Probabilistic short-term wind power forecasting based on kernel density estimators

Jérémie Juban, Lionel Fugon, Georges Kariniotakis

Probabilistic short-term wind power forecasting based on kernel density estimators

Jérémie Juban , Lionel Fugon and George Kariniotakis
École des Mines de Paris
B.P.207, F-06904 Sophia-Antipolis, France
jeremie.juban@ensmp.fr; georges.kariniotakis@ensmp.fr

Abstract

Short-term wind power forecasting tools have been developed for some time. The majority of such tools usually provide single-valued (spot) predictions. Such predictions are however often not adequate when the aim is decision-making under uncertainty. In that case there is a clear requirement by end-users to have additional information on the uncertainty of the predictions for performing efficiently functions such as reserves estimation, unit commitment, trading in electricity markets, a.o. In this paper, we propose a method for producing the complete predictive probability density function (PDF) for each time step of the prediction horizon based on the kernel density estimation technique. The performance of the proposed approach is demonstrated using real data from several wind farms. Comparisons to state-of-the-art methods from both outside and inside the wind power forecasting community are presented illustrating the performances of the proposed method.

1 Introduction

Wind power has been undergoing a rapid development in recent years. Several countries have reached already a high level of installed wind power capacity, such as Germany, Spain and, Denmark, while others follow with high rates of development. Such large-scale integration of wind power is challenging in terms of power system management. Indeed, wind is a variable resource that is difficult to predict. As an example, traditionally, additional reserves are allocated to manage this uncertainty. This increases the overall cost of the produced energy and limits the benefits of using such a renewable energy resource.

A way of reducing the uncertainty associated to wind power production is to use forecasting tools. Development of such tools has been ongoing for more than 15 years [1]. These tools are multi-step ahead forecasting models that provide information for several horizons i.e. look-ahead times. The majority of the existing forecasting tools provide a single expected value for each forecast horizon, called deterministic, *spot* or *point* forecast. The main drawback of such predictions is that no information is provided about any departure from the predicted values. This limits their use in decision-making applications, especially those based on stochastic optimization or risk assessment.

Recently, various energy-related applications have shown the benefits of using additional information on the uncertainty. For example, such information may be used to estimate the optimal level of reserves that need to be allocated to compensate wind variability [2]. Energy bidding in a day-ahead electricity market is an emerging application. It has been shown that, when trading future production on an electricity market, the use of probabilistic wind power predictions can lead to higher benefits than those obtained by only using spot forecasts [3]. Another recent use of probabilistic predictions is in weather derivative trading. Weather derivatives enable energy companies to protect themselves against weather risk. In [4], various probabilistic methods are used to forecast the conditional density of the pay-off associated to a weather derivative.

The probabilistic models that are available today for wind power forecasting concentrate on the prediction of specific quantiles or intervals. In this paper we propose an approach that is not limited to such predefined quantities but rather provides the full probability density function of the expected wind power generation for each horizon. Such predictive PDF may then be used as such or in the form of quantiles or spot predictions as required by the decision-making algorithms.

The paper initially presents a state of the art in probabilistic forecasting. Then, a probabilistic prediction model, based on kernel density estimators, is proposed. A comparison is made with other prediction approaches. The performance of the model is evaluated using real-world data from French wind farms corresponding to different terrain complexity and climatic conditions. The paper ends with some conclusions and remarks.

2 Probabilistic forecasting

2.1 Definition

Probabilistic forecasting consists in providing the future probability of one or more events. In this sense, it is generally opposed to deterministic forecasting, where a single predicted value is provided for each considered horizon. Probabilistic forecasts can be provided under different forms depending on the nature of the variable being forecast. For discrete variables (i.e. for a finite number of possible events) probabilistic forecasts are called “probability forecasts”. Various types of forecasts exist when forecasting continuous variables. A *quantile forecast* is the value such that the observation has a predefined probability to be inferior or equal to that value. *Predictive intervals* provide a lower and an upper bound between which the observed event is expected to fall with a predefined probability. In this sense, quantile forecasts can be seen as open predictive intervals. Finally, probabilistic forecasts can be provided as *predictive cumulated probability distributions* or *predictive probability density functions*, which provide a full estimation of uncertainty. The model proposed in this paper provides *predictive probability density functions* for each forecast horizon.

2.2 Overview in various fields

Probabilistic forecasting has been developed in several fields. Meteorology and economics are the two fields that have been most active in this respect. Probabilistic forecasting has spread from these fields into other fields such as hydrology and power system management. Probabilistic forecasting has been performed in meteorology for more than a century [5]. Meteorologists developed simple approaches based on class definitions. Advanced statistical procedures such as discriminant analysis and various model output statistics techniques have also been used. More recently, a novel approach called ensemble forecasting has been developed. This approach is based on numerical model perturbation. An overview of ensemble forecasting can be found in [6].

Economics and finance has generated a substantial amount of publications on probabilistic forecasting. These kinds of predictions are used in various applications such as growth output rate, unemployment, inflation rate, stock returns, etc. A wide variety of forecasting methods exist and traditionally classified as structural or non-structural. Structural approaches view and interpret data through the lens of a particular economic theory. In contrast, non-structural methods attempt to exploit the correlation between variables with little reliance on economic theory. Widely used structural models are based on dynamic stochastic differential equations e.g. the Black-Scholes model and the dynamic stochastic equilibrium model. The most often used non-structural methods are models of volatility dynamics such as ARCH [7] and GARCH [8]. Various non-parametric statistical methods have been developed such as quantile regression [9] and bootstrap resampling [10].

2.3 Wind power applications

Probabilistic forecasting of wind power output is a recent development. Two main approaches are found: the *prediction error approach* and the *direct approach*. The first approach provides probabilistic forecasts of the errors of an existing deterministic forecasting model, while the second approach concentrates on directly providing probabilistic predictions of the considered variable.

The prediction error approach "adds" uncertainty estimation to existing "spot" forecasting systems. Early approaches used global evaluation criteria (such as the standard deviation of forecast errors computed over several runs) as forecast uncertainty assessment. However, this provides constant values for a given time period. Such approaches can be seen as measuring the "climatological" uncertainty instead of the "meteorological" uncertainty. A way to provide situation-dependent uncertainty assessments is to separate the errors into classes based on the explanatory variables. The standard deviation of prediction errors can be computed for predefined classes of predicted wind power [11] or depending on weather situations [12]. The main drawback with class definition is that it introduces discontinuities. Also determining the number of classes and their width can be difficult. A way to avoid discontinuities is to use smoothing techniques. In [3] fuzzy set theory is used to overcome the problem of class discontinuity. The error distributions are associated to different fuzzy sets and are then combined using the *linear opinion pool* or the *adapted resampling* method. A conceptually different method, quantile regression based on cubic B-spline is described in [13], where quantiles of the prediction error are computed using various explanatory variables.

