Impact of Data Quality on Photovoltaic (PV) Performance Assessment

By Elena Koumpli

Doctoral Thesis

Submitted in partial fulfilment of the requirements for the award of

Doctor of Philosophy of Loughborough University

November 2017

Abstract

In this work, data quality control and mitigation tools have been developed for improving the accuracy of photovoltaic (PV) system performance assessment. These tools allow to demonstrate the impact of ignoring erroneous or lost data on performance evaluation and fault detection. The work mainly focuses on residential PV systems where monitoring is limited to recording total generation and the lack of meteorological data makes quality control in that area truly challenging. Main quality issues addressed in this work are with regards to wrong system description and missing electrical and/or meteorological data in monitoring.

An automatic detection of wrong input information such as system nominal capacity and azimuth is developed, based on statistical distributions of annual figures of PV system performance ratio (PR) and final yield. This approach is specifically useful in carrying out PV fleet analyses where only monthly or annual energy outputs are available. The evaluation is carried out based on synthetic weather data which is obtained by interpolating from a network of about 80 meteorological monitoring stations operated by the UK Meteorological Office. The procedures are used on a large PV domestic dataset, obtained by a social housing organisation, where a significant number of cases with wrong input information are found.

Data interruption is identified as another challenge in PV monitoring data, although the effect of this is particularly under-researched in the area of PV. Disregarding missing energy generation data leads to falsely estimated performance figures, which consequently may lead to false alarms on performance and/or the lack of necessary requirements for the financial revenue of a domestic system through the feed-in-tariff scheme. In this work, the effect of missing data is mitigated by applying novel data inference methods based on empirical and artificial neural network approaches, training algorithms and remotely inferred weather data. Various cases of data loss are considered and case studies from the CREST monitoring system and the domestic dataset are used as test cases. When using back-filled energy output, monthly PR estimation yields more accurate results than when including prolonged data gaps in the analysis.

Finally, to further discriminate more obscure data from system faults when higher temporal resolution data is available, a remote modelling and failure detection framework is

developed based on a physical electrical model, remote input weather data and system description extracted from PV module and inverter manufacturer datasheets. The failure detection is based on the analysis of daily profiles and long-term PR comparison of neighbouring PV systems. By employing this tool on various case studies it is seen that undetected wrong data may severely obscure fault detection, affecting PV system's lifetime. Based on the results and conclusions of this work on the employed residential dataset, essential data requirements for domestic PV monitoring are introduced as a potential contribution to existing lessons learnt in PV monitoring.

Acknowledgments

I would like to express my gratitude and deep appreciation to my supervisors Ralph Gottschalg and Paul Rowley for all their advice and support during my PhD studies. I would like to specifically thank Ralph for giving me this great opportunity to undertake a PhD in CREST, where I had the chance to be part of a lovely team.

I would like to deeply thank Tom Betts and Brian Goss for their full and continuous support all this time, even though they weren't officially my supervisors, they have always provided me with great advice whenever I needed one.

I would like to say a huge thank you to Heather Convey and John Carr from Nottingham City Homes, for their trust and support and for providing me data without which a large part of my work could not have been possible.

I am grateful to my colleague and friend Diane Palmer for our great collaboration all these years and for providing me weather data without which this project could not have succeeded. I would also like to thank Thomas Huld for our excellent collaboration during my short visit in the Joint Research Centre in Ispra, Italy.

I am grateful to my friend and colleague Ian Cole for reviewing this work and for giving me great advice on improving it. I would also like to thank my friends Michael Owen-Bellini, Farhad Anvari-Azar and Karl Bedrich for the great times that we had during our PhDs which I will never forget.

I am extremely grateful to my beloved family, Alekos, Anna and Katia for being the best family a person can have. Their love and support has given me the strength to carry out my studies and making them proud has been my motivational force at all times.

Last and most importantly I want to thank my other half and best friend George Koutsourakis for his full and continuous support all these years, for the fun and the sad times, for making tough times easy and for being one of the most charismatic scientists I have ever met.

Table of Contents

List of Figures

List of Tables

Nomenclature

V	= Voltage (Volts - V)
I	= Current (Amperes - A)
P	= Power (Watts - W)
I_0	= Diode saturation current (A)
I_{PH}	= Photogenerated current (A)
R_S	= Series resistance (Ohm - Ω)
R_{SH}	= Shunt resistance (Ohm - Ω)
q	= Elementary charge = $1.6 \cdot 10^{-19}$ in Coulombs
n	= Diode ideality factor
k	= Boltzmann constant = $1.38 \cdot 10^{-23}$ $J \cdot K^{-1}$ = $8.167 \cdot 10^{-5}$ $eV \cdot K^{-1}$
T	= Temperature (Kelvin or Celsius)
a	= Fraction of ohmic current involved in avalanche breakdown
m	= Avalanche breakdown exponent
V_{BR}	= Breakdown voltage (V)
V_{TH}	= Thermal voltage (V)
I_{SC}	= Short circuit current (A)
I_{MPP}	= Current at maximum power point (A)
V_{MPP}	= Voltage at maximum power point (A)
P_{MPP}	= Maximum power (W)
V_M	= Voltage of the module (V)
I_M	= Current of the module (A)
V_A	= Voltage of the array (V)
I_A	= Current of the array (A)
N_S	= Number of cells in series
N_P	= Number of strings in parallel
FF	= Fill factor
G	= Irradiance (W/m^2)
H	= Irradiation (Wh/m^2)
K_i	= Temperature coefficient for short-circuit current ($A \cdot K^{-1}$)

E_g	= Semiconductor bandgap (eV)
STC	= Standard Testing Conditions
E_{DC}	= Energy output at the DC side of the inverter (Watt hours – Wh)
E_{AC}	= Energy output at the AC side of the inverter (Watt hours – Wh)
P_{DC}	= Power output at the DC side of the inverter (W)
P_{AC}	= Power output at the AC side of the inverter (W)
G_{STC}	= In-plane irradiance (W/m^2) at STC = 1000 W/m^2
P_{STC}	= Nominal capacity (W_P) = peak power at STC
T_{STC}	= Module temperature at STC (K)
GHI	= Global horizontal irradiation (Wh/m^2)
POA	= Plane of array (in-plane) irradiation (Wh/m^2)
θ	= Azimuth angle (^{0})
φ	= Inclination (tilt) angle (^{0})
k_{t}	= hourly clearness index
K_{t}	=daily average clearness index
G_b	= beam irradiance (W/m^2)
G_d	= diffuse irradiance (W/m^2)
GHI	= global horizontal irradiance (W/m^2)
R_d	= diffuse irradiance transposition factor
ρ	= ground reflected albedo
R_r	= ground reflected irradiance transposition factor
z	= angle of incidence of the beam on the tilted plane
T_m	= Module temperature (K)
T_a	= Ambient temperature (K)
k_R	= empirical Ross' coefficient (K·m^2/W)
$k_1 - k_6$	= power model coefficients
η	= Array or system efficiency
η_{INV}	= instantaneous inverter efficiency
Y_A	= Array yield $= E_{DC}/P_{STC}$ (h or kWh/kW$_P$)
Y_F	= Final yield $= E_{AC}/P_{STC}$ (h or kWh/kW$_P$)
Y_R	= Reference yield (h)

L_C	= Array capture losses (h)
L_S	= System losses (h)
PR	= Performance ratio
PR_S	= System performance ratio (= PR including system losses)
PR_A	= Array performance ratio (= PR not including system losses)
$\mathrm{PR}_{\mathrm{theor}}$	= Theoretical performance ratio = 0.85
PI	= Performance index $= \frac{Actual\ final\ yield\ (in\ kWh/kW_P)}{Theoretical\ \ final\ yield\ (in\ kWh/kW_P)}$
MAD	= Median absolute deviation
MED_X	= Median of a distribution *X*
$(\mathrm{r})\mathrm{RMSE}$	= (percentage) root mean square error
$(\mathrm{r})\mathrm{MAE}$	= (percentage) mean absolute error
$(\mathrm{r})\mathrm{MBE}$	= (percentage) mean bias error

Chapter 1

Introduction

As of the end of August 2017, the overall UK solar capacity stood at about 13 GW, according to the latest report on solar PV deployment by the Department for Business, Energy and Industrial Strategy [1]. A significant 20% of this installed capacity (about 868,000 systems) comes from small scale domestic PV systems with an average peak capacity of $3kW_P$. Accurate evaluation of photovoltaic (PV) system performance is key element for the further advancement of the solar industry. This is ensured by effectively monitoring PV systems throughout operation. Without accurate data monitoring, actual field performance cannot reliably be compared to what is guaranteed, thus increasing financial risks.

A significant monitoring effort was carried out almost 17 years ago in the UK, known as the domestic field trials (DFT), which was the first wide spread monitoring of PV systems on a national level [2]. Lessons learnt, with regards to data quality in monitoring primarily included erroneous measurements, sensor shading or malfunction and interrupted monitoring. The acquisition of high resolution data of both meteorological and electrical measurements allowed for the identification of such issues and their correction, where possible. This further enabled reliable performance assessment, which was the main goal of the project. Today, the level of domestic PV monitoring is largely limited to total generation while the attention is mainly focused on the decrease of utility bills. Sophisticated monitoring solutions may be offered as extra commercial services, which are, however, not compelling to domestic PV owners due to increased costs. The lack of sufficient monitoring not only entails risks on the performance of these systems but also obstructs any efforts carried out to analyse their performance, by means of remote monitoring.

To evaluate PV performance, the most employed index is the performance ratio (PR) per the IEC (International Electrotechnical Commission) standards [3]. The accurate estimation of PR requires the knowledge of the total energy produced by the system, its nominal rating and the irradiation received. For the analysis of domestic systems, solar radiation and/or temperature are usually derived from satellite data or from nearby meteorological stations,

thus the assessments are primarily carried out remotely. There is a plethora of studies on the remote performance assessment of PV systems, such as [4]–[6]. Although data quality in solar radiation datasets has been largely researched [7], data quality in terms of accurate PV system descriptions and/or interrupted monitoring are occasionally broached but not studied in depth in recent assessments. Erroneous system descriptions in domestic PV monitoring make data interpretation and performance modelling extremely difficult and often result in unusual annual PR values and sometimes in excess of unity. Evidently, low data quality obscures the overall assessment and often hides potential faults, which have detrimental impact if not detected. The question arising at this point is to what extent can bad data be detected and/or corrected remotely so as their impact on the subsequent performance assessment is minimised. This becomes even more challenging in cases where insight into individual systems is troublesome as in PV fleet assessments.

This work aims to investigate common data quality issues in PV monitoring and to propose means to identify and remedy their impact on performance assessments. There are often cases where due to power outages, communication link failures and component faults, missing data occur. In such cases the performance evaluation will be weighted towards the period in which data are available. Thus, the estimation of the annual PR with the inclusion of prolonged missing periods would lead to biased results and a financial penalty if the PR is lower than a contractual threshold. In this work, various methodologies are developed for the inference of missing data in PV monitoring, both for domestic and non-domestic PV. The ultimate aim of these methods is to “recover” the lost data, if the operation of the PV system has not been affected by unknown faults while monitoring is interrupted.

In the context of domestic PV systems, statistical procedures for the detection of wrong PV system description are focused on PV fleet analyses, where meteorological data are inferred. Performance ratio is thereby calculated based on estimated rather than measured incident irradiation. Consequently, modelling of irradiation is affected by the installation azimuth and inclination of the PV surface. Common errors in this information in addition to modelling uncertainty compromises the quality of the performance assessment. Similar errors are found in the declared nominal capacities. Thus, in terms of domestic monitoring accurate PR estimation becomes more difficult due to limited input information and low data quality.

In addition to PR estimation, remote detection and identification of specific failures, occurring during PV system operation, is realised based on the comparison of actual to simulated output and remotely assessed weather data. The impact of the lack of data quality assessment monitoring on PR calculation and fault detection is demonstrated, focusing on domestic PV systems.

In Chapter 2, the background of the fundamental blocks required for modelling and assessing PV system performance is provided. The most important aspects of remote monitoring are discussed, such as commonly employed meteorological datasets and the models required to translate these into system-specific variables. Performance indicators and fault detection methodologies are reviewed, where the differences between domestic and large PV system monitoring are discussed.

In Chapter 3, a chain of statistical tools is developed to automatically classify PV systems based on selected quality indicators. Wrong input information is discriminated based on annual figures of performance ratio and specific yield. The proposed tools are applied on a large PV domestic dataset obtained by the Nottingham City Homes social housing association, and for a year's worth of data. The impact of missing data and wrong input information on the annual performance figures is demonstrated. Finally, an azimuth correction tool is developed based on a clear sky model and Gaussian fits of hourly energy readings, which can be used for the automatic correction of wrong input azimuth, where daily profiles are available.

In Chapter 4 novel data inference techniques are developed. Different cases of data loss are introduced, whereby each case is associated to specific PV system configurations and monitoring granularity. The accuracy of the applied back-filling methodologies is discussed for each case and the benefits from back-filling are presented in terms of the significant reduction of the bias in the calculated performance ratio. For one case of data loss, two different modelling approaches are applied, based on a commonly employed empirical model and on an artificial neural network approach respectively.

In Chapter 5 a remote failure detection framework is developed where the impact of data quality on the efficiency of remote fault detection is demonstrated. It is shown that bad data in domestic monitoring lead to false estimations of performance and entail risks with regards to the system's lifetime. Based on the overall number of issues found in the test dataset,

specific data requirements are highlighted, as a further contribution to existing lessons learnt in domestic PV monitoring.

Chapter 2

Photovoltaic system performance evaluation framework

2.1 Introduction

This chapter reviews the processes of photovoltaic (PV) system performance assessment and monitoring. Performance assessment can be differentiated from energy yield prediction as, while often used in the same context, they are different in terms of their objectives. Performance assessment aims to evaluate the outcome of an operating system and determine its efficiency under various weather or design effects at any time. Energy yield prediction most often refers to estimating the energy output of a system based on historical weather data, often expressed in annual terms, and typically carried out before the installation of a PV system at a given location.

Performance assessment of PV systems comprises an extensive framework of various sub-processes; from monitoring to modelling to performance prediction and fault detection of actual PV installations. In order to determine whether a PV system behaves as expected a comparison to a theoretical system of the same characteristics is required, as measuring power production against performance benchmarks allows to determine if and when a failure is present in the system. The outcome of this comparison is the most crucial step in the performance evaluation framework.

Modelling the theoretical output of a PV system requires a collection of models which overall take into account various effects such as meteorological variables (incident solar radiation and temperature) and system specific variables such as installation configuration and PV module performance data. Each model describes specific blocks of the procedure, starting from modelling a PV cell to a module, array and then system. In its simplest form, the transition to a PV system can be described by incorporating an inverter model which describes the output of a system at various array voltage and power levels.

By comparing actual to measured data, increased system losses can be detected, and the next step is to estimate the factors that could have caused the system to under-perform.

Ideally, failure detection should take place during monitoring a PV system in order to repair any occurring failures as soon as they appear. However, this cannot be applied with the same manner for domestic and utility scale PV systems, as monitoring granularity differs significantly between these two cases. This is because the value of monitoring in the case of a large-scale PV system is much higher, whereas for a domestic PV system installing a monitoring service is generally not considered. The granularity and type of monitoring used in each system category also dictates the applied procedures for performance assessment and fault detection for that application, which is further analysed in this chapter. Specifically for domestic PV systems the lack of climatic monitoring make remote monitoring a useful way of determining system under-performance, whereby modelled rather than measured weather data are employed, which area comprises a separate sub-modelling chain. The performance assessment framework can be summarised in the block diagram in Figure 2.1. Major blocks are the modelling of the theoretical output of a PV system, the analysis of PV monitoring data, which depends on the size and type of system, and the comparison between the above as part of the fault detection process. Each block includes its own sub-models and processes, which are described in the following sections, starting from modelling the PV system.

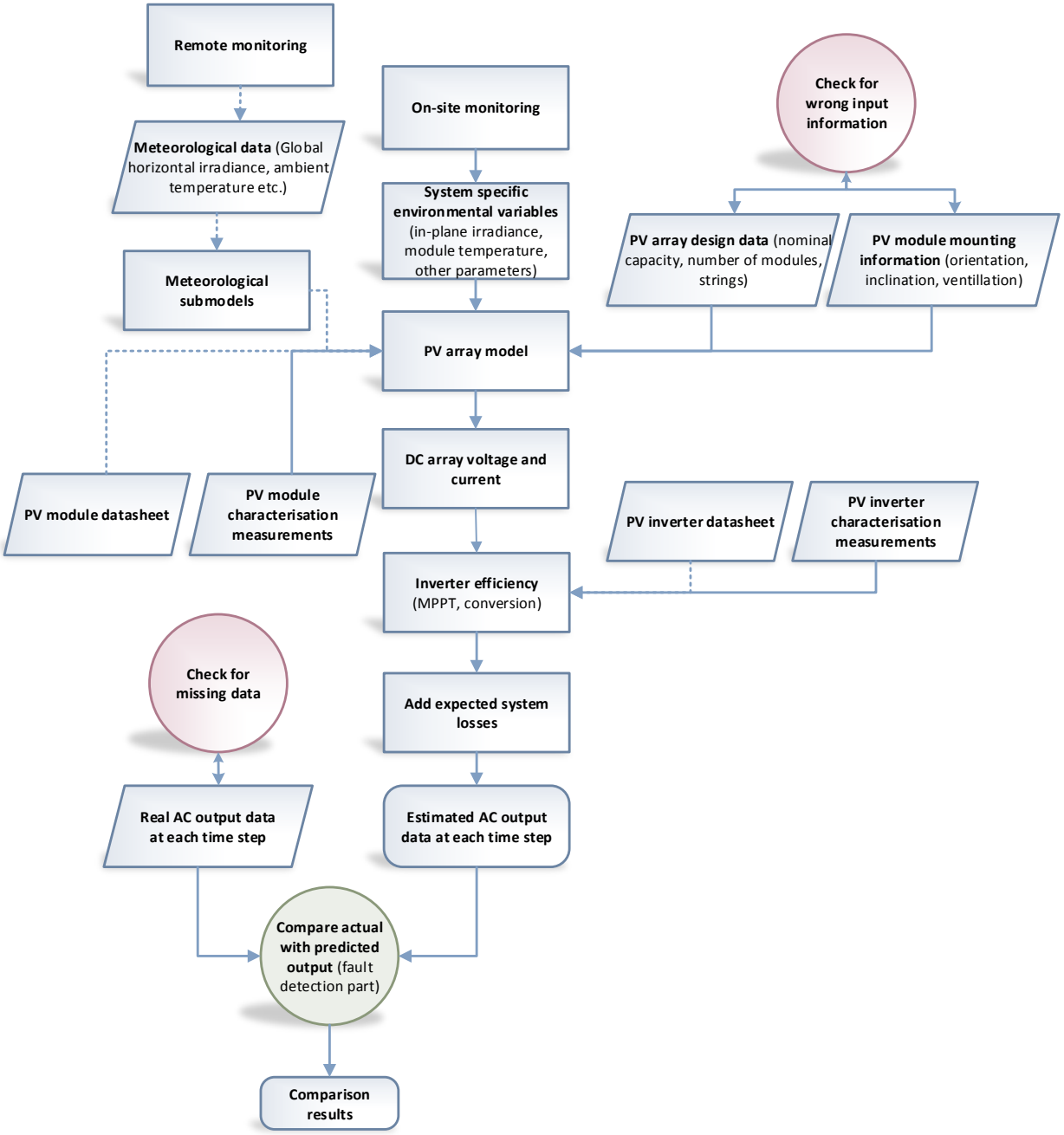


Figure 2.1. Main performance assessment blocks including the data quality functions applied in this work. The dotted lines denote optional steps when data from onsite monitoring or device characterisation measurements are not available.

2.2 Photovoltaic array performance modelling

There is a plethora of models which can be used to predict the performance of a PV array. There are two major categories of PV models; those based on the equivalent circuit representation of the cell and those based on empirical correlations. Within these categories, models are differentiated according to the complexity of the underlying physics and the fundamental parameters needed as input. A large number of models lie within these two categories, differing in complexity and the number and type of input parameters required, whether the output is the full I-V curve (parametric continuous), a set of I–V points (discrete), or just the maximum power point. In practice, there is no such thing as "best model". A significant number of models has been reviewed within the "Performance" round robins [8] showing that deviations in module modelling uncertainty were within 5%, but rather higher errors were observed when the same models were applied to describe different PV modules of the same manufacturer and technology. The choice of model becomes truly situational, where in some cases simplicity is preferred over accuracy or cases where the accuracy of modelling input parameters is more important than the model itself [9]. These two major categories are discussed in this section.

2.2.1 Device physics modelling

In the context of device physics modelling the solar cell is the basic building block of the PV array. A simplified representation of solar cell operation is described in Figure 2.2. A solar cell comprises of a junction of p-type and n-type semiconductors. The n-type semiconductor has a high concentration of electrons whereas the p-type semiconductor has a high concentration of holes. When the two types of semiconductors form a contact, holes diffuse from the p to n type region and electrons diffuse from the n to the p type region, and as a result, an electric field is formed. The charge diffusion process continues until an equilibrium is reached between the charge concentration and the developing electric field at this interface. Such a field is called the "depletion region" or the "space charge region". When radiation is absorbed, electrons and holes are generated and the electric field at the junction pushes the

carriers across the junction which separates the charges. These charge carriers are then diffused towards the metallic contacts.

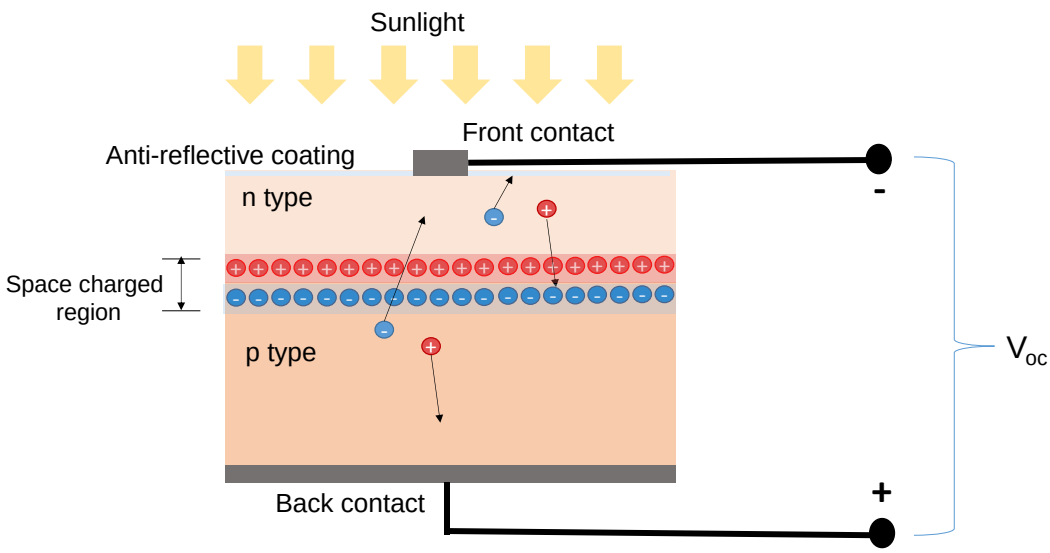


Figure 2.2. Simple schematic diagram of a solar cell, where V_{oc} is the open circuit voltage.

Charge carriers are generated at different depths according to the wavelength of the absorbed photon. Only photons with sufficient energy to create an electron–hole pair are absorbed, that is photons with energy equal or higher than the semiconductor's bandgap. Short wavelengths (higher energy photons) are absorbed near the surface and longer wavelengths (lower energy photons) are absorbed in the bulk. In the absence of sunlight, the solar cell acts as a simple diode described by Shockley diode equation [10]:

$$I_D = I_0 \left[\exp\left(\frac{\mathrm{q}V}{n\mathrm{k}T} \right) - 1 \right] \tag{2.1}$$

Where,

I_0 = diode saturation current (Amperes)

q = elementary charge = $1.6 \cdot 10^{-19}$ in Coulombs

V = Voltage (Volts)

n = diode ideality factor

k = Boltzmann constant = $1.38 \cdot 10^{-23}$ $J \cdot K^{-1}$

T = temperature (Kelvin)

For an ideal solar cell under illumination, Equation (2.1) becomes:

$$I = I_{PH} - I_D = I_{PH} - I_0 \left[\exp\left(\frac{\mathrm{q}V}{n\mathrm{k}T}\right) - 1\right] \tag{2.2}$$

Where I_{PH} is the photocurrent. Realistic solar cell operation is never ideal and so the current-voltage characteristic of the p-n junction is given by altered diode expressions, where different charge carrier recombination mechanisms are also considered. By adding parasitic resistances in Equation (2.2), it becomes:

$$I = I_{PH} - I_0 \left[e^{\frac{q(V+IR_S)}{nkT}} - 1\right] - \frac{V + IR_S}{R_{SH}} \tag{2.3}$$

Where R_S, R_{SH} are the series and shunt resistance respectively. Sources of series resistance include metal contacts and current flow resistance, while shunt resistance represents the leakage current in the p-n junction. Equation (2.3) is associated with the equivalent electrical circuit given in Figure 2.3 and is known as the *one-diode model*, which is the longest established in literature [11].

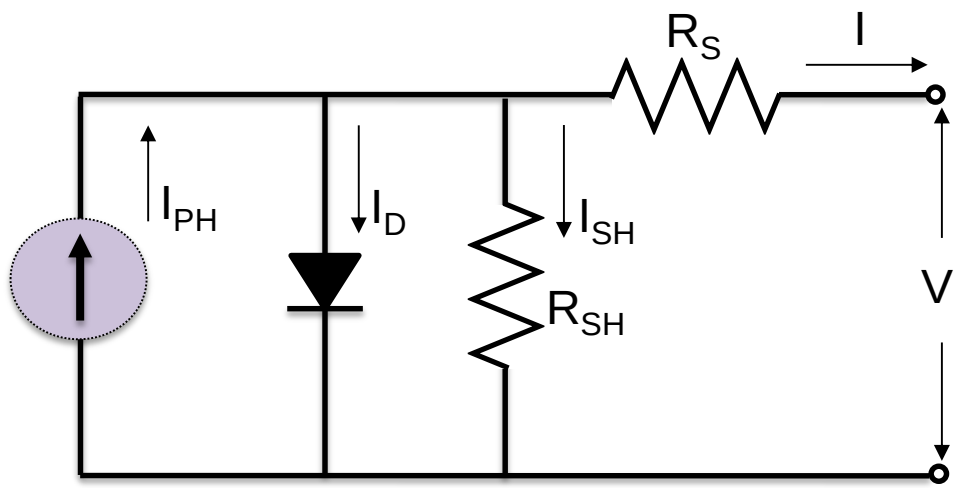


Figure 2.3. Single diode equivalent circuit

Equivalent circuit models define the entire I-V curve of a PV device (cell, module, or array) as a continuous function for a given set of operating conditions. For a good working cell, series resistance should be close to zero ($R_S \to 0$) and shunt resistance close to infinity ($R_{SH} \to \infty$).

Typical I-V, P-V curves produced by Equation (2.3) are given in Figure 2.4, where the *three characteristic points* on the I-V curve are highlighted, namely:

Short circuit current:	$I = I_{SC}$, $V = 0$
Maximum power point:	$I = I_{MPP}$, $V = V_{MPP}$, P_{MPP} = maximum power
Open circuit voltage:	$I = 0$, $V = V_{OC}$

The maximum power (P_{MPP}) that can be obtained is given by the largest area rectangle under the I-V curve:

$$P = I \cdot V \tag{2.4}$$

The fill factor (FF) is a measure of "squareness" of the IV curve and is given by:

$$FF = \frac{I_{MPP} \cdot V_{MPP}}{I_{SC}.V_{OC}} \tag{2.5}$$

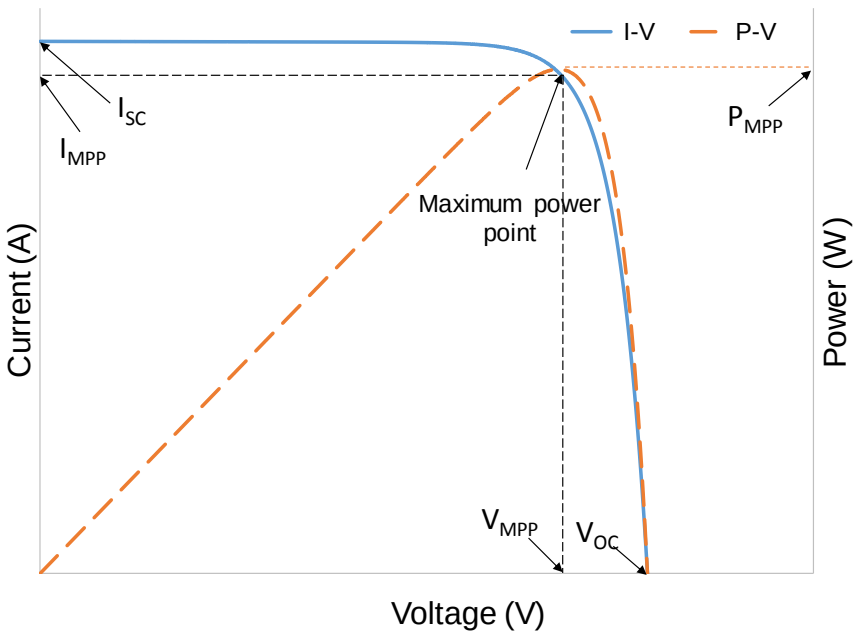


Figure 2.4. Example I-V and P-V curves of a PV device.

Series resistance has no effect on the open-circuit voltage but reduces short circuit current, when it becomes too high. Conversely, shunt resistance has no effect on the short-circuit

current, but reduces open-circuit voltage, when it becomes too low [12]. Both effects reduce the fill factor, and thus the power output of the solar cell. The one-diode model may also have a simpler form: the 4-parameter or R_S model, which completely omits the shunt resistance and therefore the number of unknown parameters is reduced to four, namely (n, R_S, I_{PH}, I_0). This is usually the case when the shunt losses are considered too small compared to the current output. Four-parameter models are obviously less complicated but it has been shown that they are unable to predict the effect of high temperature on the current, and thus they lead to a less accurate prediction when considering temperature variations and high temperatures operation [13]. The effects of series and shunt resistances are further seen in Figure 2.5.

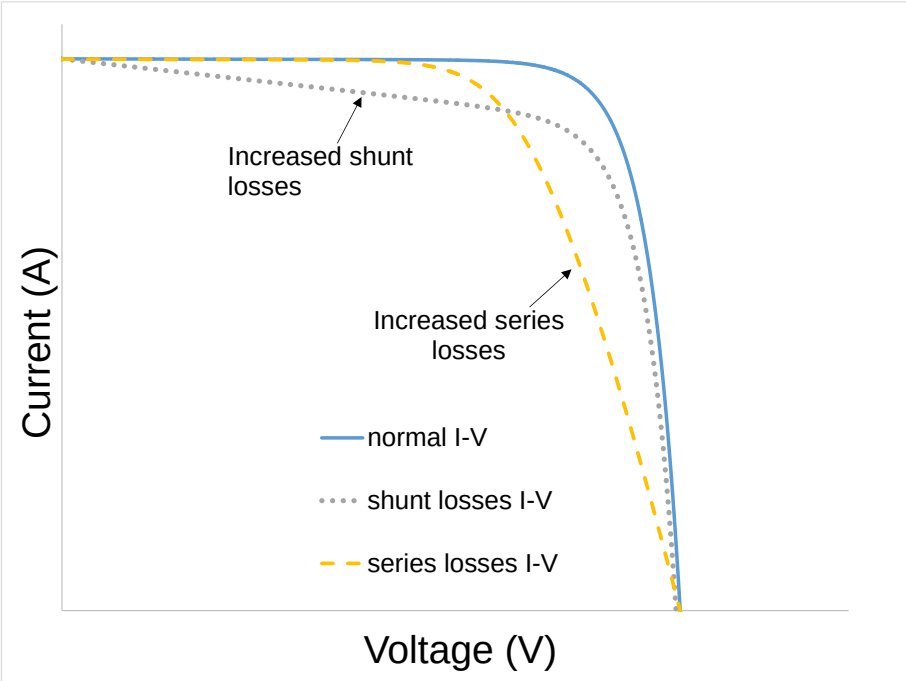


Figure 2.5. Series and shunt resistance effect on the I-V curve.

The diode ideality (quality) factor typically has a value between 1 and 2 for crystalline silicon solar cells, with $n \approx 1$ for cells dominated by recombination in the bulk (quasi-neutral) regions and $n \rightarrow 2$ when recombination in the depletion region dominates. In order to represent recombination effects in the depletion zone a second diode can be added and Equation (2.3) becomes [14]:

$$I = I_{PH} - I_{01}\left[e^{\frac{q(V+IR_S)}{n_1 kT}} - 1\right] - I_{02}\left[e^{\frac{q(V+IR_S)}{n_2 kT}} - 1\right] - \frac{V + IR_S}{R_{SH}} \qquad (2.6)$$

The one-diode model (usually) employs five parameters (n, R_S, R_{SH}, I_{PH}, I_0), also known as the 5-parameter model, while the two-diode model employs seven (n_1, n_2, R_S, R_{SH}, I_{PH}, I_{01}, I_{02}) due to the additional diode. One- and two-diode models are the most commonly employed in device physics modelling [15]–[20]. For crystalline silicon solar cells one-diode models have been found to perform well, especially for solar cells with high fill factors, i.e. low R_S and high R_{SH}. Several improvements have been proposed for the representation of cell behaviour when translating I-V data from Standard Testing Conditions (STC)(cell temperature (T_{STC} = 25 ^oC), solar irradiance (G_{STC} = 1000 W/m^2) and a spectrum equivalent to clear sky conditions at a relative air mass (*AM*) of 1.5 [21]) to other conditions of irradiance and temperature [15][17]. These apply a correction factor which can be extracted if I-V curves are provided for more than one operating conditions of temperature (including STC) if that is available. Various assumptions or modifications are also applied for describing thin film solar cells. These should consider device specific behaviours (on shunt and series resistances) and different recombination mechanisms in thin film devices as shown for amorphous silicon [22]–[24] and CIGS or CdTe solar cells [24]–[26]. Two-diode models are usually more appropriate to describe thin film solar cells [19],[27] and also show greater accuracy at low irradiance conditions. While the two-diode model is generally more accurate, it is also more complicated and requires extra computational time to be solved due to the iterative numerical optimisation problem the circuit equation presents. Thus, the one-diode model is often preferred for simple modelling applications.

An extension term can be added in both models, which includes the diode breakdown at very high negative voltages. This extension is given by [28]:

$$diode\_breakdown\_term = -\,a\left(1 - \frac{V + IR_s}{V_{BR}}\right)^{-m} \tag{2.7}$$

Where,

a = fraction of ohmic current involved in avalanche breakdown

m = avalanche breakdown exponent

V_{BR} = breakdown voltage (Volts)

This term is employed when modelling specific conditions such as mismatches between different cells connected in series or parallel. The breakdown voltage ranges between -5 and -20 Volts, depending on the device material and technology. It is generally omitted when modelling cells or modules under uniform conditions (a = 0).

To describe the electrical behaviour of a module as illustrated in Figure 2.6, Kirchhoff's laws for voltage and current are employed. So, for a series connection between N number of cells or modules or any PV device, the following laws apply for current and voltage respectively:

$$I = I_1 = I_2 = \cdots = I_N \tag{2.8}$$

$$V = V_1 + V_2 + \cdots + V_N \tag{2.9}$$

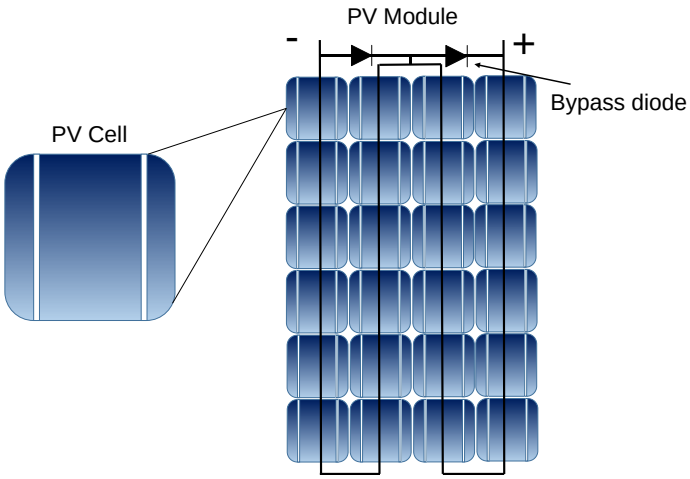


Figure 2.6. Graphical representation of a silicon PV module of 24 cells connected in series and by pass diodes. Typical silicon PV modules range from a few Watts to about 300 Watts.

Additionally, a PV string consists of N_{Series} modules connected in series and a PV array consists of $N_{Parallel}$ PV strings connected in parallel as depicted in Figure 2.7.

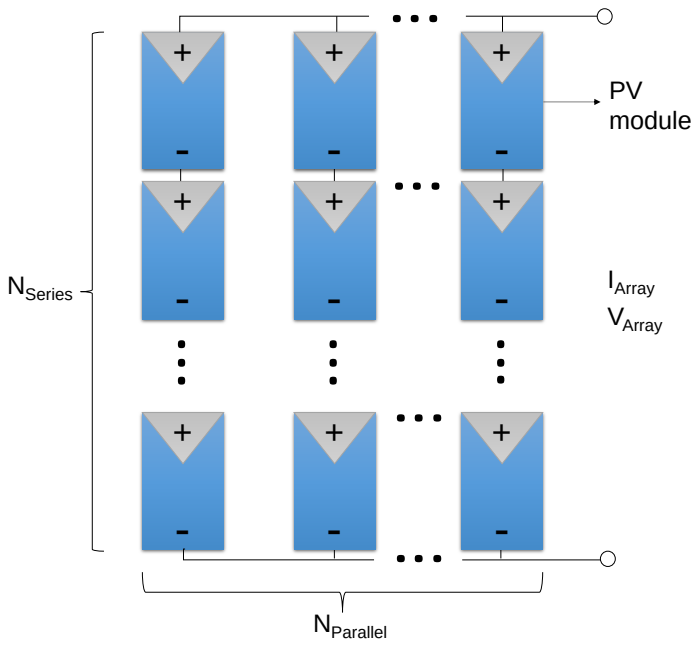


Figure 2.7. Graphical configuration of a PV array consisting of N_{Series} x $N_{Parallel}$ PV modules.

For parallel electrical connections (for example PV strings) Kirchhoff's laws are as follows:

$$I = I_1 + I_2 + \cdots + I_N \tag{2.10}$$

$$V = V_1 = V_2 = \cdots = V_N \tag{2.11}$$

Finally, one-diode and two-diode models can be used to describe a PV module/string and a PV array by slightly modifying the corresponding equations for a solar cell [29],[30]. Specifically, in the case of the one-diode model equation (2.3) can be re-written for a module as:

$$I_M = I_{PH} - I_0 \left[e^{\frac{q(V_M + I_M N_S R_S)}{n N_S k T}} - 1 \right] - \frac{V + I_M N_S R_S}{N_S R_{SH}} \tag{2.12}$$

Where,

V_M = Voltage of the module (Volts)

I_M = Current of the module (Amps)

N_S = Number of cells in series

The five modelling parameters appearing in Equation (2.12), namely n, R_S, R_{SH}, I_{PH} and I_0 imply *cell* properties whereas voltage and current describe the module (or string). For a PV array with N_S cells in series and N_P strings in parallel Equation (2.3) becomes:

$$I_A = I_{PH} - I_0 \left[e^{\frac{q(V_M + I_A \frac{N_S}{N_P} R_S)}{n N_S k T}} - 1 \right] - \frac{V + I_A \frac{N_S}{N_P} R_S}{\frac{N_S}{N_P} R_{SH}} \tag{2.13}$$

Where,

V_A = Voltage of the module (Volts)

I_A = Current of the module (Amps)

N_S = Number of cells in series

N_P = Number of strings in parallel

As before, the modelling parameters n, R_S, R_{SH}, I_{PH} and I_0 in this expression imply *cell* properties whereas voltage and current correspond to the electrical output of the array. Both equations (2.12) and (2.13) assume that modules and arrays consist of identical cells with the exact same characteristics described by the same modelling parameters under the same temperature and irradiance conditions. This assumption, however, is unrealistic. For the sake of simplicity, a PV module or a small string consisting of several modules, can be roughly described as comprising a number of identical cells operating at the same temperature and under the same irradiance at all times. However, this assumption is a compromise between modelling accuracy and complexity which grows weaker as the size and area of the array surface increases.

In reality, non-uniform conditions often occur across the area of a PV module. This can be due to the intrinsic variations of cell and module characteristics due to fabrication processes. For example, variations in rated power between different modules in a PV array can cause an average of 1.3% power loss annually based on simulations carried out for PV systems of over 250 kW_P using UK climatic data [31]. Another common reason is partial shading. Different areas across the module or array may experience different conditions of irradiance and operating temperature. Therefore, the same equation cannot describe all blocks of the PV array accurately. Mismatches in cell characteristics cause modules (and consequently entire

arrays) to operate at suboptimal conditions, generally governed by the weakest cell. In order for mismatches and inhomogeneities to be represented, *circuit based modelling* is employed. In these models the module is resolved to smaller units of substrings or cells or even sub-cells where each one is represented by its own diode characteristics added up to form the array electrical circuit [14], [28], [32], [33]. In terms of performance assessment, however, this electrical mismatch is commonly included as a factor in the estimated system losses, as further discussed in 2.4.

2.2.1.1 *Extraction of the modelling parameters*

Diode models described by Equations (2.3) and (2.6) require a number of parameters in order to be solved, varying from 5 to 7 for one- and two-diode models respectively. Thus, the accuracy of the 5 (or 7) parameters also affects the accuracy of the model and subsequently the accuracy of PV performance prediction. Yet these parameters are not readily available as they are not provided on manufacturers' data sheets. Manufacturers provide the following information [34]: open circuit voltage (V_{OC}), short-circuit current (I_{SC}), voltage (V_{MP}), and current (I_{MP}) at maximum power point (MPP), temperature coefficient of open-circuit voltage and the temperature coefficient of short-circuit current, measured at STC. PV manufacturers typically provide I-V curves at other environmental conditions for specific PV module models within the datasheet. To extract these five or seven parameters and enable modelling at a wider variety of operating conditions, there is a plethora of proposed solutions, especially for the five-parameter model starting as early as 1963 [35]. Generally, I-V parameter extraction methods can be classified into two main categories, based on the applied approach: those that employ analytical expressions and those that use curve fitting and numerical optimisations [36].

The first category is a more practical approach since it utilises information that can be taken from manufacturers' datasheets. Specifically, to solve for a number of unknown parameters, an equivalent number of equations is required. Three of these equations can be taken from the three characteristic points on the I-V curve, namely open circuit, short circuit and maximum power point. The remaining equations can be the reciprocal of slopes at the open circuit and short circuit conditions which are used to estimate shunt and series resistance or by utilising temperature coefficients or by using combinations of the above [29],

[36]–[41]. A review of five analytical models explaining their physical and mathematical assumptions is given in [17]. Model differences compared to manufacturer's I-V curves are generally insignificant, except for those models that employ several simplifications in order to reduce the computational time [42] such as assuming diode ideality factor as unity or photocurrent equal to short circuit current. Other differences are due to the assumptions made on the dependence of the modelling parameters on irradiance and temperature (see 2.2.1.2). Some analytical solutions exploit mathematical tools such as the Lambert W [43] and co-content functions [44] in order to convert Equation (2.3) into an explicit form and solve for $I = f(V)$. These show satisfactory results with relatively fast solutions but the mathematical expressions are more complex to apply.

Numerical methods employ iterative optimisation algorithms and curve fitting techniques based on experimental I-V curves. Other so-called "soft computing" approaches are popular due to their capability of handling non-linear equations relatively easily and through embedded toolboxes in popular software packages (for example MATLAB/GNU Octave) but their nature is primarily stochastic. Examples of popular optimisation algorithms employed in parameter extraction with modern computing are genetic algorithms [45] differential evolution [46], simulated annealing [47] and particle swam optimisation [48]. Other "conventional" optimisation algorithms are the Simplex and Levenberg – Marquardt methods [45]. The accuracy of the parameter extraction techniques mainly relies on three elements: the choice of the initial guess values, the optimisation algorithm and the error criterion [49]. The performance of any optimisation algorithm depends on its initial starting point, however some algorithms are more sensitive than others. Moreover, the speed of convergence and the determination of the appropriate solution also depends on the algorithm as well as the error criterion, as shown in early work [50]. That is, evaluating how close the simulated curve is to the experimental one. Solving Equation (2.3) may lead to many different solutions depending on the chosen error criterion. Furthermore, the quality and number of the measurement points are important factors that should be considered prior to applying the extraction method, as they may also affect the accuracy, for example if there is significant measurement noise in the data (instabilities in power supply or small variations in irradiance and temperature during measurements) [51].

Overall, the most important part in the list of all methods applied in literature is the achieved accuracy of the extracted parameters, which is further validated by testing these to

predict power output at various irradiance and temperature levels than STC. That is because in their majority and especially for the models which use data from manufacturer datasheets, the extraction of the modelling parameters is applied at STC. Additionally, the choice of the parameter extraction method also depends on the availability of measurement data, otherwise curve fitting techniques cannot be applied.

2.2.1.2 *Variation of modelling parameters with irradiance and temperature*

The modelling parameters namely n, R_S, R_{SH}, I_{PH} and I_0 are extracted at STC if no measurement values are available at other conditions. For the majority of the studies, only I_{PH} and I_0 are assumed to change with irradiance and temperature while n, R_S, R_{SH} remain constant (static parameters) [31]. Although this assumption may be a simplification, it describes crystalline PV modules well [15]. In the case of thin film solar cells, empirical correlations can be found for R_S and R_{SH} [49], [52] which are validated by using experimental I-V curves.

In literature, there are several empirical expressions proposed for the diode saturation current [53] but the most commonly employed Equation for I_0 is given by [54]:

$$I_0 = I_{0_{STC}} \left(\frac{T}{T_{STC}}\right)^3 exp\left[\frac{1}{k}\left(\frac{qE_{g_{STC}}}{T_{STC}} - \frac{qE_g}{T}\right)\right] \tag{2.14}$$

Although, for this expression, the energy bandgap needs to be known. The effective variation of energy bandgap with temperature is described by the following empirical expression for semiconductors [55]:

$$E_g = E_{g_{STC}} - \frac{aT^2}{T + \beta} \tag{2.15}$$

Where, α, β are empirical coefficients and E_{gSTC} is 1.121 for crystalline silicon at STC. For crystalline silicon solar cells a commonly employed equation is [54]:

$$E_g = E_{g_{STC}}(1 - 0.002677(T - T_{STC})) \tag{2.16}$$

And the photocurrent is given by I_{PH} [56] :

$$I_{PH} = I_{PH_{STC}} \frac{G}{G_{STC}} (1 + K_i(T - T_{STC})) \tag{2.17}$$

Where

G = Irradiance (W/m^2)

T = Device temperature (K)

K_i = Temperature coefficient for short-circuit current (A·K^{-1})

E_g = Semiconductor bandgap (eV)

k = Boltzmann constant = 8.167·10^{-5} eV·K^{-1}

And the temperature coefficient μ_Q, of a measured quantity $Q(T)$, is given by the following formula, at two temperatures, T_1, T_2:

$$\mu_Q = \frac{dQ}{dT} = \frac{Q_2(T_2) - Q_1(T_1)}{T_2 - T_1} \tag{2.18}$$

So far, the effect of irradiance and temperature on I_{SC} and V_{OC} is not mentioned. Short circuit current relation with irradiance is described by the same equation as (2.17) by replacing I_{PH} with I_{SC}. The relation of V_{OC} with temperature can be given by the following equation, (by setting I = 0 in Equation (2.3):

$$V_{OC} = \frac{nkT}{q} \ln\left(\frac{I_{PH}}{I_0} + 1\right) \tag{2.19}$$

The effect of irradiance and temperature on the I-V curve of a PV device is shown in the following graphs for a PV module.

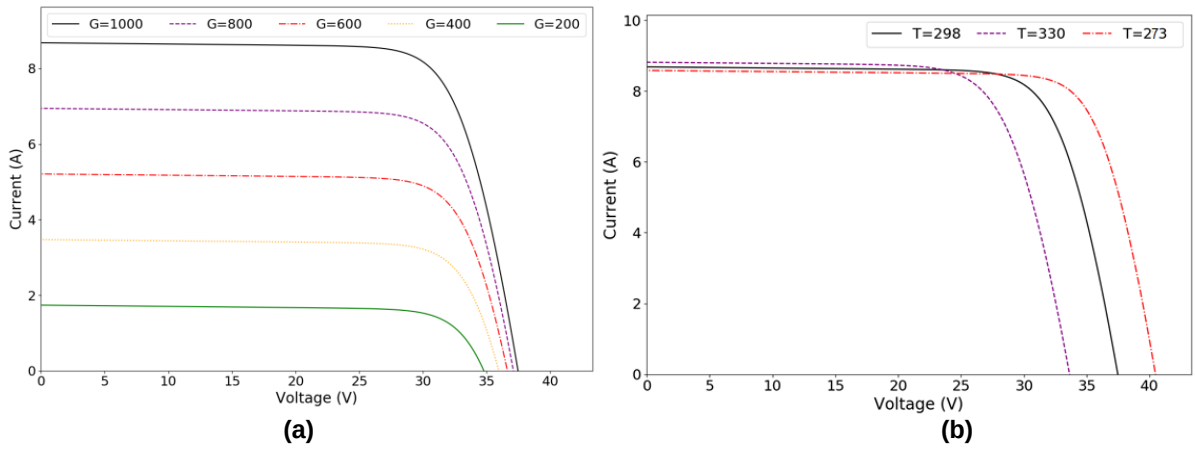


Figure 2.8. Effect of irradiance (in W/m^2) (a) and temperature (in Kelvin) (b) on the I-V curve of a PV device is shown in the following graphs for a simulated solar cell.

Measured I-V curves can also be directly translated from one condition of irradiance and temperature to another, by using *translation equations* directly for current (I) and voltage (V) without translating all the five parameters and repeating the modelling procedure. These comprise semi-empirical equations based on the one-diode model and empirical correction factors [39], [57]–[62] such as the IEC standard 60891 [41] and the module energy rating model (MER) [46]. These equations mainly employ the temperature coefficients at short-circuit current and open-circuit voltage and correction factors for temperature and irradiance determined from indoor or outdoor measurements at different conditions. The drawback of these methods is that they depend on the availability of a set of measurements carried out at certain conditions, as for example described in [60] and in [41].

2.2.2 Empirical modelling

Empirical models are quite simplistic in nature as they only estimate key points on the I-V curve, for example the maximum power point (V_{MP}, I_{MP}). Thus, the modelling procedure is by far less computationally intensive. Empirical models use a variant number of fitted parameters for the calculation of the maximum power output. As a consequence models may vary significantly in the used expression and simplicity [8], [63], [64]. Furthermore, empirical models are power (P) or efficiency (η) based. Efficiency and power are generally related with the following equation:

$$P = \eta \cdot G \cdot Area \tag{2.20}$$

Where G is the irradiance on the PV surface area. Empirical models are used to describe PV modules, for example the models proposed for Energy Rating (ER) standards [8], [65]–[67] but also for modelling larger PV arrays [4],[68]. The power output (or efficiency) is usually obtained by fitting a function of total irradiance and module temperature. In their simplest form, empirical models can be described by a simple matrix of irradiances and temperatures multiplied with a device descriptor matrix, which is obtained either by using outdoor or indoor measurements at different operating conditions.

One of the most employed empirical models is Sandia's array performance model (SAPM) [69] which calculates maximum power with an algebraically simple method, using two extra points on the I-V curve (I_x, I_{xx} at $V_{OC}/2$ and $(V_{OC}+V_{MP})/2$ respectively). This model includes spectral and angle of incidence phenomena, which are described with polynomial functions of fourth and fifth order respectively. The disadvantage of this model is that it requires many empirical parameters (about 27 parameters in total) which need to be determined using real measurements. These can otherwise be obtained from Sandia's database for known modules.

A simplified and popular version of King's model [70],[71] can be obtained for maximum power estimation where small simplifications take place, omitting spectral corrections introduced in SAPM. This model only requires in-plane irradiance (G) and module temperature (T) and predicts power output with good accuracy for crystalline silicon modules [70][65]. The power output is given by:

$$P'(G',T') = G' \cdot (1 + k_1 \mathrm{C} + k_2 C^2 + k_3 T' + k_4 C \cdot T' + k_5 T' \ln C^2 + k_6 T'^2) \tag{2.21}$$

Where,

G' = Normalised irradiance to STC = G/G_{STC}

P' = Normalised maximum power to STC = P/P_{STC}

T' = Module temperature difference from STC = $T - T_{STC}$

C = $\ln(G')$

$k_1 - k_6$ = empirical coefficients

Empirical models that only depend on G, T input and predict maximum power are called single-point efficiency models or power matrix or performance surface models as they produce a 3D surface. However, also in this case, power (P) measurements at different operating conditions of irradiance (G) and module temperature (T) must be realized in order to produce the surface and extract the coefficients for a particular module or array.

2.3 Inverter modelling

So far, it has been shown how a module is formed from PV cells, a string by connecting PV modules in series and a PV array by combining strings in parallel. A PV system as a whole includes further components, the most important of which is the inverter, which converts DC output power into AC output. A key role of the inverter is to optimise the load in order for the connected strings or arrays to operate at their maximum power point (see Figure 2.4), which is realised by an embedded maximum power point tracker (MPPT). Thus, in order to describe the performance of a PV system, it is necessary to model the behaviour of the inverter, incorporating operational losses affecting PV performance. The instantaneous conversion efficiency of the inverter is given by:

$$\eta_{\mathrm{INV}} = \frac{P_{AC}}{P_{DC}} \tag{2.22}$$

Where,

P_{AC} = AC power output (W)

P_{DC} = DC power output (W)

However, inverter efficiency is not fixed but it rather varies with *input power* and *input voltage* [72]. Typically, inverter datasheets will include maximum efficiency (usually between 95% and 98% depending on the inverter technology), maximum power point tracking efficiency (usually between 95% and 100% depending on the algorithm) as well as a weighted

efficiency to account for the operation at different input power levels, such as the Euro efficiency [73] in Europe. Efficiency curves are also provided at different levels of input voltage as seen in Figure 2.9 for a typical commercial inverter.

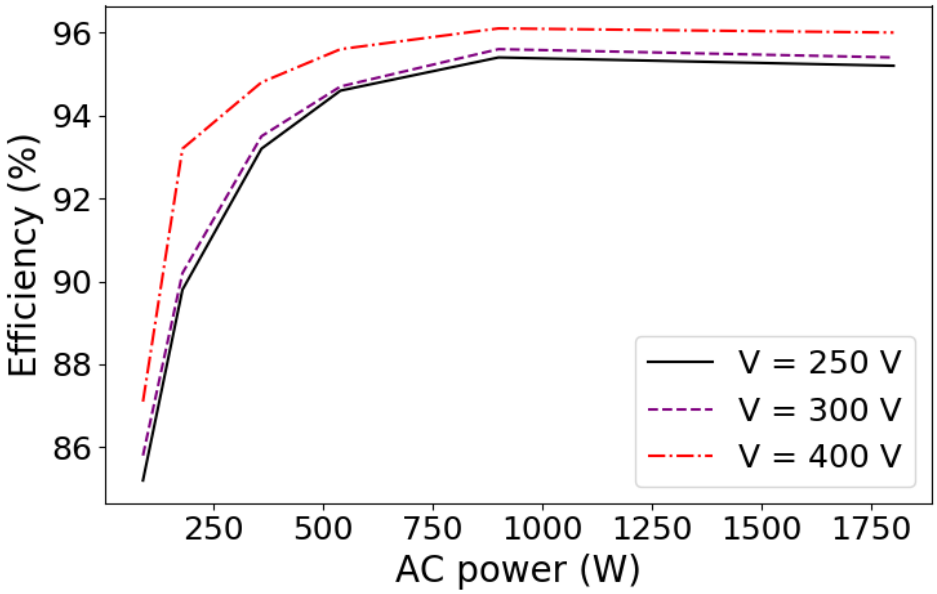


Figure 2.9. Efficiency curves with power output for different levels of input voltage for a typical commercial inverter.

