redin Revista Facultad de Ingeniería, Universidad de Antioquia, No.102, pp. 26-43, Jan-Mar 2022

Method of monitoring and detection of failures in PV system based on machine learning

Método de monitoreo y detección de fallos en el sistema fotovoltaico basado en aprendizaje automático

Darío Javier Benavides [1], Páúl Arévalo-Cordero [2, 3], Luis G. Gonzalez [2], Luis Hernández-Callejo [4*], Francisco Jurado [3], José A. Aguado [1]

[1]Campus de la Universidad de Málaga, Universidad de Málaga. Av. de Cervantes, 2. C. P. 29071. Málaga, España.
[2]Campus Tecnológico Balzay, Universidad de Cuenca. Av 12 de Abril &. C. P. 010107. Cuenca, Ecuador.
[3]Campus Científico-Tecnológico Linares, Universidad de Jaén. Ronda Sur s/n Campus, SG-318. C. P. 23700. Linares, España.
[4]Campus de la Universidad de Soria, Universidad de Valladolid. C/Plaza de Santa Cruz, 8. C. P. 42004 Soria, España.

CITE THIS ARTICLE AS:

D. J. Benavides, P. Arévalo, L. G. González, L. Hernández, F. Jurado and J. A. Aguado. "Method of monitoring and detection of failures in PV system based on machine learning techniques", *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 102, pp. 26-43, Jan-Mar 2022. [Online]. Available: https://www.doi.org/10.17533/udea.redin.20200694

ARTICLE INFO:

Received: November 25, 2019
Accepted: July 01, 2020
Available online: July 01, 2020

KEYWORDS:

Artificial intelligence; renewable energy sources; monitoring

Inteligencia artificial; fuentes de energía renovable; supervisión

ABSTRACT: Machine learning methods have been used to solve complicated practical problems in different areas and are becoming increasingly popular today. The purpose of this article is to evaluate the prediction of the energy production of three different photovoltaic systems and the supervision of measurement sensors, through Machine learning and data mining in response to the behavior of the climatic variables of the place under study. On the other hand, it also includes the implementation of the resulting models in the SCADA system through indicators, which will allow the operator to actively manage the electricity grid. It also offers a strategy in simulation and prediction in real-time of photovoltaic systems and measurement sensors in the concept of smart grids.

RESUMEN: Los métodos de aprendizaje automático se han utilizado para resolver problemas prácticos complicados en diferentes áreas y se están volviendo cada vez más populares hoy en día. El propósito de este artículo es evaluar la predecición de la producción de energía de tres sistemas fotovoltaicos diferentes y la supervision de sensores de medición, por medio un aprendizaje automático y minería de datos en respuesta al comportamiento de las variables climáticas del lugar en estudio. Por otro lado, también incluye la implementación de los modelos resultantes en el sistema SCADA por medio de indicadores, que permitirá al operador gestionar activamente la red eléctrica. Ademas ofrece una estrategia en la simulación y predicción en tiempo real de sistemas fotovoltaicos y sensores de medición en el concepto de redes inteligentes.

1. Introduction

The integration of Renewable Energies (RE) in the electric grid intensifies the complexity of electric grid management to maintain service continuity and the production -consumption balance, due to the intermittent and unpredictable nature [1, 2]. Therefore, it is necessary to focus more on research and development in government and other levels to explore RE resources and meet energy needs globally [3].

The prediction in photovoltaic (PV) production is necessary for the optimal integration of this technology in the existing power systems and is an important factor for the operators of the electric grid [2, 4] However, there are two main

* Corresponding author: Luis Hernández Callejo
E-mail: luis.hernandez.callejo@uva.es
ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.20200694

concerns about the implementation of PV systems in high penetration rates, intermittent nature and uncertainty of availability [5]. In addition, poorly functioning photovoltaic panels can cause gradual or rapid falls in the amount of energy generated. One study shows that it is possible to predict the daily power curve of a photovoltaic panel depending on the power curves of neighboring panels, by applying neural networks which allows monitoring the correct operation [6]. The precise forecast of PV production can mitigate the effects of energy quality that represent large quantities of distributed systems through the active management of electric grid and is an important feature that can help companies and operators in energy management and economic dispatch planning [2]. The power generated by a PV system at a given time is proportional to the solar radiation received by the panel. However, the radiation varies due to the seasons and for several hours of the day, depending on the geographical location and orientation of the panel [6]. Therefore, it is important that the solar radiation and the corresponding energy production be predicted, so that the operator can acquire the appropriate measures and manage the intermittency [7]. The methodology proposed in [4] is based on the implementation of the sensor in the RES operation and is considered big data technologies for the processing and analysis of data for the prediction of PV systems. In the article [8], authors presents a complete review of the forecast of the generation of PV energy based on machine learning and metaheuristic techniques, which is represented in classifiers (i) *Persistence method*, (ii) *Statistical approaches*, (*iii*) *Machine learning approaches* and (*iv*) *hybrid techniques*. In addition, according to the Classification of PV power forecasting based on time: Very short-term (1 sec - <1h), Short-term (1h - 24h), Medium-term (1 week-1 month) and Long-term (1 month-1 year).

Advanced data analysis applications with functionality and versatility allow managing energy system information to analyze and extract information, for example: improving energy quality, more efficient distribution, optimization, machine learning, among others. Under this same criterion, the slogans are corrected as new information is known [9, 10]. However, to obtain an acceptable model it is necessary to analyze a large amount of data for its training, considered thus an inconvenience in new systems and applications where this information is not yet available. Especially the technology of "Big Data" (BD) applied in the energy system that is currently in its initial stage and there is a long way to go [9].

In [11] an intelligent method of fault diagnosis for PV arrays based on an improved rotation forest algorithm has been proposed. This consists of the selection and classification of characteristics based on two rotation forest (RoF) algorithm classifiers ensemble hybridized with extreme learning machine (ELM) for fault diagnosis of PV arrays. In addition, a PV system of 9.54 kWp has proposed a new procedure for the detection and diagnosis of PV system failures, based on a red probabilistic neuronal classifier (PNN) [12].

There are several related studies of solar photovoltaic systems, from the point of view of modeling and simulation, however, this behavior in these systems is not always the same, because the climatic conditions are different in each part of the world. For this reason, it is necessary to carry out an additional study in real measured data to observe its behavior under normal operating conditions. In this study, we present a prediction model of electric power (kW) generated by three different photovoltaic systems: polycrystalline, monocrystalline and one-axis tracking. Using machine learning and data mining techniques applied in SCADA databases (Supervision Control and Data Acquisition) and climatic data obtained from the weather station, whose main objective is to establish a model using real-time indicators, which will allow establishing a comparison between actual PV production compared to the PV production of the model. Additionally, a focus has been carried out on fault detection methods with the obtained equations. The results were implemented in the SCADA system which allows the operator to obtain a better reference in the monitoring, control and detection of failures on the production of photovoltaic energy. Finally, this article is an extension of the document published at the ICSC-CITIES 2019 conference, entitled *"Machine learning data applied to monitoring PV systems: A case study*"[13]. Among the novelties of this article lie in the application of the PVS1 PVS2 and PVS3 models for the detection of faults in photovoltaic systems in a comparative way between the photovoltaic generation measured in real time and the value of the model estimated with the calculation of meteorological variables. Establishing a 20% allowable range between the actual value and the specific value to determine if a measurement failure or error has occurred with alarm indicators.

2. Applications in energy management

There are many challenges ahead in terms of the BD Technology of smart grids, such as: data integration and storage, real-time data processing technology, data compression, great technology data visualization and privacy and data security [9]. Figure 1 shows some current applications of BD and "Machine learning" focused in terms of RE management for smart cities [14]. Analytics and big data can help with processing large amounts of historical data, thus increasing the wind, solar and loading

Nomenclature

$PVS1$	photovoltaic system 1
$PVS2$	photovoltaic system 2
$PVS3$	photovoltaic system 3
$f(t)$	approximate value of the temperature
$g(t)$	solar radiation limit values
$PVS1_{Cloudless}(t)$	prediction clear days' power of the photovoltaic system 1
$PVS2_{Cloudless}(t)$	prediction clear days' power of the photovoltaic system 2
$PVS3_{Cloudless}(t)$	prediction clear days' power of the photovoltaic system 3
$P_{PVS1}(rad, temp)$	prediction of the power of the photovoltaic system 1
$P_{PVS2}(rad, temp)$	prediction of the power of the photovoltaic system 2
$P_{PVS3}(rad, temp)$	prediction of the power of the photovoltaic system 3
$P_{PVS1(real-time)}$	power of photovoltaic system 1 in real-time
$P_{PVS2(real-time)}$	power of photovoltaic system 2 in real-time
$P_{PVS3(real-time)}$	power of photovoltaic system 3 in real-time
$a0$,w	coefficient of the equations of the models Fourier 4th
$a1$,$a2$,$a3$,$a4$	coefficient of the equations of the models Fourier 4th and Sum of Sine 4th
$b1$,$b2$,$b3$,$b4$	coefficient of the equations of the models Fourier 4th and Sum of Sine 4th
$c1$,$c2$,$c3$,$c4$	coefficient of the equations of the models Sum of Sine 4th
SSE	sum of squares due to error
$RMSE$	root of the mean square error
$temp$	ambient temperature
rad	solar radiation
t	time

forecast accuracy [10]. All these topics can be analyzed from a database, through machine learning techniques and their derivatives. Therefore, progress in different energy management applications with ER sources and distributed storage systems in smart grids is under development.

Some techniques of Artificial Intelligence (AI) can be applied as an effective method to achieve the future objectives of renewable energy [3]. AI is used in almost all RE types (wind, solar, geothermal, hydroelectric, oceanic and hydrogen) for *design, optimization, control, estimation, management, distribution* and *economics*. The present and future in RE consists mainly of the development of innovative technology for optimal production from the available natural resources, in environmental awareness, and the best management and distribution system, as mentioned in previous studies. Like other domains (health, education, business, technology, industry, security, etc.), AI could help achieve the future objectives of the RE and within it also shares its study of machine learning [14].

Machine learning methods have been used to solve complicated practical problems in different areas and are becoming increasingly popular today [15]. For example, the energy system presents some important challenges for microgrid and power management of smart grids, advanced technologies that use sensors and actuators, IT&C, big data, complex analytics, etc. [10, 16]. A study using the data mining algorithm "Decision Tree" with a black-box model based on *Random Forest* proposes the detection and classification of microgrid faults [9].