Several direct probabilistic prediction approaches have also been proposed. A method to convert wind prediction error into power output uncertainty based on the derivative of the power curve is proposed in [12]. Local quantile regression is used in [14] to compute specific quantiles of the power production. A comparison of three quantile approaches, namely local quantile regression, local Gaussian modelling and, the Nadaraya-Watson estimator, is performed in [15].

2.4 Towards complete predictions

As shown in the state of the art, most probabilistic forecasting models provide predictive intervals computed from quantiles. In this paper, we propose to provide the full probability density function. Many reasons lead us to this choice. Firstly, from the full distribution all common probabilistic quantities can be extracted (e.g. spot, intervals, quantiles predictions) and this permits to avoid using multiple models to obtain each quantity. Secondly, using the full PDF enables to take better decisions. For instance, in case of bi-modal distribution (a density with two local maximums), the classical centred predictive intervals provide misleading uncertainty information of a large central uncertainty. Whereas, from the full PDF, it becomes possible to provide two prediction intervals centred on each density local maximum. Such representation is closer to the reality and permits to inform the decision-maker of these two scenarios with a smaller uncertainty. Thirdly, advanced decision-making tools may directly use the full probability density function and take into account of the full complexity of the situation.

One might think that providing the full distribution is the last step for taking optimal decisions in an environment with uncertainty. However, a limitation is identified for a full integration of such information in advanced decision-making methods. Indeed, most of these methods, like stochastic dynamic programming, are *multi-stage decisions-making* tools. Such algorithms generally need to compute scenarios of the predicted variable. However, from the predictions available today, i.e. the full distribution given for each horizon, a precise estimate of a scenario probability is impossible to compute. Indeed, full conditional predictions are necessary to compute this probability. For example, if one needs to compute the predictive probability of an event at time $t_0 + 2$ given that the event at time $t_0 + 1$ has a low value (scenario hypothesis). This probability cannot be computed directly from the prediction currently provided, i.e. prediction of the probability of this event at time $t_0 + 2$ given available information at time t_0. Providing conditional predictions is an extension of the work presented in this paper.

3 Model Input Selection

3.1 Preamble

The quantity of available information to be taken into account by wind forecasting applications can be considerable. This is amplified by the possibilities of using advances in information and communication technology. This is the case in wind power forecasting applications where, apart from the common measurements as well as the Numerical Weather Predictions (NWPs), one can also consider measurements from neighbor

sites, additional NWPs grid points or alternative models. Using all available information might potentially give us the opportunity to improve predictions. However, there are two main problems when dealing with high input dimensionality, the computational burden and the estimation quality. The problems associated with high dimensions is sometimes referred to as the *curse of dimensionality*, see for example [16].

Firstly, addressing a problem in high dimensions can rapidly become computationally intractable. Several algorithms that are easy to apply for single input become impossible to use for high dimensions. For example, in our problem, high dimensional conditional densities should be computed. In order to reduce computational burden we are proposing the using of the *kd-tree algorithm*. Kd-tree has been proposed by Bentley [17] and permits a fast computation of nearest neighbor points in the considered samples.

Second, in most statistical prediction models, the number of model parameters grows exponentially with the input dimension. In the meantime, the number of samples remains fix. Thus, the quality of the estimation of each parameter will quickly decrease leading to over-fitting the input sample. Various methods have been proposed to choose the right model order able to reduce over-fitting. Common examples are *cross-validation* and *structural risk minimization principle* [18].

In order to develop models following the well known principle of parsimony in time series forecasting, it is important at a first step of model building to select/reduce the number of input variables. The input dimension can be reduced by either combining or selecting the various input variables. In this paper, a method for the selection input variables is considered, which presents the advantage of clearly identifying the relevant inputs.

3.2 Input selection based on Information Theory

The *mutual information* is a measure from **Information Theory** introduced by Shannon. Below, the two main measures from this field are defined, namely *entropy* and *mutual information*.

3.2.1 Entropy

The entropy of a random variable enables to measure the quantity of information contained in that variable. More formally, the entropy of a random variable $\mathbf{X}$ with PDF $f_{\mathbf{X}}$ is defined as [19]:

$$H(\mathbf{X}) = \int -f_{\mathbf{X}}.log\left(f_{\mathbf{X}}\right) \tag{1}$$

3.2.2 Mutual information

The mutual information is a measure of the quantity of information contained in one random variable about another variable.

More formally, the *average mutual information* of two random variables, $\mathbf{X}$ with PDF $f_{\mathbf{X}}$ and $\mathbf{Y}$ with PDF $f_{\mathbf{Y}}$, is given by [19]:

$$I(\mathbf{X};\mathbf{Y}) = \int f_{\mathbf{X},\mathbf{Y}}.log\left(\frac{f_{\mathbf{X},\mathbf{Y}}}{f_{\mathbf{X}}.f_{\mathbf{Y}}}\right) \tag{2}$$

where $f_{\mathbf{X},\mathbf{Y}}$ is the joint PDF of $\mathbf{X}$ and $\mathbf{Y}$.

In fact, the mutual information can be seen as a measure of "distance" (Kullback-Leiber divergence) between $f_{\mathbf{X},\mathbf{Y}}$ and $f_{\mathbf{X}}.f_{\mathbf{Y}}$. So, when those two quantities equals, which correspond to the case where $\mathbf{X}$ and $\mathbf{Y}$ are independent, the distance between the distributions is null and $I(\mathbf{X};\mathbf{Y})$ equals zero. On the contrary, if $I(\mathbf{X};\mathbf{Y})$ is greater than zero, some kind of dependency is observed between the two variables. The higher the mutual information value is, the more dependency is expected between the variables.

One last interesting point is that the mutual information is bounded by $min\left(H(\mathbf{X}), H(\mathbf{Y})\right)$, this permits to normalize the mutual information as shown in the results below.

One may notice that the mutual information measures, in some way, the correlation between two variables. However, the correlation only measures the linear component of the relationship between two variables. Whereas, mutual information is able to capture the non-linearities and in this sense is more complete than simple correlation information.

