



## PV PERFORM MOD

# Methodology: Automatic PV System Fault Detection



Version 1.1  
2016

© Her Majesty the Queen in Right of Canada,  
as represented by the Minister of Natural  
Resources Canada, 2016

**Prepared by:**  
CanmetENERGY - Varennes

November 2016

**CanmetENERGY**  
*Leadership in ecoInnovation*

Canada 

**Disclaimer**

Natural Resources Canada (NRCan) is not responsible for the accuracy or completeness of the information contained in the reproduced material. NRCan shall at all times be indemnified and held harmless against any and all claims whatsoever arising out of negligence or other fault in the use of the information contained in this publication or product.

**Copyright**

Information contained in this publication or product may be reproduced, in part or in whole, and by any means, for personal or public non-commercial purposes, without charge or further permission, unless otherwise specified.

You are asked to:

- exercise due diligence in ensuring the accuracy of the materials reproduced;
- indicate the complete title of the materials reproduced, and the name of the author organization; and
- indicate that the reproduction is a copy of an official work that is published by Natural Resources Canada (NRCan) and that the reproduction has not been produced in affiliation with, or with the endorsement of, NRCan.

Commercial reproduction and distribution is prohibited except with written permission from NRCan. For more information, contact NRCan at [copyright.droitdauteur@nrcan-mcan.gc.ca](mailto:copyright.droitdauteur@nrcan-mcan.gc.ca).

# Table of Contents

1.	Introduction .....	1
1.1.	Required PV system measurements.....	2
2.	Data cleaning .....	3
2.1.	Data quality.....	3
2.2.	Data averaging .....	5
2.3.	Outliers .....	7
2.4.	Low irradiance and power values.....	9
2.5.	Irradiance intervals .....	9
2.6.	Measurements not representative of normal operation .....	9
2.7.	Summary of data cleaning steps.....	12
3.	AC power model .....	13
3.1.	Training and validation errors.....	13
4.	Fault Detection .....	15
4.1.	Normal operation limits.....	15
4.2.	Fault detection.....	15
5.	Performance ratio.....	18
5.1.	Recommendations for performance ratio calculation .....	20
5.2.	Performance ratio superior to 100% .....	20

## List of Figures

Figure 1: Power and irradiance measurements following a linear relationship .....	4
Figure 2: Power and irradiance measurements not following a linear relationship.....	5
Figure 3: 10-minute power and irradiance measurements .....	6
Figure 4: Hourly-averaged power and irradiance measurements .....	6
Figure 5: Abnormally high solar irradiance values .....	7
Figure 6: Abnormally high AC power values.....	8
Figure 7: Abnormally high PV panel temperatures .....	8
Figure 8: 1 <sup>st</sup> cleaning iteration.....	10
Figure 9: 7 <sup>th</sup> cleaning iteration .....	11
Figure 10: Training and validation errors .....	14
Figure 11: Faulty data and normal operation limits.....	16
Figure 12: Normal operation limits expressed in kW .....	17
Figure 13: Monthly performance ratios .....	19
Figure 14: Weekly performance ratios.....	19

## List of Tables

Table 1: Data cleaning iterations.....	11
--	----

# 1. Introduction

The software tool executes an automatic procedure for detecting PV system faults. Faults are defined as periods of abnormally low alternative current (AC) power production, when the system produces less power than it should at given operating conditions – solar irradiance and PV panel temperature.

This document describes the methodology and the calculations carried out automatically by the software tool.

The fault detection procedure consists in comparing the measured AC power of the current produced by the PV system to the power that the system should produce under normal (fault-free) operating conditions. Significant differences between those values – with respect to pre-determined thresholds – are considered as faults.

Fault-free power productions are calculated by a model computing the expected (normal) system AC power production. The model consists of one equation that uses solar irradiance and PV panel temperature measurements to predict the AC power production.

The performance ratio (PR) is also automatically calculated; the PR evaluates the efficiency of the PV system by comparing its real electricity production to the ideal (rated) production. The PR complements the model-based fault detection approach, by keeping track of the long-term performance of the PV system.

The fault detection procedure follows the following steps:

1. data cleaning prior to modelling – remove observations not representative of a normal PV system operation
2. development of the model predicting the AC power production that the system should have under fault-free operation
3. calculation of the PV system normal operation limits
4. identification of faults

The procedure used to identify faults is described in this report. The calculations presented here are carried out automatically by the software. The user only needs to upload a file containing the PV system measurements and enter a few parameters specific to the PV system from which the measurements were taken. For more information regarding the user input please refer to the *Help* manual.

The procedure presented here was developed and validated using data from several commercial PV systems. Plots of actual PV system data from these systems are used throughout the report to better illustrate various calculations.

## **1.1. Required PV system measurements**

The following PV system measurements are required for fault detection:

- AC power
- solar irradiance in the PV system plane
- PV panel temperature

The rated (design) direct current (DC) power is required for calculating the performance ratio.

## 2. Data cleaning

The AC power production model should represent the expected output of the PV system under normal operation conditions. Therefore, prior to model development, the data should be analyzed to identify and remove from the modeling dataset the measurements representative of a faulty PV system operation. The presence of faulty data can be caused by different factors, such as instrument or equipment malfunction. Lags can also be present in the dataset, due to improper synchronization between sampling times of different measurements.

Data quality has a direct effect on the accuracy of the fault detection procedure. In some extreme cases, the quality of the measurements is not high enough to produce a reliable model.

The data cleaning procedure contains the following steps:

1. verification of data quality
2. 1-hour averaging of original measurements
3. removal of outliers
4. removal of measurements not representative of a normal operation

Outliers represent measurements that do not follow the general trend of the data. In order to capture the behavior of the PV system according to different sunlight levels and increase the data cleaning accuracy, outliers are identified according to different irradiance intervals.

The following sections will present data quality issues and detail the data cleaning procedure.