3. Case study

These case studies were developed in the energy laboratory of the Universidad de Cuenca. Figure 2, shows the location of the energy laboratory, located in the area GTM-5 with geographic coordinates (-2.8919, -79.0385). The laboratory is made up of a 35kW photovoltaic system with a connection to the electricity grid and a weather station that records data [17].

An example of data analysis and tools applied to the modeling and simulation of a photovoltaic system of this study is presented in [18], Its structure consists of 4 stages, *data acquisition, modeling and simulation, validation* and finally its *applications*. This study analyzes its application for the monitoring and detection of PV system failures through the implementation of equations resulting from an analysis of databases in the SCADA system, considering the databases of the PV system and the weather station. In Figure 3, the structure of the example of a machine learning application in a mentioned PV system is described. For the study, a database generated by the data logger of the SCADA system is necessary.

3.1 PV System

The Photovoltaic System (PVS) consists of 136 panels of 250Wp of the Atersa brand, 60 of the monocrystalline type (15kW), 60 of the polycrystalline type (15kW) and 16 of the polycrystalline type with tracking on an axis (5kW). Each system is connected to a separate DC/AC inverter with the

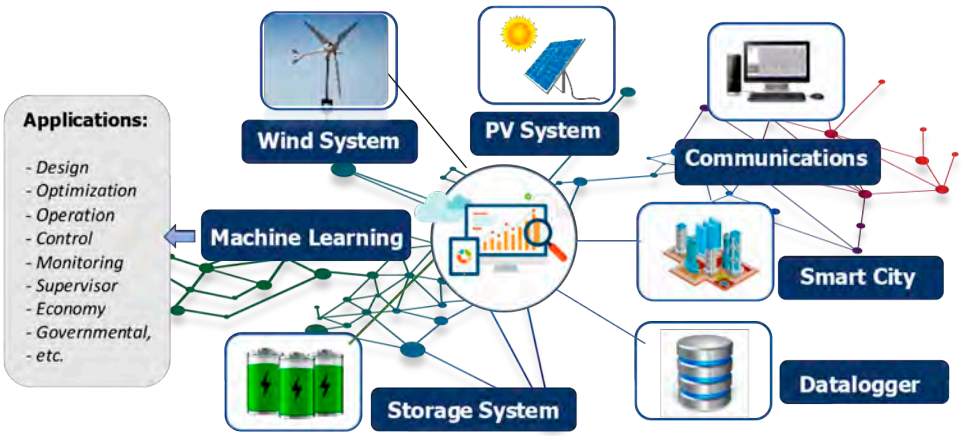


Figure 1 Machine learning applications in energy systems

Figure 2 Energy laboratory for the case study

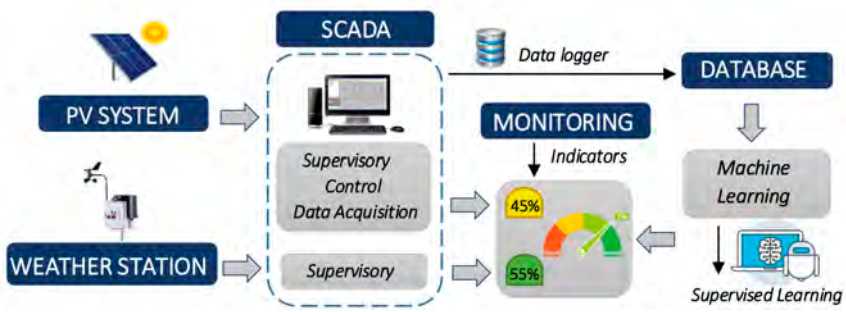


Figure 3 Example of machine learning application in a PV system

public grid (See Figure 2) [17]. Table 1 summarizes the characteristics of these systems.

3.2 Weather Station

A weather station located in the study area and at the same level of the photovoltaic installation (See Figure 2), obtains the climatic information of the zone corresponding to the variables of solar radiation (W/m^2) [measuring range, precision][1-1,250 W/m^2, ± 5%], ambient temperature (°C) [-20° to 70°C, ± 0.6°C], relative humidity (%) [20% to 100%, ± 3%], precipitation (cm) [0.01"(0.25cm), ± 2%], wind speed (m/h) [0-175 m/h, ±5%], wind direction (°) [2° increments, ± 7°] and wind gust (m/h) [0-175 m/h, ± 5%]. For the study, these variables were assigned a sampling rate of 1 minute.

3.3 Database

The database is created with the variables mentioned above and stored in the server's datalogger through the *Datalogging and Supervisory Control Module and the Database connectivity* tool of LabVIEW 2015 software with the communication interface through *NI OPC servers*. In this way, the value of the variables is acquired and their conversion from analog to digital signals for later viewing and storage. The database for the study corresponds to the one-year record.

Table 1 Characteristics of photovoltaic systems

Item	Description	Panels (series-parallel)	Type	Max. Power
1st	PVS1 Photovoltaic System 1	60 (15 × 4)	Polycrystalline	15 kW
2nd	PVS2 Photovoltaic System 2	60 (15 × 4)	Monocrystalline	15 kW
3th	PVS3 Photovoltaic System 3	16 (16 × 1)	Polycrystalline (tracking)	5 kW

3.4 Monitoring

The main objective of this research focuses on this point. It proposes an application model in the monitoring of this photovoltaic installation. It also analyzes the information of the database and its implementation of the models in the SCADA system, as a support or help in case of failures in the PV electrical system or measurement errors in the sensors as it exceeds a preset range, by means of indicators related to the variables of the PV system and the real-time weather station. In this way, it allows the electric grid operator to improve the safety and reliability of the PV system and its integration with the grid.

3.5 SCADA System

The SCADA allows the control of the PV system in activation and disconnection of the DC/AC inverters, from the electric grid, it also allows the supervision of the electrical variables of current, voltage and power in DC, current, voltage, power, (active, reactive and apparent) in AC of each PV system. Data acquisition is performed by means of measuring sensors and network analyzers, connected by means of a PLC through Modbus communication that allow reading/writing and storing the information in the local server [17].

4. Machine learning

The advancement of computing in recent years and the reduction in the cost of hardware allows us to develop some new methods of extracting information. Machine learning is a branch of artificial intelligence and deals with the construction and study of systems that can learn from data sets by giving computers the ability to learn without being explicitly programmed [1]. Machine learning identifies knowledge and patterns in data, which is currently considered one of the most useful techniques for extracting information [9]. Machine learning algorithms use computational methods to "learn" information directly from the data without relying on a predetermined equation as a model [19]. In general, there are nine most used machine learning algorithms, "including k-means, Linear Support Vector Machines (LSVM), Logistic Regression (LR), Locally Weighted Linear Regression (LWLR), Gaussian Discriminant Analysis (GDA), Back-propagation Neural Network (BPNN), Expectation Maximization (EM), Naive Bayes (NB) and Value-Added Tax (VAT). Each of the algorithms has its own characteristics and can be used under different scenarios" [9]. Several methodologies based on artificial intelligence such as machine learning, Genetic Algorithm (GA) and Neural Networks (NN) have been proposed and applied for the modeling and forecasting of solar irradiance [15]. Although the main drawback of neural networks is the long training time and many parameters that require the intervention of the user [5].

It is also important to consider that a greater amount of data and relevant information (data mining) with respect to a better study topic will improve the possibilities of finding an appropriate application model. Considering the objectives of the application of the case study, data analysis is projected through *supervised learning*, which is, creating a predictive model from known input and response data. In this type of learning two categories of algorithms are used: *classification* destined to databases that include qualitative values (words) and *regression* for quantitative (numerical) databases. Based on the numerical characteristics of the databases obtained from the SCADA system, it will be used in the *regression* category for the analysis.

Figure 4 shows the workflow to establish an ideal model defined under two criteria in *training* and *application phases*. In the training phase to effectively apply a learning technique to a performance function, it must be subdivided into 4 stages: initially enter the database, then perform a preprocessing (filters, statistical summary, cluster analysis), then define the category of supervised learning (classification or regression) and finally the model is obtained. In the *application phase*, the new database is entered, the preprocessing is performed again, the model obtained in the training phase is applied to identify the important characteristics and architectural parameters of each model, to define the predictive model. Finally, the performance of the model is evaluated [2, 19].

4.1 Training phase (75% of data)

In the application of a predictive model for the estimation of solar energy, the databases of both the PV system and weather station will be studied under the same sampling

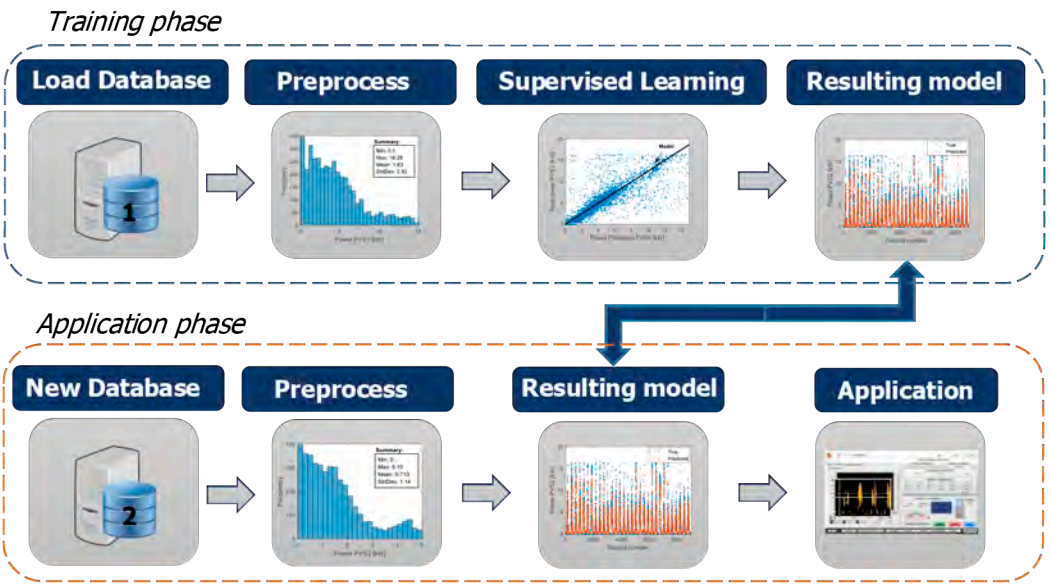
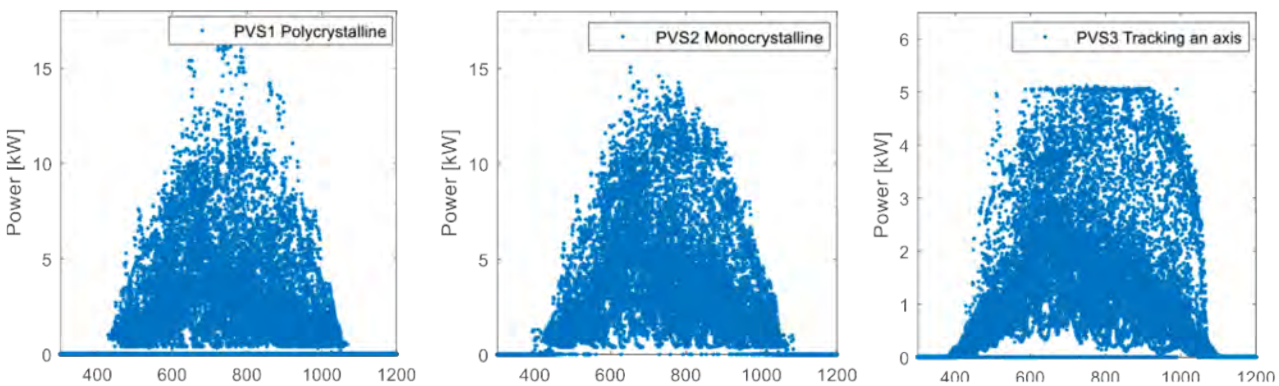


