Nonparametric Kullback-Divergence-PCA for intelligent mismatch detection and power quality monitoring for safe rooftop PV to grid integration

Azzeddine Bakdi [*a], Wahiba Bounoua [b], Saad Mekhilef [c], Laith M. Halabi [c]

bkdaznsun@gmail.com, wb.bounoua@gmail.com, saad@um.edu.my, L.halabi@outlook.com

[a] Department of Mathematics, University of Oslo, 0851 Oslo, Norway.

[b] Signals and Systems Laboratory, Institute of Electrical and Electronics Engineering, University M'Hamed Bougara of Boumerdes, Avenue of independence, 35000 Boumerdès, Algeria.

[c] Power Electronics and Renewable Energy Research Laboratory (PEARL), Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia.

Abstract:

In parallel to sustainable growth in solar fraction, continuous reductions in Photovoltaic (PV) module and installation costs fuelled a profound adoption of residential Rooftop Mounted PV (RMPV) installations already reaching grid parity. RMPVs are promoted for economic, social, and environmental factors where they not only improve energy performance and reduce greenhouse effects but also contribute to bill savings. RMPV modules and energy conversion units are frequently subject to various types of anomalies which compromise power quality and promote fire risk and safety hazards for the personnel for which reliable protection is crucial. This article analyses historical data and presents a novel design that easily integrates with data storage units of RMPV systems to automatically process real-time data streams for reliable supervision. Dominant Transformed Components (TCs) are online extracted through multiblock Principal Component Analysis (PCA), most sensitive components are selected and their time-varying characteristics are recursively estimated in a moving window using smooth Kernel Density Estimation (KDE). Novel monitoring indices are developed as preventive alarms using Kullback-Leibler Divergence (KLD). This work exploits data records during 2015-2017 from thin-film, monocrystalline, and polycrystalline RMPV energy conversion systems. Fourteen test scenarios include array faults (line-to-line, line-to-ground, transient arc faults); DC-side

mismatches (shadings, open circuits); grid-side anomalies (voltage sags, frequency variations); in addition to inverter anomalies and sensor faults.

Keywords:

Rooftop PV; Grid-connected PV; Fault detection; Principal Component Analysis; Kullback-Leibler divergence; Kernel density estimation; Power quality monitoring.

1. Introduction

Solar plants are continuously expanding as global solutions for clean energy. In the next few decades, renewable energy is going to contribute to a significant proportion of the world's electricity needs. According to financial reports, international businesses exhibit an increasing interest in buying more renewable energy to the extent that 36 corporations, government agencies, and universities have agreed to buy 3.3 Gigawatts (GW) of wind and solar power in 2018 alongside the deals of 4.8 GW in 2017 [1]. Google, the giant company, announced in 2017 that it had met the target made in 2012 to achieve a 100% consumption of clean energy by its establishments around the world [2]. This followed the deals closed to purchase 3 GW of renewable energy capacity that year [3]. In Europe, the five largest countries in electricity production (UK, Germany, France, Italy and Spain) produced 90.5 Terawatt-hours (TWh) from solar in 2015 and had 90.4 GW of installed solar capacity at the end of 2017. Worldwide, the New Policies Scenario of the World Energy Outlook 2017 is expecting a solar Photovoltaic (PV) capacity of 2067 GW by 2040, producing 3162 TWh. And in the Sustainable Development Scenario, a goal of 3246 GW solar PV capacity is set to be achieved [4].

Many countries have already reached grid parity for solar Photovoltaics (PVs) [5], for which solar power plants will continue to get built at utility-scale, but in fact, the millions of Rooftop Mounted PV (RMPV) installations make the real potential for solar fraction [6]. RMPV installations are promoted for economic [7], social [8], and environmental [9] factors. By installing RMPV solar panels, consumers will pay less to the electric utility and may even become energy producers rather than consumers only. In addition to bill savings [10], RMPV systems contribute to lowering the demand for fossil fuels and greenhouse gas emissions [11], they also reduce the Levelized Cost Of Electricity (LCOE) and the dependency on the utility grid [12]. The

RMPV modules and energy conversion units are frequently subject to faults which if remain undetected, they cause safety hazards [13] for the personnel and fire risk [14],[15]. Moreover, distribution companies are anxious about what is being injected into their grids since local malfunctions in the system cause serious problems in the grid side. Grid connection faults compromise power quality [16] and cause many grid voltage regulation issues [17], these pose several protection-related challenges [18] to avoid islanding, tripping, and interference [19].

The fast growth of sustainable energy production relies on the developments in the technology involved. Besides, the economics of PV systems play an important role in making the technology available, where an economically valuable PV system is reliable, long-lasting, and rarely prone to malfunctions. Solar plants are degrading faster than expected as a result of various defects; prompt detection of such defects in RMPV systems can guarantee the continuous healthy operation and reduce energy losses. An inclusive survey on defects in grid-connected PV systems and the most recent fault diagnosis schemes proposed in the literature is provided in [20]. Grid-connected PV systems are subjected to defects that generally occur due to equipment failures such as PV array and Maximum Power Point Tracking (MPPT) faults at the DC side, faults on the AC side (or faults at the grid level), the DC/AC inverters interface, and sensors faults as classified by [21] for which nondestructive inspection, testing, and evaluation are highly required as reported by [22]. Most of the methods proposed so far for monitoring PV systems are summarised in a review paper [23]. The authors emphasized the importance of early fault detection to prevent the risks of energy loss and disastrous fires at PV installations. Satellite-based observations are used in a remote monitoring method and compared with the expected ones to detect small grid-connected PV systems failures [24]. This method utilises solar irradiance information derived from satellites to simulate the energy yield of a PV system, which is undoubtedly less accurate than on-site measurements. Another technique [25] is based on multiple online models corresponding to different ranges of solar irradiation to predict the AC power generation that requires climate measurements; hence, this method is cost-ineffective since it involves additional sensors. The last hardware-based approaches [24, 25] are known for their increased expenses of installing and maintaining such equipment and sensors which are also subject to failures and add to the complexity of the system. On the other side, the authors in [26] proposed a fault detection framework employing the fuzzy logic systems interface and artificial neural

networks (ANN) techniques to detect different types of faults. However, those methods need further specifications on the faulty data. Additionally, outlier detection rules were adopted for monitoring PV systems in [27] and [28]; under normal operation, these rules were found to trigger false alarms and were computationally expensive. A major drawback of Artificial Intelligence (AI) methods, such as [26, 27, 28], is the requirement of labelled data in training and calibration. This, however, is practically not feasible since representative data cannot be collected during faulty operations.

In this paper, a data-driven algorithm is proposed to detect the different types of faults in grid-connected RMPV systems. Measured data of several years of operation is used for three interconnected systems, namely thin-film, monocrystalline, and polycrystalline RMPV energy conversion systems. The datasets available from the RMPV sub-systems are statistically modelled using Multiblock Principal Component Analysis (MPCA). MPCA [29] [30] consists of calculating the multivariate models of each block using the standard PCA method after dividing the variables into relevant blocks. The blocking technique utilizes just the correlations between the features in the particular block to estimate the scores, while the whole number of variables is used to estimate the scores in the standard PCA. Constructing the blocks typically depends on the system structure, [31] considered the process sections to divide the variables into blocks that describe a unit or a specific physical or chemical operation. The measured data is projected on the MPCA model to obtain reference and online Transformed Components (TCs) which describe the amount and direction of variation in a large d-dimensional space which is orthogonal. The traditional PCA and its variants are proved effective and computationally efficient for monitoring and fault detection purposes [32, 33], however, they rely on heavy assumptions such as system linearity and time-invariance (process stationarity) in the construction stage while the analysis of the principal components is based on the assumption of data following a multivariate Gaussian distribution. Unfortunately, the three assumptions do not hold in practice as it will be verified and proved experimentally in this paper.

Therefore, the most sensitive features are selected to detect and measure any operation deviation using an information gain named Kullback-Leibler (KL) divergence. Recently, KL divergence has proved its effectiveness in a considerable number of research items in multivariate process monitoring [34]. However, the KL divergence method is inefficient due to its high computational complexity that limits the application

scope to low frequency (3-minutes sampled) data as reported in [34]. [35] and [36] employed univariate KL divergence on multivariate principal scores obtained from a PCA model, this procedure was extended in [37] to incipient faults. In this work, such techniques are referred to as parametrised approximations to KL divergence since they are based on assumptions of Gaussian distributions [35-37] and Gamma distributions [38] for the original data and extracted scores. The divergence in such parametrised approaches is turned into a simple detection of changes in statistical parameters such as the process mean shift and statistical dispersion which deteriorates the design sensitivity and robustness. The presented experimental analysis will prove that such approximations are inaccurate and practically far in grid-tied RMPV systems, for which a nonparametric but computationally-efficient method must be adopted. In this work, the scores are online evaluated through a sliding window approach employing KL measure through the non-parametric smooth Kernel Density Estimation (KDE) without any assumptions on the real system or its collected data. The developed approach is based on multiblock PCA decomposition and sensitive components selection followed by actual recursive KDE and accurate KL divergence. It is proved very efficient and effective since it respectively avoids the computation burden of multivariate KL divergence and escapes the basic assumptions of PCA. To this end, a wide range of tests are implemented in this article through fourteen scenarios based on real data records. A deep analysis reflects the violation of theoretical assumptions associated with traditional approaches. The obtained results prove the potential application of the proposed developments compared to state-of-the-art methods in fault detection [39, 40].

The rest of this article is organised as follows; Section 2 details the scope of this work with descriptions of the different PV systems under study as well as their collected data and test scenarios; Section 3 then summarizes the design procedure of the proposed algorithms, which are then applied on the given systems in Section 4 and tested on the fourteen scenarios, the obtained results are discussed and compared; and finally, important remarks are drawn in a conclusion in Section 5.

2. Systems description and scope

Today's rooftop solar arrays do not only generate clean energy and reduce the dependency on grid power, but they must also provide a long-term sustainable and reliable power source, and they need to have a

long-lasting solid foundation. On the dark side, distribution companies are anxious about what is injected in their grids and mainly the power quality. Moreover, safety hazards of the system are mapped into unsafety of the personnel working or living under such utilities. Minor drawbacks of rooftop solar systems are due to the technical standards of connectivity, accessibility, and increased maintenance costs. Because of these peculiarities, RMPV systems need to be equipped with the most reliable protection schemes and need to be continuously monitored. Despite the abundant methods available in the literature, not all of them can accurately address this problem in practice.

