Energies **2013**, *6*, 662-695; doi:10.3390/en6020662

OPEN ACCESS

energies

ISSN 1996-1073

www.mdpi.com/journal/energies

Article

Quantile Forecasting of Wind Power Using Variability Indices

Georgios Anastasiades [1,2] **and Patrick McSharry** [1,2,*]

[1] Mathematical Institute, University of Oxford, 24-29 St Giles', OX1 3LB, Oxford, UK;
E-Mail: georgios.anastasiades@exeter.ox.ac.uk

[2] Smith School of Enterprise and the Environment, University of Oxford, Hayes House, 75 George Street, OX1 2BQ, Oxford, UK

* Author to whom correspondence should be addressed;
E-Mail: patrick.mcsharry@smithschool.ox.ac.uk; Tel.: +44-1865-614-943; Fax: +44-1865-614-960.

Received: 23 November 2012; in revised form: 12 January 2013 / Accepted: 22 January 2013 / Published: 5 February 2013

Abstract: Wind power forecasting techniques have received substantial attention recently due to the increasing penetration of wind energy in national power systems. While the initial focus has been on point forecasts, the need to quantify forecast uncertainty and communicate the risk of extreme ramp events has led to an interest in producing probabilistic forecasts. Using four years of wind power data from three wind farms in Denmark, we develop quantile regression models to generate short-term probabilistic forecasts from 15 min up to six hours ahead. More specifically, we investigate the potential of using various variability indices as explanatory variables in order to include the influence of changing weather regimes. These indices are extracted from the same wind power series and optimized specifically for each quantile. The forecasting performance of this approach is compared with that of appropriate benchmark models. Our results demonstrate that variability indices can increase the overall skill of the forecasts and that the level of improvement depends on the specific quantile.

Keywords: wind power forecasting; wind power variability; quantile forecasting; density forecasting; quantile regression; continuous ranked probability score; quantile loss function; check function

---

1. Introduction

Wind power is one of the fastest growing renewable energy sources (Barton and Infield [1]). According to the European Wind Energy Association (EWEA), the wind industry has had an average annual growth of 15.6% over the last 17 years (1995–2011). In 2011, 9616 MW of wind energy capacity was installed in the EU, making a total of 93957 MW, which is sufficient to supply 6.3% of the European Union's electricity. These figures represent 21.4% of new power capacity showing that wind energy continues to be a popular source of energy.

However, due to the large variability of wind speed caused by the unpredictable and dynamic nature of the earth's atmosphere, there are many fluctuations in wind power production. This inherent variability of wind speed is the main cause of the uncertainty observed in wind power generation. Recently, scientists have been directly or indirectly attempting to model this uncertainty and produce improved forecasts of wind power production.

According to Boyle [2], the most important application for wind power forecasting is to reduce the need for balancing the energy and reserve power which are needed to optimize the power plant scheduling. Moreover, wind power forecasts are used for grid operation and grid security evaluation. For maintenance and repair reasons, the grid operator needs to know current and future values of wind power for each grid area or grid connection point. Wind power forecasts are also required for small regions and individual wind farms.

The length of the relevant forecast horizon usually depends on the required application. For example, in order to schedule power generation (grid management), forecast horizons of several hours are usually sufficient, but for maintenance planning forecast horizons of several days or weeks are needed [3].

Since there is no efficient way to store wind energy, the wind power production decreases to zero if wind speed drops below a certain level known as the "cut-in speed". On the other hand, excessively strong winds can cause serious damage to the wind turbines, and hence they are automatically shut down at the "disconnection speed", leading to an abrupt decline of power generation. In addition, the wind power generated is limited by the capacity of each turbine. Therefore, it is important to produce accurate wind power forecasts for enabling the efficient operation of wind turbines and reliable integration of wind power into the national grid.

The literature of wind power forecasting starts with the work of Brown *et al.* [4] where they used autoregressive processes to model and simulate the wind speed, and then estimate the wind power by applying suitable transformations to values of wind speed. Most of the early literature focuses on producing wind power point forecasts, directly, or indirectly in the sense that the focus is on modelling the wind speed and then transforming the forecasts through a power curve [5,6]. The approach of modelling the wind speed series is found to be quite useful because in many situations researchers do not have access to wind power data due to its commercial sensitivity. This approach has as an advantage the fact that the wind speed time series is much smoother than the corresponding wind power time series. An obvious disadvantage is that, since the shape of the power curve may vary with the time of year and different environmental conditions, it is much more difficult to model this type of behaviour.

Recent research has focused on producing probabilistic or density forecasts, because the point forecast methods are not able to quantify the uncertainty related to the prediction. Point forecasts usually inform

us about the conditional expectation of wind power production, given information up to the current time and the estimated model parameters. Only a fully probabilistic framework will give us the opportunity to model the uncertainty related to the prediction, and avoid the intrinsic uncertainty involved in a point forecasting calibrated model. Up to now, the number of studies on multi-step quantile/density forecasting is relatively small compared with point forecasting.

Moeanaddin and Tong [7] estimated densities using recursive numerical methods, which are quite computationally intensive. Gneiting *et al.* [8] introduces regime-switching spaceCtime (RST) models which identify forecast regimes at a wind energy site and fit a conditional predictive model for each regime. The RST models were applied to 2-h-ahead forecasts of hourly average wind speed near the Stateline wind energy center in the U.S. Pacific Northwest. One of the most recent regime-based approaches is the one used by Trombe *et al.* [9], where they propose a general model formulation based on a statistical approach and historical wind power measurements only. The model they propose is an extension of Markov-Switching Autoregressive (MSAR) models with Generalized Autoregressive Conditional Heteroscedastic (GARCH) errors in each regime to cope with the heteroscedasticity.

Pinson [10], by introducing and applying a generalised logistic transformation, managed to produce ten-minute ahead density forecasts at the Horns Rev wind farm in Denmark. Pinson and Kariniotakis [11] described a generic method for the providing of prediction intervals of wind power generation and Sideratos and Hatziargyriou [12] proposed a novel methodology to produce probabilistic wind power forecasts using radial basis function neural networks. Taylor *et al.* [6] used statistical time series models and weather ensemble predictions to produce density forecasts for five wind farms in the United Kingdom. This is a relatively new approach for wind power forecasting that uses ensemble forecasts produced from numerical weather prediction (NWP) methods [6,13]. Moreover, Lau and McSharry [14] produced multi-step density forecasts for the aggregated wind power series in Ireland, using ARIMA-GARCH processes and exponential smoothing models. Jeon and Taylor [15] modelled the inherent uncertainty in wind speed and direction using a bivariate VARMA-GARCH model and then they modelled the stochastic relationship of wind power to wind speed using conditional kernel density (CKD) estimation. This is a rather promising semi-non-parametric model but unfortunately cannot be used as benchmark in this article because we aim to make predictions using only wind power data.

The quantile regression method [16] has been extensively used to produce wind power quantile forecasts, using a variety of explanatory variables among which are wind speed, wind direction, temperature and atmospheric pressure. Recent literature includes papers by Bremnes [17], Nielsen *et al.* [18], and Moller *et al.* [19]. More specifically, Bremnes [17] produced wind power probabilistic forecasts for a wind farm in Norway, using a local quantile regression model. The predictors used for the local quantile regression were outputs from a NWP model (HIRLAM10), and used lead times from 24 to 47 h. Nielsen *et al.* [18] used an existing wind power forecasting system (Zephyr/WPPT) and showed how the analysis of the forecast error can be used to build a model for the quantiles of the forecast error. The explanatory variables used in their quantile regression model include meteorological forecasts of air density, friction velocity, wind speed and direction from a NWP model (DMI-HIRLAM). Moreover, Moller *et al.* [19] presented a time-adaptive quantile regression algorithm (based on the simplex algorithm) which manages to outperform a static quantile regression model on a data set with wind power production. In addition, Pritchard [20], discussed ways of formulating

quantile-type models for forecasting variations in wind power within a few hours. Such models can predict quantiles of the conditional distribution of the wind power available at some future time using information presently available.

Davy *et al.* [21], proposed a new variability index that is designed to detect rapid fluctuations of wind speed or power that are sustained for a length of time, and used it as an explanatory variable in the quantile regression model they constructed. Bossavy *et al.* [22] extracted two new indices that are able to recognize and predict ramp events (A ramp event is defined as a large change in the power production of a wind farm or a collection of wind farms over a short period of time.) in the wind power series, and used them to produce quantile estimates with the quantile regression forest method as their basic forecasting system. Finally, Gneiting [23] studied the behaviour of quantiles as optimal predictors and illustrated the relevance of decision theoretic guidance in the transition from a predictive distribution to a point forecast using the Bank of England density forecasts of United Kingdom inflation rates, and probabilistic predictions of wind energy resources in the Pacific Northwest.

This article does not have as a purpose to develop models that can compete with the commercially available models that focus on forecast horizons greater than six hours (and are using NWPs). This is also the main reason we chose a very short forecast horizon (six hours), since it has been shown that statistical time series models may outperform sophisticated meteorological forecasts for short lead times within six hours [24]. In fact, NWPs are not even available (for some regions) for lead times shorter than three hours. So, as mentioned above, our choice of such a short forecast horizon is particularly useful for the assessment of grid security and operation. We would like to investigate the extent to which the use of quantile regression models with endogenous explanatory variables can improve the forecasting performance of probabilistic benchmarks such as persistence and climatology.

In this article we use wind power series from three wind farms in Denmark, to produce very short-term quantile forecasts, from 15 min up to six hours ahead. In order to produce quantile forecasts, we will use a linear quantile regression model, with explanatory variables extracted from the same wind power time series. Modelling the wind power series directly is preferable to a method based on wind speed forecasts because we avoid the uncertainty involved in transforming wind speed forecasts back to wind power forecasts using the power curve. The fact that we use only endogenous explanatory variables is also a very important practical consideration that we have taken on board to ensure the ability to apply our model to all wind farms. Power systems operators will require an approach to forecast a wide range of sites, where a collection of different wind farm owners implies that the only variable that they are guaranteed to have access to is the wind power generation over time.

Four new variability indices will be produced (extracted from the original wind power time series), which serve to capture the volatile nature of the wind power series. These indices, together with some lagged versions of the wind power series, will be used as explanatory variables in the quantile regression model. As for any regression model, we need predictions (point forecasts) for the future values of the explanatory variables in order to produce future quantile estimates. To produce these predictions we will use time series models that are able to model both the mean and the variance of the underlying series.

The motivation behind the chosen model structure is based on understanding the way that the underlying weather variability can affect the conditional predictive density of the wind power generation. We would like to keep the model structure as simple as possible and therefore assume that the probability

of observing a value of wind power below a certain level can be written as a function of some local mean plus the local variability involved in observing the specific wind power value. A linear combination of recently observed wind power values seems to be the easiest way to identify a function that can forecast the expected value of a specific quantile, given recent information. It is worth noticing that the model may be linear in parameters but the nonlinearity is attained in the explanatory variable themselves, and especially in the variability indices. In addition, the variability indices can capture the underlying weather variability, and hence help to improve the probabilistic forecasts given a certain weather regime.

The three Danish wind farms were chosen according to their monthly wind power capacity and standard deviation. We choose one high, one low, and one average variability wind farm, in order to understand better the ability of each model to produce probabilistic forecasts under different circumstances.

The indices used will be independently optimized for each of the three wind farms, using a one-fold cross validation technique. In fact, two different optimizations will take place for each wind farm: The first one will aim to minimize the Check Function Score (defined in Section 4.2) produced by a 1-step ahead quantile regression forecast, for each of 19 different quantiles. The second one will aim to minimize the averaged Check Function Score, produced by taking the average over all 24 predicted lead times (equal to six hours), for each quantile. The final forecast results will be compared with those of some widely used benchmark models (persistence distribution and unconditional distribution).

The remainder of the article is presented as follows. In Section 2 we will introduce the wind power data, and the new variability indices will be derived in Section 3. Section 4 will present the methodology behind the various models and explain ways to evaluate the resulting quantile forecasts. In Section 5 we will present the four competing quantile regression models and optimize their quantile forecast performance on the in-sample testing set. In Section 6 the out-of-sample quantile and density forecast performance of the competing quantile regression models will be assessed, and Section 7 will conclude the article.

