Accounting for Environmental Conditions in Data-Driven Wind Turbine Power Models

Ravi Pandit, David Infield, and Matilde Santos

***Abstract*—Continuous assessment of wind turbine performance is a key to maximising power generation at a very low cost. A wind turbine power curve is a non-linear function between power output and wind speed and is widely used to approach numerous problems linked to turbine operation. According to the current IEC standard, power curves are determined by a data reduction method, called binning, where hub height, wind speed and air density are considered as appropriate input parameters. However, as turbine rotors have grown in size over recent years, the impact of variations in wind speed, and thus of power output, can no longer be overlooked. Two environmental variables, namely wind shear and turbulence intensity, have the greatest impact on power output. Therefore, taking account of these factors may improve the accuracy as well as reduce the uncertainty of data-driven power curve models, which could be helpful in performance monitoring applications. This paper aims to quantify and analyse the impact of these two environmental factors on wind turbine power curves. Gaussian process (GP) is a data-driven, nonparametric based approach to power curve modelling that can incorporate these two additional environmental factors. The proposed technique's effectiveness is trained and validated using historical 10-minute average supervisory control and data acquisition (SCADA) datasets from variable speed, pitch control, and wind turbines rated at 2.5 MW. The results suggest that (i) the inclusion of the additional environmental parameters increases GP model accuracy and reduces uncertainty in estimating the power curve; (ii) a comparative study reveals that turbulence intensity has a relatively greater impact on GP model accuracy, together with uncertainty as compared to blade pitch angle. These conclusions are confirmed using performance error metrics and uncertainty calculations. The results have practical beneficial consequences for O&M related activities such as early failure detection.**

***Index Terms*—Condition monitoring, fault detection, gaussian process, power curves machine learning, SCADA data.**

I. Introduction

IN THE last decade, wind power has witnessed significant growth worldwide. By the end of 2021, the annual net wind capacity addition is expected to reach 68 GW despite the impact of the Covid-19 pandemic [1]. According to preliminary World Wind Energy Association statistics [2], 93 GW of new turbines were installed in 2020 alone, as shown in Fig. 1, where China, the United States, and Russia all set new installation records, while most European markets saw only modest expansion. With the significant increase in rotor blade size, operation and maintenance (O&M) costs of wind farms have risen correspondingly. Therefore, the wind power must be cost-effective in order to compete with traditional generation sources in the long term.

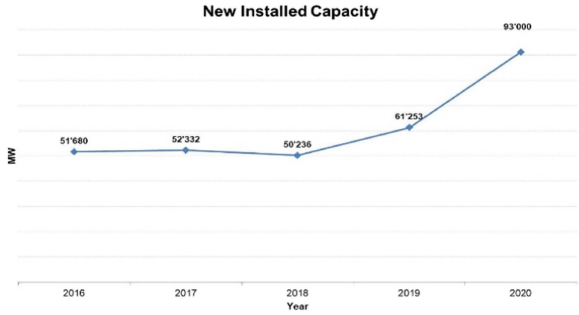

Fig. 1. New installed capacity, 2016-2020 [1].

Compared to onshore wind farms, those located offshore are less mature and face different environmental challenges (e.g., lightning and extreme winds). Because of this, the O&M cost is significantly higher offshore and is estimated to account for 20%–30% of the lifetime costs of an offshore wind farm mainly due to logistics and transportation challenges [3]. As per [4], the global cost of wind O&M is projected to hit $27.4 billion by 2025, with an expected compound annual growth rate of 8%. Unexpected component failures trigger unscheduled maintenance, which is particularly problematic for wind turbine operators. Thus, wind farm operators are gradually adopting condition-based maintenance philosophies to prevent such occurrences and minimise O&M costs by improving turbine efficiency.

Accurate monitoring of wind turbine power generation performance can support more rational maintenance planning, prevent failures [5], and reduce O&M costs [6]. Wind power curve modelling is commonly used to assess wind turbine power generation efficiency (Power Curve Grouping Group). Due to technological advancement, wind turbines are generally equipped with SCADA systems that record extensive historical and operational data from wind turbines. These data are in recent years started finding an application in performance [7] as well as condition

Manuscript received 5 October 2021; revised 21 January 2022, 20 May 2022, 30 June 2022, and 1 September 2022; accepted 1 September 2022. Date of publication 5 September 2022; date of current version 19 December 2022. This work was partially supported by the Spanish Ministry of Science, Innovation and Universities under MCI/AEI/FEDER Project under Grant RTI2018-094902-B-C21. Paper no. TSTE-01021-2021. *(Corresponding author: Ravi Pandit.)*

Ravi Pandit is a lecturer in Instrumentation and AI, School of Aerospace, Transport and Manufacturing, Cranfield University, MK43 0AL Bedford, U.K. (e-mail: ravi.pandit@cranfield.ac.uk).

David Infield is with the Electronic and Electrical Engineering, University of Strathclyde, G1 1XQ Glasgow, U.K. (e-mail: david.infield@strath.ac.uk).

Matilde Santos is with the Complutense University of Madrid, 28040 Madrid, Spain (e-mail: msantos@ucm.es).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSTE.2022.3204453.

Digital Object Identifier 10.1109/TSTE.2022.3204453

monitoring [8] tasks. However, it is worth noting that for larger turbines SCADA system suffers from prediction reliability and accuracy issues, mainly due to erroneous measurements by sensors, which may affect estimating the average power generated by such wind turbines, resulting in time-consuming maintenance plans and resource waste [9]. However, suitable, pre-processing of datasets ensures more accurate wind power calculation and increases SCADA-data-based model accuracy [10].

The International Standard IEC 61400-12-1 [11] for measuring the power curves utilises a standard data reduction technique known as 'binning'. Given a turbine at a location within a particular wind farm, this curve is referred to as a site-specific power curve for the specific wind turbine. While calculating power curves, a significant database covering a wide range of operating conditions is required, and this is generally obtained over a significant time period. It has been found that because of changes in air density, atmospheric stability and other aspects of operational conditions that influence power performance, a notable difference in power output at a given hub height wind speed is seen [12], [13], [14]. Though Section I-B of the IEC only considers hub height wind speed and air density, it is, however, worth mentioning that the IEC standard IEC 61400-12-1 (2017) also has informative annexes that discuss turbulence normalisation, based largely on the work of Axel Albers (Annex M) and on wind shear normalisation (Annex P). As a result, attention has been given to improving site-specific power curve predictions by considering various factors.

A. Related Work

For power curve modelling, various SCADA data-based parametric and nonparametric methods have been used in the past; the results show that nonparametric models generally perform better than parametric models in part due to their ease of operation and flexibility [15], [16]. Nonparametric approaches, often applying data-driven machine learning methods, have recently become popular in power curve modelling and its related applications such as performance and condition monitoring of wind turbines [8], [9] and wind farms [17], [18], where 10-minute averaged historical SCADA data have been used for training and validation purposes. Shen et al. [19], for example, suggested that a nonparametric model is better suited to working with large datasets than parametric models because it can integrate the effects of various parameters other than wind speed on power curves more easily. Wang et al. [20] proposed a copula-based joint probability model for modelling wind turbine power curves, and outliers were detected based on the derived joint probability distribution. In [21], a multi-layered neural network power curve model has been shown to verify the performances and fault diagnosis in turbines. Wind power forecasting [22], wind resource assessment [23], wind turbine site matching [24], and power system reliability assessment [25], among others, are considered to be essential applications of power curves in power systems.

