applied sciences

MDPI

Article

A Robust Wind Turbine Component Health Status Indicator

Roberto Lázaro [1,2], Julio J. Melero [1,*] and Nurseda Y. Yürüşen [1]

[1] Instituto Universitario de Investigación Mixto de la Energía y Eficiencia de los Recursos de Aragón ENERGAIA, Universidad de Zaragoza, Campus Río Ebro, Ed. CIRCE, Mariano Esquillor Gómez 15, 50018 Zaragoza, Spain; rlazaro@fcirce.es (R.L.); nursedayildirim@gmail.com (N.Y.Y.)

[2] CIRCE Centro Tecnológico, Parque Empresarial Dinamiza, Avenida Ranillas, Edificio 3D, Planta 1, 50018 Zaragoza, Spain

* Correspondence: melero@unizar.es

Abstract: Wind turbine components' failure prognosis allows wind farm owners to apply predictive maintenance techniques to their fleets. Determining the health status of a turbine's component typically requires verifying many variables that should be monitored simultaneously. The scope of this study is the selection of the more relevant variables and the generation of a health status indicator (Failure Index) to be considered as a decision criterion in Operation and Maintenance activities. The proposed methodology is based on Gaussian Mixture Copula Models (GMCMs) combined with a smoothing method (Cubic spline smoothing) to define a component's health index based on the previous behavior and relationships between the considered variables. The GMCM allows for determining the component's status in a multivariate environment, providing the selected variables' joint probability and obtaining an easy-to-track univariate health status indicator. When the health of a component is degrading, anomalous behavior becomes apparent in certain Supervisory Control and Data Acquisition (SCADA) signals. By monitoring these SCADA signals using this indicator, the proposed anomaly detection method could capture the deviations from the healthy working state. The resulting indicator shows whether any failure is likely to occur in a wind turbine component and would aid in a preventive intervention scheduling.

Keywords: wind turbine; Gaussian Mixture Copula models; failure index; health status indicator; cubic spline smoothing

Citation: Lázaro, R.; Melero, J.J.; Yürüşen, N.Y. A Robust Wind Turbine Component Health Status Indicator. *Appl. Sci.* **2024**, *14*, 7256. https://doi.org/10.3390/app14167256

Academic Editor: Francesca Scargiali

Received: 22 July 2024
Revised: 6 August 2024
Accepted: 15 August 2024
Published: 17 August 2024

1. Introduction

Wind energy is one of the fastest-growing energy sources. Both onshore and offshore wind technologies pose Operation and Maintenance (O&M) challenges for O&M decision-makers such as limited site accessibility [1], adverse and dynamic operating conditions [2] and the aging of the wind turbine fleet [3].

Lack of proper management under these challenging circumstances results in severe penalties such as major component failures, long unavailability and significant revenue losses. To avoid severe consequences, there exists a growing interest in asset digitalization and timely status tracking of wind power plants [4–6]. This interest results in new data acquisition opportunities, new data sources and the tendency to move from data silos to data sharing. The available data must be analyzed and interpreted smartly, which is highly beneficial to decision-makers. Then, smart management of O&M activities is relevant to enhancing the economic balance of a wind energy power plant. Furthermore, the success of maintenance task optimization models relies on accurate estimation of the component's remaining useful life (RUL), the impact of component degradation on system efficiency, and the accuracy of the prognosis model itself ([7,8]). Therefore, the use of health indices that can be easily monitored over time can enhance the precision of these models and facilitate the field maintenance personnel's tasks in the power plant.

Smart management of O&M activities involves planning predictive and preventative interventions, evaluation of assets performance and effective condition monitoring. Compact decision support tools are needed to transform such complex decisions into a manageable form and make consistent decisions [9].

The decision-maker responsible for managing wind farm O&M activities (wind power plant owner or O&M engineer) needs to gain in-depth knowledge of the machine and its components' health status. It is essential to perform adequate maintenance planning that maximizes asset reliability. The decision support tools must help with these needs and also provide helpful and easy-to-track indicators of anomalous working states promptly [10,11].

Different malfunction and anomaly detection techniques can be implemented in a wind turbine O&M decision support tool. The majority of the condition monitoring techniques use Supervisory Control and Data Acquisition (SCADA) signals or Condition Monitoring System (CMS) measurements as inputs. A wind turbine is a complex system with many components & sub-components. In wind turbines, the condition monitoring systems target major components such as electric and control systems, generators, hubs, blades and gearboxes, since these components are classified as critical assemblies [12]. The frequency of major wind turbine gearbox failures is generally low. However, these failures are associated with severe consequences [13].

Timely evaluation of CMS and SCADA data trends can aid decision-making on whether to inspect, repair, or replace the wind turbine component under examination [14]. There are also tools for analyzing the remaining useful lifetime of the wind turbine component. These tools use the maintenance and failure records of the component. Nevertheless, many challenges exist in conveying data sets and deriving results from such analyses. On the one hand, existing methods that use only CMS and SCADA data are limited to aid only short-term maintenance decisions. On the other hand, methods based on statistical distributions, which use only maintenance and failure records as input data, cannot provide fine temporal failure occurrence estimations. Moreover, these methods require waiting until a reasonable amount of failures have occurred before obtaining robust findings due to the input failure data requirement of the models [15]. Therefore, new methods are needed to integrate all forms of maintenance, failures and operational data to support the scheduling of component intervention associated with decision making [16,17].

Condition monitoring covers fault monitoring activities that can be performed using either signal-based or model-based techniques ([18]). In signal-based methods, in general, patterns are considered and the statistics of the signals are compared to threshold values. Then, the deviation between the thresholds and measurements is used as an indicator of abnormal behavior. In the model-based techniques, the model simulates the behavior of the component, and then the estimation of the signal is compared to real measurements. The resulting difference from this comparison is used as an indicator to detect the abnormal working of the investigated asset.

The most straightforward approach to condition monitoring is asset status tracking via control charts ([19]), developed to monitor target variables easily. Generally, control charts are based on statistical parameters, through which the average historical behavior of the signal and significant intervals is defined. Via control charts, it is possible to detect normal and abnormal working trends and patterns of a signal. However, conventional control charts have some challenging prerequisites to fulfill, such that the collected data be independent and normally distributed. Therefore, control charts generate many false alarms under real operating conditions [20]. One of the solutions for these false alarms is to remove the auto-correlation which exists in input data. To do that, Exponentially Weighted Moving average techniques (univariate, EWMA, and multivariate, MEWMA) were implemented in the control chart design process [21]. The EWMA control charts were not able to detect the faults for wind turbines, while in comparison, the MEWMA control charts could detect the faults with a half an hour lead time window in [20]. The better performance of the MEWMA, compared to the EWMA, was associated with the multivariate nature of wind turbine faults. Nevertheless, a half an hour lead time window

is very limited to schedule and/or perform maintenance activities. From the decision maker's perspective, this means simultaneously monitoring several signals together with the result that is not worth it.

Therefore, there is room for improvement in existing anomaly detection techniques for wind turbines, which are multicomponent and multivariate entities, as the existing methods lack the consideration of real working conditions and adequate integration of available SCADA signals into the anomaly indicators [22]. The applicability of the existing models to the condition monitoring data is limited due to their short lead times between the alert status and the failure time. Furthermore, wind turbine faults require the analysis of multivariate systems, so selecting the variables to be considered and their relationships is essential. The selection of these variables is crucial for better model performance, but also for a better understanding of their operation and avoiding black-box models that are difficult to interpret physically.

Depending on the component under investigation, the signals considered in model-based condition monitoring must be revised and selected carefully.

In particular, oil temperature, rotational speed, and power signals obtained from SCADA and vibration and oil debris obtained from CMS are relevant variables to detect gearbox failures [23]. Notably, vibrations require special attention, evidenced by the comprehensive review conducted in [24] or the promising results presented in [25] based on generalized multiscale Poincare plots (GMPOP) and support vector data description (SVDD). Also, in the literature oil pressure was considered as an input variable for the gearbox failure detection. In the case of the generator, variables such as generator speed, active power, or reactive power are variables that can be studied for fault detection. In the case of the blades, other variables such as wind speed, rotor speed, or the miscorrelations between wind speed and torque shaft are parameters of interest for this purpose [26].

In brief, O&M engineers need to simultaneously monitor a wide range of variables to assess the health status level of different wind turbine components. Therefore, easy-to-track summary indicators are required. These indicators can be considered as a decision criterion for intervention planning, component repair and asset remaining lifetime evaluations [10].

Furthermore, a good prognosis of the wind power plant and adapted management of the maintenance to the specific working conditions will revert very positively to the economic balance of each project. The new methodologies, which aim for "Condition Monitoring" and predictive maintenance of the assets, are becoming more promising [27–29].

Copula models are proposed in [30,31] for power curve performance analysis with applications in condition monitoring. In [32] the Gaussian Mixture Copula Model (GMCM) is proposed as a new modeling tool for power curves. The results are promising in modeling and outlier detection in power curves. It is expected that additional variables will be included to obtain a more accurate model in the future. One of the first advanced approaches in wind turbines from a multivariate environment [33], obtains a power curve multivariate model based on conditional copulas. Reference [34] explores the multivariate analysis considering generator speed, wind speed and power in multivariate models for wind turbine health monitoring. This kind of tool allows estimating the joint probability density function of considered variables in a non-parametric way.

These models can satisfy the need for a methodology to characterize the relationship between critical variables in a multivariate environment. They can be applied to power curves, and components characterization and they are suitable to be applied on preventive detection of failures.

As a summary and to support the final motivation and justification for the selection of the proposed methodology, Figure 1 and Table 1 are shown. Figure 1 illustrates the typical data used in condition monitoring, as well as the most commonly employed models. It highlights the scope of this paper, which focuses on SCADA data and probabilistic models. Table 1 presents a representative selection of the prevailing trends and dominant models in fault detection and diagnostics within wind turbine components, based on SCADA data, with applications in condition-based maintenance. SCADA data may be

a preferable choice for condition monitoring, over well-tested techniques like vibration-based, due to its availability and low cost of exploitation, providing a comprehensive view of equipment performance. Leveraging existing infrastructure, SCADA enables effective anomaly detection and predictive maintenance, enhancing operational efficiency and minimizing downtime.

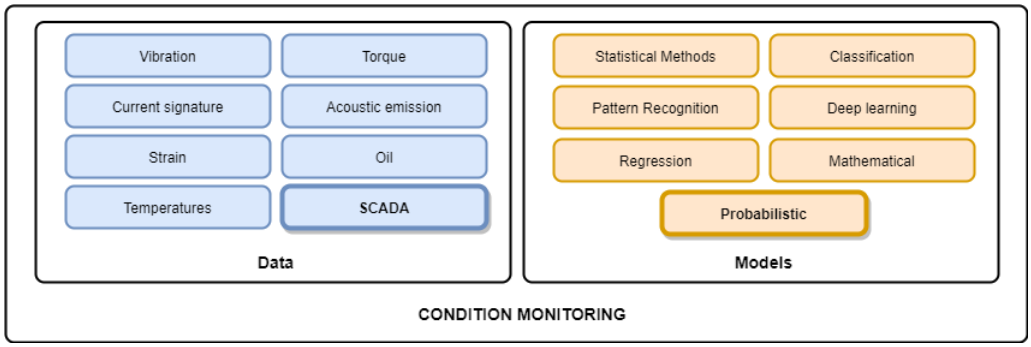


Figure 1. Condition Monitoring. Common Data & Models.

Table 1. Summary main trends in fault detection and diagnostics in wind turbines.

Ref	Models	Applications on	Main Contribution
[19]	EWMA	tower, yaw, gearbox	Simple and robust. Only on large wind farms
[20]	EWMA, MEWMA	Power output deviation, generator	Autocorrelation is challenging. Better performance when data is independent and normally distributed. Multivariate model is the best. Limited lead time window.
[21]	EWMA, MEWMA	Power Curve monitoring	Robust detection of underperformance
[22]	LSTM-SDAE XGBoost	Generator, pitch motor, gearbox, anemometer	Average classification accuracy 92%. Black box, not comprehensible
[26]	ARIMA	Gearbox	Trend propagation of the oil pump outlet pressure. Good trend predictor. Applied to one variable and one failure mode
[33,34]	Copula	Power curve	Good performance in outliers rejection and multivariate power curve
[35]	MDAE-DTAD (deep autoencoder)	generator	Multivariate model. An in-depth performance comparison is made with other similar models, Semi-supervised anomaly detection models.
[36]	TWSVM	gearbox	Better performance than standard classifiers (SVM, KNN, MLPNN, DT)
[37]	DT	wind turbine structures	Root cause of excessive vibrations applied to big data-based applications

As extracted from the consulted bibliography, Table 1, and other comprehensive and recent reviews in the field ([18,27,38]), most references on wind turbine fault detection methods are based on condition monitoring techniques, such as signal-based or model-based approaches. Furthermore, the main models used in CM in recent years appear to have focused on data-driven classification and regression models. There is also a noticeable interest in other probabilistic and deep learning models.

