A Hybrid – Machine Learning and Possibilistic – Methdology for Predicting Produced Power Using Wind Turbine SCADA Data

Maneesh Singh

Western Norway University of Applied Sciences, 5020 Bergen, Norway
maneesh.singh@hvl.no

ABSTRACT

During its operational lifetime, a wind turbine is continuously subjected to a number of aggressive environmental and operational conditions, resulting in degradation of its parts. If left unattended, these degraded components will negatively influence its performance and may lead to failure of the wind turbine. In order to mitigate the risk associated with the failure of components, a wind turbine is regularly inspected and maintained.

Currently, there are two commonly used approaches for making maintenance management (inspection and maintenance) plans. Traditional Approach utilises understanding of failure profile of the components for manually developing maintenance plan for the equipment. Condition-Based Approach utilises data collected by condition monitoring of equipment for developing dynamic maintenance plan. SCADA system offers a low-resolution condition-monitoring data that can be used for fault detection, fault diagnosis, fault quantification and fault prognosis and eventually for maintenance planning.

The monitoring data from SCADA system of a wind turbine is often afflicted with variability and uncertainty. The variability in data is the result of continuously changing environmental conditions and uncertainty arises due to imperfections in the recorded data. The uncertainty may be due to many reasons, including, inherent characteristic of sensing devices, drift in calibration with time, deterioration of sensing devices' sensitivity and response due to environmental attacks, etc.

For handling variability in monitoring data a number of parametric and non-parametric (statistical) predictive models for different aspects of a wind turbine's structure and operation have been proposed. Depending upon its type – aleatory or epistemic – an uncertainty can be handled in a number of ways. Since, the dynamic nature of wind turbine operation does not allow collection of multiple values under the same condition; hence, uncertainty is mostly epistemic in nature. Possibilistic Approach, based on Fuzzy Set Theory, is especially suitable for handling epistemic uncertainty that may arise due to imprecision or lack of statistical data.

Aim of the ongoing research is to develop a methodology for detecting sub-optimal operation of a wind turbine by comparing Measured Produced Power against Predicted Produced Power. Unfortunately, variability and uncertainty associated with the recorded data make accurate prediction of produced power challenging.

This paper presents methodologies for predicting produced power using SCADA data while simultaneously accounting for variability and uncertainty. The methodologies utilise either parametric (example, power curve) or machine learning (example, XGBoost) models for handling variability; and Possibilistic Approach for handling uncertainty.

1. INTRODUCTION

1.1. Background

The world has two conflicting needs, on one side is the need to generate and supply more energy to bring people out of poverty and improve their living standard; on the other side is the need to reduce reliance on fossil fuel so as to cut down on emissions that cause global warming. These conflicting needs have acted as a spur to find economical and clean alternative sources of energy. In recent years, wind power has become one of the major sources of alternative energy and its share is expected to continuously grow in the coming decade (Global Wind Energy Council, 2021).

Due to various financial, social ("not-in-my-backyard" syndrome), environmental (meteorological conditions) and geographical (topological features) reasons the wind turbines are often located in remote areas where they experience harsh environmental conditions. The inconsistent and aggressive environmental conditions, like, wind velocity, humidity, temperature, precipitation and icing, degrade the vulnerable components. If left unattended, these degraded components will result in deterioration of performance and at times failure. To prevent that from happening, maintenance of wind turbines is needed throughout their lifetime. It is estimated that maintenance costs comprise of a significant proportion (10-25%) of the total annual operational cost (Nilsson & Bertling, 2007).

Currently, there are two commonly used approaches for making maintenance management plans (tasks and schedules):

(a) **Traditional Approach** – In which understanding of the failure profile (failure causes, failure mechanisms, failure modes, failure rates, etc.) of components is used to develop maintenance concept and maintenance plan for the equipment.

(b) **Condition-Based Approach** – In which data, collected using condition-monitoring equipment or Supervisory Control and Data Acquisition (SCADA) systems is analysed for fault detection, fault diagnosis, fault quantification and fault prognosis and maintenance planning.

The Traditional Approach analyses structural, environmental and operational attributes to develop corrective or preventive maintenance plans. The preventive maintenance plans are often time-based, for example, preventive maintenance activities of wind turbines are normally planned at 3 to 6-month intervals based upon the age and condition of the turbine (Nilsson & Bertling, 2007). Since these time-based inspection and maintenance plans are expensive to execute, there have been efforts to develop methodologies based on formalized risk analysis, e.g., Risk Based Inspection and Maintenance or Reliability Centered Maintenance. This involves understanding failure profile and carrying out risk analysis & risk evaluation for preparing maintenance plans that are more efficient and effective than time-based or incidence-based maintenance plans (Fischer, Besnard & Bertling, 2012).

The Condition-Based Approach improves upon the inspection and maintenance plan by using condition attributes to update the equipment's risk assessment by detecting faults. This is achieved by (a) intermittent or continuous monitoring using sensors; (b) data analytics; and (c) developing condition-based maintenance plans. This approach can be applied using either (Tavner, 2012):

1. **Condition Monitoring System (CMS)** – A high-resolution specialized system for detailed analysis of the condition of a machinery by monitoring parameters like, speed, displacement, vibration and oil particles, using sensitive sensors. While specialized Condition Monitoring Systems can give accurate and detailed analysis, they are also expensive to install and use.
2. **Supervisory Control and Data Acquisition (SCADA)** – A low-resolution, usually at 10-minute intervals, standard system in every large wind turbine that monitors parameters for characterising environment, electrical, operational or structural attributes. SCADA system uses this data for controlling the wind turbine's operation after analysing its operating conditions and status. This data can also be used for deducing the health (fault detection, diagnosis and quantification) of the wind turbine.
3. **Structural Health Monitoring (SHM)** – A low-resolution system for monitoring health of a structure, including tower and foundation.

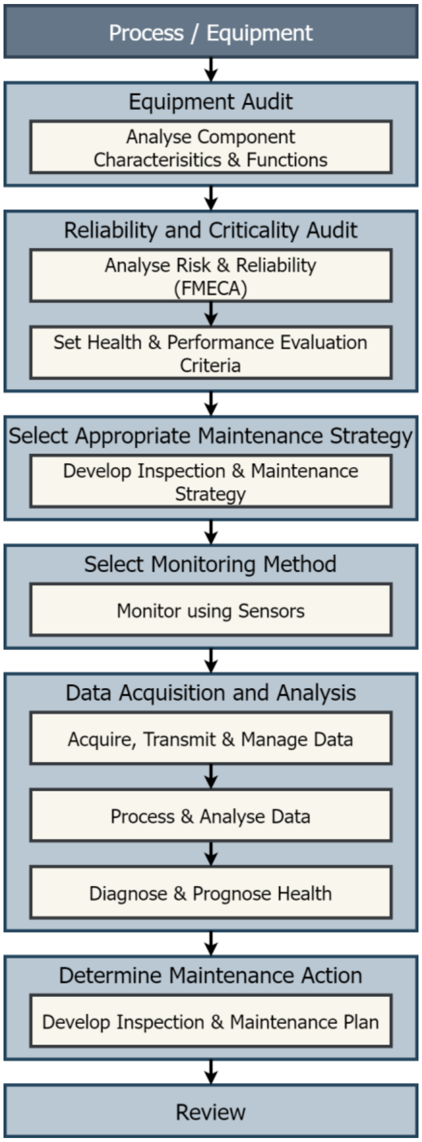
Process / Equipment
Equipment Audit
Analyse Component Characterisitics & Functions
Reliability and Criticality Audit
Analyse Risk & Reliability (FMECA)
Set Health & Performance Evaluation Criteria
Select Appropriate Maintenance Strategy
Develop Inspection & Maintenance Strategy
Select Monitoring Method
Monitor using Sensors
Data Acquisition and Analysis
Acquire, Transmit & Manage Data
Process & Analyse Data
Diagnose & Prognose Health
Determine Maintenance Action
Develop Inspection & Maintenance Plan
Review

Figure 1. Main steps of a monitoring system (Based on ISO17359).