Most manufacturers provide inverter efficiency with regards to the output power (P_{AC}) and for different input voltages: the minimum input voltage, nominal input voltage and 90% of the maximum input voltage. For operation at input voltages and power, other than those given in the datasheets, interpolation can be implemented as for example shown in Figure 2.10 (see Chapter 5).

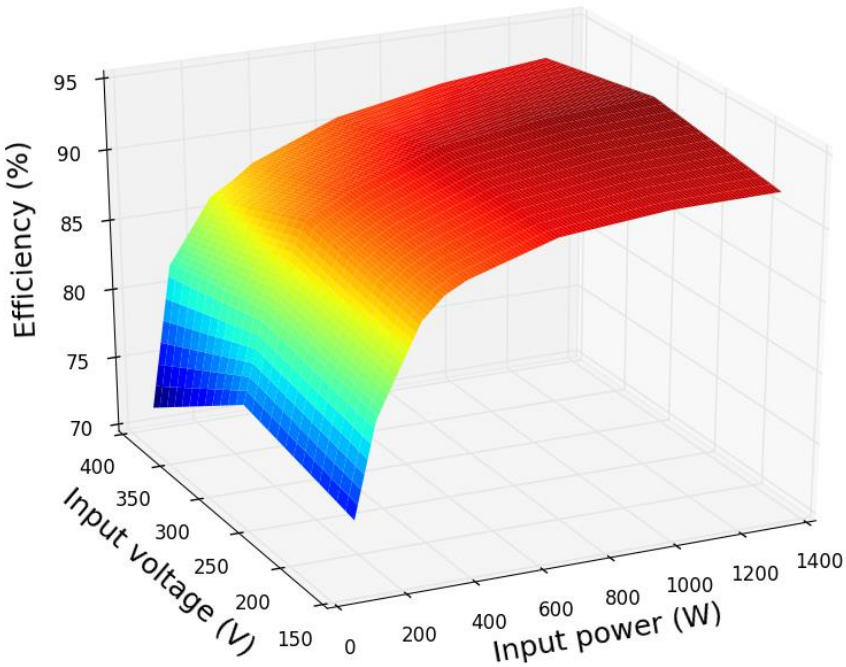


Figure 2.10. Interpolated surface of efficiency *vs* input voltage *vs* input power for a commercial inverter.

For system modelling, if efficiency curves are available at various input power and voltage levels as shown in Figure 2.9 and Figure 2.10 then this is preferred over weighted (for example Euro) efficiency while the latter is preferred over peak efficiency.

2.4 Factors affecting photovoltaic performance

The overall performance of a PV system results from the performance of its components, predominantly PV modules and inverters, which are in turn affected by a number of factors. These generate a different impact on the overall efficiency of the plant. Energy losses are present at all stages of solar energy conversion. Specifically, losses can be categorized into pre-module losses, module losses and system losses. Losses due to downtime periods should also be taken into account. A PV system comprises three main units: the PV generator unit, the string combiner unit and the power conditioning system. The PV generator unit is the PV array. The string combiner unit includes the connections and wiring between strings as well as the fuses/block diodes, which are employed for string overcurrent protection. Finally, the

main parts of the power conditioning system are the inverters, AC connections and wiring. All three units are associated with operational losses which are discussed next.

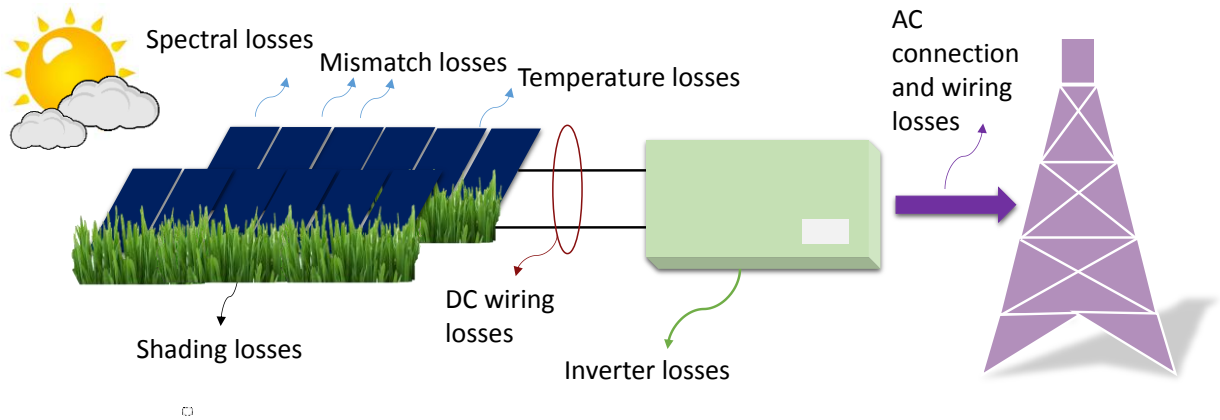


Figure 2.11. Simplified sketch demonstrating energy losses in a PV system.

Shading [74], temperature[75],[76], dust [77], module [78] and inverter [79] losses as well as electrical mismatches [31] are the most common sources of energy loss. Shade, dirt, soil, snow, reflection and spectral effects can be considered as pre-module effects since they prevent part of solar radiation from being absorbed by the modules. Shading can be either due to far (for example hills) or near objects (trees, chimneys etc.) with regards to the PV array vicinity. However, (partial) shading, dirt and soiling technically are not inherent PV system losses but installation or location specific losses and in many cases can be avoided or remedied. Module losses are due to mismatch effects, ageing and temperature as temperature can result in high conversion losses. On the system's side, wiring, maximum power point tracking, inverter and transformer are the primary loss factors.

Spectral and angle of incidence effects have a stronger dependence on the material and the type of the PV device. It has been shown in several studies that this effect is more evident for semiconductor materials with larger energy band gaps such as amorphous silicon while it is less evident in crystalline silicon modules [80] [81]. On an annual basis, spectral losses are considered only a small percentage of the energy output but this also varies amongst different technologies [81]. For crystalline silicon modules, which is predominantly used in small and large scale PV systems, an example of maximum moderate gain of +4% is observed in the winter and a very small loss of -0.8% is observed in the summer in Switzerland and countries with similar weather patterns [82].

Temperature losses are caused due to high operating module temperature with regards to STC and this effect varies for different PV technologies [83] and locations [80] as well as different mounting configurations [84] as the latter affects the natural ventilation of the PV modules. Mismatch losses are caused due to variations in the physical parameters of the modules as well as possible defects. These defects might be caused during installation, manufacturing or transportation and usually appear at the beginning of their lifetime, also known as infant failures [78]. These losses are typically about 2% and may slightly increase with PV system size [85].

Additionally, inverter sizing plays a significant role in power losses. If the inverter is undersized then power is clipped for higher irradiance levels, whereas if it is oversized, its efficiency is too low for lower irradiance levels [86]. Other losses with regards to inverter are the MPPT mismatch losses, which for modern inverters are less than 1%.

Finally, wiring losses are caused from series resistance in electrical connections between modules and strings (DC wiring losses) or between the inverter and the grid (AC wiring losses). In larger systems of several megawatts energy losses due to electrical mismatches only, are about 2% and decrease for smaller systems [31]. These effects can be minimised by choosing the appropriate cables and by reducing their length. The wiring losses under normal operation will depend on national guidelines, but are typically 1-3% for systems of several hundred Watts and above [87].

Overall losses are typically described using a Sankey diagram such as in [88]. Generally, these can be assessed by calculating the performance indicators and increase over time due to various module degradation modes. These mainly concern the semiconductor material (e.g. stresses due to temperature, humidity, thermal cycling, high voltage etc.), the cell interconnects (due to increased thermo-mechanical stresses) and the packaging material (e.g. glass breakage, cracks, browning of the encapsulant, delamination) [78][89]. Increased losses and potential component failures are detected via analysing the electrical output of the system in fault detection.

2.5 Photovoltaic performance monitoring

Monitoring the performance of PV systems, of any size, is necessary in order to detect and identify system faults as early as possible. Performance optimisation, operational efficiency

and system uptime are the main considerations for the operation and maintenance (O&M) of a PV system. System monitoring is therefore an integral part of a cost-effective O&M as it enables system owners to diagnose and repair faults, minimise downtime and mitigate revenue loss.

Monitoring systems differ for residential, commercial, or utility scale PV systems in terms of the number of utilised sensors and the monitoring granularity (e.g. sub-array or string inverters). As the system size increases, monitoring granularity becomes more critical but also economically more viable. This means that for smaller PV systems, for example domestic PV systems (≤ 4 kWp in the UK) the costs for monitoring become proportionally larger compared to the overall installation costs. For example, a relatively cheap pyranometer (Kipp & Zonen CMP3) costs about 800 £, which is currently between 9 and 12% of the overall cost of an average 3 kW_P system [90]. By omitting this cost, the near-term financial benefits are increased making PV investments more appealing, especially considering the continuous degradation of the FIT in the UK [90]. In fact, only energy (or power) readings are usually available. At the simplest level, cumulative energy production in kilowatt-hours, is recorded by electricity meters. The exact metering arrangement varies depending on whether the site has a high or low voltage connection and whether the country has a feed in tariff or net metering for renewable energy systems. These meters are usually operated by the public utility company or its contractors and may also be remotely accessible. Modern 'smart' electricity meters generally have the capability to measure a wide range of electrical parameters relating to power quality and at relatively low cost, however energy or average power is more commonly the only recorded parameter. Due to low monitoring granularity in domestic PV systems, occurring failures may require time consuming tests and expensive maintenance visits, which increases the lifetime O&M costs of the system. A large proportion of domestic systems in the UK are under an infrequent (for example every 6 months) or no maintenance plan, which increases the risks for achieving the investment potential [91].

Guidelines on the minimum requirements on monitoring data are given in the IEC standard 61724 [3]. In practice the number of available monitoring parameters may be less than indicated in this standard, since as already mentioned different monitoring approaches are applied in different scale PV systems. The monitored parameters which are normally employed are summarised in Table 2.1, grouped in three different categories of data availability.

Monitoring data availability as well as the type of collected data, namely electrical and meteorological, differs according to the applied monitoring strategy. Commonly, the type of monitoring system depends on the size and the inverter topology of the PV system it is connected to. In larger projects, stakeholders typically require a higher degree of visibility of system performance. A higher level of monitoring granularity can be provided by a dedicated string level monitoring system. DC and AC current, voltage, power and energy are recorded in addition to network and inverter status. The granularity of data collection will further depend on the number of inverters used in the array, for example using a larger number of smaller inverters will provide more detailed data than a single centralised inverter. Dividing the array into smaller parts offers faster and more efficient fault detection, as the problematic areas are more easily located, especially for larger PV plants. In some cases, inverters have isolated inputs with independent monitoring, which further reduces module mismatches and increases system efficiency. High level monitoring provides a better insight for O&M into system performance, thus minimising the risks of reduced energy production and downtime.

Table 2.1. Monitoring parameters for three common cases of data availability

Parameters	symbol	Low kWh meter only	Medium Inverter built in monitoring	*Medium +* Inverter built in monitoring with added features	*High* Detailed monitoring system
In- plane irradiance	G_{in}	-	-	✓	✓
Ambient temperature	T_a	-	-	✓	✓
Module temperature	T_m	-	-	-	✓
String current DC	I_{DC}	-	-*	-	✓
Array current DC	I_{DC}	-	✓	✓	✓
Array voltage DC	V_{DC}	-	✓	✓	✓
Array current AC	I_{AC}	-	✓	✓	✓
Array voltage AC	V_{AC}	-	✓	✓	✓
Current, Voltage at maximum power point	I_{MPP} V_{MPP}	-	✓	✓	✓
Array output power (DC)	P_{DC}	-	✓	✓	✓
AC Energy (AC)	E_{AC}	✓	✓	✓	✓

*: String current or even module DC current may be available with some systems for example those with micro inverters.

2.6 Review of photovoltaic system performance assessment

There are numerous studies in literature focused on analysing the performance of photovoltaic systems whether these are ground mounted PV plants or rooftop installations. These are focused on a single or bulk of PV installations, for case studies per country such as France [92], Belgium [93], UK [94], Germany [95], Greece [88], Spain [96] and for wider projects including more than one country [97], [98]. Country-wide projects spawned in the 1990s and early 2000s with the incentive to monitor and analyse PV rooftop installations; for example the German 1000 roof PV programme [99] and the domestic field trials in the UK [2]. Groups of domestic buildings were monitored in order to allow information to be collected on buildability, reliability, maintainability and PV performance under real operating conditions. Part of the efforts on monitoring and analysing PV performance has been dedicated to enabling the improvement of guidelines for a better planning and design of PV systems. In this context, the international energy agency (IEA) Photovoltaic Power Systems Programme has released a series of annual reports and Tasks summarising methodologies and outcomes as well as suggestions for improvement on various topics; as for example on performance evaluation of grid connected PV systems [100] and monitoring practises and fault detection [85].

Long term performance and reliability of PV installations are reviewed based on field experience and common loss factors are identified [101]–[103] based on selected performance indicators. Performance indicators are related to performance guarantees which ensure that the system will produce a certain amount of energy each year. The most common indicators used to assess the performance of a (grid connected) PV system are the following, corresponding to both DC and AC side of the system:

Final yield, $$Y_F = E_{AC}/P_{STC} \tag{2.23}$$

Reference yield, $$Y_R = H/G_{STC} \tag{2.24}$$

Array yield, $$Y_A = E_{DC}/P_{STC} \tag{2.25}$$

Array performance ratio,	$PR_A = Y_A/Y_R$	(2.26)
System performance ratio,	$PR_S = Y_F/Y_R$	(2.27)

Where,

E_{AC} = Energy output at the AC side of the inverter (in Watt hours – Wh)

E_{DC} = Energy output at the DC side of the inverter (in Watt hours – Wh)

G_{STC} = In-plane irradiance (W/m^2) at STC = 1000 W/m^2

P_{STC} = Nominal capacity (W_P) = peak power at STC

The performance ratio (PR) is a dimensionless quantity and it is the most common metric used to assess the performance of a PV system as it enables comparison of systems of different power ratings, by normalising the energy produced under actual operating conditions to the rated power at STC of the module and the incident solar radiation. The PR is usually calculated either on a monthly basis to determine performance variations within the year or annually to determine performance losses over time. Typical module PRs (MPRs) differ for various module technologies [80] or between different modules as well as location. This is because PV power output is affected by additional parameters, such as spectral effects and temperature which are not included in the PR expression but they implicitly affect system performance (see 2.4).

The final (or array) yield often referred as kWh/kW$_P$ is a common performance metric and it is usually calculated on annual terms. This indicator is directly translated in terms of produced energy normalised to the rated output of the system and allows comparing systems of different sizes. However, due to the fact that it does not include incident solar radiation it is not a suitable metric for comparing systems installed at different inclination and azimuth angles or systems in locations with significantly different solar resource. Specifically, countries with medium solar resource report an average annual kWh/kWp of 700-900 kWh/kWp [93], [99] whereas countries with higher solar resource report values over 1300 kWh/kWp [88].

These values, however, are not strictly indicative as they refer to PV systems installed at optimal angles.

To evaluate a PV plant's performance, based on losses, the most important indicators are: the final yield (Y_F), the array yield (Y_A), the reference yield (Y_R), and the performance ratio (PR) as defined by the IEC Standard 61724 [3]. Losses can then be calculated using yield results. So, system losses (L_S) reflect the inverter and transformer conversion losses, and the array capture losses (L_C) are due to the PV array losses.

$$L_C = Y_R - Y_A \tag{2.28}$$

$$L_S = Y_A - Y_F \tag{2.29}$$

Although the aforementioned indicators are most commonly employed, additional indicators are utilised in literature such as the performance index, the ratio of actual to theoretical output for e.g. in [92]. This represents conversion losses compared to a theoretical ideal system of the same characteristics but without the inverter losses. Also, system efficiency as the actual output per incident solar radiation per system area, is a very common metric [104]:

$$\eta = \frac{E_{AC}}{G \cdot Area} \tag{2.30}$$

In overall, performance evaluation can be carried out based on temporal differences, theoretical output and/or annual performance distributions for more than one system. In the first case the indicators are applied on one system each time and can be calculated on monthly and annual basis. This shows performance trends but the method is not indicative whether the system actually performs as expected. To realise that, expected output is compared to actual output employing measured weather data in a performance model [6] of choice, namely using a reference system. Finally, in the latter case a large statistical example is utilised, usually for systems in the same or similar locations (example [104]). Selected

performance indicators are presented in terms of statistical distributions. Using statistical distributions for example of PR assists in detecting outliers in large samples, namely systems with unexpected performance which then enables to then focus on the detected cases.

2.7 Performance assessment using remote weather monitoring

As mentioned earlier, in larger PV systems monitoring granularity becomes more critical and therefore higher level monitoring strategies are applied. Conversely, for smaller systems monitoring equipment is usually not financially viable and therefore, small scale projects such as domestic systems (about 4 kWp) are typically not equipped with climatic sensors, hence analysing data from these systems is primarily based on approximations to local conditions. In such cases in-plane irradiance and module temperature must be acquired with alternative and indirect methods. Global horizontal irradiance and ambient temperature can be acquired from meteorological stations as well as satellite imagery. The acquisition of localised weather data is non-trivial and requires modelling and interpolation techniques in order to determine local conditions and to then convert global horizontal irradiance and ambient temperature into in-plane irradiance and module temperature respectively.

There are various products of satellite data, and their main differences lie in their spatial, spectral, temporal and radiometric resolution [105]. Spatial resolution corresponds to the size of the image pixel that corresponds to the Earth's field size. Spectral resolution corresponds to the employed spectral channels. Additional information on the spectrum of the reflected radiation is useful in order to determine effects that have little differences in the visible range, for example snow cover and clouds. These effects can result in significant overestimation of GHI. Temporal resolution is the time between collections of images on specific locations. Finally, radiometric resolution determines the ability of a satellite image to record various levels of brightness [106].

Ground based data are acquired from meteorological ground based stations. The difference between the two sources is that global horizontal radiation is directly obtained from ground based sensors (usually pyranometers) whereas in the case of satellite imagery this information is extracted from Earth and atmosphere reflected solar radiation by applying

additional modelling algorithms. However, the spatial coverage from satellite data is larger as opposed to ground based measurements which are only available at the point of measurement. In this case, in order to obtain global horizontal irradiance at different locations, spatial interpolation algorithms need to be applied [4],[107]. To date there aren't any conclusive studies as to which source is more suitable for PV applications but the trends are rather towards using satellite data mainly due to their larger spatial coverage, data continuity and historical repository [108]. However, comparison with available ground based data is a common way of validating satellite data. A characteristic example is by using the Baseline radiation surface network stations (BSRN)[1] as per [109]. It has been found that the efficiency of satellite data decreases in high latitudes (over 50 degrees), mountains, high albedo areas (such as deserts and showy areas) and in cases of other rapidly occurring weather and aerosol concentration variations in the atmosphere [108]. So, in fact there can be combinations of both satellite and ground based measurements in order to optimise prediction accuracy [110].

Various studies have employed remote weather monitoring for the performance assessment of distributed small scale PV systems, which is also used for the detection of failures [4], [92], [93], [98], [105], [111]. In those cases, the accuracy of the solar radiation data mainly determines the accuracy of the performance assessment. Specifically, annual global horizontal irradiation can be predicted with an average error of -5 to 8% [4] using satellite data but differences can be larger and up to 4% depending on the utilised satellite source [105]. Additional modelling errors arise from the modelling steps required for the translation of global horizontal irradiation to in-plane, as discussed next. The entirety of the models tested in literature are found to underestimate in-plane irradiation compared to measured data up to 13% annually [112]–[116] which, consequently, affects the estimation of PV performance and performance ratio.

2.7.1 Translation of weather data onto system specific conditions

In this step, the acquired global horizontal irradiance (GHI) is translated to the tilted plane of the PV array. This is normally realised by employing two separate algorithms; the first

[1] The Baseline Surface Radiation Network (BSRN) is a world-wide collaboration of organizations which maintain high-quality ground measurements. BSRN stations are only two in the UK, in Camborne and Lerwick.

algorithm requires that global horizontal irradiance is separated into its components namely beam and diffuse. The second algorithm requires that both beam and diffuse components are then translated to the tilted plane in two distinct steps. The split of global horizontal radiation to its beam and diffuse components relies upon determining the clearness index, given by:

$$k_{\mathrm{t}} = \frac{GHI}{ET} \tag{2.31}$$

Where ET represents the extraterrestial radiation. Therefore, the clearness index is a measure of radiation attenuation in the atmosphere, or else “cloudiness” in the sky.

The next step is to define the ratio of diffuse irradiance to GHI with regards to the clearness index since it is:

$$\frac{G_d}{GHI} = X(k_t) \tag{2.32}$$

Where $X = f, g, h$ corresponding to three portions of the clearness index. Theoretically, clearness index can be the only predictor used to estimate the diffuse fraction of GHI [117]. That, however, only gives a one-dimensional approximation of reality. More sophisticated models include a number of predictors such as ambient temperature, humidity, sun elevation and other factors which gives a better approximation [118],[119] but also assume knowledge of more input variables, which in remote monitoring applications is not always available.

In the translation part, ideally a third step can be included which involves the ground reflected albedo on the PV surface. In this case the total in-plane irradiance is ultimately given by [56]:

$$G_{POA} = G_b \cdot cosz + G_d \cdot R_d + \rho \cdot GHI \cdot R_r \tag{2.33}$$

Where,

G_b = beam irradiance

G_d = diffuse irradiance

GHI = global horizontal irradiance

R_d = diffuse irradiance transposition factor

ρ = ground reflected albedo

R_r = ground reflected irradiance transposition factor

z = angle of incidence of the beam on the tilted plane

The transposition factors describe the ratio of the incident irradiation (diffuse or beam) on the plane of array, to the global horizontal irradiation. Beam radiation can be readily translated onto plane of array as this can be realised by applying geometric terms [56], having calculated the position of the sun in the sky (for example refer to [120] for the calculation of sun position). Conversely, diffuse radiation on plane is generally difficult to model as its spatial distribution is unknown and time dependent. The simplest model for diffuse translation is based on *isotropic sky* assumption. This essentially assumes that diffuse radiation is uniform across each point at the sky. In order to improve the accuracy of diffuse radiation translation, anisotropic models are proposed [121].

Module or cell temperature is another essential modelling parameter as it plays an important role in device physics. As already seen, the voltage of a module decreases when its operating temperature rises, while short-circuit current slightly increases. The result is a decrease in its output power. The operating temperature of photovoltaic modules under real operating conditions is important information for calculating power output. Generally, thermal models can be split according to the simplicity or complexity of their expression [122]. In their most explicit form models only depend on in plane irradiance, ambient temperature and/or wind speed [69],[84], [123]. One of the most popular simple models used for the assessment of module temperature is the Ross' model [124]. As a steady state model, it assumes that for the calculated time the intensity of solar radiation, wind and other parameters that affect PV module performance are constant. The equation is given by:

$$T_m = T_a + k_R \cdot G \tag{2.34}$$

Where,

T_m = Module temperature (K)

T_a = Ambient temperature (K)

G = In-plane irradiance (W/m^2)

k_R = empirical Ross' coefficient (K·m^2/W)

The k_R is known as Ross coefficient and it takes different values according to the mounting configuration of the module, as experimentally determined in [75]. Typical values of the k_R are given in Table 2.2.

Table 2.2. Ross coefficient values for various mounting configurations [75] [125].

Mounting configuration	Ross coefficient (K·m^2/W)
Free -standing	0.021
Flat roof	0.026
Sloped roof: well cooled	0.020
Sloped roof: not so well cooled	0.034
Sloped roof: highly integrated	0.056

These are empirical values that can be obtained from graphical representation of (T_m-T_a) against G. Ross model is sufficient in cases where irradiance and ambient temperature are the only available weather data. Simpler models can be accurate when hourly or generally lower resolution measurements are used as input [122].

Implicit models not only account for ambient temperature, in-plane irradiance and wind speed but also the energy balance between the module and its environment. Therefore, they may include a number of intrinsic factors such as heat transfer coefficients, cell absorption coefficient, glass transmittance etc. [126]–[130]. These models are generally more sophisticated and they describe module temperature more accurately in high resolution time

series since they account for non-steady state conditions and thermal lag. However, they demand a higher number of input variables, which are often not available in remote modelling.

2.8 Main factors affecting quality in remote performance assessment

The lack of a detailed data analysis in domestic monitoring often leads to two situations: the existence of system faults, such as non-generating or generally under-performing systems and the existence of undetected data quality issues, such as missing data and invalid system description. These factors can severely affect the quality of the performance assessment especially with regards the performance ratio (PR) distribution. Assuming that only total generation (E_{AC}) is available in kWh and there is no climatic monitoring, PR is defined as [101]:

$$PR = \frac{E_{AC} \cdot G_{STC}}{H \cdot P_{STC}} \tag{2.35}$$

The parameters that mostly affect the calculation of PR are the accuracy of the modelled in-plane irradiation *H* and the accuracy of the given P_{STC} for each system. Thus, the uncertainty in PR calculation (u_{PR}) can be shown as a combination of factors:

$$u_{PR} = f(u_{E_{AC}}, u_H, u_{P_{STC}}) \tag{2.36}$$

Whereby uncertainty in P_{STC} ($u_{P_{STC}}$) expresses the error in the declared nominal capacity compared to its actual value. Additionally, modelling of irradiation (*H*) is affected by the a) quality of the solar radiation data and b) the validity of the declared installation azimuth and inclination of the PV surface.

2.8.1 Erroneous system description

This is one of the most unpredictable and therefore hard to identify issues. The majority of these are caused by human error and are fairly common in similar applications [4]. Often, these are due to installers or the people entering data working under time pressure or

insufficient training. As an example, Figure 2.12 depicts the impact of different azimuth (South =0) and tilt angles on modelled in-plane irradiance. It is seen that using the wrong azimuth has a higher impact on modelling in-plane irradiance than tilt angle, where the difference is expected to be small for ±10 degrees. Azimuth may be incorrectly derived from satellite images, whilst measurements taken with magnetic compasses or smartphone sensors may be affected by factors such as the presence of nearby metalwork and magnetic declination.

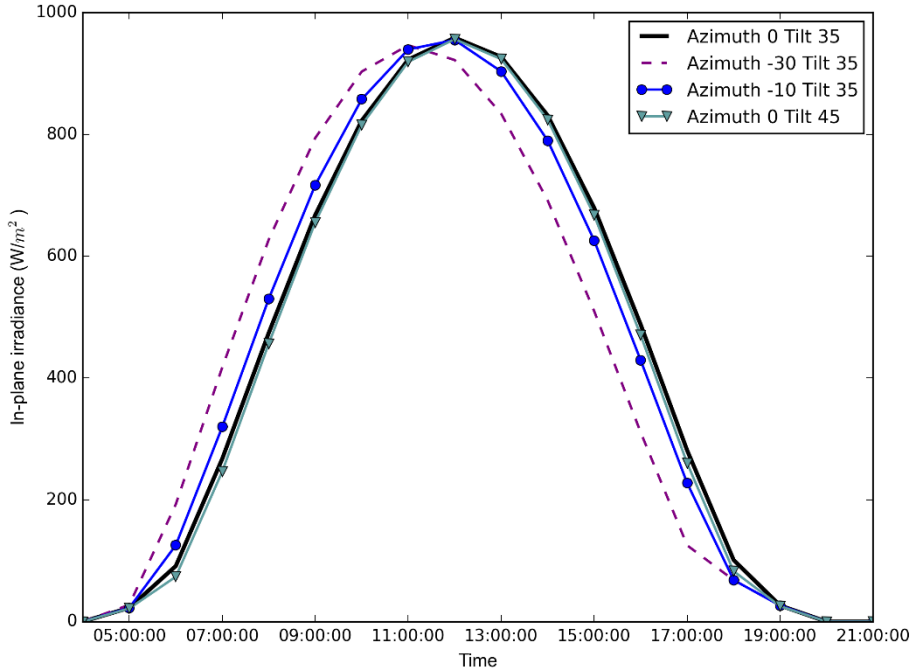


Figure 2.12. In-plane irradiance on a clear day for different cases of azimuth (0,-10,-30) and tilt angles (35, 45).

Erroneous nominal capacity also lies in this category, and such errors are generally not straightforward to verify automatically as opposed to system azimuth. The actual capacity of a PV system cannot be automatically inferred in PV fleet assessments, and has to be verified by the owner. These errors have a higher impact than expected deviations in nominal capacity which are within manufacturer's tolerance. Normally, manufacturers are required to define module nominal rating (P_{STC}) within a given tolerance (for example 0/+5 W_P)[34]. This means that, although PR may be calculated using the nominal capacity from the manufacturers' datasheet, the actual value could be slightly different. Assuming that nominal capacity is

actually higher, then the calculated PR will also be higher than the actual PR since $P_{STC_{model}} < P_{STC_{actual}}$. However, the impact of this on the PR is small compared to other possible errors discussed. In fact, assuming an average system of 8 modules of 245 W_P each, and the extreme case where every module is 5W higher than rated, the expected increase in PR is about 0.03. Thus, deviations of this size in nominal ratings do not justify why some systems report unexpectedly high PRs for example [94].

2.8.2 Timestamp mismatches

This refers to remote environment sensing, where different sources of solar radiation may use differing (temporal) reference systems. Data may be recorded at mean solar time (MST), local time (LT) or coordinated universal time (UTC), which may be a different timestamp system than the one used by the PV monitoring device. Furthermore, for hourly averaged data the timestamp may represent the middle or end of the averaging period depending on the convention used in the system or database. These factors may cause temporal mismatches, which are more evident in sub-daily analyses. Using mixed timestamps affects the disaggregation of solar radiation into beam and diffuse irradiance (via the clearness index calculation). This has a follow-on effect on the estimated in-plane irradiation. Therefore, timestamp conventions should be first examined for both solar radiation and PV system monitoring to avoid mismatches [131].

2.8.3 Remote solar radiation data

Particularly for solar radiation, the quality of performance assessment is significantly affected by the quality of the chosen solar radiation dataset [7]. Furthermore, when it comes to translating global horizontal irradiation to the plane of the PV array, the employed models introduce their own uncertainty and in-plane irradiation is generally underestimated [112]–[116]. This underestimation may derive from the chosen transposition model [115] (albedo and diffuse irradiance underestimation and/or the selection of empirical coefficients) but it appears that it is more affected by the choice of the separation model, which is location dependant [114] and finally, the combination of the above. Particularly, the accuracy of the diffuse component affects the accuracy of the transposition model. Since, there is not a

universally best performing combination between separation and transposition models, the optimal choice would rely on comparisons carried out at same or similar locations and using available measurements of both global horizontal and in-plane irradiance (direct and diffuse).

2.8.4 Missing data

The problem of missing data arises frequently in PV monitoring as also highlighted in major monitoring reports [2],[6]. Missing data often occur due to equipment failure, power outages or monitoring interruption for maintenance reasons. This means that while a system may operate normally, various monitoring parameters such as its energy generation may not be recorded for a period of time. This creates gaps in the resulting monitoring dataset. If these gaps are not treated properly, they lead to false conclusions on the system's performance, as for example when considering "no data" as "no generation". For this reason, in several studies the amount of missing data allowed in the analysed dataset is restricted. For example, in [2] the tolerance of missing electrical data is set at no more than 8% while in [6] is set at 10%, meaning that if missing data exceed that specified threshold within a monitoring period then this period is not taken into account in the performance analysis.

If missing data occur randomly within a large monitoring period (without exceeding a certain threshold), then the analysis may still be carried out for the remaining data. However, in PV monitoring these missing periods are often for consecutive weeks, which may cause a bias in the analysis due to seasonal performance variations. To correct for this bias missing data can be inferred. Inference of missing data is a major concern for example in statistical sciences, but has not been extensively applied in PV monitoring, though relevant work has been shown recently on incomplete reliability datasets [132]. Common statistical approaches for handling missing data are the maximum likelihood (ML) and multiple imputation (MI) methods. The principle of MI is based on simulating different sets of missing values to complete the data using regression models, then combine the results to a single set. ML is based on estimating missing values based on existing data by maximising a likelihood function. The aim in both cases is to infer the missing values with a minimum bias. However, both methods assume that data are missing at random, which is often not the case in PV datasets. In PV monitoring a complete dataset should reflect both the seasonal variations in performance as well as individual system characteristics. This can be realised by using a

performance model to predict system's output when it is missing, instead of a purely statistical approach, as it is further analysed in chapter 4.

2.9 Fault detection

As analysed in Section 2.5 monitoring a PV system is required in order to determine whether a system operates as expected. The definition of a fault is not really standardised but based on the definition for a module failure [78], it can be any occurrence which limits power output beyond a predicted threshold, and cannot be reversed by normal operation. Other occurrences can be situational or location and installation specific for example soiling, dirt, bird nesting or (partial) shading. These factors can potentially damage the modules if not repaired. Various methodologies have been developed for the detection of faults and other factors which increase system losses. Some methodologies also include the identification of specific faults. However, these methods usually rely on more detailed monitoring (medium+ and high in Table 2.1).

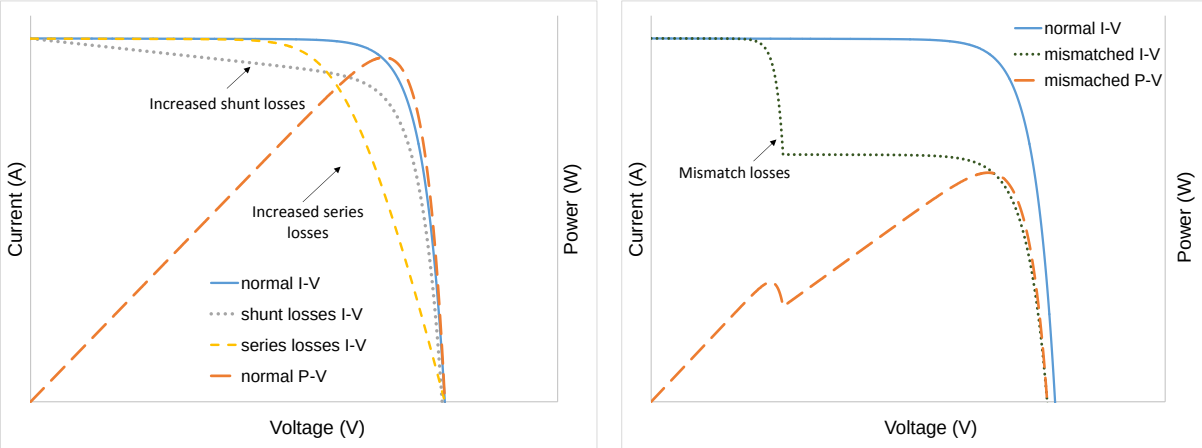
The most common approach in fault detection is analysing the electrical output of a PV system namely its power, voltage and current if these are available from monitoring and comparing this data with a reference system. In this category, fault detection may employ trained models by using past data where the system is known to perform normally. Thus, in the case of a fault, the behaviour of the system will deviate from the one described by the trained model, which will indicate the existence of a fault. Examples of this approach are the application of decision trees (DT) [133] and neural networks [134]. However, these methods require the availability of past data from normal and/or fault conditions and therefore their efficiency largely depends on the availability and the quality of the employed dataset.

Table 2.3. Commonly studied failure modes in literature

Indication	Failure modes examined	Reference
Constant energy loss (hourly, daily)	Degradation	[4]
	Soiling	[4]
	Module defects	[4], [135], [136]
	Faulty string(s)	[111],[137], [138], [139], [135], [136], [133], [140]
Varying energy loss	Mismatch (including shading)	[4], [138], [104], [136], [133]
	Power limitation (MPPT error, decreased inverter efficiency and deration)	[4], [139], [104], [136], [68], [141]
Abrupt energy loss	Snow cover	[4]
	mechanical failure	[4]
Complete outage	Defective inverter	[104], [136]
	Component failure	[104]
	Grid outage	[4]

On a module level power losses are caused due to increased series or shunt resistance which can be seen on the I-V curves of the device, in Figure 2.13 (a) and (b). Shunt losses are commonly caused by module defects and partial shading which create local heating and hot-spot appearance. Module defects are generated during installation, manufacturing or transportation and their effects usually appear at the beginning of their lifetime, also known as infant failures [78]. Series losses are commonly caused by increased series resistance at the electrical connections between modules and strings (DC wiring losses) or between the inverter and the grid (AC wiring losses). These effects can be minimised by choosing the

appropriate cables and by reducing their length. Increased series resistance may also derive from PV modules under specific effects of degradation mechanisms such as solder corrosion or mechanical stresses that cause cell cracks and also corrosion of DC connectors and junction box terminals.



(a) **(b)**

Figure 2.13. Effect on the IV curve of (a) increased shunt, series losses and (b) mismatch losses caused by possible faults in the PV array.

Energy losses are quantified for a defined time period and the result determines whether a fault exists in the system and if that's the case, the next stage is initiated, which is to identify the fault if that is possible. Hence this part relies on performance evaluation parameters and array capture and system losses as described in Section 2.4. As mentioned, array collection losses account for the DC side of the system, and they usually refer to losses due to mismatch effects, wiring, maximum power point tracking errors and temperature losses, but the largest contributors to system losses are usually inverter and transformer losses [88].

The evaluation of changes in performance can be achieved by comparing expected to actual output, for defined timespans as:

$$|Q_{meas} - Q_{sim}| = \varepsilon < threshold \tag{2.37}$$

Where Q_{meas} and Q_{sim} is the compared measured and simulated quantity respectively. For normal operation, this difference must lie within margins specified by a threshold which will discriminate between actual faults and 'false positives'. The determination of this threshold

depends on the described quantities and their actual values. Therefore, the definition of an appropriate threshold is somewhat ambiguous. In this context, (2.37) can be used for performance ratio differences [136]. In this case the threshold can be defined according to the observed daily deviations of the PR.

An alternative representation of (2.37) employs normal distributions of the difference ε over specified domains. This allows the determination of the threshold by calculating the standard deviation, σ, and the mean, ε_{mean}, of the distribution over a specified interval. The margins are then given by:

$$\varepsilon_{mean} - k \cdot \sigma < \varepsilon_{meas} < \varepsilon_{mean} + k \cdot \sigma \tag{2.38}$$

Where k (=1, 2 or 3) is the number of standard deviations used. This equation can be applied for system efficiency (k=3) [104], normalised power (k=2) [4], current and voltage values [142] or collection losses [138].

In higher resolution monitoring, I-V measurements if available can be used in order to detect PV system abnormalities in real time. Studies based on this approach employ I-V modelling to simulate theoretical output, taking into account the systems' behaviour at normal operating conditions. The efficiency of the employed diagnostic indicators depends to a great extent on the modelling accuracy. Therefore accuracy plays a great role in detecting errors as significant deviations from actual values might lead to false positive alarms or worse, increased detection thresholds and reduce the method's efficiency by producing false negatives. String monitoring enables an even better supervision of a PV system, compared to array monitoring, since module defects occurring in strings are more easily located. There are different levels of automatic detection, starting from the top level by identifying at which side of the system the fault occurred, namely AC or DC, (through to the sub-array, string and module level). To determine which side the fault is in, the power ratio, RP_{DC_AC}, can be used as an example [136]:

$$RP_{DC\_AC} = \frac{P_{DC_{sim}}/P_{DC_{meas}}}{P_{AC_{sim}}/P_{AC_{meas}}} \tag{2.39}$$

Where RP_{DC_AC} can be interpreted as follows:

RP_{DC_AC} <1 Inverter fault

RP_{DC_AC} ~1 Normal operation

RP_{DC_AC} >1 Array fault

This assumes that inverter efficiency is calculated based on the specific inverter characteristics, as it drops for low power input [143]. At low power this ratio could yield a false alarm if simulated inverter efficiency is poorly fitted to actual inverter data.

Taking (2.39), if $RP_{DC\_AC} < 1$, this indicates AC power being decreased with regards to the expected inverter output, indicating inverter malfunctioning or faulty connection. In such case, this can also be verified by the following ratio [136] for each inverter in the array:

$$R_{inv} = \frac{P_{AC_{meas}}}{P_{AC_{sim}}} \tag{2.40}$$

Or similarly, based on the residual [141]:

$$R_{inv} = (P_{AC_{sim}} - P_{AC_{meas}}) / P_{AC_{sim}} \tag{2.41}$$

If more strings are present then comparing output from different strings is an additional way of identifying an inverter fault. String faults are the most commonly tested type of faults in literature as their detection routine is relatively easy in terms of their fingerprint. In case of a string disconnection the current of the array drops significantly. So it can be detected initially by using array current and voltage [136], [138], [137]:

$$R_c = \frac{I_{DC_{sim}}}{I_{DC_{meas}}} \tag{2.42}$$

And

$$R_V = \frac{V_{DC_{sim}}}{V_{DC_{meas}}} \tag{2.43}$$

These two indicators of power losses include temperature effects (increased thermal losses) and miscellaneous collection losses such as wiring, mismatch, MPPT errors etc. If R_C >1 and R_V <1, then string disconnection is most likely the fault. But to determine the exact location, then string currents must be compared with simulated values, hence (2.42) be used at each string. Moreover, since module defects may derive from various factors and their effect on current and voltage is more implicit than it is for a faulty inverter or a disconnected string. In large strings power deterioration is relatively small (unless a significant number of modules is affected) and may not be detected in early stage. To accurately determine such faults from I-V signals, the full I-V curve would be required, but as mentioned this is generally not available from commercially operated PV systems. Instead, only maximum power point current and voltage from both DC and AC sides are used [144].

In domestic PV monitoring the situation is radically different. First, climatic data are often not available which decreases the accuracy of fault detection. Second, I-V data are not available and thus identification of faults relies on analysing energy losses in terms of occurrence and duration. In this context, Firth et al. classified faults into four categories: a) sustained zero efficiency, where generation is zero for long time periods, b) brief zero efficiency, where generation is zero for short time periods, c) shading (identified utilising a sun position algorithm) and d) non zero efficiency (and non-shading), for other faults not falling into previous categories [104]. System efficiency (on the AC side of the systems) was plotted against in-plane irradiance and a numerical approach was applied in order to define the boundaries. The outliers in the resulting graph were classified into the four fault categories. This technique relies on analysing long–term performance and hence does not enable automatic detection, but provides a rough classification of common faults such as inverter power point tracking, inverter cut off at high irradiance (power limitation), shading and total outage due to possible inverter shutdown and system isolation.

Similarly, in [4] and [145] failures are determined based on the amount, duration and variations of energy losses as well as comparison with neighbouring systems. Failures are grouped into four categories such as constant energy loss (for example degradation, module defects and soiling), varying energy loss (for example shading, inverter disconnection from the network, power limitation due to inverter), snow cover and total blackout (for example electricity network outage and inverter failure). Because more than one failure may belong in one category, failure rates are also exploited, namely the probability of one failure occurring

compared to another in the same category. The practicality of this method is more evident where only energy readings are available, which is a common case for residential systems. However, the accuracy of irradiance data can potentially decrease the method's efficiency in detecting some faults, especially on non-clear days. This is further enhanced by other quality issues which typically exist in domestic monitoring but are not effectively discussed. Fault detection can potentially be achieved without solar radiation data, only by employing statistical distributions of the performance indices of neighbouring PV systems and detecting the outliers [5]. However, the efficiency of this method primarily relies on the existence of a large number of neighbouring PV systems for the statistical analysis to be more reliable. Again, data quality in all datasets largely accounts for the accuracy of fault detection.

2.10 Chapter conclusions

This chapter reviewed the performance evaluation framework comprising several stages of modelling, performance assessment, PV monitoring and fault detection. Performance models are employed to either assess the efficiency of an existing system compared to its simulated analogue, or in yield forecasting. Modelling of a PV system can be realised by utilising device physics or empirical models. Device physics models are implicit in nature and they require the knowledge of five to seven modelling parameters in order to be solved. These parameters can be acquired by extracting this information using manufacturer datasheets or measured I-V curves. On the other hand, empirical models are more simplistic in nature but can only model specific points on the I-V curve such as the maximum power point, i.e. not the whole I-V curve. The choice of the model largely depends on the application. Where I-V data are available from monitoring, device physics modelling can be employed for the comparison of measured to simulated output and fault detection. Conversely, in cases where only energy output is available from monitoring, simple empirical models are usually used in fault detection.

An inverter model is required to convert PV array to PV system modelling. The most comprehensive approach is the one that takes into account the dependence of inverter efficiency on varying input voltage and power levels, thus an interpolation needs to be applied in two domains of operation, namely input (DC) voltage and input (DC) power, producing a

three-dimensional graph of inverter efficiency. Further inclusion of operational losses in the performance model produces a realistic output of a PV system. These losses need to be taken into account in order to define reasonable thresholds in the fault detection process. The applicability of any fault detection procedure mainly depends on monitoring granularity, in terms of applied level of monitoring namely module, string, sub-array, array or system level. More often large systems (of over 1MW$_P$) will monitor on string level while small domestic systems (of an average size of 3 kW$_P$) will only monitor on system level. Accordingly, I-V data are not available in all cases, while voltage and current might be available only at maximum power point. High-level monitored systems provide a lot more information on operational and weather data at system location. Additionally, weather monitoring is very often not applied, thus weather profiles must be generated by using information from remote ground meteorological stations or satellite data. This information has to be further translated onto system specific variables such as in-plane irradiance and module temperature by employing further models.

The quality of the performance assessment is affected by various factors which need to be further investigated such as the inclusion of gaps in monitoring and/or possible errors in PV system description which cause PR to falsely increase or decrease. These errors need to be identified and corrected to a large possible extent. This becomes even more crucial in fault detection, specifically in the case of domestic, where only power or energy output is usually available. Fault detection is often based on assumptions derived by comparisons of actual performance with pre-defined performance profiles or the comparison with neighbouring PV systems using remotely inferred climatic data. High quality in performance assessment is crucial so that fault detection efficiency is increased. The aspects of remote monitoring, choice of models and fault detection on domestic systems will be examined in the following chapters focusing on the impact of monitoring data quality.

Chapter 3

Data quality in domestic photovoltaic monitoring

3.1 Introduction

Shortfalls in data quality such as wrong system descriptors or missing data become significant obstacles when attempting robust performance assessments [5]. More importantly, the use of erroneous data may lead to inaccurate and rather damaging statements with regards the advancement of solar industry if left unrevealed (see for example a characteristic case of misleading analysis on PV energy return [146] and the response from the solar community [147]). To date, there are no studies dedicated in the quality assessment on residential PV data sets with minimum monitoring equipment, where the only data available is the technical description of the system and its total energy generation. This chapter focuses on the distinction of these common data quality issues found in domestic datasets, and their impact on the performance ratio estimation as the most established performance indicator. It is shown in the applied case study that even datasets from commercially monitored installations can suffer from unusual 'artefacts' which makes data interpretation extremely difficult. Based on the performance assessment framework summarised in Figure 2.1, the specific tasks which are considered here are highlighted in Figure 3.1. The performance ratio calculation steps by using remotely inferred solar radiation data are further described in Figure 3.1, whereby each step as well as the associated procedures and models involved are also described in this chapter.

Statistical approaches are applied based on calculated annual performance ratios and final yields (kWh per kW_P). This approach is specifically useful for PV fleet assessments, where insight into individual systems and/or higher resolution data are scarce. The utilised data sources and the applied practices in posing quality controls and correcting erroneous entries where possible are demonstrated next, using a case study of 1788 residential PV systems in Nottingham. Data quality is presented within two distinct categories: erroneous system description and missing monitoring data. Where hourly energy output data are possible to obtain, an azimuth correction algorithm is proposed.

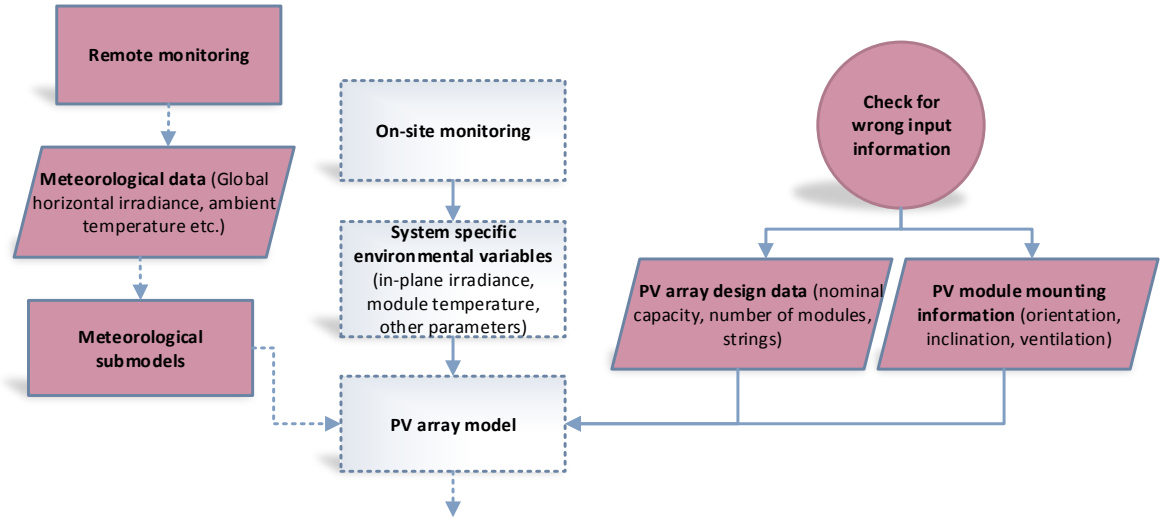


Figure 3.1. Main blocks (highlighted in fuchsia) of the overall performance assessment framework associated with the work described in this chapter.

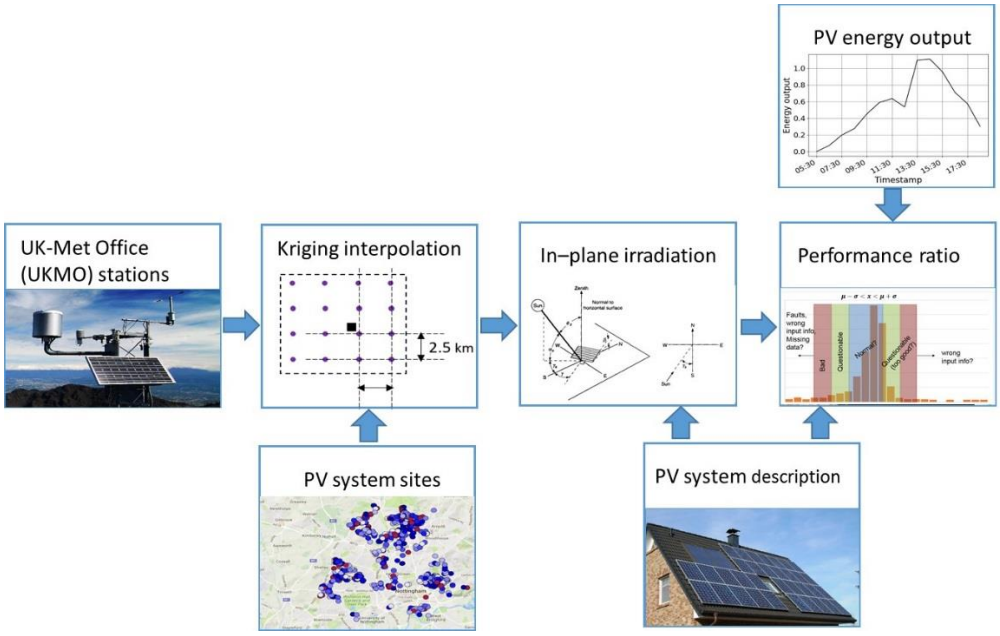


Figure 3.2. Steps applied for the calculation of performance ratio based on remotely inferred solar radiation data from the UK Met Office (UKMO) meteorological stations.

3.2 The Nottingham City Homes (NCH) dataset

Monitoring data from about 1800 domestic installations varying from 1kWp to 4kWp (see Figure 3.3) have been gathered from the systems' commercial monitoring portal for the years 2012 to 2015. The procedure for downloading and sorting this data is described in the Appendix. This data set belongs to a social housing association (Nottingham City Homes – NCH)[148] who control over 28000 rented houses in Nottingham. Essentially, the council owns the homes, and NCH manage them on its behalf but an independent monitoring company has taken over to collect this data and provide rough indications on their performance, namely categorise them into generating, non-generating and low performing systems on a daily basis. For the case study, the availability of measured performance data was in excess of 98% for the majority of systems. Information available for these systems includes location, rated (peak) capacity (kWp), PV module model, inverter model, elevation, azimuth and inclination as summarised in Table 3.1. Frequency of installation capacity (as declared) in terms of peak power is given for 1788 PV systems in Figure 3.3.

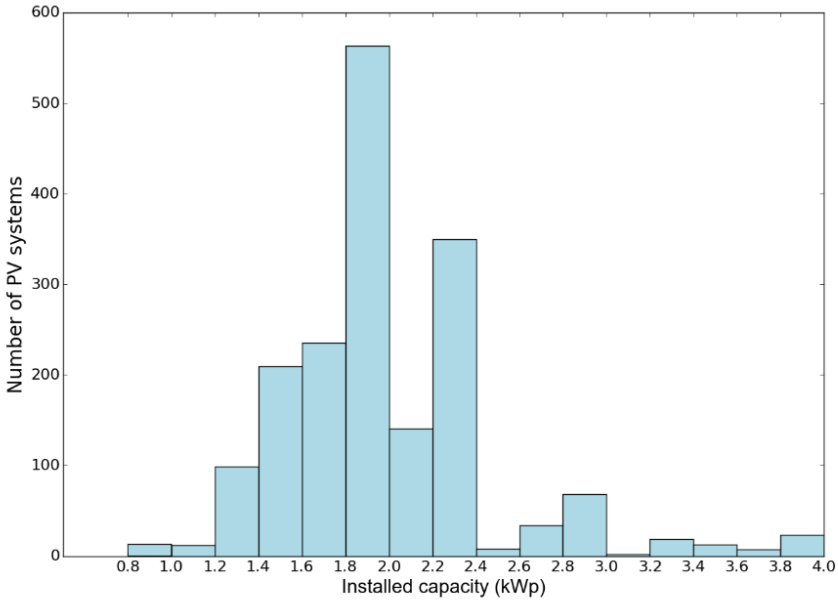


Figure 3.3. Histogram of installed capacity (in kWp) for 1788 PV systems at Nottingham, UK.

Table 3.1. Table of meta-data for the Nottingham City Homes (NCH) database.

Parameter	Comments
Home ID	Unique identifier per home
Address	-
Postcode	-
Anticipated Annual Generation (kWh)	Calculated based on past solar radiation data
Peak Rate (kWp)	Nominal capacity of the PV system
Monitoring Type	Refers to the way data are transmitted (GSM* is the only case)
Panel Configuration	Refers to different layouts in terms of the type of string connection implemented but the number of strings is not specified explicitly
Panel Manufacturer	-
Panel Model	-
Inverter Manufacturer	-
Inverter Model	-
Elevation (m)	As in distance from the ground
Bearing (degrees)	Usually taking North as a 0 degrees bearing
Inclination (degrees)	-
First Meter Reading Date	First day of operation
Monitoring Status	Active or in the process of activation

* **GSM = global system for mobile communication, for example via SMS (short message service)**

3.3 Data from the UK Met Office Integrated Data Archive System (MIDAS)

An essential step of the analysis, as shown in Figure 3.2, is to obtain solar radiation data for each PV system's location. The solar radiation data used in this study were acquired from monitoring stations operated from the UK Met office through MIDAS (Met Office Integrated Data Archive System) [149]. Hourly data of global horizontal irradiation and ambient temperature are downloaded from MIDAS and stored in a different database than the one

dedicated to the CREST outdoor monitoring system (COMS) but on the same server. The total number of the UK stations employed in the UK over 10 years of operation, is 123. Out of this number, about 75% are installed in 2005, 10% in 2015 and the rest are installed in the years between. Some stations were seemingly operational for a few years but with very low data availability. Other stations seemed to have only a few days or months of operational lifetime and these were not taken into account. The lowest operational lifetime allowed in this study is one year, namely 365 days. Finally, the total number of stations is shown in Figure 3.4 for over the years 2005 to 2015. An average number of 88 meteorological stations across the UK were utilised for the years 2012 - 2015.