### 2.1. Data quality

Data quality has a direct impact on the accuracy of the AC power production model used for detecting faults. In order to prevent the development of an inaccurate model caused by poor data quality, an initial verification is carried out before proceeding with the data cleaning. Data quality is measured in terms of the relationship between solar irradiance and power production: under normal conditions, the solar irradiance and power production measurements follow a strong linear relationship. The coefficient of determination  $R^2$ , indicating the strength of the linear relationship between the irradiance and power, is used to evaluate the data quality – an  $R^2$  value of 1 indicates that the data perfectly fit a straight regression line.

To ensure proper fault detection, only datasets with a  $R^2$  coefficient of at least 0.70 for irradiance and AC power measurements will be considered for proceeding with the fault detection procedure. As an example, Figure 1 shows a clean dataset containing irradiance and power measurements following a strong linear relationship. Figure 2 shows a plot containing a

dataset of poor quality, with a  $R^2$  coefficient of 0.37. This is caused by a low AC power measurement precision – 6 kW. Since the nominal AC power production of the PV system is 18 kW, the AC power measurement accuracy is only 33.3%, preventing the development of an accurate model.

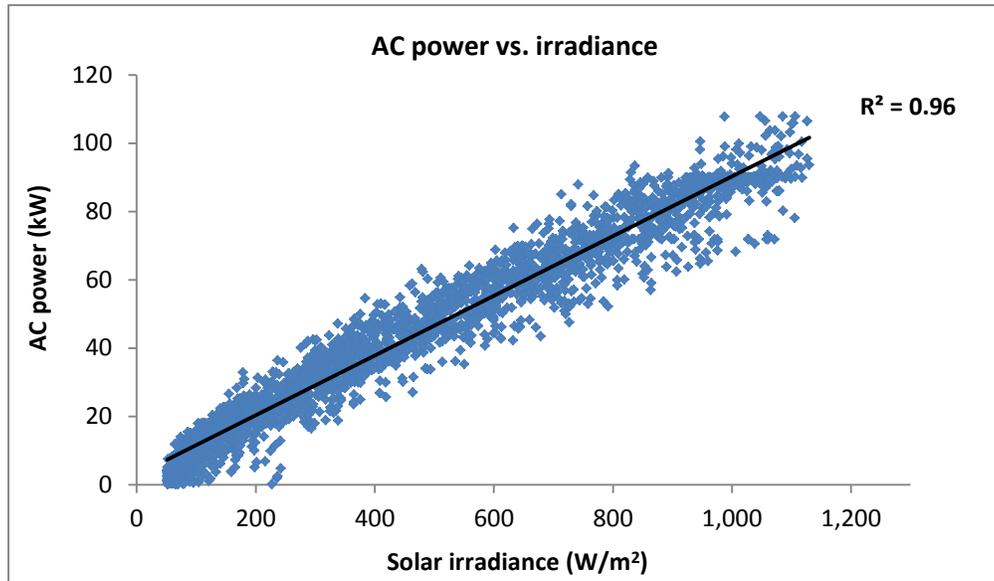


Figure 1: Power and irradiance measurements following a linear relationship

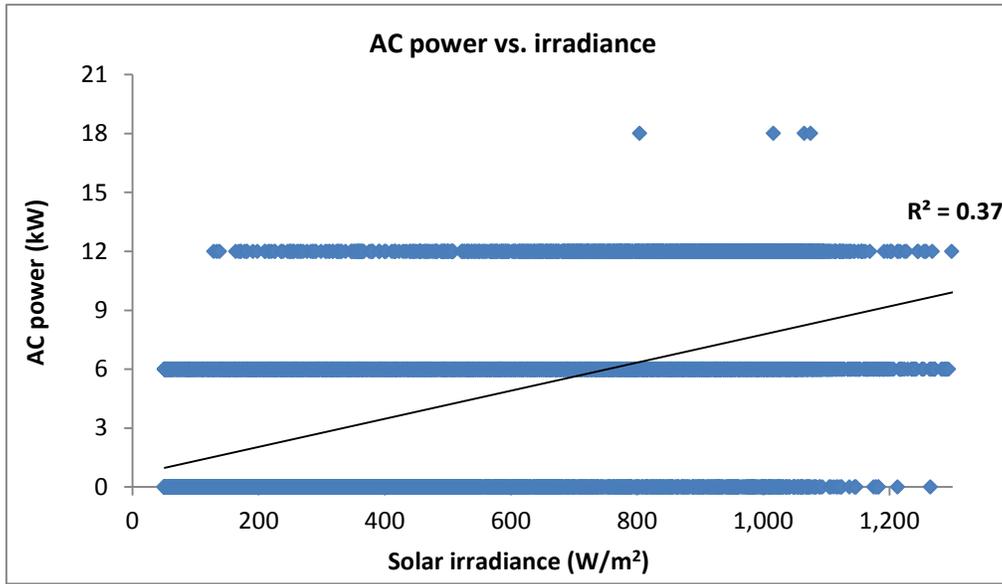


Figure 2: Power and irradiance measurements not following a linear relationship

## 2.2. Data averaging

Original measurements are averaged over 1 hour, in order to reduce measurement variability and increase model accuracy. Figure 3 shows the original measurements taken every 10 minutes, while Figure 4 shows the corresponding hourly-averaged measurements.

It can be seen that averaging the original measurement over one hour led to an increase of the  $R^2$  coefficient – from 0.78 to 0.91.