2. Wind Power Data

We use wind power data recorded at three wind farms in Denmark summarized in Table 1. These wind farms were chosen to have different amounts of wind power variability, located in different geographical regions (The 446 wind farms in Denmark are assigned to 15 different geographical regions, but no further information about the actual locations of the wind farms is disclosed), and have the smallest percentage of missing values among all available wind farms. The percentage of missing values (mostly isolated points) is found to be less than 0.025% for all three wind farms, and missing values were imputed using linear interpolation. For such a small percentage of missing values, the smoothing effect caused by using linear interpolation to impute the missing values is practically negligible.

Table 1. The Danish wind farms used in this study.

Wind farm station name	Wind power variability	Wind farm rated capacity (kW)
DØR	Low	1000
ALB	Medium	25,500
VES	High	2195

Our data sets contain wind power measurements recorded every 15 min for four years, from 1 January 2007 to 31 December 2010. The data of each wind farm is bounded between zero and the maximum capacity of the wind farms. The zero value is attained in the case of excessively strong wind, where the turbines shut down in order to prevent them from damage, or in the case of very weak wind (the cut-in wind speed, usually 3–4 ms^{-1} according to Pinson [10]). In order to facilitate comparisons between the data sets of different capacities, we normalize the wind power data of each wind farm by dividing by the total (rated) capacity, which is constant over the four years period. Hence, the data is now bounded within the interval [0,1].

We dissect the data of each farm into a set of exactly two years (2007 and 2008) for in-sample model training and calibration, and an out-of-sample testing set (the remaining two years) for out-of-sample testing and model evaluation. The in-sample set is dissected again into two sub-sets, a training set and a testing set. For the in-sample training set we use the first 1.5 years and for the in-sample testing set the remaining half year. This way, we can use a *one-fold cross validation technique* to optimize the indices introduced in Section 3, and test the performance of our final chosen model using the out-of-sample testing set.

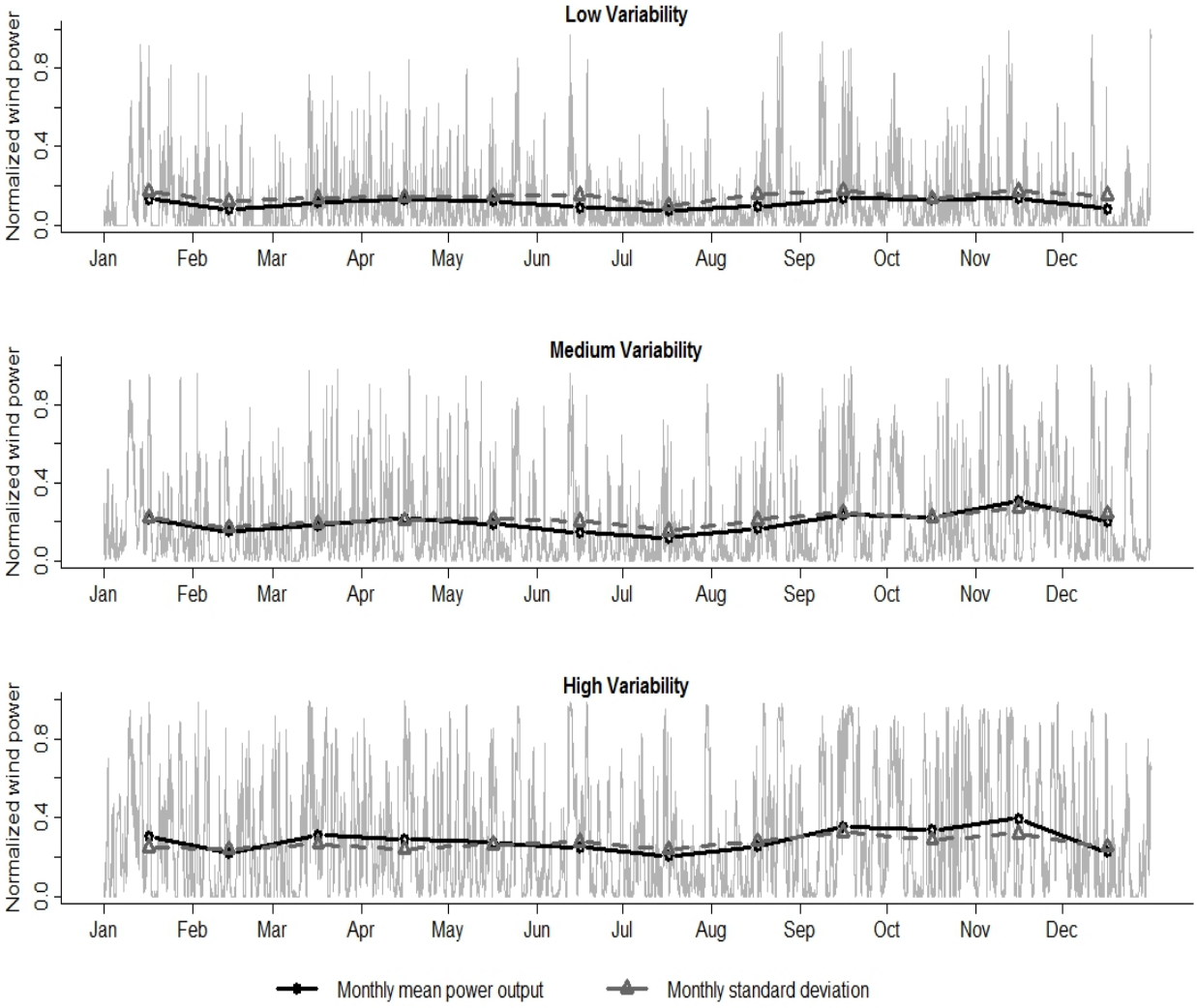
The time series plots for the year 2010, together with the monthly mean power output and standard deviation, are shown in Figure 1. The monthly mean power output and monthly standard deviation were generated by taking the mean and standard deviation of wind power, respectively, for each month over the entire four year period. As we observe, the three wind farms have different wind power variability. More specifically, the first and last wind farms of Figure 1 have the lowest and highest possible wind power variability for all four years (from all the available wind farms in Denmark), without having any significant changes (Wind power variability may change from year to year by addition of new turbines or removal (maybe for maintenance) of existing ones.) in the capacity from year to year. The second wind farm of Figure 1 was chosen to have an average (medium) wind power variability compared with the other two farms, but again without having any significant changes in the monthly capacity from year to year.

3. Indices of Wind Power Variability

Davy *et al.* [21] proposed a variability index that is designed to detect rapid fluctuations of wind speed or power that are sustained for a length of time. They defined this variability index as the standard deviation of a band-limited signal in a moving window, and they constructed such an index for a wind speed time series. This variability index depends on four parameters: the order of the filter (integer greater than one), the upper and lower frequencies of the extracted signal, and the width of

the moving window. We would like to use such an index as an explanatory variable in our quantile regression, but a proper optimization of this is too computationally expensive because of the number of parameters involved.

Figure 1. Time series plots of normalized power data for the three chosen Danish wind farms, for the year 2010. Please note that the point on the time axis labelled Jan refers to 00:00 on 1 January and similarly for every month.



Instead, we propose a parsimonious variability index which depends only on two parameters, (m, n) where $m, n \in \mathbb{N}_0 \backslash \{1\}$, and is constructed as follows. Firstly we smooth our original wind power series using an averaging window of size m, in order to obtain the smoothed wind power series,

$$r_t = \begin{cases} \frac{1}{m} \sum_{i=1}^{m} y_{t-i+1} & \text{if } m > 1 \\ y_t & \text{if } m = 0 \end{cases} \quad (1)$$

for $t \geq m$. Note that this series behaves in a fully retrospective way, in the sense that each point of the series depends only on the historical values of the original series. Since the smoothed series is $m - 1$ points smaller than the original series, we set $r_t = r_m$, for $t = 1, 2, ..., m - 1$.

Finally, the new variability index is just the standard deviation of the extracted smoothed wind power series in a moving window of width n. So, if r_t is a given point of the smoothed series, we define the new index as

$$SD_t = \begin{cases} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(r_{t-i+1} - \frac{1}{n}\sum_{j=1}^{n} r_{t-j+1}\right)^2} & \text{if } n > 1 \\ r_t & \text{if } n = 0 \end{cases} \quad (2)$$

for $t \geq n$. Again, we impute the first $n-1$ points of the series by setting $SD_t = SD_n$, for $t = 1, 2, ..., n-1$. This index can be optimized much more easily than the one proposed by Davy *et al.* [21], since it has only two parameters: the smoothing parameter m, and the variability parameter n.

By similar reasoning, we create another three variability indices. We create the smoothed wind power series, r_t, as defined by Equation (1), and then instead of finding the standard deviation we find the sample interquartile range (IQR), the 5% and the 95% sample quantiles of the smoothed series over a moving variability window (different for each series) of width n.

There are many different ways to define the quantiles of a sample. We use the definition recommended by Hyndman and Fan [25] and presented as follows. Let $R_t = \{r_{t-n+1}, ..., r_{t-1}, r_t\}$ for $t \geq n > 1$, denote the order statistics of R_t as $\{r_{(1)}, ..., r_{(n)}\}$ and let $\hat{Q}_{R_t}(p)$ denote the sample p-quantile of R_t with proportion $p \in (0, 1)$. We calculate $\hat{Q}_{R_t}(p)$ (for a chosen proportion p) by firstly plotting $r_{(k)}$ against p_k, where $p_k = \frac{k-1/3}{n+1/3}$ and $k = 1, .., n$. This plot is called a quantile plot and p_k a plotting position. Then, we use linear interpolation of $(p_k, r_{(k)})$ to get the solution $(p, \hat{Q}_{R_t}(p))$ for a chosen $0 < p < 1$. Therefore, the three new indices can be defined as:

$$IQR_t = \begin{cases} \hat{Q}_{R_t}(0.75) - \hat{Q}_{R_t}(0.25) & \text{if } n > 1 \\ r_t & \text{if } n = 0 \end{cases} \quad (3)$$

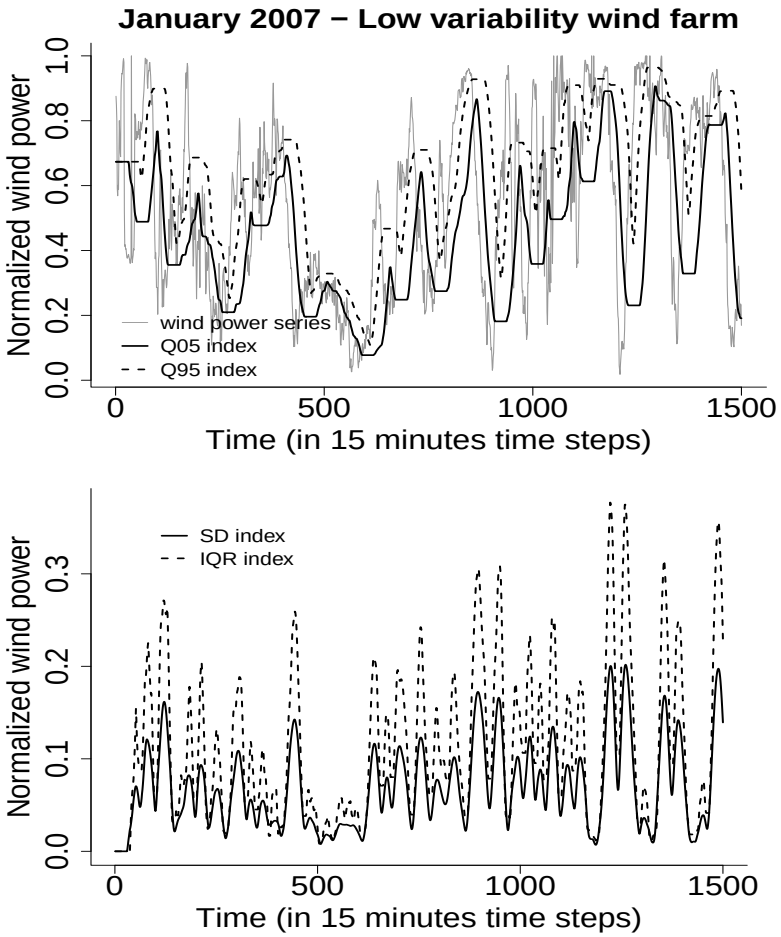
$$Q05_t = \begin{cases} \hat{Q}_{R_t}(0.05) & \text{if } n > 1 \\ r_t & \text{if } n = 0 \end{cases} \quad (4)$$

$$Q95_t = \begin{cases} \hat{Q}_{R_t}(0.95) & \text{if } n > 1 \\ r_t & \text{if } n = 0 \end{cases} \quad (5)$$

for $t \geq n$. We also impute their values for $t = 1, ..., n-1$ in a similar way as we did for the SD index. An example of the construction of the three variability wind power indices is shown in Figure 2. A first observation is that the IQR and SD indices behave similarly, but the IQR index has higher peaks than the SD index, and hence gives more emphasis to the high variability regions of the wind power series. Moreover, the Q05 and Q95 indices also behave quite similarly, capturing the two tails of the wind power distribution over a predefined window.