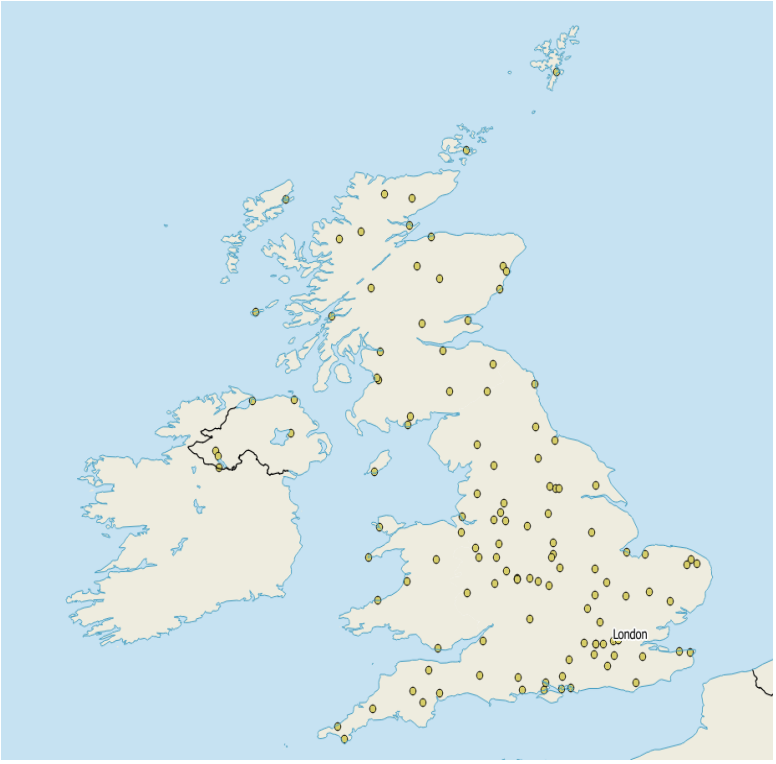


Figure 3.4. Map of the UK stations over the 11 years of operation (2005 – 2015) (QGIS image).

3.3.1 Quality controls at each met station and data averaging

The UK met office apply their own quality checks on the data prior to releasing them. These quality controls essentially include “flagging” entries with specific indicators and are applied on data during various stages of data transfer from point of observation to the final database. Values are checked based on climatological extremes, which also vary with location, and also

based on previous observations from the same source. For air temperature, checks are also applied based on comparisons against neighbouring stations.

For solar radiation, these quality checks are further complemented by modelling clear sky (using Ineichen and Perez model [150]) radiation values at every hour and every location, namely latitude and longitude of the station according to the Helioclim algorithm described in [7] such that:

$$GHI < 1.1 \cdot G_{CS} \tag{3.1}$$

Where,

G_{CS} = clear sky irradiation

GHI = global horizontal irradiation

Generally, effects such as irradiance enhancement due to cloud reflection and snow may lead to slightly increased limits in equation (3.1), however hourly data present much lower variations than instantaneous data and thus the applied limits are found sufficient [151].

For the calculation of PV output irradiance values lower than 50 W/m^2 are not taken into account. By not applying a single value restriction in this case, it is also possible to check for any timestamp mismatches between the met station irradiation data and the temporal reference system used for the analysis, which in this case is universal coordinate time (UTC).

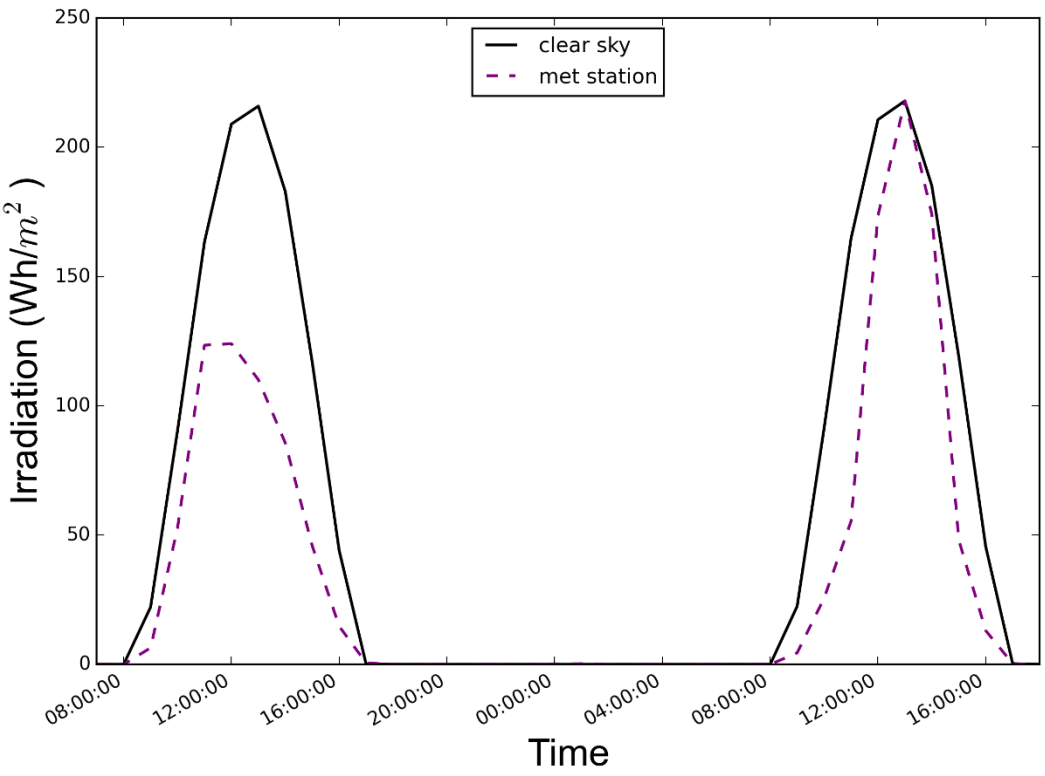


Figure 3.5. Hourly irradiation for clear sky modelled output and Loughborough met station for two days in January 2015.

Additionally, the Met Office supply their data as sums of 60-minutely readings for each hour which are stored at the end of observation time. For example, this essentially means that the average of the values between 12:00 and 13:00 would be stored in the “13:00” bin. Therefore, to be able to compare these aggregated results with clear sky modelled results the same procedure is applied for the model. Namely, minutely instantaneous values are modelled and then averaged over each hour of the year (see Figure 3.5).

Additional quality controls are applied for each station which ensure that no duplicate and abnormal entries are detected. Duplicate entries can be readily detected by applying restrictions on unique identifiers per entry. This can be achieved by simple database commands and no additional effort is required for such occurrences.

3.3.2 Spatial interpolation with kriging

Both global horizontal irradiance and ambient temperature are acquired at the measurement location i.e. the meteorological stations. However, both variables vary in time and space. In many practical applications such as in the case of domestic PV systems with no climatic monitoring, measured data are not available at the location of interest. Thus, regional interpolation techniques are employed as means to transfer data from the measurement sites to the estimation point. Available information is transferred from a number of adjacent measurement sites to the estimation site (GHI_E) through a function that represents the spatial weights according to the distances between *n* number of sites [152]:

$$GHI_E = \sum_{n=1}^{i} w_i G_i \tag{3.2}$$

Where,

w_i = weighting factor at each measurement site

G_i = irradiance at each measurement site

These weighting factors depend on the distance r_i between the measurement and the estimation site(s). For the inference of global horizontal irradiance at multiple sites, including Loughborough and Nottingham, the methodology described in [153] was applied using ordinary kriging interpolation. Different interpolation methods may give better results for different variables, station densities and climate regimes but kriging has proven to give the overall best results for the interpolation of various climate variables [107].

The estimation point is not a single site but rather a square grid of mapping points, which have a 2.5 km distance from one another, based on the applied resolution in this work. Thus, for a single geographical location, for example Loughborough, the irradiance data used are those estimated at the *nearest grid point* from Loughborough, as it is graphically described in Figure 3.6. Specifically for Loughborough at the monitoring site (latitude = 52.7, longitude = -1.2), the distance from the nearest interpolation grid point was 210 m. As implied in Equation (3.2) the higher density of meteorological stations close to the interpolation point, the more accurate kriging will be. For example, kriging for Loughborough is expected to yield better

results compared to a location in North Wales, where the meteorological stations are more scarce.

This method is very fast, as various locations can be simultaneously inferred by using the same interpolation grid and only choosing a different grid point. To establish the weighting factors (w_i) for kriging, an exponential semi-variogram (SV) is used [153]. An SV models a graph which shows the variance in measure with distance from all sampling locations and it is a significant prerequisite for the kriging process.

In summary, the following have been applied for the inference of climate variables in the cases of Loughborough and Nottingham [107].

a) Grid cell size was chosen at 2.5 km. Higher resolution would lead to slightly more accurate results but would significantly increase the computational time.
b) It was found that when using a number of stations below 30, the results became less accurate (which also depends on the complexity of the weather patterns at that particular hour). The average number of stations used for the years of study (mainly 2014 and 2015) was 88.

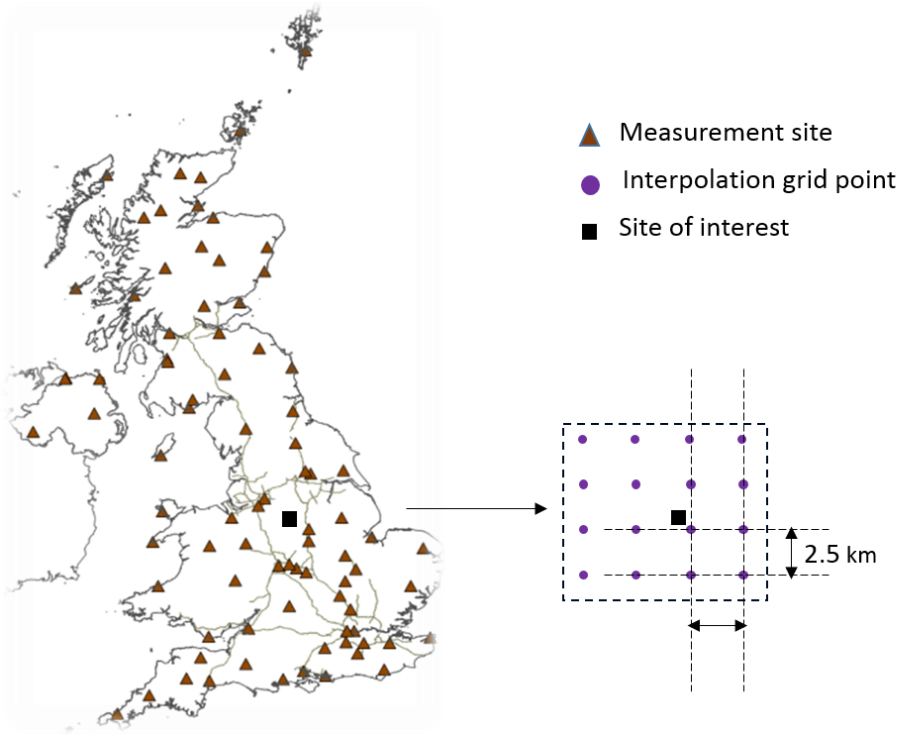


Figure 3.6. Simplified schematics of the spatial interpolation process from measurement sites to the 2.5 km square grid.

3.3.3 Modelling of in-plane irradiance

Once global horizontal irradiance is estimated for the location of interest, this has to be further separated into its beam and diffuse components and translated onto the plane of array. In this study the applied models and the related references are presented in Table 3.2, based on previous works which prove that the specific models give the most accurate results for Loughborough (and similar locations). The process can be summarised as follows:

a) Climate data (global horizontal irradiation and ambient temperature) are collected from meteorological stations across the UK at hourly time steps.
b) Both climatic variables are estimated at the location of interest using an interpolation grid derived through kriging technique.
c) Global horizontal irradiation is then separated into its beam and diffuse components.
d) Beam and diffuse components are translated onto the plane of array and added to calculate global in-plane irradiance.

Table 3.2. Employed models for the translation of inferred global horizontal radiation to plane of array.

Modelling stage	Validation studies	Original study
Separation Beam and diffuse irradiance components	[153],[154] (for Loughborough)	Boland-Ridley Lauret model (BRL) [119]
Beam and diffuse translation to plane of array (POA)	[155] (Loughborough), [113]	Hay, Perez and McKay [156] with Reindl correction [121]
Solar position algorithm	-	Reda and Andreas [120],[157]

3.4 Statistical procedure based on PV performance indicators

The applied quality checks are mainly based on two parameters, namely a) annual final yield (in kWh per kW_P) and b) annual performance ratio. The annual energy output is calculated by using the following formula [3]:

$$E_{AC} = t_r \sum_{i=1}^{n} P_{AC_i} \tag{3.3}$$

Where,

t_r = recording interval in units of hours

P_{AC_i} = power measured at each reporting period *i* (in kW)

Using two parameters, instead of one, while also looking at the correlation with each other is useful in differentiating potential data quality faults. The expected trends in both these parameters are described next as this is crucial for choosing the applied thresholds.

3.4.1 Performance ratio and specific yield

The limits for PV system PRs theoretically range from 0 to 1. Because PR is not temperature corrected, daily or hourly PR may sometimes exceed 1.0 for single (fault-free) modules [158] at optimum conditions of high irradiation, low temperature and high wind speed. These values are not typically the case for PV arrays, where greater operational losses occur and ambient conditions are hardly ever ideal. Even if temperature is taken into account, this effect would be more evident in monthly PR variations but not on an annual basis [159]. Recent studies have shown that expected system PRs are typically about 0.85 whilst higher figures (around 0.90) can generally be observed for a small sample of systems [160], typically large field based installations. Where PR is over 0.95 however, this could indicate either underestimated nominal capacity (where P_{STC} is significant higher than declared) or underestimated in-plane irradiation (an expected trend in cases where irradiance is remotely inferred). Even in this case though, it has been shown that for a moderate climate as in the UK), the module PR for c-Si is generally not higher than 0.93 [161], and this is further reduced for a PV system due to the additional losses (cabling, mismatch, inverter etc.).

Final yield in kWh/kW$_P$ is a means of comparing the energy yield potential of systems of different sizes, and typically this is calculated per annum for optimal installation angles. Countries with medium solar resource report an average annual kWh/kW$_P$ of 700-900 kWh/kW$_P$ [93][99]. A map of kWh/kW$_P$ potential for Europe and several other countries is given in [162], which for the Southern UK gives typical values ranging from 800 to 900 kWh/kW$_P$.

In the case of residential PV systems, such as those in this study, installations are distributed over different inclination and azimuth angles. For this reason, annual kWh/kW$_P$ will be distributed over a range of values (as opposed to PR). In order to neglect this diversity, instead of the absolute kWh/kW$_P$, the "normalised" kWh/kW$_P$ or performance index [92] is compared with an ideal 1kW$_P$ system which is modelled at different inclination and azimuth angles according to [162]:

$$E_{theor}/P_{STC} = \mathrm{PR}_{theor} \cdot \frac{H}{1kW/m^2} \tag{3.4}$$

Where PR_{theor} is taken as 0.85 [160] and *H* the modelled irradiation at different inclination and azimuth angles. Then combining (2.35) and (3.4):

$$\frac{E_{actual}/P_{STC}}{E_{theor}/P_{STC}} = \frac{1}{PR_{theor}} \cdot PR \tag{3.5}$$

The first term of (3.5) is defined here, as the performance index (PI) and it is used as an additional metric to detect those data outliers which cause an increased bias in either PR or final yield distributions:

$$PI = \frac{Actual\ yield\ (in\ kWh/kW_P)}{Theoretical\ \ yield\ (in\ kWh/kW_P)} \tag{3.6}$$

3.4.2 Median absolute deviation (MAD) analysis

In order to classify systems by their (system or data) quality, the Median Absolute Deviation (MAD) is used, given by:

$$\mathrm{MAD} = \mathrm{median}(|X_i - \mathrm{median}(\mathrm{X})|) \tag{3.7}$$

Where X is the applied distribution and X_i is a point in X. MAD was chosen as a more robust metric than the widely employed standard deviation, as it is also appropriate for non-normal distributions and less sensitive to outliers [163][164].

In this study, MAD is used to classify systems into categories of data and system quality respectively. The thresholds are given by the following expression [164]:

$$\mathrm{MED} - k \cdot \mathrm{MAD} < X_i < \mathrm{MED} + k \cdot \mathrm{MAD} \tag{3.8}$$

Where $MED = \mathrm{median}(\mathrm{X})$ and k takes different values according to the applied criteria.

The annual PR distribution is seen in Figure 3.7 for 2014. Due to the underestimation of in-plane irradiation, this is shifted towards higher PR values, but using the MAD limits the extreme cases can still be detected based on the applied thresholds. For PR distribution it is k = 3, for the lower region ($PR < MED_{PR}$) and k = 2 for the upper region ($PR > MED_{PR}$) [164]. For the upper region a stricter limit is applied as it is easier to distinguish wrongly declared capacity. For the distribution shown in Figure 3.7, these values corresponds to lower and upper limits of 0.68 and 1.05 respectively. Systems with very low PRs (lower than 0.68 in this case) also need to be explored further for system or data quality issues, such as over-estimated nominal capacity, missing data and zero generation due to reasons other than faults in the system, such as for example the system has been turned off by the owner or for maintenance.

Generally, the PR distribution shown in Figure 3.7 indicates that about 24% of the systems require further investigation. In this initial analysis 1% of the systems show abnormally high output, 5% show very low generation and 20% in total, show annual PR well below 0.68. Missing data will be treated separately in the following sections. The median of the PR distribution is $MED_{PR} = 0.83$.

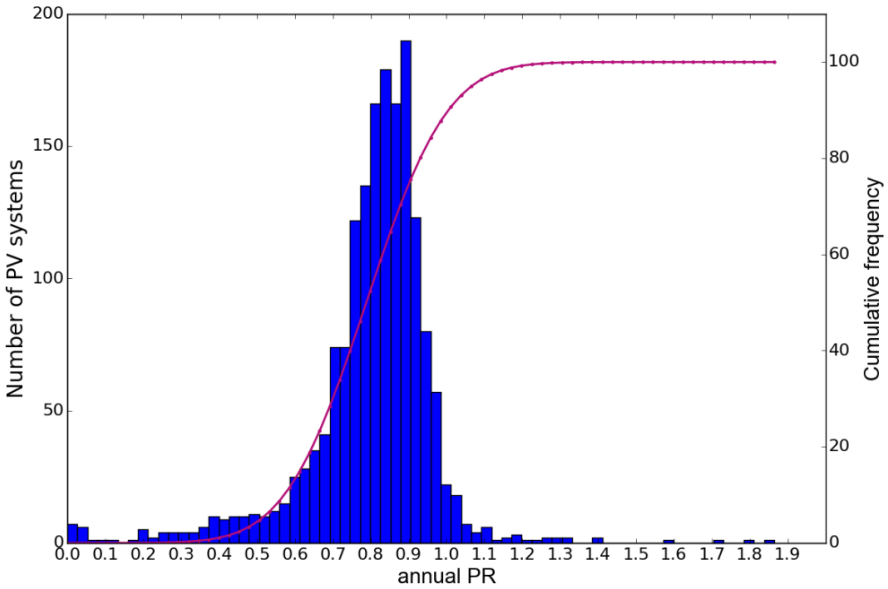


Figure 3.7. Annual performance ratio (2014) histogram using the initial dataset. The red line indicates the cumulative frequency of the PV systems.

Annual kWh/kW$_P$ distribution is shown for the same year (2014) in Figure 3.7 where the median of the kWh/kW$_P$ is MED$_{SY}$ = 855 kWh/kW$_P$. It is evident also here that about 2% of the PV systems have a suspiciously high PR and specific yield.

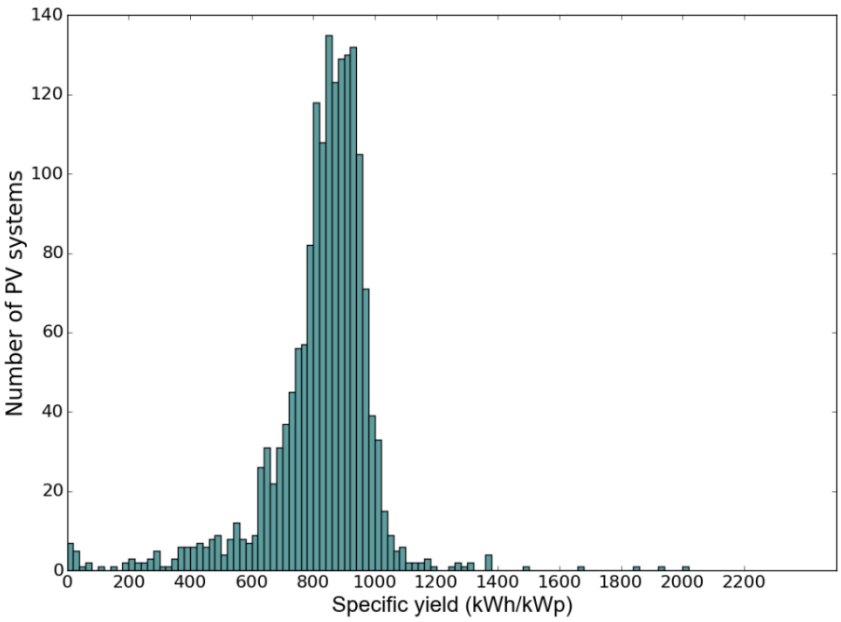


Figure 3.8. Annual kWh/kWp histogram (2014) using the initial dataset.

From (3.5) it is apparent that PR and PI have a positive correlation. By applying a linear regression between PR and PI it is possible to detect large data outliers (low correlation data points). The PR-PI linearity can be graphically realised by the coefficient of determination (R^2) as illustrated in Figure 3.9

Figure 3.9. Performance ratio versus performance index. Three cases of systems are highlighted here: low correlation, very low and very high PR and increased zero generation (prior to irradiance correction)

, where three different cases are highlighted; namely very high performance ratio, systems with increased zero generation and low correlation. The detection of the low correlation data points is then realised by applying the Student's T-test based on a 0.99 confidence level [46] on the residuals of the linear fitting. This additional index assisted in quickly identifying those cases where reference azimuth was wrongly assigned, for example by following the USA annotation which assumes south as -180 degrees instead of 0 degrees, conventionally used.

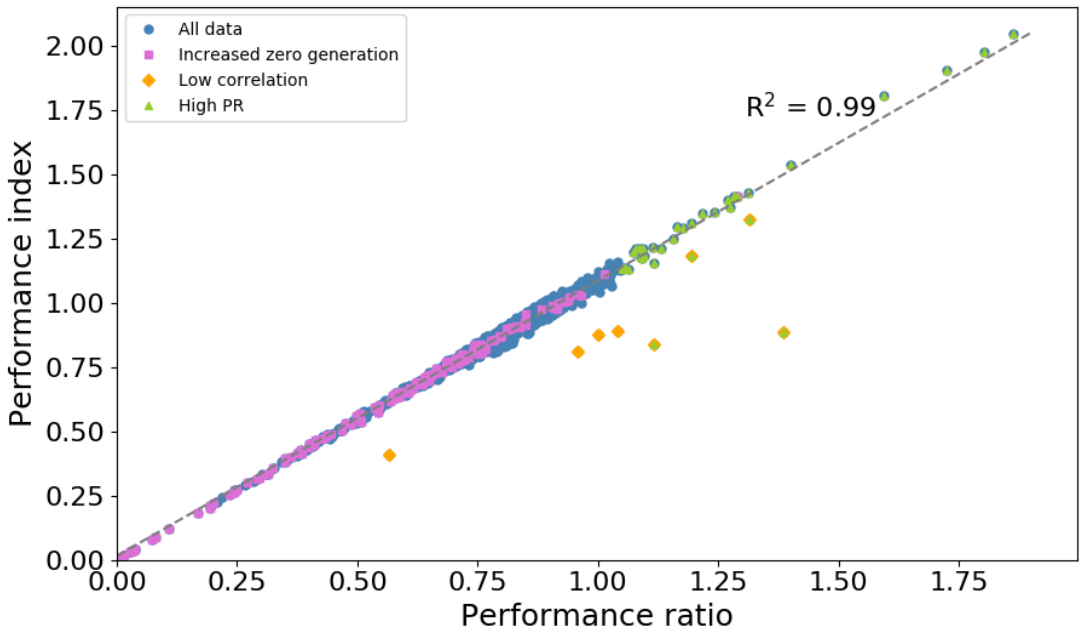


Figure 3.9. Performance ratio versus performance index. Three cases of systems are highlighted here: low correlation, very low and very high PR and increased zero generation (prior to irradiance correction)

Different correlation cases between PR and PI are differentiated as follows:

- Lower PR - high PI: This is likely to be due to overestimated irradiation at the specific system location. Such cases are unlikely to occur as irradiation is usually underestimated rather than overestimated (as discussed in 4.3.3).
- Lower PI – high PR: This is likely to be caused by underestimated irradiation at the specific system location. There are two reasons for this; either calculated irradiation is significantly lower than in reality or the system is installed at a different azimuth than declared.

- High PR – high PI: This could be due to underestimated nominal capacity.

- Low PR – low PI: This is a more complicated situation, as several factors may contribute to lower indicators. These include erroneous input information (such as overestimated nominal capacity), or low performance due to faults and/or shading. In such cases, addressing data quality issues is critical, as identifying missing data.

Finally, the combination of different flags and their priority is taken into account, for example if a system has increased missing data, it is expected to also have a lower annual PR which however does not imply the existence of a system fault, thus this case should be eliminated first. The priority in the checks goes as follows: increased missing data > low PR-PI correlation > abnormal PR > abnormal PI. This step is critical to eliminate cases where PR and PI are affected by data quality. A summary of these identifiers is given in Table 3.3 applied on the particular dataset. In the case where a system is found with (very) low PR and PI, then this system is automatically considered as a possible faulty system. Those cases of PV systems are further discussed in Chapter 5 where a (nearly) real time detection procedure is presented.

Table 3.3. Summary of identifiers based on annual records

Parameter	***Identifier***	***Description***	***Action***
PR, PI	0	Normal	No action
	1	Lower performance	Yellow alarm – check other identifiers
	2	Very low PR	Red alarm – check system - verify system description
	3	Too high PR- possibly wrong system description	Verify system description
Missing data/zero generation days	0	Normal	No action
	1	15 < days <30	Yellow alarm
	2	days >30	Red alarm – check system
PR-PI correlation	0	Normal	No action
	2	Wrong azimuth	Red alarm – check/correct angle

3.5 Identified data quality issues

3.5.1 Ambiguous models description

Initially, a number of inconsistencies was found in terms of single module capacity and the number of the installed modules for specific systems, a shortcoming which is surprising given that this type of information is normally quite straightforward to record. Moreover, both panel and inverter manufacturers' and models' names were very often mentioned with different spelling whilst implying the same model or manufacturer. To be able to sort the PV systems into groups of same manufacturers and models, one unique spelling was chosen from each category and a specific ID was assigned to it using a simple string recognition algorithm.

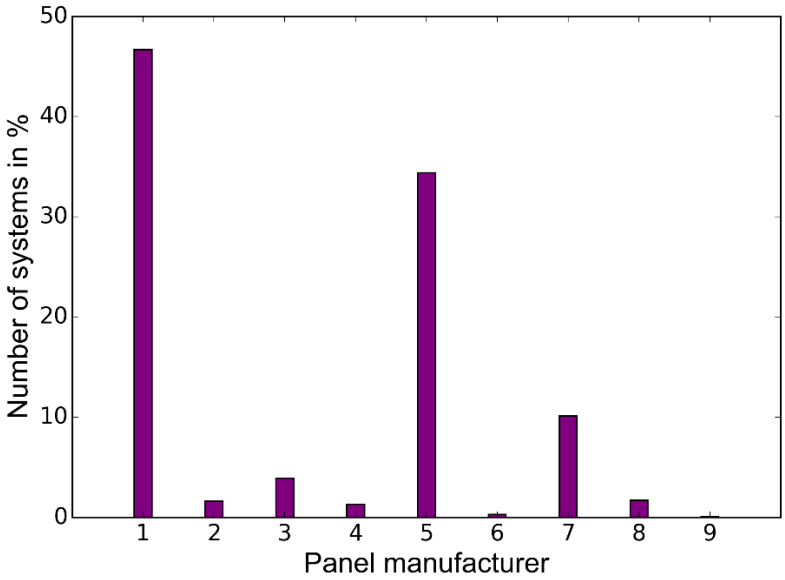


Figure 3.10. Percentage of systems per panel manufacturer. In total 9 different panel manufacturers were reported, with one of them comprising about 46% of the PV modules.

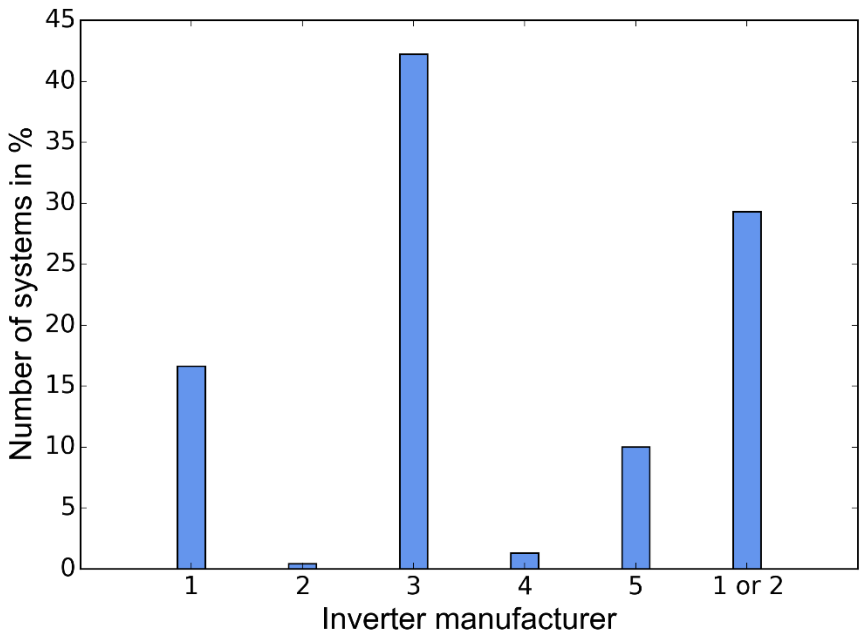


Figure 3.11. Percentage of systems per inverter manufacturer. For a large number of systems this information was an ambiguous entry such as “manufacturer 1 or 2”.

The applied re-classification procedure reduced the number of inverter and panel manufacturers and models significantly (see Figure 3.10 and Figure 3.11) to 5 and 9 respectively. Essentially, this issue indicates that hand-written entries should rather be replaced by an electronic drop-down list, as this information becomes particularly relevant when modelling the theoretical performance of each PV system.

3.5.2 Wrongly declared azimuth (azimuth) angles

This is the most frequent issue met in the dataset and as seen next, it may have a significant impact on the PR distribution. To demonstrate this case, a clear sky day was selected for each system based on daily average clearness index K_t where $K_t > 0.7$, where daily clearness index is calculated as [119]:

$$K_t = \frac{\sum_{i=1}^{24} GHI_i}{\sum_{i=1}^{24} ET_i} \tag{3.9}$$

Where,

GHI_i = global horizontal irradiation at hour *i*

ET_i = extraterrestial radiation at hour *i*

An extreme example of plane of array irradiance for system 'A' is shown in Figure 3.12. This PV system was noted as having an azimuth of 40 degrees (northeast) where the azimuth should be 220 (taking North as 0) degrees approximately. This case specifically demonstrates the mistake caused by differing azimuth conventions. Zero degrees as South are usually used by PV modelling software and in this case, whereas measuring instruments such as magnetic compasses and GPS devices take North as the zero datum.

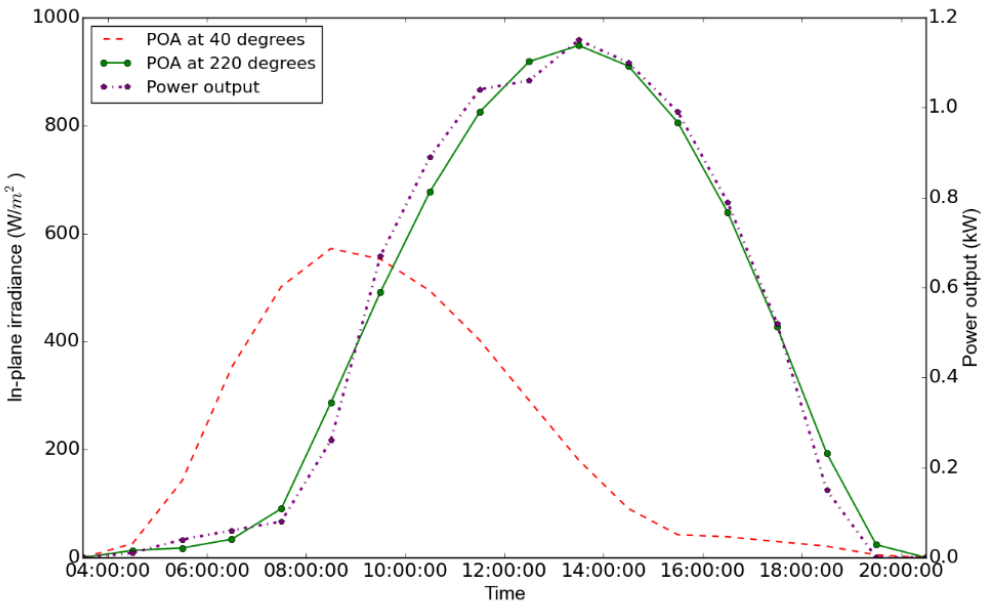


Figure 3.12. In-plane irradiance (POA) profiles at two azimuth angles and power output

This case can be easily distinguished from the PI-PR correlation. The effect of smaller deviations of up to 50 degrees on the PR which are frequently caused by human error (such as inaccurate compass readings or even lack of compass), is further visualised by simulating the annual in-plane irradiation for a PV system assuming different declared azimuths than its actual azimuth (θ_{actual}).

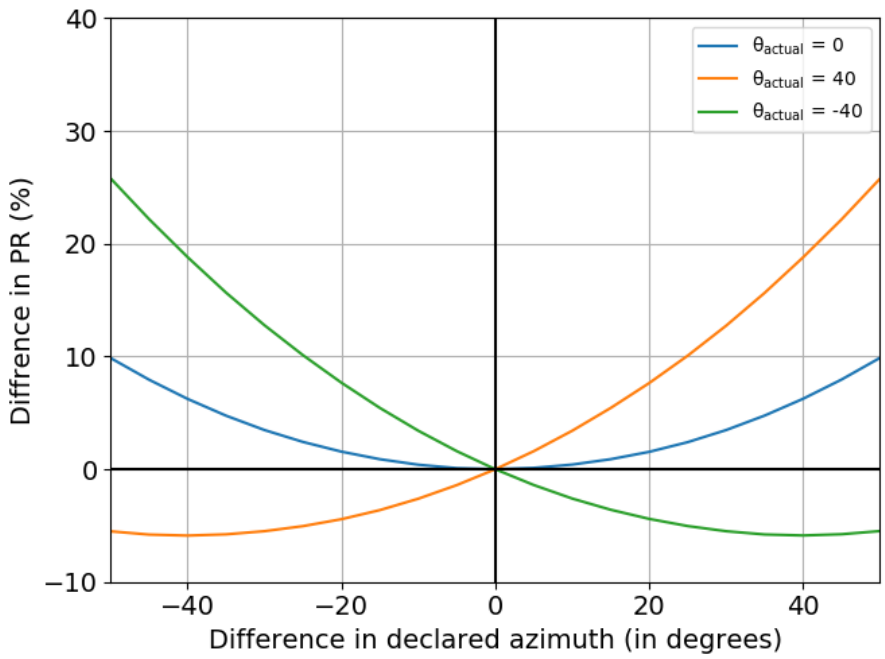


Figure 3.13. Impact of wrongly declared azimuths on the PR for different PV system azimuths (θ = 0, 40, -40).

As shown in Figure 3.13, when the difference between the declared and the actual azimuth is over ±20 degrees, the difference in estimated PR may be up to 9% depending on the actual azimuth of this system. This could potentially place that system in a different bin of the PR distribution towards higher or lower values. Eventually, if a large amount of wrongly declared systems exist in a dataset then the initial PR distribution is expected to change.

Azimuth has a higher impact (as seen in Figure 2.12) than inclination (see Figure 3.14) on modelling the PV system output. Small variations in inclination (up to 15 degrees) do not yield significant differences as opposed to azimuth (azimuth), when modelling the electrical output of a PV system. For example, in Figure 3.14 (a) and (b) a difference can be mainly seen in the afternoon for a system facing at 30 degrees east of south and in the morning for a system facing at 30 degrees west of south, respectively. Moreover, inclination angles are often binned within a certain range depending on the type of house and/or its location (for example 30 degrees is predominantly found in the dataset corresponding to a particular type of house) [165].

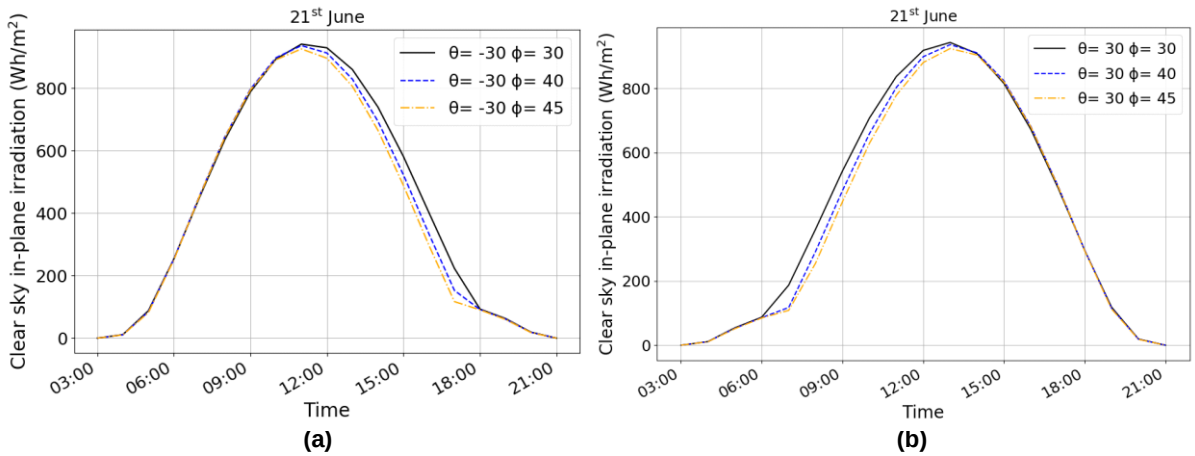


Figure 3.14. Modelled hourly clear sky in-plane irradiation for azimuth (a) θ =-30° and (b) θ =30° for three inclination(tilt) angles (φ=30°, 40°, 45°).

To correct for high azimuth deviations (over 15 degrees), a simple identification procedure is developed based on the systems' electrical output on a clear day and their location (latitude and longitude where available). It is expected that on a clear day the (ideal) electrical output of a system will follow the Gaussian curve and therefore adjusting a clear sky model to it by simply testing different azimuth angles, would indicate the azimuth which gives the best fit. The problem in most domestic systems however, is that the curve maximum is not always evident on power output, often due to partial shading or other faults, for example an undersized inverter. To overcome this obstacle, a Gaussian fitting is thereby used for the energy output of the PV system (see relative equations in the Appendix). An example of such system is given in Figure 3.15, where the energy output maximum is not evident.

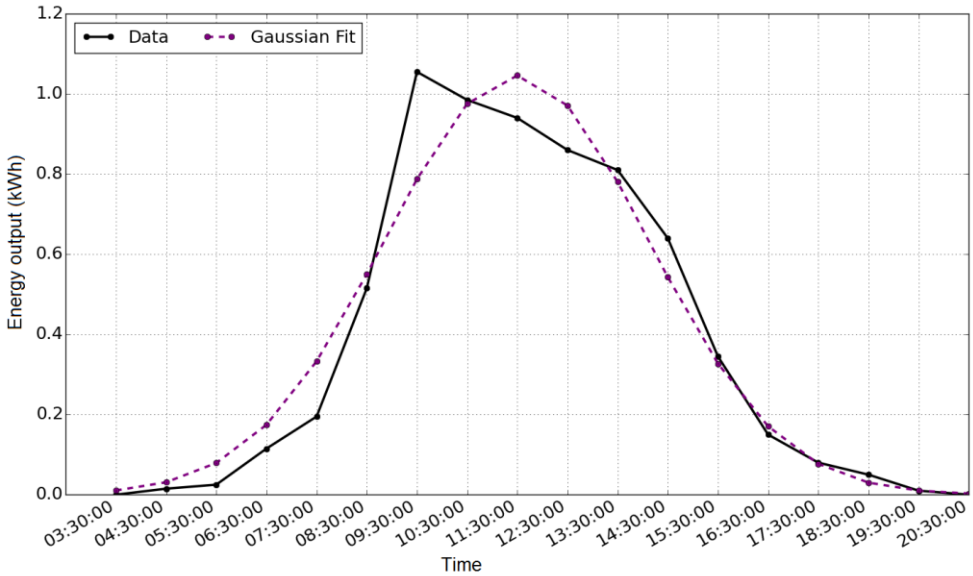


Figure 3.15. Energy output *vs* Gaussian fit for a PV system.

The tool is initially tested on the hourly energy output of a PV module taken from the CREST monitoring system, which is placed at 0 degrees due South, with an uncertainty of ± 0.59 degrees. By using the Gaussian fitting tool the azimuth of the rack of the PV module was found to be -2 degrees east of south (see Figure 3.16) which is a good agreement, considering that even up to 10 degrees the difference in hourly aggregated output is barely distinguishable.

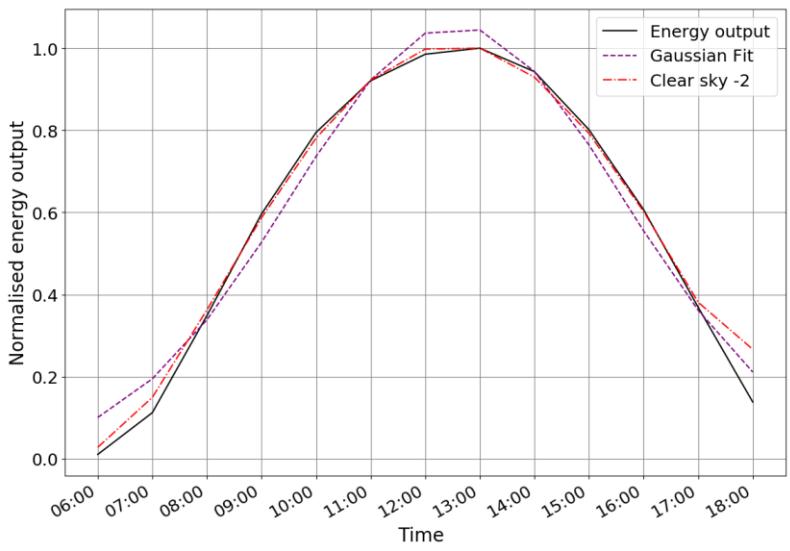
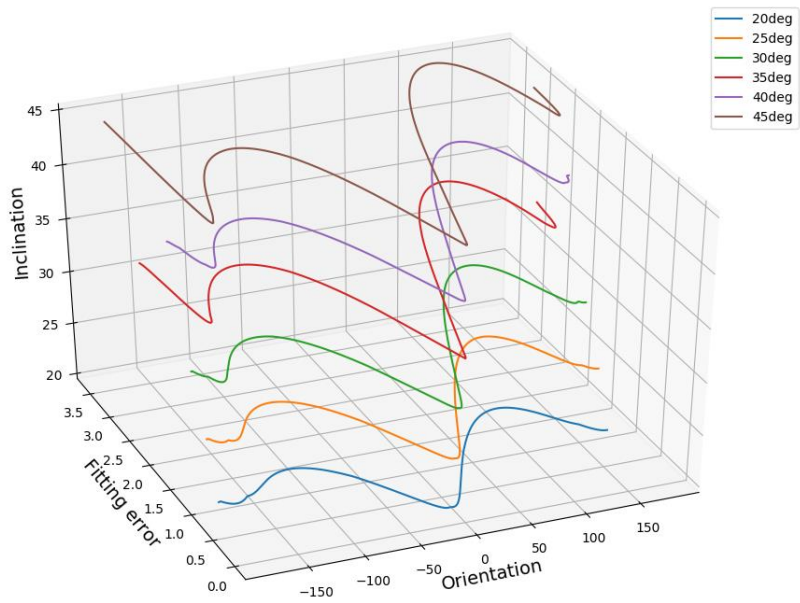
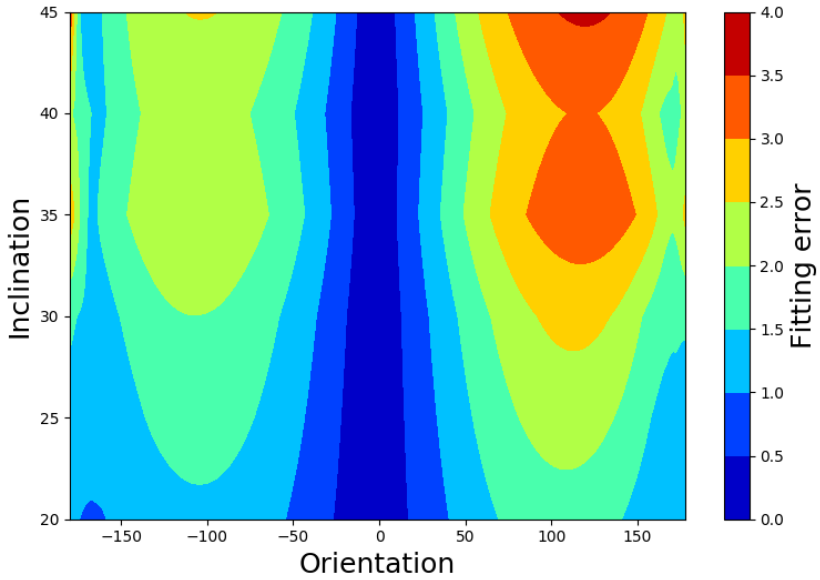


Figure 3.16. Hourly energy output, Gaussian fit and modelled clear sky irradiation for the optimum fitted azimuth (-2) corresponding to a PV module in CREST. Clear sky and energy output are normalised to their maximum values.

Repeating the same algorithm for different inclination (tilt) angles (20 to 45 degrees per steps of 5) yielded the results seen in Figure 3.17 (a) and (b), where the fitting error refers to the applied error criterion between the (fitted) Gaussian curve and the clear sky model at different azimuth (azimuth) angles. For the different inclination angles, the optimum azimuth angle was the same (-2 degrees) in all cases, though the fitting error increased (from 0.015 to 0.2) with deviation from the optimum tilt which was at 35 degrees, as expected.



(a)



(b)

Figure 3.17. Fitting error between the Gaussian and the clear sky model as a function of azimuth (azimuth) and inclination (tilt) angles in (a) 3D and (b) contour plot. The fitting error refers to the applied area criterion between the Gaussian and the clear sky model curves.

In order to find the azimuth of a wrongly declared PV system in the domestic dataset, clear sky in-plane irradiation was modelled for that system at a range of azimuth values (355

degrees at a step of 5) at three clear days, where the azimuth is chosen based on the minimum fitting error between these days. Again, both clear sky and energy output were compared on the same aggregation basis (see Appendix). An example is given in Figure 3.18 for different azimuth angles where 20 degrees west of south was found as the most possible azimuth for the particular system, which was declared at 45 degrees.

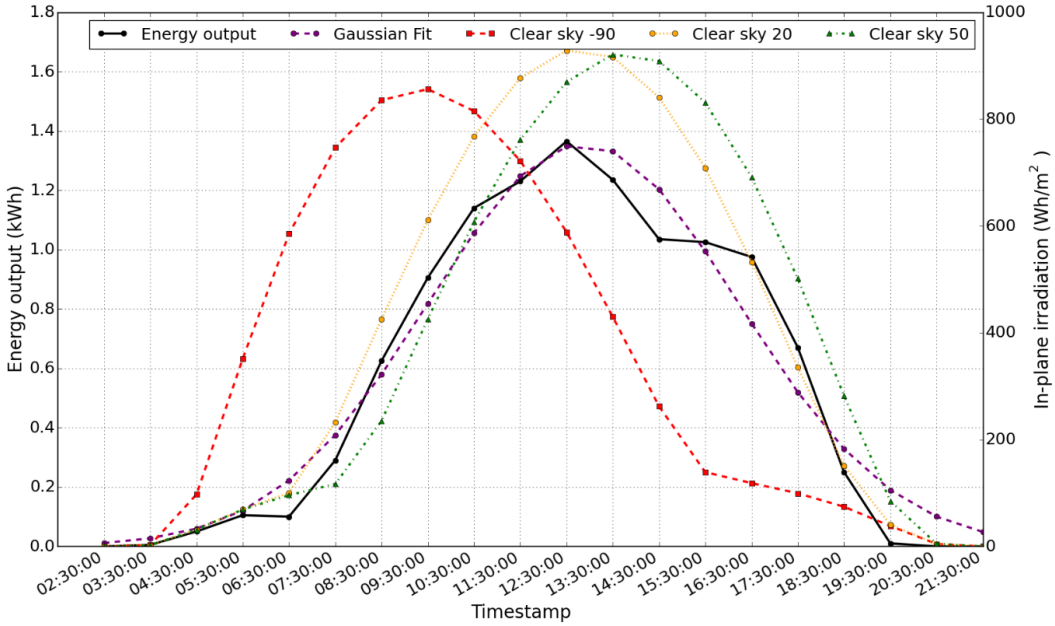


Figure 3.18. Energy output, Gaussian fit and clear sky in-plane irradiation (not normalised) for different azimuth angles (South = 0). Optimum fit was found for azimuth equal to 20 degrees west of south.

Although this is a useful tool to quickly identify the possible azimuth of a wrongly declared PV system, there is yet validation to be carried out with comparison to a robust mapping software where available. This is to eliminate cases where Gaussian fitting is not possible due to severe (partial or uniform) shading on the systems, where a large part of the energy output area is “missing”. A comparison with the LIDAR azimuth extraction method described in [166] was possible for a total of 287 PV systems, which were wrongly declared from as little as 5 up to 50 degrees. The results showed close agreement with a mean and maximum deviation between the methods of 8 and 16 degrees, respectively. Essentially, the 8 degrees difference corresponds to about half an hour where the dataset is at hourly intervals, thus it is an acceptable deviation. The frequency diagram of the difference between the declared and the extracted azimuth using the Gaussian fitting tool is shown in Figure 3.19.

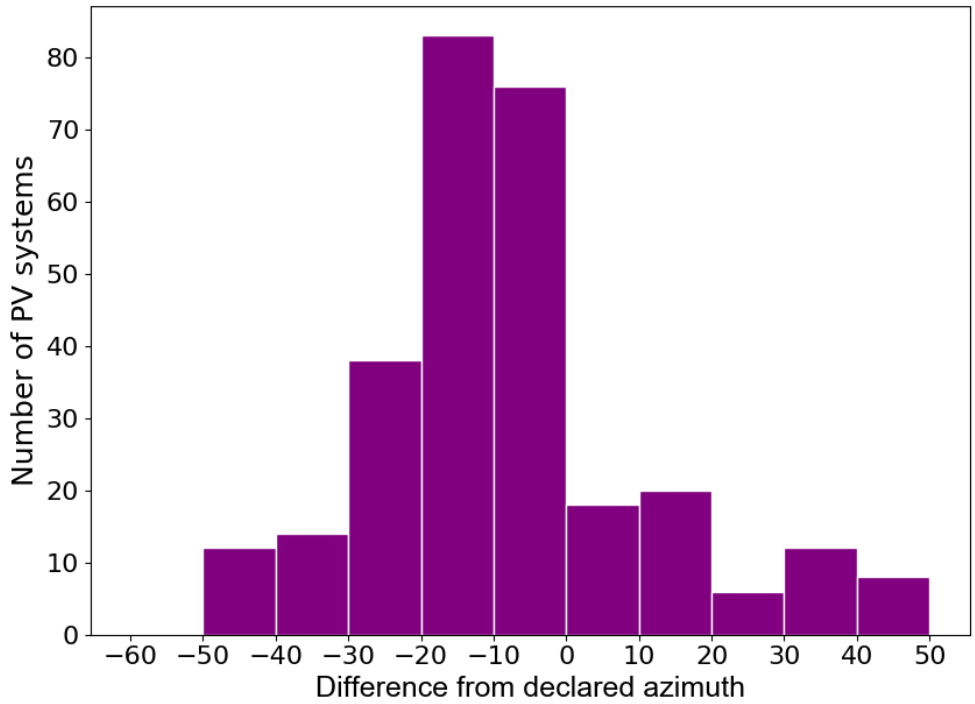


Figure 3.19. Histogram of the differences between declared and extracted azimuth using the Gaussian fitting tool for 287 PV systems.

A significant 41% of the PV systems (corresponding to 6.5% of the whole population) showed an absolute deviation from 20 to 50 degrees, which may cause a substantial difference in calculated PRs and up to 25% as shown in Figure 3.13. Considering that at least 6.5% of the PV systems in overall present this error, shows that the verification of azimuth declarations is of major importance in the final PR distribution.

3.5.3 Erroneous nominal capacities

Another case of wrong input information, such as erroneous capacity values, is often found. This can be easily distinguished based on the upper limits of both PI and PR. System 'B' is an extreme example of a system for which very high PR and specific yield is indicated (Figure 3.20). The nominal capacity of this system is about twice as high as the declared value compared with systems with the same module/inverter configuration found in the dataset with a nominal capacity of 2.9 kW_P. This value was found to give a better match with the

system's output comparing it to a modelled output on a clear day, confirming the inaccurate sizing. These two are the most extreme cases, however there are another 25 cases with underestimated nominal capacity which required verification with the owners/installers. These cases demonstrated constantly high yield and performance ratio since the first day of operation. It was further found that 69% of these cases, had the same module manufacturer, where individual module ratings were not available.

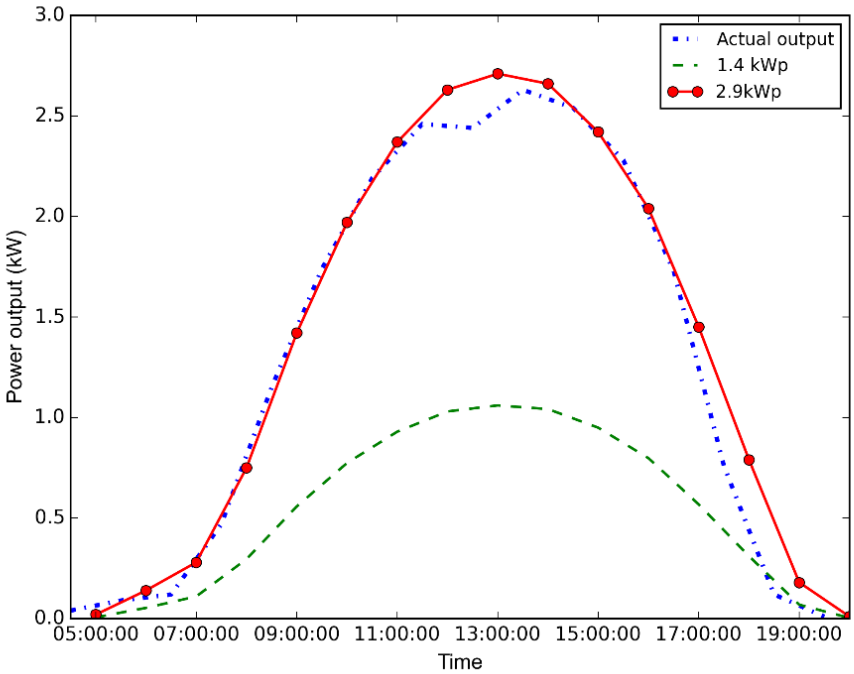


Figure 3.20. Actual and modelled output (considering about 10% system losses) for two cases of nominal capacity 1.4 and 2.9 kWp.

