradiance measurements with the reported drift values. Although, machine learning models have a better predictive ability than the physical baselines it is interesting to evaluate if the performance can increase. To further improve the model's performance, the introduction of new explanatory variables was assessed.

### 4.4.6 Feature engineering

Feature engineering (FE) was performed in accordance with Chapter 3.4. In addition, measurements of the inverter interior temperature and variables describing the sky conditions were included. However, due to RFR and KNN's weak extrapolation performance, the time in seconds since the first measurement was removed in these models. An example of extrapolation with RFR and KNN is shown in appendix B. Furthermore, the detailed results of feature engineering applied to power, current and voltage are available in appendix B. The models average score and average standard error from nested cross-validation is summarised in Figure 4.13. Although only power input is represented, the same tendencies can be seen for current and voltage which is available in appendix B.

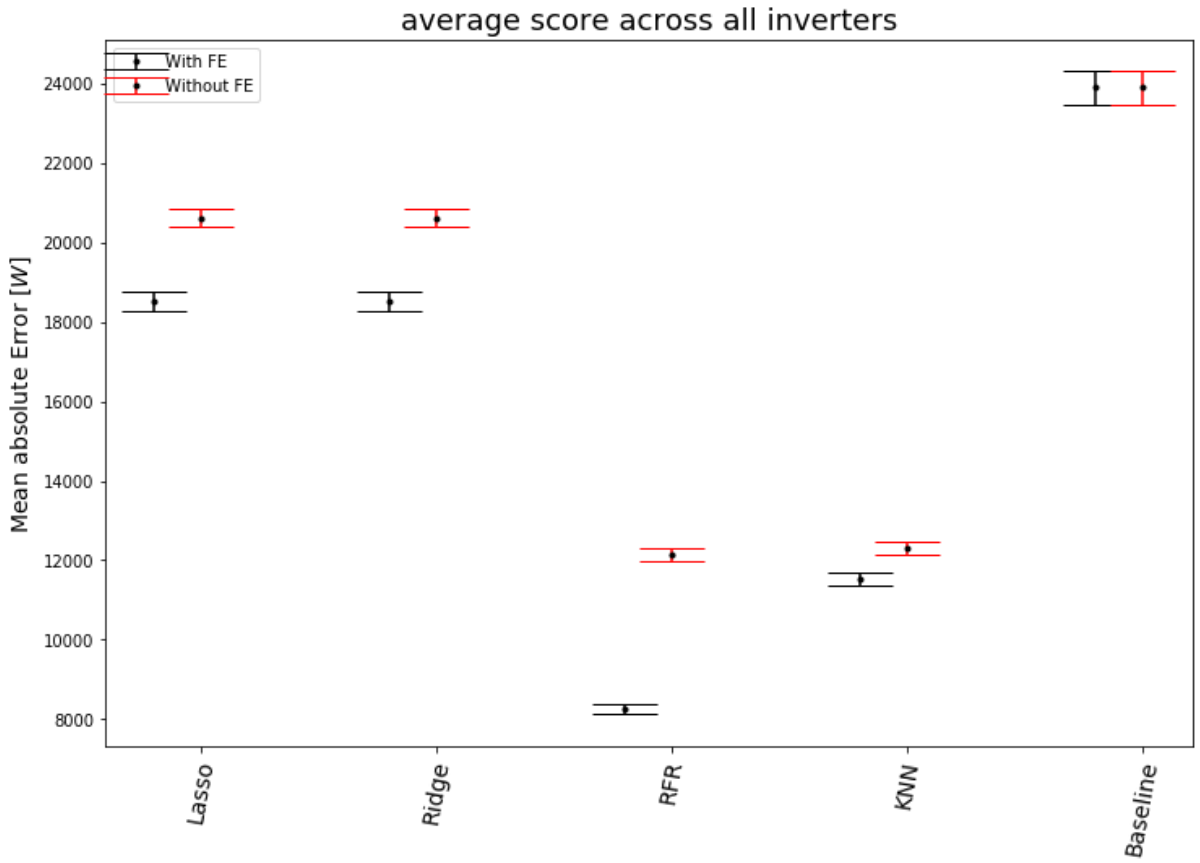



*Figure 4.13 Illustration of the score of each model with and without feature engineering. Since the baseline model only uses cell temperature and irradiance its performance does not change. However, all the other models show a significant performance improvement due to feature engineering.*

As seen from the figure, the baseline performs worst. However, the performance of all other models has increased, yet their rank is still the same. Indeed, the performance of Ridge and Lasso has increased but with the worst performance of the machine learning methods. KNN is second best, with performance increase relative to without feature engineering. Finally, RFR has the largest performance increase. In addition, the performance of RFR on the inverters whose modelling is based upon their local weather stations (INV1-INV8) is higher than the performance whose modelling is based upon the average irradiance from all weather stations (INV9-INV12). However, this difference is not clear for other methods.

By the introduction of new explanatory variables, such as the sky conditions, cosine and sine transformations of the day of the year, the hour of the day, inverter interior temperature and time since first measurements, the performance of all machine learning models increases. However, due to SFS, only the features with the best predictive ability is included in the final evaluation of the model. Although a thorough analysis of the variable's importance in each of the models is beyond the scope of the thesis, some points are worth noticing.

First, the coefficients in the linear models corresponding to the absolute time since the first measurement are negative. Thus, the models can register degradation trends in the PV system.

Furthermore, when SFS is applied to RFR the majority of the models use all the variables. However, it should be noticed that the RFR applied to power, current and voltage use different variables. The models for power use all the variables, whereas the models for voltage and current fluctuate more in the choice of parameter space. Current models tend to leave out parameters with information about the inverter and cell temperature. In addition, there are some examples where the standard deviation of the irradiance from all weather stations are not used.

Voltage models tend to leave out the detection of clear sky parameter. This can be due to the information from the detection method is non-informative, or that the model is biased and inflates the importance of continuous or high cardinality categorical variables as mentioned in Chapter 2.6. However, the models for current and power does use the variable, so it seems more likely that clear sky detection is non-informative for voltage. Furthermore, the relationship between voltage and irradiance is described as logarithmic while linear for current and power (ALQahtani, et al., 2012). Therefore, it seems reasonable that the voltage models will be more independent of the clear sky detection in contrast to power and current. With the increase in performance, the best performing machine learning models will be applied to the data and utilised to classify events. In addition, physical models will be used for comparison.

## 4.5 Event detection

Event detection was performed in accordance with the problem-solving approach presented in Chapter 3.1. The models used were both the physical baseline models, in addition to the best performing machine learning models. Event detection was applied to inverter 1 to 12.

### 4.5.4 Event detection - physical models

The threshold values presented in Figure 3.1 and Figure 3.2 were calculated as the sum of the model uncertainty and the maximum measurement uncertainty. The model uncertainty is calculated in accordance with Chapter 3.8.4 with data from 24.02.2014 to 24.02.2017 inserted to a 5-fold cross-validation. In the data, curtailment has been filtered out in addition to outliers, as described in the model evaluation chapter. For inverter 1 the model uncertainty for power is 9.42%, in accordance with equation 3.19. The power threshold is therefore 12.68%, whereas the same calculations are made for current and voltage. As a result, the threshold for current and voltage is 11.05% and 3.05% respectively. The values of model uncertainties for the different inverters are presented in Table 4-2 in accordance with equation 3.19 where z=1.

*Table 4-2 Relative model uncertainties for the inverters at the site. The tables are generated from the physical models with an irradiance threshold of 50W/m$^2$. In addition, curtailment and outliers have been filtered out. From equation 3.19 z is set to 1.*

| | **Power** | **Current** | **Voltage** |
|---|---|---|---|
| **INV1** | 8.91% ± 0.17% | 7.63% ± 0.14% | 1.99% ± 0.02% |
| **INV2** | 8.70% ± 0.16% | 7.27% ±0.13% | 2.08% ± 0.02% |
| **INV3** | 9.34% ± 0.19% | 8.03% ± 0.16% | 2.07% ± 0.02% |
| **INV4** | 10.02% ± 0.20% | 8.44% ± 0.17% | 2.09% ± 0.02% |
| **INV5** | 10.37% ± 0.21% | 8.70% ± 0.17% | 2.07% ± 0.02% |
| **INV6** | 10.22% ± 0.20% | 8.54% ± 0.17% | 2.04% ± 0.02% |
| **INV7** | 9.41% ± 0.19% | 7.96 % 0.16% | 1.95% ± 0.02% |
| **INV8** | 9.61% ± 0.20% | 7.96% ± 0.16% | 1.92% ± 0.02% |
| **INV9** | 10.00% ± 0.19% | 8.52% ± 0.16% | 1.97% ± 0.02% |
| **INV10** | 9.94% ± 0.19% | 8.57% ± 0.16% | 1.91% ± 0.02% |
| **INV11** | 9.76% ± 0.18% | 8.27% ± 0.15% | 2.08% ± 0.02% |
| **INV12** | 9.36% ± 0.18% | 8.05% ± 0.16% | 1.91% ± 0.02% |

An example of event detection is illustrated in Figure 4.14 for inverter 8. Although only inverter 8 is shown, the same tendencies apply for all the inverters. The tendency is few events and a couple of major occurrences of low current in May 2017.

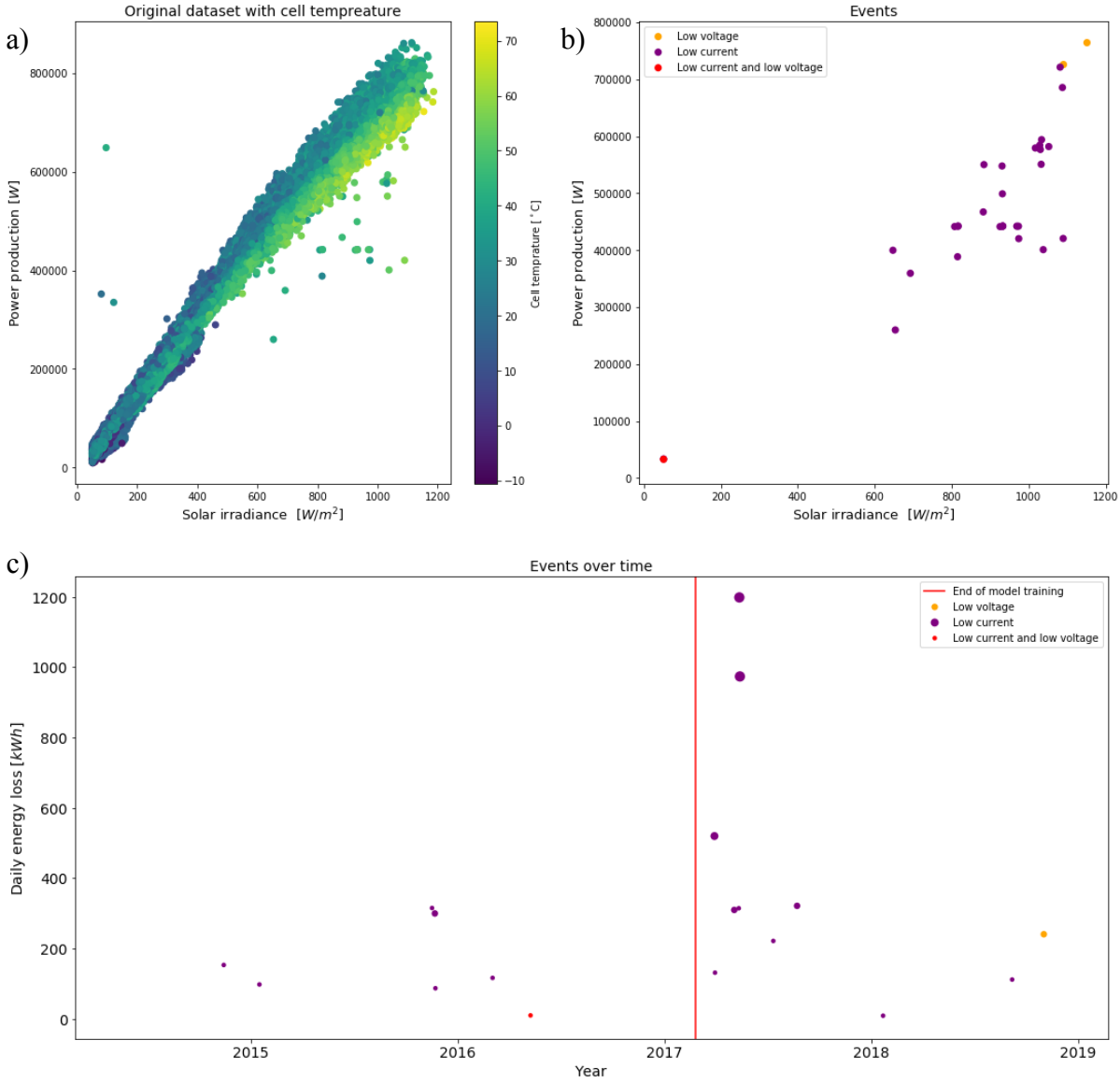



*Figure 4.14 Illustration of events detected with the used methodology for inverter 8. a) is the power -irradiance curve where the color values indicate the cell temperature. b) is the same curve but showing only the measurements where an event is detected. These events have been plotted against time in c. c) illustrates events in the time period between 8.00 and 13.00 with at least two occurrences within the day. The size of the event indicates the duration of the event for one day. The y-axis is the daily energy loss due to the specific event and the x-axis is the years from 24.02.2014 to 24.02.2019.*

The two largest low current detections are on 13. and 14. of May 2017. As is seen from the figure above, it seems to be few events in the period, either due to good operating conditions or a high threshold for events. Therefore, to reduce the threshold for events, a different irradiance threshold was applied rather than the one represented in 3.3.4. The same methodology is tested with a solar irradiance threshold value of 250W/m$^2$. The result is that the physical models performed significantly better and the model error is represented in Table 4-3.

*Table 4-3 Relative model uncertainties for the inverters at the site. The tables are generated from the physical models with an irradiance threshold of 250W/m$^2$. In addition, curtailment and outliers have been filtered out. From equation 3.19 z is set to 1.*

| | Power | Current | Voltage |
|---|---|---|---|
| **INV1** | 3.12% ± 0.05% | 2.93% ± 0.04% | 1.13% ± 0.01% |
| **INV2** | 2.99% ± 0.04% | 2.79% ± 0.04% | 1.17% ± 0.01% |
| **INV3** | 3.03% ± 0.04% | 2.68% ± 0.04% | 1.21% ± 0.01% |
| **INV4** | 3.31% ± 0.05% | 2.96% ± 0.05% | 1.13% ± 0.01% |
| **INV5** | 3.42% ± 0.05% | 2.98% ± 0.05% | 1.19% ± 0.01% |
| **INV6** | 3.36% ± 0.05% | 2.87% ± 0.05% | 1.16% ± 0.01% |
| **INV7** | 3.28% ± 0.05% | 2.83% ± 0.05% | 1.12% ± 0.01% |
| **INV8** | 3.35% ± 0.05% | 2.87% ± 0.05% | 1.12% ± 0.01% |
| **INV9** | 3.42% ± 0.05% | 3.14% ± 0.05% | 1.06% ± 0.01% |
| **INV10** | 3.42% ± 0.05% | 3.10% ± 0.05% | 1.03% ± 0.01% |
| **INV11** | 3.49% ± 0.05% | 3.20% ± 0.05% | 1.16% ± 0.01% |
| **INV12** | 3.16% ± 0.05% | 2.76% ± 0.04% | 1.09% ± 0.01% |

By comparison of Table 4-2 and Table 4-3, the reduction in model uncertainty is evident. With the change in solar irradiance threshold, the models for current and power reduces the uncertainty with roughly one-third. Furthermore, these new models are again evaluated on the data with an irradiance threshold of 250W/m$^2$. The model's ability to classify events increases and is visualised in Figure 4.16. The figure presents events detected for inverter 8. However, the same tendency applies to all inverters. The tendency is that a higher number of events are detected during the whole period. In addition, there are many occurrences of low voltage in the end of 2017 and the beginning of 2018 whereas the amount of detection of low voltage seems to increase in the period of October-December 2017 for only inverter 8. As can be seen from Figure 4.16, most of the detected events are low current. The same trend can be seen across different inverters. The distribution of the events found for inverter 8 is visualised in a pie chart in Figure 4.15.

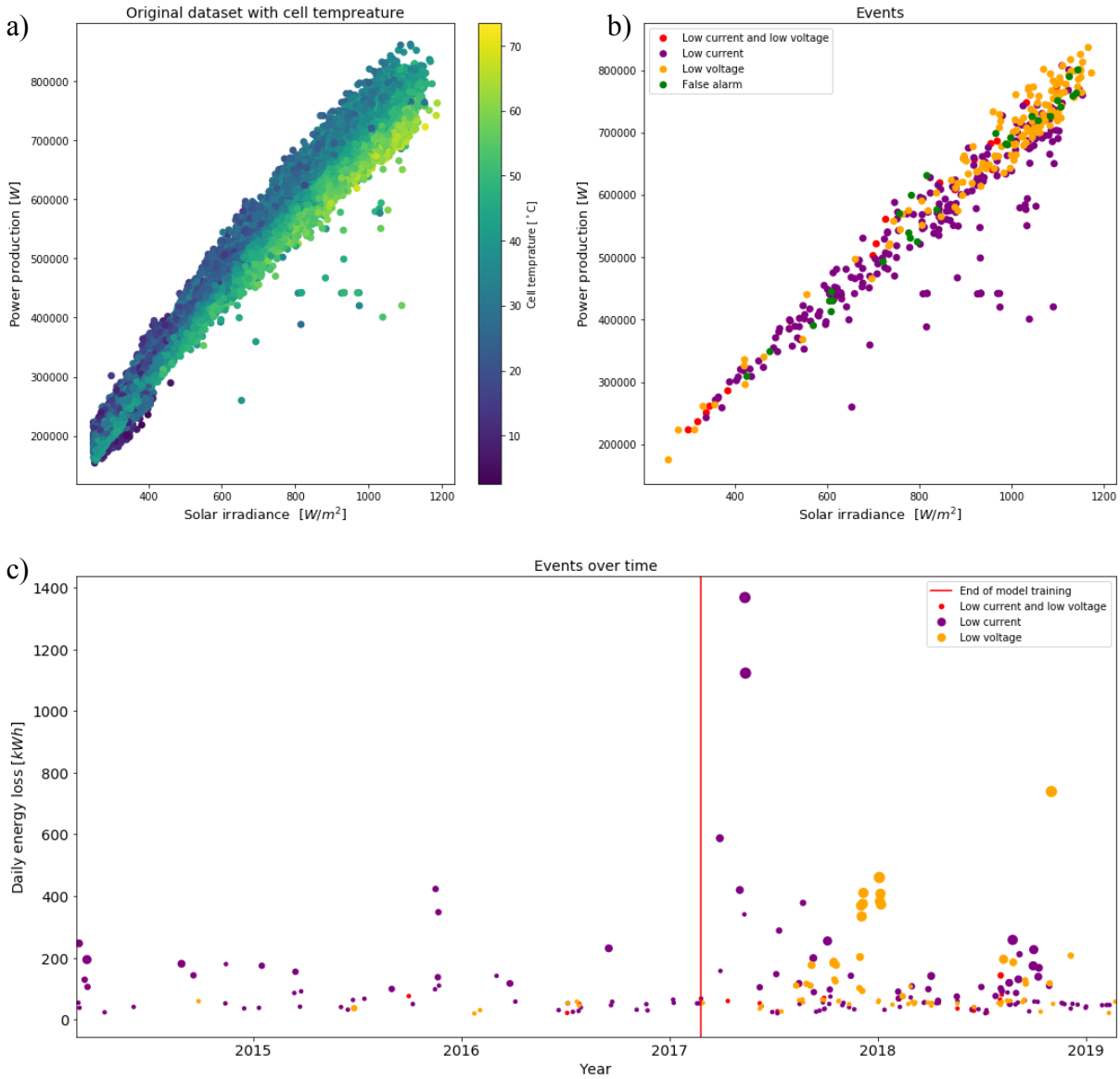



*Figure 4.16 Illustration of events detected with the used methodology for inverter 8. a) is the power -irradiance curve, where the color values indicate the cell temperature. b) is the same curve, but showing only the measurements where an event is detected. These events have been plotted against time in c. c illustrates only events in the time period between 8.00 and 13.00 with at least two occurrences within the day. The size of the event indicates the duration of the event for one day. The y-axis is the energy loss during that day due to the specific event and the x-axis is the years from 24.02.2014 to 24.02.2019. In contrast to other figures, the figure presented is for a solar irradiance threshold of 250W/m².*

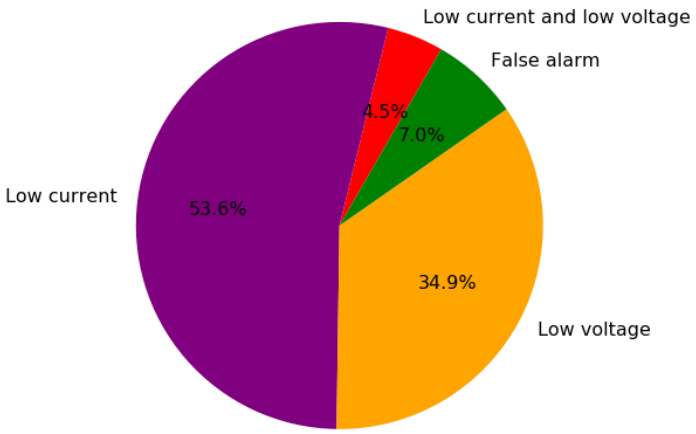



*Figure 4.15 Pie chart of the detected events on inverter 8 during the period investigated.*

Due to a large number of low current events, it is interesting to evaluate the evolution of the number of events through time for all inverters. This is illustrated in Figure 4.17.

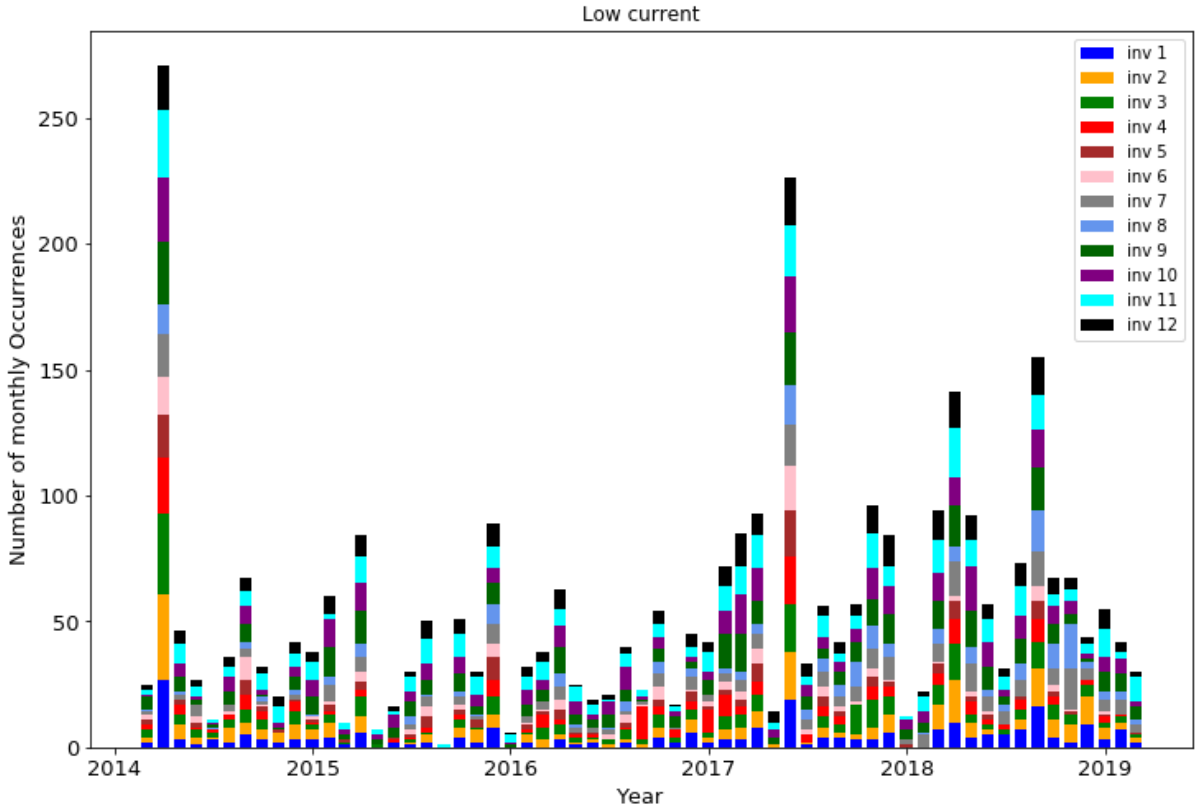



*Figure 4.17 Illustration of the number of monthly occurrences of events classified as low current from the chosen inverters at the site. Each inverter is presented with a unique colour value and the number of events for each month is shown. The figure is obtained with the physical models and an irradiance threshold of 250W/m$^2$.*

From the figure, the two months with the highest number of low current events is March 2014 and May 2017. The reason for this is that the site has been operated on half capacity on 13. and 14. May 2017. This could also to apply for February 2014 where the park operated in about 80% capacity at the beginning of the period in the sampled data. The reason for the low capacity at the beginning of the period could be because the park was established late 2013 and therefore the park was not operated at full capacity before a point in mid-March. Furthermore, due to the sampled data starting from 24.02.2014 only 5-days are accounted for. In those five days, 25 low current events were detected in February. To evaluate the performance of the physical models, Random forest is used in the next subchapter.