Wind power generation is known to be affected by many different factors, such as air density, wind shear, turbulences, etc, so the inclusion of these factors is important to improve data-driven models' accuracy. For example, Pandit et al. [13] incorporated air density as the second input variable along with hub height wind speed, for a Gaussian Process (GP) based power curve, and the result suggests significant improvement in accuracy and a reduction in uncertainty. It should be noted that considerable changes in air density (which is wind farm location-specific) will add significant uncertainty to long-term energy yield prediction from turbines. Another important observation is found in [26], where air density correction (via temperature and pressure) is suggested for improving the power curve and wind resource assessment of a wind farm. Adjusting resource measurements to ensure they represent the same air density as the one for which a power curve is considered valid is standard practice in wind resource assessment and energy yield estimation. According to existing literature [27], [28], as wind shear increases, power generation decreases (within a certain permissible range of shear exponents). Wagner et al. [27] proposed that a hub height equivalent to wind speed be used in the power curve derivation. With wind shear taken into account in this manner simulations based on a blade element model were tested to show a reduction in power curve scattering. The Wagner et al. [28] study on wind shear effect on wind turbine power curves proposed and concluded that power curves are less sensitive to shear, thus less dependent on the site.

Clifton et al. [29] proposed three methods for accounting for the effect of turbulence on wind turbine power curves. They used IEC standard binning, turbulence normalization and random forest techniques and comparative analysis showed that a random forest model can estimate the power as conditions change, and is more flexible than the alternatives examined. Turbulence intensity is another environmental factor that affects wind power generation. Bardal et al. [30] analysed the effect of turbulence intensity and wind shear on the performance of a 3 MW wind turbine using lidar data. A new technique for turbulence normalization is presented in [31]. This approach defines a zero turbulence power curve that may be used to generate power output using either a measured or reference wind speed distribution. Recent works on turbulence intensity and wind turbines can be found in [32], [33], [34], [35], [36], [37].

B. Scientific Novelty and Contribution to Knowledge

From the above works of literature, it is well established that the previously mentioned factors have significant influences on the power performance of a wind turbine. However, the IEC 61400-12-1 standard considers only wind speed and air density as the relevant input variables and ignores other operational and environmental factors that are known to influence wind power generation. Pandit et al. [38] carried out a comparative analysis of operational variables (rotor speed and blade pitch angle) and found rotor speed as a key variable that improves power curve model accuracy and uncertainty. They further extended this work in [39] and developed a fault detection algorithm incorporating rotor speed and compared this with existing techniques and concluded that including rotor speed in the algorithm increases early fault detection capability without false positive alarms. But both have ignored the importance of environmental factors on

wind turbine power performance. As a result, it is unclear if including environmental variables (wind shear and turbulence intensity) would increase data-driven power curve accuracy and reduce uncertainty. This research is vital as it can help to identify the most appropriate parameters that improve data-driven power curve model accuracy, and further, that can be useful in constructing robust fault detection algorithms for wind turbine condition/performance monitoring purposes. This paper aims to fill this knowledge gap by proposing an in-depth analysis of the impact of these two environmental variables on a power curve model based on a GP. This is the driving force behind the work that is presented here. In addition, as a practical contribution, we are suggesting key variables that if included in the power curve-based condition monitoring model may improve O&M-related activities such as early failure detection and thus, reduce costs.

The remainder of this paper is organised as follows. Section II describes the wind turbine power curve and factors affecting it, while Section III presents a description of operational wind turbine SCADA data and appropriate pre-processing methodologies. The GP methodology for power curve modelling is explained in Section IV. Section V investigates the effect of environmental variables on the GP power curve model accuracy and uncertainty, further divided into two subsections. Comparative output analyses for the GP model incorporating environmental parameters are presented in Section VI. Finally, Section VII provides conclusions drawn from the research and an indication of useful future work.

II. Wind Turbine Power Curves and Factors Affecting It

The wind shear, turbulence intensity and air density have strong relationships with topography and meteorological conditions and therefore, their impacts on the power curve model are important for wind farms condition monitoring and performance analysis purposes as described as follows.

A. Wind Speed and Air Density

A wind turbine power curve is a nonlinear graph that specifies how much electrical power the turbine will produce as a function of wind speed, as shown in Fig. 2(10-min mean values from the SCADA system). Many existing methods concentrate primarily on the power curve, although data spread, such as the probability distribution of power points in a given wind speed bin of a typical power curve, certainly contains useful information. Thus, an accurate power curve model aids wind power suppliers in capturing the performance of wind turbines. Mathematically the power curve is described by the following equation, [11]:

$$P \;=\; 0.5\,\rho A C_p\,(\lambda, \beta\,)\;\; v^3 \tag{1}$$

where ρ is air density (kg/m^3), A is swept area (m^2), and v is the hub height wind speed (m/sec). C_p is the power coefficient and depends on tip speed ratio (λ) and pitch angle (β) turbine parameters. Generally, the higher the wind speed is, the more power can be generated and this can be shown in (1) where wind power is proportional to the cube of wind speed below rated wind speed where the tip speed ratio and blade angle are fixed. It should be noted that generator losses are conventionally absorbed into C_p, unless otherwise stated. Furthermore, as shown in (1), wind power output is also directly proportional to the air density and therefore affects power production. Ambient temperature and atmospheric pressure affect the air density which consequently affects the wind turbine power curve. Therefore, as per IEC standard 61400-12-1, air density correction for variable pitch-regulated wind turbines must be made by using the following equations:

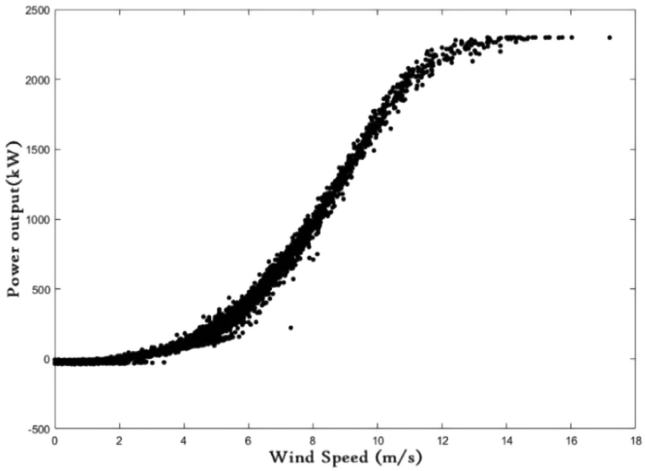

Fig. 2. Measured power curve.

$$\rho = \; 1.225 \left[\frac{288.15}{T}\right]\left[\frac{B}{1013.3}\right] \tag{2}$$

$$V_C = \; V_M\Big[\;\frac{\rho}{1.225}\;\Big]^{\frac{1}{3}} \tag{3}$$

where V_C and V_M are the corrected and measured wind speed in m/sec. It should be noted that *B* is atmospheric pressure in $mbar$, and *T* the temperature in Kelvin. It is worth noting that equation (2) represents the approximate relationship between air density, temperature and pressure; however, the impact of humidity on air density is small compared to ambient temperature. The air density (ρ) is calculated by putting the SCADA 10-minute average ambient temperature and pressure values into (2), and then corrected wind speed (V_C) is computed by (3), by inserting the calculated air density into it. Finally, power output is available as a function of corrected wind speed; this is called the air density corrected power curve, and is used in the upcoming sections.