Figure 4 Ideal workflow of Machine Learning [19]



(a) PSV1 **(b)** PSV2 **(c)** PSV3

Figure 5 Power generated during one month (July 2018)

rate corresponding to one year, which allows establishing a correct relation of variables of equitable way. Before applying a learning algorithm to the model, it is necessary to perform a preprocessing, where the database must be analyzed, for example in SCADA systems it is very frequent the loss of information from a record, either in errors of measurements of the sensors or in the discretization processing of analog reading variables. For this reason, it is necessary to be familiar with all the variables in the database, in order to verify some type of irregularities. An effective method in the case that the database has lost values, is to replace its value by the average of the previous and subsequent data in the registry.

The ideal methodology for developing optimal models of machine learning for predictions of PV energy should include a training phase (one-year), a validation phase (per month) and a test phase (per month), as shown by a similar study in [2]. Figure 5 shows the production of PV power (kW) defined as a variable to be predicted under the supervised learning criterion. The records correspond to a one-month test database (43,200 $\times$ 3 data) for the PVS1, PVS2 and PVS3 systems. The maximum generation limit for PVS1, PVS2 is 15kW and for PVS3 is 5kW. This power is limited by the inverters connected to the power grid. In this way it is possible to observe the differences in the electrical generation by the three PV systems. The PVS3 (5kW) differs from other systems, due to its structure of solar tracking on an axis, which allows obtaining a better use of solar energy. An average percentage increase of 15 % is estimated for the same installed power capacity with respect to the fixed system in a study conducted during the test phase for the month of July 2018. In general, solar sensors provide information about the relative position of

the sun. This is very useful information for the tracking system that has a better use with approximately 30% more than a fixed system according to the size of the PV system [10].

The prediction of PV generation depends largely on the study of the behavior of the system, because it plays an important role in the operational management of power grids [16]. The data of the weather station as a function of the variable to predict Power PVS1, are shown in Figure 6. It is possible to observe its correlation with the variables of solar radiation, temperature, relative humidity, wind speed, wind gust and precipitation respectively. A possible linear regression model can be clearly observed with respect to the variable of solar radiation. Another point of interest is the inverse relationship between the parameters of temperature and relative humidity. On the other hand, the variables of wind speed, wind gust, wind direction and rainfall do not show a pattern of interest in the model. This analysis can also be observed when constructing the correlation matrix between the aforementioned variables, as established in a preliminary study in [18].
Figure 7 presents a statistical summary of the meteorological variables of solar radiation and temperature, as well as the variables to predict PVS1, PVS2 and PVS3, according to what is established in the data preprocessing (See Figure 6), the maximum, minimum, average and standard deviation values are indicated. It should be noted that these values were calculated considering 24 hours a day, so that the zero values generated during the night have been suppressed in the graphs since there is no radiation and consequently no power.

In the database corresponding to the weather station, *solar radiation* and *temperature* are defined as the most influential variables on the photovoltaic prediction models, as detailed in a preliminary study [18]. Figure 8a shows the behavior of the temperature of the maximum and minimum values during the 24 hours of the day corresponding to the month of August 2018 (test phase). An average of variations between 8 - 22 °C can be observed and it is possible to define a model of the temperature by means of the "sum of sine" function defined in Equation 1 [18]:

$$f(t) = 15.87 \cdot sin(0.05 \cdot t + 0.686) + 2.664 \cdot sin(0.4139 \cdot t + 2.199) \quad (1)$$

Where $f(t)$ represents the approximate value of the temperature in the study area and "t" corresponds to the record variable (hour-minutes). In addition, it is possible to observe that the temperature decreases in the night hours and increases its maximum values with the solar radiation data during the midday hours.

A machine-learning study for the prediction of the solar radiation of the PV system, shows that it is possible to obtain a model with great approximation, using the forecast parameters the variables of temperature, relative humidity, wind speed and irradiance [15]. Using a model, it has been established that there is a large percentage of data below the curve of the solar radiation limit values $g(t)$, according to Equation 2, where "t" indicates the number of records (value between 0 – 290 for the example), obtained through machine learning in Matlab:

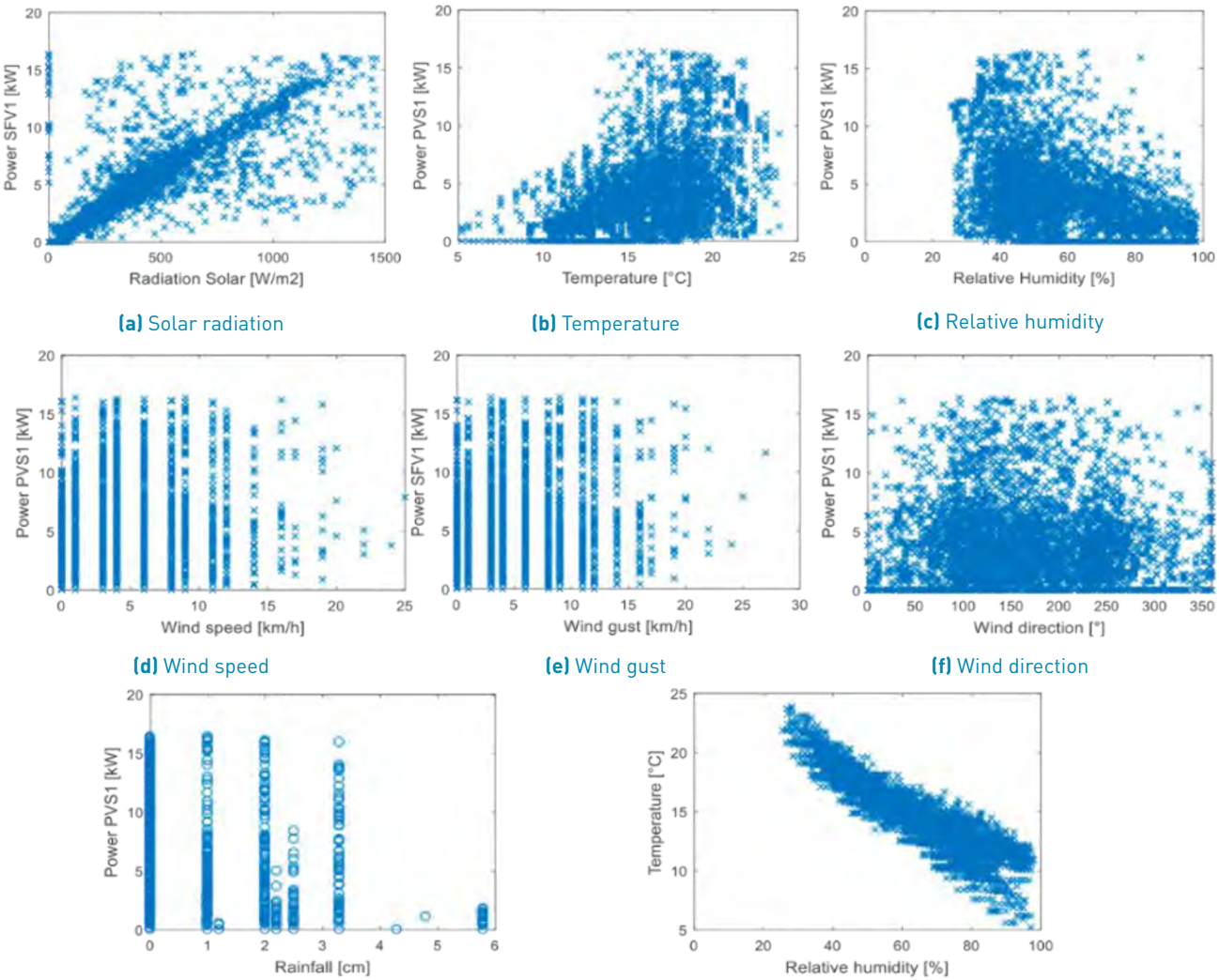
$$g(t) = 477.5 \cdot \sin(0.0017 \cdot t + 1.345) + 276.75 \cdot \sin(0.0498 \cdot t + 0.3721) + 723.25 \cdot \sin(0.0249 \cdot t - 2.16) \quad (2)$$

Figure 8a and Figure 8b, show the results of the application of Equations 1 and 2, together with the values of temperature and solar radiation respectively, corresponding to the month of August 2018 (test phase). It should be noted that the same procedure performed for the rest of the months of the year.

Although, in this study it is necessary to consider an important factor of cloudiness, since according to the State Agency of Metrology (SAMET) the areas of maximum cloudiness are in the equatorial zone and between 60 and 70º. A single cloud that passes can bring the energy production of a solar farm from the total production to the minimum and return to the whole in a matter of minutes or even seconds [7]. This information can be seen in Figure 9a, where there is a large amount of solar radiation data that drastically change during daylight hours. This effect directly affects the production of energy in PV systems. Figure 9b shows the power generated by the three photovoltaic systems under study, under practically zero cloud conditions. Mitigation measures for large drops in solar radiation, such as response to demand, storage and scheduling within an hour can only be maximized with accurate and reliable forecasting [7].