### 4.5.5 Event detection - Random Forest regressor

The threshold values presented in Figure 3.1 and Figure 3.2 was calculated as described for the physical models. Unlike the physical models, Random forest has a high prediction performance and therefore a low model uncertainty. The model uncertainty for power, current and voltage for each inverter is presented in Table 4-4.

*Table 4-4 Relative model uncertainties for the inverters at the site. The tables are generated from RFR with an irradiance threshold of 50W/m$^2$. In addition, curtailment and outliers have been filtered out. From equation 3.19 z is set to 1.*

| | Power | Current | Voltage |
|---|---|---|---|
| **INV1** | 2.49% ± 0.04% | 2.41% ± 0.04% | 0.76% ± 0.01% |
| **INV2** | 2.38% ± 0.04% | 2.35% ± 0.04% | 0.75% ± 0.01% |
| **INV3** | 2.49% ± 0.04% | 2.36% ± 0.04% | 0.75% ± 0.01% |
| **INV4** | 2.71% ± 0.04% | 2.47% ± 0.04% | 0.74% ± 0.01% |
| **INV5** | 2.89% ± 0.04% | 2.57% ± 0.04% | 0.80% ± 0.01% |
| **INV6** | 2.85% ± 0.04% | 2.49% ± 0.04% | 0.79% ± 0.01% |
| **INV7** | 2.68% ± 0.04% | 2.50% ± 0.04% | 0.75% ± 0.01% |
| **INV8** | 2.67% ± 0.04% | 2.44% ± 0.04% | 0.76% ± 0.01% |
| **INV9** | 2.87% ± 0.05% | 2.69% ± 0.05% | 0.73% ± 0.01% |
| **INV10** | 2.96% ± 0.05% | 2.76% ± 0.05% | 0.74% ± 0.01% |
| **INV11** | 2.98% ± 0.04% | 2.83% ± 0.04% | 0.77% ± 0.01% |
| **INV12** | 2.67% ± 0.04% | 2.45% ± 0.04% | 0.75% ± 0.01% |

After the threshold values for power, voltage and current have been calculated, events were detected for each inverter. The distribution of events and results for inverter 8 are visualised in Figure 4.18 and Figure 4.19.

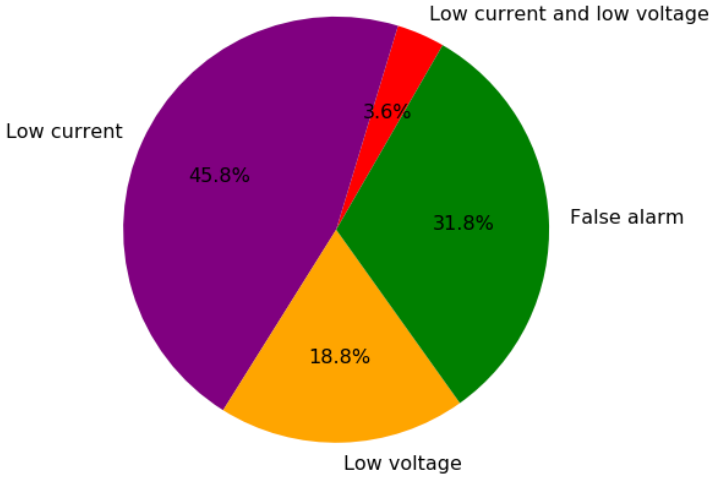



*Figure 4.18 Pie chart of the detected events in inverter 8 during the period investigated.*

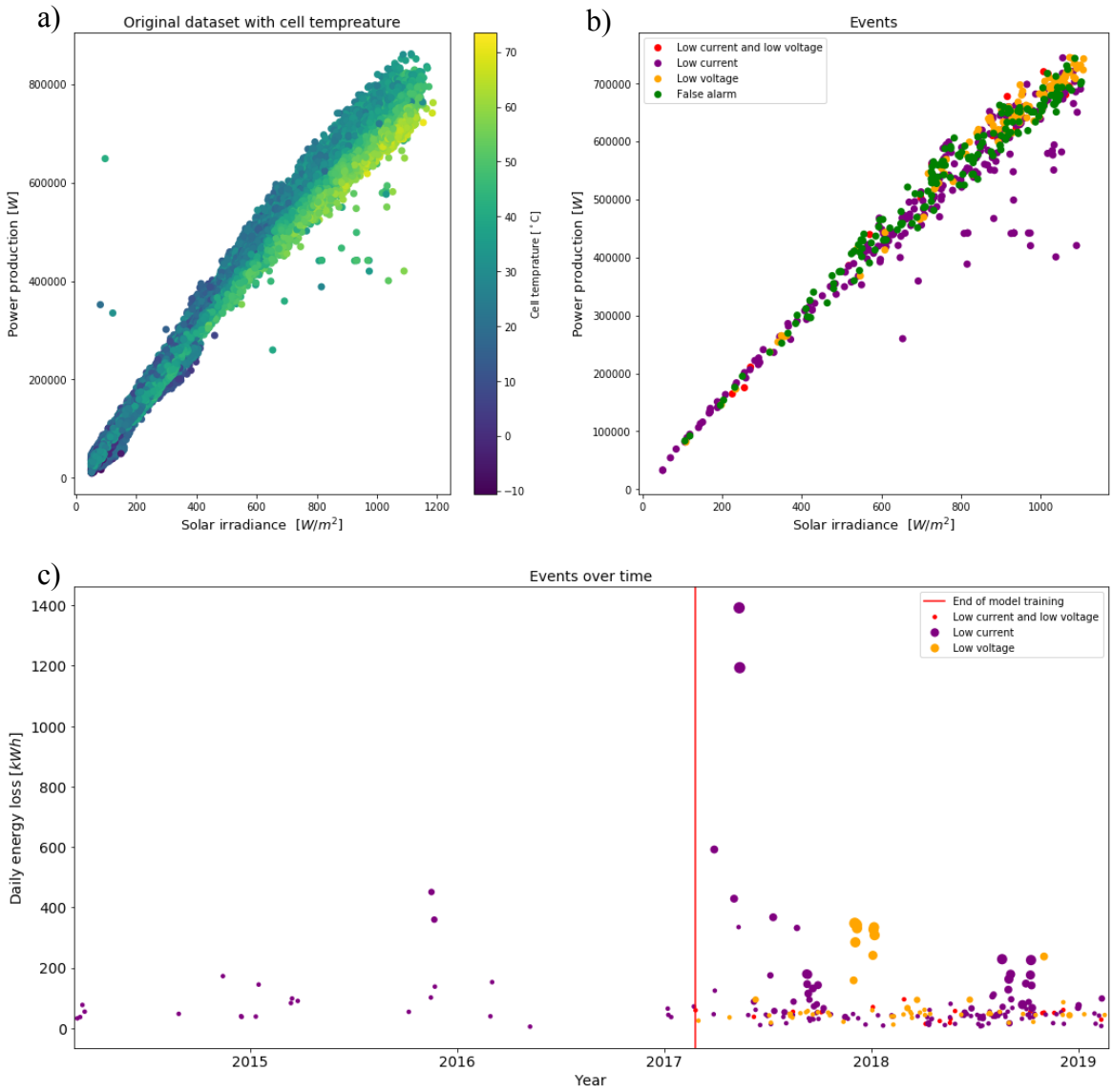



*Figure 4.19 Events detected with RFR in the period from 24.02.2014-24.02.2019 with a solar irradiance threshold of 50W/m$^2$. Figure a is the power -irradiance curve where the color values indicate the cell temperature. Moreover, figure b is same curve, but showing only the measurements where an event is detected. These events have been plotted against time in figure c. In figure c, only events in the time period between 8.00 and 13.00 are visualised with at least two occurrences within the day. The size of the event indicates the duration of the event for one day. The y-axis is the energy loss during that day due to the specific event and the x-axis is the years from 24.02.2014 to 24.02.2019.*

As can be seen from both Figure 4.16 and Figure 4.19, there seem to be many events on the inverter from August 2017 until the end of the period. This tendency is seen across the inverters on the site. Furthermore, the results are based upon a model trained on data from 2014 to 2017, so the detections made in this period are biased and should be disregarded. Although for comparison reasons they were included to compare with the physical models. Due to the high number of low current events an evaluation of the monthly occurrences was made and visualised in Figure 4.20.

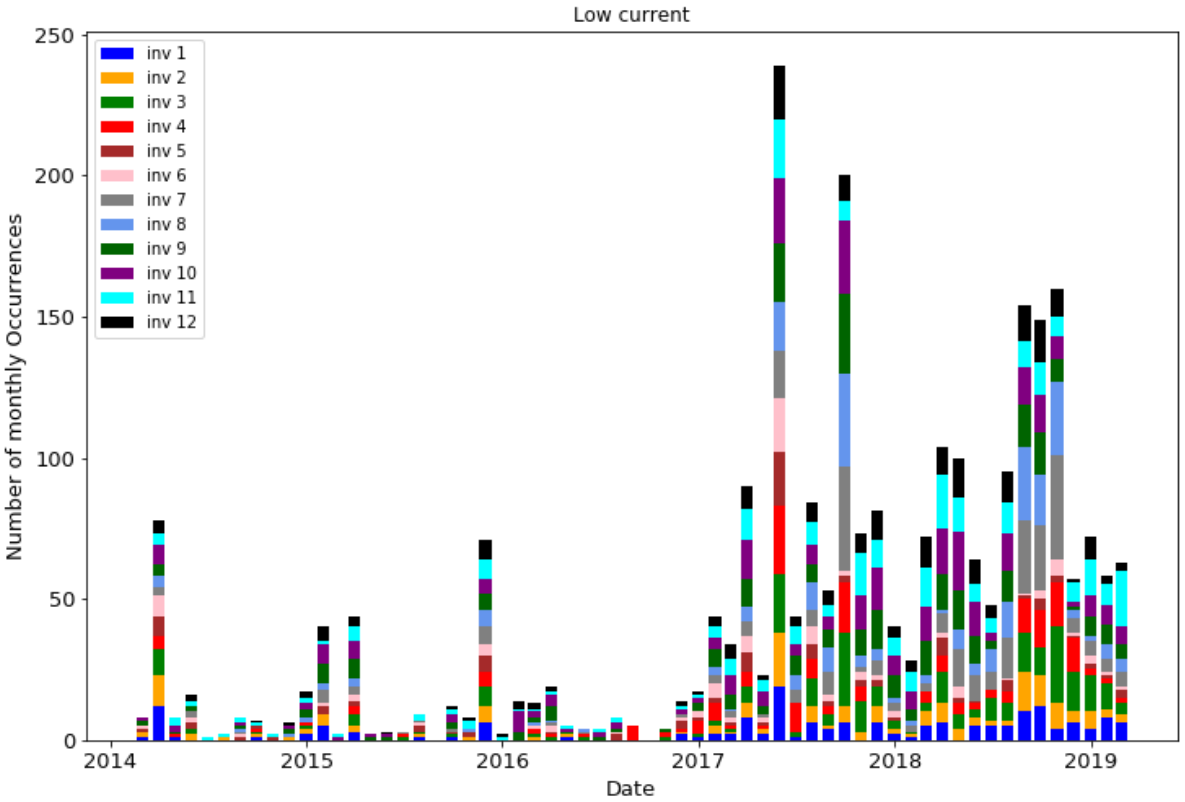



*Figure 4.20 Illustration of the number of monthly occurrences of low current events from the chosen inverters at the site. Each inverter is presented with a unique colour value and the total occurrences for each month is shown. The figure is obtained with RFR and an irradiance threshold at 50W/m$^2$.*

Due to the bias in the period of 24.02.2014-24.02.2017, this part of the figure does not provide much information. Although it is interesting to see the number of detected events after this period. As for the physical models, RFR detects the half capacity in May 2017 where both methods detect a total of 176 low current events in between 12. and 15. May on all inverters. In contrast to the physical models, RFR detects a higher number of low current events in September 2017 in addition to the period August to October 2018. This could be due to the lower model error in RFR in contrast to the physical models. In fact, the mean model error of RFR is always in between the mean and maximum measurement uncertainty for all the measurements.

Furthermore, it is plausible that the performance of RFR could increase further by increasing the hyperparameter space. The best performing hyperparameters were often a *number of estimators* equal to 400. Therefore, the model's performance could increase by the introduction of more estimators. However, it is plausible that the minimum mean uncertainty of the inverter modelling could be limited to the mean measurement uncertainty. Moreover, due to the high amount of low current events and to evaluate if the events presented are realistic, an inspection of the string performance during the period was made.

### 4.5.6 String monitoring

Due to the high ratio of low current events and to validate the performance of the methodology represented, an evaluation of the performance of each string connected to an inverter was assessed. This was done by filter the data for different types of curtailments and periods when the inverter is disconnected. Subsequently, the physical current model was trained on data from 24.02.2014 to 24.02.2015. The model was trained upon data which was filtered for power-irradiance- and I-V outliers for each string. In addition, current-irradiance outliers were detected with KNN outlier detection and filtered out before model training. Thereupon, the model evaluated all the data from 24.02.2014 to 24.02.2019 without measurements with curtailment. In this period outliers were not filtered out. Furthermore, the ratio between the measured current and estimated current is defined as the current ratio (CR). The mean CR for each day in the period between 08:00 and 13:00 was calculated. This evaluation was done for each string and the result is illustrated in Figure 4.21 for inverter 8. The labels in the figure indicate the string pair. The abbreviations are string monitor (SM) and Current channel (C)

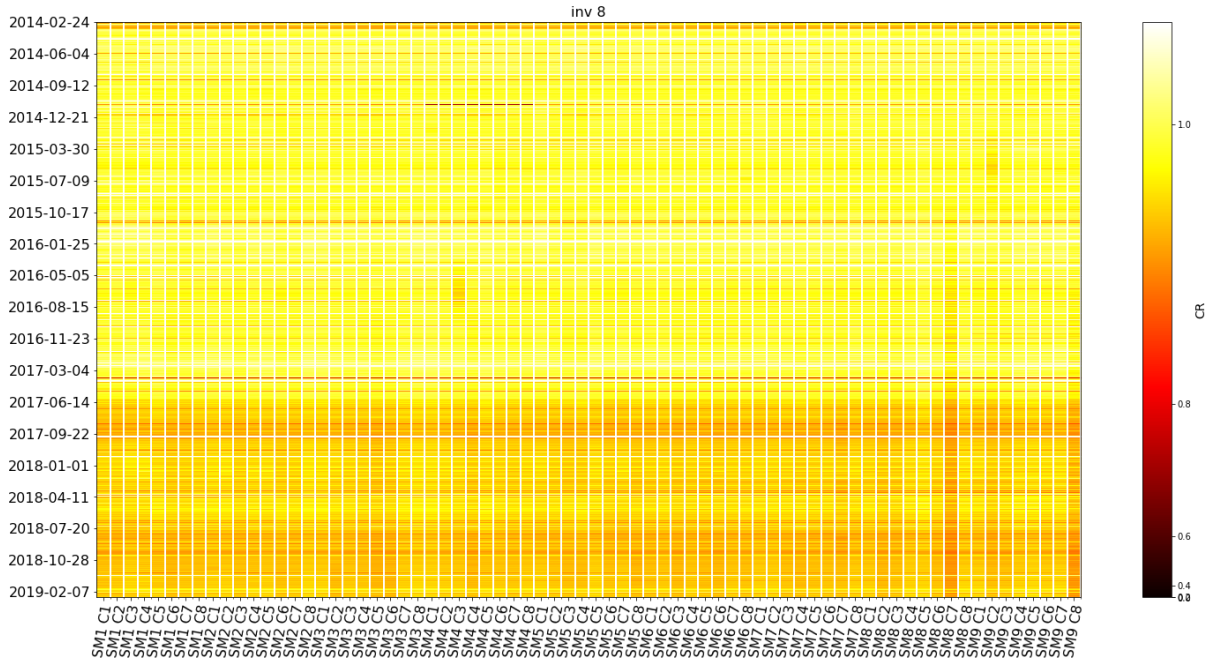



*Figure 4.21 Illustration of the physical current model applied to each of the string-pair on inverter 8. The colour value is the daily mean CR and are visualised on a logarithmic scale. The mean values of each day are inserted vertically for each string-pair in the figure.*

In Figure 4.21 the dark red line between 04.03.2017 and 14.06.2017 is for 13. and 14. of May. These power losses can also be seen in Figure 4.14, Figure 4.16 and Figure 4.19. In addition, there is a decrease in efficiency across all strings on 17.11.2015. At this day it was detected two events of low current with both RFR and the physical models. The details of the events were a substantial decrease in current from 08:00 to 10:00 and an increase in voltage in the same period. At this point inverter 2 and 6 were disconnected. Also, the event was detected across all connected inverters which could imply a form of undetected curtailment or maintenance of the power plant before the summer months.

Similar circumstances apply for most of the sudden decreases and subsequent increases in CR across all strings. However, the changes of most interest are vertical. From the figure, it can be seen a CR decrease from June 2017 on all string-pairs. This decrease can be seen across the site but with varying duration and amount. Inverter 1 experiences the same decrease but in a shorter time span as illustrated in Figure 4.22.

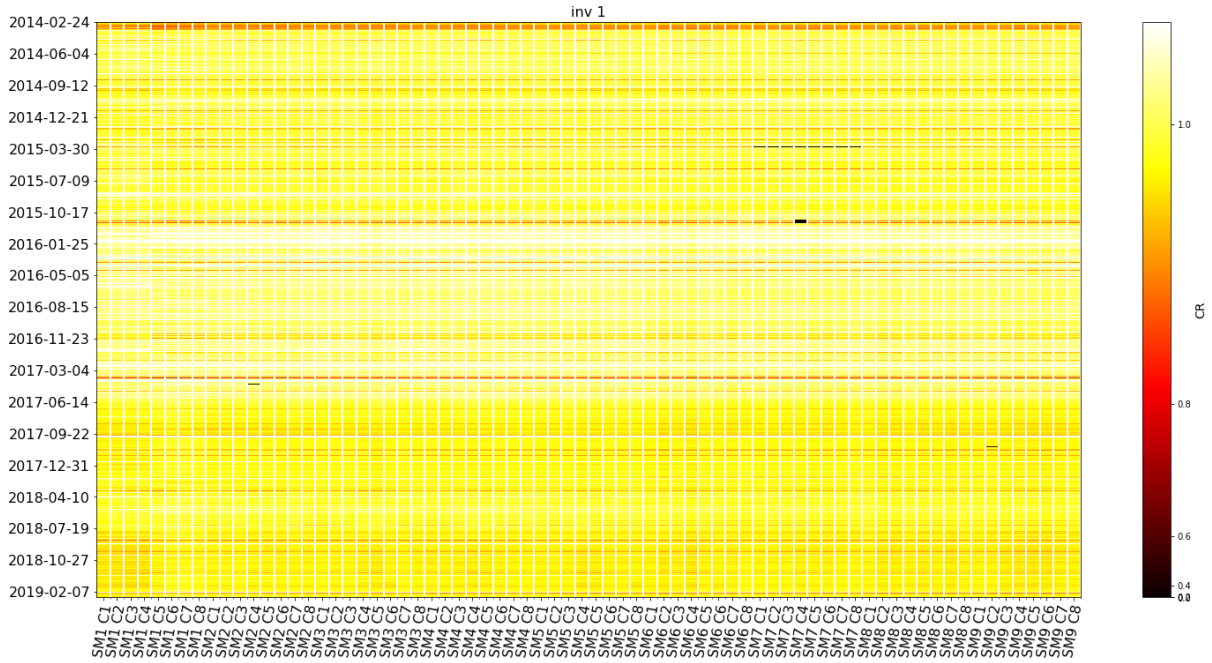



*Figure 4.22 Illustration of the physical current model applied to each of the string-pairs for inverter 1. The colour value is the daily mean CR and are visualised on a logarithmic scale. The mean values of each day are inserted vertically for each string-pair in the figure.*

Worth noticing is the gradual decrease in the performance of some strings. The variation in performance between each string can be visually enhanced by dividing the CR of each string for each day by the median CR of all strings from the same day. Thus, obtaining a relative current ratio (RCR). This is done for inverter 8 and visualised in Figure 4.23. The logarithmic scale is increased to enhance the performance decrease. By doing this, all variations that happen on inverter level and thus across all the strings in the figure, are removed. Therefore, the result will be a comparison of the performance on string level alone. In the figure, the colour-scale is adapted to show the variations in RCR.

In contrast to other techniques for event detection presented in this work, who evaluate the performance at the inverter level. The proposed method of string monitoring dives further into the details of each string-pair. Also, the evolution of the RCR for each string-pair as visualised in Figure 4.23 provides the opportunity to evaluate the development of possible faults. Thus, the operators of the power plant can detect small faults with a prolonged character. These faults are harder to detect in comparison with more severe faults which has a bigger reduction in the performance of the power plant.

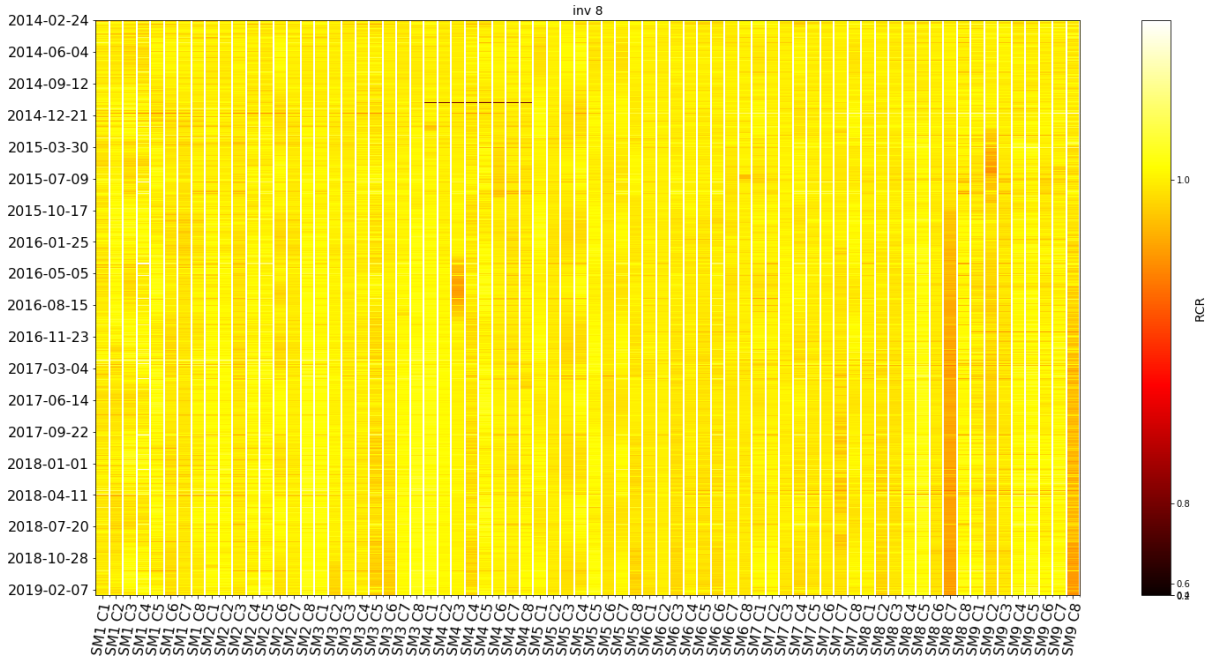



*Figure 4.23 Illustration of the physical current model applied to each of the strings in inverter 8. The colour value is the ratio between measured current and estimated current based upon the model, divided by the median CR of the sting. The colour values are visualised on a logarithmic scale. The mean values of each day are inserted vertically for each string in the figure. The seasonal decrease and increase in the RCR of some string-pairs, such as SM5 C7 and C5 could be due to inter-row shading whereas one string can shade for a subsequent string.*

For inverter 8 string monitor (SM) 8, the performance of the strings in current channel (C) 7 gradually decreases from the beginning of 2015. The same tendency can be seen in string monitor 9, current channel 8. Therefore, some strings were evaluated against infrared images taken in November 2018, with multiple matches. As a result, 4 of the top 5 thermal defects for inverter 8 detected by infrared images can be seen in Figure 4.23. The defects are on SM9 C8, SM3 C2, SM3 C6, SM 6 C1 and SM8 C7, ordered from highest to lowest thermal defect. By searching for a decrease in RCR it can be seen from SM3 C2 that the decrease in RCR increases in the spring of 2018. For SM3 C6 the RCR decrease starts in the summer of 2017. However, for SM6 C1 it is harder to detect a pattern. The reason for this could be that although a hot spot has occurred on the string, the impact on string performance varies. In addition, SM6 C1 has the 4. highest thermal defect detected from IR images. The 5. highest thermal defect from infrared images is on SM8 C7, this result emphasises that although a thermal defect has been spotted, their effect on performance can differ. From the infrared scanning, SM 9, current channel 8 had the most severe temperature difference. In the channel consisting of two strings connected in parallel, both strings had thermal defects shown in Figure 4.24.