The proposed method can be considered as a combination of both approaches, leveraging the strengths of signal-based and model-based methods. It utilizes a model-based approach by employing the GMCM model to define the wind turbine component's behavior. This model is trained using selected variables identified through a feature selection process and captures statistical relationships and dependencies among them. Additionally, it incorporates a signal-based approach by using signals obtained from the wind turbine, which can be historical or real-time data. These signals capture dynamic and changing

information of the system and serve as input for the copula model. Consequently, a univariate signal is generated, processed, and monitored. By analyzing and processing these signals alongside the GMCM model, the joint behavior of the variables can be evaluated, and the joint probability of anomalous events or behaviors can be determined.

Moreover, as stated in the existing literature, there has been limited utilization of probabilistic models, such as copulas, specifically the GMCM model, in the context of fault detection in wind turbine components. The GMCM model enables the generation of a univariate output parameter, which facilitates the transition from condition monitoring to maintenance decision-making. This capability allows for the simultaneous monitoring of multiple variables, thereby simplifying the maintenance planning decision system by minimizing the number of health indices to be monitored (ideally, one per component).

For all the aforementioned reasons, the Gaussian Mixture Copula Model (GMCM) is chosen as a solution to address the need for flexibility when working with multivariate systems and to overcome and relax the requirement of Gaussian data assumption, which is commonly imposed by many models in this field. This study aims to develop and validate a holistic methodology for generating a Failure Index (*FI*) focused on early fault detection in major wind turbine components.

To conclude and highlight the main contributions of this paper to the wind energy industry and to previous work with GMCM models, this study advances the application of the Gaussian Mixture Copula Model (GMCM) by extending its use beyond the traditional power curve modeling and outlier detection with limited input variables (primarily wind speed and power). Our approach applies the GMCM to various wind turbine components, including the gearbox, generator, and blades, which represent a broader range of entities compared to previous studies. We have developed a comprehensive variable selection methodology to identify the most influential variables for representing the health of each component, utilizing up to seven variables in the final GMCM models (and many more during the variable selection process). This enhancement improves the flexibility and scalability of the GMCM method.

The main contributions of this work are as follows:

- **Enhanced GMCM Application:** We extend the GMCM method to a variety of wind turbine components, providing a more versatile and comprehensive modeling approach.
- **Scalability Improvements:** By incorporating multiple variables through an advanced selection process, the methodology ensures scalability and robustness in fault detection across different wind turbine components.
- **Holistic Health Index Generation:** We propose a holistic GMCM-based methodology that generates a user-friendly and easily monitorable health index for the main systems of a wind turbine, facilitating effective maintenance decision-making.
- **Early Fault Detection:** The developed Failure Index (*FI*) focuses on the early detection of faults in major wind turbine components, enhancing operational efficiency and reducing downtime.

The paper is structured in the order that follows. In Section 2, selected SCADA signals, case study information, and details of learning and test data sets are introduced. It also covers the mathematical foundations of the employed methodologies. The data filtering process and feature selection are discussed, and the applied GMCM model's optimization process is explained. The presentation of the proposed framework complements this section. In Section 3, the resulting *FI* signals from the studied cases are discussed. Finally, a summary of the contributions is given in Section 4.

2. Materials and Methods

In this section, algorithms and models used to develop a failure index for the condition monitoring of wind turbine components and the proposed technique are summarized. Generators, blades and gearboxes have been tested with the proposed methodology. In Section 2.1 the description of the data considered in this study is provided. In Section 2.2,

the overall proposed methodology is shown as a combination of the techniques explained in the following subsections. A data filtering task is needed to ensure that the input observations correspond to the characteristics of the wind turbine's healthy or faulty state. Therefore, in Section 2.3, the filtering procedure and its dependencies are described. Then, the relevant feature selection techniques are discussed in Section 2.4. Subsequently, the correlation analysis is introduced, which is used to investigate associations between the signals. The next step is to develop a GMCM method and optimize its coefficients. In Section 2.5, the underlying theory of the GMCM models and the advantages of the copula models are portrayed. The process of converting the resulting log-likelihoods obtained from the GMCM models into a failure index is explained. The final technique that needs to be referred to is the signal smoothing method, and the details are shown in Section 2.6.

2.1. Data

Data were obtained from two different sites in Spain. In Site 1, there are four wind farms with 33 wind turbines each and a nominal power of 950 kW per wind turbine. In Site 2, there is one wind farm comprising 12 wind turbines with a nominal power of 3000 kW each; 10 min SCADA data, together with failure and maintenance logbooks, were used in the study.

SCADA signals had to be cleaned to make them usable. Then, each failure was assigned to the respective wind turbine (WT) component with the help of the SCADA alarms and logbooks. The method has been previously tested in site 1, where the information related to reported failures was available in different components (generator, blades and gearbox). Later, in site 2 the study focuses on gearbox failures due to the high number of reported failures of this component in the available period of this site.

Inspections and regular service work periods were excluded from the database. Available data were divided into two datasets corresponding to the learning and test periods. Then, three failure cases and nine healthy cases (generator, blades and gearbox) were modeled in site 1, and eight cases were modeled in site 2, two of them containing a gearbox failure and six without failure.

The optimal splitting of the data into the training and the test periods for supervised learning techniques ranges from 60% to 80% for the training period (being the remaining data for the test period) [39,40]. Therefore, approximately six months of data were used for learning the model and, approximately two months of data were used for testing it.

Maintenance interventions and failure history have been documented as unstructured comments in spreadsheets. A simplified summary of the maintenance history excluding routine services is provided in Table 2.

Table 2. Maintenance logbook events.

Site	Wind Turbine	Start Date	End Date	Event
1	S1_GEN_F1	1 July 2015	_	Generator failure
1	S1_BLD_F1	4 December 2014	_	Blades failure
1	S1_GBX_F1	22 May 2012	_	Gearbox failure
2	S2_GBX_F1	7 May 2014	31 July 2015	Gearbox failure
2	S2_GBX_F2	5 June 2014	1 August 2014	Gearbox oil pump pressure failure

On site 1, S1_GEN_F1, S1_BLD_F1 and S1_GBX_F1 are analyzed as generator failure, blade failure and gearbox failure cases. In addition, S1_GEN_H1, S1_GEN_H2, S1_GEN_H3, S1_BLD_H1, S1_BLD_H2, S1_BLD_H3, S1_GBX_H1, S1_GBX_H2 y S1_GBX_H3 are also studied as healthy cases for comparison.

On site 2, S2_GBX_F1 and S2_GBX_F2 are analyzed as gearbox failure cases. Similarly to site 1, six cases without failure are also studied as S2_GBX_H1, S2_GBX_H2, S2_GBX_H3, S2_GBX_H4, S2_GBX_H5, S2_GBX_H6.

Between the two sites, five failure cases and 15 healthy cases have been analyzed. This results in a ratio of three healthy cases for each failure case analyzed.

In the case of site 1, the number of available SCADA signals is limited to 16 variables and all will be considered in the study. Whereas, in the case of site 2, SCADA data provide a total of 97 signals. Taking into consideration the component to be studied and literature-based recommendations [41], 13 signals were selected to model gearbox healthy and faulty working behaviors. Table 3 and Table 4 show the list of variables of site 1 and site 2, respectively.

Table 3. List of signals of Site 1 (all variables).

Signal	Abbreviation	Unit
Gearbox bearing temperature 1	BT1	°C
Gearbox bearing temperature 2	BT2	°C
Reactive Power	Q	KVAr
Voltage Phase (1,2,3) X	$V_{l(1,2,3)}$	V
Current Intensity Phase (1,2,3) X	$I_{l(1,2,3)}$	A
Large generator temperature	GlT	°C
Small generator temperature	GsT	°C
Torque	M	Nm
Gearbox oil tank temperature	OTT	°C
Ambient temperature	AT	°C
Rotor speed	Rw	rpm
Generator speed	Gw	rpm
Power	Pow	kW
Wind speed Average	V	m/s
Wind speed Maximum	Vmax	m/s
Wind speed Deviation	Vvar	m/s

Table 4. List of signals of Site 2 (selected).

Signal	Abbreviation	Unit
Gearbox bearing temperature 1	BT1	°C
Gearbox bearing temperature 2	BT2	°C
Gearbox bearing temperature 3	BT3	°C
Power	Pow	kW
Ambient temperature	AT	°C
Gearbox oil input pressure	OIP	Bar
Gearbox oil mechanical pressure	OMP	Bar
Gearbox oil electrical pressure	OEP	Bar
Rotor speed	Rw	rpm
Generator speed	Gw	rpm
Gearbox oil inlet temperature	OIT	°C
Gearbox oil tank temperature	OTT	°C
Wind speed Average	V	m/s

In the end, a representative data sample has been considered. Later, data cleaning was applied to the signals and registration errors and duplicated data were removed. In general, the data availability values are very high after this process, exceeding 95% in most cases (except for S1_GBX_H1, which is around 93%), making it a representative data sample.

2.2. Flowchart of the FI Generation Process

The conceptual flowchart of the proposed methodology is given in Figure 2.

In this analysis, 20 cases are investigated to develop a functional failure index and test its performance. With the 12 cases of site 1 and the available variables, the model's performance has been analyzed for different components (generator, blades and gearbox). The best results have been observed in the case of the gearbox. To confirm the resulting performance in this component, a specific analysis of the gearboxes has been carried out at site 2 where eight cases were analyzed using the same methodology shown in the flowchart given in Figure 2. The methodology is the same in site 1 and in site 2. However, in site 1 it is not as critical as in site 2 because of the limited number of available variables that can affect each component.

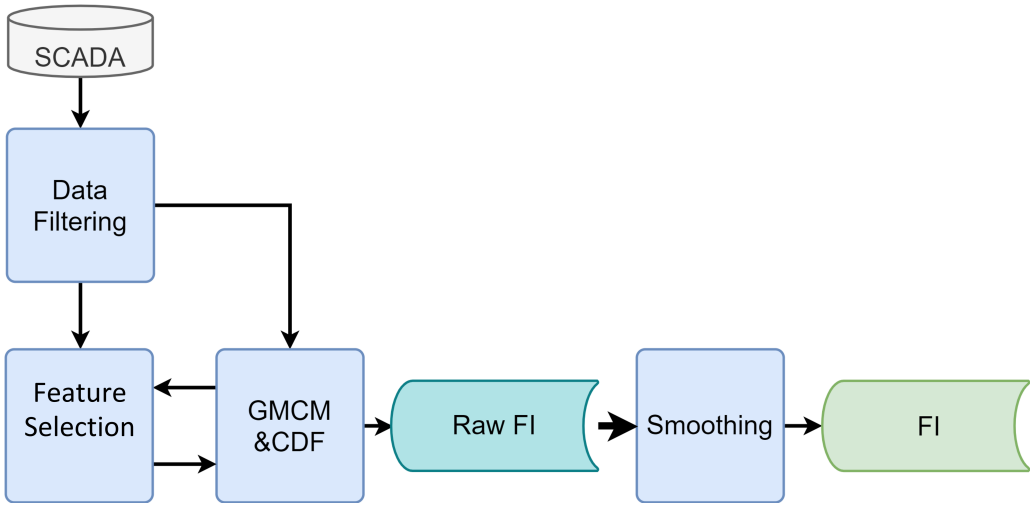


Figure 2. Diagram. Health Status Indicator Process.

First, signals are reviewed considering literature recommendations for wind turbine component failure modes [41]. Then, all available signals have been considered at site 1. They have been grouped by component by considering 10 variables related to the blade component, nine variables related to the generator and seven related to the gearbox. On the other hand, 13 signals related to the gearbox are selected on site 2.

In order to ensure that the input data in the learning period represents the healthy period of the wind turbine, the following procedure was performed. Duplicate data and zero power values are removed. Later, power values outside the normal operating range are filtered. For this purpose, data associated with lower or higher production than expected are removed using the Mahalanobis distance [42,43] between the measured power series and the estimated power for the measured wind speed considering the manufacturer's power curve.

Next, filtered input data of the SCADA signals are split as learning and test inputs for all wind turbines. Learning data are used in the GMCM and the resulting output from the GMCM (Log-likelihood density) is transformed as the inverse of the cumulative Log-Likelihood density [44] to generate the self-defined failure index (*FI*).

Then, the feature selection process is conducted. Indeed, this feature selection is applied by component and by site. This process aims at reducing the number of variables to feed the final model. Using Spearman correlation, the generated *FI* and the regressor variables are analyzed in both sites to evaluate their influence on the generated *FI* series. Then, a backward and forward process is started where a pre-selection criterion is used to assess the overall importance of the variables. As a result, considering the Spearman correlation index and the attempt to homogenize the final models, a GMCM model based on seven variables is established in the learning period and applied to all the studied cases. The seven variables selected as the most influencing for each component model and for each site are listed in Section 2.4. Later, the GMCM coefficients are saved and then used with the test data to evaluate the FI's performance for both sites. Finally, generated failure index series are smoothed using cubic spline methodology [45] to provide a clean summary signal to the decision-maker.

2.3. Filtering Process

One of the strengths of this study is the compatibility with the transition to the digitalization of assets. The flexibility of the methods to the uncertain quality of the input data is a remarkable feature in the models' behavior. For this reason, to verify the robustness and flexibility of the method, no complex filtering algorithms have been applied to the study variables. However, this task is needed to ensure a minimum quality of input data to generate the model from the learning period. Among the information available for this study, there were no SCADA alarms. In this sense, only the main information gaps, the main outliers identified in the power curve and the periods in which the wind turbine is stopped have been removed.