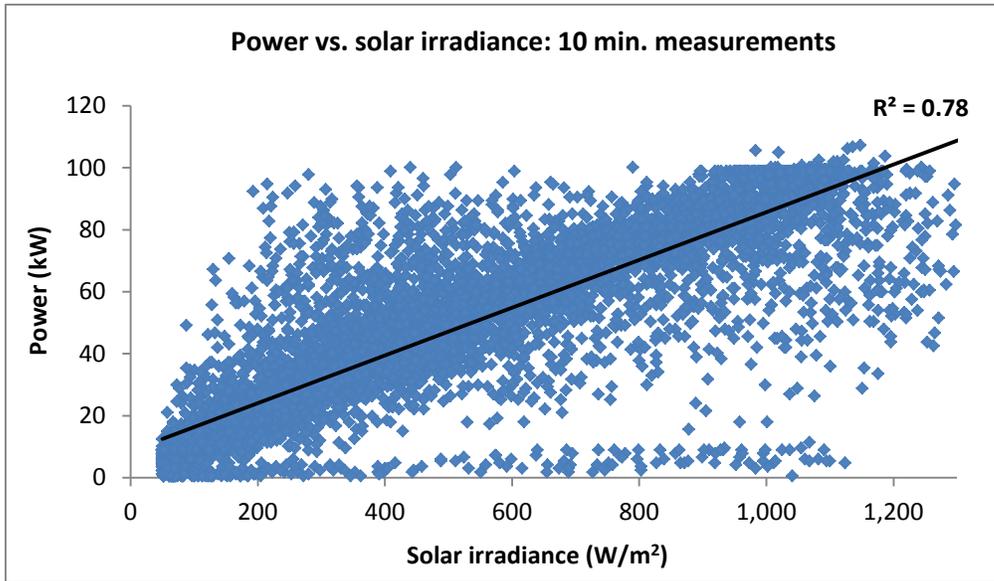


Figure 3: 10-minute power and irradiance measurements

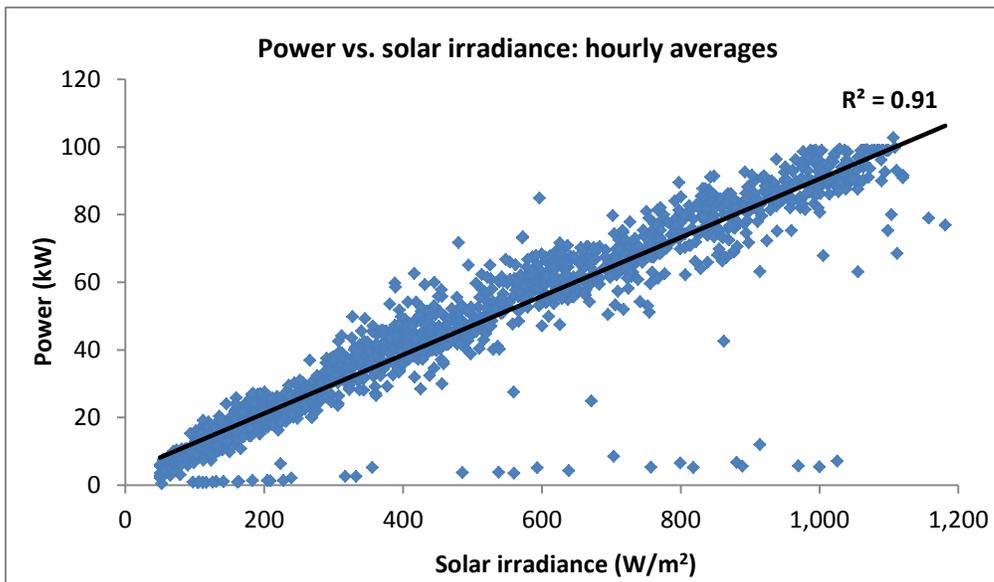


Figure 4: Hourly-averaged power and irradiance measurements

## 2.3. Outliers

Outliers represent measurements that do not follow the general trend of the data and they should be removed prior to model development. Outliers can be caused by measurement errors and instrument malfunction or miscalibration.

Three examples of outliers are given in this section: abnormally high values of solar irradiance, AC power and PV panel temperature.

Figure 5 shows that irradiance values greater than  $1,400 \text{ W/m}^2$  are measured for this particular PV system. The maximum solar irradiance in space, before entering Earth's atmosphere, is slightly superior to  $1,300 \text{ W/m}^2$ . Although higher irradiance values can be caused by light reflection, values greater than  $1,300 \text{ W/m}^2$  are eliminated from the dataset as a precaution to avoid the presence of erroneous measurements in the modelling dataset.

Figure 6 shows of measurements from a PV system with a maximum AC power production of 100 kW. However, values higher than 100 kW are present in the dataset.

Figure 7 shows a plot of PV panel temperatures; it can be seen that temperature reaches abnormally high values – around  $400^\circ\text{C}$ . Generally, the temperature of a PV panel under normal operation should not be higher than  $60^\circ\text{C}$ , in the case of rack-mounted systems, and  $90^\circ\text{C}$  in the case of building-integrated systems.