While Condition Monitoring System (CMS) provides costly but in-depth coverage, SCADA and Structural Health Monitoring (SHM) can provide cheap but wide coverage. Hence, a number of commercially available SCADA systems offer real-time data analysis, using statistical and artificial intelligence techniques, for fault detection of components. Yet, there is a need for better diagnostics, prognostics and control techniques using SCADA (Tavner, 2012; Yang et al., 2018).

Since, both – Traditional and Condition-based – Approaches have their own advantages and disadvantages, most of the maintenance planning is carried out by integrating the two approaches. The integration provides a solution that is robust, effective, and efficient. In an integrated method (Bindingsbø et al. 2023):

- failure analysis is carried out in the traditional manner, and then the results of failure profile is used judiciously to develop a maintenance strategy;
- time for inspection and maintenance of a component is adjusted based upon outcome of condition monitoring.

Figure 1 shows main steps that should be carried out to monitor a system according to ISO17359. According to the standard, condition-monitoring approach has three steps (Equipment Audit, Reliability and Criticality Audit and Select Appropriate Maintenance Strategy) that help in developing maintenance plan using the Tradition Approach. Thereafter, three more steps (Select Monitoring Method, Data Acquisition and Analysis and Determine Maintenance Action) help in improving the maintenance plan by incorporating knowledge of system’s condition.

1.2. Supervisory Control and Data Acquisition (SCADA) System

An offshore wind turbine is subjected to severe variations in the environmental and operating conditions. To continuously monitor these variations all modern wind turbines come with a Supervisory Control and Data Acquisition (SCADA) system (Pandit & Wang, 2024).

In a SCADA system, a multitude of sensors constantly monitor various meteorological and operational parameters; and the data is transmitted, processed and stored in SCADA supervisory computers. The parameters that are monitored include (Manwell, McGowan & Rogers, 2009):

- **Position** – blade pitch angle, nacelle direction
- **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
- **RPM** – rotor, generator
- **Hydraulic Characteristics** – pressure, reservoir level, flowrate
- **Environmental Characteristics** – wind speed, wind direction, temperature, humidity
- **Electrical Characteristics** – active power, reactive power, voltage, current, phase displacement, frequency

Apart from the data collected using sensors that are connected to a wind turbine, a number of data streams from nearby weather stations are also recorded.

The recorded SCADA data is analysed using different deterministic, probabilistic, Fuzzy Logic, Machine Learning, Artificial Neural Networks and Deep Learning approaches to detect, diagnose and quantify failures in the components. Information gained after analysis is used to control the process or operation (Manwell, McGowan & Rogers, 2009; Tavner, 2012).

Based on the data collected and analysed, a SCADA system can perform the following tasks (Manwell, McGowan & Rogers, 2009; Pandit & Wang, 2024):

1. **Controlling Operating Conditions** – SCADA uses the information regarding environment and grid to determine the appropriate operating conditions. It then controls the components (pitch angle, brakes, generator connection to the grid, etc.) so that the turbine operates according to the determined task schedule.
2. **Monitoring for Fault Detection** – SCADA uses the data from sensors (example, bearing temperature, hydraulic oil temperature, etc.) connected to critical components to monitor their behaviour and detect potential faults or spurious behaviour.
3. **Raising Alarm in Case of Faulty Behaviour** – If SCADA detects abnormal behaviour of a component it can raise alarm and notify the operator.
4. **Triggering Safety and Emergency Response** – In case of situations that can escalate into an accident, SCADA can disconnect turbine from the grid and activate brakes to isolate and shut down the operation.
5. **Integrating with Power Grid** – SCADA can control integration of individual wind turbine into the power grid, thereby contributing to feed and stabilisation.

1.3. Condition-based Maintenance Planning Using SCADA Data

The data acquired from SCADA can be used for fault detection, where a fault can be of various kinds, for example, degradation of components, failure of sensors, operation beyond safe operating limits, problems associated with grid. While it may be possible to detect some of these faults directly, for example, failure of sensors resulting in irrational readings, other faults may only be detected indirectly (Manwell, McGowan & Rogers, 2009).

Depending upon the type of fault, the time span between inception to potential failure could be between a few seconds (example, generator earth fault) to a few weeks (example, wear-out of gears). For the faults that have a long time span, analyses of SCADA data using appropriate models for fault diagnosis, fault quantification and finally fault prognosis may help in planning maintenance activities. These activities can be:

- triggered either when some condition indicator crosses a pre-set limit, or
- decided based on combination of Failure mode, Effect and Criticality Analysis (FMECA) with the condition analysis (fault diagnosis, quantification and prognosis) to update the existing maintenance plan.

The recommended maintenance activities may include inspection (visual, auditory, NDT), testing, service (lubrication, cleaning, repair, etc.), repair and replacement tasks. These activities may be either preventive or corrective in nature depending on whether the needed task is carried out before or after failure. Since maintenance activities are planned based on the actual monitored condition, condition-based maintenance strategy offers advantages that are associated with (Bindingsbø et al. 2023, Tavner, 2012):

- maintenance activities being carried out when required and not limited to corrective or preventive maintenance;
- not conducting unnecessary scheduled replacement of parts before their end of useful life.

In spite of these advantages, use of the Condition-Based Approach is still restricted and needs further research and development. This is because of the difficulties associated with the (Bindingsbø et al. 2023):

- quality and quantity of collected data,
- handling of imperfect (spurious, inconsistent, inaccurate, uncertain, or irrational) data collected from faulty sensors,
- interpretation of data for fault diagnosis, quantification and prognosis,
- updating of maintenance plan, and
- handling of unreliable analysis that may trigger false alarm (false positive) or failure to respond (false negative)

1.4. Methodologies for Predicting Produced Power

One of the common methods for analysing the performance of a wind turbine using SCADA data is to understand the power generation as a function of various variables, especially wind speed. A significant difference between the predicted power generation and measured power generation gives an indication of sub-optimal performance, hence, need for detailed examination. For this purpose it is essential to be able to accurately predict power generation under varying environmental and operating conditions (Pandit & Wang, 2024; Wang et al., 2016).

Power curve of a wind turbine is the unique relationship of a wind turbine between the power it generates and the environmental and operational conditions under which it operates. The power generated by a wind turbine is dependent upon the technical (example, radius of the rotor), environmental (example, wind speed, air density) and operational (example, pitch angle, angle between wind and nacelle) attributes (Manwell, McGowan & Rogers, 2009).

In a simplified power balance model, the wind power is converted to rotor power; which in turn is converted to electrical power. The efficiency of conversion of wind power to rotor power is dependent upon wind speed, air density, blade geometry, etc. Ideally, the rotor power should be converted entirely to the electrical power via its drive train system; but in reality, some power is lost as vibration and heat. The energy balance can be expressed as (Manwell, McGowan & Rogers, 2009):

$$P_{Rotor} = P_{Electrical} + P_{Vibration} + P_{Thermal} \tag{1a}$$

$$P_{Rotor} - P_{Electrical} = P_{Vibration} + P_{Thermal} \tag{1b}$$

Where:

P_{Rotor}	= Rotor power
$P_{Electrical}$	= Electrical power
$P_{Vibration}$	= Vibration power
$P_{Thermal}$	= Thermal power

Hence, an increased discrepancy between rotor power (P_{Rotor} , predicted using models) and electrical power ($P_{Electrical}$, measured) is an indication of additional loss of energy due to increase in vibrations and heat generation-dissipation. This in turn can be attributed to the falling health condition of the mechanical and electrical drive train components. Thus, analysis of produced power can be used for (Duguid, 2018):

- **Fault Detection** – While exact cause may not be easy to identify, but a significant difference may help in fault detection necessitating further investigation.
- **Suboptimal Performance Detection** – Suboptimal performance, often due to poor control, can be identified using power curve. A comparison in power generation between a local group of wind turbines may also help in identifying those units that are performing sub-optimally.

