Power Prediction from Wind Turbine SCADA Data in the Presence of Modelling and Measurement Uncertainties

Maneesh Singh

Western Norway University of Applied Sciences, 5020 Bergen, Norway

*Corresponding author: maneesh.singh@hvl.no

Received Month X, XXXX | Accepted Month X, XXXX | Posted Online Month X, XXXX

Wind turbines are continuously exposed to harsh environmental and operational conditions throughout their lifetime, leading to the gradual degradation of their components. If left unaddressed, these degraded components can adversely affect turbine performance and significantly increase the likelihood of failure. As degradation progresses, the risk of failure escalates, making it essential to implement appropriate risk control measures.

One effective risk control method involves performing inspection and monitoring activities that provide valuable insights into the condition of the structure, enabling the formulation of appropriate maintenance strategies based on accurate assessments.

Supervisory Control and Data Acquisition (SCADA) systems offer low-resolution condition-monitoring data that can be used for fault detection, diagnosis, quantification, prognosis, and maintenance planning. One commonly used method involves predicting power generation using SCADA data and comparing it against measured power generation. Significant discrepancies between predicted and measured values can indicate sub-optimal operation, natural aging, or unnatural faults.

Various predictive models, including parametric and non-parametric (statistical) approaches, have been proposed for estimating power generation. However, the imperfect nature of these models introduces uncertainties in the predicted power output. Additionally, SCADA monitoring data is prone to uncertainties arising from various sources. The presence of uncertainties from these two sources – imperfect predictive models and imperfect SCADA data – introduces uncertainty in the predicted power generation. This uncertainty complicates the process of determining whether discrepancies between measured and predicted values are significant enough to warrant maintenance actions.

Depending on the nature of uncertainty – *aleatory*, arising from inherent randomness, or *epistemic*, stemming from incomplete knowledge or limited data – different analytical approaches, like *Probabilistic* and *Possibilistic*, can be applied for effective management. Both, Probabilistic and Possibilistic, Approaches offer distinct advantages and limitations. The Possibilistic Approach, rooted in fuzzy set theory, is

particularly well-suited for addressing epistemic uncertainties, especially those caused by imprecision or sparse statistical information. This makes it especially relevant for applications such as wind turbines, where it is often challenging to construct accurate probability distribution functions for environmental parameters due to limited sensor data from hard-to-access locations.

This research focuses on developing a methodology for identifying sub-optimal operation in wind turbines by comparing Grid Produced Power (Measured Produced Power) with Predicted Produced Power. To achieve this, the paper introduces a Possibilistic Approach for power prediction that accounts for uncertainties stemming from both model imperfections and measurement errors in SCADA data. The methodology combines machine learning models, used to establish predictive relationships between environmental inputs and power output, with a Possibilistic Framework that represents uncertainty through possibility distribution functions based on fuzzy logic and interval analysis. A real-world case study using operational SCADA data demonstrates the approach, with XGBoost selected as the final predictive model due to its strong accuracy and computational efficiency.

Keywords: interval analysis, ML, measurement error, uncertainty, possibility distribution function, wind turbine

1. Introduction

Throughout its operational lifespan, a wind turbine is consistently exposed to harsh operational and environmental conditions, such as wind velocity, humidity, temperature, precipitation, and icing. These factors trigger various degradation mechanisms, including corrosion, erosion, fatigue, and deformation, which can deteriorate critical components and significantly compromise the integrity of associated structures. If these degraded components are not attended to, their performance will diminish and the likelihood of their failure will increase. Thus, as degradation progresses, the risk of failure rises, necessitating the need for implementing appropriate risk control measures involving effective and efficient asset integrity management program.

However, financial, social ("not-in-my-backyard" syndrome), environmental (e.g., meteorological conditions), and geographical (e.g., topological features) factors often necessitate placing wind turbines in remote and difficult to access locations. This remoteness significantly increases the costs of asset integrity management, with maintenance expenses estimated to account for a substantial portion (10–25%) of the total annual operational cost [1]. Hence, there is a need for developing effective, efficient and economically viable asset integrity management program for wind turbines.

One effective risk control approach involves deploying a robust asset integrity management strategy, which includes monitoring, inspection, and maintenance of structures at suitable intervals. Inspection

and monitoring activities provide valuable insights into the structural condition, enabling the application of targeted maintenance strategies throughout the turbine's lifecycle.

Currently, maintenance management (inspection and maintenance) plans are developed using two primary approaches:

- *Traditional Approach:* Relies on understanding the failure profile of components – such as failure causes, mechanisms, modes, and rates – to manually develop maintenance plans based on historical data and experience.

- *Condition-Based Approach:* Analyses data collected through condition monitoring systems for fault detection, diagnosis, quantification, and prognosis. This information is used to create dynamic maintenance plans that respond to real-time or near-real-time changes in equipment condition.

The Traditional Approach relies on examining structural, environmental, and operational attributes to formulate corrective or preventive maintenance strategies. Preventive maintenance is typically time-based; for instance, maintenance activities for wind turbines are often scheduled at intervals of 3 to 6 months, depending on the turbine's age and condition [1]. However, time-based inspection and maintenance plans can be costly to implement. To address this, methodologies rooted in formalized risk analysis, such as *Risk-Based Inspection and Maintenance* or *Reliability-Centered Maintenance*, have been developed. These methods involve understanding the failure profile and conducting risk analysis and evaluation to establish maintenance plans that are more efficient and effective than time-based or incident-driven approaches [2].

The Condition-Based Approach enhances maintenance management plans established by the Traditional Approach by utilizing real-time condition attributes to continually refine the equipment's risk assessment through fault detection. This method involves analysing data collected from intermittent or continuous monitoring using sensors for fault detection, diagnosis, prognosis, and advisory generation. By assessing the equipment's actual health status, the Condition-Based Approach enables the development of maintenance plans that are dynamically tailored to the actual condition of the equipment.

In a wind turbine Supervisory Control and Data Acquisition (SCADA) system, numerous sensors continuously monitor various meteorological and operational parameters, with data transmitted, processed, and stored in SCADA supervisory computers. The parameters monitored include:

1. *Position:* Blade pitch angle, nacelle direction.

2. *Temperature:* Nose cone, gearbox bearing, gearbox oil, hydraulic system oil, generator bearing, generator stator windings, generator split ring chamber, transformer, busbar section, inverter, controllers, VCP control boards.

3. *RPM:* Rotor speed, generator speed.

4. *Hydraulic Characteristics:* Pressure, reservoir level, flow rate.

5. *Environmental Characteristics:* Wind speed, wind direction, ambient temperature, humidity.

6. *Electrical Characteristics:* Active power, reactive power, voltage, current, phase displacement, frequency.

Additionally, data streams from nearby weather stations are often recorded to provide further insights into environmental conditions affecting turbine performance.

Despite its numerous advantages, the adoption of the Condition-Based Approach remains limited and requires further research and development. This is largely due to several challenges associated with [3]:

1. *Quality and Quantity of Collected Data:* Ensuring sufficient, accurate, and comprehensive data is critical for effective analysis, and limitations in data availability and reliability can hinder performance.

2. *Handling Imperfect Data:* Faulty sensors may produce spurious, inconsistent, inaccurate, uncertain, or irrational data, complicating the analysis and potentially leading to erroneous conclusions.

3. *Data Interpretation:* Accurately diagnosing faults, quantifying damage, and forecasting future conditions requires sophisticated analytical techniques, which can be challenging to implement effectively.

4. *Updating Maintenance Plans:* Continuously adjusting maintenance plans based on new insights from real-time monitoring is complex and resource-intensive.

5. *Managing Unreliable Analysis:* Poor-quality data or flawed analytical models can result in false alarms (false positives) or undetected faults (false negatives), undermining the credibility of the system.

Addressing these challenges is essential for enhancing the reliability, accuracy, and efficiency of Condition-Based Maintenance systems.

2 Motivation and Aim of the Research

2.1 Motivation for the Research

A drawback of the Condition-Based Approach is its dependence on monitoring data to assess the condition of wind turbine components. Monitoring through sensors generates massive amounts of data, which may be imperfect, inconsistent, or challenging to interpret accurately. These imperfections can undermine the reliability of subsequent analysis and interpretation, causing decision-makers to doubt the validity of the results. Consequently, there is reluctance to incorporate these findings into future maintenance planning. As a result, valuable data is often underutilized, leading to maintenance and inspection decisions being made without fully considering all relevant information. This underscores the need to develop systems that can effectively process and utilize monitoring data to enhance decision-making, while also transparently acknowledging the limitations and uncertainties inherent in the analysis.

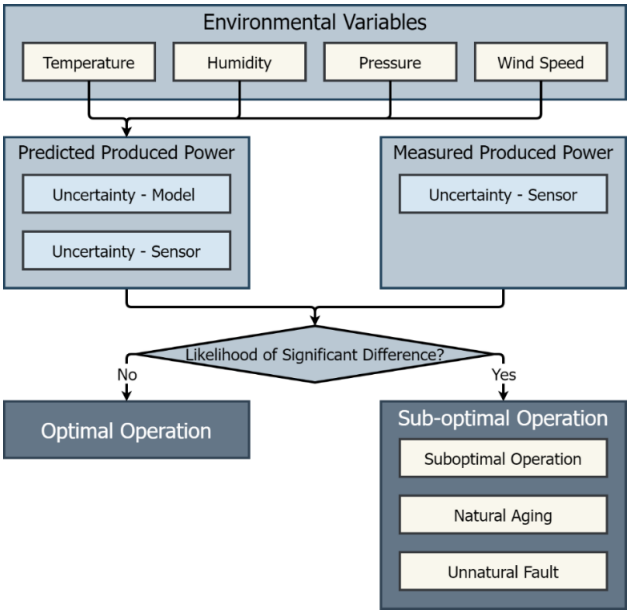

FIGURE 1. Flowchart showing the proposed fault detection methodology.

A commonly used Condition-Based Approach for evaluating wind turbine performance is by using SCADA data that involves analysing power generation as a function of various variables, particularly wind speed. A substantial discrepancy between predicted power generation and the actual measured power generation can indicate sub-optimal performance, which warrants further investigation.

Fig. 1 presents a flowchart illustrating a methodology for detecting sub-optimal power production. In this approach, the *Predicted Produced Power* can be estimated using key environmental and operational variables. If the *Grid Produced Power (Measured Produced Power)* is significantly lower than the predicted value, it suggests that the wind turbine may be operating below its optimal efficiency.

While SCADA data provides a useful basis for this analysis, the methodology has notable weaknesses arising from challenges associated with accurately predicting power generation under varying environmental and operational conditions. These weaknesses include:

1. *Lack of Reliable SCADA Data:* The accuracy of the analysis depends heavily on the quality and consistency of the SCADA data collected. Poor or inconsistent data can significantly impact the reliability of the predictions.

2. *Inadequate Models for Predicted Power Calculation:* Developing reliable models that accurately calculate Predicted Produced Power while accounting for variations and imperfections in the collected data remains challenging. Imperfect models can give erroneous and misleading values of Predicted Produced Power.

3. *Defining Significant Difference:* Determining what constitutes a "significant difference" between predicted and measured power is complicated by the inherent uncertainties in the Predicted Produced Power. Without clear criteria for significance, it becomes difficult to reliably identify sub-optimal performance.

