Forecasting Conditional Extreme Quantiles for Wind Energy

Carla Gonçalves*†, Laura Cavalcante*, Margarida Brito†‡, Ricardo J. Bessa* and João Gama*§
* INESC TEC, Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal
† FCUP, Faculty of Sciences of the University of Porto, Portugal, ‡ CMUP, Porto, Portugal
§ FEP, Faculty of Economics of the University of Porto, Portugal
Corresponding author: carla.s.goncalves@inesctec.pt

***Abstract*—Probabilistic forecasting of distribution tails (i.e., quantiles below .05 and above .95) is challenging for non-parametric approaches since data for extreme events are scarce. A poor forecast of extreme quantiles can have a high impact in various power system decision-aid problems. An alternative approach more robust to data sparsity is extreme value theory (EVT), which uses parametric functions for modelling distribution's tails. In this work, we apply conditional EVT estimators to historical data by directly combining gradient boosting trees with a truncated generalized Pareto distribution. The parametric function parameters are conditioned by covariates such as wind speed or direction from a numerical weather predictions grid. The results for a wind power plant located in Galicia, Spain, show that the proposed method outperforms state-of-the-art methods in terms of quantile score.**

***Index Terms*—Extreme quantiles, extreme value theory, forecasting uncertainty, wind energy.**

List of Acronyms and Symbols

Notation	Description
CDF	Cumulative distribution function
CRPS	Continuous ranked probability score
EVT	Extreme value theory
Exp_Tails	Exponential function
GBT	Gradient boosting tree regression
GBT_EVT	GBT combined with Hill estimator
GBT_tGPD	Proposed method combining GBT with truncated GPD
GPD	Generalized Pareto distribution
POT	Peaks-over-threshold
QR	Quantile regression
QR_EVT	QR combined with Hill estimator
QR_EVT_T	QR, Hill estimator and transformed power data
RES	Renewable energy sources
C	Installed power capacity
$H(.)$	Heaviside function
h	Sample size for tails representation, using GBT_tGPD
k	Sample size for extreme quantiles extrapolation
p	Number of covariates
n	Number of observations
X	p-dimensional vector of covariates
$\mathbf{x}$	Observed p-dimensional vector of covariates
Y	Target variable
y	Observed target variable
$Y_{1,n}...Y_{n,n}$	Ordered sample of Y
τ	Nominal proportion of a quantile, $\tau \in [0,1]$
$\rho_\tau(.)$	Pinball loss function
$\hat{Q}^{\text{exp}}(\tau\|\mathbf{x})$	Conditional quantile through exponential functions
$\hat{Q}^{\text{GBT}}(\tau\|\mathbf{x})$	Conditional quantile through a GBT model
$\hat{Q}^{\text{QR}}(\tau\|\mathbf{x})$	Conditional quantile through a QR model
$\hat{Q}^{\text{W}}(\tau\|\mathbf{x})$	Conditional extreme quantile through Weissmans estimator
$\hat{Q}_k^{\text{tGPD}}(\tau)$	Extreme quantile through POT estimator for truncated GPD
$\mathbf{1}_{(.)}$	Indicator function
$\boldsymbol{\beta}(\tau)$	QR model coefficients
$\hat{\gamma}(\mathbf{x})$	Conditional tail index estimator
$\Lambda_\lambda(.)$	Power transformation function
λ	Power parameter
$s(.)$	Similarity function between two CDF curves

The research leading to this work is being carried out as a part of the Smart4RES project (European Union's Horizon 2020, No. 864337). The sole responsibility for the content lies with the authors. It does not necessarily reflect the opinion of the Innovation and Networks Executive Agency (INEA) or the European Commission (EC), which are not responsible for any use that may be made of the information it contains. C. Gonçalves was supported by FCT within the Ph.D. grant PD/BD/128189/2016 with financing from POCH (Operational Program of Human Capital) and the EU (European Union). M.Brito was partially supported by CMUP (UID/MAT/00144/2019), which is funded by FCT with national (MCTES) and EU funds through FEDER, under the agreement PT2020.

I. Introduction

The growing integration of renewable energy sources (RES) brings new challenges to system operators and market players and robust forecasting models are fundamental for handling their variability and uncertainty. This fomented a growing interest in RES probabilistic forecasting techniques and its integration in decision-aid under risk [1].

Many satisfying methods already exist to forecast RES generation quantiles between .05 and .95, which can be parametric or non-parametric. An up-to-date literature review about RES probabilistic forecasting can be found in [2]. Parametric models assume that data is generated from a known probability distribution (e.g. Gaussian, Beta), whose parameters are estimated from the data. Non-parametric models do not make any assumptions about the shape of the probability distribution and comprise techniques such as quantile regression (QR) with radial basis functions [3], local QR [4], conditional kernel density estimation [5] and gradient boosting trees (GBT) [6]. It is also possible to find semi-parametric approaches, e.g., mixture of a censored distribution and probability masses on the upper and lower boundaries that transform wind power data

into a Gaussian distribution, whose mean and standard deviation are forecasted with a statistical model [7]; combination of linear regression, inverse (power-to-wind) transformation and censored normal distribution [8].

The main advantage of parametric methods is that the distribution's shape only depends on a few parameters, resulting in a simplified estimation and consequently requiring low computational costs. However, the choice of the parametric function is not straightforward. On the other hand, non-parametric models require a large number of observations to achieve good performance. Therefore, when estimating quantiles below .05 and above .95, non-parametric models tend to have poor performance due to data sparsity. This suggests the combination of both approaches to forecast the conditional probability function: intermediate quantiles are estimated with a non-parametric model and the extreme quantiles (or tails) with a parametric approach.