To predict power generation, a number of parametric and non-parametric (statistical) methods have been proposed (Lydia at al. 2014; Pandit, Infield & Kolios, 2019; Saint-Drenan et al., 2020; Pandit & Wang, 2024). The parametric models are based on functions that correlate different variables and are of different types. For example, linearized segmented model, polynomial power curve, 4/5-parameter logistic function, etc. are based on power equation derived from Bentz's law, which can be expressed as (Manwell, McGowan & Rogers, 2009):

$$P_{Rotor} = P_{Wind} \times C_P(\lambda, \beta) \quad (2a)$$

$$P_{Electrical} = P_{Rotor} \times \eta \quad (2b)$$

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

Where:

P_{Wind}= Wind power
η = Drive train efficiency ($generator\ power/rotor\ power$), (mechanical & electrical)
ρ = Air density
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).

The λ (tip speed ratio) can be expressed as:

$$\lambda = \frac{\Omega R}{U} \quad (3)$$

Where:

λ = Tip speed ratio
R = Radius of the wind rotor
Ω = Angular velocity (in radians/sec)

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 much below the Bentz limit and is dependent upon technical features of the turbine and environmental factors (Saint-Drenan et al., 2020).

According to the **Equation 2c**, produced electric power is proportional to the density of air and cube of wind speed. The density of air is in-turn dependent upon the ambient temperature, humidity and pressure. It can be calculated according to:

$$\rho = \rho_d + \rho_v \quad (4a)$$

$$\rho_d = \frac{P - P_v}{\left(R_{Specific,Dry\ Air} \times T_k\right)} \quad (4b)$$

$$\rho_v = \frac{P_v}{\left(R_{Specific,Water\ Vapour} \times T_k\right)} \quad (4c)$$

$$P_{sat} = 6.1078 \times 10^{\frac{7.5T}{T+237.3}} \quad (4d)$$

$$P_v = \frac{(h \times P_{sat})}{100}$$

Where:

ρ_d = Density of the dry air
ρ_v = Density of the water vapour
T = Temperature (°C)
T_K = $T + 273.15$ (Kelvin)
h = Humidity
P = Total pressure of air
P_{sat} = Saturation water vapour pressure (Tetens' Formula)
P_v = Partial pressure of water vapour
$R_{Specific,Dry\ Air}$ = Specific gas constant for dry air = 287.05 J/(kg·K)
$R_{Specific,Water\ Vapour}$ = Specific gas constant for water vapour = 461.5 J/(kg·K)

The actual operation of a wind turbine is outcome of a number of controls, for example, aerodynamic torque control, yaw orientation control, brake torque control and generator torque control, that work together to create a number of decision combinations. The final operating strategy, which is an outcome of optimisation of diverse and often contradictory goals, determines the control of individual components. These goals include, safe operation, maximising power generation, minimising vibrations, preventing structural damages, integration with grid, etc. (Manwell, McGowan & Rogers, 2009).

Due to the complexities involved in accounting for all the parameters that can effect control and operation, the parametric models are often not accurate. Hence, for predicting power generation of existing wind turbines a number of models based on Artificial Intelligence (Support Vector Machine, Gaussian Process, Random Forest and Artificial Neural Network) have been propounded These models are trained using historical SCADA data and the trained models are later used for making predictions (Ouyang et al., 2017; Pandit, Infield & Kolios, 2019).

1.5. Data Quality for Predicting Power Produced

In spite of all the precautions, the measurements recorded by SCADA system are always afflicted with imperfections or uncertainties of various kinds. Where uncertainty of measurement can be defined as *the doubt that exists about the result of any measurement* (Bell, 1999).

Since, the uncertainties arise due to multiple reasons they are also of different types. Some of them are tangible (can be quantified), while others are intangible (cannot be properly quantified). Some uncertainties can by random and others can be systematic. Because of the difficulties associated with the taxonomy of uncertainties, a number of classifications have been proposed. Unfortunately, there is no consensus regarding these classifications and the proposed classifications have not been widely accepted, resulting in confusions. Traditionally, uncertainties have been classified into two types (Manwell, McGowan & Rogers, 2009; Simon, Weber & Sallak, 2018):

- **Aleatoric** – This type of uncertainty arises due to inherent randomness or variability of the measured parameter. By repeating the measurement, it is possible to express it in terms of mean and standard deviation (interval and confidence level).
- **Epistemic** – This type of uncertainty arises due to the lack of knowledge or data. The factors that contribute to the uncertainty influence all the recorded values, hence, there is limited benefit to be gained by repeated measurement. Epistemic uncertainty can be further classified into:
    - **Bias** – It is a systematic shift from the true value.

  - **Inaccuracy** – This is the mean difference between the measured and true value of the measured variable.
  - **Imprecision** – It refers to the length of interval between which the measured values lie.
  - **Ignorance** – It arises due to limited availability of measurements or knowledge regarding precision.
  - **Incompleteness** – It arises due to missing data.
  - **Credibility** – It arises due to competence or trustworthiness during calibration, installation, etc.

  Epistemic uncertainty can be evaluated based on information like the manufacturer's specifications, past experience, expert opinion or subjective feel.

For the sake of completeness, measurements should be reported along with their corresponding uncertainties. A tangible uncertainty can be quantified using two numbers: interval (width of margin of doubt or dispersion about the mean) and confidence level (confidence that the "true" value lies with that margin. Since the uncertainties of a measurement depends upon a number of factors, it is often difficult to quantify all of them (Bell, 1999).

These uncertainties are severe for wind turbines because of the large variations taking place in the environmental conditions. Most of the errors arise due to:

- **Imperfections Caused by Sensors** – These imperfections arise because of many reasons, including, variations in the parametric values, imperfect nature (bias, noise, etc.) of the instruments, incorrect calibration, drift in the instrument calibration, measurement location, etc. They may be characterised as:
  - **Inherent Imperfections** – Since, environmental conditions constantly change, the sensors report values based on their response time, sampling rate, resolution, sensitivity and statistical analysis. Each of these behaviour introduces different types of uncertainties.
  - **Acquired Imperfections** – During its operation, a sensor is exposed to a number of environmental attacks, like, variations in impacts, wind force, temperature, humidity, condensation, frosting / icing, vibrations, oil / dirt / salt deposition, etc., resulting in its degradation.
- **Imperfection Caused by SCADA System** – In a SCADA system, values are recorded every 10 minutes, hence, the recorded data is actually not of that particular time, but a statistical value based on predefined algorithm.