Computing high dimensional mutual information is computationally intensive since it necessitates high dimensional integration. However, one may notice that mutual information can be expressed as an expectation: $E\left[log\left(\frac{f_{\mathbf{X},\mathbf{Y}}}{f_{\mathbf{X}}.f_{\mathbf{Y}}}\right)\right]$, so, the mutual information can be estimated from sample data using the law of large numbers [20]:

$$\hat{I}(\mathbf{X},Y) = \frac{1}{N}\sum_{i=1}^{N} log\left(\frac{f_{\mathbf{X},Y}(\mathbf{x}_i,y_i)}{f_{\mathbf{X}}(\mathbf{x}_i).f_Y(y_i)}\right) \tag{3}$$

thus avoiding high integration computation. Following [20], the density functions in Equation 3 are estimated using kernel density estimators.

3.3 Case study description

The aim below is to evaluate the use of the mutual information based input selection. This method is applied on the case studies of three wind farms (WF1, WF2 and, WF3) described in detail in section 5. Sixteen potential input variables from the ARPEGE Numerical Weather Prediction model (by Meteo France) are considered. These include wind speed and direction from 10m, 50m, and, 850/700 hPa levels. Also temperature, wind gust, geopotential, humidity and sea level pressure forecasts are considered. The period of study spans 18 months from July 2004 to December 2005. The forecasts are provided once a day for horizons 0 to 60 hours ahead, with a 3-hour resolution, i.e. 21 values for each meteorological variable are provided per run.

3.4 Results

In Figure 1 and Figure 2 the mutual information between the wind power production and each meteorological variables is computed for various forecast horizons for WF1 and WF2 respectively.

Firstly, as expected, wind speed and wind direction are the more relevant variables with average per horizon information content around 45 % for wind speed and 15 % for wind direction. The temperature at level 850 hPa has a noticeable information below 10 %. All other variables are independent from the wind power production and thus should not be considered for the prediction model. The levels closest to the wind turbine height (50m) are slightly more informative than other

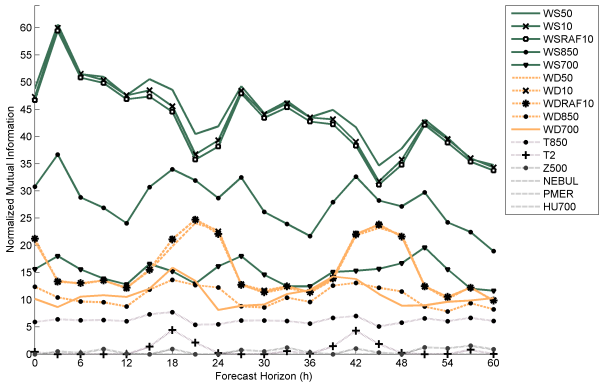

Figure 1: Mutual information computed between the measured wind power production of WF1 and several meteorological variables: wind speed (WS), wind direction (WD), wind gust (WSRAF), temperature (T), geopotential (Z), nebulosity (NEB), humidity (HUM), sea level pressure (PMER).

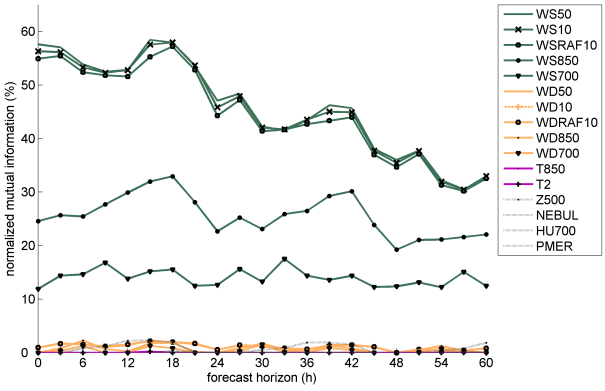

Figure 2: Mutual information between the measured wind power production of WF3 and several meteorological variables: wind speed (WS), wind direction (WD), wind gust (WSRAF), temperature (T), geopotential (Z), nebulosity (NEB), humidity (HUM), sea level pressure (PMER).

levels (10m). Secondly, as expected the information contents decreases, as the forecast horizon gets longer since far events are harder to predict than near events. We can also notice an important daily pattern along the horizons. A decrease in information content of the wind speed is observed around 18h00 (horizons +0h, +24h and, +48h). In parallel, the wind direction information content increases significantly (up to 25 %). As a consequence, we expect that the more complex situation around 18h00 will be better handled by using the increasing information coming from the wind direction.

However, evaluation of mutual information between couples of variables is not sufficient. For instance, one might be tempted to use both wind speed at 10m and 50m since they are identified as good predictors. However, the mutual information between these two variables is equal to 91 %. So, the information brought by these two variables is redundant. It is more interesting to use a variable less informative (e.g. wind direction) but with a stronger independence with variables already selected. An automated way of selecting the most informative variables is presented in [20]. This algorithm uses an heuristic in order to avoid an exhaustive search over all the combinations of explanatory variables. The algorithm starts with an empty "selection set". This set is grown by successively adding the variable that causes the greatest increase in total mutual information. The cumulated mutual information of the selected variables for WF1 is presented in Figure 3. As expected, the first selected variable is the 50m wind speed. The second selected variable is 10m (gust) wind direction. As aforementioned, this second variable is selected because it brought new information since it is both another type of variable (direction) and a different level (10m). One can notice that the 50 m wind direction is the last selected variable, which is expected, since information from both 50m level (WS50) and direction (WDRAF10) has already been considered. Finally, by considering the increase in total mutual information, the first four variables are identified as potentially informative by the algorithm. However, the variable HU700 is independent from the prediction process as shown in Figure 1. The fact that the global amount of information continues to increase significantly when HU700 is added is due to problems of different bias for different dimension in the estimation of the mutual information from samples of limited size. Such problems are studied and discussed in details in [20]. To correct this, the initial estimation of mutual information presented in Figure 1 is used to automatically remove the non-relevant variables. The result of the selection procedure leads to a final set for WF1 containing WS50, WDRAF10 and T850.