On the other side of the PR distribution (lower limits) 28 cases were found with very low output from the first year of operation. This could either indicate wrongly declared capacity or an installation fault such as system shading. Nominal capacities need to be verified as well as additional situations which will be explored for such cases, in the fault detection framework presented in Chapter 5. Due to potential faults and increased performance losses the corrections of nominal capacity are not as straightforward as wrongly declared azimuth thus, these need to be confirmed by the owner or the administrator of a system.

3.5.4 Missing data

Missing data are noted with a special character in the study to discriminate the particular issue from the non-generating systems. In terms of the actual records, the difference between zero and missing entries is that in the first case, the output of a system is noted as “0.0” throughout the day, whereas in the second case the output of a system is blank entries. Missing data affects the performance analysis of a system and may often lead to misleading results. Particularly, in the Nottingham dataset the missing data durations ranged from 5 up to 360 days and only for 2014, 60 PV systems had an average of 30% of the days missing. An overview of PV systems with missing data of more than 30% days and for the years 2012 to 2015 is given in Table 3.4.

Table 3.4. Number of systems with more than 30 days of missing data per year

Checked parameter	Year	Percentage of systems (%)
Missing data	2012	3.44
	2013	1.03
	2014	2.85
	2015	7.94

It is evident that a large number of PV systems suffer from monitoring gaps some of which extend to months of missing records. Possible reasons for this issue are communication failures or systems switched off, though these often remain unknown. The majority of these systems were also found in the lower regions of PR of less than 0.68 due to this particular issue. If missing data are treated the same way as zero generation then the effect of this on both the PI and PR distribution may not be evident for up to 10% missing days, if these are randomly distributed within a year, as graphically presented in Figure 3.21.

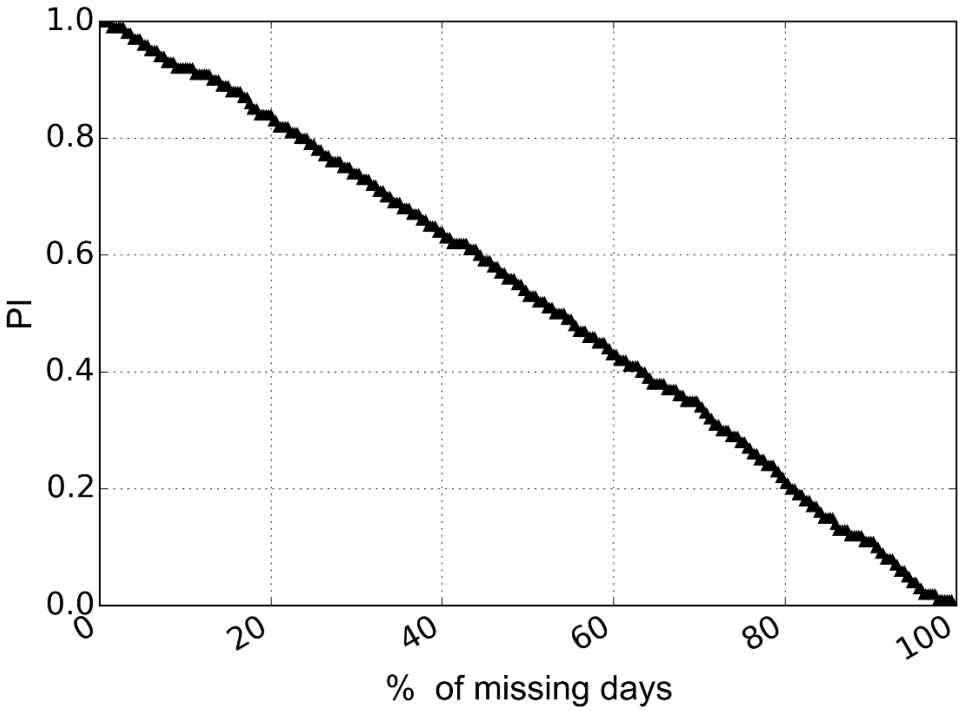


Figure 3.21. Impact of missing data on the performance index of a PV system.

However, their impact becomes more significant for missing days over 15%, as the system output starts shifting to the lower range of the PI. The same holds true for monthly analyses, i.e. ideally no more than 3-5 days should be missing. Additionally, when completely omitting missing data, the calculation of the PR is biased by the existing data points. This may lead to significant overestimation or underestimation of PR depending on the position of missing data within a year, as demonstrated in Figure 3.22 for a real PV system example. So, when high PR days are missing then annual PR is underestimated, conversely when low PR days are missing then annual PR is overestimated, due to PR seasonal variation.

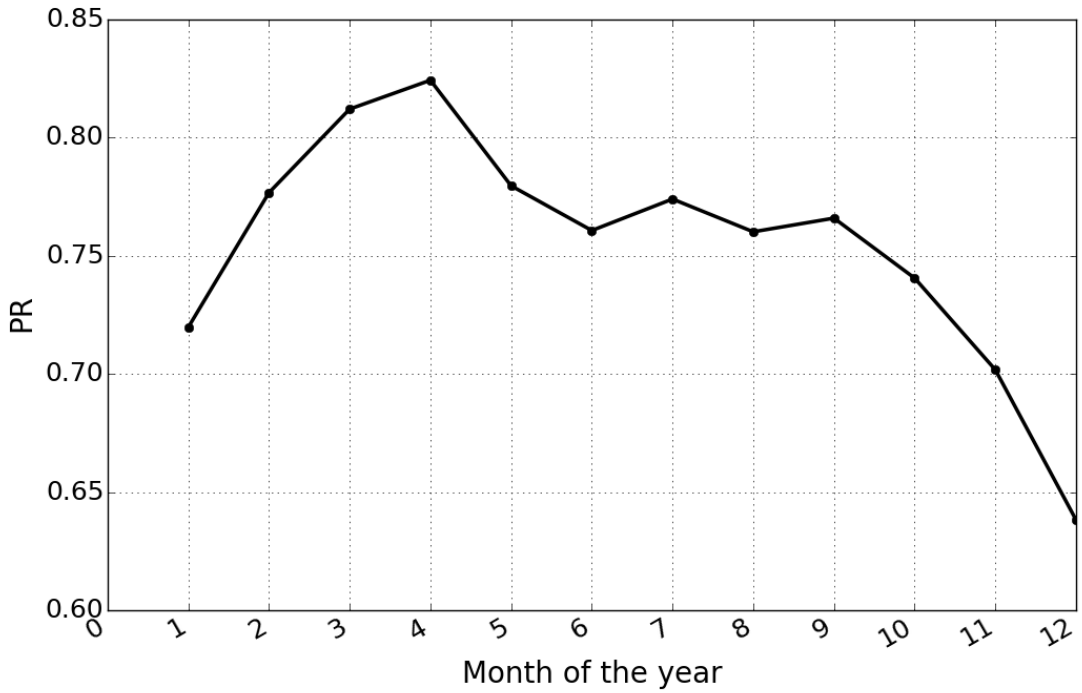


Figure 3.22. Monthly performance ratio (PR) variations for a PV system (UK field trials) with polycrystalline silicon modules.

The following chapter is dedicated to techniques proposed for filling these gaps with modelled energy output and re-calculating PR, whilst also considering the merits and limitations of *back-filling*.

3.6 Conclusions

Domestic PV systems comprise a significant percentage of the overall photovoltaic sector in the UK as well as worldwide and thus attention was specifically paid on the analysis of domestic monitoring data including the necessary quality assessment followed on the solar radiation datasets. In the analysis of 1788 PV systems in Nottingham, unexpected artefacts in data quality such as ambiguous system information, erroneous nominal capacities and installation angles and missing data were found. A significant percentage of the available information needs to be filtered, as otherwise it contributes towards odd results as seen in the initial PR distribution in Figure 3.7. This becomes even more important considering that a number of studies are focused on fleet PV system performance assessments and often the supplied data for these studies are based on monthly outputs and information given by the

owners which may or may not be verified. In the majority of these studies there is no indication of the percentage of possible missing data nor any validation with regards to the technical description of each system, even though both situations seem to affect the calculation of the main performance indicators. Even though annual indicators can be used to detect cases where there was a high deviation between declared and actual system information, a flagging system based on hourly profiles must be applied for the detection and correction of more obscure data quality issues, as further seen in Chapter 5.

Finally, missing data is identified as a common occurrence throughout system's operation and missing data of over 30% of operational days within a year, which shifts the performance indicators to lower values, without any system faults occurring. The importance of rectifying the impact of missing data on performance assessments as well as novel means to achieve this are discussed in Chapter 4.

Chapter 4

Inference of missing data in photovoltaic monitoring

4.1 Introduction

As seen in Chapter 3 missing data in monitoring is a very common occurrence. Interrupted monitoring data with prolonged gaps results in biased performance ratio (PR) results if these are ignored as seen in Chapter 2, and therefore can have a significant effect on the conclusions that are drawn from the data. This may be caused due to malfunctions such as power outages, communication failures or component faults. Although, in other fields, such as in signal processing and data science, dealing with missing data is a very popular area, to date there are no official strategies on inferring lost readings of energy output in PV monitoring. This will affect, as shown here, the PR calculated for the system and may hide incidents that would otherwise trigger warranty cases. Any attempts to back-fill data using previous dates or days from previous years are at best temporary with very high uncertainty attached to these methods. Utilising average values from dates close to the missing period may give an estimation of PR, but such methods become less relevant when missing periods are extended to several weeks. Therefore, an issue remains of how to recover lost data, and to arrive at a valid monthly and annual performance ratio.

In this chapter, different cases of data loss are considered, as each case can be often associated to specific PV system configurations and size. Most large PV systems operate independent meteorological and electrical monitoring systems. In the case of domestic PV systems only electrical monitoring is available. Thus, three different cases of data loss are considered that cover the aforementioned layouts. In the first and second cases, string monitoring may be considered as a common monitoring practice in larger PV systems. In the first case, string data are missing but weather data are still available. In the second case, both string electrical data and weather data are lost. This case could be a smaller PV system where there is only one weather monitoring station. Such cases have been reported for several PV systems for example in the UK field trials where both weather and energy output readings

may be lost [2]. Finally, in the third case a domestic monitoring system is considered, where there is no weather monitoring and electrical data are lost.

The approaches for inferring missing data are based on statistics as well as knowledge of the particular systems and weather profiles, since input weather data is a prerequisite for estimating energy output. Different cases of data loss are connected to certain types of monitoring such as in utility scale or small-scale residential PV systems. For each case, the results vary in accuracy which is assessed by comparing measured to modelled energy output data from the CREST outdoor monitoring system, the UK field trials [2] and several cases from the Nottingham City Homes dataset. Finally, the benefits from back-filling are presented in terms of reducing the bias in calculated monthly and annual performance ratio.

4.2 Statistical metrics

The prediction error of the proposed back-filling methods applied at each case is assessed by using the following statistical metrics, comparing measured with predicted (back-filled) values:

A) Root mean square error (RMSE)

$$\mathrm{RMSE} = \left(\frac{1}{\mathrm{N}}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}}(P_i - M_i)^2\right)\right)^{1/2} \tag{4.1}$$

$$\mathrm{rRMSE} = \frac{\left(\frac{1}{\mathrm{N}}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}}(P_i - M_i)^2\right)\right)^{1/2}}{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} M_i} 100\% \tag{4.2}$$

B) Mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}|P_i - M_i| \tag{4.3}$$

$$\mathrm{rMAE} = \frac{\frac{1}{\mathrm{N}}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} |P_i - M_i|\right)}{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} M_i} 100\% \tag{4.4}$$

C) Mean bias error (MBE)

$$\mathrm{MBE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (P_i - M_i) \tag{4.5}$$

$$\mathrm{rMBE} = \frac{\frac{1}{\mathrm{N}}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} (P_i - M_i)\right)}{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} M_i} 100\% \tag{4.6}$$

Where,

P_i = predicted quantity

M_i = measured quantity

N = number of predictions

The RMSE describes the scatter of the predicted data and the differences between measured and predicted values are added up by the second power, thus high deviations from measured values have a strong influence. MAE describes the absolute error, as overall difference between predicted and measured quantities and MBE indicates whether the model overestimates or underestimates the measurement value, which is also expressed as the “systematic” error of the distribution. MBEs close to zero signify an (almost) unbiased distribution. When comparing single quantities, as in the case of monthly performance, then the absolute values of all these metrics yield the same result, since N=1. Then MBE or more often rMBE, expresses the deviation of the predicted from the measured value.

Both absolute and percentage errors are given for the test year, in order to provide a better understanding of the magnitude of the error. This is because in some cases, percentage errors tend to increase significantly at lower values of the tested parameter but their effect is relatively small as the absolute difference is also small.

4.3 Error analysis in solar radiation and temperature data

4.3.1 Data from CREST outdoor monitoring system (COMS)

For the validation of the applied models and error analysis electrical and meteorological data from CREST monitoring system were used. Maximum power output and module temperature are the utilised electrical data which are stored in 1-minute resolution. Meteorological data of global horizontal, in-plane irradiance and ambient temperature are stored in 1-second resolution. These two discrete datasets are averaged in hourly time steps, and joined into one dataframe which contains the five aforementioned parameters for the time period of test. The averaging is applied since, for the rest of the analysis, hourly data are required. On this dataset simple data quality checks are applied according to [167]. The employed parameters and the corresponding quality checks are summarised in Table 4.1 which are applied on the final hourly results. Specifically, the year 2014 was chosen for the majority of the validation processes as a year with relatively high solar resource as well as high data availability from the COMS.

Table 4.1. Measured parameters obtained from COMS and the simple quality checks applied.

Measured parameter	Minimum value	Maximum value
In-plane Irradiance	0	1300 W/m^2
Global horizontal irradiance	0	1300 W/m^2
Ambient temperature	-20 ^{o}C	+50 ^{o}C
Module temperature	-20 ^{o}C	100 ^{o}C
Current – Voltage at maximum power point	0 A 0 V	-

The electrical data are taken for two PV module types from the CREST monitoring system which are described in Table 4.2.

Table 4.2. CREST PV modules used for the demonstration of training and back-filling procedures.

Metrics in absolute units	Module types	Tilt angle (°)	Azimuth	Nominal power (W)	Mounting type	Data origin
Module A	Mono-crystalline silicon (c-Si)	32.5	South	245.0	Open-rack	CREST outdoor monitoring system
Module B	Poly-crystalline silicon (pc-Si)	32.5	South	245.0	Open-rack	CREST outdoor monitoring system

4.3.2 Ambient temperature and global horizontal irradiation

In cases 2 and 3 weather data (namely solar radiation and ambient temperature) are not available from the monitoring system. Thus, these are synthesized from remote UK weather stations and the kriging interpolation technique as described in 3.3. Hence the accuracy with which these values are predicted plays a significant role in the prediction of energy output used for back-filling. The following analysis shows the accuracy of the inferred global horizontal solar radiation and ambient temperature with regards to measured data from COMS. Historical data of monthly total horizontal irradiation are compared to modelled (interpolated) data for the building (W-roof) where COMS is located. The annual analysis results have shown good agreement for the years 2011 to 2013 and are shown in Figure 4.2.

Figure 4.1. CREST outdoor monitoring facility. The modules used in this work are placed at the highest rack as the red arrow indicates.

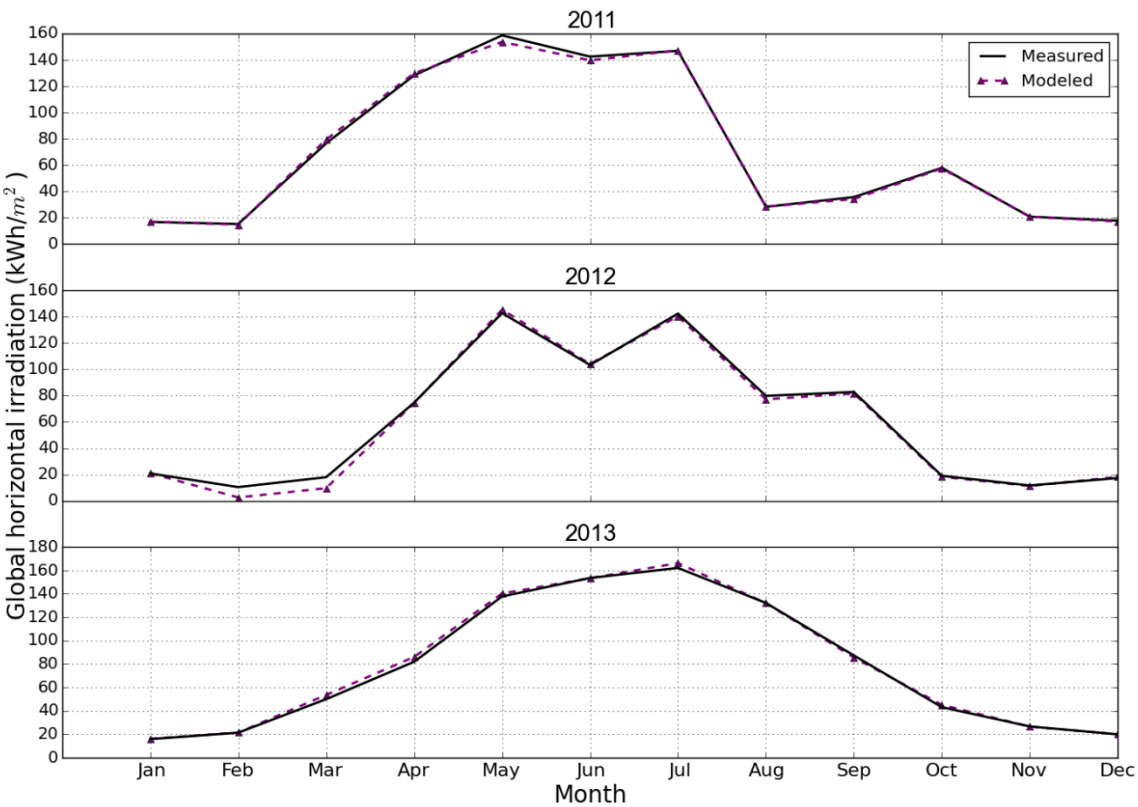


Figure 4.2. Measured versus modelled monthly global horizontal irradiation for the years 2011 to 2013.

The average monthly (absolute) deviation for the years 2011 to 2013 is approximately 1.75 kWh/m^2 but showed noticeable underestimation particularly for February and March 2012. This can be due to the particularly low irradiation conditions occurring these months compared to other years.

Table 4.3. Statistical metrics for the comparison of monthly and annual modelled and measured global horizontal irradiation for the years 2011 to 2013.

Metrics in kWh/m²	**2011**		**2012**		**2013**	
	Monthly	*Annual*	*Monthly*	*Annual*	*Monthly*	*Annual*
RMSE	1.99	5.76	3.57	18.9	2.24	14.6
MAE	1.35	"	2.32	18.9	1.59	"
MBE	-0.48	-5.76	-1.57	-18.9	1.22	14.6

Aggregated monthly and annual results are very close to measured values for global horizontal irradiation (GHI) and ambient temperature as seen in Table 4.3, and in Figure 4.3 and Figure 4.4 for the year 2014. Higher resolution analysis is carried out for the same year (2014), where hourly and daily results are also compared.

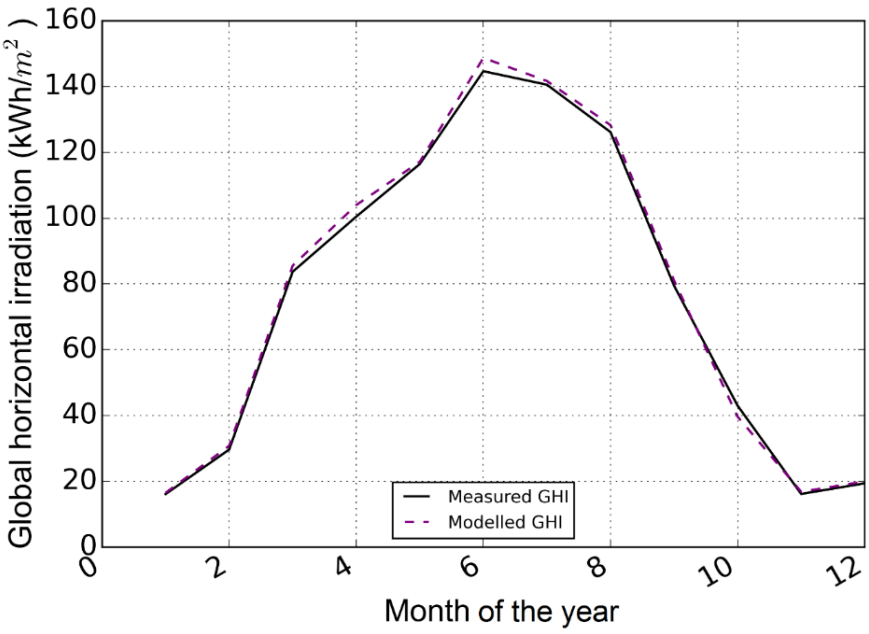


Figure 4.3. Measured versus modelled (krigging) monthly global horizontal irradiation for 2014.

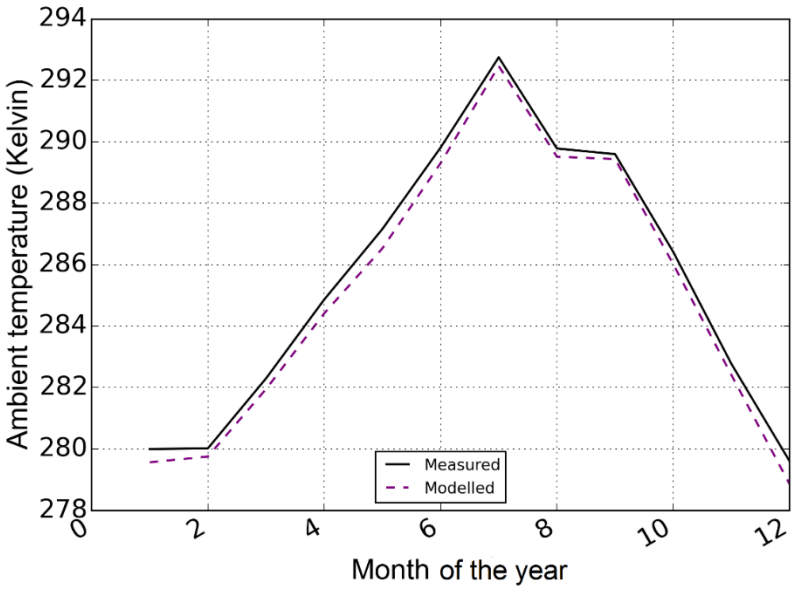


Figure 4.4. Measured versus modelled (krigging) monthly average ambient temperature for 2014.

Table 4.4. Statistical metrics for the comparison of monthly and annual modelled and measured ambient temperature and global horizontal irradiation for 2014.

Metrics in absolute units	**Ambient temperature (Kelvin)**		**Global horizontal irradiation (kWh/m^2)**	
	Monthly	*Annual*	*Monthly*	*Annual*
RMSE	0.43	0.40	2.13	14.3
MAE	0.40	0.40	1.75	14.3
MBE	-0.40	-0.40	1.20	14.3
%RMSE	0.15	0.14	2.79	1.54
%MAE	0.14	0.14	2.30	1.54
%MBE	-0.14	-0.14	1.54	1.54

At hourly temporal resolution, RMSE becomes higher and this can be demonstrated by using scatter diagrams for both global horizontal irradiation and ambient temperature. This is much more evident in solar radiation than in ambient temperature, as temperature is spatially more homogeneous, considering locations in close proximity, whereas solar

radiation depends on more complicated atmospheric conditions such as cloud movement, which are not being accounted for. A linear regression model between the modelled and measured results is also shown in Figure 4.5 and Figure 4.6.

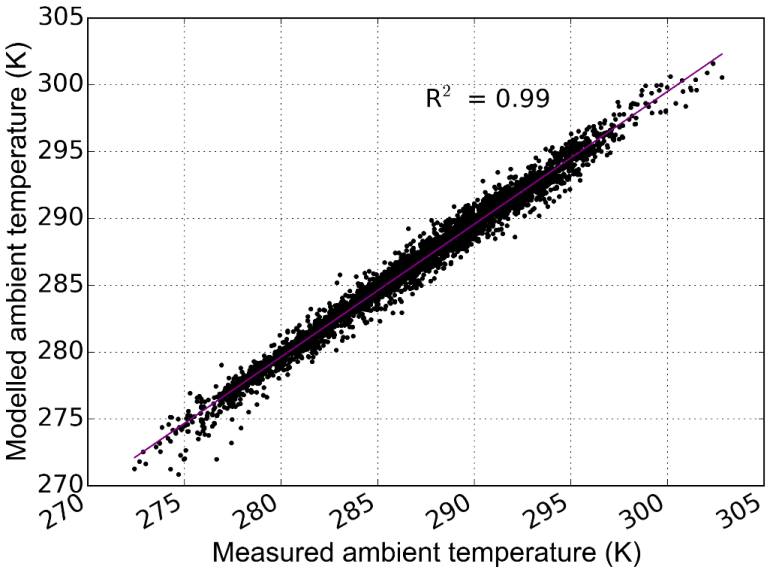


Figure 4.5. Scatter diagram of hourly measured versus modelled ambient temperature.

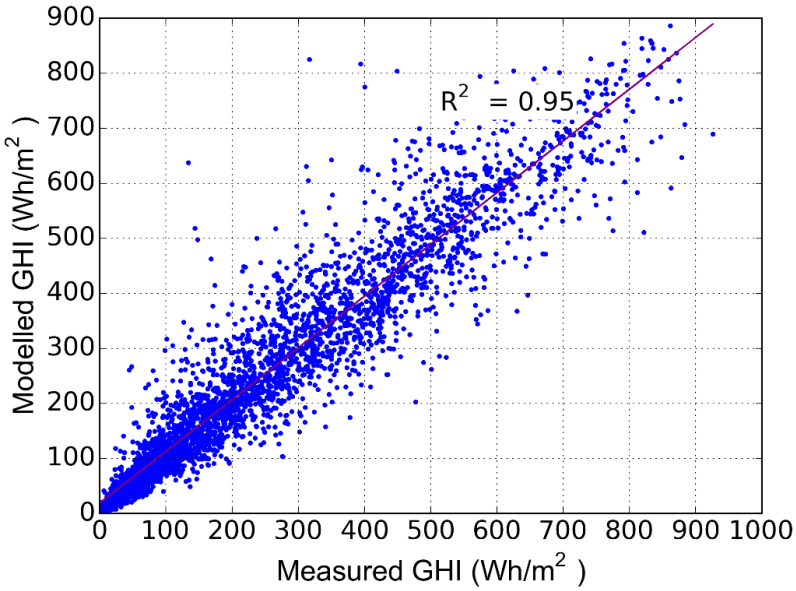


Figure 4.6. Scatter diagram of hourly measured versus global horizontal irradiation (GHI).

The R^2 coefficient of determination is a statistical measure which shows how well the regression line approximates the real data points, namely indicates the linearity between the

measurement and the model. The closer R^2 is to unity the better the regression line fits the data. Both hourly and daily statistical results are shown in Table 4.5. The low bias in hourly data indicates that the averaged values are expected to yield a better agreement between measured and modelled values.

Table 4.5. Statistical metrics for the comparison of hourly and daily modelled and measured ambient temperature and global horizontal irradiation for 2014.

Metrics in absolute units	**Ambient temperature (Kelvin)**		**Global horizontal irradiation (kWh/m^2)**	
	hourly	*daily*	*hourly*	*daily*
RMSE	0.81	0.59	0.06	0.30
MAE	0.60	0.46	0.04	0.21
MBE	-0.40	0.41	0.04	0.44
%RMSE	0.28	0.21	25.8	10.8
%MAE	0.46	0.16	16.7	7.65
%MBE	-0.14	-0.14	1.57	1.57

This random error which causes the scatter in hourly global horizontal irradiation can be justified; firstly, prior to kriging, the raw measurements are taken at minutely intervals at each meteorological station and then, they are averaged at the end of each observation hour. This procedure is carried out by the Met Office. Theoretically, the timestamps corresponding at each hour for every meteorological station are the same. This corresponds to about 88 timestamps for every daylight hour of the day, which are then to be (spatially) interpolated. In reality, however, it is likely that a given hour is represented by a normal distribution of timestamps with an average deviation of a few minutes with regards to each other. Namely, a station may have several minutes delay compared to another station. That is already a factor which may increase the random error in the prediction whose impact however, cannot be quantified and it is currently impossible to predict as raw data are unavailable at all cases (or prohibitively expensive).

A second factor is that not all stations provide readings for every day or every hour of the day. Namely, Kriging is not always applied based on the same number of stations and thus

the prediction error may sometimes be higher. This is further discussed in detail in [168]. The main factor, however, is that solar radiation is highly variant throughout the day and this variation cannot be modelled as accurately as real time on-site measurements.

4.3.3 Global plane of array (in-plane) irradiation (POA)

Next, global in-plane irradiation is compared to actual in-plane irradiation for one year of study, where in-plane is calculated from global horizontal irradiation using the separation and translation models given in 3.3.3. A noticeable deterioration in the statistic metrics is noted for in-plane irradiation, which is primarily due to the sub-models used in the process of translation of global horizontal radiation into the plane of array [112]–[116].

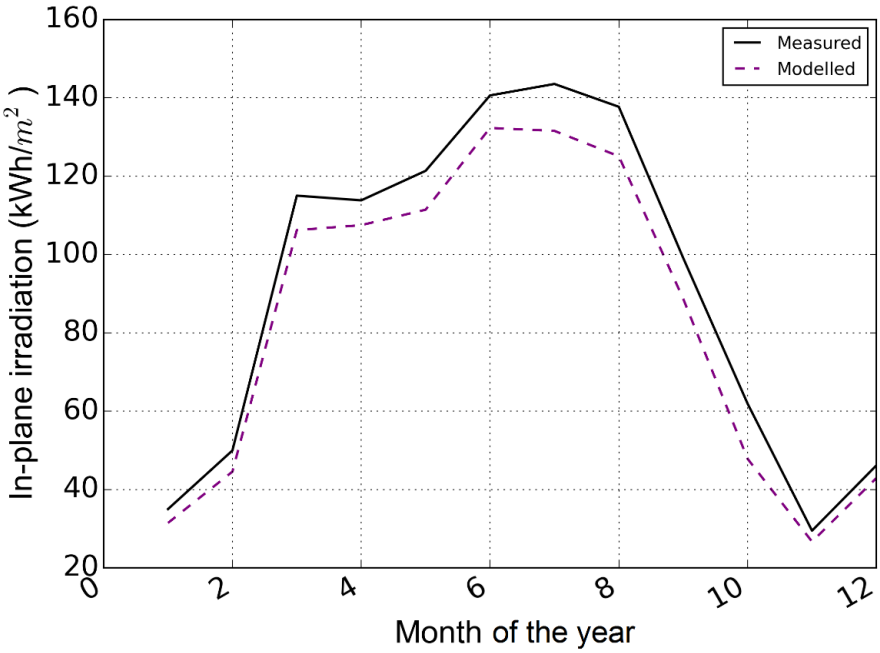


Figure 4.7. Measured versus modelled monthly global in-plane irradiation for 2014.

Table 4.6. Statistical metrics for the comparison of hourly, daily, monthly and annual modelled and measured global in-plane irradiation for 2014.

Metrics in absolute units	**In-plane irradiation (kWh/m²)**			
	hourly	*daily*	*monthly*	*annually*
RMSE	0.10	0.54	8.90	96.7
MAE	0.07	0.39	8.06	96.7
MBE	-0.03	-0.30	-8.06	-96.7
%RMSE	36.1	16.1	9.78	8.86
%MAE	24.0	11.8	8.86	8.86
%MBE	-8.86	-8.86	-8.86	-8.86

This is to be anticipated, as separation (global horizontal irradiance to beam and diffuse) and translation algorithms add a high percentage random and bias error [112],[114]. In-plane irradiation is generally underestimated, with some days giving better results than others. A further analysis based on different irradiation bins and clearness indices shows that the random error derives mainly for low irradiation and partly cloudy days as seen in Figure 4.8 Figure 4.9. The width and the number of the irradiation bins in Figure 4.8 were adjusted considering the frequency of irradiation values so that RMSE in different bins is affected by the same number of observations. Lower irradiation values present a higher percentage RMSE which however is small in terms of absolute difference. The data points in Figure 4.8, represent the calculation bias, which changes according to hourly clearness index (k_t). It seems that the method tends to underestimate higher irradiance (negative MBE) which presents a higher bias error whereas for irradiance values lower than 100 W/m^2 the result is slightly overestimated (positive MBE).

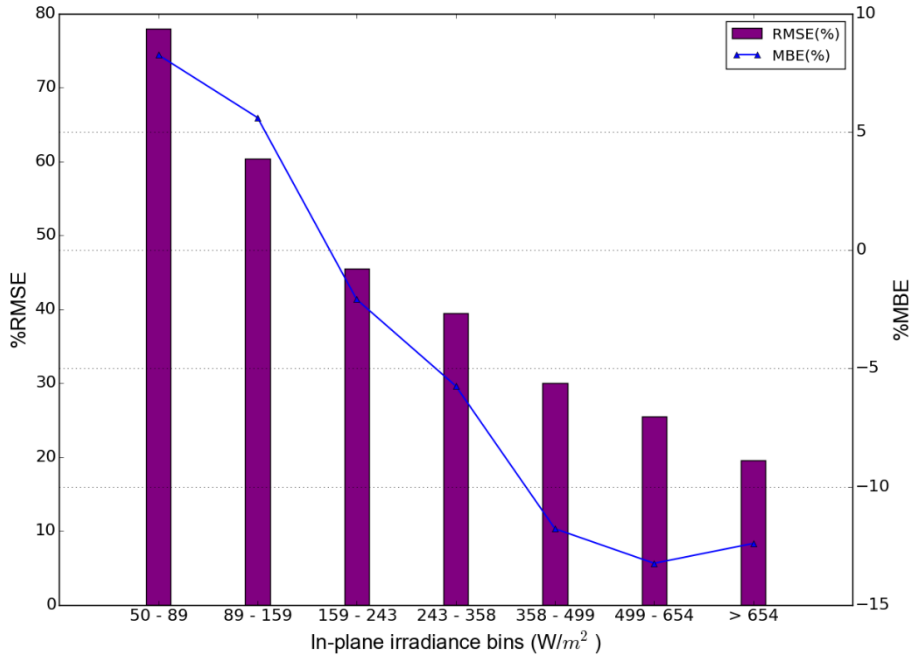


Figure 4.8. Statistical analysis on hourly irradiation bins for relative RMSE and MBE metrics

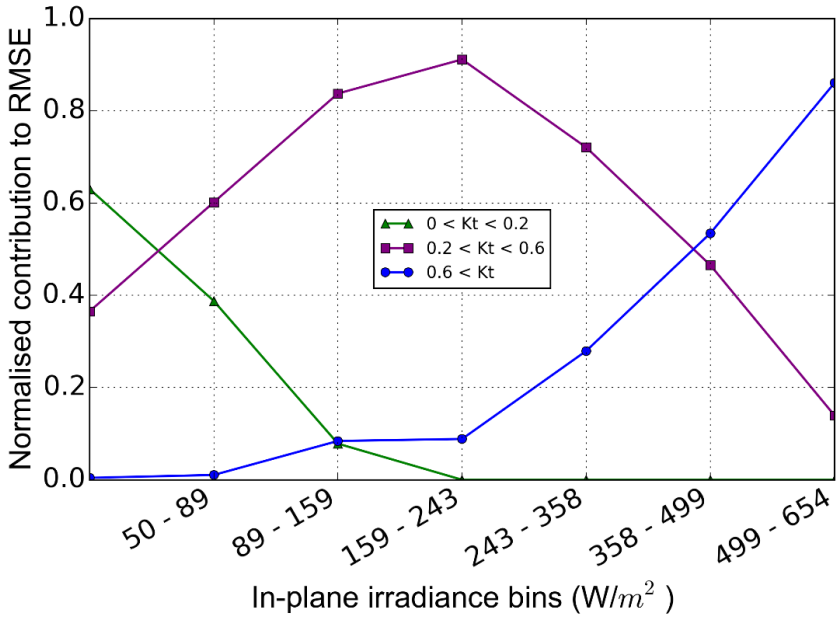


Figure 4.9. Normalised contribution to RMSE of hours with different clearness index k_t.

This was found to be due to the separation into beam and diffuse algorithms which tend to overestimate diffuse radiation for days with higher clearness index. Partly cloudy days contribute significantly to the overall random error for the majority of the bins. This is also expected considering the patterns in cloud movement which vary across the distributed

stations and is also evident in modelled GHI. In days with high daily clearness index ($K_t > 0.6$) this phenomenon is much weaker, which results in more accurate agreement between modelled and measured hourly results both for POA and GHI as seen in Figure 4.10. However, since there is only a small percentage of clear days in the UK – about 9.2% days where $K_t > 0.6$ in 2014 - the monthly and annual results are mostly affected by the days with lower clearness index and the overall bias is thus increased.

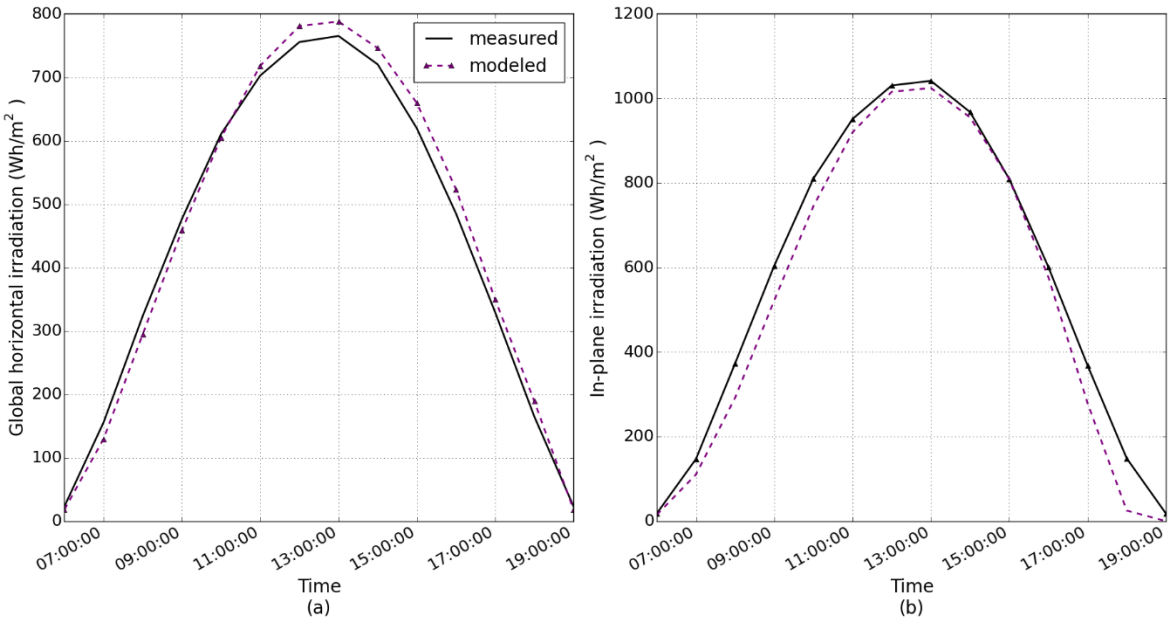


Figure 4.10. Measured versus modelled (a) global horizontal and (b) in-plane irradiation on a clear day (16-Apr-2014).

4.4 Choice of models used in back-filling

For the simulation of electrical output and module temperature (where this is not directly obtained from the monitoring system), two models are chosen respectively.

4.4.1 Electrical model

The model expressed by (4.7) and as also seen in 2.2.2 is an empirical model, which was chosen for two main reasons; a) its small number of input parameters, in-plane irradiation

and module temperature, which thus enables fast computational time and b) its coefficients can be trained and fitted to existing data.

$$P'_{\mathrm{AC}} = \mathrm{G}'(1 + \mathrm{k}_1 C + k_2 C^2 + k_3 T'_m + k_4 C T'_m + k_5 C T'_m + k_6 T'^2_m) \tag{4.7}$$

Where

G' = Normalised irradiance to STC = G/G_{STC}

P'_{AC} = Normalised maximum power to STC = P_{AC}/P_{STC}

$T_m{}'$ = Module temperature difference from STC = T_m - T_{STC}

C = $\ln(G')$

$k_1 - k_6$ = empirical coefficients

Additional reasons are that the specific model has been compared with a number of other models [65] where it was found that it performed well for a range of PV module technologies, on predicting annual energy output and it is also possible to combine measured data from many PV modules to obtain a general model for a given PV technology [70]. It is also a candidate model for its implementation in the module energy rating standards as discussed in the relevant round-robin results [8],[65],[67]. By varying the $P_{STC,}$ Equation (4.7) can be used to describe a single module to an entire PV system. In this study, defining the coefficients for a specific system is essentially training the model based on the specific system characteristics and re-using these coefficients to predict the output of the missing period as will be seen in Section 4.6.

4.4.2 Thermal model

Module temperature is calculated from in-plane irradiance and ambient temperature using the thermal model presented in [124] and Equation (2.34):

$$T_m = T_a + k_R \cdot G \tag{4.8}$$

The k_R, or Ross coefficient is the modified thermal resistance of the module, modified in terms of influence of the mounting configuration of the array [126] where typical values of

which are given in Table 2.2. Ross's model is a good choice in cases where irradiance and ambient temperature are the only available input parameters, which is the case for remotely inferred weather data. Furthermore, *k* can also be obtained using outdoor measurements of module (T_m) and ambient temperature (T_a) and in-plane irradiance. In this work, Equation (4.8) was used by taking the hourly values of irradiance as proposed in [122]. In the case where CREST data are used, *k* was obtained experimentally for each module by linear fitting of (T_m-T_a) against in-plane irradiance for one year's worth of data (see Figure 4.11). The fitted value of *k* was found to be 0.027 K·m^2/W, which is very close to the value suggested from literature for a flat roof [125]. This model does not include wind speed thus it is more suitable for describing steady state conditions and the reason why there is an increased scatter in Figure 4.11.

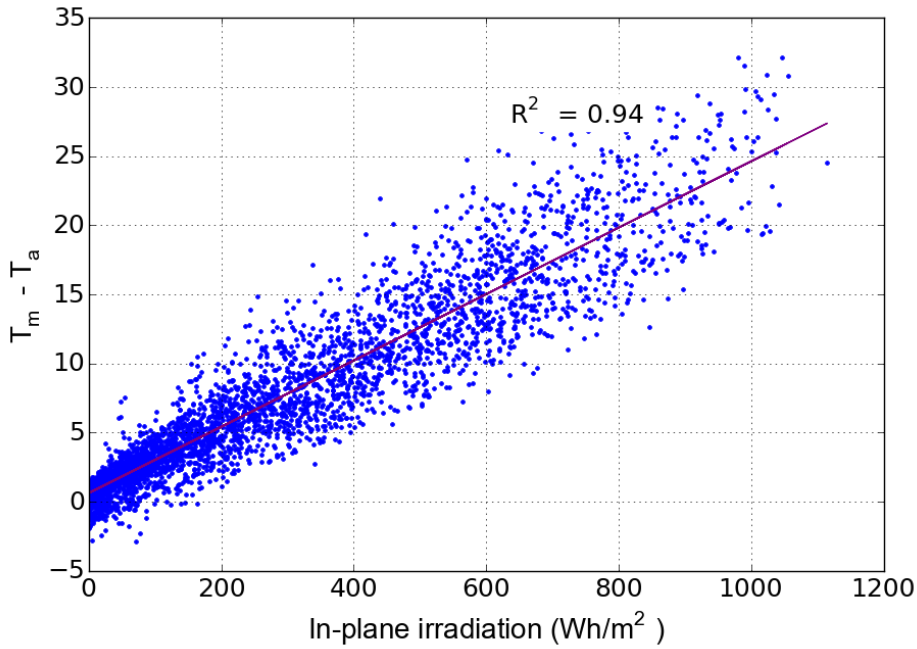


Figure 4.11. Scatter diagram of the difference between module and ambient temperature against in-plane irradiation.

4.5 Back-filling flowchart

To calculate and back-fill energy output, Equation (4.7) is employed twice. Initially, it is used with hourly measured data of in-plane irradiation, module temperature and energy output to extract the model coefficients using the training period as defined by the training algorithm described in Section 4.6. Then, it is applied again to calculate the energy output for the missing month, by using either interpolated or measured climatic data (depending on the availability of data) *only* for this period. Aggregated irradiation is calculated using hourly sums of irradiance. Module temperature, if not available for the missing period, is calculated using Equation (4.8) with interpolated irradiation and ambient temperature as input parameters, otherwise it is taken from the monitoring system. The whole procedure steps are described in the flowchart in Figure 4.12.

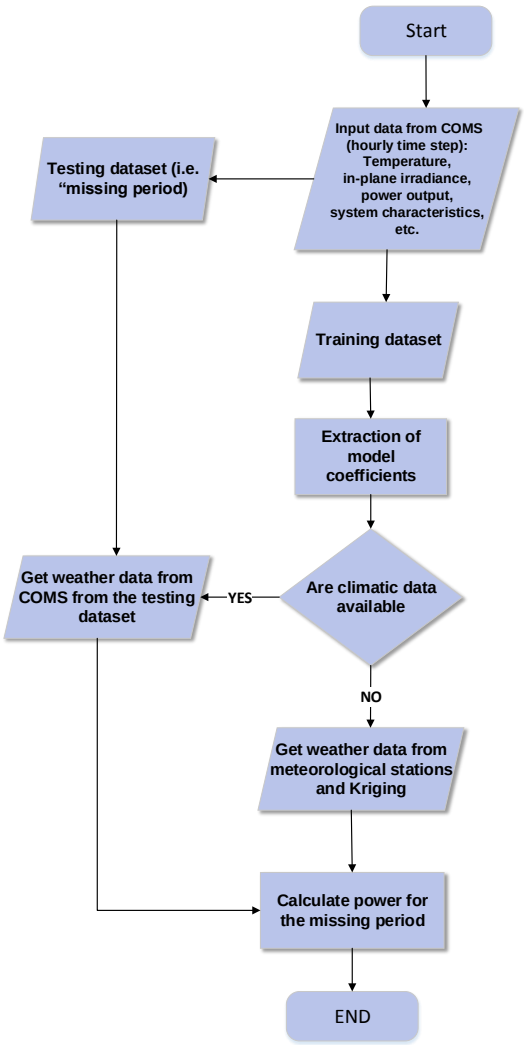


Figure 4.12. Flowchart of the back-filling process.

Finally, performance ratio is calculated based on the inferred energy output and the dataset of in-plane irradiation already employed for the back-filling of the latter.

4.6 Determination of the training set

The first step prior to applying back-filling is to define the training dataset which will be used to acquire the coefficients for (4.7). Defining the coefficients for a specific system is essentially training the model based on the specific system characteristics and re-using these coefficients to predict the output of the missing period [70]. The requirements for the training need to be determined in terms of the optimal training set's size and how recent it should be with respect to the missing period of data, in order to achieve maximum agreement between predicted and actual energy output. This is because system performance is affected by seasonal variations as well as by module technology [80], [169] and this is expected to have an impact upon the determination of the optimum training set. So, the training coefficients will vary over different PV systems or different training periods.

The process of determination for the optimum size of the training set is demonstrated using data for PV module A in Table 4.2 and assuming one missing month (June 2014). For the training, a validation period ("missing" set) is removed from the dataset and is later used for comparison with the predicted output. Hourly data of in-plane irradiation, module temperature and energy output are taken from the training set and used to fit the model (see (4.7)) by means of a Marquardt-Levenberg optimisation algorithm [170], which yields the optimum coefficients (k_1–k_6) for that training set. At each training cycle the training set size varies with adding number of days before (going backwards in time) and/or after (going forwards in time) the missing period, until the remainder of the whole year is employed.

By varying their size within a year's full of data, each training set is sorted based upon the lowest MBE achieved as well as their hourly RMSE. MBE deviation across different sets is very low and equal to 1.67Wh with the highest MBE being 1.71Wh which is only about 0.006% of the total monthly energy output. Hourly RMSE with regards to different training sets is further shown in Figure 4.13.

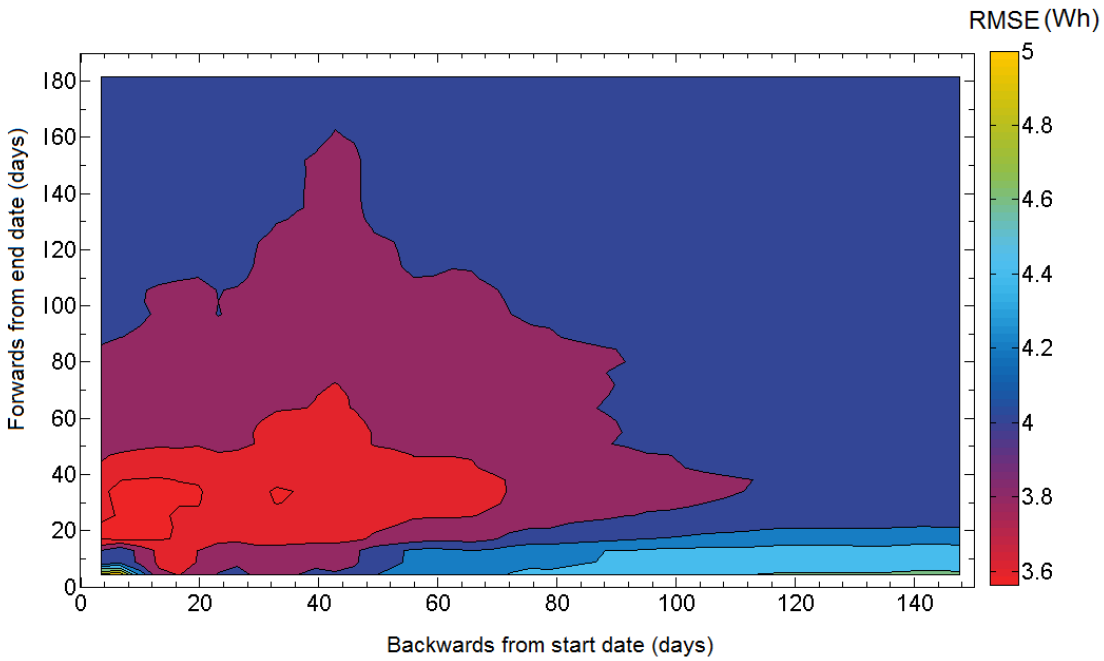


Figure 4.13. Training sets around the missing period taking as "start date" the 1st of June and "end date" the 30th of June 2014, and going backwards and forwards in time, respectively. The starting point (0,0) indicates the 1st and the 30th of June.

It can be seen in Figure 4.13 that although RMSE does not vary significantly across different training sets, there is a specific (red) area where it showed its lowest values. This area includes points that are closer to the missing period, which can be justified as seasonal dependence. The results also showed that very small training sets yielded the highest RMSE, e.g. using only several days before and after the missing period was not a sufficient data pool. This can be due to location and the local weather phenomena. Smaller training sets are expected to suffice for less variable weather patterns (for example a Mediterranean summer). The optimal training set size varies between 40 and 50 days in total, whereby good agreement is obtained with training sets varying from 20 to 25 days before and after the missing period, which is also taken into account for the back-filling cases carried out next. Repeating the process for larger missing periods of 2 consecutive months showed that this training set size is sufficient for estimating the monthly energy output with very low bias, a case which is further examined in 4.8.4. The training algorithm defined the best set of coefficients (k_1-k_6) which then provided the power surface shown in Figure 4.14.

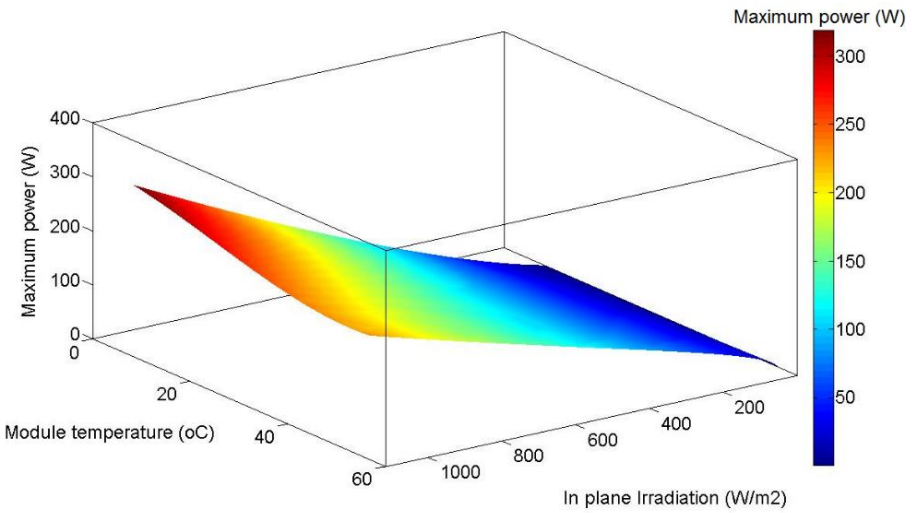


Figure 4.14. Fitting curve for the optimum training set (20 days backwards and 26 days forwards) for module A.

To further test and validate the accuracy of the selected model and its customised coefficients, simulated hourly power output is compared to measured hourly output (from COMS) for a different month (May 2014). The validation results yielded an hourly %RMSE and %MBE of 6.6 and 0.32 respectively, where absolute RMSE and MBE were 4.7 and 0.23 (in Watts) respectively. These figures give an idea of the error deriving from modelling the energy output, which is very low compared to the error contribution from all irradiation modelling steps. The result is depicted in Figure 4.15.

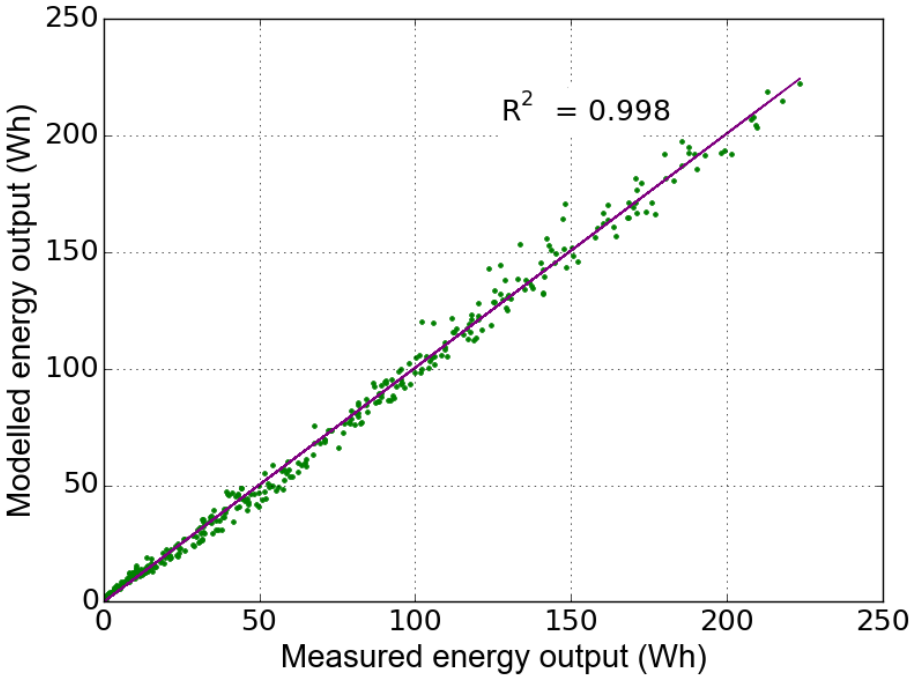


Figure 4.15. Hourly modelled versus measured energy output for a selected month (May 2014).

Finally, the training process yields valid results if no significant changes (i.e. component failures) have occurred in the PV system during its operation while no data are available (i.e. during the missing period). This is an assumption made at all cases of back-filling and it is further discussed in the final section of this chapter.

4.7 Different cases of data loss and back-filling strategies

4.7.1 First case: Missing electrical output

This case of data loss can be graphically represented by the sketch in Figure 4.16. A PV system from the UK field trials (found as Site A) was used, comprising a single string of 8 polycrystalline modules (where P_{STC} = 120W per PV module).