Addressing these weaknesses requires enhancing data quality, developing more advanced predictive models, and establishing robust criteria for evaluating discrepancies. Additionally, it is crucial to quantify uncertainties and integrate them into the decision-making process. Improving these aspects will result in more accurate performance assessments and more reliable, well-informed maintenance decisions.

2.2 Aim of the Research

Aim of the ongoing research is to develop a methodology for detecting sub-optimal operation of wind turbines by comparing Grid Produced Power (Measured Produced Power) with Predicted Produced Power. A key aspect of this research involves developing a systematic approach to account for uncertainties in the Predicted Produced Power, which arise from imperfections in SCADA data and limitations of predictive models.

2.3 Scientific Novelty and Importance of the Research

This paper presents a methodology for predicting wind turbine power output using SCADA data, incorporating a Possibilistic Approach to account for uncertainties arising from both model imperfections and measurement errors. The method integrates machine learning models, specifically XGBoost, to establish relationships between environmental inputs and power output, alongside a Fuzzy logic-based possibilistic framework that quantifies uncertainty through possibility distribution functions and interval analysis.

The Possibilistic Approach was chosen due to its key advantages over the Probabilistic Approach. First, it is well-suited for addressing epistemic uncertainties, particularly those arising from imprecise data or limited statistical information. This is especially relevant in contexts such as wind turbines, where constructing accurate probability distribution functions for environmental parameters is difficult due to sparse sensor data from remote or hard-to-reach locations. Second, the possibility measure tends to be more conservative than the probability measure, making it a valuable tool for supporting decision-making frameworks that emphasize zero-tolerance for errors.

The approach is demonstrated using publicly available data from an operational wind turbine, effectively capturing the influence of real-world data imperfections on prediction accuracy.

3 Application of SCADA Data for Fault Detection

3.1 Applications of SCADA Data for Predictive Maintenance

While SCADA systems are primarily designed for control and automation, they are closely connected to condition monitoring systems that focus on diagnostic and predictive analysis. Wind turbine SCADA systems collect extensive operational data, including parameters such as temperature, vibration, pressure, flow rate, and electrical metrics (e.g., current and voltage). This data can be effectively utilized for fault detection, encompassing various types of faults such as:

- Component degradation,
- Sensor failures,
- Operation beyond safe limits, and
- Grid-related issues.

Some faults can be directly identified through SCADA data. For example, sensor failures are often evident through irrational or out-of-range readings. However, other faults, such as gradual gear wear or structural degradation, may only be detected indirectly through changes in performance metrics or subtle deviations from expected behaviour [4].

The duration between fault initiation and potential failure can vary significantly:

- *Short-duration faults* (e.g., generator earth faults) may develop within seconds.
- *Long-duration faults* (e.g., gradual gear wear) can take weeks or months to manifest fully.

Due to the typically low-resolution nature of SCADA data, it is most effective for identifying faults with longer time spans. Recognizing this potential, many modern SCADA systems now incorporate real-time data analysis capabilities, using statistical and artificial intelligence (AI) techniques to enhance fault detection and diagnosis. While certain faults, such as sensor failures, can be directly detected, others may only be identified through indirect indicators or by applying advanced analytical techniques [4,5,6,7].

3.2 Analysing Wind Turbine Performance Using SCADA Data

The *Power Curve* of a wind turbine represents the unique relationship between the power generated and the environmental and operational conditions under which the turbine operates. It serves as a critical tool for evaluating and comparing turbine performance under various scenarios.

The power generated by a wind turbine is influenced by:

- *Technical Attributes:* Such as rotor radius, blade geometry, and drive train efficiency.
- *Environmental Attributes:* Including wind speed, air density, temperature, and turbulence intensity.
- *Operational Attributes:* Such as pitch angle, nacelle orientation, and the angle between the wind direction and nacelle.

These factors collectively determine the efficiency of power generation and are essential for developing accurate predictive models [4].

In a simplified power balance model, wind power is first converted into rotor power, which is then transformed into electrical power. The efficiency of converting wind power to rotor power depends on several factors, including wind speed, air density, blade geometry, and rotor size. Ideally, all the rotor power should be converted to electrical power through the drive train system, however, in practice, some energy is inevitably lost due to factors such as friction, vibration, and heat dissipation.

The overall energy balance, accounting for these losses, can be represented in a simplified way as [4]:

$$P_{Rotor} = P_{Electrical} + P_{Vibration} + P_{Thermal} \quad (1a)$$

$$P_{Rotor} - P_{Electrical} = P_{Vibration} + P_{Thermal} \quad (1b)$$

Where:

- P_{Rotor} = Rotor power
- $P_{Electrical}$ = Electrical power

- $P_{Vibration}$ = Vibration power
- $P_{Thermal}$ = Thermal power

Therefore, a significant discrepancy between the predicted rotor power (P_{Rotor}, calculated using models) and the measured electrical power ($P_{Electrical}$) indicates suboptimal performance. This discrepancy is often attributed to inefficient operation or increased energy losses resulting from factors such as vibration, friction, and heat generation or dissipation. Therefore, detecting significant deviations between Predicted Produced Power and Grid Produced Power can be effectively utilized for the following purposes [8]:

- *Suboptimal Operation Detection:*
    - Suboptimal performance, often caused by inefficient control mechanisms can be identified using the power curve.
    - Comparing power generation across a localized group of wind turbines can help identify individual units that are performing below expected standards, thereby facilitating targeted maintenance and optimization efforts.
- *Fault Detection:*
    - Although pinpointing the exact cause may be challenging, a substantial discrepancy between predicted and measured power generation can serve as an indicator of underlying faults, prompting further investigation.

Analysing deviations between predicted and actual power outputs provides valuable insights into the health and efficiency of wind turbine components, enabling early detection of faults and opportunities for performance improvement.

3.3 Uncertainties in Data

According to Bell [9], measurement uncertainty can be defined as the doubt that exists about the result of any measurement. This doubt arises because despite all precautions, measurements are inevitably affected by various imperfections and uncertainties. Uncertainties in SCADA measurements can arise from multiple sources, resulting in different types and classifications. These uncertainties can be:

- *Tangible (Quantifiable):* Such uncertainties can be measured and expressed numerically.
- *Intangible (Non-Quantifiable):* These are difficult to measure precisely and may only be qualitatively assessed.
- *Random:* Arising from unpredictable variations in measurement conditions.
- *Systematic:* Resulting from consistent biases or errors in measurement processes.

To ensure completeness and accuracy, measurements should be reported along with their associated uncertainties. A tangible uncertainty can be quantified using two key metrics: the interval, which represents the width of the margin of doubt or dispersion around the mean, and the confidence level, which indicates the probability that the "true" value falls within that margin. However, since measurement uncertainties are influenced by various factors, it is often challenging to account for all sources of uncertainty comprehensively [9].

Due to the complexities associated with categorizing uncertainties, various classification schemes have been proposed. However, there is no universally accepted framework, leading to inconsistencies and confusion. However, they are often categorized into two broad types [4,10]:

- *Aleatoric Uncertainty:* Aleatoric uncertainty arises from inherent randomness or natural variability within the measured parameter. It is generally quantifiable through repeated measurements and can be expressed using statistical measures such as mean and standard deviation, along with intervals and confidence levels. For example, variations in wind speed due to natural turbulence are a common source of aleatoric uncertainty.

- *Epistemic Uncertainty:* Epistemic uncertainty results from a lack of knowledge, incomplete data, or an imperfect understanding of the measurement process. Unlike aleatoric uncertainty, it affects all measured values consistently, making repeated measurements ineffective at reducing this type of uncertainty. It is often challenging to quantify precisely but can be evaluated using expert opinions, manufacturer specifications, historical data, or subjective judgment. Epistemic uncertainty can be further divided into the following subcategories:

    1. *Bias:* A consistent, systematic deviation from the true value, often introduced by faulty calibration or measurement techniques.
    2. *Inaccuracy:* The average difference between the measured value and the true value, indicating a general error in measurement.
    3. *Imprecision:* The spread or range within which measured values lie, indicating a lack of exactness.
    4. *Ignorance:* Arising from insufficient data or limited knowledge regarding measurement precision.
    5. *Incompleteness:* Occurring when relevant data is missing or unavailable.
    6. *Credibility:* Related to the reliability or trustworthiness of the measurement process, including factors such as calibration, installation, and operational competence.

Understanding and managing these uncertainties is essential for accurate fault detection, diagnosis, and predictive maintenance using SCADA systems. Recognizing the different types of uncertainties can help in formulating strategies for handling them during analysis. Evaluating their potential impacts can significantly enhance the reliability of condition monitoring and diagnostic processes.

Since epistemic uncertainty arises from knowledge gaps or incomplete data, it is typically evaluated using:

a. *Manufacturer's Specifications:* Guidelines and tolerances provided by equipment manufacturers.

b. *Past Experience:* Historical data and previously observed patterns.

c. *Expert Opinion:* Insights from skilled practitioners familiar with the measurement process.

d. *Subjective Judgment:* Personal assessment based on experience and intuition when objective data is insufficient.

3.4 Data Quality in SCADA System

Uncertainties are particularly problematic for wind turbines due to the substantial variations in environmental conditions. Most errors arise from two primary sources:

1. *Imperfections Caused by Sensors:* These imperfections occur for various reasons, including fluctuations in parametric values, instrument limitations (such as bias, noise, or drift), incorrect calibration, measurement location errors, and overall instrument degradation. They can be further categorized as:

    - Inherent Imperfections: In response to the changing environmental conditions sensors report values based on their response time, sampling rate, resolution, sensitivity, and statistical analysis. Each of these characteristics introduces unique uncertainties.

    - Acquired Imperfections: During operation, sensors are exposed to various environmental stressors such as impacts, wind force, temperature fluctuations, humidity, condensation, frosting or icing, vibrations, and the accumulation of oil, dirt, or salt. These factors contribute to gradual sensor degradation.

2. *Imperfections Caused by SCADA System:* SCADA systems typically record data at 1 to 10-minute intervals, meaning the recorded value is not an instantaneous measurement but rather a statistical estimate derived from predefined algorithms. This limitation can introduce errors, especially when rapid changes occur within those intervals.

To improve the reliability and accuracy of the SCADA data used for analysis, several corrective measures have been recommended [4,6]:

- *Use of High-Quality Sensors:* High-quality sensors should be robustly designed to withstand harsh environmental conditions and provide superior performance in terms of accuracy, precision, reliability, repeatability, and reproducibility. Sensors with improved structural integrity are less likely to degrade under adverse conditions.

- *Use of Multiple Data Streams:* Employing multiple and varied data streams enhances fault detection by cross-referencing results, thereby increasing detection probability. For example, using both vibration monitoring and debris analysis improves the reliability of detecting bearing faults. While redundancy offers benefits, the use of various sensors at different locations also improves detection probability. However, this approach may result in data overload, where the volume of collected data becomes too large to process efficiently. Additionally, the

"law of diminishing returns" indicates that deploying multiple sensors for the same task may not yield significant new information.

- *Use of Advanced Data Analytics Techniques:* Various methodologies have been developed to manage different types of uncertainties. Aleatoric uncertainty, arising from natural variability, is often addressed through Probabilistic Approaches such as statistical analysis and probability theory. Conversely, Epistemic uncertainty, resulting from knowledge gaps or incomplete data, is more effectively managed through Possibilistic Approaches, including fuzzy logic, expert systems, and Bayesian networks.