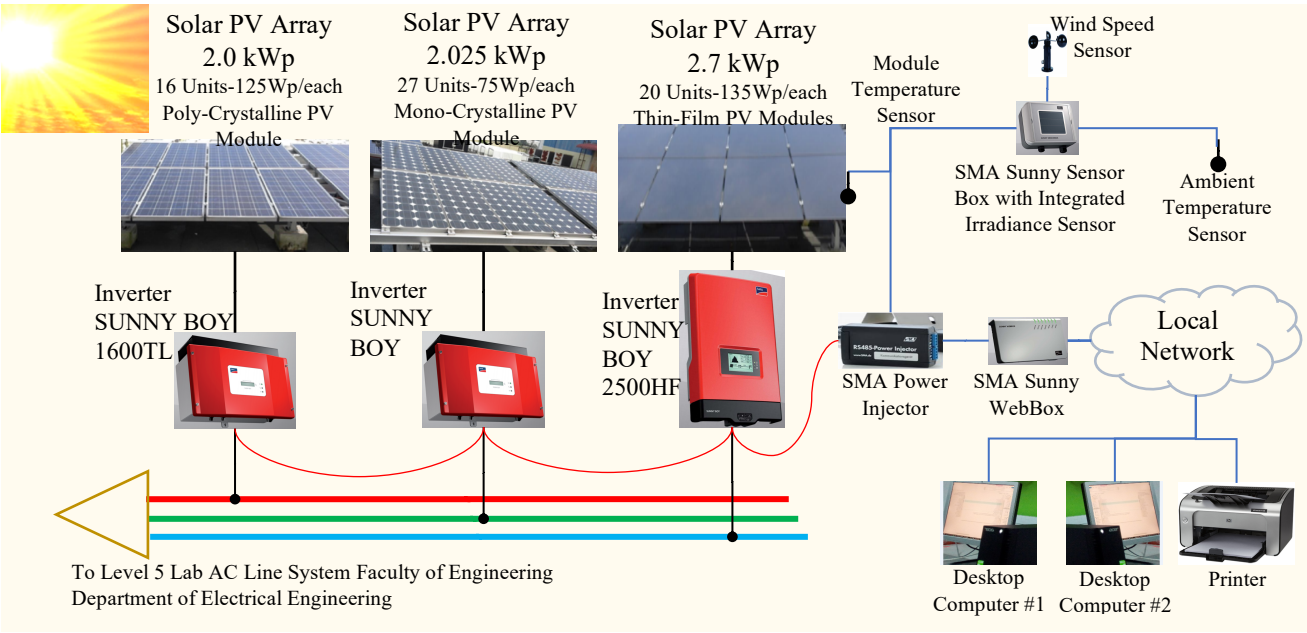
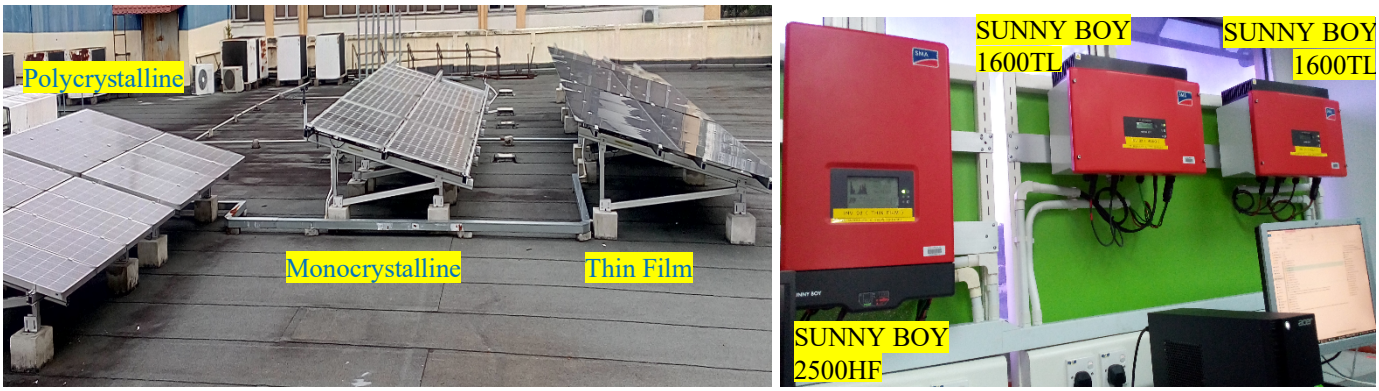


Figure 1. Overview of the three connected subsystems

A rooftop mounted PV system connected to a microgrid, is considered in this article, with collected data records over the three years 2015-2017. This medium-size installation can be seen as a collection of three main interconnected subsystems as shown in Figure 1. Their solar PV arrays are Poly-Crystalline, Mono-Crystalline, and thin film, they respectively consist of 16, 27, and 20 units as shown in the rooftop installations is Figure 2(a), their rated power is 2 kW, 2.025 kW, and 2.7 kW, respectively. In this article, S_1, S_2, and S_3 refer to the three subsystems in their respective order. The three blocks are connected to a microgrid, powering various loads of the research laboratory and synchronized with local sources and with the main grid lines though two SUNNY BOY 1600TL inverters, and SUNNY BOY 2500HF inverter as shown in Figure 2(b). Technical specifications of the SUNNY BOY 1600TL inverter can be found in its operating manual [41] and

its ratings are listed in the technical datasheet [42]. Technical specifications and ratings of the SUNNY BOY 2500HF inverter are also provided in [43, 44].



(a) (b)

Figure 2. Main components of the rooftop-mounted PV system. (a) Photovoltaic modules, (b) Grid-tied solar inverters

The three subsystems are also connected to SMA SUNNY SENSSORBOX [45, 46] to measure environmental conditions such as solar irradiance, wind speed, ambient temperature, and module temperature. The sensor box is powered through SMA power injector and integrated through a communication bus with SMA SUNNY WEBBOX [47, 48] which records data from all connected devices (sensor box and grid-tied inverters). The latter is connected to a local network and desktops to store and monitor data measurements.

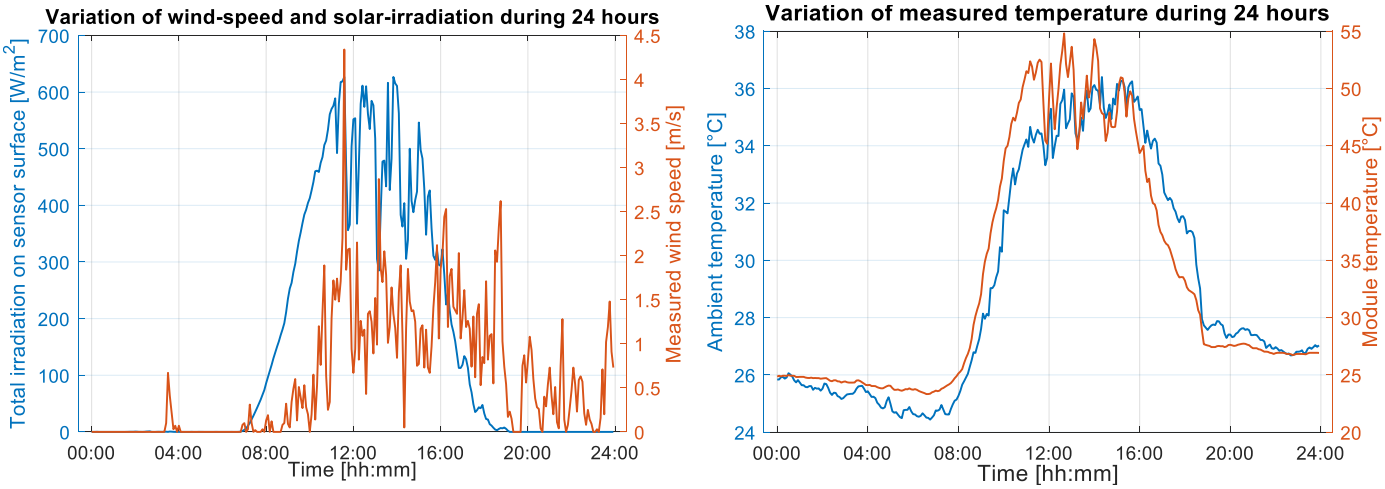


Figure 3. One-day variation of environmental conditions.

Table 1. Description of the selected monitoring variables of the PV system.

S_b	#	Name	Detailed description	Unit
Sensor Box	1	IntSolIrr	Total irradiation on sensor surface	W/m²
	2	TmpAmbC	Environment (Ambient) temperature	°C
	3	TmpMdulC	PV module temperature	°C
	4	WindVelms	Wind speed	m/s
Subsystem 1	5	Fac	Power frequency	Hz
	6	IacIst	Grid current	mA
	7	Ipv	DC current input	mA
	8	Pac	AC active power across all phases	W

	9	Uac	AC voltages (average of all string voltages)	V
	10	Upv-Ist	DC voltage input	V
	11	UpvSoll	Reference voltage	V
Subsystem 2	12	Fac1	Power frequency	Hz
	13	IacIst1	Grid current	mA
	14	Ipv1	DC current input	mA
	15	Pac1	AC active power across all phases	W
	16	Uac1	AC voltages (average of all string voltages)	V
	17	UpvIst1	DC voltage input	V
	18	UpvSoll1	Reference voltage	V
Subsystem 3	19	AMsAmp	DC current input in A	A
	20	AMsVol	DC voltage input in V	V
	21	AMsWatt	DC power input in W	W
	22	GridMsAphsA	Grid current phase L1 in A	A
	23	GridMsHz	Grid frequency in Hz	A
	24	GridMsPhVphsA	Grid voltage phase L1 in V	A
	25	GridMsTotVA	Total apparent power in VA	VA
	26	GridMsWphsA	Active power phase L1 in W	W
	27	Pac2	Delivered active power in W (total)	W

Recorded data include measurements of over 60 system variables (referred to as "measured values" in the technical description [49, 50]), operating parameters, log events, and messages. These are listed and described in [49, 50] for both inverter types. In the developed monitoring algorithm, monitoring variables are limited to the 27 fault-relevant signals as summarized in Table 1. Variables one to four are the environmental conditions measured through the sensor box, this set of variables is common for the three blocks. In addition to the external measurements, subsystems S_1, S_2, and S_3 are respectively monitored by variables 5 to 11, 12 to 18, and 19 to 27.