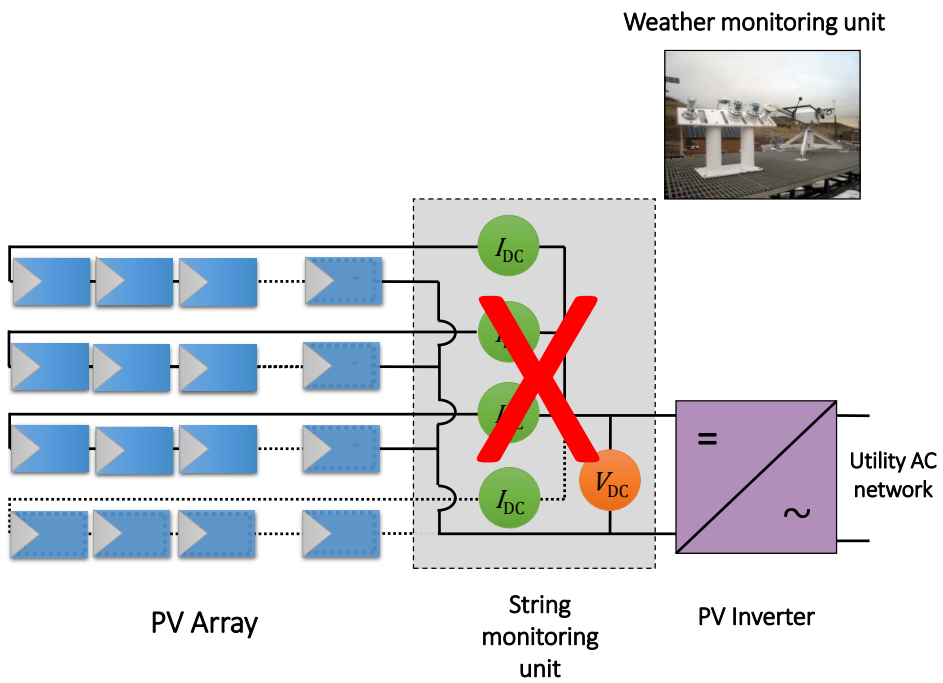


Figure 4.16. First case of missing data using a simplified sketch of string monitoring. Missing power output while weather monitoring is available.

It has been assumed that one month of DC readings is missing as a realistic case but also because it comprises a sufficiently large data pool for the demonstration of the results. Dates around the missing period i.e. the past and later days around this gap (40 days in total) are used to extract the coefficients for the model. This method gives good agreement with measurements for hourly, daily and monthly RMSE of 0.02, 0.06 and 0.02 kWh respectively (see Figure 4.17, Figure 4.18).

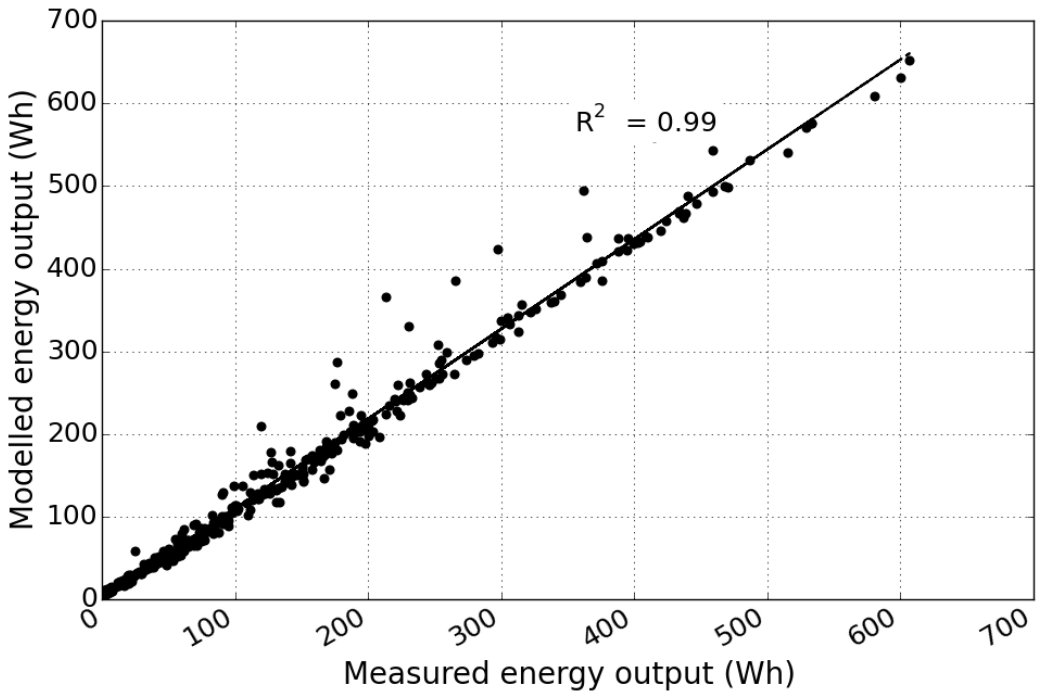


Figure 4.17. Hourly back-filled versus measured energy output for a selected month (March 2005).

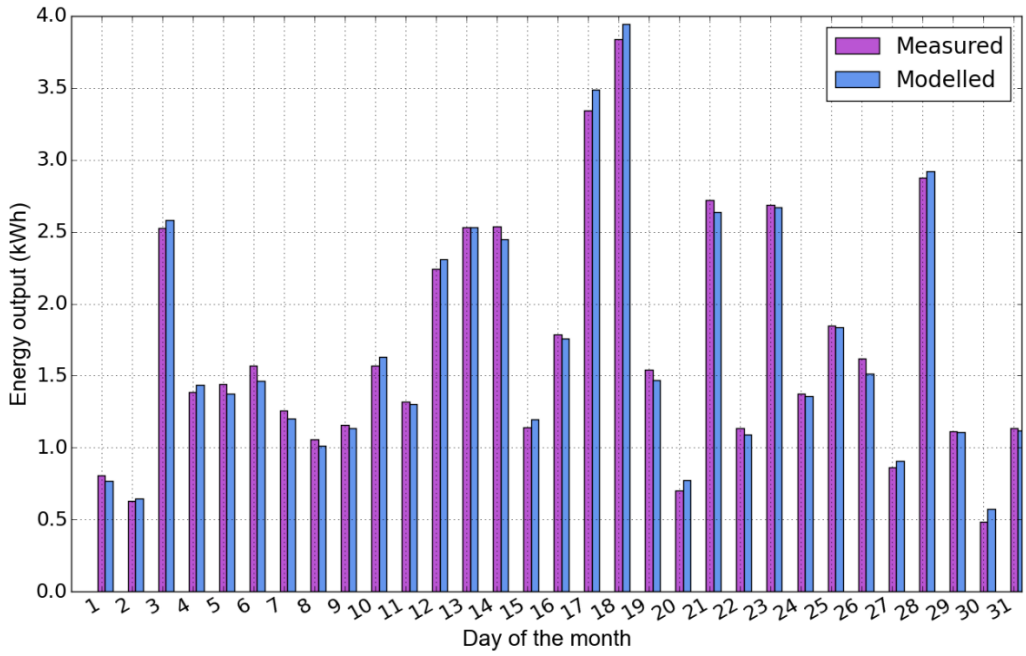


Figure 4.18. Comparison of modelled and measured energy output and PR for the missing month.

Additionally, using the replenished period to acquire the monthly PR of this system, it was found that the deviation from the actual PR (0.7545) was just about 0.0001 (namely 0.7544), thus yielding a very accurate monthly PR. Given that the training model provides accurate agreement between measured and modelled energy output (see Figure 4.15), this result is

expected. However, attention should be specifically paid when using sensor measurements from less reliable sources as these are often contaminated with random abnormalities (such as negative irradiance or extreme temperatures). Thus, the efficiency of the inferred output largely depends on the quality of the input data and specifically those random occurrences. Repeating the back-filling procedure for 10 months (February to November) in the same year, the median of the absolute differences between modelled (back-filled) and actual monthly energy output was 0.87 kWh where MAD = 0.35 kWh and the highest deviation occurring in May (equals 5 kWh). This is due to the model slightly overestimating energy output at higher irradiance (400 to 800) and elevated module temperatures (over 40 ^{0}C)[70], which often occurs in May for the particular dataset as opposed to the rest of the days used in training.

4.7.2 Second case: Missing electrical and meteorological data

In this case of data loss concurrent energy yield readings and a climatic dataset were utilised containing a period of one month of missing data (June 2014), during which *neither* (namely energy output or meteorological data) of the above information is available. This case of data loss can be graphically shown in Figure 4.19.

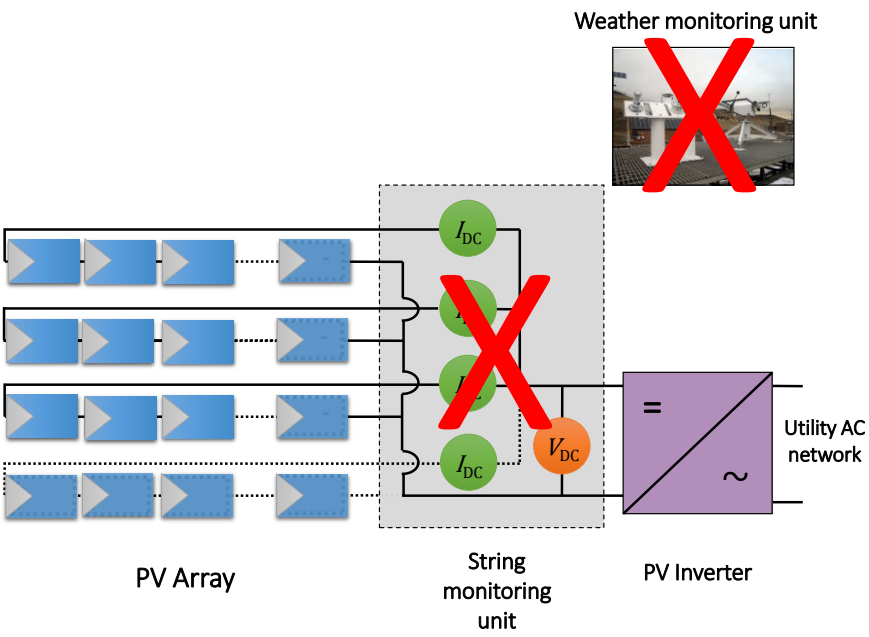


Figure 4.19. Second case of missing data using a simplified sketch of string monitoring. Both power output and weather data are lost.

In order to validate the modelling results, this period is completely removed from the initial dataset and is treated as the "missing" period. The modelled parameters of the inferred data are compared to the measured data and the results are depicted in Figure 4.20, Figure 4.21 and Figure 4.22, followed by the statistical results gathered in Table 4.7 and Table 4.8.

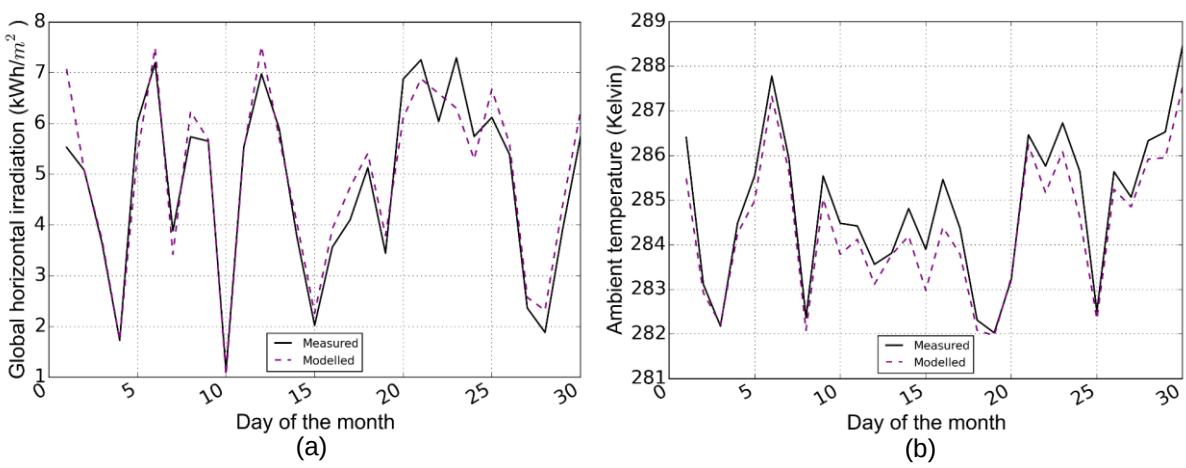


Figure 4.20. Daily interpolated (a) global horizontal irradiation (GHI) and (b) average ambient temperature as inferred for June 2014.

The majority of the days in Figure 4.20 show satisfactory results for both predicted variables. Specifically, for GHI in June, the daily RMSE is affected by those few days where higher deviation is observed. Hourly results show a higher random error of about 27% but this is to be expected for higher resolution analysis (higher than daily). For ambient temperature daily %RMSE is 4.0 with a maximum deviation as low as 1.5 ^{o}C and hourly %RMSE is 5.2 with a maximum deviation of 3 ^{o}C. The results for ambient temperature show that it can be interpolated to the location of interest with a very small MBE and RMSE.

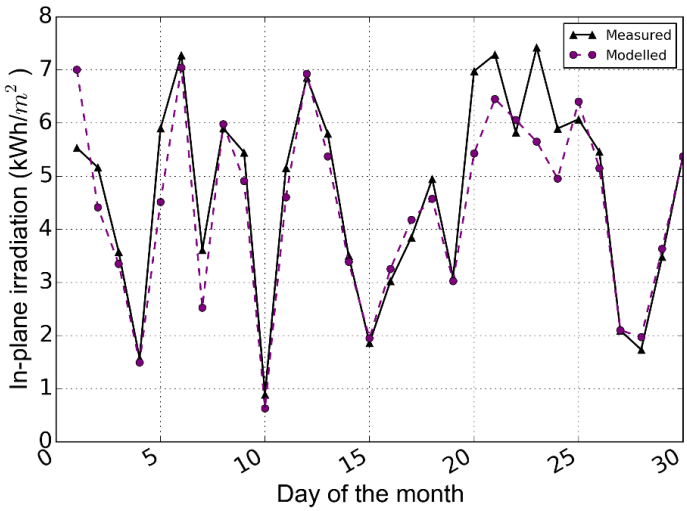


Figure 4.21. Daily in-plane irradiation as inferred from interpolated GHI for June 2014.

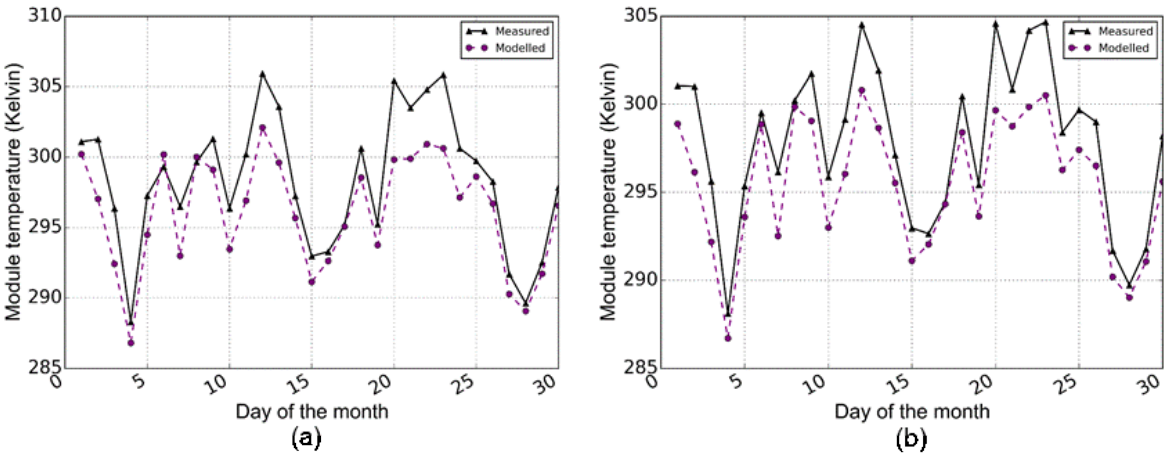


Figure 4.22. Daily average module temperature for (a) module A (c-Si) and (b) module B (pc –Si).

Table 4.7. Statistical results for in-plane irradiation and module temperature comparisons.

Modelled Parameter	Statistical metrics*	Hourly	Daily	Monthly
Global horizontal irradiation	%RMSE	27.1	10.7	2.83
	%MAE	17.7	8.4	‘’
	%MBE	2.83	2.83	2.83
Ambient temperature	%RMSE	0.30	0.24	0.16
	%MAE	0.24	0.19	‘’
	%MBE	-0.16	-0.16	-0.16
In-plane irradiation	%RMSE	33.0	14.9	5.9
	%MAE	21.7	10.5	5.9
	%MBE	-5.9	-5.9	-5.9
Module temperature (A)	%RMSE	1.30	0.93	0.75
	%MAE	1.00	0.78	‘’
	%MBE	-0.75	-0.75	-0.75
Module temperature (B)	%RMSE	1.24	0.90	0.77
	%MAE	0.96	0.78	0.77
	%MBE	-0.77	-0.77	-0.77

*The statistical metrics are only presented in relative terms for conciseness.

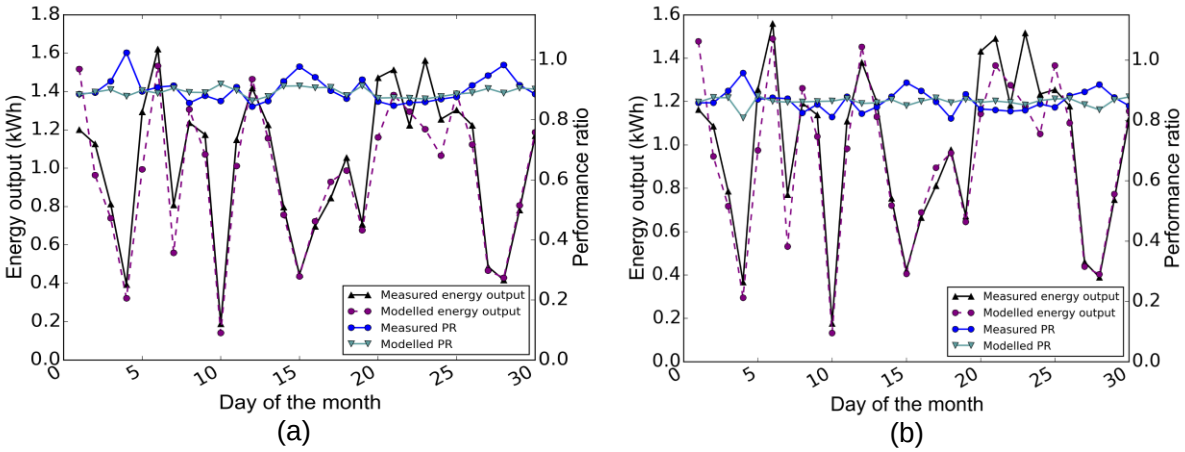


Figure 4.23. Comparison of daily modelled and measured energy output and PR for the missing month (June 2014) for (a) Module A and (b) Module B.

The MBE metric increases for module temperature due to error propagation from both in-plane irradiation (inherent underestimation) and ambient temperature, but the effect is very small with an absolute maximum deviation of about 5 degrees and average deviation of about 2 degrees. In Figure 4.23, there are days where ambient temperature is particularly low and thus actual PR may slightly exceed 1.0, leading to a small deviation from the modelled value. This is in fact an expected behaviour for the tested module types (for the UK climate at least) considering that module temperature is not significantly elevated compared to STC (around 30 ^{o}C) and irradiation is relatively high [80].

Moreover, module temperature plays a significant role in modelling the energy output, but this is not directly evident in the modelled PR. Specifically, if in-plane irradiation is overestimated, modelled energy output is affected by both in-plane irradiation and module temperature rise. The modelled result is very close to the measured energy output (where actual in-plane irradiation is lower) and thus, modelled PR is slightly lower than the measured value. This behaviour is particularly evident for days with lower average of module temperature (i.e. low ambient temperature and/or windy days).

Table 4.8 Statistical results for the back-filled energy output for module A and B.

Modelled Parameter	Statistical metrics*	Hourly	Daily	Monthly
Energy output (A)	%RMSE	33.1	14.5	5.9
	%MAE	21.4	10.8	5.9
	%MBE	-5.9	-5.9	-5.9
Energy output (B)	%RMSE	32.2	14.4	5.09
	%MAE	21.2	10.8	5.09
	%MBE	-5.1	-5.1	-5.1

*The statistical metrics are only presented in relative terms for conciseness.

In Figure 4.24 (a) and (b), the scatter diagrams of modelled and predicted in-plane irradiation and energy output, respectively, show that the discrepancy for energy output is higher than in the case depicted in Figure 4.15 with a small number of outliers, which contribute in higher RMSE values. It is, however, expected that any analysis of higher temporal resolution than daily will yield a higher random error which primarily derives from interpolated global horizontal irradiation as previously discussed.

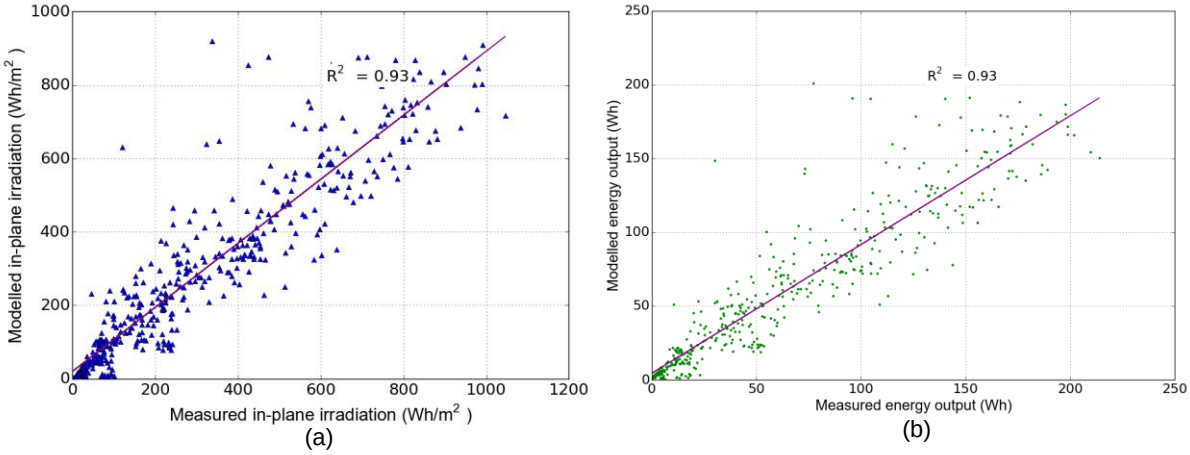


Figure 4.24. Scatter diagrams for module A, of hourly modelled and measured (a) in-plane irradiation and (b) energy output for the missing month (June 2014).

The results for energy output errors are a propagation of both in-plane irradiation and module temperature. Therefore, the negative bias which arises primarily due to in-plane irradiation is evident also in energy output. In terms of absolute RMSE for monthly energy output, that is 1.8 and 1.5 (in absolute kWh) for module A and module B respectively. The derived scatter which is evident in both in-plane irradiation and measured energy output is due to the high random error inherent in global horizontal irradiation and kriging. This discrepancy is, however, largely diminished with regards to modelled performance ratio. In fact, the monthly PR can be predicted with a very small error for both cases. More specifically, the performance ratio difference (ΔPR) for module A and module B is $\Delta PR_A = -0.0001$ and $\Delta PR_B = -0.0024$ respectively. This corresponds to a deviation as low as -0.02 % and -0.3% from the actual monthly PR values, which is insignificant considering that the maximum monthly PR deviation within a year is 5.6% for module A and system PR variations can be up to 10% throughout the year [159].

4.7.3 Third case: Loss of energy output where no climatic data are available

This case describes a situation where energy readings (only) are missing for a specific time period and climatic data are not available from the monitoring system at any time (see Figure 4.25). This is a case that is most commonly found in domestic monitoring and, as already

shown in Chapter 3, in the case of the NCH dataset, a large number of homes presented missing data which even extended to one year in a few cases.

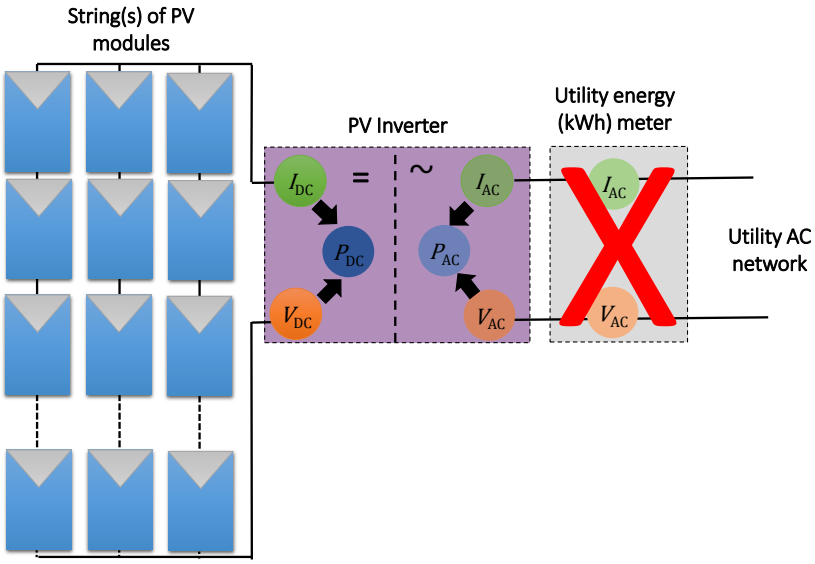


Figure 4.25. Third case of missing data on a simplified sketch of domestic monitoring. Weather monitoring is not available and power output data is lost.

The same procedure as in Case 2 is followed here but since there is no local climatic data monitoring, both training and back-filling are applied by using interpolated climatic data. Thus, essentially this case 3 is treated as case 2 but using a different dataset for training. An additional modelling feature is also used in this case, whereby *dynamic fitting coefficients* are applied to the electrical model (Equation (4.7)). The k_1-k_6 coefficients are fitted based on hourly bins of in-plane irradiation, module temperature and energy output. For this method, two weeks of missing data are back-filled for two random PV systems with high availability (see Table 4.9) and two different periods.

Table 4.9. PV system characteristics for case 3.

System	System description	Tilt angle (°)	Azimuth	Nominal power (kW)	Missing period	Data origin
A	A string of poly-crystalline silicon (pc-Si) modules	30	-30 (SE)	1.68	2 weeks (June and February)	NCH dataset
B	A string of poly-crystalline silicon (pc-Si) modules	30	10 (SW)	1.68	2 weeks (June)	NCH dataset

The back-filled daily output of system A is shown in Figure 4.26. Back-filled energy output showed good agreement with actual energy output, with daily RMSE and MBE values of 0.46 and -0.02, respectively.

Training with interpolated data diminished the negative bias of in-plane irradiation significantly. However, the random error is still evident in the hourly results as seen in Figure 4.27.

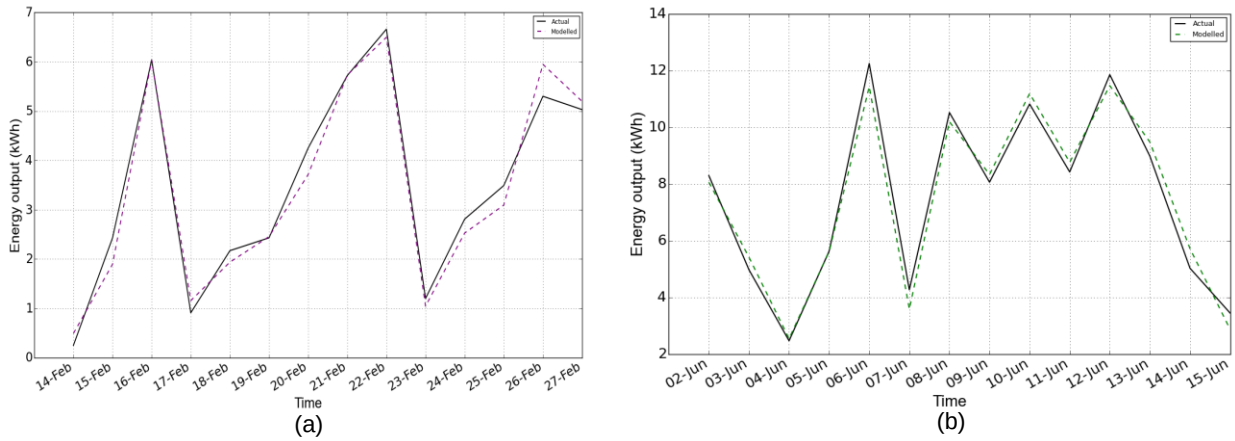


Figure 4.26. Actual and back-filled (modelled) energy output of a PV system A test case for 15 days in (a) February and (b) June.

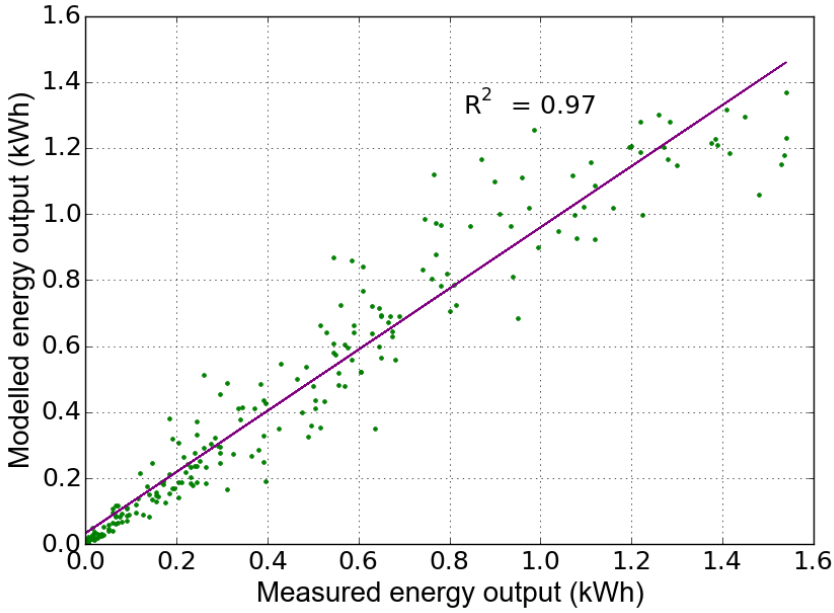


Figure 4.27. Scatter diagrams for module A, of hourly modelled and measured energy output for two missing weeks in June (2014).

Similar results are obtained for PV system B, a west facing PV system in Nottingham as seen in Figure 4.28. The corresponding statistical results for the tested cases are given in Table 4.10.

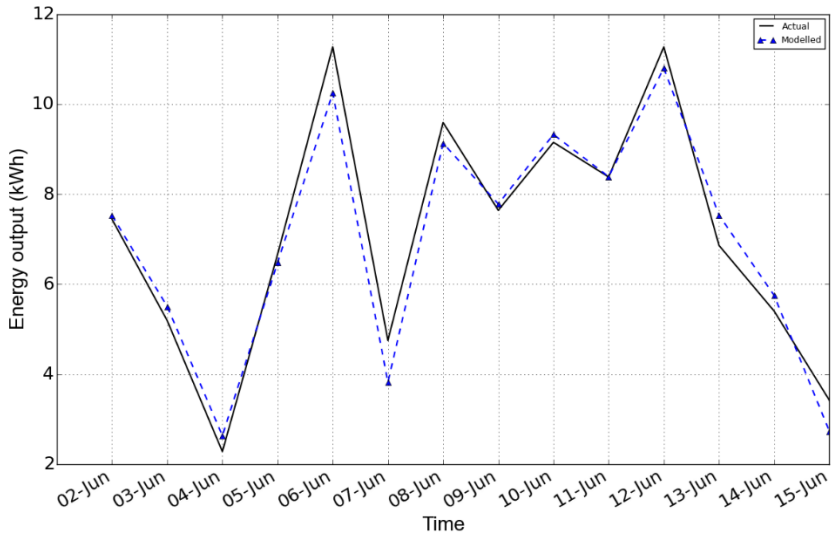


Figure 4.28. Actual and back-filled (modelled) energy output of a PV system B test case for 15 days in June 2014.

Table 4.10. Statistical results for the NCH case systems System A and System B.

System	Statistical metrics	Hourly	Daily	Monthly
System A (June)	%RMSE	23.5	6.0	0.18
	%MAE	15.5	5.3	0.18
	%MBE	-0.18	-0.18	-0.18
System A (February)	%RMSE	47.1	9.4	2.00
	%MAE	29.8	7.5	2.00
	%MBE	-2.00	-2.00	-2.00
System B (June)	%RMSE	33.2	7.2	1.7
	%MAE	22.8	5.9	1.7
	%MBE	-1.7	-1.7	-1.7

4.8 Inference of missing data by using Artificial Neural Networks

In the three aforementioned cases, it has been shown already that it is possible to back-fill missing meteorological and/or electrical data by utilizing synthetic weather data and/or electrical readings. Available data from the monitoring system are used for selected time periods surrounding the gap and good accuracy can be achieved on monthly and annual performance ratio. In the second case, however, bias deriving from inferring in-plane irradiation caused a noticeable underestimation in energy yield. This is due to the separation and translation algorithms employed to translate global horizontal irradiation to in-plane. In order to improve in-plane irradiation accuracy artificial neural networks (ANN) are employed here, replacing the two-step method, namely the separation and translation steps.

Neural networks have been employed previously in photovoltaic modelling to predict energy yield [171]–[174] energy consumption [175] solar radiation [176]–[178] and to translate global horizontal irradiance (GHI) to in-plane [179],[180] using case specific solar radiation datasets. So far, they have not been applied to infer missing performance data in solar photovoltaic datasets, which is a very common issue in PV monitoring. Moreover, the implementation and practicality of NN described in literature, depends on the availability of training data. Thus, in cases where no past data are available (for example a new location or a new PV system) conventional methods for predicting in-plane irradiation and/or energy output have to be applied instead, and so the applicability of a NN approach is decreased. In the case of missing datasets however, it is already assumed that data from past and future periods around the missing period are available and hence this is an application of NN that gains merit.

In this approach, both in-plane irradiation and energy output are predicted by using a single neural network, reducing the modelling time and effort of the overall modelling procedure. For the better comprehension of the suggested approach, basic aspects of the artificial neural networks are firstly presented, while the following sections focus on the description of the modelling procedure and the obtained results. An overall comparison between the two methods of missing data inference suggested so far is provided at the end of this chapter for case 2.

4.8.1 Basic theory of an Artificial Neural Network (ANN)

The artificial neural network (ANN), or simply neural network, is one of the most popular machine learning algorithms. The idea for their structure is evolved from the simulation of the human brain. The ANN are able to describe complex relationships between various elements in a way that often cannot be provided by mathematical models. Their key characteristic is their ability to adapt and learn. Each neural network has three critical components: node character, network topology, and learning rules [181]. Simply, a neural network can be represented by the diagram in Figure 4.29 representing only a single node or neuron.

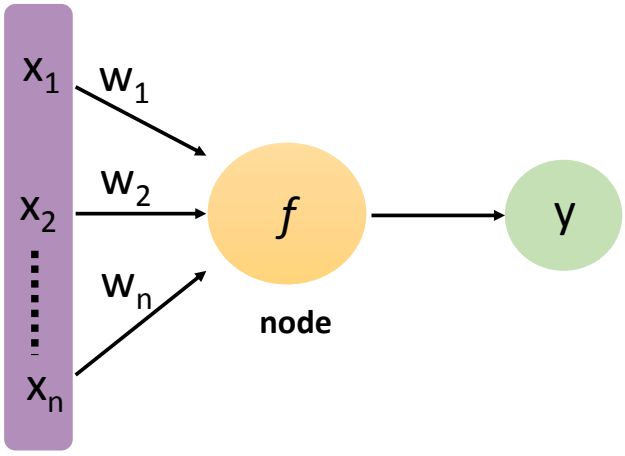


Figure 4.29. Basic single node structure with inputs (x_i), weights (w_i), transfer function (f) and output (y).

Each node receives multiple inputs from others via connections that have associated weights. When the weighted sum of inputs exceeds the threshold value of the node, it activates and passes the signal through a *transfer function* and sends it to neighbouring nodes. This can be simply described by the following mathematical expression:

$$y = f\left(\sum_{i=0}^{n} w_i x_i - T\right) \tag{4.9}$$

Where:

y = output of the node

f = transfer function

w_i = weight of input x_i

T = threshold value

The transfer function (also called *bias*) can have various forms but the most commonly employed is the type of a sigmoid function. This is usually either a logistic or a hyperbolic tangent type of sigmoid function, as shown in Figure 4.30.

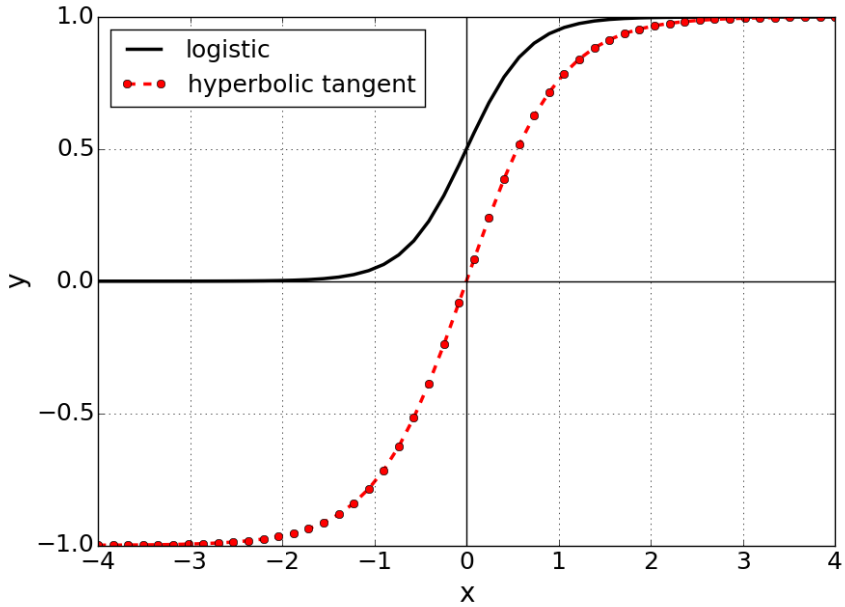


Figure 4.30. Example of two sigmoid functions as transfer functions.

The nodes are organized into linear arrays, which are called layers. A neural network structure comprises input layers, output layers, and hidden layers. In simple models, one hidden layer can be used whereas in more complicated structures there might be several hidden layers. Neural networks are also characterised for the type of connection between the nodes. So, in *feed-forward* ANNs (also called *perceptron*) the information only moves towards one direction, namely, the connections between the nodes do not form a cycle, as in the *recurrent* ANNs. In the case of a recurrent network, one input may produce a series of outputs thus, the complexity as well as the computational time of such a network increase.

During the learning phase, the weights of the inputs are configured such that the result is adjusted to the desired values. Again, there are two main categories in learning. The *supervised* and the *unsupervised* learning. In supervised learning, a training set, namely an array of inputs and an array of target outputs is given. The weights are adjusted to minimize the error between the network output and the correct output. In order to do so, most commonly, error correction methods are used whereby a back-propagation mechanism is applied. During back-propagation the weights are re-adjusted so as the error is decreased.

The error minimisation is achieved with a number of possible optimisation algorithms, such as the gradient descent function and the Levenberg-Marquardt. The error is constantly adjusted so as to succeed the best fit between inputs and targets. Normally, this error can be defined by the user in the beginning of the training process and once it is achieved the training is ended. The rate of the weight adjustment is the training rate of the network, which may become very high at complex problems. The training rate is mainly affected by the error minimisation function, but it also depends on the training set.

The ideal training set must be representative of the underlying model. A less suitable training set will not yield a reliable and general model while the training rate will increase in time. For networks using supervised learning, the network must be trained first. When the network produces the desired outputs for a series of inputs, the weights are fixed, and the network can be re-used to model the outputs for a new input dataset. Conversely, unsupervised learning does not use target output values from a training set, but the relationship is configured from the input data only.

4.8.2 Proposed ANN configuration

The training set used here aims to utilize as little information as possible, given that global horizontal irradiation is the only parameter which is usually available from the met stations or at least the only parameter which is currently interpolated. The chosen structure is a feed-forward back-propagation Levenberg-Marquardt [170] optimization algorithm. The applied configuration is chosen as being suitable for problems of relatively low complexity with few input nodes, as in studies with similar modelling requirements [171],[176]. Initially, this setup was applied within a trial and error process, but was finalised based on the obtained validation results in comparison with measured data. The structure of neural networks

applied in this work consists of three layers. One input layer, a single hidden layer and the output layer, which is demonstrated in Figure 4.31. As already mentioned, the neuron output is calculated based on the activation function. For the particular implementation the hyperbolic tangent activation function was used for the neurons in the hidden layer ($y = \tanh(n)$). Moreover, simple NN training sets are proposed which exploit as few input data as possible, increasing the method's efficiency.

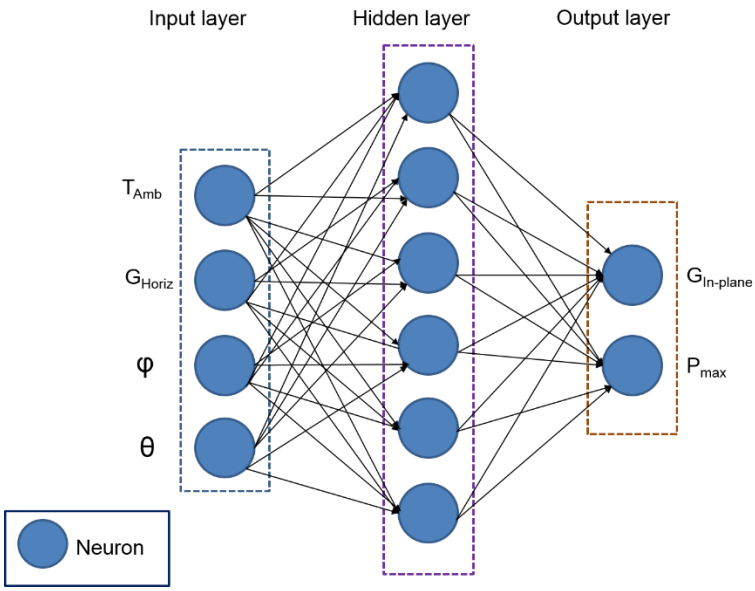


Figure 4.31. Proposed neural network architecture for the prediction of in-plane irradiance and power output. Each arrow (i) represents a connection (also called a synapsis) between two neurons (of neighbouring layers) and corresponds to a specific weight (wi).

The parameters of the training set were chosen based on the main predictor variables used in the conventional models for both in-plane irradiance and power output. These comprise global horizontal irradiance, sun position angles and ambient temperature [119]. Azimuth and zenith angles are calculated using a solar position algorithm [157]. The output layer consists of the target output, which in this case is in-plane irradiance ($G_{In\text{-}plane}$) and maximum power output (P_{max}). The number of hidden neurons was set to six, following the empirical "rules of thumb", according to which, the number of hidden neurons should not exceed the sum of the input and output neurons. Adding more neurons and/or hidden layers might result to "over-fitting", which means that the network has a poor predictive performance for data outside its training range [181]. Furthermore, the error as a function of iterations follows the exponential

decay curve. Thus, for the termination of the training, four significant digits were chosen as the stopping criterion; which gave an average of about 300 iterations for the validation process and 1000 iterations for the back-filling process, where the time required for each iteration is about 0.1 ms. Increasing iterations from that point added computational time without significantly affecting the result, as seen in Figure 4.32 for the validation process. Furthermore, increasing the number of repetitions up to 3000 or the number of hidden neurons up to 10 was found to yield random results which was a sign of over-fitting.

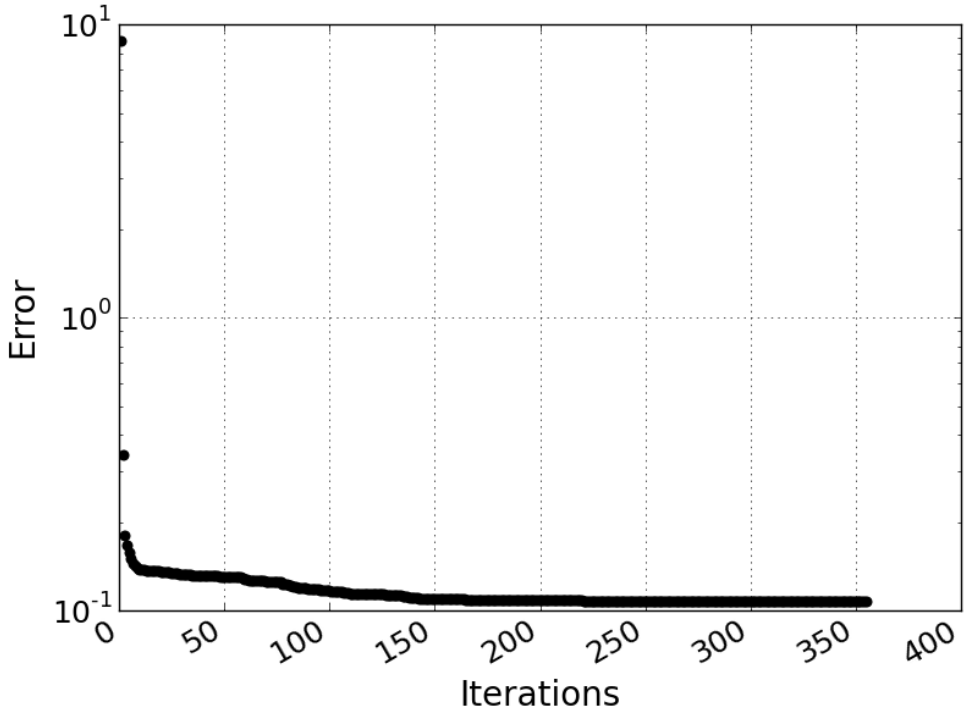


Figure 4.32. Iterations vs Error for the validation process with the applied neural network configuration.

The procedure of training and modelling with NN goes as follows: First, the data set is divided to the training and the testing (validation) set in the same manner as before. After the training is complete, the same network configuration is used to model the output of the testing (or missing) period using the same number and type of input parameters used for training. For the *validation* of the method, global horizontal irradiation (GHI) is taken from the monitoring system whereas for *back-filling* the missing period, interpolated GHI is used just as in case 2. Thus, for the case of back-filling, two different sources of horizontal irradiation

are essentially utilized; measured and interpolated horizontal irradiation. The flowchart depicted in Figure 4.33 describes the procedures for training and validating the NN as well as back-filling for a missing period also demonstrating the data sources utilised in each case.

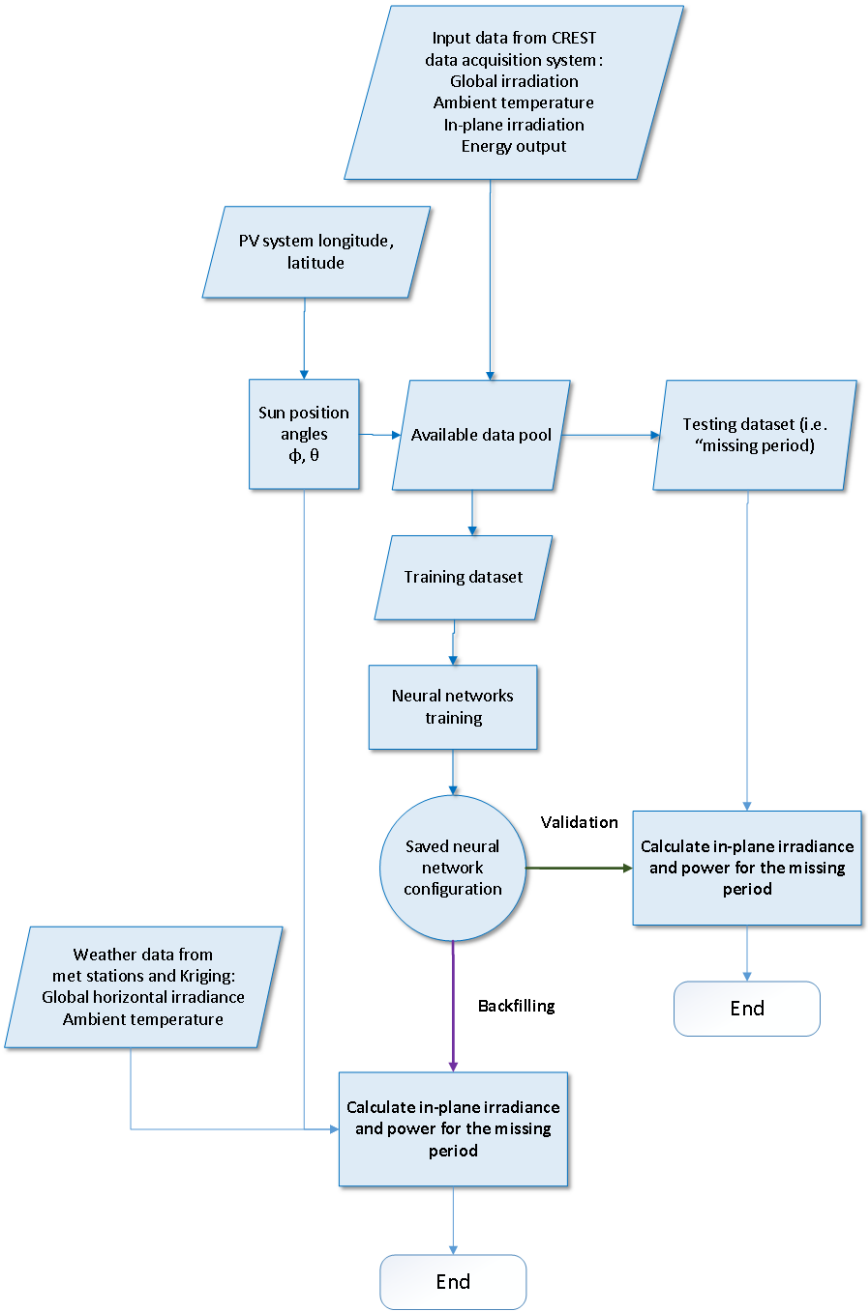


Figure 4.33. Block diagram that shows the training, validation and back-filling procedures and the utilised data sources.

4.8.3 Validation results

The validation results are shown in Figure 4.34 for in-plane irradiance. The result is compared to the two-step method, namely the conventional method comprising a separation [119] and translation algorithm [156],[121].

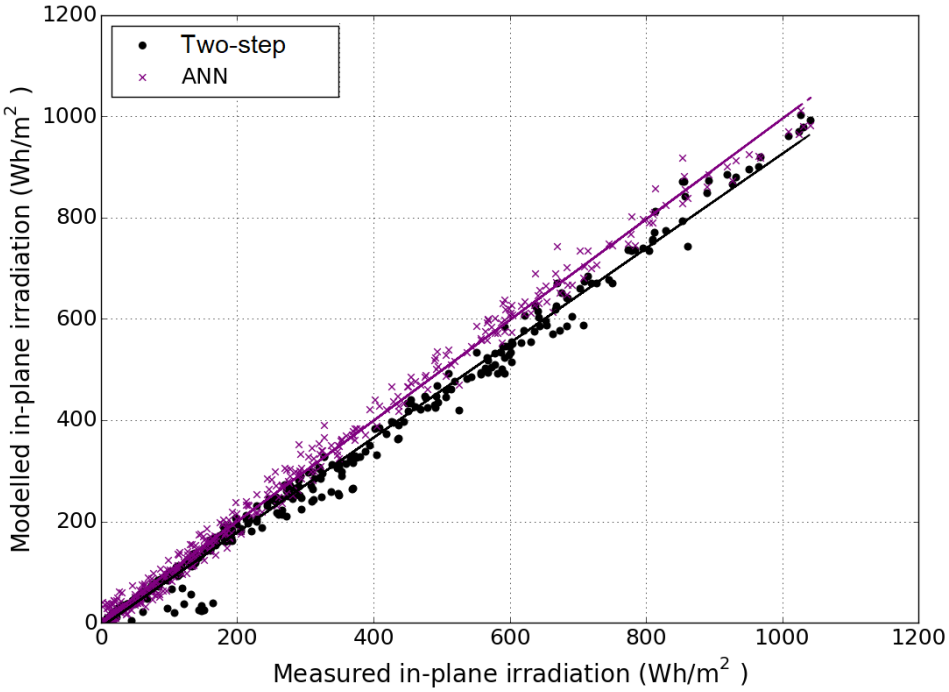


Figure 4.34. Comparison of hourly measured and predicted in-plane irradiance using neural networks (NN) and the two-step method for June 2014.

The hourly %RMSE and %MBE were 13.8 and -9.4 for the two-step method, respectively and 4.5 and -0.34 for the NN method. The two-step method consistently underestimates in-plane irradiation in comparison to the NN method, while also yielding a slightly higher RMSE. The validation results are also shown for the two silicon modules: a mono-crystalline (c-Si)(module A) and a poly-crystalline silicon (pc-Si)(module B). For the c-Si module hourly %RMSE and %MBE is 5.7 and -0.74 and for the pc-Si is 6.7 and -0.81.

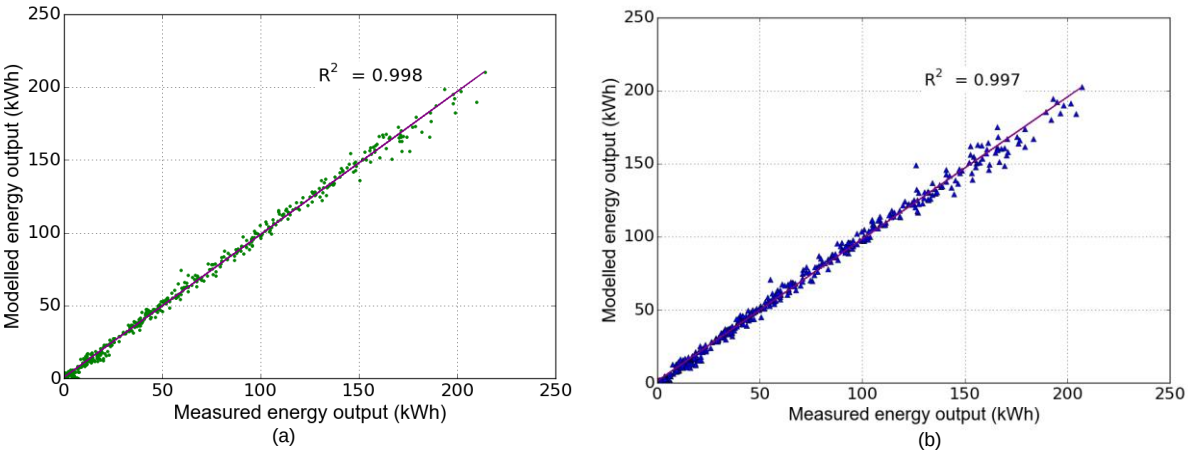


Figure 4.35. Comparison of hourly measured and predicted maximum power output using neural networks (NN) for June 2014 and for (a) a crystalline silicon module and (b) a poly-crystalline silicon module.

Furthermore, the method was also used to back-fill a different period within the year (September) and was found to give similar results with %RMSE of 5.0 and %MBE of -0.2. Additionally, a larger missing period of two consecutive months was back-filled for c-Si, using the same training set size as before. The daily results for energy output are shown in Figure 4.37.