For this purpose, the Mahalanobis Distance (MD) has been considered. MD can determine the similarity between two vectors, taking into account distances and the correlations between variables. This value has no unit of measurement and is scale-invariant. According to this method, two power series have been compared, the measured power and the power associated with the wind speed measured according to the manufacturer's curve. Filtering against the manufacturer's power curve allows for identifying the normal mode behavior. As a result, all the observations corresponding to events where the wind turbine is stopped or related to a power value different from expected will be excluded from the set of learning data.

In this process, since we are dealing with only two variables (real power output and ideal power output), the covariance matrix is a 2×2 matrix and is guaranteed to be of full rank given the imperfect correlation between the variables. This ensures that the Mahalanobis distance can be effectively applied for outlier detection. Although the distributions of these variables are not perfectly normal, the method remains robust for detecting outliers in this specific application.

The results of the applied methodology are shown in Figure 3 where power, mechanical pump pressure, bearing temperature and generator speed are plotted against wind speed. It can be seen how the main outliers are identified and removed from the set of learning data. The filtering methodology outlined above has been carried out on all the analyzed wind turbines.

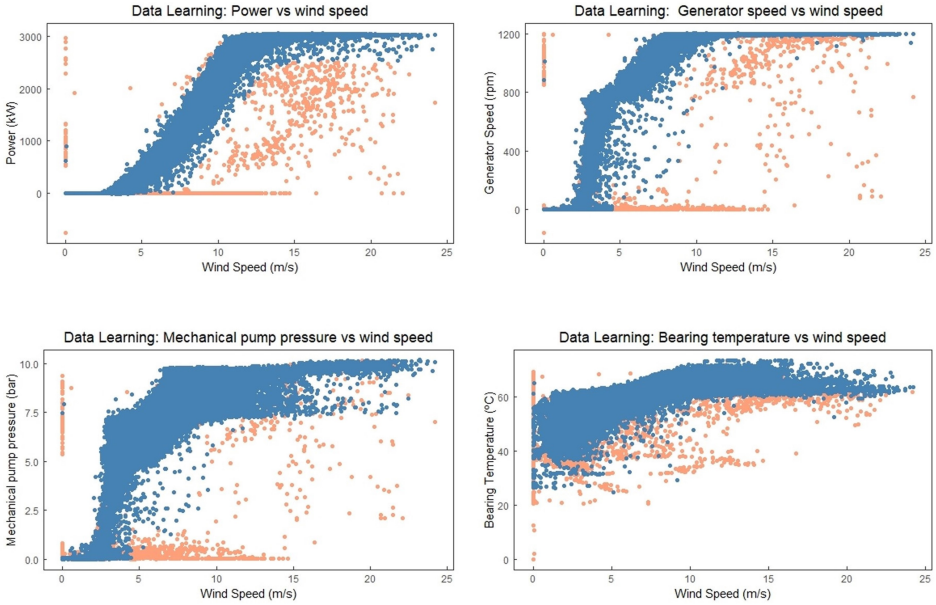


Figure 3. Accepted (blue) and Rejected (orange) learning data.

2.4. Feature Selection

Different variables were selected to monitor the generator, blades, and gearbox based on the domain expertise, data availability, and variables' physical meanings. In the case of site 1, 16 signals were available to monitor the components and 97 signals in the case of site 2. In the first stage, the behavior of the GMCM model was tested in three wind turbine components (generator, blades and gearbox) in site 1, where the availability of variables was limited. Once a potential failure detection capacity was observed in the gearbox components of the turbines from site 1, the same procedure was applied in site 2 focused on gearbox failure detection.

In order to make a selection of the most interesting set of variables to assess the health of the component, the available variables, their relationship and the experience from the literature review were taken into consideration [41]. In this sense, in site 1, 10 variables have been considered to analyze the blade's health status, nine variables in the case of the generator, and seven variables were available to track the gearbox's health status.

In site 2, the number of available signals for the gearbox analysis is higher. Therefore, some additional issues about the variables' physical meanings have been taken into account. In this respect, it was considered that the oil affects the lubrication film directly, and thanks to this, the protection of moving parts can be achieved by reducing the friction in the gearbox. Therefore, oil has to be monitored too, and oil conditions can be monitored by temperature, dielectric constant or viscosity.

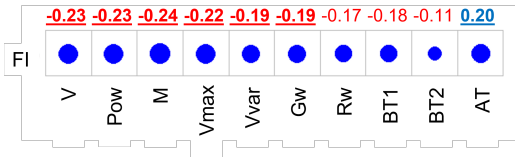
As highlighted in literature [46], there are already extensively developed analyses and procedures considering oil status for gearbox health condition monitoring. Therefore, the working conditions of the gearbox and oil status have to be taken into consideration in the gearbox health status analysis. Focusing on viscosity is a very relevant characteristic to be considered for the proper functioning of the oil. It is known that oil temperature and oil pressure directly affect the viscosity [47]. Therefore, oil pressure and temperature could be good candidates for consideration as sensitive variables when analyzing the health of a gearbox.

Hence, according to these considerations and based on the literature review and available signals, 13 variables have been selected in the first stage as significant enough to be considered in a gearbox health status tracking in site 2.

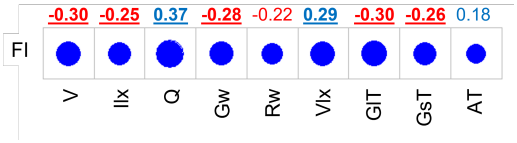
Afterward, to select the most influential variables and reduce their number when generating the model, the Spearman correlation analysis was also carried out [48]. Spearman correlation provides a test to evaluate the degree of association between variables, wherein

this Spearman correlation coefficient will acquire a higher value as the degree of association increases. Then, the *FI* and the regressor variables were analyzed to evaluate their influence on the generated *FI* series estimated from different wind turbines in the wind farm.

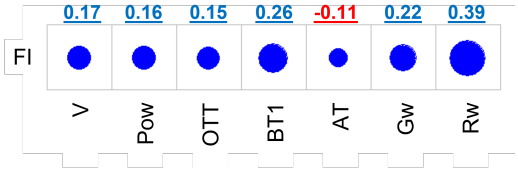
Hereafter, the findings obtained from the Spearman correlation of *FI* and the dependent variables are shown in Figures 4 and 5. These figures show the most influential variables in the failure index results allowing us to discard some of them to simplify the models. In order to harmonize the number of variables and generate the models considering the most influential of them, they were reduced to 7 to establish the final GMCM model in the learning period. For this purpose, the variables with a Spearman correlation index greater than 0.18 (V, Pow, M, Vmax, Vvar, Gw, AT) were selected in the case of the blade model (see Figure 4a). In the case of the generator model, the selected variables were associated with Spearman index values greater than 0.22 (V, Ilx, Q, Gw, Vlx, GlT, GsT) (see Figure 4b). Finally, a value greater than 0.1 was fixed for variable selection in the gearbox model in site 1 (V, Pow, OTT, BT1, AT, Gw, Rw) and site 2 (V, Gw, Rw, OEP, OMP, OIP, Pow) (see Figures 4c and 5).



(**a**) Site 1: Blades



(**b**) Site 1: Generator



(**c**) Site 1: Gearbox

Figure 4. Spearman's Correlation: *FI* vs. dependent variables, Site 1 (selected variables are underlined). Positive (blue) and negative (red) correlations are shown. Circle size reflects correlation strength.

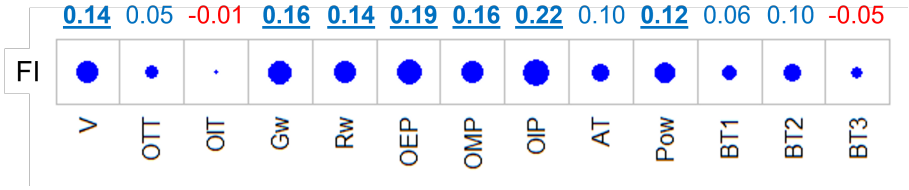


Figure 5. Spearman's Correlation: *FI* vs. dependent variables, Site 2: Gearbox (selected variables are underlined). Positive (blue) and negative (red) correlations are shown. Circle size reflects correlation strength.

In summary, the first step consists of literature-based signal selection. Then, the featured selection method is launched and redundant parameters are removed. For this task, a Failure Index (*FI*) is calculated (Section 2.5.3) for a representative set of data from the components in different wind turbines of the studied wind farm (site-dependent, technology-dependent, component-dependent) and a Spearman correlation is calculated. The most relevant variables are selected based on the Spearman correlation index to establish the optimized GMCM to apply in the test period.

Table 5 summarizes the selection process showing, for both sites, all the available signals that may influence each component. The seven selected signals for each component in each site are shown in color.

Table 5. Available and Selected Variables per Component for both sites. Selected variables are color-coded by component. Numbers 1 and 2 refer to the two studied sites.

Site	Signal	Abbreviation	Component (Selected = Color-Coded)
1	Gearbox bearing temperature 1	BT1	Blades, Gearbox
1	Gearbox bearing temperature 2	BT2	Blades, Gearbox
1	Reactive Power	Q	Generator
1	Voltage Phase $X_{(1,2 \text{ or } 3)}$	Vlx	Generator
1	Current Intensity Phase $X_{(1,2 \text{ or } 3)}$	Ilx	Generator
1	Large generator temperature	GlT	Generator
1	Small generator temperature	GsT	Generator
1	Torque	M	Blades
1	Gearbox oil tank temperature	OTT	Gearbox
1	Ambient temperature	AT	Blades, Generator, Gearbox
1	Rotor speed	Rw	Blades, Generator, Gearbox
1	Generator speed	Gw	Blades, Generator, Gearbox
1	Power	Pow	Blades, Gearbox
1	Wind speed Average	V	Blades, Generator, Gearbox
1	Wind speed Maximum	Vmax	Blades
1	Wind speed Deviation	Vvar	Blades
2	Gearbox bearing temperature 1	BT1	Gearbox
2	Gearbox bearing temperature 2	BT2	Gearbox
2	Gearbox bearing temperature 3	BT3	Gearbox
2	Power	Pow	Gearbox
2	Ambient temperature	AT	Gearbox
2	Gearbox oil input pressure	OIP	Gearbox
2	Gearbox oil mechanical pressure	OMP	Gearbox
2	Gearbox oil electrical pressure	OEP	Gearbox
2	Rotor speed	Rw	Gearbox
2	Generator speed	Gw	Gearbox
2	Gearbox oil inlet temperature	OIT	Gearbox
2	Gearbox oil tank temperature	OTT	Gearbox
2	Wind speed Average	V	Gearbox

2.5. The Gaussian Mixture Copula Model

The Gaussian Mixture Copula Model (GMCM) provides a powerful tool for modeling and understanding the joint behavior of multiple variables. Combining copulas and Gaussian mixture models allows for a flexible and accurate representation of complex dependency structures, making it a valuable approach in statistical modeling and analysis. In this study, the GMCM model is applied using the variables measured with the SCADA data in a multivariate environment. The aim is to highlight the hidden dependency structure in the wind turbine variables. The theoretical foundations of this type of copula model (GMCM) are described below.

Let $[x_i^j]_{p\times n} = (\boldsymbol{x^1}, \ldots, \boldsymbol{x^p}) = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ be p random variables with n instances each, where j identifies the variables, i the observations, and x_i^j is the i-th observation value of

the j-th variable. A single superscript indicates the index along the variables while a single subscript is used for the index along the observations. Random variables are denoted by uppercase letters and their realizations are represented by lowercase letters. Bold font is used for vectors of random variables, observations, or parameters. Plain font is used for scalars.

The Gaussian Mixture Model (GMM) is a probabilistic model used for representing complex data distributions. It assumes that the data are generated from a mixture of several Gaussian (Normal) distributions. The GMM represents the probability density function as a weighted sum of Gaussian distributions. Each Gaussian component represents a cluster or mode in the data. The probability density function (PDF) of a K-component GMM is given by:

$$f_{GMM}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{1}$$

where π_k are the mixing proportions, with $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$, while $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the component-specific vectors of means and covariance matrices, respectively. For simplicity, we use the parameter $\boldsymbol{\theta} = (\pi_1,\ldots,\pi_k,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_k)$ to jointly represent all parameters.

