Photovoltaic system monitoring and fault detection using peer systems

Alcañiz, Alba; Nikam, Maitheli M.; Snow, Yitzi; Isabella, Olindo; Ziar, Hesan

DOI
10.1002/pip.3558

Publication date
2022

Document Version
Final published version

Published in
Progress in Photovoltaics: research and applications

Received: 28 October 2021 | Revised: 4 March 2022 | Accepted: 7 March 2022

DOI: 10.1002/pip.3558

RESEARCH ARTICLE

PROGRESS IN PHOTOVOLTAICS WILEY

Photovoltaic system monitoring and fault detection using peer systems

Alba Alcañiz[1] | Maitheli M. Nikam[1] | Yitzi Snow[2] | Olindo Isabella[1] | Hesan Ziar[1]

[1]Photovoltaics Materials and Devices Group, Delft University of Technology, Delft, Netherlands

[2]Solar Monkey, The Hague, Netherlands

Correspondence
Hesan Ziar, Photovoltaics Materials and Devices Group, Delft University of Technology, Delft, Netherlands.
Email: h.ziar@tudelft.nl

Funding information
Horizon 2020 Framework Programme; Trust PV, Grant/Award Number: 952957

Abstract

Monitoring residential scale photovoltaic (PV) systems is important for maximizing the energy yield and detecting malfunctions. Analytical-based approaches are not reliable in these systems because of the lack of on-site measurements and detailed PV system specifications. In this paper, a collaborative approach is proposed which does not depend on weather data but on similar PV systems. Based on the so-called performance-to-peer approach, the aim of this work is to improve this baseline model by adding PV systems characteristics and by optimizing with machine learning techniques. The methodology has been tested in a fleet of more than 12,000 PV systems located in the Netherlands with up to 7 years of data per system. The proposed model achieves an average R^2 of 94.1% and a NRMSE of 0.05, outperforming in terms of R^2 the baseline model by 1.4 points, and the analytical approach by 3.8. The data requirements of this model are not high: With 1,700 years of PV system data with daily resolution, the maximum performance can be achieved as long as a minimum of 6 months of data per system and 100 PV systems are considered. The application of this model for fault detection and categorization has also been shown. The proposed approach has shown its strengths with respect to other methods through its ability of distinguishing between system mismatch and actual fault and of adapting to new situations via retraining.

KEYWORDS
fault detection, genetic algorithm, monitoring, peer-to-peer, photovoltaics

1 | INTRODUCTION

Motivated by the rising awareness on climate change and higher competitiveness, in recent years, photovoltaic (PV) sources have seen an increase in the share of electricity generation all around the world.[1] Solar energy is an abundant, increasingly affordable, scalable, and clean source of energy.[2] It is possible to generate electricity from solar energy using photovoltaic panels with no direct emissions; that is, no greenhouse gasses (GHG) are emitted during operation, only in the manufacturing and installation processes. With non-renewable energy sources set to decline and an exponential drop in prices of PV modules, solar energy has the potential to supply electricity that is environmentally as well as economically attractive.[3]

One of the main drawbacks of solar energy is the uncertainty and intermittent generation due to not only diurnal and seasonal Sun variations but also due to local phenomena such as clouds. This

interrupted generation may create instabilities in the power grid such as voltage fluctuations.[4] One of the ways to tackle this issue is by estimating the energy generation of these PV systems. Monitoring PV systems can also be beneficial for fault detection in order to rapidly spot any malfunction in the system.

The conventional approach employed in monitoring is to use design information along with accurate weather data.[5–7] This is suitable for large or commercial solar farms where expensive sensors can be installed on site. However, for residential PV systems, extrapolated weather data have to be used. This rarely represents the local weather as the weather stations are often far away from the actual site of PV system installation. Thus, local phenomena like a cloud passing over the area cannot be accounted for in the calculation of expected energy generation. This affects the accuracy of monitoring.

To overcome the problem of weather data, an alternative approach was developed where neighboring or peer systems are used to monitor each other. A peer system is essentially a system similar to the monitored PV system. It could be physically nearby or have similar design parameters. This way, if a cloud passes over one PV system, it will also pass over a geographically close peer system, and thus, their yields can be compared to detect any fault. Since the approach is based on historical energy yields along with the peer predictions, it will also overcome any monitoring error due to incorrect design information.

Several works can be found where the information from neighboring PV systems was employed to improve the predictions in the residential scale. The studies[8,9] employed neighboring systems to determine the irradiance variations due to cloud motion and then forecasted energy production using analytical models. Golnas et al[10] predicted the output of a PV system from a regional fleet by using data from other systems in the fleet based on historical performance correlation between the systems. Tsafarakis et al[11] developed a method for fault detection considering that the power produced by a PV system is linearly related to the power produced by neighboring PV systems. Next, Popovic and Radovanovic[12] presented a methodology for inter-system comparison between correlated PV systems to estimate the operation status of individual panel in urban surroundings.

The most relevant approach to use peer systems to monitor the PV power was presented by Leloux et al.[13,14] They defined a novel performance indicator called Performance to Peer (P2P) which was computed by comparing the energy production of several neighboring PV systems.[13,14] They proposed it as an alternative to the Performance Ratio (PR), commonly employed for monitoring and automatic fault detection. They showed that P2P was more stable than PR when monitoring 6,000 PV installations across Europe with energy output data for approximately 7 years and a temporal resolution of 10 min. Stability here was interpreted by the authors as the ability to easily distinguish between the absence and presence of a fault.

The objective of this research is to extend the work performed in Leloux et al.[14] The database will be expanded to more than 12,000 PV residential systems located in the Netherlands, with diverse system characteristics and installation age. System information will be employed beyond just geographical closeness. In contrast to some research papers,[8,9,11,12,15] weather data are not employed for finding similar systems in order to avoid their inaccuracies. Furthermore, in line with the recommendations of Leloux et al,[13] machine learning algorithms are used to optimize the model and increase its accuracy. The application of the developed model for fault detection and categorization will also be demonstrated.

The structure of the paper is as follows. Section 2 presents the employed models. The data used are briefly described in Section 3. Main results are displayed in Section 4 where the superior performance of the proposed algorithm is shown, together with its limits. The main application of the proposed model is demonstrated in Section 5 by presenting the fault detection approach and showing the fault categorization performed. Finally, Section 6 presents the main conclusions.

2 | MODELS

This section explains the main models employed in this work. Subsection 2.1 describes the P2P model developed in Leloux et al.[13] With this approach as base, several improvements are made yielding the proposed model, in subsection 2.2. Optimizations are required for the latter approach, which are explained in subsection 2.3. Finally, subsection 2.4 presents a series of models that will be used for comparison in later sections.

2.1 | Performance-to-peer model

In P2P model, the performance of one system is compared with its peers in order to monitor the former system.[13] The PV system to be monitored is defined as the focus system, whereas all other available PV systems are referred as peer systems. The model is divided into two steps: identifying the good peers among all available PV systems and calculating the expected yields by using the data of the good peers.

The energy yield data are the only characteristic employed to find the good peers for the focus system. The yields are normalized with respect to their total capacity resulting in the Capacity Utilization Factor (*CUF*) as shown in Equation 1. With the daily yields now normalized to the same scale, the Capacity Utilization Ratio (*CUR*) is calculated for each focus-to-peer system pair, according to Equation 2. Next, the weighing factor is calculated by taking the inverse of the Median Absolute Deviation (*MAD*) of *CUR* and raised to fourth power (Equation 3). The exponent was determined through a sensitivity analysis performed by Leloux et al.[13] This weighing factor is used to determine whether the peer system is a good peer system or not. The higher the weighing factor, the better the peer for the chosen focus system.

$$CUF = \frac{E_{PV}}{P \cdot T}, \tag{1}$$

where E_{PV} is the energy output of the PV system, P is the rated power of the PV system, and T is the time interval of energy output measurement.

$$CUR = \frac{CUF_{focus}}{CUF_{peer}}, \tag{2}$$

where CUF_{focus} is the CUF of the focus PV system and CUF_{peer} is the CUF of the peer PV system.

$$weighing\ factor = \frac{1}{[MAD(CUR)]^4} \tag{3}$$

The next step is to use the data of good peer systems to calculate the expected yields. Once top 10 peers are chosen as good peers, a weighted median of the peer systems' CUF with respect to their weighing factors is calculated to give the reference CUF_{ref}. CUF_{ref} can be recognized as the normalized expected yields for the focus system. Hence, the true expected yields can be estimated by reverse calculation of CUF. Additionally, the model further calculates P2P for each day by taking a ratio of the focus system CUF_{focus} to reference CUF_{ref}. This new metric is proposed instead of the conventional Performance Ratio (PR) for the monitoring of focus system.[13]

2.2 | Proposed model

The proposed model is based on the above Performance-to-Peer model. The main difference between the two is the use of a higher amount of PV system data. Moreover, system design information, daily yields, and system location are used to find good peers for the chosen focus system, not only system yields. Just like P2P, the proposed model is divided into two steps: distance calculation to identify the good peers and expected yields calculation using the data of the good peers.