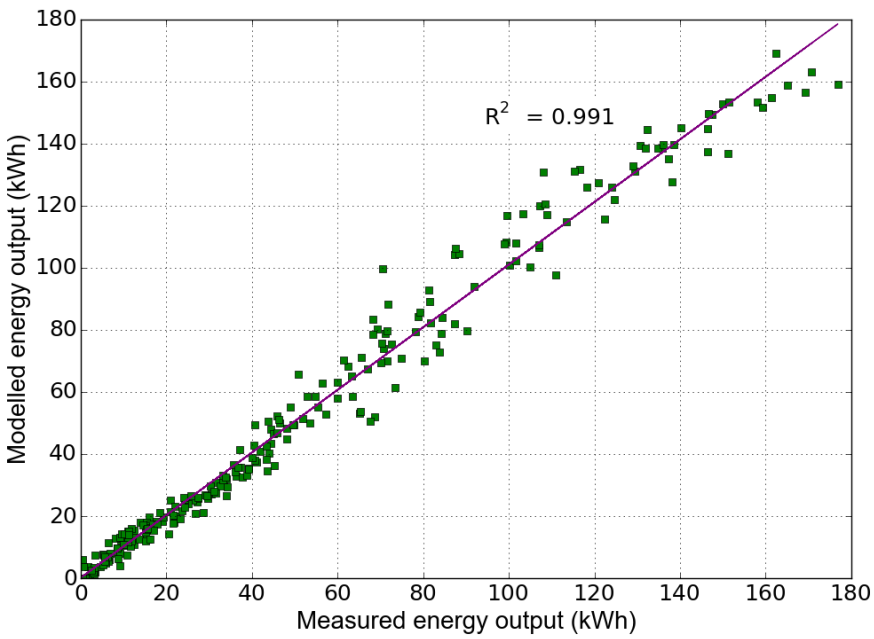


Figure 4.36. Comparison of hourly measured and predicted maximum power output using neural networks (NN) for September 2014 and for a crystalline silicon module (c-Si).

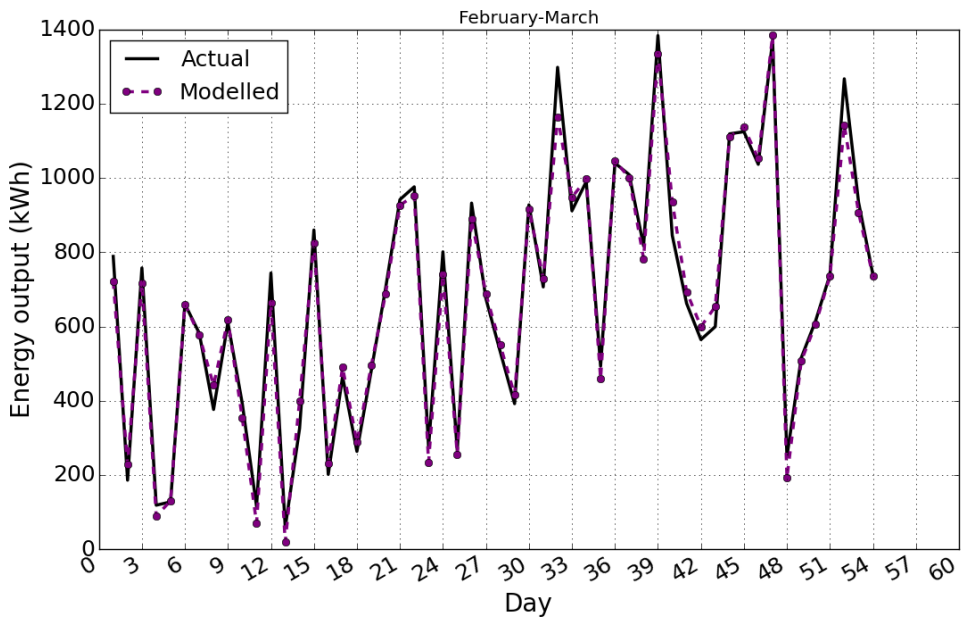


Figure 4.37. Modelled vs measured daily energy output for two consecutive months.

When the missing period increases, as in the case of back-filling two consecutive months (February and March), the same training set size can be used again, but the hourly %RMSE is

slightly increased. This is due to having a significantly larger test set where the state of the system can no longer be captured with the same efficiency in the training set. Increasing the training data pool does not improve the results significantly, while also becoming more impractical in terms of employing larger data sets. However, daily and monthly results show good agreement with measured data with daily and monthly %RMSE of 6.5 and 1.2 respectively.

4.8.4 Back-filling with remote weather data

For this step, the procedure is the same as for the validation process, with the difference that for modelling the missing period (test set), interpolated GHI is used instead of measured GHI (since it is assumed that weather data are also missing for that period). For the following cases, the testing period is the same and June is primarily chosen for the demonstration of the method as the largest testing set in the year. In-plane irradiation and energy output for c-Si are predicted and the results are shown in Figure 4.38 and Figure 4.39, respectively.

Monthly in-plane irradiation yields for the NN method a %MBE of 2.25 whereas for the two-step method it is -5.92. The result for energy output is 2.18 for the %MBE, which yields a difference from the measured monthly output of about 0.67 kWh. Thus, utilizing a neural networks approach decreases this (absolute) bias in energy output from 5.9% to 2.18% for June.

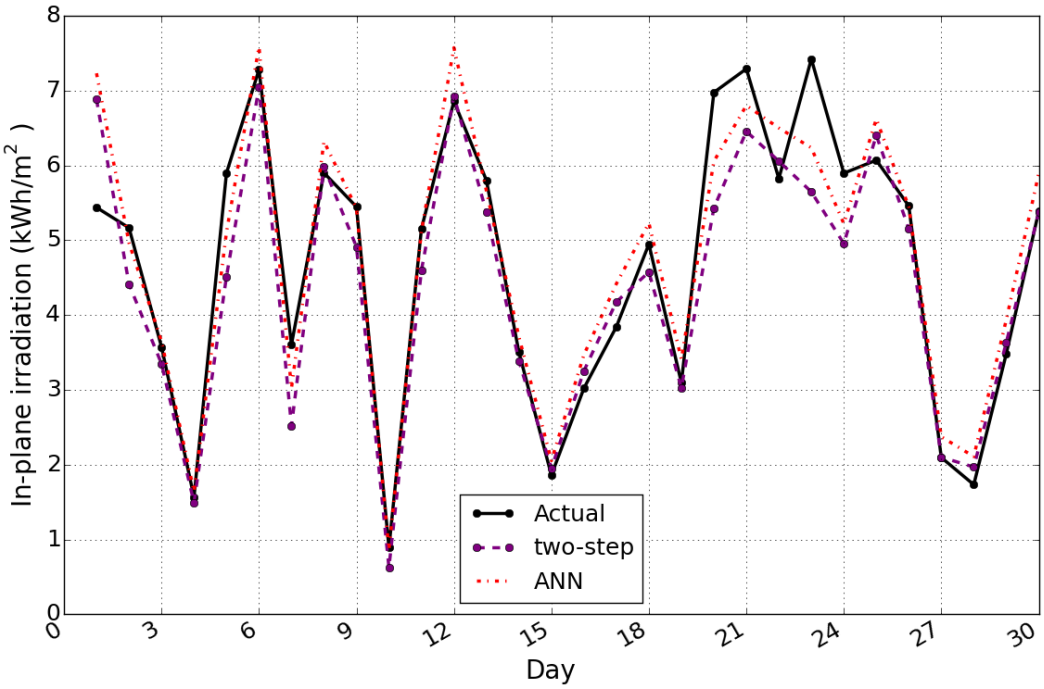


Figure 4.38. Measured vs modelled daily in-plane irradiation with neural networks (NN) and the two-step method (June).

The modelled results follow a similar trend with global horizontal irradiation. Applying the same procedure for July, where monthly interpolated GHI was very close to measured GHI (%MBE = -0.8) yields a %MBE as low as -0.07 for monthly energy output with an absolute difference from real of only 0.02 kWh.

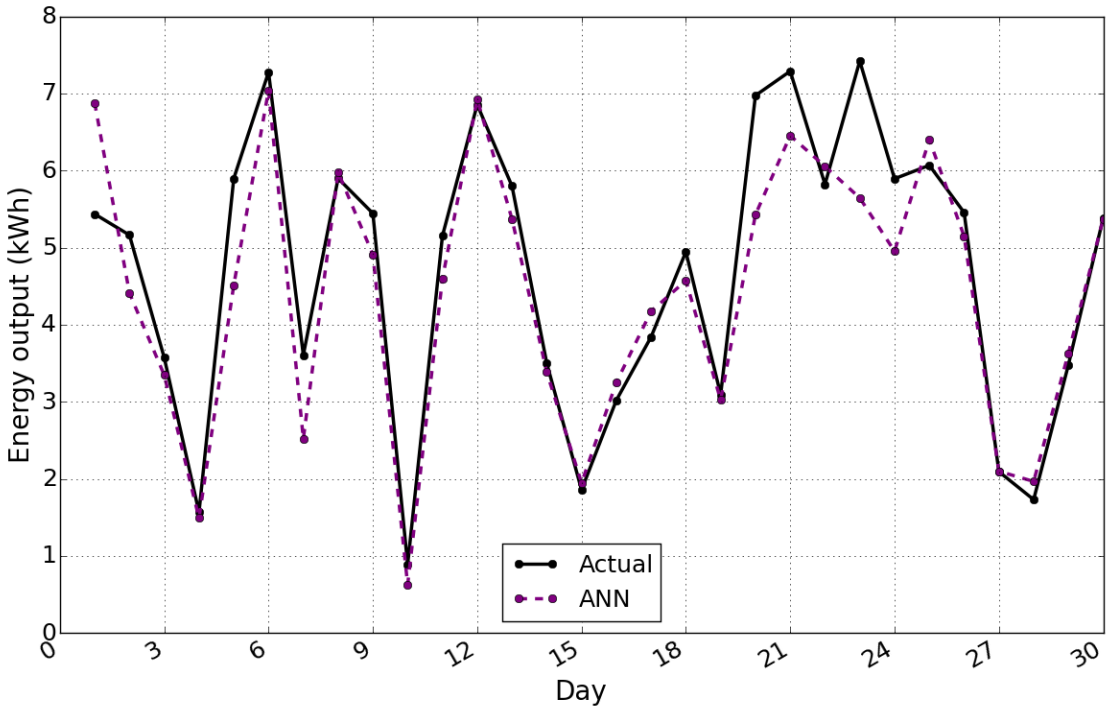


Figure 4.39. Measured vs modelled daily energy output with neural networks (ANN) (June).

Furthermore, in the demonstrated case of June, monthly modelled PR was very close to measured, both yielding a 0.89 for the month with a negligible deviation of ΔPR_{jun} = -0.0006 as with the empirical method . This deviation is even lower than the average daily (actual) PR variation. Again, this result was expected as both back-filled in-plane irradiation and energy output always have a similar bias (as a result from the same training) which is then eliminated in the performance ratio. A similar result is found for the July's example with a PR of 0.87 and a negligible deviation of ΔPR_{jul} = -0.0009. Finally, back-filling results of monthly MBEs are gathered in Table 4.11 for 7 months with high availability (> 90%) in 2014.

Table 4.11. Monthly MBE for energy output and in-plane irradiation for the two different modelling approaches, ANN and two-step.

Month	Back-filled energy output (kWh)	Two-step in-plane irradiation (kWh/m^2)	ANN in-plane irradiation (kWh/m^2)	Median & MAD for energy output (kWh)
3	1.08	-7.73	5.19	
4	1.46	-5.56	5.38	
5	-0.23	-8.43	-0.40	
6	0.67	-5.92	2.24	MED = 0.67 MAD = 0.48
7	-0.02	-8.37	0.03	
8	0.80	-9.16	1.89	
9	0.19	-10.51	0.36	

4.9 Discussion on benefits and potential limitations of back-filling

The motivation for inferring missing data, derives from the fact, that missing information may lead to biased results, especially when estimating the performance ratio of a system. As seen in the third case of back-filling, the annual performance ratio of 'PV system A' decreases by about 7% if there are two missing weeks in one month. The more days are missing, consequently, the more the estimated PR decreases. The applied back-filling method aims to replenish missing or corrupted information on energy output and performance ratio with the highest accuracy possible and from remote weather data if these are not available from the monitoring system. It can be a useful tool as a means of acquiring data that would normally be available from monitoring, therefore can be used in place of monitoring if needed. The main reason for this is that while energy readings are usually available for the total production of a system (AC meter readings or as exported to the grid), if monitoring fails, array or sub-array monitoring data are not available. The proposed method can be applied on array, sub-array or string level (each represents a specific case of data loss) and this is especially

important for enabling fault detection, thus minimizing downtime. On the other hand, if irradiation is not available then the energy output alone is not an indicator of normal operation, so in every case the PR on the DC side will have to be calculated, as one of the main indicators in PV performance assessment. Additionally, in cases of prolonged missing periods, the annual PR is also biased from seasonal effects. Thus it could be lower resulting in serious warranty cases, where even 1% lower PR might be crucial for the annual revenue of the system, and as seen in the case of a domestic PV system, PR is improved when missing days over 10% are back-filled.

The demonstrated methods showed that PR can be predicted with high accuracy in all cases. As further illustrated in Figure 4.40, after 12 days of missing data, the impact on the PR is more evident if the missing period is *not* back-filled, where "no back-fill" implies that the PR is calculated based only on the existing data of energy output and in-plane irradiation. For up to 3 days of missing data the difference is quite small and thus it is possible to still calculate monthly PR considering the monitoring fraction for a given period (ratio of hours of monitoring activity to hours in the given time period) and with a very small bias (<1%). When missing data increases over that threshold the deviation becomes more obvious. While for all the back-filling methods the maximum deviation from the actual PR is as low as 0.007, the maximum deviation for the case of the non-back-filled period is 0.03 which is about 3% of the monthly PR. In all cases the results show that it is worth back-filling the lost data both for estimating the energy output at sub-system level monitoring as well as for a more accurate prediction of the PR at the end of the month, especially when missing data exceeds the threshold of 10%.

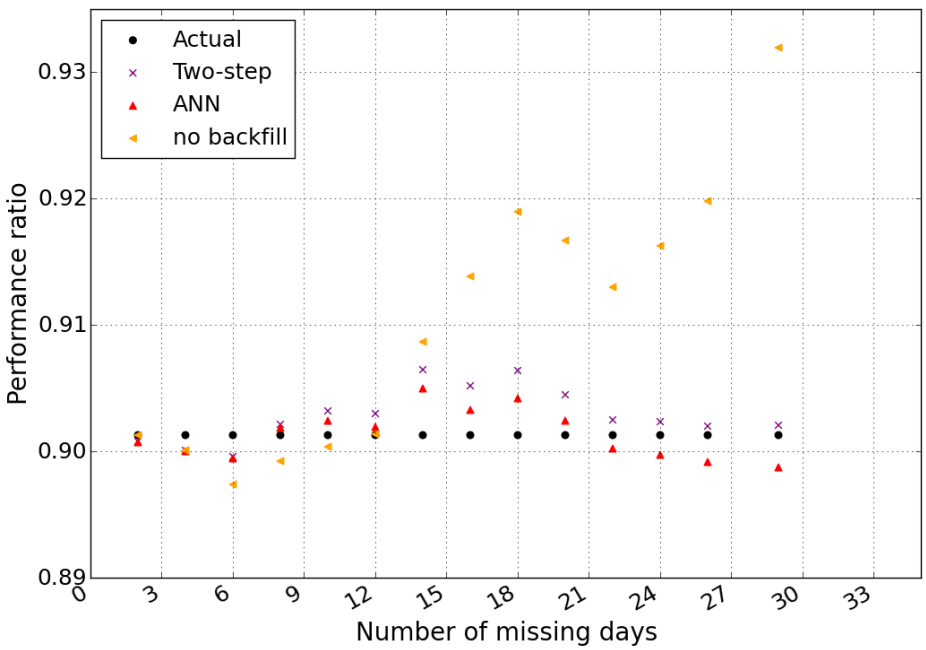


Figure 4.40. Impact of the missing days on the monthly PR with back-filled energy output and without.

However, there are cases where further considerations should be taken into account prior to back-filling. For example, in cases of more than one month missing, hourly random error increases whilst giving satisfactory results for daily and monthly results. Furthermore, in cases of very prolonged monitoring interruption (for example a year) there is increased probability of PV system faults occurring during the missing period. Under these circumstances, the 'system fingerprinting' implicit in the training will no longer capture the system operational and behavioural state and actual energy output variations will not be represented accurately with back-filling. In such contexts, more considerations must be taken into account prior to back-filling. Specifically:

- If the missing period is known to have had no generation, no back-filling should be applied as the real output is apparently zero.
- If known faults have occurred during the missing period which could cause severe under-performance then back-filling will not be as accurate and can be applied only as an indication. This further depends on several factors such as the maintenance scheme followed by the system owners or administrators, the age of the system and its record on failures. A characteristic example would be a domestic PV system

where the average time of detection and repair of a fault is about three months according to [91].

- If days from these months are used for training the model and estimating a period where the fault is fixed then, as expected, the result will not be as reliable. On the other hand, the age of the system and its record on failures indicate the likelihood of a failure occurring. That becomes more relevant when the missing period is increased.

Thus, unless there is sufficient information about the state of a system, the more data that are missing, the less accurate or riskier the back-filling becomes as failures might have taken place in the meantime [91], also taking into account that some components have higher failure rates than others [182]. Also, in cases of more than one month missing, hourly random error increases whilst giving good agreement for daily and monthly results.

Further limitations may arise from the fact that remote meteorological data may not be available or may be of limited accuracy. Using two different sources, one being of ambiguous accuracy may compromise the back-filling procedure. Thus, the overall efficiency of the method depends on a) the quality of the training input data (i.e. the measurements) b) the quality of the interpolated data and c) the quality of the training process (both for the ANN and the two-step method for case 2) which is already discussed in the training procedures. The quality of the measurements depends on the applied cleaning process (i.e. removing abnormal or unphysical values from the dataset). The poorer the quality of input data, the less meaningful become the modelling results. The quality of the interpolated data mainly depends on two factors: the sensor uncertainty of the meteorological stations' equipment and the accuracy of the applied spatial interpolation method. Whichever interpolation technique is applied, validation using measured data prior to applying back-filling is an essential part of the process, as error propagation from interpolated GHI is evident in both energy output and in-plane irradiation results.

4.10 Chapter conclusions

A methodology to replenish missing meteorological and electrical data (not) obtained during PV system operation was developed and validated for three case studies. The methods

are based on training procedures by using interpolated or local meteorological data and energy output of the PV system used at each case. A sufficient training set size was found to be about 40 days in total and by using data from the surrounding days, namely before and after the missing period. The approach is validated against data from a precision measurement system and it showed that energy output can be accurately predicted by using the proposed models and training data pool.

In case 1, where climatic data are available at all times, back-filling yields accurate agreement for both energy output and performance ratio, at all temporal resolutions tested. In case 3, both training and back-filling is applied by employing remote weather data. The results showed that back-filling yields accurate daily and monthly results for both smaller and larger training data pools (in terms of available daylight hours). The negative bias which normally derives from the inference of in-plane irradiation estimation is largely diminished in this case. However, hourly analysis yields a relatively higher RMSE which is propagated from global horizontal irradiation.

In case 2, there are noticeable differences in terms of the absolute energy production, while the estimation of the performance ratio shows excellent agreement for both PV technologies namely c-Si and poly-Si. This means that the key property for assessing system quality can be replenished accurately with the given method which is sufficient for evaluating the performance of a system in the longer term. The PR is the key parameter required for warranty verification, and the method is useful for achieving this purpose. Detailed investigation of the relative underestimation of the energy yield of a system identified that the negative bias is almost exclusively due to the irradiance translation to plane of array. Thus, further efforts focused on the correction of this part by replacing the separation and translation algorithms with neural networks and simple training procedures comprising global horizontal irradiation and sun position angles. The results showed that higher accuracy can be achieved in terms of in-plane irradiation and consequently energy output reducing the (absolute) bias in energy output by an average of about three times. Further work has been done on the impact of missing data on the monthly PR and was found that without back-filling, the monthly PR may be 3% off its actual value, whereas using the proposed back-filling methods the obtained PR lies within 0.8% of its actual value. Therefore, back-filling techniques should be applied for acquiring energy output and a more accurate monthly PR in case of data loss or corruption, as well as estimation of losses due to total blackouts.

Successful back-filling requires that zero generation and failures are excluded as possible occurrences during the missing or the training periods and especially in cases of prolonged interruption of the monitoring system, where the risk increases.

Chapter 5

Remote fault detection framework and limitations due to data quality

5.1 Introduction

It was shown in previous chapters, that poor data quality obscures the PV performance assessment results significantly. Especially in the case of wrong input information and missing data, the obtained annual performance ratio of a PV system may be well off expectations. As also demonstrated, there are ways to mitigate this impact of data quality by applying quality checks and inferring the missing data, as well as correcting wrongly declared system descriptions where possible. While data quality often creates false positive (or negative) alarms on PV systems, actual system quality is the next most worrying part. System quality refers to failures which often occur during a system's operation and which cause it to underperform, such as inverter malfunction, defect modules and string disconnection (see section 2.9 in Chapter 2). In the long term this results in decreased annual energy yield and may also lead to further damage on the affected equipment, if the fault is not detected on time. Consequently, this leads to further loss of energy production and compromises the financial status of the investment.

This chapter analyses the employed methodology for the detection of low performing systems based on the Nottingham City Homes dataset as a test case, whilst developing a failure detection framework, adapted for small-scale systems. It then demonstrates specific case studies whereby data quality significantly obscures system fault detection. The main elements of the failure detection are the modelling framework which comprises the applied models and the detection framework which is based on several applied checks and remote weather monitoring. Thus, this chapter finally focuses around the remaining steps of the performance assessment procedure, as seen in Figure 5.1.

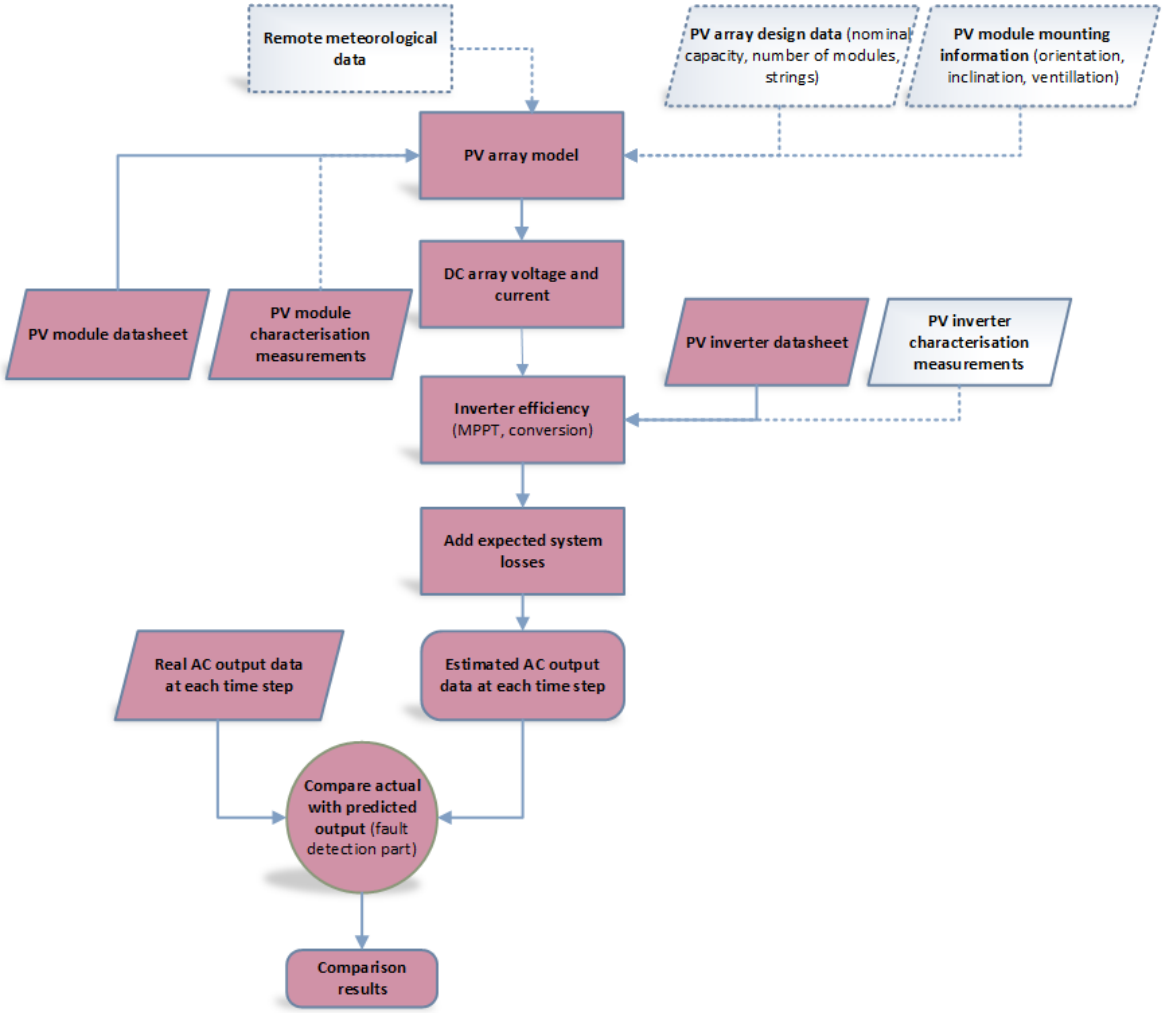


Figure 5.1. Main blocks (highlighted in fuchsia) of the overall performance assessment framework associated with the work described in this chapter.

In the present methodology, it is assumed as the most realistic case, that PV system data are available for a certain period of time which occurred in the (recent) past. Specifically, the analysis is carried out assuming that there is no live data available and the checks take place using a specific date as a starting point. This is further justified by the fact that essential climatic data from meteorological stations, become available on a quarterly basis at the best case. Generally, remote data sources cannot be accessed continuously and therefore, fault detection through data analysis is scheduled according to the availability of the employed weather datasets.

Fault detection methodology is also affected by the data availability in domestic monitoring which is rather scarce and therefore, not all fault detection studies can be applied

in reality. In Table 5.1, the most commonly employed monitoring parameters are presented and examples of the relevant studies for a comparison with the present database, as this determined the proposed fault detection framework.

Table 5.1. Monitored parameters found in failure detection routines in literature and limitations in the present dataset.

Commonly monitored parameters		Present dataset	Example studies
Electrical	Current, voltage at maximum power point, at both AC and DC side	Total energy output	[138],[137],[139],[135], [136],[133],[183],[184]
Meteorological	On-site in-plane irradiance and module temperature	Not available. It is remotely inferred	
System information (Metadata)	Overall capacity	Available but may be wrongly declared	
	Single module capacity	Not always known	
	Location, azimuth, inclination	Available but may be wrongly declared	
	Module type	Not always known	
	System topology	Two basic configurations are employed.	

For the particular dataset, the most related studies are [104], [4] and [5] in terms of their application on domestic PV sector and the employed parameters. In [104] the performance of a sample of domestic systems is analysed based on monitoring parameters on site, as part of the UK field trials programme. The detection of faults in this case is applied based on the classification of different types of observed losses and long term data analysis. Part of this methodology is also employed here based on remotely inferred irradiance and temperature. The method is further complemented with elements from [4] where remote weather

monitoring is also employed. The main difficulty here is the automatic discrimination of "normal" from faulty systems with regards *both* data and system quality. This becomes especially challenging due to the increased amount of PV systems which need to be tested and the limited or erroneous input information. In [5] a remote failure procedure is applied based on long term data analysis of neighbouring PV systems where "performance-to-peers" parameter is the basic detection metric. This is calculated based on comparing a PV system's performance with its "peers", namely its neighbours. This parameter is also adapted here in order to categorise neighbouring systems based on specified criteria. However, due to several difficulties concerning data quality, this metric is not used as the only performance indicator. Finally, the types of fingerprints (flags) that can be imposed using the failure detection framework are presented whereby insight is provided on the difficulties arising due to ambiguous or erroneous system descriptions and their implications on the long term performance and financial revenue of these systems.

5.2 The modelling framework

The modelling framework includes the system description model, namely the PV and the inverter model based on the employed PV module types and inverters (if these are available). This includes several steps and sub-procedures such as the extraction of the coefficients for the electrical model, and the inverter model which are summarised in the block diagram depicted in Figure 5.2.

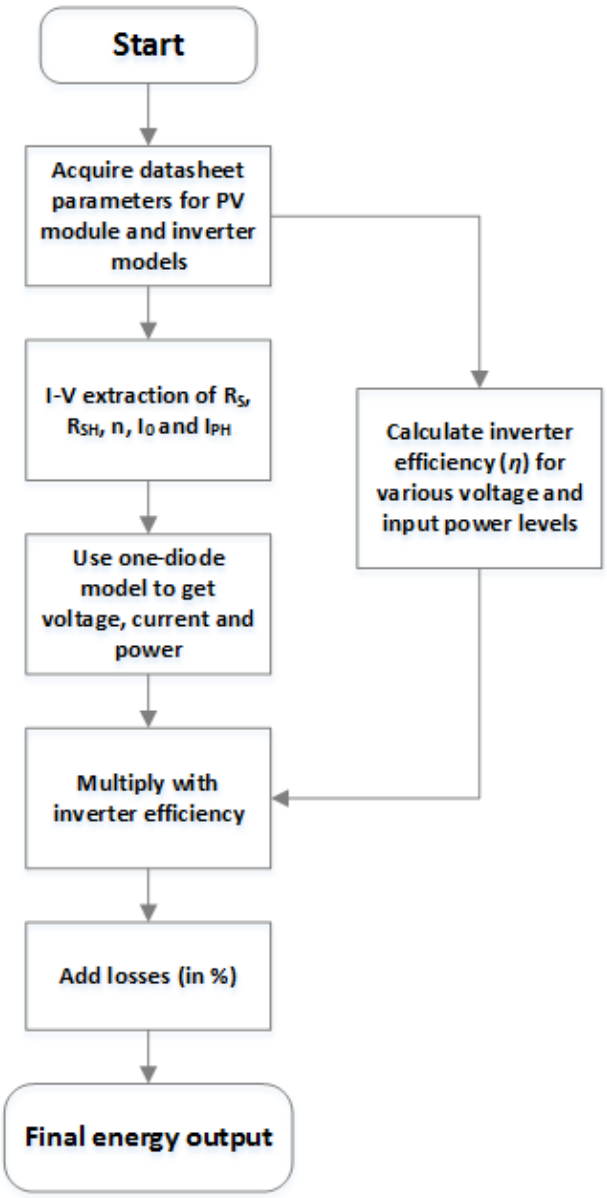


Figure 5.2. Modelling blocks for the calculation of the theoretical energy output.

The chosen electrical model is the one-diode model which is extended to describe a small PV array (< 4kW$_P$). As described in 2.2.1.1 the I-V extraction is a complex procedure which can be accomplished either by using measured I-V curves or by using manufacturer's data if experimental curves are not available. The reason for choosing the particular physical model is to estimate both current and voltage having extracted the required coefficients and not just the maximum power point. This is also necessary in order to use voltage as an input parameter to the inverter model. Choosing one-diode over the two-diode model was done for simplicity, given also that the modelled module types are polycrystalline silicon, which is

described well by the one-diode model. The array output is calculated based on module output and the string configuration of the PV array. In this dataset two simple configurations are met, either multiple strings in series or multiple strings in parallel. The most common configuration however, is the first one for the 98% of the cases. This is also the only case taken into account in this study for the reason that there is no indication for the number of parallel strings employed. Single string output current and voltage (for the maximum power point) are simply calculated based on Kirchhoff's laws (see 2.2).

Manufacturer datasheets for both the PV modules and the inverters can be found online either directly from the manufacturer's websites or by using widely available and updated databases of PV modules and inverters such as those provided by CEC [185], Photon Magazine [186], and PVSyst [187]. The I-V extraction and inverter efficiency models are described separately in the following sections.

5.2.1 Diode modelling and I-V parameter extraction

As seen already in 2.2.1.1, there are essentially two types of methodologies that can be applied for the extraction of the five parameters for the one-diode model. The analytical methods which employ the characteristic points on the I-V curve and the numerical methods which exploit the entire I-V curve. Both methods are applicable here, as these comprise independent blocks of the overall framework. In the cases where a specific module type exists in the lab, then measured curves can be used, whereas if it doesn't, then manufacturer data are used instead. There is no specific reason whether one method is better than the other, as both show equally good results. However, the meaning of the results in the case of using manufacturer datasheets depends on the quality of the supplied data. Conversely, when using experimental I-V curves, at least the quality of the measurements is a known and controllable factor. However, in most cases, experimental curves are not available and thus, using manufacturer data is a commonly employed solution, also in commercial software such as in PVSyst [187].

5.2.1.1 *Using experimental I-V curves*

The developed I-V extraction method is based on the Levenberg-Marquardt optimisation algorithm and a fitting area criterion which minimises the area between the simulated and the experimental (real) I-V curve, represented by the shaded area in Figure 5.3.

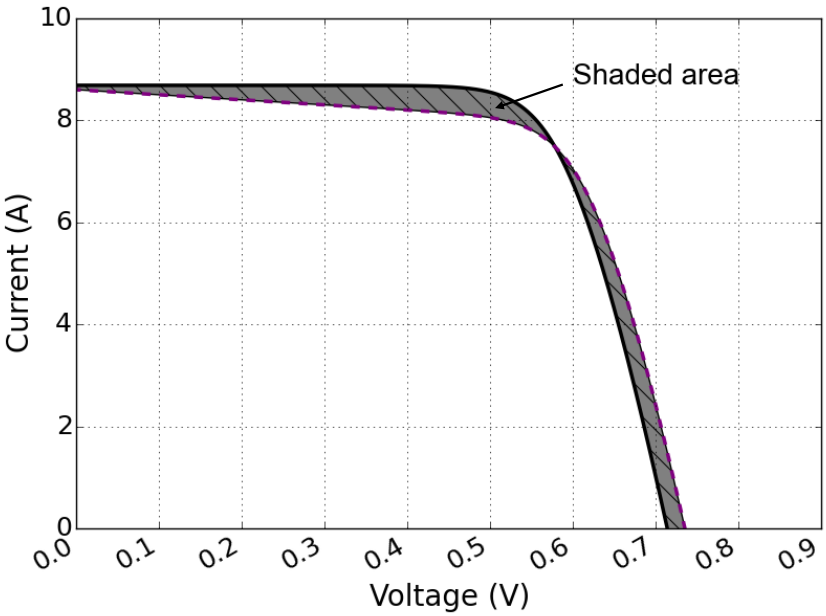


Figure 5.3. Simulated I-V curves with slightly different diode current and shunt resistances. The error criterion aims at minimising the area difference (shaded area) between the two curves.

Initially, a set of guess parameters is given, which remains the same for every type of PV device that has been tested so far. This is sufficient because the employed method does not heavily depend on the initial parameters. In the second iteration these initial parameters are immediately replaced with the new ones. Within every iteration, all parameters are adjusted simultaneously based on the Levenberg-Marquardt optimisation mechanism. In case of runtime errors -namely, algorithm failure to converge after a certain number of iterations - several control mechanisms are activated. These essentially adjust the series and shunt resistances and re-set the iteration procedure with these new guess parameters. The procedure ends when the area criterion (see 2.2.1.1) is lower than a specified threshold, here $\varepsilon = 10^{-5}$, where the two areas (defined by the simulated and the real curves respectively) are calculated numerically based on the trapezoidal rule (see Appendix).

The aim is to extract the five modelling parameters required for the one-diode model, so that these parameters are then used to model the voltage and current at maximum power point and at different conditions of irradiance and temperature. The accuracy of the algorithm is demonstrated in Figure 5.4 by fitting an I-V curve using STC indoor measurements of four 6-cell c-Si mini-modules. Good agreement between the simulated results in comparison to experimental data is obtained, as shown in Table 5.2, for the maximum power point, voltage and power.

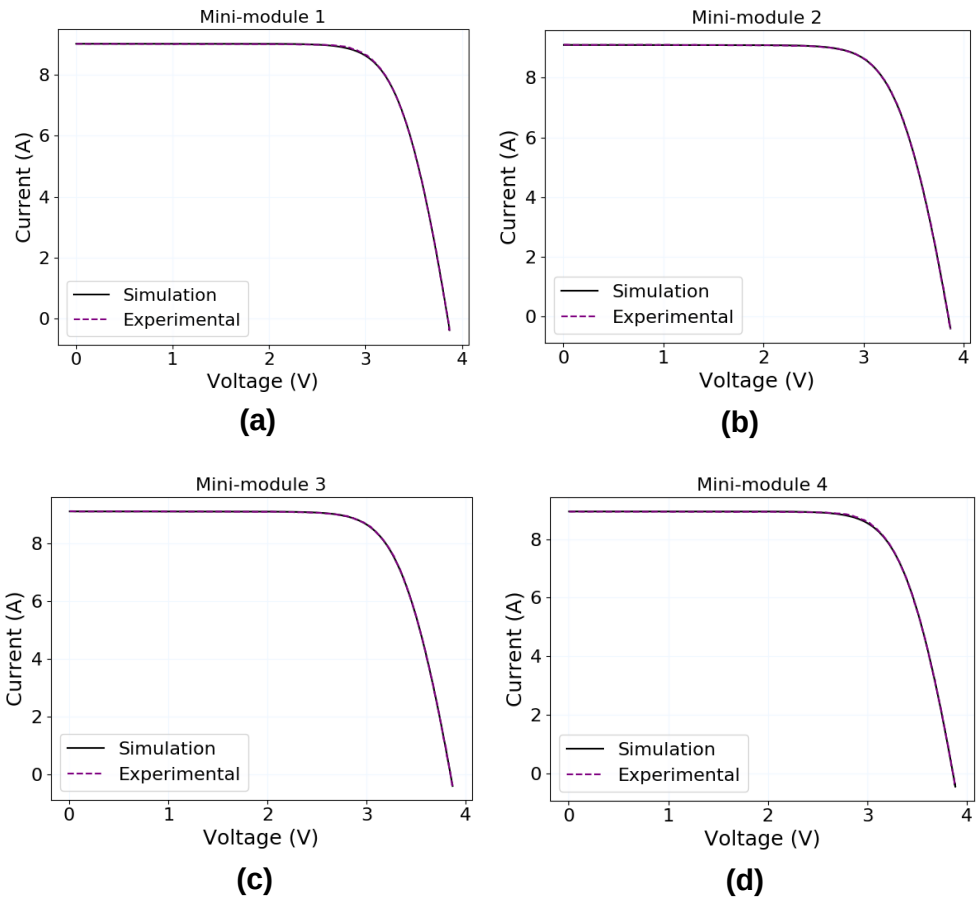


Figure 5.4. I-V curve fitting for four (a)-(d) 6-cell c-Si mini-modules at STC (G =1000 W/m^2, T = 25 ^{0}C).

Table 5.2. Comparison of measurements and simulation results at maximum power point.

	V_{MPP} (V)			P_{MPP} (W)		
	Measurement	Simulation	Error (%)	Measurement	Simulation	Error (%)
Module 1	3.07	3.07	0.00	26.11	26.02	-0.35
Module 2	3.04	3.06	0.64	26.06	25.98	-0.29
Module 3	3.08	3.05	-0.79	25.87	25.81	-0.23
Module 4	3.02	3.07	1.61	25.84	25.73	-0.44

5.2.1.2 *Using manufacturer's datasheets*

In order to model the theoretical DC output of a PV system, exploiting datasheet information for modelling was found to be the best way amongst the available options. An alternative way would be to extract module information from widely available databases like the SANDIA or the CEC module database. However, one of the difficulties in that was that some of the available modules' nominal characteristics did not exactly match the ones found in the datasheets for the same module models. Furthermore, the extracted information was based on specific parameter extraction methods, which are not described. Therefore an algorithm based on [188] for the parameter extraction for PV modules was developed. The analytical expressions used for this are based on the characteristic points supplied by the manufacturer for the Standard Testing Conditions (STC) as shown in Table 5.3. The employed equations as well as the simulation procedure are described in more detail in the Appendix.

Table 5.3. Datasheet parameters used in modelling and example values

Datasheet Parameter (Symbol) [Unit]	**Example values**
Nominal Output (P_{MP}) [W_P]	245.0
Voltage at maximum power (V_{MP}) [V]	30.2
Current at maximum power (I_{MP}) [A]	8.13
Open Circuit Voltage (V_{OC}) [V]	37.5
Short Circuit Current (I_{SC}) [A]	8.68
Temperature Coefficient of V_{OC} [%/°C]	-0.32
Temperature Coefficient of I_{SC} [%/°C]	0.047

The method has been further optimised by iterating the same steps for different values of ideality factor, whereas in [188] this is considered to be constant for simplicity. The accuracy of the method is finally tested by minimising the following expression, based on the characteristic points on the I-V curve, open circuit voltage and voltage, current and power at maximum power point:

$$\varepsilon = \sqrt{\left(\frac{\mathrm{V_{MP}} - V_{MP\,sim}}{\mathrm{V_{MP}}}\right)^2 + \left(\frac{\mathrm{V_{OC}} - V_{OC\,sim}}{\mathrm{V_{OC}}}\right)^2 + \left(\frac{\mathrm{P_{MP}} - P_{MP\,sim}}{\mathrm{P_{MP}}}\right)^2 + \left(\frac{\mathrm{I_{MP}} - I_{MP\,sim}}{\mathrm{I_{MP}}}\right)^2} \qquad (5.1)$$

Where the subscript sim denotes the corresponding simulated parameter. Here $\mathrm{I_{SC}}$ remains the same during each repetition thus it is not included in (5.1). The results are extracted from the STC and then tested against experimental (digitised) curves which are obtained from the same manufacturer datasheet but for different illumination conditions. The fitting results are demonstrated in Figure 5.5 and Figure 5.6.

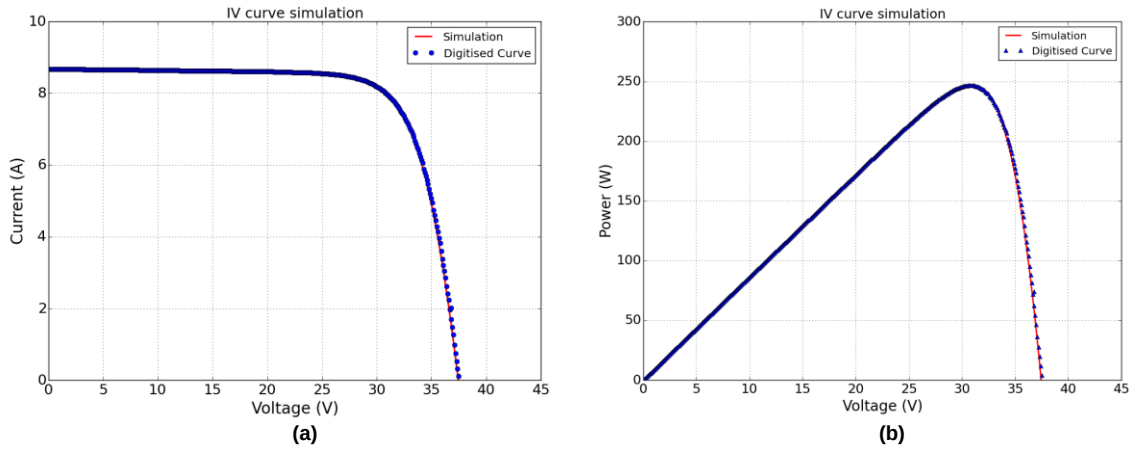


Figure 5.5. Simulated and digitised (a) current-voltage and (b) power-voltage curves for a commercial PV module.

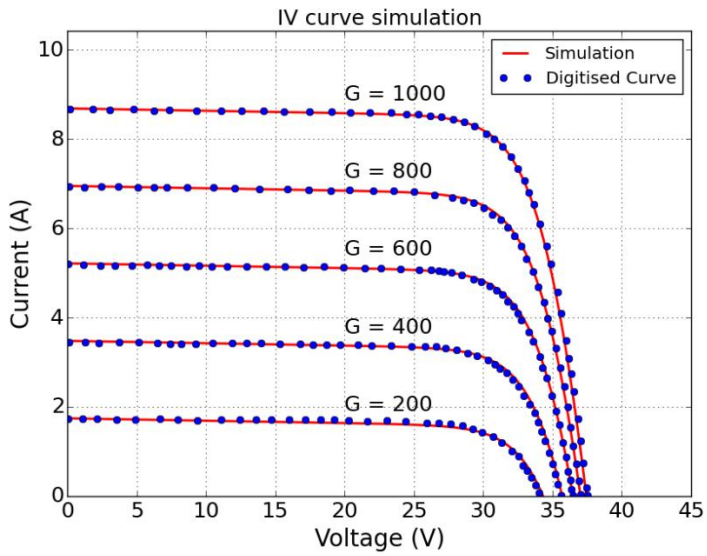


Figure 5.6. Simulated and digitised curves using the extracted parameters for a commercial PV module for different values of irradiance (G).

The extracted modelling parameters for the example commercial module is given in Table 5.4.

Table 5.4. Extracted modelling parameters for the one-diode model for a commercial module.

Modelling parameter (Symbol) [Unit]	Modelling parameters
Shunt resistance (R_S) [Ω]	190.2
Series resistance (R_{SH}) [Ω]	0.177
Diode ideality factor (n)	1.21
Photocurrent (I_{PH}) [A]	8.688
Diode saturation current (I_0) [A]	$1.34 \cdot 10^{-8}$

Finally, the comparison between experimental and simulated data points are given in Table 5.5 for the irradiation conditions shown in Figure 5.6.

Table 5.5. Comparison of measurements and simulation results at maximum power point for different levels of irradiation and constant module temperature (T=298 K).

G(W/m²)	**V_{MPP} (V)**			**P_{MPP} (W)**		
	Measurement	Simulation	Error (%)	Measurement	Simulation	Error (%)
200	28.75	28.55	-0.70	43.04	42.80	-0.56
400	29.57	29.57	0.00	92.6	92.9	0.30
600	29.89	30.47	1.96	143.3	144.1	0.60
800	30.02	30.65	2.10	193.9	195.8	0.95
1000	30.7	30.8	0.33	246.9	247.1	0.05

5.2.2 Inverter efficiency

To calculate inverter efficiency, an interpolation formula is adopted from [131] which was further developed in Python. This is based on interpolating the matrix values between different voltage limits and input power. In modelling, it is also taken into account that each inverter has a “switch on power” which is usually 50 W. Instead of separately interpolating for different voltage and input power values as suggested in [131], the desired efficiency values are found by interpolating in the two dimensions simultaneously (see Figure 5.7) using bilinear interpolation instead. To calculate this, the parameters were extracted from the

different manufacturer data sheets and were gathered to a common database for fast access during simulations. The interpolation function which yields a 3-D surface is only calculated once during the simulations and is used to infer efficiency for the intermediate input values.

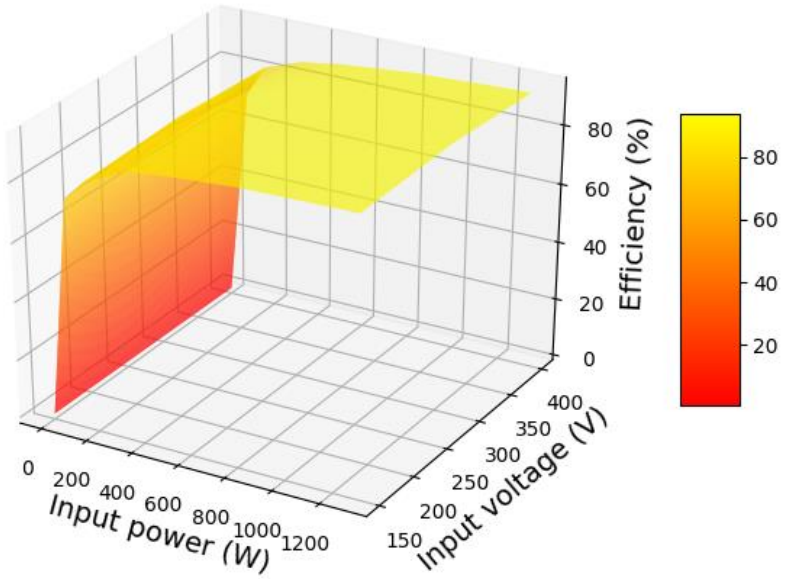


Figure 5.7. Interpolated surface of efficiency *vs* input voltage *vs* input power for a commercial inverter.

Where the efficiency curves are not available, then the Euro efficiency is used instead, which is also provided in the datasheet. In terms of fault detection, an obvious reason for underperforming systems is the inverter input mismatches for different PV systems. Namely, each inverter has its own specifications including minimum and maximum input voltage and maximum input current and power, provided by manufacturer datasheets. As also seen in Figure 5.7, inverter efficiency decreases at low irradiation and this is often mitigated by under-sizing the inverter's nominal rated power with regards to the array nominal capacity as given by:

$$sizing\_ratio = \frac{P_{DC,STC}}{P_{inv,rated}} \tag{5.2}$$

This is because the PV array generates a high proportion of its energy at low irradiances, and in that case an inverter with high rated power will more often operate below the "knee"

of its efficiency surface, as opposed to using a smaller inverter. Thus, it is common practise in countries with medium solar resource to undersize the inverters by a percentage which also depends on the azimuth of the system. Maximum annual yield can be achieved for sizing ratios between 1.1 and 1.4, in the UK [189]. Thus, for every PV system an initial check is carried out, in order to determine whether a system is over- or under-sized as over-sized inverters might also imply erroneous inverter description.

5.3 Failure detection framework with minimum input information

The framework applied in this work is suitable for small-scale installations where remote weather monitoring and averaged power or energy output are the only available data . These data are supplied by the energy meters, thus there is no information on the output of the PV array, namely prior to DC-AC conversion. For this reason, the failure detection is constrained to comparing total actual to theoretical output of the systems. Since, also irradiation and temperature are available at hourly formats, the power output readings are further aggregated into hourly time series (see Appendix for details on averaging). To account for the uncertainty deriving from the lack of on-site weather data various test domains are taken into account. Several examples as well as the employed indicators are shown for each domain for selected PV systems, where also the impact of data quality is demonstrated. The systems which are used for the application of failure detection are chosen based on the results of the applied data quality checks (see Chapter 3). Namely, PV systems with missing or ambiguous technical description (where module models are not known) cannot be taken into account. For the demonstration of the fault detection method, the systems were selected out of approximately 100 PV systems where both inverters and panels are known.

5.3.1 Hourly patterns

In this domain, modelled energy output is compared to actual energy output for every hour of the day. The checks take place on a daily basis and the results are evaluated with regards to defined upper and lower thresholds. The determination of the appropriate modelling

thresholds is also the most challenging part in this domain since it significantly affects the efficiency of the overall method. These thresholds are primarily affected by the in-plane irradiation modelling uncertainty and especially at lower intensities (<100W/m^2). Consequently, the modelled energy output also presents a random and bias error which derives from it (as seen in 4.7.3). This error can be predicted and taken into account, to some extent, based on actual measurements of in-plane irradiation for ideally the same location. Since this is not possible, however, another location (e.g. Loughborough) is used instead, which is similar in terms of climate and modelling requirements.

As already discussed in Chapter 4 the random error in the hourly analysis derives from the lack of on-site irradiance measurements and the application of Kriging interpolation. Also, the efficiency of the applied spatial interpolation mainly depends on the density of the meteorological stations around the point of estimation [168]. Both Loughborough (latitude: 52.8, longitude: -1.2) and Nottingham (latitude: 53.0, longitude: -1.15), as they are in close geographical proximity, present the same density in terms of meteorological stations in the surrounding areas. Therefore, spatial interpolation is expected to yield similar results for the two locations and thus the analysis carried out for Loughborough can be used to estimate the error in irradiation and energy output prediction in Nottingham. In terms of the translation of global horizontal irradiation to in-plane, the same models are applied for both locations. For the determination of prediction error in in-plane irradiance different irradiance bins are considered and normalised mean absolute error is used for each, based on the interpolated in-plane irradiation. The result is shown in Figure 5.8 and can be described by a logarithmic fit.

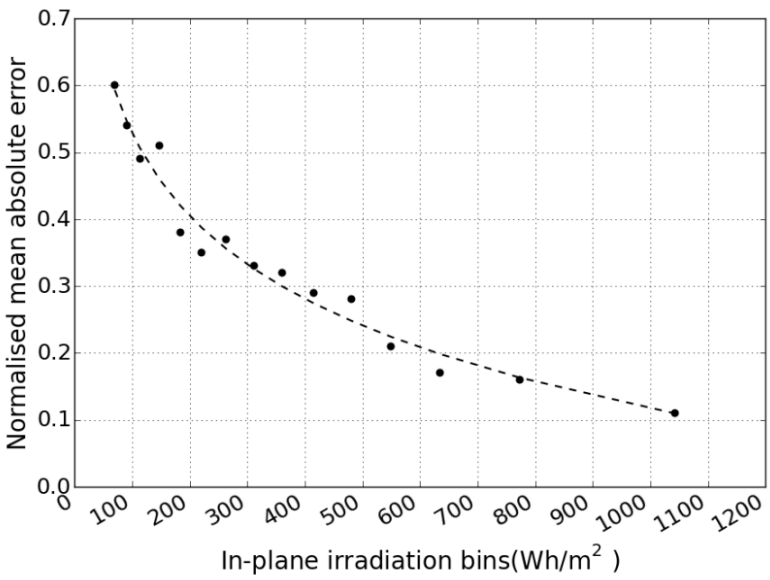


Figure 5.8. Normalised mean absolute error for different irradiation intensity bins.

The normalised mean absolute error describes the overall difference between the forecast and the actual value and includes both bias and random errors. To define the upper and lower thresholds of in-plane irradiation and hence modelled energy output the error is taken by using the normalised mean absolute error (*nMAE*) from the corresponding irradiation bin, such that:

$$G_{in-plane} - nMAE \cdot G_{in-plane} < G_{in-plane} < G_{in-plane} + nMAE \cdot G_{in-plane} \tag{5.3}$$

This yields a range of possible values between the two thresholds for PV system energy output. For a normal PV system the majority of points are within those thresholds. To further account for possible operational losses such as wiring, mismatch, soiling, maximum power point tracking, module ageing as well as possible power deviations from the nominal value the lower threshold is further reduced by 9% [4],[87],[88]. However, random occurrences of lower and higher values than the specified lower and upper thresholds respectively, can be obtained for normally operating systems. To further discriminate between the possibility of a system presenting a fault and a random occurrence, the checks are repeated over time and also checked with neighbouring systems. Furthermore, the threshold for the maximum

number of hours which are found to yield *lower than expected* output is set to 3 occurrences. This threshold was found to be sufficient for the majority of the detected cases in the particular dataset and especially for days with a small number of daylight hours, but a different threshold may be applied to other datasets. For cases where results are higher than the threshold, if the occurrences are more than 50% of the daylight hours, then this indicates that the system has possibly a wrongly declared nominal capacity. Extreme cases of this occurrence, such as double overall capacity than declared, are picked up during the data quality checks based on annual figures. Smaller deviations however are better detectable in the hourly domain. The employed indicators are given in Table 5.6, and descriptions or further actions are given for each case. A PV system with normal behaviour is presented in Figure 5.9 for a clear day. For hours with high irradiation the modelled output is well within the defined thresholds, describing the behaviour of the system more accurately. At lower intensities, the actual power output is slightly higher than the upper threshold and this is anticipated for irradiation lower than 50 Wh, where modelled losses are slightly overestimated.

Table 5.6. Table of indicators used in the failure detection based on hourly checks.

Indicators	Description	Outcome/Action
0	Energy output within thresholds	Normal
1	Over upper threshold but within 20% higher	Possible modelling underestimation
-1	Lower than lower threshold	If occurrences >3, potential fault/check the hourly distribution of flags for a given time period
2	Over upper threshold and more than 20% higher	If occurrences >50%, check input information
(-1,1)	Lower (or higher) in the morning and higher (or lower) in the afternoon	Erroneously declared azimuth/correction of installation angle (as seen in 3.5)

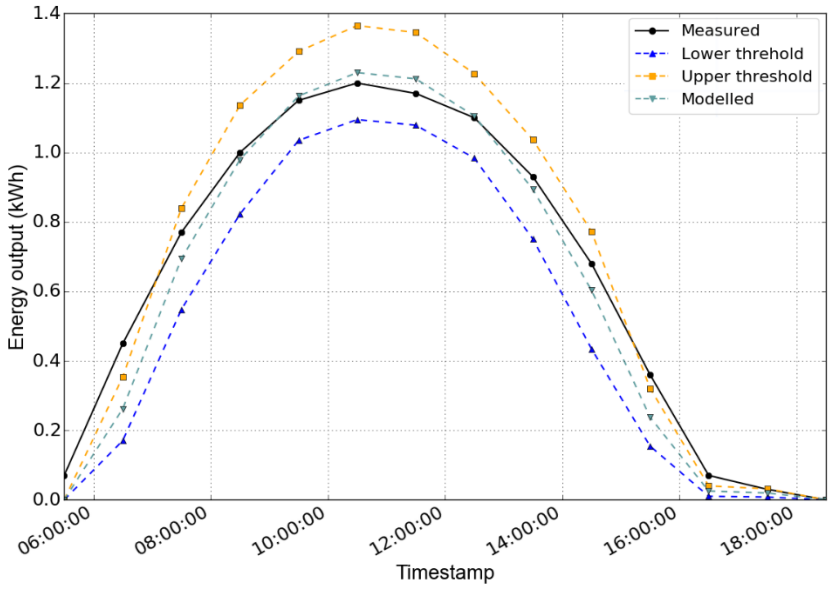


Figure 5.9. Hourly pattern for a normally operating PV system (annual (corrected) PR = 0.79).