A copula is a mathematical function that characterizes the joint distribution of multiple random variables by representing their dependence structure separately from their marginal distributions. Mathematically, a copula can be described as follows. Let F_j be the marginal cumulative distribution (CDF) of $\boldsymbol{X}^j$. The copula function $C(\boldsymbol{U}^j)$ is a multivariate function defined on the unit hypercube $[0,1]^n$ where $\boldsymbol{U}^j$ are the CDFs of $\boldsymbol{X}^j$, $\boldsymbol{U}^j = F_j(\boldsymbol{X}^j)$. The copula function $C(\boldsymbol{U}^1,\ldots,\boldsymbol{U}^p)$ maps the marginals to the joint CDF, such that for every joint distribution with continuous marginals, $F(\boldsymbol{X}^1,\ldots,\boldsymbol{X}^p)$, there exists a unique copula function such that $F(\boldsymbol{X}^1,\ldots,\boldsymbol{X}^p) = C(F_1(\boldsymbol{X}^1),\ldots,F_p(\boldsymbol{X}^p))$. It can also be shown that the corresponding joint density can be written as the copula density, c, multiplied by the individual marginal densities f_j:

$$f(\boldsymbol{x}) = c(F_1(\boldsymbol{x}^1),\ldots,F_p(\boldsymbol{x}^p)) \prod_{j=1}^{p} f_j(\boldsymbol{x}^j) \tag{2}$$

The Equation (2) illustrates how copulas provide a flexible way to construct multivariate density functions. Copulas achieve this by separating the specification of marginal distributions, f_j from the characterization of the dependence structure, c. This decoupling allows independently choosing the parametric family for each aspect, enabling the customization of the marginals and the dependency modeling according to the specific needs. From Equation (2) the copula family can be obtained as:

$$c(F_1(\boldsymbol{x}^1),\ldots,F_p(\boldsymbol{x}^p)) = \frac{f(\boldsymbol{x})}{\prod_{j=1}^{p} f_j(\boldsymbol{x}^j)} \tag{3}$$

If the joint density function is chosen to be equal to the multivariate normal density ϕ, with normal densities ϕ_j, the Gaussian copula density is given as $c_\phi = \frac{\phi(\boldsymbol{x})}{\prod_{j=1}^{p} \phi_j(\boldsymbol{x}^j)}$.

Copula-based models can be viewed as generative models where the goal is to learn the underlying distribution of the observed data and then use that distribution to generate new samples that resemble the original data. These models are defined on the CDF-transformed data, $\boldsymbol{U}^j = F_j(\boldsymbol{U}^j)$, also called pseudo or latent observations. The generative model for a Gaussian copula is defined through a Gaussian distribution on the latent CDF transformations: $\boldsymbol{X}^j = F_j^{-1}(\boldsymbol{U}^j)$; $\boldsymbol{U}^j = \Phi_j(\boldsymbol{Y}^j)$; $\boldsymbol{Y} \sim \Phi(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where Φ_j is the j^{th} marginal CDF corresponding to the multivariate normal distribution with CDF Φ, mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The Gaussian Mixture Copula Model is obtained using a GMM as the joint density function in Equation (3), which allows writing the GMCM copula density as:

$$c_f(\boldsymbol{U}, \boldsymbol{\theta}) = \frac{f_{GMM}(\boldsymbol{\Psi}_\theta^{-1}(\boldsymbol{U}))}{\prod_{j=1}^{p} \psi_j(\boldsymbol{\Psi}_{j\theta}^{-1}(\boldsymbol{U}^j))} \tag{4}$$

where $\Psi_{j\theta}$ and $\psi_{j\theta}$ are the j^{th} marginal CDF and PDF of the GMM f_{GMM}.

The generative model of the GMCM can be written as:

$$\boldsymbol{X}^j = F_j^{-1}(\boldsymbol{U}^j); \quad \boldsymbol{U}^j = \Psi_j(\boldsymbol{Y}^j); \quad \boldsymbol{Y}^j \sim \psi_{j\theta} \tag{5}$$

The likelihood of n observations from GMCM can be expressed in terms of the realizations of latent variables $\boldsymbol{Y}$:

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\prod_{j=1}^{p} \Sigma_k \psi_j(y_i^j | \mu_k^i, \Sigma_{kjj})} \tag{6}$$

As a summary of the generative process of applying the GMCM Copula, Figure 6 illustrates the transformation of variables in the real domain $\boldsymbol{X}$ to the transformed field $\boldsymbol{U}$ ([49]) followed by modeling interdependencies among the variables using the GMCM. Subsequently, a simulation is performed in the copula space, and then an inverse transformation, as described previously, is applied to obtain simulated data in the real domain. It highlights the workflow of the GMCM model and how the transformations between the copula domain and real variable domain are integrated, providing a clear visual representation of the key stages of the generative process.

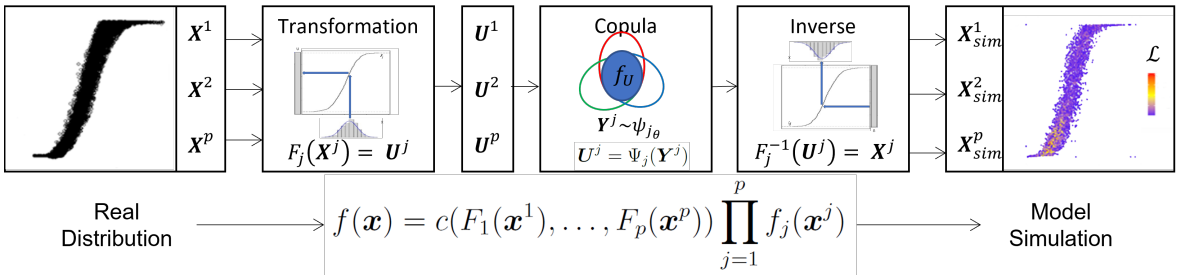


Figure 6. Multivariate copula.

As a final insight, after the theoretical foundations are exposed, the GMCM model can capture the relationship between the different distributions and estimate a joint probability. This estimation is the key to the generation of the failure index proposed in this work.

The strengths of the GMCM models support its selection [50] for the aims of this study. In general, copula models provide a flexible characterization of multivariate distributions. In this sense, GMCM models relax the assumption of working with normal distributions [51]. Therefore, these models allow flexible dependency modeling, especially of non-Gaussian data. GMCM models have been applied to deal with unsupervised pattern recognition, clustering and outlier recognition applications, and they can model many kinds of multi-modal dependencies, notably asymmetric and tail dependencies.

2.5.1. Estimation of the GMCM Parameters

One of the major challenges in using the GMCM is the determination of the parameters and it may require a high computational cost. This is conducted by finding the parameters that maximize the likelihood function given in Equation (6). The optimization of the likelihood function is not trivial and there are intrinsic problems such as convergence and identifiability of the GMCM parameters which may affect any estimation procedure [52]. In this study two optimization algorithms have been chosen, the Pseudo

Expectation Maximization algorithm (PEM) and the Nelder–Mead (NM) algorithm, which are described below.

The EM algorithm is a method for performing maximum likelihood estimation in the presence of latent variables. It does this by first estimating the values of the latent variables (Expectation step), then optimizing the model (Maximization step), and finally repeating these two steps until convergence is achieved. This is an efficient and general approach and is mostly used for density estimation with missing data, such as clustering algorithms like the Gaussian Mixture Model [53]. When the EM algorithm is applied to GMM, the inputs of the model, $(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p)$, remain fixed, as the parameters are iteratively updated through a sequence of alternating Expectation and Maximization steps. On the other hand, the inputs to a GMCM are the marginal values of the CDF, $(\boldsymbol{U}^1, \ldots, \boldsymbol{U}^p)$, which are used to obtain the values of the inverse distribution, $(\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^p)$. Since the inverse distribution functions along the margins change with each update of the parameters, so do the values of the inverse distribution. As a result, the assumption of fixed observations, made by the EM algorithm for GMM, is violated for GMCM and can not be applied directly here. In order to overcome this problem, modifications to the EM algorithm have been developed and are typically called the Pseudo Expectation Maximization (PEM) algorithm. The approach used here uses the initialization parameters to compute the pseudo-data. Then it iterates between the two EM stages, maximizing the pseudo-likelihood of obtaining new model parameters based on the pseudo-data and updating pseudo-data in the expectation stage [54]. The use of nonconstant pseudo-data in the likelihood does not guarantee the convergence of the process and can even achieve convergence to incorrect parameters.

The Nelder–Mead optimization algorithm is a popular approach for optimizing non-differentiable objective functions [55]. It is employed for global or local searches in challenging problems involving noisy, nonlinear, and multi-modal functions. A notable advantage of the Nelder–Mead algorithm is that it does not rely on function gradient information, making it applicable to situations where the gradient is unknown or difficult to compute accurately.

The NM algorithm is a simplex-based method developed to solve the optimization problem of minimizing (maximizing) a given nonlinear function $f : \mathbb{R}^n \to \mathbb{R}$ without constraints. The method uses only function values calculated at some selected points in the real multidimensional space $\mathbb{R}^n$ and does not need any gradient estimate at those points. The NM algorithm begins with a simplex S in $\mathbb{R}^n$, defined as the convex hull of $n+1$ vertices, $(x_0, \ldots, x_n) \in \mathbb{R}^n$ (a simplex is a geometric shape defined in the $\mathbb{R}^n$ domain and used as a 'vehicle' to perform the search for the optimum solution in that domain). The algorithm then starts iterating and proceeds to reshape/move this simplex, vertex by vertex, toward an optimal region in the search space. At each step, it tries one or more modifications of the current simplex and chooses one that moves it towards a 'better' region of the domain. Finally, the simplex vertex that yields the most optimal target value is returned.

Convergence

Convergence is indeed an essential issue in optimizing the model, and there is no guarantee of convergence as said before and discussed in [52]. Therefore, both methods (NM and PEM) have been tested to fit the GMCM model. To do so, 50 different starting parameters have been iterated to ensure the model's convergence in each optimization method and obtain a successful fitting.

It is remarkable that we are dealing with, a seven-variable model, and no convergence was obtained in any case when the optimization method is PEM. However, the NM method achieves successful fittings in all cases in a reasonable time, with the convergence time ranging from 65 to 232 s in the analyzed trials. As an example of these tests, the convergence times for model fitting in four of the cases from site 2 are provided in Table 6.

Table 6. Performance of optimization methods (convergence = converge., elapsed = elaps.).

	NM Algorithm		PEM Algorithm	
WTG	**Converge.**	**elaps. Time (s)**	**Converge.**	**elaps. Time (s)**
S2_GBX_F1	YES	66.30	NO	480.03
S2_GBX_F2	YES	65.06	NO	91.27
S2_GBX_H1	YES	231.97	NO	159.12
S2_GBX_H2	YES	77.97	NO	216.11

This is consistent with the findings reported in [52], where the NM method was already selected as the most robust among those tested.

The fitting algorithms were executed on a machine with the following characteristics: Processor Intel(R) Core(TM) i58250U CPU @1.6 GHz, 16 GB RAM.

Identifiability of the Parameters

The GMCM suffers from unidentifiable parameter configurations due to its invariance properties. For example, the translation invariance implies that only the relative distances between the location parameters, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$, can be deduced. This is (partially) addressed by arbitrarily anchoring the first component to $\boldsymbol{\mu}_1 = 0$. To account for scaling invariance, the variance of the first component is scaled to unity in each dimension, $\boldsymbol{\Sigma}_{1kk}$ for all k. Nevertheless, problems of identifiability may still appear in specific scenarios.

2.5.2. Model Fitting

The procedure for fitting the GMCM, including the optimization of the parameters is applied as described in [32,52]. The main steps considered to model the multivariate copula are as follows:

- Compute the scaled ranks for each column: A range scaling process is applied to the input data of the model. This helps to reduce the impact of outliers, ensures that the scales between variables are comparable, and maintains the relationship of dependency between the variables.
- Parameter initialization: suitable starting parameters for the GMCM model have to be selected. The K-means algorithm is used to choose the starting values heuristically. The parameters defined within the GMCM model are shown in Table 7. The model fitting heavily relies on the selected initial parameters. Therefore, a seed is defined at the beginning of this process, to ensure result replicability and to obtain model-fitting results for different initial parameters.
- Fitting the model: in this step, the final GMCM model parameters are estimated. Here, the optimization methods considered for model coefficients calculation are the Nelder–Mead (NM) and Pseudo-Expectation-Maximization algorithm (PEM) as described in Section 2.5.1.
- Joint probability calculation: Once the models are fitted, the joint probability calculation is conducted. This step generates the log-likelihood for each multivariate observation on the analyzed period.

Table 7. Parameters in GMCM models.

Parameter	Description
k	the number of components in the mixture
p	the dimension of the mixture distribution (input variables in the model)
π	vector of length k related to mixture proportions, where the sum of them is 1
$\boldsymbol{\mu}$	list of length k of numeric vectors of length p for each component
$\boldsymbol{\Sigma}$	list of length k of variance-covariance matrices for each component

As an example of the adjusted parameters in each case presented here, Table 8 shows all the adjusted parameters for the S2_GBX_F1 case.

Table 8. Parameters GMCM model: S2_GBX_F1.

S2_GBX_F1									
k						**Σ**			
3						comp1			
			1.000	−0.045	−0.014	−0.106	0.029	0.020	0.341
p			−0.045	1.000	−0.083	−0.198	−0.011	−0.233	−0.208
7			−0.014	−0.083	1.000	−0.013	−0.016	−0.009	−0.006
			−0.106	−0.198	−0.013	1.000	−0.007	0.009	−0.057
π			0.029	−0.011	−0.016	−0.007	1.000	0.008	0.029
comp1	comp2	comp3	0.020	−0.233	−0.009	0.009	0.008	1.000	0.160
0.193	0.503	0.304	0.341	−0.208	−0.006	−0.057	0.029	0.160	1.000
						comp2			
μ			0.837	−0.007	−0.031	−0.097	0.062	0.286	0.265
comp1	comp2	comp3	−0.007	3.308	0.126	−0.260	0.006	0.313	0.294
0	−3.242	−0.047	−0.031	0.126	2.281	−0.082	−0.100	0.096	0.167
0	0.448	4.009	−0.097	−0.260	−0.082	5.019	0.619	0.145	0.119
0	0.889	2.057	0.062	0.006	−0.100	0.619	1.537	0.260	0.255
0	0.593	−0.948	0.286	0.313	0.096	0.145	0.260	0.856	0.442
0	−2.470	−0.190	0.265	0.294	0.167	0.119	0.255	0.442	0.875
0	−3.481	0.021				comp3			
0	−3.120	0.024	1.026	0.080	−0.097	−0.050	−0.247	0.085	0.104
			0.080	0.777	−0.157	−0.122	−0.054	0.007	0.028
			−0.097	−0.157	1.168	0.031	−0.134	0.010	0.032
			−0.050	−0.122	0.031	1.333	−0.067	−0.018	0.013
			−0.247	−0.054	−0.134	−0.067	3.397	0.071	0.057
			0.085	0.007	0.010	−0.018	0.071	0.976	0.227
			0.104	0.028	0.032	0.013	0.057	0.227	0.958

In this study, 20 real cases have been analyzed. As a result, once the variable selection is made (Section 2.4), 20 GMCM models have been generated.