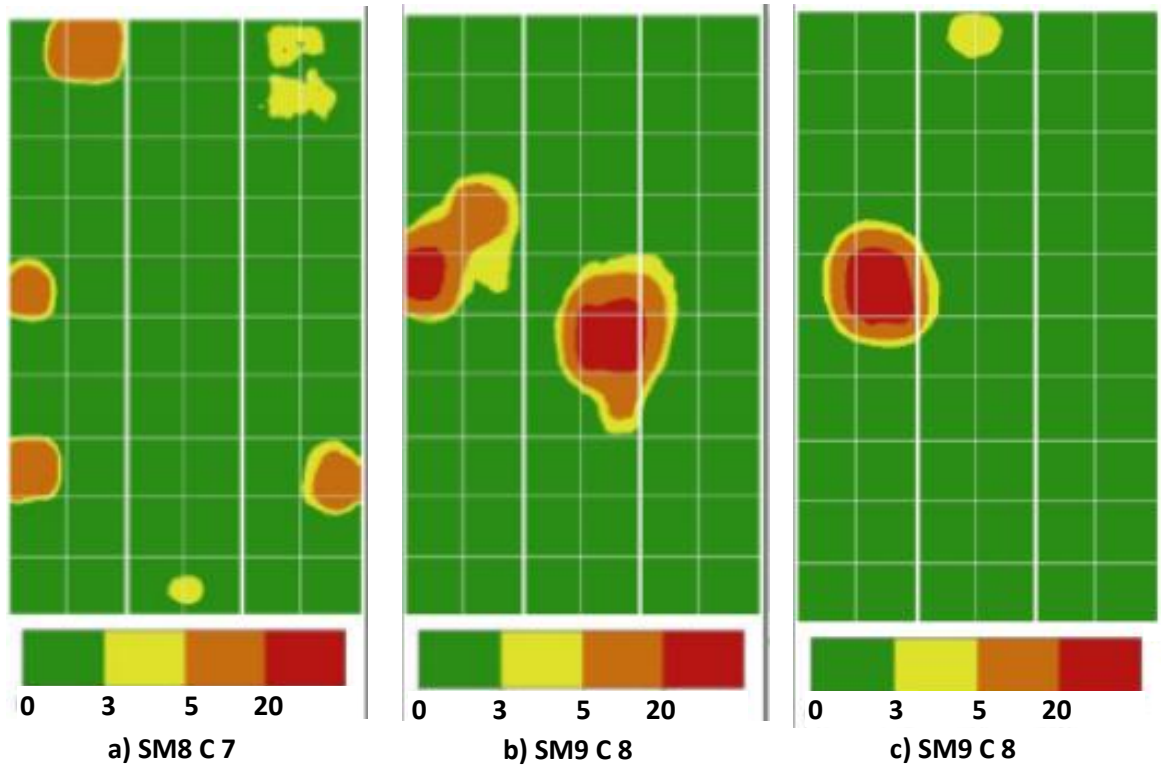



*Figure 4.24 Illustration of the detected faults samples from IR-images in November 2018. The figure highlights the thermal defects as heatmaps with the temperature difference between the average module temperature and the registered temperature from IR-images. a) illustrates string monitor 8 current channel 7, string 15 module number 13. SM9 current channel 8 string 13 module 1 is illustrated in b). The same current channel, but string 16 module 24 is illustrated in c.*

Furthermore, there are some strings-pairs which could appear to have a fault by RCR evaluation such as SM9 C2. The string-pair does not have a steady decrease in performance although it performs relatively worse in contrast to the period of 2014. The reason for this could be due to the change in RCR in the period of early 2015 to August the same year. A fault on the string could have happened with a subsequent fix in August 2015. The model, which has been trained on the string from 24.02.2014 to 24.02.2015, expect the same performance as before. However, due to technical changes, the string-pair seem to have a fault although the model may just need retraining. Nevertheless, a fault on the string-pair cannot be excluded. Therefore, in the evaluation of RCR, an important aspect is the evolution of the RCR.

Moreover, a limitation of the technique is in scenarios where a fault or multiple faults are present from the beginning. When the model is trained upon a string-pair with a constant fault it would not be able to detect this. A fault within the model training data could be due to improper installation or in the transport of the modules. However, there exist techniques that could detect such faults, due to a reduction in current from one string-pair. Due to the validation of the string monitoring system on IR images, it is interesting to evaluate the results obtained earlier with the results from string monitoring.

### 4.5.7 Comparison of detected events and string analysis

From the string analysis, there are times where all the strings experience a performance decrease. A comparison of the mean efficiency from all strings to the estimated power loss with the two event detection methods are visualised in Figure 4.25. The RFR-estimated energy loss for each day is illustrated in a), whereas the physical model's energy loss is illustrated in c) with values of kWh. To compare the energy losses with the CR from the strings, the mean CR from all the strings connected to the inverter is illustrated in b).

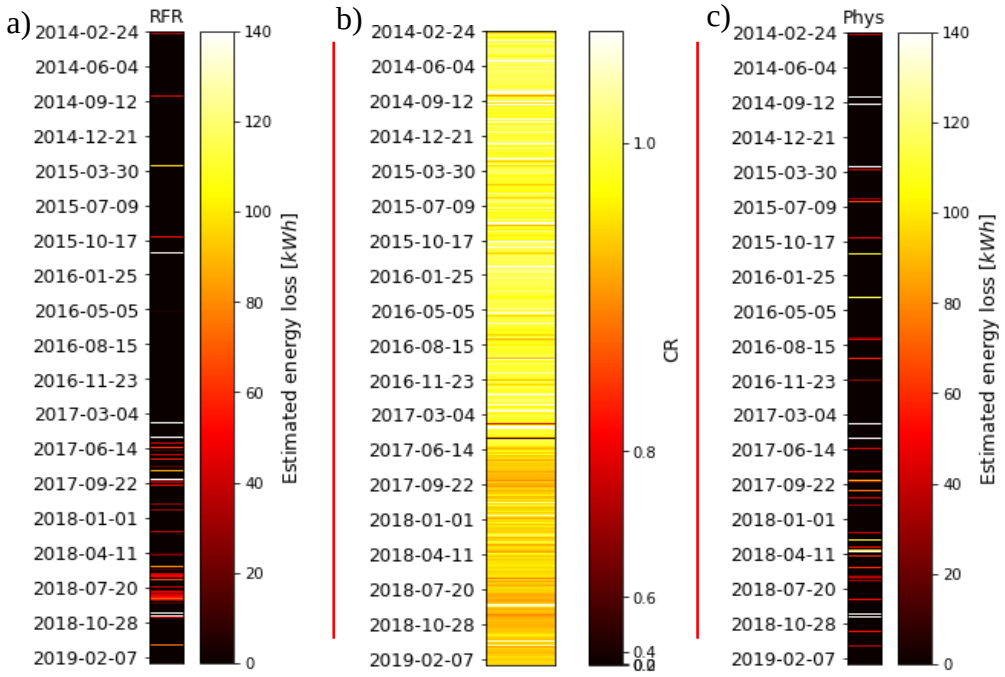



*Figure 4.25 Comparison of the three methods for power loss detection on inverter 8. Illustration of the estimated power loss with RFR (a) and the physical models (c). The estimation is based upon low current events for each day, where the colour values are in units of kWh and is visualised with a maximum value of 140kWh. b) is the average string CR from Figure 2.19. Note that RFR is trained on data from 24.02.2014-24.02.2017 and therefore the detections in this period are few and unreliable.*

From Figure 4.25, both detection algorithms detect periods when the efficiency across the strings decreases. To be specific, RFR detects more of the underperformance than the physical models due to the lower threshold values for events. This is illustrated by the increase in the frequency between events made by RFR in contrast to the physical models when the average string CR decreases in the period after 2017. However, none of the models detects all the underperformance. As mentioned in 4.5.2, many events of low current were detected in September 2017. By inspection of the average string CR in this period, an overall decrease in CR is showed by Figure 4.25. As is seen in the figure, the CR decrease starts in June and lasts to October. The same happens in the winter months of 2018 and the CR decrease lasts to November 2018. The CR decrease is small and applies to all the strings across the site. A reason for this could be due to soiling. Whereas the increase in CR around October 2017 and November 2018 could be due to rain. Heavy rainfall has proven to increase the performance of solar modules in a similar region with similar patterns for rain intensity (Øgaard, 2016). Therefore, the correlation between the increase in mean string CR and rainfall is visualised in Figure 4.26. The figure illustrates the maximum daily rain intensity and average string CR from all the strings during the period of interest.

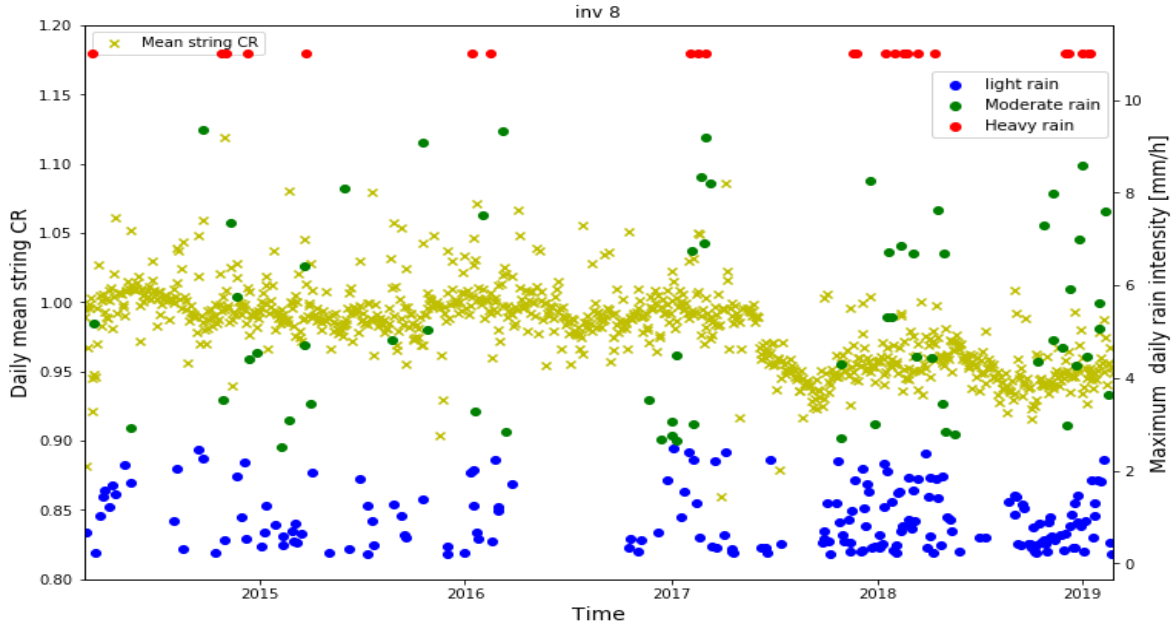



*Figure 4.26 Illustration of the mean efficiency of the strings connected to inverter 1 in addition to the maximum daily rain intensities. The values efficiency is calculated as two days averages across all strings connected to the inverter, while the rain intensity is defined as light (<2.5mm/h), moderate (<10mm/h) and heavy (>10mm/h). Furthermore, the rain intensity visualised is capped at 11 mm/h although they could be as high as over 50mm/h. The fact that the CR does not increase to its initial value indicates a deviation from the expected performance.*

The figure shows a decrease in the CR of the strings in the period of June 2017 until October 2017. Although the CR does not increase to the initial performance, an increase is shown in the period of October.2017 which can be due to an increase in rain intensity. This pattern is common in power loss estimation due to soiling (Kimber, et al., 2006). The same tendency can be seen in the winter period of 2018, with little precipitation and a decrease in CR. However, in October 2018 the rain intensity increases in addition to the CR of the string. Therefore, it seems that dust accumulates on the strings in the winter months of 2017 and 2018 and subsequently is cleaned, or partially, cleaned in the spring. This trend can also be seen across different inverters on the site. Furthermore, the difference between some of the inverters such as inverter 1 and inverter 8, could be due to their locations at the site. Inverter 8 is connected in the middle of the south-edge. Whereas inverter 1 is across the site connected to the north-west part of the plant. The tendency across the site seems to be from east to west, where the inverters in the middle of the plant are more exposed to soiling compared with inverters located closer to the edges. This could be due to less exposure to the wind which can naturally remove dust on the modules. Furthermore, dust is likely to settle in regions of low-pressure induced by high-speed wind over an inclined surface, which can occur in the middle of the site (Mani & Pillai, 2010). As shown in other studies from a similar region, the dust accumulation is probably not uniform at one site which explains the variation between inverter 1 and 8 (Øgaard, 2016).

Likewise, by analysing Figure 4.20 it seems to be spikes of low current events in the spring of 2017 and 2018. In the figure, there is a low number of low current events for inverter 1, 2, 6 and 5. These inverters are connected to the edge of the site in contrast to inverter 3, 4, 7 8 and 12, which has a relatively higher amount of low current events in this period. Inverter 9 and 10 are left out of this comparison since they are close to a nearby road.

### 4.5.8 Low voltage

Although most of the detected events are instances of low current, it is interesting to evaluate the occurrences of low voltage events. The distribution of low voltage events is illustrated in Figure 4.27a. From the figure, it can be seen a site-wide spike in events in December 2017 and January 2018. Therefore, Figure 4.27b highlights this period.

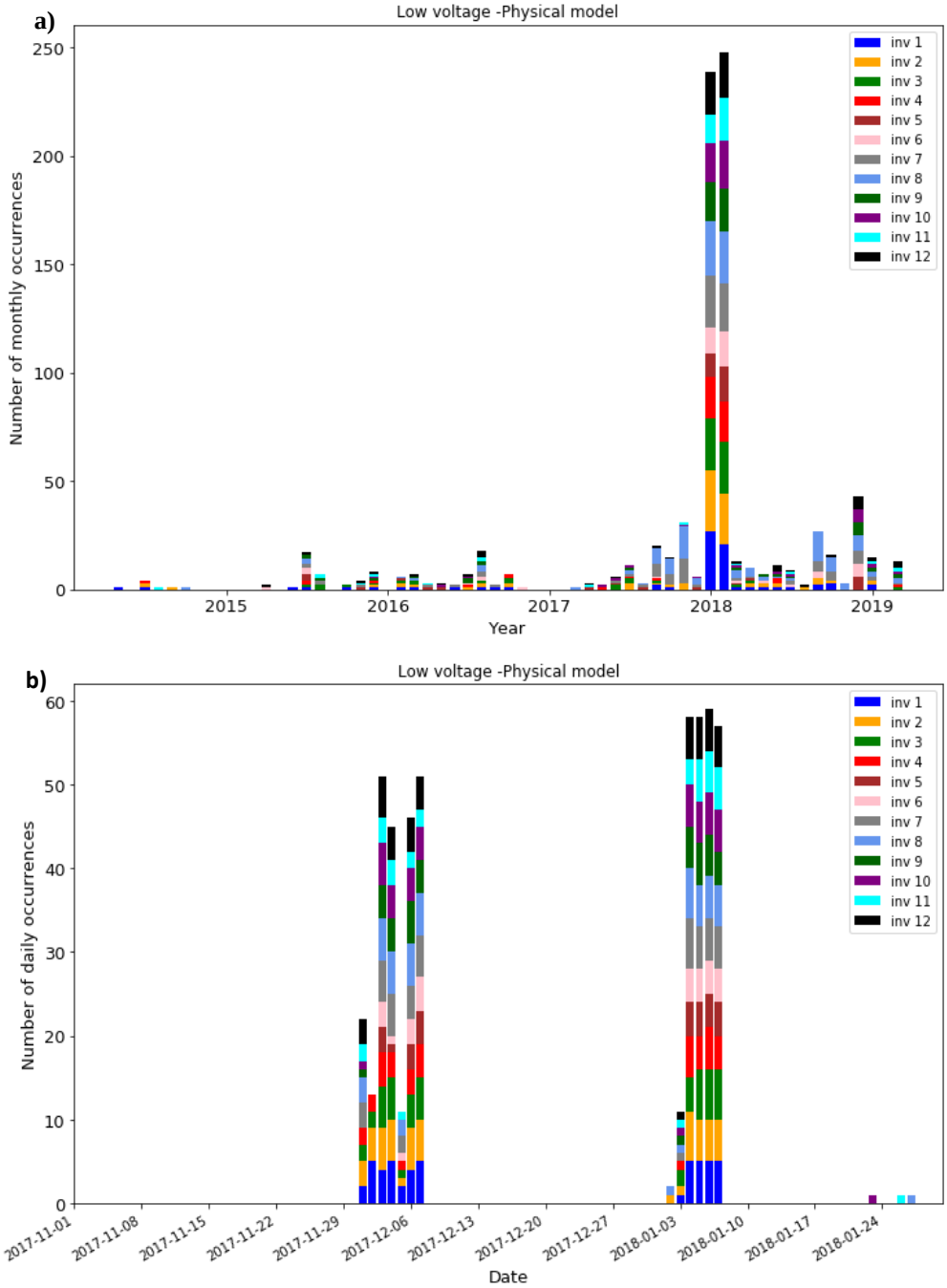



*Figure 4.27 distribution of low voltage events for each inverter. a) is from 24.02.2014 to 24.02.2019, whereas b) is for the period 01.11.2017 to 30.01.2018.*

Due to the site-wide events during the summer months, the first suspicion was that the temperature in all the solar modules had increased and therefore reduced the voltage. As a result, the voltage would be lower than the operating point of the inverter, leading to voltage

out of range. However, the first days in 01.12.2017 were cloudy and included much rain. Therefore, the measured temperature on the solar modules was relatively low. Moreover, the weather in this period could indicate deviations from normal weather and therefore the models for event detection could not handle this deviation.

Another suspicion is lightning strikes due to the impact on all inverters in addition to the bad weather in the period. From the information provided by the operators of the power plant, it was detected possible lightning events within the period of December.2017. However, it was not reported possible lightning events in January 2018 by the operators. Although, the weather on 04.01.2018 included clouds and rain which could imply the same weather conditions as for December.2017 and therefore possible lightning events. Due to the electromagnetic impulses generated from the lightning, the inverters across the power plant could deviate from the MPPT for security reasons.

### 4.5.9 Event distribution

Part of the methodology presented in this thesis is adapted from common techniques for fault classification in PV systems. The classification tools utilise the ratio between the estimated and measured values for current and voltage to classify possible faults in the same manner as in Figure 3.2 (Chouder & Silvestre, 2010) (Livera, et al., 2018). In contrast to Figure 3.2, each event would result in possible faults. Therefore, it is interesting to evaluate if the same fault diagnosis tools could be applied to utility-scale solar power plants. To illustrate the ratio of the different events, pie charts of the detected events with RFR and the physical models from all inverters have been visualised in Figure 4.28. In the figure, the distribution of the detected events for RFR is visualised in a) and c), as well as for the physical models in b) and d). Figure c and d does not include the period from 01.12.2017 to 30.01.2018.

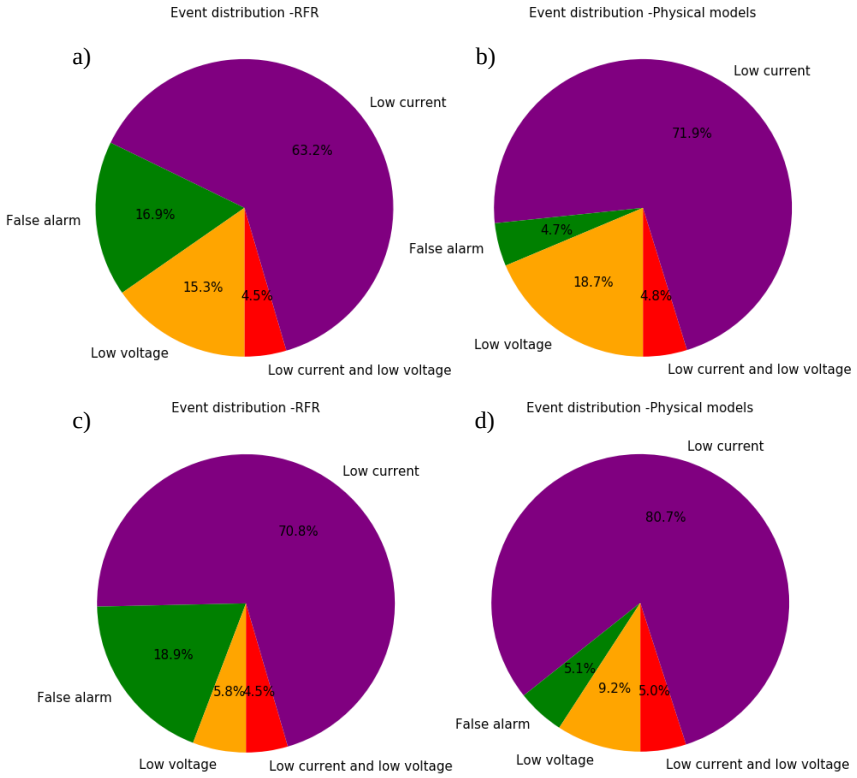



*Figure 4.28 Illustration of the distribution of events for the different detection techniques across all inverters. RFR is visualised in a) and c, whereas the physical models are illustrated in b and d. The difference between the two is that the period from 01.12.2017 to 30.01.2018 has been filtered out in figure c and d, due to the high amount of low voltage events across the whole site, which could imply bad weather.*

As seen from the figure, most of the events are low current. The high events of low current could be due to a large number of strings connected to a single inverter. In the articles referred to, the rated power is between 1 $kW_p$ and 4 $kW_p$, whereas the inverters in this thesis have a rated power of 800$KW_p$. As a result, the inverters will operate differently. Whereas an inverter with a couple of strings would decrease the voltage due to a fault to maximise the I-V curve, the same behaviour does not necessarily apply when over 150 strings are connected in parallel. In this case, a reduction in voltage would reduce the power output from all strings. Therefore, to maximise the power output from the entire array, the voltage would remain normal. Thus, the current from the faulty strings would decrease.

### 4.5.10 Applications

By comparing the current event detection from the different models in the thesis, the models are able to detect power losses on an inverter level. The methods have been shown capable to detect possible soiling and curtailments that have not been filtered out of the data. Furthermore, RFR shows less uncertainty and has been able to detect decreases in CR more accurately than the physical models after the model training period. However, these methods are not able to detect two, or four faulty strings. As an example, inverter 4 had two major reductions in CR for SM8 C1 and SM8 C3 starting at 30.12.2016 with a mean daily CR of approximately 0 for each string-pair. The subsequent days the CR increased to approximately 0.5 and lasted to 08.03.2017 when the mean CR decreased to under 0.1 for both string-pair. The subsequent day, SM8 C1 was fixed whereas SM8 C3 had a mean daily CR of approximately 0 until reparations on 20.06.2017.

