Wind power curve modeling: A probabilistic Beta regression approach

Marco Capelletti [a,*], Davide M. Raimondo [a], Giuseppe De Nicolao [a,b]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via A. Ferrata, 5, Pavia, 27100, Italy
[b] Division of Infectious Diseases I, Fondazione IRCCS Policlinico San Matteo, Viale C. Golgi, 19, Pavia, 27100, Italy

ARTICLE INFO

Keywords:
Wind turbines
Wind power curves
Statistical models
Generalized linear models

ABSTRACT

Wind turbine power curves play a key role in various aspects during the life of a wind farm. Typical uses range from wind power forecasting to wind turbine condition monitoring. This paper addresses the identification of probabilistic models of wind power curves from observed data. The main challenge is the need to handle a statistical distribution of wind energy whose shape not only may be highly skewed, but can also change with wind speed. To address these issues, we resort to the framework of Generalized Linear Models (GLMs), proposing a Beta regression approach, with constant or variable dispersion and an appropriate preconditioning step. The proposed methodology was tested on three real SCADA measurements retrieved from public datasets, including a comparison with Quantile Regression Forests (QRFs), also in terms of robustness to outliers. The results suggest that Beta regression can be a valuable resource in the development of probabilistic models for wind energy, as it provides a high degree of flexibility while preserving an interpretable structure.

1. Introduction

In the context of wind power generation, it is well known the importance of the power curve linking the generated power to the wind speed and other features such as wind direction, air density, temperature, humidity, and so on. The power curve plays a key role for several purposes, such as power forecasting, turbine condition monitoring, wind energy potential estimation and turbine selection [1–6].

Although theoretical power curves provided by manufacturers might allow for factors such as temperature and humidity, it is a common wisdom that the performance of the same piece of equipment located in different sites may vary due to differences in weather and geography. Hence the need of collecting and analyzing actual wind and power data in order to build accurate wind power curves.

A first type of power curves are those linking the generated power to measured inputs, mainly the wind speed. As an alternative, it is possible to estimate power curves that link the generated power to the forecast wind speed (and possibly other regressors). Since the uncertainty affecting the forecasts is much larger than that affecting the measured weather variables, there are some notable differences in the statistical issues arising in the estimation of these two types of power curves, see e.g. [7]. In the present paper, only the first type of problem will be investigated.

Despite the wide selection of power curve models available in the literature, it is commonly recognized that there is still room for improvement. An in-depth review of the available models and their pros and cons was provided by Wang et al. [5]. In summary, models can be divided between deterministic and probabilistic. In the first category, we find curve fitting, piecewise linear, polynomial, sigmoidal and logistic models, see [3,8–10] for some examples. A further sub-category is artificial intelligence models, including neural networks, K-Nearest Neighbors (KNN), and various regression methods, see e.g. [11–14].

As noted by Wang, deterministic models suffer from some limitations due to the specific features of wind power data. In particular, as wind speed changes, the statistical distribution of power exhibits two fundamental characteristics. First, the variance is typically heteroscedastic, that is it varies significantly. In particular, the variance may be lowest when the speed is close to the cut-in and the power is therefore close to its minimum and when the speed is in a saturated condition near the rated power. A second characteristic is the asymmetry of the power distribution. For low speeds, the tail of the power distribution is skewed to the right. Conversely, near the rated power, the distribution becomes skewed to the left.

In light of these characteristics, deterministic models, providing just a point prediction of power, appear to be insufficient. The true challenge lies in developing probabilistic models that, using wind speed and other meteorological data, go beyond single-point power predictions and instead provide a comprehensive statistical distribution, accounting for changes in variance and skewness. The superiority of these models is evident when one considers the possible uses in fault detection and forecasting, which obviously benefit from the ability to quantify and propagate uncertainty.

* Corresponding author.
E-mail address: marco.capelletti02@universitadipavia.it (M. Capelletti).

An advantage of probabilistic methods is the possibility of obtaining point forecasts optimized according to the score index. For example, one will use the mean of the power distribution, if interested in minimizing the quadratic error, and instead the median, if interested in the absolute error.

Interest in probabilistic models has grown in recent years, see e.g. [15–18]. One category of models takes into account heteroscedasticity through Gaussian distributions, where the variance depends on wind speed, while the mean can be described by a deterministic model or a Gaussian Process or a spline [5]. A limitation of these approaches is that the use of Gaussian distributions is unable to accurately represent the asymmetry of the distribution, whose skewness, as said, changes sign with varying wind speed. A more flexible approach employs a mixture of Gaussians which, thanks to its nonparametric nature, is capable of capturing both the heteroscedasticity and the asymmetry of the power distribution conditioned on wind speed [5,15]. The training hinges on a hierarchical Bayes model that requires suitable prior distributions for parameters and hyperparameters of a Gaussian mixture with a potentially large number of components.

As Wang and colleagues have recognized, there is opportunity for additional exploration in the field of probabilistic models. The primary objective of the present paper is to develop and validate a novel probabilistic model that addresses both heteroscedasticity and asymmetry in the wind power distribution. In particular, it is assumed that this distribution function can be effectively approximated by a Beta random variable. Then, the critical challenge lies in determining how the parameters of the Beta distribution are influenced by the independent variables. To efficiently address this problem, we employ Generalized Linear Models (GLMs), under the form of Beta Regression [19,20]. The paper introduces the GLM framework and discusses the steps of model building. Various implementation aspects are discussed, including the use of a variable dispersion parameter and preconditioning techniques that can exploit the manufacturer's power curve or deterministic models.

The paper is organized as follows. In Section 2, three datasets are introduced and the issue of outlier detection is addressed. The challenges posed to probabilistic models by the heteroscedastic and asymmetric distribution of wind power are illustrated in Section 3. The novel approach based on Beta Regression, a particular type of GLM, is introduced in Section 4. In Section 5, the application of the new method to the three datasets is presented, including a comparison with Quantile Regression Forests (QRFs), also in terms of robustness to outliers. A discussion of the results ends the paper. In the Appendix, details are given on three technical issues: outlier removal by the Ratio-Skewed boxplot algorithm, visual validation of conditional distribution models, and the tuning of QRFs.

2. Data description

2.1. Dataset A

This dataset pertains to a wind turbine situated in Turkey, equipped with a Supervisory Control and Data Acquisition (SCADA) system. The dataset, collected throughout the year 2018, was retrieved from the Kaggle repository [21].

Each entry of the 52,560 records corresponds to a 10-minute interval. The following variables are of interest:

1. Date/Time: the timestamp of each data point, recorded at 10-minute intervals.
2. LV ActivePower (kW): the real-time power output of the turbine.
3. Wind Speed ($\mathrm{m\,s^{-1}}$): Measured by an anemometer at the hub height of the turbine.
4. Theoretical_Power_Curve (kW): the theoretical power values that the wind turbine should generate based on the recorded wind speed. These values, determined by the turbine manufacturer, serve as a reference for evaluating the expected performance of the turbine. According to the theoretical power curve the cut-in speed and the rated one are $w_{in} = 3\,\mathrm{m\,s^{-1}}$ and $w_{rt} = 13\,\mathrm{m\,s^{-1}}$, respectively.
5. Wind Direction as degrees (°): the wind direction at the hub height of the turbine, automatically adjusted to optimize power generation based on wind direction.

The scatter plot of active power vs wind speed is plotted in the upper left panel of Fig. 1.

2.2. Dataset B

The second dataset consists of weather, turbine, and rotor features recorded from January 2018 to March 2020, retrieved from the Kaggle repository [22]. The data were collected at a 10-minute interval, resulting in a total of 118,224 entries. The following variables are of interest:

1. ActivePower (kW): the active power generated by the turbine.
2. WindSpeed ($\mathrm{m\,s^{-1}}$): the wind speed measured at the turbine location.
3. WindDirection as degrees (°): the wind direction at the turbine location.

Differently from dataset A, the theoretical power curve is not available. From the inspection of the data, presumed values of the cut-in and rated wind speed were assumed equal to $w_{in} = 2.5\,\mathrm{m\,s^{-1}}$ and $w_{rt} = 9.5\,\mathrm{m\,s^{-1}}$.

2.3. Dataset C

This dataset includes SCADA measurements of the first of 14 Senvion MM82 wind turbines forming the Penmanshiel wind farm in the United Kingdom (coordinates 55.904976, −2.291849), whose installed capacity is 28.7 MW. The dataset incorporates crucial variables, including 'Potential power default PC (kW)', which characterizes the theoretical turbine power curve. Additionally, measurements such as 'Wind direction (°)' and 'Wind speed (m/s)' that are recorded at the nacelle. 'Lost Production to Downtime and Curtailment Total (kWh)' serves as a variable indicating instances of downtime or curtailment, with non-zero values signifying recorded events. Finally, 'Power (kW)' is a key variable representing the actual power output of the turbine.