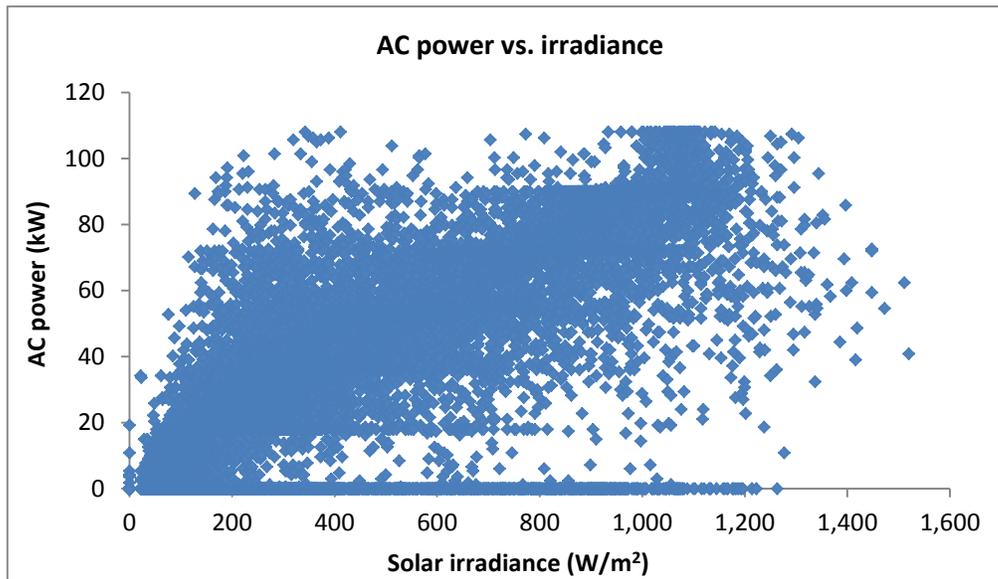


Figure 5: Abnormally high solar irradiance values

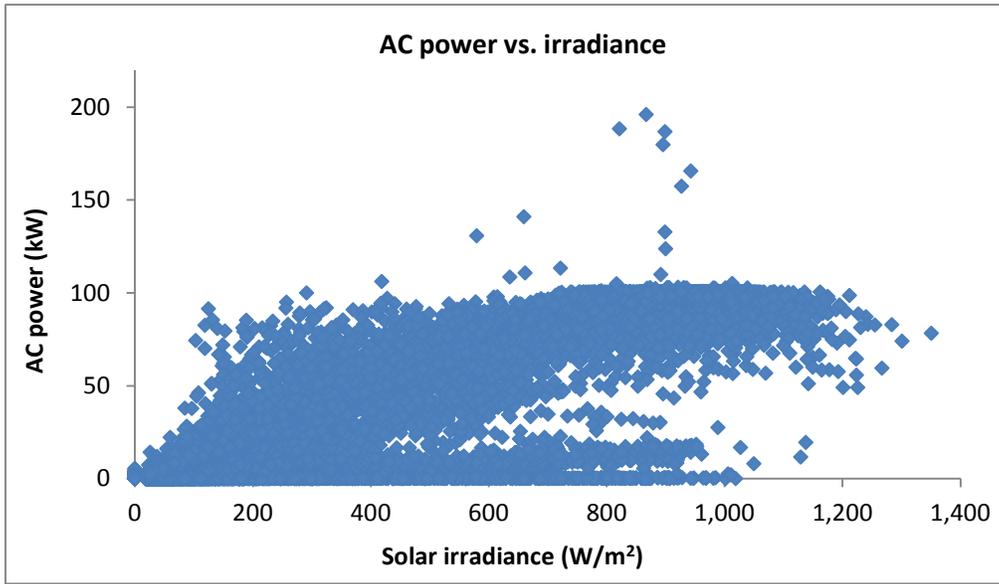


Figure 6: Abnormally high AC power values

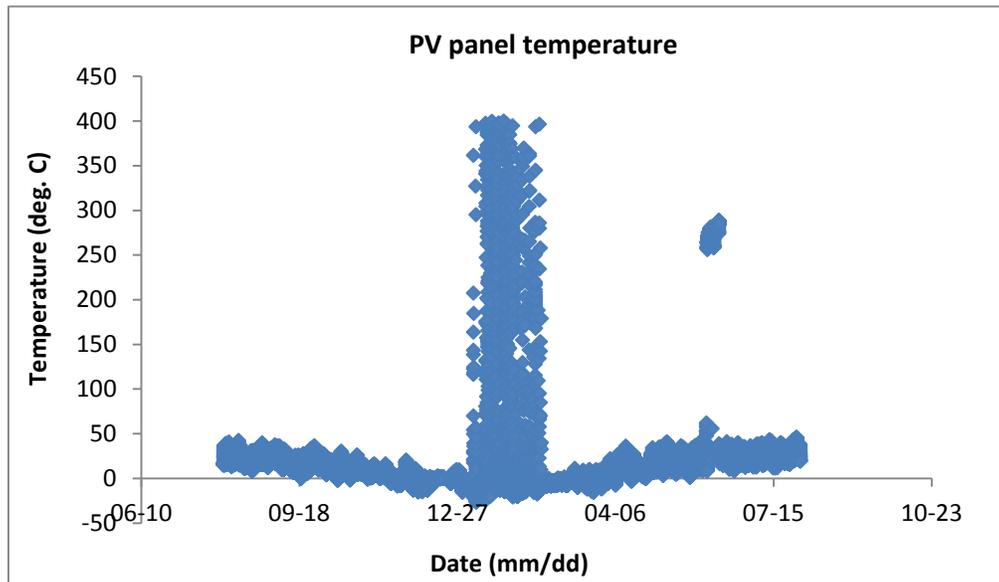


Figure 7: Abnormally high PV panel temperatures

## 2.4. Low irradiance and power values

Observations corresponding to irradiance levels less than  $50 \text{ W/m}^2$  are eliminated prior to model development, since the measurement accuracy is significantly reduced at very low sunlight levels.

Observations corresponding to very low power productions are also eliminated. Most low electricity productions are probably removed from the dataset when observations corresponding to irradiance levels less than  $50 \text{ W/m}^2$  are eliminated, but periods of faulty operation can result in low production despite high irradiance values. To simplify the subsequent outlier removal process, observations corresponding to AC power production less than 5% of nominal the PV system AC power production are eliminated from the dataset.

## 2.5. Irradiance intervals

Due to the technological quality of the solar modules, the efficiency of a PV system is dependent on the light intensity levels. At low irradiance levels – generally below  $300 \text{ W/m}^2$  – the efficiency is low; the efficiency increases with the sunlight levels, and remains relatively stable until the irradiance reaches higher values, after which the efficiency slightly drops.

In order to capture the behavior of the PV system according to different sunlight levels, the data cleaning procedure is carried out for the following irradiance intervals: 50-250, 250-500 and greater than  $500 \text{ W/m}^2$ .

## 2.6. Measurements not representative of normal operation

Under normal conditions, the solar irradiance and power production measurements follow a strong linear relationship. Therefore, observations close to the irradiance-power linear relationship line are considered as normal (fault-free) operation measurements, while observations far from this line are considered as abnormal (faulty) operation data. The faulty data corresponds to abnormal power production, when the system produces significantly less or more power than it should at the given irradiance level.

Measurements indicating that the PV system is producing less electricity than it should lie under the lower normal operation limit. Measurements lying above the upper normal operation limit can indicate problems with measuring instruments that give erroneous readings suggesting that the PV system is producing more electricity than it should.

The data cleaning is based on the irradiance and AC power values. The ratio between power and irradiance will be used to determine thresholds for identifying data representative of faulty operation. These limits will be calculated for different irradiance intervals: 50-250, 250-500 and greater than  $500 \text{ W/m}^2$ . The data cleaning limits will be calculated as follows, for each interval:

1. the ratio between AC power (kW) and irradiance ( $\text{W}/\text{m}^2$ ) is calculated
2. the average and standard deviation of those values (the ratios) are calculated
3. the limits are calculated as the *average  $\pm$  3 x standard deviation*

This is the 3 standard deviation rule (also known as the 3 sigma rule) which states that for data following a normal distribution, 99.73% of values lie within 3 standard deviations of the mean. Even if the data is not normally distributed, most of the observations will fall within the 3 standard deviation interval.