To ensure confidence in the data used for analysis, a number of corrective measures need to be taken. These include (Manwell, McGowan & Rogers, 2009; Tavner, 2012):

- **Use of High Quality Sensors** – High quality sensors should have structure that is able to withstand environmental attacks; and have superiority of performance in terms of accuracy, precision, reliability, repeatability and reproducibility.
- **Use of Multiple Data Streams** – Multiple and varied data streams can be used to confirm the same fault so that its probability of detection increases, for example, use of vibration and debris count for detecting bearing fault. Apart from the benefits of redundancy, use of different sensors at different locations increases the probability of detection. A negative side effect of this is the collection of excessive number of data streams resulting in data overload. Additionally, "law of diminishing return" dictates that use of multiple sensors for the same task may not provide any new information.
- **Use of Advanced Data Analytics Techniques** – A number of methods have been proposed to handle different types of uncertainties. While aleatoric uncertainty is often handled using the Probabilistic Approach, epistemic uncertainty can be handled using the Possibilistic Approach.

  In Possibilistic Approach, values are not regarded as "crisp point numbers" but as membership functions. By integrating Fuzzy arithmetic, that is based on extended interval analysis, with deterministic or Machine Learning models, the predicted output is not a crisp point but a Possibility Distribution Function. Comparison of this output membership function against acceptance criteria gives likelihood of failure in terms of "Possibility of Failure" and "Necessity of Failure". The advantage of using Possibility Distribution Function, over Probability Density Function, is that no preference is given to values within the range of Fuzzy interval. This suits well for the situations where the available data is sparse. The weakness of the Possibilistic Approach is its imprecise results, which may give over-conservative and, at times, uneconomical recommendations. Thus, Possibilistic Approach may be a useful tool for implementing the philosophy of zero tolerance of accidents where not only the probability but also any possibility of failure has to be eliminated (Ayyub & Klir, 2006; Ross, 2004).

2. MOTIVATION AND AIM OF THE RESEARCH

2.1. Motivation for the Research

As discussed in the previous section, performance of a wind turbine can be judged by comparing Predicted Produced Power and Measured Produced Power. A Significant Difference between the two indicates sub-optimal performance. **Figure 2** shows a flowchart of the methodology that can employed for detecting sub-optimal power production.

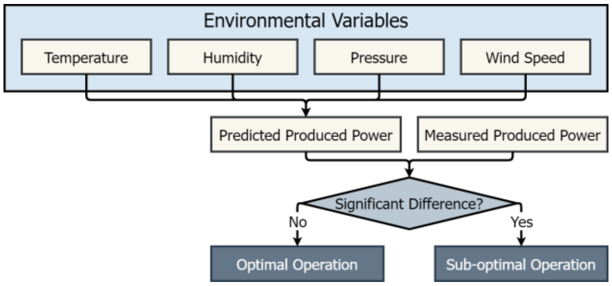

Figure 2. Flowchart showing the proposed fault detection methodology.

It may be possible to calculate Predicted Produced Power by using the four environmental variables; and if the Measured Produced Power (Grid Produced Power) is significantly less than the predicted value, there is a possibility that the wind turbine is operating sub-optimally.

While SCADA data can be used for carrying out this analysis, the methodology has some weaknesses. These weaknesses arise due to:

- lack of reliable models for calculating Predicted Produced Power taking into account all variations and imperfections in the collected data, and
- identification of what constitutes as *Significant Difference* considering the imperfections of the data.

2.2. Aim of the Research

Aim of the research is to develop a methodology for calculating Predicted Power Production using Hybrid (Machine Learning – Possibilistic) Approach while accounting for variability and uncertainty in the SCADA data.

2.3. Scientific Novelty and Importance of the Research

This paper presents work carried out to calculate Predicted Produced Power using wind turbine SCADA data using a Hybrid (Machine Learning – Possibilistic) Approach. The research includes:

- developing Machine Learning models for calculating Predicted Produced Power under varying environmental conditions, and
- handling of imperfections in the collected environmental and operating data by representing them as Fuzzy Membership Functions.

3. METHODS

3.1. SCADA Data Description

To demonstrate feasibility of the proposed methodology, SCADA data made available by the energy company EDP (2016) from four horizontal axis wind turbines located off the western coast of Africa has been used. The data has been recorded over a period of 2 years (2016 and 2017) at a 10-minute averaging interval. The datasets contain values of 76 parameters. For the mechanical components, some recorded parameters are (Bindingsbø et al. 2023):

- **Blades** – pitch angle
- **Rotor** – rpm
- **Nose Cone** – temperature
- **Nacelle** – direction, temperature
- **Generator** – rpm, bearing temperature (drive end and non-drive end), stator windings temperatures in the 3 phases, split ring chamber temperature, active power, reactive power
- **Gearbox** – bearing temperature, oil temperature
- **Hydraulic System** – oil temperature
- **High Voltage Transformer** – temperature
- **Ambient** – temperature, wind speed, wind direction

Associated dataset about meteorological conditions has also been provided for the same time instances. Failure logs containing timestamp, damaged component and associated remarks are also available. For this work, Turbine Number 7 (“T07”) has been selected for which the total number of instances are 52445 and 52294 for 2016 and 2017, respectively. The variables that have been used in the calculation of power curve are given in **Table 1**.

Figure 3a shows the effect of Ambient Wind Speed on Generator RPM. The plot can be divided into three regions – (a) Low RPM Region, where Generator RPM < 300; (b) Transition Region, where 300 < Generator RPM < 1250; and (c) High RPM Region, where 1250 rpm < Generator RPM < 1680. When the Ambient Wind Speed is below the *Cut-In Wind Speed* (4 m/s), the frequency of Generator RPM below 300 rpm is high. With the increase in Ambient Wind Speed, the wind turbine adjusts its blade pitch angle so that Generator RPM is normally above 1250 rpm. Above the *Rated Wind Speed* (12 m/s), the Generator RPM is mostly above 1650 rpm. **Figure 3b** shows the effect of Ambient Wind Speed on Grid Produced Power. When the Ambient Wind Speed is below the *Cut-In Wind Speed* (4 m/s), Grid Produced Power is either negative or less than 275 kW. With increasing Ambient Wind Speed, Grid Produced Power increases so that at the *Rated Wind Speed* (12 m/s), Grid Produced Power is mostly *Rated Power* (2000 kW). **Figure 3c** shows the effect of Generator RPM on Grid Produced Power. The figure shows that the power generation drastically increases when the Generator RPM is above 1250 rpm.

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	Grid_Prod_Power	Grd_Prod_Pwr_Avg	Power average	kW