The GMCM model estimates the joint probability density function from the multivariate data, obtaining a probability density map from the multivariate environment. In this way, it is possible to define a boundary based on the joint probability (joint log-likelihood, in this case) [52]. Based on this joint probability, it is possible to define condition monitoring parameters such as the proposed in this work (see Section 2.5.3). It allows for identifying data whose probability of fulfilling the model is lower, or in other words, to detect periods where a component is working in a failure mode.

To perform the described steps regarding the use of GMCM models, in this study, we have relied on the tools integrated within the R library "GMCM" [52].

2.5.3. Failure Index Generation

The next step consists of transforming the log-likelihood obtained from the GMCM output into a handy and understandable index to detect failures.

In general, log-likelihood provides information about the probability of belonging to the fitted model in the learning stage. In this sense, an observation for which the log-likelihood is high means that it would fit the model's behavior and, therefore, have a higher probability of belonging to it. In the case of this study, it would mean that this observation corresponds to the healthy behavior of the component. If the observation does not fulfill the model, it does not correspond to healthy behavior, and the log-likelihood value will be lower.

In order to generate a smart and comparable index to define an empirical probability of belonging to unhealthy behavior, a transformation is conducted. The transformation applied to GMCM output (log-likelihood) is based on the inverse of the cumulative empirical function [44]. This resulting parameter intends to determine the risk of failure of the component or the probability of working in an unhealthy mode. The Failure Index (*FI*) is defined in Equation (7).

$$FI(\%) = (1 - \frac{(\sum_{i=1}^{i=n} ecdf(P_i))}{n}) * 100 \quad (7)$$

where,

- P_i is the log-likelihood estimated with the copula model.
- $ecdf$ function is the empirical cumulative density function fitted in the learning period to the log-likelihood output variable. This function is used to standardize values (between 0 and 1).
- n is the number of observations

Finally, an *FI* value is generated for each observation where higher *FI* values will indicate a higher probability of operation under unhealthy conditions. Later, the GMCM coefficients are used with the test data to evaluate the performance of the *FI*.

2.6. Signal Smoothing

Lastly, the generated failure index from component selected variables is smoothed. The cubic spline function has been considered for this issue. A spline function is a curve constructed from piecewise polynomial functions, ensuring continuity at their junction points. Spline functions have numerous applications in signal smoothing, noise reduction, and trend extraction in time series analysis.

The choice of cubic splines provides a balance between flexibility and simplicity, ensuring effective smoothing without the risk of overfitting associated with higher-order splines. Specifically, the `smooth.spline` function in R, with an adjusted `spar` parameter to control the smoothness of the fit, was used. This approach has proven to be robust and practical for the dataset used. As a consequence, the generated failure index series is a smoothed [56] which provides a clean summary to the decision-maker.

The resulting series is an easy-to-interpret signal with application in failure detection in components where a multivariable environment is considered. In addition, this same index has successful applications in the early detection of component failures, as demonstrated in the paper [57].

3. Results

This section shows the results obtained for site 1 and site 2. The subsequent paragraphs provide the examination details of temporal evolution for the *FI* and the selected SCADA signals.

3.1. Results Site 1

In site 1, the available variables were limited, but the registered component failure in the maintenance logbook was available on different components (generator, blades and gearbox). Consequently, for this site, three failure cases are presented, one per component (S1_GEN_F1, S1_BLD_F1, S1_GBX_F1) and nine healthy cases (S1_GEN_H1, S1_GEN_H2, S1_GEN_H3,

S1_BLD_H1, S1_BLD_H2, S1_BLD_H3, S1_GBX_H1, S1_GBX_H2, S1_GBX_H3), three for each component (generator, blades, and gearbox).

As the first result of the behavior of the methodology applied through the values obtained from the *FI* index at this site, the graphs in Figure 7 are shown, where each row corresponds to the three healthy cases of each component. This figure displays the range of *FI* values in all the healthy cases analyzed at the site. From the analysis of this figure, the first conclusion regarding the performance of the *FI* index is drawn. In general, the healthy state of the component corresponds to *FI* values equal to or less than 50.

Subsequently, each of the failure cases studied for this site is analyzed. This analysis includes the *FI* index along with all the variables considered and associated with the component's health. Additionally, for comparative purposes, a healthy component case is included along with each failure case.

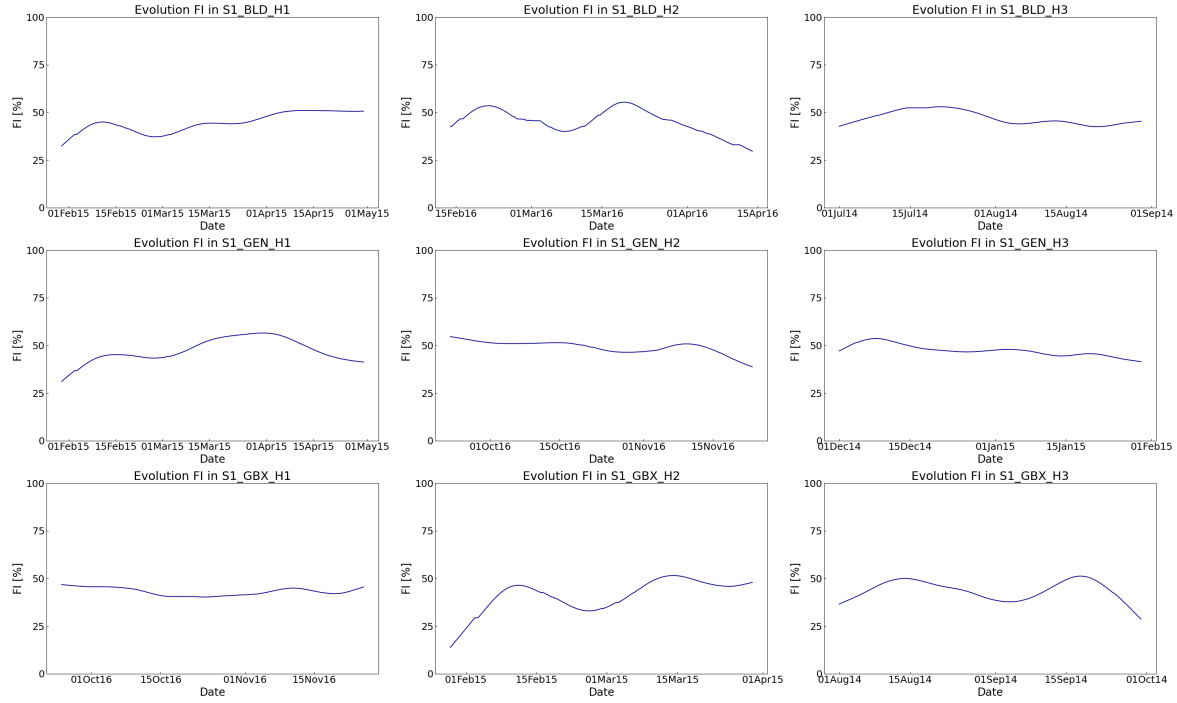


Figure 7. Failure Index Evolution (*FI*). Components in a healthy working mode. Site1 .

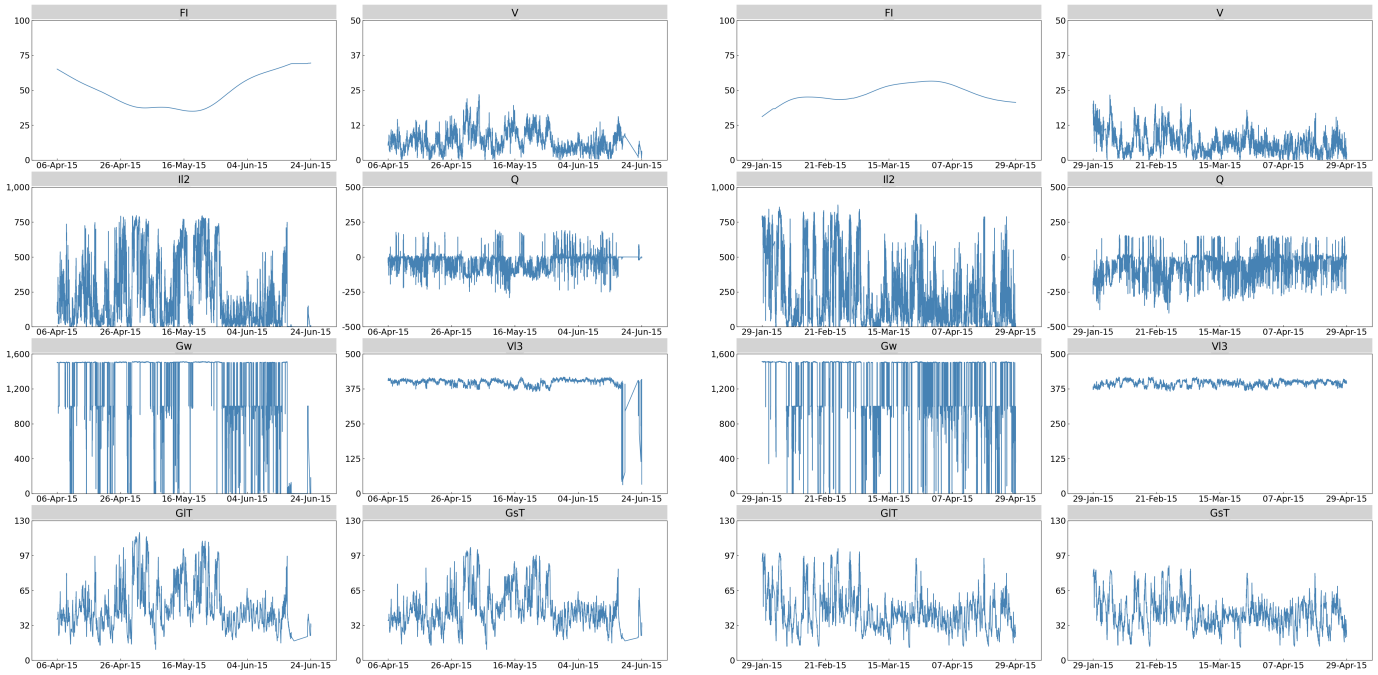
Site 1, Faulty case 1 (generator) presents the analysis performed in the faulty working period of S1_GEN_F1 (investigation of generator case). Figure 8a shows the *FI* trend and the SCADA variables under consideration.

The *FI* ranges between 65 and 70. In this case, major maintenance intervention on the generator was reported in July 2015. Before replacing the component, between 15 May and 15 June, an increase in the *FI* value between 35 and almost 70 was observed. Although similar increases were detected in the weeks prior to the test period, the model seems to detect anomalies before the intervention. Despite there being room for improvement in terms of sensitivity, the method appears to be effective in detection.

Site 1, Healthy case 1 (generator) presents the analysis performed in the healthy working period of S1_GEN_H1 (investigation of generator case). Figure 8b shows the *FI* trend and the SCADA variables under consideration.

The *FI* ranges between 30% and 60%. In accordance with the FI, the SCADA signals do not show signs of abnormal working. The SCADA signals do not exhibit significant gaps during the analysis period. There are no significant mean changes in the time series of these signals. Only from March, there is a slight decrease in the temperature of the large generator and the small generator, but it corresponds with a slight decrease in wind speed in the same period.

While the differences between the healthy case and the faulty case are not as evident as in other components analyzed in this study, they do indicate a reasonable sensitivity to anomaly detection in generators. Therefore, although it would be beneficial to continue investigating by applying complementary techniques and analyzing the data at a lower granularity, the proposed method does appear to provide useful information for the detection of anomalies in generators.



(**a**) (**b**)

Figure 8. Site 1, Time evolution Failure Index (generator). (**a**) Site 1, Faulty case 1 (generator). (**b**) Site 1, Healthy case 1 (generator).