Observations corresponding to values of the ratio between power and irradiance that are outside these limits will be considered representative of faulty operation, since they don't follow a strong irradiance-power linear relationship, and will be removed from the modelling dataset.

This data cleaning operation is iterative, in order to gradually remove all faulty data: at the end of an iteration, new cleaning limits will be calculated for the data representative of a normal operation and a new cleaning iteration starts.

Figure 8 shows data representative of normal and faulty operation after a first cleaning iteration; the cleaning limits, for the different irradiance intervals are also shown. The data cleaning for this particular PV system finished after 7 iterations; the last iteration is shown in Figure 9. The data cleaning evolution is summarized in Table 1.

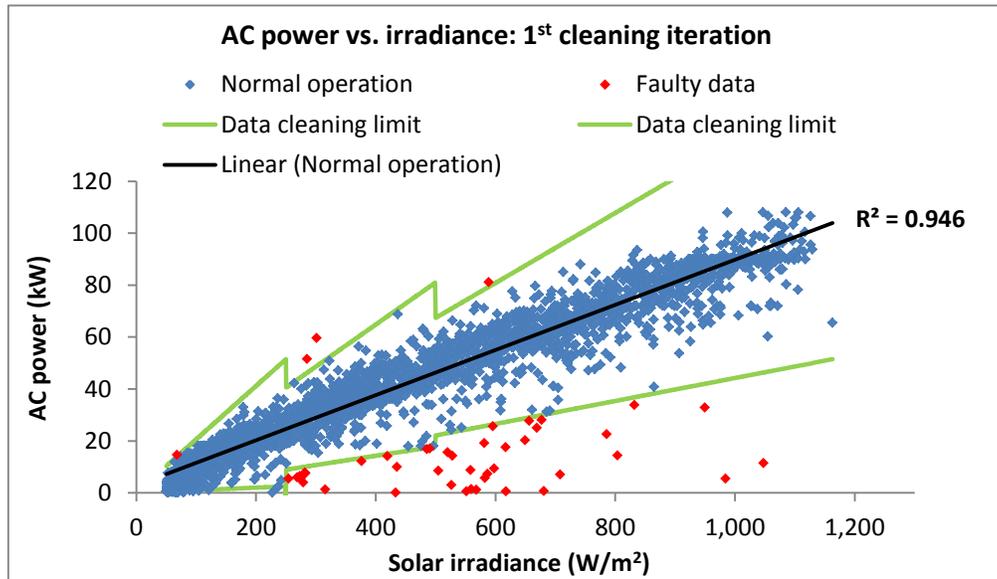


Figure 8: 1<sup>st</sup> cleaning iteration

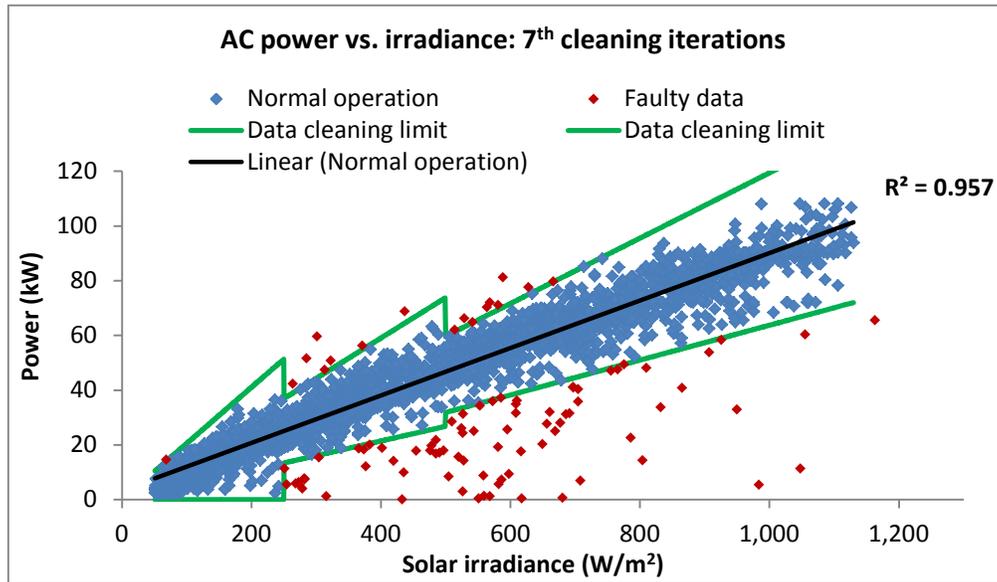


Figure 9: 7<sup>th</sup> cleaning iteration

Table 1: Data cleaning iterations

Cleaning step	R <sup>2</sup> value	R <sup>2</sup> improvement	Number of faults	% of faults (vs. total data)
Irradiance ≤ 50 W/m <sup>2</sup> , AC power ≤ 5 kW removed	0.9151	-	-	-
1 <sup>st</sup> cleaning iteration	0.9457	3.348%	45	1.48%
2 <sup>nd</sup> cleaning iteration	0.9533	0.797%	28	0.94%
3 <sup>rd</sup> cleaning iteration	0.9554	0.227%	13	0.44%
4 <sup>th</sup> cleaning iteration	0.9562	0.081%	5	0.17%
5 <sup>th</sup> cleaning iteration	0.9565	0.029%	1	0.03%
6 <sup>th</sup> cleaning iteration	0.9567	0.026%	2	0.07%
7 <sup>th</sup> cleaning iteration	0.9570	0.028%	2	0.07%

After the 6<sup>th</sup> and 7<sup>th</sup> iteration it is observed that faulty observations are still being identified. However, they correspond to irradiance values in the 50-500 W/m<sup>2</sup> interval. Since greater electricity productions occur at greater irradiance levels, it is more important to identify faults occurring at higher irradiances – superior to 500 W/m<sup>2</sup>. Moreover, the improvement in R<sup>2</sup> value is very low from the 6<sup>th</sup> to the 7<sup>th</sup> cleaning iteration, and the cleaning procedure is therefore stopped.