These indices will be properly optimized and will be used, together with some lagged values of the original power series, as explanatory variables in the quantile regression introduced in the next section. It is worth mentioning that the choice of firstly smoothing the wind power series is taken in order to take into consideration the fact that any noise may hide or alter the pattern of the underlying weather regime we wish to capture. By choosing $m = 0$ we do not remove any of the underlying noise, and hence we assume that the weather variability is fully captured by using the original wind power time series.

Figure 2. Wind power time series plot of the low variability farm, together with the four variability indices (Q05, Q95 on upper plot, and SD, IQR on lower plot). The parameters are chosen to be the same for all indices to facilitate comparison ($m = 30$ and $n = 30$).



4. Quantile Regression, Forecasting, and Evaluation Methodology

In order for the paper to be self-consistent, we include the theory of linear quantile regression in Section 4.1. In Section 4.2 we introduce the methodology we will use to evaluate the produced quantile and density forecasts.

4.1. Quantile Regression

Given a random variable, y_t, and a strictly increasing continuous CDF, $F_t(y)$, the α_i-quantile, $q_t^{(\alpha_i)}(y)$, with proportion $\alpha_i \in [0, 1]$ is defined as the value for which the probability of obtaining values of y_t below $q_t^{(\alpha_i)}$ is α_i:

$$\mathbb{P}\left(y_t < q_t^{(\alpha_i)}\right) = \alpha_i \quad or \quad q_t^{(\alpha_i)} = F_t^{-1}(\alpha_i) \tag{6}$$

Note that the notation y_t is used for denoting both the stochastic state of the random variable at time $t = 1, 2, ..., T$, and the measured value at that time for a training set of size T.

Quantile regression, introduced by Koenker and Bassett [16], models $q_t^{(\alpha_i)}$ for $\alpha_i \in [0,1]$, as a linear combination of some given explanatory variables (also called regressors or predictors). So, the α_i-quantile is modelled as:

$$\begin{aligned} q_t^{(\alpha_i)} &= \gamma_0^{(\alpha_i)} + \gamma_1^{(\alpha_i)} x_{t,1} + ... + \gamma_p^{(\alpha_i)} x_{t,p} \\ &= \gamma_0^{(\alpha_i)} + \sum_{j=1}^{p} \gamma_j^{(\alpha_i)} x_{t,j} \end{aligned} \tag{7}$$

where $\gamma_j^{(\alpha_i)}$ are unknown coefficients depending on α_i, and $x_{t,j}$ are the p known explanatory variables. In quantile regression, a regression coefficient estimates the change in a specified quantile of the response variable produced by a one unit change in the corresponding explanatory variable.

We define the *quantile loss function* [16], also known as the *check function*, for a given proportion $\alpha_i \in [0,1]$ as:

$$\begin{aligned} \rho_{\alpha_i}(u) &= \big(\alpha_i - \mathbb{1}_{\{u<0\}}\big)u \\ &= \begin{cases} \alpha_i u, & u \geq 0 \\ (\alpha_i - 1)u, & u < 0 \end{cases} \end{aligned} \tag{8}$$

where u is a given value. Then, the sample α_i-quantile can be calculated by minimizing $\sum_{t=1}^{T} \rho_{\alpha_i}(y_t - q)$ with respect to q. Hence, we can estimate the unknown coefficients, $\gamma_j^{(\alpha_i)}$, by replacing q with the right-hand side of Equation (7):

$$\hat{\gamma}^{(\alpha_i)} = \underset{\gamma}{\operatorname{argmin}} \sum_{t=1}^{T} \rho_{\alpha_i}\big\{y_t - (\gamma_0 + \gamma_1 x_{t,1} + ... + \gamma_p x_{t,p})\big\} \tag{9}$$

where $\hat{\gamma}^{(\alpha_i)}$ is a vector containing the unknown coefficients. Usually, these estimates are calculated using linear programming techniques as in Koenker and D'Orey [26].

In this article we will use quantile regression to forecast the values of quantiles with nominal proportion $\alpha_i = \{0.05, 0.10, ..., 0.95\}$, for forecast horizons $k = 1, 2, ..., 24$, measured in time steps of 15 min. We denote the forecast for the quantile with nominal proportion α_i issued at time t for forecast time $t+k$, by $\hat{q}_{t+k|t}^{(\alpha_i)}(y)$. In order to produce these forecasts, we use Equation (7), and the estimated coefficients, $\hat{\gamma}^{(\alpha_i)}$:

$$\begin{aligned} \hat{q}_{t+k|t}^{(\alpha_i)}(y) &= \gamma_0^{(\alpha_i)} + \gamma_1^{(\alpha_i)} \hat{x}_{t+k|t,1} + ... + \gamma_p^{(\alpha_i)} \hat{x}_{t+k|t,p} \\ &= \gamma_0^{(\alpha_i)} + \sum_{j=1}^{p} \gamma_j^{(\alpha_i)} \hat{x}_{t+k|t,j} \end{aligned} \tag{10}$$

where $\hat{x}_{t+k|t,j}$ for $j = 1, ..., p$ denote the forecasts of the explanatory variables $x_{t,j}$, issued at time t with lead time $t+k$.

The random variable y_t will represent the normalized wind power time series, (y_t), and the explanatory variables will be represented by time series, $(x_{t,j})$, extracted from the normalized wind power series. In order to produce the forecasts, $\hat{x}_{t+k|t,j}$, we will fit suitable time series models to the variables $(x_{t,j})$, and then predict from these models up to $t+k$ values ahead.

It is worth mentioning that by producing quantile forecasts using quantile regression, we may end up with some quantile forecasts crossing each other. This is a not very common phenomenon for so few quantile forecasts (19 in our case), but monitoring its occurrence is very important. In our analysis, whenever this phenomenon happens (it occurs very rarely because we fit the models to a large amount of data) we just shift the crossing quantile forecasts in order to keep $\hat{F}_{t+k}\big(\hat{q}^{(\alpha_i)}_{t+k|t}\big) = \alpha_i$, for $\alpha_i = \{0.05, 0.10, ..., 0.95\}$, a strictly increasing function.

4.2. Quantile and Density Forecast Evaluation

The evaluation of the quantile forecasts, for each quantile, $\alpha_i = \{0.05, 0.10, ..., 0.95\}$, will be undertaken using the quantile loss function:

The ***quantile loss function***, also known as the ***check function*** [3,27] is used to define a specific quantile of the distribution and was defined in Section 4.1, Equation (8). Hence, given a testing set of size N, we can estimate a particular quantile, $\hat{q}^{(\alpha_i)}$, with proportion α_i, using

$$\hat{q}^{(\alpha_i)} = \min_{q} \sum_{t=1}^{N} \rho_{\alpha_i}(y_t - q) \tag{11}$$

and therefore we can evaluate a series of quantile forecasts, $\hat{q}^{(\alpha_i)}_{t+k|t}$, issued at time t with lead time $t + k$ and nominal proportion α_i, using:

$$QL(k, \alpha_i) = \frac{1}{N} \sum_{t=1}^{N} \rho_{\alpha_i}(y_{t+k} - \hat{q}^{(\alpha_i)}_{t+k|t}) \tag{12}$$

This is the average over the whole testing set of the *check function score*, $\rho_{\alpha_i}(y_{t+k} - \hat{q}^{(\alpha_i)}_{t+k|t})$, for the quantile α_i, for a k-step ahead prediction. From now on we will call this function the ***Check Function (CF***), and the its score the ***Check Function Score (CFS)*** .

Using the different quantile forecasts we can also reconstruct the whole probability / cumulative forecasted distribution. We use the ***Continuous Ranked Probability Score (CRPS)*** in order to evaluate the density forecasts for each forecast horizon:

The $crps$ [28] is computed by taking the integral of the Brier scores for the associated probability forecasts at all real valued thresholds,

$$crps(\hat{F}_{t+k|t}(y), y_{t+k}) = \int_{-\infty}^{+\infty} \big(\hat{F}_{t+k|t}(y) - \mathbb{1}_{\{y \geq y_{t+k}\}}\big)^2 \, \mathrm{d}y \tag{13}$$

$$= \int_{0}^{1} QS_{\alpha_i}\big(\hat{F}^{-1}_{t+k|t}(\alpha_i), y_{t+k}\big) \, \mathrm{d}\alpha_i \tag{14}$$

where $\hat{F}_{t+k|t}(y)$ corresponds to the CDF forecast, and y_{t+k} to the corresponding verification. $\mathbb{1}_{\{y \geq y_{t+k}\}}$ is an indicator function that equals one if $y \geq y_{t+k}$ and zero otherwise. The quantile score, QS_{α_i} [29], is defined by

$$QS_{\alpha_i}(q, y) = 2\big(\alpha_i - \mathbb{1}_{\{y<q\}}\big)\big(y - q\big) \tag{15}$$

Hence, the average of these $crps$ values over each forecast-verification pair gives the CRPS for each forecast horizon k:

$$\text{CRPS}(k) = \frac{1}{N}\sum_{t=1}^{N} crps(\hat{F}_{t+k|t}(y), y_{t+k}) \tag{16}$$

$$= 2\int_0^1 QL(k, \alpha_i)\,\mathrm{d}\alpha_i \tag{17}$$

where $QL(k, \alpha_i)$ is CF defined in Equation (12). Representation (17) is useful to produce a rough estimate of the in-sample CRPS for each forecast horizon, using the CFS for each quantile. This is a rather poor approximation of the CRPS, because the number of quantiles used in this article (19 quantiles), is not large enough to produce an accurate approximation of the integral in Equation (17).

In order to find the out-of-sample CRPS for each k, we will use the following alternative representation of the $crps$, introduced by Gneiting and Raftery [29]:

$$crps(\hat{F}_{t+k|t}(y), y_{t+k}) = \mathbb{E}_{\,F}|X - y_{t+k}| - \frac{1}{2}\mathbb{E}_{\,F}|X - X'| \tag{18}$$

where X and X' are independent copies of a random variable with CDF $\hat{F}_{t+k|t}$. This representation is particularly useful when $\hat{F}$ is represented by a sample, as in our case. Then, the CPRS for each forecast horizon k is given by Equation (16).

Moreover, it will be necessary to quantify the gain/loss of some forecasting models with respect to a chosen reference model. Following McSharry *et al.* [3], this gain, denoted as an improvement with respect to the considered reference forecast system, is called a *Skill Score* and is defined as:

$$\text{Skill Score}(k) = \frac{\text{SCORE}_{\text{ref}}(k) - \text{SCORE}(k)}{\text{SCORE}_{\text{ref}}(k)} = 1 - \frac{\text{SCORE}(k)}{\text{SCORE}_{\text{ref}}(k)} \tag{19}$$

where k is the lead time of the forecast and SCORE is considered the evaluation criterion score (such as CRPS or CFS). By using the above definition we can also introduce the *Average Skill Score*. This is just the Skill Score with the scores of the competing and reference models averaged over all forecast horizons. It is defined as:

$$\text{Average Skill Score} = 1 - \frac{\sum_{k=1}^{k_{\max}} \text{SCORE}(k)}{\sum_{k=1}^{k_{\max}} \text{SCORE}_{\text{ref}}(k)} \tag{20}$$

So, when we are talking about Score, the lower the value the better the performance; but, when we are talking about Skill Score (or Average Skill Score), the higher the value the better, since we are comparing the candidate model to the reference model. Please note that the reference model will be different each time, and chosen according to the comparison we wish to make.

