



Critical comparison of power-based wind turbine fault-detection methods using a realistic framework for SCADA data simulation

Usama Aziz, Sylvie Charbonnier, Christophe Bérenguer, Alexis Lebranchu, Frederic Prevost

► To cite this version:

Usama Aziz, Sylvie Charbonnier, Christophe Bérenguer, Alexis Lebranchu, Frederic Prevost. Critical comparison of power-based wind turbine fault-detection methods using a realistic framework for SCADA data simulation. *Renewable and Sustainable Energy Reviews*, 2021, 144, pp.110961. 10.1016/j.rser.2021.110961 . hal-03187883

HAL Id: hal-03187883

<https://hal.science/hal-03187883>

Submitted on 1 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Critical comparison of power-based wind turbine fault-detection methods using a realistic framework for SCADA data simulation

Aziz, U.^{1,2,*}, Charbonnier, S.¹, Bérenguer, C.¹, Lebranchu, A.², Prevost, F.²
¹Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
²Valemo S.A.S, F-33323, Bègles, France

Abstract—Numerous power-based wind turbine (WT) fault-detection methods using supervisory control and data acquisition (SCADA) data are presented in the literature. However, their performance cannot be compared easily with one another because of the lack of a realistic benchmark. To address this concern, a novel and realistic simulation framework is presented. It utilises actual data recorded on five French wind farms located at different geographical sites and composed of WTs of different models. It was used to simulate power profiles for three-year data, generated from 25 different wind and temperature profiles on 25 different WTs. Thus, the benchmark enabled a rigorous comparison of the performances of power-based fault-detection solutions. The fault-detection performances of three detection methods were compared for four power-based fault and under-performance scenarios of various intensities. The results indicated that the fault-detection performance of a method can vary considerably depending on the environmental and operational conditions. Moreover, the most effective approach is the one that considers these operational and environmental variations in WT data. The detection performance for the four failure scenarios was also statistically analysed.

Highlights

- Lack of objective performance evaluation for power-based fault-detection methods
- Novel simulation framework for controlled performance comparison based on actual data
- Critical performance comparison of three power-based fault-detection methods
- Detailed analysis of operational and environmental effects on detection performance

Keywords: Wind turbine; Fault detection; Simulation framework; Critical comparison; Performance evaluation

Nomenclature: WT, wind turbine; O&M, operations and maintenance; SCADA, supervisory control and data acquisition; COE, cost of energy; OEM, original equipment manufacturer; PC, power curve; WS, wind speed; PA, pitch angle; RS, rotation speed; GB, gearbox; IEC, International Electrotechnical Commission; IEA, International Energy Agency; IRENA, International Renewable Energy Agency; kW, kilowatt; MW, megawatt, GW, gigawatt; MWh, megawatt-hour; ANN, artificial neural network; PD, probability of detection; PFA, probability of false alarms; Th, threshold; EST, estimated; PEM, performance evaluation matrix; NBM, normal behaviour model; RMSE, root mean square error

Word Count: (~9,500 words with prior approval of Editor-in-Chief)

1 INTRODUCTION

With the consistent increase in the global installed capacity of renewable energy systems [1], operation and maintenance teams constantly encounter the challenge of maintaining and increasing the production performance of existing wind turbine (WT) fleets [2].

Industrial operators require 24/7 production cycles, as downtime is linked to production loss and translates to financial costs for operators. Conditional maintenance is an effective method of ensuring

* = corresponding author details, aziz.usama@protonmail.com

production at a reduced maintenance cost. It relies on fault-detection methods that can detect any faults that affect a WT at an early stage.

Many WT fault-detection methods are available in the literature, but the analysis of their performance is often limited in scope. This limitation occurs because the data used to analyse the fault-detection performances are often obtained from a single geographical location and from specific machine makes or models. The fault scenarios used for validation are also limited. Additionally, the proposed methods do not subject themselves to comparison with other solutions. Hence, a convergence of solutions is not available as no approach has been demonstrated to be comprehensively performant. The key factors affecting detection performance have neither been identified nor analysed. This is because a realistic benchmark that enables the comparison of fault-detection methods on realistic and controlled data is lacking. This creates the motivation to develop a framework to compare existing solutions, quantify their detection capabilities, and establish their limitations.

1.1 Wind turbine fault detection

This study exploited supervisory control and data acquisition (SCADA) data for analysis, as SCADA systems are familiar and easily accessible. Owing to their easy accessibility, various monitoring and detection methods have been proposed in the literature. Industrial SCADA systems for WTs provide data that are normally stored at an average resolution of 10 min with measurements for different parameters.

Fault detection and monitoring approaches using SCADA data, as presented in the literature, can be classified into two categories based on some key variables. The two main parameters of interest from the health and fault-detection perspectives are ‘produced power’ and ‘temperature’ [3]. This paper focuses on methods that utilise the produced power. WTs inherently function as power generators, making the monitoring of produced power a key objective of their asset management strategy. Unwanted and unexpected decreases in the produced power are associated with non-optimal operational states or faults. For simplicity, ‘fault’ subsequently will refer to both faults and under-performing production states of a WT.

1.1.1 Power-based fault detection

Based on the understanding that faults or under-performance affect the produced power, various fault-detection approaches based on monitoring power curves have been proposed in the literature. However, power data often require pre-processing before their use in fault detection. The International Electrotechnical Commission (IEC) standard is a commonly used method for processing power curves [4]. The IEC provides a standardised method of calculating the mean power curve. The mean power curve is determined by grouping data into wind intervals called ‘bins’ and averaging the produced power within these intervals [4].

The literature on power-based fault detection can be categorised into two solution families, namely, ‘model comparison’ and ‘residual’-based approaches. Refs. [5],[6],[7] belong to the first family of methods in which two power curve models are compared. A reference power curve is first built offline on reference data under normal conditions. A second curve is built online during the observation period. Fault-detection indicators are created by comparing these two power curves. These methods require

sufficient data to build complete power curves before the comparison. Their advantage lies in the simplicity of power curve modelling, as the level of data analysis is relatively rudimentary.

The second major category of fault-detection solutions is based on generating power residuals. A residual is the difference between the expected power calculated using a model and the power measured online. Within the residual-generation methods, two subfamilies further emerge. These subfamilies are distinguished based on the selection of so-called ‘explicit’ or ‘implicit’ approaches to model the produced power. Once power production is modelled using historical data, online residuals can be generated.

The authors in [8] generated residuals through the implicit modelling of produced power using artificial neural networks (ANNs). Wind speed (WS), pitch angle (PA), rotation speed (RS), nacelle temperature, gearbox (GB) oil temperature, etc. were used as inputs. The work proposed in [9] employed cluster analysis and pattern mining using WS, PA, RS, and power, while [10] used five data mining algorithms and compared their modelling performance. In [11],[12], the authors used data mining algorithms for detection, diagnosis, and isolation, while [13],[14], [15] used Gaussian process regression to model power. Additionally, WS and density were included as inputs to improve the modelling performance. In [16], the authors used a multistage ANN to predict power. WS, density, turbulence, wind shear, wind direction, and yaw were identified as appropriate inputs to reduce modelling errors. The authors in [17] provided a literature review of machine learning methods used to monitor WT conditions, while [18] presented a comprehensive overview of strategies on power curve modelling.

The explicit modelling for residual-generation solutions proposed in the literature operates with the graphical representation of power versus another variable. When the corresponding variable is WS, the curve becomes a power curve, which is an industry standard. Other representations in the literature are generally referred to as performance curves. The authors in [19] used RS vs. power and RS vs. PA curves, while [20] used the standard power and performance curves such as WS vs. PA and WS vs. RS curves for monitoring. The authors in [21] used power-curve data and placed a graphical envelope around the shape for monitoring purposes. Another study proposed by [22] was inspired by the IEC Standard’s ‘method of bins’ [4] and used the linear operating region of the power curve, while [23] used the constructor-provided power curve as a reference model.

The strength of residual-building approaches lies in their superior online monitoring capabilities. When the offline prediction model is learned, power prediction can be performed in real-time, and fault detection becomes possible. Historic residuals can also be used to examine the evolution of the overall behaviour of the system. However, the generated residuals can be affected by environmental and operating conditions if these conditions are not properly considered by residual-generation methods.

This study focused on understanding the effect of environmental and operational variations and evaluating different strategies to account for these variations.

1.2 Problem statement

An approach to objectively compare different fault-detection methods proposed in the literature is required. For a thorough comparison, the approach must consider changes in operational and environmental conditions and their effects on residuals. Previous studies overlooked these

considerations. Frequently, the detection performance of fault-detection methods is assessed using unique, irreproducible fault scenarios that are often unique. Analysis is often performed using simulated data with additive Gaussian white noise, which is an oversimplification of reality. In summary, as evident from the literature review, the relevant publications do not address the problem of objective and critical performance evaluation of the WT fault-detection methods, which is the contribution of this study.

Representative comparisons of fault-detection methods require the use of near-real-case scenarios. This requires a realistic simulation framework and a test setup with sufficient test scenarios to include environmental variations, operational variations, different fault scenarios, and sufficient data samples to perform a rigorous analysis.

Thus, the major research questions become

- i) Can power curves be used to monitor WT conditions, despite inherent environmental and operational variations?
- ii) What type of fault-detection performance can be expected, what key factors affect the detection performance, and how?

To address these concerns, a novel approach of generating a controlled stream of data is proposed. The time-series data of the produced power were simulated as a function of wind speed and external temperature. Data streams for both fault-free and fault scenarios were generated. Realistic fault signatures were used to output the faulty data streams. This faulty data were of the desired duration and could be inserted into the fault-free data stream at the desired locations.

This study built on and extended the preliminary simulation approach proposed in [24] and the preliminary validation approach presented in [25]. The identified gap in the current approaches and the formulated research questions provided a foundation for the original contributions of this research, and are summarised as follows:

- i) The proposed simulation framework enables the production of extensive and controlled simulation data (non-white noise).
- ii) It makes a critical and objective performance evaluation of fault-detection methods, as illustrated using the three methods.
- iii) It enables the effects of operational and environmental variations on detection performance to be evaluated, as highlighted in this paper.

The remainder of this paper is organised as follows. Section II presents the proposed simulation framework in detail and briefly describes the performance evaluation criteria. Section III presents the validation benchmark setup with the source data presentation and the fault scenarios used. Three power-based fault-detection solutions are introduced and compared using the proposed benchmark. The results obtained using the simulation framework (Section II) and validation benchmark (Section III) are presented, discussed, and rigorously analysed in Section IV. The conclusions and future considerations are presented in Section V.

2 PROPOSED SIMULATION FRAMEWORK

A simulation framework is required because finding multiple, identical fault signatures on geographically distant and operationally different wind farms is often difficult. This makes isolating the effects of failures from those of operational and environmental factors difficult. Thus, a controlled but realistic simulation setup, as partially introduced in [24], is required to achieve the objectives of this study, i.e. to answer the lack of realistic and controlled data, to facilitate the evaluation of the effects of environmental and operational variations on the detection performance, and to enable a robust and extensive performance comparison of detection methods.

2.1 Simulation process synthesis

To simulate realistic power data, a two-stage process was first developed. This process relied on real power-curve data and the manifestation of various faults on the power curves as well as the dispersion of observed data owing to varying environmental and operating conditions.

- The constructor power curves, expert knowledge, and/or historical data were used as references to generate various representative power curves for fault and fault-free scenarios.
- The representative power curves generated were transformed into fault and fault-free data with the addition of real dispersion data, assuming that the dispersion data were not affected by a fault.

2.1.1 Reference power curves

The relationship between the power produced by a WT and the measured wind speed can be expressed by the curve shown in Fig. 1. Power curves are industry-wide standards; therefore, they are familiar to producers and operators. The power curve depicts the power generation behaviour under normal or fault-free scenarios, and any deviation is classified as an anomaly or a fault. Faults reduce the production capacity of a WT, which can significantly affect the power curve [21].

To begin the simulation process, the reference power curves for normal and faulty behaviours were first created. Two options for establishing these reference curves were available. The first was to use the power curve provided by a manufacturer as a reference. Fig. 1a shows the normal behaviour of a fault-free power curve, also called the constructor power curve provided by the original equipment manufacturers (OEMs) to represent the production behaviour of their WTs. This power curve depicts the production behaviour of a WT during the fault-free period. The second option was to learn a measured reference power curve from fault-free historical data using the method of bins [4]. The mean power curve shown in Fig. 1a can be represented as $(P_{\omega_i} = f(\omega_i))$, where the averaged produced power (P) in wind bin ω_i is a function of wind speeds (ω) in the same bin (i). Wind bins often have a resolution of 0.5 ms^{-1} .