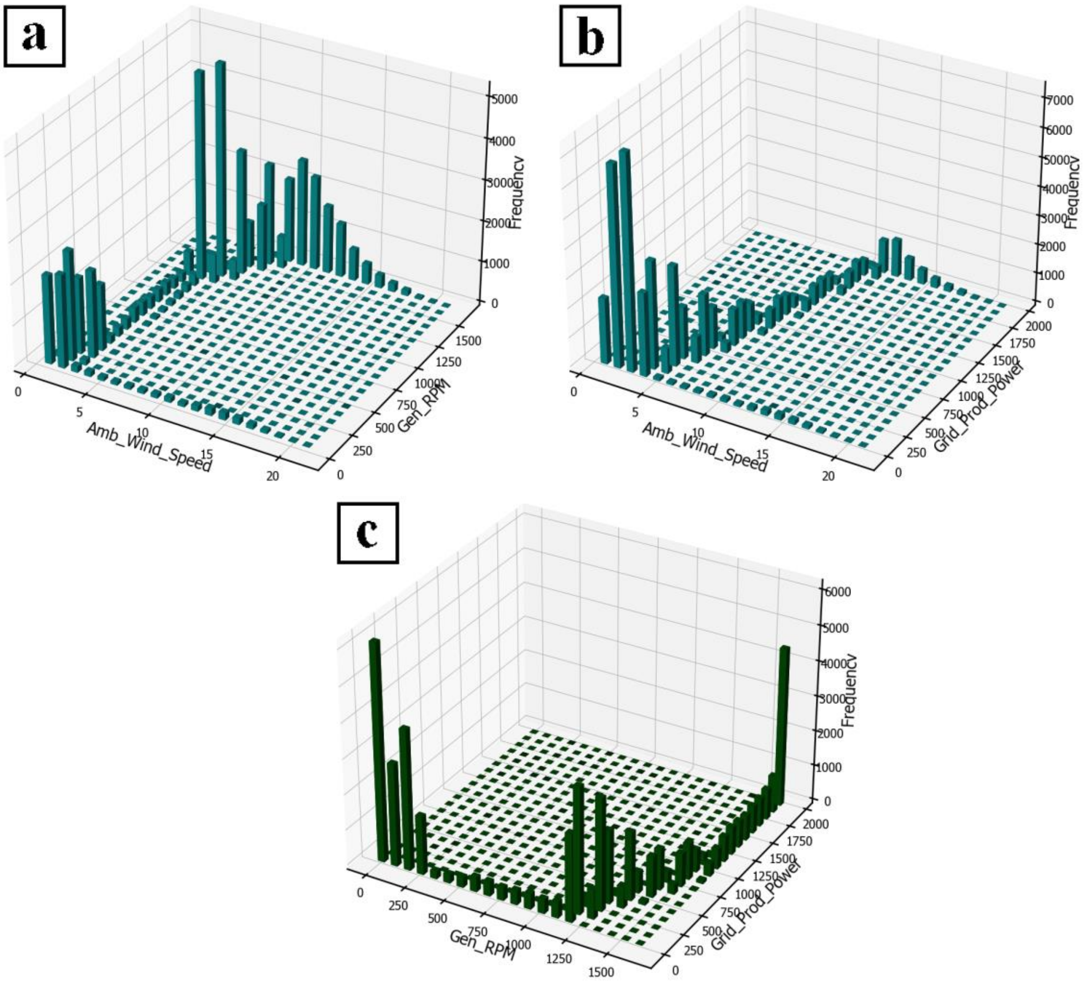

Figure 3. Relationships between Ambient Wind Speed, Generator RPM and Grid Produced Power.

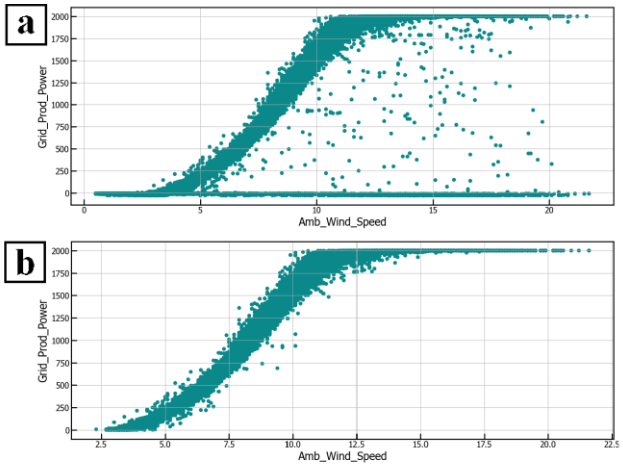

Figure 4. Plot of power generated versus wind speed using SCADA data. (a) Using raw data (b) Using data after removing outliers.

3.2. Data Pre-processing

Data pre-processing is an important step in the development of a Machine Learning model. This is to correct or remove vague, inconsistent, irrational, duplicate or missing values for algorithms to work properly (Bindingsbø et al. 2023).

SCADA data from a wind turbine also contain data that do not conform to the expected power curve and are referred to as "outliers". These outliers arise because of various explainable reasons. In this work, outliers have been identified for the following reasons:

Outlier Rule 1. *Generator RPM = 0 when Ambient Wind Speed => 4 m/s.* Even though the Wind Speed is above the *Cut-In Wind Speed* (4 m/s), the rotor does not move because the wind turbine is in the *shutdown state*. This can be because of various reasons, including the grid condition.

Outlier Rule 2. *Grid Produced Power <= 0 when Ambient Wind Speed < 4 and Generator RPM > 0.* This happens when the rpm of rotor is low, as a result of which power generation is less than the power consumed for operation. The difference is fulfilled by extracting power from grid.

Outlier Rule 3. *Grid Produced Power <= 0 when Ambient Wind Speed => 4 & Generator RPM > 0.* Even though the Wind Speed is above the *Cut-In Wind Speed* (4 m/s), the rotor is moving, power generation does not take place because the wind turbine is "free wheeling" in the *shutdown state*. This can be because of various reasons, including the grid condition.

Apart from these outlier data points, there are some more points that need to be removed. These data points have been recorded during the transition from normal operation to shutdown state or *vice versa*. These points lie scattered and can be identified using DBSCAN, a density-based clustering algorithm (Ester, Kriegel et al. 1996). Two rules that have been used for identifying the outliers are:

DBSCAN Clustering Rule 1. Ambient Wind Speed, Grid Produced Power, eps value = 2, min_samples value = 10

DBSCAN Clustering Rule 2. Ambient Wind Speed, Generator RPM, eps value = 3.45, min_samples value = 10

The results before and after cleaning are shown in **Figure 4**.

3.3. Flowchart for Predicting Produced Power

In order to develop a workable predictive model it is important to understand the process in terms of the structure, environment, and operation. **Section 1** briefly discusses some of these issues and based on this knowledge a simplified flowchart used for calculating Predicted Produced Power is shown in **Figure 5**. The figure also shows that there is a weak correlation between the environmental variables (Ambient temperature, Ambient Humidity and Ambient Pressure) and Grid Produced Power; but there is a strong correlation between Ambient Wind Speed and Grid Produced Power.

3.4. Representation of Variables as Possibility Distribution Functions

As discussed earlier, SCADA data is always encumbered by imperfections. One of the techniques that can be used for handling imperfections of the data is the Fuzzy Logic Approach. In this approach, a fuzzy variable X can be described by its Fuzzy Membership Function, instead of a Probability Density Function

In the Possibilistic Approach, a Fuzzy Membership Function can also be interpreted as a Possibility Distribution Function **(Figure 6)**. α–cut of this Possibility Distribution Function, donated by X_α, is a fuzzy interval $[x, x']$ that contains the values whose likelihood is α. The value of α can be in the range $[0,1]$. At the base, when the value of α is 0, variable has the interval within which the expected value will "certainly" lie. As the value of α increases, the interval between which the values lie decreases, but the certainty that the values will lie within this interval also decreases.

The α–cut of a fuzzy set is given by (Ayyub & Klir, 2006):

$$X_\alpha = [x, x']_\alpha = \{x \in X | x \leq x \leq x'\} \qquad (5)$$

$$\alpha \in [0,1]$$

Where:

x = Lowest real number value of the interval

x' = Highest real number value of the interval

The use of α–cut allows for the concepts of interval analysis to be used (Ayyub & Klir, 2006).