In order to formally rank and statistically justify any possible difference in the CRPS and CFS of the competing models with respect to the reference models, we will use the Amisano and Giacomini test [30] of equal forecast performance. This test is based on the statistic

$$t_{N,k} = \sqrt{N}\frac{\text{SCORE}(k) - \text{SCORE}_{\text{ref}}(k)}{\hat{\sigma}_{N,k}} \tag{21}$$

where SCORE again is considered the evaluation criterion score such as the CRPS or CFS, N is the out-of-sample size, and

$$\hat{\sigma}^2_{N,k} = \frac{1}{N-k+1} \sum_{j=-(k-1)}^{k-1} \sum_{t=1}^{1+N-k-|j|} \delta_{t,k}\delta_{t+|j|,k} \quad \text{where} \quad \delta_{t,k} = \mathrm{S}(t+k|t) - \mathrm{S}_{\text{ref}}(t+k|t) \tag{22}$$

The functions S and S_{ref} represent the before averaging scores (such as the $crps$ of Equations (13) or (18) and *check function score* defined just after Equation (12)) of the competing and reference models, respectively. Assuming suitable regularity conditions, according to Amisano and Giacomini [30], the statistic $t_{N,k}$ is asymptotically standard normal under the null hypothesis of zero expected score differentials. Small p-values of this test provide evidence that the difference in the forecast performance of the two forecasting (given a specific evaluation score) is statistically significant.

5. Optimization of the Variability Indices

In this section we will introduce four different quantile regression models, and using one-fold cross validation try to optimize their probabilistic forecasting performances. Our main goal is to evaluate whether or not the four variability indices (introduced in Section 3) can help to provide trustworthy quantile forecasts of wind power, when used as explanatory variables in the quantile regression model (7). For this purpose, we have to find the optimal set of parameters (m, n) of these indices, which provides the best quantile forecast performance, for each individual quantile. We do that using the following procedure.

For each index, we sample different combinations of parameters from the range $m, n = \{0, 8, 16, ..., 192\}$, in order to produce 625 different realizations of each index, for each wind farm. A preliminary analysis showed that creating a moving window larger than 192 time-points wide (2880 min *i.e.*, 2 days) did not increase the performance of the indices.

Then, for each set of parameters, we fit the following four different quantile regression models on the in-sample training set (of each wind farm), for each of the 19 quantiles $\alpha_i = \{0.05, 0.1, ..., 0.95\}$:

$$\begin{aligned}
\text{SD model:} \quad & q_t = \gamma_{01} + \gamma_{11} y_{t-1} + \gamma_{21} y_{t-2} + \gamma_{31} y_{t-3} + \gamma_{41} SD_t^{(\alpha_i)} \\
\text{IQR model:} \quad & q_t = \gamma_{02} + \gamma_{12} y_{t-1} + \gamma_{22} y_{t-2} + \gamma_{32} y_{t-3} + \gamma_{42} IQR_t^{(\alpha_i)} \\
\text{Q05 model:} \quad & q_t = \gamma_{03} + \gamma_{13} y_{t-1} + \gamma_{23} y_{t-2} + \gamma_{33} y_{t-3} + \gamma_{43} Q05_t^{(\alpha_i)} \\
\text{Q95 model:} \quad & q_t = \gamma_{04} + \gamma_{14} y_{t-1} + \gamma_{24} y_{t-2} + \gamma_{34} y_{t-3} + \gamma_{44} Q95_t^{(\alpha_i)}
\end{aligned} \tag{23}$$

where $q_t \equiv q_t^{(\alpha_i)}$ is defined in Equation (6), $\gamma_{hl} \equiv \gamma_{hl}^{(\alpha_i)}$ are the regression coefficients, and y_{t-j} are lagged wind power series. The choice of the number wind power series lags used as explanatory variables was taken by considering the AIC (a prediction based criterion according to Akaike [31]) of different quantile regression models which have different numbers of lags as explanatory variables. We also investigated the improvement obtained by adding to the right hand side of Equation (23) a combination of variability indices. Due to collinearity effects, the SD and IQR indices cannot coexist in the same equation. Any other combinations of the variability indices did not provide reduction to the AIC for more than 14 out of 19 quantile regression equations, at any of the three wind farm sites. Hence, we considered examining

the effect that each individual variability index will provide by being included as an explanatory variable to the quantile regression equations, as defined by Equation (6).

Moreover, we also considered adding to the right hand sides of Equation (23) a trigonometric function (also introduced in Equation (24) below) which uses two pairs of harmonics to regress wind power quantile, q_t, on the 15 min time step of the day. The addition of this function, which is used to model the diurnal component of each quantile of the wind power production at each wind farm, was not found to provide reduction to the AIC of 17 out of 19 quantile regression equations, at any of the three wind farm sites. Hence, in order to obtain parsimonious models we excluded these functions from the final models. Nevertheless we must acknowledge the fact that a diurnal effect may be relevant and very important for wind farms in other locations or countries.

The models in Equation (23) are regression models, and hence, in order to predict their responses, $q_t^{(\alpha_i)}$, we need predictions for their explanatory variables. These are just lagged versions of the original wind power series, and the different variability indices. All of these explanatory variables have similar characteristics as they result from the original wind power series. The lagged versions of the wind power series are certainly non-stationary and all 4×625 different realizations of the variability indices (for every wind farm), even though they can be much smoother (for large values of m, n) than the original wind power series, are also non-stationary.

The predictions (point forecasts) of the explanatory variables are produced using ARIMA and ARIMA (in mean)-GARCH (in variance) models. By modelling the mean of the series using an ARIMA model, we allow for its non-stationary nature, and by modelling the variance using a GARCH process we allow for its heteroskedastic nature. Due to the fact that the wind power series (and the resulting variability indices) is bounded and does not follow any known parametric distribution, one may argue that an ARIMA or an ARIMA-GARCH model may not be appropriate. A modified (This version of ARIMA-GARCH model limits the forecasts to be bounded between two specific values (zero and one in our case) ARIMA/ARIMA-GARCH model with limiter (as proposed by Chen *et al.* [32]) is used to deal with the problem of the data being bounded. Moreover, the empirical density of the differenced series is close to a Student's *t*-distribution density. Hence, we fit an ARMA/ARMA-GARCH model to the transformed series, (w_t) (or differenced variability index), assuming those data come from a Student's *t*-distribution whose parameters are estimated for each series. We incorporate this distributional assumption by assuming the resulting residual series (white noise) follows a Student's *t*-distribution.

The next step is to produce point forecasts from 15 min up to 6 h ahead ($k = 1, 2, ..., 24$), from each point of the in-sample testing set, by fitting ARIMA$(1, 1, 1)$ models to each realization of the four variability indices of the above regressions. Our choice of ARIMA$(1, 1, 1)$ model may seem unappealing and arbitrary, but was made mainly for simplicity after exploring the forecast performances of various time series models. Choosing the best ARIMA-GARCH model (according to AIC) for each of the 625 different realizations of each index (for each wind farm) is extremely computationally expensive and hence we have to make some simplifications in order to make our optimization process computationally feasible. An ARIMA(1,1,1) is able to capture the non-stationary nature of the indices, and avoid overfitting at the same time. In order to assess the goodness of the fits, we use the Ljung–Box test, and restrict our selection to the fits that do not reject the null hypothesis of this test (so the corresponding residuals are consistent with white noise).

Modelling the variance of the indices using ARCH/GARCH models (in combination with an ARIMA model for the mean) does not provide a consistent and significant improvement of the RMSE (We used the Root Mean Square Error to evaluate the point forecast performance of various time series models.) of the point forecasts. This is mainly because of the very small forecast horizon we have, and hence it suffices to use a simple ARIMA model with limiter. In order to produce point forecasts of the lagged wind power series, the model solution using AIC (results are also the same using BIC) identified an ARIMA$(0,1,2)$ - GARCH$(1,1)$ model for the low variability farm, an ARIMA$(1,1,3)$ - GARCH$(1,1)$ for the medium variability farm, and an ARIMA$(2,1,1)$ - GARCH$(1,1)$ for the high variability farm. These models have the ability to capture the heteroskedastic effects that the wind power series have, taking into account the non-linear nature of the variations. Also, these forecasts are calculated only once for all different realizations of the quantile regression models, and hence there is no point in this case to sacrifice the (small) accuracy gain for simplicity and computational efficiency. Table 2 shows the selected time series models for each wind farm and the two tests that assess their fit.

Table 2. Best fitted models for the three wind power time series according to the AIC, with Ljung–Box and LM tests p-values.

Wind farms time series	**Selected model based on AIC**	**LM test *p*-values for lags 5, 15, 25**	**LB test *p*-values for lags 5, 15, 25**
Low Var.	ARIMA$(0,1,2)$-GARCH$(1,1)$	1.00, 1.00, 1.00	0.87, 0.99, 1.00
Medium Var.	ARIMA$(1,1,3)$-GARCH$(1,1)$	1.00, 1.00, 0.95	0.53, 0.98, 1.00
High Var.	ARIMA$(2,1,1)$-GARCH$(1,1)$	1.00, 1.00, 1.00	1.00, 1.00, 1.00

After producing quantile forecasts for 24 different forecast horizons, we evaluate them (i) using the CFS of only the first step ahead forecasts; and (ii) using the CFS averaged over all forecast horizons. The results justify our inspection of better forecast performance for the models with small (smoothing and variability) moving windows. We repeat the above procedure by restricting the range of our parameters even more for each variability index, and sample every different combination of parameters from the range $m, n = \{0, 1, 2, ...50\}$.

We end up with distinct sets of parameters (for each model and wind farm) that minimize the averaged and 1-step ahead CFS of each different quantile. The CFS minimization results are shown in Tables 3–6. In general, we cannot distinguish any particular parameter pattern, but there are some features that are worth mentioning. For all the models, it is more common to have the smoothing window width (m) smaller than the variability window width (n), especially for quantiles less than or equal to the median. This pattern changes for the upper quantiles (larger than the median) where we do not observe a clear pattern. Also, on average, the parameters for the averaged over 24-steps ahead optimization are smaller than the corresponding ones of the 1-step ahead optimization.

Table 3. 1-step and averaged over 24-steps CFS optimization results for the SD model of Equation (23).

	Low Var.		**Med Var.**		**High Var.**		**Low Var.**		**Med Var.**		**High Var.**	
	m	n	m	n	m	n	m	n	m	n	m	n
α_i	**1-step optimization of SD model**						**24-steps optimization of SD model**					
0.05	0	9	0	18	2	7	2	5	3	3	2	7
0.10	0	4	0	20	2	7	0	6	3	2	2	4
0.15	0	4	0	21	0	7	0	4	2	2	3	2
0.20	0	4	0	21	0	7	3	2	2	2	2	2
0.25	2	3	0	21	2	4	2	2	0	3	2	2
0.30	2	3	2	4	2	4	0	2	0	2	0	2
0.35	2	3	2	2	0	7	0	2	0	2	0	2
0.40	5	3	2	2	5	3	0	2	0	2	0	3
0.45	5	3	0	4	6	3	0	2	0	2	7	0
0.50	28	2	6	3	9	0	0	0	7	0	7	0
0.55	14	2	0	0	2	2	0	0	0	0	0	3
0.60	0	8	13	2	2	2	0	0	0	12	0	2
0.65	0	8	2	11	2	2	5	3	0	3	0	2
0.70	0	9	0	12	3	2	5	3	0	2	0	2
0.75	0	9	0	12	0	2	5	3	2	2	2	2
0.80	0	15	0	9	0	2	3	2	2	2	2	2
0.85	0	8	0	12	0	2	2	3	3	2	3	2
0.90	0	9	2	9	0	3	0	12	3	2	3	2
0.95	2	8	0	12	0	10	3	7	0	15	0	14

Table 4. 1-step and averaged over 24-steps CFS optimization results for the IQR model of Equation (23).