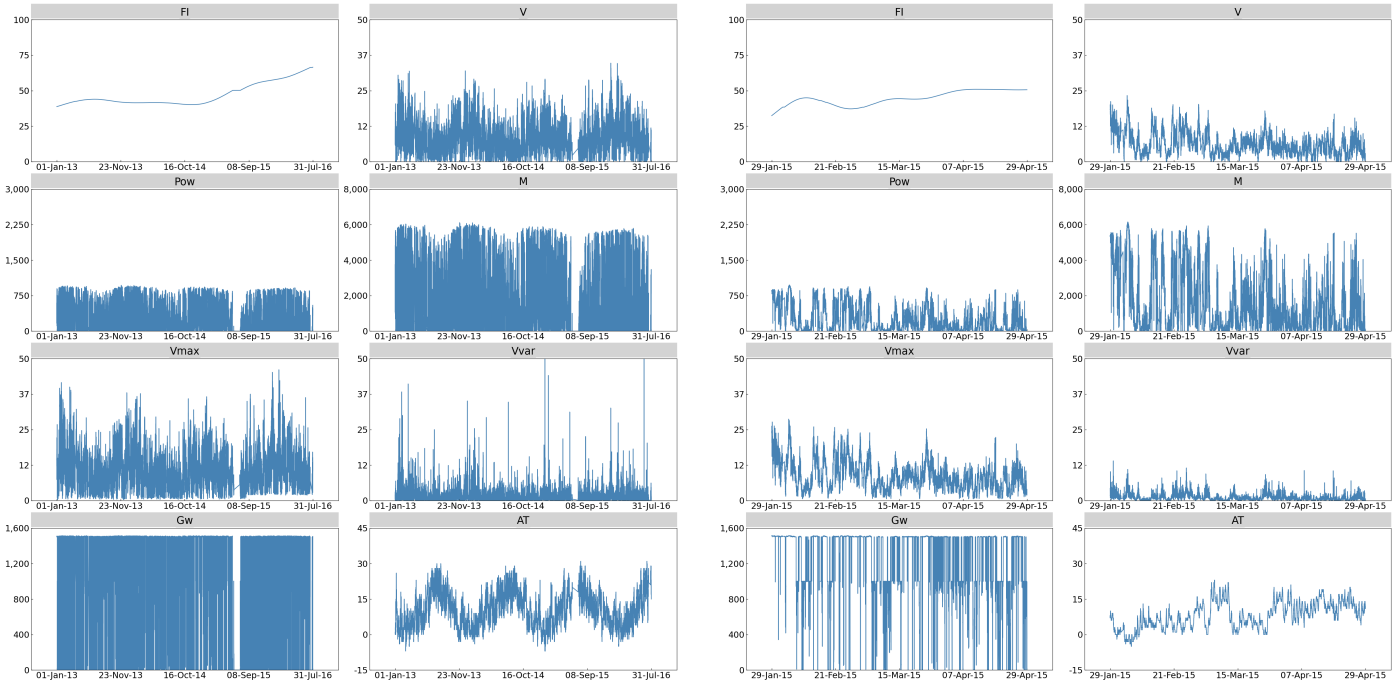
Site 1, Faulty case 2 (blades) presents the analysis performed in the faulty working period of S1_BLD_F1 (investigation of blades case). Figure 9a shows the *FI* trend and the SCADA variables under consideration.

The *FI* ranges between 40% and 66%. In this case, major maintenance interventions on blades are reported on 4 December 2014. In this case, in order to detect some degree of degradation in blades and to monitor the evolution of the *FI* index before and after major corrective actions, the component is monitored between January 2013 and July 2016. A significant increase in the *FI* value is observed from December 2014, which corresponds to the date of the major corrective.

Site 1, Healthy case 2 (blades) presents the analysis performed in the healthy working period of S1_BLD_H1 (investigation of blades case). Figure 9b shows the *FI* trend and the SCADA variables under consideration.

The *FI* ranges between 32% and 50%. In accordance with the FI, the SCADA signals do not show signs of abnormal working. The SCADA signals do not exhibit significant gaps during the analysis period. There are no significant mean changes in the time series of these signals.

Although the results between healthy and faulty cases do not seem conclusive, the method does provide sufficient sensitivity to be considered for this type of application. However, there is room for improvement in terms of its sensitivity. It would likely be interesting to investigate the considered cases further by including other variables in the model and examining their long-term evolution. This long-term analysis has already been attempted in the presented faulty case, where blade degradation could only be captured in the final stage of the available period, as the index increased with a steeper slope.



(**a**) (**b**)

Figure 9. Site 1, Time evolution Failure Index (blades). (**a**) Site 1, Faulty case 2 (blades). (**b**) Site 1, Healthy case 2 (blades).

Site 1, Faulty case 3 (gearbox) presents the analysis performed in the faulty working period of S1_GBX_F1 (investigation of gearbox case). Figure 10a shows the *FI* trend and the SCADA variables under consideration.

The *FI* ranges between 30 and 90. In this case, a major gearbox failure was reported on 25 May 2012. A notable increase in the *FI* value is detected between April 1 and April 20, where a value greater than 90 is reached. After that, a slight decrease in the *FI* value is registered until reaching a value around 60. These values can indicate an early failure detection in the gearbox.

Site 1, Healthy case 3 (gearbox) presents the analysis performed in the healthy working period of S1_GBX_H1 (investigation of gearbox case). Figure 10b shows the *FI* trend and the SCADA variables under consideration.

The *FI* ranges between 40% and 50%. In accordance with the FI, the SCADA signals do not show signs of abnormal working. The SCADA signals do not exhibit significant gaps during the analysis period. There are no significant mean changes in the time series of these signals.

As a final assessment of the gearbox case at site 1, in comparison with the case of a healthy gearbox, the differences between the analyzed healthy and faulty cases seem to be very sensitive for detecting faults in the gearbox. Therefore, the gearbox is analyzed again at site 2, where more variables are considered to generate the optimal model.

(**a**) (**b**)

Figure 10. Site 1, Time evolution Failure Index (gearbox). (**a**) Site 1, Faulty case 3 (gearbox). (**b**) Site 1, Healthy case 3 (gearbox).

3.2. Results Site 2

In the case of site 2, the results obtained from applying the proposed methodology to two faulty gearbox cases (S2_GBX_F1, S2_GBX_F2) and six gearbox healthy cases (S2_GBX_H1, S2_GBX_H2, S2_GBX_H3, S2_GBX_H4, S2_GBX_H5, S2_GBX_H6) are presented.

As an example of the behavior of the methodology applied through the values obtained from the *FI* index, the graphs in Figure 11 are shown. In these graphs, the range of *FI* values in all the healthy cases analyzed at the site can be observed. From the analysis of this figure, the expected behavior of the health index is confirmed: in general, the healthy state of the component corresponds to *FI* values equal to or less than 50.

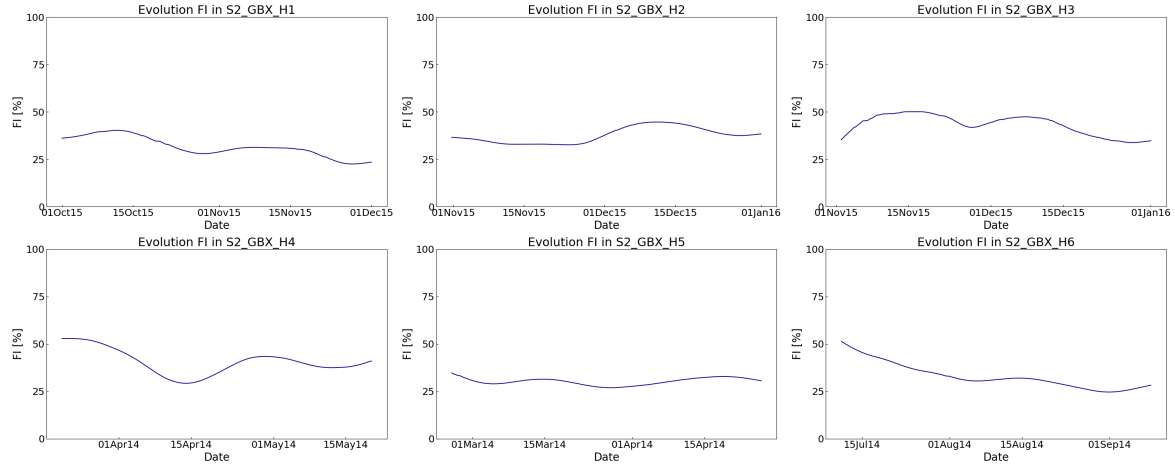


Figure 11. Failure Index Evolution (*FI*). Components in a healthy working mode. Site 2.

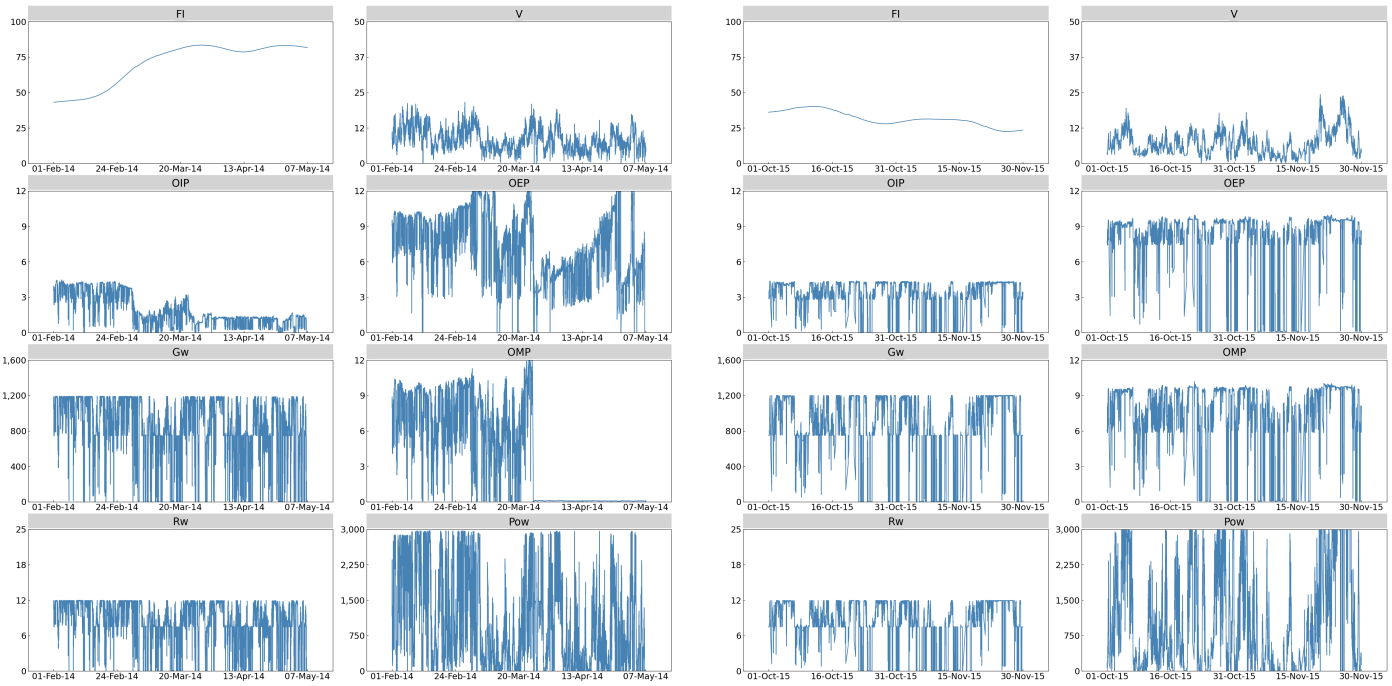
The subsequent paragraphs provide the examination details of temporal evolution for the *FI* and the selected SCADA signals.

Site 2, Faulty case 1 (gearbox): Figure 12a shows the generated *FI* signal starting from 3 months in advance to the gearbox failure occurrence in S2_GBX_F1. In this case, a major

gearbox failure was reported on 7 May 2014. From the last week of March 2014, the gearbox oil mechanical pump pressure signal had been lost. Before the occurrence of this clear anomaly, from February 2014 to the last week of March 2014, failure propagation can be tracked by the FI; hence the *FI* has a clear accelerating upward trend.

Site 2, Healthy case 1 (gearbox) presents the analysis performed in the healthy working period of S2_GBX_H2. Figure 12b shows the *FI* trend and the SCADA variables under consideration. The *FI* ranges between 21% and 40%. In accordance with the FI, the SCADA signals do not show signs of abnormal working. The SCADA signals do not exhibit significant gaps during the analysis period. There are no significant mean changes in the time series of these signals. Only during the last week of November do the means of the SCADA signals increase slightly.

As a general assessment of this case, if we compare the failure case and the healthy case, it can be seen that the range of *FI* values is notably higher in the failure case, increasing notably the closer we are to the registered time of failure. Therefore, it can be concluded that the model is sensitive enough to differentiate between healthy and failure cases.