Within this period, the events for the physical models were low string the whole day of 30.12.2016 but vanished afterwards for the physical models. Similarly, RFR detected events of low current and low current and voltage on 30.12.2016 and vanished afterwards. The events are due to a CR value of approximately 0 of all the 8 current channels connected to SM8. The CR value increases for all string-pairs except the two mentioned above. Thus, the detection methods on inverter level do not seem able to detect events based upon one or two faulty string-pairs. Although the results from RFR is highly biased in the training period. It should be noticed that the models were not able to detect low current events due to the faulty string after the model training period either. Furthermore, due to the faults being present in the model training it will be hard to detect this in the testing period. However, the models can detect suboptimal performance on inverter level due to lower performance on multiple strings. As an example, the increased events of low current in Figure 4.19 could indicate soiling. In addition, there are multiple examples where the CR of many strings decreases and the detection methods on inverter level recognises the event.

The sensitivity of event detection is based upon the threshold values presented in Figure 3.1b and Figure 3.2 and calculated as the model uncertainty plus the maximum measurement uncertainty. Although this thesis assumes that Chebyshev's inequality holds, for calculations of the model uncertainty, this is not necessarily the situation. A solution could be to estimate the model error as the error which consumes 80% of the estimated model errors with the use of quantiles. A better solution could be to tune the threshold value based upon known faults that the power plant operators aim to detect after the model training period.

Furthermore, it seems that common fault diagnosis tools do not necessarily apply to utility-scale solar power plants. The tools utilise the ratio between the estimated and measured values for current and voltage to classify possible faults (Chouder & Silvestre, 2010) (Livera, et al., 2018). The results from this thesis indicate that the inverters favour a reduction in current in a faulty string, instead of a reduction in voltage across all strings. In this thesis, the percentage of events detected as low voltage is in between 1% and 12% for all inverters when the period December.2017 and January.2018 has been filtered out.

The reduction in current for each string enables string monitoring, whereas the mean current ratio is calculated for each day and visualised for each string. Thus, enabling easy access to the performance of each string comparable with other strings for the same inverter. When the CR decreases for all the strings the power loss factors could be soiling, suboptimal performance by the inverter or unfiltered curtailments made by the power plant controller. Furthermore, to evaluate the performance of each string separately the relative current ratio

(RCR) is calculated. By evaluation of the development in RCR, power losses factors for each string-pair can be detected.

If the RCR decreases steadily through time, possible faults could be module cracking and hot spots as seen for inverter 8 SM8 C3 and SM9 C8, amongst other reasons. The behaviour of decreasing performance is typical for cracked cells and hot spots. The temperature variation under illumination and under no illumination will lead to expansion and compression of the material. Thus, the cracks can increase which can lead to a further reduction in the RCR. In contrast, a sudden decrease in RCR could be due to open-circuits and breakage of the entire module amongst other reasons. Furthermore, the evaluation of the mean CR from all the strings trough time enables detection of soiling. This trend can also be seen by analysis of the frequency in low current events as shown in Figure 4.25 by RFR, whereas a high frequency in the winter and spring months under dry seasons could indicate soiling.

The advantages of the proposed methods show application for industry usage, whereas knowledge about soiling and possible faulty strings have been derived from the data presented. The ideas presented in this thesis are to provided power plant operators with warnings about system behaviour different than normal. In addition, the methods presented provide a quantified value of the abnormalities. However, it will be up to the operators themselves to decide maintenance interventions. As a result, the methods presented will be additional set up for decision making.

# 5 Conclusion

Over a five-year time span considered in this thesis, both physical- and machine learning models have been applied to model PV systems consisting of over 150 strings connected to an inverter. The results with explanatory variables consisting of solar irradiance and solar cell temperature were that models based upon machine learning had a substantial increase in predictive ability for power, current and voltage input to the inverter. Furthermore, the solar irradiance was then calibrated for pyranometer drifting. The results from the drift calculations show a difference in drift values compared to the calibration report. When both drift values are applied to the solar irradiance, the standard deviation of MAE from the machine learning models decreases with calibration based upon the reported values. By the introduction of new features such as inverter temperature, parameters for sky conditions and others, the machine learning models experienced an increase in predictive ability. As a result, the best machine learning model had an average uncertainty between the average and maximum measurement uncertainty for power, voltage and current.

The models were later used to detect power loss events based upon reductions in power, current and voltage. The result from this analysis was that the models were able to detect power loss which could correspond to soiling in the winter and spring. Event detection was also able to detect decreases in power output across multiple strings. However, the technique was not able to detect instances where one to four strings were disconnected due to the low impact on the total performance of the inverter. Most of the events detected were classified as events due to low current. Therefore, it is plausible that common fault classification techniques which leverage the ratio between the simulated and measured voltage and current do not necessarily apply in utility-scale power plants. Due to the inverters aim to maximise the power production across over 150 strings, a decrease in voltage will affect all strings. In contrast to letting a faulty string generate less current as seen in this thesis. For this reason, an evaluation of the current from each string pair was made.

To evaluate the performance of strings the current ratio (CR) was established as the measured current divided by the estimated current. In addition, the relative current ratio (RCR) was defined as the CR divided by the median CR of all the strings connected to the same inverter. By analysing the evolution of the RCR trough time, it is possible to detect thermal defects on the string-pairs and the results were verified with infrared images. The four of the top five thermal defects from infrared images could also be seen by the proposed method, due to a decrease in RCR in the time period before the infrared images were taken.

# 6 Further work

- Increase the predictive ability of the machine learning models by an expansion of the hyperparameter space. Furthermore, it is interesting to see if other features can increase the predictive ability of the models. Another effect which can increase the predictive power of some models is to use a weighted average for solar irradiance instead of the uniform average. The weights could be based upon the distance to each of the weather stations.

- Experiment with the threshold value for solar irradiance for all the models, since it has been shown to increase the predictive ability of the physical models.

- Generate a method to weigh in the importance of previously detected events to give an estimate of what the reason for future events could be.

- Instead of analysing the period between 08:00 and 13:00, the analysis should be based upon data in a given threshold between the solar noon. Also, the models should be trained upon data only within this period.

- Work with inverter monitoring rather than the classification of events. Inverter monitoring can be done by evaluation of the ratio between measured and estimated power from each inverter. In addition, a relative ratio between all inverters should be established as the suggested RCR in this thesis.

- Experimental testing on utility-scale solar power plant inverters and inverters with a smaller scale to verify the suspicion that inverter in utility-scale solar power plants tends to favour a low current if a fault is present.

# References

EKO Instruments , 19. *MS-802 Pyranometer.* s.l., https://eko-eu.com/products/solar-energy/pyranometers/ms-802-pyranometer .

ALQahtani, A. H., S. Abuhamdeh, M. & Alsmadi, Y. M., 2012. A simplified and comprehensive approach to characterize photovoltaic system performance. *IEEE,* 17 September, pp. 1-6.

Angiulli, F. a. P. C., 2002. Fast Outlier Detection in High Dimensional Spaces. *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 15-27.

Anon., 2019. *pvlighthouse.* [Online]
Available at:
https://www2.pvlighthouse.com.au/resources/optics/spectrum%20library/spectrum%20library.aspx
[Accessed 24 March 2019].

Arno Smets, K. J. O. I. R. v. S. M. Z., 2016. *Solar Energy - The physics and engineering of photovoltaic conversion technologies and systems.* 1. ed. England: UIT Cambride, England.

Authors of scikit-learn, 2019. *github.com/scikit-learn.* [Online]
Available at: https://github.com/scikit-learn/scikit-learn/pull/13269
[Accessed 20 March 2019].

Betti, A. et al., 2017. PREDICTIVE MAINTENANCE IN PHOTOVOLTAIC PLANTS WITH A BIG DATA APPROACH. *33rd European Photovoltaic Solar Energy Conference and Exhibition (EUPVSEC),* 25-29 September, pp. 1895-1900 .

Bird, L., Cochran, J. & Wang, X., 2014. *Wind and Solar Energy Curtailment: Experience and Practices in the United States,* s.l.: National Renewable Energy.

Brehaut, C., 2016. *Megawatt-scale PV O&M and asset management 2016-2021.* Greentech Media Research, USA. : Wood mackenzie.

Burger, W. & Burge, M. J., 2016. *Digital Image Processing: An Algorithmic Introduction Using Java.* 2. ed. s.l.:Springer-Verlag London.

Bydenergy, 2011. *Datasheet BYD P6–30 Series-3BB,* Shanghai: Bydenergy.

Cawley, G. C. & Talbot, N. L. C., 2011. *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation,* Norwich, United Kingdom: Journal of Machine Learning Research 11.

Chen, C. J., 2011. *Physics of Solar Energy.* New Jersey: John Wiley & Sons, INC.

Chollet, F., 2018. *Deep learning with Python.* 1 ed. Shelter Island, NY : Manning Publications Co..

Chouder, A. & Silvestre, S., 2010. Automatic supervision and fault detection of PV systems based on Power losses analysis. *Energy Conversion and Managment,* 20 march, pp. 1929-1937.

Climate Analytics, 2018. *climateanalytics.* [Online]
Available at: https://climateanalytics.org/briefings/ratification-tracker
[Accessed 10 April 2019].

De Benedetti, M. et al., 2018. Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing*, 3 may, pp. 59-68.

Dìez-Mediavilla, M., Saldaña-Mayor, D., Dieste-Velasco, M. & Peña, D. G. C.-T., 2014. DEGRADATION IN SILICON PV MODULES: FAULT DISTRIBUTION ANALYSIS IN GRID CONNECTED. *29th European Photovoltaic Solar Energy Conference and Exhibition*, pp. 2536 - 2540.

Dobos, A. P., 2014. *PVWatts Version 5 Manual,* s.l.: National Renewable Energy Laboratory (NREL).

Dubey, S., Narotam, S. & Seshadri, B., 2013 . Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World A Review. *Energy Procedia 33* , p. 311 – 321 .

Evans, D. L., 1981. SIMPLIFIED METHOD FOR PREDICTING. *Solar Energy Vol. 27, No. 6*, 6 July, pp. 555-560.

Green, M. & Eyal, B., 2017. *Fault Prediction Using Clustering Algorithms,* s.l.: Report IEA - PVPS T13-07:2017.

Hastie, T., Tibshirani, R. & Friedman, j., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*. 2 ed. s.l.:Springer.

Havens, K. J. & Sharp, E. J., 2016. Chapter 7 - Thermal Imagers and System Considerations. In: *Thermal Imaging Techniques to Survey and Monitor Animals in the Wild.* s.l.:Elsevier Inc, pp. 101-119.

Hawkins, D. M., 1980. *Identification of Outliers.* London: Chapman and Hall.

Herrmann, W., Wiesner, W. & Vaaßen, W., 1997. HOT SPOT INVESTIGATIONS ON PV MODULES - NEW CONCEPTS FOR A TEST STANDARD AND CONSEQUENCES FOR MODULE DESIGN WITH RESPECT TO BYPASS DIODES. *Conference Record of the Twenty Sixth IEEE Photovoltaic Specialists Conference*, pp. 1129-1132.

Hinckley, A., 2017. *Campbell Scientific Pyranometers: What You Need to Know.* [Online]
Available at: https://www.campbellsci.com/blog/pyranometers-need-to-know
[Accessed 1 april 2019].

Holmgren, W. F., Hansen, C. W. & Mikofski, M. A., 2018. pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software, 3(29)*, 7 september.

iac.ethz.ch, 2019. *Institute for Atmospheric and Climate Science ETH Zürich.* [Online]
Available at: http://www.iac.ethz.ch/group/climate-and-water-cycle/research/radiation-and-the-hydrological-cycle/global-dimming-and-brightening.html
[Accessed 02 05 2019].

IEA, 2017. *CO2 emissions from fuel combustion,* s.l.: International Energy Agency.

IEA, 2017. *Global Energy & CO2 Status Report* , s.l.: International Energy Agency.

IEA, 2018. *iea.org.* [Online]
Available at: https://www.iea.org/geco/emissions/
[Accessed 11 April 2019].

International Energy Agency, 2017. *Global Energy & CO2 Status Report,* s.l.: IEA.

IPN, 2018. *Solution and strategies for operation and maintenance of utility-scale solar power plants.* Oslo: s.n.

Joint Committee for Guides in Metrology, 2008. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement,* s.l.: s.n.

Jordan, D. C. a. K. S. R., 2011. *Photovoltaic Degradation Rates—an Analytical Review,* Golden, CO 80401, USA: Wiley Online Library.

Jordan, D. C., Kurtz, S. R. V. K. & Newmiller, J., 2016. Compendium of photovoltaic degradation rates. *PROGRESS IN PHOTOVOLTAICS: RESEARCH AND APPLICATIONS*, 7 February, p. 978–989.

Kajari-Schroder, S., Kunze, I., Eitner, U. & Kontges, M., 2011. Spatial and orientational distribution of cracks in crystalline photovoltaic modules generated by mechanical load tests. *Solar Energy Materials & Solar Cells*, 19 July , p. 3054–3059.

Kimber, A., Mitchell, L. N. S. & Wenger, H., 2006. *THE EFFECT OF SOILING ON LARGE GRIDCONNECTED PHOTOVOLTAIC SYSTEMS IN CALIFORNIA AND THE SOUTHWEST REGION OF THE UNITED STATES,* Berkeley : PowerLight Corporation.

King, D. L., Boyson, W. E. & A., K. J., 2004. *Photovoltaic Array Performance Model,* Albuquerque, New Mexico 87185 and Livermore, California 94550 : Sandia National Laboratories .

Kipp & Zonen, n.d. *Pyranometers v. Reference Cells for PV Installations,* s.l.: Kipp & Zonen.

Laboratories, Sandia National, 2018. *pvpmc sandia.* [Online]
Available at: https://pvpmc.sandia.gov/modeling-steps/1-weather-design-inputs/plane-of-array-poa-irradiance/
[Accessed 27 March 2019].

Livera, A. et al., 2018. *Failure diagnosis of short- and open-circuit fault conditions in PV systems Department of Electrical and Computer Engineering, University of Cyprus ,* Nicosia, Cyprus: FOSS Research Centre for Sustainable Energy, Photovoltaic Technology Laboratory.

Lumby, B., 2015. *Utility-scale solar photovoltaic power plants: A project developer's guide.* Washington: International Finance Corporation .

Mani, M. & Pillai, R., 2010. *Impact of dust on solar photovoltaic (PV) performance: Research status, challenges and recommendations,* Bangalore: Renewable and Sustainable Energy Reviews.

Montgomery, D. C. & Runger, G. C., 2014. *Applied Statistics and Probability for Engineers.* 6. ed. Arizona: Wiley.

Myers, D. R., 2013. *Solar radiation: Practical modeling for renewable Energy applications.* 1. ed. London New York: Taylor & Francis Group, LLC.

Niclas, D. W., 2011. *sinovoltaics.com.* [Online]
Available at: https://sinovoltaics.com/learning-center/quality/standard-test-conditions-stc-definition-and-problems/
[Accessed 13 March 2019].

NRL, 2019. *National Renewable Energy Laboratory (NREL).* [Online]
Available at: https://www.nrel.gov/grid/solar-resource/assets/data/astmg173.xls
[Accessed 24 March 2019].

Olsen, p. E., 2019. *Pyranometer drifting* [Interview] (1 April 2019).

Parr, T., Turgutlu, K., Csizar, C. & Howard, J., 2018. *explained.a.* [Online]
Available at: https://explained.ai/rf-importance/index.html
[Accessed 17 April 2019].

Pedersen, H. B., 2015. *Experimental Study of soilin on photovoltaic modules in a Nordic climate,* Kjeller: Master thesis.

Pedregosa, F. et al., 2019. *github.* [Online]
Available at: https://github.com/scikit-learn/scikit-learn/tree/master/sklearn
[Accessed 25 April 2019 ].

Perez, R. et al., 2002. A NEW OPERATIONAL MODEL FOR SATELLITE-DERIVED IRRADIANCES:. *Solar Energy Vol. 73, No. 5* , 13 November, pp. 307–317,.

Platon, R., Martel, J. W. N. & Chau, T. Y., 2015. Online Fault Detection in PV Systems. *IEEE TRANSACTIONS ON SUSTAINABLE ENERGY, VOL. 6, NO. 4,* , OCTOBER , pp. 1200-1207.

Rachka, S. & Mirjalili, V., 2017. *Python Machine Learning.* 2. ed. Birmingham, UK: Packt Publishing.

Ramaprabha, R. & Mathur, B., 2008. Modelling and Simulation of Solar PV Array under Partial Shaded conditions. *ICSET*, pp. 7-11.

Raschka, S., 2018. *rasbt.github.io.* [Online]
Available at:
http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/#sequential-feature-selector
[Accessed 17 April 2019].

Raschka, S., 2018. *rasbt.github.io.* [Online]
Available at:
http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/#sequential-feature-selector
[Accessed 17 April 2019].

Reda, I. &. A. A., 2004. *Solar Position Algorithm for Solar Radiation Applications* , Cole Boulevard Golden, Colorado: National Renewable Energy Laboratory.

Reno, M. J. & Hanse, C. W., 2016. Identification of periods of clear sky irradiance in time series of GHI measurements. *Renewable Energy, v90,* 18 january, pp. 520-531.

Reno, M. J., Hanse, C. W. & Stein, J. S., 2012. *Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis,* Albuquerque, New Mexic: Sandia National Laboratories.

Rmaprabha, R. & Dr. Mathur, B., 2009. Impact of Partial Shading on Solar PV Module Containing Series Connected Cells. *International Journal of Recent Trends in Engineering, Vol 2, No. 7,* , November, pp. 56-60.

Rodrigues, S., Ramos, H. G. & Morgado-Dias, F., 2018. Machine Learning in PV Fault Detection, Diagnostics and Prognostics: A Review. *IEEE 44th Photovoltaic Specialist Conference (PVSC),* 5 November, pp. 3178-3183.

SMA Solar Technology AG, 2013. *Sunny Central CP - Deviations of measured values.* NA: SMA Solar Technology AG.

SMA Solar Technology AG, NA. *String-Monitor Unit Monitoring of string-current for Sunny Central inverters with a maximum input voltage of 1,100 V.* NA: SMA.

Solar Power Europe, 2018. *Global Market Outlook For Solar Power / 2018 - 2022,* s.l.: Solar Power Europe.

Sopori, B. et al., 2012. Understanding Light-Induced Degradation of c-Si Solar Cells. *Conference Record of the IEEE Photovoltaic Specialists Conference* , June.

Størdal, A., 2013. *pv system design and yield simulations for a farm,* Ås: NMBU.

Taylor, J. R., 1997. *An introduction to Error Analysis - the study of uncertainties in physical measurments.* 2. ed. Sausalito, California: University Science Books.

UNFCCC, U. N. F. C. o. C. C., 2015. *unfccc.* [Online]
Available at:
https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf
[Accessed 10 April 2019].

Varma, S. & Richard, S., 2006. *Bias in error estimation when using cross-validation for model selection,* Bethesda MD, USA: Biometric Research Branch, National Cancer Institute, .

Vollmer, M. & Møllman, 2010. *Infrared Thermal Imaging Fundamentals, Research and Applications.* Brandenburg: Wiley-vch Verlag GmbH & Co. KGaA.

Wier, J. T. &. T., 2006. *Renewable Energy Resources.* 2. ed. London: Taylor & Francis.

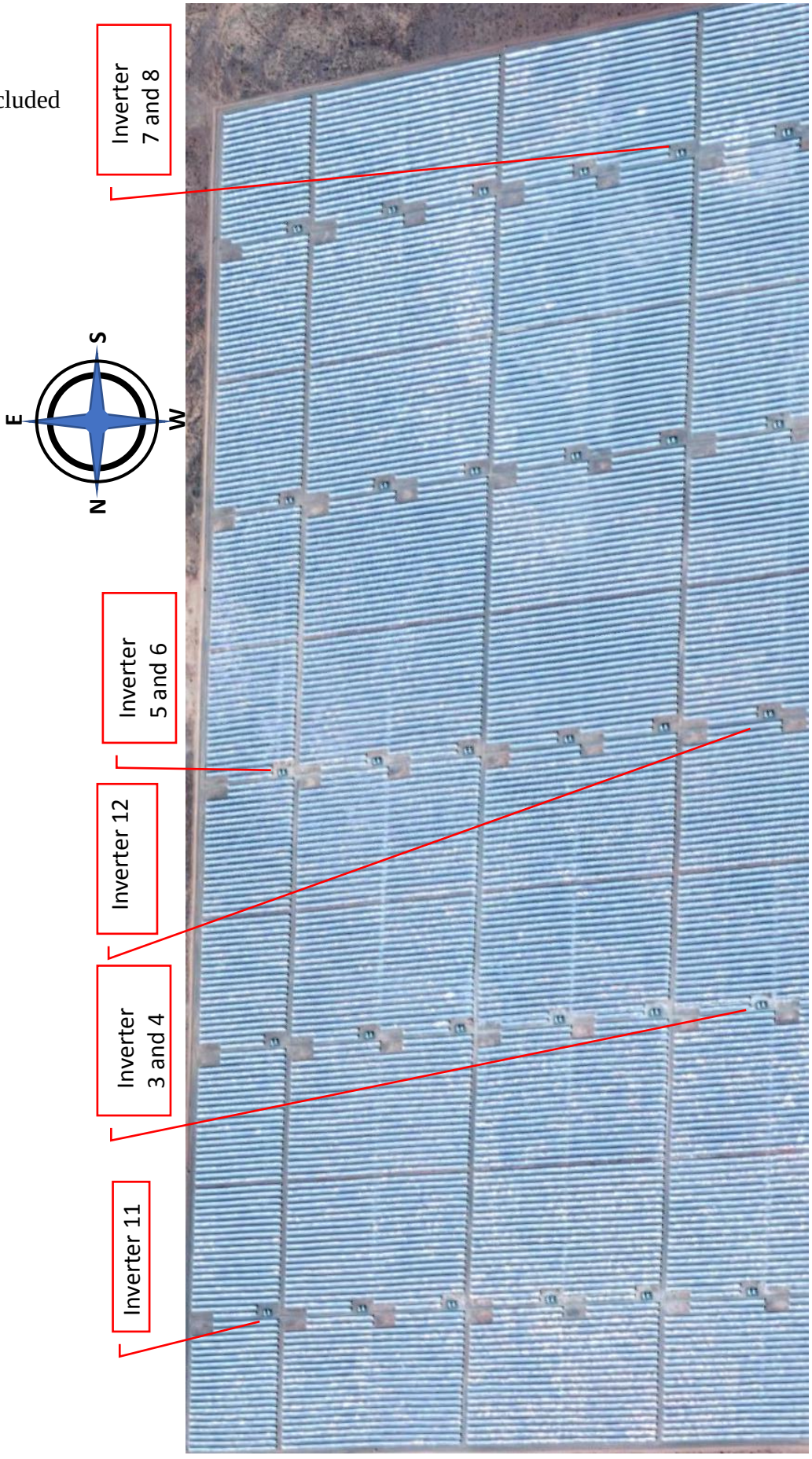
Zhao, Y., Nasrullah, Z. & Li, Z., 2019. *PyOD: A Python Toolbox for Scalable Outlier Detection,* Cornell : Cornell University.

Ødgaard, M. B., Haug, H. & Selj, J., 2018. METHODS FOR QUALITY CONTROL OF MONITORING DATA FROM COMMERCIAL PV SYSTEMS. *35th European Photovoltaic Solar Energy Conference and Exhibition*, pp. 2083- 2088.

Øgaard, M. B., 2016. *Effect of Soiling on the Performance of Photovoltaic Modules in Kalkbult, South Africa*, Ås: NMBU - Master thesis.