Implementing these corrective measures can significantly enhance data quality and reliability, thereby improving the accuracy of fault detection, diagnostics, and prognostics in wind turbine condition monitoring systems.

4 Possibilistic Approach for Handling Epistemic Uncertainties

4.1 Representation by Possibility Distribution Function

In this work, aleatoric uncertainties have not been dealt with because values of a parameter varies with time, hence, it is not possible to measure the same parameter multiple times to enable the statistical evaluation of the uncertainties.

One effective method for addressing epistemic uncertainties is through the use of the *Possibilistic Approach*. This approach represents data as membership functions rather than precise numerical values. A Fuzzy variable X is described by a *Fuzzy Membership Function*. This membership function can also be interpreted as a *Possibility Distribution Function*. **Fig. 2a** illustrates how a Possibility Distribution Function, $\Pi(x)$, represents the variable X through the degree of compatibility or truth associated with different values.

The Possibility Distribution Function, $\Pi(x)$, maps the values of the input variable X to a range between 0 and 1, where [11,12]:

- $\alpha(x) = 0$: The value is completely incompatible or impossible.
- $\alpha(x) = 1$: The value is fully compatible or completely true.
- $0 < \alpha(x) < 1$: The value is partially compatible, indicating varying degrees of possibility.

Unlike *Probability Density Functions (PDFs)*, Possibility Distribution Functions do not assign preference to any specific value within the fuzzy interval. This feature is advantageous in situations when dealing with incomplete, sparse, or vague data, where conventional statistical methods may struggle to provide reliable results.

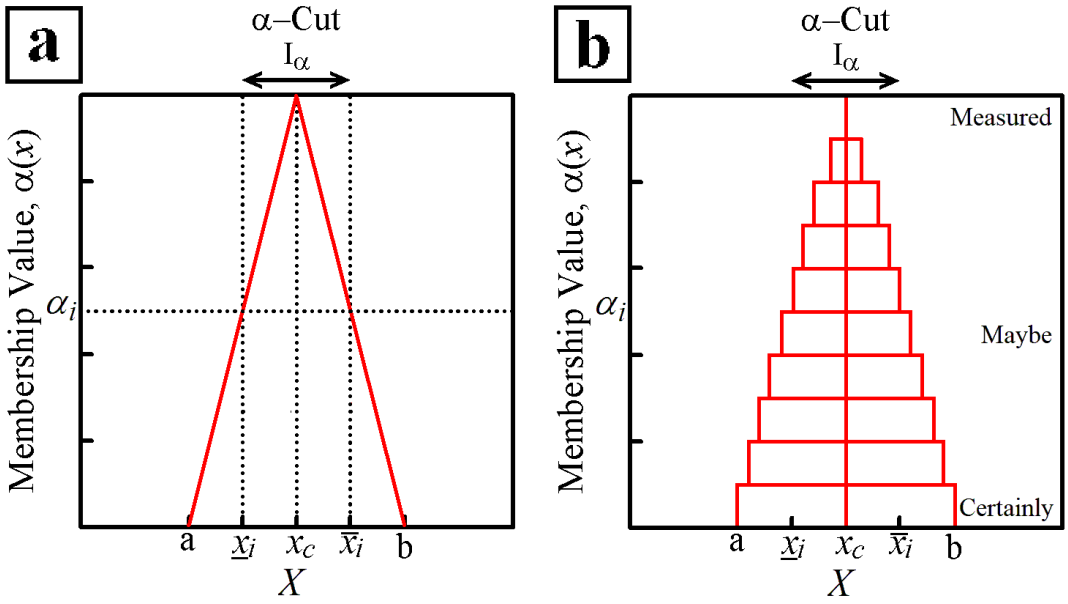

FIGURE 2. Illustration of a fuzzy subset [13, 14].

A feature of the Possibilistic Approach is the use of α–$cuts$ to represent Possibility Distribution Function. An α–cut of a Possibility Distribution Function X, denoted by X_α, is a crisp set containing all elements of X whose membership value is greater than or equal to a specified threshold α. Mathematically, this can be expressed as [11,12]:

$$X_\alpha = [\underline{x}, \overline{x}]_\alpha = \{x \in X | \underline{x} \leq x \leq \overline{x}\} \qquad (2)$$
$$\alpha \in [0,1]$$

Where:

- $\underline{x}$ = Lowest real number value of the interval
- $\overline{x}$ = Highest real number value of the interval

The value of α can be in the range [0,1]. As α increases, the interval $[\underline{x}, \overline{x}]$ becomes narrower, representing values with higher likelihood. Conversely, as the interval becomes narrower, the certainty that the true value lies within that interval decreases.

The α–cut representation allows for the extension of various properties of crisp sets to Fuzzy sets. By incrementally changing the value of α, a nested family of sets is generated **Fig. 2b**. These sets form a hierarchy where higher α-levels correspond to smaller intervals with higher degrees of possibility.

This concept is particularly useful because it allows traditional interval analysis techniques to be applied to Fuzzy sets. When performing arithmetic operations on Fuzzy variables, the interval bounds generated by the α–$cuts$ can be manipulated using established rules for interval arithmetic.

Properties of crisp sets that can be extended to Fuzzy sets through the use of α–$cuts$ are referred to as *cutworthy properties*. Such properties include operations like union, intersection, and complement, which can be adapted to work with fuzzy sets through the α–cut approach [11].

The use of α–$cuts$ offers a practical means of applying interval analysis techniques to Fuzzy sets, enhancing the ability to model and process uncertain or imperfect SCADA data. This methodology is particularly useful for complex systems like wind turbines, where data imperfections are common. The advantages include:

- *Handling Imperfections:* By representing data as Possibility Distribution Functions rather than precise points, the approach can effectively handle vague, inconsistent, or incomplete information.
- *Compatibility with Interval Analysis:* The use of α–$cuts$ allows well-established interval analysis techniques to be applied to fuzzy data.
- *Scalability:* By varying the α-level, it is possible to explore different levels of certainty and possibility, providing a flexible framework for uncertainty analysis.

Despite its numerous advantages, the Possibilistic Approach has several limitations that can affect its practical application, especially in scenarios requiring precise and economically efficient decision-making. Some of the key weaknesses include [11,12,15]:

- *Imprecise Results:* The reliance on Possibility Distribution Functions instead of precise numerical values can result in vague or overly conservative recommendations. When data is not well-defined or incomplete, the model may produce results that are too broad to be actionable or economically feasible. This lack of precision can limit the approach's effectiveness in providing clear guidance for maintenance or operational adjustments.
- *Loss of Information During Conversion:* One significant drawback of the Possibilistic Approach is the potential loss of information when converting inspection or monitoring data to Possibility Distribution Functions. During this transformation, certain nuances or details within the original data set may be overlooked or oversimplified, leading to less accurate or meaningful results. This loss of detail can negatively impact the quality of the assessment and reduce the overall reliability of the analysis.
- *Violation of Consistency in Arithmetic Operations:* The propagation of Possibility Distribution Functions through arithmetic operations can lead to inconsistencies. Unlike probability theory, which adheres to strict mathematical rules during data manipulation, Possibility Theory may produce results that are inconsistent or counterintuitive when complex calculations are performed. This violation of consistency can compromise the credibility and robustness of the analysis, particularly when handling large datasets or intricate systems.
- *Lack of Standardization:* Unlike probabilistic methods that are governed by well-established mathematical frameworks and guidelines, the Possibilistic Approach lacks universally accepted standards for implementation. This lack of formalization can result in subjective and inconsistent application, especially when

determining Possibility Distribution Functions or evaluating possibility distribution functions. The absence of standardized methodologies makes it difficult to compare results across different studies or systems, reducing the approach's generalizability.

4.2 Transforming Probability Distributions to Approximate Possibility Distributions

According to the *Consistency Principle*, a probability distribution function $P(x)$ can be transformed to a number of possibility distributions $\Pi(x)$, provided $\Pi(x) \geq P(x)$. Out of all the possibility distributions the one which is *maximally specific*, i.e. the possibility distribution that most closely preserves the amount of information of the probability distribution, is the *Optimal Possibility Distribution Function*. Developing and utilizing the optimal possibility distribution can be a complex and computationally intensive task. To address this challenge, simplified forms of possibility distribution functions have been introduced. Common examples include the *triangular*, *trapezoidal*, *truncated pseudo-triangular* and the *Gaussian distribution*. These simplified models provide a practical means of approximating the original distribution while maintaining essential characteristics [15,17,18].

Fig. 3 illustrates the general shapes of these simplified distribution functions, which are typically derived by transforming existing probability distributions into their corresponding possibility distributions.

A possibility distribution function obtained through such a transformation can be described using four key parameters:

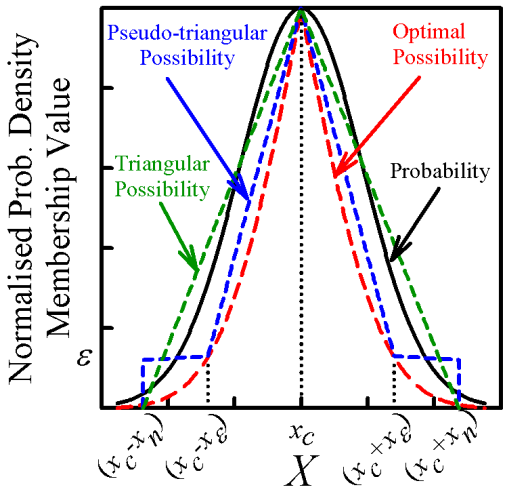

FIGURE 3. Graphical illustration of the normal probability distribution, the transformed optimal possibility distribution and the truncated pseudo-triangular possibility distribution [15,16,17].

$$For\ confidence\ level = 0.99:$$
$$\varepsilon = 0.12, x_c = x_m, x_n = x_c \pm 2.58\sigma$$

- x_c: the core value representing the peak of the possibility distribution,
- ε: the minimal possibility level considered significant,
- x_ε: the nominal limit beyond which the possibility rapidly diminishes (equal to ε),
- x_n: the threshold beyond which the possibility value becomes negligible.

For unimodal distributions, the value of x_c aligns with the mode (the most probable value, which coincides with the mean in a symmetric normal distribution) of the original Probability Density Function (PDF), denoted as x_m. Therefore:

$$x_c = x_m \tag{3}$$

The determination of x_n depends on whether the underlying distribution is bounded or unbounded:

- For *bounded distributions* – such as the triangular or uniform distribution – x_n represents the finite support of the distribution, i.e., the maximum value within which the distribution is defined.
- For *unbounded distributions*, such as the normal or lognormal distributions, it is not feasible to consider the entire domain due to their infinite tails. In these cases, x_n must be carefully selected to represent a sufficiently wide yet practical interval of the distribution.