(**a**) (**b**)

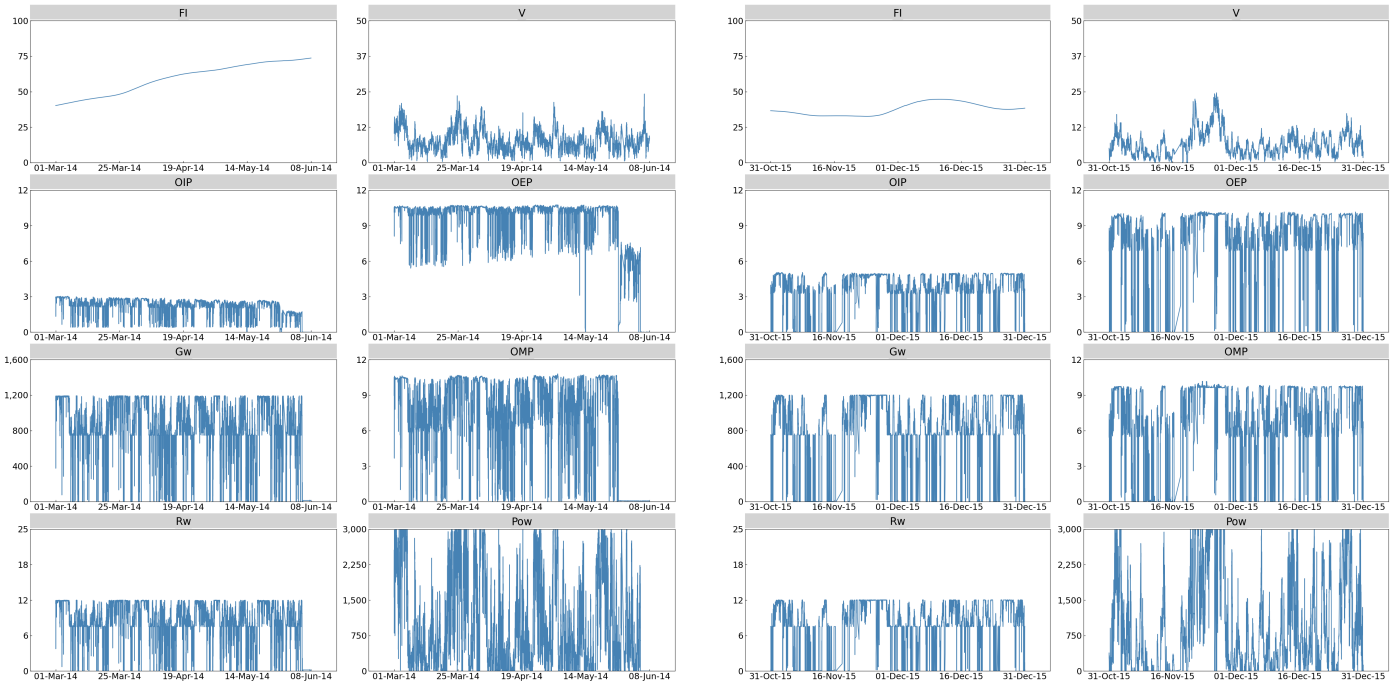
Figure 12. Site 2, Time evolution Failure Index (gearbox). (**a**) Site 2, Faulty case 1 (gearbox). (**b**) Site 2, Healthy case 1 (gearbox).

Site 2, Faulty case 2 (gearbox): Figure 13a shows the generated *FI* signal starting from 3 months in advance to the gearbox failure occurrence in S2_GBX_F2.

In this case, a major gearbox failure is reported on 5 June 2014. Tracking the SCADA signals from the last week of May 2014, it is possible to observe anomalies in gearbox oil input pressure, gearbox oil mechanical pump pressure and gearbox oil electrical pump pressure. Tracking the *FI* from March 2014 to the last week of May 2014, it is possible to foresee the failure propagation; hence *FI* ranges between 40% and 70%.

Site 2, Healthy case 2 (gearbox): covers the analysis performed for the healthy working period of WT5S2. Figure 13b shows the *FI* and the corresponding SCADA signals for the healthy case 2. The *FI* ranges between 32% and 44%. In comparison to the healthy case 1 (Figure 12b), the mean *FI* is greater. The maximum *FI* value was observed in mid-December, while the mean changes in the SCADA signals appeared in the last week of November. As also observed in healthy case 1, the mean of wind speed and the other SCADA signals increase. This similarity between the healthy cases indicates the working behavior of the

wind turbine when the wind blows between 15 and 25 m/s. Thus, it could be concluded that there were no abnormal working indicators within the analysis period.



(**a**) (**b**)

Figure 13. Site 2, Time evolution Failure Index (gearbox). (**a**) Site 2, Faulty case 2 (gearbox). (**b**) Site 2, Healthy case 2 (gearbox).

As shown in this paper, the best failure index performance has been detected for the gearbox case.

In Figure 14, a summary of the observed results for this site is presented through a box and whisker plot, aiming to show the performance of the calculated index. This allows for a visual comparison of the range of *FI* values and the monitored SCADA variables for healthy gearbox cases (in a lighter tone) and faulty gearbox cases (in a darker tone). As shown in this result for site 2, the performance of the fault index for a different site is confirmed.

Figure 14 shows that the median values of the *FI* signal (horizontal lines in each box) range between 20% and 40% in healthy cases, while in faulty cases they range between 60% and 80%. No significant variations are observed in the median wind speed measurements, which for all eight cases are in a mid-operational range between 6 and 8 m/s.

This finding is coherent with the existing literature, where the failures associated with the gearbox component showed a high number in the presence of wind speeds that were steadily in the low or high range. Alternations in wind speed did not seem to play a role in the gearbox failures (see Tables 5 and 8 in [58]). Similarly, generator speed, rotor speed, and power also do not vary significantly between the cases analyzed and observed no relationship with component health.

As for the oil input pressure, the oil mechanical pump pressure and the oil electrical pump pressure, there seems to be a slight decrease in the values compared to the healthy cases. However, although it is more accentuated in the case of the oil input pressure, this does not occur in the two failure cases studied. This issue could be due to some unknown operating mode or any type of failure mode. This should be studied further with the maintenance team. In any case, *FI* allows simultaneously evaluating the considered variables. The relationship of these predictor variables makes the failure index a good indicator of the component's health.

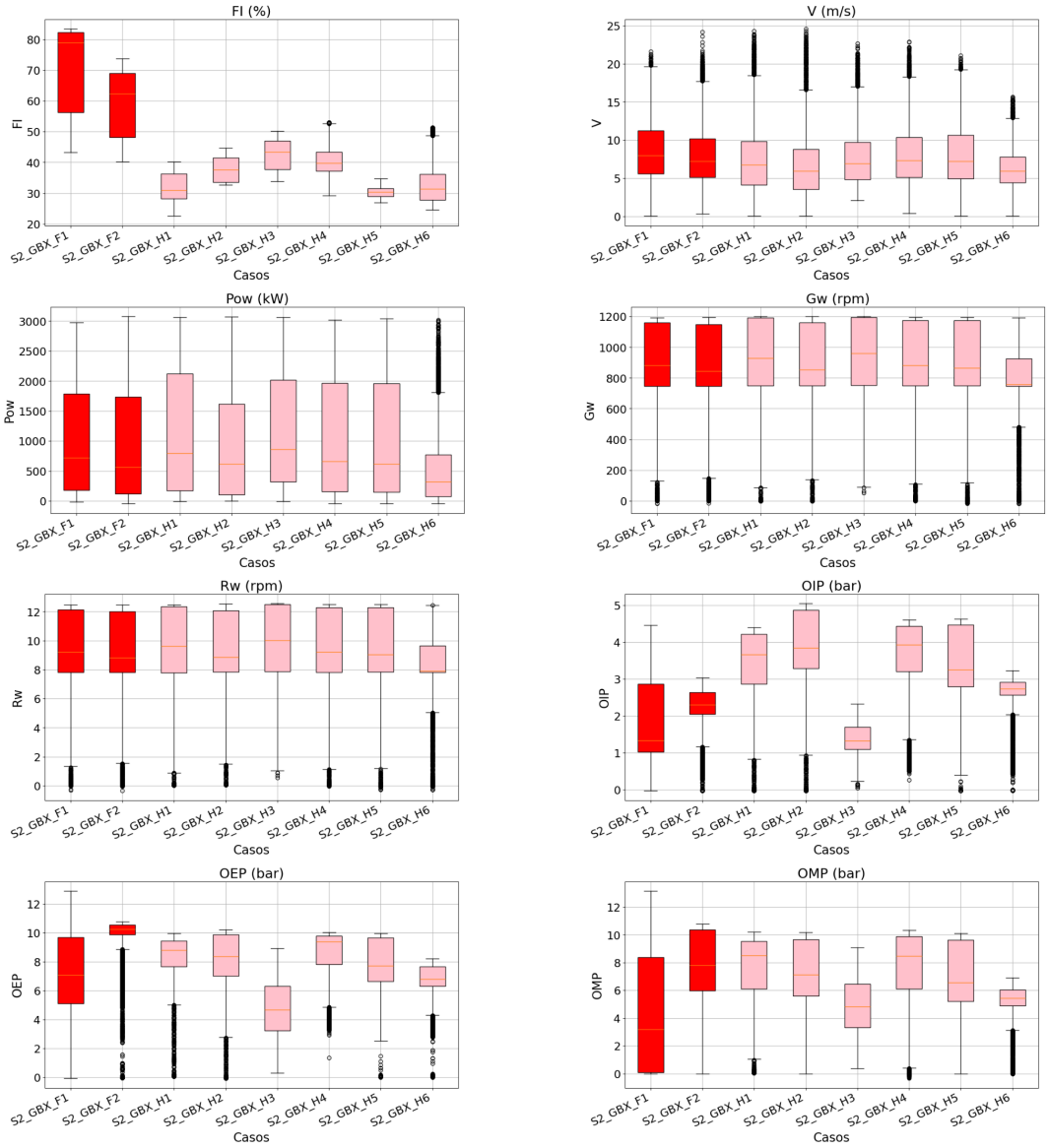


Figure 14. Sample distributions analysis: SCADA signals versus Failure Index results. Site 2 (gearbox). Dark tones represent failure cases, while light tones represent healthy cases.

4. Discussion

The main aspects of multivariable component failure estimation models are addressed and improved. A feature selection procedure and a functional and practical *FI* definition are provided for different gearbox failure modes. The proposed methodology can indicate failure propagation in advance of failure occurrence with a period of longer than two weeks. Although other components (generator and blades) have been tested by this method, the results were not as satisfactory as the gearbox case and will be studied deeper in the future. Regarding the methodology and the model used, it has been shown that it is a flexible model, capable of capturing the dependence of the variables that define the health of the components. This flexibility is demonstrated by the model's successful application to various wind turbine components, such as the gearbox, generator, and blades, each with distinct operational characteristics and failure modes. This GMCM model, applied in the proposed way, makes it possible to obtain a health index of the component that is easy to monitor, reducing the computational cost of the decision system for planning the plant's maintenance. As a drawback of the model, it has been detected that sometimes the

adjustment of the model is costly, affecting the detection sensitivity of the method in the case that the adjustment is not satisfactory. This requires the generation of several fitting scenarios to select the best one.

The cubic spline smoothing method was chosen to balance flexibility and simplicity, ensuring effective smoothing without the risk of overfitting associated with higher-order splines. While other smoothing methods may be more appropriate depending on the nature of the signal, the cubic splines here were well suited to our desired outcome.

In the next steps, this methodology will be applied in the detection of failure propagation in other components (yaw system, pitch, etc.) and other renewable energy systems like PV plants. Besides, a higher data frequency could be considered in the models to obtain better results and include different variables if available. Furthermore, this method could have applications in improving the estimation of production, taking into consideration the different working conditions captured by the copula model, and it will be tested in future works.

Author Contributions: Conceptualization, R.L., J.J.M. and N.Y.Y.; methodology, R.L., J.J.M. and N.Y.Y.; software, R.L. and N.Y.Y.; validation, R.L., J.J.M. and N.Y.Y.; formal analysis, R.L., J.J.M. and N.Y.Y.; investigation, R.L., J.J.M. and N.Y.Y.; resources, R.L.; data curation, R.L. and N.Y.Y.; writing—original draft preparation, R.L. and N.Y.Y.; writing—review and editing, R.L., J.J.M. and N.Y.Y.; visualization, R.L., J.J.M. and N.Y.Y.; supervision, J.J.M.; project administration, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to confidentiality.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GMCM	Gaussian Mixture Copula Model
FI	Failure Index
WT	Wind Turbine
CM	Condition Monitoring
SCADA	Supervisory Control and Data Acquisition
CMS	Condition Monitoring System
GMM	Gaussian Mixture Model
O&M	Operation and Maintenance
RUL	Remaining Useful Life
EWMA	Exponentially Weighted Moving average
MEWMA	Mutivariate Exponentially Weighted Moving average
GMPOP	Generalised Multiscale Poincare Plots
SVDD	Support Vector Data Description
MD	Mahalanobis Distance
PDF	Probability Density Function
CDF	Cumulative Distribution Function
PEM	Pseudo Expectation Maximization
NM	Nelder–Mead
LSTM	Long Short-Term Memory
SDAE	Stacked Denoising Autoencoder
XGBoost	Extreme Gradient Boosting
ARIMA	Autoregressive Integrated Moving Average
TWSVM	Twin Support Vector Machine

MDAE	Multi-kernel maximum mean discrepancy Deep AutoEncoder
DTAD	Dual-channel Transformer using Auxiliary Discriminator
DT	Decision Tree
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
MLPNN	Multi-Layer Perceptron Neural Network