	Low Var.		**Med Var.**		**High Var.**		**Low Var.**		**Med Var.**		**High Var.**	
	m	n	m	n	m	n	m	n	m	n	m	n
α_i	**1-step optimization of IQR model**						**24-steps optimization of IQR model**					
0.05	2	4	0	9	0	5	2	6	2	4	2	4
0.10	0	4	0	5	0	4	2	3	3	2	2	3
0.15	0	3	0	5	0	4	2	3	2	2	3	2
0.20	0	4	0	5	0	4	3	2	2	2	2	2
0.25	2	3	0	4	0	4	2	2	0	3	2	2
0.30	2	3	0	4	2	3	2	2	0	2	0	2
0.35	2	3	2	2	5	3	0	2	0	2	0	2

Table 4. *Cont.*

	Low Var.		**Med Var.**		**High Var.**		**Low Var.**		**Med Var.**		**High Var.**	
	m	n	m	n	m	n	m	n	m	n	m	n
α_i	**1-step optimization of IQR model**						**24-steps optimization of IQR model**					
0.40	5	3	2	2	5	3	0	2	0	2	0	3
0.45	5	3	0	3	6	3	0	2	0	2	7	0
0.50	28	2	6	3	9	0	0	0	7	0	7	0
0.55	14	2	0	0	2	2	0	0	0	0	0	3
0.60	11	2	13	2	2	2	0	0	0	7	0	2
0.65	2	4	0	4	2	2	5	2	0	3	0	2
0.70	0	11	0	7	3	2	5	3	0	2	0	2
0.75	0	11	0	7	0	2	5	2	2	2	2	2
0.80	0	11	0	7	0	2	3	2	2	2	2	2
0.85	0	11	0	2	0	2	2	3	3	2	3	2
0.90	0	11	0	12	0	3	0	9	3	2	3	2
0.95	2	7	0	12	4	2	5	4	0	7	0	4

Table 5. 1-step and averaged over 24-steps CFS optimization results for the Q05 model of Equation (23).

	Low Var.		**Med Var.**		**High Var.**		**Low Var.**		**Med Var.**		**High Var.**	
	m	n	m	n	m	n	m	n	m	n	m	n
α_i	**1-step optimization of Q05 model**						**24-steps optimization of Q05 model**					
0.05	0	18	48	48	2	10	0	14	0	14	7	4
0.10	0	14	15	4	2	10	0	12	3	6	3	9
0.15	0	11	25	3	2	7	0	8	0	7	0	7
0.20	0	11	35	6	2	7	0	6	0	7	0	7
0.25	0	11	24	3	0	11	0	6	0	6	0	5
0.30	0	17	24	2	2	7	0	6	0	3	0	5
0.35	2	11	24	3	2	7	0	6	0	3	0	5
0.40	2	17	25	2	0	8	0	6	0	3	0	7
0.45	2	16	24	3	2	5	0	6	0	3	5	3
0.50	2	16	13	12	6	2	0	6	0	2	5	3
0.55	21	0	0	0	6	2	2	5	0	0	5	3
0.60	21	0	2	14	9	0	0	0	7	0	0	18
0.65	17	0	2	14	0	7	2	5	7	0	0	16
0.70	16	0	2	11	0	7	2	5	7	0	0	15
0.75	16	0	2	11	0	5	2	5	0	18	0	15

Table 5. *Cont.*

	Low Var.		**Med Var.**		**High Var.**		**Low Var.**		**Med Var.**		**High Var.**	
	m	n	m	n	m	n	m	n	m	n	m	n
α_i	**1-step optimization of Q05 model**						**24-steps optimization of Q05 model**					
0.80	15	0	2	12	0	5	2	5	0	14	0	15
0.85	16	0	0	5	0	5	14	5	0	14	0	15
0.90	12	0	0	5	0	4	14	7	0	14	0	15
0.95	12	0	0	4	0	4	0	4	0	11	0	3

Table 6. 1-step and averaged over 24-steps CFS optimization results for the Q95 model of Equation (23).

	Low Var.		**Med Var.**		**High Var.**		**Low Var.**		**Med Var.**		**High Var.**	
	m	n	m	n	m	n	m	n	m	n	m	n
α_i	**1-step optimization of Q95 model**						**24-steps optimization of Q95 model**					
0.05	0	4	0	6	0	4	0	4	0	4	0	8
0.10	0	6	0	6	0	4	0	6	0	5	0	5
0.15	0	6	0	6	0	5	0	6	0	5	0	5
0.20	0	8	0	6	0	5	0	10	0	10	0	8
0.25	0	9	2	18	0	7	0	10	0	10	0	8
0.30	0	10	2	13	0	7	0	10	5	19	2	8
0.35	0	45	2	13	0	7	17	0	3	22	2	8
0.40	0	45	12	14	8	0	4	6	6	3	3	7
0.45	21	0	16	11	8	0	4	6	6	3	6	2
0.50	21	0	32	3	9	0	4	6	5	3	6	2
0.55	0	18	0	0	9	0	0	0	5	3	0	6
0.60	0	18	0	0	0	10	3	7	5	3	0	6
0.65	0	11	0	16	0	9	2	8	5	3	0	4
0.70	2	11	2	9	0	9	3	7	3	5	0	4
0.75	0	15	2	9	0	11	2	8	3	5	0	4
0.80	0	15	2	9	0	11	0	12	3	5	0	6
0.85	3	7	2	9	0	11	0	12	3	5	0	6
0.90	3	7	2	10	2	11	0	12	3	6	0	10
0.95	2	13	3	10	2	11	2	11	2	9	0	14

6. Out-of-Sample Forecast Performance Results

In this section, after fitting the four optimized models of Equation (23) to the whole two years in-sample learning set (for each farm), we will produce quantile forecasts from 15 min up to six hours ahead from each point of the out-of-sample forecasting set, and assess their forecast performance using the CFS, and the CRPS. In short, the CFS will be used to assess the skill of individual quantile forecasts, and the CRPS to assess the skill of the density forecasts (produced by using all 19 quantile forecasts).

In order to facilitate the comparison of forecast performance across different models, we will introduce two widely used *probabilistic benchmarks*:

- *Persistence distribution*: It is defined as the distribution of the last n observations. The persistence benchmark is independently optimized (by estimating n) for each wind farm, by using the same optimization methods as for the variability indices: 1-step ahead CFS minimization, and averaged over 24-steps CFS minimization. So, when the persistence is optimized using one of the two CFS minimization methods, different values of n are chosen to forecast each quantile.
- *Unconditional distribution*: We construct this benchmark by using all the past observations of the time series. This benchmark assumes that the time ordering of the observations is not relevant when attempting to predict the distribution of the response. It is also referred to as *climatology*.

The third benchmark used in this article is the quantile regression model with only the three lags of wind power series as explanatory variables. This benchmark will help us to identify the gain in forecast performance acquired by using the four variability indices and is defined as the *3-lagged series* benchmark.

Predictive distributions are often taken to be Gaussian even though the wind power series is bounded and non-negative. Moreover, in our record of wind power measurements we have values of exactly 0 and 1 and hence the predictive distributions may require point masses at 0 and 1. A convenient way to embed this property is through the use of cut-off normal predictive distributions as achieved by Sanso and Guenni [33], Allcroft and Glasbey [34], Gneiting *et al.* [35] and Pinson [10]. The fourth benchmark of this article uses a cut-off normal predictive density, $\mathcal{N}^{0,1}(\mu_{t+k|t}, \sigma^2_{t+k|t})$, and a fitted diurnal trend component to the three wind power series. The parameters $\mu_{t+k|t}$ and $\sigma_{t+k|t} > 0$ for $k = 1, ..., 24$ are called the location parameter (or predictive centre) and scale parameter (or predictive spread) of the cut-off normal density with point masses at 0 and 1. Please note that a truncated normal predictive distribution (with cut-offs at 0 and 1) has also been considered, with results very similar but worse than those of the cut-off normal predictive distribution benchmark.

The procedure to construct the fourth benchmark used in this article (also described in Gneiting *et al.* [35] and Gneiting *et al.* [8]) is as follows. At each of the three sites we firstly fit a trigonometric function,

$$y_t = a_0 + a_1 \sin\left(\frac{2\pi d(t)}{96}\right) + a_2 \cos\left(\frac{2\pi d(t)}{96}\right) + a_3 \sin\left(\frac{4\pi d(t)}{96}\right) + a_4 \cos\left(\frac{4\pi d(t)}{96}\right) \quad (24)$$

where y_t represents the normalised wind power for each farm at time t, and $d(t)$ is a repeating function that numbers the time variable (in 15 min steps) from 1 to 96 within each day. We then remove the

ordinary least square (OLS) fit from each wind power series and use the resulting residual series, denoted by ϵ_t^r, to determine the predictive centre and predictive mean of the cut-off normal predictive distribution.

More specifically we introduce the following linear autoregressive system

$$\epsilon_t^r = b_0 + b_1\epsilon_{t-1}^r + b_2\epsilon_{t-2}^r + b_3\epsilon_{t-3}^r \tag{25}$$

and use this to determine the forecasts $\hat{\epsilon}_{t+k|t}^r$ in a straightforward way, for each $k = 1, ..., 24$ (from 15 min up to 6 h ahead). Then, the predictive centre of the cut-off normal distribution is modelled as

$$\mu_{t+k|t} = \hat{y}_{t+k|t} + \hat{\epsilon}_{t+k|t}^r \tag{26}$$

where $\hat{y}_{t+k|t}$ is the forecast issued at time t with forecast horizon k for the fitted diurnal trend of Equation (24).

Finally, in order to model the predictive spread we introduce, following Gneiting *et al.* [8], the volatility function at time t:

$$v_t = \left(\frac{1}{2}\sum_{i=0}^{1}(y_{t-i} - y_{t-i-1})^2\right)^{\frac{1}{2}} \tag{27}$$

So this benchmark allows for conditional heteroskedasticity by modelling

$$v_t = c_0 + c_1 v_{t-1} \tag{28}$$

and setting the predictive spread as the forecast of v_t issued at time t for a forecast time $t + k$:

$$\sigma_{t+k|t} = \hat{v}_{t+k|t}. \tag{29}$$

These four benchmarks will be used as the reference models mentioned in Section 4.2. In the following tables we will present the evaluation results of the four models, for each evaluation criterion and optimization type. As the relative performances of the methods are similar for each of the three locations, following Taylor *et al.* [6], we will present the averaged results over the three wind farms. Moreover, we will present only the Skill and Average Skill Scores of each evaluation criterion, as we are particularly interested to quantify and statistically test (using the Anisano–Giacomini test [30]) the relative increase in forecast performance of the four competing models with respect to the four benchmarks (reference models).

6.1. Out-of-Sample Model Comparison and Evaluation-Quantile Forecasting

In this subsection we compare the out-of-sample forecast performance of the competing models for each quantile and model optimization method. We have a total of 19 quantile forecasts for each model and for two different optimization methods. Please note that in order to avoid presenting any unnecessary information, we summarise the results on the forthcoming tables by including results of 11 out of 19 quantiles (0.05, 0.10, 0.20,...,0.80, 0.90, 0.95 quantiles). Firstly, we present the results obtained using the *1-step ahead CFS* optimization, followed by the results obtained using the *averaged over 24-steps CFS* optimization. For both optimization methods, the scores will be averaged over the three wind farms because the relative performance of the models is similar across the wind farms.

6.1.1. Quantile Forecasting: 1-Step Ahead CFS Optimization

Since the models in this subsection are optimized using a 1-step ahead CFS optimization method, it makes sense to present results for the first lead time only, for each quantile and for each model. Table 7 shows the Skill CFS (as defined by Equation (19)) of the best performing model among the four competing ones and its percentage gain/loss with respect to the four reference (benchmark) models, for each quantile. Moreover, the asterisks next to the scores indicate the level of statistical significance (obtained using the Amisano–Giacomini test of Section 4.2) of the corresponding gain/loss in performance with respect to the four reference models.