Since unbounded distributions theoretically extend to infinity, it is neither practical nor necessary to represent the entire domain in the corresponding possibility distribution. A widely used method for approximating possibility distribution parameters from an unbounded probability distribution, such as the normal distribution, involves transforming it into a *Triangular Possibility Distribution Function*, $\Pi(x)$, where [15]:

$$a = x_c - k\sigma \tag{4a}$$

$$b = x_c + k\sigma \tag{4b}$$

$$\sigma = \text{Std. dev. of the original prob. dist.}$$

By defining a confidence interval to capture a substantial portion of the total probability mass, a significant interval – determined by a confidence level – is selected to effectively approximate the distribution's behaviour. For instance, selecting a *99% confidence level* provides an interval that encompasses almost the entire area under the normal curve, thereby retaining the most relevant portion of the distribution. Under this condition, $k = 2.58$. Using these bounds, a Triangular Possibility Distribution Function, $\Pi(x)$ can be constructed as:

$$\Pi(x) = \begin{cases} 0 \ if \ x \le a \ or \ x \ge b \\ \dfrac{x-a}{x_c-a} \ if \ a < x < x_c \\ \dfrac{b-x}{b-x_c} if \ x_c < x < b \\ 1 \ if \ x = x_c \end{cases} \tag{5}$$

This triangular function is a simplified representation that captures the core (mode), support (bounded interval), and spread (confidence-based width) of the original distribution. The result is a computationally efficient and interpretable possibility distribution.

However, this simplification comes with certain trade-offs. The transformation from a normal Probability Distribution Function (PDF) to a Triangular Possibility Distribution Function inherently leads to a loss of information. This occurs because the triangular function essentially approximates the original distribution as having Uniform Possibility Density within its core, disregarding the variations in probability that are present in the normal distribution's bell shape. Consequently, although this transformation is practical, it lacks the fidelity of the original distribution and may misrepresent the actual uncertainty in more sensitive analyses.

5 Methods

5.1 SCADA Data Description

To evaluate the feasibility of the proposed methodology, SCADA data provided by the energy company EDP (2016) has been utilized. This dataset comprises data collected from four horizontal-axis wind turbines located off the western coast of Africa. The data spans a two-year period (2016 and 2017) with measurements recorded at 10-minute averaging intervals. The datasets include values for 76 different parameters, covering various aspects of

turbine operation and performance. Additionally, an associated dataset containing meteorological conditions recorded at the same time intervals is provided, along with failure logs detailing timestamps, damaged components, and related remarks [3].

For this analysis, Turbine Number 7 (T07) has been selected. The variables used in the calculation of the power curve are listed in **Table 1**.

To enable model development and testing, the dataset was divided into two independent subsets: the 2016 data was used for training, while the 2017 data served as the test set to evaluate the model's predictive performance.

The total number of recorded instances for this turbine is 52,445 for 2016 and 52,294 for 2017. During the same time total number of recorded instances for metrological data is 52698 and 34832 for 2016 and 2017, respectively. Since 2016 data was used for training, the data row that does not contain all the values were dropped. It is expected that the small number of dropped rows will not make any significant effect on training.

In 2016 there were three recorded instances of failures. Since these failures were for short durations and not relevant for the analysis: high bearing temperature and high transformer temperature, these rows were retained.

Fig. 4 shows the relationships between ambient wind speed, generator RPM, and grid-produced power have been examined.

Fig. 4a shows the effect of Ambient Wind Speed on Generator RPM. The plot reveals a distinct relationship between *Ambient Wind Speed* and *Generator RPM*, which can be divided into three regions:

1. *Low Ambient Wind Speed (Ambient Wind Speed < 4 m/s):* When Ambient Wind Speed is below the *Cut-In Wind Speed* (4 m/s), the frequency of Generator RPM readings below 300 rpm is high.

TABLE 1. Selected variables used for developing the model.

Variable	Short Variable Name	Original SCADA Name	Description	Units
Timestamp			10-minute resolution	
Ambient Temperature	Amb_Temp	Amb_Temp_Avg	Average ambient temperature	ºC
Ambient Humidity	Amb_Humidity	Avg_Humidity	Average ambient relative humidity	%
Ambient Pressure	Amb_Pressure	Avg_Pressure	Average ambient pressure	millibar
Ambient Wind Speed	Amb_Wind_Speed	Amb_WindSpeed_Avg	Average windspeed within average timebase	m/s
Generator RPM	Gen_RPM	Gen_RPM_Avg	Average generator shaft / bearing rotational speed	rpm
Grid Produced Power (Measured Produced Power)	Grid_Prod_Power	Grd_Prod_Pwr_Avg	Power average	kW

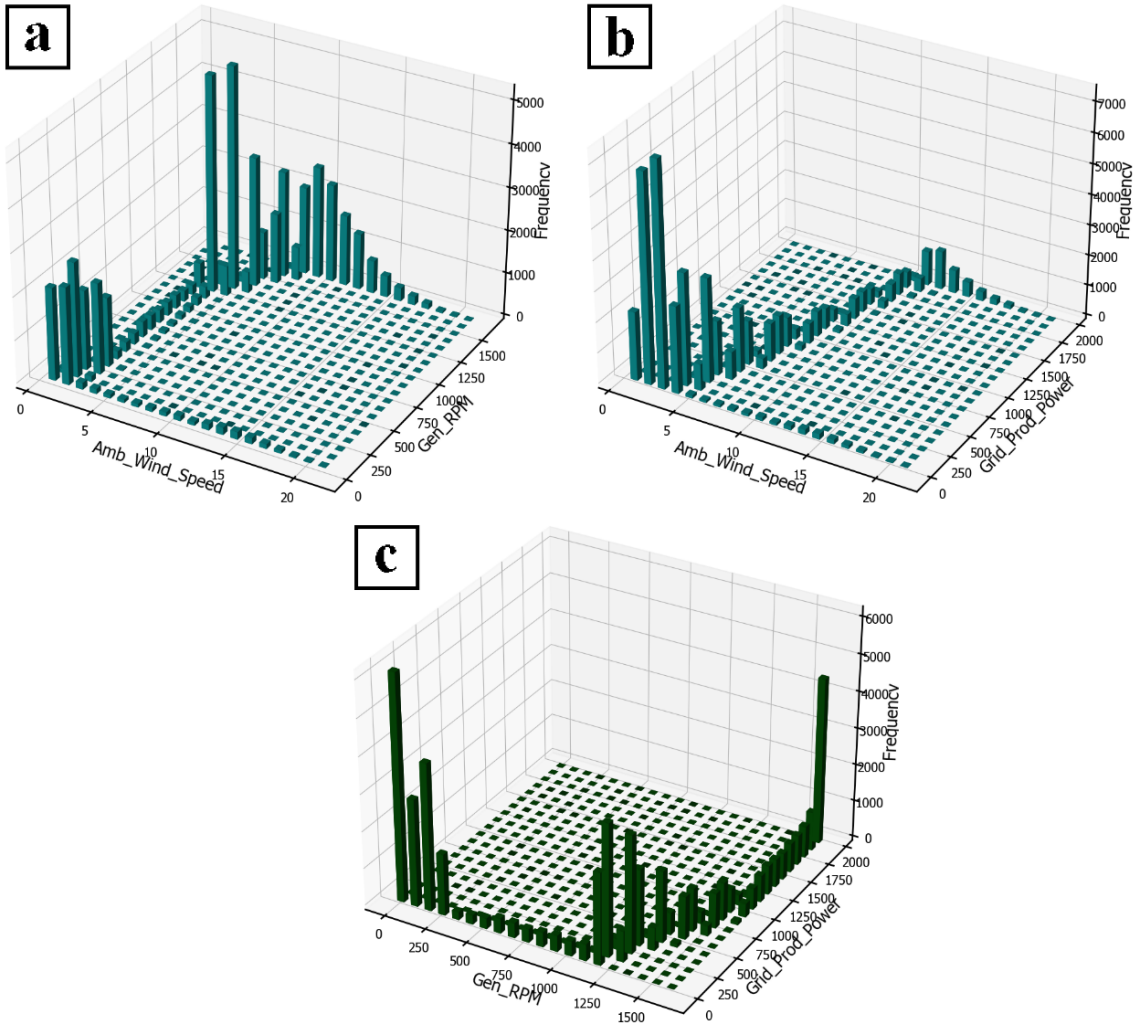

FIGURE 4. Relationships between Ambient Wind Speed, Generator RPM and Grid Produced Power [14].

2. *Transition Region (4 m/s < Ambient Wind Speed < 12 m/s):* When wind speed ranges between 4 m/s and 12 m/s (*Rated Wind Speed*) the wind turbine adjusts its blade pitch angle to reach High RPM Region. Hence, there are fewer readings in this region.
3. *High Ambient Wind Speed Region (12 m/s < Ambient Wind Speed RPM < 25 m/s):* When wind speed is above the *Rated Wind Speed* (12 m/s), the Generator RPM increases from approximately 1250 rpm to 1650 rpm. However, once the wind speed exceeds 12 m/s, the wind turbine stabilizes the Generator RPM at around 1650 rpm to ensure optimal performance and prevent mechanical stress.

Fig. 4b shows the effect of Ambient Wind Speed on Grid Produced Power. The relationship between Ambient Wind Speed and Grid Produced Power demonstrates the following patterns:

1. *Low Ambient Wind Speed (Ambient Wind Speed < 4 m/s):* When Ambient Wind Speed is below the Cut-In Wind Speed (4 m/s), Grid Produced Power is generally negative or less than 275 kW.

2. *Transition Region (4 m/s < Ambient Wind Speed < 12 m/s):* The trend highlights a strong positive correlation between wind speed and power generation until the turbine reaches its rated capacity. As Ambient Wind Speed increases, Grid Produced Power rises, reaching approximately Rated Power (2000 kW) at the Rated Wind Speed (12 m/s).
3. *High Ambient Wind Speed Region (12 m/s < Ambient Wind Speed RPM < 25 m/s):* Between the Rated Wind Speed and *Cut-off Wind Speed* the wind turbine maintains Rated Power generation.

Fig. 4c shows the effect of Generator RPM on Grid Produced Power. This plot shows a clear relationship between Generator RPM and Grid Produced Power:

1. *Low RPM Region (Generator RPM < 1250 rpm):* Grid Produced Power remains negligibly low.

2. *Transition Region (1250 rpm < Generator RPM < 1650 rpm):* Grid Produced Power increases linearly, reaching approximately 750 kW as the Generator RPM rises within this range. This indicates a direct correlation between RPM and power output within this interval.

3. *High RPM Region (Generator RPM ≈ 1650 rpm):* The Generator RPM reaches its upper operational limit, producing the maximum achievable power. Beyond this point, the turbine maintains a stable RPM to prevent mechanical stress and ensure efficient power generation.

5.2 Data Pre-processing

Data pre-processing is a critical step in the development of a Machine Learning model, aimed at improving data quality and ensuring algorithms perform effectively. It involves correcting or removing vague, inconsistent, irrational, duplicate, or missing values that may otherwise compromise model accuracy and reliability.

In the case of wind turbine SCADA data, the dataset often includes data points that deviate significantly from expected patterns, particularly the power curve, and are therefore classified as “outliers.” These outliers can arise due to various explainable factors. For this work, outliers have been identified based on the following rules [3,14]:

Outlier Rule 1. *Generator RPM = 0 when Ambient Wind Speed => 4 m/s.* Although the wind speed is above the Cut-In Wind Speed (4 m/s), the rotor remains stationary because the wind turbine is in a shutdown state. This may occur due to various reasons, including grid conditions or maintenance activities.

Outlier Rule 2. *Grid Produced Power <= 0 when Ambient Wind Speed < 4 and Generator RPM > 0*. When the rotor RPM is low, the power generated is insufficient to meet the power consumption required for operation. The deficit is compensated by drawing power from the grid, resulting in negative or zero power output.