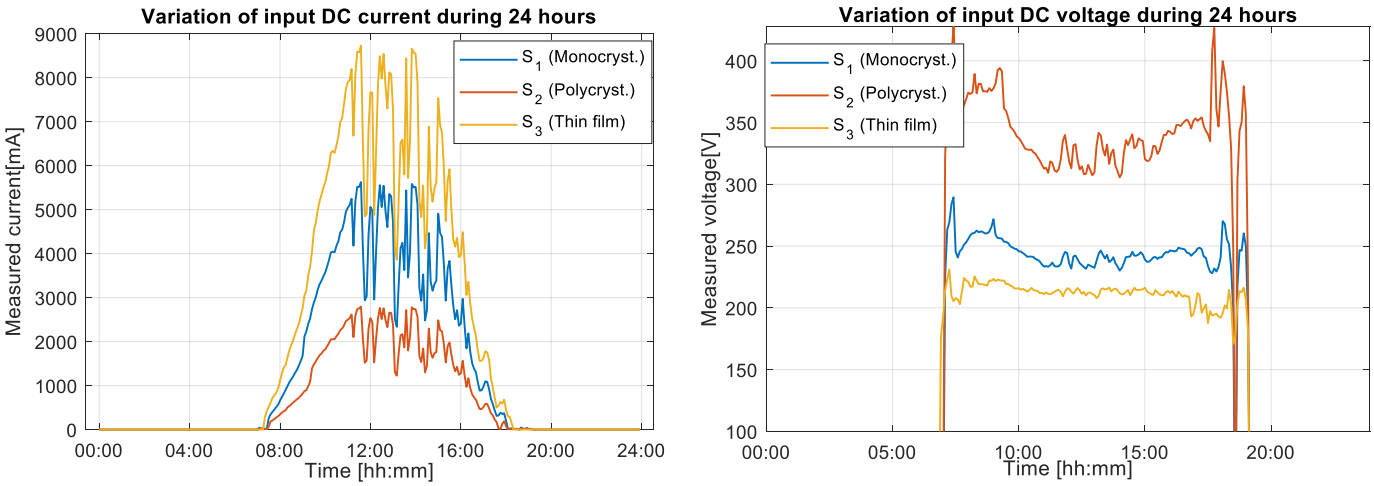


Figure 4. One-day variation of the module measurements.

Analysis of a PV system performance is highly dependent on the environmental conditions, and mainly the actual solar irradiance as well as ambient and module temperature. The influence of variations in temperature and irradiance on the PV module parameters are discussed in [51] and thermal performance of

PV modules is discussed in [52]. Authors in [15] demonstrated the strong correlation between an increase in the temperature and the change in energy production by rooftop integrated PV panels. In practice, a key aspect of environmental signals is their large (natural) variability, measurement errors, and noise as demonstrated in fig. 3 which represents the actual measurements recorded over one day. The DC-side signals, as shown in fig. 4, are highly and nonlinearly correlated with the latter variables and consequently exhibit large variability as well. In addition to the large variability within electrical signals over a day, the energy produced by the different PV arrays also exhibits a non-negligible variation over a year as shown in fig. 5 for 2017.

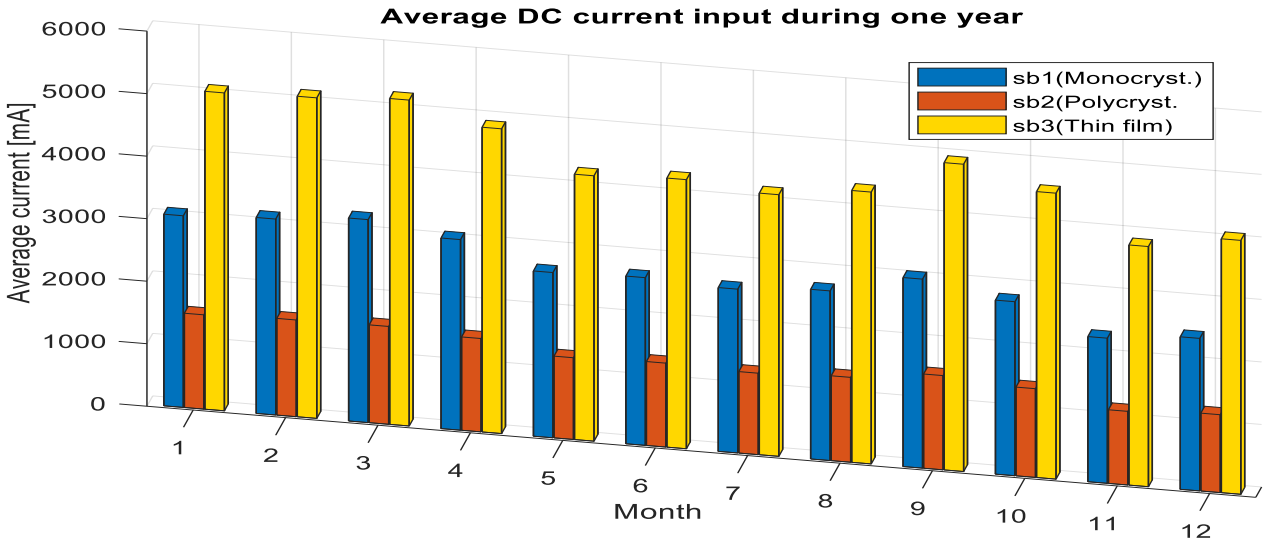


Figure 5. Variation of the monthly-mean DC-current over the year 2017.

Such strong correlations, large variations, and uncertainties cannot be ignored in practice when designing a monitoring device. Moreover, these conditions play a major role in masking symptoms of possible mismatches in the system operation, small anomalies, or unmatched power quality. These observations receive particular attention in section 3 when modelling the primary correlations of those variable with internal and grid signals through a nonparametric design. Comparisons with conventional schemes, ignoring some of such major factors, are drawn in section 4 to demonstrate the significance of such recommendations.

Table 2. PV array, Inverter, Grid, and Sensor faults.

Type	Number	Fault	location	Behaviour
Array faults (DC-side faults)	F1	Ground fault	S_1	Abrupt
	F2	Line to line fault	S_2	Abrupt
	F3	Parallel arc faults	S_3	Abrupt
	F4	Series arc faults	S_1	Abrupt
	F5	Partial shading faults	S_2	Abrupt
	F6	Short circuit	S_3	Abrupt

	F7	Open circuit	S_1	Abrupt
Inverter faults	F8	Frequency control (overshoot)	S_2	Incipient
	F9	Slow frequency control (transient)	S_3	Incipient
Grid faults (AC-side faults)	F10	Voltage sag	S_1, S_2, S_3	Intermittent
	F11	Frequency variation	S_1, S_2, S_3	Random
Sensor faults	F12	Voltage sensor fault	S_1	Random
	F13	Temperature sensor fault	S_2	Bias
	F14	Current sensor fault	S_3	Random

In this article, data records of the described system span three years of continuous operation during 2015-2017. Various scenarios are investigated in this article, as summarized in Table 2. These scenarios span physical as well as electrical and environmental faults and mismatches at the PV arrays, inverters, and grid levels. Due to the prementioned factors, the DC-side faults are hard to be detected however they have a medium severity level because of the bounded voltage and current of the PV modules. On the contrary, AC-side faults are less challenging to detect, however being highly severe, they need to be detected at their very early stages to ensure a safe microgrid integration and match the technical requirements and desired power quality. [53] provides a comprehensive review of a wide range of faults in grid-connected PV systems, causes, symptoms, and their respective protection schemes. [54] highlighted the power quality issues for building integrated renewables such as PV systems. Line to line, line to ground, and short circuit faults, in general, occur between two points of different potentials. These faults occur due to breakdown of insulation, corrosion of inductors, maintenance errors, and damage inside the PV arrays. Arc faults [55] are short-term versions of those faults and may cause permanent faults. These faults contribute to serious fire threats and safety hazards [53]. Open circuit faults may occur after some of the previous faults, and together with shading faults are considered as mismatches. Those scenarios are discussed in section 4, showing a correlation between the detection delay and the severity of each situation.

3. Data analysis and monitoring algorithms

In long terms, digitized PV plants generate vast amounts of data which reflects the historical behaviour and sharpest details within the system. Such valuable information can be exploited for artificial modelling of the PV system [56], parameter estimation and monitoring [57], and even PV power generation forecasting

[58]. A common fact is that such high-dimension data exhibits a lower statistical rank and it varies in a lower-dimensional space. Moreover, the PV system data is multivariate and includes weak as well as strong parallel and serial correlations among its variables; the data also exhibits large (DC-side) and small (grid-side) variations which are all of a significant importance and in which various anomalies exhibit different patterns to be detected by a protection system. Such information can be extracted using PCA while decorrelating the initial space to reduce the dimension of the treated problem and therefore its computational cost while considering a univariate analysis of the most sensitive TCs. Multiblock decomposition [59, 60] is used in this work for factorization and decentralized monitoring. Instead of statistical preferences, the system variables are divided into 3 blocks according to the given structure of the rooftop mounted solar system, as demonstrated in the previous section. Each subsystem is represented by a block (S_b for $b = 1,2,3$), several datasets are acquired for analysis purposes, one healthy set is recorded for constructing the basic model which is then validated and tested on a set of 14 scenarios of design and quality mismatches that the system is subject to. The healthy set, recorded during fault-free operation, is denoted as $\boldsymbol{X}_{\hbar_b}$ with $N_\hbar$ observations and l_b variables, while the very large data samples are used for validation and testing.

3.1. Multiblock PCA for PV systems modelling

Considering a raw dataset $\boldsymbol{X}_{\hbar_b}$, collected from a sub-system S_b in normal operating mode with $N_\hbar$ samples of l_b measured variables. Primarily, this large data matrix is scaled to bring all the variables, measured with different scales and units, down to zero mean (μ) and unit variance (σ^2), and hence all the variables can be treated equally during the analysis [61]. The auto-scaled data matrix is obtained as:

$$\boldsymbol{X}^{\aleph}{}_{\hbar_b} = \left[\frac{\boldsymbol{x}_{b_1} - \mu_{b,1}}{\sigma_{b,1}}; \frac{\boldsymbol{x}_{b_2} - \mu_{b,2}}{\sigma_{b,2}}; \cdots; \frac{\boldsymbol{x}_{b_{l_b}} - \mu_{b,l_b}}{\sigma_{b,l_b}}\right] \tag{1}$$

where μ_i and σ_i are the mean and the standard deviation of the i^{th} variable of the healthy data set in subsystem S_b, these can be directly estimated or updated at any stage as follows:

$$\mu_{b,i} = \frac{1}{N_\hbar}\sum_{i=1}^{N_\hbar} \boldsymbol{x}_{b_i} \quad \text{for } b = 1,2,3, \qquad i = 1, \cdots, l_b \tag{2}$$

$$\sigma_{b,i}^2 = \frac{1}{N_\hbar - 1} \sum_{i=1}^{N_\hbar} \left[\boldsymbol{x}_{b_i} - \mu_{b,i} \right]^2 \quad \text{for } b = 1,2,3, \qquad i = 1, \cdots, l_b \tag{3}$$

For simplicity, the normalized data matrix $\boldsymbol{X}^{\aleph}{}_{\hbar_b}$ is denoted $\boldsymbol{X}_{\hbar_b}$ and is given by:

$$\boldsymbol{X}_{\hbar_b} = \left[\boldsymbol{x}_{b_1} \, \boldsymbol{x}_{b_2} \, \cdots \boldsymbol{x}_{b_{l_b}} \right] \in \mathcal{R}^{N_b \times l_b} \tag{4}$$

The data transformation is based on the sample covariance matrix $\boldsymbol{\Phi}_b$ of the data, where:

$$\boldsymbol{\Phi}_b = \frac{1}{N_b - 1} \boldsymbol{X}_{\hbar_b}{}^T \boldsymbol{X}_{\hbar_b} \tag{5}$$

through the spectral decomposition of the later as:

$$\boldsymbol{\Phi}_b = \boldsymbol{P}_b \boldsymbol{\Lambda}_b \boldsymbol{P}_b{}^T \tag{6}$$

henceforth, the block loading matrix $\boldsymbol{P}_b$ is obtained with orthogonal components, i.e. $\boldsymbol{P}_b \boldsymbol{P}_b{}^T = \boldsymbol{I}_b$, constructed by the eigenvectors which represent the variations directions. $\boldsymbol{\Lambda}_b = diag\left(\lambda_{b_1}, \lambda_{b_2}, \cdots, \lambda_{b_{l_b}}\right)$ is a diagonal matrix constructed of the eigenvalues in a decreasing order $\{\lambda_{b_1} \geq \lambda_{b_2} \geq \cdots \geq \lambda_{b_{l_b}} \geq 0\}$, each eigenvalue represents the amount of variance per the corresponding direction.