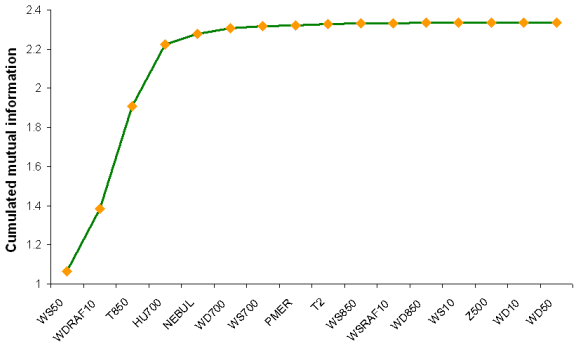

Figure 3: Cumulated mutual information of variable selection of WF1. The cumulated mutual information is plotted for the variable in selection order.

For the case studies WF2 and WF3 the selection is simpler than WF1 since from the single mutual information in Figure 2 we can deduce that only the wind speed is informative.

4 Prediction Model formulation

4.1 Preamble

In this section, two models for probabilistic wind power forecasting, *kernel density estimation* (KDE) and *quantile regression forest* are introduced and compared to the *B-Spline Quantile Regression* approach. The three models are non-parametric i.e. they do not have an hypothesis on a spe-

cific distribution family to estimate. The *Quantile Regression Forests* method is a recent method from the literature. It is design to control the effect of input uncertainty and over-fitting so the method is expected to be more robust than the two others.

In our prediction problem, two time-dependent variables are considered the hourly average power production Y_t to be predicted and a multidimensional vector of explanatory variables $\mathbf{X}_t$ summarizing all the available information up to time t. The purpose of the prediction models is to compute the distribution (quantiles or "spot" value) of Y_{t+k} given $\mathbf{X}_t$ from past data. So, the pairs of random variables (X_t, Y_{t+k}) are considered. For sake of simplicity, the pairs of past data belonging the learning set are further referred to as $(\mathbf{x}_i, y_i), i = 1..N$.

4.2 Density predictions based on KDE

4.2.1 Kernel density estimation

There are two main categories of density estimation methods: parametric and non-parametric. In the parametric framework, a distribution family is chosen, e.g. the Gaussian distribution. Then, the parameters of the distribution are estimated from the available data. In the non-parametric framework the distribution is directly estimated from the data based on a weaker hypothesis on the underlying distribution. The main drawback of the non-parametric approach is that it requires larger data sets than the parametric one to attain equivalent estimations. The main advantage is that it limits estimation errors due to incorrect hypotheses on the underlying distribution family. We have chosen a non-parametric approach, the kernel density estimation (KDE), in order to keep the prediction model as generic as possible.

The kernel density estimator computes a smooth density estimation from data samples by placing on a each sample point a function representing its contribution to the density. The distribution is obtained by summing all these contributions. The reader is referred to [21, 16, 22] for further details on kernel density estimation.

Formally, the d-dimensional multivariate kernel density estimator is given by:

$$\hat{f}(\mathbf{x}) = \frac{1}{N\,|\mathbf{H}|} \sum_{i=1}^{N} K(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)) \tag{4}$$

where $\mathbf{x}$ is the evaluation point, $\mathbf{x}_i$, $i = 1..N$ are the data samples, $\mathbf{H}$ is a $d \times d$ matrix controlling the smoothing of the estimation, K is a properly chosen kernel function. Examples of such functions are multivariate density functions.

Two parameters have to be determined, the kernel function K and the matrix $\mathbf{H}$. The choice of the kernel function has a minor role on the final quality of the estimate [16]. Following [16], we have avoided the use of the classical Normal kernel to reduce the computational overhead. We have chosen to use a biweight kernel defined by:

$$\begin{cases} K(u) = \frac{15}{16}(1-u^2)^2 & u \in [-1,1] \\ K(u) = 0 & \text{otherwise} \end{cases} \tag{5}$$

The multivariate version is simply obtained by computing the product along each component:

$$K(\mathbf{u}) = \prod_{j=1}^{d} K(u_j) \tag{6}$$

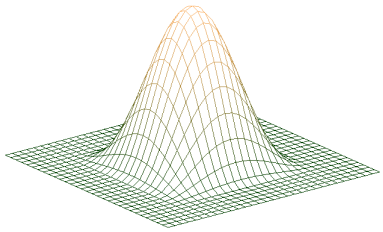

Figure 4: Two dimensional biweight kernel

The representation of a bidimentional biweight kernel function is shown in Figure 4.

The smoothing parameter $\mathbf{H}$ has a great influence on the quality of the estimated distribution. Variation around the Normal reference rule [16] has been chosen to determine the width of the H parameter.

The basic formulation of kernel density estimation presented here is adapted to estimate unbounded smooth densities. However, most of the considered variables in wind power forecasting are positive and bounded. A variety of methods have been developed in the literature for boundary correction. Reviews can be found in [16, 22, 23]. The reflection method proposed in [16] is used in this paper as a first simple approach.

4.2.2 Model formulation

Our purpose is to compute the future conditional probability density function of the variable to be predicted for time $t + k$ given the information available at time t:

$$f_{Y_{t+k}|\mathbf{x}_t} = \frac{f_{Y_{t+k},\mathbf{x}_t}}{f_{\mathbf{x}_t}} \tag{7}$$

These density functions are estimated from the data using a kernel density estimator.

$$\hat{f}_{Y_{t+k}|X_t}(y, \mathbf{x}) = \frac{1}{h} \sum_{i=1}^{N} w(\mathbf{x}, \mathbf{x}_i) K(\frac{y - y_i}{h}) \tag{8}$$

where,

$$w(\mathbf{x}, \mathbf{x}_i) = \frac{K\left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right)}{\sum_{j=1}^{N} K\left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_j)\right)}$$

and where,

- $\hat{f}_{Y_{t+h}|X_t}$ is the forecast probability density function.
- K is a multivariate biweight kernel.
- $\mathbf{H}$ and h are respectively the smoothing matrix and smoothing parameter used to control the smoothing. Parameter $\mathbf{H}$ controls the smoothing of the explanatory variables. Parameter h directly controls the smoothing of the resulting predictive PDF.

4.3 Quantile regression forests approach

Quantile regression forests is a method adapted from *Random Forests*, which rely on *classification and regression trees*. The base method used in *Quantile Regression Forests* is called classification and regression trees (CART) [24]. The goal of CART is to divide a sample of data using binary rules making the child nodes less heterogeneous than the parent nodes. Once a tree is grown it is possible to extract information from the tree structure, which makes it also a tool for data analysis. The main advantages of CART is that it permits perform a regression or a classification with high dimensional inputs. Random forest has been design to improve the CART, because the major disadvantage of the later is that it is *unstable* i.e. a small change in the training sample can generate large changes in the learned predictor (classification or regression) [25].