# Appendix A

**Site**
Whole site not included due to anonymity



# Appendix B

## Power

*Table 0-1 Mean absolute error after nested cross-validation applied to the inverters for power input to the inverter. The input to the models is solar irradiance and solar cell temperature. The values are in units of [W].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 20612.5 ± 162.8 | 20612.5 ± 162.8 | 11874.6 ± 295.5 | 12028.7 ± 251.9 | 22900.6 ± 171.4 |
| **INV2** | 20318.6 ± 347.4 | 20318.6 ± 347.4 | 11616.7 ± 220.7 | 11760.9 ± 123.2 | 22747.8 ± 478.9 |
| **INV3** | 21011.6 ± 308.9 | 21011.6 ± 308.9 | 11567.4 ± 146.4 | 11709.1 ± 200.7 | 23495.0 ± 551.8 |
| **INV4** | 20612.1 ± 290.1 | 20612.1 ± 290.1 | 11942.4 ± 127.8 | 12162.5 ± 197.2 | 25011.2 ± 266.3 |
| **INV5** | 20892.5 ± 442.6 | 20892.3 ± 442.3 | 13065.2 ± 225.5 | 13204.4 ± 321.0 | 25181.1 ± 340.3 |
| **INV6** | 20274.9 ± 434.1 | 20273.7 ± 434.3 | 12595.2 ± 239.4 | 12719.3 ± 186.5 | 24294.8 ± 265.2 |
| **INV7** | 19809.1 ± 238.4 | 19809.1 ± 238.4 | 11772.3 ± 195.2 | 11908.7 ± 289.1 | 22926.1 ± 366.6 |
| **INV8** | 19984.7 ± 194.1 | 19984.7 ± 194.1 | 12340.9 ± 182.1 | 12455.8 ± 175.6 | 22538.8 ± 179.4 |
| **INV9** | 20822.2 ± 275.3 | 20822.2 ± 275.3 | 12124.9 ± 95.4 | 12152.3 ± 93.6 | 25289.7 ± 238.3 |
| **INV10** | 20639.4 ± 307.0 | 20639.4 ± 306.9 | 12153.7 ± 127.9 | 12225.2 ± 183.1 | 24422.1 ± 498.6 |
| **INV11** | 21437.1 ± 398.0 | 21437.1 ± 398.0 | 12482.3 ± 88.3 | 12578.8 ± 44.7 | 25085.0 ± 383.0 |
| **INV12** | 20447.8 ± 193.0 | 20447.7 ± 193.0 | 11708.6 ± 160.9 | 11808.1 ± 247.2 | 22593.8 ± 300.3 |

## Current

*Table 0-2 Mean absolute error after nested cross-validation applied to the inverters for current input to the inverter. The input to the models is solar irradiance and solar cell temperature. The values are in units of [A].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 19.3 ± 0.18 | 19.3 ± 0.18 | 17.79 ± 0.49 | 18.08 ± 0.31 | 29.83 ± 0.35 |
| **INV2** | 19.25 ± 0.33 | 19.25 ± 0.33 | 17.56 ± 0.28 | 17.95 ± 0.18 | 28.82 ± 0.52 |
| **INV3** | 17.52 ± 0.31 | 17.52 ± 0.31 | 16.24 ± 0.3 | 16.69 ± 0.3 | 30.08 ± 0.71 |
| **INV4** | 17.5 ± 0.22 | 17.5 ± 0.22 | 16.73 ± 0.24 | 17.08 ± 0.29 | 32.05 ± 0.32 |
| **INV5** | 18.25 ± 0.32 | 18.25 ± 0.33 | 17.68 ± 0.36 | 18.02 ± 0.47 | 32.2 ± 0.29 |
| **INV6** | 17.33 ± 0.33 | 17.33 ± 0.33 | 16.67 ± 0.37 | 16.97 ± 0.16 | 30.78 ± 0.41 |
| **INV7** | 16.76 ± 0.24) | 16.76 ± 0.24 | 16.04 ± 0.28 | 16.4 ± 0.43 | 29.58 ± 0.44 |
| **INV8** | 17.15 ± 0.34 | 17.14 ± 0.34 | 16.44 ± 0.29 | 16.71 ± 0.21 | 28.79 ± 0.17 |
| **INV9** | 18.45 ± 0.34 | 18.46 ± 0.34 | 17.53 ± 0.26 | 17.78 ± 0.29 | 33.21 ± 0.42 |
| **INV10** | 18.05 ± 0.34 | 18.05 ± 0.34 | 17.25 ± 0.33 | 17.43 ± 0.27 | 32.32 ± 0.51 |
| **INV11** | 19.28 ± 0.45 | 19.28 ± 0.45 | 18.32 ± 0.34 | 18.59 ± 0.2 | 33.03 ± 0.64 |
| **INV12** | 17.12 ± 0.33 | 17.12 ± 0.33 | 15.74 ± 0.2 | 16.05 ± 0.24 | 29.36 ± 0.44 |

## Voltage

*Table 0-3 Mean absolute error after nested cross-validation applied to the inverters for voltage input to the inverter. The input to the models is solar irradiance and solar cell temperature. The values are in units of [V].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 11.03 ± 0.12 | 11.03 ± 0.12 | 6.22 ± 0.11 | 6.17 ± 0.11 | 13.15 ± 0.23 |
| **INV2** | 10.91 ± 0.03 | 10.91 ± 0.02 | 6.01 ± 0.12 | 5.99 ± 0.13 | 13.75 ± 0.15 |
| **INV3** | 10.81 ± 0.19 | 10.81 ± 0.19 | 6.13 ± 0.07 | 6.09 ± 0.05 | 13.73 ± 0.24 |
| **INV4** | 11.28 ± 0.17 | 11.28 ± 0.17 | 6.0 ± 0.09 | 5.95 ± 0.14 | 13.89 ± 0.27 |
| **INV5** | 10.99 ± 0.14 | 10.99 ± 0.14 | 6.47 ± 0.14 | 6.42 ± 0.14 | 13.65 ± 0.12 |
| **INV6** | 10.96 ± 0.15 | 10.96 ± 0.15 | 6.4 ± 0.12 | 6.37 ± 0.09 | 13.43 ± 0.15 |
| **INV7** | 10.59 ± 0.19 | 10.59 ± 0.19 | 6.08 ± 0.15 | 6.06 ± 0.15 | 12.73 ± 0.16 |
| **INV8** | 10.64 ± 0.1 | 10.64 ± 0.1 | 6.2 ± 0.14 | 6.19 ± 0.17 | 12.53 ± 0.23 |
| **INV9** | 10.59 ± 0.19 | 10.59 ± 0.19 | 5.73 ± 0.07 | 5.71 ± 0.09 | 12.93 ± 0.35 |
| **INV10** | 10.7 ± 0.15 | 10.7 ± 0.15 | 5.83 ± 0.09 | 5.79 ± 0.11 | 12.56 ± 0.3 |
| **INV11** | 11.67 ± 0.12 | 11.67 ± 0.12 | 6.45 ± 0.15 | 6.41 ± 0.14 | 13.78 ± 0.23 |
| **INV12** | 10.63 ± 0.18 | 10.63 ± 0.18 | 6.15 ± 0.14 | 6.14 ± 0.18 | 12.51 ± 0.28 |

## Drift- Power

*Table 0-4 Mean absolute error after nested cross-validation without SFS applied to the inverters for power input to the inverter. The drift is based upon own calculations. The values are in units of [W].*

| | Lasso | Ridge | RFR | KN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 21122.43 ± 229.7 | 21122.43 ± 229.7 | 12635.94 ± 151.32 | 12746.28 ± 94.32 | 23040.54 ± 144.0 |
| **INV2** | 20729.08 ± 351.8 | 20729.08 ± 351.8 | 12313.78 ± 194.29 | 12457.33 ± 149.48 | 22869.64 ± 339.95 |
| **INV3** | 21322.41 ± 304.0 | 21322.41 ± 303.99 | 12002.7 ± 143.9 | 12155.27 ± 149.94 | 23548.68 ± 321.18 |
| **INV4** | 21016.27 ± 209.92 | 21016.27 ± 209.92 | 12643.76 ± 162.88 | 12799.18 ± 214.87 | 25097.1 ± 227.41 |
| **INV5** | 21244.18 ± 228.87 | 21243.94 ± 228.88 | 13515.7 ± 128.7 | 13724.41 ± 130.7 | 25208.48 ± 624.02 |
| **INV6** | 20618.11 ± 243.43 | 20618.66 ± 244.33 | 13162.1 ± 197.83 | 13272.21 ± 174.38 | 24333.56 ± 152.29 |
| **INV7** | 20151.75 ± 373.82 | 20153.73 ± 370.84 | 12424.24 ± 97.3 | 12542.63 ± 62.16 | 23038.63 ± 381.69 |
| **INV8** | 20328.64 ± 100.68 | 20328.47 ± 101.08 | 12898.69 ± 168.18 | 13041.94 ± 224.87 | 22630.95 ± 504.89 |
| **INV9** | 21297.16 ± 156.28 | 21297.16 ± 156.28 | 13001.48 ± 200.34 | 13036.99 ± 166.69 | 25463.25 ± 263.08 |
| **INV10** | 21108.12 ± 269.9 | 21108.12 ± 269.9 | 12999.04 ± 206.19 | 13087.73 ± 216.82 | 24605.53 ± 444.72 |
| **INV11** | 21805.35 ± 129.82 | 21805.35 ± 129.82 | 13237.88 ± 240.24 | 13286.12 ± 267.48 | 25239.78 ± 532.42 |
| **INV12** | 20894.71 ± 199.9 | 20894.71 ± 199.9 | 12490.51 ± 149.47 | 12519.58 ± 187.45 | 22798.05 ± 184.82 |

*Table 0-5 Mean absolute error after nested cross-validation without SFS applied to the inverters for power input to the inverter. The drift is based upon the calibration report. The values are in units of [W].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 20614.1 ± 102.04 | 20614.1 ± 102.04 | 11922.99 ± 179.27 | 12075.3 ± 145.95 | 22904.17 ± 297.88 |
| **INV2** | 20317.25 ± 349.1 | 20317.25 ± 349.1 | 11607.88 ± 212.83 | 11761.05 ± 135.14 | 22745.57 ± 480.08 |
| **INV3** | 21011.58 ± 308.88 | 21011.58 ± 308.88 | 11567.71 ± 142.89 | 11709.09 ± 200.71 | 23494.96 ± 551.83 |
| **INV4** | 20612.08 ± 290.14 | 20612.08 ± 290.14 | 11944.43 ± 106.24 | 12162.47 ± 197.16 | 25011.18 ± 266.25 |
| **INV5** | 20702.97 ± 287.71 | 20702.89 ± 287.69 | 12738.3 ± 97.38 | 12986.07 ± 86.4 | 25113.9 ± 391.38 |
| **INV6** | 20107.19 ± 184.33 | 20105.71 ± 183.57 | 12238.74 ± 211.3 | 12460.95 ± 190.43 | 24233.29 ± 253.17 |
| **INV7** | 20100.81 ± 232.4 | 20100.26 ± 231.4 | 12283.96 ± 69.23 | 12459.75 ± 27.44 | 23026.01 ± 699.54 |
| **INV8** | 20281.64 ± 102.5 | 20282.01 ± 102.45 | 12810.4 ± 143.42 | 12943.11 ± 229.65 | 22621.46 ± 504.95 |
| **INV9** | 20864.56 ± 123.78 | 20864.56 ± 123.78 | 12165.14 ± 128.29 | 12209.29 ± 144.19 | 25319.5 ± 452.3 |
| **INV10** | 20676.52 ± 414.84 | 20676.52 ± 414.84 | 12236.09 ± 77.77 | 12314.95 ± 64.79 | 24459.57 ± 661.29 |
| **INV11** | 21467.29 ± 169.2 | 21467.29 ± 169.2 | 12464.33 ± 108.57 | 12650.7 ± 120.76 | 25111.14 ± 126.08 |
| **INV12** | 20477.0 ± 181.92 | 20477.0 ± 181.92 | 11777.47 ± 293.46 | 11933.64 ± 252.72 | 22622.57 ± 238.78 |

## Drift- Current

*Table 0-6 Mean absolute error after nested cross-validation without SFS applied to the inverters for current input to the inverter. The drift is based upon own calculations. The values are in units of [A].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 18.72 ± 0.33 | 18.72 ± 0.33 | 17.68 ± 0.24 | 17.9 ± 0.27 | 29.78 ± 0.13 |
| **INV2** | 18.44 ± 0.24 | 18.44 ± 0.24 | 17.41 ± 0.29 | 17.74 ± 0.25 | 28.79 ± 0.48 |
| **INV3** | 17.88 ± 0.22 | 17.88 ± 0.22 | 16.66 ± 0.17 | 17.03 ± 0.21 | 30.1 ± 0.35 |
| **INV4** | 18.33 ± 0.46 | 18.33 ± 0.46 | 17.61 ± 0.38 | 17.96 ± 0.39 | 32.19 ± 0.15 |
| **INV5** | 18.84 ± 0.28 | 18.84 ± 0.28 | 18.29 ± 0.37 | 18.56 ± 0.34 | 32.21 ± 0.69 |
| **INV6** | 17.94 ± 0.31 | 17.94 ± 0.31 | 17.34 ± 0.28 | 17.58, 0.29 | 30.75 ± 0.17 |
| **INV7** | 17.28 ± 0.46 | 17.28 ± 0.46 | 16.79 ± 0.27 | 17.04 ± 0.2 | 29.66 ± 0.42 |
| **INV8** | 17.6 ± 0.14 | 17.59 ± 0.15 | 17.03 ± 0.27 | 17.28 ± 0.27 | 28.79 ± 0.63 |
| **INV9** | 19.16 ± 0.38 | 19.17 ± 0.38 | 18.58 ± 0.28 | 18.65 ± 0.31 | 33.42 ± 0.49 |
| **INV10** | 18.72 ± 0.41 | 18.72 ± 0.41 | 17.98 ± 0.41 | 18.17 ± 0.42 | 32.49 ± 0.75 |
| **INV11** | 19.69 ± 0.5 | 19.69 ± 0.5 | 19.16 ± 0.5 | 19.42 ± 0.49 | 33.14 ± 0.79 |
| **INV12** | 17.71 ± 0.23 | 17.71 ± 0.23 | 16.66 ± 0.25 | 16.74 ± 0.28 | 29.55 ± 0.24 |

*Table 0-7 Mean absolute error after nested cross-validation without SFS applied to the inverters for current input to the inverter. The drift is based upon the calibration report. The values are in units of [A].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 19.2 ± 0.14 | 19.2 ± 0.14 | 17.8 ± 0.37 | 18.12 ± 0.29 | 29.81 ± 0.4 |
| **INV2** | 19.14 ± 0.33 | 19.14 ± 0.33 | 17.53 ± 0.26 | 17.89 ± 0.19 | 28.81 ± 0.52 |
| **INV3** | 17.52 ± 0.31 | 17.52 ± 0.31 | 16.25 ± 0.3 | 16.69 ± 0.3 | 30.08 ± 0.71 |
| **INV4** | 17.5 ± 0.22 | 17.5 ± 0.22 | 16.73 ± 0.2 | 17.08 ± 0.29 | 32.05 ± 0.32 |
| **INV5** | 18.04 ± 0.27 | 18.04 ± 0.27 | 17.47 ± 0.22 | 17.87 ± 0.24 | 32.12 ± 0.46 |
| **INV6** | 17.11 ± 0.25 | 17.11 ± 0.25 | 16.46 ± 0.28 | 16.72 ± 0.29 | 30.72 ± 0.37 |
| **INV7** | 17.18 ± 0.22 | 17.18 ± 0.22 | 16.63 ± 0.18 | 16.95 ± 0.15 | 29.64 ± 0.8 |
| **INV8** | 17.52 ± 0.14 | 17.51 ± 0.15 | 16.96 ± 0.24 | 17.19 ± 0.29 | 28.79 ± 0.62 |
| **INV9** | 18.49 ± 0.36 | 18.49 ± 0.36 | 17.58 ± 0.31 | 17.82 ± 0.29 | 33.25 ± 0.47 |
| **INV10** | 18.07 ± 0.44 | 18.08 ± 0.44 | 17.27 ± 0.18 | 17.49 ± 0.16 | 32.36 ± 0.78 |
| **INV11** | 19.28 ± 0.25 | 19.28 ± 0.25 | 18.27 ± 0.17 | 18.66 ± 0.17 | 33.06 ± 0.39 |
| **INV12** | 17.12 ± 0.48 | 17.12 ± 0.48 | 15.86 ± 0.62 | 16.15 ± 0.56 | 29.39 ± 0.52 |

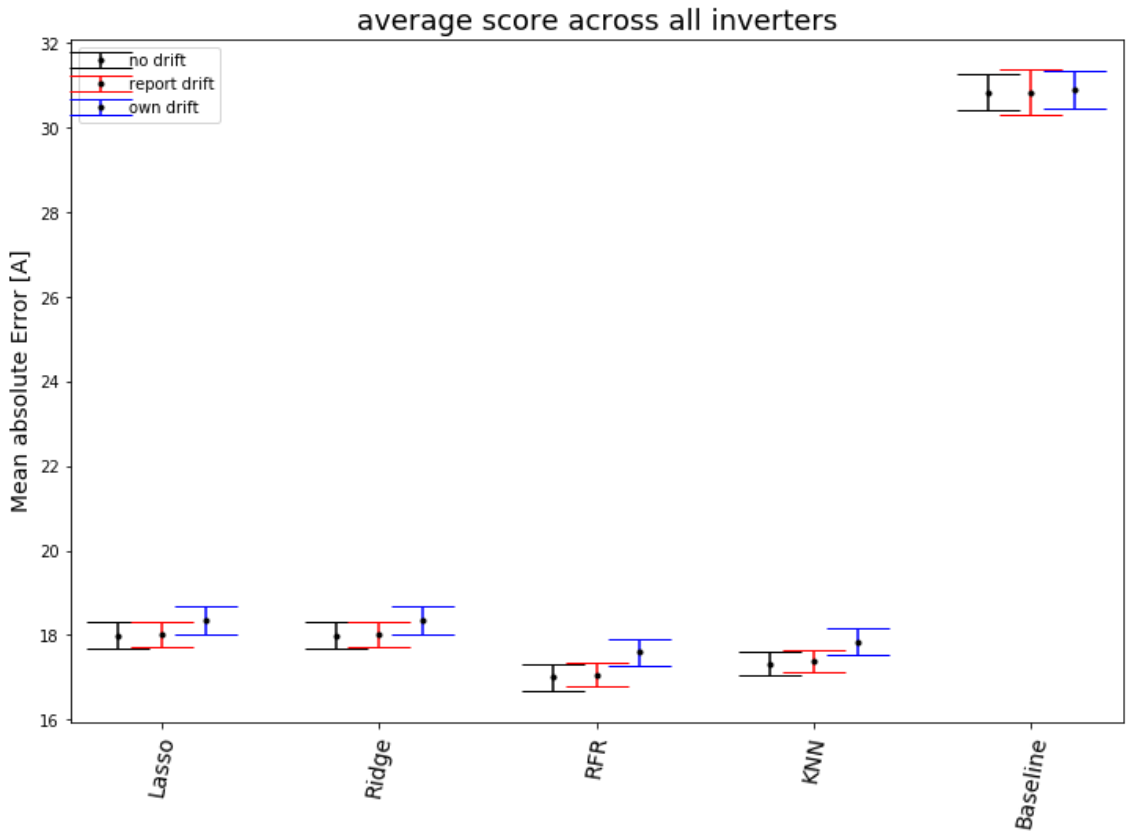



Figure 0.1 Illustration of the mean MAE across all inverters and the mean uncertainty as the errors in the plot for the five different models applied to the inverter current input. For all models three scores has been found for irradiance without calibration for drift in black (no drift), for irradiance calibration based upon own calculated drift in blue (own drift) and irradiance calibration based upon reported drift in red (report drift)

## Voltage

*Table 0-8 Mean absolute error after nested cross-validation without SFS applied to the inverters for voltage input to the inverter. The drift is based upon own calculations. The values are in units of [V].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 11.09 ± 0.15 | 11.09 ± 0.15 | 6.21 ± 0.04 | 6.18 ± 0.03 | 13.23 ± 0.31 |
| **INV2** | 10.96 ± 0.15 | 10.96 ± 0.15 | 6.06 ± 0.08 | 6.02 ± 0.08 | 13.84 ± 0.19 |
| **INV3** | 10.82 ± 0.17 | 10.82 ± 0.17 | 6.16 ± 0.04 | 6.08 ± 0.07 | 13.76 ± 0.33 |
| **INV4** | 11.31 ± 0.1 | 11.31 ± 0.1 | 6.03 ± 0.09 | 5.97 ± 0.1 | 13.92 ± 0.33 |
| **INV5** | 11.0 ± 0.17 | 11.0 ± 0.17 | 6.45 ± 0.04 | 6.39 ± 0.05 | 13.66 ± 0.31 |
| **INV6** | 10.96 ± 0.17 | 10.96 ± 0.17 | 6.42 ± 0.12 | 6.38 ± 0.14 | 13.43 ± 0.28 |
| **INV7** | 10.61 ± 0.11 | 10.61 ± 0.11 | 6.08 ± 0.08 | 6.05 ± 0.07 | 12.77 ± 0.14 |
| **INV8** | 10.66 ± 0.13 | 10.66 ± 0.13 | 6.19 ± 0.15 | 6.17 ± 0.14 | 12.56 ± 0.19 |
| **INV9** | 10.61 ± 0.21 | 10.61 ± 0.21 | 5.77 ± 0.13 | 5.74 ± 0.16 | 12.97 ± 0.24 |
| **INV10** | 10.71 ± 0.18 | 10.71 ± 0.18 | 5.85 ± 0.1 | 5.82 ± 0.08 | 12.58 ± 0.14 |
| **INV11** | 11.68 ± 0.19 | 11.68 ± 0.19 | 6.46 ± 0.07 | 6.41 ± 0.08 | 13.81 ± 0.18 |
| **INV12** | 10.64 ± 0.07 | 10.64 ± 0.07 | 6.2 ± 0.12 | 6.15 ± 0.09 | 12.53 ± 0.28 |

*Table 0-9 Mean absolute error after nested cross-validation without SFS applied to the inverters for voltage input to the inverter. The drift is based upon the calibration report. The values are in units of [V].*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 11.03 ± 0.16 | 11.03 ± 0.16 | 6.23 ± 0.1 | 6.17 ± 0.08 | 13.15 ± 0.24 |
| **INV2** | 10.91 ± 0.02 | 10.91 ± 0.02 | 6.01 ± 0.12 | 5.99 ± 0.13 | 13.75 ± 0.15 |
| **INV3** | 10.81 ± 0.19 | 10.81 ± 0.19 | 6.12 ± 0.06 | 6.09 ± 0.05 | 13.73 ± 0.24 |
| **INV4** | 11.28 ± 0.17 | 11.28 ± 0.17 | 5.99 ± 0.08 | 5.95 ± 0.14 | 13.89 ± 0.27 |
| **INV5** | 10.98 ± 0.1 | 10.98 ± 0.1 | 6.45 ± 0.09 | 6.4 ± 0.0 | 13.63 ± 0.23 |
| **INV6** | 10.96 ± 0.17 | 10.96 ± 0.17 | 6.44 ± 0.21 | 6.38 ± 0.18 | 13.41 ± 0.38 |
| **INV7** | 10.61 ± 0.2 | 10.61 ± 0.2 | 6.06 ± 0.07 | 6.04 ± 0.07 | 12.76 ± 0.35 |
| **INV8** | 10.66 ± 0.13 | 10.66 ± 0.13 | 6.2 ± 0.14 | 6.17 ± 0.14 | 12.56 ± 0.19 |
| **INV9** | 10.6 ± 0.13 | 10.6 ± 0.13 | 5.73 ± 0.11 | 5.71 ± 0.07 | 12.94 ± 0.27 |
| **INV10** | 10.7 ± 0.17 | 10.7 ± 0.17 | 5.86 ± 0.07 | 5.84 ± 0.04 | 12.57 ± 0.27 |
| **INV11** | 11.68 ± 0.12 | 11.68 ± 0.12 | 6.42 ± 0.04 | 6.4 ± 0.08 | 13.79 ± 0.39 |
| **INV12** | 10.63 ± 0.13 | 10.63 ± 0.13 | 6.17 ± 0.09 | 6.16 ± 0.06 | 12.51 ± 0.31 |