A poor forecast of extreme quantiles can have a high impact in different decision-aid problems, in particular when decision-makers are highly risk averse or the regulatory framework imposes high security levels. For instance, when setting operating reserve requirements system operators usually define risk (e.g., loss of load probability) levels below 1% [9]; the distribution's tails forecasting accuracy affects the decision quality of advanced RES bidding strategies that are based on risk metrics such as conditional value-at-risk [10]; dynamic line rating uncertainty forecasting for transmission grids also requires the use of low quantiles (e.g., 1%) [11]. Moreover, the generation of temporal and/or spatial-temporal trajectories (or random vectors) with a statistical method, such as the Gaussian copula [12], requires a full modelling of the distribution function and an accurate estimation of the tails avoids trajectories with "extreme" values. In all these use cases, it is important to underline that poor modelling of distribution' tails might lead to over and under-estimation of risk and consequently to worst decisions. This impact can be measured by metrics such as the Value of the Right Distribution that measures the difference in the cost of optimal solution, in stochastic programming, obtained with the forecasted and realized probability distribution [13].

By exploring concepts from extreme value theory (EVT), which is dedicated to characterise the stochastic behaviour of extreme values [14], the present paper proposes a novel wind power forecasting methodology, focused in improving the forecasting skill of the distribution's tails, which combines spatio-temporal information (obtained trough feature engineering), gradient boosting trees (GBT) as a non-parametric method for quantiles between .05 and .95 and the truncated generalized Pareto distribution (GPD) for the tails.

The remaining of this paper is organized as follows. Section II presents related work and identify contributions. Section III introduces the relevant statistical background of non-parametric and parametric methods. Section IV describes a novel forecasting method combining GBT with truncated GPD. Section V describes the experiments to evaluate the proposed method and conclusions are drawn in section VI.

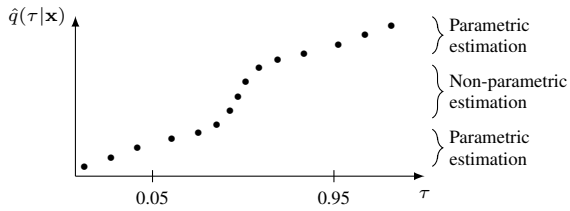

Fig. 1. The proposed method uses different estimators for intermediate and extreme quantiles.

II. Related Work and Contributions

In [15] and [16], a QR model is used to forecast the wind power quantiles from .05 to .95 and the distribution' tails are modeled using an exponential function. The exponential function requires the estimation of a single parameter that controls the tails' decay, the thickness parameter ρ. This parameter can be estimated by computing the mean of the observed power conditioned by the forecasted wind power, i.e., observed power is divided into equally populated bins according to forecasted wind power, then ρ is the average power associated to each bin. This procedure is not as flexible as those provided by an EVT estimator like GPD (used in this work), which models extreme events through distributions with two parameters (scale and shape), allowing it to estimate lightweight and heavier tails.

A two-stage EVT approach is proposed in [17] to estimate the extreme quantiles of a random variable Y conditioned by covariate X. First, the conditional quantiles are estimated with a local QR. Then, generalized extreme value distribution with a single parameter (i.e., extreme value index estimated using maximum likelihood) is applied to these non-parametrically estimated quantiles in order to construct an estimator for extreme quantiles. Similarly, the authors of [18] apply linear QR to estimate the intermediate conditional quantiles, which are then extrapolated to the upper tails by applying EVT estimators (e.g., Hill estimator) for heavy-tailed distributions (GPD is assumed). However, the conditional quantiles of Y are assumed to have a linear relation with X at the tails, which may be too restrictive in real-world applications. In order to overcome this limitation, the approach proposed in [19] works by first finding an appropriate power transformation of Y, then estimating the intermediate conditional quantiles of transformed Y using linear QR and finally extrapolating these estimates to extreme tails with EVT estimators. At the end, these quantiles are transformed back to the original scale.

More importantly, existing works only apply EVT as a post-processing step over a set of quantiles first estimated (or forecasted) by a non-parametric method [18]. However, since non-parametric models can suffer from high variability at the tails, the performance of EVT estimators may be compromised. In order to overcome this problem, we restrict non-parametric estimation to the intermediate quantiles, as depicted in Fig. 1. This estimation is then used to guide the parametric model by rating historically similar periods conditioned by the covariates.

Finally, two works proposed the use of spatio-temporal data in RES probabilistic forecasting: combination of GBT with feature engineering techniques to extract information from a grid of Numerical Weather Predictions (NWP) [6]; hierarchical forecasting models to leverage turbine-level data [20]. Both works do not deal or propose a specific methodology to forecast conditional distribution's tails.

III. Background: Non-parametric and Parametric Methods

This section presents the main statistical methods to construct the proposed method and baseline approaches. In what follows, $\mathbf{x}_i$ is the observed p-dimensional vector of covariates and y_i is the target variable, with $i \in \{1, \dots, n\}$.

A. Non-parametric Methods

1) Quantile Regression: The QR model [21] estimates the conditional quantile function of Y given X,

$$Q^{\text{QR}}(\tau|X) = \beta_0(\tau) + \beta_1(\tau)X_1 + \cdots + \beta_p(\tau)X_p, \quad (1)$$

for the nominal proportion $\tau \in [0, 1]$, by minimizing

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau\Big(y_i - \beta_0(\tau) - \sum_{j=1}^{p} \beta_j(\tau)x_{ij}\Big), \quad (2)$$

where $\hat{\boldsymbol{\beta}}(\tau) = (\hat{\beta}_0(\tau), \dots, \hat{\beta}_p(\tau))$ are unknown coefficients depending on τ, and $\rho_\tau(u)$ is the *pinball loss function* [21].

2) Gradient Boosting Trees: A GBT model for quantile forecasting is constructed by combining base learners (i.e., regression trees), f_j, recurrently on modified data,

$$Q_j^{\text{GBT}}(\tau|X) = Q_{j-1}^{\text{GBT}}(\tau|X) + \eta f_j(\tau|X). \quad (3)$$

with each regression tree f_j fitted using the negative gradients as target variable, and as part of an additive training process to minimize the *pinball loss function*

$$\hat{f}_j(\tau|X) = \arg\min_{f_j} \sum_{i=1}^{n} \rho_\tau\big(y_i, \hat{Q}_{j-1}^{\text{GBT}}(\tau|\mathbf{x}_i) + \eta f_j(\tau|\mathbf{x}_i)\big). \quad (4)$$

The initial model Q_1^{GBT} is typically the unconditional τ-quantile of $\mathbf{y}$. The challenge of GBT is to tune the different hyperparameters, which are related with the regression trees and the boosting process — see [6] for more details.