2.2.1 | Distance calculation

This first part of the proposed model calculates the similarity between two PV systems. This similarity is computed by making use of several *distances*, understood as the more alike two systems are, the lower is the distance between them. Therefore, here, the distance should not be necessarily interpreted as geographical distance between the two. Three distances are computed (one for each characteristic), normalized, and later combined. Normal standardization was applied to all distances to ensure a fair comparison.

To begin with, the *feature distance* (d_{feat}) considers the system design information. This includes the number of panels n, panel inclination θ, and panel orientation ϕ. These variables were chosen after a feature correlation analysis and experience-based selection. The model calculates the Euclidean distance for each focus-to-peer system pair using Equation 4. When computing the Euclidean distance, each of these attributes has a different weight w, representing the importance that the features have individually.

$$d_{feat} = \sqrt{w_n(n_f - n_p)^2 + w_\theta(\theta_f - \theta_p)^2 + w_\phi(\phi_f - \phi_p)^2}, \tag{4}$$

where sub-index f represents the focus PV system and sub-index p represents the peer PV system.

Next, the *yield distance* (d_{yield}) is calculated in a similar fashion as in the P2P model. The daily yields are normalized with respect to the actual energy yield in the first year of installation of the system (for systems older than 1 year) or with respect to the estimated energy yield for the typical meteorological year for the system (for systems less than a year old) in order to obtain the CUF. For each focus-to-peer system pair, the CUR is calculated for each day according to Equation 2. Since a single value is needed for the distance, median absolute deviation is calculated from all the daily CUR values.

The third and final distance is the *geographical distance* (d_{geo}). It is the physical distance between the latitude and longitude coordinates of each focus-to-peer system pair. Due to the curvature of Earth, haversine distance is used.[16]

Now that the three distances have been computed for each focus-to-peer system pair, these are combined into a *total distance* d_{tot} using the weighted sum of the distances according to Equation 5. A weighted sum is used to consider a different influence for each distance when finding good peers. Once the total distance is known for each focus-to-peer pair, the pairs with the lowest distances, that is, the peer systems most similar to focus system, are chosen as good peers.

$$d_{tot} = w_{feat} \cdot d_{feat} + w_{yield} \cdot d_{yield} + w_{geo} \cdot d_{geo} \tag{5}$$

2.2.2 | Expected yield calculation

The expected yield of the focus system is calculated using the daily yields of only the selected peer systems. Each of these peer systems has a different level of influence on the expected yields. The higher the similarity of the peer system to the focus system, the higher its influence. Since the total distance was a measure of the similarity, the weighing factor λ_p per peer system is evaluated as the reciprocal of the corresponding total distance, Equation 6.

These weighing factors along with the CUF of good peer systems are used as a weighted median to determine CUF_{ref}, as shown in Equation 7.

For N_p distinct ordered number of selected peers with $CUF_1, CUF_2, \ldots, CUF_{N_p}$ and weights $\lambda_1, \lambda_2, \ldots, \lambda_{N_p}$ such that

$$\lambda_p = \frac{1}{d_{tot,p}} \text{ restricted to } \sum_{p=1}^{N_p} \lambda_p = 1, \tag{6}$$

CUF_{ref} is the element CUF_k satisfying

$$\sum_{p=1}^{k-1}\lambda_p \le 1/2 \text{ and } \sum_{p=k+1}^{N_p}\lambda_p \le 1/2. \quad (7)$$

The reason for using a weighted median over weighted average is that a median is not influenced by an abnormal extreme value that may be present for one of the good peers. This abnormal extreme value could be a fault in a peer system, so using the median essentially helps to avoid that fault from being transferred in the predictions of the focus system. The expected yield is estimated from CUF_{ref} by reverse calculation of CUF.

Figure 1 shows the procedure to obtain the expected yield of a focus system with the proposed model. For each peer system, all the distances are computed, as just explained. From the total distance, the weights can be determined; hence, the peers can be selected. One can see some variables outside the boxes. These are the variables that need to be determined via optimization, as will be explained in the next subsection.

2.3 | Model optimization

While explaining the proposed model in the previous section, several variables came across that need to be determined in order to ensure optimum performance of the model. The unknowns are the weights w of each distance and of the individual PV system features, and the number of peers N_p that need to be chosen for estimating the expected yields. The value of these variables was determined by training the model using an optimization algorithm.

Given the complexity of the problem at hand, local optimization algorithms, such as gradient descent[17] and Nelder-Mead methods,[18] were stuck in local minima and unable to find the global optimum. Global optimization algorithms were needed. Among the tested global optimization algorithms, evolutionary algorithms gave the best results. Inside this group, particle swarm optimization (PSO) and genetic algorithm (GA) were the most promising ones. Both algorithms are relatively simple and easy to modify, and their performance depends on the problem at hand because they traverse the candidate space rather differently.[19]

In order to select between the two, a literature search was done to find the best algorithm for problems using peer-to-peer strategies. We came across the work of, who used PSO and GA for neighbor-selection in peer-to-peer networks and found out that GA obtained better results.[20] Similarly, Rehman et al[21] employed GA and PSO to optimize the peer-to-peer energy transactions in a decentralized energy trading market and saw that GA results outperformed the PSO ones. GA was also combined with PSO for device-to-device (D2D) communication in advanced communication networks.[22] The authors claimed that PSO alone can be trapped in local optima due to premature convergence; hence, a hybrid PSO-GA algorithm was proposed to find the optimum allocation of the D2D communication network's resources and avoid interference with the primary cellular network. Given the similarities of these problems with the one at hand, GA was finally chosen as the optimization algorithm.

Genetic algorithm is a global search engine inspired by natural evolution.[23] The objective of GA is to obtain the optimum value of a fitness or objective function, which represents the performance of the problem. The higher the fitness value, the better the system's performance. In our case, the objective function is to minimize the average Mean Absolute Error (MAE) between the actual yields and expected yields of the focus systems. The optimum solution is found via a