Figure 10 shows a comparison between the PV production in two completely different scenarios, considering the PV production on a normal day, compared to a completely cloudless day. A symmetric distribution can be observed at 12:00 a.m. for the PVS1 and PVS2 systems, with an amplitude difference of 1.5kW. On the other hand, the PVS3 represents an increase in the generation during the afternoon hours (13:00pm - 18:00pm). These results of symmetric data are due to the fact that the case of studies is close to the equator, where its maximum production point is around noon 12:00 pm for fixed PV systems.

In Table 2, the results of the application of some models are described in terms of cloudless conditions on the power curve of the PVS1, PVS2 and PVS3 systems, defined



(a) Solar radiation **(b)** Temperature **(c)** Relative humidity

(d) Wind speed **(e)** Wind gust **(f)** Wind direction

(g) Rainfall **(h)** Temperature vs Relative humidity

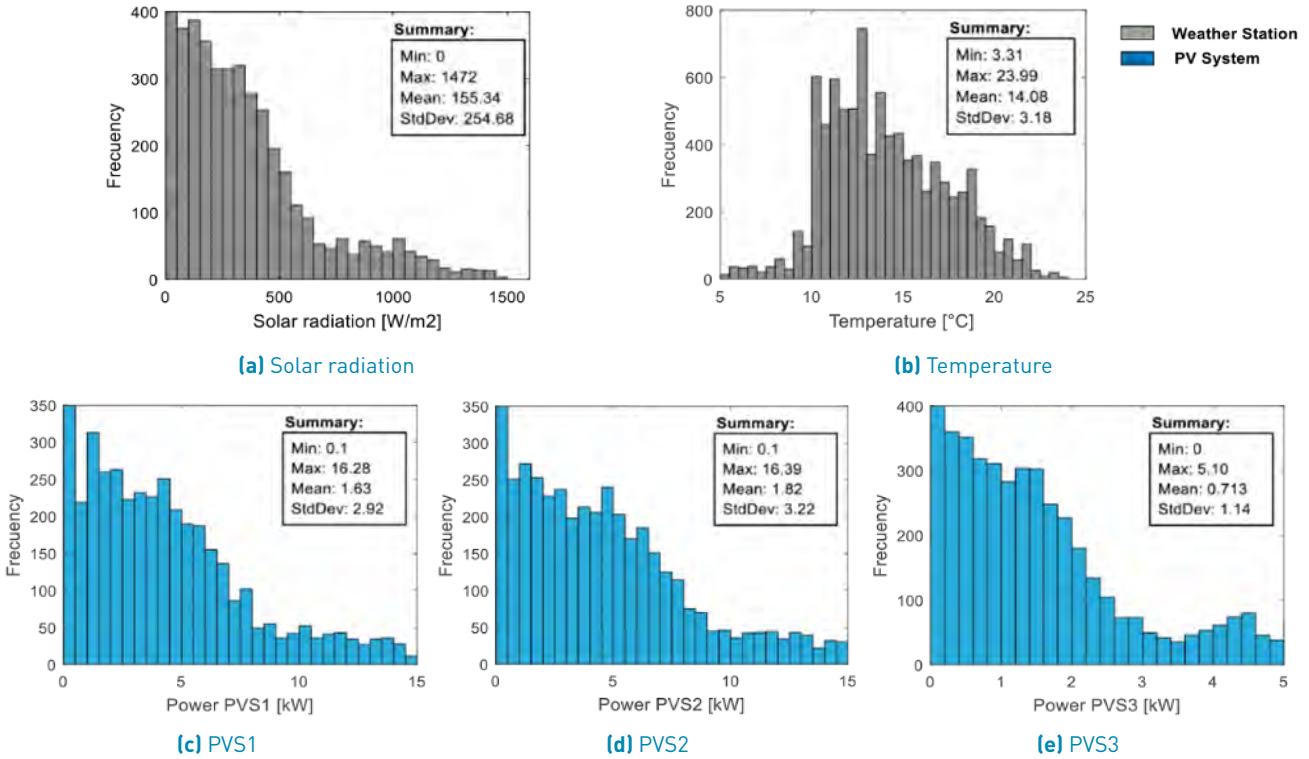
Figure 6 Power PVS1 according to the variables of the weather station

Table 2 Application of models for cloudiness PVS1, PVS2 and PVS3 systems

	Fourier 4th		Gaussian 4th		Sum of Sine 4th	
Description	Goodness of fit: SSE	RMSE:	Goodness of fit: SSE	RMSE:	Goodness of fit: SSE	RMSE:
PVS1	**98.24**	**0.262**	81.21	0.239	96.89	0.261
PVS2	**109.10**	**0.276**	76.28	0.231	113.90	0.283
PVS3	83.04	0.241	58.93	0.203	**97.28**	**0.261**

by means of the adjustment parameters of the sum of squares due to error SSE and the root of the mean square error RMSE. The RMSE parameter for prediction improves considerably when more parameters are used in the machine learning process [4]. This process was carried out using the tool Curve Fitting Toolbox™ software of MATLAB, on the models that best described the curvature based on the functions of *"Fourier"*, *"Gaussian"* and *"Sum of Sine"*, all of the 4th order and considering as variable only the time (hours-minutes).

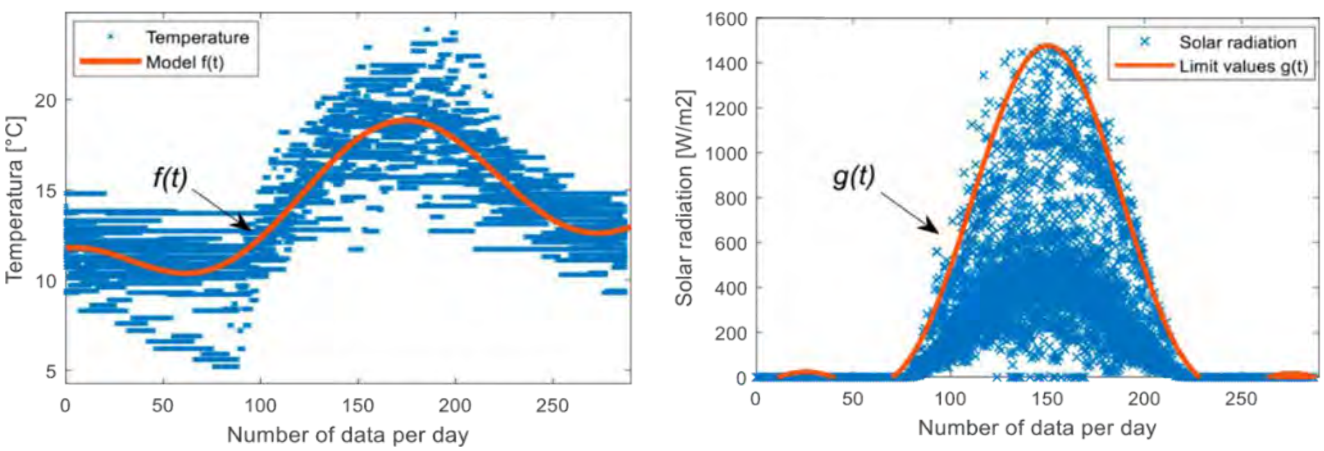
Next, according to the results in Table 2, the equations of the prediction models of completely cleared days corresponding to the PVS1, PVS2 and PVS3 systems according to Equations 3, 4 and 5, respectively, are presented.



(a) Solar radiation **(b)** Temperature

(c) PVS1 **(d)** PVS2 **(e)** PVS3

Figure 7 Histograms of the meteorological variables



(a) **(b)**

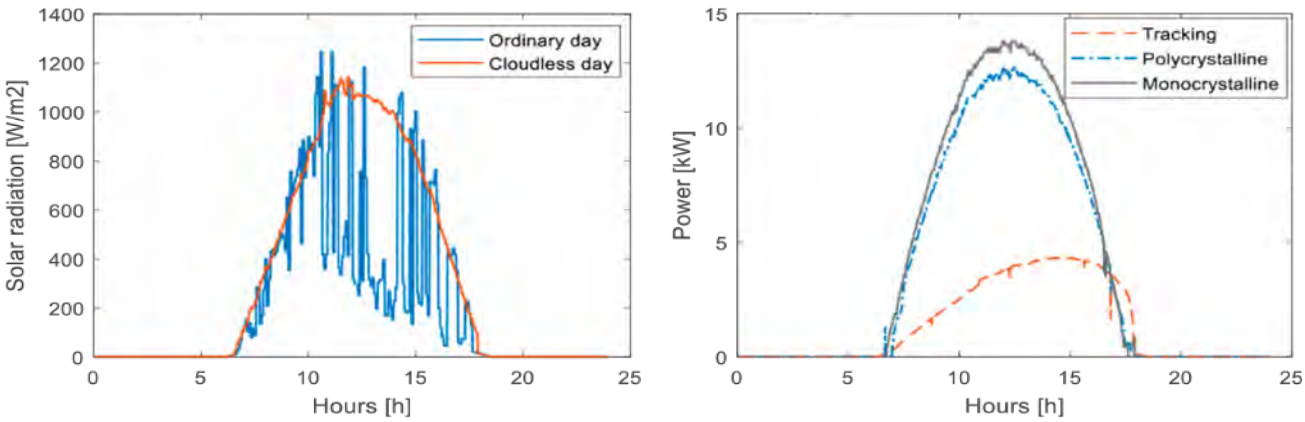
Figure 8 (a) Temperature and (b) Solar radiation, during one-month weather station (Test phase -August 2018)

PVS1 for cloudiness (General model Fourier 4th):

$$PVS1_{\text{Cloudless}}(t) = a0 + a1 \cdot \cos(t \cdot w) + b1 \cdot \sin(t \cdot w) + a2 \cdot \cos(2 \cdot t \cdot w) + b2 \cdot \sin(2 \cdot t \cdot w) + a3 \cdot \cos(3 \cdot t \cdot w) + b3 \cdot \sin(3 \cdot t \cdot w) + a4 \cdot \cos(4 \cdot t \cdot w) + b4 \cdot \sin(4 \cdot t \cdot w) \quad (3)$$