Subsequently, the data collected from the PV system blocks is transformed by PCA projection into a new matrix $\boldsymbol{T}_{\hbar_b} \in \mathcal{R}^{N_\hbar \times l_b}$ named as block score matrix of uncorrelated variables $\{\boldsymbol{t}_{\hbar b_1}, \boldsymbol{t}_{\hbar b_2}, \cdots, \boldsymbol{t}_{\hbar b_{l_b}}\}$:

$$\boldsymbol{T}_{\hbar_b} = \boldsymbol{X}_{\hbar_b} \boldsymbol{P}_b \tag{7}$$

PCA allows the provision of a set of uncorrelated variables from the original set of correlated variables. These are called Transformed Components (TCs). At this stage, l_b reference TCs components are obtained for each subsystem S_b, for $b = 1,2,3$. PCA results in a statistical model that describes the PV system data patterns and correlations given the reference data profile. Moreover, the resulting TCs (combinations of the correlated data) are independent and can be monitored individually in real-time.

3.2. Smooth KDE & Kullback-Leibler divergence

Considering the high-level uncertainty and the large variability in the PV system data compared to symptoms of anomalies as mentioned in the previous section, the Kullback-Leibler divergence (KL-divergence) [62] is adopted in this article. The idea is to develop robust and sensitive measures of any deviation in the overall PV system performance at time instance n from the nominal operation described explicitly by historical data. For

feasible computation time and resources, and for accurate monitoring purposes, this measure is calculated in the decorrelated space and selecting only the most sensitive components. The KL-divergence is an information-based measure of dissimilarity between two probability distributions $f_{\boldsymbol{X}}(\boldsymbol{x})$ and $\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})$ defined over the same random variable $\boldsymbol{X}$. KL-divergence is a special case of α-divergence functions and it is asymmetrical non-negative quantity i.e. $DKL\left[f_{\boldsymbol{X}}(\boldsymbol{x}): \tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})\right] \neq DKL\left[\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x}): f_{\boldsymbol{X}}(\boldsymbol{x})\right] \geq 0$ [63].

The KL-divergence between two probability density distributions $f_{\boldsymbol{X}}(\boldsymbol{x})$ and $\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})$ is defined as the expectation over $f_{\boldsymbol{X}}$, and it is given by:

$$DKL\left[f_{\boldsymbol{X}}(\boldsymbol{x}): \tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})\right] = \mathbb{E}_{f_X}\left[log \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})}\right] \tag{8}$$

If $\boldsymbol{X}$ is a discrete random variable then Eq. (8) reduces to:

$$DKL\left[f_{\boldsymbol{X}}(\boldsymbol{x}): \tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})\right] = \sum_{x \epsilon X} f_{\boldsymbol{X}}(\boldsymbol{x})\ log \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})} \tag{9}$$

And for continuous Random Variable $\boldsymbol{X}$:

$$DKL\left[f_{\boldsymbol{X}}(\boldsymbol{x}): \tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})\right] = \int f_{\boldsymbol{X}}(\boldsymbol{x})\ log \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})}\ dx \tag{10}$$

It is a measure of the inefficiency of assuming that the distribution is $\tilde{f}_{\boldsymbol{X}}$ when the true distribution is $f_{\boldsymbol{X}}$. In this work, $f_{\boldsymbol{X}}$ represents the reference density function created in an offline stage through the reference TCs, while $\tilde{f}_{\boldsymbol{X}}$ is the online-estimated test density function. The idea is used in this article to measure the divergence of a prevailing PV system behaviour and characteristics according to its most recent measurements for a reference data profile.

KL-divergence is widely proved efficient for monitoring purposes, authors in [37] have obtained a closed-form approximation for this measure across variables following Gaussian distribution [64]. Unfortunately, the highly sensitive information gain, in this situation, is turned into measuring deviation in the mean and variance only, this heavy assumption deteriorates the performance of this measure since data in practice does not follow a Gaussian distribution, particularly during the occurrence of anomalies. Further assumptions are made in the prementioned approach such as the limitations to linear static (time-invariant) systems. A second drawback is the density ratio estimation involves multiple parameters with optimization functions, this approach is computationally infeasible for large-scale systems, especially if the system has fast dynamics. Alternatively,

another approximation of such measure is widely used in the literature based on direct density ratio estimation, called the importance estimation, used in [65, 66]. This approach reduces relatively the computation cost with light assumptions however the estimated ratio is still multivariate and demanding, furthermore, the ratio could explode to infinity.

The prementioned lacks promote motivations of this article to design a novel approach to measure the KL-divergence to preserve its high sensitivity which is highly crucial for identifying early signs of anomalies during power generation. The multivariate problem is turned into a univariate analysis which greatly reduces the computation time and resources, moreover, monitoring only sensitive TCs greatly improves the algorithm performance.

Without any assumptions on the system or its data, the Probability Density Functions (PDFs) are recursively estimated for the block TCs by a means of a non-parametric estimation method called the Kernel Density Estimation (KDE), also knowns as Parzen windows [67] that provides a smooth estimate based on sufficient amount of data. Given a set of offline-estimated reference block TCs, represented as follows:

$$\boldsymbol{T}_{\hbar_b} = \left\{ {t_{\hbar b_i}}^{k} \right\}_{k=1;\ \ i=1}^{k=N_{\hbar};\ i=l_b} = \left\{ \boldsymbol{t}_{\hbar b_1}, \boldsymbol{t}_{\hbar b_2}, \cdots, \boldsymbol{t}_{\hbar b_{l_b}} \right\} \ , \qquad \boldsymbol{t}_{\hbar b_i} \in \mathcal{R}^{N_{\hbar} \times 1} \tag{11}$$

for all subsystems, where k is the time index. Hence ${t_{\hbar b_i}}^{k}$ is the i^{th} reference TC of sub-block S_b at time k, and the reference PDFs $f_{t_{\hbar b_i}}(t;\ h)$ can be estimated through KDE as:

$$\hat{f}_{t_{\hbar b_i}}(t;\ h) = \frac{1}{N_{\hbar} h} \sum_{k=1}^{N_{\hbar}} \mathrm{K}\left(\frac{t - {t_{\hbar b_i}}^{k}}{h} \right) \tag{12}$$

for $i = 1, \cdots,\ l_b$ for the three subsystems $b = 1, \cdots, 3$ using the smoothing kernel function:

$$\mathrm{K}(x) = \frac{1}{\sqrt{2\pi}}\ exp\left(-\frac{x^2}{2} \right) \tag{13}$$

$h > 0$ is the bandwidth that controls the smoothing and the estimation. This work adopts a bandwidth that minimizes the Mean Integrated Squared Error (MISE) [68]:

$$\mathrm{MISE}(h) = \mathbb{E}\left[\int \left(\hat{f}_{t_{\hbar b_i}}(t;\ h) - f_{t_{\hbar b_i}}(t;\ h) \right)^2 dt \right] \tag{14}$$

3.3. Performance monitoring indices

At any time index n, the measured data is first scaled using Eq. (1), however, using the same parameters obtained in Eq. (2) and Eq. (3). The scaled measurement is projected on the multiblock PCA models to estimate the online block TCs. The test TCs at a time index n, are formed by augmenting the N_t most recent samples of the TCs.

$$\boldsymbol{T}_{b,n} = \left\{t_{b_i}{}^{k}\right\}_{k=n_0;\ i=1}^{k=n;\ \ i=l_b} = \{\boldsymbol{t}_{b_1,n}, \boldsymbol{t}_{b_2,n}, \cdots, \boldsymbol{t}_{b_{l_b},n}\} \ , \qquad \boldsymbol{t}_{b_i,n} \in \mathcal{R}^{N_t \times 1} \tag{15}$$

where $n_0 = n - N_t + 1$, for the three system blocks $b = 1, 2, 3$.

The test block TCs are observed within a sliding window of length N_t, these projections (Eq. (7)) are stacked with their respective most-recent historical projections any time instance n. These online components are obtained through the online projection of PV observations on the loading vectors obtained in Eq. (6), i.e. $\boldsymbol{t}_{\boldsymbol{b}_{i,n}} = [\boldsymbol{t}_{b_i}(n),\ \boldsymbol{t}_{b_i}(n-1),\ \cdots, \boldsymbol{t}_{b_i}(n - N_t + 1)]'$. The window length is simply selected just large enough so that one TCs block can describe the instantaneous behaviour of the PV system. In other words, the number of samples in an online TC $\boldsymbol{t}_{\boldsymbol{b}_{i,n}}$ should yield a correct and representative estimation of its probability density function.

The online test density $f_{\boldsymbol{t}_{b_i,n}}(t;\ h)$ of the i^{th} reference TC of sub-block S_b at time n is estimated through KDE as:

$$\hat{f}_{\boldsymbol{t}_{b_i,n}}(t;\ h) = \frac{1}{N_t h} \sum_{k=1}^{N_t} \mathrm{K}\left(\frac{t - \boldsymbol{t}_{b_i,n}{}^{k}}{h}\right) \tag{16}$$

This estimation follows the same steps of the KDE of the reference block TCs as in Eq. (12,13,14).