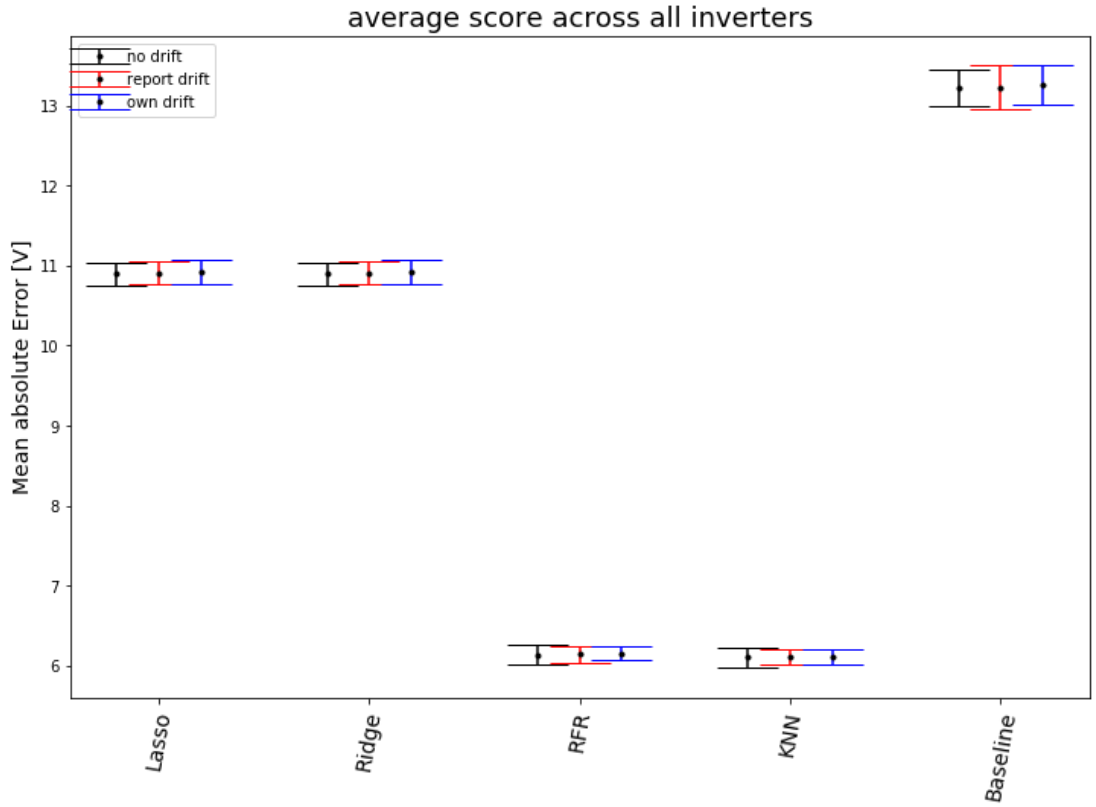



*Figure 0.2 Illustration of the mean MAE across all inverters and the mean uncertainty as the errors in the plot for the five different models applied to the inverter voltage input. For all models, three scores have been found for irradiance without calibration for drift in black (label: no drift), for irradiance calibration based upon own calculated drift in blue (label: own drift) and irradiance calibration based upon reported drift in red (label: report drift)*

## Feature engineering

## Power

*Table 0-10 Results after nested cross-validation with SFS after feature engineering applied on the voltage input to the inverter.*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 19072.9 ± 165.66 | 19072.89 ± 165.66 | 7497.03 ± 46.36 | 10752.64 ± 215.71 | 22904.17 ± 297.88 |
| **INV2** | 18436.92 ± 224.86 | 18437.02 ± 224.77 | 7303.41 ± 97.58 | 10335.09 ± 43.3 | 22745.57 ± 480.08 |
| **INV3** | 19090.28 ± 251.41 | 19083.14 ± 247.04 | 7634.07 ± 134.33 | 11149.72 ± 164.52 | 23494.96 ± 551.83 |
| **INV4** | 18303.04 ± 209.82 | 18299.85 ± 210.6 | 8212.12 ± 180.1 | 11540.58 ± 91.62 | 25011.18 ± 266.25 |
| **INV5** | 18670.27 ± 257.65 | 18663.02 ± 252.12 | 8632.71 ± 128.16 | 12355.62 ± 262.81 | 25113.9 ± 391.38 |
| **INV6** | 18007.19 ± 188.16 | 18007.19 ± 188.16 | 8270.2 ± 132.85 | 11937.61 ± 112.7 | 24233.29 ± 253.17 |
| **INV7** | 17789.49 ± 212.27 | 17789.11 ± 212.38 | 8107.82 ± 164.73 | 11779.53 ± 90.51 | 23026.01 ± 699.54 |
| **INV8** | 18017.88 ± 265.09 | 18017.88 ± 265.09 | 8004.06 ± 131.01 | 11813.31 ± 233.17 | 22621.46 ± 504.95 |
| **INV9** | 18642.0 ± 187.44 | 18642.01 ± 187.44 | 8852.28 ± 31.2 | 11605.07 ± 209.28 | 25319.5 ± 452.3 |
| **INV10** | 18340.82 ± 399.48 | 18340.83 ± 399.48 | 9079.07 ± 119.44 | 11729.87 ± 85.66 | 24459.57 ± 661.29 |
| **INV11** | 19541.67 ± 307.84 | 19534.78 ± 310.76 | 9443.45 ± 87.63 | 12111.54 ± 108.79 | 25111.14 ± 126.08 |
| **INV12** | 18271.87 ± 293.31 | 18276.76 ± 294.12 | 8172.35 ± 175.92 | 11229.74 ± 188.03 | 22622.57 ± 238.78 |

## Current

*Table 0-11 Results after nested cross-validation with SFS after feature engineering applied on the current input to the inverter. For computational expenses, only every second inverter has been calculated. Runtime: 14 hours on 7 CPU cores.*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 16.55 ± 0.17 | 16.54 ± 0.16 | 10.67 ± 0.33 | 14.51 ± 0.34 | 29.81 ± 0.4 |
| **INV3** | 15.6 ± 0.26 | 15.58 ± 0.26 | 10.09 ± 0.28 | 14.59 ± 0.34 | 30.08 ± 0.71 |
| **INV5** | 16.64 ± 0.38 | 16.64 ± 0.37 | 10.9 ± 0.26 | 15.48 ± 0.46 | 32.12 ± 0.46 |
| **INV7** | 15.31 ± 0.23 | 15.29 ± 0.25 | 10.54 ± 0.24 | 14.33 ± 0.38 | 29.64 ± 0.8 |
| **INV9** | 17.02 ± 0.47 | 16.96 ± 0.45 | 12.32 ± 0.21 | 16.36 ± 0.39 | 33.25 ± 0.47 |
| **INV11** | 18.0 ± 0.14 | 17.98 ± 0.15 | 13.39 ± 0.21 | 17.51 ± 0.39 | 33.06 ± 0.39 |

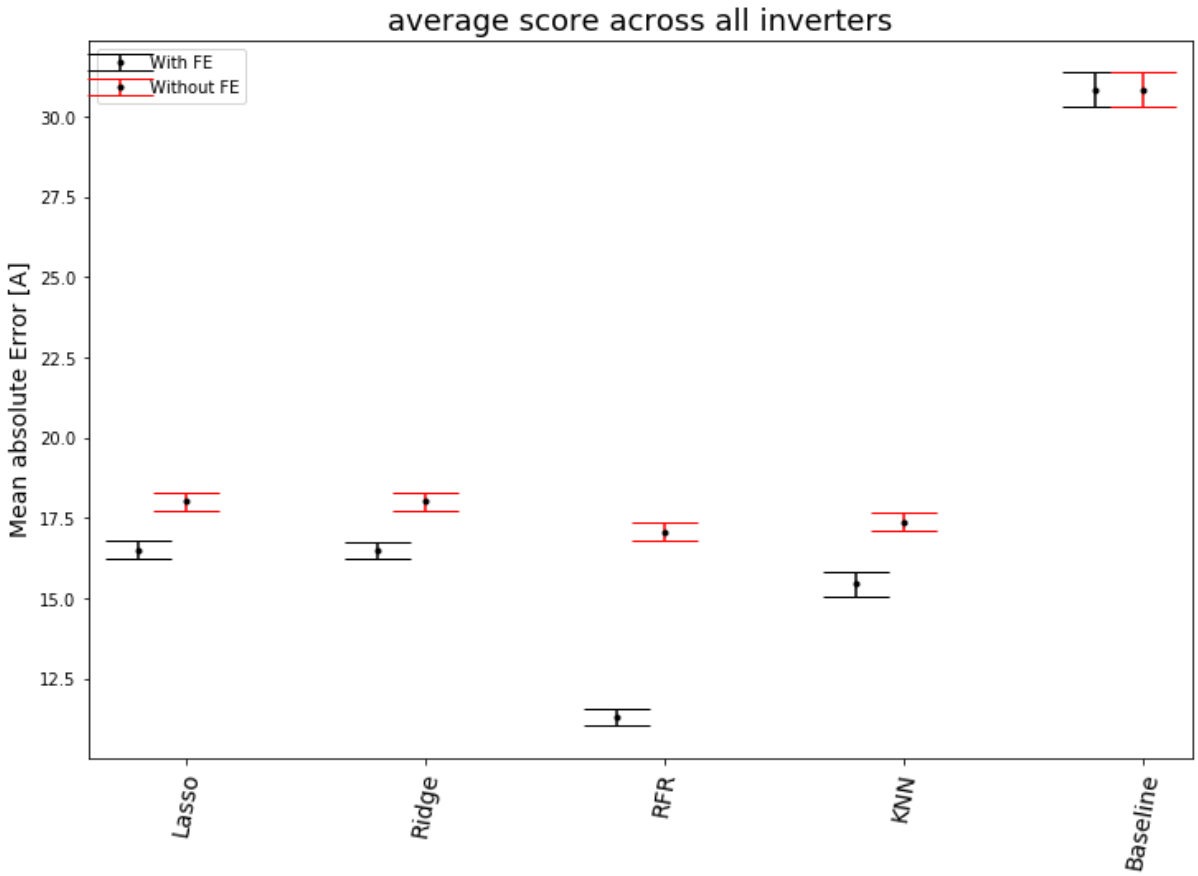



*Figure 0.3 Illustration of the score of each model with and without feature engineering applied to current input to the inverter. Since the baseline model only uses cell temperature and irradiance its performance does not change. However, all the other models show a significant performance improvement due to feature engineering.*

## Voltage

*Table 0-12 Results after nested cross-validation with SFS after feature engineering applied on the voltage input to the inverter. For computational expenses, only every second inverter has been calculated. Runtime: 14 hours on 7 CPU cores.*

| | Lasso | Ridge | RFR | KNN | Baseline |
|---|---|---|---|---|---|
| **INV1** | 9.09 ± 0.13 | 8.99 ± 0.12 | 4.92 ± 0.06 | 5.47 ± 0.07 | 13.73 ± 0.24 |
| **INV3** | 9.45 ± 0.08 | 9.29 ± 0.08 | 5.22 ± 0.04 | 5.81 ± 0.11 | 13.63 ± 0.23 |
| **INV5** | 9.13 ± 0.14 | 9.0 ± 0.18 | 4.82 ± 0.09 | 5.36 ± 0.03 | 12.76 ± 0.35 |
| **INV7** | 8.84 ± 0.12 | 8.78 ± 0.13 | 4.78 ± 0.09 | 5.33 ± 0.06 | 12.94 ± 0.27 |
| **INV9** | 10.01 ± 0.17 | 9.85 ± 0.11 | 5.12 ± 0.08 | 5.73 ± 0.05 | 13.79 ± 0.39 |

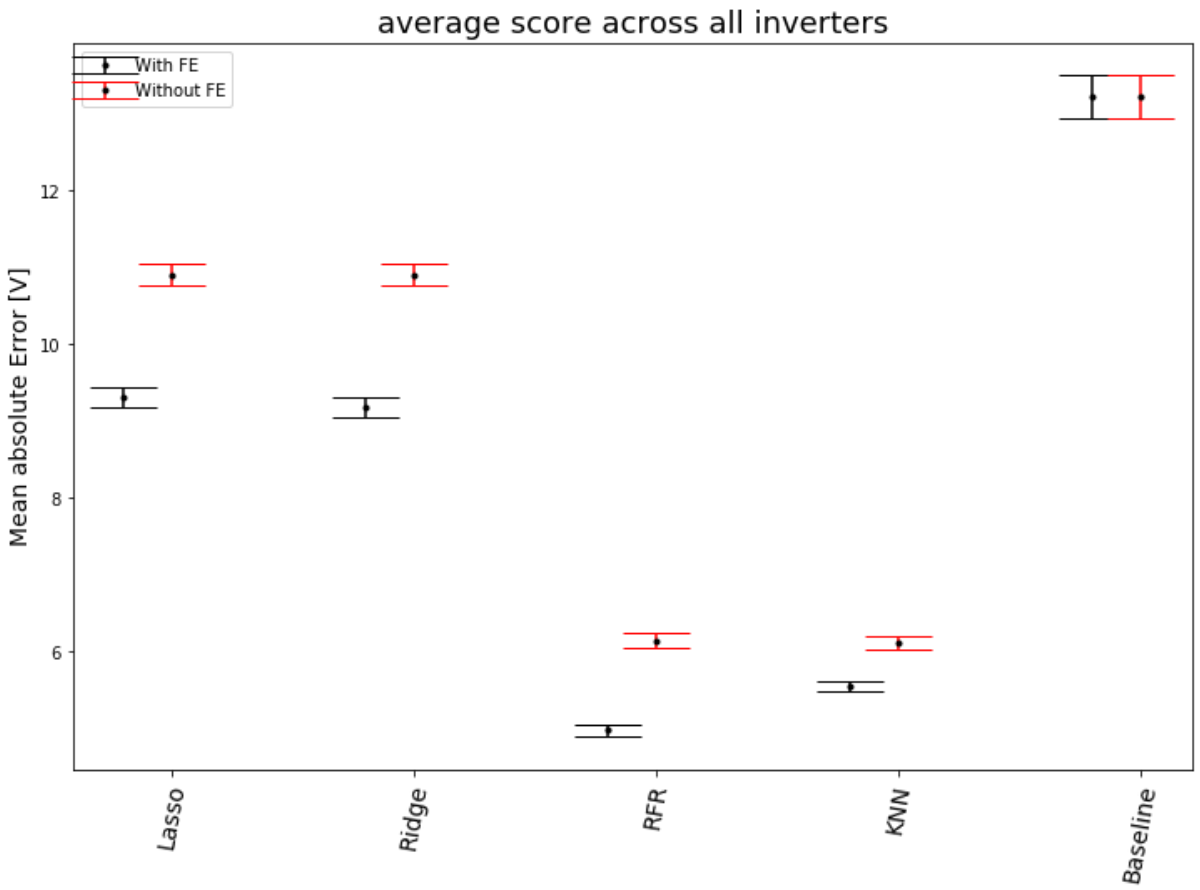



*Figure 0.4 Illustration of the score of each model with and without feature engineering applied to voltage input to the inverter. Since the baseline model only uses cell temperature and irradiance its performance does not change. However, all the other models show a significant performance improvement due to feature engineering.*

# Appendix C

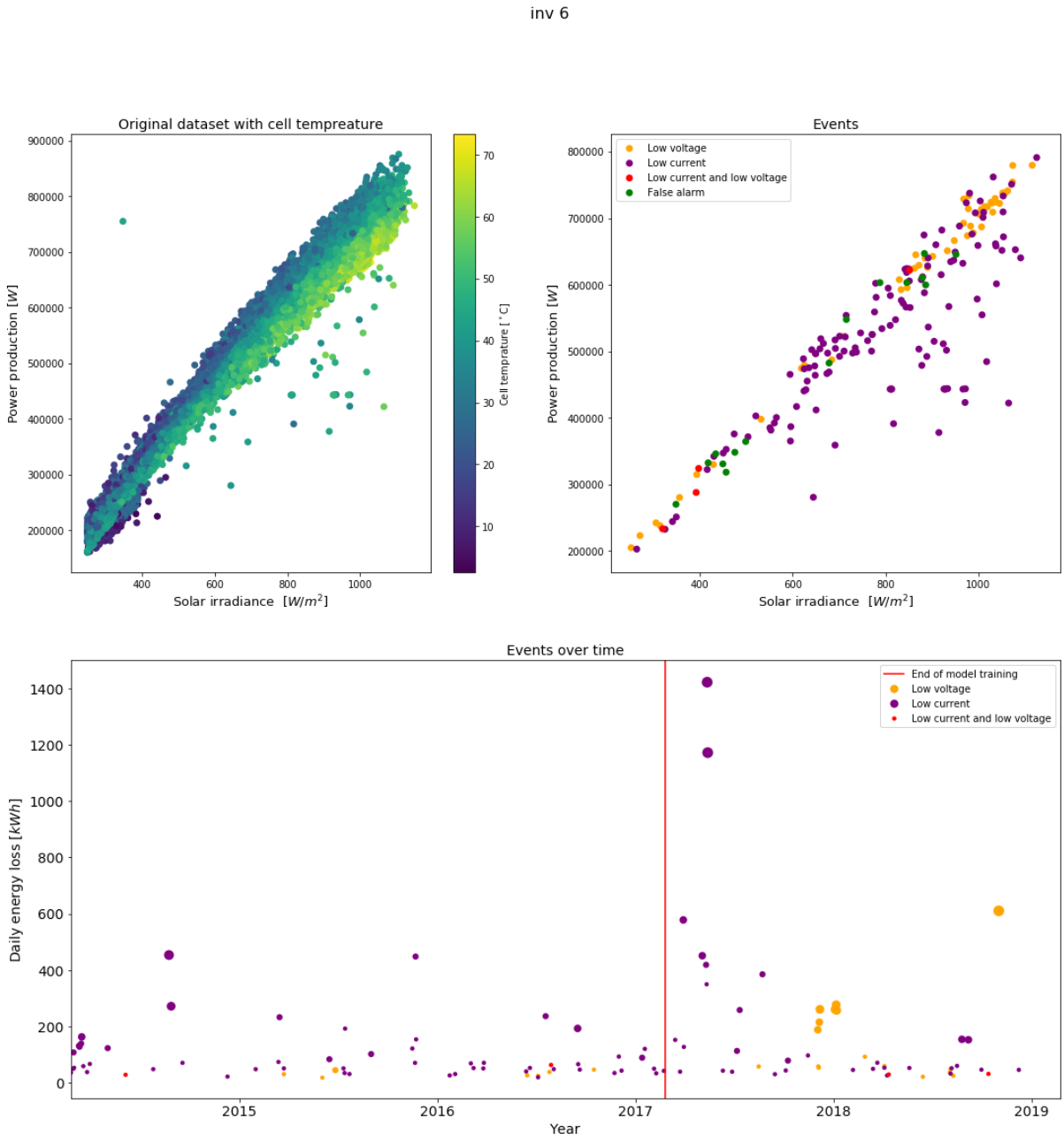



*Figure 0.1 Event detection with physical models and solar irradiance threshold of 250W/m² for Inverter 6.*

inv 6

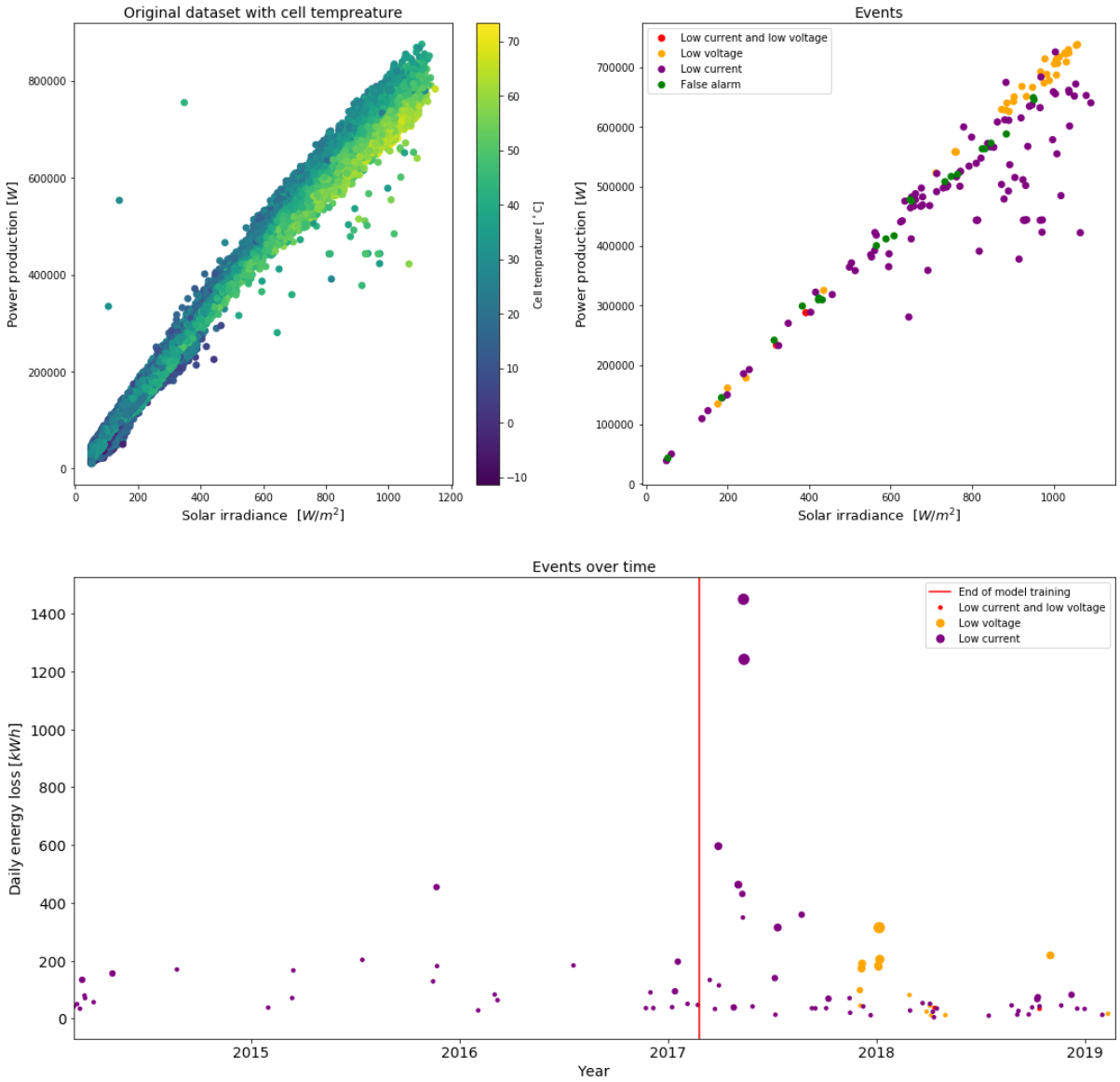



*Figure 0.2 Event detection with RFR and solar irradiance threshold of 50W/m² for Inverter 6*

Low voltage distribution -RFR

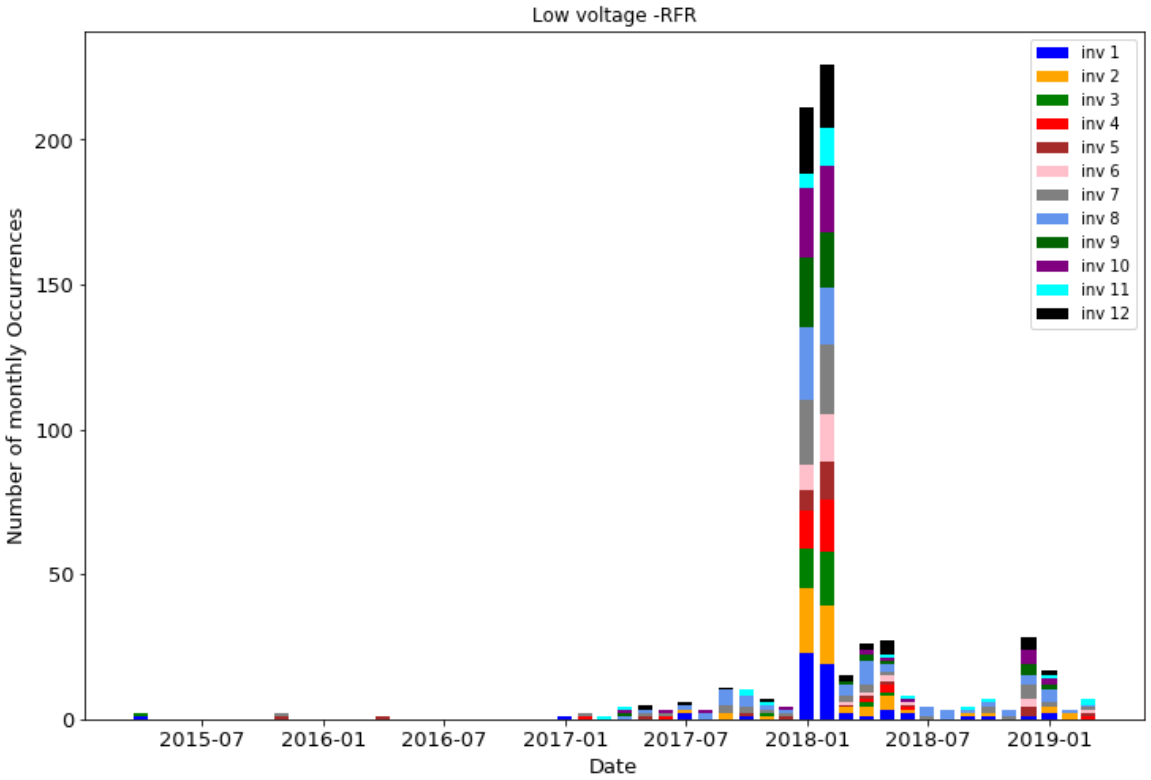



*Figure 0.3 Number of monthly occurrences of low voltage events for each inverter illustrated with different colour codes. The values on the y-axis is the sum of the occurrences each month.*