The example system given in Figure 5.10 is one case out of the 100 PV systems that were found with slightly wrongly declared azimuth (extreme cases are found where a different angle denotation system is used, as seen in 3.5) based on the hourly flags on a clear day. The flag sequence shown in Figure 5.10 (b) is characteristic where flags '-1' and '1' appear at lower sun position angles and it is readily detectable for high clearness index hours. For the same system, however, looking at a day with much lower daily irradiation (approximately 2450Wh), there is no flag indicating lower output than expected.

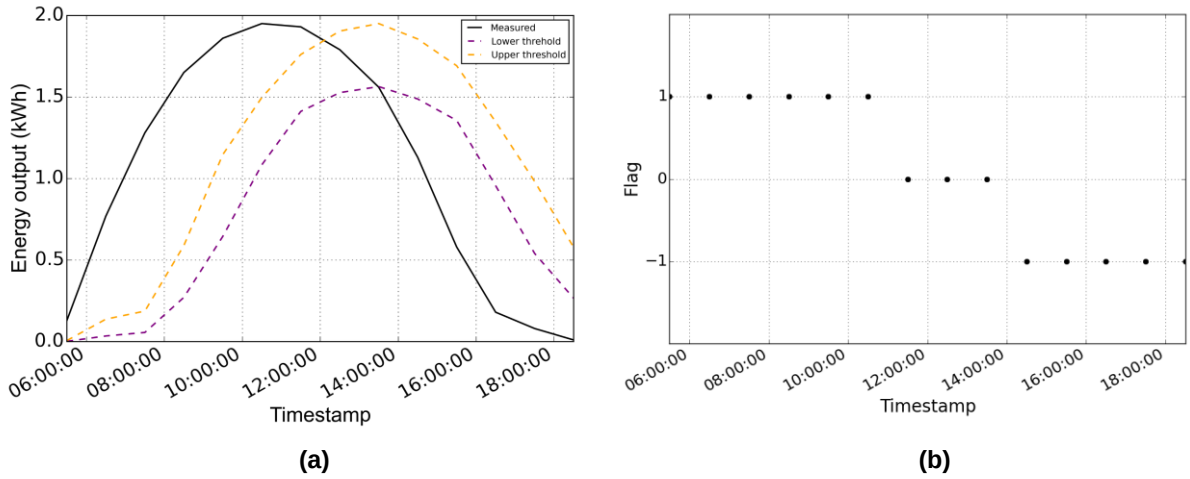


Figure 5.10. (a) Hourly energy output of a PV system where lower and upper thresholds are shown and (b) the resulting indicators (flags) for this pattern on a clear day.

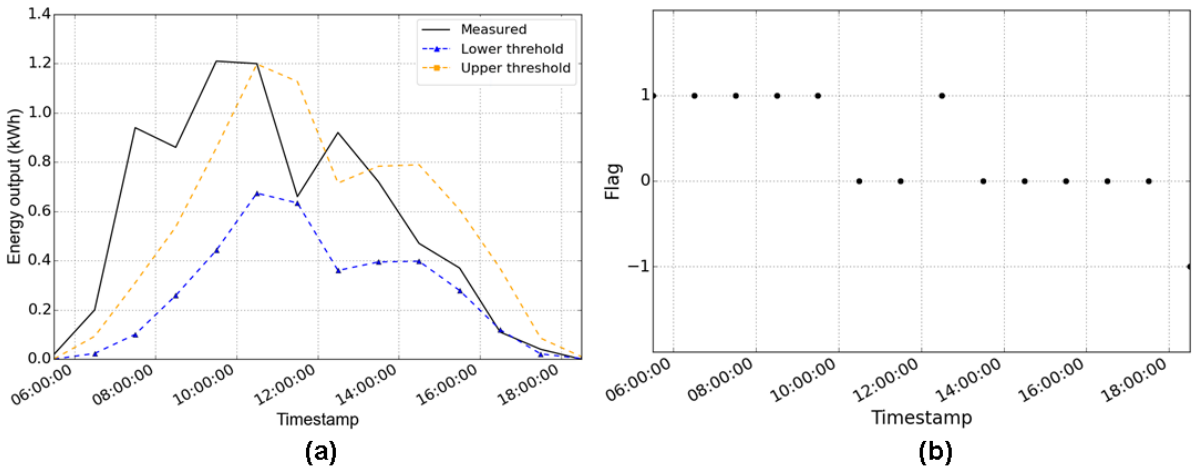


Figure 5.11. (a) Hourly energy output of a PV system where lower and upper thresholds are shown and (b) the resulting indicators (flags) for this pattern on a partly cloudy day.

In both cases, it is apparent that the particular system does not present a "normal" behaviour. However, the difference between the two days is that in the first case, a single check and visual representation are enough to indicate the issue whereas in the second case this is not particularly clear as it could also indicate wrongly declared capacity. In the second case, further checks are applied in the over-time domain, in the search of further evidence for the identification of error (in this case it is not a system fault). It was found that this particular pattern becomes obvious for days with a daily clearness index over 0.4. An example

of an under declared system is given in Figure 5.12. This particular PV system was detected due to its considerable deviation from the modelled upper threshold.

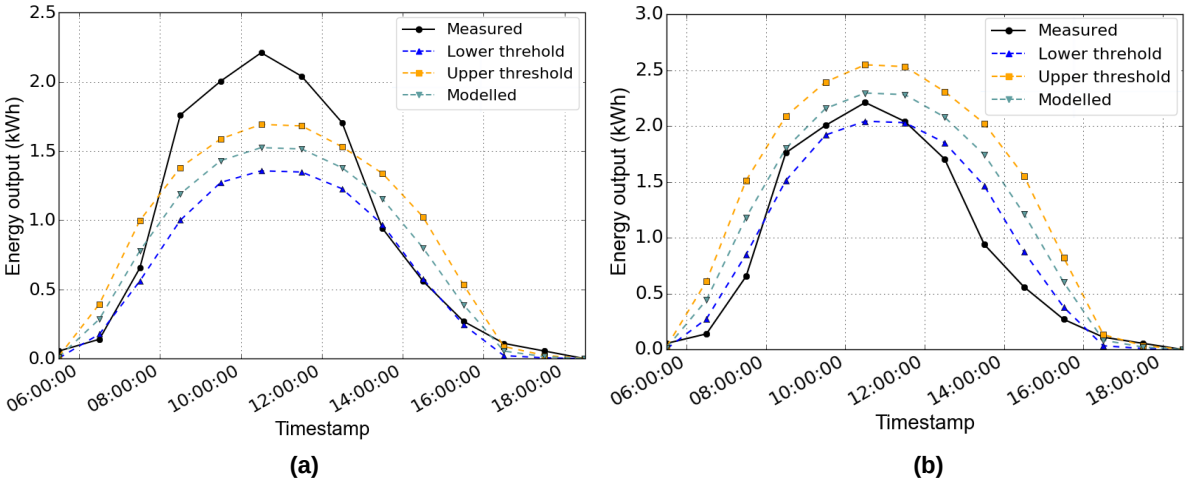


Figure 5.12. Hourly energy output of a PV system where (a) nominal capacity is higher than expected and (b) the nominal capacity is replaced with a likely value, taken from a neighbouring system.

The nominal capacity of this system is initially given as 1.88 kW_P but as also seen from Figure 5.12(a) it is apparent that the system's capacity is higher than declared. Looking at neighbouring systems with the same characteristics, a neighbouring PV system with nominal capacity of 2.82 kW_P was found. Assuming this value is a possible match and re-applying the procedure, it is found that the PV system not only is wrongly declared but could also present increased losses at low sun position angles and especially in the afternoon, as seen in Figure 5.12 (b). The wrongly declared capacity is also confirmed by noticing that the applied inverter in this case is larger than that applied for a smaller PV system (in this case 1.88 kW_P). This means that if this underestimation of nominal capacity is not detected, for example by only checking daily outputs, then the faults are not picked up as the lowered modelled output compensates for the increased losses in actual daily output.

5.3.2 Performance over-time

The over-time checks domain aims at analysing concentrated results over the period of several weeks if data are available. Specifically, the same checks as described in the hourly domain are also performed in the past and repeated for a selected number of days, initially

set at 10. This short period check aims at looking at power losses which do not depend on seasonal effects and also shows the frequency of fault indicators at particular hours of the day. However, if a clear day is not found within these 10 days, namely at least one day with an average clearness index of over 0.6, then the checks are gradually increased to up to 30 days to detect less obvious losses. The choice of keeping the number of days relatively small is to also allow for the analysis of a greater number of systems which have limited data availability. If data are available for a longer period of time then monthly analysis may also be performed for selected systems and comparison with neighbours as is discussed next. At the end of this check the frequency of occurrences is used to determine whether the system may present a fault or not and what the type of fault is, namely partial shading or increased losses throughout the day. The next case is a PV system (see Figure 5.13) which presents early shading as can be seen for two random days in winter and summer. The morning shading is more pronounced on a winter day where the sun is lower in the horizon.

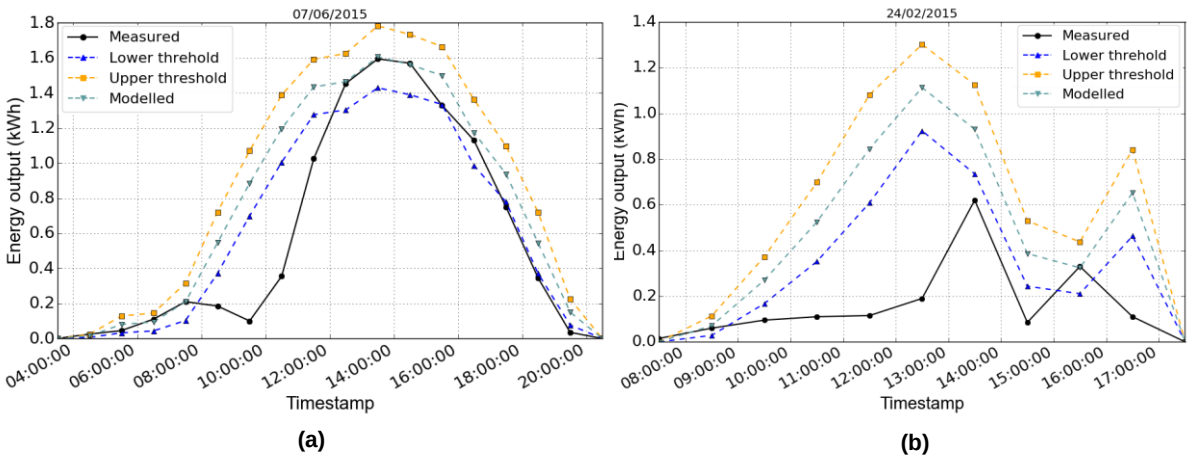


Figure 5.13. Hourly energy output of a PV system where early morning shading is indicated for (a) summer (07/06/2015) and (b) winter (24/02/2015).

The procedure of fault detection is repeated for several days in the past. An example is given for 11 days starting at 02/06/15 to 12/06/15. The indicators are binned according to the hour of the day and the clearness index for 11 days in a row. These results can be visualised by using conditional formatting for the applied thresholds (see Figure 5.14) where red colour indicates a negative flag, namely energy output is lower than expected.

hour \ K_t	0.46	0.49	0.50	0.52	0.54	0.57	0.64	0.68	0.71	0.72	0.75
4	0	0	0	-1	0	0	1	-1	0	0	0
5	0	0	-1	-1	0	0	0	0	-1	0	0
6	1	0	-1	-1	1	-1	0	0	-1	0	-1
7	1	0	-1	-1	0	-1	0	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
10	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
11	1	-1	0	0	-1	1	-1	-1	0	0	-1
12	0	0	-1	1	0	0	-1	0	0	0	0
13	-1	1	-1	-1	0	0	0	-1	0	0	-1
14	-1	-1	-1	0	-1	0	0	0	-1	-1	-1
15	-1	-1	0	0	-1	-1	-1	-1	-1	0	0
16	-1	-1	1	-1	0	0	0	-1	-1	-1	0
17	-1	0	0	-1	0	0	-1	-1	-1	-1	-1
18	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1

Figure 5.14. Plot of conditional formatting for each type of indicator per hour of the day and daily clearness index (K_t). Red colour indicates lower energy output (-1), yellow indicates higher energy output (+1) and green indicates (0) expected energy output.

The occurrences for everyday are well over the daily threshold and the concentrated results show that the particular PV system produces lower output at the majority of the days and especially from 7:30 to 11:30 and from 17:30 to 18:30. There are very few hours where hourly output exceeds the upper thresholds and mostly for days with a lower clearness index where hourly irradiation modelling uncertainty is expected to be higher. Due to irradiance uncertainty it is expected that losses which cause less than 10% loss in hourly output (see Figure 5.8) will not be detected on a day with low irradiation and thus the checks should always include at least 1 day with high clearness index (over 0.6) as in the given example.

5.3.3 Performance ranking based on neighbouring PV systems

The overall method does not rely on the comparison with neighbouring systems, at least not in the initial application of the methodology where a large part of meta-data is found to be inaccurate. It can however be used as an additional check based on daily output of neighbouring systems and over time. However certain facts have to be taken into account

first, such as that a) a large percentage of neighbouring PV systems may present data or system quality issues and therefore these will not be sufficient indicators and b) there may not be a sufficient number of neighbouring systems that can be used in comparison. In [5] a high density area is used to apply the comparison of performance indices of neighbouring PV systems, however, that is not largely the case for the particular dataset.

At the initial stage of performance assessment where data quality is still unclear, comparing with neighbouring PV systems cannot be used as a reliable metric for the discrimination of faulty from normal systems. Once the quality of all PV systems is clarified and wrongly declared information is revised and corrected, then comparing with neighbours can be used as the initial stage of the fault detection procedure followed by checking the hourly patterns for a selected period of time as described previously, and only for those systems that failed the first check. This is used as a way to identify the issue from the energy output patterns of the particular homes (especially shading or component failure).

A PV neighbourhood is defined in a radius of 150m. This is because global horizontal irradiation is expected to be the same across this distance even on partly cloudy days. The average number of neighbours in the dataset was found to be 11 PV systems for this distance which is decreased to less than 3 PV systems for the smaller distance of 50m. The minimum number of neighbouring PV systems is set to the closest 3 also taking into account an additional criterion; according to this, identical systems in terms of technical description are chosen, if these are available. Therefore, the choice of neighbours is essentially determined by the following parameters:

$$neighbours = f(technical\ description, distance, \pm 10^0\ orientation, \\ number\ of\ neighbours \geq 3) \tag{5.4}$$

At first, the neighbourhoods are determined based on these features, and those who qualify to be in the same neighbourhood, are defined as “peers” [5]. Namely, the peers comprise the group of PV systems that are compared together as an additional performance indicator.

To demonstrate the process, a PV neighbourhood of 40 PV systems is chosen (see Figure 5.15), which is the largest one found in this dataset. The PV systems are then filtered based

on the criteria in (5.4). Statistically, it is expected that at least a small percentage of these systems will be matching peers. Choosing one of those systems as a reference point, 57% of the systems showed exact match in technical characteristics within a distance of 150 meters but very few (six PV systems) were found to have similar azimuth within ±10 degrees. This is expected as the average number of neighbouring systems is 11 and only a percentage of these present similar characteristics. Their input information is also checked for possible nominal capacity and azimuth inaccuracies.

Figure 5.15. Part of the neighbourhood which includes 40 PV systems in the particular area of Nottingham.

In long term data analysis studies the results are usually based on utilising a year's worth of data. However, this is very constraining in terms of detecting a fault on a relatively timely manner. In this case, daily data are utilised but on a quarterly basis as this is a representative period between obtaining and analysing data from various sources (both electrical readings and weather data). Furthermore, comparing PV systems based on their PR alone is not the optimum solution due to high daily PR variations, as seen in Figure 5.16. However, these variations generally follow the same trend apart from few exceptions which are discriminated from normal behaviour. These are most likely due to accidental system switch offs which are often caused by residents themselves or scheduled system checks.

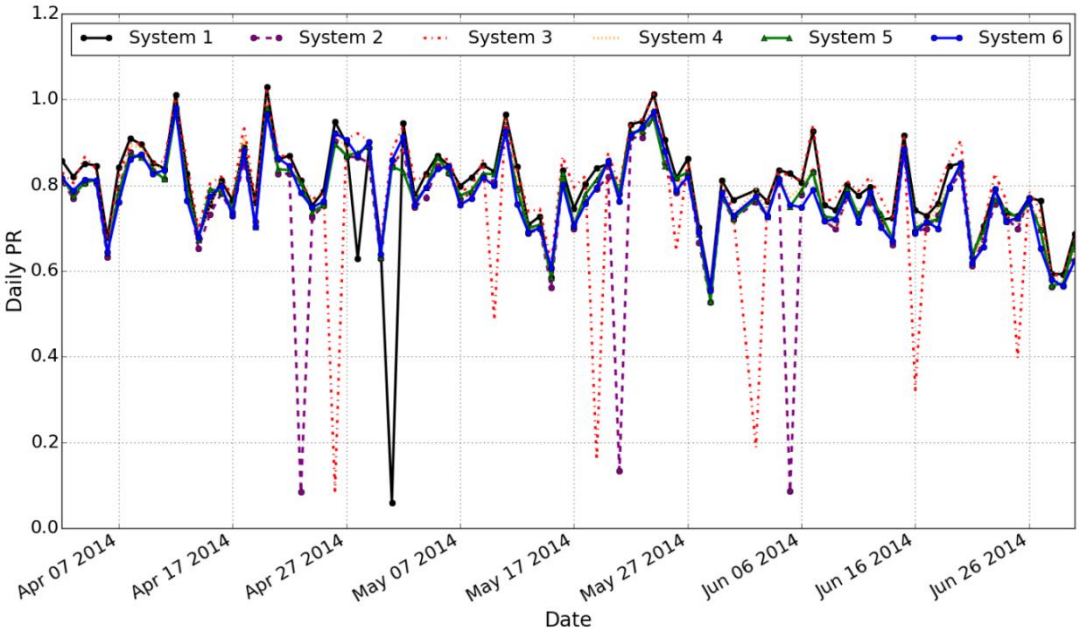


Figure 5.16. Daily performance ratio of six PV systems in the same neighbourhood for three months.

Thus, when comparing PRs among neighbouring PV systems this daily variation can be negated by using a normalised metric as a performance index. After determining the systems with inaccurate data, the following performance index is introduced which is calculated daily for each neighbour (in the same neighbourhood) and for a specified period of time:

$$PI_i = \frac{PR_i}{PR_{max}} \tag{5.5}$$

The closest the PI is to unity, the more the PV systems are correlated to each other. Applying this performance metric for the particular neighbourhood, yields the distribution seen in Figure 5.17. The abnormal values (denoted with red colour in Figure 5.17) can then be highlighted based on the median (MED) and median absolute deviation (MAD). Thus:

$$Lower\ threshold = MED - 3 \cdot MAD \tag{5.6}$$

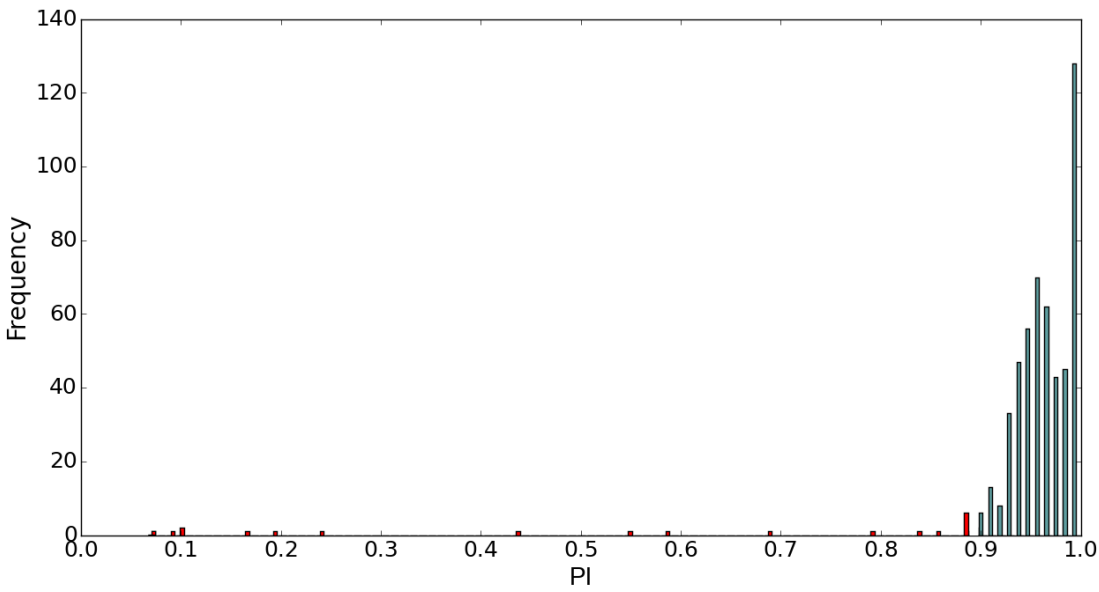


Figure 5.17. Histogram of daily normalised PIs for the six neighbouring PV systems for three months. Red colour represents the PV systems where daily PI is lower than the applied threshold, based on the median absolute deviation.

Those PV systems as well as the particular day(s) where PI is found to be abnormally low are noted in the quarterly records and further checks are applied in the previous domains to identify the error (if possible). This is further demonstrated in the following case study, where also financial implications arising from increased energy losses are considered.

5.4 The importance of early-stage quality assessment in monitoring

PV system owners and investors face failure risks in the commissioning stage (such as planning errors and installation defects) as well as during the lifetime of their system. Detecting failures within a few months of operation is important to minimise these risks at an early stage of the project. In the same way, early-stage data quality assessment is important to identify flaws in data quality which will later reduce the efficiency of fault detection and performance assessment for these systems. To further demonstrate this, a specific case study for a PV system with severe underperformance is presented. The fault detection procedure is analysed at each step and conclusions are drawn with regards to the system's return on

investment and losing the system's warranty. This PV system is found to have similar characteristics to 4 neighbouring PV systems within a distance of less than 30m. The quarterly daily PR distributions (April-June 2015) are shown in Figure 5.18. It is found that about 74 PR values were abnormally low based on (5.6), all corresponding to a particular PV system.

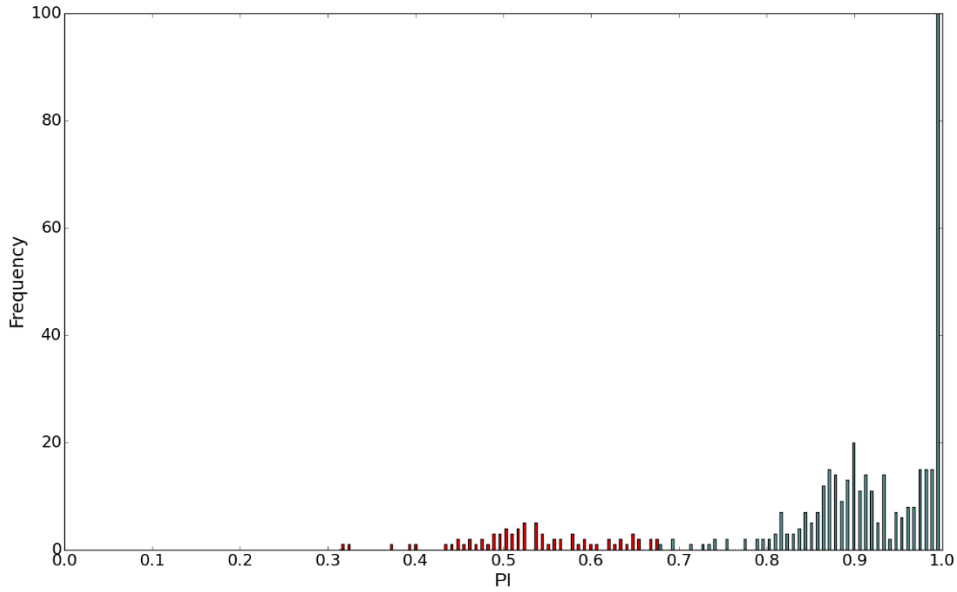


Figure 5.18. Histogram of daily normalised PIs for the six neighbouring PV systems for three months. Red colour represents the PV systems where daily PI is lower than the applied threshold, based on the median absolute deviation. 74 incidents were found to be abnormally low where all correspond to the same (faulty) system.

Figure 5.19 is obtained by comparing the hourly outputs of the faulty PV system with the model and a neighbour (of same technical description and installation dates and angles) over 10 days.

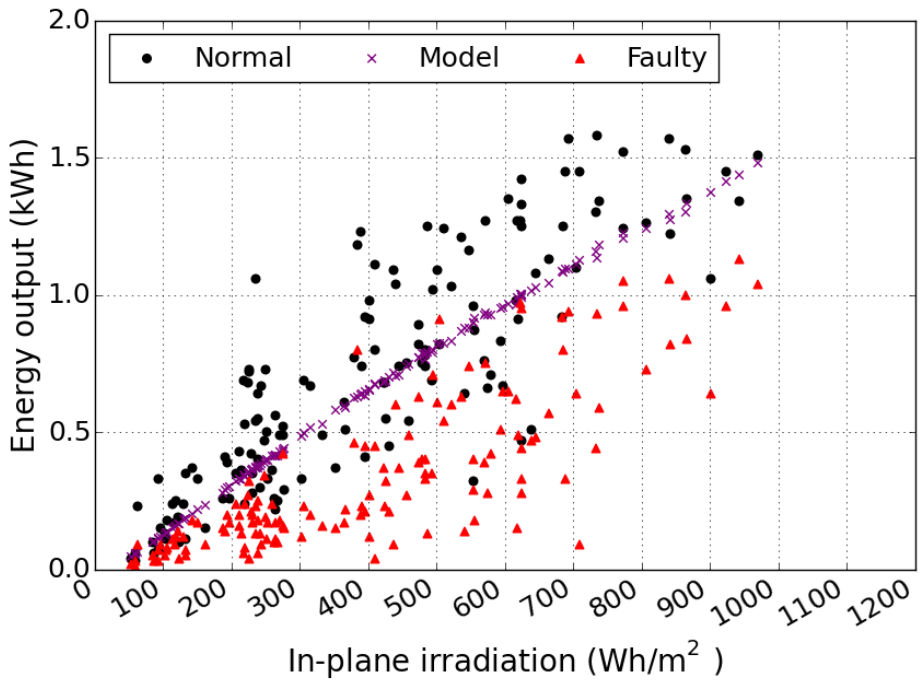


Figure 5.19. Comparison of hourly energy output vs in-plane irradiation of the model, the faulty and a normal neighbouring PV system.

From the hourly energy output of the faulty system and its comparison to its neighbour and the model, it is apparent that the particular system presents constant energy loss which can be further seen by looking at its daily profile on two days with high (K_t=0.72) and low (K_t=0.34) clearness indices respectively.

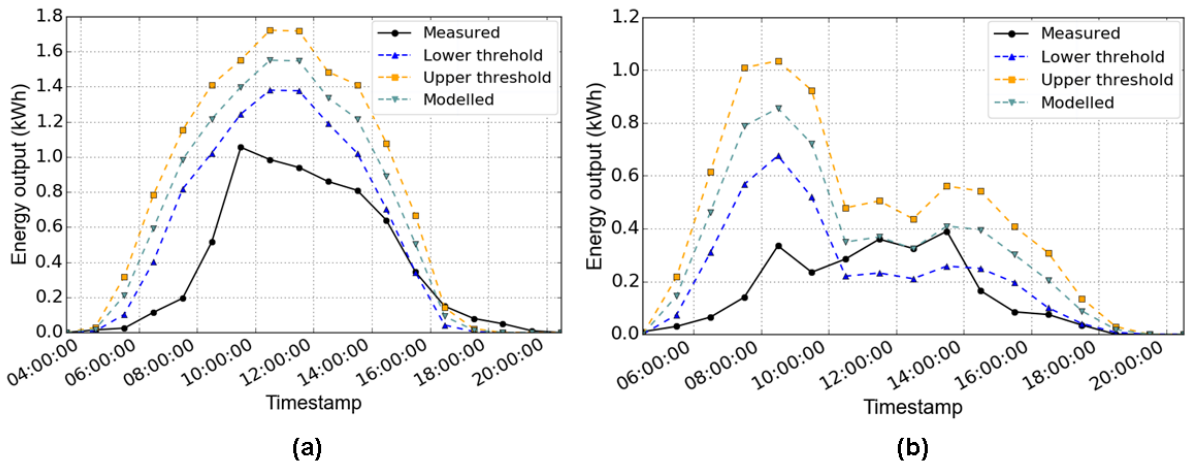


Figure 5.20. Hourly energy output of a faulty PV system on (a) a clear day and (b) a partly cloudy day.

Apart from partial shading which causes energy loss in the morning hours, constant energy loss is also noted throughout the day. According to [4] this failure footprint can be due to a module or a string defect or due to soiling. Given that all PV systems are in the same neighbourhood, soiling is unlikely to have affected only one system. String defect is a possible explanation in this case. In fact, this system consists of one single string and behaves as if 2 of its modules are not operating. If more than 2 modules are not operating then the resulting voltage would be too low compared to the minimum input voltage specifications of the particular inverter and thus zero output would be expected. Further losses can also be caused from increased module temperature. This can occur often when birds are nesting at the back of the modules and/or when the distance between the roof and the modules is very small, and thus the modules are not very well ventilated.

The average annual PR for the particular system from 2012 to 2015 is 0.40 and has shown a low annual PR since the first year of its installation. This shows that the apparent constant energy loss can be either due to occurring faults (apart from shading) as mentioned above or can be due to wrongly declared capacity, namely if the array consists of 6 modules in series instead of 8. This includes further implications such as for example the right sizing of the inverter or possible confusion with neighbouring systems, which may be over-estimated instead. The result in both cases is that the calculated PR of this system is too low.

The PR is by far the most favoured performance indicator by developers and financers because it also takes into account overall irradiation as opposed to final yield (see definitions in 2.6) [131]. This means that whichever the reason is that causes low PR estimation (monthly or annual) , the result will be equally detrimental when it comes to the evaluation of

performance guarantees [131], [159]. However, assuming that system size is indeed as declared, then this means that the particular system has produced 940 kWh less in 2015 compared to its neighbour. This would cost the owner an average of about 130£ per year based on the current value of electricity kWh unit (assuming 13.86 p/kWh[2]) and thus over 400£ in 4 years. Losses may be higher depending on the Feed-in-Tariff schemes valid at different periods of installation. For example the FiT rates in the UK until the 30th of April 2013 were at 15.44 p/kWh [190]. In this simple assumption, self –consumption is not taken into account, thus income losses may be even higher. Additionally, this does not include the maintenance costs which could arise due to potential system faults not being detected on time, increasing the lifetime costs of the PV system (also module degradation rates) and the payback period of the investment.

As discussed in Chapter 3, this is only an example case. On average, 3% of the systems (per year) show significantly low performance ratio (<0.6) including systems that may be wrongly declared (but neither checked or corrected at an early stage). Furthermore, only in 2014, 6% of the PV systems presented more than 13% days of zero generation (see Figure 5.21(a) and (b)).

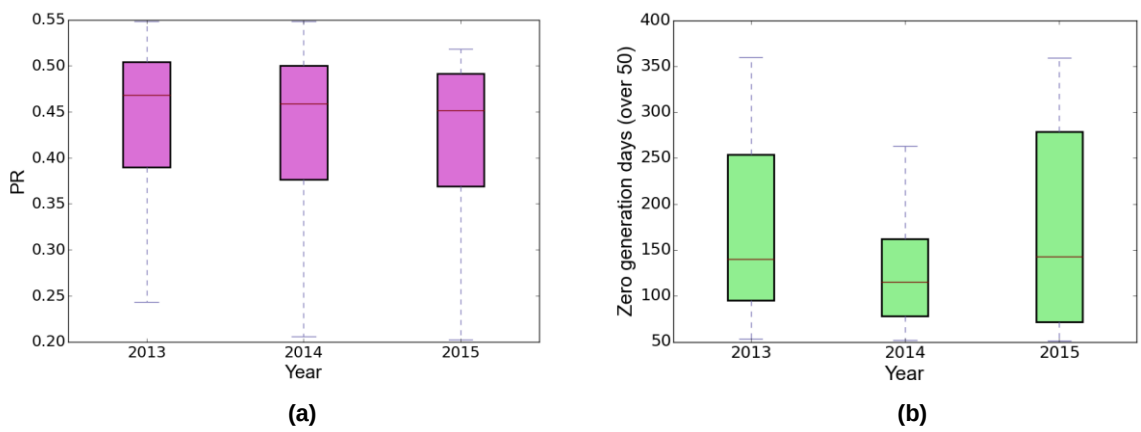


Figure 5.21 Boxplots of (a) PV systems with annual PR < 0.6 and (b) PV systems with overall zero generation over 50 days.

[2] Simple calculations based on http://www.carbon-calculator.org.uk/ .

There are over 840,000 domestic solar PV installations in the UK with an average size of 3kW$_P$ and average annual generation of 850 kWh per installed kW$_P$. Considering that the majority of small systems are not monitored, the situation met here is not specific to this dataset only and can be up-scaled to a national level corresponding to about 25000 under-performing systems and an overall cost of £3.2M/a. Thus, lost generation not only costs individual system owners revenue from the Feed-in Tariff scheme, but also adds national-level cost.

A summary of system (SQ) and data quality (DQ) issues with regards their impact on PR and accurate PR estimation, studied in this work, is presented in Table 5.7. The case of more than one issue occurring at the same time (as shown in 5.3.1) is not included in Table 5.7 as this is not a straightforward situation and can cause a combined impact on PR which cannot be predicted, as already demonstrated in this chapter. Negative impact implies reduction of PR and positive means that PR is overestimated (this relates to data quality only).

Table 5.7. Table of studied data and system quality issues related to the particular dataset and their impact on PR.

Occurrence	Type	Impact on PR (or PR estimation)
Zero generation	SQ	Negative
Missing data	DQ	Negative (if not inferred)
Low generation: presence of several faults including shading and component failures	SQ	Negative
Overestimated nominal capacity	DQ	Negative
Underestimated nominal capacity	DQ	Positive
Wrongly declared azimuth	DQ	Positive or negative

As seen in Table 5.7, most issues cause PR reduction, but when occurring together the outcome is highly unpredictable and may be either positive or negative with regards to PR estimation.

All the arising issues from limited data and data quality have a huge impact on accurate performance assessment and detection of faults. For the demonstration of the fault detection framework and the identification of very specific failures such as inverter shutdown, system isolation and shading, only a small percentage out of an average of 1750 systems (per year) could be utilised, due to limited or ambiguous input information. Out of this percentage few PV systems were found to be wrongly declared in terms of nominal capacity and azimuth. While wrong azimuth can be detected and corrected in a more straightforward manner, the estimation of actual nominal capacity is not always obvious. Based on the overall incidents discovered in the particular dataset, the minimum requirements for assuring sufficient data quality for remote monitoring applications are gathered in Table 5.8. These primarily concern the type and quality of meta-data information which should be included to every domestic installation, both on installer side as well as the monitoring company that take over.

Table 5.8. Minimum data requirements for remote performance assessment of domestic PV systems

Module level	Array level	System level	Additional information
Single module capacity	Number of strings in series and parallel	Number of inverters	Date of installation, exact postcode and address
Manufacturer information	Overall array capacity given in 4 significant digits	Inverter model	XY coordinates for utilization in mapping software
Specific module model information	Azimuth, inclination and elevation	Inverter manufacturer	Ideally a diagram of the topology or a picture of the PV system

Furthermore, instruments of measurement for both azimuth and inclination should be specified and also a particular reference point (e.g. South = 0 degrees) should be used and always declared. Overall capacity should be used as calculated and at 4 significant digits (for example 1.645 and not 1.65 kW_P). Erroneous rounding up or down of calculated capacities very often leads to wrong capacity estimations and compromises the modelling accuracy.

5.5 Chapter conclusions

A fault detection framework is developed which includes individual performance checks on a large number of PV systems and the role of data quality on accurate detection is demonstrated on selected case studies. The detection relies on the assignment of different indicators obtained by comparing actual with modelled energy outputs per system and for different time periods. Daily profiles are used to detect specific types of energy losses such as partial shading and constant losses which can be due to zero generation, string defects or inverter malfunction. Due to high irradiation modelling uncertainty, faults which cause less than 10% power loss (hourly) are not detected on days with low clearness index (<0.6). Therefore, repeating performance checks over several weeks aims at utilising at least three days with a daily clearness index over 0.6 and thus confirming the existence of a fault. An additional domain is employed whereby PV systems are classified based on their normalised performance with regards to the "best" PV system in the defined neighbourhood. The definition of a PV neighbourhood is based on several elements such as distance between different PV systems and their technical characteristics. Using PV neighbourhoods for the detection of the lowest performing systems is useful in long term data analysis, where daily performance variations are much lower than hourly variation in sub-daily analyses. However, its application is feasible only when a PV system has at least one appropriate neighbour which can be utilised in the detection framework.

It is shown that a large number of PV systems present limited or erroneous input information, an issue in domestic monitoring which has not yet drawn significant attention in literature. Such cases can neither be assessed nor used as reliable indicators for other PV systems in the neighbourhood. Moreover, due to the limited input information whereby only energy output is available, and the large number of PV systems with erroneous input information, the identification of more complicated faults cannot be applied with an automated manner. More implications arise from the combined existence of both data and system quality. Wrongly declared azimuth and nominal capacities as well as ambiguous information regarding the systems' technical characteristics comprise almost 60% of the cases in the particular dataset. More importantly, cases where low performance is obscured by

wrongly declared (underestimated) capacity are found, and these cannot be readily corrected until confirmed by the administrators/owners of the systems. An example showed that such cases continue to exist after years of systems' operation. To avoid energy losses from PV systems which are either not detected timely or PV systems which cannot be assessed due to obscuring data quality, it is necessary to apply an early-stage quality assessment ("commission" monitoring) whereby minimum requirements on data accuracy are ensured.

Chapter 6

Conclusions and future prospects

6.1 Conclusions

The main goal of this work was to highlight, analyse and remedy the impact of the most common data quality issues found in photovoltaic (PV) performance datasets, and specifically in domestic PV monitoring. The work was largely centred around the incentive of adding experience on lessons learnt in PV monitoring which will improve the accuracy of future performance assessments. The work progressed through three different phases:

- Firstly, common data quality issues in domestic PV monitoring were identified, based on a large domestic PV dataset. Statistical procedures were proposed for the automatic detection of PV systems with erroneous system descriptions, based on annual figures of performance ratio and specific yield. The impact of erroneous system descriptions on annual PR was demonstrated, using real case examples. The results showed that the accuracy of PR assessments in PV fleet analyses may be distorted by these data quality artefacts if these are not identified.

- After realising that missing data is a very common occurrence in photovoltaic monitoring, there was a strong incentive to develop a method in order to infer the missing data. This area of research has been largely developed in other fields but was introduced for the first time in the field of photovoltaics. Various methodologies were applied where different cases of missing data were presented. Both an empirical model and a model based on neural networks were presented for one of the cases. Furthermore, the role of irradiation underestimation was discussed, as it affects the accuracy with which missing data in PV energy output are inferred.

- Finally, having looked at the impact of data quality in domestic PV monitoring on annual performance assessment and then specifically at the very severe issue of missing data, the next step focused on the fault detection of PV systems. In this context, a remote failure framework was developed including several aspects in photovoltaic modelling and the application of remote weather data. Real cases were shown where limited input information, ambiguous and erroneous data were found to obscure failure detection. The results showed that the impact of data quality on fault detection can only be mitigated if data quality assessment is applied at an early stage monitoring, to achieve better and more efficient performance assessments in the future.

6.1.1 Quality assessment in domestic PV monitoring

The majority of performance assessments applied on domestic PV systems are based on scarce and remotely accessed system description data. It is therefore a question to what extent commonly employed performance indices such as the performance ratio (PR) can realistically describe PV performance if data quality is not considered. An insight of the most common quality issues found in domestic monitoring datasets was given based on a dataset of 1788 PV systems in Nottingham. Initially, a methodology based on annual performance ratio and specific yield distributions and the statistical median absolute deviation was applied to detect and classify PV systems using distinct identifiers on data and system quality. Specifically, the classification of PV systems was carried out according to identified categories of data quality, such as missing data and wrong system description, as well as system quality such as increased zero generation or low performance due to other faults.

It was found that irradiation modelling underestimation, missing data, zero generation and erroneous input information are the most influential factors on PR estimation. Annual irradiation bias was about -10% for Loughborough based on 2014 data, which increased the estimated PRs by the same percentage. This demonstrated, that any performance assessment

studies which are based on remotely inferred data should always include uncertainty factors concerning irradiance modelling.

However, the rest of the quality checks were applied based on the median absolute deviations of PR and specific yield distributions, and therefore irradiance underestimation did not affect the results. More than 1.5% of the PV systems were found with possibly overestimated capacities, and a 1.9% were found with possibly underestimated capacities, excluding those cases which may be underestimated but with faulty behaviour. As for erroneous information on installation angles, about 8 cases were found to have a ±180 degrees difference from actual azimuth. More cases of smaller deviations in azimuth are found later, as these cannot be detected through the annual analysis but through the analysis of hourly profiles on a clear day. Such issues indicated the lack of adequate documentation or installation supervision in domestic PV system installations and the necessity of addressing these issues in PR assessment studies. A significant 6.5% of the whole population showed an absolute azimuth deviation from 20 to 50 degrees, which may cause a substantial difference in calculated PRs and up to 25%. Finally, an average of 4% of the PV systems were found to have more than 30 days of missing data per year and thus, missing data was immediately identified as a factor with a high impact on the performance analysis, especially because missing data is often mistaken with zero generation.

6.1.2 Inference of missing data

In this work three cases of data loss were identified, whereby two cases are applicable to any type of PV monitoring while the third case mostly concerns domestic PV monitoring. The back-filling methodologies were mostly based on a simple empirical model and measured or inferred input weather data. The model's coefficients were extracted by fitting the model to energy output data taken from dates surrounding the gap, whose size was considered to be up to one month. In the first case, energy output is assumed to be missing but on-site weather data are available. Particularly in this case, the applied inference method yielded accurate agreement for daily and monthly energy output while almost 100% accuracy in monthly PR was achieved. In the third case it is assumed that weather monitoring is not available and energy output is missing, as often found in a domestic PV system. In this case then, the model

was trained based on inferred weather data. The validation was carried out by using two real PV systems from the Nottingham City Homes dataset. The back-filled results showed accurate agreement for daily and monthly energy output for both smaller and larger training data pools. Moreover, the negative bias deriving from the inference of in-plane irradiation modelling was largely diminished. It was shown that by using inferred energy output for two weeks of missing output, instead of completely disregarding this period, the annual PR was improved by 7.2% for one PV system.

In the second case, both energy output and weather data are considered to be missing for a month. Therefore, energy output was back-filled based on remote weather data. This unavoidably caused an underestimation of the energy yield, which is almost exclusively due to a negative bias in in-plane irradiation modelling. However, also in this case the monthly PR was predicted with a significantly low error, indicating that all the employed methodologies thus far were successful to achieving good accuracy in one of the key parameters in PV performance assessment and warranty verification.

In order to achieve a better accuracy in the back-filling of energy output for case 2, an alternative method was also developed based on an artificial neural network (ANN). The proposed configuration was very simple and it only required global horizontal irradiation, ambient temperature and sun position angles as input data and returned in-plane irradiation and energy output. Employing ANN improved the prediction accuracy of the back-filled monthly energy output by almost three times. Further Investigation on the impact of missing data on the monthly PR of a system in case 2, it was found that without back-filling the monthly PR may be 3% off its actual value, depending on the days included in the assessment, whereas using the proposed back-filling methods the obtained PR lied within 0.8% of its actual value.

The proposed back-filling techniques yielded satisfactory results in all cases and with an emphasis on daily and monthly analyses. However, these are applied on PV systems where it is known or assumed that no failures have occurred during the missing period. If this assumption does not apply then the back-filled results will be misleading. Therefore, although back-filling should be applied for a more realistic PR assessment for a short period when monitoring fails, it does not replace the importance of monitoring, which is the only way to ensure timely detection of system faults.

6.1.3 Remote failure detection framework with limited input data quality

Although there are a number of studies dealing with fault detection, the information on the factors that affect the data quality of the underlying assessments is very limited. Having detected abnormal PR values due to wrong system description data raised important concerns in cases where data quality issues cannot be easily distinguished from actual system faults. Thus, the final step was to further study the implications of data quality on remote failure detection on domestic PV systems and the arising financial implications either due to undetected faults or due to false estimation of PR. In order to achieve that, firstly a failure detection framework was developed based on manufacturer datasheets, one-diode modelling, and remotely inferred input weather data. Secondly, the efficiency of the method was demonstrated using specific case studies from the Nottingham City Homes dataset and for three different domains; based on indicators with regards to daily profiles, a selected period of time and performance comparison with neighbours. The sequence of these checks was not necessarily applied with this order, but it depended on the specific case demonstration. Daily profiles were used to detect specific types of energy losses such as partial shading and constant losses which were due to zero generation, string defects or inverter malfunction. Repeating performance checks over several weeks aimed at utilising more than one clear days but also to ensure that there were not false alarms in the initial check. That was due to high irradiation modelling uncertainty faults, due to which, less than 10% power loss are not expected to be detected on days with lower clearness index. Finally, a normalised performance index was introduced whereby PV systems in the same neighbourhood were ranked according to their performance. The lowest performing systems were distinguished based on the median absolute deviation of daily PRs for the neighbouring systems on a quarterly basis.

During fault detection, data quality issues were exposed. About 100 more cases were found with wrongly declared azimuth, which would otherwise be detected as PV systems with very low generation. Underestimated capacity in combination with system faults, created false negative alarms for one detected system. Wrongly declared azimuth and nominal capacities as well as ambiguous information regarding the systems' technical characteristics

were found to comprise almost 60% of the cases in the particular dataset. More importantly, such issues can be up-scaled to account for similar situations in the UK and worldwide [4],[5]. Data quality which obscures fault detection creating false negative alarms leads to increased energy losses and costs a significant amount to the owners of the PV systems from the feed-in- tariff scheme. Moreover, wrongly estimated PR due to data quality may have the same effect on the financial warranties of the investment and thus based on the investigation on the particular dataset, minimum data requirements were introduced at the end of this work as a pre-requisite and data quality assessment in domestic PV monitoring.

6.2 Future prospects

This work focused on the impact of data quality on PV performance assessment and fault detection, and the development of novel techniques for the inference of missing data in photovoltaic monitoring. This is expected to raise the interest for the treatment of missing data in other fields of photovoltaics such as for example missing information in long term reliability datasets, a very popular topic, as well as missing data in PV failure modes datasets. It will also alert the PV community about the importance of treating missing data, and further improving the proposed techniques. Accurate results were obtained for the inference of energy output and performance ratio. Especially the improvement of PR when missing data exceeds a certain threshold can be used as a potential tool to aid contractual agreements, where even 1% of PR loss may have a detrimental impact.

In terms of energy output prediction, uncertainty mainly derives from in-plane irradiation modelling and this effect has already drawn significant attention in literature. It is thus expected that with the development of improved models for the separation and translation of global horizontal irradiation to in-plane, the accuracy of the inference for energy output by using the empirical model will also improve further. An improvement was already achieved in this work by the replacement of the empirical model with an artificial neural network. Further improvement of the applied network configuration could potentially be achieved if other types of neural networks, or machine learning algorithms, or optimisation algorithms were applied. This field however is vast and constantly new types of machine learning algorithms or types of neural networks are reported in literature. There is thus enough room

for the improvement of the given neural network and further investigation of new machine learning algorithms to even better accuracies for the inference of missing energy output. When applying back-filling to other locations, another source of uncertainty may derive from the fact that kriging interpolation accuracy depends on the density of meteorological stations around the location of interest. Thus, the results would have to be further validated for more sites with different densities of meteorological stations. In this case the accuracy of back-filling can potentially be improved by utilising combined sources of solar radiation including satellite data, which cover larger geographical areas.

Apart from the missing data, other data quality issues were found in domestic PV monitoring such as ambiguous or erroneous input information. These issues were distinguished and their impact was demonstrated on both the annual PR assessment and fault detection by means of remote monitoring. In terms of remote fault detection, the applied methodology, including the modelling steps and the determination of applied thresholds are subject to further investigation and improvement. In addition to the applied framework, the proposed training procedures and models using past data for back-filling could also be applied here to detect potential faults in systems, once data quality has been detected and corrected. In the same context, the training can potentially include a power degradation factor where fault detection utilises more than one year's worth of data. Finally, the employed empirical and physical models could be further improved or potentially be replaced in order to accurately predict the PV output for technologies other than crystalline silicon.

In terms of the impact of data quality on fault detection efficiency, similar issues in domestic PV monitoring are only vaguely recognised in literature, and they have not been studied in depth in more recent performance assessments. Towards this direction, the findings from this work have the potential to develop more interest on the quality of the performance assessments carried out on domestic PV systems. Moreover, quality assessment prior to fault detection might be a future prerequisite as it was shown that it can obscure the analysis significantly and fault detection works which are based solely on neighbouring systems could be significantly biased due to poor input information. Consequently, this means that there should be guidelines on the quality assessment steps carried out prior to further analyses taking place and these should be explicitly mentioned in future studies. The demonstration of more real case studies in literature such as the one carried out in this work

will further recognise these issues and further potential will grow for a better monitoring in domestic PV and more accurate and responsible performance assessments.

Appendix

All the tools presented in this Appendix, along with a large number of algorithms developed throughout this work are available for anyone to use, upon contacting the author.

The trapezoidal rule

This refers to the type of the numerical integration used to calculate an area, where the area under the curve is divided to equal trapezoids as seen in Figure 0.1.

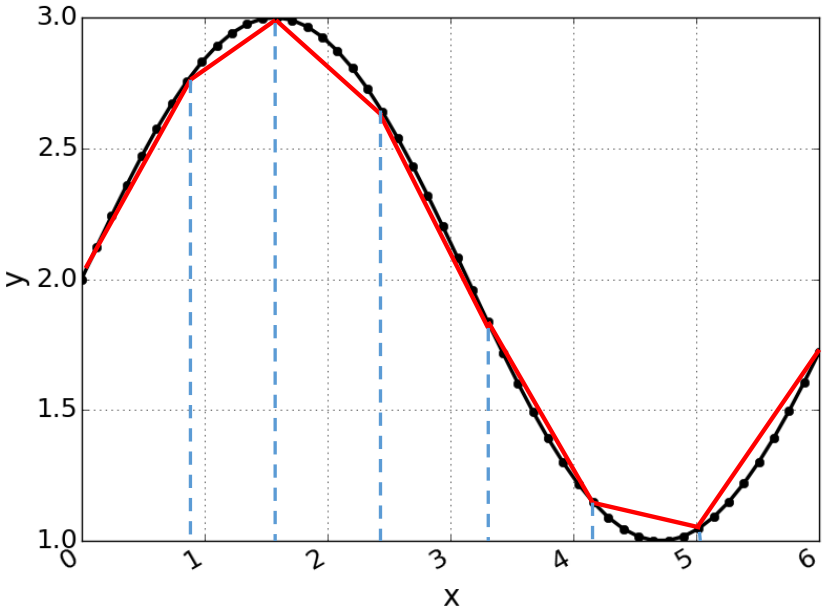


Figure 0.1. Numerical integration based on trapezoids.

So, the integral of the function of f is given by:

$$\int_a^b f(x)dx = \frac{b-a}{2N}(f(x_1) + f(x_2) + \ldots + f(x_{k+1})) \tag{0.1}$$

Where the x_k represent k intervals and $a = x_1,\ b = x_{k+1}$and:

$$h = \frac{b-a}{N} \tag{0.2}$$

Where N is the number of the trapezoids. Essentially, this is automatically determined when choosing the grid spacing value, h. The smaller this value the higher the number of the trapezoids and the more accurate the numerical integration is.

Aggregation of power output based on irradiance data

In remote climate sensing, different sources of solar radiation may use differing (temporal) reference systems. Data may be recorded at mean solar time (MST), local time (LT) or coordinated universal time (UTC), which may be a different timestamp system than the one used by the PV monitoring device. Furthermore, for hourly averaged data the timestamp may represent the middle or end of the averaging period depending on the convention used in the device or database. These factors cause temporal mismatches, which are more evident in sub-daily analyses. They can also affect the solar radiation separation modelling (via the clearness index calculation) when inferring in-plane irradiation using mixed sources of timestamps. Therefore, time reference should be first examined for both solar radiation and PV system monitoring to avoid timestamp mismatches [131]. In the NCH dataset, the supplied readings of power output correspond to half-hourly averages. The weather data, supplied by the Met Office, are given as hourly averages of minutely readings. Therefore, both datasets were used based on the middle of each hour as described in Figure 0.2.

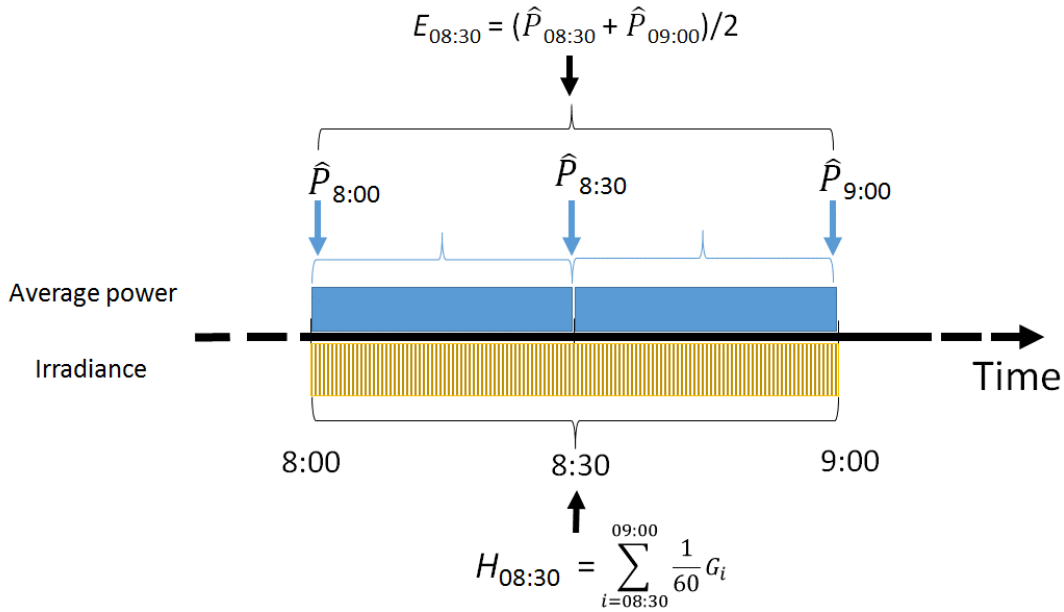


Figure 0.2. Averaging for power output and irradiation at the middle point of each hour.

Gaussian fitting of hourly energy profiles

The applied fitting error criterion is based on the equations described here, using a simple error criterion. A Gaussian fit is applied on the hourly output of the PV system, using Equation (0.7). The areas between the Gaussian and the clear sky irradiation curves are then compared based on the Equations (0.4)-(0.6), and highlighted in Figure 0.3.

$$f = a \cdot \exp\left(-\frac{(x - x_0)^2}{2\sigma^2}\right) \tag{0.3}$$

$$\varepsilon = \Delta A_{gaus\_\max\_left} \tag{0.4}$$

$$\delta = \Delta A_{gaus\_\max\_right} \tag{0.5}$$

The optimum azimuth (for a given tilt) is then chosen upon a minimising the function given by:

$$error = \sqrt{\varepsilon^2 + \delta^2} \tag{0.6}$$

The aim is to simulate clear-sky irradiation by automatically changing azimuth at a user defined step and to minimise the areas between the Gaussian and the clear-sky irradiation curves. The two areas are defined by the Gaussian maximum (which is the same as clear sky irradiation).