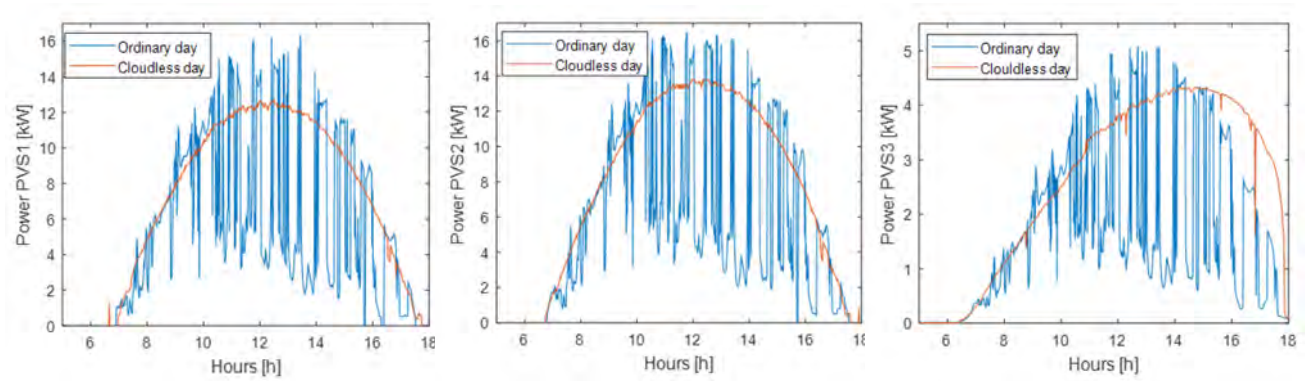
PVS2 for cloudiness (General model Fourier 4th):

$$PVS2_{\text{Cloudless}}(t) = a0 + a1 \cdot \cos(t \cdot w) + b1 \cdot \sin(t \cdot w) + a2 \cdot \cos(2 \cdot t \cdot w) + b2 \cdot \sin(2 \cdot t \cdot w) + a3 \cdot \cos(3 \cdot t \cdot w) + b3 \cdot \sin(3 \cdot t \cdot w) + a4 \cdot \cos(4 \cdot t \cdot w) + b4 \cdot \sin(4 \cdot t \cdot w) \quad (4)$$



(a) **(b)**

Figure 9 (a) Impact of cloudiness on radiation, (b) Photovoltaic power PVS1, PVS2 and PVS3 (clear day)



(a) PVS1 **(b)** PVS2 **(c)** PVS3

Figure 10 Comparison in the PV generation against an ordinary and clear day

PVS3 for cloudiness (General model Sin 4th):

$$PVS3_{\text{Cloudless}}(t) = a1\cdot\sin(b1\cdot t+c1)+a2\cdot\sin(b2\cdot t+c2) + a3\cdot\sin(b3\cdot t+c3)+a4\cdot\sin(b4\cdot t+c4) \quad (5)$$

Next, the function models for cloudiness are described and the values of the coefficients are detailed in Table 3.

A comparison between the photovoltaic generation of the PVS1, PVS2 and PVS3 systems of a completely clear day vs the result of the application of the Equations 3, 4, 5 respectively are presented in Figure 11. This shows that for days of low cloudiness, it is possible to establish a model based on the variable "time" (hours of the day) only with an excellent approximation as a special case.

4.2 Validation phase (25% of data)

Now, considering the effects of cloudiness in this study, it can be seen that it is not possible to define a model only with the variable "time", it is necessary to find patterns with respect to the meteorological variables described above. For this reason, linear regression models are defined below using temperature and solar radiation as input variables of the functions [18].

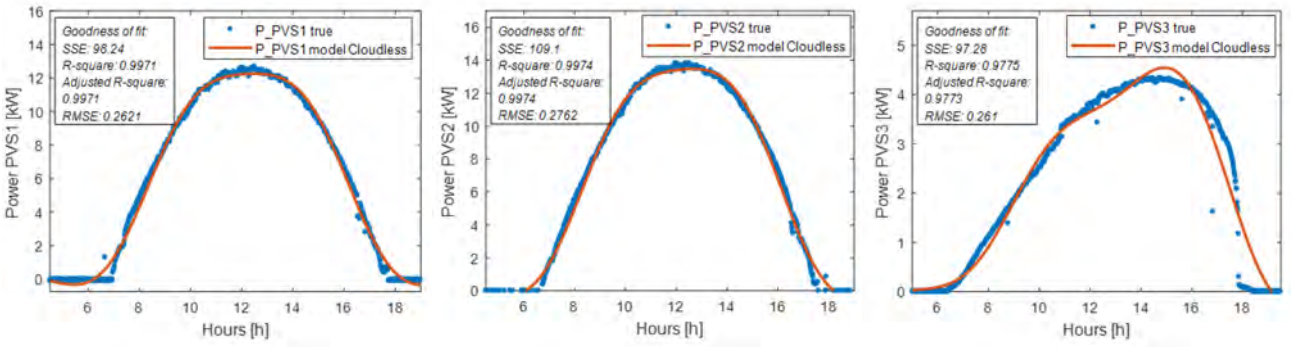
PVS1 Polycrystalline equation (83.27%)

$$P_{PVS1}(rad, temp) = 0.0088*rad+0.0999*temp-1.1393 \quad (6)$$

When applying the *linear regression model* with the variables of *solar radiation* and *temperature*, a correlation coefficient of 0.8327 was established, which demonstrates an acceptable value in the model. In Equation 6, the power generated from the PVS1 is presented as a function of the variables of solar radiation (W/m^2) and temperature (°C). Figure 12a shows the approximation of the real value and the prediction result during the *training phase*, where the values outside the line of the model represent in large part the effects caused by the crossing of clouds in the area. Subsequently, the results are presented in the

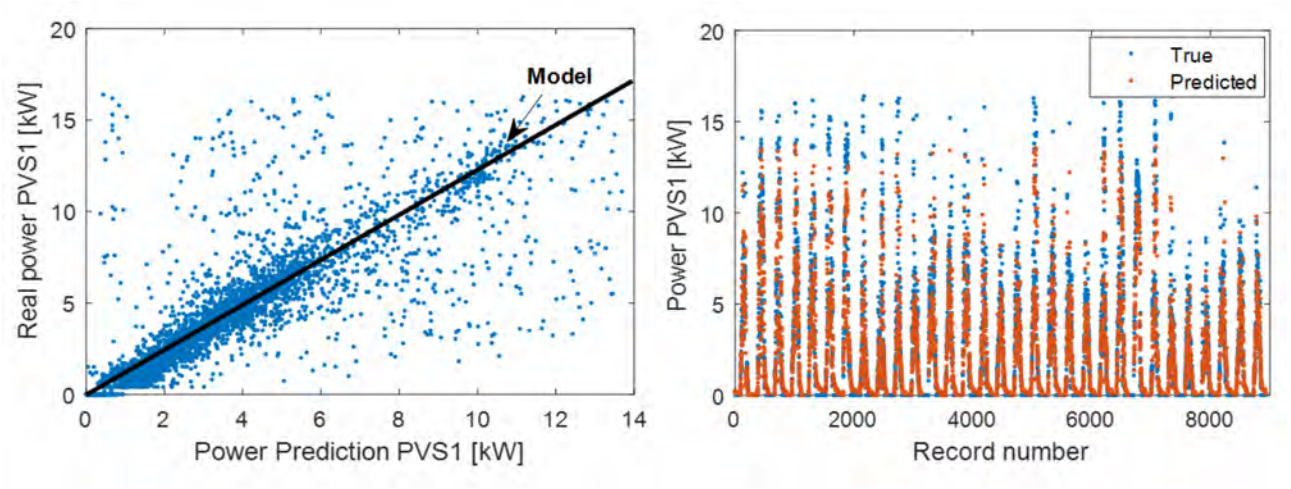
Table 3 Coefficient of the equations of the models

	PVS1 (Fourier 4th)		PVS2 (Fourier 4th)		PVS3 (Sum of Sine 4th)	
	Coefficients (with 95% confidence bounds):		Coefficients (with 95% confidence bounds):		Coefficients (with 95% confidence bounds):	
a0 =	4.27	(4.22, 4.31)	4.682	(4.634, 4.73)	-	-
a1 =	-5.76	(-5.85, -5.66)	-6.396	(-6.489, -6.304)	3.227	(3.195, 3.26)
a2 =	1.256	(1.071, 1.44)	1.443	(1.246, 1.641)	0.3615	(0.339, 0.383)
a3 =	0.0464	(0.001, 0.091)	0.1421	(0.0976, 0.1865)	1.297	(1.274, 1.32)
a4 =	0.255	(0.177, 0.333)	0.2201	(0.1318, 0.3085)	0.307	(0.286, 0.328)
b1 =	-3.192	(-3.42, -2.96)	-3.348	(-3.595, -3.101)	0.166	(0.164, 0.167)
b2 =	2.052	(2.004, 2.099)	2.21	(2.153, 2.267)	0.8523	(0.838, 0.867)
b3 =	0.3879	(0.334, 0.442)	0.4118	(0.3454, 0.4781)	0.4889	(0.483, 0.495)
b4 =	-0.5591	(-0.62, -0.502)	-0.6314	(-0.685, -0.577)	1.12	(1.1, 1.14)
c1 =	-	-	-	-	-0.6329	(-0.65, -0.611)
c2 =	-	-	-	-	-5.416	(-5.58, -5.244)
c3 =	-	-	-	-	-5.124	(-5.21, -5.039)
c4 =	-	-	-	-	-3.941	(-4.187, -3.69)
w =	0.2971	(0.294, 0.3002)	0.2964	(0.2935, 0.2994)	-	-



(a) PVS1 **(b)** PVS2 **(c)** PVS3

Figure 11 Results of the application of the models clear sky



(a) **(b)**

Figure 12 (a) Regression model applied to PVS1, (b) Phase validation PVS1

validation phase (See Figure 12b). It is possible to observe a good approximation to the real values. Under this same analysis, a criterion will be developed for PVS2 and PVS3 in Equations 7 and 8 respectively below:

PVS2 Monocrystalline equation (82.36%)

$$P_{PVS2}\,(rad, temp) = 0.0096{*}rad{+}0.1162{*}temp{-}1.3018 \quad (7)$$

Figures 13a and 13b show the application of the regression model to the PVS2 and the validation of the PVS2, respectively.