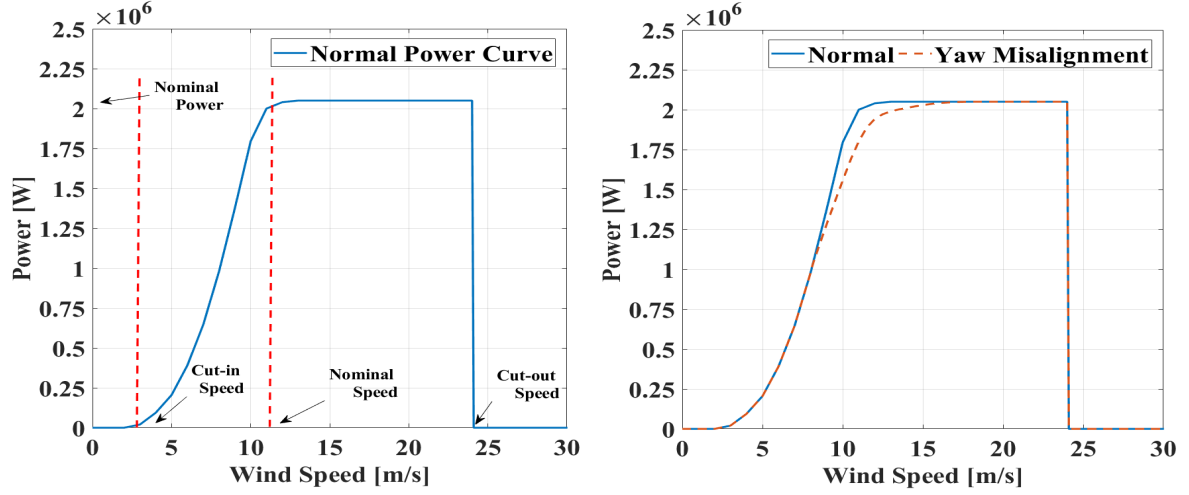


Fig. 1. Example of power curve a) an ideal fault-free power curve, also referred to as constructor power curve b) fault power curve; fault type yaw misalignment.

Similar to the normal behaviour, fault power curves replicating the behaviour of a WT power curve under ‘abnormal’ scenarios were created based on the literature review [21], expert knowledge, and analysis of real fault scenario data. Any faults affecting the shape of the power curve were simulated. As an illustration, Fig. 1b shows the power curve under a specific fault (yaw angle misalignment) and under normal conditions.

2.1.2 Learning of dispersion and reference matrix creation

The second step necessary for the realisation of the simulation process is to capture the realistic data dispersion around the power curve by calculating residuals and sorting the calculated dispersion residuals into an appropriate reference set.

To learn the dispersion profile of real data, 10-min SCADA data from actual WTs operating under normal conditions for several years were used. Fig. 2 shows the real power-curve data from a 2-MW WT operating under normal conditions for the years 2014–2016. It also shows the reference mean power curve ($P_{\omega_i}^{ref} = f(\omega_i^{ref})$) calculated using the IEC binning method [4] from the fault-free reference data ‘ref’.

Note that in Fig. 2, in contrast to the ideal power curve shown in Fig. 1, real data have a significant dispersion around the mean power curve. A significant amount of this data dispersion can be associated with variations in onsite air density, which is a consequence of onsite temperature variation [26]. The visual representation of this variation is shown in Fig. 2, which presents the dataset with colour bars associated with the corresponding onsite temperature values (θ). Fig. 2 shows the data after basic filtering. Outliers corresponding to erroneous samples (owing to measurement errors) and known maintenance actions were removed. Only data in which the WT produced power were maintained.

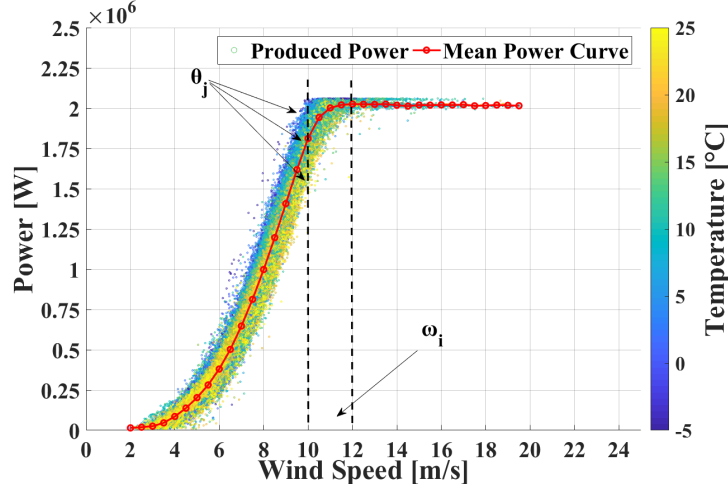


Fig. 2. Scatter plot of 10-min wind speed and power for a turbine over a 3-year period for dispersion learning; the mean power curve and colour corresponding to the temperature value of each sample are also shown.

As shown in Fig. 2, it is important to observe that in the same wind speed bin ω_i , multiple values of produced power P at different temperature values θ are present. Hence, the P values can be stored in the corresponding wind speed bin ω_i and temperature bin θ_j , where N_{ij} is the number of samples in the cell indexed by (ω_i, θ_j) . As an example of the varying temperature values, the data corresponding to the wind bin 10–12 m/s (Fig. 2) will intuitively fall within a varied range of temperature bins (Fig. 3).

This fragmentation can be elaborated further as a 2D reference of i wind bins (ω_i) and j temperature bins (θ_j) (Fig. 4), where one 2D reference bin (marked red in Fig. 4) is defined by the wind bin and the temperature bin that contain a set of N_{ij} power values denoted by $\{P_{\omega_i, \theta_j}(\ast)\}$.

The data dispersion around the mean power curve (Fig. 2) is captured as the *difference* in the mean power curve values ($(P_{\omega_i}^{ref} = f(\omega_i^{ref}))$) learned using the IEC method of bins [4] and the produced power data ($P_{\omega, \theta}$) around this mean power curve. As explained earlier, the produced power data are fragmented into i wind bins and j temperature bins. For each 2D reference bin defined by the wind bin ω_i and temperature bin θ_j , the set of residuals $\{r_{\omega_i, \theta_j}\}$ on power values is calculated using Eq. 1.

$$r_{\omega_i, \theta_j}(l) = P_{\omega_i}^{ref} - P_{\omega_i, \theta_j}(l) \quad l = 0, \dots, N_{ij} \quad (1)$$

where

N_{ij} is the number of samples within wind speed bin ω_i and temperature bin θ_j ;

$r_{\omega_i, \theta_j}(l)$ are the dispersion residuals calculated for wind speed bin ω_i and temperature bin θ_j ;

$P_{\omega_i}^{ref}$ is the mean power value in wind speed bin ω_i calculated on reference data;

$P_{\omega_i, \theta_j}(l)$ is the measured power within wind speed bin ω_i and temperature bin θ_j ;

The number N_{ij} of dispersion residuals ($r_{\omega_i, \theta_j}(l)$) in each wind and temperature bin (ω_i, θ_j) is variable. Fig. 5 shows a colour image of the number of data dispersion residuals in each cell of the 2-D reference matrix, filled using Eq. 1. The colour scale is determined by the number (N_{ij}) of the dispersion

residual samples in each cell of the 2-D reference (Fig. 4). The fault-free data dispersion residuals $r_{\omega_i, \theta_j}(l)$ calculated on the reference data are stored in a residual reference dataset (R_{ω_i, θ_j}), as presented by Eq. 2.

$$R_{\omega_i, \theta_j} = \{r_{\omega_i, \theta_j}(\ast)\} \quad (2)$$

where

R_{ω_i, θ_j} is the set of dispersion residuals sorted in wind speed bin ω_i and temperature bin θ_j as a function of ω and θ values.

Note that the resolution of each cell in this 2-D reference (R_{ω_i, θ_j}) can be varied. For the sake of simplicity, the examples presented in Figs. 3 and 4 use a bin resolution of 2 m/s by 5 °C. However, for this paper, the resolution is selected to be 0.5 m/s by 1 °C and each cell is subsequently populated by l dispersion residual ($r_{\omega_i, \theta_j}(l)$) entries. A wind bin resolution of 0.5 m/s is standard [4] for power curve analysis; however, the selected value also depends on the availability of data. A finer resolution may result in a lack of sufficient reference data per bin; thus, a careful trade-off might be required. The size of the empty 2-D reference R_{ω_i, θ_j} before the population must be consistent and representative of real data sample values.

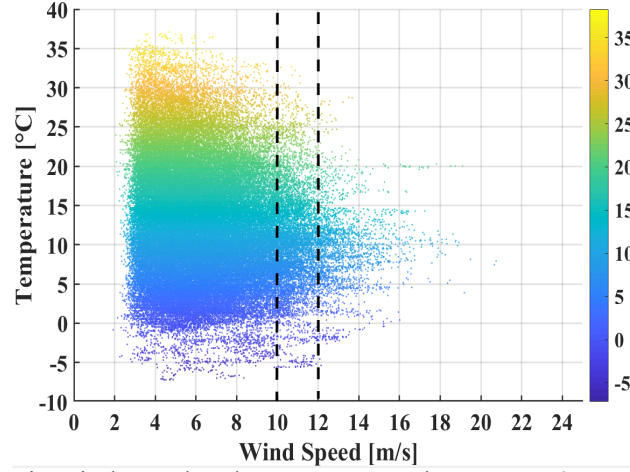


Fig. 3. Scatter plot of 10-min wind speed and temperature values over a 3-year period with wind bin 10–12 [m/s] marked with dotted lines.

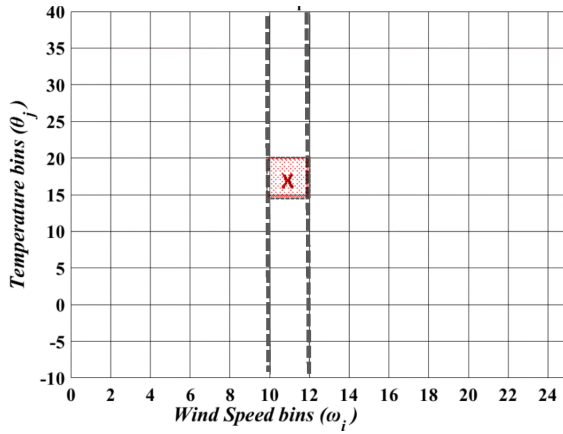


Fig. 4. 2-D ref. of i wind and j temperature bins to be filled with data corresponding to respective wind speed bin ω_i and temperature bin θ_j .

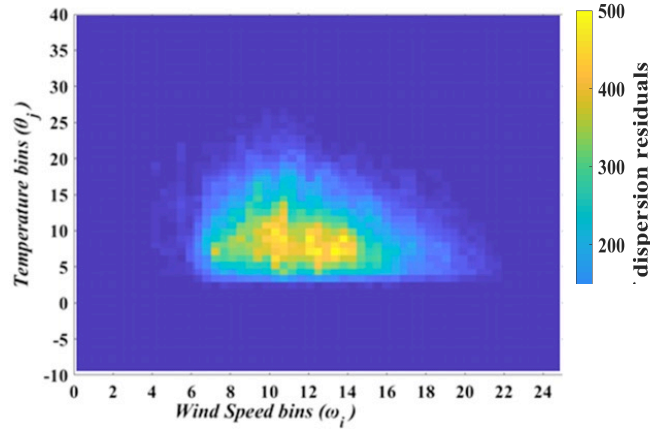


Fig. 5. Image of dispersion residuals set with colours scaled for number of residuals per wind and temperature bin (ω_i, θ_j) .

For this implementation, the reference ranges for the wind speed and temperature were selected as 0 to 25 m/s and -10 to 40 °C, respectively. The bin resolution of 0.5 m/s by 1 °C for this reference range (0 to 25 m/s and -10 to 40 °C) resulted in the size of R_{ω_i, θ_j} being 50×50 . Note that the total number of reference bins with *available and non-empty reference data* in the 2-D reference is a function of the dataset under consideration and can vary. The different pixel intensities in Fig. 6 further elaborate the varying number of samples in each reference cell. The pixel density in the image indicates that most data lie roughly within the boundaries of 4–22 m/s for wind and 2–25 °C for temperature values.

2.2 Generation of simulated data flows

2.2.1 Simulation procedure

A simulation process was now available to generate simulated data. This process relied on the following:

- ❖ Various realistic and useful reference fault power curve patterns, replicating the multiple fault scenario identified and created in Section 2.1.1.