In order to reach decisions on the state of the PV system, it is required to set up statistical hypotheses on the basis of the KL-divergence (Eq. (8,9,10)) as a monitoring statistic. The normal state, defined by the null hypothesis $\mathcal{H}_0$, is characterized by a region of non-significance where the real-time divergence is within a pre-established threshold δ. Whenever DKL diverges significantly from the region of acceptance, the owner of the RMPV system or a supervision platform would be inclined to reject the null hypothesis and accept the alternative hypothesis $\mathcal{H}_A$. The latter situation indicates the departure from nominal to the abnormal state of the PV system. Consequently, the following decision rule is formulated:

$$DKL\left[\hat{f}_{t_{\hbar b_i}}(t;\, h): \hat{f}_{\boldsymbol{t}_{b_i,n}}(t;\, h)\right] \underset{\mathcal{H}_0}{\overset{\mathcal{H}_A}{\gtrless}} \delta \tag{17}$$

Through this statement, a decision is made over the KL-divergence of the score components of each sub-block as follows:

$$\begin{cases} \mathcal{H}_0:\ DKL\left[\hat{f}_{t_{\hbar b_i}}(t;\, h): \hat{f}_{\boldsymbol{t}_{b_i,n}}(t;\, h)\right] \ \leq \delta_{b_i} \\ \mathcal{H}_A:\ DKL\left[\hat{f}_{t_{\hbar b_i}}(t;\, h): \hat{f}_{\boldsymbol{t}_{b_i,n}}(t;\, h)\right] \ > \delta_{b_i} \end{cases} \text{for } b = 1,2,3 \ \ and \ \ i = 1, \cdots, l_b \tag{18}$$

The user can also get an insight into the level of severity for a particular malfunction by how from from zero is the KL-divergence at an instance. In addition to those decisions based on hypothesis tests, the following performance monitoring indices are defined for a given sample measurement at time instance n:

$$D_{b_i}(n) = DKL\left[\hat{f}_{t_{\hbar b_i}}(t;\, h): \hat{f}_{\boldsymbol{t}_{b_i,n}}(t;\, h)\right] \tag{19}$$

for $b = 1,2,3 \ \ and \ \ i = 1, \cdots, l_b$, with respective control limits $CL_{Db_i} = \delta_{b_i}$. The previous hypothesis of Eq. (17,18) will be adapted at time n to:

$$\begin{cases} \mathcal{H}_0(n):\ D_{b_i}(n) \ \leq CL_{Db_i} \\ \mathcal{H}_A(n):\ D_{b_i}(n) \ > CL_{Db_i} \end{cases} \text{for } b = 1,2,3 \ \ and \ \ i = 1, \cdots, l_b \tag{20}$$

Notice that the null hypothesis is violated whenever any of the online block components has its test density function $\hat{f}_{\boldsymbol{t}_{b_i,n}}(t;\, h)$ diverged considerably (above the control limit) from its respective reference function $\hat{f}_{t_{\hbar b_i}}(t;\, h)$. Moreover, it will be shown in the next section that monitoring the entire RMPV system is reduced to monitoring the first and last TCs which are sensitive to large variability (in DC-side and mismatches) and small variations (AC-side and quality deviation), respectively.

The sensitive indices are observed during nominal operation conditions and their control limits are constructed through nonparametric confidence intervals based on their estimated cumulative distribution functions. Allowing a tolerable level α of false alarms ($\alpha = 1\%$), the control limits are selected to ensure a coverage probability of $(1-\alpha)\%$ of the measured samples during healthy operation conditions are flagged as safe. The thresholds of such indices are established empirically such that the PV system is considered to be operating in normal mode if KL-divergence between the estimates of the reference density function and the online test density function is approximately zero i.e. the two distributions are similar. While any dissimilarity between

the distributions will appear as a departure of DKL from the threshold and this will be regarded as an abnormality.

4. Experimental results and discussion

This section is based on the real records measured during the three years 2015 to 2017 from the three interconnected RMPV systems, installed on the roof of Power Electronics and Renewable Energy Research Laboratory (PEARL) of MALAYA University. The three blocks of the system, as described in Section 2, are connected to a microgrid, powering various loads of the research laboratory and synchronized with local sources and with the main grid lines depicted in Figures 1 and 2 above. Huge datasets are available where 27 fault relevant signals are selected as detailed in Table 1, this covers all the modules, inverters, and grid measurements for the three systems. The rich datasets are filtered to remove records with erroneous and missing values and pre-processed for noise reduction, the available data is then exploited to analyse the system behaviour and used for the design and implementation and validation of the proposed algorithms as well as tests and comparisons. Single and multiple events are investigated in 14 scenarios as described in Table 2; these are injected in different locations with various behaviours.

The data is first randomly divided into regions, the training regions for the global as well as the multiblock PCA models are of 500 samples for each signal $\boldsymbol{X}_{\hbar} \in \mathcal{R}^{500\times 27}$, while a 40000 long set is used for validation $\boldsymbol{X}_{t} \in \mathcal{R}^{40000\times 27}$, while each of the 14 test scenarios includes 5000 samples $\boldsymbol{X}_{t} \in \mathcal{R}^{5000\times 27}$. Multiblock PCA models are first constructed for the RMPV systems, each block data $\boldsymbol{X}_{\hbar b}^{500}$ is first autoscaled as given in Eq. (1,2,3), and the multiblock PCA decomposition is achieved through Eq. (4,5, 6) for the three blocks $b = 1,2,3$. The resulting TCs have their reference densities estimated through Eq. (12) with all samples of projected training data $N_{\hbar} = 500$, the respective online TCs have their PDFs estimated at each time instance through KDE as in Eq. (16) with $N_t = 300$ samples, the online TCs are compared to their reference ones and evaluated based on the dissimilarity between their PDFs through KL divergence through Eq. (10 and 17). The traditional KLD approach is expelled from performance comparison since it is completely impractical to consider recursive estimation of online to reference density ratio for such 27-dimensional space. The approximation multivariate KLD approach [34] is of high complexity for this RMPV system and data with a computation time measured around weeks in addition to out-of-memory problems.

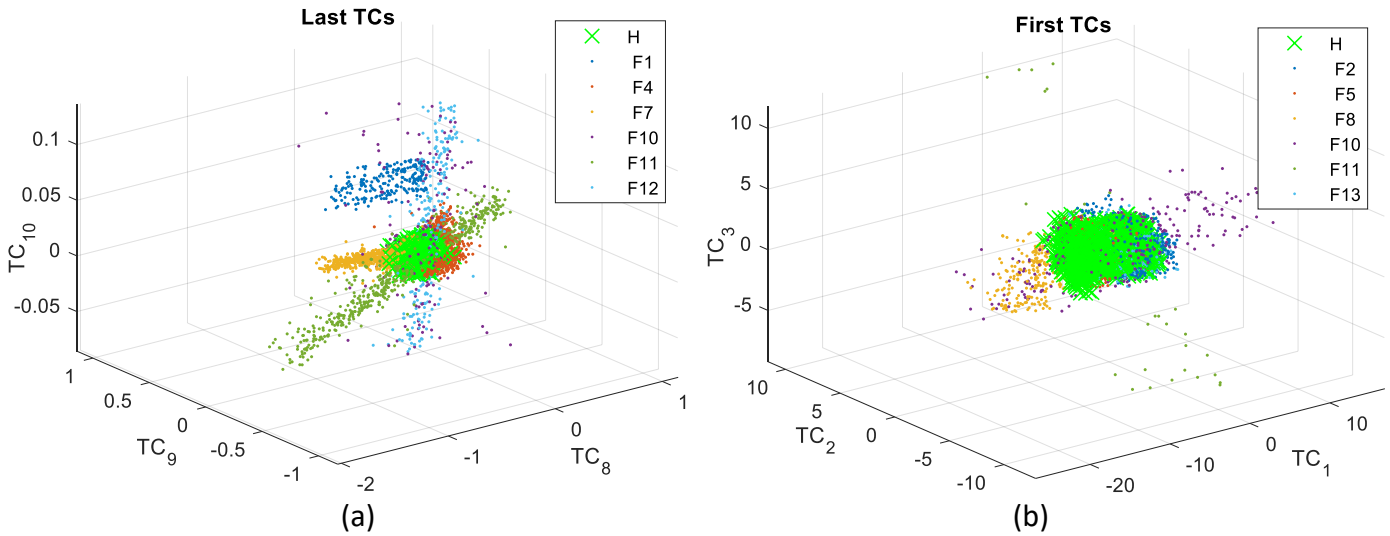


Figure 6. Normal and faulty data visualization in (a) last TCs of S_1 and (b) first TCs of S_2.

Figure 6 demonstrates some of the online block Transformed Components (TCs), obtained through projecting the online measurements from the RMPV systems on the reference artificial models, during normal and faulty operations of different scenarios. Notice the poor capability of the block TCs in discriminating the faulty ($F_.$) and fault-free (H) situations, more importantly, the components of the residual subspace are even more sensitive than those of the principal subspace showing some sort of separation of fault clusters. While the principal subspace describes the dominant variation due to natural behaviour, the residual subspace contains the negligible variation, which in most of the cases, is considered as attributed to noise and anomalies. The evaluation of the TCs in both subspaces must be very accurate for robust monitoring. MPCA decorrelates multivariate data from a high-dimensional space into univariate one-dimensional TCs which still describe the original covariance (variations and correlations). This is extremely advantageous but unfortunately cannot detect faults and quality deviations. The limitation of conventional PCA [32, 39] and parametrised KL approaches [35 - 38] is investigated in the following, while nonparametric KL divergence of the same TCs in Figure 6 will be proved very effective in the detection.

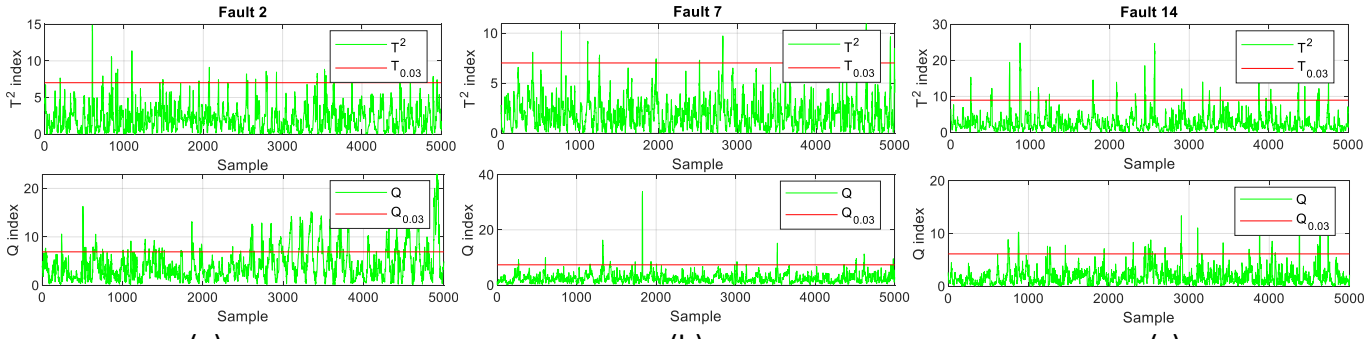


Figure 7. The poor detection performance of conventional PCA through Q and T^2 charts of [32, 39].