## 2.7. Summary of data cleaning steps

The data cleaning steps are as follows:

1. Calculate  $R^2$  coefficient between AC power and irradiance measurements; if  $R^2 < 0.70$   $\Rightarrow$  stop data cleaning, since the data quality is too poor to enable accurate modelling; else continue
2. Average measurements over 1 hour
3. Remove observations corresponding to:
  - a. irradiance  $< 50 \text{ W/m}^2$  and irradiance  $> 1,250 \text{ W/m}^2$
  - b. AC power  $\leq 5\%$  of maximal PV system AC power
  - c. AC power  $>$  maximal PV system AC power
  - d. PV panel temperature  $> 60^\circ\text{C}$  (rack-mounted) or PV panel temperature  $> 90^\circ\text{C}$  (building-integrated)
4. Calculate the  $R^2$  coefficient between irradiance and power
5. Calculate the values of the power / irradiance ratio
6. Calculate the averages and standard deviation values for the 50-250, 250-500 and greater than 500  $\text{W/m}^2$  irradiance intervals
7. Calculate data cleaning limits for each irradiance interval: average  $\pm 3 \times$  standard deviation
8. Identify and remove observations outside the data cleaning limits
9. Calculate new the  $R^2$  coefficient for the cleaned dataset
10. If  $R^2$  improvement  $< 0.1\%$  and number of faulty observations = 0  $\Rightarrow$  stop data cleaning; else repeat steps 4 to 8  
  
If  $R^2$  improvement  $< 0.1\%$  and number of faulty observations  $\neq 0$  and faulty observations occur only in the 50-500  $\text{W/m}^2$  irradiance interval  $\Rightarrow$  stop data cleaning; else repeat steps 4 to 8

It is important to note that although the cleaning limits are calculated for separate irradiance intervals, the  $R^2$  coefficient is calculated over the complete irradiance range.

### 3. AC power model

The model predicting the AC power of the current produced by the PV system under normal conditions is used to detect faults. This model uses the clean dataset obtained previously: it uses solar irradiance and PV panel temperatures to calculate the expected AC power production under fault-free conditions.

The model is based on a parametric approach to PV system modelling which calculates the AC power as follows:

$$P_{AC} = Irrad. \times [a_1 + a_2 Irrad + a_3 \log(Irrad)] \times [1 + a_4(T_{module} - 25)]$$

where  $P_{AC}$  = AC power (W)

$Irrad$  = solar irradiance (W/m<sup>2</sup>)

$T_{module}$  = PV module temperature (°C)

$a_1, a_2, a_3$  and  $a_4$  = parameters calculated so the model result is as close as possible to the measured data

In order to capture the behavior of the PV system according to sunlight levels, models for following irradiance intervals were developed: 50-250, 250-500 and greater than 500 W/m<sup>2</sup>. Models developed for those irradiance intervals are more accurate than one global model that covers the complete irradiance range.

#### 3.1. Training and validation errors

Prior to computing the predictive models, the data is separated into training and validation sets. The training dataset is used for developing the model, while the validation dataset is used to validate the model performance. If only the training data is used to determine the prediction performance, the accuracy can be over-estimated, since the model is specifically tuned to fit the training data. This is called model over-fitting, and it occurs when a model performs very well on the training dataset but is not able to generalize from the data trend and performs poorly on unseen data – the validation data.

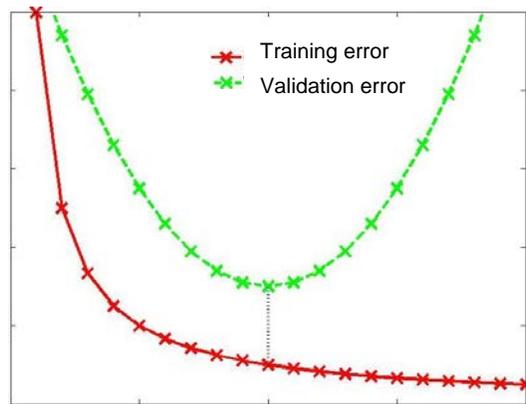
The validation dataset is obtained by setting aside a portion of the original dataset that will not be used during the training process.

The model parameters are calculated using an optimization routine that determines the parameter values that lead to a model result as close as possible to the AC power values from the training dataset.

After the model is fitted on the training data, its performance is tested on the validation data. Figure 10 shows an example of over-fitting, where the training error (red line) steadily decreases while the validation error (green line) starts to increase at some point during the modelling process.

From the available data, 30% of observations are randomly selected, in a uniformly distributed manner, as validation data. The remaining 70% of observations represent the training dataset and are used to develop the predictive model.

The models were scored in terms of the validation error, using the Coefficient of Variation of the Root Mean Square Error – CV(RMSE); it is calculated as the ratio of the Root Mean Square Error to the mean of the measured AC power values from the validation dataset.



**Figure 10: Training and validation errors**

## 4. Fault Detection

### 4.1. Normal operation limits

The fault detection is based on the models developed previously. The models calculate the expected AC power output that the PV system should have under normal operation conditions, for each irradiance interval. They are also used to calculate the system's normal operation limits.

The normal operating limits were calculated using the ratio between the measured and modeled AC power for the cleaned dataset used to develop the model. The limits were calculated using the 3 standard deviation rule:

$$\text{normal operation limits} = \text{average} \pm 3 \times \text{standard deviation}$$

where *average* and *standard deviation* are calculated using the values of the ratio *measured AC power / modeled AC power*

Normal operation limits were calculated for each irradiance interval (50-250, 250-500 and greater than 500 W/m<sup>2</sup>).

### 4.2. Fault detection

Faults are detected as follows:

- the irradiance and panel temperature measurements are used as inputs to the AC power model, and the production expected under normal operating conditions is computed by the model
- the ratio between the measured and modeled AC power is calculated
- ratios that have values outside the normal operation limits represent measurements considered as faults

The points located far from the irradiance-power straight line, previously identified in the data cleaning process as representative of a faulty operation, were used to validate this approach. They are compared to the corresponding model results to determine if they lie outside the normal operation limits.