3) Rearrangement of quantiles: Since both QR and GBT solve an optimization problem for each quantile τ independently, quantile crossing may happen, i.e. $Q(\tau_1|\mathbf{x}) < Q(\tau_2|\mathbf{x})$ for $\tau_1 > \tau_2$. Post-processing is applied to the model's output to ensure that the estimated cumulative function is monotonically non-decreasing. We can monotonize the function by considering the proportion of times the quantile $Q(\tau|\mathbf{x})$ is bellow a certain y, mathematically provided by the cumulative distribution function (CDF)

$$F(y|\mathbf{x}) = \int_0^1 \mathbf{1}_{Q(\tau|\mathbf{x}) \le y} d\tau \quad (5)$$

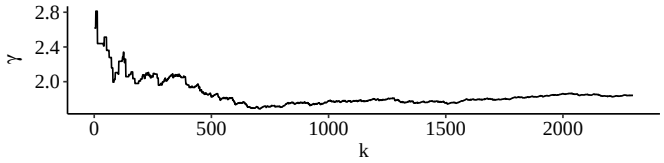

Fig. 2. Illustration of γ value in function of k. The first stable part of the plot happens for $k \approx 700$.

which is monotone at the level y, and then use its quantile function

$$\tilde{Q}(\tau|\mathbf{x}) = F^{-1}(\tau|\mathbf{x}) \quad (6)$$

which is monotone in τ [22].

B. Parametric Methods for Extreme Quantiles

1) Exponential function: In [15], distribution' tails of wind power are approximated by exponential functions. Given the estimated conditional quantiles for nominal proportion between .05 and .95, the extreme quantiles are computed as

$$\hat{Q}^{\text{exp}}(\tau|\mathbf{x}) = \begin{cases} \hat{Q}(.05|\mathbf{x})^{\frac{\ln(\frac{.05}{\rho})}{\ln(\frac{\tau}{\rho})}}, & \tau < .05, \\ C\left(1 - \left(1 - \frac{\hat{Q}(.95|\mathbf{x})}{C}\right)^{\frac{\ln(\frac{1-.95}{\rho})}{\ln(\frac{1-\tau}{\rho})}}\right), & \tau > .95, \end{cases} \quad (7)$$

where ρ corresponds to the thickness parameter for the exponential extrapolation and C is the installed capacity. Since the lower and upper tails may have different behaviors, ρ is independently estimated for each tail by maximum likelihood [16].

2) Hill-based methods: In [18] and [19], a QR model is combined with EVT estimators. First, a local QR model is used to estimate the conditional quantiles τ_j, denoted as $\hat{Q}^{\text{QR}}(\tau_j|\mathbf{x})$, $j \in \{1, ..., n - [n^\eta]\}$, for some $0 < \eta < 1$, being [u] the integer part of u. Then, using these values, extreme quantiles are computed through an adaptation of Weissman's estimator,

$$\hat{Q}^{\text{W}}(\tau|\mathbf{x}) = \left(\frac{1 - \tau_{n-k}}{1 - \tau_n}\right)^{\hat{\gamma}(\mathbf{x})} \hat{Q}^{\text{QR}}(\tau_{n-k}|\mathbf{x}), \quad (8)$$

where $\hat{\gamma}(\mathbf{x})$ is based on Hill's estimator

$$\hat{\gamma}(\mathbf{x}) = \frac{1}{k - [n^\eta]} \sum_{j=[n^\eta]}^{k} \log \frac{\hat{Q}^{\text{QR}}(\tau_{n-j}|\mathbf{x})}{\hat{Q}^{\text{QR}}(\tau_{n-k}|\mathbf{x})}. \quad (9)$$

In EVT, the selection of k is an important and challenging problem. The value k represents the effective sample size for tail extrapolation. A smaller k leads to estimators with larger variance, while larger k results in more bias, when estimating $\gamma(\mathbf{x})$. In practice, a commonly used heuristic approach for choosing k is to plot the estimated γ versus k and then choose a suitable k corresponding to the first stable part of the plot [14], see Fig. 2.

In [19], the response variable of the QR model is the power transformation $\Lambda_\lambda(.)$ of Y that aims to improve the linear relation with $\mathbf{x}$, i.e.

$$\Lambda_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0. \end{cases} \quad (10)$$

For this approach, k is estimated to minimize

$$\arg\min_{k\geq 1} \sum_{i=1}^{n} \hat{\lambda}\hat{\gamma}(\mathbf{x}_i) - \hat{\gamma}^*(\mathbf{x}_i), \tag{11}$$

where

$$\hat{\gamma}^*(\mathbf{x}) = M_{0,n}^{(1)} + 1 - \frac{1}{2}\Big(1 - \frac{(M_{0,n}^{(1)})^2}{M_{0,n}^{(2)}}\Big)^{-1} \tag{12}$$

$$M_{0,n}^{(i)} = \frac{1}{k - [n^\eta]} \sum_{j=[n^\eta]}^{k} \Big(\log \frac{\hat{Q}^{QR}(\tau_{n-j})}{\hat{Q}^{QR}(\tau_{n-k})}\Big)^i. \tag{13}$$