- ❖ A realistic dispersion profile learnt and the 2-D dispersion reference matrix created in Section 2.1.2.

To simulate the 10-min power time series, the simulation process was used with wind speed and external temperature time series ($U(k), T(k)$) measured on different WTs as inputs. ($U(k), T(k)$) were time series of wind speeds and temperatures recorded every 10 min, where k is the time index. For this implementation, $k = 0, \dots, 157824$ (~ 3 years for a 10-min sampling rate) was used.

At each timestamp (k), the values of this ‘input’ pair of environmental parameters (wind $U(k) \in \omega_i$ and temperature $T(k) \in \theta_j$) were used to select the corresponding reference power value (step 1) and the dispersion profile value (step 2). The input wind data sample ($U(k) \in \omega_i$) was used to select the corresponding reference power value (P_{ω_i}) (Section 2.1.1). Both new environmental parameters (wind $U(k) \in \omega_i$ and temperature $T(k) \in \theta_j$) were used to identify the corresponding reference wind and temperature bin (ω_i, θ_j) of the reference dispersion residual set R_{ω_i, θ_j} built earlier.

When the correct reference bin (ω_i, θ_j) was identified, a dispersion value $\tilde{r}_{\omega_i, \theta_j}(k)$ from the corresponding bin of R_{ω_i, θ_j} was randomly drawn. The value of the ($U(k), T(k)$) pair operated as a pointer to the corresponding reference wind speed and temperature bins of the 2-D reference matrix. Hence, for each ($U(k), T(k)$) pair of 10-min data samples, a dispersion residual was randomly drawn from the relevant reference bin. Note that when randomly drawn, residual sample $\tilde{r}_{\omega_i, \theta_j}(k)$ was neither removed nor replaced and was available for the next and subsequent withdrawals throughout the simulation process (in a boot-strap-like approach).

The randomly selected dispersion $\tilde{r}_{\omega_i, \theta_j}(k)$ value was then added to the corresponding reference power value (P_{ω_i} & $P_{\omega_i}^q$) for the modelling of normal and faulty behaviour, respectively. A realistically simulated time series of power data $P_{U(k), T(k)}(k)$ as a function of the ($U(k), T(k)$) pair of input data could be generated using Eq. 3 for the fault-free scenario, and Eq. 4 for the fault scenarios as follows:

$$P_{U(k), T(k)}^{\cdot\cdot}(k) = P_{\omega_i} + \tilde{r}_{\omega_i, \theta_j}(k) \quad \text{for fault-free case} \quad (3)$$

$$P_{U(k), T(k)}^{\cdot\cdot}(k) = P_{\omega_i}^q + \tilde{r}_{\omega_i, \theta_j}(k) \quad \text{for fault profiles} \quad (4)$$

$$\left. \begin{array}{l} U(k) \in \omega_i \\ T(k) \in \theta_j \end{array} \right\}$$

where

$P_{U(k), T(k)}(k)$ is the simulated power output for the input pair ($U(k), T(k)$);

P_{ω_i} is the fault-free reference power curve;

$P_{\omega_i}^q$ is the fault reference power curve ranging from 1 to q depending on the fault mode selected;

$\tilde{r}_{\omega_i, \theta_j}(k)$ is the dispersion residual selected randomly from (l) entries in the reference bin (ω_i, θ_j);

q is the number of fault modes for the reference power curves available for simulation (Fig. 10).

Fig. 6 shows the input wind and temperature time series ($U(k), T(k)$) recorded for a 2-MW WT over three years (2014–2016). This input time series was used to create new simulations of power data ($P_{U(k),T(k)}$), as detailed in Section 2.2.1.

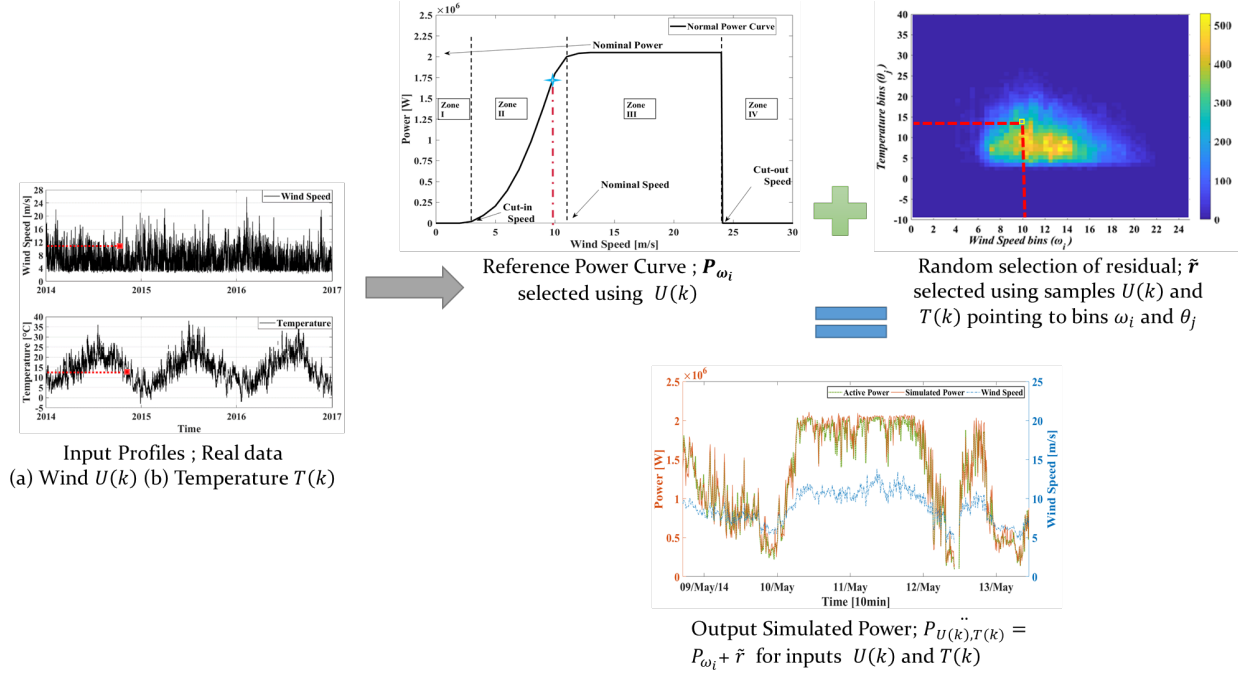


Fig. 6. Input profiles (Real data): (a) Wind $U(k)$, (b) temperature $T(k)$ used to select the reference power $P_{\omega_i}^q$ and random dispersion sample $\tilde{r}_{\omega_i, \theta_j}$ to simulate the output power ($P_{U(k),T(k)}$).

The simulation framework presented could then be used to generate a faulty and fault-free power data time series at length. The simulated data were entirely controlled and a realistic data stream could be selected based on the selected reference power curve mode. The data dispersion used replicated the dispersion observed in the real world.

2.2.2 Simulated data example

Fig. 7 shows the zoomed version of an example of simulated power against the wind samples for a one-week duration (19th to 26th April 2015). To validate that the simulated power is a response to the input wind speed time series used, some key moments are identified in the example. The horizontal red bar in Fig. 7a corresponds to the ‘nominal wind speed’ above which the produced power attained the ‘nominal power’ value, as indicated by the red line in Fig. 7b. The nominal wind speed for the example presented was 11 m/s, while the nominal power was 2.05 MW (Fig. 1). The figure shows that the moment wind speed values in Fig. 7a crossed the red line (nominal speed of 11 m/s), the simulated power in Fig. 7b attained the nominal power (2.05 MW).

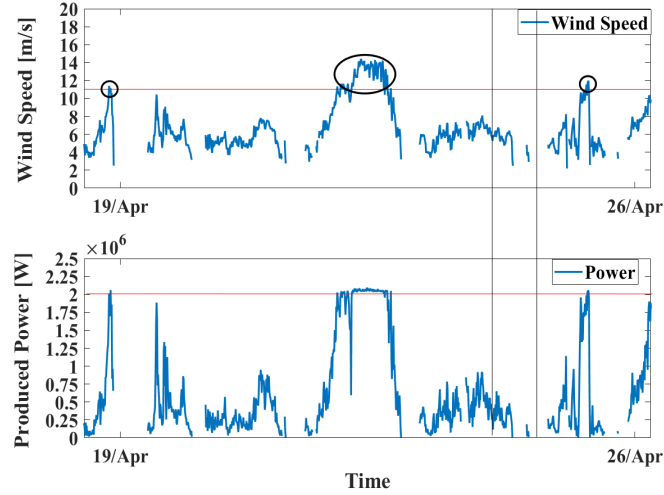


Fig. 7. (a) Input: Real wind profile and nominal speed drawn as a red line; (b) output: simulated power profile and nominal power drawn as a red line (1 week).

2.3 Evaluation of the performance of the detection methods

The time-series data generated using the aforementioned framework can be used to gauge the performance of fault-detection methods. Such an analysis is based on the evaluation of the detection performance indicators. Wind farm operators require automated monitoring methods that can generate alarms when the WTs experience faults. The detection mechanisms must be timely, robust, and with a minimal number of false alarms.

The principle of detecting the fault occurrence is based on the difference between the normal behaviour and observed behaviour of a characteristic operating variable. When this difference exceeds a predefined threshold, a fault is detected. Setting a proper detection threshold is an arduous task because fault indicators are imperfect and may vary even in fault-free scenarios. Therefore, a compromise is found between the detection of actual defects and the number of false alarms generated by the system. Despite its significance, few studies have focused on this aspect in the context of fault detection on wind turbines based on SCADA data [27].

2.3.1 Performance indicator definition

As mentioned earlier, a fault is detected by setting a threshold on the residual generated by the fault-detection methods. When the residual decreases below the detection threshold, an alarm is raised. The alarm raised can be a true detection (or true positive), i.e. an actual fault occurs on the WT, or it can be a false alarm (or false positive), i.e. no fault actually occurs on the WT. The performance of a detector can be assessed using a receiver operating characteristic (ROC) curve, which sketches the true detection probability (the probability of detecting a fault when one occurs) as a function of the false alarm probability (the probability of detecting a fault when none occurs). This curve is obtained by changing the value of the detection threshold [28].

The ROC curve is a monotonically increasing function. Changing the detection threshold value to increase the true detection probability also increases the false alarm probability. Fig. 8 shows an example of three ROC curves obtained using three different fault-detection methods, which will be

presented and used later in this paper, for a particular fault. Detector performances using ROC curves can be compared by calculating the area under the curve (AUC): the higher the AUC, the better the detector. Another solution that is frequently used in supervision is to select a particular value for the false alarm probability and compare the corresponding true detection probabilities. The higher the corresponding true detection probability, the better the detector. It is a standard approach in fault-detection theory and is called the Neyman–Pearson criterion. We selected the Neyman–Pearson criterion over the AUC because the true detection probability is easy to interpret and relevant for maintenance operators, while the AUC is not.

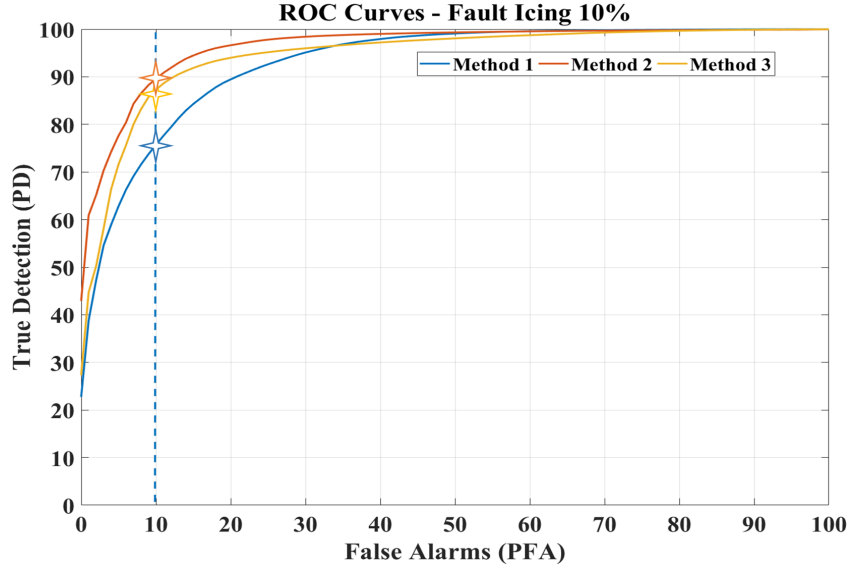


Fig. 8. Example ROC curves for Methods 1, 2, and 3 for fault type icing 10%; corresponding PD values are identified by stars against PFA 10%.