Faulty data, located outside the normal operation limits, is shown in Figure 11. Most of the faults are located below the lower limit, indicating that the system is producing less electricity than it should at the given operating conditions – solar irradiance and PV panel temperature.

However, a few points are located above the upper limit, indicating that the system is producing more electricity than it should. This can be caused by malfunctioning measuring instruments generating erroneous readings; they can also be caused by inaccurate results of the AC power model.

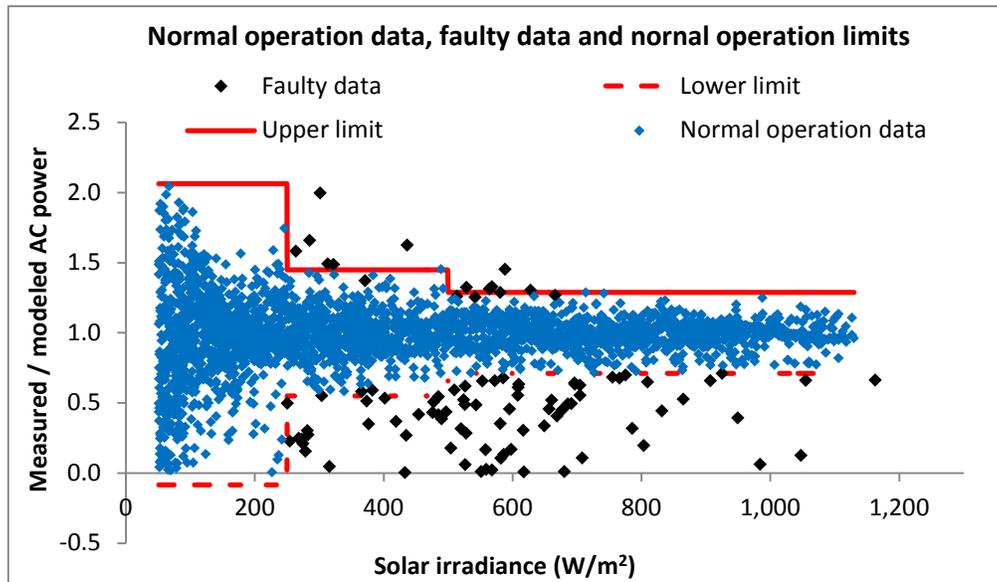


Figure 11: Faulty data and normal operation limits

For every AC power output measurement, the normal operation limits can also be expressed in kW by taking their values in terms of the ratio *measured AC power / modeled AC power* and multiplying them with the power output calculated by the model. For example:

- the modeled AC power is 95.36 kW
- the lower and upper normal operation limits, expressed in terms of the ratio *measured AC power / modeled AC power* are 0.807 and 1.193, respectively
- the normal operation limits, expressed in terms of kW, are: *lower limit* = 95.36 kW x 0.807 = 76.96 kW and *upper limit* = 95.36 kW x 1.193 = 113.76 kW

As an example, normal operation limits expressed in kW are shown in Figure 12.

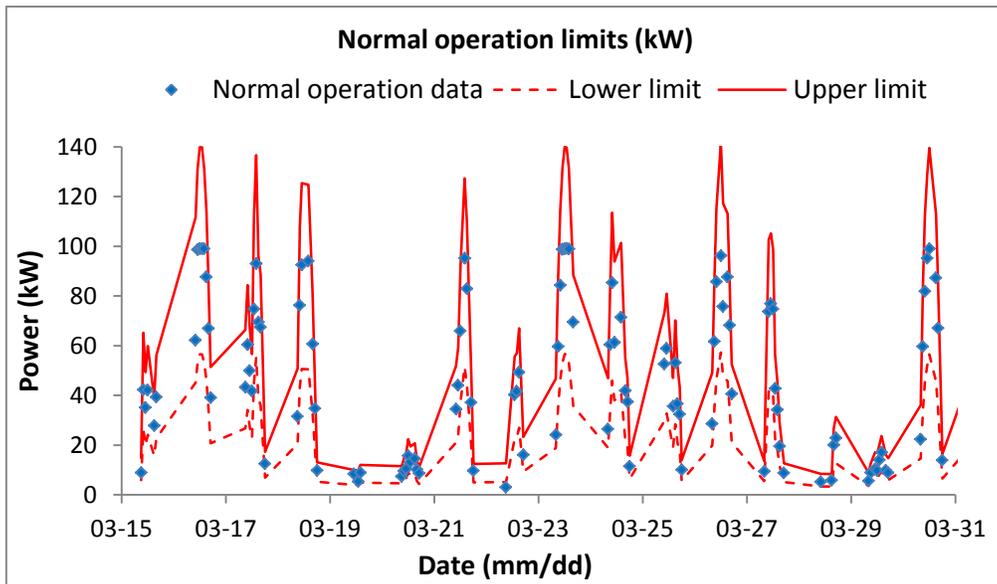


Figure 12: Normal operation limits expressed in kW

## 5. Performance ratio

The performance ratio (PR) evaluates the efficiency of the PV system by comparing its real electricity production to the ideal production at Standard Test Condition (STC) of 1,000 W/m<sup>2</sup> solar irradiance and 25<sup>0</sup>C PV module temperature – this is the rated (design) DC power output of the system. The PR is calculated over a specified period of time, and is used to monitor the system performance over time; it takes into account the electricity output and the amount of sunlight received over that period. The performance ratio *PR* is calculated as follows:

$$PR = \frac{\text{Total AC electricity production (kWh)}}{\text{Rated DC power (kW) x Total insolation (Wh/m}^2\text{) / 1,000 (W/m}^2\text{)}}$$

where *Total AC Electricity Production* = AC energy (kWh) produced over the interval for which the performance ratio is calculated

*Rated DC Power* = DC power at STC (rated, or design, DC power of the PV system, kW)

*Total insolation* = total solar energy received by the PV system over the interval for which the performance ratio is calculated (Wh/m<sup>2</sup>)

The closer the PR value for PV system approaches 1, the closer the system production is to its ideal, design production.