3) Peaks-over-threshold (POT) method with truncation: Since wind power generation is limited between 0 and installed capacity C, we observe the truncated random variable Y, $Y <= C$. The work in [23] provides an estimator for the extreme quantiles by using a random sample of Y, with independent and identically distributed observations, i.e. does not consider that Y is conditioned by covariates $\mathbf{x}$. The POT method [24] is adapted to estimate extreme quantiles from a GPD distribution affected by truncation at point C. The quantiles for Y are estimated by

$$\hat{Q}_k^{\text{tGPD}}(1-p) = Y_{n-k,n} + \frac{\hat{\sigma}_k}{\hat{\xi}_k}\left(\left[\frac{\hat{D}_{C,k} + \frac{(k+1)}{(n+1)}}{p(\hat{D}_{C,k}+1)}\right]^{\hat{\xi}_k} - 1\right), \tag{14}$$

where $Y_{1,n} < \cdots < Y_{n,n}$ is the ordered sample, $\hat{\xi}_k$ and $\hat{\sigma}_k$ are the maximum likelihood estimates adapted for truncation, and $\hat{D}_C$ the truncation odds estimator

$$\hat{D}_{C,k} = \max\left\{0, \frac{k}{n}\frac{(1+(\hat{\xi}_k/\hat{\sigma}_k)E_{1,k})^{-1/\hat{\xi}_k} - \frac{1}{k}}{1-(1+(\hat{\xi}_k/\hat{\sigma}_k)E_{1,k})^{-1/\hat{\xi}_k}}\right\}, \tag{15}$$

with $E_{j,k} = Y_{n-j+1,n} - Y_{n-k,n}$.

The GDP estimator will be used in our proposed method because (i) the shape parameter ξ allows modeling everything from extreme events with lightweight distribution (ξ<0) to events with exponential distribution (ξ=0) and events with heavy distribution (ξ>0); (ii) the existence of estimators for truncated GPD that can handle random variables with limited support like wind power.

IV. Gradient Boosting Trees with a Truncated Generalized Pareto Model

As previously discussed in section II, EVT estimators are, at present, used in post-processing steps for quantiles forecasted with a non-parametric model, i.e., the non-parametric model forecasts all quantiles (including extreme quantiles) and EVT estimators are applied to correct the forecasted distribution's tails. However, since non-parametric approaches do not properly estimate extreme quantiles due to data sparsity, the performance of EVT estimators may be compromised. In this section and to overcome this gap, we propose to apply EVT estimator to historical data directly. The selection of the relevant historical data is guided by the non-parametric model.

Our proposal consists of the following steps, also depicted in Fig. 3:

- **S1 Non-parametric estimation:** A non-parametric model $Q(\tau|\mathbf{x})$ is estimated for intermediate quantiles, $\tau \in \boldsymbol{\tau} = \{.05, .10, \ldots, .95\}$, i.e. 19 models are estimated. A rearrangement is also performed as described in (6). For a given training observation i, $(\mathbf{x}_i^{\text{tr}}, y_i^{\text{tr}})$, there is an estimation $\hat{q}_i^{\text{tr}}(\tau) = Q(\tau|\mathbf{x}_i^{\text{tr}})$.
- **S2 Non-parametric forecast:** Given a new observation $\mathbf{x}^*$, the estimation $\hat{q}^*(\tau)$ is given by the aforementioned non-parametric model Q for $\tau \in \boldsymbol{\tau}$.
- **S3 Historical similarity:** A similarity score $s(\mathbf{q}_1, \mathbf{q}_2)$ is computed between two quantile curves along several values of τ. The quantile curve $\hat{\mathbf{q}}^*$ from the new sample $\hat{\mathbf{q}}^* = [\hat{q}^*(\tau) \,|\, \tau \in \boldsymbol{\tau}]$ is compared with the quantile curve of each historical observation i, $\hat{\mathbf{q}}_i^{\text{tr}} = [\hat{q}_i^{\text{tr}}(\tau) \,|\, \tau \in \boldsymbol{\tau}]$. This similarity function is the Kolmogorov-Smirnov statistic given by

$$s(\mathbf{q}_1, \mathbf{q}_2) = \sup_{\tau} |\hat{\mathbf{q}}_1(\tau) - \hat{\mathbf{q}}_2(\tau)|. \tag{16}$$

  The new observation is scored against each historical observation, $s_i = s(\hat{\mathbf{q}}^*, \hat{\mathbf{q}}_i^{\text{tr}})$. Since both quantile curves $\hat{\mathbf{q}}^*$ and $\hat{\mathbf{q}}_i^{\text{tr}}$ are conditioned by the covariates, the selection of the similar periods through s_i is also conditioned by the covariates
- **S4 EVT data sample:** The EVT estimator is applied twice, for the lower-tail ($\tau < .05$) and the upper-tail ($\tau > .95$) quantiles. The historical values of y_i, used as the fitting sample of the EVT estimator, are selected as those corresponding to the top-h (hyperparameter) values of $s_i = s(\hat{\mathbf{q}}^*, \hat{\mathbf{q}}_i^{\text{tr}})$. To avoid quantile crossing, these values

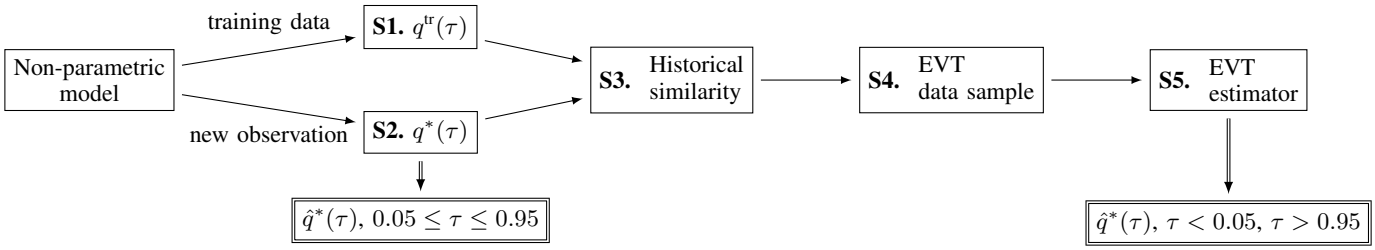

Fig. 3. Overview of the proposed forecasting model.

are further narrowed down to $y_i \leq \hat{q}^*(.05)$ and $y_i \geq \hat{q}^*(.95)$, respectively.