The temporal span of the dataset encompasses the period from January 1, 2016, to July 1, 2021, with data recorded at 10-minute intervals and organized in different .csv files by year. Further data, including site substation/PMU meter and site fiscal/grid meter information, is accessible on Zenodo [23]. For the subsequent analysis, we kept all the wind turbine data where 'Lost Production to Downtime and Curtailment Total (kWh)' is equal to zero.

3. The challenge of heteroscedasticity and asymmetry

In order to illustrate the features of the distribution of wind power conditional on different wind speeds, let us inspect datasets A and B, plotting the histograms of wind power observations associated with wind speeds belonging to specific bins, ranging from the cut-in wind speed to the rated one, see Figs. 2 and 3.

In particular, for dataset A, wind speed was binned in 9 intervals of width $1.11\,\mathrm{m\,s^{-1}}$ starting from $3\,\mathrm{m\,s^{-1}}$. Accordingly, the dataset was split in 9 subsets, yielding the 9 histograms of wind power values displayed in Fig. 2. In a similar way, for dataset B, wind speed was binned in 9 intervals of width $0.77\,\mathrm{m\,s^{-1}}$ starting from $2.5\,\mathrm{m\,s^{-1}}$ and the 9 histograms were plotted in Fig. 3.

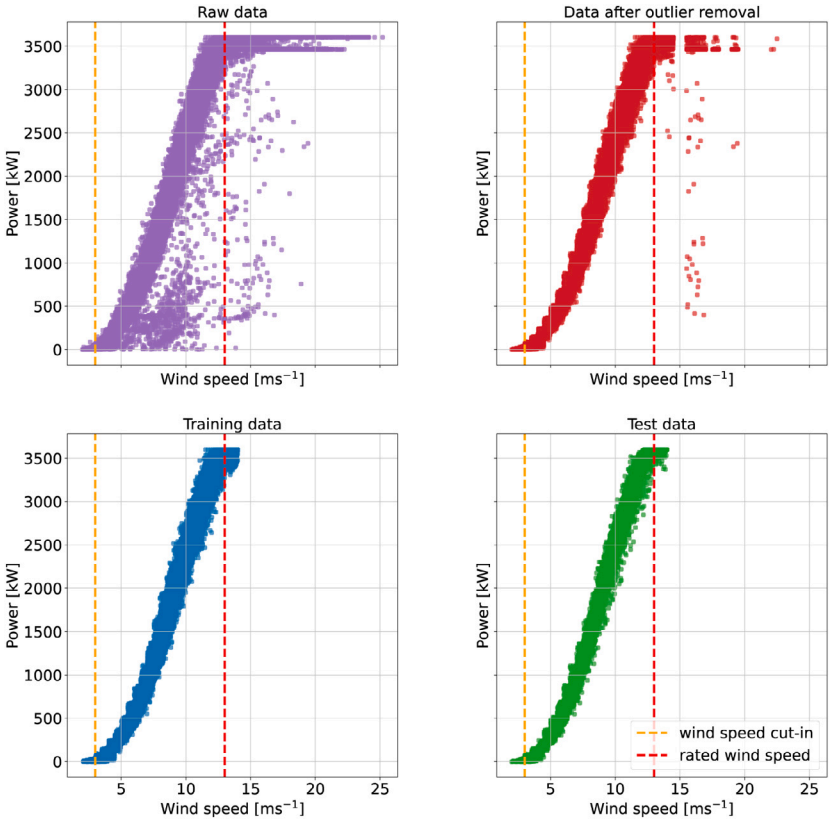

Fig. 1. Dataset A. Raw data: observed wind speed and wind power pairs from a wind turbine situated in Turkey (upper left); data after outlier removal via ratio-skewed boxplot outlier detection (upper right); training dataset (lower left); test dataset (lower right).

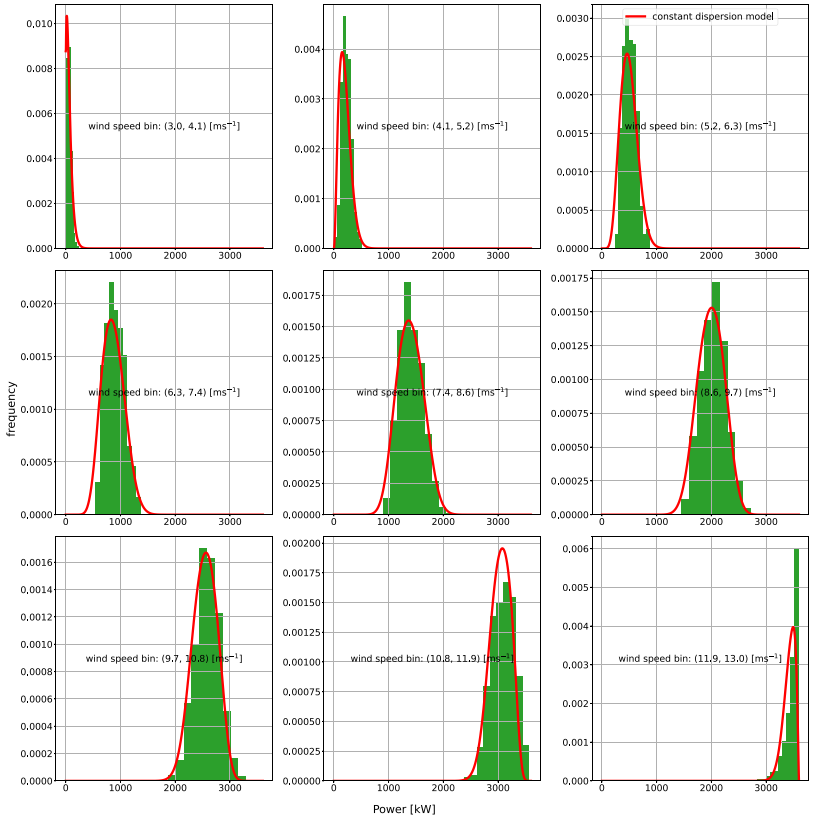

Fig. 2. Dataset A (test data): Wind power histograms (green) corresponding to 9 bins of wind speed ranging from the cut-in speed (top left panel) to the rated one (bottom right). The red curve is the distribution predicted by fitting the training data with a Beta regression model with constant dispersion, preconditioned using a spline function (model M5).

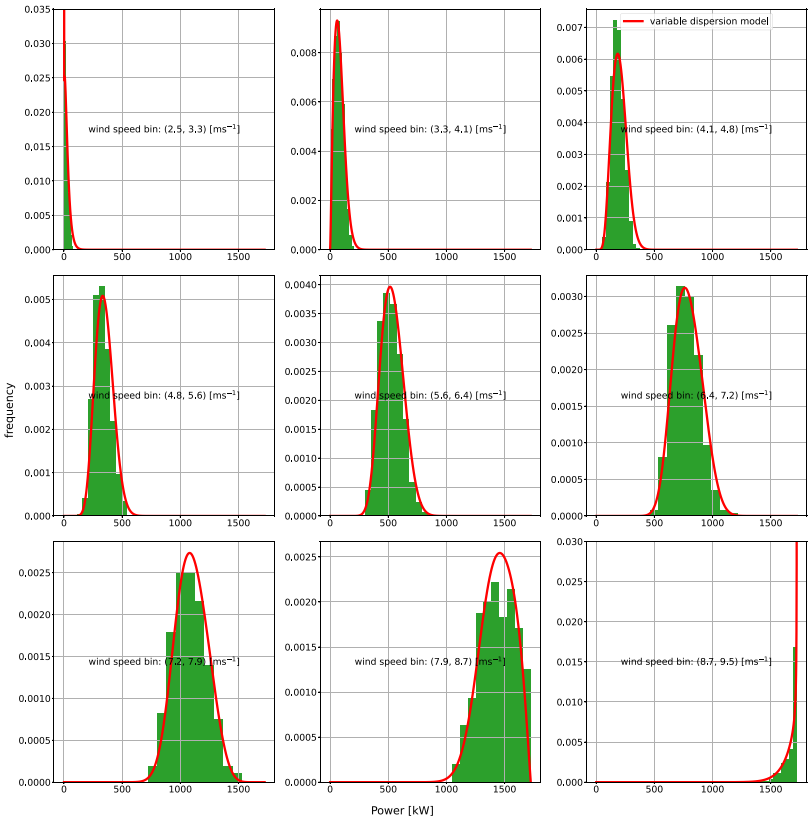

Fig. 3. Dataset B (test data): Wind power histograms (green) corresponding to 9 bins of wind speed ranging from the cut-in speed to the rated one. The red curve is the distribution predicted by fitting the training data with a Beta regression model with variable dispersion, preconditioned using a spline function (Model M6).