Table 7. The best performing model among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for each quantile. Reference models: 3-lagged series (column 3), Cut-off normal (column 4), Persistence (column 5) and Climatology (column 6). These results are outcomes from a *1-step ahead CFS* optimization, and we use the CFS only for the first predicted step. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

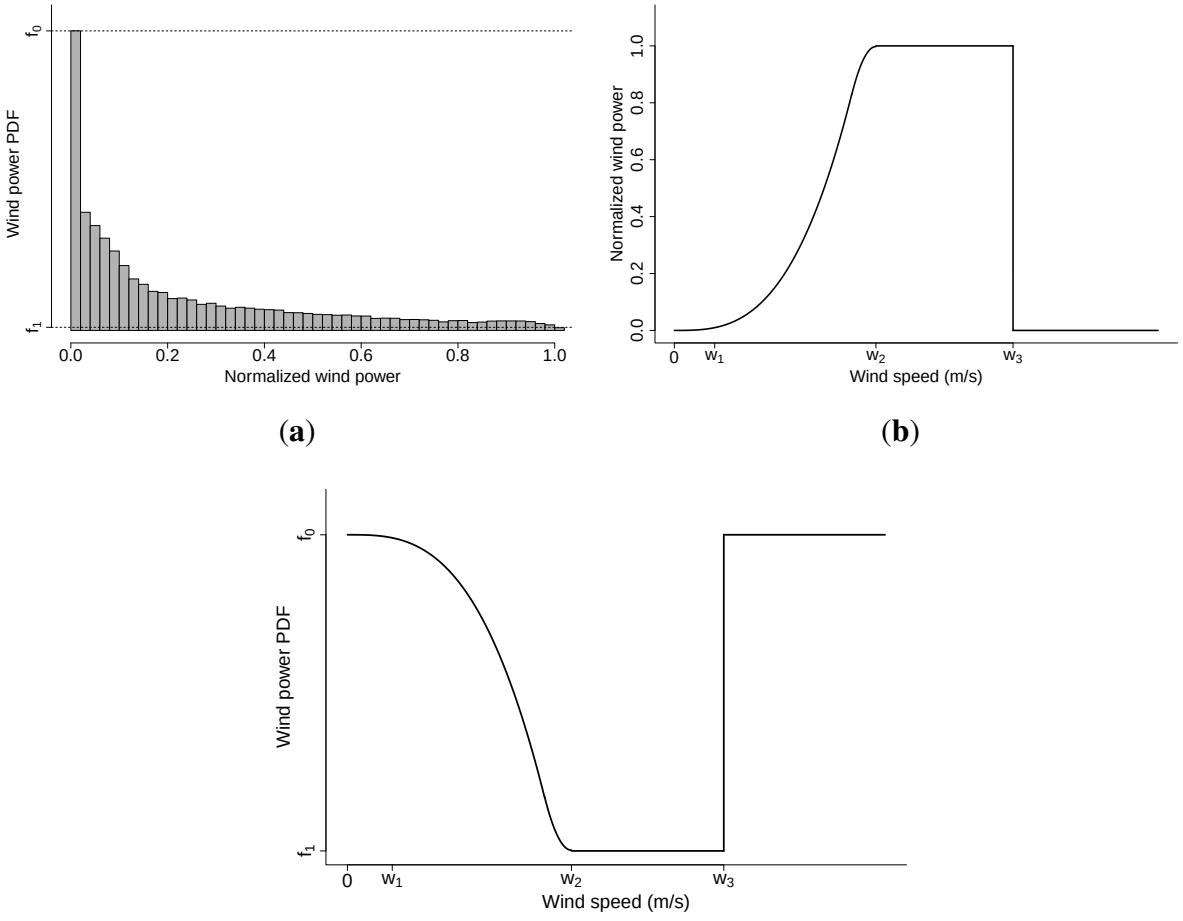
	Skill CFS (%)				
α_i	**Best model**	**3-lagged series**	**Cut-off normal**	**Persistence benchmark**	**Climatology benchmark**
0.05	Q95	4.01^{***}	88.04^{***}	54.82^{***}	65.63^{***}
0.10	Q95	2.57^{***}	81.70^{***}	53.25^{***}	71.89^{***}
0.20	Q95	1.23^{**}	73.88^{***}	55.61^{***}	78.14^{***}
0.30	SD	0.81	69.53^{***}	57.81^{***}	81.56^{***}
0.40	SD	0.22	67.07^{***}	59.18^{***}	83.70^{***}
0.50	Q95	-0.21	65.93^{***}	59.82^{***}	85.18^{***}
0.60	Q05	-0.15	65.84^{***}	59.53^{***}	86.17^{***}
0.70	Q05	0.59	67.21^{***}	58.62^{***}	86.79^{***}
0.80	SD	2.74^{***}	71.12^{***}	57.57^{***}	87.20^{***}
0.90	Q05	2.38^{***}	78.00^{***}	54.48^{***}	86.46^{***}
0.95	Q05	3.44^{***}	84.64^{***}	54.46^{***}	85.05^{***}

A general observation is that for almost all quantiles (except the 0.50–0.60 quantiles scores which have negative signs), the best forecast performance is achieved by one of the four competing models and not by the four benchmarks. The 0.05–0.10 and 0.90–0.95 quantiles form the two tails of the predictive density, and represent the rare events (such as ramps, cut offs) of a wind power series. As we observe from this table, both tails of the predictive density are quite well captured by the Q05 and Q95 models. Out of the four competing models, the lower tail of the predictive density is better predicted by the Q95

model and the upper by the Q05 model, but assuming the structure of the two variability indices used in these models, we might intuitively expect the opposite to happen.

This phenomenon can be explained by having a look at Figure 3. Figure 3(a) shows the probability density function (PDF) of the medium variability wind farm, together with the function values when the normalized wind power is equal to zero and one. Figure 3(b) shows an example of a wind power curve as presented by McSharry *et al.* [3]. On this plot we mark the "cut-in speed" (w_1), the "nominal speed" (w_2) and the "disconnection speed" (w_3). So, for very low wind speeds ($<w_1$) the wind power production is almost zero, for wind speeds greater than w_2 but less than w_3 the normalized wind power production is equal to one, and for wind speeds greater than w_3 the turbines shut down in order to prevent damage, and hence the wind power production falls again to zero. By combining these two plots, we can plot a rough estimate of the normalized wind power PDF versus the wind speed.

Figure 3. (**a**) Histogram of normalized wind power data (**b**) Deterministic power curve (**c**) Probability density function of normalized wind power data (PDF) *versus* wind speed. The equations used to reproduce (**b**) and (**c**) were taken from McSharry *et al.* [3].



We expect the 0.95 quantile of the unconditional density (not to be confused with the predictive density) to be close to the nominal (normalized) wind power value of one. But the produced wind power is driven by the actual wind speed at any given time, and hence falling from the nominal wind power

production (one) to zero can happen unexpectedly (Exceeding w_3 can happen unexpectedly, given that we do not have any information about the wind speed at any given time.) if we exceed the disconnection speed w_3 (Figure 3(b)). This results in a sudden jump from the 0.95 quantile to the 0.05 quantile of the unconditional probability density and is represented by the lower tail of the predictive density. Hence, the Q95 index which captures this sort of events can provide some extra information about the lower tail of the predictive distribution that the 3-lagged series and the Q05 index do not describe. Similarly, if the wind speed falls below w_3, we are suddenly jumping from the 0.05 to the 0.95 quantile of the unconditional density and these kind of rare events (jumping from low to high values) are represented by the upper tail of the predictive density. Therefore, the Q05 index can provide some extra information about the upper tail of the conditional predictive distribution.

In addition, Table 7 shows that the strongest benchmark for all quantiles is the 3-lagged series model. The biggest and statistically significant improvement with respect to this benchmark is achieved near the tails of the predictive density, and decays as we move towards the median. This has important practical applications because it is exactly these extreme fluctuations that are of interest to transmission system operators (TSOs). More specifically, we get a performance gain of up to 4.01% (for the 0.05 quantile, achieved by the Q95 model), which is certainly not negligible. Unfortunately, the performance gain by using one of the competing models with respect to this benchmark in order to forecast the quantiles 0.30–0.70 is negligible (statistically insignificant), and is in the range of -0.21% to 0.81%. Furthermore, there is no gain for the quantiles close to the median of the predictive density (0.50–0.60).

Since the 3-lagged series model is our strongest benchmark and the performance gain with respect to it is only worth mentioning near the tails of the predictive density, it makes sense to focus on the performance gain achieved with respect to the last two benchmarks only for quantiles near the tails. Table 7 shows that the increase in forecast performance with respect to the cut-off normal benchmark is at least 78%, which shows that the cut-off normal is not capturing the tails of the predictive distribution as well as our competing models.

Moreover, we get more than 53.25% increase in the forecast performance with respect to the persistence benchmark when we use one of the four competing models. At the tails, where the Q05 and Q95 models are more suitable, we have a gain with respect to the persistence benchmark of up to 54.82%. By using the climatology benchmark as reference model, we observe that the maximum performance gain at the tails goes up to 86.46% (for the 0.90 quantile, achieved by the Q05 model), and in general the Q05 and Q95 models manage to maintain the performance gain (at the tails) above 65.63%.

6.1.2. Quantile Forecasting: Averaged over 24-Steps CFS Optimization

This subsection has similar structure to the preceding one, but now we present the results for the models which minimize the averaged (over six hours) CFS. Because the models are optimized on their forecast behaviour over all 24 forecast horizons, it makes sense to present results with the scores averaged over the 24 horizons, for each of the 11 selected quantiles.

Table 8 is analogous to Table 7, but here we provide the averaged over 24-steps CFS optimization results. It presents the Average Skill CFS (instead of Skill CFS) as defined by Equation (20). Once more, a general observation is that for almost all quantiles (except the first two and the median), the best

forecast performance is achieved by our competing models and the strongest benchmark is the 3-lagged series model.

Table 8. The best performing model among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for each quantile. Reference models: 3-lagged series (column 3), Cut-off normal (column 4), Persistence (column 5) and Climatology (column 6). These results are outcomes from an *averaged over 24-steps CFS* optimization, and the CFS are also averaged over all 24 forecast horizons. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

	Average Skill CFS (%)				
α_i	**Best model**	**3-lagged series**	**Cut-off normal**	**Persistence benchmark**	**Climatology benchmark**
0.05	Q95	0.71	43.82^{***}	-102.14^{***}	-103.89^{***}
0.10	IQR	1.99^{**}	31.81^{***}	-29.08^{***}	-29.91^{***}
0.20	IQR	5.95^{***}	22.46^{***}	14.54^{***}	21.00^{***}
0.30	SD	4.57^{***}	15.25^{***}	18.86^{***}	38.16^{***}
0.40	IQR	2.56^{***}	10.22^{***}	20.47^{***}	46.64^{***}
0.50	Q95	-0.11	6.03^{***}	20.13^{***}	50.95^{***}
0.60	Q05	0.07	5.27^{***}	20.05^{***}	53.81^{***}
0.70	IQR	1.67^{***}	7.15^{***}	19.48^{***}	54.67^{***}
0.80	SD	3.71^{***}	11.94^{***}	17.86^{***}	52.16^{***}
0.90	SD	2.66^{***}	19.55^{***}	12.86^{***}	38.12^{***}
0.95	Q05	1.09^{**}	28.78^{***}	8.29^{***}	11.94^{***}

Table 8 also shows that the lower tail of the predictive density is quite poorly captured by our competing models, and the last two benchmarks (persistence, climatology) are performing much better than any other model. On the other hand, for all the other quantiles, the SD and IQR models have quite similar performances and manage to outperform all the benchmarks. Moreover, the performance gain by using one of the four competing models to forecast the quantiles near the median (0.50–0.60) is statistically negligible or does not exist. A final general observation is that, as mentioned for the previous optimization method, the 0.05 quantile is better predicted by the Q95 model and the 0.95 quantile by the Q05 model.

By using one of the SD or IQR models (which perform almost identically) we get a performance gain with respect to the 3-lagged series benchmark of up to 5.95% (for the 0.20 quantile, achieved by the IQR model), which is statistically significant with a p-value less than 0.001. In addition all the competing models are outperforming the cut-off normal model by at least 5.27% and attain the maximum increase in forecast performance near the tails of the predictive density (up to 43.82% achieved by the Q95 model for the 0.05 quantile).

Table 8 also shows that we have up to 20.47% (for the 0.40 quantile, achieved by the IQR model) increase of the forecast performance with respect to the persistence benchmark. The SD and IQR models maintain the gain over the persistence benchmark above 8.29% for all quantiles larger than 0.10. By using the climatology benchmark as reference model, we observe that the maximum performance gain goes up to 54.67% (for the 0.70 quantile, achieved by the IQR model), and in general the SD and IQR models can maintain the percentage performance gain with respect to the climatology benchmark above 11.94% for all quantiles larger than 0.10.

6.2. Out-of-Sample Model Comparison and Evaluation-Density Forecasting

In this subsection, we evaluate the out-of-sample density forecast performance of the competing models, for each optimization method. We will use the quantile forecasts obtained from each optimization method to reconstruct the whole predictive density, and assess its skill using the Skill CRPS or the Average Skill CRPS. Firstly, we present the results obtained using the *1-step ahead CFS* optimization, followed by the results obtained using the *averaged over 24-steps CFS* optimization. Moreover, because the relative performance of the models is similar across the wind farms, the scores will be averaged over the three wind farms.