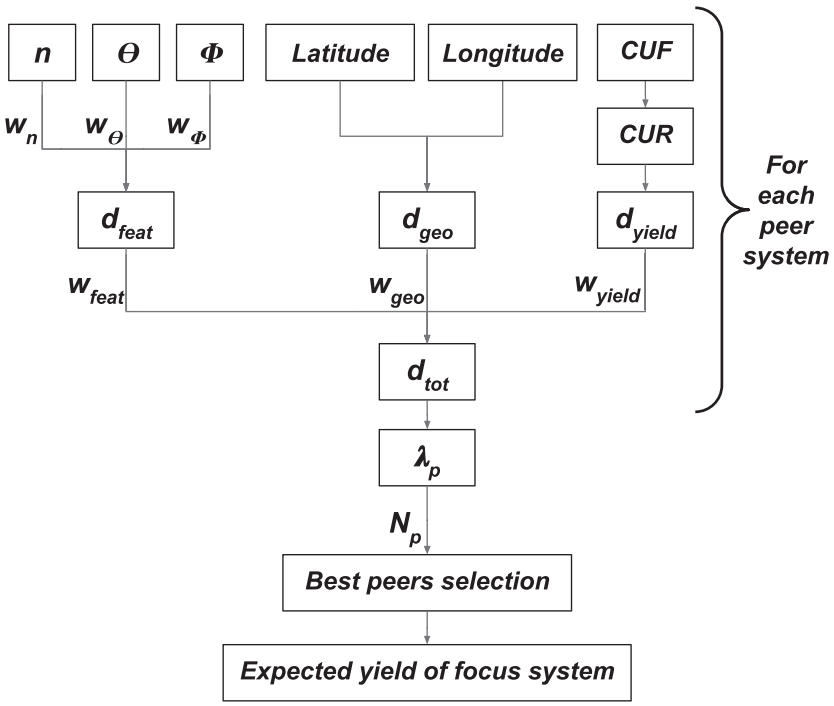

FIGURE 1 General outline of the proposed model for one focus system

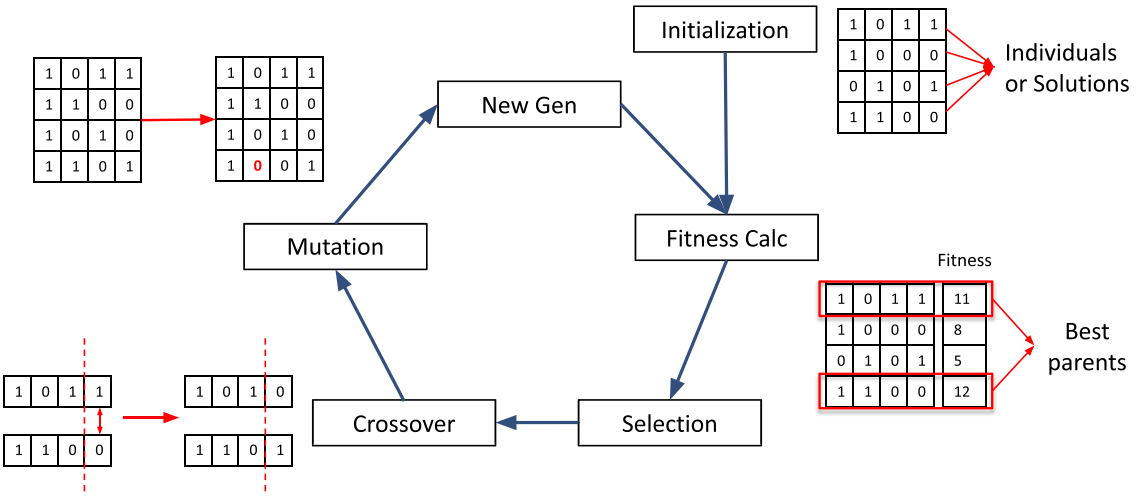

FIGURE 2 Overview of the steps of Genetic Algorithm: Initialization, Fitness Calculation, Selection, Crossover, Mutation, and New Generation creation. The cycle is repeated until a termination condition has been reached [Colour figure can be viewed at wileyonlinelibrary.com]

generational process which consists of fitness evaluation, selection, crossover, or reproduction and mutation.[24] An overview of the process can be seen in Figure 2.

The problem is initialized by a set of randomly selected *chromosomes*, which are candidate solutions to the optimization problem. This whole set is called *population*. Each of the variables to be optimized are named *genes*, so that a set of genes forms a chromosome. In the first step, the fitness function for each of these chromosomes is computed (fitness evaluation), and the ones with the highest fitness value are chosen (selection). These selected chromosomes will be combined in the reproduction step, generating the crossover population. Mutation is incorporated in the algorithm to prevent the population from stagnating at any local point. It consists of altering one or more genes in a chromosome from its original state. The offspring together with the muted chromosomes will form the next generation. The algorithm repeats the generational process until a termination condition has been reached.

There are several modifications to this GA method with more complex interactions[25,26]; however, here, we employed the simplest one with the most popular operators[27]: fitness proportionate selection method for parent selection and one-point crossover for the reproduction step.

2.4 | Baseline models

This section provides a brief explanation of the algorithms that will be used for comparison with the proposed model in Sections 4 and 5. These are models currently developed at Solar Monkey[28] for PV system monitoring and fault detection.

The current approach of PV system monitoring at Solar Monkey consists of an analytical algorithm, hereafter called *analytical* model. By making use of weather data from nearby stations and an accurate skyline profile to account for obstacles surrounding the PV modules, they are able to predict the PV power produced by each of the systems in their fleet. Details on the framework can be found in de Vries et al.[29] This approach is also employed for fault detection by using a fraction of actual yield over the expected yield.

Based also on a physics-based approach, Solar Monkey developed a method for fault detection which consisted of estimating the number of panels that would generate the actual energy yield: the *sizing yields* model. Comparing the estimated amount of panels with the actual value, overestimation and underestimation in the PV systems can be detected.

The third method for fault detection is based on historical yields, referred as the *year-over-year* model. For systems older than a year, the PV power produced is compared to that one year earlier. This model is unavailable in the first year of operation of the PV system.

3 | DATA

The data employed come from the fleet of PV systems available at Solar Monkey.[28] It consists of the daily energy yield and main characteristics of 12229 roof-top PV systems, for a period ranging from 2 months to 7 years. While finer data resolution such as hourly data can provide higher quality results, it was available only for a limited number of systems; hence, daily energy yields were employed. All systems are spread across the Netherlands.

Data cleaning was performed, consisting on removing systems with large amount of missing data, after which the total number of PV systems was reduced to 9,480. Data splitting was performed on these systems between training and testing sets. While the training set is employed for model optimization, the testing set is used for model evaluation so the results are not influenced by the training stage.[30] A 10:1 system split between training and testing set was used to reduce the computational burden while ensuring a sufficient number of systems for testing.

4 | RESULTS

This section provides the results of the proposed algorithm. In subsection 4.1, the main outcomes of the optimization of the proposed model are explained. Then, the performance of the proposed model is compared to that of the other available PV system monitoring algorithms, namely, P2P and the analytical model, in subsection 4.2. Finally, the limits of the proposed model are found and discussed in subsection 4.3.

4.1 | Optimization

Once the model was developed and the optimization algorithm was chosen, GA was trained on a set of random 5,000 systems with each having daily yields for up to 1 year. The whole dataset was not used due to computational and memory limitations, although the peers were found from the whole set of 9,480 PV systems. The results of the optimization can be found in Table 1.

The first three weights correspond to the individual characteristics of the PV systems employed to compute the feature distance. Among the three, the panel count has a very low weight, probably due to the normalization of the PV system yields. Panel tilt and panel azimuth have a 1:3 weight distribution. Regarding the weights of the three distances, the results show that the yield distance has the highest weight among all three. It is interesting to note that geographical distance has negligible weight compared to the other two distances.

The number of peers needed for accurate calculation of expected yields is in the range of 12–20 peer systems. Upon multiple simulations, it was observed that the fitness value of GA does not change significantly within a given range of number of peers. One possible reason is that the model uses a weighted median when calculating expected yields. Thus, within a certain range, the value of weighted median does not change the result by a significant amount. This topic will be further discussed in subsection 4.3.

The low weight of geographical distance seemed suspicious considering that location was the only feature employed in previously published peer-to-peer PV monitoring models.[31–33] Thus, further exploration was required. The first hypothesis was that there existed a high correlation between feature and yield distances. The Spearman correlation coefficient[34] was therefore calculated and resulted in a value of 0.53, which was not significant enough. Despite this, the optimization model was run again without the yield distance. It was found that geographical distance still had a low weight of 0.07, while now most of the weight was skewed towards feature distance. Another hypothesis considered was related to the data itself. The PV systems available are only from the Netherlands, which is a rather flat and small country. One major influence of geographical distance for this project is that it enables the peer systems to account for the changes in very local weather like clouds near the focus system, especially when the weather stations are farther away from the focus system. This is important when dealing with hourly or higher frequency energy yields. However, for this project, daily energy yields are used, and it can be postulated that in the Netherlands, the day-to-day weather is very similar throughout the country. With this logic in mind, it is possible that when dealing with a larger area or with non-uniform weather conditions in a different environment, geographical distance could have a higher impact.

In our case, this low weight of geographical distance indicates that peers are not necessarily located close to each other. As can be seen in Figure 3, peers of a focus system can be either closer to it or farther apart from the focus system. This new finding gives more flexibility to P2P approaches, especially when there are not many PV system around a contain focus PV system.

TABLE 1 Parameters of the proposed model optimized by GA

w_n	w_θ	w_ϕ	w_{feat}	w_{yield}	w_{geo}	N_p
0.01	0.24	0.75	0.13	0.87	0	12–20

4.2 | Models comparison

Once the proposed model was optimized and trained, its performance was assessed in the testing dataset. This set was composed of 500 PV systems with data up to 3 years.