Furthermore, EVT requires that the sample encompasses the entire quantile curve, therefore the remaining 90% quantiles, which correspond to $\frac{0.9h}{0.05}$ observations, are sampled from a spline interpolation constructed from the discrete $\hat{\mathbf{q}}^*$ curve. The ensuing sample is called $\mathbf{y}'$.

S5 EVT estimation: Lower-tail and upper-tail quantiles are estimated through the estimator in (14), considering the sample $\mathbf{y}'$. Since, by convention, EVT distributions are defined for quantiles close to 1, the estimation of the lower-tail is obtained by considering the sample $y_i'' = C - y_i'$. EVT estimation is performed by (14) so that forecasted values are non-negative and below the installed capacity, $0 \leq \hat{y} \leq C$.

Note that step **S3** chooses i by comparing the probability distribution $\hat{\mathbf{q}}$ of the target variable conditioned on $\mathbf{x}^*$ and $\mathbf{x}_i^{\text{tr}}$. This is different from the usual approach of choosing i by comparing $\mathbf{x}^*$ against $\mathbf{x}_i^{\text{tr}}$ directly, as in [17], which assumes that covariates have equal weight and does not take the target variable into consideration. For instance, covariate j may be uncorrelated with the target, i.e. $\text{corr}((\mathbf{x}^{\text{tr}})_j, y^{\text{tr}}) = 0$, yet it contributes to the similarity through Euclidean distance as $((\mathbf{x}_i^{\text{tr}})_j - (\mathbf{x}^*)_j)^2$.

V. Numerical Experiments

A. Data Description

The proposed method is tested with a wind power dataset from the *Sotavento* wind power plant, located in Galicia (Spain), with a total installed capacity of 17.56 MW. The dataset extends from January 1st, 2014 to September 22nd, 2016, with hourly time steps.

The NWP data was retrieved from the MeteoGalicia THREDDS server, which is a publicly available service that provides historical and daily forecasts of several weather variables. The NWP is run at 0h UTC and the time horizon is 96 hours-ahead, meaning that for each day a set of four forecasts are available for each point of the grid (one generated in the current day at 0h UTC plus three generated on the previous days). The NWP model provides forecasts for: (a) u $[m/s]$, azimuthal wind speed; (b) v $[m/s]$, meridional wind speed; (c) mod $[m/s]$, wind speed module; (d) dir $[0, 360]$, wind direction. Four model levels (0 to 3) are available, meaning a total of 16 variables in each grid point.

1) Covariates extracted from the NWP grid: The features created by the authors of [6], from a NWP grid with 13×13 equally distributed points (4 km), were used in this work and are described below. Our goal is to forecast the wind power for 24h-ahead and the majority of the covariates are constructed with the most recent NWP run.

Temporal information is represented by:

- Temporal variance for the mod variable (level 3) at the central point of the grid, computed as

$$\sigma_{\text{time}}(t) = \sqrt{\frac{\sum_{i=t+k-N_h/2}^{t+k+N_h/2}(x(i) - \overline{x})^2}{N_h - 1}}, \tag{17}$$

with $N_h = 7$.

- *Lags* and *leads*, $x(t \pm z)$, for mod and dir (level 3) at the central point of the grid, $z = 1, 2, 3$.
- Four predictions generated for mod (level 3) at the central point of the grid, for time t.

The spatial information is represented through:

- Principal Component Analysis (PCA), the best model in [6] included the PCA applied to mod and dir (levels 1, 2, 3), and to u and v (level 3) with a 95% variance threshold.
- Spatial standard deviation for mod, u and v at level 3, computed as

$$\sigma_{\text{spatial}}(t) = \sqrt{\frac{\sum_{i=1}^{N_p}(x_i(t) - \overline{x}(t))^2}{N_p - 1}}, \tag{18}$$

where N_p is the number of geographical points in the NWP grid, $x_i(t)$ is the value of variable x at time t and location i, and $\overline{x}(t)$ is the mean of variable i for all locations.

- Spatial mean computed with the grid values of mod, u and v at model levels 1, 2, 3.

2) Data Division: A sliding-window approach was used for training the models. Table I presents the four distinct test folds. Each train and test set consists of 12 and 5 months, respectively, allowing an evaluation under different conditions.

TABLE I
Time Period for Training and Testing Folds.

Fold	Train set range	Test set range
1	01/01/2014—31/12/2014	01/01/2015—31/05/2015
2	01/06/2014—31/05/2015	01/06/2015—31/10/2015
3	01/11/2015—30/10/2016	01/11/2015—31/03/2016
4	01/04/2015—31/03/2016	01/04/2016—22/09/2016

B. Evaluation Metrics

This section describes the set of metrics adopted in this work to evaluate probabilistic forecasting skill of extreme quantiles.

1) Calibration: Measures the mismatch between the empirical probabilities (or long-run quantile proportions) and nominal (or subjective) probabilities, e.g. a .25 quantile should contain 25% of the observed values lower or equal to its value. For each quantile τ, the observed proportion $\hat{\alpha}(\tau)$ of observations bellow the estimated quantile is

$$\hat{\alpha}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{y_i \leq \hat{Q}_y(\tau|\mathbf{x}_i)}. \tag{19}$$

2) Sharpness: Measures the "degree of uncertainty" of the probabilistic forecast, which numerically corresponds to compute the average interval size between two symmetric quantiles, e.g., .10 and .90 centered in the .50 quantile (median), as follows

$$\text{sharp}_Y(\tau) = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}_Y(1 - \tau|\mathbf{x}_i) - \hat{Q}_Y(\tau|\mathbf{x}_i), \tag{20}$$

for $\tau \in [0, 0.5]$.

TABLE II
EVALUATED FORECASTING MODELS.