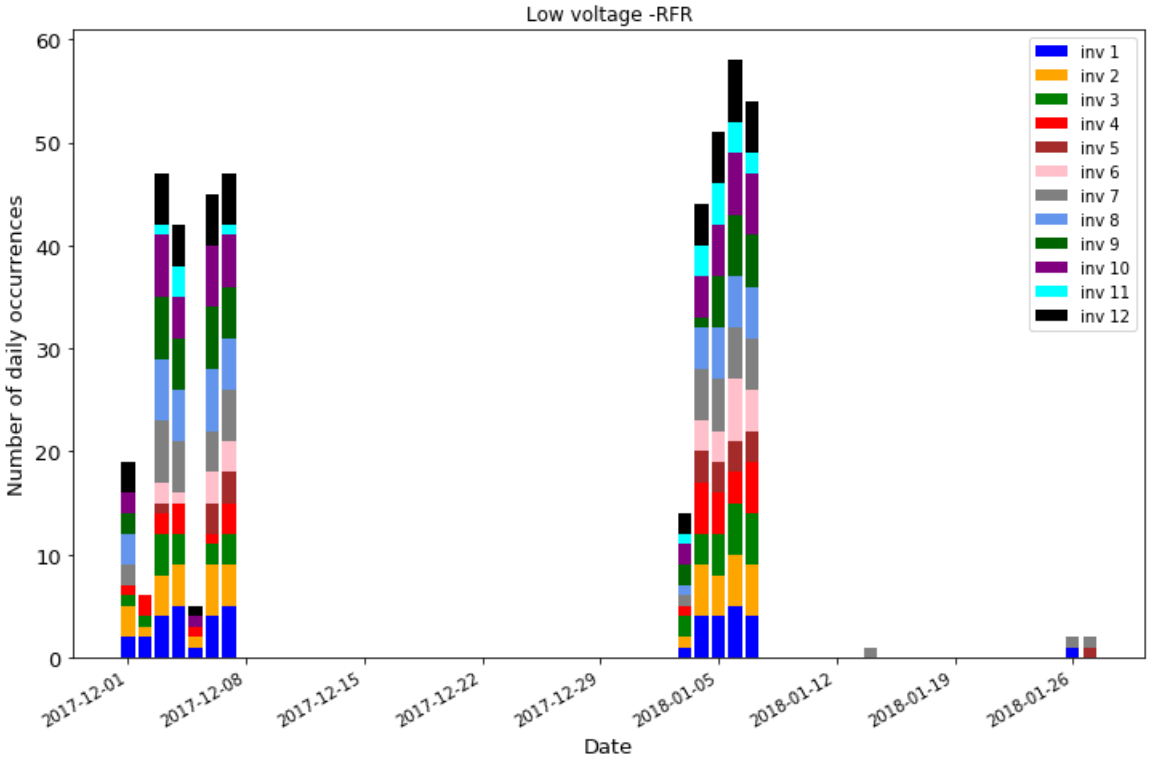
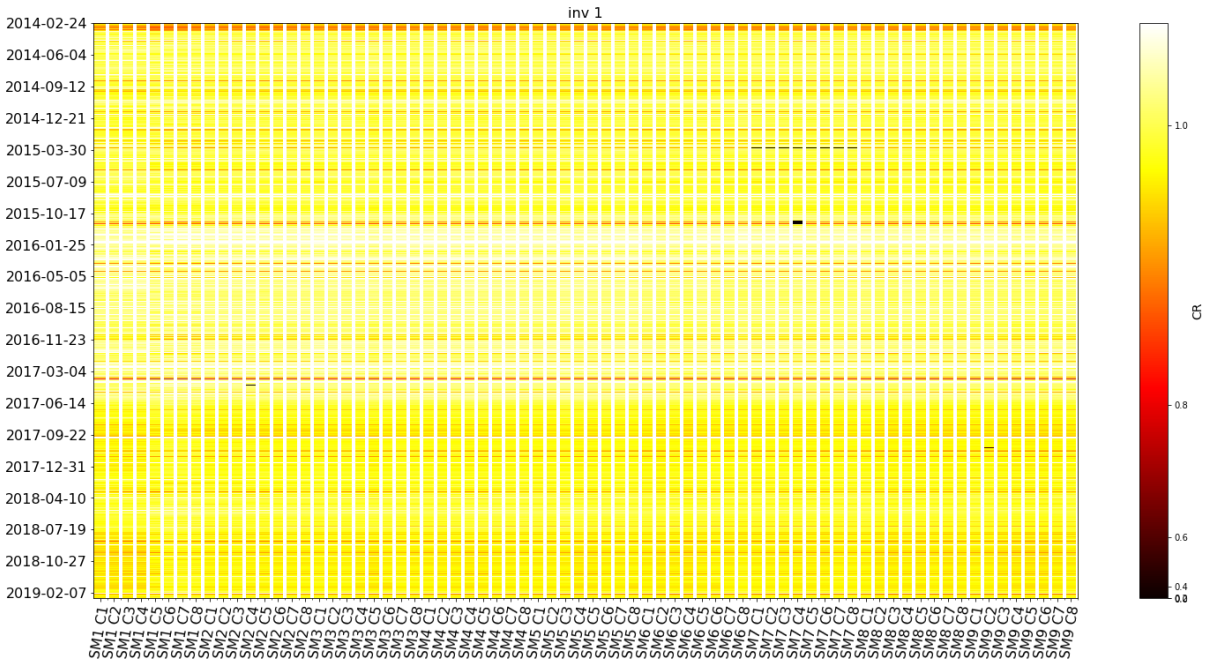
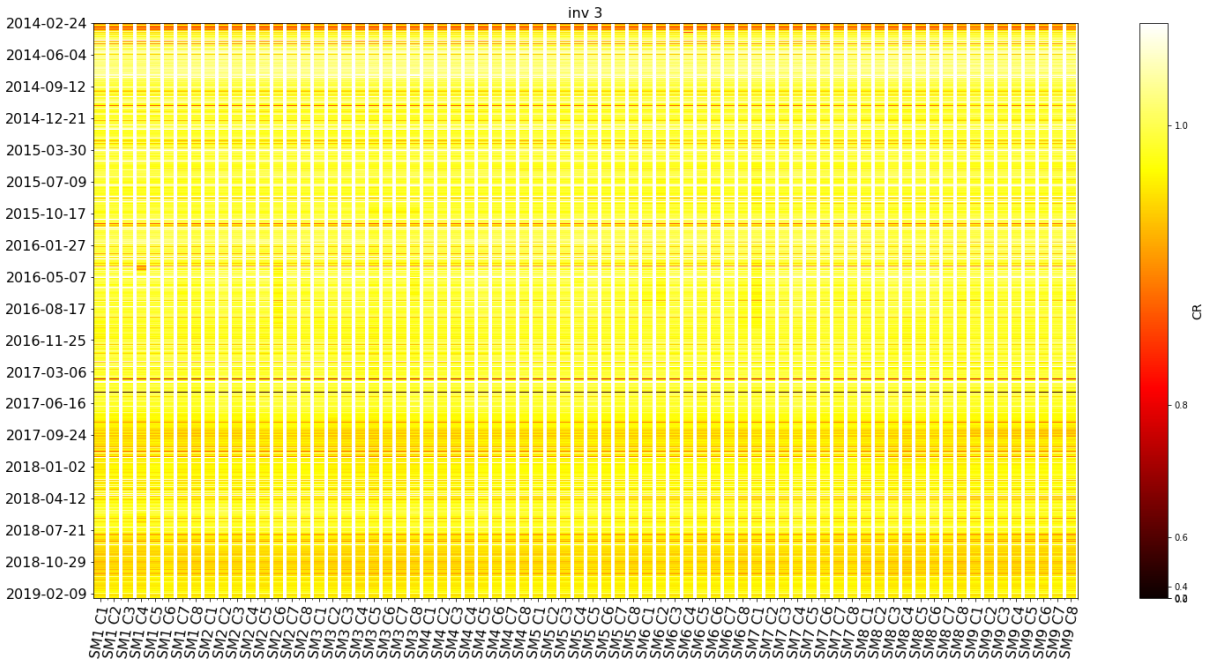


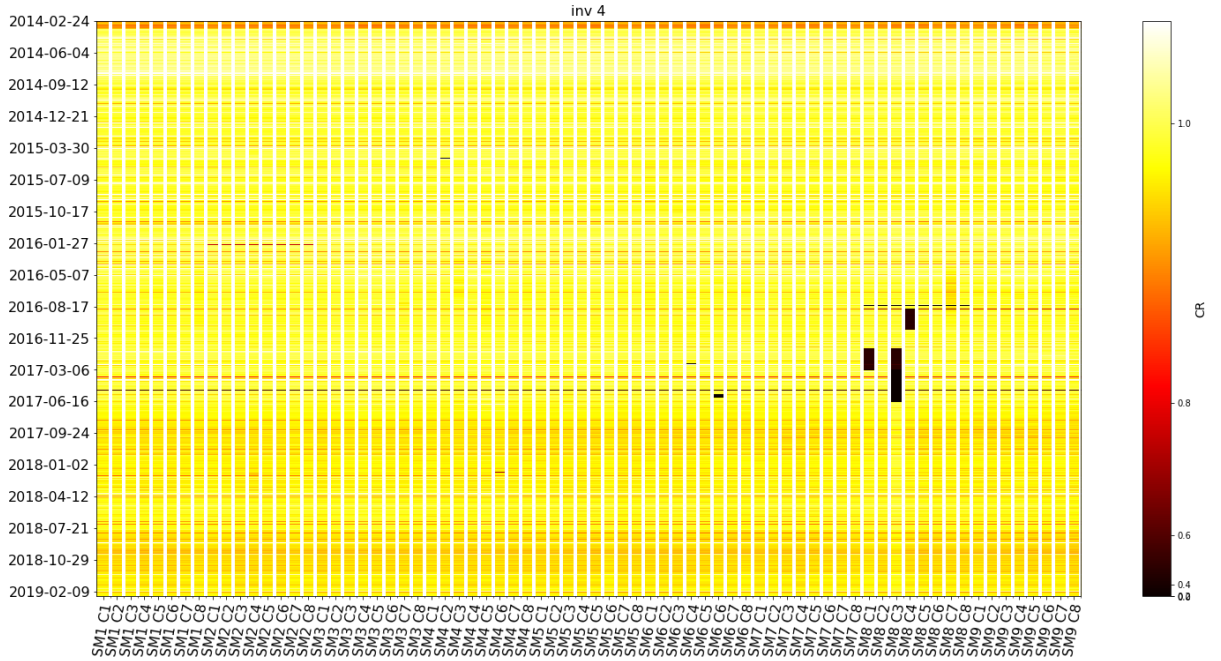

*Figure 0.4 Number of daily occurrences of low voltage events for each inverter illustrated with different colour codes. The values on the y-axis is the sum of the occurrences each day.*

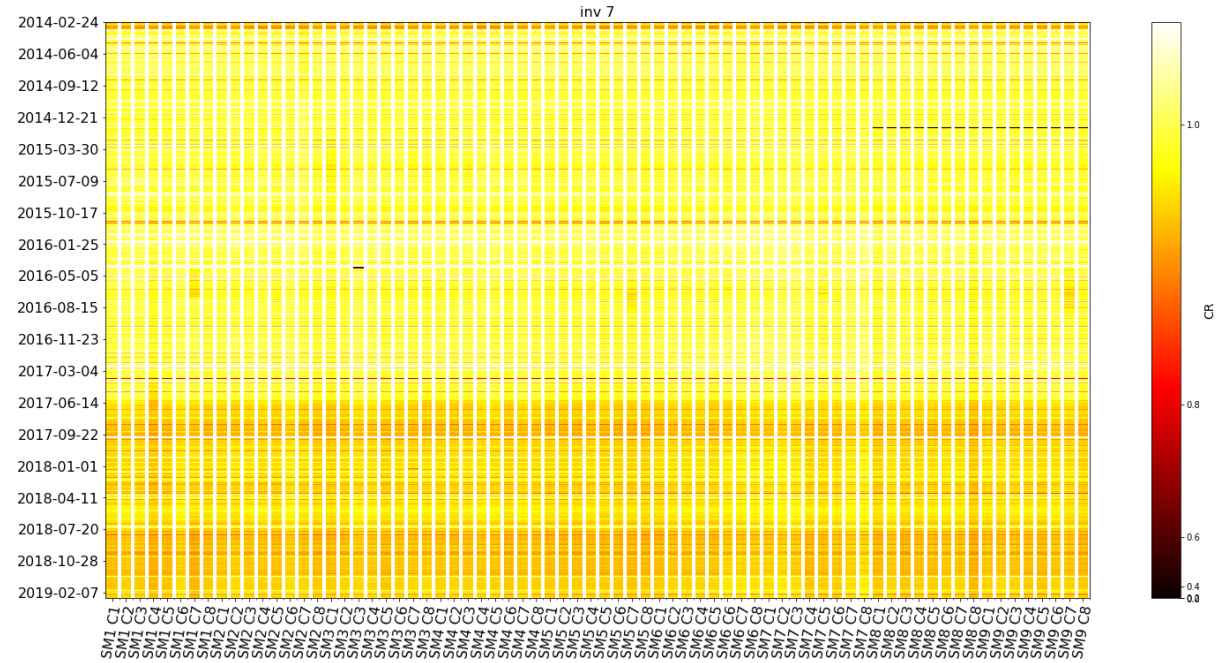
# Appendix D

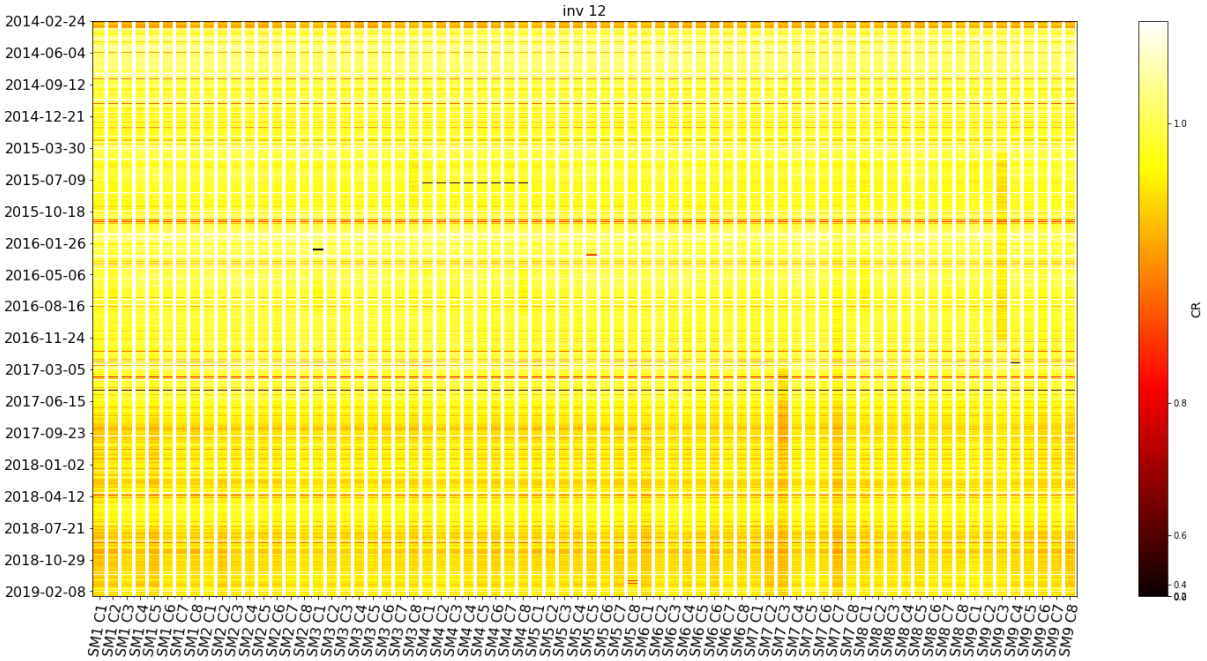
## CR for different inverters



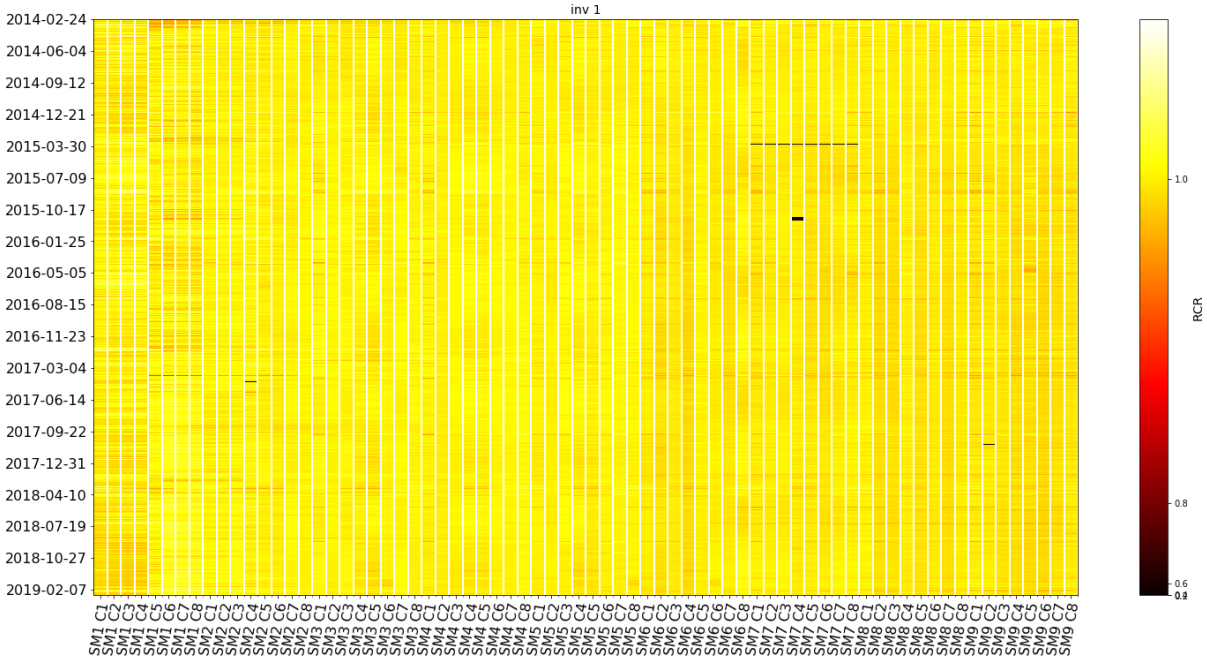



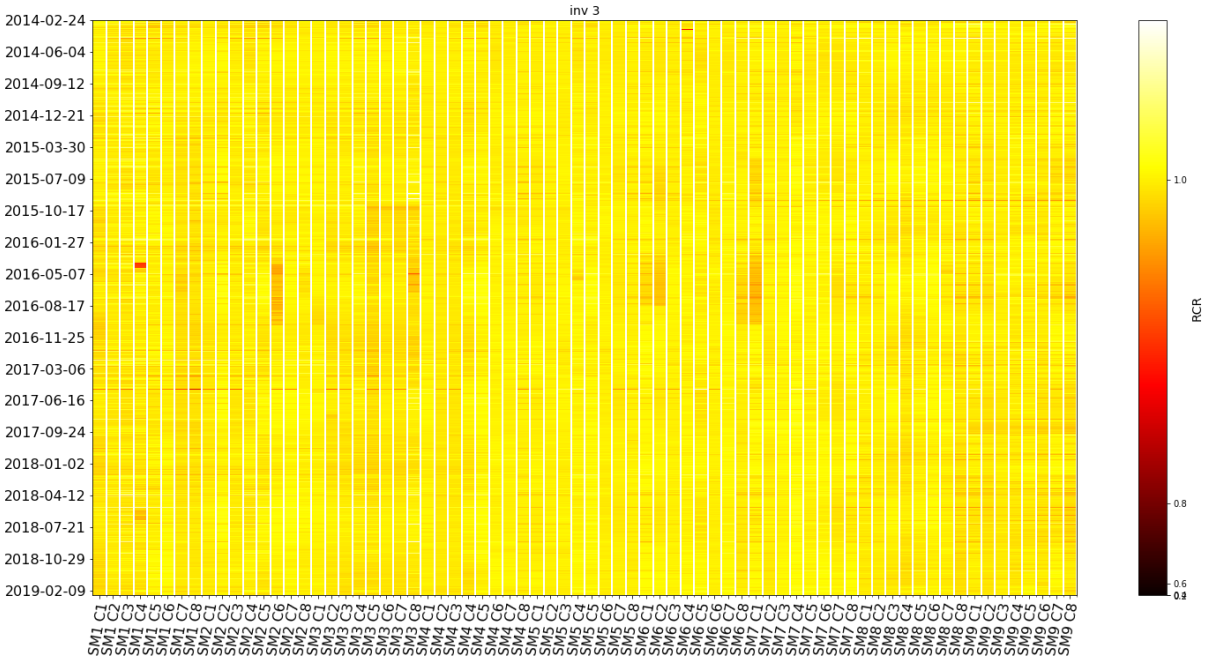

inv 4
2014-02-24
2014-06-04
2014-09-12
2014-12-21
2015-03-30
2015-07-09
2015-10-17
2016-01-27
2016-05-07
2016-08-17
2016-11-25
2017-03-06
2017-06-16
2017-09-24
2018-01-02
2018-04-12
2018-07-21
2018-10-29
2019-02-09
CR
1.0
0.8
0.6
0.4

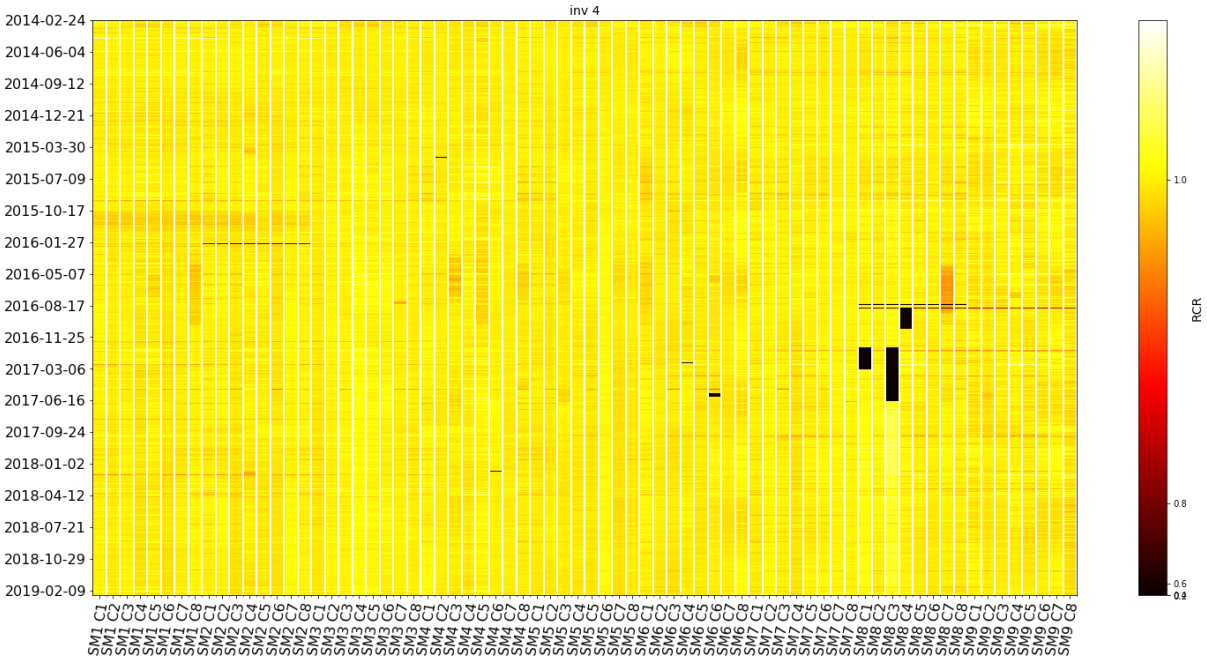

inv 7
2014-02-24
2014-06-04
2014-09-12
2014-12-21
2015-03-30
2015-07-09
2015-10-17
2016-01-25
2016-05-05
2016-08-15
2016-11-23
2017-03-04
2017-06-14
2017-09-22
2018-01-01
2018-04-11
2018-07-20
2018-10-28
2019-02-07
CR
1.0
0.8
0.6
0.4