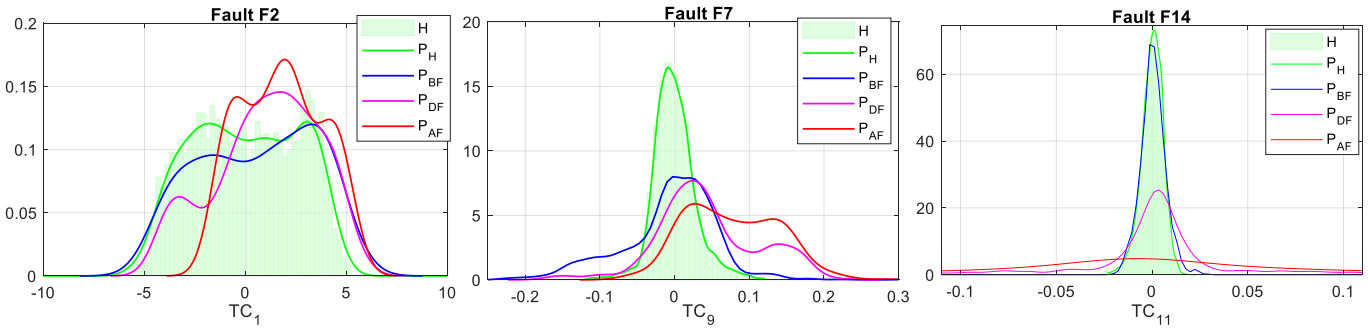


Figure 8. Reference and online densities of TCs before (BF), during (DF), and after (AF) faults.

A comparison of conventional PCA statistics such as Q and T^2 [32, 39] is made in Figures 7 and 8 across scenarios 2, 7, and 8 which respectively stand for line-to-line fault, open circuit fault, and current sensor fault in subsystems 2, 1, and 3, respectively. Figure 7 demonstrates the poor performance of both PCA Q and T^2 statistics [32, 39] to show any symptoms before and after faults are introduced at the 1500th sample, even though these statistics combine all the TCs. Figure 8 shows how far are the actual densities form Gaussian or Gamma distributions, it also highlights the time-varying characteristics of the PDFs across principle and residual block TCs. The Figure shows the real data projection histograms (H) during fault-free operation, and some smooth Kernel Density Estimates (KDE), as given in Eq. (11 to 14), at a few instances. Those KDEs are given for the indicated TCs including their PDFs during Healthy mode (P_H) that was used for training, these are the reference densities for those TCs of their corresponding subsystems, notice their overlapping with H due to the accuracy of smooth KDE. The PDF KDEs are also given in this Figure at other independent instances: Before the Fault (P_{BF}), During Fault (P_{DF}), and After Fault (P_{AF}) for the three situations. Notice first that the distributions are not normal, a condition that violates a heavy assumption of PCA and its statistics, notice also that these faults are characterized by distortions is the obtained PDFs rather than devotions of the squared distance in one direction. Approximating the PDFs of TC_1, TC_9, and TC_{11} with a Gaussian distribution, the parametrised KL approaches of [35,37] failed to detect fault F2 through TC_1 since there is no clear mean shift from and no change in variance during various experiments before and after the fault. TC_9 produces a false alarm due to the natural variation of density before the fault (P_{BF}). The same results are obtained if the parametrised Gamma distribution-based KL approach [38] is applied.

Recall Figure 6(a) where F7 cannot be classified from H along TC_9, and Figure 6(b) where TC_1 cannot discriminate F2 from H. Figure 8 shows that the dissimilarity between the PDFs at different instances

successfully discriminates the same faults along the same TCs, such little dissimilarities can be measured through the nonparametric KL divergence approach. This can be seen in the following through the clear detection of F7 through D_9.

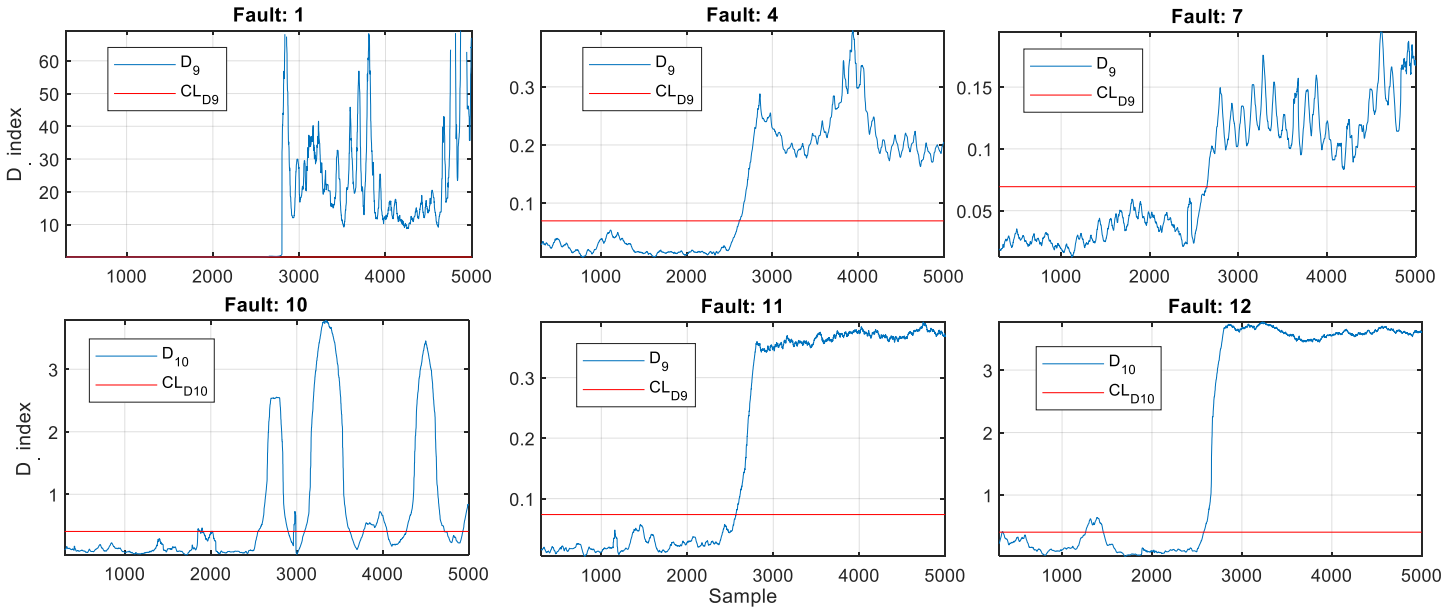


Figure 9. Response of sensitive indices $D_{\cdot}(t)$ to faults in subsystem S_1

This distortion and any small deviations, which PCA statistics failed to capture, can be accurately measured using the developed measures based on KL divergence. The PDFs of the online block TCs are estimated in the same manner as their reference densities, however, at each time new variable measurements are recorded through Eq. (15) and (16). Consequently, the KL divergence between the reference estimated PDFs of the reference TCs and their online block counterparts is evaluated each time through Eq. (10) and (19). The monitoring performance of the developed method is tested against all the scenarios as shown in Figures 9, 10, and 11 for RMPV subsystems S_1, S_2, and S_3 respectively, including those shown in Figures 7 and 8 where any performance deviation across the RMPV systems is measured with high accuracy.

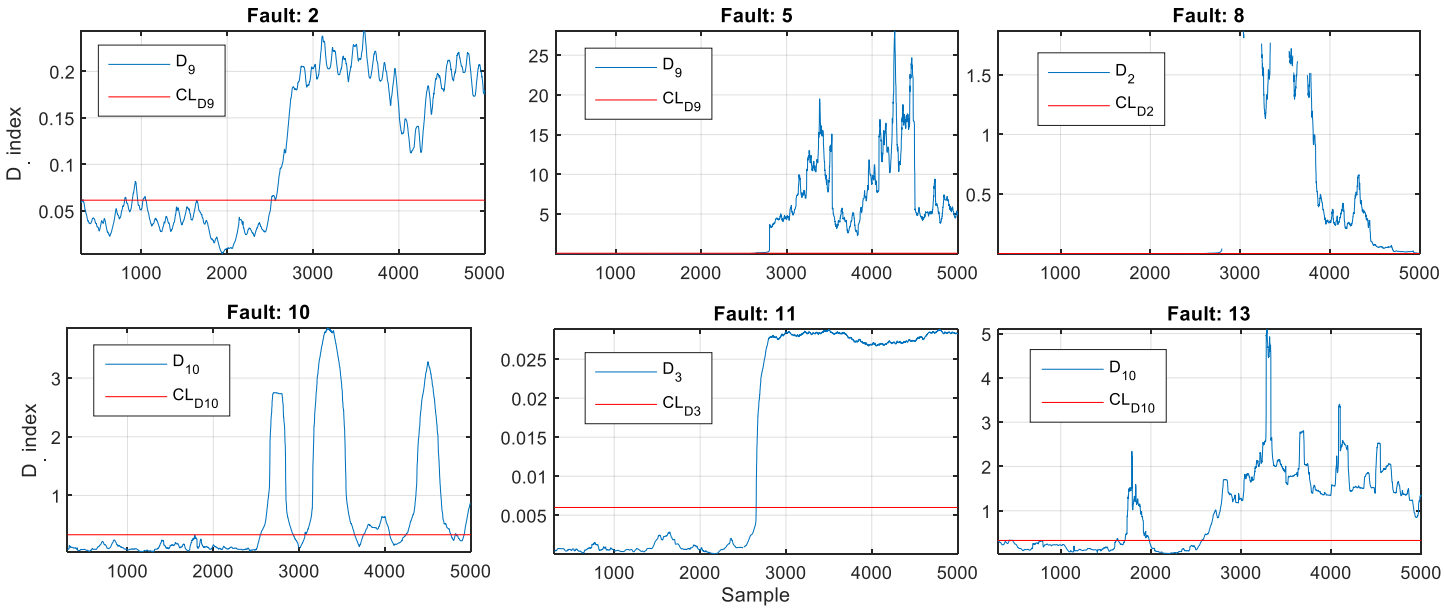


Figure 10. Response of sensitive indices $D_{\cdot}(t)$ to faults in subsystem S_2

Table 3. Detection time delay (samples) across the RMPV system scenarios.

Subsystem 1			Subsystem 2			Subsystem 3		
Fault	DTD	D_i	Fault	DTD	D_i	Fault	DTD	D_i
F1	31	9	F2	14	9	F3	67	3
F4	112	9	F5	128	9	F6	91	3
F7	146	9	F8	74	2	F9	98	2
F10	47	9	F10	47	10	F10	87	11
F11	66	8	F11	0	3	F11	29	11
F12	75	9	F13	66	10	F14	21	11

Figures 9 to 11 show the high sensitivity of the proposed deviation measures in addition to high robustness that can be tuned with some Control Limits of the Divergence ($CL_{D.}$). The fault detection alarms are then designed for the RMPV system based on the hypothesis tests as given in Eq. (17) and (18) according to their offline constructed thresholds tuned though an independent data set (Eq. (20)). The designed control limits allow for some negligible divergence attributed to measurement noise and training inaccuracies so that only a considerable divergence which is out of control at a given time point will trigger an alarm of a near-hazardous situation. The event Detection Time Delay (DTD) is then calculated as the faulty operation time before the fault is truly reported, this performance index is reported in Table 3 in addition to which monitoring index (D_i) had first triggered a fault alarm. Since the i^{th} index (D_i) has firstly detected the j^{th} fault, the i^{th} block TC is most sensitive to that fault and worth monitoring through nonparametric KL divergence for proper detection of that fault in the future.