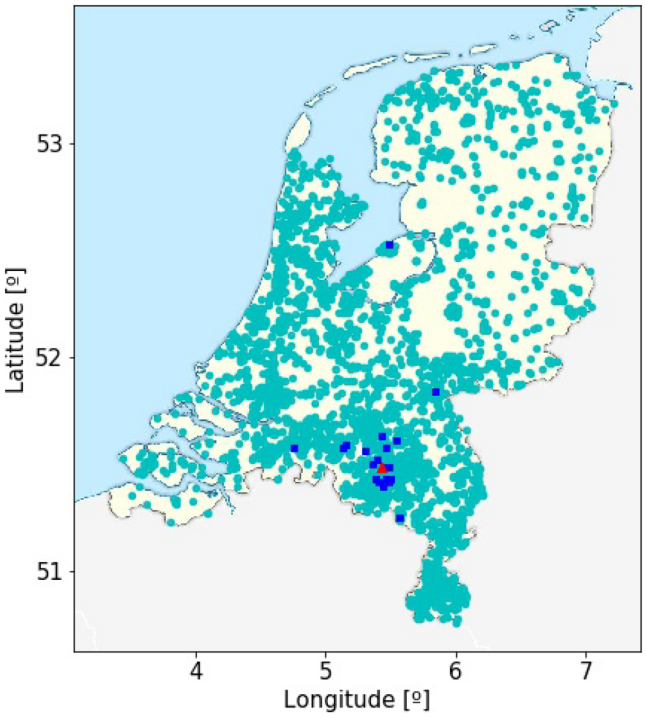

FIGURE 3 Spatial plot of good peer systems for a random focus system; focus system, red triangle; good peer systems, blue squares; all available systems, cyan circles [Colour figure can be viewed at wileyonlinelibrary.com]

In order to properly test its accuracy, the performance of the proposed model was compared to that of the P2P and the analytical model, explained in Sections 2.1 and 2.4, respectively. Two commonly used metrics were employed for this assessment: the Normalized Root Mean Squared Error (NRMSE) and the R^2 score, shown in Equations 8 and 9, respectively. For information on these metrics, the reader is referred to Zhang et al.[35] Table 2 depicts the mean value of these metrics for all systems in the testing set for the three models. Additionally, R^2 score was reinterpreted as a new metric: Percentage of Good Systems (PGS). Considering that an R^2 score higher than 85% is a good fit, PGS is the percentage of PV systems with an R^2 score higher than 85%.

$$NRMSE = \frac{1}{y_{max} - y_{min}} \sqrt{\sum_{i}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}, \tag{8}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{9}$$

where y_{max} is the maximum of actual values of y, y_{min} is the minimum of actual values of y, n is the number of observations, y_i is the actual value of y for observation i, $\hat{y}_i$ is the predicted value of y for observation i, and $\bar{y}$ is the mean of actual values of y.

All metrics show that the cooperative models have higher accuracy than the physical one. The daily yield shows that the analytical model underestimates the expected yields, while the P2P-based models are more accurate. This information is also represented in the error plot of Figure 4. Table 2 also depicts that the proposed model has better metrics than the previously published P2P model.[13] Hence, the addition of PV system data and the optimization with GA was an improvement.

TABLE 2 Metrics of the three monitoring models on the testing set

	NRMSE [–]	R^2 [%]	PGS [%]
Analytical	0.08	90.3	85.2
P2P	0.06	92.7	88.1
Proposed	0.05	94.1	92.5

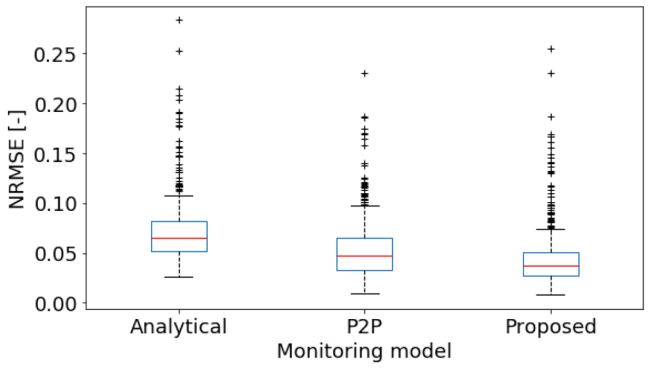

FIGURE 4 Error plot showing the NRMSE for all the test systems for each monitoring model [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | Use-case analysis

In this subsection, we elaborate on the limits of the proposed model. Characteristics such as the minimum data requirements for good performance and the sensitivity to the number of peers employed are explored. These would give a better perspective on the proposed model and show its robustness.

The first experiment consists of exploring the minimum data requirements for the model. Hence, the minimum number of PV systems required for training the model and the minimum number of days of data per PV system are determined. In this experiment, the trained systems still look for peers in the complete dataset of 9,480 PV systems. Fifty-six simulations were performed where both the amount of PV systems and the number of days of each system were tuned. Since cross-validation was not possible due to the high computational requirements, it was ensured that the training and testing sets had a similar distribution as the total dataset to have an accurate representation. The results of the experiment can be seen in a heatmap in Figure 5. The metric R^2 score was employed as a measure of performance.

The experiments show that number of systems does not have a strong influence on the metrics. However, if only a few months of data (<6 months) is available per system, the performance degrades strongly. Thus, when it comes to training the model, number of days per system is important while number of PV systems used for training is not as long as a minimum of 100–250 is available.

Although the experiment gives an understanding of the minimum data requirement for optimization, it was interesting to observe that even 100 PV systems might be fairly good enough to train. Thus, the results of the experiment were portrayed differently as seen in Figure 6. In this graph, the x axis shows the total number of years of data for each run. Thus, for grid point (4,000, 365) in the heatmap, the corresponding value on x axis would be 4,000 years of data. A trend line was generated on the experiment results as a help to the human eye. According to the trend line, data higher than 1,700 years are more than sufficient for training the model, that is, achieve R^2 score greater than 90%. This translates to 1,700 PV systems with 1 year data each or 850 PV systems with 2 years of data each and so on. It is important to note that data lower than 1,700 years can also lead to good performance, although it is not as reliable. The trend line saturates at around 4,000 years data, which is about 25% of the total data available.

The second experiment conducted was also related to the previous one, although it was way smaller. The objective was to determine the minimum data requirement when a new element is incorporated into the fleet. When a new PV system is added, energy yield data for the first 30 days of operation are sufficient to locate the good peers.

FIGURE 5 R^2 score as a function of number of days and number of training systems for minimum data requirement. The higher the R^2 score, hence the lighter the color, the better [Colour figure can be viewed at wileyonlinelibrary.com]

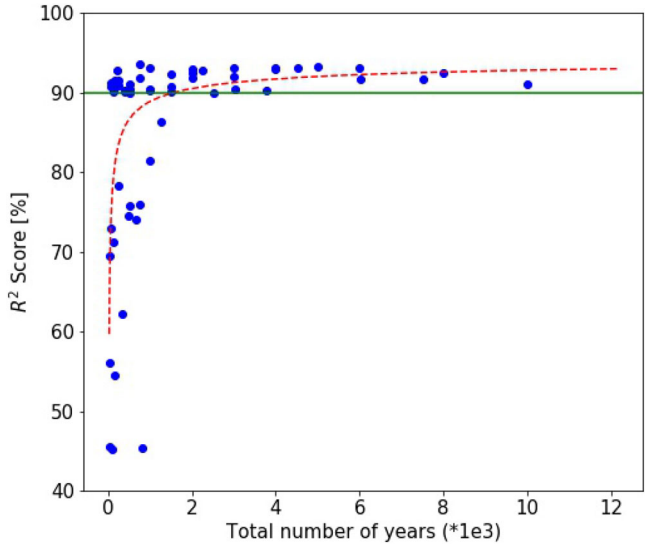

FIGURE 6 Minimum data required: R^2 score as a function of amount of data employed. The red dashed trend line is a help to the eye, while the green continuous line indicates an R^2 of 90%. For color references, refer to the web version of this article [Colour figure can be viewed at wileyonlinelibrary.com]

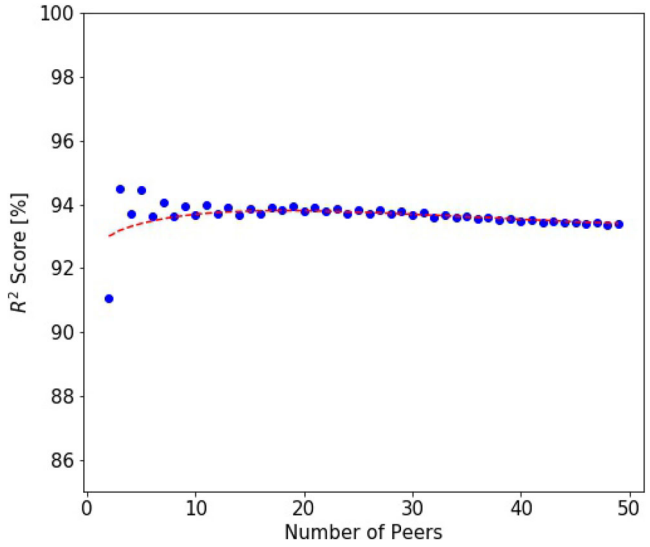

FIGURE 7 Learning curve of number of peers with respect to R^2 score [Colour figure can be viewed at wileyonlinelibrary.com]

These peers can be used for estimating the expected yields for the next one to two months with fairly good accuracy ($R^2 \approx$ 85%). After this time, good peers need to be located again with the new information available. These new good peers can be used accurately for another few months and so on. This process needs to be repeated until sufficient amount of yield data is available, at least 6–9 months. After 1 year of active monitoring, the frequency needed for distance training is low as long as the system parameters stay fairly constant.