Formally, at the end of the construction of the tree $\hat{T}(\mathbf{x})$, every leaf corresponds to a rectangular subspace of the explanatory variables X.

The deterministic prediction of a tree, given the explanatory variables $X_t = x$, is then:

$$\hat{T}(\mathbf{x}, \theta) = \sum_{i=1}^{N} w_i(\mathbf{x}, \theta) y_i \tag{9}$$

where, θ represents the tree parameters defining how the tree is grown (e.g. split points), $w_i(\mathbf{x}, \theta)$ are weight equals to a positive constant if sample X_i is classified in the same leaf as x and 0 if it is not. The positive constant is chosen such that the weights sum to one.

Various solutions have been proposed in the literature to improve stability of various unstable classification and regression algorithms such as neural networks or CART. A way to deal with prediction stability is to generate various alternative models, which slightly differs in the learning samples or in the modeling. Such methods are called *ensembles methods* in statistics and share the same philosophy as meteorological ensemble predictions from numerical weather prediction models. An overview of meteorological ensemble can be found in [6]. Common statistical ensemble methods are *bagging*, *boosting* and, *randomization*. A comparison of these three approaches can be found in [26].

Following this, *Random Forests* [27] describe an approach for generating *ensembles* of tree-structured predictors. Formally, a Random Forest consists in a collection of tree-predictors $\hat{T}(\mathbf{x}, \theta_k)$ where θ_k are independent and identically distributed (i.i.d.) random parameter vectors that determine how the tree is grown [27]. In *Random Forest*, two procedures are used to include randomization in the construction of the trees, namely, *bagging* and *random input selection*. Bagging or **b**ootstrap **agg**rega**ting** is a method for generating an ensemble of models constructed from samples bootstrap replicates [25]. These replicates are obtained by sampling uniformly with replacement from the original samples. The predictors are then combined by voting for classification or averaging for regression [25]. *Random input selection* consists in selecting at random, at each node, a small group of input variables to split on. A version using random linear combination of inputs is also presented in [27].

In the Random Forests approach, the conditional mean $E(Y|X = x)$ is approximated by the averaged prediction of K single trees, each constructed with an i.i.d. vector $\theta_k, k = 1..K$. Let $w_i(\mathbf{x})$ be the average of $w_i(\mathbf{x}, \theta_k)$ over this collection of trees:

$$w_i(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} w_i(\mathbf{x}, \theta_k) \tag{10}$$

The deterministic prediction as given by the Random Forest approach is then:

$$\hat{T}(\mathbf{x}) = \sum_{i=1}^{N} w_i(\mathbf{x}) y_i \tag{11}$$

Quantile Regression Forests is a generalization of *Random Forests* and thus give a non-parametric way of estimating conditional quantiles for high-dimensional predictor variables [28]. As aforementioned, Random Forest approximate the conditional mean $E(Y|X = x)$ by a weighted mean over the observations of the response variable Y . One could expect that the weighted observations deliver not only a good approximation to the conditional mean but also an approximation to the full conditional distribution [28]. The estimation of the cumulated distribution function of Y_{t+k}, given $X_t = x$, is given by:

$$\hat{F}_{Y_{t+k}|X_t}(y, \mathbf{x}) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}} \tag{12}$$

using the same weights $w_i(x)$ as for Random Forests, defined in Equation 10. The predictive quantiles $\hat{Q}_\alpha(x)$ are directly obtained from the cumulated distribution function.

4.4 B-Spline Quantile Regression

The *B-Spline Quantile Regression* is used here as a third benchmark model following the formulation recently proposed in the wind power forecasting literature [13]. In quantile regression proposed by Koenker and D'Orey (1987), a quantile is expressed as a linear combination of the explanatory variables:

$$\hat{Q}(\tau, \mathbf{x}) = \beta_0(\tau) + \sum_{i=1}^{D} \beta_i(\tau) x_i \tag{13}$$

where, $\hat{Q}(\tau, \mathbf{x})$ is the estimated quantile, τ the quantile level, D is the number of considered explanatory variables x_i, $i = 1..D$, β_i are the parameters to estimate.

In [13], an additive model is used instead of a simple linear combination. This approach models the relationship between the quantile and the explanatory variables as a linear combination of known basis functions (e.g. B-spline basis):

$$\hat{Q}(\tau, \mathbf{x}) = \alpha_0(\tau) + \sum_{i=1}^{D} \sum_{j=1}^{N_b} b_{ij}(x_i) \theta_{ij}(\tau) \tag{14}$$

where $b_{ij}(.)$ are the basis functions, N_b are the number of basis functions, $\theta_{ij}(\tau)$ are unknown coefficients.

The coefficients are found using linear programming methods.

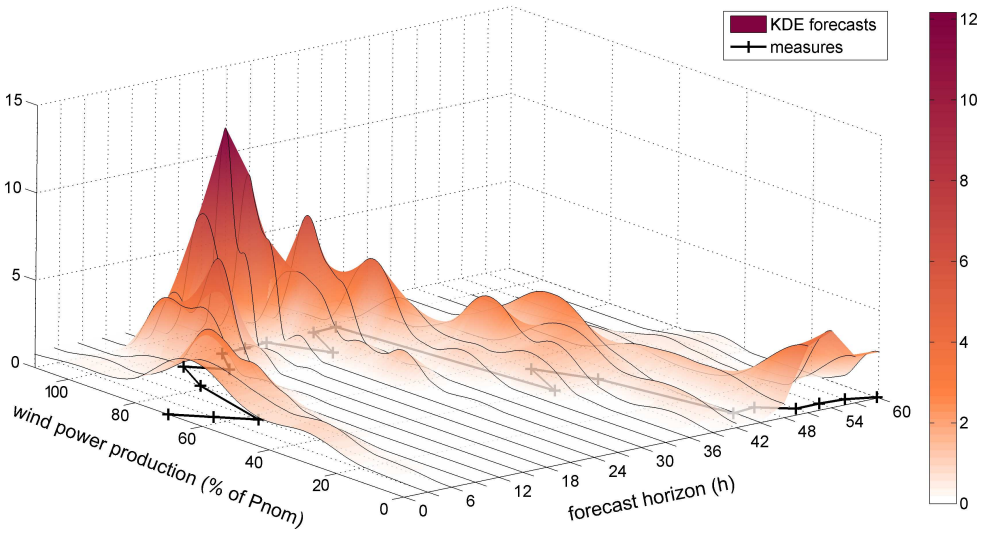

Figure 5: Example of probability density function forecasts for WF1 the 23th of November 2005. The forecast horizons ranges from +0 h to +60 h and the power production level is given in percentage of wind farm nominal power. The corresponding measured wind power production is plotted in the plane corresponding to 0 % density.