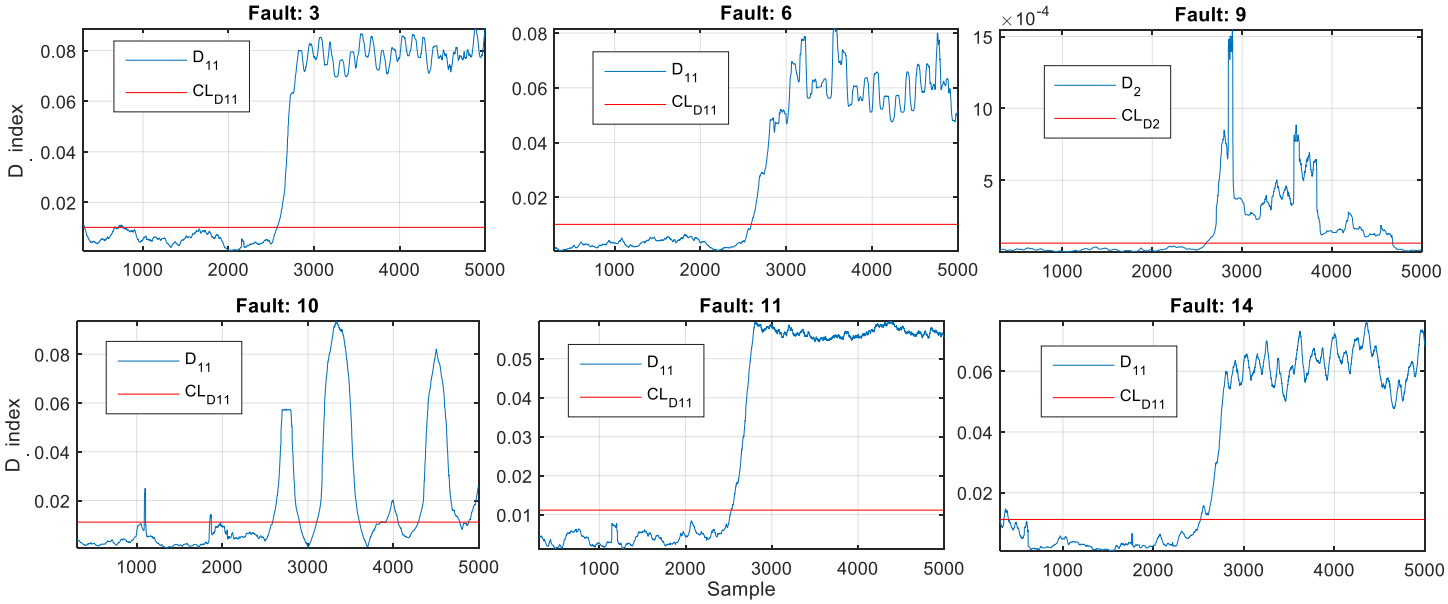


Figure 11. Response of sensitive indices $D_{\cdot}(t)$ to faults in subsystem S_3

While the conventional methods [32, 39] and [35 to 38] failed to detection dangerous faults F2 (line to line fault), F7 (open circuit fault), and F14 (feedback current sensor fault), as shown in Figures 7 and 8, these faults are successfully detected by the proposed method as demonstrated in Figures 9, 10, and 11. In addition to this superior detection performance, the proposed Kullback-divergence-MPCA approach ensures other potential applications over state-of-the-art methods as summarized in Table 4.

Table 4. Comparison of potential applications in real grid integrated RMPV.

	Proposed	[32]	[39]	[34]	[33]	[40]	[26]	[27]	[28]
Assumption-free	✓	x	x	✓	x	✓	✓	x	-
Real application	✓	x	✓	x	x	x	✓	✓	✓
Long term reliability	✓	✓	✓	✓	x	x	✓	x	✓
Not supervised	✓	✓	✓	✓	✓	✓	x	x	-
Efficient	✓	✓	✓	x	✓	✓	x	✓	-
Array mismatches	✓	x	x	x	✓	✓	✓	✓	✓
Inverter faults	✓	x	x	x	x	x	x	x	x
Grid perturbations	✓	x	x	x	x	x	x	x	x

Table 4 compares the proposed and different methods according to eight practical criteria. Assumption-free methods do not rely on any analytical models or theoretical assumptions on the system and its collected data which are reflected if an approach is investigated through real applications. Another criterion is if the approach is investigated for long term reliability and the ability to handle a large amount of data to ensure satisfactory results independently from the training instance (refer to monthly variations in Figure 5). A major concern is the requirement of labelled data for training on different types of mismatches and power quality deviations, this is an aspect of supervised approaches while the “not supervised” class includes semi-

supervised and unsupervised approaches only. Due to the fast dynamics and high-frequency high-dimensional data of grid-integrated RMPV systems, computation efficiency is important to realize the online real-time mismatch detection and power quality monitoring. A KL method such as [34] requires a computation time of several weeks in such applications with out-of-memory problems and it cannot be used in reality even if it is theoretically proved effective. More importantly, any approach must be checked for its proved reliability of detecting different types of RMPV system faults such as array mismatches, inverter faults, and grid perturbations. The different approaches are compared according to these criteria for which ✓, x, and (-) stands for positive, negative, and not-given evaluations, respectively. Notice that the presented approach outperforms the state-of-the-art methods and exhibits the most potential applications.

Conclusion

This article considers the safe operation of rooftop-mounted PV installations to avoid hazardous events and ensure a smooth injection to microgrid with good power quality. The work exploits several years of real measurements of three interconnected RMPV systems: Poly-Crystalline, Mono-Crystalline, and thin-film modules with their energy conversion systems with respective capacities of 2 kW, 2.025 kW, and 2.7 kW. Performance deviation is investigated through fourteen test scenarios which span array faults such as line to line, ground, and transient arc faults; DC-side mismatches in form of shadings and open circuits; grid-side anomalies such as voltage sags and frequency variations; in addition to inverter anomalies and sensor faults.

For this purpose, novel data-driven methodologies are developed for long-term performance monitoring and deviation measurement. Multiblock PCA is used in this article for statistical modelling and multivariate data decomposition and decorrelation to project the online measurements into block transformed components which are more sensitive and computationally efficient to analyze. Novel significant extensions are also proposed for accurate evaluation and robust alarm generation using Kullback-Leibler divergence through smooth kernel density estimation in a moving window approach.

The designed algorithms are explicitly based on multivariate analysis and information gain measure, these were applied to the analysis of large datasets of real measurements and tested against the fourteen different scenarios in the RMPV systems. While theoretical methods completely failed to detect line to line,

open circuit, and feedback current sensor faults in RMPV system, the presented design was proved highly effective for its assumption-free approach which successfully detected all faults in the experimented scenarios with acceptable performance. At the same time, the computationally-efficient algorithm is easily realized for online applications in RMPV systems. Moreover, the obtained results demonstrate the potential applications of the proposed strategies and outperform their conventional counterparts in terms of reliable indication of performance deviation with increased robustness and sensitivity.

References:

[1]. Eckhouse, B., 2018. Businesses Are Buying More Renewable Power Than Ever Before. Retrieved from Bloomberg database.website: https://www.bloomberg.com/news/articles/2018-04-30/businesses-are-buying-more-wind-and-solar-power-than-ever-before

[2]. Google., 2016. Achieving Our 100% Renewable Energy Purchasing Goal and Going Beyond. https://static.googleusercontent.com/media/www.google.com/fr//green/pdf/achieving-100-renewable-energy-purchasing-goal.pdf

[3]. Irfan, U., 2017. Energy hog Google just bought enough renewables to power its operations for the year. Retrieved from Vox.website: https://www.vox.com/energy-and-environment/2017/12/6/16734228/google-renewable-energy-wind-solar-2017

[4]. World Nuclear Association, 2017. Renewable Energy and Electricity. <http://www.world-nuclear.org/information-library/energy-and-the-environment/renewable-energy-and-electricity.aspx>

[5]. Deutsche Bank Markets Research, 2014. 2014 Outlook: Let the Second Gold Rush Begin. <https://www.deutschebank.nl/nl/docs/Solar_-_2014_Outlook_Let_the_Second_Gold_Rush_Begin.pdf>

[6]. Gagnon, P., Margolis, R., Melius, J., Phillips, C., Elmore, R., 2016. Rooftop Solar Photovoltaic Technical Potential in the United States: A Detailed Assessment. National Renewable Energy Laboratory. <https://www.nrel.gov/docs/fy16osti/65298.pdf>

[7]. Al-Saqlawi, J., Madani, K., Mac Dowell, N., 2018. Techno-economic feasibility of grid-independent residential roof-top solar PV systems in Muscat, Oman. Energy Convers. Manage. 178, 322-334.

[8]. Abreu, J., Wingartz, N., Hardy, N., 2019. New trends in solar: A comparative study assessing the attitudes towards the adoption of rooftop PV. Energy Policy. 128, 347-363.

[9]. Ma, W. W., Rasul, M. G., Liu, G., Li, M., Tan, X. H., 2016. Climate change impacts on techno-economic performance of roof PV solar system in Australia. Renew. Energ. . 88, 430-438.

[10]. Chaianong, A., Tongsopit, S., Bangviwat, A., Menke, C., 2019. Bill saving analysis of rooftop PV customers and policy implications for Thailand. Renew. Energ. 131, 422-434.

[11]. Bany Mousa, O., Kara, S., Taylor, R. A., 2019. Comparative energy and greenhouse gas assessment of industrial rooftop-integrated PV and solar thermal collectors. Appl. Energ. 241, 113-123.

[12]. Luerssen, C., Gandhi, O., Reindl, T., Sekhar, C., Cheong, D., 2019. Levelised Cost of Storage (LCOS) for solar-PV-powered cooling in the tropics. Appl. Energ. 242, 640-654.

[13]. Hu, J., Chen, W., Cai, Q., Gao, C., Zhao, B., Qiu, Z., Qu, Y., 2016. Structural behavior of the PV–ETFE cushion roof. Thin-Walled Struct. 101, 169-180.

[14]. Manzini, G., Gramazio, P., Guastella, S., Liciotti, C., Baffoni, G. L., 2015. The Fire Risk in Photovoltaic Installations - Test Protocols For Fire Behavior of PV Modules. Energy Procedia. 82, 752-758.

[15]. Poulek, V., Matuška, T., Libra, M., Kachalouski, E., Sedláček, J., 2018. Influence of increased temperature on energy production of roof integrated PV panels. Energ. Buildings. 166, 418-425.

[16]. Patsalides, M., Efthymiou, V., Stavrou, A., Georghiou, G. E., 2016. A generic transient PV system model for power quality studies. Renew. Energ. 89, 526-542.

[17]. Pullaguram, D., Bhattacharya, S., Mishra, S., Senroy, N., 2017. A Fuzzy Assisted Enhanced Control for Utility Connected Rooftop PV. IFAC-PapersOnLine. 50(1), 7693-7698.