6.2.1. Density Forecasting: 1-Step Ahead CFS Optimization

In this subsection, the models' forecast performance is optimized for only the first predicted step, so it makes sense to focus (initially) on the first lead time and present the out-of-sample Skill CRPS for the first step ahead.

Table 9 presents the out-of-sample Skill CRPS (%) for the *1-step ahead CFS* optimized models, together with significance codes for the Amisano–Giacomini test of equal forecast performance. This table shows that the best benchmark model is the 3-lagged series. That was expected because this benchmark was also the strongest one (for most quantiles) when we were looking at the quantile forecast results for the same optimization method (Section 6.1.1). The SD and IQR models behave almost identically and manage to outperform all the other benchmarks. The SD model performs slightly better than the IQR model, and managed to outperform the 3-lagged series model by 1%, the cut-off normal model by 1.48%, the persistence benchmark by 58.38% and the climatology benchmark by 84.23%.

Table 10 shows the best performing model among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for a collection of forecast horizons. For simplicity, we present the results for seven of the 24 forecast horizons. The SD model is outperforming the 3-lagged series for the first 16 forecast horizons (except for the second one) where the improvements in forecast performance are also statistically significant for a 90% significance level. For the second forecast horizon we get the maximum forecast performance gain over the 3-lagged series model (equal to 1.96%) achieved by the IQR model. The SD model also manages to outperform the cut-off normal benchmark for all forecast horizons, with all improvements in forecast performance being statistically significant for a 99% significance level.

Table 9. Out-of-sample Skill CRPS (%) (averaged over all wind farms) for the *1-step ahead CFS* optimized models. The scores are just for the first lead time. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

Performance Gain/Loss - Skill CRPS (%)				
Reference model	**SD model**	**IQR model**	**Q05 model**	**Q95 model**
3-lagged series	1.00^{**}	0.93^{**}	0.29	0.20
Cut-off normal	1.48^{***}	1.41^{***}	0.77^{*}	0.69^{*}
Persistence	58.38^{***}	58.35^{***}	58.08^{***}	58.04^{***}
Climatology	84.23^{***}	84.22^{***}	84.12^{***}	84.11^{***}

Table 10. The best performing model among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for forecast horizon k (measured in 15 min steps). Reference models: 3-lagged series (column 3), Cut-off normal (column 4), Persistence (column 5) and Climatology (column 6). These results are outcomes from a *1-step ahead CFS* optimization. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

Skill CRPS(%)					
k	**Best model**	**3-lagged series**	**Cut-off normal**	**Persistence**	**Climatology**
1	SD	1.00^{**}	1.48^{***}	58.38^{***}	84.23^{***}
2	IQR	1.96^{***}	6.79^{***}	44.59^{***}	76.78^{***}
3	SD	1.81^{***}	13.02^{***}	37.00^{***}	71.34^{***}
4	SD	1.71^{***}	14.09^{***}	31.97^{***}	66.73^{***}
8	SD	1.15^{***}	13.95^{***}	20.88^{***}	51.62^{***}
16	SD	0.63^{*}	15.12^{***}	10.94^{***}	27.55^{***}
24	SD	0.40	15.70^{***}	5.87^{***}	8.24^{***}

When the persistence and climatology benchmarks are used as reference models, Table 10 shows that the gain in forecast performance by using the SD model is at least 5.87% and 8.24%, respectively. Moreover, the noted density forecast improvements are statistically significant for all forecast horizons, for a 99% level of significance.

In addition to the above results, we carried out a marginal calibration analysis and investigated how the CRPS evolves conditional to some wind power levels. More specifically, Table 11 presents the marginal Skill CRPS (%) conditional to the normalized wind power being $\leq$0.20 or $\geq$0.80, for a collection of seven forecast horizons. We choose to focus on these specific wind power levels, because these form the two tails of the unconditional wind power density (not to be confused with the predictive density).

Table 11. The best performing model (according to the Marginal Skill CRPS) among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for forecast horizon k (measured in 15 min steps). Reference models: 3-lagged series (column 3), Cut-off normal (column 4), Persistence (column 5) and Climatology (column 6). These results are outcomes from a *1-step ahead CFS* optimization. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

Marginal Skill CRPS(%) conditional to the normalized wind power being $\leq$ 0.20					
k	**Best model**	**3-lagged series**	**Cut-off normal**	**Persistence**	**Climatology**
1	Q05	1.32^{***}	3.44^{***}	60.21^{***}	84.09^{***}
2	IQR	1.81^{***}	7.44^{***}	44.92^{***}	75.51^{***}
3	IQR	1.83^{***}	11.68^{***}	36.81^{***}	69.40^{***}
4	IQR	1.63^{***}	13.28^{***}	31.48^{***}	64.29^{***}
8	IQR	1.04^{**}	14.71^{***}	20.74^{***}	47.64^{***}
16	IQR	0.56	17.94^{***}	11.58^{***}	20.09^{***}
24	IQR	0.36	19.27^{***}	6.40^{***}	-4.30^{***}
Marginal Skill CRPS(%) conditional to the normalized wind power being $\geq$ 0.80					
1	IQR	4.95^{***}	1.31^{***}	57.74^{***}	93.52^{***}
2	IQR	3.04^{***}	13.51^{***}	48.16^{***}	91.11^{***}
3	SD	3.44^{***}	18.99^{***}	44.76^{***}	89.57^{***}
4	SD	2.40^{***}	22.23^{***}	41.11^{***}	87.88^{***}
8	SD	2.25^{***}	19.97^{***}	32.79^{***}	81.42^{***}
16	SD	1.60^{***}	17.60^{***}	23.59^{***}	68.09^{***}
24	SD	1.75^{***}	15.42^{***}	15.47^{***}	53.93^{***}

Given that the normalized wind power is less than or equal to 0.20, the IQR seems to be the best performing model for all except the first forecast horizon (where the Q05 model is performing better). For small forecast horizons we observe statistically significant improvements over all competing models. These improvements (with the exception of the cut-off normal benchmark) are getting smaller as we move to larger forecast horizons, which is perfectly reasonable because the results are the outcome of a 1-step ahead optimization. Conditioning on power levels which belong to the upper tail of the unconditional wind power density, we observe that the IQR and SD models seem to provide the largest performance gain according to the CRPS. These two models are outperforming all the benchmarks with improvements that are also statistically significant for all forecast horizons, for a 99% level of significance.

6.2.2. Density Forecasting: Averaged over 24-Steps CFS Optimization

Now we would like to assess the out-of-sample density forecast performance of the four competing models, for the *averaged over 24-steps CFS* optimization method. Our assessment criterion will be the out-of-sample Skill CRPS or Average Skill CRPS.

Initially, it makes sense to have a look at the out-of-sample Average Skill CRPS (Equation (20)) with the four benchmarks as reference models (Table 12). The IQR model outperforms the 3-lagged series benchmark by 2.45%, a considerably larger improvement than for the 1-step ahead results given in Table 9. This model also outperforms the cut-off normal benchmark by 16.13%, the persistence benchmark by 12.77% and the climatology benchmark by 39.15%. Moreover, the density forecast performance of the SD model is quite close to that of the IQR model.

Table 12. Out-of-sample Average Skill CRPS (%) (also averaged over all wind farms) for the *averaged over 24-steps CFS* optimized models. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

Performance Gain/Loss - Average Skill CRPS (%)				
Reference model	**SD model**	**IQR model**	**Q05 model**	**Q95 model**
3-lagged series	2.37^{***}	2.45^{***}	-0.07	0.42
Cut-off normal	16.06^{***}	16.13^{***}	13.96^{***}	14.38^{***}
Persistence	12.70^{***}	12.77^{***}	10.51^{***}	10.96^{***}
Climatology	40.91^{***}	40.96^{***}	39.43^{***}	39.73^{***}

Since our optimization considers all 24 forecast horizons, it will be interesting to investigate how the four competing models perform in producing density forecasts for each forecast horizon, k, from 15 min up to 6 h ahead. As for the 1-step ahead optimization case, we present the results for only a collection of seven out of 24 forecast horizons. The best competing model together with the performance gain obtained for each forecast horizon k with respect to the four benchmarks can be found in Table 13. Clearly, the best performing benchmark is the 3-lagged series model, and the IQR is the best performing model out of the four competing models.

The competing models' performances are disappointing for the first lead time, where the 3-lagged series benchmark offers a performance gain (of at least 1.57%) with respect to these models. On the other hand, for predictions larger than 30 minutes ahead (second predicted step), Table 13 shows that the IQR model manages to maintain the gain in density forecast performance with respect to the 3-lagged series model above 2.14%, with a recorded maximum of 3.59% (achieved at the fourth predicted step). Moreover, all the scores (except the first two) produce p-values which give strong evidence to reject the null hypothesis of equal forecast performance between the competing and reference model. Hence, the observed gain in forecast performance is statistically significant for a 99% significance level. The gain

in forecast performance with respect to the cut-off normal model is at least 4.96% (excluding the first lead time) and attains a maximum of 17.18% for the 24th predicted step.

Table 13. The best performing model among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for forecast horizon k (measured in 15 min steps). Reference models: 3-lagged series (column 3), Cut-off normal (column 4), Persistence (column 5) and Climatology (column 6). These results are outcomes from an *averaged over 24-steps CFS* optimization. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

		Skill CRPS(%)			
k	**Best model**	**3-lagged series**	**Cut-off normal**	**Persistence**	**Climatology**
1	Q95	-1.57^{***}	-1.07^{**}	59.08^{***}	83.82^{***}
2	IQR	0.03	4.96^{***}	44.84^{***}	76.32^{***}
3	IQR	3.28^{***}	14.33^{***}	38.53^{***}	71.77^{***}
4	IQR	3.59^{***}	15.74^{***}	33.21^{***}	67.37^{***}
8	IQR	3.34^{***}	15.86^{***}	20.02^{***}	52.69^{***}
16	IQR	2.59^{***}	16.79^{***}	6.80^{***}	28.98^{***}
24	IQR	2.14^{***}	17.18^{***}	-0.44	9.85^{***}

If we consider the persistence benchmark as the reference model (column 4 of Table 13), we note that the Skill CRPS of the best model starts at 59.08% (Q95 model) and then decays to meet approximately the performance of the persistence benchmark for the last forecast horizon. When the climatology benchmark is used as a reference model (column 5 of Table 13), we again observe a decay of the skill scores, with the performance gain remaining above 9.85% for all forecast horizons (for the IQR model).

From the results presented we conclude that this optimization method is found to produce models (mainly the IQR model) that can substantially outperform the density forecast performance of the widely used benchmarks (persistence, climatology) and fully parametric models such us the cut-off normal benchmark. Moreover the gain used by including a variability index such as the (IQR) improves considerably the performance (up to 3.59%) of a quantile regression model which uses only autoregressive terms as explanatory variables (3-lagged series benchmark).

Finally, as for the 1-step ahead optimization case, we present some marginal calibration analysis results by investigating how the CRPS evolves conditional to some wind power level. Table 14 presents the marginal Skill CRPS(%) conditional to the normalized wind power being $\leq$0.20 or $\geq$0.80, for a collection of seven forecast horizons.

Given that the normalized wind power is less or equal to 0.20, the IQR model is outperforming all the benchmarks for forecast horizons larger than two steps ahead (except the last forecast horizon of the persistence and climatology benchmarks). The Q05 seems to be the best performing model for the first two steps ahead, but still cannot outperform the 3-lagged series benchmark for the first step ahead.

The second part of this table shows that, given normalized power levels greater or equal to 0.80, the SD model is the overall best model among all the others. It manages to outperform all the benchmarks for all forecast horizons, with improvements that are also statistically significant using a 99% level of significance.

Table 14. The best performing model (according to the Marginal Skill CRPS) among the four competing ones, and its performance gain/loss with respect to the four reference (benchmark) models, for forecast horizon k (measured in 15 min steps). Reference models: 3-lagged series (column 3), Cut-off normal (column 4), Persistence (column 5) and Climatology (column 6). These results are outcomes from an *averaged over 24-steps CFS* optimization. The asterisks indicate the statistical significance of the gain/loss according to the Amisano and Giacomini test with the following significance codes for the p-value of the test: ***: $p \leq 0.01$, **: $0.01 < p \leq 0.05$, *: $0.05 < p \leq 0.1$.