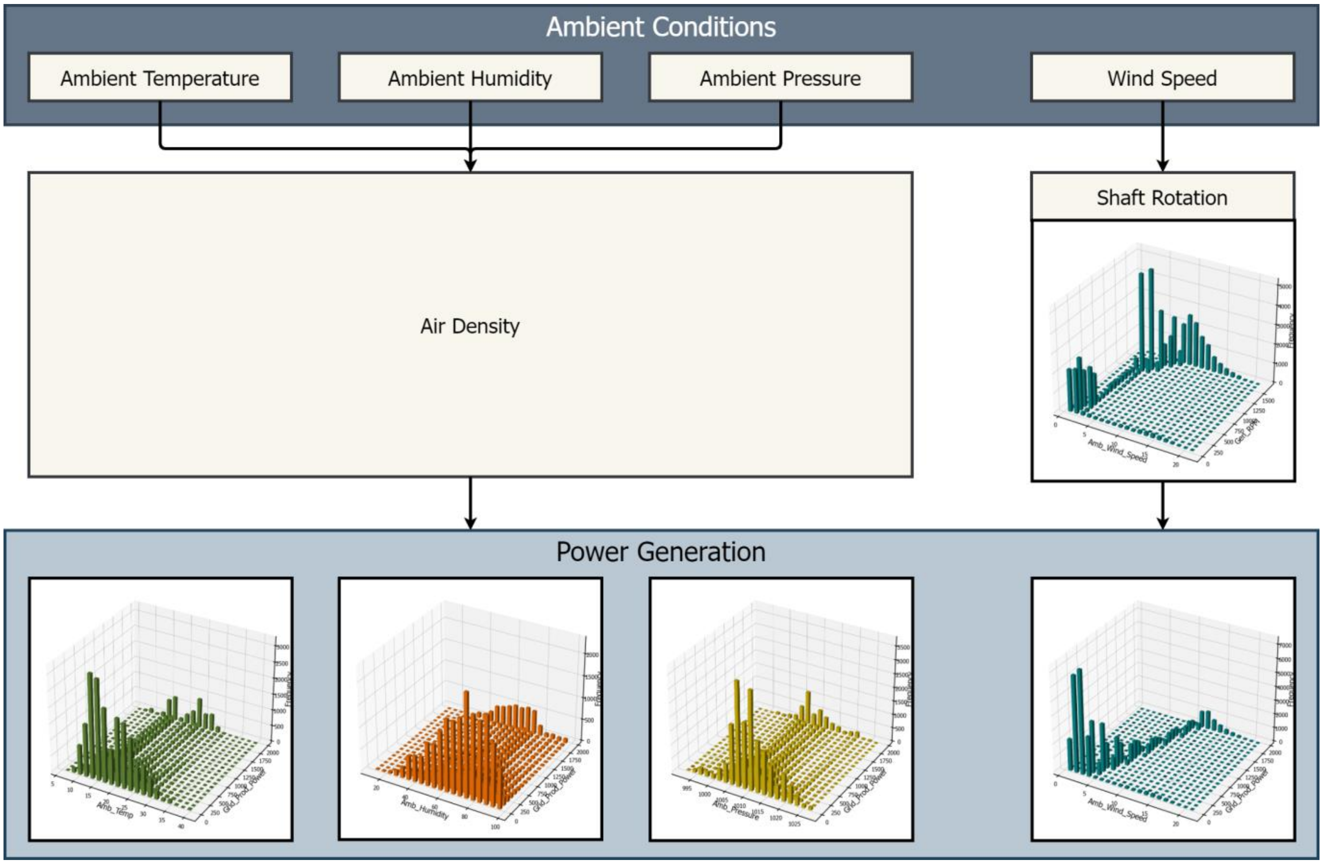

Figure 5. Flowchart showing influence of variables on the calculation of produced power.

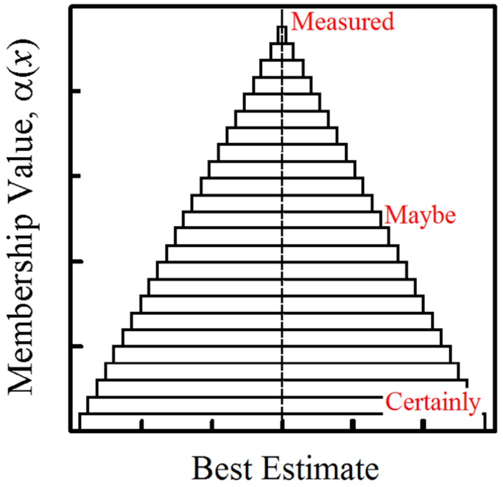

Figure 6. Conceptual illustration of possibility distribution function.

In the absence of detailed study to quantify the interval, limit values that have been used for the calculations are based on the literature and experience. For example, response time and uncertainty of a value recorded by a cup anemometer, depends upon its construction (dimensions, weight, etc.) and degree of deterioration (example, friction caused by corrosion). Under test conditions, a new anemometer can show inaccuracy of about 2%. Under working conditions, this inaccuracy may increase due to corrosion, wear, misalignment, deposition of dust, etc. (Manwell, McGowan & Rogers, 2009). Thus, at $\alpha = 0$ (interval within which the expected value “certainly” lies), the estimated limit of values around the measured values have been estimates as:

- Ambient Temperature : ±1.0ºC
- Ambient Humidity : ±1.0%
- Ambient Pressure : ±1.0 millibars
- Ambient Wind Speed : ±0.5 m/s
- Power Coefficient : 0.45 ±0.05

Possibility Distribution Function for a variable is generated by stacking α number of intervals, where the bottom layer, $\alpha = 0$, has interval range:

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

$$\begin{bmatrix}(measured\ value - estimated\ limit\ value),\\(measured\ value + estimated\ limit\ value)\end{bmatrix}$$

In the Possibilistic Approach, in order to account for the uncertainty, instead of using crisp values of environmental variables (Ambient Temperature, Humidity, Pressure and Wind Speed) as recorded by SCADA and Power Coefficient, Possibility Distribution Functions of the variables are used. Calculations are carried out using interval values at each α–cut. For each value of α, the interval values of variables are determined. Considering all the minimum and maximum values of the intervals, the minimum and maximum values of the output function are calculated using accepted equations. Different combinations that are possible are shown in **Table 2**. The results of all α–$cuts$ are stacked to build the possibility distribution function of the output function (Ayyub & Klir, 2006).

3.5. Possibilistic Approach

The calculations are done in two steps. In the first step, Possibility Distribution Function for Air Density is generated using **Equation 4**. In the second step, the Possibility Distribution Functions for Air Density, Ambient Wind Speed and $C_P(\lambda, \beta)$ are used to generate Possibility Distribution Function for Predicted Produced Power using **Equation 2**.

3.6. Hybrid (Machine Learning – Possibilistic) Approach

Development of the Hybrid (Machine Learning – Possibilistic) is done in two steps.

In the first step, different Machine Learning models are trained using training dataset and the output from the trained models are evaluated. Models that have been evaluated are:

- Linear Models – Linear Regression (LR), Lasso, Ridge, and
- Tree-based Models – Decision Trees, Random Forest (RF)
- Boosting Models – AdaBoost, XGBoost and LGBoost
- Support Vector Regression (SVR)

Out of these models, XGBoost (RMSE = 186, R^2 = 0.93, MAE = 127) has been selected because it gives acceptable fit and takes short calculation time.

In the second step, the trained model and Possibility Distribution Functions of the environmental variables are used to generate Possibility Distribution Functions for Predicted Produced Power. The calculations are carried out according to the method described in the previous section, except that the calculations are done using the trained Machine Learning model instead of the equations.

4. RESULTS AND DISCUSSION

4.1. Possibilistic Approach

4.1.1. Effect of Environmental Variables on Air Density

Figure 7 shows the results of the calculations carried out for predicting Air Density. Since Air Density increases with the increase in Ambient Pressure, but decreases with the increase in Ambient Temperature and Ambient Humidity; the graph shows seasonal variations of the Air Density. The graph also shows sensitivity to the inaccuracies of recorded values and the “true” value may lie anywhere within the $Certain - Min$ and $Certain - Max$ curves.