Both normal operation and faulty values are used to calculate the performance ratio. Prior to computing the PR values, observations containing missing values are identified; if the AC power value is missing, then the corresponding irradiance value is also deleted from the dataset. The same is done if an irradiance value is missing. This ensures the integrity of the PR calculation, since the values of power and irradiance are summed over the same interval.

Observations corresponding to irradiances lower than 50 W/m<sup>2</sup> are removed, in order to avoid measurement uncertainty at low sunlight levels.

Monthly performance ratios are shown in Figure 13. There is a drop in the system's performance for the month of May, caused by an inverter malfunction. The PR calculation can be calculated for shorter time intervals to further identify the time period when the system underperformed. Figure 14 shows PR values calculated on a weekly basis for the month of April; the problem occurred in the second week of April.

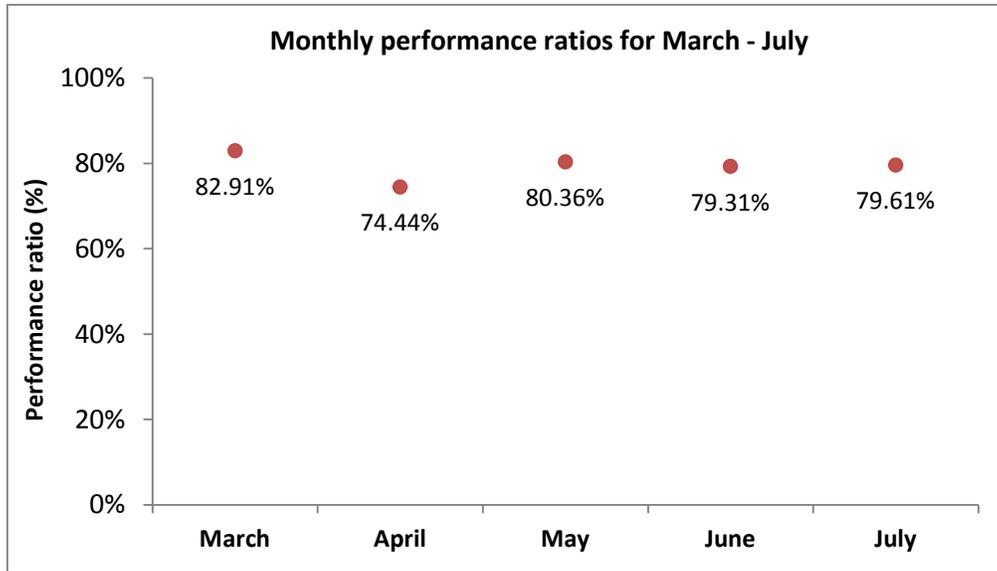


Figure 13: Monthly performance ratios

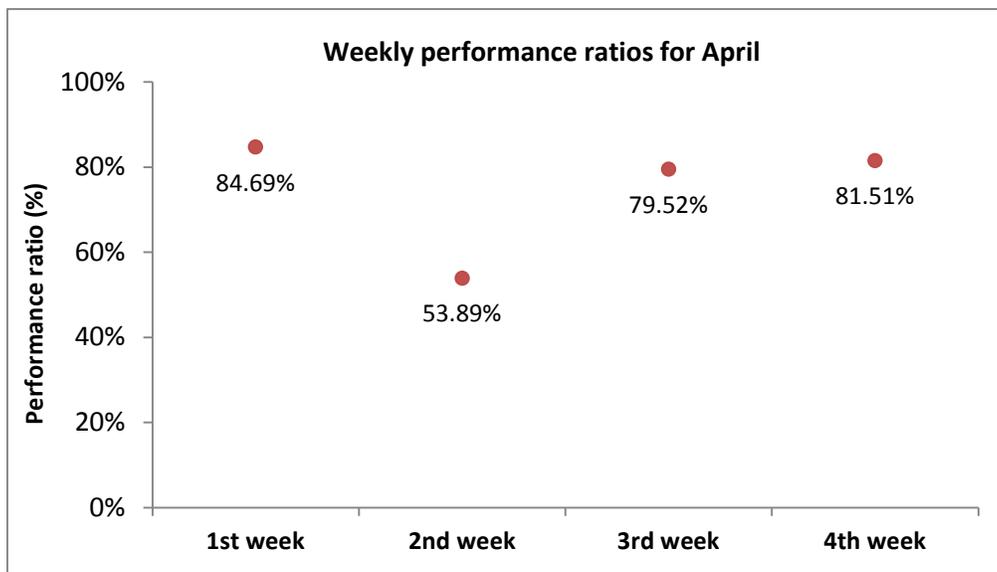


Figure 14: Weekly performance ratios

The performance ratio is a valuable tool for monitoring the PV system performance, as it indicates how closely a PV system is delivering to its design capacity; moreover, it does not require a model, only measured data.

The PR value complements the model-based method for detecting PV system faults:

- the model detects, in real time, PV system faults
- the PR value keeps track of the PV system performance, as it degrades over time – due to equipment aging or equipment malfunction, for example

## 5.1. Recommendations for performance ratio calculation

In order to properly calculate the performance ratio, the following steps are taken:

- observations corresponding to irradiances lower than  $50 \text{ W/m}^2$  are removed, in order to avoid measurement uncertainty at low sunlight levels
- all missing data is eliminated from the dataset
- data containing a large amount of missing values should not be used to calculate a PR value, as the presence of many missing values over the specified time interval will lead to a PR value not truly representative of the PV system performance over that interval
- a long enough time interval should be used to calculate a representative PR value; for example, one day might not be sufficient to properly capture the PV system behaviour and produce a representative PR value
- it is recommended that the performance ratio be calculated over a period of at least a week

## 5.2. Performance ratio superior to 100%

The performance ratio (PR) cannot be superior to 100%. If such a situation occurs, in most cases it is probably caused by erroneous irradiance measurements. A pyrometer miscalibration might be responsible for an under-estimation of the the irradiance level. In this case, the data wrongly indicates that the electricity production of the PV system takes places at lower than actual sunlight levels.