[18]. Wu, Y., Zhou, J., 2019. Risk assessment of urban rooftop distributed PV in energy performance contracting (EPC) projects: An extended HFLTS-DEMATEL fuzzy synthetic evaluation analysis. Sustain. Cities Soc. 47, 101524.

[19]. Kaur, G., Vaziri, M. Y. (2006, 18-22 June 2006). *Effects of distributed generation (DG) interconnections on protection of distribution feeders*. Paper presented at the 2006 IEEE Power Engineering Society General Meeting.

[20]. Livera, A., Theristis, M., Makrides, G., Georghiou, G. E., 2019. Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems. Renew. Energ. 133, 126-143.

[21]. Paul, M. M. R., Mahalakshmi, R., Karuppasamypandiyan, M., Bhuvanesh, A., Ganesh, R. J., 2016. Classification and Detection of Faults in Grid Connected Photovoltaic System. Int. J. Sci. Eng. Res. 7(4), 149-154.

[22]. Du, B., Yang, R., He, Y., Wang, F., Huang, S. (2017). Nondestructive inspection, testing and evaluation for Si-based, thin film and multi-junction solar cells: An overview. Renewable and Sustainable Energy Reviews, 78, 1117-1151. doi:https://doi.org/10.1016/j.rser.2017.05.017.

[23]. Mellit, A., Tina, G. M., Kalogirou, S. A., 2018. Fault detection and diagnosis methods for photovoltaic systems: A review. Renew. Sust. Energ. Rev. 91, 1-17.

[24]. Drews, A., de Keizer, A. C., Beyer, H. G., Lorenz, E., Betcke, J., van Sark, W. G. J. H. M., Heydenreich, W., Wiemken, E., Stettler, S., Toggweiler, P., Bofinger, S., Schneider, M., Heilscher, G., Heinemann, D., 2007. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Sol. Energy. 81(4), 548-564.

[25]. Platon, R., Martel, J., Woodruff, N., Chau, T. Y., 2015. Online Fault Detection in PV Systems. IEEE Trans. Sustain. Energy. 6(4), 1200-1207.

[26]. Dhimish, M., Holmes, V., Mehrdadi, B., Dales, M., 2018. Comparing Mamdani Sugeno fuzzy logic and RBF ANN network for PV fault detection. Renew. Energ. 117, 257-274.

[27]. Zhao, Y., Lehman, B., Ball, R., Mosesian, J., Palma, J. d., 2013. *Outlier detection rules for fault detection in solar photovoltaic arrays*. Paper presented at the 2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC).

[28]. Zhao, Y., Balboni, F., Arnaud, T., Mosesian, J., Ball, R., Lehman, B., 2014. *Fault experiments in a commercial-scale PV laboratory and fault detection using local outlier factor*. Paper presented at the 2014 IEEE 40th Photovoltaic Specialist Conference (PVSC).

[29]. Cherry, G. A., Qin, S. J., 2006. Multiblock principal component analysis based on a combined index for semiconductor fault detection and diagnosis. IEEE Trans. Semicond. Manuf. 19(2), 159-172.

[30]. Westerhuis, J. A., Kourti, T., MacGregor, J. F., 1998. Analysis of multiblock and hierarchical PCA and PLS models. J. Chemom. 12(5), 301-321.

[31]. Qin, S. J., Valle, S., Piovoso, M. J., 2001. On unifying multiblock analysis with application to decentralized process monitoring. J. Chemom. 15(9), 715-742.

[32]. Bakdi, A., & Kouadri, A. (2018). An improved plant-wide fault detection scheme based on PCA and adaptive threshold for reliable process monitoring: Application on the new revised model of Tennessee Eastman process. Journal of Chemometrics, 32(5), e2978. doi:10.1002/cem.2978.

[33]. Mansouri, M., Hajji, M., Trabelsi, M., Harkat, M. F., Al-khazraji, A., Livera, A., . . . Nounou, M. (2018). An effective statistical fault detection technique for grid connected photovoltaic systems based on an improved generalized likelihood ratio test. Energy, 159, 842-856. doi:https://doi.org/10.1016/j.energy.2018.06.194.

[34]. Hamadouche, A., Kouadri, A., & Bakdi, A. (2017). A modified Kullback divergence for direct fault detection in large scale systems. Journal of Process Control, 59, 28-36. doi:https://doi.org/10.1016/j.jprocont.2017.09.004.

[35]. Harmouche, J., Delpha, C., Diallo, D., 2014. Incipient fault detection and diagnosis based on Kullback–Leibler divergence using Principal Component Analysis: Part I. Signal Process. 94, 278-287.

[36]. Harmouche, J., Delpha, C., Diallo, D., 2015. Incipient fault detection and diagnosis based on Kullback–Leibler divergence using principal component analysis: Part II. Signal Process. 109, 334-344.

[37]. Chen, H., Jiang, B., Lu, N., 2018. An improved incipient fault detection method based on Kullback-Leibler divergence. ISA Trans. 79, 127-136.

[38]. Delpha, C., Diallo, D., Youssef, A., 2017. Kullback-Leibler Divergence for fault estimation and isolation: Application to Gamma distributed data. Mech. Syst. Signal Process. 93, 118-135.

[39]. Bakdi, A., Kouadri, A., & Mekhilef, S. (2019). A data-driven algorithm for online detection of component and system faults in modern wind turbines at different operating zones. Renewable and Sustainable Energy Reviews, 103, 546-555. doi:https://doi.org/10.1016/j.rser.2019.01.013.

[40]. Fezai, R., Mansouri, M., Trabelsi, M., Hajji, M., Nounou, H., & Nounou, M. (2019). Online reduced kernel GLRT technique for improved fault detection in photovoltaic systems. Energy, 179, 1133-1154. doi:https://doi.org/10.1016/j.energy.2019.05.029.

[41]. SMA Solar Technology AG, 2014. SUNNY BOY 1300TL / 1600TL / 2100TL- Operating Manual. Germany. <http://files.sma.de/dl/5684/SB13-21TL-BE-en-11.pdf>

[42]. SMA Solar Technology AG, 2012. Sunny Boy 1600TL - Specification Sheet. <https://www.solarchoice.net.au/wp-content/uploads/Sunny-boy-sma-solar-inverter-1600tl-spec-sheet.pdf>

[43]. SMA Solar Technology AG, 2011. SUNNY BOY 2000HF-US / 2500HF-US / 3000HF-US - Easy installation, simple communication and maximum performance. <http://files.sma.de/dl/9524/SB3000HFUS-DEN113412W.pdf>

[44]. SMA Solar Technology AG, 2012. SUNNY BOY 2000HF / 2500HF / 3000HF - Installation Manual.

[45]. SMA Solar Technology AG, 2010. SUNNY SENSORBOX - Installation Guide. <https://files.sma.de/dl/4148/Sensorbox-IEN100914.pdf>

[46]. SMA Solar Technology AG, 2010. SUNNY SENSORBOX - The weather station for PV plants. <https://files.sma.de/dl/4148/SENSORBOX-DEN103131W.pdf>

[47]. SMA Solar Technology AG, 2010. SUNNY WEBBOX - Remote monitoring and maintenance of large solar power plants. <https://files.sma.de/dl/2585/WEBBOX-DEN102530.pdf>

[48]. SMA America, LLC, 2013. SUNNY WEBBOX - User Manual. <https://files.sma.de/dl/4253/SWebBox-BA-US-en-34.pdf>

[49]. SMA Solar Technology AG, SUNNY BOY 3000TL / 4000TL / 5000TL - Parameters and measured values. <https://www.energymatters.com.au/images/SMA/SBNG_PAR-TEN084410.pdf>

[50]. SMA America, LLC, 2011. SUNNY BOY 2000HF-US/2500HF-US/3000HF-US - Technical Description. <https://files.sma.de/dl/9524/HF_STP_Par-eng-TUS114810.pdf>

[51]. Ibrahim, H., Anani, N., 2017. Variations of PV module parameters with irradiance and temperature. Energy Procedia. 134, 276-285.

[52]. Skandalos, N., Karamanis, D., 2016. Investigation of thermal performance of semi-transparent PV technologies. Energ. Buildings. . 124, 19-34.

[53]. Pillai, D. S., Rajasekar, N., 2018. A comprehensive review on protection challenges and fault diagnosis in PV systems. Renew. Sust. Energ. Rev. 91, 18-40.

[54]. Nair, N.-K. C., Jing, L., 2013. Power quality analysis for building integrated PV and micro wind turbine in New Zealand. Energ. Buildings. 58, 302-309.

[55]. Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., Chen, H., 2018. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energ. Convers. Manage. 178, 250-264.

[56]. Castro, R., 2018. Data-driven PV modules modelling: Comparison between equivalent electric circuit and artificial intelligence based models. Sustainable Energy Technol. Assess. 30, 230-238.

[57]. Rezk, H., Tyukhov, I., Al-Dhaifallah, M., Tikhonov, A., 2017. Performance of data acquisition system for monitoring PV system parameters. Measurement. 104, 204-211.

[58]. Malvoni, M., De Giorgi, M. G., Congedo, P. M., 2017. Forecasting of PV Power Generation using weather input data-preprocessing techniques. Energy Procedia. 126, 651-658.

[59]. Tong, C., Yan, X., 2017. A Novel Decentralized Process Monitoring Scheme Using a Modified Multiblock PCA Algorithm. IEEE T. Autom. Sci. Eng. 14(2), 1129-1138.

[60]. Worley, B., Powers, R., 2015. A sequential algorithm for multiblock orthogonal projections to latent structures. Chemom. Intell. Lab. Syst. 149, 33-39.

[61]. Cao, X. H., Stojkovic, I., Obradovic, Z., 2016. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics. 17(1), 359.

[62]. Kullback, S., 1997. Information Theory and Statistics. Courier Corporation, North Chelmsford,

[63]. Ponti, M., Kittler, J., Riva, M., Campos, T. d., Zor, C., 2017. A decision cognizant Kullback–Leibler divergence. Pattern Recognit. 61, 470-478.

[64]. Zeng, J., Kruger, U., Geluk, J., Wang, X., Xie, L., 2014. Detecting abnormal situations using the Kullback–Leibler divergence. Automatica. 50(11), 2777-2786.

[65]. Kawahara, Y., Sugiyama, M., 2012. Sequential change-point detection based on direct density-ratio estimation. Statistical Analysis and Data Mining: The ASA Data Science Journal. 5(2), 114-127.

[66]. Liu, S., Yamada, M., Collier, N., Sugiyama, M., 2013. Change-point detection in time-series data by relative density-ratio estimation. Neural Networks. 43, 72-83.

[67]. Trapero, J. R., 2016. Calculation of solar irradiation prediction intervals combining volatility and kernel density estimates. Energy. 114, 266-274.

[68]. Gramacki, A., 2018. Nonparametric Kernel Density Estimation and Its Computational Aspects, 1 ed. Springer International Publishing.