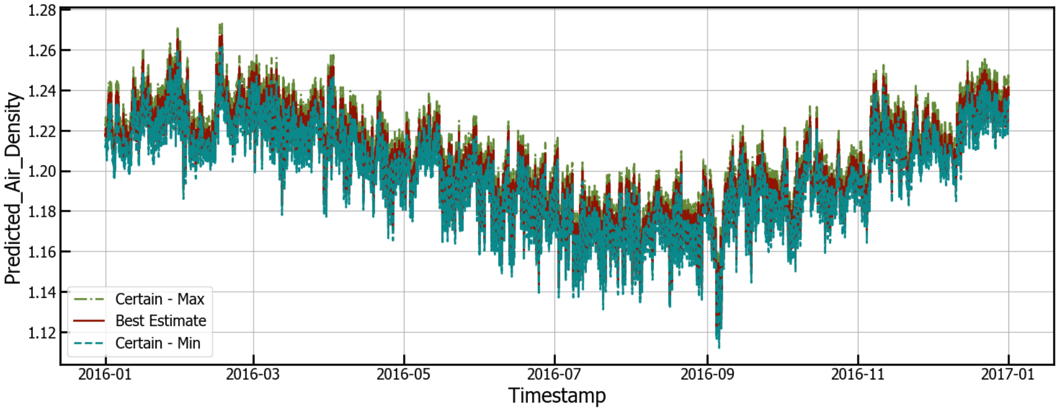

Figure 7. Seasonal variation on Predicted Air Density at $\alpha\text{–}cut = 0$.

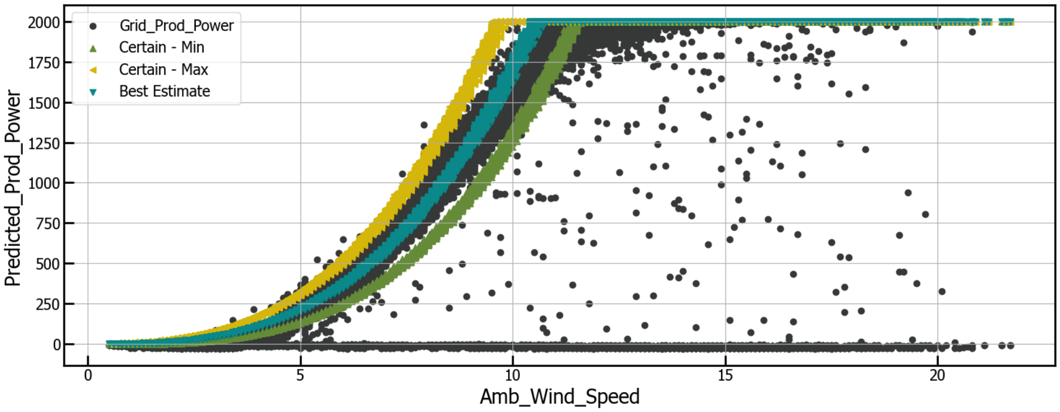

Figure 8. Effect of Ambient Wind Speed on Predicted Produced Power using Possibilistic Approach at $\alpha\text{–}cut = 0$.

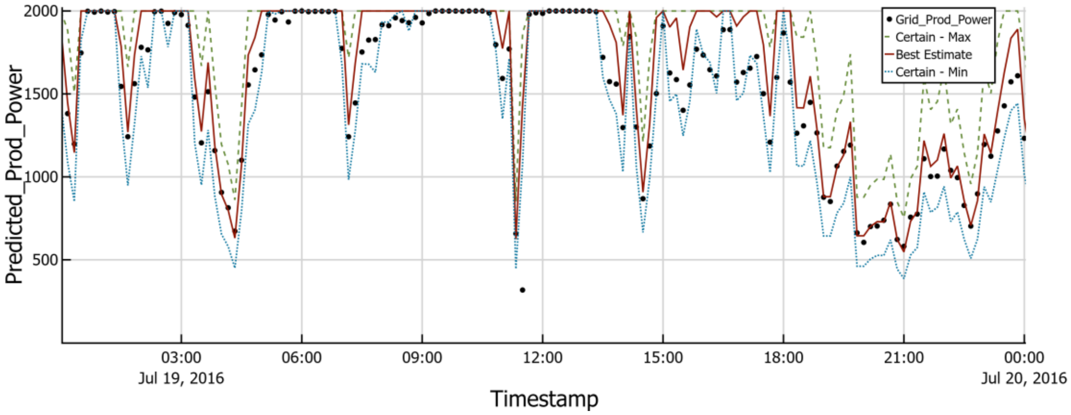

Figure 9. Plot of Grid Produced Power and Predicted Produced Power calculated using Possibilistic Approach at $\alpha\text{–}cut$=0 for a 24 hour duration (19th July, 2016).

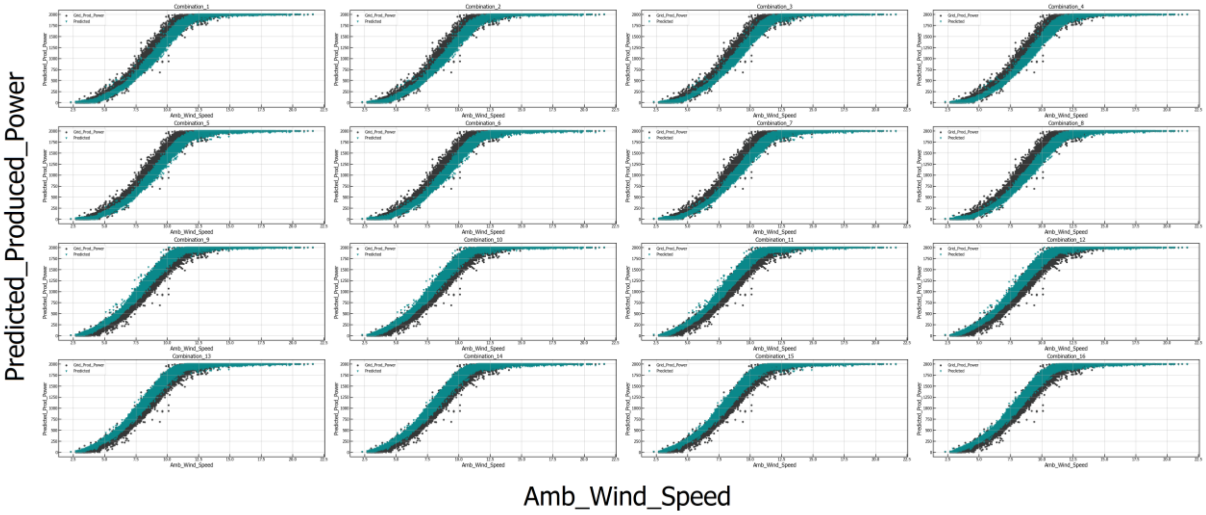

Figure 10. Predicted Produced Power using Hybrid Model for the combinations of interval values given in **Table 2** at $\alpha{-}cut$=0.