B. Wind Shear and Turbulence Intensity

Wind shear is the variation of wind speed with height, and is significant in the lower atmosphere; it is also known as wind gradient. The wind profile power law describes an approximate relationship between wind speeds at different heights, and can be used to compute the wind shear exponent as follows:

$$v/v_0 = \left(h/h_0\right)^{\alpha} \tag{4}$$

where, $v =$ the wind speed at height $h\;(m/sec)$; $v_0 =$ the wind speed at a height $h_0\;(m/sec)$, and $\alpha =$ the wind shear exponent.

It should be noted that the equation is only valid under the assumption of stable atmospheric conditions [15].

Turbulence is another environmental factor that needs to be included for an accurate site-specific wind power curve. Turbulence is quantified in the wind energy industry using a metric called turbulence intensity (TI), which is the standard deviation of horizontal wind speed divided by the average wind speed over a time period, typically 10 minutes [40]. The impact of turbulence on a typical power curve can be broken down into two parts. First, the normal 10-minute averaging of power and wind speed data have an effect. Second, the non-linearity of the power curve, when considered as a transfer function between wind speed and power, enables the produced power to be dependent mostly on the variance and the average of the wind speed [41]. Therefore, affects the different wind speed regions of the power curve. It is worth noting that topography can have a strong influence on both shear and turbulence, and this can introduce a directional dependence, while atmospheric stability can introduce a diurnal and seasonal dependence.

C. Wind Direction and Blade Pitch Angle

One of the most often utilised elements in wind power forecasting models is wind direction. However, it has less of an impact on producing wind power as compared to wind speed and is hence not covered in this study. Nevertheless, wind direction changes impact yaw control which plays an important role in minimising yaw drive misalignment to boost wind power production [42].

The blade pitch angle is recorded in the SCADA system and is used to manage the blades of a wind turbine to employ the right amount of available wind speed in order to produce controlled power making sure it does not surpass its rated power. Blade pitch angle may make a considerable effect on power estimations when wind speed is above or close to its rated value. Since most of the useful power is produced in-between cut-in and rated wind speed. Thus, blade pitch angle will have a far lesser impact on the wind power curve [38] and it is not included in this research. In short, three environmental variables, namely, a) wind shear; b) turbulence intensity and c) air density are incorporated in this paper to investigate their impacts on data-driven algorithms accuracy and uncertainty.

III Exploratory Scada Data Analysis and Filteration

SCADA records a massive amount of datasets used to reflect the operational status together with performance without additional costs. SCADA data used in this investigation comes from a 2.3 MW Siemens (SWT-2.3-108) variable pitch controlled wind turbine that is currently in operation and owned by ScottishPower Renewables. This is an onshore WT located on Eaglesham Moor, 15 km from Glasgow, Scotland. It measures 91,44 m to the tip of the blade. Each blade is 45,72 m in length. The turbine makes 17 rpm turning at 360° tearing the wind at a speed of 150 mph. Each turbine at the Whitelee windfarms has a tip height of 110 metres from ground level to the hub plus the rotor radius. The rotor blades measure 45 metres in length. The location is 11.5 kilometres broad (east-west) and 7 kilometres long (north-south), with a height of 370 metres above sea level. The wind farm's main access road is 16.5 kilometres long, with additional 70 kilometres of tracks connecting the turbines. Also, the terrain is reasonably flat but the presence of many forest patches and clearing may cause additional turbulence.

In this study, due to the lack of information about the wind direction sector, it is not possible to analyse how the inflow of the wind turbine is distorted by the wake of the adjacent wind turbines. This source of uncertainty may not allows us to consider some turbine power losses and fatigue loads caused by the wake effects, that have been ignored and all data have been used even when the turbine was exposed to wakes from neigbouring turbines.

Data are recorded 10-minute mean, maximum, and standard deviation values for over 100 variables, including timestamp, wind speed, rotor speed, power output, ambient temperature, air pressure, and so on. In this study, 10-min mean values are considered. Due to a confidentiality agreement, important turbine information is excluded; however, a sample of SCADA data used in this paper is provided in Table I.

The raw data gathered from SCADA systems may contain sensor faults and communication problems, resulting in missing data and, if not pre-processed, affecting the performance of created models. As a result, before using these data for further analysis, the first task is to filter them. Pre-processing of the raw data was done similarly to that outlined in [43]. The first step is to filter out samples with missing values or negative power values. Data points with a maximum wind speed of more than 25 m/sec are additionally filtered out because the turbine is usually shut down at this speed. Furthermore, during low-wind-speed periods, data sampling during repeated start-up or stop may have a different variance. Thus, a lower limit of output power is kept at 0 kW for data pre-processing.

Overall, criteria including timestamp missing, negative power values, out-of-range values, and turbine curtailment are utilised to filter out misleading data like that reported in [43], [39]. Table II summarises a SCADA dataset that has a starting time-stamp of "01/3/2012 00:00 PM" and ending at the time-stamp of "30/05/2012 00:00 PM". It contains 14465 measured values that were reduced to 9677 data points after filtration using the above-mentioned criteria. The air density corrected (as explained before) power curve of the filtered SCADA data is shown in Fig. 3 and will be used in the subsequent investigation.

IV. Gaussian Process Methodology

For nonlinear inference, GP is a powerful nonparametric machine learning technique for building probabilistic models of real-world problems [44]. It's a stochastic process with a joint Gaussian distribution for any finite number of collections [45]. A GP model is unique in that it allows you to describe prior distributions over functions in a straightforward way. For given training datasets $D = \{(X_n, y_n), n = 1, 2, 3, \ldots.., N\}$, where the input is $X_n \in R^{d_x}$, the output $y_n \in R$. Here d_x is the dimension of the input. The relationship between input and the target value is modelled in GP regression as:

$$y = f(X) + \in \tag{5}$$

TABLE I
SCADA DATA SAMPLES

TimeStamp	Wind speed (Avg.) m/sec	Power (Avg.) kW	Ambient temp (Avg.) ℃	Atmospheric pressure (Avg.) mbar	Blade pitch angle (Avg.) degree
01/ 03/2012 00:00:00	7.57	788.15	8.19	986.16	-0.97
01/ 03/2012 00:10:00	8.17	991.83	8.22	986.26	-0.97
01/ 03/2012 00:20:00	7.99	985.08	8.33	986.22	-0.98
01/ 03/2012 00:30:00	7.69	879.38	8.38	986.12	-0.98
01/ 03/2012 00:40:00	8.43	938.32	8.85	986.29	-0.99

TABLE II
SCADA DATA PRE-PROCESSING SUMMARY

Start timestamp	End timestamp	Measured data	Filtered data
01/3/2012 00:00 PM	30/05/2012 00:00 PM	14465	6677