Outlier Rule 3. *Grid Produced Power <= 0 when Ambient Wind Speed => 4 & Generator RPM > 0.* Even though the wind speed exceeds the Cut-In Wind Speed (4 m/s) and the rotor is rotating, power generation does not occur because the turbine is "free-wheeling" in a shutdown state. This condition could be due to grid issues, maintenance operations, or other shutdown scenarios.

Apart from the predefined outlier rules, additional anomalous data points need to be removed. These points are often recorded during transitions between normal operation and shutdown states or vice versa. This shutdown often takes place when the grid is saturated. They appear scattered in the dataset and can be effectively identified using DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a density-based clustering algorithm known for its robustness in handling noise and discovering clusters of arbitrary shapes [19].

Two specific clustering rules have been applied [3,14]:

DBSCAN Clustering Rule 1. *Ambient Wind Speed, Grid Produced Power, eps value = 2, min_samples value = 10*. To identify outliers based on the relationship between wind speed and grid-produced power

DBSCAN Clustering Rule 2. *Ambient Wind Speed, Generator RPM, eps value = 3.45, min_samples value = 10*. To detect anomalies by examining the relationship between wind speed and generator RPM

The application of these clustering rules helps to effectively isolate and eliminate noise points, thereby enhancing the integrity of the dataset.

The impact of the data pre-processing and outlier removal is illustrated in **Fig. 5**, which compares the dataset before and after cleaning. Eliminating outliers helps isolate data points that follow the power curve, ensuring that the Machine Learning model receives high-quality inputs. This refinement enhances the model's ability to accurately capture and represent the power curve, leading to more reliable predictions.

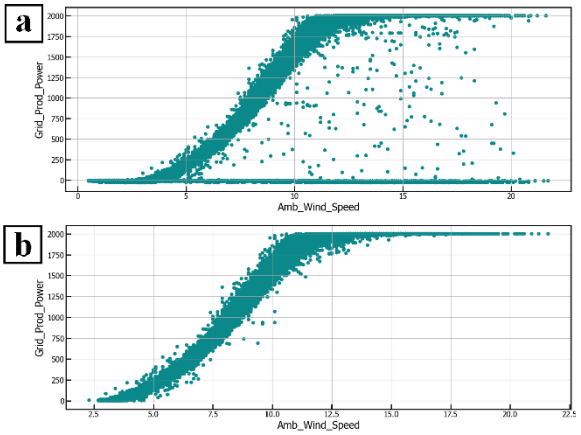

FIGURE 5. Plot of power generated versus wind speed using SCADA data. (a) Using raw data (b) Using data after removing outliers [14].

5.3 Flowchart for Predicting Produced Power

To develop an effective predictive model, it is essential to understand the process in terms of its structure, environment, and operational dynamics.

Electrical power ($P_{Electrical}$) produced can be given as [4,6]:

$$P_{Electrical} = \left(\frac{1}{2}\rho A U^3\right) \times C_P(\lambda, \beta) \times \eta \quad (6)$$

Where:

- $P_{Electrical}$ = Electrical power
- ρ = Air density, which is dependent upon the ambient temperature, humidity and pressure
- A = Rotor disc area
- U = Air velocity
- $C_P(\lambda, \beta)$ = Rotor power coefficient, it expresses the recoverable fraction of wind power and is a function of λ (tip speed ratio) and β (blade pitch angle)
- η = Drive train efficiency ($generator\ power/rotor\ power$), (mechanical & electrical)

The maximum theoretically possible rotor power coefficient, $C_{P,max}$ also called the Betz limit, can be determined to be 0.59. The actual value of $C_P(\lambda, \beta)$ is below the Bentz limit and is dependent upon technical features of the turbine and environmental factors [20].

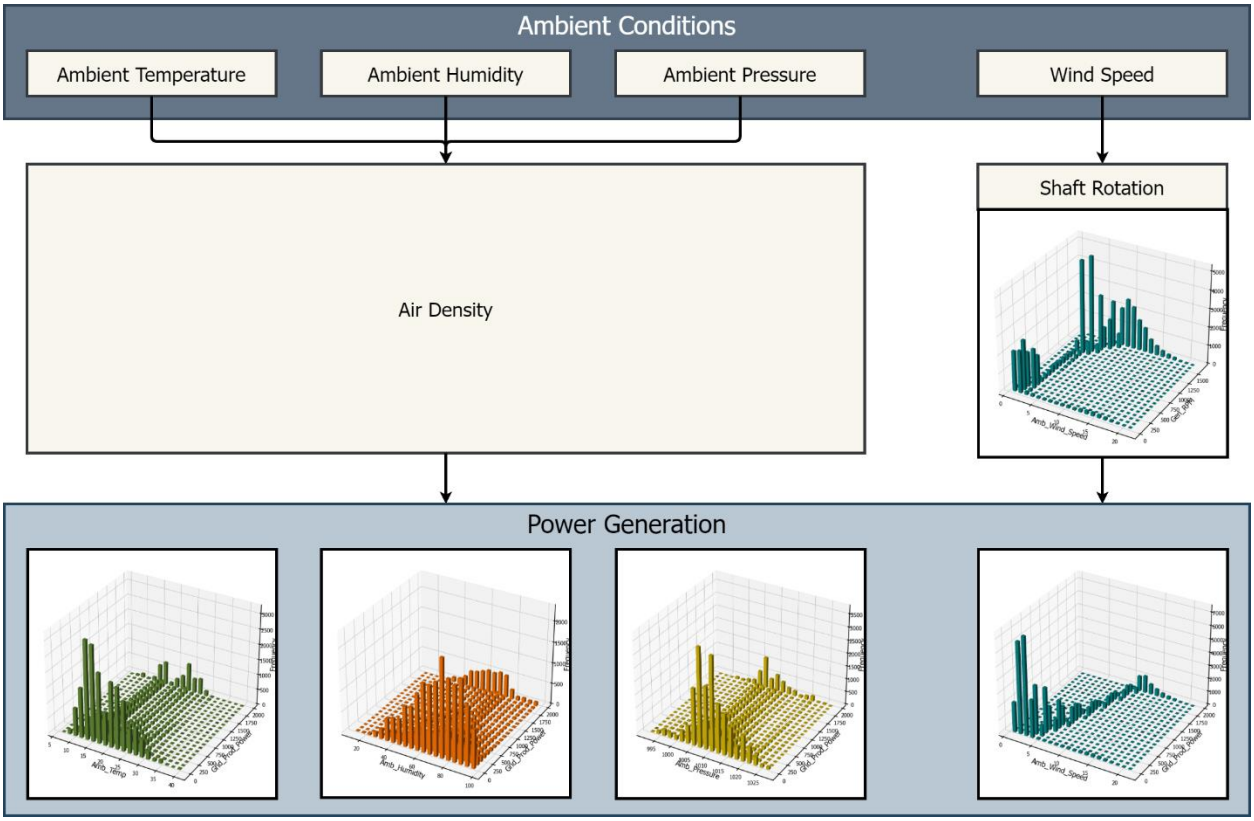

FIGURE 6. Flowchart showing influence of variables on the calculation of produced power [14].

This analysis reveals that Grid Produced Power has a strong correlation with Ambient Wind Speed, making it a pivotal factor for predictive modelling. In contrast, other environmental variables such as Ambient Temperature, Ambient Humidity, and Ambient Pressure exhibit only weak correlations with Grid Produced Power. Recognizing these differences helps in selecting the most relevant inputs for the predictive model, thereby enhancing the accuracy and reliability of power production assessments.

This insight serves as the foundation for developing a simplified flowchart to calculate Predicted Produced Power, as illustrated in **Fig. 6**. The flowchart highlights the relationships between various environmental variables and power generation, emphasizing the dominant influence of Ambient Wind Speed on power output.

5.4 Selection of Machine Learning Algorithms

In this project, several machine learning algorithms were evaluated for developing a robust predictive model. The models considered include [3, 14]:

1. Linear Models : Linear Regression (LR), Lasso, Ridge and Bayesian Ridge Regression
2. Tree-based Models : Decision Trees, Random Forest (RF)
3. Boosting Models : AdaBoost, XGBoost and LGBoost
4. Support Vector Regression (SVR)

Among these, XGBoost was selected as the final model based on its overall performance and practical advantages. It demonstrated high level of goodness-of-fit (RMSE = 186, R^2 = 0.93, MAE = 127) indicating a strong predictive accuracy. Additionally, XGBoost was favoured for its computational efficiency and relatively simple implementation (**Fig. 7**).

6 Uncertainties Due to Model

6.1 Probability Density Function of Modelling Error

Fig. 8 presents a plot of the error in Predicted Power Production as a function of the Grid Produced Power. The graph illustrates that, across the entire range of Grid Produced Power values, the prediction errors are generally centred around a mean close to zero.

When the distribution of error values is analysed, it closely follows a normal distribution with a mean of 7.44 and a standard deviation of 147.80. This suggests that the prediction model maintains a balanced performance, with errors symmetrically distributed around the mean.

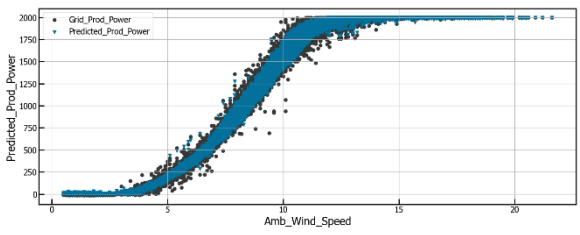

FIGURE 7. Effect of Ambient Wind Speed on Predicted Produced Power using XGBoost.

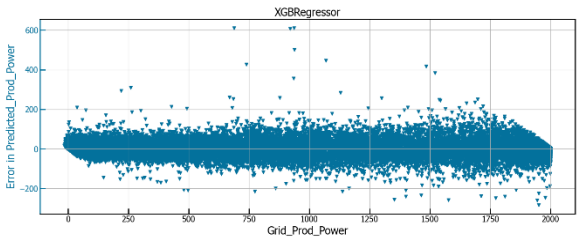

FIGURE 8. Effect of Grid Produced Power on Error in Predicted Produced Power (difference between Grid Produced Power and Predicted Produced Power) calculated using XGBoost.

6.2 Representation of Modelling Error as Possibility Distribution Function

As discussed in **Section 4.2**, the Normal Probability Distribution Function (PDF) that characterizes the error in Predicted Power Production can be systematically transformed into a Triangular Possibility Distribution Function (TPDF). This transformation provides a simplified yet practical representation of uncertainty, particularly useful in possibilistic analysis where crisp probabilities are replaced by degrees of possibility.

In this context, the error distribution – originally modeled as a normal distribution with a mean (mode) of 7.44 and a standard deviation of 147.80 – has been converted into a triangular possibility distribution. This possibility function uses the mode of the original distribution (7.44) as the mean error value due to model. The support of the triangle, which defines the full range of plausible error values, is calculated using a ±2.58σ interval around the mode. This corresponds to a 99% confidence level, ensuring that the majority of the probability mass from the original normal distribution is captured within the possibility framework.