In this paper, we propose the use of a specific point on the ROC curve as the performance indicator, we set the false alarm probability to 0.1 (or 10%) and use the corresponding true detection probability, denoted as PD_{10} , as our performance indicator. This particular value for the false alarm probability is set by constraints from the industrial application, requiring a maximum level of false alarms equal to 10% (above this level, too many unnecessary maintenance actions could be triggered). False alarms can be very costly because they result in the intervention of maintenance technicians on sites that are often remote and sometimes difficult to reach.

The following section details the estimation of the true detection and false alarm probabilities using the proposed framework.

2.3.2 Protocol for performance evaluation

A three-year simulation time-series dataset was generated: Years 1 and 2 were fault-free periods. A fault was introduced in year 3. Let Th be the detection threshold set for the fault indicator. Whenever the fault indicator crosses Th , a fault is detected.

Year 1 (*learning period*) of fault-free data is reserved for learning fault-free behaviour using the implemented fault-detection solutions.

Year 2 (*validation period, setting of Th*) is used to select the detection threshold Th_{10} corresponding to the selected 10% false alarm rate. During this year, no faults are introduced. Hence, periods when

the residual decreases below the detection threshold correspond to false alarms. The detection threshold Th_{10} is set such that the percentage of false alarms during the second year is equal to 10%.

PFA_{10} is estimated as follows:

$$PFA_{10} = \frac{\text{No. of samples crossing threshold } (Th_{10})}{\text{Total No. of data samples in validation period}} \quad (5)$$

Th_{10} is set such as $PFA_{10} = 0.1$ (10% PFA)

Year 3 (*fault period*) is used to estimate the probability of detection (PD), which is used as the performance indicator for fault detection. During this year, a fault is always present. Hence, periods when the residual decreases below the detection threshold correspond to the true detections. Note that only the PD value corresponding to a PFA of 10% is used as a performance indicator in this analysis. The indicator is denoted PD_{10} .

PD_{10} is estimated as follows:

$$PD_{10} = \frac{\text{No. of samples crossing threshold } (Th_{10})}{\text{Total No. of data samples in the fault period}} \quad (6)$$

2.3.3 Performance evaluation strategy

For a particular fault-detection method and a particular fault, PD_{10} can be estimated using several time-series data streams generated by the framework. These data streams are generated using real data from wind farms located in distinct geographical locations and WTs built by different manufacturers. Each time-series datum can be generated using the dispersion residuals from one turbine of a particular wind farm and the environmental parameters recorded on another WT. If N_T is the number of WTs, a total of N_T by N_T time-series data can be generated, and an equal number of PD_{10} indicators can be calculated. Subsequently, the performance of a particular fault-detection method on a specific fault can be evaluated using the N_T by N_T performance-indicator values. These values are stored in tabular form (Table II) called the performance evaluation matrix (PEM). Using the calculated PEMs, further analysis can be performed to compare the performances of different fault-detection approaches.

2.4 Simulation and evaluation framework overview

The proposed simulation and performance comparison framework can now be summarised. Fig. 9 shows the step-by-step methodology for the performance comparison framework.

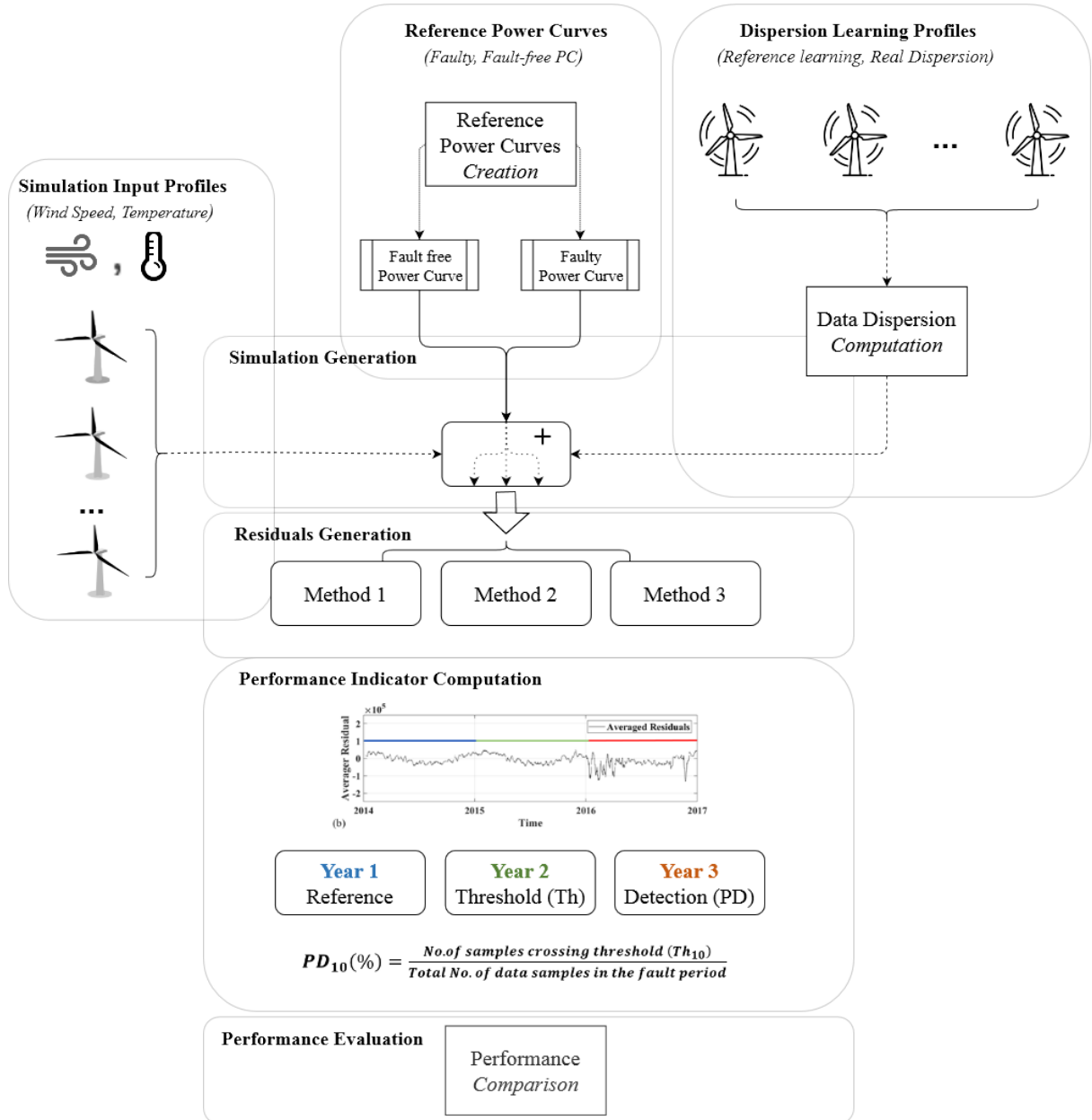


Fig. 9. Overview of the proposed simulation framework. Input profiles shown to generate simulated power. Residuals are generated and detection performance compared.

3 CASE STUDY: COMPARISON OF THREE FAULT-DETECTION METHODS ON FIVE DIFFERENT WIND FARMS

The simulation framework presented in the previous sections was used to set up a test bench for performance comparison. In this study, three fault-detection methods were selected as representative samples of the diverse existing fault-detection methods. Five different wind farms were selected for their diversity of operating and environmental conditions, ensuring technological and environmental variability. This experimental approach enables the exploration of the effects of geographical and operational variations on the performance of fault-detection methods.

3.1 Real data set description

A realistic simulation benchmark requires a deep-rooted foundation in the real dataset. Hence, 10-min interval SCADA data recorded during three years from five different wind farms of the research partners were utilised. These five wind farms provided a total of $N_T=25$ WT. One WT was either selected as a source of input environmental parameters (wind $U(k)$, temperature $T(k)$) or as a source of measured power data to learn the dispersion profile (R_{ω_i, θ_j}). Based on this, 625 ($N_T \times N_T = 25 \times 25$) simulation combinations were possible. For this analysis, each data stream had a three-year duration. The proposed simulation framework enabled the generation of an overall simulated data duration of 1875 years (625×3 years) by adding all combinations of simulation scenarios. The time-series data from years to 2014–2016 were used to generate residuals using different fault-detection methods.

Hence, five wind farms operated by research partners labelled V, L, D, S, and C were selected. The selection of the wind farms under analysis in this study ensured several key considerations: a fair representation of the research partner's fleet from mainland France, sufficient environmental variability, and sufficient operational variability. Fig. 10a shows a map of the wind farms selected from the on-shore fleet of the research partner. The geographical variability of the data source ensured sufficient environmental variability in the dataset. Fig. 10 b and c show the onsite wind and temperature distributions of these wind farms. As the figures show, the wind speed distribution for Farm S differed considerably from that of Farm V. Farm S was subjected to stronger winds, with a median value of approximately 6 m/s and maximal values up to 16 m/s. For Farm V, the median value of wind speed was 3 m/s, and the maximum value was approximately 12 m/s. Similarly, the onsite temperature was significantly colder for Farm V, with minimal values of approximately -5 °C and maximal values of 30 °C. The minimal value for Farm S was positive, and the maximal value was approximately 35 °C.



Fig. 10. (a) Geographical location of the five selected on-shore wind farms.

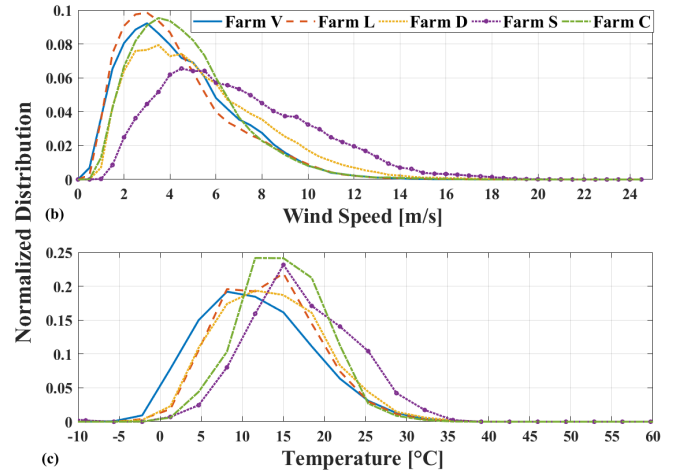


Fig. 10. (b) Wind profile distribution and (c) Temperature profile distribution of the five wind farms.

The operational variation of the dataset was ensured by the selection of data from wind farms with machines from different OEMs. Their characteristics are presented in Table I. The selected turbines had different models, hub heights, swept areas, rotor diameters, and rated powers. The wind farms ranged from three to six turbines per farm and were implanted in line, grid, L-shape, or arc configurations.

TABLE I. DATA BASE SUMMARY

Wind Farms	Wind Farm Characteristics				Model	Wind Turbine Specifications			
	# of WTs	Farm Location	Instal. Year	Farm Layout		Rated Power	Hub Height	Rotor Dia.	Swept Area
Farm V	6	Centre-North	2014	Line	SENVION MM92	2.05 MW	68.5/80/100 m	92.5 m	6,720.0 m ²
Farm L	3	Centre-East	2014	Grid	GE 2.5×L/100	2.5 MW	75/85 m	100.0 m	7,854.0 m ²
Farm D	6	Centre-North	2010	L-shape	VESTAS V90	2.0 MW	80/95/105 m	90.0 m	6,362.0 m ²
Farm S	5	South	2009	Line	ECOTECNIA 80	2.0 MW	70/80 m	80.0 m	5,027.0 m ²
Farm C	5	West	2010	Arc	SENVION MM92	2.05 MW	68.5/80/100 m	92.5 m	6,720.0 m ²

In summary, the variation in the selection of input farm data provided an opportunity to test environmental variation. The variation in data used to learn the dispersion profiles enabled the operational characteristics of the WTs to be captured.

3.2 Simulation test bench configuration

The simulated time-series data were used to set up a test bench. As mentioned earlier, one simulated time-series datum was composed of two years of fault-free data and one year of faulty data. The normal behaviour models required by each fault-detection method were learned in the first year of the simulated data stream. Year 2 of fault-free data was used to set the detection threshold to 10% of false alarms, and PD₁₀ values were calculated on year-3 faulty data. A unique fault scenario of a one-year duration was induced at the same period (year 3) for each simulated data stream.