5 Evaluation results

5.1 Case study description

Three wind farms in France, denoted as WF1, WF2, and, WF3, are considered. They are representative of various terrain and climate conditions. Hourly average power production time series are considered spanning a period of 18 months from July 2004 to December 2005.

For the same period, numerical weather predictions (NWPs) by the ARPEGE model of Meteo France are used. The forecasts are provided once a day for horizons 0 to 60 hours ahead, with a 3-hour resolution, i.e. 20 values for each meteorological variable are provided per run. The meteorological variables considered in this study are the ones selected in section 3, namely 50 meter above ground level wind speed and 10m gust wind direction.

The variable to be predicted $\mathbf{Y}_t$ is the hourly average power production of each wind farm. The explanatory variable vector ($\mathbf{X}_t$) contains the predicted wind speed and wind direction by the NWP model, the last measured wind power and the forecast horizons. The horizons of power predictions are the same as that of NWPs, which ranges from 0 to 60 hours ahead, with a 3-hour resolution. The available dataset is divided into a learning-set and a test-set comprising 1 year and 6 months of data respectively.

5.2 Predictive PDF results

An example of probability density function forecasting for WF1 is presented in Figure 5. As expected the measured wind power correspond to prediction of high density. This remains true for sharp predicted PDF. One can notice that the level of uncertainty is directly related to the power production level. For high and low production levels the prediction are sharper than for mid production level. This can be explained by the influence of the slope of the wind turbine power curve when wind speed prediction errors are converted in power production errors. Finally, most of predicted PDF are multi-modal (various local maximum). For example, the prediction at horizon +15 h has two modes with one higher maximum. The corresponding measured wind power appears near the higher mode, which makes the predictions agree with the observations. In contrary, if a uni-modal parametric distribution were used instead, the new mode would lie between the two identified modes, the predicted PDF would be larger and, thus, the observations would less agree with the predictions. The problem is similar with central predictive intervals that, by definition, ignore the multi-modality. This is one of the reasons why non-parametric estimation of the full distribution enables to take better decisions.

5.3 Comparison with deterministic approaches

The aim of this section is to evaluate the performance of the proposed models in a deterministic framework. The criteria presented are the normalized root mean square error, classically used in the wind power forecasting literature [29].

The proposed KDE is used to produce density forecasts from which spot forecasts are extracted. Here, spot forecasts based on the mean and the median of the distributions are considered. These forecasts are compared to *persistence*, which is used as base line reference model, and simply consists in using the latest observation as forecast for all horizons. Persistence is commonly used as a benchmark model in wind power forecasting. In addition, the KDE-based forecasts are compared to forecasts from the *Quantile Regression Forests* as well as *B-Spline Quantile Regression* models. The results for WF1

are shown in Figure 6. The models are evaluated using the same inputs (measured power, predicted wind speed and direction) and learning/testing configuration. The performances of the deterministic predictions from the proposed models are highly comparable. However, it is of interest to notice that, even if the smoothing parameters of the KDE approach are not optimized, both the mean and median of the PDFs show a slight improvement over respectively *Quantile Regression Forests* and *B-Spline Quantile Regression*.

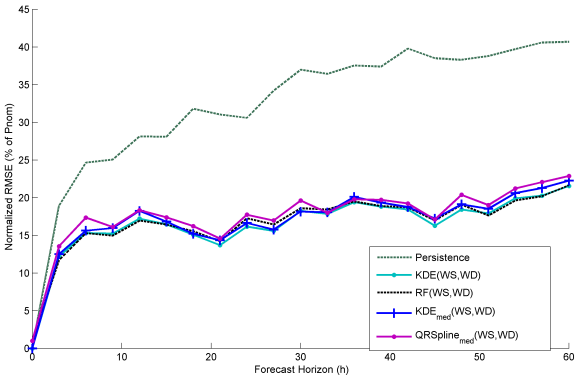

Figure 6: Normalized Root Mean Square Error for persistence, mean and median of KDE PDF and Quantile Regression Forests.

5.4 Evaluation of probabilistic predictions

5.4.1 Reliability

The "reliability" represents the ability of the probabilistic forecasting model to match the observation frequencies. For example, an 85 % predictive interval should contain 85 % of the observed values in the long run [30]. The reliability of the presented models is assessed by examining the reliability of predictive intervals and then the reliability of the full predictive PDF.

Predictive intervals with nominal coverage rates ranging from 10 % to 90 % with 10 % increments are computed from the predictive density. The choice of the 10 % increment is made so that an evaluation of such intervals is consistent with the results reported in the wind power literature. The reliability of intervals is represented by plotting the difference between the nominal and the observed coverage rates. The smaller the deviation is, the most reliable the prediction model is. A comparison of the models considered in this paper in terms of reliability is shown in Figure 7. The diagram shows the deviation from perfect reliability. The observed deviation from perfect reliability indicates that the models have a tendency to provide under-confident intervals. The order of magnitude of the results (between -2 % and 4 %) is similar to that found in the literature [3]. All the approaches presented in this paper show similar performances for most coverage rates. Concerning the KDE approach, the smoothing parameter H conditions the width of the forecast distribution. By varying the smoothing parameter one can obtain predictive intervals ranging from over-confident to under-confident ones. The overall shape of the reliability is mainly due to the fact that the smoothing parameter is the same for all classes of probabilities. The possibility to vary this parameter through time is expected to lead to important reliability improvements.

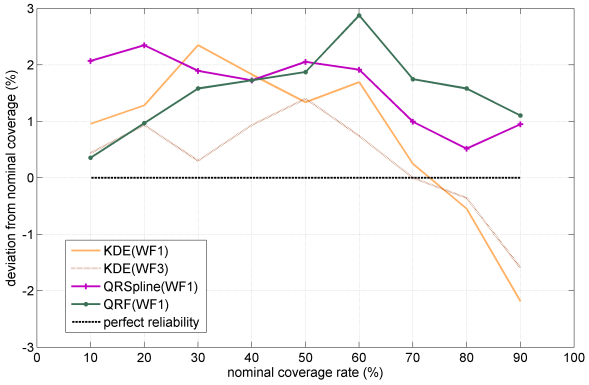

Figure 7: Reliability of predictive intervals computed from the predictive PDFs. The reliability is measured by the frequency of the observations falling within each interval.