Mathematically, the triangular possibility distribution is defined by:

- *Mode (peak):* 7.44
- *Lower bound (a):* 7.44−2.58×147.80
- *Upper bound (b):* 7.44+2.58×147.80

This transformation allows the model to account for uncertainty in a more interpretable and computationally efficient way, while still preserving the essential characteristics of the original error distribution.

An advantage of using a Possibility Distribution Function to represent uncertainty from modeling errors is that the possibility measure is inherently more conservative than the probability measure, making it well-suited for decision-making frameworks that prioritize zero-tolerance for errors.

6.3 Possibility of Error Due to Model Imperfection

Fig. 9 illustrates a comparison between the actual Grid Produced Power and the Predicted Produced Power computed at an $\alpha\text{–}cut$=0, over a continuous 24-hour period on 10th November, 2017. This $\alpha\text{–}cut$ level represents the maximum uncertainty scenario within the Possibilistic Framework, incorporating the widest possible intervals for model output.

The plot reveals that the majority of the measured power values lie within the outer bounds defined by the $Certain - Min$ and $Certain - Max$ estimate The fact that observed values remain largely within these boundaries shows that the possibilistic model effectively captures the range of possible outcomes. It demonstrates that the model is capable of accounting for potential imperfections in the predictive model itself.

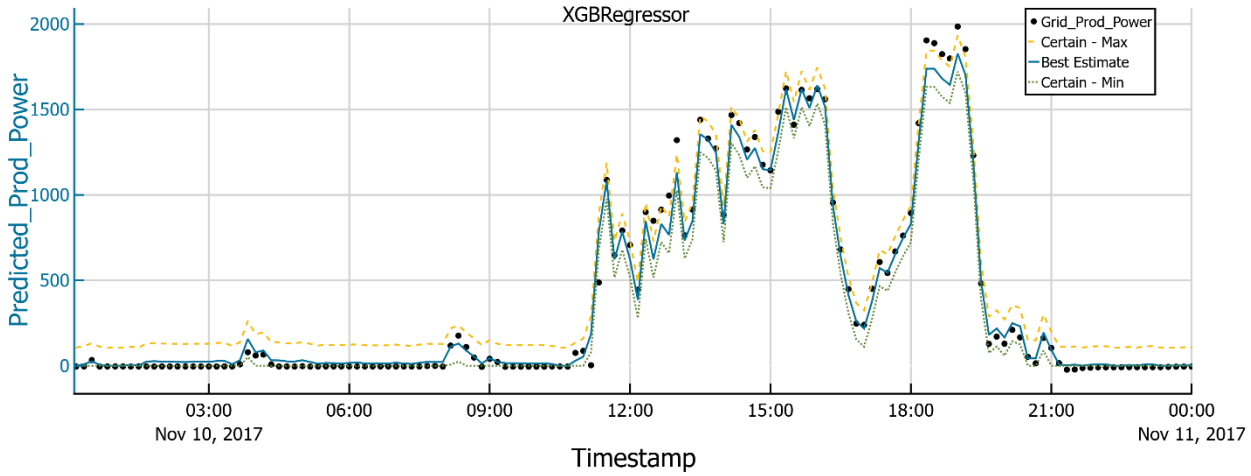

FIGURE 9. Plot of Grid Produced Power and Predicted Produced Power incorporating uncertainties due to model at $\alpha{-}cut$=0 for a 24 hour duration (10th November, 2016).

7 Uncertainties Due to Measurement

7.1 Representation of Input Variables as Possibility Distribution Functions

As outlined in **Section 3**, SCADA data is inherently affected by various imperfections, including sensor noise, inconsistencies, inaccurate readings, and missing values. These imperfections can significantly impact the reliability and accuracy of any data-driven model if not properly addressed. Developing a robust predictive model therefore requires careful pre-processing and error-handling strategies to mitigate the influence of such flaws.

Although a substantial portion of these data issues can be identified and corrected during the training phase – through methods such as data cleaning, normalization, and outlier detection – additional sources of error may still emerge during the model's deployment. In this study, the model was trained on data from 2016 and tested on data from 2017.

During this test phase, the model may encounter previously unseen patterns, shifts in turbine behaviour, or subtle inconsistencies not captured in the training data. As a result, these residual uncertainties and imperfections must be carefully considered and quantified to ensure that the model remains both accurate and resilient under real-world conditions.

In the Possibilistic Approach, instead of using fixed numerical values for environmental variables such as Ambient Temperature, Humidity, Pressure, Wind Speed, and Power Coefficient as recorded by SCADA, the approach models these variables as Triangular Possibility Distribution Functions.

A Possibility Distribution Function for a variable is constructed by stacking multiple intervals corresponding to different α-levels. The process begins with the bottom layer, where $\alpha = 0$. At this level, the interval range is defined as:

$[(measured\ value - estimated\ limit\ value), (measured\ value + estimated\ limit\ value)]$

In the absence of a detailed study to precisely quantify the interval, the estimated limit values used for calculations are derived from existing literature and practical experience. For instance, the response time and uncertainty associated with a measurement recorded by a cup anemometer depend on various factors such as its construction (e.g., dimensions, weight) and the degree of deterioration over time (e.g., friction caused by corrosion).

Under ideal test conditions, a newly calibrated anemometer may exhibit an inaccuracy of approximately 2%. However, during actual operational conditions, this inaccuracy is likely to increase due to factors such as corrosion, wear, misalignment, dust deposition, and other environmental influences [4].

Therefore, at $\alpha = 0$ (representing the interval within which the expected value is considered to "certainly" lie), the estimated limits around the measured values have been determined based on practical estimates and previous experience. These intervals account for both the inherent inaccuracies of the instruments and the additional uncertainties introduced by operational degradation.

- Ambient Temperature : ±1.0^{o}C
- Ambient Humidity : ±1.0%
- Ambient Pressure : ±1.0 milli-bars
- Ambient Wind Speed : ±0.5 m/s
- Power Coefficient : 0.45 ±0.05

7.2 Calculation of Predicted Produced Power Accounting for Measurement Errors

The calculation process within the Possibilistic Framework involves performing computations over interval values at various levels of certainty, represented by α–$cuts$. For each selected value of α, ranging from 0 to 1, the corresponding interval for each input variable is determined. The overall methodology consists of the following key steps

1. *Interval Generation:* To initiate the analysis, a specific α value within the range [0,1] is selected. For this value, the corresponding α–cut of each fuzzy number is determined, yielding an interval representation for each variable. Higher α values (closer to 1) correspond to narrower intervals, reflecting greater certainty. Conversely, lower α values (closer to 0) result in wider interval, indicating greater uncertainty in the variable's value.

2. *Combination of Input Intervals:* To thoroughly explore the output range, various combinations of input interval values are systematically analysed. Each combination corresponds to a particular set of input conditions at a given α–cut. The evaluation of these combinations, as detailed in **Table 2**, helps quantify the output's sensitivity to input uncertainty and ensures comprehensive coverage of all plausible input scenarios.

3. *Calculation of Output Intervals:* At each α–cut, the output variable – Predicted Produced Power in this case – is computed by evaluating the minimum and maximum values of the output function over all possible combinations of input intervals. These calculations are performed using the trained XGBoost model. This step ensures that the full range

of feasible outcomes is considered for the selected α-level, thereby capturing the propagation of uncertainty through the model.

4. *Stacking of Intervals to Construct the Possibility Distribution:* The intervals computed across all α–$cuts$ are then stacked to form the complete Possibility Distribution Function (PDF) of the output variable. This stacking results in a comprehensive depiction of uncertainty, ranging from the most uncertain scenarios (wide intervals at $\alpha = 0$), to the most certain predictions (narrow intervals at $\alpha = 1$). This stacking process creates a comprehensive representation of the variable's uncertainty, providing a full spectrum of possibilities from the most uncertain (broadest interval) to the most certain (narrowest interval).

By utilizing Possibility Distribution Functions, the model effectively captures and incorporates the uncertainties inherent in SCADA-recorded environmental variables. The application of intervals enables a more adaptable and realistic representation of uncertain data – particularly valuable when dealing with imprecise, inconsistent, or sparse measurements.

This approach allows the prediction process to account for varying degrees of uncertainty arising during measurement, enhancing the model's robustness and reliability, even in scenarios where precise input data is unavailable. Moreover, it ensures that the effects of measurement errors, and other data imperfections are explicitly considered and appropriately managed throughout the analysis.

TABLE 2. Possible combinations of interval values used for calculating Predicted Produced Power.

Combination	Ambient Wind Speed	Ambient Temperature	Ambient Pressure	Ambient Humidity
Combination_1	Min	Min	Min	Min
Combination_2	Min	Min	Min	Max
Combination_3	Min	Min	Max	Min
Combination_4	Min	Min	Max	Max
Combination_5	Min	Max	Min	Min
Combination_6	Min	Max	Min	Max
Combination_7	Min	Max	Max	Min
Combination_8	Min	Max	Max	Max
Combination_9	Max	Min	Min	Min
Combination_10	Max	Min	Min	Max
Combination_11	Max	Min	Max	Min
Combination_12	Max	Min	Max	Max
Combination_13	Max	Max	Min	Min
Combination_14	Max	Max	Min	Max
Combination_15	Max	Max	Max	Min
Combination_16	Max	Max	Max	Max

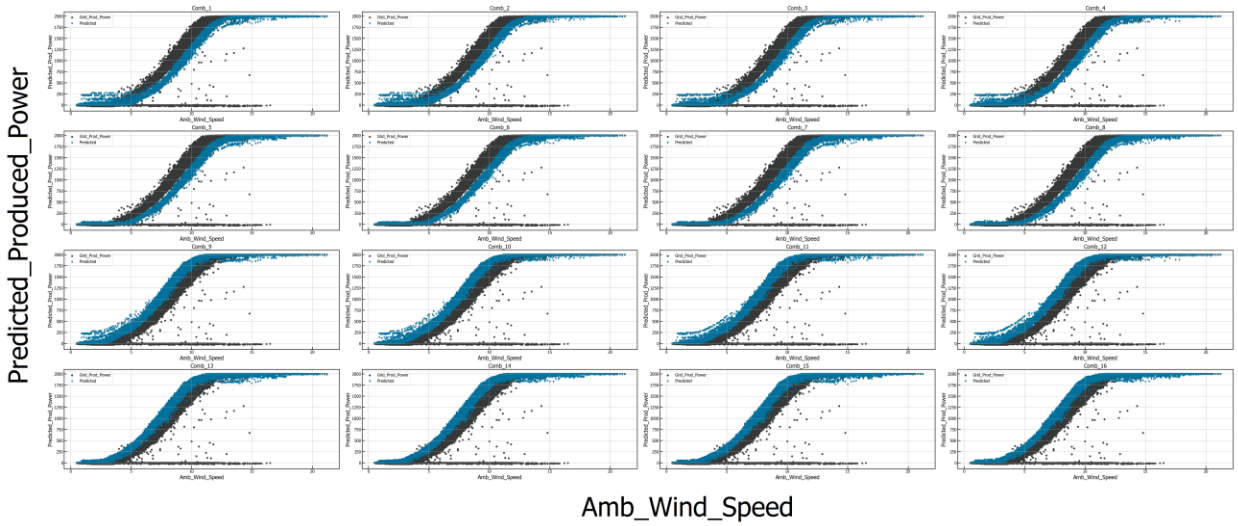

FIGURE 10. Impact of Measurement Error on Predicted Produced Power for various interval combinations listed in Table 2 at α–cut=0.