Note that the proposed simulation framework enables flexibility in the location and duration of the faults induced in the dataset. The selection of the duration of threshold, learning, and validation periods is also flexible.

3.2.1 Faults induced

Four different families of faults, including (a) icing, (b) down-rating, (c) acoustic curtailment, and (d) yaw misalignment, were used. Different intensities of these faults could be generated. The faults tested were realistic and were selected to encompass a wide spectrum of faults of interest. The corresponding power curves of these four fault scenarios are shown in Fig. 11.

- Icing (v-low, low, medium, high): This type of failure appears as a uniform degradation in the operational zone on the power curve (Fig. 11a). In the time series of produced power, a downward shift is visible for data corresponding to moderate wind speed values.
- Down-rating (1%, 3.5%, 7%, and 15%): This type of failure appears as a fixed degraded value on the power curve for higher wind speeds (Fig. 11b). In the time series of the produced power, the downward shift is only visible for data corresponding to higher wind speed values.
- Acoustic curtailment: This mode of operation regulates surrounding noise levels. The produced power is degraded at higher wind speeds, but the degradation is different from down-rating (Fig. 11c).

- Yaw misalignment: WTs are equipped with a yaw motor to ensure that the most efficient angle along the wind direction is followed. A misalignment changes the optimal behaviour of the WT, and the produced power is degraded (Fig. 11d).

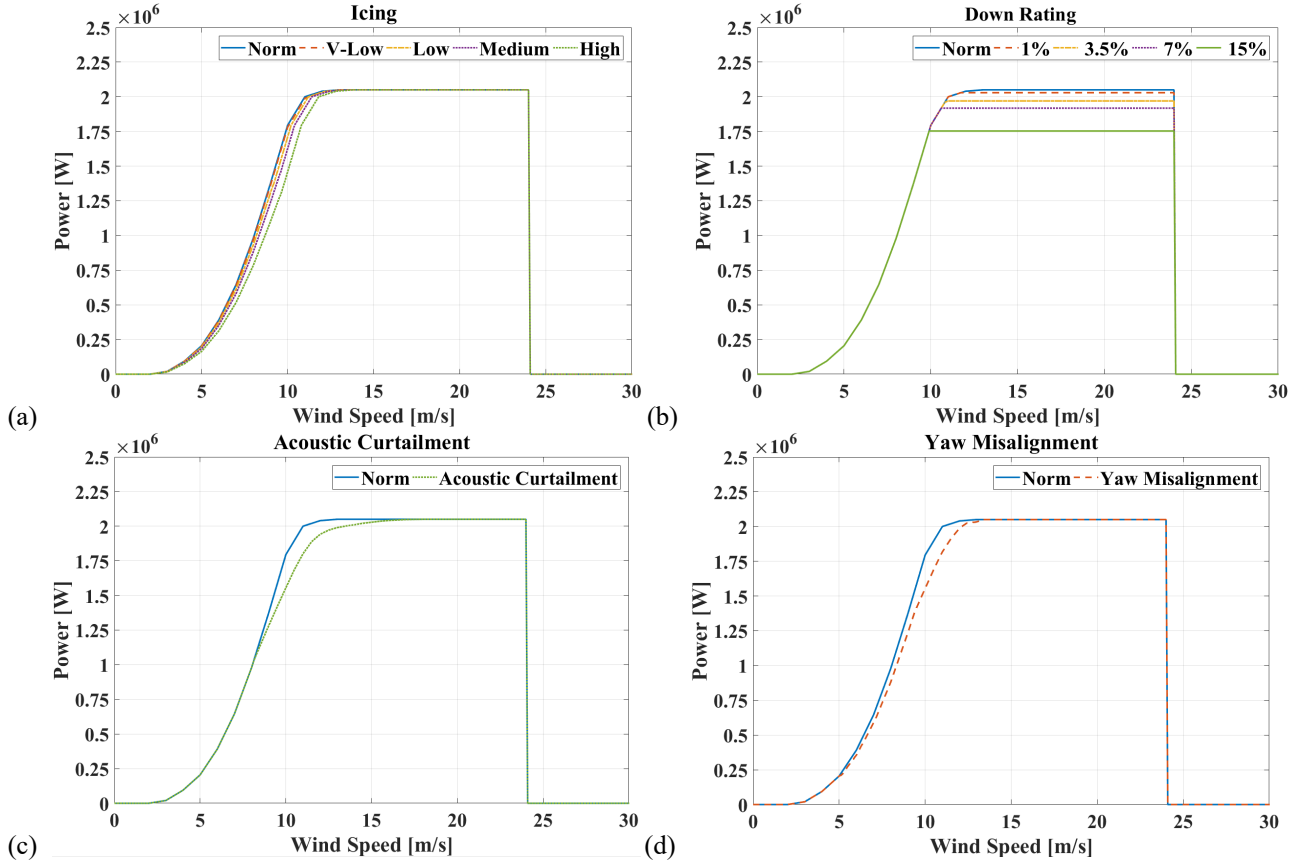


Fig. 11. Fault reference power curves for (a) icing, (b) down-rating, (c) acoustic Curtailment, and (d) yaw misalignment.

Note that down-rating and acoustic curtailment are operational modes, but unintended activation results in significant production loss; hence, they can be referred to as fault scenarios.

3.3 Evaluation of fault-detection methods

Three representative fault-detection approaches proposed in the literature were implemented in this study. All methods generated residuals by comparing the current 10-min SCADA power values to the expected values. The expected power values were predicted using a normal behaviour model derived from the power curve. The selection of these three fault-detection methods considered in this study was based on three different methods of ‘managing’ data dispersion and variability. Method 1 used only wind, Method 2 used wind, density correction, and data translation, and Method 3 used wind and air density for dispersion correction. Fig. 12 shows an overview of the different characteristics of the selected methods.

Methodologically, the first method used is a proposed residual-generation method inspired by the method of bins [4], the second is a more intricate method proposed by [22], and the third is a machine learning method based on Gaussian process regression presented by [13]. These methods are presented below in further detail.

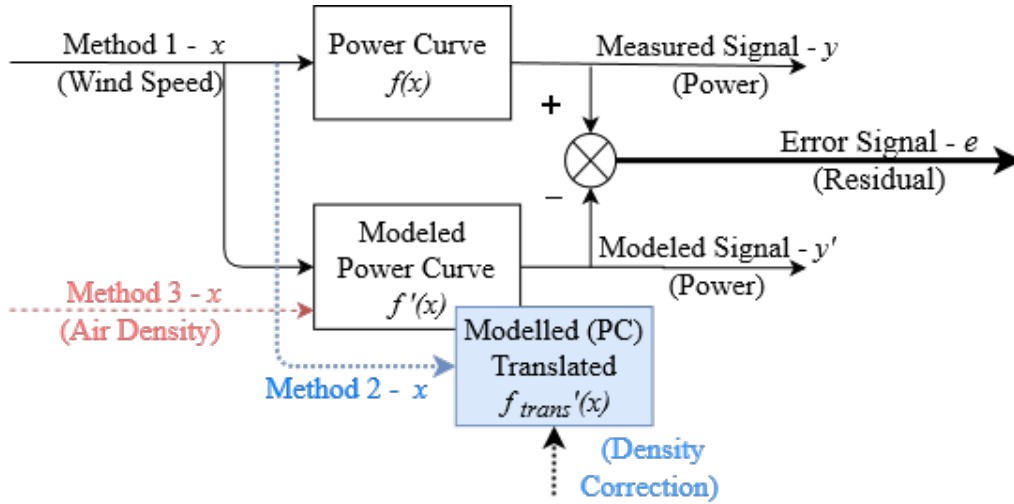


Fig. 12. Overview of three residual-generation-based fault-detection methods. Method 1 uses wind only; Method 2 uses wind, density correction, and data translation; Method 3 uses wind and air density.

3.3.1 Method 1: Approach inspired by the IEC

With this technique, the reference power curve is learned by binning data in wind speed intervals of a 0.5-m/s resolution. For each wind bin with a 0.5-m/s resolution, a reference mean value of the produced power data samples, as presented by the IEC standard [4], is computed. This in turn builds a fault-free reference power curve. The indicator for fault detection is calculated as the difference between all the produced power samples and the mean power value within each wind bin [24]. Fig. 13a shows an example of residuals calculated for a three-year data stream. The fault scenario presented in this example is a 15% down-rating during year 3 (2016–2017). The faulty period is indicated by a horizontal red bar. Fig. 13b shows the same residual with a one-week moving average.

3.3.2 Method 2: IEC-based density correction and translated data approach

The approach presented in [22] is one of the approaches of normal behaviour modelling. This technique generates a residual inspired by the IEC standard [4], but considers both environmental and operational conditions. Similar to Method 1, the data are binned into 0.5-m/s wind intervals, and the reference mean is calculated for each wind bin. However, according to the recommendation of the IEC standard [4], the data are corrected for onsite density variations and normalised to the reference density of 1.225 kg/m³. Because the reference mean lies in the centre of a wind bin, the residual is only calculated after translating all the data samples within each wind bin towards the bin centre [22]. The indicator for fault detection is calculated as the difference between all the produced power samples and the mean power value of two consecutive wind bins. Fig. 14a shows an example of residuals calculated for a three-year data stream. The fault scenario presented in this example is a 15% down-rating during year 3 (2016–2017). The faulty period is indicated by a horizontal red bar. Fig. 14b shows the same residual averaged with a one-week moving window.

3.3.3 Method 3: Approach based on Gaussian process regression

The authors in [13] used Gaussian process regression to model the WT power output. The learned model was used to predict the power output for each timestamp based on two inputs (wind speed and

air density). Fault-detection residuals were generated as the difference between the simulated power data samples and predicted reference power data. Fig. 15a shows an example of residuals calculated from a three-year data stream with an induced fault of 15% down-rating during year 3 (2016), as indicated by the red lines. Fig. 15b shows the one-week moving average of the same residual.

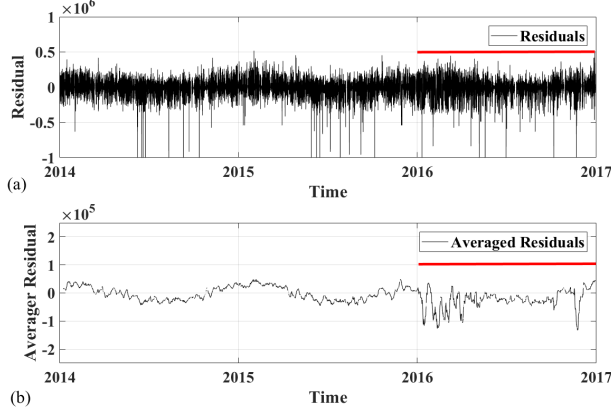


Fig. 13. Method 1 for 15% down-rating: (a) Unprocessed; (b) moving average; the red lines mark the fault period (year 3).

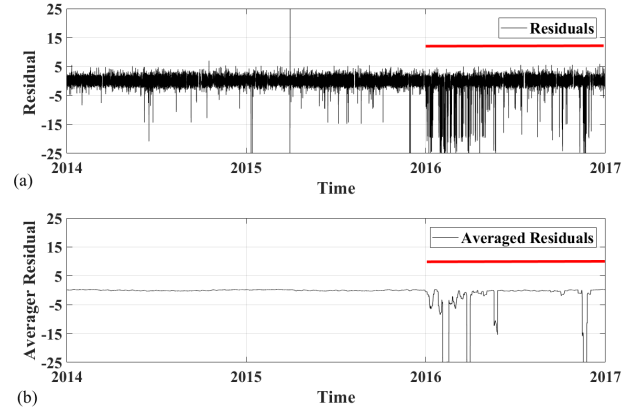


Fig. 14. Method 2 for 15% down-rating: (a) Unprocessed; (b) moving average; the red lines mark the fault period (year 3).

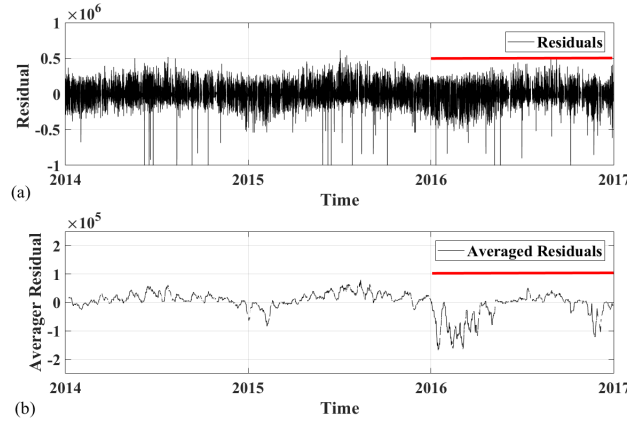


Fig. 15. Method 3 for 15% down-rating: (a) Unprocessed; (b) moving average; the red lines mark the fault period (year 3).