The visual inspection of Figs. 2 and 3 highlights the peculiar nature of the distribution of wind power. First of all, it is highly heteroscedastic in that the variance changes substantially depending on the selected wind speed bin: smaller for the extreme bins and larger for the central ones. In addition, the distribution exhibits a strong asymmetry with skewness that ranges from positive values for low speeds to negative values for wind speeds close to the rated one.

It is this heterogeneity that motivated the development of flexible error models, such as normal distributions with different variances or Gaussian mixtures [5]. Note that, compared to [5,24,25], Figs. 2 and 3 do not display forecast errors relative to some model, but display just the conditional histograms without any modeling assumption. The idea is that the statistical properties of the conditional distribution should guide the choice of the most appropriate model. As a matter of fact, the Beta-like shapes of the histograms suggest that a parsimonious model could be obtained by resorting to Beta regression, a particular type of GLM, reviewed in the next section.

4. Beta regression model with expit-spline-link

Consider the problem of modeling the distribution of a scalar target real variable Y, given a real-valued p-dimensional vector $X \in \mathbb{R}^p$ of predictors. Recall that a GLM is characterized by three elements:

1. the distribution $f(Y|X)$ of the target variable given the predictors, to be chosen within the exponential family of distributions;
2. a linear predictor $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$; hereafter, $\beta = \begin{bmatrix}\beta_0 & \dots & \beta_p\end{bmatrix}^T$ will denote the vector of regression parameters;
3. a link function $g(\cdot)$ such that $\mu = \mathbb{E}[Y|X] = g^{-1}(\eta)$.

A crucial ingredient of a GLM model is the choice of the distribution, whose features should match the physical properties of the phenomenon under investigation. In our case, we are dealing with the wind power which, by its nature, is bounded in the interval ranging from zero to the rated power of the turbine.

4.1. Constant dispersion beta regression model

If Y takes values in a known interval, by a proper scaling it can be assumed that $Y \in (0, 1)$. Then, a plausible distribution is the Beta one, which is completely specified by parameters $0 < \mu < 1$ and $\phi > 0$:

$$f(y;\mu,\phi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{(a-1)}(1-y)^{(b-1)}, \qquad 0 < y < 1$$
$$a = \mu\phi, \qquad b = \phi(1-\mu)$$

where $\Gamma(\cdot)$ denotes the gamma function. The mean and the variance are:

$$\mathbb{E}(Y) = \mu$$
$$\mathrm{Var}(Y) = \frac{\mu(1-\mu)}{1+\phi}$$

Accordingly, μ is known as the mean parameter and ϕ the precision parameter since, for fixed μ, the larger ϕ is, the smaller the variance of Y. Conversely, ϕ^{-1} plays the role of a dispersion parameter.

In our modeling problem, the mean and variance of the target variable will change as a function of the predictor vector X. A link function $g(\cdot)$ is chosen such that $\mu = g^{-1}(\eta)$ guarantees $0 < \mu < 1, \forall \eta \in \mathbb{R}$. Then, the GLM is completely specified by ϕ and

$$\eta(\beta, X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j. \tag{1}$$

Therefore, for a given ϕ, η will change as a function of X, so that through $g^{-1}(\cdot)$ also μ will change, yielding Beta distributions whose shapes vary with X.

Assume now that a training set $D = \{y_i, x_i\}$, $i = 1, \dots, n$, is available, where $x_i = \begin{bmatrix}x_{i1} & x_{i2} & \dots & x_{ip}\end{bmatrix}^T$. Then, the model parameters β and ϕ

can be obtained by numerically solving a maximum likelihood problem. More precisely, for a given link function $g(\cdot)$, the likelihood of the Beta model is

$$\mathcal{L}(\beta, \phi|D) = \prod_{i=1}^{n} f(y_i; \mu(\beta, x_i), \phi)$$

where $\mu(\beta, x_i) = g^{-1}(\eta_i), \eta_i = \eta(\beta, x_i)$.

The *Constant-dispersion Beta Regression* problem can therefore be stated as follows.

$$(\beta^{\mathrm{ML}}, \phi^{\mathrm{ML}}) = \arg\max_{\beta,\phi} \mathcal{L}(\beta, \phi|D) \tag{2}$$

subject to $\phi > 0$.

Note that the GLM provides probabilistic predictions of the target variable: indeed, for a test predictor X, the distribution of $Y|X$ will be $f(y; \mu(\beta^{\mathrm{ML}}, X), \phi^{\mathrm{ML}})$.

Depending on the preferred performance metrics, different choices of the point estimate $\hat{y}$ might be used, e.g. the conditional expectation $\mathbb{E}[Y|X]$ or the conditional median. Moreover, from the percentiles of the conditional distribution, confidence bands can be easily obtained.

4.2. Variable dispersion beta regression model

The constant dispersion Beta regression model assumes that the precision parameter ϕ does not change with X. In some cases this might be too restrictive. This motivates the introduction of a parametric model $\phi = \phi(\theta, X)$. Since ϕ is nonnegative, it is convenient to assume a regression model for $\zeta = \log(\phi)$, e.g. $\phi = \exp(\zeta)$ with $\zeta = \theta_0 + \sum_{i=1}^{m} \theta_i X_i$. Accordingly, the likelihood

$$\mathcal{L}^*(\beta, \theta|D) = \prod_{i=1}^{n} f(y_i; \mu(\beta, x_i), \phi(\theta, x_i))$$

is a function of the parameter vectors β and $\theta = \begin{bmatrix}\theta_0 & \dots & \theta_m\end{bmatrix}^T$.

The variable dispersion Beta Regression problem, is solved again via likelihood maximization, yielding the vectors $\beta^{ML} \in \mathbb{R}^{p+1}, \theta^{ML} \in \mathbb{R}^{m+1}$, from which probabilistic predictions can again be obtained.

Variable Dispersion Beta Regression Problem:

$$(\beta^{\mathrm{ML}}, \theta^{\mathrm{ML}}) = \arg\max_{\beta,\theta} \mathcal{L}^*(\beta, \theta|D) \tag{3}$$

4.3. Expit link function

As already said, the choice of the link function $g(\cdot)$, should comply with the requirement $g^{-1}(\eta) \in (0, 1), \forall \eta \in \mathbb{R}$. A common choice is the *expit* function

$$\mu = g^{-1}(\eta) = \frac{1}{1 + \exp(-\eta)},$$

equivalent to assuming that $g(\cdot)$ is the *logit* function

$$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

For the sake of simplicity, consider the case of a scalar X, e.g. the wind speed. Then, if $\eta = \beta_0 + \beta_1 X$, the expit model amounts to predicting $\mathbb{E}[Y|X]$ by the sigmoid function

$$\hat{\mu}(X; \beta) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X)}.$$

This represents a very basic model for a power curve, where the parameters β_0 and β_1 are employed to capture the shape of the curve, e.g. the steepness and the inflection point of the curve.

4.4. Preconditioning

The link function provided by the expit model may lack the flexibility needed to follow the power curves of real wind turbines. This has motivated the development of numerous alternative models with as many as ten parameters, see, for example, the review paper [5].

Although the link function could be chosen from these models, herein we propose a flexible solution through the use of a preconditioner $s(X)$ that, differently from $\eta(X)$, is a nonlinear function of X. More precisely, given a preconditioner $s(\cdot)$, the associated link function is defined as

$$\mu = g^{-1}(\eta) = \frac{1}{1 + \exp(-\eta - s)}. \tag{4}$$

The idea is that η will still be a linear regression with tunable parameters β_j, but rather than being used to explain all the variability, they are used for the fine tuning of a nominal model based on $s(\cdot)$. In particular, if a theoretical power curve $\bar{\mu}(X)$ is available, a natural preconditioner could be

$$s(X) = \log\left(\frac{\bar{\mu}(X)}{1 - \bar{\mu}(X)}\right). \tag{5}$$

In fact, from (4) and (5) we have that $\mu = \bar{\mu}$ if and only if $\eta = 0$. In other words, η is used to tune the nominal curve to the actual behavior of the turbine .

If the preconditioner has to be estimated, one can fit a suitable function $s(\cdot)$ to the turbine data sheet or experimental data. In the next subsection, natural spline functions are introduced as a way to obtain a flexible parametrization of $s(\cdot)$.

4.5. Natural spline expit model

Given a real scalar variable X, a cubic spline is a piecewise cubic polynomial function of X that is continuous up to the second derivative in the K knots. It can be parametrized as a linear combination of $K + 4$ basis functions. A cubic spline with the added constraint of extrapolating linearly outside the first and the last knot is called a natural spline and can be parametrized as a linear combinations of K basis functions:

$$s(X; \alpha) = \sum_{m=1}^{K} \alpha_m N_m(X)$$

where α_m are the K free parameters and the basis functions are:

$$\begin{aligned} N_1(X) &= 1 \\ N_2(X) &= X \\ N_{k+2}(X) &= d_k(X) - d_{K-1}(X) \\ d_k(X) &= \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k} \end{aligned}$$

with $\xi_k, k = 1, \dots, K$, the locations of the knots. The flexibility of the natural spline is governed by the number of knots, whose location can be optimized. In practice, it may be convenient to use equally spaced knots so that their number remains the only tunable hyperparameter.