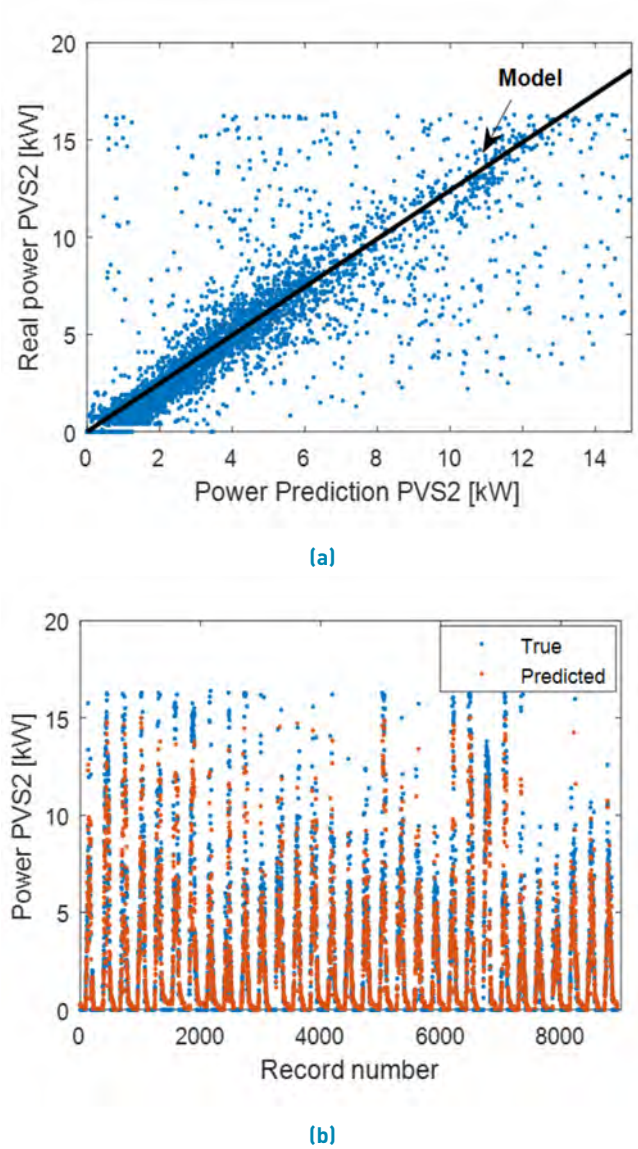


Figure 13 (a) Regression model applied to PVS2, (b) Phase validation PVS2

PVS3 Tracking on an axis equation (85.76%)

$$P_{PVS3}\,(rad, temp) = 0.0037{*}rad{+}0.0225{*}temp{-}0.1745 \quad (8)$$

Similarly, the application of the regression model to the PVS3 and the validation of the PVS3 are presented in Figures 14a and 14b, respectively.

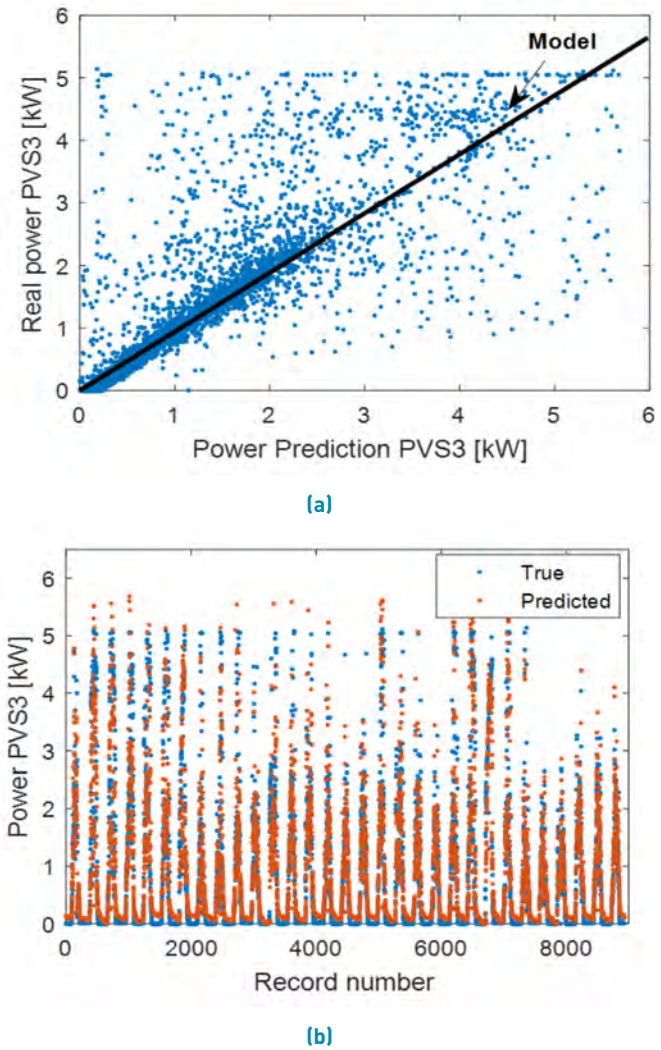


Figure 14 (a) Regression model applied to PVS3, (b) Phase validation PVS3

5. Application to SCADA

The forecast of the power output of the PV systems is necessary for the proper functioning of the electric grid or the optimal management of the energy flows that occur in the PV system [1]. Therefore, it is very important to integrate a monitoring system in real-time, which guarantees the safety in operation and the control of

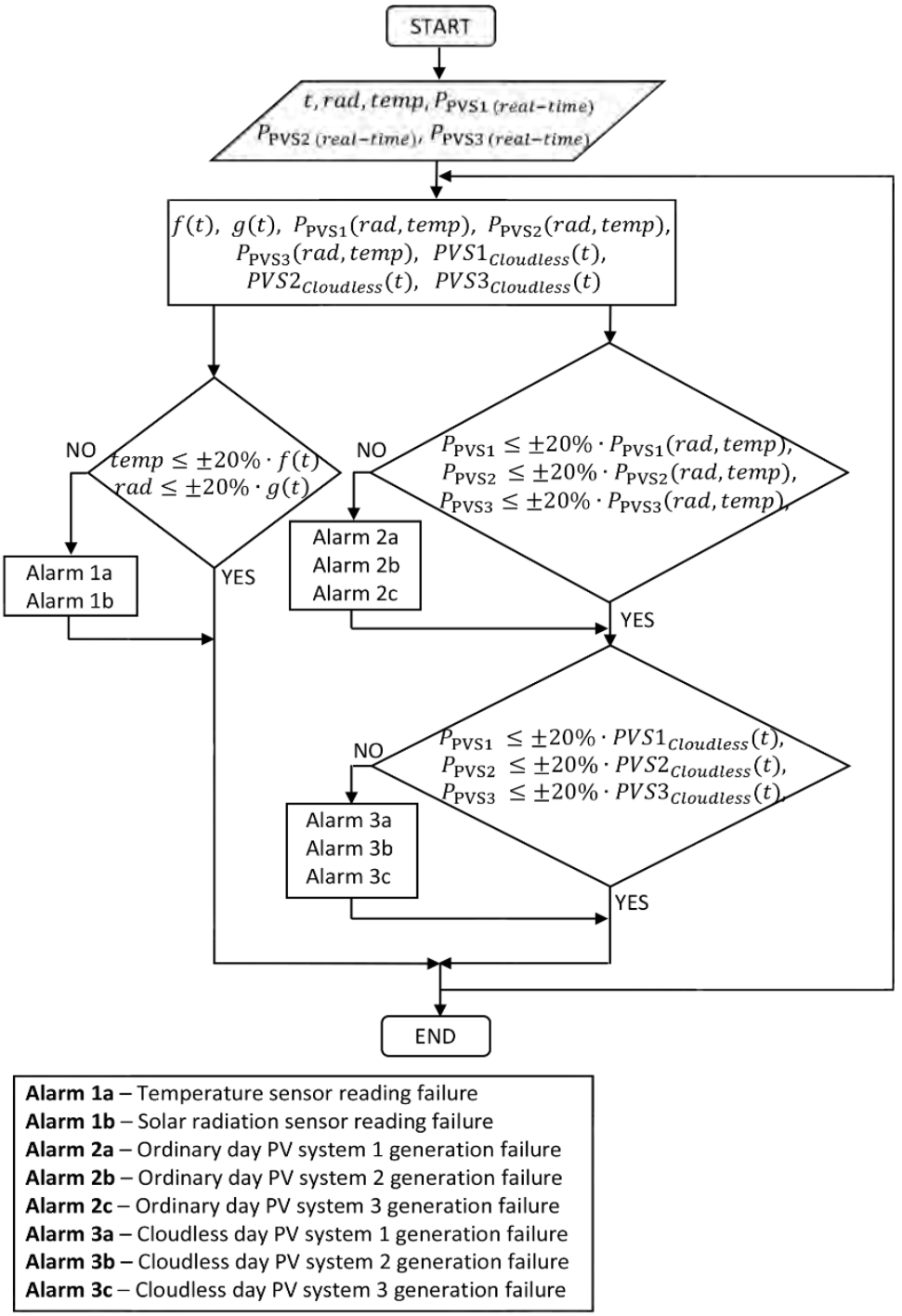


Figure 15 Algorithm for monitoring and fault detection with alarm indicators

the electrical systems supervised by the SCADA. Under this concept, the linear regression models developed in this study will be integrated. It should be noted that in order to develop a monitoring system in real-time, it is necessary to study the corresponding database. Carrying out all the steps to approximate the real value with the prediction, once the model is formulated under all possible considerations and cases, the model can be applied. A power utility operator must ensure a precise balance between electricity production and consumption at any time. This is often very difficult to maintain with a conventional and controllable energy production

system, mainly in small or non-interconnected systems [1]. Therefore, it is necessary to cover the term of real-time, although to be able to apply a predictive model in real-time, a previous analysis of databases is required. Later its application is possible including the equations of the resulting models in the SCADA system, entering the measurements in real-time, that is, the SCADA carries out the measurement, discretization and calculation in the corresponding formula of the model.

Figure 15 shows the implementation of the study algorithm, with the integration of Equations 1 to 8, corresponding to the temperature prediction. $f(t)$ and the limits of solar radiation $g(t)$. Photovoltaic production on cloudless days: $PVS1_{Cloudless}(t)$, $PVS2_{Cloudless}(t)$, and $PVS3_{Cloudless}(t)$.

Photovoltaic production on ordinary days: $P_{PVS1}(rad, temp)$, $P_{PVS2}(rad, temp)$, and $P_{PVS3}(rad, temp)$. The input parameters corresponding to time t solar radiation rad, temperature $temp$, and the values of the powers PVS1, PVS2, PVS3. These parameters are acquired in real-time from the SCADA system. The values are then processed using the models of the equations and return the calculated prediction value. When exceeding the established values, the alarms corresponding to each case are activated, as indicated in the almorithm below.