3.4 Performance evaluation matrix

The three fault-detection methods presented were applied to the 625 streams of three years of simulated data (i.e. $25 \times 25 = 625$), and their respective residuals were computed. To reduce noise, the residuals generated by all methods were smoothed by obtaining a one-week moving average, and the performance indicator (PD_{10}) was subsequently calculated as presented in Section 2.3. For each fault scenario and fault-detection method, 625 performance indicators (PD_{10}) were now available. They were stored in the PEM shown in Table II [25]. The letters in Table II are wind farm names, while the numbers (1,..., n) refer to the corresponding WTs in these wind farms. For example, V3 and D4 refer to the third and fourth WTs of Farms V and D, respectively.

The variation in environmental behaviour was ensured by the selection of various environmental profiles for simulation inputs, as presented in the rows of Table II. All the PD_{10} values in one **row** were calculated from data streams simulated from the same input environmental profile ($U(k), T(k)$),

measured on the same turbine. However, for each row, a dispersion set (R_{ω_i, θ_j}) was drawn from each of the 25 WT's (V_1 to C_n). All the PD_{10} values in one **column** were calculated from data streams simulated by 25 environment profiles ($U(k), T(k)$) (from V_1 to C_n). In each column, a dispersion profile drawn from the same WT was used (Section 2.2). Thus, extensive statistical analysis and performance comparison could be performed for each test scenario using this matrix.

TABLE II. PERFORMANCE EVALUATION MATRIX

			Operational Profiles						
			Farm-V			...	Farm-C		
Environmental Profiles			V_1	...	V_n		C_1	...	C_n
	Farm-V	V_1							
		...							
		V_n							
	\vdots								
	Farm-C	C_1							
		...							
		C_n							

4 PERFORMANCE EVALUATION: RESULTS AND DISCUSSION

The simulation framework proposed in this paper was used to evaluate the detection performance of the three fault-detection methods presented earlier. To achieve a realistic quantification of the global performance, different fault intensities were considered. The performance indicator was calculated to populate the PEM shown above. The results are presented and discussed in the following sections.

4.1 Performance evaluation according to the type of fault

4.1.1 Down-rating

Fault intensities of 1%, 3.5%, 7%, and 15% down-rating were selected for analysis. In terms of produced power, for a 2.05-MW WT, this down-rating translates to a curtailment from the rated maximum of 2.05 MW to approximately 2.03, 1.97, 1.92, and 1.75 MW, respectively. This curtailment is only activated at higher wind speeds. The selection of different fault intensity levels provides a comprehensive view of the fault-detection performance for this particular fault family.

Fig. 16 shows the mean PD_{10} value for each method, calculated for different intensities of fault mode down-rating. The mean value represented by a single bar per fault-detection method is calculated for the complete PEM (for all 625 simulations, environmental profiles, and dispersion profiles). The results indicate that, globally, the increase in fault intensity results in an increase in detection performance for all methods. However, the mean PD_{10} value for Method 2 is approximately 10% and 20% higher than that of Methods 3 and Method 1, respectively, for the fault intensities of 1%, 3.5%, and 7%. As the

fault intensity is increased to 15%, the fault signature becomes easier to detect, and the detection performance advantage of Method 2 is decreased to approximately 10% and 5% compared with Methods 1 and 3, respectively.

Note that the overall mean detection performance for fault-type down-ratings is relatively low. The PD_{10} values shown in Fig. 16 have the lowest performance value of approximately 7% for Method 1 and a fault intensity of 1%. The highest value of detection performance is 52% for Method 2 and a fault intensity of 15%. For a false alarm rate of 10%, the highest mean detection performance (PD_{10}) of 52% is relatively low. This can be explained by examining the fault signatures for the down-rating. Owing to the peculiar nature of the fault (visible only for high winds), the fault signature was not pertinent throughout the fault period (year 3). Because the PD_{10} is calculated over the complete fault period (year 3), the average detection values for the down-rating (visible only for high winds) are lower than those for a fault that was consistently visible throughout the entire fault period (year 3).

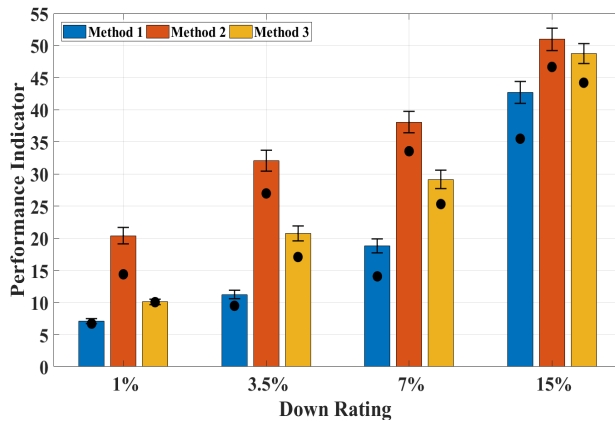


Fig. 16. Performance indicators for **down-rating (1%, 3.5%, 7%, and 15%)** for Methods 1, 2, and 3 with 95% confidence intervals and median values.

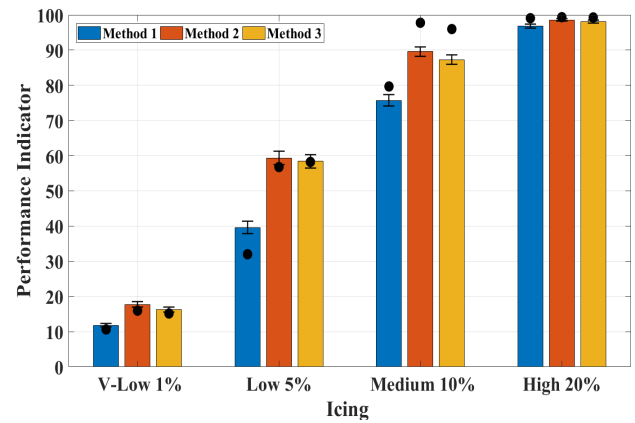


Fig. 17. Performance indicators for **icing (v-low, low, medium, and high)** for Methods 1, 2, and 3 with 95% confidence intervals and median values.

4.1.2 Icing on blades

The fault intensities (v-low, low, medium, and high icing) were selected for analysis. In terms of reduced production, for a 2.05-MW WT, these intensities translate to a reduction of approximately 1%, 5%, 10%, and 20% power, respectively. Unlike down-rating, this production loss affects all wind speeds. The hypothesis here is that this fault type changes the aerodynamics of the turbine blades and affects the production for all wind speeds. Note that although icing is a seasonal phenomenon, in this study, the fault was evaluated over the entire year (year 3). The selection of different fault intensity levels provides a comprehensive view of the fault-detection performance for this particular fault family.

Fig. 17 shows the mean PD_{10} value calculated for different intensities of the icing fault mode. The mean value is represented by a single bar for each method. The mean value shown is calculated for the complete PEM (over all the 625 simulations, environmental profiles, and dispersion profiles). The results indicate that, globally, the increase in fault intensity results in an increase in detection performance for all methods. The mean PD_{10} value for Method 2 remains ahead of that of Methods 1

and 3 for the fault intensities of 1%, 5%, 10%, and 20%. The advantage of Method 2 in detection performance compared to other Methods is reduced for the icing fault. As the fault intensity increased to 20%, the fault signature became easier to detect, and the difference in the detection performance became nominal.

Note that the overall mean detection performance for fault-type icing is high. The PD_{10} values in Fig. 17 have the highest values at approximately 97% for Method 2 and a fault intensity of 20%. For a false alarm rate (PFA) of 10%, the highest mean detection performance (PD_{10}) of 97% is very high. This can be explained by examining fault signatures for icing. The production loss for this fault type appeared as a downward ‘step’ shift in the fault period. As the fault intensity increases, the amplitude of the step shift increases, and the fault becomes easier to detect. Moreover, the icing fault is visible for all wind speeds, and the fault signature is pertinent throughout the fault period (year 3). Because the performance indicator PD_{10} was calculated over the complete fault period (year 3), the average detection values for icing (visible for all wind speeds) are higher than those for a fault that is only visible for high wind speeds (down-rating).

4.1.3 *Acoustic curtailment and yaw misalignment*

The mean detection performance results for various fault intensities of fault types down-rating and icing have been presented thus far. Two other important faults, namely, acoustic curtailment and yaw misalignment, were included in the analysis and are presented in this section.

Acoustic curtailment is a particular operational mode that is often activated at night to ensure compliance with acceptable noise levels. The activation is also a function of the wind direction and populated sectors near wind farms. In reality, this type of fault is occasional, but for the sake of this analysis, an active fault throughout the fault period (year 3) was considered. The fault signature for acoustic curtailment is only visible at high wind speeds. This is because the noise levels increase at higher wind speeds. Acoustic curtailment is a configuration activated by an operator. Because unintentional or faulty activation results in production loss, it is considered a fault.

The results (Fig. 18) indicate that for acoustic curtailment, the global value of PD_{10} for Method 2 is higher than that for Methods 1 and 3. For this fault type, the PD_{10} for Method 2 is approximately 13% higher than that of Method 1. The advantage of Method 2 in terms of detection performance compared to that of Method 1 is relatively low (3%).

Yaw misalignment occurs when the WT fails to align with the direction of the maximum wind. The control system of a WT is configured to adjust the yaw angle to enable maximum exposure to wind. Any misalignment results in suboptimal power production and is termed a fault. An 8° misalignment was used as a fault for this analysis, but the severity of this fault can vary. The fault signature of the 8° misalignment resembles the fault signature of the fault type 10% icing.

For yaw misalignment, the trend of the detection performance of Method 2 being superior continues (Fig. 18). Globally, the value of PD_{10} for Method 2 remained higher than that of Methods 1 and 3. For this fault type, the PD_{10} for Method 2 is approximately 11% higher than that of Method 1. Similar to acoustic curtailment, the advantage of Method 2 in detection performance compared with Method 1 is relatively low (3%).

Fig. 18 shows these two faults (acoustic curtailment and yaw misalignment) along with the down-rating and icing faults. The results for down-rating and icing in Fig. 18 represent the mean values across

all fault intensities. For each method, the down-rating in Fig. 18 is the mean of all values (1%, 3.5%, 7%, and 15% down-rating) presented in Fig. 16. Similarly, the results for icing in Fig. 18 are the global mean of all fault intensities (1%, 5%, 10%, and 20% icing) presented in Fig. 17.

4.1.4 Conclusion—fault-specific comparison

The performances of the three methods on each particular fault (type and fault intensity) were compared using paired t-tests. The aim was to determine whether, for a particular fault, one method (Method i) had superior performance over another (Method j), with i and j being 1, 2, or 3. Hence, two hypotheses were formulated. The null hypothesis (H_0) was that the mean value of PD_{10} is equal for methods i and j for a particular fault. The alternative hypothesis (H_1) was that the mean value of PD_{10} of Method i is higher than that of Method j for a particular fault.

For this analysis, an α risk of 1% was accepted. A conclusion is drawn for each combination of the two methods and for each fault. The results are listed in Table III. Because the methods were compared using PD_{10} values calculated for ten different faults, we used a Bonferroni correction for multiple testing. Therefore, the threshold set on the p-value to reject hypothesis H_0 was set to $\alpha/10$, i.e. 0.001. Whenever the p-value of the test is below 0.001, the mean value of PD_{10} of Method i is higher than that of Method j . Methods 2 and 3 outperform Method 1 for every fault, except for 20% high icing, which was seemingly very easy to detect. For this fault, the absolute PD_{10} values were higher than 95% for each method. Method 2 outperforms Method 3 for every fault except for low to high icing, for which the test was inconclusive.