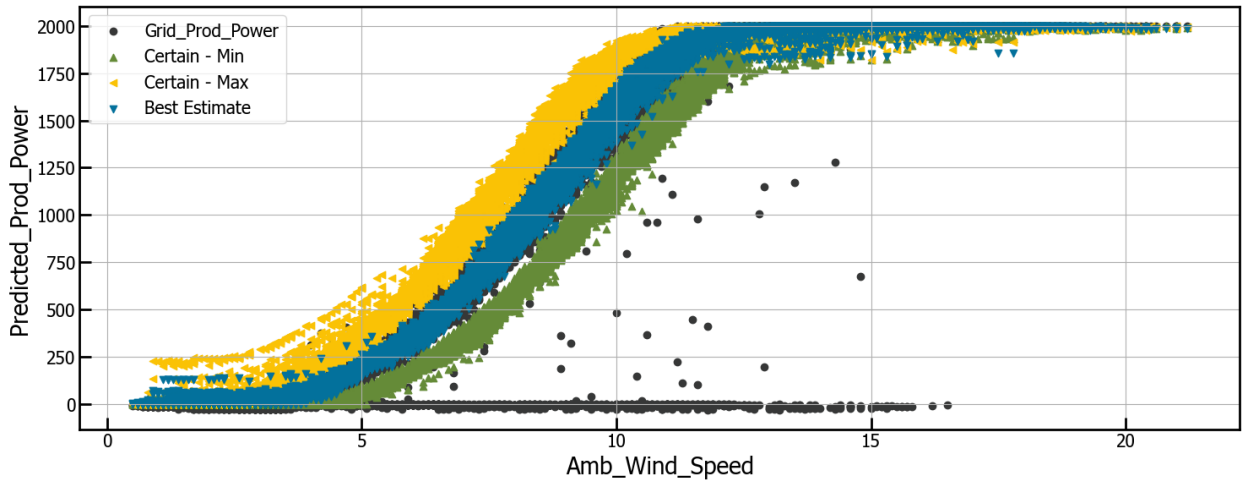

FIGURE 11. Effect of Ambient Wind Speed on Predicted Produced Power at α–$cut = 0$. $Certain - Min$ is obtained from Combination_6 and $Certain - Max$ is obtained from Combination_11.

7.3 Possibility of Errors Due to Imperfection of Measurement

Fig. 10-12 present the results of calculations performed using the Possibilistic Approach to assess the impact of measurement uncertainties on the Predicted Produced Power.

Fig. 10 illustrates the influence of the maximum and minimum interval values of environmental variables on the predicted power output. It is evident that different combinations of these input values lead to significant variations in the Predicted Produced Power. This sensitivity is largely attributed to the fact that, as described by

Equation 6, the Predicted Produced Power is proportional to the cube of the Ambient Wind Speed. Consequently, Combinations 1 through 8 yield noticeably lower power predictions compared to Combinations 9 through 16. This difference is primarily due to the higher wind speeds used in the latter combinations. Within each group (Combinations 1–8 and Combinations 9–16), the variation is relatively small, largely because the air density does not vary significantly across the combinations.

Fig. 11 focuses on the effect of Ambient Wind Speed on the Predicted Produced Power at $\alpha{-}cut$=0. This figure highlights the pronounced influence of measurement uncertainty on power prediction. The lower bound ($Certain - Min$) of the predicted power is derived from Combination 6, while the upper bound ($Certain - Max$) corresponds to Combination 11. These bounds represent the extremes of the predicted values based on plausible variations in the wind speed.

Fig. 12 presents a comparison between the actual Grid Produced Power and the Predicted Produced Power at $\alpha{-}cut$=0 over a 24-hour period (10th November, 2017). The plot shows that the measured power values generally fall within the outermost bounds defined by the $Certain - Min$ and $Certain - Max$ predictions. This suggests that the possibilistic model effectively captures the range of potential outcomes arising from measurement uncertainties in the input variables.

8 Uncertainties Due to Combination of Measurement and Model

8.1 Representation of Input Variables as Possibility Distribution Functions

In this stage of the analysis, three distinct Possibility Distribution Functions (PDFs) are employed to represent uncertainty in the input variables. These functions model both measurement errors and model-related imperfections, enabling a comprehensive uncertainty analysis in the prediction of produced power. The three PDFs used are:

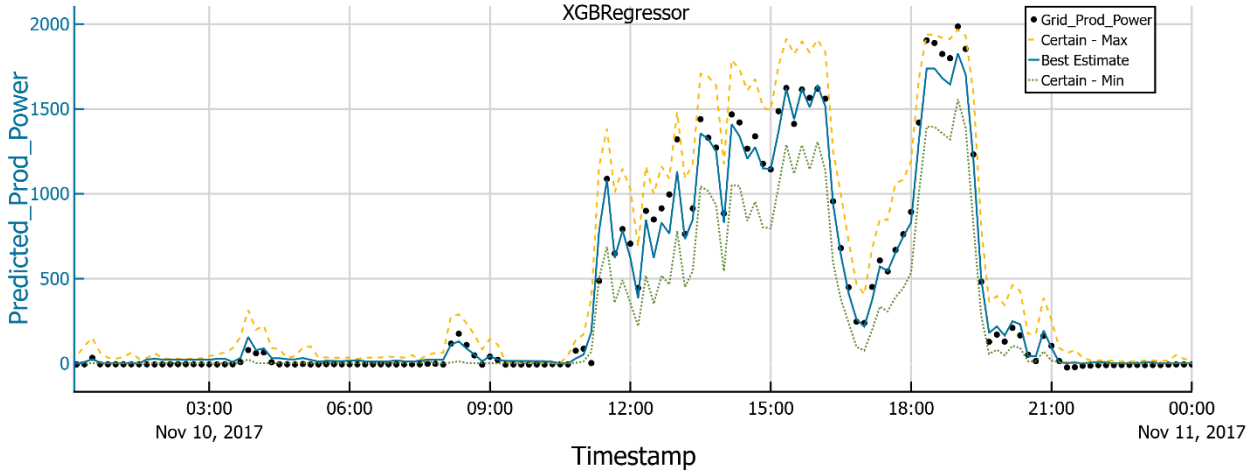

FIGURE 12. Plot of Grid Produced Power and Predicted Produced Power incorporating uncertainties due to measurement at $\alpha{-}cut$=0 for a 24 hour duration (10th November, 2016).

- *Model_Error:* This function represents the uncertainty associated with model imperfection – specifically the error arising from the limitations of the predictive model itself. It was derived based on the error analysis discussed in **Section 6.2**.

- *Measurement_Error_Combination_6 :* This function captures the lower bound of measurement uncertainty, representing the $Certain - Min$ scenario for Predicted Produced Power. It is derived from Combination 6, as detailed in **Section 7.3**.

- *Measurement_Error_Combination_1 1:* This function captures the upper bound of measurement uncertainty, representing the $Certain - Max$ scenario for Predicted Produced Power. It is derived from Combination 11, also discussed in **Section 7.3**.

These possibility distribution functions are treated as Fuzzy numbers and are evaluated across multiple α–cuts, which represent varying levels of confidence or certainty.

8.2 Calculation of Predicted Produced Power Accounting for Modelling and Measurement Errors

The combined effect of modelling error and measurement error on the Predicted Produced Power is evaluated using the principles of interval arithmetic applied at various α–cut levels of the fuzzy numbers.

Each possibility distribution function is decomposed into α–cut, which define interval ranges for each α-level ($\alpha \in [0,1]$). For two fuzzy numbers, A and B, represented at a given α–cut as:

$A_\alpha = \left|\underline{a}, \overline{a}\right|_\alpha$ and $B_\alpha = \left|\underline{b}, \overline{b}\right|_\alpha$

the interval arithmetic operations are defined as follows:

Addition:

$$A_\alpha + B_\alpha = \left[\underline{a}_\alpha + \underline{b}_\alpha, \overline{a}_\alpha + \overline{b}_\alpha\right] \quad (7a)$$

Subtraction:

$$A_\alpha - B_\alpha = \left[\underline{a}_\alpha - \overline{b}_\alpha, \overline{a}_\alpha - \underline{b}_\alpha\right] \quad (7b)$$

These operations are used to propagate uncertainty through the output variable (Predicted Produced Power) by combining the Fuzzy input intervals.

To apply these concepts of interval analysis, first a value of α is selected. For this value of α the α–cut of each possibility distribution function is determined. Considering all the values located in the α–cuts for every possibility distribution function, the minimum and maximum values of the output function are calculated. This step is repeated for all α–cuts for $\alpha \in [0,1]$. The results of all α–cuts are combined to build the fuzzy membership function of the output function.

Using this concept, the steps followed for the calculation of the Predicted Produced Power using fuzzy arithmetic are:

1. *Initialize with $\alpha = 0$:* Select a value α of the membership function starting from $\alpha = 0$.

2. *Determine α–cuts:* For each Possibility Distribution Function (Model_Error, Measurement_Error_Combination_6, and Measurement_Error_Combination_11

), determine the interval corresponding to the selected α–cut.

3. *Calculate* $Certain - Min$ *and* $Certain - Max$*:*

- The lower bound ($Certain - Min$) of Predicted Produced Power is calculated by subtracting the Model_Error interval from the Measurement_Error_Combination_6 interval.

- The upper bound ($Certain - Max$) is calculated by adding the Model_Error interval to the Measurement_Error_Combination_11 interval.

4. *Compute Output Interval:* Using the above combinations, determine the minimum and maximum values of Predicted Produced Power for the given α.

5. *Repeat for* $\alpha \in [0,1]$*:* Repeat Steps 2–4 for multiple α–cut levels (e.g., α = 0, 0.1, ..., 1.0) to span the entire range of uncertainty.

6. *Stack* α*–cut Intervals:* Combine the results of all α–cuts to reconstruct the fuzzy membership function of the Predicted Produced Power, effectively forming its Possibility Distribution Function.

This method ensures a thorough and mathematically consistent handling of uncertainty, incorporating both measurement variability and model imperfection into the final prediction.

8.3 Possibility of Errors Due to Combination of Measurement and Model

Fig. 13 illustrates the effect of Ambient Wind Speed on the Predicted Produced Power under conditions of maximum uncertainty, represented by α–cut=0. This α-level signifies the lowest level of certainty, where the input variables are allowed to vary within their widest possible bounds. The figure clearly highlights the pronounced impact of combined measurement and model-related uncertainties on the predicted power output.

At this α-level, the lower bound of the prediction range – referred to as $Certain - Min$ – is derived by combining the Model Error distribution with the Measurement Error from Combination 6, which represents the most conservative (pessimistic) estimate. Conversely, the upper bound, or $Certain - Max$, is determined by combining the Model Error with the Measurement Error from Combination 11, representing the most optimistic scenario. Together, these two boundaries define the extremes of predicted power based on plausible variations in ambient wind speed and the associated uncertainty in the modelling process.

Fig. 14 presents a time-series comparison of the actual Grid Produced Power and the Predicted Produced Power at α–cut=0 over a continuous 24-hour period on 10th November 2017. This figure serves as a validation of the possibilistic prediction approach. It shows that the majority of the measured values fall within the envelope defined by the $Certain - Min$ and $Certain - Max$ bounds. This indicates that the possibilistic model not only accommodates uncertainty but does so in a way that encompasses real-world observations, even under conditions of high variability and limited data precision.