The evaluation of the models will be obtained by means of a percentage indicator "meter" which is assigned to compare between the real value of power and the value of the prediction, in such a way that, if its value is lower than 20% (assigned value according to the results case study) implies that there is a difference of 20% in the photovoltaic production of the real value and the prediction, which must be considered by the operator of the electrical system. Under these circumstances, an alarm indicator has also been implemented, in the case of similar situations. On the other hand, the temperature model $f(t)$, it can be used to verify the measurement status of that variable. Similarly, the solar radiation limit model $g(t)$ allows checking the status of the measurement sensor, that is, verifying that the measurement value does not exceed the value of the resulting model at the set time. For example, in Figure 16 shows the monitoring algorithm for temperature measurement with the maximum and minimum values established. In this case, all values are within the established range, whereby the state of *alarm 1a* is low.

Similarly, the solar radiation limit model $g(t)$, it allows checking the status of the measurement sensor, that is, verifying that the measurement value does not exceed the value of the resulting model at the set time. If the measured values exceed the set values, it returns a state of *alarm 1b* at a high value. Otherwise, all values within the model equation return an *alarm 1b* at low value (See Figure 17).

5.1 Method for fault detection

A method of detection and diagnosis of failures proposed by Garoudja [12] consists of four main stages: **(i)** *extraction of parameters from the PV module*, **(ii)** *validation of the model*, **(iii)** *elaboration of relevant data sets and finally* **(iv)** *fault detection and diagnosis*.

In this study, the panel configuration corresponds to 15x4 (parallel series) of the PVS1 and PVS2 systems. So if a fault occurs in one of the branches, the fuse protection acts and the total power value of the PVS1 system will be reflected at the measurement point (See Figure 18).
In this way, when analyzing the photovoltaic production of two normal days (without failure, with failure). It is possible to observe the behavior of both systems compared to the prediction model (see Figure 19). For the first case, the values are within the established limit of 20%. For example: the point measurement (186, 7.18) and (187, 5.57) would indicate a variation of 7.18-5.57 = 1.61kW that corresponds to the 10.73% variation and for the second case exceeds the limit by 12.85-6.96 = 5.89kW (39.27%).

When the time of the branch failure in the PVS1 system is specifically analyzed, as shown in Figure 20 of record 1,140 approximately. The reference of PV production is higher by a large percentage and during a wide range of records. Therefore, *alarm 2a* is activated at high value, otherwise if the power value is within the percentage of 20% *alarm 2a* returns a low value.

The application of the equations has been implemented in the SCADA system (See Figure 21). By means of a modification in the LabVIEW Sotfware, the calculation in real-time has been included that allows the operator to obtain a better reference of the PV production and the detection of failures for optimal operation.

6. Results and discussion

The application of the equations obtained in this study have allowed obtaining reference values of photovoltaic power of PVS1, PVS2 and PVS3. Thus, they can vary as a system for monitoring and detecting faults when comparing the photovoltaic power generated in real time with the values calculated based on the meteorological variables of radiation and temperature.

The equations of the temperature measurement model and radiation measurement limit also allow us to

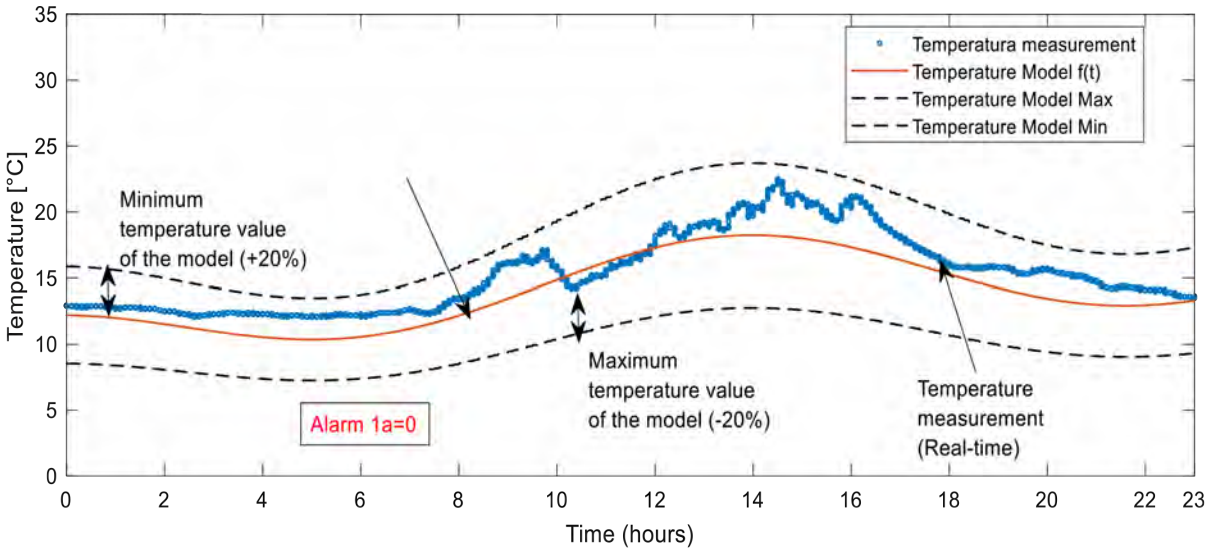


Figure 16 Monitoring algorithm for temperature measurement

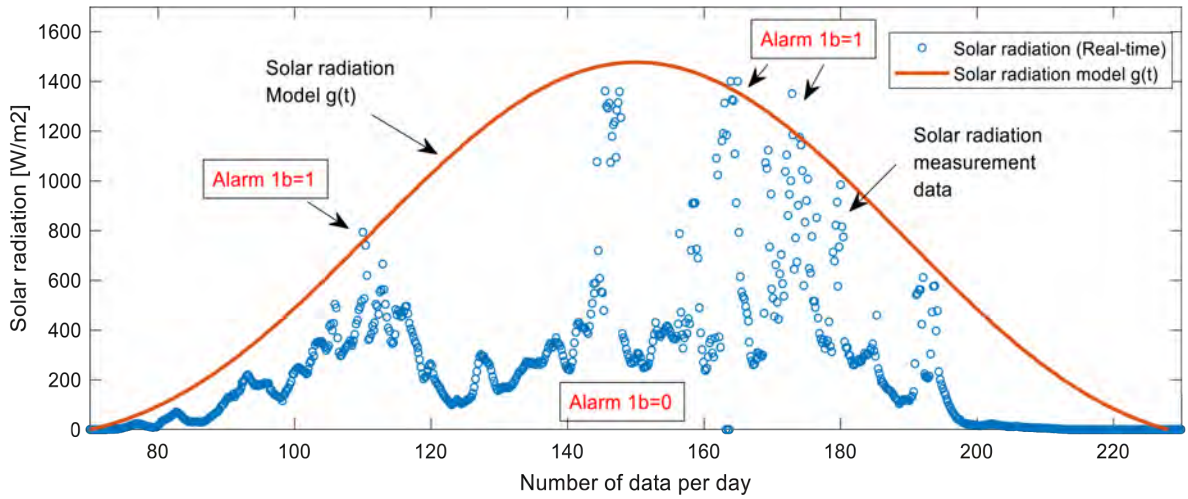


Figure 17 Monitoring algorithm for measuring solar radiation

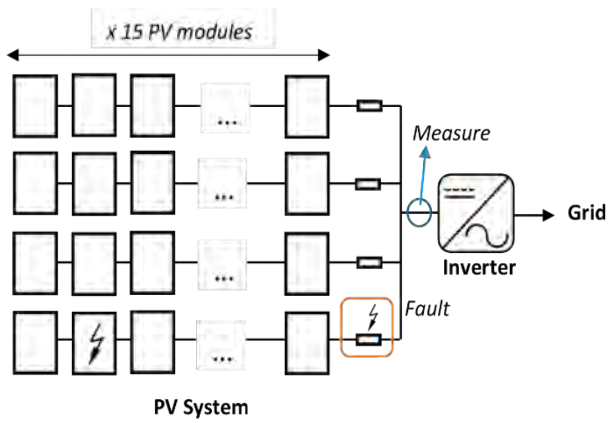


Figure 18 Fault simulation in a branch of the PVS1 system

monitor the behavior of the sensors and rule out possible measurement errors as a function of time.

A 75% training of random data has allowed us to approach the winter and summer months within the most adaptable values under those considerations.

In comparison of the PVS1 and PVS2 systems, the coefficients of the equations have presented some similarity, due to the same installed power value of 15 kW. Although a small variation in the parameter values has been determined, due to the characteristics of the monocrystalline and polycrystalline cell type. In this way, through data training they have been able to adjust to the best conditions.

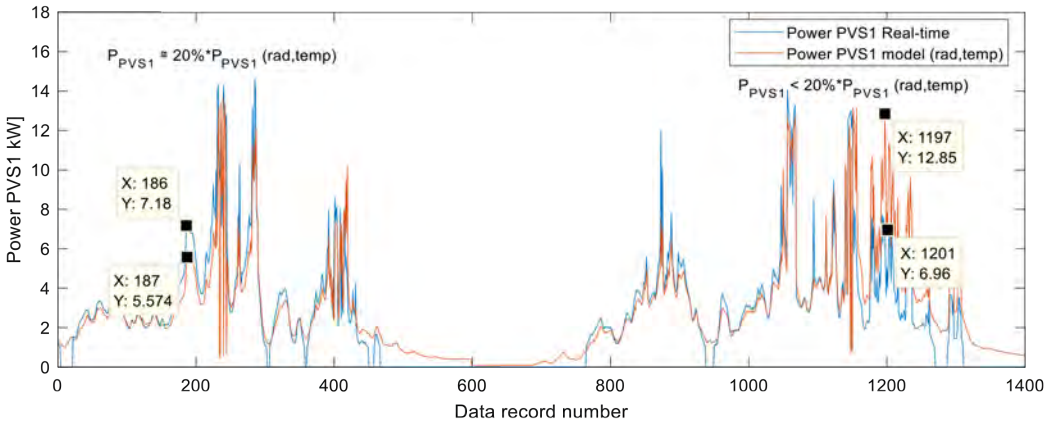


Figure 19 Algorithm application for PVS1 system (two days of registration)