Knowing that our method produces continuous densities, we also assess reliability in a continuous framework. To this end, the Probability Integral Transform (PIT) described in [31] is computed. The transformation consists in combining the series of continuous predictions with that of the observations. The resulting series is expected to be uniformly distributed over the $[0, 1]$ interval. The result of the difference between the quantiles of the PIT distribution and the quantiles of the true uniform distribution (diagonal $y = x$) is shown in Figure 8. Not surprisingly, the behavior observed is similar to the interval reliability. The negative spike for quantiles equals to 1 indicates that the predicted PDFs are too narrow and some observations fall outside the support of the distribution. However, the reliability diagrams reveal under-confident distributions (and intervals). This is partly due to the fact that a single smoothing parameter h is used for all sample points. Indeed, parameter h is too high (low) for the high (low) density part of the sample. A way to correct this bias is to enable h to vary as proposed in the literature [16].

5.4.2 Sharpness and Resolution

The sharpness represents the capacity of the forecasting model to forecast extreme probabilities (0 or 1 probabilities versus 0.5). This criterion evaluates the predictions independently of the observations. It gives an indication of the level of usefulness of the predictions. For example, a system that only provides uniformly distributed predictions is useless for decision making under uncertainty. Conversely, predictions having perfect sharpness are discrete predictions with probability one (deterministic predictions).

The sharpness is measured by the average interval size in the case of predictive intervals. The sharpness results obtained for the intervals described in the previous section are presented in Figure 9. As expected, the interval size increases with increasing nominal coverage rate. The results range from 3% up

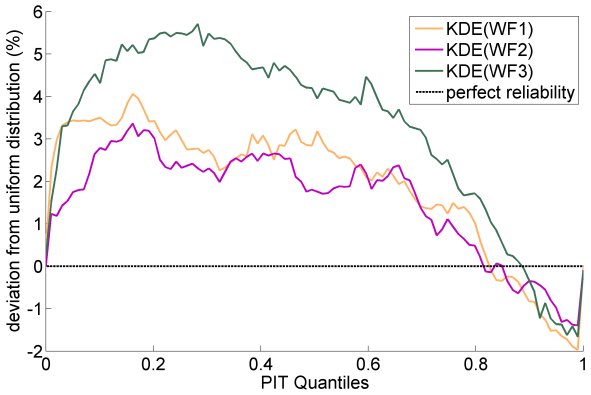

Figure 8: Difference between the quantiles of the probability integral transform of the predicted distribution and the quantiles of the uniform distribution.

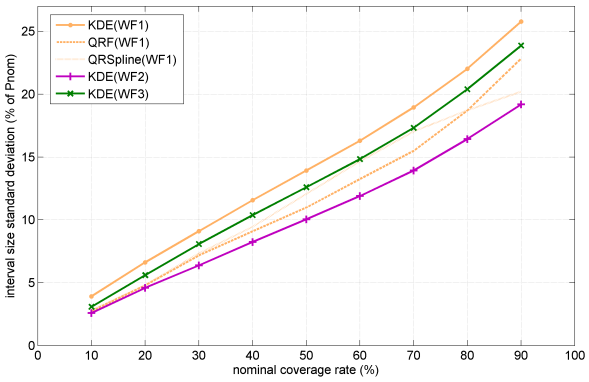

Figure 10: Resolution of predictive interval computed from the predictive PDFs. The sharpness is measured in terms interval mean size.

to 54%, which is similar to the values found in the literature [3]. The sharpness is dependent to the considered case study. For example, WF2 has a load factor significantly smaller than than WF1. For WF2 the null productions are more frequent than for WF1. These null productions are easier to forecast and leads to smaller intervals, thus making the overall sharpness smaller. Such criterion is important when comparing models on the same data but can difficultly be used to make comparison between case studies.

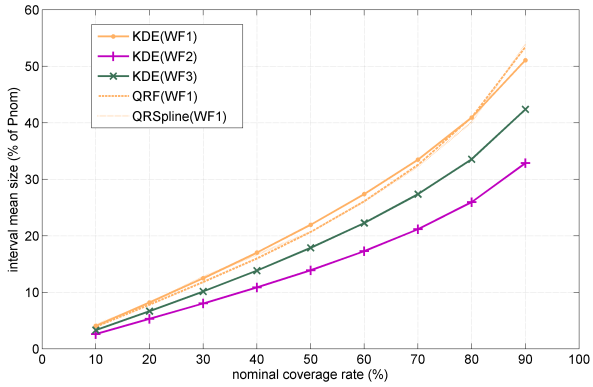

Figure 9: Sharpness of predictive interval computed from the predictive PDFs. The sharpness is measured in terms interval mean size.

Finally, another criterion used for the evaluation of probabilistic forecast is resolution. This criterion represents the capacity of the forecasting model to provide situation dependent forecasts. This criterion can be measured by the standard deviation of the interval size in case of predictive intervals [3]. The resolution for the three wind farms and the three prediction models is presented in Figure 10. Resolution and sharpness for WF1, WF2 and, WF3 shows similar results. As mentioned for sharpness, the resolution is dependent to the case study, so, WF2 which is a case study with less uncertainty (smaller sharpness), generate also less variability in interval size. Sharpness and resolution are related and tends to give the same results since they are equivalent for perfect reliability [30]. In the same way as sharpness, the resolution should be used to compare models on the same data. As opposed to sharpness, the higher the resolution is the better the model is. The KDE shows an improvement in resolution over the *B-Spline Quantile Regression* and *Quantile Regression Forests*.

5.4.3 Overall evaluation

Finally, a last criterion used to evaluate the quality of the prediction is the Continuous Ranked Probability Score (CRPS) [32]. This criterion can be seen as comprising all the previously used criteria. It serves the purpose of evaluating the forecast distribution as a whole. The main advantage of the CRPS is that it is sensitive to the entire distribution. Moreover, CRPS is expressed in the same units as the forecast variable (here % Pnom). Another interesting property is that when evaluating deterministic forecasts, CRPS is equivalent to MAE [33]. This is the reason why the CRPS evaluation shows similarities with the results of the deterministic evaluation in subsection 5.3.