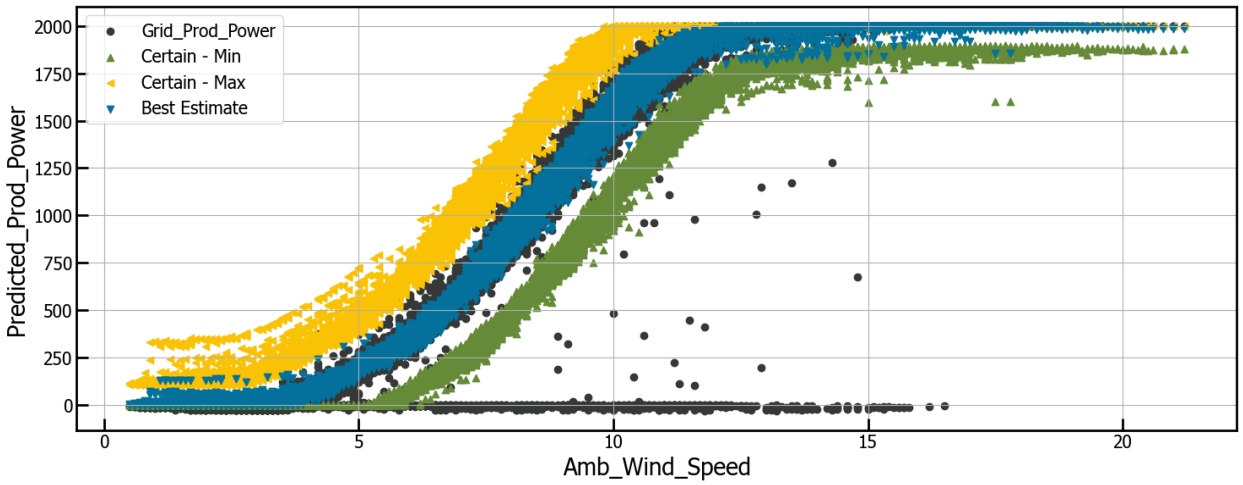

FIGURE 13. Effect of Ambient Wind Speed on Predicted Produced Power at $\alpha\text{–}cut = 0$. (a) $Certain - Min$ is obtained from Model_Error and Measurement_Error_Combination_6 and (b) $Certain - Max$ is obtained from Model_Error and Measurement_Error_Combination_11.

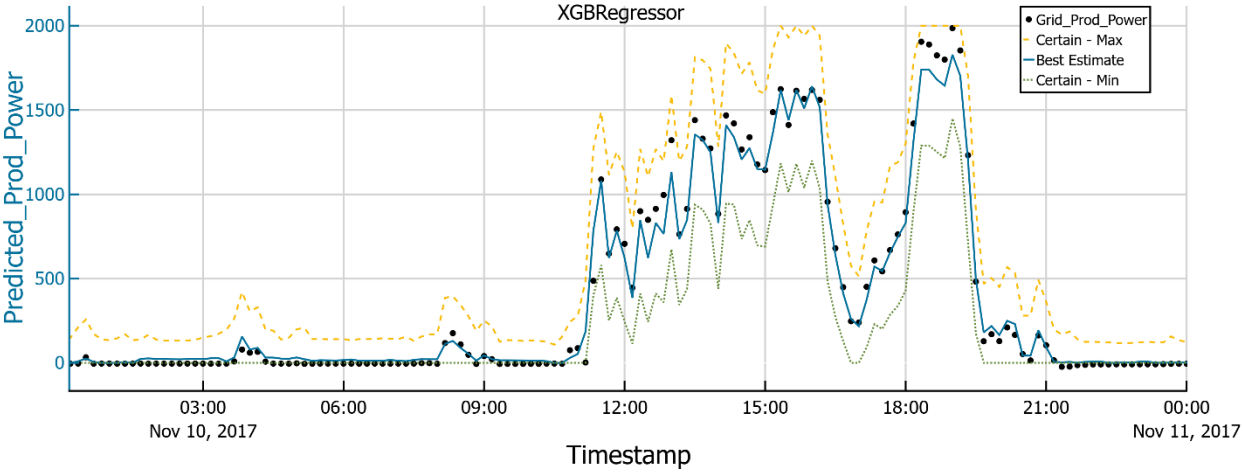

FIGURE 14. Plot of Grid Produced Power and Predicted Produced Power incorporating combined uncertainties due to measurement and model at $\alpha\text{–}cut$=0 for a 24 hour duration (10th November, 2016).

The fact that observed values remain largely within these boundaries provides evidence that the possibilistic model effectively captures the full range of possible outcomes, even under significant uncertainty. More specifically, it demonstrates that the model is capable of accounting not only for uncertainties in environmental input variables (such as wind speed and air density), but also for potential imperfections in the predictive model itself.

Overall, **Fig. 13 and 14** underscore the value of the possibilistic modelling approach in predictive power generation, particularly in operational contexts where uncertainty is

inevitable and precise measurements are difficult to obtain. This method provides a comprehensive and realistic representation of uncertainty, making it a suitable tool for decision-making in wind energy forecasting.

9 Conclusions

This paper presents a simple yet robust methodology for predicting Produced Power from wind energy systems using SCADA data, while explicitly accounting for uncertainties arising from both modelling imperfections and measurement errors. The approach integrates two complementary components:

- *Machine Learning models:* Used to construct predictive relationships between environmental input variables and power output.
- *Possibilistic Framework:* Used to systematically handle uncertainty through the use of possibility distribution functions, derived from Fuzzy Logic and interval analysis.

To demonstrate the applicability of this methodology, a real-world case study was conducted using operational SCADA data. The predictive model was trained using various machine learning algorithms, with XGBoost selected as the final model based on its superior performance and computational efficiency.

The effectiveness of the possibilistic approach was validated by comparing the Predicted Produced Power under conditions of uncertainty against the actual Grid Produced Power. The analysis showed that the observed values remained largely within the predicted $Certain - Min$ and $Certain - Max$ bounds. This result provides evidence that the possibilistic model reliably captures the full range of potential outcomes, even when input data is imprecise, incomplete, or subject to operational variability.

More specifically, the methodology demonstrates a clear capability to account for:

- Uncertainties in measurement (e.g., wind speed and air density) in SCADA instrumentation,
- Structural limitations or assumptions in the predictive machine learning model itself.

This ability to model and quantify uncertainty in both data and model behaviour makes the proposed approach especially valuable for power forecasting in real-world applications, where ideal measurement conditions are rarely met and system performance may deviate from theoretical expectations.

In summary, the integration of machine learning with possibilistic uncertainty modelling offers a practical, adaptable, and reliable solution for power prediction tasks in the presence of real-world data imperfections.

Data Availability

The datasets presented in this study can be found in online repositories given below:

https://www.edp.com/en/innovation/data/wind-farm-1-wind-turbine-scada-signals-2016

https://www.edp.com/en/wind-farm-1-wind-turbine-scada-signals-2017

References

1. Nilsson, J., and Bertling, L. Maintenance Management of Wind Power Systems Using Condition Monitoring Systems — Life Cycle Cost Analysis for Two Case Studies. IEEE Transactions on Energy Conversion, Vol. 22 (1), 223–229. (2007).

2. Fischer, K.; Besnard, F.; Bertling, L. Reliability-Centered Maintenance for Wind Turbines Based on Statistical Analysis and Practical Experience, IEEE Transactions on Energy Conversion, Vol.27 (1), p.184-195. (2012).

3. Bindingsbø, O.T., Singh, M., Øvsthus, K. and Keprate, A. Fault Detection of a Wind Turbine Generator Bearing Using Interpretable Machine Learning, Frontiers in Energy Research, 11:1284676, doi: 10.3389/fenrg.2023.1284676. (2023).

4. Manwell, J. F., McGowan, J.G. and Rogers, A.L. Wind Energy Explained — Theory, Design and Application (2nd ed.), John Wiley & Sons Ltd., ISBN 978-0-470-01500-1. (2009).

5. Pandit, R. and Wang, J. A Comprehensive Review on Enhancing Wind Turbine Applications with Advanced SCADA Data Analytics and Practical Insights, IET Renewable Power Generation, Vol. 18, pp. 722-742. (2024).

6. Tavner, P. Offshore Wind Turbines — Reliability, Availability and Maintenance, The Institution of Engineering and Technology, IET Renewable Energy Series 13, ISBN 978-1-84919-230-9. (2012).

7. Yang, W., Wei, K., Peng, Z. and Hu, W. Chapter 7, Advanced Health Condition Monitoring of Wind Turbines, W. Hu (ed.), Advanced Wind Turbine Technology, Springer International Publishing AG. (2018).

8. Duguid, L. Data Analytics in the Offshore Wind Industry – Pilot Case Study Outcomes, CATAPULT - Offshore Renewable Energy Report No. PN000229-RPT-001. https://ore.catapult.org.uk/wp-content/uploads/2018/05/Data-Analytics-in-Offshore-Wind-Pilot-Case-Study-Outcomes.pdf. (2018),

9. Bell, S. A Beginner's Guide to Uncertainty of Measurement. Issue 2, National Physical Laboratory, Report No. 11. (1999).

10. Simon, C., Weber, P. and Sallak, M. Data Uncertainty and Important Measures, John Wiley & Sons, EBOOK ISBN 9781119489351. (2018).

11. Ayyub, B.M. and Klir, G.J. Uncertainty Modeling and Analysis in Engineering and Sciences, Chapman & Hall/CRC Press, Boca Raton. (2006).

12. Ross, T. J. Fuzzy Logic with Engineering Applications, John Wiley and Sons Ltd, ISBN 9780470860748. (2004).

13. Singh, M. and Markeset, T. Hybrid Models for Handling Variability and Uncertainty in Probabilistic and Possibilistic Failure Analysis of Corroded Pipes, Engineering Failure Analysis, 42, pp. 197–209, 2014. (2014).

14. Singh, M. A Hybrid – Machine Learning and Possibilistic – Methodology for Predicting Produced Power Using Wind Turbine SCADA Data, *Proc. of the 8th European Conference of the PHM Society*, July 3-5, Prague, Czech Republic, 2024, https://papers.phmsociety.org/index.php/phme/article/view/4006. (2024).

15. Mauris G, Lasserre V, Foulloy L. A Fuzzy Approach for the Expression of Uncertainty in Measurement, Measurement, Vol. 29, pp. 165–77. (2001).

16. Singh, M. and Markeset, T. Handling of Variability in Probabilistic and Possibilistic Failure Analysis of Corroded Pipes, International Journal of Systems Assurance Engineering and Management, 5 (4), pp 503-512, Dec. 2014. (2014a).

17. Dubois D., Foulloy L., Mauris G. and Prade H. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Computing 2004, 10: 273-297.

18. Karimi I. and Hüllermeier E. Risk assessment system of natural hazards: A new approach based on fuzzy probability. Fuzzy Sets and Systems 2007, 158: 987-999.

19. Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, August 2-4, p. 226–231. (1996),

20. Saint-Drenan, Y.-M. et al. A Parametric Model for Wind Turbine Power Curves Incorporating Environmental Conditions, Renewable Energy, Vol. 157, pp. 754-768. (2020).