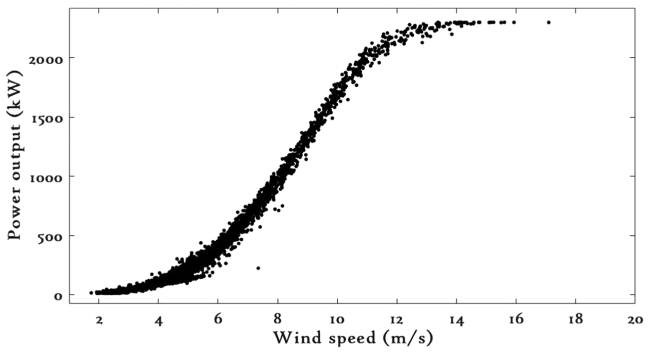

Fig. 3. Pre-processed power curve.

where f is the latent function and $\in$ represents i.i.d. (independent, identically distributed) Gaussian noise with zero mean and variance $\sigma_n^2, i.e., \in\sim N(0, \sigma_n^2)$. It should be highlighted that the target value is $y = (y_1, y_2, \ldots\ldots, y_N)^T$ while X is the input, mathematically represented as $X = [x_1, x_2, \ldots\ldots, x_N]^T$. The latent function $f(X)$ has a GP prior, which is defined by $f(X) \sim GP(m(X), k(X, X'))$. For a real process X, the mean function $m(X)$ and covariance function $k(X, X')$ can be written as:

$$m\ (X) = E\left[f\ (X)\right] \tag{6}$$

$$k\ (X, X') = E\left[\left(f\ (x) - m\ (X)\right)\left(f\ (X') - m\ (X')\right)\right] \tag{7}$$

The accuracy of a GP model is determined by the covariance function $k(X, X')$, also known as a kernel (a positive-definite function), which quantifies the similarity between two points. Even though it can be randomly chosen, the mean function $m(X)$ is commonly assumed to be zero for notational simplicity because it is preferred to centre the observed data around a zero mean. The most generally used covariance function is the squared exponential function, which is a stationary function that will be employed in this research to explain the nonlinear relationship between wind speed and wind turbine power output mathematically; it is defined as:

$$k\ (x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right) \tag{8}$$

The squared exponential covariance function is a function of Euclidean distance and to minimise the impact of noise, a noise term is added to it, as follows.

$$k\ (x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right) + \sigma_n^2 \delta\left(x, x'\right) \tag{9}$$

where σ_f^2 signifies the signal variance while the length scale, l, is used to define how quickly the covariance decreases with respect to the distance between the points. Both of these factors are commonly known as hyper-parameters. The marginal distribution over any set of input points must have a joint multivariate Gaussian distribution, according to the definition of a GP. As a result, we have:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K\left(X, X\right) & K\left(X, X_*\right) \\ K\left(X_*,\ X\right) & K\left(X_*, X_*\right) \end{bmatrix}\right) \tag{10}$$

where $K(X, X_*), k(X_*, X_*)$ are constructed by using (5), and $K\ (X_*,\ X) = K(X, X_*)^T$. From the i.i.d. noise assumption, we have that,

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_n^2 I & 0 \\ 0 & \sigma_n^2 I \end{bmatrix}\right) \tag{11}$$

Because the sums of independent Gaussian, random variables are Gaussian (11) can be modified as follows,

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K\left(X, X\right) + \sigma_n^2 I & k\left(X, X_*\right) \\ k\left(X_*, X\right) & k\left(X_*, X_*\right) + \sigma_n^2 I \end{bmatrix}\right) \tag{12}$$

The prior distribution is helpful in providing important information about the unknown parameters. The posterior distribution is created by combining the prior distribution with the probability distribution of future data, which is useful for inference and any decisions involving uncertain parameters. The posterior distribution $p(y * |y)$ reflects the likelihood of a prediction y_* given data y and is given by:

$$y_* | \mathrm{y}, \mathrm{X}, X_* \sim \mathcal{N}\left(\mu_*, \Sigma_*\right) \tag{13}$$

where μ_* is the mean vector and Σ_* is a covariance matrix. These are calculated using the following equations:

$$\mu_* = k\left(X_*, X\right)\left(K\left(X, X\right) + \sigma_n^2 I\right)^{-1} y \tag{14}$$

$$\Sigma_* = k\left(X_*, X_*\right) - k\left(X_*, X\right)\ \left(K\left(X, X\right) + \sigma_n^2 I\right)^{-1} k\left(X, X_*\right) + \sigma_n^2 I \tag{15}$$

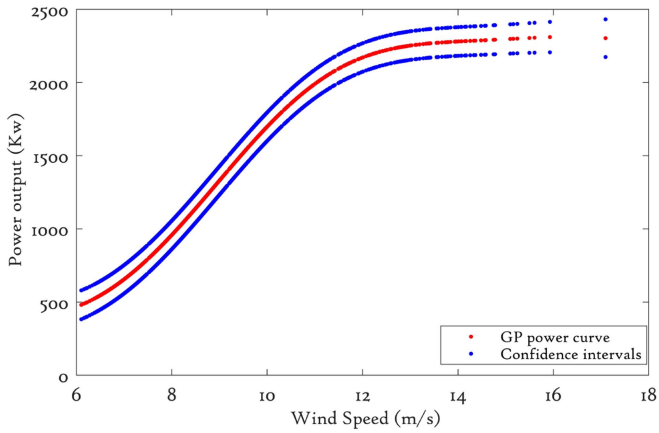

Fig. 4. Gaussian Process estimated power curve.

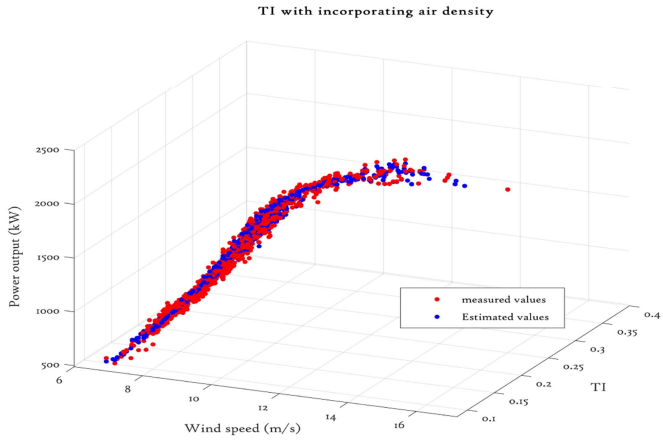

Fig. 5. GP power curve incorporating turbulence intensity.

If P is the new dataset, then, the prediction for a new sample x_{new} is easily calculated by using the following formula,

$$\hat{y}_{new} = \frac{1}{P} \sum_{p=1}^{P} \mu_p \left(x_{new} \right) \tag{16}$$

where $\mu_p(x_{new})$ is the GP regression model's prediction for a new sample dataset. The estimation of $\hat{y}_{new}$ takes into account the sampling variability caused by missing data. Five-fold cross-validation has been used to nullify the impact of missing data on model performance.

To assess the performance of GP models, confidence intervals (CIs) are found to be key and hence included in later sections of this paper. These GP CIs are model-based estimates that provide information on the uncertainty surrounding an estimate. A GP estimates CIs for the prior and posterior for each predicted mean value that represents the pointwise mean plus and minus two times the standard deviation for a given input value (corresponding to a 95% confidence interval, that corresponds to a significance level of 0.05) [13].