The third and final experiment performed explored the best value for the number of peers. As mentioned in subsection 4.1, while the optimum value of the weights w does not usually change, the variable number of peers varies considerably over different simulations. Thus, a learning curve on the number of peers was plot in Figure 7 in order to find the range of number of peers that gave high performance.

It can be observed that the simulation has the highest R^2 score when the number of peers varies between 12 and 20. Less than 12 peers cannot ensure an accurate prediction due to high dependence on only a few systems. More than 20 peers also leads to lower accuracy in predictions most likely due to additional noise resulting from not so adequate peers, although this effect is highly mitigated thanks to the use of median.

5 | FAULT DETECTION

The main objective of this work was to develop a model for monitoring residential PV systems using a peer-to-peer approach. Monitoring PV systems consists of a two-step process: calculating the expected yield for the PV system and comparing that yield with the actual one to find any faults in the system. Once completed the development of the proposed model, this section focuses on the second step: fault detection.

The typical approach for fault detection consists of first filtering out poorly performing systems, and then finding out the origin of the poor performance by inspecting the PV system characteristics. The proposed model together with the analytical, system sizing, and year-over-year ones, explained in subsection 2.4, is employed in this process. These models are mainly used to detect the poorly performing systems and classify the type of fault. To do the former, the performance factor (PF) is calculated for all the expected yields. This metric computes the ratio of actual yield to expected yield, as shown in Equation 10. When the PF is outside an experience-determined range, the system is considered faulty; hence, it is selected for further inspection.

$$Performance\ Factor = \frac{Actual\ Yield}{Expected\ Yield} \tag{10}$$

One of the most important steps in this process is to distinguish between system design mismatch and fault detection. In the case of system design mismatch, the design should be simply updated to match the actual installation. Comparatively, in the case of fault detection, there is a drop in performance due to a problem with the system. In this latter case, the owner of the PV system should be notified to solve the issue and act accordingly.

Using the above two-step process along with the expected yield and actual yield temporal plots, some PV systems were scrutinized to diagnose the most common faults encountered. Based on this analysis, a fault categorization framework was developed, whose details are provided in the next subsection. Higher focus is given to two of these faults in subsection 5.2 where the added value of using a collaborative approach is highlighted.

5.1 | Categorization

In this section, a categorization of the most common encountered faults is made. Instead of using the whole database, 120 randomly chosen PV systems were checked to gain insight into the occurrence and type of faults. They were selected to ensure a good distribution of new and old systems. Only the data for the month of June 2021 were used for categorization, although the energy yields for the entire lifetime of the PV system was looked at for inspection. The reason for choosing June was that in summer months, PV systems are expected to produce the most energy yield in the year. A malfunction in summer could lead to significantly large energy and monetary loss for the owner than a malfunction in winter. Furthermore, in winter, fault detection becomes relatively harder due to very low energy yields.

Four criteria were used to separate the systems into groups, namely, the expected yields of the analytical model, the proposed model, year-over-year, and sizing yields. The fault categories found were no fault, missing data, under-performance, over-performance, and false positive as usual faults along with additional, peer-to-peer failure. Each of these categories can have multiple combinations of the four criteria, and they may or may not lead to a different fault diagnosis. The occurrence share of each of these faults can be seen in Figure 8. An additional category is included in the pie chart for systems whose fault was detected, but its origin could not be determined.

To begin with, a baseline of no fault would be when all four criteria are within an acceptable range, here when PF is between 93% and 120%. Most of the systems (62.5%) fell within this category. There is also a small possibility for false negative, but among the

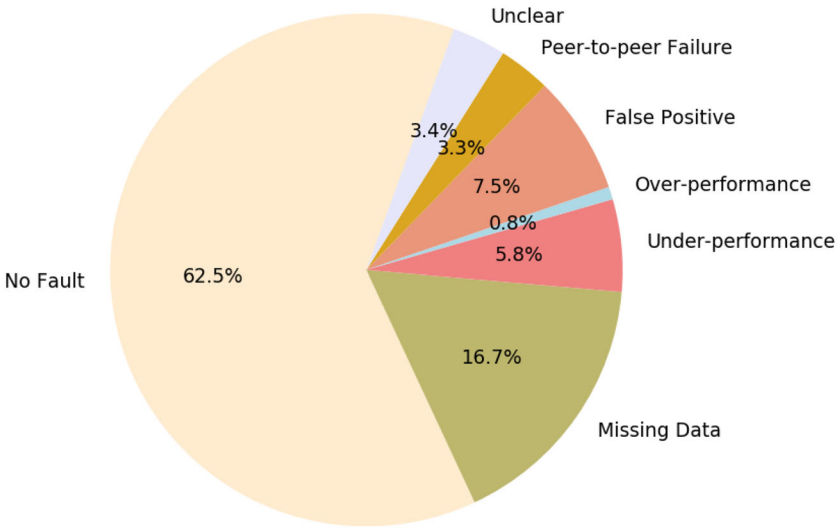

FIGURE 8 Occurrence of faults when categorization tested on 120 PV systems [Colour figure can be viewed at wileyonlinelibrary.com]

120 PV systems considered, none presented a clear case. The case of missing data was the most occurring fault detected, in 16.7% of the systems, although it is easy to detect due to automatic notice from the inverter.

In case of under-performance (5.8% of systems) and over-performance (0.8% of systems), if the fault occurred within the last year, it was reflected in all criteria. Yet if the fault persisted for longer than 1 year, year-over-year yields failed to flag the fault. Depending on the magnitude of under-performance, the diagnosis could either be a small, temporary problem, or else it could be due to broken panels or strings. In the case of broken panels, the magnitude of PF should drop by an equivalent amount. Similarly, on the rare occasion of over-performance, depending on the magnitude of the increase in actual yields, it could be a case of system size change. This should always be verified with the latest satellite images.

The fifth category is a false positive, when some of the models detect a fault that does not exist. This fault was detected in about 7.5% of the analyzed systems. It usually occurred when there was a mismatch between the actual installation and the system design details in the company database. The analytical and sizing yield models would be more easily fooled in this case due to their high dependence on the system design parameters. On the other hand, since both the year-over-year and proposed model depend more on historical yields, they are more robust against incorrect system data. Another case of false positive is when all except the year-over-year yields are within the acceptable range. Here, the possible diagnosis is that due to, for example, unusually sunny/cloudy days, the actual yields are higher/lower than last year. Since both analytical and sizing yields depend on weather data, they adjust accordingly to the unusual irradiance. Similarly, the unusual weather is experienced by most peer systems; thus, the proposed model does not flag either.

The final category and the one checked with caution were the case of peer-to-peer failure. In this case, only the proposed model detected a failure. This occurred mainly due to poor distance training related to poor quality or low availability of data. Another cause of failure occurred for systems older than 5 years with degradation in the performance. Since the distance training was done in early years of the PV system, the model does not adjust to the performance degradation over the years and, thus, detects it as a fault. Compared to that, both analytical and sizing yields include system decay, while year-over-year yield experiences the degradation in a controlled manner. The occurrence of this fault was not very high, only 3.3%, and it can generally be solved by distance retraining of the systems in order to ensure proper expected yields. Figure 9 shows the fault categorization flowchart.