TABLE III. PERFORMANCE COMPARISON RANKING (P-VALUES)

	M2 vs M1	M2 vs M3	M3 vs M1
Down-rating 1%	8.7E-63	4.0E-40	1.8E-24
Down-rating 3.5%	1.8E-127	4.0E-75	8.1E-77
Down-rating 7%	5.8E-161	5.6E-70	2.7E-98
Down-rating 15%	1.2E-131	1.8E-31	2.0E-51
Icing V-Low 1%	5.3E-44	8.7E-05	4.0E-19
Icing Low 5%	3.4E-83	0.72	7.6E-64
Icing Medium 10%	7.3E-76	0.01	8.0E-45
Icing High 20%	5.2E-02	0.95	3.2E-08
Yaw Misalignment	1.7E-44	5.6E-95	9.8E-51
Acoustic Curtailment	2.20E-49	3.0E-02	4.1E-46

One of the causes of the under-performance of Method 1 is the lack of mitigation of operational and environmental variations. For Method 1, no effort was made to reduce the data dispersion caused by these variations. Method 3 performed better than Method 1 because of the use of air density as an input during the implementation. The consideration of wind density in modelling a WT's output power has been demonstrated to reduce the RMSE by 16% [26]. Method 2 performed better than the other two methods for most of the fault scenarios evaluated. With this method, data dispersion owing to environmental variation is reduced by introducing a density correction in its implementation. The operational variation was addressed by data translation within each wind bin [22].

Finally, Fig. 18 shows the mean PD_{10} value obtained by the three methods for each type of fault, regardless of the intensity of the fault. These results agree with the conclusions of the paired t-tests. The

mean PD_{10} values of Methods 2 and 3 are clearly higher than those of Method 1, regardless of the fault type. However, as shown in Fig. 18, although the t-test enabled us to conclude that the difference in means between Methods 2 and 3 is statistically significant for acoustic curtailment and yaw misalignment, this difference is significantly small.

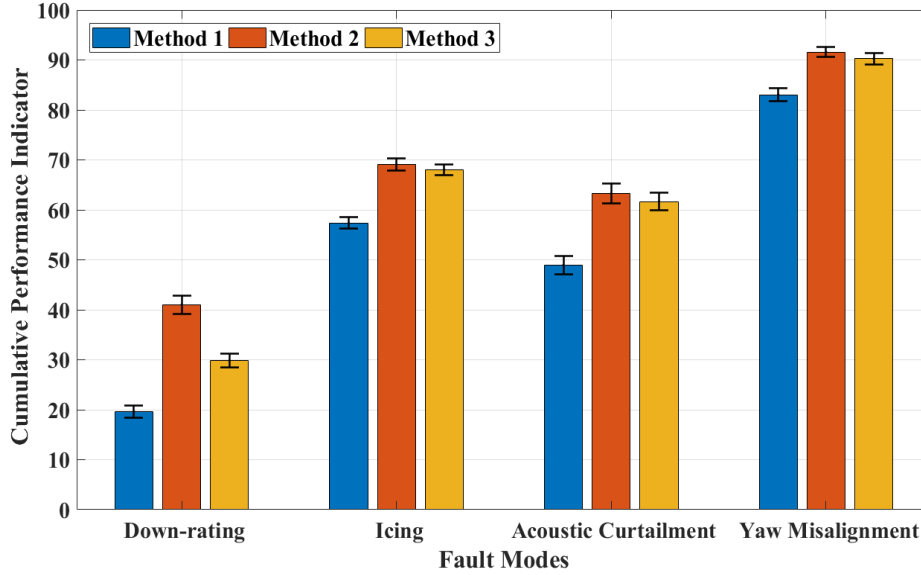


Fig. 18. Global performance indicators for all fault types (**down-rating, icing, acoustic curtailment, yaw misalignment**) for Methods 1, 2, and 3 with 95% confidence intervals.

4.2 Performance evaluation—Effects of environmental and operational variations

The proposed simulation framework enables the simulation of data flows generated by different wind and temperature profiles and produced by different WT models. This enables the effects of environmental and operational conditions on the detection performance of fault-detection methods to be evaluated.

4.2.1 Effect of environmental variation

The down-rating fault with an intensity level of 15% was selected for a meaningful analysis because it presented a fair opportunity for a comparative examination. The fault intensity level of 15% presented a borderline scenario of a sufficiently difficult detection problem for all three methods. Below this intensity, the fault signature was low such that detection became difficult, and for fault intensities above this value, the fault signatures began to be relatively easier to detect.

The rows of the PEM represent the PD_{10} computed with different environmental profiles (wind, temperature time series ($U(k)$, $T(k)$)) from different wind farms. The three PEMs obtained for a down-rating fault of 15% for each of the three fault-detection methods were appended horizontally. They are displayed in Fig. 19a as a scaled colour image for Methods 1, 2, and 3. This arrangement aims to visually evaluate the effect of environmental variations on the detection performance across all three methods. The images of the PEMs show significantly higher PD_{10} values for Wind Farm S for the 15% down-rating test scenario. This indicates that the site-specific wind and temperature profiles have a direct effect on performance detection for all methods.

To quantify this visual observation further, the mean PD_{10} values for each wind farm (for each method) were calculated. The results are presented in tabular form in Fig. 19b as mean values and their 95% confidence intervals. When compared with wind farms (V, L, D, and C), the mean detection performance for Farm S is approximately 52% higher on average for Method 1, 43% higher for Method 2, and 40% higher for Method 3.

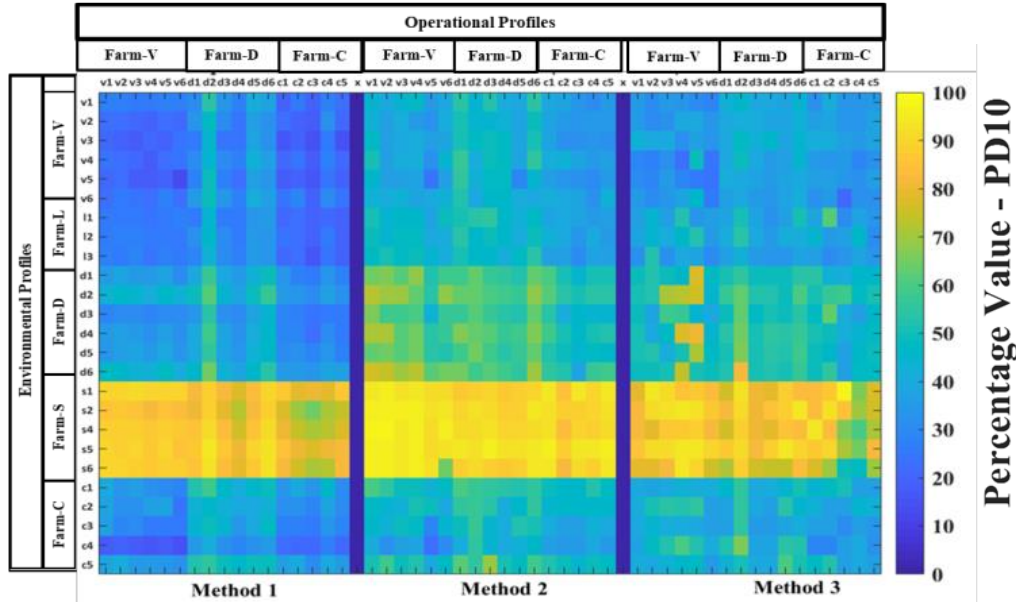


Fig. 19a. Image of performance evaluation matrices for Methods 1, 2, and 3 for a 15% down-rating.

Environmental Profiles	Method 1	Method 2	Method 3
Farm V	27 ± 1.4	41 ± 2.1	36 ± 1.5
Farm L	27 ± 1.4	44 ± 2.5	39 ± 2.5
Farm D	38 ± 1.3	58 ± 2.1	52 ± 2.1
Farm S	84 ± 1.2	92 ± 3.2	83 ± 2.1
Farm C	35 ± 1.6	47 ± 1.4	43 ± 1.4

Fig. 19b. Farm average and 95% confidence interval of the performance indicator (PD_{10}) for Methods 1, 2, and 3 for a 15% down-rating.

The effects of environmental variation on detection performance can be explained by examining the signature of the fault under observation and the distinct environmental profile of Farm S. Farm S is located in the south of France, where relatively higher wind speeds are experienced compared with the other four locations. Figs. 20 and 21 show the down-rating faults and the onsite wind distribution for all farms under observation, respectively. The distribution of wind at higher wind speeds (from 8 to 16 m/s (Fig. 21) is of interest here. As identified by the region between red dotted lines), a significantly higher number of wind speed samples can be observed for Farm S in this high wind region. The same wind region (from 8 to 16 m/s) is shown in Fig. 20. This shows that the effect of a 15% down-rating only becomes visible in the high-wind-speed regions (> 10 m/s) owing to the inherent nature of the fault signature. A larger number of high-wind-speed samples mean that during the fault period (year 3), the fault signature was excited more frequently for Farm S than for Farms V, L, D, and C. This explains the higher values of the detectability indicator PD_{10} for Farm S. The peculiar fault nature of down-rating

(fault excitation for high winds only) and the specific wind distributions (Fig. 21) explain the lower detection performance values for wind farms V, L, D, and C.

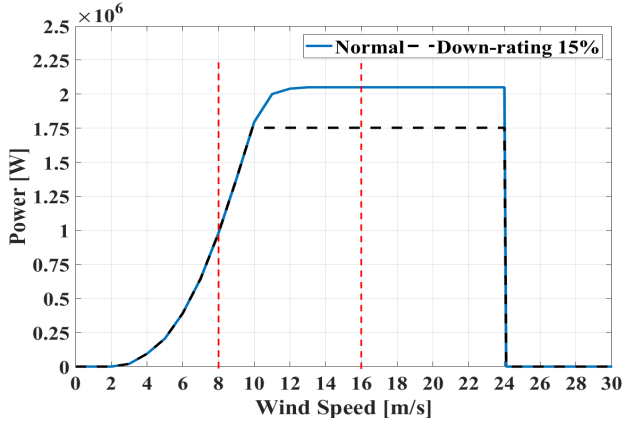


Fig. 20. Faulty power curves for down-rating (15%) with red dotted lines identifying region of interest (from 8 to 16 m/s).

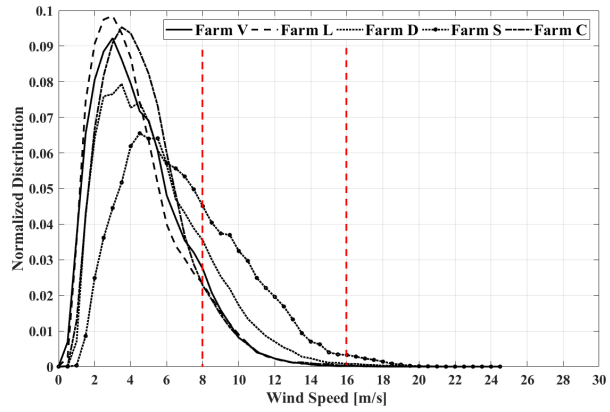


Fig. 21. Wind distribution of farms (Farms V, L, D, S, and C) with red dotted lines identifying the region of interest (from 8 to 16 m/s).

4.2.2 Effect of operational variation

Different turbines from different manufacturers can exhibit unique characteristic behaviours. The PEM provides the opportunity to observe and evaluate performance variations owing to the variability in these operational profiles. The columns of PEM represent different operational profiles (data dispersion and operational characteristics) from different manufacturers.

The icing fault is a good candidate for evaluating operational variations. As presented earlier, this fault type affects the overall behaviour of a turbine for all wind speeds. Hence, it is suitable for capturing the operational behaviour of WTs. The performance indicator PD_{10} for fault-type icing of 5% was calculated and the PEM was populated. The same process was repeated for all methods of interest, and the resulting PEMs were appended vertically. The selection of vertical concatenation ensured a better visual interpretation.

The icing fault with an intensity level of 5% was selected to calculate the performance indicators and perform meaningful analysis. The 5% fault intensity level for this type of fault presented a fair opportunity for comparative analysis. Below this intensity, the fault signature was low such that the detection became too difficult, and for fault intensities above this value, the fault signatures began to be relatively easier to detect.

Three wind farms were carefully selected for the operational analysis. Because Farms V and C have WTs from the same OEM, they became a natural candidate for the control group in this analysis. Wind Farm D was selected as the third candidate for comparison with the control group of similar turbines. Fig. 22 shows the performance indicator PD_{10} calculated for Methods 1 and 2.

For easier visual interpretation, the performance evaluation matrices are shown as an image with scaled colours for two methods of fault detection (Methods 1 and 2). The colour bar on the right-hand side shows a colour scale associated with the PD_{10} values, with higher values depicted as shades of yellow and lower values as blue. The image of the PEMs exhibits globally similar performance values for Wind Farms V and C, in contrast to Wind Farm D. This indicates that, similar to the environmental

profiles evaluated earlier, data dispersions, i.e. operational characteristics, also have a direct effect on the fault-detection performance.