The natural spline expit model of a power curve is just obtained by plugging a natural spline function of the wind speed X in the expit model, namely:

$$\hat{\mu}(X; \alpha) = \frac{1}{1 + \exp(-s(X; \alpha))}.$$

4.6. Two-step beta regression algorithm

We are now in a position to describe the two-step Beta Regression algorithm. In the first step, a flexible preconditioner is obtained via nonlinear least squares, while in the second step the preconditioner is plugged in a Beta regression GLM to enhance its descriptive capabilities. Consider first the case of a scalar X, typically the observed wind speed.

Step 1. For a given number and location of the knots, estimate the parameter vector $\alpha \in \mathbb{R}^K$ of the preconditioner $s(\cdot;\alpha)$ by least-squares regression on D of the natural spline expit model

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{N} (y_i - \hat{\mu}(x_i;\alpha))^2$$

Step 2. Solve either the constant dispersion Beta regression problem (2) or the variable dispersion Beta regression problem (3), using $\hat{s} = s(X;\hat{\alpha})$ as preconditioner in the expit link function

$$\mu = g^{-1}(\eta) = \frac{1}{1+\exp(-\eta-\hat{s})}.$$

The scheme can be easily extended to the case of a multivariate feature vector X, including for instance wind direction, pressure, humidity, etc. In such a case, Step 1 may not change if the preconditioner depends just on the wind speed. In Step 2, on the other hand, the complete vector X will be used, thus estimating a $(p+1)$-dimensional vector β instead of just β_0 and β_1.

In summary, the two-step procedure enjoys the advantages of flexible parametric models that try to follow the nonlinear shape of the power curve, but, at the same time, it preserves the accurate probabilistic description provided by the Beta GLM. In fact, the nonlinear least squares regression of the natural spline preconditioner guarantees a flexibility comparable or superior to the parametric models available in the literature. The GLM step, on the other hand, performs a fine-tuning that takes into account the inherently asymmetrical nature of wind power distribution, bounded between zero and the nominal wind power.

4.7. Quantile regression forests

Quantile Regression Forests (QRFs) are an extension of the random forests algorithm that focuses on estimating conditional quantiles of a response variable. The method involves growing an ensemble of trees and estimating the conditional distribution of the response variable by considering the weighted distribution of observed response variables.

The estimation of conditional quantiles is a key feature of QRFs, providing a non-parametric and accurate way of inferring the relationship between predictor variables and the full conditional distribution of the response variable.

Formulas and detailed mathematical expressions for the algorithm can be found in the original paper [26] where QRF is shown to outperform four other quantile regression methods: linear quantile regression with interactions and without interactions, and three variants of quantile regression trees. Here, QRF will be used as a benchmark against which Beta regression will be compared, especially with regard to robustness to outliers.

5. Results

5.1. Data cleaning

The data cleaning phase consisted of the following steps:

1. Discard observations where one or more of the variables wind speed, direction and power are missing.
2. Observations with negative or zero power, are discarded, while all power values exceeding the theoretical maximum power are set to the theoretical maximum power. The theoretical maximum power for dataset A corresponds to 3600 kW [21] while for dataset C corresponds to 2050 kW. For dataset B, since information about the theoretical curve was not available, it was empirically derived from the data and set to 1725 kW.
3. Discard observations corresponding to wind speeds less than $w_{in} - 1$ for dataset A and C and w_{in} for dataset B, where w_{in} ($\mathrm{m\,s^{-1}}$) denotes the cut-in speed, derived from the theoretical power curve (dataset A and C) or assessed from the data (dataset B).
4. Discard observations corresponding to wind speeds greater than $w_{rt}+1$ for dataset A and C and w_{rt} for dataset B, where w_{rt} ($\mathrm{m\,s^{-1}}$) denotes the rated wind speed.
5. Finally, remove outliers using the Ratio-Skewed boxplot algorithm with $\kappa = 1.5$, see Appendix A.

After the cleaning procedure, dataset A passed from 52,560 to 35,428 data points. Dataset B passed from 118,224 to 57,094 observations. Dataset C passed from 266,433 to 216,724 observations when Tukey's clipping factor $\kappa = 1.5$ was used. In order to assess robustness to outliers, the cleaning was repeated with larger values of κ and even the raw data were fed to Beta regression, see 5.6 for details.

5.2. Performance indices

The following performance indices were used to evaluate predictive accuracy and goodness of fit on both the training and test sets:

1. Weighted Mean Absolute Percentage Error (WMAPE):
$$\mathrm{WMAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|} \times 100$$
2. Mean Absolute Error (MAE):
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$
3. Root Mean Square Error (RMSE):
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$
4. Coefficient of Determination (R^2):
$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \times 100$$
5. Cross Entropy (CE), see [27]:
$$\mathrm{CE} = -\frac{1}{n}\left(\sum_{i=1}^{n} \log(f(y_i; \mu(\beta, x_i), \phi(\theta, x_i))) \right)$$

Here, $\bar{y}$ and $\bar{\hat{y}}$ represent the arithmetic mean of, y_i and $\hat{y}_i$, respectively. The flexibility offered by probabilistic model is exploited in the choice of the point estimate $\hat{y}_i = \hat{y}_i(x_i)$ which is either the median of the Beta distribution when WMAPE and MAE are evaluated or its mean when RMSE and R^2 are considered. Finally, the cross entropy, also interpretable as a negative average likelihood, is used to measure the distance of the predicted Beta distribution with respect to the observed distribution of (training or test) data.

5.3. Experiments

Nine variants of the Beta regression model were assessed. More precisely, combinations of three preconditioning strategies p_1, p_2, p_3, three regression models r_1, r_2, r_3, and two dispersion models d_1, d_2 were considered:

p_1: no preconditioning;
p_2: theoretical power curve as preconditioner;
p_3: natural cubic spline preconditioner where the number K of equally spaced knots is chosen by crossvalidation;
r_1: affine:

$$\eta = \beta_0 + \beta_1 w_s;$$

r_2: quadratic:

$$\eta = \beta_0 + \beta_1 w_s + \beta_2 w_s^2;$$

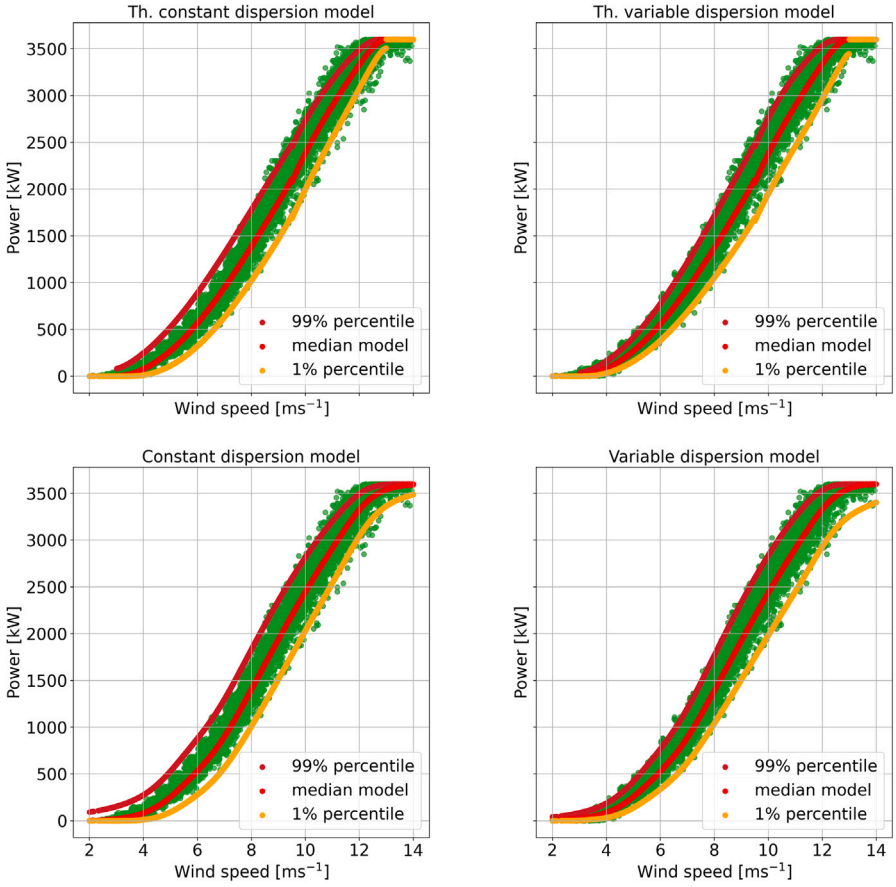

Fig. 4. Dataset A (test data): probabilistic power curve models fitted via Beta Regression with different models. Top left: model M8; top right: M9, bottom left: M5; bottom right: M6. First row: preconditioning via theoretical power curve (p_2); second row: preconditioning via spline (p_3); first column: constant dispersion (d_1); second column: variable dispersion (d_2).