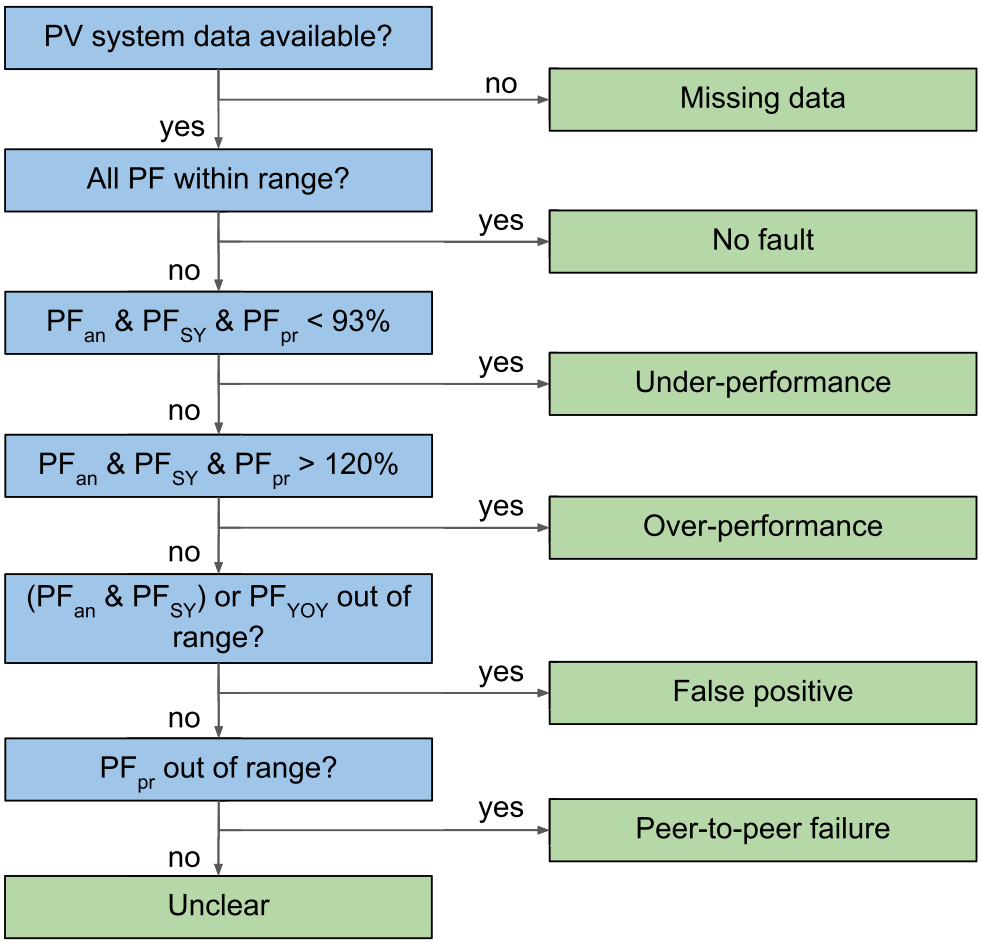

FIGURE 9 Flowchart depicting the main process of fault categorization, where "an" is analytical model, "pr" is proposed model, "YOY" is year-over-year, and "SY" is sizing yield [Colour figure can be viewed at wileyonlinelibrary.com]

It was expected to find some examples of inverter limited systems among the 120 checked. There were a few systems present whose inverter capacity was lower than the total system watt-peak, yet none of them seemed to have an energy generation high enough to be limited by the inverter. Note that while these systems represent the total dataset, the statistics should be taken with a grain of salt. Moreover, hourly yield data would have been more appropriate to properly detect inverter clipping.

5.2 | Key examples

In this subsection, some examples of PV systems with certain faults are presented. Examining these PV systems will help understanding how the collaborative approach fits into the categorization framework and can be a useful addition to fault detection. The proposed model plays a key role in the detection of over-performance and false positives, thus examples of PV systems with these faults are presented in this subsection.

5.2.1 | Common fault

This first type of example is intended to serve as baseline and to show how all the models are able to detect a common fault such as broken strings or panels. The example PV system has 58 panels, formed by separate sets of 31 panels and 27 panels which differ in both type and orientation.

Among the models described in subsection 2.4, all the fault detection checks give a red flag, suggesting that the system has been under-performing to around 40% as seen in Table 3. There is a sudden drop in the actual daily yield around the month of September 2020 as seen in the yield plots in Figure 10. As the fault occurred within the last 1 year, year-over-year model is able to flag the issue. In case this fault persists beyond September 2021, year-over-year yields will not be able to flag the fault anymore.

While all these checks simply suggest that the system is under-performing, the fact that under-performance has been consistent or in other words the performance factor being consistently around 40% suggests that only a part of the PV system has developed a fault. The sizing yields estimate the system size as 25 panels instead of 58 ($\approx$ 40%) that were installed, indicating that part of the system has likely broken down.

TABLE 3 Performance factor according to each model for a PV system with broken strings or panels

	Analytical	Year-over-Year	Sizing	Proposed
PF	37.0%	47.0%	40.1%	46.9%

5.2.2 | Over-performance

Due to the capital-intensive nature of PV systems, it can be preferred by residential owners to install their systems in steps on their roofs. Thus, a few years after a PV system has been installed, there have been instances when the owner has decided to increase the capacity of the PV system. This is usually not reported back to the monitoring company which leads to system over-performing. Considering the fault detection checks for this particular example as seen in the first row of Table 4, all models except for year-over-year flag the system as over-performing. When the yield plots in Figure 11 are checked, it is clear why this is the case. This system has been over-performing since April of 2019 by 140%

TABLE 4 Performance factor according to each model for an over-performing PV system before and after retraining

	Analytical	Year-over-year	Sizing	Proposed
Before retraining	143.0%	106.9%	136.2%	139.3%
After retraining	143.0%	106.9%	136.2%	99.5%

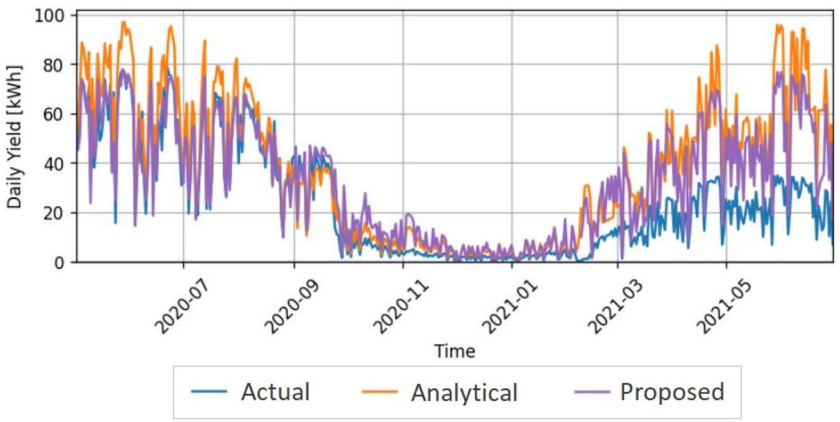

FIGURE 10 Temporal yield plots: example of a PV system with broken strings or panels. For color references refer to the web version of this article [Colour figure can be viewed at wileyonlinelibrary.com]

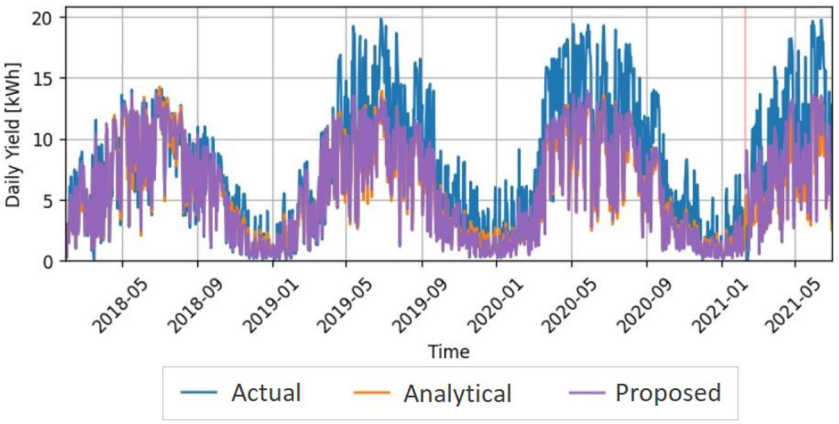

FIGURE 11 Temporal yield plots: example of over-performing PV system due to an increase in system size. For color references, refer to the web version of this article [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 12 System design developed with 8 PV panels in the company software before installation [Colour figure can be viewed at wileyonlinelibrary.com]

which, looking at the system specifications, is equivalent to a size of 11 panels instead of 8. This was verified when the design image in Figure 12 is compared with the latest satellite image in Figure 13.