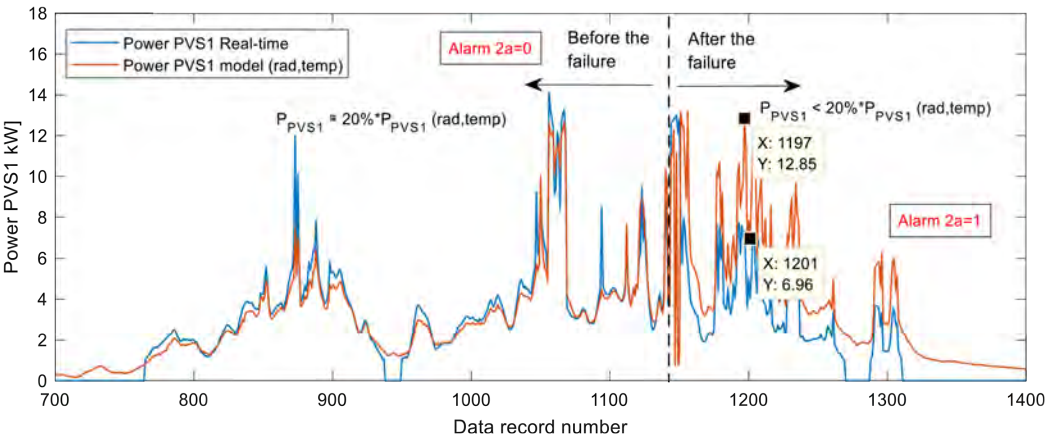


Figure 20 Algorithm application before and after a failure (12 hours of registration)

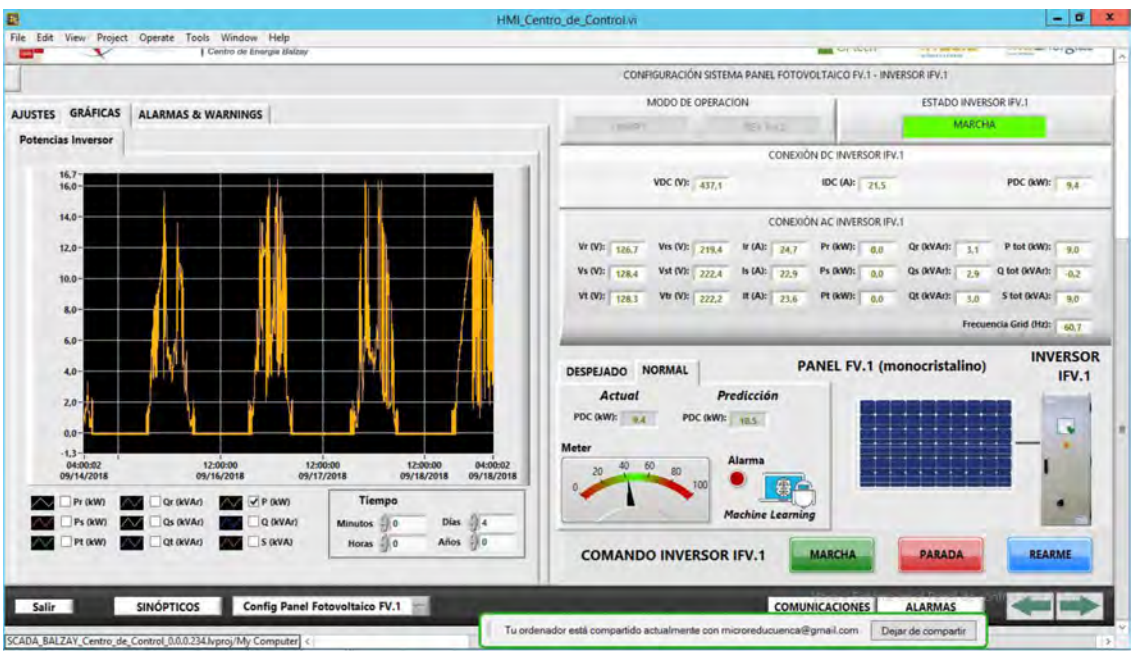


Figure 21 Example of machine learning application in SCADA

Unlike the PVS3 system considered its monitoring characteristic and low value of installed power 5kW, the parameter values have varied significantly. This is the main reason why the parameters have been adjusted differently.

7. Conclusions

As industrial development increases, automation and processes generate more data and information and require analysis, interpretation and communication. Therefore, this study has demonstrated the application of machine learning techniques in the analysis of real data and the development of predictive models.

As a result of the study, it has been shown that it is possible to predict the photovoltaic power of the three systems studied by means of regression models established with an excellent approximation.

The importance of monitoring the variables by means of the measurement sensors has allowed us to obtain adequate control of the photovoltaic system.

The application of a fault detection strategy has been demonstrated through predictive model techniques in PV systems, iIn such a way that it allows us to monitor PV systems by comparing the PV power generated in real time with the calculated values based on the meteorological variables of radiation and temperature.

Correlation coefficients of 83.27%, 82.36% and 85.76% have been obtained in the model results for the PVS1, PVS2 and PVS3 systems, respectively. With which a range of 20% has been established that allows the comparison of the calculation values with the values measured in real time.

The equations of the temperature measurement model and radiation measurement limit also allow us to monitor the behavior of the sensors and rule out possible measurement errors as a function of time. In this way, an additional means of monitoring and control of the equations using these parameters is obtained.

The implementation of the prediction models of the PV systems in the SCADA allows the monitoring of the electrical system operator in an optimal way.

Finally, the importance of the application of machine learning techniques and its wide variety in development in the field of energy management and its importance in smart grids was demonstrated.

8. Declaration of competing interest

We declare that we have no significant competing interests including financial or non-financial, professional, or personal interests interfering with the full and objective presentation of the work described in this manuscript.

9. Acknowledgements

The authors thank the CYTED Thematic Network "CIUDADES INTELIGENTES TOTALMENTE INTEGRALES, EFICIENTES Y SOSTENIBLES (CITIES)" no 518RT0558.

References

[1] C. Voyant and *et al*, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, May 2017. [Online]. Available: https://doi.org/10.1016/j.renene.2016.12.095

[2] S. Theocharides, G. Makrides, G. E. Georghiou, and A. Kyprianou, "Machine learning algorithms for photovoltaic system power output prediction," in *2018 IEEE International Energy Conference (ENERGYCON)*, Limassol, Cyprus, 2018.

[3] C. Kurien and A. K. Srivastava, "Scope of artificial intelligence techniques for exhaust emission prediction of CI engines and renewable energy applications„" *International Journal of Engineering Research in Computer Science and Engineering*, vol. 5, no. 2, pp. 456–461, Feb 2018.

[4] S. Preda, S. Vasilica, A. Bâra, and A. Belciu, "PV forecasting using support vector machine learning in a big data analytics context," *Symmetry*, vol. 10, no. 12, December 2018. [Online]. Available: https://doi.org/10.3390/sym10120748

[5] T. T. Teo, T. Logenthiran, W. L. Woo, and K. Abidi, "Forecasting of photovoltaic power using regularized ensemble extreme learning machine," in *2016 IEEE Region 10 Conference (TENCON)*, Singapore, Singapore, 2017, pp. 455–458.

[6] T. Huuhtanen and A. Jung, "Predictive maintenance of photovoltaic panels via deep learning," in *2018 IEEE Data Science Workshop (DSW)*, Lausanne, Switzerland, 2018.

[7] H. T. Pedro, C. F. Coimbra, M. David, and P. Lauret, "Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts," *Symmetry*, vol. 123, August 2018. [Online]. Available: https://doi.org/10.1016/j.renene.2018.02.006

[8] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. M. Shah, "Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques," *IET Renewable Power Generation*, vol. 13, no. 7, May 9 2019. [Online]. Available: https://doi.org/10.1049/iet-rpg.2018.5649

[9] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid – A review," *Renewable and Sustainable Energy Reviews*, vol. 79, November 2017. [Online]. Available: https://doi.org/10.1016/j.rser.2017.05.134

[10] S. Siniscalchi and F. D. Bianchi and M. De Prada and C. Ocampo, "A wind farm control strategy for power reserve maximization," *Renew. Energy*, vol. 131, February 2019. [Online]. Available: https://doi.org/10.1016/j.renene.2018.06.112

[11] F. Han and *et al*, "An intelligent fault diagnosis method for pv arrays based on an improved rotation forest algorithm," *Energy Procedia*, vol. 158, February 2019. [Online]. Available: https://doi.org/10.1016/j.egypro.2019.01.498

[12] E. Garoudja, A. Chouder, K. Kara, and S. Silvestre, "An enhanced machine learning based approach for failures detection and diagnosis of PVsystems," *Energy Procedia*, vol. 158, February 2019. [Online]. Available: https://doi.org/10.1016/j.egypro.2019.01.498

[13] D. Benavides, P. Arévalo, L. G. Gonzalez, L. Hernández, and F. Jurado, "Machine learning data applied to monitoring PV systems: A case study*," in *Ibero-American Congress of Smart Cities (ICSC-CITIES 2019)*, Soria, Spain, 2019, pp. 456–470.

[14] S. K. Jha, J. Bilalovic, A. Jha, N. Patel, and H. Zhang, "Renewable energy: Present research and future scope of Artificial Intelligence," *Renewable and Sustainable Energy Reviews*, vol. 77, September 2017. [Online]. Available: https://doi.org/10.1016/j.rser.2017.04.018

[15] J. Li, J. K. Ward, J. Tong, L. Collins, and G. Platt, "Machine learning for solar irradiance forecasting of photovoltaic system," *Renew. Energy*, vol. 90, May 2016. [Online]. Available: https://doi.org/10.1016/j.renene.2015.12.069

[16] M. K. Behera, I. Majumder, and N. Nayak, "Solar photovoltaic power forecasting using optimized modified extreme learning machine technique," *Eng. Sci. Technol. an Int. J.*, vol. 21, no. 3, June 2018. [Online]. Available: https://doi.org/10.1016/j.jestch.2018.04.013

[17] J. L. Espinoza, L. G. González, and R. Sempértegui, "Micro grid laboratory as a tool for research on non-conventional energy sources in Ecuador," in *2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, Ixtapa, Mexico, 2018.

[18] D. J. Benavides, F. Jurado, and L. G. González, "Data analysis and tools applied to modeling and simulation of a PV system in Ecuador," *Enfoque UTE*, vol. 9, no. 4, 2018. [Online]. Available: https://doi.org/10.29019/enfoqueute.v9n4.389

[19] MathWorks. Machine Learning with MATLAB. The MathWorks, Inc. Accessed Nov. 6, 2019. [Online]. Available: https://bit.ly/3hXzJzs