Table 1
Model table.

Model	M1	M2	M3	M4	M5	M6	M7	M8	M9
precond.	p_1	p_1	p_1	p_1	p_3	p_3	p_3	p_2	p_2
regr. model	r_1	r_2	r_1	r_2	r_1	r_1	r_3	r_2	r_2
disp. model	d_1	d_1	d_2	d_2	d_1	d_2	d_1	d_1	d_2

r_3: surface, i.e. a bivariate function of wind speed w_s and wind direction ψ:

$$\eta = \beta_0 + \beta_1 w_s + \beta_2 w_s \sin(\psi) + \beta_3 w_s \cos(\psi);$$

d_1: constant dispersion model (constant ϕ);
d_2: variable dispersion model ($\phi = \exp(\theta_0 + \theta_1 w_s)$).

For dataset A, all the models listed in Table 1 were tested.

For dataset B, only models M1–M7 were tested. Note that for dataset B, the theoretical power curve is not available so that models M8 and M9 could not be estimated For both datasets, the first 75% of the data were used as training set, leaving the remaining 25% as test set. Training and test data for dataset A are plotted in the lower panels of Fig. 1. Dataset C was used to assess the robustness of the Beta regression in the presence of different levels of noise in the data and compare its performance with another probabilistic method, QRF. For this purpose, model M6 was chosen.

5.4. Results on dataset A

The performances of models M1–M9 on dataset A are summarized in Table 2, where WMAPE, MAE, RMSE, and R^2 are reported for both training and test data. In each column, boldface numbers highlight the best performing models for the given index. It appears that all models achieved an R^2 above 98%. It is also observed that R^2 remained stable when passing from training to test, indicating the absence of overfitting issues. A few models exhibit a superior performance, namely M5–M7 that seem to take advantage of the flexibility provided by the spline preconditioner p_3. Overall, the bivariate model M7, accounting for both wind speed and direction, achieved the best performances, although on test data is was just marginally better that M5–M6. The weakest performances were those of M1–M4, i.e. the models without preconditioner. A somehow intermediate performance was achieved by M8–M9 that use the theoretical power curve as preconditioner.

The identified probabilistic power curves for models M5, M6, M8, M9 are displayed in Fig. 4 against the test data. The plots confirm that M5 and M6 describe remarkably well the joint density of wind speed and power.

As seen from Table 3, model M6 achieved the smallest cross entropy both on training and test. For this models, a visual validation of the conditional density of power given wind speed is provided in Fig. 5: after splitting the wind speed axis in nine bins the conditional distribution of power predicted by M6 was plotted over the histogram of the observed power values (see Appendix B for the technical details). Apparently, a very good agreement was achieved. For the sake of comparison, in Fig. 2, the conditional distributions predicted by the constant dispersion model M5 are displayed against the histograms. Although the cross entropy on test data of M5 is larger than that of M6, it is seen that the simpler model still provides a fairly good approximation.

The application of model M6 on a time window is illustrated in Fig. 6, where in the top panel the 98% interval prediction is plotted

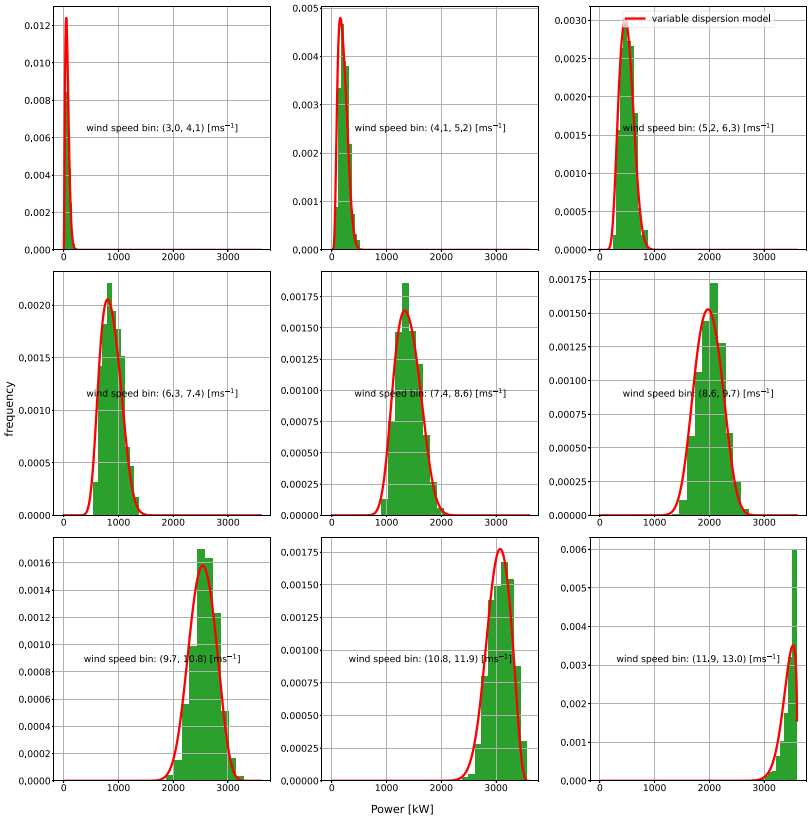

Fig. 5. Dataset A (test data): Model M6. Wind power histograms (green) corresponding to 9 bins of wind speed ranging from the cut-in speed (top left panel) to the rated one (bottom right). The red curve is the distribution predicted by fitting the training data with a Beta regression model with variable dispersion, preconditioned using a spline function.

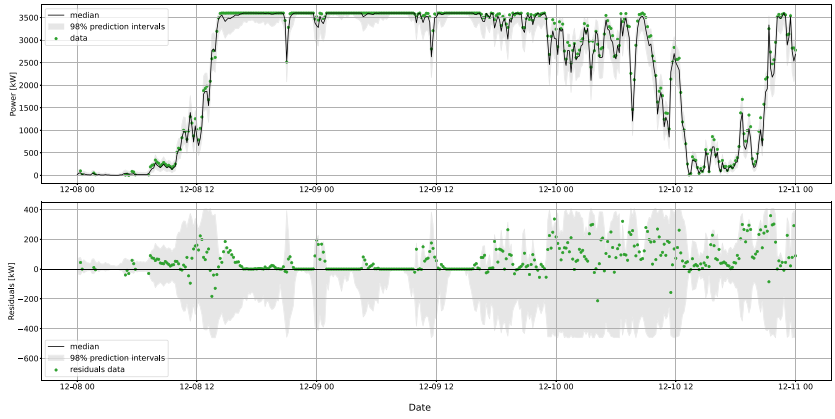

Fig. 6. Dataset A (test data): Model M6. Upper panel: time series (08/12/2018-11/12/2018) of observed wind power (green dots) and predictions (black) given by the median of the Beta regression model with variable dispersion and spline preconditioning. Lower panel: prediction errors (green dots). The pointwise 98% confidence bands are displayed as gray regions. The (asymmetrical) limits are calculated according to the Beta distribution model, so their amplitude and shape change significantly with wind speed.

against test data. In the bottom panel, the forecast errors are displayed with the corresponding time-varying 98% bands.

Finally, the model M7, accounting also for wind direction, yielded the power surface, displayed in Fig. 7 together with the test data. As seen from the plot, wind direction does not play a major role, although the performance indices in Table 2 highlight a marginal benefit coming from the inclusion of direction information.

5.5. Results on dataset B

Table 4 summarizes the results obtained, showcasing scores for WMAPE, MAE, RMSE, and R^2 on both training and test data. The R^2 values are still consistently high, hovering around 96%–98%, indicating a very good predictive capability across all experiments. Again, models M5–M7 outperform M1–M4 and the lowest cross entropy was achieved by M6, see Table 5.

The conditional density of power given wind speed predicted by model M6 is displayed in Fig. 3. Also for this dataset, the agreement between the predicted distribution and observed data is remarkably good, confirming the soundness of the modeling strategy.