Notation	Description
GBT	GBT (non-parametric model)
local_tGPD	Hill estimator and truncated GPD (Eq. (14))*
Exp_Tails	Exponential functions (Eq. (7))
QR_EVT	QR combined with Hill estimator (Eq. (8))**
QR_EVT_T	QR, Hill estimator and transformed power data (Eq. (10))**
GBT_EVT	GBT combined with Hill estimator (Eq. (8))**
GBT_tGPD	Proposed method combining GBT with truncated GPD

* applied to $b\%$ of training samples ranked by similarity
** EVT estimator used in post-processing stage

3) Continuous Ranked Probability Score (CRPS): Evaluates the forecasting skill of a probabilistic forecast in terms of the entire predictive CDF, using an omnibus scoring function that simultaneously addresses calibration and sharpness [25]. Let y be the observation, and F_Y the CDF associated with an empirical probabilistic forecast,

$$\text{CRPS}(F_Y, y) = \int_{-\infty}^{\infty} \Big(F_Y(z) - H(z-y)\Big)^2 dz, \qquad (21)$$

where H is the Heaviside function.

Although CRPS is very popular in evaluating the quality of CDF forecast, recent work in [26] concluded that the mean of the CRPS is unable to discriminate forecasts with different tails behavior since it tends to benefit distributions with smaller uncertainty intervals, even if the calibration is poor. A more suitable scoring rule, following the suggestion in [25], is the *pinball function* or quantile loss. Smaller the value of the quantile score, better the model when forecasting quantile τ.

C. Implementation Details and Baseline Models

In order to evaluate the added-value of the proposed method, the models described in Table II are compared. The implementation is performed through R and Python programming languages. The R-packages include `quantreg` [27] (for quantile regression), `Rearrangement` [28] (for quantile crossing problem) and `ReIns` [29] (for truncated GPD estimation). The GBT model was implemented in Python using the `scikit-learn` library [30].

The hyperparameters of the GBT models were estimated using the Bayesian optimization algorithm from a Python implementation [31]. A 12-fold cross-validation was employed and, since all training sets contemplate one year of data, 12-folds guarantees 12 different monthly validation scenarios. For the final evaluation, the average of monthly CRPS is considered for each training set in the optimization process.

The local_tGPD benchmark is a naive model. The EVT estimator in (14) is applied to a $b\%$ of training samples listed in ascending order according to the Euclidean distance between $\mathbf{x}_i^{\text{tr}}$ and $\mathbf{x}^*$. The hyperparameter b was determined through grid-search from 5% to 50%, with increments of 5% and set to 15%. This model is used to assess if the mapping between covariates (e.g., weather forecasts) and target variable is important (as discussed in section IV).

Finally, the estimators in (8) and (14), used in QR_EVT and GBT_tGPD respectively, require the selection of the value of k for each time step. We followed the heuristic approach for choosing the first stable part of the plot of γ versus k. The stable part is found by computing a moving average on the differences of γ. In our approach, we selected the hyperparameter h by grid-search from 50 to 500, with increments of 50, with $h = 200$ being the best.

TABLE III
RELATIVE QUANTILE LOSS IMPROVEMENT [%] OVER THE BASELINE MODELS, CONSIDERING THE EXTREME QUANTILES $\boldsymbol{\tau}_e$.

Folds	Fold 1	Fold 2	Fold 3	Fold 4	W.Avg.
GBT	5.40	1.97	7.03	0.12	3.76
local_tGPD	22.27	29.34	21.71	27.80	26.25
Exp_Tails	12.87	11.03	9.44	14.79	12.55
QR_EVT	10.16	7.10	4.56	8.90	8.21
QR_EVT_T	12.39	7.20	10.78	8.55	10.39
GBT_EVT	12.20	9.06	9.33	5.03	9.75

TABLE IV
QUANTILE LOSS FOR EACH MODEL (LOWER IS BETTER).

$\boldsymbol{\tau}$	.001	.005	.01	.99	.995	.999
GBT	3.20	15.49	29.60	52.65	30.98	10.60
local_tGPD	3.16	15.74	31.05	84.52	45.21	9.69
Exp_Tails	8.63	20.95	32.47	53.14	32.26	9.43
QR_EVT	3.14	15.64	29.67	54.90	32.17	8.89
QR_EVT_T	3.19	15.55	29.84	59.27	34.48	9.68
GBT_EVT	3.17	15.72	31.97	67.13	35.23	8.45
GBT_tGPD†	**3.13**	**15.28**	**29.30**	**50.35**	**28.23**	**8.01**

†the proposed method

TABLE V
DESCRIPTIVE STATISTICS FOR THE OBSERVED WIND POWER (% OF INSTALLED CAPACITY).

Fold	Fold1		Fold 2		Fold 3		Fold 4	
	Train	Test	Train	Test	Train	Test	Train	Test
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q(.25)	1.4	2.8	0.8	0.4	1.2	5.7	2.3	0.1
Q(.5)	11.0	12.5	9.0	6.7	9.4	19.6	11.6	5.3
Q(.75)	32.1	33.6	26.6	23.0	28.6	41.3	31.6	16.5
Max	93.2	90.3	90.3	87.1	90.3	88.3	88.3	77.3

D. Forecasting Skill Evaluation

Since the GBT model performs better for power data, due to the nonlinear relationship between wind and power, GBT is used to estimate quantiles between .05 and .95. The proposed model is then used to estimate the quantiles $\boldsymbol{\tau}_e = \{.001, .005, .01, .02, .03, .04, .96, .97, .98, .99, .995, .999\}$.

Table III summarizes the relative quantile score improvement obtained by GBT_tGPD over the baseline models. Quantile score is computed by considering the extreme quantiles for nominal proportions $\boldsymbol{\tau}_e$.

The GBT_tGPD improvement is greater than 3.5% for all testing folds, except over GBT. Table IV shows a finer-grained view of the quantile loss for the most extreme quantiles, averaged over the testing folds. It can be noticed that the improvement of the proposed method is slightly higher for the upper quantiles, but, all in all, the proposed method shows the best results.

The statistics of the wind power generation for the train and test periods are summarized in Table V. Two factors might justify the different improvements obtained in the four folds:

the variability of the wind power values and the differences between train and test data distributions. When high variability is associated with different distributions for train and test sets, as is the case of fold 3, the selection of 200 observations results on more dispersed power measurements and, consequently, the EVT estimator has longer tails.