Marginal Skill CRPS(%) conditional to the normalized wind power being $\leq$ 0.20					
k	**Best model**	**3-lagged series**	**Cut-off normal**	**Persistence**	**Climatology**
1	Q05	-0.13	2.02^{***}	59.10^{***}	83.86^{***}
2	Q05	1.20^{***}	6.87^{***}	42.75^{***}	75.36^{***}
3	IQR	3.45^{***}	13.14^{***}	34.89^{***}	69.90^{***}
4	IQR	3.93^{***}	15.32^{***}	29.05^{***}	65.13^{***}
8	IQR	3.81^{***}	17.10^{***}	15.50^{***}	49.10^{***}
16	IQR	3.12^{***}	20.06^{***}	1.77^{***}	22.15^{***}
24	IQR	2.71^{***}	21.18^{***}	-6.61^{***}	-1.84^{***}
Marginal Skill CRPS(%) conditional to the normalized wind power being $\geq$ 0.80					
1	SD	6.27^{***}	2.68^{***}	66.58^{***}	93.61^{***}
2	SD	4.92^{***}	15.18^{***}	58.01^{***}	91.28^{***}
3	SD	1.78^{***}	17.60^{***}	52.60^{***}	89.39^{***}
4	SD	1.33^{***}	21.38^{***}	49.18^{***}	87.75^{***}
8	SD	1.95^{***}	19.73^{***}	39.04^{***}	81.36^{***}
16	SD	1.95^{***}	17.89^{***}	27.42^{***}	68.20^{***}
24	SD	2.13^{***}	15.75^{***}	17.99^{***}	54.11^{***}

7. Conclusions

In this paper we showed how to produce wind power quantile and density forecasts, for lead times from 15 minutes up to six hours ahead, using three different univariate wind power series. This was achieved by introducing innovative variability indices, which are able to capture the volatile behaviour of the wind power series.

We used linear (in parameters) quantile regression as our main tool for producing quantile forecasts for 19 different quantiles, with three lagged versions of the wind power series as the main explanatory

variables. Four models were proposed, each one having as a fourth explanatory variable one of the four extracted variability indices.

In order for the final results to be consistent, we used data from three wind farms in Denmark, each one chosen to have different wind power variability (low, medium and high). We investigated four years of wind power data, with a 15 min resolution, for each wind farm. The first two years were used for estimating the parameters of the models, and the final two years for out-of-sample forecast evaluation.

All four quantile regression models were optimized using the in-sample training data set, in order to find their specific set of indices' parameters, (m, n), which minimizes (i) the first lead time CFS and (ii) the Average CFS over all forecast horizons, for each individual quantile.

Our main goal was to evaluate how well these models performed compared with the cut-off normal, persistence and unconditional distribution (climatology) probabilistic benchmarks. It is worth mentioning that persistence is a strong yet simple benchmark for very short forecast horizons, and was optimized using the same cost (optimization) functions as the four regression models. The use of a cut-off normal benchmark provided a good comparison between a fully parametric model (as the cut-off normal model) and the non-parametric quantile regression models used in this article.

The fourth and strongest benchmark used was a quantile regression model with three lags of the original series as explanatory variables. The comparison of the competing models with this benchmark provides evidence of how useful our extracted variability indices are for forecasting wind power production. The individual (out-of-sample) quantile forecasts were evaluated using the Skill or Average Skill CFS for direct comparison between the competing models and the benchmarks. The density forecasts of the models were evaluated using the Skill or Average Skill CRPS.

In the following we summarize the quantile and density forecasts results found using the two different types of model optimization:

Quantile forecasting: 1-step ahead CFS optimization

- The best competing models are the Q05 and Q95 models, which outperform our best benchmark (3-lagged series) by a maximum of 3.44% (0.95 quantile) and 4.01% (0.05 quantile), respectively.
- The largest gain in performance with respect to the best benchmark is noticed when forecasting the quantiles which form the tails of the conditional predictive density. In addition, the Q05 model performs better for the upper tail, and the Q95 model for the lower tail.
- The best quantile regression models for each forecast horizon manage to maintain the performance gain with respect to the cut-off normal, persistence and climatology benchmarks above 65.73%, 53.25% and 65.63%, respectively.

Quantile forecasting: Averaged over 24-steps CFS optimization

- The SD and IQR models have the best quantile forecast performance, with similar CFS. They manage to maintain the performance gain with respect to the best benchmark (3-lagged series) above 1.99% for 11 out of 19 quantiles. The maximum Skill CFS is 5.95%, and is achieved by the IQR model for the 0.20 quantile.
- The SD and IQR models maintain the performance gain with respect to the cut-off normal, persistence and climatology benchmarks above 5.25%, 12.86% and 21.00%, respectively, for

15 out of 19 quantiles. The performance gain by using one of the two quantile models over the persistence and climatology benchmarks is much lower (or does not exist) for predicting the tails (0.05, 0.10,0.90, 0.95 quantiles) than for predicting the quantiles close to the median of the conditional density.

Density forecasting: 1-step ahead CFS optimization

- The best competing model is the SD model, which has almost equal density forecast performance with the IQR model. It manages to outperform the best benchmark (3-lagged series) by 1.00% (improvement which is statistically significant for a 95% significance level), for the first lead time. All four competing models manage to outperform the cut-off normal, persistence and climatology benchmarks by at least 0.69%, 58.04% and 84.11%, respectively, for the first lead time.
- Across all 24 forecast horizons, the average gain in forecast performance using the SD or IQR model with respect to the best benchmark is statistically significant (using a 90% significance level) for the first 16 forecast horizons. Moreover, these two models manage to outperform the cut-off normal persistence and climatology benchmarks by at least 1.48%, 5.87% and 8.24%, respectively.

Density forecasting: Averaged over 24-steps CFS optimization

- The IQR model is the best competing model, and manages to outperform the best benchmark (3-lagged series) by, on average (over all forecast horizons), 2.45%. It also outperforms the cut-off normal, persistence and climatology benchmarks by, on average, 16.13%, 12.77% and 40.96%, respectively.
- Across all 24 forecast horizons (excluding the first two lead times), the IQR model manages to maintain a performance gain over the best benchmark by more than 2.14%. Moreover, the noted improvements in density forecast performance are statistically significant for 22 out of 24 forecast horizons, for a 99% significance level.

Acknowledgements

The authors would like to thank Energinet.dk for data provision and support. This work has been partly supported by the European Commission under the SafeWind project (ENK7-CT2008-213740), Her Majesty's Government and an IBM Innovation Award.

References

1. Barton, J.; Infield, D. Energy storage and its use with intermittent renewable energy. *Energy Convers. IEEE Trans.* **2004**, *19*, 441–448.
2. Boyle, G. *Renewable Electricity and the Grid: The Challenge of Variability*; Earthscan: London, UK, 2007.
3. McSharry, P.; Pinson, P.; Gerard, R. *Methodology for the Evaluation of Probabilistic Forecasts*; SafeWind Report, 2009.
4. Brown, B.G.; Katz, R.W.; Murphy, A.H. Time series models to simulate and forecast wind speed and wind power. *J. Appl. Meteorol.* **1984**, *23*, 1184–1195.

5. Sanchez, I. Short term prediction of wind energy production. *Int. J. Forecast.* **2006**, *22*, 43–56.
6. Taylor, J.W.; McSharry, P.E.; Buizza, R. Wind power density forecasting using ensemble predictions and time series models. *IEEE Trans. Energy Convers.* **2009**, *24*, 775–782.
7. Moeanaddin, R.; Tong, H. Numerical evaluation of distributions in non-linear autoregression. *J. Time Series Anal.* **1990**, *11*, 33–48.
8. Gneiting, T.; Larson, K.; Westrick, K.; Genton, M.G.; Aldrich, E. Calibrated probabilistic forecasting at the stateline wind energy center. *J. Am. Stat. Assoc.* **2006**, *101*, 968–979.
9. Trombe, P.J.; Pinson, P.; Madsen, H. A general probabilistic forecasting framework for offshore wind power fluctuations. *Energies* **2012**, *5*, 621–657.
10. Pinson, P. Very short-term probabilistic forecasting of wind power with generalized logit-normal distributions. *J. R. Stat. Soc. C* **2012**, *61*, 555–576.
11. Pinson, P.; Kariniotakis, G. Conditional prediction intervals of wind power generation. *Power Syst. IEEE Trans.* **2010**, *25*, 1845–1856.
12. Sideratos, G.; Hatziargyriou, N. Probabilistic wind power forecasting using radial basis function neural networks. *Power Syst. IEEE Trans.* **2012**, *27*, 1788–1796.
13. Pinson, P.; Madsen, H. Ensemble-based probabilistic forecasting at Horns Rev. *Wind Energy* **2009**, *12*, 137–155.
14. Lau, A.; McSharry, P. Approaches for multi-step density forecasts with application to aggregated wind power. *Ann. Appl. Stat.* **2010**, *4*, 1311–1341.
15. Jeon, J.; Taylor, J.W. Using conditional kernel density estimation for wind power density forecasting. *J. Am. Stat. Assoc.* **2012**, *107*, 66–79.
16. Koenker, R.; Bassett, G. Regression quantiles. *Econometrica* **1978**, *46*, 33–50.
17. Bremnes, J. Probabilistic wind power forecasts using local quantile regression. *Wind Energy* **2004**, *7*, 47–54.
18. Nielsen, H.; Madsen, H.; Nielsen, T. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy* **2006**, *9*, 95–108.
19. Moller, J.; Nielsen, H.; Madsen, H. Time-adaptive quantile regression. *Proc. Windpower* **2008**, *52*, 1292–1303.
20. Pritchard, G. Short-term variations in wind power: Some quantile-type models for probabilistic forecasting. *Wind Energy* **2011**, *14*, 255–269.
21. Davy, R.; Milton, J.; Russell, C.; Coppin, P. Statistical downscaling of wind variability from meteorological fields. *Bound.-Layer Meteorol.* **2010**, *135*, 161–175.
22. Bossavy, A.; Girard, R.; Kariniotakis, G. Forecasting ramps of wind power production with numerical weather prediction ensembles. *Wind Energy* **2012**, *16*, 51–63.
23. Gneiting, T. Quantiles as optimal point forecasts. *Int. J. Forecast.* **2011**, *27*, 197– 207.
24. Milligan, M.; Schwartz, M.; Wan, Y. Statistical Wind Power Forecasting Models: Results of U.S. Wind Farms. In *Proceedings of 2004 American Meteorological Society Annual Meeting*, Seattle, WA, USA, 11–15 January 2004.
25. Hyndman, R.J.; Fan, Y. Sample quantiles in statistical packages. *Am. Stat.* **1996**, *50*, 361–365.
26. Koenker, R.; D'Orey, V. Computing regression quantiles. *Appl. Stat.* **1987**, *36*, 383–393.
27. Li, Y.; Zhu, J. L1-Norm Quantile Regression. *J. Comput. Graph. Stat.* **2008**, *17*, 1–23.

28. Matheson, J.; Winkler, R. Scoring rules for continuous probability distributions. *Manag. Sci.* **1976**, *22*, 1087–1096.
29. Gneiting, T.; Raftery, A. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
30. Amisano, G.; Giacomini, R. Comparing density forecasts via weighted likelihood ratio tests. *J. Bus. Econ. Stat.* **2007**, *25*, 177–190.
31. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
32. Chen, P.; Pedersen, T.; Bak-Jensen, B.; Chen, Z. ARIMA-based time series model of stochastic wind power generation. *Power Syst. IEEE Trans. Power Syst.* **2010**, *25*, 667–676.
33. Sanso, B.; Guenni, L. A nonstationary multisite model for rainfall. *J. Am. Stat. Assoc.* **2000**, *95*, 1089–1100.
34. Allcroft, D.J.; Glasbey, C.A. A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation. *J. R. Stat. Soc. C (Appl. Stat.)* **2003**, *52*, 487–498.
35. Gneiting, T.; Larson, K.; Westrick, K.; Genton, M.G.; Aldrich, E. *Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center: The Regime-Switching Space-Time (RST) Method*; Technical Report; Department of Statistics, University of Washington: Seattle, WA, USA, 2004.