5.6. Results on dataset C

The purpose of the experiments on dataset C was to assess the robustness of the Beta regression-model M6 in the presence of different levels of noise in the data and to compare its performance with that of a

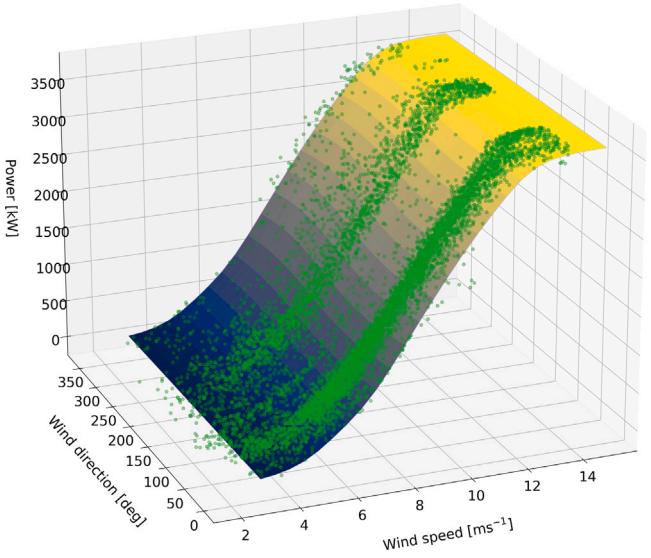

Fig. 7. Dataset A (test data): Model M7. Observed data (green dots) against power surface fitted by a Beta regression model with constant dispersion and spline preconditioning. Power depends mainly on wind speed and is only marginally affected by wind direction.

Table 2
Scores on dataset A.

Model	WMAPE (%)		MAE (kW)		RMSE (kW)		R^2 (%)	
	Median		Median		Mean		Mean	
	Train	Test	Train	Test	Train	Test	Train	Test
M1	7.24	6.59	103.6	111.6	142.1	141.6	98.64	98.52
M2	7.19	6.67	102.9	113.1	140.3	143.1	98.65	98.54
M3	7.16	6.57	102.5	111.4	139.9	143.0	98.65	98.54
M4	7.19	6.54	102.9	110.9	142.9	147.9	98.52	98.39
M5	5.52	**5.25**	79.0	**88.9**	113.1	124.2	99.07	98.92
M6	5.48	5.29	78.4	89.6	112.5	126.7	99.08	98.91
M7	**5.33**	**5.25**	**76.3**	89	**109.9**	**123.0**	**99.13**	**98.94**
M8	5.71	5.54	81.71	93.9	116.8	134.9	99.01	98.75
M9	5.66	5.54	81.1	93.9	116.6	136.5	99.00	98.74

Table 3
Cross entropy of dataset A models.

Model	Cross entropy	
	Train	Test
M1	−1.88	−1.67
M2	−1.91	−1.77
M3	−2.02	−1.94
M4	−2.05	−1.94
M5	−2.18	−1.98
M6	**−2.33**	**−2.23**
M7	−2.21	−2.01
M8	−1.85	−1.83
M9	−2.00	−1.94

flexible nonparametric probabilistic method, i.e. QRF [26]. It is known that QRF requires careful hyperparameter tuning to avoid overfitting, a task that has been addressed with the help of the Python library Optuna, see Appendix C for the details.

A range of values of the clipping factor κ were used to tune the selectivity of outlier removal, namely $\kappa \in \{1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}$. Moreover, Beta regression and QRF were applied also to raw data.

In Tables 6 and 7, some performance indices on training and test data are given with respect to different values of the clipping factor κ, considering also the case of no data-cleaning (raw data). In terms of WMAPE, MAE, RMSE and R^2, the performances of the QRF are just marginally better than those of Beta regression. In order to test the

Table 4
Scores on dataset B.

Model	WMAPE (%)		MAE (kW)		RMSE (kW)		R^2 (%)	
	Median		Median		Mean		Mean	
	Train	Test	Train	Test	Train	Test	Train	Test
M1	11.77	13.78	75.2	71.6	98.9	96.7	96.66	96.19
M2	10.61	11.98	67.9	62.2	88.9	83.4	97.26	96.80
M3	10.70	12.37	68.4	64.2	92.3	87.4	96.93	96.45
M4	10.61	12.23	67.9	63.5	91.0	86.2	97.03	96.57
M5	8.29	9.69	53.0	50.3	72.7	68.4	98.10	97.76
M6	8.24	**9.51**	52.7	**49.4**	72.3	**66.9**	98.11	97.78
M7	**8.14**	**9.51**	**52.1**	**49.4**	**71.6**	67.1	**98.16**	**97.85**

Table 5
Cross entropy of dataset B models.

Model	Cross entropy	
	Train	Test
M1	−1.51	−1.56
M2	−1.62	−1.65
M3	−1.70	−1.73
M4	−1.70	−1.73
M5	−1.87	−1.90
M6	**−1.97**	**−2.02**
M7	−1.89	−1.91

Table 6
Scores on dataset C - Beta regression model with variable dispersion.

Clipping factor κ	WMAPE (%)		MAE (kW)		RMSE (kW)		R^2 (%)	
	Median		Median		Mean		Mean	
	Train	Test	Train	Test	Train	Test	Train	Test
1.5	5.39	5.71	41.07	41.33	55.35	56.41	99.27	99.28
2	5.47	5.80	41.63	42.01	56.50	57.67	99.24	99.25
2.5	5.50	5.82	41.87	42.17	57.15	58.01	99.22	99.24
3	5.51	5.84	41.95	42.33	57.37	58.38	99.22	99.23
3.5	5.53	5.85	42.12	42.37	58.00	58.51	99.20	99.22
4	5.54	5.87	42.19	42.53	58.41	59.10	99.19	99.21
4.5	5.54	5.88	42.23	42.56	58.55	59.17	99.19	99.20
5	5.55	5.92	42.29	42.86	58.78	60.49	99.18	99.17
Raw	5.56	6.66	42.66	46.90	60.25	97.80	99.14	97.79

Table 7
Scores on dataset C - QRF.

Clipping factor κ	WMAPE (%)		MAE (kW)		RMSE (kW)		R^2 (%)	
	Median		Median		Mean		Mean	
	Train	Test	Train	Test	Train	Test	Train	Test
1.5	5.16	5.48	39.33	39.63	54.40	55.75	99.29	99.29
2	5.29	5.64	40.31	40.83	55.91	57.46	99.25	99.25
2.5	5.29	5.61	40.27	40.64	56.28	57.44	99.24	99.25
3	5.28	5.60	40.23	40.60	56.45	57.84	99.23	99.24
3.5	5.34	5.66	40.67	40.99	57.27	58.15	99.21	99.23
4	5.31	5.67	40.46	41.05	57.44	58.68	99.21	99.21
4.5	5.34	5.69	40.74	41.18	57.80	58.95	99.20	99.21
5	5.32	5.71	40.57	41.32	57.86	60.03	99.19	99.18
Raw	5.33	6.40	40.91	45.09	59.25	96.82	99.16	97.81

robustness of the estimated percentiles, in Figs. 8 and 9 we plot the 98% confidence intervals obtained from data cleaned with $\kappa = 1.5$, $\kappa = 5$ and from the raw data. As can be seen from the three lower panels, the Beta regression proves to be robust even in the presence of an increasing fraction of outliers. On the contrary, looking at the top three panels, it can be seen that the lower confidence limit estimated by QRF becomes unreliable as the fraction of outliers increases. It is also worth noting that QRF takes significantly longer to run.

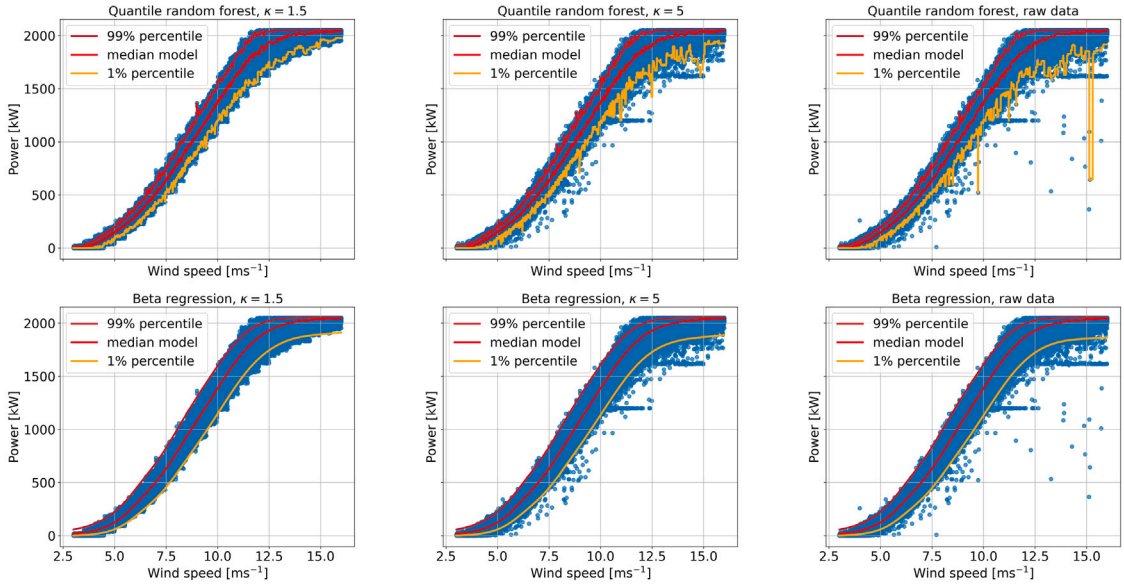

Fig. 8. Dataset C (training data): Comparison between QRF (upper panels) and Beta Regression-Model M6 (lower panels) for decreasing selectivities of outlier removal: clipping factor $\kappa = 1.5$ (left), $\kappa = 5.0$ (center), and direct use of raw data (left).