Fig. 4 complements the previous analysis by showing the calibration values for each model. The numerical values of the calibration deviation are also presented in Table VI. For the upper tail, the GBT_tGPD model exhibits almost perfect calibration for all quantiles. In the lower tail, it produces a lower overestimation of the quantiles. However, when considering all quantiles, QR-based models are the most well calibrated models. Yet, when analysing the sharpness of the forecast intervals generated by these methods in Fig. 5, these methods show that the better calibration comes at the cost of a higher amplitude (i.e., lower sharpness), which is a trade-off well-known in the forecasting literature. The lower sharpness from GBT_EVT, QR_EVT_T and QR_EVT is justified by the fact that the Hill estimator is more suitable for heavy-tailed distributions.

For illustrative purposes, the most extreme forecasted quantiles (i.e., .001 and .999) obtained with GBT, Exp_Tails and GBT_tGPD are depicted in Fig. 6. The Exp_Tails model was chosen since it is the model with the lowest sharpness. This plot clearly shows that GBT_tGPD has a better calibration than Exp_Tails, but wider intervals, and also shows a higher temporal variability of the forecast generated by GBT_tGPD.

The baseline model GBT shows small sharpness for all nominal coverage rates (between 92% and 99%) except the most extreme one (99.8%), as depicted in Fig. 5. The small sharpness is explained by the fact that GBT fails to capture the variability for the most extreme quantiles. The forecast of the lower quantiles is particularly bad with values very close to zero, as depicted in Fig. 6.

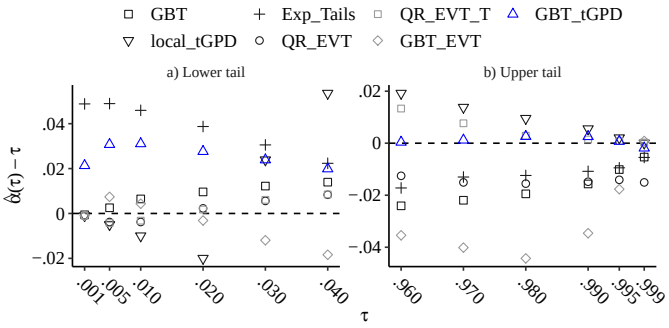

Fig. 4. Deviation between nominal and empirical quantiles for all folds. Dashed black line represents perfect calibration.

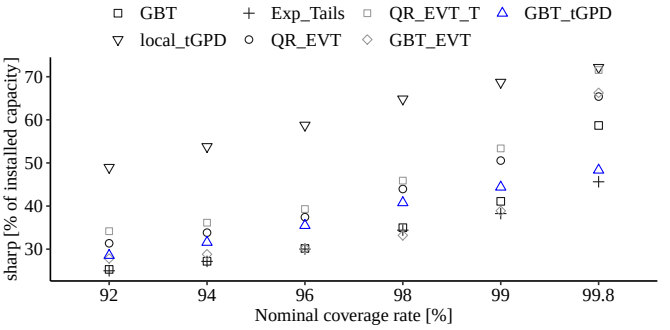

Fig. 5. Sharpness results for all folds.

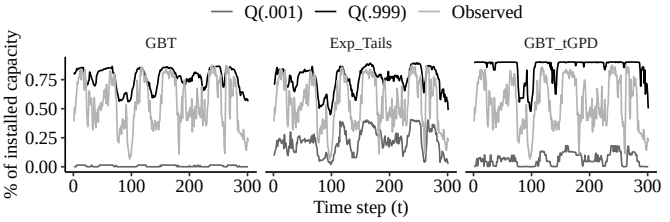

Fig. 6. Illustrative forecast of extreme quantile for GBT, Exp_Tails and GBT_tGPD.

VI. Conclusions

Accurate forecasting of distribution tails remains a challenge in the RES forecasting literature since are often associated with data sparsity. Furthermore, information from the tails is of major importance in power system operation (e.g., reserve capacity setting, dynamic line rating) and RES market trading. For this reason, concepts were borrowed from EVT for truncated variables and combined with a non-parametric wind power forecasting framework that includes features created from spatial-temporal information.

Two major benefits are provided by this work: (a) covariates are used to produce conditional forecasts of quantiles without any limitation in the number of variables; (b) the parametric EVT-based estimator can be combined with any non-parametric model (artificial neural networks, GBT, random forests, etc.) without any major modification. Moreover, the results for a wind farm located in Galicia, Spain, show that the proposed method can provide sharp and calibrated forecasts (important to avoid over- and under-estimation of risk) and outperforms state-of-the-art methods in terms of the quantile score. Finally, the proposed method can be transposed to other use cases in the energy sector, such as risk management in portfolio's future returns and study grid resilience to adverse weather events.

Topics for future work are: (i) inclusion of information from weather ensembles, as additional covariates, in order to exploit its capability to capture extreme events with a physically-based approach; (ii) generalization of the proposed method to other energy-related time series, e.g., solar power and electricity price; (iii) new proper scoring rules are needed to evaluate the forecasting skill of extreme (rare) events (see [32] for instance).

References

[1] R. Bessa, C. Möhrlen, V. Fundel, M. Siefert, J. Browell, S. Gaidi, B.-M. Hodge, U. Cali, and G. Kariniotakis, "Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry," *Energies*, vol. 10, no. 9, p. 1402, 2017.

[2] C. Sweeney, R. Bessa, J. Browell, and P. Pinson, "The future of forecasting for renewable energy," *Wiley Interdisciplinary Reviews: Energy and Environment*, In Press, 2019.

TABLE VI
DEVIATION BETWEEN NOMINAL (τ) AND EMPIRICAL ($\hat{\alpha}(\tau)$) QUANTILES.