References

1. Kang, J.; Sobral, J.; Soares, C.G. Review of condition-based maintenance strategies for offshore wind energy. *J. Mar. Sci. Appl.* **2019**, *18*, 1–16. [CrossRef]
2. Ha, J.M.; Oh, H.; Park, J.; Youn, B.D. Classification of operating conditions of wind turbines for a class-wise condition monitoring strategy. *Renew. Energy* **2017**, *103*, 594–605. [CrossRef]
3. Ziegler, L.; Gonzalez, E.; Rubert, T.; Smolka, U.; Melero, J.J. Lifetime extension of onshore wind turbines: A review covering Germany, Spain, Denmark, and the UK. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1261–1271. [CrossRef]
4. Tao, F.; Zhang, M.; Liu, Y.; Nee, A. Digital twin driven prognostics and health management for complex equipment. *CIRP Ann.* **2018**, *67*, 169–172. [CrossRef]
5. Leahy, K.; Gallagher, C.; Bruton, K.; O'Donovan, P.; O'Sullivan, D.T. Automatically identifying and predicting unplanned wind turbine stoppages using scada and alarms system data: Case study and results. *J. Phys. Conf. Ser.* **2017**, *926*, 012011. [CrossRef]
6. Leahy, K.; Gallagher, C.; O'Donovan, P.; O'Sullivan, D.T. Issues with data quality for wind turbine condition monitoring and reliability analyses. *Energies* **2019**, *12*, 201. [CrossRef]
7. He, R.; Tian, Z.; Wang, Y.; Zuo, M.; Guo, Z. Condition-based maintenance optimization for multi-component systems considering prognostic information and degraded working efficiency. *Reliab. Eng. Syst. Saf.* **2023**, *234*, 109167. [CrossRef]
8. Van Horenbeek, A.; Van Ostaeyen, J.; Duflou, J.R.; Pintelon, L. Quantifying the added value of an imperfectly performing condition monitoring system—Application to a wind turbine gearbox. *Reliab. Eng. Syst. Saf.* **2013**, *111*, 45–57. [CrossRef]
9. Zio, E. Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliab. Eng. Syst. Saf.* **2022**, *218*, 108119. [CrossRef]
10. Medjaher, K.; Zerhouni, N.; Baklouti, J. Data-driven prognostics based on health indicator construction: Application to PRONOSTIA's data. In Proceedings of the 2013 European Control Conference, ECC 2013, Zurich, Switzerland, 17–19 July 2013; pp. 1451–1456. [CrossRef]
11. Zhou, Q.; Xiong, T.; Wang, M.; Xiang, C.; Xu, Q. Diagnosis and early warning of wind turbine faults based on cluster analysis theory and modified ANFIS. *Energies* **2017**, *10*, 898. [CrossRef]
12. Artigao, E.; Martín-Martínez, S.; Honrubia-Escribano, A.; Gómez-Lázaro, E. Wind turbine reliability: A comprehensive review towards effective condition monitoring development. *Appl. Energy* **2018**, *228*, 1569–1583. [CrossRef]
13. Reder, M.; Gonzalez, E.; Melero, J.J. Wind Turbine Failure Analysis—Targeting current problems in Failure Data Analysis. *J. Phys. Conf. Ser.* **2016**, *753*, 072027. [CrossRef]
14. Igba, J.; Alemzadeh, K.; Henningsen, K.; Durugbo, C. Effect of preventive maintenance intervals on reliability and maintenance costs of wind turbine gearboxes. *Wind. Energy* **2014**, *18*, 2013–2024. [CrossRef]
15. Igba, J.; Alemzadeh, K.; Durugbo, C.; Henningsen, K. Performance assessment of wind turbine gearboxes using in-service data: Current approaches and future trends. *Renew. Sustain. Energy Rev.* **2015**, *50*, 144–159. [CrossRef]
16. Shafiee, M.; Sørensen, J.D. Maintenance optimization and inspection planning of wind energy assets: Models, methods and strategies. *Reliab. Eng. Syst. Saf.* **2019**, *192*, 105993. [CrossRef]
17. Yürüşen, N.Y.; Rowley, P.N.; Watson, S.J.; Melero, J.J. Automated wind turbine maintenance scheduling. *Reliab. Eng. Syst. Saf.* **2020**, *200*, 106965. [CrossRef]
18. Badihi, H.; Zhang, Y.; Jiang, B.; Pillay, P.; Rakheja, S. A Comprehensive Review on Signal-Based and Model-Based Condition Monitoring of Wind Turbines: Fault Diagnosis and Lifetime Prognosis. *Proc. IEEE* **2022**, *110*, 754–806. [CrossRef]
19. Cambron, P.; Masson, C.; Tahan, A.; Pelletier, F. Control chart monitoring of wind turbine generators using the statistical inertia of a wind farm average. *Renew. Energy* **2018**, *116*, 88–98. [CrossRef]
20. Yang, H.H.; Huang, M.L.; Lai, C.M.; Jin, J.R. An approach combining data mining and control charts-based model for fault detection in wind turbines. *Renew. Energy* **2018**, *115*, 808–816. [CrossRef]
21. Cambron, P.; Lepvrier, R.; Masson, C.; Tahan, A.; Pelletier, F. Power curve monitoring using weighted moving average control charts. *Renew. Energy* **2016**, *94*, 126–135. [CrossRef]
22. Zhang, C.; Hu, D.; Yang, T. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. *Reliab. Eng. Syst. Saf.* **2022**, *222*, 108445. [CrossRef]
23. Feng, Y.; Qiu, Y.; Crabtree, C.J.; Long, H.; Tavner, P.J. Monitoring wind turbine gearboxes. *Wind Energy* **2013**, *16*, 728–740. [CrossRef]
24. Wang, T.; Han, Q.; Chu, F.; Feng, Z. Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review. *Mech. Syst. Signal Process.* **2019**, *126*, 662–685. [CrossRef]
25. Shao, K.; He, Y.; Xing, Z.; Du, B. Detecting wind turbine anomalies using nonlinear dynamic parameters-assisted machine learning with normal samples. *Reliab. Eng. Syst. Saf.* **2023**, *233*, 109092. [CrossRef]

26. Yang, Y.; Bai, Y.; Li, C.; Yang, Y.N. Application Research of ARIMA Model in Wind Turbine Gearbox Fault Trend Prediction. In Proceedings of the 2018 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Xi'an, China, 15–17 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 520–526. [CrossRef]
27. Pandit, R.; Astolfi, D.; Hong, J.; Infield, D.; Santos, M. SCADA data for wind turbine data-driven condition/performance monitoring: A review on state-of-art, challenges and future trends. *Wind Eng.* **2023**, *47*, 422–441. [CrossRef]
28. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review IET Renewable Power Generation Using SCADA data for wind turbine condition monitoring—A review. *Generation* **2017**, *11*, 382–394. [CrossRef]
29. Stetco, A.; Nenadic, G.; Flynn, D.; Barnes, M.; Zhao, X.; Dinmohammadi, F.; Keane, J.; Robu, V. Machine learning methods for wind turbine condition monitoring: A review. *Renew. Energy* **2018**, *133*, 620–635. [CrossRef]
30. Gill, S.; Stephen, B.; Galloway, S. Wind turbine condition assessment through power curve copula modeling. *IEEE Trans. Sustain. Energy* **2012**, *3*, 94–101. [CrossRef]
31. Stephen, B.; Galloway, S.J.; McMillan, D.; Hill, D.C.; Infield, D.G. A copula model of wind turbine performance. *IEEE Trans. Power Syst.* **2011**, *26*, 965–966. [CrossRef]
32. Wang, Y.; Infield, D.G.; Stephen, B.; Galloway, S.J. Copula-based model for wind turbine power curve outlier rejection. *Wind Energy* **2014**, *17*, 1677–1688. [CrossRef]
33. Bai, G.; Fleck, B.; Zuo, M.J. A stochastic power curve for wind turbines with reduced variability using conditional copula. *Wind Energy* **2016**, *19*, 1519–1534. [CrossRef]
34. Song, Z.; Zhang, Z.; Jiang, Y.; Zhu, J. Wind turbine health state monitoring based on a Bayesian data-driven approach. *Renew. Energy* **2018**, *125*, 172–181. [CrossRef]
35. Zheng, M.; Man, J.; Wang, D.; Chen, Y.; Li, Q.; Liu, Y. Semi-supervised multivariate time series anomaly detection for wind turbines using generator SCADA data. *Reliab. Eng. Syst. Saf.* **2023**, *235*, 109235. [CrossRef]
36. Dhiman, H.S.; Deb, D.; Muyeen, S.M.; Kamwa, I. Wind Turbine Gearbox Anomaly Detection Based on Adaptive Threshold and Twin Support Vector Machines. *IEEE Trans. Energy Convers.* **2021**, *36*, 3462–3469. [CrossRef]
37. Abdallah, I.; Dertimanis, V.; Mylonas, H.; Tatsis, K.; Chatzi, E.; Dervilis, N.; Worden, K.; Maguire, E. Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data. In *Safety and Reliability—Safe Societies in a Changing World—Proceedings of the 28th International European Safety and Reliability Conference, ESREL 2018, Trondheim, Norway, 17–21 June 2018*; Department of Civil, Environmental and Geomatic Engineering, ETH Zürich: Zürich, Switzerland, 2018; pp. 3053–3062. [CrossRef]
38. Farrar, N.O.; Ali, M.H.; Dasgupta, D. Artificial Intelligence and Machine Learning in Grid Connected Wind Turbine Control Systems: A Comprehensive Review. *Energies* **2023**, *16*, 1530. [CrossRef]
39. Pentreath, N. *Machine Learning with Spark*; Packt Publishing Ltd.: Birmingham, UK, 2015.
40. Whitenack, D. *Machine Learning with Go: Implement Regression, Classification, Clustering, Time-Series Models, Neural Networks, and More Using the Go Programming Language*; Packt Publishing Ltd.: Birmingham, UK, 2017.
41. Yang, W.; Court, R.; Jiang, J. Wind turbine condition monitoring by the approach of SCADA data analysis. *Renew. Energy* **2013**, *53*, 365–376. [CrossRef]
42. Bangalore, P.; Tjernberg, L.B. An artificial neural network approach for early fault detection of gearbox bearings. *IEEE Trans. Smart Grid* **2015**, *6*, 980–987. [CrossRef]
43. Niu, G.; Singh, S.; Holland, S.W.; Pecht, M. Health monitoring of electronic products based on Mahalanobis distance and Weibull decision metrics. *Microelectron. Reliab.* **2011**, *51*, 279–284. [CrossRef]
44. Pontoppidan, N.; Larsen, J. Unsupervised condition change detection in large diesel engines. In Proceedings of the 2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), Toulouse, France, 17–19 September 2003; IEEE: Piscataway, NJ, USA, 2003; pp. 565–574. [CrossRef]
45. Chouldechova, A.; Hastie, T. Generalized Additive Model Selection, *arXiv* **2015**. arXiv:1506.03850.
46. Coronado, D.; Kupferschmidt, C. Assessment and Validation of Oil Sensor Systems for On-line Oil Condition Monitoring of Wind Turbine Gearboxes. *Procedia Technol.* **2014**, *15*, 747–754. [CrossRef]
47. Lenard, J.G. 9—Tribology. In *Primer on Flat Rolling*, 2nd ed.; Lenard, J.G., Ed.; Elsevier: Oxford, UK, 2014; pp. 193–266. [CrossRef]
48. Cohen, J.; Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2002. [CrossRef]
49. Meucci, A. A Short, Comprehensive, Practical Guide to Copulas Univariate results. *GARP Risk Prof.* **2011**, 22–27. [CrossRef]
50. Kasa, S.R.; Rajan, V. Automatic Differentiation in Mixture Models. *arXiv* **2018**, arXiv:1812.05928. [CrossRef]
51. Tewari, A.; Giering, M.J.; Raghunathan, A. Parametric characterization of multimodal distributions with non-Gaussian modes. In Proceedings of the IEEE International Conference on Data Mining, ICDM, Vancouver, BC, Canada, 11–14 December 2011; pp. 286–292. [CrossRef]
52. Bilgrau, A.E.; Eriksen, P.S.; Rasmussen, J.G.; Johnsen, H.E.; Dybkaer, K.; Boegsted, M. GMCM: Unsupervised Clustering and Meta-Analysis Using Gaussian Mixture Copula Models. *J. Stat. Softw.* **2016**, *70*, 1–23. [CrossRef]
53. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–38. [CrossRef]
54. Li, Q.; Brown, J.B.; Huang, H.; Bickel, P.J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **2011**, *5*, 1752–1779. [CrossRef]

55. Nelder, J.A.; Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **1965**, *7*, 308–313. [CrossRef]
56. Pollock, D. Chapter 11—Smoothing with Cubic Splines. In *Handbook of Time Series Analysis, Signal Processing, and Dynamics*; Pollock, D., Ed.; Signal Processing and its Applications, Academic Press, Queen Mary and Westfield College The University of London UK: London, UK, 1999; pp. 293–322. [CrossRef]
57. Lázaro, R.; Yürüşen, N.Y.; Melero, J.J. Determining Remaining Lifetime of Wind Turbine Gearbox Using a Health Status Indicator Signal. *J. Phys. Conf. Ser.* **2020**, *1618*, 022037. [CrossRef]
58. Reder, M.; Yürüşen, N.Y.; Melero, J.J. Data-driven learning framework for associating weather conditions and wind turbine failures. *Reliab. Eng. Syst. Saf.* **2018**, *169*, 554–569. [CrossRef]