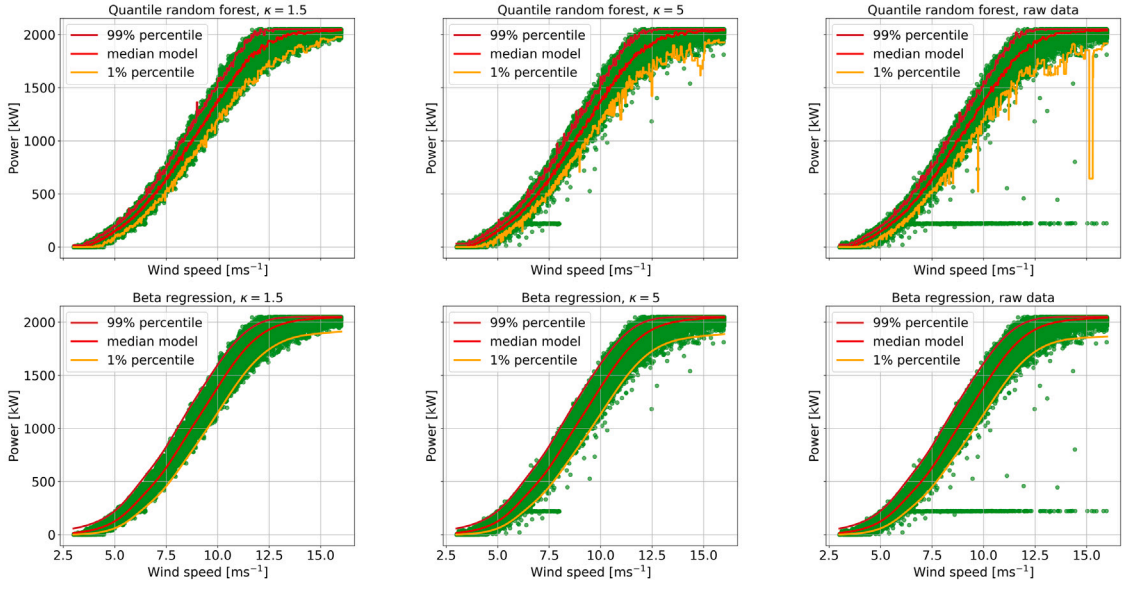

Fig. 9. Dataset C (test data): Comparison between QRF (upper panels) and Beta Regression-Model M6 (lower panels) for decreasing selectivities of outlier removal: clipping factor $\kappa = 1.5$ (left), $\kappa = 5.0$ (center), and direct use of raw data (left).

6. Discussion

As explained in Section 3, the major challenges of probabilistic power curve modeling are the heteroscedasticity and asymmetry of the power distribution conditional on the wind speed. The results showed that a Beta regression model can effectively address both issues.

Resorting to a GLM brings the advantage of interpretability and ease of inclusion of additional features, e.g. wind direction, but also temperature, humidity and so on. One may think that simplicity comes at the expense of flexibility. It is here that comes into play the role of preconditioning. It is not just a tool to gain elasticity, but also a transparent way to incorporate prior knowledge into the model. In fact, a reference power curve, possibly provided by the manufacturer, can be used as preconditioner. Alternatively, a preconditioner can be chosen from the many parametric and nonparametric power curve models available in the literature.

The results obtained on three public datasets convey valuable insights. First, the importance of including a preconditioner compared to its absence. In particular, the spline preconditioner proves very effective, but using the theoretical curve is still a viable alternative, leading to a less complex model. The role of the preconditioner is also valuable in the context of the rapid deployment of probabilistic power curves for new sites with scarce historical data. Theoretical power curves or curves from similar plants could be employed to complement the available data in order to get a model without having to wait for a long data collection.

With probabilistic modeling it is possible to adapt the point estimate to the desired performance index. While RMSE and R^2 may call for the conditional mean, the WMAPE and MAE may be better accommodated by the conditional median. By these choices it was seen that the ranking of the models was reasonably insensitive to the choice of the index.

Beta regression compared favorably with a nonparametric probabilistic model such as QRFs. In fact the point estimates, mean and median, perform almost the same in terms of WMAPE, MAE, RMSE and R^2, while the confidence limits of Beta Regression were definitely more robust when different levels of selectivity were tested for outlier removal.

The validation experiments showed the potential benefit of including an additional feature like the wind direction, although it did not give a decisive advantage because wind speed remains the main determinant of produced power. It is expected that the ease of including additional features may prove particularly advantageous if the Beta regression approach were extended to the use of weather forecasts. In that case, forecasts are used in place of observed wind speed and direction and the inclusion of further features may become crucial.

The proposed Beta regression model could prove useful for the monitoring of wind turbines and wind farms, e.g. for fault detection purposes. Another use could regard the design of probabilistic wind power forecasts based on the indirect approach, which plugs the forecast wind speed in a power curve model.

The validation on three public datasets has been successful, but further experience needs to be accumulated to assess the general validity of this framework. Some confidence in the wide applicability comes from the flexibility of the Beta distribution, especially when its dispersion parameter can be variable.

CRediT authorship contribution statement

Marco Capelletti: Data curation, Conceptualization, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing, Supervision, Validation. **Davide M. Raimondo:** Funding acquisition, Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Giuseppe De Nicolao:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been partially supported by the Italian Ministry of University and Research in the framework of the 2017 PRIN, Grant no. 2017YKXYXJ.

Appendix A. Outlier detection

A variety of techniques have been proposed for detecting outliers in power wind vs wind speed data [5]. Among them, one might mention quantile-based methods such as outlier boxplot ones, see e.g. [28], that filter data in two steps, both vertically and horizontally. We performed outlier detection on the vertical axis only, by means of outlier boxplots applied on subsets selected by binning the x-axis, i.e. wind speed.

In view of heteroscedasticity and skewness of wind power, the Ratio-Skewed boxplot [29] was preferred to standard boxplots, whose outlier detection performances tend to degrade on asymmetric distributions. The Ratio-Skewed boxplot is a modification of the traditional boxplot that can be used with any univariate dataset, either symmetric or skewed.

To apply the Ratio-Skewed boxplot algorithm, the data were divided into bins. For each bin $B = [x_{\min}, x_{\max}]$, we considered the subset $\{(x_i, y_i) | x_i \in B\}$ and computed the 1st quartile Q_1, the median Q_2, the 3rd quartile Q_3 of the $\{y_i\}$ coordinates, as well as the interquartile range $H = Q_3 - Q_1$. Next, the Bowley's Coefficient B_c, a sample quartile-based measure of skewness, was computed as

$$B_c = \frac{Q_3 + Q_1 - 2Q_2}{H}$$

The lower and upper skewness adjustment factors, R_L and R_U were calculated as:

$$R_L = \frac{1 - B_c}{1 + B_c}, \qquad R_U = \frac{1 + B_c}{1 - B_c}$$

An observation y_i was labeled as outlier if $y_i \notin [Q_1 - \kappa H R_L, Q_3 + \kappa H R_U]$, where κ is a suitable clipping factor, that, following Tukey, is typically set equal to 1.5. Alternatively, the clipping factor can be adjusted, e.g. by using larger values to reduce the selectivity of outlier removal. When the distribution is symmetric, the Bowley's Coefficient B_c becomes zero. In this case, $R_L = R_U = 1$ so that for $\kappa = 1$ the Ratio-Skewed boxplot reverts to the traditional boxplot. Regarding the determination of widths for the wind speed bins, in [30] different widths were tested, ranging from $0.1\,\mathrm{m\,s^{-1}}$ to $1\,\mathrm{m\,s^{-1}}$. The optimum value of the amplitude of the single sub-wind interval was reported to be equal to $0.5\,\mathrm{m\,s^{-1}}$, which was also adopted in our study. For dataset A, the scatter plot of wind power vs wind speed after outlier removal is displayed in the upper right panel of Fig. 1.