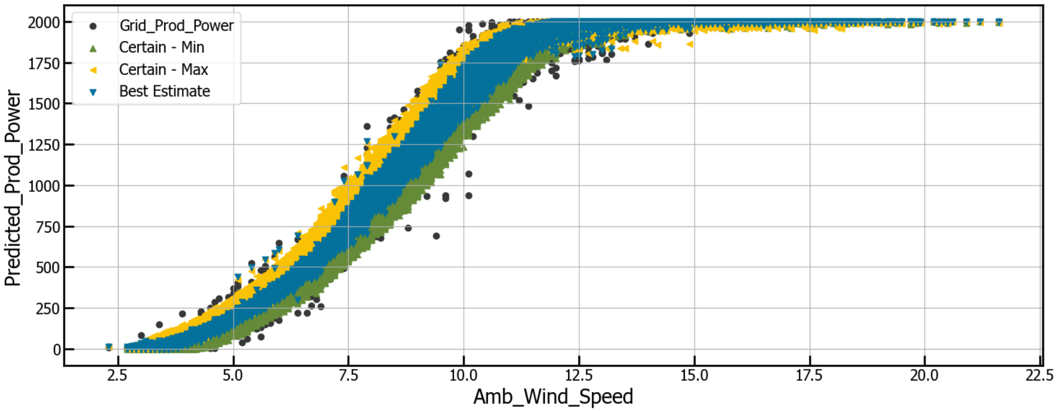

Figure 11. Effect of Ambient Wind Speed on Predicted Produced Power using Hybrid Approach at $\alpha{-}cut = 0$. $Certain - Min$ is obtained from Combination_6 and $Certain - Max$ is obtained from Combination_11.

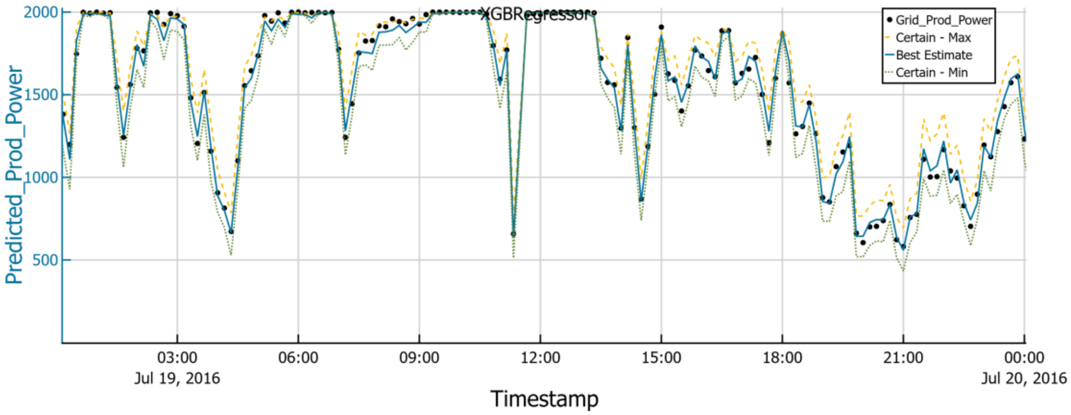

Figure 12. Plot of Grid Produced Power and Predicted Produced Power calculated using Hybrid Approach at $\alpha{-}cut$=0 for a 24 hour duration (19th July, 2016).

4.1.2. Effect of Environmental Variables on Predicted Produced Power

Figure 8 shows the effect of Ambient Wind Speed on the Predicted Produced Power. The graph shows that:

- Power curve developed according to the **Equation 2** does not follow the actual trend. A better model, as proposed by Saint-Drenan, Y.-M. et al. (2020), may give better result.
- Spread of measured Grid Produced Power at a particular wind speed has not been accounted for. The spread can arise due to various reasons, like, control of the operation and imperfections in measurements.
- Predicted produced power is sensitive to the inaccuracies of recorded values and the “true” value may lie anywhere within the $Certain - Min$ and $Certain - Max$ curves.

Figure 9 shows plot of Predicted Produced Power and Grid Produced Power for a 24-hour duration (19th July, 2016). The graph shows that measured values generally lie within the boundaries set by $Certain - Min$ and $Certain - Max$ values.

4.2. Hybrid (Machine Learning – Possibilistic) Approach

Figures 10-12 show the results of calculations carried out using Hybrid (Machine Learning – Possibilistic) Approach. **Figure 10** shows the effect of max and min interval values of environmental variables on Predicted Produced Power. The figure shows that combinations have significant effect on the Predicted Produced Power.

According to **Equation 2**, Predicted Produced Power is proportional to cube of Ambient Wind Speed. Hence, Combination_1 to Combination_8 show lower values of Predicted Produced Power as compared to Combination_9 to Combination_16. Within these two sets of combinations, the differences are small because of the relatively small differences in the calculated air density.

Figure 11 shows the effect of Ambient Wind Speed on Predicted Produced Power using Hybrid Approach at $\alpha{-}cut = 0$. The figure shows significant effect of measurement uncertainties on the predicted values. $Certain - Min$ is obtained from Combination_6 and $Certain - Max$ is obtained from Combination_11.

Figure 12 shows plot of Grid Produced Power and Predicted Produced Power calculated using hybrid approach at $\alpha{-}cut$=0 for a 24-hour duration (19th July, 2016). The graph shows that measured values generally lie within the outer most boundaries set by $Certain - Min$ and $Certain - Max$ values.

A comparison between **Figure 9** and **Figure 12** shows that, in general, (a) Machine Learning model fits better than the parametric model; and (b) the difference between $Certain - Max$ and $Certain - Min$ in the Hybrid Model is less than that in the Possibilistic Model.

5. CONCLUSIONS

This paper presents a simple yet robust methodologies for calculating Predicted Produced Power using SCADA data while accounting for variability and uncertainty. The methodologies utilise either parametric or Machine Learning models for handling variability; and Possibilistic Approach for handling uncertainty. As a case study, the idea has been demonstrated using real-life SCADA data.

To take the research work further, the following tasks have been identified:

- The models do not account for effect of control measures of the wind turbine on produced power. Since, these measures can significantly effect power generation (López-Queija et al., 2022); models that account for control measures need to be used.
- Grid Produced Power has been assumed to have crisp values, but in reality measurement of Grid Produced Power is also afflicted with uncertainties. Hence, calculations need to be done by representing it by a Possibility Distribution Function.
- Having obtained Possibility Distribution Functions of Predicted Produced Power and Grid Produced Power, *Likelihood of Sub-optimal Performance* can be determined using the concepts of Possibility and Necessity Measures.

DATA AVAILABILITY

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

- https://www.edp.com/en/wind-turbine-scada-signals-2016
- https://www.edp.com/en/innovation/open-data/wind-turbinescada-signals-2017.

REFERENCES

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

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

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

Duguid, L. (2018), 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

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996), 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

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

Global Wind Energy Council (2021). "Global Wind Report 2021", available at: https://gwec.net/wp-content/uploads/2021/03/GWEC-Global-Wind-Report-2021.pdf

López-Queija, J., Robles, E., Jugo, J., Alonso-Quesada, S. (2022). Review of Control Technologies for Floating Offshore Wind Turbines, Renewable and Sustainable Energy Reviews Vol. 167, 112787

Lydia, M., Kumar, Suresh Kumar, S., Selvakumar, A. I., Prem Kumar, G. E. (2014). A Comprehensive Review on Wind Turbine Power Curve Modeling Techniques, Renewable & Sustainable Energy Reviews, Vol. 30, pp.452-460

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

Nilsson, J., and Bertling, L. (2007). 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

Ouyang, T., Kusiak, A., He, Y. (2017). Modeling Wind-Turbine Power Curve: A Data Partitioning and Mining Approach, Renewable Energy, Vol. 102, pp. 1-8

Pandit, R. and Wang, J. (2024). 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

Pandit, R. K., Infield, D. and Kolios, A. (2019). Comparison of Advanced Non-Parametric Models for Wind Turbine Power Curves, IET Renewable Power Generation, Vol. 13(9), pp. 1503-1510

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

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

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

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

Wang, S., Huang, Y., Li, L., Liu, C. (2016). Wind Turbines Abnormality Detection Through Analysis of Wind Farm Power Curves, Measurement, Vol. 93, pp. 178–188

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