τ	GBT	local_tGPD	Exp_Tails	QR_EVT	QR_EVT_T	GBT_EVT	GBT_tGPD
.001	-.001	-.001	.049	-.001	-.001	-.001	.021
.005	.003	-.005	.049	-.004	-.004	.007	.031
.01	.006	-.010	.046	-.004	-.003	.004	.031
.02	.010	-.020	.039	.002	.002	-.003	.028
.03	.012	.024	.031	.006	.006	-.012	.024
.04	.014	.054	.022	.008	.008	-.018	.020
.96	-.024	.019	-.017	-.013	.013	-.035	.000
.97	-.022	.014	-.013	-.015	.008	-.040	.001
.98	-.019	.010	-.012	-.016	.003	-.044	.003
.99	-.016	.006	-.011	-.015	.001	-.035	.003
.995	-.010	.002	-.009	-.014	.001	-.018	.001
.999	-.005	.000	-.005	-.015	.000	.001	-.002

[3] R. Juban, H. Ohlsson, M. Maasoumy, L. Poirier, and J. Z. Kolter, "A multiple quantile regression approach to the wind, solar, and price tracks of GEFCom2014," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1094–1102, 2016.

[4] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 7, no. 1, pp. 47–54, 2004.

[5] R. Bessa, V. Miranda, A. Botterud, J. Wang, and E. M. Constantinescu, "Time adaptive conditional kernel density estimation for wind power forecasting," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 4, pp. 660–669, 2012.

[6] J. R. Andrade and R. J. Bessa, "Improving renewable energy forecasting with a grid of numerical weather predictions," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 4, pp. 1571–1580, 2017.

[7] P. Pinson, "Very short-term probabilistic forecasting of wind power with generalized logit-normal distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 61, no. 4, pp. 555–576, 2012.

[8] J. W. Messner, A. Zeileis, J. Broecker, and G. J. Mayr, "Probabilistic wind power forecasts with an inverse power curve transformation and censored regression," *Wind Energy*, vol. 17, no. 11, pp. 1753–1766, 2013.

[9] M. A. Matos and R. J. Bessa, "Setting the operating reserve using probabilistic wind power forecasts," *IEEE Transactions on Power Systems*, vol. 26, no. 2, pp. 594–603, 2010.

[10] A. Botterud, J. Wang, Z. Zhou, R. Bessa, H. Keko, J. Akilimali, and V. Miranda, "Wind power trading under uncertainty in LMP markets," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 894–903, 2012.

[11] R. Dupin, "Prévision du dynamic line rating et impact sur la gestion du systéme électrique," Ph.D. dissertation, MINES ParisTech, PSL Research University, Paris, France, Jul. 2018.

[12] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl, "From probabilistic forecasts to statistical scenarios of short-term wind power production," *Wind Energy*, vol. 12, no. 1, pp. 51–62, 2009.

[13] M. Cagnolari, "The value of the right distribution for the newsvendor problem and a bike-sharing problem," Ph.D. dissertation, University of Bergamo, May 2017.

[14] L. De Haan and A. Ferreira, *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[15] P. D. Andersen, "Optimal trading strategies for a wind-storage power system under market conditions," Master's thesis, Technical University of Denmark, Lyngby, Denmark, 2009.

[16] M. Matos, R. J. Bessa, C. Gonçalves, L. Cavalcante, V. Miranda, N. Machado, P. Marques, and F. Matos, "Setting the maximum import net transfer capacity under extreme res integration scenarios," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2016, pp. 1–7.

[17] J. Beirlant, T. D. Wet, and Y. Goegebeur, "Nonparametric estimation of extreme conditional quantiles," *Journal of statistical computation and simulation*, vol. 74, no. 8, pp. 567–580, 2004.

[18] H. J. Wang, D. Li, and X. He, "Estimation of high conditional quantiles for heavy-tailed distributions," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1453–1464, 2012.

[19] H. J. Wang and D. Li, "Estimation of extreme conditional quantiles through power transformation," *Journal of the American Statistical Association*, vol. 108, no. 503, pp. 1062–1074, 2013.

[20] C. Gilbert, J. Browell, and D. McMillan, "Leveraging turbine-level data for improved probabilistic wind power forecasting," *IEEE Transactions on Sustainable Energy*, In Press, 2019.

[21] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[22] V. Chernozhukov, I. Fernández-Val, and A. Galichon, "Quantile and probability curves without crossing," *Econometrica*, vol. 78, no. 3, pp. 1093–1125, 2010.

[23] J. Beirlant, I. F. Alves, T. Reynkens *et al.*, "Fitting tails affected by truncation," *Electronic Journal of Statistics*, vol. 11, no. 1, pp. 2026–2065, 2017.

[24] A. J. McNeil and T. Saladin, "The peaks over thresholds method for estimating high quantiles of loss distributions," in *Proceedings of 28th International ASTIN Colloquium*, 1997, pp. 23–43.

[25] P. Friederichs and T. L. Thorarinsdottir, "Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction," *Environmetrics*, vol. 23, no. 7, pp. 579–594, 2012.

[26] M. Taillardat, A.-L. Fougères, P. Naveau, and R. de Fondeville, "Extreme events evaluation using CRPS distributions," *arXiv preprint arXiv:1905.04022*, 2019.

[27] R. Koenker, *quantreg: Quantile Regression*, 2018, R package version 5.38. [Online]. Available: https://CRAN.R-project.org/package=quantreg

[28] W. Graybill, M. Chen, V. Chernozhukov, I. Fernandez-Val, and A. Galichon, *Rearrangement: Monotonize Point and Interval Functional Estimates by Rearrangement*, 2016, R package version 2.1. [Online]. Available: https://CRAN.R-project.org/package=Rearrangement

[29] T. Reynkens and R. Verbelen, *ReIns: Functions from "Reinsurance: Actuarial and Statistical Aspects"*, 2018, R package version 1.0.8. [Online]. Available: https://CRAN.R-project.org/package=ReIns

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] F. Nogueira, "Python bayesian optimization implementation," http://github.com/fmfn/BayesianOptimization.

[32] S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, "Forecaster's dilemma: Extreme events and forecast evaluation," *Statistical Science*, vol. 32, no. 1, pp. 106–127, 2017.