To quantify this visual observation further, the mean PD_{10} values for each wind farm (and for each method) were calculated. The results are presented in a tabular form in Fig. 22b. When comparing Wind Farms V and C, the mean detection performance remains approximately within one standard deviation of each other. On the contrary, the mean detection values for Farm D was approximately 35% higher for Method 1 and 45% higher for Method 2 on average. This made the detection performance for Farm D twice as much as that for Farms V and C. The improved performance of Farm D can be associated with the specific operational performance of machines from this OEM. For Farm D, the data dispersion was minimal, and the observed power values closely followed the expected power curve.

Similar to the environmental variation analysis, the mean performance indicator (PD_{10}) values for Method 2, as presented in Fig. 22b, are relatively higher. The same is visually represented in Fig. 22a through visibly lighter shades for Method 2. This suggests that Method 2 performed globally better than Method 1 for this fault type. Quantitatively, the gain of Method 2 for the specific fault example is approximately 20%. This observation of performance gain for Method 2 corresponds to the conclusions drawn in the fault-specific performance evaluation discussed earlier.

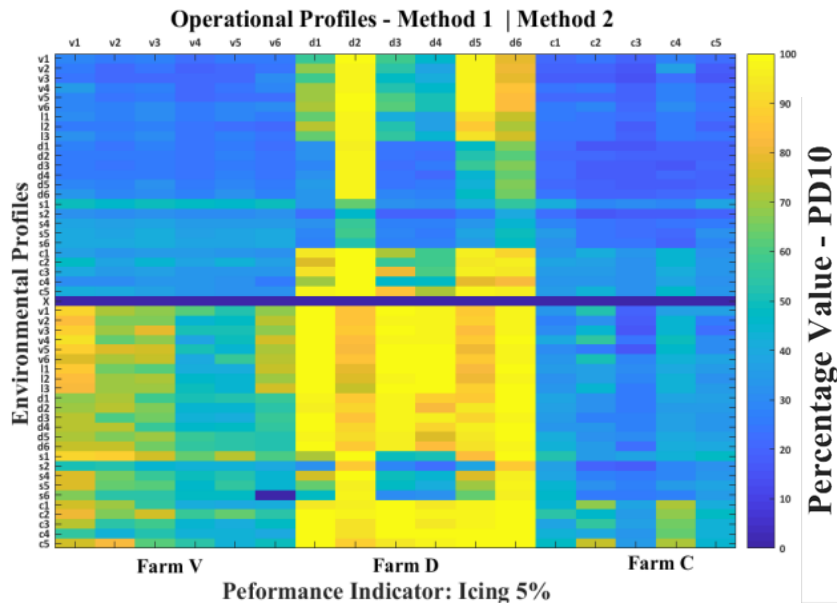


Fig. 22a. Image of performance evaluation matrices for Methods 1, 2, and 3 for a 15% down-rating.

	Farm V	Farm D	Farm C
Method 1	31 ± 1.2	62 ± 3.2	25 ± 1.2
Method 2	43 ± 2.2	89 ± 3.1	37 ± 2.0

Fig. 22b. Farm average and 95% confidence interval of PD_{10} for Methods 1, 2, and 3 for a 15% down-rating.

5 CONCLUSION

In this paper, we propose a simulation framework that realistically replicates the dispersion around a power curve to enable large sets of data flows in various environmental and operational conditions to be generated. This enables the performance of fault-detection algorithms to be compared and the key factors that may affect the performance of the algorithms to be identified. The framework used a large set of available data, recorded from five different French wind farms, composed of WTs of different models and located in different geographical areas and therefore subject to different environmental conditions. In total, 625 power profiles were generated.

The main finding of this research is the extent to which environmental and operational conditions affect the performance of the methods. For certain faults, we observed that our performance indicator could be increased by up to 50% when the power profiles were generated with the wind and temperature profiles of a particular wind farm compared with the others. This clearly shows that equitably comparing two fault-detection methods in the literature using a single profile of the power produced is impossible.

The proposed framework is fully robust and can be used to validate and analyse other methods and evaluate more fault signatures, provided that these methods use the power produced, the wind speed and temperature, and provided the fault affects the turbine power curve. For future research, the framework presented could be used to compare the performance of additional wind farm configurations (turbines in configurations such as square, diamond, scattered, etc.). Other phenomena of interest could include the analysis of detection performance as a function of the time of the year. The concept involves evaluating whether fault occurrence in the winter is different from fault occurrence in the summer. Additionally, the analysis of the effect of the duration of fault occurrence could be of interest.

In summary, the unique contribution of this study is that it demonstrated that the comparison of fault-detection methods based on a single example of a fault on a particular WT on a particular farm is insufficient to determine whether one method is superior to the other. However, most studies conducted over the last decade on WT fault detection have adopted this method. To prove our point, we used an extensive database of fault-free SCADA data recorded on WTs of different OEMs installed in various geographical sites. We created a realistic simulation framework and proposed a method of comparing the performances of different fault-detection methods. Thus, we demonstrated that there is great variability in the values of the performance indicators of a particular fault-detection method, even when applied to the same fault. We consider that these findings are of value to both academia and the industry and can aid the design of objective performance evaluation procedures for fault-detection methods for WTs.

ACKNOWLEDGEMENTS

This research was supported by ANRT (Association nationale de la recherche et de la technologie), France, through a collaboration between the industrial partners VALEMO, France and GIPSA-Lab, France under CIFRE CONVENTION N° 2016/0749.

REFERENCES

- [1] S. Koebrich, T. Tian, and E. Chen, “2017 Renewable Energy Data Book: Including Data and Trends for Energy Storage and Electric Vehicles,” p. 142.
- [2] P. J. Tavner, *Offshore wind turbines: reliability, availability and maintenance*. London, U.K: Institution of Engineering and Technology, 2012.
- [3] M. Lydia, S. S. Kumar, A. I. Selvakumar, and G. E. Prem Kumar, “A comprehensive review on wind turbine power curve modeling techniques,” *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 452–460, Feb. 2014, doi: 10.1016/j.rser.2013.10.030.
- [4] IEC 61400-12-1, “Power Performance Measurements of Electricity Producing Wind Turbines.” 2005.
- [5] W. Yang, R. Court, and J. Jiang, “Wind turbine condition monitoring by the approach of SCADA data analysis,” *Renewable Energy*, vol. 53, pp. 365–376, May 2013, doi: 10.1016/j.renene.2012.11.030.
- [6] R. J. de Andrade Vieira and M. A. Sanz-Bobi, “Power curve modelling of a wind turbine for monitoring its behaviour,” in *Renewable Energy Research and Applications (ICRERA), 2015 International Conference on*, 2015, pp. 1052–1057, Accessed: Apr. 07, 2017. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/7418571/>.
- [7] X. Jia, C. Jin, M. Buzza, W. Wang, and J. Lee, “Wind turbine performance degradation assessment based on a novel similarity metric for machine performance curves,” *Renewable Energy*, vol. 99, pp. 1191–1201, Dec. 2016, doi: 10.1016/j.renene.2016.08.018.
- [8] S.-Y. Kim, I.-H. Ra, and S.-H. Kim, “Design of wind turbine fault detection system based on performance curve,” in *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*, 2012, pp. 2033–2036, Accessed: Apr. 07, 2017. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/6505401/>.
- [9] F. Eljorje, S. Kim, and J. Lee, “A Wind Turbine Fault Detection Approach Based on Cluster Analysis and Frequent Pattern Mining,” *TIIS*, vol. 8, no. 2, pp. 664–677, 2014.
- [10] A. Kusiak, H. Zheng, and Z. Song, “Models for monitoring wind farm power,” *Renewable Energy*, vol. 34, no. 3, pp. 583–590, Mar. 2009, doi: 10.1016/j.renene.2008.05.032.
- [11] A. Kusiak and W. Li, “The prediction and diagnosis of wind turbine faults,” *Renewable Energy*, vol. 36, no. 1, pp. 16–23, Jan. 2011, doi: 10.1016/j.renene.2010.05.014.
- [12] M. Schlechtingen and I. F. Santos, “Using Data-Mining Approaches for Wind Turbine Power Curve Monitoring: A Comparative Study,” *IEEE TRANSACTIONS ON SUSTAINABLE ENERGY*, vol. 4, no. 3, p. 9, 2013.
- [13] S. Butler, J. Ringwood, and F. O’Connor, “Exploiting SCADA system data for wind turbine performance monitoring,” in *Control and fault-tolerant systems (systol), 2013 conference on*, 2013, pp. 389–394, Accessed: Apr. 10, 2017. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/6693951/>.
- [14] R. Pandit and D. Infield, “Gaussian Process Operational Curves for Wind Turbine Condition Monitoring,” *Energies*, vol. 11, no. 7, p. 1631, Jun. 2018, doi: 10.3390/en11071631.
- [15] T. J. Rogers *et al.*, “Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian Process regression,” *Renewable Energy*, vol. 148, pp. 1124–1136, Apr. 2020, doi: 10.1016/j.renene.2019.09.145.
- [16] F. Pelletier, C. Masson, and A. Tahan, “Wind turbine power curve modelling using artificial neural network,” *Renewable Energy*, vol. 89, pp. 207–214, Apr. 2016, doi: 10.1016/j.renene.2015.11.065.
- [17] A. Stetco *et al.*, “Machine learning methods for wind turbine condition monitoring: A review,” *Renewable Energy*, vol. 133, pp. 620–635, Apr. 2019, doi: 10.1016/j.renene.2018.10.047.
- [18] Y. Wang, Q. Hu, L. Li, A. M. Foley, and D. Srinivasan, “Approaches to wind power curve modeling: A review and discussion,” *Renewable and Sustainable Energy Reviews*, vol. 116, p. 109422, Dec. 2019, doi: 10.1016/j.rser.2019.109422.
- [19] R. Bi, C. Zhou, and D. M. Hepburn, “Detection and classification of faults in pitch-regulated wind turbine generators using normal behaviour models based on performance curves,” *Renewable Energy*, vol. 105, pp. 674–688, May 2017, doi: 10.1016/j.renene.2016.12.075.
- [20] A. Kusiak and A. Verma, “Monitoring Wind Farms With Performance Curves,” *IEEE Transactions on Sustainable Energy*, vol. 4, no. 1, pp. 192–199, Jan. 2013, doi: 10.1109/TSTE.2012.2212470.
- [21] J.-Y. Park, J.-K. Lee, K.-Y. Oh, and J.-S. Lee, “Development of a Novel Power Curve Monitoring Method for Wind Turbines and Its Field Tests,” *IEEE Transactions on Energy Conversion*, vol. 29, no. 1, pp. 119–128, Mar. 2014, doi: 10.1109/TEC.2013.2294893.
- [22] P. Cambron, R. Lepvrier, C. Masson, A. Tahan, and F. Pelletier, “Power curve monitoring using weighted moving average control charts,” *Renewable Energy*, vol. 94, pp. 126–135, Aug. 2016, doi: 10.1016/j.renene.2016.03.031.
- [23] O. Uluyol, G. Parthasarathy, W. Foslien, and K. Kim, “Power curve analytic for wind turbine performance monitoring and prognostics,” in *Annual conference of the prognostics and health management society*, 2011, vol. 2, pp. 1–8, Accessed: Apr. 07, 2017. [Online]. Available: http://72.27.231.73/sites/phmsociety.org/files/phm_submission/2011/phmc_11_049.pdf.

- [24] U. Aziz, S. Charbonnier, C. Bérenguer, A. Lebranchu, and F. Prevost, "Simulation of wind turbine faulty production profiles and performance assessment of fault monitoring methods," p. 8, 2018.
- [25] U. Aziz, S. Charbonnier, C. Berenguer, A. Lebranchu, and F. Prevost, "SCADA data based realistic simulation framework to evaluate environmental impact on performance of wind turbine condition monitoring systems," in *2019 4th Conference on Control and Fault Tolerant Systems (SysTol)*, Casablanca, Morocco, Sep. 2019, pp. 360–365, doi: 10.1109/SYSTOL.2019.8864769.
- [26] Z. Farkas, "Considering air density in wind power production," *arXiv preprint arXiv:1103.2198*, 2011, Accessed: Apr. 10, 2017. [Online]. Available: <https://arxiv.org/abs/1103.2198>.
- [27] A. Lebranchu, S. Charbonnier, C. Bérenguer, and F. Prévost, "A combined mono- and multi-turbine approach for fault indicator synthesis and wind turbine monitoring using SCADA data," *ISA Transactions*, vol. 87, pp. 272–281, Apr. 2019, doi: 10.1016/j.isatra.2018.11.041.
- [28] H. L. Van Trees, *Detection, estimation, and modulation theory. Part I, Part I.*, New York; Chichester: Wiley, 2001.