Power curves were produced with the GP models outlined above (using MATLAB and Python) on the filtered SCADA datasets (of Section III), and the results are shown in Fig. 4. The result shows the GP power curve with confidence intervals (CIs) and suggests that the GP is able to estimate the power curve accurately. Indeed, the GP model overlaps the curve except for some red points at high wind speed where it is possible to differentiate them. Note that CIs are smaller between the 6 m/sec to 12 m/sec wind speed region as compared to other regions. This is because of the concentration of data in this region. Also, for low wind speed, GP estimates negative power indicating GP assumptions may not work with low wind speed data. A GP model like any machine learning technique suffers from a cubic inversion problem that affects its accuracy together with uncertainty and increases the computation cost. The posterior conditional distribution for a given observation is defined mathematically in [46], and there is a covariance matrix component, associated with the inverse matrix operation which leads to the mathematical challenge of inverting an $n \times n$ matrix (and this goes approximately with $\mathrm{O(n^3)}$, where n is the number of data points). This is the cubic inversion challenge, and therefore, for effective GP modelling, maintaining a balance between the numbers of data points employed and the cost of computation is important.

V. Incorporating Environmental Parameters

The covariance function, as previously stated, is the heart of the GP model and is used to represent the similarity between two points. The variance of each variable along the leading diagonal is given by the general covariance matrix, while the off-diagonal elements measure the correlations between the individual variables and are mathematically expressed as follows:

$$k\ (x, x') = \begin{bmatrix} k\left(x_1, x_1\right) & \cdots & k\left(x_1, x_n\right) \\ \vdots & \ddots & \vdots \\ k\left(x_n, x_1\right) & \cdots & k\left(x_n, x_n\right) \end{bmatrix}$$

Where k is of size $n \times n$, and being n the number of input parameters is considered, and it must be symmetric and non-negative semidefinite.

Because of the intrinsically multivariate nature of a general GP, a large number of predictors can be incorporated into the GP model to examine their impact on GP model accuracy and uncertainty. Hence, environmental parameters (air density, wind shear and turbulence intensity) can be included along with wind speed to train the GP model, and the results are shown in a 3D scatter plot to analyse the effect of incorporated variables on the GP power curve accuracy.

A. Accounting for the Effect of Turbulence Intensity

The ratio of the standard deviation of wind speed to the mean wind speed is known as turbulence intensity (TI). TI is found to be critical in wind turbine construction, design and aerodynamic load calculations; however, its impact on wind power generation is significant as well [47]. TI is incorporated as an extra input variable together with wind speed for GP power curve modelling to analyse its impact on GP accuracy. Figs. 5 and 6 show a 3D scatter plot of the predicted GP power curve (red: measured, blue: estimated) where TI is incorporated with,

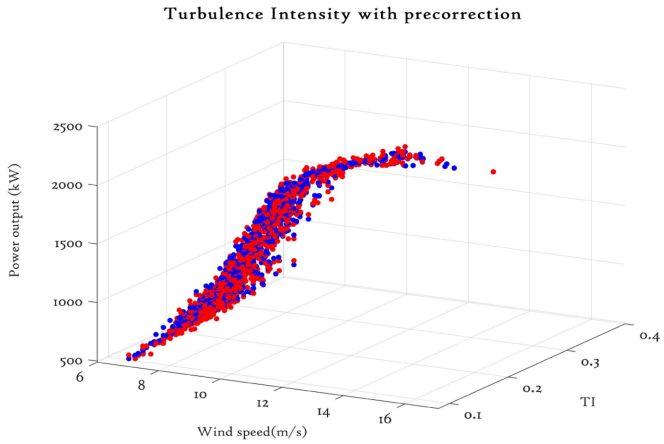

Fig. 6. Pre-corrected air density GP power curve incorporating TI [red: measured, blue: estimated].

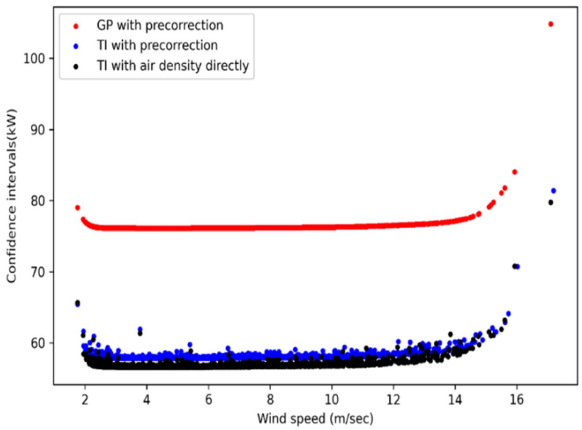

Fig. 7. Uncertainty quantification when turbulence intensity is included.

a) without pre-correction but with air density, and b) with air density pre-correction. It should be noted that pre-correction means air density correction as per (2) and (3). If we compare Figs. 5 and 6, results suggest that without pre-correction but with air density further improves GP model accuracy, which is slightly better than the air density pre-correction approach. That is, by adding air density directly into the model without precorrection together with turbulence intensity, the accuracy of the power curve improves instead of adding only turbulence intensity as an extra input.

This is further confirmed by Fig. 7 which shows the calculated CIs as a function of wind speed used for uncertainty analysis. If TI is incorporated without pre-correction but with air density in the GP model, there is an improvement in model uncertainty as shown in Fig. 7. Nevertheless, whether TI is included with air density or not, the improvement in the GP model will always be higher than the IEC standards prescribed pre-correction approach, as illustrated in Fig. 7.

B. Accounting for the Effect of Wind Shear

As already explained, wind shear is an environmental phenomenon that influences wind power generation. A shear exponent (α) that links wind speeds at two different heights is used to

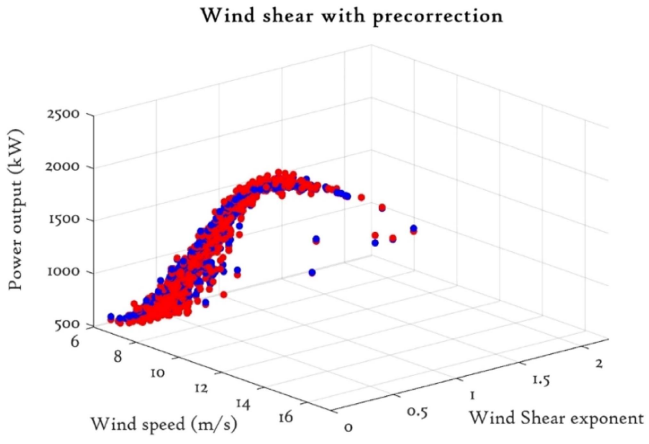

Fig. 8. GP power curve incorporating wind shear with pre-correction[red: measured, blue: estimated].

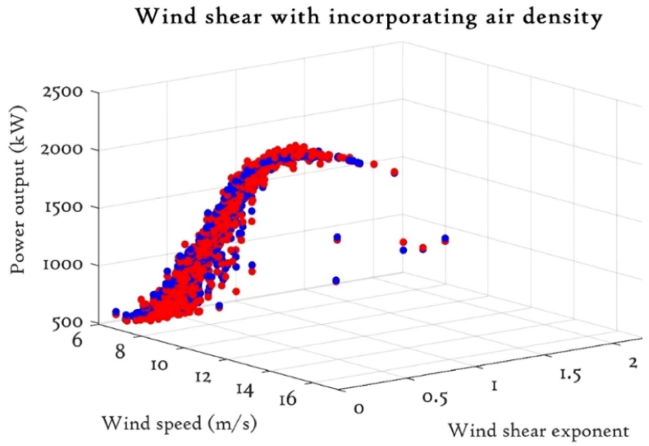

Fig. 9. Pre-corrected air density GP power curve incorporating wind shear [red: measured, blue: estimated].

quantify wind shear and it is calculated once in 10 minutes using the power-law, (4), where v_0 is taken from the SCADA system at $\mathrm{h}_0 = 27$ m while v is at $\mathrm{h} = 66$ m. Thereafter, calculated α values are added as an extra input variable alongside wind speed to estimate the power curve using GPs.

Figs. 8 and 9 present 3D scattered plots of the estimated GP power curve with and without air density pre-correction, respectively. In both cases, it is seen that the inclusion of wind shear improves accuracy. Furthermore, incorporating air density directly into the model improves accuracy and uncertainty. This is demonstrated in the uncertainty analysis, where estimated CIs are plotted against corrected wind speed, as shown in Fig. 10. It is possible to see how incorporating both air density and wind shear results in improved uncertainty across all wind speed ranges in terms of CI. However, improvement is not as significant as turbulence intensity. These results will be quantified in Table IV.

VI. Comparative Studies

This section presents a comparative assessment of incorporating environmental characteristics into GP power curve models,

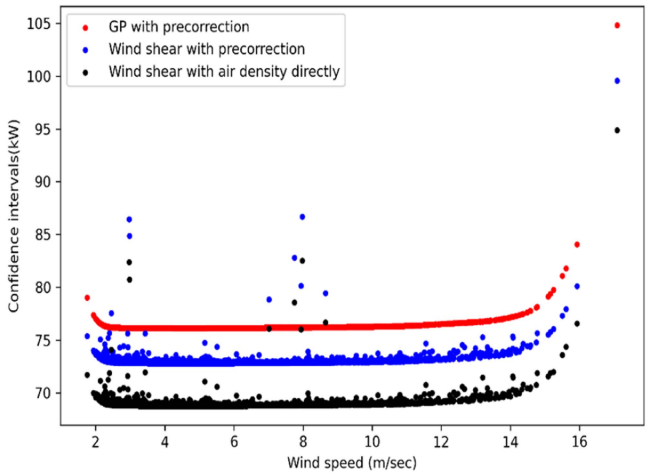

Fig. 10. Uncertainty quantification when wind shear is included.

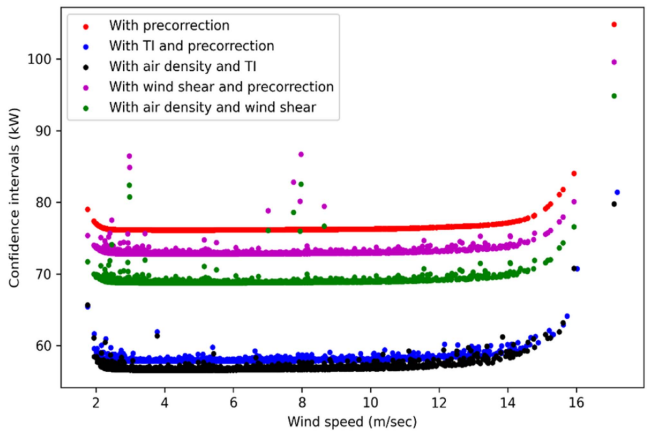

Fig. 11. Comparative performance analysis.

TABLE III
PERFORMANCE METRICS EQUATIONS

Metric	Equation
RMSE	$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i' - y_i)^2}{n}}$
MAE	$\text{MAE} = \frac{\sum_{i=1}^{n}\text{abs}(y_i' - y_i)}{n}$

TABLE IV
PERFORMANCE METRICS RESULTS

GP Models	RMSE (kW)	MAE (kW)	R^2
Without wind shear and turbulence intensity	47.56	36.52	.951
With pre-correction and turbulence intensity	35.36	29.11	.987
With TI and air density directly incorporated without pre-correction	33.23	27.02	.991
With pre-correction and wind shear	41.36	33.91	.977
With wind shear and air density directly incorporated without pre-correction	38.36	30.11	.982
With incorporated air density without pre-correction, TI and wind shear	**30.03**	**23.13**	**.998**

using uncertainty analysis and error metrics (RMSE, MAE and R^2) to assess the performance.

A. Uncertainty Quantification

As previously stated, confidence intervals (CIs) are a useful tool for assessing model uncertainty and precision. Therefore, estimated CIs of GPs incorporating environmental variables into models were plotted as a function of wind speed as shown in Fig. 11.

TI and air density directly included in the GP power curve model suggest the largest improvement in uncertainty across all wind speed regions, unlike the other models. Furthermore, adding air density without pre-correction also improves GP model accuracy but is not as significant as compared to turbulence intensity and wind shear, as shown in Fig. 11. This conclusion is further demonstrated by the calculated numerical values of performance error metrics in an upcoming section.

B. Performance Error Metrics

The root mean squared error (RMSE) and mean absolute error (MAE) are presented in Table III as performance metrics to quantify the performance of the proposed models incorporating environmental conditions, where n is the number of observed samples, y_i' and $y_i (i = 1, 2, \ldots, n)$ denote the estimated values and measure values, respectively. They measure the deviation between the estimated and the measured values. In general, lower values of these metrics reflect higher model accuracy.

The coefficient of determination (R^2) is another useful metric and is defined as the square of the correlation between predicted output and actual value; it reflects how close the data are to the fitted regression (hence always in between 0 to 1 with values closer to 1 indicating better fitting of the model to the data). It is defined as $R^2 = 1 - \frac{SSE}{TSS}$, where SSE is the sum of squared errors and TSS is the total sum of squares.

The calculated values of these performance error metrics for GP models are tabulated in Table IV. The calculated RMSE and MAE values for turbulence intensity incorporated in the GP model record a significant improvement in accuracy. The $R^2 = 00.991$ for turbulence intensity and air density directly incorporated in the GP model is relatively close to 1 and thus suggests a highly robust model. In the case of the wind shear based GP model, calculated values of RMSE and MAE are significantly higher as compared to the turbulence intensity-based GP model. The largest improvement in performance metrics is being recorded (see bolded results of Table IV) when turbulence

intensity, wind shear, and air density are all incorporated directly into the GP model. These performance error metrics values are consistent with Fig. 11.