Appendix B. Visual validation of conditional distribution model

In this appendix we discuss the visual validation of a conditional distribution model. More precisely, assume that the pairs $\{x_i, y_i\}, i = 1, \dots, n$, are sampled from a joint distribution $f(x, y)$. We look for a method to visually inspect if a function $h(y, x)$ describes well the conditional distribution

$$f(y|x) = \frac{f(x, y)}{f(x)}.$$

In order to visualize the empirical conditional distribution of $y|x$ around $x = \bar{x}$, we can plot the histogram of $\{y_i | x_i \in B\}$, where $B = [\bar{x} - \delta/2, \bar{x} + \delta/2]$ denotes a bin on the x-axis. We let $\mathcal{B} = \{i | x_i \in B\}$, denote the index set and $n_B = |\mathcal{B}|$ the sample size of observations lying in the bin.

A simple and direct check would be comparing such histogram with $h(y, \bar{x})$. This, however, may not work if the shape of $f(\cdot|x)$ is very sensitive to small changes of x. In that case, one should narrow the bin B by choosing a small δ, but n_B may result too small to obtain a meaningful histogram.

As a matter of fact, the distribution of $\{y_i | x_i \in B\}$ is well approximated by $f(\cdot|\bar{x})$ only if δ is small. The conditional distribution of $y_i|x_i$ accounting for the dispersion of x_i around $\bar{x}$ involves an averaging described by the following mixture

$$f_B(\cdot|\bar{x}) = \frac{1}{n_B} \sum_{i \in \mathcal{B}} f(\cdot|x_i).$$

Therefore, even when δ is not small, the visual inspection can still be performed if the histogram of y_i is superimposed to the mixture

$$h_B(\cdot, \bar{x}) = \frac{1}{n_B} \sum_{i \in \mathcal{B}} h(\cdot, x_i).$$

In Figs. 2, 3, and 5, the distributions $h_B(\cdot, \bar{x})$ are plotted as red curves thus enabling the comparison with the histograms of test data falling in the wind speed bins.

Appendix C. Hyperparameter tuning of quantile regression forest

QRF is a non-parametric method known for its competitiveness but which requires careful hyperparameter tuning to avoid overfitting. A cross-validation procedure was performed for the following hyperparameters:

- 'criterion': {'squared_error', 'absolute_error'}
- 'n_estimators': [1, 200]

- ‘min_samples_split’: [100, 1000]
- ‘min_samples_leaf’: [20, 1000]
- ‘max_depth’: [1, 50]

Here, ‘criterion’ measures split quality, with ‘squared_error’ minimizing mean squared error (L_2 loss) and ‘absolute_error’ minimizing mean absolute error (L_1 loss). ‘n_estimators’ is the number of trees, ‘min_samples_split’ is the minimum samples for node split, ‘min_samples_leaf’ is the minimum samples at a leaf node, and ‘max_depth’ is the maximum tree depth.

Since the training data exceeded 156,000 samples (the precise number depends on the choice of clipping factor κ, see Appendix A), and the QRF with ‘absolute_error’ criterion had high runtimes, 15% of the training data (approximately 23,000 samples) were randomly selected for cross-validation using Optuna [31]. Optuna, an open-source Python library, employs Bayesian optimization to efficiently search and tune hyperparameters. The number of Optuna trials was set to 50 to explore different combinations of hyperparameters.

References

[1] Z. Hameed, Y. Hong, Y. Cho, S. Ahn, C. Song, Condition monitoring and fault detection of wind turbines and related algorithms: A review, Renew. Sustain. Energy Rev. 13 (1) (2009) 1–39.

[2] M. Lydia, S.S. Kumar, A.I. Selvakumar, G.E.P. Kumar, Wind resource estimation using wind speed and power curve models, Renew. Energy 83 (2015) 425–434.

[3] V. Sohoni, S. Gupta, R. Nema, et al., A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems, J. Energy 2016 (2016).

[4] M. Lydia, S.S. Kumar, A.I. Selvakumar, G.E.P. Kumar, A comprehensive review on wind turbine power curve modeling techniques, Renew. Sustain. Energy Rev. 30 (2014) 452–460.

[5] Y. Wang, Q. Hu, L. Li, A.M. Foley, D. Srinivasan, Approaches to wind power curve modeling: A review and discussion, Renew. Sustain. Energy Rev. 116 (2019) 109422.

[6] M. Lydia, A.I. Selvakumar, S.S. Kumar, G.E.P. Kumar, Advanced algorithms for wind turbine power curve modeling, IEEE Trans. Sustain. Energy 4 (3) (2013) 827–835.

[7] M. Capelletti, D.M. Raimondo, G. De Nicolao, Regression dilution effects in wind power prediction from wind speed forecasts, in: 2022 IEEE Conference on Control Technology and Applications, CCTA, IEEE, 2022, pp. 137–143.

[8] M. Javadi, A.M. Malyscheff, D. Wu, C. Kang, J.N. Jiang, An algorithm for practical power curve estimation of wind turbines, CSEE J. Power Energy Syst. 4 (1) (2018) 93–102.

[9] Y. Yan, L.A. Osadciw, G. Benson, E. White, Inverse data transformation for change detection in wind turbine diagnostics, in: 2009 Canadian Conference on Electrical and Computer Engineering, IEEE, 2009, pp. 944–949.

[10] D. Villanueva, A. Feijóo, Comparison of logistic functions for modeling wind turbine power curves, Electr. Power Syst. Res. 155 (2018) 281–288.

[11] A. Marvuglia, A. Messineo, Monitoring of wind farms’ power curves using machine learning techniques, Appl. Energy 98 (2012) 574–583.

[12] A. Kusiak, H. Zheng, Z. Song, On-line monitoring of power curves, Renew. Energy 34 (6) (2009) 1487–1493.

[13] T. Ouyang, A. Kusiak, Y. He, Modeling wind-turbine power curve: A data partitioning and mining approach, Renew. Energy 102 (2017) 1–8.

[14] O. Janssens, N. Noppe, C. Devriendt, R. Van de Walle, S. Van Hoecke, Data-driven multivariate power curve modeling of offshore wind turbines, Eng. Appl. Artif. Intell. 55 (2016) 331–338.

[15] Y. Wang, Q. Hu, D. Srinivasan, Z. Wang, Wind power curve modeling and wind power forecasting with inconsistent data, IEEE Trans. Sustain. Energy 10 (1) (2018) 16–25.

[16] Y. Wang, Y. Li, R. Zou, A.M. Foley, D. Al Kez, D. Song, Q. Hu, D. Srinivasan, Sparse heteroscedastic multiple spline regression models for wind turbine power curve modeling, IEEE Trans. Sustain. Energy 12 (1) (2020) 191–201.

[17] Y. Wang, Q. Hu, S. Pei, Wind power curve modeling with asymmetric error distribution, IEEE Trans. Sustain. Energy 11 (3) (2019) 1199–1209.

[18] R. Zou, J. Yang, Y. Wang, F. Liu, M. Essaaidi, D. Srinivasan, Wind turbine power curve modeling using an asymmetric error characteristic-based loss function and a hybrid intelligent optimizer, Appl. Energy 304 (2021) 117707.

[19] S. Ferrari, F. Cribari-Neto, Beta regression for modelling rates and proportions, J. Appl. Stat. 31 (7) (2004) 799–815.

[20] F. Cribari-Neto, A. Zeileis, Beta regression in R, J. Stat. Softw. 34 (2010) 1–24.

[21] B. Erisen, 2018 Scada data of a wind turbine in Turkey, 2023, Available online at: https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset (last accessed on 09.20.2023).

[22] S. Bhaskarpandit, Kaggle scada data of a wind turbine, 2023, Available online at: https://www.kaggle.com/datasets/theforcecoder/wind-power-forecasting (last accessed on 09.20.2023).

[23] C. Plumley, Penmanshiel wind farm data, 2023, http://dx.doi.org/10.5281/zenodo.8253010.

[24] K. Bruninx, E. Delarue, A statistical description of the error on wind power forecasts for probabilistic reserve sizing, IEEE Trans. Sustain. Energy 5 (3) (2014) 995–1002.

[25] S. Tewari, C.J. Geyer, N. Mohan, A statistical model for wind power forecast error and its application to the estimation of penalties in liberalized markets, IEEE Trans. Power Syst. 26 (4) (2011) 2031–2039.

[26] N. Meinshausen, G. Ridgeway, Quantile regression forests, J. Mach. Learn. Res. 7 (6) (2006).

[27] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

[28] Y. Zhao, L. Ye, W. Wang, H. Sun, Y. Ju, Y. Tang, Data-driven correction approach to refine power curve of wind farm under wind curtailment, IEEE Trans. Sustain. Energy 9 (1) (2017) 95–105.

[29] M. Walker, Y. Dovoedo, S. Chakraborti, C. Hilton, An improved boxplot for univariate data, Amer. Statist. 72 (4) (2018) 348–353.

[30] M. Marčiukaitis, I. Žutautaitė, L. Martišauskas, B. Jokšas, G. Gecevičius, A. Sfetsos, Non-linear regression model for wind turbine power curve, Renew. Energy 113 (2017) 732–741.

[31] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.