A particular advantage that the proposed model has in such scenario is that it can be tweaked by changing the time period over which the distance training is done. In the above case, distance training was done prior to the system change. Since system size change is not a malfunction in the system, this system can be retrained with the new PV system size to absorb this change in system. Hence, distance training can be done after the system change, in order to find different peer systems that indicate this focus system is performing as expected. The results of retraining can be seen in Figure 14, and in the second row of Table 4. While the analytical and sizing yields still suggest that the system is over-performing, the proposed model does not flag any faults. This distance retraining is especially important if a true fault occurs now, since it will be flagged by the proposed model while the analytical one might suggest only a lesser over-performance.

5.2.3 | False positive

One of the strengths of using peer-to-peer yields for fault detection is the ability to detect false positives. A false positive occurs when the analytical model suggests an issue with the system despite the system not having any particular malfunctions. As illustrative example, the chosen PV system has 14 panels according to the database. Furthermore, from the fault detection checks in Table 5, most models suggest that the system is under-performing. When these three checks give a red flag, the first conclusion would be that there might be a few broken panels. However, the proposed model suggests that the system is performing perfectly fine with a 97% performance factor. To make sense of this discrepancy, the yield plots are inspected.

From Figure 15, it can be deduced that the system has been under-performing since the day it was installed. Since the proposed model was trained during this alleged under-performance, it does not consider this to be any fault. From the yield plot, it can be deduced that there is a mismatch between the system design details and the

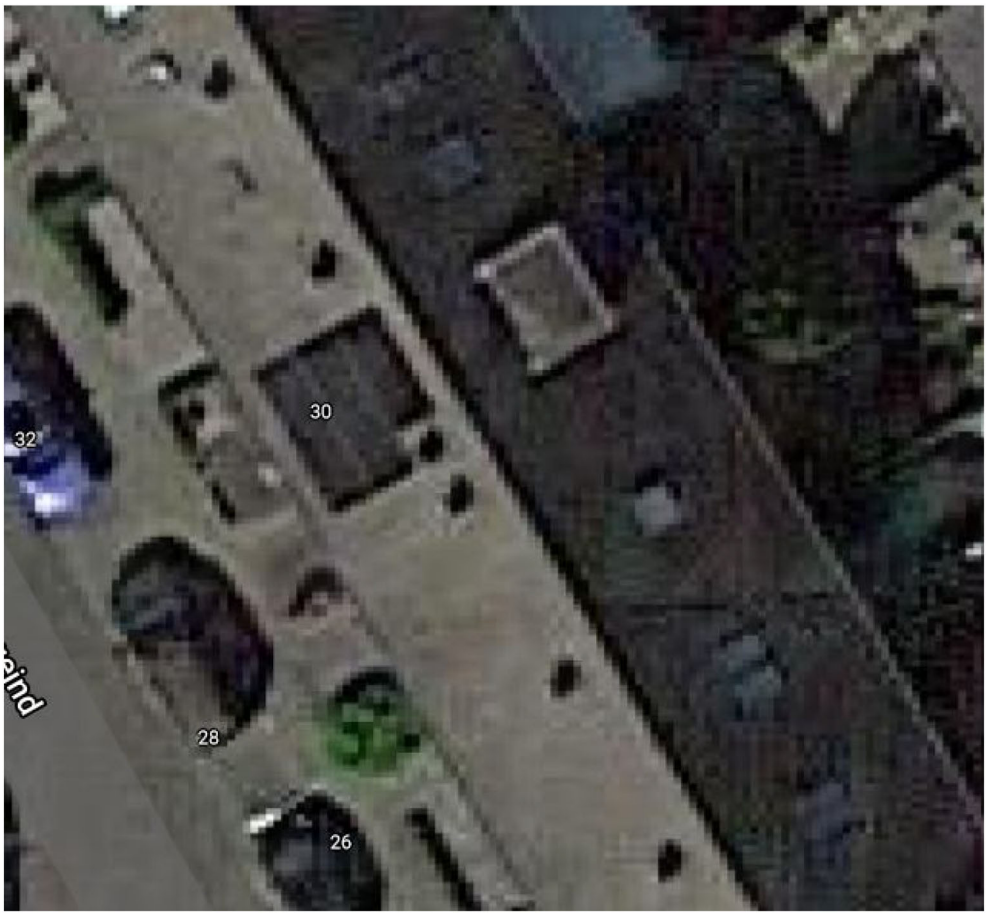

FIGURE 13 Latest satellite image available (2020) for the PV system with 11 panels [Colour figure can be viewed at wileyonlinelibrary.com]

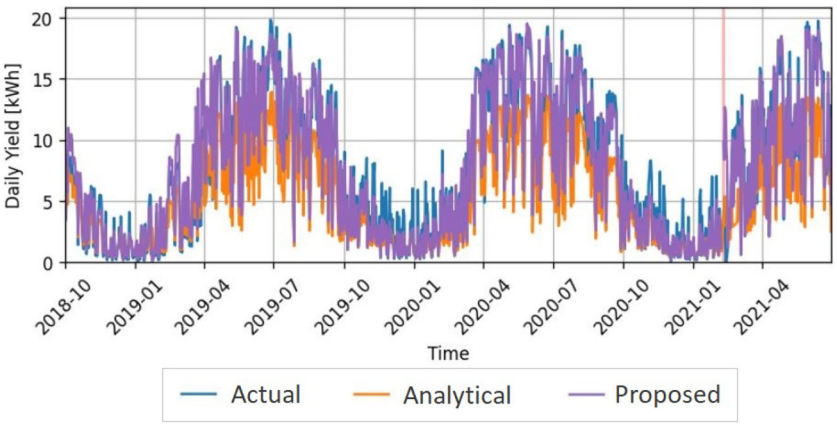

FIGURE 14 Temporal yield plots: over-performing PV system after retraining to absorb the increase in system size. For color references, refer to the web version of this article [Colour figure can be viewed at wileyonlinelibrary.com]

actual installation. This suggests that the under-performance flag by the analytical models is a false positive as the system has been performing as historically expected according to the proposed model and there is no malfunction in the system that might degrade its performance.

This kind of mismatch fault is the cause for many persistently under-performing systems. These systems are a time-sink monitoring companies, and thus, it is key to simplify the diagnosis. The proposed approach can be very handy in such scenarios.

The categorization framework and the examples given are based on the theory, discussions, and experience. Unfortunately, there was no possibility to validate the conclusions as this would entail physically visiting 120 PV installation sites all over the country or contacting the corresponding home owners. Nonetheless, the objective for undertaking fault detection was to show how the proposed approach can be used for fault detection.

TABLE 5 Performance factor according to each model for a PV system when a false positive is detected

	Analytical	Year-over-year	Sizing	Proposed
PF	53.7%	82.0%	55.8%	97.0%

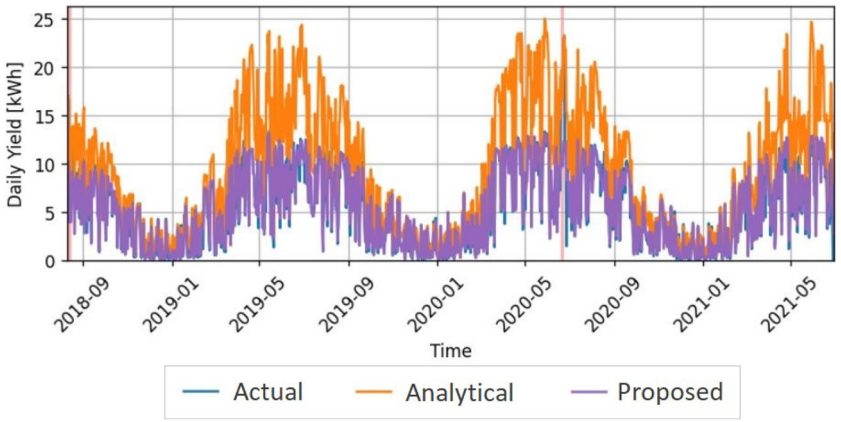

FIGURE 15 Temporal yield plots: example of false positive PV system. For color references, refer to the web version of this article [Colour figure can be viewed at wileyonlinelibrary.com]

6 | CONCLUSION

In this work, we have developed a model based on comparisons between neighboring and similar installations for PV system monitoring and fault detection. The model was based on a previously developed performance to peer approach, which has been improved by the addition of PV systems parameters and the use of machine learning techniques.