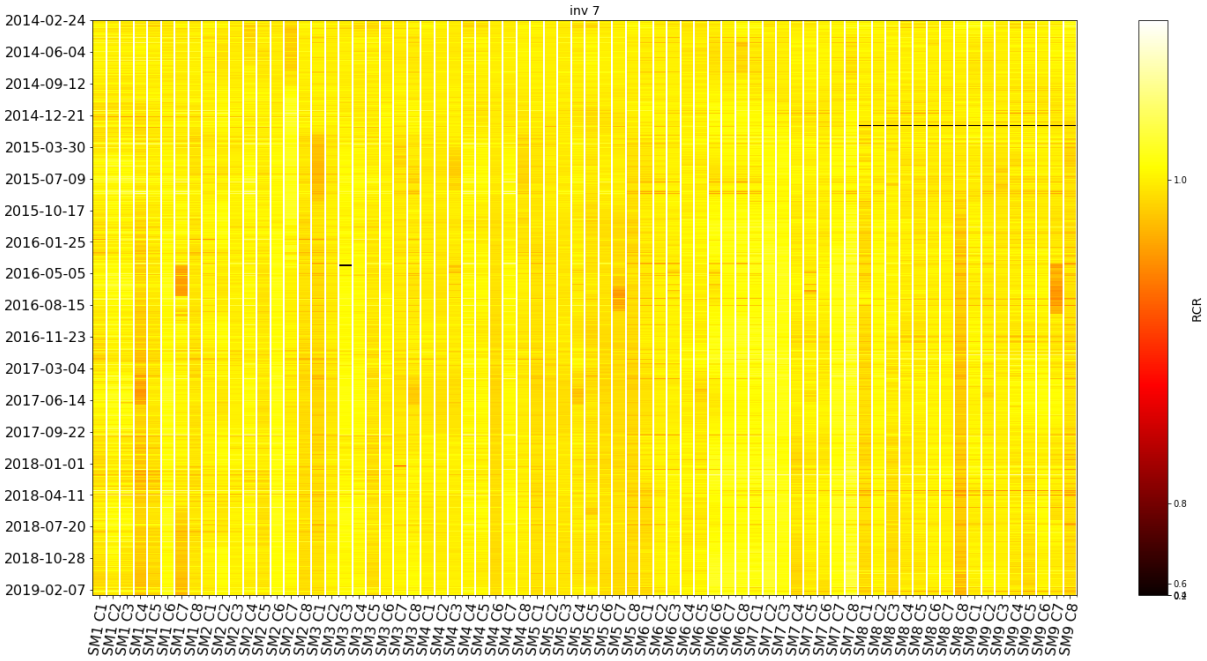


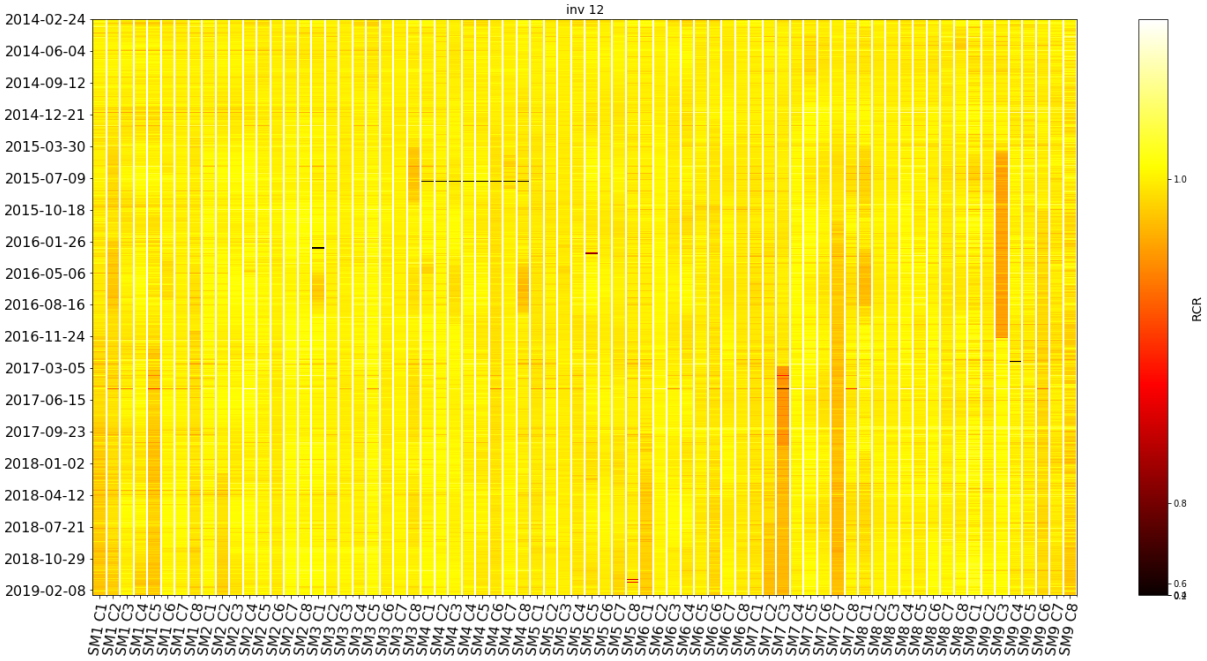

RCR for different inverters



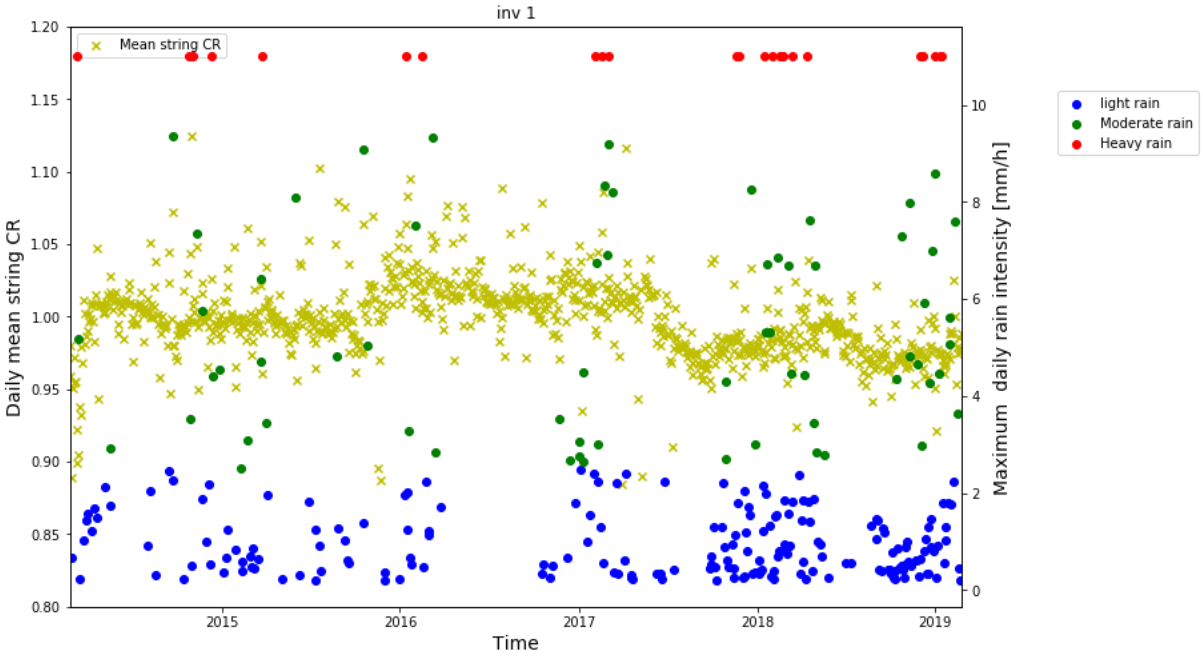

inv 3
2014-02-24
2014-06-04
2014-09-12
2014-12-21
2015-03-30
2015-07-09
2015-10-17
2016-01-27
2016-05-07
2016-08-17
2016-11-25
2017-03-06
2017-06-16
2017-09-24
2018-01-02
2018-04-12
2018-07-21
2018-10-29
2019-02-09
RCR
1.0
0.8
0.6

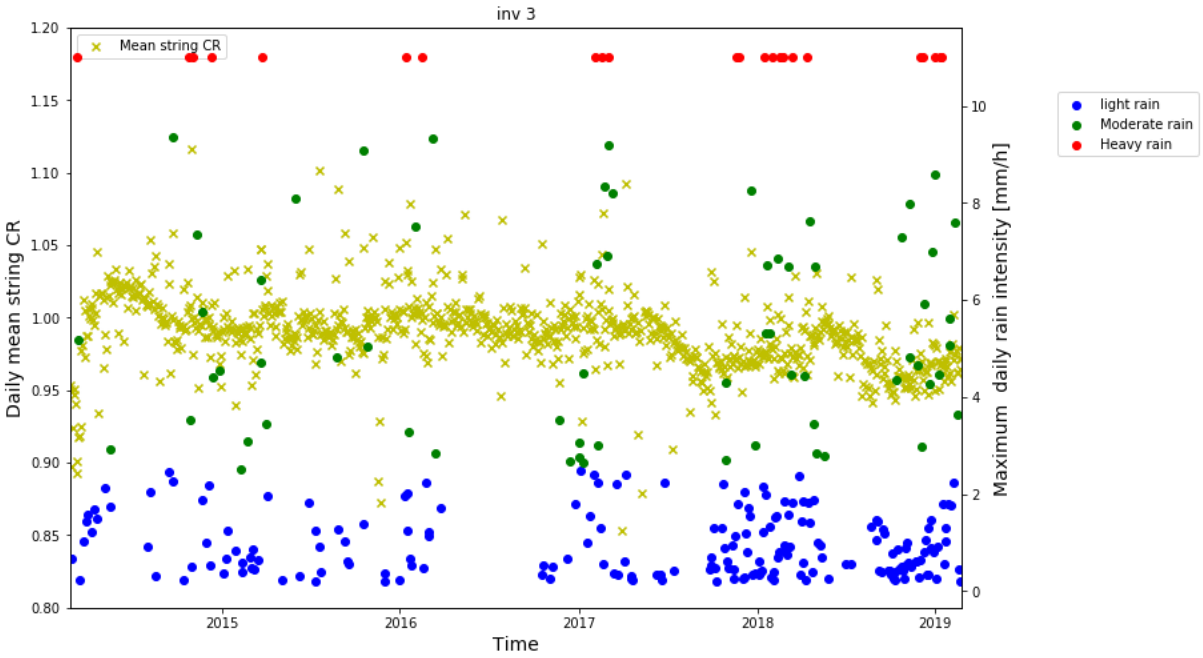

inv 4
2014-02-24
2014-06-04
2014-09-12
2014-12-21
2015-03-30
2015-07-09
2015-10-17
2016-01-27
2016-05-07
2016-08-17
2016-11-25
2017-03-06
2017-06-16
2017-09-24
2018-01-02
2018-04-12
2018-07-21
2018-10-29
2019-02-09
RCR
1.0
0.8
0.6

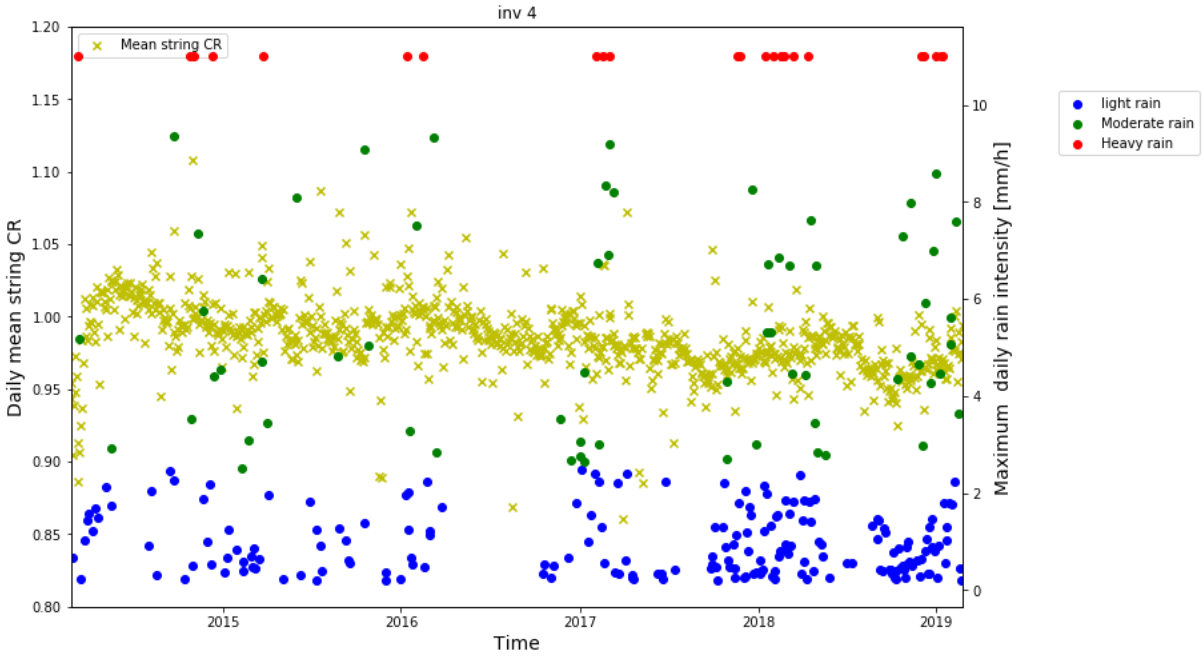

inv 7
2014-02-24
2014-06-04
2014-09-12
2014-12-21
2015-03-30
2015-07-09
2015-10-17
2016-01-25
2016-05-05
2016-08-15
2016-11-23
2017-03-04
2017-06-14
2017-09-22
2018-01-01
2018-04-11
2018-07-20
2018-10-28
2019-02-07
RCR
1.0
0.8
0.6

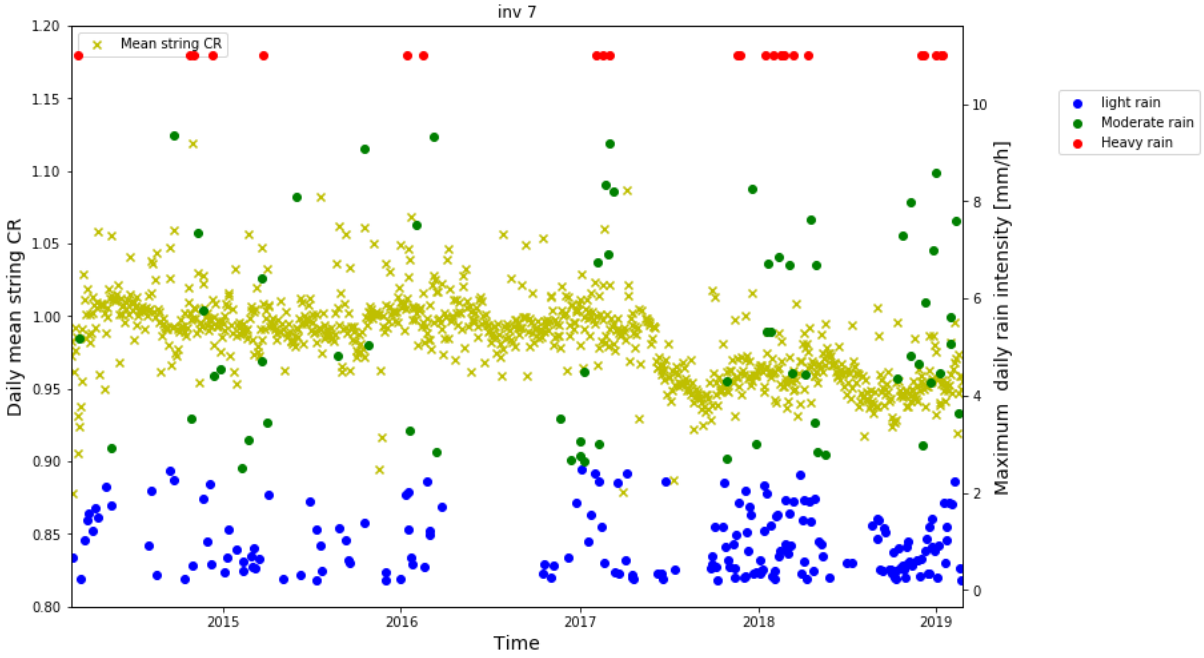

inv 12
2014-02-24
2014-06-04
2014-09-12
2014-12-21
2015-03-30
2015-07-09
2015-10-18
2016-01-26
2016-05-06
2016-08-16
2016-11-24
2017-03-05
2017-06-15
2017-09-23
2018-01-02
2018-04-12
2018-07-21
2018-10-29
2019-02-08
RCR
1.0
0.8
0.6

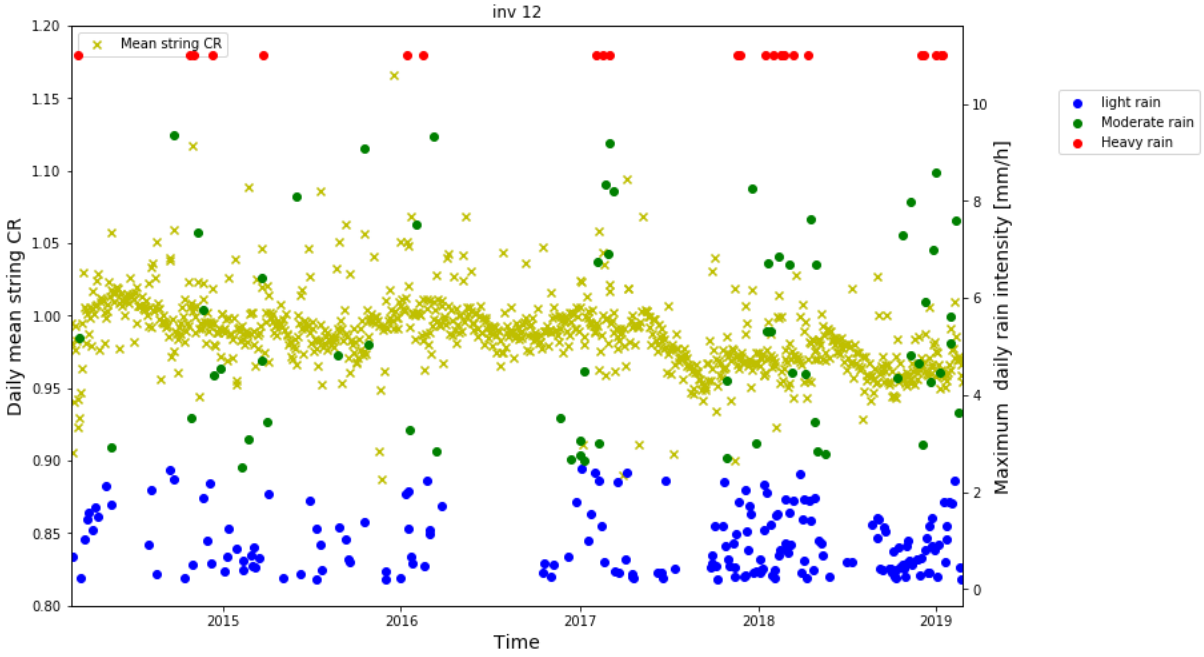
**Mean CR across all strings with rain intensity**




inv 4
Mean string CR
light rain
Moderate rain
Heavy rain
Daily mean string CR
Maximum daily rain intensity [mm/h]
Time
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
10
8
6
4
2
0
2015
2016
2017
2018
2019


inv 7
Mean string CR
light rain
Moderate rain
Heavy rain
Daily mean string CR
Maximum daily rain intensity [mm/h]
Time
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
10
8
6
4
2
0
2015
2016
2017
2018
2019


inv 12
Mean string CR
light rain
Moderate rain
Heavy rain
Daily mean string CR
Maximum daily rain intensity [mm/h]
Time
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
10
8
6
4
2
0
2015
2016
2017
2018
2019

# Appendix E

To visualise the extrapolation abilities of Random Forest Regressor and KNN regressor a simple example is shown the figure below. The explanatory variable is [0,99] and the response value is two times x. RFR and KNN have been trained upon the first 75 samples and the last 25 in the range of [75,99] are left out to be predicted upon.

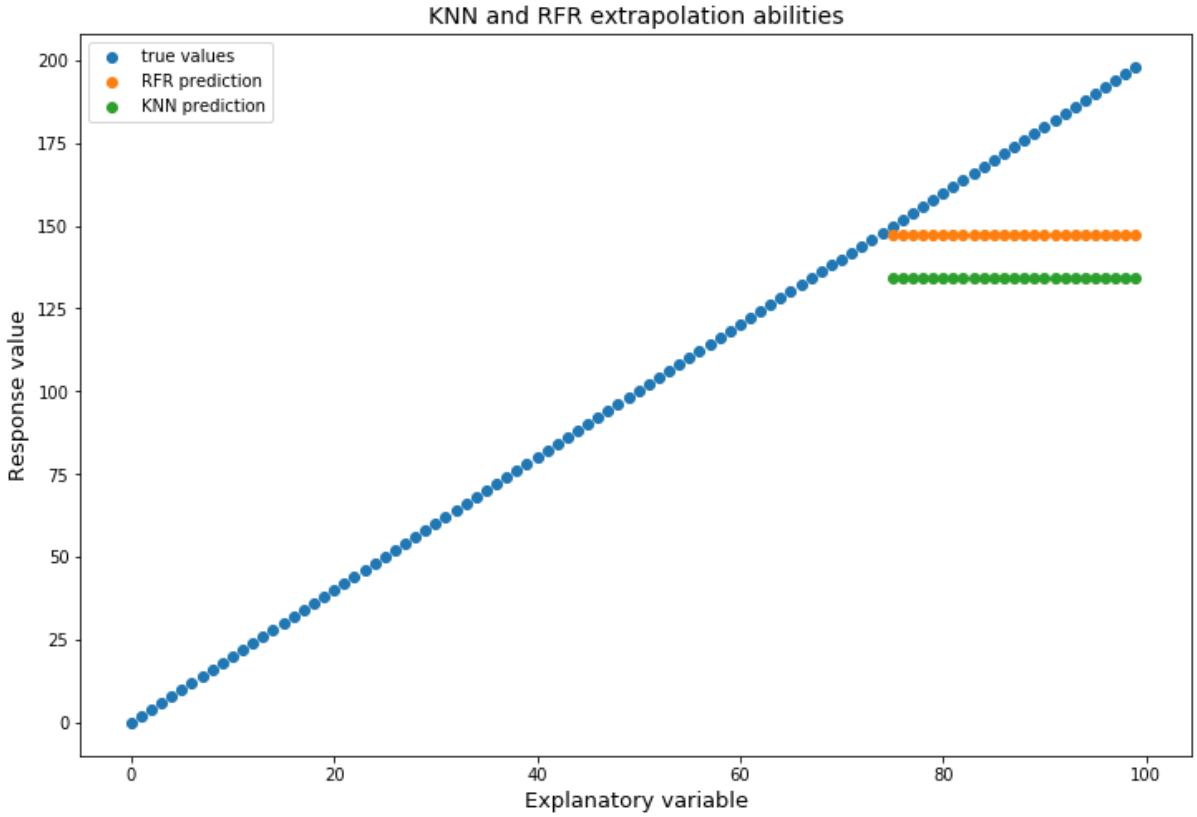
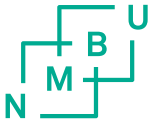



*Figure 0.1 Illustration of the extrapolation abilities of RFR and KNN regressor. The models are trained with explanatory variables in the range of 0 to 74. Subsequently, the models have been tested on the last 25 samples. RFR predictions are visualised in yellow, KNN predictions are visualised in green and the true response values are in blue.*

As can be seen from the figure, when the two models are presented variables beyond the range trained upon, the models predict the last known response value. By looking back on the example presented in the decision tree of random forest, the model which is trained upon a cell temperature in the range of 0 to 74 will, given the same irradiance, predict the same response value of 180 if the cell temperature increases from 46 to 80. For KNN the situation is more intuitive due to inner workings of the model. When given a value beyond the range trained upon the model will look at the K-nearest neighbours in the training data and predict the mean value of these. Therefore, if the explanatory variable increases from 75 to 100, it will still look at the K-nearest neighbours in the training data, which are the same in both situations.




**Norges miljø- og biovitenskapelige universitet**
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway