Wind Power Probabilistic Forecast in the Reproducing Kernel Hilbert Space

Cristobal Gallego-Castillo
Alvaro Cuerva-Tejero
DAVE, Universidad Politécnica de Madrid
Madrid, Spain
{cristobaljose.gallego, alvaro.cuerva}@upm.es

Ricardo J. Bessa
Laura Cavalcante
INESC Technology and Science
Porto, Portugal
{ricardo.j.bessa, laura.l.cavalcante}@inesctec.pt

***Abstract*—Wind power probabilistic forecast is a key input in decision-making problems under risk, such as stochastic unit commitment, operating reserve setting and electricity market bidding. While the majority of the probabilistic forecasting methods are based on quantile regression, the associated limitations call for new approaches. This paper described a new quantile regression model based on the Reproducing Kernel Hilbert Space (RKHS) framework. In particular, two versions of the model, off-line and on-line, were implemented and tested for a real wind farm. Results showed the superiority of the on-line approach in terms of performance, robustness and computational cost. Additionally, it was observed that, in the presence of correlated data, the optimal on-line learning may cause unreliable modelling. Potential solutions to this effect are also described and implemented in the paper.**

***Index Terms*—On-line, probabilistic forecast, quantile regression, Reproducing Kernel Hilbert Space (RKHS), wind power**

I. Introduction

Presently, the increasing share of wind power in the generation portfolio of several control areas is demanding for a revision of the operational practices and management tools [1]. For instance, the Transmission System Operators (TSO) of Portugal and Spain studied alternative methods for setting the operating reserve requirements based on wind power uncertainty forecasts [2][3]. In fact, wind power uncertainty forecast is a vital input in decision-making problems under risk, such as stochastic unit commitment [4], operating reserve setting [5] and electricity market bidding [6].

The current wind power forecasting literature is rich in statistical and machine learning applied to the point (or deterministic) forecast problem. For probabilistic forecasting, the algorithms are mainly based in three classes of models [7][8]: (a) conditional kernel density estimation (KDE), (b) semi-parametric regression and (c) quantile regression. It is important to stress that other representations for the wind power uncertainty are also possible, such as ramp forecasting [9] and temporal trajectories (or short-term scenarios) [10].

Two examples of conditional KDE algorithms are: (a) time-adaptive quantile-copula estimator that produces density forecasts for the next hours using Numerical Weather Predictions (NWP) as inputs and explores the non-parametric copula for modelling the dependency between wind speed/direction and power [11]; (b) two-stage approach that, firstly, uses a vector autoregressive moving average-generalized autoregressive conditional heteroscedastic (VARMA-GARCH) model to capture wind speed and direction uncertainty forecast, secondly, employs conditional KDE to model the relationship between wind speed/direction and power (i.e., the power curve) [12].

One work about semi-parametric regression is presented in [13], which proposes the use of generalized logit-Normal distribution to enable a full characterization of the forecasted densities by their location and scale parameters. Dynamic models based on classical time series models (e.g., autoregressive model) are proposed for the location and scale parameters.

The majority of the methods based on quantile regression employed to model the non-linear relation between wind speed and power use two well-known techniques, local regression (or varying coefficients) [14] and additive models with splines [15]. The main limitation of local quantile regression is that the computational time increases significantly with the number of predictors and it is also prone to overfitting. The additive models require a correct choice of the splines for different types of variables (e.g., categorical, circular) and a hyperparameter is needed to each predictor variable.

This paper proposes a new quantile regression model based on kernel methods. Kernel methods are a class of algorithms oriented to pattern analysis that have been applied to a number of problems, involving classification, regression and time series forecasting (see [16] and references therein). The presented model implements quantile regression in the Reproducing Kernel Hilbert Space (RKHS) according to the framework described in [17]. In this framework, the data from the input space is transformed to the feature space using a kernel matrix. In other words, this means transforming a non-linear space into a high dimensional linear space where the

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013, and the DAVE – Departamento de Aeronaves y Vehculos Espaciales (UPM).

classical linear quantile regression technique can be applied. This paper presents two original contributions:

- First application of the linear quantile regression in the RKHS to the wind power probabilistic forecasting problem. Furthermore, it establishes a connection between quantile regression and recent research in kernel learning theory [18];
- Propose an on-line version based on the stochastic gradient descent method (inspired by [19]), in contrast to the off-line version based on solving a quadratic optimization problem with the interior-point method.

Optimal model parameters are obtained through k-fold cross-validation. The role and meaning of the model parameters is studied for one real wind farm. Furthermore, the on-line and off-line approaches are also compared.

The remaining of the paper is organized as follows: section II provides a description of the quantile regression models in the RKHS, for both off and on-line standpoints. Section III describes the employed data and the setup of the experiment. The obtained results are presented and discussed in section IV. Finally, the paper ends with concluding remarks in section V.

II. Description of the models

Let consider a number of observations in the form $(\mathbf{x}_t, p_t) \in \mathbb{R}^n \times \mathbb{R}^+$, where p_t is the wind power output of a wind farm or portfolio at time t, and $\mathbf{x}_t$ is a vector with n explanatory variables. In order to obtain a probabilistic forecast of p_t (in the form of a set of quantiles), a number of quantile regression models can be implemented. A quantile regression model establishes a functional relationship between $\mathbf{x}$ and the τ-th quantile of the wind power output, denoted by q^τ, where $\tau \in [0, 1]$. From the definition of quantile, it holds that $P(p \leq q^\tau) = \tau$. Without loss of generality, a quantile regression model can be written as follows:

$$q^\tau(\mathbf{x}) = f(\mathbf{x}) + b, \tag{1}$$

where b is a bias term and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a non-linear function.

In the following, it is assumed that $f(\mathbf{x})$ is linear in a certain feature space given by the feature map $\varphi : \mathbb{R}^n \to \mathcal{F}$, that is:

$$f(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle. \tag{2}$$

It is also assumed that f belongs to $\mathcal{H}$, a Reproducing Kernel Hilbert Space (RKHS) defined by the reproducing kernel (also referred to as kernel matrix) $k(\mathbf{x}_1, \mathbf{x}_2) = \langle \varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2) \rangle$. From that, the so-called reproducing property holds [18]:

$$\langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = f(\mathbf{x}). \tag{3}$$

In order to set the model parameters, $\mathbf{w}$ and b, the following regularised risk functional, evaluated over N samples, is to be minimised:

$$R_{1:N} := \frac{1}{N} \sum_{t=1}^{N} l_\tau(p_t, q^\tau(\mathbf{x}_t)) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}}, \tag{4}$$

where, according to quantile regression theory [17], $l_\tau(p_t, q^\tau(\mathbf{x}_t))$ is the pinball function, given by:

$$l^\tau(p_t, q^\tau(\mathbf{x}_t)) = \begin{cases} \tau \cdot (p_t - q^\tau(\mathbf{x}_t)) & \text{if } p_t \geq q^\tau(\mathbf{x}_t) \\ (\tau - 1) \cdot (p_t - q^\tau(\mathbf{x}_t)) & \text{if } p_t < q^\tau(\mathbf{x}_t) \end{cases}, \tag{5}$$

and $\|\cdot\|^2_{\mathcal{H}}$ is the norm in the RKHS, which is employed to measure the complexity of the function f. The so-called regularization parameter, λ, provides a means to handle the balance between the committed forecast error (first term in the right side of (4) and the function complexity (second term), that is, the balance between bias and variance when estimating $q^\tau(\mathbf{x})$.

Finally, it is noted that the reproducing kernel, $k(\cdot, \cdot)$, must meet certain conditions to be considered an admissible kernel. A number of admissible kernels are provided in [20]. From here on, we assume $k(\cdot, \cdot)$ to be the radial basis function kernel, given by:

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\sigma \|\mathbf{x}_1 - \mathbf{x}_2\|^2\right), \tag{6}$$

where σ is a parameter related to the kernel width. According to [21], this kernel is a general purpose kernel well-suited for situations in which no prior knowledge about the data is available. Note that, with the definition provided in (6), small σ values lead to large kernel widths. It is also remarked that the optimal σ value needs to be properly assessed during the modelling stage, as it has a direct impact on the aforementioned bias/variance balance.

In the following sections, the presented quantile regression model is particularised for two different learning strategies: the off-line and the on-line learning.

A. Off-line model

Under the off-line learning standpoint, the model for the τ-th quantile, $q^\tau_{off}(\mathbf{x})$, is obtained from the minimisation of (4) for a given training set with N_0 samples $(\mathbf{x}_t, p_t)$, with $1 \leq t \leq N_0$. After this process, the model remains fixed and is solely employed to generate forecasts.

From the Support Vector Machine literature (see [20], among others), the problem at hand can be manipulated so that the model can be written in the form of a kernel expansion:

$$q^\tau_{off}(\mathbf{x}) = \sum_{i=1}^{N_0} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b, \tag{7}$$

where the $N_0 + 1$ model coefficients, $\{b, \alpha_1, \alpha_2, ..., \alpha_{N_0}\}$, are obtained by solving the dual optimisation problem of (4). To this end, the interior point algorithm available in the R package *kernlab* [21] will be employed. The power of (7) is that, once

the model is determined, generating a forecast merely requires N_0 evaluations of $k(\mathbf{x}_i, \mathbf{x})$, whose computational cost increases fairly little with the dimension of $\mathbf{x}$.

Taking all that into consideration, given a training set with N_0 samples and specific values for the parameters σ and λ, the off-line model $q^{\tau}_{off}(\mathbf{x})$ can be determined. In order to illustrate the impact of these parameters on the aforementioned bias/variance balance, Fig. 1 has been performed. This figure shows the obtained quantiles $q^{\tau=0.2}_{off}$ and $q^{\tau=0.8}_{off}$ conditioned to the forecast wind speed (see section III for details on the employed data), with parameter values favouring either the bias (small σ and large λ values, blue lines) or the variance (large σ and small λ values, orange lines). Since the underlying function between wind speed and power is the well-known power curve, the figure clearly reveals the need for appropriate σ and λ values.

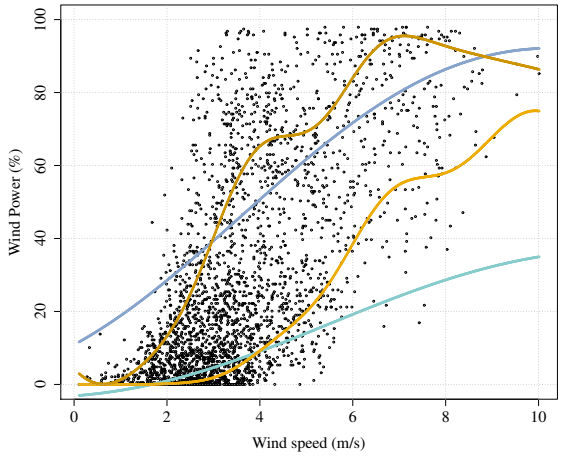

Figure 1. Conditional quantiles $\tau = 0.2$ and $\tau = 0.8$ of the wind power reflecting two situations: high bias (blue lines) and high variance (orange lines).

B. On-line model

On-line learning is an incremental process in which the model integrates information as new observations are available. One of the main advantages of this strategy in the case of wind power forecasting is that the model is able to account for smooth variations of the underlying dynamics of the wind power output across time, which are likely to happen because of meteorological seasonalities and the wind turbine aging. Consequently, the quantile model evolves over time, from an initial arbitrary state, let say $q^{\tau}_{on,1}(\mathbf{x}) = b_0$, to a certain state at time instant t, described by:

$$q^{\tau}_{on,t}(\mathbf{x}) = \sum_{i=1}^{t-1} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b_{t-1}, \tag{8}$$

according to data delivering and a set of predefined learning rules. These learning rules are to be obtained by applying the stochastic gradient descent in Hilbert Space [19]. Stochastic gradient descent means that only the most recent error is included in the cost function, which collapses into $R_{t:t}$ (see (4)). Since the minimisation occurs in the RKHS, the gradient is computed with respect to the function q^{τ}_{on}. Mathematically:

$$q^{\tau}_{on,t+1} = q^{\tau}_{on,t} - \eta \left. \frac{\partial R_{t:t}}{\partial q^{\tau}_{on}} \right|_{q^{\tau}_{on}=q^{\tau}_{on,t}}, \tag{9}$$

where η is the learning rate, which is assumed to be constant in this work.

From (9), and making use of (3)–(5), and (8), the following rules for updating the model are obtained:

$$\alpha_t := \begin{cases} \eta\tau & \text{if } p_t > q^{\tau}_{on,t}(\mathbf{x}_t) \\ \eta(\tau - 1) & \text{if } p_t < q^{\tau}_{on,t}(\mathbf{x}_t) \\ 0 & \text{if } p_t = q^{\tau}_{on,t}(\mathbf{x}_t) \end{cases}, \tag{10}$$

$$\alpha_i := (1 - \eta\lambda)\alpha_i \quad \text{for } i < t, \tag{11}$$

$$b_t := b_{t-1}. \tag{12}$$

According to the previous description, given a stream of samples $(\mathbf{x}_t, p_t)$ and the aforementioned learning rules, there are three parameters to assess in order to have the on-line quantile model completely defined. These are η (the learning rate), σ (the kernel parameter) and λ. We note that, while the original role of λ was tuning the bias/variance balance (see (4)), in the on-line version this parameter only affects the *memory* of the model, that is, the rate at which a certain α_{t_0} tends to zero for $t > t_0$. Actually, according to (11), this effect is given by the product $\eta\lambda$, which can be considered as a forgetting factor. If $\eta\lambda = 0$ the model does not forget the information captured from any past samples, regardless how old the sample is. If $\eta\lambda$ is too large, the model forgets too quickly, making it impossible to capture other relationships different from $q^{\tau}_{on,t}(\mathbf{x}) = b_0$.

Another issue has to do with the expansion length. According to (8), the number of terms in the expansion grows linearly with t, increasing the computational memory requirements. To avoid this, [19] proposes expansion truncation by dropping the oldest samples, given that the α_i coefficients decrease with time as $(1 - \eta\lambda)^t$. For example, considering a forgetting factor $\eta\lambda = 10^{-5}$, a certain coefficient α_i falls by 99.5% after 530.000 time steps. Another option is to explore the extent to which the contribution of certain samples to the model can be approximated by linear combinations of the contribution of another samples, so that not every sample must translate into a new term in (8). This idea is the base of the method proposed in [22], where a discussion on several sparsification methods can be found.

The choice for b_0 has relatively low impact on the model performance as long as the learning process occurs successfully within the span of $\mathbf{x}$ before test. However, specific choices

for b_0 may entail additional advantages in specific situations. In this work, we propose:

$$b_0 = Q^{\tau}_{train}, \tag{13}$$

where Q^{τ}_{train} is the (non-conditioned) quantile τ of the wind power time series during a training set. The rational for this choice is that, in case of a long sequence of missing data, the forgetting process makes the quantile estimates tend to the non-conditioned quantile, which actually represents a classical reference model in wind power forecasting (referred to as *climatology*). Hence, in such situations, the model would still perform as a reference. This property also applies to regions of the span of $\mathbf{x}$ where data are observed with low frequency (rare or extreme events).

For illustrative purposes, Fig. 2 shows the on-line model of the median, $q^{\tau=0.5}_{on,t}(\mathbf{x})$, obtained at different time instants.

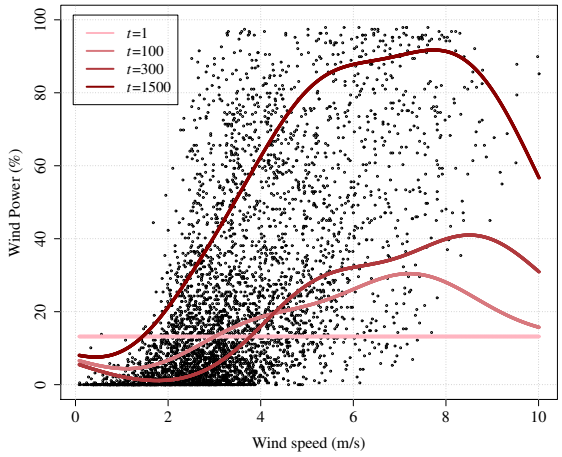

Figure 2. Quantile $\tau = 0.5$ of the expected wind power conditioned to the forecast wind speed for different time instants (simulation for $\eta = 0.05$, $\lambda = 1.61e-03$ and $\sigma = 1$).

III. Experiment setup

The case-study consists in one real wind farm from the Global Energy Forecasting Competition (GEFCOM 2012 - first wind farm) dataset, which is freely available in [23]. Three years of data are available and consists of historical power measurements and weather predictions extracted from the European Centre for Medium-range Weather Forecasts model (ECMWF) with hourly time resolution.

In this study, a number of time series covering a six-month period were considered. The involved variables are:

- wind power generation $\{p_t\}$
- wind speed forecast, $\{\widehat{ws}_t\}$
- wind direction forecast, $\{\widehat{wd}_t\}$
- time of day, $\{h_t\}$

The wind power is referred to the nominal power, so that the power records belong to the interval [0, 1]. The rest of the variables were standardized (zero mean and unit variance) in order to put all predictors on a common scale. Standardizing is a common practice in forecasting as it helps remove the impact of the variable scale on the regression process. In view of (6), this step is deemed to be of particular importance, since different predictor scales may hamper the optimisation of the parameter σ. It is also noted that the two last variables are circular variables.

Six models were implemented according to the learning strategy (off/on-line) and the regression variables contained in $\mathbf{x}_t$ (see Table I for details and nomenclature). Each model actually comprises 19 quantile regression submodels, with $\tau = \{0.05, 0.10, ..., 0.95\}$.

Because forecasts of wind speed and direction are provided each 12 hours in the GEFCOM dataset, the time lag between the availability of the forecast $\mathbf{x}_t$ and the observation p_t varies from one hour to 12 hours, according to the time of the day. This means that, for the case of the off-line model, the prediction horizon follows a similar scheme (i.e. the prediction horizon is 1 hour for daytime 13:00, 2 hours for daytime 14:00, and so on up to 12 hours for daytime 00:00). However, for the on-line model, even when the same reasoning for the lag between $\mathbf{x}_t$ and p_t holds, it should be noted that the expansion given in (8) entails that quantile forecasts for time t were generated by a model that has already assimilated (learned from) the sample $(\mathbf{x}_{t-1}, p_{t-1})$, making the prediction horizon equal to one hour for every daytime because the previous power observation must be available. Taking all that into account, the described benchmark exercise should be considered for a prediction horizon of one hour ahead.

TABLE I. Regression variables of the different models

Model	$\widehat{ws}_t$	$\sin(\widehat{wd}_t)$	$\cos(\widehat{wd}_t)$	$\sin(h_t)$	$\cos(h_t)$
$M^{(1)}_{off}$	x				
$M^{(2)}_{off}$	x	x	x		
$M^{(3)}_{off}$	x	x	x	x	x
$M^{(1)}_{on}$	x				
$M^{(2)}_{on}$	x	x	x		
$M^{(3)}_{on}$	x	x	x	x	x

Concerning the model parameters, ten values for σ and λ were considered, giving 100 potential configurations for each of the three off-line models. For the case of the on-line models, this number rose up to 500, since five learning rates were also considered. The considered parameter values are:

- $\sigma \in [1 \cdot 10^{-2}, ..., 1 \cdot 10^{1}]$, in logarithmic scale
- $\lambda \in [3.47 \cdot 10^{-2}, ..., 3.47 \cdot 10^{-5}]$, in logarithmic scale
- $\eta \in [0.01, 0.05, 0.1, 0.5, 1]$

The optimal values were picked up through k-fold cross-validation with three folds. The performance of the models was evaluated in terms of Continuous Rank Probability Score (CRPS). The CRPS provides an average performance of how well probabilistic forecasts compares with observations. This criterion associates lower values to better performances, zero being the best mark possible. The CRPS is given by:

$$CRPS = \frac{1}{N}\sum_{t=1}^{N}\int_{-\infty}^{\infty}\left(F_t(p) - H_{p_t}(p)\right)^2 \, dp \qquad (14)$$

where $F_t(p)$ is the predicted cumulative distribution function for time t, $H_{p_t}(p)$ is the Heaviside function located at the observation p_t, and N is the number of evaluated forecasts. For the case of quantile regression models, the CRPS can be estimated from a set of quantiles using the loss function from (5) [24].

IV. Results and Discussion

In this section, several results obtained from the experiment described in Section III are provided and discussed.

A first set of results has to do with some aspects of the training process of the models, specifically the computational time[1] and the robustness of the employed algorithms. Table II shows the averaged computational time in hours, t_c, employed to train a single configuration of the model (i.e., 19 quantile submodels for specific values of the model parameters). It is reminded that there are 100 configurations for each off-line model (ten values for λ by ten values for σ) and 500 for each on-line model (100 by five learning rates). Interestingly, on-line models required computational times two orders of magnitude lower than off-line models, which represents a clear advantage of the former. We note in passing that the computational time (specially for the off-line approach) was found to grow notably with the time length of the dataset, this being the reason why the experiment was limited to a six month period.

For the case of the off-line models, the optimization through interior-point method resulted unsuccessful in some cases. This fact could be attributed to singularities when combining specific model configurations with the available data that may cause ill-conditioning that often prevents the standard Cholesky factorization from getting an acceptable approximate solution. The percentage of theses cases is reflected by *Break* in Table II. While a lack of robustness was found for $\mathrm{M}_{off}^{(1)}$, an important decrease of *Break* was observed when the problem was unfolded over higher dimensions (higher number of predictors). On the other hand, the simplicity of the mathematical formulation behind the on-line model prevents such type of problems, resulting in higher robustness in comparative terms.

TABLE II. Computational time and breaks observed during the training process.

	Off-line		On-line	
	t_c (h)	Break (%)	t_c (h)	Break (%)
$\mathrm{M}_{on/off}^{(1)}$	4.65	63%	0.05	-
$\mathrm{M}_{on/off}^{(2)}$	3.70	7%	0.05	-
$\mathrm{M}_{on/off}^{(3)}$	3.35	1%	0.06	-

[1] Results for an Intel Core i7-2600 CPU 3.40 GHz with 8GB of RAM.

Next, results concerning the forecasting performance of the models are provided. Fig. 3 shows the results attained for the best configuration of the six models. Straight lines represent the CRPS obtained for the three off-line models, $\mathrm{M}_{off}^{(i)}$ for $1 \leq i \leq 3$), while curves were employed for the case of the on-line models, reflecting the dependency with the learning rate.

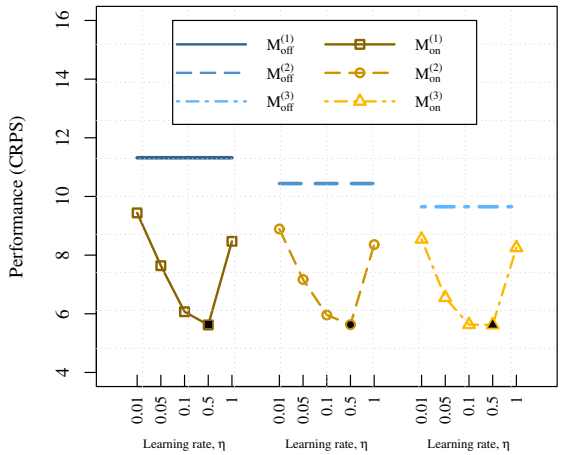

Figure 3. Performance of the off-line models (straight lines) and on-line models (dotted lines). For the on-line models, the filled dot represents the minimum CRPS.

Concerning the results obtained for the off-line models, it can be seen that adding more variables improved systematically the performance. It is also observed that the on-line models overcame the off-line models with no exception. In addition, the following remarks should be noted:

i The minimum CRPS values obtained for the three on-line models (that is, the three filled dots in Fig. 3) are very similar, reflecting that adding more information did not contribute to a better performance.

ii Increasing learning rates provided, regardless the on-line model considered, better results up to a certain point, in such a way that the optimal learning rate was found to be $\eta = 0.5$ in all the cases. Considering (10) and that the objective variable (normalised wind power, p_t) ranges between 0 and 1, such learning rate is likely to entail large modifications at each time step of the quantile models $q_{on,t}^{\tau}(\mathbf{x}_t)$ in the surroundings of $\mathbf{x}_t$ (the notion of surroundings is related to the kernel width).

iii The optimal values for σ were found to be very low (see Table III above), which translates into very wide kernel widths (see (6)).

Motivated by these facts, a detailed analysis on the models was performed. This analysis revealed that on-line models actually were not capturing reliable underlying relationships between inputs and outputs with smooth variations over time. Instead, the modelled relationships between inputs and quantile forecasts were found to be arbitrary functions experimenting strong fluctuations at each time step, according to the last forecast error committed. This behaviour could be expected from the observed large learning rates and large kernel widths.

The reason why this behaviour resulted in a better performance could be due to the fact that the models were able to exploit the error correlation. Error correlation means that the forecast error of the model at time t is likely to be very similar to the forecast error at time $t+1$. This effect derives from wind power correlation, which is a well-known phenomenon observed within a few hours in wind power dynamics [25]. In addition, the fact that the on-line learning algorithm was built on the basis of the stochastic gradient descent (i.e. the minimisation of the last forecast error committed) is likely to strengthen this effect. Taking all that into account, it is understandable that the obtained optimal parameters, that is, those providing the best performance, were those allowing such behaviour of the models.

To our knowledge, a fair comparison between off and on-line models cannot be performed unless (i) the off-line model is designed to capture power correlation as well (for instance, adding p_{t-1} as predictor), or (ii) the on-line model is prevented from exploiting error correlation. In view of the computational cost associated to off-line training, the second choice was adopted. With this purpose, the kernel expansion described in (8) was replaced by:

$$q^{\tau}_{on,t}(\mathbf{x}_t) = \sum_{i=1}^{t-12} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b_{t-12}, \tag{15}$$

It is important to note that equation (15) restricts the on-line learning process in the sense that, in order to generate the power forecast for time t, the model has assimilated samples only up to time $t-12$. However, it is important to mention that samples $(\mathbf{x}_t, p_t)$ employed to feed the new models are exactly the same than those employed to train the other models. In other words, placing conditions on the kernel expansion does not necessarily translates into a larger prediction horizon, which is given in this case by the time lag between the availability of forecast $\mathbf{x}_t$ and time t. Consequently, and following the reasoning given in section III, the benchmark prediction horizon remains at one hour.

Three new on-line models (referred to as $\mathrm{M}^{(\cdot)}_{on+12}$) were implemented according to (15). Now, reliable relationships between inputs and outputs evolving smoothly over time were obtained. Table III compares $\mathrm{M}^{(\cdot)}_{on}$ and $\mathrm{M}^{(\cdot)}_{on+12}$ models by showing optimal parameters and performance.

Concerning the model parameters, a remarkable reduction experimented in the optimal learning rate and kernel width can be observed. This is in good agreement with the aforementioned comments. In terms of model performance, results show that, once the error correlation effect was avoided, the model performance decreased notably, as it could be expected. Interestingly, the new on-line models still overcome the off-line models, as it can be observed from the last column, showing the improvement over the off-line model (IoOff). The IoOff is defined as follows:

TABLE III. COMPARISON BETWEEN MODELS $\mathrm{M}^{(\cdot)}_{on}$ AND $\mathrm{M}^{(\cdot)}_{on+12}$

	σ	λ	η	CRPS	IoOff
$\mathrm{M}^{(1)}_{on}$	0.0215	$3.4674 \cdot 10^{-3}$	0.5	5.62	50.35%
$\mathrm{M}^{(2)}_{on}$	0.0100	$7.4703 \cdot 10^{-3}$	0.5	5.63	46.07%
$\mathrm{M}^{(3)}_{on}$	0.0215	$7.4703 \cdot 10^{-3}$	0.5	5.62	41.76%
$\mathrm{M}^{(1)}_{on+12}$	0.2154	$1.6092 \cdot 10^{-4}$	0.01	10.22	9.71%
$\mathrm{M}^{(2)}_{on+12}$	0.2154	$3.4674 \cdot 10^{-5}$	0.01	9.58	8.24%
$\mathrm{M}^{(3)}_{on+12}$	0.2154	$7.4703 \cdot 10^{-4}$	0.05	8.84	8.40%

$$\mathrm{IoOff} = 100 \frac{\mathrm{CRPS}(\mathrm{M}^{(\cdot)}_{off}) - \mathrm{CRPS}(\mathrm{M}^{(\cdot)}_{on/on+12})}{\mathrm{CRPS}(\mathrm{M}^{(\cdot)}_{off})} \tag{16}$$

Results also show that, similar to what was observed for the off-line models, adding more predictors contributed to a better modelling (the CRPS evolved from 10.22 to 8.84), while the improvement with respect to the off-line model was slightly reduced (from 9.71% to 8.40%).

For illustrative purposes, Fig. 4 shows the probabilistic forecasts provided by the best off-line model, $\mathrm{M}^{(3)}_{off}$ (top), on-line model, $\mathrm{M}^{(3)}_{on}$ (middle) and on-line model with limited kernel expansion, $\mathrm{M}^{(3)}_{on+12}$ (bottom) during a one week period. It can be appraised how correlated data led to a bad modelling for the case of model $\mathrm{M}^{(3)}_{on}$, which shows important fluctuations in the obtained probabilistic forecasts, and how this effect was removed by acting on the expansion length (model $\mathrm{M}^{(3)}_{on+12}$).

V. CONCLUSIONS

This work presented the first application of the linear quantile regression in the Reproducing Kernel Hilbert Space (RKHS) to the wind power probabilistic forecast problem. It explores the RKHS framework for modelling the non-linear mapping between wind speed and power with a simple linear quantile regression model.

Two versions of the model (off-line and on-line) were implemented and tested for a real wind farm. Results showed the superiority of the on-line approach in terms of performance, robustness and computational cost. Additionally, quantile regression in RKHS was deemed suitable for problems with multi-dimensional explanatory variables, as the models provided better probabilistic forecasts when power quantiles were conditioned to all the available variables (wind speed forecast, wind direction forecast and time of the day).

An important remark has to do with the optimisation of the on-line learning process. It was observed that, in the presence of correlated data (as wind power generation data), the on-line learning strategy may tend to exploit this feature instead of capturing reliable underlying relationships between inputs and outputs. This followed as a consequence of combining the stochastic gradient descent algorithm together with correlated

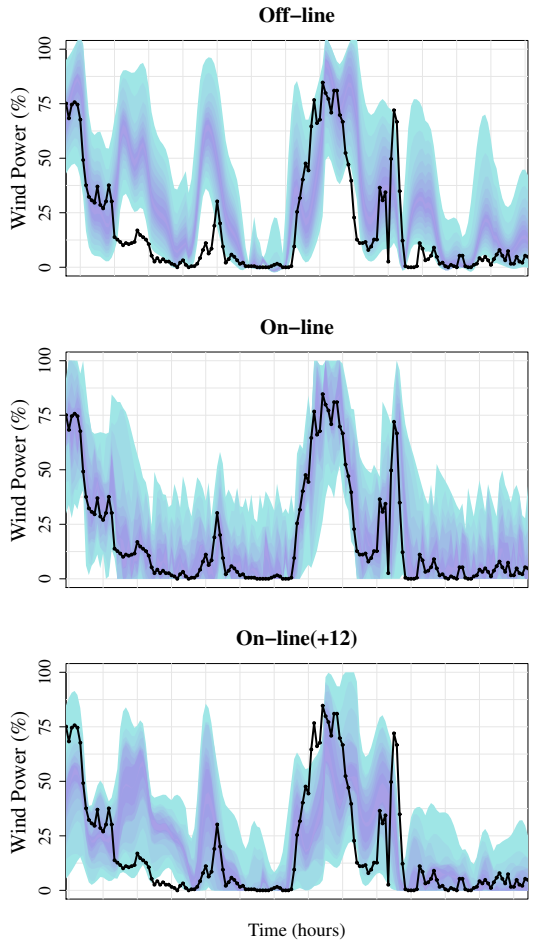

Figure 4. Probabilistic forecasts provided by the different models during one week.