The viability of the proposed model has been demonstrated in the fleet of Solar Monkey consisting of more than 12,000 residential PV systems with up to 7 years of data per system. The developed model showed an average R^2 score of 94.1% and normalized root mean squared error of 0.05 on all tested PV systems. This implied an improvement in terms of R^2 score of 1.4 percentage points with respect to the baseline performance-to-peer model and of 3.8 points with respect to the analytical one. This superiority with respect to analytical models was thanks to its independence on inaccurate weather data and lower dependence on PV system parameters. A use-case analysis was also performed to find the limits of the proposed model. It was discovered that 1,700 years of data were required for proper model training, with a minimum of 6 months of data per system and 100 PV systems.

The usage of the developed model for fault detection and categorization has also been demonstrated. This model has the strength of distinguishing from incorrect PV system information and actual faults. Moreover, distance and peer retraining provide the flexibility to adapt the model to changes occurred in the PV systems. Although validation of these faults was not possible, the proposed model has demonstrated to be a good tool in combination with other developed models for fault detection and diagnosis.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the "Increase Friendly Integration of Reliable PV plants considering different market segments," under Grant Agreement 952957, Trust PV.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Alba Alcañiz https://orcid.org/0000-0002-8786-7832
Olindo Isabella https://orcid.org/0000-0001-7673-0163
Hesan Ziar https://orcid.org/0000-0002-9913-2315

REFERENCES

1. groupInternational Energy Agency. Snapshot of Global PV Markets; 2021. http://www.iea-pvps.org
2. Smets A, Jäger K, Isabella O, van Swaaij R, Zeman M. Solar energy. In: The Physics and Engineering of Photovoltaic Conversion, Technologies and Systems, Vol. 20; 2012.
3. SolarPower Europe. EU Market Outlook for Solar Power 2020-2024; 2020. http://www.solarpowereurope.org
4. Refaat SS, Abu-Rub H, Sanfilippo AP, Mohamed A. Impact of grid-tied large-scale photovoltaic system on dynamic voltage stability of electric power grids. *IET Renew Power Gener*. 2018;12(2):157-164.
5. Antonanzas J, Osorio N, Escobar R, Urraca R, Martinez-de Pison FJ, Antonanzas-Torres F. Review of photovoltaic power forecasting. *Sol Energy*. 2016;136:78-111.
6. Kumar DS, Yagli GM, Kashyap M, Srinivasan D. Solar irradiance resource and forecasting: a comprehensive review. *IET Renew Power Gener*. 2020;14(10):1641-1656.
7. Sobri S, Koohi-Kamali S, Rahim NA. Solar photovoltaic generation forecasting methods: a review. *Energy Convers Manag*. 2018;156:459-497.
8. Elsinga B, van Sark W. Spatial power fluctuation correlations in urban rooftop photovoltaic systems. *Prog Photovolt Res Appl*. 2015;23(10):1390-1397.
9. Lonij VPA, Brooks AE, Cronin AD, Leuthold M, Koch K. Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. *Sol Energy*. 2013;97:58-66. https://doi.org/10.1016/j.solener.2013.08.002
10. Golnas A, Bryan J, Wimbrow R, Hansen C, Voss S. Performance assessment without pyranometers: Predicting energy output based on historical correlation. In: IEEE Photovoltaic Specialists Conference; 2011:2006-2010. https://doi.org/10.1109/PVSC.2011.6186347
11. Tsafarakis O, Sinapis K, Van Sark WG. PV system performance evaluation by clustering production data to normal and non-normal

operation. *Energies*. 2018:11(4). https://doi.org/10.3390/en11040977. [Online]. Available: http://www.mdpi.com/journal/energies
12. Popovic I, Radovanovic I. Methodology for detection of photovoltaic systems underperformance operation based on the correlation of irradiance estimates of neighboring systems. *J Renew Sustain Energy*. 2018;10(5):53701. https://doi.org/10.1063/1.5042579
13. Leloux J, Narvarte L, Desportes A, Trebosc D. Performance to peers (P2P): a benchmark approach to fault detections applied to photovoltaic system fleets. *Sol Energy*. 2020;202:522-539. https://doi.org/10.1016/j.solener.2020.03.015
14. Leloux J, Narvarte L, Luna A, Desportes A. Automatic fault detection on bipv systems without solar irradiation data. In: 29th European Photovoltaic Solar Energy Conference and Exhibition; 2014.
15. Elsinga B, van Sark WG. Short-term peer-to-peer solar forecasting in a network of photovoltaic systems. *Appl. Energy*. 2017;206:1464-1483.
16. Korn GA, Korn TM. *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. Vol 15. Dover Publications; 1961:421. https://doi.org/10.2307/2003035
17. Stojiljkovic M. Stochastic gradient descent algorithm with Python and NumPy. *Real Python*. 2021. Available: https://realpython.com/gradient-descent-algorithm-python/
18. Nelder JA, Mead R. A simplex method for function minimization. *Comput J*. 1965;7(4):308-313.
19. Panda S, Padhy NP. Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design. *Appl Soft Comput J*. 2008;8(4):1418-1427. https://doi.org/10.1016/j.asoc.2007.10.009
20. Abraham A, Yue B, Xian C, Liu H, Pant M. Multi-objective peer-to-peer neighbor-selection strategy using genetic algorithm; 2007:443-451. https://doi.org/10.1007/978-3-540-77220-0_41
21. Rehman S, Khan B, Arif J, Ullah Z, Aljuhani AJ, Alhindi A, Ali SM. Bi-directional mutual energy trade between smart grid and energy districts using renewable energy credits. *Sensors*. 2021;21(9):3088.
22. Sun S, Kim K-Y, Shin O-S, Shin Y. Device-to-device resource allocation in lte-advanced networks by hybrid particle swarm optimization and genetic algorithm. *Peer-to-Peer Netw Appl*. 2016;9(5):945-954.
23. Holland JH, et al. *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*: MIT press; 1992.
24. Mitchell M. Genetic algorithms: an overview. In: Complex. Vol 1. Citeseer; 1995:31-39.
25. Blickle T, Thiele L. A comparison of selection schemes used in evolutionary algorithms. *Evol Comput*. 1996;4(4):361-394.
26. Shir OM. Niching in evolutionary algorithms. In: Rozenberg G, Bäck T, Kok JN, eds. *Handbook of Natural Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012:1035-1069. https://doi.org/10.1007/978-3-540-92910-932
27. Back T. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford university press; 1996.
28. Solar Monkey - Market leader in solar panel software. https://solarmonkey.io/
29. de Vries TNC, Bronkhorst J, Vermeer M, et al. A quick-scan method to assess photovoltaic rooftop potential based on aerial imagery and LiDAR. *Sol Energy*. 2020;209(February):96-107. https://doi.org/10.1016/j.solener.2020.07.035
30. Bishop CM. Pattern recognition. *Mach Learn*. 2006;128(9).
31. Berdugo V, Chaussin C, Dubus L, Hebrail G, Leboucher V. Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems. Next Generation Data Mining Summit (NGDM11), Athènes, Greece, 4 September 2011. https://hal.univ-lille.fr/INRIA/hal-02278607
32. Vaz AGR, Elsinga B, van Sark WGJHM, Brito MC. An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in Utrecht, the Netherlands. *Renew Energy*. 2016;85:631-641. https://doi.org/10.1016/j.renene.2015.06.061
33. Yang C, Xie L. A novel arx-based multi-scale spatio-temporal solar power forecast model. In: 2012 North American Power Symposium (naps) IEEE; 2012:1-6.
34. Spearman C. The proof and measurement of association between two things. *Am J Psychol*. 1987;100(3/4):441-471.
35. Zhang J, Florita A, Hodge B-M, Lu S, Hamann HF, Banunarayanan V, Brockway AM. A suite of metrics for assessing the performance of solar power forecasting. *Sol Energy*. 2015;111:157-175.

How to cite this article: Alcañiz A, Nikam MM, Snow Y, Isabella O, Ziar H. Photovoltaic system monitoring and fault detection using peer systems. *Prog Photovolt Res Appl*. 2022; 1-15. doi:10.1002/pip.3558