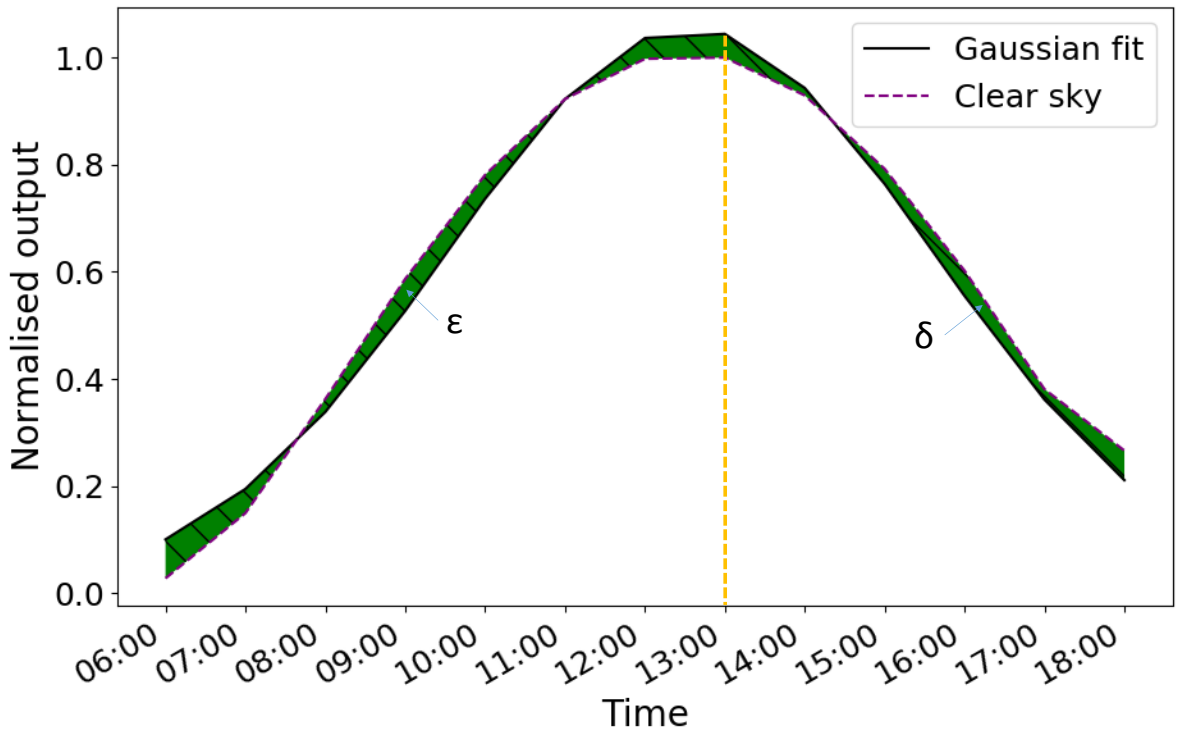


Figure 0.3. Gaussian fit and clear sky irradiation curves for the optimum azimuth angle (-2). The shaded area is the difference between the two curves.

Analytical I-V parameter expressions

For the parameter extraction the method proposed in [188] was employed and by using the following analytical expressions for initial values of fill factor, series and shunt resistance and diode saturation current:

$$V_{OC} = \frac{I_{MPP} V_{MPP}}{FF \cdot I_{SC_{STC}}} \tag{0.7}$$

$$I_{PH_{STC}} = I_{SC_{STC}}((R_S + R_{SH})/R_{SH}) \tag{0.8}$$

$$I_{O_{STC}} = I_{SC_{STC}}/(\exp\left(\frac{V_{OC_{STC}}}{nV_{TH}}\right) - 1) \tag{0.9}$$

$$A = \left|V_{MPP_{STC}} \cdot I_{PH_{STC}} \cdot I_{O_{STC}} \cdot exp\left(V_{MPP_{\mathrm{STC}}} + I_{MPP_{STC}} \cdot R_S\right) \cdot \mathrm{q}/(N_S nkT)\right) + V_{MPP_{STC}} \cdot I_{O_{STC}} - P_{MPP}\right| \tag{0.10}$$

And

$$I_{O_{STC}} = I_{SC_{STC}}/(exp\left(\frac{V_{OC_{STC}}}{nV_{TH}}\right) - 1) \tag{0.11}$$

$$R_{SH} = V_{MPP_{STC}}\left(V_{MPP_{STC}} + I_{MPP_{STC}} \cdot R_S\right)/A \tag{0.12}$$

Where

$$V_{TH} = N_S kT/q \tag{0.13}$$

And

V_{OC}	=	Open circuit voltage
I_{MPP}	=	Current at maximum power point
V_{MPP}	=	Voltage at maximum power point
P_{MPP}	=	maximum power
$I_{O_{STC}}$	=	Diode saturation current at STC
$I_{SC_{STC}}$	=	Short circuit current at STC
R_S	=	Series resistance
R_{SH}	=	Shunt resistance
$V_{MPP_{\mathrm{STC}}}$	=	Voltage at maximum power point at STC
$I_{MPP_{STC}}$	=	Current at maximum power point at STC

FF = Fill factor

The initial conditions for the initiation of the iterative process are:

$$R_S = 0, \qquad n = 1 \tag{0.14}$$

$$R_{SH} = \frac{V_{MPP_{STC}}}{I_{SC_{STC}} - I_{MPP_{STC}}} - \frac{V_{OC_{STC}} - V_{MPP_{STC}}}{I_{MPP_{STC}}} \tag{0.15}$$

$$I_{O_{STC}} = I_{SC_{STC}} / \left(exp\left(\frac{V_{OC}}{nV_{TH}}\right) - 1 \right) \tag{0.16}$$

If no optimum solution is reached for a specified number of loops (4000) then the process is aborted. The process is repeated for increasing series resistance and ideality factor and the final result includes the five modelling parameters, and the obtained V_{OC}, I_{SC} and P_{MPP}.

The web automation tool in Python

In order to acquire the Nottingham City Homes dataset a **web automation tool** was developed in Python as well as additional tools for the organising of the files and the import into the local database for further processing. Particularly, this set of data is hosted by a monitoring company where access is granted for specific users, namely the administrators and owners of the PV systems. Higher level users can access more than one systems and such a license was also employed here. However, the website navigation did not allow direct querying of the database for more than one homes at the same time, thus data from every single system would have to be manually downloaded per day and per year for 4 years for 1788 systems. This means that without the execution of an automatic routine/query the same procedure would have to be repeated for over 2.6 million times, corresponding to about 1.5 years of manual work, which is rather impossible. Instead a web automation tool was developed in Python which enabled fast downloading of more than 2.6 million CSV files in less than one month, and timed to only run over night and stop at a particular time in the morning so as to avoid overloading the hosting server during the day. This automated routine basically

resembles the moves of a human user in a much faster and much more organised and consistent way.

Organising and importing files into the database

Each downloaded CSV file from the monitoring portal of Nottingham City Homes had a particular format and thus the manipulation of files was employed by repeating the same routine which was based on parsing each file, detecting the date written inside the file as well as the unique identifier of the system (system ID) and renaming and sorting each file based on this sequence. Thus, the sorting of the numerous files was based on fast file parsing and string/date recognition algorithms. Files were then merged based on each system ID for every single day of operation for the particular system and automatically imported into the database (see Appendix) for safe storage and easier manipulation. The procedure is graphically described in Figure 0.4.

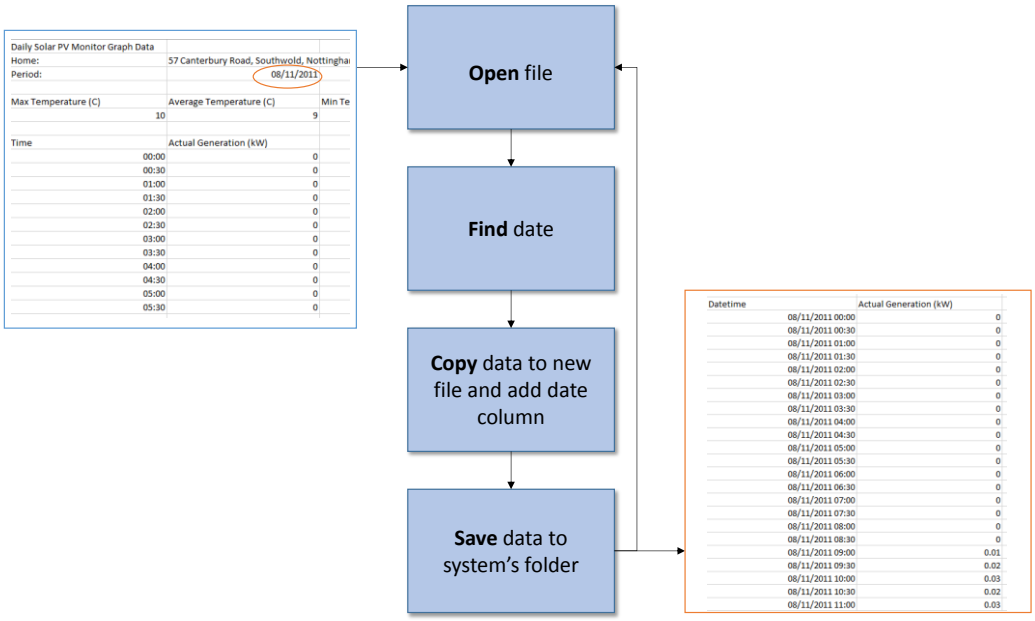


Figure 0.4. Timestamp creation procedure per system per file prior to final importing into the database.

References

[1] BEIS, "Department for Business, Energy & Industrial Strategy." [Online]. Available: https://www.gov.uk/government/organisations/department-for-business-energy-and-industrial-strategy. [Accessed: 18-Oct-2017].

[2] BRE, "Domestic photovoltaic field trials - final technical report," Building Research Establishment, Watford, UK, 2006.

[3] IEC standard 61724, "Photovoltaic system performance monitoring — Guidelines for measurement , data exchange and analysis," 1998.

[4] A. Drews *et al.*, "Monitoring and remote failure detection of grid-connected PV systems based on satellite observations," *Sol. Energy*, vol. 81, no. 4, pp. 548–564, Apr. 2007.

[5] J. Leloux, L. Narvarte, A. Luna, and A. Desportes, "Automatic fault detection on BIPV systems without solar irradiation data," *29 th Eur. Photovolt. Sol. Energy Conf. Exhib.*, no. September, pp. 1–7, 2014.

[6] S. Kurtz, J. Newmiller, T. Dierauf, A. Kimber, J. Mckee, and R. Flottemesch, "Analysis of Photovoltaic System Energy Performance Evaluation Method," United States. Dept. of Energy, NREL/TP-5200-60628, 2013.

[7] S. Younes, R. Claywell, and T. Muneer, "Quality control of solar radiation data: Present status and proposed new approaches," *Energy*, vol. 30, no. 9 SPEC. ISS., pp. 1533–1549, 2005.

[8] Y. N. Friesen, R. Gottschalg, H.G.Beyer, S. Williams, A. Guerin de Montgareuil, N. van der Borg, W.G.J.H.M. van Sark, T. Huld, B. Müller, A.C. de Keizer, "Intercomparison Of Different Energy Prediction Methods Within The European Project 'Performance' - Results Of The 1st Round Robin," in *22nd European Photovoltaic Solar Energy Conference*, 2007, no. September, pp. 3–7.

[9] Jyotirmoy Roy, T. R. Betts, and R. Gottschalg, "Accuracy of Energy Yield Prediction of Photovoltaic Modules," *Jpn. J. Appl. Phys.*, vol. 51, no. 10S, 2012.

[10] William Shockley, "Electrons and Holes in Semiconductors." second ed, D. Van Nostrand, Inc., New York, 1950.

[11] H. S. Rauschenbach, *Solar cell array design handbook: the principles and technology of photovoltaic energy conversion*. New York: Springer Science & Business Media, 1980.

[12] J. L. Gray, "The Physics of the Solar Cell," in *Handbook of Photovoltaic Science and Engineering*, John Wiley & Sons, Ltd, 2011, pp. 82–129.

[13] A. N. Celik and N. Acikgoz, "Modelling and experimental verification of the operating current of mono-crystalline photovoltaic modules using four- and five-parameter models," *Appl. Energy*, vol. 84, no. 1, pp. 1–15, Jan. 2007.

[14] V. Quaschning and R. Hanitsch, "Numerical simulation of current-voltage characteristics of photovoltaic systems with shaded solar cells," *Sol. Energy*, vol. 56, no. 6, pp. 513–520, 1996.

[15] V. Lo Brano, A. Orioli, G. Ciulla, and A. Di Gangi, "An improved five-parameter model for photovoltaic modules," *Sol. Energy Mater. Sol. Cells*, vol. 94, no. 8, pp. 1358–1370, 2010.

[16] H. A. B. Siddique, P. Xu, and R. W. De Doncker, "Parameter extraction algorithm for one-diode model of PV panels based on datasheet values," *4th Int. Conf. Clean Electr. Power Renew. Energy Resour. Impact, ICCEP 2013*, pp. 7–13, 2013.

[17] G. Ciulla, V. Lo Brano, V. Di Dio, and G. Cipriani, "A comparison of different one-diode

models for the representation of I–V characteristic of a PV cell," *Renew. Sustain. Energy Rev.*, vol. 32, pp. 684–696, Apr. 2014.

[18] K. Ishaque, Z. Salam, and H. Taheri, "Simple, fast and accurate two-diode model for photovoltaic modules," *Sol. Energy Mater. Sol. Cells*, vol. 95, no. 2, pp. 586–594, Feb. 2011.

[19] B. Werner, W. Kolodenny, M. Prorok, A. Dziedzic, and T. Zdanowicz, "Electrical modeling of CIGS thin-film solar cells working in natural conditions," *Sol. Energy Mater. Sol. Cells*, vol. 95, no. 9, pp. 2583–2587, 2011.

[20] N. Barth, R. Jovanovic, S. Ahzi, and M. A. Khaleel, "PV panel single and double diode models: Optimization of the parameters and temperature dependence," *Sol. Energy Mater. Sol. Cells*, vol. 148, pp. 87–98, 2016.

[21] IEC standard 61215, "Crystalline Silicon Terrestrial Photovoltaic (PV) Modules: Design Qualification and Type Approval," 2002.

[22] U. Stutenbaeumer and B. Mesfin, "Equivalent model of monocrystalline, polycrystalline and amorphous silicon solar cells," *Renew. Energy*, vol. 18, no. 4, pp. 501–512, 1999.

[23] J. Merten, J. M. Asensi, C. Voz, A. V. Shah, R. Platz, and J. Andreu, "Improved equivalent circuit and analytical model for amorphous silicon solar cells and modules," *IEEE Trans. Electron Devices*, vol. 45, no. 2, pp. 423–429, 1998.

[24] A. Mermoud and T. Lejeune, "Performance assessment of a simulation model for PV modules of any available technology," in *25th European Photovoltaic Solar Energy Conference*, 2010, no. September, pp. 6–10.

[25] M. Burgelman, J. Verschraegen, S. Degrave, and P. Nollet, "Modeling thin-film PV devices," *Prog. Photovoltaics Res. Appl.*, vol. 12, no. 23, pp. 143–153, 2004.

[26] R. Gottschalg, D. G. Infield, and M. J. Kearney, "Experimental study of variations of the solar spectrum of relevance to thin film solar cells," *Sol. Energy Mater. Sol. …*, vol. 79, no. 4, pp. 527–537, Sep. 2003.

[27] F. Attivissimo, F. Adamo, A. Carullo, A. M. L. Lanzolla, F. Spertino, and A. Vallan, "On the performance of the double-diode model in estimating the maximum power point for different photovoltaic technologies," *Meas. J. Int. Meas. Confed.*, vol. 46, no. 9, pp. 3549–3559, 2013.

[28] J. W. Bishop, "Computer simulation of the effects of electrical mismatches in photovoltaic cell interconnection circuits," *Sol. Cells*, vol. 25, no. 1, pp. 73–89, 1988.

[29] T. Ma, H. Yang, and L. Lu, "Development of a model to simulate the performance characteristics of crystalline silicon photovoltaic modules / strings / arrays," *Sol. Energy*, vol. 100, pp. 31–41, 2014.

[30] H. Tian, F. Mancilla-David, K. Ellis, E. Muljadi, and P. Jenkins, "A cell-to-module-to-array detailed model for photovoltaic panels," *Sol. Energy*, vol. 86, no. 9, pp. 2695–2706, Sep. 2012.

[31] J. P. Vargas, B. Goss, and R. Gottschalg, "Large scale PV systems under non-uniform and fault conditions," *Sol. Energy*, vol. 116, pp. 303–313, 2015.

[32] X. Wu, M. Bliss, A. Sinha, T. R. Betts, R. Gupta, and R. Gottschalg, "Accelerated Spatially Resolved Electrical Simulation of Photovoltaic Devices Using Photovoltaic-Oriented Nodal Analysis," *IEEE Trans. Electron Devices*, vol. 62, no. 5, pp. 1390–1398, 2015.

[33] L. Castañer and S. Silvestre, *Modelling Photovoltaic Systems using PSpice*. John Wiley & Sons Ltd, 2002.

[34] BSI EN 50380:2003-09, "Datasheet and nameplate information for photovoltaic modules," vol. 3, 2003.

[35] M. Wolf and H. Rauschenbach, “Series resistance effects on solar cell measurements,” *Adv. Energy Convers.*, vol. 3, no. 2, pp. 455–479, Apr. 1963.
[36] D. S. H. Chan and J. C. H. Phang, “Analytical methods for the extraction of solar-cell single- and double-diode model parameters from I-V characteristics,” *IEEE Trans. Electron Devices*, vol. 34, no. 2, pp. 286–293, 1987.
[37] H. Tian, F. Mancilla-David, K. Ellis, E. Muljadi, and P. Jenkins, “A cell-to-module-to-array detailed model for photovoltaic panels,” *Sol. Energy*, vol. 86, no. 9, pp. 2695–2706, Sep. 2012.
[38] W. De Soto, S. a. Klein, and W. a. Beckman, “Improvement and validation of a model for photovoltaic array performance,” *Sol. Energy*, vol. 80, no. 1, pp. 78–88, Jan. 2006.
[39] K. Ding, J. Zhang, X. Bian, and J. Xu, “A simplified model for photovoltaic modules based on improved translation equations,” *Sol. Energy*, vol. 101, no. July 2015, pp. 40–52, 2014.
[40] J. Bai, S. Liu, Y. Hao, Z. Zhang, M. Jiang, and Y. Zhang, “Development of a new compound method to extract the five parameters of PV modules,” *Energy Convers. Manag.*, vol. 79, pp. 294–303, Mar. 2014.
[41] A. Laudani, F. Riganti Fulginei, and A. Salvini, “Identification of the one-diode model for photovoltaic modules from datasheet values,” *Sol. Energy*, vol. 108, pp. 432–446, 2014.
[42] G. Farivar and B. Asaei, “Photovoltaic module single diode model parameters extraction based on manufacturer datasheet parameters,” *PECon2010 - 2010 IEEE Int. Conf. Power Energy*, no. 2, pp. 929–934, 2010.
[43] C. W. Hansen, “Parameter estimation for single diode models of photovoltaic modules,” 2015.
[44] A. Ortiz-Conde, F. J. G. Sánchez, and J. Muci, “New method to extract the model parameters of solar cells from the explicit analytic solutions of their illuminated characteristics,” *Sol. Energy Mater. Sol. Cells*, vol. 90, no. 3, pp. 352–361, 2006.
[45] J. A. Jervase, H. Bourdoucen, and A. Al-Lawati, “Solar cell parameter extraction using genetic algorithms,” *Meas. Sci. Technol.*, vol. 12, no. 11, pp. 1922–1925, Nov. 2001.
[46] K. Ishaque, Z. Salam, S. Mekhilef, and A. Shamsudin, “Parameter extraction of solar photovoltaic modules using penalty-based differential evolution,” *Appl. Energy*, vol. 99, pp. 297–308, 2012.
[47] K. M. El-Naggar, M. R. AlRashidi, M. F. AlHajri, and A. K. Al-Othman, “Simulated Annealing algorithm for photovoltaic parameters identification,” *Sol. Energy*, vol. 86, no. 1, pp. 266–274, 2012.
[48] V. Khanna, B. K. Das, D. Bisht, Vandana, and P. K. Singh, “A three diode model for industrial solar cells and estimation of solar cell parameters using PSO algorithm,” *Renew. Energy*, vol. 78, pp. 105–113, Jun. 2015.
[49] R. Gottschalg, “Environmental Influences on the Performance of Thin Film Solar Cells,” Loughborough University, 2001.
[50] J. C. H. Phang and D. S. H. Chan, “A review of curve fitting error criteria for solar cell I-V characteristics,” *Sol. Cells*, vol. 18, no. 1, pp. 1–12, 1986.
[51] R. Gottschalg, M. Rommel, D. G. Infield, and M. J. Kearney, “The influence of the measurement environment on the accuracy of the extraction of the physical parameters of solar cells,” *Meas. Sci. Technol.*, vol. 10, no. 9, pp. 796–804, 1999.
[52] A. P. Dobos and S. M. MacAlpine, “Procedure for applying IEC-61853 test data to a single diode model,” in *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, 2014, pp. 2846–2849.

[53] V. J. Chin, Z. Salam, and K. Ishaque, “Cell modelling and model parameters estimation techniques for photovoltaic simulator application: A review,” *Appl. Energy*, vol. 154, pp. 500–519, 2015.

[54] W. De Soto, S. Klein, and W. Beckman, “Improvement and validation of a model for photovoltaic array performance,” *Sol. Energy*, vol. 80, no. 1, pp. 78–88, Jan. 2006.

[55] B. Van Zeghbroeck, *Principles of Semiconductor Devices*. 2011.

[56] J. A. Duffie and W. A. Beckman, “Solar Engineering of Thermal Processes.” second ed. John Wiley & Sons, Inc., New York, 1991.

[57] “BS EN 60891: Photovoltaic devices — Procedures for temperature and irradiance corrections to measured I-V characteristics,” 2010.

[58] A. J. Bühler, F. Perin Gasparin, and A. Krenzinger, “Post-processing data of measured I-V curves of photovoltaic devices,” *Renew. Energy*, vol. 68, pp. 602–610, 2014.

[59] K. Ding, X. Bian, H. Liu, and T. Peng, “A MATLAB-simulink-based PV module model and its application under conditions of nonuniform irradiance,” *IEEE Trans. Energy Convers.*, vol. 27, no. 4, pp. 864–872, 2012.

[60] B. Marion, “A method for modeling the current-voltage curve of a PV module for outdoor conditions,” *Prog. Photovoltaics Res. Appl.*, vol. 10, no. 3, pp. 205–214, 2002.

[61] B. Kroposki, W. Marion, D. King, W. Boyson, and J. Kratochvil, “Comparison of Module Performance Characterization Methods for Energy Production Comparison of Module Performance Characterization Methods for Energy Production,” 2000.

[62] B. Marion, B. Kroposki, K. Emery, D. Myers, J. del Cueto, and C. Osterwald, “Validation of a Photovoltaic Module Energy Ratings Procedure at NREL,” *Nrel*, no. August, 1999.

[63] E. D. D. R.P. Kenny, G. Friesen, D. Chianese, A. Bernasconi, “Energy Rating of PV Modules: Comparison of Methods and Approach,” in *3rd World Conference on Photovoltaic Energy Conversion, Osaka, Japan*, 2003, pp. 5–8.

[64] K. Ding, Z. Ye, and T. Reindl, “Comparison of Parameterisation Models for the Estimation of the Maximum Power Output of PV Modules,” *Energy Procedia*, vol. 25, pp. 101–107, Jan. 2012.

[65] S. Dittmann *et al.*, “Results Of The Third Modelling Round Robin Within The European Project „Performance” – Comparison Of Module Energy Rating Methods,” in *25th European Photovoltaic Solar Energy Conference and Exhibition / 5th World Conference on Photovoltaic Energy Conversion, Valencia, Spain, September 2010*, 2010, pp. 4333–4338.

[66] G. Friesen, D. Chianese, S. Rezzonico, A. Realini, N. Cereghetti, and E. Bura, “Matrix method for energy rating calculations of PV modules,” in *PV in Europe: From PV Technology to Energy Solutions*, 2002, no. November 2015, pp. 3–6.

[67] G. Friesen *et al.*, “Intercomparison of Different Energy Prediction Methods within the European Project ‘Performance’ - Results on the 2nd Round Robin,” *24th EU-PVSEC*, no. September, pp. 3189–3197, 2009.

[68] R. Platon, J. Martel, N. Woodruff, and T. Y. Chau, “Online Fault Detection in PV Systems,” *IEEE Trans. Sustain. ENERGY*, vol. 6, no. 4, pp. 1200–1207, 2015.

[69] J. A. Kratochvil, W. E. Boyson, and D. L. King, “Photovoltaic array performance model,” SAND2004-3535, Sandia National Laboratories (SNL), Aug. 2004.

[70] T. Huld *et al.*, “A power-rating model for crystalline silicon PV modules,” *Sol. Energy Mater. Sol. Cells*, vol. 95, no. 12, pp. 3359–3369, 2011.

[71] T. Huld, E. Salis, A. Pozza, W. Herrmann, and H. Müllejans, “Photovoltaic energy rating data sets for Europe,” *Sol. Energy*, vol. 133, pp. 349–362, 2016.

[72] W. N. Macêdo and R. Zilles, “Operational results of grid-connected photovoltaic system with different inverter’s sizing factors (ISF),” *Prog. Photovoltaics Res. Appl.*, vol. 15, no. 4, pp. 337–352, Jun. 2007.
[73] B. Bletterie, R. Bründlinger, and G. Lauss, “On the characterisation of PV inverters’ efficiency-introduction to the concept of achievable efficiency,” *Prog. Photovoltaics Res. Appl.*, vol. 19, no. 4, pp. 423–435, Jun. 2011.
[74] B. Goss, I. Cole, T. Betts, and R. Gottschalg, “Irradiance modelling for individual cells of shaded solar photovoltaic arrays,” *Sol. Energy*, vol. 110, pp. 410–419, 2014.
[75] T. Nordmann and L. Clavadetscher, “Understanding temperature effects on PV system performance,” *Energy Convers.*, vol. 3, pp. 2–5, 2003.
[76] J. K. Kaldellis, M. Kapsali, and K. a. Kavadias, “Temperature and wind speed impact on the efficiency of PV installations. Experience obtained from outdoor measurements in Greece,” *Renew. Energy*, vol. 66, pp. 612–624, Jun. 2014.
[77] S. a. Kalogirou, R. Agathokleous, and G. Panayiotou, “On-site PV characterization and the effect of soiling on their performance,” *Energy*, vol. 51, pp. 439–446, Mar. 2013.
[78] M. Köntges *et al.*, “Review of Failures of Photovoltaic Modules,” 2014.
[79] R. Kaplar *et al.*, “PV inverter performance and reliability: What is the role of the IGBT?,” *2011 37th IEEE Photovolt. Spec. Conf.*, pp. 001842–001847, Jun. 2011.
[80] R. Gottschalg, T. R. Betts, A. Eeles, S. R. Williams, and J. Zhu, “Influences on the energy delivery of thin film photovoltaic modules,” *Sol. Energy Mater. Sol. Cells*, vol. 119, pp. 169–180, Dec. 2013.
[81] D. Dirnberger, G. Blackburn, B. Müller, and C. Reise, “On the impact of solar spectral irradiance on the yield of different PV technologies,” *Sol. Energy Mater. Sol. Cells*, vol. 132, pp. 431–442, 2014.
[82] A. Virtuani, D. Strepparava, and G. Friesen, “A simple approach to model the performance of photovoltaic solar modules in operation,” *Sol. Energy*, vol. 120, pp. 439–449, 2015.
[83] A. Virtuani, D. Pavanello, and G. Friesen, “Overview of Temperature Coefficients of Different Thin Film Photovoltaic Technologies,” in *25th European Photovoltaic Solar Energy Conference and Exhibition / 5th World Conference on Photovoltaic Energy Conversion, 6-10 September 2010, Valencia, Spain*, 2010, no. JANUARY, pp. 4248–4252.
[84] E. Skoplaki and J. a. Palyvos, “On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations,” *Sol. Energy*, vol. 83, no. 5, pp. 614–624, 2009.
[85] A. Woyte, M. Richter, D. Moser, M. Green, S. Mau, and H. G. Beyer, “Analytical Monitoring of Grid-connected Photovoltaic Systems,” IEA-PVPS T13-03:2014, 2014.
[86] F. Jackson, *Planning and installing Photovoltaic Systems*, 2nd ed. London: Earthscan, 2007.
[87] M. Díez-Mediavilla, C. Alonso-Tristán, M. C. Rodríguez-Amigo, T. García-Calderón, and M. I. Dieste-Velasco, “Performance analysis of PV plants: Optimization for improving profitability,” *Energy Convers. Manag.*, vol. 54, no. 1, pp. 17–23, Feb. 2012.
[88] E. Kymakis, S. Kalykakis, and T. M. Papazoglou, “Performance analysis of a grid connected photovoltaic park on the island of Crete,” *Energy Convers. Manag.*, vol. 50, no. 3, pp. 433–438, Mar. 2009.
[89] V. Sharma and S. S. Chandel, “Performance and degradation analysis for long term reliability of solar photovoltaic systems: A review,” *Renew. Sustain. Energy Rev.*, vol. 27, pp. 753–767, Nov. 2013.

[90] G. G. Pillai, G. A. Putrus, T. Georgitsioti, and N. M. Pearsall, “Near-term economic benefits from grid-connected residential PV (photovoltaic) systems,” *Energy*, vol. 68, pp. 832–843, 2014.

[91] M. Perdue and R. Gottschalg, “Energy yields of small grid connected photovoltaic system: effects of component reliability and maintenance,” *IET Renew. Power Gener.*, vol. 9, no. 5, pp. 432–437, Jul. 2015.

[92] J. Leloux, L. Narvarte, and D. Trebosc, “Review of the performance of residential PV systems in France,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 2, pp. 1369–1376, 2012.

[93] J. Leloux, L. Narvarte, and D. Trebosc, “Review of the performance of residential PV systems in Belgium,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 1, pp. 178–184, Jan. 2012.

[94] J. Taylor, J. Leloux, A. M. Everard, J. Briggs, and A. Buckley, “Monitoring thousands of distributed PV systems in the UK : Energy production and performance,” in *11th Photovoltaic Science Application and Technology (PVSAT-11)*, 2015, pp. 77–80.

[95] U. Jahn and W. Nasse, “Operational performance of grid-connected PV systems on buildings in Germany,” *Prog. Photovoltaics Res. Appl.*, vol. 12, no. 6, pp. 441–448, Sep. 2004.

[96] M. Díez-Mediavilla, M. I. Dieste-Velasco, M. C. Rodríguez-Amigo, T. García-Calderón, and C. Alonso-Tristán, “Performance of grid-tied PV facilities: A case study based on real data,” *Energy Convers. Manag.*, vol. 76, pp. 893–898, Dec. 2013.

[97] S. Mau and U. Jahn, “Performance analysis of grid-connected PV systems,” in *21st European Photovoltaic Solar Energy Conference*, 2006, vol. 43, no. 0, pp. 4–8.

[98] J. Leloux *et al.*, “Monitoring 30 ,000 PV systems in Europe : Performance , Faults, and State of the Art,” in *31st European Photovoltaic Solar Energy Conference and Exhibition*, 2015, no. September, pp. 1574–1582.

[99] B. Decker and U. Jahn, “Performance of 170 grid connected PV plants in northern Germany - Analysis of yields and optimization potentials,” in *Solar Energy*, 1997, vol. 59, no. 4–6–6 pt 4, pp. 127–133.

[100] IEA-PVPS T7-08, “Reliability Study of Grid Connected PV Systems Field Experience and Recommended Design Practice Task 7,” 2002.

[101] A. Woyte, M. Richter, D. Moser, M. Green, S. Mau, and H. G. Beyer, “Analytical Monitoring of Grid-connected Photovoltaic Systems,” Brussels, 2014.

[102] A. Woyte, M. Richter, D. Moser, S. Mau, N. H. Reich, and U. Jahn, “Monitoring of Photovoltaic Systems: Good Practices and Systematic Analyes,” *28th Eur. PV Sol. Energy Conf. Exhib.*, 2013.

[103] M. a. Eltawil and Z. Zhao, “Grid-connected photovoltaic power systems: Technical and potential problems-A review,” *Renew. Sustain. Energy Rev.*, vol. 14, no. 1, pp. 112–129, 2010.

[104] S. K. Firth, K. J. Lomas, and S. J. Rees, “A simple model of PV system performance and its use in fault detection,” *Sol. Energy*, vol. 84, no. 4, pp. 624–635, Apr. 2010.

[105] A. Drews, H. G. Beyer, and U. Rindelhardt, “Quality of performance assessment of PV plants based on irradiation maps,” *Sol. Energy*, vol. 82, no. 11, pp. 1067–1075, Nov. 2008.

[106] J. Polo, “Solar global horizontal and direct normal irradiation maps in Spain derived from geostationary satellites,” *J. Atmos. Solar-Terrestrial Phys.*, vol. 130–131, pp. 81–88, 2015.

[107] N. Hofstra, M. Haylock, M. New, P. Jones, and C. Frei, “Comparison of six methods for the interpolation of daily, European climate data,” *J. Geophys. Res.*, vol. 113, no. D21,

pp. 1–19, Nov. 2008.
[108] M. Suri and T. Cebecauer, “Satellite-Based Solar Resource Data: Model Validation Statistics Versus User’S Uncertainty,” *ASES Sol. 2014 Conf.*, no. July, pp. 7–9, 2014.
[109] T. Huld, E. Dunlop, H. G. Beyer, and R. Gottschalg, “Data sets for energy rating of photovoltaic modules,” *Sol. Energy*, vol. 93, pp. 267–279, Jul. 2013.
[110] T. Cebecauer and M. Suri, “Site-adaptation of satellite-based DNI and GHI time series: Overview and SolarGIS approach,” *AIP Conf. Proc.*, vol. 1734, 2016.
[111] S. Stettler, P. Toggweiler, and J. Remund, “SPYCE: SATELLITE PHOTOVOLTAIC YIELD CONTROL AND EVALUATION,” in *21st European Photovoltaic Solar Energy Conference*, 2006, no. September, pp. 2613–2616.
[112] J. D. Mondol, Y. G. Yohanis, and B. Norton, “Solar radiation modelling for the simulation of photovoltaic systems,” *Renew. Energy*, vol. 33, no. 5, pp. 1109–1120, 2008.
[113] C. A. Gueymard, “Direct and indirect uncertainties in the prediction of tilted irradiance for solar engineering applications,” *Sol. Energy*, vol. 83, no. 3, pp. 432–444, 2009.
[114] M. Lave, W. Hayes, A. Pohl, and C. W. Hansen, “Evaluation of Global Horizontal Irradiance to Plane-of-Array Irradiance Models at Locations Across the United States,” *IEEE J. Photovoltaics*, vol. 5, no. 2, pp. 597–606, Mar. 2015.
[115] Y. Xie and M. Sengupta, “Diagnosing Model Errors in Simulation of Solar Radiation on Inclined Surfaces,” *43rd IEEE Photovolt. Spec. Conf.*, pp. 1022–1025, 2016.
[116] J. K. Copper, a. B. Sproul, and S. Jarnason, “Photovoltaic (PV) performance modelling in the absence of onsite measured plane of array irradiance (POA) and module temperature,” *Renew. Energy*, vol. 86, pp. 760–769, 2016.
[117] D. Erbs, S. Klein, and J. Duffie, “Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation,” *Sol. energy*, vol. 28, no. 4, 1982.
[118] D. T. Reindl, W. A. Beckman, and J. A. Duffle, “Diffuse fraction correlations,” no. I, 1990.
[119] B. Ridley, J. Boland, and P. Lauret, “Modelling of diffuse solar fraction with multiple predictors,” *Renew. Energy*, vol. 35, no. 2, pp. 478–483, Feb. 2010.
[120] I. Reda and A. Andreas, “Solar position algorithm for solar radiation applications,” Golden, Colorado, United States, 2008.
[121] D. Reindl, W. Beckman, and J. Duffie, “Evaluation of hourly tilted surface radiation models,” *Sol. Energy*, vol. 45, no. 1, pp. 9–17, 1990.
[122] P. M. Segado, J. Carretero, and M. Sidrach-de-Cardona, “Models to predict the operating temperature of different photovoltaic modules in outdoor conditions,” *Prog. Photovoltaics Res. Appl.*, vol. 23, no. 10, pp. 1267–1282, 2014.
[123] E. Skoplaki, A. G. Boudouvis, and J. A. Palyvos, “A simple correlation for the operating temperature of photovoltaic modules of arbitrary mounting,” *Sol. Energy Mater. Sol. Cells*, vol. 92, no. 11, pp. 1393–1402, 2008.
[124] J. Ross, R. G., “Interface design considerations for terrestrial solar cell modules,” in *12th IEEE Photovoltaic Specialist Conference*, 1976, pp. 801–806.
[125] E. Skoplaki, A. G. Boudouvis, and J. A. Palyvos, “A simple correlation for the operating temperature of photovoltaic modules of arbitrary mounting,” *Sol. Energy Mater. Sol. Cells*, vol. 92, no. 11, pp. 1393–1402, 2008.
[126] E. Skoplaki and J. A. Palyvos, “Operating temperature of photovoltaic modules: A survey of pertinent correlations,” *Renew. Energy*, vol. 34, no. 1, pp. 23–29, Jan. 2009.
[127] A. J. Veldhuis, A. M. Nobre, I. M. Peters, T. Reindl, R. Ruther, and A. H. M. E. Reinders, “An Empirical Model for Rack-Mounted PV Module Temperatures for Southeast Asian Locations Evaluated for Minute Time Scales,” *IEEE J. Photovoltaics*, vol. 5, no. 3, pp.

774–782, May 2015.
[128] M. Fuentes, “A simplified thermal model for flat-plate photovoltaic arrays,” New Mexico, USA, 1987.
[129] M. Mattei, G. Notton, C. Cristofari, M. Muselli, and P. Poggi, “Calculation of the polycrystalline PV module temperature using a simple method of energy balance,” *Renew. Energy*, vol. 31, no. 4, pp. 553–567, 2006.
[130] D. Infield, L. Mei, and U. Eicker, “Thermal performance estimation for ventilated PV facades,” *Sol. Energy*, vol. 76, no. 1–3, pp. 93–98, 2004.
[131] B. Goss, “Design process optimisation of solar photovoltaic systems,” Loughborough University, 2015.
[132] A. Livera, A. Phinikarides, M. Theristis, G. Makrides, and G. E. Georghiou, “Impact of missing data on the estimation of photovoltaic degradation rate,” in *NREL/SNL/BNL Photovoltaic Reliability Workshops*, 2017.
[133] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian, and R. Lyons, “Decision tree-based fault detection and classification in solar photovoltaic arrays,” *2012 Twenty-Seventh Annu. IEEE Appl. Power Electron. Conf. Expo.*, pp. 93–99, 2012.
[134] Syafaruddin, E. Karatepe, and T. Hiyama, “Controlling of artificial neural network for fault diagnosis of photovoltaic array,” *2011 16th Int. Conf. Intell. Syst. Appl. to Power Syst.*, pp. 1–6, 2011.
[135] K.-H. Chao, S.-H. Ho, and M.-H. Wang, “Modeling and fault diagnosis of a photovoltaic system,” *Electr. Power Syst. Res.*, vol. 78, no. 1, pp. 97–105, Jan. 2008.
[136] W. Chine, A. Mellit, A. M. Pavan, and S. A. Kalogirou, “Fault detection method for grid-connected photovoltaic plants,” *Renew. Energy*, vol. 66, pp. 99–110, 2014.
[137] M. Tadj, K. Benmouiza, A. Cheknane, and S. Silvestre, “Improving the performance of PV systems by faults detection using GISTEL approach,” *Energy Convers. Manag.*, vol. 80, pp. 298–304, 2014.
[138] a. Chouder and S. Silvestre, “Automatic supervision and fault detection of PV systems based on power losses analysis,” *Energy Convers. Manag.*, vol. 51, no. 10, pp. 1929–1937, Oct. 2010.
[139] S. Silvestre, M. A. Da Silva, A. Chouder, D. Guasch, and E. Karatepe, “New procedure for fault detection in grid connected PV systems based on the evaluation of current and voltage indicators,” *Energy Convers. Manag.*, vol. 86, pp. 241–249, 2014.
[140] Y. Hu, B. Gao, X. Song, G. Y. Tian, K. Li, and X. He, “Photovoltaic fault detection using a parameter based model,” *Sol. Energy*, vol. 96, pp. 96–102, 2013.
[141] C. Ventura and G. M. Tina, “Utility scale photovoltaic plant indices and models for on-line monitoring and fault detection purposes,” *Electr. Power Syst. Res.*, vol. 136, pp. 43–56, 2016.
[142] S. Silvestre, A. Chouder, and E. Karatepe, “Automatic fault detection in grid connected PV systems,” *Sol. Energy*, vol. 94, pp. 119–127, Aug. 2013.
[143] J. D. Mondol, Y. G. Yohanis, and B. Norton, “Optimal sizing of array and inverter for grid-connected photovoltaic systems,” *Sol. Energy*, vol. 80, no. 12, pp. 1517–1539, 2006.
[144] a. Chouder and S. Silvestre, “Analysis Model of Mismatch Power Losses in PV Systems,” *J. Sol. Energy Eng.*, vol. 131, no. 2, p. 24504, 2009.
[145] S. Stettler, “Failure Detection Routine for Grid-Connected PV Systems,” in *Proc. 20th European Photovoltaic Solar Energy …*, 2005, pp. 2490–2493.
[146] F. Ferroni, A. Guekos, and R. J. Hopkirk, “Further considerations to: Energy Return on

Energy Invested (ERoEI) for photovoltaic solar systems in regions of moderate insolation," *Energy Policy*, vol. 107, pp. 498–505, 2017.
[147] M. Raugei *et al.*, "Energy Return on Energy Invested (ERoEI) for photovoltaic solar systems in regions of moderate insolation: A comprehensive response," *Energy Policy*, vol. 102, no. January, pp. 377–384, 2017.
[148] "Nottingham City Homes." [Online]. Available: http://www.nottinghamcityhomes.org.uk/. [Accessed: 25-Jan-2017].
[149] MetOffice, "Met Office Integrated Data Archive System (MIDAS)," *NCAS British Atmospheric Data Centre*. 2013.
[150] P. Ineichen and R. Perez, "A new airmass independent formulation for the linke turbidity coefficient," *Sol. Energy*, vol. 73, no. 3, pp. 151–157, 2002.
[151] K. Gibson, I. R. Cole, B. Goss, T. R. Betts, and R. Gottschalg, "Compensation of temporal averaging bias in solar irradiance data," *IET Renew. Power Gener.*, pp. 1–7, 2017.
[152] Z. Sen, *Solar energy fundamentals and modeling techniques*. 2008.
[153] D. Palmer, I. Cole, T. R. Betts, and R. Gottschalg, "Interpolating and estimating horizontal diffuse solar irradiation to provide UK-wide coverage: selection of the best performing models," *Energies*, 2017.
[154] P. Rowley *et al.*, "Multi-domain analysis of photovoltaic impacts via integrated spatial and probabilistic modelling," *IET Renew. Power Gener.*, vol. 9, no. 5, pp. 424–431, 2015.
[155] J. Zhu, T. Betts, and R. Gottschalg, "Accuracy Assessment of Models Estimating Total Irradiance on Inclined Planes in Loughborough," in *4th Photovoltaic Science Application and Technology (PVSAT-4)*, 2008, pp. 207–210.
[156] J. E. Hay, R. Perez, and D. C. McKay, "'Estimating Solar Irradiance on Inclined Surfaces: A Review and Assessment of Methodologies,'" *Int. J. Sol. Energy*, vol. 4, no. 5, pp. 321–324, 1986.
[157] I. Reda and A. Andreas, "Solar position algorithm for solar radiation applications," *Sol. Energy*, vol. 76, no. 5, pp. 577–589, 2004.
[158] R. Gottschalg, T. R. Betts, a. Eeles, S. R. Williams, and J. Zhu, "Influences on the energy delivery of thin film photovoltaic modules," *Sol. Energy Mater. Sol. Cells*, vol. 119, pp. 169–180, Dec. 2013.
[159] T. Dierauf, A. Growitz, S. Kurtz, and C. Hansen, "Weather-Corrected Performance Ratio," 2013.
[160] N. H. Reich, B. Mueller, A. Armbruster, W. G. J. H. M. van Sark, K. Kiefer, and C. Reise, "Performance ratio revisited: is PR > 90% realistic?," *Prog. Photovoltaics Res. Appl.*, vol. 20, no. 6, pp. 717–726, Sep. 2012.
[161] T. Huld and A. M. Gracia Amillo, "Estimating PV module performance over large geographical regions: The role of irradiance, air temperature, wind speed and solar spectrum," *Energies*, vol. 8, no. 6, pp. 5159–5181, 2015.
[162] M. Šúri, T. a. Huld, E. D. Dunlop, and H. a. Ossenbrink, "Potential of solar electricity generation in the European Union member states and candidate countries," *Sol. Energy*, vol. 81, no. 10, pp. 1295–1305, 2007.
[163] B. Iglewicz and D. Hoaglin, "Volume 16: How to Detect and Handle Outliers," in *The ASQC Basic References in Quality Control: Statistical Techniques*, vol. 16, 1993.
[164] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.
[165] B. Goss, *Choosing Solar Electricity: A Guide to Photovoltaic System*. Centre for

Alternative Technology, 2010.
[166] D. Palmer, I. Cole, T. Betts, and R. Gottschalg, “Assessment of potential for photovoltaic roof installations by extraction of roof tilt from light detection and ranging data and aggregation to census geography,” *IET Renew. Power Gener.*, vol. 10, no. 4, pp. 467–473, 2016.
[167] T. R. Betts, “Investigation of photovoltaic device operation under varying spectral conditions,” 2005.
[168] D. Palmer, I. Cole, T. Betts, and R. Gottschalg, “Interpolating and Estimating Horizontal Diffuse Solar Irradiation to Provide UK-Wide Coverage: Selection of the Best Performing Models,” *Energies*, vol. 10, no. 2, p. 181, Feb. 2017.
[169] A. J. Carr and T. L. Pryor, “A comparison of the performance of different PV module types in temperate climates,” *Sol. Energy*, vol. 76, no. 1–3, pp. 285–294, Jan. 2004.
[170] M. I. A. Lourakis, “A Brief Description of the Levenberg-Marquardt Algorithm Implemened by levmar,” *Institute of Computer Science, Foundation for Research and Technology - Hellas, 2005*. pp. 1–6.
[171] F. Almonacid, C. Rus, P. Pérez-Higueras, and L. Hontoria, “Calculation of the energy provided by a PV generator. Comparative study: Conventional methods vs. artificial neural networks,” *Energy*, vol. 36, no. 1, pp. 375–384, Jan. 2011.
[172] F. Almonacid, C. Rus, P. J. Pérez, and L. Hontoria, “Estimation of the energy of a PV generator using artificial neural network,” *Renew. Energy*, vol. 34, no. 12, pp. 2743–2750, Dec. 2009.
[173] F. Bonanno, G. Capizzi, G. Graditi, C. Napoli, and G. M. Tina, “A radial basis function neural network based approach for the electrical characteristics estimation of a photovoltaic module,” *Appl. Energy*, vol. 97, pp. 956–961, 2012.
[174] A. Mellit and A. M. Pavan, “Performance prediction of 20kWp grid-connected photovoltaic plant at Trieste (Italy) using artificial neural network,” *Energy Convers. Manag.*, vol. 51, no. 12, pp. 2431–2441, Dec. 2010.
[175] R. Tapakis, S. Michaelides, and a. G. Charalambides, “Computations of diffuse fraction of global irradiance: Part 2 - Neural Networks,” *Sol. Energy*, 2015.
[176] A. Mellit and A. M. Pavan, “A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy,” *Sol. Energy*, vol. 84, no. 5, pp. 807–821, 2010.
[177] C. Renno, F. Petito, and a. Gatto, “Artificial neural network models for predicting the solar radiation as input of a concentrating photovoltaic system,” *Energy Convers. Manag.*, vol. 106, pp. 999–1012, 2015.
[178] A. Azadeh, A. Maghsoudi, and S. Sohrabkhani, “An integrated artificial neural networks approach for predicting global radiation,” *Energy Convers. Manag.*, vol. 50, no. 6, pp. 1497–1505, 2009.
[179] G. Notton, C. Paoli, S. Vasileva, M. L. Nivet, J. L. Canaletti, and C. Cristofari, “Estimation of hourly global solar irradiation on tilted planes from horizontal one using artificial neural networks,” *Energy*, vol. 39, no. 1, pp. 166–179, 2012.
[180] A. N. Celik and T. Muneer, “Neural network based method for conversion of solar radiation data,” *Energy Convers. Manag.*, vol. 67, pp. 117–124, 2013.
[181] H. M. Cartwright, “Artificial Neural Networks in Biology and Chemistry—The Evolution of a New Analytical Tool,” in *Artificial Neural Networks*, 2008, pp. 1–13.
[182] A. Colli, “Failure mode and effect analysis for photovoltaic systems,” *Renew. Sustain. Energy Rev.*, vol. 50, pp. 804–809, 2015.

[183] O. Hachana, G. M. Tina, and K. E. Hemsas, “PV array fault diagnostic technique for BIPV systems,” *Energy Build.*, vol. 126, pp. 263–274, 2016.
[184] D. Sera, R. Teodorescu, and P. Rodriguez, “Photovoltaic module diagnostics by series resistance monitoring and temperature and rated power estimation,” *Proc. - 34th Annu. Conf. IEEE Ind. Electron. Soc. IECON 2008*, pp. 2195–2199, 2008.
[185] California Energy Commission (CEC), “Lists of Eligible Equipment,” *CEC website*. 2017.
[186] PHOTON International GmbH, “PHOTON Databases.” 2017.
[187] A. Mermoud, “PVSYST: a user-friendly software for PV systems simulation ,” *12th European Photovoltaic Solar Energy Conference*. James and James Science Publishers, Amsterdam, Netherlands, 1994.
[188] M. Villalva, J. Gazoli, and E. Filho, “Comprehensive Approach to Modeling and Simulation of Photovoltaic Arrays,” *IEEE Trans. Power Electron.*, vol. 24, no. 5, pp. 1198–1208, 2009.
[189] I. Balouktsis, J. Zhu, R. Bründlinger, T. R. Betts, and R. Gottschalg, “Optimised Inverter Sizing in the UK,” in *Proceedings of the 4th Photovoltaic Science, Application and Technology (PVSAT) Conference*, 2008, p. tbc.
[190] T. Georgitsioti, N. Pearsall, and I. Forbes, “Simplified levelised cost of the domestic photovoltaic energy in the UK: the importance of the feed-in tariff scheme,” *IET Renew. Power Gener.*, vol. 8, no. 5, pp. 451–458, 2014.

Publications and achievements

Book chapter

Brian Goss, Ian Cole, Elena Koubli, Diane Palmer, Tom Betts, Ralph Gottschalg: *Modelling and prediction of PV module energy yield*. The Performance of Photovoltaic (PV) Systems: Modelling, Measurement and Assessment, Edited by Nicola Pearsall, 11/2016: chapter 4: pages 103-132; Elsevier Science & Technology, ISBN: 9781782423362, DOI:10.1016/B978-1-78242-336-2.00004-5

Journal publications

E. Koubli, D. Palmer, P. Rowley, R. Gottschalg: *Inference of missing data in photovoltaic monitoring datasets*. IET Renewable Power Generation 01/2016; DOI:10.1049/iet-rpg.2015.0355

R. Urraca, A. M. Gracia-Amillo, E. Koubli, T. Huld, J. Trentmann, A. Riihelä, A. V. Lindfors, D. Palmer, R. Gottschalg, and F. Antonanzas-Torres, "Extensive validation of CM SAF surface radiation products over Europe," *Remote Sens. Environ.*, vol. 199, pp. 171–186, Sep. 2017.

D. Palmer, E. Koubli, T. Betts, R. Gottschalg*: The UK Solar Farm Fleet: a challenge for the National Grid?* Energies, 2017.

D. Palmer, E. Koubli, I. Cole, T. Betts, R. Gottschalg: *Comparison of Solar Radiation and PV Generation Variability: System Dispersion in the UK*. IET Renewable Power Generation 03/2017; DOI:10.1049/iet-rpg.2016.0768

Conference publications

Elena Koubli, Diane Palmer, Paul Rowley, Ralph Gottschalg: *Remote monitoring and failure detection for distributed small-scale PV systems.* 13th Photovoltaic Science Application and Technology (PVSAT-13), Bangor, UK; 04/2017

Elena Koubli, Diane Palmer, Tom Betts, Paul Rowley, Ralph Gottschalg: *Inference of Missing PV Monitoring Data using Neural Networks*. 43rd IEEE Photovoltaic Specialists Conference, Portland, USA; 06/2016, DOI:10.1109/PVSC.2016.7750305

Elena Koubli, Diane Palmer, Paul Rowley, Ralph Gottschalg: *Assessment of PV System Performance with Incomplete Monitoring Data*. 31st European Photovoltaic Solar Energy Conference and Exhibition, Hamburg, Germany; 09/2015

Elena Koubli, Diane Palmer, Paul Rowley, Ralph Gottschalg: *Replenishing Deficient Datasets in PV System Monitoring*. 11th Photovoltaic Science Application and Technology (PVSAT-11), Leeds, UK; 04/2015

Elena Koubli, Diane Palmer, Paul Rowley, Ralph Gottschalg: *Investigating Two thousand PV rooftop Systems in the UK: Performance Analysis and Fault Diagnosis.* 12th Photovoltaic Science Application and Technology (PVSAT-12), Liverpool, UK; 04/2016

Diane Palmer, Elena Koubli, I. Cole, T. Betts Ralph Gottschalg: *Satellite or Ground-based Irradiation Data: which is closer to reality?.* 13th Photovoltaic Science Application and Technology (PVSAT-12), Bangor, UK; 04/2017

Diane Palmer, Elena Koubli, Ralph Gottschalg: *Space and Time Analysis of Irradiation variation cross the UK: A 10 Year Study of Solar Farm Yield.* 12th Photovoltaic Science Application and Technology (PVSAT-12), Liverpool, UK; 04/2016

Diane Palmer, Elena Koubli, Paul Rowley, Ralph Gottschalg: *Comparison of Solar Radiation and PV Generation Variability with System Dispersion in the UK:* 12th Photovoltaic Science Application and Technology (PVSAT-12), Liverpool, UK; 04/2016

Ian Cole, Diane Palmer, Brian Goss, Elena Koubli, Tom Betts, Murray Thomson and Ralph Gottschalg. *Impact Analysis of Irradiance Dataset Selection on Photovoltaic System Energy Yield Modelling*, in: 1st International Conference on Large-Scale Grid Integration of Renewable Energy in India. New Dehli, India; 09/2017

Workshop presentations

2nd International Workshop in Energy Generation of PV Systems –CREST, September 2014

E. Koubli and D. Palmer "*Predicting and Mapping UK PV Performance Variation*".

PV2025 workshop (EPSRC project: EP/K02227X/1)

E. Koubli "*Investigating Two Thousand PV Rooftops: From Data Quality to Performance Assessment and Fault Detection*".

Nominations, awards and proposals

- Awarded the mobility grant researcher (RMG) fund within the European project PhotoClass (March 2016).

- Awarded enterprise project proposal within Loughborough University with title: "*PV-DuDES: Photovoltaic Due Diligence Evaluation Services"* with Dr Tom Betts and Prof Ralph Gottschalg (April 2017), based on the work carried out on the Nottingham City Homes dataset.

- Nominated for the best student paper award in the 43rd IEEE Photovoltaic Specialists Conference, Portland, USA, 2016 for the paper with title: "*Inference of Missing PV Monitoring Data using Neural Networks*."