The results obtained through the use of the CRPS criterion are depicted in Figure 11. Various input variables are considered, has shown in section 3. One might notice that when WDRAF10 is added a slight improvement is observed. However, as we add more variables the results are getting worst for several horizons. This is partly due to the fact that the parameter H is not optimized. However, even with an optimized H there will remain the problem of the *curse of dimensionality*. The more dimensions we wish to consider, the more data we need in order to learn the parameters and get a precise estimate. Also, the risk of over-fitting is reinforced if the explanatory variables are highly correlated.

6 Conclusions

Deterministic short-term wind power forecasting techniques have been developed for the last 15 years. Recently, several probabilistic approaches started to appear due to their interest for optimal decision making when it comes to large-scale wind power integration. Nevertheless, probabilistic methods only provide particular quantiles, or moments of the predictive distribution. The approach presented in this paper provides the complete predictive distribution. It is a non-parametric

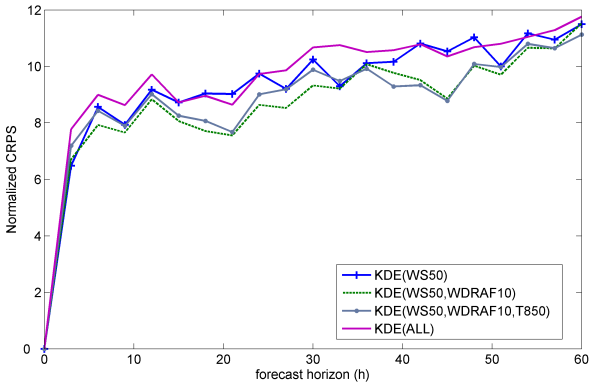

Figure 11: Continuous Ranked Probability Score evaluation results for WF1 using different input

approach based on kernel density estimation with a discrete-continuous mixed model. A wide range of forecasts products can be derived from the complete predictive distribution computed by the model: spot forecasts, quantile forecasts and, interval forecasts.

The performance of these derived forecasts has been compared to that of other forecasting models. When compared to a spot forecasting model from the literature, the probabilistic model compares very favourably. Further, when examining the derived quantiles, the values found for different performance criteria are very similar to those found in the probabilistic wind power forecasting literature.

From a more practical point of view, although the model is based on kernel density estimation, the algorithm has proved to be computationally efficient. Its computation time was found to be in the order of that of other (even deterministic) wind power forecasting models.

The paper has provided an encouraging result. Improvements can be expected by optimizing the value of the smoothing parameter of the model or by considering different smoothing parameters for different regions of the input hyperspace. Another improvement, which can be considered, is on-line tuning of the model, where the smoothing parameters evolve through time in order to take into account the non-stationary nature of wind power production.

The paper has provided a methodology to estimate the model order and select its input based on a mutual information criterion. The method is validated using real word data from European wind farms.

Acknowledgements

The authors would like to thank EDF and Météo France for providing the data for the various case studies. This work was performed in the frame of project ENSEOLE, funded in part by ADEME, the French Environment and Energy Management Agency.

References

[1] G. Giebel, G. Kariniotakis, and R. Brownsword. The state-of-the-art in short-term prediction of wind power - from a danish perspective. In *Proceedings of the 4th International Workshop on Large-Scale Integration of Wind Power and Transmission Networks for Offshore Wind Farms*, Billund, Denmark, 21-23 October 2003.

[2] R. Doherty and M. O'Malley. A new approach to quantify reserve demand in systems with significant installed wind capacity. *Power Systems, IEEE Transactions on*, 20(2):587–595, 2005.

[3] Pierre Pinson. *Estimation of the Uncertainty in Wind Power Forecasting*. PhD dissertation, École des Mines de Paris, 2006.

[4] James W. Taylor and Roberto Buizza. Density forecasting for weather derivative pricing. *International Journal of Forecasting*, 22(1):29–42, 2006.

[5] Allan H. Murphy and Robert L. Winkler. Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79:489–500, 1984.

[6] Joel K. Sivillo, Jon E. Ahlquist, and Zoltan Toth. An ensemble forecasting primer. *Weather and Forecasting*, 12(4):809–818, December 1997.

[7] Robert F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of u.k. inflation. *Econometrica*, 50:987–1008, 1982.

[8] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[9] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[10] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, New York, 2 edition, 1993.

[11] Armin Luig, Stefan Bofinger, and Hans Georg Beyer. Analysis of confidence intervals for the prediction of the regional wind power output. In *Proceedings of the European Wind Energy Conference*, Copenhagen, 2001.

[12] Matthias Lange. *Analysis of the Uncertainty of Wind Power Predictions*. PhD dissertation, Carl von Ossietzky Oldenburg University, 2003.

[13] Henrik Aalborg Nielsen, Henrik Madsen, and Torben Skov Nielsen. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy*, 9(1-2):95–108, 2006.

[14] J. B. Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7(1):47–54, 2004.

[15] J. B. Bremnes. A comparison of a few statistical models for making quantile wind power forecasts. *Wind Energy*, 9(1-2):3–11, 2006.

[16] David W. Scott. *Multivariate Density Estimation*. probability and mathematical statistics. Wiley, New York, 1992.

[17] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975.

[18] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, New York, second edition, 2000.

[19] Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[20] B. Bonnlander. *Nonparametric selection of input variables for connectionist learning*. PhD thesis, University of Colorado Department of Computer Science, 1996.

[21] Bernard W. Silverman. *Density Estimation Silverman*. Chapman & Hall/CRC, London, 1 edition, 1986.

[22] Wand M.P. and Jones M.C. *Kernel Smoothing*. Chapman & Hall, London, 1995.

[23] R.J. Karunamuni and Alberts T. on boundary correction in kernel density estimation. *Statistical Methodology*, 2:191–212, 2005.

[24] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

[25] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996.

[26] Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, August 2000.

[27] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[28] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, June 2006.

[29] Henrik Madsen, Pierre Pinson, George Kariniotakis, Henrik Aa. Nielsen, and Torben S. Nielsen. Standardizing the performance evaluation of shortterm wind power prediction models. *Wind Engineering*, 29:475–489(15), December 2005.

[30] Ian T. Jolliffe and David B. Stephenson, editors. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, New York, March 2003.

[31] Michael P. Clements. *Evaluating Econometric Forecasts of Economic and Financial Variables*. Palgrave Texts in Econometrics. Palgrave, 2005.

[32] J.E. Matheson and R.L. Winkler. Scoring rules for continuous probability distributions. *Management Sciences*, 22:1087–1095, 1976.

[33] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.