data. Modellers should be aware of this effect in order to give proper interpretation to the obtained models and the related performances. Potential solutions to this problem were also described in the paper.

The following topics will be covered in a future work: (a) implementation of the models within the framework of operational wind power forecasting, and benchmark with reference quantile regression models, (b) development of non-gradient based on-line learning methods, and (c) time-adaptive tunning of the model's hyperparameters.

REFERENCES

[1] R. J. Bessa, C. L. Moreira, B. Silva, and M. A. Matos, "Handling renewable energy variability and uncertainty in power systems operation," *Wiley Interdisciplinary Reviews: Energy and Environment*, vol. 3, no. 2, pp. 156–178, March/April 2014.

[2] R. Bessa, M. Matos, I. Costa, L. Bremermann, I. Franchin, R. Pestana, N. Machado, H.-P. Waldl, and C. Wichmann, "Reserve setting and steady-state security assessment using wind power uncertainty forecast: a case study," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 4, pp. 827–837, October 2012.

[3] H. Holttinen, M. Milligan, E. Ela, N. Menemenlis, J. Dobschinski, B. Rawn, R. Bessa, D. Flynn, E. Lazaro, and N. Detlefsen, "Methodologies to determine operating reserves due to increased wind power," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 4, pp. 713–723, October 2012.

[4] J. Wang, A. Botterud, R. Bessa, H. Keko, V. Miranda, J. Akilimali, L. Carvalho, and D. Issicaba, "Wind power forecasting uncertainty and unit commitment," *Applied Energy*, vol. 88, no. 11, pp. 4014–4023, November 2011.

[5] M. A. Matos and R. Bessa, "Setting the operating reserve using probabilistic wind power forecasts," *IEEE Transactions on Power Systems*, vol. 26, no. 2, pp. 594–603, May 2011.

[6] 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, May 2012.

[7] C. Monteiro, R. Bessa, V.Miranda, A. Botterud, J.Wang, and G. Conzelmann, "Wind power forecasting: state-of-the-art 2009," Argonne National Laboratory, Tech. Rep. Report ANL/DIS-10-1, November 2009.

[8] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[9] C. Gallego-Castillo, A. Cuerva-Tejero, and O. Lopez-Garcia, "A review on the recent history of wind power ramp forecasting," *Renewable & Sustainable Energy Reviews*, vol. 52, pp. 1148–1157, 2015.

[10] 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, January 2009.

[11] R. Bessa, V. Miranda, A. Botterud, Z. Zhou, and J. Wang, "Time-adaptive quantile-copula for wind power probabilistic forecasting," *Renewable Energy*, vol. 40, no. 1, pp. 29–39, April 2012.

[12] J. Jeon and J. W. Taylor, "Using conditional kernel density estimation for wind power density forecasting," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 66–79, March 2012.

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

[14] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy*, vol. 7, no. 1, pp. 47–54, January/March 2004.

[15] H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts," *Wind Energy*, vol. 9, no. 1-2, pp. 95–108, January/April 2006.

[16] G. Rubio, H. Pomares, L. J. Herrera, and I. Rojas, *Computational and Ambient Intelligence: 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. Kernel Methods Applied to Time Series Forecasting, pp. 782–789.

[17] I. Takeuchi, Q. Le, T. Sears, and A. Smola, "Nonparametric quantile estimation," *Journal of Machine Learning Research*, vol. 7, pp. 1231–1264, 2006.

[18] W. Liu, J. C. Principe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley, 2010.

[19] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, August 2004.

[20] A. Smola and B. Schølkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[21] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[22] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, Aug 2004.

[23] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 357–363, April-June 2014.

[24] G. Anastasiades and P. McSharry, "Quantile forecasting of wind power using variability indices," *Energies*, vol. 6, no. 2, pp. 662–695, 2013.

[25] T. Nielsen, A. Joensen, H. Madsen, L. Landberg, and G. Giebel, "A new reference for wind power forecasting," *Wind Energy*, vol. 1, no. 1, pp. 29–34, 1998.