VII. Conclusion and Suggested Future Work

In this paper, the impact of environmental conditions on data-driven wind turbine power curve models is reported. The results obtained with this dataset show that including environmental variables improves power curve accuracy and reduces model uncertainty without undue additional complexity. A multivariate data-driven Gaussian Process technique was applied to estimate the power curve with environmental variables (air density, wind shear, and turbulence intensity) incorporated as extra inputs alongside wind speed. The results demonstrate clearly that the inclusion of these additional parameters significantly improves model accuracy and reduces uncertainty, as shown in Figs. 4 to 8. A comparative analysis was undertaken to identify the most significant parameters in terms of impact on GP model accuracy. This study concluded that the inclusion of turbulence intensity makes the greatest improvement. Indeed, turbulence intensity is an important parameter that improves GP power curve accuracy and uncertainty in terms not only of RMSE but of MAE (around 15% regarding the second-best case). This accuracy further improves if air density is directly added to the model together with turbulence intensity as shown in Fig. 11. Calculated values of performance error metrics (Table IV) also support these conclusions.

The core findings of this research revealed that by using environmental variables, industrial practitioners can improve modelling performance for condition monitoring-related activities. Moreover, incorporating key environmental variables without additional computation and high processing power demand will help enhance early fault detection algorithms and, thereby, optimize O&M decisions and reduce costs. The inclusion of environmental variables makes the confidence interval narrower, which means it has a better ability to detect incoming faulty data points if they are not within confidence interval limits. This is kept for future study.

Future research will apply the findings of this study to alternative data-driven algorithms in order to evaluate, more generally, the impact of environmental variables on power curve modelling accuracy and uncertainty. Even more, how higher resolution data (e.g., blade pitch angle) may affect data-driven fault detection algorithm accuracy will be also analysed. Finally, other machine learning techniques-based models will be explored.

References

[1] IEA Report entitled, "Renewables 2020–Analysis," 2020. Accessed: Aug. 1, 2021. [Online]. Available: https://www.iea.org/reports/renewables-2020/wind

[2] WWEA Report Entitled, "Worldwide wind capacity reaches 744 gigawatts – An unprecedented 93 gigawatts added in 2020." 2021. Accessed: Jul. 1, 2021. [Online]. Available: https://wwindea.org/worldwide-wind-capacity-reaches-744-gigawatts/

[3] R. Martin, I. Lazakis, S. Barbouchi, and L. Johanning, "Sensitivity analysis of offshore wind farm operation and maintenance cost and availability," *Renewable Energy*, vol. 85, pp. 1226–1236, 2016, doi: 10.1016/j.renene.2015.07.078.

[4] Global wind operations & maintenance market to double by 2025, 2017. Accessed: May 1, 2021. [Online]. Available: http://www.climateaction.org

[5] Y. Zhao, L. Ye, W. Wang, H. Sun, Y. Ju, and Y. Tang, "Data-Driven correction approach to refine power curve of wind farm under wind curtailment," *IEEE Trans. Sustain. Energy*, vol. 9, no. 1, pp. 95–105, Jan. 2018, doi: 10.1109/TSTE.2017.2717021.

[6] J. Nilsson and L. Bertling, "Maintenance management of wind power systems using condition monitoring systems—Life cycle cost analysis for two case studies," *IEEE Trans. Energy Convers.*, vol. 22, no. 1, pp. 223–229, Mar. 2007.

[7] Y. Wang, Q. Hu, D. Srinivasan, and Z. Wang, "Wind power curve modeling and wind power forecasting with inconsistent data," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 16–25, Jan. 2019, doi: 10.1109/TSTE.2018.2820198.

[8] X. Liu, J. Du, and Z.-S. Ye, "A condition monitoring and fault isolation system for wind turbine based on SCADA data," *IEEE Trans. Ind. Inform.*, vol. 18, no. 2, pp. 986–995, Feb. 2022, doi: 10.1109/TII.2021.3075239.

[9] D Astolfi, R Pandit, L Celesti, M Vedovelli, A Lombardi, and L. Terzi, "Data-Driven assessment of wind turbine performance decline with age and interpretation based on comparative test case analysis," *Sensors*, vol. 22, no. 9, 2022, Art. no. 3180, doi: 10.3390/s22093180.

[10] J. Tautz-Weinert and S. J Watson, "Using SCADA data for wind turbine condition monitoring—A review," *IET Renewable Power Gener.*, vol. 11, pp. 382–394, 2017.

[11] Z. Wang, L. Wang, and C. Huang, "A fast abnormal data cleaning algorithm for performance evaluation of wind turbine," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5006512, doi: 10.1109/TIM.2020.3044719.

[12] *Wind Turbines—Part 12–1: Power Performance Measurements of Electricity Producing Wind Turbines*, IEC 61400-12-1:2017, IEC, Geneva, Switzerland, 2017.

[13] Y. Hu, Y. Qiao, J. Liu, and H. Zhu, "Adaptive confidence boundary modeling of wind turbine power curve using SCADA data and its application," *IEEE Trans. Sustain. Energy*, vol. 10, no. 3, pp. 1330–1341, Jul. 2019, doi: 10.1109/TSTE.2018.2866543.

[14] R. K. Pandit, D. Infield, and J. Carroll, "Incorporating air density into a gaussian process wind turbine power curve model for improving fitting accuracy," *Wind Energy*, vol. 22, pp. 302–315, 2019, doi: 10.1002/we.2285.

[15] A. Feijóo and D. Villanueva, "Four-parameter models for wind farm power curves and power probability density functions," *IEEE Trans. Sustain. Energy*, vol. 8, no. 4, pp. 1783–1784, Oct. 2017, doi: 10.1109/TSTE.2017.2698199.

[16] D. Astolfi, R. Pandit, L. Celesti, A. Lombardi, and L. Terzi, "SCADA data analysis for long-term wind turbine performance assessment: A case study," *Sustain. Energy Technol. Assessments*, vol. 52, 2022, Art. no. 102357, doi: 10.1016/j.seta.2022.102357.

[17] D. Villanueva and A. Feijóo, "Normal-based model for true power curves of wind turbines," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1005–1011, Jul. 2016, doi: 10.1109/TSTE.2016.2515264.

[18] A. Kusiak and A. Verma, "Monitoring wind farms with performance curves," *IEEE Trans. Sustain. Energy*, vol. 4, no. 1, pp. 192–199, Jan. 2013.

[19] Y. Wang et al., "Sparse heteroscedastic multiple spline regression models for wind turbine power curve modeling," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 191–201, Jan. 2021, doi: 10.1109/TSTE.2020.2988683.

[20] X. Shen, X. Fu, and C. Zhou, "A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 46–54, Jan. 2019, doi: 10.1109/TSTE.2018.2822682.

[21] Y. Wang, D. G. Infield, B. Stephen, and S. J. Galloway, "Copula-based model for wind turbine power curve outlier rejection," *Wind Energy*, vol. 17, no. 11, pp. 1677–1688, Nov. 2014, doi: 10.1002/we.1661.

[22] G. Ciulla, A. D'Amico, V. Di Dio, and V. Lo Brano, "Modelling and analysis of real-world wind turbine power curves: Assessing deviations from nominal curve by neural networks," *Renewable Energy*, vol. 140, pp. 477–492, 2019, doi: 10.1016/j.renene.2019.03.075.

[23] M. Xu, P. Pinson, Z. Lu, Y. Qiao, and Y. Min, "Adaptive robust polynomial regression for power curve modeling with application to wind power forecasting," *Wind Energy*, vol. 19, pp. 2321–2336, Dec. 2016.

[24] S. Sarkar and V. Ajjarapu, "MW resource assessment model for a hybrid energy conversion system with wind and solar resources," *IEEE Trans. Sustain. Energy*, vol. 2, no. 4, pp. 383–391, Oct. 2011.

[25] S. H. Jangamshetti and V. G. Rau, "Normalized power curves as a tool for identification of optimum wind turbine generator parameters," *IEEE Trans. Energy Convers.*, vol. 16, no. 3, pp. 283–288, Sep. 2001.

[26] B. Hu, Y. Li, H. Yang, and H. Wang, "Wind speed model based on kernel density estimation and its application in reliability assessment of generating systems," *J. Modern Power Syst. Clean Energy*, vol. 5, pp. 220–227, 2017.
[27] D. Zahariea and D. Husaru, "Atmospheric air density analysis with Meteo-40S wind monitoring system," in *Proc. 21st Innov. Manuf. Eng. Energy Int. Conf.*, 2017, Art. no. 07020, doi: 10.1051/matecconf/201711207020.
[28] R. Wagner, I. Antoniou, S. M. Pedersen, M. Courtney, and H. E. Jørgensen, "The influence of the wind speed profle on windturbine performance measurements," *Wind Energy*, vol. 12, pp. 348–362, 2009.
[29] R. Wagner, M. Courtney, J. Gottschall, and P. Lindelöw-Marsden, "Accounting for the speed shear in wind turbine power performance measurement," *Wind Energy*, vol. 14, pp. 993–1004, 2011.
[30] A. Clifton and R Wagner, "Accounting for the effect of turbulence on wind turbine power curves," *J. Phys. Conf. Ser.*, vol. 524, 2014, Art. no. 012109, doi: 10.1088/1742-6596/524/1/012109.
[31] L. Bardal, L. Sætran, and E. Wangsness, "Performance test of a 3 MW wind turbine – effects of shear and turbulence," *Energy Procedia*, vol. 80, pp. 83–91, 2015, doi: 10.1016/j.egypro.2015.11.410.
[32] IEC, 2013, *CDV IEC 61400-12-1 Power Performance Measurements of Electricity Producing Wind Turbines*.
[33] G. Ren, J. Liu, J. Wan, F. Li, Y. Guo, and D. Yu, "The analysis of turbulence intensity based on wind speed data in onshore wind farms," *Renewable Energy*, vol. 123, pp. 756–766, 2018.
[34] M.S. Siddiqui et al., "Effect of turbulence intensity on the performance of an offshore vertical axis wind turbine," *Energy Procedia*, vol. 80, pp. 312–320, 2015.
[35] I. Malael, V. Dragan, and G.B. Gherman, "Turbulence intensity effects on the vertical axis wind turbine starting efficiency," *Ann. DAAAM Proc.*, vol. 26, no. 1, pp. 0974–0979, 2015.
[36] N. Carpman, "Turbulence intensity in complex environments and its influence on small wind turbines," M.Sc. thesis, Dept. Earth Sci., Uppsala Univ., Uppsala, Sweden, 2011.
[37] G. Ren, J. Liu, J. Wan, F. Li, Y. Guo, and D. Yu, "The analysis of turbulence intensity based on wind speed data in onshore wind farms," *Renewable Energy*, vol. 123, pp. 756–766, 2018.
[38] F. Castellani, D. Astolfi, P. Sdringola, S. Proietti, and L Terzi, "Analyzing wind turbine directional behavior: SCADA data mining techniques for efficiency and power assessment," *Appl. Energy*, vol. 185, pp. 1076–1086, 2017, doi: 10.1016/j.apenergy.2015.12.049.
[39] R. K. Pandit, D. Infield, and A. Kolios, "Gaussian process power curve models incorporating wind turbine operational variables," *Energy Rep.*, vol. 6, pp. 1658–1669, 2020.
[40] R. Pandit, D. Infield, and T. Dodwell, "Operational variables for improving industrial wind turbine yaw misalignment early fault detection capabilities using data-driven techniques," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 2508108, doi: 10.1109/TIM.2021.3073698.
[41] L. M. Bardal and L. R. Sætran, "Influence of turbulence intensity on wind turbine power curves," *Energy Procedia*, vol. 137, pp. 553–558, 2017.
[42] S. Díaz, J. A. Carta, and A. Castañeda, "Influence of the variation of meteorological and operational parameters on estimation of the power output of a wind farm with active power control," *Renewable Energy*, vol. 159, pp. 812–826, 2020.
[43] N. Mittelmeier and M. Kühn, "Determination of optimal wind turbine alignment into the wind and detection of alignment changes with SCADA data," *Wind Energy Sci.*, vol. 3, pp. 395–408, 2018, doi: 10.5194/wes-3-395-2018.
[44] P. Bangalore, S. Letzgus, D. Karlsson, and M. Patriksson, "An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox," *Wind Energy*, vol. 20, no. 8, pp. 1421–1438, 2017.
[45] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *J. Mach. Learn. Res.*, vol. 11, pp. 3011–3015, 2010.
[46] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[47] J. Gonzalez, "Linear-time inverse covariance matrix estimation in Gaussian processes," 2017, *arXiv:1610.08035v4*.
[48] A. Albers, T. Jakobi, R. Rohden, and J. Stoltenjohannes, "Influence of meteorological variables on measured wind turbine power curves," in *Proc. Eur. Wind Energy Conf. Exhib.*, 2007, pp. 525–546.

Ravi Pandit received the bachelor's degree from Jadavpur University, Kolkata, India, in 2009, the master's degree from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2011, and the Ph.D. degree from the University of Strathclyde, Glasgow, U.K., in 2019. He is currently a Lecturer of instrumentation and AI with the Centre for Life-Cycle Engineering and Management, School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, U.K. From 2011 to 2016, he was an Assistant Professor with Jadavpur University and the Vellore Institute of Technology, Vellore, India. His research interests include data-driven applications on offshore wind including condition monitoring, predictive maintenance, forecasting and prediction, and SCADA data statistical analysis. He has more than five years of direct research experience in his research areas.

David Infield received the B.A. degree in mathematics and physics from the University of Lancaster, Lancaster, U.K., and the Ph.D. degree in applied mathematics from the University of Kent, Canterbury, U.K.

From 1982 to 1993, he was with Rutherford Appleton Laboratory, Oxfordshire, U.K., researching into wind electricity systems. From 1993 to 2007, he was with Loughborough University, Leicestershire, U.K., where he established CREST, the Centre for Renewable Energy Systems Technology. He is currently a Research Professor of renewable energy technologies with the Institute for Energy, University of Strathclyde.

Matilde Santos was born in Madrid, Spain. She received the B.Sc. and M.Sc. degrees in physics (computer engineering) and the Ph.D. degree in physics from the University Complutense of Madrid, Madrid, Spain. She is currently a Full Professor of system engineering and automatic control. She is a Member of the European Academy of Sciences and Arts. She has authored or coauthored many papers in international scientific journals and several book chapters. She has supervised more than ten Ph.Ds. Her main research interests include intelligent control (fuzzy and neuro-fuzzy), modeling and simulation, autonomous (guided) vehicles, and wind energy. She has worked on several national, European and international research projects, leading some of them. She is currently a member of the editorial board of prestigious journals and the Editor-In-Chief Assistant of one of them.